Characterization of Pathogen-Driven Selection at *B4galnt2* in House Mice

Dissertation

in fulfillment of the requirements for the degree "Dr. rer. nat."

of the Faculty of Mathematics and Natural Sciences

at Kiel University

Submitted by

Marie Vallier

Kiel, May 2017

First referee: Pr.Dr. John Baines
Second referee: Pr.Dr. Hinrich Schulenburg

Date of the oral examination: 12 May 2017
Approved for publication: 12 May 2017

Table of Contents

Zusammenfassung

B4galnt2 ist eine Blutgruppen-assoziierte Glykosyltransferase, die in Hausmäusen *cis*-regulatorische Variation hinsichtlich der gewebsspezifischen Expression aufweist. Das Wildtypallel, das beispielsweise im Laborstamm C57BL/6J auftritt, führt zur Expression von *B4galnt2* im Intestinaltrakt, die auch bei anderen Vertebraten zu finden ist. Eine alternative Klasse, die sich im Stamm RIIIS/J und weiteren Mäusen findet, resultiert in der Expression in den Blutgefäßen, die einen Phänotyp zur Folge hat, der dem Typ 1 von-Willebrand-Syndrom ähnelt, die eine häufige Blutgerinnungsstörung bei Menschen darstellt. Vorherige Studien zeigten, dass die unterschiedlichen *B4galnt2*-Allele in der Maus langfristig einer *Balancing Selection* unterliegen und die Variation der *B4galnt2*-Expression die Interaktion zwischen Wirt und Mikroorganismen im Darm beeinflusst. Das impliziert, dass die Nachteile durch eine höhere Blutungsdauer in Mäusen, die das RIIIS/J-Allel tragen, durch Vorteile in der Resistenz gegenüber gastrointestinalen Pathogenen aufgewogen werden. Die Bedingungen unter denen ein solcher Ausgleich zur langfristigen Erhaltung krankheits-assoziierter Variation führt, sind jedoch bislang ungeklärt.

Um die potentiell pathogenbedingten Selektionsmechanismen zu verstehen und charakterisieren, die in der Natur auf *B4glant2* wirken, habe ich zunächst ein mathematisches Modell entwickelt, das auf einem evolutionären Spiel mit einem modifizierten Wright-Fisher-Prozess basiert und für diploide Individuen angepasst wurde. Dabei habe ich mich insbesondere auf heterozygote Mäuse konzentriert, die *B4galnt2* in Blutgefäßen und Gastrointestinaltrakt exprimieren. Im Vergleich von simulierten und freilebenden Populationen zeigt sich, dass die Frequenzen der Genotypen, die in der Natur beobachtet werden, durch pathogenbedingte Selektion hervorgerufen werden können, falls (i) die Nachteile der Gerinnungsstörung etwa halb so groß sind wie die durch Infektion sowie (ii) sowohl heterozygote als auch das RIIIS/J-Allel tragende Individuen gegen Infektionen geschützt sind. Die Resistenz der heterozygoten Individuen zeigt, dass eine dominante protektive Eigenschaft des RIIIS/J-Allels wahrscheinlicher ist als eine protektive Wirkung des Verlustes der intestinalen Expression. Der Hintergrund der dominanten protektiven Funktion des RIIIS/J-Allels bleibt jedoch unklar, da das Modell impliziert, dass die damit assoziierte vaskuläre Expression nicht unbedingt mit einer Resistenz gegen Pathogene verbunden ist.

Weiterhin habe ich mittels „*Deep Phenotyping*“ in mehr als 200 neugefangenen Mäusen aus Südfrankreich, wo mittlere Häufigkeiten des RIIIS/J-Allels auftreten, auf die Identifikation potentieller Pathogene abgezielt, die die Selektion von *B4galnt2*-Varianten antreiben. Durch die

Analyse von genetischen Mustern, Entzündungszeichen und der intestinalen mikrobiellen Gemeinschaften, konnte ich mehrere Bakteriengattungen mit Mustern in Verbindung bringen, die mit genotypabhängigen Wirt-Pathogen-Interaktionen konsistent sind. Besonders Arten aus der Gattung *Morganella* sind wahrscheinliche Kandidaten, da diese bekannte opportunistische Pathogene einschließt und zudem Häufigkeiten, Prävalenzen und Aktivitätsmuster mit einer erhöhten Entzündung in Mäusen einhergehen, die eine intestinale Expression von *B4galnt2* aufweisen. Zudem konnte ich die entsprechende *Morganella*-Art identifizieren, die eine neue Art der *Morganella morganii*-Gruppe darstellt und über Virulenz-assoziierte Gene verfügt, die nicht in anderen *Morganella*-Arten vorkommen und für ihr Potential über genotypabhängige Wirt-Pathogen-Interaktionen die Selektion von *B4galnt2* voranzutreiben, verantwortlich sein könnten.

Insgesamt hat meine Arbeit dazu beigetragen, neue Einsichten in mögliche evolutionäre Dynamiken von *B4galnt2* in wildlebenden Hausmaus-Populationen zu entwickeln, die zeigen, dass die Pathogen-abhängige Selektion ein wahrscheinlicher Grund für die Erhaltung beider *B4galnt2*-Allele in der Natur ist. Weiterhin bietet meine Arbeit auch über Glykosyltransferasen der Maus hinaus Perspektiven für den Einsatz der hier entwickelten Methoden, die leicht für andere biologische Modelle generalisierbar sind.

Abstract

B4galnt2 is a blood group-related glycosyltransferase that displays cis-regulatory variation for its tissue-specific expression patterns in house mice. The wild type allele, found e.g. in the C57BL/6J laboratory mouse strain, directs intestinal expression of *B4galnt2*, which is the pattern observed among vertebrates, including humans. An alternative allele class found in the RIIIS/J strain and other mice instead drives expression in blood vessels, which leads to a phenotype similar to type 1 von Willebrand disease (VWD), a common human bleeding disorder. Previous studies showed that alternative *B4galnt2* alleles are subject to long-term balancing selection in mice and that variation in *B4galnt2* expression influences host-microbe interactions in the intestine. This suggests that the cost of prolonged bleeding in RIIIS/J allele-bearing mice might be outweighed by benefits associated with resistance against gastrointestinal pathogens. However, the conditions under which such trade-offs could lead to the long-term maintenance of disease-associated variation at *B4galnt2* are unclear.

To understand and characterize the potential pathogen-driven selection acting on *B4galnt2* in the wild, I first developed a mathematical model based on an evolutionary game framework with a modified Wright-Fisher process, adjusted to implement diploid individuals. In particular, I focused on heterozygous mice, which express *B4galnt2* in both blood vessels and the gastrointestinal tract. By comparing simulated to natural populations, I found that the genotype frequencies observed in nature can be produced by pathogen-driven selection when (i) the fitness cost of bleeding is roughly half that of infection and (ii) both heterozygotes and RIIIS/J homozygotes are protected against infection. The resistance of the heterozygote individuals indicates that a dominant protective function of the RIIIS/J allele is more likely than a protective loss of intestinal expression. However, the nature of the dominant protective function of the RIIIS/J allele remains unknown, as the model suggests that the associated vascular expression is not necessarily linked to the pathogen resistance.

Furthermore, I aimed to identify potential pathogens driving the selection at *B4galnt2* by sampling and phenotyping over 200 newly collected mice from Southern France, where an intermediate frequency of the RIIIS/J allele is present. Through the multilayer analysis of genetic patterns, signs of inflammation, and intestinal microbial communities, I could associate several bacterial genera to patterns consistent with genotype-dependent host-pathogen interaction. One genus in particular, *Morganella*, is a likely candidate as it is a well-known opportunistic pathogen and its abundance, prevalence and activity patterns are associated with increased inflammation

in mice with intestinal expression of *B4galnt2*. Finally, I could identify the relevant species of *Morganella*, which represents a new subspecies of the *Morganella morganii* group, and possesses virulence-related genes absent from the other *Morganella* species, which may account for its potential to drive selection at *B4galnt2* via genotype-dependent host-pathogen interactions.

In conclusion, my work provides new insights into the potential evolutionary dynamics taking place at *B4galnt2* in wild populations of house mice, showing that pathogen-driven selection is a likely cause for the maintenance of both *B4galnt2* alleles in the wild. Moreover, my work could be applied beyond the scope of murine glycosyltransferases, as the methods that I developed can easily be generalized to other biological models.

Introduction

Von Willebrand disease (vWd) is a common bleeding disorder characterized by a defect of coagulation due to low plasma levels of the von Willebrand factor (vWf), which can be caused by regulatory mutations outside of the vWf gene, or a defective vWf, which can result from mutations inside the vWf gene itself. In a mouse model of the vWd - the laboratory strain RIIIS/J – the disease is caused by a mutation called Modifier of von Willebrand factor 1 (MvWf1) (*Mohlke, Purkayastha et al. 1999*). This cis-regulatory mutation is located upstream of *B4galnt2*, a blood-group related glycosyltransferase on chromosome 11, and consists of a large haplotype block of approximately 30 kb, highly divergent (>2%) between the wild type C57BL/6J haplotype and the RIIIS/J haplotype, with a peak at ~12% of divergence about 10 kb upstream of *B4galnt2* (*Johnsen, Levy et al.*). These two haplotypes differ by numerous SNPs, insertions and deletions, resulting in an RIIIS/J haplotype 10 kb smaller than the C57BL/6J allele (*Johnsen, Levy et al.*). In C57BL/6J mice, as in humans, *B4galnt2* is expressed in the gastrointestinal epithelium, where it takes part in the glycosylation of the mucosa. The MvWf1 mutation in RIIIS/J mice changes the tissue specificity of *B4galnt2* to the vascular endothelium, where vWf becomes one of its substrate. The addition of a *B4galnt2*-specific GalNac residue on vWf accelerate its clearance leading to low plasma levels of vWf, producing a bleeding disorder similar to vWd in humans (*Mohlke, Purkayastha et al.*). In RIIIS/J mice, the plasma level of vWf is twenty times lower than in the wild type strain C57BL/6J (*Mohlke, Purkayastha et al.*).

The RIIIS/J haplotype was identified in 13 laboratory mouse strains out of 59 tested (*Johnsen, Levy et al.*), suggesting it is a common polymorphism. But wild mice are constantly exposed to various sources of injury, making it reasonable to consider a mutation provoking a bleeding disorder as a fitness cost for the concerned individuals, leading to low RIIIS/J allele frequencies in the wild. As such, the apparent frequency of the RIIIS/J allele in laboratory strains could be an artifact of their creation as all laboratory strains, although originally derived from wild caught individuals, were inbred for decades, potentially fixing mutations otherwise rare in nature, or creating new mutations that do not exist in the wild (*Silver 1995*). However, 5 of the 13 strains bearing the RIIIS/J haplotype are "wild-derived" strains (*Johnsen, Levy et al.*), suggesting that this polymorphism might be relevant for wild populations of house mice and indeed, three studies (see below) identified high frequencies of RIIIS/J allele in wild populations of mice from the genus *Mus*, despite the potential fitness cost of bleeding disorder inherent to the RIIIS/J allele.

First, the RIIIS/J haplotype was observed in frequencies varying from 22 to 39% in wild population of *M. m. domesticus* from France, Cameroon, and the United-States, while it couldn't be found in the German population studied (*Johnsen, Teschke et al.*). Focusing on the German and French populations at the DNA level, a high nucleotide diversity in the French population of *M. m. domesticus* ~10 kb upstream of *B4galnt2* can be observed, while the nucleotide diversity at this locus in the German population is low. This corresponds to the MvWf1 mutation, already found to be highly divergent between RIIIS/J and C57BL/6J haplotypes. Interestingly, in the French population, microsatellite loci linked to the RIIIS/J allele show highly reduced heterozygosity in the 20 kb region directly upstream of *B4galnt2*, compared to the same region for the microsatellite alleles linked to the C57BL/6J haplotype. This suggests a recent increase (low microsatellite diversity in RIIIS/J alleles) of an old allele (high nucleotide divergence between haplotypes). Moreover, based on this genetic information, the estimated age of the RIIIS/J allele is 97 years based on microsatellite variation *vs.* 70 000 years based on allele frequency assuming neutral drift. This implies that some kind of selective pressure is acting on *B4galnt2* to keep the RIIIS/J allele at high frequency in the French population.

Moreover, the RIIIS/J haplotype class was also identified in various species belonging to the genus *Mus (M. m. musculus, M. m. domesticus, Mus spretus and Mus castaneus)*, with allele frequencies reaching over 80% in one population (*Linnenbrink, Johnsen et al.*). All studied populations show high nucleotide diversity 10 kb upstream of *B4galnt2*, where the MvWf1 mutation lies, and in most cases a significantly positive Tajima's D. This statistical result, taken with the observed trans-species polymorphism, is consistent with long-term balancing selection, which appears to have maintained both *B4galnt2* haplotype classes in wild populations for at least 2.8 My, which is the time to the last common ancestor of the investigated populations.

Finally, a striking pattern of allele frequency distribution was identified in different populations of *M. m. domesticus* from France and Germany (*Linnenbrink 2012*). In the south and west of France, the RIIIS/J allele frequencies are high, varying from 34 to 45%, whereas in the northeast of France and in Germany, the RIIIS/J allele was rare, varying from 0 to 4%. This pattern could not be explained by population structure as determined by mitochondrial D-loop haplotypes or 18 neutral microsatellite loci with STRUCTURE (*Pritchard, Stephens et al.*), nor environmental cues (*Linnenbrink 2012*, *Linnenbrink, Wang et al. 2013*). Therefore, it appears that the selective pressure(s) that might be responsible for the maintenance of the RIIIS/J allele in wild populations of *M. m. domesticus* are geographically limited.

These three studies show that despite the expected fitness cost that a bleeding disorder represents for wild mice, the RIIIS/J allele is not only present at high frequencies in wild populations of *Mus* species, but also show signs of long-term balancing selection, and recent positive selection. This implies that the fitness cost due to the bleeding disorder must be balanced by an unknown benefit, which could reside in the loss of gastrointestinal expression of *B4galnt2* or in the gain of vascular expression, or both.

Several studies investigating signs of balancing selection in the human genome have identified genes that are mainly involved in immunity in its broadest sense (*Andres, Hubisz et al. 2009*, *Andrés 2011*, *Leffler, Gao et al. 2013*). Moreover, several other blood-group related genes showing signs of balancing selection seem to be involved in host pathogen interactions (*Fumagalli, Cagliani et al. 2009*, *Segurel, Gao et al. 2013*), suggesting that *B4galnt2*, being under balancing selection itself, could be involved in host-pathogen interactions as well. Given the role that *B4galnt2* plays in the glycosylation of the gastrointestinal mucosa, where a tremendous amount of bacteria and other microorganism reside, continuously using the various glycans available in the mucus as an anchor or source of carbon, it is sensible that its absence should have a non-negligible effect on the resident microbiota, and potential microbial pathogens. In fact, two studies showed the importance of *B4galnt2's* expression pattern on the composition of the gastrointestinal microbial communities.

First, the study of C57BL/6J wild type mice and their *B4galnt2*-knockout counterparts (*Staubach, Kunzel et al. 2012*) showed that the composition of the gut microbiota differs according to the expression level of *B4galnt2* in the gut. Moreover, indicator species -- bacterial species that are specific for a given "habitat" -- were found for both genotypes, and the indicators of *B4galnt2*-expressing mice contain known pathogenic genera such as *Helicobacter, Campylobacter, Shigella* and *Citrobacter*, suggesting that the absence of *B4galnt2* expression in the gut could be protective against bacterial pathogens.

A second study repeated the experiment (*Rausch, Steck et al. 2015*), again showing that whether *B4galnt2* is expressed in the gastrointestinal tract or not changes the composition and diversity of the resident microbial communities. Moreover, model infections with *Salmonella typhimurium* showed that mice with no *B4galnt2* gastrointestinal expression are to some extent protected against this pathogen, as they show lower signs of inflammation and lower colonization of *Salmonella*.

Furthermore, some bacterial pathogens such as *Staphylococcus aureus* (*McAdow, Missiakas et al. 2012*) and *Helicobacter pylori* (*Byrne, Kerrigan et al. 2003*) are known to bind vWf as a mechanism of invasion and/or evasion from the host immune system, potentially leaving RIIIS/J mice protected, as their level of plasma vWf is twenty times lower than in C56BL6/J mice (*Mohlke, Purkayastha et al.*), weakening the pathogenic potential of these bacteria.

In conclusion, both circumstantial and experimental evidence point towards pathogen-driven selection as responsible selective force maintaining the polymorphism of the cis-regulatory region of *B4galnt2* in mice. This type of selection could explain the distribution of *B4galnt2* alleles in *M. m. domesticus* populations from Western Europe, e.g. if the implicated pathogen is absent from the German and North-eastern French populations while abundant in the south and west.

In this context, my PhD thesis aims to understand and characterize the pathogen-driven selection potentially responsible for the long-term maintenance of both murine *B4galnt2* alleles in wild populations of *Mus* species. To this end, I first confirmed the signatures of selection acting on the *B4galnt2* cis-regulatory region in Western European *M. m. domesticus*. Next, I used mathematical modeling to understand under which conditions pathogen-driven selection might maintain both *B4galnt2* murine alleles in mouse populations. And finally, I performed a broad study of *M. m. domesticus* from Southwest France and identified candidate pathogens that could potentially contribute to the signatures of selection at *B4galnt2* in natural populations.

Chapter I:

Evaluating the maintenance of disease-associated variation at the blood group-related gene *B4galnt2* in house mice. *

** Vallier M, Abou Chakra M, Hindersin L, Linnenbrink M, Traulsen A, Baines JF.*

Submitted to BMC Evolutionary Biology

Introduction

Von Willebrand disease (VWD) is a common human bleeding disorder characterized by a defect of coagulation caused either by low plasma levels of von Willebrand factor (VWF) or a dysfunctional VWF. In a mouse model of VWD – the laboratory strain RIIIS/J – the disease is caused by a cis-regulatory mutation at the *B4galnt2* gene, a blood group related glycosyltransferase (*Mohlke, Nichols et al. 1999*, *Mohlke, Purkayastha et al. 1999*). This mutation switches the usual expression pattern of *B4galnt2* in the gastrointestinal (GI) epithelium, as observed in the wild type strain C57BL/6J and other vertebrates (*Stuckenholz, Lu et al. 2009*), to the vascular endothelium in the RIIIS/J strain. Vascular expression of *B4galnt2* leads to aberrant glycosylation of VWF, resulting in its accelerated clearance from circulation. Accordingly, RIIIS/J mice have up to twenty times lower plasma levels of VWF than C57BL/6J mice (*Mohlke, Purkayastha et al. 1999*).

Despite the expected fitness cost of prolonged bleeding times for wild animals, the RIIIS/J allele is found in high frequencies in various wild populations of house mice and their relatives (*Johnsen, Teschke et al. 2009*, *Linnenbrink, Johnsen et al. 2011*). Furthermore these populations show signs of long term balancing selection maintaining both C57BL/6J and RIIIS/J allele classes for at least 2.8 Million years. Further, in a previous survey of *Mus musculus domesticus* populations, a partial selective sweep revealed a recent increase in RIIIS/J allele frequency in a population from Southern France, while the allele was absent from a German population (*Johnsen, Teschke et al.*). This suggests that selective force(s) operating on *B4galnt2* alleles in Western Europe may differ according to space and/or time.

Genome-wide scans for balancing selection in the human genome (*Andres, Hubisz et al. 2009*, *Andrés 2011*, *Leffler, Gao et al. 2013*) identified a moderate number of genomic regions, but nearly all of them are involved in immunity *lato sensu*, supporting the hypothesis that *B4galnt2* could be involved in host-pathogen interactions, as shown for other blood-group related genes (*Fumagalli, Cagliani et al. 2009*, *Segurel, Gao et al. 2013*). Laboratory experiments show that the absence of *B4galnt2*-associated GalNac residues on the GI mucosa results in an altered resident microbiota (*Staubach, Kunzel et al.*), and that this modified GI microbiota confers lower susceptibility to a model of *Salmonella typhimurium* infection (*Rausch, Steck et al. 2015*). On the other hand, bacteria such as *Staphylococcus aureus* are known to use VWF to invade the host and escape the immune system (*McAdow, Missiakas et al. 2012*, *Thomer, Schneewind et al. 2013*). Although experimental evidence is lacking, *S. aureus*'s ability to utilize VWF could be

compromised in RIIIS/J allele-bearing mice due to the low plasma levels of VWF, and hence lead to protection against this pathogen. Thus, potential benefits of the RIIIS/J allele could reside either in the *gain* of vascular expression and/or in the *loss* of GI expression in mice homozygous for the RIIIS/J allele, which would be associated with resistance against systemic and/or intestinal pathogens, respectively.

Under the above hypothesis, heterozygous mice are of particular interest, as they express *B4galnt2* in both blood vessels and the GI tract (i.e. the two allele classes influence on tissue-specific expression patterns is co-dominant) (*Johnsen, Teschke et al. 2009*), potentially incurring both the cost of bleeding and pathogen susceptibility. It is known that heterozygous mice have the same bleeding phenotype as RIIIS/J homozygotes (i.e. the RIIIS/J allele's effect on VWF is dominant) (*Mohlke, Purkayastha et al. 1999*, *Johnsen, Teschke et al. 2009*), although we have little information concerning the susceptibility of heterozygous mice to pathogens in the wild. Indeed, heterozygous mice could have the same level of protection as the RIIIS/J homozygotes (e.g. in the case of *S. aureus* using VWF directly to infect), whereas on the other hand they could display similar susceptibility to gut pathogens as the C57BL/6J homozygotes (e.g. when a gut pathogen utilizes *B4galnt2* specific GalNac residues in the mucosa). Finally, heterozygous mice could have an intermediate phenotype compared to both homozygotes in terms of resistance or susceptibility to pathogens.

In this study, we set out to determine the conditions under which the trade-off between prolonged bleeding times and pathogen susceptibility leads to the maintenance of the RIIIS/J allele, and in addition extended a previous geographic survey of *B4galnt2* allele frequencies (*Johnsen, Teschke et al.*) to characterize spatial selection across Western Europe in more detail. Accordingly, we modeled the interaction between host and pathogen using an evolutionary game with a Wright-Fisher process (*Imhof and Nowak 2006*). Since mice are diploid sexual organisms, we modified the Wright-Fisher "asexual" random process to include diploid reproduction. Pathogens were modeled as an environmental variable, being either present or absent, with the possibility for the environment to change regularly from one state to another. Alternatively, we also relax this assumption and model the pathogens such that their population depends on frequencies of susceptible hosts. Although simplified, the model provides a method to disentangle the effects of genotypic costs and environmental variability on the host population. Moreover, the environmental model resembles a "trench warfare" dynamic (i.e. advances and retreats of resistance allele frequency due to costs in the absence of a pathogen (*Stahl, Dwyer et al. 1999*, *Woolhouse, Webster et al. 2002*), which we might expect in the context of balancing selection acting on resistance/susceptibility alleles as may be the case at *B4galnt2* (*Johnsen, Teschke et al.*).

To identify the model parameters that best explain the natural population dynamics, we compared the simulated populations to the observed wild populations. We found that the genotype frequencies observed in nature were best explained by a model where heterozygotes are protected against infection with a pathogen in a frequency-dependent manner, and the cost of bleeding being half that of infection.

Results

I. Wild mice

First, in order to further characterize the intriguing geographic pattern of RIIIS/J allele frequency observed by *Johnsen et al.* (*Johnsen, Teschke et al. 2009*), we typed *B4galnt2* allele classes using the same diagnostic PCR fragment in a set of eight wild population collections spread across France and Germany (*Linnenbrink, Wang et al.*). These populations represent six new locations, in addition to a resampling of the two locations previously analyzed by *Johnsen et al.* (*Johnsen, Teschke et al.*). This reveals an intriguing pattern of distribution of the RIIIS/J allele: it is nearly absent in the north and east of France and in Germany, but it is consistently >30% in three local populations in the south and west of France (figure 1A).

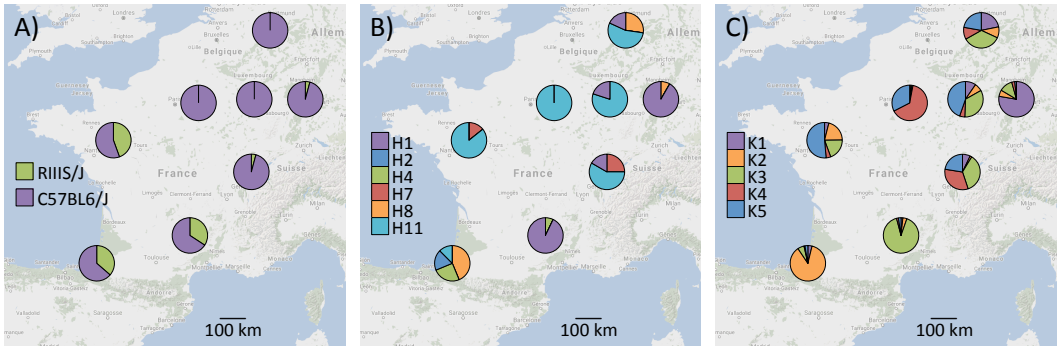

Figure 1: *B4galnt2* allele frequencies and population structure. **A)** *B4galnt2* allele frequency, **B)** Mitochondrial D-loop haplogroup, based on the haplogroup nomenclature from *Bonhomme et al.* (*Bonhomme, Orth et al. 2011*), **C)** Distribution of genetic clusters based on STRUCTURE analysis applied to 18 neutral microsatellite markers, colors indicate genetic clusters.

As we previously attributed differences in RIIIS/J allele frequency between two of these locations to a recent, partial selective sweep (*Johnsen, Teschke et al.*), we evaluated whether this pattern holds in this broader dataset. We thus typed 12 microsatellite loci linked to *B4galnt2* and resolved their haplotypic phase with respect to the RIIIS/J and C57BL/6J alleles as previously described (*Johnsen, Teschke et al.*). This reveals a near identical pattern (figure 2), whereby the expected heterozygosity of the two loci located closest to the cis-regulatory mutation of *B4galnt2* (-30 kb and 0 kb) is very low on the RIIIS/J background, while high and close to the genome average (as determined by 18 unlinked microsatellites) (*Linnenbrink, Wang et al.*) on the C57BL/6J background.

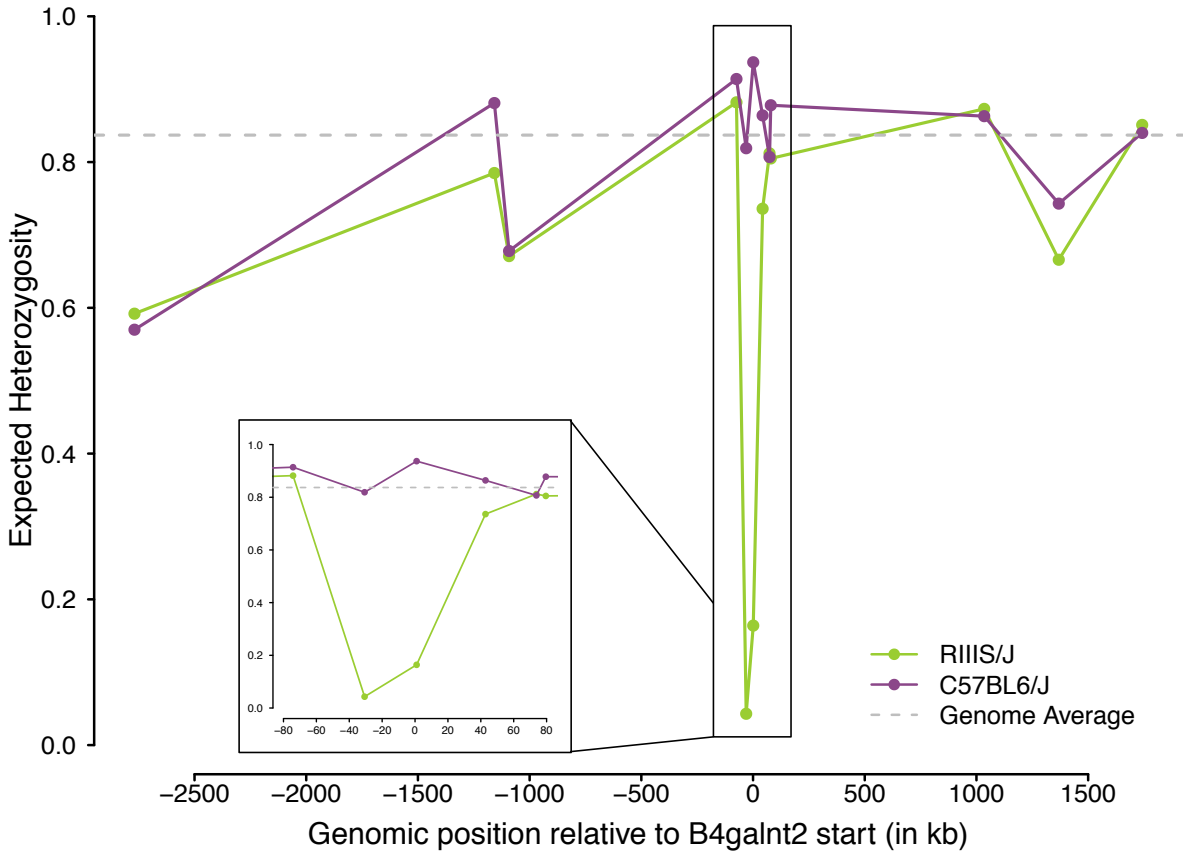

Figure 2: Expected heterozygosity in eight wild mice populations. Data are presented for 12 phased microsatellite loci spanning 160 kb surrounding the *B4galnt2* start position. The data are stratified for RIIIS/J (green) and C57BL/6J (purple) allele classes. The genome average (dotted gray lines) is determined by 18 unlinked microsatellite loci.

Due to the extremely high level of nucleotide divergence observed between the RIIIS/J and C57BL/6J alleles, it is possible, however, that the two microsatellites displaying very low heterozygosity on the RIIIS/J background could have experienced e.g. one or more mutation interrupting their repeats and thus changing their mutation rate. Thus, we also performed direct Sanger sequencing of these individuals, which reveals no evidence of interruption. Rather, these two dinucleotide loci display a repeat number within the range of the alleles found on the C57BL/6J background, but with very few alleles (two and four alleles at each locus, respectively). Thus, the pattern of a drastic, local reduction of microsatellite variability near the cis-regulatory mutation on the background of the RIIIS/J allele is most consistent with a partial selective sweep.

A second alternative is that the above-mentioned geographic pattern of allele frequency distribution could also be related to underlying population structure. Indeed, different waves of migration led to the colonization of Western Europe by house mice (*Bonhomme, Rivals et al. 2007*,

Gabriel, Johannesdottir et al. 2010, *Harr, Karakoc et al. 2016*): one coming from the east through modern day Turkey and Greece, and another from the south through North Africa and Spain. These migration routes led to the distinct maternal lineages present in Northern Europe and the Mediterranean basin (*Bonhomme, Orth et al. 2011*, *Jones, Johannesdottir et al. 2011*). To test whether the distribution of *B4galnt2* alleles might be explained by population structure, we compared the observed allele frequencies to the previously established distribution of the mitochondrial D-loop haplogroups (figure 1B) and the genetic clusters identified by 18 nuclear microsatellite markers (figure 1C) (*Linnenbrink, Wang et al. 2013*). This reveals little to no correspondence, e.g. some local populations dominated by the same mitochondrial haplogroup and/or genetic cluster display contrasting RIIIS/J frequencies, and on the other hand, local populations with similar RIIIS/J frequencies display contrasting haplogroups and/or genetic clusters. Thus, the observed pattern of RIIIS/J allele frequency appears to have little to do with underlying population structure.

Taken together, these results confirm and extend those of *Johnsen et al.* (*Johnsen, Teschke et al. 2009*): a partial selective sweep visible through the *B4galnt2*-linked microsatellites indicates that the RIIIS/J allele recently rose in frequency in Southwestern France, most likely due to the action of strong natural selection.

II. Model

II.1 Constant environment

Although a constant environment is very unlikely in nature, the study of this limiting case allows us to test the behavior of our model. We chose to vary the costs of bleeding and infection from 0 to 1, incremented by steps of 0.2.

In a constant environment with no pathogen (figure 3), the cost of infection c_i is irrelevant, and only the cost of bleeding c_b influences the outcome of the simulation. Thus, the three investigated values of the infection cost for the heterozygotes c_h lead to the same results, as $c_h=c_i=0$. When $c_b=0$, we have a neutral state (i.e. all individuals have the same fitness) leading to ~50% heterozygotes and ~25% of each homozygote. When $c_b>0$, as expected the CC individuals dominate the population, representing over 80% of the individuals, whereas the RR individuals are very close to 0 and the heterozygotes remain in low frequency (<20%). These proportions depend on the value of c_b. Indeed, when c_b increases, the selection strength increases,

particularly on the heterozygotes, leading to a deviation from Hardy-Weinberg equilibrium (HWE) and an excess of the favored homozygotes -- the CC individuals.

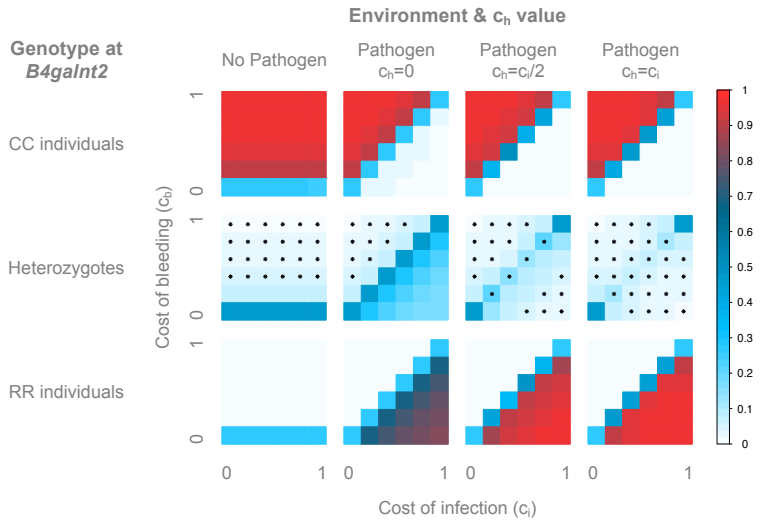

Figure 3: Average genotype frequencies in the model with a constant environment. The frequencies are displayed according to the cost of bleeding (y axis) and of infection (x axis). For the non-pathogenic environment (left panel), all c_h values are equivalent since c_i=c_h=0 in all cases. For the pathogenic environment models with varying c_h values are shown. The average genotype frequencies across 50 simulations, each with 10000 generations, are displayed. The frequencies are color-coded according to the legend on the right. Stars indicate an excess of homozygotes.

In a constant environment with a pathogen, the value of c_h has a great influence on the population frequencies, as it is dependent on the cost of infection. With c_h=0 (figure 3), heterozygotes have the same fitness as the RR individuals, and the neutral state is reached whenever c_b=c_i. Consistently, when c_b>c_i, the CC individuals dominate the population and when c_i>c_b, the RR individuals and heterozygotes represent the majority. Interestingly, when the difference in costs becomes too high, the selection becomes so strong that the population deviates from HWE with an excess of homozygotes. However, this is only true when c_b>c_i but not when c_b<c_i, indicating the asymmetry of the system that translates to a stronger effect of c_b compared to that of c_i. When c_h>0 (figure 3) the heterozygotes have a lower fitness than both homozygotes, resulting in the population being mostly composed of the favored homozygotes (RR individuals when c_b<c_i and CC individuals when c_i<c_b). Notably, the neutral state is reached only for the two extreme cases were all individuals have the same fitness (c_b=c_i=0 and c_b=c_i=1) and not for every c_b=c_i as in the previous model. This can be explained by the strong selection acting on the heterozygotes when c_h>0, as they bear the dual cost of bleeding and infection. This leads to a deviation from HWE and an excess of homozygotes. As previously observed, this deviation is also present when the difference in costs becomes too strong, but this time for both c_i>c_b and c_i<c_b. However, due to the asymmetry of the system when c_h=c_i/2, the difference in costs must be

stronger for $c_i > c_b$ to lead to a deviation from HWE than for $c_b > c_i$. For $c_h = c_i$, the system becomes symmetrical: c_i and c_b have the same effect on the selection strength, leading to a deviation from HWE for the same difference in costs when $c_i > c_b$ as when $c_b > c_i$.

II.2 Changing environment

To approach the trench warfare dynamics that may be relevant for the putative host-pathogen interactions involving *B4galnt2*, we modeled the pathogen as an exogenous variable, being either present or absent from the environment. This property of the environment was regularly alternated according to host generations: the environment switches from pathogenic to non-pathogenic and back every S host generations. We investigated a broad range of switching frequencies: every 1, 10, 50, 100, 500, 1000 and 5000 host generations. For all values of c_h, the two rapid switching frequencies (1 & 10) show similar results, as do the intermediate (50 & 100) and slow ones (500 onwards), thus, we display the results for 1, 50 and 500.

First, we observe for $c_h = 0$ (figure 4A) that the parameter space is divided in two distinct regions where a given genotype is favored, as in the case for the constant pathogenic environment described previously. The boundary line is however different, as it does not represent a neutral state, but is still characterized by the coexistence of the three genotypes. Interestingly, the boundary lies around $c_b = c_i/2$ for rapid switching, but approaches the $c_b = c_i$ line for slower frequencies. Of note, this boundary region is important as it regularly occurs in the subsequent analyses and is in most cases characterized by the coexistence of all three genotypes, with only one- or a combination of two genotypes being favored above- or below the boundary region, respectively. In the case of rapid switching (figure 4A), the results are qualitatively very similar to the constant pathogenic environment: the boundary line approach a neutral state, since we have ~ 50% heterozygotes and ~25% of each homozygote; above this line, the population is dominated by CC individuals, and below this line, the RR individuals and heterozygotes represent the majority. The position of the line is however not the same. Indeed, the average payoff of the CC individuals in this model is $c_i/2$, as half of the time these individuals bear no cost, and the other half they bear the cost of infection c_i. This pushes the boundary line to $c_b = c_i/2$ rather than $c_b = c_i$ as in the constant pathogenic environment. We observe, like in the constant pathogenic environment, a deviation from HWE for $c_b >> c_i/2$. In this rapid model, the heterozygotes can be seen as an allelic pool that helps the system maintain both alleles in the population, and ensure the transition between the two homozygous states. For slower frequencies of environmental (pathogenic) change (figure 4A), the delay between switches is long enough for the alleles to fix,

and for each period the system reaches the characteristics of the corresponding constant environment, therefore bringing the boundary line back towards $c_b=c_i$, similar to the constant pathogenic environment. Moreover, under these conditions heterozygotes are no longer needed to maintain both alleles in the population, as the selection is strong enough to recover the alleles from very low frequencies. It appears that the heterozygotes even suffer from stronger selective pressure than the homozygotes, as we observe a deviation from HWE with an excess of homozygotes already with low fitness costs. Moreover, the asymmetry of the system is different compared to the constant environment. Indeed, in a non-changing environment, the influence of c_b compared to c_i on the selection strength is stronger, leading to deviations from HWE for smaller differences in costs when $c_i<c_b$ than when $c_b<c_i$. However, in this fluctuating environment a certain difference in costs is needed above the boundary line to lead to a deviation from HWE, as for the constant environment, but below the boundary it seems that only the value of c_b is important.

For $c_h=c_i/2$ (figure 4B), the selection on the heterozygotes is stronger than on the homozygotes, as already observed under the constant environment. This leads to the "disappearance" of heterozygotes on the boundary line when c_b and c_i increase and an excess of homozygotes. The position of the line is however similar to that of the $c_h=0$ model: it lies around $c_b=c_i/2$ for the rapid environmental changes and approaches $c_b=c_i$ for slowly fluctuating environments. For the rapidly switching environment (figure 4B), when $c_b>c_i/2$ the CC individuals dominate the population, as observed for $c_h=0$. When $c_b<c_i/2$ however, the RR individuals dominate alone and not in conjunction with the heterozygotes as for $c_h=0$. This is due to the lower fitness of the heterozygotes. A similar pattern is observed for the intermediate environment (figure 4B), but the selection appears to be stronger for below the boundary, as we observe an excess of homozygotes. For the slowly switching environment (figure 4B), we still observe the dominance of the CC individuals above the boundary, but below the line the RR individuals do not take over and rather coexist with the CC individuals. The heterozygotes, however, are still in very low frequency due to their low fitness, leading to an excess of homozygotes. This is again due to the asymmetry of the system: the cost of bleeding is always present but the cost of infection is present only half of the time, leading to a stronger selective pressure from the bleeding phenotype than from infection.

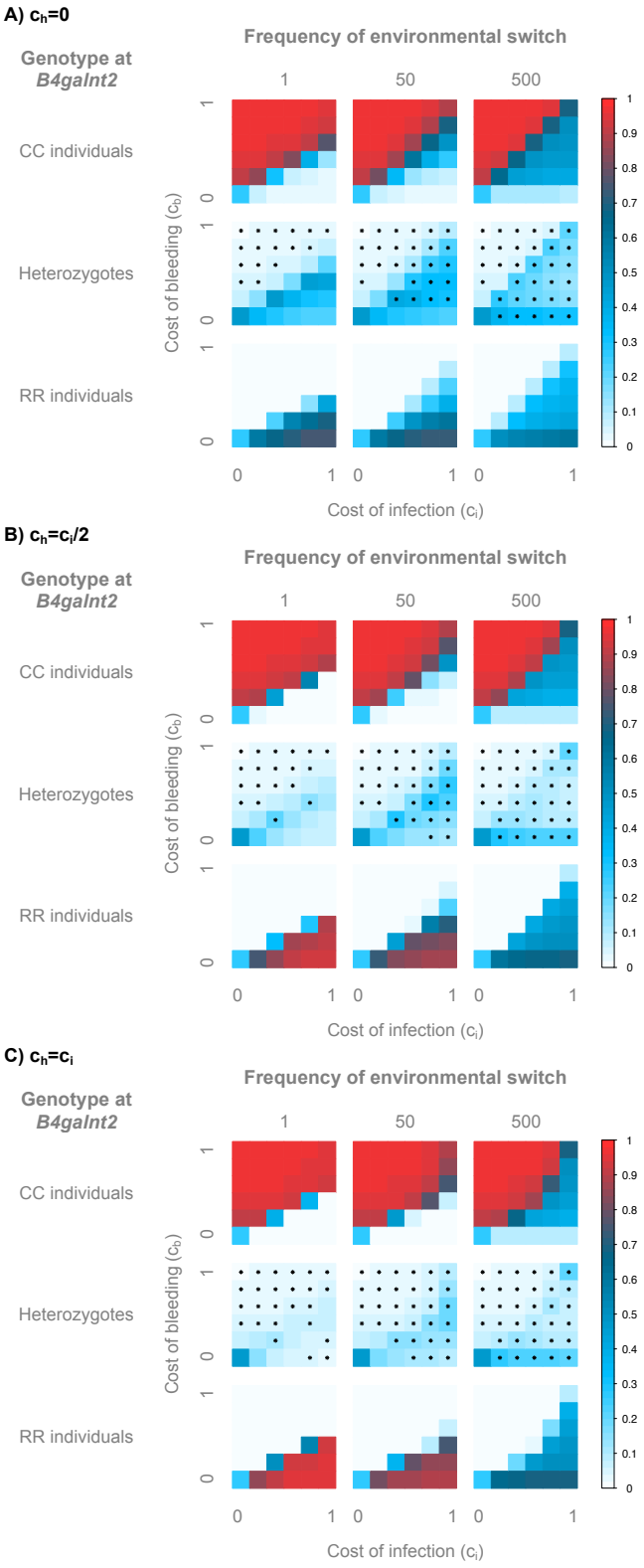

Figure 4: Average genotype frequencies in the model with a switching environment. The frequencies are displayed according to the frequency of environmental change expressed in host generations, the cost of bleeding (y axis) and of infection (x axis). The average genotype frequencies across 100 simulations, each with 10000 generations, are displayed for c_h=0 (A), c_h=c_i/2 (B) and c_h=c_i (C). The frequencies are color-coded according to legend on the right. Stars indicate an excess of homozygotes.

For $c_h=c_i$ (figure 4C), the trend is similar to that of $c_h=c_i/2$, but the selective pressure is stronger on the heterozygotes than with $c_h=c_i/2$, leading to deviations from HWE with lower values of c_b and c_i. Interestingly, we observe that the boundary line is no longer characterized by a linear $c_b=c_i/2$ relationship, but rather takes an exponential distribution. This might be due to the non-additive dual cost of the heterozygous mice.

II.3 Similarity to natural populations

One important goal of constructing our model is to compare its results to the pattern of *B4galnt2* allele frequencies observed among wild populations of mouse species belonging to the genus *Mus*, in order to understand the selective forces maintaining disease-associated variation at this locus. Accordingly, we evaluated a broad collection of populations (summarized in table 1) from the current- and two previous studies (*Johnsen, Teschke et al.* , *Linnenbrink, Johnsen et al.*): *Johnsen et al.* (*Johnsen, Teschke et al. 2009*) and the current study provide data from a total 10 *M. m. domesticus* populations from Europe, Africa and North America, whereas *Linnenbrink et al.* (*Linnenbrink, Johnsen et al. 2011*) added an ancestral *M. m. domesticus* population (Iran) and data from other house mouse subspecies and their relatives, including *M. m. musculus* (Kazakhstan), *M. m. castaneus* (India) and *M. spretus* (Spain).

The study by *Linnenbrink et al.* (*Linnenbrink, Johnsen et al. 2011*) revealed that greater allelic and functional diversity is present at *B4galnt2* than that previously observed in derived *M. m. domesticus* populations. Indeed, the *M. m. domesticus* population from Iran and *M. spretus* population from Spain both display a modified RIIIS/J allele, which appears to turn off gastrointestinal expression of *B4galnt2* without turning on vascular expression. Interestingly, the frequency of this modified RIIIS/J allele class is higher than in any of the derived *M. m. domesticus* populations, which is consistent with it having the potential to be beneficial against infections without incurring the cost of prolonged bleeding times. Further, the *M. m. musculus* population form Kazakhstan contains yet another allele class at low frequency, termed "CRK", which appears to be a recombinant allele driving expression in both the GI tract and blood vessels. For simplicity, however, we did not consider the Kazakh population containing this low frequency CRK class in the analysis. The *M. m. castaneus* population from India was also excluded, as no functional data on *B4galnt2* expression patterns is available for this population/subspecies.

Table 1: **Description of the wild house mice populations used in this study.**

Population		Species	Location	Study	Sample Size				RIIIS/J Allele Frequency	Hardy-Weinberg
ID	Group				RR	RC	CC	Total		
DE	A	Mmd	Cologne-Bohn (DE)	i	0	0	36	36	0.00	Equilibrium
CB	A	Mmd	Cologne-Bohn (DE)	iii	0	0	15	15	0.00	Equilibrium
DB	A	Mmd	Divonne-lès-Bains (FR)	iii	0	1	11	12	0.04	Equilibrium
LO	A	Mmd	Louan-Villegruis-Fontaine (FR)	iii	0	0	12	12	0.00	Equilibrium
NA	A	Mmd	Nancy (FR)	iii	0	0	12	12	0.00	Equilibrium
SL	A	Mmd	Schömberg (DE)	iii	0	1	11	12	0.04	Equilibrium
Overall	**A**	--	--	--	**0**	**2**	**97**	**99**	**0.01**	**Equilibrium**
MC	B	Mmd	Massif Central (FR)	iii	1	11	7	19	0.34	Equilibrium
ES	B	Mmd	Espelette (FR)	iii	3	10	9	22	0.36	Equilibrium
AN	B	Mmd	Anger (FR)	iii	4	8	6	18	0.44	Equilibrium
CH	B	Mmd	Chicago (USA)	i	1	3	6	10	0.25	Equilibrium
CA	B	Mmd	Cameroon	i	2	18	9	29	0.38	Equilibrium
FR	B	Mmd	Massif Central (FR)	i	12	16	22	50	0.40	Homozygotes Excess
Overall	**B**	--	--	--	**23**	**66**	**59**	**148**	**0.38**	**Equilibrium**
IR	C	Mmd	Iran	ii	10	5	2	17	0.74	Equilibrium
SP	C	Ms	Spain	ii	19	7	1	27	0.83	Equilibrium
Overall	**C**	--	--	--	**29**	**12**	**3**	**44**	**0.80**	**Equilibrium**

Data from populations belonging to *Mus musculus domesticus* (Mmd) or *Mus spretus* (Ms) were included from (i) Johnsen et al. 2009 (*Johnsen, Teschke et al. 2009*), (ii) Linnenbrink et al. 2011 (*Linnenbrink, Johnsen et al. 2011*) and (iii) Linnenbrink 2013 et al. (*Linnenbrink, Wang et al. 2013*). The populations are categorized into three groups: A) populations with low RIIIS/J allele frequency, suspected to be in a non-pathogenic environment, B) populations with intermediate RIIIS/J allele frequency, suspected to be in a pathogenic environment and C) populations with high frequency of a modified RIIIS/J allele assumed to carry no bleeding cost, and suspected to be in a pathogenic environment. The sample size is given with the number of individuals of each genotype (RR for RIIIS/J homozygotes, CC for C57BL/6J homozygotes and RC for heterozygotes), and the total number of mice. The corresponding RIIIS/J allele frequency and whether the population significantly deviates from HWE are also indicated.

Thus, we ultimately grouped the included populations into three categories, summarized in table 1:

A. Populations where the RIIIS/J allele is either absent or its frequency is very low, which are assumed to be in a non-pathogenic environment. These include five M. *m. domesticus* populations across Germany and Northeastern France, one of which (Cologne-Bonn) was sampled twice;

B. Populations displaying intermediate RIIIS/J allele frequencies, which are assumed to be in a pathogenic environment. These include one M. *m. domesticus* population from North America, one M. *m. domesticus* population from Africa, and three M. *m. domesticus* populations from Southwestern France, one of which (Massif Central) was sampled twice;

C. Populations with a modified RIIIS/J allele that are assumed to carry no bleeding cost and likewise assumed to be in a pathogenic environment. These include one M. *m. domesticus* population from Iran and one M. *spretus* population from Spain.

To evaluate which model and parameters best explain the observations in natural populations, we estimated the similarity between the simulated and observed genotype frequencies (see Methods).

First, we observe for the populations assumed to be in a non-pathogenic environment (population group "A") (figure 5A) that the simulations from the constant, non-pathogenic environment match well with the observed populations whenever $c_b>0$ and for every value of c_h. Notably, when c_b is very strong (>0.2) the population deviates from HWE. Further, the simulations from the constant pathogenic environment and the changing environments all match well with the observed data above the boundary line, for every value of c_h. This might be explained by the asymmetry of the model, which generally favors CC individuals, yielding all simulations above the boundary to closely match the observed group A populations. Notably, the majority of the simulated population deviate from HWE, since only a small cost window above the boundary line is in equilibrium for the constant pathogenic environment and the rapidly switching one, when $c_h<c_i$. For the slowly switching environments, only small values of c_b leave the population in equilibrium.

For the populations assumed to be in a pathogenic environment (population group "B") (figure 5B), we observe that the constant non-pathogenic environment does not explain the observed data very well, only in the case where c_b=0, which is likely unrealistic. For the constant pathogenic- and the switching environment, we observe that in general the best match to the real populations is at the boundary line, with the populations simulated in a rapidly changing environment for c_h=0, reaching the highest similarity to the observed populations. Notably, only the constant pathogenic- and rapidly switching environments with c_h=0 maintains HWE while providing relatively high similarity to the observed populations.

Finally, for the populations without bleeding phenotype (population group "C") (figure 5C), we observe that the constant non-pathogenic environment is unlikely to explain the observed data: for c_b>0 the similarity is between 40 and 50%, although it reaches 70% for c_b=0. The constant pathogenic- and switching environment best explain the data below the boundary line, and it seems, as for the pathogenic populations, that these environments, are more likely to fit the real populations with c_h=0 than with c_h>0, as their genotype frequencies are very close to the observed ones (>90% similarity). The intermediate environment (S=50) with c_h=0 and the slowly switching environment (S=500) with all c_h values both fit the populations well for null or very low values of c_b, which is also consistent with our hypothesis that these mice carry no cost of bleeding. Notably, only the slowly switching environment produces high similarity with an excess of homozygotes, whereas the other environments produce high similarity while maintaining HWE.

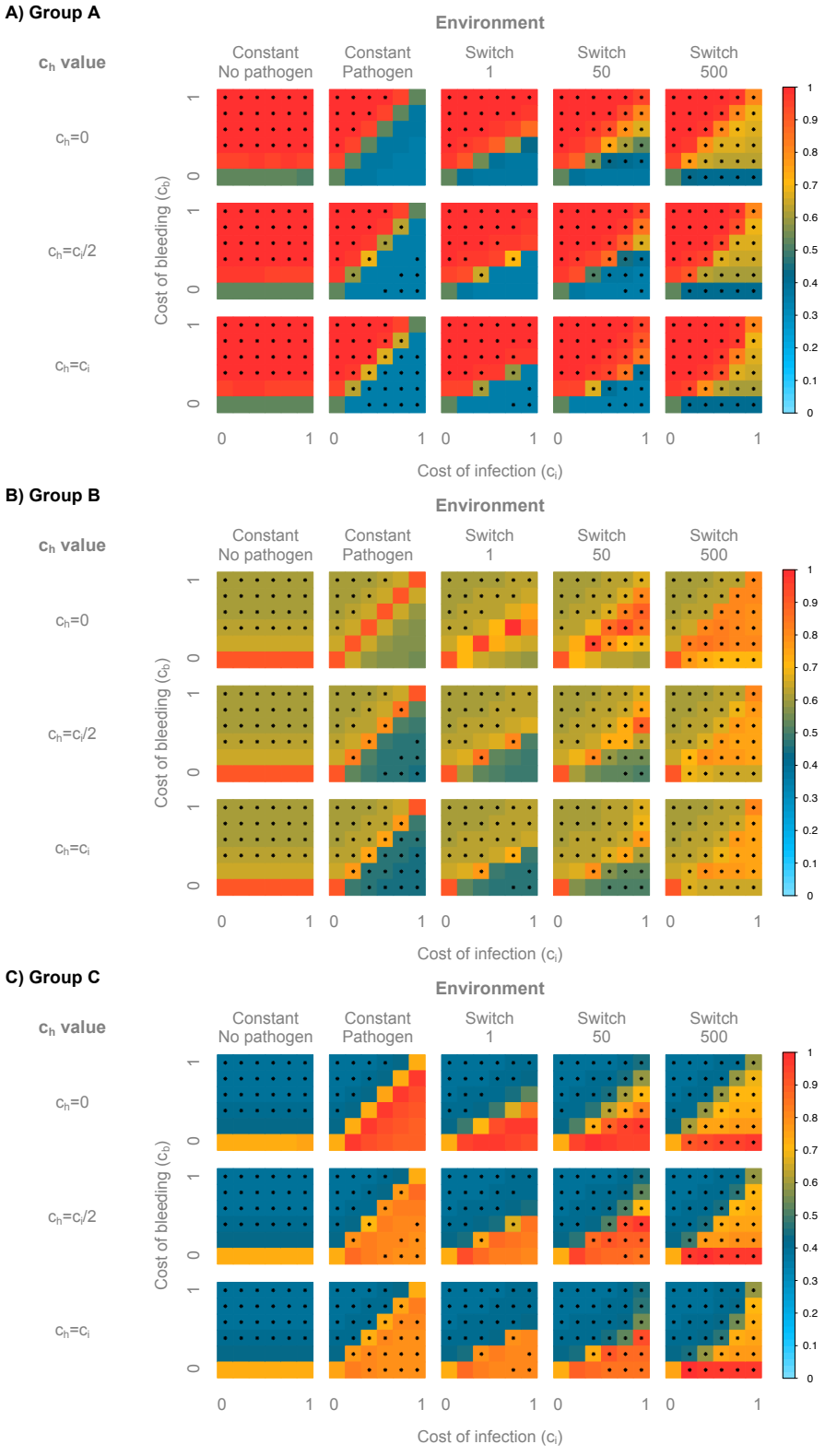

Figure 5: Similarity of the simulated populations to the natural populations. A) Similarity to populations from Group A, B) Similarity to populations from Group B, C) Similarity to populations from Group C. The similarity is displayed according to the value of c_h, the cost of bleeding (y axis) and of infection (x axis), and the modeled environment (constant with- or without pathogen, and switching between a pathogenic- and non pathogenic environment every 1, 50 or 500 host generations). The similarity is color-coded according to the legend on the right. Stars denote an excess of homozygotes. Full similarity is achieved when all genotype frequencies coincide.

Since all studied populations bear similar *B4galnt2* alleles, with the exception of the modified RIIIS/J allele found in group C, it is reasonable to assume that the populations will bear largely the same fitness costs. This applies to the C57BL/6J allele class in all three groups regarding the cost of infection, c_i. Similarly, we can consider population groups A and B to carry the same cost of bleeding, $c_b>0$, whereas this is expected to be zero for group C. These assumptions allow us to further identify combinations of costs of bleeding and infection that might take place in nature, by comparing the similarity values across the three groups. First, group B can be seen as the "limiting factor" since they are approached by the simulated populations only at the boundary line in a rapidly changing environment and with $c_h=0$. This reduces the space of possible cost parameters to $c_b=c_i/2$, excluding the special case of $c_b=c_i=0$. Population groups A and C are however never approached by the simulated populations at the boundary, but always above or below this line, respectively. This suggests that group A is in a constant environment without a relevant pathogen. There are however multiple possibilities for group C: assuming that c_b is indeed zero in these populations, a constant environment, a rapidly changing environment, and an intermediate frequency environment are all capable of producing the observed genotype frequencies. This is limited to small non-zero values of c_i in the constant environment, but is true for all $c_i>0$ for the switching environments.

II.4 Pathogen as frequency

The models we investigated so far are important to understand the behavior of the system, approach trench warfare dynamics and model seasonal changes, but another key biological aspect is the reaction of a pathogen to the changes in host genotype frequencies. Thus, to address this aspect we modified our model to let the pathogen population vary according to the host population. For this, we express the pathogen as the proportion of susceptible individuals in the host population. Interestingly, this model does not lead to a trench-warfare dynamic, but quickly reaches an equilibrium that is maintained over 10000 host generations.

First, we observe that the average population (figure 6) is relatively similar to the fast and intermediate environment (S=1; S=50). The selection strength appears however weaker, as more populations remain in HWE compared to the switching environments. For $c_h=0$, the boundary line lies around $c_b=c_i/2$, as for the rapidly switching environment. Above this line CC individuals dominate the population, whereas below the line RR individuals and heterozygotes share the majority. As for the rapidly switching environment, when c_b becomes too high compared to c_i, the population deviates from HWE with an excess of homozygotes. For $c_h=c_i/2$ however, the

boundary appears to move towards $c_b=c_i$. CC individuals are dominant above this line, and below the line RR individuals dominate, as the selective pressures on the heterozygotes are stronger when $c_h>0$. In contrary to the switching environment, where the boundary shows a deviation from HWE, we observe deviations from HWE in this model only above this line, suggesting that the selective pressures might be weaker in this model. For $c_h=c_i$, the results are very similar to $c_h=c_i/2$. However, with the selection strength being stronger on the heterozygotes, we see more deviations from HWE and a higher proportion of homozygotes above and below the boundary, as we previously observed for switching environments.

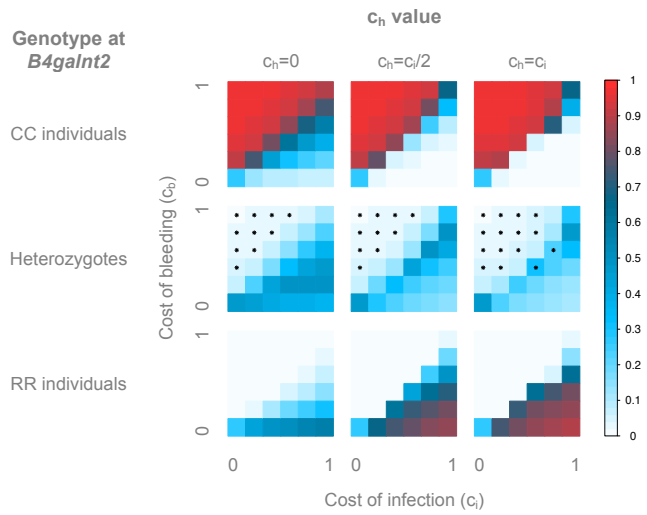

Figure 6: Average genotype frequencies in the model with a frequency-dependent environment. The frequencies are displayed according to the value of c_h, the cost of bleeding (y axis) and of infection (x axis). The average frequencies across 100 simulations, each with 10000 generations, are displayed. The frequencies are color-coded according to the legend on the right. Stars indicate an excess of homozygotes.

Given the resemblance of the frequency-dependent model to the switching environments in terms of genotype frequencies, we might expect similar results concerning the fit to the real populations. Indeed, we observe (figure 7) that the populations assumed to be in a non-pathogenic environment, group A, are best approximated when the costs are above the boundary line, regardless of the value of c_h. For the populations assumed to be in a pathogenic environment, group B, the model fits best around the boundary in general, and in particular for $c_h=0$. For the populations with no bleeding phenotype, group C, the models fit best below the boundary: for the special case of $c_b=0$ with $c_h=0$, and in a narrow space just below it with $c_h>0$.

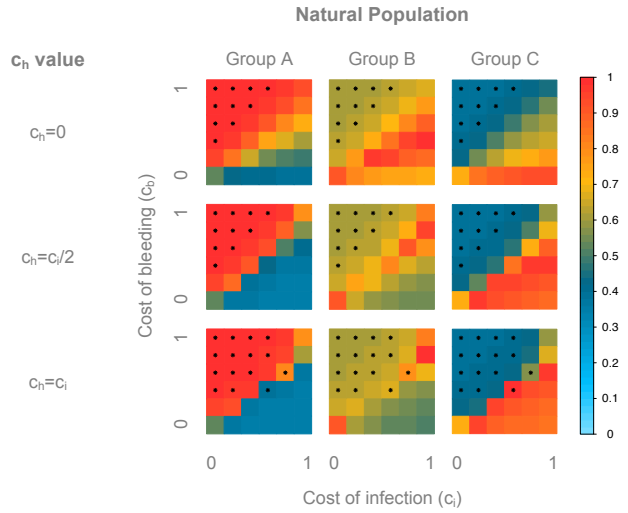

Figure 7: Similarity of the natural populations to the populations simulated in the model with frequency-dependent environment. The similarity is displayed according to the value of c_h, the cost of bleeding (y axis) and of infection (x axis). The similarity is color-coded according to the legend on the right. Stars indicate an excess of homozygotes. Full similarity is achieved when all genotype frequencies coincide.

As for the switching environment, comparing the results across populations enables us to infer the best overall model. Again, groups B and C are only compatible with c_h=0. Group C is best approximated by the model for any c_i>0.2, whereas group B is best approximated for c_b=c_i/2 with c_i>0.2. Group A is however not compatible with the two other population categories, as they are approximated by any c_b>c_i/2 for c_h=0, suggesting once again that they are associated with a non-pathogenic environment.

In conclusion, although conceptually very different from the exogenously changing environment, the frequency-dependent model leads to similar results as the rapidly changing environment. In both models it appears that the model with c_h=0 is more plausible, with costs within the non-zero range of c_b=c_i/2.

II.5 Hardy-Weinberg based process

Our modified Wright-Fisher based process represents a very good approximation of the real populations when considering genotype frequencies. In addition, we developed an alternative process based on HWE that calculates the expected number of individuals from each genotype based on the weighted fitness, which also represents a computationally faster model, as there is only one calculation and few random steps per generation.

Interestingly, this model leads to remarkably similar results to those obtained from the Wright-Fisher based process. The genotype frequencies are indeed very similar for both population dynamics model (constant environment: figure S1; switching environments: figure S2, frequency-dependent environment: figure S4) and consequently, the comparison to the natural populations is also very similar to the Wright-Fisher based process (constant & switching environment: figure S3, frequency-dependent environment: figure S5).

This model however represents a stronger selection regime, as the number of random steps is highly reduced compared to the Wright-Fisher based process. This translates into heterozygote frequencies being lower in the HWE-based process than in the random process for otherwise equal model parameters (c_b, c_i, c_h, S), which consequently leads to an excess of homozygotes.

II.6 Effect of mutations/migration

The results presented thus far were produced using a mutation rate $\mu=0.005$, which in our system can also be viewed as a proxy for migration, as the two *B4galnt2* alleles considered are highly divergent including numerous SNPs and indels, which renders direct mutation from one functional allele class to the other unlikely. Since this mutation/migration rate could significantly impact the results, we investigated its effects using the HWE-based process.

For the constant environments (figure 8), we observe a relatively similar pattern as with mutation. In a constant non-pathogenic environment, we observe a neutral state for $c_b=0$, and CC individuals dominate the population when $c_b>0$. For constant pathogenic environments, CC individuals dominate when $c_b>c_i$ for every value of c_h, as already observed in the model with mutations. RR individuals dominate when $c_b<c_i$, also for every value of c_h, which differs from the models with mutation since this was not the case for $c_h=0$, where both RR individuals and heterozygotes were abundant. In general, the model without mutation represents a very strong selection regime, as heterozygotes are nearly absent from every model, except for the neutral states, where all genotypes have equal fitness ($c_b=0$ for constant non-pathogenic environment, $c_b=c_i$ for constant pathogenic environment when $c_h=0$, $c_b=c_i=0$ and $c_b=c_i=1$ for constant pathogenic environment when $c_h>0$).

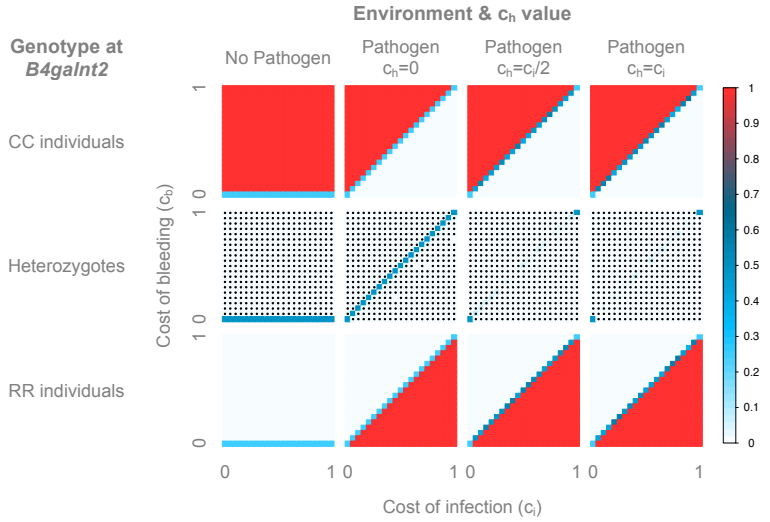

Figure 8: Average genotype frequencies in the model with a constant environment, without mutation. The frequencies are displayed according to the cost of bleeding (y axis) and of infection (x axis). For the non-pathogenic environment (left panel) all c_h values are equivalent since c_i=c_h=0 in all cases. For the pathogenic environment, models with varying c_h values are shown. The average genotype frequencies across 100 simulations using the HWE-process without mutation, each with 10000 generations, are displayed. The frequencies are color-coded according to the legend on the right. Stars indicate an excess of homozygotes.

Although the starting environment has little influence on the long-term dynamics in the models with mutation, it has potentially strong consequences in the models without mutation. Therefore, it is necessary to distinguish the models that started in a pathogenic environment from those that started in a non-pathogenic environment for the exogenously changing environments.

For the rapidly switching environment (figure 9A), the starting environment has little influence on the results. These are however different from the models with mutation. Although the boundary line is at the same position, the heterozygotes are mostly in much lower frequency: they are limited to a narrow region around c_b=c_i/2 for c_h=0, and nearly absent for c_h>0 (except for the neutral state c_b=c_i=0), leaving the population to be dominated by the homozygotes.

For the intermediate environment (figure 9B), the starting environment has limited influence on the model with c_h=0, but a strong influence on the models with c_h>0. For the c_h=0, the boundary line is at c_b=c_i/2, as for the model with mutation, but like for the constant and rapid environment, the heterozygotes are very low in frequency except on the boundary. For c_h=c_i/2, the simulations beginning in a non-pathogenic environment are relatively similar to those with mutation, but again the heterozygotes are nearly absent. For the simulations beginning with a pathogenic environment, the results are very similar to those from the constant pathogenic environment. Finally, for c_h=c_i, both starting environments look very similar to the corresponding constant environment.

For the slowly switching environment (figure 9C), the starting environment strongly influences all models. For all values of c_h, the models beginning in a non-pathogenic environment resemble the constant non-pathogenic environment, with the CC individuals dominating the population for nearly all $c_b>0$. For the simulations beginning in the pathogenic environment however, only models with $c_h>0$ resemble the constant pathogenic environment. The model with $c_h=0$ on the other hand differs, as the RR individuals dominate the population only when $c_b<c_i/2$, while both homozygotes are present when $c_i/2<c_b<c_i$. This can be explained by the presence of heterozygote individuals. When the population begins in a pathogenic environment for $c_h=0$, both RR individuals and heterozygotes are favored over CC individuals, leading to an initial phase where both RR individuals and heterozygotes are in high frequency in the population. The presence of heterozygotes allows the reappearance of CC individuals when the environment switches, and they subsequently become favored over both other genotypes. When $c_h>0$ heterozygotes disappear quickly from the population as they have a lower fitness than both homozygotes. Thus, when the environment changes, the population is composed of only one homozygote genotype, and due to the absence of mutation the other genotypes are unable to return, hence the resemblance to the constant environment.

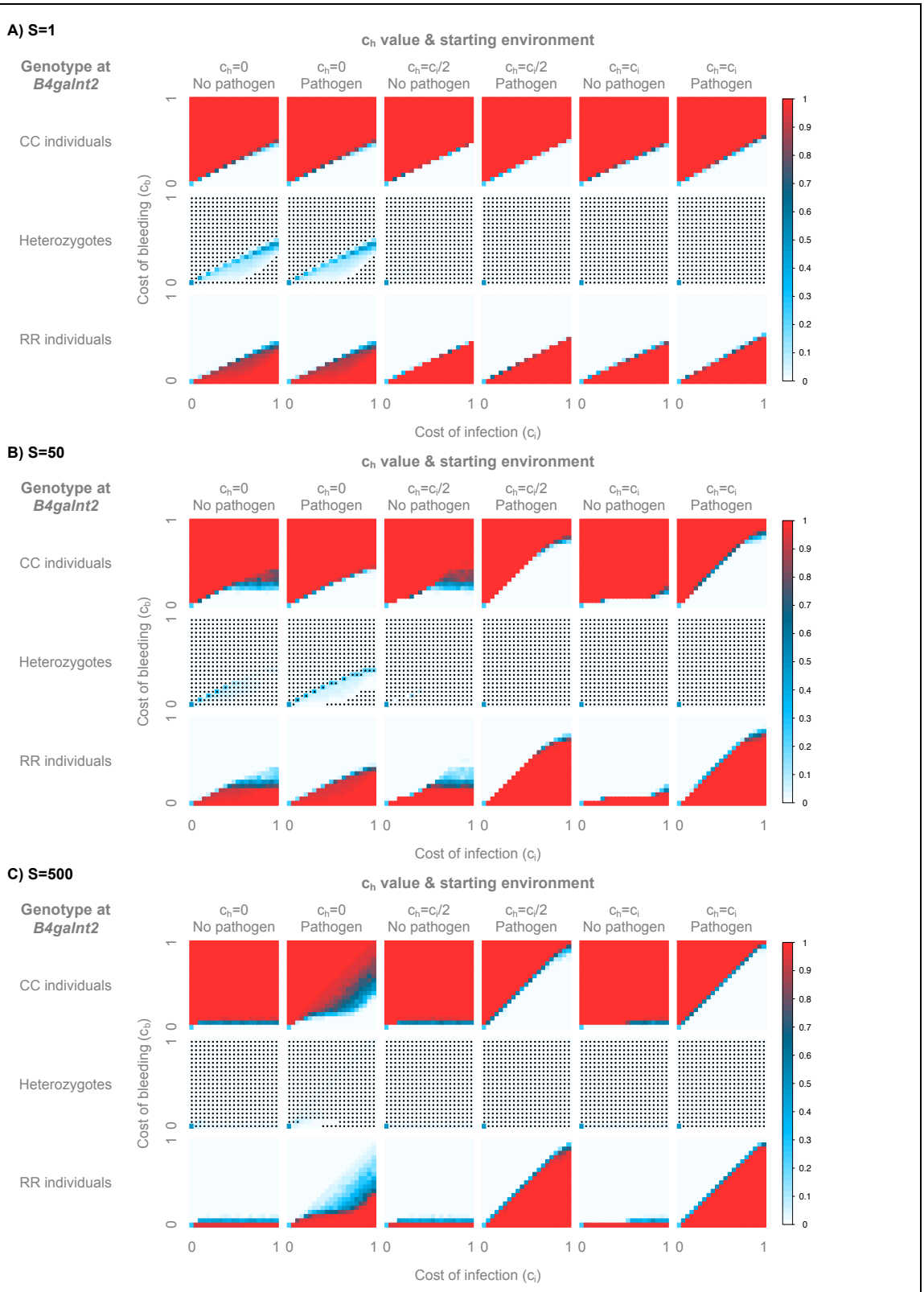

Figure 9: Average genotype frequencies in the model with a switching environment, without mutation. The frequencies are displayed according to the value of c_h, the starting environment, the cost of bleeding (y axis) and the cost of infection (x axis). The environment switches every A) 1 host-generation B) 50 host-generations or C) 500 host-generations. The average genotype frequencies across 100 simulations using the HWE-process without mutation, each with 10000 generations, are displayed. The frequencies are color-coded according to the legend on the right. Stars indicate an excess of homozygotes.

In contrast to the switching environments, the frequency-dependent environment without mutation (figure 10) is quite similar to the one with mutation. However, more populations deviate from HWE, and heterozygotes are in lower frequencies compared to the model with mutation.

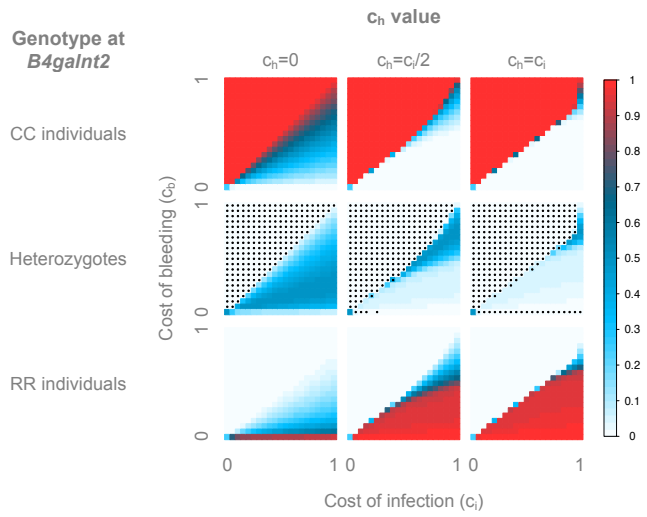

Figure 10: Average genotype frequencies in the model with a frequency-dependent environment, without mutation. The frequencies are displayed according to the value of c_h, the cost of bleeding (y axis) and of infection (x axis). The average genotype frequencies across 100 simulations using the HWE-process without mutation, each with 10000 generations, are displayed. The frequencies are color-coded according to the legend on the right. Stars indicate an excess of homozygotes.

In conclusion, mutation/migration appear to be dispensable to the maintenance of all three genotypes in the population only in the rapid and intermediate environments, both with c_h=0 or in a frequency-dependent environment. Moreover, the starting environment, which has a negligible effect on the population frequencies when mutation is allowed, seems to dictate the genotype frequencies for intermediate and slowly switching environments in the absence of mutation.

Discussion

The study of polymorphism at *B4galnt2* in house mice provides an interesting opportunity to elucidate the selective forces leading to the maintenance of disease-associated variation in nature. Previous studies revealed the action of long-term balancing selection on the one hand (*Linnenbrink, Johnsen et al. 2011*), in addition to dynamics on more recent timescales in the present study. Systematic review of signatures of balancing selection in the human genome identified genes involved in immunity *lato sensu* (*Andres, Hubisz et al. 2009*, *Andrés 2011*, *Leffler, Gao et al. 2013*), indicating that host-pathogen interactions could be among the forces maintaining allelic diversity at *B4galnt2*. In this study, we set out to better understand the nature of potential trade-offs between resistance against pathogens and susceptibility to prolonged bleeding times surrounding variation at *B4galnt2*.

First, by extending our previous geographic survey of *B4galnt2* alleles across Western Europe from two- (*Johnsen, Teschke et al. 2009*) to eight locations (*Linnenbrink, Wang et al. 2013*), we discovered an intriguing pattern of *B4galnt2* allele distribution, with northern populations being nearly devoid of the RIIIS/J allele in contrast to southwestern populations, which display intermediate RIIIS/J allele frequencies. Comparing this distribution of *B4galnt2* allele frequencies to population structure based on either mtDNA haplotypes or unlinked nuclear microsatellite markers (*Linnenbrink, Wang et al. 2013*) reveals little- to no correspondence. In contrast, we confirmed the pattern of a partial selective sweep of the RIIIS/J allele (*Johnsen, Teschke et al. 2009*) in two additional populations, indicating that the high frequency of the RIIIS/J allele in the Southwestern populations is most consistent with the action of recent natural selection.

To further investigate host-pathogen interactions as a possible driver of selection at *B4galnt2*, we developed a mathematical model derived from the classical Wright-Fisher process, to which we added random mating. This was necessary as existing mathematical models, including the Wright-Fisher process are applicable to haploid hosts. Moreover, heterozygotes individuals are of particular interest in this study system, as they express *B4galnt2* both in the blood vessels and the gastrointestinal tract, leaving them with a potential dual cost of prolonged bleeding and infection susceptibility. As our model focuses on the mechanisms maintaining host genotypes in a population, we treated the pathogen as an environmental probability p that can be constant (p=1 - all susceptible hosts are infected; p=0 - no host is infected), fluctuating (switching between 0 and 1 with frequency S expressed in host generations) or dependent on the proportion of susceptible individuals in the host population. We ran a very broad set of simulations, aiming to

maximize the possible parameter combinations to fully understand the model and its behavior, and ultimately determine which parameters can lead to population frequencies similar to those observed in the wild.

Importantly, for nearly all combinations of parameters studied, we observe a similar dichotomy between regions of parameter space where a given genotype(s) is favored. The boundary between these regions can be observed at $c_b=c_i/2$ or $c_b=c_i$ (depending on the values of c_h and S), and is characterized by the coexistence of all three genotypes. The model is however asymmetric, as above the boundary, CC individuals always dominate the population, while below the line, RR individuals dominate the population either alone or in conjunction with heterozygotes (depending on the values of c_h and S).

Interestingly, the two different population dynamic models used, a Wright-Fisher process with random mating or a HWE-based process, lead to remarkably similar results. The average genotype frequencies are similar in both models, with the differences being characterized by the variation around the mean being much larger in the random process, which contains 3N random steps per generation, while the HWE-based process contains only a few. This result is quite valuable as it allows the use of the HWE-based model to quickly screen the parameter space to identify interesting parameter combinations, which can subsequently be run under the random process, which is much slower to compute due to its speed being proportional to population size.

Not surprisingly, the mutation rate can have a strong influence on the model, although the extent of this influence greatly depends on the environmental model considered. First, for constant-, frequency-dependent- or rapidly switching environments, the effect of mutation is rather weak. Indeed, for constant environments the direction of selection does not change over time, leading to the near fixation of a favored genotype, as the small number of unfavoured individuals produced through mutation is quickly removed by selection. In contrast, for rapidly switching environments the favored individuals do not have the time to reach dominance in the population before the environment changes, leading to the maintenance of all three genotypes. In this case, the switch from one homozygote to the other when the environment changes is ensured partly through the mating of heterozygotes and partly through mutation, which explains the limited influence of the latter on the population frequencies. For frequency-dependent environments, the system quickly reaches an equilibrium, which stems from the fitness of the different genotypes with little input from the mutation rate. Second, for slowly fluctuating environments the absence of mutation drastically influences the population. Indeed, when the environment changes infrequently, the favored genotype has sufficient time to reach dominance

such that when the environment changes, only mutation can rescue the other genotypes in the population, leading to completely different population dynamics with- or without mutation. With mutation the system is reversible, and the genotypes can alternate according to the environment. Without mutation the population becomes rapidly fixed for one homozygous genotype and can no longer change, even when the environment switches between pathogenic and non-pathogenic states.

In the context of long-term balancing selection at *B4galnt2*, we use the mutation rate as a proxy for migration, since the high divergence between *B4galnt2* haplotypes and likely complex nature of the regulatory sequences separating them (*Johnsen, Levy et al.*) make it unlikely that one allele would directly mutate to another. Moreover, *M. m. domesticus* populations in Western Europe display very little population structure on a continental scale (*Salcedo, Geraldes et al. 2007*), thus, models with migration are clearly more realistic.

Lastly, we compared the results of our models to the current and previous surveys (*Johnsen, Teschke et al. 2009*, *Linnenbrink, Johnsen et al. 2011*, *Linnenbrink, Wang et al. 2013*) of DNA sequence polymorphism at *B4galnt2,* which we grouped in three categories: populations with very low RIIIS/J allele frequency, which are suspected to be in a non-pathogenic environment (group A); populations with intermediate RIIIS/J allele frequencies, which are suspected to be in a pathogenic environment (group B); and finally populations with high frequencies of a modified RIIIS/J allele that does not carry a cost of bleeding, which are suspected to be in a pathogenic environment (group C). For group B, the best fitting models are those with a rapidly switching environment and the frequency-dependent environment, each with costs of bleeding and infection on the boundary line ($c_b=c_i/2$), excluding the special case of $c_b=c_i=0$. For group C, the best fitting models are those with rapid- and intermediate switching environments and the frequency-dependent environment, for costs below the boundary ($c_b<c_i/2$) and particularly for $c_b=0$, which corresponds to prior knowledge of the expression phenotype for the modified RIIIS/J allele. For group A, any environment is suitable to explain the observed data, for any costs above the boundary.

With the exception of group C harboring a modified RIIIS/J allele class, the remaining populations studied share related alleles and their corresponding expression patterns, suggesting that they may have similar costs of bleeding and infection. For the three population groups to be compatible, the costs must follow a $c_b=c_i/2$ relationship, excluding the $c_b=c_i=0$ special case; group A must be in a constant non-pathogenic environment; groups B & C must be in a rapidly changing environment or frequency-dependent environment; and finally group C must have $c_b=0$.

It is notable that the model predictions perfectly match our hypotheses for groups A and C. Moreover, the predictions for group B (a frequency-dependent or rapidly switching environment) represent two biologically relevant models, the first one taking the response of the pathogen into account, while the second may reflect seasonal changes.

Interestingly, the better fitting model is that where heterozygotes and RIIIS/J homozygotes are protected against bacterial infections (c_h=0). Although our previous analysis of both commensal gut bacterial communities (*Staubach, Kunzel et al. 2012*) and an experimental model of infectious gastroenteritis (*S. typhimurium*; (*Rausch, Steck et al. 2015*)) suggest a potential benefit of the removal of *B4galnt2* expression in the GI tract, we note that *S. typhimurium* is not a naturally occurring mouse gut pathogen and requires antibiotic pre-treatment in order to cause intestinal pathology. On the other hand, although it played a comparatively smaller role, blood vessel expression driven by the RIIIS/J allele does appear to provide a small degree of protection in the *S. typhimurium* model, which might be associated with increased mucus thickness (*Rausch, Steck et al. 2015*). This indicates that the potential benefit of vascular *B4galnt2* expression does not reside solely in the blood vessels -- as could be the case with e.g. systemic infection with *Staphylococcus* -- but also in the gastrointestinal tract, where it seems to have a protective effect against *S. typhimurium* colonization (*Rausch, Steck et al. 2015*).

Although our results appear to be in agreement with previous research, they raise new questions regarding the potential mechanism(s) of protection against pathogens involving *B4galnt2*. Indeed, if the heterozygotes experience the same degree of protection against infection as RIIIS/J homozygotes (c_h=0), it implies that the benefit lies in the vascular expression of *B4galnt2* and not in the absence of gastrointestinal expression. However, group C carries a modified RIIIS/J allele that turns off GI expression without turning the vascular expression on (*Linnenbrink, Johnsen et al. 2011*). Consequently the RIIIS/J allele would presumably lack the protection provided by vessel expression. This suggests that the RIIIS/J allele might have another, yet unknown function(s) that leads to protection against pathogens.

Conclusion

In conclusion, by comparing the results of our models to the patterns of polymorphism at *B4galnt2* in natural populations and considering the still limited functional information available for this gene, we are able to recognize the long-term maintenance of the RIIIS/J allele through host-pathogens interactions as a viable hypothesis if its fitness costs due to prolonged bleeding time are roughly half those of being susceptible to a given pathogen. Further, our models identify that a putative dominant-, yet unknown protective function of the RIIIS/J allele appears to be more likely than a protective loss of GI expression in RIIIS/J homozygotes, which may help guide future experiments. Lastly, our model developed here may be used for numerous other biological scenarios, as it does not depend on explicit assumptions regarding a given gene or phenotype, but could be applied to any other diploid model where two co-dominant alleles are maintained by fluctuating selection.

Methods

I. Wild mice

We genotyped the *B4galnt2* locus in newly collected mice described in *Linnenbrink et al.* (*Linnenbrink, Wang et al. 2013*) by sequencing a previously developed diagnostic PCR product following the procedure described in *Johnsen et al.* (*Johnsen, Teschke et al. 2009*). Sequences were edited in Seqman (included in DNASTAR, Inc., Madison, Wisc.) and aligned to the homologous sequences from RIIIS/J (GenBank EF372924) and C57BL/6J (NCBI build 36) using the ClustalW algorithm (*Thompson, Higgins et al. 1994*) included in MEGA 4.0.2 (*Tamura, Dudley et al. 2007*).

We further typed 12 microsatellite loci located around *B4galnt2* cis-regulatory mutation as described previously (*Johnsen, Teschke et al. 2009*). The alleles were called with GENEIOUS 7.0 (Biomatters Ltd) and the haplotypic phase was reconstructed with PHASE 2.1 (*Stephens, Smith et al. 2001*). The algorithm was run 5 times with 10.000 iterations, a thinning interval of 100 and a burn-in of 10.000, and the best output was chosen based on the "goodness of fit". Microsatellite gene diversity estimates were calculated using GenoDive 2.0 (*Meirmans and Van Tienderen 2004*). The two microsatellite loci with highly reduced diversity - located at -30 kb and 0 kb from *B4galnt2* start position - were additionally sequenced using the same primer pairs and PCR conditions as for their typing; the sequencing was performed as for the *B4galnt2* Fragment 5, and the sequences were analyzed in GENEIOUS 7.0 (Biomatters Ltd). Finally, the STRUCTURE analysis included was taken from the output of Linnenbrink et al. (*Linnenbrink, Wang et al. 2013*).

II. Model

II.1 Principle

We modeled the interaction between mouse hosts and pathogens as an evolutionary game (*Hofbauer, Schuster et al. 1982*, *Broom and Rychtář 2013*). Evolutionary game theory uses mathematical models assuming that a genotype with a high fitness (given by the payoff from the interaction) has a high probability to spread within a population (*Hofbauer, Schuster et al. 1982*). More precisely, we investigated whether the presence of a pathogen can lead to the

maintenance of the two murine alleles of *B4galnt2*. In short, *B4galnt2* is a glycosyltransferase expressed either in the gastrointestinal epithelium (C57BL/6J allele) or in the vascular endothelium (RIIIS/J allele). Although the second allele causes prolonged bleeding times, most likely at a significant cost to wild mice, both alleles have been maintained by balancing selection for over 2.8 My. Our working hypothesis is that this maintenance may be due to a protective effect of the RIIIS/J allele against pathogen(s), where protection could result from the loss of gastro-intestinal expression and/or from the gain of vascular expression.

II.2 Pathogen

As our goal is to understand whether the presence of a pathogen can lead to the maintenance of the host alleles in wild populations, we modeled the pathogen as being present with probability p. If p=1, the pathogen is overwhelmingly present and every susceptible host is infected; if p=0, no host is infected; if 0<p<1, the burden of infection for the susceptible hosts is proportional to p. This method allows us to focus on the host population and avoid the many assumptions we should make if we were to model a dynamic and co-evolving pathogen population (e.g. generation time relative to host generation time, population size, transmission mode, transmission efficiency…).

Under the hypothesis of balancing selection due to a trade-off between prolonged bleeding time and pathogen resistance/tolerance, we expect a trench warfare dynamic: (i) the frequency of susceptible hosts increases in the absence of the pathogen due to the cost of resistance, (ii) as the number of susceptible hosts increases, the pathogen population grows, favoring the resistant hosts, (iii) as the number of resistant hosts increases, the pathogen population decays, favoring the susceptible hosts, and the cycle continues (*Stahl, Dwyer et al. 1999*, *Woolhouse, Webster et al. 2002*). To approximate this phenomenon, we let the environment vary between a state where no pathogen is present (p=0) and a state where pathogens are overwhelming (p=1). This environmental switch (S) is based on the host generations so that the environment changes every S host generations. Varying S allows us to investigate different rates of evolution.

Alternatively, we approximated p with the proportion of susceptible hosts present in the population. This may represent a more "natural" model, as we do not externally force the switch from a pathogenic to non-pathogenic environment.

II.3 Host

Considering the host population and our focus on the gene *B4galnt2*, we have two allelic states: R represents the vascular endothelium expression allele (RIIIS/J) and C represents the gastrointestinal epithelium expression allele (C57BL/6J). These alleles can be combined into three possible genotypes - RR, RC and CC - and determine the payoff of an individual. RR individuals carry a cost of bleeding c_b. CC individuals carry a cost of infection c_i in a pathogenic environment and no cost in a pathogen-free environment. RC individuals present an interesting case as they express *B4galnt2* in both tissues, potentially carrying both costs. We previously demonstrated that heterozygous mice display the same bleeding phenotype as the homozygous RR individuals (*Mohlke, Nichols et al. 1999*); hence, they carry the same cost of bleeding c_b. However, we have no evidence that the intestinal phenotype of the heterozygotes is equivalent to that of CC individuals, so we defined a separate infection cost for the heterozygotes c_h, which we can vary to explore different phenotypes. We investigated three possibilities. First $c_h=0$, where heterozygotes do not become infected by the pathogen, like the RR individuals. This corresponds to the hypothesis of protection through the gain of vascular expression. Then $c_h=c_i$, where heterozygotes are infected the same as the CC individuals. This corresponds to the hypothesis of protection through the loss of gastrointestinal expression. Finally $c_h=c_i/2$ represents a state where heterozygotes have an intermediate phenotype to that of both homozygotes.

Following these definitions, we can build the following payoff matrix, where the maximum payoff is 1, and to which the costs of the respective genotypes are withdrawn (similarly to the payoff used by Tellier *et al.* (*Tellier, Moreno-Gamez et al. 2014*)):

π	R	C
R	$(1-c_b)$	$(1-c_b)*(1-c_h*p)$
C	$(1-c_b)*(1-c_h*p)$	$(1-c_i*p)$

In this payoff matrix, c_b is the cost of bleeding, c_i is the cost of infection for CC individuals, and c_h is the cost of infection for heterozygotes. Finally, p is the pathogen probability, defined as 0 or 1 in the exogenously changing environment, or the proportion of susceptible hosts in the frequency dependent environment. Finally, we used an exponential fitness mapping, leading to the following fitness matrix:

f	R	C
R	$\exp(1-c_b)$	$\exp((1-c_b)*(1-c_h*p))$
C	$\exp((1-c_b)*(1-c_h*p))$	$\exp(1-c_i*p)$

II.4 Population dynamics

Our model constrains the host population to a constant size N. We assume that each mouse transmits their strategy at a probability proportional to the fitness of the whole population. This corresponds to an evolutionary game in a Wright-Fisher process (*Imhof and Nowak 2006*). However, since mice are diploid sexual organisms, we added additional steps to the typical haploid "asexual" Wright-Fisher process. First, we selected one individual based on fitness; second we randomly (no mate choice) selected another individual without replacement; third, given the genotypes of the parents, we drew one offspring at random from the set of possible offspring. This process is repeated N times, so that the population always consists of non-overlapping generations. As a result, this method contains 3N random steps per generation.

Alternatively, we calculated the expected genotypes of the offspring population from the parent population weighted by its fitness, using Hardy-Weinberg equilibrium (HWE). First we calculated the weight W of each genotype according to their population frequencies P and their fitness *f* (1). Then we calculated the weighted allele frequencies A (2), and finally the offspring genotype frequencies O (3).

$$\begin{aligned} W_{RR} &= P_{RR} * f_{RR} \\ W_{RC} &= P_{RC} * f_{RC} \\ W_{CC} &= P_{CC} * f_{CC} \end{aligned} \quad (1)$$

$$\begin{aligned} A_C &= (W_{CC} * 2 + W_{RC})/2N \\ A_R &= (W_{RR} * 2 + W_{RC})/2N \end{aligned} \quad (2)$$

$$\begin{aligned} O_{RR} &= A_R^2 * N \\ O_{RC} &= 2 * A_R * A_C * N \\ O_{CC} &= A_C^2 * N \end{aligned} \quad (3)$$

As this calculation creates non-integer values, the results were rounded to the next lower integer and the difference to N was adjusted by randomly adding/removing individuals of any genotype. Hence this method contains only few random steps per generation.

Finally, the new generation - obtained either by the HWE-based or the Wright-Fisher like process - is allowed to mutate with a probability μ taken from a Poisson distribution. In our case, this is a proxy for migration, as the two *B4galnt2* alleles are highly divergent and thus unlikely to easily mutate from one state to the other.

II.5 Simulations

We ran every model with a local population size of 500 and a mutation rate of 0.005 over 10000 generations. All simulations were started with a random population that consisted of roughly 1/3 of each genotype. We varied the genotype-specific costs c_b, c_h and c_i, the environmental switch S, the starting environment and the definition of the pathogen. Each parameter combination was repeated 50 to 100 times. All the parameter combinations tested are summarized below:

Population Dynamics: Wright-Fisher process with random mating

- $c_h=0$, $c_h=c_i/2$, $c_h=c_i$
- c_b and c_i from 0 to 1 by 0.2 steps
- Mutation rate: $\mu=0.005$
- Environment:
    - Constant: 100 iterations for both environments
    - Frequency-dependent: 100 iterations
    - Switching environment:
        - Switch frequencies: 1, 10, 50, 100, 500, 1000, 5000
        - 50 iterations starting with a pathogenic environment + 50 starting with a non-pathogenic environment

Population Dynamics: HWE-based process

- $c_h=0$, $c_h=c_i/2$, $c_h=c_i$
- c_b and c_i from 0 to 1 by 0.05 steps
- Mutation rate: $\mu=0$, $\mu=0.005$
- Environment:
    - Constant: 100 iterations for both environments
    - Frequency-dependent: 100 iterations
    - Switching environment:
        - Switch frequencies: 1, 10, 50, 100, 250, 500, 750, 1000
        - 100 iterations starting with a pathogenic environment + 100 starting with a non-pathogenic environment

II.6 Results

We present two aspects of the model results: first the average population frequencies are the frequencies of each genotype in the host population averaged across the 10000 generations and across the 50 to 100 repetitions; second, the comparison to the real data consists of the average similarity (S) between the simulated (F) and the observed (O) population frequencies for each genotype (RR for RIIIS/J homozygotes, CC for C57BL/6J homozygotes and RC for heterozygotes):

$$S = 1-(abs(F_{RR}-O_{RR}) + abs(F_{RC}-O_{RC}) + abs(F_{CC}-O_{CC}))/3$$

The figures were produced in R using the reshape (*Wickham 2007*) and corrplot (*Taiyun Wei 2016*) packages.

Supplementary figures

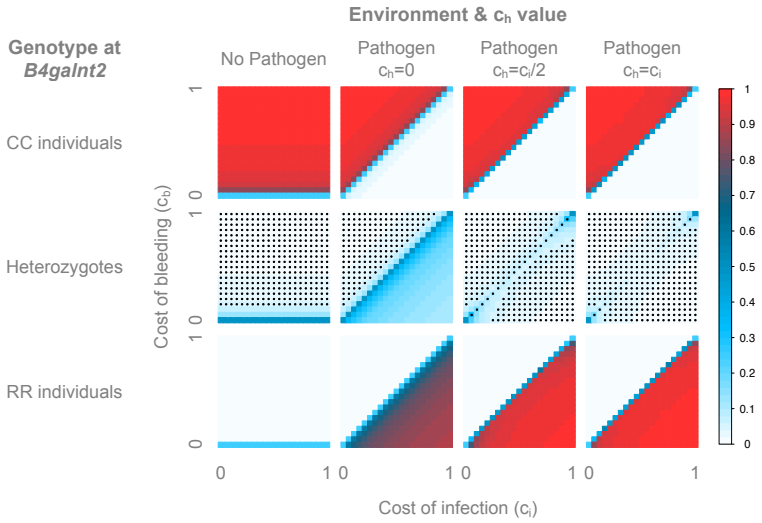

Figure S1: Average genotype frequencies in the model with a constant environment. The frequencies are displayed according to the cost of bleeding (y axis) and of infection (x axis). For the non-pathogenic environment (left panel), all c_h values are equivalent since c_i=c_h=0 in all cases. For the pathogenic environment however, the different models of c_h values are shown. The average genotype frequencies across 100 simulations using the HWE-process, each with 10000 generations, are displayed. The frequencies are color-coded according to the legend on the right. Stars indicate an excess of homozygotes.

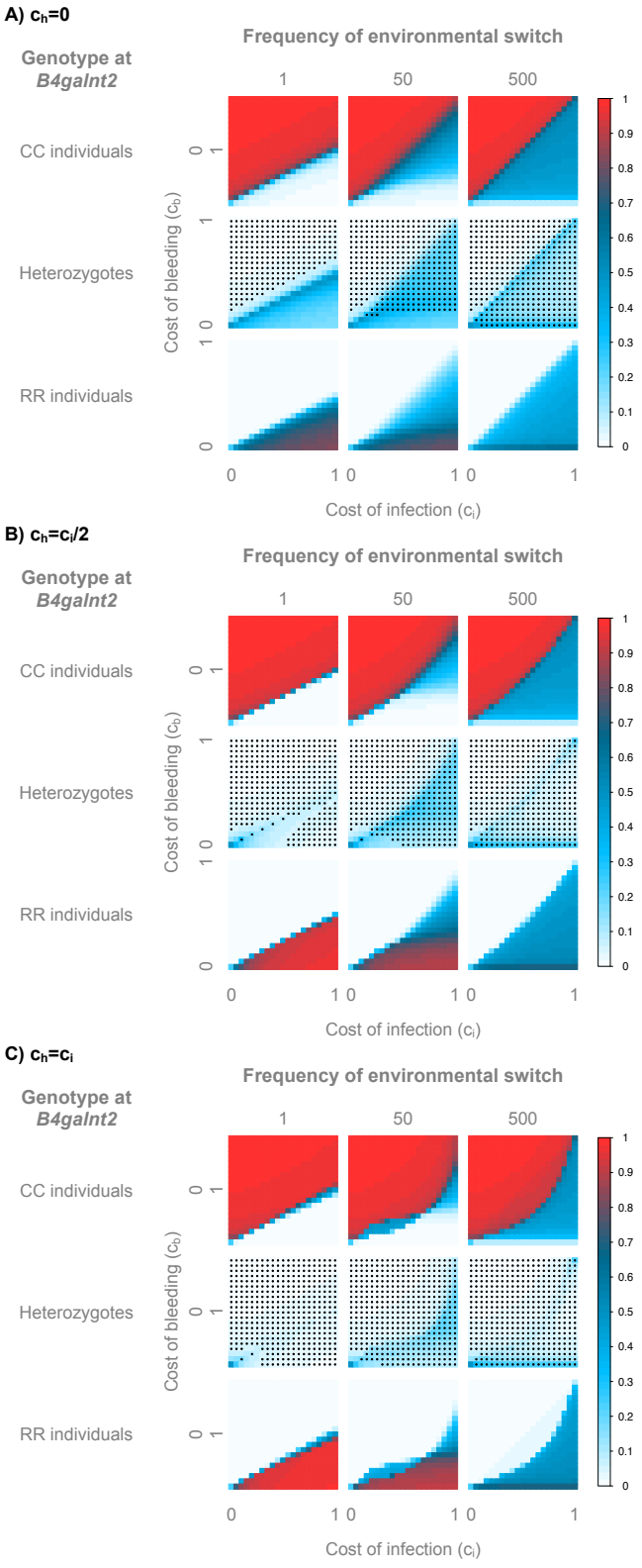

Figure S2: Average genotype frequencies in the model with a switching environment. The frequencies are displayed according to the frequency of environmental change expressed in host generations, the cost of bleeding (y axis) and of infection (x axis). The average genotype frequencies across 200 simulations using the HWE-process, each with 10000 generations, are displayed for c_h=0 (A), c_h=c_i/2 (B) and c_h=c_i (C). The frequencies are color-coded according to the legend on the right. Stars indicate an excess of homozygotes.

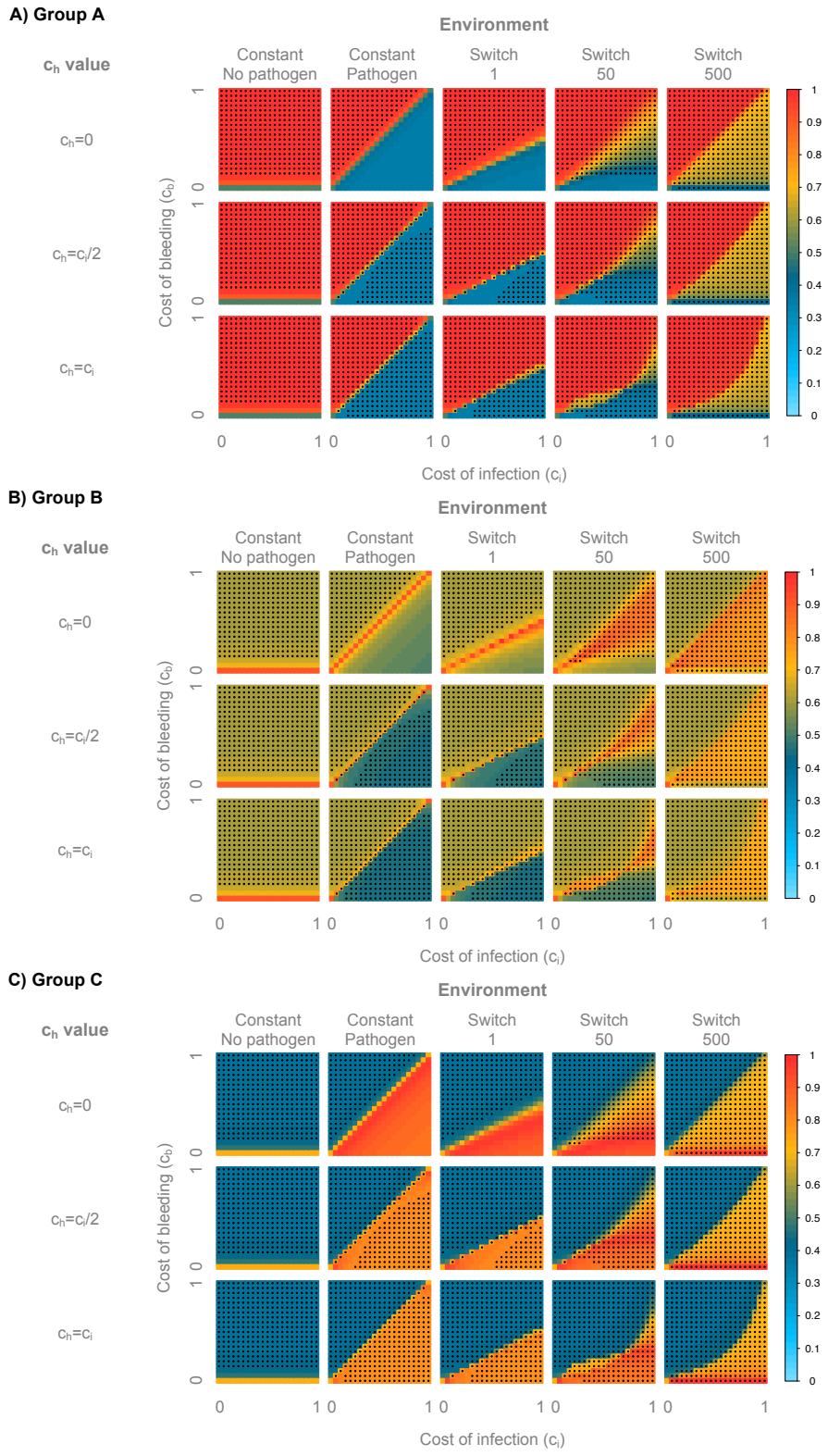

Figure S3: Similarity of the populations simulated with the HWE-process to the natural populations. A) Similarity to populations from Group A, B) Similarity to populations from Group B, C) Similarity to populations from Group C. The similarity is displayed according to the value of c_h, the cost of bleeding (y axis) and of infection (x axis), and the modeled environment (constant with or without pathogen, and switching between pathogenic and non pathogenic every 1, 50 or 500 host generations). The similarity is color-coded according to the legend on the right. Stars indicate an excess of homozygotes. Full similarity is achieved when all genotype frequencies coincide.

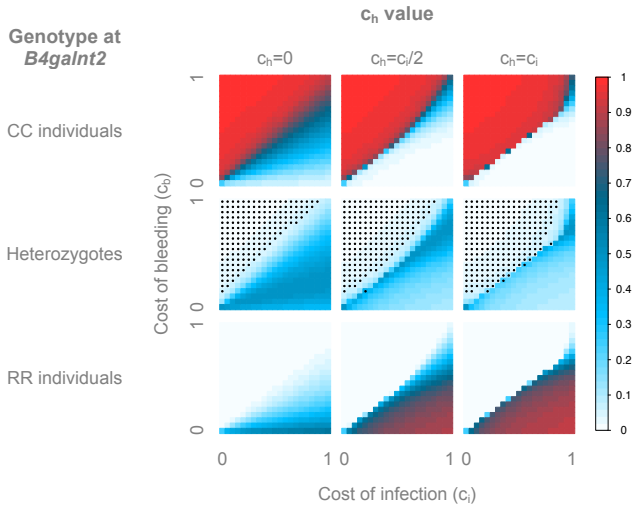

Figure S4: Average genotype frequencies in the model with a frequency-dependent environment. The frequencies are displayed according to the value of c_h, the cost of bleeding (y axis) and of infection (x axis). The average genotype frequencies across 100 simulations using the HWE-process, each with 10000 generations, are displayed. The frequencies are color-coded according to the legend on the right. Stars indicate an excess of homozygotes.

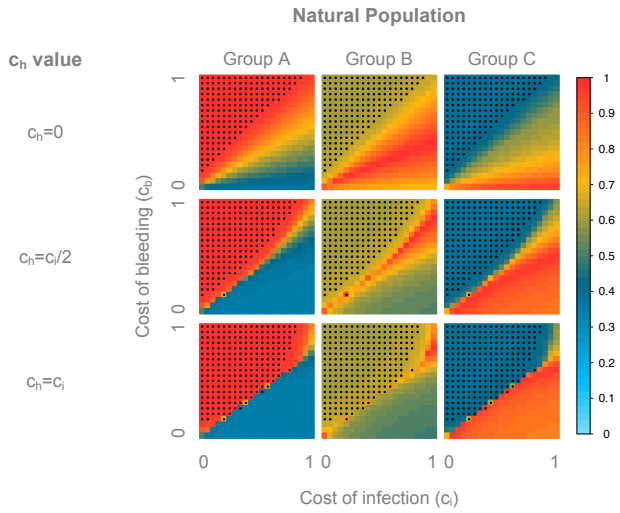

Figure S5: Similarity of the natural populations to the populations simulated in the model with a frequency-dependent environment, using the HWE-process. The similarity is displayed according to the value of c_h, the cost of bleeding (y axis) and of infection (x axis). The similarity is color-coded according to the legend on the right. Stars indicate an excess of homozygotes.

Chapter II:

Pathometagenomic analysis of a natural house mouse population: identifying candidate pathogens by 16S rRNA gene metabarcoding

Introduction

In the previous chapter, I confirmed the signs of selection observed in wild populations of *Mus musculus domesticus* and determined that pathogen-driven selection can maintain both alleles of *B4galnt2* in the host populations if the heterozygotes are protected against infection, i.e. if the vascular expression of *B4galnt2* carries the benefit of resistance against pathogens. These results, together with the previously published studies discussed in the general introduction, are very good arguments in favor of the pathogen-driven selection, but more direct evidence is needed.

To tackle this issue, I captured over 200 wild mice in Espelette, one of the French populations where a high RIIIS/J allele frequency was previously identified in the mice from *Linnenbrink et al.* (*Linnenbrink, Wang et al. 2013*), collected numerous organs, and performed an extensive analysis of various data, in order to identify potential candidate pathogens driving the selection at *B4galnt2*. In order to first describe these newly collected samples, I sequenced the mitochondrial D-loop, typed neutral microsatellite loci and finally performed genotyping at the *B4galnt2* region. Next, in order to identify candidate pathogens, I evaluated the health status of the mice, which I obtained through an extensive scoring of inflammation from histological slides from systemic organs and sections of the intestine. The strongest correlation with *B4galnt2* genotype was in the cecum, where RIIIS/J homozygotes seem to have lower prevalence of inflammation than C57BL/6J homozygotes, as well as lower inflammation scores. Additionally, I used qPCR to estimate the relative expression of two immune genes in the cecum, but failed to identify a correlation with *B4galnt2* genotype, as appeared for the histological scores. Finally, I used NGS sequencing of the V1-V2 region of the 16SrRNA gene to gain insight on the composition of the gastrointestinal microbiota of these newly collected mice. I focused on two parts of the intestine, the cecum and distal colon, and combined the microbial information with the *B4galnt2* genotype and inflammation data in order to identify potential candidate pathogens. This novel combination of methods allowed me to identify several candidate pathogens that could contribute to the selection acting on *B4glant2* in the wild. Among the candidates are *Citrobacter*, *Morganella* and *Proteus*, three genera known for their pathogenicity. Additionally, to investigate the alternative hypothesis of protection against a blood-pathogen, I used a PCR-based method to detect the presence of bacteria in blood samples, using bacterial universal primers for the 16SrRNA gene. Surprisingly, bacteria were present in a high proportion of mice, although I could not find a correlation with *B4galnt2* genotype, even when focusing on only one genus (*Staphylococcus*).

Results

I. Mouse collection

I collected 217 mice over a 5-weeks field work, coming from 34 farms around Espelette, France, which clusters into 12 groups of geographically close farms (figure 1).

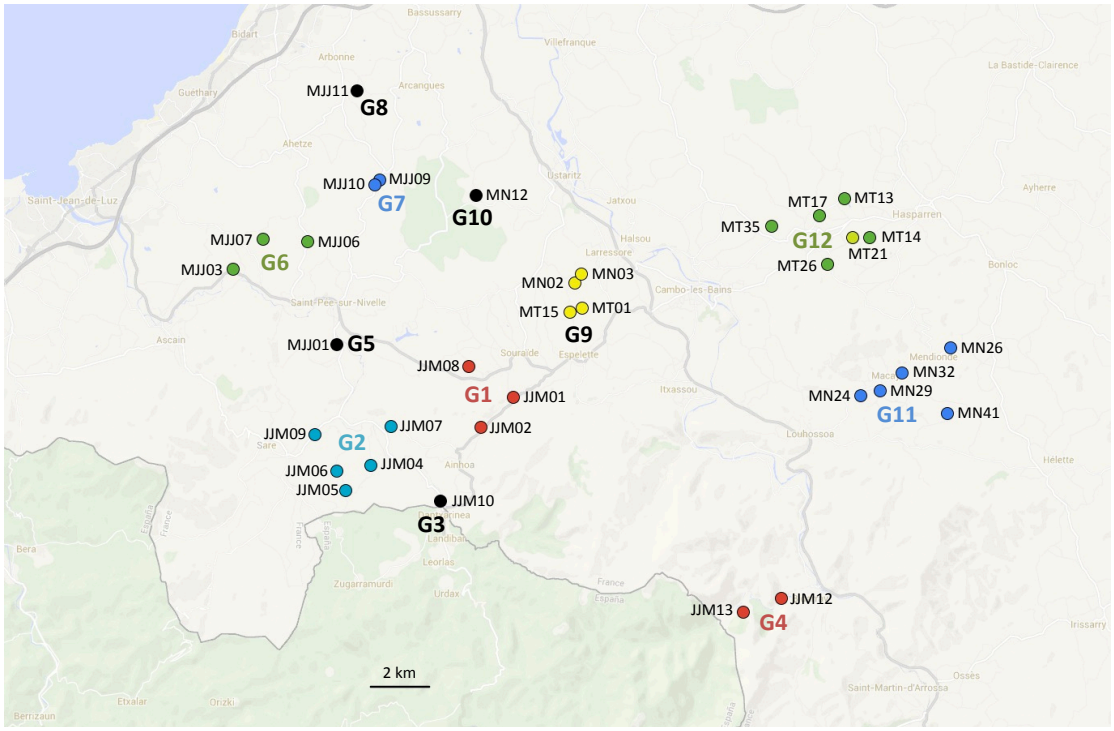

Figure 1: Geographic localization of the 34 farms around Espelette, FR, with the groups coded by colors, the colors are recycled for distant groups. The farm MT21 is displayed separately from the group G12, as from the genetic and microbial analysis, this farm behaves differently.

First, to verify that all the mice collected are *Mus musculus domesticus*, I sequenced an 885 bp portion of the mitochondrial D-loop and compared it to reference sequences of *Mus musculus domesticus*, *Mus spretus* and *Mus spicilegus* (*Linnenbrink, Wang et al. 2013*). I could confirm the *Mus musculus domesticus* classification for 216 mice. Moreover, I compared these mtDNA haplotypes to those from the previously described mice (Linnenbrink et al. (*Linnenbrink 2012*, *Linnenbrink, Wang et al. 2013*), thereafter referred to as "Linnenbrink mice") (figure 2), and found that the new Espelette samples (Espelette 2) cluster into 5 haplogroups, 4 of which where already present in the Espelette mice from the "Linnenbrink mice" (Espelette), the last one

was also present in the "Linnenbrink mice", but not in Espelette. This indicates that both datasets are comparable, despite the time period separating both collections (2009 vs. 2013).

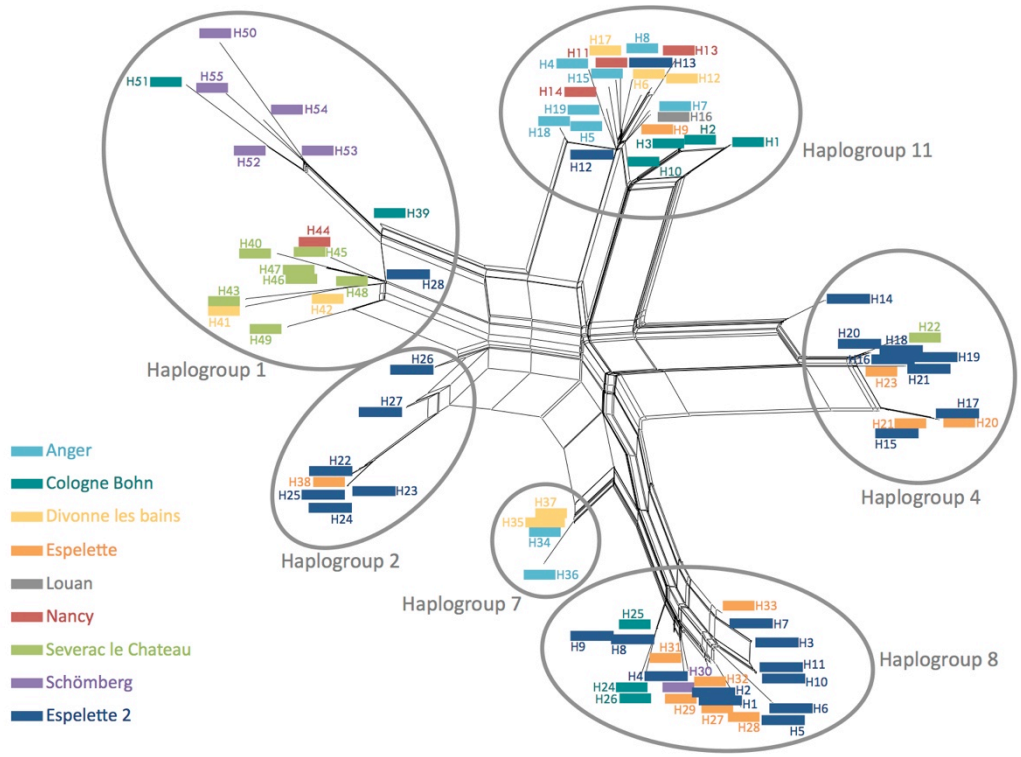

Figure 2: NeighboorNet network of mitochondrial D-loop haplotypes of the "Linnenbrink mice" (Anger, Cologne-Bohn, Divonne-lès-bains, Espelette, Louan, Nancy, Sévérac-le-Château and Schömberg) and the mice I collected in Espelette (Espelette 2). The haplogroup numbering corresponds to that of Bonhomme et al (*Bonhomme, Orth et al. 2011*).

Next, I evaluated the genetic structure of the newly collected mice. I first used KINGROUP (*Konovalov, Manning et al. 2004*) to estimate the relatedness between individuals, based on 18 neutral microsatellite loci previously described (*Thomas, Moller et al. 2007*, *Linnenbrink, Wang et al. 2013*). I found some degree of inbreeding within farms or groups of farms, as mice tend to be more related to individuals from the same farm/group compared to mice from a different farm/group (figure 3). The farm MT21 represents a particular case, as it was the only farm with a completely closed barn, making it safe from predators, providing an *ad libitum* food source with no threats from the farmer. When most farms show intermediate levels of inbreeding, the farm MT21 showed high level of inbreeding, and lower relatedness to other farms compared to the average between farm relatedness (figure 3). Thus, I separated this farm from the cluster G12 for later analysis, and named it G21.

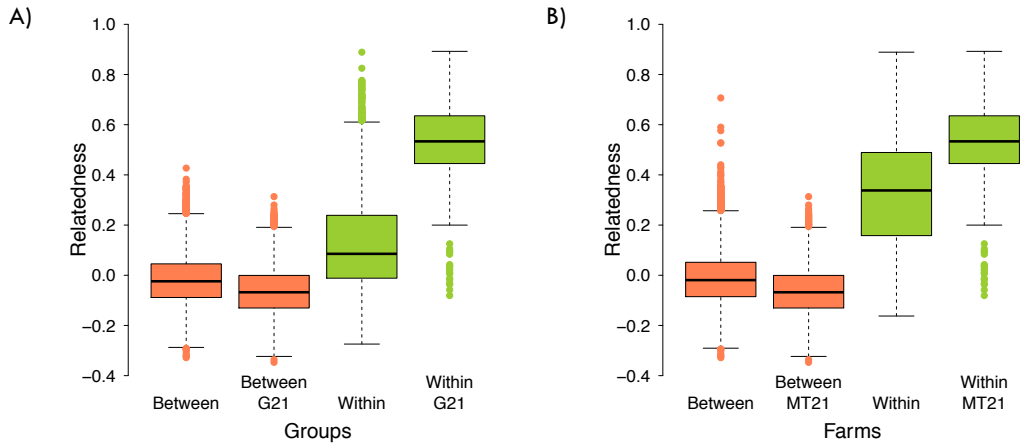

Figure 3: Pairwise relatedness, as estimated by the Kinship method from the KINGROUP software (*Konovalov, Manning et al.*), between/within mice from the same geographical groups **(A)** or farms **(B)**. "Between" refers to the relatedness between two mice that were caught in different farms/groups and that do not belong the MT21 farm. "Between_G21"/"between_MT21" refers to the relatedness between mice from the MT21 farms and mice from another farm. "Within" refers to the relatedness between mice from the same farm/group but not MT21. "Within_MT21" refers to the relatedness between mice that are both from the MT21 farm. All pairwise comparisons of the groups are highly significant ($p<0.001$).

Then, I estimated the population substructure using STRUCTURE (*Pritchard, Stephens et al. 2000*) and found that the best number of genetic clusters to be 13. Based on the cluster membership of each individual, I defined 16 homogenous and distinct populations (figure 4A). As expected from the estimated relatedness, there is some overlap between the populations from the STRUCTURE analysis and the geographical groups. Indeed, if two mice belong to the same genetic population (figure 4B), they also belong to the same geographical group (figure 4C). However, the reverse is not true, as there can be multiple genetic populations present within one geographical group. The admixed populations P07 and P14 are however spread across many geographical groups.

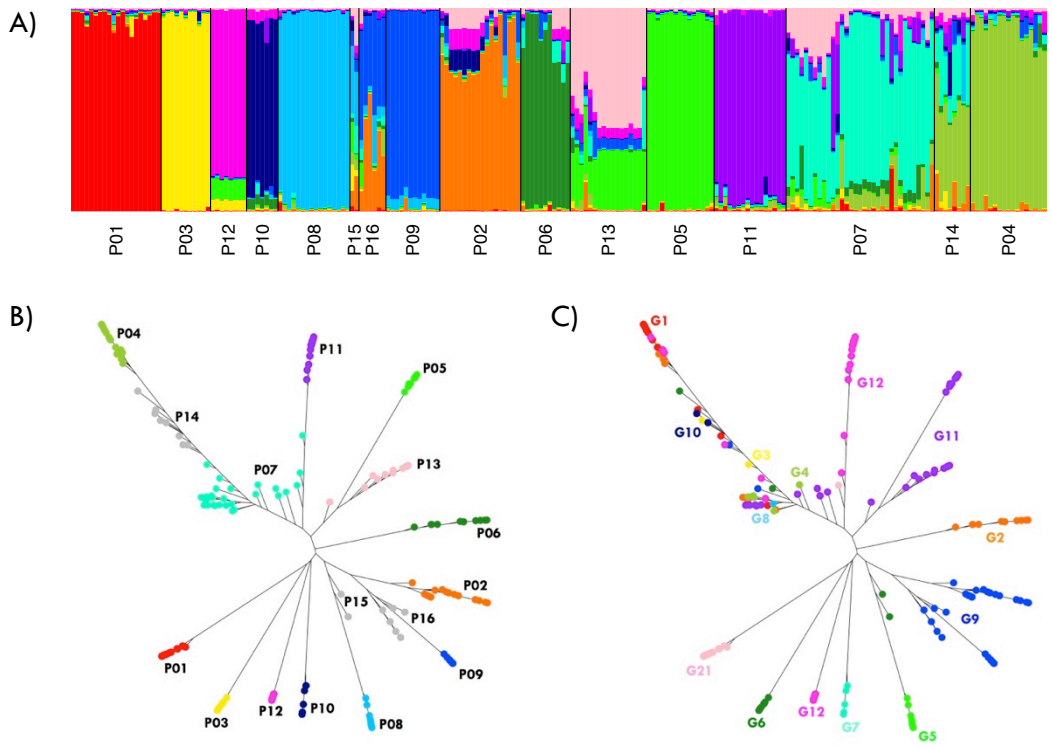

Figure 4: Visualization of the STRUCTURE (*Pritchard, Stephens et al. 2000*) results based on 18 neutral microsatellite markers. The results are the average of the 20 runs for K=13, determined as the best K by Structure Harvester (*Evanno, Regnaut et al. 2005*, *Earl and Vonholdt 2012*). The average was computed with CLUMPP (*Jakobsson and Rosenberg 2007*). **A)** STRUCTURE plot showing the cluster membership of each mouse (columns) and the deduced genetic population. **B)** Neighbor-joining tree based on the cluster membership from the STRUCTURE analysis. The color code refers to the genetic population described in A) and uses the color of the major cluster, except for the admixed populations P14, P15 and P16 in grey. **C)** Neighbor-joining tree based on the cluster membership from the STRUCTURE analysis. The color code refers to the geographic groups.

To further characterize the correlation between geography and genetics, I used a mantel test with the Pearson coefficient to test whether the geographic distance -- distance between farms in kilometers, expanded to the mice -- was correlated to the genetic distance, both at the genome level -- using the chord distance based on the 18 neutral microsatellite loci -- and for the mitochondrial genome -- using the p-distance based on the mitochondrial D-loop haplotypes. All three factors show a highly significant correlation with each other, but the nature of these correlations is difficult to grasp (figure 5). Indeed, the mitochondrial D-loop haplotype data do not seem to correlate very well with either factor. It might be that the extremely high number of data points inflates the significance of the correlations. The genetic to geography correlation, however, is consistent with the results from the STRUCTURE analysis (*Pritchard, Stephens et al. 2000*): if two mice are genetically similar, they belong to the same farm, but if two mice come from the same farm, they can be genetically distinct. It appears that at this small geographical scale, there is no isolation by distance, as mice from different farms tend to be genetically different, but this difference does not increase with the geographic distance.

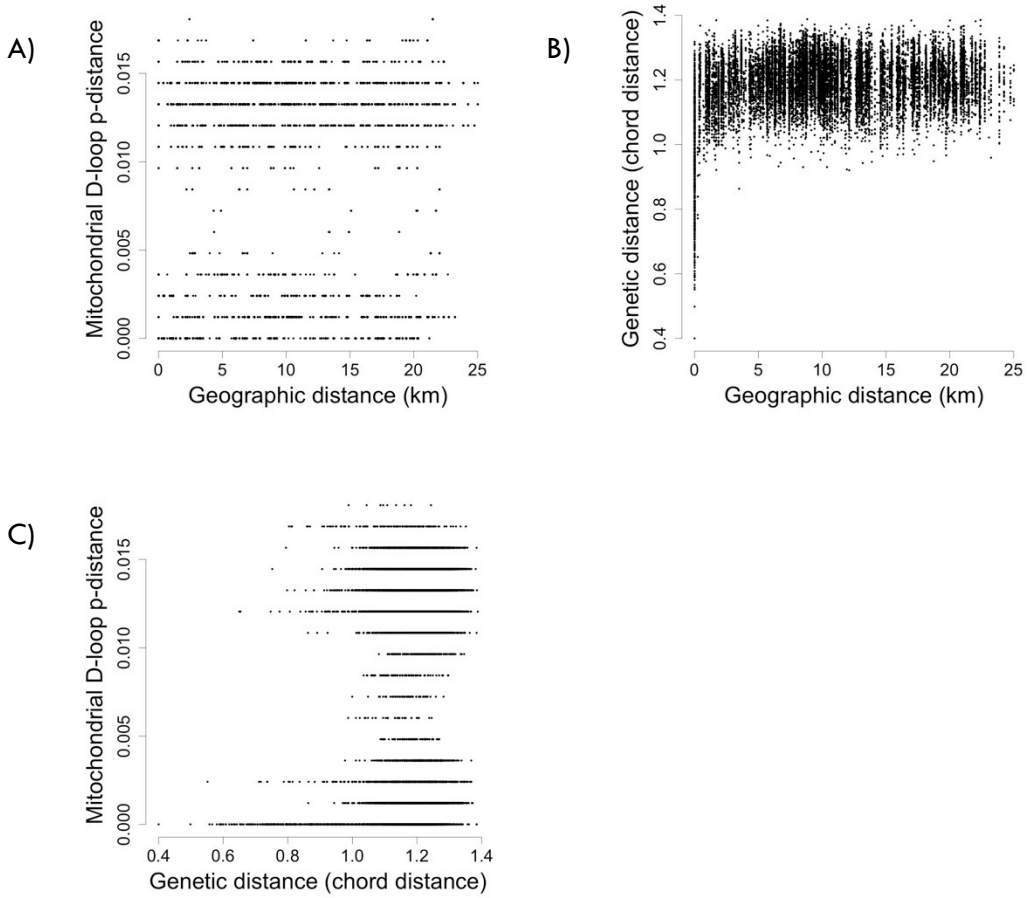

Figure 5: Correlation between the geographic distance in kilometer and the mitochondrial D-loop haplotype p-distance **(A)** or the genetic distance based on 18 neutral microsatellite loci **(B)**, and correlation between the mitochondrial D-loop haplotype p-distance and the genetic distance based on 18 neutral microsatellite loci **(C)**. All correlations are highly significant ($p<0.001$) with the mantel test, both with Pearson and Spearman correlation factors.

Next, I genotyped the *B4galnt2* cis-regulatory region for the 217 collected mice, using the diagnostic fragment 5, and identified 30 mice as RIIIS/J homozygotes, 125 as C57BL/6J homozygotes and 62 as heterozygotes. This corresponds to an RIIIS/J allele frequency of ~28%. It slightly differs from the frequency found in the "Linnenbrink mice" which was ~36% in Espelette, but given the important difference in sample size (217 versus 22) and the time period between both collections (4 years); this cannot be regarded as significant. The allele frequency per clusters of farms is highly heterogeneous, with frequencies varying all the way from 0 to 100%, but I could not identify a bias in the geographic distribution of *B4galnt2* alleles around Espelette (figure 6), which suggests that geography is not a confounding factor for any potential effect of *B4galnt2* on the microbiota.

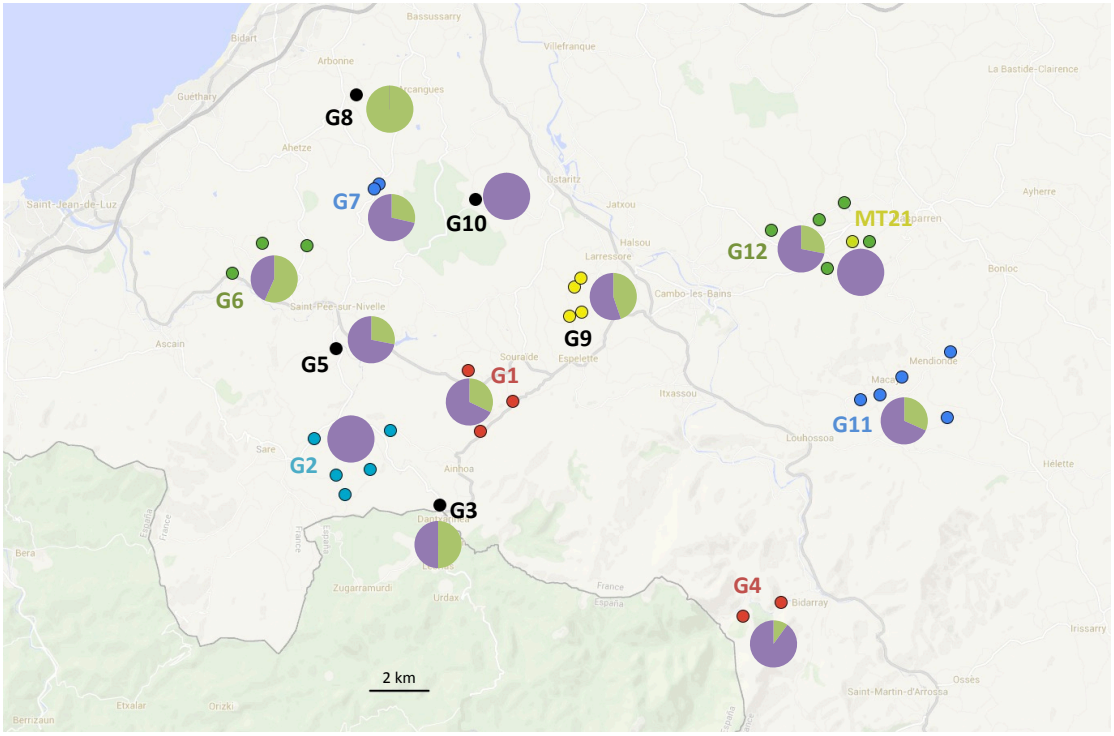

Figure 6: Geographic distribution of the *B4galnt2* allele frequencies (purple for C57BL/6J and green for RIIIS/J), expressed for each geographic group of mice. The farm MT21 is separated from the G12.

As the presence of farm animals or the use of poison by the farmers could influence the composition of the mice microbiota, I collected as much information as possible about each single farm (table 1). The nature of this data makes it however difficult to use. We can nonetheless conclude that *B4galnt2* allele frequencies do not seem to correlate with the farm characteristics, even when combining farms into geographical groups (table 2), which allows me to largely exclude the farm characteristics as a confounding factor for any potential effect of *B4galnt2* on the gut microbiota of the mice.

Finally, I typed the 12 *B4galnt2*-linked microsatellite loci previously described in Chapter 1, and observed the same partial selective sweep as in the Linnenbrink mice, with only two and three alleles on the RIIIS/J background for the loci ~30 kb upstream and at the start of *B4galnt2*, respectively (figure 7A), confirming once more that the RIIIS/J allele is subject to selection. The divergence between haplotypes is once again very high within the selective sweep (figure 7B), showing that the microsatellite alleles are very different between *B4galnt2* haplotypes around the mutation locus, but quite similar at distant loci, suggesting homogenization through recombination outside of the region under selection.

Table 1: Summary of the farm compositions in relationship to the *B4galnt2* genotype frequencies of the resident mice.

Population	Group	N	RIIISJ frequency		Hay	Cats	Dogs	Cow	Goat	Sheep	Pig	Horse	Donkey	Chicken	Turkey	Duck	Rabbit	Poison
MJJ11	G8	1	100.0	high	1		1	1										
MJJ06	G6	11	77.3	high	1	1	1	1						1				
MT15	G9	3	66.7	high	1			1										
MN32	G11	15	60.0	high	1		1			1								
MN02	G9	7	57.1	high	1	1	1	1										
MT01	G9	16	56.3	high	1		1					1						
JJM10	G3	2	50.0	high	1	1	1	1										1
MJJ09	G7	2	50.0	high	1	1		1										
MT35	G12	13	50.0	high	1		1			1	1							1
JJM02	G1	10	35.0	medium	1		1	1	1	1	1	1		1				
MJJ01	G5	16	28.1	medium	1		1	1			1							
JJM01	G1	2	25.0	medium	1								1	1				
JJM08	G1	2	25.0	medium	1	1	1		1	1								1
MN29	G11	2	25.0	medium	1	1	1	1										
MT13	G12	8	25.0	medium	1					1								
MJJ10	G7	5	20.0	medium	1	1	1	1						1	1	1		
MN03	G9	12	16.7	low	1		1	1										
MN26	G11	15	16.7	low			1	1										
MN41	G11	7	14.3	low	1	1	1	1										
JJM12	G4	4	12.5	low	1	1	1	1		1				1				
MT17	G12	7	7.1	low	1	1	1	1				1		1			1	
JJM04	G2	2	0.0	null	1	1	1			1				1				
JJM05	G2	4	0.0	null	1		1			1								
JJM06	G2	4	0.0	null	1		1			1				1				1
JJM07	G2	2	0.0	null	1	1		1										
JJM09	G2	11	0.0	null	1	1						1						
JJM13	G4	1	0.0	null	1	1	1		1	1		1						
MJJ03	G6	1	0.0	null	1		1	1						1				
MJJ07	G6	3	0.0	null	1			1			1							
MN12	G10	2	0.0	null	1	1	1				1			1			1	
MN24	G11	2	0.0	null	1		1	1										
MT14	G12	2	0.0	null		1	1			1								1
MT21	G12-21	21	0.0	null	1	1	1	1			1			1				
MT26	G12	2	0.0	null	1	1	1	1		1								

Table 2: Summary of the compositions of groups of farms in relationship to the *B4galnt2* genotype frequencies of the resident mice.

Group	N	RIIISJ frequency		Hay	Cats	Dogs	Cow	Goat	Sheep	Pig	Horse	Donkey	Chicken	Turkey	Duck	Rabbit	Poison
G8	1	100.0	high	1		1	1										
G6	15	56.7	high	1	1	1	1			1			1				
G3	2	50.0	high	1	1	1	1										1
G9	38	44.7	high	1	1	1	1				1						
G1	14	32.1	medium	1	1	1	1	1	1	1	1	1	1				1
G11	41	31.7	medium	1	1	1	1		1								
G7	7	28.6	medium	1	1	1	1						1	1	1		
G5	16	28.1	medium	1		1	1			1							
G12	32	28.1	medium	1	1	1	1		1	1	1		1			1	1
G4	5	10.0	low	1	1	1	1	1	1		1		1				
G10	2	0.0	null	1	1	1				1			1			1	
G2	23	0.0	null	1	1	1	1		1		1		1				1
G21	21	0.0	null	1	1	1	1			1			1				

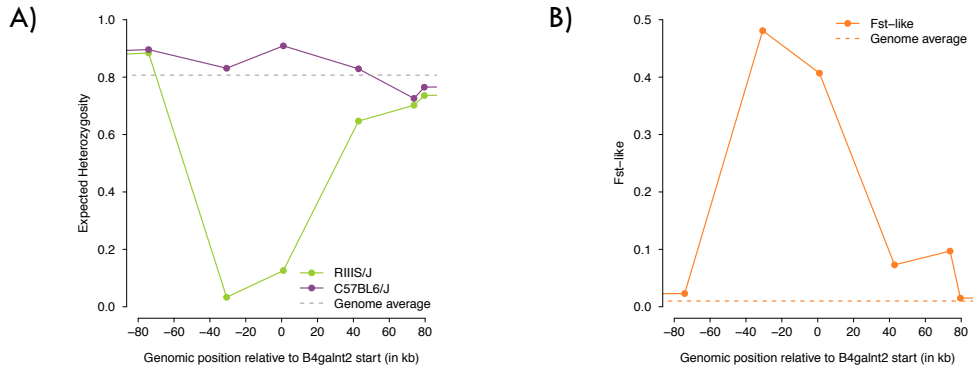

Figure 7: Microsatellite diversity in a 160 kb region around *B4galnt2* start position. The haplotypic phase was reconstructed to express variability according to *B4galnt2* allele. **A)** Expected heterozygosity. **B)** Fst-like measure of divergence.

II. Inflammation in mice as determined by histology

In order to assess the health status of the mice with regard to their *B4galnt2* genotype, I used the inflammation scores deduced from the level of inflammation visible on histological slides from various organs, reflecting either systemic inflammation (liver and spleen) or intestinal inflammation (ileum, cecum, colon proximal and colon distal).

In the liver, two main measures were used, the level of cellular toxic changes, and the general level of inflammation. The level of toxic changes is significantly higher in RIIIS/J homozygote mice compared to both C57BL/6J homozygote and heterozygote mice (figure 8A), but its prevalence does not differ between the three genotypes (figure 8B). For the general level of inflammation, the strength of inflammation does not differ significantly between genotypes (figure 8C), but the prevalence of inflammation is significantly lower in the RIIIS/J homozygote mice compared to heterozygotes only (figure 8D). This pattern seems however only driven by RIIIS/J homozygote mice that belong to the geographical group G6, as they have significantly higher levels of toxic changes than any other groups (figure 9A), and they show no inflammation, when all other groups have very high inflammation prevalence (figure 9B).

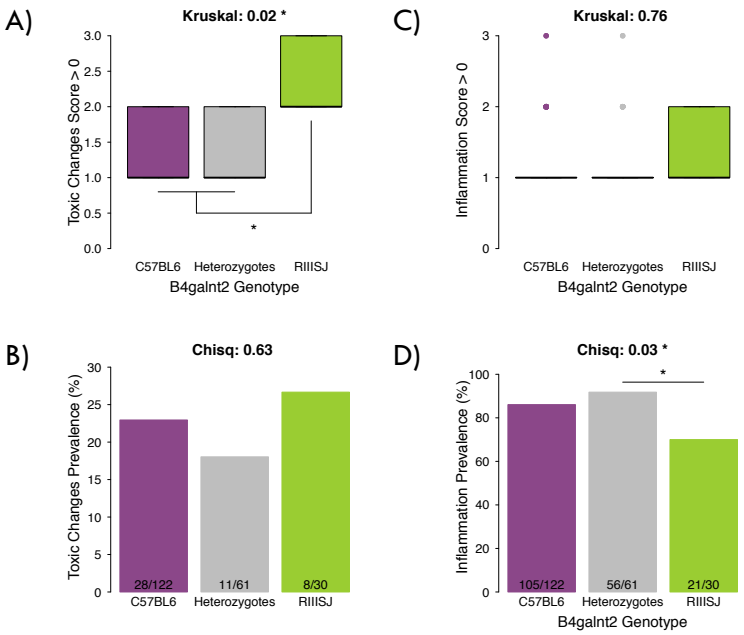

Figure 8: Liver histology parameters according to *B4galnt2* genotype. **A)** Strength of cellular toxic changes. **B)** Prevalence of cellular toxic changes. **C)** Strength of inflammation. **D)** Prevalence of inflammation. The p-values in the titles refer to the global test when stars in the graph refer to the pairwise group comparisons.

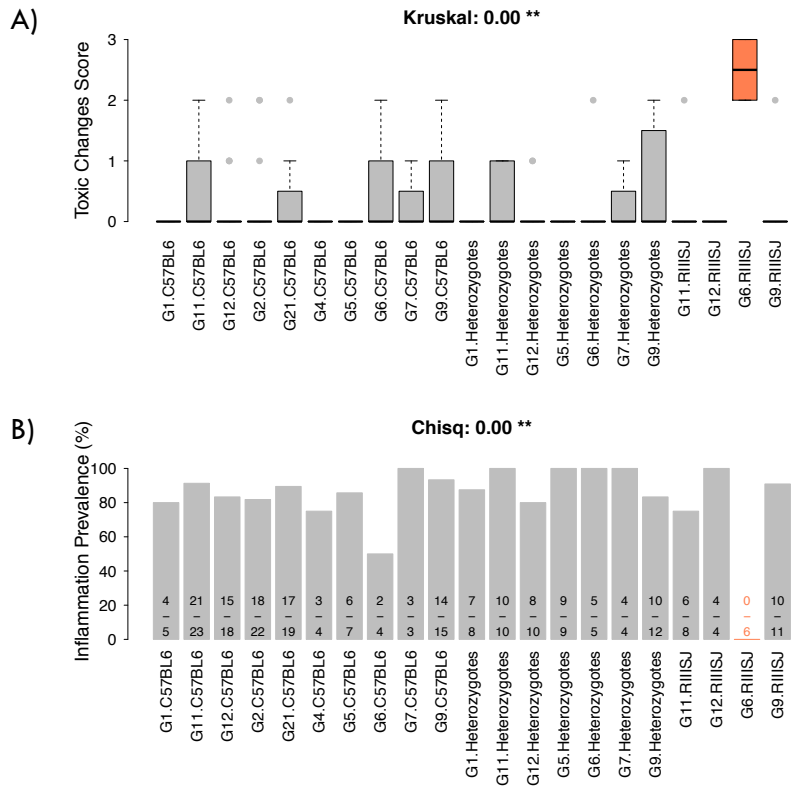

Figure 9: Liver histology parameters according to *B4galnt2* genotype and geographical group. **A)** Strength of cellular toxic changes. **B)** Prevalence of inflammation. The numbers in the bars indicate the sample sizes. The p-values in the titles refer to the global tests.

In the spleen, I used the measure of general inflammation, and the level of erythropoiesis and granulopoiesis, since they might be more important in RIIIS/J homozygote mice and heterozygotes, as a compensating mechanism for the increased blood loss due to the von Willebrand-like disease. Although no results reach significance, the general trend seems to be an increase in the prevalence of inflammation, erythropoiesis and granulopoiesis from the C57BL/6J homozygotes to the RIIIS/J homozygotes, with heterozygotes having an intermediate prevalence (figure 10).

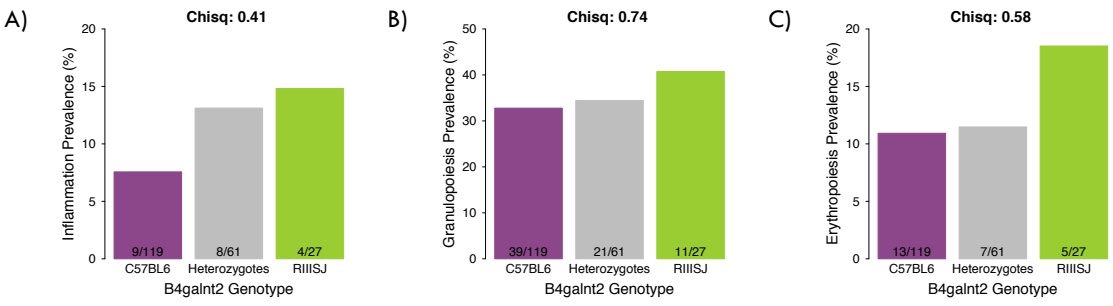

Figure 10: Spleen histology parameters according to *B4galnt2* genotype. **A)** Prevalence of inflammation. **B)** Prevalence of granulopoiesis. **C)** Prevalence of erythropoiesis. The p-values in the titles refer to the global tests.

For the individual sections of the intestine, the inflammation score is the sum of the levels of desquamation, necrosis and infiltration of polymorphonuclear leukocytes in the mucosa and the submucosa. In the ileum, the inflammation score do not differ according to *B4galnt2* genotype when taking all data into account (figure 11A). However, this kind of data generally presents a lot of zero values that can skew such analyses. To circumvent this issue, I chose to split the data in two sets of information: first the non-zero inflammation scores (figure 11B), presenting the strength of inflammation when present, then the prevalence of inflammation (figure 11C). With this data, it becomes clearer that there is no difference in prevalence of inflammation between the genotypes, but that when inflamed, the RIIIS/J homozygotes tend to have lower inflammation than the C57BL/6J homozygotes, with the heterozygotes being more similar to the C57BL/6J homozygotes. In the proximal colon, heterozygotes seem to have lower inflammation than both homozygotes (figure 12A). This effect however is no longer significant when splitting the data into inflammation score (figure 12B) and prevalence (figure 12C). In the distal colon (figure 13), no

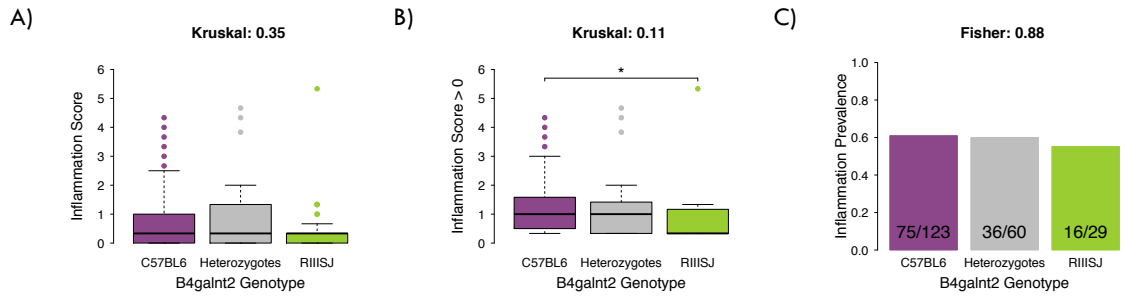

Figure 11: Ileum inflammation according to *B4galnt2* genotype. **A)** Inflammation score, defined as the sum of necrosis, desquamation, and infiltration of polymorphonuclear leucocytes in the mucosa and in the submucosa. **B)** Non-zero inflammation score. **C)** Prevalence of inflammation. The p-values in the titles refer to the global tests, while the stars in the graph refers to the pairwise tests, with the following code: p<0.001 ***, p<0.01 **, p<0.05 *.

Figure 12: Colon proximal inflammation according to *B4galnt2* genotype. **A)** Inflammation score, defined as the sum of necrosis, desquamation, and infiltration of polymorphonuclear leucocytes in the mucosa and in the submucosa. **B)** Non-zero inflammation score. **C)** Prevalence of inflammation. The p-values in the titles refer to the global tests, while the stars in the graph refers to the pairwise tests, with the following code: p<0.001 ***, p<0.01 **, p<0.05 *.

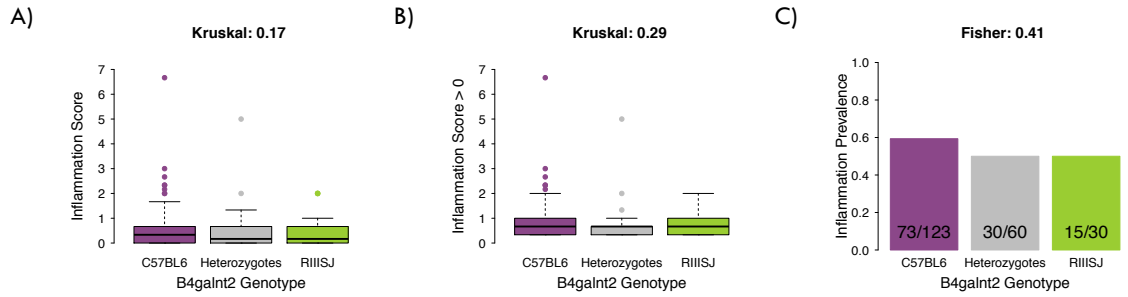

Figure 13: Colon distal inflammation according to *B4galnt2* genotype. **A)** Inflammation score, defined as the sum of necrosis, desquamation, and infiltration of polymorphonuclear leucocytes in the mucosa and in the submucosa. **B)** Non-zero inflammation score. **C)** Prevalence of inflammation. The p-values in the titles refer to the global tests, while the stars in the graph refers to the pairwise tests, with the following code: p<0.001 ***, p<0.01 **, p<0.05 *.

The strongest effect, however, is in the cecum (figure 14), as the RIIIS/J homozygotes have significantly lower inflammation than C57BL/6J homozygotes in the complete data (figure 14A), but also when splitting inflammation score (figure 14B) and prevalence (figure 14C). The heterozygotes seem to have an intermediate phenotype, but it reaches significance only in the full dataset. This strong genotype effect seems however to be driven by one group of mice, the G21 or farm MT21, which shows much stronger inflammation than any other group of mice (figure 15). If I remove this farm from the analysis, the trend is still present (figure 16), but only the difference in prevalence reaches significance.

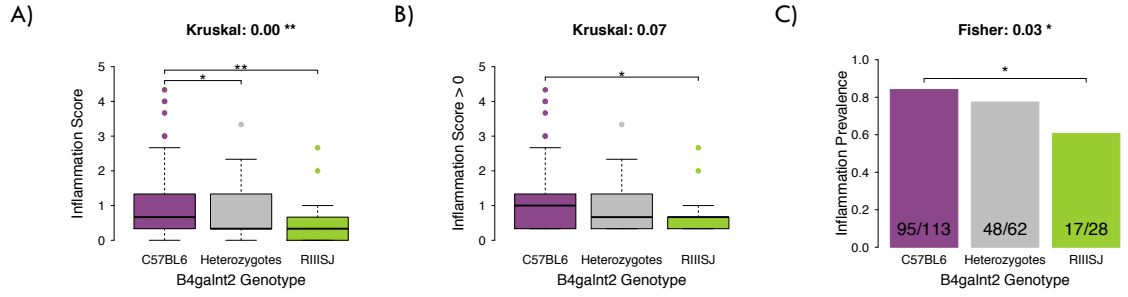

Figure 14: Cecum inflammation according to *B4galnt2* genotype. A) Inflammation score, defined as the sum of necrosis, desquamation, and infiltration of polymorphonuclear leucocytes in the mucosa and in the submucosa. B) Non-zero inflammation score. C) Prevalence of inflammation. The p-values in the titles refer to the global tests, while the stars in the graph refers to the pairwise tests, with the following code: p<0.001 ***, p<0.01 **, p<0.05 *.

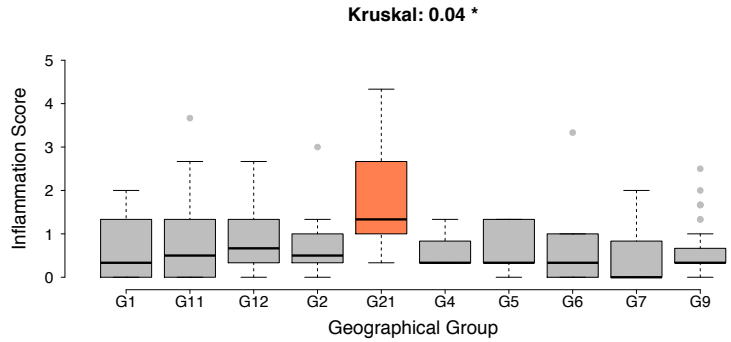

Figure 15: Cecum inflammation according to the geographical groups. The Inflammation score is defined as the sum of necrosis, desquamation, and infiltration of polymorphonuclear leucocytes in the mucosa and in the submucosa. The p-value in the title refers to the global tests.

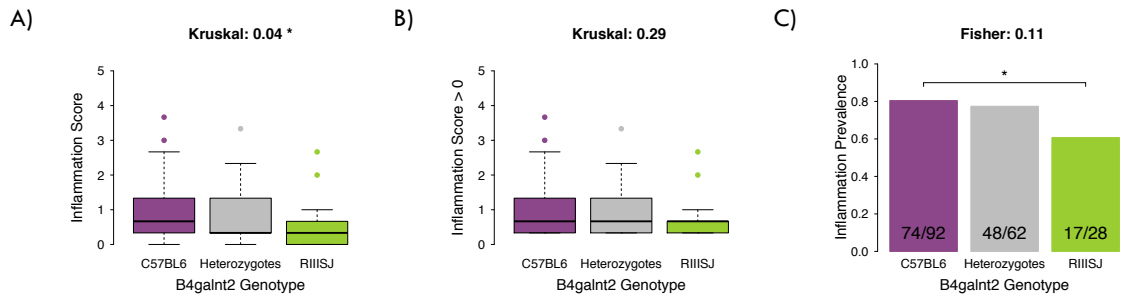

Figure 16: Cecum inflammation according to *B4galnt2* genotype, without the mice from the MT21 farm. A) Inflammation score, defined as the sum of necrosis, desquamation, and infiltration of polymorphonuclear leucocytes in the mucosa and in the submucosa. B) Non-zero inflammation score. C) Prevalence of inflammation. The p-values in the titles refer to the global tests, while the stars in the graph refers to the pairwise tests, with the following code: p<0.001 ***, p<0.01 **, p<0.05 *.

In conclusion, the systemic inflammation level, as determined by histological scores do not seem to significantly correlate with *B4galnt2* genotype, as the only detectable effect is driven by only one genotype in only one geographical group of mice, which do not appear to have peculiar condition that would justify this pattern. The intestinal inflammation however, does correlate with *B4galnt2* genotype, but the direction varies according to the part of the intestine, with heterozygotes being less inflamed than both homozygotes in the colon proximal, while RIIIS/J homozygotes are less inflamed than C57BL/6J homozygotes in the ileum and the cecum, even if this effect is to some extent driven by the peculiar farm MT21.

III. Inflammation in mice as determined by expression of immune genes in the cecum

To gain a deeper understanding of the inflammation in the cecum, where the strongest effect is detected by histology, I performed qPCR on two immune genes: MCP1 and IFNgamma. As both genes are known inflammation markers, not unexpectedly they show some degree of correlation (figure 17), although the strength of the correlation seems to be dependent on *B4galnt2* genotype. Surprisingly, both immune genes relative expression show only limited correlation to the inflammation in the cecum as determined by histology, when taking the strength of inflammation into account (figure 18). However, when considering the histological score as binary (i.e. either inflamed or healthy), the correlation with both immune genes becomes visible and significant (figure 19), despite no correlation to *B4galnt2* genotype, even when looking at the interaction of *B4galnt2* genotype and cecal inflammation. In conclusion, the two immune genes studied display limited correspondence to the inflammation score deduced from histology, and no correlation to *B4galnt2* genotype. As such, the histology measures might represent a better indicator for the purposes of this study than the relative expression of MCP1 and IFNgamma, so I chose to use only the histological score for further analysis.

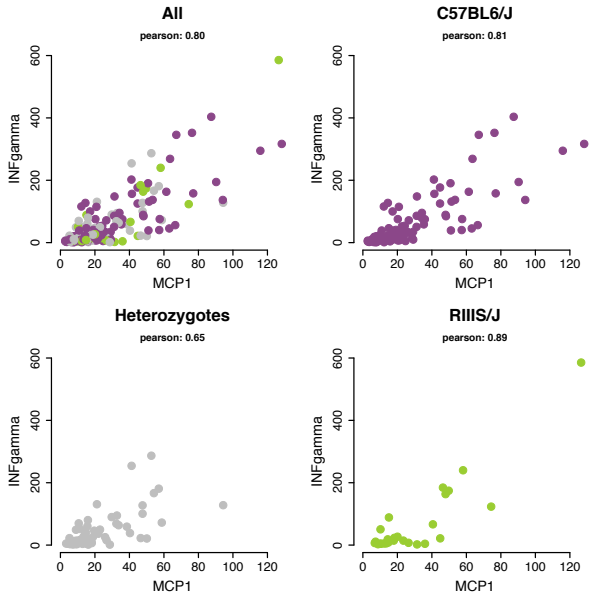

Figure 17: Correlation of the relative expression of two inflammatory genes, INFgamma and MCP1, in the cecum. These relative expressions were measured by qPCR, using HPRT1 as reference gene and the ΔΔCt method, with the sample with the lowest expression as reference to set the minimum value to one.

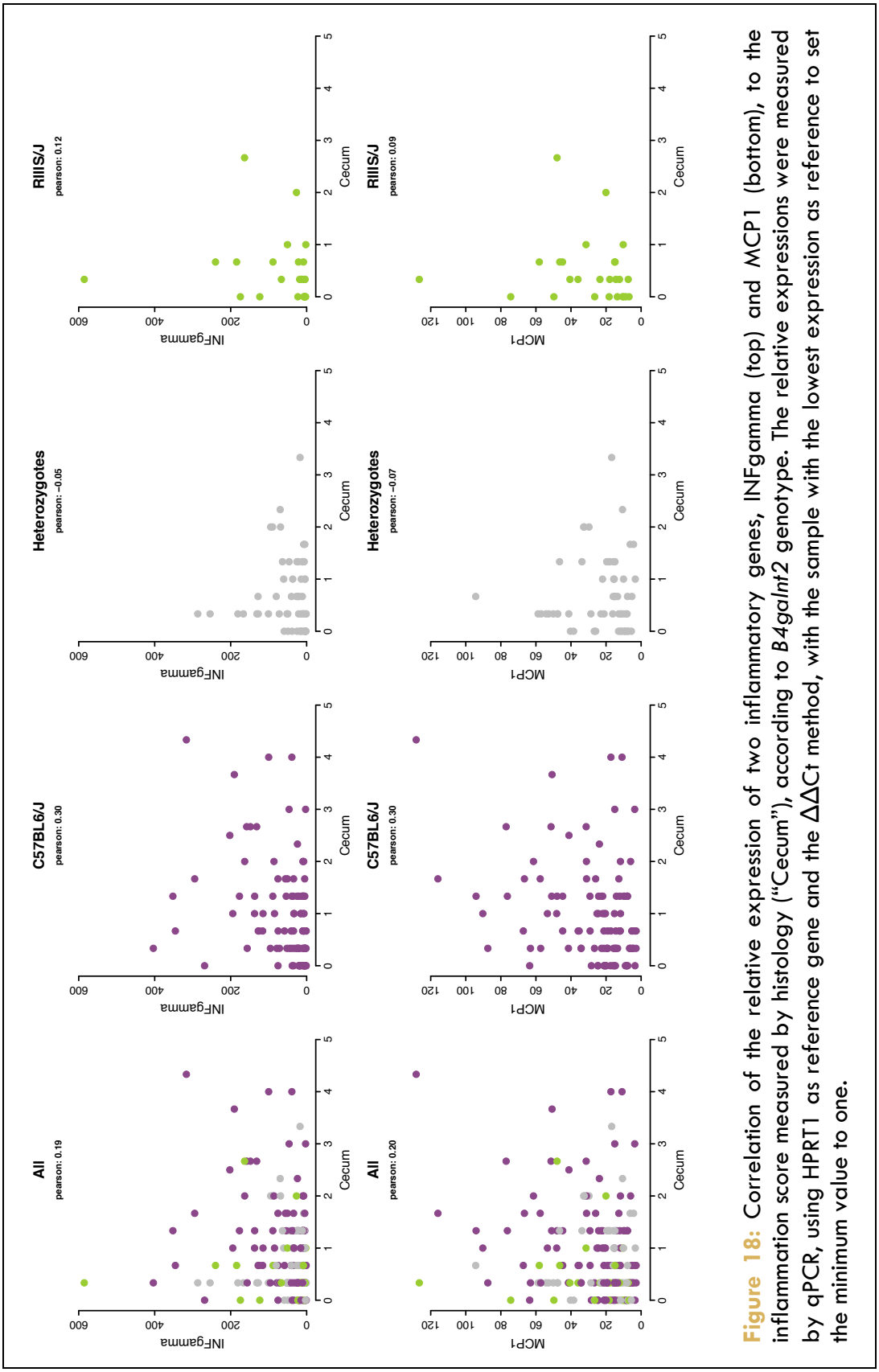

Figure 18: Correlation of the relative expression of two inflammatory genes, INFgamma (top) and MCP1 (bottom), to the inflammation score measured by histology ("Cecum"), according to *B4galnt2* genotype. The relative expressions were measured by qPCR, using HPRT1 as reference gene and the ΔΔCt method, with the sample with the lowest expression as reference to set the minimum value to one.

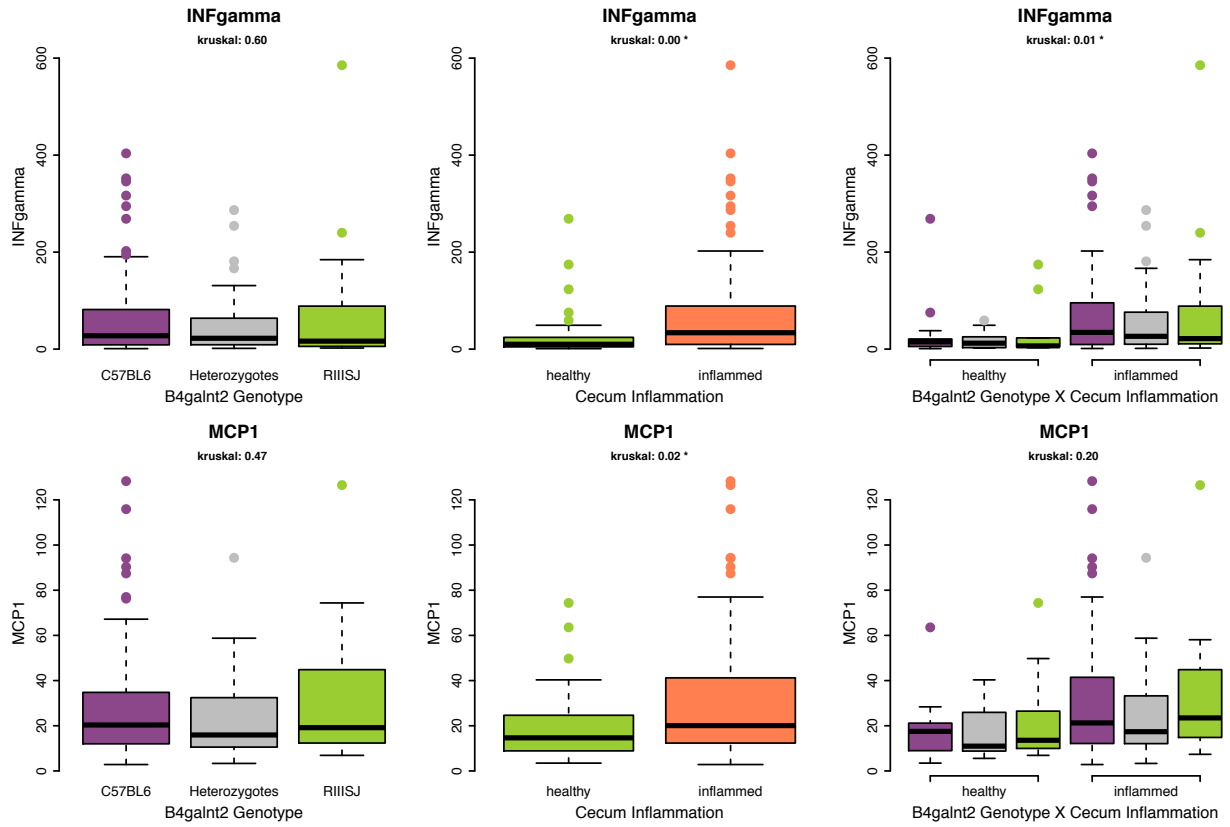

Figure 19: Correlation of the relative expression of two inflammatory genes, INFgamma (top) and MCP1 (bottom), in the cecum to *B4galnt2* genotype (left), cecum inflammation as measured through histology (middle) and their interaction (right). The relative expressions were measured by qPCR, using HPRT1 as reference gene and the ΔΔCt method, with the sample with the lowest expression as reference to set the minimum value to one.

IV. Cecum microbiota

After establishing the relationship between *B4galnt2* genotype and inflammation in the cecum, I used this information in conjunction with microbiota sequencing data in order to identify potential pathogens linked to *B4galnt2* genotype. Accordingly, I sequenced the V1-V2 region of the 16S rRNA gene to assess the composition of the cecum microbiota, both at the DNA and RNA level. DNA typically reveals every present bacteria, dead or alive, active and inactive, as well as "passers by" that might be brought by the food. RNA reveals active bacteria that might be growing or undergoing several cellular processes, and should present a more "functional" picture than the DNA. Additionally, I chose to express the activity of bacteria as the ratio of RNA to DNA. This method, like the reference gene in a qPCR, normalizes the activity of the bacteria (RNA) to the number of bacteria present (DNA). This however yield numbers difficult to compare between samples, and I thus chose to normalize the activity within samples, as if the entire community had an activity of 1, spread across different taxa. For the activity, I present only abundance-based

analysis (Bray-Curtis dissimilarity) since prevalence data for the activity are equal to the prevalence data for the RNA, due to the calculation of the activity. Indeed when dividing RNA by DNA, only the numerator (RNA) determines whether the resulting ratio is null or positive.

IV.1 SourceTracker

One pitfall of any microbial community analysis is the risk of contamination through the ubiquitous bacteria present in the environment. To ensure that my data are free of contamination, I use negative controls during the DNA extraction, as well as during the PCR. All PCR negative controls (i.e. "water blanks") were required to be negative, but despite the precautions taken during laboratory procedures, some batches of samples did produce signal in the negative extraction control (i.e. material was either introduced during extraction or was already present in the reagents). This, however, is typically not a serious issue when dealing with gut samples, as the bacterial load, and thus "signal-to-noise" ratio, is very high (*Salter, Cox et al. 2014*). Nonetheless, I used SourceTracker (*Knights, Kuczynski et al. 2011*) to estimate the potential contamination of my samples, and fortunately, only two samples, one each in the RNA and DNA, showed strong contamination (figure 20) and were subsequently removed them from the analysis, along with their RNA/DNA counterparts.

Figure 20: Estimation of the contribution of contamination to the microbiota composition of each sample, as produced by SourceTracker (*Knights, Kuczynski et al. 2011*). The left graph presents the results for the extraction 11 ("E11") in the DNA samples, and the right graph presents the data for the extraction 4 ("E04") in the RNA samples. Each bar represents one sample that was extracted in the corresponding extraction. The color red shows the estimated amount of contamination from the current extraction, the color blue shows the estimated amount of contamination from other extractions, and the "unknown" green refers to the contribution of the sample.

IV.2 Experimental variables

Another pitfall of microbial community analysis is the abundance of confounding factors. Indeed, bacteria are present everywhere, and communities can be influenced by environment, diet, hygiene, and of course genetics of the hosts. Since my goal is to find bacteria that correlate with *B4galnt2* genotype and inflammation, I first had to ensure that other factors are not responsible for putative *B4galnt2* genotype effects.

First, I evaluated experimental variables such as the sequencing library the samples were run in, the date of capture and dissection of the mice, the team that dissected them and the extraction batch. All variables had a significant and non-negligible effect on the microbiota, both at the DNA (figure S1) and the RNA (figure S2) level. These parameters are, however, partially confounded with the farms the mice were caught from. Farms are the wild equivalent of cages in laboratory conditions, whose effects are well documented. Since mice in farms tend to be in close contact with each other, in a relatively close environment, the microbiota of mice coming from he same farm is likely to be more similar than to other mice. Thus, I repeated the analysis, but this time conditioning on the farms. This time, in the DNA (figure S3), only the MiSeq Library was still significant, while no other factor remained significant in the RNA (figure S4). For the DNA, if we look at the library, we clearly see one batch separating from the rest (library 4 in pink). As this represents only a small amount of samples, I decided to exclude them from the analysis. When repeating the analysis without these samples, no factors were significant, neither in the DNA (figure S5) nor in the RNA (figure S6).

Next, I analyzed the microbiota with regard to the farms. As expected, there is a strong farm influence on the microbiota (figure S7) at the DNA, RNA and activity level, based on both count data (Bray-Curtis diversity measure) and prevalence data (Jaccard diversity measure). One farms sticks out for DNA and RNA, both for Bray-Curtis: the farms MT21 (in red). Given that this farm already presents peculiarities in terms of environment, genetics and intestinal inflammation, I chose to exclude those samples from the analysis (figure S8).

Finally, I looked at intrinsic variables, such as genetic structure and gender, but also weight, length and BMI, which can serve as proxies for the age of the mice, a parameter known to influence the composition of the microbiota in mammals. At the DNA level (figure S9), there is no influence of the genetic cluster, the mitochondrial D-loop haplotype/haplogroup or the gender. At the RNA level, however (figure S10), the mitochondrial D-loop shows significant influence on the microbiota, both at the haplotype level and at the less precise haplogroup level (as defined in

Bonhomme et al. (*Bonhomme, Orth et al. 2011*)), for Bray-Curtis and Jaccard dissimilarities. At the activity level (figure S11), the mitochondrial D-loop is only significant at the less precise haplogroup level. For the body characteristics, no factor significantly influences the microbiota at the DNA level (figure S12), but the weight and BMI significantly influences the microbiota at the RNA (figure S13) and activity level (figure S14), the explained variance is however quite low.

In conclusion, beside the strong influence of the farms, as one might expect from microbial analysis, only the mitochondrial D-loop seems to have a significant and strong influence on the microbiota. This influence is however only seen at the RNA and activity level.

IV.3 *B4galnt2* & Inflammation

Now that I verified the influence of potential confounding factors, I can test whether *B4galnt2* genotype (figure S15), cecal inflammation (figures S16 & S17) and their interaction (figure S18) influences the microbiota. None of these factors show significant correlation with the microbiota composition, and the variance explained by the ordination axes is very low. This indicates that the overall composition of the microbiota does not differ according to *B4galnt2* genotype or inflammation in these wild samples, suggesting that if *B4galnt2* has an effect on pathogen resistance/susceptibility as hypothesized, it is likely to be a precise effect, only targeting one or few bacteria but not changing the composition of the microbiota dramatically.

IV.4 Indicator species analysis

To identify candidate pathogen(s) that might be driving selection at *B4galnt2*, I performed an indicator species analysis on species-level OTUs as well as on the genus level of classification. For this, I used multiple techniques: (i) indicator analysis *per se*, using the function MULTIPATT of the package INDICSPECIES (*De Caceres and Legendre 2009*, *De Caceres, Legendre et al. 2010*), (ii) the function simper from the vegan package (*Dixon 2003*), (iii) Kruskal-Wallis tests on abundance data and (iv) chi square tests on prevalence data. To tackle the issue of spurious zeros in the data set, I performed the Kruskal-Wallis test on the complete data set, but also on the non-zero data set, similar to the histology analysis. I performed these analysis using *B4galnt2* genotype, cecal inflammation and their interaction as explanatory variables. For cecal inflammation I categorized the data in two ways: first with the two categories "healthy" (inflammation score = 0) and "inflamed" (inflammation score > 0); then with the three categories "null" (inflammation score = 0), "low" (inflammation score <= 1), and "high" (inflammation score > 1).

Importantly, these analyses allowed me to identify three indicator genera and four indicator species. The indicator genera are *Citrobacter*, *Morganella* and *Proteus*, while the species-level indicator OTUs belong to *Allistipes*, unclassified *Bacteroidales*, *Morganella* and *Proteus*. The indicator OTUs belonging to *Morganella* and *Proteus* make up the abundance of nearly all their respective genera, while the remaining OTUs belonging to these genera are single occurrences of very low abundance. Although I detected a strong influence of the farms and the mitochondrial D-loop on the microbiota, I didn't include these parameter in the tests, as the small sample size per farm/haplotype would have drastically reduced the detection power. However, I verified *a posteriori* that the association of the indicator OTU to *B4galnt2* genotype and/or inflammation was not confounded by any other factors (tables S1-S5).

For ***Citrobacter*** at the DNA level (figure 21.1), no correlation reaches significance, but a higher prevalence is nonetheless observed in C57BL/6J homozygotes than in heterozygotes, and RIIIS/J homozygotes do not present any *Citrobacter*. The abundance also seems higher in C57BL/6J homozygotes than in heterozygotes. Moreover, it seems that the abundance of *Citrobacter* is higher in inflamed mice compared to healthy mice, although its prevalence is not. In the RNA (figure 21.2), the trend is globally the same as in the DNA, although one RIIIS/J mouse has *Citrobacter*, and the prevalence seems lower in inflamed mice than in healthy mice. This time, the difference between genotypes reaches significance in one MULTIPATT test. It is at the activity level (figure 21.3) that *Citrobacter* passes the threshold I used for defining indicator species, as the correlation with inflammation is significant in multiple tests. Interestingly, the activity of *Citrobacter* is lower in inflamed mice compared to healthy mice. This might indicate that inflammation in response to *Citrobacter* is indeed effective in fighting against this pathogen, and/or different strains are present in inflamed vs. healthy mice. An alternative explanation would be that *Citrobacter* can act as a probiotic, keeping mice healthy if active enough in the microbial community. However, the fact that *Citrobacter* is more abundant (DNA) but less active (RNA/DNA ration) in inflamed mice compared to healthy mice might suggest some sort of quorum sensing, which *Citrobacter* is known to use (*Gupta, Moller et al. 2016*), although quorum sensing has only been shown to increase virulence, not decrease it (*Gupta, Moller et al. 2016*). Thus, the observation may be more consistent with the former hypothesis. It is in any case an important observation to find *Citrobacter*, a common member of the intestinal flora but also well characterized pathogenic genus, as indicator species in wild mice, although the trend in the *B4galnt2* genotype correlation is not significant, as it was already identified as *B4galnt2* indicator in an earlier study (*Staubach, Kunzel et al. 2012*).

Figure 21.1

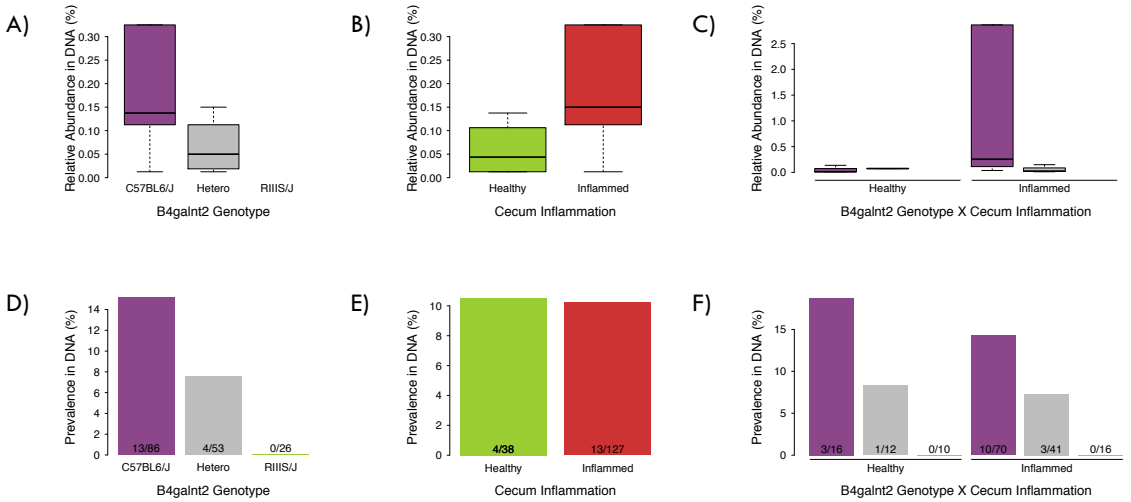

Figure 21.2

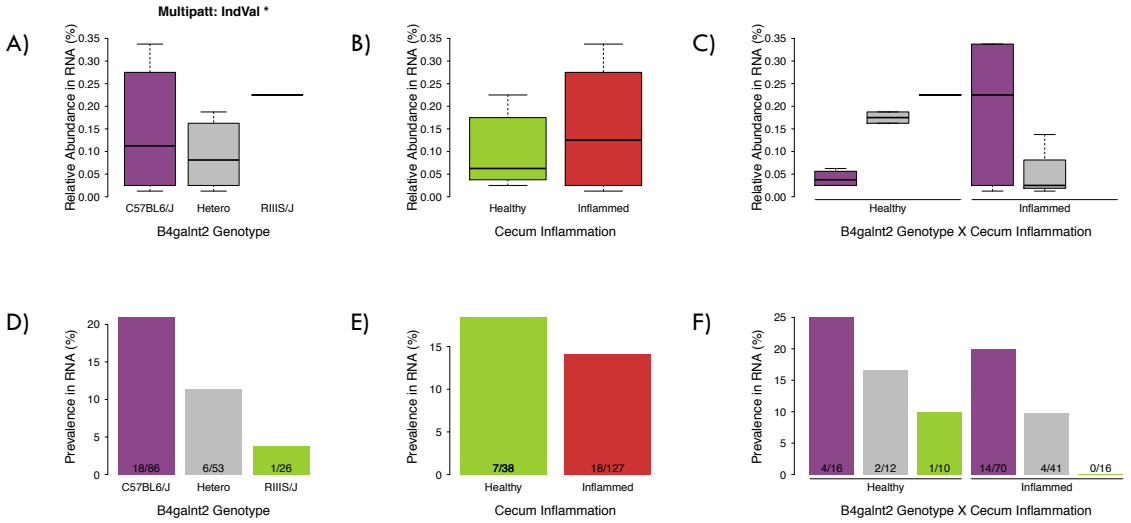

Figure 21.3

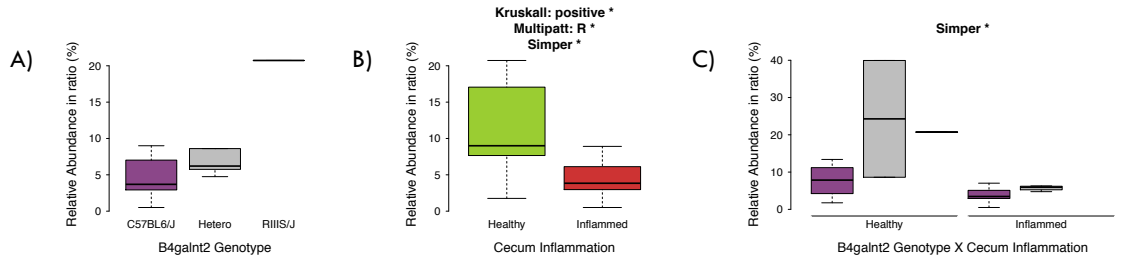

Figure 21: Non-zero relative abundance (ABC) and prevalence (DEF) of the indicator genus *Citrobacter* (*Proteobacteria; Gammaproteobacteria; Enterobacteriales; Enterobacteriaceae*) in the DNA (21.1) and RNA (21.2) samples, and at the activity level (21.3), according to the *B4galnt2* genotype (AD), the cecum inflammation (BE), and their interaction (CF). The test performed were three indicator analysis functions: multipatt R, multipatt IndVal, and simper; the Kruskal-Wallis test on all values and only on non-zero values; and a chi square test on the prevalence. The results of the tests are reported in the title of the graphs with the following code: p<=0.001 ***, p<=0.01 **, p<=0.05 *. Outliers are not displayed.

For the OTU belonging the genus ***Alistipes*** (Otu000311) at the DNA level (figure 22.1), the correlation with the genotype reaches significance for the complete Kruskal-Wallis, and the chi square test. As for *Citrobacter*, Otu000311 is not present in the RIIIS/J homozygotes, and more prevalent in the C57BL/6J homozygotes than the heterozygotes. There seem to be a trend in the prevalence for inflammation, with more inflamed mice carrying Otu000311 than healthy mice, but when looking at the interaction between *B4galnt2* and cecal inflammation, it appears different for C57BL/6J homozygotes than for the heterozygotes. At the RNA level (figure 22.2), Otu0003111 is truly an indicator species for *B4galnt2* genotype, with multiple tests reaching significance. The trend is the same as in the DNA: the prevalence is higher in C57BL/6J homozygotes than in heterozygotes, and no RIIIS/J homozygotes carry Otu000311, but there doesn't seem to be a difference in relative abundance between genotypes. The difference in prevalence between inflamed and healthy mice is stronger than in the DNA, but still doesn't reach significance. At the activity level (figure 22.3), Otu000311 is an indicator for *B4galnt2* genotype; however, it is obvious that only the prevalence plays a role, as the positive activity is the same for C57BL/6J homozygotes as for heterozygotes. *Alistipes* is a bacterial genus common to intestinal microbiota, and although not known for pathogenic behavior, its excess has been linked to colorectal cancer (*Borges-Canha, Portela-Cidade et al. 2015*). Its higher activity in healthy mice than in inflamed mice might suggest a probiotic behavior or susceptibility to inflammation in inflamed mice, but the sample size in healthy mice is too small to make confident inferences.

Figure 22.1

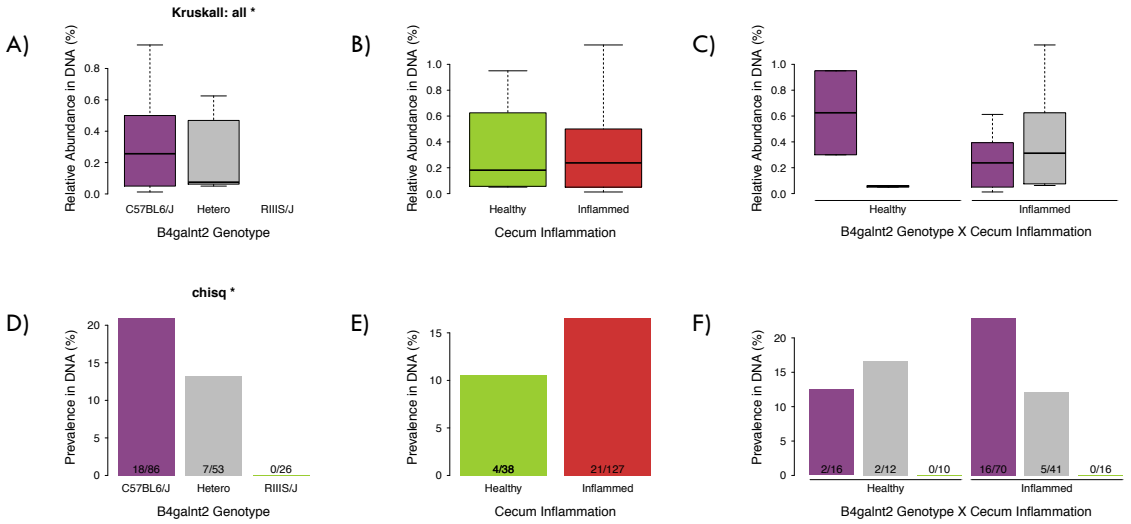

Figure 22.2

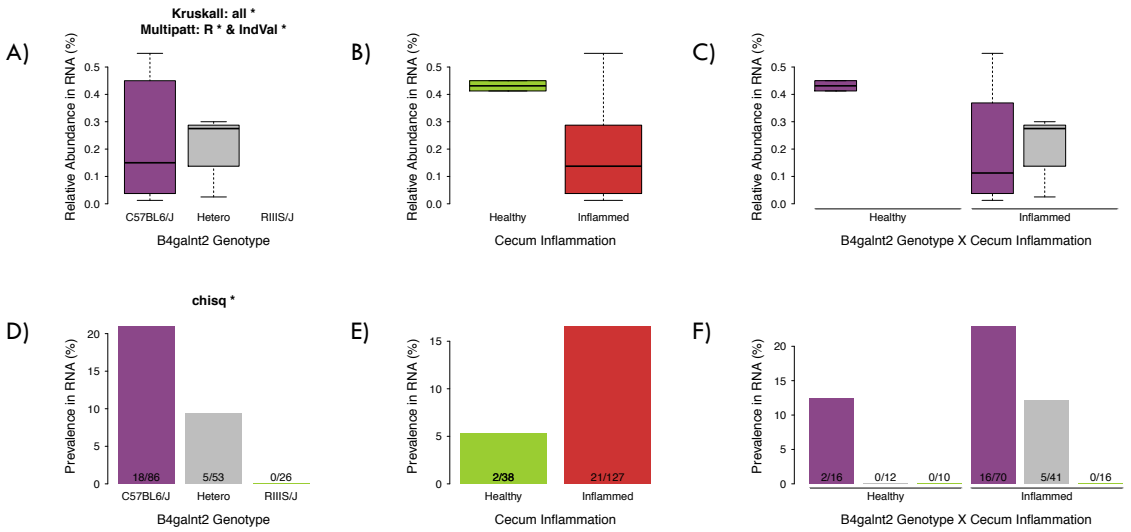

Figure 22.3

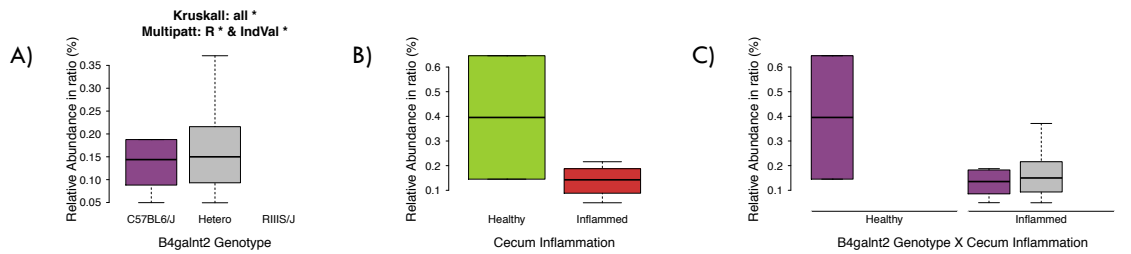

Figure 22: Non-zero relative abundance (ABC) and prevalence (DEF) of indicator species **Otu000311** (*Bacteroidetes; Bacteroidia; Bacteroidales; Rikenellaceae; **Alistipes***) in the DNA (21.1) and RNA (21.2) samples, and at the activity level (21.3), according to the *B4galnt2* genotype (AD), the cecum inflammation (BE), and their interaction (CF). The test performed were three indicator analysis functions: multipatt R, multipatt IndVal, and simper; the Kruskal-Wallis test on all values and only on non-zero values; and a chi square test on the prevalence. The results of the tests are reported in the title of the graphs with the following code: p<=0.001 ***, p<=0.01 **, p<=0.05 *. Outliers are not displayed.

The OTU belonging to the class ***Bacteroidales*** (Otu000322) at the DNA level (figure 23.1) is an indicator species for inflammation. Interestingly, the prevalence is much higher in highly inflamed mice compared to mice with no- or low inflammation, but the relative abundance goes in the opposite direction. There might be a trend with *B4galnt2* genotype, as the prevalence of Otu000322 is higher in C57BL/6J homozygotes than in heterozygotes and RIIIS/J homozygotes, but the sample size is too small for confident inference. At the RNA level (figure 23.2), the trend in prevalence with regard to cecal inflammation is the same as in the DNA, but as fewer samples carry Otu000322 at the RNA level, the difference no longer reaches significance. Interestingly, the relative abundance of this OTU in the RNA follows the opposite trend as in the DNA, as it is more abundant in highly inflamed mice compared to healthy and lowly inflamed mice, but again, the difference is too small to reach significance. At the activity level (figure 23.3), interestingly, the difference between inflammation levels is significant in the simper test, and the relative activity is higher in highly inflamed mice compared to healthy and inflamed mice. This might suggest pathogenic behavior, with higher prevalence and higher activity of the bacteria in highly inflamed mice. It is however difficult to go deeper into the understanding of this OTU, as it is only classified at the class level, and *Bacteroidales* is a broad taxa with numerous species and diverse behavior, making it difficult to infer whether this OTU could be considered pathogenic or not.

Figure 23.1

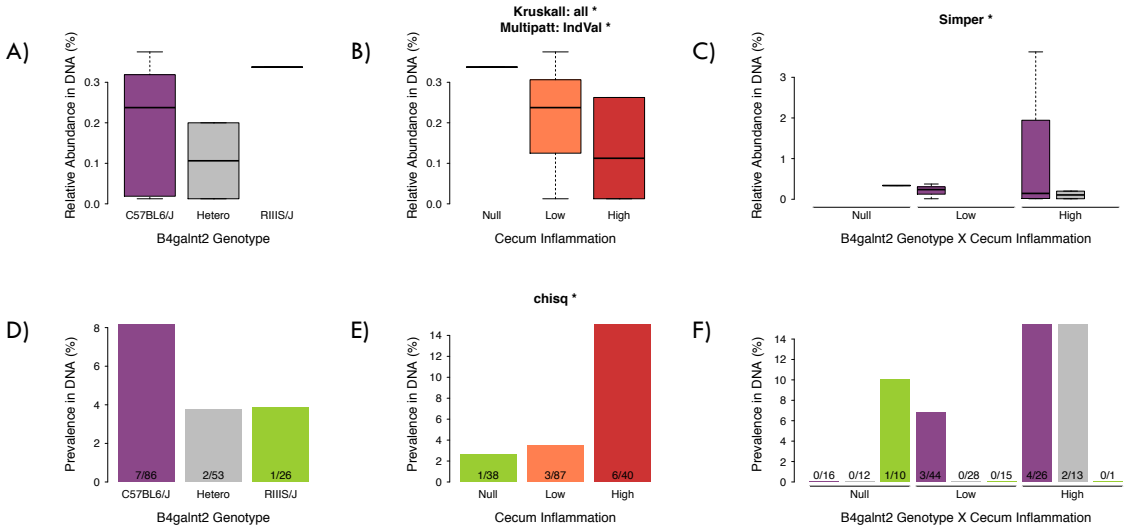

Figure 23.2

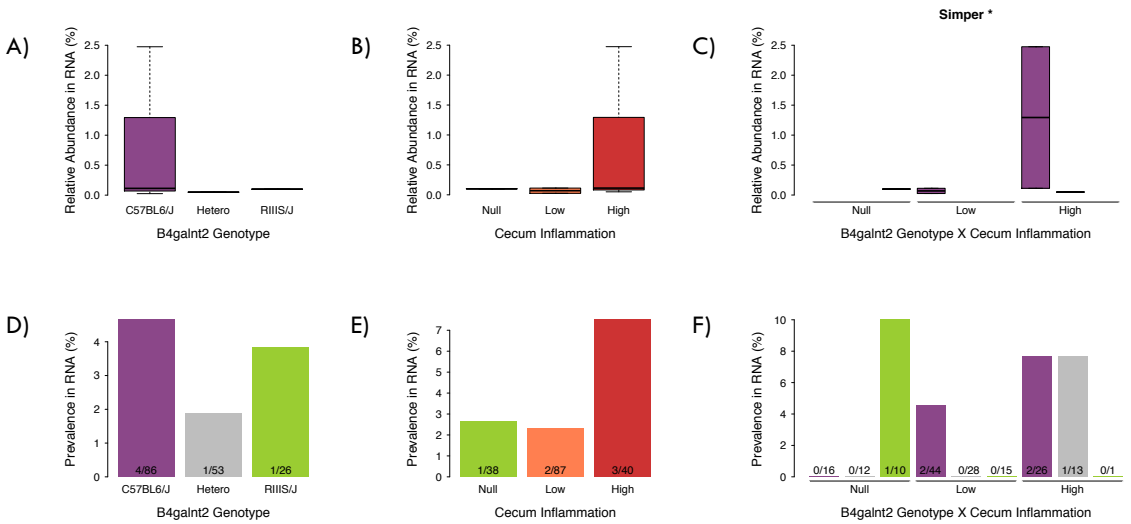

Figure 23.3

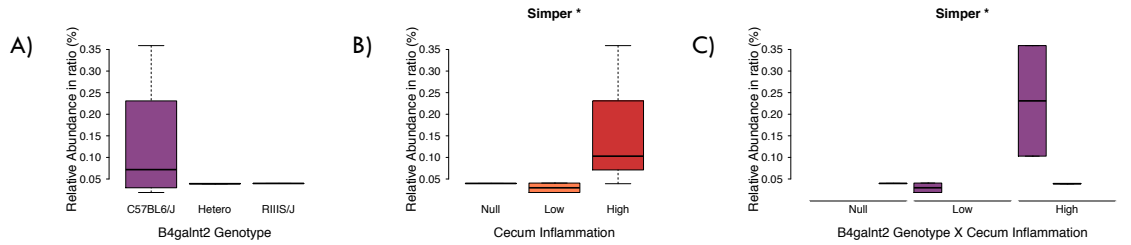

Figure 23: Non-zero relative abundance (ABC) and prevalence (DEF) of indicator species **Otu000322** (*Bacteroidetes; Bacteroidia; Bacteroidales; unclassified; unclassified*) in the DNA (21.1) and RNA (21.2) samples, and at the activity level (21.3), according to the *B4galnt2* genotype (AD), the cecum inflammation (BE), and their interaction (CF). The test performed were three indicator analysis functions: multipatt R, multipatt IndVal, and simper; the Kruskal-Wallis test on all values and only on non-zero values; and a chi square test on the prevalence. The results of the tests are reported in the title of the graphs with the following code: p<=0.001 ***, p<=0.01 **, p<=0.05 *. Outliers are not displayed.

The OTU belonging to the genus ***Morganella*** (Otu000463) does not show any significant correlations at the DNA level (figure 24.1), although a trend in prevalence can be seen with *B4galnt2* genotype and cecal inflammation, since only one RIIIS/J mouse carries Otu00463, and this mouse is healthy, while twelve C57BL/6J homozygotes an two heterozygotes mice carry *Morganella*, and they are all inflamed. Moreover, it seems that the prevalence is higher in highly inflamed mice compared to mildly inflamed mice. At the RNA level (figure 24.2), the trend is the same as in the DNA, although no correlations reach significance. It is at the activity level (figure 24.3) that *Morganella* is an indicator species for cecal inflammation. Moreover, the interaction between *B4galnt2* genotype and cecal inflammation also reaches significance in one test. For this OTU, the activity is much higher in highly inflamed mice than in mildly inflamed mice, and there seems to be a correlation with *B4galnt2* genotype as well, since no RIIIS/J homozygotes carry active *Morganella*, while all C57BL/6J and heterozygous mice that carry active *Morganella* are inflamed. *Morganella* is a common member of intestinal flora, but it is also a well-known opportunistic pathogen. The fact that it is more active in highly inflamed mice compared to mildly inflamed mice, and absent from healthy mice suggests that it could have a pathogenic behavior in the Espelette mice.

The OTU belonging to the genus ***Proteus*** (Otu000204) at the DNA level (figure 25.1) shows a significant correlation with the interaction between *B4galnt2* genotype and cecal inflammation, but does not reach the threshold for indicator species. Similarly to *Morganella*, *Proteus* is more prevalent in C57BL/6J homozygotes than in heterozygotes and absent from RIIIS/J homozygotes. Moreover, it is more prevalent in highly inflamed mice than in mildly inflamed ones, and absent from healthy mice. The relative abundance follows the same trend as the prevalence. At the RNA level (figure 25.2), the trend is the same as for the DNA, but Otu000204 is more prevalent in highly inflamed mice, leading cecal inflammation to significance for one MULTIPATT test. At the activity level (figure 25.3), Otu000204 is an indicator for cecal inflammation, with highly inflamed mice carrying a more active *Proteus* than lowly inflamed mice. Similar to *Morganella*, *Proteus* is a normal member of intestinal microbiota and a well-known opportunistic pathogen. Its increased activity in highly inflamed mice might thus indicate a pathogenic behavior.

Figure 24.1

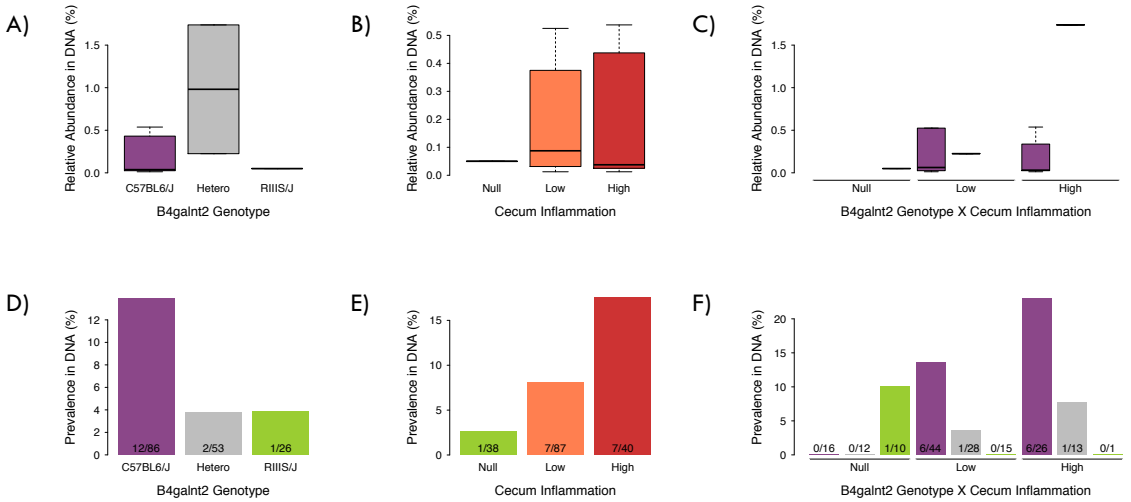

Figure 24.2

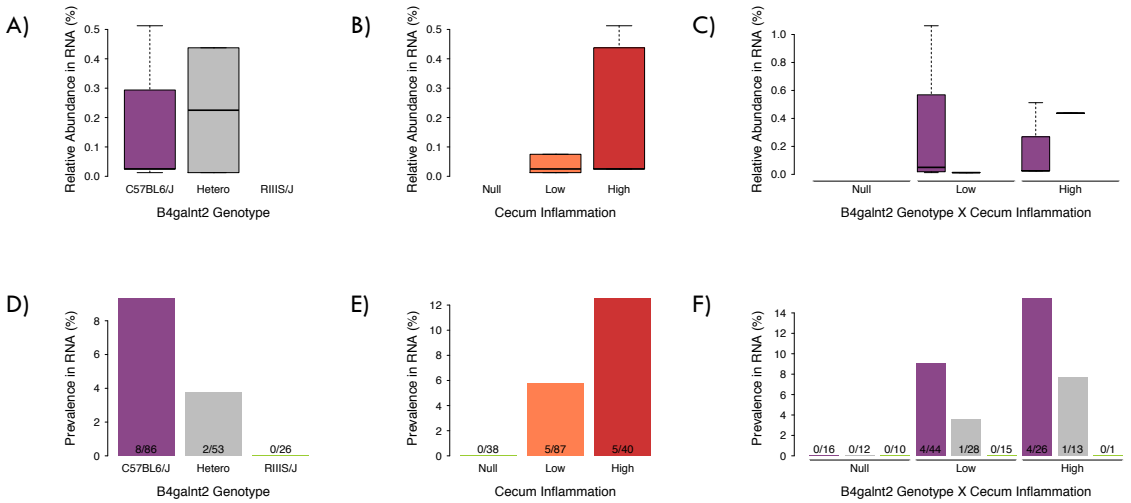

Figure 24.3

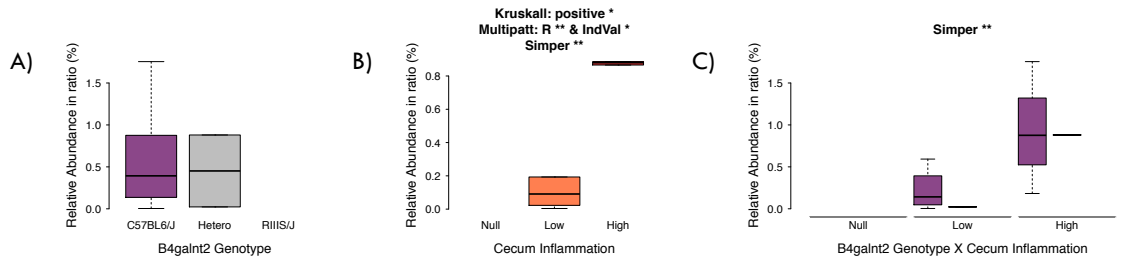

Figure 24: Non-zero relative abundance **(ABC)** and prevalence **(DEF)** of the indicator species **Otu000463** (*Proteobacteria; Gammaproteobacteria; Enterobacteriales; Enterobacteriaceae; **Morganella***) in the DNA **(21.1)** and RNA **(21.2)** samples, and at the activity level **(21.3)**, according to the *B4galnt2* genotype **(AD)**, the cecum inflammation **(BE)**, and their interaction **(CF)**. The test performed were three indicator analysis functions: multipatt R, multipatt IndVal, and simper; the Kruskal-Wallis test on all values and only on non-zero values; and a chi square test on the prevalence. The results of the tests are reported in the title of the graphs with the following code: p<=0.001 ***, p<=0.01 **, p<=0.05 *. Outliers are not displayed.

Figure 25.1

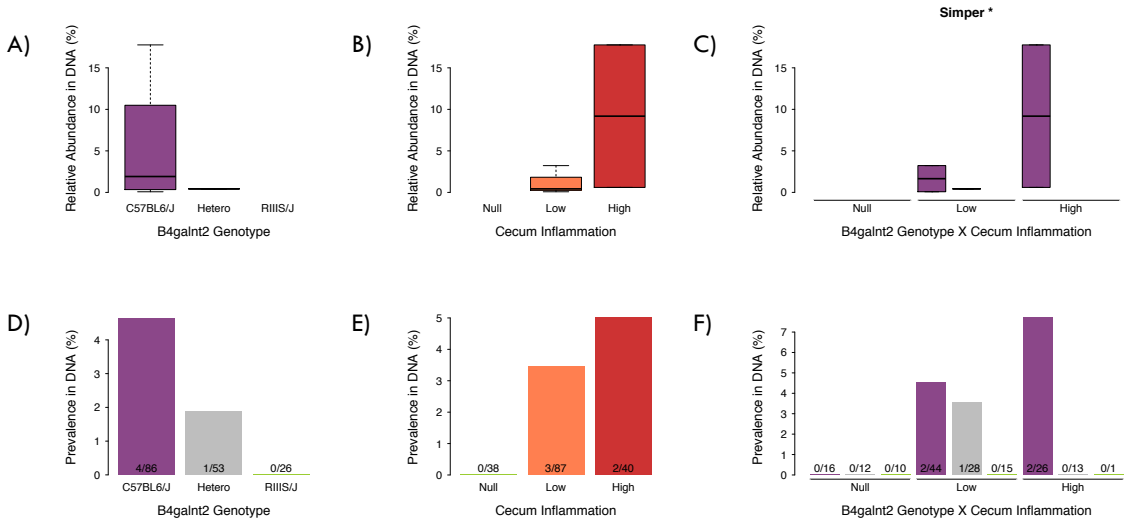

Figure 25.2

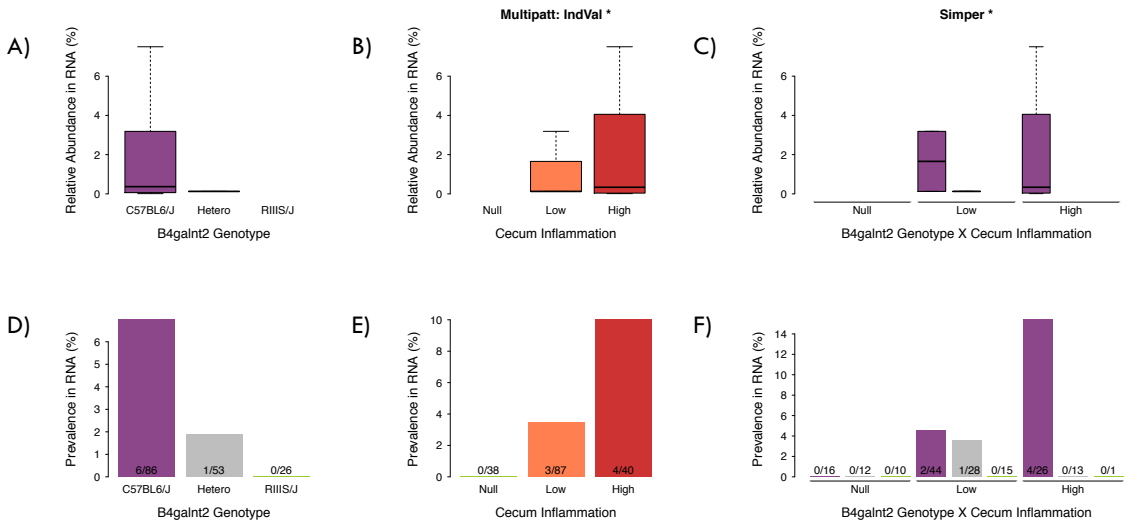

Figure 25.3

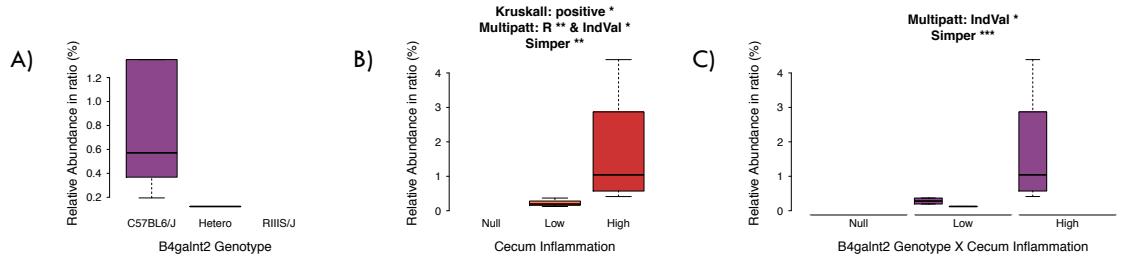

Figure 25: Non-zero relative abundance (ABC) and prevalence (DEF) of the indicator species **Otu000204** (*Proteobacteria; Gammaproteobacteria; Enterobacteriales; Enterobacteriaceae;* ***Proteus***) in the DNA (21.1) and RNA (21.2) samples, and at the activity level (21.3), according to the *B4galnt2* genotype (AD), the cecum inflammation (BE), and their interaction (CF). The test performed were three indicator analysis functions: multipatt R, multipatt IndVal, and simper; the Kruskal-Wallis test on all values and only on non-zero values; and a chi square test on the prevalence. The results of the tests are reported in the title of the graphs with the following code: p<=0.001 ***, p<=0.01 **, p<=0.05 *. Outliers are not displayed.

In **conclusion**, I found five indicators with regard to inflammation and/or *B4galnt2* genotype, including *Citrobacter, Alistipes* (Otu000311), unclassified *Bacteroidales* (Otu000322), *Morganella* (Otu000463) and *Proteus* (Otu000204). *Morganella* and *Proteus* appear to be especially promising as candidate pathogens, as they are indicator species at the activity level, suggesting a more functional role than the unclassified *Bacteroidales* OTU, which is an indicator only at the DNA level and represents a broader taxonomic group that makes it more difficult to characterize.

V. Colon microbiota

Although no significant correlation can be observed between *B4galnt2* genotype and inflammation in the distal colon, it is nonetheless interesting to analyze its microbiota with regards to *B4galnt2* genotype and inflammation level as a comparison point to the cecum analysis, as different parts of the intestine have diverse properties which represent various environmental conditions for the resident bacterial communities.

V.1 Helicobacter

While preparing the colon library for the 16S rRNA gene profiling, I was faced with a peculiar issue at the PCR level: a high proportion of samples showed a seemingly nonspecific band greater than the expected 400 bp product: one band at ~600 bp that I called "Mittel" and one at ~700 bp that I called "Oben". The concern was that these bands were quite prevalent, and for DNA samples outcompeted the correct 400 bp band. Out of the total 217 samples, 71 had an "Oben" band, 15 had a "Mittel" band, and 22 had both in the DNA samples. In the RNA, 105 showed an "Oben" band, 34 had a "Mittel" band and 8 had both. The overall prevalence of "Mittel" and "Oben" bands do not correlate with *B4galnt2* genotype.

To identify these unspecific bands, I gel purified them and sequenced them using classical Sanger sequencing. Unfortunately, only a fraction could be sequenced successfully. Interestingly, these products all belonged to the bacterial 16S rRNA gene, classified as *Helicobacter* via the online RDP classifier (*Cole, Wang et al. 2014*) and confirmed by NCBI Blast. The particularity of these sequences is that they have an insert of ~200 bp for the "Mittel" band and ~300 bp for the "Oben" band in the middle of the V1-V2 region of the 16SrRNA gene. From the literature, four *Helicobacter* species are known to have an insert in the same region: *H. macacae* and *H. bilis*

have an insert of 170 bp and 185 bp, respectively, while *H. mastomyrinus* and *H. fennelliae* have an insert of 294 bp and 358 bp, respectively.

The "Mittel" band represents three distinct sequences (figure 26, table 3), which group into two strains: "Mittel 2" clearly belongs to *H. bilis*, as there is 99.8% sequence identity (1 SNP) between both strains, while the two sequences of "Mittel 1" cannot be attributed to any know *Helicobacter* species, as the closest sequence is that of *H. bilis* which is already <82% sequence identity (>94 SNPs), way beyond any species binning threshold.

Table 3: Pairwise identity between the "Mittel" Helicobacter strains and the references H. bilis (Gene ID U18766) and H. macacae (Gene ID AF333338) at the 16S rRNA gene. The upper triangle contains the number of SNPs while the lower triangle contains the percentage identity. The diagonal contain the size of the insert for each strain.

	H.bilis	Mittel 2	Mittel 1.1	Mittel 1.2	H.macacae
H.bilis	185 bp	1	95	94	123
Mittel 2	99.8%	185 bp	95	94	124
Mittel 1.1	81.3%	81.3%	164 bp	1	127
Mittel 1.2	81.5%	81.5%	99.8%	164 bp	128
H.macacae	76.1%	75.9%	75.1%	74.9%	170 bp

The "Oben" band represents eight sequences (figure 27, table 4), which group into two strains. The six sequences of the "Oben 1" strain match very well to the *H. mastomyrinus* sequence, with a pairwise sequence identity between 99% and 99.8%. The two "Oben 9" sequences also match relatively well to the *H. mastomyrinus* sequence, with ~95% sequence identity, but they can be distinguished from the reference by a 20 bp deletion in the middle of the region, suggesting they are a different strain of the same species.

Table 4: Pairwise identity between the "Oben" Helicobacter strains and the reference H. mastomyrinus (Gene ID AY742307) at the 16S rRNA gene. The upper triangle contains the number of SNPs while the lower triangle contains the percentage identity. The diagonal contain the size of the insert for each strain.

	H.mastomyrinus	Oben 1.1	Oben 1.2	Oben 1.3	Oben 1.4	Oben 1.5	Oben 1.6	Oben 9.1	Oben 9.2
H.mastomyrinus	294 bp	1	2	2	2	3	5	22	26
Oben 1.1	99.8%	293 bp	1	1	1	2	4	21	25
Oben 1.2	99.6%	99.8%	293 bp	2	2	1	5	20	26
Oben 1.3	99.6%	99.8%	99.6%	293 bp	2	3	5	22	24
Oben 1.4	99.6%	99.8%	99.6%	99.6%	293 bp	3	5	22	26
Oben 1.5	99.4%	99.6%	99.8%	99.4%	99.4%	293 bp	6	21	27
Oben 1.6	99.0%	99.2%	99.0%	99.0%	99.0%	98.8%	293 bp	25	27
Oben 9.1	95.5%	95.7%	95.9%	95.5%	95.5%	95.7%	94.8%	273 bp	6
Oben 9.2	94.6%	94.8%	94.6%	95.0%	94.6%	94.4%	94.4%	98.7%	273 bp

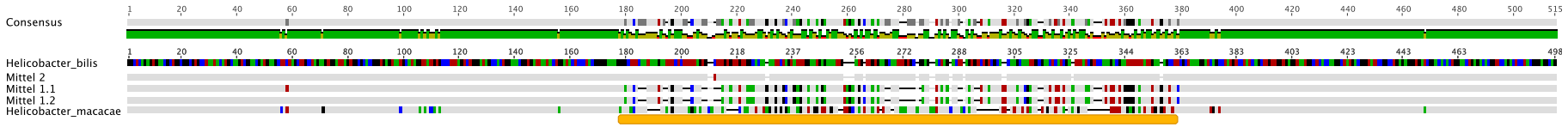

Figure 26: Alignment of the sequences belonging to the "Mittel" Helicobacter strains and the reference sequences of H. bilis (Gene ID U18766) and H. macacae (Gene ID AF333338). The sequence of H.bilis is used as reference in the alignment; bases in grey are identical to the reference while colored bases indicate SNPs. The yellow box shows the position of the insert relative to other Helicobacter species. The insert is 185bp long for H. bilis and Mittel 2, 164bp long for Mittel 1.1 and 1.2 and 170bp long for H. macacae.

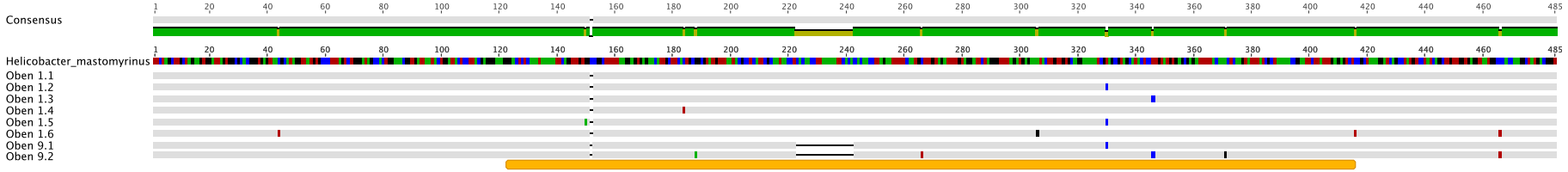

Figure 27: Alignment of the sequences belonging to the "Oben" Helicobacter strains and the reference sequences of H. mastomyrinus (Gene ID AY742307). The sequence of H. mastomyrinus is used as reference in the alignment; bases in grey are identical to the reference while colored bases indicate SNPs. The yellow box shows the position of the insert relative to other Helicobacter species. The insert is 294 bp long for H. mastomyrinus, 293 bp long for the Oben 1 sequences and 273 bp long for the Oben 9 sequences.

It is important to note that no other *Helicobacter* species were detected by the analysis of the expected 400 bp bands.

Considering the sequenced PCR products, only the "Mittel 1" strain from DNA samples was prevalent enough for me to test for correlations with *B4galnt2* genotype and colon inflammation. Interestingly, this strain is more prevalent in RIIIS/J homozygotes compared to C57BL/6J homozygotes (figure 28), with the heterozygotes at an intermediate level. Moreover, the majority of the RIIIS/J homozygotes carrying this *Helicobacter* strain are healthy, while all C57BL/6J homozygotes carrying "Mittel 1" are inflamed. To have a different perspective on the data, I expressed them in different way, looking at the level and prevalence of inflammation according to the presence/absence of "Mittel 1" and *B4galnt2* genotype (figure 29). There does not appear to be any influence of this *Helicobacter* strain on the inflammation level, but the inflammation prevalence seems higher in C57BL/6J homozygotes with *Helicobacter* compared to C57BL/6J homozygotes without "Mittel 1", while it seems to be the opposite trend in RIIIS/J homozygotes, although no results are significant due to small sample sizes.

In conclusion, it is quite intriguing to discover that for about half of the samples, the colon microbiota when viewed at the DNA level is dominated by one or more strains of *Helicobacter*, a well-characterized pathogenic genus. Moreover, it seems that one of these strains, "Mittel 1", might represent a good candidate, as it seems to be linked to inflammation, specifically in C57BL/6J homozygotes. However, due to the low success rate of sequencing, the results need to be interpreted with caution and would need further characterization to be able to assess the true correlation between *B4galnt2* genotype, colon inflammation and the prevalence of *Helicobacter* species.

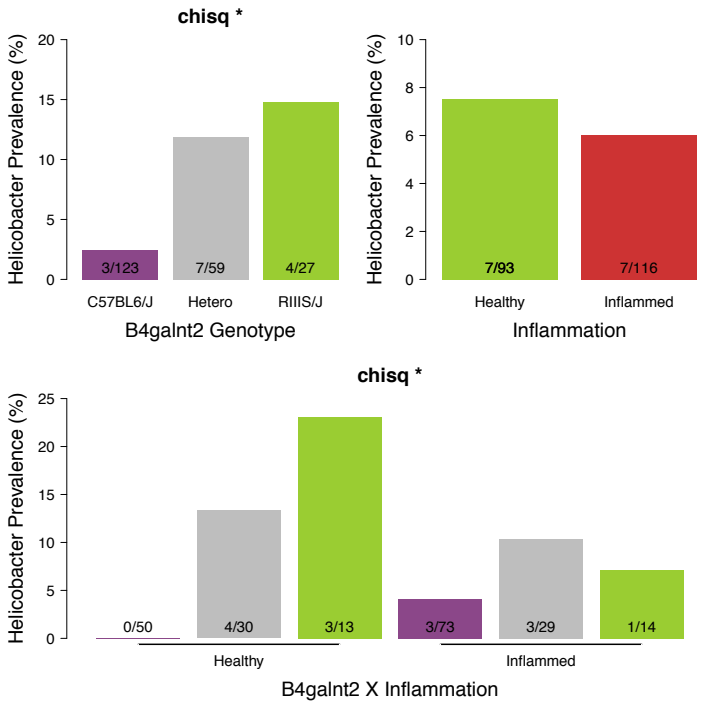

Figure 28: Prevalence of the Helicobacter strain "Mittel 1" in the colon at the DNA level, according to *B4galnt2* genotype, colon inflammation and their interaction. The association was tested by chi square test, the significance is reported in the title as follow: p<0.001 ***, p<0.01 **, p<0.05 *.

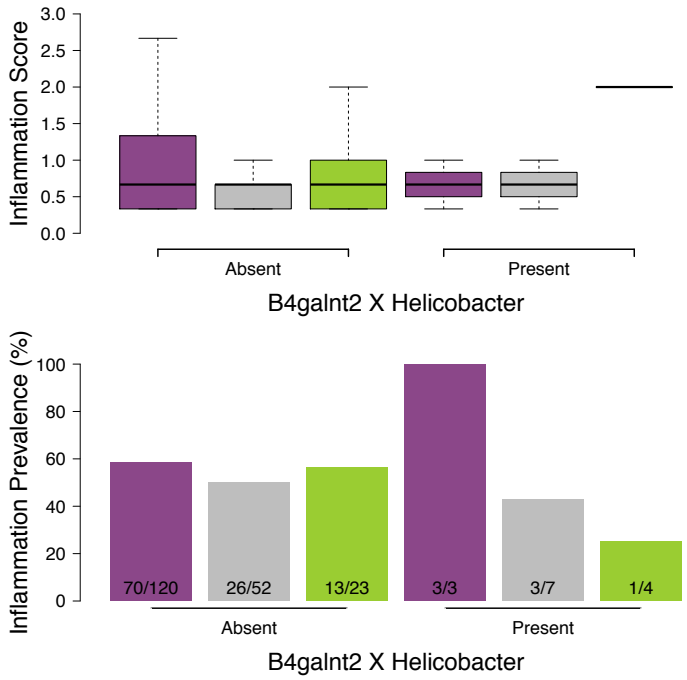

Figure 29: Non-zero inflammation score (top) and inflammation prevalence (bottom) in the colon at the DNA level, depending on the presence of the Helicobacter strain "Mittel 1" (Absent/Present), and according to *B4galnt2* genotype (C57BL/6J in purple, heterozygotes in grey and RIIIS/J in green). The association was tested by chi square test, the significance is reported in the title as follow: p<0.001 ***, p<0.01 **, p<0.05 *. The association was tested with Kruskal-Wallis test for the inflammation score and chi square test for the prevalence, no test reached significance.

V.2 Experimental variables

For the 16SrRNA library, as for the cecum, I first verified the influence of experimental variables on the composition of the colonic microbiota. Similar to the cecal microbiota, all experimental variables show significant and non-negligible influence on the microbial community, both at the DNA (figure S19) and the RNA (figure S20) level. Moreover, the fact that I had to gel purify a large amount of DNA samples due to the presence of strong *Helicobacter* bands seems to have a strong effect on the microbiota (figure S21).

When correcting for the farm influence, at the DNA level (figure S22), surprisingly, the MiSeq library (both for the Bray-Curtis and Jaccard dissimilarities) and the date of capture of the mice (only Jaccard) remain significant. At the RNA level (figure S23), the library is no longer significant, but the date of capture (only Jaccard) remains associated to the intestinal microbiota, as well as the date of dissection (only Jaccard). These influences appear however weak, as there is no clear separation of the groups along the significant axes. Whether the DNA samples were gel purified or not remains however significant (figure S24), and a clear separation of the samples that were extracted from the agarose gel (blue & green) from those that were directly used (red) is visible. This effect is however unlikely to be due to the gel purification itself, but is more likely to reflect the difference between *Helicobacter*-dominated communities and *Helicobacter*-poor communities. Indeed, technically speaking, the gel purification can influence slightly the Jaccard diversity measure, since extractions are always the source of minimal material loss, which impact rare species more strongly than abundant ones, leading to a reduced number of detected species, but the Bray-Curtis diversity measure should not be influenced, as it relies more on abundant species. On the other hand, given the various microbial interactions that takes place is such communities, it is not surprising that a community dominated by only one species (here *Helicobacter*) would be very different than communities were this bacteria is not present.

Obviously, as in the cecum, the farm has a strong influence on the microbiota (figure S25). Although the peculiarity of the MT21 farm is not as strong in the colon as it was in the cecum (dark pink), I chose to remove it from the analysis to be consistent between both intestinal parts (figure S26). Notably, while the variance explained by the farms in the cecum decreased when removing the MT21 farm, it increases in the colon, except for the Bray-Curtis dissimilarity at the DNA level.

Finally, I tested whether intrinsic variables influence the colon microbiota. At the DNA level (figure S27), the genetic population, mitochondrial D-loop haplotype and haplogroup and the gender of the mice have no significant association with the microbiota. These parameters also

display no association to the microbiota at the RNA (figure S28) and activity (figure S29) levels. The mice parameters, weight, length and BMI are also not significantly associated with the microbiota at the DNA level (figure S30), but weight and BMI are significant at the RNA level (figure S31). This is however limited to Jaccard dissimilarities, and the effect is limited to ~1.5% of the variation. At the activity level (figure S32), only the correlation of the Jaccard dissimilarity to the BMI remains significant.

In conclusion, the factor influencing the microbiota the most is the farm of origin, as observed for the cecum. In the colon, the mitochondrial D-loop does not appear to have an influence like it had in the cecum, but the weight and the BMI do have a limited influence in the colon, at the RNA and activity level, only for prevalence data (Jaccard dissimilarity). The main difference between the colonic and the cecal microbiota is the high prevalence of *Helicobacter*-dominated microbiota in the colon and at the DNA level, which seem to have a small (~3%) but highly significant influence on the remaining bacteria.

V.3 *B4galnt2* & Inflammation

As for the cecal microbiota, the main goal of the analysis is to test for an influence of *B4galnt2* genotype and inflammation. Similar to the cecal microbiota, neither *B4galnt2* genotype (figure S33), inflammation studied as categorical (figure S34)- or quantitative variable (figure S35), nor their interaction (figure S36) have a significant influence on the colon microbiota, suggesting again that the role of *B4galnt2* might be more limited, influencing only a few bacteria rather than the entire flora.

V.4 Indicator species analysis

Using the same collection of tests as in the cecum to detect indicator species for the colonic microbiota, I could identify one indicator genus and seven indicator OTUs. The indicator genus is *Citrobacter*, and the indicator OTUs belong to unclassified *Porphyromonadaceae*, unclassified *Lachnospiraceae*, unclassified *Ruminococcaceae*, *Coprobacillus*, and *Morganella*. As for the cecum analysis, I verified *a posteriori* that the association was not confounded by other factors (tables S6-S14).

For ***Citrobacter*** (figure 30.1) at the DNA level, the correlation with *B4galn2* and the interaction between *B4galnt2* and inflammation is significant only for one test, but we can see the same trend as for the cecum: *Citrobacter* is more prevalent in the C57BL/6J homozygotes than in the heterozygotes, and it is absent from the RIIIS/J homozygotes. For inflammation, however, it seems more prevalent in the inflamed C57BL/6J homozygotes than in the healthy C57BL/6J homozygotes, although the relative abundance displays the opposite pattern. At the RNA level (figure 30.2), *Citrobacter* is an indicator species for *B4galnt2* genotype, as the prevalence is higher in the C57BL/6J homozygotes than in the heterozygotes, and there is only one occurrence in the RIIIS/J homozygotes. For inflammation, however, the trend seems to be reversed, as the prevalence in healthy mice is similar to that of inflamed mice, and only the genotype seems to influence it, though the relative abundance seems slightly higher in inflamed mice compared to healthy mice. At the activity level (figure 30.3), the correlation with *B4galnt2* genotype is significant in the complete Kruskal-Wallis test, which is probably due to the difference in prevalence in the RNA.

The OTU belonging to **unclassified *Porphyromonadaceae*** (Otu000089) is an indicator species for inflammation and the interaction between inflammation and *B4galnt2* genotype at the DNA level (figure 31.1). Indeed, it is more prevalent and more abundant in highly inflamed mice compared to mildly inflamed or healthy mice. This trend is especially true for the C57BL/6J homozygotes, as all heterozygotes carrying Otu000089 are healthy, and it is absent from the RIIIS/J homozygotes. At the RNA level (figure 31.2), the OTU is still significant for two tests, for the inflammation and the interaction between inflammation and *B4galnt2* genotype, and the trend is completely similar to that of the DNA, although there is now one occurrence in a healthy RIIIS/J mouse. At the activity level (figure 31.3), the trend is different, as it seems now that the inflamed heterozygotes have a higher abundance than all other types. This might suggest, if one considers this OTU as potential candidate pathogen, that heterozygote mice (and potentially RIIIS/J homozygotes) requires a higher abundance of the infectious bacteria to be inflamed compared to C57BL/6J homozygotes, pointing to a tolerance mechanism rather than resistance.

Figure 30.1

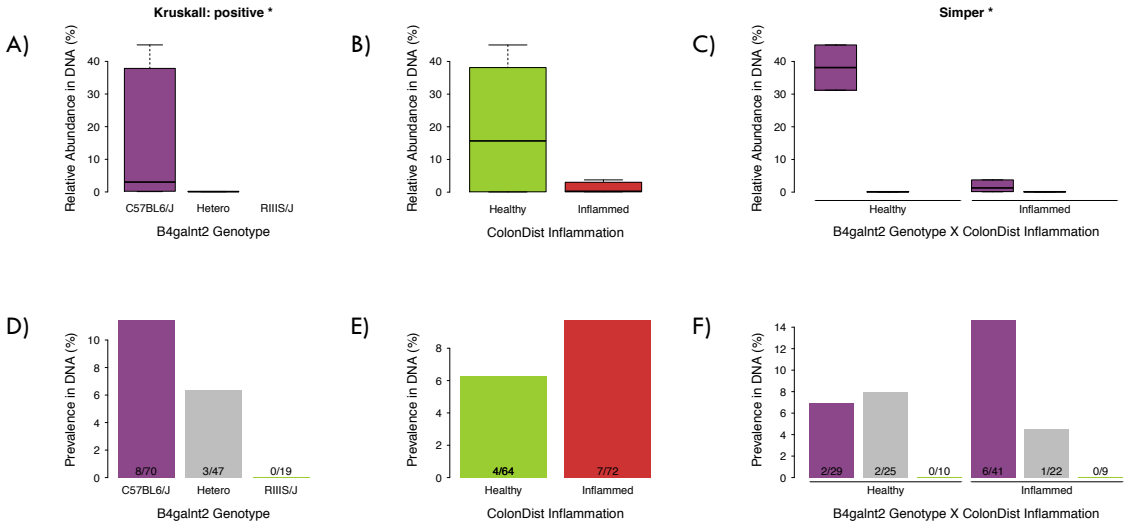

Figure 30.2

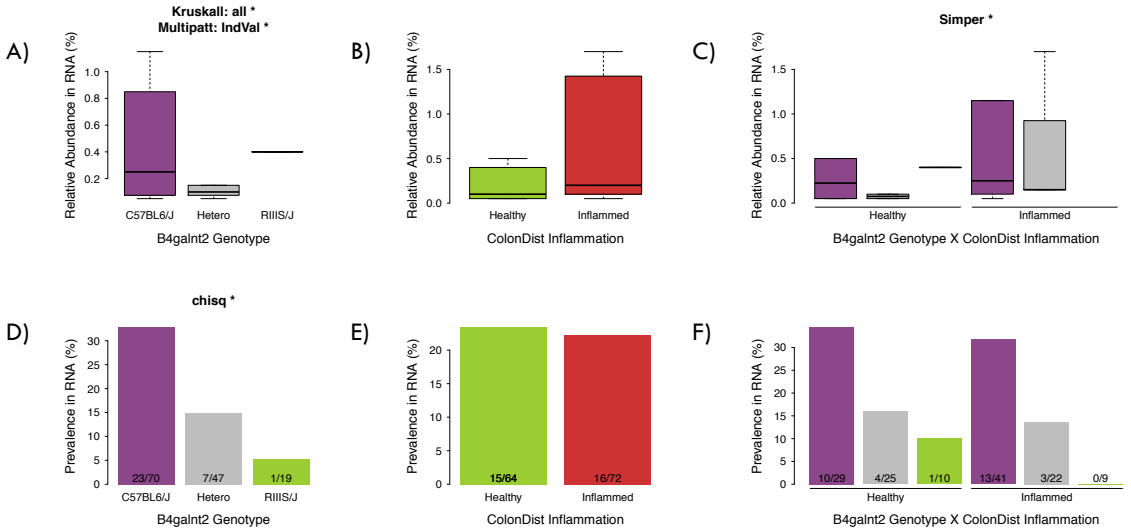

Figure 30.3

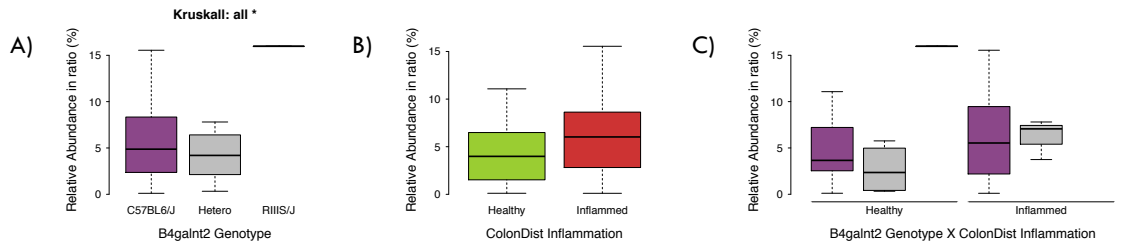

Figure 30: Non-zero relative abundance **(ABC)** and prevalence **(DEF)** of the indicator genus *Citrobacter* (*Proteobacteria; Gammaproteobacteria; Enterobacteriales; Enterobacteriaceae*) in the DNA **(21.1)** and RNA **(21.2)** samples, and at the activity level **(21.3)**, according to the *B4galnt2* genotype **(AD)**, the cecum inflammation **(BE)**, and their interaction **(CF)**. The test performed were three indicator analysis functions: multipatt R, multipatt IndVal, and simper; the Kruskal-Wallis test on all values and only on non-zero values; and a chi square test on the prevalence. The results of the tests are reported in the title of the graphs with the following code: p<=0.001 ***, p<=0.01 **, p<=0.05 *. Outliers are not displayed.

Figure 31.1

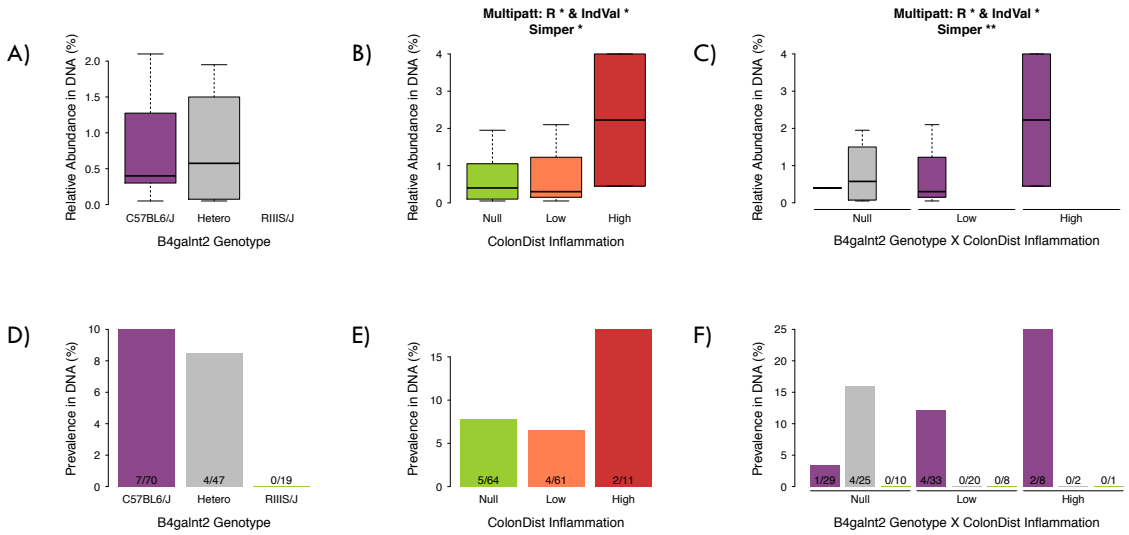

Figure 31.2

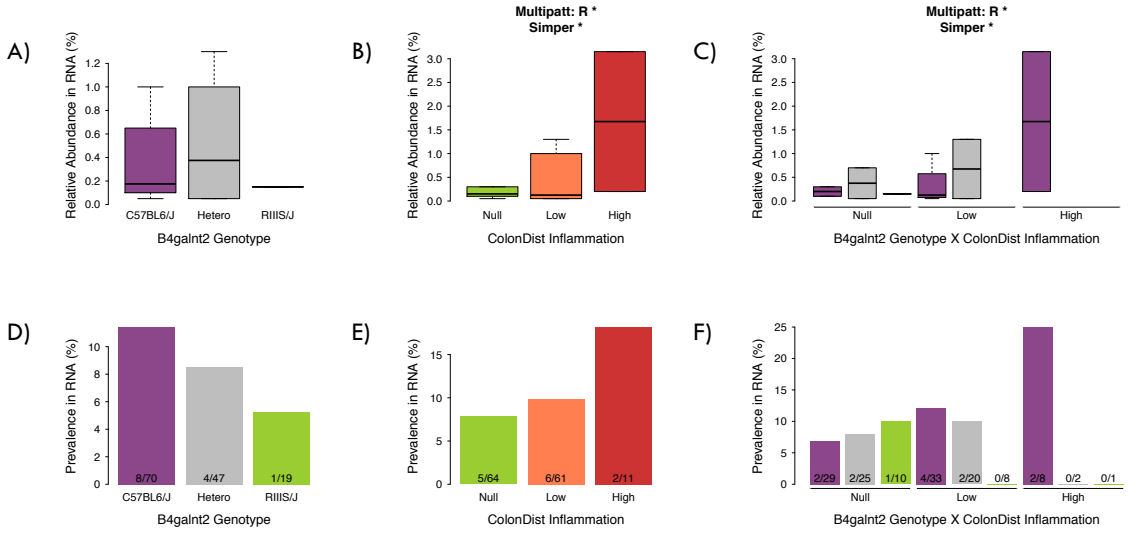

Figure 31.3

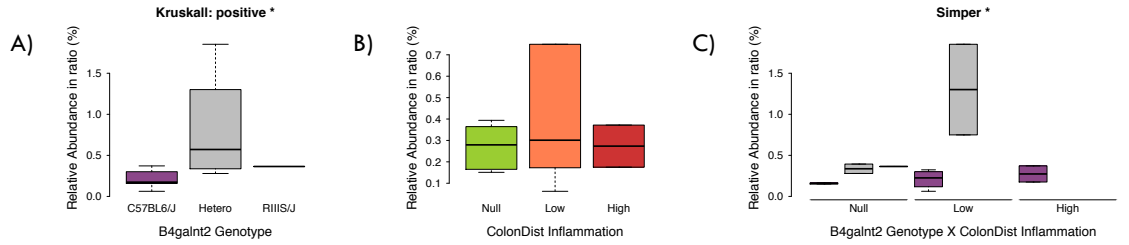

Figure 31: Non-zero relative abundance (ABC) and prevalence (DEF) of the indicator species **Otu000089** (*Bacteroidetes; Bacteroidia; Bacteroidales; **Porphyromonadaceae**; unclassified*) in the DNA (21.1) and RNA (21.2) samples, and at the activity level (21.3), according to the *B4galnt2* genotype (AD), the cecum inflammation (BE), and their interaction (CF). The test performed were three indicator analysis functions: multipatt R, multipatt IndVal, and simper; the Kruskal-Wallis test on all values and only on non-zero values; and a chi square test on the prevalence. The results of the tests are reported in the title of the graphs with the following code: p<=0.001 ***, p<=0.01 **, p<=0.05 *. Outliers are not displayed.

For the first OTU belonging to **unclassified *Lachnospiraceae*** (Otu000198), at the DNA level (figure 32.1), it seems that the prevalence is higher in C57BL/6J compared to heterozygotes and RIIIS/J homozygotes. Moreover, the majority of the C57BL/6J carrying this OTU are inflamed, while the RIIIS/J individual carrying it is healthy. At the RNA level (figure 32.2), Otu000198 is an indicator for the interaction between *B4galnt2* and inflammation, as all RIIIS/J individuals carrying the bacteria are healthy, while all C57BL/6J homozygotes carrying it are inflamed, and the single occurrence in a heterozygote is in a healthy one. At the activity level (figure 32.3), the OTU is an indicator for *B4galnt2* genotype and its interaction with inflammation. This is however likely due only to the difference in prevalence, since the relative activities are in the same range.

The second OTU belonging to **unclassified *Lachnospiraceae*** (Otu000521), is present only at the DNA level (figure 33), and is an indicator for inflammation and its interaction with *B4galnt2* genotype. This OTU is more prevalent and more abundant in the inflamed C57BL/6J compared to the mildly inflamed C57BL/6J homozygotes and the heterozygotes. It is moreover absent from the RIIIS/J homozygotes.

The last OTU belonging to **unclassified *Lachnospiraceae*** (Otu000293) is not significantly associated to *B4galnt2* genotype, inflammation or their interaction at the DNA level (figure 34.1). The trend seems to be that this OTU is more prevalent in RIIIS/J homozygotes and in healthy or mildly inflamed mice. It might however be more abundant in C57BL/6J homozygotes. At the RNA level however (figure 34.2), the trend is very different, as it is more prevalent in highly inflamed C57BL/6J homozygotes, and this pattern is significant at the activity level (figure 34.3).

The OTU belonging to **unclassified *Ruminococcaceae*** (Otu000408) is present only at the RNA level (figure 35.1), where it is indicator for inflammation and its interaction with *B4galnt2* genotype. This bacterium is both more prevalent and more abundant in the highly inflamed C57BL/6J homozygotes, while the two RIIIS/J homozygotes and two heterozygotes carrying it are healthy. These results are also true at the activity level (figure 35.2), and perhaps even stronger.

The OTU belonging to ***Coprobacillus*** (Otu000276) is highly associated to inflammation at the DNA level (figure 36.1), as it is only present in inflamed mice, but there doesn't seem to be an effect of genotype. The trend is the same at the RNA (figure 36.2) and activity (figure 36.3) level, but the sample size is smaller and does not reach significance.

Figure 32.1

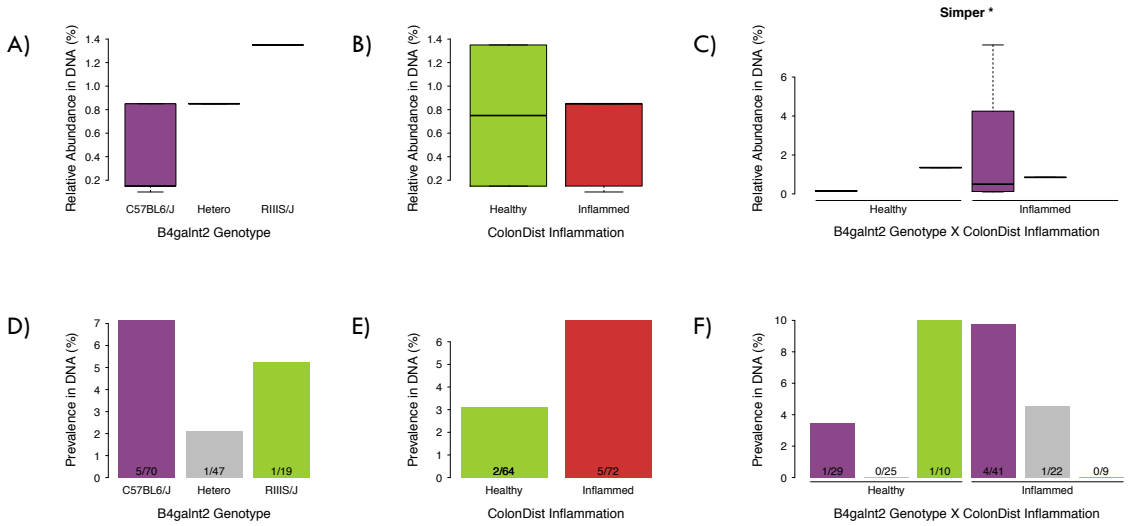

Figure 32.2

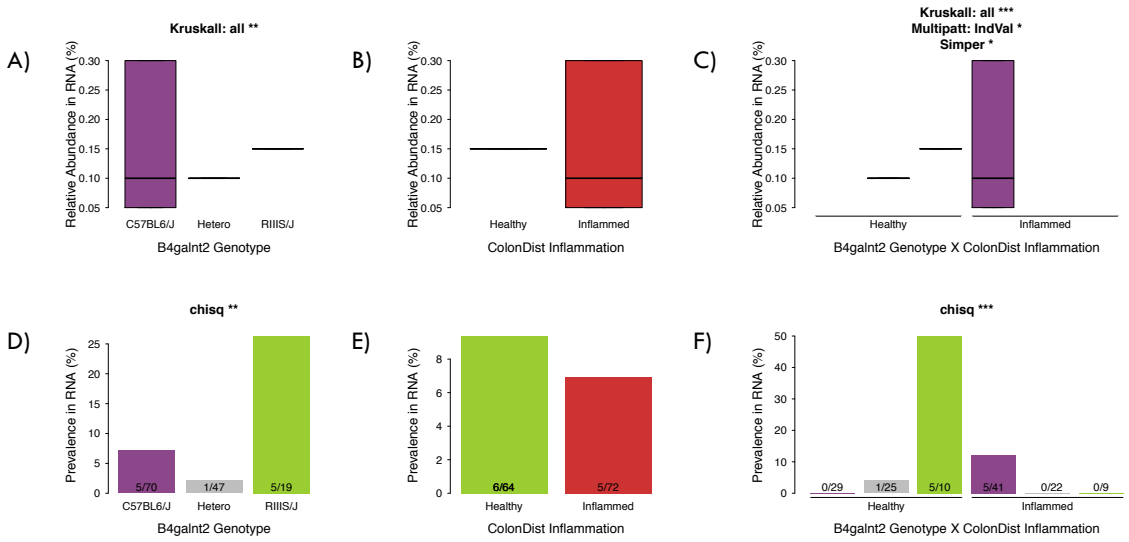

Figure 32.3

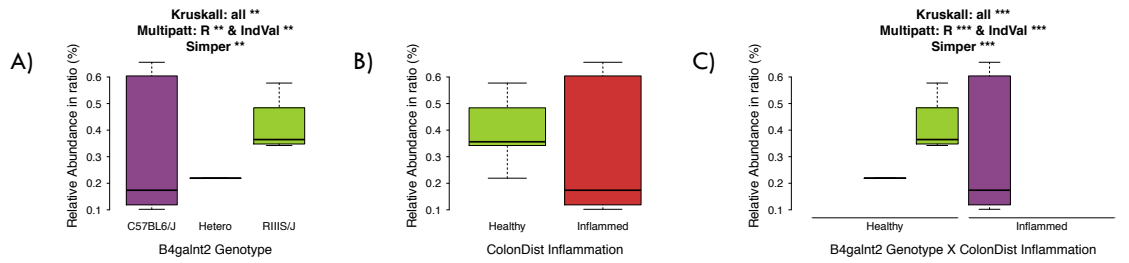

Figure 32: Non-zero relative abundance (ABC) and prevalence (DEF) of the indicator species **Otu000198** (*Firmicutes; Clostridia; Clostridiales; Lachnospiraceae; unclassified*) in the DNA (21.1) and RNA (21.2) samples, and at the activity level (21.3), according to the *B4galnt2* genotype (AD), the cecum inflammation (BE), and their interaction (CF). The test performed were three indicator analysis functions: multipatt R, multipatt IndVal, and simper; the Kruskal-Wallis test on all values and only on non-zero values; and a chi square test on the prevalence. The results of the tests are reported in the title of the graphs with the following code: p<=0.001 ***, p<=0.01 **, p<=0.05 *. Outliers are not displayed.

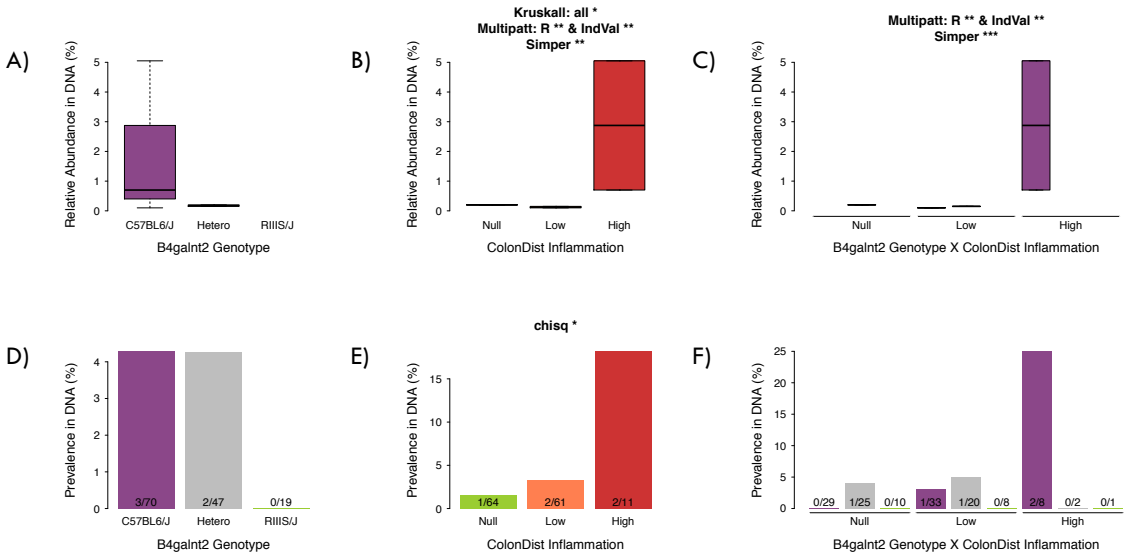

Figure 33: Non-zero relative abundance **(ABC)** and prevalence **(DEF)** of the indicator species **Otu000521** (*Firmicutes; Clostridia; Clostridiales; **Lachnospiraceae**; unclassified*) in the DNA samples, according to the *B4galnt2* genotype **(AD),** the cecum inflammation **(BE)**, and their interaction **(CF)**. The test performed were three indicator analysis functions: multipatt R, multipatt IndVal, and simper; the Kruskal-Wallis test on all values and only on non-zero values; and a chi square test on the prevalence. The results of the tests are reported in the title of the graphs with the following code: p<=0.001 ***, p<=0.01 **, p<=0.05 *. Outliers are not displayed.

Figure 34.1

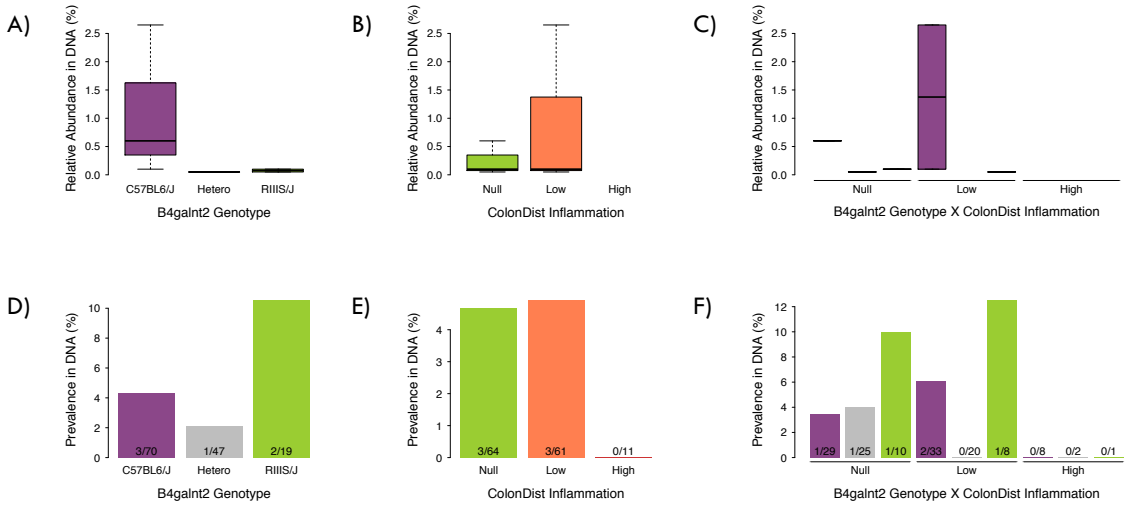

Figure 34.2

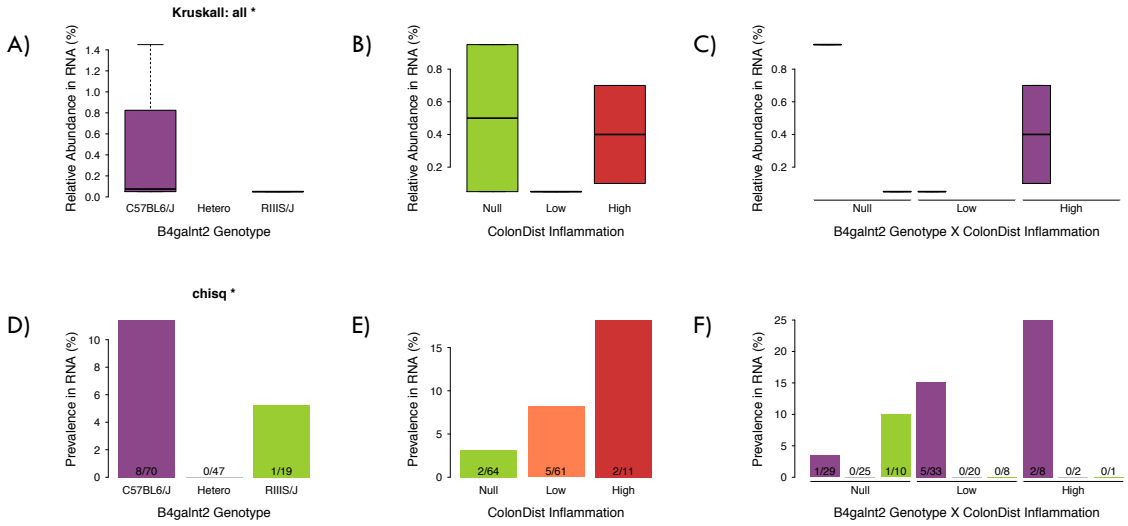

Figure 34.3

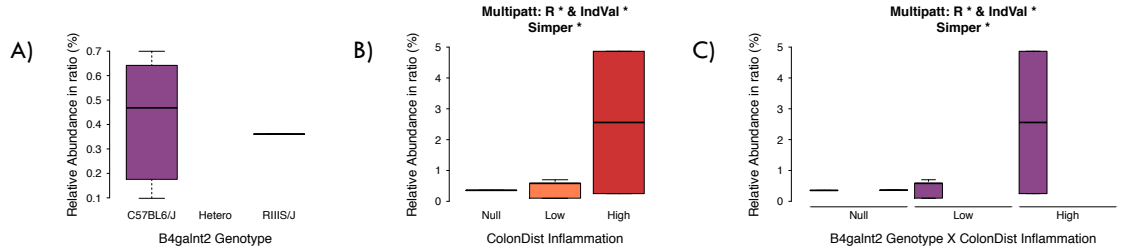

Figure 34: Non-zero relative abundance (ABC) and prevalence (DEF) of the indicator species **Otu000293** (*Firmicutes; Clostridia; Clostridiales; Lachnospiraceae; unclassified*) in the DNA (21.1) and RNA (21.2) samples, and at the activity level (21.3), according to the *B4galnt2* genotype (AD), the cecum inflammation (BE), and their interaction (CF). The test performed were three indicator analysis functions: multipatt R, multipatt IndVal, and simper; the Kruskal-Wallis test on all values and only on non-zero values; and a chi square test on the prevalence. The results of the tests are reported in the title of the graphs with the following code: p<=0.001 ***, p<=0.01 **, p<=0.05 *. Outliers are not displayed.

Figure 35.1

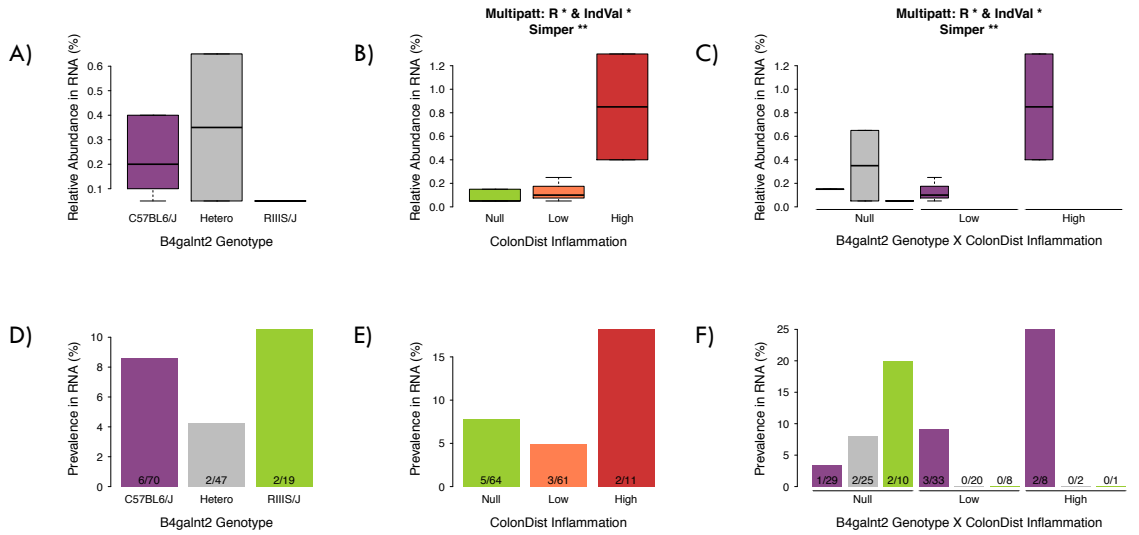

Figure 35.2

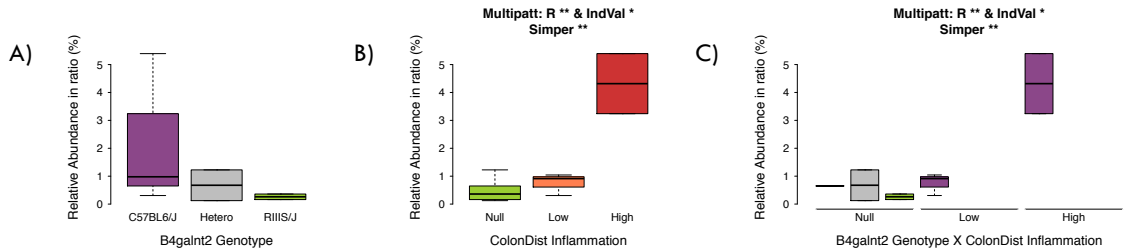

Figure 35: Non-zero relative abundance **(ABC)** and prevalence **(DEF)** of the indicator species **Otu000408** (*Firmicutes; Clostridia; Clostridiales; **Ruminococcaceae**; unclassified*) in the RNA **(21.1)** samples, and at the activity level **(21.2)**, according to the *B4galnt2* genotype **(AD)**, the cecum inflammation **(BE)**, and their interaction **(CF)**. The test performed were three indicator analysis functions: multipatt R, multipatt IndVal, and simper; the Kruskal-Wallis test on all values and only on non-zero values; and a chi square test on the prevalence. The results of the tests are reported in the title of the graphs with the following code: p<=0.001 ***, p<=0.01 **, p<=0.05 *. Outliers are not displayed.

Figure 36.1

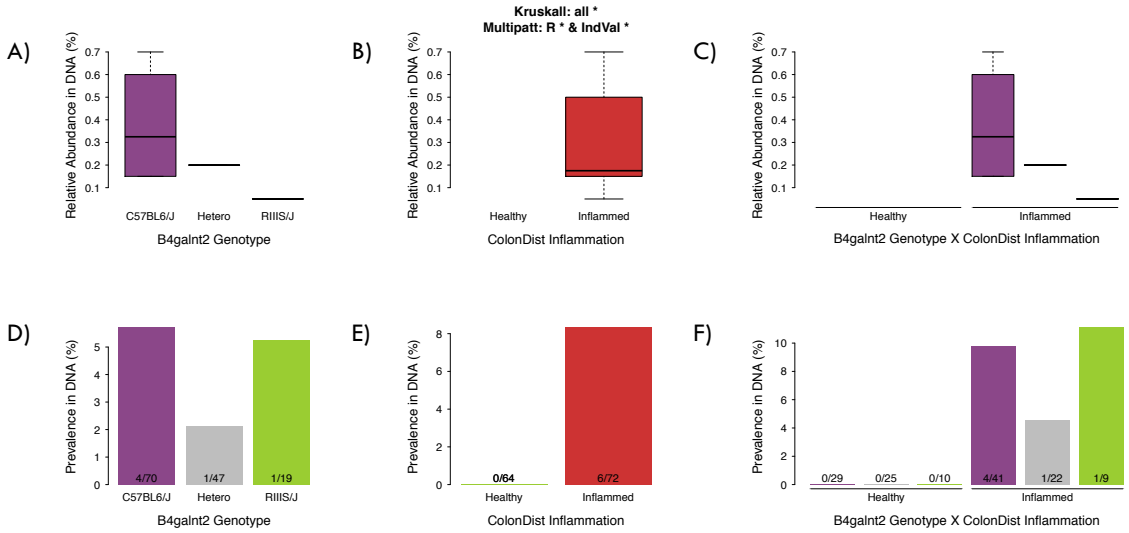

Figure 36.2

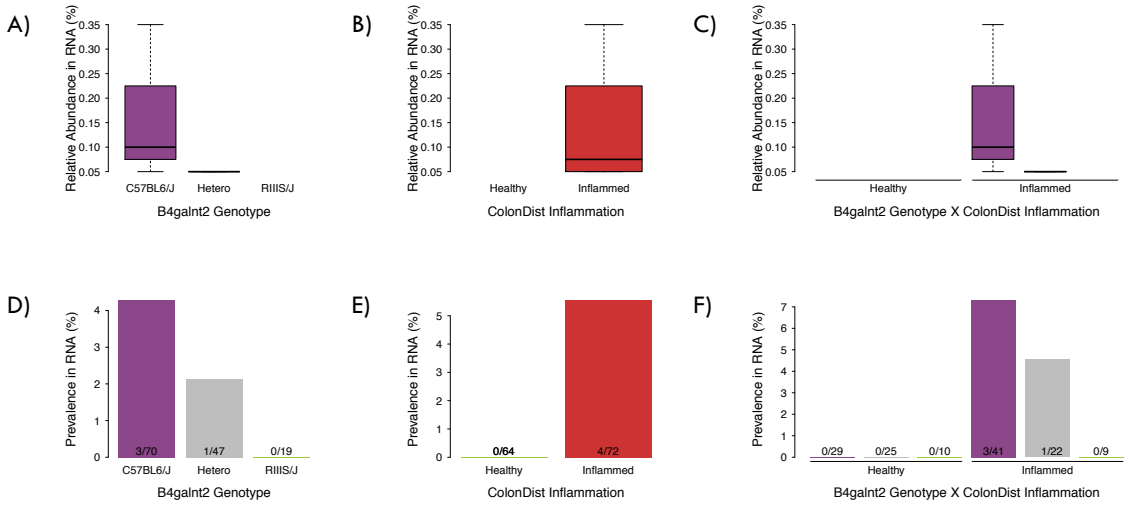

Figure 36.3

Figure 36: Non-zero relative abundance **(ABC)** and prevalence **(DEF)** of the indicator species **Otu000276** (*Firmicutes; Erysipelotrichia; Erysipelotrichales; Erysipelotrichaceae; Coprobacillus*) in the DNA **(21.1)** and RNA **(21.2)** samples, and at the activity level **(21.3)**, according to the *B4galnt2* genotype **(AD)**, the cecum inflammation **(BE)**, and their interaction **(CF)**. The test performed were three indicator analysis functions: multipatt R, multipatt IndVal, and simper; the Kruskal-Wallis test on all values and only on non-zero values; and a chi square test on the prevalence. The results of the tests are reported in the title of the graphs with the following code: p<=0.001 ***, p<=0.01 **, p<=0.05 *. Outliers are not displayed.

Interestingly, the OTU belonging to ***Morganella*** is the same identified as indicator species in the cecum (Otu000463), but its behavior in the colon differs to that in the cecum. Indeed, at the DNA level (figure 37.1), *Morganella* is more prevalent in the highly inflamed RIIIS/J homozygotes, although it seems to be less abundant in RIIIS/J homozygotes compared to C57BL/6J homozygotes. At the RNA level (figure 37.2), the trend is perfectly similar, but it is only at the activity level (figure 37.3) that it is an indicator species for inflammation, with higher prevalence and higher activity in highly inflamed mice compared to mildly inflamed and healthy mice. In the colon, *Morganella* is thus still an indicator for inflammation, but it does not follow the same relationship to *B4galnt2* genotype as in the cecum.

Additionally, I present the results for the other interesting candidate from the cecum, ***Proteus*** (Otu000204), although it displays no significant association in the colon. At the DNA level (figure 38.1), as for *Morganella*, the prevalence is now higher in RIIIS/J homozygotes than in C57BL/6J homozygotes, the overall prevalence being lower than in the cecum. For inflammation, the relationship is the same as in the cecum: it is more prevalent in highly inflamed mice. For the interaction, it follows the same trend as in the cecum, as both occurrences in C57BL/6J homozygotes are in inflamed mice, while the single occurrence in RIIIS/J homozygotes is in a healthy one. However, the sample sizes are too small to reach significance. The trend is similar at the RNA (figure 38.2)- and activity levels (figure 38.3). It appears that *Proteus* is more active in inflamed mice, both mildly and highly inflamed, compared to healthy mice.

In conclusion, in the colon *Citrobacter* is an indicator of *B4galnt2* genotype, but not of inflammation. All indicator OTUs are associated with inflammation and/or the interaction between inflammation and *B4galnt2* genotype. Most of them seem to have a similar correlation with *B4galnt2* genotype: the bacteria are more prevalent and/or more abundant in inflamed C57BL/6J homozygotes, while absent from RIIIS/J homozygotes or present in healthy RIIIS/J homozygotes; only *Morganella* shows a different pattern, as it is more abundant in inflamed RIIIS/J homozygotes, which is the opposite to the genotype correlation in the cecum.

Figure 37.1

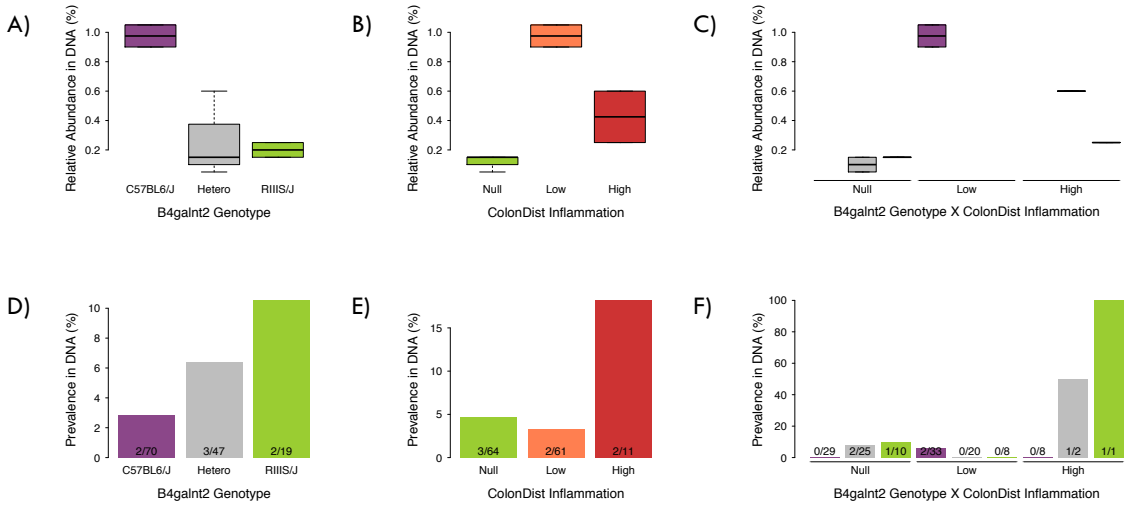

Figure 37.2

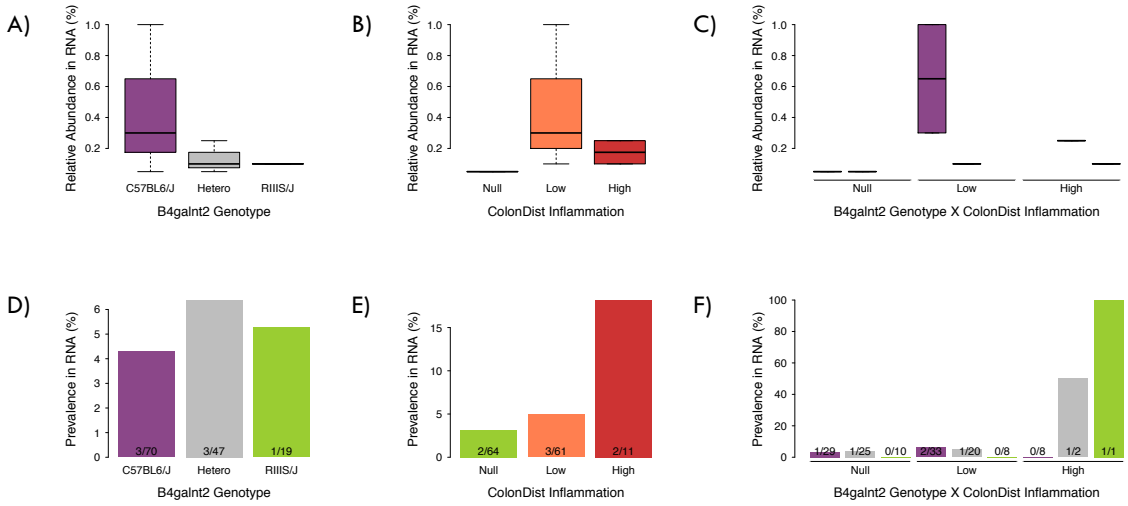

Figure 37.3

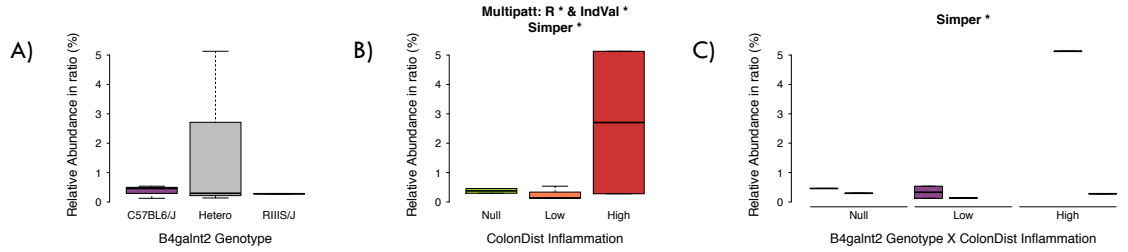

Figure 37: Non-zero relative abundance (ABC) and prevalence (DEF) of the indicator species **Otu000463** (*Proteobacteria; Gammaproteobacteria; Enterobacteriales; Enterobacteriaceae;* ***Morganella***) in the DNA (21.1) and RNA (21.2) samples, and at the activity level (21.3), according to the *B4galnt2* genotype (AD), the cecum inflammation (BE), and their interaction (CF). The test performed were three indicator analysis functions: multipatt R, multipatt IndVal, and simper; the Kruskal-Wallis test on all values and only on non-zero values; and a chi square test on the prevalence. The results of the tests are reported in the title of the graphs with the following code: p<=0.001 ***, p<=0.01 **, p<=0.05 *. Outliers are not displayed.

Figure 38.1

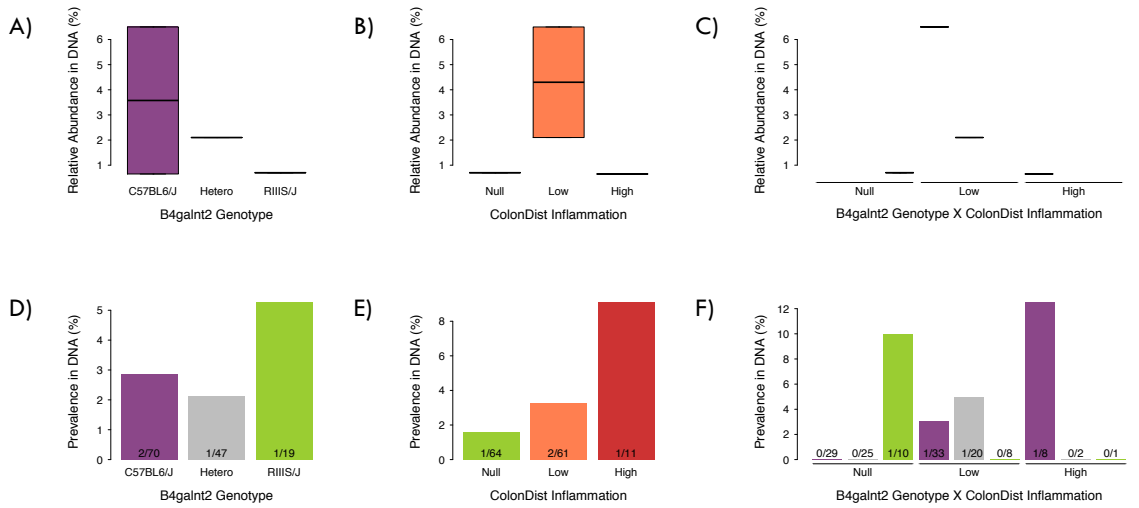

Figure 38.2

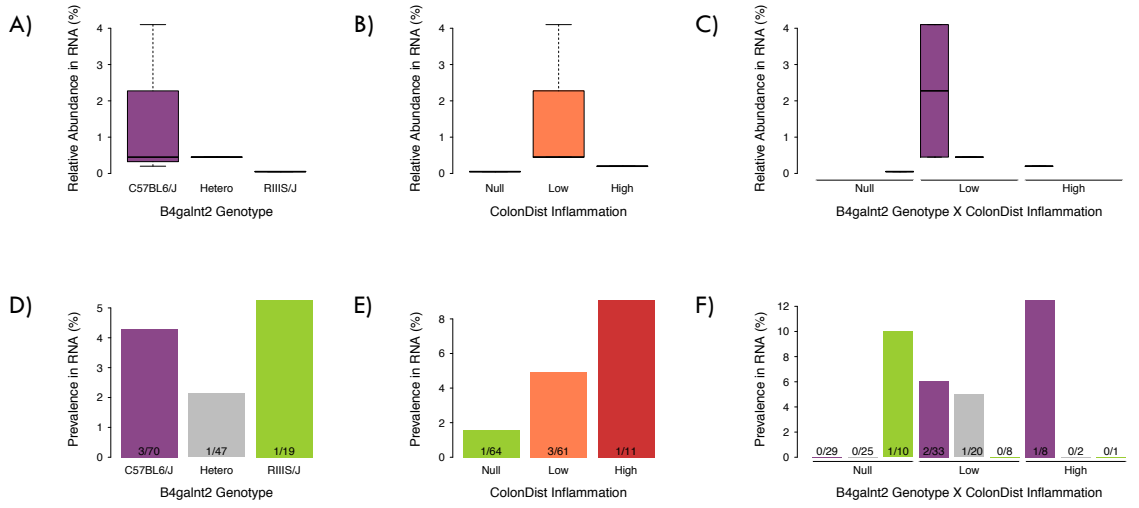

Figure 38.3

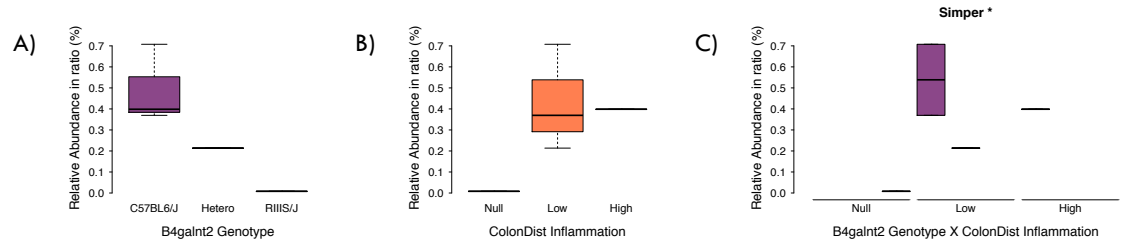

Figure 38: Non-zero relative abundance **(ABC)** and prevalence **(DEF)** of the indicator species **Otu000204** (*Proteobacteria; Gammaproteobacteria; Enterobacteriales; Enterobacteriaceae; **Proteus***) in the DNA **(21.1)** and RNA **(21.2)** samples, and at the activity level **(21.3)**, according to the *B4galnt2* genotype **(AD)**, the cecum inflammation **(BE)**, and their interaction **(CF)**. The test performed were three indicator analysis functions: multipatt R, multipatt IndVal, and simper; the Kruskal-Wallis test on all values and only on non-zero values; and a chi square test on the prevalence. The results of the tests are reported in the title of the graphs with the following code: p<=0.001 ***, p<=0.01 **, p<=0.05 *. Outliers are not displayed.

VI. Blood pathogens

Finally, although the presence of candidate pathogens in the intestine of the mice is compelling, it does not rule out the alternative hypothesis of protection against systemic infections. To investigate whether a blood pathogen could contribute to the selection acting at *B4galnt2*, e.g. through the protection of mice carrying the vascular allele (RIIIS/J homozygotes and heterozygotes), I used the universal primers 27F and 1492R to amplify and sequence the full length 16SrRNA gene from the DNA extracted from blood samples. A surprisingly high proportion of mice revealed traces of bacterial DNA in their blood stream (figure 39A), with between 50 and 60% of mice from each genotype with detectable amount of bacterial DNA. This proportion does not seem to differ according to *B4galnt2* genotype, and if a trend is to be seen, it is the opposite as what I expected: blood bacteria are more prevalent in RIIIS/J homozygotes than in C57BL/6J homozygotes. Moreover, there seems to be some diversity in the bacteria present, as I could not directly sequence most of the samples due to the presence of multiple strains. Nonetheless, I obtained seven good sequences, which is very small compared to the 120 samples showing amplification, five of them classify to *Staphylococcus* (three mice are C57BL/6J, two are heterozygotes), one to *Lactobacillus* (in a heterozygote mouse), and the last one to *Leuconostoc* (in a C57BL/6J individual). It is interesting to find members of *Staphylococcus*, since it is one genus known to interact with von Willebrand factor, and could thus interact with *B4galnt2* genotype. However, to be able to understand the relationship between specific bacterial infection and *B4galnt2*, I need to identify all strains present. Unfortunately, the high number of samples with amplification did not allow me to do clone libraries in order to identify all the bacteria present.

Since *Staphylococcus* is an interesting bacterium to investigate in the context of *B4galnt2* genotype, I used primers from the literature (*Mason, Blevins et al. 2001*), specific for *Staphylococcus*, to gain a better insight into this genus. The majority (72/93) of samples yielded good quality sequences, allowing me to test for correlation with *B4galnt2* genotype (figure 39B). Although the comparison between genotype does not reach significance, it seems that the trend is that fewer RIIIS/J homozygotes carry *Staphylococcus* compared to C57BL/6J homozygotes, which agrees with the theory that RIIIS/J homozygotes should be protected against *Staphylococcus* due to less- and/or modified von Willebrand factor, which *S. aureus* uses in its infection mechanism. The heterozygotes however seem to have a higher prevalence of *Staphylococcus* than both homozygotes, suggesting some sort of interaction between the alleles.

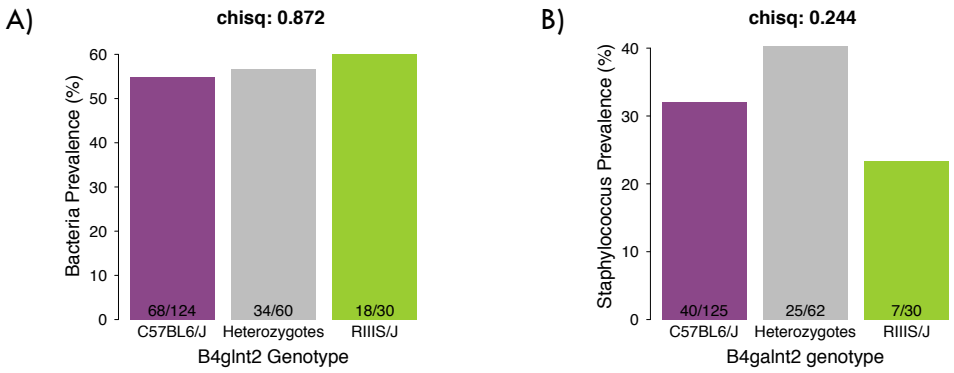

Figure 39: Prevalence of bacteria in the blood. A) General prevalence of bacteria as detected by amplification of the nearly full-length 16SrRNA gene with the universal primers 27F and 1492R. No sequencing was done to confirm the detection. B) Prevalence of Staphylococcus bacteria, as detected by amplification and sequencing of a 791 bp fragment of the 16SrRNA gene, with Staphylococcus specific primers (*Mason, Blevins et al. 2001*).

In conclusion, the alternative hypothesis of protection against blood pathogen cannot be ruled out, since a high proportion of mice carry bacterial DNA in their blood stream. One candidate might be *Staphylococcus*, which show a non-significant trend between their prevalence and *B4galnt2* genotype.

Conclusion

The biology and genetics of *B4galnt2* in house mice has been studied now for over 20 years: from the first identification of the mutation linked to the reduced plasma level of von Willebrand factor in RIIIS/J laboratory mice (*Mohlke, Nichols et al. 1996*), to the latest study showing its implication in pathogen susceptibility (*Rausch, Steck et al. 2015*); two studies showed that the allele provoking the costly bleeding disorder is abundantly present in various species of the genus *Mus*, and that it shows signs of recent positive selection in a *Mus musculus domesticus* Population from South France (*Johnsen, Teschke et al. 2009*), while showing signs of long-term balancing selection at the genus level (*Linnenbrink, Johnsen et al. 2011*). The nature of *B4galnt2*, a glycosyltransferase, suggests that the unknown benefit balancing the cost of a bleeding disorder in wild mice might be the interaction with microbes in the intestine, and indeed two studies showed the influence of *B4galnt2* intestinal expression on the composition of the microbiota (*Staubach, Kunzel et al. 2012*, *Rausch, Steck et al. 2015*), the latter of which even showing its influence on susceptibility to a model of *Salmonella typhimurium* infection. These results provide reasonable evidence that both *B4galnt2* murine alleles were maintained through balancing selection in various populations of *Mus* species, due to the threshold between bleeding disorder and pathogen resistance; however, one key component is still missing: what pathogen might be driving the selection at *B4galnt2*?

In order to answer this question, I performed an extensive five-week field trip to south France, and collected over 200 mice, aiming to identify the pathogen driving the selection at *B4galnt2*. To do that, I needed a number of information about the mice: *B4galnt2* genotype, health status, and microbial communities. As a proxy for the health status of the mice, I explored two methods: scoring of inflammation markers in histological slides, and qPCR of immune genes. Surprisingly, although both measures seem to correlate to some extent, their relationship to *B4galnt2* genotype is different: the inflammation scores from the histological slides show correlation to *B4galnt2* genotype in the intestine, while the level of immune gene expression does not. This could come from the fact that the histological scores are broader, since they are the sum of various markers, but also more precise since it is a direct visual assessment of the local signs of inflammation. In comparison, the immune gene expression is more targeted, since it measures only one actor of the inflammation, but more general in the sense that it is assessed at the level of whole organ, although a specific cell subset might be responsible. In the wild, it appears that inflammation occurs at high prevalence, which could be expected from wild animals that are in constant contact with numerous microorganisms, many of which have pathogenic potential. In this

context, *B4galnt2* might have only a targeted influence on the defense against infection and thus a limited influence on the inflammation levels, influence that might only be detected through the broad but local assessment of inflammation by histological scoring.

Interestingly, the correlation between *B4galnt2* and inflammation, as estimated by the histological scores, is dependent on the organ. Indeed, the systemic organs (liver and spleen), as well as the distal colon do not show any correlation with genotype, while the ileum, cecum and proximal colon do. This is however not entirely surprising. First, systemic and intestinal infection represent very different cases, as the first is not expected to be very prevalent, as blood stream infections generally have dire consequences, while the second might be more prevalent, since the intestine is a contact zone with the external environment and more importantly plentiful of microorganisms, which are constantly interacting with the host immune system (among other pathways). Finally, the intestine is not a homogenous organ; the different sections have different chemical and physical properties that represent divergent ecosystems in terms of microbial communities, which are not expected to behave similarly. Moreover, the fact that systemic inflammation shows no relationship to *B4galnt2* while intestinal inflammation does points in the direction of my main hypothesis -- protection of the RIIIS/J allele against intestinal pathogens -- rather than the alternative hypothesis of protection against blood pathogens.

In order to identify candidate pathogens, I sequenced the bacterial 16S rRNA gene to investigate the microbial community of the cecum and distal colon. To have a complete picture of the communities, I sequenced at both the DNA and RNA level. The advantage of the RNA over the DNA is that it is expected to represent more biologically relevant bacteria. Indeed, with DNA we detect every bacterium present in the sample, regardless of whether it is part of the resident community or coming from external sources, or whether it comes from living bacteria or remains of dead microorganism. Using the RNA allows, at the very least, to focus on living bacteria, and might also be less biased by "passers by", which are probably less active than resident microbes, as they are not in their preferred environment. It does not however distinguish between growth-related and non-growth related activities, but is associated to all protein-producing activities. Moreover, the raw level of RNA is biased by the number of bacteria present, and need thus to be normalized to DNA level (+1 to avoid mathematical issues with zeros), the same way that gene expression is normalized to house keeping genes in qPCR assays. This ratio however yields to highly heterogeneous values between samples, as it has virtually no upper limit. Indeed, the upper limit is a technical one, due to the sequencing depth: in the extreme case of a bacteria representing 100% of the community at the RNA level but not detected at the DNA level, the value of the activity would be the number of sequences of the RNA sample aka the sequencing

depth. Although unrealistic, this extreme example illustrates the problem of no upper limit to the activity value, which makes it difficult to compare samples, as their activity scales might be different. As I couldn't find any standard method in the literature to tackle this issue, I chose to normalize this ratio by samples, to express the importance of each bacterium relative to the total volume of activity of the microbial community.

Using these three markers of microbial communities (RNA, DNA and activity) and comparing single bacterial species prevalence, abundance and activity to the genotype of the mice and their inflammation status, I could identify several promising candidate pathogens. Moreover, it seems that the use of RNA, and especially of the activity ratio is especially relevant, since it allowed me to detect some interesting candidates that would have been missed had I used only DNA. To identify potential candidates, I used a combination of different tests, which is not a widespread approach, but it maximizes the chances of finding candidates as each test has a different focus, some relying more on the abundance data and some relying more on the prevalence data. Among all the candidate pathogens I identified, *Morganella* is probably the most promising one, as its pattern of prevalence, abundance and activity is very clear with regard to both cecal and colonic inflammation and it is a known opportunistic pathogen. Its relationship to *B4galnt2* is however different between the cecum and colon, with higher prevalence in C57BL/6J homozygotes in the cecum and higher prevalence in RIIIS/J homozygotes in the colon, suggesting that the hypothesized protective effect of the RIIIS/J allele may be complex.

Moreover, the high prevalence of bacterial DNA in the blood stream of the studied mice, and in particular the observed trend of *Staphylococcus* prevalence, also gives strength to the alternative hypothesis of protection against blood pathogens.

In conclusion, the novel combination of methods I applied lead to the discovery of several promising candidate pathogens that might play a role in the maintenance of *B4galnt2* alleles in the wild. These results strengthen the hypothesis that the cost of prolonged bleeding in RIIIS/J homozygotes is balanced out by protection against pathogen, but it also suggest that the protection conferred by the RIIIS/J allele might be more complex than previously thought, with different effects in the various regions of the intestine and in the blood stream.

Methods

I. Field work

In order to further investigate the relationship between *B4galnt2* expression, gut microbiota, and inflammation status, I went back to Espelette, France, where a high frequency of the RIIIS/J allele was previously identified (*Linnenbrink 2012*). In 5 weeks, we collected samples from 217 mice caught in 34 farms. The work was organized over a two-day period: the first day we prepared the traps and went to different farms to place them, the next day we came back to the farms to collect the traps and dissect the live-caught mice. The dissection was organized between four people:

- The first person was in charge of the euthanasia using CO_2;
- The second person measured the mouse (body and tail length, weight), determined its gender and collected ears;
- The third opened the mouse and collected mesenteric lymph node, liver, spleen and jejunum;
- The last collected cecum, colon, ileum and feces from the colon.

The different organs collected were stored under various conditions, for different purposes, detailed in table 5. The histological samples were used to estimate the inflammation status of the mice. Samples stabilized in RNAlater were first stored at 4°C for 24h, then the RNAlater was pipetted out and the samples were stored at -20°C. The cryosolution was a solution containing culture media (Heart-Brain infusion), glycerol (final concentration of 20%), sodium D-isoascorbate (an oxygen scavenger, at a final concentration of 3.2 $g.L^{-1}$), and resazurin salt, a color indicator for the anaerobic condition. To be sure this solution remains anoxic, we added solubilized sodium hydrosulfide (another oxygen scavenger) until the medium lost the pink coloration specific to oxidized solution.

A summary of the 34 farms is presented in table 6. I used the "haversine" formula to calculate the distances between farms from their GPS coordinates. I then grouped the farms in "families" when they were less then 2km apart, based on previously published sampling strategies (*Ihle, Ravaoarimanana et al. 2006*), and groups of bigger size containing farms that appeared to cluster on the map (figure 1).

Table 5: Summary of the organs taken from the mice in the field, the storage conditions and intended purposes.

Organ	Storage Solution	Storage Temperature	Purpose
Ears	--	-20°C	Genotyping, Microsatellite, D-loop
Mesenteric lymph node	Formalin	4°C	Histology
Liver	Formalin	4°C	Histology
Spleen	Formalin	4°C	Histology
Jejunum	Formalin	4°C	Histology
Ileum	Formalin	4°C	Histology
Colon1 (proximal)	Formalin	4°C	Histology
Colon2 (distal)	RNA later	-20°C	16S rRNA profiling
Colon3 (distal)	Formalin	4°C	Histology
Cecum1	RNA later	-20°C	16S rRNA profiling
Cecum2	Cryosolution	-80°C	Bacterial cultivation
Cecum3	Formalin	4°C	Histology
Cecum4	AllProtect	4°C	16S rRNA profiling & qPCR
Blood	--	-80°C	Detection of systemic pathogens
Feces	PBS	-20°C	Macroparasite detection

Table 6: Location of the 34 farms where wild mice were collected near Espelette, FR. The coordinates are given in latitude and longitude. Families are groups of farm that are closer than 2km to each other; groups are bigger groups of farms that are geographically close to each other.

Farm	Latitude	Longitude	Family	Group	#Mice
JJM01	43.327778	-1.476222	F01	G01	2
JJM02	43.318543	-1.490102	F02	G01	10
JJM08	43.337143	-1.495154	F06	G01	2
JJM04	43.306704	-1.536797	F03	G02	2
JJM05	43.298985	-1.547350	F03	G02	4
JJM06	43.304951	-1.551148	F03	G02	4
JJM07	43.318737	-1.528070	F05	G02	2
JJM09	43.316138	-1.560482	F07	G02	11
JJM10	43.295746	-1.507033	F08	G03	2
JJM12	43.265708	-1.362697	F09	G04	4
JJM13	43.261364	-1.378737	F10	G04	1
MJJ01	43.343913	-1.551134	F11	G05	16
MJJ03	43.367031	-1.595092	F12	G06	1
MJJ06	43.375537	-1.563524	F13	G06	11
MJJ07	43.376255	-1.582535	F14	G06	3
MJJ09	43.394596	-1.532853	F15	G07	2
MJJ10	43.393006	-1.534956	F15	G07	5
MJJ11	43.421932	-1.542395	F16	G08	1
MN02	43.362902	-1.450289	F17	G09	7
MN03	43.365562	-1.447467	F17	G09	12
MT01	43.355250	-1.447263	F17	G09	16
MT15	43.353908	-1.452392	F17	G09	3
MN12	43.389935	-1.492119	F18	G10	2
MN24	43.328051	-1.328891	F19	G11	2
MN29	43.329674	-1.320866	F19	G11	2
MN32	43.335121	-1.311468	F19	G11	15
MN26	43.343003	-1.290997	F20	G11	15
MN41	43.322619	-1.292134	F21	G11	7
MT13	43.388949	-1.335757	F22	G12	8
MT14	43.376863	-1.325136	F22	G12	2
MT17	43.383507	-1.346379	F22	G12	7
MT26	43.368674	-1.342967	F22	G12	2
MT35	43.380325	-1.366892	F23	G12	13
MT21	43.376832	-1.332410	F22	G12-21	21

II. DNA/RNA/Protein extractions

Total DNA was extracted from the ears using the DNeasy Blood & Tissue Kit from Qiagen. Each sample was quantified individually using a NanoDrop (ThermoScientific).

DNA, RNA & Proteins were extracted from the cecum using the Qiagen AllPrep DNA/RNA/Protein Mini kit. First the tissues were washed in PBS in order to remove as much AllProtect reagent as possible. Then the tissues were disrupted and homogenized in 600 µL of RLT 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. The lysates were then filtered with the QIAshredder spin column (Qiagen). Finally, DNA, RNA and Protein were extracted using the Qiagen AllPrep DNA/RNA/Protein Mini kit following the manufacturer's instructions. After elution, the RNA was treated with DNase and repurified using sodium acetate and ethanol, as explained in the manufacturer's guide.

DNA and RNA were extracted from the colon. First the tissue was disrupted and homogenized in 600 µL of RLT 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 before extraction using the AllPrep DNA/RNA 96 kit from Qiagen following the manufacturer's instructions

In both the DNA/RNA/Proteins and DNA/RNA extractions, β-mercaptoethanol was replaced by a non-toxic alternative: Tris (2-carboxyethyl) phosphine hydrochloride (TCEP). cDNA was synthesized using the High-Capacity cDNA Reverse Transcription Kit (Applied Biosystems) and the manufacturer's instructions.

Microbial DNA was extracted from the blood samples using the MolYsis™ Complete5 without β-mercaptoethanol (Molzym), following the manufacturer's instructions.

III. *B4galnt2* & mitochondrial D-loop Genotyping & Sequencing

The diagnostic "fragment 5" of *B4galnt2* (*Johnsen, Teschke et al. 2009*) was sequenced as follow: PCR reactions were 5µL Multiplex PCR kit (Qiagen), 1µL primer (10µM), 2µL water and 1µL DNA template (20ng/µL). The amplification program is presented in table 7.

PCR products were then treated with ExoSap as follow: 2.5µl PCR product and 1µl ExoSap, incubated 15 min at 37°C then 15 min at 80°C. The sequencing reaction was 1.75µl sequencing buffer, 0.5µl Big Dye, 6.75µl H2O, 0.5µl primer, and 0.5µl ExoSap treated PCR product. The sequencing program is presented in table 8.

Table 7: Amplification program for *B4galnt2* genotyping.

Temperature	Time	Cycles
95°C	15 min	
94°C	30 sec	x 42
63°C	90 sec	
72°C	90 sec	
72°C	10 min	
12°C	∞	

Table 8: Sequencing program for *B4galnt2* genotyping.

Temperature	Time	Cycles
96°C	1 min	
96°C	10 sec	x 30
55°C	15 sec	
60°C	4 min	
12°C	∞	

The sequencing run was finally performed on an ABI 3730 automated sequencer (Applied Biosystems), and sequences were edited in GENEIOUS 7.0 (Biomatters Ltd) and transferred in MEGA 5 to align with RIIIS/J and C57BL/6J reference sequences.

An 885 bp portion of the mitochondrial D-loop was sequenced as described by (*Prager, Sage et al. 1993*), sequences were edited in GENEIOUS 7.0 (Biomatters Ltd) and transferred in MEGA 5 to align with *Mus musculus domesticus*, *Mus spretus* and *Mus spicilegus* reference sequences, to verify that all samples were *Mus musculus domesticus*, and with samples from the previous French and German collection (*Linnenbrink, Wang et al. 2013*).

IV. Microsatellites

IV.1 Typing of the microsatellites

Two sets of microsatellites loci were used. First, 12 linked loci located around the start position of *B4galnt2* (*Johnsen, Teschke et al. 2009*) were used to study the microsatellite variation associated to *B4galnt2* alleles. Second, a set of 18 neutral unlinked loci (*Thomas, Moller et al. 2007*, *Linnenbrink, Wang et al. 2013*) was used to estimate the background variation of

microsatellites along the genome. In both cases, the forward primer was labeled either with FAM or HEX to allow multiplexing. PCR reactions were 2.5µL Multiplex PCR kit (Qiagen), 0.1µL Primer (10µM), water to 3µL and 2µL DNA template (5ng/µL). The amplification program is presented in table 9.

Table 9: Amplification program for microsatellite typing.

Temperature	Time	Cycles
95°C	15 min	
95°C	30 sec	x 28
60°C	90 sec	
72°C	90 sec	
72°C	10 min	
12°C	∞	

PCR products were then diluted 1:20 in water and 1µL of diluted product was added to 10µL HiDi formamide and 0.1µL of 500 ROX size standard before denaturation (90°C 2 min; 20°C 5 min; 12°C ∞; ice shortly). The sequencing run was finally performed on an ABI 3730 automated sequencer (Applied Biosystems). The alleles were called using GENEIOUS 7.0 (Biomatters Ltd) with the microsatellite plug-in.

IV.2 Haplotype reconstruction

To be able to study the microsatellite variation at the 12 *B4galnt2*-linked loci with respect to the *B4galnt2* allele, I had to reconstruct the haplotype phase of these markers. First, I converted the PCR product length obtained from GENEIOUS into number of repeats, using the following relationship: Number of repeats = (amplicon length - flanking region length) / 2. Then, I submitted these repeat data to PHASE 2.1 (*Stephens, Smith et al. 2001*). I ran PHASE multiple times with different seeds for the pseudo-random number generator. In order to compare the runs, I followed the manual's recommendations and created two python scripts to 1) estimate if the runs are consistent with each other based on the _freqs output file, 2) choose the best run using the _monitor output file. Ultimately, I ran PHASE with a number of iterations of 10000, a thinning interval of 100 and a burn-in of 10000 (100 times the default values).

IV.3. Microsatellite analysis

I used the software GenoDive 2.0 (*Meirmans and Van Tienderen 2004*) to analyze the *B4galnt2*-linked microsatellite variation. I calculated the expected heterozygosity for each population and an Fst analogue (*Excoffier, Smouse et al. 1992*, *Michalakis and Excoffier 1996*), both with regard to the *B4galnt2* genotype. The obtained results were transferred to Excel or R for graphical visualization.

To be able to investigate whether the genetic background and the genetic architecture of the studied population had any influence on the intestinal microbiota, I used the data from the 18 neutral microsatellite markers to determine the population structure using STRUCTURE (*Pritchard, Stephens et al. 2000*), first with a 10,000 burn-in period and 50,000 Markov Chain Monte Carlo (MCMC) simulations with 20 iterations per number of clusters (K) for K from 2 to 40, then with a 500,000 burn-in period and 1,000,000 Markov Chain Monte Carlo (MCMC) simulations with 20 iterations per number of clusters (K) for K from 2 to 18. I used Structure Harvester to determine the best value of K (*Evanno, Regnaut et al. 2005*, *Earl and Vonholdt 2012*), and CLUMPP (*Jakobsson and Rosenberg 2007*) to compare and average the results of the 20 runs for the best K and R to plot the results. Most of the individuals showed >80% membership to one cluster, allowing the categorization of the structure output. To determine the category that admixed individuals belongs to, I used the R package vegan (*Dixon 2003*), PHANGORN (*Schliep 2011*) and ape (*Paradis, Claude et al. 2004*, *Paradis 2012*), to 1) build a distance matrix between individuals based on the STRUCTURE output, using VEGDIST with the Euclidean distance; 2) build a neighbor-joining tree from this distance matrix; 3) visualize the tree to determine the best clustering of individuals (figure 4B). Additionally, I used GenoDive 2.0 (*Meirmans and Van Tienderen 2004*) to calculate the chord distance between each individuals, and KINGROUP (*Konovalov, Manning et al. 2004*) to estimate the relatedness between individuals, using all five methods available (Kinship, Konovalov & Heg, Lynch & Ritland, Maximum Likelihood, Wang).

V. Histology & inflammation scores

Histological samples were used to assess the degree of inflammation affecting the mice, potentially correlating with bacterial infection. The inflammation score for the Ileum, jejunum, cecum, colon proximal and colon distal inflammation was calculated as the sum of desquamation (score between 0 and 3), infiltration of polymorphonuclear (PMN) leukocytes in the mucosa (score

from 0 to 3) or in the submucosa (score from 0 to 3) and necrosis (score from 0 to 3), for a total score ranging from 0 to 12. The scoring of the samples was done three times independently and blind to the mouse genotype. The average of the three scores was used for further analysis.

For the spleen, the amount of granulopoiesis, white pulp, red pulp, hematopoiesis, erythropoiesis spleen, general inflammation and signs of lymphoma were observed; and for the liver, the amount of glycogen, general inflammation, toxic changes and intranuclear inclusion bodies. The scoring was done blind to the mouse genotype.

The scores were exported into R along with the *B4galnt2* genotype information and tested using the non-parametric tests of Wilcoxon and Kruskal-Wallis, for quantitative data, and Chi Square for prevalence data.

VI. Cecum4 qPCR of immune genes

To have complementary information on the immune status of the mice, I performed qPCR of two immune genes on the cecum samples. I used the primer/probe approach to allow multiplexing of the genes, thus reducing the amount of PCR needed and increasing the reliability of the assay since target gene and control gene were measured in the same reaction. I chose assays from the company IDT (Integrated DNA Technologies):

- Mm.PT.58.29815602 is specific for Hprt1 that I labeled with Hex
- Mm.PT.58.41152792 is specific for Interferon Gamma (IFNgamma) that I labeled with FAM
- Mm.PT.58.42151692 is specific for CCL2/MCP1 that I labeled with Cy5

I solubilized the primer/probe assays in 500µl of water to obtain 20x solutions. As advised by IDT, I used the Agilent Brilliant Probe Multiplex Master Mix (Cat No 600553) that is designed to amplify up to four targets in one reaction. The PCR protocol was 0.5µl of each target gene assay (MCP1 and IFNgamma), 0.2µl of the control gene assay (HPRT1), 5µl of the Agilent master mix, 2µl of water. 8µl of that mix was distributed in each well with 2µl cDNA template. For each samples I ran 3 technical replicates. I ran the qPCR on the PikoReal 96 Real-Time PCR System, using program presented in table 10.

Table 10: Amplification program for Cecum4 immune genes qPCR.

Temperature	Time	Cycles
95°C	10 min	
95°C	15 sec	x 50
60°C	60 sec *	
60°C	30 sec	
Melt ramp 60-95°C *		
4°C	10 sec	

* Data acquisition

I used the PikoReal Software to generate the Cq data, using the individual Cq method. I then exported the Cq data, and processed them in Excel. First, I calculated the Cq difference between target and control gene. Then I calculated the average target-control difference of the three technical replicates. I excluded samples for which one or two replicates didn't work and those with high standard deviation between replicates. Then I used the $2^{-\Delta\Delta CT}$ method for relative quantification. For both genes, I used the sample with the highest CT value (i.e. the sample with the lowest gene expression) as reference, so that the lowest expression value is 1.

I exported the relative quantities into R where I did Pearson correlations for the comparison of quantitative variables (IFNgamma, MCP1 and Cecum4 inflammation score) and Kruskal-Wallis tests with the categories.

VII. 16S rDNA/rRNA profiling for the cecal and colonic microbiota

VII.1 PCR & NGS Sequencing

The primers 27F and 338R were used to amplify the V1-V2 regions of the 16S rDNA gene from the **cecum 4** and **colon 2** samples, from the DNA and cDNA. The primers were barcoded to allow multiplexing. PCR reactions were 10.25µl H_2O, 5µL buffer, 0.50µl dNTPs, 0.25µl Taq polymerase (Phusion High-Fidelity DNA Polymerase - Thermo Scientific), 4µl of 2µM Primer and 1µl DNA/cDNA template. The PCR program is presented in table 11.

Table 11: Amplification program for 16S rRNA profiling.

Temperature	Time	Cycles
98°C	30 sec	
98°C	9 sec	x 30
55°C	1 min	
72°C	90 sec	
72°C	10 min	
12°C	∞	

The PCR products were quantified on the gel using the GelDoc XR+ (BioRad). The samples were mixed in subpools in equal amounts of DNA. The subpools were purified through gel extraction with the MiniElute kit (Qiagen). Purified subpools were quantified with the fluorescence NanoDrop (Thermo Scientific), and mixed in a final library so that each sample has the same final concentration. The Library was sequenced on an Illumina MiSeq machine using the v2 kit with 2x250 bp reads.

For the colon 2, about half of the samples displayed additional bands of bigger size (~500 bp and ~700 bp) in much higher quantity than the expected ~400 bp V1-V2 band, rendering it impossible to use directly. For these samples, the individual bands were purified through gel extraction with the MiniElute kit (Qiagen). The extracted ~400 bp product was used as template for an additional PCR as described above, which was treated as the non gel-purified samples, and combined in the same MiSeq library. The additional bands were sequenced through classical Sanger sequencing as described for *B4galnt2* genotyping, and identified as *Helicobacter* species by the online RDP classifier using the training set 14 (*Cole, Wang et al. 2014*). A more precise classification was obtained by comparison to type strain reference sequences (table 12).

Table 12: Summary of the reference 16S rRNA genes used for phylogenetic analysis.

Genus	Species	Gene Identification
Helicobacter	acinonychis	M88148
Helicobacter	anseris	DQ415545
Helicobacter	aurati	AF297868
Helicobacter	baculiformis	EF070342
Helicobacter	bilis	U18766
Helicobacter	brantae	DQ415546
Helicobacter	canadensis	AM998803
Helicobacter	canis	AY631945
Helicobacter	cholecystus	AY686606
Helicobacter	cynogastricus	DQ004689
Helicobacter	equorum	AM998804
Helicobacter	felis	M57398
Helicobacter	fennelliae	M88154
Helicobacter	heilmannii	HM625820
Helicobacter	hepaticus	U07574
Helicobacter	macacae	AF333338
Helicobacter	marmotae	AF333341
Helicobacter	mastomyrinus	AY742307
Helicobacter	mesocricetorum	AF072471
Helicobacter	muridarum	M80205
Helicobacter	mustelae	M35048

VII.2 Sequences processing

First I generated the demultiplexed Fastq files using CASAVA (Illumina) allowing no mismatches in the barcodes, and using the "eamss" algorithm that attributes the lowest quality value to the terminal bases of a read that have unreliable quality and base calls, allowing for later trimming of low quality endings of any read. Then I designed a pipeline for quality filtering and sequence processing using different softwares: USEARCH v7 (*Edgar 2010*), UCHIME (*Edgar, Haas et al. 2011*), MOTHUR v.1.33.3 (*Schloss, Westcott et al. 2009*), FastX Toolkit (*Hannon 2010*) and self-made python scripts. The pipeline is as follow:

1) Merge the forward and reverse read, filter the quality.

- *usearch -fastq_mergepairs:* trim the reads at the first base with a quality below 5 (allows to remove low quality tails), keep only trimmed reads that are 200 bp or longer, allow only one mismatch between forward and reverse reads, keep only merged reads (forward aligned with reverse) that are between 300 and 350 bp and that have an overlap of at least 100 bp.
- *fastXtoolkit -fastq_quality_filter:* keep only sequences that have 99% of theirs bases with a quality of 30 or higher.
- *usearch -fastq_filter:* keep only sequences that have an expected error of 0.1 or lower.

2) Detect the chimeras using two databases (Chimera Slayer Gold and RDP 9), remove the chimeras detected by UCHIME using a self-made python script.

- *usearch -uchime_ref:* detect the chimeras in the data set by comparing with the chimera slayer gold database on one hand, and the rdp9 gold data base on the other hand, both available for download on the UCHIME website.
- *python:* remove all the detected chimeras from the data set.

3) Create the group file needed by MOTHUR using a self-made python script and concatenate all sample fasta in one fasta file.

For this step, I combined data from the cecum 4 and colon 2 to allow direct comparison of the two data sets. I also combined the published sequences from the "Linnenbrink mice" (*Linnenbrink, Wang et al. 2013*), and reprocessed them to be comparable to my data sets, although they were sequenced on the Roche 454 platform which reduces the comparability.

4) Continue the sequence processing in MOTHUR, following a customized version of the MiSeq SOP from Patrick Schloss, available on the MOTHUR website.

- *screen.seqs:* Remove sequences that have homopolymer longer than 6 bp.

- *unique.seqs:* Reduce the fasta file to only unique sequences to ease the computation of downstream steps.
- *align.seqs:* Align sequences to the silva bacteria reference database version 119. Prior to alignment, the database was reduced to only the V1-V2 region using the pcr.seqs command with the "oligo" option, "start" and "end" option.
- *screen.seqs:* Remove aligned sequences shorter than 300 bp or longer than 350 bp, or that do not align with the reference (start and end option).
- *filter.seqs:* Filter aligned sequences to remove columns containing only gaps.
- *unique.seqs:* Reduce the aligned fasta file to only unique sequences to ease the computation of downstream steps.
- *pre.cluster:* bin sequences that have 3 differences (i.e. 1% difference) or less together to account for PCR and sequencing errors.
- *classify.seqs:* classify the sequences by comparison to the Ribosomal Database Project (RDP) training set 14 (Cole, Wang et al. 2014), with a threshold of 60% confidence.
- *remove.lineage:* remove sequences that were not identified as bacteria ("unknown").
- *split.abund:* keep sequences that are present in 3 or more copies in the data set to remove PCR and sequencing errors.
- *cluster.split:* bin the sequences into operational taxonomic units (OTU), using the classification information to split the data set into order level clusters (taxlevel=4) and a cutoff of 9% differences.
- *get.groups:* separate the data set according to the organ the samples belong to.
- *merge.groups:* some samples were repeated in different MiSeq library. After verifying in Excel or R that the repeats of the same sample were consistent, I merged them to increase the read depth of the concerned samples.
- *sub.sample:* normalize the read depth for each sample at 8000 reads for the cecum 4 and 2000 for the colon 2, as for unknown reason, the sequencing depth was much lower for the colon than for the cecum. As the “Linnenbrink mice” were sequenced on the Roche-454 machine, the coverage is much lower so I normalized at 700 reads per sample.
- *list.seqs, get.seqs, make.shared, classify.otu:* reduce the OTU files to the same sequences as in the normalized fasta files, create the OTU count tables to be used in downstream community analysis with R. Obtain a classification of the OTUs based on the classification of the sequences that belongs in the OTU, using a threshold of 51%.

The OTU tables obtained from MOTHUR were exported in R for downstream analyses.

VII.3 OTU tables analysis

First I defined a core microbiome, which presents at least two non-negligible benefits: first it removes bacterial groups that would never reach significance in any statistical test due to their rareness, thus reducing the number of tests and improving the computational burden of the downstream analyses. Second, it reduces the noise of the community-wise signal, thus improving the detection of any effect from the factors studied (host genetic, environment...). Moreover, the core microbiota corresponds to a biological concept now mostly accepted in the scientific community, which defines the community of resident microbiota that have long-term interaction with the host, not accounting for the "passers by", which interact only briefly with the host and the resident microbiota. How to define such a core microbiota in community analysis is however subject to debate, and no consensus was reached in the scientific community. I chose to define the core microbiota as OTUs that are present in 5 or more samples, and that reach 1% of the community in relative abundance in at least one sample. I defined the core for DNA and cDNA separately.

I defined the activity of the bacterial species as the ratio between the abundances in the cDNA relative to the abundances in the DNA. This normalizes the activity of the bacteria detected via the cDNA analysis to the number of present bacteria detected by the DNA analysis. To avoid issues with divisions by zero, I added 1 to the DNA abundances. As this ratio leads to highly variable range of values between OTUs and samples, which are difficult to compare, I normalized the activity per sample so that each sample's community has a total activity of 1.

For the cecal microbiota, some extraction negative controls showed signs of contamination so I used the package SourceTracker (*Knights, Kuczynski et al. 2011*) to identify potentially contaminated samples. This identified two contaminated samples that I removed from the dataset for downstream analyses.

For Beta diversity measures, I used the vegan package (*Dixon 2003*). First I used VEGDIST to obtain a distance matrix of my samples based on the OTU table, using Bray-Curtis (takes into account abundances) and Jaccard indices (only consider prevalence). Then I used the CMDSCALE and CAPSCALE functions to perform principal coordinate analysis (PCoA) and constrained ordinations respectively. To test for the influence of environmental and host factors on the microbiota, I used Adonis and ENVFIT on the PCoA, and the multivariate analysis of variance ANOVA on the constrained ordination, both by term and by axis. For all tests I used 1000 permutations. When the explanatory variables were distances (e.g. geographic, genetic

distances...), I performed a mantel test on the distance matrices to see if they were significantly correlated with the Bray-Curtis or Jaccard distances. I also tested whether the distances within/between groups were different using the non-parametric Kruskal-Wallis test.

For the identification of candidate pathogens, I performed an indicator species analysis on the OTU table to identify the candidate pathogen using the package INDICSPECIES (*De Caceres and Legendre 2009*). I tested the two methods implemented in the function MULTIPATT: the indicator values (INDVAL) (*Dufrene and Legendre 1997*, *De Caceres, Legendre et al. 2010*) and the biserial correlation coefficient (r) (*Chytry, Tichy et al. 2002*, *Tichy and Chytry 2006*). For both methods I used 10.000 permutations and the version of the algorithm that accounts for unbalanced design. I also used the alternative indicator species analysis "simper" from the package vegan (*Dixon 2003*), with 10.000 permutations. Additionally, I used Kruskal-Wallis test on the full OTU table, and on a reduced OTU table including only non-zero occurrences, coupled with a X^2 test on the prevalence data. I selected those species that were significant for at least three of the tests I conducted, plotted their relative abundance and prevalence, and examined all the graphs to find those with signals that would be consistent with our pathogen-driven hypothesis.

VIII. Investigating blood pathogens

To test whether the alternative hypothesis of protection against blood pathogen could be true, I screened the blood samples using the universal 16S rDNA primers 27F and 1492R. PCR reactions were 10.25µl H_2O, 5µL buffer, 0.50µl dNTPs, 0.25µl Taq polymerase (Phusion High-Fidelity DNA Polymerase - Thermo Scientific), 4µl of 2µM Primer and 2µl DNA template. The PCR program is presented in table 13. The PCR product was treated with ExoSap and sequenced following the same protocols as for the *B4galnt2* genotyping.

Table 13: Amplification program for 16S rRNA screening of blood samples.

Temperature	Time	Cycles
98°C	30 sec	
98°C	9 sec	x 30
55°C	1 min	
72°C	90 sec	
72°C	10 min	
12°C	∞	

As some of the identified bacteria belong to *Staphylococcus*, which is one of the pathogen known to attach von Willebrand factor as an invasion mechanism, I screened the samples again using *Staphylococcus*-specific 16S rDNA primers from the literature (*Mason, Blevins et al. 2001*). The PCR was performed using the GoTaq DNA Polymerase from Promega, with the following protocol: 0.2µL of 10 µM primers, 0.2µL dNTPs, 0.8µL $MgCl_2$, 0.1µL Taq, 5.5µL H_2O, with the PCR program presented in table 14. The PCR product was treated with ExoSap and sequenced following the same protocols as for the *B4galnt2* genotyping.

Table 14: Amplification program for Staphylococcus-specific screening of blood samples.

Temperature	Time	Cycles
94°C	3 min	
94°C	90 sec	x 35
55°C	60 sec	
72°C	60 sec	
72°C	10 min	
12°C	∞	

I edited the sequenced in GENEIOUS 7.0 (Biomatters Ltd) and exported the fasta file to MOTHUR (*Schloss, Westcott et al. 2009*) where I aligned the sequences to the silva v119 database, before binning the sequences into OTUs using the functions dist.seq and cluster. I transferred the data to R where I did a chi square test on the prevalence data with respect to *B4galnt2* genotype.

Supplementary figures

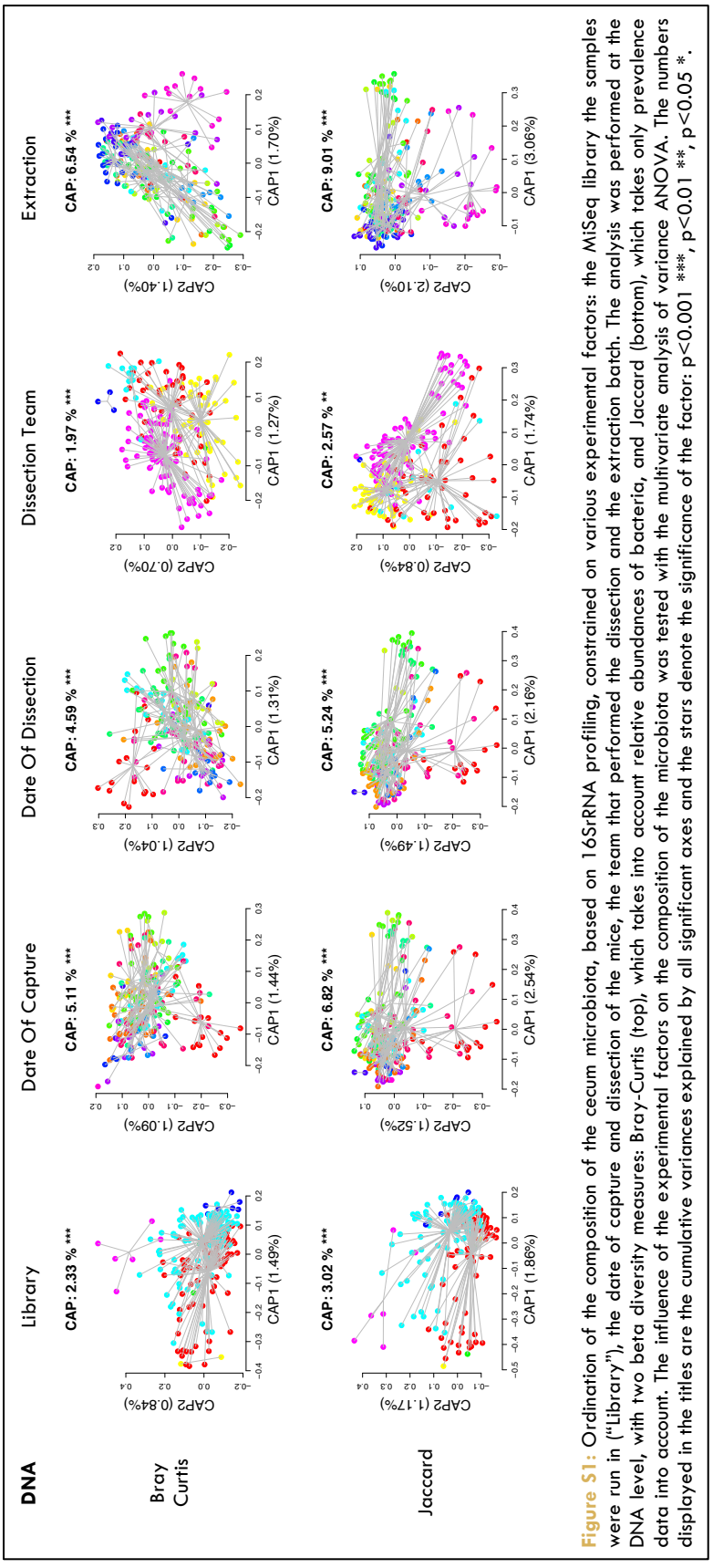

Figure S1: Ordination of the composition of the cecum microbiota, based on 16SrRNA profiling, constrained on various experimental factors: the MiSeq library the samples were run in ("Library"), the date of capture and dissection of the mice, the team that performed the dissection and the extraction batch. The analysis was performed at the DNA level, with two beta diversity measures: Bray-Curtis (top), which takes into account relative abundances of bacteria, and Jaccard (bottom), which takes only prevalence data into account. The influence of the experimental factors on the composition of the microbiota was tested with the multivariate analysis of variance ANOVA. The numbers displayed in the titles are the cumulative variances explained by all significant axes and the stars denote the significance of the factor: $p<0.001$ ***, $p<0.01$ **, $p<0.05$ *.

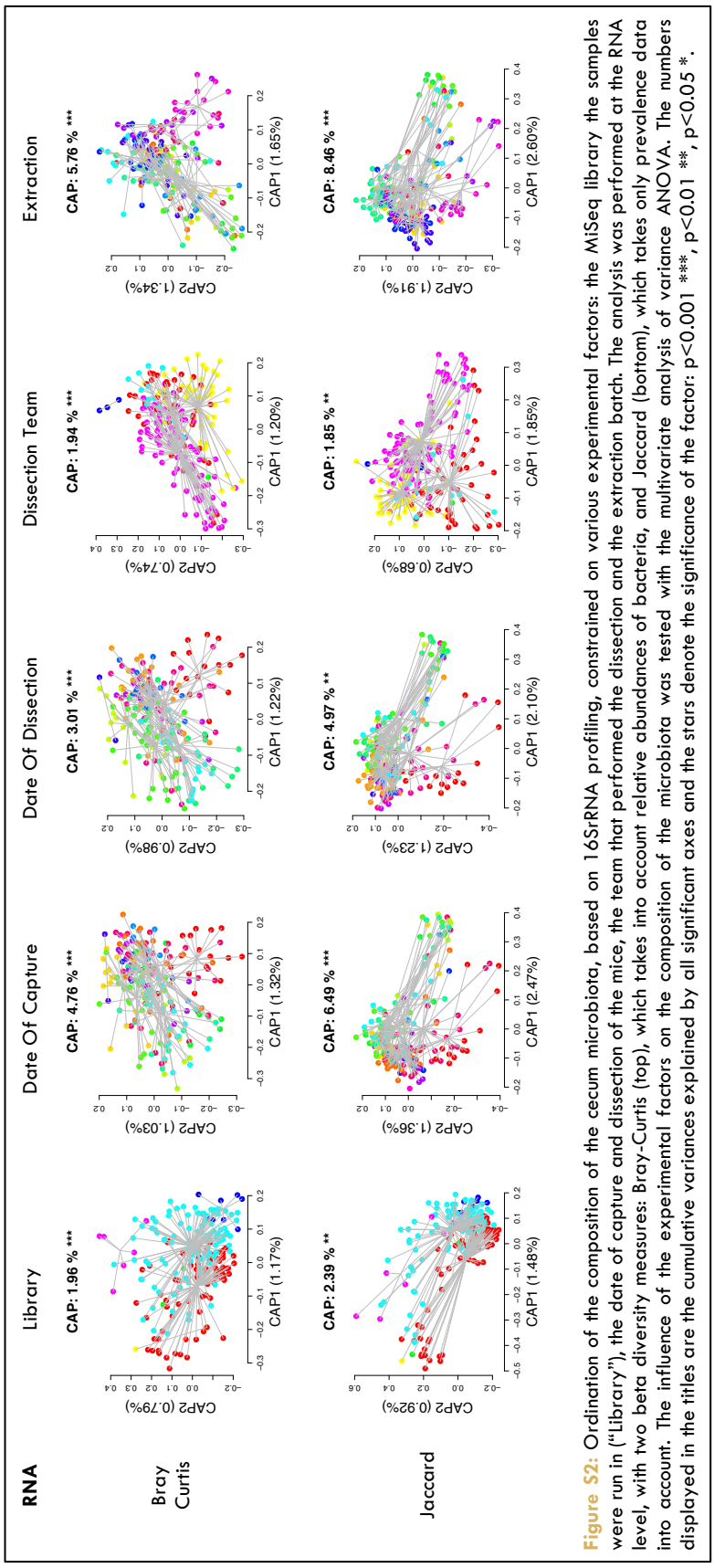

Figure S2: Ordination of the composition of the cecum microbiota, based on 16SrRNA profiling, constrained on various experimental factors: the MiSeq library the samples were run in ("Library"), the date of capture and dissection of the mice, the team that performed the dissection and the extraction batch. The analysis was performed at the RNA level, with two beta diversity measures: Bray-Curtis (top), which takes into account relative abundances of bacteria, and Jaccard (bottom), which takes only prevalence data into account. The influence of the experimental factors on the composition of the microbiota was tested with the multivariate analysis of variance ANOVA. The numbers displayed in the titles are the cumulative variances explained by all significant axes and the stars denote the significance of the factor: $p<0.001$ ***, $p<0.01$ **, $p<0.05$ *.

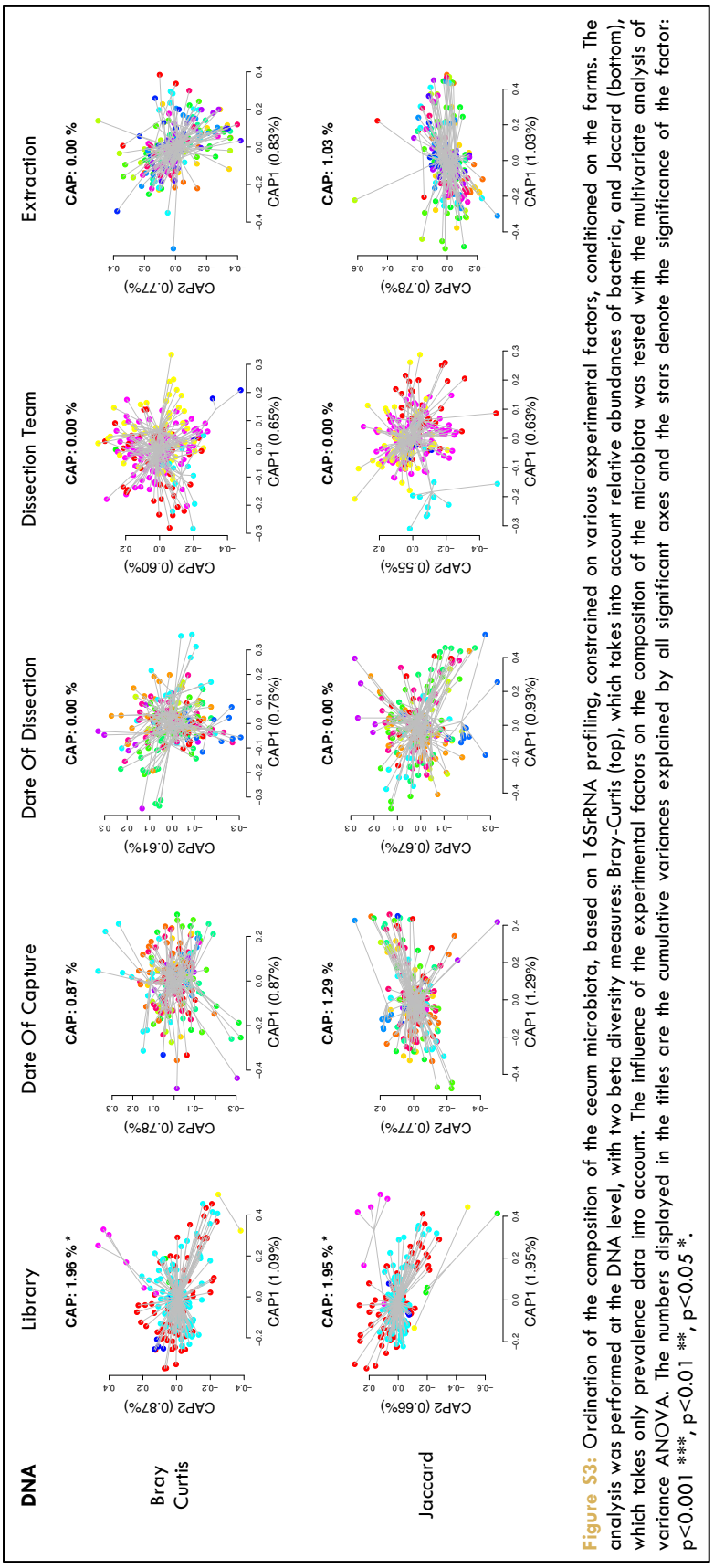

Figure S3: Ordination of the cecum microbiota, based on 16SrRNA profiling, constrained on various experimental factors, conditioned on the farms. The analysis was performed at the DNA level, with two beta diversity measures: Bray-Curtis (top), which takes into account relative abundances of bacteria, and Jaccard (bottom), which takes only prevalence data into account. The influence of the experimental factors on the composition of the microbiota was tested with the multivariate analysis of variance ANOVA. The numbers displayed in the titles are the cumulative variances explained by all significant axes and the stars denote the significance of the factor: $p<0.001$ ***, $p<0.01$ **, $p<0.05$ *.

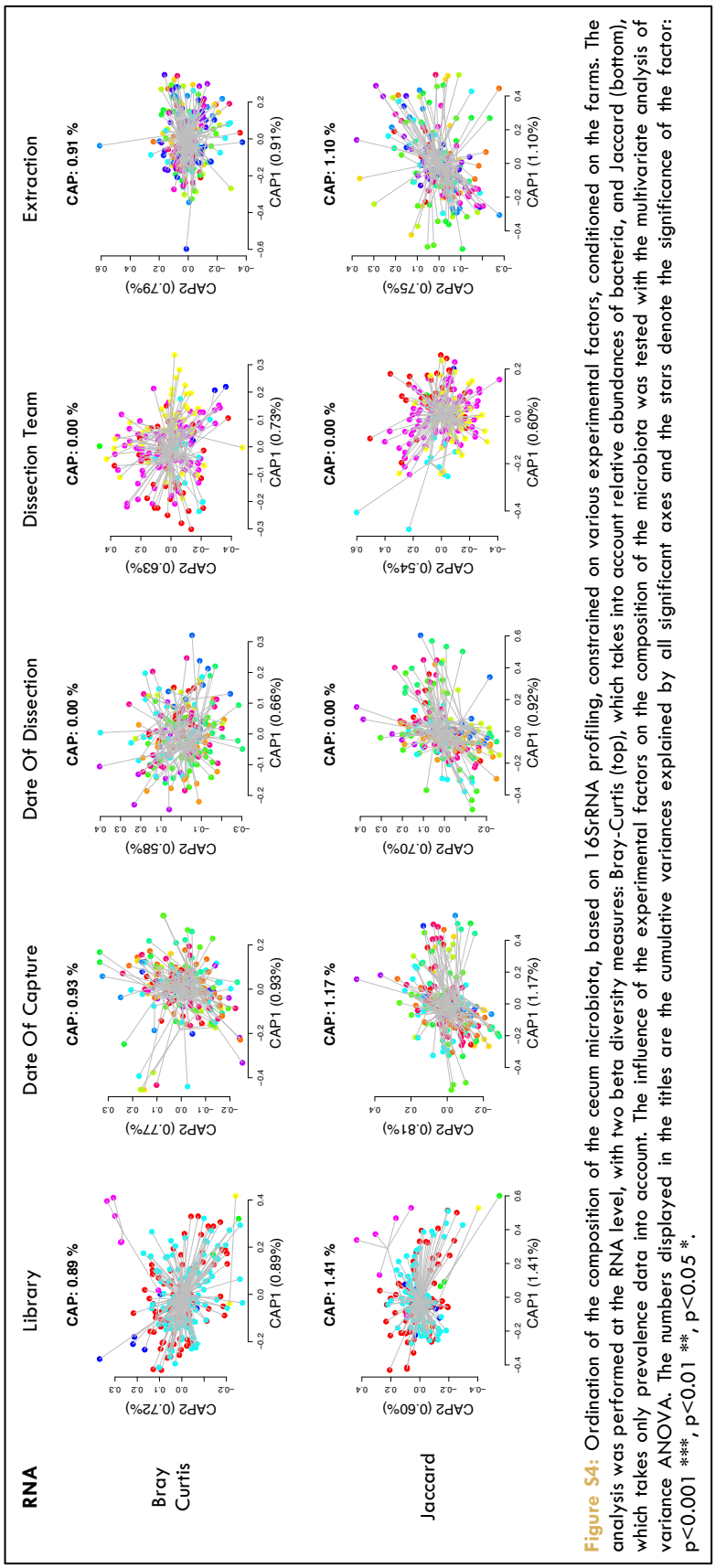

Figure S4: Ordination of the composition of the cecum microbiota, based on 16SrRNA profiling, constrained on various experimental factors, conditioned on the farms. The analysis was performed at the RNA level, with two beta diversity measures: Bray-Curtis (top), which takes into account relative abundances of bacteria, and Jaccard (bottom), which takes only prevalence data into account. The influence of the experimental factors on the composition of the microbiota was tested with the multivariate analysis of variance ANOVA. The numbers displayed in the titles are the cumulative variances explained by all significant axes and the stars denote the significance of the factor: $p<0.001$ ***, $p<0.01$ **, $p<0.05$ *.

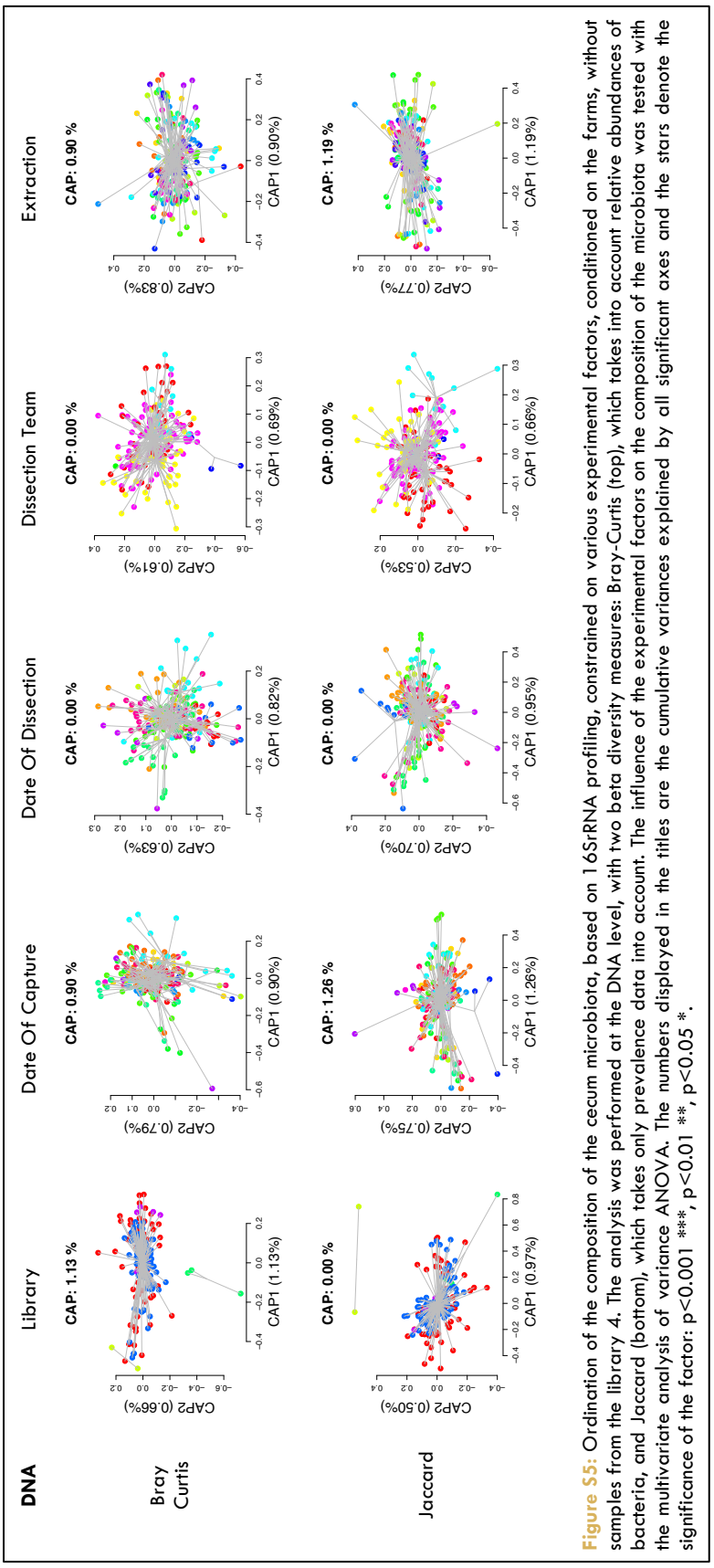

Figure S5: Ordination of the composition of the cecum microbiota, based on 16SrRNA profiling, constrained on various experimental factors, conditioned on the farms, without samples from the library 4. The analysis was performed at the DNA level, with two beta diversity measures: Bray-Curtis (top), which takes into account relative abundances of bacteria, and Jaccard (bottom), which takes only prevalence data into account. The influence of the experimental factors on the composition of the microbiota was tested with the multivariate analysis of variance ANOVA. The numbers displayed in the titles are the cumulative variances explained by all significant axes and the stars denote the significance of the factor: p<0.001 ***, p<0.01 **, p<0.05 *.

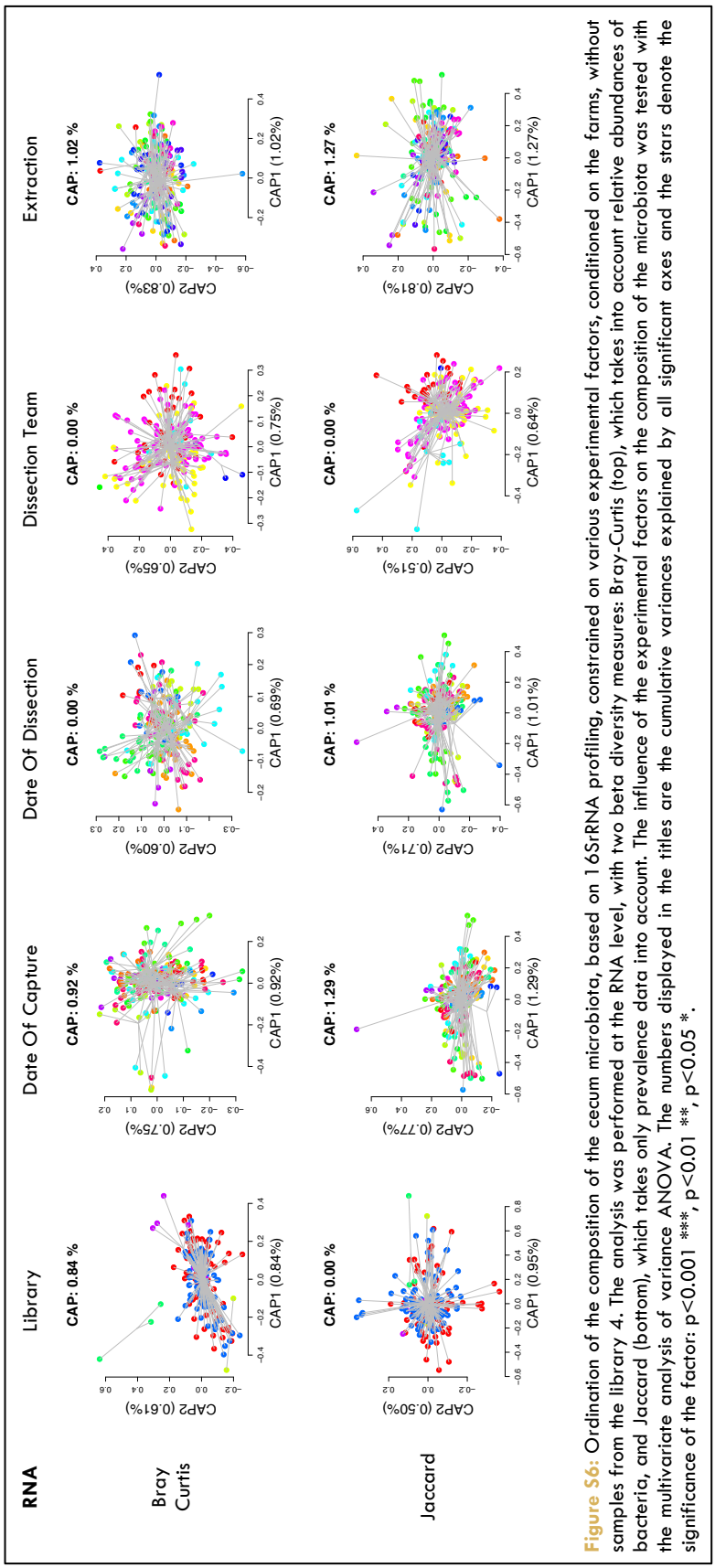

Figure S6: Ordination of the composition of the cecum microbiota, based on 16SrRNA profiling, constrained on various experimental factors, conditioned on the farms, without samples from the library 4. The analysis was performed at the RNA level, with two beta diversity measures: Bray-Curtis (top), which takes into account relative abundances of bacteria, and Jaccard (bottom), which takes only prevalence data into account. The influence of the experimental factors on the composition of the microbiota was tested with the multivariate analysis of variance ANOVA. The numbers displayed in the titles are the cumulative variances explained by all significant axes and the stars denote the significance of the factor: p<0.001 ***, p<0.01 **, p<0.05 *.

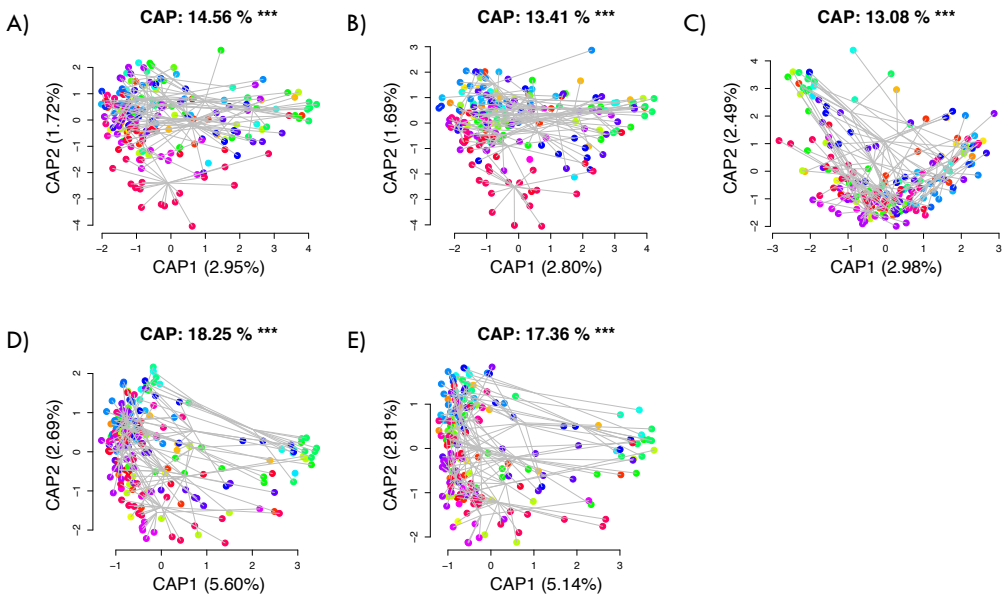

Figure S7: Ordination of the composition of the cecum microbiota, based on 16SrRNA profiling, constrained on the farms. The analysis was performed at the DNA (A, D) and RNA (B, E) level, as well as with the activity (C), defined as the ratio between RNA and DNA, normalized by sample. Two beta diversity measures were used: Bray-Curtis (A, B, C), which takes into account relative abundances of bacteria, and Jaccard (D, E), which takes only prevalence data into account. The influence of the farms on the composition of the microbiota was tested with the multivariate analysis of variance ANOVA. The numbers displayed in the titles are the cumulative variances explained by all significant axes and the stars denote the significance of the factor: $p<0.001$ ***, $p<0.01$ **, $p<0.05$ *.

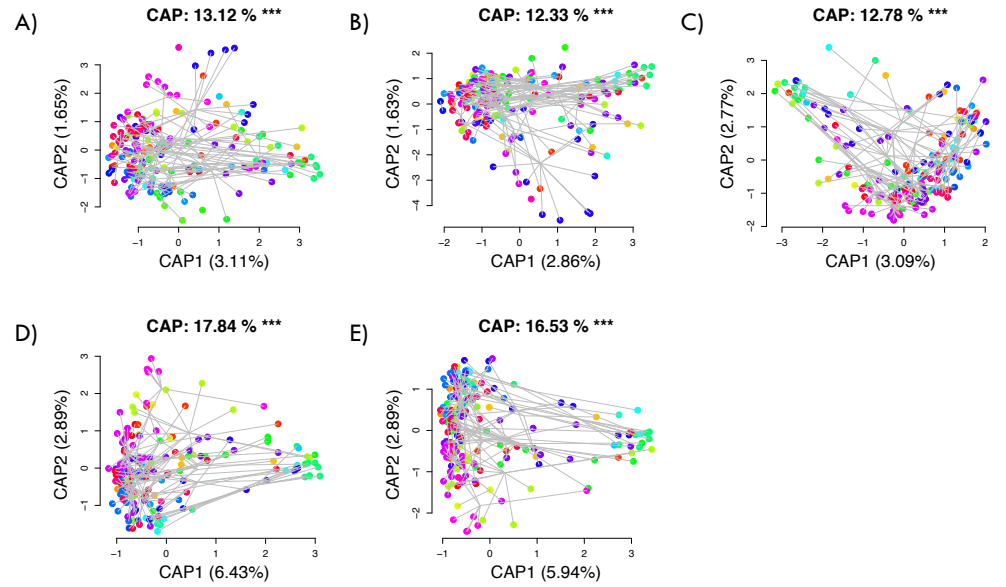

Figure S8: Ordination of the composition of the cecum microbiota, based on 16SrRNA profiling, constrained on the farms, without samples from the farm MT21. The analysis was performed at the DNA (A, D) and RNA (B, E) level, as well as with the activity (C), defined as the ratio between RNA and DNA, normalized by sample. Two beta diversity measures were used: Bray-Curtis (A, B, C), which takes into account relative abundances of bacteria, and Jaccard (D, E), which takes only prevalence data into account. The influence of the farms on the composition of the microbiota was tested with the multivariate analysis of variance ANOVA. The numbers displayed in the titles are the cumulative variances explained by all significant axes and the stars denote the significance of the factor: $p<0.001$ ***, $p<0.01$ **, $p<0.05$ *.

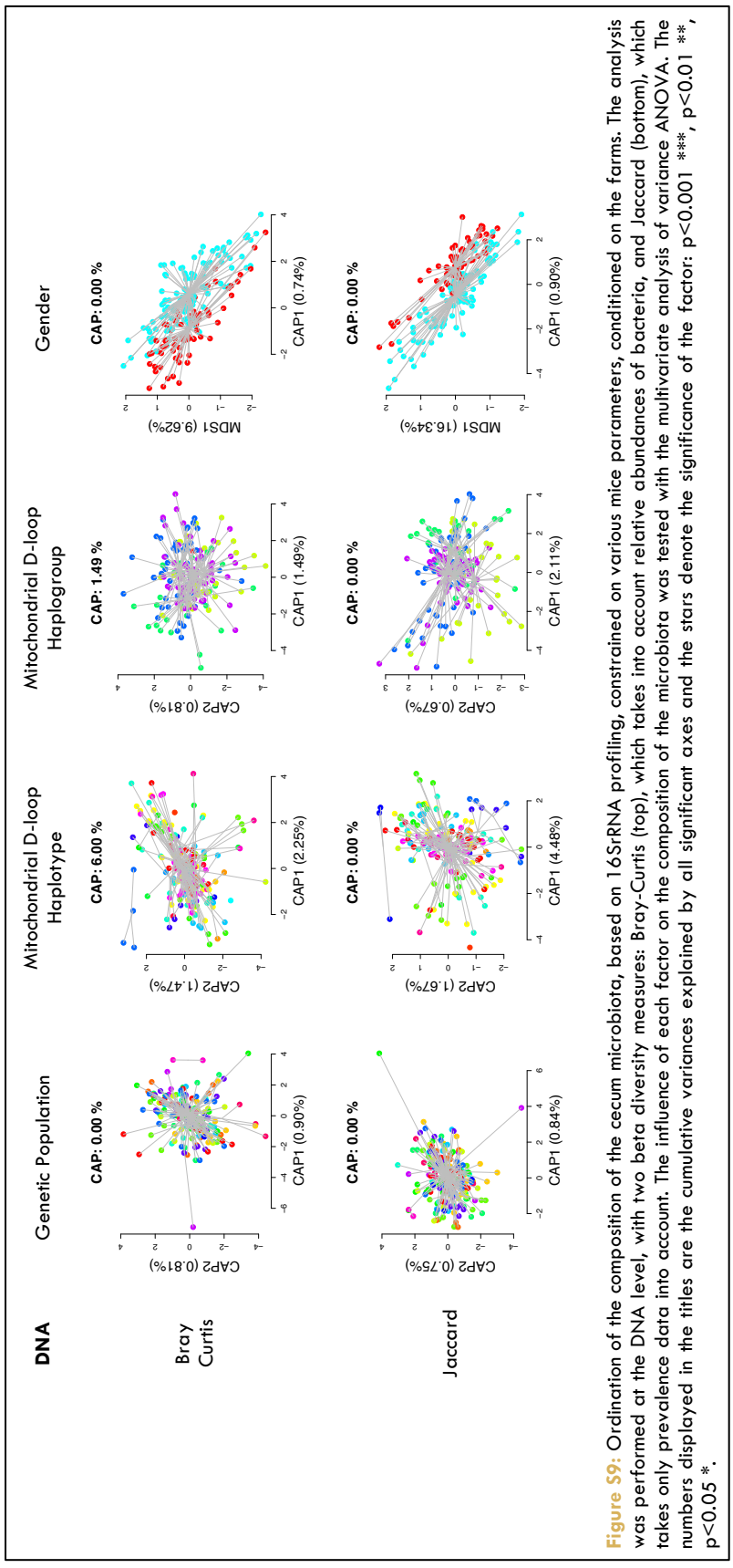

Figure S9: Ordination of the composition of the cecum microbiota, based on 16SrRNA profiling, constrained on various mice parameters, conditioned on the farms. The analysis was performed at the DNA level, with two beta diversity measures: Bray-Curtis (top), which takes into account relative abundances of bacteria, and Jaccard (bottom), which takes only prevalence data into account. The influence of each factor on the composition of the microbiota was tested with the multivariate analysis of variance ANOVA. The numbers displayed in the titles are the cumulative variances explained by all significant axes and the stars denote the significance of the factor: $p<0.001$ ***, $p<0.01$ **, $p<0.05$ *.

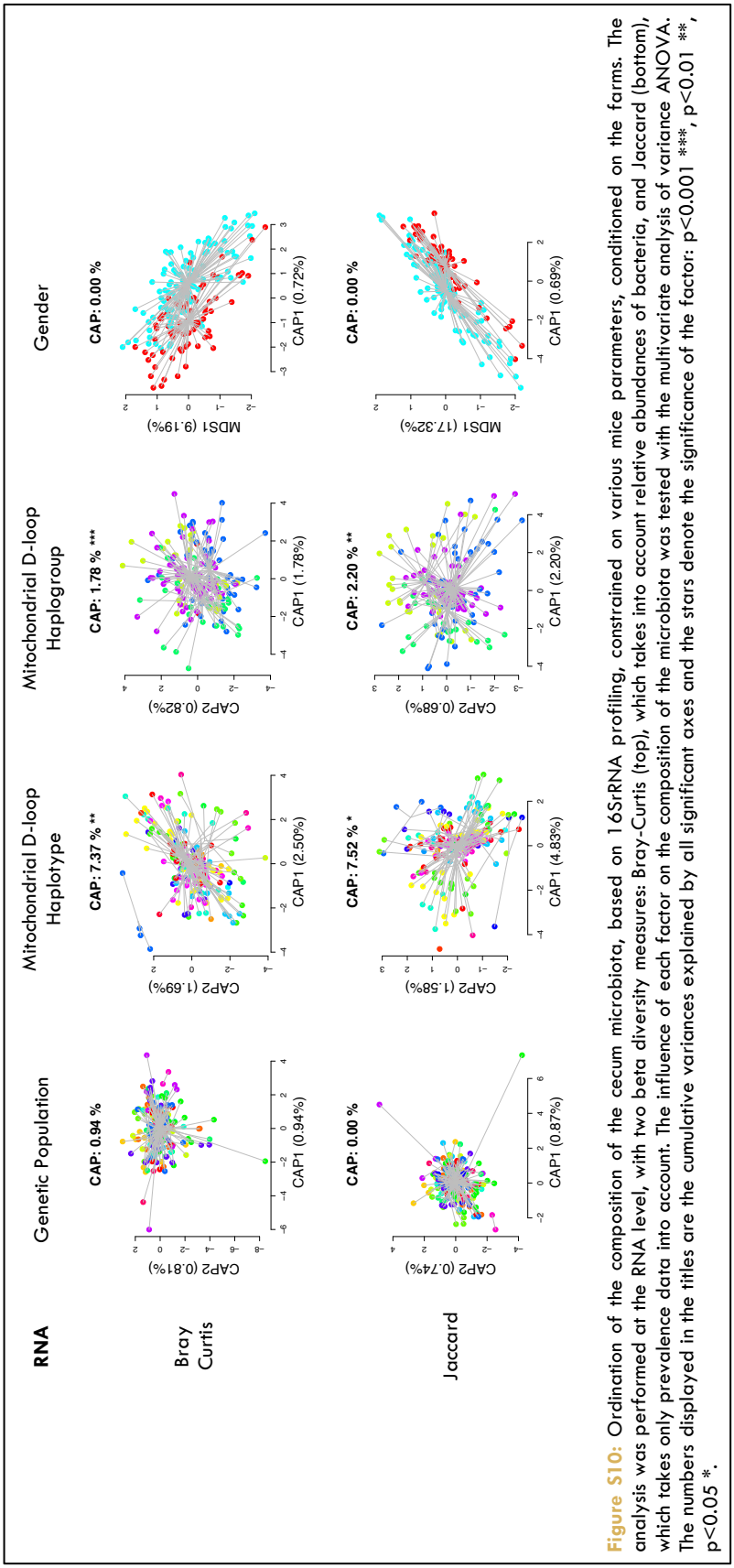

Figure S10: Ordination of the composition of the cecum microbiota, based on 16SrRNA profiling, constrained on various mice parameters, conditioned on the farms. The analysis was performed at the RNA level, with two beta diversity measures: Bray-Curtis (top), which takes into account relative abundances of bacteria, and Jaccard (bottom), which takes only prevalence data into account. The influence of each factor on the composition of the microbiota was tested with the multivariate analysis of variance ANOVA. The numbers displayed in the titles are the cumulative variances explained by all significant axes and the stars denote the significance of the factor: $p<0.001$ ***, $p<0.01$ **, $p<0.05$ *.

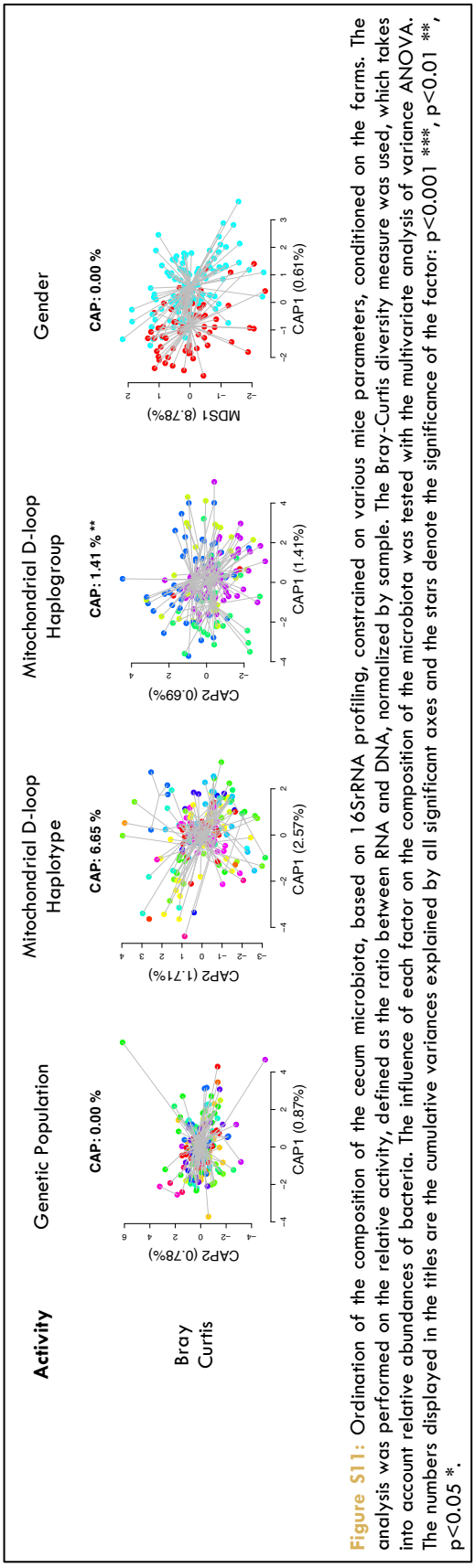

Figure S11: Ordination of the composition of the cecum microbiota, based on 16SrRNA profiling, constrained on various mice parameters, conditioned on the farms. The analysis was performed on the relative activity, defined as the ratio between RNA and DNA, normalized by sample. The Bray-Curtis diversity measure was used, which takes into account relative abundances of bacteria. The influence of each factor on the composition of the microbiota was tested with the multivariate analysis of variance ANOVA. The numbers displayed in the titles are the cumulative variances explained by all significant axes and the stars denote the significance of the factor: $p<0.001$ ***, $p<0.01$ **, $p<0.05$ *.

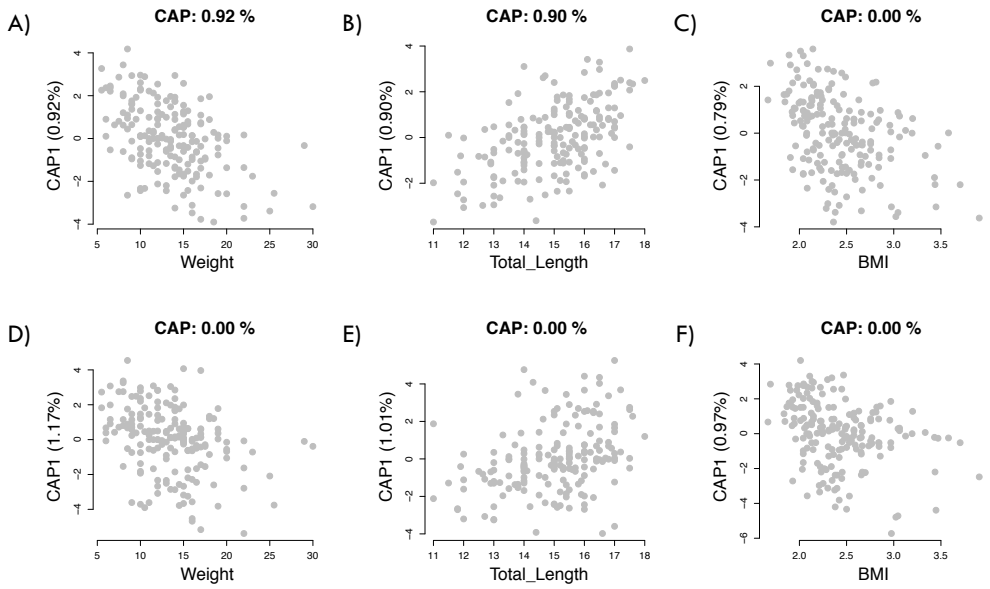

Figure S12: Ordination of the composition of the cecum microbiota, based on 16SrRNA profiling, constrained on various mice parameters, conditioned on the farms. The first axis of the ordination is plotted against the mice parameters. The analysis was performed at the DNA level, with two beta diversity measures: Bray-Curtis (A, B, C), which takes into account relative abundances of bacteria, and Jaccard (D, E, F), which takes only prevalence data into account. The influence of each factor on the composition of the microbiota was tested with the multivariate analysis of variance ANOVA. The numbers displayed in the titles are the cumulative variances explained by all significant axes and the stars denote the significance of the factor: p<0.001 ***, p<0.01 **, p<0.05 *.

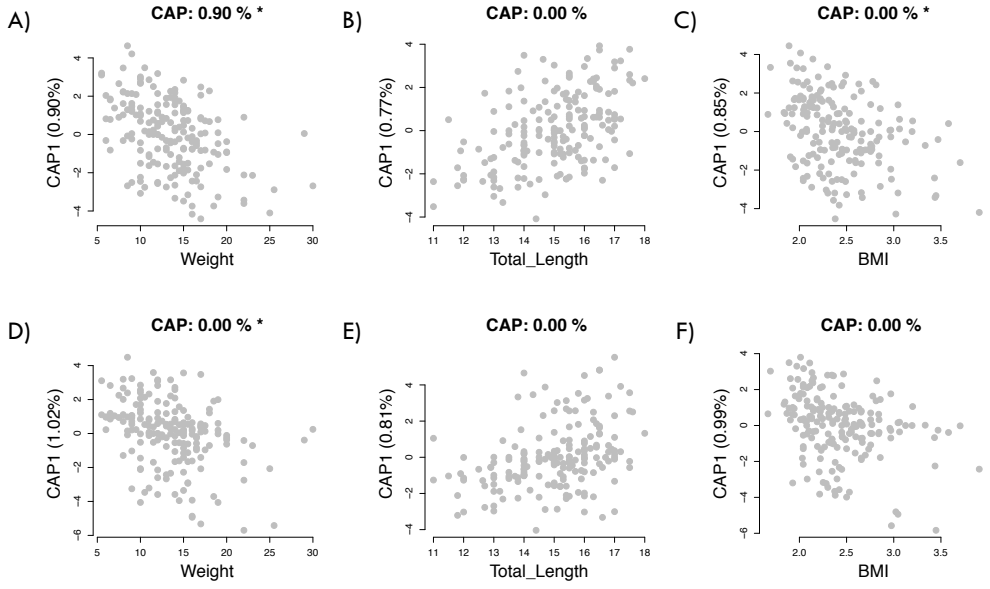

Figure S13: Ordination of the composition of the cecum microbiota, based on 16SrRNA profiling, constrained on various mice parameters, conditioned on the farms. The first axis of the ordination is plotted against the mice parameters. The analysis was performed at the RNA level, with two beta diversity measures: Bray-Curtis (A, B, C), which takes into account relative abundances of bacteria, and Jaccard (D, E, F), which takes only prevalence data into account. The influence of each factor on the composition of the microbiota was tested with the multivariate analysis of variance ANOVA. The numbers displayed in the titles are the cumulative variances explained by all significant axes and the stars denote the significance of the factor: p<0.001 ***, p<0.01 **, p<0.05 *.

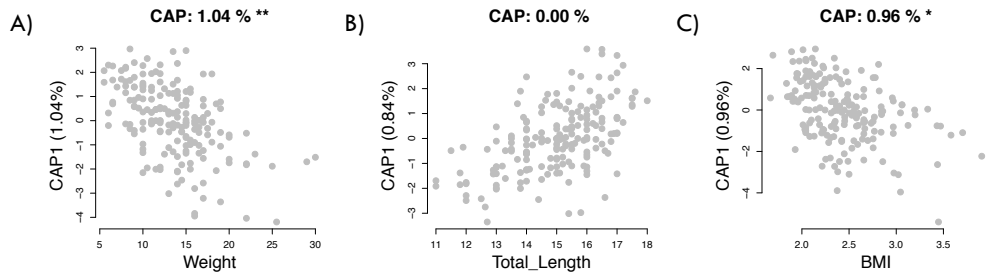

Figure S14: Ordination of the composition of the cecum microbiota, based on 16SrRNA profiling, constrained on various mice parameters, conditioned on the farms. The first axis of the ordination is plotted against the mice parameters. The analysis was performed on the relative activity, defined as the ratio between RNA and DNA, normalized by sample. The Bray-Curtis diversity measure was used, which takes into account relative abundances of bacteria. The influence of each factor on the composition of the microbiota was tested with the multivariate analysis of variance ANOVA. The numbers displayed in the titles are the cumulative variances explained by all significant axes and the stars denote the significance of the factor: p<0.001 ***, p<0.01 **, p<0.05 *.

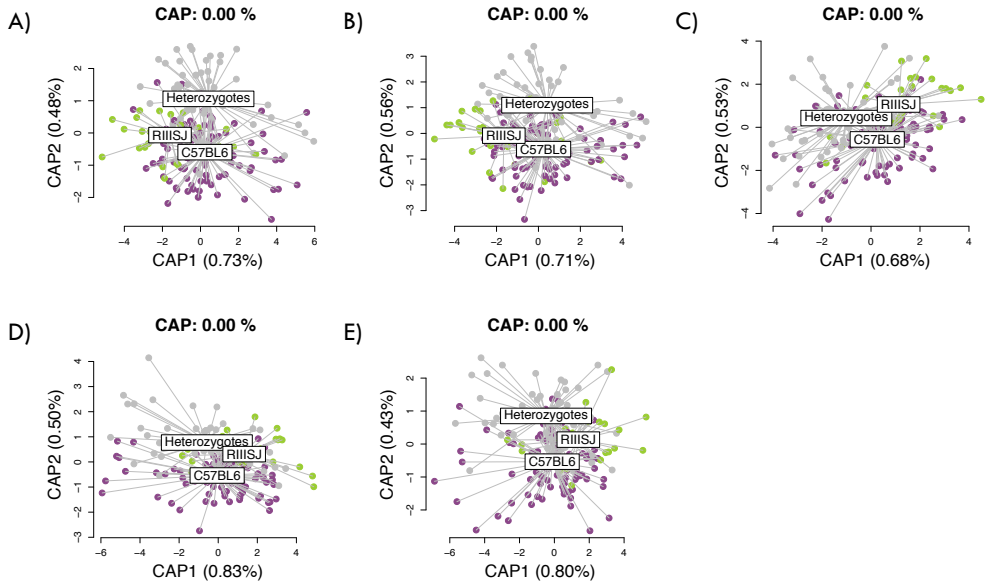

Figure S15: Ordination of the composition of the cecum microbiota, based on 16SrRNA profiling, constrained on *B4galnt2* genotype, conditioned on the farms. The analysis was performed at the DNA (A, D) and the RNA (B, E) level, as well as on the activity (C), defined as the ratio between RNA and DNA, normalized by sample. Two beta diversity measures were used: Bray-Curtis (A, B, C), which takes into account relative abundances of bacteria, and Jaccard (D, E), which takes only prevalence data into account. The influence of *B4galnt2* genotype on the composition of the microbiota was tested with the multivariate analysis of variance ANOVA. The numbers displayed in the titles are the cumulative variances explained by all significant axes and the stars denote the significance of the factor: p<0.001 ***, p<0.01 **, p<0.05 *.

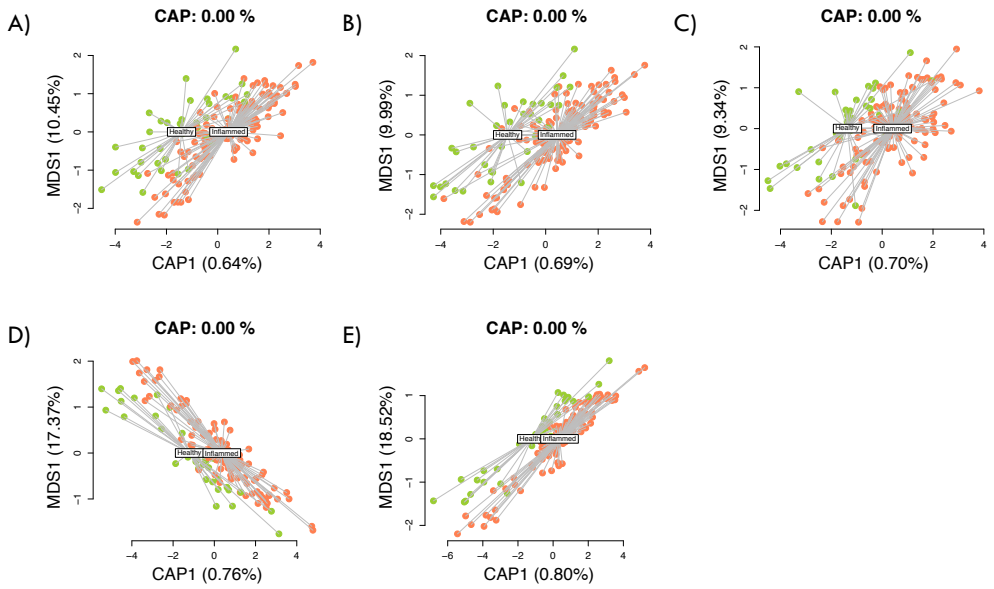

Figure S16: Ordination of the composition of the cecum microbiota, based on 16SrRNA profiling, constrained on cecum inflammation, taken as categorical variable, conditioned on the farms. The analysis was performed at the DNA (A, D) and the RNA (B, E) level, as well as on the activity (C), defined as the ratio between RNA and DNA, normalized by sample. Two beta diversity measures were used: Bray-Curtis (A, B, C), which takes into account relative abundances of bacteria, and Jaccard (D, E), which takes only prevalence data into account. The influence of the cecum inflammation on the composition of the microbiota was tested with the multivariate analysis of variance ANOVA. The numbers displayed in the titles are the cumulative variances explained by all significant axes and the stars denote the significance of the factor: p<0.001 ***, p<0.01 **, p<0.05 *.

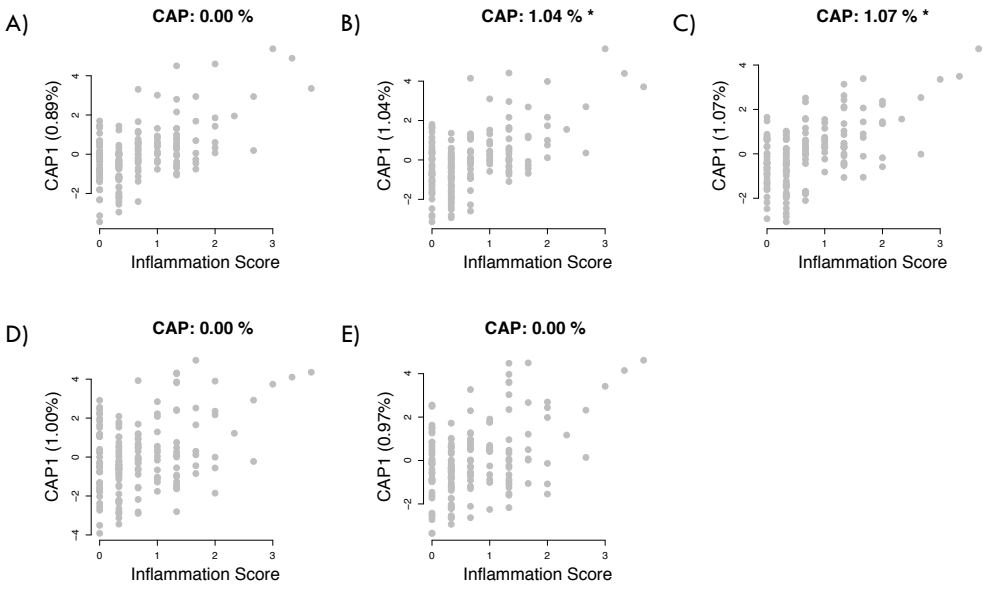

Figure S17: Ordination of the composition of the cecum microbiota, based on 16SrRNA profiling, constrained on cecum inflammation, taken as quantitative variable, conditioned on the farms. The analysis was performed at the DNA (A, D) and the RNA (B, E) level, as well as on the activity (C), defined as the ratio between RNA and DNA, normalized by sample. Two beta diversity measures were used: Bray-Curtis (A, B, C), which takes into account relative abundances of bacteria, and Jaccard (D, E), which takes only prevalence data into account. The influence of the cecum inflammation on the composition of the microbiota was tested with the multivariate analysis of variance ANOVA. The numbers displayed in the titles are the cumulative variances explained by all significant axes and the stars denote the significance of the factor: p<0.001 ***, p<0.01 **, p<0.05 *.

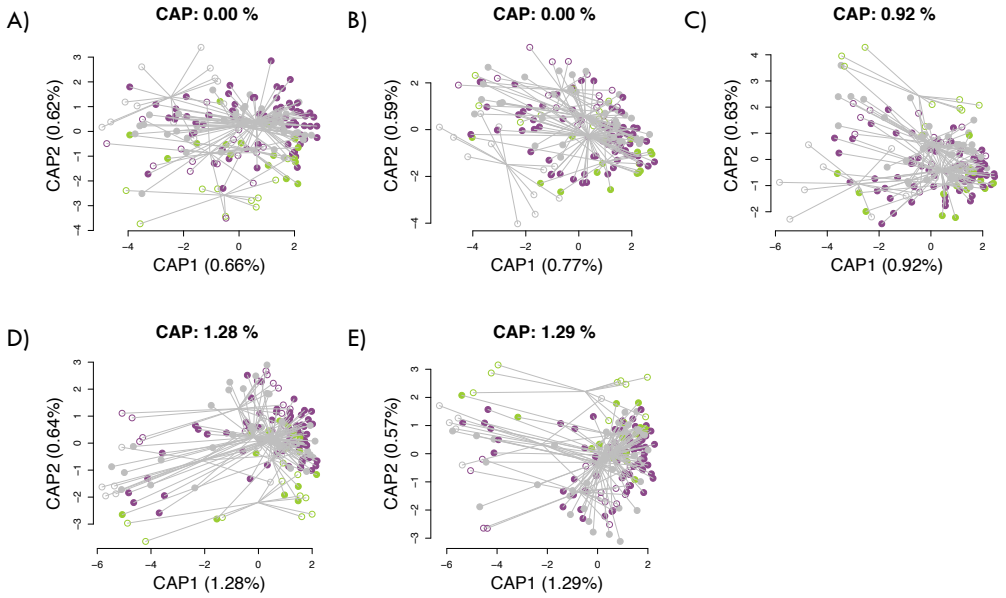

Figure S18: Ordination of the composition of the cecum microbiota, based on 16SrRNA profiling, constrained on the interaction between *B4galnt2* genotype and cecum inflammation, taken as categorical variable, conditioned on the farms. The color code refers to *B4galnt2* genotype: C57BL6/J in purple, RIIIS/J in green and heterozygotes in grey. The symbol refers to the inflammation: full circle for inflamed and empty circles for healthy. The analysis was performed at the DNA (A, D) and the RNA (B, E) level, as well as on the activity (C), defined as the ratio between RNA and DNA, normalized by sample. Two beta diversity measures were used: Bray-Curtis (A, B, C), which takes into account relative abundances of bacteria, and Jaccard (D, E), which takes only prevalence data into account. The influence of the interaction between *B4galnt2* genotype and cecum inflammation on the composition of the microbiota was tested with the multivariate analysis of variance ANOVA. The numbers displayed in the titles are the cumulative variances explained by all significant axes and the stars denote the significance of the factor: $p<0.001$ ***, $p<0.01$ **, $p<0.05$ *.

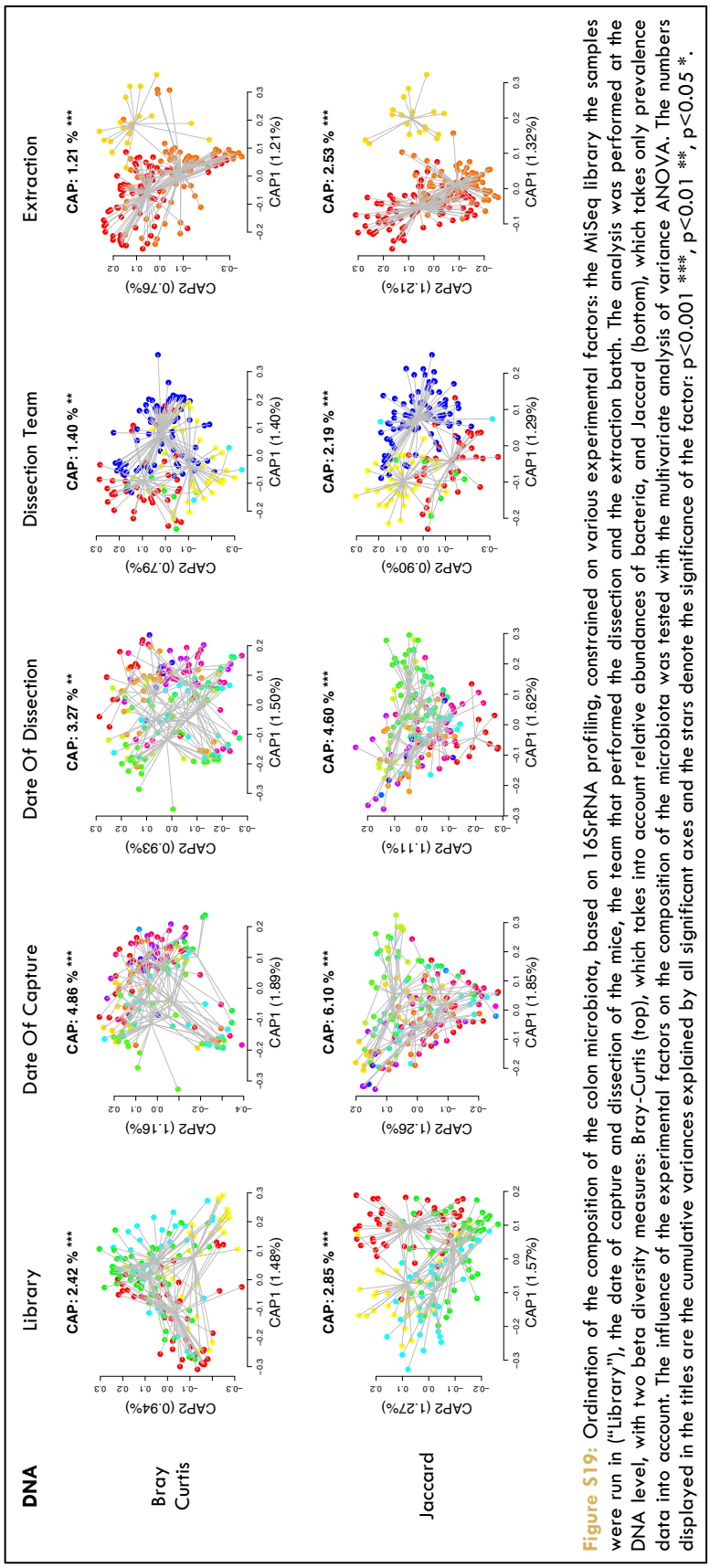

Figure S19: Ordination of the composition of the colon microbiota, based on 16SrRNA profiling, constrained on various experimental factors: the MiSeq library the samples were run in ("Library"), the date of capture and dissection of the mice, the team that performed the dissection and the extraction batch. The analysis was performed at the DNA level, with two beta diversity measures: Bray-Curtis (top), which takes into account relative abundances of bacteria, and Jaccard (bottom), which takes only prevalence data into account. The influence of the experimental factors on the composition of the microbiota was tested with the multivariate analysis of variance ANOVA. The numbers displayed in the titles are the cumulative variances explained by all significant axes and the stars denote the significance of the factor: $p<0.001$ ***, $p<0.01$ **, $p<0.05$ *.

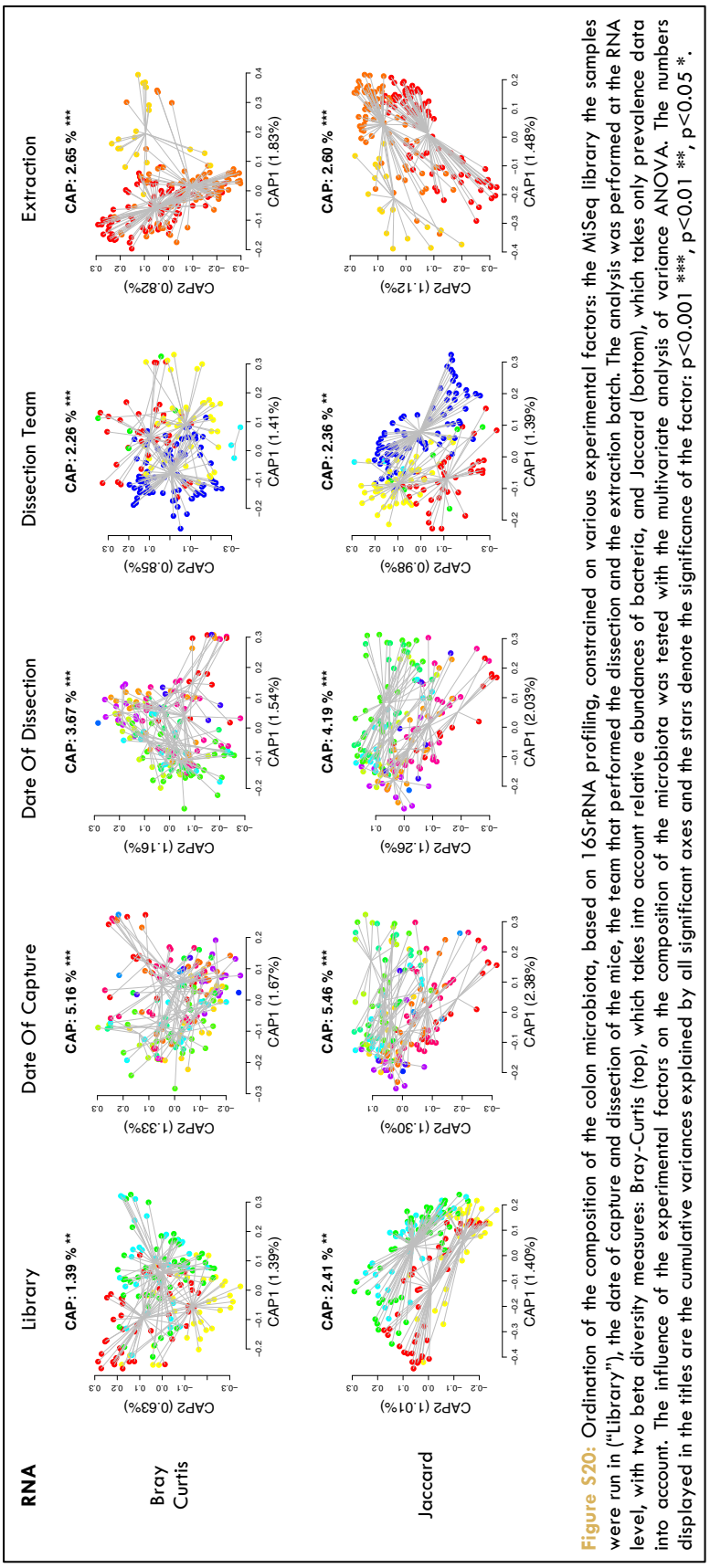

Figure S20: Ordination of the composition of the colon microbiota, based on 16SrRNA profiling, constrained on various experimental factors: the MiSeq library the samples were run in ("Library"), the date of capture and dissection of the mice, the team that performed the dissection and the extraction batch. The analysis was performed at the RNA level, with two beta diversity measures: Bray-Curtis (top), which takes into account relative abundances of bacteria, and Jaccard (bottom), which takes only prevalence data into account. The influence of the experimental factors on the composition of the microbiota was tested with the multivariate analysis of variance ANOVA. The numbers displayed in the titles are the cumulative variances explained by all significant axes and the stars denote the significance of the factor: $p<0.001$ ***, $p<0.01$ **, $p<0.05$ *.

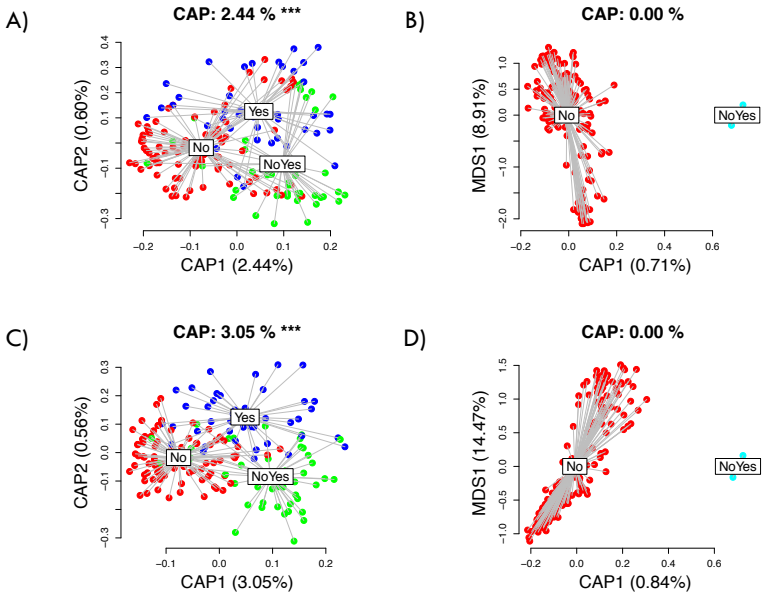

Figure S21: Ordination of the composition of the colon microbiota, based on 16SrRNA profiling, constrained on whether the 400bp band was gel extracted prior to the library run or not. The analysis was performed at the DNA (A, C) and RNA (B, D) level, with two beta diversity measures: Bray-Curtis (A, B), which takes into account relative abundances of bacteria, and Jaccard (C, D), which takes only prevalence data into account. In the DNA, the "NoYes" samples refer to samples that where repeated and where gel extracted in one of the repeat but not in the other. The influence of the extraction on the composition of the microbiota was tested with the multivariate analysis of variance ANOVA. The numbers displayed in the titles are the cumulative variances explained by all significant axes and the stars denote the significance of the factor: $p<0.001$ ***, $p<0.01$ **, $p<0.05$ *.

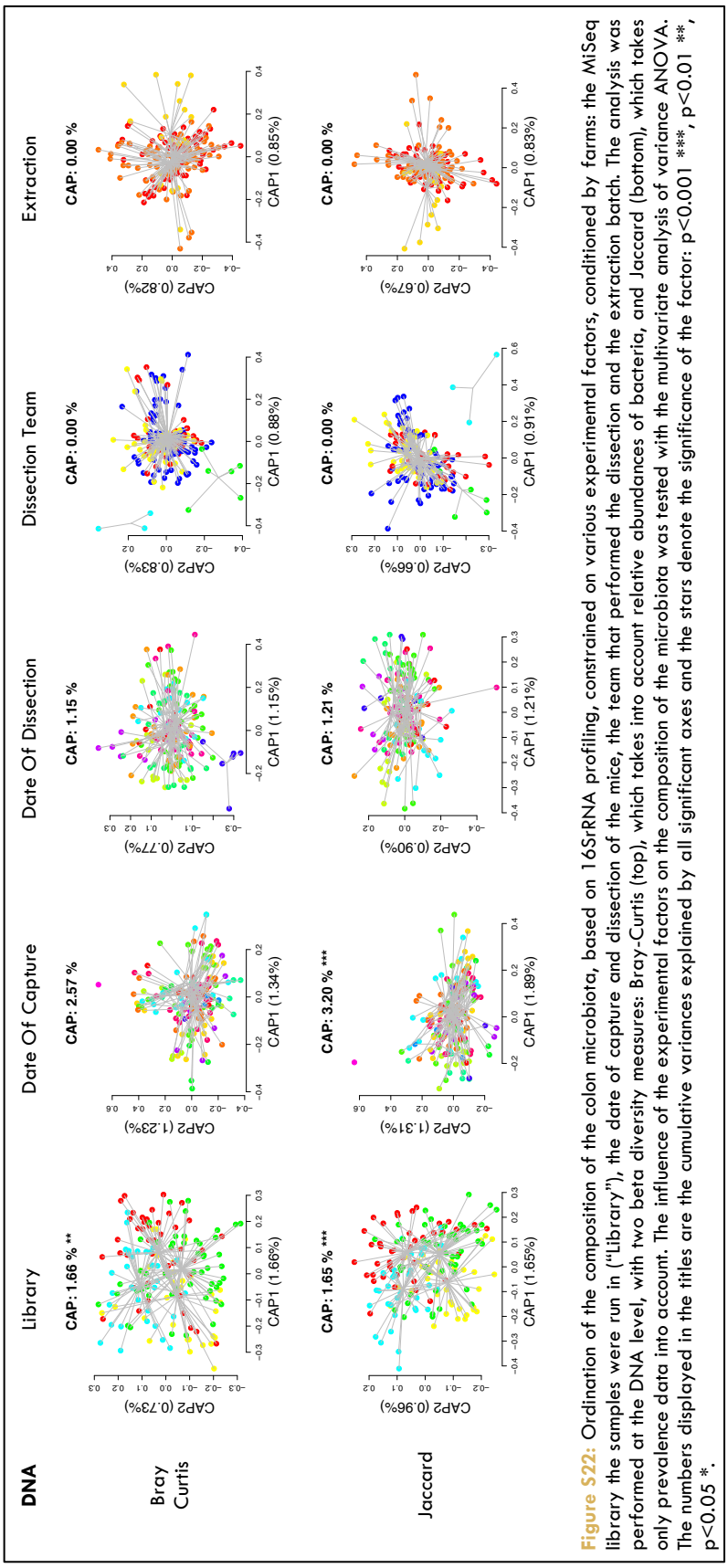

Figure S22: Ordination of the composition of the colon microbiota, based on 16SrRNA profiling, constrained on various experimental factors, conditioned by farms: the MiSeq library the samples were run in ("Library"), the date of capture and dissection of the mice, the team that performed the dissection and the extraction batch. The analysis was performed at the DNA level, with two beta diversity measures: Bray-Curtis (top), which takes into account relative abundances of bacteria, and Jaccard (bottom), which takes only prevalence data into account. The influence of the experimental factors on the composition of the microbiota was tested with the multivariate analysis of variance ANOVA. The numbers displayed in the titles are the cumulative variances explained by all significant axes and the stars denote the significance of the factor: $p<0.001$ ***, $p<0.01$ **, $p<0.05$ *.

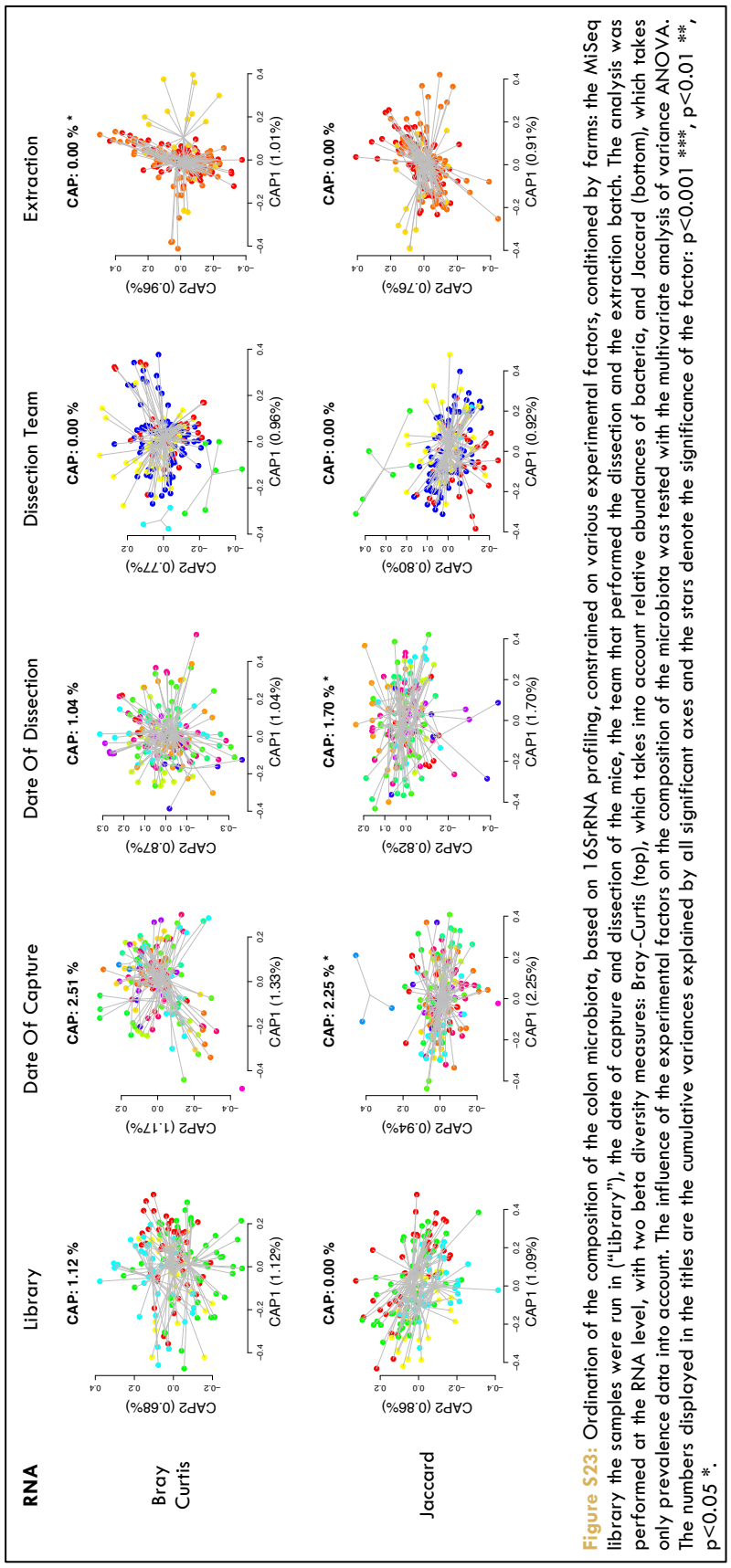

Figure S23: Ordination of the composition of the colon microbiota, based on 16SrRNA profiling, constrained on various experimental factors, conditioned by farms: the MiSeq library the samples were run in ("Library"), the date of capture and dissection of the mice, the team that performed the dissection and the extraction batch. The analysis was performed at the RNA level, with two beta diversity measures: Bray-Curtis (top), which takes into account relative abundances of bacteria, and Jaccard (bottom), which takes only prevalence data into account. The influence of the experimental factors on the composition of the microbiota was tested with the multivariate analysis of variance ANOVA. The numbers displayed in the titles are the cumulative variances explained by all significant axes and the stars denote the significance of the factor: $p<0.001$ ***, $p<0.01$ **, $p<0.05$ *.

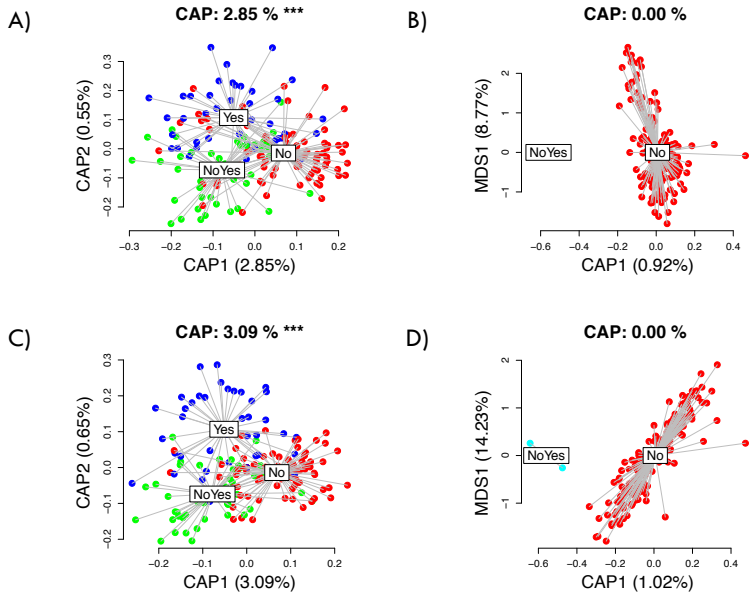

Figure S24: Ordination of the composition of the colon microbiota, based on 16SrRNA profiling, constrained on whether the 400bp band was gel extracted prior to the library run or not, conditioned on the farms. The analysis was performed at the DNA (A, C) and RNA (B, D) level, with two beta diversity measures: Bray-Curtis (A, B), which takes into account relative abundances of bacteria, and Jaccard (C, D), which takes only prevalence data into account. In the DNA, the "NoYes" samples refer to samples that where repeated and where gel extracted in one of the repeat but not in the other. The influence of the extraction on the composition of the microbiota was tested with the multivariate analysis of variance ANOVA. The numbers displayed in the titles are the cumulative variances explained by all significant axes and the stars denote the significance of the factor: p<0.001 ***, p<0.01 **, p<0.05 *.

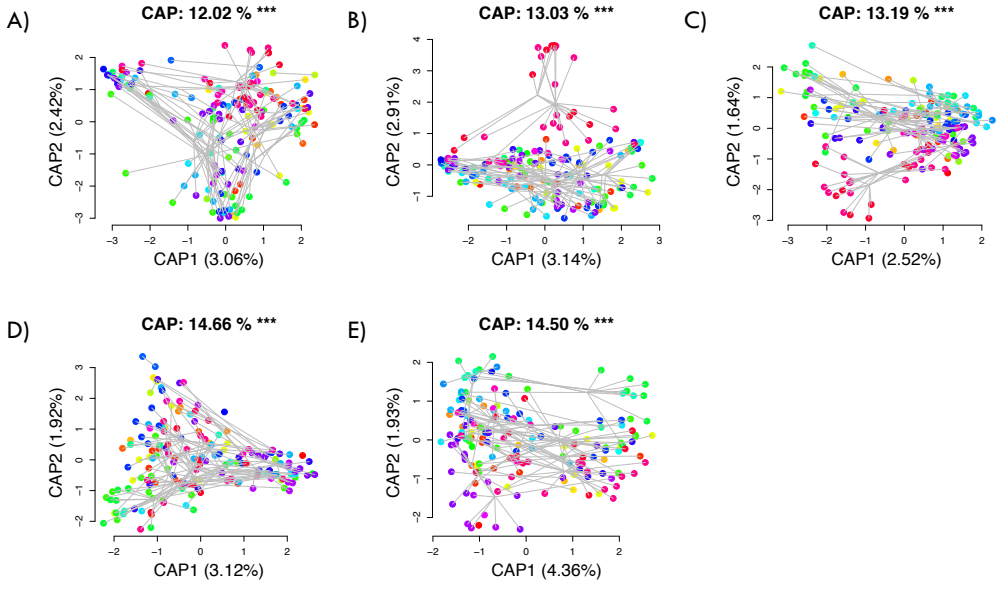

Figure S25: Ordination of the composition of the colon microbiota, based on 16SrRNA profiling, constrained on the farms. The analysis was performed at the DNA (A, D) and RNA (B, E) level, as well as with the activity (C), defined as the ratio between RNA and DNA, normalized by sample. Two beta diversity measures were used: Bray-Curtis (A, B, C), which takes into account relative abundances of bacteria, and Jaccard (D, E), which takes only prevalence data into account. The influence of the farms on the composition of the microbiota was tested with the multivariate analysis of variance ANOVA. The numbers displayed in the titles are the cumulative variances explained by all significant axes and the stars denote the significance of the factor: p<0.001 ***, p<0.01 **, p<0.05 *.

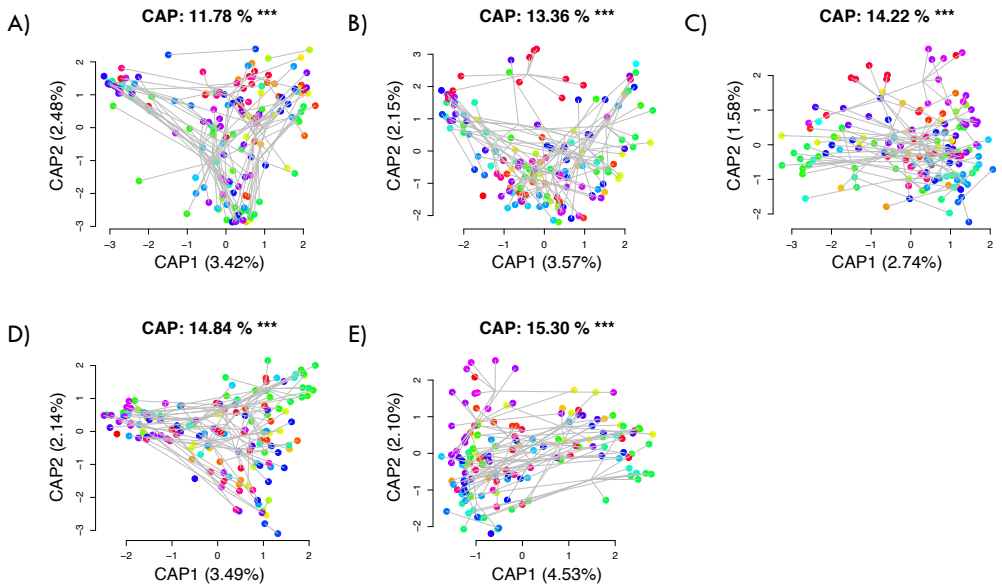

Figure S26: Ordination of the composition of the colon microbiota, based on 16SrRNA profiling, constrained on the farms, without the mice from the MT21 farm. The analysis was performed at the DNA (A, D) and RNA (B, E) level, as well as with the activity (C), defined as the ratio between RNA and DNA, normalized by sample. Two beta diversity measures were used: Bray-Curtis (A, B, C), which takes into account relative abundances of bacteria, and Jaccard (D, E), which takes only prevalence data into account. The influence of the farms on the composition of the microbiota was tested with the multivariate analysis of variance ANOVA. The numbers displayed in the titles are the cumulative variances explained by all significant axes and the stars denote the significance of the factor: $p<0.001$ ***, $p<0.01$ **, $p<0.05$ *.

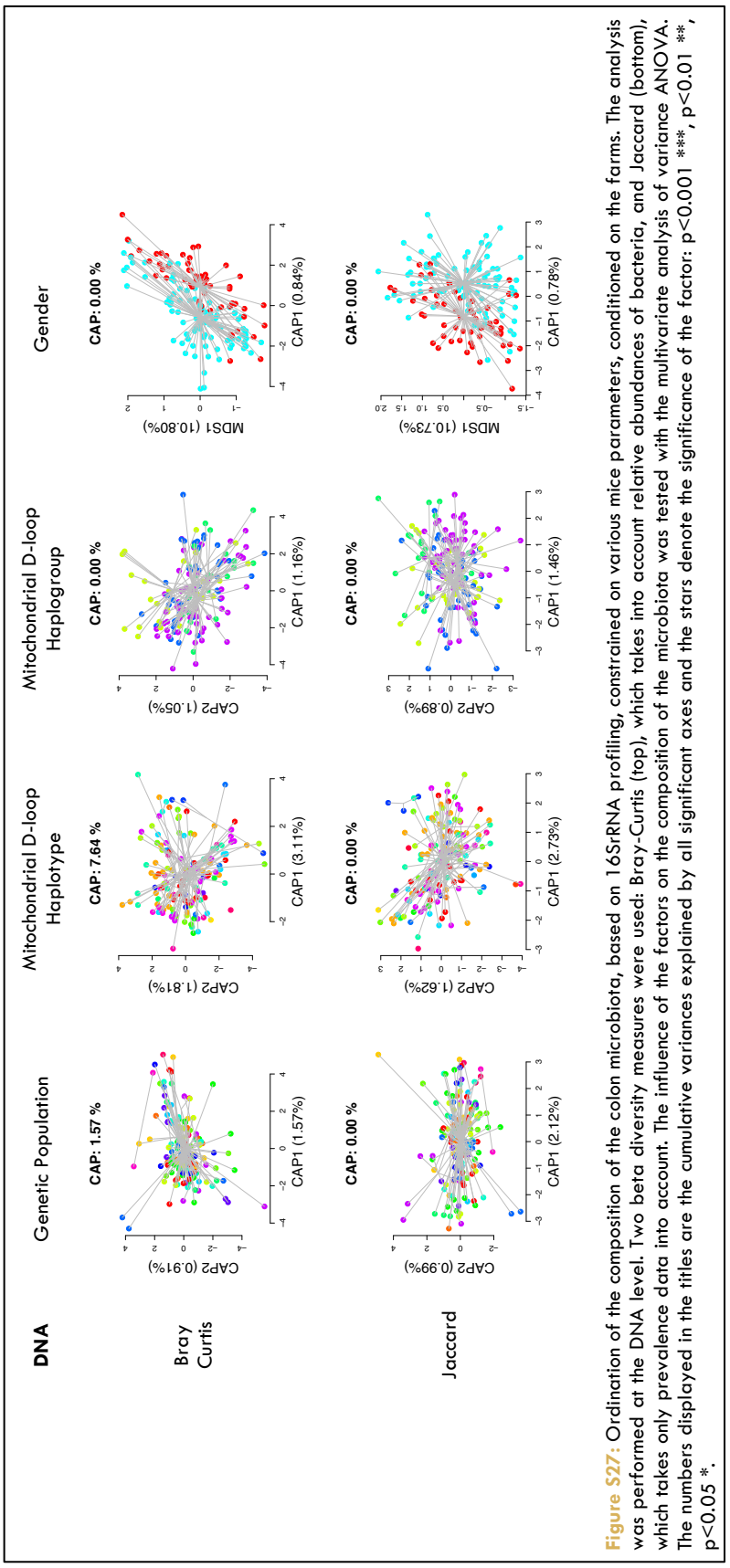

Figure S27: Ordination of the composition of the colon microbiota, based on 16SrRNA profiling, constrained on various mice parameters, conditioned on the farms. The analysis was performed at the DNA level. Two beta diversity measures were used: Bray-Curtis (top), which takes into account relative abundances of bacteria, and Jaccard (bottom), which takes only prevalence data into account. The influence of the factors on the composition of the microbiota was tested with the multivariate analysis of variance ANOVA. The numbers displayed in the titles are the cumulative variances explained by all significant axes and the stars denote the significance of the factor: $p<0.001$ ***, $p<0.01$ **, $p<0.05$ *.

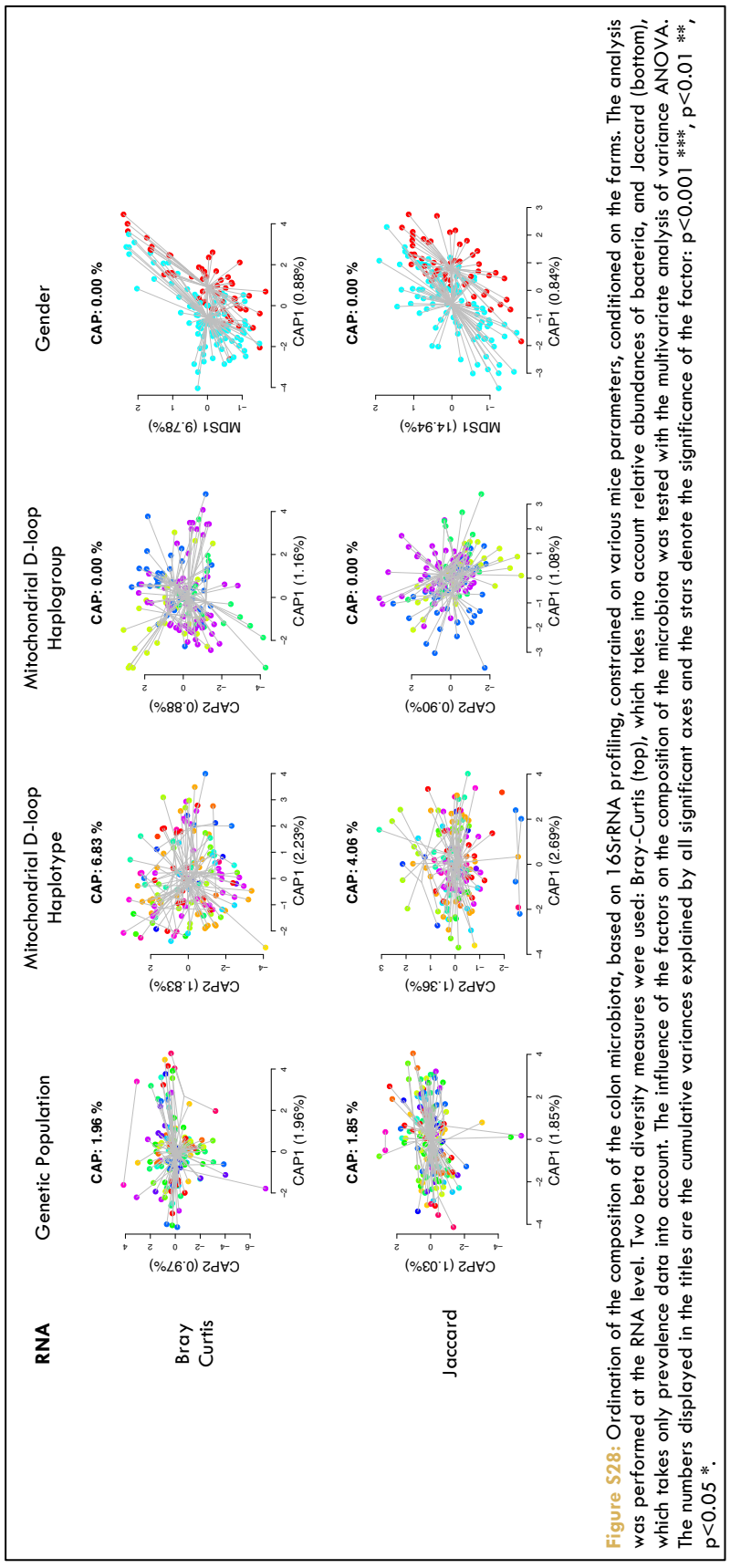

Figure S28: Ordination of the composition of the colon microbiota, based on 16SrRNA profiling, constrained on various mice parameters, conditioned on the farms. The analysis was performed at the RNA level. Two beta diversity measures were used: Bray-Curtis (top), which takes into account relative abundances of bacteria, and Jaccard (bottom), which takes only prevalence data into account. The influence of the factors on the composition of the microbiota was tested with the multivariate analysis of variance ANOVA. The numbers displayed in the titles are the cumulative variances explained by all significant axes and the stars denote the significance of the factor: $p<0.001$ ***, $p<0.01$ **, $p<0.05$ *.

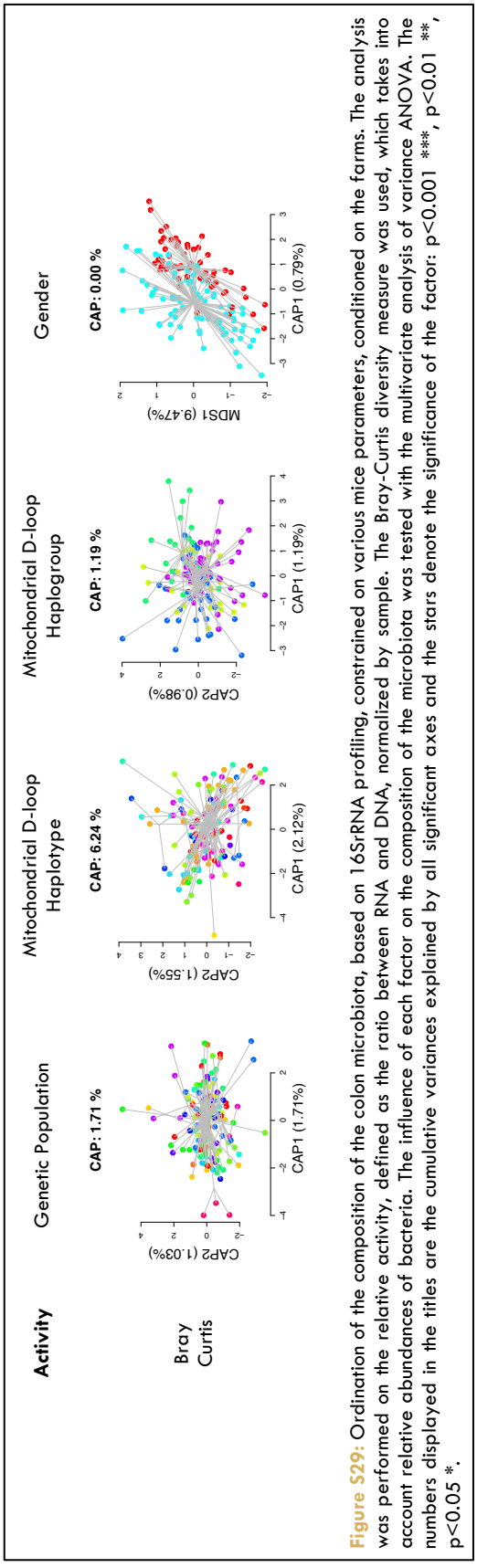

Figure S29: Ordination of the composition of the colon microbiota, based on 16SrRNA profiling, constrained on various mice parameters, conditioned on the farms. The analysis was performed on the relative activity, defined as the ratio between RNA and DNA, normalized by sample. The Bray-Curtis diversity measure was used, which takes into account relative abundances of bacteria. The influence of each factor on the composition of the microbiota was tested with the multivariate analysis of variance ANOVA. The numbers displayed in the titles are the cumulative variances explained by all significant axes and the stars denote the significance of the factor: $p<0.001$ ***, $p<0.01$ **, $p<0.05$ *.

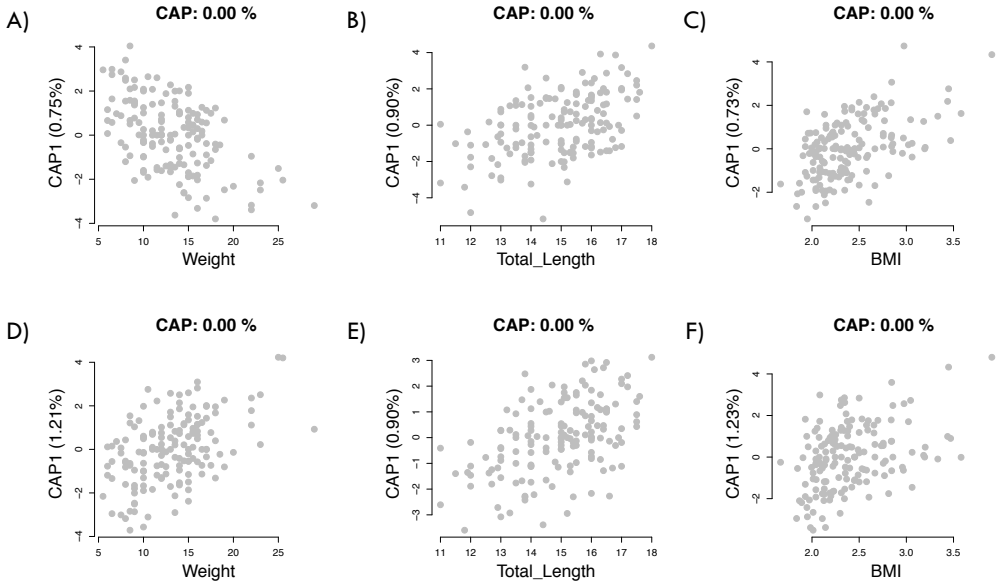

Figure S30: Ordination of the composition of the colon microbiota, based on 16SrRNA profiling, constrained on various mice parameters, conditioned on the farms. The first axis of the ordination is plotted against the mice parameters. The analysis was performed at the DNA level, with two beta diversity measures: Bray-Curtis (A, B, C), which takes into account relative abundances of bacteria, and Jaccard (D, E, F), which takes only prevalence data into account. The influence of each factor on the composition of the microbiota was tested with the multivariate analysis of variance ANOVA. The numbers displayed in the titles are the cumulative variances explained by all significant axes and the stars denote the significance of the factor: p<0.001 ***, p<0.01 **, p<0.05 *.

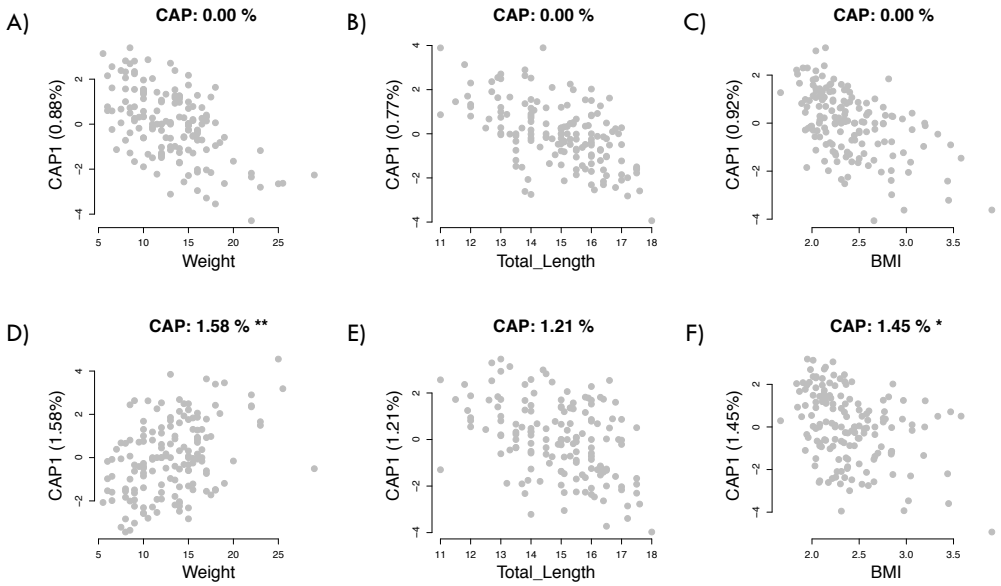

Figure S31: Ordination of the composition of the colon microbiota, based on 16SrRNA profiling, constrained on various mice parameters, conditioned on the farms. The first axis of the ordination is plotted against the mice parameters. The analysis was performed at the RNA level, with two beta diversity measures: Bray-Curtis (A, B, C), which takes into account relative abundances of bacteria, and Jaccard (D, E, F), which takes only prevalence data into account. The influence of each factor on the composition of the microbiota was tested with the multivariate analysis of variance ANOVA. The numbers displayed in the titles are the cumulative variances explained by all significant axes and the stars denote the significance of the factor: p<0.001 ***, p<0.01 **, p<0.05 *.

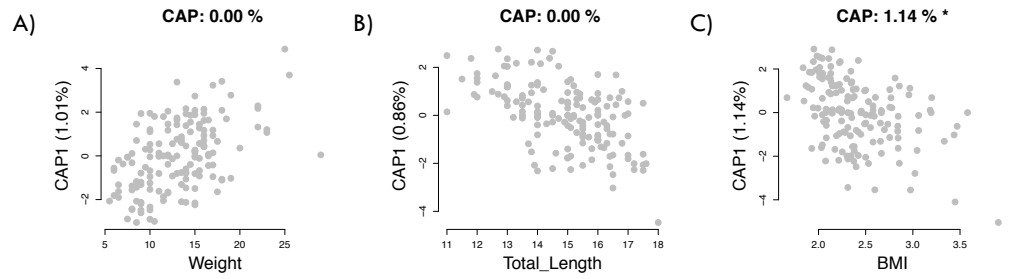

Figure S32: Ordination of the composition of the colon microbiota, based on 16SrRNA profiling, constrained on various mice parameters, conditioned on the farms. The first axis of the ordination is plotted against the mice parameters. The analysis was performed on the relative activity, defined as the ratio between RNA and DNA, normalized by sample. The Bray-Curtis diversity measure was used, which takes into account relative abundances of bacteria. The influence of each factor on the composition of the microbiota was tested with the multivariate analysis of variance ANOVA. The numbers displayed in the titles are the cumulative variances explained by all significant axes and the stars denote the significance of the factor: p<0.001 ***, p<0.01 **, p<0.05 *.

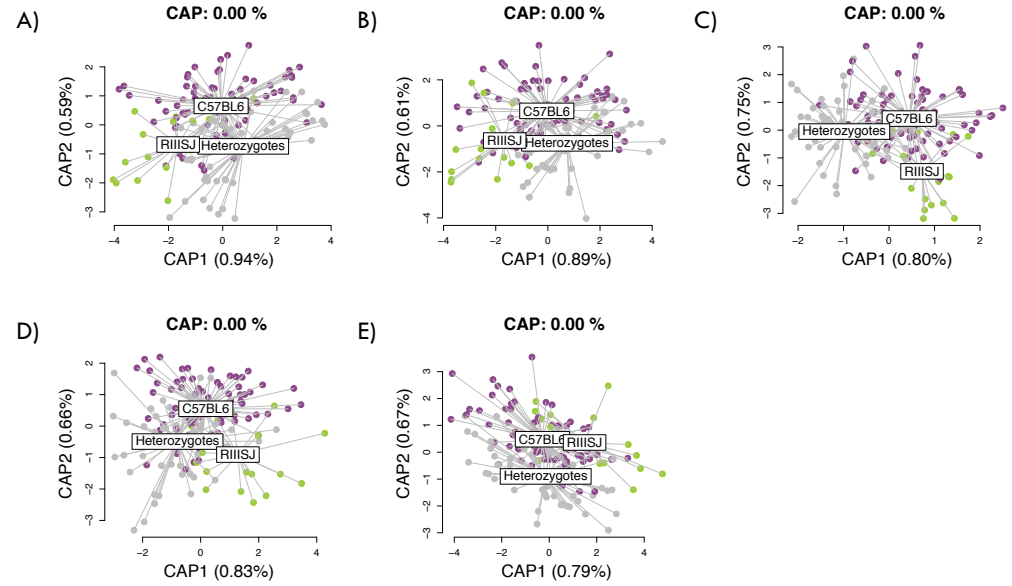

Figure S33: Ordination of the composition of the colon microbiota, based on 16SrRNA profiling, constrained on *B4galnt2* genotype, conditioned on the farms. The analysis was performed at the DNA (A, D) and the RNA (B, E) level, as well as on the activity (C), defined as the ratio between RNA and DNA, normalized by sample. Two beta diversity measures were used: Bray-Curtis (A, B, C), which takes into account relative abundances of bacteria, and Jaccard (D, E), which takes only prevalence data into account. The influence of *B4galnt2* genotype on the composition of the microbiota was tested with the multivariate analysis of variance ANOVA. The numbers displayed in the titles are the cumulative variances explained by all significant axes and the stars denote the significance of the factor: p<0.001 ***, p<0.01 **, p<0.05 *.

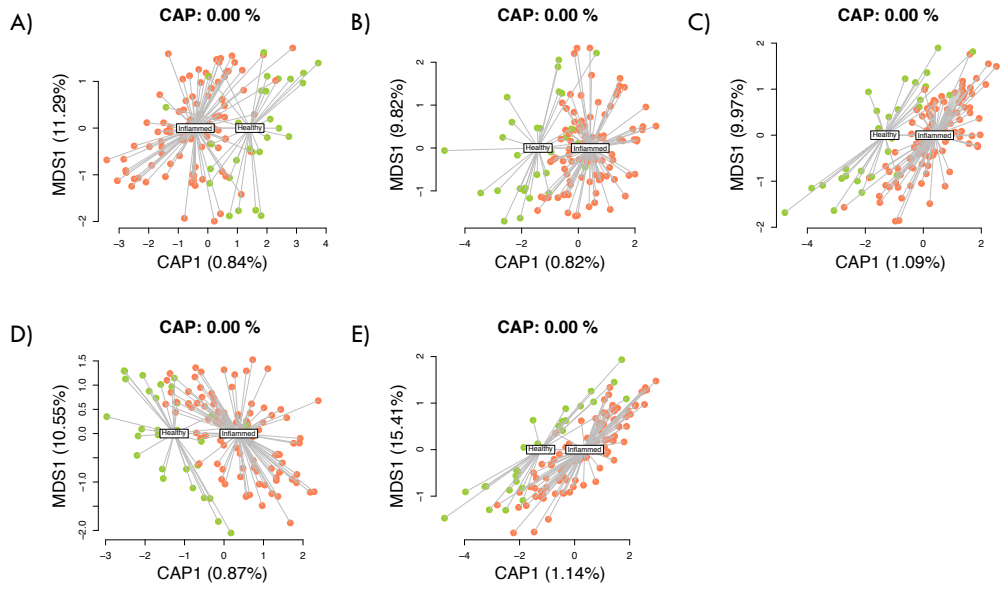

Figure S34: Ordination of the composition of the colon microbiota, based on 16SrRNA profiling, constrained on cecum inflammation, taken as categorical variable, conditioned on the farms. The analysis was performed at the DNA (A, D) and the RNA (B, E) level, as well as on the activity (C), defined as the ratio between RNA and DNA, normalized by sample. Two beta diversity measures were used: Bray-Curtis (A, B, C), which takes into account relative abundances of bacteria, and Jaccard (D, E), which takes only prevalence data into account. The influence of the cecum inflammation on the composition of the microbiota was tested with the multivariate analysis of variance ANOVA. The numbers displayed in the titles are the cumulative variances explained by all significant axes and the stars denote the significance of the factor: p<0.001 ***, p<0.01 **, p<0.05 *.

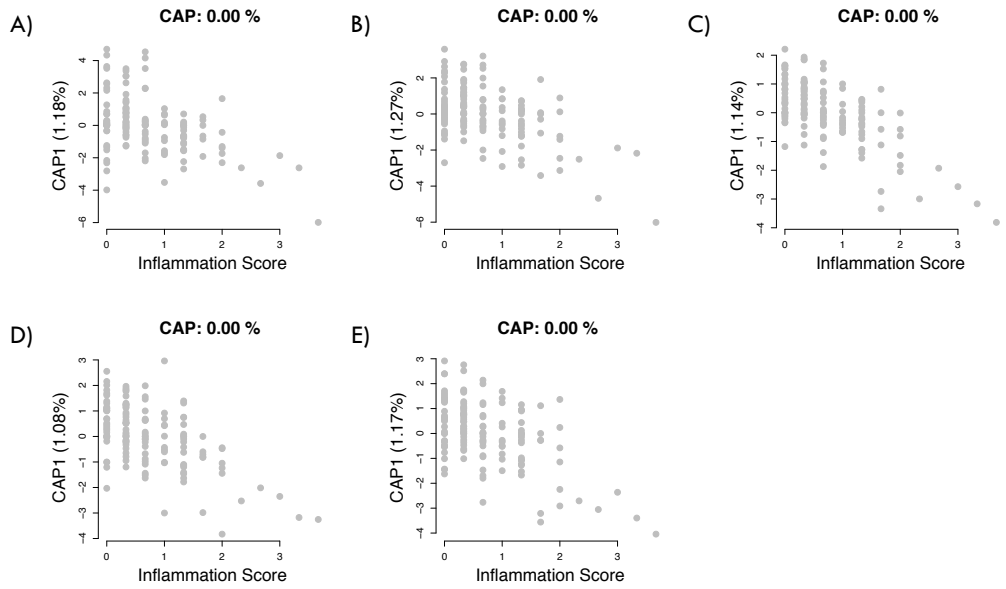

Figure S35: Ordination of the composition of the colon microbiota, based on 16SrRNA profiling, constrained on cecum inflammation, taken as quantitative variable, conditioned on the farms. The analysis was performed at the DNA (A, D) and the RNA (B, E) level, as well as on the activity (C), defined as the ratio between RNA and DNA, normalized by sample. Two beta diversity measures were used: Bray-Curtis (A, B, C), which takes into account relative abundances of bacteria, and Jaccard (D, E), which takes only prevalence data into account. The influence of the cecum inflammation on the composition of the microbiota was tested with the multivariate analysis of variance ANOVA. The numbers displayed in the titles are the cumulative variances explained by all significant axes and the stars denote the significance of the factor: p<0.001 ***, p<0.01 **, p<0.05 *.

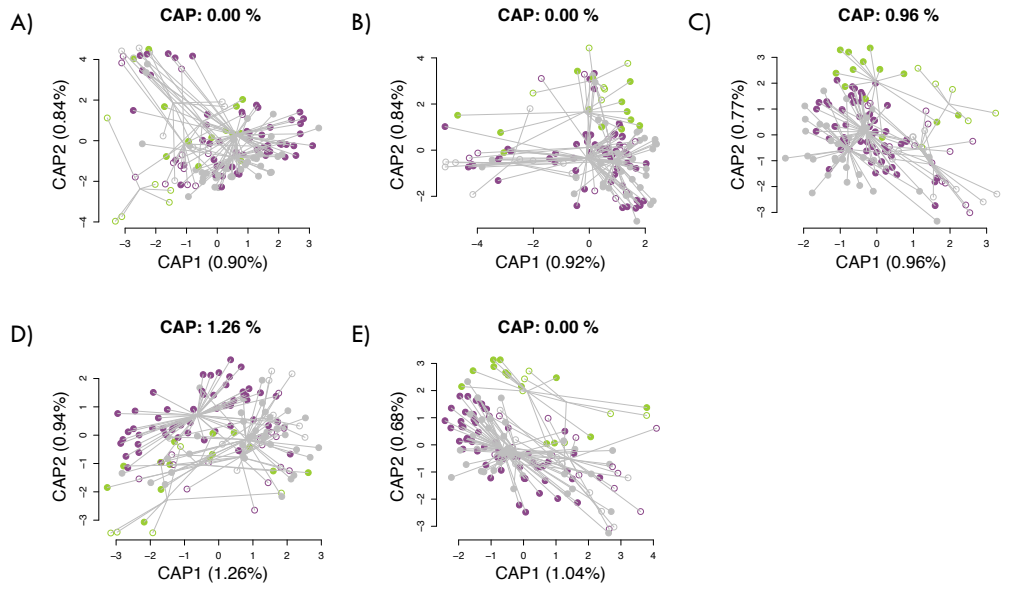

Figure S36: Ordination of the composition of the colon microbiota, based on 16SrRNA profiling, constrained on the interaction between B4galnt2 genotype and cecum inflammation, taken as categorical variable, conditioned on the farms. The colour code refers to B4galnt2 genotype: C57BL6/J in purple, RIIIS/J in green and heterozygotes in grey. The symbol refers to the inflammation: full circle for inflamed and empty circles for healthy. The analysis was performed at the DNA (A, D) and the RNA (B, E) level, as well as on the activity (C), defined as the ratio between RNA and DNA, normalized by sample. Two beta diversity measures were used: Bray-Curtis (A, B, C), which takes into account relative abundances of bacteria, and Jaccard (D, E), which takes only prevalence data into account. The influence of the interaction between *B4galnt2* genotype and cecum inflammation on the composition of the microbiota was tested with the multivariate analysis of variance ANOVA. The numbers displayed in the titles are the cumulative variances explained by all significant axes and the stars denote the significance of the factor: $p<0.001$ ***, $p<0.01$ **, $p<0.05$ *.

Supplementary tables

Table S1: Information about the mice carrying the indicator genus *Citrobacter* in the cecum. The haplotype h1_2 is an uncertain haplotype that might be h1 or h2, as only one SNP at the beginning of the sequence distinguish these two haplotypes, and it was not resolved for the sample MT3513.

Mice ID	*B4galnt2*	Farm	Geographical Group	Genetic Cluster	Mitochondrial D-loop Haplotype	Mitochondrial D-loop Haplogroup	Cecum Inflammation	Prevalence
JJM0503	C57BL/6J	JJM05	G2	P07	h27	H2	Healthy	RNA
JJM0602	C57BL/6J	JJM06	G2	P07	h17	H4	Healthy	Both
MJJ0104	C57BL/6J	MJJ01	G5	P08	h12	H11	Healthy	RNA
MJJ1004	C57BL/6J	MJJ10	G7	P10	h20	H4	Healthy	Both
MN4101	C57BL/6J	MN41	G11	P07	h15	H4	Healthy	DNA
JJM0401	C57BL/6J	JJM04	G2	P06	h7	H8	Inflamed	RNA
JJM0603	C57BL/6J	JJM06	G2	P07	h10	H8	Inflamed	Both
JJM0604	C57BL/6J	JJM06	G2	P07	h17	H4	Inflamed	Both
JJM0901	C57BL/6J	JJM09	G2	P06	h7	H8	Inflamed	Both
JJM0903	C57BL/6J	JJM09	G2	P07	h12	H11	Inflamed	Both
JJM0904	C57BL/6J	JJM09	G2	P06	h27	H2	Inflamed	Both
JJM0908	C57BL/6J	JJM09	G2	P06	h27	H2	Inflamed	Both
JJM0910	C57BL/6J	JJM09	G2	P06	h26	H2	Inflamed	RNA
MN0303	C57BL/6J	MN03	G9	P09	h5	H8	Inflamed	RNA
MN2603	C57BL/6J	MN26	G11	P05	h24	H2	Inflamed	Both
MN3213	C57BL/6J	MN32	G11	P13	h22	H2	Inflamed	Both
MN4102	C57BL/6J	MN41	G11	P07	h22	H2	Inflamed	RNA
MT1704	C57BL/6J	MT17	G12	P07	h1	H8	Inflamed	Both
MT3513	C57BL/6J	MT35	G12	P11	h1_2	H8	Inflamed	Both
MJJ0606	Heterozygotes	MJJ06	G6	P03	h16	H4	Healthy	RNA
MJJ0902	Heterozygotes	MJJ09	G7	P10	h20	H4	Healthy	Both
JJM0209	Heterozygotes	JJM02	G1	P04	h20	H4	Inflamed	RNA
JJM1002	Heterozygotes	JJM10	G3	P14	h23	H2	Inflamed	DNA
MJJ0115	Heterozygotes	MJJ01	G5	P08	h12	H11	Inflamed	RNA
MJJ0601	Heterozygotes	MJJ06	G6	P03	h16	H4	Inflamed	Both
MJJ0610	Heterozygotes	MJJ06	G6	P03	h16	H4	Inflamed	RNA
MT1501	Heterozygotes	MT15	G9	P16	h5	H8	Inflamed	DNA
MN3212	RIIIS/J	MN32	G11	P13	h12	H11	Healthy	RNA

Table S2: Information about the mice carrying the indicator species Otu000311 (*Bacteroidetes; Bacteroidia; Bacteroidales; Rikenellaceae; Alistipes*) in the cecum.

Mice ID	*B4galnt2*	Farm	Geographical Group	Genetic Cluster	Mitochondrial D-loop Haplotype	Mitochondrial D-loop Haplogroup	Cecum Inflammation	Prevalence
JJM0208	C57BL/6J	JJM02	G1	P04	h20	H4	Healthy	Both
MT1301	C57BL/6J	MT13	G12	P12	h1	H8	Healthy	Both
JJM0204	C57BL/6J	JJM02	G1	P07	h9	H8	Inflamed	Both
JJM0401	C57BL/6J	JJM04	G2	P06	h7	H8	Inflamed	Both
JJM0402	C57BL/6J	JJM04	G2	P04	h3	H8	Inflamed	Both
JJM0802	C57BL/6J	JJM08	G1	P04	h15	H4	Inflamed	Both
JJM0902	C57BL/6J	JJM09	G2	P06	h27	H2	Inflamed	Both
JJM0904	C57BL/6J	JJM09	G2	P06	h27	H2	Inflamed	Both
JJM0905	C57BL/6J	JJM09	G2	P06	h27	H2	Inflamed	Both
JJM0906	C57BL/6J	JJM09	G2	P06	h27	H2	Inflamed	Both
JJM0908	C57BL/6J	JJM09	G2	P06	h27	H2	Inflamed	Both
JJM1201	C57BL/6J	JJM12	G4	P07	h12	H11	Inflamed	Both
MN4104	C57BL/6J	MN41	G11	P07	h22	H2	Inflamed	Both
MN4105	C57BL/6J	MN41	G11	P07	h22	H2	Inflamed	Both
MT1302	C57BL/6J	MT13	G12	P12	h1	H8	Inflamed	Both
MT1308	C57BL/6J	MT13	G12	P12	h1	H8	Inflamed	Both
MT1401	C57BL/6J	MT14	G12	P04	h22	H2	Inflamed	Both
MT1701	C57BL/6J	MT17	G12	P04	h1	H8	Inflamed	Both
JJM0205	Heterozygotes	JJM02	G1	P04	h21	H4	Healthy	DNA
MT1307	Heterozygotes	MT13	G12	P12	h1	H8	Healthy	DNA
JJM0210	Heterozygotes	JJM02	G1	P04	h20	H4	Inflamed	Both
MN4107	Heterozygotes	MN41	G11	P13	h22	H2	Inflamed	Both
MT1303	Heterozygotes	MT13	G12	P12	h1	H8	Inflamed	Both
MT1304	Heterozygotes	MT13	G12	P12	h1	H8	Inflamed	Both
MT1306	Heterozygotes	MT13	G12	P12	h1	H8	Inflamed	Both

Table S3: Information about the mice carrying the indicator species Otu000322 (*Bacteroidetes; Bacteroidia; Bacteroidales; unclassified; unclassified*) in the cecum.

Mice ID	*B4galnt2*	Farm	Geographical Group	Genetic Cluster	Mitochondrial D-loop Haplotype	Haplotype	Cecum Inflammation	Prevalence
JJM0402	C57BL/6J	JJM04	G2	P04	h3	H8	Inflamed	Both
JJM0903	C57BL/6J	JJM09	G2	P07	h12	H11	Inflamed	Both
MN0301	C57BL/6J	MN03	G9	P09	h5	H8	Inflamed	DNA
MN0308	C57BL/6J	MN03	G9	P09	h5	H8	Inflamed	DNA
MN0309	C57BL/6J	MN03	G9	P09	h5	H8	Inflamed	Both
MN2402	C57BL/6J	MN24	G11	P07	h2	H8	Inflamed	Both
MN4102	C57BL/6J	MN41	G11	P07	h22	H2	Inflamed	DNA
MN0205	Heterozygotes	MN02	G9	P16	h6	H8	Inflamed	DNA
MN4103	Heterozygotes	MN41	G11	P07	h22	H2	Inflamed	Both
MN0302	RIIIS/J	MN03	G9	P09	h5	H8	Healthy	Both

Table S4: Information about the mice carrying the indicator species Otu000463 (*Proteobacteria; Gammaproteobacteria; Enterobacteriales; Enterobacteriaceae; Morganella*) in the cecum. The haplotype h1_2 is an uncertain haplotype that might be h1 or h2, as only one SNP at the beginning of the sequence distinguish these two haplotypes, and it was not resolved for the samples MT3513.

Mice ID	*B4galnt2*	Farm	Geographical Group	Genetic Cluster	Mitochondrial D-loop Haplotype	Haplogroup	Cecum Inflammation	Prevalence
JJM0901	C57BL/6J	JJM09	G2	P06	h7	H8	Inflamed	Both
JJM0903	C57BL/6J	JJM09	G2	P07	h12	H11	Inflamed	RNA
JJM0904	C57BL/6J	JJM09	G2	P06	h27	H2	Inflamed	Both
JJM0908	C57BL/6J	JJM09	G2	P06	h27	H2	Inflamed	Both
JJM0910	C57BL/6J	JJM09	G2	P06	h26	H2	Inflamed	Both
JJM0912	C57BL/6J	JJM09	G2	P06	h26	H2	Inflamed	Both
MJJ0117	C57BL/6J	MJJ01	G5	P08	h12	H11	Inflamed	DNA
MN2605	C57BL/6J	MN26	G11	P05	h15	H4	Inflamed	DNA
MN3204	C57BL/6J	MN32	G11	P13	h22	H2	Inflamed	DNA
MN3209	C57BL/6J	MN32	G11	P13	h12	H11	Inflamed	DNA
MN3213	C57BL/6J	MN32	G11	P13	h22	H2	Inflamed	Both
MN4104	C57BL/6J	MN41	G11	P07	h22	H2	Inflamed	DNA
MT3513	C57BL/6J	MT35	G12	P11	h1_2	H8	Inflamed	Both
MJJ0601	Heterozygotes	MJJ06	G6	P03	h16	H4	Inflamed	Both
MN3210	Heterozygotes	MN32	G11	P13	h12	H11	Inflamed	Both
MN3212	RIIIS/J	MN32	G11	P13	h12	H11	Healthy	DNA

Table S5: Information about the mice carrying the indicator species Otu000204 (*Proteobacteria; Gammaproteobacteria; Enterobacteriales; Enterobacteriaceae; Proteus*) in the cecum.

Mice ID	*B4galnt2*	Farm	Geographical Group	Genetic Cluster	Mitochondrial D-loop Haplotype	Mitochondrial D-loop Haplogroup	Cecum Inflammation	Prevalence
JJM0903	C57BL/6J	JJM09	G2	P07	h12	H11	Inflamed	RNA
JJM0904	C57BL/6J	JJM09	G2	P06	h27	H2	Inflamed	Both
JJM0910	C57BL/6J	JJM09	G2	P06	h26	H2	Inflamed	Both
MJJ0117	C57BL/6J	MJJ01	G5	P08	h12	H11	Inflamed	Both
MN3209	C57BL/6J	MN32	G11	P13	h12	H11	Inflamed	RNA
MN3213	C57BL/6J	MN32	G11	P13	h22	H2	Inflamed	Both
MN3210	Heterozygotes	MN32	G11	P13	h12	H11	Inflamed	Both

Table S6: Information about the mice carrying the indicator genus *Citrobacter* in the colon. The haplotype h1_2 is an uncertain haplotype that might be h1 or h2, as only one SNP at the beginning of the sequence distinguish these two haplotypes, and it was not resolved for the sample MT3513.

Mice ID	*B4galnt2*	Farm	Geographical Group	Genetic Cluster	Mitochondrial D-loop Haplotype	Mitochondrial D-loop Haplogroup	Gel Purification	Colon Inflammation	Prevalence
JJM0603	C57BL/6J	JJM06	G2	P07	h10	H8	No	Healthy	Both
JJM0908	C57BL/6J	JJM09	G2	P06	h27	H2	No	Healthy	Both
MT1702	C57BL/6J	MT17	G12	P07	h1	H8	Yes	Healthy	RNA
MT3508	C57BL/6J	MT35	G12	P11	h1	H8	No	Healthy	RNA
JJM0802	C57BL/6J	JJM08	G1	P04	h15	H4	Yes	Healthy	RNA
MN2402	C57BL/6J	MN24	G11	P07	h2	H8	Yes	Healthy	RNA
MN4102	C57BL/6J	MN41	G11	P07	h22	H2	Yes	Healthy	RNA
JJM0902	C57BL/6J	JJM09	G2	P06	h27	H2	Yes	Healthy	RNA
MN0303	C57BL/6J	MN03	G9	P09	h5	H8	No	Healthy	RNA
MN0305	C57BL/6J	MN03	G9	P09	h5	H8	No	Healthy	RNA
MT1704	C57BL/6J	MT17	G12	P07	h1	H8	No	Inflamed	Both
MN2603	C57BL/6J	MN26	G11	P05	h24	H2	No	Inflamed	Both
JJM0904	C57BL/6J	JJM09	G2	P06	h27	H2	Yes	Inflamed	Both
JJM0901	C57BL/6J	JJM09	G2	P06	h7	H8	No	Inflamed	Both
JJM0903	C57BL/6J	JJM09	G2	P07	h12	H11	No	Inflamed	DNA
JJM0401	C57BL/6J	JJM04	G2	P06	h7	H8	Yes	Inflamed	DNA
MT1701	C57BL/6J	MT17	G12	P04	h1	H8	Yes	Inflamed	RNA
MT3513	C57BL/6J	MT35	G12	P11	h1_2	H8	No	Inflamed	RNA
JJM0601	C57BL/6J	JJM06	G2	P07	h10	H8	Yes	Inflamed	RNA
JJM1201	C57BL/6J	JJM12	G4	P07	h12	H11	Yes	Inflamed	RNA
MJJ0111	C57BL/6J	MJJ01	G5	P08	h12	H11	Yes	Inflamed	RNA
MT1707	C57BL/6J	MT17	G12	P07	h12	H11	Yes	Inflamed	RNA
JJM0602	C57BL/6J	JJM06	G2	P07	h17	H4	Yes	Inflamed	RNA
JJM0910	C57BL/6J	JJM09	G2	P06	h26	H2	Yes	Inflamed	RNA
MN0304	C57BL/6J	MN03	G9	P09	h5	H8	Yes	Inflamed	RNA
MT1306	Heterozygotes	MT13	G12	P12	h1	H8	No	Healthy	Both
MJJ0901	Heterozygotes	MJJ09	G7	P10	h20	H4	Yes	Healthy	DNA
MJJ0610	Heterozygotes	MJJ06	G6	P03	h16	H4	No	Healthy	RNA
MT1501	Heterozygotes	MT15	G9	P16	h5	H8	Yes	Healthy	RNA
MJJ0902	Heterozygotes	MJJ09	G7	P10	h20	H4	Yes	Inflamed	Both
MT1705	Heterozygotes	MT17	G12	P11	h12	H11	Yes	Inflamed	RNA
JJM1002	Heterozygotes	JJM10	G3	P14	h23	H2	Yes	Inflamed	RNA
MN3212	RIIIS/J	MN32	G11	P13	h12	H11	No	Healthy	RNA
MJJ0608	RIIIS/J	MJJ06	G6	P03	h16	H4	No	Healthy	RNA

Table S7: Information about the mice carrying the indicator species Otu000089 (*Bacteroidetes; Bacteroidia; Bacteroidales; Porphyromonadaceae; unclassified*) in the colon.

Mice ID	*B4galnt2*	Farm	Geographical Group	Genetic Cluster	Mitochondrial D-loop Haplotype	Mitochondrial D-loop Haplogroup	Gel Purification	Colon Inflammation	Prevalence
MN0206	C57BL/6J	MN02	G9	P07	h5	H8	Yes	Healthy	Both
MN0301	C57BL/6J	MN03	G9	P09	h5	H8	Yes	Healthy	RNA
MT1302	C57BL/6J	MT13	G12	P12	h1	H8	No	Inflamed	Both
JJM1201	C57BL/6J	JJM12	G4	P07	h12	H11	Yes	Inflamed	Both
MN3208	C57BL/6J	MN32	G11	P13	h12	H11	Yes	Inflamed	Both
MT1401	C57BL/6J	MT14	G12	P04	h22	H2	No	Inflamed	Both
MN0309	C57BL/6J	MN03	G9	P09	h5	H8	No	Inflamed	Both
JJM0903	C57BL/6J	JJM09	G2	P07	h12	H11	No	Inflamed	DNA
MT1707	C57BL/6J	MT17	G12	P07	h12	H11	Yes	Inflamed	RNA
MT1303	Heterozygotes	MT13	G12	P12	h1	H8	No	Healthy	Both
MT1306	Heterozygotes	MT13	G12	P12	h1	H8	No	Healthy	Both
MT3506	Heterozygotes	MT35	G12	P11	h12	H11	Yes	Healthy	DNA
JJM0205	Heterozygotes	JJM02	G1	P04	h21	H4	Yes	Healthy	DNA
MT3510	Heterozygotes	MT35	G12	P11	h12	H11	Yes	Inflamed	RNA
JJM0203A	Heterozygotes	JJM02	G1	P04	h21	H4	Yes	Inflamed	RNA
MN0207	RIIIS/J	MN02	G9	P02	h5	H8	Yes	Healthy	RNA

Table S8: Information about the mice carrying the indicator species Otu000198 (*Firmicutes; Clostridia; Clostridiales; Lachnospiraceae; unclassified*) in the colon.

Mice ID	*B4galnt2*	Farm	Geographical Group	Genetic Cluster	Mitochondrial D-loop Haplotype	Mitochondrial D-loop Haplogroup	Gel Purification	Colon Inflammation	Prevalence
MN0202	C57BL/6J	MN02	G9	P14	h5	H8	Yes	Healthy	DNA
MJJ0101	C57BL/6J	MJJ01	G5	P08	h12	H11	No	Inflamed	Both
MJJ0103	C57BL/6J	MJJ01	G5	P08	h12	H11	No	Inflamed	Both
MN0304	C57BL/6J	MN03	G9	P09	h5	H8	Yes	Inflamed	Both
MT1401	C57BL/6J	MT14	G12	P04	h22	H2	No	Inflamed	DNA
MT2603	C57BL/6J	MT26	G12	P07	h12	H11	Yes	Inflamed	RNA
MN0311	C57BL/6J	MN03	G9	P09	h5	H8	Yes	Inflamed	RNA
JJM0205	Heterozygotes	JJM02	G1	P04	h21	H4	Yes	Healthy	RNA
MT3505	Heterozygotes	MT35	G12	P11	h15	H4	No	Inflamed	DNA
MJJ0608	RIIIS/J	MJJ06	G6	P03	h16	H4	No	Healthy	Both
MJJ0611	RIIIS/J	MJJ06	G6	P03	h16	H4	Yes	Healthy	RNA
MN0201	RIIIS/J	MN02	G9	P16	h20	H4	Yes	Healthy	RNA
MN0207	RIIIS/J	MN02	G9	P02	h5	H8	Yes	Healthy	RNA
MN0302	RIIIS/J	MN03	G9	P09	h5	H8	Yes	Healthy	RNA

Table S9: Information about the mice carrying the indicator species Otu000521 (*Firmicutes; Clostridia; Clostridiales; Lachnospiraceae; unclassified*) in the colon.

Mice ID	*B4galnt2*	Farm	Geographical Group	Genetic Cluster	Mitochondrial D-loop Haplotype	Mitochondrial D-loop Haplogroup	Gel Purification	Colon Inflammation	Prevalence
MJJ0101	C57BL/6J	MJJ01	G5	P08	h12	H11	No	Inflamed	DNA
MJJ0103	C57BL/6J	MJJ01	G5	P08	h12	H11	No	Inflamed	DNA
JJM0912	C57BL/6J	JJM09	G2	P06	h26	H2	Yes	Inflamed	DNA
MN2901	Heterozygotes	MN29	G11	P07	h14	H4	No	Healthy	DNA
MT3505	Heterozygotes	MT35	G12	P11	h15	H4	No	Inflamed	DNA

Table S10: Information about the mice carrying the indicator species Otu000293 (*Firmicutes; Clostridia; Clostridiales; Lachnospiraceae; unclassified*) in the colon.

Mice ID	*B4galnt2*	Farm	Geographical Group	Genetic Cluster	Mitochondrial D-loop Haplotype	Mitochondrial D-loop Haplogroup	Gel Purification	Colon Inflammation	Prevalence
MJJ1003	C57BL/6J	MJJ10	G7	P10	h20	H4	Yes	Healthy	Both
MT1308	C57BL/6J	MT13	G12	P12	h1	H8	No	Inflamed	Both
MN0304	C57BL/6J	MN03	G9	P09	h5	H8	Yes	Inflamed	DNA
MT1701	C57BL/6J	MT17	G12	P04	h1	H8	Yes	Inflamed	RNA
JJM0601	C57BL/6J	JJM06	G2	P07	h10	H8	Yes	Inflamed	RNA
MJJ0114	C57BL/6J	MJJ01	G5	P08	h12	H11	Yes	Inflamed	RNA
JJM0912	C57BL/6J	JJM09	G2	P06	h26	H2	Yes	Inflamed	RNA
JJM0401	C57BL/6J	JJM04	G2	P06	h7	H8	Yes	Inflamed	RNA
MJJ0702	C57BL/6J	MJJ07	G6	P14	h7	H8	No	Inflamed	RNA
JJM1202	Heterozygotes	JJM12	G4	P07	h20	H4	No	Healthy	DNA
MN0201	RIIIS/J	MN02	G9	P16	h20	H4	Yes	Healthy	DNA
MJJ0608	RIIIS/J	MJJ06	G6	P03	h16	H4	No	Healthy	RNA
MN0203	RIIIS/J	MN02	G9	P02	h6	H8	No	Inflamed	DNA

Table S11: Information about the mice carrying the indicator species Otu000408 (*Firmicutes; Clostridia; Clostridiales; Ruminococcaceae; unclassified*) in the colon.

Mice ID	*B4galnt2*	Farm	Geographical Group	Genetic Cluster	Mitochondrial D-loop Haplotype	Mitochondrial D-loop Haplogroup	Gel Purification	Colon Inflammation	Prevalence
MN0308	C57BL/6J	MN03	G9	P09	h5	H8	Yes	Healthy	RNA
MT1308	C57BL/6J	MT13	G12	P12	h1	H8	No	Inflamed	RNA
MJJ0101	C57BL/6J	MJJ01	G5	P08	h12	H11	No	Inflamed	RNA
MJJ0103	C57BL/6J	MJJ01	G5	P08	h12	H11	No	Inflamed	RNA
JJM0912	C57BL/6J	JJM09	G2	P06	h26	H2	Yes	Inflamed	RNA
MN0304	C57BL/6J	MN03	G9	P09	h5	H8	Yes	Inflamed	RNA
MJJ0109	Heterozygotes	MJJ01	G5	P08	h12	H11	Yes	Healthy	RNA
JJM0203B	Heterozygotes	JJM02	G1	P04	h20	H4	Yes	Healthy	RNA
MJJ0608	RIIIS/J	MJJ06	G6	P03	h16	H4	No	Healthy	RNA
MJJ0611	RIIIS/J	MJJ06	G6	P03	h16	H4	Yes	Healthy	RNA

Table S12: Information about the mice carrying the indicator species Otu000276 (*Firmicutes; Erysipelotrichia; Erysipelotrichales; Erysipelotrichaceae; Coprobacillus*) in the colon.

Mice ID	*B4galnt2*	Farm	Geographical Group	Genetic Cluster	Mitochondrial D-loop Haplotype	Mitochondrial D-loop Haplogroup	Gel Purification	Colon Inflammation	Prevalence
MJJ0103	C57BL/6J	MJJ01	G5	P08	h12	H11	No	Inflamed	Both
MN3204	C57BL/6J	MN32	G11	P13	h22	H2	No	Inflamed	Both
MT1401	C57BL/6J	MT14	G12	P04	h22	H2	No	Inflamed	Both
MN2603	C57BL/6J	MN26	G11	P05	h24	H2	No	Inflamed	DNA
MJJ0112	Heterozygotes	MJJ01	G5	P08	h12	H11	No	Inflamed	Both
MT3504	RIIIS/J	MT35	G12	P11	h1	H8	No	Inflamed	DNA

Table S13: Information about the mice carrying the indicator species Otu000463 (*Proteobacteria; Gammaproteobacteria; Enterobacteriales; Enterobacteriaceae; Morganella*) in the colon. The haplotype h1_2 is an uncertain haplotype that might be h1 or h2, as only one SNP at the beginning of the sequence distinguish these two haplotypes, and it was not resolved for the sample MT3513.

Mice ID	*B4galnt2*	Farm	Geographical Group	Genetic Cluster	Mitochondrial D-loop Haplotype	Mitochondrial D-loop Haplogroup	Gel Purification	Colon Inflammation	Prevalence
JJM0902	C57BL/6J	JJM09	G2	P06	h27	H2	Yes	Healthy	RNA
MT3513	C57BL/6J	MT35	G12	P11	h1_2	H8	No	Inflamed	Both
JJM0904	C57BL/6J	JJM09	G2	P06	h27	H2	Yes	Inflamed	Both
MT1306	Heterozygotes	MT13	G12	P12	h1	H8	No	Healthy	Both
MT1501	Heterozygotes	MT15	G9	P16	h5	H8	Yes	Healthy	DNA
MJJ0601	Heterozygotes	MJJ06	G6	P03	h16	H4	No	Inflamed	Both
JJM1002	Heterozygotes	JJM10	G3	P14	h23	H2	Yes	Inflamed	RNA
MN3212	RIIIS/J	MN32	G11	P13	h12	H11	No	Healthy	DNA
MJJ0602	RIIIS/J	MJJ06	G6	P03	h16	H4	Yes	Inflamed	Both

Table S14: Information about the mice carrying the indicator species Otu000204 (*Proteobacteria; Gammaproteobacteria; Enterobacteriales; Enterobacteriaceae; Proteus*) in the colon.

Mice ID	*B4galnt2*	Farm	Geographical Group	Genetic Cluster	Mitochondrial D-loop Haplotype	Mitochondrial D-loop Haplogroup	Gel Purification	Colon Inflammation	Prevalence
MJJ0117	C57BL/6J	MJJ01	G5	P08	h12	H11	No	Inflamed	Both
JJM0904	C57BL/6J	JJM09	G2	P06	h27	H2	Yes	Inflamed	Both
JJM0910	C57BL/6J	JJM09	G2	P06	h26	H2	Yes	Inflamed	RNA
MN0205	Heterozygotes	MN02	G9	P16	h6	H8	No	Inflamed	Both
MN0207	RIIIS/J	MN02	G9	P02	h5	H8	Yes	Healthy	Both

Chapter III:

Characterization of candidate pathogens influenced by *B4galnt2* genotype

Introduction

In Chapter II, I used a novel approach that allowed me to identify several candidate pathogens that could be responsible for the selection acting on *B4glant2* in the wild. Among the candidates are *Citrobacter*, *Morganella* and *Proteus*, which I decided to further characterize, since they were identified at the species-level OTU or the genus level, which renders precise identification of the bacteria more feasible, and most importantly, they are three genera known for their pathogenicity, which makes them the most promising candidate pathogens.

In this chapter, I aimed to characterize the candidates identified in chapter II. First, I performed PCR-based cloning and sequencing of a nearly full-length 16SrRNA in order to have a more precise classification of the candidates. Then, I performed whole genome sequencing on the candidates' isolates -- obtained through collaboration with Guntram Grassl from the samples I took in the wild -- to gain insight in their phylogenetic classification, and to assess their pathogenic potential.

I could identify my candidate *Citrobacter* as belonging to *Citrobacter freundii* and the *Morganella* candidates are members of the *Morganella morganii* species, with two colonies representing new strains of the *Morganella morganii morganii* subspecies and four other colonies belonging to a totally new subspecies of *Morganella morganii*. This new subspecies, which seem to correspond the OTU that I identified, shows features belonging to pathogenicity functions, reinforcing its potential role as pathogen.

Results

Although I found interesting candidates in both the cecum and colon, the indicator OTU in the colon appear more difficult to characterize further, as their pattern is not as clear cut as in the cecum, and they generally belong to a broad taxonomic group, making it harder to characterize deeply. This is why I focused on the cecum indicators, and chose to characterize *Citrobacter* further, as it was previously observe as indicator of the *B4galnt2* genotype in laboratory mice, and *Proteus* and *Morganella*, both seeming to be convincing candidate pathogens driving inflammation specifically in C57BL/6J mice.

I. *Citrobacter*

During my analysis of indicator species, I was puzzled by one fact: *Citrobacter*, although relatively abundant at the genus level, is not present at the OTU level (i.e. there is no OTU belonging to *Citrobacter*). I tried using less stringent threshold for the classification of the sequences and the OTU, but it didn't change the results for *Citrobacter*. In fact, the problem stem from the *Enterobacteriaceae* group not being very divergent, leading many sequences classified to other genera within the *Enterobacteriaceae* to be binned together with *Citrobacter* sequences, so that *Citrobacter* do not reach the threshold for classification (figure 1).

I nonetheless continued to further characterize the *Citrobacter* bacteria present in my samples. First I used specific primer pairs to amplify, clone and sequence a nearly full-length 16SrRNA gene, to gain deeper insight in the taxonomy of my *Citrobacter*, which cluster together with *Citrobacter freundii* (figure 2, table 1). Furthermore, three *Citrobacter* colonies could be isolated from my cecum samples, of which I could sequence the whole genome. From the genome assemblies, I extracted the 16SrRNA gene to compare to the previously cloned sequences. It is known that bacteria often have multiple copies of the 16SrRNA gene, and in particular the Enterobacteriaceae are known to carry on average seven copies of this gene in their genome (*Stoddard, Smith et al. 2015*). From the whole genome assembly, I could resolve at least five alleles of the 16SrRNA gene, but I couldn't determine the full phase of the alleles, since the distance between the SNPs at the beginning of the gene and those at the end exceeded the read length. Because of this, I had to split the sequence in two portion of certain phase (figure 3, tables 2 & 3). The two portions leads to different phylogenetic trees, but both agree that the genes from

the isolated colonies, and the genes obtained through PCR, cloning and sequencing cluster together, and close to *Citrobacter freundii*. To have a final classification of my *Citrobacter* candidate, I used ANDI (*Haubold, Klotzl et al. 2015*) to estimate the genome-wide distance between pairs of strains, and R to build a phylogenetic tree based on that distance (*Dixon 2003*, *Paradis, Claude et al. 2004*, *Schliep 2011*). With this method, all three *Citrobacter* isolates show 100% identity, and they cluster with *Citrobacter freundii*, with ~95% identity (figure 4, table 4). Other *Citrobacter* species cluster much further away at ~85% identity, when the outgroup *Morganella*, is on average at 77.5% identity to the *Citrobacter* species. Although some *Citrobacter* species are missing from this whole genome phylogeny, as not every species has a fully sequenced genome available in public databases, it is safe to conclude that my *Citrobacter* candidate is a new strain of *Citrobacter freundii*.

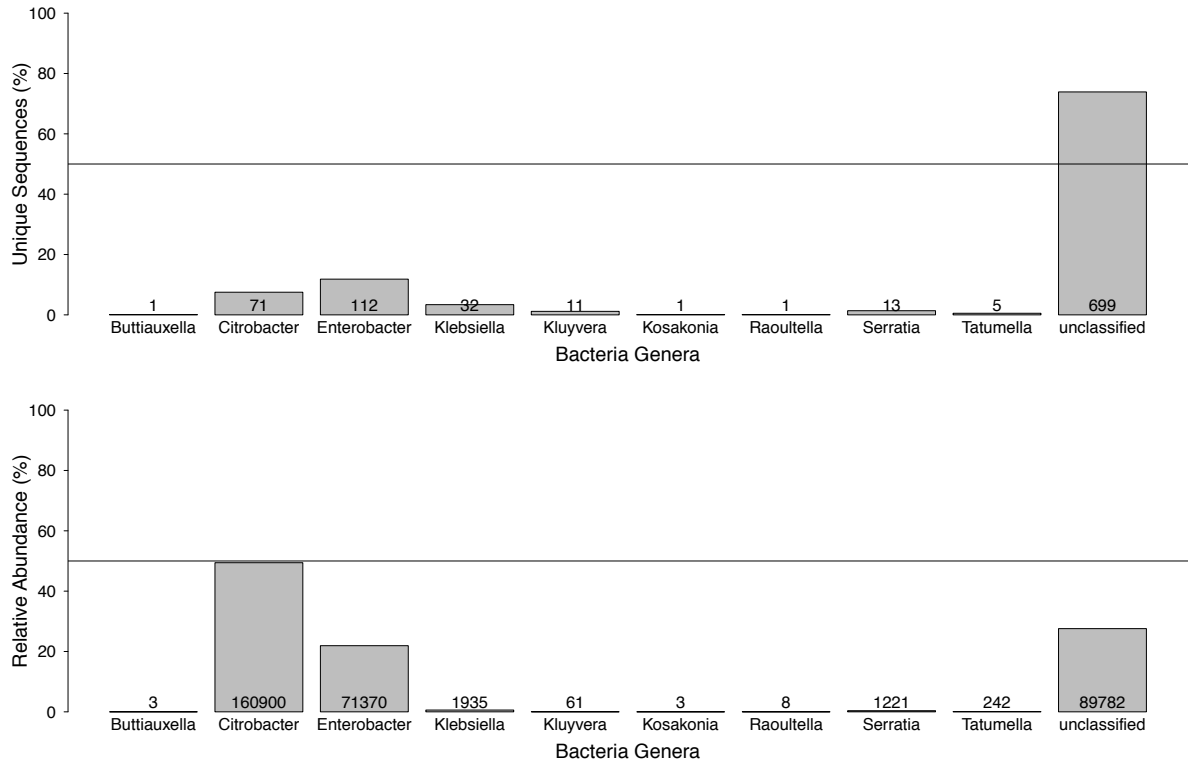

Figure 1: Distribution of the sequences belonging to the major *Citrobacter* OTU according to their classification at the genus level. The top graph shows the percentage of unique sequences while the bottom graph shows the their relative abundance.

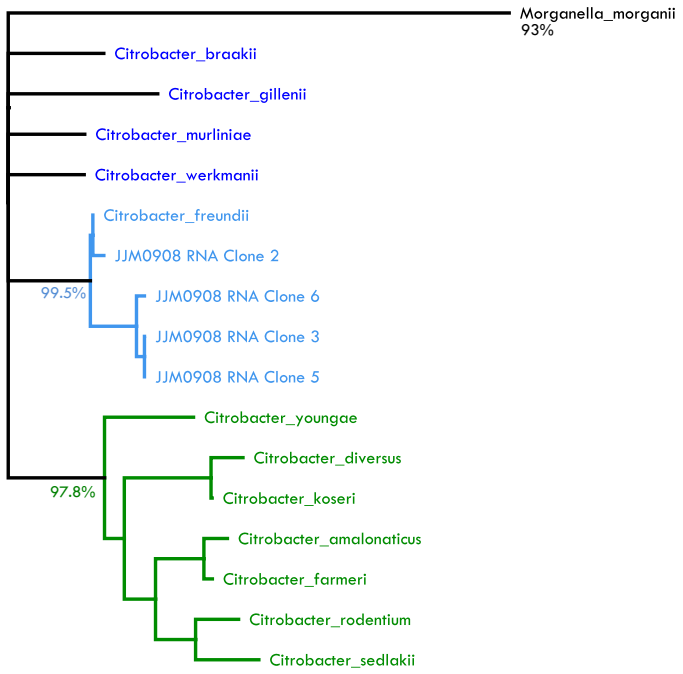

Figure 2: Neighbor-joining tree of the *Citrobacter* 16SrRNA gene of clones obtained from one RNA sample (JJM0908) and the *Citrobacter* references. The tree is based on a 1146 bp alignment of the 16SrRNA gene, amplified with the primers 27F and ArsR (*Hubert, Erban et al. 2015*). *Morganella morganii morganii* was used as outgroup. Bootstrapping was used with 10.000 iterations and a threshold of 50% support. The percentages refer to the average pairwise identity between members of the corresponding phylogenetic cluster.

Table 1: Pairwise identity of the *Citrobacter* 16S rRNA gene of clones obtained from one RNA sample (JJM0908) and the *Citrobacter* references. The table is based on a 1146 bp alignment of the 16S rRNA gene, amplified with the primers 27F and ArsR (*Hubert, Erban et al. 2015*). *Morganella morganii morganii* was used as outgroup. The upper triangle contains the number of SNPs while the lower triangle contains the percentage identity.

	Morganella morganii	*Citrobacter braakii*	*Citrobacter gillenii*	*Citrobacter murliniae*	*Citrobacter werkmanii*	*Citrobacter freundii*	JJM0908 RNA Clone 2	JJM0908 RNA Clone 6	JJM0908 RNA Clone 3	JJM0908 RNA Clone 5	*Citrobacter youngae*	*Citrobacter diversus*	*Citrobacter koseri*	*Citrobacter amalonaticus*	*Citrobacter farmeri*	*Citrobacter rodentium*	*Citrobacter sedlakii*
Morganella morganii		71	84	69	72	73	73	79	79	79	90	81	80	78	82	90	95
Citrobacter braakii	93.5		11	7	13	8	10	14	14	14	19	31	32	30	30	37	35
Citrobacter gillenii	92.7	99.0		16	22	13	15	20	20	20	21	42	42	40	39	44	40
Citrobacter murliniae	94.0	99.4	98.6		11	5	7	12	12	12	25	39	40	32	34	38	43
Citrobacter werkmanii	93.8	98.9	98.2	99.2		11	11	18	18	18	33	40	42	38	36	40	44
Citrobacter freundii	93.6	99.3	98.9	99.6	99.2		2	7	7	7	24	41	41	35	35	37	43
JJM0908 RNA Clone 2	93.6	99.1	98.7	99.4	99.2	99.8		9	9	9	26	43	42	37	37	39	45
JJM0908 RNA Clone 6	93.1	98.7	98.3	99.0	98.6	99.4	99.2		2	2	27	44	44	40	40	37	46
JJM0908 RNA Clone 3	93.1	98.7	98.3	99.0	98.6	99.4	99.2	99.8		0	27	44	44	40	40	37	46
JJM0908 RNA Clone 5	93.1	98.7	98.3	99.0	98.6	99.4	99.2	99.8	100.0		27	44	44	40	40	37	46
Citrobacter youngae	92.1	98.3	98.2	97.8	97.2	97.9	97.7	97.6	97.6	97.6		32	34	33	33	30	20
Citrobacter diversus	92.9	97.2	96.3	96.6	96.6	96.4	96.2	96.2	96.2	96.2	97.2		9	32	30	32	25
Citrobacter koseri	93.1	97.3	96.5	96.6	96.6	96.6	96.5	96.4	96.4	96.4	97.3	99.4		32	33	37	30
Citrobacter amalonaticus	93.2	97.3	96.5	97.2	96.8	96.9	96.8	96.5	96.5	96.5	97.1	97.2	97.4		6	23	24
Citrobacter farmeri	92.9	97.3	96.7	97.1	96.9	97.0	96.8	96.6	96.6	96.6	97.2	97.4	97.4	99.5		21	22
Citrobacter rodentium	92.2	96.7	96.2	96.7	96.7	96.8	96.6	96.8	96.8	96.8	97.4	97.2	97.1	98.0	98.3		16
Citrobacter sedlakii	91.8	96.9	96.6	96.3	96.3	96.3	96.1	96.0	96.0	96.0	98.3	97.9	97.7	97.9	98.2	98.7	

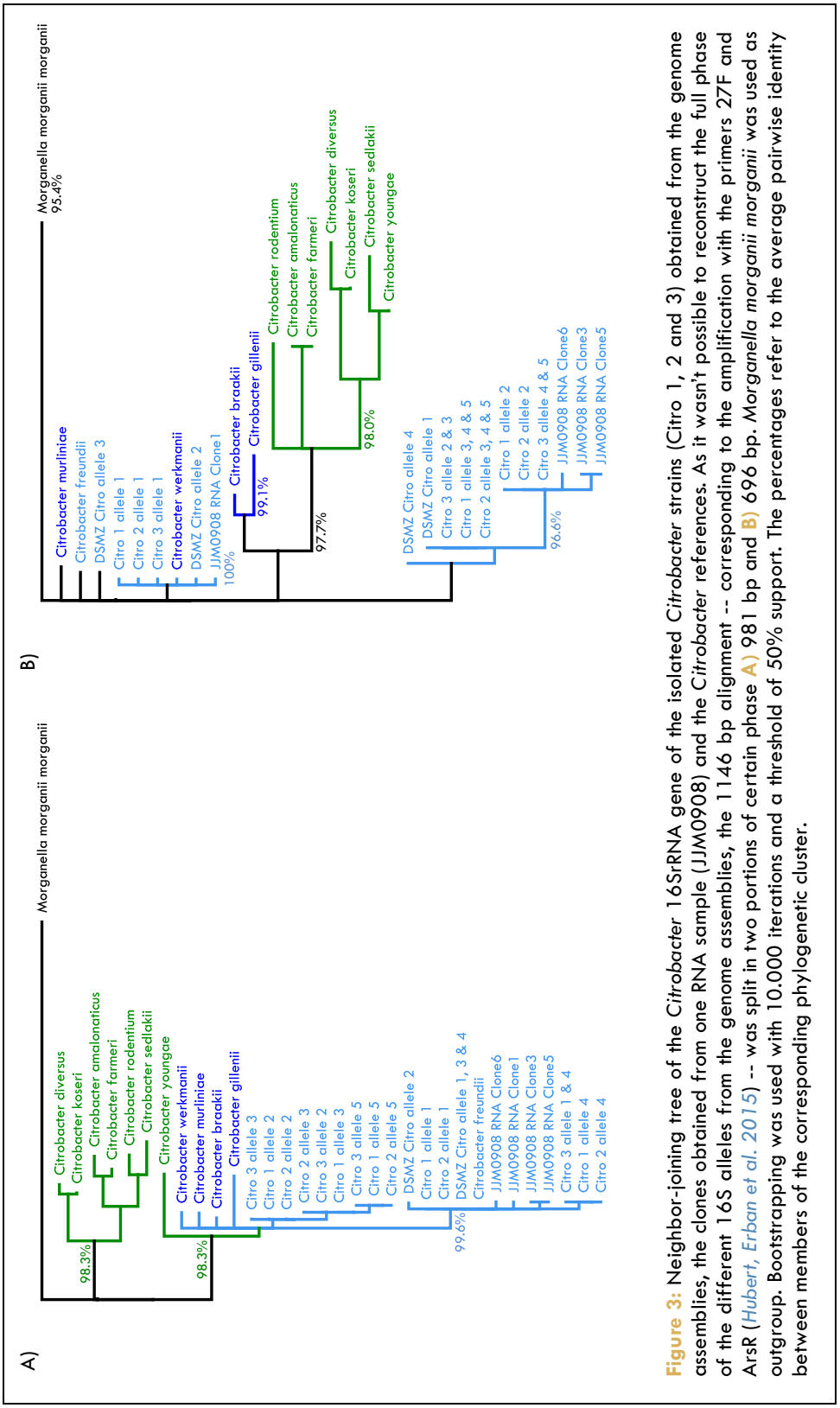

Figure 3: Neighbor-joining tree of the *Citrobacter* 16SrRNA gene of the isolated *Citrobacter* strains (Citro 1, 2 and 3) obtained from the genome assemblies, the clones obtained from one RNA sample (JJM0908) and the *Citrobacter* references. As it wasn't possible to reconstruct the full phase of the different 16S alleles from the genome assemblies, the 1146 bp alignment -- corresponding to the amplification with the primers 27F and ArsR (*Hubert, Erban et al. 2015*) -- was split in two portions of certain phase **A)** 981 bp and **B)** 696 bp. *Morganella morganii morganii* was used as outgroup. Bootstrapping was used with 10.000 iterations and a threshold of 50% support. The percentages refer to the average pairwise identity between members of the corresponding phylogenetic cluster.

Table 2: Pairwise identity of the *Citrobacter* 16S rRNA gene of the isolated *Citrobacter* strains (Citro 1, 2 and 3) obtained from the genome assemblies, the clones obtained from one RNA sample (JJM0908), the DSMZ *Citrobacter freundii* type strain and the *Citrobacter* references. This table present only the first 981 bp of the 1146 bp alignment, as it wasn't possible to reconstruct the full phase of the different 16S alleles from the genome assemblies, the full alignment was split in two portions of certain phase. *Morganella morganii morganii* was used as outgroup. The upper triangle contains the number of SNPs while the lower triangle contains the percentage identity.

	Morganella morganii morganii	*Citrobacter diversus*	*Citrobacter koseri*	*Citrobacter amalonaticus*	*Citrobacter farmeri*	*Citrobacter rodentium*	*Citrobacter sedlakii*	*Citrobacter youngae*	*Citrobacter werkmanii*	*Citrobacter murliniae*	*Citrobacter braakii*	*Citrobacter gillenii*	Citro 3 allele 3	Citro 1 allele 2	Citro 2 allele 2	Citro 2 allele 3	Citro 3 allele 2	Citro 1 allele 3	Citro 3 allele 5	Citro 1 allele 5	Citro 2 allele 5	DSMZ Citro allele 2	Citro 1 allele 1	Citro 2 allele 1	DSMZ Citro allele 1, 3 & 4	*Citrobacter freundii*	JJM0908 RNA Clone6	JJM0908 RNA Clone1	JJM0908 RNA Clone3	JJM0908 RNA Clone5	Citro 3 allele 1 & 4	Citro 1 allele 4	Citro 2 allele 4
Morganella morganii morganii		65	65	69	73	79	82	77	67	63	62	75	65	65	65	66	66	66	67	67	67	68	67	67	67	67	68	68	68	68	68	68	68
Citrobacter diversus	93.4		3	23	21	23	19	26	28	29	23	33	28	28	28	29	29	29	30	30	30	31	30	30	30	30	31	31	31	31	31	31	31
Citrobacter koseri	93.4	99.8		22	23	25	22	26	31	29	23	33	28	28	28	29	29	29	30	30	30	31	30	30	30	30	31	31	31	31	31	31	31
Citrobacter amalonaticus	93.0	97.7	97.8		6	17	18	27	33	29	28	35	31	31	31	32	32	32	33	33	33	32	31	31	31	31	31	32	32	32	32	32	32
Citrobacter farmeri	92.6	97.9	97.8	99.5		15	16	27	31	31	28	34	31	31	31	32	32	32	33	33	33	31	31	31	31	31	31	32	32	32	32	32	32
Citrobacter rodentium	91.9	97.7	97.6	98.3	98.5		9	23	33	33	30	35	31	31	31	30	30	30	31	31	31	32	31	31	31	31	31	32	32	32	32	32	32	32
Citrobacter sedlakii	91.7	98.1	97.9	98.2	98.5	99.1		20	35	36	30	34	35	35	35	34	34	34	35	35	35	36	35	35	35	35	36	36	36	36	36	36	36
Citrobacter youngae	92.2	97.3	97.5	97.2	97.3	97.7	98.0		24	18	14	15	16	16	16	17	17	17	18	18	18	17	16	16	16	16	16	17	17	17	17	17	17	17
Citrobacter werkmanii	93.2	97.3	97.1	96.8	96.9	96.8	96.6	97.7		9	8	18	8	8	8	9	9	9	10	10	10	11	10	10	10	10	11	11	11	11	11	11	11
Citrobacter murliniae	93.6	97.0	97.1	97.0	96.9	96.6	96.4	98.2	99.2		4	12	2	2	2	3	3	3	4	4	4	5	4	4	4	4	5	5	5	5	5	5	5
Citrobacter braakii	93.3	97.5	97.6	97.0	97.1	96.8	96.8	98.5	99.2	99.6		8	2	2	2	3	3	3	4	4	4	5	4	4	4	4	5	5	5	5	5	5	5
Citrobacter gillenii	92.4	96.6	96.7	96.4	96.6	96.4	96.6	98.5	98.3	98.8	99.1		10	10	10	11	11	11	12	12	12	10	10	10	10	10	11	11	11	11	11	11	11
Citro 3 allele 3	93.4	97.1	97.2	96.8	96.9	96.8	96.5	98.4	99.3	99.8	99.8	99.0		0	0	1	1	1	2	2	2	3	2	2	2	2	3	3	3	3	3	3	3
Citro 1 allele 2	93.4	97.1	97.2	96.8	96.9	96.8	96.5	98.4	99.3	99.8	99.8	99.0	100.0		0	1	1	1	2	2	2	3	2	2	2	2	3	3	3	3	3	3	3
Citro 2 allele 2	93.4	97.1	97.2	96.8	96.9	96.8	96.5	98.4	99.3	99.8	99.8	99.0	100.0	100.0		1	1	1	2	2	2	3	2	2	2	2	3	3	3	3	3	3	3
Citro 2 allele 3	93.3	97.0	97.1	96.7	96.8	96.9	96.6	98.3	99.2	99.7	99.7	98.9	99.9	99.9	99.9		0	0	1	1	1	4	3	3	3	3	4	4	4	4	4	4	4
Citro 3 allele 2	93.3	97.0	97.1	96.7	96.8	96.9	96.6	98.3	99.2	99.7	99.7	98.9	99.9	99.9	99.9	100.0		0	1	1	1	4	3	3	3	3	4	4	4	4	4	4	4
Citro 1 allele 3	93.3	97.0	97.1	96.7	96.8	96.9	96.6	98.3	99.2	99.7	99.7	98.9	99.9	99.9	99.9	100.0	100.0		1	1	1	4	3	3	3	3	4	4	4	4	4	4	4
Citro 3 allele 5	93.2	96.9	97.0	96.6	96.7	96.8	96.5	98.2	99.1	99.6	99.6	98.8	99.8	99.8	99.8	99.9	99.9	99.9		0	0	5	4	4	4	4	5	5	5	5	5	5	5
Citro 1 allele 5	93.2	96.9	97.0	96.6	96.7	96.8	96.5	98.2	99.1	99.6	99.6	98.8	99.8	99.8	99.8	99.9	99.9	99.9	100.0		0	5	4	4	4	4	5	5	5	5	5	5	5
Citro 2 allele 5	93.2	96.9	97.0	96.6	96.7	96.8	96.5	98.2	99.1	99.6	99.6	98.8	99.8	99.8	99.8	99.9	99.9	99.9	100.0	100.0		5	4	4	4	4	5	5	5	5	5	5	5
DSMZ Citro allele 2	93.1	96.8	96.9	96.7	96.9	96.7	96.4	98.3	99.0	99.5	99.5	99.0	99.7	99.7	99.7	99.6	99.6	99.6	99.5	99.5	99.5		1	1	1	1	2	2	2	2	2	2	2
Citro 1 allele 1	93.2	96.9	97.0	96.8	96.9	96.8	96.5	98.4	99.1	99.6	99.6	99.0	99.8	99.8	99.8	99.7	99.7	99.7	99.6	99.6	99.6	99.9		0	0	0	1	1	1	1	1	1	1
Citro 2 allele 1	93.2	96.9	97.0	96.8	96.9	96.8	96.5	98.4	99.1	99.6	99.6	99.0	99.8	99.8	99.8	99.7	99.7	99.7	99.6	99.6	99.6	99.9	100.0		0	0	1	1	1	1	1	1	1
DSMZ Citro allele 1, 3 & 4	93.2	96.9	97.0	96.8	96.9	96.8	96.5	98.4	99.1	99.6	99.6	99.0	99.8	99.8	99.8	99.7	99.7	99.7	99.6	99.6	99.6	99.9	100.0	100.0		0	1	1	1	1	1	1	1
Citrobacter freundii	93.2	96.9	97.0	96.8	96.9	96.8	96.5	98.4	99.1	99.6	99.6	99.0	99.8	99.8	99.8	99.7	99.7	99.7	99.6	99.6	99.6	99.9	100.0	100.0	100.0		1	1	1	1	1	1	1
JJM0908 RNA Clone6	93.1	96.8	96.9	96.7	96.8	96.7	96.4	98.3	99.0	99.5	99.5	98.9	99.7	99.7	99.7	99.6	99.6	99.6	99.5	99.5	99.5	99.8	99.9	99.9	99.9	99.9		2	2	2	2	2	2
JJM0908 RNA Clone1	93.1	96.8	96.9	96.7	96.8	96.7	96.4	98.3	99.0	99.5	99.5	98.9	99.7	99.7	99.7	99.6	99.6	99.6	99.5	99.5	99.5	99.8	99.9	99.9	99.9	99.9	99.8		2	2	2	2	2
JJM0908 RNA Clone3	93.1	96.8	96.9	96.7	96.8	96.7	96.4	98.3	99.0	99.5	99.5	98.9	99.7	99.7	99.7	99.6	99.6	99.6	99.5	99.5	99.5	99.8	99.9	99.9	99.9	99.9	99.8	99.8		0	2	2	2
JJM0908 RNA Clone5	93.1	96.8	96.9	96.7	96.8	96.7	96.4	98.3	99.0	99.5	99.5	98.9	99.7	99.7	99.7	99.6	99.6	99.6	99.5	99.5	99.5	99.8	99.9	99.9	99.9	99.9	99.8	99.8	100.0		2	2	2
Citro 3 allele 1 & 4	93.1	96.8	96.9	96.7	96.8	96.7	96.4	98.3	99.0	99.5	99.5	98.9	99.7	99.7	99.7	99.6	99.6	99.6	99.5	99.5	99.5	99.8	99.9	99.9	99.9	99.9	99.8	99.8	99.8	99.8		0	0
Citro 1 allele 4	93.1	96.8	96.9	96.7	96.8	96.7	96.4	98.3	99.0	99.5	99.5	98.9	99.7	99.7	99.7	99.6	99.6	99.6	99.5	99.5	99.5	99.8	99.9	99.9	99.9	99.9	99.8	99.8	99.8	99.8	100.0		0
Citro 2 allele 4	93.1	96.8	96.9	96.7	96.8	96.7	96.4	98.3	99.0	99.5	99.5	98.9	99.7	99.7	99.7	99.6	99.6	99.6	99.5	99.5	99.5	99.8	99.9	99.9	99.9	99.9	99.8	99.8	99.8	99.8	100.0	100.0	

Table 3: Pairwise identity of the *Citrobacter* 16S rRNA gene of the isolated *Citrobacter* strains (Citro 1, 2 and 3) obtained from the genome assemblies, the clones obtained from one RNA sample (JJM0908), the DSMZ *Citrobacter freundii* type strain and the *Citrobacter* references. This table present only the last 696 bp of the 1146 bp alignment, as it wasn't possible to reconstruct the full phase of the different 16S alleles from the genome assemblies, the full alignment was split in two portions of certain phase. *Morganella morganii morganii* was used as outgroup. The upper triangle contains the number of SNPs while the lower triangle contains the percentage identity.

	Morganella morganii morganii	*Citrobacter murliniae*	*Citrobacter freundii*	DSMZ Citro allele 3	Citro 1 allele 1	Citro 2 allele 1	Citro 3 allele 1	*Citrobacter werkmanii*	DSMZ Citro allele 2	JJM0908 RNA Clone1	*Citrobacter braakii*	*Citrobacter gillenii*	*Citrobacter rodentium*	*Citrobacter amalonaticus*	*Citrobacter farmeri*	*Citrobacter diversus*	*Citrobacter koseri*	*Citrobacter sedlakii*	*Citrobacter youngae*	DSMZ Citro allele 4	DSMZ Citro allele 1	Citro 3 allele 2 & 3	Citro 1 allele 3, 4 & 5	Citro 2 allele 3, 4 & 5	Citro 1 allele 2	Citro 2 allele 2	Citro 3 allele 4 & 5	JJM0908 RNA Clone6	JJM0908 RNA Clone3	JJM0908 RNA Clone5
Morganella morganii morganii		27	27	27	26	26	26	26	26	26	32	35	42	40	40	40	39	47	43	27	28	28	28	28	32	32	32	33	33	33
Citrobacter murliniae	96.1		1	1	2	2	2	2	2	2	5	9	17	15	15	23	24	22	18	2	3	3	3	3	7	7	7	8	8	8
Citrobacter freundii	96.1	99.9		0	1	1	1	1	1	1	6	8	18	16	16	24	24	23	19	1	2	2	2	2	6	6	6	7	7	7
DSMZ Citro allele 3	96.1	99.9	100.0		1	1	1	1	1	1	6	8	18	16	16	24	24	23	19	1	2	2	2	2	6	6	6	7	7	7
Citro 1 allele 1	96.3	99.7	99.9	99.9		0	0	0	0	0	7	9	19	17	17	25	24	24	20	2	3	3	3	3	7	7	7	8	8	8
Citro 2 allele 1	96.3	99.7	99.9	99.9	100.0		0	0	0	0	7	9	19	17	17	25	24	24	20	2	3	3	3	3	7	7	7	8	8	8
Citro 3 allele 1	96.3	99.7	99.9	99.9	100.0	100.0		0	0	0	7	9	19	17	17	25	24	24	20	2	3	3	3	3	7	7	7	8	8	8
Citrobacter werkmanii	96.3	99.7	99.9	99.9	100.0	100.0	100.0		0	0	7	9	19	17	17	25	24	24	20	2	3	3	3	3	7	7	7	8	8	8
DSMZ Citro allele 2	96.3	99.7	99.9	99.9	100.0	100.0	100.0	100.0		0	7	9	19	17	17	25	24	24	20	2	3	3	3	3	7	7	7	8	8	8
JJM0908 RNA Clone1	96.3	99.7	99.9	99.9	100.0	100.0	100.0	100.0	100.0		7	9	19	17	17	25	24	24	20	2	3	3	3	3	7	7	7	8	8	8
Citrobacter braakii	95.4	99.3	99.1	99.1	99.0	99.0	99.0	99.0	99.0	99.0		6	21	14	14	19	20	18	14	7	8	8	8	8	11	11	11	12	12	12
Citrobacter gillenii	95.0	98.7	98.8	98.8	98.7	98.7	98.7	98.7	98.7	98.7	99.1		26	20	20	22	22	21	17	9	10	10	10	10	14	14	14	15	15	15
Citrobacter rodentium	94.0	97.6	97.5	97.5	97.3	97.3	97.3	97.3	97.3	97.3	97.1	96.3		12	12	22	25	14	16	19	20	20	20	20	17	17	17	18	18	18
Citrobacter amalonaticus	94.3	97.8	97.7	97.7	97.6	97.6	97.6	97.6	97.6	97.6	98.0	97.1	98.3		0	20	21	13	15	17	18	18	18	18	20	20	20	21	21	21
Citrobacter farmeri	94.3	97.8	97.7	97.7	97.6	97.6	97.6	97.6	97.6	97.6	98.0	97.1	98.3	100.0		20	21	13	15	17	18	18	18	18	20	20	20	21	21	21
Citrobacter diversus	94.3	96.7	96.5	96.5	96.4	96.4	96.4	96.4	96.4	96.4	97.3	96.8	96.9	97.1	97.1		6	16	14	25	26	26	26	26	26	26	26	27	27	27
Citrobacter koseri	94.5	96.7	96.7	96.7	96.7	96.7	96.7	96.7	96.7	96.7	97.3	97.0	96.8	97.3	97.3	99.4		18	16	25	26	26	26	26	26	26	26	27	27	27
Citrobacter sedlakii	93.2	96.8	96.7	96.7	96.5	96.5	96.5	96.5	96.5	96.5	97.4	97.0	98.1	98.1	98.1	97.7	97.7		4	24	25	25	25	25	25	25	25	26	26	26
Citrobacter youngae	93.8	97.4	97.3	97.3	97.1	97.1	97.1	97.1	97.1	97.1	98.0	97.6	97.8	97.8	97.8	98.0	98.0	99.4		20	21	21	21	21	21	21	21	22	22	22
DSMZ Citro allele 4	96.1	99.7	99.9	99.9	99.7	99.7	99.7	99.7	99.7	99.7	99.0	98.7	97.3	97.6	97.6	96.4	96.5	96.5	97.1		1	1	1	1	5	5	5	6	6	6
DSMZ Citro allele 1	96.0	99.6	99.7	99.7	99.6	99.6	99.6	99.6	99.6	99.6	98.8	98.6	97.2	97.4	97.4	96.3	96.4	96.4	97.0	99.9		0	0	0	4	4	4	5	5	5
Citro 3 allele 2 & 3	96.0	99.6	99.7	99.7	99.6	99.6	99.6	99.6	99.6	99.6	98.8	98.6	97.2	97.4	97.4	96.3	96.4	96.4	97.0	99.9	100.0		0	0	4	4	4	5	5	5
Citro 1 allele 3, 4 & 5	96.0	99.6	99.7	99.7	99.6	99.6	99.6	99.6	99.6	99.6	98.8	98.6	97.2	97.4	97.4	96.3	96.4	96.4	97.0	99.9	100.0	100.0		0	4	4	4	5	5	5
Citro 2 allele 3, 4 & 5	96.0	99.6	99.7	99.7	99.6	99.6	99.6	99.6	99.6	99.6	98.8	98.6	97.2	97.4	97.4	96.3	96.4	96.4	97.0	99.9	100.0	100.0	100.0		4	4	4	5	5	5
Citro 1 allele 2	95.4	99.0	99.1	99.1	99.0	99.0	99.0	99.0	99.0	99.0	98.4	98.0	97.6	97.1	97.1	96.3	96.5	96.4	97.0	99.3	99.4	99.4	99.4	99.4		0	0	1	1	1
Citro 2 allele 2	95.4	99.0	99.1	99.1	99.0	99.0	99.0	99.0	99.0	99.0	98.4	98.0	97.6	97.1	97.1	96.3	96.5	96.4	97.0	99.3	99.4	99.4	99.4	99.4	100.0		0	1	1	1
Citro 3 allele 4 & 5	95.4	99.0	99.1	99.1	99.0	99.0	99.0	99.0	99.0	99.0	98.4	98.0	97.6	97.1	97.1	96.3	96.5	96.4	97.0	99.3	99.4	99.4	99.4	99.4	100.0	100.0		1	1	1
JJM0908 RNA Clone6	95.3	98.8	99.0	99.0	98.8	98.8	98.8	98.8	98.8	98.8	98.3	97.8	97.5	97.0	97.0	96.1	96.3	96.3	96.8	99.1	99.3	99.3	99.3	99.3	99.9	99.9	99.9		2	2
JJM0908 RNA Clone3	95.3	98.8	99.0	99.0	98.8	98.8	98.8	98.8	98.8	98.8	98.3	97.8	97.5	97.0	97.0	96.1	96.3	96.3	96.8	99.1	99.3	99.3	99.3	99.3	99.9	99.9	99.9	99.7		0
JJM0908 RNA Clone5	95.3	98.8	99.0	99.0	98.8	98.8	98.8	98.8	98.8	98.8	98.3	97.8	97.5	97.0	97.0	96.1	96.3	96.3	96.8	99.1	99.3	99.3	99.3	99.3	99.9	99.9	99.9	99.7	100.0	

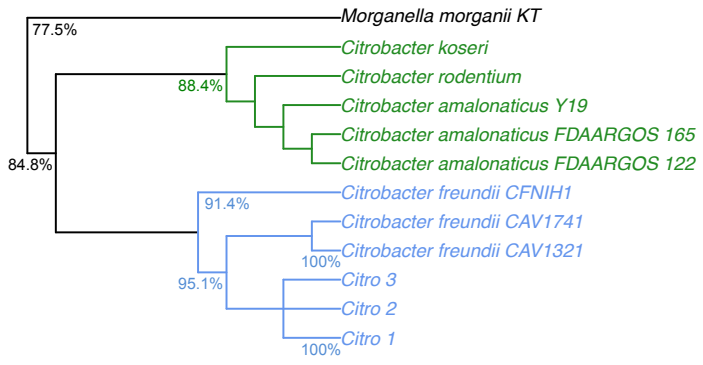

Figure 4: Neighbor-joining tree of the isolated *Citrobacter* strains and the *Citrobacter* references, based on the whole genome identities estimated by ANDI (*Haubold, Klotzl et al.*) *Morganella morganii morganii* KT was used as outgroup. Bootstrapping was used with 10.000 iterations and a threshold of 50% support. The percentages refer to the average pairwise identity between members of the corresponding phylogenetic cluster.

Table 4: Pairwise identity of the isolated *Citrobacter* strains and the *Citrobacter* references, based on the whole genome identities estimated by ANDI (*Haubold, Klotzl et al. 2015*) *Morganella morganii morganii* KT was used as outgroup.

	Morganella morganii KT	*Citrobacter koseri*	*Citrobacter rodentium*	*Citrobacter amalonaticus Y19*	*Citrobacter amalonaticus FDAARGOS 165*	*Citrobacter amalonaticus FDAARGOS 122*	*Citrobacter freundii CFNIH1*	*Citrobacter freundii CAV1741*	*Citrobacter freundii CAV1321*	Citro 3	Citro 2
Citrobacter koseri	77.3										
Citrobacter rodentium	77.3	85.9									
Citrobacter amalonaticus Y19	77.0	85.6	86.5								
Citrobacter amalonaticus FDAARGOS 165	77.0	85.6	86.4	92.5							
Citrobacter amalonaticus FDAARGOS 122	77.6	85.7	86.6	93.2	95.7						
Citrobacter freundii CFNIH1	77.6	85.4	84.3	84.8	85.0	84.6					
Citrobacter freundii CAV1741	77.4	85.1	84.2	84.9	84.7	84.7	91.3				
Citrobacter freundii CAV1321	77.7	85.1	84.2	84.9	84.7	84.5	91.3	100.0			
Citro 3	78.0	85.4	84.4	84.6	84.7	84.9	91.5	95.1	95.1		
Citro 2	78.0	85.4	84.4	84.6	84.7	84.9	91.5	95.1	95.1	100.0	
Citro 1	78.1	85.4	84.4	84.6	84.7	84.9	91.5	95.1	95.1	100.0	100.0

II. *Proteus*

For *Proteus*, three colonies were successfully isolated from my samples. Similar to the *Citrobacter* assemblies, I extracted the 16SrRNA gene and identified three potential alleles. Again, the phase between SNPs at the beginning of the sequence and at the end could not be directly determined, as the distance between them was higher than the read length, but I could infer it based on variant frequencies. In any case, the three colonies present the same three variable positions with similar variant frequencies, suggesting a 100% identity between the three colonies. Indeed, these alleles cluster together (figure 5, table 5), and together in a bigger cluster including all strains of *Proteus mirabilis* and *Proteus vulgaris*, one strain of *Proteus hauseri*, and *Proteus penneri*. Within this super-cluster, the sequences have on average 99.4% pairwise identity, which is much greater than the generally admitted 97% species-level threshold. This suggests that the 16SrRNA gene might not be divergent enough to study the phylogeny of *Proteus* species. Unfortunately, only one species, *Proteus mirabilis*, has its full genome available on public databases, which prevents me from performing a full-genome phylogeny, which might have yielded more insight into the taxonomy of my candidates than the 16SrRNA gene alone.

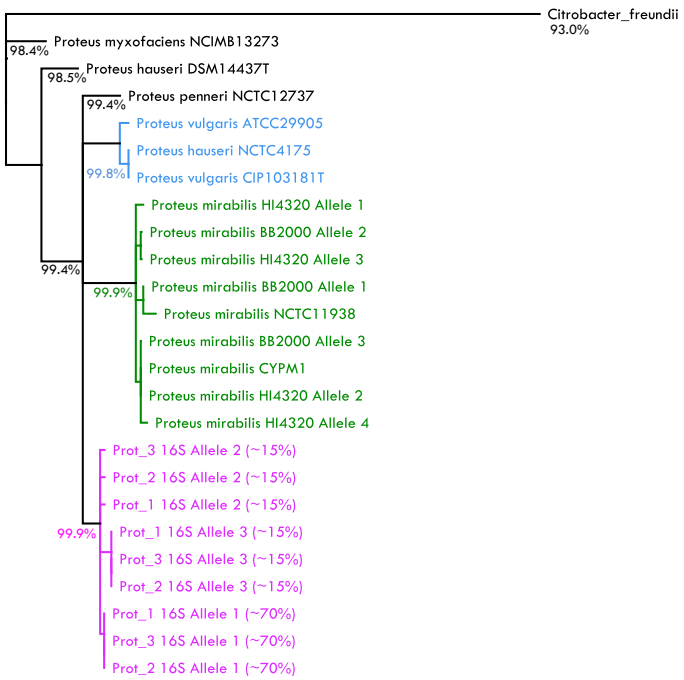

Figure 5: Neighbor-joining tree of the Proteus 16SrRNA gene of the isolated *Proteus* strains (Prot 1 to 3) obtained from the genome assemblies, and the *Proteus* references. The 1474 bp alignment corresponds to the amplification with the universal primers 27F and 1492R. *Citrobacter freundii* was used as outgroup. Bootstrapping was used with 10.000 iterations and a threshold of 50% support. The percentages refer to the average pairwise identity between members of the corresponding phylogenetic cluster.

Table 5: Pairwise identity of the *Proteus* 16S rRNA gene of the isolated *Proteus* strains (Prot 1 to 3) obtained from the genome assemblies, and the *Proteus* references. The table is based on a 1474 bp alignment of the 16S rRNA gene, corresponding to an amplification with the universal primers 27F and 1492R. *Citrobacter freundii* was used as outgroup. The upper triangle contains the number of SNPs while the lower triangle contains the percentage identity.

	Citrobacter_freundii	*Proteus myxofaciens NCIMB13273*	*Proteus hauseri DSM14437T*	*Proteus penneri NCTC12737*	*Proteus vulgaris ATCC29905*	*Proteus hauseri NCTC4175*	*Proteus vulgaris CIP103181T*	*Proteus mirabilis HI4320 Allele 1*	*Proteus mirabilis BB2000 Allele 2*	*Proteus mirabilis HI4320 Allele 3*	*Proteus mirabilis BB2000 Allele 1*	*Proteus mirabilis NCTC11938*	*Proteus mirabilis BB2000 Allele 3*	*Proteus mirabilis CYPM1*	*Proteus mirabilis HI4320 Allele 2*	*Proteus mirabilis HI4320 Allele 4*	Prot_3 16S Allele 2 (~15%)	Prot_2 16S Allele 2 (~15%)	Prot_1 16S Allele 2 (~15%)	Prot_1 16S Allele 3 (~15%)	Prot_3 16S Allele 3 (~15%)	Prot_2 16S Allele 3 (~15%)	Prot_1 16S Allele 1 (~70%)	Prot_3 16S Allele 1 (~70%)	Prot_2 16S Allele 1 (~70%)
Citrobacter_freundii		93	95	95	108	109	109	106	106	106	106	109	105	105	105	106	102	102	102	102	102	102	103	103	103
Proteus myxofaciens NCIMB13273	93.7		23	19	24	27	27	27	27	27	27	30	28	28	28	29	19	19	19	19	19	19	20	20	20
Proteus hauseri DSM14437T	93.6	98.5		24	23	24	24	23	24	24	23	26	22	22	22	23	19	19	19	19	19	19	20	20	20
Proteus penneri NCTC12737	93.1	98.6	98.3		7	10	10	11	10	10	11	14	12	12	12	13	6	6	6	7	7	7	6	6	6
Proteus vulgaris ATCC29905	92.7	98.4	98.4	99.5		5	5	14	15	15	14	17	15	15	15	16	10	10	10	10	10	10	9	9	9
Proteus hauseri NCTC4175	92.6	98.2	98.4	99.3	99.7		0	15	16	16	15	16	14	14	14	15	9	9	9	11	11	11	8	8	8
Proteus vulgaris CIP103181T	92.6	98.2	98.4	99.3	99.7	100.0		15	16	16	15	16	14	14	14	15	9	9	9	11	11	11	8	8	8
Proteus mirabilis HI4320 Allele 1	92.8	98.2	98.4	99.2	99.0	99.0	99.0		1	1	0	3	1	1	1	2	14	14	14	16	16	16	13	13	13
Proteus mirabilis BB2000 Allele 2	92.8	98.2	98.4	99.3	99.0	98.9	98.9	99.9		0	1	4	2	2	2	3	13	13	13	15	15	15	12	12	12
Proteus mirabilis HI4320 Allele 3	92.8	98.2	98.4	99.3	99.0	98.9	98.9	99.9	100.0		1	4	2	2	2	3	13	13	13	15	15	15	12	12	12
Proteus mirabilis BB2000 Allele 1	92.8	98.2	98.4	99.2	99.0	99.0	99.0	100.0	99.9	99.9		3	1	1	1	2	14	14	14	16	16	16	13	13	13
Proteus mirabilis NCTC11938	92.6	98.0	98.2	99.0	98.8	98.9	98.9	99.8	99.7	99.7	99.8		4	4	4	5	17	17	17	19	19	19	16	16	16
Proteus mirabilis BB2000 Allele 3	92.9	98.1	98.5	99.1	99.0	99.0	99.0	99.9	99.9	99.9	99.9	99.7		0	0	1	13	13	13	15	15	15	12	12	12
Proteus mirabilis CYPM1	92.9	98.1	98.5	99.1	99.0	99.0	99.0	99.9	99.9	99.9	99.9	99.7	100.0		0	1	13	13	13	15	15	15	12	12	12
Proteus mirabilis HI4320 Allele 2	92.9	98.1	98.5	99.1	99.0	99.0	99.0	99.9	99.9	99.9	99.9	99.7	100.0	100.0		1	13	13	13	15	15	15	12	12	12
Proteus mirabilis HI4320 Allele 4	92.8	98.0	98.4	99.1	98.9	99.0	99.0	99.9	99.8	99.8	99.9	99.7	99.9	99.9	99.9		14	14	14	16	16	16	13	13	13
Prot_3 16S Allele 2 (~15%)	93.1	98.7	98.7	99.6	99.3	99.4	99.4	99.0	99.1	99.1	99.0	98.8	99.1	99.1	99.1	99.0		0	0	2	2	2	1	1	1
Prot_2 16S Allele 2 (~15%)	93.1	98.7	98.7	99.6	99.3	99.4	99.4	99.0	99.1	99.1	99.0	98.8	99.1	99.1	99.1	99.0	100.0		0	2	2	2	1	1	1
Prot_1 16S Allele 2 (~15%)	93.1	98.7	98.7	99.6	99.3	99.4	99.4	99.0	99.1	99.1	99.0	98.8	99.1	99.1	99.1	99.0	100.0	100.0		2	2	2	1	1	1
Prot_1 16S Allele 3 (~15%)	93.1	98.7	98.7	99.5	99.3	99.3	99.3	98.9	99.0	99.0	98.9	98.7	99.0	99.0	99.0	98.9	99.9	99.9	99.9		0	0	3	3	3
Prot_3 16S Allele 3 (~15%)	93.1	98.7	98.7	99.5	99.3	99.3	99.3	98.9	99.0	99.0	98.9	98.7	99.0	99.0	99.0	98.9	99.9	99.9	99.9	100.0		0	3	3	3
Prot_2 16S Allele 3 (~15%)	93.1	98.7	98.7	99.5	99.3	99.3	99.3	98.9	99.0	99.0	98.9	98.7	99.0	99.0	99.0	98.9	99.9	99.9	99.9	100.0	100.0		3	3	3
Prot_1 16S Allele 1 (~70%)	93.0	98.7	98.6	99.6	99.4	99.5	99.5	99.1	99.2	99.2	99.1	98.9	99.2	99.2	99.2	99.1	99.9	99.9	99.9	99.8	99.8	99.8		0	0
Prot_3 16S Allele 1 (~70%)	93.0	98.7	98.6	99.6	99.4	99.5	99.5	99.1	99.2	99.2	99.1	98.9	99.2	99.2	99.2	99.1	99.9	99.9	99.9	99.8	99.8	99.8	100.0		0
Prot_2 16S Allele 1 (~70%)	93.0	98.7	98.6	99.6	99.4	99.5	99.5	99.1	99.2	99.2	99.1	98.9	99.2	99.2	99.2	99.1	99.9	99.9	99.9	99.8	99.8	99.8	100.0	100.0	

III. *Morganella*

For *Morganella*, I first amplified, cloned and sequenced a long fragment of the 16SrRNA gene using specific primers. With this method, I could identify two distinct strains (figure 6, table 6), one cluster at 99.1% identity to the *Morganella morganii* cluster, suggesting it might be a new strain of this species, while the other clusters at 98.4% identity of the *Morganella morganii* cluster, suggesting it might be a entirely new species of *Morganella*, since the other known species of *Morganella*, *Morganella psychrotolerans*, clusters also at 98.4% identity to the other *Morganella* sequences. This might suggest again that the 16SrRNA gene might be too similar within the *Morganella* genus to fully resolve the phylogeny, as known species appears to be closer than the usual species-level threshold of 97%.

Six colonies of *Morganella* could be isolated from my samples, of which I performed whole-genome sequencing, as for *Citrobacter* and *Proteus*. I again extracted the 16SrRNA gene from the assemblies in order to compare these isolated bacteria to the clone libraries. As for *Citrobacter* and *Proteus*, multiple copies of the 16SrRNA are present, and I could resolve between three and five alleles for each isolate. For some cases the phase could not be determined with 100% certainty, as the distance between the SNPs exceeded the read length, so I inferred it based on the variant frequencies. I found the same two cluster as previously (figure 7, table 7), with four isolates belonging to a potentially new species (Morg 2, 4, 5 & 6), and two isolates belonging to the *Morganella morganii* cluster (Morg 1 & 3). The last two might represent different strains. Moreover, these alleles cluster relatively well with the sequences obtained through PCR, cloning and sequencing.

To deepen the classification of my candidates, I again used ANDI (*Haubold, Klotzl et al. 2015*) to estimate the pairwise distance between whole-genomes, and used that estimate to build a phylogenetic tree in R (*Dixon 2003*, *Paradis, Claude et al. 2004*, *Schliep 2011*). With this method, both clusters are still visible, but the "new species" appear now between M. *morganii sibonii* and M. *morganii morganii*, which suggest that it might be a new subspecies of the M. *morganii* species rather than an entirely new species (figure 8, table 8). The two other isolates still cluster relatively close to M. *morganii morganii*, which suggest they might be new strain(s) of this subspecies.

To further understand the differences between my candidates, I annotated their genomes with RAST (*Aziz, Bartels et al. 2008*, *Overbeek, Olson et al. 2014*). First, I used the pairwise

distances between protein sequences of predicted genes (figure 9, table 9): each circular plot uses a different strain as reference, and the color refers to the percentage of protein identity, defined by open reading-frame. First, we can observe that Morg 1 and Morg 3 are relatively close to each other and to the different M. *morganii morganii* reference strains, but not as close as the various M. *morganii morganii* reference strains are to each other. This confirms that they are probably different strains of the same species, as they are much more distant to the other *Morganella* isolates, M. *morganii sibonii* and M. *psychrotolerans*. Interestingly, M. *morganii sibonii* seems equally distant to all M. *morganii* strains, while M. *psychrotolerans* is quite distant to all other proteomes. It is quite obvious at the protein level that Morg 2, 4, 5 & 6 represent the same strain, and probably a new subspecies of M. *morganii*, since they have nearly 100% identity with each other, and they are equally distant to all other M. *morganii* strains, but not as far as M. *psychrotolerans*.

Furthermore, I used the information from RAST concerning the gene content of my genomes, using the SEED system, and performed an ordination on the presence/absence data (figure 10). In the broader gene categories (Subcategories), as expected, Morg 2, 4, 5 & 6 cluster together and away from the other genomes, the three M. *morganii morganii* strains cluster together, and away from the other genomes, and Morg 1 & 3, M. *morganii sibonii* and M. *psychrotolerans* cluster together and away from the other genomes. Interestingly, when looking at more precise gene categories (Subsystems and Roles), all M. *morganii* subspecies collapse together, leaving only M. *psychrotolerans* and Morg 2, 4, 5 & 6 separate. This suggests that on a functional level, Morg 2, 4, 5 & 6 are as distant to the M. *morganii* cluster than they are to M. *psychrotolerans*.

Notably, although the three isolated strains (Morg 2456, Morg 1 and Morg 3) seem to be present in the samples, when comparing to the major *Morganella* OTU from the 16SrRNA library (figure 11), it appears that the indicator Otu000463, is closest to the new strain Morg2456 (3 SNPs) than to Morg 1 (6 SNPs) and Morg 3 (5 SNPs), suggesting that this new strain might be the candidate pathogen I identified. Moreover, theses new strains carry specific genes that are absent from nearly every other strains, some of which have a link to pathogenicity or toxicity (table 10), which reinforce it as potential source of infection/pathology.

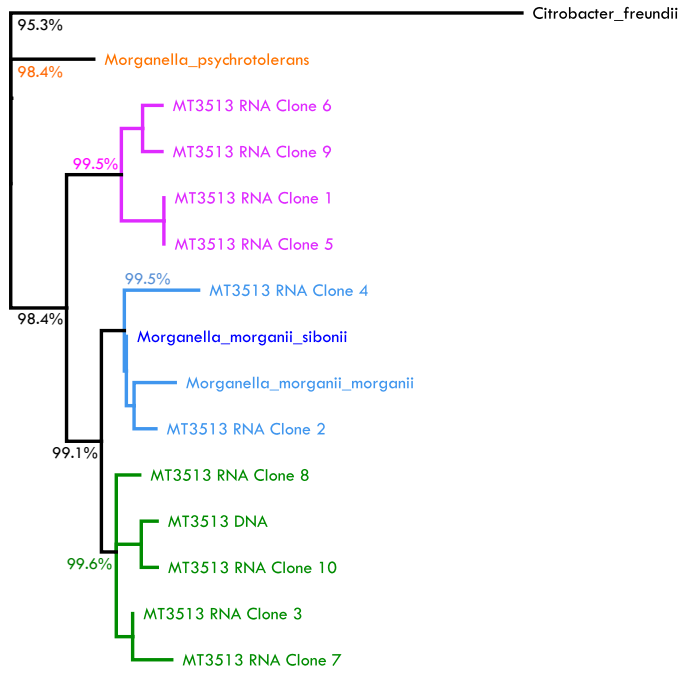

Figure 6: Neighbor-joining tree of the *Morganella* 16SrRNA gene of clones obtained from one sample (MT3513) and the *Morganella* references. The tree is based on a 713 bp alignment of the 16SrRNA gene, amplified with the primers ArsF and ArsR (*Hubert, Erban et al. 2015*). *Citrobacter freundii* was used as outgroup. Bootstrapping was used with 10.000 iterations and a threshold of 50% support. The percentages refer to the average pairwise identity between members of the corresponding phylogenetic cluster.

Table 6: Pairwise identity of the *Morganella* 16S rRNA gene of clones obtained from one sample (MT3513) and the *Morganella* references. The table is based on a 713 bp alignment of the 16S rRNA gene, amplified with the primers ArsF and ArsR (*Hubert, Erban et al.*). *Citrobacter freundii* was used as outgroup. The upper triangle contains the number of SNPs while the lower triangle contains the percentage identity.

	Citrobacter freundii	*Morganella psychrotolerans*	MT3513 RNA Clone 6	MT3513 RNA Clone 9	MT3513 RNA Clone 1	MT3513 RNA Clone 5	MT3513 RNA Clone 4	*Morganella morganii sibonii*	*Morganella morganii morganii*	MT3513 RNA Clone 2	MT3513 RNA Clone 8	MT3513 DNA	MT3513 RNA Clone 10	MT3513 RNA Clone 3	MT3513 RNA Clone 7
Citrobacter freundii		30	35	35	35	35	36	32	34	33	33	33	33	32	34
Morganella psychrotolerans	95.8		10	10	9	9	14	10	12	11	12	12	12	11	13
MT3513 RNA Clone 6	95.1	98.6		2	5	5	15	12	14	13	9	9	9	9	11
MT3513 RNA Clone 9	95.1	98.6	99.7		5	5	15	12	14	13	9	9	9	9	11
MT3513 RNA Clone 1	95.1	98.7	99.3	99.3		0	10	10	11	10	12	12	12	12	14
MT3513 RNA Clone 5	95.1	98.7	99.3	99.3	100.0		10	10	11	10	12	12	12	12	14
MT3513 RNA Clone 4	95.0	98.0	97.9	97.9	98.6	98.6		5	6	5	6	8	8	7	9
Morganella morganii sibonii	95.5	98.7	98.4	98.4	98.7	98.7	99.4		3	2	4	6	6	3	5
Morganella morganii morganii	95.2	98.3	98.0	98.0	98.5	98.5	99.2	99.6		3	6	8	8	5	7
MT3513 RNA Clone 2	95.4	98.5	98.2	98.2	98.6	98.6	99.3	99.8	99.6		5	7	7	4	6
MT3513 RNA Clone 8	95.4	98.3	98.7	98.7	98.3	98.3	99.2	99.5	99.2	99.3		2	2	1	3
MT3513 DNA	95.4	98.3	98.7	98.7	98.3	98.3	98.9	99.2	98.9	99.0	99.7		2	3	5
MT3513 RNA Clone 10	95.4	98.3	98.7	98.7	98.3	98.3	98.9	99.2	98.9	99.0	99.7	99.7		3	5
MT3513 RNA Clone 3	95.5	98.5	98.7	98.7	98.3	98.3	99.0	99.6	99.3	99.4	99.9	99.6	99.6		2
MT3513 RNA Clone 7	95.2	98.2	98.5	98.5	98.0	98.0	98.7	99.4	99.0	99.2	99.6	99.3	99.3	99.7	

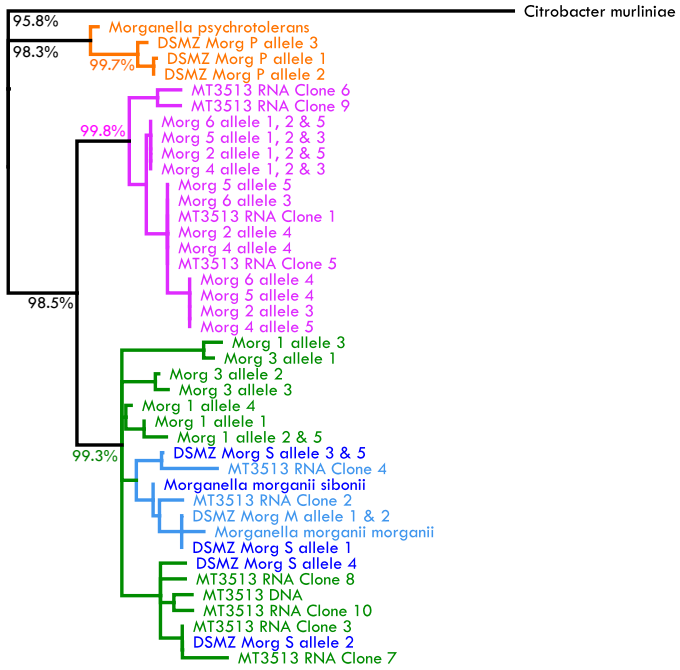

Figure 7: Neighbor-joining tree of the *Morganella* 16SrRNA gene of the isolated *Morganella* strains (Morg 1 to 6) obtained from the genome assemblies, the clones obtained from one sample (MT3513) and the *Morganella* references. The 713 bp alignment corresponds to the amplification with the primers ArsF and ArsR (*Hubert, Erban et al. 2015*). *Citrobacter murliniae* was used as outgroup. Bootstrapping was used with 10.000 iterations and a threshold of 50% support. The percentages refer to the average pairwise identity between members of the corresponding phylogenetic cluster.

Table 7: Pairwise identity of the *Morganella* 16S rRNA gene of the isolated *Morganella* strains (Morg 1 to 6) obtained from the genome assemblies, the clones obtained from one sample (MT3513), the DSMZ type strains and the *Morganella* references. The table is based on a 713 bp alignment of the 16S rRNA gene, amplified with the primers ArsF and ArsR (*Hubert, Erban et al.*). *Citrobacter freundii* was used as outgroup. The upper triangle contains the number of SNPs while the lower triangle contains the percentage identity.

	Citrobacter murliniae	*Morganella psychrotolerans*	DSMZ Morg P allele 3	DSMZ Morg P allele 1	DSMZ Morg P allele 2	MT3513 RNA Clone 6	MT3513 RNA Clone 9	Morg 6 allele 1, 2 & 5	Morg 5 allele 1, 2 & 3	Morg 2 allele 1, 2 & 5	Morg 4 allele 1, 2 & 3	Morg 5 allele 5	Morg 6 allele 3	MT3513 RNA Clone 1	Morg 2 allele 4	Morg 4 allele 4	MT3513 RNA Clone 5	Morg 6 allele 4	Morg 5 allele 4	Morg 2 allele 3	Morg 4 allele 5	Morg 1 allele 3	Morg 3 allele 1	Morg 3 allele 2	Morg 3 allele 3	Morg 1 allele 4	Morg 1 allele 1	Morg 1 allele 2 & 5	DSMZ Morg S allele 3 & 5	MT3513 RNA Clone 4	*Morganella morganii siboniii*	MT3513 RNA Clone 2	DSMZ Morg M allele 1 & 2	*Morganella morganii morganii*	DSMZ Morg S allele 1	DSMZ Morg S allele 4	MT3513 RNA Clone 8	MT3513 DNA	MT3513 RNA Clone 10	MT3513 RNA Clone 3	DSMZ Morg S allele 2	MT3513 RNA Clone 7
Citrobacter murliniae		26	29	30	29	31	31	30	30	30	30	31	31	31	31	31	31	32	32	32	32	31	33	29	30	29	29	30	29	32	28	29	29	30	29	29	29	29	29	28	28	30
Morganella psychrotolerans	96.4		3	4	3	10	10	8	8	8	8	9	9	9	9	9	9	10	10	10	10	12	14	12	13	10	10	11	11	14	10	11	11	12	11	12	12	12	12	11	11	13
DSMZ Morg P allele 3	95.9	99.6		1	2	12	12	10	10	10	10	11	11	11	11	11	11	12	12	12	12	15	16	14	15	12	12	13	12	15	13	14	14	15	14	13	13	13	13	14	14	16
DSMZ Morg P allele 1	95.8	99.4	99.9		1	12	12	10	10	10	10	11	11	11	11	11	11	12	12	12	12	16	16	14	15	12	12	13	13	16	14	15	15	16	15	14	14	14	14	15	15	17
DSMZ Morg P allele 2	95.9	99.6	99.7	99.9		12	12	10	10	10	10	11	11	11	11	11	11	12	12	12	12	15	16	14	15	12	12	13	14	17	13	14	14	15	14	15	15	15	15	14	14	16
MT3513 RNA Clone 6	95.7	98.6	98.3	98.3	98.3		2	4	4	4	4	5	5	5	5	5	5	6	6	6	6	14	12	6	7	8	9	10	12	15	12	13	13	14	13	9	9	9	9	9	9	11
MT3513 RNA Clone 9	95.7	98.6	98.3	98.3	98.3	99.7		4	4	4	4	5	5	5	5	5	5	6	6	6	6	14	12	6	7	8	9	10	12	15	12	13	13	14	13	9	9	9	9	9	9	11
Morg 6 allele 1, 2 & 5	95.8	98.9	98.6	98.6	98.6	99.4	99.4		0	0	0	1	1	1	1	1	1	2	2	2	2	12	10	8	9	6	7	8	8	11	9	9	9	10	9	11	11	11	11	11	11	13
Morg 5 allele 1, 2 & 3	95.8	98.9	98.6	98.6	98.6	99.4	99.4	100.0		0	0	1	1	1	1	1	1	2	2	2	2	12	10	8	9	6	7	8	8	11	9	9	9	10	9	11	11	11	11	11	11	13
Morg 2 allele 1, 2 & 5	95.8	98.9	98.6	98.6	98.6	99.4	99.4	100.0	100.0		0	1	1	1	1	1	1	2	2	2	2	12	10	8	9	6	7	8	8	11	9	9	9	10	9	11	11	11	11	11	11	13
Morg 4 allele 1, 2 & 3	95.8	98.9	98.6	98.6	98.6	99.4	99.4	100.0	100.0	100.0		1	1	1	1	1	1	2	2	2	2	12	10	8	9	6	7	8	8	11	9	9	9	10	9	11	11	11	11	11	11	13
Morg 5 allele 5	95.7	98.7	98.5	98.5	98.5	99.3	99.3	99.9	99.9	99.9	99.9		0	0	0	0	0	1	1	1	1	12	10	9	8	7	8	9	9	10	10	10	10	11	10	12	12	12	12	12	12	14
Morg 6 allele 3	95.7	98.7	98.5	98.5	98.5	99.3	99.3	99.9	99.9	99.9	99.9	100.0		0	0	0	0	1	1	1	1	12	10	9	8	7	8	9	9	10	10	10	10	11	10	12	12	12	12	12	12	14
MT3513 RNA Clone 1	95.7	98.7	98.5	98.5	98.5	99.3	99.3	99.9	99.9	99.9	99.9	100.0	100.0		0	0	0	1	1	1	1	12	10	9	8	7	8	9	9	10	10	10	10	11	10	12	12	12	12	12	12	14
Morg 2 allele 4	95.7	98.7	98.5	98.5	98.5	99.3	99.3	99.9	99.9	99.9	99.9	100.0	100.0	100.0		0	0	1	1	1	1	12	10	9	8	7	8	9	9	10	10	10	10	11	10	12	12	12	12	12	12	14
Morg 4 allele 4	95.7	98.7	98.5	98.5	98.5	99.3	99.3	99.9	99.9	99.9	99.9	100.0	100.0	100.0	100.0		0	1	1	1	1	12	10	9	8	7	8	9	9	10	10	10	10	11	10	12	12	12	12	12	12	14
MT3513 RNA Clone 5	95.7	98.7	98.5	98.5	98.5	99.3	99.3	99.9	99.9	99.9	99.9	100.0	100.0	100.0	100.0	100.0		1	1	1	1	12	10	9	8	7	8	9	9	10	10	10	10	11	10	12	12	12	12	12	12	14
Morg 6 allele 4	95.5	98.6	98.3	98.3	98.3	99.2	99.2	99.7	99.7	99.7	99.7	99.9	99.9	99.9	99.9	99.9	99.9		0	0	0	13	11	10	9	8	9	10	10	11	11	11	11	12	11	13	13	13	13	13	13	15
Morg 5 allele 4	95.5	98.6	98.3	98.3	98.3	99.2	99.2	99.7	99.7	99.7	99.7	99.9	99.9	99.9	99.9	99.9	99.9	100.0		0	0	13	11	10	9	8	9	10	10	11	11	11	11	12	11	13	13	13	13	13	13	15
Morg 2 allele 3	95.5	98.6	98.3	98.3	98.3	99.2	99.2	99.7	99.7	99.7	99.7	99.9	99.9	99.9	99.9	99.9	99.9	100.0	100.0		0	13	11	10	9	8	9	10	10	11	11	11	11	12	11	13	13	13	13	13	13	15
Morg 4 allele 5	95.5	98.6	98.3	98.3	98.3	99.2	99.2	99.7	99.7	99.7	99.7	99.9	99.9	99.9	99.9	99.9	99.9	100.0	100.0	100.0		13	11	10	9	8	9	10	10	11	11	11	11	12	11	13	13	13	13	13	13	15
Morg 1 allele 3	95.7	98.3	97.9	97.8	97.9	98.0	98.0	98.3	98.3	98.3	98.3	98.3	98.3	98.3	98.3	98.3	98.3	98.2	98.2	98.2	98.2		2	8	8	6	6	7	7	9	6	7	7	8	7	8	8	8	8	7	7	9
Morg 3 allele 1	95.4	98.0	97.8	97.8	97.8	98.3	98.3	98.6	98.6	98.6	98.6	98.6	98.6	98.6	98.6	98.6	98.6	98.5	98.5	98.5	98.5	99.7		6	6	4	5	6	8	10	8	9	9	10	9	9	9	9	9	9	9	11
Morg 3 allele 2	95.9	98.3	98.0	98.0	98.0	99.2	99.2	98.9	98.9	98.9	98.9	98.7	98.7	98.7	98.7	98.7	98.7	98.6	98.6	98.6	98.6	98.9	99.2		1	2	3	4	6	9	6	7	7	8	7	3	3	3	3	3	3	5
Morg 3 allele 3	95.8	98.2	97.9	97.9	97.9	99.0	99.0	98.7	98.7	98.7	98.7	98.9	98.9	98.9	98.9	98.9	98.9	98.7	98.7	98.7	98.7	98.9	99.2	99.9		3	4	5	7	8	7	8	8	9	8	4	4	4	4	4	4	6
Morg 1 allele 4	95.9	98.6	98.3	98.3	98.3	98.9	98.9	99.2	99.2	99.2	99.0	99.0	99.0	99.0	99.0	99.0	98.9	98.9	98.9	98.9	98.9	99.2	99.4	99.7	99.6		1	2	4	7	4	5	5	6	5	5	5	5	5	5	5	5	7
Morg 1 allele 1	95.9	98.6	98.3	98.3	98.3	98.7	98.7	99.0	99.0	99.0	99.0	98.9	98.9	98.9	98.9	98.9	98.9	98.7	98.7	98.7	98.7	99.2	99.3	99.6	99.4	99.9		1	4	7	4	5	5	6	5	5	5	5	5	5	5	7
Morg 1 allele 2 & 5	95.8	98.5	98.2	98.2	98.2	98.6	98.6	98.9	98.9	98.9	98.9	98.7	98.7	98.7	98.7	98.7	98.7	98.6	98.6	98.6	98.6	99.0	99.2	99.4	99.3	99.7	99.9		5	8	5	6	6	7	6	6	6	6	6	6	6	8
DSMZ Morg S allele 3 & 5	95.9	98.5	98.3	98.2	98.0	98.3	98.3	98.9	98.9	98.9	98.9	98.7	98.7	98.7	98.7	98.7	98.7	98.6	98.6	98.6	98.6	99.0	98.9	99.2	99.0	99.4	99.4	99.3		3	2	2	2	3	2	3	3	5	5	4	4	6
MT3513 RNA Clone 4	95.5	98.0	97.9	97.8	97.6	97.9	97.9	98.5	98.5	98.5	98.5	98.6	98.6	98.6	98.6	98.6	98.6	98.5	98.5	98.5	98.5	98.7	98.6	98.7	98.9	99.0	99.0	98.9	99.6		5	5	5	6	5	6	6	8	8	7	7	9
Morganella morganii siboniii	96.1	98.7	98.2	98.1	98.2	98.4	98.4	98.8	98.8	98.8	98.8	98.7	98.7	98.7	98.7	98.7	98.7	98.5	98.5	98.5	98.5	99.2	98.9	99.2	99.1	99.5	99.5	99.4	99.8	99.4		2	2	3	2	4	4	6	6	3	3	5
MT3513 RNA Clone 2	95.9	98.5	98.0	97.9	98.0	98.2	98.2	98.7	98.7	98.7	98.7	98.6	98.6	98.6	98.6	98.6	98.6	98.5	98.5	98.5	98.5	99.0	98.7	99.0	98.9	99.3	99.3	99.2	99.7	99.3	99.8		2	3	2	5	5	7	7	4	4	6
DSMZ Morg M allele 1 & 2	95.9	98.5	98.0	97.9	98.0	98.2	98.2	98.7	98.7	98.7	98.7	98.6	98.6	98.6	98.6	98.6	98.6	98.5	98.5	98.5	98.5	99.0	98.7	99.0	98.9	99.3	99.3	99.2	99.7	99.3	99.8	99.7		1	0	5	5	7	7	4	4	6
Morganella morganii morganii	95.8	98.3	97.9	97.8	97.9	98.0	98.0	98.6	98.6	98.6	98.6	98.5	98.5	98.5	98.5	98.5	98.5	98.3	98.3	98.3	98.3	98.9	98.6	98.9	98.7	99.2	99.2	99.0	99.6	99.2	99.6	99.6	99.9		1	6	6	8	8	5	5	7
DSMZ Morg S allele 1	95.9	98.5	98.0	97.9	98.0	98.2	98.2	98.7	98.7	98.7	98.7	98.6	98.6	98.6	98.6	98.6	98.6	98.5	98.5	98.5	98.5	99.0	98.7	99.0	98.9	99.3	99.3	99.2	99.7	99.3	99.8	99.7	100.0	99.9		5	5	7	7	4	4	6
DSMZ Morg S allele 4	95.9	98.3	98.2	98.0	97.9	98.7	98.7	98.5	98.5	98.5	98.5	98.3	98.3	98.3	98.3	98.3	98.3	98.2	98.2	98.2	98.2	98.9	98.7	99.6	99.4	99.3	99.3	99.2	99.6	99.2	99.5	99.3	99.3	99.2	99.3		0	2	2	1	1	3
MT3513 RNA Clone 8	95.9	98.3	98.2	98.0	97.9	98.7	98.7	98.5	98.5	98.5	98.5	98.3	98.3	98.3	98.3	98.3	98.3	98.2	98.2	98.2	98.2	98.9	98.7	99.6	99.4	99.3	99.3	99.2	99.6	99.2	99.5	99.3	99.3	99.2	99.3	100.0		2	2	1	1	3
MT3513 DNA	95.9	98.3	98.2	98.0	97.9	98.7	98.7	98.5	98.5	98.5	98.5	98.3	98.3	98.3	98.3	98.3	98.3	98.2	98.2	98.2	98.2	98.9	98.7	99.6	99.4	99.3	99.3	99.2	99.6	99.2	99.5	99.3	99.3	99.2	99.3	99.7	99.7		2	3	3	5
MT3513 RNA Clone 10	95.9	98.3	98.2	98.0	97.9	98.7	98.7	98.5	98.5	98.5	98.5	98.3	98.3	98.3	98.3	98.3	98.3	98.2	98.2	98.2	98.2	98.9	98.7	99.6	99.4	99.3	99.3	99.2	99.6	99.2	99.5	99.3	99.3	99.2	99.3	99.7	99.7	99.7		3	3	5
MT3513 RNA Clone 3	96.1	98.5	98.0	97.9	98.0	98.7	98.7	98.5	98.5	98.5	98.5	98.3	98.3	98.3	98.3	98.3	98.3	98.2	98.2	98.2	98.2	99.0	98.7	99.6	99.4	99.3	99.3	99.2	99.4	99.0	99.6	99.4	99.4	99.3	99.4	99.9	99.9	99.6	99.6		0	2
DSMZ Morg S allele 2	96.1	98.5	98.0	97.9	98.0	98.7	98.7	98.5	98.5	98.5	98.5	98.3	98.3	98.3	98.3	98.3	98.3	98.2	98.2	98.2	98.2	99.0	98.7	99.6	99.4	99.3	99.3	99.2	99.4	99.0	99.6	99.4	99.4	99.3	99.4	99.9	99.9	99.6	99.6	100.0		2
MT3513 RNA Clone 7	95.8	98.2	97.8	97.6	97.8	98.5	98.5	98.2	98.2	98.2	98.2	98.0	98.0	98.0	98.0	98.0	98.0	97.9	97.9	97.9	97.9	98.7	98.5	99.3	99.2	99.0	99.0	98.9	99.2	98.7	99.4	99.2	99.2	99.0	99.2	99.6	99.6	99.3	99.3	99.7	99.7	

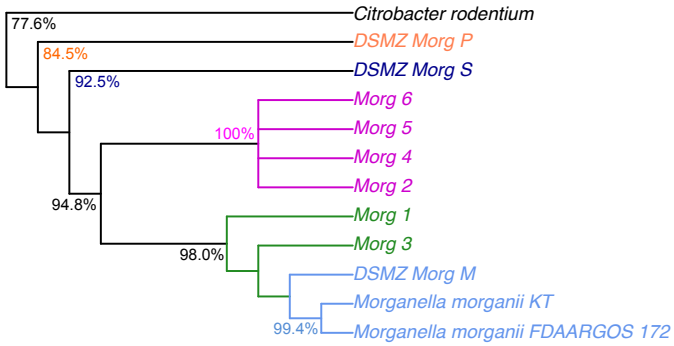

Figure 8: Neighbor-joining tree of the isolated *Morganella* strains and the *Morganella* references, based on the whole genome identities estimated by ANDI (*Haubold, Klotzl et al.*) *Citrobacter rodentium* was used as outgroup. Bootstrapping was used with 10.000 iterations and a threshold of 50% support. The percentages refer to the average pairwise identity between members of the corresponding phylogenetic cluster.

Table 8: Pairwise identity of the isolated *Morganella* strains, the DSMZ type strains and the *Morganella* references, based on the whole genome identities estimated by ANDI (*Haubold, Klotzl et al. 2015*) *Morganella morganii morganii* KT was used as outgroup.

	Citrobacter rodentium	DSMZ Morg P	DSMZ Morg S	Morg 6	Morg 5	Morg 4	Morg 2	Morg 1	Morg 3	DSMZ Morg M	*Morganella morganii KT*
DSMZ Morg P	76.2										
DSMZ Morg S	77.8	84.4									
Morg 6	78.0	84.5	92.5								
Morg 5	78.1	84.6	92.5	100.0							
Morg 4	78.0	84.6	92.5	100.0	100.0						
Morg 2	78.0	84.5	92.5	100.0	100.0	100.0					
Morg 1	77.5	84.5	92.5	94.7	94.7	94.7	94.7				
Morg 3	77.9	84.6	92.5	94.8	94.8	94.8	94.8	97.2			
DSMZ Morg M	77.9	84.6	92.5	94.8	94.8	94.8	94.8	97.2	97.6		
Morganella morganii KT	77.3	84.6	92.6	94.8	94.7	94.8	94.8	97.1	97.6	99.1	
Morganella morganii FDAARGOS 172	77.2	84.6	92.6	94.8	94.8	94.8	94.8	97.1	97.7	99.1	99.9

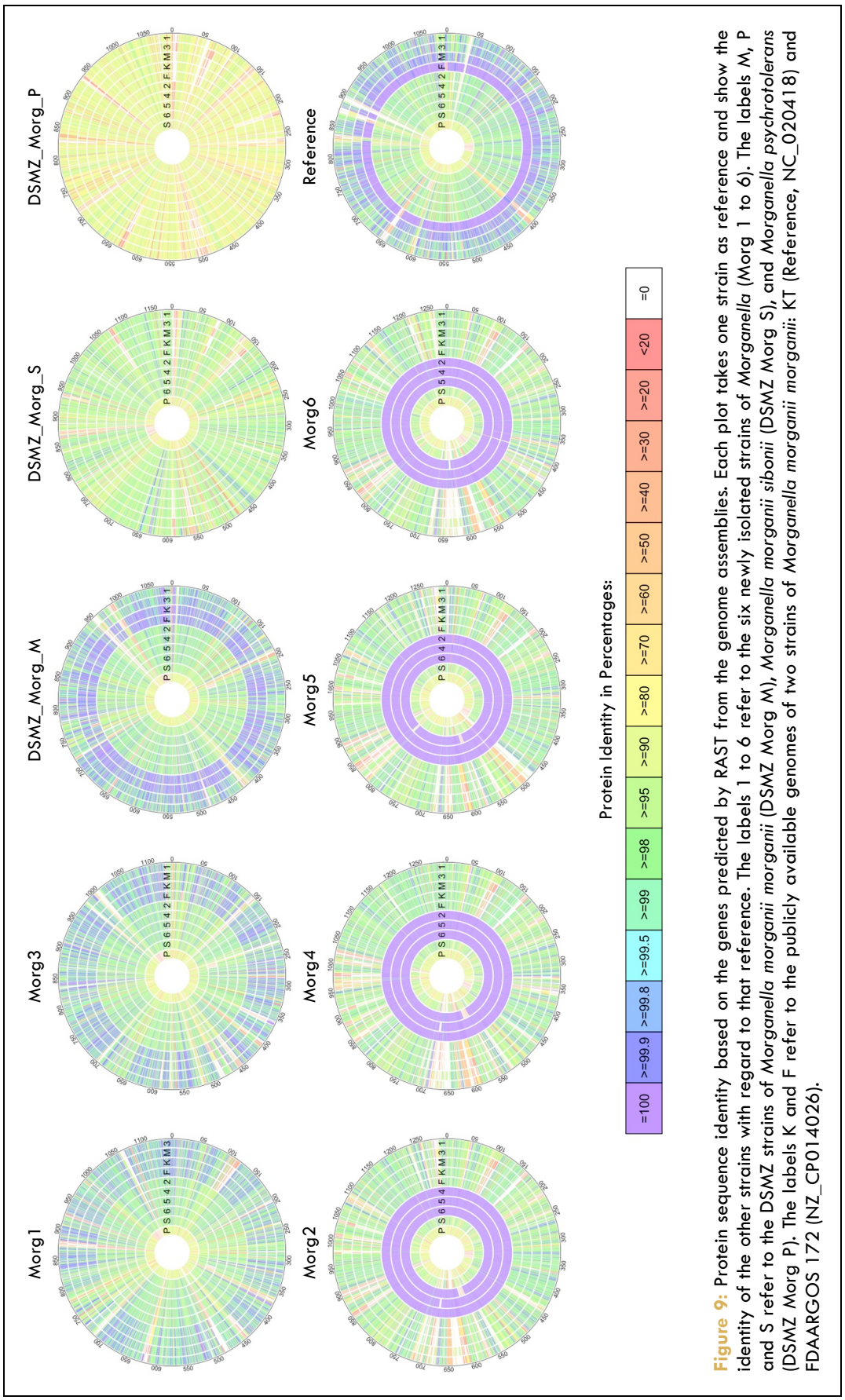

Figure 9: Protein sequence identity based on the genes predicted by RAST from the genome assemblies. Each plot takes one strain as reference and show the identity of the other strains with regard to that reference. The labels 1 to 6 refer to the six newly isolated strains of *Morganella* (Morg 1 to 6). The labels M, P and S refer to the DSMZ strains of *Morganella morganii morganii* (DSMZ Morg M), *Morganella morganii sibonii* (DSMZ Morg S), and *Morganella psychrotolerans* (DSMZ Morg P). The labels K and F refer to the publicly available genomes of two strains of *Morganella morganii morganii*: KT (Reference, NC_020418) and FDAARGOS 172 (NZ_CP014026).

Table 9: Pairwise identity of the *Morganella* proteome from the isolated *Morganella* strains, the DSMZ type strains and the *Morganella* references. These date are based on the genes predicted by RAST based on the genome assemblies.

	Morg_p	Morg_s	Morg_2	Morg_4	Morg_5	Morg_6	Morg_1	Morg_3	Ref_K	Ref_F	Morg_m
Morg_p		71.67	70.11	70.28	70.38	69.40	69.51	69.24	68.03	68.88	68.18
Morg_s	66.33		80.67	80.76	80.75	79.73	79.66	78.86	77.67	78.36	77.47
Morg_2	58.97	73.02		99.52	99.09	97.62	74.03	73.64	70.38	71.70	70.12
Morg_4	58.58	72.68	99.38		98.97	97.48	73.53	73.18	70.28	71.62	69.88
Morg_5	58.82	72.91	99.25	99.27		97.44	73.81	73.38	70.63	71.98	70.18
Morg_6	58.81	73.01	99.25	99.23	98.86		73.96	73.45	70.46	71.82	70.12
Morg_1	65.24	80.72	82.83	82.97	82.99	81.91		84.20	82.18	83.72	82.17
Morg_3	66.42	80.99	83.38	83.44	83.41	82.28	85.49		84.69	84.95	85.29
Ref_K	67.96	83.78	85.01	84.96	85.05	83.75	87.76	88.94		94.99	92.28
Ref_F	66.75	82.22	83.08	83.07	83.16	81.91	86.20	86.23	92.45		89.97
Morg_m	68.65	84.06	84.68	84.83	84.86	83.65	87.63	89.83	92.41	93.33	

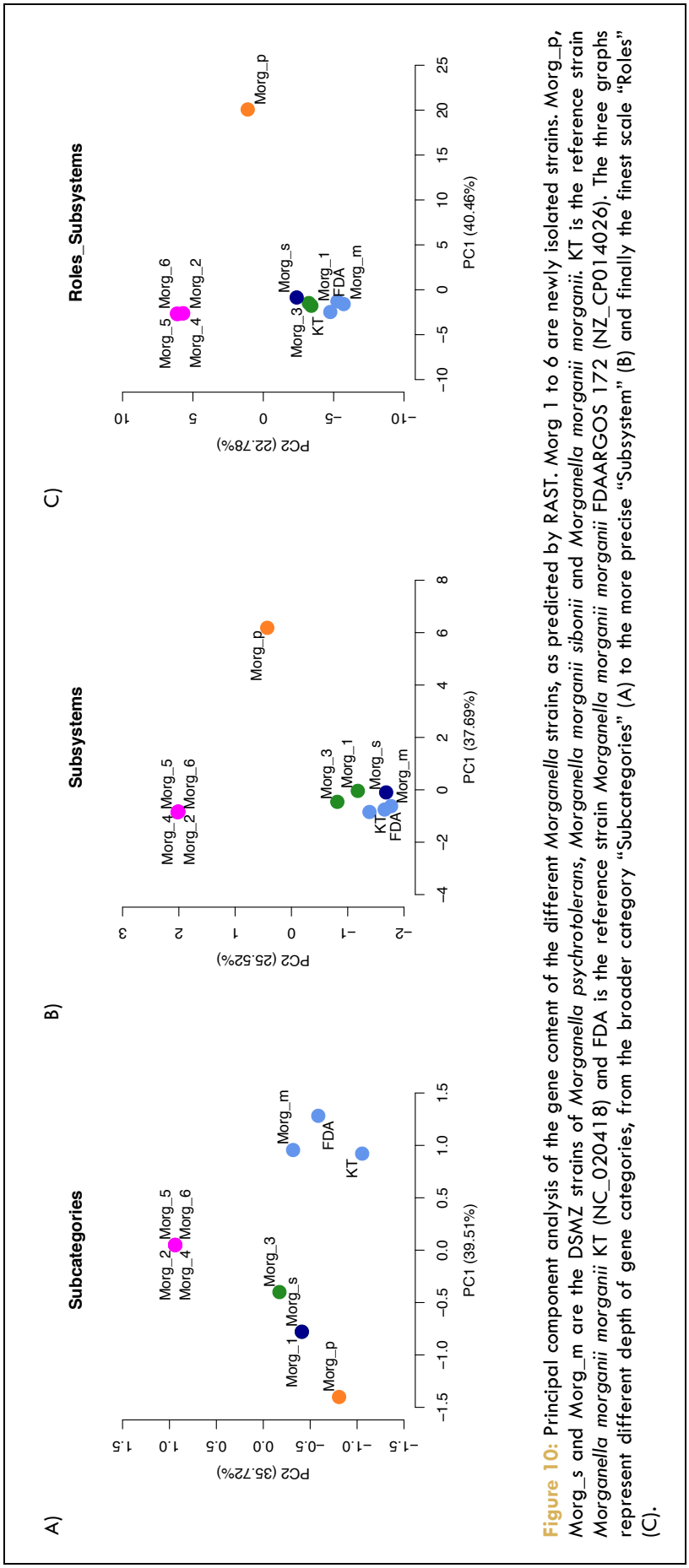

Figure 10: Principal component analysis of the gene content of the different *Morganella* strains, as predicted by RAST. Morg 1 to 6 are newly isolated strains. Morg_p, Morg_s and Morg_m are the DSMZ strains of *Morganella psychrotolerans*, *Morganella morganii sibonii* and *Morganella morganii morganii*. KT is the reference strain *Morganella morganii morganii* KT (NC_020418) and FDA is the reference strain *Morganella morganii morganii* FDAARGOS 172 (NZ_CP014026). The three graphs represent different depth of gene categories, from the broader category "Subcategories" (A) to the more precise "Subsystem" (B) and finally the finest scale "Roles" (C).

Figure 11: Alignment of the *Morganella* OTU sequence (Otu000463, used as reference) from the 16SrRNA library to the corresponding sequences of the isolated clones of *Morganella* (the consensus sequence was used). Sequence identity is displayed in grey while colored bars represent SNPs.

Table 10: Subset of gene functions present only in the Morg 2456 isolates or in *Morganella morganii morganii* that are related to pathogenicity or toxicity. The functional categories are those from the RAST annotation system (*Aziz, Bartels et al. 2008*, *Overbeek, Olson et al. 2014*).

Category	Subcategory	Subsystem	Role	Specificity
Regulation and Cell signaling	Programmed Cell Death and Toxin antitoxin Systems	Toxin antitoxin replicon stabilization systems	HigB toxin protein	Morg2456
Regulation and Cell signaling	Programmed Cell Death and Toxin antitoxin Systems	Toxin antitoxin replicon stabilization systems	RelE/StbE replicon stabilization toxin	Morg2456
Regulation and Cell signaling	Programmed Cell Death and Toxin antitoxin Systems	Toxin antitoxin replicon stabilization systems	YkfI toxin protein	Morg2456
Phages Prophages Transposable elements Plasmids	Pathogenicity islands	Listeria Pathogenicity Island LIPI 1 extended	Phosphatidylinositol specific phospholipase C (EC 4.6.1.13)	Morg2456 + *Morganella morganii morganii*

Conclusion

In Chapter II, I used 16S rRNA gene metabarcoding of the microbial communities that I compared to various attributes of the mice, which lead me to identify several promising candidate pathogens. Among these candidate pathogens, *Morganella*, *Proteus* and *Citrobacter*, represent promising cases. These results are however only correlations, which need to be confirmed by experiments.

To tackle this issue, I used two methods to deepen the phylogenetic identification of my candidate pathogens. First by sequencing a nearly full-length 16S rRNA gene, second, by whole genome sequencing. These methods allowed me to identify the candidate *Citrobacter* as a member of the *Citrobacter freundii* species, while the candidate *Morganella* represents a new species, potentially a new subspecies of *Morganella morganii*. Unfortunately, I couldn't resolve the phylogeny of my candidate *Proteus*, due to a lack of publicly available reference genomes. Interestingly, for *Morganella*, it seems that the new subspecies represents the best candidate, as the main OTU from the V1-V2 16S rRNA library is closest to that strain -- although this should be interpreted with caution since it is a quite short fragment to infer phylogeny within a relatively lowly divergent genus -- and it carries specific genes with toxicity-related functions.

These results certainly increase our confidence in the role that *Morganella* could play in maintaining diversity at *B4galnt2* in wild populations of house mice, but more experimental evidence are needed. Our collaborator Guntram Grassl at Hannover Medical School is currently in the process of confirming the pathogenic behavior of *Morganella*, via different methods. First through the *Morganella*-specific antibody staining of my intestinal histology slides, to 1) confirm the presence of *Morganella* detected by sequencing, and 2) determine the localization of *Morganella* in the cecum; second, by the infection of organoids by the *Morganella* strains isolated from my samples; finally by the in vivo infection experiment, also with my *Morganella* isolates.

Of note, the preliminary results of the antibody staining of *Morganella* in the cecum histological slides show that 1) *Morganella* was detected in an inflamed C57BL/6J individual for which I detected the bacteria by sequencing, while no signal was detected in a sample without *Morganella* 16SrRNA sequences, suggesting that the sequencing results are reliable; 2) in this individual, *Morganella* was present inside the epithelium, which is a typical pathogenic behavior, as commensal stay in the mucus layer. Of course this is only one individual, but it supports the potential pathogenic behavior of *Morganella*.

Methods

To further characterize the candidate pathogens identified with the microbial community analysis in Chapter II, I first sequenced a longer part of the 16S rRNA gene to obtain better phylogenic resolution than with only the V1-V2 region. Then, I performed whole-genome sequencing of the candidates that were isolated from my samples by Guntram Grassl.

I. 16S rRNA gene sequencing

For the 16S rRNA sequencing, I used primers from the literature (*Hubert, Erban et al. 2015*) and primers that I designed using the online tool Primer3 (*Koressaar and Remm 2007*, *Untergasser, Cutcutache et al. 2012*), the list of primers is presented in table 11. I tested the specificity of my primer pairs using the *in silico* PCR from the RDP website (*Cole, Wang et al. 2014*). The PCR was performed using the GoTaq DNA Polymerase from Promega, with the following protocol: 0.2µL of 10 µM primers, 0.2µL dNTPs, 0.8µL $MgCl_2$, 0.1µL Taq, 5.5µL H_2O, with the PCR program presented in table 12.

As positive controls, I ordered DNA from the three *Morganella* type strains and from the *Citrobacter freundii* type strain from the "*Deutsche Sammlung von Mikroorganismen und Zellkulturen*" (DSMZ), which I processed together with my samples and in the same experimental conditions (table 13).

The PCR product was diluted to ~30ng/µl and used as template for cloning using the CloneJet PCR cloning kit from ThermoScientific and One Shot TOP10 Chemically Competent E. coli from Invitrogen. After 24h growth at 37°C, the colonies were picked and directly added into the PCR reaction mix, using the external primer pairs (ArsF-ArsR2 or Morg1F-Morg2R for *Morganella* and 27F-ArsR2 for *Citrobacter*). Then nested PCR was performed using the internal primers. All PCR products were treated with ExoSap and sequenced using the appropriate primer following the same protocols as for the *B4galnt2* genotyping. I edited and assembled the sequences in GENEIOUS 7.0 (Biomatters Ltd).

Table 11: Summary of the primer pairs used to sequence nearly full-length 16S rDNA genes from the candidate bacteria. ArsF and ArsR were obtained from (*Hubert, Erban et al. 2015*) while other primers were newly designed.

Primer F		Primer R		Amplicon Size (bp)	Position (bp)	Target	Specificity
ArsF	GGGTTGTAAAGTACTTTCAGTCGT	ArsR2	GTAGCCCTRCTCGTAAGGGCC	805	416-1220	*Morganella*	*Morganella*
ArsF	GGGTTGTAAAGTACTTTCAGTCGT	Morg2R	TGAGTTTTAACCTTGCGGCC	491	416-906	Internal	--
Morg7F	ATTCGATGCAACGCGAAGAA	ArsR2	GTAGCCCTRCTCGTAAGGGCC	269	952-1220	Internal	--
Morg2F	CCTAACACATGCAAGTCGGG	Morg1R	CGCTTCCCTTTGTATACGCC	1215	40-1254	*Morganella*	*Morganella* + *Clostridium*
Morg2F	CCTAACACATGCAAGTCGGG	338R	TGCTGCCTCCCGTAGGAGT	310	40-349	Internal	--
Morg7F	ATTCGATGCAACGCGAAGAA	Morg1R	CGCTTCCCTTTGTATACGCC	303	952-1254	Internal	--
27F	AGAGTTTGATCCNTGGCTCAG	ArsR2	GTAGCCCTRCTCGTAAGGGCC	1220	2-1221	*Citrobacter*	*Citrobacter* + *Escherichia-Shigella* + *Bacteroides*
27F	AGAGTTTGATCCNTGGCTCAG	338R	TGCTGCCTCCCGTAGGAGT	350	2-351	Internal	--
Citro11F	ATTCGATGCAACGCGAAGAA	ArsR2	GTAGCCCTRCTCGTAAGGGCC	268	954-1221	Internal	--
Citro2F	GGAGGGTGCAAGCGTTAATC	Citro4R	TGAGTTTTAACCTTGCGGCC	377	532-908	Internal	--
Citro5F	TCCAGGTGTAGCGGTGAAAT	Citro3R	CCCCACCTTCCTCCAGTTTA	510	673-1182	Internal	--

I used the Geneious Tree Builder to assess the phylogeny of my candidates, using the HKY genetic distance model and the Neighbor-joining tree building method. I used the reference sequences of the *Morganella* and *Citrobacter* type strains obtained from StrainInfo (*Verslyppe, De Smet et al. 2014*) and summarized in table 13. I used the bootstrap method for resampling of the tree with 10,000 iterations and a threshold of 50% support, using *Citrobacter freundii* as an outgroup for *Morganella* and *Morganella morganii morganii* as an outgroup for *Citrobacter*.

Table 12: Amplification program for candidate-specific PCR.

Temperature	Time	Cycles
98°C	5 min	
98°C	30 sec	x 35
62°C	90 sec	
72°C	90 sec	
72°C	10 min	
12°C	∞	

Table 13: Summary of the reference 16S rDNA genes used for phylogenetic analysis, together with the reference strains obtained from the "Deutsche Sammlung von Mikroorganismen und Zellkulturen" (DSM).

Genus	Species	Strain	Gene Identification DSMZ Cat. No.
Citrobacter	*Amalonaticus*	--	334084827
Citrobacter	*Braakii*	--	573008384
Citrobacter	*Diversus*	--	3169782
Citrobacter	*Farmeri*	--	3169781
Citrobacter	*Freundii*	--	4581981
Citrobacter	*Gillenii*	--	3169777
Citrobacter	*Koseri*	--	325975904
Citrobacter	*Murliniae*	--	3169779
Citrobacter	*Rodentium*	--	3169773
Citrobacter	*Sedlakii*	--	3169774
Citrobacter	*Werkmanii*	--	3169783
Citrobacter	*Youngae*	--	157073727
Citrobacter	*Freundii*	--	DSM 30039
Proteus	*Hauseri*	DSM14437T	FR733709
Proteus	*Hauseri*	NCT4175	DQ885262
Proteus	*Mirabilis*	--	DQ885256
Proteus	*Myxofaciens*	--	DQ885259
Proteus	*Penneri*	--	DQ885258
Proteus	*Vulgaris*	ATCC29905	DQ885257
Proteus	*Vulgaris*	CIP103181T	AJ301683
Morganella	*Psychrotolerans*	--	86451969
Morganella	*Morganii morganii*	--	15551726
Morganella	*Morganii sibonii*	--	86451980
Morganella	*Psychrotolerans*	--	DSM 17886
Morganella	*Morganii morganii*	--	DSM 30164
Morganella	*Morganii sibonii*	--	DSM 14850

II. Isolation of candidate bacteria

Our collaboration partner Guntram Grassl from Hannover Medical School isolated single colonies of my candidate pathogens from the cecum samples I collected in the wild. I selected the samples based on the abundance of the candidate pathogen relative to other *Enterobacteriaceae* to maximize the chances of recovery.

For the isolation of *Citrobacter*, the following protocol was used:
Cecum2 solutions were diluted 1:10 in PBS/40% glycerol and homogenized. Serial dilutions were plated on MacConkey and Columbia Blood agar (aerobic and anaerobic cultivation over night at 37°C). Single colonies were picked on MacConkey and blood agar (aerobic and anaerobic cultivation over night at 37°C). Fresh colonies were identified using MALDI-TOF: 6 strains belonged to E. coli and 3 to *Citrobacter*.

For the isolation of *Morganella* and *Proteus*, the following protocol was used:
Cecum2 solutions were diluted 1:10 in PBS/40% glycerol and homogenized. 5µl of homogenate was inoculated in 2.5ml of LB medium containing the following antibiotics either alone or combined: oxacillin (20µg/ml), vancomycin (20µg/ml), erythromycin (20µg/ml), and cefaclor (20µg/ml). The liquid cultures were incubated over night at 37°C with shaking. Serial dilutions were plated on MacConkey, LB and Columbia Blood agar. Single colonies were picked and spread on MacConkey and Columbia Blood agar. Fresh colonies were identified using MALDI-TOF: 3 strains belonged to *Proteus* and 6 to *Morganella*, which displayed 2 different colony morphologies: rough and round colonies stemming from the initial selection with all 4 antibiotics; lighter colonies stemming from the initial selection with cefaclor only. *Proteus* colonies also came from the initial selection with cefaclor only.

III. Whole genome sequencing of the isolated candidates

I prepared the library for whole genome sequencing using the Nextera XT DNA Library Prep kit from Illumina. The multiplexed library ran on an Illumina MiSeq, with the V2 kit with 2x250 bp read length. I demultiplexed the output using CASAVA (Illumina) allowing no mismatches in the barcodes, and using the "eamss" algorithm. I merged forward and reverse

reads using usearch v8.1.1861 (*Edgar 2010*), with the following options: truncate the read at the first base with quality 5 or below; the truncated read should be 100 bp or bigger, and the overlap between forward and reverse should be 100 bp or bigger. I evaluated the quality of the libraries using FastQC.

I first tested many assembly softwares (the Geneious assembler, Tadpole as implemented in Geneious, Cap3 (*Huang and Madan 1999*), and Velvet (*Zerbino and Birney 2008*)) and compared them with Quast (*Gurevich, Saveliev et al. 2013*). I generally found that Velvet and the Geneious assembler outperform the other assemblers, and give similar results. Although Velvet is much faster than the Geneious assembler, I chose to use the later, as I could improve greatly the assembly quality using increasing sensitivity: first I used the "Medium-Low Sensitivity" on the merged and quality filtered reads, then the "Medium Sensitivity" on the first assembly and unassembled reads, finally the "High Sensitivity" on the second assembly and the unassembled reads. This allowed me to recover a good proportion of the genomes with a small amount of contigs, and with low number of unassembled reads. Furthermore, the Geneious assembler output consist of contigs, with all the reads aligned, which allows me to verify the assembly, assess precisely the coverage along the genome, and detect variants, when Velvet's results consist of consensus sequences with an average coverage, which can mask regions of high/low coverage.

To further improve the quality of my assemblies, I used the Geneious tools to detect regions of low coverage (<10 reads) and variations/SNPs (minimum coverage 10 and minimum variant frequencies 0.1). I then inspected the detected regions manually, trimmed the low coverage regions, and verified that the variants represent true variation and not sequencing errors. I found the true variants to cluster within regions of very high coverage compared to the rest of the genome; probably indicating duplicated genes that are too similar to be separated by the assembler. Finally, I kept only contigs that were formed of 100 reads or more, with an average coverage of 10 or more.

I exported consensus sequences as fasta files using the majority option, and annotated the genomes with the online tool RAST (*Aziz, Bartels et al. 2008*, *Overbeek, Olson et al. 2014*). I used the "classic RAST" annotation scheme, with the "RAST" gene caller, and used the options "fix error" and "backfill gaps". Finally, I estimated the evolutionary distances between my candidates and the reference genomes (table 14) using ANDI (*Haubold, Klotzl et al. 2015*) with both Jukes-Cantor and Kimura model, with 10,000 iterations. I imported the distance matrices obtained from ANDI in R via the function AS.DIST from the VEGAN package (*Dixon 2003*), build a neighbor joining tree for each iteration using the function NJ from the PHANGORN package (*Schliep 2011*), and

finally used the package ape (*Paradis, Claude et al. 2004*) to build a consensus tree from the 10,000 iterations and plot this consensus tree. For this, I excluded the plasmids, as they are not transmitted only vertically, and thus could bias the phylogenetic reconstruction.

Table 14: Summary of the reference genomes and plasmids used for phylogenetic and functional analysis.

Genus	Species	Strain	Plasmid	Identification Number
Citrobacter	*Amalonaticus*	FDAARGOS122	--	NZ_CP014015
Citrobacter	*Amalonaticus*	FDAARGOS165	--	NZ_CP014070
Citrobacter	*Amalonaticus*	Y19	--	NZ_CP011132
Citrobacter	*Amalonaticus*	Y19	pY19	NZ_CP011133
Citrobacter	*Freundii*	CAV1321	--	NZ_CP011612
Citrobacter	*Freundii*	CAV1321	pCAV1321-71	NZ_CP011609
Citrobacter	*Freundii*	CAV1321	pCAV1321-135	NZ_CP011610
Citrobacter	*Freundii*	CAV1321	pCAV1321-1916	NZ_CP011603
Citrobacter	*Freundii*	CAV1321	pCAV1321-3223	NZ_CP011604
Citrobacter	*Freundii*	CAV1321	pCAV1321-3820	NZ_CP011605
Citrobacter	*Freundii*	CAV1321	pCAV1321-4310	NZ_CP011606
Citrobacter	*Freundii*	CAV1321	pCAV1321-4938	NZ_CP011607
Citrobacter	*Freundii*	CAV1321	pKPC_CAV1321-45	NZ_CP011608
Citrobacter	*Freundii*	CAV1321	pKPC_CAV1321-244	NZ_CP011611
Citrobacter	*Freundii*	CAV1741	--	NZ_CP011657
Citrobacter	*Freundii*	CAV1741	pCAV1741-16	NZ_CP011653
Citrobacter	*Freundii*	CAV1741	pCAV1741-101	NZ_CP011654
Citrobacter	*Freundii*	CAV1741	pCAV1741-110	NZ_CP011655
Citrobacter	*Freundii*	CAV1741	pCAV1741-1916	NZ_CP011651
Citrobacter	*Freundii*	CAV1741	pCAV1741-3223	NZ_CP011652
Citrobacter	*Freundii*	CAV1741	pKPC_CAV1741	NZ_CP011656
Citrobacter	*Freundii*	CFNIH1	--	NZ_CP007557
Citrobacter	*Freundii*	CFNIH1	pKEC-a3c	NZ_CP007558
Citrobacter	*Freundii*	P10159	--	NZ_CP012554
Citrobacter	*Koseri*	ATCCBAA895	--	NC_009792
Citrobacter	*Koseri*	ATCCBAA895	pCKO2	NC_009794
Citrobacter	*Koseri*	ATCCBAA895	pCKO3	NC_009793
Citrobacter	*Rodentium*	ICC168	--	NC_013716
Citrobacter	*Rodentium*	ICC168	pCROD1	NC_013717
Citrobacter	*Rodentium*	ICC168	pCROD2	NC_013718
Citrobacter	*Rodentium*	ICC168	pCROD3	NC_013719
Proteus	*Mirabilis*	BB2000	--	NC_022000
Proteus	*Mirabilis*	CYPM1	--	NZ_CP012674
Proteus	*Mirabilis*	HI4320	--	NC_010554
Morganella	*Morganii*	CGH69	pCGH69	NC_021524
Morganella	*Morganii*	FDAARGOS172	--	NZ_CP014026
Morganella	*Morganii*	KT	--	NC_020418
Morganella	*Morganii*	M60	pM60	NC_023901
Morganella	*Morganii*	M203	R485	NC_016036

In RAST, I used the comparison tools to evaluate the differences between the strains. I exported the table data from the sequence-based comparison and imported in R where I used the package CIRCLIZE (*Gu, Gu et al. 2014*) to plot the circular comparisons. I also exported the features tables for each strain and imported it in R. I modified the data to obtain a table similar to the OTU tables for microbial analysis, and used the package vegan (*Dixon 2003*) to 1) build a Euclidean distance matrix, 2) perform PCoA on the distance matrix. Additionally, I exported the list of features that are strain-specific (present in all colonies of one strain and absent in all colonies of another strain).

General conclusion

The glycosyltransferase *B4galnt2* displays cis-regulatory variation particular to house mice: the wild type C57BL/6J allele directs intestinal expression, which is the pattern in other vertebrates, including humans, while the alternative RIIIS/J allele drives expression in blood vessels, inducing a bleeding disorder similar to the human von Willebrand disease. Twenty years of research have shown that the RIIIS/J haplotype is not only an artifact of the artificial breeding of laboratory mouse strains, but represents naturally occurring variation, which is widespread in wild populations of various *Mus* species despite the expected fitness cost of the RIIIS/J-associated bleeding disorder. Moreover, molecular evidence indicates that long-term balancing selection has maintained these alleles in the wild for at least 2.8 Million years, and that recent selection lead to the increase of RIIIS/J frequency in one population of South France.

The fact that systematic reviews of balancing selection in the human genome have identified genes that are mainly involved in immunity in its broadest sense (*Andres, Hubisz et al. 2009*, *Andrés 2011*, *Leffler, Gao et al. 2013*), together with the demonstrated involvement of *B4galnt2* in host-microbe interaction (*Staubach, Kunzel et al. 2012*, *Rausch, Steck et al. 2015*), suggest that host-pathogen interactions might be responsible for the long-term maintenance of *B4galnt2* alleles in the wild. For example, the modified glycosylation of the gastrointestinal mucus layer in RIIIS/J carrying mice compared to C57BL/6J carrying mice could confer protection against gastrointestinal pathogen(s) and thus compensate the fitness cost stemming from the prolonged bleeding times.

Although evidences from both laboratory and wild mice experiments indicate that host-pathogen(s) interactions are a likely cause of the maintenance of *B4galnt2* variation in natural populations, the conditions under which such trade-offs could lead to the long-term maintenance of disease-associated variation at *B4galnt2* remained however unclear. To resolve the population dynamics of *B4galnt2* in the wild, I addressed the maintenance of its murine allele via various approaches. First, I used available resources to determine whether the signs of recent selection observed in a French population of wild *Mus musculus domesticus* (*Johnsen, Teschke et al. 2009*) was detectable for other populations of *M. m. domesticus* from France and Germany (*Linnenbrink 2012*, *Linnenbrink, Wang et al. 2013*). I could confirm the recent action of selection in the increase of RIIIS/J allele frequency in three populations of southwestern France, suggesting that geographically limited selective forces are responsible for the maintenance of variation at *B4galnt2* in southwestern populations. These forces could for example come from an

environmental pathogen, present in southwestern France but absent from the northeast of France and from Germany. Then, I used mathematical modeling to evaluate whether pathogen-driven selection might be a plausible cause of maintenance of the disease-associated RIIIS/J allele. I based my model on the well-known Wright-Fischer process, but modified it to consider the mating of diploid hosts. I explored different models with regard to the phenotype of heterozygous mice, as they express *B4galnt2* both in the gastrointestinal epithelium and the vascular endothelium, thus potentially carrying both costs of pathogen susceptibility and prolonged bleeding. By comparing the simulated genotype frequencies to those of the natural populations, I could demonstrate that long-term maintenance of both *B4galnt2* alleles can be caused by pathogen-driven selection, if the fitness cost of prolonged bleeding is roughly half that of infection, and if the heterozygotes and RIIIS/J homozygotes are resistant/tolerant to the relevant pathogen. This indicates that a dominant protective function of the RIIIS/J allele is more likely to drive the maintenance of both *B4galnt2* alleles than a protective loss of intestinal *B4galnt2* expression.

Next, I collected over 200 mice from Southwest France in an attempt at identifying pathogen(s) that could be responsible for the maintenance of the RIIIS/J allele in wild populations. For each mouse, I linked *B4galnt2* genotype, inflammation status and gastrointestinal microbial community composition, and through this thorough analysis, I could identify several promising candidates. Among those potential pathogens, two are of particular interest: (i) *Citrobacter*, which I identified at the genus level only, is a well known member of the intestinal flora, and was already associated to *B4galnt2* genotype in laboratory experiments (*Staubach, Kunzel et al. 2012*) and (ii) *Morganella*, which I identified to the resolution of a species-level OTU, and is a well known opportunistic pathogen. Finally, I could identify the relevant species of *Morganella*, which represents a new subspecies of the *Morganella morganii* group and possesses virulence-related genes absent from the other *Morganella* species, which may account for its potential to drive selection at *B4galnt2* via genotype-dependent host-pathogen interactions.

Interestingly, although both the theoretical- and experimental analyses suggest that pathogen-driven selection could be responsible for the long-term maintenance of both *B4galnt2* alleles in the wild, the model indicates that the protective function of the RIIIS/J allele should be dominant, while the analysis of wild mice mostly suggest a co-dominant effect, with heterozygotes displaying an intermediate phenotype intermediate to that of both homozygotes (although this varies with the considered organ/bacterial species). One explanation for this disagreement resides in the joint resistance/tolerance to both intestinal and systemic pathogens. Indeed, my analysis of the bacterial traces in the blood of the mice could not rule out the association between systemic pathogens and *B4galnt2* genotype, suggesting that *B4galnt2* could influence not only

gastrointestinal-, but also systemic pathogens. The model however, does not explicitly consider whether the pathogen invades through the gastrointestinal tract or through the blood vessels, but considers a cost of infection, which could represent the net fitness cost resulting from both intestinal- and vascular infection. This net fitness effect could appear as a dominant phenotype, while each of its component may have a different behavior. Another possibility would be that the wild populations have not yet reached equilibrium with regard to a balance between tradeoffs, making the comparison of the simulated- to the observed populations less informative. However, this may not be very likely, as the allele frequencies appear to be stable, with no significant changes observed between the first and second sampling of the Massif Central (MC) and Espelette (ES) populations ((*Johnsen, Teschke et al. 2009*) sampled MC in 2005, (*Linnenbrink 2012*, *Linnenbrink, Wang et al. 2013*) sampled MC and ES in 2009, I sampled ES in 2013). Yet another explanation might reside in the non-specificity of my measure of inflammation with regard to its cause(s). Indeed, wild mice are subject to all sorts of immune challenges that can cause inflammation, which might create experimental noise interfering with the observed phenotype.

In conclusion, my work provides new insights into the potential evolutionary dynamics taking place at *B4galnt2* in wild populations of house mice, showing that pathogen-driven selection is a likely cause for the maintenance of both *B4galnt2* alleles in the wild. Moreover, my work could be applied beyond the scope of murine glycosyltransferases, as the methods that I developed can easily be generalized to other biological models. Finally, this work will yield follow up projects to: (i) test whether *Morganella's* potential pathogenic behavior is dependent on the *B4galnt2* genoytpe *in vivo*; (ii) investigate the impact of *B4galnt2* on the functions performed by the intestinal microbial community via shot-gun metagenomics; and (iii) explore the presence of candidate pathogens in ancient mouse feces samples in an attempt to retrace the history of the host-pathogen interaction in the French populations of *Mus musculus domesticus*.

Acknowledgment

First of all, I would like to thank my parents. You are the reason I persevered through all the though times, you sacrificed yourselves to allow me to follow my path, since I was a child you told me repeatedly "we don't care what you do, we will always support you, just do something you enjoy" and these words have pushed me to surpass myself to obtain what I want. There are no words to say how grateful I am to you. Je vous aime.

Then, I would like to thank two of my schoolteachers, which shared their passion for biology with me, and gave me the incentive to pursue a carrier in science: Mr. Sannier, my biology teacher in high school, who managed to get dissipated teenagers to be interested in biology, and Mr. Mottet, who was the irreplaceable master of the "classe prépa BCPST".

Next I would like to thank the professors that allowed me to discover academic research by taking me as intern in their labs when I knew nothing about pipets and PCRs, and helped me grow as a scientist: Pr.Dr. Sahil Adriouch, Pr.Dr. John Baines, and Pr.Dr. Frank Jiggins. John, you deserve a special place, since you not only gave me a chance when I was still an inexperienced bachelor student, but you also provided me with a great working place to continue my studies and finally become a PhD. You offered me great opportunities to perform high-level science, in a field I am particularly interested in, and you not only supervised me on my main thesis project, but also guided me through the development of my own work, allowing me to become the accomplished scientist I am today.

Finally, I want to thank the Baines' Group, in particular Silke Carstensen and Sven Künzel, you are the heart of this group, and most importantly, Meriem Beleheouane, we supported each other through our PhDs, we argued a bit, discussed a lot, complained about French politics, and killed a lot of mice (but we saved Freddy!), but most importantly, we became true friends. You will always have a special place in my heart, and I hope we both succeed in our scientific career and our private lives.

Curriculum Vitae

Personal information

Name: Marie Vallier

Date of birth: September 25th 1989

Place of birth: Rouen, France

Current residence: Kiel, Germany

Education

September 2004 - June 2007: Scientific A-level equivalent ("Baccalauréat" - Biology major) obtained with the distinction "good" ("mention bien") at the Lycée de la Vallée du Cailly, Déville-lès-Rouen, France.

September 2007 - July 2009: Bachelor's Degree Equivalent ("CPGE-BCPST") in Biology, Chemistry, Physics and Earth Sciences at the Lycée Pierre Corneille, Rouen, France.

September 2009 - August 2012: Master's Degree ("Ingénieur") in Biotechnologies and Pharmacology at Polytech'Nice-Sophia, Sophia-Antipolis, France.

First internship (July 2010) in the immunology unit of the Inserm U905 of Rouen, France, under the supervision of Dr. Sahil Adriouch, working on the development of a murine model of polymiosite.

Second internship (June-August 2011) at the Max Planck Institute for Evolutionary Biology of Plön, Germany & the Institute for Experimental Medicine, UKSH, Kiel, Germany, under the supervision of Pr. Dr. John Baines, working on the impact of the beta-defensin 3 on the composition of the intestinal microbiota in mice.

Master's Thesis (March-September 2012) in the Genetics Department of Cambridge University, UK, under the supervision of Pr. Dr. Frank Jiggins, working on the host-pathogen coevolution using the Drosophila - sigma viruses biological model.

January 2013 - current: PhD student at the Max Planck Institute for Evolutionary Biology of Plön, Germany and the Christian-Albrechts University, Kiel, Germany, inside the International Max Planck Research School (IMPRS) for Evolutionary Biology; working on pathogen-driven selection in the context of *B4galnt2* in wild mice.

Publications

Published

1) Multicenter quality assessment of 16S ribosomal DNA-sequencing for microbiome analyses reveals high inter-center variability. Hiergeist A, Reischl U, Priority Program 1656 Intestinal Microbiota Consortium/ quality assessment participants*, Gessner A. * Baines JF and **Vallier M** are contributing members. ***International Journal of Medical Microbiology 2016,*** *Vol. 306 No. 5, pp. 334-342.* doi:10.1016/j.ijmm.2016.03.005.

2) Analysis of intestinal microbiota in hybrid house mice reveals evolutionary divergence in a vertebrate hologenome. Wang J, Kalyan S, Steck N, Turner LM, Harr B, Künzel S, **Vallier M**, Häsler R, Franke A, Oberg H, Ibrahim SM, Grassl GA, Kabelitz D, Baines JF. ***Nature Communications 2015,*** *Vol. 6 No. 6440*

Submitted

1) Evaluating the maintenance of disease-associated variation at the blood group-related gene *B4galnt2* in house mice. **Vallier M**, Hindersin L, Abou Chakra M, Linnenbrink M, Traulsen A, Baines JF. *Submitted to* ***BMC Evolutionary Biology***

2) MHC diversity in wild mice is driven by the local generation of new alleles in meta-populational demes. Linnenbrink M, Teschke M. Montero I, **Vallier M**, Tautz D. *Submitted to* ***Molecular Ecology****, under revision.*

Affidavit

Declaration

I hereby 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: Evaluating the maintenance of disease-associated variation at the blood group-related gene *B4galnt2* in house mice. Marie Vallier, Laura Hindersin, Maria Abou Chakra and John Baines designed the study. Miriam Linnenbrink provided mouse population samples and sequenced *B4galnt2* diagnostic fragment 5 for the present study. Marie Vallier typed, sequenced and analyzed the 12 *B4galnt2*-linked microsatellites loci for the present study. Marie Vallier, Laura Hindersin, Maria Abou Chakra and Arne Traulsen developed the models. Marie Vallier implemented the computational model and performed all analysis. Marie Vallier and John Baines wrote the paper, with significant input from Laura Hindersin, Maria Abou Chakra and Arne Traulsen.

Chapter II: Pathometagenomics: identifying candidate pathogen by 16S rRNA metabarcoding. Marie Vallier, John Baines and Guntram Grassl designed the study. Marie Vallier planned, organized and supervised the fieldwork. Marie Vallier performed the fieldwork together with Theresa Arlt, Meriem Beleheouanne, Janine Braun, Jana Neckelmann, Nina Reinhardt, Jan Schubert and Janice Seidel. Marie Vallier performed all lab work (with support

from Theresa Arlt, Silke Carstensen and Jan Schubert) apart from the histological slides processing and scoring, which was done by Janine Braun, Marina Pils and Jan Schubert. Marie Vallier performed all analysis and wrote the chapter, with editing from John Baines.

Chapter III: Characterization of the candidate pathogens. Marie Vallier performed the 16S rRNA gene analysis, with technical support from Silke Carstensen and Jan Schubert. Guntram Grassl isolated bacterial clones from the samples collected by Marie Vallier, and extracted the DNA. Marie Vallier performed the whole genome sequencing together with Cornelia Burghardt. Marie Vallier performed all analysis and wrote the chapter, with editing from John Baines.

Plön, March 2017,

Marie Vallier

Pr. Dr. John F. Baines

Bibliography

Andrés, A. M. (2011). "Balancing Selection in the Human Genome." eLS.

Andres, A. M., M. J. Hubisz, A. Indap, D. G. Torgerson, J. D. Degenhardt, A. R. Boyko, R. N. Gutenkunst, T. J. White, E. D. Green, C. D. Bustamante, A. G. Clark and R. Nielsen (2009). "Targets of Balancing Selection in the Human Genome." Molecular Biology and Evolution **26**(12): 2755-2764.

Aziz, R. K., D. Bartels, A. A. Best, M. DeJongh, T. Disz, R. A. Edwards, K. Formsma, S. Gerdes, E. M. Glass, M. Kubal, F. Meyer, G. J. Olsen, R. Olson, A. L. Osterman, R. A. Overbeek, L. K. McNeil, D. Paarmann, T. Paczian, B. Parrello, G. D. Pusch, C. Reich, R. Stevens, O. Vassieva, V. Vonstein, A. Wilke and O. Zagnitko (2008). "The RAST Server: rapid annotations using subsystems technology." BMC Genomics **9**: 75.

Bonhomme, F., A. Orth, T. Cucchi, H. Rajabi-Maham, J. Catalan, P. Boursot, J. C. Auffray and J. Britton-Davidian (2011). "Genetic differentiation of the house mouse around the Mediterranean basin: matrilineal footprints of early and late colonization." Proc Biol Sci **278**(1708): 1034-1043.

Bonhomme, F., E. Rivals, A. Orth, G. R. Grant, A. J. Jeffreys and P. R. J. Bois (2007). "Species-wide distribution of highly polymorphic minisatellite markers suggests past and present genetic exchanges among House Mouse subspecies." Genome Biology **8**(5).

Borges-Canha, M., J. P. Portela-Cidade, M. Dinis-Ribeiro, A. F. Leite-Moreira and P. Pimentel-Nunes (2015). "Role of colonic microbiota in colorectal carcinogenesis: a systematic review." Rev Esp Enferm Dig **107**(11): 659-671.

Broom, M. and J. Rychtář (2013). Game-theoretical models in biology. Boca Raton, FL, CRC Press, Taylor and Francis Group.

Byrne, M. F., S. W. Kerrigan, P. A. Corcoran, J. C. Atherton, F. E. Murray, D. J. Fitzgerald and D. M. Cox (2003). "Helicobacter pylori binds von Willebrand factor and interacts with GPIb to induce platelet aggregation." Gastroenterology **124**(7): 1846-1854.

Chytry, M., L. Tichy, J. Holt and Z. Botta-Dukat (2002). "Determination of diagnostic species with statistical fidelity measures." Journal of Vegetation Science **13**(1): 79-90.

Cole, J. R., Q. Wang, J. A. Fish, B. L. Chai, D. M. McGarrell, Y. N. Sun, C. T. Brown, A. Porras-Alfaro, C. R. Kuske and J. M. Tiedje (2014). "Ribosomal Database Project: data and tools for high throughput rRNA analysis." Nucleic Acids Research **42**(D1): D633-D642.

De Caceres, M. and P. Legendre (2009). "Associations between species and groups of sites: indices and statistical inference." Ecology **90**(12): 3566-3574.

De Caceres, M., P. Legendre and M. Moretti (2010). "Improving indicator species analysis by combining groups of sites." Oikos **119**(10): 1674-1684.

Dixon, P. (2003). "VEGAN, a package of R functions for community ecology." Journal of Vegetation Science **14**(6): 927-930.

Dufrene, M. and P. Legendre (1997). "Species assemblages and indicator species: The need for a flexible asymmetrical approach." Ecological Monographs **67**(3): 345-366.

Earl, D. A. and B. M. Vonholdt (2012). "STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method." Conservation Genetics Resources **4**(2): 359-361.

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

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

Evanno, G., S. Regnaut and J. Goudet (2005). "Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study." Molecular Ecology **14**(8): 2611-2620.

Excoffier, L., P. E. Smouse and J. M. Quattro (1992). "Analysis of Molecular Variance Inferred from Metric Distances among DNA Haplotypes - Application to Human Mitochondrial-DNA Restriction Data." Genetics **131**(2): 479-491.

Fumagalli, M., R. Cagliani, U. Pozzoli, S. Riva, G. P. Comi, G. Menozzi, N. Bresolin and M. Sironi (2009). "Widespread balancing selection and pathogen-driven selection at blood group antigen genes." Genome Res **19**(2): 199-212.

Gabriel, S. I., F. Johannesdottir, E. P. Jones and J. B. Searle (2010). "Colonization, mouse-style." Bmc Biology **8**.

Gu, Z., L. Gu, R. Eils, M. Schlesner and B. Brors (2014). "circlize Implements and enhances circular visualization in R." Bioinformatics **30**(19): 2811-2812.

Gupta, Y., S. Moller, M. Witte, M. Belheouane, T. Sezin, M. Hirose, A. Vorobyev, F. Niesar, J. Bischof, R. J. Ludwig, D. Zillikens, C. D. Sadik, T. Restle, R. Hasler, J. F. Baines and S. M. Ibrahim (2016). "Dissecting genetics of cutaneous miRNA in a mouse model of an autoimmune blistering disease." BMC Genomics **17**(1): 112.

Gurevich, A., V. Saveliev, N. Vyahhi and G. Tesler (2013). "QUAST: quality assessment tool for genome assemblies." Bioinformatics **29**(8): 1072-1075.

Hannon. (2010). from *http://hannonlab.cshl.edu/fastx_toolkit/*.

Harr, B., E. Karakoc, R. Neme, M. Teschke, C. Pfeifle, Z. Pezer, H. Babiker, M. Linnenbrink, I. Montero, R. Scavetta, M. R. Abai, M. P. Molins, M. Schlegel, R. G. Ulrich, J. Altmuller, M. Franitza, A. Buntge, S. Kunzel and D. Tautz (2016).

"Genomic resources for wild populations of the house mouse, Mus musculus and its close relative Mus spretus." Sci Data **3**: 160075.
Haubold, B., F. Klotzl and P. Pfaffelhuber (2015). "andi: fast and accurate estimation of evolutionary distances between closely related genomes." Bioinformatics **31**(8): 1169-1175.
Hofbauer, J., P. Schuster and K. Sigmund (1982). "Game Dynamics in Mendelian Populations." Biological Cybernetics **43**(1): 51-57.
Huang, X. and A. Madan (1999). "CAP3: A DNA sequence assembly program." Genome Res **9**(9): 868-877.
Hubert, J., T. Erban, M. Kamler, J. Kopecky, M. Nesvorna, S. Hejdankova, D. Titera, J. Tyl and L. Zurek (2015). "Bacteria detected in the honeybee parasitic mite Varroa destructor collected from beehive winter debris." J Appl Microbiol **119**(3): 640-654.
Ihle, S., I. Ravaoarimanana, M. Thomas and D. Tautz (2006). "An analysis of signatures of selective sweeps in natural populations of the house mouse." Molecular Biology and Evolution **23**(4): 790-797.
Imhof, L. A. and M. A. Nowak (2006). "Evolutionary game dynamics in a Wright-Fisher process." J Math Biol **52**(5): 667-681.
Jakobsson, M. and N. A. Rosenberg (2007). "CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure." Bioinformatics **23**(14): 1801-1806.
Johnsen, J. M., G. G. Levy, R. J. Westrick, P. K. Tucker and D. Ginsburg (2008). "The endothelial-specific regulatory mutation, Mvwf1, is a common mouse founder allele." Mamm Genome **19**(1): 32-40.
Johnsen, J. M., M. Teschke, P. Pavlidis, B. M. McGee, D. Tautz, D. Ginsburg and J. F. Baines (2009). "Selection on cis-regulatory variation at B4galnt2 and its influence on von Willebrand factor in house mice." Mol Biol Evol **26**(3): 567-578.
Jones, E. P., F. Johannesdottir, I. Gunduz, M. B. Richards and J. B. Searle (2011). "The expansion of the house mouse into north-western Europe." Journal of Zoology **283**(4): 257-268.
Knights, D., J. Kuczynski, E. S. Charlson, J. Zaneveld, M. C. Mozer, R. G. Collman, F. D. Bushman, R. Knight and S. T. Kelley (2011). "Bayesian community-wide culture-independent microbial source tracking." Nat Methods **8**(9): 761-763.
Konovalov, D. A., C. Manning and M. T. Henshaw (2004). "KINGROUP: a program for pedigree relationship reconstruction and kin group assignments using genetic markers." Molecular Ecology Notes **4**(4): 779-782.
Koressaar, T. and M. Remm (2007). "Enhancements and modifications of primer design program Primer3." Bioinformatics **23**(10): 1289-1291.
Leffler, E. M., Z. Y. Gao, S. Pfeifer, L. Segurel, A. Auton, O. Venn, R. Bowden, R. Bontrop, J. D. Wall, G. Sella, P. Donnelly, G. McVean and M. Przeworski (2013). "Multiple Instances of Ancient Balancing Selection Shared Between Humans and Chimpanzees." Science **339**(6127): 1578-1582.
Linnenbrink, M. (2012). Population genetic and functional analysis of the B4galnt2 gene in the genus Mus (Rodentia; Muridae). PhD, Christian-Albrechts-Universität zu Kiel.
Linnenbrink, M., J. M. Johnsen, I. Montero, C. R. Brzezinski, B. Harr and J. F. Baines (2011). "Long-term balancing selection at the blood group-related gene B4galnt2 in the genus Mus (Rodentia; Muridae)." Mol Biol Evol **28**(11): 2999-3003.
Linnenbrink, M., J. Wang, E. A. Hardouin, S. Kunzel, D. Metzler and J. F. Baines (2013). "The role of biogeography in shaping diversity of the intestinal microbiota in house mice." Mol Ecol **22**(7): 1904-1916.
Mason, W. J., J. S. Blevins, K. Beenken, N. Wibowo, N. Ojha and M. S. Smeltzer (2001). "Multiplex PCR protocol for the diagnosis of staphylococcal infection." J Clin Microbiol **39**(9): 3332-3338.
McAdow, M., D. M. Missiakas and O. Schneewind (2012). "Staphylococcus aureus secretes coagulase and von Willebrand factor binding protein to modify the coagulation cascade and establish host infections." J Innate Immun **4**(2): 141-148.
Meirmans, P. G. and P. H. Van Tienderen (2004). "GENOTYPE and GENODIVE: two programs for the analysis of genetic diversity of asexual organisms." Molecular Ecology Notes **4**(4): 792-794.
Michalakis, Y. and L. Excoffier (1996). "A generic estimation of population subdivision using distances between alleles with special reference for microsatellite loci." Genetics **142**(3): 1061-1064.
Mohlke, K. L., W. C. Nichols and D. Ginsburg (1999). "The molecular basis of von Willebrand disease." Int J Clin Lab Res **29**(1): 1-7.
Mohlke, K. L., W. C. Nichols, R. J. Westrick, E. K. Novak, K. A. Cooney, R. T. Swank and D. Ginsburg (1996). "A novel modifier gene for plasma von Willebrand factor level maps to distal mouse chromosome 11." Proc Natl Acad Sci U S A **93**(26): 15352-15357.
Mohlke, K. L., A. A. Purkayastha, R. J. Westrick, P. L. Smith, B. Petryniak, J. B. Lowe and D. Ginsburg (1999). "Mvwf, a dominant modifier of murine von Willebrand factor, results from altered lineage-specific expression of a glycosyltransferase." Cell **96**(1): 111-120.
Overbeek, R., R. Olson, G. D. Pusch, G. J. Olsen, J. J. Davis, T. Disz, R. A. Edwards, S. Gerdes, B. Parrello, M. Shukla, V. Vonstein, A. R. Wattam, F. Xia and R. Stevens (2014). "The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST)." Nucleic Acids Res **42**(Database issue): D206-214.
Paradis, E. (2012). "Analysis of Phylogenetics and Evolution with R, Second Edition." Analysis of Phylogenetics and Evolution with R, Second Edition: 1-386.
Paradis, E., J. Claude and K. Strimmer (2004). "APE: Analyses of Phylogenetics and Evolution in R language." Bioinformatics **20**(2): 289-290.

Prager, E. M., R. D. Sage, U. Gyllensten, W. K. Thomas, R. Hubner, C. S. Jones, L. Noble, J. B. Searle and A. C. Wilson (1993). "Mitochondrial-DNA Sequence Diversity and the Colonization of Scandinavia by House Mice from East Holsten." Biological Journal of the Linnean Society **50**(2): 85-122.
Pritchard, J. K., M. Stephens and P. Donnelly (2000). "Inference of population structure using multilocus genotype data." Genetics **155**(2): 945-959.
Rausch, P., N. Steck, A. Suwandi, J. A. Seidel, S. Kunzel, K. Bhullar, M. Basic, A. Bleich, J. M. Johnsen, B. A. Vallance, J. F. Baines and G. A. Grassl (2015). "Expression of the Blood-Group-Related Gene B4galnt2 Alters Susceptibility to Salmonella Infection." PLoS Pathog **11**(7): e1005008.
Salcedo, T., A. Geraldes and M. W. Nachman (2007). "Nucleotide variation in wild and inbred mice." Genetics **177**(4): 2277-2291.
Salter, S. J., M. J. Cox, E. M. Turek, S. T. Calus, W. O. Cookson, M. F. Moffatt, P. Turner, J. Parkhill, N. J. Loman and A. W. Walker (2014). "Reagent and laboratory contamination can critically impact sequence-based microbiome analyses." Bmc Biology **12**.
Schliep, K. P. (2011). "phangorn: phylogenetic analysis in R." Bioinformatics **27**(4): 592-593.
Schloss, P. D., S. L. Westcott, T. Ryabin, J. R. Hall, M. Hartmann, E. B. Hollister, R. A. Lesniewski, B. B. Oakley, D. H. Parks, C. J. Robinson, J. W. Sahl, B. Stres, G. G. Thallinger, D. J. Van Horn and C. F. Weber (2009). "Introducing mothur: Open-Source, Platform-Independent, Community-Supported Software for Describing and Comparing Microbial Communities." Applied and Environmental Microbiology **75**(23): 7537-7541.
Segurel, L., Z. Gao and M. Przeworski (2013). "Ancestry runs deeper than blood: the evolutionary history of ABO points to cryptic variation of functional importance." Bioessays **35**(10): 862-867.
Silver, L. M. (1995). Mouse genetics : concepts and applications. New York, Oxford University Press.
Stahl, E. A., G. Dwyer, R. Mauricio, M. Kreitman and J. Bergelson (1999). "Dynamics of disease resistance polymorphism at the Rpm1 locus of Arabidopsis." Nature **400**(6745): 667-671.
Staubach, F., S. Kunzel, A. C. Baines, A. Yee, B. M. McGee, F. Backhed, J. F. Baines and J. M. Johnsen (2012). "Expression of the blood-group-related glycosyltransferase B4galnt2 influences the intestinal microbiota in mice." ISME J **6**(7): 1345-1355.
Stephens, M., N. J. Smith and P. Donnelly (2001). "A new statistical method for haplotype reconstruction from population data." Am J Hum Genet **68**(4): 978-989.
Stoddard, S. F., B. J. Smith, R. Hein, B. R. Roller and T. M. Schmidt (2015). "rrnDB: improved tools for interpreting rRNA gene abundance in bacteria and archaea and a new foundation for future development." Nucleic Acids Res **43**(Database issue): D593-598.
Stuckenholz, C., L. Lu, P. Thakur, N. Kaminski and N. Bahary (2009). "FACS-assisted microarray profiling implicates novel genes and pathways in zebrafish gastrointestinal tract development." Gastroenterology **137**(4): 1321-1332.
Taiyun Wei, V. S. (2016). corrplot: Visualization of a Correlation Matrix.
Tamura, K., J. Dudley, M. Nei and S. Kumar (2007). "MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0." Mol Biol Evol **24**(8): 1596-1599.
Tellier, A., S. Moreno-Gamez and W. Stephan (2014). "Speed of adaptation and genomic footprints of host-parasite coevolution under arms race and trench warfare dynamics." Evolution **68**(8): 2211-2224.
Thomas, M., F. Moller, T. Wiehe and D. Tautz (2007). "A pooling approach to detect signatures of selective sweeps in genome scans using microsatellites." Molecular Ecology Notes **7**(3): 400-403.
Thomer, L., O. Schneewind and D. Missiakas (2013). "Multiple ligands of von Willebrand factor-binding protein (vWbp) promote Staphylococcus aureus clot formation in human plasma." J Biol Chem **288**(39): 28283-28292.
Thompson, J. D., D. G. Higgins and T. J. Gibson (1994). "CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice." Nucleic Acids Res **22**(22): 4673-4680.
Tichy, L. and M. Chytry (2006). "Statistical determination of diagnostic species for site groups of unequal size." Journal of Vegetation Science **17**(6): 809-818.
Untergasser, A., I. Cutcutache, T. Koressaar, J. Ye, B. C. Faircloth, M. Remm and S. G. Rozen (2012). "Primer3--new capabilities and interfaces." Nucleic Acids Res **40**(15): e115.
Verslyppe, B., W. De Smet, B. De Baets, P. De Vos and P. Dawyndt (2014). "StrainInfo introduces electronic passports for microorganisms." Syst Appl Microbiol **37**(1): 42-50.
Wickham, H. (2007). "Reshaping data with the reshape package." Journal of Statistical Software **21**(12): 1-20.
Woolhouse, M. E., J. P. Webster, E. Domingo, B. Charlesworth and B. R. Levin (2002). "Biological and biomedical implications of the co-evolution of pathogens and their hosts." Nat Genet **32**(4): 569-577.
Zerbino, D. R. and E. Birney (2008). "Velvet: algorithms for de novo short read assembly using de Bruijn graphs." Genome Res **18**(5): 821-829.