# Detecting genes involved in selective sweeps within populations of the house mouse species complex - a multi-locus candidate gene approach -

**Inaugural – Dissertation**

zur

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

vorgelegt von

**Sonja Ihle**

aus
Augsburg

Köln, 2003

| | |
|---|---|
| **Berichterstatter:** | **Prof. Dr. Diethard Tautz** |
| | **Prof. Dr. Thomas Wiehe** |
| **Tag der mündlichen Prüfung:** | **13.02.2004** |

**List of Contents** **Page**

**Acknowledgements** ..... **1**

**Zusammenfassung** ..... **3**

**Abstract** ..... **5**

**Declaration** ..... **7**

**Chapter I** ..... **9**
Mouse populations with indistinguishable D-loop haplotypes can be differentiated by their unique proportion of shared microsatellite alleles even in recently founded populations

**Chapter II** ..... **33**
Screening multiple microsatellite loci identifies selective sweep genes by a hitchhiking approach

**Chapter III** ..... **69**
The antimicrobial ß-Defensin 6 gene was subjected to a selective sweep event caused by differences in the regulatory regions.

**Literature** ..... **99**

**Erklärung** ..... **108**

**Lebenslauf** ..... **109**

# Danksagung

Ich danke Professor Dr. Diethard Tautz, dass er mir die Möglichkeit gegeben hat, in seiner Arbeitsgruppe dieses Projekt durchzuführen. Seine innovativen Ideen und Impulse, seine immerwährende Diskussionsbereitschaft und Unterstützung haben den Erfolg dieser Arbeit maßgeblich mitbestimmt. Sein Vertrauen in das Gelingen des Projekts und auch in meine Arbeit haben mir über viele Schwierigkeiten hinweggeholfen.
Ich danke Professor Dr. Thomas Wiehe für die hilfreichen Gespräche zur Datenauswertung und für die Übernahme des zweiten Gutachtens. Vielen Dank an Professor Dr. Hartmut Arndt für die Übernahme des Prüfungsvorsitzes und Dr. Röbbe Wünschier als Beisitzer in meinem Prüfungskommittee.

Vielen Dank an Susi Krächter für die gemeinsame Planung, Organisation und Durchführung unserer Mausfang-Exkursionen, die den Grundstein für diese Arbeit gelegt haben. Die gemeinsamen Erlebnisse in den entlegensten „Maushabitaten“ dieser Welt werden uns ewig in Erinnerung bleiben!
Mein besonderer Dank gilt den vielen „Mäusefängern“ vor Ort, die uns die „Mauskontakte“ vermittelt haben: Barbara Dod und Louis Rouquet in Frankreich; Uli Schliewen und Solo Robespierre in Kamerun, Paul Pfander in Kasachstan, Jaroslav Pialek und seine Familie, so wie Barbora Bimova und Eva Bozikova in Tschechien.
Weiterhin danke ich Francois Bonhomme, Pierre Boursot und Annie Orth, die uns mit Mausproben aus der Wildmauszucht des “Conservatoire genetique de souris sauvage” in Montpellier versorgt haben, sowie Pavel Munclinger, der uns die ersten Mausproben aus Tschechien geschickt hat. Bettina Harr steuerte Mausproben aus den USA bei.
Weiterhin danke ich Bettina Harr für ihre Unterstützung bei der Datenauswertung. Durch ihre Ideen, Ratschläge und ihre ständige Bereitschaft, meine Fragen zu beantworten, hat sie sehr viel zum Gelingen dieser Arbeit beigetragen.
Vielen Dank an Patricy de Andrade Sales für ihre Unterstützung bei der DNA-Isolierung und Genotypisierung, sowie Susi Krächter für ihre Hilfe bei der „Highthroughput"-Genotypisierung und –Sequenzierung.
Vielen Dank an Iary Ravaoarimanana, die mir Daten des „random microsatellite screen“ zur Verfügung stellte.
Ich danke Chris Voolstra für seine unermüdliche Bereitschaft, meinen Computer am Laufen zu halten; ebenso Alex Pozhitkov und Till Bayer, die einsprangen, wenn weiterhin Not am

Computerfachmann war. Alex Pozhitkov war stets eine wichtige Hilfe in statistischen Fragen. Röbbe Wünschier danke ich für den "Schnellkurs" in Linux und ebenso für seine Unterstützung in Computerfragen.
Martin Gajewski danke ich für die Bereitstellung der Taq Polymerase, mit der ein Großteil der Proben typisiert wurde, sowie für seine Hilfe in labortechnischen Fragen. Ebenso danke ich Wim Damen, der immer ein offenes Ohr für meine Fragen hatte und mich in die Light Cycler Technologie einwies.
Arne Nolte danke ich für seine stete Diskussionsbereitschaft und die vielen Gespräche über populationsgenetische und sonstige ungeklärte Rätsel dieser Welt. Mit seinen Ideen und seiner Motivation hat er über manches Problem hinweggeholfen.
Kathrin Schwertz und Viktoria Rivkin danke ich für ihre unermüdliche Hilfe in der Küche. Zu Highthroughput-Zeiten müssen sie unser Labor und die nicht enden wollenden leeren Spitzenkästen gehaßt haben.
Eva Siegmund und Gerti Meyer zu Altenschildesche danke ich für ihre Geduld, die sie ständig für unsere nicht ein- und ausgetragenen Rechnungen aufbringen und ihre Hilfe bei Bestellungen und vielen anderen organisatorischen Fragen.
Vielen Dank an Heidi Fußwinkel, die viel für das Mausprojekt organisiert hat, und immer für alle Probleme da war.
Weiterhin danke ich allen Mitgliedern der Abteilung Evolutionsgenetik für die doch meistens gute Stimmung und den täglichen Spaß, den wir miteinander bei der Arbeit haben. Allen voran natürlich meinen Kollegen bzw. Kolleginnen des Pop-Labs!

Meinen Eltern danke ich für ihre wohlwollende Unterstützung in der ganzen Zeit. Neben der mentalen Unterstützung stellten sie mit immer ein Auto zum Mäusefangen zur Verfügung. Meinem Vater danke ich vor allem auch für die unermüdliche Hilfe bei der Korrektur der Arbeit.

Mein besonderer Dank gilt der Volkswagenstiftung, die das „Selective Sweep Projekt“ in verschiedenen Populationen der Hausmaus finanziell unterstützt.

# Zusammenfassung

Die Suche und Identifizierung von Genen, die an Selektions- und Anpassungsprozessen beteiligt sind, haben sich zu einem Hauptforschungsschwerpunkt in der Evolutionsbiologie entwickelt, um die molekularen Mechanismen zu verstehen, die der Evolution, Anpassung und auch der Artbildung zu Grunde liegen. Ziel dieser Studie war es, "Selective Sweeps" in der Hausmaus zu finden. "Selective Sweeps" finden an Gen-Orten statt, die durch eine neu erworbene Mutation einen Selektionsvorteil erworben haben und dadurch andere Allele am gleichen Gen-Ort aus dem Gen-Pool verdrängen. Als Untersuchungsobjekt wurde die Hausmaus gewählt, da einerseits die komplette genomische Sequenz zur Verfügung steht und sie ein wichtiger Modellorganismus in der biomedizinischen Forschung ist und andererseits die Ökologie und Phylogenie freilebender Populationen gut untersucht sind. Zwei *Mus musculus* und vier *Mus domesticus* Populationen wurden nach einem Beprobungssystem gefangen, das vermeidet, nah verwandte Tiere aus der selben Familiengruppe zu fangen.

Die phylogenetische Analyse zeigte, dass die *Mus musculus* Populationen aus Kasachstan und Tschechien ziemlich alte Populationen darstellen, während die *Mus domesticus* Populationen aus Frankreich, Deutschland, Kamerun und USA einen relativ jungen Ursprung haben. Daher scheint Westeuropa entgegen anderer Annahmen wesentlich später von der Hausmaus besiedelt worden zu sein. Zudem hat die Ausbreitung der Art durch den Menschen die Besiedlungsgeschichte der Hausmaus stark beeinflußt. Die Sequenzierung des mitochondriellen D-loops ermöglichte die Unterscheidung zwischen den beiden Mauslinien, aber der Marker konnte nicht zwischen den Populationen innerhalb einer Linie trennen. Dahingegen konnten die Populationen durch Typisierung von 81 Microsatelliten Loci gut unterschieden werden und jedes Individuum wurde genau seiner Herkunftspopulation zugewiesen.

Während ihrer Besiedlungsgeschichte wurde die Maus vermutlich mit verschiedenen Selektionsdrücken konfrontiert. Durch Anwendung der lnRV und lnRH Statistik auf die typisierten Mikrosatelliten Loci wurden 9 potentielle „Selective Sweep“-Loci identifiziert. In Abhängigkeit von der Mutationsrate der Mikrosatelliten wurden „Selective Sweeps“ auf unterschiedlichen taxonomischen Ebenen entweder in einzelnen Populationen oder in den einzelnen Mauslinien detektiert. Die Mikrosatelliten waren so gewählt, dass sie direkt an Gene gekoppelt sind, die eine Rolle bei Umweltinteraktionen spielen. Die beobachteten "Sweep Loci" waren an Gene gekoppelt, die an Immunantworten oder an olfaktorischer Sinneswahrnehmung beteiligt sind oder differentiell exprimiert werden.

In einem zweiten Ansatz wurden 106 zufällig im Genom verteilte Microsatelliten-Loci typisiert und auch hier wiesen einige Loci reduzierte Variabilität auf, die auf ein potentielles „Selective Sweep“ Ereignis hinweisen.

Allerdings bedarf es stets zusätzlicher Methoden, um „Selective Sweep“-Ereignisse zu bestätigen. Daher konzentrierten sich die weiteren Analysen auf das Immungen ß-Defensin 6, das möglicherweise einen „Selective Sweep“ in *Mus domesticus* verursacht hat. Das Ergebnis wurde durch reduzierte Nucleotid-Diversität in einem „Selective Sweep“-Fenster von 22 kb bestätigt. ß-Defensin 6 ist ein kationisches Peptid und spielt eine wichtige Rolle bei der Immunantwort auf bakterielle Infektionen. Die vorteilhafte Mutation, die den „Selective Sweep“ verursacht hat, trat vermutlich in den regulatorischen Bereichen auf und veränderte die Induktion des Gens und vermutlich damit auch die Spezifität der Immunantwort. Quantitative PCR Experimente zeigten, dass der Expressionslevel von ß-Defensin 6 in verschiedenen Organen von *Mus domesticus* Tieren etwas niedriger ist als im Vergleich zu *Mus musculus* Tieren. Dennoch können erst klare Aussagen über mögliche Veränderungen in der Genexpression gemacht werden, wenn die Expression unter kontrollierten Laborbedingungen mit gezüchteten Labortieren nochmal untersucht wird.

Diese Arbeit zeigte, dass ein genomweiter Mikrosatelliten-Screen ein sinnvoller Ansatz ist, Gene zu identifizieren, die in den verschieden Hausmaus-Population unterschiedlich selektiert wurden. Andere Methoden wie Sequenzierungen oder quantitative PCR Experimente müssen jedoch angewendet werden, um die funktionalen Unterschiede zu finden, die den „Selective Sweep“ verursacht haben.

# Abstract

The search for genes involved in selection and adaptation has become a main research subject in evolutionary biology to understand the molecular mechanisms of evolution, adaptation and speciation. The aim of this study was to identify selective sweep events in the house mouse species complex. Selective sweeps occur at loci that have acquired a new advantageous mutation and due to a selective advantage the new mutation sweeps through the gene-pool of a population replacing other alleles at this locus.

The house mouse was chosen as a study object because the full genomic sequence for this model species of biomedical research is available and also the ecology and phylogenetic history of natural populations are fairly well known. Two *Mus musculus* and four *Mus domesticus* populations were caught under a stringent sampling regime avoiding the sampling of closely related individuals. Phylogenetic analysis of these populations showed that the two *Mus musculus* populations from Kazakhstan and the Czech Republic were rather old while the *Mus domesticus* populations from Germany, France, Cameroon and USA seemed to be of relatively recent origin claiming that the house mouse colonization of Western Europe is younger than previously assumed and that human mediated long distance transport strongly influenced mouse colonization patterns. Mitochondrial D-Loop sequencing distinguished between the two mouse lineages but was not suitable to differentiate populations within the lineages. In contrast to this, genotyping of 81 microsatellite loci separated nicely the different populations and proved that each individual was assigned to its population of origin.

During colonization times the mice were most likely confronted with all kinds of different selection pressures. By applying the lnRV and lnRH statistics to the microsatellite screen nine potential sweep loci were identified. Depending on the mutation rate of the microsatellites selective sweeps were observed at different divergence levels either in single populations or in whole lineages. The microsatellite loci were chosen to be located in close phyical linkage to gene which are known to be involved in environmental interactions. The genes linked to the observed sweep loci play a role in immune response, in olfactory recognition, or they are known to be differentially expressed.

In another approach 106 randomly chosen microsatellite loci were genotyped and several loci exhibited reduced variability which is characteristic for a potential selective sweep. However, additional methods have to be applied to verify selective sweep events.

Therefore, within this study the further analysis focused on the immune response gene “ß-Defensin 6” that showed a selective sweep in the *Mus domesticus* lineage. The result was

confirmed by reduced nucleotide diversity within a selective sweep window of 22 kb around the locus. The ß-Defensin 6 gene is a cationic peptide expressed in several organs and plays an important role in microbial defense. The advantageous mutation that caused the sweep occurred probably in the regulatory regions of the gene, altering the induction pathways and potentially shaped the specifity of the immune response to infections in *Mus domesticus*. Quantitative PCR experiments confirmed that slightly lower expression of ß-Defensin 6 are found in wild caught *Mus domesticus* animals than in *Mus musculus* animals. However, inferences about differences in the regulation of gene expression between the lineages have to be tested under more controlled conditions with animals raised in the laboratory.

Nevertheless, this study proved that a genome wide microsatellite locus screen is a valuable tool to identify genes which were shaped by selection in different populations of the house mouse species complex. Additional methods such as sequencing the genomic regions and quantitative PCR experiments are necessary to reveal functional differences that have caused the selective sweep event.

## Declaration

The data of the “random microsatellite screen” used in Chapter II were generated and analysed by Iary Ravaoarimanana and kindly offered to be included in this analysis.

# Mouse populations with indistinguishable D-loop haplotypes can be differentiated by their unique proportion of shared microsatellite alleles even in recently founded populations

## Introduction

In spite of the ongoing taxonomic debate about the house mouse species complex (Auffray et al. 1990, Sage et al. 1993, Din et al. 1996, Ferris et al. 1983) there is general agreement that the house mouse has diverged within the genus *Mus* about one million years ago. A further radiation of the complex 0.5 MYA resulted in the today recognized different taxa: *Mus domesticus, Mus musculus, Mus castaneus, Mus molossinus* and the not clearly defined *Mus bactrianus. Mus molossinus* is of hybrid origin between *Mus castaneus* and *Mus musculus* (Boursot et al. 1993). The status of the taxa is still unresolved; depending on the authors taxa are either classified as species or subspecies and new lineages are continuously described, e.g. *Mus gentilulus* by Prager et al. (1998) and Duplantier et al. (2002). To avoid the problematic application of species definitions I refer to the different lineages as different taxa.

Figure 1 shows the molecular phylogeny of the genus *Mus* with *Apodemus* and *Rattus* as outgroups according to Boursot et al. (1993). The relationships of the sister taxa towards the house mouse are not clearly resolved and they are assumed to be either *Mus spretus*, *Mus macedonicus* or *Mus spicilegus*. The first two species occur in sympatry with *Mus domesticus* while *Mus spicilegus* lives sympatrically with *Mus musculus*. All these species are still known to hybridise occasionally with the house mouse (Boursot et al. 1993, Guenet & Bonhomme 2003).

A unique feature of these sister taxa is that they retained their aboriginal life style while all house mouse taxa have evolved commensal lineages (Sage et al. 1993). The time of the evolution of the commensalistic life style is still open. Ferris et al. (1983) suggest that mice were associated with the hominid lineage already in preagricultural times. Berry et al. (1992) claim the first real house mouse fossil from human settlements to be 80.000 years old supporting the more broadly accepted view of multiple origins of commensalism within the different lineages as the separation of the lineages predates the evolution of the human associated life style (Boursot 1993, Sage et al. 1993).

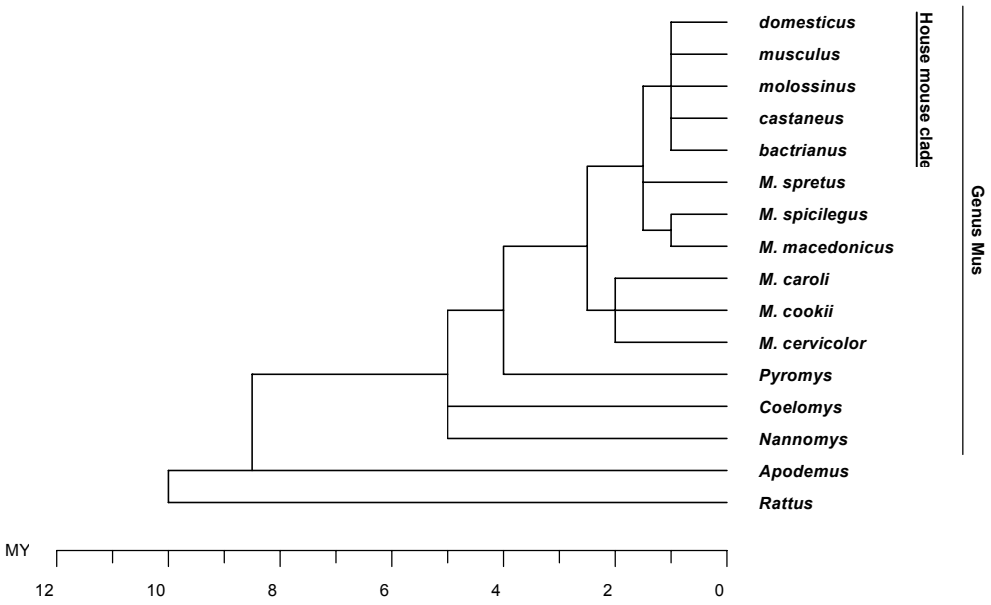



**Figure 1:** Molecular Phylogeny of the Genus *Mus* based on scnDNA/DNA hybridisations (taken from Boursot 1993)

Comparisons of the abundance of fossils at sites predating and post-dating human settlements suggest that the association with humans has led to an increase in density of house mice and therefore the commensalism represents a selective advantage (Hesse 1979, reviewed in Sage 1981).

Due to competition in areas of sympatry with other rodents house mice occur nowadays mainly within human settlements. Only in areas where other mouse species are absent the house mouse is able to establish feral populations. Only *Mus castaneus* seems to live exclusively commensal (Boursot et al. 1993, Sage 1981, Auffray et al.1988).

The colonization history of the house mouse to their present range all over the world is strongly connected to human migrations and can clearly be distinguished from the distribution pattern of the aboriginal species (Sage 1981). The strongest indication for the connections of mouse and human migration is the colonization of the New World continents such as North and South America and Australia as well as tropical Africa and many Atlantic and Pacific islands where mice were absent before the human invasions. As especially the Europeans have extensively roamed since the fifteenth century the western European house mouse *Mus*

*domesticus* is the main species that has colonized these continents via human-mediated long-distance transport (Boursot et al. 1993, Sage 1981).

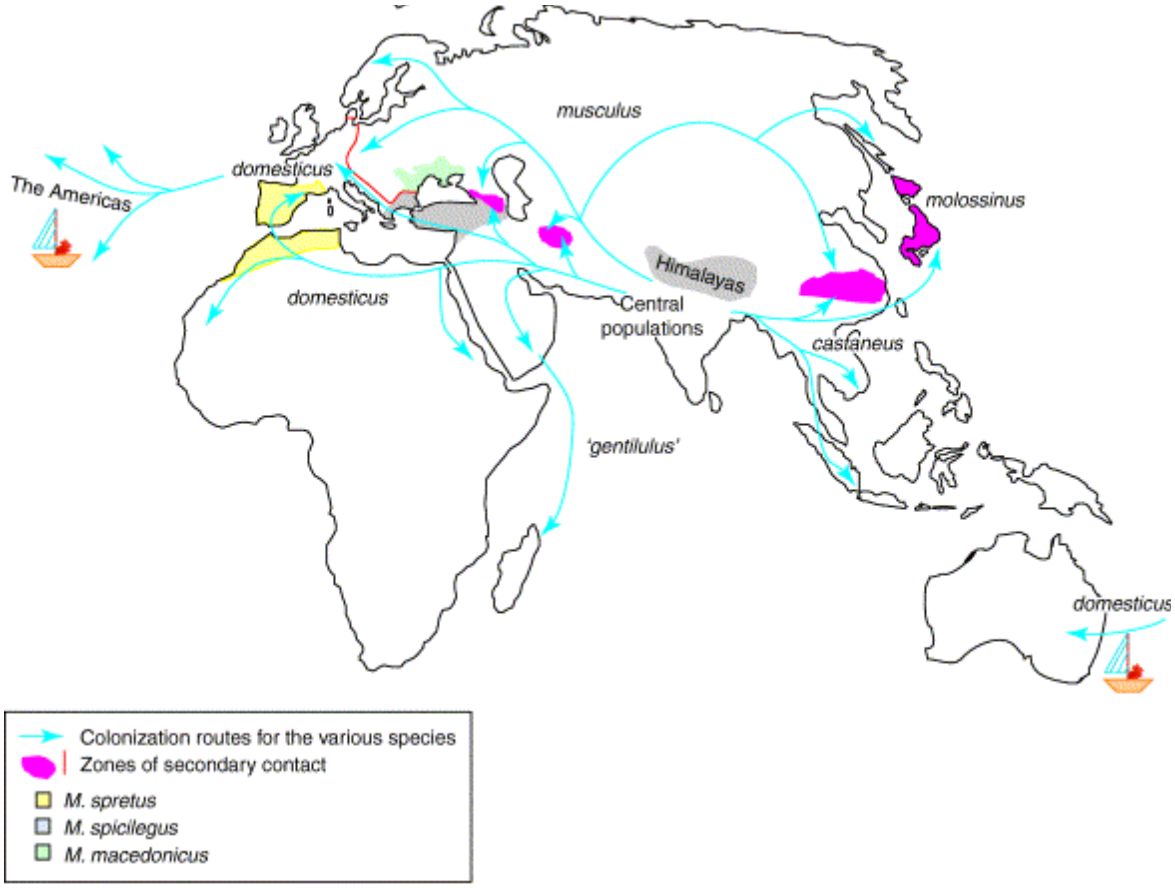



**Figure 2:** Colonization pathways of the house mouse and closely related species (Guenet & Bonhomme 2003).

Figure 2 illustrates the colonization pathways of the house mouse lineages and the current distribution of the commensal and aboriginal species. While the youngest migration pathways can easily be dated according to human colonization events during the last five hundred years the older colonization pathways are very difficult to date. Estimations of the divergence times of the main mitochondrial DNA lineages suggest that the initial radiation of the house mouse occurred 0.9 MYA which is consistent with the absence of house mouse species in the fossil record before these times (Boursot et al. 1993). The central populations in northern India, Pakistan and Iran present very polymorphic populations that cannot be grouped in any of the existing taxa. They represent the ancient lineages and it is assumed that the ancestors of the house mouse complex have evolved in this region. Boursot et al. (1996) suggest their origin to be in northern India, while Prager et al. (1998) suggest a more westward one. From there the species spread over Southeast Asia and evolved into *Mus castaneus.* While northern Eurasia and eastern Europe were penetrated by the *Mus musculus* lineage, the Middle East, the Mediterranean countries and western Europe were colonized by the *Mus domesticus* lineage.

The dates of the colonization of the western Mediterranean countries are highly debatable and almost no data exist for the Asian colonization. Concerning *Mus domesticus* the fossil record suggests its presence in Israel approximately 80.000 years ago (Berry & Bronson 1992). However, other literature cites house mouse fossils in the middle East not older than 12,000 years BP only (Bonhomme 1993). The area around the Mediterranean Sea and north-west Europe was not inhabited by mice 4000 to 2800 years ago, fossil evidence even claim younger colonization (Cucchi, oral presentation at ECM4, Brno). Concerning colonization of central Europe almost no details are available except that it was colonized after the last glaciation by a south-western invasion of *Mus domesticus* and by a north-eastern invasion of *Mus musculus* encountering a hybrid zone running in a bent line from northern Denmark to the Caspian Sea (Hunt & Selander 1973, Bonhomme et al. 1994).

As the distribution of mice is strongly associated with human mediated transport the colonization history of the house mouse species complex was continuously reshaped by several invasion events and ongoing long distance transports both complicating the resolution of the mouse migration history.

Since the house mouse has colonized nearly all places of the world, it was obviously able to adapt to many different environments which subsequently led to differentiation and speciation.

So far many studies were conducted to determine the origin of the house mouse species complex and the subsequent differentiation into the different lineages by analysing nuclear (Din et al. 1996, Munclinger et al. 2002) and mitochondrial markers (Boursot et al. 1996, Yonekawa et al. 1994, Ferris et al. 1983, Prager et al. 1996, Prager et al. 1998, Duplantier et al. 2002). As nuclear marker diagnostic allozymes were used to distinguish between the taxa and to identify the width and the amount of introgression across hybrid zones (Munclinger et al. 2002). On the basis of mitochondrial markers it was shown that the today recognized taxa represent monophyletic lineages. However, the resolution of mitochondrial markers is too weak to differentiate between populations within one lineage.

The weak resolution of the mitochondrial DNA might either be due to the fact that the successive migration events during mice colonization resulted in a constant admixture of lineages within the taxa or that the history of mouse colonization is too young to be traced by

the divergent evolution rate of mitochondrial markers and no lineage sorting has occurred so far. If the latter situation is true faster evolving microsatellite markers should be able to trace the recent population history and divergence (Bowcock et al. 1994, Goldstein et al. 1995a, Richard et al. 2001, Schlötterer 2001).

Hence, six wild caught populations of *Mus musculus* and *Mus domesticus* were genotyped for 81 microsatellite loci to find out whether the populations within the lineages can be distinguished by their microsatellite alleles, whether they form distinct homogeneous groups and which factors may have caused the divergence. Among these populations also two young populations from Cameroon and the United States were analysed which are known to be younger than 500 years. They are suitable entities to identify the level of variability of young populations and the time they required for their differentiation. In addition to the wild caught populations four captive bred populations were added to the analysis including one *Mus musculus* population, one *Mus domesticus* population, one *Mus castaneus* population and one lineage from Iran which remained unassigned to any of the today recognized taxa.

Mice are known to live in small family groups of four to twelve individuals with home ranges rarely exceeding a radius of 2 km. For strictly indoor living mice home ranges can even be restricted to a few square meters (Berry & Bronson 1992, Sage 1981). Additionally, the longevity of such demes is very short and leads to high rates of extinction and recolonization events. Hauffe et al. (2000) calculated extinction rates of 56% of a deme per year and found very low migration rates within their study populations. While their study populations consisted of populations with different Robertsonian fusion Singleton et al. (1983) found similar results for mouse populations with acrocentric chromosomes and they estimated the life span of their mouse demes to range between two and seven month. This indicates that mouse demes generally have a low life span independent of the chromosomal races.

In order to have a representative sample of mouse populations within a given area and to avoid the influence of short term specific family effects I sampled individuals from different demes which are not related to each other.

## Materials and Methods

### Samples

Mouse populations from Germany, France, Cameroon and Kazakhstan were sampled according to the following sampling scheme: in each country mice were caught at about 50 different locations; trapped mice were assigned to different location only if the trapping sites were at least 300 m apart from each other; mouse traps were put up in private houses, barns or stables. Depending on the shape of the landscape, the infrastructure of the area and the availability of suitable trapping sites the sizes of the trapping areas differ between the various countries. Mice from the USA (provided by Bettina Harr) were sampled under the same sampling scheme. Mice from the Czech Republic (provided by Pavel Munclinger) were sampled under a different sampling regime and the sampling represent mice from only a few sites, but the sites themselves are distributed over nearly the whole Czech Republic. Additionally, animals from four captive breeds of wild derived strains from Spain, Iran, Georgia and India were provided by Francois Bonhomme and Pierre Boursot from the wild mice breeding centre "Conservatoire genetique de souris sauvage" in Montpellier.
A detailed list of the characteristics of the trapping sites from the different countries is listed in Table 1.

**Table 1:** list of populations, their geographic origin and size of the sampling areas

| Population | Town | Coordinates | Area Size | Samples (total) |
|---|---|---|---|---|
| **Germany** | around Bonn and Cologne | 50°45'N - 51°N<br>6°45'E - 7°E | 40 x 70 | 59 |
| **France** | Severac-le-Chateau, Massif Central | 44°15'N - 44°30'N<br>2°45'E - 3°15'E | 20 x 20 km | 88 |
| **Cameroon** | Kumba | 4°15'N - 4°30'N<br>9°15'E - 9°45'E | 15 x 20 km | 72 |
| **Kazakhstan** | Almaty | 43°N<br>77°E | 30 x 45 km | 73 |
| **Czech Republic** | | 49°N - 50°30'N<br>12°E - 16°E | 400 km x 200 km (three main sampling patches) | 50 |
| **USA** | Chicago | | 15 x 30 km | 14 |
| **Spain** | Barcelona | 41°50'N<br>2°E | ? | 12 |
| **Iran** | Birdjand | 33°N59E | ? | 12 |
| **Georgia** | Alazani | 42°N46°E | ? | 12 |
| **India** | Masinagudi | ? | ? | 13 |

Mice were caught in death traps. For DNA isolation mice were dissected: the spleen was directly dissolved in 5 ml of SDS based Hom-buffer (80 mM EDTA (pH 8), 0.5% SDS, 100 mM Tris (pH 8)) without proteinase K. Liver, kidneys, heart, lung and testis were stored in 100% Ethanol as tissue backup.

**Isolation of DNA and preparation for genotyping**

DNA was isolated in 15 ml falcon tubes by standard salt extraction procedures. Only DNA from the Cameroon samples were isolated by using the extraction kit NucleoSpin® Blood XL (Macherey - Nagel)**.** Dried DNA pellets were resolved in 500 to 1000 µl of 1 x TE. The DNA was stored at -20°C. For the genotyping procedures the DNA was diluted to 5 ng per µl and stored in 96 well-plates for high throughput processing.

**D-Loop sequencing**

With D-Loop Primers from Prager et al. (1993) with the following sequences:

Forward-Primer: 5′-CAT TAC TCT GGT CTT GTA AAC C
Reverse-Primer: 5′-GCC AGG ACC AAA CCT TTG TGT

50 µl PCR reaction were performed by the following standard PCR conditions: final concentration of 0.5 ng/µl DNA-template; 0.06 µmol/µl dNTP; 0.045 µmol/µl $MgCl_2$; 0.1 µl/µl of 10 x PCR-buffer; 0.4 pmol/µl of forward and reverse primer and 0.04 units/µl of EuroBioTaq polymerase or 0.01 µl/µl of selfmade Taq polymerase provided by Martin Gajewski.

For the amplification the DNA template was denatured for 5’ at 94°C followed by 35 cycles of 45’’at 94°C denaturation, 45’’ at 59°C annealing, 1’30’’ at 72°C extension and a final extension step of 5’ at 72°C.

PCR products were purified using ultra-free filters from Millipore according to the manufacturers protocol. Products were rediluted in 10 mM Tris (pH 8) and 100 ng PCR product were directly added to a 10 µl cycle sequencing reaction using 3 µl of ET-Terminator Ready-Reaction Mix (Amersham Biosciences) and 5 pmol of either the reverse or the forward primer. Cycle sequencing was performed with 35 cycles with 20’’ at 95°C, 15’’ at 58°C and 1’ at 60°C. Sephadex G-50 columns were used for clean-up of the reaction.

Sequencing reactions were performed on a MegaBace 1000 capillary sequencer (Amersham - Molecular Dynamics) by the following injection protocol: DNA-injection at 3kV for 60'' followed by a runtime of 200' at 7 kV.

Raw sequence data were base-called by the Cimarron 2.19.5 Slim Phredify base-caller provided within the MegaBace Sequencing Analysis Software Version 2.1. For sequence alignment the program Seqman of the DNAstar package from Lasergene (DNASTAR, Inc.) was used. Aligned sequences were analysed with the program Arlequin (Schneider et al. 2000) and DNASP 3.51 (Rozas & Rozas 1999).

Phylogenetic reconstruction of the samples was executed by the neighbour-joining tree algorithm implemented in the program MEGA 2.1 using the Kimura's two parameter model (Kumar et al. 2001). For outgroup comparison the D-Loop sequences of *Mus spretus* (Accession number: U47539) and *Mus spicilegus* (Accession number: U47536) were downloaded from Genbank, NCBI.

**Genotyping 10 mouse populations with 81 microsatellite loci**

Microsatellite loci were selected from the mouse genetic mapping panel as part of the database of the Whitehead Institute (Dietrich et al. 1994, Dietrich et al. 1996, Copeland et al. 1993) and from the NCBI LocusLink database (http://www.ncbi.nlm.nih.gov/LocusLink/) or they were detected within genomic sequences available on the NCBI nucleotide database (http://www.ncbi.nlm.nih.gov/). The characteristics of the latter microsatellite loci are that they are in close vicinity to coding genes.

Primer sequences for the amplification of these loci were directly taken from the databases or they were designed either by hand in the flanking regions of the repeat-units or by downloading the relevant microsatellite loci and their surrounding regions into the primer design software FPCR (http://www.biocenter.helsinki.fi/bi/bare-1_html/manual.htm). The forward-primer was labelled with a fluorescence 5'-modification such as Fam, Hex or Tet and ordered from Metabion.

Microsatellite loci of all individuals were genotyped by the already mentioned standard PCR-conditions with varying annealing temperatures. PCR-products of three different loci of the same individual which were labelled with different fluorescence dyes, diluted in a scale of

1:15, pooled and mixed with an internal ET-Rox size standard (Amersham) and run in the genotyping mode on the MegaBace 1000 capillary sequencer.

Allele calling was performed with the software Genetic Profiler 2.0 (Amersham) and the data were transferred into an Excel spreadsheet for further analysis. General gene diversity estimates were calculated using the Microsatellite toolkit (Park 2001) and the program Fstat (Goudet 1995).

Deviations from Hardy-Weinberg Expectations were evaluated by the program Arlequin (Schneider et al. 2000). To prove significant differences in the gene diversity one factor analysis of variance with the software package SPSS 10.0 for Windows was performed. In order to identify homogeneous groups the Scheffé procedure was applied which is a Post-Hoc range test that groups the pairwise comparisons into significantly different and non-different groups.

Different genetic distances such as Nei's Gst (Nei 1972), Slatkin's Rst (Slatkin 1995), $\Delta\mu^2$ (Goldstein et al. 1995b) and the proportions of shared alleles (Bowcock et al. 1994) were calculated with the program Microsat (Minch et al. 1995). Significance of the distances was evaluated by performing 100 bootstraps for each data set. Tree topology was inferred with the programs Neighbour and Consense from the Phylip 3.5c software package (Felsenstein 1993) and visualized by the program TreeViewPPC 1.6.6 (Page 1996) and MEGA 2.1. (Kumar et al. 2001).

## Results

### Mitochondrial D-loop sequence analysis

The mitochondrial D-loop was sequenced for a subset of 274 individuals. The size of the processed and aligned sequences is 977 bp including insertions and deletions. The Iranian sequences contain a 76 bp duplicated insertion. At the homologous position a duplication within two *Mus musculus* samples from the Czech Republic is found. Shorter insertions appear in *Mus domesticus* samples. Insertions and deletions are excluded from the further analysis and subsequently the results of the total alignment are based on a sequence stretch of 810 bp. In the alignment 99 polymorphic sites are detected and 25 of them are singletons. In total 78 haplotypes are found.

**Table 2:** Results of the D-loop sequencing analysis in the different populations

| | Indivi-duals | sequence length | different haplotypes | haplotype diversity | Nucleotide diversity | Shared haplo-types between populations |
|---|---|---|---|---|---|---|
| **Kazakhstan** | 45 | 884 | 19 | 0.829 | 0.00349 | |
| **USA** | 11 | 882 | 7 | 0.909 | 0.00924 | 1 shared with Cameroon |
| **India** | 9 | 880 | 1 | 0 | 0 | |
| **Germany** | 45 | 835 | 23 | 0.934 | 0.01004 | 2 shared with Cameroon |
| **Spain** | 12 | 884 | 1 | 0 | 0 | 1 shared with France |
| **Czech Republic** | 18 | 878 | 8 | 0.669 | 0.01272 | |
| **Cameroon** | 59 | 869 | 8 | 0.062 | 0.00539 | 1 shared with USA, 2 shared with Germany |
| **France** | 57 | 875 | 16 | 0.862 | 0.0023 | 1 shared with Spain |
| **Georgia** | 8 | 882 | 2 | 0.25 | 0.00028 | |
| **Iran** | 10 | 956 | 1 | 0 | 0 | |
| **Complete alignment** | 274 | 810 | 78 | 0.945 | 0.01991 | |

The populations of the *Mus domesticus* clade share only four haplotypes. The highest proportion of haplotypes can be found in the German population followed by the Czech Republic. The Czech Republic exhibits the highest nucleotide diversity because the population contains *Mus domesticus* mitochondrial lineages from the hybrid zone. Apart from the lab strain populations the Cameroon population shows the lowest haplotype diversity. The

lab strains contain one or two haplotypes. Haplotypes are used to reconstruct the phylogeny of the populations. The result is shown in figure 3.

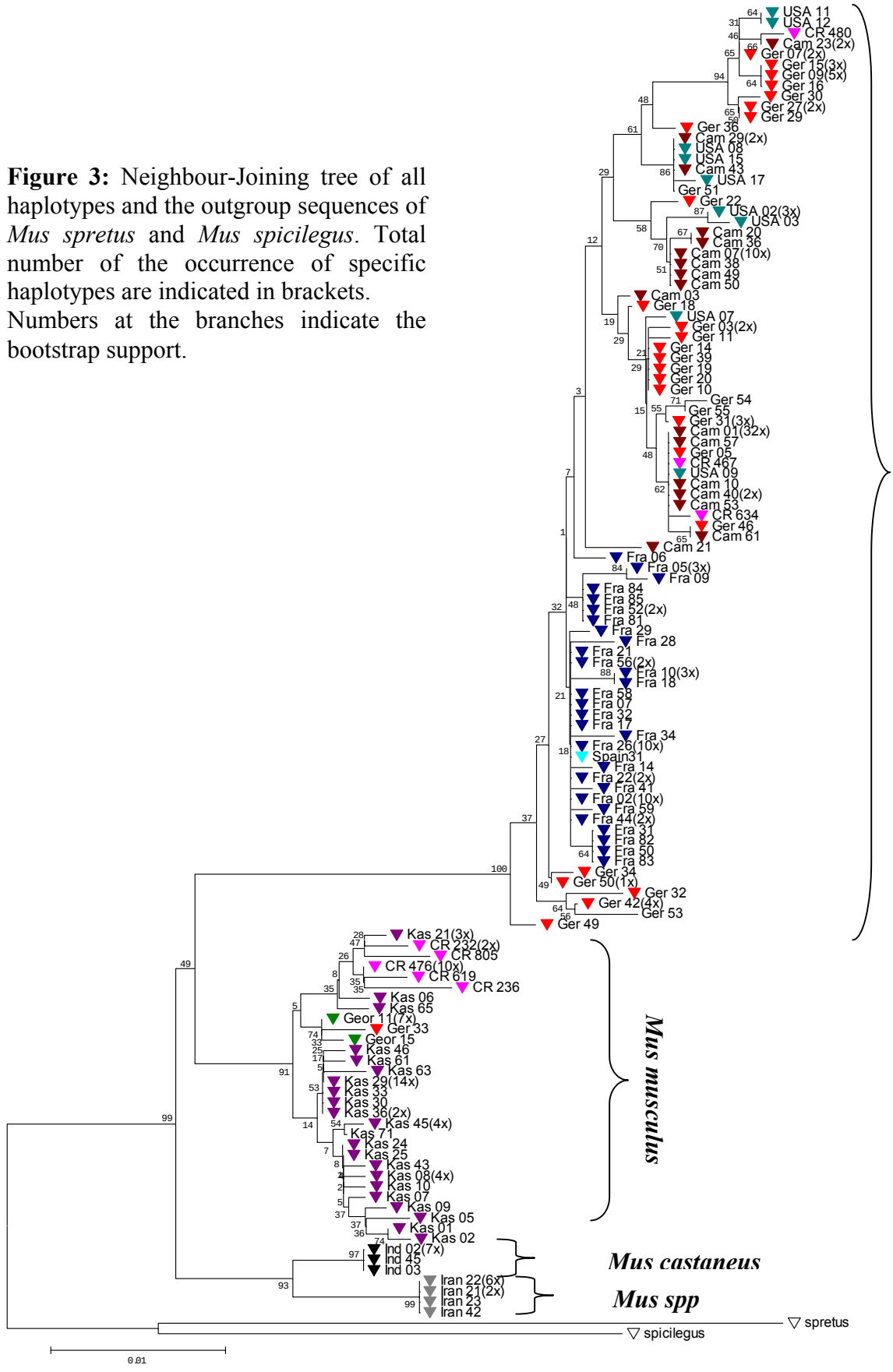



**Figure 3:** Neighbour-Joining tree of all haplotypes and the outgroup sequences of *Mus spretus* and *Mus spicilegus*. Total number of the occurrence of specific haplotypes are indicated in brackets. Numbers at the branches indicate the bootstrap support.

The Neighbour-Joining tree reveals a good separation of the different taxa. High bootstrap values of 99 support the separation of the house mouse complex from the outgroup species *Mus spretus* and *Mus spicilegus*. All lineages within the house mouse comlex are also clearly distinguishable from each other supported by high bootstrap values above 90. Only the status of *Mus musculus* as the sister group of *Mus domesticus* is poorly supported by a low bootstrap of 49.

Within the lineages very low bootstrap values are found. The *musculus* clade contains only one highly supported cluster from Kazakhstan, but within the other branch a mixture of animals from Czech Republic, Georgia and Kazakhstan exists.

Within the *domesticus* clade the resolution is even weaker. The French and Spanish animals apparently form a distinct group but the accompanying bootstrap support is low. Mice from Germany, Czech Republic, Cameroon and the USA are completely mixed and form several clusters.

Generally, the D-loop is a suitable marker to trace population differentiation between mouse lineages. For within lineage differentiations the resolution of the D-loop is weak although haplotype sharing between populations is rather low. The *Mus domesticus* lineage exhibits the deepest clades within the cluster while the branch lengths of the *Mus musculus* cluster are half as long. The differentiation between the Iranian and the Indian samples are comparable to the separation within the *Mus musculus* cluster stressing that the two lineages from Iran and India are relatively closely related. Blasting the Iranian sequence suggests a close similarity to *Mus castaneus*.

**Microsatellite loci analysis: general population characteristics and phylogenetic inferences**

In total 337 animals were genotyped for 81 loci including the same samples that were used for D-loop sequencing. General gene diversity calculations are represented in Table 3. The highest gene diversity is found in the population from Kazakhstan with an average of 12.5 alleles per locus and an expected heterozygosity of 0.69. The heterozygosity of the population from France, Germany and the Czech Republic is comparable to the population from Kazakhstan although the average number of alleles per locus is between 9 and 10 alleles. The population from Kazakhstan possesses many low frequency alleles at several loci resulting in a higher number of alleles overall which do not affect the heterozygosity. Despite of the small sample size the population from the USA has a heterozygosity of 0.62 and an average number

of alleles of 5.18 and is comparable to the Cameroon population. The Cameroon population exhibits the lowest gene diversity of all wild caught populations. Again the lab strain populations show very low diversities.

**Table 3:** General gene diversity parameters averaged over 81 microsatellite loci.

| Population | Sample size | Loci typed | Unbiased Hz | Obs Hz | No Alleles | No Alleles SD | Loci non HWE |
|---|---|---|---|---|---|---|---|
| **Cameroon** | 68 | 81 | 0.5134 | 0.3917 | 6.67 | 3.73 | 51 |
| **Czech Republic** | 42 | 79 | 0.7054 | 0.3828 | 9.08 | 4.92 | 74 |
| **France** | 64 | 81 | 0.6454 | 0.4764 | 9.98 | 6.30 | 60 |
| **Georgia** | 8 | 81 | 0.3592 | 0.3166 | 2.16 | 0.89 | 7 |
| **Germany** | 55 | 81 | 0.6524 | 0.4359 | 10.04 | 6.58 | 68 |
| **India** | 6 | 78 | 0.2910 | 0.2342 | 1.94 | 1.18 | 4 |
| **Iran** | 11 | 80 | 0.4203 | 0.3401 | 2.55 | 1.09 | 17 |
| **Kazakhstan** | 59 | 81 | 0.6994 | 0.5671 | 12.67 | 7.21 | 68 |
| **Spain** | 12 | 81 | 0.4718 | 0.4043 | 3.27 | 1.60 | 13 |
| **USA** | 12 | 79 | 0.6158 | 0.4782 | 5.23 | 2.85 | 21 |
| **combined** | 337 | 81 | 0.776 | 0.446 | 21.58 | | |

Among the wild caught population 60 to 90% of the loci show a significant deviation from Hardy-Weinberg Expectation. The lab populations show less loci deviating from HWE since a lot of loci are monomorphic and the animals were bred in the laboratory under completely different outbreeding conditions than the wild caught animals.

In the figures 4a and 4b the significant differences in heterozygosity and average number of alleles between the populations are visualized. The one factor analysis of variance reveals significant differences in average heterozygosity between the populations ($F = 21.784$; $df = 9$; $p < 0.001$) and the average number of alleles ($F = 64.665$; $df = 9$; $p < 0.001$). Horizontal lines above the bars mark the homogeneous groups with no significant differences in number of alleles or heterozygosity. Among the wild caught populations significant differences in heterozygosity are found between the populations of Cameroon on one side and Kazakhstan and Czech Republic on the other side; all other comparisons are not significantly different. The lab strains are mainly homogeneous and significantly different from all wild caught populations except for the Cameroon population which exhibits also a very low gene diversity.

Similar results are found for the number of alleles per locus: the Cameroon population has a significant lower number of alleles compared to France, Germany and Kazakhstan, but a

significant higher number of alleles compared to the lab strains. The population from the Czech Republic is significantly different from Kazakhstan.

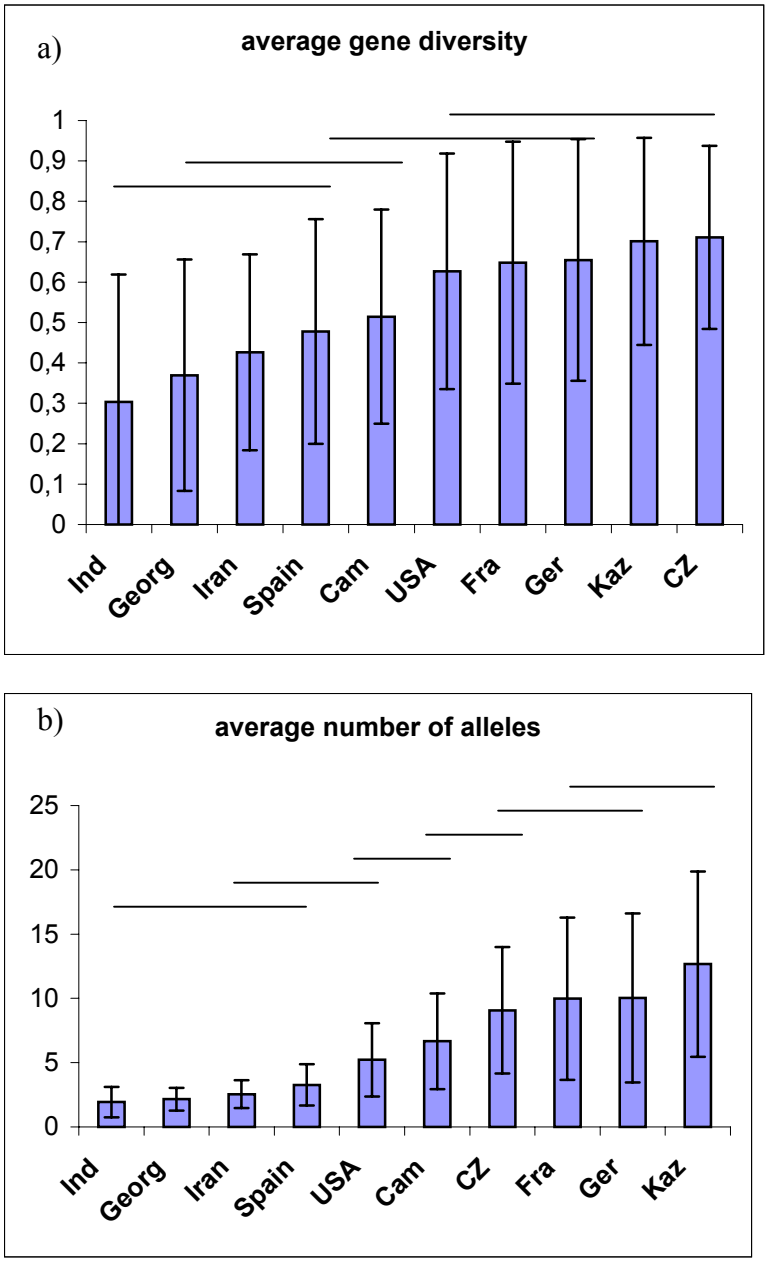



**Figure 4: a)** Average gene diversity per population; **b)** average number of alleles per population; populations which are not significantly different are indicated by horizontal bars above the columns.

Since Hardy-Weinberg Expectations were violated at most of the loci no test was performed for population differentiation such as Fst estimations. To infer whether the populations consist of separate units which are different from each other a similarity index based on the proportion of shared alleles between all pairs of individuals was calculated.

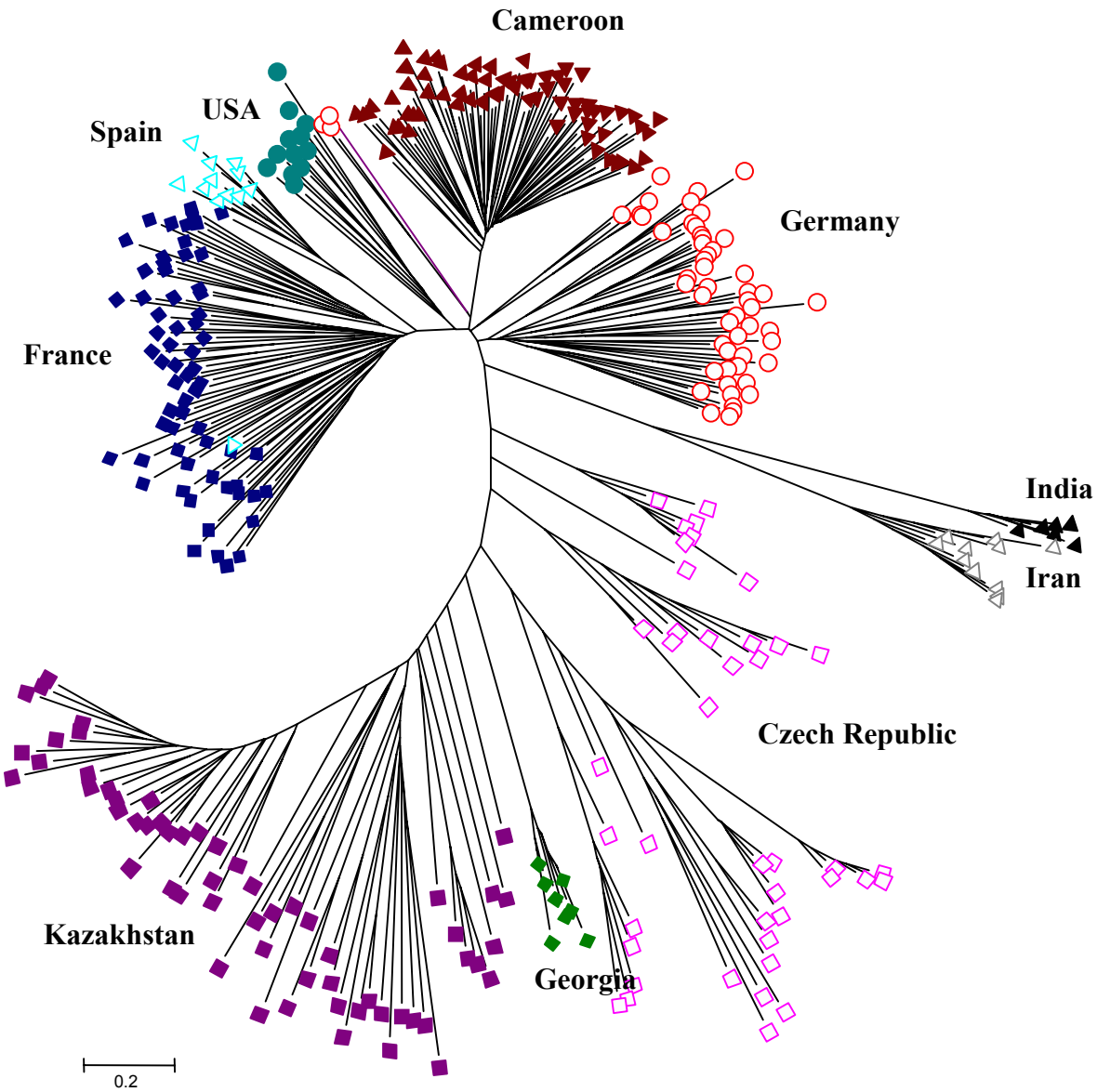



**Figure 5:** Neighbour-Joining tree based on the calculation of the proportion of shared alleles calculated for all individual comparisons

The allele sharing tree based on individuals reveals that every mouse is closer related to individuals of its population of origin, thus allowing clear distinctions between the populations. Every population can be characterised by a unique set of alleles. Only the populations from the Czech Republic and from Germany contain paraphyletic groups.

For the Czech Republic population these paraphyletic groups correspond to different sampling sites and therefore may be regarded as separate populations. For the German population the paraphyletic branches contain samples from the extreme boundaries of the sampling area, but they are still mixed with animals from the geographic centre of the area and hence cannot be regarded as independent populations.

Nevertheless, microsatellite loci are regarded as a useful tool to discriminate between various populations and to assign individuals to their populations of origin.

To infer phylogeny from microsatellite data several distance estimators were calculated and the resulting trees are shown in the following figures.

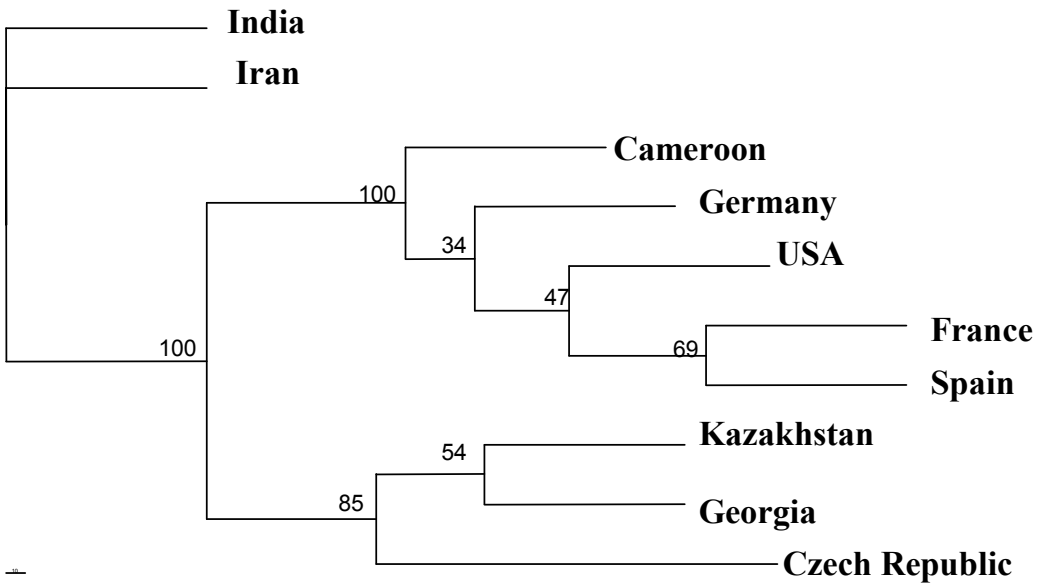



**Figure 5a:** Neighbour-Joining tree based on the proportion of shared alleles between populations (100 bootstraps)

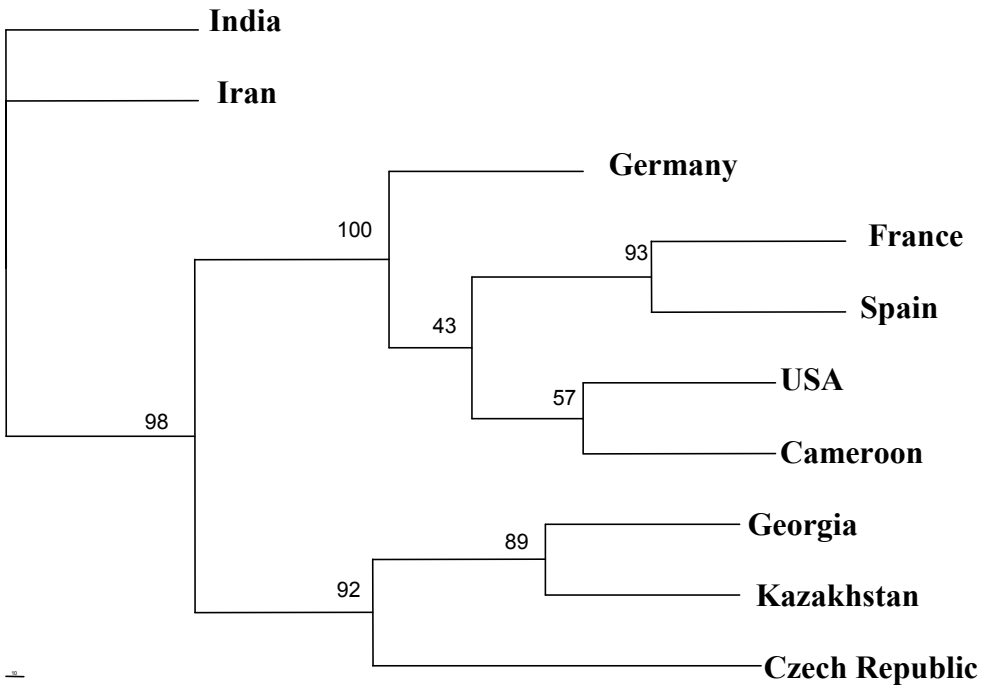



**Figure 5b:** Neighbour-Joining tree based on Nei's distance (100 bootstraps)

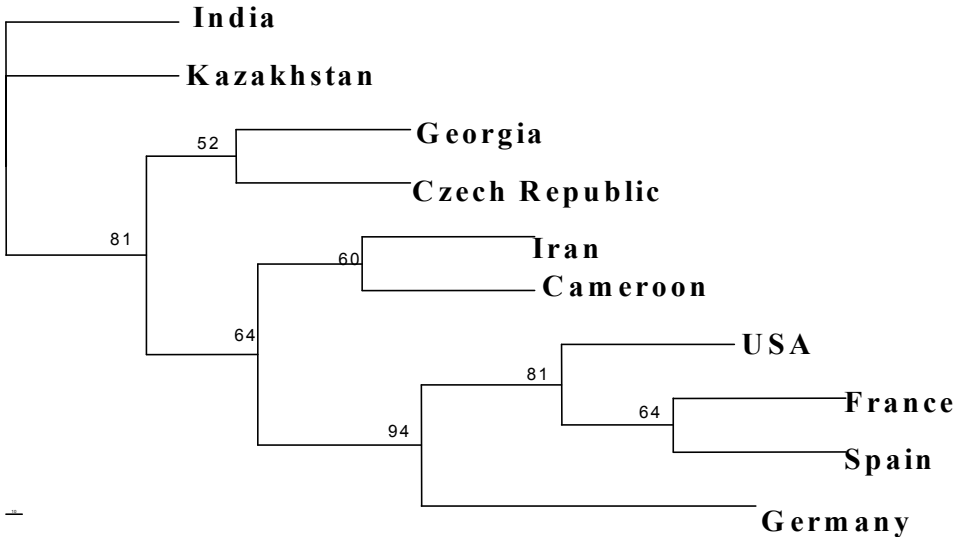



**Figure 5c:** Neighbour-Joining tree based on Slatkin's Rst distance (100 bootstraps)

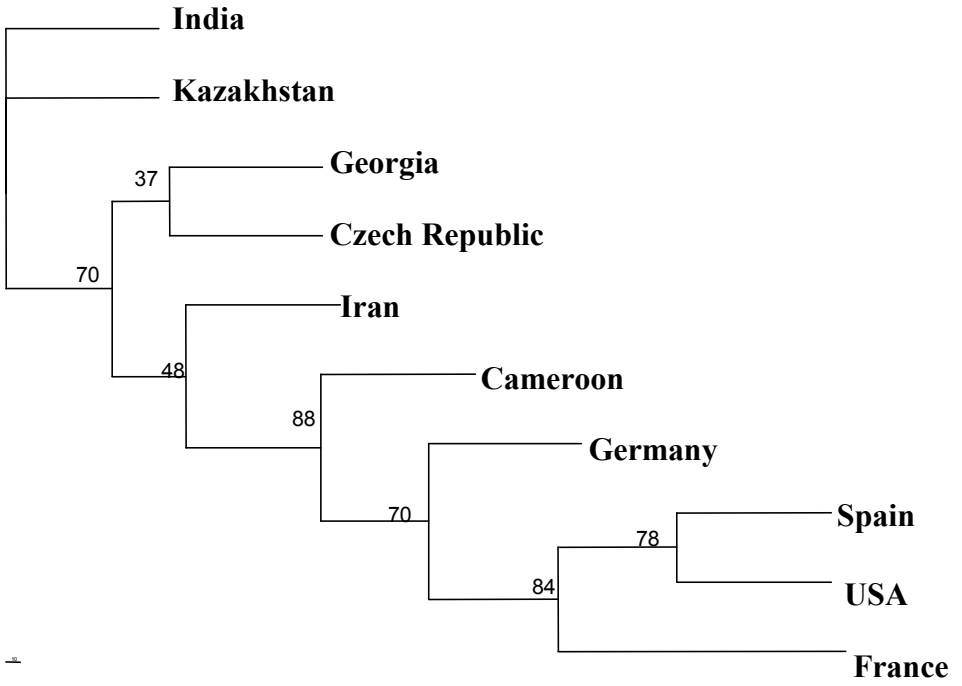



**Figure 5d:** Neighbour-Joining tree based on $\Delta\mu^2$ distance (100 bootstraps)

The Nei's distance tree reflects the same topology as the individual allele sharing tree. High bootstrap values support the divergence between the lineages. Although within the lineages populations are separated by relatively low bootstrap values. Similar results are achieved by calculating the proportion of shared alleles between the populations. Only the topology within the *domesticus* clade differs from Nei's distance. Bootstrap values within the *domesticus* clade are low.

Analyses of the trees based on Rst and $\Delta\mu^2$-values cannot resolve the differences between the lineages. All bootstrap values are very low and a separation of the different taxa is not possible. Hence, these methods are considered not suitable to analyse this dataset.

Based on the allele sharing tree and on Nei's distance estimation deep clades between the different lineages are visible.
Within the lineages the *Mus musculus* animals from Kazakhstan and Georgia cluster very closely together followed by the geographically more distant mice from the Czech Republic.
Within the *Mus domesticus* clade the resolution is not obvious. Only France and Spain form a distinct cluster which was already evident in the D-loop analysis.
In the allele sharing tree based on individuals Germany appears to be the basal population from which the other populations evolved whereas in the allele sharing tree based on populations the Cameroon population seems to be more basal. Bootstrap values within the clades for this tree are low and do not allow any inferences about the true relationships.
The topology for *Mus domesticus* in the allele sharing tree based on individuals is starlike and the branch lengths are shorter than in the *Mus musculus* clade suggesting that the populations are younger. The shortest branch lengths are found in the Cameroon population. The short branch lengths within the *Mus domesticus* clade contradicts the results of the D-loop tree. Shorter branch lengths suggest younger populations while the D-loops sequences contained older lineages.
There is no correlation with the geographic distance between the populations as the geographically closest populations from Germany and France exhibit the greatest genetic distance.

## Discussion

### Inferences from the D-loop sequence analysis

As shown in other studies (Boursot et al. 1993, Prager et al. 1998, Duplantier et al. 2002) sequencing D-loops is a valuable tool to identify different lineages within the house mouse species complex. D-loop sequencing of the ten wild caught and lab derived mouse populations analysed in this study allowed to assign mice to the different taxa such as *Mus musculus*, *Mus domesticus* and *Mus castaneus*. In particular in the hybrid zone within the Czech Republic mitochondrial markers allowed to assign the samples to one of the two clades (Munclinger et al. 2002).
The Indian and Iranian samples cluster closely together and form the basal sister group to the *musculus-domesticus* clades. However, these results do not provide any significant contribution to the ongoing discussion about the localisation of the phylogenetic origin of the house mouse being either in northern India (Boursot et al. 1996) or closer to the Arabian peninsular (Prager et al. 1998), because both populations appear to be equally basal to the *Mus domesticus* and *Mus musculus* clades in the allele sharing distance and Nei's distance calculations.

Within the clades D-loop sequences do not resolve differences between the populations. Only the animals from France and some from Kazakhstan form distinct groups while all other groups are fairly strongly intermixed stressing the fact that some of the populations are still very young and lineage sorting has not occurred. This strong mixture of haplotypes from different countries corresponds to the studies of Prager et al. (1996) who analysed samples from a broad area spanning from Egypt to east Asia. They found monophyletic lineages for the different taxa but no correlation with isolation by distance of the different haplotypes. Although only four out of 78 haplotypes were shared between the populations the many differences between them consisted of a majority of single nucleotide exchanges which do not contribute strongly to the genetic distance. Additionally, some of those nucleotide exchanges were singletons which may be regarded as very recent mutation events or as sequencing artefacts.

The populations from the Czech Republic and especially from Cameroon show very low haplotype frequencies. For the Czech Republic this result might be explained by the differences in the sampling design: although the sampling area covered the biggest region the

number of "different locations" - according to the my definition of different locations - within this area is rather low. Animals caught from the same location are likely to be related and members of the same deme.
For the Cameroon populations the extreme low number of haplotypes is probably due to the founder effect of the young African population.

The populations from the wild mouse strains kept in the breeding centre consisted of only one or two haplotypes which might be due to the fact that the original samples came from single localities. Under captive breeding conditions with a low effective population size haplotypes are likely to be lost by genetic drift.

The longest branches within the D-loop tree are found in the *domesticus* clade which might lead to the interpretation that *Mus domesticus* is more diverse and therefore older than the *Mus musculus* clade as suggested by Boursot et al. (1996) and Prager et al. (1996). Since my *domesticus* samples are known to be taken from rather young populations (Boursot et al. 1993) I assume that the depth of the clades in my samples is dependent on the origin of the different ancestral founders and is therefore a random result. Inferences about ages of populations based on diversity are only reasonable if more representative sampling is done in continuous regions as for example in Prager et al. (1998).

**The use of microsatellite markers in phylogeny and population history of house mouse populations**

To infer phylogeny within the clades I tried to get a resolution with faster evolving microsatellite markers. Depending on the different time frames, the number of samples and loci the different estimators perform differentially well to describe phylogenetic relationships between populations (Goldstein et al. 1995a, Paetkaeu et al. 1997, Richard et al. 2001). For the mouse data set the Rst- and $\Delta\mu^2$-distance estimators gave very bad resolutions of the divergence between lineages and populations: Lineages previously clearly identified by D-loop analysis were mixed again; generally bootstrap values were very low. Rst- and $\Delta\mu^2$-estimators are based on the stepwise mutation model and measure, in particular, the variance in repeat numbers (Richard et al. 2001). These distance estimators produce wrong results, because of violated assumptions of the mutation model, very tight constraints on alleles distribution, or very old separation of populations (Goldstein et al. 1995a, Richard et al. 2001, Paetkau et al. 1997). Constraints on allele distribution result in the effect that after sufficient

time of divergence all distance measurements applied to these loci will reach maximum values and, therefore, are not suitable to infer phylogenies of very distantly related taxa (Goldstein et al. 1995). Concerning my populations these estimates did not resolve any phylogeny because of high differences in mutation rates and deviations from the stepwise mutation model in the used microsatellites.

The different distance estimators mainly affected the clustering of the captive bred strains. These populations are characterised by their low effective population size indicating that the distance estimators perform differentially well for different population sizes (Schlötterer 2001). Removing these populations from the analysis results in more consistent phylogenies for all different distance estimators (data not shown).

Best results were achieved by applying a simple similarity index based on the proportion of shared alleles as no assumptions about mutation mechanisms or populations sizes are implied. These estimators were applied to populations as well as to individuals and both applications led to similar results. Only Cameroon and Germany changed places in the two allele sharing trees leaving all other topologies unchanged. The reason for the difference in the two trees is that the German population consists of several paraphyletic clusters which is shown in the allele sharing tree based on individuals. In the individual allele sharing tree the Cameroon population is a direct sister group to one paraphyletic group of the German population indicating a close relationship to the German samples. This relationship can be problematic for the resolution of the true relationship in the allele sharing tree based on populations.

Iran and India can be regarded as basal populations from which the other two clades descended. The separation between *musculus* and *domesticus* is very well supported by high bootstrap values in the allele sharing and Nei's distance tree. Within the *musculus* clade there is a clear divergence between Czech Republic and the Kazakhstan-Georgia cluster which would partially correspond to an isolation by distance pattern. The allele diversity is very high and especially the branches in the individual allele sharing tree are very long compared to the *domesticus* branches suggesting an older origin of these populations. This contradicts the interpretation of the D-loop results with longer branches in the *Mus domesticus* populations. Nevertheless, I assume that the D-loop results are an artefact of the sampling design. The microsatellite analysis appears to be more reliable because for the *Mus domesticus* populations four populations, collected under my stringent sampling regime, exhibit the same

pattern in the individual allele sharing tree with a starlike phylogeny. This phylogeny would correspond to the assumption that all *domesticus* populations are rather young. Interestingly, there is no real difference in variability and shape of the phylogeny between the European populations and the "overseas" populations which are known to be younger than 500 years. The European samples are probably younger than the proposed 4000 to 2800 years corresponding with the assumption of Cucchi (oral presentation at ECM4, Brno). To verify this assumption it would be necessary to sample older *Mus domesticus* populations from the near East which are known to be at least 10.000 years old.

In contrast to the starlike phylogenies of the *Mus domesticus* samples the Kazakhstan population exhibits a much more structured allele sharing tree. No sampling artefact can be responsible for this as the Kazakhstan population was collected under my stringent sampling regime and the size of the sampling area is comparable to the areas in France and Germany. Therefore, I assume that the Kazakhstan population is older than the *domesticus* populations and possesses a higher gene diversity and a longer genealogy of the microsatellite alleles. No inferences can be made about the Czech population because of the different sampling regime.

A young origin of the *Mus domesticus* clade also corresponds to the low resolution of the cluster with microsatellite alleles. No real inferences can be made about the relationships between the examined populations. D-loops are shared between Germany, Cameroon and the USA suggesting that these populations are more closely related to each other which, however, is only partially supported by microsatellite analysis. Based on the D-loop and microsatellite analysis France and Spain seem to form distinct groups. The France samples were caught in a remote area in the Centre of the Massif Central where agricultural land-use has persisted for a couple of hundred years or even longer, thus allowing relative constant development of mouse populations. As the turnover of demes is very high it is very probable that in such a remote area recolonization takes place from neighbouring demes (Singleton 1992). Thus, in remote areas the chance of long distance dispersal is lower which could explain the distinct D-loops and microsatellite clusters for the France population.

In contrast, the German samples belong to an area which has definitely undergone many structural changes during the past centuries. Agricultural land-use was not as persistent as in the French area resulting in less stable mouse populations and presumably in more possibilities of long distance transport and colonization events. Interestingly these populations

share more haplotypes with the "overseas" populations in Cameroon and USA leading to the assumption that they are either more closely related to each other than anticipated or that they were affected in similar ways by recent and still ongoing colonization events.

Concerning the lab strains I also found distinct clusters for all populations. Due to their reduced effective population size, the small number of individuals and the different sampling regime no further inferences about age and general diversity estimates compared to the wild caught populations can be made.

Summarising, I found that the *Mus musculus* populations can be nicely separated by microsatellite analysis and also inferences about phylogeography can be made. As the *Mus domesticus* populations are rather young no clear phylogeny can be inferred from microsatellite analysis.

Concerning the diversity of the studied mouse populations the wild caught populations exhibit comparable gene diversities. Only the Cameroon population shows a significantly reduced number of alleles and heterozygosity. Contrary, mice from the USA show higher diversities although the sample size is much smaller. The low gene diversity and the short branches in the allele sharing tree of the Cameroon population indicate that this population is very young and was recently founded.
The lab populations exhibit the lowest gene diversities due to the small number of individuals and the small effective population size under breeding conditions. They were valuable as outgroups for my phylogenetic estimates but with these populations no further estimates about their age and general population genetic parameters are possible.

Thus, microsatellites are a valuable tool to show that mice within one area can be distinguished by their microsatellite alleles as almost each mouse of the 337 samples could be assigned to its original population by calculating its individual allele sharing index among all individuals.

This result is very important because most of the loci deviated from Hardy-Weinberg Equilibrium which is always an indicator for subdivided populations. Since based on their family structure mouse populations are strongly subdivided it was very important to prove that demes within a neighbourhood are closer related to each other than to those from distant

regions. Although demes are short living entities mouse populations of a whole region develop as homogeneous units because of short distance migration and recolonization by neighbouring demes. Thus, mice from one area are assumed to have evolved under the same environmental demands and selection pressures which naturally differ from those in other areas. Most probably these different influences may have left signatures of selection within the different mouse populations. Since phylogenetic analysis revealed that most of the wild caught populations form distinct units and exhibit similar gene diversities these populations can be used to study the differences within the genomes which are a result of different selection pressures.

# Screening multiple microsatellite loci identifies selective sweep genes by a hitchhiking approach

## Introduction

Evolution and natural selection is a permanent process which in the past has reshaped organisms and continues to form and change existing organisms. Nowadays, populations represent just a snapshot of the evolution and their acquired characteristics do not allow inferences about the evolutionary forces that have formed them (Grant & Grant 2002). Although the results of natural selection are clearly visible in terms of different species and various adaptations within species, almost nothing is known about the actual mechanisms of evolution that created the different traits (Rieseberg et al. 2002). The mechanisms of selection and the molecular evolution have been a dominant issue of discussion in the past three decades and many population genetic theories have been developed (Kimura 1983, Nei 1987, Maynard Smith & Haigh 1974). On the one hand, neutral mutations within a species result in DNA variations which are proportional to the amount of divergence between species; on the other hand, non-neutral mutations lead to differentiation independent of intra-specific variation. Non-neutral mutations under negative or positive selection pressure are either eliminated from a gene-pool or they increase in frequency within populations. These phenomena are known as background selection or selective sweeps (Maynard Smith & Haigh 1974). In both cases, these effects leave traces in the genomes because the elimination or the increase in frequency of a mutation also affect the flanking regions; this effect is called hitchhiking event (Maynard Smith & Haigh 1974, Fay & Wu 2000). Hitchhiking events result in reduced polymorphism in the flanking regions of the advantageous mutation (Kaplan et al. 1989). The region exhibiting the reduced polymorphism is referred to as the selective sweep window. Although it is still unclear to which extent neutral, positive and negative selection are acting upon populations and which molecular mechanisms are responsible for adaptations (Fay & Wu 2001, Schlötterer 2002a) these theories allowed the development of several tests for the identification of regions and genes under selection within populations and species.

In the past decade the development of molecular techniques became a powerful tool in population genetics, and they were applied to verify the theoretical models in natural populations. First approaches used single candidate genes that were already known to be shaped by environmental selection pressure. Changes within the genomic region of a

candidate gene were directly linked to a change of the phenotype. These genes were, for example, resistance genes in mosquitoes (Yan et al. 1998), in rats (Kohn et al. 2000, Kohn et al. 2003) and in the Malaria parasites *Plasmodium falciparum* (Wootton et al 2002, Nair et al. 2003). Other genes identified to be subjected to positive selection and selective sweeps were involved in reproductive isolation between species, for example the Sdic gene (Nurminsky et al. 1998) and the Nup96 gene (Presgrave et al. 2003) in *Drosophila*, or they were involved in lineage specific new characteristics such as the Foxp2 gene which is relevant for the human ability to develop language and speech (Enard et al. 2002).

All these studies dealt with the monogenic traits that were shaped under extreme selection pressures and some of these studies even identified species specific genes. Nevertheless, selective sweeps and hitchhiking events can influence polygenic traits and occur under weaker selection pressure in local populations of a species. They may be used to identify genomic regions exposed to recent selection pressures although the different phenotypes are not easily measurable (Schlötterer 2003). As genome wide approaches with neutral markers already have identified multiple regions exhibiting traces of hitchhiking events this method seems to be a tool to identify "new" i.e. not yet identified genes which play a role in adaptation processes: first evidence resulted from screening neutral markers which exhibited reduction of polymorphism or skews in the allele frequency distributions independent on demographic events (Schlötterer et al. 1997, Payseur et al. 2002, Vigouroux et al. 2002, Kauer et al. 2002, Kayser et al. 2003).

Several methodological concepts were developed to identify regions subjected to selection and the different methods and statistics were reviewed by Fay & Wu (in press), Nurminsky (2001) and Schlötterer (2003).

The molecular methods to identify regions under selection are either based on sequencing approaches or on typing microsatellite markers. Both methods require different statistical treatments. Sequencing approaches offer the possibility to identify ancestral and derived states of a polymorphism by sequencing the common ancestor of two populations or a closely related taxon as an outgroup. Several statistics as the H-Test (Fay & Wu 2000) or the composite likelihood test (Kim & Stephan 2002) apply these characteristics to detect local signatures of selection. Those tests analyse the frequency of new derived mutations, and calculate whether the frequency spectrum follows neutral or non-neutral expectations. Similar

approaches which are only based on the frequency of polymorphisms were developed by Tajima (1989) and Fu & Li (1993). The Hudson-Kreitman Aguadé test compares interspecific divergence to intraspecific polymorphism which correlate under neutral conditions (Hudson et al. 1987).

Microsatellites require very different tests. Since many different alleles exist at one locus the ancestral states of microsatellites are difficult to determine. Wiehe (1998) showed that microsatellite variation is effected by selection due to the hitchhiking effect but the effect is dependent on the time to fixation of the linked favourable mutation. High mutation rates quickly obscure the hitchhiking effect. Hence microsatellites only show effects of selective sweeps within a certain time frame which is dependent on the mutation and recombination rate.

The expected effect of a selective sweep event on a microsatellite locus is that the locus shows reduced variability within the population which was affected by the selective sweep event. To identify such a locus inferences about the neutral variability are necessary and demographic factors such as general low variabilities must be taken into account. Since mutation rates differ strongly between loci (Ellegren 2000) neutral assumption for each microsatellite are difficult to infer. Schlötterer (2002b) developed a statistical test that is independent of locus specific or demographic effects and solely relies on the comparison of gene diversities of different populations which are likely to be affected by different selection pressures: the lnRV and lnRH statistics are based on direct pairwise comparisons of variance in repeat units (V) or heterozygosities (H) at multiple loci. Loci at the same genomic location are assumed to evolve under the same mechanisms and same constraints. Due to the pairwise comparisons locus specific effects such as mutation rates are eliminated. Demographic effects are controlled for by inclusion of multiple loci because all loci of one population are equally affected by a demographic event.

This approach allows the identification of extreme differences at one locus independent of demographic effects which might be the result of a locus specific selection event.

Other tests like the Lewontin-Krakauer test apply genetic distance estimators such as the Fst-values per locus to identify loci for which distance estimators deviate extremely from the other loci (Baer 1999). Since the variance of Fst values per locus under neutrality is substantially high this could lead to many false positive sweeps if this method is solely applied (Schlötterer 2003). Within populations, deviation from neutrality can be inferred from

the expected and observed allele frequency distribution as done in a study on human microsatellite allele distributions by Payseur et al. (2002).
Other approaches rely on the fact that selective sweeps show extended linkage disequilibrium as found in studies of Gilad et al. (2002), Kohn et al. (2000) and Sabeti et al. (2002).

The growing amount of theoretical knowledge in combination with today's high throughput genotyping procedures allows to address the question to which extent natural populations are affected by selection and, in particular, which genes are responsible for the different adaptations or even speciation events. To study these questions the house mouse was selected as a model system because

- the species has colonized many different habitats all over the world (chapter 1) and was certainly confronted with different selection pressures which must have left traces within the genome of the different populations;
- the house mouse species complex is relatively young. It diverged from the other *Mus* taxa less than one million years ago and even younger splits in the different lineages such as *Mus domesticus* and *Mus musculus* occurred. The new world continents are known to be colonized by the house mouse within the last centuries forming very young entities (chapter 1);
- the first draft of the full genomic sequence became available in 2002 at public database such as ENSEMBL (http://www.ensembl.org/) and NCBI (http://www.ncbi.nlm.nih.gov/genome/guide/mouse/ );
- the house mouse is a model system in biomedical research and many genetic, biochemical and physiological information about the species is available;
- the ecology of wild house mice has been extensively studied.

All these factors are an important basis for studying and understanding natural selection and the molecular mechanisms underlying the adaptations of the house mouse to different environments.

To identify selective sweeps I chose genotyping of microsatellite loci because genotyping costs are lower and the processing of samples is faster than in sequencing approaches. In total, 81 microsatellite loci in six wild caught populations from Germany, France, Cameroon, USA, Czech Republic and Kazakhstan were screened. Detailed information about the populations was given in chapter 1. Most microsatellite loci were directly linked to genes that are likely to play a role in response to the abiotic or biotic environment: resistance genes are likely to be

shaped by different pathogenic environments; genes involved in environmental perception such as olfactory receptor genes; saliva genes which are involved in social interactions such as mate choice; genes that are differentially regulated as fast responses to changing environments and genes that are already known to play a role in species barriers such as hybrid sterility genes. In addition, four microsatellites were chosen randomly from the genetic map of the Whitehead Institute (http://www-genome.wi.mit.edu/cgi-bin/mouse/index). I call this approach "multi-locus candidate gene approach". This approach was chosen because the extend of sweep windows are still a matter of debate and the direct physical linkage to the candidate locus increases the chance of detecting a sweep event (Nair et al. 2003, Glinka et al. 2003, Saez et al. 2003). In a separate approach, 106 random microsatellite loci, selected from the genetic maps of the mouse (Dietrich et al. 1996), were genotyped in four populations from Germany, France, Cameroon and Kazakhstan, which were also used in the multi-locus candidate gene approach. This approach intended to evaluate the likelihood of detecting selective sweep events by randomly chosen microsatellite loci.

For statistical analysis the lnRV and lnRH tests were chosen since they proved to be robust to small demographic differences such as recent bottlenecks and differences in mutation rates of the microsatellite loci (Schlötterer 2002b).

The aim of these microsatellite screens was:

- to evaluate whether microsatellite genotyping is a suitable tool to identify traces of selection within different populations of the house mouse;
- to determine which kind of genes are affected by selective sweeps;
- to evaluate the usefulness of the "multi-locus candidate gene approach" versus the "random microsatellite approach";
- to identify possible differences in the performance of the different microsatellite markers depending on their characteristics such as repeat length and variability.

## Materials and Methods

Six wild caught mouse populations were used to search for selective sweep events. Two populations were *Mus musculus* types from Kazakhstan and the Czech Republic and four populations were *Mus domesticus* types from Germany, France, Cameroon and the USA. Details about the sampling design, the phylogeny and population history of these populations are presented in chapter 1. The analysis for selective sweeps is based on the same 81 microsatellite loci that were also used to infer the phylogenetic relationships. These microsatellites were selected from genomic sequences in the flanking regions or within introns of genes which are likely to respond to environmental selection pressures. Genes were chosen from the following databases:

- Pubmed, Locuslink and Nucleotide database (http://www.ncbi.nlm.nih.gov/) were screened with the following keywords: “interferon”, “resistance”, “speciation”, “behaviour”, “imprinting”, “olfactorious”, “saliva”.
- Differentially expressed genes from a preliminary micro-array experiment comparing differential expressions from different mouse taxa.

Genomic sequences of these genes were downloaded and screened for microsatellite repeats by applying a text search for all different kinds of di-, tri- and tetra-nucleotide repeats. An overview of the used loci, their search history, sequence characteristics and location on the chromosomes is given in Table 1. This screen will be referred to as the “multi-locus candidate gene approach”.
Additionally, 48 animals each of the populations from Kazakhstan, Germany, France and Cameroon were screened for 106 microsatellite loci. Theses marker were extracted from the genetic map of the Whitehead Institute: 41 microsatellites were located on chromosome 1 and selected to represent equal centiMorgan locations; 39 markers were located on chromosome 9, and 26 markers were located on chromosome 15. This screen is referred to as the "random microsatellite screen".

**Table 1:** List of used loci containing short tandem repeats within their sequences, the search history of the loci, the Genebank accession number, the characteristics of the repeat types, the location on the chromosome and the distance of the repetitive sequences to the coding regions of the genes.

| internal number | search history | Accession | Name | size | Number of repeats | Chromo-some | Location of microsatellite locus |
|---|---|---|---|---|---|---|---|
| P1 | Genebank: Interferon | M37707 | Ly-6E/A gene | 267 | $(TGGG)_4$ | 15 | within genomic sequence; max 1 kb distance to exon |
| P7 | | M32489 | Mouse interferon consensus sequence binding protein mRNA | 684 | indel | 8 | within genomic sequence; max 1 kb distance to exon |
| P8 | | AJ299405 | Microtubule associated protein 44 (Mtap44gene) | 160 | $(CA)_{24}$ | 3 | within genomic sequence; max 1 kb distance to exon |
| P10 | | L09126 | Calcium-independent nitricoxide synthase (iNOS) | 200 | $(GT)_{36}$ | 11 | within genomic sequence; max 1 kb distance to exon |
| P11 | | L23806 | Nitricoxide synthase (Nos-1) | 249 | $(GT)_{36}$ | 11 | within genomic sequence; max 1 kb distance to exon |
| P12 | | X07640 | Cell surface glycoprotein Mac-1 alpha-chain | 217 | $(GT)(GA)_{48}$ | 7 | within genomic sequence; max 1 kb distance to exon |
| P14 | | U20949 | Lymphoid-specific interferon regulatory factor (LSIRF) gene | 180 | (GTT)(GGT): 97 bp | 13 | within genomic sequence; max 1 kb distance to exon |
| P19 | | X01973 | Interferon alpha 4 (MuIFN-alpha4) | 185 | $(GTTT)(ATTT)_{12}$ | 4 | within genomic sequence; max 1 kb distance to exon |
| P25 | | NM_008331 | Interferon-induced protein with tetratricopeptide repeats1 (Ifit1) | 183 | $(GTT)_9$ | 19 | within genomic sequence; max 1 kb distance to exon |
| P26 | | U06237 | Interferon alpha/beta receptor (IFNAR) gene | 196 | (GTn): 47 bp | 16 | within genomic sequence; max 1 kb distance to exon |
| P27 | | U19119 | G-protein-like LRG-47 | 156 | $(TAAA)_8$ | 11 | within genomic sequence; max 1 kb distance to exon |
| P29 | Locuslink: Interferon | D17Mit71 | MS next to gene protein kinase, interferon-inducible double stranded RNA dependent | 106 | $(GT)_{23}$ | 17 | same linkage group |
| P31 | | D4Mit302 | MS next to interferon alpha family | 108 | $(TG)_{23}$ | 4 | 0.1 cM distance (MGI backcross map) |
| P32 | | D1Mit349 | MS next to interferon activated gene 201 (Ifi201) | 118 | $(TG)_{26}$ | 1 | 0.6 cM distance (MGI backcross map) |
| P33 | randomly chosen | D1Mit1 | D1Mit1 | 125 | $(CA)_{24}$ | 1 | |
| P34 | | D1Mit100 | D1Mit100 | 244 | $(TG)_{22}(TA)_{48}$ | 1 | |
| P35 | | D1Mit200 | D1Mit200 | 199 | $(TAGA)_{19}$(TA-GT-CA):58 bp; interrupted | 1 | |
| P36 | differentially expressed | AB051897 | Scya6, Scya9, Scya16-ps, Scya5 genes | 338 | $(CT)(GT)_{96}$ | 11 | 8 kb upstream Scy6 |
| P37 | | D83329 | Prostaglandin D2 synthase | 291 | $(TA)(CA)_{100}$ | 2 | 5 kb upstream |
| P38 | | D78343 | Ig gamma-chain, secrete-type and membrane-bound | 225 | $(GA)_{48}$ | 12 | within genomic sequence; max 1 kb distance to exon |
| P42 | | AF021345 | Plasma selenoprotein P (SELP) gene | 290 | $(GT)_{12}(CA)_{23}(CAGACA)_4$ | 15 | within genomic sequence; max 1 kb distance to exon |
| P44 | | D78344 | Ig gamma-chains, partial cds | 304 | $(TC)(GT)_{43}$ | 14 | 4 kb distnat from exons |
| P45 | | M65161 | Pro-alpha1 (II) collagen chain gene | 267 | $(GA)_{36}$ | 11 | within intron; max 1 kb from exon |
| P46 | | M17376 | Alpha-1-acid glycoprotein I (AGP-1) gene | 350 | $(GT)_{28}$ | 4 | within genomic sequence; max 1 kb distance to exon |
| P47 | | AJ271004 | TFF3/ITF gene for Trefoil Factor 3/Intestinal Trefoil Factor protein | 541 | $(GA)(CA)_{150}$ | 2 | within genomic sequence; max 1 kb distance to exon |
| P49 | | X56790 | Growth factor inducible immediate early gene cyr61 | 352 | $(CA)_{27}$ | 3 | within genomic sequence; max 1 kb distance to exon |

| internal number | search history | Accession | Name | size | Number of repeats | Chromo-some | Location of microsatellite locus |
|---|---|---|---|---|---|---|---|
| P50 | Locuslink: Interferon, linked MS | D3Mit19 | MS next to NDV-induced circulating interferon (Ifi1) | 159 | $(CA)_{23}$ | 3 | 0.1 cM distance (MGI backcross map) |
| P51 | | D17Mit200 | MS next to interferon activated gene 15 (Ifi15) | 123 | $(CA)_{12}$ | 17 | 0.3 cM distance (MGI backcross map) |
| P52 | | D10Mit187 | MS next to interferon gamma | 254 | $(TG)_{29}$ | 10 | 0.5 cM distance (MGI backcross map) |
| P53 | | D6Mit47 | MS next to interferon regulatory factor 5 (Irf5) | 189 | $(CA)_{15}$ | 17 | 0.3 cM distance (MGI backcross map) |
| P54 | | D14Mit261 | MS next to interferon dependent positive acting transcription factor 3 gamma (Isgf3g) | 92 | $(GT)_{18}$ | 5 | same linkage group (MGI backcross map) |
| P55 | | D5Mit89 | MS next to chemokine (C-X-C motif) ligand 9 (Cxcl9) | 147 | $(CA)_{33}$ interrrupted | 6 | same linkage group (MGI backcross map) |
| P56 | | D11Mit140 | MS next to interleukin 3 (Ili3) | 129 | $(GT)_{19}$ | 9 | 0.5 cM distance (MGI backcross map) |
| P57 | differentially expressed | M73741 | Alpha-B2-crystallin gene | 328 | $(CT)_{20}$ | 4 | within genomic sequence; max 1 kb distance to exon |
| P58 | | M36120 | Keratin 19 gene | 348 | $(CAGA)_{11}$ | 15 | within genomic sequence; max 1 kb distance to exon |
| P59 | | AF071001 | PHR1 (Phr1) gene | 281 | $(GT)_{27}$ | 19 | 4 kb upstream |
| P61 | | X14061 | Beta-globin complex DNA for y, bh0, bh1, b1 and b2 genes, bh2 and bh3 pseudogenes | 363 | $(CA)_{15}$ | 7 | within intron; max 1 kb from exon |
| P62 | Locuslink: Interferon; MS within gene sequence | U44903 | GTP binding protein (IRG-47) gene | 369 | CAA/CAAAA: 63 bp | 11 | within genomic sequence; max 1 kb distance to exon |
| P63 | | NM_008331 | Interferon-induced protein with tetratricopeptide repeats 1 (Ifit1) | 388 | $(GTT)_{10}$ | 19 | within genomic sequence; max 1 kb distance to exon |
| P66 | | U06237 | Interferon alpha/beta receptor (IFNAR) gene | 387 | (GTTT): 59 bp, interrupted | 16 | within genomic sequence; max 1 kb distance to exon |
| P67 | randomly chosen | D17Mit23 | D17Mit23 | 137 | (GTTT-GTT): 49 bp, interrupted | 17 | |
| P71 | Genebank: speciation | NM_011445 | SRY-box containing gene 6 (Sox6) | 197 | $(CAA)_5$ | 7 | within genomic sequence; max 1 kb distance to exon |
| P72 | | NM_011444 | SRY-box containing gene 5 (Sox5) | 162 | $(GAAA)_7$ | 6 | within genomic sequence; max 1 kb distance to exon |
| P73 | | AF070933 | YFVB sex determining region of Y protein | 265 | (GA/GGAA): 114 bp | Y | within genomic sequence; max 1 kb distance to exon |
| P76 | | AF342999 | Axonemal dynein heavy chain 8 Dnahc8 gene | 330 | (CAA/CAAAA): 71 bp | 17 | 2 kb upstream |
| P83 | | M32352 | Mouse renin (Ren-1-d) gene | 238 | $(CA)_{18}$ | 1 | within genomic sequence; max 1 kb distance to exon |
| P85 | | U84291 | Ornithine decarboxylase antizyme gene | 232 | $(GA)_{39}$ | 10 | within genomic sequence; max 1 kb distance to exon |
| P92 | | AF363577 | Haplotype t axonemal dynein heavy chain 8 short form 2 (Dnahc8) | 299 | (TCC): 45 bp, interupted | 17 | within genomic sequence; max 1 kb distance to exon |
| P94 | | NM_013486 | CD2 antigen (Cd2) | 184 | (TCC): 45 bp, interupted | 3 | within genomic sequence; max 1 kb distance to exon |
| P103 | Pubmed: saliva gene | AJ300673 | Beta-defensin 8 | 326 | $(GT)(CT)_{100}$ interrupted | 8 | within genomic sequence; max 1 kb distance to exon |
| P104 | | AB063110 | Beta-defensin 6 | 327 | $(GT)(CT)_{100}$ interrupted | 8 | within genomic sequence; max 1 kb distance to exon |
| P106 | | D29794 | T cell receptor gamma chain | 337 | $(GT)(CT)_{90}$ | 13 | within genomic sequence; max 1 kb distance to exon |
| P108 | | X68699 | Psp gene for parotid secretory protein | 237 | $(GT)_{25}$ | 2 | within genomic sequence; max 1 kb distance to exon |

| internal number | search history | Accession | Name | size | Number of repeats | Chromo-some | Location of microsatellite locus |
|---|---|---|---|---|---|---|---|
| P116 | | U82375 | MSG2alpha, beta, gamma, delta and epsilon salivary protein (Vcs2) | 324 | $(GTTT)(ATTT)_{24}$ | 5 | within genomic sequence; max 1 kb distance to exon |
| P118 | Genebank: olfactory genes | AC091802 | Chromosome 2 clone RP23-52P17 | 282 | (TC/TTCC): 96 bp | 2 | |
| P119 | | AC091747 | Chromosome 19 clone RP23-64I16 | 263 | GTT/GTTT: 62 bp | 19 | |
| P120 | | AC091745 | RP23-52M7 | 238 | $(CA)_{23}$ | 2 | |
| P121 | | AC091743 | RP23-71E10 | 361 | (TA): 120 bp, interrupted | 10 | |
| P122 | | AL136158 | RP21-538M10 | 381 | (CTTT/CCTT): 128 bp | 17 | |
| P126 | | AF133300 | MOR 3'Beta1, MOR 3'Beta2, MOR 3'Beta3, MOR 3'Beta4, Cbx3 pseudogene, MOR 3'Beta5 and MOR 3'Beta6 genes | 215 | $(GT)_{18}(GA)_{15}$ | 7 | 21 kb upstream to exon |
| P127 | | Y09167 | MATH4B gene | 319 | $(GA)_{22}$ | 10 | within genomic sequence; max 1 kb distance to exon |
| P129 | | AJ251155 | Nasal embryonic LHRH factor (Nelf-pending) | 219 | $(GT)_{23}$ | 4 | 6 kb upstream |
| P133 | | NM_008192 | Guanylyl cyclase 2e (Gucy2e) | 253 | (GACA): 60 bp | 11 | within genomic sequence; max 1 kb distance to exon |
| P139 | | AF016619 | Rb-8 neural cell adhesion molecule short form precursor (RNCAM) | 194 | (CCTCT): 37 bp | 16 | within genomic sequence; max 1 kb distance to exon |
| P141 | | U49391 | Cyclic nucleotide-gated olfactory channel protein gene | 327 | (GTTT/GT): 35 bp | X | within genomic sequence; max 1 kb distance to exon |
| P142 | | X92969 | OR23 gene | 333 | $(TC/GT)_{38}$ | 1 | within genomic sequence; max 1 kb distance to exon |
| P146 | | U01213 | Olfactory marker protein (OMP) gene | 312 | (CCT): 112 bp | 7 | within genomic sequence; max 1 kb distance to exon |
| P148 | Genebank: imprinting | AB007765 | Mest gene | 249 | $(GT)_{25}$ | 6 | within genomic sequence; max 1 kb distance to exon |
| P149 | | NM_010514 | Insulin-like growth factor 2 (Igf2), mRNA | 212 | $(CA)_{18}$ | 7 | within genomic sequence; max 1 kb distance to exon |
| P150 | | AF049091 | H19 and muscle-specific Nctc1 genes | 158 | $(GT)_{18}$ | 7 | 10 kb downstream |
| P151 | | AB030734 | Peg8/Igf2as | 148 | $(TA)_{14}$ | 7 | within genomic sequence; max 1 kb distance to exon |
| P153 | | AF130348 | Zfp127 protein gene | 195 | $(CA)_{14}$ | 7 | 4 kb upstream |
| P154 | | AF198619 | Snrpn gene | 217 | (GA/GAA): 46 bp | 7 | within genomic sequence; max 1 kb distance to exon |
| P155 | | AF081460 | Small nuclear ribonucleoprotein N gene | 187 | $(TG)_{23}$ | 7 | within genomic sequence; max 1 kb distance to exon |
| P156 | | D78349 | Preproadrenomedullin | 203 | $(TG)_{20}$ | 7 | within genomic sequence; max 1 kb distance to exon |
| P157 | | U84903 | L23 mitochondrial-related protein (L23mrp) gene | 188 | (TG): 16 bp, interrupted | 7 | within genomic sequence; max 1 kb distance to exon |
| P158 | | AF139595 | Mash2 gene | 168 | $(CA)_{31}$ | 7 | within genomic sequence; max 1 kb distance to exon |
| P159 | | AJ251788 | Tssc6 gene | 255 | $(CA)_{21}$ | 7 | 5 kb upstream |
| P160 | | AP001916 | Clone:B131C | 223 | $(CA)_{20}$ | 7 | |
| P111a | linked MS to olfactory gene | D15Mit243 | Aquaporin 5 (Aqp5) | 125 | $(GT)_{22}$ | 15 | same linkage group (MGI backcross map) |
| P140a | linked MS to saliva gene | D14Mit203 | Olf-1/EBF-like-3 transcription factor (O/E-3) | 149 | $(GT)_{30}$ | 14 | 0.2 cM distance (MGI backcross map) |

**Statistical analysis**

General gene diversity estimates were calculated by using the Microsatellite Toolkit (Park 2001) and the progam MS Analyser (Dieringer & Schlötterer 2003). For all microsatellites of the "multi-locus candidate gene approach" the length of the repeat region was calculated by subtracting the smallest from the largest allele found among all samples. Markers were classified into pure dinucleotide repeat units, interrupted repeats, and tri- and tetranucleotide repeat units and analysed for any correlation of repeat type or length of the repetitive sequence with the variability indices such as expected heterozygosity, variance in repeat units and number of alleles. Kruskal-Wallis tests and regressions of the program SPSS 10.0. were applied to find significant differences.

To identify loci under selection pressure the lnRV and lnRH statistics according to Schlötterer (2002b) and Kauer et al. (2003) were applied. This approach is based on the following equations:

$$\ln RV = \ln \frac{Var\,\mathrm{Re}(loc1(pop1))}{Var\,\mathrm{Re}(loc1(pop2))}$$

*Var*Re = variance in repeat units
Loc = microsatellite locus
pop = population

$$\ln RH = \ln \frac{\left(\frac{1}{1-H(loc1, pop1)}\right)^2 - 1}{\left(\frac{1}{1-H(loc1, pop2)}\right)^2 - 1}$$

$H$ = heterozygosity

Based on the estimator:
$H = 1 - (1 / (1+2\theta)^{1/2})$
(Ohta & Kimura 1973)

The lnRV and lnRH ratios are calculated for all 81 microsatellite loci of the "multi-locus candidate gene approach" and all 106 loci of the "random microsatellite screen". For the "multi-locus candidate gene approach" the pairwise lnRV and lnRH values were calculated by comparing all single populations against each other resulting in 15 different comparisons; additionally the *Mus domesticus* populations from Germany, France, Cameroon and USA were pooled, as well as the *Mus musculus* populations from Czech Republic and Kazakhstan and both groups were compared against each other. For the "random microsatellite screen" the same analysis were performed for the subset of the populations from Germany, France, Cameroon and Kazakhstan.

Pooling of the samples allows to identify sweeps that occurred after the split of the species into the eastern and the western house mouse while the pairwise population comparisons intends to identify sweeps specific for single populations.

Simulations done by Schlötterer (2002b) predict that under neutrality the amount of all lnRV and lnRH values will follow a normal distribution. Due to the pairwise comparisons the statistic is robust to differences in mutation rates at single loci and to differences in constraints because these factors are eliminated by the ratio. The statistic is also not affected by differences in effective population size as for example small bottleneck events (Schlötterer 2002b). Assuming that most microsatellite markers will behave in a neutral manner, the distribution can be regarded as a test distribution against which single values can be compared to. Loci with lnRV or lnRH values below or above the 95%-interval are potential sweep loci. The statistics was applied to the different level of comparisons:

1) comparison of all loci between pooled *Mus musculus* and *Mus domesticus* animals to identify lineage specific sweeps;
2) pairwise comparisons of each population against all others for all loci to identify population specific sweeps;
3) summary of all lnRV or lnRH values from 2) to evaluate whether the same loci are identified as in the single pairwise comparisons.

For all levels the distribution of lnRV and lnRH values was tested for normality with the Kolmogorov-Smirnov goodness of fit test. Descriptive statistics and histograms were calculated with the SPSS 10.0 for Windows software package. The characteristics of all distributions such as the standard deviation were compared to each other. All loci that show values above or below the 95% confidence interval were listed.

Different characteristics were applied to distinguish “real” selective sweep loci from false positive results by analysing the number of extreme values per locus. Selective sweeps are assumed to occur either in a single population or in a monophyletic clade; exceptions from this rule are independent sweep events for the same locus. Therefore, sweep loci should exhibit extreme values in multiple comparisons. For loci exhibiting multiple extreme values the allele frequency distributions were analysed. Sweep loci are assumed to show reduced polymorphism in a single group of population which results in a high frequency of single alleles. Such effects should be detectable in allele frequency diagrams.

Generally, the used microsatellite loci were analysed whether the occurrence of extreme lnRV or lnRH values also depend on the characteristics of the repeat units. Chi-Square tests were applied to test for statistical differences; Kruskall-Wallis tests were performed to test for significant differences between the number of extreme values per locus dependent on the repeat type.

One potential sweep locus that followed the above mentioned criteria was examined for its sweep window size: the genomic sequence around the sweep locus was downloaded from the ENSEMBLE database, and short tandem repeat sequences were identified in the flanking regions with the program Tandem Repeats Finder (Benson 1999). Five microsatellite loci located within 50 kb upstream of the locus and 6 microsatellite loci located within 90 kb downstream were genotyped in all populations and the lnRV and lnRH values were calculated to identify extreme outliers within the same sweep window.

## Results

### General gene diversity of the different populations and characteristics of the different microsatellite markers

The "random microsatellite screen" and the "multi-locus candidate gene approach" were analysed separately and subsequently compared to each other. Contrary to the "random microsatellite screen", general gene diversity seems to be lower in the "multi-locus candidate gene approach" which might be caused by the fact that this screen has included loci with generally shorter repeat regions. Nevertheless, the characteristics of the populations with respect to each other are the same in both screens: the population from Cameroon exhibits the lowest gene diversity in both screens, followed by the USA population which was only screened in the "multi-locus candidate gene approach". The populations from France, Germany, Czech Republic and Kazakhstan exhibit relative equal gene diversities; only Kazakhstan has the highest number of alleles in both screens. The number of individuals from the USA is rather low and resulted in the lowest number of alleles. Still the characteristics of the USA population are comparable to the Cameroon population as analysed in detail in chapter 1, and subsequently, the population is also included in the search for selective sweep events. Gene diversities per population and per screen are summarised in Table 2.

**Table 2:** Characteristics and gene diversity estimates of the genotyped populations.

| Population | Sample size | Loci typed | Unbiased He | Obs He | No Alleles |
|---|---|---|---|---|---|
| **"Multi-locus candidate gene approach"** | | | | | |
| Cameroon | 68 | 81 | 0.5134 | 0.3917 | 6.67 |
| Czech Republic | 42 | 79 | 0.7054 | 0.3828 | 9.08 |
| France | 64 | 81 | 0.6454 | 0.4764 | 9.98 |
| Germany | 55 | 81 | 0.6524 | 0.4359 | 10.04 |
| Kazakhstan | 59 | 81 | 0.6994 | 0.5671 | 12.67 |
| USA | 12 | 79 | 0.6158 | 0.4782 | 5.23 |
| **"Random microsatellite screen"** | | | | | |
| Cameroon | 48 | 104 | 0.6722 | 0.4916 | 7.54 |
| France | 48 | 104 | 0.7853 | 0.5605 | 10.81 |
| Germany | 48 | 106 | 0.7991 | 0.5288 | 10.64 |
| Kazakhstan | 48 | 106 | 0.8111 | 0.6219 | 12.76 |

The microsatellite loci selected for the "multi-locus candidate gene approach" show very different characteristics: some markers contain long stretches of up to 100 dinucleotide repeats comparable to the markers used in the "random microsatellite screen", while other markers contain interruptions within the repetitive elements. Markers with tri- or

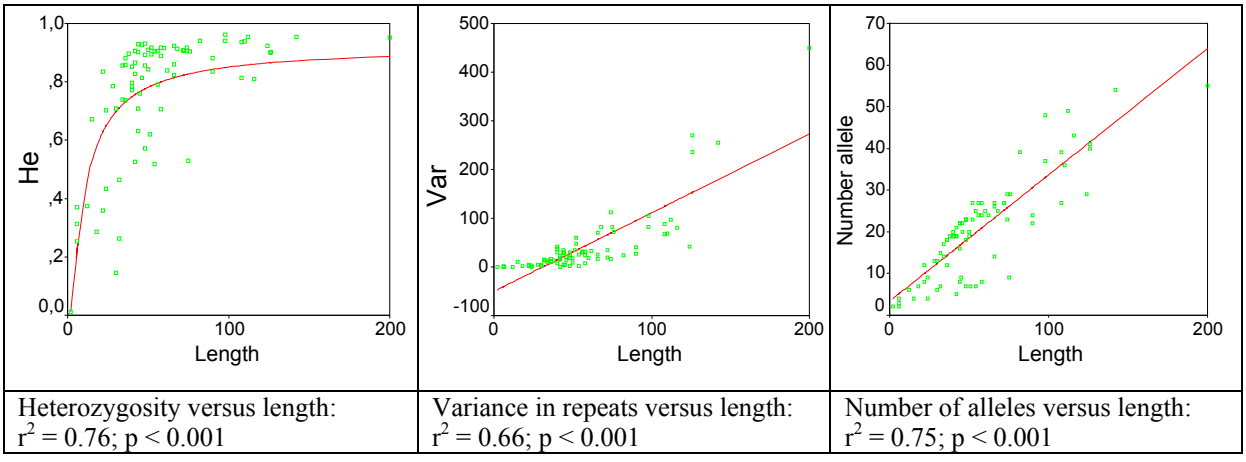
tetranucleotide repeats never contain more than 12 repeat units. The characteristics of the repeats strongly affect the gene diversity of these loci. The highest diversity is found in the dinucleotide repeats and the lowest diversity in the interrupted repeats (Table 3). Kruskall-Wallace tests revealed highly significant differences of variability depending on the repeat type.

**Table 3**: Results of the Kruskall-Wallace Test for differences between heterozygosity, variance in repeat units, and alleles for the different marker classes pure dinucleotides, interrupted repeats and pure tri- and tetranucleotide repeats.

| | Repeat type | N | Mean | Chi-Quadrat | df | Significance |
|---|---|---|---|---|---|---|
| **Hetero-zygosity** | Pure Dinucleotides | 48 | 0.87 | 25.844 | 2 | 0.000*** |
| | Interrupted repeats | 21 | 0.57 | | | |
| | Pure Tri-, Tetranucleotides | 12 | 0.67 | | | |
| **Variance repeat units** | Pure Dinucleotides | 48 | 51.37 | 12.485 | 2 | 0.002** |
| | Interrupted repeats | 21 | 19.06 | | | |
| | Pure Tri-, Tetranucleotides | 12 | 42.06 | | | |
| **Number of alleles** | Pure Dinucleotides | 48 | 25.29 | 24.549 | 2 | 0.000*** |
| | Interrupted repeats | 21 | 12.57 | | | |
| | Pure Tri-, Tetranucleotides | 12 | 15.92 | | | |

Kruskall-Wallace Test * level of Significance

Strong correlation is apparent in the total length of the repeat region and the variability indices as shown in Figure 1. Longer repeat stretches correlate significantly with higher variability.




| Heterozygosity versus length: $r^2 = 0.76$; $p < 0.001$ | Variance in repeats versus length: $r^2 = 0.66$; $p < 0.001$ | Number of alleles versus length: $r^2 = 0.75$; $p < 0.001$ |
|---|---|---|

**Figure 1:** Correlation of the total length of the repeat region with the different gene diversity estimates.

### lnRV and lnRH analysis

The lnRV and lnRH statistics were applied to the "multi-locus candidate gene approach" and the "random microsatellite screen". Since monomorphic loci cannot be included into the calculations of the ratios one artificial allele with one repeat unit difference was added to

monomorphic loci in the following population: within locus P1 (France, Germany, USA), P26 (Cameroon, Germany, USA), P62 (France), P66 (France, USA), P71 (Czech Republic, France, Germany, Kazakhstan, USA), P92 (Cameroon, France, Germany, USA), P94 (France, USA), P103 (Germany, USA), P104 (Cameroon), P154 (Germany), P157 (France, Germany, USA).

lnRV and lnRH calculations were done on different levels:

a) pooling the *Mus musculus* type and *Mus domesticus* type populations; b) population pairwise comparisons by comparing each population against all others, and c) summary of all values of the population pairwise comparisons. For the "random microsatellite screen" only a) and c) were performed. The resulting distributions were checked for normal distribution and the characteristics such as mean, standard deviation and the 5% and 95% percentils were calculated (Table 4).

**Table 4:** Characteristics of all pairwise lnRV and lnRH distributions. Table a) shows the summaries of all combined pairwise comparisons and the results of the pooled analysis for the "random microsatellite screen" and the "multi-locus candidate gene approach"; b) shows the results of all comparisons of the "multi-locus candidate gene approach" of "within lineages" comparisons only comparing *Mus domesticus* to *Mus domesticus* populations or *Mus musculus* to *Mus musculus*; c) shows the "between lineage" comparisons: *Mus domesticus* populations were compared to *Mus musculus* populations.

| a) | Random microsatellite screen | | | | Multi-locus candidate gene approach | | | |
|---|---|---|---|---|---|---|---|---|
| | Combined pairwise comparisons | | dom-mus pooled | | Combined pairwise comparisons | | dom-mus pooled | |
| **N** | 626 | | 106 | | 1195 | | 81 | |
| | lnRV | LnRH | LnRV | lnRH | lnRV | LnRH | lnRV | lnRH |
| **Mean** | 0.33 | 0.31 | -0.33 | 0.28 | -0.04 | -0.39 | -0.85 | -0.98 |
| **Stdev** | 1.29 | 1.56 | 1.15 | 1.23 | 2.19 | 1.73 | 1.80 | 1.77 |
| **Lower 5% border** | -1.76 | -2.38 | -2.13 | -1.75 | -3.67 | -3.41 | -5.40 | -4.94 |
| **Upper 95% border** | 2.53 | 2.78 | 1.79 | 2.26 | 2.99 | 2.31 | 1.59 | 1.79 |
| **KS_Z-value** | 0.97 | 0.69 | 0.61 | 0.66 | 1.93 | 1.02 | 1.03 | 0.81 |
| **Significance** | 0.31 | 0.72 | 0.85 | 0.78 | 0*** | 0.25 | 0.24 | 0.53 |

*** = highly significant deviation from normal distribution

| b) | Multi-locus candidate gene approach: pairwise population comparison within lineages | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Cam-Fra | | Cam-Ger | | Cam-USA | | Fra-Ger | | Fra-USA | | Ger-USA | | CZ-Kaz | |
| **N** | 81 | | 81 | | 79 | | 81 | | 79 | | 79 | | 79 | |
| | lnRV | lnRH | lnRV | lnRH | lnRV | LnRH | lnRV | lnRH | lnRV | LnRH | lnRV | lnRH | lnRV | lnRH |
| **Mean** | -0.02 | -1.13 | -0.11 | -1.28 | -0.04 | -1.08 | -0.09 | -0.15 | -0.04 | 0 | 0.06 | 0.22 | -0.07 | 0.15 |
| **Stdev** | 1.92 | 1.52 | 1.75 | 1.50 | 2.19 | 1.46 | 1.64 | 0.96 | 1.72 | 1.41 | 1.64 | 1.37 | 1.47 | 1.46 |
| **Lower 5% border** | -2.62 | -3.60 | -2.90 | -3.80 | -3.91 | -3.61 | -2.29 | -1.99 | -3.45 | -2.05 | -2.96 | -1.94 | -2.44 | -1.85 |
| **Upper 95% border** | 2.77 | 1.56 | 2.96 | 1.36 | 3.73 | 0.91 | 1.56 | 1.41 | 2.17 | 2.59 | 1.94 | 2.90 | 1.88 | 2.57 |
| **KS_Z-value** | 0.79 | 0.59 | 0.94 | 0.34 | 0.90 | 0.49 | 0.77 | 0.62 | 0.79 | 0.56 | 1.04 | 0.54 | 1.09 | 1.0 |
| **Significance** | 0.57 | 0.88 | 0.34 | 1 | 0.40 | 0.97 | 0.60 | 0.84 | 0.56 | 0.91 | 0.23 | 0.94 | 0.18 | 0.26 |

| c) | Multi-locus candidate gene approach: pairwise population comparison between lineages | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Cam-CZ | | Cam-Kaz | | CZ-Fra | | CZ-Ger | | CZ-USA | | Fra-Kaz | | Ger-Kaz | | Kaz-USA | |
| N | 79 | | 81 | | 79 | | 79 | | 77 | | 81 | | 81 | | 79 | |
| | lnRV | lnRH | lnRV | lnRH | lnRV | lnRH | lnRV | lnRH | lnRV | lnRH | lnRV | lnRH | lnRV | lnRH | lnRV | lnRH |
| Mean | -1.05 | -1.57 | -1.08 | -1.66 | 0.98 | 0.41 | 0.89 | 0.26 | 0.97 | 0.52 | -1.06 | -0.53 | -0.97 | -0.37 | 1.07 | 0.63 |
| Stdev | 1.98 | 1.53 | 2.27 | 1.85 | 2.23 | 1.65 | 1.98 | 1.61 | 2.43 | 1.45 | 2.51 | 1.98 | 2.35 | 1.90 | 2.59 | 1.71 |
| Lower 5% border | -5.26 | -4.43 | -5.81 | -4.17 | -2.07 | -1.63 | -1.85 | -2.17 | -2.52 | -1.66 | -7.59 | -4.52 | -6.81 | -4.03 | -2.11 | -2.34 |
| Upper 95% border | 1.83 | 1.39 | 1.91 | 1.96 | 6.49 | 4.42 | 5.42 | 4.07 | 5.88 | 2.77 | 2.12 | 2.55 | 1.82 | 2.96 | 7.62 | 3.27 |
| KS Z-value | 0.72 | 0.42 | 0.82 | 1.01 | 0.77 | 1.17 | 0.87 | 0.85 | 1.22 | 0.55 | 1.02 | 0.87 | 1.06 | 1.13 | 0.54 | 0.67 |
| Significance | 0.67 | 0.99 | 0.51 | 0.25 | 0.60 | 0.13 | 0.44 | 0.47 | 0.10 | 0.92 | 0.25 | 0.44 | 0.21 | 0.16 | 0.93 | 0.77 |

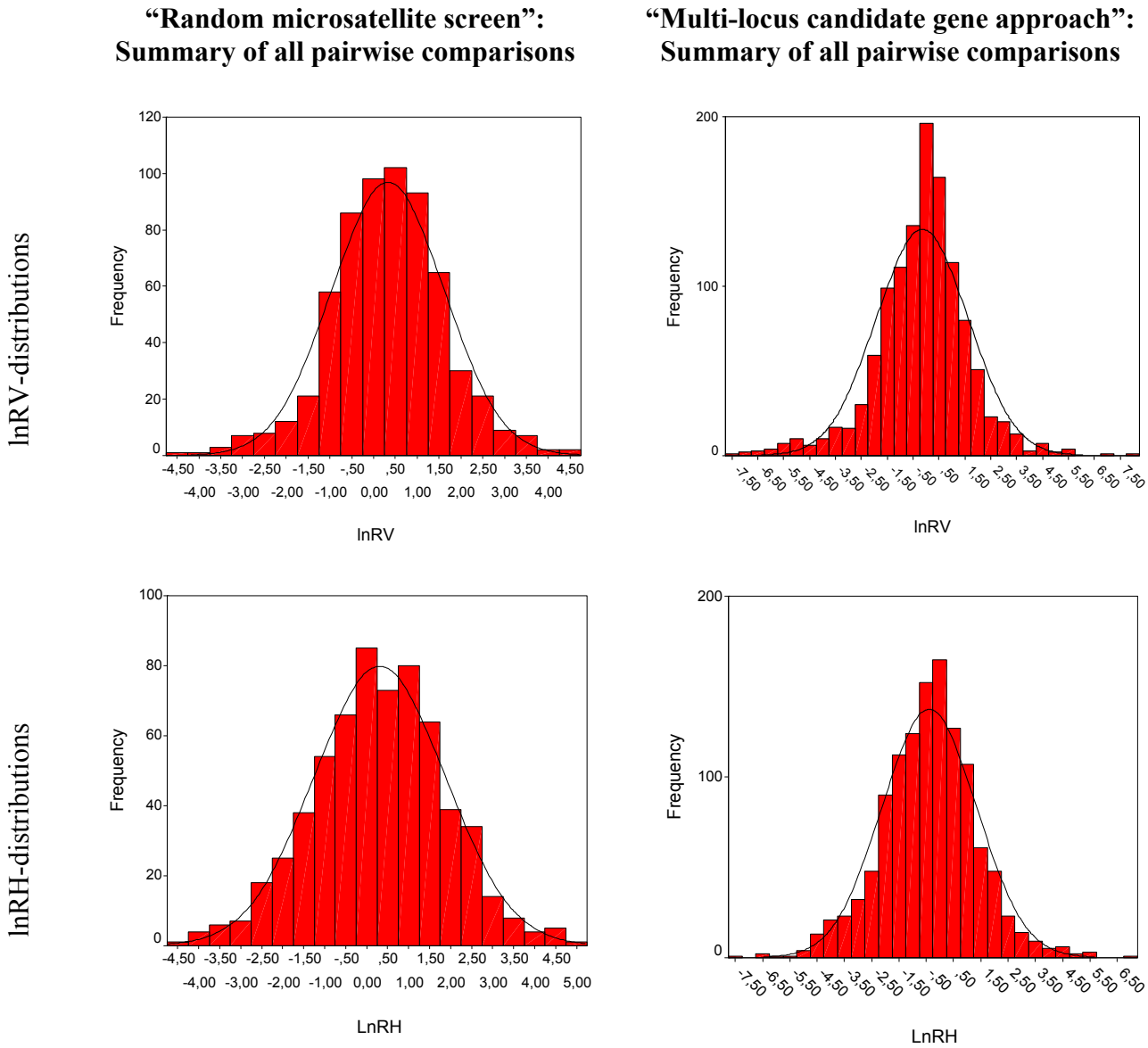



**Figure 2:** Distribution of lnRV and lnRH values calculated for the "random microsatellite screen" and "multi-locus candidate gene approach". The distributions represent the summary of all pairwise comparisons between the separate populations.

All lnRV and lnRH distributions resulted in normal distributions except for the lnRV distribution which consists of all pairwise results from the "multi-locus candidate gene

approach". The deviation from normal distribution resulted from too many extreme values at the edges of the distribution which is shown in Figure 2. These extreme lnRV values correspond to high differences in variability at one locus in two populations. Therefore, these loci can be regarded as potential sweep loci. All loci below the 5% or above the 95% threshold of the normal distribution can be regarded as significantly different because they exceed the threshold which is characteristic for a normal distribution. The comparisons were performed on different levels by comparing populations of different divergence times. The thresholds for each pairwise distribution differ from each other depending on which populations were compared.

In general, the distributions of the "random microsatellite screen" is narrower compared to the distribution of the "multi-locus candidate gene approach". Even the combined distribution of all pairwise lnRV calculations is normally distributed indicating that almost no extreme differences exist between the loci. Therefore most of the observed extreme differences can probably be considered as normal statistical deviations. The "multi-locus candidate gene approach" was analysed in more detail because the distribution of all combined pairwise comparisons deviated from a normal distribution while all pairwise comparisons themselves were normally distributed.

It is difficult to set a general threshold for the identification of a sweep locus as the threshold values are influenced by the general characteristics of the populations. This can be illustrated by analysing the pairwise comparisons including the Cameroon population: the Cameroon population generally exhibits a lower gene diversity. As a consequence, all comparisons with this population in the numerator are slightly skewed to more negative values. Comparing lnRV and lnRH values from the Cameroon pairwise comparisons against the summary of all pairwise distributions of the "multi-locus candidate gene approach" lead to wrong results because lnRV values below -3.67 which are well within the confidence interval of most of the Cameroon pairwise distributions lie beyond the 5% threshold of the combined distribution (Table 4). Therefore, the lower gene diversity of the Cameroon population has to be taken into account.

Generally, the distribution resulting from "between lineage" comparisons (comparing a *Mus musculus* population against a *Mus domesticus* population) results in a broader distribution. This was inferred by comparing the standard deviation for the "between lineage" comparisons

to the "within lineage" comparisons (comparing *Mus domesticus* with *Mus domesticus* populations and *Mus musculus* with *Mus musculus* populations) (T = -4.477; df = 13; p = 0.001 for LnRV; T = -3.337; df = 13; p = 0.005 for lnRH). Depending on the level of divergence time different thresholds must be applied. Accordingly, all loci with extreme differences in either the lnRV or the lnRH distributions are listed in Table 5.

**Table 5:** List of all loci exhibiting extreme values in any lnRH and lnRV calculation. Populations mentioned in the cells carry the lower variance or heterozygosity; populations in bold and curved are significant in the single pairwise comparison and in the summarized pairwise comparison.

| | | | | **within lineages pairwise population comparisons** | | | | | | | **between lineages pairwise population comparisons** | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Level of population comparison | | summarised pairwise distribution | dom-mus pooled | Cam-Fra | Cam-Ger | Cam-USA | Fra-Ger | Fra-USA | Ger-USA | CZ-Kaz | Cam-CZ | Cam-Kaz | CZ-Fra | CZ-Ger | CZ-USA | Fra-Kaz | Ger-Kaz | Kaz-USA |
| P1 | lnRH | 4 | | ***Fra*** | Ger | | | | | | | | | | | ***Fra*** | ***Ger*** | USA |
| | lnRV | 1 | | | | | | | | | | | | | | | | |
| P8 | lnRH | 1 | | | | ***Cam*** | | | | | | | | | | | | |
| P12 | lnRH | 4 | | | | | | | | CZ | | | | | | | | USA |
| P14 | lnRH | 2 | | | | | | | | CZ | | | | | | | | |
| P19 | lnRH | | | | Ger | | | | | | | | | | | | | |
| P25 | lnRH | 2 | | Fra | Ger | ***USA*** | | | | | CZ | ***Kaz*** | | | | | | |
| | lnRV | 1 | 1 | | | | | | | Kaz | | Kaz | | | | Kaz | ***Kaz*** | Kaz |
| P26 | lnRH | 8 | 1 | | | | | | | | ***Cam*** | | ***Fra*** | ***Ger*** | ***USA*** | | | |
| | lnRV | 5 | 1 | | | | | | | | | | | | | | | |
| P29 | lnRH | 2 | | | | | | | | | | | | | | | | |
| | lnRV | 1 | | | | | | | | | | | | | | | | |
| P31 | lnRH | 1 | | | | | | | | | | | | | | | | |
| | lnRV | 1 | | | | | | | | ***CZ*** | | | | | | | | Kaz |
| P32 | lnRH | | | | | | | | USA | | | | | | | | | |
| | lnRV | | | | | | Fra | | | | | | | | | | | |
| P34 | lnRH | | | | | | | | | | | | CZ | | | | | |
| P35 | lnRH | 1 | | | | | | ***USA*** | | | | | | | | | | |
| P36 | lnRV | 2 | | | | | | | | | | | | | | | | |
| P37 | lnRH | 1 | | | ***Cam*** | | | | | | | | | | | | | |
| | lnRV | | | | | | | | | | | | CZ | CZ | | | | |
| P38 | lnRH | 1 | | | | | | ***USA*** | | | | | | | | | | |
| | lnRV | 3 | | | | | | ***USA*** | USA | | | | | | | | | |
| P42 | lnRH | 2 | | | | | | | | | | | | | CZ | ***Kaz*** | ***Kaz*** | Kaz |
| P44 | lnRH | | | | | | | | | CZ | | | | | | | | |
| | lnRV | | | | | | | | | CZ | CZ | | CZ | CZ | CZ | | | |
| P45 | lnRH | 4 | | | | | | ***USA*** | ***USA*** | | | | CZ | CZ | | ***Kaz*** | ***Kaz*** | |
| | lnRV | | | | | | | | | | | Kaz | | | | Kaz | Kaz | |
| P49 | lnRH | 1 | | | | | | | | | | | | | | | | |
| P51 | lnRH | 6 | | | | | | | ***USA*** | | | | | | | | | |
| P55 | lnRV | 4 | | | | ***Cam*** | Ger | | ***Ger*** | | | | | | | | | |
| P57 | lnRV | | | Cam | | | | | | | | | | | | | | |
| P61 | lnRH | 3 | | ***Cam*** | | | | | | | ***Cam*** | ***Cam*** | | | | | | |
| | lnRV | 1 | | | Cam | ***Cam*** | | | | | | | | | CZ | | | |
| P62 | lnRH | 5 | | | | | Fra | ***Fra*** | Ger | | | | | | | ***Fra*** | | |
| P63 | lnRV | | 1 | | | | | | | | | | | | | | | |
| P66 | lnRH | 5 | | | | | ***Fra*** | | | | ***Cam*** | | ***Fra*** | | ***USA*** | | | |
| | lnRV | 7 | | | ***Cam*** | | ***Fra*** | | ***USA*** | | | | | | | | | |

| Level of population comparison | | summarised pairwise distribution | dom-mus | Cam-Fra | Cam-Ger | Cam-USA | Fra-Ger | Fra-USA | Ger-USA | CZ-Kaz | Cam-CZ | Cam-Kaz | CZ-Fra | CZ-Ger | CZ-USA | Fra-Kaz | Ger-Kaz | Kaz-USA |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| P67 | lnRH | 5 | 1 | | | | | | | ***Kaz*** | CZ | ***Kaz*** | | | | ***Kaz*** | ***Ger*** | ***Kaz*** |
| P72 | lnRH | 1 | 1 | | | | | | | | ***CZ*** | | CZ | CZ | CZ | | | |
| | lnRV | 4 | | | | | | | Ger | ***CZ*** | CZ | | CZ | CZ | ***CZ*** | | | |
| P73 | lnRH | 2 | | | | | | | | | | | | | | | | |
| | lnRV | 2 | | | | | Fra | Fra | | | | | | | | | | |
| P76 | lnRH | 4 | | ***Cam*** | ***Cam*** | ***Cam*** | | | | | | | | | | | | |
| | lnRV | 5 | | ***Cam*** | ***Cam*** | ***Cam*** | Ger | | | | ***Cam*** | ***Cam*** | | | | | | |
| P83 | lnRH | 1 | | | | | | | | | | Kaz | | | | | | |
| P85 | lnRH | 4 | | | | | | ***USA*** | ***USA*** | | | | | | | | | |
| P92 | lnRH | 6 | 1 | | | | | Fra | Ger | | ***Cam*** | ***Cam*** | ***Fra*** | ***Ger*** | | | ***Ger*** | |
| | lnRV | 7 | | ***Fra*** | ***Ger*** | | | | | | CZ | Kaz | | | | | | |
| P94 | lnRH | | 1 | | | | | | | | | | | | | | | |
| | lnRV | 8 | | ***Fra*** | ***Ger*** | ***USA*** | | | | | CZ | | | | | | | |
| P103 | lnRH | 1 | | | | | Ger | | | | | | | | | | Ger | |
| | lnRV | 7 | | ***Cam*** | | | ***Ger*** | ***USA*** | | ***CZ*** | | | | | | | | |
| P104 | lnRH | 9 | | ***Fra*** | ***Ger*** | ***USA*** | | | | | | ***Cam*** | ***Fra*** | ***Ger*** | ***USA*** | ***Fra*** | ***Ger*** | ***USA*** |
| | lnRV | 6 | 1 | Fra | | | | | | | ***Cam*** | | ***Fra*** | ***Ger*** | | ***Fra*** | ***Ger*** | |
| P106 | lnRH | 1 | | | | | | | | | | | CZ | | | | | |
| P108 | lnRV | 2 | | | | | | | | Kaz | | | | | | | | |
| P116 | lnRH | 1 | | | | | | | | | | | | | | | | |
| P118 | lnRH | | | | | | Ger | | | | | | | | | | | |
| P120 | lnRH | 2 | | Fra | | | | | | | | | | | | | | |
| P121 | lnRH | 2 | | Fra | | | Fra | | | | CZ | ***Kaz*** | | CZ | | | ***Kaz*** | Kaz |
| | lnRV | 1 | | | | | | | | ***Kaz*** | | Kaz | | | | | Kaz | Kaz |
| P127 | lnRH | 5 | | ***Cam*** | ***Cam*** | ***Cam*** | | | | | | | | | | | | |
| P129 | lnRH | 1 | 1 | | | | | | | | | | | | | | | |
| | lnRV | | 1 | | | | | | | | | | | | | Kaz | Kaz | Kaz |
| P140 | lnRH | | | | Cam | | | | | | | | | | | | | |
| P141 | lnRH | 5 | | | | | | | | ***CZ*** | | ***Cam*** | | | CZ | ***Fra*** | ***Ger*** | ***USA*** |
| | lnRV | 6 | | | | ***Cam*** | | | | ***CZ*** | | ***Cam*** | | | ***CZ*** | | | |
| P142 | lnRH | | | | | | | Ger | | | | | | | | | | |
| | lnRV | | | | | | | | | | | | | | | Cam | | |
| P149 | lnRV | | | | | | | | | | | | | CZ | | | | |
| P150 | lnRH | | | | | | Fra | | | | | | | | | | | |
| P153 | lnRH | | | | | | Ger | | | | | | | | | | | |
| P154 | lnRH | 7 | | | | | | ***Fra*** | ***Ger*** | | | | | ***Ger*** | | | | |
| | lnRV | 2 | | | | | | | | | | | | | | | | |
| P156 | lnRH | | | | | | | | | | | | | | | Kaz | | |
| P157 | lnRH | 3 | | | | | | Fra | Ger | | | | | | | | | |
| | lnRV | 3 | 1 | | | | | | | | | | | | | | | |
| P158 | lnRH | 3 | | | | | | | | | | | | | | | | |
| P159 | lnRV | | | | | | | Fra | | | | | | | | | | |
| P160 | lnRH | 1 | | | | | | | | | | | | | | | | ***USA*** |
| | lnRV | 6 | | | ***Ger*** | ***USA*** | Ger | USA | | | | | | | | | | |

In total 56 loci show extreme differences: 33 genes exhibit extreme lnRV values and 49 loci exhibit extreme lnRH values. Different levels of comparisons result in different identifications of sweep loci. By pooling the *Mus musculus* populations and the *Mus domesticus* populations population specific differences disappear. For example locus P76 shows an extreme reduction in polymorphism in the Cameroon population in five lnRV and three lnRH comparisons. These differences disappear in the comparison of the pooled *musculus* and *domesticus* samples (Table 5). The loci P92 and P94 are rather monomorphic and contain a maximum of

three alleles leading even to contradictive results for the lnRV and lnRH calculations and are also considered as artefacts.

Extreme values which just occurr in one or two comparisons are regarded as normal statistical deviations and accordingly 32 loci are excluded from the sweep locus list above. Additionally, to apply very conservative assumptions loci are considered as potential sweep loci only if they reveal extreme values in at least three comparisons in the same population. Applying these characteristics to Table 5, I suggest that the following loci can be regarded as potential sweep loci: the Interferon-induced protein with tetratricopeptide repeats1 (Ifit1) (P25), the Interferon alpha/beta receptor gene (P26), the ß-Defensin 6 gene (P104), the nasal embryonic LHRH factor (P129) and the randomly chosen microsatellite locus (P67) seem to contain sweep loci in the *domesticus-musculus* comparison. P26 and P104 are rather monomorphic in the *Mus domesticus* lineage while P25 and P67 and P129 lost variability in *Mus musculus*. Other loci show reduction in polymorphism in multiple comparisons in just one population: locus P61 within the Beta-globin complex, the Axonemal dynein heavy chain 8 Dnahc8 gene (P76) and the MATH4B gene (P127) show reduction in the Cameroon population. The microsatellite P44 within the Ig gamma-chains, and the locus P72 within the SRY-box containing gene 5 (Sox5) have potentially swept within the Czech population. Kazakhstan mice show reduction in polymorphism at the following loci: P42 within the Plasma selenoprotein P (SELP) gene, P45 within the Pro-alpha1 (II) collagen chain gene and P121 which is located in the genomic sequences containing olfactory genes. P141 within the Cyclic nucleotide-gated olfactory channel protein gene exhibits extreme values in nearly all populations compared to the Kazakhstan population.
For all those loci the allele distributions were analysed and the distributions are shown in figures 3a - 3d.

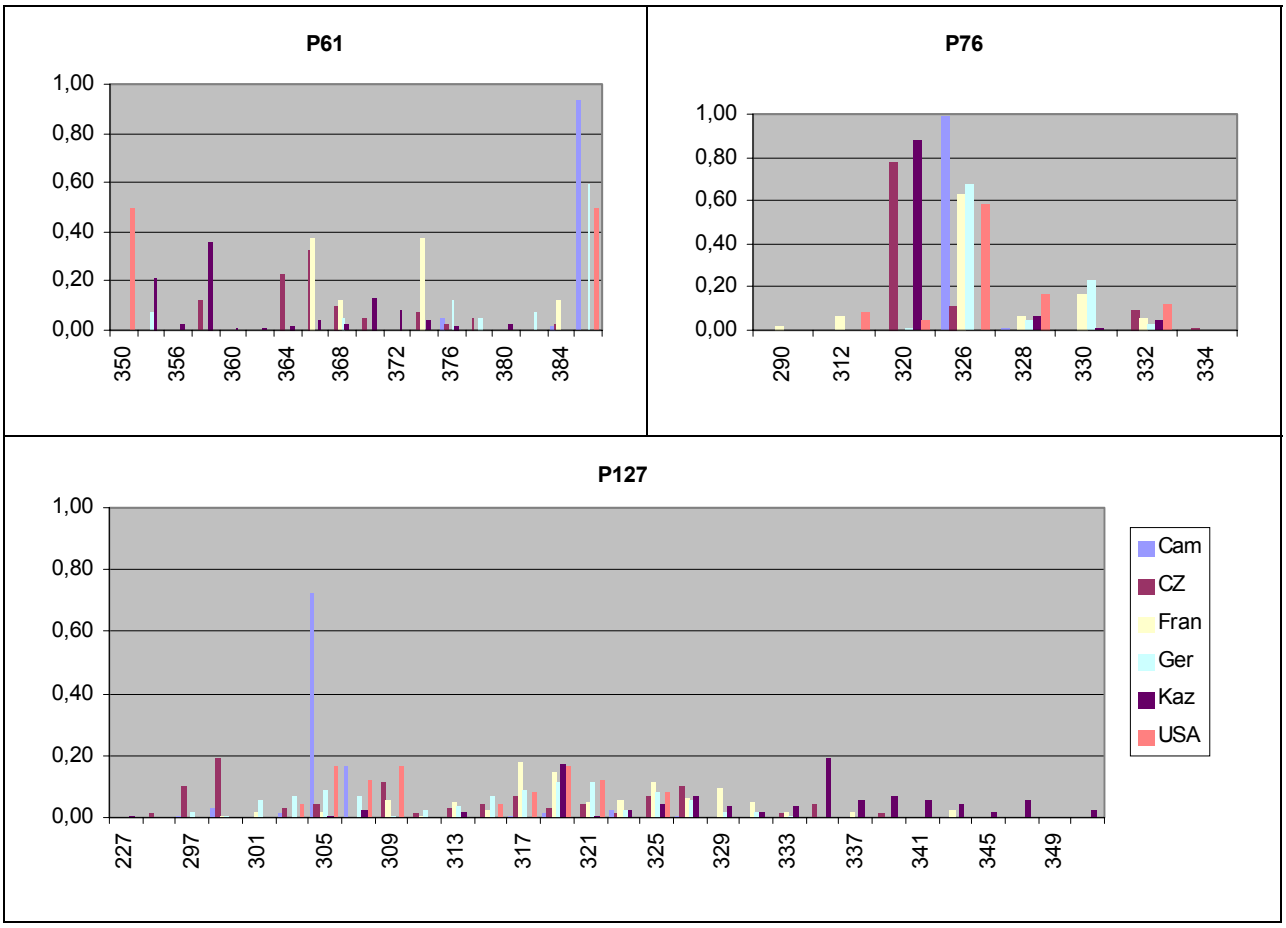



**Figure 3a:** Allele distributions of the sweep loci within Cameroon mice.

The locus P61 and P76 are nearly monomorphic in the Cameroon population, while they are relatively polymorphic in the other populations. The possibility that they have lost variability due to a locus specific selective event is well supported by the lnRV and lnRH calculations. The locus P127 also shows a striking pattern with one very frequent allele in the Cameroon population but only the lnRH calculation recognizes this locus as a potential sweep locus.

The lnRV and lnRH calculations are differentially influenced by the different parameters: the lnRV distribution is strongly dependent on the size distribution of the alleles. Interrupted allele distributions can result in high variances although the general gene diversity is rather low. This seems to be the case at locus P127: although the locus seems to exhibit a low gene diversity and shows one major allele with 305 bp in the Cameroon population some rare alleles of 297 and 327 bp are found and increase the variance at this locus substantially. The lnRH calculation, in contrast, is less dependent on the allele distribution and just takes into account the frequencies of the alleles which results in extreme lnRH values for locus P127.

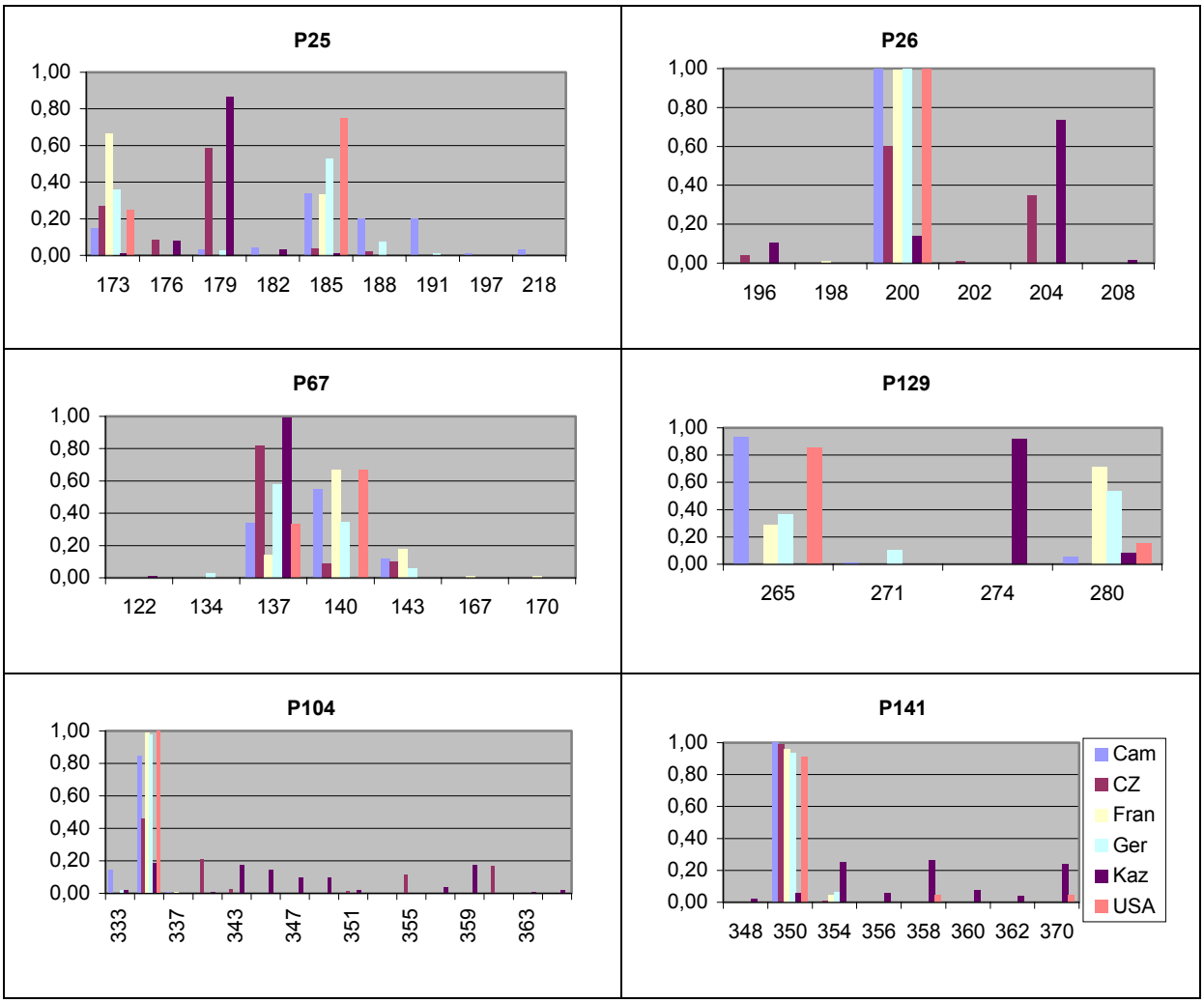



**Figure 3b:** Allele distributions of the *domesticus-musculus* sweep loci.

Figure 3b shows the sweep loci which seem to be lineage specific and occur either in all *Mus domesticus* populations or all *Mus musculus* populations. Only locus P141 exhibits extremely high polymorphism in the Kazakhstan population while all other populations have swept. Interestingly all "lineage specific sweep loci" show a rather low number of total alleles per locus. Locus P129 shows a very low diversity and the sweep was just recognized by the lnRV statistics. The result therefore seems to be an effect of the gap in the allele distribution for the *domesticus* clade; all populations have one major allele for this locus and the extreme value for P129 seems to be an artefact. Generally, loci with a low number of alleles tend to exhibit extreme lnRV and lnRH values easily because differences in the allele distributions directly result in higher differences of heterozygosity and variance of repeats.

Therefore, only loci P26, P104 and P141 exhibit the expected allele distribution for a selective sweep event with an extreme reduction of diversity in one clade and substantial variation in the other populations.

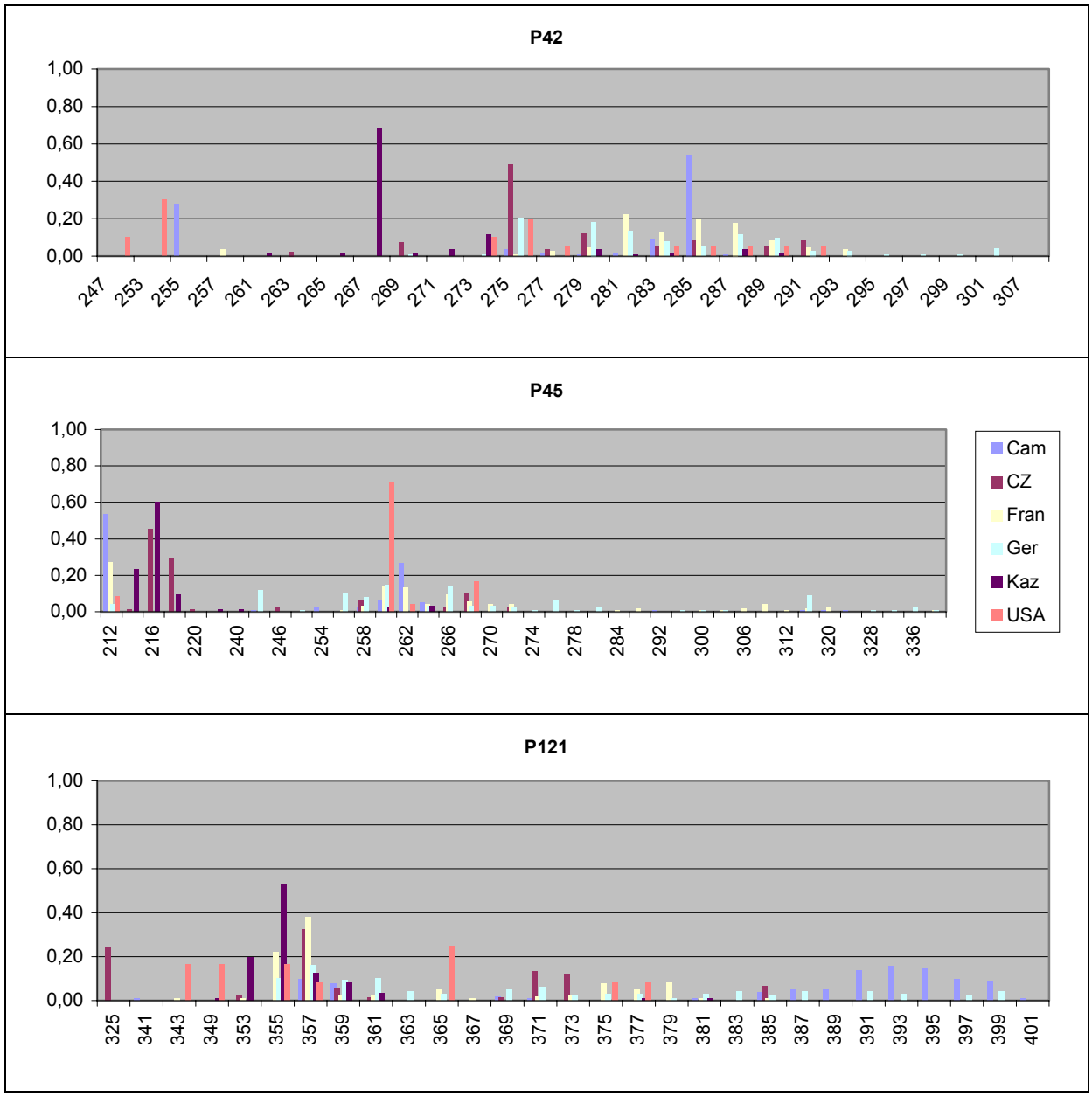



**Figure 3c:** Allele distributions of the sweep loci recognized in the Kazakhstan animals.

The allele distributions for potential sweeps in the Kazakhstan population follow the expectation for selective sweep events. P42 is just recognized by the lnRH statistics but shows extreme reduction of polymorphism within the Kazakhstan population compared to the other rather polymorphic populations. P45 exhibits extreme lnRV and lnRH values for Kazakhstan but also for the USA animals indicating two possible independent events at this locus. P121 is also significant in the lnRV and lnRH values and shows the expected sweep pattern.

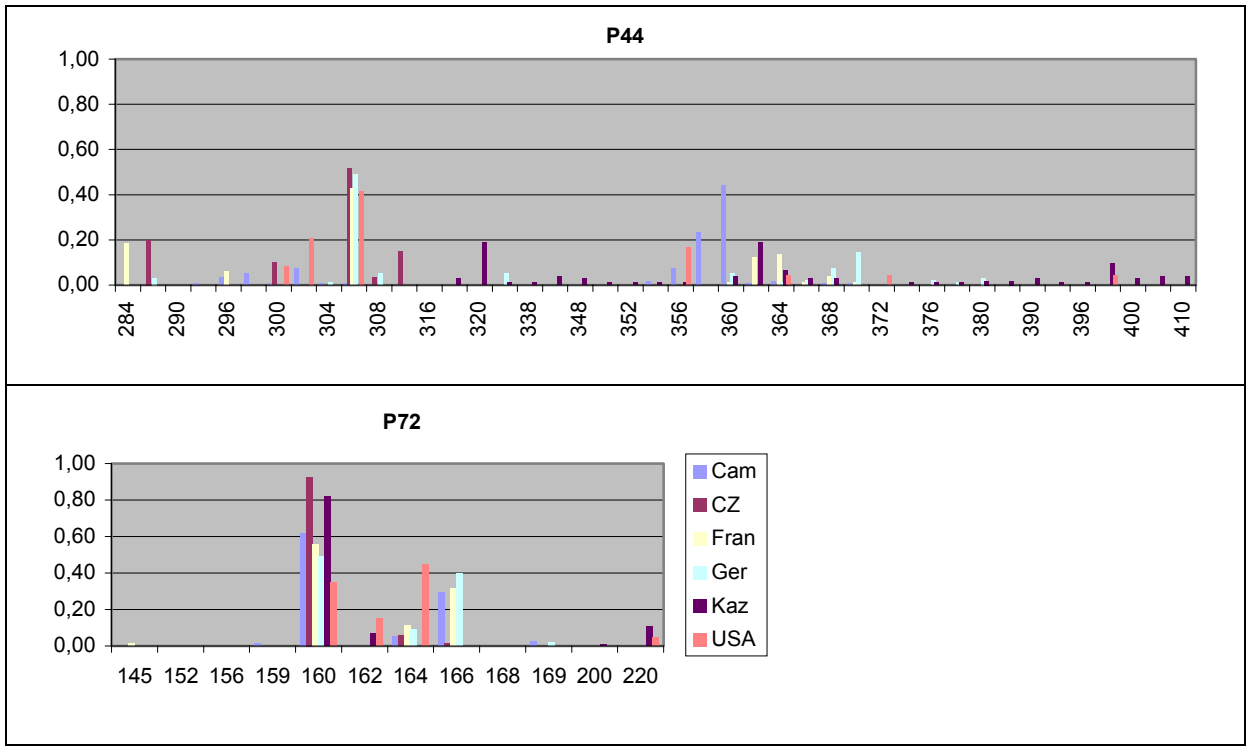



**Figure 3d:** Allele distributions of the sweep loci recognized in the population from the Czech Republic.

The selective sweeps for the Czech Republic are highly debatable because, in all cases, the allele with the highest frequency in the Czech Republic is also the allele with the highest frequency in the other populations.

In general, extreme lnRV or lnRH values are at the first indicators for possible sweep events but the sweep event has to be confirmed by further methods such as analysis of the allele distributions. Summarising the analysis of the calculations and the inspections of the allele frequency distributions the following loci fulfil the criteria of selective sweep events: three genes known to be differentially expressed carrying the microsatellite loci P42, P45 and P61; three loci within olfactory genes carrying the microsatellite loci P121, P127 and P141; one saliva gene containing locus P104; one interferon related gene containing locus P26 and one speciation gene carrying P76.

The “random microsatellite screen” was not analysed in such detail. The analysis of the summary of all pairwise comparisons reveals that from 106 microsatellite loci, 32 loci show lnRV-values exceeding the 95%-threshold of the lnRV distribution; nine of those loci occur in three or more comparisons. For the lnRH distribution 33 loci exhibit extreme values exceeding the 95%-threshold; eight of them in at least three comparisons. Extreme lnRV and

lnRH values occur simultaneously in eleven comparisons and seven of these loci exhibit extreme values in more than four lnRV and lnRH comparisons.

**Correlation between repeat type and extreme lnRV and lnRH values**

Since the variance in repeat units and the heterozygosity is dependent on the repeat type, I tested whether the repeat types also influence the occurrences of extreme lnRV or lnRH values. In summary, extreme lnRV and lnRH values are detected in 21 dinucleotide repeat units, 14 interrupted repeat units, and 6 tri- or tetranucleotide repeat units. However, Chi-Square tests reveal no significant differences in the occurrence of extreme or non-extreme values depending on the repeat type (data not shown). Generally, more loci exhibit extreme lnRH than lnRV values independent of the repeat type.

Significant differences are found by comparing the number of extreme comparisons per locus in dependence of the repeat type: mean number of extreme lnRV values are lowest in dinucleotide repeat markers and highest in tri- or tetranucleotide repeat markers. The lnRH values are less dependent on the repeat type, only the comparison of the dinucleotide repeats with the interrupted repeats results in significant different results (results of significance analysis are not shown).

**Table 5:** Occurrence of extreme values in di-, tri-, tetra and interrupted repeats.

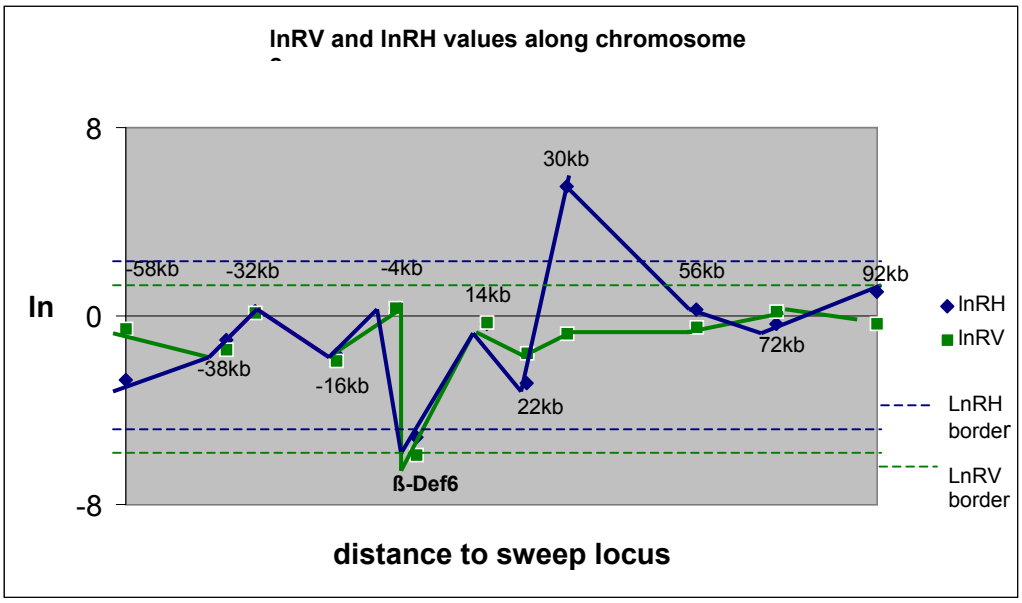
| | **Number of loci** | | **lnRV** | | **LnRH** | |
|---|---|---|---|---|---|---|
| | non-extreme Com-parisons | Extreme Com-parisons | number of loci | Mean number of extreme comparisons per locus | number of loci | Mean number of extreme comparisons per locus |
| **Dinucleotid Repeats** | 27 | 21 | 14 | 2.14 | 22 | 1.64 |
| **Interrupted repeats** | 7 | 14 | 9 | 3.78 | 13 | 4.4 |
| **Tri-/Tetranu-cleotid Repeats** | 6 | 6 | 3 | 5.7 | 6 | 3.2 |

Interrupted repeats and tri- and tetranucleotide repeat markers correlate with lower gene diversities. As already mentioned allele distributions based on a smaller number of alleles are more likely to exhibit extreme lnRV or lnRH values as small changes in the allele distributions directly correlate with big differences in the variance of repeat units or the heterozygosity, which explains the higher number of extreme values for those repeats.

**Delimitation of the selective sweep window**

Another proof for a selective sweep event is the occurrence of a sweep window due to the hitchhiking effect. To identify the size of a sweep window eleven microsatellites around the ß-Defensin 6 gene were genotyped in a region of 150 kb around the sweep locus. Since my results suggest that the ß-Defensin 6 gene has swept in the *Mus domesticus* lineage samples were pooled for the two lineages and the lnRV and lnRH values for the flanking microsatellites were calculated. The results are presented in Figure 4.

**Figure 4:** lnRV and lnRH results of microsatellite loci flanking the ß-Defensin 6 locus. The dashed lines indicate the upper and lower border of the 95% confidence interval for the pooled lnRV and lnRH calculations.




The reduction of polymorphism at the ß-Defensin 6 locus is exceeding the 95%-threshold for the lnRV and lnRH calculations. Nevertheless, no reduction in polymorphism is visible in the flanking regions except 30 kb downstream of the ß-Defensin 6 locus a strong reduction of heterozygosity for *Mus musculus* occurs. Most of the flanking microsatellite loci consist of pure dinucleotide repeat units except for one tri- (maximum length of 24 bp) and one tetranucleotide (maximal length of 64 bp) repeat marker. All dinucleotide repeat markers have repeat regions of at least 32 bp. Therefore, those flanking microsatellite loci have a higher mutation rate compared to locus P104 which has a repeat length of 32 bp and shows interruptions within the repeat units.

Thus for the ß-Defensin 6 locus the sweep window is either very narrow or the flanking microsatellite loci have evolved a new polymorphic allele spectrum after the sweep event obscuring the traces of selection.

## Discussion

### Identification of selective sweep loci and their differentiation from false positive results

Genotyping multiple microsatellite loci is a valuable tool to identify regions within the genome which are affected by selection (Schlötterer et 1997, Payseur et al. 2002, Schlötterer 2002b). The aim of the present study was to detect selective sweeps in different populations of the house mouse species complex by genotyping 81 microsatellite loci directly linked to genes that may play a role in adaptations, and by a "random microsatellite screen" of 106 markers. By using the lnRV and lnRH statistics of Schlötterer (2002b), I was able to identify several loci which might have swept as indicated by extreme values. Particularly, in the summary of all pairwise population comparisons, the lnRV statistics showed a deviation from the normal distribution which was caused by a high amount of extreme values at the edges of the distribution. These extreme values could not be considered as normal statistical deviations.
In contrast to this, the "random microsatellite screen" exhibited no deviation from the Gaussian distribution indicating that the number of observed extreme values was lower and within the range of normal statistical deviations. Also for the lnRH calculations all distributions did not show significant deviations from normality.

A general problem of multi-locus analysis is to distinguish "real" sweep loci from false positive results since, generally, values exceeding 95% confidence interval at the lower and upper end of the distribution are assumed to be potential sweep loci. Therefore, all multi-locus screens try to verify their results by further methods. One important criteria is that a "real" sweep locus must exhibit significant results in multiple comparisons (Schlötterer 2002b); in other studies two different methods such as calculating Rst and lnRV values are used to confirm sweep loci (Kayser et al. 2003). Screening of markers in the flanking region to identify the size of the region that was subjected to the hitchhiking event are also applied by several authors (Kohn et al. 2000, Harr et al. 2002, Payseur et al. 2002, Nair et al. 2003). Sequencing of the suspicious locus is also promising to verify sweep events because different statistical analysis tools can be applied to sequence data (Harr et al. 2002).

The statistical schemes applied in the studies of Vigouroux et al. (2002) and Payseur et al. (2002) are based on neutral expectations of allele frequency distributions and the estimations are either derived from the stepwise mutation model (Payseur et al. 2002) or from the infinite

allele model. As previously discussed in chapter 1, the application of estimators based on the microsatellite mutation models to the markers used in my screen lead to false results in the genetic distance estimations and, therefore, I concluded that these models are not applicable to estimate neutral allele distributions for my markers. In contrast, the lnRV and lnRH calculations are not based on any modelling parameters or apriori assumptions thus reducing the likelihood of false estimations. Multiple simulations proved that the lnRV statistic is robust to differences in mutation rates, different population sizes and demographic deviations (Schlötterer 2002b) which are, indeed, strong arguments for the application of this approach.

To identify sweep loci in this screen I applied the lnRV and lnRH calculations and verified the results by the analysis of multiple pairwise comparisons of each population against all the others and by inspection of the allele frequency distributions of promising loci. For the locus P104 I also investigated the variability of the flanking regions to identify the size of the sweep window.

Identifying multiple extreme comparisons for one population is a problematic issue because threshold values of the different comparisons vary substantially. One general outcome of the population pairwise comparisons was that the pairwise comparisons either within the *Mus musculus* or the *Mus domesticus* lineages result in narrower normal distributions with a lower standard deviation than the between lineage comparisons between a *Mus musculus* population against a *Mus domesticus* population. This is probably a result of neutral divergence between the lineages. Therefore, sweep loci that are different between *Mus musculus* and *Mus domesticus* must exhibit very extreme differences to exceed the already high "neutral" divergence rate between the taxa.

By identifying sweeps the levels of different divergence times have to be taken into account: within the sampled populations I have the possibility to study lineage specific sweeps by comparing pooled *musculus* and pooled *domesticus* samples against each other. These lineage specific comparisons can be verified by looking for extreme values in the pairwise population comparisons: as expected my results prove that extreme loci in the *musculus-domesticus* comparison also showed extreme values in the pairwise comparisons between the single populations except for locus P63 (results are listed in Table 5).

Generally, loci that showed extreme values in the *musculus-domesticus* comparisons were loci which consist of a relatively low number of alleles. Two explanations exist for this phenomenon: on the one hand loci with a low number of alleles are prone to exhibit extreme

values just because slight differences in the allele distribution have major effects on the diversity indices. For the loci P129, P67 and P63 (distribution not shown) this seemed to be the case because none of these loci exhibit a typical sweep with one major allele in one population and a distinct allele distribution compared to the other populations. On the other hand loci with a shorter repeat length also have a lower mutation rate (Schug et al. 1997) and are better suited to retain older sweeps such as sweeps after the divergence of the *Mus musculus* and the *Mus domesticus* lineages. This was probably the case for locus P104.
Except for locus P129 all significant microsatellite loci for the *musculus-domesticus* comparison consist of non-dinucleotide repeat units which correlate with lower diversity and a lower mutation rate (Ellegren 2000, Balloux and Lougon-Moulin 2002 and references therein). Thus, these loci are suitable to detect older sweep events.

On a different divergence level the sampled populations offer the opportunity to study population specific sweeps. Populations within *Mus domesticus* or within *Mus musculus* are more similar to each other and show narrower distributions of the standard deviation. Therefore, values which are extreme for the within lineage comparisons would be obscured by the between lineage distribution. These pairwise population comparisons also take demographic parameters into account: the Cameroon population showed the lowest gene diversity in all comparisons, therefore, by calculating the lnRV or lnRH values with Cameroon in the numerator the values will always be slightly biased to negative values shifting the whole distribution to negative values. Extreme values indicative for a Cameroon sweep, therefore, have to be very negative to exceed the 95% threshold at the lower end of the distribution. Accordingly, inferring sweeps for Cameroon by just using the threshold of the combined pairwise distribution would result in many “false” positives which show extreme values due to the general low gene diversity of the Cameroon population.
Hence, to apply very conservative estimates the extended criteria for sweep detection based on pairwise population comparisons was that at least three pairwise comparisons should show extreme reduction in polymorphism in the same population. By doing so, I detected 9 population specific sweep loci.
After analysing the allele distributions the Czech specific sweep loci were considered as “false” positives because the major allele of the distribution was also a major allele within the other populations. However this finding does not correspond with the expectation of a selective sweep event. For a population specific sweep the selective sweep allele is expected to exhibit a high frequency just in the population where the sweep occurs.

In contrast to the results of the *musculus-domesticus* comparisons four out of seven microsatellite loci that exhibit population specific extreme values consists of pure dinucleotide repeat units. Therefore, dinucleotide repeat markers with faster mutation rates are better suited to trace sweeps which have occurred in single populations. Especially, within the Cameroon population three loci were found that show the expected pattern of a selective sweep event.
Still no differences were found for example between the German and the France population although phylogenetic inferences in chapter 1 suggest that the populations have evolved their unique allele distributions. Some loci of the lnRV and lnRH calculations showed extreme values in the pairwise comparison but were excluded from further analysis due to the stringent criteria of at least three extreme values for one population. The lack of sweep loci between the German and France population may be caused by sharing more or less the same ecological habitat, or the method of identifying real sweeps was too stringent and has removed some potential sweep loci from further analysis.
This problem could be solved by genotyping more local populations to increase the possibility of further pairwise comparisons.

**Comparison of the "multi-locus candidate gene approach" to the "random microsatellite screen" and other multi-locus screens in different organisms**

The "random microsatellite screen" generally had a narrower distribution compared to the "multi-locus candidate gene approach". This result may either be caused that mainly pure dinucleotide repeats were implemented in the random screen or due to the fact that only one *Mus musculus* population, instead of two *Mus musculus* population as in the "candidate gene approach", was incorporated which reduced the comparisons between populations of different lineages and the number of comparisons with different values due to neutral divergence. The analysis for the "random microsatellite screen" was not done in such detail as for the "multi-locus candidate approach". Still, about seven loci out of 106 were identified to exhibit extreme values in multiple comparisons in the lnRV and lnRH distribution. Compared to the "candidate gene approach" with nine out of 81 significant loci, the number is lower. Since in the "candidate gene approach" the microsatellite loci were close to genes which might be involved in adaptation and selection a higher number of extreme values compared to a random microsatellite screen was expected.
Compared to other studies the identification of nine potential sweep loci out of 81 is rather high. Vigouroux et al. (2002) performed a multi-locus candidate screen in maize and identified 15 loci out of 501 exhibiting evidence for selection. Payseur et al. (2002) analysed

5,257 microsatellites within humans and found 43 sweep windows which included several linked markers with skews in their frequency distribution. These skews were an indicator of non-neutral evolution. Kayser et al. (2003) identified eleven out of 332 microsatellite loci as potential sweep candidates in human populations. However, similar to my data Schlötterer et al. (1997) identified one microsatellite locus out of 10 which showed reduced variability in a random microsatellite screen in *Drosophila melanogaster*. Another screen of Schlötterer (2002b) identified four loci by screening 94 random microsatellite loci in the human genome. Two of them showed promising features which might have been shaped by a selective sweep event. Considering that the "multi-locus candidate gene approach" that I applied for the investigation of the mouse populations was a candidate approach with genes that were pre-chosen because of their potential response to selection, I expected that the number of potential sweep loci exceed the number of sweeps compared to those in random microsatellite screens. In addition, the close vicinity of these microsatellite loci to the coding regions increases the probability that they might be located in the same sweep window as an advantageous mutation.

**Analysis of selective sweep windows and factors influencing the detection of sweep regions**

Evidently, all loci with extreme values must be evaluated by additional means to proof that the individual extreme value is caused by a selective sweep event. Genotyping microsatellite loci in close vicinity to each other is a suitable tool to identify the size of the sweep region (Harr et al. 2002, Kayser et al. 2003). Nevertheless, such an analysis requires the use of microsatellite loci with identical characteristics such as same repeat structure and same length of the repetitive region in order to ensure that they evolve under the same mutation rate (Ellegren 2000). In the case of the ß-Defensin 6 locus the flanking microsatellite loci were not suitable to trace the signatures of selection. They mainly consist of large stretches of dinucleotide repeat units, which correlate with a higher mutation rate. Thus, since the ß-Defensin locus was assumed to be a sweep in the *Mus domesticus* lineage the sweep event backdates too long and the traces were obscured by the high mutation rates of the flanking markers.

Another unknown factor influencing the detection of a selective sweep event is the size of the sweep window. The size essentially depends on the recombination rate of a region and on the strength of selection (Kim & Stephan 2002). The recombination rate can be estimated by comparing the physical (http://www.ensembl.org/Mus_musculus/) and the genetic map

(Copeland et al. 1993, Dietrich et al. 1994, Dietrich et al. 1996). Nevertheless, the selection coefficient remains unknown. For the genomic region around the ß-Defensin 6 locus the recombination rate was rather low: for a region of 2 Mb no recombination was found in the Copeland and Jenkins backcross map (Copeland et al. 1993). Therefore, a larger sweep window was expected which contradicts the genotyping results of the flanking microsatellite loci. Nevertheless, since the sweep is assumed to be old and occurred at the split of the *domesticus* and the *musculus* lineage sufficient time has past to recover diversity in the flanking region, particularly, as the flanking microsatellite loci are faster evolving than the core marker within the ß-Defensin 6 gene. Nothing can be inferred about the strength of the selection. A weak selection coefficient also results in a narrower sweep window. In other studies the size of the sweep windows differ substantially: Nair et al. (2003) identified a selective sweep window in *Plasmodium falciparum* to be as long as 100 kb which corresponded to 6 cM in this species. Glinka et al. (in press) even detected selective sweeps every 140 kb by scanning fragments on the *Drosophila* X-chromosome. Saez et al. (2003) estimated the size of a sweep window to be around 41 kb to 54 kb. Instead of this, Nachman found the extent of linkage equilibrium to be much smaller than 50 kb (Nachman et al. 2003). These differences depend on the sweep characteristics such as time of the sweep event, strength of the selection coefficient and the recombination rate of the genomic region.

Analysing the size of sweep windows is an important tool to develop guidelines for genome wide screens for selective sweeps: large sweep windows will easily be traced by genotyping random markers. In contrast, small sweep windows can be missed if markers are screened in low density. Generally, the source of a selective sweep event will be located within a functional region of a gene. Hence, a candidate locus screen of all potential genes within a genome is probably the method of choice to discover most of the important sweep events.

**Influence of microsatellite repeat types on the discovery of selective sweep events**

The occurrence of extreme lnRV or lnRH values was independent on the repeat unit types used for this screen. Still there was a significant difference in the number of extreme values per locus for the different repeat types: non-dinucleotide repeat markers proved to be significant in more comparisons than dinucleotide repeat markers. As dinucleotide repeat markers are known to have higher mutation rates than other repeat types (Ellegren 2000), the microsatellite loci are differentially suitable for different levels of divergence: older sweeps can only be traced with slower evolving markers as tri-, tetra- or interrupted repeat markers

which was also proven in this study; younger sweeps are recognizable with faster evolving markers as dinucleotide repeat loci.

**Type of genes that are involved in selective sweep event**

The nine microsatellites that have potentially been involved in a selective sweep event were closely linked to the following genes:

P42 was linked to the Plasma selenoprotein P (SELP) gene and P45 Pro-alpha1 (II) collagen chain gene; both genes were included in this analysis because a preliminary micro-array experiment showed differential expression of these genes between different mouse lineages (unpublished results). Selenoproteins are known to be differentially regulated and play a role in health, for example by protection of tissues from antioxidant injuries and in other regulatory pathways (Burk et al. 2003, Moustafa et al. 2003); Collagen genes are known to be highly conserved between mice and humans (Metsaranta et al. 1991), and they are extensively studied in biomedical research but their evolutionary significance is unclear so far. The beta-globin complex containing the microsatellite locus P61 was also identified via differential expression. The beta-globin complex has already been involved in evolutionary studies 20 years ago (Konkel et al. 1979) and is still an interesting candidate for differential selection and adaptation.

The olfactory senses in the house mouse are strongly developed and are important for the perception of the environment, therefore, olfactory genes were generally assumed to be good candidates for selection. Three olfactory genes were identified as potential sweep loci. The microsatellite locus P121 is located within the olfactory receptor gene family (AC091743). Locus P127 is located within MATH4B gene. This gene is involved in the activation cascade of bHLH regulators in olfactory neuron progenitors (Cau et al. 1997). Locus P141 is located within a cyclic nucleotide-gated olfactory channel protein gene (Ruiz et al. 1996).

Saliva is known to play a role in recognition in mice. One salivary androgen binding protein has already been identified to have evolved under strong divergent evolution between different mouse lineages and proved to play a role in female choice preferences (Karn & Laukaitis 2003, Talley et al. 2001). From screening six microsatellite loci within genomic sequences of different saliva genes the locus P104 showed a selective sweep. P104 is located within the ß-Defensin 6 gene; ß-Defensins are cationic peptides playing a role in microbial defence and are expressed in the mucosal surface (Bals et al. 1999).

Other sweep loci were connected to an interferon alpha/beta receptor (IFNAR). Interferons are generally involved in pathogen mediated immune response (Boehm et al. 1997) and can be considered to be continuously shaped by selection.
Locus P76 within the axonemal dynein heavy chain gene is already known to cause hybrid sterility in crosses with *Mus spretus* and might play a role in divergent evolution and separation of taxa (Fossella et al. 2000).
All those genes are very likely to be shaped by differential selection pressures and are good candidates for selective sweep events.

**Summary and Conclusion**

Microsatellite loci are suitable tools to detect regions under selection. The "random microsatellite screen" detected seven interesting regions within the house mouse genome. These regions were not further analysed for possible functional genes within close vicinity. With the "multi-locus candidate gene approach" nine genes revealed extreme values in multiple comparisons and are promising to be real sweeps. However, the search for the advantageous mutation that has led to the selective sweep must employ other approaches such as sequencing the functional regions around the significant microsatellite loci. Microsatellite loci are considered to be neutral markers, therefore, they will only be affected by hitchhiking along with an advantageous flanking mutation. Thus, sequencing is the tool to analyse functional regions and to identify the changes which are responsible for the selective advantage.

To find selective sweeps at the different levels of population comparisons the analysis must be separated for the different levels of divergence times between populations; all kinds of different microsatellite repeat types should be genotyped to trace older and younger sweep events. Screening more microsatellite loci in close vicinity to each other allow to proof potential sweep loci by identifying the sweep windows.

For the investigated populations it would not only be necessary to extend the level of divergence times but also to have multiple population of the same level to extend the level of pairwise comparisons. As discussed in detail in chapter 1 the *Mus domesticus* populations seem to be rather young. To determine further sweeps within the *Mus domesticus* clade it would be beneficial to have an older population as a reference to identify loci under selection in the younger populations. Older *Mus domesticus* populations are assumed to occur in the

near East and Arabian peninsula (chapter 1). To improve the results for the *Mus musculus* clade the current Czech population should be replaced by a new population to be collected under the same sampling scheme as all other populations, and to catch a third population to allow additional comparisons within the *Mus musculus* clade.

# The antimicrobial peptide ß-Defensin 6 was subjected to a selective sweep event caused by differences in the regulatory regions

## Introduction

Genomewide scans of microsatellite loci are valuable tools to identify regions subjected to selection as already discussed in chapter 2. Still, verification by other methods is necessary to distinguish false positives from real sweep loci (Schlötterer 2003, Harr et al. 2002). The multi-locus candidate gene approach identified the ß-Defensin 6 gene as a potential sweep locus. Nevertheless, by genotyping flanking microsatellite loci no hint for a sweep window around the locus was detectable. This result was either due to a very narrow sweep window or to higher mutation rates of the flanking microsatellite loci obscuring the traces of the selective sweep. Subsequently, the results of the selective sweep event had to be verified by sequencing the region around the ß-Defensin gene. In addition, the size of the sweep window had to be determined by sequencing several flanking fragments in distances of 5 to 20 kb to each other along the chromosome 8. Microsatellite mutation rates vary substantially: in mammals mutation rates ranging from $10^{-2}$ to $10^{-5}$ per locus and per generation were found (Weber & Wong 1993, Schug et al. 1998 and references therein) and in Drosophila even lower mutation rates of $10^{-6}$ were detected (Schlötterer 1998, Schug 1997). In contrast, mutation rates of nucleotide sequences vary less and are assumed to be around $10^{-9}$ per locus and per generation (Crow 1993). Therefore, nucleotide diversity is superior to trace a selective sweep because the flanking sequences will recover diversity under the same mutation rate as the sweep region itself, and thus allowing the identification of the selective sweep window.

The assumption of similar mutation rates and the limited number of alleles per nucleotide position facilitates the development of statistical tests and, so far, more methods have been developed for sequence data than for microsatellite data such as the HKA test, Tajima's D, H-Tests etc. (Fay & Wu 2001 and references therein). Prediction about neutral and non-neutral behaviour are more reliable for sequence data than for microsatellite data, and selective sweeps are retained over a longer time frame. By sequencing an outgroup, for example *Mus spretus* as the sister taxon of the house mouse species complex, the ancestral or derived state of a nucleotide polymorphism can be inferred and subsequently allow inferences about the behaviour of new derived mutations and their spread throughout the genome.

Therefore, sequencing is the preferred method to verify the results of the microsatellite screen. Since sequencing is more labour and cost intensive, the sequencing approach was applied to the ß-Defensin 6 locus only because this locus based on its features of extreme reduction in polymorphism in the *domesticus* lineage and high number of alleles in the *musculus* sample was assumed to display the classical expectation of a selective sweep event.

In contrast to microsatellite regions which presumably evolve neutrally and are just shaped by selection due to the hitchhiking effect, sequencing probably allows the identification of the advantageous changes responsible for the sweep because they can consist of functional and non-functional regions. Advantageous changes are expected to occur in functional regions of the genome such as RNA, or protein coding regions, or regulatory regions such as promoters and transcription factor binding sites. For example amino acid changes were identified in the *pfcrt* gene responsible for chloroquine resistance in malaria parasites (Wootton et al. 2002), and in the FOXP2 gene which in comparison to the very conserved FOXP2 gene in other apes exhibits two amino acid substitutions in the human lineage. The changes in the human lineage may be responsible for the development of articulation and speech in humans (Enard et al. 2002). Generally, the hypothesis is that a selective sweep is the consequence of a selective advantage either of a new mutation or due to changes in the environment which then suddenly favour a certain allele of an existing polymorphism. Due to the ubiquity of the genetic code changes in the coding regions can easily be correlated with changes in the amino acid composition of a protein and in the potentially different function of the gene. Statistics such as the ratio of synonymous to non-synonymous substitutions or the McDonald and Kreitman test deal with such changes in the coding region (Fay & Wu 2001, Nekrutenko et al. 2001, McDonald & Kreitman 1991). The ratio of synonymous to non-synonymous substitutions compares the frequency of nucleotide changes resulting in silent or in replacement substitutions in amino acid coding sequences. Under neutrality equal frequencies are expected while non-neutral mutations can be recognized by non-synonymous changes of nucleotides which are either more frequent under positive selection or less frequent under background selection (Fay & Wu 2001, Yang 2001). The McDonald and Kreitman test compares synonymous and replacement substitutions within and between species similar to the Hudson-Kreitman-Aguadé test for non-coding nucleotide sequences (Hudson et al. 1987, Mc Donald & Kreitman 1991).

Concerning regulatory regions of a gene inferences about advantageous changes are less obvious (Zhang & Gerstein 2003, Qiu 2003). The identification of promoters and *cis*-regulatory elements from sequence data is still a challenging task for bioinformatics because regulatory elements consist of a wide variety of different transcription factor binding sites in different combinations and vary in size from ten up to thousand base pairs (Lenhard et al. 2003, Qui 2003, Zhang et al. 2003). As complete genome sequences of complex organisms became available the importance of gene regulation and regulatory networks became more and more a subject of evolutionary studies because the complexity of different organisms are likely to depend more on differential gene regulation than on the actual number of genes (Markstein & Levine 2002). Approaches to detect transcription factor binding sites are based on the identification of the core motifs within sequence stretches. As the core motifs consist of four to six fully conserved base pairs only (Zhang & Gerstein 2003) concepts were developed to trace sequences for combinations of such informative motifs. Generally, regulatory regions consist of a variety of many transcription factor binding sites and, therefore, only sites which contain several binding elements (Qiu 2001) are considered as true regulatory sites.

Based on the availability of sequence data from different species, evolutionary concepts can be applied to identify regulatory regions as functional active regions which are subject to selective constraints and which will probably remain more conserved between two species than non-functional regions. This approach is called phylogenetic footprinting and is embedded in several software programs which try to identify functional regions by aligning sequences of different species (Zhang & Gerstin 2003). In case of the murine ß-Defensin 6 gene I applied the concept of phylogenetic footprinting by aligning the sequence with the outgroup *Mus spretus* and by searching the rat genome databases for orthologous sequences. Alignments with the different organisms then allow the identification of conserved regions containing potential transcription factor binding sites. The identification of such regions in the non-protein coding regions of the ß-Defensin 6 gene would then allow the search for functional differences in gene regulation between the different lineages *Mus musculus* and *Mus domesticus*.

Still, these computational approaches can only predict functional regions. To prove whether the ß-Defensin 6 gene is really differentially expressed in the two mouse lineages, the expression levels of mRNAs of different organs from wild caught mice of both lineages were

analysed. As all other ß-Defensin genes, the ß-Defensin 6 gene is a cationic peptide that is involved in microbial defence in various epithelial tissues. In total 14 murine ß-Defensin genes are known and located in two clusters on chromosome 8 (Maxwell et al. 2003, Morrison et al. 2003). All these genes consist of two exons and specifically the second exon seems to be subjected to strong positive selection. The strong selection pressure is due to the role of this gene family as defence against different bacterial pathogens (Morrison et al. 2003). Different ß-Defensin genes show tissue specific expressions (Jia et al. 2000, Yamaguchi et al. 2001) and are switched on or off by different regulatory pathways. While some ß-Defensin genes such as ß-Defensin 4 seem to be constitutively expressed other ß-Defensin genes are inducibly expressed by microbial infections (Burd et al. 2002, Bals et al. 1999, Morrison et al. 2002). The ß-Defensin 6 gene is the only gene of this family that is expressed in the skeletal muscle which might indicate a special physiological role (Yamaguchi et al. 2001). Therefore, the ß-Defensin 6 gene is an attractive candidate gene which is most likely involved in a selective sweep event. By measuring the expression level of the gene in various organs, I tried to identify possible differences in gene expression between the two lineages which might be a hint for differential regulation of the gene that was shaped by differential selection pressures.

Evidence from the literature indicates that all ß-Defensin genes are subject to positive selection pressure (Morrison et al. 2003). Now, the aim of this approach was to find out whether the gene has also differentially evolved in the lineages of *Mus domesticus* and *Mus musculus* causing the selective sweep in the *Mus domesticus* populations.

Summarising the aims of this chapter are:

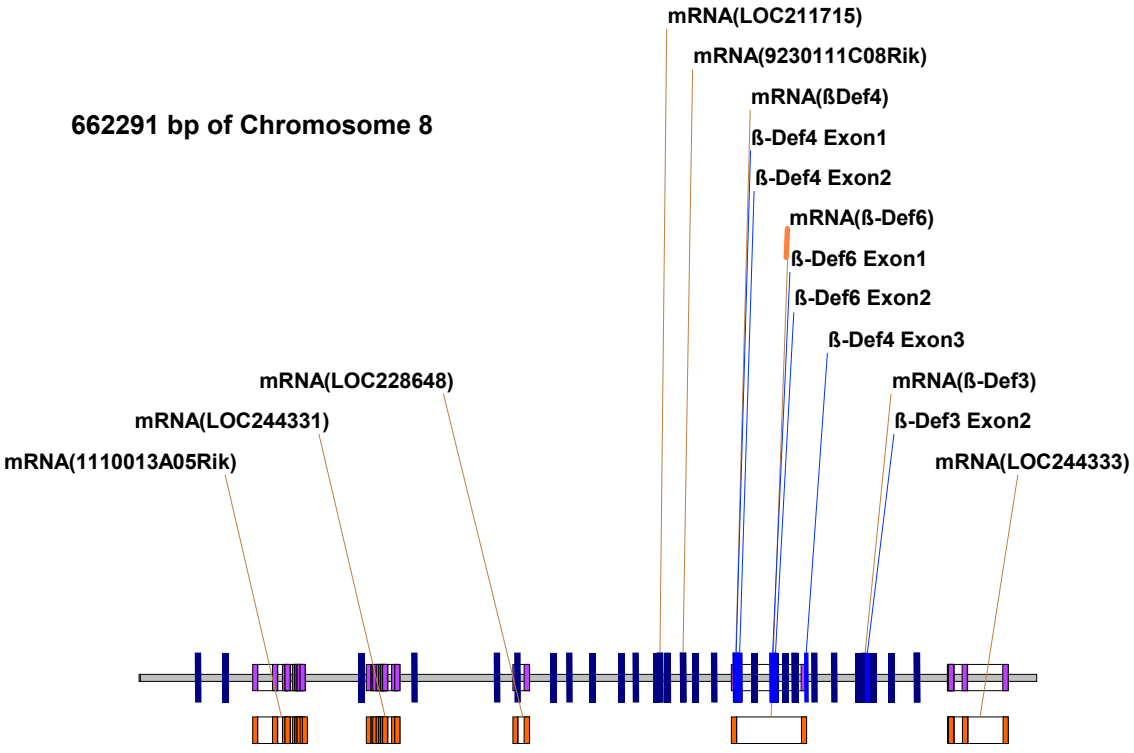
- to verify the selective sweep that was found by the microsatellite analysis as described in chapter 2;
- to determine the size of the sweep window by sequencing the flanking regions;
- to identify the advantageous changes within functional regions of the gene and their potential consequences such as altered amino acid sequences or changed regulatory patterns of the gene in the different populations;
- to prove that functional differences result in differential expressions of the gene in the different lineages.

## Materials and Methods

### Analysis of nucleotide diversity of the ß-Defensin 6 gene and the flanking region

660 kb of genomic sequence from the mouse chromosome 8 including the ß-Defensin 6 locus were downloaded from the ENSEMBL database. Since 38.6% of the mouse genome consists of repetitive elements such as LINEs, SINEs and transposable elements (Waterston et al. 2002) I applied the internet based program RepeatMasker (Smit & Green, RepeatMasker at http://ftp.genome.washington.edu/RM/RepeatMasker.html) to identify repetitive regions and to avoid sequencing them. Within the non-repetitive sequences primers were designed to amplify stretches of 500 to 1,000 bp. The amplicons were chosen in such a way that they were equally distributed over the whole region of 600 kb with a higher density around the ß-Defensin 6 locus (Figure 1). About 4,000 bp of the genomic region around the two ß-Defensin 6 exons were also sequenced. Primer design, PCR amplification and sequencing of these stretches were performed following the methods described in the previous chapters.

**Figure 1**: 662 kb of Chromosome 8 surrounding the ß-Defensin 6 gene. Annotated genes are displayed within the sequence. Dark blue bars represent the location of sequenced fragments. Light blue and purple bars represent the exon and coding sequences of genes.



In total 25 fragments of maximal 800 bp each and two fragments of 2,600 bp and 1,200 bp containing the ß-Defensin 6 exons were sequenced in 28 animals: selecting always five animals from Kazakhstan, Czech Republic, Cameroon, France, Germany and three animals from the USA. For outgroup comparisons four animals from *Mus spretus* were sequenced for each fragment. Among the 25 shorter fragments three fragments contained exon sequences of the ß-Defensin 3 and ß-Defensin 4 gene.
The sequences were aligned with the program Seqman of the Lasergene expert sequence analysis software (DNASTAR, Inc.) and imported into the program DNASP 3.51 (Rozas & Rozas 1999). As direct sequencing was performed there was the possibility of sequencing heterozygote individuals resulting in a base-calling for two different nucleotides at one locus. Most of the sequence analysis programs cannot deal with the heterozygote annotations, therefore, all sequences were duplicated and each nucleotide state was assigned to one of the duplicated sequences.

Gene diversity and the number of polymorphic sites were calculated separately for the two lineages *Mus musculus* and *Mus domesticus* with the program DNASP 3.51 (Rozas & Rozas 1999). By comparing all polymorphic sites against the *Mus spretus* sequences the ancestral or derived state for each polymorphism was inferred.

The multi-locus Hudson-Kreitman-Aguadé test developed by Jody Hey (http://lifesci.rutgers.edu/~heylab) was performed to test for non-neutral divergence between the house mouse lineages and *Mus spretus* and non-neutral polymorphism rates by analysing the variability within species. Theory predicts that under neutrality the interspecific divergence rate correlates with the intraspecific polymorphism.

The exon sequences were analysed for synonymous and non-synonymous nucleotide exchanges. The McDonald-Kreitman test was applied to test for non-neutral nucleotide exchanges within the coding and non-coding regions of the ß-Defensin 6 locus by comparing the house mouse lineages against *Mus spretus*. Nucleotide changes within coding regions were analysed for the Ka/Ks ratio. Ka is the proportion of non-synonymous changes at non-synonymous sites and Ks is the proportion of synonymous changes at synonymous sites. Both analysis are implemented in the program DNASP 3.51 (Rozas & Rozas 1999). The exon sequences were blasted against the rat genome to identify a possible orthologous gene.

**Analysis of regulatory regions of the ß-Defensin 6 locus**

To identify possible regulatory regions of the ß-Defensin 6 gene two *Mus domesticus* and two *Mus musculus* consensus sequences were created containing all possible nucleotide alleles. These sequences were aligned with one *Mus spretus* sequences and subjected to the program ConSite (http://www.phylofoot.org/ described in Lenhard et al. 2003). The program recognizes conserved stretches between the two species which are likely to contain functional regions. These regions are analysed for transcription factor binding sites. The search for transcription factor binding sites was performed with a pre-selection of vertebrate specific transcription factors.
The different *Mus musculus* and *Mus domesticus* sequences were aligned separately with the *Mus spretus* sequence to identify differences in possible transcription factor binding sites due to single nucleotide polymorphisms in the coding regions.

**Test for differential expression of ß-Defensin gene analysed in several organs**

RNA was isolated from 6 different organs from six *Mus domesticus* and six *Mus musculus* animals. The animals were caught in Germany and Czech Republic, kept in the laboratory for 48 hours, killed by applying a $CO_2$ atmosphere and directly dissected. The organs were shock frozen in liquid nitrogen and stored at -80°C. The following organs were used for measuring expression-levels of ß-Defensin 6: lung, tongue, salivary glands, esophagus, trachea and skeletal muscle from the hind leg.

For isolation of RNA the organs were homogenised in Trizol (Gibco) and total RNA isolation was performed following the manufacturers protocol. Total RNA was quantified by spectrophotometry and 5 µg of total RNA were used for reverse transcriptase reaction with the Superscript II RT-Polymerase Kit (Invitrogen) following the manufacturers protocol. Real-time PCR was performed using the SYBR Green Kit from Qiagen and reactions were run on the Light Cyler (Roche) by applying the following PCR program:

| | | |
|---|---|---|
| **1) Initial denaturing:** | 95°C for 15’ | |
| **2) Denaturing :**<br>**3) Annealing:**<br>**4) Elongation:**<br>**5) Measuring:** | 94°C for 8’’<br>58°C for 20’’<br>72°C for 8’’<br>82°C for 1’’ | During step 5: Fluorescence measurement |
| Repetition of step 2 to 5: 45 times | | |
| **6) melting curve** | Increase from 58°C to 95°C with 0,1°C/s | Permanent fluorescence measurement |
| **7) cooling** | 95°C to 45°C | |

The SYBR green is a fluorescence dye binding to double stranded DNA. The Light Cycler technology allows to quantify the amount of double stranded DNA such as amplified PCR products by measuring the fluorescence after each PCR cycle.

PCR was performed for the ß-Defensin 6 gene using intron spanning primers. Additionally two housekeeping genes Glyceraldehyd-3-phosphate dehydrogenase and TATA box binding protein were amplified.

**Table 1**: list of primer sequences used for the amplification of the ß-Defensin gene and the two housekeeping genes

| Gene | Accession | Primer |
|---|---|---|
| **ß-Defensin 6** | NM_054074 | ß-Def6-F: GTCATGAAGATCCATTACCTGC<br>ß-Def6-R: ACCCAGTCGAAAACCTCCATTGC |
| **Glyceraldehyd-3-phosphate dehydrogenase** | M32599 | GAPDH-F: CATCTTGGGCTACACTGAGG<br>GAPDH-R: GGAGGCCATGTAGGCCATG |
| **TATA box binding protein (Tbp)** | NM_013684 | TBP-F: TGCACCGTTGCCAGGCACC<br>TBP-R: TCAGCATTTCTTGCACGAAGTGC |

The theoretical prediction for PCR reactions are that each PCR cycle duplicates the amount of the PCR template. The expectation is that the amount of template increases exponentially until a certain saturation is reached due to depletion of dNTPs and reduced efficiency of the polymerase. Thus, the increase in PCR product follows a sigmoidal curve.

To compare the amount of template within different samples all probes were amplified in the same Light Cycler run. Early increase of the curve corresponds to higher template concentrations. For relative quantification of the different template concentrations a line was drawn through the parallel slopes of all sigmoidal curves. The crossing points of the slopes with this line correspond to the number of PCR-cycles that were necessary to reach the same amount of PCR product and are negatively correlated with the amount of the starting template.

The amount of template can differ due to RT-PCR efficiency and different mRNA concentration of the samples, independent on the original expression level. Therefore, normalisation with housekeeping genes is necessary. The expectation is that housekeeping genes are constitutively expressed and exhibit equal concentration in the same organs of different animals. Thus, they can be used as references to normalise expression levels of non-housekeeping genes by the following equation:

$$Expr_{(target\ gene)} = 2^{-\left(\frac{CP(target\ gene)}{CP(housekeeping\ gene)}\right)}$$

Expr = relative expression level of the target gene
CP = crossing point of the slopes with the crossing line

Samples were analysed for expression or non-expression of the ß-Defensin gene. Significant differences between the lineages of *Mus musculus* and *Mus domesticus* were inferred by calculating a chi-square test implemented in Excel separately for each organ and also summarised over all samples.

Differences in quantified expression levels between *Mus musculus* and *Mus domesticus* were analysed by the Mann-Whitney-U-Test implemented in SPSS 10.0.

## Results

### Nucleotide diversity along chromosome 8

In total 19711 bp were sequenced in each of about 56 chromosomes of the two lineages *Mus musculus* and *Mus domesticus*. The distribution along the chromosome, the length of the fragments and the diversity estimates per fragment and per lineage are listed in Table 2. The number of sequenced chromosomes per taxon is a duplication of the number of sequenced animals because of heterozygote animals. The polymorphic loci are the number of single nucleotide polymorphisms per sequenced fragment. The fixed derived alleles are monomorphic sites within one taxon which exhibit a derived allelic state compared to the outgroup *Mus spretus* but they are polymorphic in the *musculus* and *domesticus* comparison.

The nucleotide diversity per fragment along the chromosome for *Mus musculus* and *Mus domesticus* and the comparison of the two lineages are shown in Figure 2.

The *Mus domesticus* samples generally exhibit lower gene diversity especially in the region from 300 kb upstream to 45 kb downstream of the ß-Defensin 6 gene. Outside this regions as well as at single fragments within the region higher nucleotide diversity for *Mus domesticus* is found indicating that the lower nucleotide diversity is not a general feature of *Mus domesticus*. The highest proportion of fragments with extreme reduced nucleotide diversity is found from 10 kb upstream to 10 kb downstream of the ß-Defensin 6 gene indicating a whole window of reduced nucleotide diversity in this area while in the other region the nucleotide diversity is more fluctuating. The results for this window are very reliable as long stretches were sequenced within this area: in an area of 22 kb 5,000 bp are sequenced while in other regions of the same size just one fragment of about 600 bp is sequenced. More fragments around the ß-Defensin 6 locus were sequenced to identify the size of a possible sweep window. From this analysis I assume a sweep window of roughly 22 kb around the ß-Defensin 6 locus because outside this window are stretches with higher nucleotide diversity. As shown in Figure 2c the area around the ß-Defensin 6 locus in *Mus domesticus* exhibits not only the lowest nucleotide diversity in comparison to *Mus musculus* but also in comparisons with other fragments within *Mus domesticus*: the $\pi$-values of the sequences within the window are lower than the average $\pi$-value of *Mus domesticus*.

Other regions along chromosome 8 were not as densely sequenced and the differences in nucleotide diversity between *Mus musculus* and *Mus domesticus* are not that extreme as compared within the region of the ß-Defensin 6 gene.

**Table 2:** List of sequenced fragments in *Mus musculus* and *Mus domesticus* and their relative distance in kb to the ß-Defensin 6 locus. Diversity indices were calculated by the program DNASP 3.51.

| | Rel. distance to ß-Def6 [kb] | Sequence length | ***Mus musculus*** | | | | | ***Mus domesticus*** | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | Number of Chro-mosomes | Nucleotide Diversity | Polymor-phic Loci | fixed derived alleles | Insertion/Deletion | Number of Chro-mosomes | Nucleotide Diversity | Polymor-phic Loci | fixed derived alleles | Insertion/Deletion |
| **SNP8-9** | -1400 | 473 | 18 | 0.00398 | 9 | 3 | | 32 | 0.00621 | 11 | 1 | |
| **SNP8-8** | -403.1 | 546 | 20 | 0.00053 | 2 | 9 | | 32 | 0.00585 | 22 | 0 | |
| **SNP8-7** | -302.1 | 644 | 16 | 0.00378 | 10 | 1 | I/D | 34 | 0.0029 | 6 | 1 | I/D |
| **SNP8-6d** | -263.2 | 724 | 8 | 0.00219 | 3 | | I/D | 32 | 0.00104 | 4 | | I/D |
| **SNP8-6** | -202.1 | 799 | 18 | 0.00532 | 14 | | | 32 | 0.00043 | 2 | 9 | |
| **SNP8-5ef** | -187 | 582 | 20 | 0.00378 | 7 | | | 36 | 0.00313 | 8 | | |
| **SNP8-5d2** | -160.5 | 725 | 12 | 0.00196 | 4 | | | 28 | 0.0007 | 2 | | |
| **SNP8-5cd** | -148.6 | 550 | 10 | 0.00162 | 2 | | | 36 | 0.0002 | 2 | | |
| **SNP8-4c** | -110 | 671 | 18 | 0.0033 | 6 | | | 30 | 0.00055 | 2 | | |
| **SNP8-4ab** | -99.5 | 875 | 16 | 0.00222 | 6 | 1 | | 28 | 0.00032 | 3 | 1 | |
| **SNP8-4a** | -84.6 | 564 | 16 | 0.00356 | 5 | | | 28 | 0.00136 | 3 | 1 | |
| **SNP8-3c** | -76 | 726 | 14 | 0.00605 | 10 | | | 30 | 0.00033 | 1 | 1 | |
| **SNP8-3a** | -55.1 | 578 | 18 | 0.00204 | 5 | | I/D | 30 | 0.00087 | 3 | 2 | |
| **SNP8-2a** | -41.5 | 750 | 14 | 0.00035 | 1 | 1 | | 28 | 0.00121 | 2 | | |
| **Defb4-Exon1** | -25.8 | 801 | 14 | 0.00344 | 7 | | | 26 | 0.00055 | 2 | 1 | |
| **Defb4-Exon2** | -23 | 549 | 16 | 0.00131 | 4 | | I/D | 14 | 0.00236 | 4 | 1 | |
| **SNP8-1** | -12 | 544 | 20 | 0.00577 | 7 | 1 | I/D | 36 | 0.00032 | 2 | 1 | |
| **Defb6-1** | 0 | 2596 | 18 | 0.00479 | 32 | | | 28 | 0.00019 | 4 | 7 (1 only dom) | |
| **Defb6-2** | 2.9 | 1152 | 12 | 0.01098 | 28 | | I/D | 26 | 0.00171 | 5 | 10 | I/D |
| **SNP8-0a** | 10.8 | 697 | 14 | 0.00787 | 6 | | I/D | 26 | 0.00256 | 5 | 1 | |
| **SNP8-2down2** | 16.8 | 648 | 10 | 0.00274 | 5 | 2 | | 32 | 0.00322 | 10 | 2 | |
| **SNP8-3down** | 32.4 | 628 | 12 | 0.00321 | 6 | | I/D | 32 | 0.0008 | 4 | 2 | |
| **SNP8-4down** | 47.1 | 697 | 12 | 0.00724 | 11 | 2 | | 14 | 0.00429 | 8 | 1 | |
| **SNP8-5down** | 65.1 | 496 | 8 | 0.00276 | 4 | 3 (2 only mus) | | 32 | 0.00555 | 9 | 1 | |
| **Defb3_Exon2** | 71.4 | 555 | 18 | 0.00138 | 8 | 1 | | 32 | 0.00085 | 2 | 3 | |
| **SNP8-7down** | 89.2 | 562 | 18 | 0.03466 | 42 | | I/D | 32 | 0.01327 | 23 | 24 | |
| **SNP8-8down** | 108.1 | 579 | 16 | 0.00923 | 12 | 1 | I/D | 32 | 0.00695 | 21 | | I/D |
| **Total** | 1500 kb | 19711 | 406 | | 256 | 25 | | 798 | | 178 | 71 | |
| **Average** | | 730 | 15 | 0.00504 | 9.63 | 2.27 | | 29.56 | 0.0025 | 5.81 | 3.55 | |

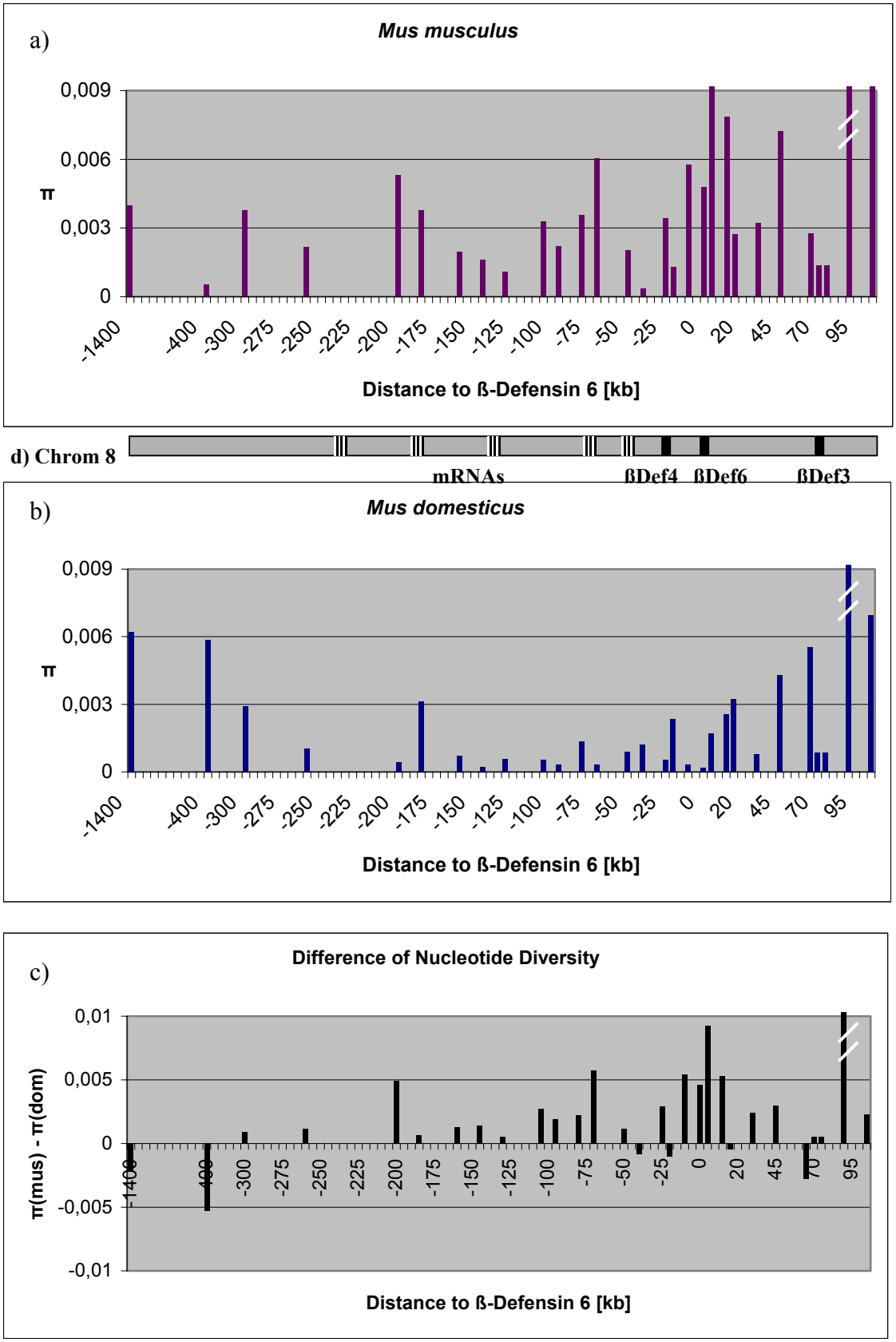



**Figure 2:** Nucleotide diversity calculated per fragment in a) *Mus musculus* and b) *Mus domesticus*; in c) the difference in nucleotide diversity for *Mus musculus* – *Mus domesticus*; in d) the location of ß-Defensin exons and other annotated potential genes on chromosome 8.

An interesting result of the sequencing approach is the high proportion of fixed derived alleles in the *Mus domesticus* samples. While the average number of polymorphic loci is 40% smaller in *Mus domesticus* than in *Mus musculus* the number of fixed derived alleles is 1.5 times higher. All these fixed derived sites are SNPs which are polymorphic in *Mus musculus* but fixed in *Mus domesticus* except for one site which is a new mutation specific for the *Mus domesticus* lineage. The highest proportion of fixed derived alleles is found in SNP8-7down, SNP8-6, ß-Defensin 6-1 and ß-Defensin 6-2; the ß-Defensin 6-1 locus contains the *Mus domesticus* specific SNP. Locus SNP8-7down exhibits, in general, strange results such as a very high nucleotide diversity and a very high number of heterozygote individuals and, therefore, needs further analysis for example a check for a possible duplication of the sequence. Therefore the ß-Defensin 6 locus is among the sites with most of the fixations which represent a typical feature of a selective sweep event.

**Test for selection on chromosome 8**

With 20 of the sequenced fragments the multi-locus HKA-test was performed. The results reveal a significant deviation from neutrality in the comparison of *Mus spretus* to *Mus domesticus* ($\chi2 = 71.6594$; $dF = 19$; $p < 0.001$); no such deviation was found in the comparison *Mus musculus* and *Mus spretus* ($\chi2 = 21.147$; $dF = 19$; $p = 0.329$). The results are listed in Table 3. The average of the expected divergence between *Mus domesticus* and *Mus spretus* with 19.6 differences per fragment is the same as the averaged divergence between *Mus musculus* and *Mus spretus* with 19.2 which indicates that the significant $\chi2$ results are due to differences within the sequenced fragments of the *Mus domesticus* samples. The $\chi2$ value is the sum of deviations of the within species polymorphism and of the divergence between the taxa. Therefore, loci with the highest values strongly influence the high $\chi2$ value and are responsible for the deviation from neutrality. For *Mus domesticus* the largest deviations are found for ß-Defensin 6-1, SNP8-8 and SNP8-8down. While the first locus exhibits much less polymorphism than expected the last two loci exhibit much higher polymorphism than expected. The results of the last two loci are caused by high numbers of nucleotide exchanges at low frequency; the frequencies themselves are not considered in the HKA-test. Exclusion of these fragments from the analysis still results in a significant $\chi2$ test indicating that these fragments are not the only reason for a significant HKA test result (data not shown). Generally, many fragments of *Mus domesticus* show higher deviations indicating that the whole region is affected by non-neutral evolution. The highest deviation is found at the ß-Defensin 6-1 locus proving that selection and selective sweeps have acted on this locus.

In contrast to this no such extreme deviation is found in any fragment of the *Mus musculus* screen indicating that no selection process affected this region.

**Table 3:** Result of the HKA-test: a) *Mus domesticus* compared with *Mus spretus*; b) *Mus musculus* compared with *Mus spretus*.

| a) | Polymorphism within *Mus domesticus* | | | | Divergence between *Mus domesticus* and *Mus spretus* | | | |
|---|---|---|---|---|---|---|---|---|
| **Locus** | observed | expected | variance | Deviation | observed | Expected | variance | deviation |
| **ß-Def6-1** | 4 | 19.48 | 59.76 | 4.007 | 70 | 54.52 | 79.57 | 3.01 |
| **ß-Def6-2** | 5 | 9.86 | 20.57 | 1.147 | 33 | 28.14 | 34.82 | 0.678 |
| **SNP8-0a** | 6 | 2.85 | 3.75 | 2.639 | 5 | 8.15 | 8.71 | 1.137 |
| **SNP8-1** | 2 | 7.99 | 14.0 | 2.566 | 27 | 21.01 | 24.72 | 1.454 |
| **SNP8-2a** | 2 | 1.05 | 1.17 | 0.767 | 2 | 2.95 | 3.02 | 0.297 |
| **SNP8-2down** | 10 | 5.4 | 8.3 | 2.553 | 10 | 14.6 | 16.4 | 1.292 |
| **SNP8-3a** | 3 | 4.27 | 6.14 | 0.261 | 13 | 11.73 | 12.89 | 0.124 |
| **SNP8-3c** | 1 | 4.00 | 5.64 | 1.595 | 14 | 11 | 12.02 | 0.749 |
| **SNP8-4a** | 3 | 2.35 | 2.95 | 0.142 | 6 | 6.65 | 7.02 | 0.06 |
| **SNP8-4ab** | 3 | 9.21 | 18.22 | 2.117 | 32 | 25.79 | 31.39 | 1.229 |
| **SNP8-4c** | 2 | 6.67 | 11.23 | 1.939 | 23 | 18.33 | 21.17 | 1.029 |
| **SNP8-4down** | 8 | 4.52 | 7.69 | 1.576 | 12 | 15.48 | 17.5 | 0.692 |
| **SNP8-5cd2** | 2 | 5.24 | 7.82 | 1.341 | 17 | 13.76 | 15.36 | 0.683 |
| **SNP8-5down** | 9 | 4.86 | 7.21 | 2.381 | 9 | 13.14 | 14.6 | 1.175 |
| **SNP8-5ef** | 8 | 6.89 | 11.36 | 0.108 | 17 | 18.11 | 20.87 | 0.059 |
| **SNP8-6** | 2 | 6.44 | 10.63 | 1.854 | 22 | 17.56 | 20.16 | 0.978 |
| **SNP8-7** | 6 | 5.46 | 8.34 | 0.035 | 14 | 14.54 | 16.32 | 0.018 |
| **SNP8-7down** | 23 | 17.44 | 48.16 | 0.642 | 42 | 47.56 | 66.62 | 0.464 |
| **SNP8-8** | 22 | 8.91 | 16.80 | 10.208 | 11 | 24.09 | 28.98 | 5.915 |
| **SNP8-8down** | 21 | 9.12 | 17.53 | 8.049 | 13 | 24.88 | 30.09 | 4.688 |
| **average** | 7.1 | 7.1005 | **Sum:** | 45.927 | 19.6 | 19.5985 | **Sum:** | 25.731 |

| b) | Polymorphism within *Mus musculus* | | | | Divergence between *Mus musculus* and *Mus spretus* | | | |
|---|---|---|---|---|---|---|---|---|
| **Locus** | observed | expected | variance | deviation | observed | expected | variance | deviation |
| **ß-Def6-1** | 32 | 36.24 | 212.56 | 0.085 | 66 | 61.76 | 172.8 | 0.104 |
| **ß-Def6-2** | 28 | 20.74 | 94.26 | 0.559 | 33 | 40.26 | 87.44 | 0.602 |
| **SNP8-0a** | 6 | 4.57 | 7.82 | 0.26 | 7 | 8.43 | 10.49 | 0.194 |
| **SNP8-1** | 7 | 12.82 | 33.63 | 1.007 | 27 | 21.18 | 34.24 | 0.99 |
| **SNP8-2a** | 1 | 1.06 | 1.23 | 0.002 | 2 | 1.94 | 2.05 | 0.001 |
| **SNP8-2down** | 5 | 4.56 | 8.55 | 0.023 | 9 | 9.44 | 12.04 | 0.016 |
| **SNP8-3a** | 5 | 6.29 | 11.59 | 0.143 | 12 | 10.71 | 14.05 | 0.118 |
| **SNP8-3c** | 10 | 9.35 | 23.55 | 0.018 | 17 | 17.65 | 26.73 | 0.016 |
| **SNP8-4a** | 5 | 5.06 | 8.74 | 0 | 9 | 8.94 | 11.27 | 0 |
| **SNP8-4ab** | 6 | 14.1 | 42.63 | 1.539 | 33 | 24.9 | 42.95 | 1.527 |
| **SNP8-4c** | 6 | 11.47 | 29.11 | 1.026 | 25 | 19.53 | 30.65 | 0.975 |
| **SNP8-4down** | 11 | 8.16 | 19.54 | 0.412 | 13 | 15.84 | 23.14 | 0.348 |
| **SNP8-5cd2** | 2 | 6.19 | 13.55 | 1.293 | 17 | 12.81 | 17.6 | 0.996 |
| **SNP8-5down** | 4 | 4.29 | 8.44 | 0.01 | 10 | 9.71 | 12.45 | 0.007 |
| **SNP8-5ef** | 7 | 9.05 | 19.42 | 0.216 | 17 | 14.95 | 21.46 | 0.196 |
| **SNP8-6** | 14 | 11.1 | 27.62 | 0.305 | 16 | 18.9 | 29.31 | 0.288 |
| **SNP8-7** | 10 | 9.76 | 23.44 | 0.002 | 17 | 17.24 | 25.89 | 0.002 |
| **SNP8-7down** | 42 | 25.97 | 119.5 | 2.149 | 29 | 45.03 | 104.05 | 2.468 |
| **SNP8-8** | 2 | 5.66 | 9.71 | 1.377 | 13 | 9.34 | 11.89 | 1.125 |
| **SNP8-8down** | 12 | 8.56 | 19.5 | 0.605 | 12 | 15.44 | 22.37 | 0.528 |
| **average** | 10.75 | 10.75 | **Sum:** | 11.03 | 19.2 | 19.2 | **Sum:** | 10.5 |

**Analysis of the coding regions**

The reduced polymorphism at the locus ß-Defensin 6 and the HKA test revealed that this locus has undergone non-neutral changes in the *Mus domesticus* lineage such as a selective sweep event. To identify the reason for the non-neutral evolution sequences of the exons of the locus were checked for any non-synonymous nucleotide exchanges.

The first exon consists of 19 codons which contained no nucleotide polymorphism in the two house mouse lineages. The same is true for the second exon consisting of 45 codons. Accordingly, the amino acid coding sequences do not contain any advantageous changes responsible for the selective sweep event. In contrast to this, the sequences of three *Mus spretus* animals were polymorphic with three intraspecific SNPs resulting in 2 amino acid replacements and 6 interspecific SNPs resulting in four amino acid replacements in exon 1 and three intraspecific polymorphism resulting in three replacement polymorphisms in exon 2 and one interspecific replacement polymorphism.

```
Nucleotide sequence Exon1

Cons     --* -*- --- --- --- -** --- --* -** --- *-- --- --- *-- --- --- --- --- ---
Dom/Mus ATG AAG ATC CAT TAC CTG CTC TTT GCC TTT ATC CTG GTG ATG CTG TCT CCA CTT GCA
Spre1   ATG AGG ATC CAT TAC CAT CTC TTC GCA TTT CTC CTG GTG CTG CTG TCT CCA CTT GCA
Spre2   ATG AGG ATC CAT TAC CAT CTC TTC GAA TTT CTC CTG GTG CTG CTG TCT CCA CTT GCA
Spre3   ATA AGG ATC CAT TAC CAT CTC TTT GCA TTT CTC CTG GTG CTG CTG TCT CCA CTT GCA



Amino acid sequence Exon1

Cons      **---*--*-*--*-----
Dom/Mus  MKIHYLLFAFILVMLSPLA
Spre1    MRIHYHLFAFLLVLLSPLA
Spre2    MRIHYHLFEFLLVLLSPLA
Spre3    IRIHYHLFAFLLVLLSPLA
```

```
Nucleotide sequence Exon2

Cons    --- --- --- --- --- --- --- --- -*- --- --- --- --- --- --- --- --- --- --- ---
Dom/Mus GCC TTT TCC CAA TTA ATC AAC AGT CCA GTA ACA TGC ATG AGC TAT GGA GGC TCA TGC
Spre1   GCC TTT TCC CAA TTA ATC AAC AGT CCA GTA ACA TGC ATG AGC TAT GGA GGC TCA TGC
Spre2   GCC TTT TCC CAA TTA ATC AAC AGT CCA GTA ACA TGC ATG AGC TAT GGA GGC TCA TGC
Spre3   GCC TTT TCC CAA TTA ATC AAC AGT CTA GTA ACA TGC ATG AGC TAT GGA GGC TCA TGC

--- -*- --- --- --- --- --- --- -*- --- --- --- --- --- --- --- --- --- --- ---
CAG CGT TCA TGC AAT GGA GGT TTT CGA CTG GGT GGC CAT TGT GGC CAT CCT AAA ATC AGA TGC
CAG CGT TCA TGC AAT GGA GGT TTT CGA CTG GGT GGC CAT TGT GGC CAT CCT AAA ATC AGA TGC
CAG CAT TCA TGC AAT GGA GGT TTT CAA CTG GGT GGC CAT TGT GGC CAT CCT AAA ATC AGA TGC
CAG CGT TCA TGC AAT GGA GGT TTT CGA CTG GGT GGC CAT TGT GGC CAT CCT AAA ATC AGA TGC

--- -*- --- --- ---
TGC CGC AGA AAA TAG
TGC CAC AGA AAA TAG
TGC CAC AGA AAA TAG
TGC CAC AGA AAA TAG



Amino acid sequence Exon2

Cons    --------*-----------*-------*-------------*--
Dom/Mus AFSQLINSPVTCMSYGGSCQRSCNGGFRLGGHCGHPKIRCCRRK
Spre1   AFSQLINSPVTCMSYGGSCQRSCNGGFRLGGHCGHPKIRCCHRK
Spre2   AFSQLINSPVTCMSYGGSCQHSCNGGFQLGGHCGHPKIRCCHRK
Spre3   AFSQLINSLVTCMSYGGSCQRSCNGGFRLGGHCGHPKIRCCHRK
```

The McDonald-Kreitman test was applied to the whole region of the ß-Defensin 6-1 and ß-Defensin 6-2 loci to check for selection between the different mouse species. For both fragments the McDonald-Kreitman test revealed no significant results (G-value = 0.639; p = 0.42415 for part 1 and G = 0.003; p = 0.95366 for part 2) indicating that the divergence at the loci is compatible with neutral divergence between the taxa *Mus spretus* and *Mus musculus/domesticus* although functional regions are always assumed to be under selective constraints.

The Ka/Ks test was applied to the coding sequences to identify the rate of divergence between the species. The sequence of the *Mus musculus*/*domesticus* lineage was tested against the three *Mus spretus* sequences revealing 49 silent positions and 143 non-synonymous positions. The number of silent differences ranged from 1.5 to 2.5 and the number of non-synonymous differences was between 5.5 and 9. In the three comparisons the Ka/Ks ratios ranged from 0.7 to 1.8. Ka/Ks values above 1 are usually a sign for strong positive selection which contradicts the results of the McDonald-Kreitman test.

The exon sequences were blasted against the Rat genome database. Exon 1 matched with one part of the ß-Defensin 2 precursor gene (NM_022544) with 82.46% sequence identity but a

very low E-value on rat chromosome 16 within the cluster of ß-Defensin genes. The second exon also matched with a high E-value of $8.9e^{-7}$ but no coding regions corresponded to the match which indicates that the mouse ß-Defensin 6 gene does not directly correspond to a ß-Defensin gene in the rat genome.

**Analysis of regulatory regions**

The above analysis shows that the advantageous changes have occurred after the divergence of the *Mus musculus* and the *Mus domesticus* lineage. Since no differences were found in the amino acid coding regions between the lineages the expectation is to find changes in the regulatory regions of the gene.

In total 45 transcription factor binding sites were identified by the program ConSite (Table 4). Nineteen TFs are mouse specific transcription factors and 26 are specific for vertebrates. 25 transcription factor binding sites showed differences by comparing the *Mus musculus* and *Mus domesticus* sequences. In 24 cases *Mus musculus* showed variability in transcription factor binding sites with the presence of the site in one and absence in the other *Mus musculus* sequence while the *Mus domesticus* samples were monomorphic. In nine cases the transcription factor binding site was lost while in fifteen cases the transcription factor binding site was present in the *Mus domesticus* samples. Eight transcription factor groups were in close vicinity of less than 50 bp of twelve fixed derived alleles of the *Mus domesticus* sequences correlating with the loss of binding sites in four groups. Two of these groups contained three fixed derived alleles in a sequence stretch of 30 bp. Two fixed derived alleles correspond to a loss of the binding site and two fixed derived alleles showed no correlation with gain or loss of a binding site.

**Table 4:** List of potential transcription factor binding sites and their localization identified by the program ConSite. Depending which sequence was aligned to the *Mus spretus* sequence different TFs were identified. The last column indicates the location of the fixed derived mutations in the *Mus domesticus* samples; in bold letters are sites less than 50 bp distant to a TF binding site. Light grey and white shadings indicate same location with domains for several TFs. Dark grey shading indicates the amino acid coding part of the gene.

| origine of Transcription factors | Transcription factors | Sequence motif | dom1 | dom2 | mus1 | mus2 | Location of fixed derived allele (dom) |
|---|---|---|---|---|---|---|---|
| Rat TF | HNF-3beta | ATGTAAACATTC | 63 | | | 63 | |
| Human TF | FREAC-4 | GTAAACAT | 65 | | | 65 | |
| Mouse TF | Sox-5 | AAACATT | 67 | | | 67 | 125 |
| Mouse TF | SOX17 | GAGAATGCA | | | 295 | | |
| Human TF | E4BP4 | ATAGGTAACAT | | | 325 | | **311** |
| Human TF | E4BP4 | AGGTAACATAA | | | 327 | | 509 |
| Human TF | HFH-3 | GCATATTTGCTT | 715 | 715 | 715 | 715 | |
| Rat TF | HNF-3beta | GCATATTTGCTT | 715 | 715 | 715 | 715 | |
| Mouse TF | Sox-5 | AATGTTA | 750 | 750 | 750 | | |
| Mouse TF | Sox-5 | AGTGTTA | 759 | 759 | 759 | | |
| Mouse TF | Sox-5 | GGACAAT | 837 | | 837 | | **835** |
| Mouse TF | SOX17 | GACAATAGG | 838 | | 838 | | |
| Mouse TF | Sox-5 | TAGCAAT | 857 | 857 | 857 | 857 | |
| Human TF | RORalfa-1 | AGGCAGGTCA | 947 | 947 | 947 | 947 | |
| Human TF | AML-1 | TACCACAGA | 1034 | 1034 | 1034 | 1034 | |
| Mouse TF | SOX17 | ATCATTGTG | 1596 | 1596 | 1596 | | |
| Human TF | AML-1 | ATTGTGTTT | 1599 | 1599 | 1599 | | |
| Mouse TF | Sox-5 | ATTGTGT | 1599 | 1599 | 1599 | | |
| Human TF | HFH-3 | TTGTGTTTGTAT | 1600 | 1600 | 1600 | | |
| Human TF | HFH-3 | AACCAAATATGC | 1850 | 1853 | 1838 | | **1825** |
| Rat TF | HNF-3beta | AACCAAATATGC | 1850 | 1853 | 1838 | | |
| Human TF | FREAC-4 | CTTTTTAC | 1861 | 1864 | 1849 | 1864 | |
| Human TF | RORalfa-1 | TTACCTTGAA | 1865 | 1868 | 1853 | 1868 | |
| Mouse TF | c-FOS | TCATTCAC | 1891 | 1894 | 1879 | 1894 | |
| Mouse TF | Brachyury | TTCACAGCTAA | 1894 | 1897 | 1882 | 1897 | |
| Rat TF | USF | TCACATG | 1911 | 1914 | 1899 | 1914 | |
| Mouse TF | ARNT | CACATG | 1912 | 1915 | 1900 | 1915 | |
| Mouse TF | n-MYC | CACATG | 1912 | 1915 | 1900 | 1915 | |
| Human TF | FREAC-4 | TTAAACAT | 1981 | 1984 | 1969 | 1984 | **2005**, 2170 |
| Human TF | FREAC-2 | TAAATGTAAAGAAG | | | 2620 | | 2430 |
| Human TF | FREAC-4 | GTAAAGAA | | | 2625 | | **2650** |
| Human TF | FREAC-4 | TTAAACAA | 2763 | 2766 | 2751 | | |
| Mouse TF | Sox-5 | AAACAAG | 2765 | 2768 | 2753 | | |
| Human TF | RXR-VDR | GGGTCAAATACTTCA | 2789 | 2792 | 2777 | | |
| Mouse TF | Sox-5 | GAAAAAT | | | 2828 | | **2820, 2830, 2840** |
| Vertebrate TF | Thing1-E47 | CAGCCAGACA | 2996 | 2999 | | 2999 | |
| Mouse TF | c-FOS | GTGAGGCA | 3054 | | 3042 | 3057 | |
| Human TF | Thing1-E47 | AAACCAGAAC | 3307 | 3310 | | 3310 | |
| Human TF | AML-1 | AACCAGAAC | 3308 | 3311 | 3372 | 3311 | |
| Human TF | Myf | CAGCATCAGCGG | 3326 | 3329 | | 3329 | |
| Human TF | HFH-3 | AAACAGATATCA | 3362 | 3365 | | 3365 | **3390** |
| Mouse TF | Sox-5 | CAACAGT | 3512 | 3515 | 3507 | | 3415 |
| Mouse TF | SOX17 | GCCATTGTG | 3586 | 3589 | 3581 | 3589 | |
| Human TF | AML-1 | ATTGTGGCC | 3589 | 3592 | 3584 | 3592 | |
| Mouse TF | Sox-5 | ATTGTGG | 3589 | 3592 | 3584 | 3592 | |
| Human TF | USF | CAACGTG | | | 3645 | | |
| Mouse TF | ARNT | AACGTG | | | 3646 | | **3655, 3675, 3685** |
| Mouse TF | n-MYC | AACGTG | | | 3646 | | |
| Mouse TF | c-FOS | GTGAATAA | | | 3649 | | |

**Differential Expression of the ß-Defensin 6 gene in various epithelial organs**

To test the reliability of the real-time Light Cycler PCR results the PCR of the ß-Defensin 6 gene was repeated three times in all twelve esophagus and twelve trachea organs; the housekeeping genes GAPDH was repeated two times in all organs. The line crossing the sigmoidal curves of all PCR reaction was manually drawn through the parallel part of the curves and the crossing points with this line were determined (Figure 3).

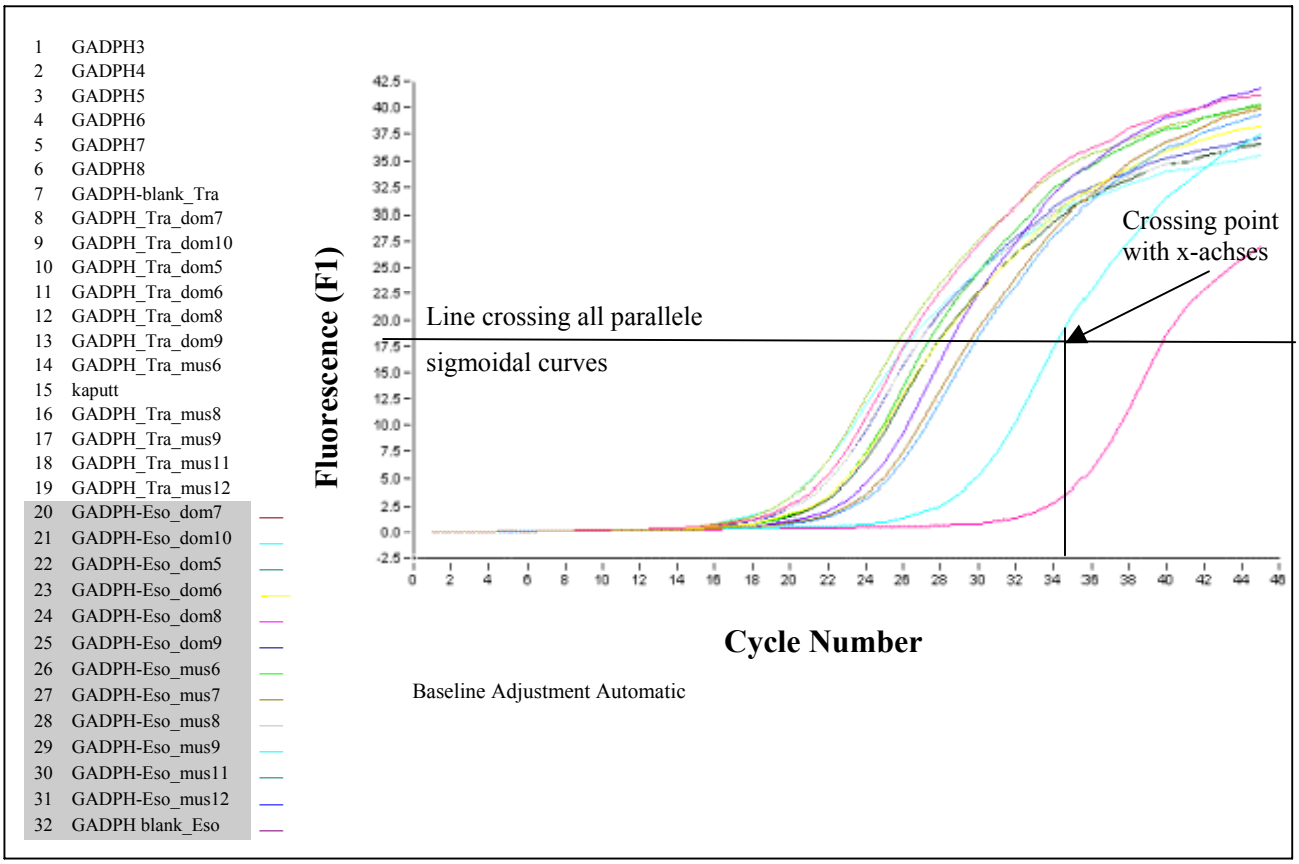



**Figure 3:** Light cycler results for the amplification of GAPDH in the esophagus of *Mus musculus* and *Mus domesticus* and demonstration of the determination of the crossing points for all samples. The coloured lines correspond to the increase in double stranded PCR product per PCR cycle in 6 *Mus musculus* and 6 *Mus domesticus* animals. The colours and the corresponding samples are listed in the dark underlayed part of the legend. The pink curve represents the amplification results of the blank probe. Increase in product is very delayed and correspond to random amplifications of the PCR primers. The cycle numbers at the crossing points are a relative measurement of the initial amount of template. Lower cycle number correspond to higher template concentrations, since less cycles are needed to reach an equal amount of PCR-product that is reached in all probes at the crossing line.

For the replicate PCR runs the mean of the crossing points for all samples and the standard deviations were calculated and are shown as error bars in figure 4.

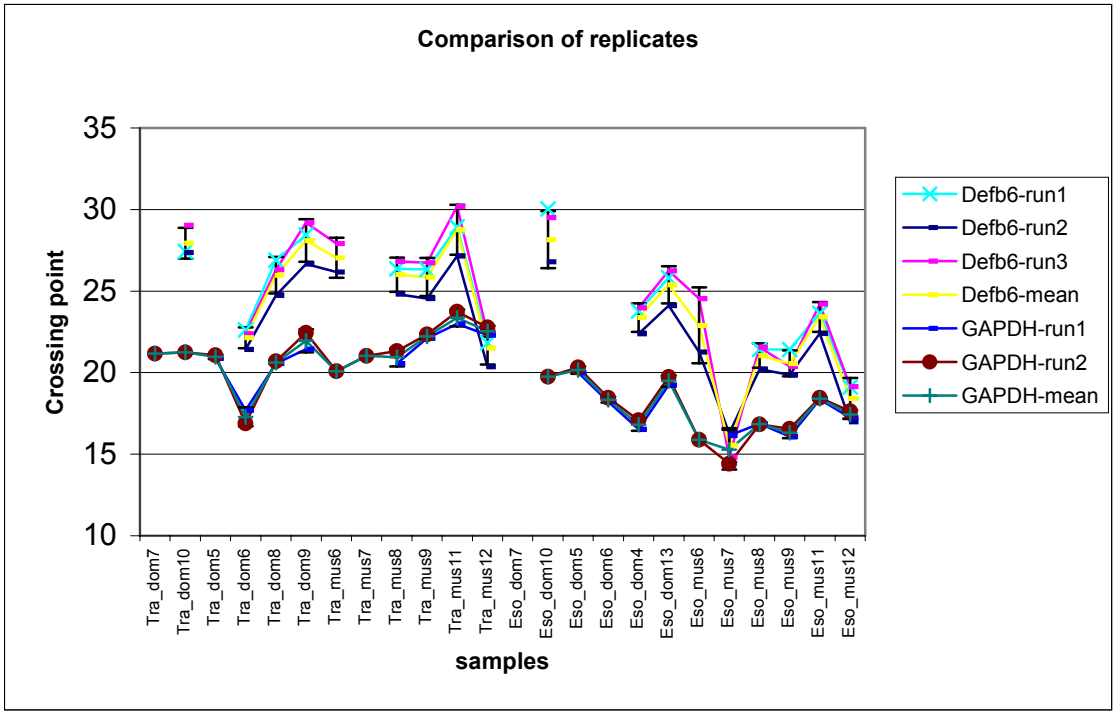



**Figure 4:** Comparisons of expression levels in replicate runs per sample of the ß-Defensin 6 and the GAPDH gene in trachea and esophagus. On the x-axes the organs and sample names are listed (Tra = trachea; Eso = esophagus).

For the two runs of the housekeeping gene the results are nearly identical and standard deviations are very small with a maximum of 1.2 in the sample Eso-mus 7. For the ß-Defensin 6 gene the standard deviations are higher with a maximum of 2.3 in sample Eso-mus 6. The behaviour of all replicates within each run is nearly the same: comparing the results within each run the relative differences between all samples remain the same. Only the crossing points are slightly different which is an effect of the manual choice of the crossing line. By analysing the behaviour of replicates within each run and between the runs, main differences are found between the samples indicating that the differences are sample specific and the results of each run are highly reproducible. The only exception is the sample Eso-mus 6 which exhibits different results in two runs. Generally, the standard deviation of the ß-Defensin 6 samples averaged over all samples within one run are between 3.2 to 4.1 and much higher than the standard deviation averaged over the replicate results per sample ranging from 0.74 to 2.3. The run conditions therefore seem to be very stable and differences are not dependent on PCR conditions but on the characteristics of the samples themselves.

The real-time PCR results of the other organs were based on single PCR runs only and are listed in Table 5. The housekeeping genes were expressed in all organs except for the esophagus sample from *Mus domesticus* 7 which showed no expression of any organ and no expression of TBP was found in the lung of *Mus domesticus* 8. GAPDH was expressed in this

sample. In the tongue ß-Defensin 6 was expressed in all samples; in the other organs ß-Defensin 6 was expressed in few animals only.

**Table 5:** Results of the expression of the ß-Defensin 6 gene in the different organs of *Mus musculus* and *Mus domesticus*. In the first columns the number of animals is listed which show expression or non-expression of the gene; in the last columns the average expression levels normalised with the two housekeeping genes and averaged over all animals from *Mus musculus* or *Mus domesticus* are listed.

| | | ***domesticus*** | ***musculus*** | **Chi-Square** | **Mean normalised expression level (GAPDH)** | | **Mean normalised expression level (TBP)** | |
|---|---|---|---|---|---|---|---|---|
| | | | | | dom | mus | dom | mus |
| **Trachea** | expr | 4 | 5 | P = 0.505 | M 0.41<br>Std 0.007 | M 0.044<br>Std 0.046 | M 0.63<br>Std 0.195 | M 0.63<br>Std 0.030 |
| | non-expr | 2 | 1 | | | | | |
| **Esophagus** | expr | 3 | 6 | P = 0.087 | M 0.386<br>Std 0.17 | M 0.432<br>Std 0.047 | M 0.638<br>Std 0.002 | M 0.687<br>Std 0.035 |
| | non-expr | 2 | 0 | | | | | |
| **Lung** | expr | 2 | 3 | P = 0.74 | M 0.433<br>Std 0.011 | M 0.469<br>Std 0.026 | M 0.603<br>Std 0.016 | M 0.63<br>Std 0.035 |
| | non-expr | 4 | 3 | | | | | |
| **Salivary Glands** | expr | 1 | 4 | P = 0.079 | M 0.524<br>Std / | M 0.472<br>Std 0.021 | M 0.649<br>Std / | M 0.603<br>Std 0.015 |
| | non-expr | 5 | 2 | | | | | |
| **Tongue** | expr | 6 | 6 | Not determined | M 0.451<br>Std 0.05 | M 0.48<br>Std 0.061 | M 0.673<br>Std 0.010 | M 0.69<br>Std 0.022 |
| | non-expr | 0 | 0 | | | | | |
| **Skeletal Muscle** | expr | 2 | 3 | P = 0.558 | M 0.402<br>Std 0.085 | M 0.262<br>Std 0.051 | M 0.63<br>Std 0.036 | M 0.59<br>Std 0.003 |
| | non-expr | 4 | 3 | | | | | |
| **total** | Expr | 18 | 27 | P = 0.039 | | | | |
| | Non-expr | 17 | 9 | | | | | |

The expression and non-expression of ß-Defensin 6 was tested for significant differences between the two taxa. For the organs separately none of the comparisons was significant although the expression patterns in the esophagus and the salivary glands were nearly significant. The total expression revealed a significant different expression with a lower expression of ß-Defensin 6 in *Mus domesticus*. No hint was found that the expression of ß-Defensin 6 is dependent on the individual because ß-Defensin 6 expression was not consistently switched on or off in the different organs of one individual.

To determine the level of expression the results were normalised with the two housekeeping genes. The mean expression levels per organ and per species was calculated by averaging the expressions of the samples where the gene showed expression.

Mann-Whitney-U test was applied to test whether the levels of expressions differed between the two lineages. The calculations were done separately for the two housekeeping genes. A significant result in expression level was detected for the esophagus samples: the *Mus domesticus* animals showed a significant lower expression level than the *Mus musculus* animals ($Z = -2.196$; $p = 0.028$ for GAPDH normalisation; $Z = -2.745$; $p = 0.006$ for TBP normalisation). All other organs showed no significant differences in expression levels.

## Discussion

### Confirming the selective sweep for the ß-Defensin 6 locus by a sequencing approach

In many studies the analysis of nucleotide sequences has been used to confirm the positive results of multi-locus microsatellite screens for traces of selection (Harr et al. 2002, Vigouroux et al. 2002). While Vigouroux et al. (2002) sequenced the promising sweep locus alone, Harr et al. (2002) additionally sequenced fragments in the flanking regions. In this study about selective sweeps in the house mouse quite extensive sequencing was performed along chromosome 8 in order to identify the sweep window, its size and the general fluctuation of nucleotide diversity along chromosome 8 within the mouse genome.
Summarising the findings over all sequenced fragments *Mus domesticus* exhibits a lower nucleotide diversity than *Mus musculus* which corresponds to the results of the microsatellite screen for the locus P104 located within the intron of the ß-Defensin 6 gene (chapter 2).
Additionally, by analysing the nucleotide diversity per fragment substantial fluctuation in diversity is obvious. Fluctuation of nucleotide diversity is a normal pattern along a recombining chromosome and high variability is expected under neutrality due to recombination (Kim & Stephan 2002). Therefore, it is straightforward to differentiate between regions with reduced variability due to selective sweeps and hitchhiking events, and sequences with reduced variability due to chance alone (Kim & Stephan 2002).

To verify the results of the microsatellite screen and to identify the size of the sweep window a higher proportion of fragments were sequenced around the ß-Defensin 6 locus, in contrast to less fragments in other regions located in a certain distance from the sweep locus. By comparing *Mus domesticus* against *Mus musculus* a window of reduced variability ranging from 12 kb upstream to 10 kb downstream of the ß-Defensin 6 locus was identified. The borders of the window are marked by fragments of higher nucleotide diversity of *Mus domesticus* versus *Mus musculus*. Microsatellite analysis identified this region as a potential sweep locus and the sequencing approach verified the results of reduced nucleotide diversity within the *Mus domesticus* lineage and even allowed to estimate the size of the sweep window to be between 22 kb and 30 kb. The estimated size of this sweep window corresponds to sweep window sizes in *Drosophila melanogaster* with 41 to 54 kb (Saez et al. 2003) and 28 kb (Harr et al. 2002). Other sequenced fragments along chromosome 8 revealed several patterns of reduced nucleotide diversity. The results were based on local sequence stretches of 500 to 800 bp while the results of the ß-Defensin 6 locus were based on sequencing 5 kb

within a region of 22 kb. Nevertheless, in the chromosomal region of 300 kb upstream of the ß-Defensin locus there are annotations of five putative genes which are derived by automated computational analysis using gene prediction methods such as BLAST or genome scan (information within NW_000383, Genbank Accession number for sequence stretch on chromosome 8). These sequence stretches, therefore, may contain functional regions such as unknown genes which might also be potential targets for selection. The fluctuation of nucleotide polymorphism upstream and downstream of the ß-Defensin 6 gene can also be the result of multiple selection events at different sites that have independently shaped the diversity along the chromosome.

As the size of the sweep window is influenced by many factors such as strength of selection, mutation rates, recombination and time passed since the selective sweep event it is very difficult to predict sizes of sweep windows (Kim & Stephan 2002, Fay & Wu in press). The advantage of nucleotide diversity is that the mutation rates of sequence stretches are more similar to each other and the estimation of the sweep window size is more reliable than with microsatellites which show high variations in mutation rates. Consequently, the search for selective sweeps by sequencing random fragments within the genome only the recombination rate of the relevant region has to be taken into account and the sequence fragments have to be distributed in equal centiMorgan distances. In contrast to this, by screening microsatellite loci the mutation rate has to be considered in addition. Due to their high variability microsatellite loci are better suited to identify recent sweeps but their traces will be obscured relatively fast (Schlötterer 2003). Sequence data are superior to identify ancient sweep events (Glinka et al. in press) as is the case for the ß-Defensin 6 locus which apparently has undergone a selective sweep event after the separation of the *Mus musculus* and the *Mus domesticus* lineage.

**Test for selection by comparisons with the outgroup *Mus spretus***

Although the microsatellite screen and the sequencing approach confirmed the reduced variability at the ß-Defensin 6 locus the multi-locus Hudson-Kreitman-Aguadé test was applied to check for selection. The HKA test compares the intraspecific nucleotide polymorphism within populations against the interspecific nucleotide divergence between two species. Under neutrality polymorphism and divergence should be equal because the same fragments are assumed to have evolved neutrally under the same mutation mechanism (Kimura 1983). *Mus musculus* and *Mus domesticus* were both tested against sequences of *Mus spretus* which diverged from the house mouse about 2 Mya ago (Bonhomme 1993). The

comparison of *Mus musculus* against *Mus spretus* followed neutral expectations while the comparison of *Mus domesticus* against *Mus spretus* significantly deviated from neutrality. For *Mus domesticus* almost all fragments showed higher deviations than the *Mus musculus* fragments and the $\chi 2$ test indicated that selection is acting on multiple fragments in this region. The highest deviations were found in the intraspecific comparisons indicating that the selective process has acted within *Mus domesticus*. Some fragments showed high deviations which are probably artificial results because they are based on elevated nucleotide diversity only due to one or two animals carrying different nucleotides. For locus SNP8-8 one *Mus domesticus* animals carried a *Mus musculus* sequence. This animal increased the general nucleotide polymorphism at this locus artificially and the sequence of this locus requires revaluation. SNP8-8down and SNP8-7down generally exhibit extremely high nucleotide diversities based on many heterozygote animals. Since in the other sequences heterozygote animals were rare, the results of these fragments have to be reanalysed for duplications of the fragments or for any other effects leading to this extremely high polymorphism.

Even after eliminating these fragments from the HKA-analysis the results remain significant (data not shown). The highest deviation is then found in the first part of the genomic sequence of the ß-Defensin 6 gene but also other loci especially in the downstream region of the ß-Defensin 6 gene exhibit high deviations. Starting with the ß-Defensin 4 gene 25 kb upstream to ß-Defensin 6 there is a cluster of ß-Defensin genes stretching over several 100 kb downstream (Maxwell et al 2003, Morrison et al 2003). Comparing the human and the mouse ß-Defensin cluster other studies found out that the cluster has evolved through duplication and then further developed under strong directional selection (Maxwell et al. 2003). Therefore, it is not an unexpected result if additionally to the ß-Defensin 6 gene other regions of the ß-Defensin cluster are shaped by selection.

Only *Mus domesticus* showed traces of selection within this cluster and none were found within *Mus musculus*. Compared to the *Mus musculus* populations the *Mus domesticus* populations are rather young (chapter 1). Maybe the cluster was reshaped again during the colonization of new habitats while within the old *Mus musculus* populations these traces became already obscured by mutation and recombination events.

**Search for the advantageous mutation**

Multiple evidence showed that the region around the ß-Defensin 6 gene was subject to selection within the *Mus domesticus* clade. Now the challenge was to find the advantageous mutation that has swept through the populations eliminating other polymorphisms. One characteristic of such an advantageous mutation is that it must occur in a functional region either in the amino acid coding sequence or in the regulatory regions. Simultaneously, alleles in non-functional regions become monomorphic just because of the hitchhiking effect. Sequencing of the two exons of the ß-Defensin 6 gene revealed no nucleotide substitutions. All sequenced animals from the *Mus musculus* and *Mus domesticus* lineages were completely monomorphic even at the silent positions of the amino acid coding sequence indicating that the advantageous mutation did not occur within the coding sequence.

By comparing the sequence against the *Mus spretus* animals evidence of positive selection was found by calculation the Ka/Ks ratios. If the number of replacement substitutions exceeds the number of synonymous substitutions this is usually an indication for positive Darwinian selection (Yang 2002). Generally Ka/Ks values of orthologous genes of mice and humans are around 0.18 and only 1% of all genes are known to result in Ka/Ks values above 1 and those genes are mainly involved in sexual selection or disease resistance (Fay & Wu (a), in press) which is also true for the ß-Defensin 6 gene. Morrison et al. (2003) analysed the ß-Defensin gene cluster in different rodent species and discovered that most of the genes evolved by duplication and subsequent positive selection. However, they were not able to amplify the second exon of ß-Defensin 6 in *Mus spretus*. In this study I succeeded to amplify and to sequence the whole genomic region of the ß-Defensin 6 gene in *Mus spretus*. For the amino acid coding sequences high polymorphism rates were detected which even revealed different amino acid exchanges within four *Mus spretus* animals. Polymorphism of the ß-Defensin peptides within one species is not reported in the literature (Morrison et al. 2003, Maxwell et al. 2003) probably because only one or few animals of the relevant species were sequenced.

The McDonald Kreitman test showed no significant deviation from neutral expectation and the genomic sequence within *Mus spretus* evolved without any selective constraints. Normally, functional genes are always under selective controls which are either positive or negative: positive selection is expected to increase the number of amino acid substitutions but with little impact on polymorphism while negative selection affects the amino acid polymorphism but not divergence (Fay & Wu (a), in press).

This result indicates that the high values of the Ka/Ks test are just a result of unconstraint evolution indicating that the gene in *Mus spretus* is not subjected to strong directional selection.

By blasting the exon sequences against the rat database only one coding sequence for the first exon to the ß-Defensin 2 precursor gene in the rat genome was found. The second exon showed a highly significant BLAST match but no coding region was annotated for this sequence which indicates that a similar sequence stretch exists in the rat genome but does not contain a functional gene.

This result corresponds well with the results of Morrison et al. (2003) who amplified the ß-Defensin 6 exons only in *Mus domesticus*, *Mus musculus*, *Mus castaneus* and in *Mus caroli*, but not in more distantly related rodents. In contrast to this, other ß-Defensin genes were found in many different rodent species indicating that those genes evolved early during the divergence time of the Genus *Mus*. Morrison et al. (2003) suggest that the ß-Defensin 3, ß-Defensin 5 and ß-Defensin 6 gene evolved by duplication of the ß-Defensin 4 gene four to six million years ago and were shaped by strong selection pressure.

*Mus caroli* is more distantly related to the house mouse species complex than *Mus spretus* and the ß-Defensin 6 gene showed eight amino acid changes which correspond to a very high divergence rate. In *Mus spretus* four inter- and intraspecific amino acid substitutions for the second exon were found but differed from the substitution in *Mus caroli*, thus indicating again that the ß-Defensin 6 gene evolved differently in *Mus spretus*.

Within the house mouse the gene presumably acquired novel physiological roles because of its special expression pattern in the skeletal muscle (Yamaguchi et al. 2001). As the gene is rather young and shows special features in the house mouse species complex, it is very likely that the gene is still being shaped by further selection within the different lineages of the house mouse species complex.

**Analysis of the regulatory regions**

The history of duplication of the ß-Defensin cluster and of the positive selection pressure shaping the whole gene family does not explain why a selective sweep window was identified around the ß-Defensin 6 gene for *Mus domesticus* and not for *Mus musculus*. Although the amino acid coding regions exhibited no differences between the lineages the hypothesis is that advantageous changes occurred in the regulatory regions and altered the expression patterns of the gene. Especially the regulatory regions of duplicated genes are under strong selective

control as reviewed by Ohta (2003), because strict regulation is often an important factor for the survival of a duplicated gene.

Therefore, all *Mus musculus* and *Mus domesticus* sequences were analysed for possible transcription factor binding sites. This approach was based on the concept of phylogenetic footprinting that assumes that functional regions are under selective constraints and are more conserved between two species than non-functional regions (Lenhard et al. 2003, Zhang & Gerstein 2003). The house mouse sequences were aligned to *Mus spretus*. Within the aligned sequences several conserved stretches between the species and about 45 putative transcription factor binding sites were identified. To detect the advantageous mutation the analysis concentrated on the differences between *Mus musculus* and *Mus domesticus*. The most important difference was that *Mus musculus* was mainly polymorphic for the presence or absence of transcription factor binding sites while *Mus domesticus* was more monomorphic showing either presence or absence of the same binding site in all sequences. The regulation of transcription in *Mus domesticus* seemed to be more strictly controlled as a possible response to a selection pressure which caused the precise activation of the ß-Defensin 6 gene.

This interpretation follows the idea that a selective sweep is a response to a changing environment which suddenly favours one allele of an existing polymorphism. Another theory about selective sweeps assumes that a new advantageous mutation occurs within a population which suddenly changes its properties without changes in the environment. Such mutations are recognized as derived variants.

For *Mus domesticus* 17 fixed derived alleles were identified within the ß-Defensin 6 region. The transcription factor analysis showed that eleven of all these alleles occurred in close vicinity to putative transcription factor binding regions and can thus play a role in the altered regulation of the expression of the gene.

Phylogenetic footprinting allows the identification of possible functional regions but, so far, the program ConSite is only able to detect about 68% of experimentally confirmed transcription factor binding sites (Lenhard et al. 2003). These computational approaches always produce a lot of false positive results and the identification of a transcription factor binding site does not always correspond to a transcription factor binding site in vivo (Zhang & Gerstein 2003).

Although no differences were found in the coding sequence between *Mus musculus* and *Mus domesticus* quite a lot of differences appeared in potential transcription factor binding sites indicating that the advantageous mutation which caused the selective sweep can be one of the single nucleotide changes that altered possible regulatory sites. Of particular interest are those sites where *Mus domesticus* exhibits fixed derived alleles in high density: for example around the position 2828 three fixed derived variants were found that led to the complete loss of the Sox-5 binding site and this site is definitely an excellent candidate for an advantageous mutation.

**Differential expression of ß-Defensin 6**

Gene regulation within an organism is one of the most complicated processes responsible for the ontogenetic development of individuals; on a broad scale the evolution of those networks is probably a major contributor to animal diversity (Rast 2003). Disturbances of the regulatory system can lead to severe malfunctions and diseases and, therefore, the regulatory system has to be under stringent control. Gradually the understanding of the architecture of these regulatory networks is coming into the focus of research (Rast 2003).
On the other hand flexible regulation of genes allows the organism to respond quickly to the environment and especially genes that play a role in environmental interactions are often inducible genes and expressed by certain signals only.

The ß-Defensin cluster consists of genes that are differentially regulated: some genes are just expressed if they are induced by certain pathogens while other genes are constitutively expressed (Bals et al. 1999, Jia et al. 2000, Yamaguchi et al. 2001, Burd et al. 2002). The ß-Defensin 6 gene is known to be expressed in the esophagus, tongue, trachea and skeletal muscle and is inducible expressed in the lung (Yamaguchi et al. 2001). As I assume that the selective sweep of the ß-Defensin 6 gene within *Mus domesticus* is based on differential regulation the expression level of the gene in various organs of the two lineages *Mus domesticus* and *Mus musculus* were analysed. Generally, less expression of ß-Defensin 6 was found in *Mus domesticus*. This result would correspond nicely to the fact that within *Mus domesticus* the potential transcription factor binding sites are either present or absent but not as polymorphic as in the *Mus musculus* samples. Subsequently, I assume that the expression is more specifically regulated in *Mus domesticus*. Concerning the different organs it was found that ß-Defensin was constitutively expressed in the tongue of all animals while in the other organs ß-Defensin expression was found in some animals only. For the esophagus the

expression level was found to be slightly significantly lower in *Mus domesticus* than in *Mus musculus*.
Still, the results of the expression level of the ß-Defensin 6 gene in *Mus domesticus* and *Mus musculus* showed a lot of variance and no clear differences were found in the regulation of the gene in the two lineages.

The problem of measuring gene expression levels is that gene expression can deviate substantially depending on the state of the animal. For the quantitative PCR experiments wild caught animals were used. Although kept in the lab for 48 hours the state of the animals remained unclear: no information could be obtained about age, state of health and general physiological conditions. Particularly, for a gene as the ß-Defensin gene the expression can differ significantly dependent on any kind of bacterial infections. Therefore, to really get an answer whether the expression of the ß-Defensin 6 gene is differentially regulated within the species the experiment must be repeated with lab raised *Mus musculus* and *Mus domesticus* animals with well known status. Additionally, I expect differences in the regulation of the gene which might be visible only if the animals are subjected to bacterial infections. Therefore, a final answer whether the gene is differentially regulated can only be inferred after testing the response of lab raised animals of the two lineages to bacterial infections under controlled conditions.

**Summary and Conclusion**

Sequencing several fragments along chromosome 8 revealed that the region around the ß-Defensin 6 gene was shaped by selection within *Mus domesticus*. Other regions, in particular, regions within the ß-Defensin cluster also showed traces of selection which are either caused by an extended hitchhiking event or by several independent selective sweep events. Comparisons of the coding sequences revealed no differences between *Mus musculus* and *Mus domesticus*. Nevertheless, all ß-Defensin genes have evolved under strong positive selection within rodents and, therefore, are promising targets for further selection events. The ß-Defensin 6 gene is a rather young gene within the house mouse species complex. Although no changes were found in the coding regions the analysis of potential transcription factor binding sites revealed many differences between *Mus musculus* and *Mus domesticus* mainly resulting in monomorphic occurrence of transcription factor binding sites within the *Mus domesticus* clade. Although selective changes in regulatory regions are difficult to proof the analysis of the expression levels of the gene within different organs revealed significant

differences independent of the unknown state of the animals which certainly increased the variance of the results and reduced the chance to detect significant differences.

My search for a selective sweep event within the house mouse complex identified the ß-Defensin 6 gene as a sweep locus. Although the functional differences are not completely resolved, this study showed that the method is successful to determine an interesting candidate gene which was already identified by other research groups to play a role in evolution and adaptation at a higher taxonomic level.

## Literature


- Auffray JC, Marshall, JT, Thaler L, Bonhomme F (1990) Focus on the nomenclature of European species of Mus. Mouse Genome. 88: 7-8.

- Auffray, JC, Tchernov E, Nevo E (1988) Origine du commensalisme de la souris domestique (Mus musculus domesticus) vis-à-vis de l'homme. C.R. Academy de Science Paris 307 (série III): 517-22.

- Baer CF (1999) Among-locus variation in Fst: fisch, allozymes and the Lewontin-Krakauer Test revisited. Genetics 152: 653-659.

- Balloux F, Lugon-Moulin N (2002) The estimation of population differentiation with microsatellite markers. Mol Ecol. 11(2): 155-65.

- Bals R, Wang X, Meegalla RL, Wattler S, Weiner DJ, Nehls MC, Wilson JM (1999) Mouse beta-defensin 3 is an inducible antimicrobial peptide expressed in the epithelia of multiple organs. Inf. Imm. 67(7): 3542-3547.

- Benson G. (1999) Tandem repeats finder: a program to analyse DNA sequences. Nucleic Acids Research. 27(2): 573-580.

- Berry RJ, Bronson FH (1992) Life histroy and bioeconomy of the house mouse. Biological Reviews. 67: 519-550.

- Boehm U, Klamp T, Groot M, Howard JC (1997) Cellular response to Interferon-γ. Annual Revies of Immunology 15: 749-795.

- Bonhomme F, Anand R, Darviche D, Din W, Boursot P (1994) The house mouse as a ring species? Pp. 13-23 in Genetics in Wild Mice (Moriwaki K, et al eds) Japan Science Society Press, Tokyo.

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

- Boursot P, Din W, Anand R, Darviche D, Dod B, VonDeimling F, Talwar GP, Bonhomme F (1996) Origin and radiation of the house mouse: Mitochondrial DNA phylogeny. Journal of Evolutionary Biology. 9 (4): 391-415.

- Bowcock AM, Ruiz-Linares A, Tomfohrde J (1994) High resolution of human evolutionary trees with polymorphic microsatellites. Nature. 386: 455-457.

- Burd RS, Furrer JL, Sullivan J, Smith AL (2002) Murine beta-defensin-3 is an inducible peptide with limited tissue expression and broad-spectrum anti-microbial activity. Shock. 18(5): 461-4.

- Burk RF, Hill KE, Motley AK (2003) Selenoprotein metabolism and function: evidence for more than one function for selenoprotein P. J Nutr.133(5 Suppl 1): 1517-20.

- Cau E, Gradwohl G, Fode C, Guillemot F (1997) Mash1 activates a cascade of bHLH regulators in olfactory neuron progenitors. Development.124(8): 1611-21.

- Copeland, N.G., D.J. Gilbert, N.A. Jenkins, J.H. Nadeau, J.T. Eppig, L.J. Maltais, J.C. Miller, W.F. Dietrich, R.G. Steen, S.E. Lincoln, A. Weaver, D.C. Joyce, M. Merchant, M. Wessel, H. Katz, L.D. Stein, M.P. Reeve, M.J. Daly, R.D. Dredge, A. Marquis, N. Goodman, E.S. Lander (1993) Genome Maps IV. Science 262: 67.

- Crow, J.F. (1993) How much do we know about spontaneous human mutation rates? Environ. Mol. Mutagen. 21: 122–129.

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

- Dietrich, W.F., J. Miller, R. Steen, M.A. Merchant, D. Damron-Boles, Z. Husain, R. Dredge, M.J. Daly, K.A. Ingalls, T.J. O'Conner, C.A. Evans, M.M. DeAngelis, D.M. Levinson, L. Kruglyak, N. Goodman, N.G. Copeland, N.A. Jenkins, T.L. Hawkins, L. Stein, D.C. Page, & E.S. Lander (1996) A comprehensive genetic map of the mouse genome. Nature 380: 149-152.

- Dietrich, W.F. et al. (1994) A genetic map of the mouse with 4,006 simple sequence length polymorphisms. Nature Genetics 7: 220-245.

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

- Duplantier JM, Orth A, Catalan J, Bonhomme F (2002) Evidence for a mitochondrial lineage originating from the Arabian peninsula in the Madagascar house mouse (Mus musculus). Heredity 89: 154-158.

- Ellegren H (2000) Microsatellite mutations in the germline: implications for evolutionary inference. Trends Genet. 16(12):551-8.

- Enard W, Przeworski M, Fisher SE, Lai CSL, Wiebe V, Kitano T, Monaco AP, Pääbo S (2002) Molecular evolution of FOXP2, a gene involved in speech and language. Nature. 418: 869-872.

- Fay, J. C. and C.-I Wu. Detecting hitchhiking from patterns of DNA polymorphism. Preprint of book chapter available at: http://crimp.lbl.gov/pubs.html .

- Fay, J. C. and C.-I Wu (a). Sequence divergence, functional constraint and selection in protein evolution. Preprint of book chapter available at: http://crimp.lbl.gov/pubs.html

- Fay JC, Wu CI (2000) Hitchhiking under positive Darwinian selection. Genetics. 155: 1405-1413.

- Fay JC, Wu CI (2001) The neutral theory in the genomic era. Curr Opin Genet Dev. 11(6): 642-6.

- Fay JC, Wyckoff GJ, Wu CI (2001) Positive and negative selection on the human genome. Genetics. 158: 1227-1234.

- Felsenstein, J., 1993. PHYLIP (Phylogeny Inference Package) version 3.5c. Distributed by the author. Department of Genetics, University of Washington, Seattle.

- Ferris SD, Sage RD, Prager EM, Ritte U, Wilson AC (1983) Mitochondrial DNA Evolution in mice. Genetics. 105: 681-721.

- Fossella J, Samant SA, Silver LM, King SM, Vaughan KT, Olds-Clarke P, Johnson KA, Mikami A, Vallee RB, Pilder SH (2000) An axonemal dynein at the Hybrid Sterility 6 locus: implications for t haplotype-specific male sterility and the evolution of species barriers. Mamm Genome. 11(1): 8-15.

- Fu YX, LI WH. (1993). Statistical tests of neutrality of mutations. Genetics 133: 693-709

- Gilad Y, Rosenberg, S, Przeworski M, Lancet D, Skorecki K (2002) Evidence for positive selection and population structure at the human MAO-A gene. Proc Natl Acad Sci U S A. 99(2): 862-867.

- Glinka S, Ometto L, Mousset S, Stephan W, De Lorenzo D (2003) Demography and natural selection have shaped genetic variation in Drosophila melanogaster: a multi-locus approach. Genetics. In press.

- Goldstein DB, Linares AR, Cavalli-Sforza LL, Feldman MW (1995a) An evaluation of genetic distances for use with microsatellite loci. Genetics. 139: 462-471.

- Goldstein DB, Linares AR, Cavalli-Sforza LL, Feldman MW (1995b). Genetic absolute dating based on microsatellites and the origine of modern humans. Proc Natl Acad Sci USA. 92: 6723-6727.

- Goudet J (1995). Fstat version 1.2: a computer program to calculate Fstatistics. Journal of Heredity. 86(6): 485-486.

- Grant PR, Grant BR (2002) Unpredictable evolution in a 30-year study of Darwin's finches. Science. 296(5568): 707-11.

- Guenet JL, Bonhomme F (2003) Wild mice: an ever-increasing contribution to a popular mammalian model. Trend in Genetics. 19 (1): 24-31.

- Harr B, Kauer M, Schlotterer C (2002) Hitchhiking mapping: a population-based fine-mapping strategy for adaptive mutations in Drosophilamelanogaster. Proc Natl Acad Sci U S A. 99(20):12949-54.

- Hauffe HC, Pialek J, Searle JB (2000) The house mouse chromosomal hybrid zone in Valtellina (SO): a summary of past and present research. Hystrix. 11(2):17-25.

- Hudson RR, Kreitman M, Aguade M (1987). A test of neutral molecular evolution based on nucleotide data. Genetics 116: 153-159.

- Hunt WG, Selander RK (1973) Biochemical genetics od hybridisation in European house mice. Heredity. 31(1): 11-33.

- Jia HP, Wowk SA, Schutte BC, Lee SK, Vivado A, Tack BF, Bevins CL, McCray PB (2000) A novel murine beta-defensin expressed in tongue, esophagus and trachea. J Biol Chem. 275(43): 33314-33320.

- Kaplan NL, Hudson RR, Langley CH (1989) The "hitchhiking effect" revisited. Genetics. 123: 887-899.

- Karn RC, Laukaitis CM (2003) Characterization of two forms of mouse salivary androgen-binding protein (ABP): implications for evolutionary relationships and ligand-binding function. Biochemistry. 42(23): 7162-70.

- Kauer M, Zangerl B, Dieringer D, Schlotterer C (2002) Chromosomal patterns of microsatellite variability contrast sharply in African and non-African populations of Drosophila melanogaster. Genetics. 160(1): 247-56.

- Kauer M, Dieringer D, Schlötterer C (2003) A microsatellite variability screen for positive selection associated with the 'out of Africa' habitat expansion of Drosophila melanogaster. Genetics: in press.

- Kayser M, Brauer S, Stoneking M (2003) A genome scan to detect candidate regions influenced by local natural selection in human populations. Mol Biol Evol. 20(6): 893-900.

- Kim Y, Stephan W (2002) Detecting a local signature of genetic hitchhiking along a recombining chromosome. Genetics. 160(2): 765-77.

- Kimura M (1983) The neutral theory of molecular evolution. Cambridge University Press, Cambridge, UK.

- Kohn MH, Pelz HJ, Wayne RK (2000) Natural selction mapping of the warfarin-resistance gene. Proc Natl Acad Sci USA. 97(14): 7911-7915.

- Kohn MH, Pelz HJ, Wayne RK (2003) Locus-specific genetic differentiation at Rw among warfarin-restistant rat (rattus norvegicus) populations.Genetics. 164: 1055-1070.

- Konkel DA, Maizel JV Jr, Leder P (1979) The evolution and sequence comparison of two recently diverged mouse chromosomal beta--globin genes. Cell. 18(3): 865-73.

- Kumar S, Tamura K, Jakobsen IB, Nei M (2001) MEGA2: Molecular Evolutionary Genetics Analysis software, Arizona State University, Tempe, Arizona, USA.

- Lenhard B, Sandelin A, Mendoza L, Engstrom P, Jareborg N, Wasserman WW (2003) Identification of conserved regulatory elements by comparative genome analysis. J Biol. 2(2): 13.

- Markstein M, Levine M (2002) Decoding cis-regulatory DNAs in the Drosophila genome. Curr Opin Genet Dev. 12(5): 601-6.

- Maxwell AI, Morrison GM, Dorin JR (2003) Rapid sequence divergence in mammalian beta-defensins by adaptive evolution. Mol Immunol. 40(7): 413-21.

- Maynard Smith J, Haigh J (1974) The hitchhiking effect of a favorable gene. Genet. Res. 23: 23-35.

- McDonald JH, Kreitman M (1991). Adaptive protein evolution at the Adh locus in Drosophila. Nature 351: 652-654.

- Metsaranta M, Toman D, de Crombrugghe B, Vuorio E (1991) Mouse type II collagen gene. Complete nucleotide sequence, exon structure, and alternative splicing. J Biol Chem. 266(25): 16862-9.

- Minch, E., Ruiz-Linares, A., Goldstein, D. B. et al., 1995 Microsat (version 1.5d): a program for calculating statistics on microsatellite allele data.

- Morrison GM, Rolfe M, Kilanowski FM, Cross SH, Dorin JR (2002) Identification and characterization of a novel murine beta-defensin- related gene. Mamm. Genome. 13(8): 445-51.

- Morrison GM, Semple CA, Kilanowski FM, Hill RE, Dorin JR (2003) Signal sequence conservation and mature peptide divergence within subgroups of the murine beta-defensin gene family. Mol Biol Evol. 20(3): 460-70.

- Moustafa ME, Kumaraswamy E, Zhong N, Rao M, Carlson BA, Hatfield DL(2003) Models for assessing the role of selenoproteins in health.J. Nutr. 133.(7 Suppl): 2494-2496.

- Munclinger p, Bozikova E, Sugerkova M, Pialek J, Macholan M (2002) Genetic variation in house mice (Mus, Muridae, Rodentia) from Czech and Slovak Republics. Folia Zoologica. 51(2): 81-92.

- Nachman MW, Hoekstra HE, D'Agostino SL (2003) The genetic basis of adaptive melanism in pocket mice. Proc Natl Acad Sci U S A. 100(9): 5268-73.

- Nair S, Williams JT, Brockman A, Paiphun L, Mayxay M, Newton PN, Guthmann JP, Smithuis FM, Hien TT, White NJ, Nosten F, Anderson TJ (2003) A selective sweep driven by pyrimethamine treatment in southeast asian malaria parasites. Mol Biol Evol. 20(9):1526-36.

- Nei M (1972) Genetic distances between populations. Am Nat. 106: 283-292.

- Nei M (1978) Molecular evolutionary genetics. Columbia University Press, New York.

- Nekrutenko A, Makova KD, Li WH (2001) The KA/KS ratio test for assessing the protein-coding potential of genomic regions: an empirical simulation study. Genome Research 12: 198-202

- Nurminsky DI, Nurminskaya MV, De Aguiar D, Hartl DL (1998) Selective sweep of a newly evolved sperm-specific gene in Drosophila. Nature. 396: 572-575.

- Ohta T, Kimura M (1973) A model of mutation appropriate to estimate the number of electrophoretically detectable alleles in a finite population. Genet. Res. 22: 201-204.

- Ohta T (2003) Evolution by gene duplication revisited: differentiation of regulatory elements versus proteins. Genetica. 118(2-3): 209-16.

- Paetkau D, Waits LP, Clarkson PL, Craighead L, Strobeck C (1997) An empircal evaluation of genetic distance statistics using microsatellite data from bear (Ursidae) populations. Genetics. 147: 1943-1957.

- Page RDM (1996) TREEVIEW: An application to display phylogenetic trees on personal computers. Computer Applications in the Biosciences 12: 357-358.

- Park, SDE (2001) Trypanotolerance in West African Cattle and the Population Genetic Effects of Selection (Ph.D. thesis). University of Dublin.

- Payseur BA, Cutter AD, Nachman MW (2002) Searching for evidence of positive selection in the human genome using patterns of microsatellite variability. Mol Biol Evol. 19(7): 1143-53.

- Prager EM, Sage RD, Gyllensten U, Thomas WK, Hubner R, Jones CS, Noble L, Searle JB, Wilson AC (1993). itochondrial-DNA sequence diversita and the colonization of scandinavia by house mice from East Holsten. Biol. Journ. Linn. Soc. 50(2): 85-122

- Prager EM, Orrego C, Sage RD (1998) Genetic variation and phylogeography of central Asian and other house mice, including a major new mitochondrial lineage in Yemen. Genetics. 150(2): 835-61.

- Prager EM, Tichy H, Sage RD (1996) Mitochondrial DNA sequence variation in the eastern house mouse, Mus musculus: comparison with other house mice and report of a 75-bp tandem repeat. Genetics. 143(1): 427-46.

- Presgraves DC, Balagopalan L, Abmayr SM, Orr HA (2003) Adaptive evolution drives divergence of a hybrid inviability gene between two species of Drosophila. Nature. 423(6941): 715-9.

- Qiu P (2003) Recent advances in computational promoter analysis in understanding the transcriptional regulatory network. Biochem Biophys Res Commun. 309(3): 495-501.

- Rast JP (2003) Development gene networks and evolution. J Struct Funct Genomics. 3(1-4): 225-34.

- Richard M, Thorpe RS (2001) Can microsatellites be used to infer phylogenies? Evidence from population affinities of the western Canary Island lizard (Gallotia galloti). Mol Phyl Ecol. 20(3): 351-360.

- Rozas J, Rozas R (1999). DNASP version3: an integrated program for population genetics and molecular evolution analysis. Bioinformatics 15: 174-175.

- Rieseberg LH, Widmer A, Arntz M, Burke JM (2002) Directional selection is the primary cause of phenotypic diversification. Proc Natl Acad Sci USA. 99(19) 12242-12245.

- Ruiz ML, London B, Nadal-Ginard B (1996) Cloning and characterization of an olfactory cyclic nucleotide-gated channel expressed in mouse heart. J Mol Cell Cardiol. 28(7):1453-61.

- Sabeti PC, Reich DE, Higgins JM, Levine HZP, Richter DJ, Schaffner SF, Gabriel SB, Platko JV, Patterson NJ, McDonal GJ, Ackerman HC, Campbell SJ, Altshuler D, Cooper R, Kwiatkowski D, Ward R, Lander ES (2002) Detecting recent positive selection in the human genome from haplotype structure. Nature. 419: 832-837.

- Saez AG, Tatarenkov A, Barrio E, Becerra NH, Ayala FJ (2003) Patterns of DNA sequence polymorphism at Sod vicinities in Drosophila melanogaster: unraveling the footprint of a recent selective sweep. Proc Natl Acad Sci U S A. 100(4): 1793-8.

- Sage RD (1981) Wild mice, pp 39-90 in The mouse in bimedical research, Vol 1 (Foster HL, Small JD, Fox, JG, eds) Academic Press, New York.

- Sage RD, Atchley WR, Capanna E (1993) House Mice as Models in Systematic Biology Systematic Biology. 42(4): 523-561.

- Schlötterer C, Vogl C, Tautz D (1997) Polymorphism and locus-specific effects on polymorphism at microsatellite loci in natural Drosophila melanogaster populations. Genetics. 146(1): 309-20.

- Schlötterer C. et al. (1998) High mutation rate of a long microsatellite allele in Drosophila melanogaster provides evidence for allele-specific mutation rates. Mol. Biol. Evol. 15: 1269–1274

- Schlötterer C (2001) Genealogical inference of closely related species based on microsatellites. Genet Res. 78: 209-212.

- Schlötterer C (2002a) Towards a molecular characterization of adaptation in local populations. Curr Opin Genet Dev. 12(6): 683-7.

- Schlötterer C (2002b) A microsatellite-based multilocus screen for the identification of local selective sweeps. Genetics. 160(2):753-63.

- Schlötterer C (2003) Hitchhiking mapping – functional genomics from the population genetic perspective. Trends in Genetics. 19(1): 32-38.

- Schneider S, Roessli D, Excoffier L (2000) Arlequin: A software for population genetics data analysis. Ver 2.000. Genetics and Biometry Lab, Dept. of Anthropology, University of Geneva.

- Schug MD, Mackay TF, Aquadro CF (1997) Low mutation rates of microsatellite loci in Drosophila melanogaster. Nat Genet. 15(1): 99-102.

- Schug MD, Hutter CM, Wetterstrand KA, Gaudette MS, Mackay TF, Aquadro CF (1998) The mutation rates of di-, tri-and tetranucleotide repeats in Drosophila melanogaster. Mol Biol Evol. 15(12): 1751-60.

- Schug MD, Hutter CM, Noor, MAF, Aquadro CF (1998). Mutation and evolution of microsatellite in Drosophila melanogaster. Genetica. 102/103: 359-367.

- Singleton GR(1983) The social and genetic structure of a natural colony of house mice, Mus musculus, at Healesville Wildlife Sanctuary. Australian Journal of Zoology. 31: 155-66.

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

- Smit, AFA & Green, P. RepeatMasker at http://ftp.genome.washington.edu/RM/RepeatMasker.html

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

- Talley HM, Laukaitis CM, Karn RC (2001) Female preference for male saliva: implications for sexual isolation of Mus musculus subspecies. Evolution Int J Org Evoluion. 55(3): 631-4.

- Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, Agarwal P, Agarwala R, Ainscough R, Alexandersson M, An P, Antonarakis SE, Attwood J, Baertsch R, Bailey J, Barlow K, Beck S, Berry E, Birren B, Bloom T, Bork P, Botcherby M, Bray N, Brent MR, Brown DG, Brown SD, Bult C, Burton J, Butler J, Campbell RD, Carninci P, Cawley S, Chiaromonte F, Chinwalla AT, Church DM, Clamp M, Clee C, Collins FS, Cook LL, Copley RR (2002) Initial sequencing and comparative analysis of the mouse genome. Nature. 420(6915): 520-62.

- Weber JL, Wong C (1993) Mutation of short tandem repeats. Hum. Mol. Genet. 2: 1123-1128.

- Wiehe T (1998) The effect of selective sweeps on the variance of the allele distribution of a linked multiallele locus: hitchhiking of microsatellites. Theor Popul Biol. 53(3): 272-83.

- Wooton JC, Feng X, Ferdig MT, Cooper RA, Mu J, Baruch DI, Magill AJ, Su X (2002) Genetic diversity and chloroquine selective sweeps in Plasmodium falciparum. Nature. 418: 320-323.

- Vigouroux Y, McMullen M, Hittinger CT, Houchins K, Schulz L, Kresovich S, Matsuoka Y, Doebley J (2002) Identifying genes of agronomic importance in maize by screening microsatellites for evidence of selection during domestication. Proc Natl Acad Sci USA. 99: 9650-9655.

- Yamaguchi Y, Fukuhara S, Nagase T, Tomita T, Hitomi S, Kimura S, Kurihara H, Ouchi Y (2001) A novel mouse beta-defensin, mBD-6, predominantly expressed in skeletal muscle. J Biol Chem. 276(34): 31510-4.

- Yan G, Chadee DD, Severson DW (1998) Evidence for genetic hitchhiking effect associated with insecticide resistance in Aedes aegypti. Genetics. 148: 793-800.

- Yang Z (2002) Inference of selection from multiple species alignments. Curr Opin Genet Dev. 12(6): 688-94.

- Yonekawa H, Takahama S, Gotoh O, Miyashita N, Moriwaki K (1994) Genetic diversity and geographic distribution of Mus musculus subspecies based on the polymorphism of mitochondrial DNA. Pp 25-40 in Genetics in Wild Mice (Moriwaki K, et al eds) Japan Science Society Press, Tokyo.

- Zhang Z, Gerstein M (2003) Of mice and men: phylogenetic footprinting aids the discovery of regulatory elements. J Biol. 2(2): 11.

- Zhang LH, Liu DP, Liang CC (2003) Finding regulatory sequences. Int J Biochem Cell Biol. 35(1): 95-103.

## Online Citations of programs and databases:

- ConSite: http://www.phylofoot.org/

- HKA-test: http://lifesci.rutgers.edu/~heylab

- RepeatMasker: http://ftp.genome.washington.edu/RM/RepeatMasker.html

- FPCR: http://www.biocenter.helsinki.fi/bi/bare-1_html/manual.htm

- NCBI: http://www.ncbi.nlm.nih.gov/

- ENSEMBL http://www.ensembl.org/

- Whitehead Institute: http://www-genome.wi.mit.edu/cgi-bin/mouse/index

# 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 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 Herrn Prof. Dr. Diethard Tautz betreut worden.

Sonja Ihle

# Lebenslauf


| | |
|---|---|
| **Name:** | Sonja Ihle |
| **Anschrift:** | Kölnstraße 159<br>53111 Bonn |
| **Geburtstag, und –ort:** | 29.10.1973 in Augsburg |
| **Staatsangehörigkeit:** | Deutsch |


## Schulausbildung:

| | |
|---|---|
| 1980 – 1982 | Gemeinschaftsgrundschule Kaiserslautern |
| 1982 – 1984 | Gemeinschaftsgrundschule Neunkirchen |
| 1984 – 1993 | Antoniuskolleg Neunkirchen |

## Auslandsaufenthalt:

| | |
|---|---|
| Juni 1993 – Januar 1994 | Frankreich: internationales Workcamp des christlichen Friedensdienstes |

## Hochschulausbildung:

| | |
|---|---|
| SS 1994 | Studium der Medizin an der Rheinischen Friedrich-Wilhelms-Universität Bonn |
| WS 1994 – WS 1996 | Grundstudium der Biologie an der Rheinischen Friedrich-Wilhelms-Universität Bonn |
| Februar 1997 – Januar 1998 | Auslandsstudium an der University of New South Wales, Sydney, Australien |
| SS 1998 – SS 2000 | Hauptstudium der Biologie an der Rheinischen Friedrich-Wilhelms-Universität Bonn |
| Juni 2000 | Abschluss des Diploms in Biologie am Institut für Zoologie, Abteilung Ethologie bei Frau Prof. Dr. Rasa, Rheinische Friedrich-Wilhelms-Universität Bonn |
| August 2000 - Februar 2004 | Promotion am Institut für Genetik, Lehrstuhl für Evolutionsgenetik bei Herr Prof. Dr. Tautz, Universität zu Köln |
| Februar 2004 | Vorraussichtlicher Abschluss der Promotion |