Gene Expression Divergence between Subspecies of the House Mouse and the Contribution to Reproductive Isolation

Inaugural – Dissertation

zur
Erlangung des Doktorgrades
der Mathematisch-Naturwissenschaftlichen Fakultät
der Universität zu Köln

vorgelegt von

Ruth Rottscheidt

aus Bad Kreuznach

Köln, 2007

Berichterstatter:	**Dr. Bettina Harr**
	Prof. Dr. Diethard Tautz
Tag der letzten mündlichen Prüfung:	**19. Februar 2008**

Table of Contents

Danksagung ... IV

Zusammenfassung ... VI

Abstract ... VIII

Declaration ... X

1 General Introduction ... 1

1.1 Species, speciation and the evolution of reproductive isolation ... 1

1.2 Hybrid incompatibilities associated with gene expression differences ... 6

1.3 The house mouse ... 7

1.4 Aim of the study ... 13

2 Genome wide expression patterns in house mouse subspecies and their F1 hybrids ... 14

2.1 Introduction ... 14

2.2 Materials and Methods ... 17

2.2.1 Animals ... 17

2.2.2 Sample preparation ... 17

2.2.3 Microarray analysis ... 18

2.2.4 Data analysis ... 18

2.3 Results ... 24

2.3.1 Number of differentially expressed genes ... 24

2.3.2 Additivity/nonadditivity of transcripts differentially expressed between the subspecies ... 28

2.3.3 Nonadditive expression in F1 hybrids at transcripts that are not differentially expressed between the parents ... 30

2.3.4 Contribution of sequence divergence to expression change ... 31

2.3.5 Functional categories of transcripts ... 34

2.4 Discussion ... 35

3 Confirmation of microarray candidate genes: factors affecting expression data in cross-subspecies analyses ... 42

3.1 Introduction ... 42

3.2 Materials and Methods ... 45

3.2.1 Candidate Genes ... 45

3.2.2 Animals........ 46
3.2.3 Sample preparation........ 47
3.2.4 cDNA synthesis 47
3.2.5 Quantitative real-time PCR (qRT-PCR)........ 47
3.2.6 Analysis of qRT-PCR data 49
3.2.7 Sequencing of Affymetrix GeneChip® target regions........ 49
3.2.8 Sequencing of amplicon regions........ 51
3.3 Results........ 52
3.3.1 qRT-PCR using TaqMan® gene expression assays........ 52
3.3.2 qRT-PCR using Sybr Green fluorescent dye........ 54
3.3.3 Sequencing of Affymetrix GeneChip® target regions........ 55
3.3.4 Sequencing of the amplicon regions 58
3.3.5 Combining results – evidence for alternative splicing........ 59
3.4 Discussion 59
3.4.1 Comparison of two qRT-PCR methods........ 60
3.4.2 Sequence differences affecting microarray probe binding........ 63
3.4.3 Alternative splicing as contributor to expression differences........ 65
3.4.4 Conclusion........ 66
4 Characterization of candidate genes from a gene expression screen – insights from hybrid zone and population genetic analyses........ 67
4.1 Introduction 67
4.1.1 Protein and gene expression evolution........ 67
4.1.2 Detecting signatures of selection 68
4.1.3 *Cis*- versus *trans*-regulatory evolution........ 70
4.1.4 Hybrid zone analyses 70
4.1.5 Aim of the study........ 72
4.2 Materials and Methods........ 73
4.2.1 Animals........ 73
4.2.2 Sample preparation........ 74
4.2.3 cDNA synthesis 74
4.2.4 DNA sequencing........ 75
4.2.5 Gene expression analysis........ 75
4.2.6 Sequence data analysis 76

4.2.7 Correlation of expression-phenotype and genotype – determination of *cis*-or *trans*-acting factors 79
4.2.8 Assessment of functional consequences for candidate genes from hybrid zone analysis 80
4.3 Results 80
4.3.1 Gene expression analysis using Sybr Green 80
4.3.2 Sequence analysis........ 81
4.3.3 Correlation of expression-phenotype and genotype – determination of *cis*-or *trans*-acting factors 87
4.3.4 Assessment of functional consequences for candidate genes from hybrid zone analysis 89
4.4 Discussion 92
4.4.1 Protein coding versus regulatory evolution 93
4.4.2 *Cis*- versus *trans*-regulatory changes 99
4.4.3 Functional consequences of candidate genes in a hybrid zone........102
4.4.4 Conclusion 106
5 Literature 108
6 Supplement 127
7 Digital Supplement 152
Erklärung 153
Lebenslauf........154

Danksagung

Ich möchte mich bei all jenen bedanken, die mich auf unterschiedliche Art und Weise während der letzten Jahre unterstützt und damit dazu beigetragen haben, dass diese Arbeit nun in dieser Form vorliegt.

Mein erster Dank gilt Dr. Bettina Harr, die mir diese Arbeit ermöglichte. Mit ihren Ideen und ihrer Unterstützung hat sie maβgeblich zum Erfolg dieses Projektes beigetragen. Durch ihre Anregungen und durch viele Diskussionen hat sie mein Verständis für die molekulare Evolution stark erweitert.

Prof. Dr. Diethard Tautz möchte ich für die Übernahme des zweiten Gutachtens und für die Zeit in seiner Arbeitsgruppe danken. Die Möglichkeiten zu forschen, sich auszutauschen und zu informieren hätten idealer nicht sein können.

Weiterhin möchte ich Prof. Dr. Hartmut Arndt für die Übernahme des Prüfungsvorsitzes und Dr. Wim Damen für den Beisitz in meinem Prüfungskomitee danken.

Herzlich danken möchte ich Christian Voolstra, Meike Teschke und Fabian Staubach, die mir mit ihrer ständigen Diskussionsbereitschaft und unzähligen Erklärungen sehr geholfen haben. Vor allem Chris und Meike waren immer bereit, sich mit meinen Fragen auseinander zu setzen, und waren mir auch eine groβe Hilfe beim Korrekturlesen der Arbeit. Till Bayer war immer für mich da, wenn es um Computer- und andere technische Fragen ging. Auβerdem war er der perfekte Reisebegleiter und Mäusefänger in Spanien und Bayern. Birgit Schmitz war eine groβe Unterstützung bei Präparationen, RNA/DNA-Extraktionen, Sequenzierungen und bei der Organisation der mouse-facility. Vielen Dank auch an Christine Pfeiffle und Susanne Kipp für ihren unermüdlichen Einsatz in der mouse-facility und an zahlreiche Mauspfleger und Mauspflegerinnen, deren Arbeit von enormer Wichtigkeit für dieses Projekt war.

Der gesamten Arbeitsgruppe möchte ich für die gute Arbeitsatmosphäre danken. Neben viel Spaβ hatte ich auch immer das Gefühl einer Zusammengehörigkeit, was nicht selbstverständlich ist. Besonderer Dank gilt den Mädels Meike Teschke, Kathryn Stemshorn und Evelyn Schwager für die sportlichen und auch unsportlichen Mittagspausen und für die vielen guten Gespräche, Till Bayer für alles – Gespräche, Hilfe, Kuchen, Spaβ, u.v.m. –, Tobias Heinen und Fabian Staubach für das Quäntchen kölschen Frohsinns im Labor und Maarten Hilbrant für die entspannenden Stunden mit Tapas, Rotwein und Salsamusik.

Meine Freunde waren immer eine groβe Unterstützung, vor allem Petra, Rebekka und Alex. Vielen Dank für all die kleinen und groβen Dinge, die Ihr im Laufe der Zeit für mich getan habt.

Meiner Familie möchte ich einen ganz besonderen Dank aussprechen: Ihr habt mich immer unterstützt, ermutigt und Interesse am Fortschritt meiner Arbeit gezeigt. Wann immer ich Euch gebraucht habe – Ihr wart für mich da!

Zusammenfassung

Dem biologischen Artkonzept zufolge werden Arten als Gruppen von Individuen definiert, die sich miteinander fortpflanzen und von anderen Gruppen reproduktiv isoliert sind. Eine verbreitete Form reproduktiver Isolation im Tierreich ist intrinsische postzygotische Isolation durch Hybridensterilität oder erhöhte Hybridensterblichkeit. Der allgemeinen Überzeugung zufolge werden Fehlfunktionen in Hybriden durch epistatische Interaktionen zwischen inkompatiblen elterlichen Allelen verschiedener Loci (sogenannte Dobzhansky-Muller-Inkompatibilitäten) verursacht. Die Identifizierung von Genen, die zur reproduktiven Isolation von Taxa beitragen, ist wichtig für das Verständnis des Artbildungsprozesses, allerdings hat sich dies in der Vergangenheit als schwierig erwiesen.

Bei der Entstehung postzygotischer Isolation scheint die Evolution von Genregulation eine wichtige Rolle zu spielen. Deshalb wurden in der vorliegenden Studie mittels Microarrays genomweite Genexpressionsdaten erhoben, um regulatorische Unterschiede zwischen drei Unterarten der Hausmaus, *Mus musculus musculus*, *M. m. domesticus* und *M. m. castaneus*, zu identifizieren. Die Analyse von drei verschiedenen Organen (Gehirn, Leber und Testis) der jeweiligen Unterarten und ihrer reziproken Hybriden erlaubte die Bestimmung des Vererbungsmodus für Genexpressionsunterschiede in den F1-Hybriden. Der gröβte Teil der Transkripte zeigt additive Expression in den Hybriden; nur wenige sind dominant oder überdominant exprimiert, mit Ausnahme einer Kreuzung, die viele missexprimierte Gene im Testis aufweist. Drei verschiedene Analysemethoden sowie Kontrollexperimente bestätigen, dass additive Vererbung von Genexpressionsunterschieden vorherrschend zu sein scheint. Gene, die differentiell exprimierten Transkripten zugrunde liegen, sind vielversprechende Kandidaten, die zu reproduktiver Isolation durch regulatorische Inkompatibilitäten beitragen könnten.

Einge Transkripte mit Expressionsunterschieden zwischen *M. m. musculus* und *M. m. domesticus* wurden für weiterführende Untersuchungen ausgewählt. Die Validierung der Expressionsniveaus dieser Gene mittels quantitativer Real-Time PCR hat die Notwendigkeit verdeutlicht, Microarray-Daten durch unabhängige Methoden zu konfirmieren. Die Ergebnisse zeigen klar, dass Expressionsdaten, auch beim Vergleich sehr nah miteinander verwandter Taxa, durch Unterschiede in der

Gensequenz beeinflusst werden können, wobei dies für Microarray-Daten gleichermaβen gilt wie für nachfolgende Validierungsmethoden.

Unterschiede in der Evolution von Genexpression zwischen Taxa sind nicht zwangsläufig funktional bedingt. Aus diesem Grund wurde in der vorliegenden Studie versucht, funktionale Konsequenzen solcher Unterschiede mittels der Analyse von Individuen aus der natürlichen *musculus-domesticus* Hybridzone in Bayern für zwei ausgewählte Kandidaten-Gene nachzuweisen. Für diese zwei Gene, die einen groβen Expressionsunterschied zwischen den Elternarten zeigen, konnte kein Anzeichen dafür festgestellt werden, dass sie einen Beitrag zur reproduktiven Isolation der beiden Unterarten leisten, da sie in Bezug auf Introgression nicht limitiert zu sein scheinen. Die Kombination der Hybridzonenuntersuchung mit Ergebnissen aus populationsgenetischen Analysen lässt vermuten, dass eine adaptive Introgression der Allele, die das hohe Expressionslevel verursachen, wahrscheinlicher ist. Für beide Gene konnte nachgewiesen werden, dass der Phänotyp, der mit einem hohem Expressionsniveau einhergeht, das abgeleitete Merkmal darstellt und mit reduziertem Nucleotid-Polymorphismus und negativen Tajima's D-Werten assoziiert ist. Die Evolution von regulatorischer und Protein-kodierender Sequenz scheint für beide Gene voneinander entkoppelt und die Expressionsunterschiede ein Resultat von *cis*- und nicht *trans*-regulatorischen Änderungen zu sein.

Abstract

Under the Biological Species Concept species are groups of interbreeding individuals that are reproductively isolated from other such groups. A common form of isolation in animals is intrinsic postzygotic isolation via hybrid sterility/inviability. There is a strong consensus that hybrid dysfunctions are caused by epistatic interactions between incompatible alleles from different loci (Dobzhansky-Muller incompatibilities). The identification of genes that contribute to reproductive isolation between taxa is critical to the understanding of the process of speciation, but identifying such genes has proven to be difficult.

It appears that regulatory evolution might play an important role in postzygotic isolation and the formation of species. The present study employs a whole genome microarray approach to identify genes with regulatory differences between three subspecies of the house mouse, *Mus musculus musculus*, *M. m. domesticus* and *M. m. castaneus*. The within-locus mode of inheritance for gene expression was assessed for three different tissues (brain, liver and testis) by studying the subspecies and their male reciprocal F1 hybrids. The vast majority of transcripts are additively expressed in the hybrids with only few transcripts showing dominance or overdominance in expression except for one direction of one cross, which shows large misexpression in the testis. The reliability of the observed pattern was ensured by three different analysis methods as well as control experiments. The results suggest that additivity is the general mode of inheritance regarding gene expression changes between house mouse subspecies. Differentially expressed transcripts provide promising candidate genes that could be related to reproductive isolation through regulatory incompatibilities.

Several transcripts with expression differences between *M. m. musculus* and *M. m. domesticus* were selected for further investigation. A validation approach using quantitative Real-Time PCR strongly emphasizes the need for confirmation of microarray candidate genes. The results show that sequence differences even between closely related taxa have the potential to influence expression data from both microarray and follow-up validation approaches.

As divergent gene expression evolution between taxa may be entirely neutral, samples from a transect of the natural *musculus-domesticus* hybrid zone in Bavaria were analyzed in order to assess functional consequences for two candidate genes. Both genes show large expression differences between the subspecies. The analysis revealed that it is unlikely that the two genes contribute to reproductive isolation between the subspecies as no sign of limited introgression is evident. Rather, the hybrid zone approach in combination with population genetic analyses suggests adaptive introgression of those alleles that are associated with high expression levels. In both cases, the high expression phenotype represents the derived state and is associated with reduced levels of nucleotide polymorphism and a negative Tajima's D. For both genes, regulatory and protein-coding evolution is decoupled and the expression difference results from *cis*- rather than *trans*-acting changes.

Declaration

The project was initiated by Bettina Harr. In the course of the project I profited from the work of several other persons, whose contribution I want to acknowledge in the following.

Chapter 2

The inbred-strains of *M. m. musculus* and *M. m. domesticus* were provided by J. Pialek from the Department of Population Biology, Institute of Vertebrate Biology, Academy of Sciences of the Czech Republic; *M. m. castaneus* were donated by A. Orth and F. Bonhomme from the Laboratory of Genome, Populations, Interactions, and Adaptation in Montpellier, France. The Affymetrix GeneChip® experiments were conducted by P. Nürnberg and C. Becker from the Cologne Center for Genomics and data analysis was performed by Bettina Harr.

Chapter 3

Meike Teschke provided the German *M. m. domesticus* mice used for the experiments. Wild-caught *M. m. musculus* mice were provided by K. Musolf from the Konrad-Lorenz-Institute for Ethology in Vienna.

Chapter 4

The reference data set used for the hybrid zone analysis consists of samples provided by Bettina Harr, Sonja Ihle, Rick Scavetta, Meike Teschke and myself. Analysis of these data was performed by Meike Teschke. Population samples of *M. m. musculus* and *M. m. domesticus* used for sequencing analysis were collected and provided by Sonja Ihle, Rick Scavetta and Meike Teschke. Affymetrix GeneChip® experiments were conducted by P. Nürnberg and C. Becker from the Cologne Center for Genomics and analysis of the GeneChip® data was performed by Bettina Harr.

1 General Introduction

1.1 Species, speciation and the evolution of reproductive isolation

The definition of a species and the question of how speciation – the splitting of one species into two – takes place has been discussed intensively and controversial since the publishing of Darwin's "The origin of species" (1859), the book which set the basis for the Modern Synthesis. The founders of the Modern Synthesis substantially changed our understanding of what a species is and therefore what speciation means. Dobzhansky and Mayr in particular promoted the "Biological Species Concept" (BSC). Mayr provided the most famous definition of the BSC, which in its latest version states that "species are groups of interbreeding natural populations that are reproductively isolated from other such groups" (Mayr 1995). Thus, species are characterized by reproductive isolation instead of being defined by morphological differences. This species concept is in large parts a result of Dobzhansky's observation that hybrids between *Drosophila* sibling species display hybrid sterility, hybrid inviability and assortative mating (Dobzhansky 1937); an observation which laid the foundation for the formulation of his reproductive isolation species concept, which was later incorporated into Mayr's definition of a species (Mayr 1942). Reproductive isolation is associated with isolating barriers, "those biological features of organisms that impede the exchange of genes with members of other populations" (Coyne and Orr 2004, p. 29), which can be divided into premating isolation and postmating isolation, while the latter furthermore discriminates prezygotic and postzygotic barriers. All forms, which include various mechanisms, ensure the genetic distinctness of species and that they can undergo independent evolutionary fates (Orr and Presgraves 2000). Postzygotic isolation refers to sterile or inviable hybrids between two species, while the other forms prevent the occurrence of interspecies hybrids. The BSC has been widely understood as requiring absolute barriers to gene flow between taxa, i.e. no fertile hybrids exist (e.g. Mayr 1963, Barton and Hewitt 1985). Other scientists consider species as entities, which retain their distinctness in sympatry even if occasional hybridization takes place (Grant 1971, Wright 1978). Coyne and Orr (2004) suggest using a "sliding scale" to define the species status. Taxa with substantial gene flow (despite morphological distinctness) are not seen as species, but “as reproductive barriers become stronger, taxa become more and more “species-like”” until reproductive isolation is complete and taxa are “good species”. This requires, as they admit themselves, somewhat subjective decisions about the species status. However, the primary aim of speciation studies is to understand the underlying mechanisms of the speciation process. Therefore the question if a species is a “good species” may be of secondary interest.

Since the early 1980’s the interest in speciation studies has shifted towards a genetical perspective of how speciation can be understood (Orr and Presgraves 2000, Orr et al. 2004, Wu and Ting 2004, Mallet 2006). Most studies that are aimed at uncovering genes which are involved in the process of speciation deal with the analysis of intrinsic postzygotic reproductive isolation, i.e. the occurrence of sterile or inviable hybrids between two species. The term “speciation gene” has been widely applied for genes that cause intrinsic postzygotic reproductive isolation and is somewhat unfortunately named, since it strongly implies that genes that reduce hybrid fitness are the *cause* of speciation (Mallet 2006, Orr 2005). It has been suggested that reproductive isolation evolves as a by-product of differential adaptation of populations (Mayr 1963), which would mean that “speciation genes” are genes that represent aspects of differential adaptation (driven by natural or sexual selection (Albert and Schluter 2005)) which then reduce fitness of hybrids (Wu 2001). Orr and Presgraves (2000) suggest using the term “speciation gene” for *any* gene that reduces hybrid fitness, as many of the hybrid problems accumulate after the attainment of complete reproductive isolation, since it has been demonstrated that assortative mating, hybrid sterility and inviability increase gradually with genetic distance (Coyne and Orr 1997). Analysis of such genes nevertheless will help to shed light on the “mystery” of speciation.

How is intrinsic postzygotic isolation caused on a genic basis? Dobzhansky (1937) and Muller (1940) presented experimental evidence that sterility and inviability in certain crosses of species crosses is caused by incompatibilities between different loci. Today a large body of evidence has been obtained that shows that hybrid dysfunctions are indeed due to interactions of different loci (e.g. Orr and Coyne 1989, Wittbrodt et al. 1989, Gadeau et al. 1999, Presgraves 2003, Sweigart et al. 2007). The Dobzhansky-Muller model explains how such between-locus incompatibilities between populations can evolve within populations without selection acting against any intermediate step. Incompatibilities arise as pleiotropic by-products of the divergence of genomes of geographically separate lineages: alleles from different loci that increase fitness in a pure-species genetic background fail to interact properly when brought together in a hybrid genomic background (Figure 1.1). In each of the two populations displayed in Figure 1.1, a different mutation occurs and goes to fixation, each yielding a fully viable and fertile genotype. Brought together in a hybrid genome, there is no guarantee that the combination of these alleles functions correctly, as they have never been tested in combination by natural selection. The combination may result in sterility or inviability. A distinct feature of hybrid incompatibilities therefore is that they require epistasis: nonadditive interactions between alleles at different loci (Dobzhansky 1936A, 1937, Muller 1940, 1942). It is important to note that a two-locus incompatibility would be the simplest case: many more loci might be necessary to cause sterility/inviability. Furthermore a genic incompatibility may only have slight effects and complete reproductive isolation may require the cumulative effect of many incompatibilities (Coyne and Orr 2004).

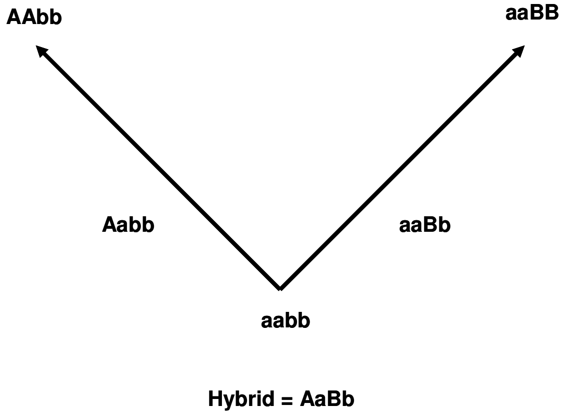


Figure 1.1: The Dobzhansky-Muller model. The common ancestor of the two species is shown at the bottom, time runs upwards. The model shows how a genic incompatibility between two loci can evolve unopposed by natural selection. Figure adapted from Coyne and Orr 2004.

To understand the genetic architecture of intrinsic postzygotic isolation, one has to identify the interacting genes that cause hybrid sterility or inviability. Despite much effort, only recently genes have been identified. The reason for the difficulties in identifying such genes is that "the traits of interest, hybrid sterility and inviability, are by their nature barriers to crossing and thus are refractory to standard genetic approaches" (Presgraves 2003). Only since actual genes have been identified, it is possible to address fundamental questions about the factors that cause reproductive isolation like "which normal function have these genes within species?", "do they typically belong to a specific class?", "are they rapidly evolving?", "are they adaptively evolving?", and so on (Orr et al. 2004). Four of the five identified hybrid incompatibility genes were found in *Drosophila (*see Orr et al. 2004, Wu and Ting 2004, Orr 2005 for a review and Masly et al. 2006). *OdsH*, an X-linked homeobox gene (a presumed transcription factor), is misexpressed in the testes of *D. simulans* x *D. mauritiana* hybrids and causes male sterility in backcrossed hybrids (Ting et al. 1998, Sun et al. 2004, Ting et al. 2004). *Hmr*, which encodes a transcriptional regulator of or related to the MYB family, causes male lethality and female sterility in *D. melanogaster* x *D. simulans* hybrids, likely due to disrupted gene regulation (Barbash et al. 2003). *Lhr* has been identified as an interacting partner, but the interaction of two genes alone is insufficient to cause hybrid lethality, additional genes seem to be involved (Brideau et al. 2006). The *D. simulans* allele of *Nup96*, a nuclear pore protein, causes lethality when combined with a hemizygous *D. melanogaster* X-chromosome (Presgraves 2003). *JYAlpha* (encoding a male fertility-essential Na^+-K^+ ATPase) was recently found to cause sterility in *D. melanogaster* x *D. simulans* hybrids (Masly et al. 2006). The only gene causing postzygotic isolation identified in a vertebrate so far is *Xmrk-2* in hybrids of *Xiphorus maculatus* and *X. helleri*. *Xmrk*-2 encodes a novel receptor tyrosine kinase, which is, while found as a duplicate gene on the *X. maculatus* X-chromosome, absent from the *X. helleri* X-chromosome (Malitschek et al. 1995, Wittbrodt et al. 1989, Schartl et al. 1999). *Xmrk*-2 is misexpressed in hybrids, causing tumor formation and ultimately lethality. What is missing to date is evidence that the identified "speciation genes" did play a role in the early stages of speciation or if incompatibilities have accumulated subsequent to the species split.

Genes involved in hybrid sterility and inviability are characterized by special dominance relations. There is good evidence that the genes causing reproductive isolation are, on average, partially recessive as predicted by the dominance theory of Haldane's rule (see for example Orr 1997, Coyne and Orr 2004). For hybrid inviability the evidence for recessivity is strong (Orr 1993, Presgraves 2003). In an extensive deficiency mapping experiment for hybrid inviability in *Drosophila*, it was shown that recessive-recessive interactions highly outnumber dominant ones (Presgraves 2003). Also for hybrid sterility evidence for recessive action of genes has been obtained (Orr 1992, 1997, Hollocher 1996, Sawamura 2000). Recessive speciation genes can obviously only contribute to the sterility and inviability of F2 or backcross hybrids (when genes have become homozygous), unless the incompatibilities involve X-linked genes. The dominance theory is now widely accepted to explain the phenomenon of Haldane's rule (Haldane 1922), the observation that the heterogametic sex suffers from more severe hybrid problems than the homogametic sex. As long as some fraction of the genes causing hybrid problems is recessive and also X-linked genes are involved, hybrid males should fare worse than hybrid females in cases for males are the heterogametic sex. They are afflicted by all X-linked genes involved in hybrid incompatibilities (dominant and recessive) whereas females are affected only by those that are fairly dominant (Coyne and Orr 2004).

Several conclusions can be drawn from the few "speciation genes" identified so far, although it is hard to generalize from a sample of that size. All these genes have normal functions within the species. No evidence has been obtained so far that reproductive isolation involves novel genetic factors like e.g. repetitive DNA sequences (Orr 2005). Furthermore, genes causing intrinsic postzygotic isolation seem not to belong to a single functional class. Although some of them are involved in transcriptional regulation (*OdsH*, *Hmr*) – and it remains possible that disruption of gene regulation is a common cause of hybrid problems – some are not. They are also not invariably duplicated genes; sometimes they are members of duplicate gene families (*Xmrk*-2, *OdsH*) sometimes they are single-copy genes. Additionally, speciation genes seem to be rapidly evolving (Orr 2005). This is what one would expect assuming that genes that cause disruptions in hybrids are those, which have diverged most between species. Even more important, it could be demonstrated for three of the genes that they diverge due to positive selection (Ting et al. 1998, Presgraves 2003, Barbash et al. 2004), which supports one of the central tenets of the neoDarwinian view of speciation – that reproductive isolation results from natural selection within species.

1.2 Hybrid incompatibilities associated with gene expression differences

Genome-wide expression profiling by microarrays provides new insights about mechanisms that lead to evolutionary changes (Ranz and Machado 2006). It has long been speculated that many species differences arise from regulatory differences (King and Wilson 1975) and despite the limited information concerning patterns, rates and mechanisms of change at the regulatory level it is at least obvious that changes in transcriptional regulation comprise a quantitatively and qualitatively significant component of the genetic basis for evolutionary change (Wray 2007). Theoretical and empirical studies give support to the hypothesis that failures in the regulation of gene expression may contribute to hybrid dysfunctions (Orr and Presgraves 2000, Ortíz-Barrientos et al. 2007). Such regulatory incompatibilities in hybrids arise from differences that have accumulated in the regulatory sequences between different lineages due to compensatory mutations in order to maintain constant levels of gene expression (Ranz et al. 2004). In a hybrid genetic background, where these alleles have not previously occurred together, such changes do not longer compensate and the interaction may produce novel expression patterns (Landry et al. 2007), i.e. over- or under-expression (misexpression). Many parameters affect the level of transcription, such as number of transcription factor binding sites, abundance of transcription factors and their affinity to binding sites and their interactions. Thus, there are multiple opportunities for the accumulation of changes between species. Such failures seem to be disproportionally greater in regulatory pathways containing rapidly evolving genes, particularly those involved in transcription factor-binding site interactions (Ortíz-Barrientos et al. 2007). Also, simulations by Johnson and Porter (2000) suggest that hybrid incompatibilities often arise as a consequence of divergence of regulatory genetic pathways between populations, due to adaptation processes. Empirically, it has been shown that three of the five genes associated with hybrid sterility and inviability appear to be related to regulatory changes (Malitschek et al. 1995, Ting et al. 1998, Barbash et al. 2003). Furthermore, several recent studies of genome-wide patterns of gene expression in pure species and sterile male hybrids suggest that hybrids show disruptions in gene expression (Michalak and Noor 2003, Ranz et al. 2004, Haerty and Singh 2006, Moehring et al. 2007, see Landry et al. 2007 for a review). Particularly, male biased genes have repeatedly shown to have erratic gene expression patterns in hybrids (Michalak and Noor 2003, Ranz et al. 2004, Haerty and Singh 2006). It has been shown that sex-biased genes, especially male-biased genes with testis specificity or relation to accessory gland proteins, evolve exceptionally rapidly at the expression level in *Drosophila* (Meiklejohn et al. 2003, Ranz et al. 2003, Nuzhdin et al. 2004). Therefore it can be expected that the regulation of such genes exhibits greater regulatory incompatibilities in hybrids than other genes. Vice versa, the observation that male-biased genes are disproportionally affected by gene expression disruption emphasizes the importance of rapid evolution in the divergence of species and the acquisition of reproductive isolation (Ranz et al. 2004).

Microarray analysis, as a reverse-genetics approach, offers an alternative or supplement to classical forward-genetics approaches in identifying genes which are involved in reproductive isolation. Genome-wide expression profiling can rapidly assay many genes in the genome simultaneously for their expression levels and identify potential genes and genetic pathways responsible for hybrids' misregulation. Recent analyses have shown that microarrays are highly instrumental in identifying regulatory differences between species and their hybrids (e.g. Michalak and Noor 2003, Ranz et al. 2004, Haerty and Singh 2006) and that many of the identified genes are likely to be involved in the generation of reproductive barriers (e.g. Michalak and Noor 2004). However, such distorted patterns of gene expression do not necessarily contribute to fitness reduction and reproductive isolation; they may also yield innovative phenotypes (Landry et al. 2007). Nevertheless, it seems likely that the accumulation of regulatory incompatibilities due to divergence and the architecture of transcriptional networks can directly influence the process of speciation (Porter and Johnson 2002, Johnson and Porter 2007).

1.3 The house mouse

The house mouse *Mus musculus* is one of the best studied model systems and particular well suited for evolutionary genetics studies. First, the complete genome sequence of a laboratory strain is available (Waterston et al. 2002) and second, various aspects like development, anatomy, pathology, behavior and ecology are intensively studied so that inferences between genomic and phenotypic patterns can be made (Galtier et al. 2004). Besides laboratory strains, wild-derived mouse strains can be kept in the laboratory (Guénet and Bonhomme 2003) allowing assessment of naturally occurring genetic variation.

The mouse phylogeny and colonization history is well documented (e.g. Boursot et al. 1993). The genus *Mus* is thought to have emerged ~ 3 million years ago (Guénet and Bonhomme 2003) and the house mouse *Mus musculus*, a polytypic species, split up into the different subspecies less than one million years ago (Figure 1.2). It is thought that these subspecies originated on the Indian subcontinent from which they radiated outwards (Din et al. 1996) and have now spread across the world. Best known and described are the three subspecies *M. m. musculus*, the central house mouse, *M. m. domesticus*, the western house mouse and *M. m. castaneus*, the eastern house mouse, which can be distinguished morphologically (Boursot et al. 1993). They have been described as subspecies because they are only partially reproductively isolated from each other and hybrid zones exist in regions of secondary contact. *M. m. musculus* is found in northern Asia as well as in Eastern Europe. *M. m. domesticus* is found in Western Europe and was introduced through human transport to Africa, America and Australia. The habitat of *M. m. castaneus* spans from Sri Lanka to South East Asia, including the Indian-Malayan archipelago (Figure 1.3). Due to competition in areas of sympatry with other rodents, house mice occur mainly in association with humans. Only in areas where other mouse species are absent they are able to exist in feral populations. *M. m. castaneus* is the only subspecies which lives exclusively commensal (Sage 1981, Boursot et al. 1993). Today's distribution of the house mouse is clearly connected to the migration of the humans over the world. America and Australia as well as many islands lacked mice before humans started to enter those regions. Since Europeans have extensively roamed the oceans, the western house mouse, *M. m. domesticus*, is particular widespread over the world, though long-distance passive transport of mice is not restricted to *M. m. domesticus* (Boursot et al. 1993).

The genome of the sequenced laboratory strain C57/BL6J is a mixture of the three subspecies, *M. m. domesticus*, *M. m. musculus* and *M. m. castaneus* (Wade et al. 2002, Yang et al. 2007). According to the most recent analysis, *M. m. domesticus* provides up to 92% of the genome, *musculus* <7% and *castaneus* <1% (Yang et al. 2007).

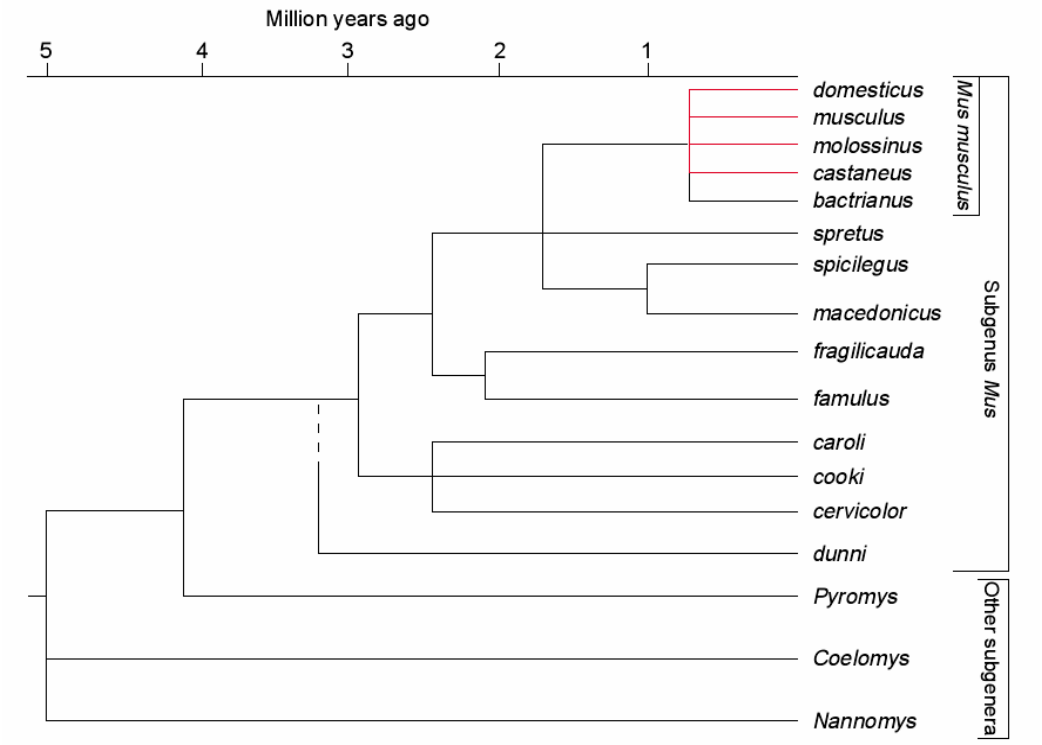


Figure 1.2: Evolutionary tree of the genus *Mus*. The last node refers to the polytypic species *Mus musculus* (Figure from Guénet and Bonhomme 2003).

The fact that the house mouse subspecies are not completely reproductively isolated from each other makes them ideal for speciation studies. To study the origin and the evolution of reproductive barriers, it is intuitive to focus on study systems that are not completely isolated. Thereby it can be discriminated between causes and consequences of the genetic isolation (Macholán et al. 2007). Several areas of secondary contact – hybrid zones – occur between the subspecies (Figure 1.3). *M. m. musculus* and *M. m. domesticus* meet in Europe, in the Caucasus, and in a region southeast of the Caspian Sea. For *M. m. musculus* and *M. m. castaneus* a contact zone in China and Japan has been described. In Japan, these two subspecies have hybridized extensively, giving rise to a unique population often referred to as *M. m. molossinus* (Yonekawa et al. 1988).

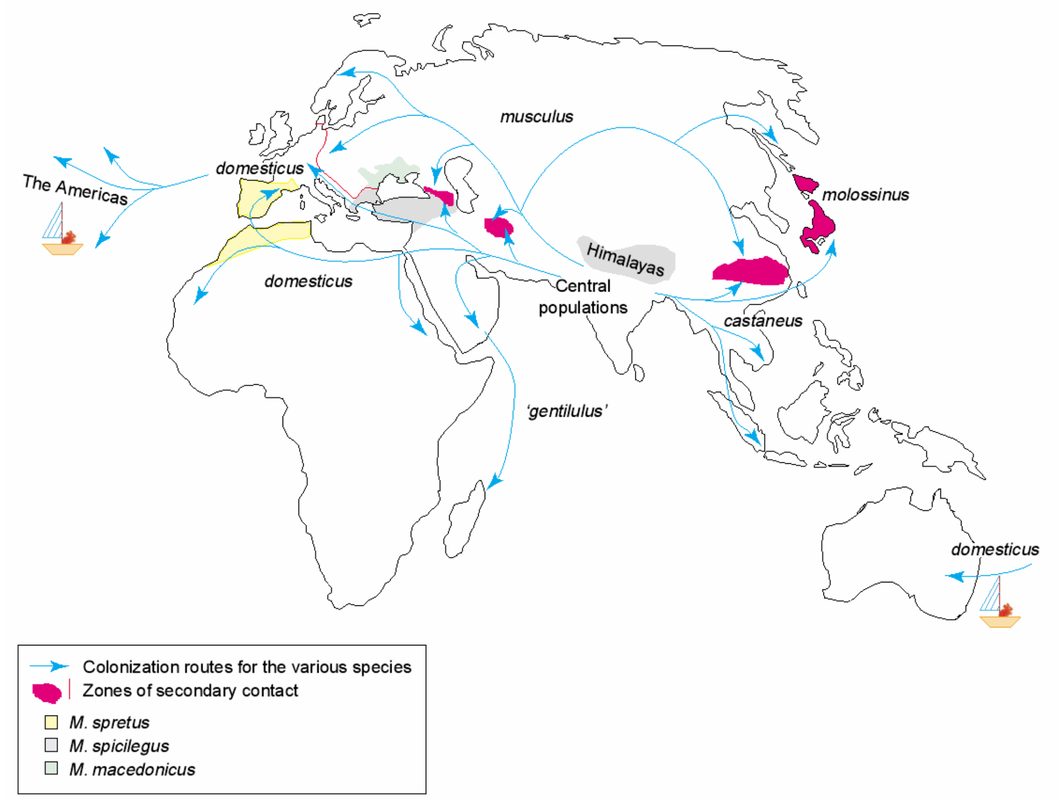


Figure 1.3: Geographical distribution of the different species of the genus *Mus* and routes of colonization. Mice of the American and Australian continents were imported by humans during colonization (Figure from Guénet and Bonhomme 2003).

The European hybrid zone between *domesticus* and *musculus* stretches from the Jutland peninsula to the Bulgarian coast of the Black sea (Boursot et al. 1993, Sage et al. 1993). It has formed as a consequence of the movement of *M. m. domesticus* into Western Europe within the last 3000 years (Cucchi et al. 2005). Several transects along the 2,400 km length of the hybrid zone have been studied (e.g. Hunt and Selander 1973, Sage et al. 1986, Tucker et al. 1992, Dod et al. 1993, 2005, Prager et al. 1993, Payseur et al. 2004, Raufaste et al. 2005, Macholán et al. 2007, Teeter et al. 2007), which has yielded insights into the genetic basis of reproductive isolation between the two subspecies. Heterogeneity in levels of introgression between genomic regions has been shown (Tucker et al. 1992, Boursot et al. 1993, Dod et al. 1993, Sage et al. 1993, Payseur et al. 2004); thus the genome of the house mouse seems to be "semipermable", with some regions tolerant of gene flow between species and others not (Payseur et al. 2004). Particular for the X chromosomes limited introgression was found (Tucker et al. 1992, Payseur et al. 2004, Macholán et al. 2007). Changes in allele frequency at diagnostic markers occur very rapidly relative to the geographic extent of the species ranges (Tucker et al. 1992, Dod et al. 1993, Payseur et al. 2004), which suggests that the hybrid zone is maintained by a balance between selection against hybrids and dispersal (Barton and Gale 1993). Secondary contact of the subspecies seems to have occurred recently (3000-10,000 years ago (Auffray et al. 1990, Cucchi et al. 2005)) in relation to the estimated divergence time, which is thought to be consistent with the accumulation of reproductive isolation in allopatry according to the Dobzhansky-Muller model (Payseur and Hoekstra 2005).

For hybrids between *musculus* and *domesticus* in the natural hybrid zone, elevated parasite loads, suggesting general reduced fitness, have been observed (Sage et al. 1986, Moulia et al. 1991, 1993) and reduced testis size has been documented (Britton-Davidian et al. 2005). Complementary to hybrid zone studies, several laboratory experiments with both inbred and wild outbred mice showed that *M. m. domesticus* and *M. m. musculus* are isolated by hybrid male sterility (Forejt and Ivanyi 1975, Alibert et al. 1997, Storchová et al. 2004, Britton-Davidian et al. 2005, Vyskocilová et al. 2005, Good et al. 2007). In some cases, crosses between the subspecies produce fully fertile hybrid males (Vanlerberghe et al. 1986) or variation in the strength of sterility occurs (Alibert et al. 1997, Britton-Davidian et al. 2005). Furthermore evidence exists that male F1 hybrid sterility is asymmetric, depending on the origin of the X chromosome (Britton-Davidian et al. 2005, Good et al. 2007). For some crosses there is also evidence for female sterility (Britton-Davidian et al. 2005). Another contributor to reproductive isolation between *musculus* and *domesticus* may be assortative mating. Mate choice experiments hint to behavioral differences between the subspecies which may be important in preventing hybridization in some crosses (Smadja and Ganem 2002, Ganem et al. 2005). The genetic basis of male hybrid sterility seems to be fairly complex (Good et al. 2007). Studies which involve classical inbred *domesticus* strains identified two sets of Dobzhansky-Muller incompatibilities between the two subspecies. One set of dominant-autosomal interactions, which include one or more tightly linked loci on chromosome 17 (*Hst1*, Forejt and Ivanyi 1975, Vyskocilová et al. 2005). The other set involves the interaction of the X and autosomes (Storchová et al. 2004, Britton-Davidian et al. 2005, Oka et al. 2007), suggesting a crucial role of the X chromosome in reproductive isolation between *M. m. domesticus* and *M. m. musculus*. Only recently, Good et al. (2007) discovered that at least a third set of Dobzhansky-Muller incompatibilities occurs in wild derived strains of *M. m. musculus* and *M. m. domesticus*. Hybrid zone analyses have identified multiple markers on the X-chromosome with reduced gene flow across the zone (Tucker et al. 1992; Dod et al. 1993; Payseur et al. 2004, Macholán et al. 2007). These regions probably harbor genes involved in reproductive isolation. Only recently, in a detailed genetic survey, Teeter et al. (2007) identified several regions in the genome which may be involved in reproductive isolation. They found genes associated with a variety of biological processes, including such as reproductive physiology, behavior and physiological and immune response. So far, no specific gene has been identified that contributes to reproductive isolation between house mouse subspecies. For hybrids between *M. m. domesticus* and *M. spretus* a strong hybrid sterility candidate locus has been identified. *Dnahc8*, an axonemal dynein protein expressed in the testis, has been mapped to the site of the hybrid sterility locus 6, which has been shown to be involved in hybrid sterility between the two species (Fossella et al. 2000).

1.4 Aim of the study

The study seeks to identify genes that are involved in postzygotic reproductive isolation in the house mouse *Mus musculus* with a genome wide expression screen. The house mouse is particularly suited for studying the early stages of speciation, since reproductive isolation is not complete yet and natural hybrids occur. I used a combination of two classical approaches: a survey of laboratory F1 hybrids between taxa and an analysis of animals from a natural hybrid zone. An Affymetrix GeneChip® expression analysis of three house mouse subspecies in comparison to their reciprocal F1 hybrids for three different tissues was conducted to identify candidate genes for regulatory hybrid incompatibilities. Transcripts with expression differences between the parental subspecies, and between the hybrids and their parents, respectively, are candidates which might be involved in reproductive isolation. After validation of candidate genes with enlarged datasets, a hybrid zone analysis was intended to infer if expression differences between parental subspecies and their hybrids are connected to a reduction in fitness, i.e. to be of functional consequence. Samples from a transect of the natural hybrid zone of *M. m. musculus* and *M. m. domesticus* in Bavaria were used to make inferences about the fitness effects of specific gene combinations in a hybrid genomic background by considering the introgression of “expression phenotypes” across the hybrid zone. Genes with negative fitness effects should not spread very far across the hybrid zone in comparison to neutral genes. Additionally, sequence based population genetic analyses for chosen candidate genes were applied to evaluate the genetic basis and evolutionary forces that might contribute to differences in gene expression.

2 Genome wide expression patterns in house mouse subspecies and their F1 hybrids

2.1 Introduction

Microarray-based gene expression profiling has become widely applied as tool to uncover the genetic architecture of quantitative traits in the past several years. Transcript levels are treated as phenotype and their intermediate position between DNA sequence polymorphism and organismal phenotype can be seen as “bridge” between the two in mapping studies (Rockman and Kruglyak 2006). The genetic correlation between expression levels and an organism’s phenotype points to underlying molecular pathways. A co-localization of QTLs for expression and the organism’s phenotype facilitate the identification of causal mutations. “Expression quantitative trait locus” mapping (eQTL) is therefore highly suited to answer basic questions about the number of loci underlying variation in expression phenotypes, the proportion of heritable variance and the genetic complexity of traits (Brem et al. 2002, Schadt et al. 2003, Morley et al. 2004, Brem and Kruglyak 2005, West et al. 2007). Despite of the usefulness in medical and agricultural genetics, uncovering general principles of the genetic architecture of expression traits will be helpful to reveal the basic forces responsible for phenotypic variation and evolution. In contrast to classical QTL mapping studies, eQTL mapping allows the assessment of thousands of traits simultaneously (Cui et al. 2006, Rockman and Kruglyak 2006). Therefore, a detailed description of possible architectures can be made and an unbiased set of traits in comparison to pre-selected single traits is provided. EQTL mapping studies can also be applied to assign variation in transcript abundance to differences in *cis*- respective *trans*-acting sites (Schadt et al. 2003, Yvert et al. 2003, Morley et al. 2004). Significant levels of *cis*-polymorphism, controlling individual genes have been detected as well as evidence for clustered *trans*-eQTLs that simultaneously regulate a large fraction of the transcriptome has been revealed in yeast, mice and humans. Recent studies indicate that expression levels are heritable and can be under multigenic control (Brem et al. 2002, Schadt et al. 2003, Brem and Kruglyak 2005). Despite much effort, the relationship between transcript level-variation and downstream organismal phenotypic trait variation is not well understood (Mackay

2001, Gibson and Weir 2005).

Two approaches are used to uncover the genetic basis of expression traits. One relates DNA polymorphism to differences in expression levels (Brem et al. 2002, Schadt et al. 2003, Doss et al. 2005, Storey et al. 2005). These studies revealed that expression traits are often affected by multiple underlying loci and interactions among them (Rockman and Kruglyak 2006). In yeast, the complex inheritance of transcript levels has been uncovered by detecting significant levels of nonadditive genetic variance, epistatic interactions and transgressive segregation (Brem and Kruglyak 2005). The second approach determines the within-locus mode of inheritance by comparing the transcript abundance of F1 hybrids to that of the parents and assesses if the expression is intermediate (additive) to that of the parents or not. A prevalence of nonadditively over additively expressed transcripts would suggest complex inheritance of expression traits. Recent studies treating this issue have come to inconclusive results concerning the prevalence of additivity and nonadditivity, respectively. Several studies found pervasive within-locus additivity of expression traits in F1 hybrids of laboratory mice, *Drosophila* and maize (Cui et al. 2006, Hughes et al. 2006, Stupar and Springer 2006, Swanson-Wagner et al. 2006). In contrast, two studies found most transcripts to be nonadditively expressed in *Drosophila* (Gibson et al. 2004) and the Pacific Oyster (Hedgecock et al. 2007), a third one reported similar numbers of additivity and non-additivity in *Arabidopsis* (Vulysteke et al. 2005), suggesting to consider epistasis as pervasive aspect of genetic architecture.

Comparative gene expression profiling of hybrids and their parents has also been applied to identify genes which have the potential to cause hybrid dysfunctions via regulatory incompatibilities. It is generally assumed that divergent evolution in two lineages leads to coevolution of genes and that crosses between species can reveal such coevolution (Landry et al. 2007). The formation of hybrids combines alleles that have not been previously occurred together and the interaction may generate new phenotypes. Uncovering the molecular basis of such nonadditive, epistatic interactions is central to the understanding of the evolution of hybrid incompatibilities (Dobzhansky-Muller incompatibilities (Dobzhansky 1937, Muller 1940). Theoretical and empirical studies suggest the importance of gene regulation for hybrid incompatibilities (Orr and Presgraves 2000, see Ortíz-Barrientos et al. 2007 for a review). Regulatory incompatibilities are defined as interactions of the transcriptional networks that lead to novel expression phenotypes in interspecific hybrids (Landry et al. 2007). Several studies in *Drosophila* show that disruptions in transcriptional regulation may indeed be associated with hybrid incompatibilities such as hybrid sterility (Michalak and Noor 2003, 2004, Ranz et al. 2004, Haerty and Singh 2006, Moehring et al. 2007). Michalak and Noor (2003, 2004) for example found several genes severely underexpressed in hybrids between *D. simulans* and *D. mauritiana*, which are associated with spermatogenesis and other male-specific phenotypes. Still in the 5th generation of backcross hybrids, sterility and misexpression of these transcripts was strongly correlated, which makes it possible that genes involved in spermatogenesis cause sterility in these hybrids. Interestingly, a large proportion of transcripts with disrupted expression in hybrids do not show expression level divergence between the parental species, suggesting that these genes are under stabilizing selection (Ranz et al. 2004, Haerty and Singh 2006). This finding underlines that species often accumulate divergent cryptic genetic changes in coding regions as well as regulatory regions that are revealed only if the two divergent regulatory architectures are brought together in a single individual (Moehring et al. 2007, Ortíz-Barrientos et al. 2007). If such disruptions in gene expression are the cause and not the consequence of hybrid incompatibilities, high-throughput "reverse genetics" approaches, like microarray analyses, have the potential to identify candidate genes or pathways that contribute to reproductive isolation via regulatory incompatibilities and thus to speciation (Noor and Feder 2006).

The present analysis was conducted to assess the within-locus mode of inheritance of expression differences between three subspecies of the house mouse, *M. m. domesticus*, *M. m. musculus* and *M. m. castaneus*. Differentially regulated transcripts between the subspecies and between the F1 hybrids and their parents, respectively, provide candidate genes, which can be involved in regulatory hybrid incompatibilities. Transcripts with expression levels in the hybrids significantly over- or under-expressed in comparison to the parents may be particular interesting with regard to hybrid incompatibilities. It has to be stressed that a hybrid incompatibility gene is any gene that contributes to reduced fitness in hybrids (Orr and Presgraves 2000).

2.2 Materials and Methods

2.2.1 Animals

Wild-derived strains of *M. m. domesticus, M. m. musculus* and *M. m. castaneus* were used for the microarray analysis. For *M. m. musculus*, strain JPC 2821 from the Czech Republic, for *M. m. domesticus*, strain JPC 2705 from Germany and for *M. m. castaneus*, strain CIM (from India), was used for the experiments. Animals from *domesticus* and *musculus* were collected in the wild and inbred by brother-sister matings for ~13 generations in the laboratory of J. Pialek in the Department of Population Biology in Studenec (Czech Republic). The *castaneus* strain was provided by A. Orth and F. Bonhomme and has been kept in the Laboratory of Genome, Populations, Interactions, Adaptation, Montpellier in France for more than 30 generations. Reciprocal crosses between the wild-derived-inbred subspecies were set up to obtain F1 hybrids (Table 2.1). *M. m. musculus* will be abbreviated as *mus, M. m. domesticus* as *dom* and *M. m. castaneus* as *cas* in figures and tables in the following.

Table 2.1: Breeding setup for F1 hybrids.

mother		father	F1 hybrid
M. m. musculus	x	*M. m. domesticus*	*mus-dom*
M. m. domesticus	x	*M. m. musculus*	*dom-mus*
M. m. castaneus	x	*M. m. musculus*	*cas-mus*
M. m. musculus	x	*M. m. castaneus*	*mus-cas*

Three tissues each (brain, liver, testis) of two males each cross and direction and each of the parental pure subspecies were used for the experiment (see Supplement 1). Altogether 42 microarray experiments were performed; 14 animals were analyzed for three tissues each. All mice were raised under identical standard laboratory conditions and were sacrificed at the age of 6-8 weeks.

2.2.2 Sample preparation

Animals were sacrificed using CO_2. Organs were excised and immediately snap frozen in liquid nitrogen. RNA was extracted using Trizol® (Invitrogen, Carlsbad, CA) following the manufacturer’s protocol. Quality and integrity of the total RNA was controlled by using the Agilent Technologies 2100 Bioanalyzer and the RNA 6000 Nano LabChip® Kit (Agilent Technologies Waldbronn, Germany). Only samples with RNA integrity numbers (RIN) >8.0 were used for analysis.

2.2.3 Microarray analysis

Expression profiles were determined for over 39,000 mouse transcripts using the Mouse Genome 430 2.0 Affymetrix GeneChip®. The biotin-labeled target synthesis started from 1 µg of total RNA following the Ambion Message Amp II aRNA amplification Kit protocol. After hybridization the GeneChips® were washed, using an Affymetrix GeneChip® fluidic station 450, stained with SA-PE (Streptavidin R-phycoerythrin conjugate) and read using an Affymetrix GCS 300 G scanner.

2.2.4 Data analysis

Data processing and statistical analyses were performed using the statistical language R. Raw signal intensities were normalized and summarized according to the standard Affymetrix MA Suite 5.0 algorithm using the program Bioconductor (http://www.bioconductor.org/). Signal intensities were ln-transformed prior to statistical analyses. MA Suite 5.0 expression values have been submitted to the Gene Expression Omnibus (accession number GSE9338). All analyzed samples were defined as "groups"; a "group" consisting of either *domesticus*, *musculus*, *castaneus*, F1 hybrids from one direction of the cross, or F1 hybrids from the other direction of the cross. Transcripts are defined as "expressed" in a group if the average expression level among the replicate samples within a group is >500. Applying a threshold of >500 yields about 11,000 expressed transcripts per group and tissue, a value that corresponds to previous studies (Su et al. 2004). The expression levels in the hybrids were determined as additive or nonadditive for two different groups:

1. transcripts that are differentially expressed between the parental subspecies (see Figure 2.1 a, b)
2. those transcripts, which do not differ in expression between the parental lines, but which are expressed differently in the F1 hybrids relative to both parents (Figure 1c):

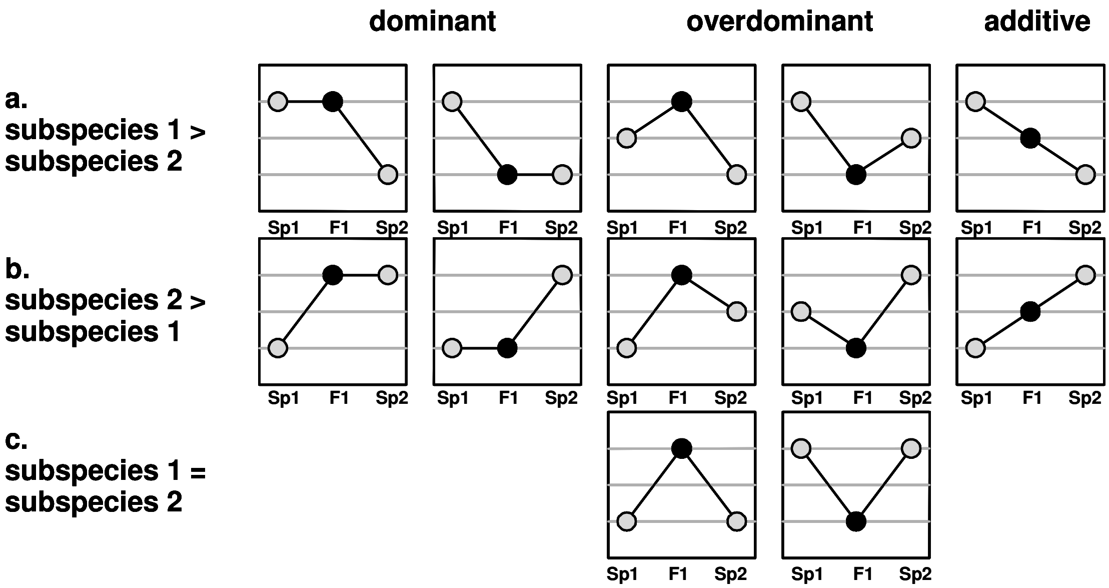


Figure 2.1: Schematic representation of within-locus modes of gene action in the comparison between house mouse subspecies and their F1 hybrids. Sp = subspecies, F1 = F1 hybrid.

Additivity/nonadditivity of transcripts differentially expressed between subspecies

To identify differentially expressed transcripts between each parental comparison (*domesticus* vs. *musculus* and *musculus* vs. *castaneus*), a standard One-Way ANOVA was performed for each tissue, applying a False Discovery Rate (FDR; (Reiner et al. 2003)) of <10%. A FDR of <10% corresponds to a threshold of p<0.0000053 in the brain, p<0.00025 in the liver and p<0.002 in the testis for the *domesticus-musculus* comparison. For the *musculus-castaneus* comparison this corresponds to p<0.002 in the testis and no transcript at all is significant in brain and liver under this criterion (see results). Since preliminary experiments using quantitative real-time-PCR of one gene differentially expressed between *musculus* and *domesticus* in the liver (Ces1) suggest that transcripts up to a p-value level of 0.06 on the microarray can be confirmed on independent samples (data not shown), the selection of differentially expressed genes for each tissue was repeated using arbitrarily significance values of p<0.01 and p<0.05.

Significant transcripts were considered separately according to whether they were expressed (>500) in both subspecies or just in one of the subspecies within a given comparison. Since expression levels <500 likely include transcripts that are not expressed at all, additivity of transcript expression cannot be revealed in these cases.

However, if the transcript is absent in one parental subspecies but present in the other, it can at least be determined whether the expression status in the hybrid is consistent with dominance or not.

Three methods were used to determine the mode of inheritance in the hybrids for the transcripts differentially expressed between the parental subspecies for all tissues separately. The average expression level across replicates of the subspecies with lower expression is defined as x, the average expression level of the subspecies with higher expression as y, and the expression in the hybrid as h. The analysis was performed with untransformed signal intensity values for all three methods.

Fixed threshold method

Following Gibson et al. (2004), additivity is defined as $y / 1.25 > h > x * 1.25$, for all h. Dominance is defined as $1.25 * x > h > x / 1.25$ or $1.25 * y > h > y / 1.25$ for all h, and overdominance as $h < x / 1.25$ or $h > y * 1.25$ for all h. Using this model, only transcripts with an expression difference of >2.5 can show additivity and analysis is restricted to only these transcripts.

The fixed threshold method yields the problem that as the fold difference between two subspecies increases, the zone where additivity will be accepted also becomes larger. To correct for that bias, a second method was applied.

Fractional threshold method

This method is based on the relative difference in expression levels between the parental subspecies. Additivity is defined as $(x+y)/2 - \varepsilon < h < (x+y)/2 + \varepsilon$, for all h, where ε was set to 5%, 10%, 15% and 20% of the difference between x and y (e.g., $\varepsilon = (y-x)*0.15$). Dominance was set to be $x - \varepsilon/2 < h < x + \varepsilon/2$ or $y - \varepsilon/2 < h < y + \varepsilon/2$ for all h, so that for each transcript the interval over which either dominance or additivity is accepted is equivalent. Overdominance was set to be outside of the range of dominance, i.e., $h < x - \varepsilon/2$ or $h > y + \varepsilon/2$ for all h. Only genes with an expression difference of >2.5 were included in the analysis to make the fractional threshold method comparable with the fixed threshold method.

For both methods the mode of inheritance in F1 hybrids was determined by

1. treating F1 hybrids from both directions of the cross together (i.e., all four replicates had to be consistent) and

2. treating both directions of the cross separately to account for possible maternal

effects, sex-linkage and directional dominance.

Distribution of dominance effects

Additionally to the classification of transcripts into arbitrary mode-of-inheritance classes, the distribution of dominance effects was determined for both comparisons (*domesticus* vs. *musculus* and *musculus* vs. *castaneus*). It was calculated as “d/a”, where “a” is half the difference in expression level between the parental strains and “d” is the difference in expression level between the F1 hybrid (average across replicates independent of the direction of cross) and the average of the parental strains (Falconer and Mackay 1996; Gibson et al. 2004). If the expression level of the transcript in the hybrid is exactly intermediate between the parental strains d = 0. A d/a value of 0 corresponds to perfect within-locus additivity, |d/a| = 1 to complete dominance and |d/a|>1 to overdominance. All differentially expressed transcripts between the parental strains (p<0.01 and p<0.05), with average expression levels >500 in at least one of the parental subspecies, were considered. In contrast to the analyses where transcripts are assigned to discrete additive and nonadditive classes, it is not required that the differentially expressed transcript shows a fold-change >2.5 between the parental subspecies.

Nonadditive expression of transcripts not differentially expressed in the subspecies

Pairwise Tukey post hoc tests were used to assess parent-F1 hybrid offspring expression differences for the transcripts, which are not differentially expressed between the parental subspecies. Only those transcripts were selected in which F1 hybrids have significant (p<0.05) and 1.25-fold higher or lower expression compared to both parents. This analysis was performed for F1 hybrids from both directions of the cross together as well as from each direction of the cross separately. For transcripts with expression levels <500 in both of the parental subspecies, only those crosses were analyzed in which the F1 hybrids from at least one direction of the cross had average expression levels >500.

Contribution of sequence divergence to expression difference: Determination of subspecies origin of each probe set

The probe sets of the Mouse Genome 430 2.0 Affymetrix GeneChip® are designed based on the sequence of the laboratory inbred strain C57BL/6J. The genome of this strain represents a mixture of genetic contributions from three subspecies *M. m.*

musculus, *M. m. domesticus* and *M. m. castaneus* (Wade et al. 2002, Yang et al. 2007), whereas *M. m. domesticus* provides up to 92% of the strain C57BL/6J sequence, *musculus* <7% and *castaneus* <1% (Yang et al. 2007). Therefore it is possible that sequences differences (i.e. SNPs, Single Nucleotide Polymorphisms) between the wild subspecies translate into differences in hybridization efficiency (elevated or decreased signal intensities) of specific probe sets; i.e. one can expect that a probe set hybridizes more efficiently with that subspecies the particular probe set sequence was derived from. Although sequence divergence between the subspecies is generally low (typically 1% for non-coding DNA at autosomal loci (Harr 2006)) it could potentially enhance or alleviate expression differences. This problem is probably most serious if other subspecies are compared to *domesticus*, since *domesticus* contributed most of the genetic material to C57BL/6J. Therefore this problem was specifically investigated with respect to the *domesticus-musculus* comparison. To get an estimate about the contribution of such subspecies sequence differences to expression data, probe sets were assigned to a "subspecies" status and analyses were performed considering this status. More than eight million SNP loci distributed all over the genome were downloaded from the Perlegene website (http://mouse.perlegen.com/mouse/download.html). These SNPs have been typed in 15 commonly used inbred strains of the laboratory house mouse. The data from these strains can be combined with the known sequence of strain C57BL/6J (Waterston et al. 2002) at the respective SNP position. Two out of the 15 strains are so-called "wild-derived" strains, one of which belongs to *musculus* (PWD/PhJ) and one to *domesticus* (WSB/EiJ). SNPs that distinguish the *musculus* from the *domesticus* strain were used to assign the subspecies origin of each probe set in the C57BL/6J strain. The frequency of *musculus*-like or *domesticus*-like SNPs in a 20 kb region surrounding the location of the transcript that is targeted by a specific probe set was calculated. A probe set was called *musculus*-like if >60% of the SNPs in the region matched the *musculus* strain and *domesticus*-like if >60% of the SNPs matched the *domesticus* strain.

Data were divided into two sets:

1. transcripts that are located in "*musculus*-like" regions but show higher expression in *M. m. domesticus* and
2. transcripts which are located in "*domesticus*-like" regions but show higher

expression in *M. m. musculus*.

For these transcripts, sequence divergence is unlikely to explain the expression differences, thus representing a measure for the contribution of sequence divergence to differences in signal intensities. Because no wild-derived strain of *castaneus* is available from the Perlegene dataset, genomic regions in the genome of C57BL/6J that stem from *castaneus* could not be assigned. Since only a small proportion of the genome originates from *castaneus*, one can assume this limitation to be not serious.

Functional annotation

"PANTHER" (http://www.pantherdb.org/tools/genexAnalysis.jsp) was used to find overrepresentation of biological functions relative to the full gene content of the house mouse genome. This analysis was conducted for transcripts differentially expressed between the two subspecies (fold change >2.5, p<0.05), identified by the fractional method (15%). Analysis was performed for two groups:

1. transcripts additively expressed in F1 hybrids from both directions of the cross
2. transcripts differentially expressed between the parents independent of expression status in F1 hybrids.

2.3 Results

2.3.1 Number of differentially expressed genes

About half of all the transcripts (20,811) represented on the Affymetrix GeneChip® have signal intensities >500 and are therefore called “expressed” in at least one of the three tissues in at least one of the groups (*domesticus*, *musculus*, *castaneus*, *dom-mus*, *mus-dom*, *cas-mus*, *mus-cas*), ca. 1/4 (12,130) is called expressed in at least one of the tissues in at least one of the three subspecies. Transcript numbers are quite similar among tissues with brain 13,584, liver 11,637 and testis 12,130 expressed transcripts in at least one of the subspecies. Comparing the number of expressed transcripts among the three subspecies reveals almost identical values for the three different tissues (Table 2.2).

Table 2.2: Number of transcripts showing signal intensities >500 in the different subspecies for brain, liver and testis.

tissue	***M. m. domesticus***	***M. m. musculus***	***M. m. castaneus***
brain	12,142	12,248	12,183
liver	10,083	10,243	10,101
testis	10,552	10,480	10,774

Table 2.3 shows transcripts differentially expressed between *M. m. domesticus* and *M. m. musculus* and between *M. m. musculus* and *M. m. castaneus* for different significance levels. At significance level of p<0.05 and considering all transcripts irrespective of the fold change, approximately similar numbers of differentially expressed transcripts in each of the three organs are found. Assuming an average of ~12,000 transcripts expressed in at least one subspecies and tissue, 600 should be assigned significantly different by chance at p<0.05. Instead roughly three times as many significant different transcripts in each tissue occur. The number of transcripts differentially expressed between the subspecies drops noticeably when only transcripts with a larger than 2-fold change are considered. For brain 445, liver 683 and testis 847 transcripts are differentially expressed between the *domesticus* and *musculus*. Between *musculus* and *castaneus*, 435 transcripts are differentially expressed in the brain, 691 in liver and 827 in testis. But even for 5-fold expression

difference between the subspecies, a relatively large number of significantly different transcripts can be identified.

Table 2.3: Number of transcripts differentially expressed between the parental subspecies *M. m. domesticus* and *M. m. musculus*, and *M. m. castaneus* and *M. m. musculus* at different ANOVA p-value thresholds and at different magnitudes of change.

tissue	**Magnitude of change**	**ANOVA p<0.05**	**ANOVA p<0.01**	**ANOVA p<10% FDR**	**ANOVA p<0.05**	**ANOVA p<0.01**	**ANOVA p<10% FDR**
		domesticus* vs. *musculus			***castaneus* vs. *musculus***		
brain	all	1943	573	1	1819	516	0
	2-fold	445	208	0	435	192	0
	2.5-fold	274	146	0	281	137	0
	5-fold	96	63	0	98	54	0
liver	all	1777	490	9	1808	463	0
	2-fold	683	249	7	691	228	0
	2.5-fold	465	182	6	464	164	0
	5-fold	179	90	3	185	85	0
testis	all	2563	882	218	2514	822	215
	2-fold	847	431	128	827	401	131
	2.5-fold	568	355	104	546	295	106
	5-fold	218	151	55	197	132	61

Differences between the tissues are apparent if the significance level is reduced (Figure 2.2); for p equal to the False Discovery Rate of 10%, a strong overrepresentation of differentially expressed transcripts in the testis is apparent. While only one transcript is differentially expressed in the brain and nine in the liver, 218 transcripts show up different in the testis between *domesticus* and *musculus*. For the *castaneus-musculus* comparison significant different transcripts were only found in the testis (213).

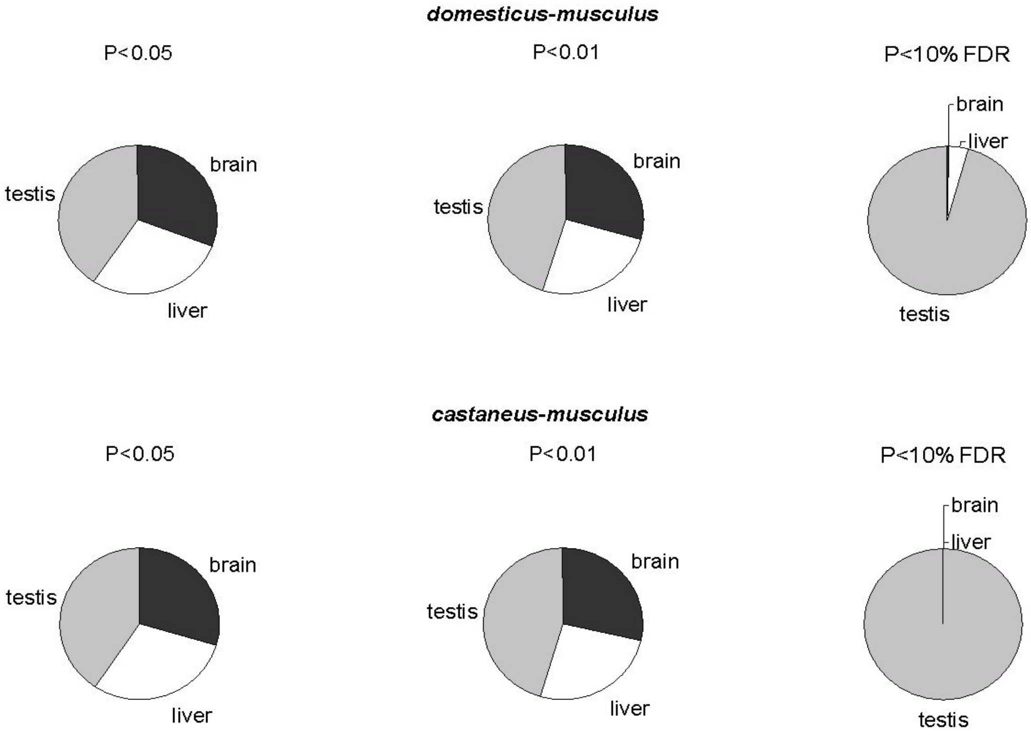


Figure 2.2: Number of all differentially expressed transcripts between *M. m. domesticus* and *M. m. musculus*, and *M. m. castaneus* and *M. m. musculus* for different tissues and significance levels.

Most of the transcripts are differentially expressed in just one tissue within a comparison (Figure 2.3), even though about half of the differentially expressed transcripts in each tissue are detectably expressed in at least one additional tissue. This means that a large proportion of differentially expressed genes are not tissue specific (data not shown).

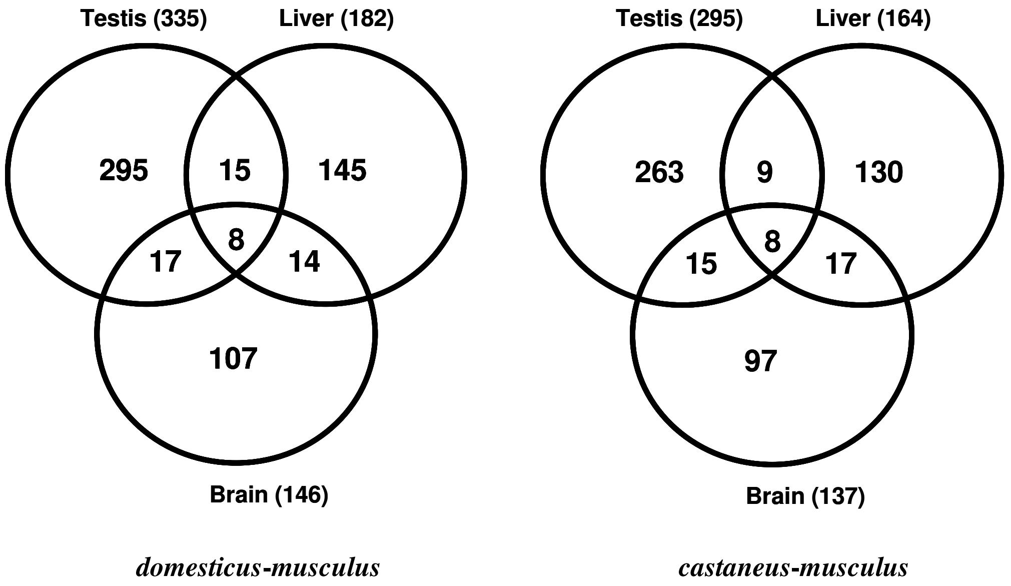


Figure 2.3: Overlap of transcripts differentially expressed between parental subspecies (p<0.01, fold-change >2.5) among the different tissues.

Figure 2.4 shows the overlap of differentially expressed transcripts between the subspecies comparisons (i.e. *domesticus-musculus* and *castaneus-musculus*) for the three tissues. Interestingly, between one third and one half (depending on the tissue) of the transcripts differentially expressed between *M. m. domesticus* and *M. m. musculus* are also differentially expressed between *M. m. castaneus* and *M. m. musculus*. This suggests that the expression phenotype of these transcripts is *musculus*-specific.

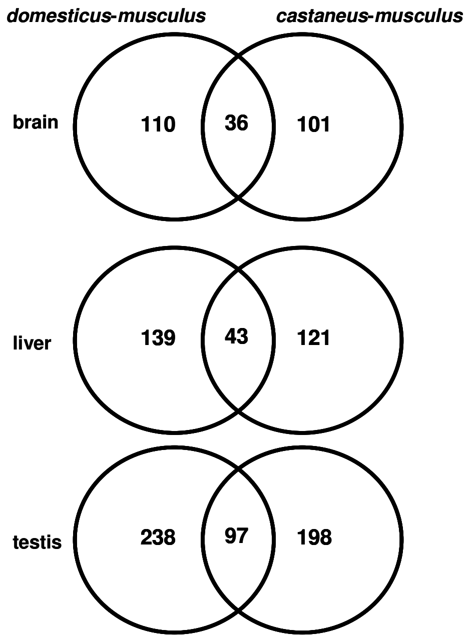


Figure 2.4: Overlap of differentially expressed transcripts between *M. m. domesticus* and *M. m. musculus*, and *M. m. castaneus* and *M. m. musculus* within a tissue (p<0.01, fold-change >2.5 for each comparison).

2.3.2 Additivity/nonadditivity of transcripts differentially expressed between the subspecies

Three different methods were used to infer the mode of inheritance (i.e. additivity, dominance, and overdominance) for the analyzed transcripts.

Fixed/fractional threshold method

Table 2.4 and Table 2.5 show the number of transcripts identified for each class for the fixed threshold method and the fractional threshold method (ε corresponds to 15% of the parental difference in expression). Numbers are shown separately for transcripts with expression levels of >500 in both parental strains and those with an expression level of >500 in one of the two parents. For this group, additive expression in the F1 hybrid can not be unambiguously determined, since expression level <500 may mean no expression at all in one of the parental subspecies. Therefore, those putatively additively expressed transcripts are depicted in brackets in the tables.

Table 2.4: Number of additively, dominantly and overdominantly expressed transcripts among the differentially expressed transcripts between *M. m. musculus* and *M. m. domesticus* in different tissues (fold-change >2.5, ANOVA p<0.05). Numbers in brackets indicate that additive expression of a transcript cannot unequivocally be identified as the transcript has signal intensities <500 in one of the subspecies and may therefore be not expressed.

	Additive			**Dominant**			**Overdominant**		
	dom-mus	***mus-dom***	***mus-dom* AND *dom-mus***	***dom-mus***	***mus-dom***	***mus-dom* AND *dom-mus***	***dom-mus***	***mus-dom***	***mus-dom* AND *dom-mus***
Fixed threshold method, expressed transcripts in *domesticus* AND *musculus*									
Brain	39	41	31	5	5	1	0	0	0
Liver	104	113	84	17	10	1	2	1	0
Testis	97	104	80	11	10	4	1	0	0
Fixed threshold, expressed transcripts in *domesticus* OR *musculus*									
Brain	(112)	(108)	(96)	2	14	0	0	1	0
Liver	(137)	(135)	(112)	13	8	2	1	2	0
Testis	(276)	(278)	(244)	13	15	4	0	2	0
Fractional threshold (15%), expressed transcripts in *domesticus* AND *musculus*									
Brain	27	20	13	0	0	0	0	0	0
Liver	57	56	24	6	2	0	2	1	0
Testis	58	66	39	0	1	0	1	0	0
Fractional threshold (15%), expressed transcripts in *domesticus* OR *musculus*									
Brain	(67)	(60)	(34)	3	3	0	0	0	0
Liver	(74)	(63)	(36)	2	3	0	3	1	1
Testis	(148)	(146)	(80)	0	1	0	1	2	0

Dominant and overdominantly expressed transcripts are rare; most transcripts show additivity in expression. In general, fewer transcripts that can be classified in either of the modes of inheritance are observed with the fractional threshold method.

Nevertheless, the fraction of additive versus nonadditive effects is almost identical for both methods, and changes little with varying values of ε in the fractional method (data not shown). In all cases, transcripts were classified into the same mode-of-inheritance categories by both methods. The relative frequency of additive effects is independent of whether it was identified in F1 hybrids from one or both directions of the cross combined. Furthermore the percentage of transcripts showing additivity was similar across tissues: more than 80% of the transcripts are additively expressed and <10% dominantly and overdominantly, respectively. One exception is present: in the *musculus*-mother x *castaneus*-father cross solely for the testis, a much higher fraction of nonadditively expressed transcripts occurs. Between 30% and 65% of the assigned transcripts are either dominantly or overdominantly expressed, depending on the method.

Table 2.5: Number of additively, dominantly and overdominantly expressed transcripts among the differentially expressed transcripts between *M. m. castaneus* and *M. m. musculus* in different tissues (fold-change >2.5, ANOVA p<0.05). Numbers in brackets indicate that additive expression of a transcript cannot unequivocally be identified as the transcript has signal intensities <500 in one of the subspecies and may therefore be not expressed.

	Additive			**Dominant**			**Overdominant**		
	cas-mus	***mus-cas***	***cas-mus*** **AND** ***mus-cas***	***cas-mus***	***mus-cas***	***cas-mus*** **AND** ***mus-cas***	***cas-mus***	***mus-cas***	***cas-mus*** **AND** ***mus-cas***
Fixed threshold method, expressed transcripts in *castaneus* AND *musculus*									
Brain	36	38	25	5	2	0	0	1	0
Liver	100	95	70	15	14	5	0	3	0
Testis	76	55	36	12	24	4	1	11	1
Fixed threshold, expressed transcripts in *castaneus* OR *musculus*									
Brain	(199)	(179)	(151)	15	17	3	0	1	0
Liver	(301)	(282)	(226)	37	35	10	0	3	0
Testis	(388)	(264)	(214)	30	88	7	2	26	1
Fractional threshold (15%), expressed transcripts in *castaneus* AND *musculus*									
Brain	15	25	4	0	1	0	0	1	0
Liver	28	33	13	1	2	0	1	4	1
Testis	34	17	8	4	9	0	1	22	1
Fractional threshold (15%), expressed transcripts in *castaneus* OR *musculus*									
Brain	(104)	(89)	(47)	1	2	0	1	2	0
Liver	(96)	(100)	(42)	6	10	0	9	1	1
Testis	(192)	(65)	(34)	8	32	0	2	55	1

The analysis of transcripts with signal intensities >500 in only one of the parental subspecies yields similar results. Dominance and overdominance are rare and most transcripts show additivity in expression. This large fraction of transcripts with putatively intermediate expression suggests that transcripts with expression levels

<500 are more likely very low expressed rather than absent.

Distribution of dominance effects

Also the analysis which considers the distribution of dominance values shows a preponderance of additive gene expression. Most of the d/a values fall into the -0.5 to +0.5 interval (Figure 2.5). This analysis considers all transcripts which are differentially expressed at the given p-values, not only those with >2.5 fold difference between the parental subspecies (like for the fixed and fractional threshold method).

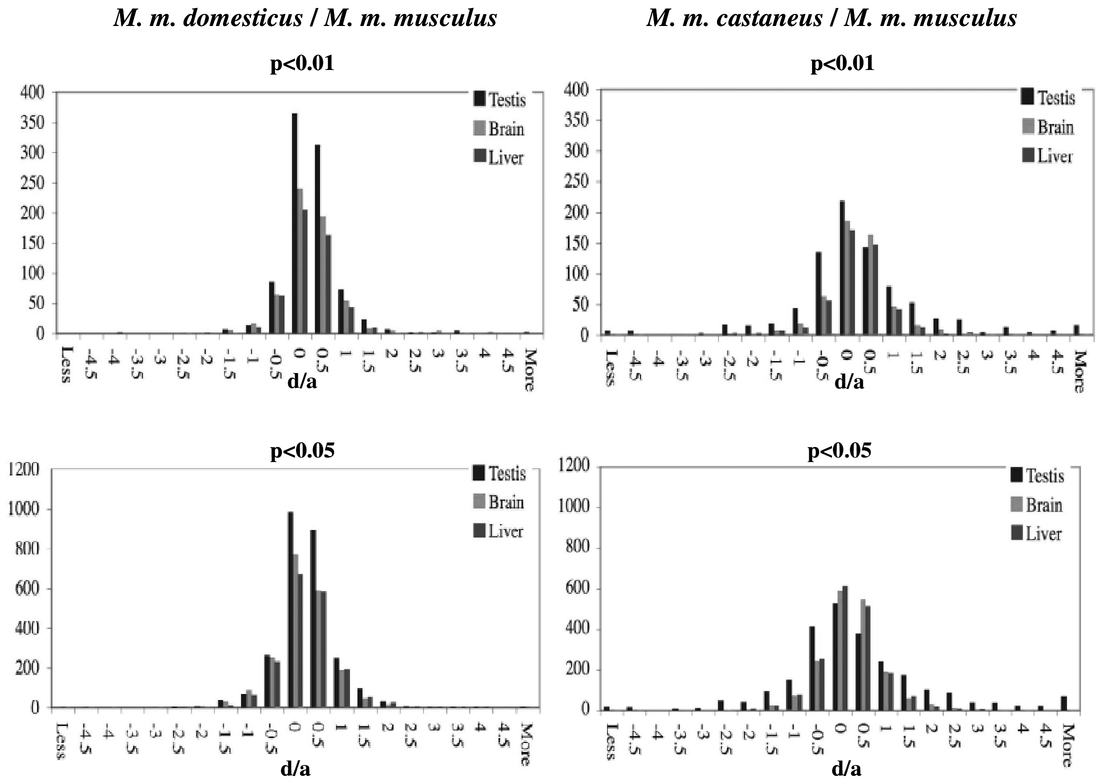


Figure 2.5: Distribution of dominance values in F1 hybrids for differentially expressed transcripts between *M. m. domesticus and M. m. musculus*, and *M. m. castaneus* and *M. m. musculus* for ANOVA p value of <0.01 and p<0.05, respectively. A d value of 0 indicates exact intermediacy of expression.

2.3.3 Nonadditive expression in F1 hybrids at transcripts that are not differentially expressed between the parents

This analysis considers all transcripts that did not show a significant difference in expression between the parental subspecies. To be included in the analysis the transcripts had to have signal intensities >500 in at least one of the classes (i.e. one of the parental subspecies or at least one of the cross directions). As an average of 11,000 transcripts is called expressed in each tissue and ca. 2,000 of them are differentially expressed between the two subspecies at p<0.05, ~ 9,000 transcripts per tissue are included in this analysis. Only very few transcripts in F1 hybrids are found to differ significantly (p<0.05) at least 1.25 fold from the parental subspecies (Table 2.6), less than 0.05%. Thus, also this analysis shows that nonadditive transcript expression is rare except again for the *musculus*-mother x *castaneus*-father cross in the testis, where ca. 1/3 of the transcripts show significant non-additivity.

Table 2.6: Number of transcripts nonadditively expressed in F1 hybrids compared to the parental subspecies.

	dom-mus	***mus-dom***	***dom-mus* AND *mus-dom***	***mus-cas***	***cas-mus***	***cas-mus* AND *mus-cas***
Brain	7	19	1	23	3	2
Liver	23	44	5	15	33	7
Testis	15	17	3	2798	12	16

Summarizing, all analyses yield the same result: additivity is most common and non-additivity is rare, with the exception of one cross in the testis (*musculus*-mother x *castaneus*-father).

2.3.4 Contribution of sequence divergence to expression change

If sequence divergence between *M. m. domesticus* and *M. m. musculus* had the potential to influence the microarray result, one would expect that probe sets hybridize more efficiently to *domesticus* in all regions with sequence differences, since the Affymetrix probe sets are based predominantly on *domesticus* DNA (C57BL6/J strain). To elucidate this potential problem, differentially expressed transcripts were counted separately for cases where *musculus* shows higher expression than *domesticus* and vice versa (Table 2.7). At low significance levels, no difference between the numbers is apparent. For high significance values, more transcripts in *domesticus* show higher expression than in *musculus*. This suggests that sequence divergence may influence the detection of differentially expressed transcripts.

Table 2.7: Number of transcripts differentially expressed between the parental subspecies, *M. m. musculus* and *M. m. domesticus* at different p-value thresholds and at different magnitudes of change, separately for the cases where *domesticus* shows higher expression than *musculus* (*domesticus* > *musculus*) and vice versa (*musculus* > *domesticus*).

tissue	Magnitude of change	ANOVA p<0.05		ANOVA p<0.01		ANOVA p<10% FDR	
		mus>dom	*dom>mus*	*mus>dom*	*dom>mus*	*mus>dom*	*dom>mus*
brain	all	956	987	267	306	1	0
	2-fold	159	286	69	139	0	0
	2.5-fold	90	184	45	101	0	0
	5-fold	29	67	21	42	0	0
liver	all	913	864	253	237	3	6
	2-fold	315	369	112	137	2	5
	2.5-fold	210	255	98	66	2	4
	5-fold	81	98	42	48	2	1
testis	all	1269	1294	417	465	87	131
	2-fold	331	516	164	267	41	87
	2.5-fold	209	360	124	211	33	71
	5-fold	76	142	56	95	17	38

To test for this potential bias, each probe set was assigned as *domesticus*- or *musculus*-like and analysis was restricted to two sets of transcripts. The first consists of differentially expressed transcripts which have a *musculus*-like probe set but high expression in *domesticus*, and the other with *domesticus*-like probe set but high expression in *musculus*. For both cases, expression differences are highly unlikely explained by sequence divergence. Table 2.8 shows additivity, dominance and overdominance for this subset of transcripts. The result is similar to the analysis including all transcripts: additivity is the predominant mode of in heritance.

Table 2.8. Number of additively, dominantly and overdominantly expressed transcripts among the differentially expressed transcripts between *M. m. musculus* and *M. m. domesticus* (fold-change >2.5, ANOVA p <0.05) in different tissues, considering the "*musculus*"/"*domesticus*"-likeness of the probe set. Numbers in brackets indicate that additive expression of a transcript cannot unequivocally be identified as the transcript is not expressed in one of the subspecies.

	Additive			**Dominant**			**Overdominant**		
	dom-mus	***mus-dom***	***mus-dom* AND *dom-mus***	***dom-mus***	***mus-dom***	***mus-dom* AND *dom-mus***	***dom-mus***	***mus-dom***	***mus-dom* AND *dom-mus***
Fixed threshold method, expressed transcripts in *domesticus* AND *musculus*									
Brain	12	13	11	1	3	1	0	0	0
Liver	45	45	39	9	3	0	0	0	0
Testis	39	39	33	2	3	2	0	0	0
Fixed threshold, expressed transcripts in *domesticus* OR *musculus*									
Brain	(67)	(71)	(60)	2	6	1	0	0	0
Liver	(127)	(123)	(102)	22	12	4	0	2	0
Testis	(165)	(166)	(142)	6	12	4	1	1	0
Fractional threshold (15%), expressed transcripts in *domesticus* AND *musculus*									
Brain	8	10	7	0	0	0	0	0	0
Liver	24	17	8	4	1	0	0	0	0
Testis	20	24	13	0	0	0	1	0	0
Fractional threshold (15%), expressed transcripts in *domesticus* OR *musculus*									
Brain	(40)	(45)	(22)	1	1	0	0	0	0
Liver	(52)	(41)	(20)	8	4	1	3	4	2
Testis	(70)	(80)	(38)	0	0	0	1	1	0

The comparison of *castaneus* and *musculus* can be considered as a second test to see if sequence divergence influences the outcome of the analysis. Both subspecies are more closely related to each other than they are to *M. m. domesticus* (Prager et al. 1998) therefore they are equally diverged from the subspecies that provides most of the genome of C57BL/6J. Thus, the analysis should not be severely affected by sequence differences. The result of this comparison (Table 2.5) shows that most transcripts are additively expressed (with the exception of the *musculus*-mother x *castaneus*-father cross in testis).

Summarizing, the surplus of additively expressed genes in hybrids of subspecies of the house mouse seems not to be substantially affected by sequence divergence between them.

Transcripts differentially expressed between the parental subspecies were analyzed with respect to maternal or paternal effects. Therefore both directions of the cross were analyzed separately and only those transcripts were considered that were dominantly expressed in both directions of the cross. If the expression level in the F1 hybrids always matched that of the mother this was classified as maternal effect. The expression level of a F1 hybrid for a paternal-effect transcript had to match that of the father. This analysis was restricted to transcripts that are not sex-linked. One region on chromosome 7 (~59 Mb, NCBI Build 36, gene Snrnp) seems to show a paternal effect in the brain of F1 hybrids from both the *musculus-domesticus* and the *musculus-castaneus* cross. The *musculus-domesticus* cross shows the same paternal effect also in the liver. No maternal-effect transcripts were found.

2.3.5 Functional categories of transcripts

For the transcripts additively expressed as assessed by the fractional threshold method ($p<0.05$, fold change >2.5, $\varepsilon = 15\%$), no overrepresentation of certain biological processes was found in brain and testis. For liver, metabolic functions were significantly overrepresented in *domesticus-musculus* hybrids, for the *castaneus-musculus* hybrids, oxygenase and acetyltransferase were overrepresented (after Bonferroni correction).

For transcripts differentially expressed between the two subspecies ($p<0.05$, fold change >2.5) independent of the expression status in the F1 hybrids (i.e. all differentially expressed transcripts are considered), significant overrepresentation of certain biological processes are found in all three tissues and in both comparisons of the subspecies (Supplement 2, Supplement 3). For the *domesticus-musculus* comparison, similar biological processes are overrepresented in brain and testis (transcripts involved in intracellular protein traffic); in liver a high preponderance of transcripts involved in lipid, fatty acid and steroid metabolism is apparent (Supplement 2). Similar patterns are obvious in the liver for the *castaneus-musculus* comparison. For brain and testis different patterns of overrepresentation are found, like carbohydrate metabolism in brain and detoxification in testis (Supplement 3).

The analysis for overrepresentation of certain biological processes was also conducted for the large number of transcripts nonadditively expressed in the *musculus*-mother x *castaneus*-father cross (Supplement 4) in the testis. Various processes are overrepresented. Interestingly, “spermatogenesis and motility” and “gametogenesis” are overrepresented with high significance values ($p = 5 \times 10^{-6}$).

2.4 Discussion

The expression levels in subspecies of the house mouse and their reciprocal F1 hybrids were assayed in three different tissues (brain, liver and testis). These tissues were chosen as representatives of different aspects of an organism's phenotype. Voolstra et al. (2007) suggested that expression differences in the brain may reflect behavioral differences, changes in the liver general metabolic differences and changes in testis differences in reproduction. For all analyzed tissues and both subspecies-cross comparisons (with the exception of the *musculus*-mother x *castaneus*-father cross), the analysis of within-locus mode of inheritance shows substantial overrepresentation of genes additively expressed in F1 hybrids in comparison to the parental subspecies, if transcripts with a significant expression difference between the parental subspecies are considered. Three different methods were applied to infer the mode of inheritance and all of them yielded the same result. The fixed threshold method has the tendency to shift the results towards detecting more additively expressed transcripts if the fold change between the parental subspecies becomes higher; the fractional method in contrast yields the potential to assign more overdominance with decreasing fold changes between the parental subspecies. Since both methods yield the same results although they would affect the results in opposite directions and the third method, which assesses the distribution mode of inheritance classes, is also consistent with a preponderance of additivity, it is likely that the pattern is real rather than a methodological or technical effect. Furthermore, a control experiment was performed to ensure that observed effects do not result from sequence divergence between the subspecies, since it is known that there is potential of sequence differences to result in lower hybridization efficiency (Chismar et al. 2002, Gilad et al. 2005, Ji et al. 2004). Also this experiment confirmed that additivity is the prevalent mode of inheritance. In addition, for genes showing no expression difference between the parental subspecies, only a negligible proportion of transcripts show nonadditive gene expression in hybrids.

Functional annotation of the additively expressed transcripts did not reveal an overrepresentation of certain functional categories. While there is an overrepresentation of metabolic functions for additively expressed transcripts in the liver, the same overrepresentation is also found if all transcripts that are differentially expressed between the parents are considered. This suggests that this overrepresentation is a liver-specific phenomenon rather than attributed to additivity.

Studies which have addressed the question of additivity/nonadditivity of gene expression systematically come to inconclusive results. Two found evidence for substantial nonadditivity in gene expression (Gibson et al. 2004, Hedgecock et al. 2007), four found pervasive additivity (Cui et al. 2006, Hughes et al. 2006, Stupar and Springer 2006, Swanson-Wagner et al. 2006) and one similar numbers of additively and dominantly expressed transcripts (Vuylsteke et al. 2005). Hughes et al. (2006) suggested two explanations for the discrepancy of results of their *Drosophila* study to those of Gibson et al. (2004). One reason could be that differences occur because of inbreeding. Natural populations of *D. melanogaster* (Hughes et al. 2006) show a preponderance of additivity while strongly inbred lines (Gibson et al. 2004) show nonadditivity to be prevailing. It seems plausible to relate overdominant or dominant gene expression to heterosis, a phenomenon which describes the better performance of crosses between inbred strains with regard to e.g. biomass, speed of development and fertility than both parents (Comings and MacMurray 2000). Heterosis is most commonly attributed to dominance (masking of recessive deleterious alleles in the heterozygous) or overdominance (superiority of heterozygous at loci affecting fitness). However, this seems not to be a general explanation for the disagreement of the studies, since other studies also used highly inbred lines despite observing nonadditivity (Cui et al. 2006, Stupar and Springer 2006, Swanson-Wagner et al. 2006). Secondly, it might be that the genetic architecture depends in the taxonomic level at which the variation is investigated. Hughes et al. (2006) assume that the more diverged the parents are the greater the nonadditivity in gene expression becomes. While this might to some extend contribute, it does not solely explain differences, since within-species studies have documented frequent nonadditivity (Gibson et al. 2004, Hedgecock et al. 2007, Vuylsteke et al. 2005). However, it seems equally likely that, as taxa diverge and a trait becomes influenced by more and more genes, additivity of expression becomes more and more prevalent, provided that no direct dominance occurs. This would apply up to some certain threshold after which incompatible alleles lead to massive misexpression in F1 hybrids.

The most plausible explanation for the differences in genetic architecture between the different studies seems to be the analysis method of gene expression, since several methodological differences are relevant. One critical point is that the definition of additivity depends on the measurement scale, such that on certain scales a transcripts' expression level might be different from others. Indeed, the different studies apply various measurement scales. To minimize this effect in the present analysis, three different analysis methods have been used, which gave similar results. While only a small number of replicates was used in the current study, the result was also very robust with respect to changes in p-values and fold change levels (data not shown). Thus, the preponderance of additive gene action in the house mouse seems convincing and not due to measurement artifacts. A second critical point may be difference in the used biological material; some studies use single organs while others pool different tissues by using whole organisms. That differences in tissues indeed might have an effect is illustrated by the study of Ranz et al. (2004), who found a high proportion of nonadditivity in *D. melanogaster* x *D. simulans* hybrids when comparing the whole body of the flies and more additivity if only the head was surveyed. However, the differences in used tissue are no general explanation since also Hughes et al. (2006) used the whole body and found additivity to be prevalent. A third factor is that studies differ due to differences in the technical setup. The most critical point here is the used microarray platform. Recent studies comparing the performance of platforms come to inconsistent results. Patterson et al. (2006) do not find differences in performance among platforms while other studies conclude that one-color microarray generally outperform two color-microarrays (De Reynies et al. 2006, Kuo et al. 2006).

Supplement 5 list studies that have systematically assayed patterns of inheritance in gene expression plus four additional studies, which have less rigorously treated this issue but allow a qualitative evaluation. Neither the level of inbreeding nor differences in divergence times are systematically associated with the frequency of additivity, although all might contribute to the differences in the mode of inheritance. The most striking correlate of additivity is the analysis platform used. None of the four studies, which have found a preponderance of nonadditivity, used Affymetrix microarrays, but two-color microarrays or sequence based expression measures. Therefore, methodological difference may be a major factor responsible for different results concerning genome-wide inheritance patterns. This would imply that additivity may be a general attribute of most (but not all) genes differentially expressed between taxa and that divergence level, degree of inbreeding and most probably also the complexity of tissues is no major contributor.

The screen of expression patterns of inter-subspecies hybrids was conducted to identify appropriate candidate genes, which might be involved in reproductive isolation. Based on the theory of regulatory incompatibilities in hybrids (see Landry et al. 2007 for a review), strongly misregulated genes, i.e. over- or undertranscribed in comparison to both parents (Moehring et al. 2007), are most promising since those genes suggest regulatory failures in the F1 hybrid generation. The analysis of expression patterns did reveal only few misregulated genes. This result is independent of whether the parents differ in expression or not (1 to 16 genes, if both hybrid combinations have to differ from both parents, depending on tissue and comparison). Only one exception occurs that is the cross of *musculus*-mother x *castaneus*-father, for which a relative large proportion of nonadditively expressed genes have been observed (~30% of the ~9,000 expressed transcripts that do not differ between the parents and 10-20% (depending on method) of the ~300 transcripts that show a difference between the parents).

Studies examining gene expression to uncover regulatory incompatibilities in hybrids and their parents have mainly focused on *Drosophila* crosses (Michalak and Noor 2003, 2004, Ranz et al. 2004, Haerty and Singh 2006, Moehring et al. 2007). These studies find at least in part large proportions of misexpressed genes (5-65% of all expressed transcripts) in the hybrids in comparison to the parents; especially male-biased genes are disproportionally misexpressed (Michalak and Noor 2003, 2004, Ranz et al. 2004, Haerty and Singh 2006). Specifically, these studies suggest that a significant fraction of the divergence in gene expression is cryptic, with more differences present between hybrids and the parents at the regulatory level as predicted from the examination of parental phenotypes only (True and Haag 2001). Therefore, the relative low number of misexpressed genes in house mouse hybrids is rather unexpected. However, several factors have to be considered when comparing such studies and thinking about regulatory incompatibilities in general. First, microarrays merely score a new phenotype – the amount of transcripts in individuals – and the connection of this trait to genotype and the final phenotype (like e.g. hybrid sterility) is unclear (Ortíz-Barrientos et al. 2007). A large difference in expression (or many differentially expressed genes) does not necessarily indicate that theses genes at the same time cause problems in hybrids. Similarly, some hybrid dysfunctions may not be associated with any expression difference. Furthermore, the number of incompatibilities, even if many misexpressed genes have been identified, is not automatically large. Numerous changes could be due to few major regulatory genes whose effects cascade through the gene expression (Landry et al. 2007). Therefore, the few identified over- or underexpressed transcripts in the house mouse hybrids may be major contributors to F1 hybrid unfitness.

An important difference between the studies in *Drosophila* and the present study is that *Drosophila* between-species hybrids display sterility, respectively that they do not hybridize in nature. Therefore it seems intuitive to find misregulated genes with effect on hybrid fitness, especially those related to reproduction. In contrast, the reproductive status of house mouse hybrids is not clear; fertility of the used crosses has not been tested. Male hybrid sterility has been demonstrated in natural hybrids and in laboratory crosses between both inbred and wild outbred mice (Forejt and Ivanyi 1975, Storchová et al. 2004, Britton-Davidian et al. 2005, Vyskocilová et al. 2005, Good et al. 2007), whereas crosses with other laboratory strains yield fully fertile offspring or display variation in strength of sterility (Vanlerberghe et al. 1986, Forejt 1996, Alibert et al. 1997, Britton-Davidian et al. 2005). Thus, reproductive isolation through hybrid sterility frequently occurs but is not complete. Furthermore, different hybrid sterility genes segregate within geographical separate populations. This indicates that hybrid male sterility emerges frequently in *M. musculus* populations and that the genetic basis of reproductive isolation depends on the population and even the individual tested (Vyskocilova et al. 2005, Good et al. 2007). Thus, the lack of misregulated genes in most of the surveyed F1 hybrids might mean that they are not sterile and also relatively fit with regard to other potential hybrid disadvantages.

House mouse subspecies are less divergent and display different geographical settings than the *Drosophila* species. It has been assumed that the homogenizing effect of gene flow between taxa prevents or delays the accumulation of regulatory incompatibilities, since the populations are forced to evolve towards the same compensatory genotypes (Porter and Johnson 2002). The ranges of either *domesticus* and *musculus,* and *castaneus* and *musculus* overlap and gene flow exists. As a consequence, the number of incompatibilities should be smaller than for strictly separated taxa (Hughes et al. 2006). However, for one of the analyzed crosses a large number of misexpressed genes could be observed despite existing gene flow between the subspecies. Therefore, this explanation is unlikely of general relevance for the present data.

Not only genes displaying misexpression but also intermediate expression levels in hybrids might be linked to hybrid incompatibilities as such expression levels can already be detrimental in F1 hybrids. Additionally, such genes have the potential to be important isolating factors in F2 and later backcross generations, which has been observed for some house mouse crosses (Oka et al. 2007). The heterozygous state of alleles at interacting loci might not be sufficient to cause hybrid dysfunctions and effects might occur only when homozygous allele combinations come to interact in later generations (Orr and Presgraves 2000). It has been suggested that alleles involved in hybrid incompatibilities are, on average, partially recessive (Orr 1997), which means that these genes lower fitness in hybrids far more when homozygous or hemizygous than heterozygous. Empirical evidence for this “dominance theory” has been obtained from a *Drosophila* deficiency mapping study, where recessive interactions were found to vastly outnumber dominant ones (Presgraves 2003).

One cross (*musculus*-mother x *castaneus*-father) displayed intensive nonadditivity in gene expression in the testis, i.e. many genes are misexpressed in the hybrid in comparison to both parents. This pervasive misexpression is restricted to the testis, which rules out methodological problems to have caused this pattern. Interestingly, biological processes related to spermatogenesis and gametogenesis are significantly overrepresented among these genes. Thus, a link of misexpression in this cross to hybrid sterility seems plausible. The high number of misexpressed genes could mean that many genes are involved in sterility as well as that the effects of a few master control genes have cascaded through the regulatory network and influenced the expression of many genes (Landry et al. 2007). The extensive misexpression in the testis could represent the crossing of a certain threshold at which genic incompatibilities between taxa result in low hybrid fitness. Regarding the theory that incompatibilities affect fertility before viability (Coyne and Orr 2004) the observed pattern in this cross is conceivable.

Whether reproductive isolation between subspecies of the house mouse is related to a fast rate of divergence in gene expression in reproductive organs (i.e. testis), as it is suggested by some models of sexual selection and sexual conflict (Rice 1998), has been previously tested by Voolstra et al. (2007) by comparing the expression patterns of different tissues and with respect to different divergence levels (i.e. between species and between subspecies). They found only few genes differentially expressed in the testis between the subspecies. In contrast, the present study, observes most differentially expressed transcripts in the testis especially at high significance thresholds. One reason for the different outcomes of the studies may be methodological. Voolstra et al. (2007) used two color microarrays and unrelated wild-caught animals, which have been exposed to standard laboratory conditions for only few days, whereas the current analysis was performed with Affymetrix GeneChips® and laboratory-raised, to a large extend inbred strains of each subspecies, thus excluding natural variation to influence gene expression. A second factor, which can explain the differences, is that hybrid sterility affects genes that segregate within populations. Several hybrid sterility loci within populations but without fixed differences between populations have been found in crosses between a laboratory strain (mainly *domesticus* origin) and wild *M. m. musculus* (Vyskocilova et al. 2005), between wild-derived inbred strains of *M. m. musculus* and *M. m. domesticus* (Good et al. 2007) and in *Mimulus* species (Sweigart et al. 2007). Thus, if the different wild-caught animals in Voolstra et al.'s (2007) study were of different genotypes, the consequent within-subspecies variability would obscure between-subspecies differences, while in inbred-animals such within-subspecies differences can be neglected.

At this point of the analysis it is unclear, whether any of the identified genes (misexpressed or differentially expressed between the parental subspecies) contributes to reproductive isolation and which reasons account for the differences between the analyses in *Drosophila* and the present study. However, the identified genes provide valuable candidate genes for further analyses addressing different aspects of expression, phenotypic and population genetics analyses that may shed light on the genetic basis of reproductive isolation between house mouse subspecies.

3 Confirmation of microarray candidate genes: factors affecting expression data in cross-subspecies analyses

3.1 Introduction

In the past several years, the technique of DNA microarrays has opened the possibility to move apart from “one gene in one experiment” analyses towards surveying the genome of an organism as a whole, thus giving the opportunity of unraveling interactions of genes and get a more complex picture of an organisms’ organization. With DNA microarrays one can conduct large-scale comparative gene expression profiles between biological samples by simultaneously surveying expression in thousands of genes. However, as they are very sensitive to various artifacts, a follow-up confirmation of observed effects is strongly recommended (Chuaqui et al. 2002, Miron et al. 2006, Morey et al. 2006). Several factors may influence results from microarray studies. One critical factor is that the relationship between probe sequence, target concentration and signal intensity currently is not completely understood (Draghici et al. 2006). As probes are designed as perfect complement to a specific region of a transcript, the probe should capture a specific number of these transcripts and thus the signal intensity of a probe should be proportional to the concentration of the transcript. The number of molecules which are bound to the probe depends to a large degree on the sequence affinity of the probes; affinity being a consequence of the actual sequence stretch participating in the binding. Recent studies showed that it is very hard to predict correct DNA-RNA binding affinities: studies conducted in solution have shown that a single base-mismatch in a probe can stabilize or destabilize the RNA-DNA depending on the identity of mismatch, its position in the helix and its neighboring base pairs (Kierzek et al. 1999).

A recent study by Pozhitkov et al. (2006) has determined the effect of the type of mismatch and type of neighboring nucleotide on signal intensity on DNA arrays and found, that both factors had significant effects on the normalized intensity values as well as that there are interactions among the two factors. Thus, signal intensities can be confounded by all these factors. Furthermore Pozhitkov et al. (2007) stated that “… there is little evidence supporting the notion that the known thermodynamic parameters accurately predict signal intensity values of duplexes on oligonucleotide DNA arrays. This makes it highly questionable if the current thermodynamic parameters which are applied to design probe oligonucleotides are useful", respectively it is likely that any analyses reveal a proportion of false positive results.

Another important contributor to confounded expression data is alternative splicing. A recent study has estimated that about 45% of the mouse genome is alternatively spliced (Modrek and Lee 2003) and therefore a major factor potentially influencing the data. If a probe set only refers to one exon, expression changes occurring only in certain splice variants may be obscured by detecting all splice variants (if the probes are located in an exon present in all the alternative transcripts) or in the reverse case (if the probes are located in an exon which is only present in certain splice variants), the specific transcript will be measured but the others will be ignored.

Cross-hybridization (probe sequence is not strictly complementary to the target sequence) also seems to be a significant contributor to skewed signal intensities. It has been shown that numerous probes produce cross-hybridization signals both in cDNA and oligonucleotide microarrays (Wu et al. 2005, Zhang et al. 2005) whereas even short stretches of sequence complementary seem to be sufficient to make hybridization of unrelated sequences possible. Evaluation of how significant such effects may be is not easy as for example for Affymetrix arrays single cross-hybridizations of probes can be out-weighted by the remaining probes of a probe set. Cross-hybridization also depends on the abundance as well as on the affinities of the specific and the non-specific target (non-specific target must be present in sufficient concentrations to influence the true signal).

Furthermore expression and copy number variation can be confounded. Evidence has been presented that increased copy number can be positively correlated with gene expression levels (McCarroll et al. 2006). The relative abundance of a specific transcript can be overestimated due to duplications that increase the potential for the cross-hybridization of highly related sequences to specific probes in the microarray.

A factor which has rarely been considered yet is, if the comparison of closely related taxa (and therefore existence of sequence differences) may be critical, if one of the taxa shows higher divergence from the sequence represented on the microarray than the other. For cross-species analyses, using species different from the Chip sequence, there is a general agreement that the array sensitivity and thus the accuracy of the analysis decreases with increasing sequence differences between the species analyzed, which implies that cross-species analyses yield significantly more false-negatives genes that appear not to be expressed although they are, than same-species analyses (Chismar et al. 2002, Ji et al. 2004). Chismar et al. (2002) for example hybridized *Macaca* to a human GeneChip® and had 50% less detected transcripts than for human samples, while Ji and coworkers found overall low hybridization signals for cattle, pig and dog when hybridized to a human GeneChip® (Ji et al. 2004). Enard et al. (2002) compared the transcriptome of humans, chimpanzees and orangutans to a human GeneChip® and three mouse species to a mouse GeneChip®. Both comparisons revealed that the expression levels were reduced for that species the chip was not designed for. Although different methods have been developed to account for this type of bias in cross-species analyses (e.g. Khaitovich et al. 2004, Ranz et al. 2003), overall, most studies so far have not had an effective way to estimate or correct for the effect of sequence mismatches on array hybridization. The study of Gilad et al. (2005) again affirmed the effect of sequence divergence on expression levels, even for very close related species such as human-chimpanzee, by applying a multispecies cDNA array and stated that "…sequence divergence… cannot be safely ignored in direct cross-species comparisons". It is however questionable, if sequence divergence as less as between house mouse subspecies, which diverged from each other 0.5 to 0.8 million years ago (Guénet and Bonhomme 2003), has a substantial effect on the accuracy of expression analysis.

All the numerous parameters which potentially influence microarray data show the necessity to confirm any obtained candidate gene with an independent method. Additionally to the need to check for such artifacts, it is important to determine whether the observed expression profile is a general biological feature of the sample under study rather than a property of the sampled entity. Therefore it is necessary to critically evaluate a larger and more extensive sample set with an independent method.

Quantitative real-time-PCR (qRT-PCR) is the choice of many for quantitatively measuring specific mRNAs as, once established, the method is rapid, relatively inexpensive and requires minimal starting template (Walker 2002). However, data achieved from microarrays and qRT-PCR often show inconsistencies, as both methods have their own pitfalls. To date no golden standard for validation of microarray data exists (Morey et al. 2006), thus choosing a method strongly depends on the scientific question. In literature a wide range of the correlation between microarray data and qRT-PCR data exists (Morey et al. 2006). For genes with high fold changes (i.e. >2) accordance has found to be high (Etienne et al. 2004). Most common, genes for validation are chosen regarding high expression differences between two samples as it is expected that high fold changes are more likely of biological significance.

Here, I evaluate different factors which may confound microarray data or results from quantitative real-time-PCR if comparing different subspecies and tissues of house mouse subspecies. From an expression screen, several genes with high fold changes between the two house mouse subspecies *M. m. domesticus* and *M. m. musculus* were chosen for validation via qRT-PCR. Both TaqMan® Gene Expression assays and Sybr Green was used for the confirmation approach and compared for their utility. Sequence comparisons were conducted to evaluate the contribution of subspecies sequence differences for both microarray and follow-up qRT-PCR analyses.

3.2 Materials and Methods

3.2.1 Candidate Genes

Gene expression analysis with Affymetrix GeneChip® 430.2 (analysis described in detail in 2.2.4) yielded 431 genes differentially expressed in testis between *M. m. musculus* and *M. m. domesticus* (ANOVA p-value <0.01, fold change minimum between subspecies = 2). From these genes a set of ten genes was chosen (Table 3.1) for confirmation with quantitative real-time-PCR involving a larger sample set. Those genes showed a minimum average fold change between subspecies of 2.5 with intermediate average expression levels of both hybrids (i.e. additively expressed, see 2.2.4).

Table 3.1: Candidate genes used for the confirmation experiment and fold changes between *M. m. domesticus* and *M. m. musculus*. Positive fold changes refer to *domesticus* being higher expressed, negative to *musculus* being higher expressed.

Affymetrix ID	Gene symbol	Refseq	Average fold change from GeneChip®	Expression difference in
1460235_at	Scarb2	NM_007644.2	-19.8566	testis
1421113_at	Pga5	NM_021453.2	-11.0281	testis
1424060_at	Neil3	NM_146208.1	7.242553	testis
1453203_at	1700011K15Rik	NM_029294.1	20.19957	testis
1455939_x_at	Srp14	NM_009273.4	23.41926	brain, liver, testis
1419715_at	1700029F12Rik	NM_025585.1	207.2982	testis
1429661_at	Rhobtb3	NM_028493.1	12.97842	testis
1428437_at	Lsm14a	NM_025948	-7.35	brain, liver, testis
1417515_at	Lsm10	NM_138721	-7.81	testis
1452877_at	2700029M09Rik	XM_910498	-7.79	testis

3.2.2 Animals

The confirmation of the GeneChip® results was done with the same animals used in the microarray study as well as a larger set of wild-derived inbred mice (hybrids and parents) and wild-caught *M. m. domesticus* and *M. m. musculus*. For *M. m. musculus*, strain JPC 2821 from the Czech Republic and for *M. m. domesticus*, strain JPC 2705 from Germany was used for the experiments. JPC 2821 and JPC 2705 were collected in the wild and inbred by brother-sister matings for ~13 generations in the laboratory of J. Pialek in the Department of Population Biology in Studenec (Czech Republic). Reciprocal crosses between the wild-derived-inbred subspecies were set up to obtain F1 hybrids (Table 3.2). *M. m. musculus* will be abbreviated as "*mus*" and *M. m. domesticus* as "*dom*" in figures and tables in the following.

Table 3.2: Breeding setup for F1 hybrids.

mother		father	F1 hybrid
M. m. musculus	x	*M. m. domesticus*	*mus-dom*
M. m. domesticus	x	*M. m. musculus*	*dom-mus*

Altogether seven inbred-males for *M. m. domesticus* and four for *M. m. musculus*, six *mus-dom* hybrids and 3 *dom-mus* hybrids were included in the analysis. Additional, four pairs each of wild-caught animals of *M. m. domesticus* from Germany (provided by M. Teschke) and *M. m. musculus* from Austria (provided by K. Musolf, Vienna) were used to obtain F1 offspring in the laboratory (at least one male offspring per locality) which were also included in the analysis (Supplement 1). F1 offspring was used instead of directly caught animals in order to exclude variation in gene expression due to differences in age, health condition, food resources etc. As it is known that mice live in small family groups with home ranges restricted to 2 km respectively strictly indoor living mice even not moving more than a few square meters (Berry and Bronson 1992, Pocock et al. 2005) a specific sampling scheme was applied to get mice from different demes and to ensure unrelatedness of sampled mice. Trapped mice were assigned to different localities only if the trapping sites were at least 300 m apart from each other. Animals were captured in live traps and transferred to the laboratory. Breeding pairs derived from the same locality were set up to obtain F1 offspring. All mice were held under standard laboratory conditions and males were sacrificed at the age of 6-8 weeks.

3.2.3 Sample preparation

Animals were sacrificed using CO_2. Organs were excised and immediately snap frozen in liquid nitrogen. Organs were stored until RNA extraction at -80 degrees. RNA from testis was extracted using Trizol® (Invitrogen, Carlsbad, CA) following the manufacturer’s protocol. Quality and integrity of the total RNA was controlled by using the Agilent Technologies 2100 Bioanalyzer and the RNA 6000 Nano LabChip® Kit (Agilent Technologies Waldbronn, Germany). Only samples with RNA integrity numbers (RIN) >8.0 were used for analysis.

3.2.4 cDNA synthesis

RNA was DNase-treated prior to cDNA synthesis using Ambion’s DNA-*free*™ following the manufacturer’s protocol. RNA was reverse transcribed using random hexamers (Fermentas) and the ThermoScript Revese Transcriptase Kit (Invitrogen, Carlsbad, CA) using 1 µg RNA as starting material according to the manufacturer’s protocol.

3.2.5 Quantitative real-time PCR (qRT-PCR)

Confirmation analysis for the candidate genes was performed for two different data sets, one with an extended set of inbred, which is used to exclude sample variability as having caused results. The other sample set consists of wild animals. Repeating the analysis with wild animals shall ensure that the observed result is a general subspecies effect rather than an inbreeding effect.

A. Using TaqMan® Gene Expression Assays

For seven genes the confirmation of the GeneChip® expression differences was done using TaqMan® Gene Expression Assays from Applied Biosystems (Supplement 6). PCR amplicons were obtained from ABI's 'Assay on Demand' selection and reactions were performed using 2 µl of a 1:10 diluted cDNA in 8 µl reaction volume, containing 4 µl TaqMan® Universal Mastermix, 0.4 µl TaqMan® Gene Expression Assay and 1.6 µl H_2O. For each individual a single cDNA synthesis was performed, which was used in triplicate in the qRT-PCR. Averaged C_T values from each qRT-PCR reaction of the target gene were standardized relative to the endogenous control of the same sample. Eif4g2 (eukaryotic translation initiation factor 4, gamma 2) was used as endogenous control as it showed highly similar expression levels and the lowest standard deviation across individuals, subspecies and different tissues in the GeneChip® analysis, including three different subspecies and tissues, respectively (data not shown). Real-time PCR reactions were performed using the ABI PRISM 7900HT sequence detection system (Applied Biosystems) using 384 well plates with recommended thermal cycling protocols and 40 cycles of amplification. Threshold cycle (CT) values were determined using the supplied sequence detection system (SDS 2.1.1) software package.

B. Using Sybr Green fluorescent dye

All ten candidate genes (Table 3.1, Supplement 6) were used for the Sybr Green experiment. QuantiFast® Sybr Green PCR Kit (Qiagen) was used in a 10 µl reaction volume according to the manufacturer's protocol with 2 µl of a 1:40 dilution of cDNA as starting material. Again, samples were performed in triplicate. Averaged CT values from each qRT-PCR reaction of the target gene were standardized relative to the endogenous control, Eif4g2, of the same sample. Primers (Supplement 7) were designed in a way that the PCR product of each candidate gene represented exactly a section of the target region of the Affymetrix GeneChip®, whereas it was assured that primer sequences contained no differences (i.e. SNPs) between the two subspecies (sequencing of the target region is described in 3.2.7) to exclude differences in primer affinity. Melting curve analysis as well as agarose gel electrophoresis was done to check for the specificity of the PCR products. Experiments were run on an ABI Prism 7900HT Sequence Detection System following the provider's recommendation concerning the thermal cycling protocol, using 384 well plates and running SDS 2.1.1.

3.2.6 Analysis of qRT-PCR data

For each qRT-PCR reaction the averaged C_T value of the endogenous control was subtracted from the averaged C_T value of the target gene, yielding the ΔC_T value. These values were then averaged across the triplicate samples within an individual, yielding the average ΔC_T value. Outliers among triplicates were removed if they raised the standard deviation above 0.3.

The following formula was used to calculate the fold-change between the two subspecies:

$$\text{fold-change} = 2^{-(\overline{\Delta C_T\_dom} - \overline{\Delta C_T\_mus})},$$

where $\overline{\Delta C_T\_dom}$ is the average ΔC_T value across the *M. m. domesticus* individuals and $\overline{\Delta C_T\_mus}$ is the average ΔC_T value across individuals of *M. m. musculus*. In case that *M. m. domesticus* expression levels are smaller than *M. m. musculus* values, fold changes are shown as the negative reciprocal value for reasons of convenience (i.e. the direction of change can be directly deduced).

ΔC_T values were used in a Kruskall-Wallis test to test for expression differences between the three groups *M. m. musculus*, *M. m. domesticus* and the hybrids, respectively. Subsequent pair wise Mann-Whitney-U tests were performed to identify significant pair wise differences.

3.2.7 Sequencing of Affymetrix GeneChip® target regions

A GeneChip® probe array consists of a number of probe cells where each probe cell contains a unique probe. The Affymetrix Mouse Genome 430 2.0 GeneChip® has 45,000 probe sets which analyze the expression level of over 39,000 transcripts and variants from over 34,000 well characterized mouse genes. A probe set is composed of eleven pairs of 25-mer oligonucleotide probes, which measures the expression for a section of the full-length sequence of a specific mRNA and is therefore complementary to a target sequence derived from that specific mRNA. Probes are tiled in probe pairs as Perfect Match (PM) and a Mismatch (MM) (Figure 3.1). The sequence for PM and MM are the same, except for a change to the Watson-Crick complement in the middle of the MM probe sequence. The reason for including a MM probe is to provide a value that comprises most of the background cross-hybridization and stray signal affecting the PM probe. To define a measure of expression representing the amount of the corresponding mRNA it is necessary to summarize probe signal intensities for each probe set. Different algorithms are available for summarizing the signal intensities. The signal value is calculated from the combined, background-adjusted, PM and MM values of the probe set if the standard Affymetrix MA Suite 5.0 is used (described in detail in “Statistical Algorithms Description Document” provided by Affymetrix 2002), RMA (Irizarry et al. 2003B) considers PM oligos only (for a comparison of different algorithms refer to Irizarry et al. 2003A and Irizarry et al. 2003B). The newest generations of Affymetrix expression quantitation refrains from including mismatches as it was shown in several studies that mismatch signal intensities often exceed perfect match intensities and therefore one would expect that including mismatch oligos in the normalization methods contributes to noise rather than reducing it (Harr and Schlötterer 2006). Especially for low-intensity genes using of MM data has been found to be unreliable (Irizarry et al. 2003A, Qin et al. 2006).

The eleven probes of a probe set cover the target sequence (between 150 and 500 bp long) to varying degrees since the probes of a set overlap to greater or lesser extend, depending on the target sequence size (usually overlap is the higher the shorter the sequence is).

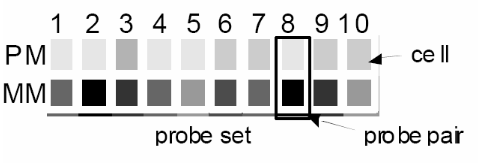


Figure 3.1: GeneChip® array design. Picture derived from Affymetrix statistical algorithms descriptions document.

The probe sets of the Mouse Genome 430 2.0 Affymetrix GeneChip® are designed based on the sequence of the laboratory inbred strain C57BL/6J. The genome of this strain represents a mixture of genetic contributions from the three subspecies *M. m. musculus*, *M. m. domesticus* and *M. m. castaneus* (Wade et al. 2002, Yang et al. 2007), whereas *M. m. domesticus* provides up to 92% of the sequence (Yang et al. 2007). Therefore it is possible that sequences differences (i.e. SNPs or alternative transcripts) between the wild subspecies translate into differences in hybridization efficiency (elevated or decreased signal intensities) of specific probe sets; i.e. one can expect that a probe set hybridizes more efficiently with that subspecies the particular probe set sequence was derived from. To control for this, the target sequence of all candidate genes was sequenced for all *M. m. musculus* and *M. m. domesticus* samples used for the GeneChip® analysis (Table 3.1, Supplement 1).

The sequence of the target region for each candidate gene is available from the Affymetrix homepage (http://affymetrix.com/index.affx). Whole sequences for the candidate genes were downloaded from ENSEMBL (http://www.ensembl.org). Primers were designed to amplify the whole target region (Supplement 7). PCR was setup in a 10 µl volume using the Qiagen® Multiplex PCR Kit following the manufacturer's protocol with 1 µl of 1:10 diluted cDNA as starting material. PCR products were purified by adding 0.012 µl Exonuclease I (Biolabs® 20 U/ml) and 0.045 µl Shrimp Alkaline Phosphatase (Promega, 1U/ µl) per 10 µl reaction and heating 20 minutes for 37ºC followed by 72ºC for 15 minutes. PCR products were sequenced in both directions using the BigDye sequencing chemistry on an ABI3730 automated sequencer. Sequences assembly and analysis was performed using the program CodonCode Aligner 2.0. Heterozygous positions were identified through visual inspection of all sequence positions for which the automatic nucleotide calls were ambiguous. All sequences were aligned using CLUSTAL W (Thompson et al. 1994). All sequences are provided in the Digital Supplement.

Differences (i.e. SNPs) to the target sequences provided by Affymetrix are identified and related to the subspecies-specific expression levels and analyzed with respect to their relevance for hybridization efficiencies. Because some probes of a probe set are overlapping, it may be that one SNP in the target region influences the overall expression level of a probe set more than one time since more than one probe binds to that region. Therefore the (background-corrected) PM signal intensities of each single probe were related to the accordance of the target sequence with the probe sequence to see if multiple hit of a SNP more likely influence signal intensities.

3.2.8 Sequencing of amplicon regions

In a similar way as SNPs or alternative transcripts affect hybridization efficiencies of the microarray sequence, differences between subspecies may influence the proper binding of the primers and the probe of the Applied Biosystems TaqMan® assays, which are designed on basis of the inbred strain C57BL/6J, and therefore potentially lead to artifacts. To test for this the amplicon region of all seven candidate genes analyzed with TaqMan® Gene Expression assays as well as the endogenous control was sequenced for all samples used for the qRT-PCR (Supplement 6) to confirm, that only differences in mRNA levels are captured by the assays. The assay location and the amplicon length are provided on the Applied Biosystems webpage (https://www2.appliedbiosystems.com). The sequence for each candidate gene was downloaded from ENSEMBL (http://www.ensembl.org). Primers were designed to amplify the whole amplicon region (Supplement 7). Amplicon lengths range from 60 to 150 bp, including primers of approximately 20 bp length and a probe of 12-20 bp length. As the exact position of the primer/probe position is not provided by Applied Biosystems, one can only conclude if a SNP at any position in the amplicon region potentially influences primer/probe binding, if a primer or probe is actually affected remains speculative. But at least for the primers, conclusions can be made with a certain probability as those are located at the beginning/end of the amplicon. For PCR and sequencing setup refer to 3.2.7.

3.3 Results

3.3.1 qRT-PCR using TaqMan® gene expression assays

Table 3.3 shows the results of the seven candidate genes analyzed with qRT-PCR using TaqMan® gene expression assays, considering fold changes for the samples used for the GeneChip®, for the inbred samples and the wild samples separately. Only for Neil3 the expression difference between the two subspecies as assessed by the Chip (fold change = 7.2) could be confirmed with an extended data set, showing a similar result for the Chip samples with a fold change of 8.9, a reduction in fold change with more inbred samples and wild samples, respectively, though still clearly differentially expressed. For 1700011K15Rik the fold change (20.2) could only be confirmed for the Chip samples (13.9) and the inbred animals (9.6). For all other candidate genes the expression difference, as deduced from the Chip, could not be confirmed and showed the opposite direction of change. No data were obtained for Pga5 for all samples, i.e. this transcript is not or very low expressed.

Table 3.3: Fold changes between *domesticus* and *musculus* for candidate genes inferred from TaqMan® qRT-PCR analysis (FC=fold change).

Gene	**FC GeneChip® *dom/mus* Chip**	**FC qRT-PCR *dom/mus* Chip-samples**	**FC qRT-PCR *dom/mus* inbred samples**	**FC qRT-PCR *dom/mus* wild samples**	**Confirmation of Chip result inbred samples**	**Confirmation of Chip result wild samples**
Scarb2	-19.8	2.19	2.0	1.2	no	no
Pga5	-11	no amplification	no amplification	no amplification	no	no
Neil3	7.2	8.9	4.5	3.4	yes	yes
1700011 K15Rik	20.2	13.9	9.6	-1.4	yes	no
Srp14	23.4	-1.7	-1.5	-1.3	no	no
1700029 F12Rik	207.3	-1.4	-2	-2.5	no	no
Rhobtb3	12.9	-1.1	-1.2	-1.2	no	no

Analysis for significant differences between the two subspecies respectively the hybrids was only performed for genes, for which the fold change from the microarray could be confirmed, by using the ΔC_T values in a Kruskall-Wallis test. To identify the contribution of the three groups to the significant value, subsequent pair wise Mann-Whitney-U tests were performed. Both Neil3 and 1700011K15Rik are significantly differentially expressed between the three groups for the inbred sample set ($p<0.05$). Pair wise comparisons show that the parental subspecies differ significantly from each other as do the hybrids from both parents (i.e. intermediate, $p<0.05$). For 1700011K15Rik the averaged expression difference of all hybrids differs by 2.3 fold from inbred *domesticus* and by 3.8 fold from inbred *musculus*, for Neil3 hybrids differ 2.7 fold from *domesticus* and 2.4 fold from *musculus*, thus clearly showing intermediate expression levels. For Neil3 the three groups also differ significantly in the wild sample set ($p<0.05$) but the pair wise comparison shows that this significance can only be attributed to a significant difference between the parental subspecies and a difference between the hybrids and *M. m. musculus* ($p<0.05$, fold change 3.7). The expression level of the hybrids therefore follows the expression of *M. m. domesticus* (i.e. the hybrids are not intermediate in expression). Supplement 8 shows boxplots for the expression differences of Neil3 and 1700011K15Rik of the subspecies and hybrids, respectively. Altogether, only two out of seven candidate genes (29%) can be confirmed for the animals used for the microarray and only for Neil3 the expression difference between *domesticus* and *musculus* (as derived from the Chip) can be validated also for the wild sample set.

3.3.2 qRT-PCR using Sybr Green fluorescent dye

Quantitative real-time PCR with Sybr Green was performed for all but one of the candidate genes (n=10). For Pga5 the establishment of adequate primers and therefore amplification was not possible. For all the other genes amplification was successful and primers gave specific products as checked by melting curve analysis and agarose gel electrophoresis.

Fold changes from the microarray could not be validated for Scarb2, Srp14, Rhobtb3 and Lsm14 for a larger data set, neither for the inbred nor for the wild samples, regarding the magnitude respectively the direction of change. For the other 5 genes the fold changes from the microarray analysis could be validated for both data sets (Table 3.4). Though a fold change could not be calculated for 1700029F12Rik, the qRT-PCR result supports the Chip outcome as no expression at all could be detected for *M. m. musculus*, thus *M. m. domesticus* being much higher expressed.

Table 3.4: Fold changes between *domesticus* and *musculus* for candidate genes inferred from Sybr Green qRT-PCR analysis (FC=fold change).

Gene	FC *dom/mus* Chip	FC qRT-PCR *dom/mus* Chip-samples	FC qRT-PCR *dom/mus* inbred samples	FC qRT-PCR *dom/mus* wild samples	Confirmation of Chip result inbred samples	Confirmation of Chip result wild samples
Scarb2	-19.8	1.133388	-1.66232	8.755755	no	no
Neil3	7.2	30.66613	37.72493	153.6626	yes	yes
1700011 K15Rik	20.2	11.92926	11.11592	2.899065	yes	yes
Srp14	23.4	-1.40169	-1.54796	1.15394	no	no
1700029 F12Rik	207.3	*mus* not expressed	*mus* not expressed	*mus* not expressed	yes	yes
Rhobtb3	12.9	1.08103	-1.06672	1.201301	no	no
Lsm14a	-7.35	1.380218	1.538537	1.02772	no	no
Lsm10	-7.81	-13.3617	-11.1333	-7.52597	yes	yes
270002909Rik	-7.79	-2.85265	-4.09687	-4.15821	yes	yes

For all genes for which the fold change from the microarray could be validated via qRT-PCR, this expression difference is highly significant across the three groups (Kruskall-Wallis test, p<0.05) irrespective of whether only inbred samples or wild samples were considered (Table 3.5). Pair wise Mann-Whitney-U tests showed significant differences for all comparisons except for *musculus*-hybrid in 2700029M09Rik (inbred samples) and *domesticus*-hybrid in 1700011K15Rik (wild samples). For 1700011K15Rik the non-significant difference of the averaged hybrid expression levels to *domesticus* is attributed to one extreme outlier each in *domesticus* and in the hybrids, which confounds the result. Apart from significance the expression level for the hybrids is clearly intermediate. For all other genes/data sets the hybrid data show undoubtedly intermediate expression. For Neil3 the hybrids show clear and significant intermediate expression but with a shift to *M. m. domesticus* expression levels for both data set. This is in accordance with the result from TaqMan® analysis as for the wild sample set the expression of the hybrids follows *M. m. domesticus*. Supplement 9 shows boxplots for all genes with significant differences.

Table 3.5: Significance analysis for genes that could be confirmed with qRT-PCR using Sybr Green. ΔCT values were used for analysis. Table shows the significance p-values.

Gene	**Kruskall-Wallis significance p-value**	**Mann-Whitney-U test significance p-value**	**Mann-Whitney-U test significance p-value**	**Mann-Whitney-U test significance p-value**	**Kruskall-Wallis significance p-value**	**Mann-Whitney-U test significance p-value**	**Mann-Whitney-U test significance p-value**	**Mann-Whitney-U test significance p-value**
	inbred samples	*dom-mus*	*dom-hybrid*	*mus-hybrid*	**wild samples**	*dom-mus*	*dom-hybrid*	*mus-hybrid*
Neil3	0.00045	0.00952	0.00040	0.00280	0.000156	0.002165	0.0004	0.0004
1700011 K15Rik	0.00096	0.01587	0.00200	0.00280	0.001066	0.002165	0.113487	0.0004
1700029 F12Rik	*mus* not expressed	*mus* not expressed	0.00067	*mus* not expressed	*mus* not expressed	*mus* not expressed	0.000666	*mus* not expressed
Lsm10	0.00307	0.01587	0.00794	0.01587	0.000811	0.002165	0.004329	0.004329
2700029 M09Rik	0.00638	0.02381	0.00480	0.20909	0.000156	0.002165	0.0004	0.0004

Summarizing 50% of the candidate genes from the screen, tested via Sybr Green qRT-PCR show up as reliable candidates for further analysis.

To compare the validation rates of the two qRT-PCR methods, only those transcripts were considered, which have been analyzed with both methods (Scarb2, Pga5, Neil3, 1700011K15, Srp14, 1700029F12Rik, Rhobtb3). Analysis with TaqMan® Gene Expression assays yielded in a validation rate of 29% and with Sybr Green in 43%.

3.3.3 Sequencing of Affymetrix GeneChip® target regions

Contribution of sequence differences to hybridization efficiency

The target region as derived from the Affymetrix website was sequenced for all eleven candidate genes for the samples, which were used for the microarray experiment (Supplement 1). This analysis served as a test if sequence differences between the two subspecies cause differences in hybridization efficiency and therefore signal intensities. For all but one candidate gene sequencing results were obtained from cDNA for both subspecies. Amplification of cDNA was not possible for 1700029 F12Rik for *M. m. musculus.* Surveying the ten remaining transcripts, one can say that for 80% at least one SNP between the two subspecies is detected in the whole target region (Table 3.6). If only transcripts are considered, which have probes that indeed would capture an existing SNP (the target region is only partly covered by probes), 60% (six transcripts) may be affected by differences in hybridization efficiency (Neil3, Srp14, Rhobtb3, Lsm14, Lsm10 and 2700029M09Rik, Table 3.6). A more detailed analysis, in which the signal intensities of the single probes are correlated to deviation from the C57BL/6J sequence, reveals that only for two genes (Srp14, Lsm14) the overall expression level is influenced by SNPs, both cases, where several probes capture the SNP (Supplement 10). The six probes which are affected by the SNPs in the sequence of Srp14, exhibit clearly reduced signal intensities in *musculus*, whereas the non-affected probes show intensities comparable to *domesticus*. For Lsm14 the same pattern is observed for probe 1-5 with reduced signal intensities in *M. m. domesticus*, whereas probe 9-11 do not show reduced signal intensity in *domesticus* (but signal intensity is low in both subspecies). However, the non-affected probes clearly show no serious difference in signal intensity between the two subspecies. Thus, the overall expression difference between the two subspecies, as deduced from the microarray analysis, for this two genes is clearly due to an artifact and the overall signal intensity difference only caused through the affected probes. For Neil3, Rhobtb3, Lsm10 and 2700029M09Rik, the SNP either does not influence the signal intensities for the affected probe or the overall result for the subspecies comparison. For Neil3 *domesticus* shows even higher signal intensities irrespective of SNPs affecting probe binding, underlining that a mismatch does not necessarily lead to poorer hybridization efficiency.

Position and type of mismatch

From this analysis no general conclusion about the effect of the position and the type of the SNP in a duplex can be made. There is a slight tendency for signal intensities to be more seriously affected if the SNP is positioned in the middle of a probe sequence. Both Srp14 and Lsm14 exhibit the SNP more to the middle of the probe sequence, whereas in transcripts for which the SNP has less impact the SNP is positioned more to the 3’ or 5’ end of the probe sequence. According to Pozhitkov et al. (2006) two groups of mismatches with clearly separated extremes exist. GA and GG mismatches destabilize duplexes most; TC, TU and TG mismatches destabilize duplexes the least, all other types of mismatches resulting in medium destabilization. Both Srp14 and Lsm14 have types of mismatch which fall into the strong respectively the medium class (GA and CA), thus the alteration in signal intensity may be influenced by the type of mismatch. But on the contrary, also Neil3, which has no overall changed signal intensity, has mismatches which fall into the class of strong and medium mismatch.

Altogether, the analysis shows two out of eleven picked candidate genes clearly as false positives due to sequence differences between the two subspecies.

Table 3.6: Results of Affymetrix target region sequencing.

Gene	**High overall expression (MAS 5.0)**	**SNP in target region in**	**No. SNPs**	**No. probes affected by SNP**	**SNP changes overall expression**	**Size PCR product**	**SNP position**
Scarb2	*mus*	–	–	–	no	789	–
Pga5	*mus*	–	–	–	no	744	–
Neil3	*dom*	*dom*	2	3	no	286	122 244
1700011 K15Rik	*dom*	*mus*	1	–	no	691	153
Srp14	*dom*	*mus*	3	6	yes	393	189 299 303
1700029 F12Rik	*dom*	*dom* SNP free	?	?	?	591	?
Rhobtb3	*dom*	*dom*	1	1	no	688	488
Lsm14a	*mus*	*dom*	2	8	yes	577	77 522
Lsm10	*mus*	*mus*	3	2	no	612	132 385 397
2700029 M09Rik	*mus*	*dom*	2	1	no	629	148

3.3.4 Sequencing of the amplicon regions

To test whether sequence differences between the subspecies influence proper binding of TaqMan® gene expression assays, all amplicon regions were sequenced to control for such artifacts. Table 3.7 shows number of sequence differences (for details refer to Supplement 11).

Table 3.7: Sequencing results for TaqMan® amplicon regions of the candidate genes.

Gene symbol	Assay ID	SNP	Fixed/ polymorph
Scarb2	Mm00446978_m1	0	no
Pga5	Mm00480598_m1	1	fixed
Neil3	Mm00467596_m1	1	polymorphic
1700011 K15Rik	Mm00661433_s1	2	fixed
Srp14	Mm00726104_s1	2	polymorphic
1700029 F12Rik	Mm00481622_m1	0	no
Rhobtb3	Mm00712630_m1	0	no
Eif4g2	Mm00469036_m1	0	no

For 50% of all sequenced amplicon regions, at least one SNP could be identified between the two subspecies (all samples used for qRT-PCR were sequenced), either being fixed or polymorphic. The endogenous control Eif4g2 shows no sequence differences between the two subspecies. For 1700011K15Rik it is quite certain that the SNP affects the F-primer-binding as two SNPs are located within the first 20 bp of the amplicon sequence but to which extend and with which consequence remains open. For the other three candidate genes it is not possible to infer if a SNP really affects primer/probe-binding but at least a certain probability exists. Amplification of cDNA was not possible for Pga5 for both *M. m. domesticus* and *M. m. musculus*, results are deduced from sequencing genomic DNA. Consistent with the fact that the assays are designed on basis of the laboratory mouse strain C57BL/6J, whose genome is mainly derived from *domesticus*, more SNPs are found in the *musculus* samples (Supplement 11). For all genes tested with TaqMan® assays, there was no overlap of the assay sequence and the Affymetrix target region. In all but two genes the assays sequence is located upstream of the Affymetrix target region, even in different exons. For Srp14 and 1700011K15Rik the assays sequence shows also no overlap with the target sequence but is located within the same exon.

Summarizing, for half of the sequenced amplicon regions at least one SNP could be identified in one of the subspecies, thus having the potential to influence proper primer/probe binding.

3.3.5 Combining results – evidence for alternative splicing

For Pga5 qRT-PCR with TaqMan® gene expression assay approaches yielded no results for both subspecies, whereas cDNA sequencing of the Affymetrix target region was possible, supporting that this gene is expressed. PCR of cDNA for the TaqMan® amplicon region failed, suggesting that this part of the sequence is not expressed in both subspecies. Eventually, both subspecies express a transcript different from the transcript on which the Chip is designed, which lacks sequence 5' upstream of the Affymetrix target region. Since the establishment of adequate primers for the Sybr Green qRT-PCR was not possible, no conclusion about the expression difference obtained by the GeneChip® is possible. For 1700029F12Rik, amplification of cDNA in *M. m. musculus* with the Sybr Green approach was not possible (Table 3.4), additionally sequencing, respectively PCR, with cDNA for the Affymetrix target region did not reveal any result (Table 3.6). It is therefore possible, that the GeneChip® expression difference results from either the absence of the target sequence on cDNA level in *M. m. musculus* (alternative spliced transcript) or complete missing of the transcript. PCR/Sequencing of the amplicon region of the TaqMan® (which lies 5' upstream of the Affymetrix target region) was possible for *M. m. musculus*; also amplification with the TaqMan® approach succeeded, thus the gene is expressed. The fact that the expression difference from the GeneChip® is not reflected by this qRT-PCR method, underlines that *musculus* has a different transcript which is not captured by the microarray.

3.4 Discussion

Factors which may influence oligonucleotide microarray data and the derived candidate genes from such experiments are various and hardly predictable as even the physics of oligonucleotide arrays are not sufficiently understood (Morey et al. 2006, Pozhitkov et al. 2007). Also follow-up-confirmation methods such as qRT-PCR are not free of pitfalls including artifact-creating factors such as amplification bias (Chuaqui et al. 2002), mispriming and the formation of primer dimers (Bustin 2002) and the changing efficiency in later cycles (Freeman et al. 1999). In essence, different factors may have a contribution to false-positive microarray outcomes but also the validation method may not reveal completely unbiased results. Accordingly, validation rates for microarray data with qRT-PCR exhibit wide ranges in literature, varying from as low as 23% (Jurata et al. 2004) up to 94% (Bosotti et al. 2007), depending on factors like the type of microarray, used sample, applied normalization and statistic, type of qRT-PCR, magnitude of fold-changes and more. The present analysis particular stresses the importance of sequence differences between the closely related subspecies of the house mouse, *M. m. domesticus* and *M. m. musculus*.

3.4.1 Comparison of two qRT-PCR methods

The validation of candidate genes derived from a microarray screen with two different approaches revealed that a substantial fraction of candidate genes are false positives and not suited for further analyses.

Validation rate

Achieving the same expression data for the samples used for microarray with an independent method ensures that no measurement error caused the difference, whereas a validation for a larger set of samples corrects for errors due to sample variability (Allison et al. 2006). The qRT-PCR-validation was conducted with two different methods: 1. premade *Mus musculus* TaqMan® gene expression assays for the genes of interest were used and 2. primers were designed for use with Sybr Green (after sequencing the region of interest to exclude sequence differences between the subspecies). The results for seven candidate genes which were tested with both methods differ but show overlap with a confirmation rate of 29% for the TaqMan® approach (two validated transcripts) and 43% for the Sybr Green approach (three validated transcripts). Data achieved from microarrays and qRT-PCR often show inconsistencies, a wide range of correlation between the two methods exist (Morey et al. 2006). The validation rate of 29% (43%) achieved in this study seems to be relative low compared to other studies. Validation rates up to 94% are reported (e.g. Rajeevan et al. 2001, de Vos et al. 2003, Bosotti et al. 2007). However, many of those analyses are cell-line experiments and therefore highly controllable and don’t include species or subspecies comparisons. Furthermore candidates mostly have been picked with respect to high overall signal intensities and fold-changes between the samples (for which it is known that the concordance of microarray and qRT-PCR data are high (Etienne et al. 2004)) and different microarray and qRT-PCR methods have been applied, respectively. In general it seems not reasonable to compare validation rates across different experiments as too many differences in experiment conduction, used techniques etc., are involved. Fold changes for the confirmed candidate genes do not correspond perfectly with the fold changes of the microarray, either being inflated or deflated. Differences in the expression levels when comparing data from microarrays and qRT-PCR analyses are generally observed (Draghici et al. 2006); especially for genes with low expression on the microarray little agreement between these methods has been found (Kuo et al. 2006). This was attributed as stochastic variation due to the low transcript abundance in both microarray and validation procedures. The exact cause of compression/decompression is not clear and the magnitude of the effect is not predictable. Overall, the magnitude of change was in an equivalent range for the analyzed transcripts for both used qRT-PCR methods. This outcome is in accordance to an experiment of Andersen et al. (2006), who compared four different qRT-PCR methods and showed that none of these methods performed significantly better concerning sensitivity and yielded in similar C_T values. Sybr Green performed slightly better regarding the rate of confirmation than TaqMan® assays in the present study, which can be attributed to identifying identical transcript structures like in the microarray analysis. The overall outcome suggests that a substantial fraction of candidate genes are false positives and not suited for further analyses.

In the following, factors which contribute to these skewed results are discussed. Other factors, like cross-hybridization, non-specific hybridization, incorrect probe design parameters or even washing off specific targets over non-specific ones (Pozhitkov et al. 2007), may also have contributed to deviations of qRT-PCR data and microarray data but cannot be disentangled by this comparative analysis; they contribute to varying and unpredictable degrees.

Sequence differences affecting qRT-PCR results

Subspecies-differences like SNPs and alternative transcripts may be very critical for the follow-up confirmation with qRT-PCR. This becomes obvious, if one compares the two approaches and considering sequence data of the amplicon region of the TaqMan® assay. If designing ones' own primers and probes (as for the Sybr Green approach) one can exclude that differences in the sequence between two samples influence the amplification efficiency but if using pre-designed primers and probes such as TaqMan® "assay on demand", for *Mus musculus,* amplification bias possibly alter the results. The sequencing of the TaqMan® amplicon regions showed 50% of the genes with sequence differences (i.e. SNPs) between the two subspecies, thus being a potential important problem in a validation approach. A parallel study also records pre-designed assays as prone to amplification differences if the assays are not specifically designed for the species under study (Fabian Staubach, personal communication, unpublished data). Because only two of the seven candidate genes (Neil3, 1700011K15Rik), tested with both approaches have been validated, conclusions about a contribution of SNPs can only be made concerning these two transcripts. The direction and the magnitude of the fold change via qRT-PCR was similar to the microarray result for the samples used for the GeneChip® and the inbred samples, the confirmation also holding true if statistic tests are performed. Inconsistencies between the two methods become evident, if considering the data set which includes wild *domesticus* and *musculus* samples. While for Sybr Green both transcripts show clear and significant expression differences with intermediate hybrid expression, deviation from these outcomes are observed for the TaqMan® approach. Inbreeding and other biological effects can be ruled out as solely having created this effect, since the Sybr Green approach confirms the microarray results for both sample sets. Most probable differences in primer/probe binding between the two subspecies and the hybrids, respectively, are responsible for the deviating outcome as for both transcripts SNPs in the TaqMan® amplicon region are identified, either being fixed or polymorphic. At least for 1700011K15Rik two SNPs have the potential to influence the proper binding of the F-primer, if and to which extend the SNPs influence the amplification cannot be resolved. For Neil3 (and also the rest of the tested transcripts) the actual contribution of sequence differences remains unclear.

One has of course to consider that the two qRT-PCR methods differ substantially. TaqMan® Gene Expression Assays are general assumed to be more accurate and Sybr Green more sensitive, binding any double-stranded DNA in the reaction (including genomic DNA, primer-dimers and other non-specific reaction products) which may results in an overestimation of the target concentration, producing higher fold changes. However, cDNA was DNAse-treated prior to qRT-PCR and melting curve analysis was performed to check for specificity of the PCR-product, so that the more pronounced expression differences in the Sybr Green analysis seem unlikely be due to inexactness of the approach.

Nevertheless at least for subspecies with marginal sequence differences, it is obvious that if an expression difference exists (at least with a certain magnitude), it can be detected irrespective of sequence deviations. But it is also obvious that if very accurate results are needed or small differences shall be detected, the contribution of such sequence differences in a sensitive method like qRT-PCR should not be neglected. Therefore the specific design of primers and probes, after a sequence comparison of the target, seems to be the appropriate consequence one should draw from these observations, especially with regard to alternative splicing (see 3.4.3).

3.4.2 Sequence differences affecting microarray probe binding

Sequence analysis of the Affymetrix GeneChip® target region for the two subspecies *M. m. domesticus* and *M. m. musculus* revealed that a substantial fraction of differences in signal intensities are due to sequence differences rather than differences in mRNA abundance and that even in a narrow phylogenetic comparison (as between subspecies) are error-prone. Microarrays designed for another species than that under study (cross-species analyses) are frequently used. Provided that both species share enough sequence similarity, RNA transcripts for one species will hybridize efficiently with the arrayed sequence of another species. It is generally assumed that the accuracy of the data decreases with increasing sequence divergence (Nieto-Díaz et al. 2007). This implies that such analyses result in more false-negative genes that appear not to be expressed although they are really being expressed, than same-species analyses. It is therefore intuitive that also subspecies-sequence differences may falsify expression data. Differences in sequence and subsequent differences in hybridization efficiency/signal intensity may be assigned falsely as expression difference. For two transcripts the signal intensities for probes which capture the sequence difference were evidently altered resulting in an overall expression difference for these transcripts. In contrast to the transcripts, for which SNPs in the probe sequence did not lead to altered signal intensities, in these two probe sets several probes overlapped such that a single SNP was captured numerous times adding up to an overall altered signal intensity. It is therefore possible that a certain “critical number” of probes must be affected to result in an overall expression difference, respectively that one or two altered probe signals can be out-weighed by the remaining ones. Overall the results follow the basic assumption concerning probe binding that is, that deviations from the C57BL/6J sequence result in poorer hybridization efficiency and therefore signal intensity. Probes, which capture SNPs, generally show reduced signal intensities in comparison to the signal intensities of the non-affected probes. But also transcripts deviating from this general trend have been identified, where the signal intensities are not severely altered by a SNP respectively showing even higher signal intensity irrespective of a SNP. This is in accordance with the observation that a mismatch in a probe does not necessarily lead to poorer hybridization efficiency but may also enhance the binding (Kierzek et al. 1999).

In general several factors influence the binding affinity of target and probe. It varies depending on the position and type of mismatch as well as neighbouring nucleotides in a probe (Pozhitkov et al. 2006); in addition there is also interaction of theses factors. The most significant contributor to decreased signal intensities is the position of the mismatch, the more it is moved away from 3' or 5' end to the middle the stronger the reduction in signal intensity. Surveying the affected transcripts, this assumption is confirmed. Both transcripts with strongly altered signal intensity have the SNP more in the middle of the probe sequence, whereas in transcripts for which the SNP has less impact, it is positioned more to the 3' or 5' end of the probe sequence. Furthermore for those two transcripts mismatch types are detected which fall in the class of highly and medium destabilizing duplexes. However, also for transcripts, for which no seriously altered change in expression level in probes affected by sequence differences could be detected, this classes of mismatch type (highly and medium destabilizing) has been found. Thus, the type of mismatch alone cannot explain why some SNPs alter signal intensities and some do not. It seems most likely that the falsified result for Srp14 and Lsm14 is due to high overlap of probe sequences which results in numerous caption of the sequence difference such as it can not be out-weighed by the remaining probes. Furthermore the position and the type of the mismatch may play a role, so that for these two transcripts several factors influence the signal intensity.

Of course, one has to be aware that this analysis is purely descriptive and lacks any statistic validation. However, general observations of other research can be at least comprehended if not confirmed. Furthermore this analysis only considers the PM probe sets as only those are provided by Affymetrix, so that it remains unclear which effect subspecies sequence differences have in respect to MM probe sets (if a SNP for example leads to a perfect match in a MM probe) and how the overall signal intensities from a MAS 5.0 analysis would be influenced. Nevertheless, for a first approximation the applied approach (with only considering PM probes and signal intensities) seems to be adequate.

The sequencing of the microarray target region for the two subspecies clearly shows that even for comparisons of samples with low sequence divergence, any obtained result from a microarray analysis has to be validated with an independent method as a considerable amount of expression differences may be a mere consequence of sequences differences. However, it seems clear that, if a pronounced expression difference between the subspecies exists, one can detect it, irrespective of sequence variation. Consequently, false-negative results because of subspecies sequence differences seem to be more improbable than false-positives which contribute – at least in this experiment – with 20% to the microarray candidate genes.

3.4.3 Alternative splicing as contributor to expression differences

Two transcripts (Pga5 and 1700029F12Rik) show inconsistencies when comparing the outcomes of the two qRT-PCR-methods with the Affymetrix GeneChip® results and the data obtained by sequencing the target and amplicon region, respectively. Combining the results suggests that alternative spliced transcripts have a potential to influence results of expression analyses. This is in accordance with the assumption of other authors who see discrepancies between different microarray platforms mainly caused through alternative splice variants in conjunction with cross-hybridization (Draghici et al. 2006) respectively think that the contribution of alternative splice variants to gene expression data has been largely overlooked until now (Gershon 2005). Expression differences due to alternative splice variants may be unnoticed in a follow-up validation if primers do not refer to the position of the microarray target, although some authors see it as advantage to design primers deviating from the microarray target sequence and state that only by doing so the actual expression of each gene is validated and not simply the signal detected by the microarray (Bosotti et al. 2007). But it seems likely that, especially for comparisons across samples with a certain sequence divergence (like subspecies) a validation approach is more promising if the identical region (3' region) like the microarray is surveyed rather than using any part of the sequence as Etienne et al. (2004) found that "…increased distance between the location of the PCR primers and microarray probes on a given gene also decreases the correlation between the two methods”. Thus, designing primers and probes specifically for the microarrays’ target region, is definitely the validation approach “closer” to the microarrays’ outcome, as the identical transcripts structures like on the microarray are detected.

3.4.4 Conclusion

The overall result of the validation experiment with comparison of two methods strongly emphasizes the need for confirmation of microarray data as a significant proportion of candidate genes seems to be assigned falsely due to artifacts. The comparison shows that sequence differences due to divergence must be considered when contrasting samples in microarray and/or follow-up validation approaches as there is potential to influence the results. Furthermore, alternative spliced transcripts have the potential to influence results and expression differences, respectively. Therefore it is suggested to specifically design primers and probes for qRT-PCR validation if species/subspecies are used for which inventoried assays were not designed for.

4 Characterization of candidate genes from a gene expression screen – insights from hybrid zone and population genetic analyses

4.1 Introduction

4.1.1 Protein and gene expression evolution

The relationship between protein and regulatory sequence evolution is a central question in molecular evolution. The observation that many proteins among distantly related taxa are highly conserved has raised the theory that most of the differences between taxa must arise from regulatory differences. Thus, changes in gene regulation and protein structure have been conventionally treated as independent modes of evolution (King and Wilson 1975). Indeed, many cases are known in which regulatory changes contribute to phenotypic difference (see Wray 2007 for a review). In humans, over 100 are known exclusively to affect diverse aspects of behavior, physiology and disease (Rockman and Wray 2002, Knight 2005). Likewise, several studies have identified rapidly evolving proteins. For example, evidence has been obtained for rapid evolution of amino acid sequences (ca. 5-9% of genes studied) in the hominid lineage, including genes involved in immune response (Nielsen et al. 2005). However, as a variety of factors may influence both protein and gene expression evolution (e.g. protein-protein interactions), it has been suggested, that protein evolution might be coupled with divergence in gene expression (Lemos et al. 2005A). But the relationship between sequence evolution and divergence in gene expression is controversial. In *Drosophila*, a positive correlation between the rate of protein evolution (i.e. the rate of amino acid substitution) and divergence in expression levels has been found for male-biased genes (Nuzhdin et al. 2004, Lemos et al. 2005A). Furthermore, Gu et al. (2002) and Makova and Li (2003) were able to show a positive correlation between divergence in protein sequences and mRNA levels between gene duplicates in yeast and human, respectively. In contrast, Wagner (2000) was unable to find an association between divergence in protein coding sequences and mRNA levels in gene duplicates in yeast. Additionally, Jordan et al. (2004) found no association in a human-mouse study, which led them conclude that the two modes of evolution (regulatory and structural) are largely decoupled. Differences in gene expression are –

as sequence differences – heritable and therefore a possible target of selection (Schadt et al. 2003, Morley et al. 2004, Gibson and Weir 2005). However, we just begin to understand the evolutionary mechanisms such as Darwinian selection and random genetic drift, which underlie differences in expression patterns, on a genomic level (Holloway et al. 2007). Recent studies have proposed a neutral model of transcriptome evolution, i.e. the transcriptome divergence accumulates approximately linear with time and changes in expression can be explained mainly neutrally (Khaitovich et al. 2004, 2005). Other studies suggest that stabilizing selection is the major force acting on gene expression levels, thus only a small fraction may be shaped by directional selection (Rifkin et al. 2003, Denver et al. 2005, Lemos et al. 2005B).

To sum up, to date little is granted about the evolution of gene regulation in general and of its relationship to protein evolution (Castillo-Davis et al. 2004, Andolfatto 2005).

4.1.2 Detecting signatures of selection

Several methods to detect signatures of selection are described, polymorphism based and divergence based methods. Polymorphism based methods are appropriate to detect single recent adaptive events while divergence based methods are more suitable to detect recurrent selection (see Jensen et al. 2007 for a review). Divergence based tests rely on manifestation of evolutionary forces at the target itself; traditionally these test have been almost exclusively used for the analysis of coding regions.

Increase or decrease in genetic variation in a population as a consequence of selection leads to specific patterns or “signatures” that are left in the DNA sequence. Such signatures can be identified through comparison with what is expected under a neutral scenario, where the amount of change is determined solely by the mutation rate. The neutral theory states, that most observed genetic variation within and between species is neutral, i.e. has no effect on the fitness of an individual (Kimura 1968) and thus changes are the consequence of chance alone (genetic drift). It is assumed that genetic variants, which change the amino acid sequence of a protein (non-synonymous substitutions), are on average deleterious and thus less likely to become fixed than a synonymous (silent) change (Hurst 2002). Nevertheless, through recurrent positive selection, function-altering mutations can eventually become fixed in a population. Identification of such a signature can be revealed by comparing DNA sequences between species. Commonly used tests include the Ka/Ks (dN/dS) test, which compares the rate of non-synonymous to synonymous changes (Nei and Gojobori 1986, Hurst 2002, Fay and Wu 2003) and the McDonald-Kreitman-test, which assesses synonymous and non-synonymous changes within and between species (McDonald and Kreitman 1991).

Other tests are intended to detect signatures of recent selection. Non-neutral mutations either rise or decrease in frequency in a population, depending on if they are exposed to positive or negative selection. Either of these modes of selection leaves a signature in the nearby genomic region, as the increase or decrease in frequency of a mutation is carried over to the linked neutral region. This effect is called "hitchhiking" if neutral linked variation is reduced due to positive selection at a nearby locus (Maynard Smith and Haigh 1974, Fay and Wu 2000). A typical recovery pattern after such a "selective sweep" event is an excess of rare alleles which is the result from new mutations. Such regions of overall low diversity with an excess of rare alleles can be detected by several statistical tests such as Tajima's D (Tajima 1989). If a site experiences strong purifying selection, linked neutral variation is also weeded out along with it, producing a region of overall low variability, an effect which is called "background selection" (Charlesworth et al. 1993). As "background" selection reduces linked variability and effects of demographic history (such as bottlenecks) may result in patterns similar to a "selective sweep" (Haddrill et al. 2005), one of the major challenges is, to disentangle the different potential origins of reduced sequence variability (Otto 2000). To overcome these limitations, Fay and Wu (2000) developed a test, which evaluates high and intermediate frequency new derived alleles (which is consistent with hitchhiking but not background selection). However, since significant results can be caused by advantageous mutations in either the coding region or regulatory elements in vicinity of the investigated regions, further analysis is needed to pinpoint the specific locus of selection.

The identification of regulatory regions (i.e. promoters and transcription-factor-binding sites) is still a challenge, since promoters are diverse and even well-known motifs are not fully conserved. In addition, each gene contains a set of unique combinations of transcription factor binding sites, which may vary substantially in size (Lenhard et al. 2003, Qiu 2003). Thus, identification of beneficial mutations in regulatory regions is less obvious than for protein-coding regions.

4.1.3 *Cis*- versus *trans*-regulatory evolution

Expression differences can arise from *cis*- or *trans*-regulatory changes. *Cis*-acting elements regulate the expression of genes on the same strand. Promotors, which regulate transcription, typically lie 5' of the start site of transcription. Elements that control mRNA stability and degradation are primarily located in 3' regions (Holloway et al. 2007). In contrast, *trans*-regulatory elements regulate genes distant from the gene they are transcribed from (e.g. transcription factors). The contribution of both to changes in gene expression remains largely unknown to date (Wray et al. 2003). Recent studies are coming to inconsistent results. *Trans*-acting factors as the predominant contributor to expression variance within species were identified in yeast (Yvert et al. 2003), humans (Morley et al. 2004), *Drosophila simulans* (Wayne et al. 2004) and *Caenorhabditis elegans* (Denver et al. 2005). In contrast, large surveys in mouse (Cowles et al. 2002, Doss et al. 2005), human (Yan et al. 2002) and *Drosophila* (Wittkopp et al. 2004) revealed *cis*-regulatory factors as the main source for gene expression variances. The different outcomes of these studies were at least partly attributed to differences in methodologies and statistics. Apart from inconsistencies regarding the contribution of *cis*- vs. *trans*-regulatory effects to differences in expression levels, it remains unclear from theses studies, whether any of these differences has an adaptive effect (Hoekstra and Coyne 2007). However, uncovering the genetic basis of variable gene expression is the first step for identifying the specific underlying nucleotide polymorphism causing the expression difference (Wittkopp et al. 2004).

4.1.4 Hybrid zone analyses

Under the "Biological Species Concept" (Mayr 1995), speciation is synonymous with the evolution of reproductive barriers, since reproductive isolation "…ensures that species remain genetically distinct and…can undergo independent evolutionary fates" (Orr and Presgraves 2000). One form of reproductive isolation is intrinsic postzygotic reproductive isolation, which refers to sterile or inviable hybrids from a cross of two different species and is thought to arise because alleles from different species, brought together in a hybrid genomic background, fail to interact properly (hybrid incompatibilities) (Dobzhansky 1936B). A distinct feature of hybrid sterility and inviability is that they require epistasis: nonadditive interactions between alleles at different loci (Dobzhansky 1937, Muller 1940, Turelli and Orr 2000). A hybrid incompatibility or speciation gene is, by definition one, that causes some degree of ecological, sexual or post-mating isolation between species (Wu and Ting 2004). Thus, identifying and describing hybrid incompatibility genes will provide insight in the process of speciation in general. To distinguish between causes and consequences of genetic isolation, it seems plausible to focus on cases where reproductive isolation is not complete yet (Macholán et al. 2007). Naturally occurring hybrid zones, contact zones of genetically distinct populations, are optimally suited for this purpose. The study of hybrid zones has contributed substantially to the understanding of how speciation is realized genetically; e.g. the number of loci underlying reproductive isolation has been estimated (Szymura and Barton 1986, Barton and Gale 1993). Furthermore, individual genomic regions putatively involved in reproductive isolation were identified by investigating patterns of differential introgression (Barton and Hewitt 1985, Rieseberg et al. 1999, Payseur et al. 2004, Macholán et al. 2007) through proposing that loci, which are involved in causing reproductive isolation, will introgress at lower rates than loci having a "neutral" effect in a hybrid. Hybrid zones thus provide a "natural laboratory" (Barton and Hewitt 1989) in which different genotypes are created and directly tested with respect to their adaptive value. Genes, which are significant for the speciation process, and therefore detrimental in a hybrid genomic background, should not move far across the hybrid zone (i.e. introgression should be restricted in comparison to neutral loci) (Barton and Hewitt 1985). The width of the spread of such genes is a direct estimate for the deleteriousness of a gene in a hybrid genomic background (Szymura and Barton 1986). To date only few such genes, which contribute to reproductive isolation via hybrid sterility, inviability and reduced fitness, have been identified (reviewed in Orr et al. 2004, Orr 2005) and hybrid zone analyses so far have never been used to actually identify single genes.

The two mouse subspecies *M. m. domesticus* and *M. m. musculus* are only partially reproductively isolated from each other and form a hybrid zone that represents a region of secondary contact and stretches across Europe from the Jutland peninsula to the Bulgarian coast of the Black sea (Boursot et al. 1993, Sage et al. 1993). The hybrid zone formed after the movement of *M. m. domesticus* into Western Europe within the last 3000 years (Cucchi et al. 2005). Several transects along the

2400 km length of the hybrid zone have been studied (e.g. Hunt and Selander 1973, Sage et al. 1986, Tucker et al. 1992, Prager et al. 1993, Dod et al. 1993, 2005, Payseur et al. 2004, Raufaste et al. 2005, Macholán et al. 2007, Teeter et al. 2007), two of which particular intensively, one in Denmark (e.g. Dod et al. 1993, 2005) and one in southern Germany near Munich (Tucker et al. 1992, Payseur et al. 2004, Payseur and Nachman 2005). The southern hybrid zone has an estimated width of 20 kilometers (Sage et al. 1986).

4.1.5 Aim of the study

The goal of the present study is to characterize individual loci that might be involved in the process of reproductive isolation (via hybrid unfitness) by analyzing hybrids from a natural hybrid zone of *M. m. domesticus* and *M. m. musculus* as well as pure subspecies. Both theoretical and empirical studies suggest that the regulation of gene expression may contribute to hybrid dysfunctions (Orr and Presgraves 2000**,** Ortíz-Barrientos). The hypothesis is that genes, which show expression divergence between the two subspecies, may have functional consequences in the hybrids and contribute to reproductive isolation. To uncover such regulatory incompatibilities, a microarray analysis has been performed to screen for genes which are differentially regulated between the two subspecies and their F1 hybrids, respectively (see chapter 2 for details). Two genes (Lsm10 and Neil3) were selected for further analysis, which exhibit a high fold change in expression in the testis between the two subspecies and show intermediate expression in laboratory-bred F1 hybrids. As gene expression divergence itself does not necessarily have a functional consequence and may also be selectively neutral, the hybrid zone analysis offers the opportunity to infer fitness costs – and therefore the contribution to reproductive isolation – associated with the expression change by estimating the patterns of differential introgression of expression phenotypes across the hybrid zone. Expression divergence of identified candidate genes may be a result of genetic drift or alternatively of positive selection due to adaptation to different habitats. Since the two subspecies diverged only recently, signatures of selective sweep events are expected to still be observable.

To understand and identify the evolutionary mechanisms and forces, which act on these genes, several population genetics and hybrid zone analyses were performed for the two candidate genes.

The aim of the analysis was to

1. identify changes which contribute to the expression difference of the candidate genes, and to elucidate if it is coupled with protein sequence evolution.
2. determine if the mutations that cause an expression difference between the two subspecies have a *cis*- or *trans*-regulatory basis.
3. assess if the expression difference is caused through an adaptive process.
4. infer fitness costs of the expression differences and the potential to contribute to reproductive isolation.

4.2 Materials and Methods

4.2.1 Animals

The population genetic analyses for the two subspecies *M. m. musculus* and *M. m. domesticus* included the same samples as for the microarray screen, respectively the follow-up confirmation (see 2.2.1, 3.2.2) as well as additional samples from different populations to get a complete picture of the factors involved in the subspecies/hybrid expression difference.

For *M. m. domesticus* unrelated mice have been collected in Germany (near Bonn, provided by M. Teschke) and in Iran (Ahvaz, provided by R. Scavetta), the latter representing a more ancestral population. *M. m. musculus* stem from Kazakhstan (Almati area, ancestral population), Czech Republic (near Námest and Oslavou) (both sampled and provided by S. Ihle) and Austria (Vienna, kindly provided by K. Musolf, Vienna). All mice were caught at least 300 m apart from each other, thus representing unrelated animals. It is known that for strictly indoor living mice home ranges are restricted to a few square meters (Berry and Bronson 1992, Pocock et al. 2005). Mouse traps were put up in private houses, barns or stables. For the German population, samples from ten, for the Austrian from three and for the other populations from eight different sites were used for analysis. One *Mus caroli* sample from Thailand was used for outgroup comparisons (kindly provided by A. Orth and F. Bonhomme).

To asses the genetic basis which underlies mutations that potentially cause expression differences (4.2.7) and to identify functional consequences for candidate genes

(4.2.8), a hybrid zone analysis was performed. For this analysis pure subspecies as well as natural hybrids between *M. m. domesticus* and *M. m. musculus* were used for expression and sequence analysis. These animals represent backcrossed hybrids and were directly collected in the hybrid zone in Bavaria in September 2005 referring to the sampling area that was investigated by R. Sage in 1985 (Sage et al. 1986, Payseur et al. 2004). To ensure standardized conditions for expression analyses F1 offspring generated in the laboratory was used. Breeding pairs with animals from the same locality were set up or if a pair was not available from the same spot, pairs were set up according to closest proximity. Altogether two male F1 offspring each of nine breeding pairs (representing 6 different localities) were obtained; additionally two pregnant females were directly caught in the field, which gave birth to one respectively three males. Furthermore five more males directly from the field were used for analysis. Overall 27 male hybrid animals were used, representing eight different localities. Supplement 13 shows the sampling sites relative to the hybrid zone. For the pure subspecies four pairs each of wild-caught animals of *M. m. domesticus* from Germany and *M. m. musculus* from Austria were used to obtain F1 offspring in the laboratory (at least one male offspring per locality). All mice were held under standard laboratory conditions and male mice were sacrificed at the age of 6-8 weeks. Supplement 12 and Supplement 13 list the characteristics of the sampling sites and give an overview of the used samples.

4.2.2 Sample preparation

Animals used for the expression analysis were sacrificed using CO_2. Organs were excised and immediately snap frozen in liquid nitrogen. Organs were stored at -80 degrees. RNA from testis was extracted using Trizol® (Invitrogen, Carlsbad, CA) following the manufacturer's protocol. Quality and integrity of the total RNA was controlled by using the Agilent Technologies 2100 Bioanalyzer and the RNA 6000 Nano LabChip® Kit (Agilent Technologies Waldbronn, Germany). Only samples with RNA integrity numbers (RIN) >8.0 were used for analysis. DNA was extracted in 15 ml falcon tubes by standard salt extraction procedures. Dried pellets were dissolved in 500 to 1000µl of 1xTE.

4.2.3 cDNA synthesis

RNA was DNase-treated prior to cDNA synthesis using Ambion's DNA-*free*™

following the manufacturer's protocol. RNA was reverse transcribed using random hexamers (Fermentas) and the ThermoScript Revese Transcriptase Kit (Invitrogen, Carlsbad, CA) using 1 µg RNA as starting material according to the manufacturer's protocol.

4.2.4 DNA sequencing

The sequence of for both Lsm10 (NM_138721) and Neil3 (NM_146208) was downloaded from ENSEMBL (http://www.ensembl.org). For Lsm10 only exon 2 is protein-coding (369 bp). Therefore this gene was sequenced from genomic DNA for 20 *M. m. domesticu*s and 24 *M. m. musculus* samples. For Neil3, amplification from cDNA was performed for several fragments solely for the German *M. m. domesticus* and Austrian *M. m. musculus* as only for those samples cDNA was available. Fragments which could not be amplified from cDNA, potential due to alternative splicing, were amplified from genomic DNA. Additionally, for both genes 300-600 bp long fragments of non-coding regions were sequenced (four fragments for Lsm10 and two for Neil3) for all subspecies samples. Supplement 14 gives detailed information about the primers and fragments' position in relation to the location of the genes.

PCR was setup in a 10 µl volume using the Qiagen® Multiplex PCR Kit following the manufacturer's protocol with 1 µl of 1:10 diluted cDNA as starting material, respectively 10 µg of genomic DNA. PCR products were purified by adding 0.012 µl Exonuclease I (Biolabs® 20 U/ml) and 0.045 µl Shrimp Alkaline Phosphatase (Promega, 1U/µl) per 10 µl reaction and heating 20 minutes for 37ºC followed by 72ºC for 15 minutes. PCR products were sequenced in both directions using the BigDye sequencing chemistry on an ABI3730 automated sequencer. All sequences are provided in the Digital Supplement.

4.2.5 Gene expression analysis

Affymetrix GeneChip® analysis

Expression profiles for testis from four *M. m. domesticus*, four *M. m. musculus* (Supplement 12) and 27 hybrids (Supplement 16) were determined for over 39,000 mouse transcripts using the Mouse Genome 430 2.0 Affymetrix GeneChip®. The biotin-labeled target synthesis started from 1 µg of total RNA following the Ambion Message Amp II aRNA amplification Kit protocol. After hybridization the GeneChips® were washed, using an Affymetrix GeneChip® fluidic station 450, stained with SA-PE (Streptavidin R-phycoerythrin conjugate) and read using an Affymetrix GCS 300 G scanner. Raw signal intensities were normalized and summarized according to the standard Affymetrix MA Suite 5.0 algorithm using the affy-package of Bioconductor (http://www.bioconductor.org/).

Quantitative qRT-PCR using Sybr Green

For *M. caroli* the expression levels for Lsm10 and Neil3 were achieved through qRT-PCR analysis. To have an estimate for the expression level, two samples each of *M. m. domesticus* and *M. m. musculus* were also analyzed. QuantiFast® Sybr Green PCR Kit (Qiagen) was used in a 10 µl reaction volume according to the manufacturer's protocol with 2 µl of a 1:40 dilution of cDNA as starting material. Samples were performed in triplicate and averaged C_T values from each qRT-PCR reaction of the target gene were standardized relative to the endogenous control, Eif4g2, of the same sample, yielding the ΔC_T value. Eif4g2 (eukaryotic translation initiation factor 4, gamma 2) was used as endogenous control as it showed highly similar expression levels and the lowest standard deviation across individuals, subspecies and different tissues in the GeneChip® analysis, including three different subspecies and tissues, respectively (data not shown). Primers were designed ensuring that the primer sequence contained no difference between the species/subspecies (primers are listed in Supplement 7). Melting curve analysis as well as agarose gel electrophoresis was done to check for the specificity of the PCR products. Experiments were run on an ABI Prism 7900HT Sequence Detection System following the provider's recommendation concerning the thermal cycling protocol, using 384 well plates and running SDS 2.1.1. For each qRT-PCR reaction the averaged C_T value of the endogenous control was subtracted from the averaged C_T value of the target gene, yielding the ΔC_T value. These values were then averaged across the triplicate samples within an individual, yielding the average ΔC_T value. Outliers among triplicates were removed if they raised the standard deviation above 0.3.

4.2.6 Sequence data analysis

Sequences assembly and analysis was performed using the program CodonCode Aligner 2.0. Heterozygous positions were identified through visual inspection of all sequence positions for which the automatic nucleotide calls were ambiguous. All sequences were aligned using CLUSTAL W (Thompson et al. 1994). Sequence analyses were performed using tests implemented in DnaSP 4.10.9 (Rozas et al. 2003). To deal with heterozygous positions, all sequences were duplicated, and each nucleotide state was randomly assigned to one of the duplicated sequences.

Analysis of the coding sequence

To test whether the expression difference between the two subspecies is coupled with accelerated protein evolution, the coding sequence of Lsm10 and Neil3 for the two subspecies and *M. caroli* was analyzed.

The McDonald Kreitman test (McDonald and Kreitman 1991) compares synonymous and replacement substitutions within and between species, assuming that under neutrality the ratio of divergence to polymorphism should be the same in both synonymous and replacement sites. Furthermore, the ratio of synonymous to replacement sites should be the same in the divergence and the polymorphic categories. Deviations are considered as deviations from neutrality and indicative for selective forces acting on the protein.

Ka/Ks analyses assess positive or negative selection in protein coding sequences by comparing codons and comparing the rate of amino acid substitutions to the rate of synonymous substitutions. Under neutrality, the number of non-synonymous changes at each possible non-synonymous site is the same as the number of synonymous changes per synonymous site; that is, Ka/Ks = 1. Deviations from 1 give hints to the selective forces acting on the protein. Ka/Ks ratio< 1 indicates purifying selection with non-synonymous changes being less frequent than synonymous and Ka/Ks >1 hints to positive selection with non-synonymous changes being more frequent (Hurst 2002, Fay and Wu 2003). *M. m. domesticus* was compared to *M. m. musculus* and both subspecies to *M. caroli*. The Ka/Ks test is often assumed to be too conservative in detecting proteins that have evolved under positive selection between two species (Fay and Wu 2003). Some regions of constraint within a protein are likely maintained during the evolution of a new or improved function of a protein, which will lower the overall rate of amino acid substitution within a protein below the neutral rate unless the adaptive regions evolve at a rate fast enough to bring the average Ka/Ks of the entire protein above one. To get an estimate if different regions of the protein exhibit differences in selective pressure, a sliding window analysis for the genes studied here, has been performed. Windows of 100 sites width with steps of 10 bp were used. By sequencing analysis of *M. caroli* as outgroup, from which the *M. musculus* subspecies were separated about three million years ago (Guénet and Bonhomme 2003), the ancestral or derived state of nucleotide polymorphisms can be inferred and conclusions about the behavior of newly derived mutations and their spread throughout the genome are made possible.

Analysis of non-coding sequence

Evidence for a recent selective event, either in the protein coding or the regulatory region, can be obtained indirectly through the analysis of the flanking, non-coding region of a gene. If the action of positive selection can be demonstrated in a chromosomal region where a differentially expressed gene is located, it might be that the expression change has a functional consequence to the organism (Harr et al. 2006). To uncover potential signatures of adaptation, the 5' upstream region of both candidate genes was analyzed. Several basic population genetics parameters were estimated to asses the within population diversity in 5' upstream regions, namely Watterson's θ_W (Watterson 1975), π (Nei and Li 1979) and Tajima's D. The Tajima's D-statistic (Tajima 1989) evaluates low-frequency and intermediate frequency sites in a sample. Negative values indicate population expansion, strong purifying selection or recovery after a selective sweep. Positive values result from an excess of polymorphisms and may indicate the presence of population structure, balancing selection, or weak/incomplete bottlenecks (Ometto et al. 2005).

Analysis was performed by considering

1. samples from different populations of a subspecies as a functional unit, since F_{st} analysis (Hudson et al. 1992) of seven autosomal loci (data from Baines and Harr 2007) reveals no general high genetic differentiation between the populations (data not shown) and

2. populations of subspecies separately to disentangle potential population effects from subspecies effects.

The different fragments of the sequenced upstream region were concatenated for all analyses and furthermore considered solely as concatenation results in loss of samples size as sequencing of some samples failed for some fragments.

4.2.7 Correlation of expression-phenotype and genotype – determination of *cis*- or *trans*-acting factors

To determine if the differential expression is caused by a change in *cis* or in *trans*, a correlation analysis between the expression for the two candidate genes and the corresponding genotype at the given locus of natural hybrids (which represent backcrosses of different degrees) was performed. The rationale is that if an expression difference is solely caused by a *cis*-mutation, a hybrid's expression level should correspond to the underlying genotype at that locus. Particular meaningful for this analysis are samples, for which the locus-specific genotype and expression level departures from the overall hybrid-genotype (e.g. overall *domesticus* genotype shows *musculus* locus-specific genotype and expression level). For samples, where both – locus-specific and overall genotype – are identical, also *trans*-acting factors could have caused corresponding genotype and expression level. For comparison reasons four animals each of the pure subspecies and laboratory F1 hybrids were also included in this analysis. Prerequisite for this analysis is the identification of a diagnostic difference (SNP) between the two subspecies in the region of interest. Supplement 15 lists the primer sequences for fragments with fixed differences between *M. m. domesticus* and *M. m. musculus*. These fragments were then sequenced for all 27 Bavarian samples which have been used for the expression analysis (refer to Digital Supplement for sequence results) for both candidate genes, and the hybrid's locus-specific genotype was associated with the expression levels for the respective gene as derived from the Affymetrix GeneChip® analysis.

The overall genotype or "hybrid index" is calculated as the "*musculus*"-allele frequency per sample in comparison to "*domesticus*"-alleles, averaged over several loci and indicates, to which degree a hybrid consists of which genome. A hybrid index of 0 refers to a pure *domesticus* genotype, whereas a hybrid index of 1 indicates a pure *musculus* genotype. SNP genotyping analysis with the Beckman Coulter GenomeLab SNPstream platform was performed following standard protocols to estimate the overall genotype for 63 Bavarian hybrid samples (including the 27 samples used for expression and follow-up analysis, Supplement 16) by using 14 loci with fixed differences between *M. m. domesticus* and *M. m. musculus* derived from a reference data set (Harr, Ihle, Rottscheidt, Scavetta, Teschke, unpublished data). Allele frequencies were summed up over all loci per samples, yielding the overall hybrid index per sample.

4.2.8 Assessment of functional consequences for candidate genes from hybrid zone analysis

Hybrid zone analysis is particular useful to get an estimate for fitness cost associated with a differentially expressed gene. Expression differences between subspecies might have negative fitness consequences when brought together in a hybrid genome. For genes which contribute to reproductive isolation (reduced fitness), thus having a deleterious effect in hybrids, subspecies-specific alleles are considered to introgress at lower rates across a hybrid zone than genes with neutral or adaptive effect.

If considering gene expression changes caused by *cis*-acting factors, this analysis can be performed on a genomic basis. A diagnostic SNP between the two parental subspecies (with close connection to the candidate gene) is used as a representative for the unknown mutation, which causes the expression difference. The genotype for the particular candidate gene, derived from the SNP analysis, is related to the overall hybrid index of an animal. A deviation of a candidate locus-genotype from the overall hybrid-index gives an estimate for positive or negative fitness effect of the gene. A higher frequency of a locus specific genotype gives a hint to elevated introgression, associated with a potential adaptive effect while absence of a specific genotype indicates negative effects of that genotype, i.e. an isolating factor. For this “accordance analysis” samples were attributed as pure “*musculus*-like”, “*domesticus*-like” or “*hybrid*-like” for reasons of clarity rather than depicting the correct allele frequencies. This analysis was performed for the 27 hybrids used for the Affymetrix analysis. For candidate genes with unclear genetic basis, fitness costs have been deduced from relating the “expression phenotype” (i.e. the expression level) of the respective gene to the overall hybrid index of the samples.

4.3 Results

4.3.1 Gene expression analysis using Sybr Green

Two samples of each *M. m. domesticus* and *M. m. musculus* and one sample of *M. caroli* have been analyzed via qRT-PCR for Lsm10 and Neil3. Table 4.1 shows the averaged ΔC_T values (relative to the endogenous control) for the samples. Lsm10 shows low expression in *M. caroli*, almost identical to the expression level of *M. m. domesticus*. The expression level from Neil3 is also low in *M. caroli*, comparable to the expression level of *M. m. musculus*.

Table 4.1: Averaged relative expression levels from qRT-PCR using Sybr Green for *M. m. domesticus*, *M. m. musculus* and *M. caroli*. Note that low values refer to high expression levels.

	Relative expression level (ΔC_T)		
	M. m. domesticus	*M. m. musculus*	*M. caroli*
Lsm10	4.5	0.71	4.47
Neil3	3.35	8.3	9.5

4.3.2 Sequence analysis

Analysis of the coding sequence

Lsm10

The McDonald-Kreitman test could not be performed between *M. m. domesticus* and *M. m. musculus* as no polymorphic sites in the coding region are present in both subspecies. A single fixed synonymous substitution between the two subspecies was identified. No non-synonymous substitution was identified; hence the result of the Ka/Ks analysis is 0.

In the comparison between *M. m. domesticus*, *M. m. musculus* and *M. caroli*, six respectively five synonymous substitutions and one non-synonymous substitution are present, resulting in a Ka/Ks ratio of 0.052 and 0.063, respectively.

Neil3

The McDonald-Kreitman test revealed no significant result for the *domesticus-musculus* comparison (G-value = 0.895, p = 0.34421), indicating that the divergence between the two subspecies has a neutral basis (Table 4.2).

Table 4.2: Number of replacement and synonymous substitutions for fixed differences between subspecies and polymorphism within subspecies for Neil3.

	fixed differences	**polymorphism**
Synonymous	2	2
Non-synonymous	4	12

For Neil3 an overall Ka/Ks ratio of 0.933 was calculated for the *domesticus-musculus* comparison. As an overall Ka/Ks ratio of >1 as an indicator of positive selection is assumed to be very conservative (Fay and Wu 2003), this result may hint to positive selection on specific regions. The sliding window analysis did not identify any region with a Ka/Ks ration >1 (Figure 4.1), therefore the Ka/Ks test yielding the same result as the McDonald-Kreitman test.

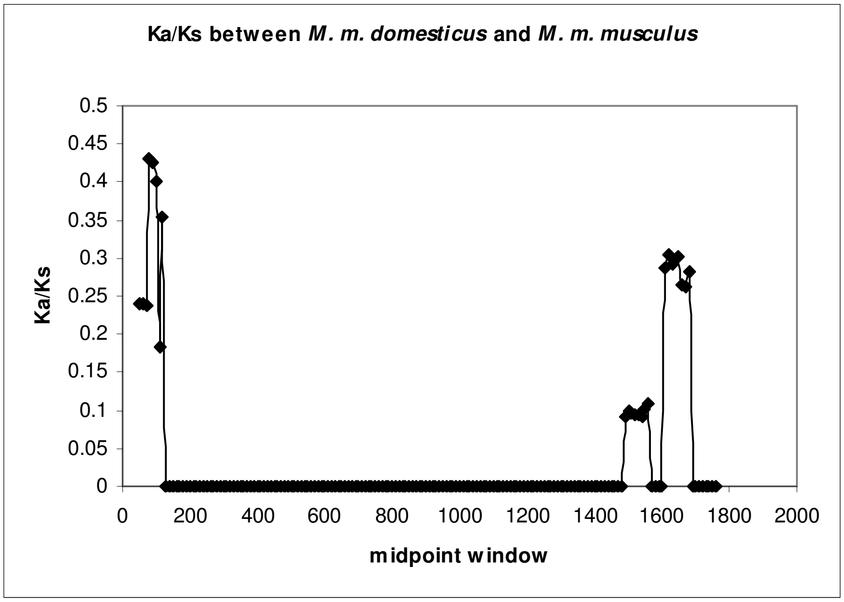


Figure 4.1: Sliding window Ka/Ks analysis for Neil3 between *M. m. domesticus* and *M. m. musculus*. Windows of 100 bp size were used with a step size of 10 bp.

In the comparison between *M. caroli* and *M. m. domesticus* 26 synonymous and 25 non-synonymous substitutions occur, 24 synonymous and 27 non-synonymous between *M. caroli* and *M. m. musculus*. The Ka/Ks analysis between *M. caroli*, *M. m. domesticus* and *M. m. musculus* yields Ka/Ks values considerably lower than that of the *M. m. domesticus* and *M. m. musculus* comparison with Ka/Ks = 0.325 and Ka/Ks = 0.342. In contrast, the sliding window analysis revealed a region with a drastic peak in divergence with Ka/Ks ratio of 2.13 (midpoint window 1380) in both comparisons (*musculus* vs. *caroli* and *domesticus* vs. *caroli*), which may be a hint towards positive selection for this region (Figure 4.2 and Figure 4.3). Also a second region shows Ka/Ks ration higher than one (midpoint window 870-900).

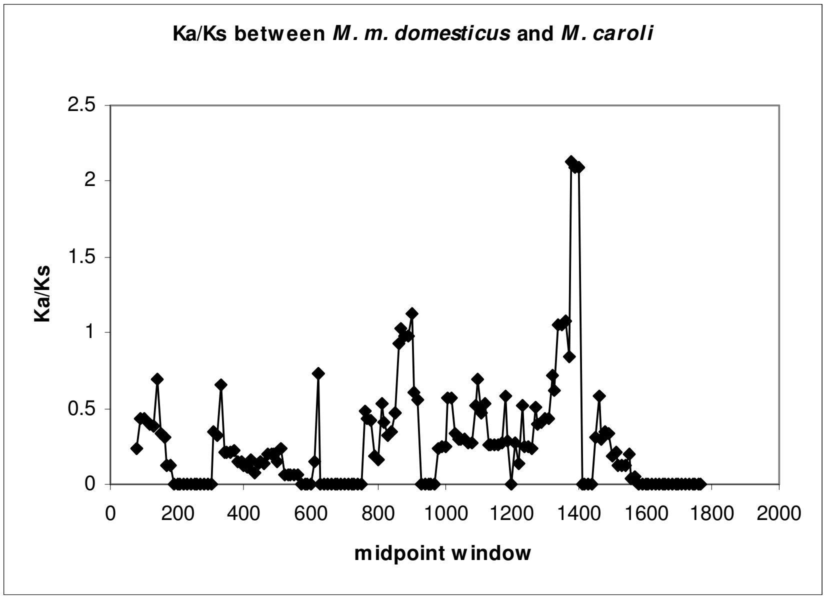


Figure 4.2: Sliding window Ka/Ks analysis for Neil3 between *M. m. domesticus* and *M. caroli*. Windows of 100 bp size were used with a step size of 10 bp.

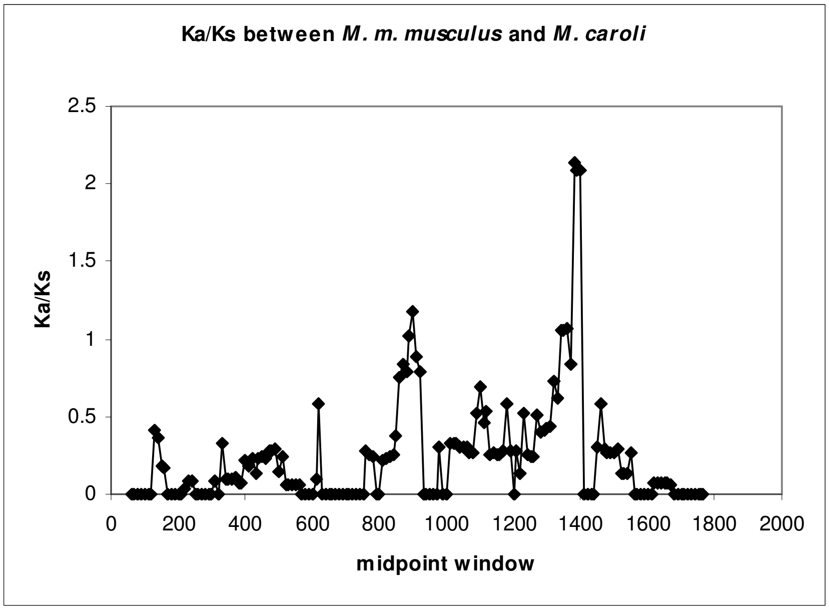


Figure 4.3: Sliding window Ka/Ks analysis for Neil3 between *M. m. musculus* and *M. caroli*. Windows of 100 bp size were used with a step size of 10 bp.

Analysis of non-coding sequence

Lsm10

Comparing nucleotide polymorphism in *M. m. domesticus* and *M. m. musculus* for the concatenated fragments of the upstream region of Lsm10 reveals general low values for both subspecies (Table 4.3). π and θ_W are considerably lower in *M. m. domesticus* than in *M. m. musculus*. As the analysis with the concatenated fragments experience a loss of power due to the reduced sample size (not all fragments were successfully sequenced for all samples) the analysis was repeated with the single fragments. The overall outcome can be confirmed with the single analyses: π and θ_W are general low for *M. m. domesticus* and higher for *M. m. musculus*. Altogether several fixed differences between the two subspecies are evident. Seven differences can be assumed to be fixed (sum of differences from fragment 1 and 3). The direct comparison of the single fragments with the concatenated underlines the potential of false assumptions due to differences in sampling size, with reductions in sampling size increasing the potential to assign polymorphic sites with relatively high frequencies as monomorphic in a population sample. Thus, it is clear that 13 fixed differences (as deduced from the concatenated fragment analysis) is an overestimation and seven differences are more likely a realistic number. However, the remaining five differences are present at very high frequencies.

To test for a selective sweep hypothesis, Tajima's D analysis was performed. For none of the subspecies, irrespective of whether the concatenated fragment was analyzed or the separate fragments (with larger sample size), a significant result was obtained, which indicates that evolution of this region is in accordance with a neutral model. Although the overall nucleotide polymorphism is higher in *M. m. musculus*, negative Tajima's D values are observed, which implies that there are more rare mutations than expected by the number of segregating sites. *M. m. domesticus* shows positive values for all fragments except for fragment 2 with slightly negative value. The overall outcome of the subspecies-wide analysis also shows up for the analysis of the separate populations of the two subspecies (Supplement 18). Nucleotide diversity is noticeably higher in *M. m. musculus* populations than in the *M. m. domesticus* populations, the latter showing no polymorphism at all for single fragments. In general the *domesticus* populations do not differ substantially in the rate of polymorphism. For the *musculus* populations slightly higher rates of polymorphism are observed in the Kazakh population for the concatenated fragment. Looking at the single fragments shows that higher nucleotide polymorphism is no general feature of the samples from Kazakhstan. It is due to a bias, caused through reduction in the sample size by concatenating the single fragments. The sample set from Vienna exhibits no polymorphic sites. As this sample set is very small and includes only three different sample localities, this result can be most likely attributed to a sampling bias.

Table 4.3: Nucleotide polymorphism in the 5' upstream flanking region of Lsm10 in *M. m. domesticus* and *M. m. musculus*. Values are given for three fragments separately and fragments concatenated, respectively.

Fragment	1	2	3	concatenated fragments
Relative distance to locus (kb)	4.8	4.2	2.6	
Number of sites	441	1136	555	2132
Number of shared mutations	0	0	0	0
Number of fixed differences between subspecies	2	0	5	13
M. m. domesticus				
Number of chromosomes	34	32	36	28
Number of segregating sites	1	2	1	4
π	0.00102	0.00042	0.00044	0.00058
θ_W per site	0.00055	0.00044	0.00043	0.00048
Tajima’s D	1.2644	-0.09237	0.0298	0.53505
Significance	>0.10	>0.10	>0.10	>0.10
M. m. musculus				
Number of chromosomes	40	44	38	36
Number of segregating sites	5	10	1	9
π	0.00186	0.00096	0.00035	0.00073
θ_W per site	0.00267	0.00202	0.00043	0.00102
Tajima’s D	-0.77398	-1.54526	-0.27123	-0.83286
Significance	>0.10	>0.10	>0.10	>0.10

Neil3

The analysis of two 5’ upstream fragments of Neil3 reveals overall higher nucleotide diversity than for Lsm10 for both subspecies (Table 4.4). Averaged π of the concatenated fragments is 0.00206 for *M. m. domesticus* and 0.00577 for *M. m. musculus*. θ_W is 0.00298 for *domesticus* and 0.00486 for *musculus*, respectively. π and θ_W are approximately three times lower in *M. m. domesticus* than in *M. m. musculus* irrespective of considering the concatenated sequence or the two fragments separately

Table 4.4: Nucleotide polymorphism in the 5' upstream flanking region of Neil3 in *M. m. domesticus* and *M. m. musculus*. Values are given for two fragments separately and fragments concatenated, respectively.

Fragment	1	2	concatenated fragments
Relative distance to locus (kb)	3.6	2.6	
Number of sites	538	340	878
Number of shared mutations	1	4	5
Number of fixed differences between subspecies	0	0	0
M. m. domesticus			
Number of chromosomes	30	38	30
Number of segregating sites	4	7	10
π	0.00151	0.0029	0.00206
θ_W per site	0.00188	0.0049	0.00298
Tajima’s D	-0.49814	-1.14312	-0.97787
Significance	>0.10	>0.10	>0.10
M. m. musculus			
Number of chromosomes	34	44	34
Number of segregating sites	6	11	17
π	0.0032	0.00999	0.00577
θ_W per site	0.00284	0.00744	0.00486
Tajima’s D	0.34904	0.9827	0.62502
Significance	>0.10	>0.10	>0.10

No fixed differences between the two subspecies are evident in the upstream region but five shared mutations. Reduction in sample size due to concatenation of the fragments has no obvious effect in terms of assigning polymorphism at high frequency as fixed differences between the subspecies. Tajima's D results in negative values for *M. m. domesticus* and positive for *M. m. musculus* for all analyzed sequence fragments, but significant results could not be obtained for either of the subspecies and fragments. Considering the populations of the two subspecies separately (Supplement 19), higher nucleotide polymorphism in the Iranian than the German population is evident. Nucleotide polymorphism data show no significant difference between the samples from Kazakhstan and the Czech Republic. Samples from Vienna show no nucleotide polymorphism at all – again probably due to the very limited population sample.

4.3.3 Correlation of expression-phenotype and genotype – determination of *cis*- or *trans*-acting factors

Figure 4.4 and Figure 4.5 show the expression levels from the GeneChip® analysis for the two candidate genes for 27 individuals from the Bavarian hybrid zone as well as four samples each of the pure subspecies and laboratory F1 hybrids in relation to the locus-specific genotype of the samples. Samples were assigned as "*domesticus*", "*musculus*" or "*hybrid*" depending on the diagnostic allele for the specific locus (see Supplement 17). Pure subspecies and laboratory F1 hybrids are also depicted such that the signal intensities of the Bavarian individuals can be directly compared to them.

For Lsm10 the expression (low like *domesticus*, high like *musculus* and intermediate like *hybrids*) of the Bavarian samples correlates perfectly with the locus-specific genotype, i.e. the species-specific SNP is indicative of the expression for that locus and vice versa. 22 of the samples show a *musculus*-like genotype/expression (i.e. SNP/expression), two a *domesticus*-like and three a *hybrid*-like status. Expression of Lsm10 in the Bavarian hybrids is exactly in the range of the pure subspecies expression (refer to Supplement 17 for details). Hybrids carrying the *musculus*-SNP are high, those carrying the *domesticus*-SNP are low and heterozygotes are intermediate in expression. Of the 27 samples, most show an overall genotype of *musculus*; only eight have an average *domesticus* genotype. Three of these show high expression according to their *musculus* locus-specific genotype and three intermediate in accordance to the heterozygous genotype (Supplement 16, Supplement 17). These six animals are particularly useful for the analysis, since it can be excluded that the expression is caused through a *trans*-acting factor derived from the *domesticus* genome, which would have led to lowered expression. Hence, it is likely that a pure *cis*-effect is responsible for the expression difference between the subspecies.

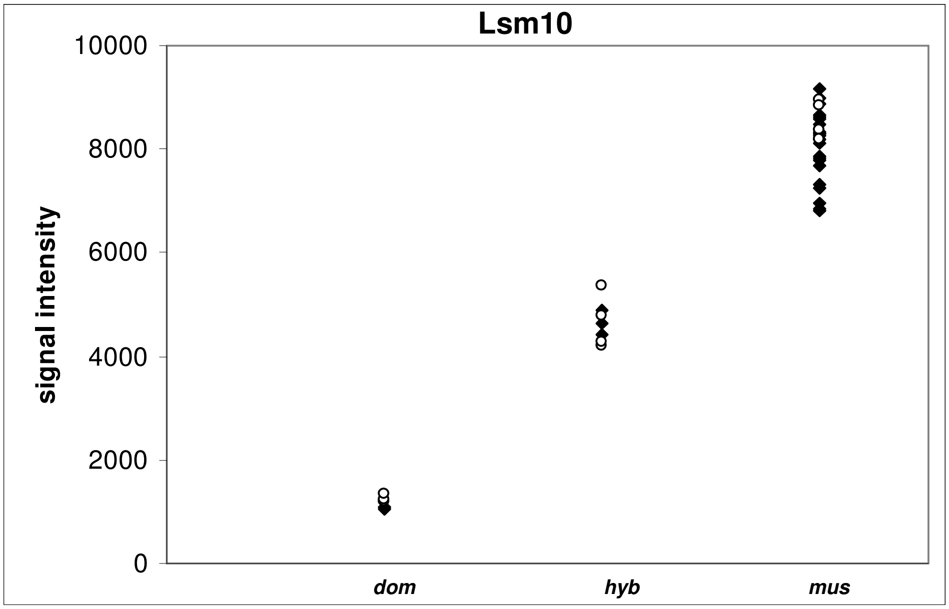


Figure 4.4: Correlation between Affymetrix GeneChip® signal intensities and genotype for Lsm10. *Dom* refers to animals homozygous for the *domesticus* SNP, *mus* to animals homozygous for the *musculus* SNP and *hyb* animals are heterozygous. Pure *domesticus*, *musculus* and laboratory F1 hybrids are indicated by circles.

For Neil3 the expression-phenotype does not exactly correspond to the locus-specific genotype. Four of 14 *domesticus*-like genotypes do not show *domesticus*-like expression (high expression) and are intermediate expressed; one animal with heterozygous genotype exhibits *domesticus*-like (high expression) (indicated by diamonds in Figure 4.5). Solely for the *musculus*-like genotypes a perfect correlation of genotype and expression level was found. For 15 of the samples the overall genotype is not identical with the locus-specific genotype. 13 animals with overall *musculus* genotype exist, eight of which have a *domesticus* locus-specific genotype and five are heterozygous. For four of the eight, the expression is high (*domesticus*-like) and for four intermediate (*hybrid*-like). Two samples show an average *domesticus* genotype, both heterozygous for the locus-specific SNP. One sample exhibits high expression and the other intermediate (Supplement 16, Supplement 17). Thus, for five samples the expression does not follow the locus-specific genotype.

For this analysis, animals showing signal intensities higher than 600 have been considered as *domesticus*-like, since the pure *M. m. domesticus* animal with the lowest expression exhibits a signal intensity of 609. Overall, *hybrid*-like and *domesticus*-like expression levels are not as clearly differentiated as *hybrid*-like and *musculus*-like*,* the hybrids showing intermediate expression but with a tendency towards *domesticus*-like expression levels. A similar tendency was observed by the qRT-PCR analysis from chapter 3.3.1 and 3.3.2, where *M. m. domesticus*, *M. m. musculus* and the artificial F1 hybrids all were significantly differentially expressed, but the hybrids showing expression levels closer to *M. m. domesticus*.

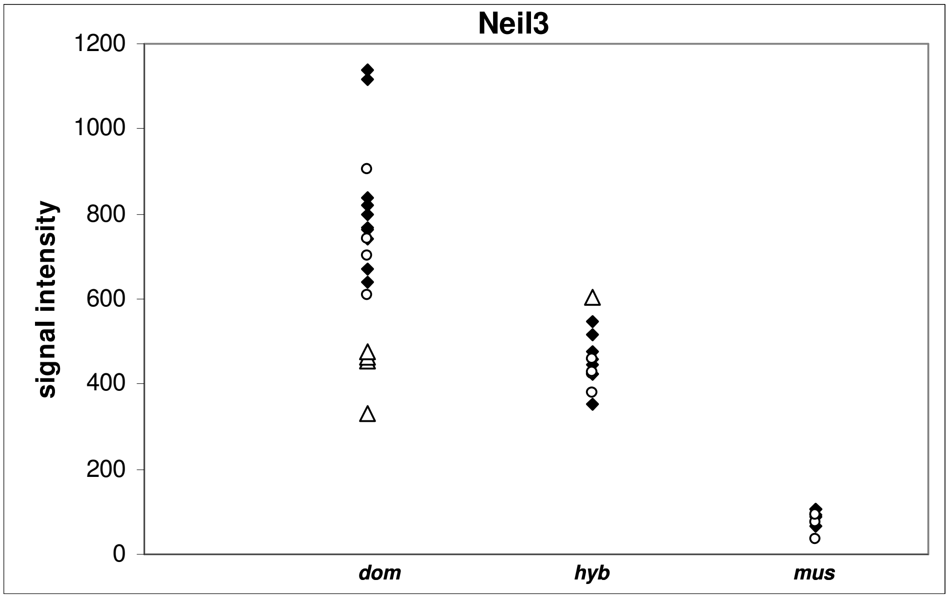


Figure 4.5 Correlation between Affymetrix GeneChip® signal intensities and genotype for Neil3. *Dom* refers to animals homozygous for the *domesticus* SNP, *mus* to animals homozygous for the *musculus* SNP and *hyb* animals are heterozygous. Animals which show genotype-phenotype correlation are indicated by diamonds, those with phenotype-genotype deviation by triangles. Pure *domesticus*, *musculus* and laboratory F1 hybrids are indicated by circles.

4.3.4 Assessment of functional consequences for candidate genes from hybrid zone analysis

Figure 4.6 shows the overall genotype (average over 14 loci) of 63 Bavarian individuals in relation to the geographic position they are derived from. Samples are ordered from west to east in the figure, the center of the hybrid zone position has been placed referring to the estimation from Sage et al. (1986) (also see Supplement 13). Overall 19 samples collected western and 44 eastern of the center of the hybrid zone have been genotyped. The typed animals represent hybrids with complex ancestry, no animal heterozygous for all loci (indicative of F1 hybrid) has been found. No animal, homozygous for all typed loci was found at the "wrong" side of the hybrid zone. One can clearly see the gradual trend in the genotypes, samples changing from *domesticus*-like to more and more *musculus*-like the more eastern the sample has been collected. Only three pure *domesticus* western of the center of the zone and one pure *musculus* eastern of the center have been genotyped.

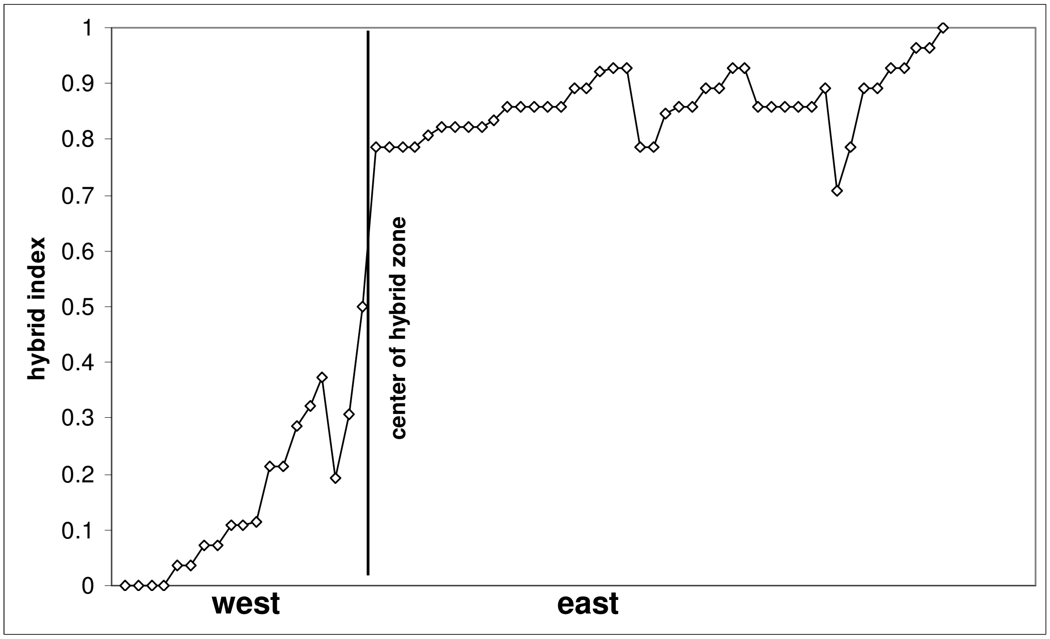


Figure 4.6: Overall hybrid index for all Bavarian individuals from the hybrid zone. A transect of approximately 20 km width has been sampled; the vertical line represents the center of the hybrid zone according to Sage (1986). Hybrid index = 0 refers to pure *domesticus*-genotype, 1 to pure *musculus*-genotypes. Individuals are sorted from the most western to the most eastern sample locality.

In Figure 4.7 the locus-specific hybrid indices (genotypes) for Lsm10 of the 27 Bavarian samples, which were used for the expression analysis are depicted in relation to the geographic position the samples are derived from. All samples in the eastern part of the hybrid zone show the *musculus*-like genotype, consistent with the overall hybrid index. In the western part of the hybrid zone, all but two samples have are at least heterozygous if not homozygous for the *musculus*-allele, indicating introgression of the *musculus*-allele into the *domesticus* genomic background. As the expression levels of the hybrids correlate perfectly with the genotypes (Figure 4.4), this means that all but two hybrids exhibit at least intermediate if not high expression of the transcript, thus the expression-phenotype *musculus*-like or *hybrid*-like moving westwards. Intermediate expression occurs only in three animals, in animals which have an overall genetic composition of *domesticus*.

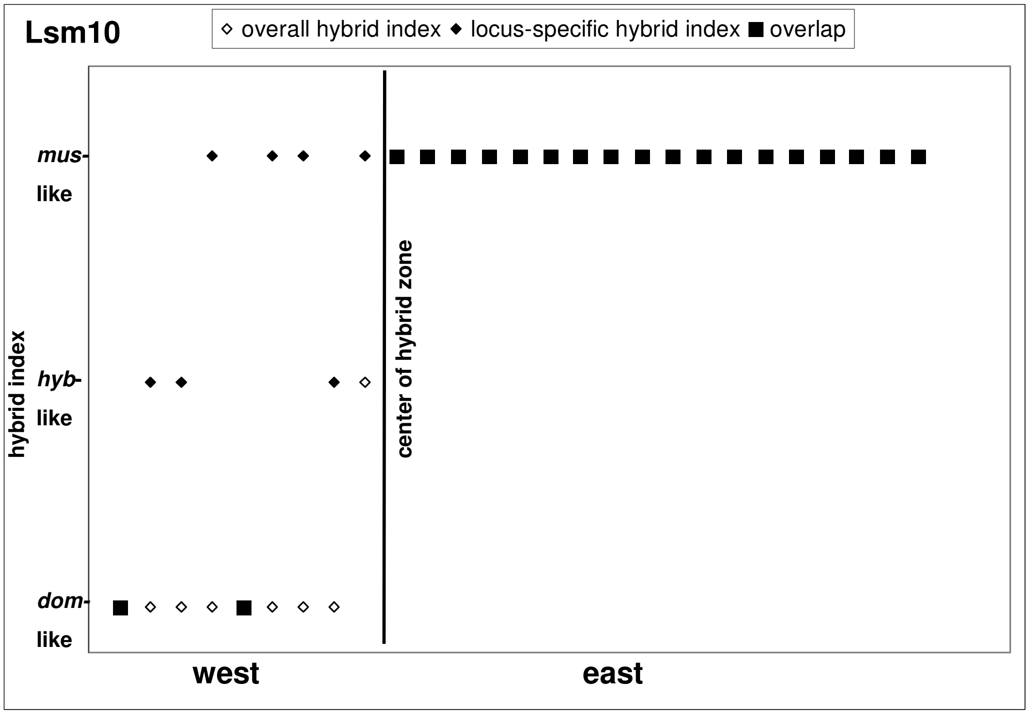


Figure 4.7: Locus-specific hybrid index for Lsm10 in relation to the overall hybrid index. Individuals are sorted from the most western to the most eastern sample locality. Samples for which the overall hybrid index overlaps with the locus-specific hybrid index are indicated in squares. Note that the categories on the y-axis correspond to the expression level (low, intermediate, high) in the respective pure subspecies/F1 hybrid.

As for Neil3 the expression does not correspond 100% to the genotypes of the hybrids (i.e. the change in expression is not due to a *cis* difference) (Figure 4.5), a direct deduction about fitness effects of the respective expression phenotypes (range over the hybrid zone) cannot be made from the genotypes. Therefore the signal intensities (representing the expression phenotype) of all hybrids have been related to the overall hybrid indices (Figure 4.8). For pure subspecies the high expression of this transcript is observed for M. *m. domesticus*, while *M. m. musculus* is very low expressed. The expression of the hybrids across the hybrid zone in most cases also is high (*domesticus*-like) or intermediate (*hybrid*-like). Only four animals show the low expression phenotype of *musculus*. As more samples eastern of the center of the hybrid zone have been collected (with overall hybrid indices *musculus*-like), the occurrence of that many *domesticus*-like expression levels as well as the intermediate hybrid expression level across the hybrid zone suggest introgression of the *domesticus* allele into the *musculus* genomic background.

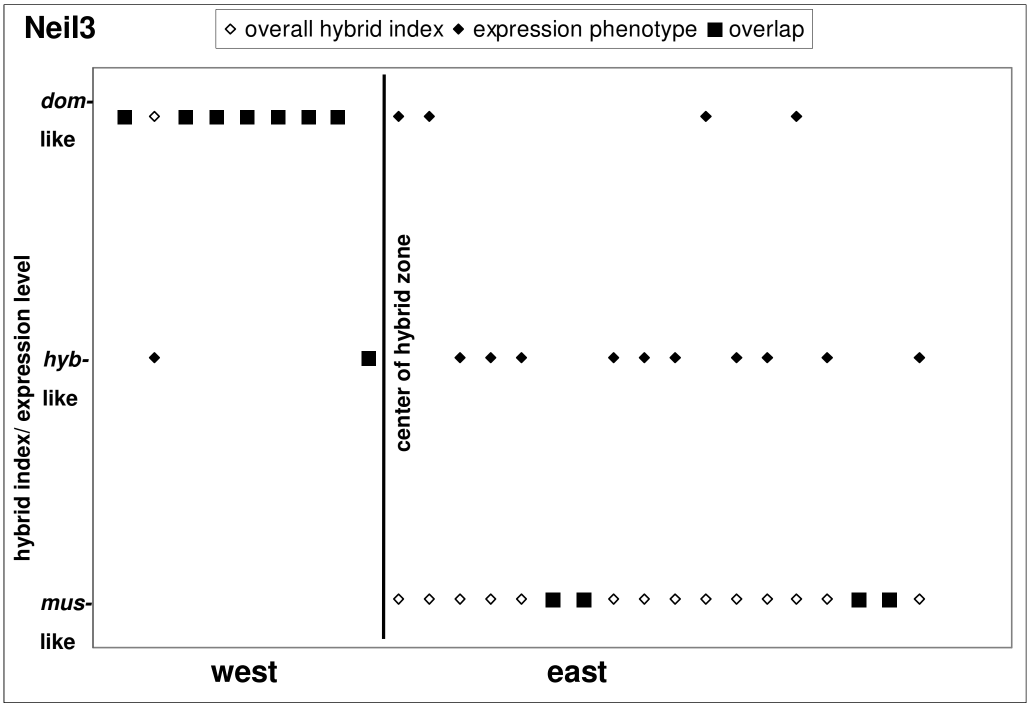


Figure 4.8: Expression phenotypes for Neil3 in relation to the overall hybrid index. Individuals are sorted from the most western to the most eastern sample locality. Samples for which the overall hybrid index overlaps with the locus-specific expression phenotype are indicated in squares. Note that the categories on the y-axis correspond to the expression level (low, intermediate, high) in the respective pure subspecies/F1 hybrid.

4.4 Discussion

The two genes, Lsm10 and Neil3, which have been characterized in detail in this analysis, showed up as differentially expressed between *M. m. domesticus* and *M. m. musculus* in an expression screen via Affymetrix GeneChip®. Laboratory-bred F1 hybrids between the two subspecies are intermediate in expression. A follow-up validation via qRT-PCR confirmed the expression differences (refer to chapter 2 and 3 for details). These genes provide candidates which may be involved in the speciation process in the house mouse by contributing to reproductive isolation. Population genetic and hybrid zone analyses have been applied to uncover the genetic basis and the potential evolutionary forces that led to the differential expression levels and to evaluate the potential of these genes to be involved in reproductive isolation.

4.4.1 Protein coding versus regulatory evolution

Lsm10

Lsm10 (U7 snRNP-specific Sm-like protein), located on chromosome 4 (125,774 kb - 125,776 kb), codes for a short protein and possesses two exons with exon 2 being protein coding only. It plays, like all eukaryotic Sm and Sm-like proteins, an important role in mRNA metabolism (Pillai et al. 2001). They associate with RNA to form the core domain of the U7 small ribonucleoprotein particle (snRNP). SnRNPs are considered as *trans*-factors, which are involved in a variety of RNA processing events including pre-mRNA splicing, teleomere replication and mRNA degradation (Smith et al. 1991, Müller et al. 2000).

The sequence analysis of the protein-coding region of Lsm10 shows that the expression change between *M. m. domesticus* and *M. m. musculus* is not accompanied by rapid evolution of the protein. No non-synonymous change was identified, between the two subspecies and only one fixed synonymous difference. As synonymous changes are assumed to be neutral (Fay and Wu 2003), the gene is identical on the protein-coding level, which indicates strong selective constraint on the coding region. This is further supported by the fact that not a single polymorphic site was identified in 20 *domesticus* and 22 *musculus* samples. Since *M. m. domesticus* and *M. m. musculus* are relatively closely related (estimated divergence time of one million years (Guénet and Bonhomme 2003)), the overall Ka/Ks of 0 may not be meaningful. The Ka/Ks test was also performed by contrasting both subspecies to *M. caroli*, which has diverged from the two subspecies three million years ago (Guénet and Bonhomme 2003). The Ka/Ks ratios from this comparison further confirm the strong selective constraint on this gene. This is in accordance to the assumption that amino-acid changes are mostly deleterious to the function of the protein and that purifying selection acts on them (Hurst 2002). Therefore, it can be excluded that protein evolution is coupled with the expression difference between the two subspecies. If genes acted as integrated units in which protein sequence and expression patterns are coupled, one would expect that strong stabilizing selection, as for Lsm10, would lead to highly conserved protein sequence and expression levels (Castillo-Davis et al. 2004). That regulatory evolution may be of more importance than protein evolution has a long standing history (King and Wilson 1975), since many analyses have shown the conservation of proteins among distantly related taxa despite high phenotypic dissimilarities. In contrast, less is known about the contribution of positive selection to these changes. For *Drosophila*, Andolfatto (2005) recently has estimated that the contribution of non-coding DNA to adaptive evolution is one order of magnitude larger than that of amino acid changes. Also other studies have illustrated the importance of adaptive evolution acting on regulatory sequences (Kohn et al. 2004, Begun et al. 2007).

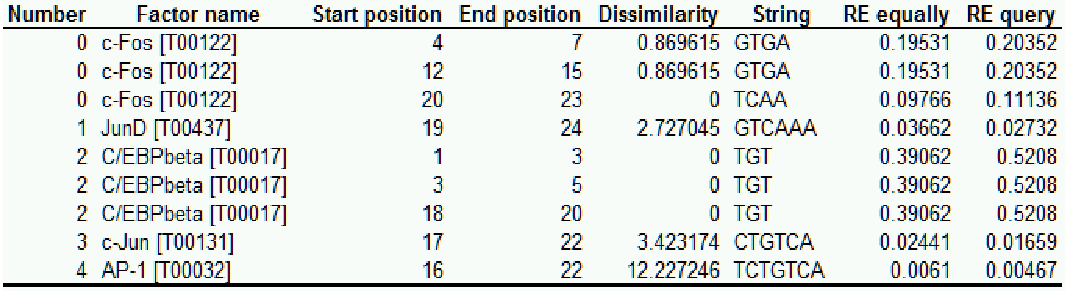
Analysis of the 5'upstream region was performed to uncover potential signatures of recent selection. Since positive selection of the protein coding sequence is highly unlikely because not a single non-synonymous substitution occurs, signatures of selection might be indicative for an adaptive change in regulatory elements. Nucleotide diversity for Lsm10 is in general low for both subspecies, with π overall being slightly higher for *musculus* than *domesticus* and θ_W being about two times higher in *musculus*. Ihle et al. (2006) found an average π in a 20kb region of 0.0025 for a potential sweep region in *M. m. domesticus*, which is three respectively four times higher. Also Baines and Harr (2007) reported overall higher π for different populations each of *domesticus* and *musculus* (π = 0.00155 for average over the population from Kazakhstan and Czech Republic and 0.002255 for average over Iranian and German population) for seven autosomal loci. The low levels of variability in both subspecies may indicate that the upstream sequence exhibits general high selective constraint. The sequenced fragments are located within 5 kb upstream of the coding region. Gaffney and Keightley (2006) showed that there are three times as many selectively constrained sites within non-coding DNA than in coding DNA, and that the constraint of non-coding DNA does not decrease before a distance of 5 kb from the known genic region. Tajima's D analysis revealed overall negative D values for *musculus* and positive D values for *domesticus*. Nevertheless, both results do not deviate from the neutral expectation. This result is independent of whether the concatenated fragment or the single fragments are considered. Furthermore there is no significant difference in nucleotide polymorphism between the populations of the two subspecies, thus any observed pattern seems to be a subspecies effect rather than a population effect. As both D values do not reveal significant results, the observed pattern may indicate neutral evolution of the region, which would also imply that the expression difference between the subspecies has a neutral basis. In principle it could also be that the expression difference is caused through a *trans*-effect, which would not be detected by sequencing the upstream non-coding region but this can be ruled out by the the results of the *cis*-*trans*-analysis discussed in paragraph 4.4.2. However, it is very difficult to achieve significance for Tajima's D. In general it is assumed that Tajima's D has too little power without sample sizes upwards 50 alleles (Simonsen et al. 1995), furthermore the overall reduced polymorphism in this region may hinder the achievement of significance. Thus, despite the non-significance, the pattern of nucleotide polymorphism in *M. m. musculus* may be indicative of positive selection acting on Lsm10. The negative D value hints to the occurrence of rare alleles, which should be in access in populations which recently exhibited a selective sweep (Tajima 1989, Braverman et al. 1995). Interestingly, the variability in direct vicinity to the coding sequence is zero and increases with increasing distance to the gene, a pattern which is typical for a selective sweep, since it is assumed that with increasing distance, more variability can be observed due to recombination events during the fixation process (Maynard Smith and Haigh 1974). For *domesticus* no such difference exists. Another indicator for adaptive evolution in this region is the occurrence of seven fixed differences between the two subspecies although overall only 13 segregating sites are identified between them. Thus, there is a high divergence in the potential regulatory region which is consistent with the difference in expression. Occurrence of high frequency derived alleles in the flanking region, as tested by Fay and Wu's H (Fay and Wu 2000), would give more support to the selective sweep hypothesis. However, amplification of the upstream targets was not possible for *M. caroli*, which is supportive of high divergence and therefore directional evolution in the regulatory region. High expression of Lsm10 represents the derived state, as *M. caroli* also shows low expression of the transcript. With regard to the observation that increases in gene expression are more often associated with adaptive evolution than decreases (Holloway et al. 2007), this may be an additional hint to a positive selection event in *M. m. musculus*.

In summary, a coupling of protein coding sequence differences and expression divergence between *M. m. domesticus* and *M. m. musculus* can be excluded and it is most likely, that the expression difference has been mediated via a change in a *cis*-regulatory region. Whether this change is associated with an adaptive event or attributable to a neutral process remains unclear, but some results hint towards a selective sweep event in *M. m. musculus*.

Interestingly, a 25 nucleotide repeat (gtgtgtgagtcggtgatctgtcaaa) in the upstream region of *M. m. musculus* was identified that coincides with the higher expression of *musculus* individuals in comparison to *M. m. domesticus*. The insertion lies about 4.6 kb upstream and is repeated between 2-4 times in *musculus* individuals while it occurs only one time in *domesticus* and also *M. caroli*, the latter showing low expression for Lsm10. The number of repeats differs between the populations; the Austrian *musculus* samples used for the expression analysis showing three repeats. A search for putative transcription factor binding sites (TFBS) in *M. musculus* for this region was conducted with PROMO (http://alggen.lsi.upc.es/cgi-bin/promo_v3/promo /promoinit.cgi?dirDB=TF_8.3), which uses TBS defined in the TRANSFAC database to construct specific binding site weight matrices for TFBS prediction (Messeguer et al. 2002). Several motifs that lie in this region have been identified (Table 4.5). PROMO is different from other TFBS-prediction programs in that it provides expectation values for the binding sites and hence a way to control significance. The RE value gives the expectation to find the same hit in a random sequence of 1kb length. Furthermore, the similarity of the hit sequence to the matrix is taken into account. Three binding sites are identified below a threshold of 0.05: JunD, c-Jun and AP-1 (Table 4.5). The inducible transcriptional complex AP-1 is composed of c-Fos and c-Jun proteins (Jochum et al. 2001). Three c-Fos binding sites are also identified in the sequence stretch, although with RE values > 0.05. Taken together, the insert contains binding sites for c-Fos and c-Jun that constitute the AP-1 complex and have been shown to act together in situ (Jochum et al. 2001). The software does not take cooperation between binding sites into account for calculating expectation values. If two events A and B are independent, then the probability of both events is the product of the probabilities for each event. The probability of 3 x c-Fos and 1 x c-Jun TFBSs to appear in this sequence is < 10^{-4} (0.20352 * 0.20352 * 0.11136 * 0.01659 = 0.00008). Hence, finding the association of these binding sites by chance is even less likely and it is tempting to speculate on functional significance, e.g. it would be interesting to see if the number of repeats correlates with expression height. In this scenario, insertion of this sequence would act as a transcriptional enhancer. As Lsm10 is important for the formation of the small ribonucleoprotein particle, a part of the splicesosomal machinery, this could have a significant downstream effect.

Nevertheless, the specific connection between activation of the splicesosomal machinery (Lsm10) by AP-1 (as the data indicate) remains to be elucidated.

Table 4.5: Potential TFBS in the 5' upstream region of Lsm10.



Number	Factor name	Start position	End position	Dissimilarity	String	RE equally	RE query
0	c-Fos [T00122]	4	7	0.869615	GTGA	0.19531	0.20352
0	c-Fos [T00122]	12	15	0.869615	GTGA	0.19531	0.20352
0	c-Fos [T00122]	20	23	0	TCAA	0.09766	0.11136
1	JunD [T00437]	19	24	2.727045	GTCAAA	0.03662	0.02732
2	C/EBPbeta [T00017]	1	3	0	TGT	0.39062	0.5208
2	C/EBPbeta [T00017]	3	5	0	TGT	0.39062	0.5208
2	C/EBPbeta [T00017]	18	20	0	TGT	0.39062	0.5208
3	c-Jun [T00131]	17	22	3.423174	CTGTCA	0.02441	0.01659
4	AP-1 [T00032]	16	22	12.227246	TCTGTCA	0.0061	0.00467

Neil3

Neil3 (Nei like3 (E. coli)), located on chromosome 8 (54,672 kb - 54,724 kb), has ten exons and codes for a protein with 606 amino acids. It is a putative DNA glycosylase. DNA glycosylases are involved in DNA replication, recombination and DNA repair.

Two tests were applied to evaluate if protein-coding evolution is coupled with the expression difference between *M. m. domesticus* and *M. m. musculus*. The McDonald-Kreitman test revealed no significant result, suggesting no deviation from the neutral expectation. The overall Ka/Ks test on the other hand yields a ratio of 0.933. It is generally assumed that a cutoff of one as indicator of positive selection is very conservative, as in most cases not the whole sequence might be under positive selection. Most probably, only a segment of a sequence that codes for a defined protein domain is adapting. Thus, Ka/Ks ratios higher than one are quite rare, in general only about 1% of genes have a Ka/Ks ratio greater one (Fay and Wu 2003). For mouse-rat comparison of 363 genes the average Ka/Ks ratio was estimated to be 0.14 (with a range of 0.05-0.2) (Wolfe and Sharp 1993). Hence, the Ka/Ks ratio between *domesticus* and *musculus* could hint to positive selection in some regions of the gene. A sliding window analysis did not support this idea; no region with significantly elevated Ka/Ks was identified. The combination of both tests suggests neutral or constraint evolution of the coding region between *domesticus* and *musculus* rather than positive evolution. The Ka/Ks analysis between *M. caroli* and the two *M. musculus* subspecies on the other hand, yields values that are about half the magnitude of the *domesticus-musculus* comparison. For both comparisons, the sliding window analysis revealed one region with a strongly elevated Ka/Ks ratio (and a second region with Ka/Ks slightly higher than one), which may reflect rapid change for this region of the gene in the *musculus*-subspecies. Specific regions of elevated Ka/Ks have been often shown to be functional important (e.g. drug resistance mutations in HIV (Chen et al. 2004)). Interestingly this region maps to exon eight, for which some evidence for an alternative spliced transcript in at least *domesticus* has been revealed from the analysis of different cDNAs for this region (data not shown). Alternative splicing is thought to contribute significantly to phenotypic complexity by allowing a single locus to produce multiple and possible functionally distinct proteins. Xing and Lee (2005) showed that alternative splicing can relax amino acid pressure without affecting neighboring exons in the same gene, thus creating evolutionary hotspots in which one part of a protein sequence is allowed to accumulate amino acid mutations at a higher rate than the rest of the protein. Because new functions may arise from the insertion or deletion of an exon, it has been suggested that alternatively spliced exons can accelerate gene evolution (Chen et al. 2006). A Northern Blot analysis failed to decipher the exact transcript structure, and if it is a unique feature of *domesticus* only, but has shown that at least two transcripts (one of ca. 2 kb and the other of ca. 5kb, data not shown) exist in both subspecies and *M. caroli*, underlining that different transcripts could indeed originate from this locus. It is also possible, that the expression difference between the subspecies from the GeneChip® is influenced by the detection of differently spliced transcripts. If this is the case, it is obvious that the pattern is complex and several different transcripts have to be involved. The two which have been identified with Northern Blot analysis showed no expression difference, thus minimum a third one, (which is not detected by the Northern Blot probe) has to be involved.

The overall nucleotide diversity of the non-coding region for Neil3 is six times higher regarding π and five times higher regarding θ_W than in Lsm10. The π values (0.00206 for *domesticus* and 0.0057 for *musculus*) are also higher than those reported by others (Ihle et al. 2006, Baines and Harr 2007). These higher polymorphism levels may point towards an overall less constraint on the non-coding region for this gene, suggesting neutral evolution of the *cis*-regulatory region. In general, the nucleotide polymorphism is three times lower in *domesticus* than in *musculus*, also an overall negative Tajima's D was found in *domesticus*. This could be a signature of a recent selective sweep in *domesticus* and would suggest - since Neil3 is low expressed in *M. caroli* - an adaptive advantage of high expression, most probably due to *cis*-

regulatory differences. However, Tajima's D test did not reveal results significant different from zero for both subspecies and the separate populations, respectively, which is indicative of neutral evolution for both coding and non-coding sequence rather than selective forces acting on Neil3. Nevertheless, the regulatory difference may also be mediated through *trans*-acting factors, which would not be detected in the flanking region of the gene (Holloway et al. 2007).

Recapitulatory, the mode of evolution for the expression differences between *domesticus* and *musculus* for Neil3 is not clear. The expression difference seems to be decoupled from protein evolution, since recurrent selection of the protein sequence can be ruled out; but modifications of the protein (including alternatively spliced transcripts) may have contributed to the evolution between the *musculus* subspecies and *M. caroli*. It cannot be excluded that alternatively spliced transcripts have caused the differences in signal intensity of the microarray. Complete decipherment of the transcript structures would be a first step to uncover the involved factors. Assuming a *cis*-regulatory change, high expression of the gene as in *domesticus* may be adaptive, since some evidence for a selective sweep event exists. More support for this scenario comes from the hybrid zone analysis and will be discussed in paragraph 4.4.3.

4.4.2 *Cis*- versus *trans*-regulatory changes

Lsm10

The *cis*-*trans*-analysis suggests that the expression difference for Lsm10 is mediated by a change in *cis*. For all hybrids the expression follows the locus-specific genotype (as determined from a diagnostic SNP analysis between the parental subspecies).Thus, animals with a *domesticus*-like genotype show low and animals with a *musculus*-like genotype show high expression. Furthermore individuals with a *hybrid*-like genotype show clearly intermediate expression, with very low variance in expression. This is only expected if the change is attributed to differences in *cis* (i.e. changes in promoter, respectively TFBS). Even stronger support of the idea comes from the analysis of samples, for which the overall genotype deviates from the locus-specific genotype/expression. For those samples, a *trans*-acting factor acting on expression can be definitely ruled out. Since all samples show congruence between their locus-specific genotype and expression phenotype, the expression difference between the two subspecies is most likely due to a change in *cis*. This result fits to the observation of general high selective constraint on the 5'upstream region (see 4.4.1). TFBSs that can potentially be involved were identified, which further underlines that the expression differences is caused by a *cis*-regulatory change. For *Drosophila* and mice, *cis*-effects have been shown as dominant contributor to interspecies expression differences (Wittkopp et al. 2004, Doss et al. 2005), although other studies come to different results (e.g. Yvert et al. 2003, Morley et al. 2004).

However, what is missing to date is the evidence of whether *cis*-regulatory differences are also more likely to be adaptive than *trans*-differences (Hoekstra and Coyne 2007), although several authors (e.g. Caroll et al. 2001) are convinced that *cis*-regulatory regions have a higher probability of being adaptive than a change in a protein-coding region (and thus a putatively *trans*-acting factor like transcription factors). *Cis*-regulatory changes are believed to be freer from negative pleiotropic effects on fitness. To date, only few examples exist, in which an adaptive change is connected to a *cis*-regulatory change. The three most cited (Sucena and Stern 2000, Shapiro et al. 2004, Gompel et al. 2005) lack the final identification of the involved mutation (and therefore the final proof for *cis*-regulation). Harr et al. (2006) find convincing evidence for the fact that an expression difference in the MKK7 gene between *M. m. domesticus* and *M. m. musculus* is adaptive, most likely due to a *cis*-effect. Nevertheless, the contribution of *trans*-acting factors cannot be completely ruled out. In contrast, for Lsm10 the contribution of a *cis*-acting factor to the difference in expression is obvious from the hybrid zone analysis. What is unclear so far is, if the higher expression in *M. m. domesticus* is caused by an adaptive event, though some evidence for positive selection exists. It remains open, what functional consequences the up-regulation of this gene, which in itself has a strong potential to affect many downstream genes, has for the individual.

Neil3

For Neil3 the underlying genetic basis of the expression difference between the two subspecies and their hybrids is unclear. The expression difference is predominantly due to a change in *cis* (accompanied by a change in *trans* with a lesser contribution); a pure *trans*-acting change can be ruled out. This is deduced from the analysis of animals for which the overall genotype is not identical with the locus-specific genotype. If the expression change was due to a pure *trans*-effect, all of the 15 animals most likely would exhibit expression according to their overall genotype; a pure *cis*-effect would lead to perfect correlation between locus-specific genotype and expression levels for all of them. Since five samples show no correlation between the locus-specific genotype and the expression, a pure *cis*-effect is just as unlikely as a pure *trans*-effect. Thus, the observed pattern may indicate that the expression difference between the two subspecies is predominately due to a difference in *cis,* with a minor contribution of a change in *trans*. A contribution of a *trans* effect would also explain the higher variance in expression (in comparison to Lsm10), if one considers that different combinations of genotypes for the involved loci are present in the hybrids. That changes in *cis* are often accompanied by change is *trans* has also been ascertained by Wittkopp et al. (2004). The authors conducted an expression study with interspecific hybrids of two *Drosophila* species Almost all genes surveyed had differences in *cis*-regulation, but for 55% of the genes, this change was accompanied by a change in *trans* with varying contribution to the expression variation of the latter. Several other studies found gene regulation to be complex and to have many loci involved in expression variation (see Whitehead and Crawford 2006 for a review of studies). However, further analysis is needed to clearly decipher the genetic basis of the expression difference of Neil3. An analysis considering allele-specific expression (like in Wittkopp et al. 2004) of the parental subspecies and F1 hybrids, where the divergence in gene expression between the parents is compared to the difference in alleles in F1 hybrids, can be helpful to infer the relative contribution of both factors. Since the two alleles and their *cis*-regulatory elements derived from the parents share the same pool of *trans*-acting factors, unequal abundance of transcripts of the two alleles in the F1 hybrid suggests the presence of the genetic variation acting in *cis*. If *trans*-regulation diverges between species, the relative allelic expression in hybrids will differ from the relative gene expression between the species.

A second explanation for the observed pattern exists, which also may have a synergistic effect. It may be that the used diagnostic SNP (identified in the 3' region) is not completely fixed. In over two kb upstream region no fixed difference has been identified in 14 *domesticus* and 18 *musculus* individuals. It is possible, that the putatively diagnostic SNP is present in homozygous form at high frequencies in the respective populations but still not fixed. For the correlation analysis, this could lead to assigning animals carrying chromosomes from both species as homozygous and vice versa. This could create the exact pattern that is found, e.g. apparent *domesticus* genotype individuals with "*hybrid*"-like expression. Under this scenario, the expression difference between the two subspecies would be most likely caused by a change of a *cis*-acting factor.

In summary, a change solely in *trans* can be excluded and a *cis*-effect seems to be the major (if not the only, assuming non-fixation of the diagnostic SNP) contributor to the expression difference. Further analysis is needed to completely uncover the contribution of *cis*- and *trans*-effects to the expression difference between the two subspecies for Neil3.

4.4.3 Functional consequences of candidate genes in a hybrid zone

The analysis of the genotypes over 14 loci for the Bavarian samples, suggests that the hybrid zone seems to have remained stable since the calculation from Sage et al. (1986). They estimated the width of the central cline of the hybrid zone to be 20 kilometers; over this distance the average hybrid indices changed from predominately *domesticus* to *musculus* genotypes. The analyzed samples exhibit average genotypes which fit to the localities they were derived from, related to the supposed center of the hybrid zone, showing *domesticus*-like genotypes one the western part of the center of the zone and *musculus*-like genotypes on the eastern part of the zone. This result is in accordance with what Barton and Hewitt (1985, 1989) stated for most of the hybrid zones studied. They argued that hybrid zones are barriers to gene flow and characterized by sharp changes in allele frequencies at the center of the zone so that each allele tends to be associated with other alleles from the parental population. In contrast to other surveys of hybrid zones, this analysis accounts for genotypes averaged over several loci of single animals, while usually population samples from different localities and allele frequencies of single loci are considered (Sage et al. 1986, Payseur et al. 2004, Macholán et al. 2007,). Thus, the analysis holds the potential, that single outliers confound the results. Nevertheless, the observed pattern shows an abrupt change from more *domesticus* to more *musculus* genotype at the center of the zone and pure subspecies at the tails of the zone, which leads to the conclusion that the consideration of single animals as representatives for population-allele frequencies is feasible for the applied approach and that the hybrid zone has remained considerably stable. For the Danish hybrid zone it is known that it has not moved since the 1960s (Raufaste et al. 2005) and also Macholán et al. (2007) found strong evidence for the stability of the hybrid zone.

Patterns of expression phenotypes across the hybrid zone

It is thought that differential movement of alleles across the hybrid zone would reflect differences in fitness of particular heterospecific gene combinations (Dod et al. 1993). The hybrid zone analysis of the two candidate genes suggests that the expression difference between the parental subspecies seems not to be of direct significance for reproductive isolation. If this were the case, one would have expected that alleles respectively expression phenotypes indicative for one subspecies do not move into the overall genomic background of the other subspecies because in hybrid populations natural selection is thought to remove deleterious heterospecific combination of genes that cause functional disruptions (Payseur and Place 2007). Since for both genes high or intermediate expression is frequently found in the opposite average genotype (in which they are supposed to be low expressed), selection seems not to act against these allele combinations. However, this result should be treated carefully as the analysis poses some problems. "Classical" hybrid zone approaches, trying to identify regions of the genome involved in reproductive isolation in mice, evaluate patterns of differential introgression by statistically comparing changes of allele frequencies along a geographical cline between several markers (e.g. Payseur et al. 2004, Dod et al. 2005, Macholán et al. 2007, Teeter et al. 2007). Such studies consider extensive population samples from many localities. The shape of the different clines (cline width and center of the cline) is used to estimate selection parameters of loci. Loci which deviate from an average (neutral) pattern (e.g. very narrow cline width) may have been exposed to selection. Payseur et al. (2004) identified one locus on the X-chromosome in their study for which they found a very low cline width, which they attributed to this locus being a candidate region for reproductive isolation; also others identified patterns of limited introgression for X-chromosomal loci (Tucker et al. 1992, Dod et al. 1993) or very narrow clines for autosomal loci (Teeter et al. 2007). The present analysis does not consider cline width nor does it compare different loci. Inferences about differences in introgression are made from the observation of single or few individuals per sample locality. Only few localities are considered (four on each side of the presumed center of the hybrid zone). Furthermore, the sampled transect was quite narrow with a width of 20 km and with a bias towards eastern samples. Inferences about introgression of alleles are made by considering the relation of the locus-specific genotype/expression phenotype of a candidate gene to an individual's overall genotype. If a locus-specific genotype/expression phenotype is frequently found in the overall genomic background of the respective other subspecies, this is seen as "introgression", respectively, if the locus-specific genotype/expression phenotype coincidences with the overall genotype, selection acts against these alleles in a foreign genetic background. This analysis is purely descriptive, no comparison with neutral loci was done and the sample set is small and biased. Therefore, one has to be careful if referring to "introgression" since the analysis holds the potential to be strongly influenced by effects of single samples. Recently, Macholán et al. (2007) postulated that analysis in a two-dimensional space, dense sampling and rigorous statistical treatment of data is essential for hybrid zone analyses to make inferences about loci involved in reproductive isolation. Thus, the present analysis might be an oversimplification of the complex basis of factors which have to be taken into consideration. Nevertheless, the analysis should give a rough idea about fitness effects of the specific alleles in a foreign genetic background. No abrupt change, as in the overall genotype, is observed for the locus specific genotypes/expression phenotypes, even samples with almost pure overall genotype (at the very western and eastern tail of the hybrid zone, respectively) show expression levels of the respective other subspecies. From these observations, it seems at least not likely that both genes act as isolating factors and contribute to reproductive isolation between the two subspecies.

Studies on plants indicate that hybridization can provide a source of genetic variation for adaptation (Arnold et al. 1991, Rieseberg 1991, 1997, Rieseberg et al. 2003) and patterns of different introgression of alleles can also be used to identify adaptive gene flow by uncovering genomic regions that mix between species at unusually high rates (Payseur et al. 2004). Comparison of many loci is necessary to differentiate between a neutral and an adaptive locus. In contrast, the present analysis combines patterns of genotype/expression-phenotype-deviation with population genetic data to gain information about the fitness effects of the expression differences between the two subspecies. An advantage of surveying expression phenotypes over the hybrid zone is that one is independent of diagnostic marker, which, if they are not 100% diagnostic, may confound cline estimates (Payseur et al. 2004). Furthermore, expression is an intermediate state between DNA polymorphism and an organism's phenotype and may be more directly linked to consequences for individuals. For both genes the up-regulation of the transcript represents the derived state. Interestingly, for both the high or at least intermediate expression is frequently found in the genomic background where it would not be expected to occur, suggesting adaptive introgression of *musculus* into *domesticus* for Lsm10 and *domesticus* into *musculus* for Neil3. The pattern is particular striking for Neil3, for which almost all overall more *musculus*-like samples exhibit at least intermediate expression although those samples should more likely be low expressed, assuming detrimental effects of the heterospecific gene combinations. Since more samples from the eastern side of the zone have been collected, the evidence for adaptive introgression of *domesticus* into *musculus* in Neil3 is more conclusive than the reverse case for Lsm10. Evidence for adaptive introgression into a foreign genomic background has been revealed from the Bavarian hybrid zone. By comparison of allele frequency clines, Payseur et al. (2004) identified one locus, *Xist*, which is unusually polymorphic on the *musculus* side on the hybrid zone. They suggest that *domesticus* alleles of a gene mapping to the *Xist* region may outcompete the *musculus* allele in an overall *musculus* genomic background and are experiencing positive selection in a heterospecific genomic background. Teeter et al. (2007) identified several autosomal markers with asymmetrical broad clines, with high frequencies of *domesticus* alleles on the *musculus* side of the hybrid zone. The associated genes are related to cell signaling, olfaction and pheromone response which play important roles in survival and reproduction and are likely targets of positive selection. For both genes the pattern of potential adaptive introgression fits the population genetic data. Reduced nucleotide diversity, respectively negative Tajimas's D values are obvious in the subspecies with the high expression, also suggesting that this expression status might be an advantage. Of course this analysis has some caveats: since the sampling is biased towards more *musculus* samples and no comparison with other loci was done, all observed patterns might also show a neutral scenario. However, if the effect was neutral rather than adaptive, one would expect variation in gene expression on the respective other side of the hybrid zone as well, even though for Neil3 such a statement is difficult because of the small number of samples western of the center of the zone. Asymmetric transition of diagnostic alleles has been reported from all parts of the hybrid zone so far (Tucker et al. 1992, Dod et al. 1993, Payseur et al. 2004, Raufaste et al. 2005, Macholán et al. 2007, Teeter et al. 2007). Despite the Czech transect of the zone (Macholán et al. 2007), this asymmetric introgression most often occurs from *domesticus* into *musculus*. One explanation for this asymmetry is that if intrinsic genomic incompatibilities between *M. m. domestic*us alleles and *M. m. musculus* alleles occur, more *M. m. musculus* alleles are incompatible with a *M. m. domesticus* genetic background than vice versa. Furthermore, behavioral differences between the subspecies might be responsible. It is known that wild *M. m. musculus* prefer homspecific urine signals (Smadja and Ganem 2002, 2005) but no such preference was found in *M. m. domesticus*. A lack of preference for conspecific mates in *M. m. domesticus* would indicate that they mate more likely with hybrids or *M. m. musculus* than vice versa. Furthermore *M. m. domesticus* males usually dominate less aggressive *M. m. musculus* males (Van Zegeren and Oortmerssen 1981). However, these differences would skew clines to the *musculus* side, which would be no explanation for the pattern observed in Lsm10.

One could argue that the pre-selection of candidate genes was not appropriate to identify genes that are contributing to reproductive isolation. The analyzed genes are intermediate in expression in F1 hybrids and therefore not strongly misregulated (i.e. higher or lower in expression in comparison to both parents), according to “classical” regulatory hybrid incompatibilities (Landry et al. 2007), and thus eventually less likely to have a negative fitness effect in hybrids. Nevertheless, while intermediate expressed genes might cause reduced fitness already in F1 hybrids, they might furthermore be important isolating factors in F2 or backcrosses (Oka et al. 2007). The heterozygous state may not be deleterious enough to cut off gene flow and effects may only occur, when alleles from single loci or combinations of loci become homozygous in F2 or later generations (Orr and Presgraves 2000). However, a contribution to reproductive isolation of the two analyzed candidate genes seems unlikely, although the observed pattern of expression across the hybrid zone and the deduced conclusions should be treated with care. The analysis (which considers hybrids apart from F1) in combination with population genetic data more likely suggests adaptive introgression of the phenotype with the high expression level into the genomic background which is associated with low transcript expression.

4.4.4 Conclusion

Although the approach applied here did not succeed to completely uncover the evolutionary mechanisms acting on and contributing to expression differences between the two subspecies *M. m. domesticus* and *M. m. musculus*, some insights have been gained. A coupling of protein and regulatory evolution can be excluded for both genes and *cis*-regulatory rather than *trans*-acting changes appear to contribute to the expression variation. Furthermore, evidence exists that the high expression phenotype represents an adaptive advantage, which seems to introgress into the respective other subspecies, rather than being involved in the process of reproductive isolation between the two house mouse subspecies. Further analysis (e.g. sequence analyses including outgroups) is needed to substantiate these conclusions and to gain more support for the selective sweep hypothesis as well as hybrid zone analyses with larger population sample sizes. Considering allele frequencies of populations rather than single animals lowers the chance that patterns, caused by random effects, are deemed real.

5 Literature

Affymetrix statistical algorithms description document. http://www.affymetrix.com/support/technical/whitepapers/sadd_whitepaper.pdf.

Albert, A. J. and D. Schluter (2005): Selection and the origin of species. Current Biology 15 (8): 283-288.

Alibert, P., Fel-Clair, F., Manolakou, K., Britton-Davidian, J. and J.-C. Auffray (1997): Developmental stability, fitness, and trait size in laboratory hybrids between European subspecies of the house mouse. Evolution 51:1284-1295.

Allison, D. B., Cui, X., Page, G. P. and M. Sabripour (2006): Microarray data analysis: from disarray to consolidation and consensus. Nature Reviews Genetics 7 (1): 55-65.

Andersen, C. B., Holst-Jensen, A., Berdal, K. G., Thorstensen, T. and T. Tengs (2006): Equal performance of TaqMan, MGB, Molecular Beacon, and SYBR Green-based detection assays in detection and quantification of roundup ready soybean. Journal of Agriculture and Food Chemistry 54: 9658-9663.

Andolfatto, P. (2005): Adaptive evolution of non-coding DNA in *Drosophila*. Nature 437: 1149-1152.

Arnold, S. J., Buckner, C. M. and J. J. Robinson (1991): Pollen-mediated introgression and hybrid speciation in Louisiana irises. Proceedings of the National Academy of Sciences of the United States of America 88: 1398-1402.

Auffray, J.-C., Vanlerberghe, F. and J. Britton-Davidian (1990): The house mouse progression in Eurasia: a palaeontological and archaeozoological approach. Biological Journal of the Linnean Society 41: 13-25.

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

Barbash, D. A., Siino, D. F., Tarone, A. M. and J. A. Roote (2003): A rapidly evolving MYB-related protein causes species isolation in *Drosophila*. Proceedings of the National Academy of Sciences of the United States of America 100: 5302-5307.

Barbash, D. A., Awadalla, P. and A. M. Tarone (2004): Functional divergence caused by ancient positive selection of a *Drosophila* hybrid incompatibility locus. PLoS Biology 2: 839-848.

Barton, N. H. and G. M. Hewitt (1985): Analysis of hybrid zones. Annual Review of Ecology and Systematics 16: 113-148.

Barton, N. H. and G. M. Hewitt (1989): Adaptation, speciation and hybrid zones. Nature 341: 497-503.

Barton, N. H. and K. S. Gale (1993): Genetic analysis of hybrid zones. Pp. 13-45 in Harrison, R. G. (ed.): Hybrid zones and the evolutionary process. Oxford University Press, Oxford, U.K.

Begun, D. J., Holloway, A. K., Stevens, K., Hillier, L. W., Poh, Y., Hahn, M. W., Nista, P. M., Jones, C. D., Kern, A. D., Dewey, C. N., Pachter, L., Myers, E. and C. H. Langley (2007): Population genomics: whole-genome analysis of polymorphism and divergence in *Drosophila simulans*. PLoS Biology 5 (11): 1-26.

Berry, R. J. and F. H. Bronson (1992): Life-history and Bioeconomy of the House Mouse. Biological Reviews of the Cambridge Philosophical Society 67 (4): 519-550.

Bosotti, R., Locatelli, G., Healy, S., Scacheri, E., Sartori, L., Mercurio, C., Calogero, R. and A. Isacchi (2007): Cross platform microarray analysis for robust identification of differentially expressed genes. BMC Bioinformatics 8: S5.

Boursot, P., Auffray, J. C., Britton-Davidian, J. and F. Bonhomme (1993): The evolution of house mice. Trends in Ecology and Evolution 24: 119-152.

Braverman, J. M., Hudson, R. R., Kaplan, N. L., Langley, C. H. and W. Stephan (1995): The hitchhiking effect on the site frequency spectrum of DNA polymorphisms. Genetics 140: 783-96.

Brem, R. B., Yvert, G., Clinton, R. and L. Kruglyak (2002): Genetic dissection of transcriptional regulation in budding yeast. Science 296: 752-755.

Brem, R. B. and L. Kruglyak (2005): The landscape of genetic complexity across 5,700 gene expression traits in yeast. Proceedings of the National Academy of Sciences of the United States of America 102: 1572-1577.

Brideau, N. J., Flores, H. A., Wang, J., Maheshwari, S., Wang, X. and D. A. Barbash (2006): Two Dobzhansky-Muller genes interact to cause hybrid lethality in *Drosophila*. Science 314: 1292-1295.

Britton-Davidian, J., Fel-Clair, F., Lopez, J., Alibert, P. and P. Boursot (2005): Postzygotic isolation between the two European subspecies of the house mouse: estimates from fertility patterns in wild and laboratory-bred hybrids. Biological Journal of the Linnean Society 84: 379-393.

Bustin, S. (2002): Quantification of mRNA using real-time reverse transcription PCR (RT-PCR): trends and problems. Journal of Molecular Endocrinology 29: 23-39.

Carroll, S. B., Grenier, J. K. and S. D. Weatherbee (2001): From DNA to diversity: molecular genetics and the evolution of animal design. Blackwell Publishing, Malden, MA.

Castillo-Davis, C. I., Hartl, D. L. and G. Achaz (2004): *Cis*-regulatory and protein evolution in orthologous and duplicate genes. Genome Research 14: 1530-1536.

Charlesworth, B., Morgan, M. T. and D. Charlesworth (1993): The effect of deleterious mutations on neutral molecular variation. Genetics 134: 1289-1303.

Chen, F.-C., Wang, S.-S., Chen, C.-J., Li, W.-H. and T.-J. Chuang (2006): Alternatively and constitutively spliced exons are subject to different evolutionary forces. Molecular Biology and Evolution 23 (3): 675-682.

Chen, L., Perlina, A. and C. J. Lee (2004): Positive selection detection in 40,000 human immunodeficiency virus (HIV) type 1 sequences automatically identifies drug resistance and positive fitness mutations in HIV protease and reverse transcriptase. Journal of Virology 78 (7): 3722-3732.

Chismar, J. D., Mondala, T., Fox, H. S., Roberts, E., Langford, D., Masliah, E., Salomon, D. R. and S. R. Head (2002): Analysis of result variability from highdensity oligonucleotide arrays comparing same-species and cross-species hybridizations. BioTechniques 33: 516-524.

Chuaqui, R. F., Bonner, R. F., Best, C. J. M., Gillespie, J. W., Flaig, M. J., Hewitt, S. M., Phillips, J. L., Krizman, D. B., Tangrea, M. A., Ahram, M., Linehan, W. M., Knezevic, V. and M. R. Emmert-Buck (2002): Post-analysis follow-up and validation of microarray experiments. Nature Genetics Supplement 32: 509-514.

Comings, D. E. and J. P. MacMurray (2000): Molecular heterosis: a review. Molecular Genetics Metabolism 71: 19-31.

Cowles, C. R., Hirschhorn, J. N., Altshuler, D. and E. S. Lander (2002): Detection of regulatory variation in mouse genes. Nature Genetics 32: 432-437.

Coyne, J. A. and H. A. Orr (1997): "Patterns of speciation in *Drosophila*" revisited. Evolution 51: 295–303.

Coyne, J. A. and H. A. Orr (2004): Speciation. Sinauer Associates, Massachusetts, USA.

Cucchi, T., Vigne, J.-D. and J.-C. Auffray (2005): First occurrence of the house mouse (*Mus musculus domesticus* Schwarz & Schwarz, 1943) in the Western Mediterranean: a zooarchaeological revision of subfossil occurrences. Biological Journal of the Linnean Society 84: 429-446.

Cui, X., Affourtit, J., Shockley, K. R., Woo, Y. and G. A. Churchill (2006): Inheritance patterns of transcript levels in F1 hybrid mice. Genetics 174: 627-637.

Darwin, C. (1859): On the origin of species by means of natural selection or the preservation of favored races in the struggle for life. J. Murray, London.

De Reynies, A., Geromin, D., Cayuela, J. M., Petel, F., Dessen, P., Sigaux, F. and D. S. Rickman (2006): Comparison of the latest commercial short and long oligonucleotide microarray technologies. BMC Genomics 7: 51.

de Vos, S., Hofmann, W.-K., Grogan, T. M., Krug, U., Schrage, M., Miller, T. P., Braun, J. G., Wachsman, W., Koeffler, H. P. and J. W. Said (2003): Gene expression profile of serial samples of transformed B-cell lymphomas. Laboratory Investigation 83 (2): 275-281.

Denver, D. R., Morris, K., Streelman, J. T., Kim, S. K., Lynch, M. and W. K. Thomas (2005): The transcriptional consequences of mutation and natural selection in *Caenorhabditis elegans*. Nature Genetics 37 (5): 544-548.

Din, W., Anand, R., Boursot, P., Darviche, D., Dad, B., Jouvin-Marche, E., Orth, A., Talwar, G. P., Cazenave, P.-A. and F. Bonhomme (1996): Origin and radiation of the house mouse: Clues from nuclear genes. Journal of Evolutionary Biology 9 (5): 519-539.

Dobzhansky, T. (1936A): Studies on hybrid sterility. II. Localization of sterility factors in *Drosophila pseudoobscura* hybrids. Genetics 21: 113-135.

Dobzhansky, T. (1936B): Studies on hybrid sterility: II. Localization of sterility factors in *Drosophila virilis* Sturt. X *lummei* Hackmann hybrids. Genetics 21: 112-135.

Dobzhansky, T. (1937): Genetics and the Origin of Species. New York: Columbia University Press.

Dod, B., Jermiin, L. S., Boursot, P., Chapman, V. H., Tonnes-Nielsen, J. and F. Bonhomme (1993): Counterselection on sex chromosomes in the *Mus musculus* European hybrid zone. Journal of Evolutionary Biology 6:529-546.

Dod, B., Smadja, C., Karn, R. C. and P. Boursot (2005): Testing for selection on the androgen-binding protein in the Danish mouse hybrid zone. Biological Journal of the Linnean Society 84: 447-459.

Doss, S., Schadt, E. E., Drake, T. A. and A. J. Lusis (2005): *Cis*-acting expression quantitative trait loci in mice. Genome Research 15: 681-691.

Draghici, S., Khatri, P., Eklund, A. C. and Z. Szallasi (2006): Reliability and reproducibility issues in DNA microarray measurements. Trends in Genetics 22 (2): 101-109.

Enard, W., Khaitovich, P., Klose, J., Zöllner, S., Heissig, F., Giavalisco, P., Nieselt-Struwe, K., Muchmore, E., Varki, A., Ravid, R., Doxiadis, G. M., Bontrop, R. E. and S. Pääbo (2002): Intra- and interspecifc variation in primate gene expression patterns. Science 296: 340-343.

Etienne, W., Meyer, M. H., Peppers, J. and R. A. Meyer (2004): Comparison of mRNA gene expression by RT-PCR and DNA microarray. BioTechniques 36: 618-621.

Falconer, D. S. and T. F. C. Mackay (1996): Introduction to Quantitative Genetics. Longman, Essex, UK.

Fay, J. C. and C. I. Wu (2000): Hitchhiking under positive Darwinian selection. Genetics 155: 1405-1413.

Fay, J. C. and C. I. Wu (2003): Sequence divergence, functional constraint, and selection in protein evolution. Annual Review of Genomics and Human Genetics 4: 213-35.

Forejt, J. and P. Ivanyi (1975): Genetic studies on male sterility of hybrids between laboratory and wild mice (*Mus musculus* L.). Genetical Research 24: 189-206.

Forejt, J. (1996): Hybrid sterility in the mouse. Trends in Genetics 12: 412-417.

Fossella, J. S., Samant, A., Silver, L. M., King, S. M., Vaughan, K. T., Olds-Clarke, P., Johnson, K. A., Mikami, A., Vallee, R. B. and S. H. Pilder (2000): An axonemal dynein at the hybrid sterility 6 locus: implications for t haplotype-specific male sterility and the evolution of species barriers. Mammalian Genome 11: 8-15.

Freeman, W. M., Walker, S. J. and K. E. Vrana (1999): Quantitative RT-PCR: pitfalls and potential. BioTechniques 26: 112-125.

Gadeau, J., Page, R. E. and J. H. Werren (1999): Mapping of hybrid incompatibility loci in *Nasonia*. Genetics 153: 1731-1741.

Gaffney, D. A. and P. D. Keightley (2006): Genomic selective constraints in murid noncoding DNA. PLoS Biology 2 (11): 1912-1923.

Galtier, N., Bonhomme, F., Moulia, C., Belkhir, K., Caminade, P., Desmarais, E., Duquesne, J. J., Orth, A., Dod, B. and P. Boursot (2004): Mouse biodiversity in the genomic era. Cytogenetic and Genome Research 105: 385-394.

Ganem, G., Ginane, C., Ostrowski, M.-F. and A. Orth (2005): Assessment of mate preference in the house mouse with reference to investigations on assortative mating. Biological Journal of the Linnean Society 84: 461-471.

Gershon, D. (2005): More than gene expression. Nature 437: 1195-1198.

Gibson, G., Riley-Berger, R., Harshman, L., Kopp, A., Vacha, S., Nuzhdin, S. and M. Wayne (2004): Extensive sex-specific nonadditivity of gene expression in *Drosophila melanogaster*. Genetics 167: 1791-1799.

Gibson, G. and B. Weir (2005): The quantitative genetics of transcription. Trends in Genetics 21: 616-623.

Gilad, Y., Rifkin, S. A., Bertone, P., Gerstein, M. and K. P. White (2005): Multi-species microarrays reveal the effect of sequence divergence on gene expression profiles. Genome Research 15: 674-680.

Gompel, N., Prud'homme, B., Wittkopp, P. J., Kassner, V. A. and S. B. Carroll (2005): Chance caught on the wing: *Cis*-regulatory evolution and the origin of pigment patterns in *Drosophila*. Nature 433: 481-487.

Good, J. M., Handel, M. A. and M. W. Nachman (2007): Asymmetry and polymorphism of hybrid male sterility during the early stages of speciation in house mice. Evolution: in press.

Grant, V. (1971): Plant speciation. Columbia University Press, New York.

Gu, Z., Nicolae, D., Lu, H. H. and W. H. Li (2002): Rapid divergence in expression between duplicate genes inferred from microarray data. Trends in Genetics 18: 609-613.

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

Haddrill, P. R., Thornton, K. R., Charlesworth, B. and P. Andolfatto (2005): Multilocus patterns of nucleotide variability and the demographic and selection history of *Drosophila melanogaster* populations. Genome Research 15: 790-799.

Haerty, W. and R. S. Singh (2006): Gene regulation divergence is a major contributor to the evolution of Dobzhansky-Muller incompatibilities between species of *Drosophila*. Molecular Biology and Evolution 23: 1707-1714.

Haldane, J. B. S. (1922): Sex-ratio and unidirectional sterility in hybrid animals. Journal of Genetics 12: 101-109.

Harr, B. (2006): Genomic islands of differentiation between house mouse subspecies. Genome Research 16: 730-737.

Harr, B. and C. Schlötterer (2006): Comparison of algorithms for the analysis of Affymetrix microarray data as evaluated by co-expression of genes in known operons. Nucleic Acids Research 34 (2): e8.

Harr, B., Voolstra, C., Heinen, T. J., Baines, J. F., Rottscheidt, R., Ihle, S., Muller, W., Bonhomme, F. and D. Tautz (2006): A change of expression in the conserved signaling gene MKK7 is associated with a selective sweep in the western house mouse *Mus musculus domesticus*. Journal of Evolutionary Biology 19: 1486-1496.

Hedgecock, D., Lin, J. Z., Decola, S., Haudenschild, C. D., Meyer, E., Manahan, D. T. and B. Bowen (2007): Transcriptomic analysis of growth heterosis in larval Pacific oysters (*Crassostrea gigas*). Proceedings of the National Academy of Sciences of the United States of America 104: 2313-2318.

Hey, J. and R. M. Kliman (1993): Population genetics and phylogenetics of DNA sequence variation at multiple loci within the *Drosophila melanogaster* species complex. Molecular Biology and Evolution 10: 804-822.

Hoekstra, H. E. and J. A. Coyne (2007): The locus of evolution: Evo Devo and the genetics of adaptation. Evolution 61 (5): 995-1016.

Hollocher, H. and C.-I. Wu (1996): The genetics of reproductive isolation in the *Drosophila simulans* clade: X vs. autosomal effects and male vs. female effects. Genetics 143: 1243-1255.

Holloway, A. K., Lawniczak, M. K. N., Mezey, J. G., Begun, D. J. and C. D. Jones (2007): Adaptive Gene Expression Divergence Inferred from Population Genomics. PLoS Genetics 3 (10): 2007-2013.

Hudson, R. R., Slatkin, M. and W. P. Maddison (1992): Estimation of levels of gene flow from DNA sequence data. Genetics 132: 583-589.

Hughes, K. A., Ayroles, J. F., Reedy, M. M., Drnevich, J. M., Rowe, K. C., Ruedi, E. A., Cáceres, C. E. and K. N. Paige (2006): Segregating variation in the transcriptome: *cis*-regulation and additivity of effects. Genetics 173: 1347-1355.

Hunt, W. G. and R. K. Selander (1973): Biochemical genetics of hybridisation in European house mice. Heredity 31: 11-33.

Hurst, L. D. (2002): The Ka/Ks ratio: diagnosing the form of sequence evolution. Trends in Genetics 18 (9): 286-287.

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

Irizarry, R. A., Bolstad, B. M., Collin, F., Cope, L. M., Hobbs, B. and T. P. Speed (2003A): Summaries of Affymetrix GeneChip probe level data. Nucleic Acids Research 31: e15.

Irizarry, R. A., Hobbs, B., Collin, F., Beazer-Barclay, Y. D., Antonellis, K. J., Scherf, U. and T. P. Speed (2003B): Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics 4 (2): 249-264.

Jensen, J. D., Wong, A. and C. F. Aquadro (2007): Approaches for identifying targets of positive selection. Trends in Genetics 32 (11): 568-577.

Ji, W., Zhou, W., Gregg, K., Yu, N. and S. Davis (2004): A method for cross-species gene expression analysis with high-density oligonucleotide arrays. Nucleic Acids Research 32 (11): e93.

Jochum, W., Passegue, E. and E. F. Wagner (2001): AP-1 in mouse development and tumorigenesis. Oncogene 20 (19): 2401-2412.

Johnson, N. A. and A. H. Porter (2000): Rapid speciation via parallel, directional selection on regulatory genetic pathways. Journal of Theoretical Biology 205: 527-542.

Johnson, N. A. and A. H. Porter (2007): Evolution of branched regulatory genetic pathways: directional selection on pleiotropic loci accelerates developmental system drift. Genetica 129: 57-70.

Jordan, I. K., Marino-Ramirez, L., Wolf, Y. I. and E. V. Koonin (2004): Conservation and coevolution in the scale-free human gene coexpression network. Molecular Biology and Evolution 21: 2058-2070.

Jurata, L. W., Bukhman, Y. V., Charles, V., Capriglione, F., Bullard, J., Lemire, A. L., Mohammed, A., Pham, Q., Laeng, P., Brockman, J. A. and C. A. Altar (2004): Comparison of microarray-based mRNA profiling technologies for identification of psychiatric disease and drug signatures. Journal of Neuroscience Methods 138 (1-2): 173-188.

Karn, R. C., Orth, A., Bonhomme, F. and P. Boursot (2002): The complex history of a gene proposed to participate in a sexual isolation mechanism in house mice. Molecular Biology and Evolution 19: 462-471.

Khaitovich, P., Weiss, G., Lachmann, M., Hellmann, I., Enard, W., Muetzel, B., Wirkner, U., Ansorge, W. and S. Pääbo (2004): A neutral model of transcriptome evolution. PLoS Biology 2: 682-689.

Khaitovich, P., Pääbo, S. and G. Weiss (2005): Toward a neutral evolutionary model of gene expression. Genetics 170: 929-939.

Kierzek, R., Burkard, M. E. and D. H. Turner (1999): Thermodynamics of single mismatches in RNA duplexes. Biochemistry 38: 14214-14223.

Kimura, M. (1968): Evolutionary rate at the molecular level. Nature 217: 624-626.

King, M. C. and A. C. Wilson (1975): Evolution at two levels in humans and chimpanzees. Science 188: 107-116.

Knight, J. C. (2005): Regulatory polymorphisms underlying complex disease traits. Journal of Molecular Medicine 83: 97-109.

Kohn, M., Fang, S. and C. Wu (2004): Inference of positive and negative selection on the 5' regulatory regions of *Drosophila* genes. Molecular Biology and Evolution 21 (2): 374-383.

Kuo, W. P., Liu, F., Trimarchi, J., Punzo, C., Lombardi, M., Sarang, J., Whipple, M. E., Maysuria, M., Serikawa, K., Lee, S. Y., McCrann, D., Kang, J., Shearstone, J. R., Burke, J., Park, D. J., Wang, X., Rector, T. L., Ricciardi-Castagnoli, P., Perrin, S., Choi, S., Bumgarner, R., Kim, J. H., Short III, G. F., Freeman, M. W., Seed, B., Jensen, R., Church, G. M., Hovig, E., Cepko, C. L., Park, P., Ohno-Machado, L. and T.-K. Jenssen (2006): A sequence-oriented comparison of gene expression measurements across different hybridization-based technologies. Nature Biotechnology 24 (7): 832-840.

Landry, C. R., Hartl, D. L. and J. M. Ranz (2007): Genome clashes in hybrids: insights from gene expression. Heredity 99: 483-493.

Lemos, B., Bettencourt, B. R., Meiklejohn, C. D. and D. L. Hartl (2005A): Evolution of proteins and gene expression levels are coupled in *Drosophila* and are independently associated with mRNA abundance, protein length, and number of protein-protein interactions. Molecular Biology and Evolution 22: 1345-1354.

Lemos, B., Meiklejohn, C. D., Ca´ Ceres, M. and D. L. Hartl (2005B): Rates of divergence in gene expression profiles of primates, mice, and flies: stabilizing selection and variability among functional categories. Evolution 59 (1): 126-137.

Lenhard, B., Sandelin, A., Mendoza, L., Engström, P., Jareborg, N. and W. W Wasserman (2003): Identification of conserved genome analysis. Journal of Biology 2: 13.

Macholán, M., Munclinger, P., Šugerková, M., Dufková, P., Bímová, B., Božíková, E., Zima, J. and J. Piálek (2007): Genetic analysis of autosomal and X-linked markers across a mouse hybrid zone. Evolution 61 (4): 746-771.

Mackay, T. F. (2001): The genetic architecture of quantitative traits. Annual Review of Genetics 35: 303-339.

Makova, K. D. and W. H. Li (2003): Divergence in the spatial pattern of gene expression between duplicate genes. Genome Research 13: 1638-1645.

Malitschek, B., Fornzler, D. and M. Schartl (1995): Melanoma formation in *Xiphophorus*: A model system for the role of receptor tyrosine kinases in tumorigenesis. BioEssays 17: 1017-1023.

Mallet, J. (2006): What does *Drosophila* genetics tell us about speciation? Trends in Ecology and Evolution 21 (7): 386-393.

Masly, J. P., Jones, C. D., Noor, M. A. F., Locke, J. and H. A. Orr (2006): Gene transposition as a novel cause of hybrid male sterility. Science 313: 1448-1450.

Mayr, E. (1942): Systematics and the Origin of Species. Columbia Univ. Press, New York.

Mayr, E. (1963): Animal Species and Evolution. The Belknap, Cambridge, Massachusetts.

Mayr, E. (1995): Species, classification, and evolution. Pp. 3-12 in Arai, R., Kato, M. and Y. Doi (eds.): Biodiversity and evolution. National Science Museum Foundation, Tokyo.

McCarroll, S. A., Hadnott, T. N., Perry, G. H., Sabeti, P. C., Zody, M. C., Barrett, J. C., Dallaire, S., Gabriel, S. B., Lee, C., Daly, M. J., Altshuler, D. M. and The International HapMap Consortium (2006). Common deletion polymorphisms in the human genome. Nature Genetics 38: 86-92.

McDonald, J. H. and M. Kreitman (1991): Adaptive protein evolution at the A*dh* locus in *Drosophila*. Nature 351: 652-654.

Meiklejohn, C. D., Parsch, J., Ranz, J. M. and D. Hartl (2003): Rapid evolution of male-biased gene expression in Drosophila. Proceedings of the National Academy of Sciences of the United States of America 100: 9894-9899.

Messeguer, X., Escudero, R., Farré, D., Nuñez, O., Martínez, J. and M. M. Albà (2002): PROMO: detection of known transcription regulatory elements using species-tailored searches. Bioinformatics 18 (2): 333-334.

Michalak, P. and M. A. Noor (2003): Genome-wide patterns of expression in *Drosophila* pure species and hybrid males. Molecular Biology and Evolution 20: 1070-1076.

Michalak, P. and M. A. Noor (2004): Association of misexpression with sterility in hybrids of *Drosophila simulans* and *D. mauritiana*. Journal of Molecular Evolution 59: 277-282.

Miron, M., Woody, O. Z., Marcil, A., Murie, C., Sladek, R. and R. Nadon (2006): A methodology for global validation of microarray experiments. BMC Bioinformatics 7: 333.

Modrek, B. and C. J. Lee (2003): Alternative splicing in the human, mouse and rat genome is associated with an increased frequency of exon creation and or loss. Nature Genetics 34 (2): 177-180.

Moehring, A. J., Teeter, K. C. and M. A. Noor (2007): Genome-wide patterns of expression in *Drosophila* pure species and hybrid males. II. Examination of multiple-species hybridizations, platforms, and life cycle stages. Molecular Biology and Evolution 24: 137-145.

Morey, J. S., Ryan, J. C. and F. M. van Dolah (2006): Microarray validation: factors influencing correlation between oligonucleotide microarrays and real-time PCR. Biological Procedures Online 8: 175–193.

Morley, M., Molony, C. M., Weber, T. M., Devlin, J. L., Ewens, K. G., Spielman, R. S. and V. G. Chueng (2004): Genetic analysis of genome-wide variation in human gene expression. Nature 430: 743-747.

Moulia, C., Aussel, J. P., Bonhomme, F., Boursot, P., Nielsen, J. J. and F. Renaud (1991): Wormy mice in a hybrid zone: a genetic control of susceptibility to parasite infection. Journal of Evolutionary Biology 4: 679-687.

Moulia, C., Lebrun, N., Dallas, J., Orth, A. and F. Renaud (1993): Experimental evidence of genetic determinism in high susceptibility to intestinal pinworm infection in mice - a hybrid zone model. Parasitology 106: 387-393.

Müller, B., Link, J. and C. Smythei (2000): Assembly of U7 small nuclear ribonucleoprotein particle and histone RNA 3' processing in *Xenopus* egg extracts. Journal of Biological Chemistry 275 (32): 24284-24293.

Muller, H. J. (1940): Bearing of the *Drosophila* work on systematics. Pp.185-286 in Huxley (ed.): The new systematics. Claredon Press, Oxford.

Muller, H. J. (1942): Isolating mechanisms, evolution, and temperature. Biology Symposium 6: 71-125.

Nei, M. and W. H. Li (1979): Mathematical model for studying genetic variation in terms of restriction endonucleases. Proceedings of the National Academy of Sciences of the United States of America 76: 5269-5273.

Nei, M. and T. Gojobori (1986): Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Molecular Biology and Evolution 3 (5): 418-426.

Nielsen, R., Bustamante, C., Clark, A. G., Glanowski, S., Sackton, T. B., Hubisz, M. J., Fledel-Alon, A., Tanenbaum, D. M., Civello, D., White, T. J., Sninsky J. J., Adams, M. D. and M. Cargill (2005): A scan for positively selected genes in the genomes of humans and chimpanzees. PLoS Biology 3: 976-985.

Nieto-Díaz, M., Pita-Thomas, W. and M. Nieto-Sampedro (2007): Cross-species analysis of gene expression in non-model mammals: reproducibility of hybridization on high density oligonucleotide microarrays. BMC Genomics 8: 89.

Noor, M. A. F. and J. L. Feder (2006): Speciation genetics: evolving approaches. Nature Genetics 7: 851-861.

Nuzhdin, S. V., Wayne, M. L., Harmon, K. L. and L. M. McIntyre (2004): Common patterns of evolution of gene expression level and protein sequence in *Drosophila*. Molecular Biology and Evolution 21: 1308-1317.

Oka, A., Aoto, T., Totsuka, Y., Takahashi, R., Ueda, M., Mita, A., Sakurai-Yamatani, N., Yamamoto, H., Kuriki, S., Takagi, N., Moriwaki, K. and T. Shiroishi (2007): Disruption of genetic interaction between two autosomal regions and the X chromosome causes reproductive isolation between mouse strains derived from different subspecies. Genetics 175: 185-197.

Ometto, L., Glinka, S., De Lorenzo, D. and W. Stephan (2005): Inferring the effects of demography and selection on *Drosophila melanogaster* populations from a chromosome-wide scan of DNA variation. Molecular Biology and Evolution 22: 2119-2130.

Orr, H. A. (1992): Mapping and characterization of a speciation gene' in *Drosophila*. Genetical Research 59: 73-80.

Orr, H. A. (1993): Haldane's rule has multiple genetic causes. Nature 361: 532-533.

Orr, H. A (1997): Haldane's rule. Annual Review of Ecology and Systematics 28: 195-218.

Orr, H. A. (2005): The genetic basis of postzygotic reproductive isolation: insights from *Drosophila*. Proceedings of the National Academy of Sciences of the United States of America 102: 6522-6526.

Orr, H. A. and J. A. Coyne (1989): The genetics of postzygotic isolation in the *Drosophila viridis* group. Genetics 121: 527-537.

Orr, H. A. and D. C. Presgraves (2000): Speciation by postzygotic isolation: forces, genes and molecules. BioAssays 22: 1085-1094.

Orr, H. A., Masly, J. P. and D. C. Presgraves (2004): Speciation genes. Current Opinion in Genetics and Development 14: 675-679.

Ortíz-Barrientos, D., Counterman, B. A. and M. A. F. Noor (2007): Gene expression divergence and the origin of hybrid dysfunctions. Genetica 129: 71-81.

Otto, S. P. (2000): Detecting the form of selection from DNA sequence data. Trends in Genetics 16 (12): 526-529.

Patterson, T. A., Lobenhofer, E. K., Fulmer-Smentek, S. B., Collins, P. J., Chu, T. M., Bao, W., Fang, H., Kawasaki, E. S., Hager, J., Tikhonova, I. R., Walker, S. J., Zhang, L., Hurban, P., de Longueville, F., Fuscoe, J. C., Tong, W., Shi, L. and R. D. Wolfinger (2006): Performance comparison of one-color and two-color platforms within the MicroArray Quality Control (MAQC) project. Nature Biotechnology 24: 1140-1150.

Payseur, B. A., Krenz, J. G. and M. W. Nachman (2004): Differential patterns of introgression across the X chromosome in a hybrid zone between two species of house mice. Evolution 58 (9): 2064-2078.

Payseur, B. A. and M. W. Nachman (2005): The genomics of speciation: investigating the molecular correlates of X chromosome introgression across the hybrid zone between *Mus domesticus* and *Mus musculus*. Biological Journal of the Linnean Society 84: 523-534.

Payseur, B. and H. E. Hoekstra (2005): Signatures of reproductive isolation in patterns of single nucleotide diversity across inbred strains of mice. Genetics 171: 1905-1916.

Payseur, B. A. and M. Place (2007): Searching the genomes of inbred mouse strains for incompatibilities that reproductively isolate their wild relatives. Journal of Heredity 98 (2): 115-122.

Pillai, R. S., Will, C. L., Lührmann, R., Schümperli, D. and B. Müller (2001): Purified U7 snRNPs lack the Sm proteins D1 and D2 but contain Lsm10, a new 14kDa Sm D1-like protein. The EMBO Journal 20 (19): 5470-5479.

Pocock, M. J. O., Hauffe, H. C. and J. B. Searle (2005): Dispersal in house mice. Biological Journal of the Linnean Society 84 (3): 565-583.

Porter, A. H. and N. A. Johnson (2002): Speciation despite gene flow when developmental pathways evolve. Evolution 56: 2103-2111.

Pozhitkov, A., Noble, P. A., Domazet-Loso, T., Nolte, A. W., Sonnenberg, R., Staehler, P., Beier, M. and D. Tautz (2006): Tests of rRNA hybridization to microarrays suggest that hybridization characteristics of oligonucleotide probes for species discrimination cannot be predicted. Nucleic Acids Research 34 (9): e66.

Pozhitkov, A. E., Tautz, D. and P. A. Noble (2007): Oligonucleotide microarrays: widely applied, poorly understood. Briefings in functional Genomics and Proteomics 6 (2): 141-148.

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

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

Presgraves, D. C. (2003): A fine-scale genetic analysis of hybrid incompatibilities in *Drosophila*. Genetics 163: 955-972.

Qin, L. X., Beyer, R. P., Hudson, F. N., Linford, N. J., Morris, D. E. and K. F. Kerr (2006): Evaluation of methods for oligonucleotide array data via quantitative real-time PCR. BMC Bioinformatics 7: 23.

Qiu, P. (2003): Recent advances in computational promoter analysis in understanding the transcriptional regulatory network. Biochemical and Biophysical Research Communications 309: 495-501.

Rajeevan, M. S., Vernon, S. D., Taysavang, N. and E. R. Unger (2001): Validation of array-based gene expression profiles by real-time (kinetic) RT-PCR. Journal of Molecular Diagnostics 3 (1): 26-31.

Ranz, J. M., Castillo-Davis, C. I., Meikeljohn, C. D. and D. L. Hartl (2003): Sex-dependent gene expression and evolution of the *Drosophila* transcriptome. Science 300: 1742–1745.

Ranz, J. M., Namgyal, K., Gibson, G. and D. L. Hartl (2004): Anomalies in the expression profile of interspecific hybrids of *Drosophila melanogaster* and *Drosophila simulans*. Genome Research 14: 373-379.

Ranz, J. M. and C. A. Machado (2006): Uncovering evolutionary patterns of gene expression using microarrays. Trends in Ecology and Evolution 21 (1): 29-37.

Raufaste, N., Orth, A., Belkhir, K., Senet, D., Smadja, C., Bird, S. J. E., Bonhomme, F., Dod, B. and P. Boursot (2005): Inferences of selection and migration in the Danish house mouse hybrid zone. Biological Journal of the Linnean Society 84: 593-616.

Reiland, J. and M. A. Noor (2002): Little qualitative RNA misexpression in sterile male F1 hybrids of *Drosophila pseudoobscura* and *D. persimilis*. BMC Evolutionay Biology 2: 16.

Reiner, A., Yekutieli, D. and Y. Benjamini (2003): Identifying differentially expressed genes using false discovery rate controlling procedures. Bioinformatics 19: 368-375.

Rice, W. R. (1998): Intergenomic conflict, interlocus antagonistic coevolution, and the evolution of reproductive isolation. Pp. 261-270 in Howard, D. J. and S. Berlocher (eds.): Endless Forms: Species and Speciation. Oxford University Press, New York.

Rieseberg, L. H. (1991): Homoploid reticulate evolution in *Helianthus* (Asteraceae): Evidence from ribosomal genes. American Journal of Botany 78: 1218-1237.

Rieseberg, L. H. (1997): Hybrid origin of plant species. Annual Review of Ecology and Systematics 28: 359-389.

Rieseberg, L. H., Whitton, J. and K. Gardner (1999): Hybrid zones and the genetic architecture of a barrier to gene flow between two sunflower species. Genetics 152: 713-727.

Rieseberg, L. H., Raymond, O., Rosenthal, D. M., Lai, Z., Livingstone, K., Nakazato, T., Durphy, J. L., Schwarzbach, A. E., Donovan, L. A. and C. Lexer (2003): Major ecological transitions in wild sunflowers facilitated by hybridization. Science 301: 1211-1216.

Rifkin, S. A., Kim, J. and K. P. White (2003): Evolution of gene expression in the *Drosophila melanogaster* subgroup. Nature Genetics 33: 138-144.

Rockman, M. V. and G. A. Wray (2002): Abundant raw material for *cis*-regulatory evolution in humans. Molecular Biology and Evolution 19: 199-204.

Rockman, M. V. and L. Kruglyak (2006): Genetics of global gene expression. Nature Reviews Genetics 7: 862-872.

Rozas, J., Sánchez-DelBarrio, J. C., Messeguer, X. and R. Rozas (2003): DnaSP, DNA polymorphism analyses by the coalescent and other methods. Bioinformatics 19: 2496-2497.

Sage, R. D. (1981): Wild mice. Pp 39-90 in Foster, H. L., Small, J. D. and J. G. Fox (eds.): The mouse in biomedical research, Vol. 1. Academic Press, New York.

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

Sage, R. D., Atchley, W. R. and E. Capanna (1993): House mice as models in systematic biology. Systematic Biology 42: 523-561.

Sawamura, K., Davis, A. W. and C.-I. Wu (2000): Genetic analysis of speciation by means of introgression into *Drosophila melanogaster*. Proceedings of the National Academy of Sciences of the United States: 2652-2655.

Schadt, E. E., Monks, S. A., Drake, T. A, Aldons, J. L., Chek, N., Colinayok, V., Ruff, T. G., Milligan, S. B., Lamb, J. R., Cavet, G., Linsley, P. S., Mao, M., Stoughton, R. B. and S. H. Friend (2003): Genetics of gene expression surveyed in maize, mouse and man. Nature 422: 297-302.

Schartl, M., Hornung, U., Gutbrod, H., Volff, J.-N. and J. Wittbrodt (1999): Melanoma loss-of-function mutants in *Xiphophorus* caused by *Xmrk*-oncogene deletion and gene disruption by a transposable element. Genetics 153: 1385-1394.

Shapiro, M. D., Marks, M. E., Peichel, C. L., Blackman, B. K., Nereng, K. S., Jonsson, B., Schluter, D. and D. M. Kingsley (2004): Genetic and developmental basis of evolutionary pelvic reduction in threespine sticklebacks. Nature 428: 717-723.

Simonsen, K. L., Churchill, G. A. and C. F. Aquadro (1995): Properties of statistical tests of neutrality for DNA polymorphism data. Genetics 141 (1): 413-429.

Smadja, C. and G. Ganem (2002): Subspecies recognition in the house mouse: a study of two populations from the border of a hybrid zone. Behavioral Ecology 13: 312-320.

Smadja, C. and G. Ganem (2005): Asymmetrical reproductive character displacement in the house mouse. Journal of Evolutionary Biology 18: 1485-1493.

Smith, H. O., Tabiti, K., Schaffner, G., Soldati, D., Albrecht, U. and M. L. Birnstiel (1991): Two-step affinity purification of U7 small nuclear ribonucleoprotein particles using complementary biotinylated 2'-O-methyl oligoribonucleotides. Proceedings of the National Academy of Sciences of the United States of America 88: 9784-9788.

Storchová, R., Gregorová, S., Buckiová, D., Kyselová, V., Divina, P. and J. Forejt (2004): Genetic analysis of X-linked hybrid sterility in the house mouse. Mammalian Genome 15: 515-524.

Storey, J. D., Akey, J. M. and L. Kruglyak (2005): Multiple locus linkage analysis of genomewide expression in yeast. PLoS Biology 3: e267.

Stupar, R. M. and N. M. Springer (2006): *Cis*-transcriptional variation in maize inbred lines B73 and Mo17 leads to additive expression patterns in the F1 hybrid. Genetics 173: 2199-2210.

Su, A. L., Wiltshire, T., Batalov, S., Lapp, H., Ching, K. A., Block, D., Zhang, J., Soden, R., Hayakawa, M., Kreiman, G., Cooke, M. P., Walker, J. R. and John B.

Hogenesch (2004): A gene atlas of the mouse and human protein-encoding transcriptomes. Proceedings of the National Academy of Sciences of the United States of America 101: 6062-6067.

Sucena, E. and D. L. Stern (2000): Divergence of larval morphology between Drosophila sechellia and its sibling species caused by *cis*-regulatory evolution of ovo/shaven-baby. Proceedings of the National Academy of Sciences of the United States of America 97: 4530-4534.

Sun, S., Ting, C.-T. and C.-I. Wu (2004): The normal function of a speciation gene, *Odysseus*, and its hybrid sterility effect. Science 305: 81-83.

Swanson-Wagner, R. A., Jia, Y., DeCook, R., Borsuk, L. A., Nettleton, D. and P. S. Schnable (2006): All possible modes of gene action are observed in a global comparison of gene expression in a maize F1 hybrid and its inbred parents. Proceedings of the National Academy of Sciences of the United States of America 103: 6805-6810.

Sweigart, A. L., Mason, A. R. and J. H. Willis (2007): Natural variation for a hybrid incompatibility between two species of *Mimulus*. Evolution 61: 141-151.

Szymura, J. M. and N. H. Barton (1986): Genetic analysis of a hybrid zone between the fire-bellied toads, *Bombina bombina* and *Bombina variegata*, near Cracow in southern Poland. Evolution 40: 1141-1159.

Tajima, F. (1989): Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123: 585-595.

Teeter, K. C., Payseur, B. A., Harris, L. W., Bakewell, M. A., Thibodeau, L. M., O'Brien, J. E., Krenz, J. G., Sans-Fuentes, M. A., Nachman, M. W. and P. K. Tucker (2007): Genome-wide patterns of gene flow across a house mouse hybrid zone. Genome Research: doi:10.1101/gr.6757907.

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

Ting, C.-T., Tsaur, S.-C., Wu, M.-L. and C.-I. Wu (1998): A rapidly evolving homeobox at the site of a hybrid sterility gene. Science 282: 1501-1504.

Ting, C.-T., Tsauer, S. C., Sun, S., Brown, W.E., Chen, Y. C., Patel, N. H. and C.-I. Wu (2004): Gene duplication and speciation in *Drosophila*: evidence from the *Odysseus* locus. Proceedings of the National Academy of Sciences of the United States of America 101: 12232-12235.

True, J. R. and E. S. Haag (2001): Developmental system drift and flexibility in evolutionary trajectories. Evolution and Development 3: 109-119.

Tucker, P. K., Sage, R. D., Warner, J., Wilson, A. C. and E. M. Eicher (1992): Abrupt cline for sex chromosomes in a hybrid zone between two species of mice. Evolution 46: 1146-1163.

Turelli, M. and H. A. Orr (2000): Dominance, epistasis and the genetics of postzygotic isolation. Genetics 154: 1663-1679.

Vanlerberghe, F., Dod, B., Boursot, P., Bellis, M. and F. Bonhomme (1986): Absence of Y chromosome introgression across the hybrid zone between *Mus musculus domesticus* and *Mus musculus musculus*. Genetical Research 48: 191-197.

Van Zegeren, K. and G. A. van Oortmerssen (1981): Frontier disputes between the West- and East-European house mouse in Schleswig-Holstein, West Germany. Zeitschrift für Säugetierkunde 46: 363-369.

Voolstra, C., Tautz, D., Farbrother, P., Eichinger, L. and B. Harr (2007): Contrasting evolution of expression differences in the testis between species and subspecies of the house mouse. Genome Research 17: 42-49.

Vuylsteke, M., van Eeuwijk, F., van Hummelen, P., Kuiper, M. and M. Zabeau (2005): Genetic analysis of variation in gene expression in *Arabidopsis thaliana*. Genetics 171: 1267-1275.

Vyskocilova, M., Trachtulec, Z., Forejt, J. and J. Pialek (2005): Does geography matter in hybrid sterility in house mice? Biological Journal of the Linnean Society 84: 663-674.

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

Wagner, A. (2000): Decoupled evolution of coding region and mRNA expression patterns after gene duplication: implications for the neutralist-selectionist debate. Proceedings of the National Academy of Sciences of the United States of America 97: 6579-6584.

Walker, N. J. (2002): A technique whose time has come. Science 296: 557-559.

Waterston, R. H., Lindblad-Toh, K., Birney, E. et al. (2002): Initial sequencing and comparative analysis of the mouse genome. Nature 420: 520-562.

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

Wayne, M. L., Pan, Y. J., Nuzhdin, S. V. and L. M. McIntyre (2004): Additivity and *trans*-acting effects on gene expression in male *Drosophila simulans*. Genetics 168: 1413-1420.

West, M. A. L., Kim, K., Kliebenstein, D. J., van Leeuwen, H., Michelmore, R. W., Doerge, R. W. and D. A. St. Clair (2007): Global eQTL mapping reveals the complex genetic architecture of transcript-level variation in *Arabidopsis*. Genetics 175: 1441-1450.

Whitehead, A. and D. L. Crawford (2006): Variation within and among species in gene expression: raw material for evolution. Molecular Ecology 15: 1197-1201.

Wittbrodt, J., Adam, D., Malitschek, B., Mäueler, W., Raulf, F., Telling, A., Robertson, S. M. and M. Schartl (1989): Novel putative receptor tyrosine kinase encoded by the melanoma-inducing *Tu* locus in *Xiphophorus*. Nature 341: 415-421.

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

Wolfe, K. H. and P. M. Sharp (1993): Mammalian gene evolution: nucleotide sequence divergence between mouse and rat. Journal of Molecular Evolution 37: 441-456.

Wray, G. A., Hahn, M. W., Abouheif, E., Balhoff, J. P., Pizer, M., Rockman, M. V. and L. A. Romano (2003): The evolution of transcriptional regulation in eukaryotes. Molecular Biology and Evolution 20 (9): 1377-1419.

Wray, G. A. (2007): The evolutionary significance of *cis*-regulatory mutations. Nature Reviews Genetics 8: 206-216.

Wright, S. (1978): Evolution and the genetics of populations. Vol. 4: Variability within and among natural populations. University of Chicago Press, Chicago.

Wu, C.-I. (2001): The genic view of speciation. Journal of Evolutionary Biology 14: 851-865.

Wu, C. I. and C. T. Ting (2004): Genes and speciation. Nature Reviews Genetics 5: 114-122.

Wu, C., Carta, R. and L. Zhang (2005): Sequence dependence of cross-hybridization on short oligo microarrays. Nucleic Acids Research 33 (9): e84.

Xing, Y. and C. Lee (2005): Evidence of functional selection pressure for alternative splicing events that accelerate evolution of protein subsequences. Proceedings of the National Academy of Sciences of the United States of America 102 (38): 13526-13531.

Yan, H., Yuan, W., Velculescu, V. E., Vogelstein, B. and K. W. Kinzler (2002): Allelic variation in human gene expression. Science 297 (5584): 1143.

Yang, H., Bell, T. A., Churchill, G. A. and F. Pardo-Manuel De Villena (2007): On the subspecific origin of the laboratory mouse. Nature Genetics 39: 1100-1107.

Yonekawa, H., Moriwaki, K., Gotoh, O., Miyashita, N., Matsushima, Y., Shi, L. M., Cho, W. S., Zhen, X. L. and Y. Tagashira (1988): Hybrid origin of Japanese mice *Mus musculus molossinus*: evidence from restriction analysis of mitochondrial DNA. Molecular Biology and Evolution 5: 63-78.

Yvert, G., Brem, R. B., Whittle, J., Akey, J. M., Foss, E., Smith, E. N., Mackelprang, R. and L. Kruglyak (2003): *Trans*-acting regulatory variation in *Saccharomyces cerevisiae* and the role of transcription factors. Nature Genetics 35: 57-64.

Zhang, J., Finney, R. P., Clifford, R. J., Derr, L. K. and K. H. Buetow (2005): Detecting false expression signals in highdensity oligonucleotide arrays by an in silico approach. Genomics 85: 297-308.

6 Supplement

- Supplement 1: Used mouse samples for microarray and qRT-PCR analysis.
- Supplement 2: Overrepresented biological processes among the differentially expressed transcripts between *M. m. domesticus* and *M. m. musculus*.
- Supplement 3: Overrepresented biological processes among the differentially expressed transcripts between *M. m. castaneus* and *M. m. musculus.*
- Supplement 4: Overrepresented biological processes among the nonadditively expressed transcripts between F1 hybrid from a cross between *M. m. musculus* mother with *M. m. castaneus* father and both parents in the testis.
- Supplement 5: Literature review on studies differentiating additivity and nonadditivity in gene expression.
- Supplement 6: Candidate genes used for qRT-PCR confirmation experiment.
- Supplement 7: Primer sequences of the different experiments.
- Supplement 8: Results of TaqMan® Gene Expression assay qRT-PCR analysis.
- Supplement 9: Results of Sybr Green qRT-PCR analysis.
- Supplement 10: Differences in single probes related to signal intensities.
- Supplement 11: Sequencing results of amplicon regions.
- Supplement 12: List of samples and geographic origin of the animal material used for sequencing.
- Supplement 13: Overview of sampling area and trapping localities for the hybrid zone analysis.
- Supplement 14: Primer combination and sequences for Lsm10 and Neil3.
- Supplement 15: Primer used for expression-phenotype/genotype correlative analysis.
- Supplement 16: Samples used for hybrid zone analysis.
- Supplement 17: Locus-specific genotype in relation to expression phenotype.
- Supplement 18: Nucleotide polymorphism in the 5' upstream flanking region of Lsm10 in *M. m. domesticus* and *M. m. musculus* populations.
- Supplement 19: Nucleotide polymorphism in the 5' upstream flanking region of Neil3 in *M. m. domesticus* and *M. m. musculus* populations.

Supplement 1: Used mouse samples for microarray and qRT-PCR analysis.

Subspecies	Generation	Geographic origin	Sample name	experiment
M. m. musculus, JPC 2821	~13	Studenec, Czech Republic, wild-derived inbred	24.01	qRT-PCR
			24.02	GeneChip®, qRT-PCR
			44.01	GeneChip® , qRT-PCR
			44.02	qRT-PCR
M. m. domesticus, JPC 2705	~13	Straas, Germany, wild-derived inbred	20.01	qRT-PCR
			20.02	qRT-PCR
			20.03	GeneChip®, qRT-PCR
			20.04	GeneChip®, qRT-PCR
			23.01	qRT-PCR
			23.02	qRT-PCR
			23.03	qRT-PCR
M. m. castaneus, CIM	~30	Masinagudi, India, wild derived inbred	07.02	GeneChip®,
			07.03	GeneChip®
mus-dom hybrid	1		14.01	GeneChip®, qRT-PCR
			14.02	qRT-PCR
			14.03	GeneChip®, qRT-PCR
			17.01	qRT-PCR
			17.02	qRT-PCR
			17.03	qRT-PCR
dom-mus hybrid	1		29.01	qRT-PCR
			29.02	GeneChip®, qRT-PCR
			29.03	GeneChip®, qRT-PCR
cas-mus hybrid	1		38.02	GeneChip®,
			38.03	GeneChip®
mus-cas hybrid	1		05.01	GeneChip®
			05.02	GeneChip®
M. m. domesticus	1, parents collected in the wild	Heimerzheim, Germany	TP1a.1	qRT-PCR
		Heimerzheim, Germany	TP1a.2	qRT-PCR
		Ardenberg, Germany	TP3a.1	qRT-PCR
		Ardenberg, Germany	TP3a.2	qRT-PCR
		Arzdorf, Gemany	TP5.1	qRT-PCR
		Niederbachem, Germany	TP8b.2	qRT-PCR
M. m. musculus	1, parents collected in the wild	Gänserndorf, Austria	W1a.1	qRT-PCR
		Vienna, Austria	W3.1	qRT-PCR
		Vienna, Austria	W3.3	qRT-PCR
		Vienna, Austria	W4a.1	qRT-PCR
		Vienna, Austria	W4b.1	qRT-PCR
		Punkersdorf, Austria	W6a.1	qRT-PCR

Supplement 2: Overrepresented biological processes among the differentially expressed transcripts between *M. m. domesticus* and *M. m. musculus* (p<0.05, fold-change >2.5), independent of expression status in the F1 hybrids. Bonferroni-corrected p-values are depicted.

Biological process	**Observed number of genes in category**	**Expected number of genes in category**	**p-value**
brain			
intracellular protein traffic	18	7.01	0.00849
protein modification	22	8.73	0.01070
liver			
lipid, fatty acid and steroid metabolism	39	10.55	<0.00001
steroid metabolism	23	3.37	<0.00001
cholesterol metabolism	11	0.97	<0.00001
electron transport	16	4.3	0.00030
other metabolism	22	7.52	0.00031
detoxification	8	1.13	0.00329
proteolysis	31	13.81	0.00446
protein metabolism and modification	69	45.83	0.01060
coenzyme and prosthetic group metabolism	9	2.14	0.01110
immunity and defense	42	24.16	0.01190
carbohydrate metabolism	18	7.3	0.01540
testis			
protein targeting and localization	11	3.12	0.01170
intracellular protein traffic	29	15.01	0.02120

Supplement 3: Overrepresented biological processes among the differentially expressed transcripts between *M. m. castaneus and M. m. musculus* (p<0.05, fold-change >2.5), independent of expression status in the F1 hybrids. Bonferroni-corrected p-values are depicted.

Biological process	**Observed number of genes in category**	**Expected number of genes in category**	**p-value**
brain			
carbohydrate metabolism	16	4.29	0.000257
other polysaccharide metabolism	7	1.05	0.015200
liver			
lipid, fatty acid and steroid metabolism	39	11.11	<0.000001
steroid metabolism	18	3.55	0.000005
other transport	8	0.76	0.000189
other metabolism	23	7.92	0.000225
immunity and defense	49	25.43	0.000302
cholesterol metabolism	8	1.02	0.002260
coenzyme and prosthetic group metabolism	10	2.25	0.003470
transport	36	18.38	0.003600
carbohydrate metabolism	20	7.68	0.003840

steroid hormone-mediated signaling	6	0.56	0.004890
amino acid metabolism	11	3.13	0.012000
protein metabolism and modification	71	48.25	0.017400
testis			
detoxification	10	1.33	0.000195

Supplement 4: Overrepresented biological processes among the nonadditively expressed transcripts between F1 hybrid from a cross between *M. m. musculus* mother with *M. m. castaneus* father and both parents in the testis. Bonferroni-corrected p-values are depicted.

Biological process	**Observed number of genes in category**	**Expected number of genes in category**	**p-value**
intracellular protein traffic	164	83.05	$2.26x10^{14}$
cell structure and motility	153	89.74	$1.03\ x10^{8}$
cell cycle	140	79.95	$1.12\ x10^{8}$
protein modification	173	103.42	$1.33\ x10^{8}$
spermatogenesis and motility	37	12.81	$4.96\ x10^{6}$
gametogenesis	51	21.48	$5.29\ x10^{6}$
nucleoside, nucleotide and nucleic acid metabolism	393	306.36	$5.50\ x10^{6}$
other intracellular protein traffic	21	5.25	$2.47\ x10^{5}$
protein metabolism and modification	385	303.81	$2.52\ x10^{5}$
pre-mRNA processing	56	26.41	$4.41\ x10^{5}$
lipid, fatty acid and steroid metabolism	112	69.93	$4.66\ x10^{5}$
protein targeting and localization	40	17.26	$5.66\ x10^{5}$
protein phosphorylation	100	59.67	$1.63\ x10^{4}$
cell structure	90	53.94	$4.93\ x10^{4}$
mRNA splicing	43	19.57	$5.68\ x10^{4}$
lipid metabolism	31	12.33	$7.71\ x10^{4}$
general vesicle transport	45	21.48	$8.17\ x10^{4}$
miscellaneous	30	13.44	$2.00\ x10^{3}$
mitosis	55	29.75	$2.76\ x10^{3}$
exocytosis	32	14.24	$4.78\ x10^{3}$
other intracellular signaling cascade	38	18.54	$8.92\ x10^{3}$
protein folding	32	14.8	$9.65\ x10^{3}$
intracellular signaling cascade	106	72.31	$1.38\ x10^{2}$
carbohydrate metabolism	73	48.37	$1.56\ x10^{2}$

Supplement 5: Literature review on studies differentiating additivity and nonadditivity in gene expression. Shaded in grey indicate the studies that have less rigorously investigated the topic of additivity/nonadditivity.

Organism	Divergence time (generations)	Reference	Method	One dye vs. two dyes	Inbreeding status	Tissue complexity	Additivity vs. nonadditivity
Studies consistent with additivity predominant							
Drosophila	within species	Hughes et al. 2006	Affymetrix microarray	one dye	no	whole fly	add>>nonadd
Maize	within species	Stupar and Springer 2006	Affymetrix microarray	one dye	high	whole plant	add>>nonadd
Maize	within species	Swanson-Wagner et al. 2006	cDNA microarray	two dyes	high	all above ground tissues	add>>nonadd
M. musculus lab strains	within species	Cui et al. 2006	Affymetrix microarray	one dye	high	single tissue	add>>nonadd
M. musculus wild strains	2 - 3 Mio*	this study	Affymetrix microarray	one dye	partially	single tissue	add>>nonadd
D. melanogaster vs. *D. simulans*	12.5 - 34 Mio[#]	Ranz et al. 2004	cDNA microarray	two dyes	high	head	add>nonadd
D. simulans vs. *D. mauritiana*	2.5 - 8 Mio[#]	Michalak and Noor 2003	Affymetrix microarray	one dye	high	whole fly	nonadd small
D. persimilis vs. *D. pseudoobscura*	2.5 -5 Mio[#]	Reiland and Noor 2002	differential display PCR	no dye	high	whole fly	nonadd small
Studies consistent with nonadditivity predominant							
Drosophila	within species	Gibson et al. 2004	Agilent microarray	two dyes	high	whole fly	nonadd>add
Arabidopsis	within species	Vuylsteke et al. 2005	cDNA microarray	two dyes	high	first leaf pair	add = nonadd
Oyster	within species	Hedgecock et al. 2007	massive parallel signature sequencing	no dye	partially (f=0.375)	1-2 Mio larvae	nonadd>>add
D. melanogaster vs *D. simulans*	12.5 - 34 Mio[#]	Ranz et al. 2004	cDNA microarray	two dyes	high	body	nonadd>add

* assuming 2-3 generations/year (Karn et al. 2002)

[#] assuming 5-10 generations/year (Hey and Kliman 1993)

Supplement 6: Candidate genes used for qRT-PCR confirmation experiment.

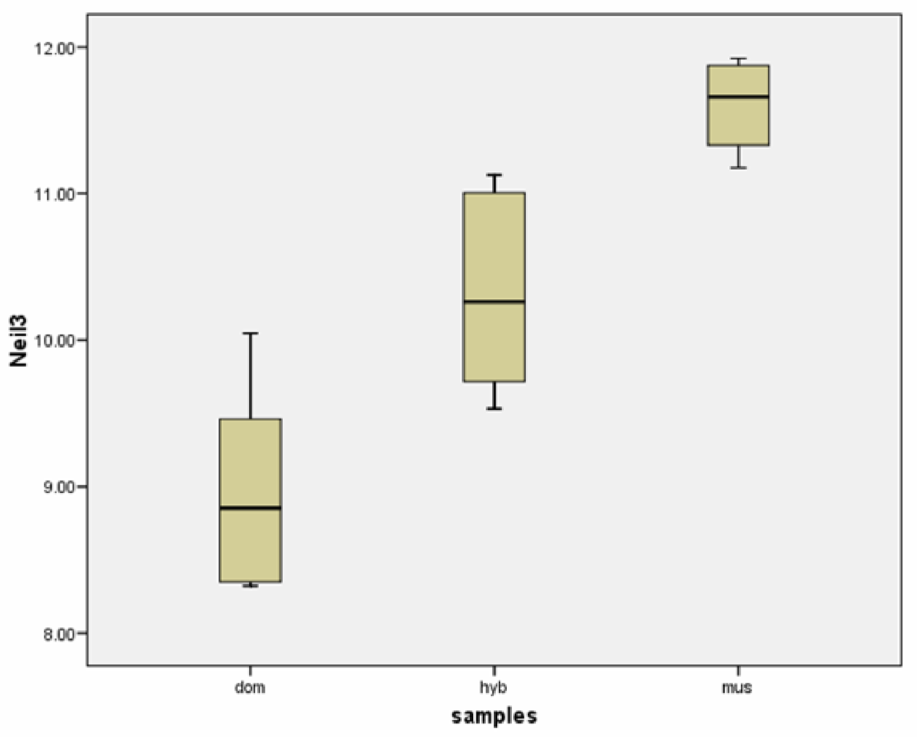
Affymetrix ID	Gene symbol	Refseq	Assay ID	Experiment
1460235_at	Scarb2	NM_007644.2	Mm00446978_m1	TaqMan®, Sybr Green
1421113_at	Pga5	NM_021453.2	Mm00480598_m1	TaqMan®, Sybr Green
1424060_at	Neil3	NM_146208.1	Mm00467596_m1	TaqMan®, Sybr Green
1453203_at	1700011K15Rik	NM_029294.1	Mm00661433_s1	TaqMan®, Sybr Green
1455939_x_at	Srp14	NM_009273.4	Mm00726104_s1	TaqMan®, Sybr Green
1419715_at	1700029F12Rik	NM_025585.1	Mm00481622_m1	TaqMan®, Sybr Green
1429661_at	Rhobtb3	NM_028493.1	Mm00712630_m1	TaqMan®, Sybr Green
1428437_at	Lsm14a	NM_025948	–	Sybr Green
1417515_at	Lsm10	NM_138721	–	Sybr Green
1452877_at	2700029M09Rik	XM_910498	–	Sybr Green
Endogenous control	Eif4g2	NM_013507.2	Mm00469036_m1	TaqMan®, Sybr Green

Supplement 7: Primer sequences of the different experiments.

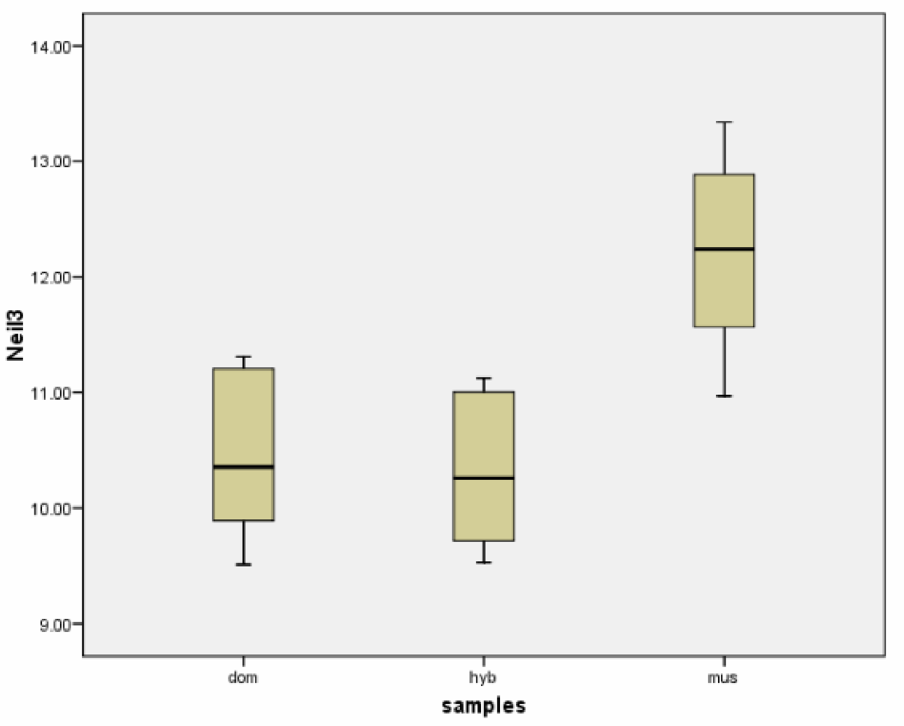
RefSeq	Name	F-Primer	Sequence	R-Primer	Sequence	Experiment
NM_007644	Scarb2	1394	tttaccaagccgacgagaag	1395	gcccaaccacaaaaagtttc	Sybr Green
NM_021453	Pga5	1514	tgctgatgctaggtggagtg	1515	cctcctcaaagttgctcct	Sybr Green
NM_146208	Neil3	1778	caaggggaggcagttttatg	1779	ttcataatggagcgcttgc	Sybr Green
NM_029294	1700011 K15Rik	1495	tgaatgtggattttgccttg	1496	atcatggcgatgtcaatcac	Sybr Green
NM_009273	Srp14	1499	gcaaaccagcacagtgacag	1500	caaaccttacggctggaatc	Sybr Green
NM_025585	1700029 F12Rik	1516	gacctccaaaacccgtgac	1517	gcaggccagatttagagcac	Sybr Green
NM_028493	Rhobtb3	1497	caaaagcctgaatttcaagacc	1498	gcaatagggcaacgtaaagg	Sybr Green
NM_025948	Lsm14a	1503	tccatatcccagatggtgtg	1504	tcacagccagagttgacgac	Sybr Green

NM_138721	Lsm10	1501	tcatgcatagtgtatggacttcg	1502	gagtcattagaaggccacagg	Sybr Green
XM_910498	2700029 M09Rik	1505	caggcttggttgttccagtag	1506	aacagggcagcttgttcatc	Sybr Green
NM_013507	Eif4g2	1523	agcagaagcacgtcagtaagg	1524	gcacataaagcccattactagagg	Sybr Green
NM_007644	Scarb2	1535	tcaggggtgttgaacatcag	1536	atctacaacagggggtcacg	GeneChip$^{®}$ target region
NM_021453	Pga5	1398	agaacctggcattttcatgg	1401	tggatgaaatatgccctgtg	GeneChip$^{®}$ target region
NM_146208	Neil3	1402	ccagctcacacttctgcaac	1405	ttcataatggagcgcttgc	GeneChip$^{®}$ target region
NM_029294	1700011 K15Rik	1406	agccagctgtgctgaagtg	1407	caaagacgccagaatgaaatg	GeneChip$^{®}$ target region
NM_009273	Srp14	1410	gcaaaccagcacagtgacag	1509	ccacctgtaggataacacaactttt	GeneChip$^{®}$ target region
NM_025585	1700029 F12Rik	1413	gctacaccccaaaactgacg	1414	ggtcccccattaccaagatag	GeneChip$^{®}$ target region
NM_028493	Rhobtb3	1416	tcatcacacagctgcagagc	1417	gaatcccatggttacatttgg	GeneChip$^{®}$ target region
NM_025948	Lsm14a	1420	gtcagtgtgctgaggagcag	1421	tgaaatctttggatctctttattcc	GeneChip$^{®}$ target region
NM_138721	Lsm10	1424	accttcgggatgagagtgtg	1425	gaaaagaaacaaatgccaaaaag	GeneChip$^{®}$ target region
XM_910498	2700029 M09Rik	1429	acaatgtgtttgctgccatc	1430	taaaaacaaagccccgtctg	GeneChip$^{®}$ target region
NM_007644	Scarb2	1481	cagaaggcggtagaccagac	1482	tccacgacagtcaacagagg	TaqMan$^{®}$ amplicon region
NM_021453	Pga5	1561 1567	gagaggcacctggcacttac aggtgtggttcctctggttg	1566 1568	ggaggttatgaaggccactg tctcagtccctgctccctac	TaqMan$^{®}$ amplicon region
NM_146208	Neil3	1485	tcttcatccggctgttaagg	1486	caaaccacacaggaccactg	TaqMan$^{®}$ amplicon region
NM_029294	1700011 K15Rik	1487	atccttgctgaagccatcc	1488	ttgaagtttcccacggactc	TaqMan$^{®}$ amplicon region
NM_009273	Srp14	1410	gcaaaccagcacagtgacag	1509	ccacctgtaggataacacaactttt	TaqMan$^{®}$ amplicon region
NM_025585	1700029 F12Rik	1489	tgtgacagtgaagccaccac	1490	tgtggacagctggtcttctg	TaqMan$^{®}$ amplicon region
NM_028493	Rhobtb3	1491	acatgccccctgtgtacttc	1492	caagctgaggaggtcgaatc	TaqMan$^{®}$ amplicon region
NM_013507	Eif4g2	870	cggtgaaggcttttcatttc	871	aggctttgtccacaatcagc	TaqMan$^{®}$ amplicon region

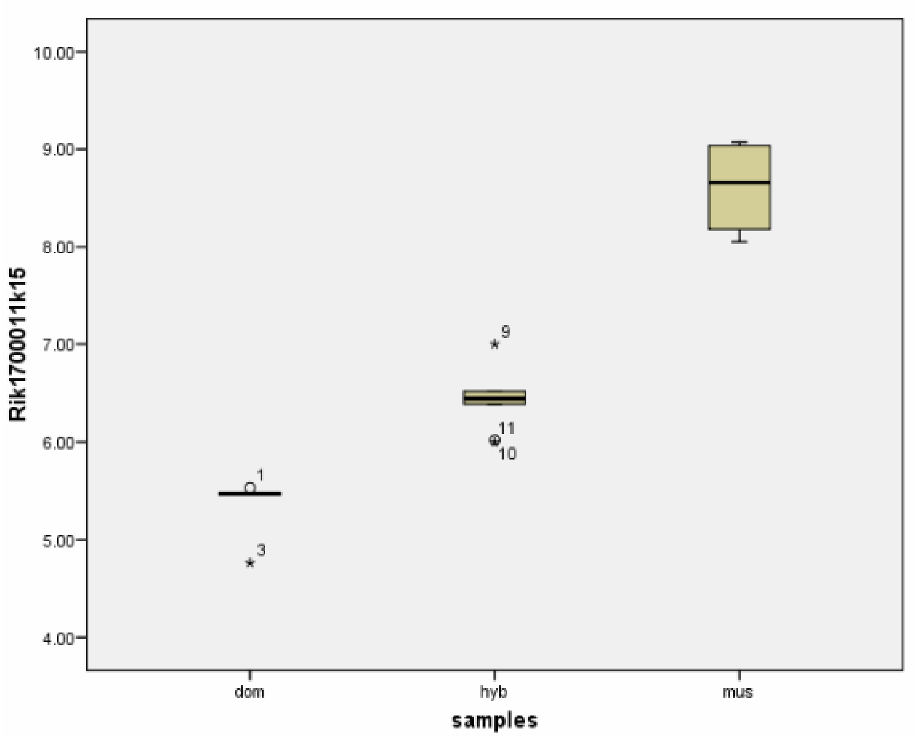
Supplement 8: Boxplots showing results of TaqMan® Gene Expression assay qRT-PCR analysis. Y-axis shows the averaged ΔCT values.



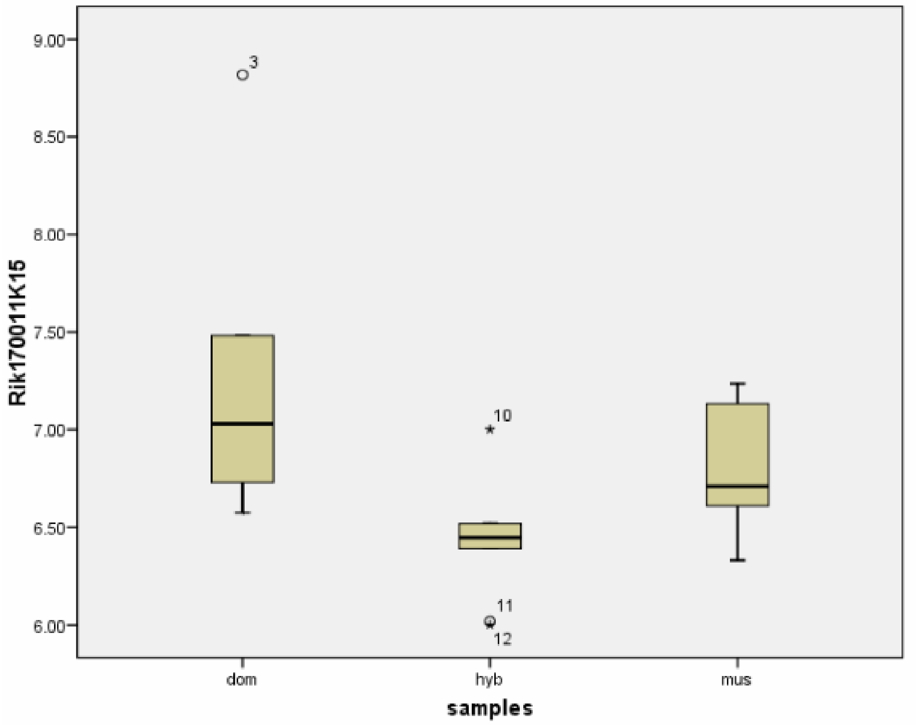
TaqMan® Gene Expression assay qRT-PCR analysis with inbred-animal samples set.



TaqMan® Gene Expression assay qRT-PCR analysis with wild-animal sample set.

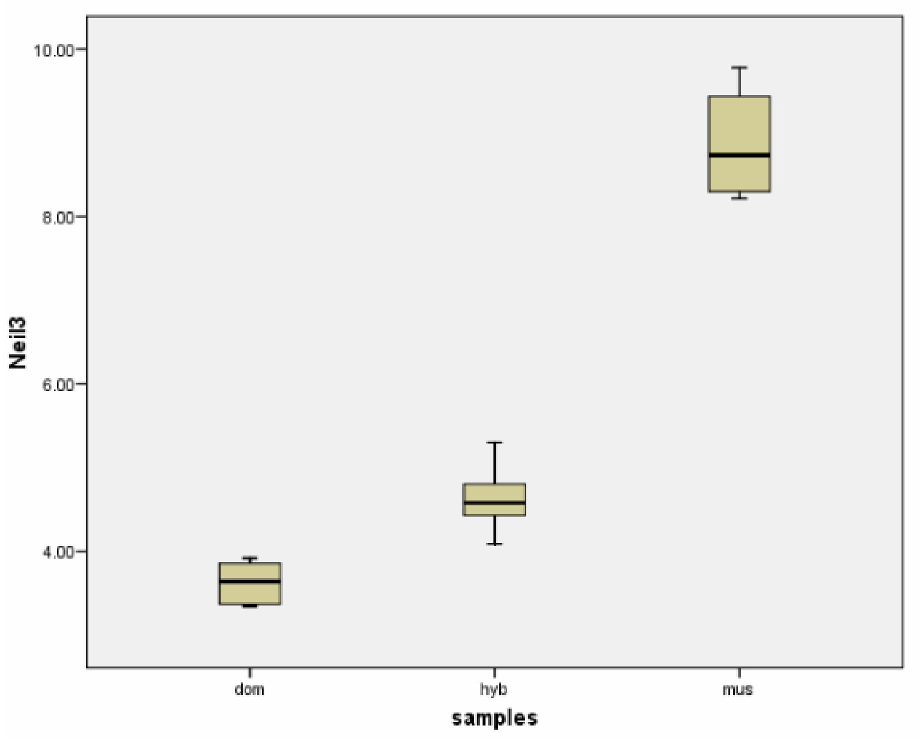


TaqMan® Gene Expression assay qRT-PCR analysis with inbred-animal samples set.

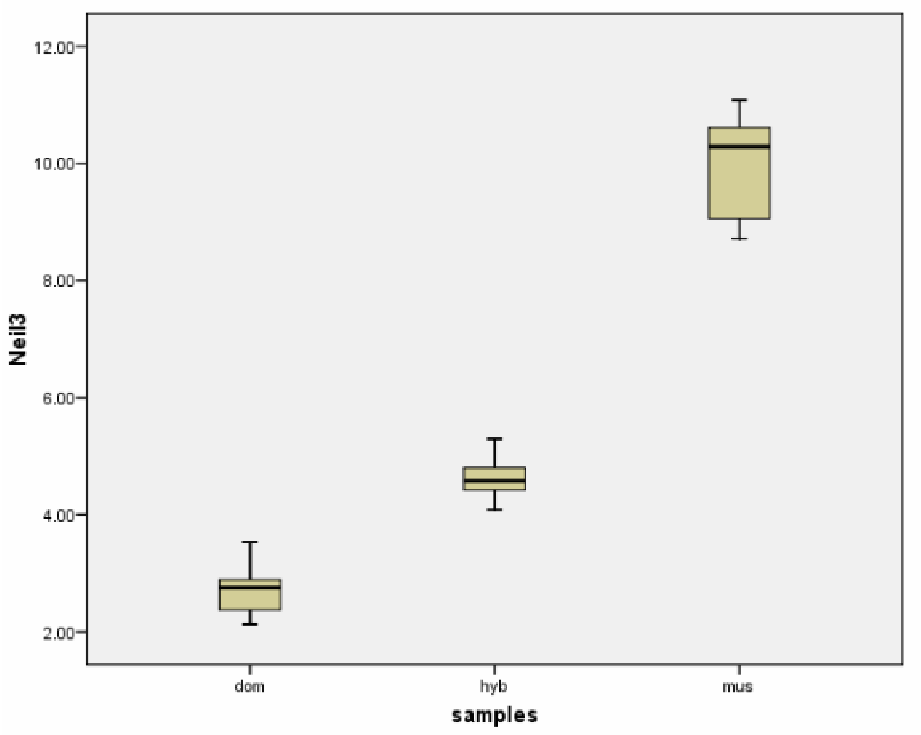


TaqMan® Gene Expression assay qRT-PCR analysis with wild-animal sample set.

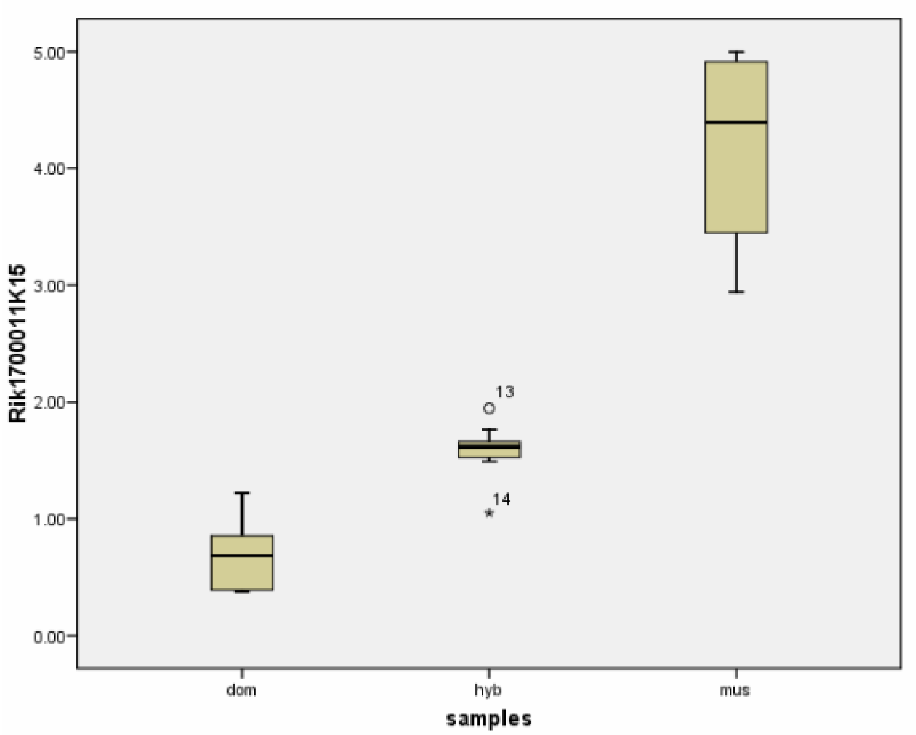
Supplement 9: Boxplots showing results of Sybr Green qRT-PCR analysis. Y-axis shows the averaged ΔC_T values.



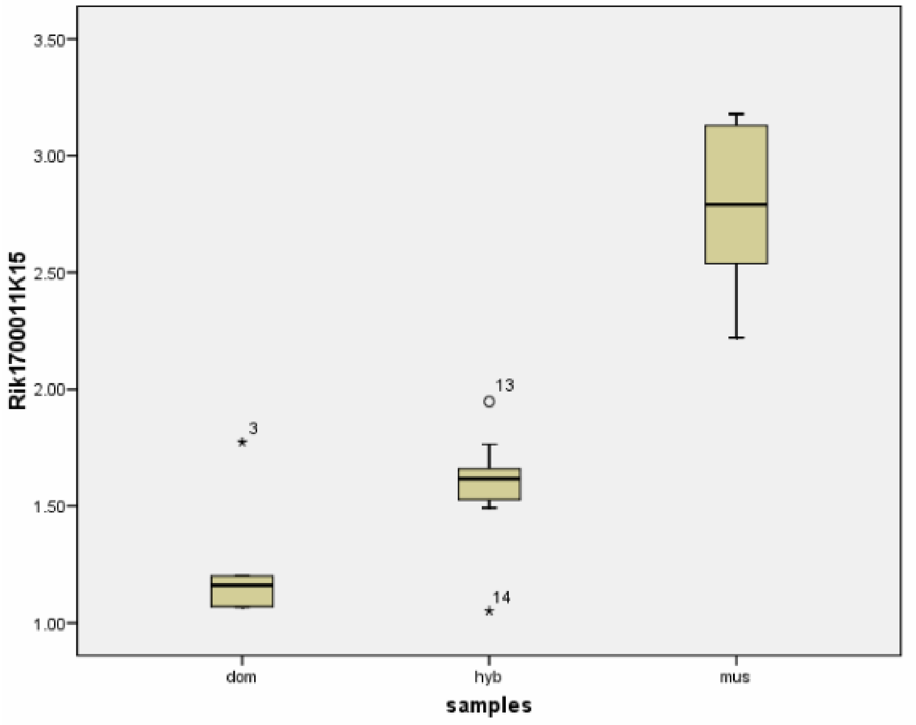
Sybr Green qRT-PCR analysis with inbred-animal samples set.



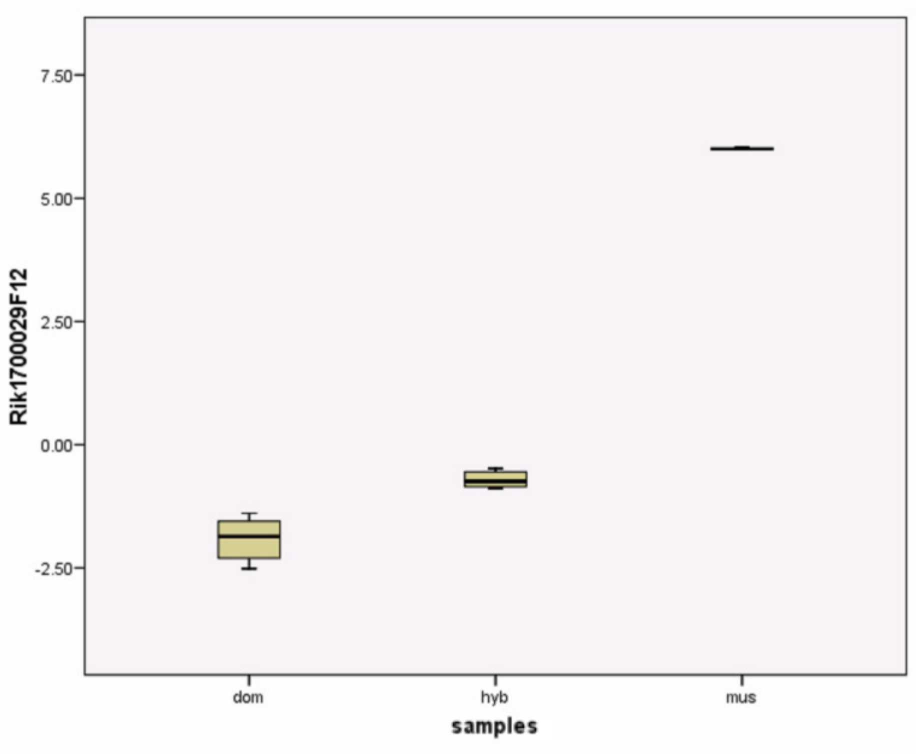
Sybr Green qRT-PCR analysis with wild-animal sample set.



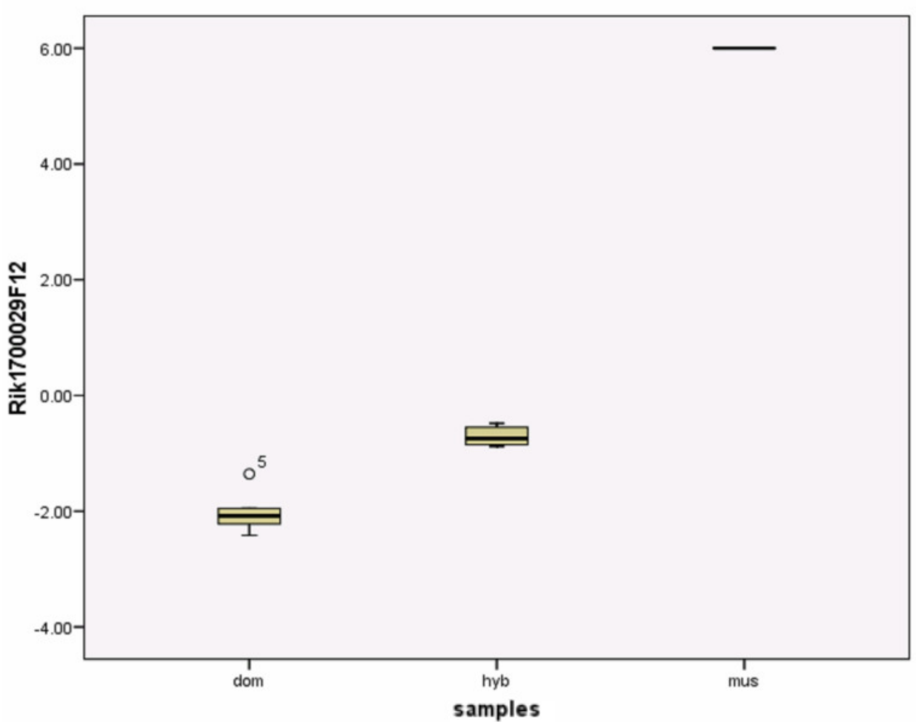
Sybr Green qRT-PCR analysis with inbred-animal samples set.



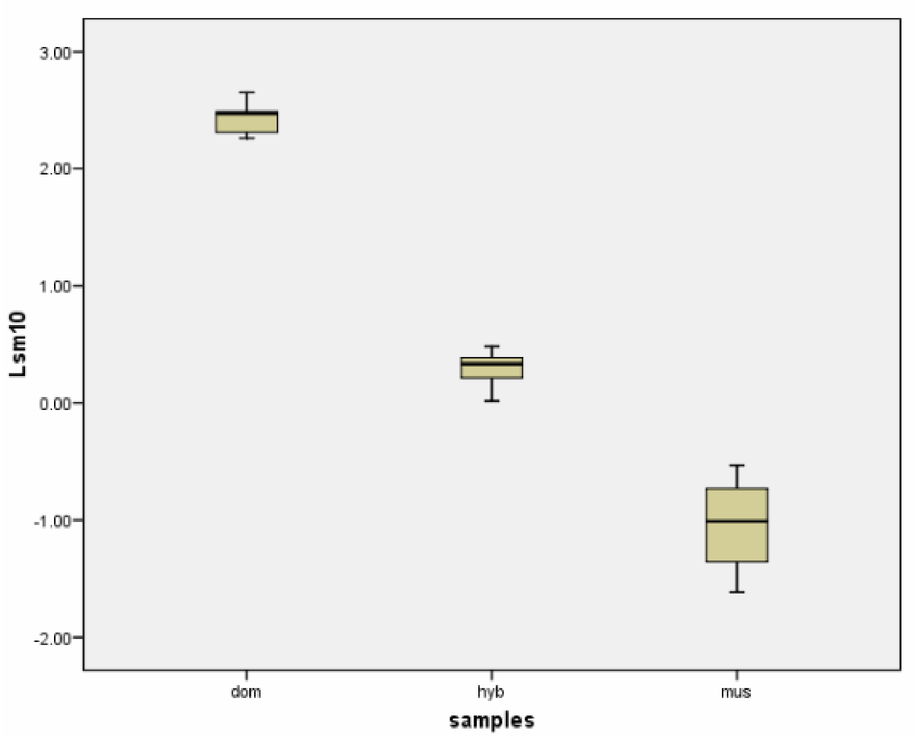
Sybr Green qRT-PCR analysis with wild-animal sample set.



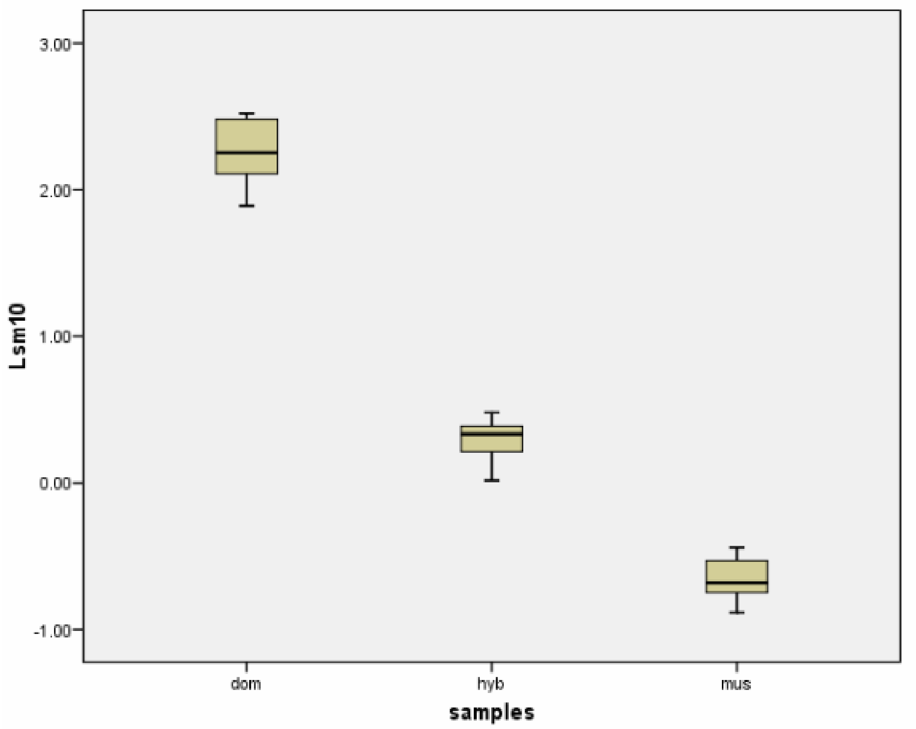
Sybr Green qRT-PCR analysis with inbred-animal samples set. *M. m. musculus* is not expressed and arbitrarily set to an averaged ΔC_T value of 6.



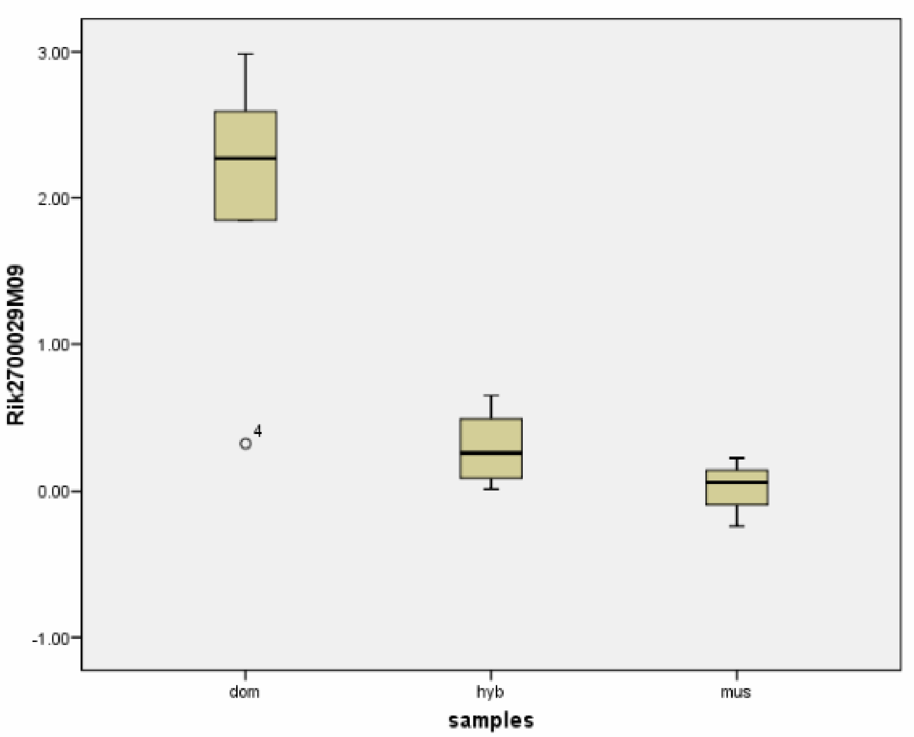
Sybr Green qRT-PCR analysis with wild-animal sample set. *M. m. musculus* is not expressed and arbitrarily set to an averaged ΔC_T value of 6.



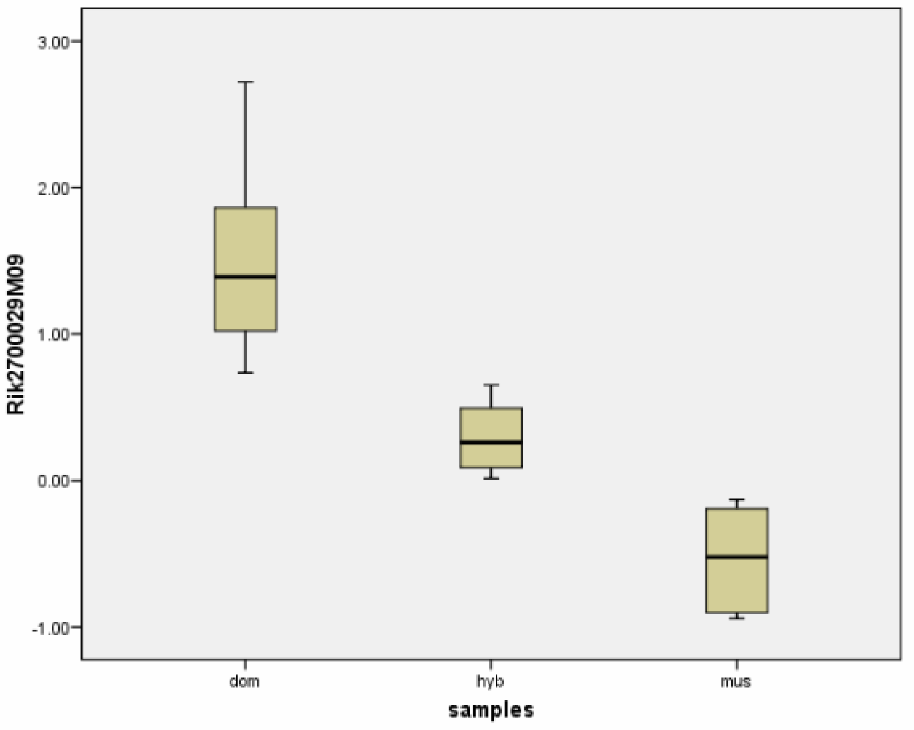
Sybr Green qRT-PCR analysis with inbred-animal samples set.



Sybr Green qRT-PCR analysis with wild-animal sample set.



Sybr Green qRT-PCR analysis with inbred-animal samples set.



Sybr Green qRT-PCR analysis with wild-animal sample set.

Supplement 10: Differences in single probes related to signal intensities. SNPs are indicated in bold in the probe sequence.

		Signal intensities					
Gene symbol	**Probes 1-11**	**20.03 (*dom*)**	**20.04 (*dom*)**	**24.02 (*mus*)**	**44.01 (*mus*)**	**SNP in**	**high overall expression (MAS 5.0)**
Neil3							
	ATGGCAGCCCTCTGTGCAA**G**ATGCA	4152	5156	3766	6134	*dom*	*dom*
	GCAAGATGCACCACCGCCGTTGTGT	2370	2809	1866	2913		
	GCCGTTGTGTTCTCCGAGTTGTGAG	1394	1844	535	696		
	GTGTTCTCCGAGTTGTGAGGAAAGA	1137	1174	116	180		
	GGAGGCAGTTTTATGCCTGTTCTCT	310	365	155	110		
	AGTTTTATGCCTGTTCTCTGCCGAG	4117	3732	3239	3168		
	TTATGCCTGTTCTCTGCCGAGAGGA	4587	5587	3359	4537		
	TTCTCTGCCGAGAGGAGCACAGTGC	397	352	158	100		
	GAGCACAGTGCGGATTTTTTGAATG	270	415	16	18		
	GCGGATTTTTTGAATGGGCAGA**C**CT	602	706	66	35	*dom*	*dom*
	GATTTTTTGAATGGGCAGA**C**CTGTC	180	131	102	72	*dom*	*dom*
Srp14							
	TGTGTGGCT**G**GATATTCTTAGATTC	9058	10012	2622	1792	*mus*	*dom*
	TATTCTTAGATTCCACCCGTAAGGT	493	905	837	1369		
	TCCAGGCTAGCTGCTTTTTTTCCTA	9710	10191	10571	11127		
	CCAGGCTAGCTGCTTTTTTTCCTAC	11272	10664	13310	12337		
	CAGGCTAGCTGCTTTTTTTCCTACC	13457	11469	14778	12439		
	TTTTTCCTACCCTATTTATGACAGT	14048	13900	15239	13052		
	AGTAACTA**C**TAA**C**TCTAGATGGTAC	13548	14752	21	8	*mus*	*dom*
	GTAACTA**C**TAA**C**TCTAGATGGTACA	13424	15826	8	4	*mus*	*dom*
	AACTA**C**TAA**C**TCTAGATGGTACAGT	14143	17308	4	4	*mus*	*dom*
	ACTA**C**TAA**C**TCTAGATGGTACAGTT	11742	14697	57	20	*mus*	*dom*
	CTA**C**TAA**C**TCTAGATGGTACAGTTA	11373	13879	17	44	*mus*	*dom*
Rhobtb3							
	AAGTTTCACCACTCAGACTGCCTTT	1848	1665	613	557		
	ATTTCATCGCCACTAACTACCTCAT	3521	2438	673	372		
	CACTAACTACCTCATATTCAGCCAA	3006	2379	490	248		
	CTACCTCATATTCAGCCAAAAGCCT	3075	2876	420	316		
	GAATTTCAAGACCTTTCAGTGGAAG	1145	1637	112	126		
	TCACCATATAGCTTCTTACTAGTTA	943	930	159	97		
	GCTTCTTACTAGTTATCTTTTAAAG	397	579	49	130		
	AAAGCTACTACAGACTAAATTTATG	82	81	11	10		
	ACTAAATTTATGTTTCA**C**CTAGAAT	5	16	3	3	*dom*	*dom*
	ATATAAAACTGTTAGACATTGGGTT	63	107	4	4		
	CCCTATTGCTTTAATAACATCAAGA	225	227	59	28		
Lsm14							
	TTTTTGATTTCCTTTGTA**C**TGTTTG	33	72	574	712	*dom*	*mus*
	TTGATTTCCTTTGTA**C**TGTTTGGAC	86	122	1266	1259	*dom*	*mus*
	GATTTCCTTTGTA**C**TGTTTGGACTA	126	210	1693	1455	*dom*	*mus*
	TTCCTTTGTA**C**TGTTTGGACTAAAG	41	56	835	834	*dom*	*mus*
	TGTA**C**TGTTTGGACTAAAGTGAAGA	37	12	296	344	*dom*	*mus*
	TATGGACTTCGTTTTGTAGCAAATC	1893	1829	2426	2123		

	GGACTTCGTTTTGTAGCAAATCACA	2299	2137	2844	2249		
	CTTCGTTTTGTAGCAAATCACAGGA	635	631	908	663		
	AACTATAGTGAAAAAGATCTGGC**C**A	9	18	29	3	*dom*	*mus*
	GTGAAAAAGATCTGGC**C**ACTTTGTG	167	193	75	39	*dom*	*mus*
	TGAAAAAGATCTGGC**C**ACTTTGTGT	153	210	33	26	*dom*	*mus*
Lsm10							
	TTTCATGAATATCCGCCTGGCCAAT	927	706	6326	4325		
	AGTTGGATGACCTTTTTGTGAC**C**GG	1075	1298	6207	7243	*dom*	*mus*
	TTGTGACCGGTCGTAACGTCCGATA	688	645	2026	1619		
	GATACGTCCATATCCCAGATGGTGT	1729	2083	12884	14792		
	TGGTGTGGACATCACTGCTACTATT	1593	1355	11226	9667		
	ACCGTGTGCGCAACTTTGGTGGCAA	851	836	8914	7424		
	GGTCAAGGTCGTCGAGAGTTCCCCT	2170	2085	14934	14119		
	AGGGACCTCCGGTGTTCTGAGCCAA	1784	1659	10134	7449		
	GTAACAGACATGTCACAGCGGCCTT	464	423	6083	4139		
	GGTGACAATCCCTCTTTTGGATCAT	2362	2794	14329	15514		
	TGGAGTGAATCACCTCTAACGTCCC	1586	1341	16359	10660		
2700029 M09Rik							
	GAAAACTAGGATACTCACTGGAACA	373	507	4028	2431		
	GCAGGCTTGGTTGTTCCAGTAGATA	1625	1671	9717	8534		
	GTTGGTTACCGAGAGCTCCCTGAAA	936	860	5285	4224		
	AGGCAGTTGT**C**GACGCTGCAAGCGA	285	487	284	240	*dom*	*mus*
	GAGACTGAAAGCATTCGCTCCCATT	383	421	2941	2914		
	TGATGACTTTTGTGCAGTTTGCTAA	2073	2370	10228	11905		
	GAATGGACCTCTTTTGCTATGGCTC	2583	2871	11535	10831		
	GCTATGGCTCTCATTATTTTCACAA	1578	1345	9997	6672		
	GTCAGCTTTTACCTCTTGCGTATAA	2337	2068	11888	9996		
	GAGAACATAGACCAGCTTGCAGGAT	909	844	5292	4487		
	AAGCTGCCCTGTTAGTGCAGTGCTT	1574	1973	5989	6936		

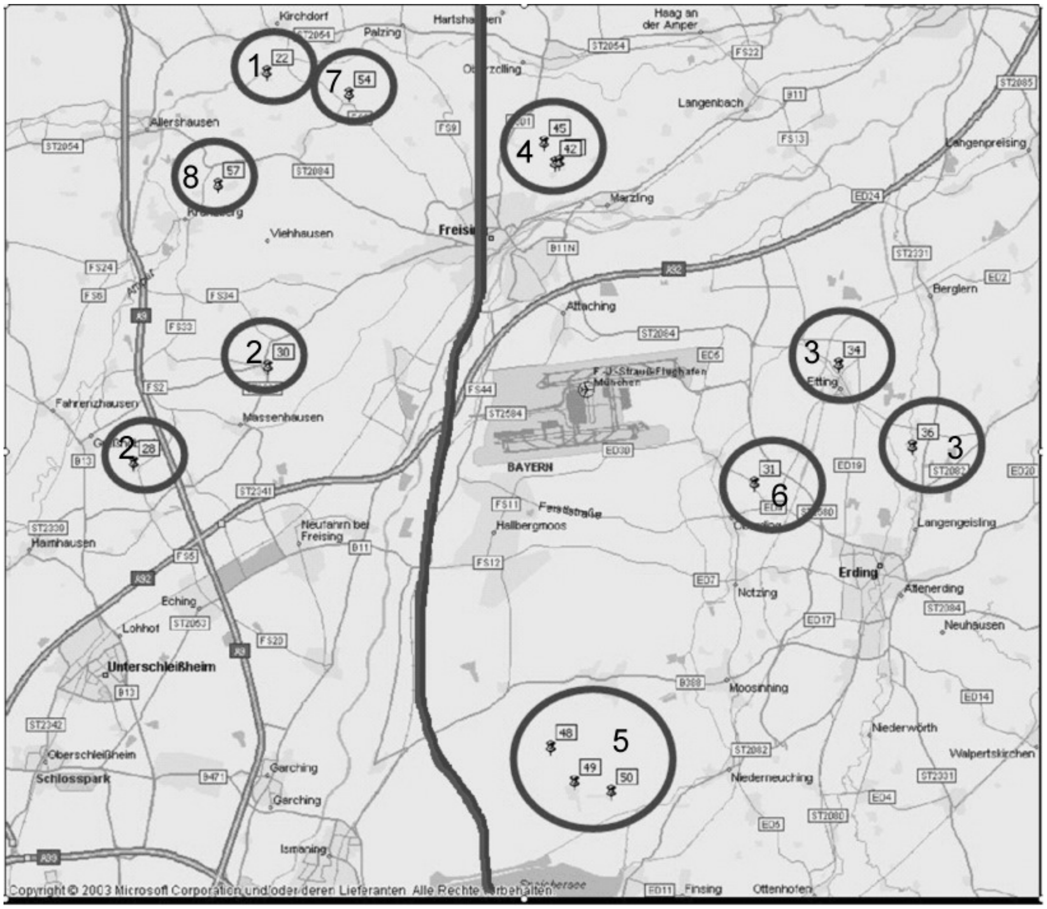
Supplement 11: Sequencing results of amplicon regions.

Gene symbol	RefSeq	Assay ID	Amplicon position	Amplicon length	Size PCR product	SNP position	SNP presumably in primer (position in amplicon)	SNP in
Scarb2	NM_007644	Mm00446978_m1	575	134	347	–	–	–
Pga5	NM_021453	Mm00480598_m1	155	59	1: 512 2: 360	– 2: 99	– ?, R-primer (60)	– *mus*
Neil3	NM_146208	Mm00467596_m1	972	75	369	248	?, F-primer (23)	*dom*
1700011 K15Rik	NM_029294	Mm00661433_s1	1171	142	396	96, 110	Yes, F-primer (5, 19)	*mus*
Srp14	NM_009273	Mm00726104_s1	576	76	394	160, 189	?, F-primer (21) probe (50)	160: polymorph in *dom* 189: polymorph in *mus*
1700029 F12Rik	NM_025585	Mm00481622_m1	368	67	232	–	–	–
Rhobtb3	NM_028493	Mm00712630_m1	865	123	244	–	–	–
Eif4g2	NM_013507	Mm00469036_m1	439	81	461	–	–	–

Supplement 12: List of samples and geographic origin of the animal material used for sequencing. Samples also used for expression analysis are additionally indicated.

Species	Country	Town	Coordinates	Samples	Different localities	Sample ID’s	Experiment
M. m. domesticus	Germany	Bonn	50°45’N - 51°N 6°45’E - 7°E	12	9	TP1a.1	microarray
						TP1a.2	
						TP3a.1	microarray
						TP3a.2	
						TP4a	
						TP5.1	microarray
						TP7	
						TP8.1	
						TP8b.2	microarray
						TP10.1	
						TP12.1	
						TP17	
M. m. domesticus	Iran	Ahvaz	31°13’N - 31° 39’N 48°37’ E - 50°E	8	8	AH2.2	
						AH3.1	
						AH4	
						AH5.1	
						AH6.2	
						AH8.3	
						AH9.2	
						AH21.2	
M. m. musculus	Czech Republic	Námest and Oslavou	49° N – 50° 30’ N 12° E- 16° E	8	8	CR02	
						CR03	
						CR15	
						CR16	
						CR28	
						CR57	
						CR61	
						CR63	
M. m. musculus	Kazakhstan	Almati	43°N 77°E	8	8	Al30	
						Al32	
						Al36	
						Al41	
						Al47	
						Al49	
						Al64	
						Al72	
M. m. musculus	Austria	Vienna	48° N 21’ - 48° 30 N 16° 26’E - 16° 71’	6	3	W1a.1	microarray
						W.3.1	microarray
						W3.3	
						W4a.1	microarray
						W4b.1	
						W6a.1	microarray
M. caroli	Thailand	Khorat	?	1	1	KTH	qRT-PCR
Laboratory F1 hybrids				4		14.01	microarray
						14.03	microarray
						29.02	microarray
						29.03	microarray

Supplement 13: Overview of sampling area and trapping localities for the hybrid zone analysis. The position of the center of the hybrid zone was placed according to Sage (1986).



Supplement 14: Primer combination and sequences for Lsm10 and Neil3.

Name	DNA	Exon	F-Primer	Sequence	R-Primer	Sequence
Lsm10	genomic DNA	Coding region	1781	gtgtgctcctgcccacag	1763	gggccacattctagaccaag
	genomic DNA	4800 bp upstream	1784	tgatgctccatactgtcacagac	1785	gcacgtgtagaaggcagagg
	genomic DNA	4200 bp upstream	1764	ttatcctctgccttctacacg	1765	tcactgcccggttcttactc
	genomic DNA	3600 bp upstream	1786	caatagttgctcaaagactgctg	1787	ataggtctgcgtccctgtcc
	genomic DNA	2600 bp upstream	1768	gcagcaagcacaaaaatgag	1769	catgtaccactactgcccaac
Neil3	cDNA	1	1684	cgggttctgtgactcctttc	1653	ggctctaggtcttgggaacc
	cDNA	1, 2, 3, 4	1663	gccagggtgtacactgaatg	1669	aggcaacaccctctgatcc
	cDNA	1, 2, 3, 4	1665	gctgctccaatgaatgctaag	1669	aggcaacaccctctgatcc

	cDNA	4, 5, 6, 7	1670	ggtggaaagccaacagagag	1671	caagcatcacaggccttagc
	cDNA	4, 5, 6, 7	1672	tcagaaattcggtggaaagc	1673	caaaccacacaggaccactg
	cDNA	7, 8, 9	1706	cagtggtcctgtgtggtttg	1707	ctgccatcacttttggaatg
	cDNA	7, 8, 9	1670	ggtggaaagccaacagagag	1709	caatttccattcctggtcca
	cDNA	8, 9, 10	1678	ttggccagaaagaaaagagc	1666	attagcatcctggaacaatttc
	cDNA	8, 9, 10	1678	ttggccagaaagaaaagagc	1664	ggaacaatttccattcctggtc
	cDNA	10	1654	tggaccaggaatggaaattg	1655	tgcgatgtaaacataatctcctg
	genomic DNA	2	1715	atgaacctggtcgaggagtc	1716	caacaacagacgtgacactgg
	genomic DNA	3	1717	taggtcttctgggcgaagtg	1718	ttgctacaattcccacatcc
	genomic DNA	4	1721	tgggaattgtagcaatgagc	1722	agattcccatttcagcatgg
	genomic DNA	7	1727	agtttgacagcaccgagtagg	1728	ggctggcaatttctctaaacc
	genomic DNA	8	1731	caaatctatttcaagtagtcagcatac	1732	taaccatggcttgaaaaacc
	genomic DNA	9	1733	atgggcagagtggtaagagc	1734	attattggccgcacttgttc
	genomic DNA	10	1737	tcctctgctcagcactgttc	1738	tgcgatgtaaacataatctcctg
	genomic DNA	2600 bp upstream	1640	cagtggcccagatataggaaag	1641	atacccccagaagtgtcagg
	genomic DNA	3600 bp upstream	1642	atcagtgtcccaggatcagg	1643	tttccatcatagttcaatgaaacc

Supplement 15: Primer used for expression-phenotype/genotype correlative analysis. SNP position is counted from first base of F-primer, number in brackets refers to position in the fasta file of the digital supplement.

Name	Fragment size	Position SNP	F-primer	Sequence	R-primer	Sequence
Lsm10	1142	516 (225), 769 (478), 781, (490)	1780	gttccaggccacacttgttt	1763	gggccacattctagaccaag
Neil3	693	353 (224), 550 (421)	1644	tttagtgcctgggaggaatg	1645	ccaaaggactggacagtgatg

Supplement 16: Samples used for hybrid zone analysis. An overall hybrid index of 0 refers to a pure *domesticus* genotype, a hybrid index of 1 to a pure *musculus* genotype.

ID hybrid	Locality	Overall hybrid index	ID father	Overall hybrid index	ID mother	Overall hybrid index	Comments
Bay1c.1	1	0.3214	Bay22.3	0.3750	Bay22.1	0.2143	
Bay1c.2	1	0.2857	Bay22.3	0.3750	Bay22.1	0.2143	
Bay2.1	2	0.0714	Bay30.1	0.0000	Bay28.1	0.1071	
Bay2.2	2	0.0357	Bay30.1	0.0000	Bay28.1	0.1071	
Bay3a.1	3	0.7857	Bay36.3	1.0000	Bay34.1	0.9286	
Bay3b.1	3	0.9643	Bay36.3	1.0000	Bay34.1	0.9286	
Bay4a.1	3	0.9286	Bay36.2	0.8929	Bay36.1	0.9643	
Bay4b.1	3	0.8929	Bay36.2	0.8929	Bay36.1	0.9643	
Bay6.1	4	0.8333	Bay42.2	0.8571	Bay42.4	0.9231	
Bay6.2	4	0.8077	Bay42.2	0.8571	Bay42.4	0.9231	
Bay7.1	4	0.9286	Bay42.6	0.8214	Bay42.5	0.8214	
Bay7.2	4	no data	Bay42.6	0.8214	Bay42.5	0.8214	
Bay9.1	5	0.9286	Bay50.1	0.8929	Bay48.1	0.9286	
Bay9.2	5	0.8929	Bay50.1	0.8929	Bay48.1	0.9286	
Bay11.1	5	0.7857	Bay49.4	no data	Bay49.2	0.7857	
Bay11.2	5	0.8571	Bay49.4	no data	Bay49.2	0.7857	
Bay14.1	2	0.0000	Bay30.2	0.0000	Bay28.2	0.0357	
Bay14.2	2	0.0000	Bay30.2	0.0000	Bay28.2	0.0357	
Bay31.3	6	0.8571	Bay31.1	0.8571	–	–	brought mother pregnant from field
Bay31.4	6	0.8571	Bay31.1	0.8571	–	–	brought mother pregnant from field
Bay31.6	6	0.8571	Bay31.1	0.8571	–	–	brought mother pregnant from field

Bay41.2	6	0.8571	–	–	–	–	caught in field
Bay45.8	4	no data	Bay45.1	0.8571	–	–	brought mother pregnant from field
Bay49.3	5	0.8571	–	–	–	–	caught in field
Bay54.1	7	0.1923	–	–	–	–	caught in field
Bay54.2	7	0.5000	–	–	–	–	caught in field
Bay57.2	8	0.2143	–	–	–	–	caught in field

Supplement 17: Locus-specific genotype in relation to expression-phenotype. Deviations of genotype and expression-phenotype are indicated in bold.

Transcript	**Sample ID**	**SNP position and genotype**			**Locus-genotype**	**Signal intensity GeneChip®**	**Expression like**	**Expression phenotype follows genotype**
Lsm10		**516**	**769**	**781**				
	B1c.1	aa	tt	cc	*mus*	7836	*mus*	yes
	B1c.2	aa	tt	cc	*mus*	8286	*mus*	yes
	B2.1	cc	cc	gg	*dom*	1083	*dom*	yes
	B2.2	ca	ct	cg	*hyb*	4897	*hyb*	yes
	B3a.1	aa	tt	cc	*mus*	8978	*mus*	yes
	B3b.1	aa	tt	cc	*mus*	8340	*mus*	yes
	B4a.1	aa	tt	cc	*mus*	8487	*mus*	yes
	B4b.1	aa	tt	cc	*mus*	8588	*mus*	yes
	B6.1	aa	tt	cc	*mus*	7681	*mus*	yes
	B6.2	aa	tt	cc	*mus*	7311	*mus*	yes
	B7.1	aa	tt	cc	*mus*	8640	*mus*	yes
	B7.2	aa	tt	cc	*mus*	8126	*mus*	yes
	B9.1	aa	tt	cc	*mus*	8300	*mus*	yes
	B9.2	aa	tt	cc	*mus*	8256	*mus*	yes
	B11.1	aa	tt	cc	*mus*	8186	*mus*	yes
	B11.2	aa	tt	cc	*mus*	7245	*mus*	yes
	B14.1	ca	ct	cg	*hyb*	4418	*hyb*	yes
	B14.2	aa	tt	cc	*mus*	6811	*mus*	yes
	B31.3	aa	tt	cc	*mus*	6954	*mus*	yes
	B31.4	aa	tt	cc	*mus*	9160	*mus*	yes
	B31.6	aa	tt	cc	*mus*	8883	*mus*	yes
	B41.2	aa	tt	cc	*mus*	8669	*mus*	yes
	B49.3	aa	tt	cc	*mus*	6859	*mus*	yes
	B54.1	ca	ct	cg	*hyb*	4630	*hyb*	yes
	B54.2	aa	tt	cc	*mus*	7795	*mus*	yes
	B57.2	cc	cc	gg	*dom*	1038	*dom*	yes
	B45.8	aa	tt	cc	*mus*	7868	*mus*	yes
	TP1a.1	cc	cc	gg	*dom*	1179	*dom*	yes

	TP3a.1	cc	cc	gg	*dom*	1201	*dom*	Yes
	TP5.1	cc	cc	gg	*dom*	1239	*dom*	yes
	TP8b.2	cc	cc	gg	*dom*	1345	*dom*	yes
	W1.1	aa	tt	cc	*mus*	8367	*mus*	yes
	W3.1	aa	tt	cc	*mus*	8192	*mus*	yes
	W4a.1	aa	tt	cc	*mus*	8956	*mus*	yes
	W6a	aa	tt	cc	*mus*	8824	*mus*	yes
	14.01	ca	ct	cg	*hyb*	5349	*hyb*	yes
	14.03	ca	ct	cg	*hyb*	4213	*hyb*	yes
	29.02	ca	ct	cg	*hyb*	4271	*hyb*	yes
	29.03	ca	ct	cg	*hyb*	4765	*hyb*	yes
Neil3		**353**	**550**					
	B1c.1	tt	cc		*dom*	740	*dom*	yes
	B1c.2	tt	cc		*dom*	641	*dom*	yes
	B2.1	tt	cc		*dom*	767	*dom*	yes
	B2.2	tt	cc		*dom*	1117	*dom*	yes
	B3a.1	ct	ct		*hyb*	545	*hyb*	yes
	B3b.1	tt	cc		*mus*	106	*mus*	yes
	B4a.1	tt	cc		*mus*	90	*mus*	yes
	B4b.1	ct	ct		*hyb*	457	*hyb*	yes
	B6.1	tt	cc		*dom*	837	*dom*	yes
	B6.2	tt	cc		*dom*	765	*dom*	yes
	B7.1	ct	ct		*hyb*	421	*hyb*	yes
	B7.2	ct	ct		*hyb*	475	*hyb*	yes
	B9.1	tt	cc		*mus*	104	*mus*	yes
	B9.2	tt	cc		mus	67	*mus*	yes
	B11.1	tt	cc		*dom*	**453**	*hyb*	**no**
	B11.2	tt	cc		*dom*	**463**	*hyb*	**no**
	B14.1	ct	ct		*hyb*	**603**	*dom*	**no**
	B14.2	ct	ct		*hyb*	515	*hyb*	no
	B31.3	tt	cc		*dom*	800	*dom*	yes
	B31.4	tt	cc		*dom*	**329**	*hyb*	**no**
	B31.6	ct	ct		*hyb*	423	*hyb*	yes
	B41.2	tt	cc		*dom*	1136	*dom*	yes
	B49.3	ct	ct		*hyb*	353	*hyb*	yes
	B54.1	tt	cc		*dom*	670	*dom*	yes
	B54.2	ct	ct		*hyb*	447	*hyb*	yes
	B57.2	tt	cc		*dom*	822	*dom*	yes
	B45.8	tt	cc		*dom*	**474**	*hyb*	**no**
	TP1a.1	tt	cc		*dom*	904	*dom*	yes
	TP3a.1	tt	cc		*dom*	609	*dom*	yes
	TP5.1	tt	cc		*dom*	699	*dom*	yes
	TP8b.2	tt	cc		*dom*	741	*dom*	yes
	W1.1	cc	tt		*mus*	36	*mus*	yes
	W3.1	cc	tt		*mus*	91	*mus*	yes
	W4a	cc	tt		*mus*	76	*mus*	yes
	W6a	cc	tt		*mus*	91	*mus*	yes
	14.01	ct	ct		*hyb*	425	*hyb*	yes
	14.03	ct	ct		*hyb*	457	*hyb*	yes
	29.02	ct	ct		*hyb*	381	*hyb*	yes
	29.03	ct	ct		*hyb*	426	*hyb*	yes

Supplement 18: Nucleotide polymorphism in the 5' upstream flanking region of Lsm10 in *M. m. domesticus* and *M. m. musculus* populations. Values are given for three fragments separately and fragments concatenated, respectively.

Fragment	**1**	**2**	**3**	**Concatenated fragments**
Relative distance to locus (kb)	4.8	4.2	2.6	
Number of sites	441	1136	555	2132
M. m. domesticus				
Samples from Iran				
Number of chromosomes	14	12	14	8
Number of segregating sites	1	0	1	2
π	0.00082	0	0.00089	0.00039
θ_W per site	0.00071	0	0.00057	0.00036
Tajima’s D	0.3244	0	1.21219	0.24178
Significance	>0.10	–	>0.10	>0.10
Samples from Germany				
Number of chromosomes	20	20	22	20
Number of segregating sites	0	2	0	2
π	0	0.00059	0	0.00031
θ_W per site	0	0.0005	0	0.00026
Tajima’s D	0	0.43538	0	0.43538
Significance	–	>0.10	–	>0.10
M. m. musculus				
Samples from Kazakhstan				
Number of chromosomes	14	16	14	12
Number of segregating sites	4	5	0	7
π	0.00267	0.00098	0	0.00099
θ_W per site	0.00285	0.00133	0	0.00109
Tajima’s D	-0.21471	-0.8533	0	-0.33218
Significance	>0.10	>0.10	–	>0.10
Samples from Czech Republic				
Number of chromosomes	14	16	14	14
Number of segregating sites	3	7	1	5

π	0.00177	0.00114	0.00079	0.00064
θ_W per site	0.00214	0.00186	0.00057	0.00074
Tajima’s D	-0.52939	-1.36612	0.84228	-0.46161
Significance	>0.10	>0.10	>0.10	>0.10
Samples from Vienna	12	12	10	10
Number of chromosomes				
Number of segregating sites	0	0	0	0
π	0	0	0	0
θ_W per site	0	0	0	0
Tajima’s D	0	0	0	0
Significance	–	–	–	–

Supplement 19: Nucleotide polymorphism in the 5' upstream flanking region of Neil3 in *M. m. domesticus* and *M. m. musculus* populations. Values are given for two fragments separately and fragments concatenated, respectively.

Fragment	**1**	**2**	**Concatenated fragments**
Relative distance to locus (kb)	3.6	2.6	
Number of sites	340	538	878
M. m. domesticus			
Samples from Iran			
Number of alleles	14	12	12
Number of segregating sites	6	3	9
π	0.00517	0.00177	0.00299
θ_W per site	0.00555	0.00185	0.00313
Tajima’s D	-0.24444	-0.12836	-0.1791
Significance	>0.10	>0.10	>0.10
Samples from Germany			
Number of alleles	24	18	18
Number of segregating sites	2	2	4
π	0.00148	0.00118	0.00145
θ_W per site	0.00158	0.00108	0.00133
Tajima’s D	-0.1312	0.22041	0.27089

Significance	>0.10	>0.10	>0.10
M. m. musculus			
Samples from Kazakhstan			
Number of alleles	16	14	14
Number of segregating sites	7	8	11
π	0.00811	0.00521	0.00607
θ_W per site	0.0062	0.00468	0.00537
Tajima's D	1.08364	0.42958	0.52986
Significance	>0.10	>0.10	>0.10
Samples from Czech Republic			
Number of alleles	16	12	12
Number of segregating sites	8	5	14
π	0.00917	0.00399	0.00588
θ_W per site	0.00709	0.00321	0.00503
Tajima's D	1.05533	0.9015	0.71386
Significance	>0.10	>0.10	>0.10
Samples from Vienna			
Number of alleles	12	8	8
Number of segregating sites	0	0	0
π	0	0	0
θ_W per site	0	0	0
Tajima's D	0	0	0
Significance	–	–	–

7 Digital Supplement

- Supplement 1: Sequence data of Affymetrix target regions.
- Supplement 2: Sequence data of TaqMan® amplicon regions.
- Supplement 3: Sequence data of coding region.
- Supplement 4: Sequence data of non-coding regions.
- Supplement 5: Diagnostic SNP for genotype/expression-phenotype association.
- Supplement 6: qRT-PCR data.

Erklärung

Ich versichere, dass ich die von mir vorgelegte Dissertation selbständig angefertigt, die benutzten Quellen und Hilfsmittel vollständig angegeben und die Stellen der Arbeit - einschließlich Tabellen, Karten und Abbildungen -, die anderen Werken im Wortlaut oder dem Sinn nach entnommen sind, in jedem Einzelfall als Entlehnung kenntlich gemacht habe; dass diese Dissertation noch keiner anderen Fakultät oder Universität zur Prüfung vorgelegen hat; dass sie - abgesehen von unten angegebenen Teilpublikationen - noch nicht veröffentlicht worden ist sowie, dass ich eine solche Veröffentlichung vor Abschluss des Promotionsverfahrens nicht vornehmen werde. Die Bestimmungen dieser Promotionsordnung sind mir bekannt. Die von mir vorgelegte Dissertation ist von Frau Dr. Bettina Harr betreut worden.

Ruth Rottscheidt

Köln, den 18.12.2007

Teilpublikationen:

Die folgende Publikation basiert auf Teilen dieser Arbeit und entspricht Kapitel 2:

Rottscheidt, R. and B. Harr (2007): Extensive additivity of gene expression differentiates subspecies of the house mouse. Genetics 177: 1553-1567.

Lebenslauf

Name:	Ruth Rottscheidt
Anschrift:	Kaiser-Karl-Ring 13 53111 Bonn
Geburtsdaten:	23.09.1975 in Bad Kreuznach
Staatsangehörigkeit:	deutsch
Schulausbildung:	
1982 – 1986	Gemeinschaftsgrundschule Swisttal-Odendorf
1986 – 1995	Privates St. Joseph Gymnasium, Rheinbach
1995	Abitur
Hochschulausbildung:	
10/1995 – 01/2002	Studium der Biologie Rheinische Friedrich-Wilhelms-Universität, Bonn
28. Januar 2002	Abschluss des Diploms in Biologie Forschungsmuseum Alexander Koenig Abteilung für Herpetologie Prof. Dr. Wolfgang Böhme Rheinische Friedrich-Wilhelms-Universität, Bonn
11/2002 – 04/2004	Angestellte in der medizinisch-wissenschaftlichen Abteilung bei Rodisma Med-Pharma GmbH, Köln
05/2004 – 02/2008	Promotion am Institut für Genetik Lehrstuhl für Evolutionsgenetik Prof. Dr. Diethard Tautz Universität zu Köln
19. Februar 2008	Voraussichtlicher Abschluss der Promotion

Köln, 18.12.2007

Ruth Rottscheidt