# Population genetic and functional analysis of the *B4galnt2* gene in the genus *Mus* (Rodentia; Muridae)

Miriam Linnenbrink

Dissertation
zur Erlangung des Doktorgrades
an der Mathematisch-Naturwissenschaftlichen Fakultät
der Christian-Albrechts-Universität zu Kiel

vorgelegt von
Miriam Linnenbrink

Kiel, den 3. April 2012

# Population genetic and functional analysis of the *B4galnt2* gene in the genus *Mus* (Rodentia; Muridae)

Miriam Linnenbrink

Dissertation
zur Erlangung des Doktorgrades
an der Mathematisch-Naturwissenschaftlichen Fakultät
der Christian-Albrechts-Universität zu Kiel

vorgelegt von
Miriam Linnenbrink

Kiel, den 3. April 2012

Referent: Prof. Dr. John Baines
Koreferent: Prof. Dr. Hinrich Schulenburg

Tag der mündlichen Prüfung: 25.05.2012
Zum Druck genehmigt: 25.05.2012

gez.: Prof. Dr. Kipp, Dekan

Nichts in der Geschichte des Lebens ist beständiger als der Wandel.

*Charles Darwin*

*Allen, die zu mir gehören…*

# Table of Contents

**Table of Contents** **II**

**List of Tables** **III**

**List of Figures** **V**

**Acknowledgements** **VII**

**Zusammenfassung** **IX**

**Summary** **XI**

**Declaration of Author's Contribution** **XIII**

**General Introduction** **1**

1. The History of a Model Organism . . . . . 1
2. Population Genetics . . . . . 4
    2.1 The History of Population Genetics . . . . . 4
    2.2 The Main Principles . . . . . 4
    2.3 Balancing Selection . . . . . 6
    2.4 An Example - *B4galnt2* as Candidate Gene . . . . . 7
3. Metagenomics . . . . . 9
    3.1 A Global View on Metagenomics . . . . . 9
    3.2 An Introduction to the (Gut) Microbiome . . . . . 10
4. Scope of the Thesis . . . . . 12

**Chapter 1 - Long-term balancing selection at the blood group-related gene *B4galnt2* in the genus *Mus* (Rodentia;Muridae)** **15**

**Chapter 2 - The Role of Biogeography in Shaping Diversity of the Intestinal Microbiota in House Mice** **23**

**Chapter 3 - Population Dynamics at *B4galnt2* and its Influence on the Intestinal Microbiota in Natural Populations of the Western House Mouse** **39**

**Bibliography** **56**

**Appendix** **75**

Appendix A - Online Material Chapter 1 . . . . . 75
Appendix B - Supplementary Materials Chapter 2 . . . . . 81
Appendix C - Supplementary Information Chapter 3 . . . . . 82
Appendix D - Primerlist . . . . . 83

**Affidavit** **85**

**Curriculum Vitae** **87**

# List of Tables

1.1 Linkage Disequilibrium. . . . . 17
1.2 Relationship between genotype (allele class) and DBA lectin staining pattern. . . 19
2.1 General information on sampling locations and molecular data. . . . . 25
2.2 Summary of mitochondrial D-loop and microsatellite variation. . . . . 31
3.1 General information on sampling locations and sample sizes. . . . . 42
3.2 Summary statistics for *B4galnt2* Upstream Region (Fragment #5). . . . . 47
A-S1 Polymorphism and divergence across the *B4galnt2* upstream gene region. . . . . 77
A-S2 Polymorphism and divergence at seven unlinked reference loci in *M. spretus*. . . 78
B-S1 Microsatellite information. . . . . 81
D-S1 List of all primers used in this study. . . . . 83

# List of Figures

I-1 Phylogeny of the genus *Mus*. . . . . 2
I-2 Colonization routes of the different (sub)species of the genus *Mus*. . . . . 3
I-3 Sliding window analysis of the nucleotide divergence between the sequences of the labstrains RIIIS/J and C57BL6/J in the gene region of *B4galnt2*. . . . . 9
I-4 Sequence based metagenomic projects from 2003 - 2008. . . . . 10
1.1 (A) Nucleotide diversity and (B) Tajima's *D* across the *B4galnt2* upstream gene region. . . . . 18
1.2 Neighbor-Joining trees of *B4galnt2* upstream regions. . . . . 20
2.1 Sampling locations across Western Europe. . . . . 29
2.2 NeighbourNet network of 100 mitochondrial D-loop sequences. . . . . 30
2.3 Genetic clusters identified by STRUCTURE for K=5. . . . . 31
2.4 Phylogenetic tree of bacterial 16S rRNA gene sequences. . . . . 32
2.5 Geographic location-constrained principle coordinate analysis (PCoA) of Bray-Curtis dissimilarity. . . . . 33
2.6 Heatmap of Indicator OTUs. . . . . 35
3.1 Analysis of host population structure. . . . . 45
3.2 Distribution of C57BL6/J- (gut expression) and RIIIS/J- (blood vessel expression) haplotype classes in sampled populations. . . . . 46
3.3 NeighbourNet network of fragement #5. . . . . 47
3.4 Bacterial taxa significantly correlated with *B4galnt2* genotype. . . . . 48
3.5 Population structure according to underlying *B4galnt2* haplotype frequencies. . . 51
A-S1 Gene region of *B4galnt2* and the distribution of all sequenced fragments in its upstream region. . . . . 78
A-S2 Summary of polymorphic sites across *B4galnt2* gene region. . . . . 79
A-S3 Summary of sequence comparison between *B4galnt2* expression classes. . . . . 80
B-S1 Heatmap of indicator OTUs according to maternal transmission. . . . . 81
C-S1 Neighbour-Joining tree of all D-loop sequences. . . . . 82

# Acknowledgements

An erster Stelle danke ich John Baines, dass er mir die Möglichkeit gab, bei ihm zu promovieren. Er brachte mir das Vertrauen entgegen, dass aus einer Verhaltensbiologin auch eine Populationsgenetikerin werden kann. Zudem respektierte er meinen Wunsch in München bleiben zu wollen, als er in in den hohen Norden berufen wurde, was die Betreuung nicht immer vereinfachte. Die Arbeit an meinem Projekt hat mir immer viel Freude bereitet und durch die exzellente Betreuung wurde ich zu einer besseren Wissenschaftlerin.
John Parsch danke ich, da er mich als ”Gast”-Mitglied in seiner Gruppe in München behielt und mir einen tollen Arbeitsplatz, sowohl im Labor als auch im Büro zur Verfügung stellte und jederzeit bei Fragen hilfreiche Antworten hatte.
Lena Müller danke ich für eine unvergessliche Zeit im Büro. Sie hatte jederzeit ein offenes Ohr und gemeinsame Diskussionen halfen mir oft. Genauso bedanke ich mich bei Lisha Naduvilezhath für die schöne Zeit; oft hat sie mir bei Programmierfragen unter die Arme gegriffen und nicht zuletzt danke für die Latex-Vorlage. Ihre gute Laune war immer ansteckend und hat so manche Arbeit erleichtert. Ana Catalán versüßte mir die Zeit im Büro durch Aufforderungen zum Drücken oder auch Verwechslungen von Nashörnchen und Eichhörnchen. Danke auch Amanda Glaser für die schönen Pausen. Genauso bedanke ich mich auch bei Sonja Grath, meiner Ex-Büro-Mitinsassin, immer die Ruhe in Person und immer für einen da, auch aus der Ferne. Immer gerne werde ich mich an Claus Kemkemer oder auch Sägezahn-Claus erinnern, der mit seinem positiven Denken manche Probleme hat verschwinden lassen. Anja Hörger ist für mich ein wissenschaftliches Vorbild. Ich habe endlich jemanden gefunden, der meinen Musikgeschmack teilt! Iris Fischer machte mir immer Mut und baute mich (”it’s just a letter”) in manch schwerer Phase immer wieder auf. Iris, es war ’ne lustige Zeit mit dir! Stephan Hutter war mir oft eine große Hilfe in Perl- und anderen wissenschaftlichen Belangen, genauso wie Martin Hutzenthaler und sein R-Kurs. Mit Stephan und Martin, zusammen mit Claus und Paul erlebte ich außerdem lange Schafkopfabende, zum Abschalten von der Arbeit genau das Richtige. Stefan Laurent, Iris Fischer und Aurélien Tellier hatten immer hilfreiche Ideen,

wenn ich wieder einmal meine Gedanken über die Populationsstruktur sortieren musste. Vielen Dank auch Dirk Metzler, der mir neben Unterstützung bei diversen Analysenfragen auch immer mit R und Latex geholfen hat, danke auch an Meike Wittmann für ihre Unterstützung. Mit Susi Voigt zusammen verbrachte ich viele Stunden mit der genauen Rezeptur für die besten Seifenblasen. Winfried Hense, danke für deine steten schokoladige Konzentrationshilfen. Großer Dank gilt Ingrid Kroiss, ohne die ich das Desaster mit der Reisekostenstelle nie überstanden hätte. Danke auch an Hilde Lainer, Hedwig Gebhard, Gisela Brinkmann, Ana Vrljic und Simone Lange, die immer für mich da waren, mir mit Tipps und Tricks im Labor geholfen und mich oft ermutigt haben. Danke an die gesamte Evolutionsbio Gruppe in München, es war eine schöne Zeit!
Nicht zu Vergessen natürlich auch meine Kollegen an der CAU Kiel bzw. am Max Planck Institut in Plön. Allen voran bedanke ich mich hier bei Christine Pfeifle für ihre wertvollen Tips in allen Mäuseangelegenheiten, aber auch für ihre riesige Gastfreundschaft mit der sie und Jan von Rönn mich immer bei sich aufgenommen haben. Ich denke gerne an die Tatortabende zurück, nicht zuletzt wegen der kulinarischen Köstlichkeiten, die sie und Arne Nolte immer bereitstellten. Danke Inka für die schö ne Zeit in Plön und die super Zusammenarbeit! Heike und Till, danke für die Unterstützung mit meinen Mäusen in Plön und auch bei der Vorbereitung von diversen Mausfangaktionen. Inka Montero danke ich für die sehr gute Zusammenarbeit. Auch sie hat wesentlich dazu beigetragen, dass ich mich in Plön wohlgefühlt habe. Silke, Sven und Katja, auf euch war und ist immer Verlass! Katja, danke auch für den schönen Ausflug an die Nordsee und die Erweiterung meines Kunstverständnisses. Jun und Emilie - Prost! Auf unser paper! Ich möchte mich ausserdem bei allen bedanken, die zum Erfolg der Mausfangaktionen (Hermann, Ann-Kathrin, Philipp, Ben, Emilie und Knut) beigetragen haben. Danke Knut für RÜ1.
Mama, Papa, ihr habt immer an mich geglaubt, ihr seid die Besten! Genauso wie meine Großeltern, die mich außerdem immer mit aktuellen Zeitungsartikeln über Mäuse und Mausmakis versorgen und Opa, der sogar fast zu meinem Mausfangteam gehört hätte!
Hermann, du bist mir der allerliebste Mäusefänger! Wochenlang hast du mit mir Bauernhöfe durchstöbert und dir oft geduldig meine wildesten Theorien über Mäuse angehört und fleißig mitdiskutiert. Inzwischen studierst du sogar selber den Habitus der Maus (danke Martina).

# Zusammenfassung

Das Gen *B4galnt2* ist eine, mit Blutgruppen in Zusammenhang stehende Glykosyltransferase, die bei den meisten Säugetieren im Magen-Darm Trakt exprimiert wird. Die genaue Funktion des Gens ist unbekannt. In Hausmäusen führt eine Mutation in der *cis*-regulatorischen Region zu einer gewebsspezifischen Veränderung der Expression, der Expressionsort ist also entweder der Magen-Darm Trakt und/oder die Blutgefäße (C57BL6/J Allel (Expression im Epithelgewebe) oder RIIIS/J Allel (Expression im Endothelgewebe)). Die Expression des Gens *B4galnt2* in den Blutgefäßen führt zu einem Phänotyp, der der menschlichen Bluterkrankheit, der Von Willebrand Krankheit ähnelt. Die Auswirkungen der veränderten Darmexpression sind jedoch unbekannt. Die hier vorliegende Dissertation verbindet die Populationsgenetik mit der Metagenomik. Neben populationsgenetischen Untersuchungen in Bezug auf das Gen *B4galnt2* beleuchteten wir zudem die Einflüsse biogeografischer Faktoren auf die Zusammensetzung von Darmbakteriengemeinschaften.

Die populationsgenetische Untersuchung des Gens *B4galnt2* in natürlichen Populationen von allen drei Unterarten von *Mus musculus* und der Schwesterart *M. spretus* zeigte, dass die verschiedenen Allele, die in der Region vor dem Gen vorkommen, schon seit langer Zeit erhalten werden. Verschiedene Expressionsmuster konnten bis zur Aufspaltung von der Art *M. famulus*, also mehr als 2,8 Millionen Jahre zurückverfolgt werden. Die Tatsache, dass Darmexpression in allen Mäusen mit wenigstens einem C57BL6/J Allel konserviert ist, führt zu der Annahme, dass der Darm ein Angriffspunkt für Selektion ist. Bereits bekannt ist, dass, abhängig von der An-/Abwesenheit bestimmter Zucker im Magen-Darm Trakt, sowohl symbiontische/kommensalistische, als auch pathogene Bakterien beeinflusst werden. Diese Tatsache, zusammen mit der Langzeit-Erhaltung verschiedener Allelklassen, und somit unterschiedlicher Expressionsmuster lässt vermuten, dass Wirt-Pathogen Interaktionen sehr wahrscheinlich eine Rolle in der evolutionären Dynamik des Gens spielen.

Zusätzlich untersuchten wir die die Effekte von Geografie, Populationsstruktur des Wirts und den Einfluss der Vererbung durch das Muttertier auf die Mikrobiota, da bisher noch

nicht viel über die Einflüsse biogeografischer Faktoren auf die Darmmikrobiota bekannt war. Mit ”high throughput pyrosequencing” erhielten wir die Sequenzen des bakteriellen 16S rRNA Gens zur Beschreibung der Darmbakteriengemeinschaften von Mäusen aus acht verschiedenen Populationen aus Westeuropa. Mit Hilfe von Mikrosatelliten konnten wir die Populationsstruktur der Hausmäuse untersuchen und die mitochondriale DNA erlaubte uns Aussagen über die mütterliche Weitergabe mancher Bakterien zu treffen. Es stellte sich heraus, dass die Lage des Fangorts (d.h. geografischer Effekt) wesentlich die Zusammensetzung der Darmbakteriengemeinschaften beeinflusst. Aber auch die Populationsstruktur des Wirts und die zusammenhängenden Effekte aus Geografie und Substruktur zeigten deutliche Auswirkungen. Zusätzlich konnten wir noch Bakterienarten, gehörig zu den Genera *Helicobacter*, *Robinsoniella* und *Bacteroides*, identifizieren, die in signifikantem Zusammenhang mit all jenen Faktoren stehen.
Desweiteren erörterten wir die Verteilung der RIIIS/J und C57BL6/J Allelfrequenzen in den acht Mauspopulationen. Dabei stellte sich heraus, dass das RIIIS/J Allel mit mittlerer Frequenz im Süd-Westen Frankreichs auftritt und nach Mittelfrankreich hin deutlich abfällt. Dies lässt vermuten, dass unterschiedliche selektive Faktoren in den verschiedenen Populationen vorherrschen. Um Einblicke in die Dynamik der Allelfrequenzen zu gewinnen, untersuchten wir das Verteilungsmuster in Bezug auf die Populationsstruktur der Mäuse (welche wir über Mikrosatelliten und die mt D-loop Sequenzen abschätzten). Zudem versuchten wir, über die Analyse der Darmbakteriengemeinschaften in Bezug auf die verschiedenen Genotypen mögliche phänotypische Konsequenzen der *B4galnt2* Expression zu erhalten. Es zeigte sich, dass wahrscheinlich lokale Anpassung die Ursache für die großen Unterschiede in den Allelfrequenzen ist. Die zugrundeliegende Populationsstruktur, die durch verschiedene Kolonisationswellen gebildet hätte werden können und somit auch für die Differenzen in Allelfrequenzen verantwortlich sein könnte, zeigte sich als eher unwahrscheinlich. Außerdem konnten wir den Einfluss der *B4galnt2* Expression auf die Bakteriengemeinschaften im Darm bestätigen, was auch schon für Labormäuse gezeigt wurde. Eine direkte Assoziation zwischen Genotyp und Pathogen konnte nicht nachgewiesen werden.
Dadurch, dass wir die Langzeit-Erhaltung verschiedener Expressionsmuster des *B4galnt2* Gens nachweisen und auch die Rolle verschiedener biogeografischer Faktoren auf die Darmbakteriengemeinschaften näher beschreiben konnten, ließen sich die evolutionäre Geschichte des *B4galnt2* Gens und seine phänotypischen Auswirkungen genauer charakterisieren. Die ungeklärten Fragen im Zusammenhang mit dem Kosten-Nutzen-Ausgleich der beiden Allele benötigen für ihre Beantwortung aber weitere Untersuchungen und vorallem das Wissen um die genaue Funktion und damit den Nutzen des RIIIS/J Allels.

# Summary

*B4galnt2* is a blood group-related glycosyltransferase expressed in the gastrointestinal (GI) tract of most mammals. However, *cis*-regulatory variation at *B4galnt2* in house mice leads to different tissue-specific expression patterns affecting intestinal epithelium and vascular endothelium according to the alleles present in the direct upstream region (”C57BL6/J”’ or ”‘RIIIS/J” allele, respectively). Blood vessel expression of *B4galnt2* leads to a phenotype in mice very similar to a common human bleeding disorder, von Willebrand disease, but the consequences of altered expression in the intestine are unknown. This dissertation combines poulation genetics with metagenomics with respect to the *B4galnt2* gene in natural populations of mice and enlightens the role of biogeography in shaping intestinal micriobial communities in general.

The population genetic study of the *B4galnt2* gene in all three subspecies of *M. musculus* (*i.e. M. m. domesticus*, *M. m. musculus* and *M. m. castaneus*) and the sister species *M. spretus* revealed the long-term maintenance of different allele classes present in the direct upstream region of *B4galnt2*. Varying expression patterns could be identified to be present for $> 2.8$ MY, since the divergence of *M. famulus*. The finding, that gut expression was conserved in all mice exhibiting the C57BL6/J allele lead to the suggestion the the gut is a likely target of selection. It is known, that glycosylation profiles in the GI tract can influence both symbiotic/commensal and pathogenic bacteria. Together with the long-term maintenance of alleles conferring differences in *B4galnt2* expression suggests that host-pathogen interactions in the gut are likely involved.

As not much was known about the biogeographic influences on the intestinal microbiota we described and thus shed light on the effect of different factors (*i.e.* geography, host population structure, maternal transmission) on them. We performed a survey of eight house mouse populations throughout western Europe, by applying high throughput pyrosequencing of the bacterial 16S rRNA gene, we obtained the microbiota composition of those mice, microsatellites were used for estimating hosts population structure and sequencing of the mitochondrial D-loop region for the inference of maternal inheritance. Geography was found to be the most dominant factor shaping the bacterial composition,

followed by host population structure and the interaction of both. Additionally we could identify several bacterial ”species” which showed significant correlation to the underlying population structure of their host, maternal lineages and geography.
We also performed a survey of *B4galnt2* allele frequencies of the eight sampled populations in France and Germany. We detected a clear pattern of allele frequency distribution with a clear decline of the RIIIS/J allele frequency in central France, where we would thus locate the selective pressure(s). To shed light on the population dynamics concerning *B4galnt2* allele frequencies we analysed the frequency distribution pattern according to population substructure (as estimated by microsatellites and mtDNA). To follow up on the question of possible phenotypic consequences of *B4galnt2* expression, likely on the intestinal microbiota, we analysed the gut microbiota composition according to *B4galnt2* genotype by keeping in mind the newly gained insights of Chapter 2. This survey allowed us to infer that local adaptation is the most likely explanation for these dramatic differences in allele frequencies, rather than population substructure (*e.g.* due to different colonization waves). We could also confirm the influence of *B4galnt2* expression on the microbiota composition in wild-caught mice as already described for lab mice. A direct pathogen-genotype association could not be detected, which might be due to our sampling size. By identifying the long-term maintenance of different *B4galnt2* expression patterns and elucidating the role of biogeography on the intestinal microbiota we were able to further characterize the evolutionary forces acting at *B4galnt2*. We shed light on the complex dynamics acting at the *B4galnt2* gene, but, until the beneficial consequence(s) of exhibiting the RIIIS/J allele is known, the story of *B4galnt2* will remain a mystery and has to be further explored.

# Declaration of Author's Contribution

In this thesis, I present my doctoral research. The design of the whole project was done by my supervisor Prof. Dr. John Baines and myself. The interpretations of all the different results were achieved during numerous discussions. Practical laboratory work as well as the major parts of the data analysis was conducted by me, with some exceptions:

**Chapter 1:**

This chapter is published as a letter in Molecular Biology and Evolution (2011), authors contributed as follows. The study on long-term balancing selection at the gene *B4galnt2* was done in collaboration with Inka Montero, who assisted with the mousework and Jill M. Johnsen together with Christine R. Brzezinski who perfomed the DBA lectin staining. Bettina Harr provided the samples of *Mus spretus*. I did the sequencing and the analysis and wrote the first draft of the manuscript. Jill M. Johnsen and Bettina Harr had helpful comments on the manuscript. John Baines was the supervisor of this project. He wrote the final manuscript. All authors read and approved the final manuscript.

**Chapter 2:**

Sampling all the french and german mouse populations has been done by me with the help of Hermann Autengruber, Emilie Hardouin, Knut Albrecht, Urs Benedikt Müller, Philipp Rausch and Ann-Kathrin Jarms. Silke Carstensen assisted me in the lab, Katja Cloppenborg-Schmidt and Sven Künzel prepared the samples for and conducted the 454 runs. Together with Jun Wang I wrote the first draft of the manuscript. Dirk Metzler helped during the analyses on population structure. Emilie Hardouin did the microsatellite genotyping and helped during the discussions on population structure. Jun Wang mainly processed the 454 data and worked on the microbiota analyses. John Baines was the supervisor of this chapter. The manuscript was written by me, Jun Wang and John

Baines. All authors read and approved the final manuscript.

**Chapter 3:**

All lab work and analyses were done by myself, except for the analyses on the indicator species analysis, which has been performed by Jun Wang. John Baines supervised the project.

The results of my thesis have contributed to the following publications:

**Linnenbrink M**, Johnsen J, Montero I, Brzezinski C, Harr B, Baines JF (2011) Long-term balancing selection at the blood group-related gene *B4galnt2* in the genus Mus (Rodentia; Muridae). Molecular Biology and Evolution, **28**, 2999-3003.

**Linnenbrink M**, Wang J, Hardouin EA, Künzel S, Metzler D and Baines JF. The Role of Biogeography in Shaping Diversity of the Intestinal Microbiota in House Mice (Molecular Ecology, in Review)

Miriam Linnenbrink John F. Baines

# General Introduction

This dissertation comprises work of two different fields in biology - population genetics and metagenomics. What all three projects have in common is the model organism - the house mouse *Mus musculus*. Population genetics will be applied to the candidate gene *B4galnt2* in natural populations of all subspecies and a sister species of the house mouse. The metagenomics part concentrates on the microbial community in the gut of natural populations throughout France and Germany. In this introduction I will first introduce the house mouse as model organism. Thereafter I will focus on the background of population genetics and the candidate gene *B4galnt2* as well as on metagenomics and the gut microbiome.

## 1. The History of a Model Organism

The genus *Mus* emerged ∼5 MYA and can be divided into several subgenera as *Nannomys*, *Coelomys*, *Pyromys* and the subgenus *Mus* per se (Figure I-1, Guénet and Bonhomme (2003)). The subgenus *Mus* comprises several species from Asia (*M. caroli*, *M. cooki* and *M. cervicolor*), India (*M. famulus*) and Thailand (*M. fragilicauda*) which split ∼2-3 MYA. Parallel to the *M. musculus* complex *M. spretus* and *M.spicilegus*/*M. macedonicus* share a common ancestor ∼1.5 MYA. *M. musculus* is thought to have originated on the Indian subcontinent where the three major subspecies (*M. m. domesticus*, *M. m. musculus* and *M. m. castaneus*) split ∼0.5 MYA (Boursot *et al.*, 1993; Guénet and Bonhomme, 2003).The house mouse (*Mus musculus*, Schwarz & Schwarz 1943) is a commensal species which followed humans during their spread all over the world (Figure I-2) since the Neolithic (Auffray *et al.*, 2001; Cucchi *et al.*, 2005). While *M. m. castaneus* colonized the East and *M. m. musculus* central and eastern Europe, *M. m. domesticus* arrived in France over the Mediterranean see via the seaway and colonized western Europe ∼3000 years ago (based on palaeontological fossil records (Cucchi *et al.*, 2005)). From there it spread to

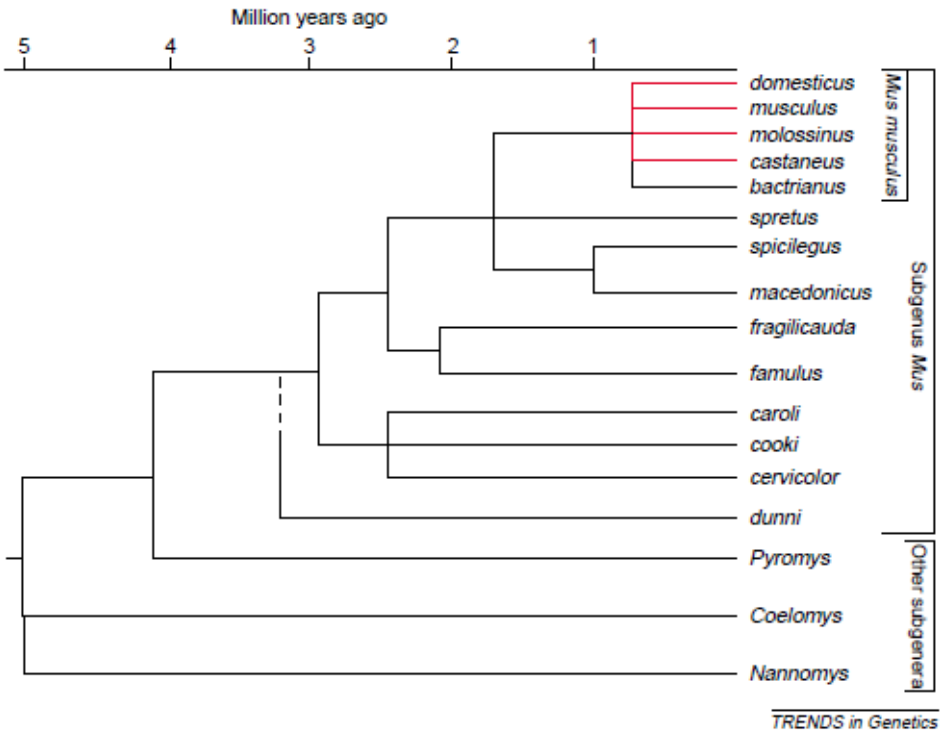


Figure I-1: Phylogeny of the genus *Mus* (taken from Guénet and Bonhomme (2003)).

America and the rest of the world in the last hundreds of years. The three subspecies of *M. musculus* are not completely genetically distinct. Hybrid zones exist between *M. m. musculus*/ *M. m. domesticus* and *M. m. musculus* /*M. m. castaneus* (Figure I-2). The hybrids between *M. m. musculus* and *M. m. castaneus* are forming one single population in Japan and are often referred to as *M. m. molossinus* (Yonekawa *et al.*, 1988). Even *M. m. domesticus* and *M. m. castaneus* could potentially form fertile hybrid offspring, but no contact zones exist between the two subspecies.

Its potential to reproduce in captivity, its small size, the short generation time and easy handling made the house mouse a very popular animal model in science, be this in biology or medicine (Kile and Hilton, 2005; Guénet and Bonhomme, 2003). Not forgetting the fact, that the house mouse is a mammal, which makes studies comparable to higher organisms, especially to humans. The first labstrains were set up by C.C. Little in 1921, based on a stock of Miss Abby Lathorp, those mice are the founders of the C57BL inbred line (http://www.informatics.jax.org/morsebook/index.shtml), the most common inbred mouse strain. Most inbred mouse strains are of polyphyletic origin of the three subspecies of *M. musculus*. *M. m. domesticus* is the primary source of labstrains, also *M. m. castaneus* and *M. m. musculus* contribute to their genetic set up (Wade and Daly, 2005;

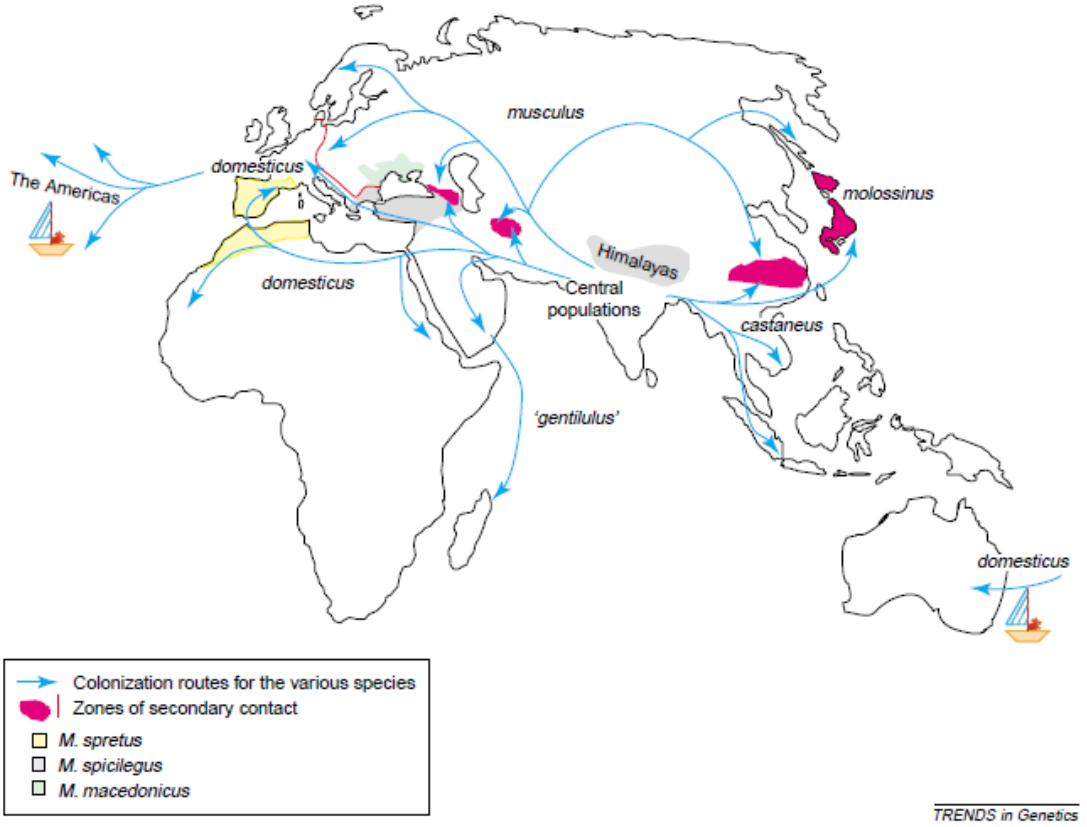


Figure I-2: Colonization routes of the different (sub)species of the genus *Mus* (taken from Guénet and Bonhomme (2003)).

Wade *et al.*, 2002). In the last decades the evolutionary history of mice has been studied extensively (Boursot *et al.*, 1993; Bonhomme *et al.*, 2011; Boursot *et al.*, 1996), as well as its behaviour (Mackintosh, 1970; Montero, 2010) and genetics (see for review Guénet and Bonhomme (2003); Kile and Hilton (2005)). In addition the whole genome sequence has been available for a decade (Mouse Genome Sequencing Consortium 2002) and thus makes it even more a perfect model system for studies in evolutionary biology.

In addition to the house mouse we included in our studies individuals of the species *M. spretus* (western Mediterranean mouse, Lataste 1883). Its species range is north Africa/ Portugal/ Spain/ southern France (Figure I-2). Thus, it occurs sympatrically with *M. m. domesticus* but they do not interbreed in nature (Palomo *et al.*, 2009), although this is possible in the lab. Contrary to the house mouse they are not commensal and also differ in some morphological traits, like coat color, body size or tail length (Palomo *et al.*, 2009).

# 2. Population Genetics

## 2.1 The History of Population Genetics

The basic principles of ”evolution is driven by natural selection” and that ”all organisms originate from a common ancestor but gradually modified over time” proposed by Darwin in his book ”On the Origin of Species” (1859), together with the letters of Gregor Mendel describing the three laws of inheritance (1865), served as basis for S. Wright, R.A. Fisher and J.B.S. Haldane to found the field of classical population genetics at the beginning of the $20^{th}$ century. Later on several other fields of biology beside genetics, also botany, morphology, palaeontology and ecology have been integrated to the concept of population genetics and was called ”The Modern Synthesis” (Huxley, 1942) a nowadays well-accepted concept of evolution. Bringing this together, today ”Population geneticists spend most of their time doing one of two things: describing the genetic structure of populations or theorizing on the evolutionary forces acting on populations” (Gillespie, 2004). Population genetics is defined as the study of the change of allele frequencies within populations.

## 2.2 The Main Principles

A central goal of evolutionary biology is to understand the forces underlying variation within and between populations and species. After Hartl and Clark (2007) a population is defined as a ”group of organisms of the same species living within a sufficiently restricted geographical area so that any member can potentially mate with any other member of the opposite sex”. The starting point for population geneticists constitutes the Hardy-Weinberg principle. This principle from $\sim$1908 states that under the assumptions of diploidy, sexual reproduction, discrete generations, infinite population size, no selection and random mating, genotype frequencies can be derived from the allele frequencies.

AA: $p^2$ Aa: 2pq aa: $q^2$ (see Hartl and Clark (2007))

Whereas AA, Aa and aa are the zygotes of any generation, $p^2$, 2pq and $q^2$ are the genotype frequencies and p and q the allele frequencies of ”A” or ”a”, respectively, from the previous generation. The allele frequencies of p and q add up to 1 (see Hartl and Clark (2007)). Random mating causes the same allele frequencies in the next generation as in the one before. This results in a constant distribution of allele frequencies that is due to Mendelian inheritance if no other evolutionary forces are present. Deviations from this so called Hardy-Weinberg Equilibrium thus facilitate the identification of allele frequency

changes.

Which reasons are possible for this deviation? The evolutionary forces shaping the genetic variation in populations are demography, random genetic drift, mutation, recombination and selection. Demography deals with the history of populations such as population bottlenecks, population growth, shrinkage or migration/gene flow due to environmental/social influences, thereby influencing allele frequencies. To assess important information on the demographic history of a population, population structure can give useful insights, for which Wright's $F_{st}$ (Wright, 1951) is a widely used estimator. Nowadays programs exist to estimate population structure, *e.g.* STRUCTURE, a "model-based clustering method for inferring population structure using genotype data consisting of unlinked markers" (Pritchard *et al.*, 2000; Falush *et al.*, 2003a). A second evolutionary force is genetic drift, which stands for random, neutral changes of allele frequencies over time. It removes genetic variation from populations. Mutation and recombination are two factors which bring in new alleles, hence, restoring allelic variation in populations. Some time after evolution and natural selection became commonly accepted theories, in 1983 Motoo Kimura presented his profound ideas of the "Neutral Theory of Molecular Evolution". The neutral theory does not exclude Darwin's idea of selection but it states that most changes that occur at the molecular level do not affect the individuals fitness, thus are (nearly) neutral and get fixed or lost by random genetic drift. The variation in a population in the end is dependent on the population size and the rate in which new mutations enter the genome. The neutral theory can, as population size and mutation rate affect the whole genome, thus be seen as null hypothesis in population genetics. Neutrality tests have been developed, so that observed data from natural populations can be tested against the neutral expectations. Either the null hypothesis of neutrality can be accepted, thus mutation and demography are the driving forces for molecular evolution, or discarded, leaving selection as a possible influencing factor. Tajima's *D* (Tajima, 1989) for example is a neutrality test, that is based on the site frequency spectrum, thus, on the nucleotide diversity which can be analyzed with two estimators: 1) $\pi$ (Tajima, 1983), which is sensitive to the allele frequency and takes the pairwise average between sequences and 2) $\theta_{Watterson}$ (Watterson, 1975), which solely counts the number of segregating sites, corrected by sample size. If only one locus is considered, Tajima's *D* equals 0 when all neutral expectations are met, Tajima's $D < 0$ can indicate directional selection or population expansion, and values $> 0$ can indicate balancing selection or population admixture.

Mainly two different types of selection can be distinguished (Hartl and Clark, 2007) - negative/purifying and positive selection. Negative selection removes detrimental alleles from populations. Positive selection either increases the allele frequency of beneficial

alleles (directional selection) or maintains allelic variation in populations (balancing selection). Selection can act either on coding sequences (*e.g.* Bustamante *et al.* (2005); Chamary *et al.* (2006)) resulting in changed proteins/ protein functions or on non-coding DNA, where selection can act on the level of gene expression, *e.g.* though changes in the *cis* - or *trans*- regulatory regions. (Wittkopp *et al.*, 2004; Hahn, 2007; Wray, 2007). The relative contribution of changes in the *cis*-regulatory versus the protein-coding regions of genes to the adaptive process has been the subject of recent debate (Carroll, 2005; Hoekstra and Coyne, 2007; Wray, 2007). Cowles *et al.* (2002) tried to quantify the extent of regulatory variation in inbred mouse strains. Even though it is hard to identify *cis*-regulatory elements as they can be quite far away of its targeted gene (Kleinjan and van Heyningen, 2005) and expression can differ depending on developmental stage and tissue, they still detected a significant amount of variation. A list of studies of individual genes under purifying and directional selection (for sexual signaling and local adaptation) and balancing selection, oftentimes due to host- parasite interactions (with variation in resistance to pathogens and/or susceptibility to disease) and along environmental clines has been reviewed in Hahn (2007). For example, a positively selected polymorphism in the *cis*-regulatory region of the *Interleukin-4* (*IL4*) locus in humans increases the inducibility of expression, leading to a more rapid response against harmful pathogens, but is associated with diseases such as asthma and atopic dermatitis, where the immune system appears to be inappropriately stimulated by otherwise harmless agents in the environment (Rockman *et al.*, 2003). Bamshad *et al.* (2002) detected a strong signature of balancing selection in the *cis*-regulatory region of the *CCR5* gene in humans. A strong haplotype structure could be identified in the direct 5' region, resulting in varying degrees of susceptibility to HIV-1 and the progression of AIDS.

## 2.3 Balancing Selection

As mentioned in the previous section 2.2, balancing selection maintains genetic variation in populations (see Charlesworth (2006) for review), whereby polymorphisms are preserved significantly longer than under random genetic drift. Different types of balancing selection can be distinguished. Under temporal and/or spatial variation in the environment either the one or the other homozygous allelic state can be advantageous (frequency dependent selection (Stahl *et al.*, 1999)) or heterozygotes serve a higher fitness in populations (overdominance; Kojima (1971); Allison (1954)). Classical examples for balancing selection are the MHC system in mammals (Apanius *et al.*, 1997), the *Rpm1* locus in

plants (Stahl *et al.*, 1999) or the human *ABO* gene (Calafell *et al.*, 2008), as well as the (although debated) textbook example of the peppered moth in England during the industrial melanism (Kettlewell, 1973; Cook, 2003). Balancing selection could also be identified in many other cases in plants (*e.g.* Filatov and Charlesworth (1999); Kamau and Charlesworth (2005), in insects (*e.g.* Fitzpatrick *et al.* (2007); Wheat *et al.* (2010); Cho *et al.* (2006), and mammals (Hiwatashi *et al.*, 2010; Newman *et al.*, 2006; Bamshad *et al.*, 2002). The *Oasb1* locus in house mice gives an example of long-term balancing selection where the maintenance of different alleles predates several speciation events and results in trans-species polymorphism (Ferguson *et al.*, 2008). Oftentimes the retention of various alleles is due to the identification or response to pathogens and therefore can influence the susceptibility/resistance to diseases (Hahn, 2007; Apanius *et al.*, 1997; Ferguson *et al.*, 2008; Andrés *et al.*, 2009; Fumagalli *et al.*, 2009). Andrés *et al.* (2009) surveyed the human genome specifically for genes under balancing selection. They detected an excess of disease related genes such as the *HLA-B* gene in the *MHC* and the *FUT2* locus, a blood group associated gene. Fumagalli *et al.* (2009) concentrated on balancing selection on blood group antigen genes (BGAs) and reported for several BGAs a correlation between ”pathogen richness” and different BGA gene variants. Another target of balancing selection is the gene *B4galnt2* ($\beta$-1,4-N-acetylgalactosaminyltransferase-2) locus in mice, which is a blood group related glycosyltransferase and will be discussed further in the next section.

## 2.4 An Example - *B4galnt2* as Candidate Gene

*B4galnt2* adds ”GalNAc” sugars to proteins and lipids and is responsible for the last step in the Sda and Cad blood antigen synthesis (Montiel *et al.*, 2003). The actual function of the gene is unknown yet. *B4galnt2* expression in the gastrointestinal tract (GI tract) is highly conserved in vertebrates from fish to humans (Stuckenholz *et al.*, 2009; Montiel *et al.*, 2003). Interestingly, some mice exhibit, beside the common gut expression, gene expression in the endothelium (blood vessel). This includes the inbred mouse strain RIIIS/J as well as some natural populations of the western house mouse (Johnsen *et al.*, 2008, 2009).

A tissue specific switch of *B4galnt2* expression from epithelial to endothelial cells, the site of von Willebrand Factor (VWF) synthesis (Jaffe *et al.*, 1973), leads to a detrimental phenotype, which is similar to the human bleeding disorder von Willebrand Disease type 1 (VWD). In humans, this is the most common inherited bleeding disorder with a preva-

lence of 1% (Rodeghiero *et al.*, 1987). The expression of *B4galnt2* in the blood vessels results in rapid clearance of VWF from the bloodstream (Mohlke *et al.*, 1996, 1999). As the VWF is an essential protein in the blood coagulation system this results in prolonged bleeding (Sweeney *et al.*, 1990). VWD in general has also been found to be present in a variety of mammals, besides in mice also in horses and calves (Brooks *et al.*, 1991; Sullivan *et al.*, 1994).

On the molecular level, Mohlke *et al.* (1996) located the region responsible for this tissue specific switch to chromosome 11, termed the mutation "Modifier von Willebrand Factor 1" (MvWF1) and showed that this autosomal dominant effect is independent of the vWF gene itself, which lies on chromosome 6. In 2008 Johnsen and colleagues found a highly conserved haplotype block spanning ∼97kb in 12 inbred mouse strains which also carry the MvWF1 which resembles clearly the sequence of the RIIIS/J strain. They also could narrow the *cis*-regulatory mutation to lie ∼30kb upstream of *B4galnt2*. Changes in the coding region could not be detected. A refined look on the sequence level reveals a clear peak of divergence (up to 12%) between the sequences of the lab strains RIIIS/J (RIII) and C57BL6/J (C57) in the *cis*-regulatory region (Johnsen *et al.*, 2008) which is several fold higher than the expected difference between inbred mouse strains (Wade *et al.*, 2002). Johnsen *et al.* (2009) detected a strong haplotype structure in the upstream region of *B4galnt2*, corresponding to the sequences of the C57 and RIII labstrains, respectively. Interestingly, they detected those two haplotype classes and with it the bleeding phenotype also in natural populations. Population genetic analyses revealed balancing selection as the most likely explanation for this unusual pattern of sequence divergence at *B4galnt2* (Figure I-3), even if introgression (*i.e.* the introduction of genetic material from one gene pool into another) cannot be completely ruled out (Johnsen *et al.*, 2009). Finally, all these observations raise several important evolutionary questions regarding the origin and maintenance of the two divergent haplotypes at *B4galnt2* and their phenotypic consequences. Concerning the phenotypic consequences two possibilities have to be considered, either selection acts in the gut epithelium and/or in the blood vessel endothelium. As blood-group related glycosyltransferases influence the attachment and colonization of both pathogenic and symbiotic/commensal microorganisms in the GI tract (Sonnenburg *et al.*, 2005), this may be one reason they are frequent targets of selection (Fumagalli *et al.*, 2010). Furthermore, it has been demonstrated that individual members of the microbiota forage upon glycans present on host mucosal surfaces when not provided dietary sources of polysaccharides (Bäckhed *et al.*, 2005). Thus, blood-group related glycosyltransferases represent good candidates for describing the evolutionary forces shaping host-microbiota interactions. Variation in those glycosyltransferases, and hence the composition of the

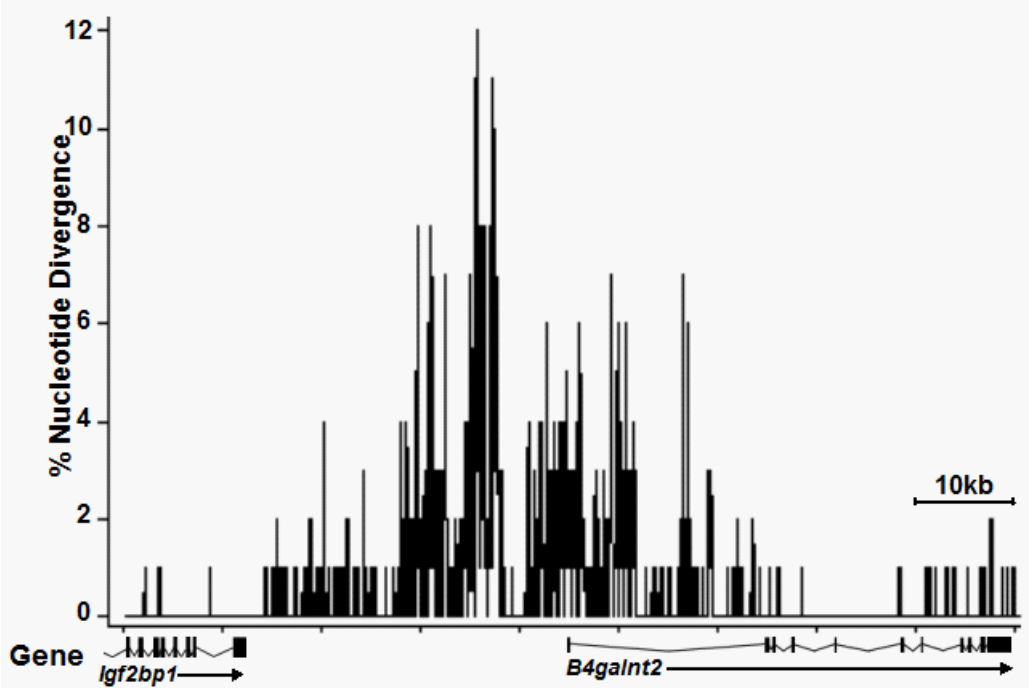


Figure I-3: Sliding window analysis of the nucleotide divergence between the sequences of the labstrains RIIIS/J and C57BL6/J in the gene region of *B4galnt2* (figure taken from Johnsen *et al.* (2008)).

sugars present in the GI tract, may have significant influences on the gut ecosystem, although this remains largely unexplored. This information makes the GI tract a good candidate to investigate the phenotypic consequences of the different *B4galnt2* alleles.

# 3. Metagenomics

## 3.1 A Global View on Metagenomics

In 1998 Jo Handelsman and colleagues first introduced the word ”metagenomics”, today also known as environmental genomics or community genomics. Put in wider terms, ”Metagenomics has also been more broadly defined as any type of analysis of DNA obtained directly from the environment” (Hugenholtz and Tyson, 2008). Microbiology took the first steps to explore the world of microorganisms, but as it is largely limited to the culturing of organisms and which is not possible fot the majority of microbes, a lot of information on microbial variation is invisible. After Riesenfeld *et al.* (2004) and Mongodin *et al.* (2005) culturing of 99% of the microorganisms is not possible by standard techniques. With advances in technological methods (*e.g.* pyrosequencing) genome se-

quences or at least portions of them (for bacteria mostly the 16S rRNA gene) from whole communities inhabiting the same enviromnment can be assessed. Culturing and even PCR-based methods in some cases can thereby be circumvented (Handelsman, 2005). With this large amount of information a new impression of the microbial world is possible. In general also viruses and archae can and already have been included in metagenomic studies (see Figure I-4), but most often metagenomics is referred to in relation to bacterial communities, as it also is in this dissertation, where we particularly focus on the gut microbiome.

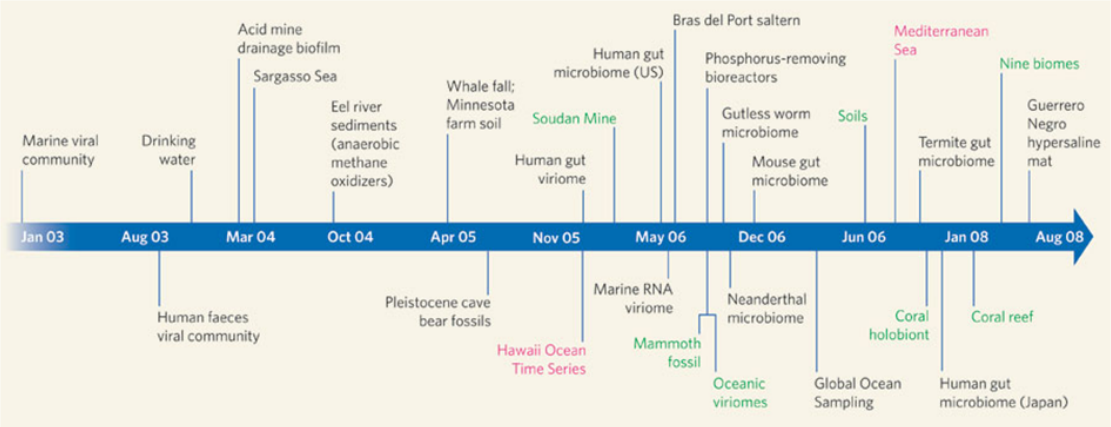


Figure I-4: Sequence based metagenomic projects from 2003 - 2008 (techniques used: in black shotgun sequencing, in red fosmid library sequencing and in green pyrosequencing; figure taken from Hugenholtz and Tyson (2008)).

## 3.2 An Introduction to the (Gut) Microbiome

Antoni van Leeuwenhoek (1632-1723) first discovered the so called ”animalcula”, today known as bacteria. He described them as broadly distributed and very diverse organisms and identified three different forms and their occurences ranging from rain water to the human mouth and intestine (reviewed in Smit and Heniger (1975)). After the field of microbiology was established many scientists started thinking about those tiny cohabitants of our world, meanwhile more than 10,000 species (http://www.bacterio.cict.fr/number.html) have been described. In the soil they decompose dead plant or animal material to their constituents, and cyanobacteria form together with green algae the phytoplankton which is an important nutrition for a diverse set of water organisms. Bacteria also show up on and in organisms like vertebrates, including humans (*e.g.* on skin and hair or in lungs, mouth, the GI tract, *etc.*) and form own communites. Great diversity of the community composition among different body parts of a healthy human has been described (Spor

*et al.*, 2011), even one single organ can house differentiated communities according to the location sampled (general: Costello *et al.* (2009), skin: Grice *et al.* (2009), GI tract: Staubach *et al.* (2012)). Especially the GI tract became of outstanding interest as the bacterial communities also within the different parts of it are highly diverse ($\sim$800 species and $\sim$10-fold more strains; Bäckhed *et al.* (2005)) and reach the highest cell densities known for any microbial habitat (up to 1012 cells/ml; Whitman *et al.* (1998)). So gut communites can also be described as "worlds within worlds" (Ley *et al.*, 2008) or as our microbial organ (Bäckhed *et al.*, 2005). The importance of the gut microbiome for vertebrates, including humans has been studied on several levels as these microbial communites are highly involved in physiological processes of their hosts as the metabolism of nutrients and organic substrates (Hooper *et al.*, 2002), the development of the intestinal epithelium (Falk *et al.*, 1998), the detection of foreign pathogens (Stecher *et al.*, 2010) and the maturation and development of the immune system. They also tend to influence the behaviour of their hosts (Heijtz *et al.*, 2011). Already recognized by Robert Koch 1876, bacteria don't live exclusively commensalistic or even mutualistic (*e.g. Bacteroides plebeius* (Hehemann *et al.*, 2010)) with their hosts, some rather occur as harmful agents (*e.g. Bacillus anthracis*, *Mycobacterium tuberculosis*). They can influence the health/disease status of humans and other vertebrates and are associated with diseases like type 2 diabetes, obesity, inflammatory bowel disease and many more (see Round and Mazmanian (2009) for Review), either due to a shift in their bacteria compositions, *i.e.* dysbiosis, or due to different community compositions. Despite their important roles gut bacteria achieved, still considerable variation exists between host individuals (Eckburg *et al.*, 2005). Little is known regarding the impact of the range of variation on host fitness.

Besides those effects of the microbiota on their hosts there is a second side of the coin - the effects of the host on the microbiota, regarding the relative roles of the contributing factors to this community variation (see also Walter and Ley (2011) for Review). The first step of bacteria colonizing vertebrates happens at birth where the first "basic" community is transmitted from the mother to the newborn (Koenig *et al.*, 2011; Palmer *et al.*, 2007). Even differences in the newborns and also children's gut communities can be observed between natural born offspring and offspring born via C-section (Koenig *et al.*, 2011; Dominguez-Bello *et al.*, 2010; Salminen *et al.*, 2004). Furthermore, the colonization history in total shapes microbial diversity (Emerson and Gillespie, 2008; Cavender-Bares *et al.*, 2009), colonization of *e.g.* environmental bacteria is thought to be influenced by the already resident community (Stecher *et al.*, 2010). Of course also the local environment the host is exposed to plays an important role in the composition of the gut microbiota (Spor *et al.*, 2011), and together with it the available local diet can cause changes to the

gut flora (*e.g.* Hehemann *et al.* (2010); De Filippo *et al.* (2010)). For vertebrates it could be shown that according to their diet categories (herbivore, omnivore, carnivore) there are essential differences in the gut morphology as well as in the fecal bacteria community (Ley *et al.*, 2008). Broader scale analysis of many mammalian hosts reveals that other individual bacterial lineages are found in multiple host species, suggesting they may be more promiscuous/environmentally-acquired (Ley *et al.*, 2008). Last but not least, the host itself is a contributing factor to its own microbiota. As already said above, the different body parts serve as different environments within the host and so assemble different bacteria communities. Also comparative studies between humans differing by varying degrees of genetic relatedness suggest a strong influence of host genotype (Zoetendal *et al.*, 2001; Rehman *et al.*, 2011; Sonnenburg *et al.*, 2005; Benson *et al.*, 2010), as does the co-evolution and -diversification of individual lineages together with their hosts (Falush *et al.*, 2003b; Oh *et al.*, 2010). A recent study of closely related hominid species examined the correspondence between intestinal communities and host phylogeny based on mitochondrial DNA (mtDNA) (Ochman *et al.*, 2010). A list of single host genes influencing the gut community can be found in Spor *et al.* (2011). It is likely that colonization of the intestinal environment arises from a spectrum of lineages, ranging from strong host-specific to promiscuous associations, although their relative contribution to overall community composition and structure remains largely unclear.

## 4. Scope of the Thesis

Overall this thesis aims to characterize the selective forces acting on *B4galnt2* and describes the dynamics of natural populations of house mice. Furthermore the gut microbiota communities of eight house mouse populations will be characterized and analyzed with respect to host genetics and the environment.

More specifically, the work comprises molecular insights of the *B4galnt2* gene locus in the three subspecies of *Mus musculus* and *Mus spretus*, as well as gene expression analyses for known and alternative alleles in the direct flanking region of the gene (Chapter 1). The fact, that sequence divergence between the two haplotypes at *B4galnt2* is much higher (up to 12%; Johnsen *et al.* (2008)) than the sequence divergence between *M. m. domesticus* and *M. spretus* ($\sim$2%; Galtier *et al.* (2004)) raises the question whether long-term balancing selection is acting in the upstream region of *B4galnt2*. For the second and third chapters eight house mouse populations in western Europe were sampled. Our dataset consisting of a panel of microsatellite loci, mitochondrial D-loop sequences and sampling location information allows us to shed light on the underlying population struc-

ture and the influence of genetic and environmental factors on the bacterial composition in the intestine. The second chapter serves as a basis for interpreting the anaylses concerning the intestinal microbiota with respect to the *B4galnt2* gene, which is part of chapter 3. With the refined information on allele frequency distribution of *B4galnt2* from France and Germany and the information about population substructure and the role of biogeography on the intestinal microbiota we address the question, whether local adaptation is a viable explanation for the great difference in allele frequencies or whether this is better explained by population substructure. We also analyze the gut microbiota with respect to the *B4galnt2* genotype in wild-caught mice and posed the question of whether a phenotypic influence is observable among highly variable microbioal communities.

# Chapter 1

# Long-term balancing selection at the blood group-related gene *B4galnt2* in the genus *Mus* (Rodentia; Muridae)

**Linnenbrink M**, Johnsen J, Montero I, Brzezinski C, Harr B, Baines JF (2011) Long-term balancing selection at the blood group-related gene *B4galnt2* in the genus *Mus* (Rodentia; Muridae). Molecular Biology and Evolution, **28**, 2999-3003.

## Abstract

Recent surveys of the human genome have highlighted the significance of balancing selection in relation to understanding the evolutionary origins of disease-associated variation. *Cis*-regulatory variation at the blood group-related glycosyltransferase *B4galnt2* is associated with a phenotype in mice that closely resembles a common human bleeding disorder, von Willebrand disease. In this study, we have performed a survey of the 5' flanking region of the *B4galnt2* gene in several *Mus musculus* subspecies and *M. spretus*. Our results reveal a clear pattern of trans-species polymorphism and indicate that allele classes conferring alternative tissue-specific expression patterns have been maintained for $> 2.8$ million years in the genus *Mus*. Furthermore, analysis of *B4galnt2* expression patterns revealed the presence of an additional functional class of alleles, supporting a role for gastrointestinal phenotypes in the long-term maintenance of expression variation at this gene.

The phenomenon of balancing selection, whereby natural selection acts to maintain multiple alleles in a population, may arise by a number of diverse processes including a heterozygote advantage, frequency dependent selection (Kojima, 1971) and temporal or spatial variation in selective pressures (Hedrick, 1976). Although the frequency and overall impact of balancing selection on the levels of diversity in natural populations has been a subject of debate since the collection of the first polymorphism data (Lewontin and Hubby, 1966), a diverse set of individual examples exists (see Charlesworth (2006) for a review). Some of the first genome-level studies cast doubt on the existence of appreciable balancing selection in the human genome (Asthana *et al.*, 2005; Bubb *et al.*, 2006). However, a recent landmark study of polymorphism data in the human genome demonstrates the clear significance of this mode of selection (Andrés *et al.*, 2009). Among their conservative list of 60 targets of balancing selection, roughly a third are known to be associated with human disease.

Studies of DNA sequence variation surrounding the blood group-related glycosyltranferase $\beta$ -1,4-N-acetylgalactosaminyltransferase-2 (*B4galnt2*) gene in house mice have uncovered the presence of two divergent haplotypes, one corresponding closely to the sequence of the RIIIS/J (RIII) inbred mouse strain and the other to the C57BL6/J (C57) strain (Johnsen *et al.*, 2008, 2009). The RIII allele carries a *cis*-regulatory mutation that directs a remarkable tissue-specific switch in *B4galnt2* gene expression from its more common site in intestinal epithelium (observed in C57) to vascular endothelium. Vascular expression of *B4galnt2* results in the aberrant glycosylation of the clotting protein von Willebrand Factor (VWF), leading to accelerated VWF clearance from circulation and low VWF levels (Mohlke *et al.*, 1999) similar to the common human bleeding disorder, Type 1 von Willebrand disease (Sweeney *et al.*, 1990).

Both the RIII allele and the C57 allele are found in inbred mouse strains and natural *Mus musculus domesticus* populations, where the relationship between *B4galnt2* genotype and tissue-specific expression was confirmed (Johnsen *et al.*, 2008, 2009). Striking signatures of natural selection were present in the populations studied by Johnsen *et al.* (2009), and simulation analyses revealed introgression alone as an unlikely explanation for these patterns. We proposed long-term balancing selection as the most likely explanation, but direct support for this hypothesis was lacking, as the study surveyed only a single subspecies.

To determine whether long-term balancing selection played a role in the generation of extreme sequence divergence (up to 8%) between *B4galnt2* haplotypes, we extended our previous survey to ancestral populations of three house mice subspecies (*M. m. musculus*, *M. m. domesticus* and *M. m. castaneus*), and the more distantly related *M. spretus*

(see Supplemental Material). The haplotypes previously described in *M. m. domesticus* from France extended over ∼60 kilobases. However, due to the ample opportunity for recombination, the signatures of long-term balancing selection are predicted to localize to narrow regions (Charlesworth, 2006). Thus, we increased the density of sequence fragments spanning the peak of polymorphism observed by Johnsen *et al.* (2009) and added the additional fragments to our previous data (Figure A-S1 and Table A-S1). In addition, we sequenced seven reference loci in the *M. spretus* sample (Table A-S2), for which data from all other populations was available (Baines and Harr, 2007) and thus provide the first information on DNA sequence polymorphism in this species.

As previously observed in *M. m. domesticus* from France, a peak of polymorphism approximately 10 kb upstream of the *B4galnt2* start codon is present in all species and populations, which displays a minimum of 3-fold, up to ∼13-fold higher levels compared to the panel of reference loci (Figure 1.1 a).In three of the five populations, the fragments with elevated polymorphism also display significantly positive Tajima's *D* (Tajima, 1989) values (Figure 1.1b). After closer inspection, the difference in Tajima's *D* between these and the remaining two populations is clearly due to differences in the frequency of divergent haplotypes (see below).

To analyze the pattern of haplotype variation, we estimated the phase of diploid sequences (Stephens *et al.*, 2001; Stephens and Donnelly, 2003). Two divergent haplotype classes similar in sequence to the RIIIS/J and C57BL6/J inbred mouse strains are ubiquitously present (Figure A-S2). However, the extent of linkage disequlibrium (LD) differed by population (Table 1.1). LD was highest in the *M. m. domesticus* population from France

Table 1.1: Linkage Disequilibrium.

| Population | Informative Sites[a] | Average $r^{2b}$ | Pairwise Comparisons | Significant Comparisons[c](%) | Range of Significant SNPs (bp) |
|---|---|---|---|---|---|
| MC | 51 | 0.9 | 1128 | 1035 (91.76 %) | 18878 bp |
| IR | 35 | 0.78 | 561 | 440 (87.43 %) | 12141 bp |
| KH | 68 | 0.55 | 2211 | 1322 (59.8 %) | 18933 bp |
| IN | 63 | 0.3 | 1953 | 472 (24.17%) | 17239 bp |
| SP | 75 | 0.43 | 2278 | 1288 (56.54 %) | 13655 bp |

[a] Sites with at least two copies of the rarer variant present in the sample.

[b] Composite genotypic $r^2$, the squared correlation of genotypic indicators at two loci in diploid individuals, was calculated using the composite_LD function submitted to the Bioperl project (Stajich *et al.*, 2002) as described in Johnsen *et al.* (2009).

[c] Based on the $\chi^2$ test.

(average $r^2$ = 0.9), followed by Iran (average $r^2$ = 0.78). Values from the other subspecies/species were comparatively lower (range 0.3 - 0.55). Due to the high number of

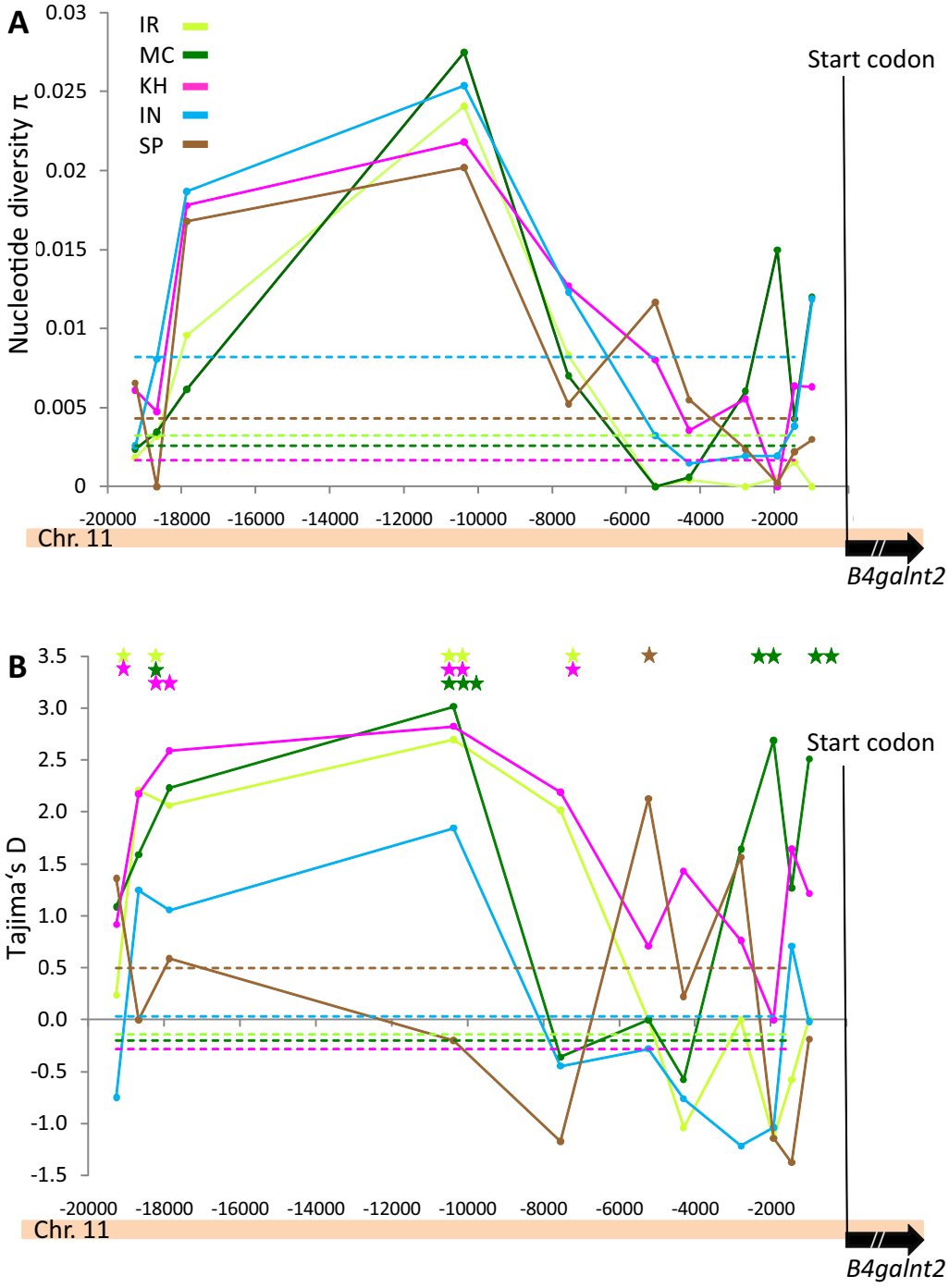


Figure 1.1: (A) Nucleotide diversity and (B) Tajima's *D* across the *B4galnt2* upstream gene region. Populations analyzed were *M. m. domesticus* from Iran (IR, light green) and France (MC, dark green), *M. m. musculus* (KH, pink), *M. m. castaneus* (IN, blue) and *M. spretus* (SP, brown). Dashed lines represent average values at seven autosomal reference loci (Baines and Harr, 2007) and this study (*M. spretus*). $^{*}\ p < 0.05$,$^{**}\ p < 0.01$,$^{***}\ p < 0.001$

pairwise comparisons no association is significant after Bonferroni correction.
To further investigate the relationship between haplotypes, we examined representative sequence fragments using phylogenetic analysis. For this we included a single individual of *M. famulus*, which shares a common ancestor with the studied taxa approximately 2.8 Myr ago (Ferguson *et al.*, 2008). The sequences at $\sim$ -2 kb and $\sim$ -20 kb cluster largely according to species (Figure 1.2a,c).However, the pattern observed $\sim$ 10 kb upstream of the start codon is particularly striking (Figure 1.2b). Although the three house mouse subspecies and *M. spretus* share a common ancestor 1.4 Myr ago (Ferguson *et al.*, 2008), the sequences clearly cluster according to allele class rather than by species, demonstrating a clear pattern of trans-species polymorphism. Furthermore, the single *M. famulus* is most closely related to the RIII allele class.
To test whether the *B4galnt2* expression patterns conferred by the RIII and C57 haplotype classes are conserved in other populations and species, we performed Dolichos biflorus (DBA) lectin staining, which is specific for *B4galnt2*-carbohydrate residues (Johnsen *et al.*, 2008, 2009). In wild-derived mice from *M. m. domesticus* (Iran) and *M. spretus* (Spain) we compared staining patterns with respect to *B4galnt2* genotype using amplicon #5 ($\sim$ -10 kb upstream) as a diagnostic marker (Table 1.2).Wild-derived *M. m. musculus* (Kazakhstan) individuals harbored a recombinant C57 and RIII haplotype, which we termed "CRK" class. Thus, additional amplicons (as in the population data; Table A-S1) were sequenced in those individuals in order to correlate haplotype classes with their expression phenotype (Table 1.2).

Table 1.2: Relationship between genotype (allele class) and DBA lectin staining pattern.

| Population | Allele Class | Gut+/Vessel- | Gut-/Vessel+ | Gut+/Vessel+ | Gut-/Vessel- |
|---|---|---|---|---|---|
| MC | C57/C57 | 8[a] | | | |
| (Johnsen et al. 2009) | C57/RIII | | | 10 | |
| | RIII/RIII | | 5 | | |
| | CRK/CRK | | | | |
| IR | C57/C57 | 2 | | | |
| | C57/RIII | 5 | | | |
| | RIII/RIII | | | | 10 |
| | CRK/CRK | | | | |
| SP | C57/C57 | 1 | | | |
| | C57/RIII | 7 | | | |
| | RIII/RIII | | | | 19 |
| | CRK/CRK | | | | |
| KH | C57/C57 | 1 | | | |
| | C57/RIII | | | 11 | |
| | RIII/RIII | | 6 | | |
| | CRK/CRK | | | 3 | |

[a]Numbers indicate the sample size of each genotype tested per population- /species-of-origin.

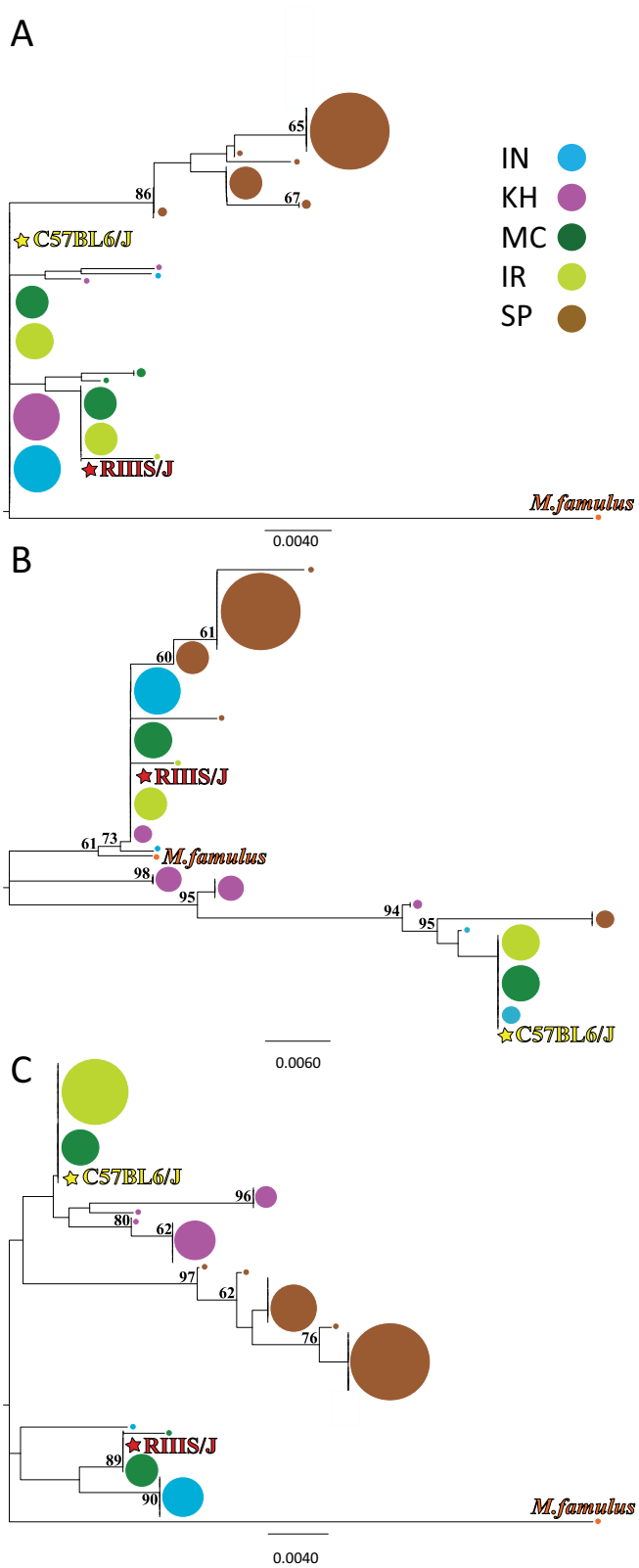


Figure 1.2: Neighbor-Joining trees of *B4galnt2* upstream regions. Sizes of the circles are proportional to the number of occurrences. Sequences of RIIIS/J and C57BL6/J were included for reference and *M. famulus* as an outgroup. Trees are from sequences located (A) $\sim$ 2 kb-(B) $\sim$ 10 kb-(C) and $\sim$20 kb upstream of *B4galnt2* (fragments 6.2, 5 and 3.5 in Table A-S1).

As expected, all individuals homo- or heterozygous for the C57 allele class exhibited lectin staining in the GI tract, and all individuals homozygous for the RIII allele class exhibited loss of GI staining. However, in contrast to the blood vessel positive DBA lectin pattern observed in RIII-homozygous individuals from France (Johnsen *et al.*, 2009) and Kazakhstan, all individuals with the RIII allele class from Iran and *M. spretus* failed to display a blood vessel staining pattern. Thus, a third functional class of alleles is present, which confers neither GI- nor blood vessel *B4galnt2* expression, but is related to the RIII-class alleles at the sequence level. To investigate whether this loss of expression might be due to a loss-of-function mutation, we sequenced all *B4galnt2* exons, intronic flanking sequences and UTRs of four *M. spretus* (one C57 and three RIII homozygotes) and five *M. m. domesticus* (Iran) individuals (two C57 and three RIII homozygotes), but detected no variants predicted to disrupt the transcript or protein (Figure A-S3). Surprisingly, CRK homozygotes displayed both bowel and blood vessel lectin staining indistinguishable from RIII-C57 heterozygotes, supporting a modular model of tissue-specific *B4galnt2* gene regulation in which two or more GI or vascular-specific regulatory elements lie within distinct genomic regions.

We here report a pattern of both great haplotypic and functional diversity at *B4galnt2* across the genus *Mus*. These results have important implications regarding the nature of the selective forces maintaining variation at *B4galnt2*. Together with extreme divergence between haplotypes, elevated polymorphism and significantly positive Tajima's *D* values, the presence of these two distinct haplotype classes in all subspecies of *Mus musculus* and the closely related species *M. spretus* provides strong evidence of long-term balancing selection. This adds to a growing list of examples of this mode of selection in mice such as $\beta$-globin (Storz *et al.*, 2007) and the *Oas1b* locus associated with West Nile virus infection (Ferguson *et al.*, 2008).

Interestingly, the common pattern across all species, subspecies and populations included in this study is the presence of B4galnt2-GalNAc residues on the GI tracts of any individual with the C57 allele class, and the loss of these residues in all individuals homozygous for the RIII allele class. Although the function of *B4galnt2* is unknown, gene expression is conserved in the GI tract in vertebrates from fish (Stuckenholz *et al.*, 2009) to humans (Montiel *et al.*, 2003). We speculate that selection on glycosylation in the gut is a contributing factor to the long-term maintenance of this variation, likely by altering glycan-specific host-pathogen interactions.

## Supplementary Materials (see Appendix A):

Animal material, Table A-S1, Table A-S2 and Figure A-S1, Figure A-S2 and Figure A-S3. A list of all primer pairs used in this study can be found in Appendix D (Table D-S1).

## Acknowledgements

We wish to thank Anja Hörger for helpful discussion, Katja Cloppenborg-Schmidt for technical assistance and an anonymous reviewer for helpful comments. This work was supported by American Heart Association Award 0575033N (JMJ), Puget Sound Blood Center (JMJ and CRB), and Deutsche Forschungsgemeinschaft (DFG) Grant BA2863/2-1 and Excellence Cluster ”Inflammation at Interfaces” (JFB).

# Chapter 2

# The Role of Biogeography in Shaping Diversity of the Intestinal Microbiota in House Mice

**Linnenbrink M**, Jun Wang, Emilie A. Hardouin, Sven Künzel, Dirk Metzler and John F. Baines
(Molecular Ecology, in Review)

## Abstract

The microbial communities inhabiting the mammalian intestinal tract play an important role in diverse aspects of host biology. However, little is known regarding the forces shaping variation in these communities and its influence on host fitness. To shed light on the contribution of host genetics, transmission and environmental factors on the diversity in microbial communities between individuals, we performed a survey of intestinal microbial communities in a panel of 100 house mice derived from eight locations across western Europe using high throughput pyrosequencing of the bacterial 16S rRNA gene. The host factors studied included population structure estimated by microsatellite loci, mitochondrial DNA and geography. We found geography, host population structure and their interaction to be critical to understanding the overall pattern of divergence in the composition and structure of the intestinal microbiota, although to different degrees, with geography being the strongest determinant. In addition, we identified individual bacterial species belonging to genera including *Bacteroides*, *Helicobacter* and *Robinsoniella* that displayed correlations with host genetic clusters, maternal lineages and geographic locations, enabling general properties of their life-history to be inferred.

## Introduction

Together with pioneering gnotobiotic animal studies (Bäckhed *et al.*, 2004; Rawls *et al.*, 2006; Turnbaugh *et al.*, 2009), the recent expansion of culture-independent analyses of bacterial communities has highlighted the biological significance those inhabiting vertebrate hosts, in particular the intestinal tract, so much so as to consider the intestinal microbiota a "forgotten organ" (O'Hara and Shanahan, 2006) or a "malleable third genome" (Carroll *et al.*, 2009). Although the majority of the details surrounding host - microbiota interactions have yet to be described, it is clear that the microbiota play an important role in diverse processes of host biology including as the metabolism of nutrients and organic substrates (Hooper *et al.*, 2002), the development of the intestinal epithelium (Falk *et al.*, 1998), the detection of foreign pathogens (Stecher *et al.*, 2010), the maturation and development of the immune system (Hooper, 2001; Round and Mazmanian, 2009) and even brain development and behavior (Heijtz *et al.*, 2011). The intestinal microbiota are highly diverse, composed of $\sim$800 species and $\sim$10-fold more strains (Bäckhed *et al.*, 2005) and despite their important roles also vary considerably between individuals (Eckburg *et al.*, 2005). While gross imbalances in these communities (dysbioses) are linked to inflammatory bowel disease (IBD) (Ott *et al.*, 2004; Round and Mazmanian, 2009), little is known regarding the impact of the "normal" range of variation on host fitness, nor regarding the relative roles of the contributing factors to this variation, be they genetic and/or environmental (Ley *et al.*, 2006). Comparative studies between humans differing by varying degrees of genetic relatedness suggest a strong influence of host genotype (Zoetendal *et al.*, 2001), as does the co-evolution and -diversification of individual lineages together with their hosts (Oh *et al.*, 2010). However, other individual phylotypes are found in multiple host species, suggesting they may be more promiscuous /environmentally-acquired (Ley *et al.*, 2008). Thus, it is likely that colonization of the intestinal environment arises from a spectrum of such lineages, ranging from strong host-specific to promiscuous associations. A recent study of humans and four species of great apes revealed a correspondence between host phylogeny based on mitochondrial DNA (mtDNA) and the relationship of fecal communities to one another (Ochman *et al.*, 2010). However, the mtDNA phylogeny explained only 25% of the variation in the microbial community tree, leaving the rest to be explained by other factors such as geography, diet and health status. Furthermore, the relationship to host phylogeny based on mtDNA may also be confounded if maternal transmission plays a large role in community assembly.

In this study, we have performed a large biogeographical survey of natural *M. m. domesticus* populations in order to simultaneously assess the contribution of host genetics,

maternal transmission and geography to the diversity in intestinal microbial communities between individuals, *i.e.* beta diversity. For each of 100 individual mice sampled from eight populations spanning a continuous area of Germany and France, we collected data from a panel of microsatellite loci, sequenced the mitochondrial D-loop and profiled the intestinal microbiota by performing barcoded 454 pyrosequencing of the bacterial 16S rRNA gene. By performing an analysis of host population structure and geography, we found both variables to be significantly correlated to the patterns of divergence in the composition and structure of the intestinal microbiota, with geography being the stronger determinant. However, we also identified individual bacterial species among these communities that displayed correlations with host genetic clusters and maternal lineages in addition to geographic locations.

## Material and Methods

### Animal Material and Tissue Sampling

121 mice were sampled in barns and stables in eight geographic locations throughout a continuous area of Germany and France in the summers of 2009 and 2010 (Table 2.1). The sampling strategy described by Ihle *et al.* (2006) was followed, maintaining a mini-

Table 2.1: General information on sampling locations and molecular data.

| | | Location | | | | |
|---|---|---|---|---|---|---|
| Area | Year | Country | Latitude | Longitude | Microsatellites and mtDNA | 16S rRNA |
| Cologne/Bonn | 2010 | Germany | 51° 0’ 6.00”N | 7° 2’18.00”E | 15[a] | 11 |
| Schömberg/Langenbrand | 2010 | Germany | 48°47’31.80”N | 8°38’ 7.38”E | 12 | 12 |
| Severac le Château | 2009 | France | 44°19’ 6.13”N | 3° 3’57.00”E | 18 | 14 |
| Espelette | 2009 | France | 43°21’10.49”N | 1°26’49.34”W | 22 | 16 |
| Angers | 2009 | France | 47°27’11.48”N | 0°35’41.77”W | 18 | 14 |
| Nancy | 2010 | France | 48°39’32.39”N | 6° 8’29.41”E | 12 | 10 |
| Louan-Villegruis | 2010 | France | 48°37’57.53”N | 3°29’ 4.10”E | 12 | 11 |
| Divonne les Bains | 2010 | France | 46°22’35.04”N | 6° 7’12.77”E | 12 | 12 |

[a] Numbers indicate the sample size per analysis

mum distance of one kilometer between sampling sites, only single individuals from any given sampling site were included in the analysis. Thus, 121 distinct sampling sites were included. The cecum was the location of the gastrointestinal tract chosen for study due to its large size, diverse microbial communities and important role in hindgut fermentation. Mice were dissected directly in the field and the cecum tissue was carefully flayed,

separated from its contents and stored in 1.5ml ice cold RNALater for future processing according the manufacturer's instructions.

### DNA Extraction

Mouse DNA was isolated from ear punches using the Dneasy Blood & Tissue Kit (Qiagen). Bacterial DNA was extracted from cecum tissue using a modification of the QIAmp DNA stool mini kit (Qiagen). Tissue samples were transferred to 2 ml screw-cap tubes containing 50 mg each of 0.1 mm, 0.5 mm and 1 mm glass beads (BioSpec Products). Tubes containing beads were treated with UV exposure for 2 hours prior to extraction. After adding 1.4 ml ASL lysis buffer, samples were subjected to bead beating using the Precellys (Peqlab) for 3 x 15 sec at 6500 rpm. Samples were then heated to 95C for 10 minutes, after which the manufacturer's protocol was followed.

### Molecular Analysis of Mouse DNA

#### D-loop Sequencing

A $\sim$950 bp portion of the mitochondrial D-loop was sequenced as described by Prager *et al.* (1993). Sequencing reactions were performed using ABI Big Dye v3.1 sequencing chemistry (Applied Biosystems, Foster City, CA) and run on an ABI 3730 automated sequencer. Sequence chromatograms were edited using Seqman (DNASTAR, Inc., Madison, Wisc.) and aligned using ClustalW (Thompson *et al.*, 1994) in the program MEGA 4.0.2 (Tamura *et al.*, 2007).

#### Microsatellite Genotyping

A set of 18 unlinked autosomal microsatellites described in Thomas *et al.* (2007) were chosen to perform analyses of population structure. These include: Chr01_25, Chr02_01, Chr03_21, Chr03_24, Chr04_31, Chr05_15, Chr05_45, Chr07_38, Chr08_11, Chr09_20, Chr11_64, Chr12_05, Chr13_22, Chr14_16, Chr16_21, Chr17_09, Chr18_08 and Chr19_08 (Hardouin *et al.*, 2010). PCR reactions containing forward primers labeled with either FAM or HEX were performed on 10 ng DNA template using a Multiplex PCR kit (Qiagen). PCR products were then diluted 1:20 in water and 1$\mu$l of diluted product was added to 10$\mu$l HiDi formamide and 0.1$\mu$l of 500 ROX size standard before running on an ABI 3730 automated sequencer (Applied Biosystems, Foster City, CA). The alleles were analyzed using GeneMapper 4.0 (Applied Biosystems, Foster City, CA).

### Population Genetic Analysis

For the analysis of the mitochondrial DNA we used one sequence of *M. m. domesticus* (GenBank accession number: AM182648) as a reference and one sequence each of *M. spretus* (GenBank accession number: U47539), *M. spicilegus* (GenBank accession number: U47536) and *M. famulus* as outgroups. A NeighbourNet network was constructed using the program SplitsTree (Huson, 1998; Huson and Bryant, 2006). For microsatellites, the observed and expected heterozygosities, average number of alleles were calculated using Arlequin (Excoffier *et al.*, 2005) and Cavalli-Sforza Chord distance (CAS) was calculated using Microsatellite Analyser (MSA, Dieringer and Schlötterer (2003)).

### Population Structure

Two measures of genetic differentiation were used: Slatkin's $R_{st}$ (Slatkin, 1995) for microsatellites and $\gamma_{st}$ for mitochondrial D-loop sequence. Slatkin's $R_{st}$ is a measure analogous to Wright's $F_{st}$ (Wright, 1951) that is adapted to the high rate of stepwise mutations occurring at microsatellites. $\gamma_{st}$ is a direct measure of differentiation between populations based on the mean number of pairwise differences. Isolation by distance (Wright, 1943) was tested for by applying a Mantel test to both datasets. Population substructure was investigated using the software STRUCTURE version 2.3.1 (Pritchard *et al.*, 2000; Falush *et al.*, 2003a). The parameters used were 500,000 burn-in period and 1,000,000 MCMC simulations with four iterations per number of clusters (K) for K equals 2 to 12. For the choice of K we applied the criterion of Evanno *et al.* (2005).

### Pyrosequencing of the 16S rRNA Gene and Sequence Processing

The 27F-338R primer pair spanning the V1 and V2 hypervariable regions of the bacterial 16S rRNA was used for PCR amplification and barcoded pyrosequencing on the 454 GS-FLX platform with Titanium sequencing chemistry as described in Rausch *et al.* (2011). Raw sequences were filtered using Mothur (Schloss *et al.*, 2009) requiring sequences to have a mean quality score $> 20$ and minimum length of 250 bp. Exact matches of MID barcodes were used to assign sequences to samples. Chimera detection and sequence clustering of operational taxonomic units (OTUs) were carried out uisng Usearch/Uchime (Edgar, 2010, 2011). Sequences were classified using RDP classifier (Wang *et al.*, 2007). A phylogenetic tree of the bacterial genera with abundance $> 0.1\%$ was generated in ARB (Ludwig *et al.*, 2004).

### Statistical Analyses on the Gut Microbial Community

Bacterial community analyses including the Bray-Curtis dissimilarity index, analysis of dissimilarity (Adonis) and constrained ordination of bacterial communities were carried out using the "Vegan" R package (R Development Core Team, 2011; Oksanen *et al.*, 2011). Phylogenetic alpha-diversity (Faith, 1992) and beta-diversity (Unifrac, Lozupone and Knight (2005)) were calculated using a tree produced by FastTree (Price *et al.*, 2009) implemented in Mothur. To test the significance of factors contributing to the divergence in bacterial communities, we used the R package "ncf" (Bjornstad, 2012) to perform Mantel and partial Mantel tests between the Bray-Curtis index, genetic distance (CAS) and geographic distance. The SIMPER method implemented in the Primer-E package (Clarke and Gorley, 2006) was applied to 96% similarity level OTUs to test for specific taxa that are correlated to genetic clusters defined by the STRUCTURE analysis (mice with $\geq$80% probability of belonging to one founder population were considered to belong to this genetic cluster) and/or sampling location (geography). The SIMPER method identifies OTUs contributing to similarity within- and dissimilarity between groups and ranks their contribution. For our analysis, significant OTUs (ANOVA; $p < 0.05$ after FDR correction) contained within the top 80% of the cumulative rank for both similarity within- and dissimilarity between groups were chosen, and a consensus taxonomy at the genus level was made for each OTU. To identify taxa that are correlated to the maternal lineage, OTUs were analyzed with respect to a maximum likelihood tree of the mtDNA sequences (generated using Mega 5 under the default settings (Tamura *et al.*, 2011) using Blomberg's K estimation (Blomberg *et al.*, 2003) implemented in the R package "ape" (Paradis *et al.*, 2012).

## Results

In total, we sampled 121 *M. m. domesticus* individuals from eight geographic locations throughout a continuous area of Germany and France in the summers of 2009 and 2010 (Fig. 2.1). To avoid the influence of inbreeding within nests, only single mice from distinct sampling sites ($> 1$ km from one another) were included (Ihle *et al.*, 2006). Bacterial 16S rRNA gene profiles meeting our inclusion criteria (*i.e.* a minimum of 650 reads; see below) were generated for 100 individuals and the majority of the analyses are based on this subset. However, mtDNA sequences and microsatellite data were generated for all samples and were utilized where appropriate (*i.e.* all samples were included in the STRUCTURE analysis to aid in assigning individuals to populations and taxonomic sta-

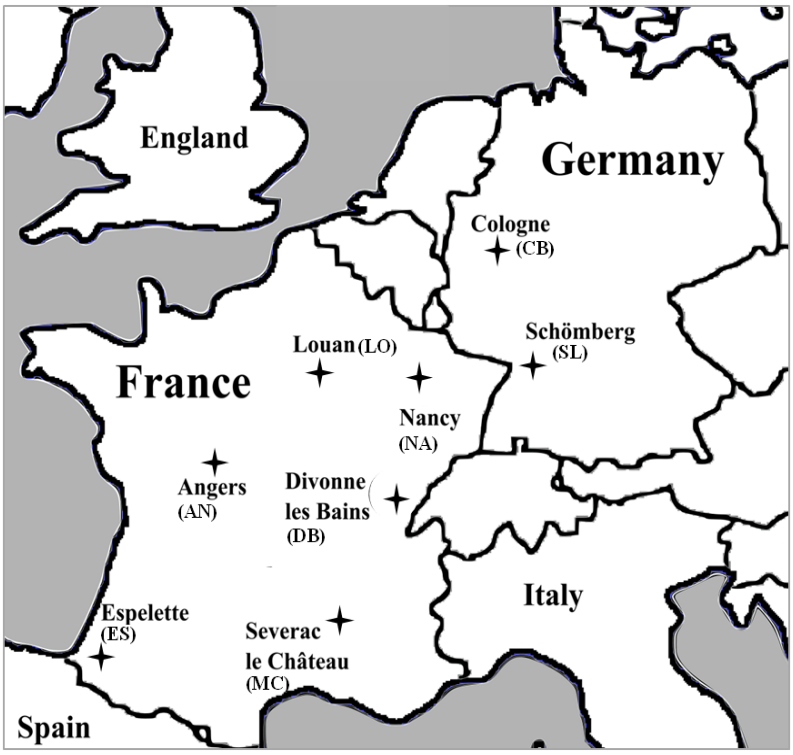


Figure 2.1: Sampling locations across Western Europe.

tus was confirmed using mtDNA; see below, Table 2.1).

### Host Population Structure

To confirm the taxonomic status and evaluate population structure based on the maternal lineage, we sequenced ∼950bp of the mitochondrial D-loop region of all 121 mice. However, because sufficient 16S rRNA sequence coverage for bacterial community analysis was achieved for only a subset of 100 individuals, further mtDNA analyses were limited to this subset. A total of 45 haplotypes were present in the 100 fully analyzed individuals (Table 2.2). A NeighbourNet network of the sequences reveals the presence of six closely related haplogroups that correspond to a subset of the haplogroups described in a much larger survey of *M. m. domesticus* including 1313 mtDNA sequences (Bonhomme *et al.*, 2011) (Fig. 2.2). An analysis of $\gamma_{st}$ with respect to geographic distance revealed no evidence of isolation by distance (Spearman's rank correlation; r = 0.0925, p = 0.30817).
To evaluate population structure based on the nuclear genome, we analyzed 18 unlinked microsatellite loci (Table 2.2) using STRUCTURE (Pritchard *et al.*, 2000; Falush *et al.*, 2003a). After conducting four independent runs for each number of clusters (K) ranging from 2 to 12, we applied the criterion of Evanno *et al.* (2005) and chose K = 5 for

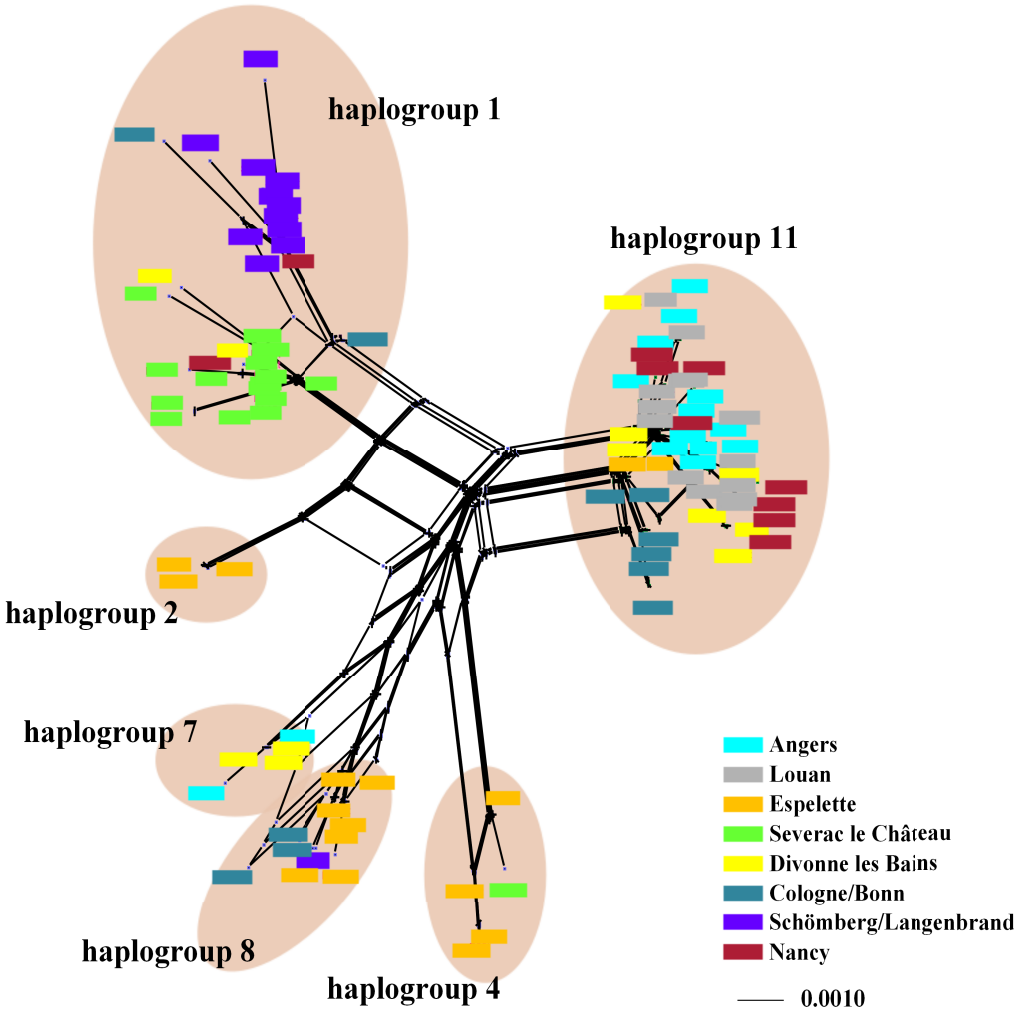


Figure 2.2: NeighbourNet network of 100 mitochondrial D-loop sequences. The numbering of the haplogroups corresponds to that of Bonhomme *et al.* (2011).

Table 2.2: Summary of mitochondrial D-loop and microsatellite variation.

| | D-loop sequences | | 18 autosomal microsatellite loci | | | |
|---|---|---|---|---|---|---|
| population | $N_{ind}$ | $N_{haplo}$ | N | $H_{exp}$ | $H_{obs}$ | $A_{av}$ |
| DB | 12 | 7 | 24.00 | 0.90 | 0.60 | 8.44 |
| LO | 11 | 3 | 23.56 | 0.79 | 0.67 | 6.83 |
| NA | 10 | 7 | 23.78 | 0.84 | 0.67 | 8.22 |
| SL | 12 | 6 | 22.56 | 0.81 | 0.54 | 7.78 |
| AN | 14 | 8 | 35.44 | 0.83 | 0.61 | 9.17 |
| ES | 16 | 8 | 42.44 | 0.79 | 0.59 | 9.72 |
| CB | 11 | 9 | 29.33 | 0.87 | 0.62 | 10.06 |
| MC | 14 | 6 | 35.67 | 0.86 | 0.71 | 10.06 |
| mean | 12.50 | 6.75 | 29.60 | 0.84 | 0.63 | 8.78 |

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

the number of clusters (Fig. 2.3). The two southernmost localities in France (Severac le

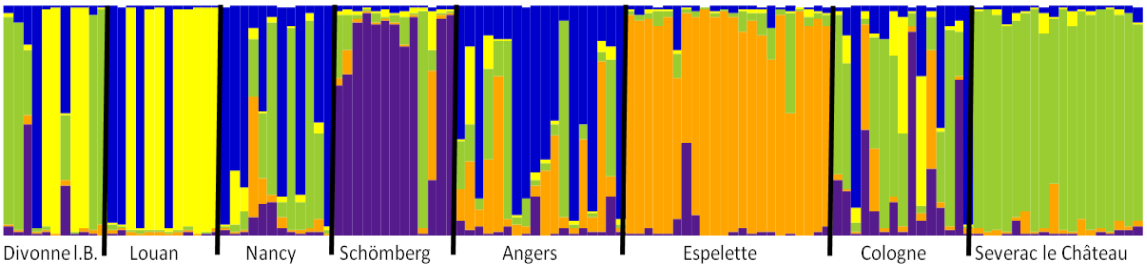


Figure 2.3: Genetic clusters identified by STRUCTURE for K=5.

Château and Espelette) and Schömberg/Langenbrand displayed the most homogeneity in ancestry, being composed of nearly single clusters, while the other localities displayed multiple clusters and/or admixed individuals. A test for isolation by distance based on $R_{st}$ revealed no significant correlation with geographic distance (Spearman's rank correlation; r = 0.125, p = 0.27987).

### Analysis and Structuring of the Gut Microbiota

For the analysis of intestinal bacterial communities, we set a cutoff of 650 reads as the minimal sequence coverage for individual mice. This resulted in a total of 212,699 sequences in the dataset, giving an average of over 2,000 reads per individual. In total 16

phyla were identified, with the three most abundant being Bacteroidetes (43.89%), Proteobacteria (25.91%) and Firmicutes (25.59%) (Fig. 2.4). Two other phyla, the Spirochae-

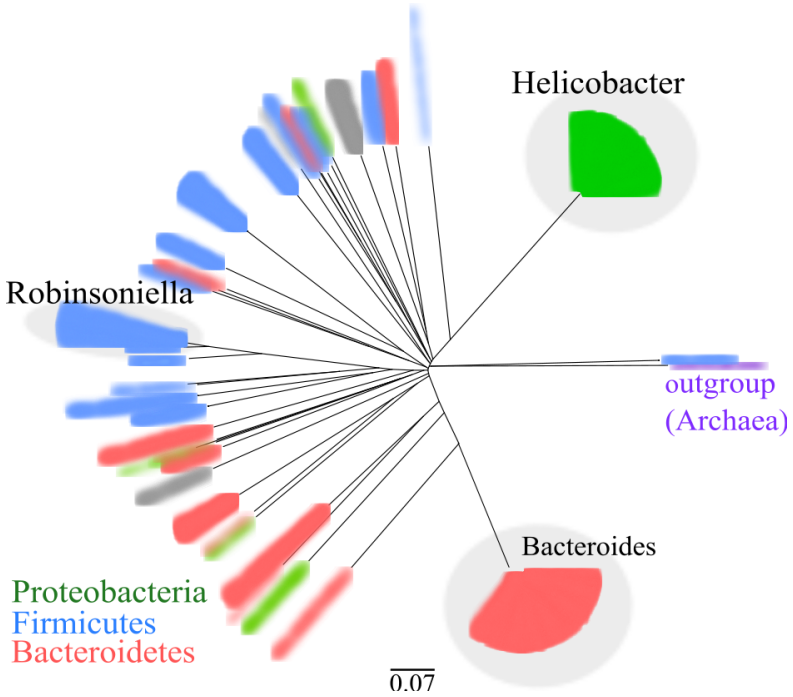


Figure 2.4: Phylogenetic tree of bacterial 16S rRNA gene sequences. The width of each taxonomic group denotes its relative abundance in the entire dataset (n = 100 individuals).

tes and Fusobacteria, were abundant only in single individuals. One individual from Angers displayed ∼39% Spirochaetes, and Fusobacteria were abundant in single individuals from Cologne/Bonn (∼26%) and Divonne les Bains (∼30%). On the genus level, the most abundant taxa were the *Bacteroides* (31.67%), *Helicobacter* (24.03%) and *Robinsoniella* (4.85%), which were also respectively the most abundant members of the three major phyla Bacteroidetes, Proteobacteria and Firmicutes (Fig. 2.4).

To evaluate the overall differences in bacterial community composition and structure between individuals and sampling locations, we first analyzed beta diversity using the Bray-Curtis dissimilarity index. This measure provides an estimate of community divergence based on the relative abundance of taxa (OTUs) between samples. Using a geographic location-constrained principle coordinate analysis (PCoA) of Bray-Curtis dissimilarity, the bacterial communities form four distinguishable clusters (Adonis $r^r$ = 0.1289, p = 0.001, see Fig. 2.5). The DB, LO, SL, NA localities group together, AN and CB form a smaller cluster, while MC and ES form two individual clusters.

In order to test the contribution of individual variables that may contribute to overall bacterial community divergence, we first performed Mantel tests to estimate the variation in

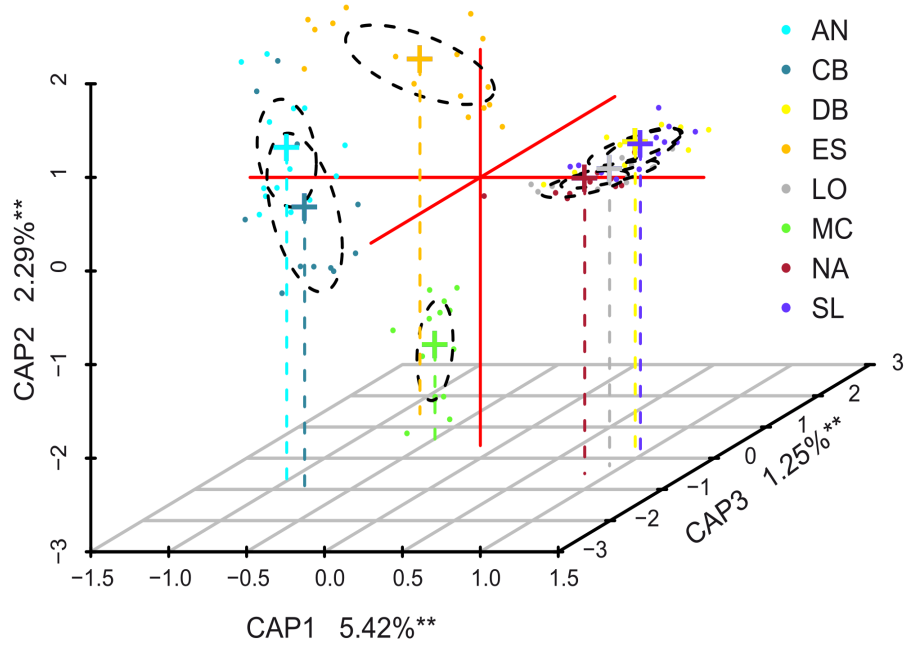


Figure 2.5: Geographic location-constrained principle coordinate analysis (PCoA) of Bray-Curtis dissimilarity. The first 3 axes are presented, which account for 8.96% of the total variation. ** $p < 0.01$.

the Bray-Curtis index that can be explained by the genetic distance of the host and the geographic distance between the individual sampling sites. This revealed significant contributions of both geographic distance and genetic distance as measured by the Cavalli-Sforza Chord distance (geographic distance: r = 0.2284, p = 0.0019; CAS: r = 0.0829, p = 0.0199). However, because genetic distance is correlated to geographic distance (r = 0.2305, p = 0.0019), partial mantel tests are necessary to distinguish the contribution of each single factor while controlling for the effect of the other. Both factors remain significantly correlated to the Bray-Curtis index, although the effect of geographic distance is more pronounced (geographic distance: r = 0.2158, p = 0.0019; CAS: r = 0.0319, p = 0.041).

While the above analysis reveals significant contributions of host genetics and geography to overall community divergence, individual bacterial lineages within communities may display contrasting patterns of association if they differ in factors such as their degree of host specialization and/or mode of transmission. Thus, to identify OTUs displaying associations to geographic location or genetic clusters, we applied the SIMPER method (Clarke, 1993), which focuses on the importance of individual taxa to the overall similarity within and between groups and ranks their contribution. Applying this method to geographic locations identified 56 significantly correlated OTUs belonging mainly to the

genera *Bacteroides* (23 OTUs) and *Helicobacter* (11 OTUs) in addition to twelve other genera containing one to four OTUs each (Fig. 2.6a). The analysis with respect to population substructure was performed using the five genetic clusters identified by STRUCTURE, revealing 23 OTUs also dominated by *Bacteroides* (13 OTUs) and seven other genera (Fig. 2.6b). As the two factors population structure and geography are in some cases closely related (*e.g.* ES and MC), it is also important to analyze their interaction. This revealed 30 OTUs, again dominated by *Bacteroides* (21 OTUs) in addition to six other genera (Fig. 2.6c).

Another possible factor to consider in our dataset is maternal transmission, which if present may be reflected by the mtDNA phylogeny. To identify OTUs whose abundance evolves along mtDNA lineages, we applied the phylogenetic method of (Blomberg *et al.*, 2003), which determines whether traits display significant evolutionary signals based on the expectation given the topology, branch lengths, and a Brownian motion model of evolution. Using this method we identified 15 OTUs with significant phylogenetic signals, with one to three OTUs belonging to each of twelve genera. Eight of these OTUs are uniquely associated with maternal transmission, while seven others are shared with the SIMPER analysis of geography and/or population structure (see Appendix B Fig. B-S1).

## Discussion

Determining the relative role of the environment and genetics on the gut microbiota is a challenging undertaking due to the complexity of these communities and the interaction of factors contributing to diversity. Several different categories of studies addressing this question have been performed including human twin studies, comparison and experimental manipulation of mouse lines, quantitative trait loci (QTL) mapping and analyses of single host genes (reviewed by Spor *et al.* (2011)). Although different twin studies have offered some conflicting results, by in large each of these approaches yields support for a role of host genetics. Using a different comparative evolutionary approach, Ochman *et al.* (2010) revealed a correspondence between the gut microbiota and host phylogeny among closely related hominid species, providing evidence for vertical inheritance among genetically differentiated hosts. Our study complements these previous studies by analyzing a large number of mice in their natural environment, with varying degrees of relatedness and spanning a large geographic area including 100 unique sampling sites. In addition, our focus on the cecal mucosal-associated community likely lends increased power to detect genetic effects compared to the lumen or feces, due to its more intimate association with the host (Spor *et al.*, 2011).

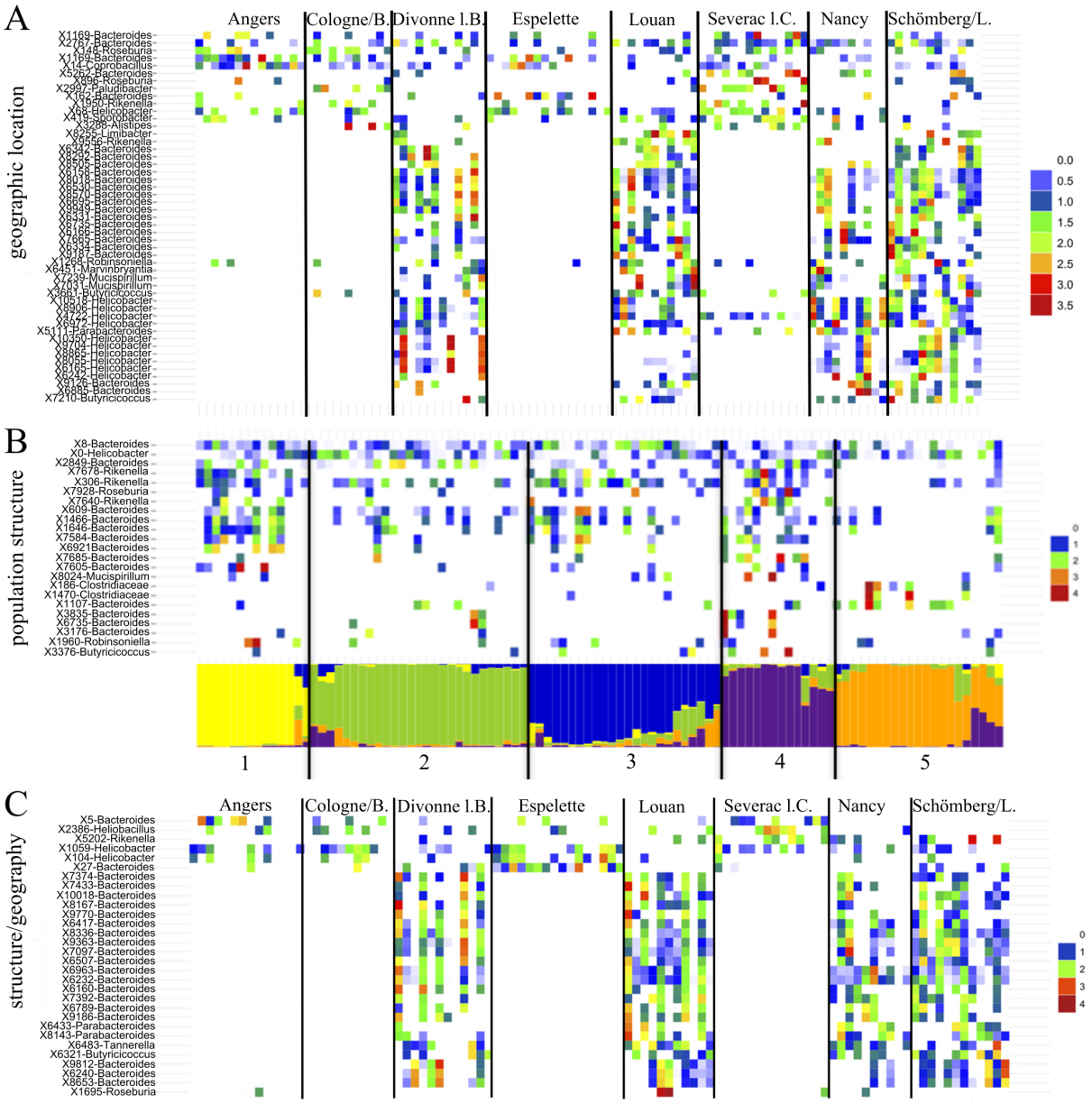


Figure 2.6: Heatmap of Indicator OTUs. Log relative abundances are displayed for OTUs with significant associations with A) geographic location, B) genetic clusters as determined by STRUCTURE and C) their interaction.

Overall, we determined the largest contributing factor to microbial diversity between individuals to be geography, as measured by the distance between sampling sites. This parameter can be viewed as an approximation of environmental effects, as it represents the sum of differences between sampling sites at multiple variables such as local weather patterns, availability of food sources, etc. Roughly speaking, this environmental component, which accounts for ∼22% of the variation in the Bray-Curtis index, is comparable to the estimate provided by the QTL study of Benson *et al.* (2010), who found environmental (litter and maternal) effects to account for 26% of the variation in abundances of a core set of bacterial taxa. Host genetic distance is also significantly correlated to the Bray-Curtis index, although it explains nearly seven-fold less variation (∼3.2%). It should be noted, however, that mice colonized Europe relatively recently, *i.e.* roughly 3,000 years ago (Cucchi *et al.*, 2005), and despite the presence of population structure or differences in historical colonization patterns (Bonhomme *et al.*, 2011), are still closely related. Thus, sufficient time for genetic differences with substantial effects on the microbiota to accumulate may be lacking in our study.

Our analysis of geographic *versus* genetic distances was performed on the individual level, but we also analyzed the structuring of our host samples using the clustering procedure provided by STRUCTURE (Pritchard *et al.*, 2000; Falush *et al.*, 2003a), which assigns individuals to populations without a priori knowledge of the population units. In contrast to absolute measures of genetic distance such as the Chord distance, this method provides no information regarding the relationship of populations to one another. Nonetheless, some nominal correspondence between the genetic clusters identified and the degree of separation of microbial communities based on the Bray-Curtis index exist. For example, two of the genetically most homogenous sampling regions (Severac le Château and Espelette) also show up as significantly distinct with respect to clustering based on the Bray-Curtis index, whereas the sampling regions displaying evidence of admixture do not (Fig. 2.4 and Fig. 2.5). Thus, it appears that similarity in microbial communities may in some cases reflect underlying population substructure but may also be similarly obscured by complex demographic histories. In addition to using biogeography as a means to estimate the relative contribution of the environment and genetics, we also posed the alternative questions of which specific taxa within the community contribute most to similarity within geographic regions, genetic clusters and mtDNA lineages. As expected if geography plays a predominant role, we identified the largest number of significantly correlated OTUs with respect to this variable, most of which belong to the genus *Bacteroides*. Fewer OTUs were significantly correlated to the genetic clusters identified by STRUCTURE (Pritchard *et al.*, 2000; Falush *et al.*, 2003a), but surprisingly they belonged largely

to the same genera (*Bacteroides* and *Helicobacter*, Fig. 2.6a,b). We also identified taxa, most belonging to *Bacteroides*, whose abundance vary with respect to genetic clusters as well as geography, suggesting an interaction between genetics and the environment (Fig. 2.6c). Indeed, these genera are highly diverse in these wild mouse samples (see Fig. 2.4), and it is tempting to speculate that they contain members that differ in their life-history strategy (*e.g.* generalists *versus* specialists) or mode of transmission (*e.g.* environmental *versus* horizontal).

Due to the known influence of maternal transmission on fecal community clustering (Ley *et al.*, 2005; Benson *et al.*, 2010), we sought to determine whether its signature is detectable on a broad scale and identify the community members most likely following this mode of transmission. Using the K statistic of Blomberg *et al.* (2003), we identified numerous species that displayed evidence of evolving along the mtDNA phylogeny. In contrast to the analysis of geography and genetic clusters based on nuclear markers, members of *Bacteroides* and/or *Helicobacter* were present, but did not dominate the identified taxa, and half the species identified were uniquely correlated to mtDNA evolution. A previous analysis of intestinal communities from the perspective of host development and ecological succession revealed that Bacteroidetes increase in abundance and become stable members of the community particularly after weaning (Rehman *et al.*, 2011). This may in part explain the more even distribution between Firmicutes and Bacteroidetes among the OTUs correlated to the mtDNA phylogeny and suggest those identified as candidates for some of the initial colonizers transferred via intimate contact with the mother.

Our analysis of the cecal mucosal communities in natural house mouse populations provides several important points of information for understanding the ecology and evolution of host-microbiota interactions. Despite our choice of a mucosal-associated community, which might be expected to display a higher degree of host association, the influence of host genetics in our study was still comparatively small. However, a detailed analysis of biogeography and mtDNA lineages enabled us to identify many interesting candidate species for future study. Although the genera *Bacteroides* and *Helicobacter* are well known and studied, our analysis of wild mice indicates that the genus *Robinsionella* is also likely of great significance, as it is the most abundant genus present among the Firmicutes and displays significant associations with respect to host biogeography and maternal trasmission. Further analyses including variation in intestinal bacterial communities both within and between mouse host subspecies may shed additional light on the nature of coevolution in this important model organism.

## Data Accessibility

All DNA sequences will be deposited in GenBank.
The bacterial 16S rRNA gene sequences reported in this paper will be deposited in the European Nucleotide Archive.

## Supplementary Materials (see Appendix B):

Table B-S1, Figure B-S1
A list of all primer pairs used in this study can be found in Appendix D (Table D-S1).

## Acknowledgements

We wish to thank Hermann Autengruber, Urs Benedikt Müller, Knut Albrecht, Philipp Rausch and Ann Kathrin Jarms for assistance in field work, Iris Fischer and Aurélien Tellier for helpful discussion regarding the analysis and Katja Cloppenborg-Schmidt and Silke Carstensen for excellent technical assistance. This work was supported by DFG Grants ME3134/3 to D.M. and BA 2863/2-1, BA 2863/2-2 and the Excellence Cluster ”Inflammation at Interfaces” to J.F.B.

## Author Contributions Box

ML, EAH and SK performed research. ML, JW, DM and JFB analyzed data. JFB conceived the research. ML, JW and JFB wrote the paper.

# Chapter 3

# Population Dynamics at *B4galnt2* and its Influence on the Intestinal Microbiota in Natural Populations of the Western House Mouse

**Miriam Linnenbrink** and John F. Baines

## Introduction

As already mentioned in the General Introduction, in some populations of house mice, *B4galnt2* is either expressed in the intestinal epithelium or blood vessel endothelium (Mohlke *et al.*, 1999), depending on its genotype in a $\sim$30kb upstream region (Johnsen *et al.*, 2008). This tissue specific switch of gene expression is due to a *cis*-regulatory mutation in the region flanking Exon 1 of *B4galnt2*. Expression in the blood vessel causes a harmful effect on the blood coagulation and results in a bleeding disorder called von Willebrand Disease (Mohlke *et al.*, 1999). To answer the question if those two allele classes (RIII or C57) and therefore the bleeding disorder is present in natural populations of house mice, Johnsen *et al.* (2009) sequenced one diagnostic fragment lying within the *cis*- regulatory region. The RIII allele class was present at intermediate frequency in three out of four tested populations (France (39%), Chicago (25%) and Cameroon (22%), Germany(0%)). The associated phenotype of a mild bleeding diathesis was confirmed by measuring VWF levels (Johnsen *et al.*, 2009). Interestingly the German population from Cologne/Bonn was monomorphic for the C57 allele and therefore the bleeding disorder could not be detected. Further analysis of the French and German populations revealed that the RIII allele class likely experienced a very recent increase in frequency in the French population.

The "trench warfare" hypothesis proposed by (Stahl *et al.*, 1999) describes a scenario where variation is maintained in a population. Changes in allele frequencies due to selection vary over space and time. Specifically, if a resistance allele carries a cost, it will increase in frequency in the presence of the pathogen and decrease in its absence. In the case of *B4galnt2* this hypothesis seems like a plausible explanation for the intriguing distribution of *B4galnt2* haplotypes in natural populations (Johnsen *et al.*, 2009).

In Chapter 1, we confirmed the long term maintenance of both allele classes since the species split of *M. famulus* ($> 2.8$ MYA) and identified different expression patterns in the subspecies of *M.musculus* and *M. spretus* (Linnenbrink *et al.*, 2011). Based on the fact, that *B4galnt2* expression in the gut is conserved (Stuckenholz *et al.*, 2009; Montiel *et al.*, 2003), yet is turned off in mice homozygous for the RIII allele class (Linnenbrink *et al.*, 2011), an intestinal phenotype seems a likely underlying factor contributing to the long-term maintenance of variation at this gene. Patterns of balancing selection and the involvement in pathogen resistance have already been observed for *ABO* (Stajich and Hahn, 2005; Fry *et al.*, 2008; Calafell *et al.*, 2008) and *FUT2* (Andrés *et al.*, 2009) in humans, both glycosyltransferases, expressed in the GI tract. Besides the information from studies of the *B4galnt2* gene itself, also the fact that host glycans play an important

role in host-microbe interactions (Bishop and Gagneux, 2007) underpins the importance to test if *B4galnt2* expression patterns influence the microbial composition in the intestine due to the presence or absence of GalNAc sugars. The contribution of single genes to the micobiota composition has already been found in several studies (Benson *et al.*, 2010; Vijay-Kumar *et al.*, 2010; Peterson *et al.*, 2007; Rausch *et al.*, 2011). Staubach *et al.* (2012) investigated the microbiota of *B4galnt2* of knockout vs. wildtype mice on a C57BL6/J (gut expression) background, and detected habitat preference of some bacterial species (OTUs) according to genotype and suggested that differing *B4galnt2* expression leads to changes in susceptibility to diseases in the gastrointestinal tract. Although the influence of *B4galnt2* on the microbiota could be detected in lab mice, information from wild mouse populations is still lacking and needs to be further explored.

This chapter aims to address the general evolutionary question of why disease-associated variation is maintained in natural populations. Therefore we use a population genetics approach to further describe the evolutionary forces contributing to differences in allele frequency throughout France and Germany and specifically ask the question whether population substructure or local adaptation is most involved. To this end we conducted a large biogeographical survey of *B4galnt2* allele frequencies in Western Europe where we include additional six populations from Germany and France (to the already known populations described by Johnsen *et al.* (2009). In addition, we expand the study of Staubach *et al.* (2012) to natural populations of house mice and analyze the intestinal microbiota from two French populations according to *B4galnt2* genotype. Chapter 2 will be taken into account with regard to population structure and the general influence of genetics and environment on the intestinal microbial communities of these mice.

## Material and Methods

In this Chapter we performed a geographic survey of *B4galnt2* haplotype frequencies. To approach the question why the disease associated RIII allele is maintained in some populations and not in others, we analyzed this distribution pattern with regard to host population substructure. Population structure has already been analyzed in Chapter 2 with regard to its role in influencing the intestinal microbiota, additional analyses have been performed for this chapter. As the bleeding phenotype is not restricted to lab mouse strains, but also present in nature, maybe the influence of *B4galnt2* appears between genotype and associated bacteria/pathogens. Here we investigate the same eight house mouse populations as we did in Chapter 2. Thus, some parts of this Material and Methods section will overlap with that of Chapter 2 and I will refer back where it is appropriate.

### Animal Material, Tissue Sampling and DNA Extraction

See Materials and Methods Chapter 2 and Table 3.1.

Table 3.1: General information on sampling locations and sample sizes.

| Area | Year | Country | Microsatellites and mtDNA | 16S rRNA | *B4galnt2* Genotype |
|---|---|---|---|---|---|
| Cologne/Bonn | 2010 | Germany | 15[a] | 11 | 15 |
| Schömberg/Langenbrand | 2010 | Germany | 12 | 12 | 12 |
| Severac le Château | 2009 | France | 18 | 14 | 19 |
| Espelette | 2009 | France | 22 | 16 | 22 |
| Angers | 2009 | France | 18 | 14 | 18 |
| Nancy | 2010 | France | 12 | 10 | 12 |
| Louan-Villegruis | 2010 | France | 12 | 11 | 12 |
| Divonne les Bains | 2010 | France | 12 | 12 | 12 |

[a] Numbers indicate the sample size per analysis

### Genotyping at *B4galnt2* and 18 Microsatellite Loci

#### Genotyping the Candidate *B4galnt2* Gene Locus

PCR and sequencing to assess the presence of *B4galnt2* haplotypes has been performed as described in Johnsen *et al.* (2009)(*i.e.* fragment "5", length ∼355bp). We used standard Sanger sequencing (Sanger *et al.*, 1992), using PCR amplified DNA as a template and ABI Big Dye sequencing chemistry (Applied Biosystems, Foster City, CA). The sequencing reactions were run on an ABI 3730 automated sequencer. The sequences were edited with

Seqman (included in DNASTAR, Inc., Madison, Wisc.) and aligned with the algorithm ClustalW (Thompson *et al.*, 1994), included in the program MEGA 4.0.2 (Tamura *et al.*, 2007).

Microsatellite Genotyping

See Material and Methods Chapter 2.

**Population Genetic Analyses**

Population Structure

See Material and Methods Chapter 2.
Additionally we calculated the proportion of shared alleles (Dps) using the program MSA (Dieringer & Schltterer 2003).

Population Genetic Analysis of the Candidate Locus *B4galnt2*

All summary statistics were calculated in DNASp 5.0 (Librado and Rozas, 2009), (a) number of segregating sites S; (b) Watterson's estimator $\theta_W$ (Watterson, 1975), based on the number of segregating sites in the sample; (c) $\pi$ (Tajima, 1983) the average number of pairwise differences in the sample; (d) Tajima's *D* statistics (Tajima, 1989) and (e) the divergence (Jukes and Cantor, 1969).

**Pyrosequencing of the 16S rRNA Gene and Sequence Processing**

See Material and Methods Chapter 2.
To test for specific OTUs that are correlated to the genotype of the *B4galnt2 cis*-regulatory region we used the the Primer-E package "SIMPER" (Clarke and Gorley, 2006) on 96% similarity level of the OTUs (corresponds to $\sim$ species level).

## Results

During three field seasons in the years 2009 and 2010, we collected eight house mouse populations throughout Germany and France (see Figure 2.1 and Table 3.1). To avoid biased results due to inbreeding we included in our analyses just one individual per farm, with a between farm distance of at least 1km. For the population genetic analyses on the mitochondrial DNA and the *B4galnt2* candidate locus we analyzed 121 and 122 individuals, respectively, analyses on the 18 autosomal microsatellite loci were performed on 121 individuals. For metagenomic purposes we analyzed 100 individuals (see Table 3.1).

### Population Structure

As described in Chapter 2 the analysis of the population substructure of the mouse populations has been carried out using two different genetic markers, the mitochondrial D-loop and 18 microsatellites. We sequenced $\sim$950 bp of the mitochondrial D-loop and constructed a Neighbor Joining tree (see Appendix C Figure C-S1), including reference sequences of *Mus musculus domesticus*, one sequence each of *M. spretus*, *M. spicilegus* and *M. famulus* as outgroups, with which we confirmed the taxonomic status of all sampled mice as being *M.m.domesticus*. To obtain a more detailed picture of the distribution of D-loop haploptes, we analyzed our dataset with respect to that of Bonhomme *et al.* (2011). This analysis displayed the presence of six haplogroups present in our data (see Figure 2.2).

We further analysed a panel of 18 microsatellite loci, to investigate the underlying population structure. In addition to population genetic parameters (see Table 2.2) we assessed information on populaion structure by applying the program STRUCTURE (Pritchard *et al.*, 2000; Falush *et al.*, 2003a) to our dataset. We compared all STRUCTURE results for the different number of clusters (K) and also compared the four independent runs of each K. After applying the criterion of Evanno *et al.* (2005) and after all comparisons among the different runs, we chose for K = 5 to be the number of clusters which best fits to our data (Figure 3.1a). Also with K $>$ 5 the general pattern of population structuring remains the same, the individuals were assigned consistently stable to the different clusters. STRUCTURE detected a homogenous ancestry for the populations MC, ES and SL. All other populations either showed mixed ancestry and/or admixed individuals.

Figure 3.1b shows an allele sharing tree, which is based on the proportion of shared alleles (Dps). This tree reflects the relationship between the populations and patterns of gene flow can be better detected than with mitochondrial D-loop and the SRUCTURE analysis. The short inner branches demonstrate the close relationship of all populations. Still,

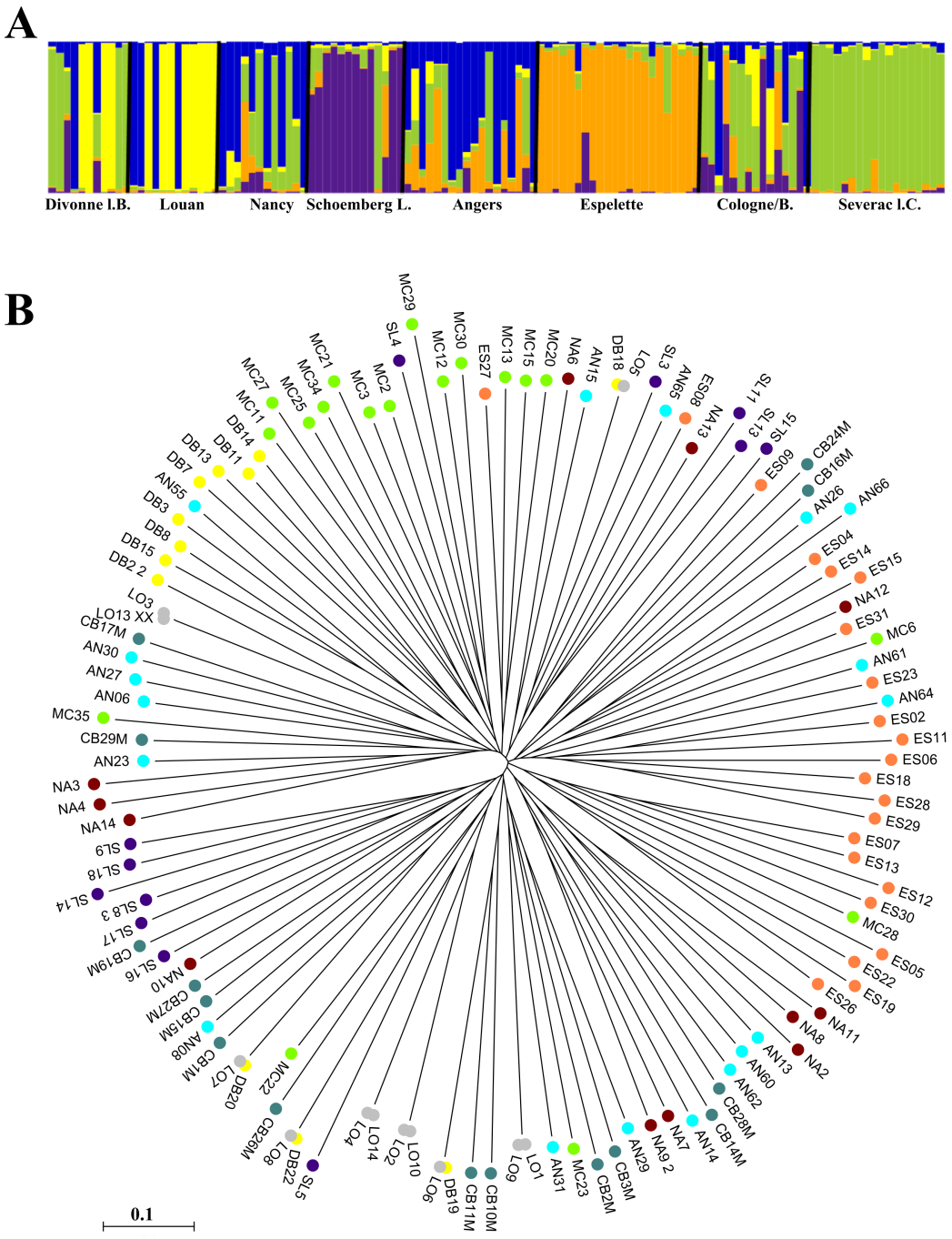


Figure 3.1: Analysis of host population structure using A) STRUCTURE (K=5) and B) the proportion of shared alleles (Dps)

this tree reflects mainly the same population structure and underlines the STRUCTURE outcome. Comparable to the pattern observed at the mitochondrial D-loop, also the microsatellites do not display a homogenous genetic background for all of the populations, even if some populations can be assigned to single haplogroups/clusters.

### Analysis of *B4galnt2* Haplotype Frequencies

To determine the frequency of alternative *B4galnt2* haplotypes in these newly collected samples, a diagnostic PCR fragment was sequenced (*i.e.* fragment #5, see (Johnsen *et al.*, 2009) and also Chapter 1). We screened all populations for the presence/absence of the RIII allele. The RIII allele was present in the populations AN, ES and MC in intermediate frequency (44%, 36% and 34%, respectively) and at very low frequencies in DB and SL (4% in each population) (Figure 3.2). In CB, NA and LO the RIII allele could not be

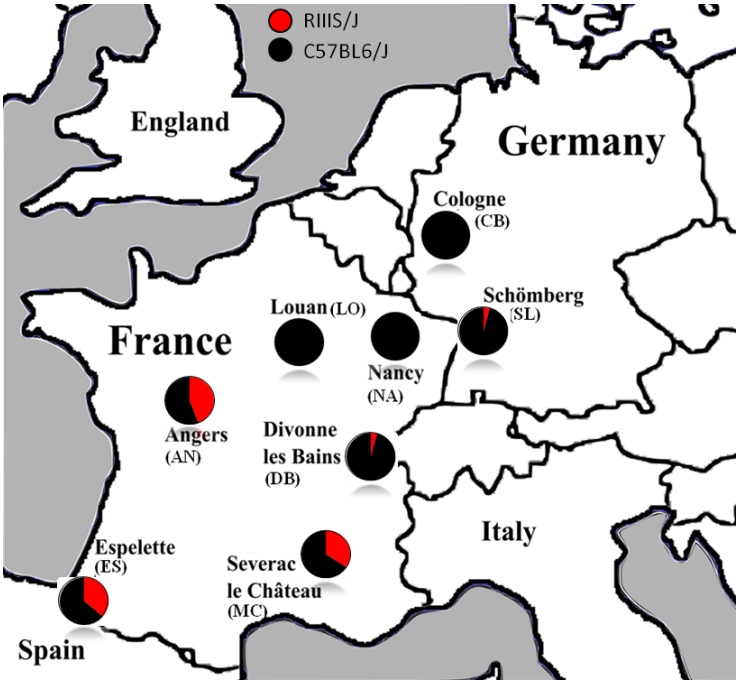


Figure 3.2: Distribution of C57BL6/J- (gut expression) and RIIIS/J- (blood vessel expression) haplotype classes in sampled populations

identified at all. The relationship between those two haplotype classes is visualized in a NeighbourNet network (see Figure 3.3). Obviously there are two main haplotypes with a few variants for each. The C57 haplotype is less variable than the RIII haplotype.We also calculated summary statistics for all populations (see Table 3.2). The most striking result were the extremely high values for Tajima's *D* of up to 3.6 in AN, in ES and MC

Table 3.2: Summary statistics for *B4galnt2* Upstream Region (Fragment #5).

| Population | Fragment | N | S | Hap | $\theta_W$ | $\pi$ | Tajima's *D* | | K |
|---|---|---|---|---|---|---|---|---|---|
| AN | 5 | 18 | 15 | 4 | 0.0263 | 0.0125 | 3.5813 | *** | 0.0356 |
| DB | 5 | 12 | 19 | 4 | 0.0058 | 0.0158 | -2.2818 | ** | 0.0524 |
| ES | 5 | 22 | 18 | 7 | 0.0224 | 0.0129 | 2.3768 | * | 0.0385 |
| LO | 5 | 12 | 1 | 2 | 0.0007 | 0.0006 | 0.1387 | | 0.0483 |
| MC | 5 | 19 | 15 | 5 | 0.0214 | 0.0111 | 2.9989 | ** | 0.0392 |
| NA | 5 | 12 | 0 | 1 | 0 | n.a. | n.d. | n.d. | 0.0479 |
| SL | 5 | 12 | 15 | 2 | 0.004 | 0.0128 | -2.434 | ** | 0.0527 |
| CB | 5 | 15 | 0 | 1 | 0 | n.a. | n.d. | n.d. | 0.0479 |

Tajima's *D* of 2.4 and 3, respectively. In those populations both allele classes are present at intermediate frequency. The high values for Tajima's *D* indicate balancing selection, for which evidence was given already for the population of MC described in (Johnsen *et al.*, 2009). The significant negative values for Tajima's *D* in SL and DB result from differences in the frequency of divergent haplotypes, namely one RIII allele per population. The allele frequencies do not deviate from Hardy Weinberg Equilibrium.

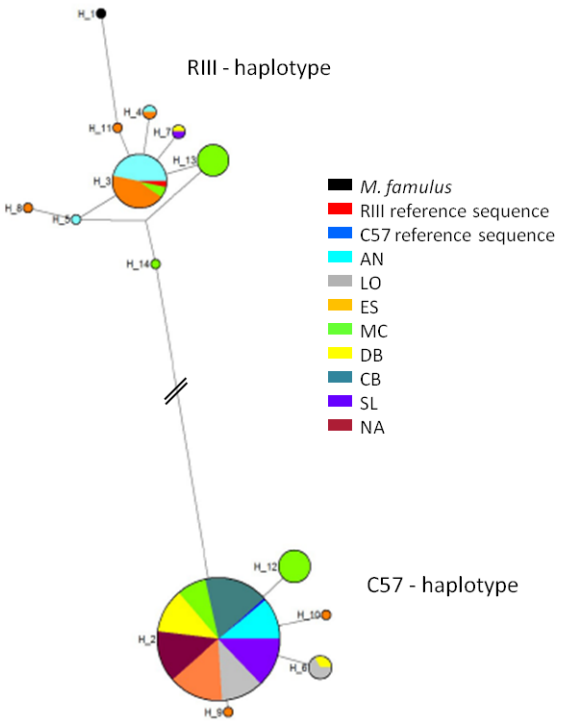


Figure 3.3: NeighbourNet network of fragement #5 in the *cis*-regulatory region of *B4galnt2*, representing both allele classes and their within-class variants

### *B4galnt2* Genotype Dependent OTUs

The analysis of *B4galnt2* dependent OTUs was restricted to two populations (MC and AN) in which homozygotes for the RIIIS/J and C57BL/6 as well as heterozygotes are present. Sufficient numbers of each genotype were not available for the other populations. With a corrected similarity level of 96% in total 15 OTUs could be identified to show specific correlations to the one or other haplotype (see Figure 3.4). OTUs

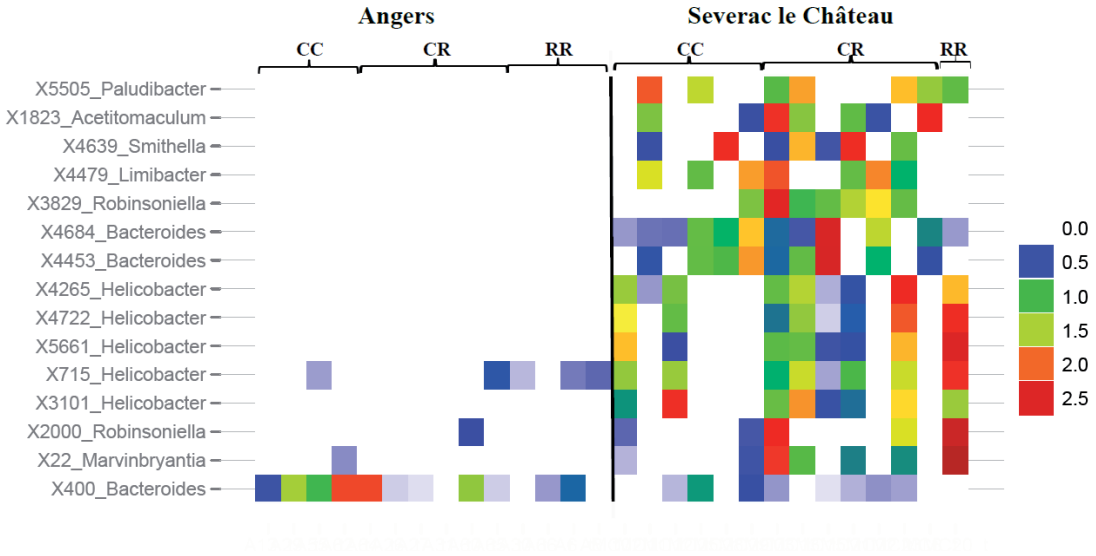


Figure 3.4: Bacterial taxa significantly correlated with *B4galnt2* genotype.

which show significant correlations with the RIIIS/J genotype were also identified with SIMPER method implemented in the Primer-E package (Clarke and Gorley, 2006). In summary, RIIIS/J homogzygotes have enriched *Helicobacter* OTUs (one OTU in Angers and four OTUs in the Severac le Château), while *Bacteroides* were enriched in heterozygotes. This general pattern also shows location-specificity, with different OTUs from *Bacteroides*/*Helicobacter* being involved in the two locations. Other OTUs belonging to *Paludibacter*, *Acetitomaculum*, *Smithella*, *Limibacter*, *Robinsoniella* and *Marvinbryantia* are majorly present in the Severac le Château and enriched in heterozygotes, while they are less frequent or missing in Angers.

## Discussion

To expand the existing knowledge about haplotype frequency and nucleotide polymorphism data (Johnsen *et al.*, 2009; Baines and Harr, 2007) we sampled mice from six localities throughout Germany and France. Additionally we resampled the two already extensively described populations from Cologne/Bonn and the Massif Central (Johnsen *et al.*, 2009; Baines and Harr, 2007; Ihle *et al.*, 2006) to primarily get tissue samples for the metagenomic analysis and to confirm former results on the presence/absence of RIII haplotypes in those two populations (Johnsen *et al.*, 2009). As Johnsen *et al.* (2009) detected the strong haplotype structure in the upstream region of *B4galnt2*, consisting of mainly two allele classes RIII and C57, and defined a diagnostic PCR fragment that gives reliable information about the expression change of *B4galnt2* from the epi- to the endothelium in the French population, we determined the fine scale distribution of *B4galnt2* haplotypes in all eight populations. A clear pattern of allele frequency distribution throughout western Europe, *i.e.* France and Germany, could be detected. This pattern raises two important evolutionary questions regarding the maintenance of these two alleles and the phenotypic consequences. Do allele frequency differences coincide with local selective pressures and/or aspects of population structure? What are the consequences of maintaining the disease associated allele and do we find influences of *B4galnt2* expression on the intestinal microbiota? To address these questions, detailed information on population structure is important to interpret the dynamics at *B4galnt2*. Thus, I will discuss population structure, as estimated via the mitochondrial D-loop and 18 neutral microsatellites, with regard to *B4galnt2* haplotype frequencies. Further I will concentrate on the influence of *B4galnt2* on the intestinal microbiota.

### Population Dynamics at *B4galnt2*

#### Population Structure

Population substructure could be detected using two different markers - the mitochondrial D-loop region and microsatellite loci. The analysis of the D-loop sequences revealed six different haplogroups, which we numbered according to the assignment of the larger number of haplogroups described by (Bonhomme *et al.*, 2011), who analyzed 1313 sequences (Figure 2.2). LO and NA share mainly one haplogroup (#11), as well as MC and SL (#1). The two populations of MC and CB show consistent results with Ihle *et al.* (2006), where MC showed very similar D-loop sequences (haplogroup #1) and CB is much more diverse, as is ES. A different pattern of population structure, was detected on the basis

of 18 microsatellites. Observed heterozygosities are in all populations lower than the expected values and are, for the populations MC and CB, consistent with the results of Ihle *et al.* (2006) and Hardouin *et al.* (2010). This reduced heterozygosity could result from inbreeding within nests, which is a common factor in the social system of mice. To avoid a bias of our results due to inbreeding, we included in our analysis just one individual per farm, with a distance between the farms of $> 1$km. The STRUCTURE analysis highlighted the differentiation of some populations (ES, SL and MC) in comparison to the other populations. The allele sharing tree revealed close relationship between all populations, still, consistent with STRUCTURE, some populations like ES, MC and SL are more genetically homogenous than others. Population structure is oftentimes due to man made habitat fragmentation (*e.g.* Yamamoto *et al.* (2004); Lee *et al.* (2012)). Habitat fragmentation due to humans (urban landscapes) is not expected to influence house mice because of their commensal biology. The observed population structure could have resulted from different colonization waves and/or simple isolation between populations.

Population Structure vs. Local Selective Pressure(s)

Regarding the candidate gene *B4galnt2* we screened all populations for the presence or absence of the two allele classes (RIII and C57) in its upstream region. We could confirm the results of (Johnsen *et al.*, 2009) for MC and CB. The presence of the RIII allele at intermediate frequency in the populations MC, ES and AN and low frequency or even absence of that allele in all other populations (DB, SL and NA, CB, LO) reveals a clear decline of the allele frequency in central France. This shows a clear geographic pattern with respect to the distribution of alternative disease (RIIIS/J) and wild type (C57BL6/J) haplotype classes (Figure 3.2). Evolutionary processes can be influenced either by population dynamics, as well as species range/ geographical distribution of populations and/or selection (de Meaux and Mitchell-Olds, 2003). To shed light on the question which forces lead to the pattern of haplotype distribution at *B4galnt2* we will discuss population structure, geographic range and selection as possible explanations.

In Figure 3.5 we demonstrate the results of population structure estimated by D-loop sequences and microsatellites according to the haplotype distribution. To visualize the STRUCTURE result according to *B4galnt2* allele frequencies we assigned the presence of a genetic cluster to a population (geographic location, *e.g.* MC) if the average proportion of ancestry for this cluster in a population was $> 15\%$. In both cases, the pattern of substructure based on the D-loop and microstallites did not match the allele frequency distribution pattern. Populations with the same genetic background (*e.g.* MC and SL for

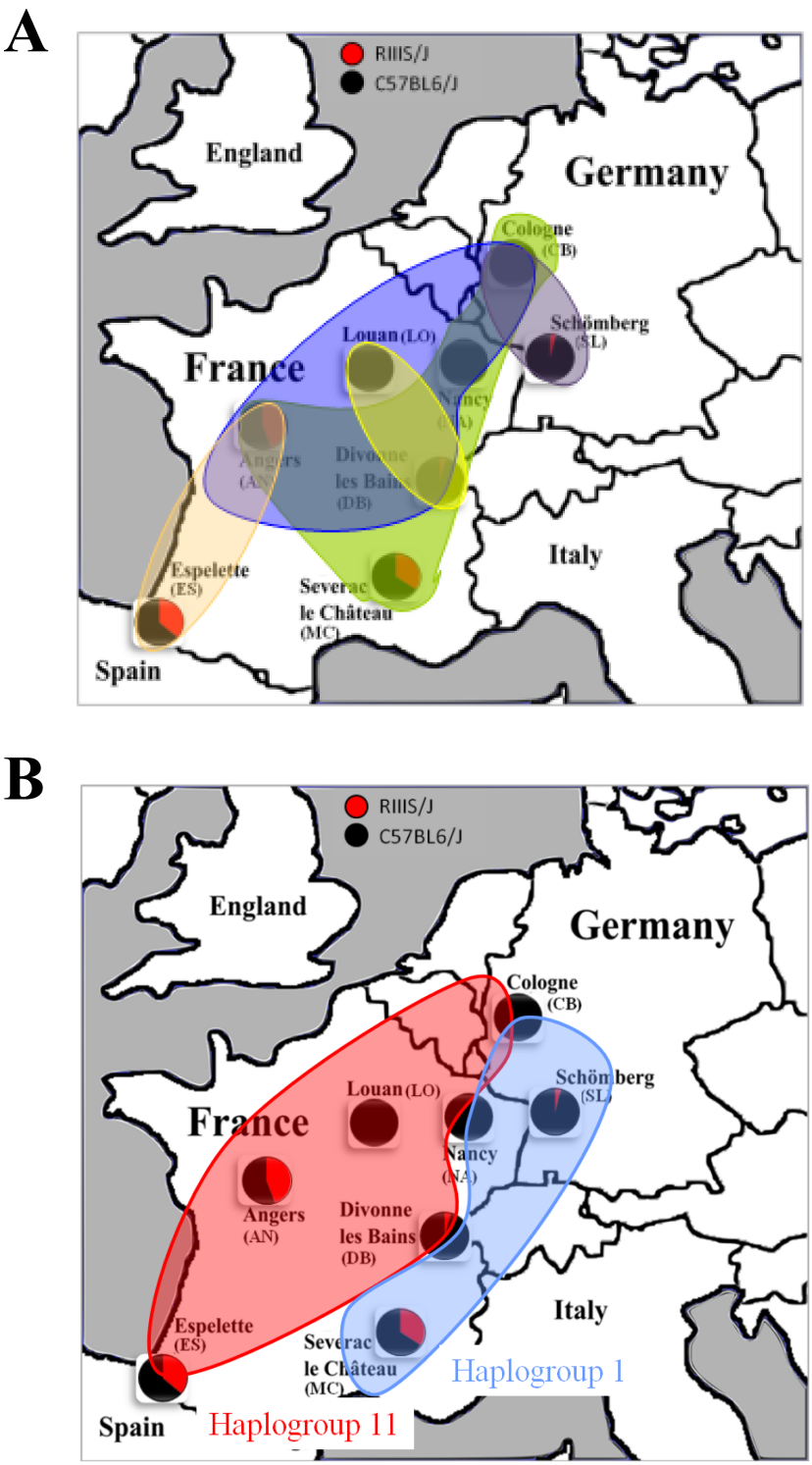


Figure 3.5: A) Pattern of population structure/clusters identified by STRUCTURE according to the *B4galnt2*- allele frequency distribution and B) and the distribution of haplogroups (estimated with the mitochondrial D-loop sequences; only the two main haplogroups 1 and 11 shown in this figure) ; Piecharts represent the frequency of the RIII and C57 allele, respectively.

mt D-loop or AN and NA as estimated by microsatellites) have very different frequencies of *B4galnt2* alleles. Likewise, other populations with very *similar B4galnt2* allele frequencies *differ* in their genetic setup (ES and MC differ in their homogeneity of D-loop and microsatellites). This makes population structure a very unlikely explanation for this outstanding pattern of haplotype distribution at the *B4galnt2* gene and supports the hypothesis of local selective pressure.

Together with indications for balancing selection acting at *B4galnt2* in the population MC (Johnsen *et al.*, 2009; Linnenbrink *et al.*, 2011) (see Chapter 1) and population genetic analyses of that locus of all populations used in this Chapter, it seems there are different selective pressures present throughout western Europe. None of the populations show deviations from Hardy-Weinberg Equilibrium, which gives evidence that either heterozygote advantage is not the driving force for the maintenance of both allele classes or that the selective force(s) are present but transient (*i.e.* deviations from Hardy-Weinberg Equilibrium can disappear after one generation). The selective force(s) seem more likely to be either temporal/spatial changes or be frequency dependent. As this could possibly be the presence or absence of pathogens we included the *B4galnt2* gene as candidate gene possibly influencing the gut microbiota (see next section). As proposed by Johnsen *et al.* (2009), the ”trench warfare” hypothesis (Stahl *et al.*, 1999) or also ”recycled polymorphism” (Holub, 2001) is one potential explanation for the ongoing population genetic processes at *B4galnt2*. The maintenance of two allele classes at one single locus in the host could be explained by a change in the frequency of occurrence of a certain pathogen. Setting the populations CB and MC from Johnsen *et al.* (2009) in the bigger picture of additional six populations throughout Germany and France, new conclusions can be drawn. The survey of *B4galnt2* haplotype frequencies seems to be a combination of local adaptation to some geographical dependent factors (*e.g.* the presence/absence of a pathogen) which constitutes the inter-population allele frequency differences and balancing selection acting. Likewise, balancing selection could act within those local populations with intermediated RIII alleles (*e.g.* fluctuating allele frequencies according to the frequency of pathogen occurrence), or, seeing all populations together as one (*i.e.* metapopulation), both alleles are maintained, the certain pattern of allele frequency distribution, still, may just be because of the spatial distribution of the sampling locations.

The trench warfare hyptothesis (Stahl *et al.*, 1999) would still represent one possible and likely explanation for this intriguing pattern of allele frequency distribution. The varying allele frequencies over western Europe may be a snapshot, where MC, ES and AN represent a status with high RIII allele frequencies, CB, NA and LO miss RIII alleles and DB and SL would represent an intermediate step. Also local adaptation remains as possible

explanation for the differing allele frequencies. The presence or absence of a pathogen could lead to the maintenance of both allele classes in some populations but not in others. As both explanations are not mutually exclusive, more sampling locations and sampling over a long period of time could help to disentangle both possiblities. Without knowing the beneficial consequences of exhibiting the RIII allele at *B4galnt2* this will remain an open question.

### Consequences of *B4galnt2* Expression Change on the Gut Microbiota Communities

Several loci in the mouse genome have been already mapped for controlling certain groups of bacteria (Benson *et al.*, 2010). The strong contribution of host genetics to the set up of gut bacteria has already been found in humans (Zoetendal *et al.*, 2001) and mice (Rehman *et al.*, 2011; Rausch *et al.*, 2011). As listed in the review of Spor *et al.* (2011), many examples for single genes influencing the microbiota exist.

*B4galnt2* is a glycosyltransferase whose GI expression of GalNAc residues might serve as binding sites for pathogens or as nutrient source to feed on (Hoskins and Boulding, 1976; Sonnenburg *et al.*, 2005) and could result in differences in the susceptibility to pathogens in the gastrointestinal tract (Robinson *et al.*, 1971).

In this study we investigate the influence of the gene *B4galnt2* on the intestinal microbiota in wild-caught animals from natural populations. Evidence of such interactions was provided by Staubach *et al.* (2012) in lab mice, where wild type and *B4galnt2*-knockout mice were profiled.

Our data revealed individual bacterial OTUs belonging to *Helicobacter*, *Bacteroides* and others, whose abundances were significantly correlated to *B4galnt2* genotype. (Figure 3.4). Unfortunately our analysis was restricted to two populations, namely AN and MC, which were the only populations with sufficient numbers of all three genotypes to be analyzed. Within the two populations AN and MC differences in habitat preference (*i.e.* genotype) can be seen, but an overall pattern is not visible. As shown in Chapter 2, geography is the strongest factor, determining the bacterial communities in the gut, together with host population structure and the interaction of both. Thus, it might be, that the general pattern of *B4galnt2* genotype influence on the intestinal microbiota is hidden by stronger determinants, as we analysed natural populations and not animals raised under controlled laboratory conditions.

Still, similar to the observations of Staubach *et al.* (2012), different strains of *Helicobacter* are significantly correlated to *B4galnt2* genotype (Figure 3.4), also in wild- caught mice. This suggests that strains belonging to the same genus differ in their biological niche, as defined by *B4galnt2* genotype and thus the presence or absence of GalNAc

sugars. *Helicobacter* might be, amongst others (*e.g. Bacteroides*), a good candidate for pathogen-driven selection. *Helicobacter* spp. are well known pathogenic bacteria which naturally infect rodents (Fox *et al.*, 1995; Lee *et al.*, 1992; Simmons *et al.*, 2000) including mice (Feng *et al.*, 2005), mostly cause infection in the intestine. As reported by Aspholm-Hurtig *et al.* (2004), binding of *Helicobacter pylori* strains is determined by the occurrence of sugar residues at the blood group antigens H1 and Lewis b in the gastric mucosa in humans. To determine the phenotypic consequences in of *B4galnt2* on the gut microbiota in more detail many more individuals are needed to get a reliable answer. Sampling more individuals from one location, even from one single farm, where the RIII haplotype is known to occur, would make the investigation of the gut microbiota and *B4galnt2* habitat preference easier as geographical and other environmental effects (*e.g.* diet, climatic changes) and population structure as influencing factors can be ruled out.

## Conclusions and Future Perspective

This chapter elucidated important evolutionary questions regarding the blood group related gene *B4galnt2*. The *B4galnt2* gene displays recent as well as long term signatures of selection in several *M.musculus* subspecies and species. Great differences in the RIII allele frequency distribution throughout France and Germany are likely due to local selective pressure(s) (possibly because of pathogen occurrence). If the differentiation between the populations is due to local selective pressures we would locate the divide somewhere in central France. As proposed in Chapter 1 selection might act on glycosylation in the gut and thus, influence possible attachment sites or nutrient sources for pathogens. This might be a contributing factor to the long-term maintenance of both allele classes and thus expression variation.

As *B4galnt2* is a blood group related gene and thus represents a good candidate for host-pathogen interaction, host gut microbiota has been profiled according to *B4galnt2* genotype. Thus, for full understanding of the evolutionary forces governing *B4galnt2* more localities have to be sampled in central France, along the ”gradient” of RIII allele frequency. To ascertain the phenotypic consequences on the gut microbiota of the different *B4glant2* expression profiles much more work remains to be done. This comprises the identification of good bacterial candidates for pathogen driven selection, followed by isolating and culturing of bacterial strains in the lab. To confirm the relationship between host *B4galnt2* genotype, candidate bacteria and the phenotypic consequence infection experiments have to be performed.

For several reasons, the intestine seems very likely to be the site of selection acting, most

likely due to host-pathogen interactions, determined by the presence or absence of certain sugar residues in the gut. But also the endothelium may not be forgotten as possible target of selection as the *cis*-regulatory mutation that leads to *B4galnt2* expression in the blood vessel may also confer resistance in this tissue. Several examples of pathogens interacting with blood platelets exist (reviewed by Fitzgerald *et al.* (2006)), also some of them have been associated with attachment to the VWF (*e.g. Helicobacter pylori* (Byrne *et al.*, 2003). VWF is also known to be the attachment site for *Staphylococcus aureus* during endothelial infection (Herrmann *et al.*, 1997) by the binding adhesin Protein A (Hartleib *et al.*, 2000). Thus altering the glycosylation status of VWF might either change the intensity or even inhibit bacterial infection, *e.g.* of *S. aureus*.

## Supplementary Information

See Appendix C for Supplementary Figure C-S1.
See Appendix D for Primers used in this Chapter.

# Bibliography

Allison, A. (1954). Protection Afforded by Sickle Cell Trait against subtertian malarial infection. *British Medical Journal*, pages 290–294.

Andrés, A. M., Hubisz, M. J., Indap, A., Torgerson, D. G., Degenhardt, J. D., Boyko, A. R., Gutenkunst, R. N., White, T. J., Green, E. D., Bustamante, C. D., Clark, A. G., and Nielsen, R. (2009). Targets of balancing selection in the human genome. *Molecular Biology and Evolution*, 26(12):2755–64.

Apanius, V., Penn, D., Slev, P. R., Ruff, L. R., and Potts, W. K. (1997). The nature of selection on the major histocompatibility complex. *Critical Reviews in Immunology*, 17(2):179–224.

Aspholm-Hurtig, M., Dailide, G., Lahmann, M., Kalia, A., Ilver, D., Roche, N., Vikström, S., Sjöström, R., Lindén, S., Bäckström, A., Lundberg, C., Arnqvist, A., Mahdavi, J., Nilsson, U. J., Velapatiño, B., Gilman, R. H., Gerhard, M., Alarcon, T., López-Brea, M., Nakazawa, T., Fox, J. G., Correa, P., Dominguez-Bello, M. G., Perez-Perez, G. I., Blaser, M. J., Normark, S., Carlstedt, I., Oscarson, S., Teneberg, S., Berg, D. E., and Borén, T. (2004). Functional adaptation of BabA, the *H. pylori ABO* blood group antigen binding adhesin. *Science*, 305(5683):519–22.

Asthana, S., Schmidt, S., and Sunyaev, S. (2005). A limited role for balancing selection. *Trends in Genetics*, 21(1):30–32.

Auffray, J., Fontanillas, P., Catalan, J., and Britton-Davidian, J. (2001). Developmental stability in house mice heterozygous for single Robertsonian fusions. *Journal of Heredity*, 92(1):23.

Bäckhed, F., Ding, H., Wang, T., Hooper, L. V., Koh, G. Y., Nagy, A., Semenkovich, C. F., and Gordon, J. I. (2004). The gut microbiota as an environmental factor that regulates fat storage. *Proceedings of the National Academy of Sciences of the United States of America*, 101(44):15718–23.

Bäckhed, F., Ley, R. E., Sonnenburg, J. L., Peterson, D. a., and Gordon, J. I. (2005). Host-bacterial mutualism in the human intestine. *Science*, 307(5717):1915–20.

Baines, J. F. and Harr, B. (2007). Reduced X-linked diversity in derived populations of house mice. *Genetics*, 175(4):1911–1921.

Bamshad, M. J., Mummidi, S., Gonzalez, E., Ahuja, S. S., Dunn, D. M., Watkins, W. S., Wooding, S., Stone, A. C., Jorde, L. B., Weiss, R. B., and Ahuja, S. K. (2002). A strong signature of balancing selection in the 5' *cis*-regulatory region of *CCR5*. *Proceedings of the National Academy of Sciences of the United States of America*, 99(16):10539–44.

Benson, A. K., Kelly, S. A., Legge, R., Ma, F., Low, S. J., Kim, J., Zhang, M., Oh, P. L., Nehrenberg, D., Hua, K., Kachman, S. D., Moriyama, E. N., Walter, J., Peterson, D. A., and Pomp, D. (2010). Individuality in gut microbiota composition is a complex polygenic trait shaped by multiple environmental and host genetic factors. *Proceedings of the National Academy of Sciences of the United States of America*, 107(44):18933–18938.

Bishop, J. R. and Gagneux, P. (2007). Evolution of carbohydrate antigens–microbial forces shaping host glycomes? *Glycobiology*, 17(5):23R–34R.

Bjornstad, O. (2012). Package "ncf".

Blomberg, S. P., Garland, T., and Ives, A. R. (2003). Testing for phylogenetic signal in comparative data: behavioral traits are more labile. *Evolution*, 57(4):717–45.

Bonhomme, F., Orth, A., Cucchi, T., Rajabi-Maham, H., Catalan, J., Boursot, P., Auffray, J.-C., and Britton-Davidian, J. (2011). Genetic differentiation of the house mouse around the Mediterranean basin: matrilineal footprints of early and late colonization. *Proceedings of the National Academy of Sciences*, 278(1708):1034–43.

Boursot, P., Auffray, J. C., Britton-Davidian, J., and Bonhomme, F. (1993). The Evolution of House Mice. *Annual Review of Ecology and Systematics*, 24(1):119–152.

Boursot, P., Din, W., Anand, R., Darviche, D., Dod, B., Von Deimling, F., Talwar, G., and Bonhomme, F. (1996). Origin and radiation of the house mouse: mitochondrial DNA phylogeny. *Journal of Evolutionary Biology*, 9(4):391–415.

Brooks, M., Leith, G. S., Allen, A. K., Woods, P. R., Benson, R. E., and Dodds, W. J. (1991). Bleeding disorder (von Willebrand disease) in a quarter horse. *Journal of the American Veterinary Medical Association*, 198(1):114–6.

Bubb, K. L., Bovee, D., Buckley, D., Haugen, E., Kibukawa, M., Paddock, M., Palmieri, A., Subramanian, S., Zhou, Y., Kaul, R., Green, P., and Olson, M. V. (2006). Scan of human genome reveals no new Loci under ancient balancing selection. *Genetics*, 173(4):2165–77.

Bustamante, C. D., Fledel-Alon, A., Williamson, S., Nielsen, R., Hubisz, M. T., Glanowski, S., Tanenbaum, D. M., White, T. J., Sninsky, J. J., Hernandez, R. D., Civello, D., Adams, M. D., Cargill, M., and Clark, A. G. (2005). Natural selection on protein-coding genes in the human genome. *Nature*, 437(7062):1153–7.

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

Calafell, F., Roubinet, F., Ramírez-Soriano, A., Saitou, N., Bertranpetit, J., and Blancher, A. (2008). Evolutionary dynamics of the human ABO gene. *Human Genetics*, 124(2):123–35.

Carroll, I. M., Threadgill, D. W., and Threadgill, D. S. (2009). The gastrointestinal microbiome: a malleable, third genome of mammals. *Mammalian Genome*, 20(7):395–403.

Carroll, S. B. (2005). Evolution at two levels: on genes and form. *PLoS Biology*, 3(7):e245.

Cavender-Bares, J., Kozak, K. H., Fine, P. V. a., and Kembel, S. W. (2009). The merging of community ecology and phylogenetic biology. *Ecology Letters*, 12(7):693–715.

Chamary, J. V., Parmley, J. L., and Hurst, L. D. (2006). Hearing silence: non-neutral evolution at synonymous sites in mammals. *Nature Reviews. Genetics*, 7(2):98–108.

Charlesworth, D. (2006). Balancing selection and its effects on sequences in nearby genome regions. *PLoS Genetics*, 2(4):e64.

Cho, S., Huang, Z. Y., Green, D. R., Smith, D. R., and Zhang, J. (2006). Evolution of the complementary sex-determination gene of honey bees: balancing selection and trans-species polymorphisms. *Genome Research*, 16(11):1366–75.

Clarke, K. (1993). Nonparametric multivariate analyses of changes in community structure. *Australian Journal of Ecology*, (1988):117–143.

Clarke, K. and Gorley, R. (2006). PRIMER v6: User Manual/Tutorial. PRIMER-E.

Cook, L. M. (2003). The rise and fall of the Carbonaria form of the peppered moth. *The Quarterly Review of Biology*, 78(4):399–417.

Costello, E. K., Lauber, C. L., Hamady, M., Fierer, N., Gordon, J. I., and Knight, R. (2009). Bacterial community variation in human body habitats across space and time. *Science*, 326(5960):1694–7.

Cowles, C. R., Hirschhorn, J. N., Altshuler, D., and Lander, E. S. (2002). Detection of regulatory variation in mouse genes. *Nature Genetics*, 32(3):432–7.

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

Darwin, C. (1859). *On the origin of species by means of natural selection.* John Murray, London.

De Filippo, C., Cavalieri, D., Di Paola, M., Ramazzotti, M., Poullet, J. B., Massart, S., Collini, S., Pieraccini, G., and Lionetti, P. (2010). Impact of diet in shaping gut microbiota revealed by a comparative study in children from Europe and rural Africa. *Proceedings of the National Academy of Sciences of the United States of America*, 107(33):14691–6.

de Meaux, J. and Mitchell-Olds, T. (2003). Evolution of plant resistance at the molecular level: ecological context of species interactions. *Heredity*, 91(4):345–52.

Dieringer, D. and Schlötterer, C. (2003). Microsatellite analyser (MSA): a platform independent analysis tool for large microsatellite data sets. *Molecular Ecology Notes*, 3(1):167–169.

Dominguez-Bello, M. G., Costello, E. K., Contreras, M., Magris, M., Hidalgo, G., Fierer, N., and Knight, R. (2010). Delivery mode shapes the acquisition and structure of the initial microbiota across multiple body habitats in newborns. *Proceedings of the National Academy of Sciences of the United States of America*, 107(26):11971–5.

Eckburg, P. B., Bik, E. M., Bernstein, C. N., Purdom, E., Dethlefsen, L., Sargent, M., Gill, S. R., Nelson, K. E., and Relman, D. A. (2005). Diversity of the human intestinal microbial flora. *Science*, 308(5728):1635–8.

Edgar, R. C. (2010). Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*, 26(19):2460–1.

Edgar, R. C. (2011). Usearch.

Emerson, B. C. and Gillespie, R. G. (2008). Phylogenetic analysis of community assembly and structure over space and time. *Trends in Ecology and Evolution*, 23(11):619–30.

Evanno, G., Regnaut, S., and Goudet, J. (2005). Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Molecular Ecology*, 14(8):2611–20.

Excoffier, L., Estoup, A., and Cornuet, J.-M. (2005). Bayesian analysis of an admixture model with mutations and arbitrarily linked markers. *Genetics*, 169(3):1727–38.

Faith, D. (1992). Conservation evaluation and phylogenetic diversity. *Biological Conservation*, 61(1):1–10.

Falk, P. G., Hooper, L. V., Midtvedt, T., and Gordon, J. I. (1998). Creating and maintaining the gastrointestinal ecosystem: what we know and need to know from gnotobiology. *Microbiology and Molecular Biology Reviews*, 62(4):1157–70.

Falush, D., Stephens, M., and Pritchard, J. K. (2003a). Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics*, 164(4):1567–1587.

Falush, D., Wirth, T., Linz, B., Pritchard, J. K., Stephens, M., Kidd, M., Blaser, M. J., Graham, D. Y., Vacher, S., Perez-Perez, G. I., Yamaoka, Y., Mégraud, F., Otto, K., Reichard, U., Katzowitsch, E., Wang, X., Achtman, M., and Suerbaum, S. (2003b). Traces of human migrations in Helicobacter pylori populations. *Science*, 299(5612):1582–5.

Feng, S., Ku, K., Hodzic, E., Lorenzana, E., Freet, K., and Barthold, S. W. (2005). Differential Detection of Five Mouse-Infecting *Helicobacter* Species by Multiplex PCR. *Clinical and Vaccine Immunology*.

Ferguson, W., Dvora, S., Gallo, J., Orth, A., and Boissinot, S. (2008). Long-term balancing selection at the West Nile virus resistance gene, *Oas1b*, maintains transspecific polymorphisms in the house mouse. *Molecular Biology and Evolution*, 25(8):1609–1618.

Filatov, D. A. and Charlesworth, D. (1999). DNA polymorphism, haplotype structure and balancing selection in the Leavenworthia *PgiC* locus. *Genetics*, 153(3):1423–34.

Fitzgerald, J. R., Foster, T. J., and Cox, D. (2006). The interaction of bacterial pathogens with platelets. *Nature Reviews. Microbiology*, 4(6):445–57.

Fitzpatrick, M. J., Feder, E., Rowe, L., and Sokolowski, M. B. (2007). Maintaining a behaviour polymorphism by frequency-dependent selection on a single gene. *Nature Letters*, 447:210–212.

Fox, J. G., Yan, L. L., Dewhirst, F. E., Paster, B. J., Shames, B., Murphy, J. C., Hayward, A., Belcher, J. C., and Mendes, E. N. (1995). Helicobacter bilis sp. nov., a novel Helicobacter species isolated from bile, livers, and intestines of aged, inbred mice. *Journal of Clinical Microbiology*, 33(2):445–54.

Fry, A. E., Griffiths, M. J., Auburn, S., Diakite, M., Forton, J. T., Green, A., Richardson, A., Wilson, J., Jallow, M., Sisay-Joof, F., Pinder, M., Peshu, N., Williams, T. N., Marsh, K., Molyneux, M. E., Taylor, T. E., Rockett, K. A., and Kwiatkowski, D. P. (2008). Common variation in the *ABO* glycosyltransferase is associated with susceptibility to severe *Plasmodium falciparum* malaria. *Human Molecular Genetics*, 17(4):567–76.

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

Fumagalli, M., Cagliani, R., Riva, S., Pozzoli, U., Biasin, M., Piacentini, L., Comi, G. P., Bresolin, N., Clerici, M., and Sironi, M. (2010). Population genetics of *IFIH1*: ancient population structure, local selection, and implications for susceptibility to type 1 diabetes. *Molecular Biology and Evolution*, 27(11):2555–66.

Galtier, N., Bonhomme, F., Moulia, C., Belkhir, K., Caminade, P., Desmarais, E., Duquesne, J. J., Orth, A., Dod, B., and Boursot, P. (2004). Mouse biodiversity in the genomic era. *Cytogenetic and Genome Research*, 105(2-4):385–94.

Gillespie, J. H. (2004). *Population Genetics - A Concise Guide*. Johns Hopkins, Baltimore and London, 2nd edition.

Grice, E. A., Kong, H. H., Conlan, S., Deming, C. B., Davis, J., Young, A. C., Bouffard, G. G., Blakesley, R. W., Murray, P. R., Green, E. D., Turner, M. L., and Segre, J. A. (2009). Topographical and temporal diversity of the human skin microbiome. *Science*, 324(5931):1190–2.

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

Hahn, M. W. (2007). Detecting natural selection on cis-regulatory DNA. *Genetica*, 129(1):7–18.

Handelsman, J. (2005). Sorting out metagenomes. *Nature Biotechnology*, 23(1):38–9.

Handelsman, J., Rondon, M. R., Brady, S. F., Clardy, J., and Goodman, R. M. (1998). Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. *Chemistry and Biology*, 5(10):R245–9.

Hardouin, E. A., Chapuis, J.-L., Stevens, M. I., van Vuuren, J. B., Quillfeldt, P., Scavetta, R. J., Teschke, M., and Tautz, D. (2010). House mouse colonization patterns on the sub-Antarctic Kerguelen Archipelago suggest singular primary invasions and resilience against re-invasion. *BMC Evolutionary Biology*, 10(1):325.

Hartl, D. L. and Clark, A. G. (2007). *Principles of Population Genetics*. Sinauer Associates, Inc. Publishers, Sunderland, Massachusetts, 4th edition.

Hartleib, J., Köhler, N., Dickinson, R. B., Chhatwal, G. S., Sixma, J. J., Hartford, O. M., Foster, T. J., Peters, G., Kehrel, B. E., and Herrmann, M. (2000). Protein A is the von Willebrand factor binding protein on *Staphylococcus aureus*. *Blood*, 96:2149–2156.

Hedrick, P. W. (1976). Genetic variation in a heterogeneous environment. II. Temporal heterogeneity and directional selection. *Genetics*, 84(1):145–157.

Hehemann, J.-H., Correc, G., Barbeyron, T., Helbert, W., Czjzek, M., and Michel, G. (2010). Transfer of carbohydrate-active enzymes from marine bacteria to Japanese gut microbiota. *Nature*, 464(7290):908–12.

Heijtz, R. D., Wang, S., Anuar, F., Qian, Y., Björkholm, B., Samuelsson, A., Hibberd, M. L., Forssberg, H., and Pettersson, S. (2011). Normal gut microbiota modulates brain development and behavior. *Proceedings of the National Academy of Sciences of the United States of America*, 108(7):3047–52.

Herrmann, M., Hartleib, J., Kehrel, B., Montgomery, R. R., Sixma, J. J., and Peters, G. (1997). Interaction of von Willebrand factor with *Staphylococcus aureus*. *The Journal of Infectious Diseases*, 176(4):984–991.

Hiwatashi, T., Okabe, Y., Tsutsui, T., Hiramatsu, C., Melin, A. D., Oota, H., Schaffner, C. M., Aureli, F., Fedigan, L. M., Innan, H., and Kawamura, S. (2010). An explicit signature of balancing selection for color-vision variation in new world monkeys. *Molecular Biology and Evolution*, 27(2):453–64.

Hoekstra, H. E. and Coyne, J. A. (2007). The locus of evolution: evo devo and the genetics of adaptation. *Evolution*, 61(5):995–1016.

Holub, E. B. (2001). The arms race is ancient history in Arabidopsis, the wildflower. *Genetics*, 2(7):516–527.

Hooper, L. V. (2001). Commensal Host-Bacterial Relationships in the Gut. *Science*, 292(5519):1115–1118.

Hooper, L. V., Midtvedt, T., and Gordon, J. I. (2002). How host-microbial interactions shape the nutrient environment of the mammalian intestine. *Annual Review of Nutrition*, 22:283–307.

Hoskins, L. C. and Boulding, E. T. (1976). Degradation of blood group antigens in human colon ecosystems. II. A gene interaction in man that affects the fecal population density of certain enteric bacteria. *The Journal of Clinical Investigation*, 57(1):74–82.

Hugenholtz, P. and Tyson, G. W. (2008). Microbiology: Metagenomics. *Nature*, 455(7212):481–3.

Huson, D. H. (1998). SplitsTree: analyzing and visualizing evolutionary data. *Bioinformatics*, 14(1):68–73.

Huson, D. H. and Bryant, D. (2006). Application of phylogenetic networks in evolutionary studies. *Molecular Biology and Evolution*, 23(2):254–67.

Huxley, S. J. S. (1942). *Evolution: The Modern Synthesis.* Allan and Unwin, London.

Ihle, S., Ravaoarimanana, I., Thomas, M., and Tautz, D. (2006). An analysis of signatures of selective sweeps in natural populations of the house mouse. *Molecular Biology and Evolution*, 23(4):790–797.

Jaffe, E. A., Hoyer, L. W., and Nachman, R. L. (1973). Synthesis of antihemophilic factor antigen by cultured human endothelial cells. *The Journal of Clinical Investigation*, 52(11):2757–64.

Johnsen, J. M., Levy, G. G., Westrick, R. J., Tucker, P. K., and Ginsburg, D. (2008). The endothelial-specific regulatory mutation, Mvwf1, is a common mouse founder allele. *Mammalian Genome*, 19(1):32–40.

Johnsen, J. M., Teschke, M., Pavlidis, P., McGee, B. B. M., Tautz, D., Baines, J. F., and Ginsburg, D. (2009). Selection on *cis*-regulatory variation at *B4galnt2* and its influence on von Willebrand factor in house mice. *Molecular Biology and Evolution*, 26(3):567–78.

Jukes, T. H. and Cantor, C. R. (1969). *Evolution of protein molecules*. Academic Press, New York.

Kamau, E. and Charlesworth, D. (2005). Balancing selection and low recombination affect diversity near the self-incompatibility loci of the plant Arabidopsis lyrata. *Current biology : CB*, 15(19):1773–8.

Kettlewell, H. (1973). *The Evolution of Melanism.* Clarendon Press, Oxford.

Kile, B. T. and Hilton, D. J. (2005). The art and design of genetic screens: mouse. *Nature Reviews. Genetics*, 6(7):557–67.

Kimura, M. (1983). *The Neutral Theory of Molecular Evolution*. Cambridge University Press, Cambridge.

Kleinjan, D. A. and van Heyningen, V. (2005). Long-range control of gene expression: emerging mechanisms and disruption in disease. *American Journal of Human Genetics*, 76(1):8–32.

Koch, R. (1876). Die Ätiologie der Milzbrand-Krankheit , begründet auf die Entwicklungsgeschichte des Bacillus Anthracis. *Cohns Beiträge zur Biologie der Pflanzen*, 2(2):5–27.

Koenig, J. E., Spor, A., Scalfone, N., Fricker, A. D., Stombaugh, J., Knight, R., Angenent, L. T., and Ley, R. E. (2011). Succession of microbial consortia in the developing infant gut microbiome. *Proceedings of the National Academy of Sciences of the United States of America*, 108 Suppl:4578–85.

Kojima, K. (1971). Is There a Constant Fitness Value for a Given Genotype? NO! *Evolution*, 25(2):281–285.

Lee, A., Phillips, M. W., O'Rourke, J. L., Paster, B. J., Dewhirst, F. E., Fraser, G. J., Fox, J. G., Sly, L. I., Romaniuk, P. J., and Trust, T. J. (1992). *Helicobacter muridarum* sp. nov., a microaerophilic helical bacterium with a novel ultrastructure isolated from the intestinal mucosa of rodents. *International Journal of Systematic Bacteriology*, 42(1):27–36.

Lee, J. S., Ruell, E. W., Boydston, E. E., Lyren, L. M., Alonso, R. S., Troyer, J. L., Crooks, K. R., and Vandewoude, S. (2012). Gene flow and pathogen transmission among bobcats (*Lynx rufus*) in a fragmented urban landscape. *Molecular Ecology*.

Lewontin, R. and Hubby, J. (1966). A molecular approach to the study of genic heterozygosity in natural populations. II. Amount of variation and degree of heterozygosity in natural populations of Drosophila pseudoobscura. *Genetics*, 54(2):595–609.

Ley, R. E., Bäckhed, F., Turnbaugh, P., Lozupone, C. A., Knight, R. D., and Gordon, J. I. (2005). Obesity alters gut microbial ecology. *Proceedings of the National Academy of Sciences of the United States of America*, 102(31):11070–5.

Ley, R. E., Lozupone, C. A., Hamady, M., Knight, R., and Gordon, J. I. (2008). Worlds within worlds: evolution of the vertebrate gut microbiota. *Nature reviews. Microbiology*, 6(10):776–88.

Ley, R. E., Peterson, D. A., and Gordon, J. I. (2006). Ecological and evolutionary forces shaping microbial diversity in the human intestine. *Cell*, 124(4):837–48.

Librado, P. and Rozas, J. (2009). DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. *Bioinformatics*, 25(11):1451–1452.

Linnenbrink, M., Johnsen, J. M., Montero, I., Brzezinski, C. R., Harr, B., and Baines, J. F. (2011). Long-term balancing selection at the blood group-related gene *B4galnt2* in the genus *Mus* (Rodentia; Muridae). *Molecular Biology and Evolution*, 28(11):2999–3003.

Lozupone, C. and Knight, R. (2005). UniFrac: a new phylogenetic method for comparing microbial communities. *Applied and Environmental Microbiology*, 71(12):8228–35.

Ludwig, W., Strunk, O., Westram, R., Richter, L., Meier, H., Buchner, A., Lai, T., Steppi, S., Jobb, G., Förster, W., and Others (2004). ARB: a software environment for sequence data. *Nucleic Acids Research*, 32(4):1363–1371.

Mackintosh, J. (1970). Territory formation by laboratory mice. *Animal Behaviour*, 18:177–183.

Mendel, G. (1865). Versuche über Pflanzen-Hybriden. Verhandlungen des Naturforschenden Vereins zu Brünn. pages 3–47.

Mohlke, K. L., Nichols, W. C., Westrick, R. J., Novak, E. K., Cooney, K. A., Swank, R. T., and Ginsburg, D. (1996). A novel modifier gene for plasma von Willebrand factor level maps to distal mouse chromosome 11. *Proceedings of the National Academy of Sciences of the United States of America*, 93(26):15352–7.

Mohlke, K. L., Purkayastha, A. A., Westrick, R. J., Smith, P. L., Petryniak, B., Lowe, J. B., and Ginsburg, D. (1999). Mvwf, a dominant modifier of murine von Willebrand factor, results from altered lineage-specific expression of a glycosyltransferase. *Cell*, 96(1):111–20.

Mongodin, E. F., Emerson, J. B., and Nelson, K. E. (2005). Microbial metagenomics. *Genome biology*, 6(10):347.

Montero, I. (2010). *Mate choice and reproductive strategies in recently diverged populations of the house mouse ( Mus musculus domesticus )*. PhD thesis, Christian-Albrechts University Kiel.

Montiel, M.-D., Krzewinskis-Recchi, M.-A., Delannoy, P., and Harduin-Lepers, A. (2003). Molecular cloning , gene organization and expression of the human UDP-GalNAc : Neu5Ac$\alpha$2-3Gal$\beta$-R $\beta$1 , 4- N -acetylgalactosaminyltransferase responsible for the biosynthesis of the blood group Sd a / Cad antigen : evidence for an unusual extended cytoplasmi. *Biochemical Journal*, 373:369–379.

Newman, R. M., Hall, L., Connole, M., Chen, G.-L., Sato, S., Yuste, E., Diehl, W., Hunter, E., Kaur, A., Miller, G. M., and Johnson, W. E. (2006). Balancing selection and the evolution of functional polymorphism in Old World monkey *TRIM5alpha*. *Proceedings of the National Academy of Sciences of the United States of America*, 103(50):19134–9.

Ochman, H., Worobey, M., Kuo, C.-H., Ndjango, J.-B. N., Peeters, M., Hahn, B. H., and Hugenholtz, P. (2010). Evolutionary relationships of wild hominids recapitulated by gut microbial communities. *PLoS Biology*, 8(11):e1000546.

Oh, P. L., Benson, A. K., Peterson, D. A., Patil, P. B., Moriyama, E. N., Roos, S., and Walter, J. (2010). Diversification of the gut symbiont *Lactobacillus reuteri* as a result of host-driven evolution. *The ISME Journal*, 4(3):377–87.

O'Hara, A. M. and Shanahan, F. (2006). The gut flora as a forgotten organ. *EMBO Reports*, 7(7):688–93.

Oksanen, J., Blanchet, F. G., Kindt, R., Minchin, P. R., Hara, R. B. O., Simpson, G. L., Soly, P., Stevens, M. H. H., and Wagner, H. (2011). Package ”vegan”.

Ott, S., Musfeldt, M., Wenderoth, D., Hampe, J., Brant, O., Fölsch, U., Timmis, K., and Schreiber, S. (2004). Reduction in diversity of the colonic mucosa associated bacterial microflora in patients with active inflammatory bowel disease. *Gut*, 53(5):685–693.

Palmer, C., Bik, E. M., DiGiulio, D. B., Relman, D. A., and Brown, P. O. (2007). Development of the human infant intestinal microbiota. *PLoS Biology*, 5(7):e177.

Palomo, L. J., Justo, E. R., and Vargas, J. M. (2009). *Mus spretus* (Rodentia: Muridae). *Mammalian Species*, 840:1–10.

Paradis, E., Bolker, B., Claude, J., Cuong, H. S., Desper, R., Du, B., Dutheil, J., Gascuel, O., Heibl, C., Lawson, D., Lefort, V., Lemon, J., Noel, Y., Nylander, J., Popescu, A.-a., Schliep, K., Strimmer, K., and Vienne, D. D. (2012). Package ”ape”.

Peterson, D. a., McNulty, N. P., Guruge, J. L., and Gordon, J. I. (2007). IgA response to symbiotic bacteria as a mediator of gut homeostasis. *Cell Host and Microbe*, 2(5):328–39.

Prager, E. M., Sage, R. D., Gyllensten, U. L. F., Thomas, W. K., Hubner, R., Jones, C. S., Noble, L. E. S., Searle, J. B., and Wilson, A. C. (1993). Mitochondrial DNA sequence diversity and the colonization of Scandinavia by house mice from East Holstein. *Biological Journal of the Linnean Society*, 50(2):85–122.

Price, M. N., Dehal, P. S., and Arkin, A. P. (2009). FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. *Molecular Biology and Evolution*, 26(7):1641–50.

Pritchard, J. K., Stephens, M., and Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics*, 155(2):945–59.

R Development Core Team (2011). R: A language and environment computing.

Rausch, P., Rehman, A., Künzel, S., Häsler, R., Ott, S. J., Schreiber, S., Rosenstiel, P., Franke, A., and Baines, J. F. (2011). Colonic mucosa-associated microbiota is influenced by an interaction of Crohn disease and *FUT2* (Secretor) genotype. *Proceedings of the National Academy of Sciences of the United States of America*, 108(47):19030–5.

Rawls, J. F., Mahowald, M. A., Ley, R. E., and Gordon, J. I. (2006). Reciprocal gut microbiota transplants from zebrafish and mice to germ-free recipients reveal host habitat selection. *Cell*, 127(2):423–33.

Rehman, A., Sina, C., Gavrilova, O., Häsler, R., Ott, S., Baines, J. F., Schreiber, S., and Rosenstiel, P. (2011). *Nod2* is essential for temporal development of intestinal microbial communities. *Gut*, pages 1–10.

Riesenfeld, C. S., Schloss, P. D., and Handelsman, J. (2004). Metagenomics: genomic analysis of microbial communities. *Annual Review of Genetics*, 38:525–52.

Robinson, M. G., Tolchin, D., and Halpern, C. (1971). Enteric bacterial agents and the ABO blood groups. *American Journal of Human Genetics*, 23(2):135–45.

Rockman, M. V., Hahn, M. W., Soranzo, N., Goldstein, D. B., and Wray, G. A. (2003). Positive Selection on a Human-Specific Transcription Factor Binding Site Regulating IL4 Expression. *Current Biology*, 13(23):2118–2123.

Rodeghiero, F., Castaman, G., and Dini, E. (1987). Epidemiological investigation of the prevalence of von Willebrand's disease. *Blood*, 69(2):454–459.

Round, J. L. and Mazmanian, S. K. (2009). The gut microbiota shapes intestinal immune responses during health and disease. *Nature Reviews. Immunology*, 9(5):313–23.

Salminen, S., Gibson, G., McCartney, A., and Isolauri, E. (2004). Influence of mode of delivery on gut microbiota composition in seven year old children. *Gut*, 53(9):1388–1389.

Sanger, F., Nicklen, S., and Coulson, A. R. (1992). DNA sequencing with chain-terminating inhibitors. 1977. *Biotechnology*, 24(12):104–8.

Schloss, P. D., Westcott, S. L., Ryabin, T., Hall, J. R., Hartmann, M., Hollister, E. B., Lesniewski, R. a., Oakley, B. B., Parks, D. H., Robinson, C. J., Sahl, J. W., Stres, B., Thallinger, G. G., Van Horn, D. J., and Weber, C. F. (2009). Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Applied and Environmental Microbiology*, 75(23):7537–41.

Simmons, J. H., Riley, L. K., Besch-Williford, C. L., and Franklin, L. (2000). *Helicobacter mesocricetorum* sp . nov ., a Novel *Helicobacter* Isolated from the Feces of Syrian Hamsters. *Journal of Clinical Microbiology*, 38(5):1811–1817.

Slatkin, M. (1995). A measure of population subdivision based on microsatellite allele frequencies. *Genetics*, 139(1):457–462.

Smit, P. and Heniger, J. (1975). Antoni van Leeuwenhoek (1632-1723) and the discovery of bacteria. *Antonie van Leeuwenhoek*, 41:217–228.

Sonnenburg, J. L., Xu, J., Leip, D. D., Chen, C.-H., Westover, B. P., Weatherford, J., Buhler, J. D., and Gordon, J. I. (2005). Glycan foraging in vivo by an intestine-adapted bacterial symbiont. *Science*, 307(5717):1955–9.

Spor, A., Koren, O., and Ley, R. (2011). Unravelling the effects of the environment and host genotype on the gut microbiome. *Nature Reviews. Microbiology*, 9(4):279–90.

Stahl, E. A., Dwyer, G., Mauricio, R., Kreitman, M., and Bergelson, J. (1999). Dynamics of disease resistance polymorphism at the *Rpm1* locus of Arabidopsis. *Nature*, 400(6745):667–71.

Stajich, J. E., Block, D., Boulez, K., Brenner, S. E., Chervitz, S. a., Dagdigian, C., Fuellen, G., Gilbert, J. G. R., Korf, I., Lapp, H., Lehväslaiho, H., Matsalla, C., Mungall, C. J., Osborne, B. I., Pocock, M. R., Schattner, P., Senger, M., Stein, L. D., Stupka, E., Wilkinson, M. D., and Birney, E. (2002). The Bioperl toolkit: Perl modules for the life sciences. *Genome Research*, 12(10):1611–8.

Stajich, J. E. and Hahn, M. W. (2005). Disentangling the effects of demography and selection in human history. *Molecular Biology and Evolution*, 22(1):63–73.

Staubach, F., Künzel, S., Baines, A. C., Yee, A., McGee, B. M., Bäckhed, F., Baines, J. F., and Johnsen, J. M. (2012). Expression of the blood-group-related glycosyltransferase *B4galnt2* influences the intestinal microbiota in mice. *The ISME Journal*, pages 1–11.

Stecher, B., Chaffron, S., Käppeli, R., Hapfelmeier, S., Freedrich, S., Weber, T. C., Kirundi, J., Suar, M., McCoy, K. D., von Mering, C., Macpherson, A. J., and Hardt, W.-D. (2010). Like will to like: abundances of closely related species can predict susceptibility to intestinal colonization by pathogenic and commensal bacteria. *PLoS Pathogens*, 6(1):e1000711.

Stephens, M. and Donnelly, P. (2003). A comparison of bayesian methods for haplotype reconstruction from population genotype data. *The American Journal of Human Genetics*, 73(5):1162–1169.

Stephens, M., Smith, N. J., and Donnelly, P. (2001). A new statistical method for haplotype reconstruction from population data. *The American Journal of Human Genetics*, 68(4):978–989.

Storz, J. F., Baze, M., Waite, J. L., Hoffmann, F. G., Opazo, J. C., and Hayes, J. P. (2007). Complex signatures of selection and gene conversion in the duplicated globin genes of house mice. *Genetics*, 177(1):481–500.

Stuckenholz, C., Lu, L., Thakur, P., Kaminski, N., and Bahary, N. (2009). FACS-assisted microarray profiling implicates novel genes and pathways in zebrafish gastrointestinal tract development. *Gastroenterology*, 137(4):1321–32.

Sullivan, P. S., Grubbs, S. T., Olchowy, T. W., Andrews, F. M., White, J. G., Catalfamo, J. L., Dodd, P. A., and McDonald, T. P. (1994). Bleeding diathesis associated with variant von Willebrand factor in a Simmental calf. *Journal of the American Veterinary Medical Association*, 205(12):1763–1766.

Sweeney, J. D., Novak, E. K., Reddington, M., Takeuchi, K. H., and Swank, R. T. (1990). The RIIIS/J inbred mouse strain as a model for von Willebrand disease. *Blood*, 76(11):2258–65.

Tajima, F. (1983). Evolutionary relationship of DNA sequences in finite populations. *Genetics*, 105:437–460.

Tajima, F. (1989). Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*, 123(3):585.

Tamura, K., Dudley, J., Nei, M., and Kumar, S. (2007). MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. *Molecular Biology and Evolution*, 24(8):1596–9.

Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M., and Kumar, S. (2011). MEGA5: Molecular Evolutionary Genetics Analysis using Maximum Likelihood, Evolutionary Distance, and Maximum Parsimony Methods. *Molecular Biology and Evolution*, 28(10):2731–2739.

Thomas, M., Möller, F., Wiehe, T., and Tautz, D. (2007). A pooling approach to detect signatures of selective sweeps in genome scans using microsatellites. *Molecular Ecology*, 7(3):400–403.

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

Turnbaugh, P. J., Hamady, M., Yatsunenko, T., Cantarel, B. L., Duncan, A., Ley, R. E., Sogin, M. L., Jones, W. J., Roe, B. a., Affourtit, J. P., Egholm, M., Henrissat, B., Heath, A. C., Knight, R., and Gordon, J. I. (2009). A core gut microbiome in obese and lean twins. *Nature*, 457(7228):480–4.

Vijay-Kumar, M., Aitken, J. D., Carvalho, F. a., Cullender, T. C., Mwangi, S., Srinivasan, S., Sitaraman, S. V., Knight, R., Ley, R. E., and Gewirtz, A. T. (2010). Metabolic syndrome and altered gut microbiota in mice lacking Toll-like receptor 5. *Science*, 328(5975):228–31.

Wade, C., Kulbokas, E., Kirby, A., Zody, M., Mullikin, J., Lander, E., Lindblad-Toh, K., and Daly, M. (2002). The mosaic structure of variation in the laboratory mouse genome. *Nature*, 420(6915):574–578.

Wade, C. M. and Daly, M. J. (2005). Genetic variation in laboratory mice. *Nature Genetics*, 37(11):1175–80.

Walter, J. and Ley, R. E. (2011). The Human Gut Microbiome: Ecology and Recent Evolutionary Changes. *Annual Review of Microbiology*, (June):411–429.

Wang, Q., Garrity, G. M., Tiedje, J. M., and Cole, J. R. (2007). Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Applied and Environmental Microbiology*, 73(16):5261–7.

Watterson, G. A. (1975). On the number of segregating sites in genetical models without recombination. *Theoretical Population Biology*, 7(2):256–276.

Weir, B. S., Hill, W. G., and Cardon, L. R. (2004). Allelic association patterns for a dense SNP map. *Genetic Epidemiology*, 27(4):442–50.

Wheat, C. W., Haag, C. R., Marden, J. H., Hanski, I., and Frilander, M. J. (2010). Nucleotide polymorphism at a gene (*Pgi*) under balancing selection in a butterfly metapopulation. *Molecular Biology and Evolution*, 27(2):267–81.

Whitman, W., Coleman, D., and Wiebe, W. (1998). Prokaryotes: the unseen majority. *Proceedings of the National Academy of Sciences of the United States of America*, 95(12):6578–6583.

Wittkopp, P., Haerum, B., and Clark, A. (2004). Evolutionary changes in cis and trans gene regulation. *Nature*, 430(6995):85–88.

Wray, G. A. (2007). The evolutionary significance of cis-regulatory mutations. *Nature Reviews. Genetics*, 8(3):206–16.

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

Wright, S. (1951). The genetical structure of populations. *Annals of Eugenics*, 15(1):323–354.

Yamamoto, S., Morita, K., Koizumi, I., and Maekawa, K. (2004). Genetic differentiation of white-spotted charr (*Salvelinus leucomaenis*) populations after habitat fragmentation : Spatial temporal changes in gene frequencies. *Conservation Genetics*, 5:529–538.

Yonekawa, H., Moriwaki, K., Gotoh, O., Myashita, N., Matsushimaj, Y., Shi, L., Cho, W. S., Zhen, X.-l., and Tagashira, Y. (1988). Hybrid Origin of Japanese Mice "molossinus": Evidence from Restriction Analysis of Mitochondrial DNA. *Molecular Biology and Evolution*, 5(1):63–78.

Zoetendal, E., Akkermans, A., Akkermans-van Vliet, W., de Visser, J., and de Vos, W. (2001). The Host Genotype Affects the Bacterial Community in the Human Gastrointestinal Tract. *Microbial Ecology in Health and Disease*, 13(3):129–134.

# Appendix

## Appendix A - Online Material Chapter 1

All Online Material can also be found on:
http://mbe.oxfordjournals.org/content/early/2011/09/26/molbev.msr150/suppl/DC1

1. Animal Material
2. Table A-S1
3. Table A-S2
4. Figure A-S1
5. Figure A-S2
6. Figure A-S3

Figure A-S2 and A-S3 are printed in this Appendix just as ”sketch”. For an appropriate view please look at the Online Material on the homepage of Molecular Biology and Evolution (see above). Both Figures are too big to show them in an appropriate size in this Appendix.

**Animal material.**

The population samples of *M. m. domesticus* from the Massif Central region of France, *M. m. castaneus* from India and *M. m. musculus* from Kazakhstan were comprised wild-caught individuals as previously described (Ihle *et al.*, 2006; Baines and Harr, 2007). The *M. spretus* population was sampled in and around the vicinity of Madrid, Spain in 2004. A newly collected sample of *M. m. domesticus* from the same locations in Iran as described in Baines and Harr (2007), for which a wild-derived colony was available for functional studies, was kindly provided by Rick Scavetta. Wild-derived colonies of all species/populations used in this study, with the exception of *M. m. castaneus* from India and *M. famulus*, are maintained at the breeding facility of the Max Planck Institute for Evolutionary Biology in Plön, Germany, and were used for functional analysis of *B4galnt2* expression patterns (DBA lectin staining). DNA from a single *M. famulus* individual (kindly provided by F. Bonhomme) was sequenced as an outgroup for additional *B4galnt2* sequence fragments.

Table A-S1: Polymorphism and divergence across the *B4galnt2* upstream gene region.

| Fragment (position[a]) | population | sample size | length | $S^b$ | $\pi^c$(%) | $\theta^d$ (%) | $K^e$ (%) | Tajima's *D* | |
|---|---|---|---|---|---|---|---|---|---|
| 3.1 (-19.3 kb) | IN | 16 | 679 | 4 | 0.26 | 0.34 | 3.63 | -0.75 | |
| | IR | 16 | 679 | 2 | 0.18 | 0.17 | 3.00 | 0.24 | |
| | MC | 16 | 679 | 2 | 0.24 | 0.17 | 3.00 | 1.09 | |
| | KH | 16 | 679 | 6 | 0.61 | 0.48 | 3.96 | 0.92 | |
| | SP | 32 | 679 | 4 | 0.66 | 0.43 | 5.16 | 1.36 | |
| 4 (-18.7 kb) | IN | 16 | 671 | 9 | 0.81 | 0.60 | 2.86 | 1.25 | |
| | IR | 16 | 671 | 3 | 0.32 | 0.18 | 2.79 | 2.20 | * |
| | MC | 16 | 671 | 4 | 0.34 | 0.23 | 2.88 | 1.59 | |
| | KH | 16 | 671 | 5 | 0.48 | 0.29 | 3.16 | 2.17 | * |
| | SP | 32 | 671 | 0 | 0.00 | n.a. | 2.66 | n.d. | |
| 4.1 (-17.9 kb) | IN | 16 | 260 | 12 | 1.87 | 1.46 | 3.79 | 1.06 | |
| | IR | 16 | 260 | 4 | 0.96 | 0.58 | 5.33 | 2.06 | * |
| | MC | 16 | 260 | 3 | 0.62 | 0.35 | 4.56 | 2.23 | * |
| | KH | 16 | 260 | 9 | 1.78 | 1.04 | 3.62 | 2.59 | ** |
| | SP | 32 | 260 | 10 | 1.68 | 1.41 | 1.00 | 0.59 | |
| 5 (-10.4 kb) | IN | 16 | 399 | 15 | 2.54 | 1.73 | 2.28 | 1.85 | |
| | IR | 16 | 399 | 16 | 2.41 | 1.43 | 3.18 | 2.69 | ** |
| | MC | 16 | 399 | 15 | 2.75 | 1.55 | 3.34 | 3.02 | *** |
| | KH | 16 | 399 | 14 | 2.18 | 1.26 | 2.78 | 2.82 | ** |
| | SP | 32 | 399 | 25 | 2.02 | 2.14 | 3.45 | -0.20 | |
| 5.1 (-7.6 kb) | IN | 16 | 611 | 24 | 1.23 | 1.38 | 2.36 | -0.45 | |
| | IR | 16 | 611 | 9 | 0.83 | 0.54 | 1.36 | 2.01 | * |
| | MC | 12 | 611 | 14 | 0.70 | 0.77 | 3.14 | -0.36 | |
| | KH | 16 | 611 | 15 | 1.27 | 0.82 | 2.35 | 2.19 | * |
| | SP | 32 | 611 | 9 | 0.52 | 0.85 | 0.58 | -1.17 | |
| 5.2 (-5.2 kb) | IN | 16 | 312 | 3 | 0.32 | 0.36 | 1.82 | -0.28 | |
| | IR | 16 | 312 | 0 | 0.00 | n.a. | 2.01 | n.d. | |
| | MC | 16 | 312 | 0 | 0.00 | n.a. | 2.02 | n.d. | |
| | KH | 6 | 312 | 5 | 0.80 | 0.71 | 2.15 | 0.71 | |
| | SP | 32 | 312 | 7 | 1.17 | 0.67 | 2.91 | 2.13 | * |
| 5.3 (-4.3 kb) | IN | 16 | 862 | 5 | 0.15 | 0.19 | 2.31 | -0.76 | |
| | IR | 16 | 862 | 2 | 0.04 | 0.07 | 2.37 | -1.04 | |
| | MC | 16 | 862 | 2 | 0.06 | 0.07 | 2.37 | -0.58 | |
| | KH | 16 | 862 | 7 | 0.36 | 0.25 | 2.25 | 1.43 | |
| | SP | 32 | 862 | 17 | 0.55 | 0.51 | 2.80 | 0.22 | |
| 5.4 (-2.8 kb) | IN | 16 | 506 | 5 | 0.19 | 0.31 | 1.73 | -1.22 | |
| | IR | 16 | 506 | 0 | 0.00 | n.a. | 1.64 | n.d. | |
| | MC | 16 | 506 | 5 | 0.61 | 0.40 | 2.19 | 1.64 | |
| | KH | 16 | 506 | 4 | 0.56 | 0.45 | 1.59 | 0.76 | |
| | SP | 32 | 506 | 2 | 0.24 | 0.13 | 2.19 | 1.56 | |
| 6.1.1 (-1.9 kb) | IN | 16 | 381 | 2 | 0.19 | 0.32 | | -1.04 | |
| | IR | 12 | 381 | 1 | 0.06 | 0.11 | | -1.14 | |
| | MC | 16 | 381 | 8 | 1.50 | 0.86 | | 2.69 | ** |
| | KH | 14 | 381 | 0 | 0.00 | n.a. | | n.d. | |
| | SP | 32 | 381 | 1 | 0.02 | 0.08 | | -1.14 | |
| 6.1.2 (-1.4 kb) | IN | 16 | 354 | 3 | 0.38 | 0.31 | | 0.71 | |
| | IR | 16 | 354 | 2 | 0.16 | 0.20 | | -0.58 | |
| | MC | 16 | 354 | 3 | 0.43 | 0.30 | | 1.27 | |
| | KH | 16 | 354 | 3 | 0.64 | 0.41 | | 1.64 | |
| | SP | 32 | 354 | 5 | 0.22 | 0.45 | | -1.38 | |
| 6.2 (-1 kb) | IN | 16 | 690 | 7 | 1.19 | 1.19 | 3.76 | -0.02 | |
| | IR | 16 | 690 | 0 | 0.00 | n.a. | 4.39 | n.d. | |
| | MC | 16 | 690 | 12 | 1.20 | 0.73 | 4.53 | 2.51 | ** |
| | KH | 16 | 690 | 9 | 0.63 | 0.48 | 5.00 | 1.21 | |
| | SP | 32 | 690 | 6 | 0.30 | 0.32 | 5.35 | -0.19 | |

* $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$

[a] Position relative to *B4galnt2* start codon

[b] S = the number of segregating sites

[c] $\pi$ = nucleotide diversity (Tajima, 1983)

[d] $\theta$ = nucleotide diversity (Watterson, 1975)

[e] K = divergence

IN = *M. m. castaneus*

IR = *M. m. domesticus* from Iran

MC = *M. m. domesticus* from France

KH = *M. m. musculus*

SP = *M. spretus*

Table A-S2: Polymorphism and divergence at seven unlinked reference loci in *M. spretus*.

| Locus | sample size | length | S[a] | $\pi$[b](%) | $\theta$[c] (%) | K[d] (%) | Tajima's *D* | |
|---|---|---|---|---|---|---|---|---|
| Sfrp1 | 16 | 1231 | 20 | 0.70 | 0.40 | 0.70 | 2.54 | * |
| Fut10 | 16 | 544 | 9 | 0.31 | 0.46 | 0.50 | -0.99 | |
| Nkd1 | 16 | 1431 | 23 | 0.51 | 0.42 | 0.32 | 0.78 | |
| Nudt7 | 16 | 1445 | 31 | 0.54 | 0.53 | 0.47 | 0.03 | |
| Ggh | 16 | 579 | 11 | 0.74 | 0.47 | 0.94 | 1.81 | |
| Melk | 16 | 1151 | 7 | 0.17 | 0.15 | 0.08 | 0.40 | |
| Pum1 | 16 | 1283 | 3 | 0.03 | 0.06 | 3.77 | -1.08 | |
| SUM | | 7664 | 104 | | | | | |
| Mean (SE) | | | | 0.43 (0.10) | 0.36 (0.07) | 0.97 (0.48) | 0.5 (0.50) | |

* $p < 0.05$
[a] S = the number of segregating sites
[b] $\pi$ = nucleotide diversity (Tajima, 1983)
[c] $\theta$ = nucleotide diversity (Watterson, 1975)
[d] K = divergence

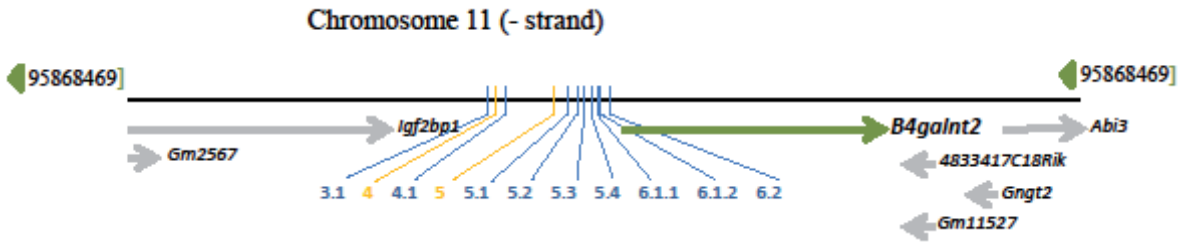


Figure A-S1: **Gene region of *B4galnt2* and the distribution of all sequenced fragments in its upstream region.** Fragments marked in orange indicate the original fragments of Johnsen *et al.* (2009), blue fragments indicate the newly designed fragments. The distance between the 5'- and 3'-most fragments is ∼20kb.

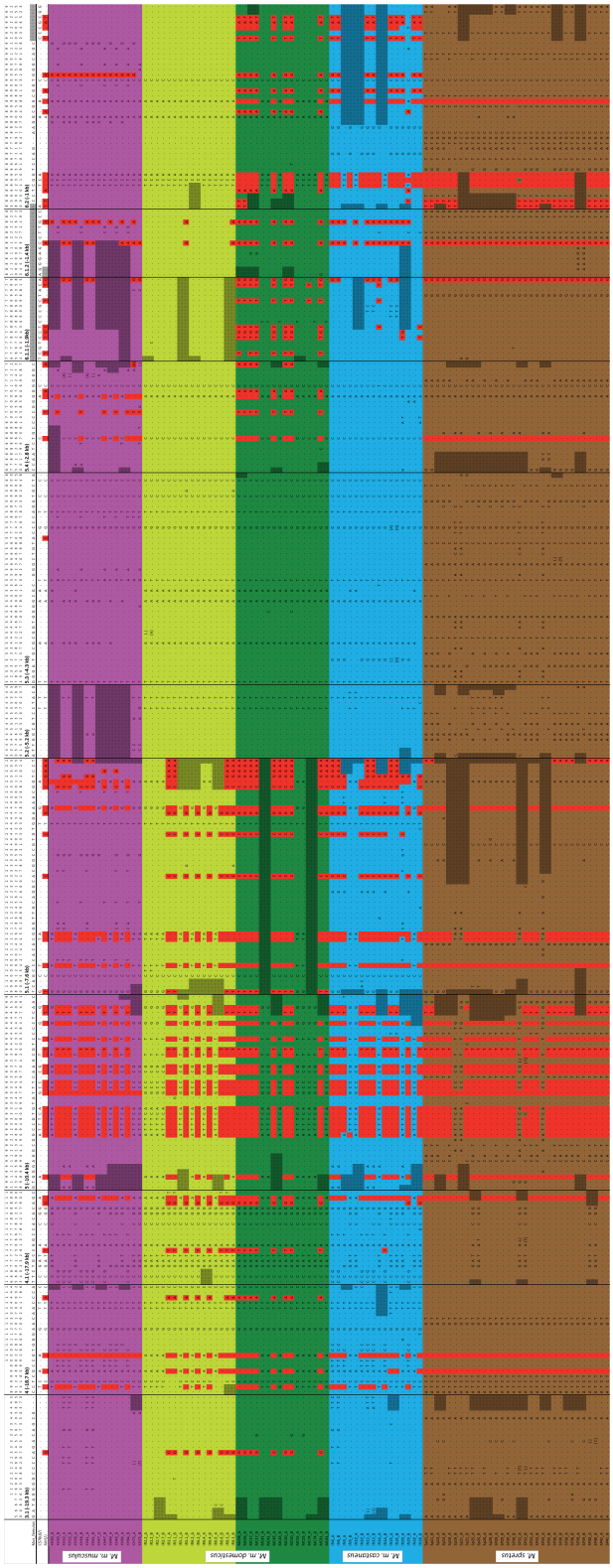


Figure A-S2: **Summary of polymorphic sites across *B4galnt2* gene region.** Haplotypic phase was determined using PHASE version 2.1 (Stephens *et al.*, 2001; Stephens and Donnelly, 2003). The populations included are from *M. m. domesticus* from Iran (IR) and France (MC), *M. m. castaneus* from India (IN), *M. m. musculus* from Kazakhstan (KH) and *M. spretus* from Spain (SP). The sequences of the inbred strains C57BL6/J and RIIIS/J as well as a *M. famulus* individual were included for reference. Uncertain phase estimates ($< 90\%$) are shown in brackets.

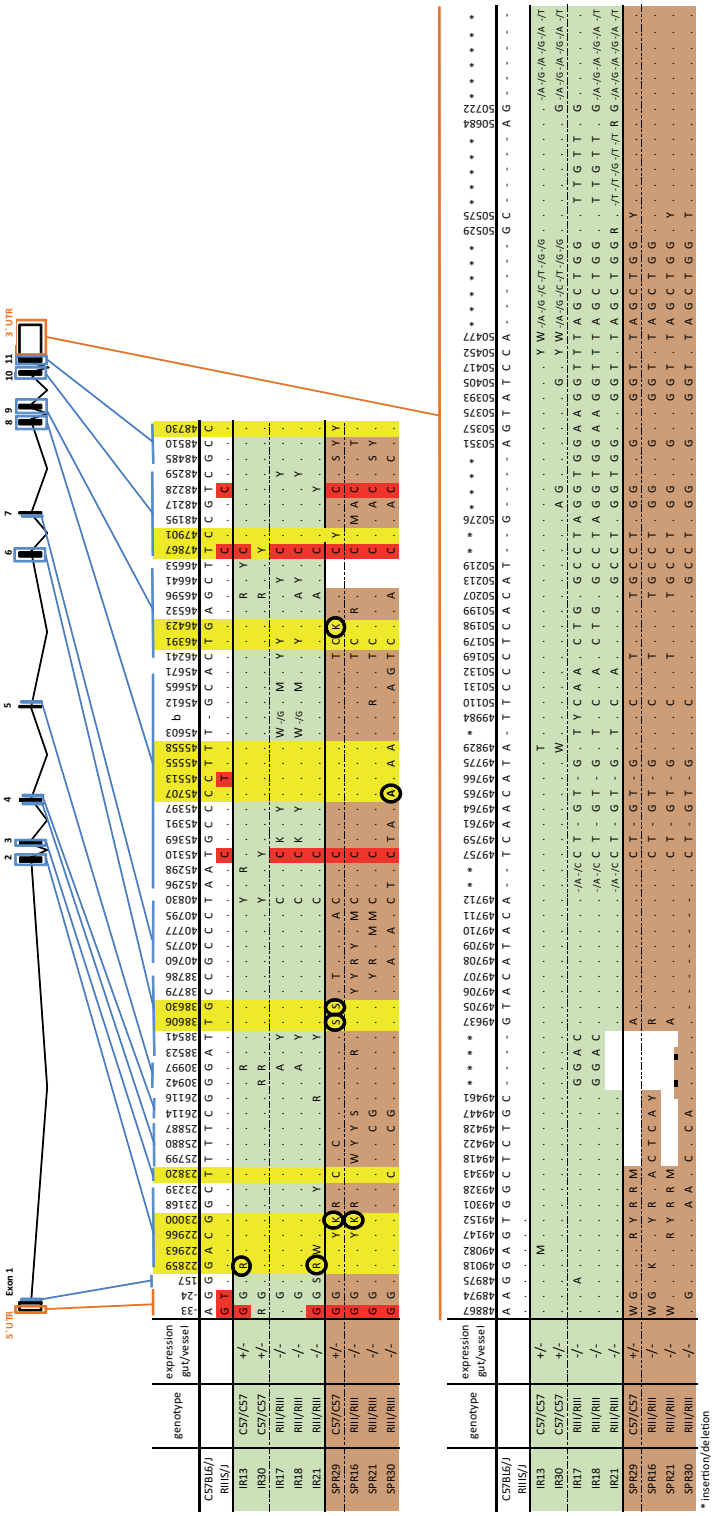


Figure A-S3: **Summary of sequence comparison between *B4galnt2* expression classes.** Coding regions are highlighted in yellow and differences between the C57BL6/J and RIIIS/J reference strains in red. Fixed differences between IR and SP are not shown. Circles indicate amino acid changes. Shaded sites indicate missing data.

# Appendix B - Supplementary Materials Chapter 2

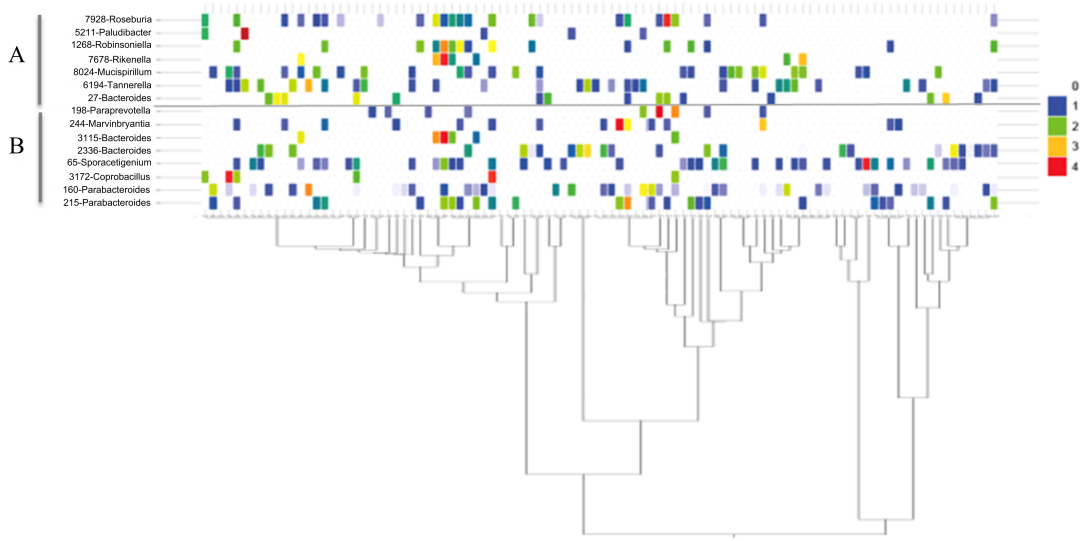


Figure B-S1: Log relative abundances of OTUs displaying a significant association to A) the interaction between mtDNA and geography and/or population structure or B) exclusively to mtDNA

Table B-S1: Microsatellite information.

Microsatellite information will be uploaded to the journal's homepage as soon as the paper is accepted for publication.

# Appendix C - Supplementary Information Chapter 3


Figure C-S1: Neighbour-Joining tree of the D-loop sequences from all mice used in this study, as well as sequences from *M. m. domesticus*, *M. m. musculus*, *M. spicilegus*, *M. spretus* and *M. famulus*, used as reference/outgroups.

# Appendix D - Primerlist

Table D-S1: List of all primers used in this study.

| Primer to amplify *B4galnt2* Exons + UTR regions | | Primer to amplify the fragments upstream *B4galnt2* | |
|---|---|---|---|
| Name | Primer 5'–> 3' | Name | Primer 5'–>3' |
| Exon1_F | GTAAATACAGCGTTAGAACCTGAC | Fragment_3.1_F | AGAAAGGTGCTATTCAGGGAG |
| Exon1_R | ACAACTCGTTTCTGATCCTTCTG | Fragment_3.1_R | TATGTGAGTACACTGTGGCTG |
| Exon2_F | TGCAGCGTTACTGTCTACACC | Fragment_4_F | TGCTGGGTTGAAAGCCTGTTAGC |
| Exon2_R | TTGGCAGCTAGTCCCATTAGG | Fragment_4_R | ACCTTTATTTTCCATGACCTCTG |
| Exon3_F | TCTTTAGTCCCTCCCTTCATCC | Fragment_4.1_F | AGTTTGGTTCACAGAAGTTAGC |
| Exon3_R | TATACACAAAGCCCTGCAAATC | Fragment_4.1_R | GGATGTATAAATAGGACTGAACGG |
| Exon4_F | TATAAAGGGCTTGTCAGGGTTG | Fragment_5_F | TCCCACAAGGTGTATGTGATG |
| Exon4_R | GAACTTGACATTGGGCTAGGAC | Fragment_5_R | TGTTCCTTTGATCCTGTCTC |
| Exon5_F | CTTGGAAGAGAACCAGTTCAGC | Fragment_5.1_F | AGGGGCACTTCACAGAAATCTT |
| Exon5_R | CTCCCTACAGTCACTCCAAAGG | | TTTAACAGAGTTTACAAAGG |
| Exon6_F | CATCCTGAGTACTTGCCTTTCC | Fragment_5.1_R | TTGTGGCAATGCAGTATTATGAA |
| Exon6_R | ATTGGGTGCAGTAAATAGGCAG | | TATTATGAAACTGTCTCTTAC |
| Exon7_F | ACTGTCCTGGTCTTGTCATTGG | | TTGTGGCAATGCACTATTATGAA |
| Exon7_R | ATATCCCAACAGGGCCTCCTAC | Fragment_5.2_F | GTCATCAAACATGTTGTTCCTC |
| Exon8_F | TAAACAACTCCTGCAACTCCTG | Fragment_5.2_R | CTACTGCAGAGTGTATATGC |
| | TGCAACTCCTGTGTACAGAAGG | Fragment_5.3_F | GTGGCAGAAGTTGTTATAATCC |
| Exon8_R | TCTGCATCAGAACAAGAGAAGC | Fragment_5.3_R | CCTTTCCTCATCTTTAACTTGG |
| | ATGTTTGGCGGAAGATAGGAGG | Fragment_5.4_F | AGTAACAGTAGGCTCCTTTGAG |
| Exon9_F | AGCCTCTGATCCAGAAAGTCAC | Fragment_5.4_R | AAGTTCAGTCTGCTAGCCAC |
| | TATCACTGAGCGCAAATGTTCC | Fragment_6.1_F | TTTAAGGGAAGGAGAAAGTGAG |
| Exon9_R | ATGTGACCGAACAAGAACATGG | Fragment_6.1_R | CTACTGTTGTTCTGATCCTGC |
| | CTCCCTCTGACACTAACATTCC | Fragment_6_2_R | AAATGCAAGCCATCTCTGCT |
| Exon10_F | TGAACACTGATCTGCTCAGACC | | ATCAGAACAACAGTAGGATCC |
| Exon10_R | AGAAAGGTAGAGGGCTGAGCTGC | | ATCCCATTCTCGTCTCAACG |
| Exon11_F | ACAGTCCCATTTAGTGGAACCC | | ACAAATGCGGAACGATTTCG |
| Exon11_R | GTGCTGTGATTAAAGGTGTGCC | Fragment_6_2_F | AGGGTTTCACAAACACAAATGC |
| 3UTRpart1_F | GCACTCCAGTACTTCAAGAACC | | CAGCCCCGACGACAGTTCTGTG |
| | ACCACCAAGTAAGAACACCACC | | |
| 3UTRpart1_R | ATCTGTAATGGGATCTGATGC | Primer to amplify the mitochondrial D-loop | |
| | GGTTCTGAGTTCAATTCCCAGC | Name | Primer 5'–> 3' |
| 3UTRpart2_F | AAGACTGACCTCACACTCATGG | D-loop F | CATTACTCTGGTCTTGTAAACC |
| 3UTRpart2_R | GGATCTCTGAATTCAAGGCTGG | D-loop R | GCCAGGACCAAACCTTTGTGT |
| 3UTRpart2seqF | AGCCATCTCTCTAGTCTCTAGC | | |
| 3UTRpart3_F | AAGAGCAGTCCATCCTCTTACC | | |
| 3UTRpart3_R | TAACAGACTTGGTCACTGAGCC | | |

# Affidavit

I declare that the dissertation in form and content and except for advices given by my supervisor constitutes my own work. The first chapter: Long-Term Balancing Selection at the Blood Group-Related Gene *B4galnt2* in the Genus *Mus* (Rodentia; Muridae) published for publication in Molecular Biology and Evolution in 2011. The second chapter: The Role of Biogeography in Shaping Diversity of the Intestinal Microbiota in House Mice was submitted to Molecular Ecology in March 2012. This work has been undertaken in compliance with the German Research Foundations (Deutsche Forschungsgemeinschaft, DFG) rules of good academic practice.

Kiel, $3^{rd}$ of April 2012 Miriam Linnenbrink

# Curriculum Vitae

**Miriam Linnenbrink**
Date of Birth: 26th of April 1983
linnenbrink@evolbio.mpg.de
miriam.linnenbrink@yahoo.de
Nationality: German
Fürstenrieder Strasse 19, 80687 München
Home Town: Munich

## EDUCATION

**Since 11/2008** PhD Study

Ludwig-Maximilians-University, Munich, Germany - Section of Evolutionary Biology
Christian Albrechts University, Kiel, Germany - Section of Evolutionary Genomics
Max Planck Institute for Evolutionary Biology, Plön, Germany

**10/2002 - 01/2008** Diploma Biology at the Technical University of Munich

Thesis: ”Interspecific dominance between two sympatric mouse lemurs (*Microcebus murinus and M. ravelobensis*) in northwestern Madagascar” (together with the University of Veterinary Medicine, Hannover, Germany)
Diploma research included 3 month field work in north west Madagascar
Main Focus: Zoology, Animal Ecology, Behavioral Ecology, Ecotoxicology

## RESEARCH EXPERIENCE AND FIELDWORK

**02/2012**
HiWi for an ”Evolutionary Biology” Bachelor Course at the LMU Munich - Evolutionary and Functional Genomics

**Since 11/2008 - 04/2012**
PhD Student of Biology
Subjects: Molecular Evolution, Population Genetics, natural populations of house mice. Including 3 field seasons in France and Germany

**08/2008 - 10/2008**
Lab/Field Assistant at the Technical University of Munich - Institute of Ecotoxicology

**04/2007 - 07/2007**
Field work in north west Madagascar for my Diploma thesis

**05/2006 - 02/2007**
Lab Assistant (molecular biology) at the Technical University of Munich - Institute of Zoology

## PUBLICATIONS

### 2011

Thorén S, **Linnenbrink M**, Radespiel U. Different competitive potential in two coexisting mouse lemur species in northwestern Madagascar. Am. J. Phys. Anthropol. 145, 156-162.

**Linnenbrink M**, Johnsen JM, Montero I, Brzezinski CR, Harr B, Baines JF . Long-term balancing selection at the blood group-related gene *B4galnt2* in the genus Mus (Rodentia; Muridae). Mol.Biol.Evol 28, 2999-3003. doi: 10.1093/molbev/msr150

**Linnenbrink M**, Jun Wang, Emilie A. Hardouin, Sven Künzel, Dirk Metzler and John F. Baines. The Role of Biogeography in Shaping Diversity of the Intestinal Microbiota in House Mice. (Molecular Ecology, in Review)

# CONFERENCES and PRESENTATIONS

**2008**

**Linnenbrink, M.**; Thorén, S.; Radespiel, U.: talk: Interspecific dominance between two sympatric mouse lemurs (*Microcebus murinus* and *M. ravelobensis*) in northwestern Madagascar Ethologische Gesellschaft (Hrsg.): Jahrestagung der Ethologischen Gesellschaft, abstracts Regensburg, $20^{th}$-$22^{nd}$ February 2008; S. 37

**2009**

Thorén, S.; **Linnenbrink, M.**; Radespiel, U.: poster: Experiments on inter-specific food competition in two coexisting mouse lemur species in north-western Madagascar Folia Primatologica 80 (2) $3^{rd}$ Congress of the European Federation for Primatology, Zürich, $12^{th}$ -$15^{th}$ August 2009; Basel: Karger, 2009, S. 145 ISSN 0015-5713

**Linnenbrink, Miriam** talk: Selection on expression variation at *B4galnt2* in natural populations of house mice $7^{th}$ Aquavit Symposium, $14^{th}$-$15^{th}$ May 2009, MPI for Evolutionary Biology, Plön, Germany.

**2010**

**Linnenbrink, Miriam**; Montero, Inka and Baines, John F. poster: Selection on expression variation at *B4galnt2* in natural populations of house mice, 43rd Population Genetics Group Meeting, $5^{th}$ - $8^{th}$ January 2010, in Liverpool, UK

**Miriam Linnenbrink** talk: Expression variation at *B4galnt2* is associated with a bleeding disorder in house mice - which evolutionary forces governs this process? Next-Generation Sequencing: New Chances and Challenges for Evolutionary Genetics, March $10^{th}$ - $12^{th}$, 2010, LMU Biocenter Martinsried

**Miriam Linnenbrink** talk: Expression variation at *B4galnt2* is associated with a bleeding disorder in house mice - which evolutionary forces govern this process? Evolution on an Island, Second Status Symposium in Evolutionary Biology, $9^{th}$ - $12^{th}$ May 2010 Fraueninsel, Lake Chiemsee, Germany

**2011**

**Miriam Linnenbrink**, Jill M. Johnsen, Inka Montero, Christine R. Brzezinski, Bettina Harr, John F. Baines poster: Long-term balancing selection at the blood group-related gene *B4galnt2* in the genus Mus (Rodentia; Muridae) $13^{th}$ Congress of the European Society for Evolutionary Biology, $20^{th}$-$25^{th}$ August, Tübingen, Germany

**Miriam Linnenbrink** talk: Long-term balancing selection at the blood group-related gene *B4galnt2* in the genus Mus (Rodentia; Muridae) Otto Warburg International Summer School and Research Symposium 2011 on Evolutionary Genomics, $14^{th}$ - $22^{nd}$ September 2011, Berlin, Germany

**Miriam Linnenbrink** talk: Long-term balancing selection at *B4galnt2* in natural populations of house mice $5^{th}$ EES conference, $11^{th}$ and $12^{th}$ october 2011, Munich, Germany

## GRANTS

**2009** EES travel fund for joining the $43^{rd}$ Population Genetics Group Meeting in Liverpool, UK

**2010** EES travel fund for joining the $13^{th}$ Congress of the European Society for Evolutionary Biology, Tuebingen, Germany