Tracing Signatures of Positive Selection in Natural Populations of the House Mouse

Inaugural – Dissertation

Zur Erlangung des Doktorgrades
der Mathematisch-Naturwissenschaftlichen Fakultät
der Christian-Albrechts-Universität zu Kiel

vorgelegt von

Anna Büntge

Plön, 2010

Berichterstatter: **Prof. Dr. Diethard Tautz**

Prof. Dr. Hinrich Schulenburg

Tag der letzten mündlichen Prüfung: **13. Juli 2010**

Zum Druck genehmigt: **Kiel,**

Der Dekan

List of Contents

List of Figures

List of Tables

List of Supplement

List of Digital Supplement

Chapter 2

Chapter 3-4

Zusammenfassung

Die Aufklärung der genetischen Grundlagen evoutionärer Anpassung steht im Fokus vieler evolutionsbiologischer Studien. Besonders wichtig ist hierbei die Identifizierung genomischer Regionen, die Hinweise auf natürliche Selektion zeigen; dies gestattet die Abschätzung wichtiger Selektionsparameter und identifiziert Gene, die an Adaptationsprozessen beteiligt sind. Die vorliegende Studie beschäftigt sich mit der Detektion positiver Selektionsereignisse im Genom der Hausmaus (*Mus musculus*). Zwei grundlegend unterschiedliche Herangehensweisen wurden in dieser Arbeit angewendet.

Bei der sogenannten ‚Kandidatengen-Analyse' werden Gene auf Selektionsmerkmale getestet, bei denen aufgrund ihrer Funktionalität vorausgesetzt wird mit einem bestimmten Phänotyp assoziiert zu sein. Dieser Ansatz wurde im ersten Teil der Arbeit benutzt, um Entgiftungsgene auf Adaptation zu untersuchen. Als Kandidaten wurden hierzu Mitglieder der Cytochrom P450 (Cyp450) Genfamilie gewählt, die bekanntermaßen eine zentrale Rolle beim Abbau von Umweltgiften einnehmen. Zwei Populationen aus unterschiedlichen ökologischen Zusammenhängen (und damit unterschiedlichen Ernährungsbedingungen) sollten Hinweise auf Spuren von Anpassung in dieser Gengruppe liefern. Insgesamt konnten drei Cyp450 Gene identifiziert werden, die deutliche Merkmale positiver Selektion tragen. Diese zeigten sowohl Hinweise auf Veränderung an *cis*-regulierende Elemente, wie auch in der Proteinsequenz. Die Gene variieren stark in ihrem evolutionsgeschichtlichen Alter, was darauf hiweist, dass jüngst erfolgte positive Selektion sowohl junge als auch alte Gene betreffen kann; ein Indiz für stetige Anpassung. Auffällig ist, dass alle drei Gene in der gleichen Population als unter Selektion stehend identifiziert wurden. Das Ergebnis bekräftigt die Annahme, dass diese Population auf einen ernährungsbedingten Selektionsdruck reagiert. Die Tatsache, dass die Gene auf unterschiedlichen Chromosomen liegen, kann nicht nur als unabhängige Bestätigung adaptiver Prozesse gesehen werden, sondern deutet auch auf mehrfach unabhängige Selektionsereignisse in der Cytochrom P450 Genfamilie hin.

Da die Gene in der oben beschriebenen Analyse auf Grund bestimmter Voraussagen *a priori* ausgewählt werden, kann diese nur zur Validierung von Annahmen genutzt werden. ‚Unerwartete' Bereiche im Genom oder Gene, die mit

komplexen Phänotypen verknüpft sind, können mit dem sogenannten ‚Genome screen' identifiziert werden. Hierbei werden ganze Genome oder einzelne Abschnitte systematisch nach bestimmten Selektionsmustern, sogenannten ‚selective sweeps', durchsucht. Da dieses Verfahren die Untersuchung einer Vielzahl von Markern und Individuen erfordert, bedarf es eines geeigneten Analyseverfahrens.

Im zweiten Teil der Arbeit wird die Etablierung einer neuen Methode vorgestellt, die einen hohen Duchsatz von Mikrosatelliten-Markern in kurzer Zeit erlaubt. Die Anwendung neuer Hochdurchsatz-Sequenziertechnologie, ermöglicht es viele hundert Marker simultan zu prozessieren. Es konnte gezeigt werden, dass die erzielten Ergebnisse qualitativ mit denen herkömmlicher Typisierungsmethoden vergleichbar sind.

Mit der neuen Methode wurde das Chromosom 19 des Mausgenoms vollständig nach positiver Selektion durchsucht. Ausgehend von der Annahme dass die charakteristischen Selektionsmuster in einem Fenster von ca 50 kb detektiert werden können, wurden Marker mit diesem Abstand sequenziert und analysiert. Dieser ‚screen' wurde in verschiedenen Populationen zweier Unterarten der Hausmaus (*M. m. musculus* und *M. m. domesticus*) durchgeführt. Die Daten erlauben es sowohl Selektionsparameter abzuschätzen, als auch potentielle Kandidatengene ausfindig zu machen.

Ein detallierter Populationsvergleich ermöglichte die Schätzung der Selektionsfrequenz auf mindestens ein adaptives Ereignis alle 70 Generationen. Zudem lassen die Daten darauf schließen, dass diese Ereignisse in den meisten Fällen mit einem schwachen Selektionsdruck einhergehen.

Interessanterweise zeigten sich im Vergleich zwischen unterschiedlichen Populationen Regionen, bei denen ‚unterartspezifische' Merkmale positiver Selektion auftreten. Dabei tragen Populationen einer Unterart in den gleichen genomischen Regionen identische Merkmale. Dies deutet darauf hin, dass ein Austausch vorteilhafter Mutationen auch zwischen räumlich getrennten Fortpflanzungseinheiten potentiell möglich ist.

Abstract

Understanding the genetic basis of positive selection in natural populations is one of the primary goals in evolutionary biology. Central to this aim is the identification of genomic regions that have been affected by natural selection. Two different approaches allow the investigation of traces in the genome left by selection, which have been both used to look for positive selection in natural populations of the house mouse (*Mus musculus*).

One way is to assess candidate genes selected *a priori*. Such a candidate gene approach defines genes of interest based on a given phenotype, *i.e.* the genes are chosen on the basis of function in biochemical pathways that are relevant to specific phenotypes. This '*top-down*' approach is advantageous if a well-defined association persists between the trait of interest and the underlying gene.

In the first part of this thesis a candidate gene approach was used to study selection on detoxification genes. The analysis was based on two populations of house mice encountering different ecosystems and therefore are thought to have different demands of dietary response. Looking for adaptations in detoxification abilities I conducted a population based comparison of Cytochrome P450 (Cyp450) genes. These genes encode for detoxification enzymes and have already been shown to harbor an important source of adaptations to cope with xenobiotic compounds in different organisms.

Clear indication for positive selection on three Cyp450 genes was found; both, selection on *cis*-regulatory elements was evident, as well as changes on protein level. Notably all three genes showed signs for selection in one of the investigated populations. Furthermore the affected genes are located on different chromosomes, supporting independent selective events within this gene family. This strongly indicates that the respective population evolved genetic responses to specific dietary compounds.

However investigation of *a priori* chosen genes can only respond to previously held ideas, but cannot identify 'unpredicted' genes. Detection of previously unidentified or unsuspected genes that contribute to adaptation can be achieved by systematically screening the entire genome. Thereby whole chromosomes are scanned

for ‘valleys’ of reduced heterozygosity (selective sweeps), a characteristic pattern caused by positive selection. In this case no *a priori* assumptions concerning the potential importance of genes or chromosomal regions are made before the scan is started; the genes are selected ‘*bottom-up*’.

In the second part of the study, I present a genome screen for selection in different house mouse populations. Since the detection of polymorphic variants requires testing multiple individuals for several populations, a complete genome scan requires usage of a large number of markers. First a newly established method is described which facilitates high throughput analysis of microsatellites. A next generation sequencing based approach using the 454 technology was established which allows processing thousands of microsatellite loci simultaneously. I show that the obtained results are reliable and that the novel approach is a useful alternative to standard procedures.

The above described sweep signatures are modified by several parameters such as the recombination rate and the selection coefficient. To reveal deeper insights into the basic parameters of positive selection and detection of chromosomal regions which might be target sites for selection, a genome screen was conducted including different populations of two house mouse subspecies (*M. m. musculus* and *M. m. domesticus*). I used the newly established method to investigate approximately 1,000 microsatellite markers on chromosome 19 in all populations.

A detailed analysis of the candidate loci, identified by single comparisons, revealed results on the frequency of selective sweeps, and the putative origin of selected variants. Significant deviations of the sweep regions from the neutral state are statistically supported. Based on these results, I calculated that there was at least one positive selection event per 70 generations in each lineage. Furthermore, since only two sweeps indicate a broader sweep size than 80 kb, I conclude that positive selection is generally driven by alleles providing weak beneficial impact.

Investigation of subspecies specific sweeps revealed shared signatures of selection between spatially and genetically distinct populations. This strongly indicates that beneficial mutations are potentially shared even among separated entities.

Declaration

The design of the whole project was done together with my supervisors Meike Teschke and Diethard Tautz. The interpretations of all the different results were acquired during numerous discussions. Practical laboratory work as well as the major parts of the data analysis was conducted by me, with some exceptions:

Chapter 1

Bernhard Haubold helped generating the dotplots. Gene expression analysis was performed by Jarek Bryk.

Chapter 2

The software pipeline for the automatic search for microsatellites '*msfinder.pl*' as well as the analysis tool was written by Till Bayer. Heinke Buhtz and Cornelia Burghardt worked on a part of the microsatellite PCRs.

1 General Introduction

1.1 Natural selection – A short introduction

Selection is widely accepted as one of the principal forces shaping phenotypic variation within populations (*e.g.* Cain 1951; Schluter 2000; Rieseberg et al. 2002). Thus the basic desire to learn more about evolutionary processes raised the interest in detecting genes, or genomic regions, that have been targeted by natural selection (Sabeti et al. 2007; Oleksyk et al. 2010). One of the main effects of selection is to modify the levels of variability within and between species and thereby leaving characteristic signatures in the genome. Tracing these signals is of central interest to understand how natural selection shapes genetic variation among species. Natural selection drives different processes whereas each mechanism leaves a characteristic signature in the genome. Commonly, the terminology of positive selection, negative selection, purifying selection, and diversifying selection is used where positive or directional selection is acting upon new advantageous mutations, negative selection removes deleterious alleles, balancing selection is acting on the maintenance of multiple alleles, and when two or more extreme phenotypic values are favored simultaneously it refers to diversifying selection.

Since directional selection plays a major role in phenotypic diversification (Rieseberg et al. 2002), identification of chromosomal regions targeted by directional selection is of central interest to understand the bases of adaptations. The aim of the present study is the identification of positive selection acting in natural populations of the house mouse to gain insights into the process of adaptation of recently derived populations.

Besides natural selection genetic drift is the main force in evolution. The original neutral theory proposes that mutations have no effect on the fitness and consequently have their fate dictated by chance alone (Chamary et al. 2006). The theory was proposed initially by Haldane who argued that, if all differences between species were due to selection, the rate of observed polymorphisms would cause too high cost of selection to be tolerated by populations. Thus the cost of selection would place an upper limit on the rate of evolution. Haldane (1957) estimated that the limit

for a diploid population was one gene substitution per 300 generations. Since the observed rates of molecular evolution appeared to be too fast to be explained by natural selection (Lewontin 1974) the cost of selection were used as main argument for the neutral theory.

However the number of observed polymorphisms among species cannot be explained by neutrality alone, *e.g*. according to the theory species that have large populations should show much higher levels of polymorphism than small populations which is not observed (Lewontin 1974). Why the observed polymorphism levels are relatively invariant remains unclear, but it is likely that linked selection rather than genetic drift is the major force generating these patterns in many natural populations (Gillespie 2000; Chamary et al. 2006). Maynard Smith and Haigh (1974) showed that in large populations the joint fixation of variants linked to the selected sites could reduce neutral diversity much more than random genetic drift (Barton 2000).

Assessing the amount of natural selection shaping genetic variation is one step to clarify the interaction of natural selection and random drift at molecular level (Orr 2005). Since theoretically all mutations, even those that are advantageous, are at risk to be lost by random genetic drift, drift and natural selection do not act in isolation in natural populations. However, the degree to which alleles are affected by drift or selection varies - the more strongly selected an advantageous mutation is, the less likely it is to be lost (Eyre-Walker and Keightley 2007). To assess the amount of natural selection shaping genetic variation as well as to identify genes targeted by selection is of central interest in molecular population genetics to distinguish neutral molecular variation from variation caused by selection (Nielsen 2005).

After a favorable mutation arises, the affected allele rises in frequency in the population, at a speed that depends on the selection coefficient. This process will alter the frequencies of alleles at closely linked loci. In the most extreme case, a single favorable mutation arises at a site which is completely linked to a neutral polymorphic locus. Fixation of the beneficial allele in the population will be accompanied by the fixation of the respective neutral variant that was present in the chromosome carrying the favorable mutation. This will result in reduction of heterozygosity of neutral polymorphism, to an extent which varies with distance from the substituted locus (Maynard Smith and Haigh 1974) (Figure 1.1).

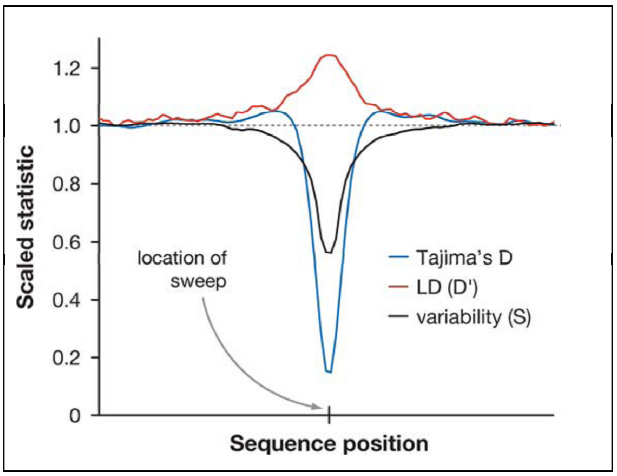


Figure 1.1 Reduction of variable sites in a region of a strong selective sweep: Decrease of Tajima's D while Linkage Disequilirium is increased. Statistics are based on calculations right after the advantageous allele has reached fixation in a population. Dotted line = neutral expectation. Picture taken from (Nielsen 2005).

The rise of frequency due to joint fixation of linked neutral loci is called genetic hitchhiking and the 'footprint' that is left in the genome is referred to as a selective sweep (Figure 1.2). Identification of a sequence region that is not under direct selection, but shows such a decreased variability in comparison to other sequences, can be taken as indication that the linked site has been under selection (Slatkin 1995). Screening for signatures of selective sweeps by comparing variability levels between populations is termed 'hitchhiking mapping' (Harr et al. 2002; Schlötterer 2003). And to date a number of studies have been published regarding the investigation of selective sweeps under varied aspects (e.g. Teshima et al. 2006; Storz et al. 2004; Jensen et al. 2005; Schweinsberg and Durrett 2005).

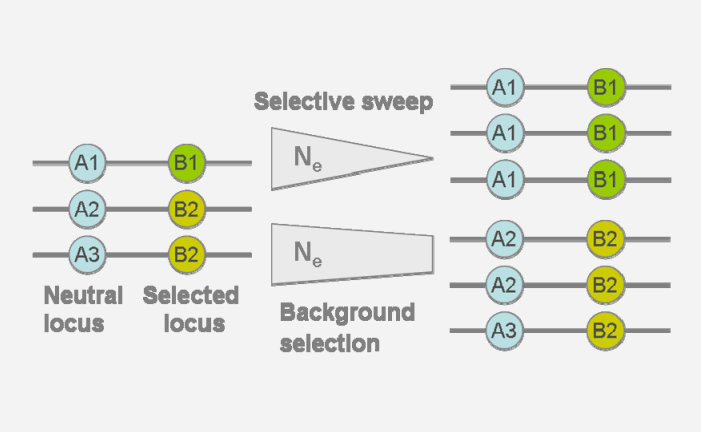


Figure 1.2 The Effects of a Selective Sweep and Background Selection with Complete Linkage. Selective sweep: Fixation of the beneficial variant B1, A2 and A3 are lost, the effective population size (Ne) is severely reduced. Background selection: The deleterious variant B1 is removed, A1 is lost, Ne is slightly reduced (Modified after Charlesworth 2010).

Hence, natural selection leaves characteristic footprints in the genome which can be identified with different methods. The shape of such a selective sweep is mainly determined by the local recombination rate and the selection coefficient (Maynard Smith and Haigh 1974), the former being negatively associated with the

length, the latter positively. Thus in regions of high recombination, the hitchhiking pattern will be disrupted by recombined variation (Fay and Wu 2000). Hence positive correlation between levels of nucleotide diversity and recombination rate may be interpreted as evidence of recurrent selective sweeps (Begun and Aquadro 1993; Andolfatto and Przeworski 2000; Excoffier et al. 2009).

After the beneficial allele reached fixation within the population the footprint is gradually lost by reconstituting variability. This 'recovery' pattern is characterized by an excess of new mutations at low frequencies. Thus, the timeframe in which the pattern of positive selection will be observable in a hitchhiking mapping approach depends on the mutation rate of the investigated neutral marker. Single nucleotide polymorphisms (SNPs) and microsatellites (short tandemly repeated sequences of 1-6 bp in length) are commonly used as such markers. While the mutation rates for SNPs are relatively constant [although it may differ between regions of the genome (Wolfe 1989) [about 2.5×10^{-8} in humans (Nachman and Crowell 2000), 2.1×10^{-8} in mice (Nachman 1997), and 3×10^{-9} in insects (Andolfatto and Przeworski 2000)], the mutation rate of microsatellites is highly variable. It is mainly determined by the repeat pattern and the number of repeat units. In general the mutation rates rises with repeat number (Ellegren 2004). The average mutation rates of microsatellites are estimated to be several orders of magnitude higher than those of SNPs (Schug et al. 1997). Hence, a single beneficial substitution can only be detected in polymorphism data as long as it occurred recently. Przeworski (2002) estimates that the signature of selection can be inferred up to about 10^4 generations in humans and about 10^6 in *Drosophila melanogaster* based on SNPs level. Concerning microsatellites, sweep patterns are expected to be blurred even more quickly by new mutations due to their high mutation rates.

Another important aspect in interpreting sweep patterns is that due to the inconsistency of the microsatellite mutation rates, different loci are not comparable with each other. To circumvent this problem, polymorphisms are deducted by comparing the same microsatellite locus among different populations or species (Schlötterer 2002). By such comparisons two key parameters can be calculated: the variance in repeat number (V) as a measure of variability at the locus (Goldstein and Clark 1995), and the expected heterozygosity (H) (Nei 1978). It has been shown that the logarithm of the ratio between the two values for either V or H ($\ln RV$ and $\ln RH$)

follows a normal distribution, if the microsatellites evolve neutrally (Kauer et al. 2003; Schlötterer 2002). Hence, this allows for detection of loci that depart from the expected null hypothesis with a defined probability. Thereby loci are identified that vary extremely in polymorphism between the two compared entities and are likely linked to sites under selection. In contrast to using the variance, the heterozygosity seems to have a higher power to detect such loci due to the smaller variance for this parameter (Kauer et al. 2003).

The advantages and drawbacks of the hitchhiking approach are quite well understood (Teshima et al. 2006; Thornton et al. 2007). The problem that is considered as most severe is that demographic effects, such as a dramatic, recent population expansion, can produce patterns in the genome that closely mimic the patterns of selective sweeps. Since bottlenecks readily lead to large variances in the genealogical (coalescent) history of samples from different loci along a chromosome (Hermisson 2009) their traces can severely impact the inference of selection (Jensen et al. 2005). Hence, for 'hitchhiking mapping' it is important to take the demographic history of the populations into account.

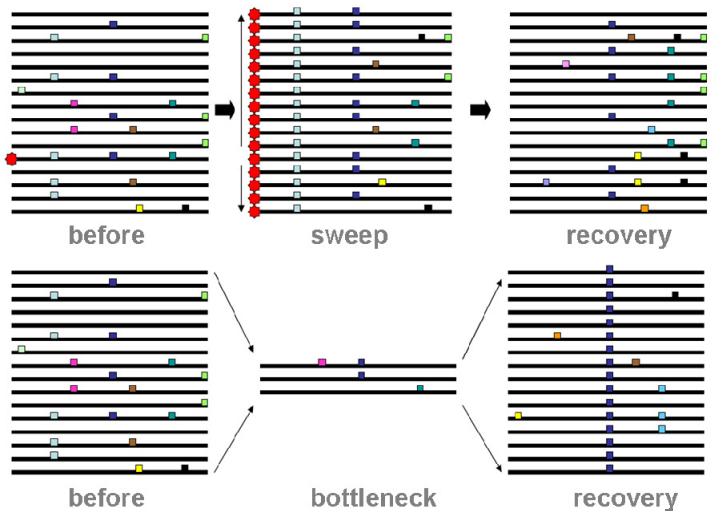


Figure 1.3 Bottleneck events may create patterns that resemble those of selective sweeps. The figure displays the shift in polymorphism for a sweep and a bottleneck scenario. Each line presents a chromosome and colored boxes different alleles. At the top: A beneficial mutation arises in a population (red star) and spreads through the population, whereas linked sites are also raised in frequency. After the sweep the population reconstitutes variation. At the bottom: Population size is drastically reduced by a bottleneck event, whereas variation is randomly reduced. Through population expansion a random allele is raised to high frequency, variation starts to recover. Picture is taken from Reed (2007).

In the present study two different approaches were used to identify genes or chromosomal regions that are under selection. First a candidate gene approach, in which the investigated genes are *a priori* expected to be under positive selection. For this approach genes involved in detoxification of xenobiotic compounds were investigated, namely genes belonging to the Cytochrome P450 gene superfamily. Second, a genome approach was conducted, where more or less randomly selected loci throughout a chromosome were examined which allows the identification of 'unpredicted' regions of selection.

1.2 The house mouse

The house mouse (*Mus musculus*) is one of the key species for research objectives in basic science and has been used to study many aspects of biological questions (Boursot et al. 1993). Two main features characterize the species as a particular suitable organism for investigating the genetic basis of adaptations. First, its evolutionary history is well known since the mouse phylogeny and history has been intensely studied (Boursot et al. 1993; Guenet and Bonhomme 2003). Second, the nearly complete genome sequence is available since 2002 (Mouse Genome Sequencing Consortium 2002) and excellent genetic and genomic resources are available.

The house mouse is the most recent phylogenetic offshoot of the genus *Mus* (Figure 1.4). It evolved on the Indian subcontinent, from where it radiated in several directions to form the well-described peripheral subspecies (*M. m. domesticus, M. m. musculus* and *M. m. castaneus*) (Din et al. 1996). As commensals to humans the species was historically spread all over the world. Fossil findings suggest two independent colonization routes: A northern one via central and northern Europe which is interpreted as being the *M. m. musculus* continental route whereas the southern route passing the western Mediterranean Sea is attributed to the *M. m. domesticus* Mediterranean inflow (Auffray et al. 1990).

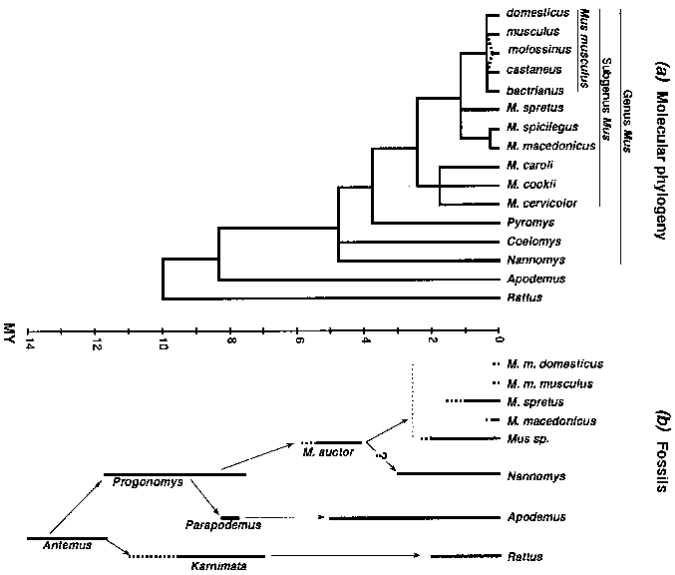


Figure 1.4 Evolutionary tree of the genus *Mus*. The time scale is based on single copy nuclear DNA hybridization studies and is calibrated with the separation of *Mus* and *Rattus*, estimated at 10 Myr ago (taken from Boursot et al. 1993).

Nowadays the nominate subspecies *M. m. musculus* is found all over northern Asia as well as in Eastern Central and Scandinavian Europe; *M. m. domesticus* has its contemporary range in Western Europe, the Near East, Northern Africa, and was recently introduced by humans into the New World, Subsaharan Africa and Australia; the third subspecies *M. m. castaneus,* spread all over South East Asia (Boursot et al. 1993). As displayed in Figure 1.5 several natural hybrid zones exist thus none of the subspecies are completely isolated genetically.

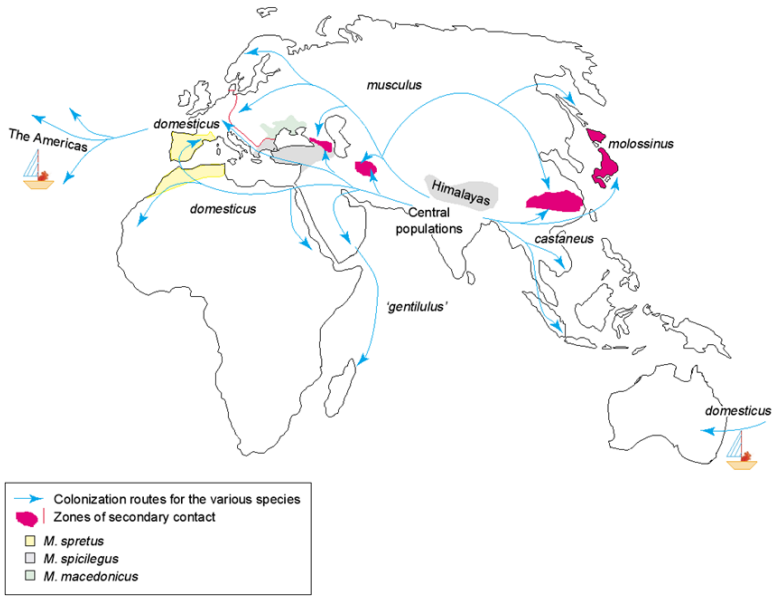


Figure 1.5 Geographical distribution and colonization routes of the different species of the genus *Mus*. (picture from Guénet and Bonhomme 2003).

Its adopted commensal existence with humans afforded the transportation all over the world and migrating with humans facilitated the exploitation of a variety of

new niches and environments. The acknowledged history of repeated successful colonization and the accompanied adaptations to new environments highlight this species as a perfect model system for evolutionary research.

Furthermore, the house mouse has become the most common laboratory animal and hence a broad range of genetic tools are available. However, the laboratory strains of mice do not stem from a single wild population, but are mixtures of different subspecies with the largest contribution from *Mus musculus domesticus*, an intermediate contribution from *Mus musculus musculus*, and a small contribution from *Mus musculus castaneus* (Wade et al. 2002). Sakai et al. (2005) demonstrated that the '*domesticus*' background of most common laboratory mouse strains (one of which is C57BL/6J, the strain used for the genome sequence assembly) is mainly derived from the Western European lineage, on which this study is mainly focused on.

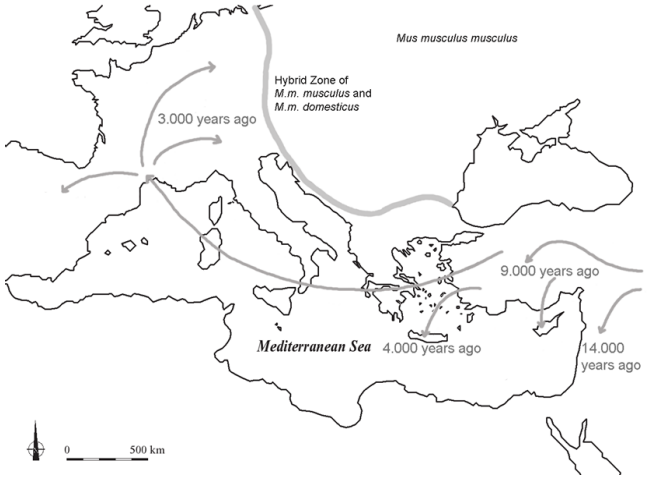


Figure 1.6 Colonization of Middle Europe by the house mouse *Mus musculus domestius* based on data by Cucchi et al. (2005). Successful colonization of Western Europe after incensement of human settlements about 3,000 years ago.

According to fossil evidence, the Western European mice entered Europe from a southern colonization route (Figure 1.6). Based on analysis of palaeontological records the invasion of Western Europe (Cucchi et al. 2005) passed through different phases associated to human movement. Sustained successful colonization of Western Europe finally occurred between the Bronze and Iron Age. Two main factors supported the appearance of the house mouse in Mediterranean and North-Western Europe: i) increase of human settlements, which enhanced the presence and vacuity of ecological niches available for anthropophilous species. The anthropization of the environment should further have lead to decreased predation and interspecific competition as well as an increase of the food availability for mice and hence provided protection against meteorological variation and climatic change (Cucchi et al. 2005). ii) Increase of sea trading. Since the colonization of Mediterranean Europe

occurred later than the appearance of agriculture, agricultural activities is not the only important determining factor. It is suggested that intensification of see trading in the Bronze Age, *i.e.* increase of passive sea transport by humans, plays also a major role in the colonization process (Auffray et al. 1990).

The focal Western European populations of *M. m. domesticus* (one from the Cologne-Bonn-Area and the other one from the Massif Central) investigated in this study are expected to have split upon arrival in Middle Europe about 3,000 years ago, which would correspond to a maximum of 18,000 generations separation time (assuming 3 generations per year and no gene flow after the first split) (Karn et al. 2002). The analysis of the second part comprises two additional '*domesticus*' populations [one very young population from the recently colonized Subsaharan Africa (Cameroon) and one ancestral population sampled in the Near East (Iran)] as well as two populations of the subspecies *M. m. musculus* [one presumably old population from Central Asia (Kazakhstan), and one European '*musculus*' sample (Czech Republic)]. Four of the sample populations have been analyzed in a previous study where Ihle et al. (2005) demonstrated clear distinction of the populations. Since no significant gene flow is observed, effective migration between populations can be excluded (Figure 1.7, the population from the Czech Republic is not included) (Ihle et al. 2005).

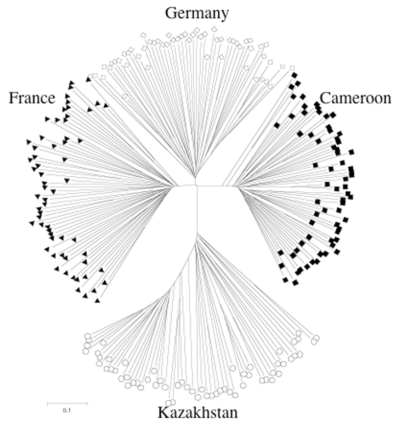


Figure 1.7 Allele sharing tree based on more than 200 microsatellites. The data clearly separates the two Western European populations, as well as samples from Cameroon. The Kazakh population representing the subspecies *M. m. musculus*, is clearly distinct from the *M. m. domesticus* populations by a longer branch (taken from Ihle et al. 2005).

To recapitulate, the presented model system is convenient to investigate fundamental questions of evolutionary biology. Several advantages highlight the particular usefulness of the German and French populations: the history is well documented, the populations are genetically distinct, their maximum divergence time is known, both samples represent derived populations and finally the availability of a nearly complete laboratory mouse genome sequence as described above.

1.3 Aim of the study

The following aims were addressed in this study:

- In Chapter 2, I systematically investigate genes of the Cytochrome P450 superfamily for signs of positive selection. A large group of these oxidase enzymes function in detoxification of xenobiotic compounds. It has been shown in different organisms that they harbor an important potential for adaptation to environmental toxins.

- Genome wide scans for signatures of positive selection based on variability comparisons between natural populations require large amounts of polymorphism data. To process large scale analysis of microsatellites a new high throughput routine was established based on next generation sequencing. In Chapter 3, the new method is presented. I could show that the established routine serves as a convenient alternative to single locus typing, which enables fast processing and analysis for large numbers of loci.

- Applying the described routine, I systematically screened chromosome 19 for signatures of selective sweeps by comparing variability levels of microsatellites between natural populations, with the aim to estimate the minimal frequency of positive selection that occurs in natural populations. Further, comparisons between several, distinct populations were included to allow the identification of inter-population dynamics of selection.

2 Detoxifier under Selection - Investigation of Cytochrome P450 Genes in Populations of Wild House Mice

2.1 Introduction

According to the ecological theory of adaptive radiation, populations that encounter differences in ecosystems are exposed to distinct selection pressures. For example differential availability of resources causes selection leading to specific combinations of advantageous traits for efficient resource exploitation (Schluter 2000). Although the assumption that divergence in environments causes phenotypic differentiation is not controversial, only few data are available for this. Hence, elucidating the genetic basis of adaptive population divergence is a central goal in evolutionary biology (Storz 2005). Application of molecular methods allows the identification of specific genes that underlie adaptive genetic variation.

Since land colonization by animal species about 400 MY ago, plants begun to biosynthesize poisonous chemicals, plant secondary metabolites (PSMs), to deter animal predators (Lewis 2001). In this co-evolutionary 'warfare' animal species that feed on plants have evolved mechanisms to cope with PSMs such as phenolics and terpenes (Bryant et al. 1992). Ingestion of PSM might cause subacute or chronic toxicosis and even relatively nontoxic substances would eventually produce nonspecific adverse effects at high concentrations. Therefore most PSMs that are absorbed from the gut must be metabolized and excreted from the body (Foley and Hume 1987). The detoxification process is energetically expensive and there is always a limit to the degree to which enzymes catalyze such reactions. Thus, integration of new plant species in the alimental spectrum will always induce new chemical defense challenges for wild animals (Palo and Robbins 1991). In terms of natural selection, a forager will gain greater fitness by having either a higher capacity for detoxification of certain PSM or a broader range for substance tolerance. Thus, the variety of chemical compounds ingested is expected to cause certain adaptive patterns in the metabolism (Foley et al. 1995).

As mice have a high demand for biotransformation of PSMs, I hypothesize

that there should be adaptation driven by dietary selection. Here genes involved in detoxification abilities are compared between two distinct populations of *Mus musculus domesticus*. As these populations are exposed to different ecosystems, it is likely that they encounter different food compounds, due to differences in vegetation. Genes that encode for cytochrome P450 (Cyp450) enzymes are such candidates to investigate adaptation in detoxification ability. This large group of oxidase enzymes that function either in the metabolism of endogenous molecules or act in detoxification of xenobiotic compounds is found in all domains of life (Thomas 2007).

Substrate and functional diversity is considered to be the consequence of evolutionary adaptation driven by different metabolic or environmental demands in different organisms (Fu et al. 2009). Particularly those Cyp450 genes, which act in decomposition of xenobiotic compounds evolve fast and are known for their high degree of interspecies and intraspecies variability (Gonzalez and Nebert 1990; Thomas 2007). Evolution of this high diversity has been suggested to be linked to the historical occurrence of important evolutionary events such as the animal-plant divergence (Lewis 2001). While animals began using plants as a food source, plants began to develop defense responses. The evolution of the common plant-animal Cyp450 ancestor was driven by the result of continuous molecular coevolution of plants producing phytoalexins and animals responding with new enzymes to detoxify these chemicals (Gonzalez and Nebert 1990; Fu et al. 2009)

An excessive expansion of new P450 genes via gene duplication was observed throughout various taxa (Foley et al. 1995). In most cases fixation of new genes is ultimately driven by selective advantages (Nebert et al. 1989). It has been proposed that at least four Cyp450 gene families have evolved and diverged in animals due to their exposure to plant metabolites during the last one billion years (Nebert and Gonzalez 1987). Besides constantly ongoing modifications via temporary gene duplications and deletions within Cyp450 gene families, a tremendous genetic modification was provoked by adaptation to the new terrestrial environment about 400 million years ago (Ingelman-Sundberg 2005, Thomas 2007). Several examples for dietary selection in Cyp450 genes can be found in the literature. Ingelman-Sundberg et al. (1999) suggested that adaptation in the *CYP2D6* gene enabled the development of alkaloid resistance in humans. An example for rapid genetic

adaptation in Cyp450 genes as a response to dietary components was observed in the fruit fly where a specific CypP450 gene (*Cyp6g1*) is associated with DDT resistance (Daborn et al. 2002).

Besides the metabolism of PSMs, Cyp450 genes function in detoxification of another considerable toxic substrate group that is taken up with food: fungal secondary metabolites such as aflatoxins. These carcinogenic mycotoxins are produced by a variety of molds, mainly *Aspergillus flavus* and *Aspergillus parasiticus*. Molds are ubiquitous in nature and contaminate a vast array of organic substrates including crop species (Machida and Gomi 2010; Farombi 2006). It has been shown that multiple Cyp450 isoenzymes contribute to catalyze the degradation of aflatoxins (Eaton and Gallagher 1994).

Cyp450 are main components of the Phase I detoxification system, which is generally the first enzymatic defense against foreign compounds. In a typical Phase I reaction, a cytochrome P450 enzyme adds a reactive group to the respective substrate. Thereby molecules are generated which may be more toxic than the parent molecule. Thus, if the molecules are not further metabolized by Phase II conjugation, they may have pathological effects (Ioannides, 1996). Several studies suggest an increasing risk to multiple diseases due to induced Phase I reaction in accociation with decreased Phase II activities (Meyer 1990; Lee 1995). One example is the most carcinogenic aflatoxin B1 (AFB_1). Cyp450-mediated oxidation of AFB_1 is considered to be the dominant route for epoxidation. During this process, AFB_1 is activated to the carcinogenic AFB_1exo-8,9-epoxide primarily by cytochrome enzymes, particularly *Cyp3a4* (Eaton and Gallagher 1994). However, *Cyp3a4* and other P450s also oxidize AFB_1 to less dangerous products (Guengerich et al. 1998). Thus adaptive advantages are expected to arise alternatively in improving the detoxification efficiency or diverting reactions to a less dangerous route.

In this study I investigate two distinct natural populations of the house mouse to test whether there has been positive selection within the Cyp450 gene superfamily. In mice Cyp450 genes that encode for enzymes acting on metabolism of endogenous compounds appear in seven gene clusters with several related genes and pseudogenes that are tandemly repeated (Nelson et al. 2004). To trace signatures of adaptation to environmental toxins a microsatellite screen within the seven mouse Cyp450 gene clusters was performed. Signs of selection were assessed by comparisons of

microsatellite variability between the sample populations. Population or locus-specific reduction of variability at polymorphic loci can be taken as indication of positive selection at linked sites (selective sweeps) and hence offers a way to identify genes that have been recently involved in adaptation (Harr et al. 2002). Signs of selective sweeps were estimated based on Schlötterer's (ln*RH*) statistics (Schlötterer 2002). It has been shown that this ratio statistic is quite robust against large fluctuations in mutation rates and population size (Kauer et al. 2003). To gain further insight into the structure of the Cyp450 clusters as well as to get a rough estimate of the age of duplication events, dotplots were generated for all Cyp450 gene clusters. Furthermore, differences in gene expression of Cyp450s were compared between the respective populations. For candidate genes additional sequence data were obtained to assess SNP distribution patterns.

The here compared German and French population of *M. m. domesticus* have split about 3,000 years ago and it has been shown that the populations are effectively distinct. The French population sample originated from Southern France (Massif Central), the German from Western Germany (Cologne/Bonn Region). Since the ecosystems of the two sampling locations differ in several abiotic factors, such as climate and altitude, they will vary in vegetation as well. Hence it is reasonable to assume that differences in the food plants and the associated PSMs and fungi spectrum affect the dietary responsiveness of these populations (Kottek et al. 2006).

In addition to the previously described analysis of the Cyp450 gene clusters, expression levels of elements that play a central role in mediating the induction of hepatic Cyp450s were investigated. Three nuclear receptor superfamily members, constitutive androstane receptor (CAR), nuclear pregnane X receptor (PXR) and peroxisome proliferator-activated receptor (PPAR), are centrally involved modulators of hepatic Cyp450s belonging to families Cyp2, Cyp3, and Cyp4 (Waxman 1999). These families are activated by a diverse set of xenotiotic substrates (Akiyama and Gonzalez, 2003). Another important factor for the constitutive expression of Cyp450s is hepatocyte nuclear factor 4 (HNF4 α). Activating effects on particular Cyp450 promoters from several species have been detected for this member of the receptor superfamily (Jover et al. 2001). While PXR and CAR are the primary transcription factors coordinating induced expression of the enzymes and proteins regulating oxidative, conjugative and transport phases of endobiotic and xenobiotic metabolism,

HNF4α can modify the PXR/CAR response (Echchgadda et al. 2007).

Additionally aminolevulinic acid synthase 1 (*ALAS1*) was taken into the expression analysis. This enzyme is the first and rate-limiting enzyme in the mammalian heme biosynthetic pathway and is induced in response to various xenobiotics (May et al. 1995). As Cyp450s require heme as a cofactor to oxidize substrates, this enzyme can be considered as a key factor for Cyp450 efficiency.

2.2 Methods

Population samples of *M. m. domesticus* from France and Germany consist of wild-caught individuals. For detailed information about the samples used in this study see Ihle et al. (2006). DNA was isolated by standard salt extraction procedure and subsequently stored at - 20°C. For further processing, concentration of DNA was adjusted to 5 ng/µl and preserved in a 96 well plate.

For RNA preparation (used for gene expression analysis) F1 cage mice generated from wild-caught individuals were raised under standardized conditions and sacrificed at the same age. Only unrelated males were taken into the analysis. Tissues were frozen in liquid nitrogen immediately after dissection and stored at - 80°C. RNA was extracted following a protocol using Trizol. Expression profiling was carried out using two different Microarray platforms, [GeneChip® Mouse Genome 430A 2.0 Array from Affymetrix and Agilent-014868 Whole Mouse Genome Microarray 4x44K (G4122F)]. Processing for the Affymetrix Microarray was outsourced (CCG – University of Cologne), whereas completion of Agilent Microarray was performed following the manufacturers protocol in house.

Sequences of the seven Cyp450 gene clusters were downloaded from GenBank and screened for microsatellite loci applying the program ‘tandem repeats finder’ (Benson 1999). Primer design was carried out using ‘FastPCR’ software (Kalendar et al. 2009). To assure uniqueness of the PCR products each primer pair was blasted using NCBI/Primer-BLAST (NCBI 2008). Microsatellites were amplified by PCR using fluorescently labeled forward primers following the QIAGEN (Valencia, CA) multiplex kit manual (cat. no. 206143) and run on an ABI3730. For

analysis GeneMapper software 4.0 (Applied Biosystems 2009) was used. In each population 40 single individuals were genotyped.

For some candidates further sequence data was gathered. Primer design and amplification was performed as described above. Sequence analysis was achieved and edited using CodonCodeAligner software 2.0.4 (CodonCode Corporation 2009).

Data Analysis: Dotplots were generated for each region using the computer program '*drawDotPlot.awk*' (Haubold 2008) which drives several software tools [blastall, formatdb (BLAST package); seqret [EMBOSS]; GNU plotutils package] for rapidly constructing dot plots. Here, the sequence of each cluster was plotted to itself (E-value=200). The age of some duplicated genes was estimated based on the number of synonymous substitution and the mutation rate (Suyama et al. 2006). The calculation assumes a mutation rate of 3.06×10^{-9} per bp per year and a generation time of 3 generations per year (Nachman 1997; Karn 2002).

Gene diversity analysis for genotyped microsatellite loci were carried out using MSAnalyzer 3.15 (Dieringer and Schlötterer 2003). To compare levels of variability between the populations ln*RH* statistics were applied (Kauer et al. 2003). Outlier were estimated by normalizing the individual ln*RH* values based on an independent reference data set of 64 microsatellites representing a genomewide variability distribution of *M. m. domesticus*. The reference polymorphism data reveal a distribution of gene diversity characterized by: mean 0.0875, standard deviation 0.8584 (Teschke et al. 2008).

Basic polymorphism statistics as well as haplotype analysis were performed using DnaSP 4 software. Reconstruction of the phase of diploid individuals was conducted with DnaSP 4 software (Rozas et al. 2003) using the algorithm provided in PHASE. Applying the default settings the output probability threshold for haplotypes is defined as 90%. For each population several measures of diversity were computed. Wattersons θ_W based on the number of segregating sites, and π (Nei and Li 1979), which estimates the per-site heterozygosity derived from the average number of pairwise sequence differences, as well as k, which reflects the average number of nucleotide differences. To test whether the frequency spectrum of mutations deviates from the standard neutral model Tajima's D (Tajima 1989) was calculated, which considers the difference between θ_W and π. Under neutrality, the test should be close to 0. Deviations from neutrality of the observed number of singleton polymorphisms

were estimated using Fu and Li's D statistic (Fu and Li 1993).

Coalescent simulations were performed including the number of segregating sites (S) and intermediate recombination (R). Recombination rates were estimated based on the radiation hybrid (RH) map (Rowe et al. 2003) and assuming an effective population size of 58,000 (Salcedo et al. 2007).

Data for gene expression was generated for three tissues: liver, brain and testis. The data set comprises six individuals of each population. Expression data were analyzed at gene level, *i.e.* signals of different transcripts were combined related to associated genes. For both platforms expression analysis was performed in R. (1) Agilent using packages Agi4x44PreProcess, limma, qvalue, and annotation package mgug4122a.db. (2) Affymetrix using the Affy package.

Raw signals were normalized and corrected for background noise. Probes were only considered as expressed if expression was detected in at least 3 samples. Differentially expressed genes were identified using moderated t-test statistics. Genes are considered as significantly differentially expressed if both platforms show an overlap of significant p-values (Bryk et al. unpublished data).

2.3 Results

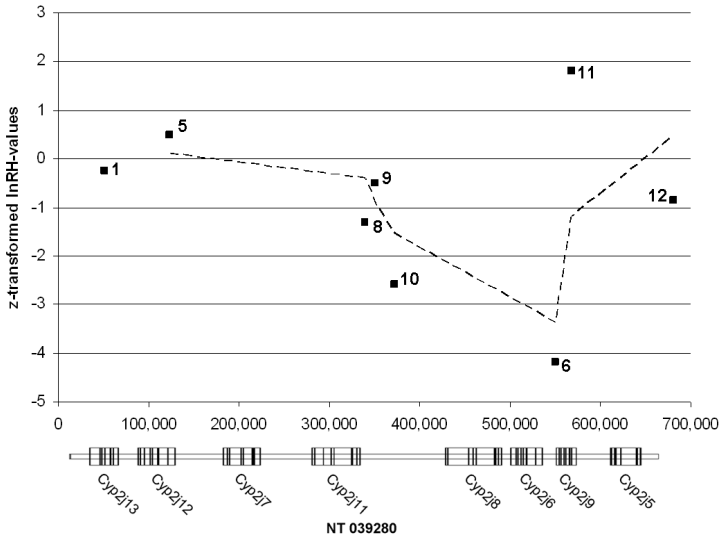
Microsatellite scan: In total 35 microsatellite loci were investigated within the seven Cyp450 gene clusters. To avoid ambiguous information caused by recent duplication events, only those loci were selected for which unique primers could be designed. According to ln*RH* values four outlier loci ($p< 0.05$) were detected in three different clusters: one in the Cyp2c (see digital supplementary Chapter 2) and Cyp2j respectively and two loci in the Cyp3a cluster.

Expression data: Analysis of gene expression data revealed two genes that show significant differences in expression ($p<0.05$) between the focal populations: *Cyp2j6* which was observed as differentially expressed in liver and brain respectively and *Cyp3a13* which was only differentially expressed in liver (Figure 2.3 and Figure 2.6). As the Cyp2j and the Cyp3a clusters already appeared as outliers in the microsatellite scan further analysis focused on these clusters.

2.3.1 Cyp2j gene cluster

This gene cluster is located on chromosome 4 in the mouse genome. In total eight microsatellite loci distributed over the whole Cyp2j cluster were investigated; seven were analyzed in the course of the microsatellite scan. Since *Cyp2j6* has been observed as differentially expressed, one additional marker (locus 6) was added in proximity to this gene. Among the eight analyzed markers locus 10 and 6 were detected as outliers (Figure 2.1), where both loci show reduction of variability in the French population (Figure 2.2).

The allele frequencies are plotted in Figure 2.2 and show the distribution of polymorphism in the two populations. At locus 6 the shortest allele is almost completely fixed within the French population, whereas the sweep allele of locus 10 is of intermediate length and less abundant among the sampled chromosomes. However, both loci display a classical sweep pattern in the French population.



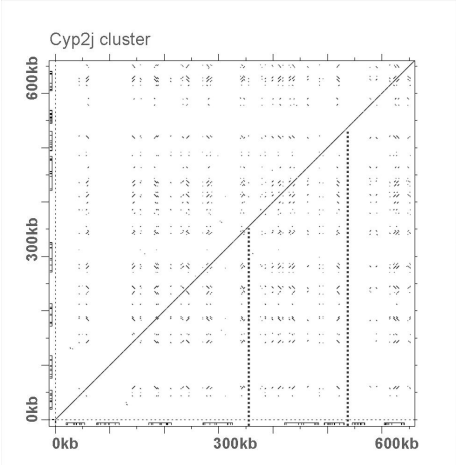


Figure 2.1 Observed ln*RH* values along the Cyp2j cluster. The length of the region is denoted in bp. Genes located within the cluster are displayed above. Left: Dotplot of the focal chromosomal region. Dotted lines indicate the locations of the two outlier loci 10 and 6. Genes are illustrated as boxes at the axes.

Notably, besides these two loci, reduction of variability is also reflected by non outlier loci in the French population, *e.g.* at microsatellite loci 8 and 12 the French population displays a limited allele spectrum compared to the German sample. Furthermore, allele distribution at locus 9 displays a pattern similar to a recovery pattern in the French population. Especially long microsatellites which are expected to have high mutation rates are at risk loosing their significant sweep pattern rapidly because of new mutations after fixation. Hence, it is likely that this locus has been recently affected by selection but has already recovered from the sweep. Apart of locus 11 which is almost completely monomorphic in both populations, none of the investigated loci showed reduced variability within the German population.

Possible patterns of gene duplication among Cyp2j genes are illustrated in the presented dotplot (Figure 2.1). Here, the sequence of the Cyp2j cluster was plotted

against itself. Every line in the dotplot marks an identical string in the respective complement. The two dotted lines indicate the positions of the outlier loci 10 and 6. Neither of the sites shows a specific pattern in the dotplot. Note that a very high E-value was chosen to generate the dotplot to detect only highly similar sequences. Since there are only a few short dashes visible among the complete sequence, this indicates that the duplication events in this cluster are relatively old and the associated genes are already diverged. Estimations of the age of duplication events based on substitution rates in coding sequence confirm this assumption, *e.g.* reconstructing the age of the duplication event between the initial gene in this subfamily *Cyp2j5* [based on (Thomas 2007)] and *Cyp2j6* for example leads to a distance of approximately 20 MY between these genes.

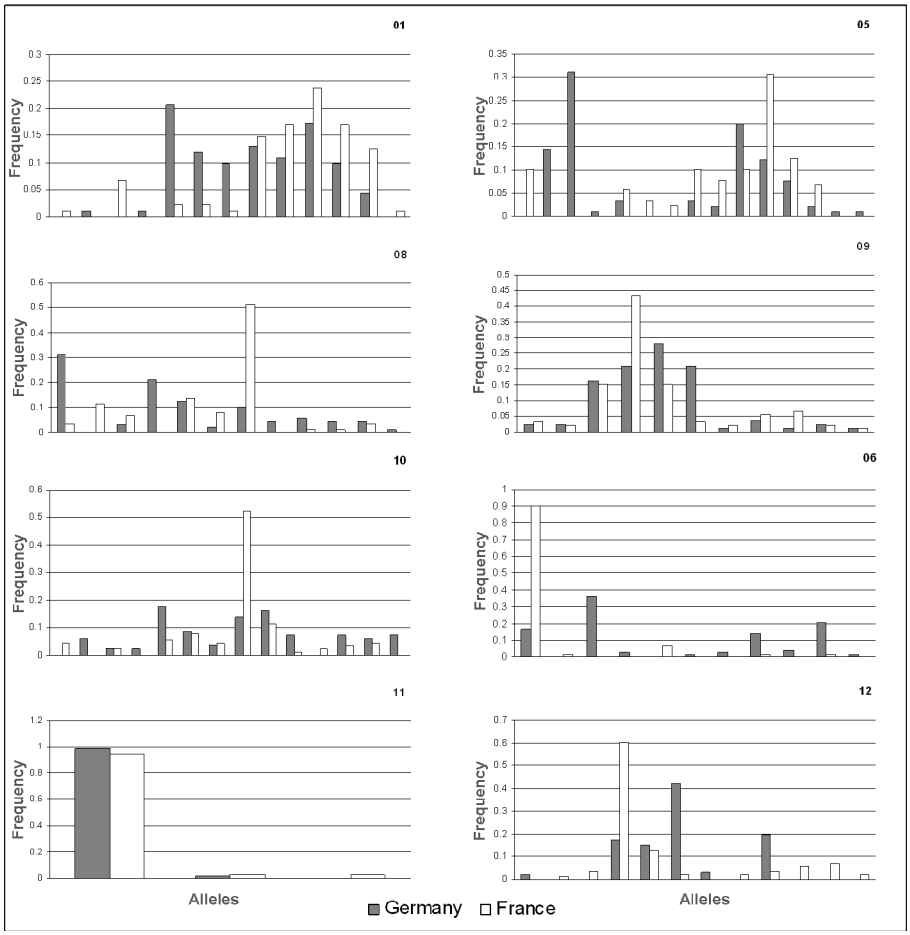


Figure 2.2 Allelefrequencies of the German and French population at microsatellite loci investigated in the Cyp2j cluster.

As mentioned above significant differences in expression are observed for *Cyp2j6* in liver and brain respectively (Figure 2.3). In both tissues the French population shows a significantly lower expression compared to the German population. This observation is consistent on both platforms. Among the German samples expression of *Cyp2j6* is 1.5 fold increased in brain, and 2 fold in liver respectively. Since all mice were kept under standardized conditions and sacrificed at similar ages expression differences due to external factors can be excluded.

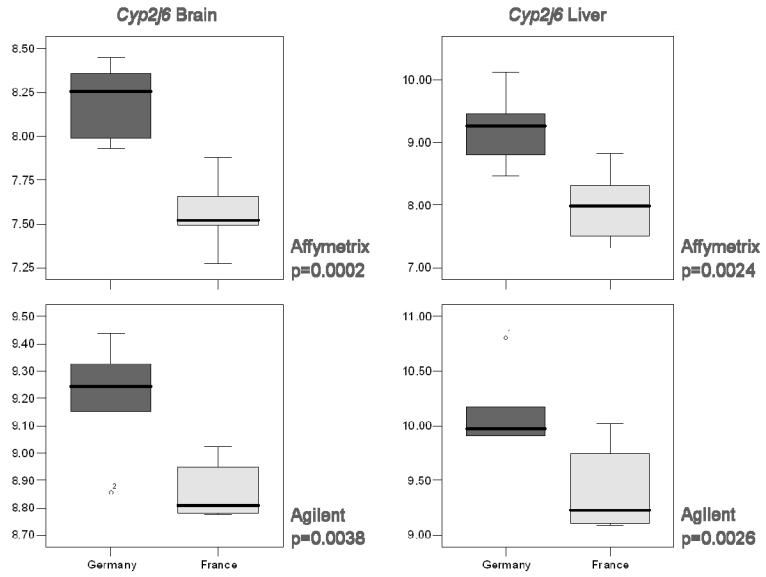


Figure 2.3 Exprssion differences of *Cyp2j6* between German and French samples in brain and liver. Expression values are log transformed. Differences in gene expression are significant in both tissues. Above: Affimetrix; Below: Agilent.

Observed gene expression data as well as ln*RH* values provide an indication for selection having affected *Cyp2j6* in the French population. To study this further, sequence polymorphisms were assessed in the respective region. Therefore 1,181 bases were analyzed in 10 French and 8 German individuals including coding and non-coding sequence. 14 segregating sites (S) (Table 2-1) were observed, 12 among the French and 3 among the German samples. For neither of the populations Tajima's D shows any sign of selection. However, the French population revealed a significant Fu and Li's D*.

Table 2-1 Polymorphism data at Cyp2j6 obtained from 10 French and 8 German individuals. S = Number of segregating sites, π = Tajima's nucleotide diversity, Θ = Wattersons nucleotide diversity per site, k = number of nucleotide differences, Hd = Haplotype diversity, h = number of haplotypes.

Population	Sample size	Length (bp)	S	π	Θ	k	Hd	h	Tajima's D	Fu and Li's D*
France	10	1181	12	0.0027	0.00287	3.21	0.45*	6*	-0.18244	1.46*
Germany	8	1181	3	0.0008	0.00077	0.93	0.68	5	0.06703	-0.04

		17408	17476	17684	17685	17704	17751	17777	17948	18161	18375*	27498	27539	27561	27590
G1	H1	C	C	T	T	C	G	A	A	T	A	A	G	C	T
G2	H2	.	A	.	.	.	.	.	.	.	.	.	.	.	.
G3	H2	.	A	.	.	.	.	.	.	.	.	.	.	.	.
G4	H2	.	A	.	.	.	.	.	.	.	.	.	.	.	.
G5	H2	.	A	.	.	.	.	.	.	.	.	.	.	.	.
G6	H2	.	A	.	.	.	.	.	.	.	.	.	.	.	.
G7	H2	.	A	.	.	.	.	.	.	.	.	.	.	.	.
G8	H2	.	A	.	.	.	.	.	.	.	.	.	.	.	.
G9	H2	.	A	.	.	.	.	.	.	.	.	.	.	.	.
G10	H2	.	A	.	.	.	.	.	.	.	.	.	.	.	.
G11	H3	.	A	.	.	.	.	T	.	.	.	.	.	.	.
G12	H3	.	A	.	.	.	.	T	.	.	.	.	.	.	.
G13	H4	.	A	.	.	.	.	T	.	.	.	G	.	.	.
G14	H4	.	A	.	.	.	.	T	.	.	.	G	.	.	.
G15	H5	.	A	.	.	.	.	.	.	.	.	G	.	.	.
G16	H5	.	A	.	.	.	.	.	.	.	.	G	.	.	.
F1	H1	.	.	.	.	.	.	.	.	.	.	.	.	.	.
F2	H6	.	.	.	.	.	.	.	.	.	.	.	.	T	.
F3	H7	.	.	.	.	.	.	.	.	.	.	.	C	T	G
F4	H8	A	A	A	G	T	A	.	T	C	G	.	.	.	.
F5	H9	A	A	A	G	T	A	.	T	C	G	.	C	T	.
F6	H10	A	A	A	G	T	A	.	T	C	G	.	C	T	G
F7	H10	A	A	A	G	T	A	.	T	C	G	.	C	T	G
F8	H10	A	A	A	G	T	A	.	T	C	G	.	C	T	G
F9	H10	A	A	A	G	T	A	.	T	C	G	.	C	T	G
F10	H10	A	A	A	G	T	A	.	T	C	G	.	C	T	G
F11	H10	A	A	A	G	T	A	.	T	C	G	.	C	T	G
F12	H10	A	A	A	G	T	A	.	T	C	G	.	C	T	G
F13	H10	A	A	A	G	T	A	.	T	C	G	.	C	T	G
F14	H10	A	A	A	G	T	A	.	T	C	G	.	C	T	G
F15	H10	A	A	A	G	T	A	.	T	C	G	.	C	T	G
F16	H10	A	A	A	G	T	A	.	T	C	G	.	C	T	G
F17	H10	A	A	A	G	T	A	.	T	C	G	.	C	T	G
F18	H10	A	A	A	G	T	A	.	T	C	G	.	C	T	G
F19	H10	A	A	A	G	T	A	.	T	C	G	.	C	T	G
F20	H10	A	A	A	G	T	A	.	T	C	G	.	C	T	G

Table 2-2 Observed haplotypes at *Cyp2j6* among 8 German (G1-16) and 10 French (F1-20) samples. In total 10 different haplotypes are observed. Closely related haplotypes are colored in grey scale. Numbers indicate the position in bp of the segregating sites in relation to the genomic sequence of the gene.* = synonymous change

Comparing polymorphism data the French population appears more variable. It exhibits four times more segregating sites which results in higher nucleotide diversity (π and Θ) as well as a higher number of nucleotide differences (k). In contrast, coalescent analysis reveals less variability in the French population (Table 2-1). Here we find a significant reduction in haplotype diversity (Hd) ($P<0.01$) as well as significance in the number of haplotypes in relation to segregating sites (h) ($P<0.02$). None of these parameters were observed significant in the German

population.

Among the German samples three main haplotypes appear [coloured in greyscale (Table 2-2)], which split into five at position number 27,498 (positions relate to the genomic sequence of *Cyp2j6*, see digital supplement Chapter 2). Up to this position the French population divides only into two major haplotypes that split into six taking the last four polymorphic sites into account. Since there is a distance of approximately 10 kb between the segregating sites 18,375 and 27,498 the enhanced haplotype variability at the last four polymorphic sites presumably result from subsequent recombination events.

Focusing on the primary part of the sequence, one shared haplotype is observed, marked as light grey, between Germany and France which might be the ancestral one. In Germany this haplotype occurs with a frequency of 62.5% but was only observed with 15% frequency in the French population. Here, the black coloured haplotype is the most common one.

To investigate potential changes on the protein level all nine exons of *Cyp2j6* were subsequently sequenced. No nonsynonymous substitution was detected in any of the examined sequences.

Table 2-3 Expression data obtained from both platforms and the respective haplotypes for each individual at *Cyp2j6*. Expression values are log transformed. Left: Data shown for liver; Right: Data shown for brain.

Cyp2j6 liver	Affymetrix	Agilent	Haplotypes	
Germany 1	10.12	10.80	H2	H2
Germany 2	9.46	9.91	H4	H4
Germany 3	8.47	9.94	H2	H3
Germany 4	9.18	10.00		
Germany 5	9.36	10.17	H5	H5
Germany 6	8.81	9.91	H2	H2
France 1	8.02	9.28	H10	H10
France 2	8.32	9.74	H10	H10
France 3	7.33	9.11	H10	H10
France 4	7.95	9.09	H7	H10
France 5	7.50	9.18	H10	H10
France 6	8.83	10.02	H6	H9

Cyp2j6 brain	Affymetrix	Agilent	Haplotypes	
Germany 1	8.35	9.33	H2	H2
Germany 2	8.45	8.86	H4	H4
Germany 3	8.16	9.27	H2	H3
Germany 4	8.36	9.44		
Germany 5	7.93	9.22	H5	H5
Germany 6	7.99	9.15	H2	H2
France 1	7.66	8.78	H10	H10
France 2	7.88	8.78	H10	H10
France 3	7.27	8.83	H10	H10
France 4	7.50	8.78	H7	H10
France 5	7.50	8.95	H10	H10
France 6	7.54	9.02	H6	H9

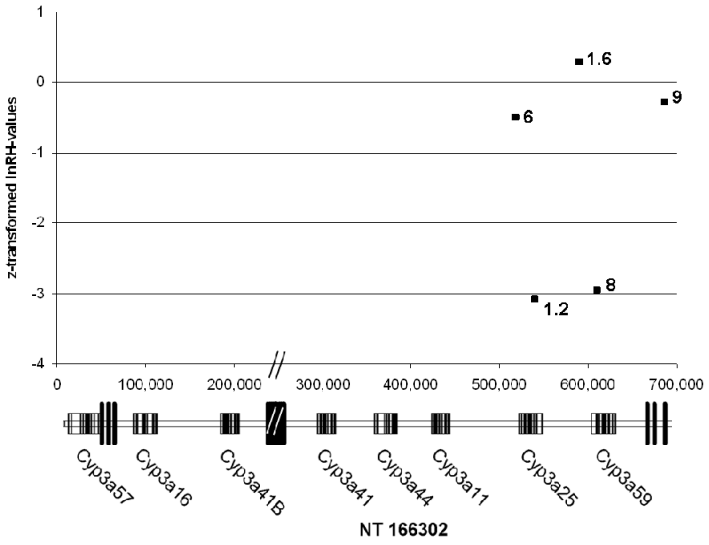
Gene expression data in combination with the corresponding haplotypes of the respective six individuals revealed that except sample number 6 all French individuals exhibit the haplotype H10 (Table 2-3). Individual number 6 is heterozygote for H6

and H9. Remarkably, this individual shows an almost 2 fold higher expression in liver compared to the other French samples; note that expression values are log transformed. This result was confirmed on both platforms. In the brain sample individual 6 shows slightly higher expression values on Agilent as well, but these findings are inconsistent between the platforms.

2.3.2 Cyp3a gene cluster

Due to recent duplication events this cluster exhibits strong sequence similarity which restricted options for microsatellite analysis; as mentioned, only microsatellites for which unique PCR products could be generated were incorporated into the study. Moreover parts of the Cyp3a cluster were not analyzed as there are still annotation gaps in the reference genome. One gene *Cyp3a13* had to be separately investigated since this gene is located in 8Mb distance to the remaining Cyp3a family members and will be presented separately.

According to ln*RH* values two of the examined loci (1.2 and 8) are detected as outliers ($p < 0.05$). Similar to the findings in the Cyp2j cluster both loci reveal a sweep pattern in the French population (Figure 2.4). As shown in the corresponding dotplot these loci are located within the two recently duplicated, and inverted, genes *Cyp3a25* (locus 1.2) and *Cyp3a59* (locus 8) (Figure 2.4). Thus linkage of the marker and the respective gene is likely. The estimated age of the duplication event is approximately 1-2 MY. Comparisons to the initial gene of this cluster, *Cyp3a13* [(sequence alignment obtained from Thomas (2007)], revealed that that *Cyp3a59* originated from *Cyp3a25*.



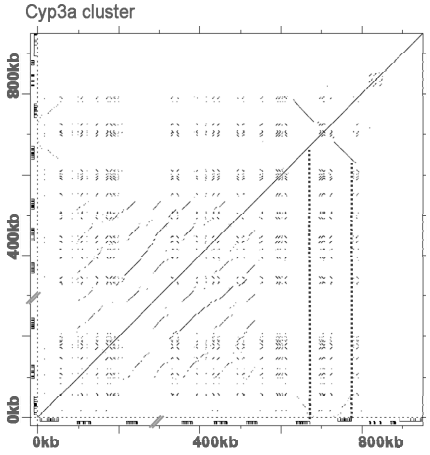


Figure 2.4 Observed ln*RH* values along the Cyp3a cluster. The length of the region is denoted in bp. Genes located within the cluster are displayed above. Left: Dotplot of the investigated region. Dotted lines indicate the locations of the two outlier loci 1.2 and 8. Genes are illustrated as boxes at the axes.

Both outlier loci 1.2 and 8 obtain allele frequencies which resemble a classical sweep pattern where one allele is almost fixed among the chromosomes (Figure 2.5). In both cases the shortest allele is at high frequency. The observed allele distribution at locus 6 resembles a recovery pattern in the French population and hence might be taken as another indication for a recent sweep event. Concordantly this locus is located in the proximity to the outlier locus 1.2 and supports the assumption of selection acting at gene *Cyp3a25*. No hint for unequal distribution of variability between the two populations was observed at loci 1.6 and 9.

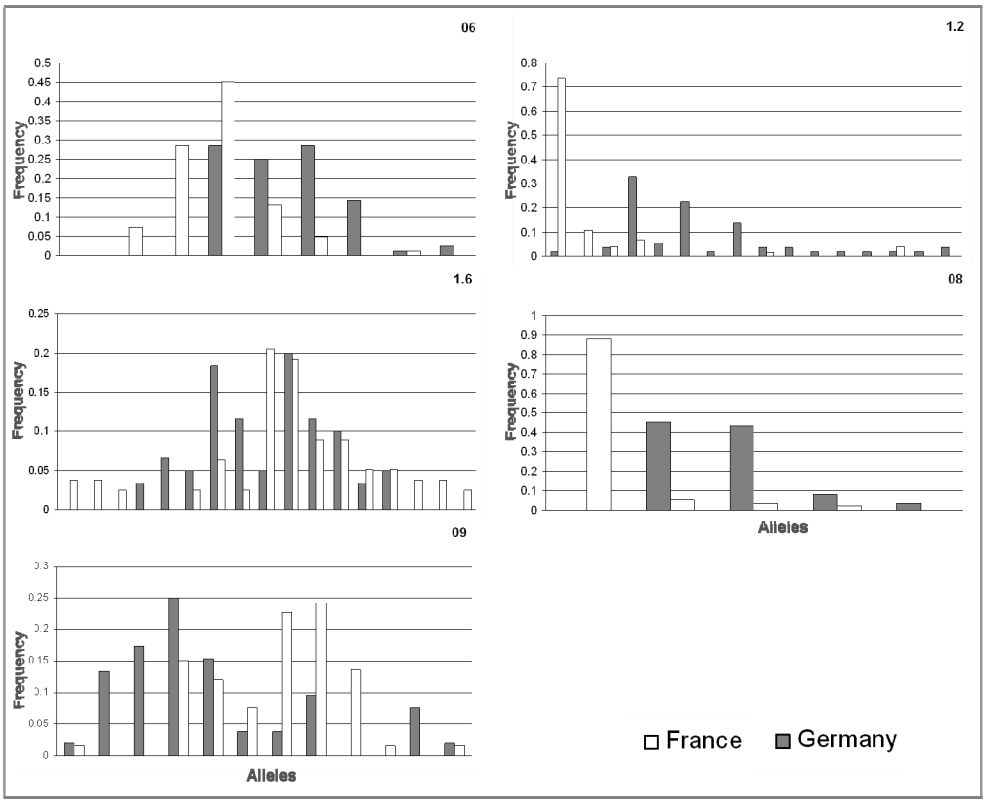


Figure 2.5 Allele frequency distributions at the five microsatellite loci investigated in the Cyp3a cluster.

For sequence analysis of gene *Cyp3a25* two fragments of cDNA were analyzed. One of approximately 500 bp including exon 1-6 (F1), and another one of about 400 bp which comprised exon 10, 11 and 12 (F2) (Table 2-1).

Table 2-4 Polymorphism data for two sequence fragments of *Cyp3a25* and one fragment of *Cyp3a13* for 10 French and 9 German individuals. S = Number of segregating sites, π = Tajima's nucleotide diversity, Θ = Wattersons nucleotide diversity per site, k = number of nucleotide differences, Hd = Haplotype diversity, h = number of haplotypes.

Fragment	Population	Sample size	Length (bp)	S	π	Θ	k	Hd	h	Tajima's D
Cyp3a25	France	10	486	8	0.0028	0.0046	1.34	0.36*	4	-1.37
F1	Germany	9	486	10	0.0063	0.0060	3.08	0.88	9	0.21
Cyp3a25	France	10	392	2	0.0016	0.0015	0.61	0.35	3	0.17
F2	Germany	9	392	1	0.0003	0.0007	0.19	0.19	2	-0.59
Cyp3a13	France	10	581	6	0.0050	0.0029	2.91	0.68	5	2.29*
	Germany	9	581	6	0.0042	0.0030	2.41	0.73	4	1.26

Both populations exhibit a high number of polymorphic sites in F1, leading to rather high values of θ in this fragment. Notably a fairly high number of changes are nonsynonymous; five out of twelve polymorphisms are nonsynonymous in the

German population and four in the French population, respectively. Comparison of pairwise differences reveals lower values in the French population leading to clearly negative values of Tajima's D. As only two segregating sites were observed in F2, polymorphism analysis is insufficiently informative for this fragment.

		141**	145**	167*	233**	285**	317**	320*	323*	325*	345*	346**	444*
G1	H1	G	C	A	A	G	A	G	G	T	C	A	C
G2	H1	.	.	.	.	.	.	.	.	.	.	.	.
G3	H1	.	.	.	.	.	.	.	.	.	.	.	.
G4	H1	.	.	.	.	.	.	.	.	.	.	.	.
G5	H1	.	.	.	.	.	.	.	.	.	.	.	.
G6	H2	T	.	.	.	.	.	.	.	.	.	.	.
G7	H2	T	.	.	.	.	.	.	.	.	.	.	.
G8	H3	.	.	.	.	.	.	.	C	.	.	.	T
G9	H4	.	.	.	.	.	.	.	C	.	T	.	.
G10	H5	.	.	.	T	.	.	.	C	.	T	.	.
G11	H6	.	.	.	T	.	.	.	C	.	T	T	.
G12	H7	.	.	T	T	.	.	.	C	.	T	T	.
G13	H7	.	.	T	T	.	.	.	C	.	T	T	.
G14	H8	.	.	.	.	A	.	A	.	.	.	.	.
G15	H8	.	.	.	.	A	.	A	.	.	.	.	.
G16	H8	.	.	.	.	A	.	A	.	.	.	.	.
G17	H8	.	.	.	.	A	.	A	.	.	.	.	.
G18	H9	.	.	.	.	A	T	A	C	.	.	.	.
F1	H1	.	.	.	.	.	.	.	.	.	.	.	.
F2	H1	.	.	.	.	.	.	.	.	.	.	.	.
F3	H1	.	.	.	.	.	.	.	.	.	.	.	.
F4	H1	.	.	.	.	.	.	.	.	.	.	.	.
F5	H1	.	.	.	.	.	.	.	.	.	.	.	.
F6	H1	.	.	.	.	.	.	.	.	.	.	.	.
F7	H1	.	.	.	.	.	.	.	.	.	.	.	.
F8	H1	.	.	.	.	.	.	.	.	.	.	.	.
F9	H1	.	.	.	.	.	.	.	.	.	.	.	.
F10	H1	.	.	.	.	.	.	.	.	.	.	.	.
F11	H1	.	.	.	.	.	.	.	.	.	.	.	.
F12	H1	.	.	.	.	.	.	.	.	.	.	.	.
F13	H1	.	.	.	.	.	.	.	.	.	.	.	.
F14	H1	.	.	.	.	.	.	.	.	.	.	.	.
F15	H1	.	.	.	.	.	.	.	.	.	.	.	.
F16	H1	.	.	.	.	.	.	.	.	.	.	.	.
F17	H2	T	.	.	.	.	.	.	.	.	.	.	.
F18	H2	T	.	.	.	.	.	.	.	.	.	.	.
F19	H7	.	.	T	T	.	.	.	C	.	T	T	.
F20	H10	.	T	T	T	.	.	.	C	A	T	T	.

Table 2-5 Observed haplotypes for *Cyp3a25* among 9 German (G1-18) and 10 French (F1-20) samples. The data are based on Fragment 1 which comprises 500 bp of cDNA including exon 1-6 (which equates to approx. 16,000 bp genomic DNA). In total 10 different haplotypes are observed. Numbers corresponding to the positions based on mRNA (*Cyp3a25* sequence information is added to the digital supplement, Chapter 2). * = synonymous change, ** = nonsynonymous change.

Although both populations exhibit similar numbers of segregating sites, the number of haplotypes greatly differs between the two populations in F1. Whereas the German population exhibits nine different haplotypes, the French population shows only four (Table 2-4 and Table 2-5). Since these investigated fragments were obtained from cDNA, coalescent simulations for both fragments were based on the number of segregating sites and computed without any recombination. Even though this is the most conservative model the French population revealed significance in haptotype diversity ($p<0.02$). Hence, these data support the assumption of selection acting on *Cyp3a25* in the French population. Analysis of expression data of *Cyp3a25* was not possible as the probe sets did not match the respective gene sequence unambiguously. Therefore, gene expression data for this gene had to be excluded from the analysis.

Separate investigation of *Cyp3a13* indicates signals of selection on the basis of microsatellite analysis as well as expression data. At the investigated microsatellite locus next to *Cyp3a13* the variability is clearly reduced in the German population ($p<0.05$). Notably a high number of low frequency alleles are present in both populations (Figure 2.6) which resembles an allele pattern generated under balancing selection. On gene expression level *Cyp3a13* exhibits significantly higher RNA concentration in the liver samples of the German population (Figure 2.6).

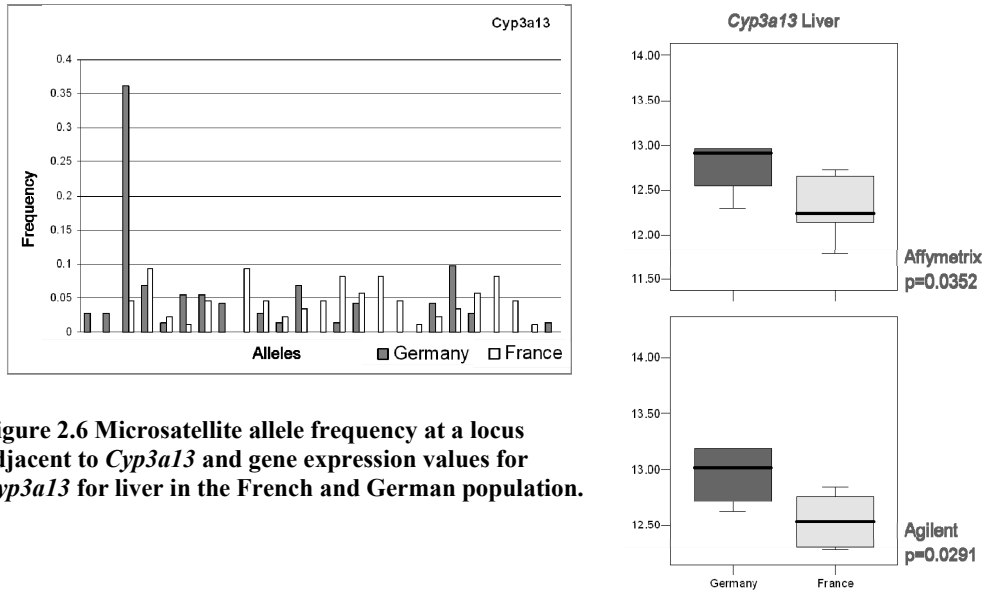


Figure 2.6 Microsatellite allele frequency at a locus adjacent to *Cyp3a13* and gene expression values for *Cyp3a13* for liver in the French and German population.

Sequence data was analyzed for 10 French and 9 German individuals. Approximately 600 bp of non-coding DNA was sequenced in the upstream region of *Cyp3a13*. Both populations are polymorphic at 6 sites and they do not show any remarkable differences in nucleotide diversity or nucleotide differences (Table 2-4). Hence the data do not indicate any signs of positive selection in the German population. Contrarily in the French population Tajima's D test results in significantly positive values, *i.e.* an indication for balancing selection.

2.3.3 Expression of hepatic Cyp450 regulatory genes

In addition to the investigation of Cyp450 genes, expression of Cyp450 regulatory genes which affect hepatic Cyp450 activity in response to xenobiotic compounds was surveyed in the two focal populations (HNF4α, PXR, PPAR and CAR). Three nuclear receptors were expressed in the liver samples of both populations (Figure 2.7), whereas PPAR was not expressed in any population.

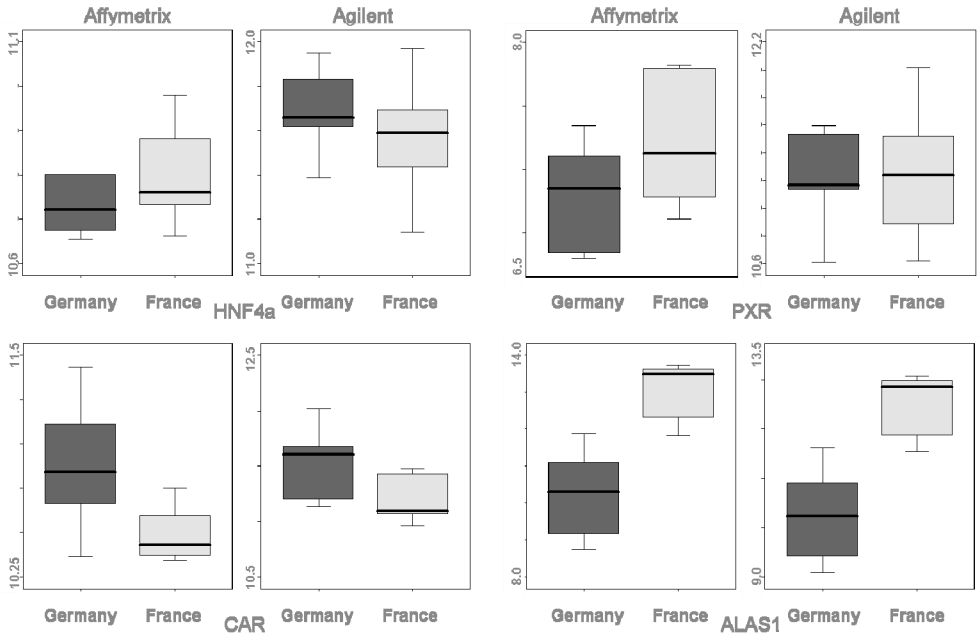


Figure 2.7 Expression of Cyp450 regulatory genes in liver for the French and German samples on both platforms.

While HNF4α and PXR did not show any differences in expression, the concentration of CAR transcripts were approximately 1.5 fold higher in the German population. This difference in expression was significant on both platforms ($p<0.05$).

Finally the heme synthesizing enzyme *ALAS1* was investigated. For this enzyme the expression of the French population is remarkably higher than in the German population (6-fold increase Agilent and 9-fold increase Affymetrix). *ALAS1* was also observed as expressed in the other tissues (brain and testis) but contrarily to liver both populations show equal levels of expression for this enzyme among these tissues.

2.4 Discussion

To gain more insights into the process of adaptation to different ecosystems a candidate gene approach was conducted comprising genes of the Cyp450 gene superfamily which code for detoxification enzymes. Tracing adaptive events was based on a microsatellite screen as well as comparisons of gene expression and sequence data between two house mouse populations. According to ln*RH* statistics microsatellite analysis revealed signs for selection in three out of the seven Cyp450 gene clusters, Cyp2c, Cyp2j and Cyp3a. Taking a closer look at these gene families, some particular genes were found to be candidates that have been influenced by natural selection. For *Cyp2j6*, *Cyp3a25*, *Cyp3a59* and *Cyp3a13* evidence for adaptive changes could be supported with different methods of analysis.

Cyp2j cluster: In mice genes of the Cyp2j subfamily are clustered on chromosome 4. To date, eight functional genes are described (*Cyp2j5, Cyp2j6, Cyp2j7, Cyp2j8, Cyp2j9, Cyp2j11, Cyp2j12* and *Cyp2j13*) each of them consisting of 9 exons, all arranged in the same orientation (Nelson et al. 2004; Ioannides 2008). (For detailed information refer to http://drnelson.utmem.edu/CytochromeP450.html) Several indications for positive selection having acted in this cluster were found in the French population.

Analysis of microsatellite allele frequencies shows a loss of variability exclusively in the French population at several loci. In total, five loci displayed an

allele pattern which can be interpreted as a selective sweep. Disregarding locus 11, which is monomorphic and hence not informative, they are located contiguously around *Cyp2j6* and *Cyp2j9*. According to ln*RH* values they generate a classical valley pattern with the tipping point right next to *Cyp2j6*. It should be mentioned that the data have not been corrected for multiple testing, *i.e*. microsatellite loci were chosen as outliers on a 5 % level but are not statistically firm. Moreover three of the five putative sweep loci do not display p-values <0.05. However the high mutation rate of microsatellites keeps the time frame in which a selective pattern is detectable with the statistic rather small. With time after a certain allele was driven to fixation the '*footprint*' of selection is gradually blurred by new arising mutations. An allele pattern arises, which is characterized by an increase of new alleles, which occur in a few-repeat-steps distance around the sweep allele. Such 'recovery pattern' was for example found at locus 9.

Similarly the allelic distribution at locus 8 and 12 might have been generated by recurring mutations and recombination events. This observation is concordant with the finding that recombination isolates selective sites from neighboring chromosomal regions, *i.e.* the size of a sweep and the magnitude of the reduction in variability expected after a selective sweep depends crucially on the rate of recombination (Schlötterer 2002). Thus, although the microsatellite data did not show statistical significance, the allelic patterns of the investigated loci clearly indicate a selective event having affected the chromosomal region containing *Cyp2j6* and *Cyp2j9*.

The available gene expression analysis provides further information on the sweep candidate genes. Cyp2j genes are mainly expressed in the epithelial cells of the small intestine and at lower levels in the liver, heart, lung, brain, and kidney (Scarborough et al. 1999). Concordantly both genes, *Cyp2j6* and *Cyp2j9*, were detected to be expressed in liver and brain but not in testis samples. While *Cyp2j9* does not show any differences in expression values between the French and German population, *Cyp2j6* expression significantly differs in liver and brain between the two sample sets. Since the investigated tissues are not primarily responsible for the first-pass metabolism of nutrients and xenobiotics, as it is the intestinal epithelium (van den Bosch et al. 2007), and accordingly show less expression intensity for *Cyp2j6*, an adaptive change targeting gene expression especially in brain would intuitively not be expected. Hence it might be speculated that the observed shift in gene expression

values is generated as a 'by-product' of an adaptive change in the small intestine. However this assumption is in conflict to the notion that significant changes of expression levels are modulated in a tissue-specific manner (Whitehead and Crawford 2005; Staubach et al. 2010). It has been shown that expression of Cyp450 genes is highly variable between tissues and inducers of Cyp450 genes are acting selectively on different tissues (Qiang Xie et al. 2000; Maglich et al. 2002). Hence deducing gene expression levels for the small intestine is not possible. Still, the data clearly display changes in gene expression in liver and brain which may be taken as additional indication for selection having acted at this gene. Even though it is unclear whether the shift in expression itself was the target of selection, or of if the expression changes are rather a side effect of an adaptive change.

Polymorphism and haplotype data confirm the microsatellite data. On a first glimpse the French population seems more variable in basic polymorphism analysis, *e.g*. it comprises many segregating sites which result in increased nucleotide diversity (π and Θ), and values of Tajima's D would suggest that both populations evolved under neutrality. However, Fu and Li's D test revealed significant values in the French population, which indicates a significant excess of singletons in this population, and thus might hint to positive selection.

The inconsistency between the tests can be explained by the fraction of time in which Tajima's Test detects positive selection. As Tajima's statistic computes the difference between π and Θ and the average number of mutations between pairs in the sample (Tajima 1989), this test will give significant results for positive selection if the ratio between π and Θ is high. This is obtained if, for example, one allele is at high frequency and few are at low frequency. In terms of allele frequency distribution Tajima's D traces recovery patterns to identify positive selection. The observed allele frequencies of microsatellite locus 6 indicate that this locus has not reached fixation so far. Although one allele is almost completely fixed in the French population some alleles with larger size still persist at this locus that are not expected to be retained from new mutations, since they are several mutation steps apart. This suggests that the selected site did not sweep through the whole population yet. Furthermore, haplotype analysis revealed another indication for a selective event. Despite the high sequence divergence in the French population, haplotype analysis pointed into the opposite direction: both, haplotype diversity and the number of haplotypes are significantly

reduced. Hence, the high diversity in the French population results from a distinct but prevalent haplotype (see Chapter 2.3.1, Table 2-2). With respect to the haplotypes the French samples split into two types: the abundant haplotype version, which is rather distinct from the German samples, and another, rare type which is similar to the common German haplotype. Since the latter one is present in both populations this one may be assumed to be the ancestral type; the first is most likely the haplotype which has originated and spread through the French population. The combination of high sequence divergence but low variation within the derived haplotype could be explained by selection having operated on an allele arisen in a partially isolated population.

Both, reduction in microsatellite variability as well as low variation within the most abundant haplotype, were examined in the 5´ part of the gene. This observation as well as an increase of haplotype diversity in the upstream region of exon4 hints towards the strongest signal of selection is located in the first part of the gene. No nonsynonymous amino acid changes were detected in any of the nine exons, supporting the hypothesis that selection acted rather on alteration of expression than on protein level, namely on a *cis*-regulatory element. That changes in *cis*-regulatory sequences could be an important source for genetic adaptation is widely acknowledged (Wilson et al. 1974; Tautz 2000; Wray 2007).

Combination of haplotype data with gene expression data indicates that gene expression of French individuals, which carry the German-like haplotype, correspond to German expression levels. This may imply that the change in gene expression is linked to the adaptive event. As has been said, it is unclear whether the gene expression itself was targeted by selection, or if linkage of expression level to the beneficial mutation invoked the change. In other words, that the expression level 'swept' as a selective 'by-product'.

However since the finding that selection rather acts on the 5' part of the gene, *i.e*. in the promoter region and no evidence for changes in the protein sequence could be found, it may be speculated that selection acted on altering the expression level; more precisely on a downregulation of this gene. *Cyp2j6* provides oxidative activity of xenobiotics by metabolizing the aromatic amine benzphetamine to formaldehyde. As has been described in the introduction, metabolism of toxic compounds may also lead to generation of pathological molecules, especially in reduction of Phase II

reaction. Hence, especially if a less poisonous alternative for the detoxification of aromatic amine is available, a downregulation of *Cyp2j6* could become selectively beneficial.

Cyp3a cluster: The region which comprises the Cyp3a family is located on chromosome 5 in the mouse genome. So far seven functional genes (*Cyp3a13, Cyp3a57, Cyp3a16, Cyp3a41, Cyp3a11, Cyp3a25* and *Cyp3a59*) and some pseudogenes have been described. Genes are highly duplicated and arranged in different directions. Moreover, this cluster is expanded over a large range, *e.g. Cyp3a13* is located in approximately 8 Mb distance to the remaining genes. Therefore, *Cyp3a13* was analyzed separately. High sequence similarities, which are illustrated in the dotplot, complicate the investigation of this cluster. Since the annotated mouse genome is not fully complete, ongoing sequence modification in this region can be observed and still, there are some gaps in the reference sequence. The present study reveals several indications for positive selection having acted in this rather young gene cluster.

Due to high sequence similarities caused by the recent duplication events, only six microsatellites could be studied within this cluster. Still, three of them show signatures of selection. Two of them (1.2 and 8) are located in an intron region within a gene and another in close proximity to *Cyp3a13*. Locus 8 is located in the intron region flanked by exon 11 and 12 of *Cyp3a59*; locus 1.2 is located adjacently to the 5`region of *Cyp3a25*. Both loci showed significant reduction of variability in the French population. Corresponding allele frequencies reveal similar allele spectra for both loci whereas the shortest allele is at high frequency. As already seen at locus 6 in the Cyp2j cluster some alleles of larger size are still present in the French population. Due to microsatellite mutation behavior, short alleles have the tendency to mutate to longer allele size (*focal length*) (*e.g*. Calabrese and Sainudiin 2005) Since the shortest allele is fixed in both cases and the larger ones are distributed in a way that fits the stepwise mutation manner, the alleles presumably emerged after fixation of the sweep allele. Thus the data may represent the state right after the spread of the beneficial mutation through the population at which variation is recovered by new mutations at the focal sites.

Further evidence of adaptative changes within these genes was obtained by sequence analysis. Two fragments of cDNA were examined of *Cyp3a25*. One

including the 5' part of the gene and the other the 3'end. While the fragment located in the 5' region shows a relatively high polymorphism in both populations, only two polymorphic sites were observed in the fragment that includes the 3' end. The primary sequence reveals indications of recent selection affecting *Cyp3a25* in the French population. First, significant haplotype diversity was detected with coalescent analysis. Second, Tajima's D revealed clearly negative results, indicating positive selection. It should be noted, that the results for Tajima's D were not significant, but due to the young population split, the test is not expected to be very powerful. As has been described above, Tajima's D detects 'recovery' patterns. The negative values are thus confirming the above described microsatellite pattern. However, the number of newly arisen microsatellite alleles is very low which indicates that the 'recovery-process' only recently started. Considering the lower mutation rate of SNPs (Nachman 1997), the time since the fixation of the beneficial mutation would have been too short to generate a pattern on SNP level at which values for Tajima's D become significant.

Regarding the high rate of nonsynonymous changes within the 5' region it is likely that there are advantageous mutations among these changes. Four splice variants are described for *Cyp3a25* and for three of them a protein product is described (Ensemble 2010). All three proteins contain at least the first four exons. Hence, transcribed variants should be affected by the aminoacid change.

Observing the last three exons of *Cyp3a25* only few polymorphic sites were detected in both populations, which are not sufficient to draw statistical conclusions.

Due to sequence similarity designing unambiguous primers to sequence *Cyp3a59* was not possible. These similarities are clearly illustrated in the dotplot. Estimates about the age of the duplication event confirm that *Cyp3a59* recently evolved from *Cyp3a25*. It is known that gene duplication is one mechanism for the evolution of new gene functions (*e.g.* Ohno 1970, Lynch and Conery 2000). If two gene copies are present the theory suggests three alternative outcomes: (a) nonfunctionalization: one of the copies loses its function and becomes a pseudogene; (b) neofunctionalization: one copy may acquire a novel, beneficial function and thus will be preserved by natural selection while the other copy retains the original function; (c) subfunctionalization: both copies become partially compromised by mutation accumulation to the point at which their total capacity is reduced to the level

of the single gene-copy, *i.e.* the copies 'share' the initial gene function (Wen-Hsiung Li 1997; Lynch and Force 2000). The present case of *Cyp3a25* and *Cyp3a59* might be interpreted as scenario (b). Observing signs for selection may be taken as a hint towards a neofunctionalization, *i.e.* one gene copy gains a new function. This hypothesis would be further supported by the supposed changes on protein level. Unfortunately gene expression data had to be excluded from the analysis since the probe sets for the gene did not match the respective gene sequences unambiguously.

Another candidate gene was identified within the Cyp3a family: *Cyp3a13*. Here, the adjacent microsatellite locus was significantly reduced in variability for the German population. In contrast to the other loci that showed significant ln*RH* values, a high variability persisted at this locus in the affected population. This pattern cannot only be explained by detecting the selective event at an early stage. Moreover maintenance of high variability at a certain locus could be interpreted as a sign for balancing selection. However, while the French pattern would resemble the classical allele pattern for balancing selection, it would not be expected to find one allele in such exceptional high frequency.

Cyp3a enzymes are mainly present in liver and gut and are induced by various xenobiotics (Moore et al. 2000). Similarly *Cyp3a13* exhibit a wide tissue distribution with predominant expression in liver (Anakk et al. 2003). Concordantly, expression for *Cyp3a13* was detected in liver and brain, respectively. Significantly higher expression values were observed in the German population in liver, whereas brain showed no difference between the two populations. Tissue specific changes in gene expression are commonly observed (Staubach et al. 2010) and might be taken as indication for selection. While no further sign for positive selection could be detected, basic polymorphism analysis revealed significantly positive values for Tajima's D, *i.e.* further indication for balancing selection acting in the French population.

As already mentioned expression of Cyp450 genes can substantially vary among tissues and is mediated in a tissue specific manner (Anakk et al. 2003). Nuclear receptors play an important role in detecting xenobiotics and stimulating genes encoding cytochrome P450 enzymes (Waxman 1999). Hence, expression levels of four nuclear receptors that are centrally involved in mediating expression of hepatic Cyp450 genes were additionally examined (HFN4α, CAR, PXR and PPARA).

The orphan nuclear receptor HNF4α is an important regulator which

coordinates nuclear-receptor−mediated response to xenobiotics in the Cyp3 family (Tirona et al. 2003). PXR and CAR are activated by a wide range of xenobiotics (Maglich et al. 2002). PXR was originally shown to regulate the expression of Cyp3A isozymes and is currently known to regulate the expression of several additional genes involved in xenobiotic metabolism. Cross-talk is observed for the two receptors, where PXR regulates CAR expression (Maglich et al. 2002). PPARA is a receptor which is, among others, involved in down-regulation of *Cyp3a13* expression (Anakk et al. 2003).

Furthermore *ALAS1* was included in the analysis, since it is known to be a rate limiting enzyme in heme biosynthesis (Yin et al. 2006). Cyp450s require heme as a cofactor to oxidize substrates and it has been shown, that *ALAS1* is induced in response to various xenobiotics (May et al. 1995).

The expression data did not show significant differences in expression levels for the two receptors HFN4α and PXR between populations. Notably, PPARA could not be detected in any of the samples. However, amounts of CAR and *ALAS1* gene transcripts differed significantly. While CAR was found to be more highly expressed in the German liver samples, *ALAS1* expression was drastically higher in liver tissue of the French population. Expression for *ALAS1* was also observed in brain and testis where both populations show similar expression values, *i.e*. the shift in expression seems to be tissue specific. Note that all specimens were kept under standardized conditions. Thus observed expression differences cannot result from exposure to different substrates. Furthermore all mice were sacrificed at similar times of the day. Therefore the possibility that shifts in expression may be due to daily oscillation, such as has been observed for hepatic mRNA expression of nuclear receptor CAR (Kanno et al. 2004), can be rejected.

Whether the differences in gene expression result from selective events can only be speculated. The here applied expression assays only detect the presence of a certain gene transcript. Thus the amount of quantitative biological activity of the corresponding protein remains unclear. However, concerning the orphan receptor CAR, higher protein availability could enhance the potential of Cyp3a activation and therefore improve the responsiveness to xenobiotics (Wen Xie et al. 2000). Relating to *ALAS1* higher expression could have a considerable impact on Cyp450 gene efficiency. The protein is described as rate limiting factor for heme biosynthesis (May

et al. 1995), which is a key compound for Cyp450 substrate oxidation (Guengerich 2007). Despite all possible benefits, it should be pointed out that alteration of Cyp450 induction which is associated with shifts of regulatory activity, would comprise a high risk: it has been shown that the perturbation of endogenous regulatory circuits potentially implicates pathophysiological consequences (Waxman 1999).

To recapitulate, the present study used different molecular methods to examine selection in two distinct populations in response to xenobiotics. Microsatellite as well as polymorphism analysis and expression data point towards a selective event having acted in at least four Cyp450 genes that function in detoxification of xenobiotic compounds: *Cyp2j6, Cyp3a25, Cyp3a59* and *Cyp3a13*. While the latter one is most likely affected by balancing selection the other three reveal clear indication for positive selection. Both, evidence for selection on *cis*-regulatory elements as well as at the protein level have been found. Furthermore, the time since the respective genes originated substantially varies, *i.e*. *Cyp2j6* and *Cyp3a13* are old gene copies whereas *Cyp3a25* and *Cyp3a59* originated from a recent duplication event. Haplotype tests as well as ln*RH* statistics are methods that detect relatively young selective events (Fay and Wu 2000; Otto 2000), nevertheless, recent positive selection was traced in both, a rather old gene duplicate as well as in two young gene copies. Hence coping with dietary components does not seem to affect specifically novel gene copies. While there is evidence for selection acting on a *cis*-regulatory element in *Cyp2j6*, which represents an already diverged gene copy, adaptive changes in the recently duplicated gene copy *Cyp3a25* presumably result from altered protein sequence. This might suggest that selection on old gene duplicates, which have already evolved a unique gene function, rather modifies expression characteristics, so the function will not be lost, whereas recently duplicated genes still provide the adaptive potential to evolve new gene functions. Further finding signs for selection in evolutionary old genes might be interpreted related to the ‘Red Queen model’ which describes a continuous need for adaptations due to a permanent arms race between parasites and their hosts or, as in this case, between plants and animals feeding on them (Van Vaalen 1973).

Finally it should be noted that all signs for selection were detected in the French population. Target genes belong to two different gene clusters (Cyp2j and

Cyp3a), *i.e*. they are functionally independent. Thus at least two selective events point independently into the same direction: The French population having evolved genetic responses to environmental factors. Finding signals for adaptation denotes that there had to be a driving selection pressure. As described previously, potential challenging factors are plant secondary metabolites and plant associated fungi. Hence it may be speculated that the French population had to cope with specific dietary compounds which led to the genetic change. Adaptation to specific pesticides or insecticides as it has been shown for *Drosophila* (Amichot et al. 2004), is unlikely in this case. Regarding the comparatively long generation times of mice, the time since these chemicals have been used would not have been sufficient for a beneficial mutation to sweep through the population; unless it would have been invoked by very strong selection pressure. This can be rejected since recent, strong selection coefficients would lead to the joint fixation of large chromosomal regions, *i.e*. to stronger signs of selection.

3 Sequencing Microsatellites using 454 Techniques

3.1 Introduction

During the past decade microsatellites have become one of the most popular molecular markers used in a wide range of fundamental and applied fields of biology and medicine (Tautz 1989; Bowcock et al. 1994; Hirschhorn and Daly 2005). High polymorphism and the relative ease of typing are two features that characterize microsatellites as suitable markers for the application in a wide range of studies including forensics, molecular epidemiology, parasitology, population and conservation genetics, genetic mapping and genetic dissection of complex traits (Ellegren 2004; Lucchini et al. 2002; Richard et al. 2000).

However, especially in studies that depend on high throughput for many different loci their use can be costly in two ways. First, conventional microsatellie typing requires the application of fluorescently labeled primers which is comparatively expensive. And even though the adjustment of product size and combination of different labels allows multiplex analysis, only few microsatellites can be run simultaneously. Second, analysis of microsatellite gels is fairly time consuming. Although there are user-friendly analysis programs available, each newly established microsatellite locus has to be manually inspected (Johansson et al. 2003).

The aim of this study is to present a time and cost efficient way of screening high numbers of microsatellites using a next generation sequencing platform. The 'Genome Sequencer™ FLX System from 454 Life Sciences™ and Roche Applied Science' is capable of generating more than one million high-quality reads per run and read lengths up to 400 bases. Hence, this system allows potentially sequencing of several hundreds of microsatellite loci for a multitude of individuals in a single run. Furthermore, since sequence information rather than band pattern information is available, it is possible to perform the analysis of the output sequences entirely automatically. For this study an algorithm was designed which determines microsatellite alleles and calculates allele frequencies without any manual inspection required. These data can be directly applied to basic population genetic statistics.

As a preliminary study about 800 microsatellite loci were sequenced using 454 techniques in six samples of house mouse (*Mus musculus)*. To verify the results several reference loci were typed using conventional protocols and results from both methods were subsequently compared. I found that the 454 approach produces proper output for analyzing microsatellites and that the data obtained by either method are fairly consistent. Moreover, sequencing microsatellites extends the range of standard statistics which can be calculated. Thus, sequence availability allows the estimation of parameters such as recombination and mutation rates (RR and μ) or components of F-statistics.

3.2 The Method

A new time and cost efficient method to carry out high throughput analysis of microsatellites was established based on the 454 sequencing technique. Therefore, tools were developed to perform time consuming steps in an automated manner. (1) A program was designed to search for microsatellites and associated primers in a given sequence under user defined conditions. To obtain full length sequences of the amplicons, the length for PCR products was restricted to 300 bp, considering an average read length of 400 bp for the applied GS FLX Titanium (Roche 2010). (2) Allele calling and basic microsatellite analysis was performed in an automated manner, applying a specially designed computer program. To verify the data obtained with this novel procedure, a subset of loci was additionally typed and conventionally analyzed serving as reference data.

In total two 454 runs were performed. One pilot run to test whether the method is practical and a second run with slightly modified conditions to improve the output.

Primerdesign: The program '*msfinder.pl*' involves mainly two steps which are illustrated in Figure 3.1. First, the Perl script searches for microsatellites in a given sequence. Any sequence that is present as a continuous string can be used as input. Optionally, the microsatellites can be filtered after various criteria specified by the user. After selecting suitable microsatellites, PCR primers are created under certain

conditions to amplify these loci. The primer sections are screened such that there are no secondary repeats or mononucleotide stretches and subsequently checked for unintended additional PCR products. Second, after a list of all possible loci is generated, any number of loci is picked from these in a way that the markers are evenly distributed along the target sequence.

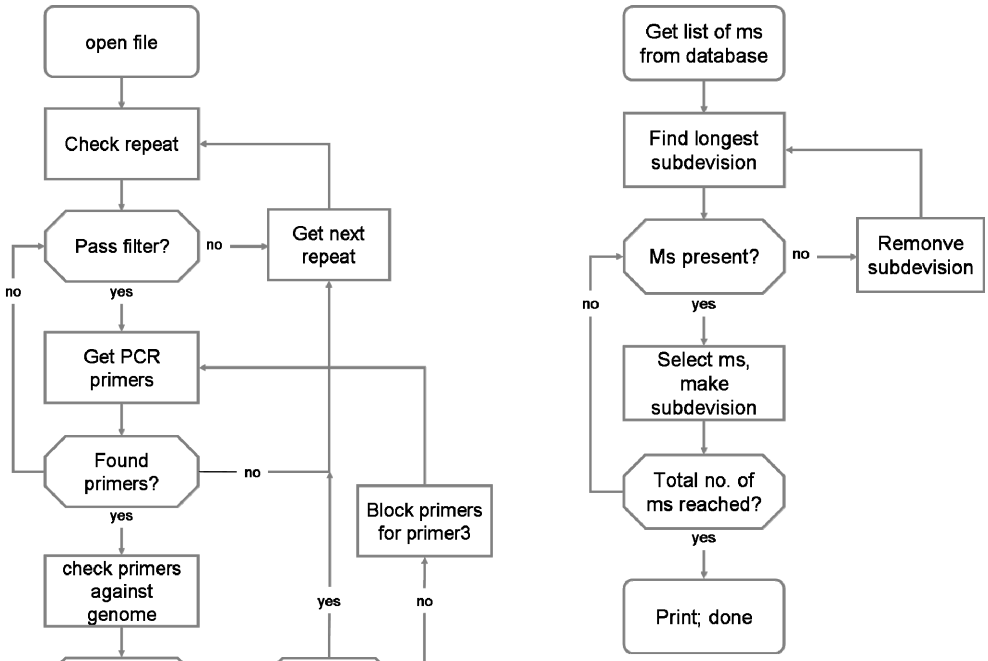


Figure 3.1 The script '*msfinder.pl*' contains two main steps: First (left), primer design for all microsatellites that fit the given criteria. This step consists of the following tasks: extract appropriate microsatellites from database, get a sequence window around the microsatellite from the input file, check for other repeats within the selected sequence, get PCR primers from Primer3. Second (above), a defined number of microsatellites is selected from the output of the above series, including the following steps: find the microsatellite closest to the center of the sequence, save this microsatellite to the output list, bisect sequence at the microsatellite location, repeat for all smaller sequences, print output file.

The script accesses a certain set of programs. To run '*msfinder.pl*' the following tools are required: trf404.linux.exe of the Tandem Repeats Finder (trf) program (or the 64 bit version) (Benson 1999), primer3_core.exe of the Primer3 package (Rozen and Skaletsky 1999), re‐PCR of the NCBI e‐PCR package, famap

and fahash to generate hash file (Schuler 1997), SQLite3. In addition some Perl modules that are not included in the standard distribution have to be installed: DBI and DBD::SQLite, Bio::DB::Fasta from the BioPerl package and Config::Easy.

Microsatellite criteria: Microsatellites picked from the database were filtered according to the following conditions: Repeat unit length was adjusted such that only di- tri- and tetranucleotides were chosen. The number of repeats was set to a minimum of 7 repeats and a maximum of 20 in the first run and 13 in the second, respectively. Furthermore only microsatellites with a minimum percent alignment match of 85% are selected. The percent alignment match predicts the correctness of microsatellites based on occurring SNPs or indels.

Sample preparation: Different sample sets of the house mouse (*Mus musculus*) were incorporated into the study. Each sample set consists of pooled DNA from 40 individuals. DNA was isolated by standard salt extraction and stored at –20°C. For further processing, the concentration of DNA was adjusted to 2 ng/µl and subsequently pooled. Amplification was done following the QIAGEN (Valencia, CA) multiplex kit manual (cat. no. 206143). Single PCR products were pooled for each sample set resulting in six solutions containing all amplicons of one sample set. Salt precipitation was performed and the purified DNA was run on an agarose gel. For gel extraction bands of expected product size (150-300 bp) were cut out and the QIAGEN (Valencia, CA) MinElute PCR Purification Kit (cat. no. 28004) was applied to extract the DNA following manufacturer's instructions.

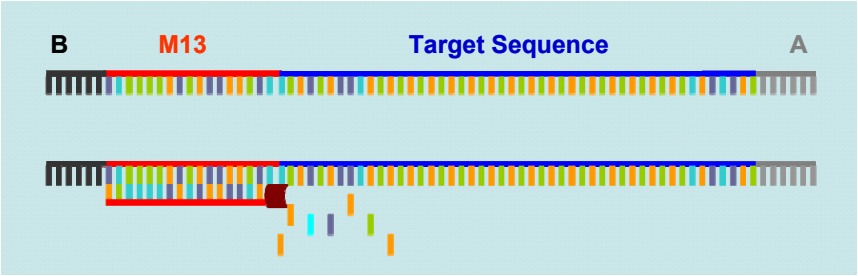


Figure 3.2 Additional single PCR step turns the single stranded library into double stranded to prevent agglutination of the microsatellite amplicons. A modified B-adaptor construct was used to optimize binding of the adaptor to the bead. An M13 primer is used as a spacer between the target sequence and the B-adaptor and serves as the priming site for the single PCR step.

Sequencing: 454 sequencing was conducted using *GS FLX Titanium* series reagents. Library preparation was performed following the manufacturer's protocol but using modified B-adaptors. Modification of the B-adaptors was developed

because capturing the sequences by binding the standard B-adaptors to the bead failed in the first run. Microsatellite amplicons have the tendency of sticking together and thus can form DNA complexes which could cover the 454 adaptors. To avoid agglutination of the B-adaptor with the amplicons, a single PCR step was performed turning the single stranded library into double stranded templates. In order that the B-adaptor can directly bind to the bead an M13 primer was integrated into the B-adaptor to serve as a priming site for the previously mentioned PCR step (Figure 3.2). After amplicon preparation, the libraries were sequenced using the 454 Sequencing System (Margulies et al. 2005).

Sequence analysis: The analysis involves three steps: matching the 454 sequences to a reference, checking reads for microsatellites and controlling the read length. First the 454 output is blasted against a reference database which contains all expected PCR products. As blast cutoff an E-value of $1e^{-35}$ was chosen. Reads that match two different reference microsatellites with E-values less than $1e^{-10}$ apart are discarded, otherwise the top hit is used for matching. The option –F, filtering of low complexity sequences, is turned off during the blast search to extend hits past repeats. If two or more high-scoring segment pairs make up one hit, they are fused into a single hit.

After that, the matched reads are checked for microsatellites using trf once more. All resulting repeats are tested for accordance with the reference data. Since SNPs at the beginning of the repeat may change the repeat unit as reported by trf, for example from ATT to TAT, all permutations of the reference repeat are considered. Checking the read length is straight forward by querying 8 bases following the detected repeat. In this way it is ensured that the microsatellite is flanked by non-microsatellite sequence, *i.e.* the complete microsatellite allele is detected and artificial alleles due to truncated reads are excluded.

Reads that passed the filter were used for standard microsatellite analysis. Once again the trf was applied to count the number of repeat units in each sequence which are considered as alleles. For further analysis the sequences had to pass one last filter which excludes all loci below a certain coverage, to assure that a representative allele sample is obtained from the pooled DNA. Here a minimum of 20 reads per microsatellite locus was chosen to assure a proper representation of alleles. Based on this final data set, allele frequencies were calculated for each locus.

Sample preparation of the reference set: Sample preparation was identical to the previously described procedure. Forward primers were replaced by 6-FAM and HEX labeled ones. Two primer pairs (one of each label) were run together in one reaction. The amplification products were typed on an ABI3730. Allele calling was performed using GeneMapper software 4.0 (Applied Biosystems 2009) followed by visual inspection of the electropherograms. Allele frequencies were calculated applying MSAnalyzer 3.15 (Dieringer and Schlötterer 2003).

3.3 Results

Two 454 runs were processed, run1 with a total set of 1,000 loci and run2 with approximately 800 loci, respectively. Each set of loci was performed with pooled DNA of six different sample sets of the house mouse (*Mus musculus*). The 454 plate was split into separate lanes where each sample was assigned to one lane.

Out of 80,000 potential reads, on average 70,000 raw reads (*i.e.* quality checked by the internal base caller) were obtained per population in run1. Quanitatively the second run performed worse whith a middling amount of 54,000 reads. In contrast, the read quality was higher, thus after a stringent internal filtering process (see Chapter 3.2) more suitable reads retained for run2 (appropriate reads run1 = 30,000; run2 = 38,000).

The histogram shown in Figure 3.3 displays the distribution of reads that passed the filtering process among the used loci. It can be observed that the reads are not equally shared between the loci and the number of reads assigned to each locus varies a lot, from below 10 to over 400. Notably, there are many loci at the extreme ends of the distribution in run1, while the coverage is more evenly distributed in run2. This affects especially loci that are found in classes of low coverage. Since loci with a poor coverage will not properly display allele representation, all loci are discarded below a coverage cutoff of 20 reads. The figure shows that substantially more loci are lost due to low coverage in run1 than in the second run.

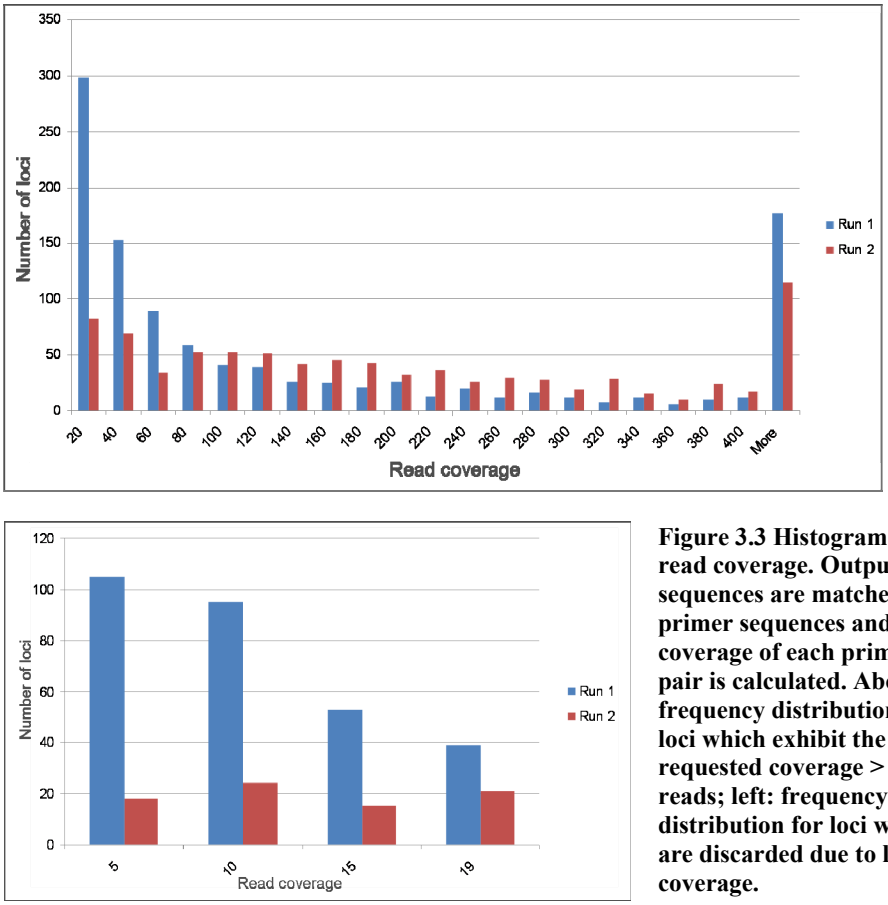


Figure 3.3 Histogram of read coverage. Output sequences are matched to primer sequences and the coverage of each primer pair is calculated. Above: frequency distribution for loci which exhibit the requested coverage > 20 reads; left: frequency distribution for loci which are discarded due to low coverage.

The final number of loci which are suitable for analysis are shown in table Table 3-1. The second run comprises considerably more suitable reads, which results in approximately 20 % more loci that can be properly analyzed. This is most likely an effect of modified microsatellite criteria used in the second run, which will be described later on.

Table 3-1 Number of reads and microsatellite loci which could be taken into the analysis for both 454 runs.

	Total reads	Suitable reads [%]	Total loci	Loci coverage > 20	%
Run 1	69,919	43	1,068	369	37
Run 2	53,915	70	833	468	59

Read extraction: Table 3-2 displays the different filtering steps that were used to extract only informative reads (see Chapter 3.2). Furthermore it shows the percentage of reads that failed in each step for each run. Approximately 50 % of the reads can be matched and analyzed as described, the other half gets lost during the different filtering steps. The main fraction of reads is discarded during the blast

search. About 20-30 % of all reads are filtered out because they are not caught during this step. These are mainly too short reads or artefactual PCR products. The E-value of $1e^{-35}$ which was chosen for the matching process is rather high assuring monitoring only unambiguous hits. At the same time such a high E-value discriminates against very short sequences. Since the reads are matched against an internal database which contains only the target sequences, side products of the PCR as well as contamination are discarded during this step. Regarding run1, another relatively high fraction of reads (10 %) is lost because they are too short in a sense that they lack the requested 8 bp after the microsatellite repeat pattern. Due to modified settings in the microsatellite search, which will be described later, this number could be substantially reduced in run2.

Table 3-2 Filtering steps that were used to extract proper reads and the percentage of reads that failed during each step.

Read Filtering Steps	Discarded Run1 (%)	Discarded Run2 (%)
Matching		
multiple blast hits	01	05
no blast hit	30	23
no repeat at all	04	05
Repeat check		
repeat of wrong type	02	04
more than one repeat	01	03
Length check		
too short	10	05
Total	48	45

To fathom the fate of reads that cannot be assigned to any of the reference loci by the initial blast step an additional blast search was performed. All reads that did not match any reference locus were blasted with a considerably lower E-value ($1e^{-12}$) against the reference loci once more. Thereby reads that exhibit the expected PCR product but are too short to be caught previously are detected. Subsequently all reads that could not be assigned in this step were blasted against the Mouse genome database and next, if no hit was detected, against the NCBI non-redundant (nr) database. The results are shown in Table 3-3. One sample from run1 is presented as an example. Note that the following quantities refer to a subset of total reads, namely only those that had no blast hit. Applying the lower blast cutoff, more than 50 % of the formally unassigned reads match the reference database, *i.e.* these reads contain the designated sequence but are too short. The blast against the mouse genome

reveals 20 % of PCR side products and 15 % external DNA contamination.

Table 3-3 The proportion of reads that did not match in the first blast search were subsequently blasted with changed conditions and to other databases.

Database	match (%) E-value <$1e^{-12}$	no match (%) E-value <$1e^{-12}$
Reference loci	58	42
Mouse genome	20	23
NCBI nr database	15	08

As mentioned previously, in run1 a relatively high fraction of reads was lost because they lack the requested 8 bp after the repeat pattern. Absence of sequence after the microsatellite is predominantly caused by truncation of the read directly after the repeat, apparently because the base caller cannot read the sequence following the repeat. One likely explanation for this observation is slippage occurring during 454-amplification on the bead.

Since slippage increases with repeat size the probability of reads to be truncated is expected to be higher for reads containing long microsatellite stretches. To test for this, sequence length after the repeat pattern is plotted against the number of repeats (Figure 3.4). Here again, one sample out of run1 was taken as an example. Note that in the microsatellite search the criterion for maximum repeat number was set to 20 repeats. In the final data repeat numbers up to 80 are observed, indicating that repeat numbers for the wild mice considerably deviate from the annotated repeat numbers in the NCBI database.

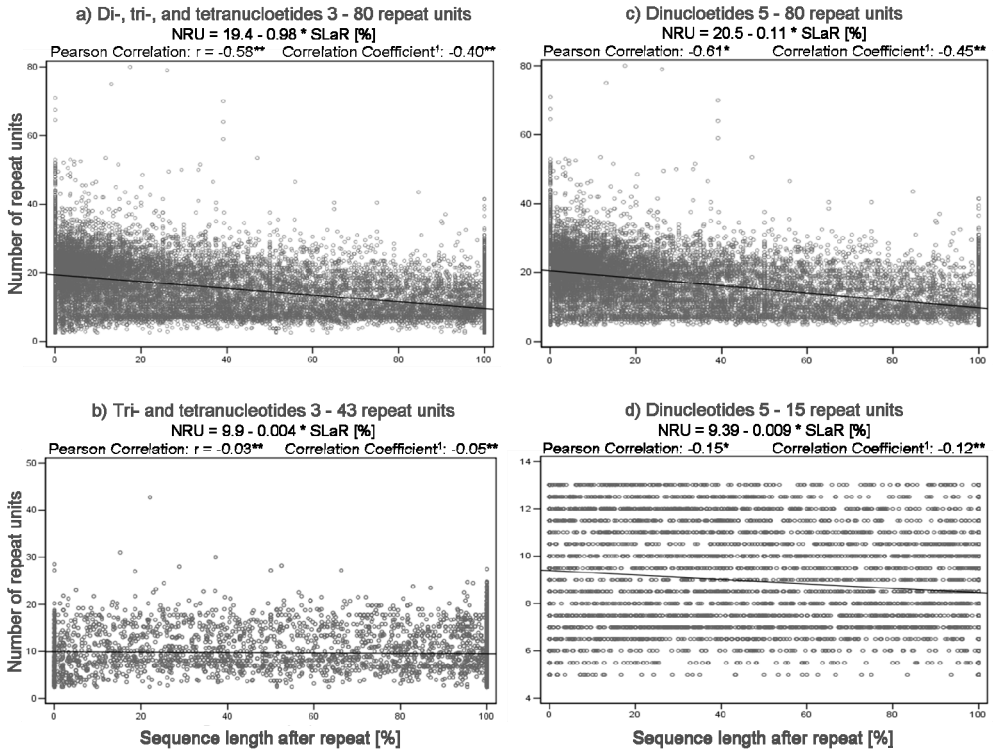


Figure 3.4 a-d Sequence length after the repeat pattern for different nucleotides is plotted against the number of repeats. Regression line was fit assuming a linear model. NRU = Number of Repeat Units, SLaR = Sequence Length after Repeat. [1] = Kendall's Tau b.

To display the degree of reads that are lost due to truncation a regression line was fit through the data based on a linear model. Since the data does not fit a normal distribution (see Supplement 1) Kandall's Tau b was computed as a nonparametric test in addition to the Pearson Correlation. Even though more complex models might fit the data better the simple model was chosen to illustrate the shift in the rate of truncation between the different repeat classes. Four different tests were performed here (Figure 3.4 a-d). In Figure 3.4 a) the read length after the repeat pattern of all reads obtained from one sample is plotted against the number of repeat units (NRU) of the respective sequence. This test comprises all three categories of repeats which were considered in the study (di-, tri-, and tetranucleotides). As expected the correlation coefficients indicate a negative correlation between the two parameters. In Figure 3.4 b and c investigation of dinucleotides was separated from tri- and tetranucloetides. While the correlation between repeat number and readable sequence increases if taking only dinucleotides into account, the effect is not observed for tri- and tetranucleotides. Lastly a subset of dinucleotides was tested for which only sequences with repeat numbers up to 13 were considered (Figure 3.4 d). Compared to

Figure 3.4 c were all dinucleotides were investigated the correlation is strongly reduced. As aforementioned all tested groups show a negative correlation but to different degrees. Hence the probability of truncation increases with expanded numbers of repeat units whereby the effect is strongest in dinucleotides.

Differences between the three subgroups are obvious if focusing on the slope of the regression lines, which can be interpreted as the 'loss ratio' of reads. This slope is rather steep regarding dinucleotides including all repeat numbers (Figure 3.4 c). In contrast the slope of the 'loss ratio' decreases if only looking at lower repeat numbers in dinucleotides (Figure 3.4 d) and is almost completely dissolved in tri- and tetranucleotides (Figure 3.4 b).

Fully sequenced reads are included in the figures to demonstrate that despite this trend long microsatellite streches that are completely sequenced are still present in the data. However, although fully sequenced reads containing more than 30 repeat steps are observed for dinucleotides, the relative fraction of reads with higher repeat number is strongly reduced. Thus 40 % of reads which contain 10-20 repeats are completely sequenced while complete sequence for only 5 % is observed for reads containing 20-30 repeat steps (data not shown).

With respect to this outcome the conditions for searching microsatellites for analysis were modified in run2. Here the maximum number of repeats for microsatellites was restricted to 13. As mentioned above, the options for the microsatellite search can only be adjusted in relation to the reference genome. As shown, the repeat numbers of the reference data may deviate from the samples used in this study. Hence, setting a maximum repeat number will not completely exclude long microsatellite streches, but the average amount of shorter repeats will be elevated.

Additionally each repeat type was investigated individually. The following Table 3-4 contains the number of loci that passed the internal matching process (detected loci) and the number of loci that obtain a read coverage > 20. Whereas most loci are still detected after the filtering process, the majority of loci gets lost applying the coverage cutoff. Although a higher risk of truncation was observed for dinucleotides (Figure 3.4) the relative numbers of loci which achieve the required coverage are similar in each class. In both runs tetranucleotides perfomed slightly worse than the other repeat types.

Table 3-4 Number of different microsatellite types taken into the study and the proportion which can be taken into analysis.

Repeat type	Reference set	Detected loci	%	Loci coverage >20	%
Dinucleotide run1	677	658	97	263	39
Trinucleotide run1	119	116	98	43	36
Tetranucleotide run1	302	249	97	87	29
Dinucleotide run2	479	471	98	285	60
Trinucleotide run2	117	115	98	73	62
Tetranucleotide run2	260	259	100	151	58

Read analysis: Detection of microsatellite alleles was performed automatically. Whereas calling alleles in conventional typing measures the length of total PCR product, sequencing of the microsatellites allows focusing on the repeat sequence to detect alleles that result specifically from repeat number differences (see Chapter 3.2).

To verify the results of this approach, some loci were subsequently typed individually for 40 individuals and allele frequencies of both methods were compared. One example of each repeat type (di-, tri-, and tetranucleotides) is displayed for both methods in Figure 3.5 (further pictures of allele frequency comparisons are added to the digital supplement: Chapter3-4: Allelefrequencies). For each locus the detected alleles and the respective frequencies are illustrated for two different samples (Sample A and Sample B). Names of the alleles correspond with their distance to each other in base pairs. As only present alleles are illustrated some allele classes may be missing, *i.e*. adjacently displayed alleles might be more than one mutational step apart. On the other hand, alleles can be observed for which the distance is less than one repeat step. Alleles labeled with a star indicate assumed PCR artifacts caused by slippage events during sample preparation for 454 sequencing. While manual allele calling allows discrimination of slippage alleles, it is not possible to distinguish between ‘real’ alleles and artifacts in the automated search.

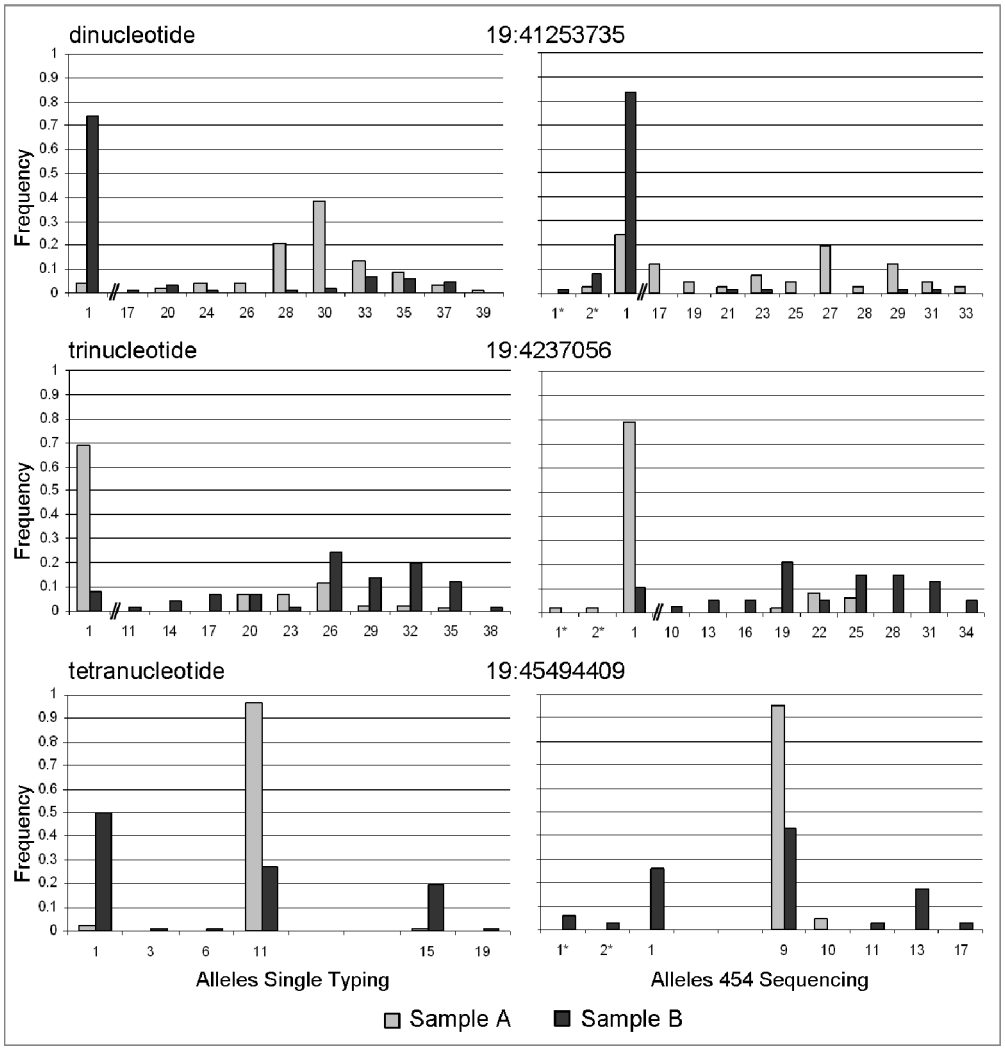


Figure 3.5 Comparison of allele frequencies of two sample sets obtained by allele typing (left) and 454 sequencing (right). One example is shown for each repeat type. Alleles are named after their distance to each other in bp. Grey: SampleA; black: SampleB. *= assumed PCR artifact

In the following observed allele frequencies will be compared between the two methods locus by locus: The microsatellite locus 19:41253735 contains a dinucleotide repeat pattern. Disregarding the two assumed slippage alleles, both methods revealed 11 different alleles. Whereas the methods are concordant in the number of alleles, they differ in allele frequency and size: Single typing revealed alleles of longer size than 454 sequencing. While the most abundant allele of Sample B is present in similar frequencies, unequal allele distributions are observed for the polymorphic Sample A.

Assessing the sequence of this locus, indels are detected in the flanking sequence of the microsatellite. Figure 3.6 displays the microsatellite repeat and one flanking indel in Sample A. As the indel occurs in sequences that contain different

microsatellite alleles, the repeat is not associated with a certain microsatellite allele. Hence, if the locus is typed by conventional means, the indel affects allele size irregularly which leads to an unspecific shift in allele frequency. In the present case the indel would lead to elongated alleles. Corresponding alleles of longer size were detected in the single typing run compared to 454 sequencing. Furthermore, such indels explain the different patterns of allele frequencies in the polymorphic Sample A. Single typing revealed higher frequencies for allele 30 and 33 than 454 sequencing for the corresponding alleles 29 and 33. This shift of allele frequency towards higher numbers of longer alleles may be associated with indels in the flanking region of the microsatellite.

19:412537	A	A	A	G	G	G	G	-	-	-	-	-	(TG)19	C	A	T	G	A	T	T	G
A 1	∘	∘	.	∘	∘	∘	.	T	T	A	G	G	(TG)19	∘	∘	∘	∘	∘	∘	∘	∘
A 2	∘	.	.	∘	∘	∘	.	T	T	A	G	G	(TG)19	∘	∘	∘	∘	∘	∘	∘	∘
A 3	∘	∘	∘	∘	∘	∘	∘	-	-	-	-	-	(TG)18	∘	∘	∘	∘	∘	∘	∘	∘
A 4	∘	∘	∘	∘	∘	∘	∘	-	-	-	-	-	(TG)16	∘	∘	∘	∘	∘	∘	∘	∘
A 5	∘	∘	.	∘	∘	∘	.	T	T	A	G	G	(TG)16	∘	∘	∘	∘	∘	∘	∘	∘
A 6	∘	∘	.	∘	∘	∘	.	T	T	A	G	G	(TG)15	∘	∘	∘	∘	∘	∘	∘	∘
A 7	∘	∘	∘	∘	∘	∘	∘	-	-	-	-	-	(TG)15	∘	∘	∘	∘	∘	∘	∘	∘
A 8	∘	∘	∘	∘	∘	∘	∘	-	-	-	-	-	(TG)6	∘	∘	∘	∘	∘	∘	∘	∘
A 9	∘	∘	∘	∘	∘	∘	.	T	T	A	G	-	(TG)5	∘	∘	∘	∘	∘	∘	∘	∘
B 1	∘	∘	∘	∘	∘	∘	∘	-	-	-	-	-	(TG)6	∘	∘	∘	∘	∘	∘	∘	∘
B 2	∘	∘	∘	∘	∘	∘	∘	-	-	-	-	-	(TG)6	∘	∘	∘	∘	∘	∘	∘	∘
B 3	∘	∘	∘	∘	∘	∘	∘	-	-	-	-	-	(TG)6	∘	∘	∘	∘	∘	∘	∘	∘
B 4	∘	∘	∘	∘	∘	∘	∘	-	-	-	-	-	(TG)6	∘	∘	∘	∘	∘	∘	∘	∘
B 5	∘	∘	∘	∘	∘	∘	∘	-	-	-	-	-	(TG)6	∘	∘	∘	∘	∘	∘	∘	∘
B 6	∘	∘	∘	∘	∘	∘	∘	-	-	-	-	-	(TG)6	∘	∘	∘	∘	∘	∘	∘	∘
B 7	∘	∘	∘	∘	∘	∘	∘	-	-	-	-	-	(TG)5	∘	∘	∘	∘	∘	∘	∘	∘
B 8	∘	∘	∘	∘	∘	∘	∘	-	-	-	-	-	(TG)5	∘	∘	∘	∘	∘	∘	∘	∘
B 9	∘	∘	∘	∘	∘	∘	∘	-	-	-	-	-	(TG)5	∘	∘	∘	∘	∘	∘	∘	∘

Figure 3.6 Sequence data at microsatellite locus 19:41253735 for nine individuals of Sample A and B. Investigation revealed indels which affect the PCR product length.

At locus 19:4237056 both methods show a consistently stepwise mutation patterns for the microsatellite alleles where the allele size is shifted by one basepair between both methods. Besides allele 38, which is detected by typing in a very low frequency but lacks in the 454 data, all other alleles are present in both analysis; slippage alleles are again disregarded. Similar to the previous locus the most frequent allele 1 of Sample A is observed in similar frequencies and again differences occur mainly in the less abundant alleles of Sample B. Here allele 19 appears with a frequency of 20 % in 454 sequencing data while the corresponding allele 20 reaches

less than 10 % in the single typing run. Investigation of the sequences does not supply a reason for this shift.

Three main alleles are detected at the tetranucleotide repeat locus 19:45494409 with both methods. Whereas the observed alleles are concordant in relative distances to each other, the 454 alleles are constantly two basepairs shorter compared to those obtained by single typing. Note that allele size is based on the actual distance between the observed alleles, *i.e*. this effect cannot be explained by shifts in the flanking sequence and the cause of this observation is unclear.

Again, the most abundant allele is observed in similar frequencies while deviating rates are reported for less frequent variants between the two methods. Some alleles are lacking in either one of the methods, all present at very low frequencies. Allele 10 and 11 in the 454 sequencing data result from a one-basepair indel within the microsatellite sequence (data not shown). Since these alleles do not occur in the single typing data this might be due to erroneous base calling of the 454 sequences or insufficiency in detecting undersized differences with the allele calling software. The observed differences in allele frequency in Sample B might be caused by the lower number of chromosomes typed in the 454 experiment. The 454 sequencing analysis encompasses approximately 30 sequences, hence every sequence constitutes a relatively high fraction to the allele frequency distribution and even small shifts in the number of alleles will cause rather strong changes in overall frequency.

Although some differences in the results are apparent between the two approaches, the general consistency of the results concerning the number of alleles and allele frequencies is good. Although allele size varies in some cases, analogous alleles are identified. With regard to the most frequent allele, results are robust in both cases. Less abundant alleles show higher variation in frequency distribution but still, the results reflect representative outcomes. Note that two different techniques of assessing alleles are compared here. Hence, detection of slightly deviating results is expected. As mentioned previously, manual typing of microsatellites comprises the complete PCR product for determining alleles and shifts in the flanking regions can lead to changes in observed allele size. In contrast, the sequencing method is currently specifically focused to determine the length of the microsatellite region itself.

3.4 Conclusions

The presented results demonstrate that despite some basal problems, next generation sequencing proved to serve as a suitable tool for running high throughput analysis of microsatellites. Results obtained by this method are consistent to those generated with conventional typing methods. Since hundreds of markers are processed simultaneously and sequencing output can be directly applied to an automated analysis program, the method is particularly time efficient for high throughput processing of microsatellites. Further, 454 sequencing achieves a cost benefit at certain number of investigated loci.

However, there is still potential of enhancing efficiency of this method. One major issue is improving the rate of reads which can be taken into the analysis. About 50 % of produced reads is lost during the process of read analysis. Most of the output has to be discarded because the reads cannot be matched with the rather high E-value used in the analysis. As shown a high fraction of sequences can be matched with a lower E-value, indicating that these are short reads. As the analysis requires a certain sequence length to obtain precise matching results it is not recommended to adjust the E-value to increase the output of matched reads. Unfortunately it is still unclear why so many short reads are generated. Hence, solving this problem is a major goal as this harbors a substantial source to improve the amount of analyzable data.

Furthermore, reads are lost in the blast search because of obtaining erroneous sequences which result from non specific PCR amplification as well as contamination. To minimize PCR side products a special blast step which checks the PCR products for artifacts was incorporated into the primer design program. Additionally, a gel purification step was performed extracting only target size products, which should prevent unwanted PCR products. Nevertheless almost 10 % of reads is due to PCR artifacts. Although there might be opportunities to reduce the fraction of PCR artifacts even more, such as refining annealing temperature for optimal primer binding or even more comprehensive blast searches, complete inhibition of PCR side products is very time consuming and probably exceeds the benefits in most cases.

Another rather high proportion of reads was lost in run1 due to truncation of the read, *i.e.* the base caller cannot read the sequence after the repeat. Since correctness of alleles cannot be assured for truncated reads, only alleles that were

flanked by leastwise 8 bp of non-repeat sequence were taken into the analysis. Correlation between the copy number of microsatellite repeats and the sequence length which can be read after the repeat was observed. This finding gives a hint towards truncation of reads due to slippage occurring on the bead at which the DNA molecule binds during high throughput sequencing. Slippage taking place during amplification of DNA on the bead results in unequal length of sequences. In other words the sequences produced on the same bead are partially shifted after the repeat and hence cannot be decoded by the base caller. Whereas tri- and tetranucleotides do not seem to be affected by slippage, dinucleotides above 13 repeats show a rather strong association between the length of the microsatellite stretch and the amount of decipherable sequence. The observation that slippage rates are highest in dinucleotides and relative to the number or repeat units is in line to reports in the literature (Kruglyak et al. 1998).

In this regard it was expected to find a decline in the relative number of dinucleotide microsatellite loci that were taken into the final analysis. In contrast, similar numbers were observed for di – and trinucleotides that passed all filtering steps including the coverage threshold. Surprisingly, on average tetranucleotides tended to performed worse than di- and tetranucleotides, indicating that there are further undetected effects affecting the PCR products. However, adjusting the permitted repeat number to 13 for dinucleotides in the options for the microsatellite search substancially enhanced the outcome of run2. The proportion of reads that failed in the length check was considerably reduced and approximately 20 % more loci could be analyzed.

To date, the main limitation in high-throughput microsatellite genotyping is the required manual editing of allele calls. Still, programs for automated allele calling have limited capability and accurate data can only be assured by manual inspection (Matsumoto et al. 2004). The presented results demonstrate that completely automated execution of allele calling is possible concerning data generated with 454 techniques. In contrast to conventional typing, at which alleles are based on the total size of the PCR product, alleles are associated with assessed numbers of microsatellite repeats. For that purpose a routine was established which basically identifies the microsatellite motif and subsequently counts the repeats. To survey the results obtained by this method several loci were additionally typed in a conventional

manner.

Overall, consistency between the results obtained by the two compared methods was observed. Although frequencies of certain alleles can vary among the methods due to method specific differences, results generated by either method are reliable. Again, it should be pointed out that two procedures are compared here which generate different types of data. Whereas manual allele revision allows discrimination against slippage alleles, this is not feasible regarding the automated manner. As a consequence an increase in the number of artificial low frequency alleles is observed which leads to a higher overall heterozygosity. Sequencing the PCR product allows examination of the microsatellite segment itself, whereas allele typing implicates the whole PCR product into the analysis, including parts that would not be sequenced. Thus, typed alleles can include indels, such as shown for locus 19:41253735. In this case the indel is not associated with certain microsatellite alleles and thus causes shifts in allele frequencies which lead to inconsistencies between the methods.

To sum up both methods generate reliable data, pointing out that differences in the results between the compared analysis procedures are expected and do not invalidate either one of the approaches. Assessing capacities of each method, it should be emphasized that the data achieved by sequencing is much more comprehensive. For example sequence data allow calculation of mutation rates of particular microsatellite loci in addition to number and size of detected alleles. Estimates about the mutation rates can be useful when assessing selection at linked loci as it is one important parameter shaping the pattern of selection (Maynard Smith & Haigh 1974; Wiehe & Stephan 1993). As the mutation rate is highly sensitive to repeat number and SNPs occurring within the repeat sequence, only sequencing the microsatellite locus will provide the required information to evaluate this parameter (Jin et al. 1996; Schlötterer et al. 1998).

In the end the decision which method to take always comes to weighing the costs against the benefits. It has been shown here that although some parameters can still be improved the presented method is a reliable and time effective substitute for conducting high throughput analysis of microsatellites. The quantitative threshold when the methods gets not only time but also cost efficient is highly variable and depends on the individual costs of material and personnel expenses. Two main factors determine the profitability of the new method. First, the amount of loci investigated,

which directly corresponds to the primer costs. If many primers are ordered, the costs for labeling may increase such that already the pure lable costs make 454 sequencing profitable. Second the number of individuals investigated which determines the costs for sample preparation. In the present study individuals were pooled for sample processing in the introduced method and results were validated against individually typed samples. Studies have shown that results obtained by typing pooled DNA require confirmation by subsequent typing on an individual level (Thomas et al. 2007). This may lead to a steep increase of costs for PCR amplification and typing.

Assuming rather low consumable costs as in the present case the sheer primer costs make the sequencing method profitable. This holds also when taking the failures of primers in the 454 run into account. Hence, despite all difficulties, the method turned out to be a valuable alternative which can be both time and cost efficient.

4 First Step Towards a Complete Genome Screen for Selective Sweeps in House Mice Using Microsatellites

4.1 Introduction

More than half a century ago population genetics started as an attempt to understand evolutionary change on a quantitative level (Lewontin 1974). A problem that still remains is quantifying the relative contribution of natural selection in shaping the genetic variation observed among living organisms (Nielsen 2005). Since evolution is taking place within populations, a key to study it is to analyze the change of allele frequencies within populations (Ohta 1992). Although it is consensus that evolutionary changes are mainly caused by natural drift and natural selection (Lewontin 1974; Kimura 1983), the relative importance of these factors is still not clear and has been debated since decades.

The neutral theory of molecular evolution was introduced by Kimura in the late 1960s (Kimura 1968). It proposes that the vast majority of evolutionary changes at the molecular level is caused by random drift of neutral mutants and so have their fate dictated by chance alone. Contrarily, the fraction of selectionists states that a large proportion of the observed variation does affect the fitness of the organisms and is subject to Darwinian selection (Maynard Smith and Haigh 1974b). Although most modern evolutionary biologists agree on the compatibility of both theories, still contradictory observations of genetic variation in natural populations are to be explained. Selected substitutions at one locus can produce stochastic dynamics that are remarkably like those of genetic drift. Gillespie (2000) considers linked selection rather than genetic drift as being the major stochastic force in many natural populations. The author shows that hitchhiking may explain invariant levels of polymorphism that are observed in natural populations (Sella et al. 2009; Gillespie 2000).

Most attempts to estimate the average frequency of positive selection are based on comparative sequence analysis between species (Smith and Eyre-Walker

2002, Birne and Eyre-Walker 2004, Bazykin et al. 2004). As these studies are mainly focused on investigations of coding regions to estimate the frequency of adaptive amino acid substitutions they are able to measure protein evolution of genes which are under ongoing constraint (e.g. Fay and Wu 2001, Fay et al. 2002). However they miss recent selective events and all adaptive events that go along with changes in *cis*-regulatory systems.

An alternative approach to such comparative analysis involves screening genome-wide patterns of DNA polymorphism to detect the locus-specific signature of positive selection (Luikart et al. 2003; Schlötterer 2003). Thereby chromosomal regions that harbor adaptive mutations are identified by exploiting theoretical predictions about the effects of positive selection on patterns of neutral genetic variation at linked sites ('hitchhiking') (Kim and Stephan 2002; Przeworski 2002; Storz 2005). While positive selection drives an adaptive mutation to fixation, the selected site may be repeatedly recombined with new genetic backgrounds which disentangles the hitchhiking effect so that the molecular footprint is lost over time. Important parameters that determine the strength of the effect are selection intensity, rates of recombination and mutation as well as population size (Maynard Smith and Haigh 1974; Kaplan et al. 1989; Wiehe and Stephan 1993).

In the last years numerous studies used the identification of selective sweep signatures to trace adaptation in natural populations. Since demographic events, such as bottleneck events or population expansion, can resemble patterns of genetic hitchhiking an interpretative challenge is associated with polymorphism-based neutrality tests (Tajima 1989; Fu and Li 1993; Barton 2000; Fay and Wu 2000; Kim and Stephan 2003; Andolfatto 2001). For example Haddrill et al. (2005) demonstrated that a reasonable number of signatures of positive selection observed in derived populations of *Drosophila* are caused by a bottleneck connected to the 'out-of-Africa' expansion (Kauer 2003).

While demographic processes will have relatively uniform effects across the entire genome, the effects of selection are generally expected to be locus-specific and can be inferred from patterns of variation at linked sites (Cavalli-Sforza 1966; Lewontin & Krakauer 1973). In order to disentangle the effects of demography and selection it is therefore necessary to screen patterns of DNA variability at multiple, unlinked loci (Storz 2005). However, a massive bottleneck followed by an extreme

population expansion makes it difficult to distinguish footprints of positive selection from numerous types of artifacts (Eyre-Walker 2002), often caused by demographic effects within and between populations. Hence, one crucial disadvantage in many studies screening for positive selection events in natural populations is the lack of demographic data.

My approach is mainly focused on the comparison of two distinct populations of the European house mouse *Mus musculus domesticus*. In the case of spatially separated populations that inhabit different environments, it is possible to identify chromosomal regions involved in adaptive divergence by comparing relative levels of variation among multiple, unlinked loci (e.g. Charlesworth et al. 1997; Storz 2005). Levels of within-population diversity will be reduced by hitchhiking in the population which harbors the locally adaptive allele, whereas increased variation will be observed in the comparative population.

The comparative analysis of several populations for detecting selective sweeps requires selection of suitable polymorphic markers (Schlötterer 2002). Furthermore the establishment of a high throughput routine is needed, since the empirical detection of polymorphic variants requires tests of multiple individuals at various markers.
Several features highlight microsatellites as appropriate tools for the application in such studies. For example microsatellites are present at high numbers in most species which enables their application in a broad spectrum of organisms (Tautz and Renz 1984), they generally comprise high polymorphism and are relatively easy to investigate (Tautz 1989; Schlötterer et al. 1997, Huttley et al. 1999, Kohn et al. 2000, Harr et al. 2002, Kauer et al. 2003, Kayser et al. 2003, Payseur et al. 2002, Schlötterer 2002, Storz et al. 2004). Further, in contrast to SNPs, their mutation profile is multi-allelic, which enhances their information content. Finally the high mutation rate makes them particularly well-suited for the characterization of very recent selective events (Schlötterer 2003). The availability of the complete genome sequence of the house mouse provides thousands of microsatellites throughout the genome and thus allows working with large number of marker. As the chromosomal locations of microsatellites are known, flanking regions can be directly investigated to detect nearby genes.

In a previous study Teschke et al. (2008) performed a genome screen for signatures of selective sweeps based on microsatellites in natural populations of the

house mouse. The screen included 1,000 microsatellite loci that were typed throughout the entire genome. Candidate loci for positive selection were identified by pairwise comparisons of microsatellite polymorphism between populations. Subsequently every locus present in a 100 kb region around the candidate loci was typed. In this manner the size and shape of the selective sweep were determined. The authors showed that the average window size of a sweep is about 50 kb in the respective populations.

Based on these results, my aim was to conduct another genome screen of microsatellites distributed at about 50 kb intervals from each other. Thereby, each selective sweep which is detectable with the chosen marker system in a continuous stretch of DNA will be identified. The result of this study is expected to provide profound information about basic parameters of selective events in natural populations. Further the data will allow addressing the fundamental questions about the course of adaptation. How frequent are positive selection events? To what extend do selected sites shape neutral genetic variation? Are the events driven by strong or weak selection? Do most advantageous mutations involve single genes of large phenotypic effect (‘major’ genes) or induce small beneficial effects? Do adaptations generally involve new mutations or standing genetic variation?

For reasons of comparability the same populations of house mouse were investigated as described by Teschke et al. (2008). In total six different populations composed of 40 individuals each were investigated; four populations of the subspecies *M. m. domesticus* [one from the Cologne-Bonn area (Germany), one from the Massif Central (France), one from Iran and one from Cameroon]; two populations of the subspecies *M. m. musculus* were sampled (one from the Czech Republic and the other one sampled in Kazakhstan) (see Chapter 1.2). For detailed information about the sampling setup see Ihle et al. (2006).

The present study is mainly focused on the analysis of the two Western European populations Germany and France along with their ancestral population, Iran. Genes influenced by positive selection are identified by the characteristic footprint of reduced variability at linked neutral loci due to genetic hitchhiking (Maynard Smith and Haigh 1974, Ohta and Kimura 1975, Slatkin 1995). Conspicuous outliers of population-specific reduction of variability are detected by applying Schlötterer’s ratio statistic (ln*RH*) which is based on pairwise comparisons of the

relative heterozygosity between populations (Schlötterer and Dieringer 2005). Accordingly, loci showing reduced polymorphism in one population but not in others are marked as candidates for selective sweeps. As a pilot study almost 1,000 microsatellite loci were investigated throughout the chromosome 19 to determine variability levels at these loci in the six natural populations. Microsatellites were identified based on the mouse genome (Pruitt et al. 2006) applying a new software tool, '*msfinder.pl*' which filters a given sequence for simple sequence repeats under user defined conditions (see Chapter 3.2). To process this large number of loci within a reasonable time, a new analysis approach was established using high throughput sequencing (see Chapter 3).

4.2 Material and Methods

To carry out the microsatellite screen of chromosome 19 next generation sequencing technique was used which enables high throughput analysis of microsatellites. Population-specific pools of samples from six different origins were analyzed. For detailed information about sampling manner see Ihle et al. (2006). DNA was normalized to 2 ng/µl and equal volumes of 40 individuals were pooled.

Primers were designed based on the Mouse Genome Build 37 using the program '*msfinder.pl*' (see Chapter 3.2) and ordered from Metabion (primer sequences are provided in the digital supplement). Amplification was performed using Quiagen kit (Cat.No. 206143) and following the protocol instructions. All reactions were prepared in 10 µl volumes using 20 ng of pooled DNA template.

PCR products were subsequently pooled per population and DNA was precipitated. After running the amplicons on a 1.2 % agarose gel bands of expected PCR product size were cut out and the DNA was purified from the gel using Quiagen MinElute PCR Purification Kit (cat. no. 28004). Purified PCR products were run on a 454 sequencer using *GS FLX Titanium* series reagents. For modifications of the manufacturer's protocol see Chapter 3.2. Base calling algorithm was included in the 454 software suite by Roche.

Additionally a program was set up which filters the 454 output under several

conditions so that the generated results can be directly used for microsatellite analysis. For detailed information about sequence analysis see Chapter 3.2. Analysis of microsatellites comprises locus- and population-specific assignment of alleles followed by calculation of allelefrequency and expected heterozygosity. The generated results can be directly used for ln*RH* statistics to detect population-specific reductions of variability (Kauer et al. 2003).

Using 454 sequences for microsatellite analysis reveals generally higher results of expected heterozygosities than compared to conventional typing methods (see Chapter 3.3). To allow for comparisons of results obtained in this study with data generated in a previous genome screen accomplished by Teschke et al. (2008), candidate loci were subsequently genotyped.

Estimation of significance: Gene diversity estimates (expected heterozygosity) were calculated for all loci and corrected for sample size by n/(n-1) where n is the number of analyzed chromosomes. As it is not possible to assign the reads obtained by 454 sequencing to certain individuals each sequence is expected to reflect an individual chromosome if less than 80 sequences per locus are present. In the presence of more than 80 reads the sample size was corrected to the maximum number of actual chromosomes. To assure an effectual number of sampled chromosomes a coverage cutoff of 20 reads was set.

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

Expected heterozygosity, corrected for sample size:

***H**(n/n-1)**

H = heterozygosity

Based on the estimator:

$H = 1 - (1 / (1+2q)^{1/2})$

(Ohta & Kimura 1973)

n= number of analyzed chromosomes

As mentioned above, population-specific reductions of variability are detected by applying the ln*RH* test. The statistic is based on comparisons of the relative gene diversity of single loci between two populations (Schlötterer and Dieringer 2005), whereas the expected heterozygosity is taken as variability measurement. In total six sample populations were investigated. The subspecies *M. m. domesticus* is

represented by respectively one population from Germany, France, Iran and Cameroon. The populations from Czech Republic and Kazakhstan belong to the subspecies *M. m. musculus*. This results in seven possible pairwise comparisons between populations of the same subspecies: Germany-France, Germany-Iran, Germany-Cameroon, France-Iran, France-Cameroon, Iran-Cameroon and Czech Republic-Kazakhstan.

The significance of ln*RH* values is estimated by performing a z-transfomation (z = (x – mean) / standard deviation), a standard normal distribution is approximated and p-values of the investigated candidate loci are inferred from this distribution. Besides the results obtained for *M. m. musculus* populations the observed ln*RH*-values do not significantly deviate from a normal distribution (Kolmogorov-Smirnov test: Germany-France Z=1.29, N=769, p= 0.071; Germany-Iran Z=0.99, N=780, p= 0.284; Germany-Cameroon Z=0.84, N=753, p= 0.481; France-Iran Z=1.04, N=755, p= 0.235; France-Cameroon Z=1.11, N=729, p= 0.17, Iran-Cameroon Z=1.16, N=781, p= 0.134 and Czech Republic-Kazakhstan Z=1.6, N=609, p= 0.012). Detailed information about neutrality tests are provided in Supplement 2. For genotyped samples an independent set of 64 'neutrally evolving' microsatellites (collected by Ihle et al. 2006) was used as a reference to normalize the data. Estimations of mean and standard deviation are based on these 64 loci [(mean 0.0875, standard deviation 0.8584) see Teschke et al. (2008)].

Considerable outliers were detected by selecting loci with p-values ≤ 0.05 (z= <-1.96 and >1.96). For the Germany-France comparison, all candidate loci were subsequently genotyped and tested against the above mentioned reference data. Because false positives accumulate due to multiple testing the data were corrected by a stringent Bonferroni-adjustment. Significant p-values of the focal German and French populations result in $p < 6*10^{-5}$ after correction.

4.3 Results

Chromosomal regions that harbor patterns of positive selection are identified by comparing the genetic variability of neutral markers between different populations. The number of markers that were taken into the applied ln*RH* statistics and the number of resultant candidate loci for all possible pairwise comparisons between the six sample populations are displayed in Table 4-1. Since only markers which showed a minimum coverage of 20 reads in each population were analyzed (see Chapter 3.2), the numbers of investigated loci compared between different populations differ somewhat. Further information of all resultant candidate loci can be found in the table of Supplement 5, which provides a list of ln*RH* values for all loci at which $p<0.05$ and the Ensembl Gene ID of the closest gene to the respective marker. All markers are consistently named after their location ± 300bp on chromosome 19 (Mouse Genome Build 37).

Calculating ln*RH* values based on 454 sequences yields the problem that PCR artifacts cannot be removed in the same way as it is possible by manual inspection of individually typed samples. Thus artificial products are considered as additional alleles in the analysis which increases the observed heterozygosity, *i.e.* absolute ln*RH* values will be generally smaller in this type of analysis (see Chapter 3.3). Especially the allelic patterns of loci which harbor low variability will be blurred by artificial alleles. Since the statistic identifies loci at which pairwise heterozygosities vary extremely, loci which come out as significant despite the occurrence of artificial alleles show generally an even stronger signal after single typing. For example p-values for significant loci of the Germany France comparison, which will be discussed later, ranged between $p = 0.045- 0.0002$ based on 454 sequences, whereas single typing revealed $p < 6*10^{-5}$ for all of them. Due to the high discrepancy of p-values resulting from 454 sequences in comparison to single typing, correction for multiple testing was only performed for the candidate loci of the Germany France comparison, for which ln*RH* values of individually typed loci were present.

Since subsequent typing was only performed for the Germany France comparison, loci were regarded as potential sweep candidates if $p < 0.05$ for the other population comparisons. Notably selecting the 5 % tails of the distribution reflects almost exactly 5 % of the total amount of investigated loci (Table 4-1). However,

selection of the upper and lower tails of the distribution will result in selecting extreme outliers. Furthermore, as a general reduction of p-values is observed using the 454 sequencing approach, the rate of false positives is expected to be rather low. In the following, outlier loci selected on a 5 % level will be termed 'candidate loci', whereas loci which remained significant after the correction for multiple testing will be referred to as 'significant loci'.

Table 4-1 Number of analyzed markers and resultant candidate loci for all pairwise population comparisons.

Pairwise comparison	Ger-Fra	Ger-Iran	Ger-Cam	Fra-Iran	Fra-Cam	Iran-Cam	CR-Kazak
no of analysed markers	756	766	739	742	716	765	598
no of candidate loci	39	47	38	41	42	36	39
candidate loci per population	23/16	27/20	16/22	23/18	20/22	10/26	22/17

Distribution of ln*RH* values along the chromosome 19 for all appropriate markers and pairwise comparisons are provided in the digital supplementary. Furthermore the supplementary material contains figures of allele frequencies for all candidate loci as well as a number of additional loci. One has to keep in mind that it has been shown that ln*RH* values for the *M. m. domesticus* comparison do not follow a normal distribution (see Chapter 4.2 Supplement 2). However, the data is incorporated into the analysis, noticing that the results have to be interpreted carefully.

The following analysis will be mainly focused on the Germany-France comparison. Furthermore the comparison of the Western European populations (Germany and France) versus the ancestral Iranian one as well as the comparison of candidate loci between the subspecies *M. m. musculus* and *M. m. domesticus* will be presented. Since the actual microsatellite repeats might deviate from the denoted database in length or quality, *i.e.* the pattern could be disrupted by indels or point mutations microsatellite sequences were inspected to confirm identified sweep patterns. The observed microsatellite repeat motifs for the respective sweep alleles are shown in Table 4-5. Related alignments of relevant sequences are provided in the digital supplementary material (Chapter 3-4: Sweep_Alignments).

4.3.1 The Germany France comparison

After a stringent Bonferroni-adjustment 6 of the 756 loci that were investigated in the two focal *M. m. domesticus* populations remained significant.

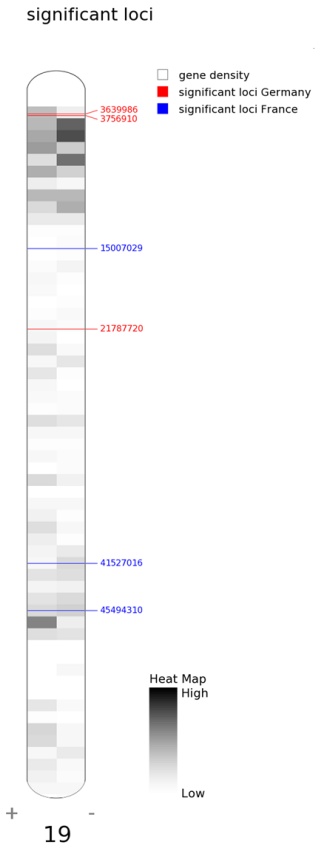


Figure 4.1 Significant loci along chromosome 19. Blue markers indicate sweep patterns detected in the French, red markers in the German population. Grey shading displays gene density referring to the heat map.

These loci are equally distributed between both populations (Figure 4.1). Two of the German sweep loci are located in close proximity to each other, *i.e.* they may be physically linked. The other significant loci are distributed across the chromosome and are found in regions of relatively high as well as low gene density. Recombination rates in these focal chromosomal regions vary between 0.75 to 2.52 cM/Mb (Table 4-2).

The six loci differ in repeat type, and repeat length ranges from 7-9 repeat units for the most frequent alleles (Table 4-2). Figure 4.2 displays the allele frequencies for both populations at the respective loci.

Table 4-2 Significant loci according to the ln*RH* test statistics after Bonferroni-correction, expected heterozygosities, physical position, recombination rate (taken from Jensen-Seaman et al. 2004) and number of repeat units of the sweep allele are displayed. Denoted p-values are taken from ln*RH* values which resulted from single typing. *RUN=Repeat unit number

Marker name	Physical Position [kb]	Recombination rate [cM/Mb] (5 Mb window)	Exp.Het. Ger.	Exp. Het. Fra.	Ln*RH* z-value (p<0.00006)	Sweep in	Repeat type	RUN* of sweep allele
3639986	3,640	1.56	0.046	0.647	-5.068	Ger	dinucleotide	8
3756910	3,757	1.56	0.026	0.621	-5.568	Ger	tetranucleotide	9
15007029	15,007	0.75	0.481	0.022	4.675	Fra	dinucleotide	8
21787720	21,788	0.87	0.023	0.717	-6.481	Ger	trinucleotide	7
41527016	41,527	1.08	0.681	0.104	4.076	Fra	dinucleotide	8
45494310	54,494	2.52	0.661	0.087	4.152	Fra	tetranucleotide	8

In two cases (3756910 and 21787720) the shortest alleles are fixed. In each of the three loci 3639986, 15007029 and 41527016 the third longest allele is the most abundant one, however the shorter ones appear in very low frequencies. Only at locus 45494310 a shorter allele than the major one is detected in a aomparatively high frequency. At all sweep loci more than 90 % of the chromosomes carry the major allele and in all cases the fixed allele is also present in the non-sweep population.

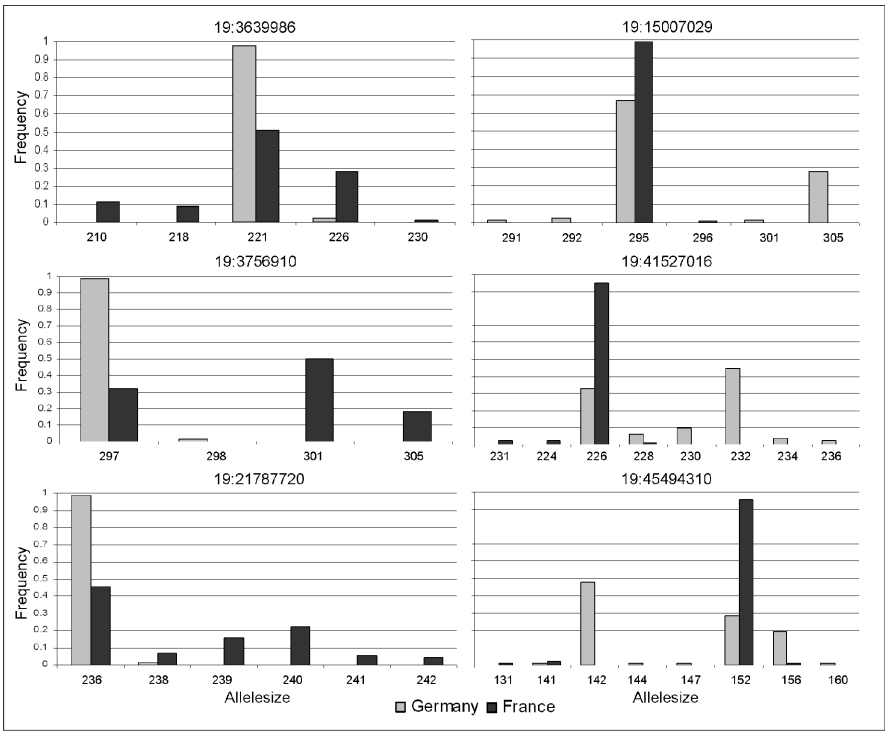


Figure 4.2 Allele frequencies for significant loci between the German and French populations. Left: sweep in the German population; right: sweep in the French population.

Concerning the occupied allele classes, deviations from a stringent stepwise

mutation pattern can be observed. This can be either due to inaccuracy of the used length standard, *e.g*. at the tetrarepeat locus 3639986 where a mutation step of five basepairs is observed between allele 221 and 226, or may be due to actual non-stepwise mutation events. Examination of the microsatellite sequence for these alleles in the Germans samples reveals proper stepwise mutation (see digital supplementary Chapter3-4: Sweep_Alignments). However, sequences are detected among the German and French samples which changed partially or fully to trinucleotides [$(AAC)_N$ instead of $(AAAC)_N$], *i.e*. do in fact not mutate stepwise or contain a completely distinct repeat pattern. Certainly this affects the detected length of microsatellite alleles in a non-stepwise manner.

Notably at locus 15007029 and 45494310 the German population contains two major alleles which are several mutation steps apart from each other, while no alleles are observed in the allele classes in between them. Under the assumption that the mutation profile of the non sweep microsatellite will be discrete such a pattern would not be expected.

Two further loci (3702999 and 3731289) were investigated in between the two significant outliers of the German population 3639986 and 3756910 (see digital supplementary Chapter3-4: Allelefrequencies_Germany-France.pdf, pp. 12, 13.). Consistently, both of them show a reduction of variability in the German population. At locus 3702999 the high frequency allele is observed with 80 % frequency and ln*RH* statistic results in $p < 0.05$ (see Supplement 5). Again, one of the shortest alleles is fixed and repeat unit numbers are rather small. Contrarily, repeat unit numbers at the neighboring locus 3731289 range from 12-27 and one of the intermediate alleles is at high frequency in the German population. Although ln*RH* values are not considerably reduced at this locus, such an allele pattern indicates a recent sweep signal which recovered variability by subsequent mutations. The presence of sweep evidence in additional, adjacently investigated markers confirms the confidence in the sweep signal at the respective chromosomal region. Furthermore, detection of four contiguous loci all displaying a sweep pattern in the German population indicates that all loci belong to one single sweep spanning a minimum range of 120 kb.

Similarly next to the marker at locus 15007029, at which a sweep signal is detected in the French population, an additional contiguous outlier (15056726) consistently displays a sweep pattern in the French population. Again these loci are

more likely to belong to one single sweep than two independent ones, which confirms the sweep pattern, as well as suggests a minimum sweep size of 50 kb.

With respect to the microsatellite sequences two loci are identified which contain a discontinuous motif. At locus 3756910 the $(GAAT)_N$ motif changed to $(GAAT)_1(GAAA)_3(GAAT)_N$ in both populations, *i.e.* the slippage mutation rate of this microsatellite is expected to be strongly reduced. Investigation of sequences at locus 21787720 revealed that the detected variability in the French population does not completely result from the microsatellite itself but from observed indels in the flanking region (see digital supplementary Chapter3-4: Sweep_Alignments).

4.3.2 The Iran France Germany comparison

As described previously, the subspecies *M. m. domesticus* has its origin in the north of the Indian subcontinent and then followed the colonization route via the Middle East to Europe. Hence, the Iranian sample is expected to represent the ancestral population of this subspecies. In the following analysis the European populations Germany and France are compared to the Iranian one. To detect loci that show a sweep pattern in multiple populations all ln*RH* values were compared within the three populations. Candidate loci were selected if two populations show a consistent signal in reduced variability ($p < 0.05$). An overview of the physical position, the repeat types and the numbers of repeat units of the sweep alleles for these loci is provided in Table 3-4. Again, the candidate loci are distributed along the whole chromosome. Most of the sweep alleles range between 5-11 microsatellite repeats. Only at locus 6550201 a very short allele is fixed.

Table 4-3 Candidate loci according to the ln*RH* test statistics, physical position, and number of repeat units of the sweep allele. *RUN=Repeat unit number (rounded)

Marker name	Physical Position [kb]	Sweep in	Repeat type	RUN* of sweep allele
18892241	18,892	Europe	tetranucleotide	11
23976316	23,976	Europe	dinucleotide	9
40868525	40,869	Europe	trinucleotide	7
6550201	6,550	Europe	tetranucleotide	4
41527152	41,527	Iran-Fra	dinucleotide	8
4237056	4,237	Iran-Fra	tetranucleotide	7
53078741	53,079	Iran-Fra	trinucleotide	7
54108514	54,109	Iran-Fra	trinucleotide	8
10965776	10,966	Iran-Ger	dinucleotide	9
3640065	3,640	Iran-Ger	dinucleotide	5 / 8
3756910	3,757	Iran-Ger	tetranucleotide	9
39007010	39,007	Iran-Ger	dinucleotide	9
47097417	47,097	Iran-Ger	dinucleotide	8
47254575	47,254	Iran-Ger	tetranucleotide	8

Germany and France compared to Iran

Figure 4.3 illustrates allele frequencies of loci which show sweep patterns according to ln*RH* in the German and French population versus the Iranian one. As such patterns might be due to selection having affected specifically the populations of the Western European continent they will be named ‘European-sweeps’.

Alleles of loci analyzed through 454 sequencing are specified based on their repeat unit length (see Chapter 3.2). As already described, the stepwise mutation profile of microsatellites may be interrupted by indels or point mutations, which is frequently observed.

The Iranian population, here representing the non-sweep population, shows high polymorphism at all four loci. In all cases the European populations share the same high frequency allele. All sweep alleles are present also in the Iranian population, nevertheless in very low frequencies. Beside the sweep alleles, the three populations share further ones among them, also variants which are several mutation steps apart. At three loci (18892241, 23976316 and 6550201) one of the shortest

alleles is fixed. Only at locus 40868525, a variety of shorter alleles are present beside the sweep allele.

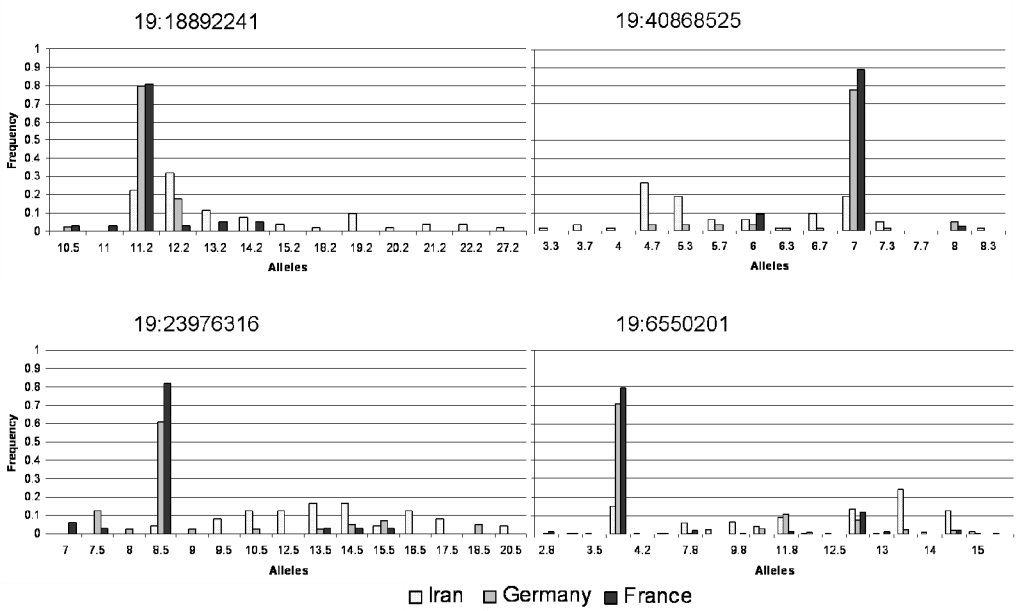


Figure 4.3 Sweep loci in the German and French population but not in the Iranian. Alleles are named after observed numbers of repeat units.

While all French sweep alleles are almost completely fixed (80-90 % frequency), the frequency of German sweep alleles are consistently found at lower levels (60 -80 %). At locus 6550094 the fixed allele contains only four repeats, *i.e*. the mutation rate is probably too low for new mutations to arise. Hence, this locus should not be considered as a good sweep candidate. The other sweep loci contain longer alleles that vary from 7-11 motif repeats. Examination of the sequence at locus 18892241 revealed a point mutation within the microsatellite in all three populations which changed the repeat motif to $(AGAT)_N AGGT(AGAT)_2$. In fact this will reduce the mutation rate but since the sweep allele still harbors eight perfect repeats it should still have the potential to mutate rather frequently.

France and Iran versus Germany

In this section loci are presented which show a sweep signal in the French and Iranian population but not in the German one. As this includes again a comparison between Germany and France, one locus (41527152) appears in this analysis which was already presented above (4.3.1).

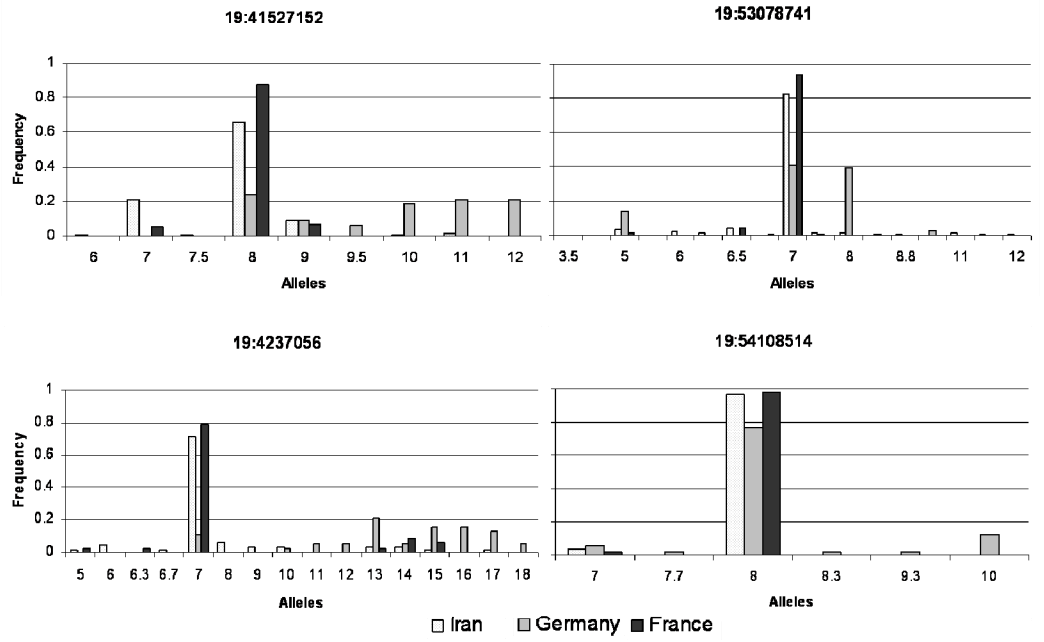


Figure 4.4 Sweep loci in the Iranian and French population but not in the German one. Alleles are named after observed numbers of repeat units.

Similar to the previous results, consistently identical high frequency alleles are observed among both sweep populations, France and Iran, and again the allele is also present within the non-sweep population, Germany. At locus 54108514 an exceptionally high frequency was observed for the sweep allele in the German population. Investigation of the sequence at this locus revealed an interrupted repeat pattern. Further, locus 53078687 revealed a disrupted microsatellite motif. Both repeats are modified such that the mutation rates are expected to be extremely low (see Table 4-5).

Germany and Iran versus France

In the following all loci are presented that show reduction of variability in the German and Iranian population but not in the French one. Again this takes into

account the German France comparison and two loci (3639986 and 3765910) are shown here which were already presented in the previous section (Chapter 4.3.1).

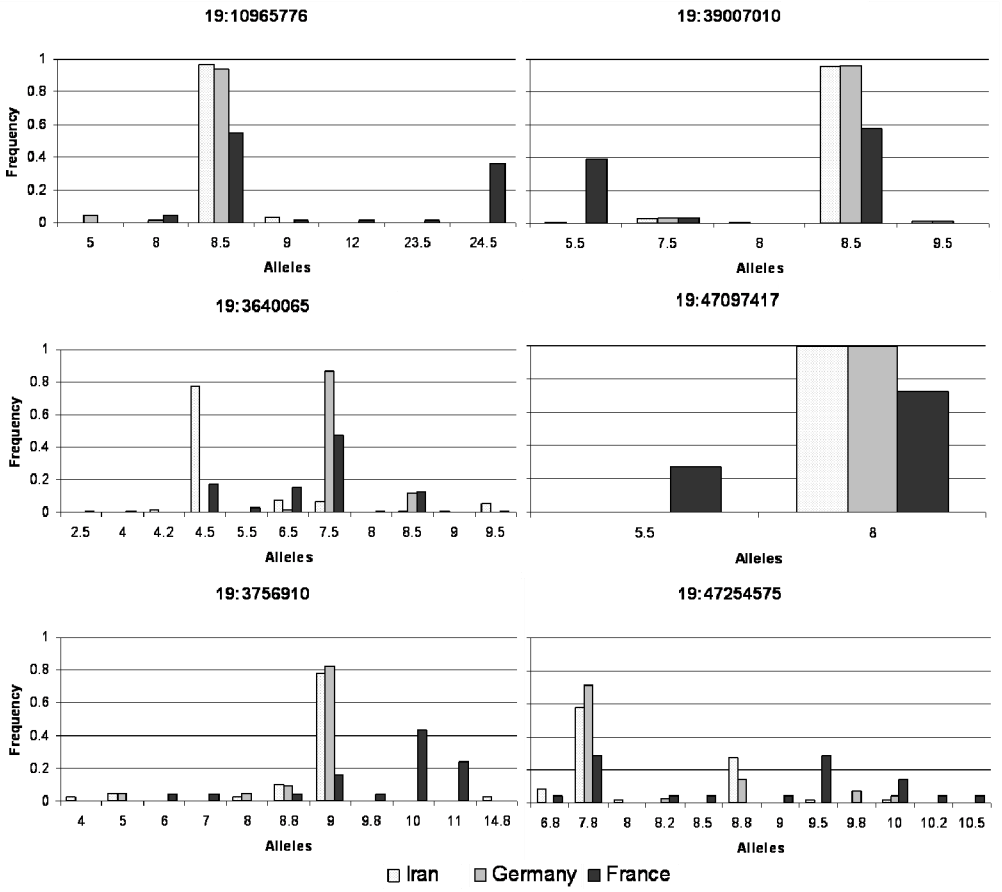


Figure 4.5 Sweep loci in the Iranian and German population but not in the French one. Alleles are named after observed numbers of repeat units.

In all but one case the Iranian and the German population harbor the same high frequency alleles. At this locus 3639986 two different alleles are fixed among the sweep populations whereas the Iranian sweep allele is not present in the German population. The fixation of population specific sweep alleles can be taken as an indication for parallel evolution. Similar to the other results all sweep alleles are also present in the non sweep population, France.

With respect to the repeat motifs beside locus 3639986 all other loci should not be regarded as adequate sweep candidates since they consistently comprise a defective repeat pattern, *i.e.* the reduction in variability is rather due to low mutation rates than genetic hitchhiking (Table 4-5).

4.3.3 Comparison between *M. m. domesticus* and *M. m. musculus*

Based on ln*RH* values populations of the subspecies *M. m. domesticus* (‘*domesticus*’) and *M. m. musculus* (‘*musculus*’) were screened for candidate loci indicating selective sweeps. Again, candidates were chosen on a 5 % level. Results are presented in Table 4-4. The first part displays loci which show overlapping sweep signals in at least one population among both subspecies. The second part comprises loci indicating sweep signals in all populations of either one of the subspecies. Due to the Cameroon population history, at which bottleneck effects cannot be excluded, the sample was excluded from the ‘*musculus* – *domesticus*-comparison’. Similar to the previously described cases candidate loci are dispersed along the entire chromosome (Table 4-4).

Table 4-4 Candidate loci according to the ln*RH* test statistics between *M. m. musculus* and *M. m. demesticus* populations. Physical position and number of repeat units of the sweep allele are presented. If more than two sweep populations are observed ‘//’ seperates the groups according to their sweep allele. *RUN=Repeat unit number (rounded)

Marker name	Physical Position [kb]	Sweep in	Repeat type	RUN* of sweep allele
10893668	10,894	Kaz / Fra	dinucleotide	14
17764383	17,764	CR // Iran / Ger / Fra / Cam	tetranucleotide	9 / 10
20092550	20,093	CR // Iran / Ger / Fra / Cam	tetranucleotide	15 / 13
22597753	22,598	Kaz Fra // Cam	tetranucleotide	9 / 7
25634125	25,634	CR // Fra	tetranucleotide	6 / 7
29340520	29,341	CR // Iran / Ger / Fra / Cam	trinucleotide	7 / 6
27998962	27,999	*musculus*	trinucleotide	9
29148250	29,148	*musculus*	tetranucleotide	7
7508132	7,508	*musculus*	trinucleotide	4
25540928	23,541	*domesticus*	dinucleotide	9
34111578	34,116	*domesticus*	trinucleotide	7
37705557	37,701	*domesticus*	dinucleotide	8

Relating to the first part of the table, almost all cases contain two different sweep alleles and fixed variants appear ‘subspecies-specific’, *i.e.* the same allele is not observed at high frequency among two populations belonging to different subspecies. Examination of the allele sequences at locus 10893668 revealed that the French and Kazakh sweep allele in fact have the same length but differ in sequence.

Hence, also in this case two different alleles are at high frequency. The numbers of repeat units vary between 6-15 repeats. Notably four of the six candidate loci are tetranucleotide repeats. Illustrated allele frequencies are provided in the digital supplement (Chapter3-4: Allelefrequencies).

In three of the six cases perfectly repeated microsatellites are observed (10893566, 25634104 and 29340438). At two of them (10893566 and 25634104) the French and either one of the '*musculus*' populations represent the sweep populations. At locus 29340438 all populations except Kazakhstan show a reduction of variability. At locus 17764383 disruption of the repeat pattern was observed but since a fraction of at least six units remained perfectly repeated the locus still harbors the potential of generating new alleles. The microsatellite motif of the remaining two loci 2009255 and 22597600 are such disrupted that mutation rates are expected to be substantially low. Hence the observed allele pattern is rather explained by sequence characteristics than evolutionary drive.

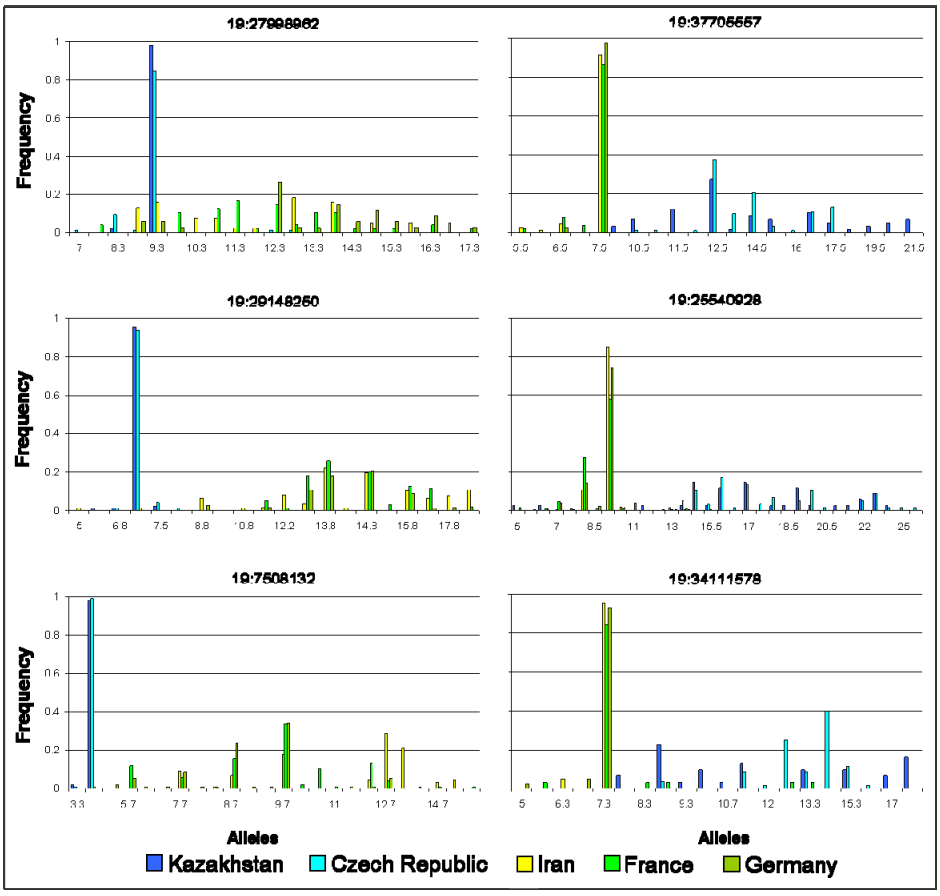


Figure 4.6 Sweep loci in the *M. m. domesticus* and *M. m. musculus* subspecies. Alleles are named after observed numbers of repeat units. Left: sweep in the *M. m. musculus* subspecies; right: sweep in the *M. m. domesticus* subspecies.

Among subspecies specific sweeps at which all populations of one subspecies have swept but none of the others, consistently the same high frequency allele is detected throughout the sweep populations (Figure 4.6). In all cases one of the shortest alleles is fixed. Only at locus 25540928 a shorter allele than the sweep variant appears rather frequently in the French population. While the high frequency alleles are almost completely fixed within the sweep subspecies, a broad allele spectrum is consistently observed among all populations of the non sweep subspecies.

At locus 7508132 a very short trinucleotide allele is fixed among the '*musculus*' populations. Thus, in this case the sweep pattern is likely due to the low mutation rate for the respective allele. At the other five loci allele sizes range from 7-9 repeats. Investigation of the microsatellite sequences revealed mismatches within the repeat motifs in the remaining two '*musculus*' sweep loci. At locus 29148250 an indel was identified and at locus 27998962 the sequence is interrupted twice by two point mutations (Table 4-5), *i.e*. all three '*musculus*' sweep patterns apparently result from low mutation rates. In contrast no irregularities were observed among the '*domesticus*' microsatellite sequences.

Sweep alleles of the '*musculus*' subspecies are present among populations of the '*domesticus*' subspecies even though in very low frequencies. In contrast no sharing of the sweep alleles among the subspecies was observed within the inverted cases.

Table 4-5 Microsatellite Sequences of sweep alleles of presented loci. CR = Czech Republic.

Marker	Population	Sweep Allele Sequence
3639986	Germany	$(AAAC)_8$
	France	$(AAAC)_N$
	Iran	$(AAAC)_5$
3756910	Germany	$(GAAT)_1(GAAA)_3(GAAT)_5$
	France	$(GAAT)_1(GAAA)_3(GAAT)_{5-7}$
	Iran	$(GAAT)_1(GAAA)_3(GAAT)_5$
15007029	Germany	$(GA)_7$
	France	$(GA)_N$
21787720	Germany	$(TGG)_5GG(TGG)_1$
	France	$(TGG)_5GG(TGG)_1$ INDEL
41527016	Germany	$(AT)_N$
	France	$(AT)_8$
	Iran	$(AT)_8$
45494310	Germany	$(AAGT)_N$
	France	$(AAGT)_7$
18892241	Germany	$(AGAT)_8AGGT(AGAT)_2$
	France	$(AGAT)_8AGGT(AGAT)_2$
	Iran	$(AGAT)_NAGGT(AGAT)_2$
23976126	Germany	$(AC)_9$
	France	$(AC)_9$
	Iran	$(AC)_N$
40868525	Germany	$(AAG)_7$
	France	$(AAG)_7$
	Iran	$(AAG)_N$
6550094	Germany	$(TGGA)_4$
	France	$(TGGA)_4$
	Iran	$(TGGA)_4TGAA(TGGA)_N$; $(TGGA)_N$
4237056	Germany	$(AAT)_N$
	France	$(AAT)_7$
	Iran	$(AAT)_7$
53078687	Germany	$(GTGC)_2GTGT(GTGC)_N$
	France	$(GTGC)_2GTGT(GTGC)_4$
	Iran	$(GTGC)_2GTGT(GTGC)_4$
54108514	Germany	$(GGA)_5(GA)(GGA)_N$; $(GGA)_N$
	France	$(GGA)_5(GA)(GGA)_2$
	Iran	$(GGA)_5(GA)(GGA)_2$
10965776		highly repetitive region
39006970	Germany	$(GT)_5GA(GT)_2$
	France	$(GT)_5GA(GT)_2$ and $(GT)_5GAA(TG)_2$
	Iran	$(GT)_5GA(GT)_2$
47097417	Germany	$(CT)_5C(CT)_2$
	France	$(CT)_5C(CT)_N$
	Iran	$(CT)_5C(CT)_2$
47254537	Germany	$(AAAC)_4(AAAT)(AAAC)_2AAAA$
	France	$(AAAC)_N(AAAT)(AAAC)_2AAAA$
	Iran	$(AAAC)_4(AAAT)(AAAC)_2AAAA$
7508027	musculus	$(TTG)_3TTT$
	domesticus	$(TTG)_N$

29148215	musculus	$(GGAA)_4A(GGAA)_3$
	domesticus	$(GGAA)_N$
27998962	musculus	$(TAA)_1TAG(TAA)_4CAA(TAA)_2$
	domesticus	$(TAA)_N$
34111578	musculus	$(AAC)_N$
	domesticus	$(AAC)_7$
37705557	musculus	$(AC)_N$
	domesticus	$(AC)_8$
25540775	musculus	$(AC)_N$
	domesticus	$(AC)_9$
10893566	Fance	$(AC)_{14}$
	Germany	$(AC)_N$
	Iran	$(AC)_N$
	Cameroon	$(AC)_N$
	CR	$(AC)_N$
	Kazakhstan	$(AC)_{10}TC(AC)_3$
17764383	Fance	$(GTCT)_2GTTT(GTCT)_7$
	Germany	$(GTCT)_2GTTT(GTCT)_7$
	Iran	$(GTCT)_2GTTT(GTCT)_7$
	Cameroon	$(GTCT)_2GTTT(GTCT)_7$
	CR	$(GTCT)_2GTTT(GTCT)_6$
	Kazakhstan	$(GTCT)_2GTTT(GTCT)_N$
20092550	hochrepetetive sequenz	besser nicht analysieren
22597600	Fance	$(TTGT)_9$
	Germany	$(TTGT)_N$
	Iran	$(TTGT)_N$
	Cameroon	$(TTGT)_9$
	CR	$(TTGT)_NTT(TTGT)_1$
	Kazakhstan	$(TTGT)_6TT(TTGT)_1$
25634104	Fance	$(AGAC)_7$
	Germany	$(AGAC)_N$
	Iran	$(AGAC)_N$
	Cameroon	$(AGAC)_N$
	CR	$(AGAC)_6$
	Kazakhstan	$(AGAC)_N$
29340438	France	(AGG)6
	Germany	(AGG)6
	Iran	(AGG)6
	Ka	(AGG)6
	CR	(AGG)7
	Ka	(AGG)N

4.3.4 Candidate loci distributed over the chromosome 19

After revision of allele frequency distribution and microsatellite sequences of the 29 presented candidate loci 17 remained putative sweep candidates.

Figure 4.7 displays the distribution of these candidates along the chromosome 19.

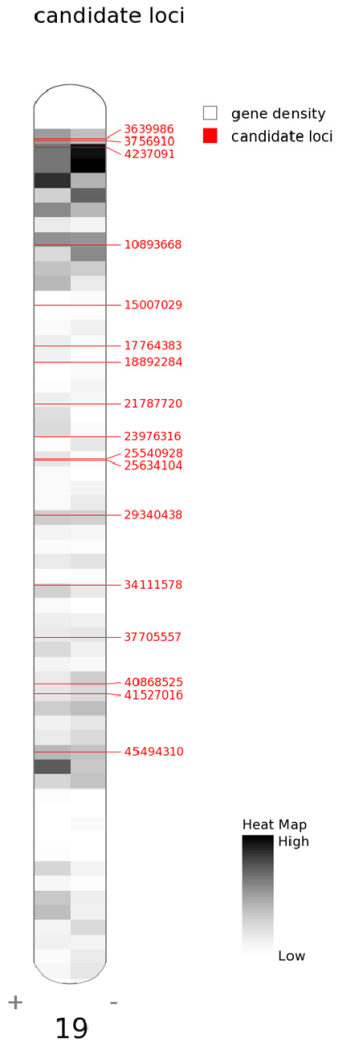


Figure 4.7 Distribution of candidate loci throughout the chromosome 19.

Overall loci are scattered throughout the entire chromosome.

Except one locus (15007029) which is observed in a region without any genes, all others candidates are located nearby genes whereas gene density varies, *i.e.* no association between candidate marker density and gene density is observed. A list of all genes flanking the residual candidate loci containing gene ID and functional information is added to the supplementary material (Supplement 3) as well as pictures of the respective chromosomal regions (Supplement 4). In two cases candidate loci are clustered within one chromosomal region. First the already described loci 3639986 and 3756910 (see Chapter 4.3.1 and 4.3.2) both sweep loci in the German and Iranian population. The marker 3639986 is located directly in the gene *Lrp5* supporting evidence for this gene to be shaped by selection. Secondly loci 25540928 and 25634104 for which a sweep pattern is observed among the '*domesticus*' populations (Chapter 4.3.3) and the French and Czech Republic populations respectively; both loci flanking the same gene *Dmrt1* indicating that the respective populations underwent an adaptive process at this gene.

4.4 Discussion

The present study is mainly focused on the analysis of the French and German house mouse populations. Concerning the investigation of natural selection these two populations are of special interest for several reasons. 1) The split of the two populations is relatively recent which is an important aspect concerning estimates of selection frequency. The distance of time in which information of adaptive events is stored in microsatellite markers is very much restricted by their mutation rate. At the same time the genetic distance must be large enough so that beneficial mutation can arise and be driven to fixation. Since the divergence time between the two focal populations does not exceed 18,000 generations, it is a fairly young split. Hence patterns of selection events that occurred after the split will still be maintained. At the same time the split is old enough so that evolution had enough time to act on the two populations. Thus the German French comparison is well suited for the identification of chromosomal regions that harbor recent adaptive mutations in natural populations. 2) The demography of the respective populations is well known and patterns of major demographic events that would be expected to mimic patterns of positive selection can be excluded. The two compared populations have very similar population parameters, *i.e.* it is not a comparison between an ancestral and a derived population, but two that have recently split from each other. Comparisons of heterozygosities between the focal populations and an ancestral one at 118 randomly chosen loci throughout the genome support that a bottleneck event between the two populations is unlikely (Teschke 2006).

Furthermore, if one of the populations would have undergone a bottleneck after the colonization of Western Europe, this would have had genome wide effects. Since the time was too short for sufficient mutations to occur which could recover genome variability, a bottleneck event would be reflected in the average heterozygosity. Moreover even if a bottleneck event would have occurred previously to the split the signatures would have been wiped out by continuous mutations. Finally if loci became randomly fixed during the colonization process, they would be present in both populations and therefore be canceled out in the ln*RH* statistic. Hence, doubts about the authenticity of the identified sweep patterns due to possible bottleneck or drift events can be rejected.

Given the time frame in which the respective populations were established, the data allows a rough estimate of the frequency of positive selection events in natural populations in the house mouse. Teschke et al. (2008) propose that the core window size of selective sweeps range between 20-100 kb, *i.e*. the size of the sweep which is detectable with the stringent screening procedure. Assuming that the generation split has occurred 3,000 years ago (Cucchi et al. 2005) and a generation time of three generations per year (Karn et al. 2002) relates to 9,000 generationsin in each lineage after the split. If the screen would have detected 50 kb windows (proposed as average core window size by Techke et al. 2008) the chromosome 19 is almost completely sampled which translates to 2 % of the mouse genome. Given that 2 and 3 significant regions (note that 2 significant loci are expected to be linked in the German populatione) were detected in each population, results in 125 sweeps having occurred since the population split, or one selective event every 70 generations.

The here assumed detectable core window size of 50 kb appears reliable. Analyzable markers were located on average in 80 kb distance among the chromosome. Only in one case a considerably larger sweep window than 80 kb was detected (see locus 3639986 and 3756910). Thus, most of the sweep windows are less than 80 kb in size. These findings are fairly consistent with the results obtained in the genome screen from Teschke et al. (2008), where the authers propose a sweep rate of one sweep evry 30-45generations. Both assessments are higher than the commonly stated frequency of 1 sweep in 250 generations which was initially proposed by Haldane (1957).

However the here proposed frequency of positive selection is still assumed to be underestimated and should be rather seen as a minimum rate of adaptive events due to the following reasons: sweeps were identified using ln*RH* statistics which detects selective patterns only if the pairwise differences of heterozygosities vary extremely. In other words, a sweep is only detected right after one allele was driven to fixation in one population and no new mutations have arisen yet. Thus the period in which a selective event will be identified is fairly small and the time a selective pattern will be maintained depends very much on the slippage or mutation rate of the flanking microsatellite. This is mainly determined by the number of repeat units, its motif and the motif length. Slippage rates are supposed to be approximately zero in very short repeated regions, or are significantly reduced if interrupted by even a single point

mutation; supporting a low probability for new mutations to arise in short or disrupted stretches (Sainudiin et al. 2004; Lai and Sun 2003). Generally speaking the mutation rate ascends with the number of perfectly repeated microsatellite units.

Hence if a short allele is fixed, the sweep pattern persists for a longer time, whereas long repeat stretches recover variability within a relatively short timeframe. Since only those loci which display an accurate sweep signature will be identified as sweep candidates, particularly recent adaptive events will be detected. Further, short microsatellite loci will be predominantly sampled, as accumulation of new mutations is slow, which prevents the pattern to be blurred by new alleles. In contrast sweeps linked to long, fast mutating alleles, are more likely to be missed in this analysis. Finally, regions under strong selection will be primarily detected since recent sweep patterns generated under weak selection would not have had the time to build up a significant sweep signature.

Estimations concerning the frequency of selection are based on statistically corrected ln*RH* values. As has been said, the ln*RH* statistic only selects for specific, or classical, allele patterns in a certain time frame within a sweep event. If these results are corrected for multiple testing this effect gets even stronger, *i.e.* the section in which a selective event will be identified is even more constricted. Consequently the effect of overrepresented sweeps associated to short alleles will get stronger.

The classical model of a selective sweep assumes that the beneficial allele was created only once by mutation and only a single copy of the beneficial allele contribute to fixation. This scenario would generate the classical pattern of a sweep which is also referred as a 'hard sweep'. This assumption may not hold if selection acts on standing genetic variation or if adaptation occurs from recurrent mutation or migration (Pennings and Hermisson 2006). If multiple allele copies of the flanking neutral marker are associated with the positively selected variant, the signature of selection will differ from the classical signature and consequently these 'soft selective sweeps' will also remain undetected (Przeworski et al. 2005). Similarly, if recombination occurs in the early stage of a sweep event, two different alleles will be linked to the selected site and the sweep pattern will be shifted towards multiple frequent alleles instead of a single one (Fay and Wu 2000).

Previous estimates of the frequency of selection have been made by comparison of genome data in *Drosophila* (Smith and Eyre-Walker 2002, Andolfatto

2005). Depending on the consideration of non-coding DNA estimates of amino acid substitution driven by natural selection greatly vary between 0.02-10 in 100 generations. Since many experimental approaches suffer from only detecting relatively strongly selected mutations, estimating the strength of selection is still of central importance to reveal the exact frequency of natural selection (Wright and Andolfatto 2008). Many attempts try to answer this question by using aminoacid substitution models (Yang and Nielsen 2008; Chamary et al. 2006; Hurst 2006; Eyre-Walker and Keightley 2007). But although comparative genomics approaches started to incorporate highly conserved noncoding sequences into their analysis, comparisons are mainly focused on using diverged species (Siepel et al. 2005; Wright and Andolfatto 2008) in which adaptive events that result from *cis*-regulatory modifications are difficult to identify.

However, in the young investigated populations the footprints of adaptation which go along with amino acid substitutions as well as changes in *cis*-regulatory elements are still present. As both types of adaptive events will be identified with the present screens for signatures of selective sweeps the frequency of detected selection in this screen is expected to be higher.

As discussed above the time window within a sweep pattern is identified by ln*RH* statistic is mainly determined by the mutation rate of the microsatellite. Concordantly a large fraction of candidate loci detected in this analysis are fixed for a short allele and contain tri- or tetranucleotide repeats which are less prone to slippage than dinucleotides. In consideration of the fact that slow mutating alleles have a higher potential to be detected by the statistic the correction for multiple testing applying to ln*RH* values should be revised. Since absolute ln*RH* values are not automatically associated to the sweep reliability, correction for multiple testing leads rather to selection of slow mutating alleles than to the removal of false positives. Hence the statistic is suitable for the identification of extreme cases but levels of significance should be interpreted in the context to individual mutation rates.

In addition to the selection coefficient and the mutation rate the chromosomal size of a selective sweep depends on the degree of linkage between adjacent sites measured as rates of recombination per physical distance (e.g. cM/Mb). Referring to positive selection, the size of the region which is fixed around the selected site is proportional to the strength of selection, and the rate of recombination (Aquadro et al.

2001), *i.e*. occurrence of large 'footprints' is expected in areas of lower recombination and vice versa.

In this approach investigated markers were sampled systematically in a 50 kb distance. Since no additional flanking loci have been analyzed around the candidate loci, the actual size of the identified sweeps cannot be determined. As has been said before, the average distance of investigated loci was 80 kb for the comparison of the German and French populations. Linked candidate loci were only observed twice, indicating that the sweep size does not exceed 80 kb in most cases. Given that the majority of detected 'footprints' of selection are rather small, suggests that the selection coefficients are roughly in balance with the recombination rates, effectively isolating positive selection on one locus from neighboring loci.

One cluster of linked candidate loci was observed in between the two loci 3639986 and 3756910. In total four contiguous loci were detected in this region, consistently displaying a sweep pattern in the German population. These loci are expected to belong to one single selective event whereby the minimum sweep size is 120 kb. This indicates strong selection acting at this locus because only a very strong sweep is capable of severely reducing larger regions (Jeffrey et a. 2008). Likewise one linked sweep candidate was detected next to the locus 15007029 which classifies the size of the sweep of about 50-100 kb. Recombination rates taken from Jensen-Seaman et al. (2004) highly vary between these two sweeps. As expected, low rates are observed at locus 15007029; in contrast the large sweep around locus 3756910 was detected in a region of comparatively high recombination again, pointing towards exceptional, strong selection coefficients at this site.

It should be mentioned that some large sweep regions might not be identified if the sweep comprises markers with high mutation rates. As has been described 'recovery patterns' are not detected by the ln*RH* statistic. Thus, if large sweep regions are associated with markers of high mutation rates, even in regions of low recombination the selective pattern will be blurred by new mutations. For example at locus 3731289, which is flanked by candidate loci, a long allele was fixed and due to ongoing mutations the sweep pattern is already recovered. Furthermore, large sweep regions might not be detected if both compared populations are affected by selection. In this case the loci will be canceled out in the statistic similar as described for bottleneck or drift effects during the colonization process. However, recently strong

selection affecting large chromosomal regions would have been revealed in comparison of the two Western European populations against the ancestral Iranian one. Since only four 'European sweeps' were detected, supports the presumption of strong selection being rare.

Among the German French comparison two loci (15007029 and 45494310) displayed an allele pattern where two main alleles are present in the non sweep population which are several mutation steps apart without almost any alleles in between. Such pattern would not be assumed for an uninfluenced mutating microsatellite. Since both loci contain perfect repeat units and no interference was observed in the flanking sequence the allele pattern cannot be explained by sequence characteristics. Instead such pattern would be generated if both populations would have been under selection where either two allele copies were linked to the positively selected mutation or a recombination event happened during the early phase of the sweep event in the German population. As has been discussed 'soft sweeps' will generally not be detected in this analysis. However if 'early' recombination happens only in one population or multiple alleles are linked to the beneficial site in the one population but not in the other both scenarios could generate sweep patterns that resemble classical sweeps.

Having mentioned the difficulties that go along with the statistic, one main advantage of the presented sweep approach using microsatellite sequences should be pointed out in the following. As has been said maintenance of the sweep pattern is mainly determined by the mutation rate which is basically allele specific, whereas not only the length but also the correctness of the repeat is essential. For example compared to pure AC repeats microsatellites interrupted by even a single point mutation exhibit a twofold decrease in their mutation rate (Sainudiin et al. 2004). Taking microsatellite information about the quality and length simply from the mouse genome of the available databases is insufficient as these data might deviate from natural populations. In Chapter 3.3 for example it has been shown that the observed lengths of the microsatellites in the present data are on average longer than expected from the database which could be for instance due to annotation bias.

A number of genome scans for the identification of candidate genes have been performed based on microsatellites but since the high throughput Sanger sequencing

is very time consuming for microsatellites most of them are based on sequence information taken from available databases (Schlötterer 2002; Storz et al. 2004; Teschke et al. 2008).

In this approach the actual microsatellite sequences were used to affirm the reliability of detected candidate loci. Referring to ln*RH* values 29 loci were selected as putative sweep candidates. The selection comprises both, loci which resulted from within subspecies comparisons as well as between subspecies. A closer look at the microsatellite sequences of the respective loci revealed that at 12 loci the observed absence of variability is more likely due to specific characteristics of the microsatellite sequence than to genetic hitchhiking. Two of these cases contain very short microsatellite stretches whereas the motif is only four times repeated. The other loci are discarded as being reliable candidates since they do not fulfill the criteria of a perfectly repeated mictosatellite motif. In all cases the observed microsatellite pattern is altered such that the mutation or slippage rate is expected to be close to zero. If the mutation rate is very low drift cannot be excluded as driving factor for the observed sweep pattern as the locus would not recover variability in an adequate time period. Thus the candidates for selective sweeps should be assured by either examination of the microsatellite sequence or typing of additional flanking markers.

Having supported sweep reliability for candidate loci based on sequence level, the remaining candidates will be discussed in the following: four microsatellite loci showed a sweep pattern in both Western European populations in comparison to the ancestral Iranian one. In all cases the same allele was fixed in the respective loci, a pattern as it would have been generated by drift or bottleneck events during the colonization process. As mentioned before persistency of such patterns over a long time requires low mutation rates of the fixed alleles. This applies only for the tetranucleotide microsatellite 6550201 at which the sweep allele contains only four repeats. Microsatellite stretches at the other three loci are longer so that the time since the split would have been sufficient for new mutations to reconstitute variability. Hence it would not be expected to find such low values of heterozygosity maintained over a long time period. Furthermore, it is very unlikely that the same allele is linked to the selected site if the sweep occurred under parallel evolution within both populations.

Similar patterns were detected comparing the German and French populations

to the Iranian. This analysis revealed sweep overlaps between the French and Iranian as well as the German and Iranian population. While the latter case might be explained by parallel evolution, since two different alleles are fixed in the respective populations (see locus 3640065, Figure 4.5), this does not apply to the sweep overlap of the Iranian and French population. Here two candidate loci (41527152 and 4237056) were identified harboring the same sweep allele in both populations. Again, this can not be explained by low mutation rates since neither the microsatellite stretches nor the flanking regions displayed any irregularities in either one of the populations.

Moreover three reliable subspecies specific sweeps were detected, where three '*domesticus*' populations were compared to both '*musculus*' populations. Similar to the previously described cases all populations harbor the same sweep allele. In general, subspecies wide reduction of variability is explained by either the loss of the microsatellite or point mutations which lower the mutation rate as described above. For the '*musculus*' sweeps in fact low mutation rates were observed due to short repeat units or interrupted microsatellite motifs. However all three '*domesticus*' sweeps are fixed for a perfectly repeated di- or trinucleotide containing up to 9 repeats, *i.e*. the maintenance of one high frequent allele can not be explained by low mutation rates.

The observation of identical alleles being fixed among genetically distinct populations strongly indicates that the alleles have spread through the populations. This indicates that beneficial mutations have the potential of spreading even across spatially distinct populations, although it has been shown that the populations are well separated and no ongoing migration is observed based on neutral markers (see Figure 1.7).

The spread of mutations between genetically separated populations of geographically very distant locations has rarely been described. For house mice Harr (2006) suggested introgression of genetic material across the subspecies boundaries based on SNPs. In agreement Bonhomme et al. (2007) found examples of identical or nearly identical alleles of hypervariable minisatellite loci across subspecies and in geographically very distant locations, suggesting recurrent gene flow between already differentiated entities. So far examples for the spread of beneficial mutations through distinct populations have been described for insecticide resistance genes in

Drosophila and *Culex pipiens* (Daborn et al. 2002; Raymond et al. 1998). However selection coefficients on insecticide resistance are considerably high and the spread of the mutation between populations is facilitated by the fact that most insecticide-treated areas are connected by plane or other transportation systems that are suitable for passive migration (Raymond et al. 1998). To confirm the finding of rather frequent allele exchange between the subpopulations, as a next step flanking regions need to be sequenced to investigate the haplotypes. Observation of the fixed alleles nested within identical haplotypes among the populations, would provide further evidence for the migration of the respective allele.

In addition to the fixation of the same sweep allele between several populations, sweep overlaps between populations were also observed for different fixed alleles. Beside the locus 3640065 at which two different alleles are fixed in the German and Iranian population, loci showing signals of selection in several populations were also detected between the Kazakh and French population as well as the Czech Republic and the '*domesticus*' populations. Especially for the sweep shared by the French and Kazakh population (10893566) low mutation rates can be rejected as explanation for the maintenance of low variability since both populations are fixed for a rather long dinucleotide allele (see Table 4-4). Given that different alleles are linked to the selected site, at least in each subspecies these cases can be taken as indication for parallel evolution.

Finding such a pattern between young invaded populations, like the Western European ones, and old ancestral populations like, Kazakhstan and Czech Republic, support the conclusion that positive selection is not exclusively associated with strong environmental changes, *e.g.* invasions into new habitats, but an ongoing background process which affects all populations.

With regards to putative candidate genes it was assumed that the closted gene is targeted by selection. This assumption might not hold if several genes are within the investigated region or if selection acts on *trans*-regulatory elements. However, in most cases the microsatellite analysis points towards a certain gene where the marker is located in close proximity or directly within the gene. This may be taken as indication for the respective gene to be the target site. The detected genes do not belong to a particular functional category, but rather seem to reflect a more or less random subset of genes. Notably a lot of reliable sweep candidate genes are involved

in a variety of processes and have multiple functions (see Supplement 3) indicating pleiotropic effects. In addition signaling and developmental genes were repeatedly detected which are reported to have broad pleiotropic effects (Doebley and Lukens 1998) (Artieri, Haerty, and Singh 2009). For example the candidate gene *Dmrt1* contributes to developmental and metabolic processes and has been shown to be essential for testicular differentiation in vertebrates (Shinseog et al. 2007). Another sweep candidate, gene *Lrp5*, is participating in a variety of different biological processes comprising cell adhesion, cellular process and reproduction and cooperates in two different pathways, Alzheimer disease-presenilin pathway and *Wnt*-signaling pathway. Intuitively, pleiotropy creates more constraints on a protein, attributable to its more diverse function involving more functional residues. Supported by Fisher's idea, that more pleiotropic mutations are less likely to be advantageous due to a higher chance that a positive alteration in one module could be unfavorable for another (Fisher and Bennett 1999), genes of high pleiotropy are expected to be under strong stabilizing selection (Hodgkin 1998; Xionglei He and Zhang 2006). At the same time species may be forced to evolve by altering the connections and interactions among interacting entities to preserve their cohesive functions over vast stretches of evolutionary time (Fraser 2005; Hartwell et al. 1999). Since the amount of standing variation should be low due to purifying selection this rather suggests positive selection acting via new mutations. Compatible to this hypothesis, the sweep patterns of apparently conserved genes resemble the classical pattern of a classical hard sweep.

The relatively high number of sweeps detected with a fairly conservative approach is consistent with a number of studies that indicate adaptation might be common on the genomic scale (Smith and Eyre-Walker 2002; Andolfatto 2007; Fay et al. 2002). However as has been said only classical sweep patters are revealed with the applied statistics. One major benefit of the presented approach is the access to the actual microsatellite sequences. The next step will be the performance of an additional analysis based on the sequence data. Mutation rates will be estimated using a mutation model as described by (Sainudiin et al. 2004) which can be used to estimate the likelihoods for locus specific allele distribution at each site. This will increase the spectrum of observed sweep patters since recovery patterns as well as ‘soft sweeps’ will be detectable and hence supply deeper insights into the process of adaptation.

5 References

Auffray, J., Vanlerberghe, F., Britton-Davidian, J., 1990. The house mouse progression in Eurasia: a palaeontological and archaeozoological approach. Biological Journal of the Linnean Society, 41(1-3), 13-25.

Anakk, S., Kalsotra, A., Shen, Q., Vu, M.T., Staudinger, J.L., Davies, P.J.A., Strobel, H.W., 2003. Genomic characterization and regulation of CYP3a13: role of xenobiotics and nuclear receptors. FASEB Journal, 17(12), 1736-1738.

Andolfatto, P., and Przeworski, M., 2000. A Genome-Wide Departure From the Standard Neutral Model in Natural Populations of Drosophila. Genetics, 156(1), 257-268.

Andolfatto, P., 2001. Adaptive hitchhiking effects on genome variability. Current Opinion in Genetics & Development, 11(6), 635-641.

Andolfatto, P., 2007. Hitchhiking effects of recurrent beneficial amino acid substitutions in the Drosophila melanogaster genome. Genome Research, 17, 1755-1762.

Applied Biosystems, 2009. GeneMapper Software. URL: https://products.appliedbiosystems.com/ab/en/US/adirect/ab?cmd=catNavigate2&catID=600798&tab=DetailInfo.

Aquadro, C.F., Bauer DuMont, V., Reed, F.A., 2001. Genome-wide variation in the human and fruitfly: a comparison. Current Opinion in Genetics & Development, 11(6), 627-634.

Artieri, C., Haerty, W., Singh, R., 2009. Ontogeny and phylogeny: molecular signatures of selection, constraint, and temporal pleiotropy in the development of Drosophila. BMC Biology, 7(1), 42-55.

Barton, N.H., 2000. Genetic hitchhiking. Philosophical Transactions of the Royal Society B: Biological Sciences 355(1403), 1553-1562.

Bazykin, G.A., Fyodor A.K., Aleksey Y.O., Shamil S., Alexey S.K., 2004. Positive selection at sites of multiple amino acid replacements since rat-mouse divergence. Nature 429(6991), 558-562.

Begun, D.J., Charles F.A., 1993. African and North American populations of Drosophila melanogaster are very different at the DNA level. Nature 365(6446), 548-550.

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

Bierne N., Eyre-Walker A., 2004. The Genomic Rate of Adaptive Amino Acid Substitution in Drosophila. Mol Biol Evol 21(7), 1350-1360.

Bonhomme, F., Rivals, E., Orth, A., Grant, G., Jeffreys, A., Bois, P.R., 2007. Species-wide distribution of highly polymorphic minisatellite markers suggests past and present genetic exchanges among house mouse subspecies. Genome Biology, 8(5), R80.

van den Bosch, H.M., Bünger, M., de Groot, P.J., van der Meijde, J. Hooiveld, G.J.E.J., Müller, M., 2007. Gene expression of transporters and phase I/II metabolic enzymes in murine small intestine during fasting. BMC Genomics 8, 267-267.

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

Bowcock, A.M., Ruiz-Linares, A., Tomfohrde, J., Minch, E., Kidd, J.R., Cavalli-Sforza, L.L., 1994. High resolution of human evolutionary trees with polymorphic microsatellites. Nature, 368(6470), 455-457.

Bryant, J.P., Reichardt, P.B., Clausen, T.P., Provenza, F.D., Kuropat, P.J., 1992. Woody plant-mammal interactions. In: G. Rosenthal & M. Berenbaum, eds. Herbivores, their interactions with secondary plant metabolites. Academic Press, New York,(433-371).

Cain, A.J., 1951. So-called Non-adaptive or Neutral Characters in Evolution. Nature 168(4271), 424.

Calabrese, P., Raazesh S., 2005. Models of Microsatellite Evolution. In Statistical Methods in Molecular Evolution,(290-305).

Cavalli-Sforza, L.L., 1966. Population Structure and Human Evolution. Proceedings of the Royal Society of London. Series B, Biological Sciences 164(995), 362-379.

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

Charlesworth, B. and Charlesworth, D., 2010. <ahref="http://science.jrank.org/pages/48503/Hitchhiking.html">Hitchhiking - Hitchhiking by Adaptive Mutations, Background Selection

Clark, A.G., Glanowski, S., Nielsen, R., Thomas, P.D, Kejariwal, A., Todd, M.A., Tanenbaum, D.M., 2003. Inferring Nonneutral Evolution from Human-Chimp-Mouse Orthologous Gene Trios. Science, 302(5652), 1960-1963.

Codon Code Corporation, 2009. Better Software for DNA Sequencing. URL: http://www.codoncode.com/aligner/.

Daborn, P.J., Yen, J.L., Bogwitz, M.R., Le Goff, G., Feil, E., Jeffers, S. & Tijet, N., 2002. A Single P450 Allele Associated with Insecticide Resistance in Drosophila. Science, 297(5590), 2253-2256.

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

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

Doebley, J., Lukens, L., 1998. Transcriptional Regulators and the Evolution of Plant Form. Plant Cell, 10(7), 1075-1082.

Eaton, D., Gallagher, E., 1994. Mechanisms of Aflatoxin Carcinogenesis. Annual Review of Pharmacology and Toxicology, 34, 135-172.

Echchgadda, I., Song, C.S., Oh, T., Ahmed, M., De La Cruz, I.J. & Chatterjee, B., 2007. The Xenobiotic-sensing Nuclear Receptors PXR, CAR and Orphan Nuclear Receptor HNF-4{alpha} in the Regulation of Human Steroid-/Bile Acid-Sulfotransferase. Molecular Endocrinology, 21, 2099-2111.

Ellegren, H., 2004. Microsatellites: simple sequences with complex evolution. Nature Review Genetics, 5(6), 435-445.

Ensemble, 2010. Splice variants Cyp3a25. URL: http://Mar2010.archive.ensembl.org/Mus_musculus/Gene/Splice?g=ENSMUSG00000029630.

Eyre-Walker, A., Keightley, P.D., 2007. The distribution of fitness effects of new mutations. Nature Review Genetics 8(8), 610-618.

Excoffier, L, T Hofer, and M Foll. 2009. Detecting loci under selection in a hierarchically structured population. Heredity 103(4), 285-298.

Farombi, E., 2006. Review-Aflatoxin contamination of foods in developing countries: Implications for hepatocellular carcinoma and chemopreventive strategies. African Journal of Biotechnology, 5(1), 1-14.

Fay, J.C., Wu, C.I., 2000. Hitchhiking under positive Darwinian selection. Genetics, 155(3), 1405-1413.

Fay, J.C., Wyckoff, G.J., Wu, C., 2002. Testing the neutral theory of molecular evolution with genomic data from Drosophila. Nature, 415(6875), 1024-1026.

Fisher, S.R.A., Bennett, J.H., 1999. The genetical theory of natural selection, Oxford University Press.

Foley, W.J., Hume, I.D., 1987. Nitrogen Requirements and Urea Metabolism in Two Arboreal Marsupials, the Greater Glider (Petauroides volans) and the Brushtail Possum (Trichosurus vulpecula), Fed Eucalyptus Foliage. Physiological Zoology, 60(2), 241-250.

Foley, W.J., McLean, S. & Cork, S.J., 1995. Consequences of biotransformation of plant secondary metabolites on acid-base metabolism in mammals—A final common pathway? Journal of Chemical Ecology , 21(6), 721-743.

Fraser, H.B., 2005. Modularity and evolutionary constraint on proteins. Nature Genetics 37(4), 351-352.

Fu, C., Xiong, J., Miao, W., 2009. Genome-wide identification and characterization of cytochrome P450 monooxygenase genes in the ciliate Tetrahymena thermophila. BMC Genomics, 10, 208-208.

Fu, Y. X., 1997. Statistical Tests of Neutrality of Mutations against Population Growth, Hitchhiking and Background Selection. Genetics 147,(2), 915-925.

Gillespie, J.H., 2000. Genetic drift in an infinite population. The pseudohitchhiking model. Genetics 155(2), 909-919.

Goldstein, D.B., Clark, A.G., 1995. Microsatellite variation in North American populations of Drosophila melanogaster. Nucleic Acids Research 23(19), 3882-3886.

Gonzalez, F.J., Nebert, D.W., 1990. Evolution of the P450 gene superfamily: animal-plant 'warfare', molecular drive and human genetic differences in drug oxidation. Trends in Genetics, 6, 182-186.

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

Guengerich, F.P., Johnson, W.W., Shimada, T., Ueng, Y.-F., Yamazaki, H., Langouët, S., 1998. Activation and detoxication of aflatoxin B1. Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis, 402(1-2), 121-128.

Haddrill, P.R., Thornton K.R., Charlesworth B., Andolfatto, P., 2005. . Genome Research 15, 790-799.

Haldane, J. 1957. The cost of natural selection. Journal of Genetics 55(3), 511-524.

Harr, B., Kauer, M., Schlötterer, C., 2002. Hitchhiking mapping: A population-based fine-mapping strategy for adaptive mutations in Drosophilamelanogaster. Proceedings of the National Academy of Sciences of the United States of America, 99(20), 12949-12954.

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

Hartwell, L.H., Hopfield, J.J., Leibler, S., Murray, A.W., 1999. From molecular to modular cell biology. Nature, 402, C47-C52.

Haubold, B., 2008. 'drawDotPlot.awk'. Computerprogram (unpublished).

He, X., Zhang, J., 2006. Toward a Molecular Understanding of Pleiotropy. Genetics, 173(4), 1885-1891.

Hermisson, J. 2009. Who believes in whole-genome scans for selection[quest]. Heredity 103(4), (online): 283-284.

Hirschhorn, J.N., Daly, M.J., 2005. Genome-wide association studies for common diseases and complex traits. Nature Review Genetics, 6(2), 95-108.

Hodgkin, J., 1998. Seven types of pleiotropy. International Journal of Developmental Biology, 42(3), 501-505.

Huttley, G.A., Smith, M.W., Carrington, M., O'Brien, S.J. 1999. A Scan for Linkage Disequilibrium Across the Human Genome. Genetics 152(4), 1711-1722.

Ihle, S., Ravaoarimanana, I., Thomas, M., Tautz, D., 2006. An Analysis of Signatures of Selective Sweeps in Natural Populations of the House Mouse. Molecular Biology and Evolution, 23(4), 790-797.

Ingelman-Sundberg, M., 2005. The human genome project and novel aspects of cytochrome P450 research. Toxicology and Applied Pharmacology, 207(2, Supplement 1), 52-56.

Ingelman-Sundberg, M., Oscarson, M., McLellan, R.A., 1999. Polymorphic human cytochrome P450 enzymes: an opportunity for individualized drug treatment. Trends in Pharmacological Sciences, 20(8), 342-349.

Ioannides, C., 2008. Cytochromes P450, Royal Society of Chemistry, RSC Publishing, Cambridge.

Jensen, J.D., Kim, K., Bauer DuMont, V., Aquadro, C.F., Bustamante, C.D. 2005. Distinguishing Between Selective Sweeps and Demography Using DNA Polymorphism Data. Genetics 170(3), 1401-1410.

Jensen, J.D., Thornton, K.R., Andolfatto, P., 2008. An Approximate Bayesian Estimator Suggests Strong, Recurrent Selective Sweeps in Drosophila. PLoS Genetics, 4(9), e1000198.

Jensen-Seaman, M.I., Furey, T.S., Payseur, B.A., Lu, Y., Roskin, K.M., Chen, C.-F., Thomas, M.A., Haussler, D. & Jacob, H.J., 2004. Comparative Recombination Rates in the Rat, Mouse, and Human Genomes. Genome Research, 14(4), 528-538.

Johansson, Å., Karlsson, P., Gyllensten, U., 2003. A novel method for automatic genotyping of microsatellite markers based on parametric pattern recognition. Human Genetics, 113(4), 316-324.

Jover, R., Bort, R., Gómez-Lechón, M.J., Castell, J.V., 2001. Cytochrome P450 regulation by hepatocyte nuclear factor 4 in human hepatocytes: A study using adenovirus-mediated antisense targeting. Hepatology, 33(3), 668-675.

Kalendar, R., Lee, D. & Schulman, A., 2009. Fast PCR Software for PCR Primer and Probe Design and Repeat Search. In A. Mansour, ed. Genes, Genomes and Genomics, 3 (Special Issue 1), 1-14.

Kanno, Y., Otsuka, S., Hiromasa, T., Nakahama, T., and Yoshio Inouye. 2004. Diurnal difference in CAR mRNA expression. Nuclear Receptor 2(1), 6

Kaplan, N.L., Hudson, R.R., Langley, C.H., 1989. The ``Hitchhiking Effect'' Revisited. Genetics 123(4), 887-899.

Karn, R.C., Orth, A., Bonhomme, F., Boursot, P., 2002. The Complex History of a Gene Proposed to Participate in a Sexual Isolation Mechanism in House Mice. Mol Biol Evol 19(4), 462-471.

Kauer, M.O., Dieringer, D., Schlotterer, C., 2003. A Microsatellite Variability Screen for Positive Selection Associated With the "Out of Africa" Habitat Expansion of Drosophila melanogaster. Genetics, 165(3), 1137-1148.

Kayser, M., Brauer, S., Stoneking, M., 2003. A Genome Scan to Detect Candidate Regions Influenced by Local Natural Selection in Human Populations. Molecular Biology and Evolution, 20(6), 893-900.

Keightley, P.D., Smith, N.G.C., Gaffney, D., Eyre-Walker, A., 2002. Quantifying the Slightly Deleterious Mutation Model of Molecular Evolution. Molecular Biology and Evolution, 19, 2142-2149.

Khaitovich, P., Paabo, S., Weiss, G., 2005. Toward a Neutral Evolutionary Model of Gene Expression. Genetics 170(2), 929-939

Kim, Y., Stephan, W., 2003. Selective Sweeps in the Presence of Interference Among Partially Linked Loci. Genetics 164(1), 389-398.

Kim, S., Bardwell, V.J., Zarkower, D., 2007. Cell type-autonomous and non-autonomous requirements for Dmrt1 in postnatal testis differentiation. Developmental Biology, 307(2), 314-327.

Kimura, M., 1968. Evolutionary Rate at the Molecular Level. Nature, 217, 624-626.

Kohn, M.H., Pelz, H.J., Wayne, R.K., 2000. Natural selection mapping of the warfarin-resistance gene. Proceedings of the National Academy of Sciences of the United States of America 97(14), 7911-7915.

Kottek, M., Grieser, J., Beck, C., Rudolf, B., Rubel, F., 2006. World Map of the Köppen-Geiger climate classification updated. Meteorologische Zeitschrift,15(3), 259-263.

Kruglyak, S., Durrett, R.T., Schug, M.D., Aquadro. C.F., 1998. Equilibrium distributions of microsatellite repeat length resulting from a balance between slippage events and point mutations. Proceedings of the National Academy of Sciences of the United States of America, 95(18), 10774-10778.

Lai, Y., Sun, F., 2003. The Relationship Between Microsatellite Slippage Mutation Rate and the Number of Repeat Units. Molecular Biology and Evolution, 20(12), 2123-2131.

Lewis, D.F.V., 2001. Guide to cytochromes P450, CRC Press, Routledge, UK

Lewontin, R.C., Krakauer, J., 1973. Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphisms. Genetics 74(1), 175-195.

Lewontin, R. C., 1974. The genetic basis of evolutionary change. Columbia biological series, no. 25. New York: Columbia University Press.

Lucchini, V., Fabbri, E., Marucco, F., Ricci, S., Boitani, L., Randi, E., 2002. Noninvasive molecular tracking of colonizing wolf Canis lupus packs in the western Italian Alps. Molecular Ecology, 11(5), 857-868.

Luikart, G., England, P.R., Tallmon, D., Jordan, S., Taberlet, P,. 2003. The power and promise of population genomics: from genotyping to genome typing. Nat Rev Genet 4(12), 981-994.

Lynch, M., Conery, J.S., 2000. The Evolutionary Fate and Consequences of Duplicate Genes. Science 290(5494), 1151-1155.

Ma, J., Alyce Bradbury, J., King, L., Maronpot, R., Davis, L.S., Breyer, M.D., Zeldin, D.C., 2002. Molecular cloning and characterization of mouse CYP2J6, an unstable cytochrome P450 isoform. Biochemical Pharmacology, 64(10), 1447-1460.

Machida, M., Gomi, K., 2010. Aspergillus: Molecular Biology and Genomics. Caister Academic Press.

Maglich, J.M, Stoltz, C.M., Goodwin, B., Hawkins-Brown, D., Moore, J.T., Kliewer, S.A., 2002. Nuclear pregnane x receptor and constitutive androstane receptor regulate overlapping but distinct sets of genes involved in xenobiotic detoxification. Molecular Pharmacology, 62(3), 638-646.

Margulies, M., Egholm, M., Altman, W.E., Attiya, S., Bader, J.S., Bemben, L.A., Berka, J., 2005. Genome Sequencing in Open Microfabricated High Density Picoliter Reactors. Nature, 437(7057), 376-380.

May, B.K, Dogra, S.C., Sadlon, T.J., Bhasker, C.R., Cox, T.C., Bottomley, S.S., 1995. Molecular regulation of heme biosynthesis in higher vertebrates. Progress in Nucleic Acid Research and Molecular Biolog,y 51, 1-51.

MGSC, 2002. Initial sequencing and comparative analysis of the mouse genome. Nature, 420(6915), 520-562.

Moore, L.B., Parks, D.J., Jones, S.A., Bledsoe, R.K., Consler, T.G., Stimmel, J.B., Goodwin, B., 2000. Orphan Nuclear Receptors Constitutive Androstane Receptor and Pregnane X Receptor Share Xenobiotic and Steroid Ligands. journal of biological chemistry 275, 15122-27.

Nachman, M.W., 1997. Patterns of DNA Variability at X-Linked Loci in Mus Domesticus. Genetics, 147(3), 1303-1316.

Nachman, M.W., Crowell, S.L., 2000. Estimate of the Mutation Rate per Nucleotide in Humans. Genetics 156(1), 297-304.

NCBI, 2008. Primer-Blast. URL: http://www.ncbi.nlm.nih.gov/tools/primer-blast/index.cgi?LINK_LOC=BlastNews.

Nebert, D., Gonzalez, F., 1987. P450 Genes: Structure, Evolution, and Regulation. Annual. Reviews of Biochemistry, 56, 945-993.

Nebert, D., Nelson, D., Feyereisen, R., 1989. Evolution of the cytochrome P450 genes. Xenobiotica, 19(10), 1149-1160.

Nei, M., 1978. Estimation Of Average Heterozygosity And Genetic Distance From A Small Number Of Individuals. Genetics 89(3), 583-590.

Nelson, D.R., Zeldin, D.C., Hoffman, S.M., Maltais, L.J., Wain, H.M., Nebert, D.W., 2004. Comparison of cytochrome P450 (CYP) genes from the mouse and human genomes, including nomenclature recommendations for genes, pseudogenes and alternative-splice variants. Pharmacogenetics, 14(1), 1-18.

Nielsen, R., 2005. Molecular signatures of natural selection. Annual Review of Genetics, 39, 197-218.

Ohta, T., Kimura, M. 1975. The Effect of Selected Linked Locus on Heterozygosity of Neutral Alleles (the Hitch-Hiking Effect). Genetical Research 25(03), 313-325.

Ohta, T., 1992. The nearly neutral theory of molecular evolution. Annual Review of Ecology and Systematics, 23, 263-286.

Oleksyk, T.K., Smith, M.W., O'Brien, S.J., 2010. Genome-wide scans for footprints of natural selection. Philosophical Transactions of the Royal Society B: Biological Sciences 12: 185-205.

Palo, R.T., Robbins, C.T., 1991. Plant defenses against mammalian herbivory. CRC Press.

Pálsson, B, Pálsson, F., Perlin, M., Gudbjartsson, H., Stefánsson, K., Gulcher, J., 1999. Using quality measures to facilitate allele calling in high-throughput genotyping. Genome Research, 9(10), 1002-1012.

Payseur, B. A., Cutter, A. D., Nachman, M.W., 2002. Searching for Evidence of Positive Selection in the Human Genome Using Patterns of Microsatellite Variability. Molecular Biology And Evolution. 19, 1143-1153.

Pennings, P.S., Hermisson, J., 2006. Soft Sweeps III: The Signature of Positive Selection from Recurrent Mutation. PLoS Genetics, 2(12), e186.

Pruitt, K.D., Tatusova, T., Maglott, D.R., 2006. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Research, gkl842.

Przeworski, M., 2002. The Signature of Positive Selection at Randomly Chosen Loci. Genetics 160(3), 1179-1189.

Przeworski, M., Coop, G., Wall, J., 2005. The signature of positive selection on standing genetic variation. Evolution, 59(11), 2312-2323.

Qiang Xie, Q.-Y.Z., Zhang, Y., Su, T., Gu, J., Kaminsky, L.S. & Ding, X., 2000. Induction of Mouse CYP2J by Pyrazole in the Eye, Kidney, Liver, Lung, Olfactory Mucosa, and Small Intestine, but Not in the Heart. Drug Metabolism and Disposition, 28(11), 1311–1316.

Raymond, M., Chevillon, C., Guillemaud, T., Lenormand, T., Pasteur, N., 1998. An overview of the evolution of overproduced esterases in the mosquito Culex pipiens. Philosophical Transactions of the Royal Society B: Biological Sciences, 353(1376), 1707-1711.

Richard, S.M., Bailliet, G., Paez, G.L., Bianchi, M.S., Peltomaki, P. & Bianchi, N.O. 2000. Nuclear and Mitochondrial Genome Instability in Human Breast Cancer. Cancer Research, 60(15), 4231-4237.

Widmer, A., Arntz, A.M., Burke, J.M., & Rieseberg, L.H., 2002. Directional Selection Is the Primary Cause of Phenotypic Diversification. Proceedings of the National Academy of Sciences of the United States of America, 99(19), 17.

Roche, 2010. 454 Sequencing, Products and Solutions. URL: http://www.454.com/products-solutions/experimental-design-options/long-single-reads.asp.

Rowe, L.B., Barter, M.E., Kelmenson, J.A. & Eppig, J.T., 2003. The Comprehensive Mouse Radiation Hybrid Map Densely Cross-Referenced to the

Recombination Map: A Tool to Support the Sequence Assemblies. Genome Research, 13, 122-133.

Rozas, J., Sánchez-DelBarrio, J.C., Messeguer, X., Rozas, R., 2003. DnaSP, DNA polymorphism analyses by the coalescent and other methods. Bioinformatics (Oxford, England), 19(18), 2496-2497.

Rozen, S. & Skaletsky, H., 1999. Primer3 on the WWW for General Users and for Biologist Programmers. Bioinformatics Methods and Protocols, 132, 365-386.

Sabeti, P.C., Reich, D.E., Higgins, J.M., Levine, H.Z.P., Richter, D.J., Schaffner, S.F., Gabriel, S.B. et al., 2002. Detecting recent positive selection in the human genome from haplotype structure. Nature, 419, 832-837.

Varilly, P., Fry, B., Lohmueller, J., Hostetter, E., Cotsapas, C., Xie, X., Byrne, E. H., ... Sabeti, P. C., 2007. Genome-wide detection and characterization of positive selection in human populations. Nature, 449(7164), 913-8.

Sainudiin, R., Durrett, R.T., Aquadro, C.F. & Nielsen, R., 2004. Microsatellite Mutation Models: Insights From a Comparison of Humans and Chimpanzees. Genetic,s 168(1), 383-395.

Salcedo, T., Geraldes, A., Nachman, M.W., 2007. Nucleotide Variation in Wild and Inbred Mice. Genetics, 177(4), 2277-2291.

Schlötterer, C., 2002. A Microsatellite-Based Multilocus Screen for the Identification of Local Selective Sweeps. Genetics, 160(2), 753-763.

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), 309-320.

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

Schlötterer, C., Dieringer, D., 2005. Molecular Biology and Evolution In Selective Sweep. pp. 55-64. URL: http://dx.doi.org/10.1007/0-387-27651-3_5

Schluter, D., 2000. The ecology of adaptive radiation. Oxford University Press.

Mackay, T.F.C., Aquadro, C.F., Schug, M.D., 1997. Low mutation rates of microsatellite loci in Drosophila melanogaster. Nature Genetics, 15(1), 99.

Schug, M.D., Mackay, T.F.C., Aquadro, C.F., 1997. Low mutation rates of microsatellite loci in Drosophila melanogaster. Nature Genetics 15(1), 99-102.

Schuler, G.D., 1997. Sequence mapping by electronic PCR. Genome Research, 7(5), 541-550.

Schweinsberg, J., Durrett R., 2005. Random partitions approximating the coalescence of lineages during a selective sweep. The Annals of Applied Probability 15(3), 1591-1651.

Sella, G., Petrov, D.A., Przeworski, M., Andolfatto, P., 2009. Pervasive Natural Selection in the Drosophila Genome? PLoS Genetics, 5(6), e1000495.

Siepel, A., Bejerano, G., Pedersen, J.S., Hinrichs, A.S., Hou, M.H., Rosenbloom, K., Clawson, H., 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Research, 15, 1034-1050.

Slatkin, M., 1995. Hitchhiking and associative overdominance at a microsatellite locus. Molecular Biology and Evolution 12(3), 473-480.

Smith, J.M., Haigh, J., 1974. The Hitch-Hiking Effect of a Favourable Gene. Genetical Research, 23(01), 23-35.

Smith, N.G.C., Eyre-Walker, A., 2002. Adaptive protein evolution in Drosophila. Nature, 415(6875), 1022-1024.

Staubach, F., Teschke, M., Voolstra, C.R., Wolf, J.B.W., Tautz, D., 2010. A test of the neutral model of expression change in natural populations of house mouse subspecies. Evolution 64(2), 549-560.

Storz, J.F., 2005. Using genome scans of DNA polymorphism to infer adaptive population divergence. Molecular Ecology, 14(3), 671-688.

Storz, J.F., Payseur, B.A., Nachman, M.W., 2004. Genome Scans of DNA Variability in Humans Reveal Evidence for Selective Sweeps Outside of Africa. Molecular Biology and Evolution, 21(9), 1800-1811.

Suyama, M., Torrents, D., Bork, P., 2006. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. Nucleic Acids Research, 34(Suppl. 2), W609-612.

Tajima, F., 1989. Statistical Method for Testing the Neutral Mutation Hypothesis by DNA Polymorphism. Genetics 123, no. 3 (November 1): 585-595.

Tautz, D., 1989. Hypervariability of simple sequences as a general source for polymorphic DNA markers. Nucleic Acids Research, 17(16), 6463-71.

Tautz, D., 2000. Evolution of transcriptional regulation. Current Opinion in Genetics & Development 10, no. 5 (October 1): 575-579. doi:10.1016/S0959-437X(00)00130-1.

Tautz, D., Renz. M., 1984. Simple sequences are ubiquitous repetitive components of eukaryotic genomes. Nucl. Acids Res. 12, no. 10 (May 25): 4127-4138.

Teschke, M. 2006. A Systematic Assessment of Signatures of Positive Selection Events in Natural Populations of the House Mouse. University of Cologne.

Teschke, M., Mukabayire, O., Wiehe, T., Tautz, D., 2008. Identification of Selective Sweeps in Closely Related Populations of the House Mouse Based on Microsatellite Scans. Genetic,s 180(3), 1537-1545.

Teshima, K.M., Coop, G., Przeworski, M., 2006. How reliable are empirical genomic scans for selective sweeps? Genome Research 16: 702-712.

Thomas, J.H., 2007. Rapid birth-death evolution specific to xenobiotic cytochrome P450 genes in vertebrates. PLoS Genetics, 3(5), e67.

Thornton, K.R., Jensen, J.D., Becquet, C., Andolfatto, P., 2007. Progress and prospects in mapping recent selection in the genome. Heredity 98(6), 340-348.

Tirona, R.G., Lee, W., Leake, B.F., Lan, L.B., Cline, C.B., Lamba, V., Parviz, F., Duncan S.A., Inoue, Y., Gonzalez, F.J., Schuetz, E.G., Kim, R.B., 2003. The orphan nuclear receptor HNF4alpha determines PXR- and CAR-mediated xenobiotic induction of CYP3A4. Nature Medicine, 9(2), 220-224.

Van Vaalen, L., 1973. A new evolutionary law. Evol Theory 1: 1–30.

Voight, B.F., Kudaravalli, S., Wen, X., Pritchard, J.K., 2006. A Map of Recent Positive Selection in the Human Genome. PLoS Biology, 4(3), e72.

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

Waxman, D.J., 1999. P450 Gene Induction by Structurally Diverse Xenochemicals: Central Role of Nuclear Receptors CAR, PXR, and PPAR. Archives of Biochemistry and Biophysics, 369(1), 11-23.

Whitehead, A., Crawford, D.L., 2005. Variation in tissue-specific gene expression among natural populations. Genome Biology 6(2), R13-R13.

Wiehe, T.H., Stephan, W., 1993. Analysis of a genetic hitchhiking model, and its application to DNA polymorphism data from Drosophila melanogaster. Mol Biol Evol 10(4), 842-854.

Wilson, A.C., Maxson, L.R., Sarich, V.M., 1974. Two Types of Molecular Evolution. Evidence from Studies of Interspecific Hybridization. Proceedings of the National Academy of Sciences of the United States of America 71(7), 2843-2847.

Wolfe, K.H., Sharp, P.M., Li, W.H., 1989. Mutation rates differ among regions of the mammalian genome. Nature 337(6204), 283-285.

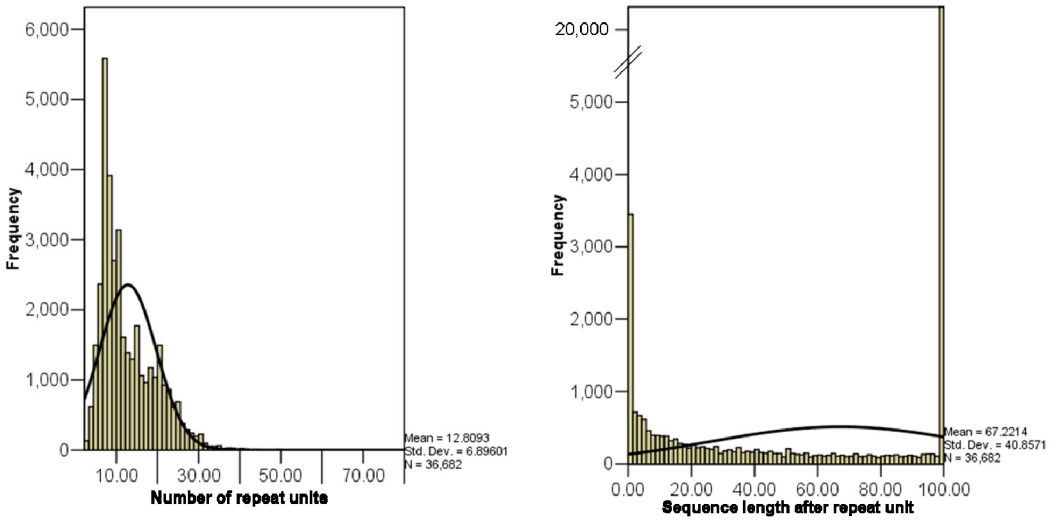
Wray, G.A., 2007. The evolutionary significance of cis-regulatory mutations. Nat Rev Genet 8(3), 206-216.

Wright, S.I., Andolfatto, P., 2008. The Impact of Natural Selection on the Genome: Emerging Patterns in Drosophila and Arabidopsis. Annual Review of Ecology, Evolution, and Systematics, 39(1), 193-213.

Yang, Z., Nielsen, R., 2008. Mutation-Selection Models of Codon Substitution and Their Use to Estimate Selective Strengths on Codon Usage. Molecular Biology and Evolution, 25(3), 568-579.

6 Supplement

Supplement 1 Test of neutrality for data used in Figure 3.4.

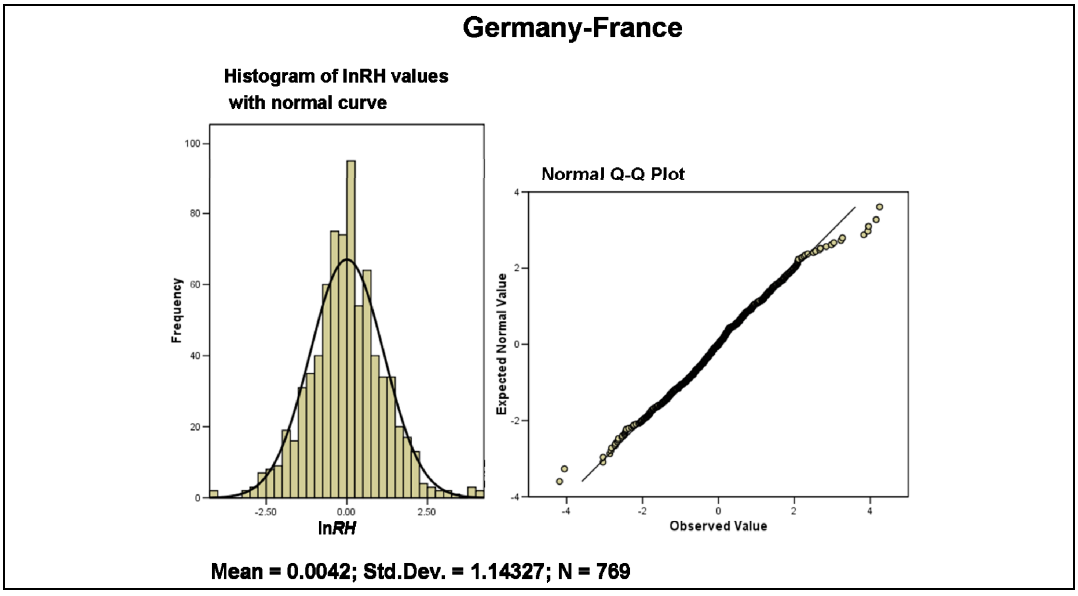


N=36,682

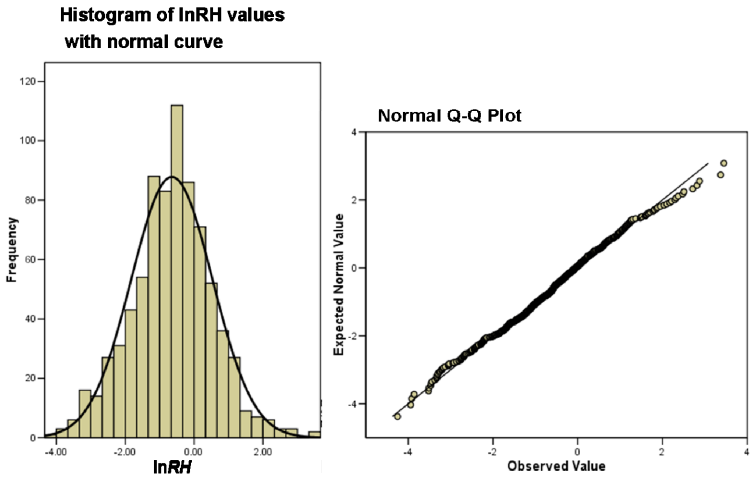
Number of repeat units: Z=28.47, p=0.000

Sequence length after repeat unit: Z=63.22, p=0.000

Supplement 2 Distribution of ln*RH* values for pairwise comparisons.

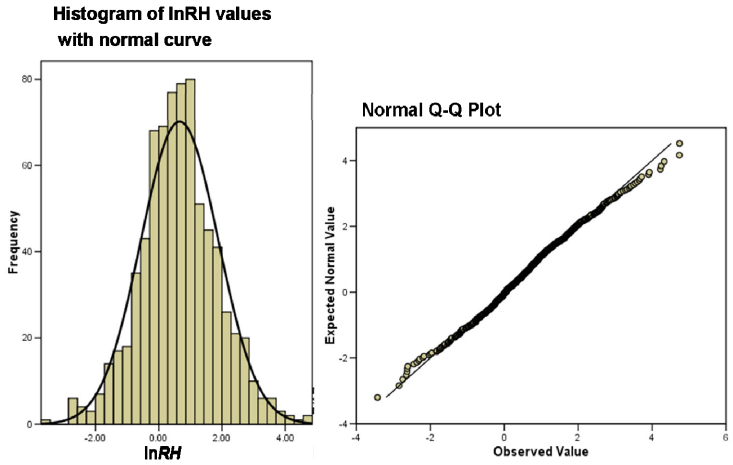


Germany-Iran



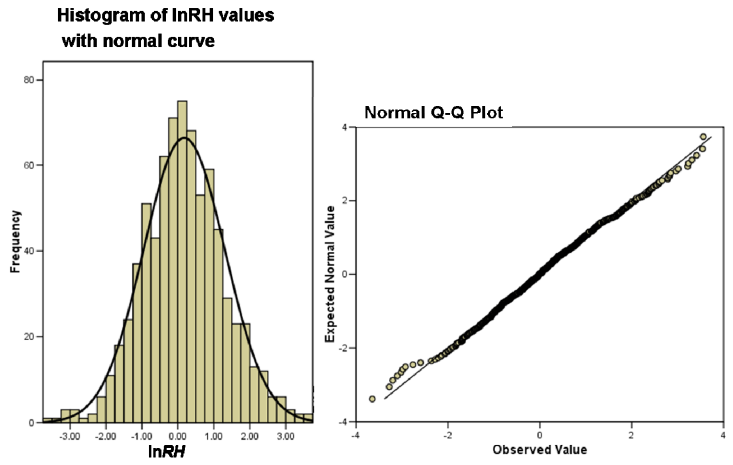
Mean = -0.6473; Std.Dev. = 1.18127; N = 780

Iran-France



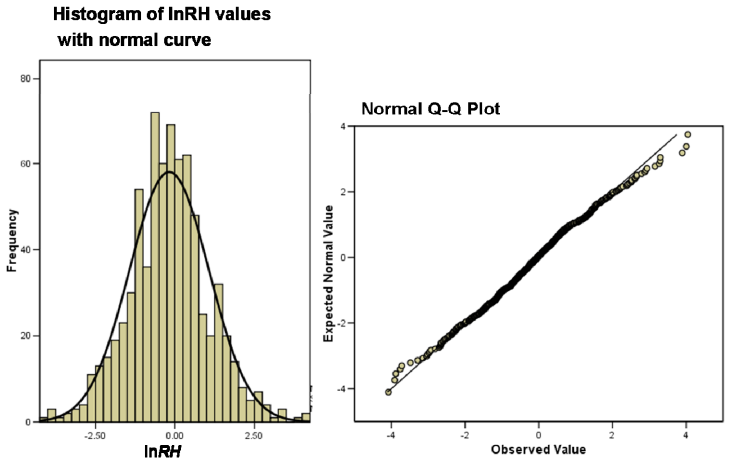
Mean = 0.6601; Std.Dev. = 1.122681; N = 755

Germany-Cameron



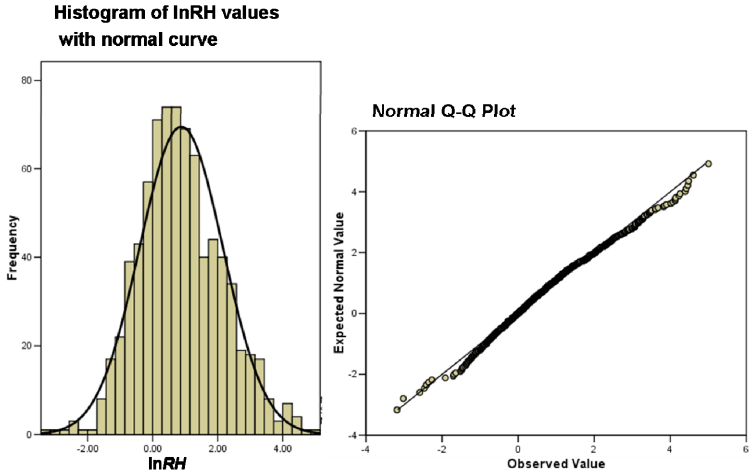
Mean = 0.1752; Std.Dev. = 1.13192; N = 753

Cameron-France



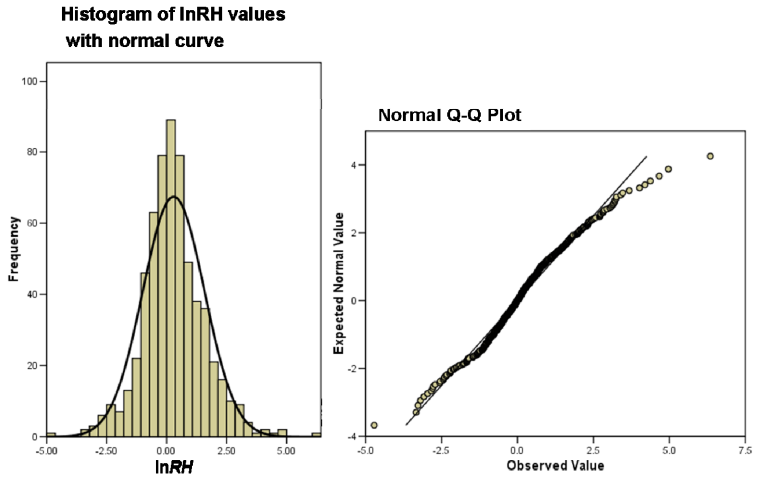
Mean = -0.1805; Std.Dev. = 1.125237; N = 729

Iran-Cameron



Mean = 0.8764; Std.Dev. = 1.28158; N = 781

Czech Republic-Kazakhstan



Mean = 0.2904; Std.Dev. = 1.28666; N = 609

Supplement 3 Table of putative candidate genes.

Genes adjacent to candidate loci in '*Mus musculus domesticus*' comparisons

Marker	S-P	Closest gene	Gene ID ENSMUSG 000000...	Gene type	Process	Component	Function
3639986	G/I	–rp5	24913	KPC	anterior/poster or pattern formation, cell migration involved in gastrulation	integral to membrane, membrane	receptor activity
3756910	G/I	Suv420h1	45098	KPC	chromatin modification, histone methylation	condensed nuclear chromosome, centromeric region	histone methyltransferase activity (H4-K20 specific), histone-lysine N-methyltransferase activity
42370 91	F/I	Pold4	24854	KPC	DNA replication, DNA-dependent DNA replication	nucleus	DNA-directed DNA polymerase activity, nucleotidyltransferase activity
15007029	F	AC116999.2-200	63586	NPG			
18892284	G/F	Trpm6	24727	KPC	calcium ion transport, ion transport	apical plasma membrane, brush border membrane	ATP binding, calcium channel activity
21787720	G	Fam108b	47368	KPC	biological process	extracellular region	hydrolase activity
23976316	G/F	Apba1	24897	KPC	gamma-aminobutyric acid secretion, glutamate secretion	plasma membrane	PDZ domain binding, phosphatidylinositol-4,5-bisphosphate binding
40868525	G/F	Conj	25010	KPC	biological process	nucleus	molecular function
41527016	F/I	AC112271.4	44065	KPC			
		–cor	25019	KPC	negative regulation of transcription from RNA polymerase II promoter, regulation of transcription	nucleus	DNA binding, receptor activity
45494310	F	Btrc	25217	KPC	branching involved in mammary gland duct morphogenesis, mammary gland epithelial cell proliferation	cytoplasm, nucleus	protein binding

S-P = Sweep Population; KPC=known protein coding; NPG=Novel pseudogene; G= Germany; F= France; I= Iran

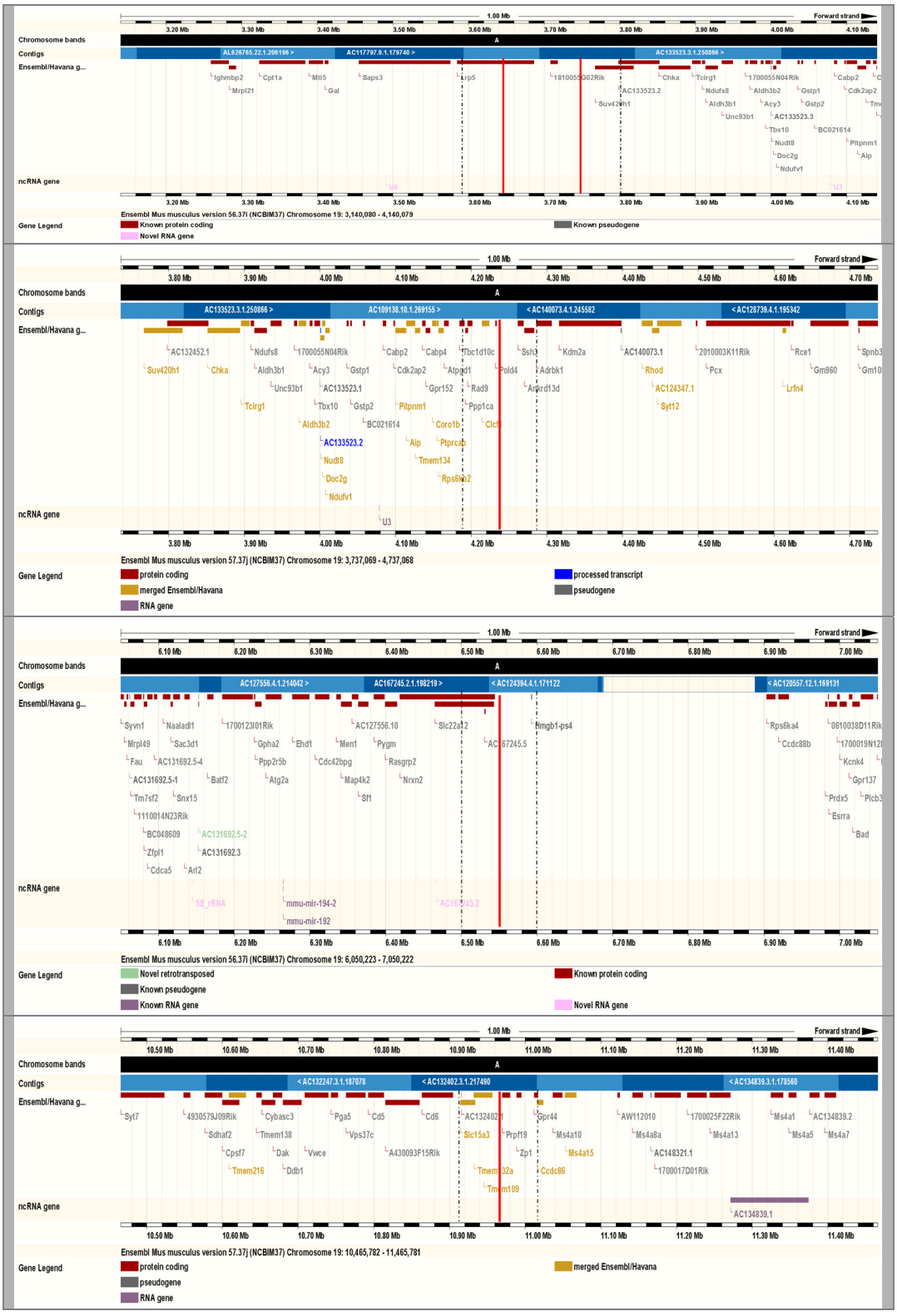
Genes adjacent to candidate loci in subspecies comparisons

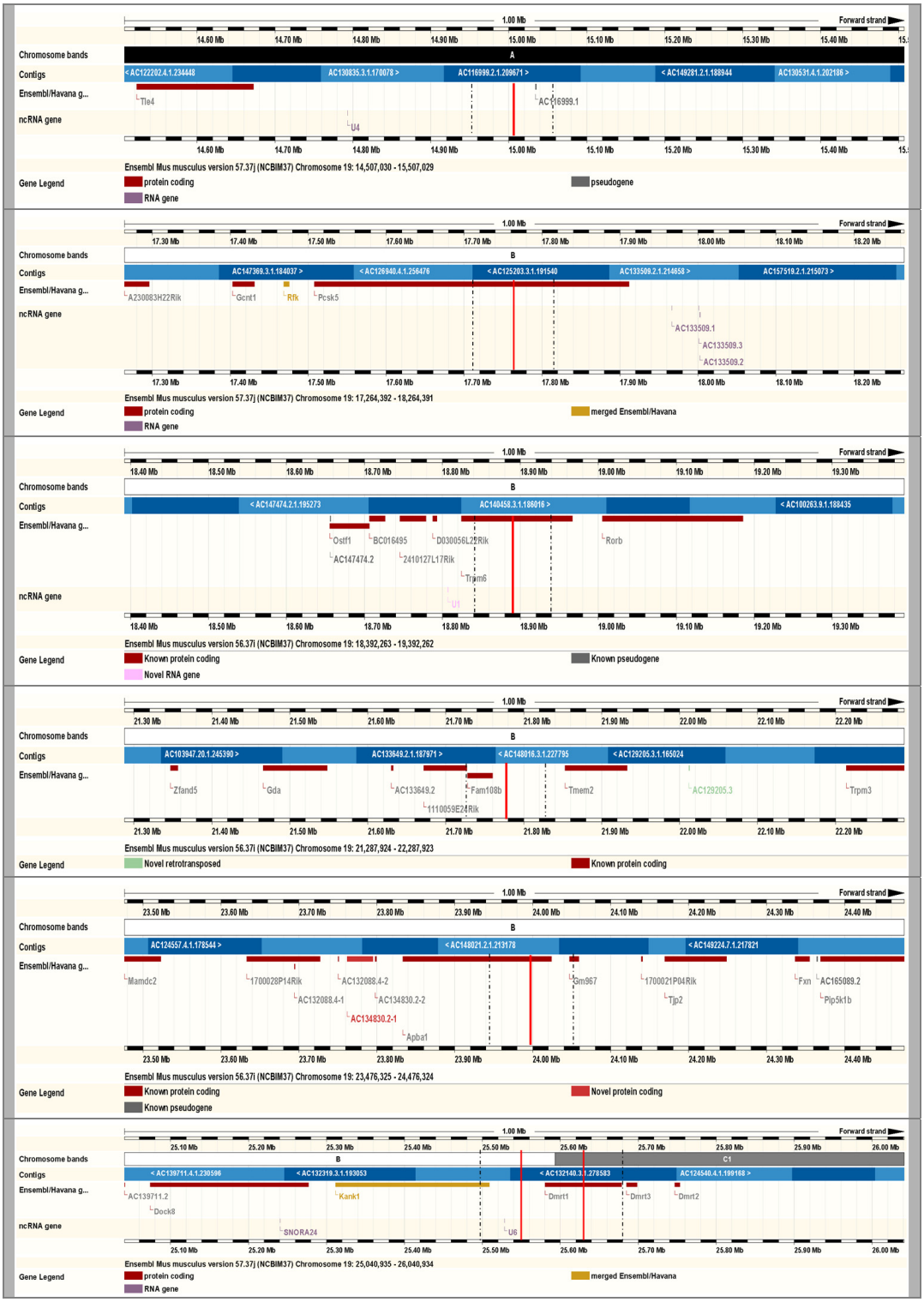
Marker	S-P	Closest gene	Gene ID ENSMUSG 000000...	Gene type	Process	Component	Function
10893868	K/F	Cd6	24670	KPC	cell adhesion, peptidyl-tyrosine phosphorylation	external side of plasma membrane, integral to membrane	protein binding, protein kinase binding
		Slc15a3	24737	KPC	oligopeptide transport, peptide transport	integral to membrane	symporter activity, transporter activity
17764383	CR/dom	Pcsk5	24713	KPC	anterior/posterior pattern formation, cytokine biosynthetic process...	extracellular region, extracellular space...	endopeptidase activity, hydrolase activity...
25540928	dom	Kank1	32702	KPC	negative regulation of actin filament polymerization	cellular component	molecular function
		Dmrt1	24837	KPC	cell differentiation, intracellular signaling pathway	nucleus	DNA binding, metal ion binding
25634104	CR/F	Dmrt1	24837	KPC	cell differentiation, intracellular signaling pathway	nucleus	DNA binding, metal ion binding
29340438	CR/dom	Jak2-201	25705	KPC	activation of MAPKK activity, apoptosis, JAK-STAT cascade..	caveola, cytoplasm...	acetylcholine receptor binding, ATP binding, interleukin-12 receptor binding...
34111578	dom	Lipk	24771	KPC	lipid catabolic process, lipid metabolic process	extracellular region	hydrolase activity
37705557	dom	Exoc3	53799	KPC	exocytosis, protein transport, vesicle docking during exocytosis	exocyst	
		Cyp26c1	62432	KPC	anterior/posterior pattern formation, central nervous system development		metal ion binding, monooxygenase activity

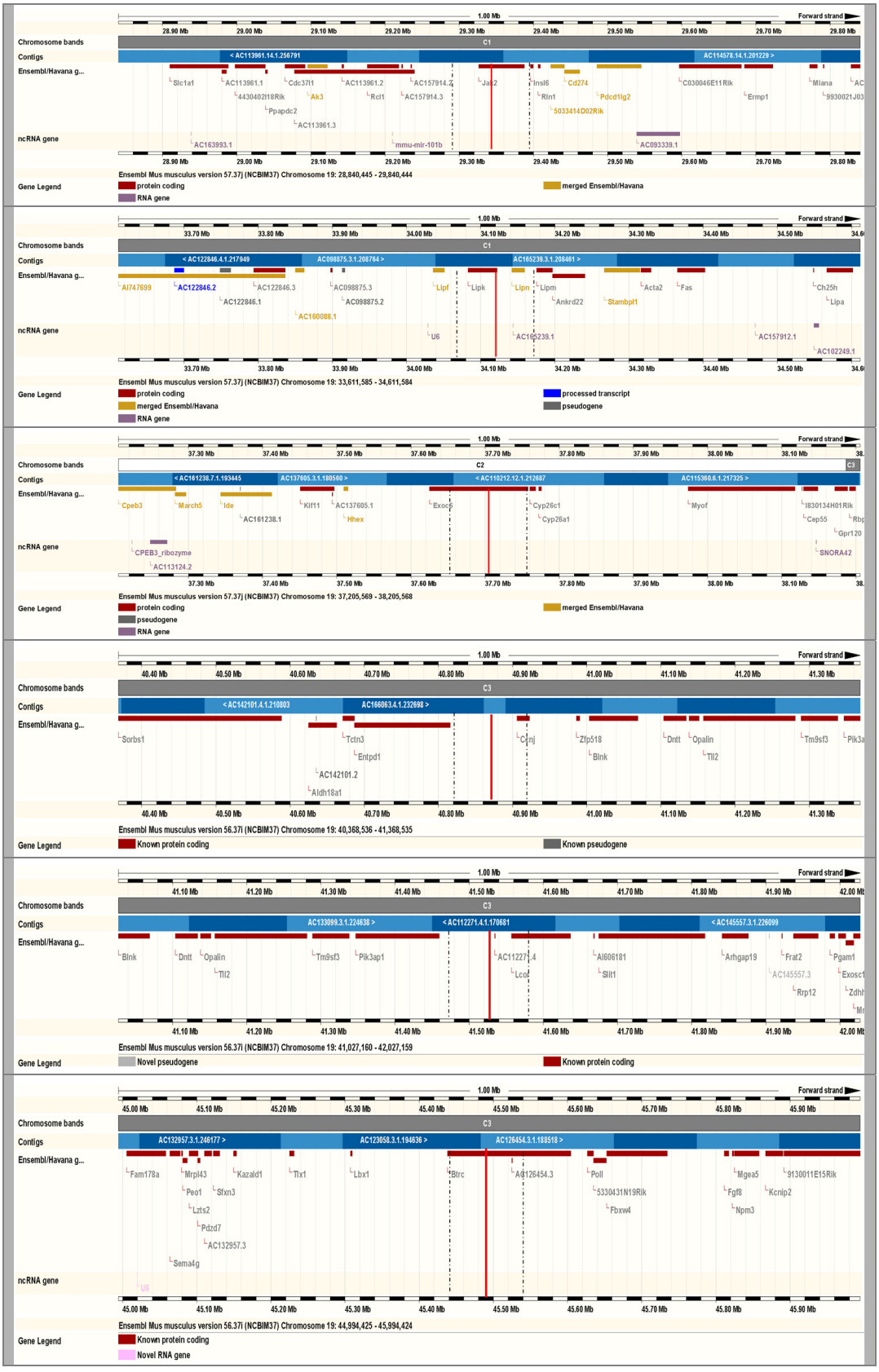
S-P = Sweep Population; KPC=known protein coding; NPG=Novel pseudogene; K = Kazakhstan; F= France; CR = Czech Republic; dom = '*domesticus*'

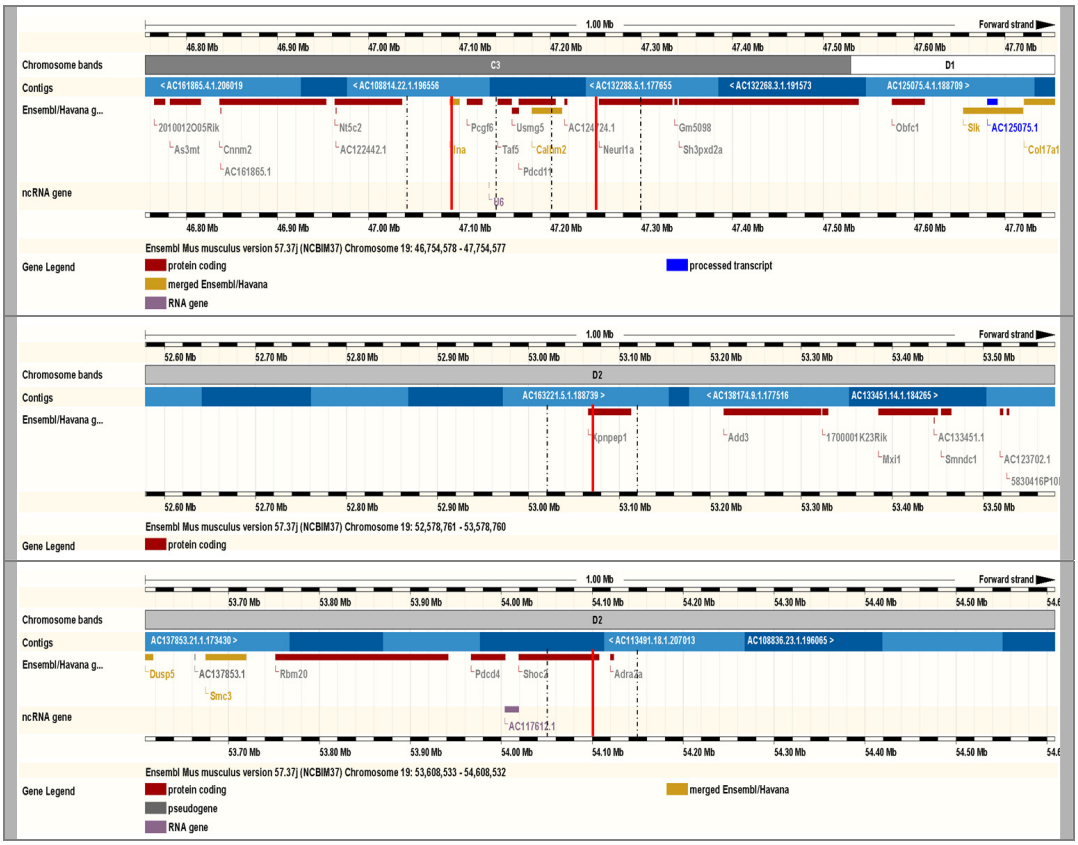
Supplement 4 Pictures of chromosomal regions of the candidate loci.

Red lines indicate candidate locus position. Black dotted lines mark the 50 kb window around the respective loci.




Chromosome bands
Contigs
Ensembl/Havana g...
ncRNA gene
Gene Legend
Forward strand
Ensembl Mus musculus version 57.37j (NCBIM37) Chromosome 19: 14,507,030 - 15,507,029
protein coding
pseudogene
RNA gene
Ensembl Mus musculus version 57.37j (NCBIM37) Chromosome 19: 17,264,392 - 18,264,391
merged Ensembl/Havana
Ensembl Mus musculus version 56.37i (NCBIM37) Chromosome 19: 18,392,263 - 19,392,262
Known protein coding
Known pseudogene
Novel RNA gene
Ensembl Mus musculus version 56.37i (NCBIM37) Chromosome 19: 21,287,924 - 22,287,923
Novel retrotransposed
Known protein coding
Ensembl Mus musculus version 56.37i (NCBIM37) Chromosome 19: 23,476,325 - 24,476,324
Novel protein coding
Known pseudogene
Ensembl Mus musculus version 57.37j (NCBIM37) Chromosome 19: 25,040,935 - 26,040,934


Chromosome bands
Contigs
Ensembl/Havana g...
ncRNA gene
Gene Legend
Ensembl Mus musculus version 57.37j (NCBIM37) Chromosome 19: 28,840,445 - 29,840,444
protein coding
merged Ensembl/Havana
RNA gene
Ensembl Mus musculus version 57.37j (NCBIM37) Chromosome 19: 33,611,585 - 34,611,584
processed transcript
pseudogene
Ensembl Mus musculus version 57.37j (NCBIM37) Chromosome 19: 37,205,569 - 38,205,568
Ensembl Mus musculus version 56.37l (NCBIM37) Chromosome 19: 40,368,536 - 41,368,535
Known protein coding
Known pseudogene
Ensembl Mus musculus version 56.37l (NCBIM37) Chromosome 19: 41,027,160 - 42,027,159
Novel pseudogene
Ensembl Mus musculus version 56.37l (NCBIM37) Chromosome 19: 44,994,425 - 45,994,424
Novel RNA gene


1.00 Mb
Forward strand
Chromosome bands
Contigs
Ensembl/Havana g...
ncRNA gene
Gene Legend
protein coding
merged Ensembl/Havana
RNA gene
processed transcript
Ensembl Mus musculus version 57.37j (NCBIM37) Chromosome 19: 46,754,578 - 47,754,577
Ensembl Mus musculus version 57.37j (NCBIM37) Chromosome 19: 52,578,761 - 53,578,760
Ensembl Mus musculus version 57.37j (NCBIM37) Chromosome 19: 53,608,533 - 54,608,532
pseudogene

Supplement 5 Table of ln*RH* values of all candidate loci.

Locus	ln*RH*-value	Population with reduced variability	Ensembl Gene ID of closest gene
19:3369257..3369279	2.21	CR (Kazakh-CR)	ENSMUSG00000024900
19:3404465..3404488	2.06	Cam (Iran-Cam)	ENSMUSG00000024905
19:3639986..3640094	-2.38	Ger (Ger-Fra)	ENSMUSG00000024913
19:3639986..3640094	-2.03	Iran (Iran-Fra)	ENSMUSG00000024913
19:3702999..3703039	-2.67	Ger (Ger-Fra)	ENSMUSG00000085960
19:3756910..3756945	-2.16	Ger (Ger-Fra)	ENSMUSG00000035372
19:3756910..3756945	-2.31	Iran (Iran-Fra)	ENSMUSG00000035372
19:3815428..3815450	-2.17	Ger (Ger-Cam)	ENSMUSG00000045098
19:3815428..3815450	2.07	Fra (Fra-Cam)	ENSMUSG00000045098
19:3917362..3917395	-3.39	Ger (Ger-Cam)	ENSMUSG00000059734
19:4237056..4237091	3.45	Iran (Ger-Iran)	ENSMUSG00000024854
19:4237056..4237091	-2.72	Iran (Iran-Cam)	ENSMUSG00000024854
19:4237056..4237091	3.48	Fra (Ger-Fra)	ENSMUSG00000024854
19:4237056..4237091	2.67	Fra (Fra-Cam)	ENSMUSG00000024854
19:5442311..5442334	-2.68	Iran (Iran-Fra)	ENSMUSG00000047423
19:5562085..5562116	2.19	CR (Kazakh-CR)	ENSMUSG00000050657
19:5808084..5808114	-2.14	Iran (Iran-Fra)	ENSMUSG00000024941
19:6011655..6011692	2.26	Fra (Ger-Fra)	ENSMUSG00000024942
19:6011655..6011692	2.68	Fra (Iran-Fra)	ENSMUSG00000024942
19:6066117..6066135	-2.60	Ger (Ger-Cam)	ENSMUSG00000024799
19:6142082..6142107	-3.57	Ger (Ger-Fra)	ENSMUSG00000024944
19:6550201..6550243	-2.05	Ger (Ger-Iran)	ENSMUSG00000085694
19:6550201..6550243	2.65	Cam (Iran-Cam)	ENSMUSG00000085694
19:6550201..6550243	2.51	Fra (Iran-Fra)	ENSMUSG00000085694
19:8002952..8002980	2.23	Iran (Ger-Iran)	ENSMUSG00000067656
19:8066743..8066776	-2.06	Ger (Ger-Cam)	ENSMUSG00000080393
19:8705554..8705589	-2.74	Ger (Ger-Cam)	ENSMUSG00000024650
19:8705554..8705589	-2.01	Ger (Ger-Fra)	ENSMUSG00000024650
19:8948514..8948544	-2.28	Cam (Fra-Cam)	ENSMUSG00000071655
19:9030248..9030289	2.33	Cam (Iran-Cam)	ENSMUSG00000071646
19:9176262..9176279	-2.37	Kazakh (Kazakh-CR)	ENSMUSG00000024653
19:9507834..9507847	2.11	Iran (Ger-Iran)	ENSMUSG00000067608
19:10119956..10119977	-2.27	Ger (Ger-Iran)	ENSMUSG00000024663
19:10119956..10119977	2.03	Cam (Iran-Cam)	ENSMUSG00000024663
19:10144285..10144301	-2.75	Ger (Ger-Iran)	ENSMUSG00000079879
19:10447252..10447284	-2.42	Kazakh (Kazakh-CR)	ENSMUSG00000035735
19:10447252..10447284	2.17	Cam (Ger-Cam)	ENSMUSG00000035735
19:10893668..10893699	-2.80	Kazakh (Kazakh-CR)	ENSMUSG00000024670
19:10893668..10893699	2.41	Fra (Iran-Fra)	ENSMUSG00000024670
19:10910545..10910579	-2.39	Ger (Ger-Iran)	ENSMUSG00000024670
19:10965776..10965791	-2.49	Ger (Ger-Fra)	ENSMUSG00000034659
19:10965776..10965791	-3.34	Iran (Iran-Fra)	ENSMUSG00000034659
19:11257475..11257499	2.63	Cam (Iran-Cam)	ENSMUSG00000024728
19:11257475..11257499	-1.97	Cam (Fra-Cam)	ENSMUSG00000024728
19:11806477..11806491	-2.20	Iran (Iran-Cam)	ENSMUSG00000055933
19:12596772..12596791	1.96	CR (Kazakh-CR)	ENSMUSG00000039982
19:12844382..12844399	-2.32	Ger (Ger-Fra)	ENSMUSG00000084981
19:13629436..13629453	1.99	Fra (Ger-Fra)	ENSMUSG00000063485

19:13629436..13629453	2.76	Fra (Fra-Cam)	ENSMUSG00000063485
19:13788551..13788569	-3.06	Ger (Ger-Cam)	ENSMUSG00000051156
19:13788551..13788569	2.16	Fra (Fra-Cam)	ENSMUSG00000051156
19:15007029..15007043	2.00	Cam (Ger-Cam)	ENSMUSG00000084675
19:15007029..15007043	2.21	Fra (Ger-Fra)	ENSMUSG00000084675
19:15056726..15056742	2.03	Fra (Ger-Fra)	ENSMUSG00000063586
19:15101188..15101213	-2.60	Kazakh (Kazakh-CR)	ENSMUSG00000063586
19:15339761..15339783	-2.95	Cam (Fra-Cam)	ENSMUSG00000063586
19:15997482..15997498	-2.34	Iran (Iran-Fra)	ENSMUSG00000024640
19:16179215..16179236	2.36	Fra (Ger-Fra)	ENSMUSG00000049247
19:16179215..16179236	2.46	Fra (Iran-Fra)	ENSMUSG00000049247
19:16644216..16644230	2.01	Cam (Iran-Cam)	ENSMUSG00000024697
19:16644216..16644230	-2.64	Cam (Fra-Cam)	ENSMUSG00000024697
19:17558568..17558595	-2.83	Ger (Ger-Cam)	ENSMUSG00000024712
19:17558568..17558595	1.99	Fra (Iran-Fra)	ENSMUSG00000024712
19:17558568..17558595	3.26	Fra (Fra-Cam)	ENSMUSG00000024712
19:17620193..17620271	-2.19	Kazakh (Kazakh-CR)	ENSMUSG00000024712
19:17764383..17764422	1.97	CR (Kazakh-CR)	ENSMUSG00000024713
19:17764383..17764422	2.31	Cam (Iran-Cam)	ENSMUSG00000024713
19:18296190..18296211	1.97	Cam (Iran-Cam)	ENSMUSG00000076303
19:18296190..18296211	2.51	Fra (Ger-Fra)	ENSMUSG00000076303
19:18296190..18296211	2.66	Fra (Iran-Fra)	ENSMUSG00000076303
19:18478514..18478545	-2.41	Ger (Ger-Iran)	ENSMUSG00000076303
19:18572618..18572633	2.21	Cam (Ger-Cam)	ENSMUSG00000076303
19:18572618..18572633	2.55	Cam (Iran-Cam)	ENSMUSG00000076303
19:18572618..18572633	2.04	Fra (Iran-Fra)	ENSMUSG00000076303
19:18790052..18790068	1.99	Iran (Ger-Iran)	ENSMUSG00000024726
19:18892241..18892284	-2.24	Ger (Ger-Iran)	ENSMUSG00000064941
19:18892241..18892284	2.76	Cam (Iran-Cam)	ENSMUSG00000064941
19:18892241..18892284	2.10	Fra (Iran-Fra)	ENSMUSG00000064941
19:18944120..18944137	1.98	Iran (Ger-Iran)	ENSMUSG00000024727
19:19438541..19438555	-2.04	Ger (Ger-Iran)	ENSMUSG00000036192
19:19685266..19685290	-2.05	Ger (Ger-Cam)	ENSMUSG00000036192
19:20092550..20092595	2.89	CR (Kazakh-CR)	ENSMUSG00000036192
19:20092550..20092595	1.98	Fra (Iran-Fra)	ENSMUSG00000036192
19:21270931..21270968	2.04	Fra (Iran-Fra)	ENSMUSG00000024749
19:21487272..21487319	2.24	Fra (Fra-Cam)	ENSMUSG00000024750
19:21787913..21787933	-2.67	Ger (Ger-Fra)	ENSMUSG00000047368
19:22519965..22519986	-2.19	Ger (Ger-Iran)	ENSMUSG00000045104
19:22519965..22519986	-2.34	Ger (Ger-Fra)	ENSMUSG00000045104
19:22597753..22597789	-2.76	Kazakh (Kazakh-CR)	ENSMUSG00000045104
19:22597753..22597789	2.99	Cam (Ger-Cam)	ENSMUSG00000045104
19:22597753..22597789	2.63	Fra (Ger-Fra)	ENSMUSG00000045104
19:22807141..22807178	2.26	CR (Kazakh-CR)	ENSMUSG00000052387
19:22952262..22952287	2.41	CR (Kazakh-CR)	ENSMUSG00000065507
19:23130121..23130172	-2.41	Iran (Iran-Fra)	ENSMUSG00000087169
19:23282473..23282501	1.97	Fra (Iran-Fra)	ENSMUSG00000033863
19:23383740..23383759	2.36	Cam (Ger-Cam)	ENSMUSG00000077247
19:23976316..23976332	-2.19	Ger (Ger-Iran)	ENSMUSG00000024897
19:23976316..23976332	3.35	Fra (Iran-Fra)	ENSMUSG00000024897
19:24008755..24008786	2.27	Fra (Fra-Cam)	ENSMUSG00000024897
19:24200352..24200381	-2.17	Iran (Iran-Fra)	ENSMUSG00000024819

19:24453964..24453998	3.75	Fra (Ger-Fra)	ENSMUSG00000082107
19:24453964..24453998	2.79	Fra (Fra-Cam)	ENSMUSG00000082107
19:25176838..25176855	-2.29	Ger (Ger-Iran)	ENSMUSG00000052085
19:25176838..25176855	-1.97	Ger (Ger-Cam)	ENSMUSG00000052085
19:25634125..25634153	2.29	CR (Kazakh-CR)	ENSMUSG00000024837
19:25634125..25634153	2.12	Fra (Fra-Cam)	ENSMUSG00000024837
19:25693559..25693584	-2.41	Ger (Ger-Iran)	ENSMUSG00000042372
19:26371473..26371500	-2.05	Ger (Ger-Iran)	ENSMUSG00000084589
19:26596252..26596285	2.05	Cam (Ger-Cam)	ENSMUSG00000084589
19:26596252..26596285	2.08	Fra (Ger-Fra)	ENSMUSG00000084589
19:26907490..26907521	2.53	Cam (Ger-Cam)	ENSMUSG00000024921
19:26907490..26907521	2.02	Cam (Iran-Cam)	ENSMUSG00000024921
19:26907490..26907521	2.85	Fra (Ger-Fra)	ENSMUSG00000024921
19:26907490..26907521	2.47	Fra (Iran-Fra)	ENSMUSG00000024921
19:27091192..27091227	2.17	Cam (Ger-Cam)	ENSMUSG00000074913
19:27185716..27185748	-1.98	Cam (Fra-Cam)	ENSMUSG00000074913
19:27223402..27223433	3.02	Iran (Ger-Iran)	ENSMUSG00000074913
19:27223402..27223433	-2.70	Iran (Iran-Fra)	ENSMUSG00000074913
19:27475963..27475993	2.31	Iran (Ger-Iran)	ENSMUSG00000047298
19:27536846..27536869	-2.69	Kazakh (Kazakh-CR)	ENSMUSG00000032546
19:27672780..27672843	2.05	Cam (Iran-Cam)	ENSMUSG00000032546
19:28244095..28244115	2.71	Cam (Ger-Cam)	ENSMUSG00000040929
19:28244095..28244115	-3.11	Cam (Fra-Cam)	ENSMUSG00000040929
19:28655614..28655654	-2.19	Cam (Fra-Cam)	ENSMUSG00000052942
19:28692083..28692121	1.99	CR (Kazakh-CR)	ENSMUSG00000052942
19:28759203..28759220	2.71	Iran (Ger-Iran)	ENSMUSG00000052942
19:28759203..28759220	-1.98	Iran (Iran-Cam)	ENSMUSG00000052942
19:28972814..28972832	-2.23	Cam (Fra-Cam)	ENSMUSG00000024935
19:29029707..29029742	-2.25	Ger (Ger-Cam)	ENSMUSG00000064202
19:29167067..29167113	2.48	Cam (Ger-Cam)	ENSMUSG00000063754
19:29167067..29167113	-2.01	Cam (Fra-Cam)	ENSMUSG00000063754
19:29220167..29220193	2.82	Cam (Iran-Cam)	ENSMUSG00000065556
19:29220167..29220193	-2.09	Cam (Fra-Cam)	ENSMUSG00000065556
19:29269485..29269530	-2.15	Ger (Ger-Iran)	ENSMUSG00000066530
19:29340520..29340540	3.03	CR (Kazakh-CR)	ENSMUSG00000066530
19:29340520..29340540	1.99	Cam (Ger-Cam)	ENSMUSG00000066530
19:30291717..30291753	-2.80	Ger (Ger-Cam)	ENSMUSG00000058607
19:30291717..30291753	-2.18	Ger (Ger-Fra)	ENSMUSG00000058607
19:30333128..30333141	2.36	Iran (Ger-Iran)	ENSMUSG00000024863
19:30333128..30333141	-2.86	Iran (Iran-Fra)	ENSMUSG00000024863
19:30693390..30693412	2.46	Iran (Ger-Iran)	ENSMUSG00000024868
19:30693390..30693412	-3.19	Iran (Iran-Cam)	ENSMUSG00000024868
19:30899742..30899764	-2.02	Iran (Iran-Cam)	ENSMUSG00000084432
19:30899742..30899764	-2.68	Iran (Iran-Fra)	ENSMUSG00000084432
19:31084040..31084076	2.11	Iran (Ger-Iran)	ENSMUSG00000084432
19:31555794..31555811	2.35	Cam (Ger-Cam)	ENSMUSG00000052920
19:31958826..31958855	2.18	Fra (Iran-Fra)	ENSMUSG00000052920
19:32208811..32208845	2.15	Cam (Ger-Cam)	ENSMUSG00000024887
19:32208811..32208845	2.81	Cam (Iran-Cam)	ENSMUSG00000024887
19:32208811..32208845	2.26	Fra (Iran-Fra)	ENSMUSG00000024887
19:32612845..32612860	-1.97	Cam (Fra-Cam)	ENSMUSG00000024896
19:32749084..32749119	-2.79	Ger (Ger-Iran)	ENSMUSG00000024899

19:32749084..32749119	-1.97	Ger (Ger-Fra)	ENSMUSG00000024899
19:33597118..33597131	2.03	Cam (Ger-Cam)	ENSMUSG00000079344
19:33655943..33655966	2.02	Iran (Ger-Iran)	ENSMUSG00000079344
19:34056943..34056956	2.17	Fra (Fra-Cam)	ENSMUSG00000024768
19:34306930..34306948	-2.13	Ger (Ger-Fra)	ENSMUSG00000024776
19:34595205..34595224	1.97	Fra (Fra-Cam)	ENSMUSG00000024781
19:34806687..34806717	-2.17	Ger (Ger-Iran)	ENSMUSG00000009378
19:35391608..35391628	2.06	CR (Kazakh-CR)	ENSMUSG00000024795
19:35656320..35656360	2.42	Iran (Ger-Iran)	ENSMUSG00000024795
19:35656320..35656360	2.32	Cam (Ger-Cam)	ENSMUSG00000024795
19:36312893..36312929	2.72	Cam (Ger-Cam)	ENSMUSG00000024803
19:36426299..36426329	2.13	CR (Kazakh-CR)	ENSMUSG00000024803
19:37270299..37270329	-2.29	Ger (Ger-Iran)	ENSMUSG00000085432
19:37454170..37454207	-2.46	Ger (Ger-Iran)	ENSMUSG00000056999
19:37454170..37454207	-2.48	Ger (Ger-Fra)	ENSMUSG00000056999
19:37705557..37705571	2.28	Fra (Fra-Cam)	ENSMUSG00000053799
19:38111231..38111277	-1.97	Ger (Ger-Fra)	ENSMUSG00000048612
19:38665376..38665400	-2.17	Ger (Ger-Fra)	ENSMUSG00000044026
19:38918356..38918380	-1.97	Cam (Fra-Cam)	ENSMUSG00000024999
19:39007010..39007026	-2.51	Ger (Ger-Fra)	ENSMUSG00000048720
19:39007010..39007026	-2.78	Iran (Iran-Fra)	ENSMUSG00000048720
19:39007010..39007026	-2.21	Cam (Fra-Cam)	ENSMUSG00000048720
19:40313456..40313485	3.63	CR (Kazakh-CR)	ENSMUSG00000067224
19:40408273..40408289	-1.99	Ger (Ger-Cam)	ENSMUSG00000055044
19:40443022..40443046	-2.26	Ger (Ger-Iran)	ENSMUSG00000055044
19:40443022..40443046	-2.33	Ger (Ger-Fra)	ENSMUSG00000055044
19:40564118..40564138	-2.99	Ger (Ger-Cam)	ENSMUSG00000025006
19:40564118..40564138	2.09	Fra (Fra-Cam)	ENSMUSG00000025006
19:40868525..40868545	-2.22	Ger (Ger-Iran)	ENSMUSG00000048120
19:40868525..40868545	2.55	Cam (Iran-Cam)	ENSMUSG00000048120
19:40868525..40868545	3.01	Fra (Iran-Fra)	ENSMUSG00000048120
19:40924849..40924876	2.23	CR (Kazakh-CR)	ENSMUSG00000025010
19:40959877..40959891	2.06	Fra (Fra-Cam)	ENSMUSG00000087183
19:41180794..41180810	-2.32	Kazakh (Kazakh-CR)	ENSMUSG00000050121
19:41253735..41253749	-3.08	Ger (Ger-Iran)	ENSMUSG00000025013
19:41253735..41253749	-3.68	Ger (Ger-Fra)	ENSMUSG00000025013
19:41253735..41253749	-2.27	Cam (Fra-Cam)	ENSMUSG00000025013
19:41409278..41409318	-2.45	Ger (Ger-Iran)	ENSMUSG00000025017
19:41527152..41527167	2.55	Iran (Ger-Iran)	ENSMUSG00000025017
19:41527152..41527167	-2.59	Iran (Iran-Cam)	ENSMUSG00000025017
19:41527152..41527167	3.49	Fra (Ger-Fra)	ENSMUSG00000025017
19:41527152..41527167	3.39	Fra (Fra-Cam)	ENSMUSG00000025017
19:41625550..41625566	1.98	Fra (Iran-Fra)	ENSMUSG00000025019
19:41625550..41625566	2.50	Fra (Fra-Cam)	ENSMUSG00000025019
19:41738428..41738441	-2.67	Iran (Iran-Fra)	ENSMUSG00000074873
19:42771211..42771239	-2.12	Ger (Ger-Iran)	ENSMUSG00000025185
19:42771211..42771239	2.19	Cam (Iran-Cam)	ENSMUSG00000025185
19:42841216..42841238	-2.38	Kazakh (Kazakh-CR)	ENSMUSG00000060224
19:43055584..43055616	3.52	Iran (Ger-Iran)	ENSMUSG00000025188
19:43055584..43055616	-2.47	Iran (Iran-Cam)	ENSMUSG00000025188
19:43055584..43055616	-2.43	Iran (Iran-Fra)	ENSMUSG00000025188
19:43435016..43435032	1.96	Cam (Iran-Cam)	ENSMUSG00000047509

19:44317939..44317974	-2.20	Kazakh (Kazakh-CR)	ENSMUSG00000025202
19:44580671..44580717	2.46	Cam (Iran-Cam)	ENSMUSG00000036961
19:44954849..44954867	-2.51	Kazakh (Kazakh-CR)	ENSMUSG00000058350
19:45020465..45020481	-1.97	Iran (Iran-Fra)	ENSMUSG00000065159
19:45068522..45068541	-2.24	Ger (Ger-Fra)	ENSMUSG00000085356
19:45320017..45320040	2.37	Cam (Iran-Cam)	ENSMUSG00000025216
19:45494409..45494439	3.67	Fra (Ger-Fra)	ENSMUSG00000025216
19:45494409..45494439	3.35	Fra (Iran-Fra)	ENSMUSG00000025216
19:45494409..45494439	2.45	Fra (Fra-Cam)	ENSMUSG00000025216
19:45918152..45918187	2.24	Fra (Fra-Cam)	ENSMUSG00000025221
19:46186660..46186687	-3.87	Kazakh (Kazakh-CR)	ENSMUSG00000015176
19:46283984..46284022	2.46	CR (Kazakh-CR)	ENSMUSG00000025229
19:46746176..46746220	2.79	Cam (Ger-Cam)	ENSMUSG00000003555
19:46746176..46746220	-2.48	Cam (Fra-Cam)	ENSMUSG00000003555
19:46789603..46789623	-2.04	Kazakh (Kazakh-CR)	ENSMUSG00000062376
19:46812959..46812999	-2.98	Cam (Fra-Cam)	ENSMUSG00000003559
19:47097417..47097431	-2.90	Ger (Ger-Cam)	ENSMUSG00000034336
19:47097417..47097431	-2.22	Ger (Ger-Fra)	ENSMUSG00000034336
19:47097417..47097431	-3.06	Iran (Iran-Cam)	ENSMUSG00000034336
19:47097417..47097431	-2.54	Iran (Iran-Fra)	ENSMUSG00000034336
19:47136578..47136613	-2.09	Kazakh (Kazakh-CR)	ENSMUSG00000064421
19:47254575..47254605	-2.39	Ger (Ger-Fra)	ENSMUSG00000079258
19:47254575..47254605	-2.20	Iran (Iran-Fra)	ENSMUSG00000079258
19:47261704..47261741	-2.11	Ger (Ger-Cam)	ENSMUSG00000079258
19:47351706..47351740	-2.45	Ger (Ger-Cam)	ENSMUSG00000078104
19:47825160..47825174	-2.15	Ger (Ger-Fra)	ENSMUSG00000044948
19:47956748..47956777	-2.08	Kazakh (Kazakh-CR)	ENSMUSG00000025069
19:48336800..48336820	4.69	CR (Kazakh-CR)	ENSMUSG00000046585
19:49050697..49050710	2.68	Iran (Ger-Iran)	ENSMUSG00000063434
19:49081222..49081271	-1.99	Cam (Fra-Cam)	ENSMUSG00000063434
19:49296181..49296199	-2.08	Ger (Ger-Fra)	ENSMUSG00000063434
19:49456672..49456715	2.16	CR (Kazakh-CR)	ENSMUSG00000063434
19:50175175..50175191	1.97	Cam (Ger-Cam)	ENSMUSG00000062083
19:50480042..50480079	-2.11	Kazakh (Kazakh-CR)	ENSMUSG00000062083
19:51203263..51203285	2.79	Fra (Fra-Cam)	ENSMUSG00000086005
19:51363698..51363745	2.09	Cam (Ger-Cam)	ENSMUSG00000065143
19:51363698..51363745	-2.86	Cam (Fra-Cam)	ENSMUSG00000065143
19:51963980..51964019	2.14	Cam (Iran-Cam)	ENSMUSG00000065143
19:51963980..51964019	2.94	Fra (Iran-Fra)	ENSMUSG00000065143
19:52322158..52322188	3.01	Cam (Ger-Cam)	ENSMUSG00000065143
19:52322158..52322188	2.56	Cam (Iran-Cam)	ENSMUSG00000065143
19:52322158..52322188	-2.37	Cam (Fra-Cam)	ENSMUSG00000065143
19:52580038..52580073	2.32	Cam (Ger-Cam)	ENSMUSG00000035804
19:52580038..52580073	2.92	Cam (Iran-Cam)	ENSMUSG00000035804
19:52580038..52580073	2.37	Fra (Iran-Fra)	ENSMUSG00000035804
19:53078741..53078768	2.16	Iran (Ger-Iran)	ENSMUSG00000035804
19:53078741..53078768	2.89	Fra (Ger-Fra)	ENSMUSG00000035804
19:53078741..53078768	2.47	Fra (Fra-Cam)	ENSMUSG00000035804
19:53670729..53670755	2.18	Iran (Ger-Iran)	ENSMUSG00000071497
19:53670729..53670755	-2.03	Iran (Iran-Cam)	ENSMUSG00000071497
19:53690223..53690238	3.17	CR (Kazakh-CR)	ENSMUSG00000071497
19:53954576..53954589	2.38	Fra (Fra-Cam)	ENSMUSG00000043639

19:54108514..54108536	2.59	Iran (Ger-Iran)	ENSMUSG00000024976
19:54108514..54108536	2.69	Fra (Ger-Fra)	ENSMUSG00000024976
19:54360514..54360534	-2.33	Ger (Ger-Iran)	ENSMUSG00000033717
19:55724966..55724996	2.63	CR (Kazakh-CR)	ENSMUSG00000024983
19:56418047..56418069	-2.54	Iran (Iran-Cam)	ENSMUSG00000025075
19:56557969..56558002	2.88	Cam (Ger-Cam)	ENSMUSG00000035818
19:56557969..56558002	-2.82	Cam (Fra-Cam)	ENSMUSG00000035818
19:57094758..57094801	-1.96	Ger (Ger-Iran)	ENSMUSG00000025083
19:57094758..57094801	2.50	Cam (Iran-Cam)	ENSMUSG00000025083
19:57217008..57217034	2.33	Fra (Iran-Fra)	ENSMUSG00000078103
19:57280217..57280239	2.08	Iran (Ger-Iran)	ENSMUSG00000025085
19:57280217..57280239	-2.47	Iran (Iran-Fra)	ENSMUSG00000025085
19:57335240..57335269	-2.02	Kazakh (Kazakh-CR)	ENSMUSG00000025085
19:57428398..57428414	2.15	Cam (Iran-Cam)	ENSMUSG00000085096
19:57874453..57874484	3.39	CR (Kazakh-CR)	ENSMUSG00000025086
19:57926214..57926243	-2.82	Ger (Ger-Iran)	ENSMUSG00000025086
19:57926214..57926243	3.24	Cam (Iran-Cam)	ENSMUSG00000025086
19:58211993..58212028	2.27	Fra (Iran-Fra)	ENSMUSG00000054843
19:58506274..58506305	2.26	CR (Kazakh-CR)	ENSMUSG00000025089
19:59114746..59114762	2.30	CR (Kazakh-CR)	ENSMUSG00000041362
19:60421766..60421786	-2.05	Ger (Ger-Iran)	ENSMUSG00000043623
19:60421766..60421786	2.77	Cam (Iran-Cam)	ENSMUSG00000043623
19:60727785..60727817	3.38	Fra (Ger-Fra)	ENSMUSG00000033417
19:60727785..60727817	2.92	Fra (Iran-Fra)	ENSMUSG00000033417
19:60727785..60727817	3.35	Fra (Fra-Cam)	ENSMUSG00000033417
19:60887718..60887746	-2.29	Ger (Ger-Iran)	ENSMUSG00000074740

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.

Plön, den 06. Juni 2010 Anna Büntge