Duplicated genes and their relevance in the process of speciation

I n a u g u r a l - D i s s e r t a t i o n

Zur
Erlangung des Doktorgrades
der Mathemathsch-Naturwissenschaftlichen Fakultät
der Universität zu Köln
vorgelegt von

Till Bayer

aus Köln

Köln, **2009**

Berichterstatter:	Dr. Bettina Harr
	Prof. Dr. Diethard Tautz
Tag der mündlichen Prüfung:	23.6.2008

Table of contents

Danksagung

Zunächst einmal möchte ich Bettina Harr danken, dafür dass sie mir die Möglichkeit gegeben hat in ihrere Gruppe zu arbeiten und zu promovieren. Dank geht auch an Diethard Tautz, nicht nur für die Erstellung des Zweitgutachtens.

Während meiner Arbeit in den AGs Harr und Tautz habe ich unter den Kollegen viele Freunde gefunden. Manuel Aranda hat mir schon in der Diplomarbeit grundlegende Labortechniken beigebracht. Er und Chris Voolstra, mit dem ich lange ein Labor geteilt habe und auf dessen hochoptimierte Protokolle ich zurückgreifen konnte, sind mir gute Freunde geworden. Es hat auch Spass gemacht mit Ruth Rottscheidt zusammen zu arbeiten, auf die ich mich im Laboralltag immer verlassen konnte.

Viel von dem was ich über Mikrosatelliten weiss habe ich im gemeinsamen Kampf gegen den MegaBACE von Arne Nolte gelernt. Sowohl im Labor wie auch ausserhalb habe ich immer gerne Zeit mit Henrique Souza, Kathryn Stemshorn, Fabian Staubach, Tobias Heinen, Meike Teschke und Suma Choorapoikayil verbracht, ob bei Diskussionen der Probleme der Molekularbiologie, oder beim Klettern, Skifahren, Badminton spielen und Kuchen essen.
Auch Jochen Wolf hat den Laboralltag mit seinen Kurzbesuchen in Köln immer aufgeheitert, und war immer für interessante Diskussionen zu haben.

Nach der Zeit des Zusammenschreibens in Plön bin ich vor allem zwei Kolleginnen zum Dank verpflichtet: Leslie Turner, fürs Korrekturlesen und dafür dass sie mich in die Feinheiten der PAML Analyse eingeweiht hat, und Anna Büntge, für Unterstützung moralischer, logistischer und kulinarischer Art.

Gemeinsam haben alle Kollegen und Freunde der AGs Harr, Tautz und Damen die Zeit in der Genetik interessant, unterhaltsam, und erfolgreich gestaltet.

Last but not least danke ich auch meinen Eltern die mich vom ersten bis zum 19. Semester meiner Unilaufbahn immer aufs Beste unterstützt haben, und meinem Bruder Jan, der immer Interesse am Fortgang der Doktorarbeit gezeigt hat.

Declaration

Bettina Harr developed the ideas for the projects presented here. All laboratory work and data analysis was performed by me, with some exceptions:

Chapter 1

Birgit Schmitz worked on a part of the microsatellite PCRs.

Chapter 3

Leslie Turner helped with the PAML analysis of *Dnahc8*.

Zusammenfassung

Ein integraler Bestanteil der Evolution, die Entstehung von Arten, ist nicht im gleichen Mass verstanden wie andere Bereiche der Evolutionsbiologie. Da Speziation ein fundamentaler Prozess ist durch den die grosse Artenvielfalt der Erde ensteht, wäre es sehr wünschenswert ein besseres Verständnis dieses Prozesses zu gewinnen. Um dies zu erreichen müssen die molekularen Grundlagen des Speziationsprozesses im Detail erforscht werden.

Diese Arbeit verfolgt zwei Ansätze um zum Verständnis des Speziazionsprozesses beizutragen: Ein Screen nach positiv selektierten, jungen Duplikaten, und eine tiefergehenede Analyse des potentiellen Speziationsgens *Dnahc8*.

Junge Duplikate können, wenn sie in nur einer von zwei divergierenden Population unter positiver Selektion stehen, durch schnelle evolutionäre Entwicklung eine Inkompatiblität zwischen den beiden enstehenden Spezies erzeugen. Um Kandidatengene mit dieser Eigenschaft zu finden wurde eine Suche nach Mikrosatelliten durchgeführt die Spuren eines Selektionsereignisses aufweisen, einem sogenannten ‚selective sweep'. Dazu wurden Mikrosatelliten in der Nähe der Gene typisiert und auf reduzierte Variabilität untersucht. Mittels einer für diese Frage entwickelten Methode, der *lnRH* Statistik, wurden selective sweep Loci im Vergleich zwischen Spezies oder Populationen identifiziert. Das Ergebnis des Screens sind 13 Kandidatenloci im (Sub)species Vergleich, und 15 im Vergleich zwischen den Populationen. Ein Vergleich zwischen den ermittelten *lnRH* Werten und der synonymen Substitutionsrate (K_S) zeigt dass die jüngsten der duplizierten Gene nicht unter positiver Selektion zu stehen scheinen.

Um zu testen ob das tatsächliche Duplikationsalter mit dem Alter welches auf der Basis der synonymen Subtitutionsraten angenommen wird korreliert, wurde für einen Teil der Duplikatpaare überprüft ob sie in verschiedenen Mausspezies vorhanden sind. In den meisten Fällen nimmt der K_S Wert mit dem tatsächlichen Alter der Duplikation zu, drei Genpaare haben jedoch sehr geringe K_S Werte die nicht zu ihrem Alter passen. Genkonversion könnte ein Grund für dieses Ergebnis sein.

Es ist bekannt dass das *Dnahc8* Gen *Mus spretus* und *Mus mus domesticus* reproduktiv voneinander isoliert. Es kodiert für ein axonemales Dyein Protein das am

Aufbau des Spermienflagellums beteiligt ist; demenstprechend haben hybride Tiere deformierte, immobilisiert Spermien. Für diese Arbeit wurde die gesamte kodierende Sequenz von *Dnahc8* in sechs Mausspezies sequenziert, um die schon bekannte Sequenz aus *M. m. domesticus* zu ergänzen. Ausserdem wurden mehrere Exons für ein Populationssample von *M. spretus*, und das gesamte Gen von zehn *M. m. domesticus* Tieren sequeziert. Die Daten wurden mit verschiedenen Tests auf positive Selektion untersucht: Eine Art von Test basiert auf Polymorphismus Daten, die zweite auf der grössten Wahrscheinlichkeit für kodonbasierte Modelle die verschiedene Arten von Selekion erlauben. Es wurden Hinweise gefunden dass der Selektionsdruck auf *Dnahc8* in der *M. m. domesticus* Linie relaxiert ist, es konnte jedoch mit keiner Methode positive Selektion nachgewiesen werden.

Neben den Sequenzdaten wurde auch die Expression mittels quantitativer Echzeit PCR in acht Geweben von sieben Mausarten gemessen. Starke Unterschiede in der Expression zwischen *M. spretus* und *M. m. domesticus* wurden gefunden, *Dnahc8* wird in *M. spretus* achtfach niedriger exprimiert. Zusammengenommen weisen diese Resultate darauf hin dass der Grund für die Inkompatibilität in Hybriden eher in der Genregulation zu suchen ist als in der Aminosäuresequenz.

Abstract

An integral part of evolution, the formation of species, is less well understood than other areas of evolutionary biology. Speciation is a fundamental process that creates the great diversity of species in the world, and a deeper insight into its mechanisms is highly desirable. To achieve this goal, the molecular basis of speciation must be elucidated and characterized in detail.

This study uses two approaches to contribute to the understanding of the genetics of speciation: A screen for positively selected, young duplicated genes, and in depth analysis of a proposed 'speciation gene', *Dnahc8*.

Young duplicated genes may be positively selected in only one of two diverging populations, and through rapid change create an incompatibility between the two emerging species. To find candidates of this type a microsatellite screen for selective sweeps is conducted, in which microsatellite loci close to the genes in question are typed and assayed for reduced variability. Using a measure developed for this problem, the *lnRH* statistic, subspecies or populations are compared, and selective sweep loci identified. The screen results in thirteen candidate sweep loci in (sub)species comparisons, and fifteen between populations of the same species. Furthermore, comparisons of *lnRH* values to synonymous substitution rates (K_S) of genes show that the youngest duplicated genes of the set do not seem to be evolving under positive selection.

To test whether the duplication time correlates with the divergence time estimated from the synonymous substitutions rate, a subset of duplicate pairs was tested for presence in different mouse species. For most duplicates, K_S between copies increases with age, but three pairs have very low K_S values that do not correspond to their age. Gene conversion is discussed as a possible explanation for this result.

The *Dnahc8* gene is already known to cause reproductive isolation between *Mus spretus* and *Mus mus domesticus*. It encodes an axonemal dynein protein that is involved in sperm tail formation, and hybrid animals have deformed, immotile sperm. Here, the entire coding sequence of *Dnahc8* is determined of six mouse species in addition to the *M. m. domesticus* sequence already known. In addition, several exons are sequenced in a population sample of *M. spretus* and the full-length gene sequenced for ten *M. m. domesticus*. Tests for positive selection based on

polymorphism data and codon-based maximum likelihood methods are performed with this data. There is evidence that *Dnahc8* may be evolving under relaxed constraint in the lineage of *M. m. domesticus*. However, no significant evidence for positive selection could be found using any method. In addition to the sequence data, quantitative real time PCR is used to measure the level of expression in eight tissues of seven mouse species. Large differences in expression pattern are identified between *M. spretus* and *M. m. domesticus*: *Dnahc8* expression is eight fold lower in *M. spretus* testis compared to *M. m. domesticus*. Together these results suggest that the nature of the incompatibility caused by *Dnahc8* may lie in gene regulation rather than differences in the amino acid sequence.

1 Introduction

Most of the areas of the study of evolution have been developed to a great extent and in detail since Darwin laid the foundations for this research (Darwin 1859). This is not as true for one process in evolution that was and is a major mechanism in shaping the plant and animal world as we see it now: speciation. The character of the forces that drive one more or less homogenous group of organisms to split into two or more distinct groups have long been disputed, as the details of speciation events are not clear. As it is a very slow process, it can usually only be scrutinized in retrospect, and the very nature of species is often that they do not hybridize, making the usual genetic approaches to study molecular processes very difficult.

There are also some theoretical problems: Even the term species can be defined in many different ways: based on phylogeny, morphology or reproductive status (Coyne, Orr et al. 1988). The most used and accepted of these definitions is the biological species concept, which describes species as “groups of actually or potentially interbreeding natural populations which are reproductively isolated from other such groups” (Mayr 1942). This focus on reproductive isolation as the most important and useful defining aspect of species has turned out to be true especially in the advent of population genetics.

Reproductive isolation can be caused by either prezygotic or postzygotic effects. The former includes behavioral (e.g. mate recognition), mechanical (e.g. incompatible mating organs), and gametic (e.g. incompatibilities between egg and sperm) factors. Postzygotic isolation includes hybrid inviability and hybrid sterility. While postzygotic isolation in plants is often caused by polyploidy (Masterson 1994), in animals genic effects seem to play an more important role (Coyne and Orr 1998). Thus far, research on the genetics of speciation has focused on postzygotic isolation.

While many advances have been made, there are still major questions that are debated when it comes to speciation. Many of these questions concern the importance of alternative mechanisms that probably both play a role, but to what extent is not known. For one, there is a lot of ongoing discussion whether mutations in cis or regulatory changes in trans contribute to evolution and speciation (Hoekstra and Coyne 2007).

There are quite a few studied examples of gene duplications that are involved in speciation, especially the 'speciation genes', but it is still not know how important the contribution of duplicated genes to speciation is. It seems reasonable to expect that duplication processes are often involved, as hybrid incompatibility genes often evolve rapidly, and redundant genes provide this freedom by lack of selective pressure.

In this chapter I will summarize the current understanding of the genetics of speciation and examples of genes known to play a role in speciation and reproductive isolation, and the evolutionary relevance of gene duplication and the theories that have been developed. I will also review the methods of selection detection that are relevant to this work, and give an introduction to the house mouse as a model system.

1.1 The genetics of speciation

The conceptual difficulties with understanding how reproductive isolation could evolve through changes at a single locus were resolved with the independent realization by Bateson, Dobzhansky and Muller that two populations can become reproductively separated when two interacting genes are involved (Bateson and Mendel 1909; Dobzhansky 1937; Muller 1942); if novel alleles of both genes are fixed in the two populations the genes may not be able to interact normally when combined in hybrid offspring. This two locus model, termed the Dobzhansky Muller model, marked the start of the research of the genetic causes of speciation.

The key to unraveling the many questions concerning speciation lies in understanding the molecular details involved in reproductive isolation. While it takes too long to observe the speciation process directly, it is possible to identify the genetic processes that maintain species boundaries. These hybrid incompatibility genes are commonly termed speciation genes (Orr and Presgraves 2000), although it is not clear if they contributed to the initial evolution of reproductive isolation. Thus far, four hybrid incompatibility genes have been identified (Orr, Masly et al. 2004).

The first gene in this category that was identified confers hybrid sterility. Called *Odysseus site homeobox* (*OdsH*), it encodes a transcription factor with a homeobox domain which is expressed in testis tissue. Its effect was found in hybrids between the fruit fly species *Drosophila simulans* and *Drosophila mauritiana*, the hybrid males from this cross are sterile (Coyne and Charlesworth 1986; Ting, Tsaur et al. 1998). In

the sterile F1 males, *OdsH* seems to confer only to about 50% of the infertility phenotype, thus other factors must also be involved (Perez and Wu 1995; Ting, Tsaur et al. 1998). The effect also is only evident in young males, and diminishes as they get older, suggesting that *OdsH* may be involved in accelerating the maturation of sperm (Sun 2003). *OdsH* has been found to evolve rapidly under positive selection in one of the species, *D. mauritiana*, as has often been reported for testis specific genes (see (Ellegren and Parsch 2007) for review).The *OdsH* gene originated from a gene duplication of the ancestral gene *Unc-4* (Sun, Ting et al. 2004), which is expressed in the embryonic stages and in neuronal tissue. After the duplication event, the expression of *OdsH* became confined to the testis. Thus, change in this hybrid sterility gene involved a duplication in addition to rapid evolution at the amino acid sequence level.

While *OdsH* belongs to the class of the hybrid sterility genes, other genes have been found that cause inviability in hybrids. The first one analyzed in detail shows its effects in hybrid fish; progeny of a cross between *Xiphophorus maculatus* and *X. helleri* often develop fatal melanomas. The molecular cause of this phenotype has been traced to over expression of a receptor tyrosine kinase encoded on the X chromosome, *Xmrk-2* (Wittbrodt, Adam et al. 1989; Schartl, Dimitrijevic et al. 1994). Overexpression of *Xmrk-2* is the result of a Dobzhansky Muller incompatibility with another locus that represses its expression. *Xmrk-2* is a partial copy of its paralog *Xmrk-1*, and additionally acquired mutations following duplication render the protein constitutively active (Gomez, Wellbrock et al. 2001). While the protein coding part comes from *Xmrk-1*, the regulatory sequence that controls its expression originates from a second gene. This promoter interacts with a repressor, thus *Xmrk-2* has acquired a novel regulatory mechanism. The detrimental phenotype occurs in hybrids which do not have a functional allele of the repressor.

Another hybrid inviability gene, *hybrid male rescue* (*Hmr*), has been found in crosses between *Drosophila* species. *Hmr* is a transcription factor located on the X chromosome. The hybrid lethality effect of *Hmr* is confined to male hybrids, which do not live beyond the transition from larval to pupal stage: females have reduced fertility and also become sensitive to high temperatures (Barbash, Roote et al. 2000). It has been shown that the dosage of *Hmr* is correlated with the strength of the inviability phenotype (Barbash, Roote et al. 2000; Orr and Irving 2000; Barbash, Siino et al. 2003). These effects show when *D. melanogaster* is crossed with one of its

sister species *D. sechellia, D. simulans,* or *D. mauritiana*, and there is evidence that *Hmr* has diverged at the sequence level between *D. melanogaster*and the latter three species. The large extent of divergence has been attributed to positive selection (Barbash, Awadalla et al. 2004).

The fourth hybrid inviability gene discovered *nucleoporin 96* (*Nup96*), could also be shown to have a period of evolution under positive selection in its history (Presgraves and Stephan 2007). The incompatibility involving *Nup96* is again between *D. melanogaster* and *D. simulans* (Presgraves, Balagopalan et al. 2003). The protein is present in all eukaryotes from yeast to humans, and forms a part of the nuclear pore complex located in the nuclear membrane, which regulates the passage of macromolecules through the membrane.

Although very limited, this set of four speciation genes allows drawing some preliminary conclusions about the genetic process of speciation. One of the unsolved questions concerning speciation is whether incompatibilities evolve neutrally or under selection. For all genes mentioned before positive selection could be shown, either recent or in the gene's history. Additionally, all genes evolve very fast, and *Hmr* and *Nup96* show evidence of adaptive protein evolution. It is not known in what way these adaptive processes relate to the species' environmental conditions. Further, it has been postulated that functional constraint has to be relaxed in order to have such a rapid rate of sequence divergence (Wu and Ting 2004). One way this could be happen is through gene duplication. Indeed, duplication events during the evolution of *OdsH* and *Xmrk-2* have contributed to the changes that cause them to act as 'speciation genes'.

Finally, despite the fact that all the genes show sequence divergence that may explain the incompatibilities, three of four genes are also directly involved in gene expression, with two being transcription factors. In the case of *Xmrk-2* the change in expression seems to be the main contributor towards the hybrid lethality phenotype, while *OdsH* may cause hybrid infertility through regulating other genes (Michalak and Noor 2004).

This recent progress in identifying genetic causes of reproductive isolation is the first evidence for speciation mechanisms at the molecular level. Nonetheless these data are still rather limited, and more examples have to be analyzed in detail to draw conclusions. For one, more taxa should be investigated, as three of the four 'speciation genes' summarized above are found in *Drosophila*. This ongoing research

is required to confirm and expand on the preliminary conclusions that are discussed today.

1.2 Gene duplication

Gene duplication is the major mechanism that enables the increase in complexity of organisms. The importance of duplication events was first mentioned in the 1930s (Haldane 1933; Muller 1935), but general acceptance of the role of duplication in evolution came through the ideas of Ohno (Ohno 1970). In his book he laid the theoretical groundwork, and postulated that novel genes can arise through copies of those already in existence. Extensive research on the mechanisms and effects of gene duplication events since that time has provided a more comprehensive understanding of this important type of evolutionary change.

To be an important contributor to genome evolution, duplication events have to occur with some frequency. And indeed, Lynch and Conery's analysis of several genome sequences comes to the conclusion that genes are copied at a rate of about 0.01 duplications per gene per million years (Lynch and Conery 2000), or 0.009 specifically for humans (Lynch and Conery 2003).

This overall rate reflects three very different mechanisms by which genetic material is multiplied and moved around the genome contribute to that rate. Unequal crossing over during recombination leads to fragments duplicated in tandem, and the affected stretch of DNA may include one or more genes, or parts thereof. Segmental duplications are much larger, encompassing 1 kb to more than 200 kb, and do not always occur within just one chromosome (Samonte and Eichler 2002). The third and most dramatic mechanism is whole genome duplication; evidence for genome duplications is reported in a variety of taxa, among them yeast (Wolfe and Shields 1997; Kellis, Birren et al. 2004; Scannell, Byrne et al. 2006), fish (Christoffels, Koh et al. 2004; Jaillon, Aury et al. 2004) and other vertebrates (Dehal and Boore 2005).

The availability of whole genome sequences makes the large contribution of the duplication mechanisms to genome structure and organization hard to deny; for example it is estimated that over 60% of human genes came into existence through some kind of a duplication event (Li, Gu et al. 2001), and now form gene families or pairs. Segmental duplications larger than 1 kb and with high similarity (>90%) alone are thought to make up 4% of the genome (Zhang, Lu et al. 2005). Whole genome

duplication can be expected to have an even stronger effect on the genome, and indeed in yeast about 16% of the protein coding genes have a paralog (Seoighe and Wolfe 1999). But where do the duplicates of all the other genes go?

This question leads to the fairly complex theories of the fate of duplicate genes after the fact. As with mutations, the mechanisms that create duplicate genes act more or less randomly, and evolutionary forces only influence the events thereafter. They have accumulated so many mutations over time that they can no longer be distinguished from other noncoding sequence. This pattern of "nonfunctionalization" is the most common outcome because mutations that are detrimental to gene function and fitness are much more common than those with a positive effect (Lynch and Walsh 1998; Lynch and Conery 2000; Harrison, Hegyi et al. 2002).

There are three other possible fates for duplicate genes: (1) Conservation - the duplicate copy is retained, (2) Subfunctionalization – the gene copy is optimized to perform one of multiple functions of the original, and (3) neofunctionalization – the new copy acquires a completely new function. Different models proposed to explain the evolution of gene duplicates vary in terms of the frequency and importance of these three outcomes.

The conservation of two or more identical open reading frames in one genome, without any one of them acquiring deleterious mutations, can only be explained by an advantage of producing a high amount of the protein encoded (Nowak, Boerlijst et al. 1997). Examples of this include rRNA and histone genes.

Subfunctionalization entails changes to the gene in the domain of protein function or expression. The latter may be a temporal or spatial difference in expression as compared to the original gene. For this subfunctionalization to occur the original gene does not have to change in the same timeframe as the copy.

In contrast, in the duplication-degeneration-complementation (DCC) model (Force, Lynch et al. 1999), deleterious mutations in one gene copy result in selective pressure to maintain the duplicate. If these mutations occur in both genes simultaneously, but affect different subfunctions or promoters responsible for expression in different tissues, the genes are stabilized, as both become necessary for the organism. The model also applies both to cases of subfunctionalization and neofunctionalization.

The third major model of gene duplication is the innovation, amplification and divergence model (Bergthorsson, Andersson et al. 2007). In this model, duplicated

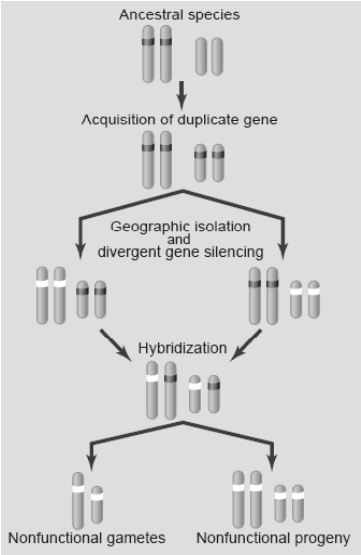


Figure 1: Divergent silencing of duplicate genes contributes to reproductive isolation (Lynch 2002).

genes acquire a new function in addition to their main purpose. If this new function is beneficial an increased amount of gene product will be favorable, providing a possibility for stabilization of the duplication similar to the case of gene conservation. The redundant copy is positively selected, and may spread through the population. Under this selection pressure it can then gain mutations that make the subfunction more efficient, and turning it from the minor side effect that it represented in the mother gene into the major function of the copy. This recent model does not suffer some problems identified with earlier models, such as the fact that neofunctionalization is unlikely to occur under the DCC model. Also, the rate of duplication retention via the 'mutation during nonfunctionality' model is very low, because of loss due to drift and null mutations.

Neofunctionalization has with good reason been thought to be much less likely than other forms of duplicate evolution (Lynch, O'Hely et al. 2001). Although less likely, there are several known examples of duplicate genes that have acquired a novel function. One example is the evolution of a RNAse A gene duplicate in the lineage of humans and old world monkeys which acquired an anti bacterial function that does not seem to require the ribonuclease activity (Rosenberg 1995; Zhang, Rosenberg et al. 1998). The authors of a recent study in *Drosophila* found a tandemly duplicated pair of transcription factors of the Polycomb group of proteins. There is evidence that strong positive selection drives neofunctionalization in one of the copies, despite the fact that they still share strong similarity (Beisswanger and Stephan 2008).

Gene duplication enables fast evolutionary changes. For example, it was recently suggested that the puzzling result of an earlier experiment, which suggests that mutation happens faster at loci under selection (Cairns, Overbaugh et al. 1988), can be better explained through gene duplication events (Hendrickson, Slechta et al. 2002). The ability to adapt rapidly may be beneficial, such as when species encounter rapidly changing environmental conditions or invade a new habitat.In addition to facilitating fast adaptation, gene duplication can also lead to reproductive isolation, and thus

speciation (Lynch and Conery 2000). One model put forth by Lynch is displayed in Figure 1 (Lynch 2002). Given the high rate of duplicate birth and death mentioned earlier, it is quite likely that a species splits into two new species after a gene has, by chance, just been copied on another chromosome. With the nonfunctionalization rate being as high as it is, the two new species may each loose the other duplicate and retain one functional copy; with a probability of 0.5. This results is a form of postzygotic isolation, as one quarter of all gametes of hybrid offspring will lack the gene function completely. As this happens with more than one gene, and genes that are of vital importance, it may lead to complete reproductive isolation, and thus separate species. In fact, a hybrid incompatibility caused by the translocation of a gene to a different chromosome has recently been identified in *Drosophila* (Masly, Jones et al. 2006). The gene duplication must not necessarily take place before isolation, geographic or otherwise, duplication can also happen and be resolved in both populations independently. This is a very fast mechanism by which populations can become reproductively isolated on a genetic level, and it is theoretically independent of outside factors such as selection pressure.

1.3 Selection

Since the introduction of the neutral model (Kimura 1983), a major focus of evolutionary studies has been the attempt to track down evidence for positive selection in the genome. This line of research is important, because positive selection is the driving force of Darwinian evolution, but succeeding in it has proven difficult, since selection has to be distinguished from neutral processes including genetic drift and demographic effects. Two major approaches are used to detect positive selection from sequence data.

The first set of methods for detecting selection relies on comparison between nonsynonymous mutations, which change the amino acid sequence of a protein and synonymous, or silent, mutations. Synonymous sites in coding sequences are assumed to evolve neutrally, and this ‘baseline’ rate of synonymous changes per synonymous site (K_S) is compared to the nonsynonymous rate (K_A). The ratio of those two values (K_A/K_S) is used to infer the type of evolutionary pressure the sequence is be under: A ratio of one implies the sequence evolves neutrally, a ratio greater than one indicates there is positive selection, and values below one signify negative or purifying

selection. As so often though the practice is much more complicated than the theory. Several different methods exist to calculate the values, incorporating various fine tuning techniques (Li, Wu et al. 1985; Nei and Gojobori 1986; Pamilo and Bianchi 1993). The interpretation of K_A/K_S ratios is not straightforward, and has been shown not to work in all scenarios (Crandall, Kelsey et al. 1999). In addition, the use of K_A/K_S estimates for entire genes to infer selection is very conservative, as selection generally targets only a subset of amino acid sites in a gene.

The most widely implemented method used to detect selection is the maximum likelihood approach implemented in the PAML package (Yang 2007). This approach involves comparison of various models that allow K_A/K_S ratios to vary among codon sites in the gene and/or among branches of a phylogenetic tree.

While the K_A/K_S ratio methods may not be easy to implement, they have the great advantage that only one representative sequence of each species in the comparison is needed. The detection of positive selection based on sequence data and synonymous and nonsynonymous changes can also be achieved by contrasting those changes within and between species (McDonald and Kreitman 1991). The McDonald-Kreitman test yields a simple table that can be assessed for significance using a Chi-square test.

The second major approach for detecting selection involves looking at the neutral variation around a gene or locus of interest. While noncoding areas of the genome are assumed to evolve neutrally, unless they are part of a promoter or enhancer, they may still be affected by positive selection events, in a manner dependent on the genomic distance to the selected locus. After a positive mutation occurs, the affected allele rises in frequency in the population, at a speed that depends on the selection coefficient. The variable of neutral sites near to the selected locus also become more common among individuals, an effect that has been termed ‘hitchhiking’ (Smith and Haigh 1974). If one can find sequence that is not under selection directly, but shows such a decreased variability in comparison to other sequences, it can be inferred that a linked site may have been under selection (Slatkin 1995). The length of this fraction of the genome with reduced variability depends on the selection coefficient and the recombination rate, the former being positively associated with the length, the latter negatively. Markers commonly used to detect sequence variability include single nucleotide polymorphisms (SNPs), and microsatellite loci. SNPs are more or less

uniformly distributed in the genome, and their mutation rate is constant (although it may differ between regions of the genome (Wolfe, Sharp et al. 1989)).

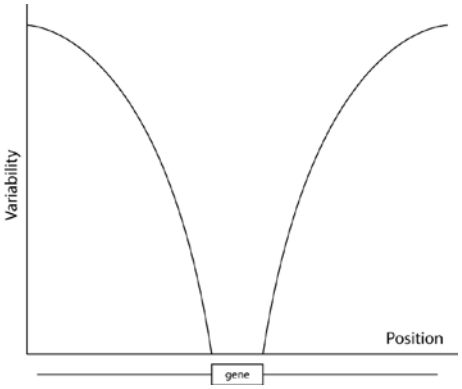


Figure 2: Schematic diagram depicting the neutral variability at a 'sweep' locus

Microsatellites have several advantages over SNPs in hitchhiking studies. The mutation rate, which is different for all individual microsatellites, is not one of them. However, it can be roughly predicted by the length of the repeat, and the microsatellites can be searched for in published genome sequences, without any population sequence data. Also, microsatellites mutate in discrete steps, according to the length of the repeated pattern (Ellegren 2004). This property makes analysis very straightforward, as only the length of a PCR fragment needs to be measured as opposed to sequencing, making it easier to study large amounts of individual animals and loci. However, as the microsatellite mutation rate depends on the type of repeat unit and total microsatellite length, different loci are not comparable. To circumvent this problem the same microsatellite locus is compared among different populations or species, thus independence from the mutation rate is gained (Schlötterer 2002).

When analyzing the data of a given microsatellite locus in multiple individuals, two key values can be calculated: The variance in repeat number (V) as a measure for the variability of the locus (Goldstein and Clark 1995), and the expected heterozygosity (Nei 1978). If has been shown that, with data from two populations, the logarithm of the ratio between the two values for either V or H (lnRV and lnRH) follows a normal distribution, if the microsatellites evolve neutrally (Schlötterer 2002; Kauer, Dieringer et al. 2003). This makes it possible to detect loci within a dataset that depart from the expected null hypothesis, neutrality, with a defined probability. These extreme loci are likely linked to sites under selection. Using the heterozygosity seems to have a higher power due to the smaller variance for this parameter (Kauer, Dieringer et al. 2003).

In this study, methods based on both of the major approaches to detecting selection are utilized: a microsatellite survey is used in a large screen for duplicated genes under selection, and K_A/K_S measurements are used to evaluate the evolutionary dynamics of a candidate gene, *Dnahc8*, in greater detail.

1.4 Mouse model

The house mouse is an excellent model organism for the study of evolution in general, and speciation in particular, for several reasons.

First, it is one of the most widespread mammals, occurring nearly everywhere on the planet. *Mus musculus* has achieved this almost ubiquitous distribution through human activity; it is a commensal rodent, living in areas inhabited by humans. The colonization history of the mouse reflects this fact; it is thought to have originated in northern India (Guenet and Bonhomme 2003), and traveled from there following agriculturally active human settlement (Cucchi, Vigne et al. 2005). The phylogeny is well established, made possible through the good fossil record and extensive morphological data (Boursot, Auffray et al. 1993; Lundrigan, Jansa et al. 2002; Guenet and Bonhomme 2003). The so called *Mus musculus* subspecies group includes 5 members (Figure 3).

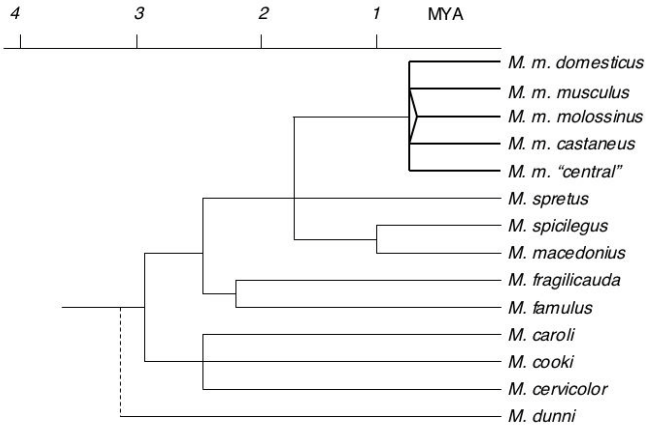


Figure 3: Phylogenetic tree for the subgenus *Mus* adapted from (Guenet and Bonhomme 2003).

M. m. castaneus colonized the Asian continent eastward and is now found in Southeast Asia. *M. m. domesticus* and *M. m. musculus* both inhabit Europe, where their ranges meet, forming a north-south hybrid zone. These two subspecies arrived in Europe by different routes, *M. m. musculus* expanded north of India to Kazakhstan and colonized eastern Europe, while *M. m. domesticus* moved westward through Iran to the western Europe and northern Africa (Boursot, Auffray et al. 1993). The *Mus musculus* subspecies group is of special interest to speciation studies, because its members are very closely related, and their invasion of the world is understood in some detail. For this study, the *Mus musculus* group is ideal because it allows for two

levels of comparison in the first part of this study: between the different subspecies, and between ancestral and derived populations within a subspecies.

In addition to the *M. musculus* group, other mouse species provide part of the data used in this study. The closest relative to the house mouse is *M. spretus*, which occurs in sympatry with it in Spain and northern Africa. *M. macedonicus*' range covers the Balkans to the Near East, while *M. macedonicus* is found in the area from eastern Austria into the Ukraine. The Indian subcontinent is the range of *M. famulus*, *M. caroli* as well as *M. cookii* are found in East Asia.

Second, several practical aspects make the house mouse a convenient model system: They have a short generation time and are easy to keep under laboratory conditions. Wild individuals are not hard to catch in their habitat. For genetic studies the availability of the whole genome sequence is greatly advantageous, facilitating the use of a wide array of bioinformatics tools. The sequenced laboratory mouse strain is a genetic mixture between the three subspecies *M. m. domesticus*, *M. m.musculus* and *M. m. castaneus*, with the first providing the largest fraction of the genome (Wade, Kulbokas et al. 2002; Wade and Daly 2005).

Finally, several loci have been found that confer hybrid incompatibility between house mouse (sub) species. These incompatibilities follow Haldane's rule, that is, sterility is observed in males (the heterogametic sex). Hybrid sterility loci (Hst) have been identified in crosses between *M. m. musculus* and *M. m. domesticus* (including wild and laboratory strains), and between *M. m. domesticus* and *M. spretus*.

Hybrid sterility 1 (*Hst1*) is responsible for an incompatibility between *M. m. musculus* and *M. m. domesticus* (in the form of the lab strain C57BL/6) and has been mapped to a region containing only two genes, which is located in the proximal region of chromosome 17, in the area defined by the t-haplotype inversions, which have been implicated in hybrid male sterility among other effects (see detailed explanation below) (Forejt, Vincek et al. 1991; Trachtulec, Mihola et al. 2005; Vyskocilova, Trachtulec et al. 2005). The candidate genes are the TATA-binding protein (*Tbp*) and proteasome subunit beta 1 (*Psmb1*), although no coding differences have been found in alleles of the strains involved (Trachtulec, Mihola et al. 2005). *Hst1* is variable in both mouse species in nature, and only some allele combinations lead to male sterility (Forejt 1996; Vyskocilova, Trachtulec et al. 2005).

Another hybrid sterility locus, named *Hstx1*, was also found in crosses of the mainly *M. m. domesticus* lab strain C57BL/6 and the *M. m. musculus* derived strain PWD/Ph.

As the name hints, the locus resides on the X chromosome. Introgressing the *M. m. musculus* allele into a *M. m. domesticus* background leads to abnormal sperm, lower sperm count, and lower testis weight in the male F1 animals. A QTL analysis showed that the *Hstx1* factor alone is not responsible, one or more linked loci of *M. m. musculus* origin must also be present (Storchova, Gregorova et al. 2004).

The other pair that has been studied for hybrid incompatibility effects, *M. m. domesticus* and *M. spretus*, is further along in the speciation process; individuals rarely hybridize in nature despite their overlap in range. All male F1 mice from this cross are sterile, but females are of normal fertility (Bonhomme, Martin et al. 1978).

One locus implicated in this incompatibility lies close to the pseudoautosomal region of the X Chromosome and has been named *Hst3* (Guenet, Nagamine et al. 1990). Closer analysis of this locus points to a possible chromosomal effect rather than a genic one, namely structural incompatibility of the X and Y chromosomes at their respective pseudoautosomal regions (Matsuda, Hirobe et al. 1991).

A group of three tightly linked loci, *Hst4, 5* and *6* have been found in the t-haplotype region of chromosome 17. These loci were found in a study that backcrossed *M. spretus* chromosome 17 into a *M. m. domesticus* background originally to survey the possible origin of the t-haplotype allele (Pilder, Hammer et al. 1991). One of these, *Hst6*, has been scrutinized in detail.

The t-haplotype is a structural variant of chomosome 17 that features four large, non-overlapping inversions which prevent recombination with the wild type alleles (Silver and Artzt 1981; Hammer, Schimenti et al. 1989). The t-haplotype does not follow mendelian inheritance, instead it shows transmission ratio distortion, where 90% of all male offspring inherit the t allele of Chromosome 17 from heterozygous +/t fathers (Hammer and Silver 1993). Nonetheless the frequency of the t allele in wild mouse populations is only about 20% (Ardlie and Silver 1996), the spread of the t-haplotype is kept at bay because homozygous males are sterile. While transmission ratio distortion and sterility are complex at the molecular level and still not fully understood, the t/t phenotype of deformed sperm seems to be the major contributor to male sterility. In particular, the motility of the sperm is affected by a flagellar waveform defect, a condition that was named “curlicue” (Olds-Clarke and Johnson 1993).

Although three of the four t-haplotype inversions are not inverted in respect to the *M. spretus* chromosome, *M. m. domesticus* mice heterozygous for the t-haplotype and

chromosome 17 from *M. spretus* are also sterile. This observation enabled a fine mapping strategy to be devised to uncover the sterility factor (Pilder, Hammer et al. 1991; Samant, Fossella et al. 1999). This research revealed that t-haplotype *M. m. domesticus* both homozygous and heterozygous for the *M. spretus* allele of *Hst6* are sterile. In both cases sperm motility was hindered by nonfunctional or even absent flagella, caused by defect assembly of flagellar structures in early stages of spermiogenesis (Phillips, Pilder et al. 1993; Pilder, Olds-Clarke et al. 1993). Further mapping narrowed the identity of *Hst6* down to one candidate gene, *Dnahc8* (Fossella, Samant et al. 2000). *Dnahc8* encodes a dynein heavy chain protein of 3480 amino acids in length, encoded in 79 exons. *Dnahc8* belongs to a large gene family present in a wide range of organisms. Its expression is testis specific in *M. m. domesticus*, but not in *M. spretus*, according to Samant et. al. (2002). Expression as well as translation take place before the process of spermatogenesis, fitting the hypothesis that it plays a part in formation of the sperm tail (Samant, Ogunkua et al. 2002).

The fact that the *M. spretus* allele of *Dnahc8* is not functional in a *M. m. domesticus* background makes it a very good candidate for a “speciation gene”. In this study, population genetic data from *M. m. domesticus* are used to investigate the evolutionary dynamics of this candidate gene in detail.

1.5 Aims of this study

The first part of this study aims to provide data on duplicated genes under selection in the mouse model system. Microsatellite loci in the vicinity of duplicates are amplified and the length allele identified in a large sample of different mouse subspecies and populations to find loci that show a signature of a selective sweep. Genes that show such a sweep signature in one subspecies or population are interesting because they may contribute to the speciation process by causing hybrid incompatibilities.

The microsatellite screen makes some assumptions about duplicate gene pairs that will be tested in the second part. The presence of a duplicate can be tested through a PCR scheme, which is done for several mouse species. The data gathered in this process will make it possible to determine the age of a duplication event, and compare it to the age estimated by the K_S value. Also, this test will show if the genes tested are present in all mouse species.

While the first party of this study is a search for candidate genes, the third part focuses on one gene, *Dnahc8*. It has been shown to be a hybrid sterility gene in *M. m. domesticus* and *M. spretus* hybrids, but it is not known what the mechanism of this incompatibility is. To answer the question if the effect is caused by changes in the coding sequence or regulatory effects, sequence data and expression levels will be analyzed in several species.

2 A screen of duplicated genes for positive selection

2.1 Introduction

Duplicated genes have been implicated to play an important role in speciation, as they allow genes to gain new functions or regulative changes very quickly through lifting constraint. Such duplicates may develop this altered function only in one population of a species, giving rise to possible hybrid incompatibility.

This study tries to detect duplicated genes that have undergone a selective sweep in subspecies or populations of the house mouse by analyzing the variability of microsatellite loci.

2.2 Materials and Methods

2.2.1 Search for duplicated genes and associated microsatellite loci

Duplicated genes were searched in the available genome sequence for *Mus muculus* as provided by the EnsEMBL project (http://ensembl.org). The database release 31 was used for the search as described here. All genes are sorted into gene families in the EnsEMBL database, a feature that is computed by generating similarity data among all proteins through blastp, and then clustering the sequences with the help of the TRIBE-MCL algorithm (Enright, Van Dongen et al. 2002). A list of all genes with their family ID was downloaded via the EnsMart service, and filtered for those families that contained only two single genes. The number of gene pairs in the resulting list was 1628 (see digital supplement). A large number of duplicated genes would not be found with this method, as they are grouped in larger families. On the other hand, this study focuses on young duplicates, which are represented in the set of genes found, and for the purposes needed here a complete list of all recently duplicated genes is not needed.

The next step in the analysis was the determination of the age of the duplication through the K_S value. For all genes the protein and nucleotide sequences were downloaded from EnsEMBL. The protein sequences of duplicate pairs were aligned with the muscle software (Edgar 2004), and the nucleotide data was aligned

subsequently on the basis of the protein sequence alignment by means of the tranalign program, part of the EMBOSS package (Rice, Longden et al. 2000). This ensures the resulting nucleotide alignment does not contain codon shifts in any one sequence, which would render the calculation of synonymous and nonsynonymous replacements impossible. This calculation was performed with a software program that implements the method of W. H. Li (1993) to calculate K_S and K_A.

To obtain a subset of gene duplicate pairs that can be presumed to have originated during or after the split that gave rise to the house mouse subspecies group, a cutoff value of $K_S < 0.1$ was chosen. The calculation assumes a mutation rate of 2.1×10^{-8} per base pair per generation (Nachman 1997), a generation time of 2 per year, and that the split occurred 0.5 – 1 million years ago (Guenet and Bonhomme 2003).

For a subset of the genes, the duplication mechanism that led to their existence was determined. Retrotransposed genes have lost all introns in one of the copies, this was tested manually by visible inspection. Whether gene duplication was caused by a large scale segmental duplication was assessed by searching for the genes in question in the non human segmental duplication database as available on the internet (Cheung, Wilson et al. 2003).

2.2.2 Microsatellite analysis

2.2.2.1 Mouse population samples

DNA from four mouse (sub)species, *Mus. m. musculus*, *M. m. domesticus*, *M. m castaneus* and *M. spretus* was used. For *M. m. musculus* genetic material from a Czech (35 individuals) and a Kazakh (36) population was available, the latter deemed the more ancestral one. The *M. m. domesticus* mice studied were caught in West Germany (37), the French Massiv Central (63) and Chicago in the USA (16). The *M. m. castaneus* subspecies was also represented in one ancestral population from India (46), and one presumably derived population from Taiwan (12). All *Mus spretus* mice originate from central Spain (46).

The animals were caught according to a sampling scheme detailed by Ihle et. al. (2006), to minimize the number of inbred individuals in the sample.

2.2.2.2 Microsatellite selection and primer design

To be able to draw conclusions from microsatellite allele patterns as to the evolution of genes, the repeat loci must be located in the vicinity of the gene, at a distance that is dependent on the selection coefficient and the local recombination rate, which, over time, breaks down the selective sweep pattern (Smith and Haigh 1974; Fay and Wu 2000). Per default, a window of 10 kilobases around the gene was searched for suitable microsatellite loci. These loci were chosen on the basis of the repeat pattern, bi-, tri- and tetranucleotide repeat units in the range of 9 to 25 repeats. Other criteria were the uniformity of the pattern, and the possible PCR product of the locus: Loci whose repeat pattern contains irregularities mutate much slower that the total length would suggest, and there may be other repeats close by that make generation of a PCR product with just the desired locus impossible. Detection, filtering and generation of a primer pair for PCR was automated via a perl script, the software used for finding the microsatellites in sequence data was Tandem Repeat Finder (Benson 1999). The primer design was facilitated by the primer3 software (Rozen and Skaletsky 2000), with settings that aim at primers with a melting temperature of 60°C and a optimum length of 20 basepairs.

The list of microsatellites and PCR primers that enclose them was checked by hand for suitable combinations. Six of these were grouped into a multiplex, two loci that differed in PCR product size by a few hundred basepairs were assigned the same of the three available fluorescent dyes. All primer sequences are available in supplement 1.

2.2.2.3 Microsatellite PCR and analysis

All PCRs were done in a total volume of 10 µl, with ca. 20 ng genomic DNA as a template. The chemistry used was either EuroBio DNA Polymerase and the corresponding buffers, or the Qiagen Multiplex PCR kit (Hilden, Germany) as suggested in the instructions. The primer oligonucleotides were obtained from Metabion (Martinsried, Germany), and in a few cases from Sigma (Munich, Germany). The primers were labeled with the fluorescent dyes HEX, FAM and TET, enabling the analysis on an MegaBACE automated capillary sequencer (Amersham,

USA) for size determination. To this end the PCR reaction was diluted 1:20 and one µl was mixed with 15 µl water containing ET550 ROX size standard.

2.2.2.4 Analysis of the microsatellite data

Raw sequencer data was manually allele typed in the Genetic Profiler program (Amersham, USA).

The program *MS analyzer* (Dieringer and Schlötterer 2003) was run on the raw data of the microsatellite allele sizes. It calculates various parameters, among them the heterozygosity and the variance in repeat number of each locus and population. The Hardy-Weinberg exact test was performed using the *Genepop* software (Rousset 2008).

An allele sharing tree was calculated with the program 'neighbor' available in the Phylip package (Felsenstein 1989) on the basis of distance data calculated by *MS analyzer* and visualized with the 'tree explorer' program, part of MEGA (Tamura, Dudley et al. 2007).

To find selective sweep candidate loci, the *lnRH* values were calculated as a measure of relative variability levels, after Kauer et. al. (2003). The *lnRH* statistic when calculated for a range of neutrally evolving microsatellite loci follows a normal distribution. Thus, the values for all loci within a given comparison were Z-transformed and tested whether they conformed to a normal distribution by the Shapiro Wilks test, performed in the Statistica software package (StatSoft, USA). The 95% confidence interval for this data set ranges from -1.96 to 1.96, and values outside of this interval are significant at the 0.05 level.

All raw data and input files used for *MS analyzer* are available in the digital supplement.

2.3 Results

2.3.1 Properties of duplicated gene dataset

Using the procedure detailed earlier, 1628 two-member gene families, or duplicate gene pairs, were found. The ratio of synonymous changes per synonymous site (KS) was calculated for all pairs, the distribution is shown in Figure 4.

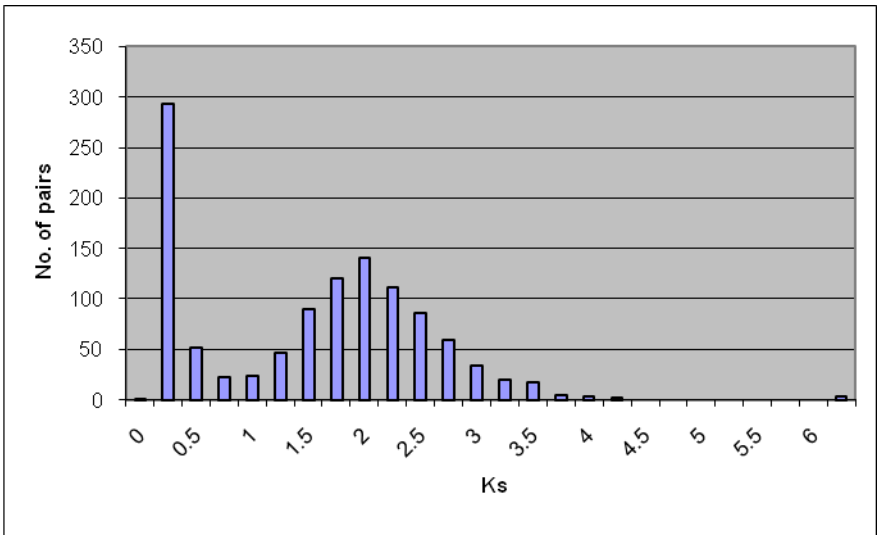


Figure 4: K_S distribution of 1131 gene pairs

The majority of the duplications are young, with a K_S value of between 0 and 0.5. Another peak of duplication lies around a K_S of 2. A linear accumulation of silent substitutions makes it possible to calculate the age of the duplication event (Lynch and Conery 2000). Assuming an average mutation rate of $4*10^{-9}$ per base pair per year in the mouse (Waterston, Lindblad-Toh et al. 2002), this translates to an age of these events of 250 Myr ($2 / 4*10^{-9} / 2$). (A similar rate, $3.06*10^{-9}$, has been found by Nachman et al. (Nachman 1997)). The same pattern, with many very young duplications and a peak of duplication at a more ancient time, was found in other studies using human gene families (Gu, Wang et al. 2002; Cotton and Page 2005). The time estimates of the wave of duplications are much higher than found here, 550 and 500 Myr for both studies, respectively. However, the method used to infer age in these studies is fundamentally different from the one used above. In both cases neighbor-joining trees of gene families with sequence data in various vertebrate species are used to map duplication events to the vertebrate phylogeny. Using

speciation events, whose age is known from the fossil record, as calibration points the age of duplications is determined from the phylogenetic interval it falls into.

I found similar distribution patterns for the K_S of duplicated genes in various mammal species, *Xenopus*, and chicken (data not shown).

The origin of the subspecies is estimated to have occurred less than one million years ago, that of *M. spretus* less than two million years (Guenet and Bonhomme 2003). The so far unexplained phenomenon of the overrepresentation of genes in the age class around 250 Myr (K_S of 2) does thus not affect the young duplicate pairs analyzed here.

In order to restrict the analysis to young duplicates, only those with a K_S of 0.1 (10 Myr) or less were selected; this restriction reduces the dataset to 157 gene pairs, excluding genes with a K_S value of zero. They were classified as retrotransposed or duplicated by other mechanisms through the absence or presence of introns in one of the genes. Of the 157 genes, 129 lack introns and thus likely arose through retrotransposition. A great majority of the retrotransposed genes are located on different chromosomes from their corresponding gene, as opposed to those pairs duplicated by other mechanisms (Table 1).

Table 1: Chromosomal location of duplicate pairs by duplication mechanism

	same chromosome	different chromosome
Retrotransposed	9	118
Other dupl. mech.	12	10

2.3.2 Microsatellite screen for selection

Microsatellites with di-, tri-, and tetra-nucleotide repeats were selected for typing. The mean heterozygosity was compared among the types, to ensure no repeat type is much more monomorphic than others, making it unsuitable for selection tests (Figure 5). There are no significant differences between the heterozygosities when tested with an ANOVA ($p = 0.2$).

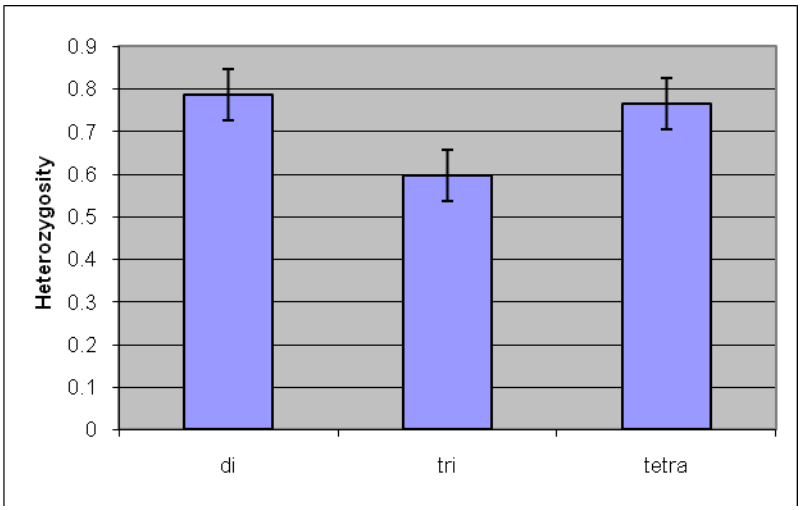


Figure 5: Mean heterozygosity of the different microsatellite repeat types. n is 36 for di-, 8 for tri-, and 25 for the tetranucleotide repeats, all (sub)species data is pooled.

The heterozygosity was also averaged over the four different subspecies. The heterozygosity means over the subspecies show similar values, with *Mus spretus* having the least genetic variation (Figure 5), but the differences are not significant when tested with an ANOVA ($p = 0.34$).

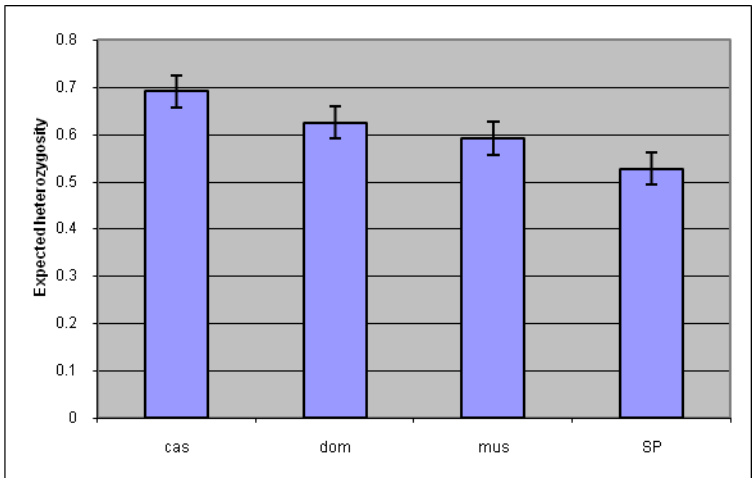


Figure 6: Average expected heterozygosity for each of the subspecies. cas: *M. m. castaneus*; dom: *M. m. domesticus*; mus: *M. m. musculus*; SP: *Mus spretus*

General population genetic parameters derived from the microsatellite data are presented in Table 2. The *M. m. castaneus* subspecies sample shows the highest heterozygosity as well as variance in repeat number, which is expected under the assumption that a large fraction of the animals from this subspecies represents the

most ancestral population, the one from India (46 of 58 total) (Din, Anand et al. 1996). The expected heterozygosity is significantly higher than the observed heterozygosity in all but one case, as has been found previously in a similar study (Ihle, Ravaoarimanana et al. 2006). Ihle et. al. assume the reason for this difference is the social system of mice, which often inbreed within a deme. For this reason a similar sampling scheme was employed that aims to sample individual mice from single demes only, and the expected heterozygosity was used for further analysis, assuming all alleles are thus from a panmictic pool.

Table 2: Population genetic parameters of the subspecies studied

		Heterozygosity				
	No. of animals	observed	expected	Hardy-Weinberg exact test p	Variance in repeat number	Avg. no. of alleles
M. m. castaneus	58	0.47	0.69	<0.001	18.26	12.36
M. m. domesticus	116	0.39	0.62	<0.001	10.34	9.16
M. m. musculus	71	0.43	0.59	<0.001	9.94	8.75
M. spretus	46	0.40	0.53	<0.001	8.35	7.22

An allele sharing tree was calculated for all individuals which is shown in Figure 7. All subspecies are clearly separated in the tree, and all individuals are grouped into populations according to their geographic origin.

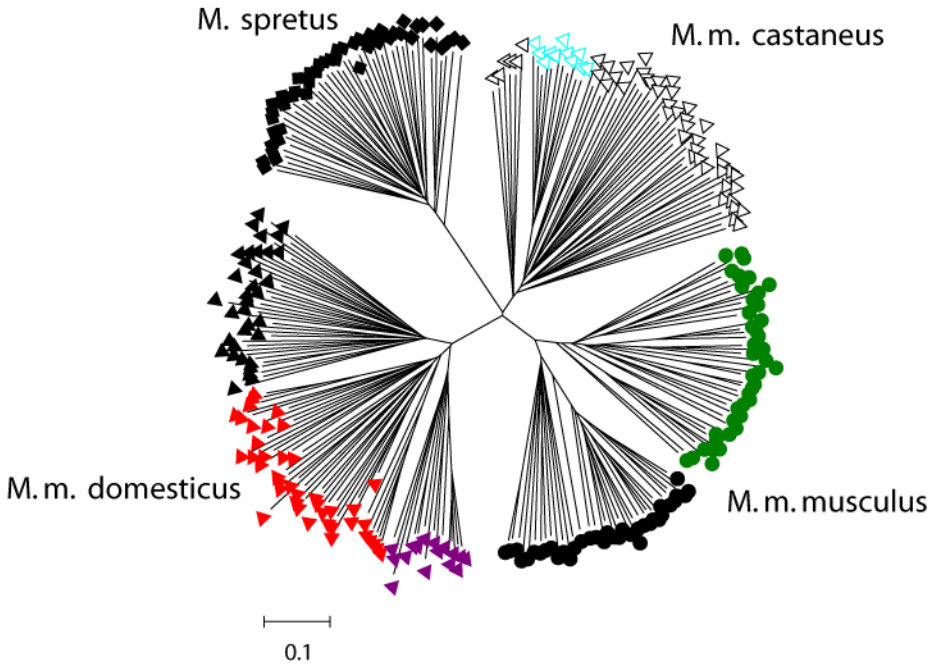


Figure 7: Allele sharing tree based on 69 microsatellites in the four (sub)species. Black circles, Kazakhstan; green circles, Czech Republic; open triangles, India; turquoise open triangles, Taiwan; black squares, *M. spretus*; black triangles, France; red triangles, Germany; purple triangles. Chicago.

The trend that the more ancestral populations have the higher heterozygosity and variability is also evident on the population level. Table 3 shows that the populations from Kazakhstan and India have the highest heterozygosity, and India a noticeably high difference between the expected and observed value, while it is low for the Kazakhstan data as compared to the other populations.

Table 3: Population genetic parameters of the populations studied

		Heterozygosity				
	No. of animals	observed	expected	Hardy-Weinberg exact test p	Variance in repeat number	Avg. no. of alleles
Kazakhstan	36	0.47	0.54	<0.001	7.84	7.06
Czech Republic	35	0.38	0.54	<0.001	8.44	5.66
Chicago	16	0.39	0.52	<0.001	5.1	4.03
France	63	0.41	0.58	<0.001	8.65	6.74
Germany	37	0.38	0.58	<0.001	9.1	6.48
India	46	0.48	0.69	<0.001	16.91	11.59
Taiwan	12	0.40	0.53	<0.001	13.48	4.68
M. spretus	46	0.40	0.53	<0.001	8.35	7.22

2.3.3 Screen for selective sweeps

Tests for selective sweeps at microsatellite loci close to genes were conducted at two levels: The (sub)species level and the population level within a subspecies. Table 4 shows the results for the former set of comparisons. Unfortunately, most of the potential sweep loci will have to be further analyzed with a larger dataset, as the *lnRH* statistic presumes a normal distribution, but only the *M. spretus – M. m. castaneus* data is normally distributed as tested by the Shapiro-Wilk W test.

Table 4: Number of potential selective sweep loci found in (sub)species comparisons

Comparison A - B	No. of loci tested	No. of loci significant* in Population A	Population B	Shapiro-Wilk W, p
M. castaneus - M. m. domesticus	69	2	2	0.94852, 0.00656
M. m. castaneus - M. m. musculus	69	2	3	0.91595, 0.00020
M. m. domesticus - M. m. musculus	69	1	3	0.91690, 0.00021
M. spretus - M. m. castaneus	67	3	1	0.97759, 0.26728
M. spretus - M. m. domesticus	66	4	1	0.95646, 0.02100
M. spretus - M. m. musculus	66	4	3	0.90603, 0.00011

*significance is determined by a z-value above or below the 95% confidence interval of -1.96 - 1.96

All the loci found significant according to the criteria detailed earlier are listed with their expected heterozygosity, based on which lnRH was calculated, and variance in repeat number data in Table 5.

Table 5: Detailed variance data for all loci tested significant in the (sub)species comparisons

For the Ensembl IDs, the leading ENSMUSG was omitted. For calculations of *lnRH* with heterozygosity values of zero in either subspecies, one 'dummy allele' was introduced. Negative *lnRH* values indicate a selective sweep in the species named first. A * indicates p values also significant when Bonferroni correction is applied.

Gene name	Ensembl ID	lnRH	p	expected heterozygosity		variance in repeat number	
M. m. castaneus – M. m. domesticus				*M. m. castaneus*	*M. m. domesticus*	*M. m. castaneus*	*M. m. domesticus*
2310037I24Rik	22992	-2.577	0.00494	0.178	0.761	0.590	117.963
n/a	47509	3.054	0.00131	0.714	0.011	240.452	4.096
BC003885	53740	2.720	0.00326	0.647	0.000	53.368	0.000
n/a	58676	-3.298	0.00048*	0.222	0.887	2.158	195.309
M. m. castaneus – M. m. musculus				*M. m. castaneus*	*M. m. musculus*	*M. m. castaneus*	*M. m. musculus*
Ndufs3	05510	2.830	0.00233	0.723	0.000	73.222	0.000
Mterf	40429	-2.952	0.00159	0.130	0.768	9.262	33.960
Ndufs5	43062	2.640	0.00415	0.682	0.013	15.640	0.240
BC003885	53740	2.523	0.00587	0.647	0.000	53.368	0.000
n/a	58676	-3.475	0.00026*	0.222	0.895	2.158	156.425
M. m. domesticus – M. m. musculus				*M. m. domesticus*	*M. m. musculus*	*M. m. domesticus*	*M. m. musculus*
Ndufs3	05510	3.800	0.00007*	0.758	0.000	37.306	0.000
NM_025868	27081	-2.405	0.00798	0.315	0.850	1.797	79.991
Ndufs5	43062	2.903	0.00187	0.547	0.013	2.315	0.240
Ddx3x	60628	2.611	0.00453	0.466	0.000	0.933	0.000
M. spretus – M. m. castaneus				*M. spretus*	*M. m. castaneus*	*M. spretus*	*M. m. castaneus*
Ndufs3	05510	-2.012	0.02275	0.000	0.723	0.000	73.222
2410127L17Rik	24726	-2.759	0.00289	0.000	0.872	0.000	221.225
NM_025868	27081	-2.371	0.00889	0.044	0.866	0.504	619.262
Mrps33	44111	2.664	0.00391	0.509	0.000	10.843	0.000
M. spretus – M. m. domesticus				*M. spretus*	*M. m. domesticus*	*M. spretus*	*M. m. domesticus*
Ndufs3	05510	-2.012	0.02275	0.000	0.758	0.000	37.306
2410127L17Rik	24726	-2.759	0.00289	0.000	0.823	0.000	47.424
n/a	43192	-1.971	0.02442	0.078	0.818	3.226	157.237
n/a	48989	2.144	0.01618	0.486	0.000	15.557	0.000
n/a	58676	-2.947	0.00164	0.000	0.887	0.000	195.309
M. spretus – M. m. musculus				*M. spretus*	*M. m. musculus*	*M. spretus*	*M. m. musculus*
2410127L17Rik	24726	-2.260	0.01191	0.000	0.685	0.000	12.547
NM_025868	27081	-2.619	0.0044	0.044	0.850	0.504	79.991
n/a	43192	-2.584	0.00480	0.078	0.885	3.226	153.498
Mrps33	44111	2.124	0.017	0.509	0.000	10.843	0.000
n/a	48989	2.077	0.01786	0.486	0.013	15.557	0.416
n/a	58676	-3.254	0.00058*	0.000	0.895	0.000	156.425
Ddx3x	60628	2.346	0.00939	0.609	0.000	44.546	0.000

Some of the genes analyzed are not listed in the Ensembl database with a name or function, but where this information is available it is listed in Table 6. Seven of the

genes listed are not associated with a know function yet, while the others are of different functional categories. Three genes are mitochondrial proteins, one factor that is involved in transcription, and one NADH dehydrogenase mitochondrial precursor, *Ndufs3*. Interestingly, another NADH dehydrogenase enzyme, *Ndufs5*, appears on the list, which is not the duplicate of *Ndufs3*, but belongs to a different pair.

Table 6: Detailed information of loci significant at the species level and the associated genes

Name	Ensembl ID	Function	Chomo-some	MS type	No. of alleles	copy/ org	retro-ransp.
Ndufs3	05510	NADH dehydrogenase iron-sulfur protein 3, mitochondrial precursor	2	CAC	13	org	yes
2310037I24 Rik	22992	Uncharacterized protein C12orf41 homolog	15	GA	13	org	yes
2410127L17 Rik	24726	Uncharacterized protein C9orf41 homolog	19	TTTA	13	org	yes
NM_025868	27081	RIKEN cDNA 2310042M24 gene	17	TC	39	copy	yes
Mterf	40429	Transcription termination factor, mitochondrial precursor	5	TTTG	8	n/a	no
Ndufs5	43062	NADH dehydrogenase (ubiquinone) Fe-S protein 5	16	GT	11	copy	yes
n/a	43192	n/a	8	AGT	21	copy	yes
Mrps33	44111	Mitochondrial 28S ribosomal protein S33	7	TAC	5	copy	yes
n/a	47509	n/a	19	TCC	9	copy	yes
n/a	48989	n/a	5	CT	6	copy	yes
BC003885	53740	Probable ribosome biogenesis protein RLP24	18	TG	7	copy	yes
n/a	58676	n/a	1	AT	26	copy	yes
Ddx3x	60628	DEAD/H (Asp-Glu-Ala-Asp/His) box polypeptide 3, X-linked	X	CA	15	org	yes

Almost all gene pairs were duplicated via retrotransposition, which makes it possible to determine if the gene associated with the microsatellite locus tested is the copy or original. The copies number twice as many as the original genes among the ones listed, but this difference is not significant considering the total number of genes tested (χ^2 test, $\chi^2 = 0.804$; df = 1; p = 0.37).

No pattern is evident in the characteristics of microsatellite loci in the list, neither in the type of repeat unit, as all types from dinucleotides to tetranucleotides are present, nor in the total number of alleles, which varies from 5 to 26.

The same microsatellite dataset was also analyzed grouped by populations within a subspecies rather than by (sub)species. Populations that belong to one subspecies and could be seen as ancestral and derived were compared, except for the comparison of Germany and France populations, for which none can be classified as ancestral. The results in Table 7 show the same range of numbers for selective sweep loci found, between two and five in a comparison. The highest number, four loci, were detected in the American population from Chicago. In contrast to the dataset divided into (sub)species, the population's distribution of *lnRH* values are closer to being normal distributions, for India – Taiwan and Chicago – Germany the Shapiro-Wilk W test is even far from significant. The India – Taiwan comparison also stands out with the least number of sweep candidate loci.

Table 7: Number of potential selective sweep loci found in the different population comparisons

		No. of loci significant* in		
Comparison A - B	**No. of loci tested**	**Population A**	**Population B**	**Shapiro-Wilk W, p**
Kazakhstan - Czech Republic	68	3	2	0.96139, 0.03375
Germany - France	69	1	3	0.96407, 0.04462
India - Taiwan	67	0	1	0.98877, 0.80988
Chicago - Germany	68	4	2	0.98033, 0.35803

*significance is determined by a z-value above or below the 95% confidence interval of -1.96 - 1.96

Table 8 provides the detailed variability data for the single loci in the population comparisons. On the population level there are fewer potential sweep loci with just one detected allele than was the case among the subspecies.

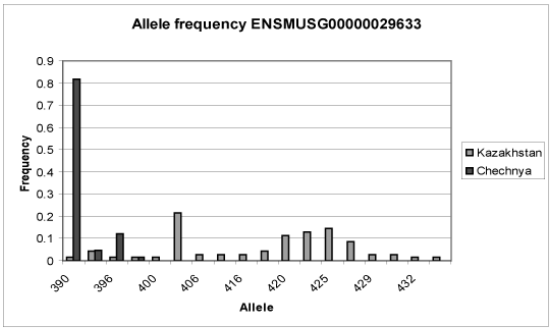
Table 8: Detailed variability data for all loci tested significant in the population comparisons

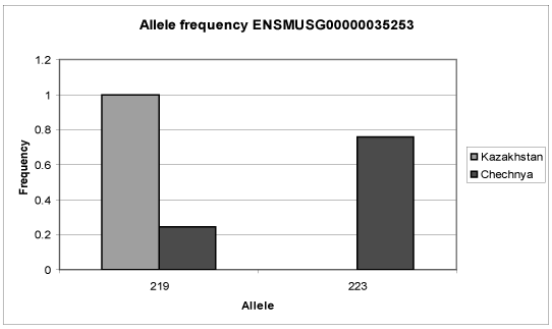
For the Ensembl IDs, the leading ENSMUSG was omitted. For calculations of *lnRH* with heterozygosity values of zero in either subspecies, one 'dummy allele' was introduced. Negative *lnRH* values indicate a selective sweep in the species named first. A * indicates p values also significant when Bonferroni corrected.

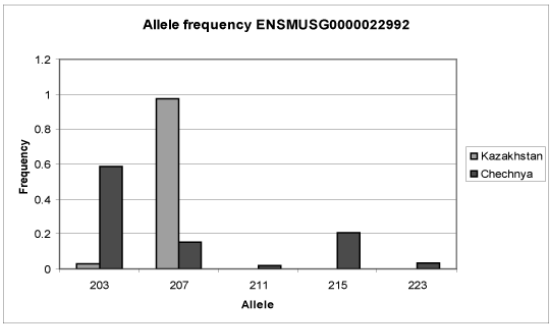
Gene name	Ensembl ID	lnRH	p	expected heterozygosity		variance in repeat number	
Kazakhstan – Czech Republic				Kazakhstan	Czech Republic	Kazakhstan	Czech Republic
0610010K06Rik	29633	3.441	0.00029*	0.901	0.318	149.326	12.469
n/a	35253	-2.305	0.01072	0.000	0.373	0.000	2.985
2310037I24Rik	22992	-2.670	0.00397	0.055	0.598	1.753	32.276
Cox5b	37359	-2.151	0.01578	0.000	0.339	0.000	2.715
n/a	39078	2.027	0.02275	0.851	0.523	656.709	1125.381
Germany – France				Germany	France	Germany	France
Tfb2m	26492	2.910	0.00181	0.902	0.327	88.827	45.599
NM_025868	27081	-2.903	0.00187	0.029	0.575	0.514	3.469
1700066M21Rik	38323	2.503	0.00621	0.757	0.138	37.476	4.711
Txndc14	50043	2.581	0.00494	0.847	0.715	53.594	18.297
India – Taiwan				India	Taiwan	India	Taiwan
n/a	46849	2.206	0.01355	0.855	0.000	111.813	0.000
Chicago – Germany				Chicago	Germany	Chicago	Germany
Gps1	25156	1.985	0.025	0.484	0.083	3.871	0.666
Tfb2m	26492	-2.182	0.01463	0.469	0.902	31.840	88.827
Ndufs5	43062	-2.011	0.02275	0.000	0.591	0.000	2.242
n/a	53178	-2.041	0.02275	0.618	0.866	13.594	109.879
EG627927	56815	2.070	0.02222	0.280	0.000	3.899	0.000
Mrps36	57202	-1.965	0.02442	0.397	0.861	27.145	65.897

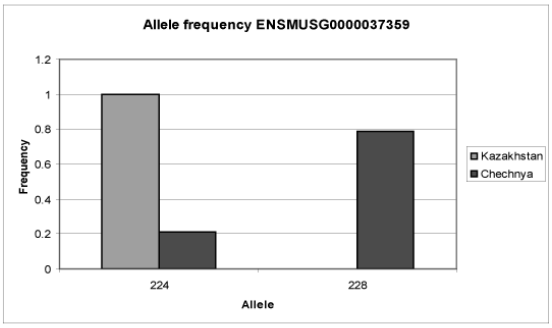
The *lnRH* statistic works only on the basis of heterozygosity, thus the allele frequency spectrum gives additional information for the evaluation of a locus as a selective sweep candidate, as a significant *lnRH* test may be based on very few alleles in both populations. For all loci in the above table these distributions are shown in histograms in Figure 8.

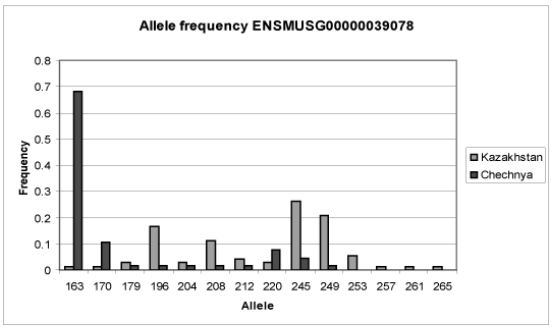
Three of the comparisons only rely on two alleles and a very uneven frequency distribution of these, and two further comparisons consider only three and four alleles, respectively. The distributions also show that in many cases the sweep allele is at the lower size range of the microsatellite. This is clearly the case in eight of the presented loci, only one locus does not follow this trend with the sweep allele in the midrange (ENSMUSG00000050043).

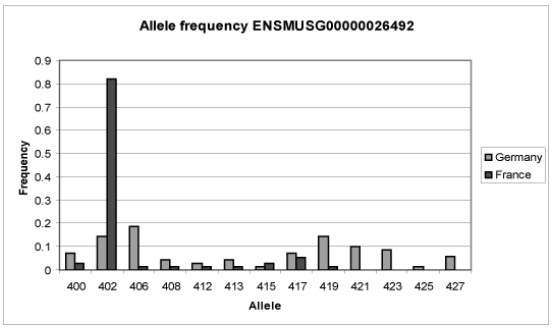

Allele frequency ENSMUSG00000029633
Frequency
Allele
Kazakhstan
Chechnya

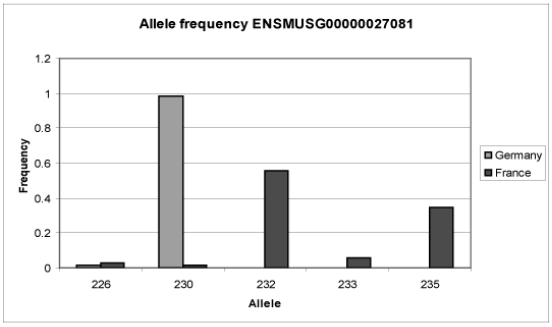

Allele frequency ENSMUSG00000035253
Frequency
Allele
Kazakhstan
Chechnya

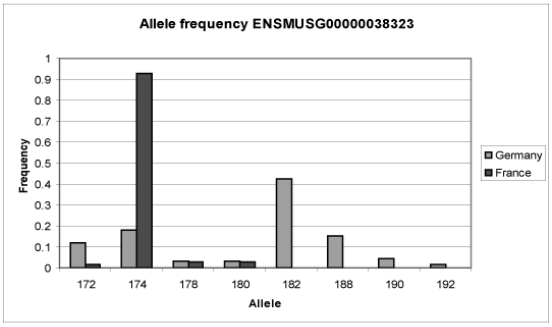

Allele frequency ENSMUSG0000022992
Frequency
Allele
Kazakhstan
Chechnya

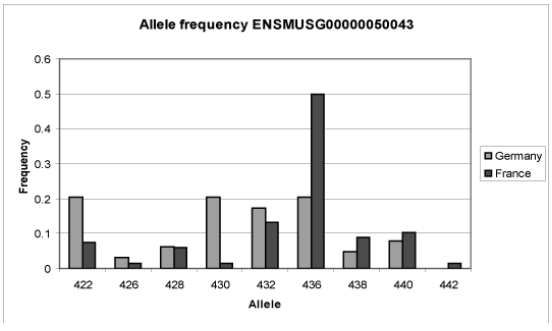

Allele frequency ENSMUSG0000037359
Frequency
Allele
Kazakhstan
Chechnya

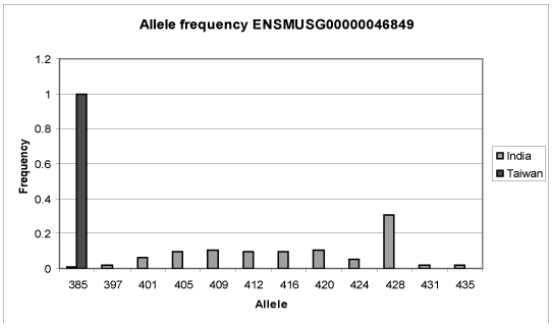

Allele frequency ENSMUSG00000039078
Frequency
Allele
Kazakhstan
Chechnya


Allele frequency ENSMUSG00000026492
Frequency
Allele
Germany
France


Allele frequency ENSMUSG00000027081
Frequency
Allele
Germany
France


Allele frequency ENSMUSG00000038323
Frequency
Allele
Germany
France


Allele frequency ENSMUSG00000050043
Frequency
Allele
Germany
France


Allele frequency ENSMUSG00000046849
Frequency
Allele
India
Taiwan

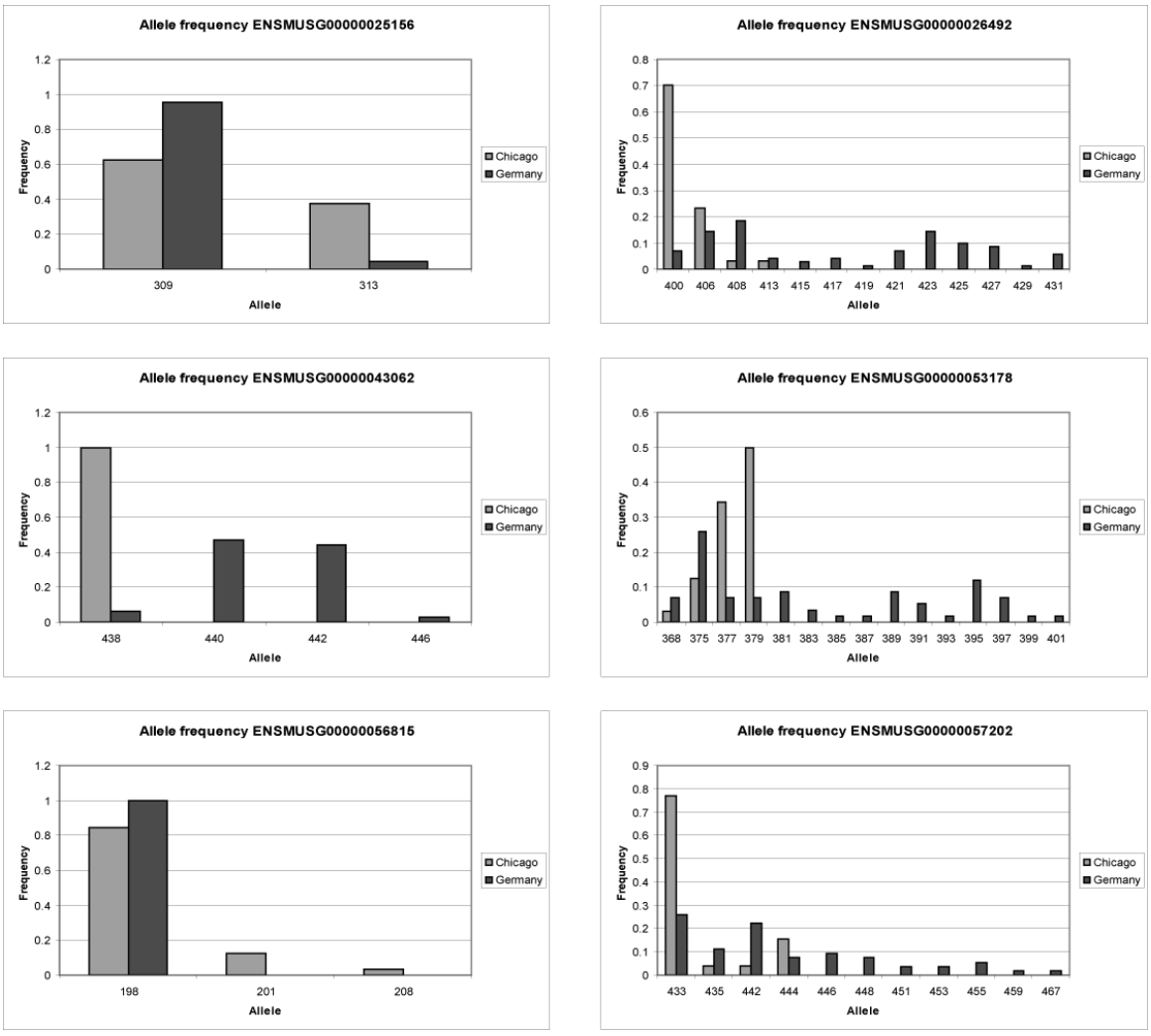


Figure 8: Histograms of allele frequency distribution of the loci listed in Table 8.

The set genes associated with sweep candidates in the population comparisons is, as the species comparison set, not biased toward any chromosomes, and mostly consists of genes duplicated through retrotransposition. The trend seen previously that more copy genes are singled out with the microsatellite screen is reversed in this dataset, eight original genes are opposed to five copies.

The microsatellites themselves also show no bias towards any one repeat unit type or a pattern in the number of alleles.

Table 9: Detailed information of loci significant at the population level and the associated genes

Name	Ensembl ID	Function	Chromo-some	MS type	No. of alleles	copy/ org	retrotr.
Gps1	25156	COP9 signalosome complex subunit 1 (Signalosome subunit 1)	11	GTTT	9	org	yes
0610010K06 Rik	29633	RIKEN cDNA 0610010K06 gene	6	ATAA	28	copy	yes
n/a	35253	n/a	13	AAAC	8	org	yes
2310037I24 Rik	22992	Uncharacterized protein C12orf41 homolog	15	GA	13	org	yes
Tfb2m	26492	Mitochondrial dimethyladenosine transferase 2, mitochondrial precursor	1	AG	24	org	yes
NM_025868	27081	RIKEN cDNA 2310042M24 gene	17	TC	39	copy	yes
Cox5b	37359	cytochrome c oxidase, subunit Vb	13	GTTT	8	copy	yes
1700066M21 Rik	38323	Adult male testis cDNA, RIKEN full-length enriched library, clone:1700066M21	1	TG	20	org	yes
n/a	39078	n/a	11	TATC	24	org	yes
Ndufs5	43062	NADH dehydrogenase (ubiquinone) Fe-S protein 5	16	GT	11	copy	yes
n/a	46849	n/a	4	GGAA	18	copy	yes
Txndc14	50043	Thioredoxin domain-containing protein 14 precursor.	2	AC	15	org	yes
n/a	53178	Mitochondrial transcription termination factor-like precursor	5	AC	25	n/a	no
EG627927	56815	predicted gene, EG627927	X	TGT	6	n/a	no
Mrps36	57202	Mitochondrial 28S ribosomal protein S36	13	AC	23	org	yes

2.3.4 Age of duplicates in correlation to *lnRH*

It can be assumed that duplicated genes do not experience positive selection right after the duplication event, as the theories outlined in the introduction predict. The age of a duplication can be estimated by measuring the divergence between both copies at neutral sites, for example by measuring the number of substitutions per silent site (K_S). The K_S can then be compared to the *lnRH* value of the associated microsatellite. Figure 9 shows scatter plots for these data for all (sub)species comparisons.

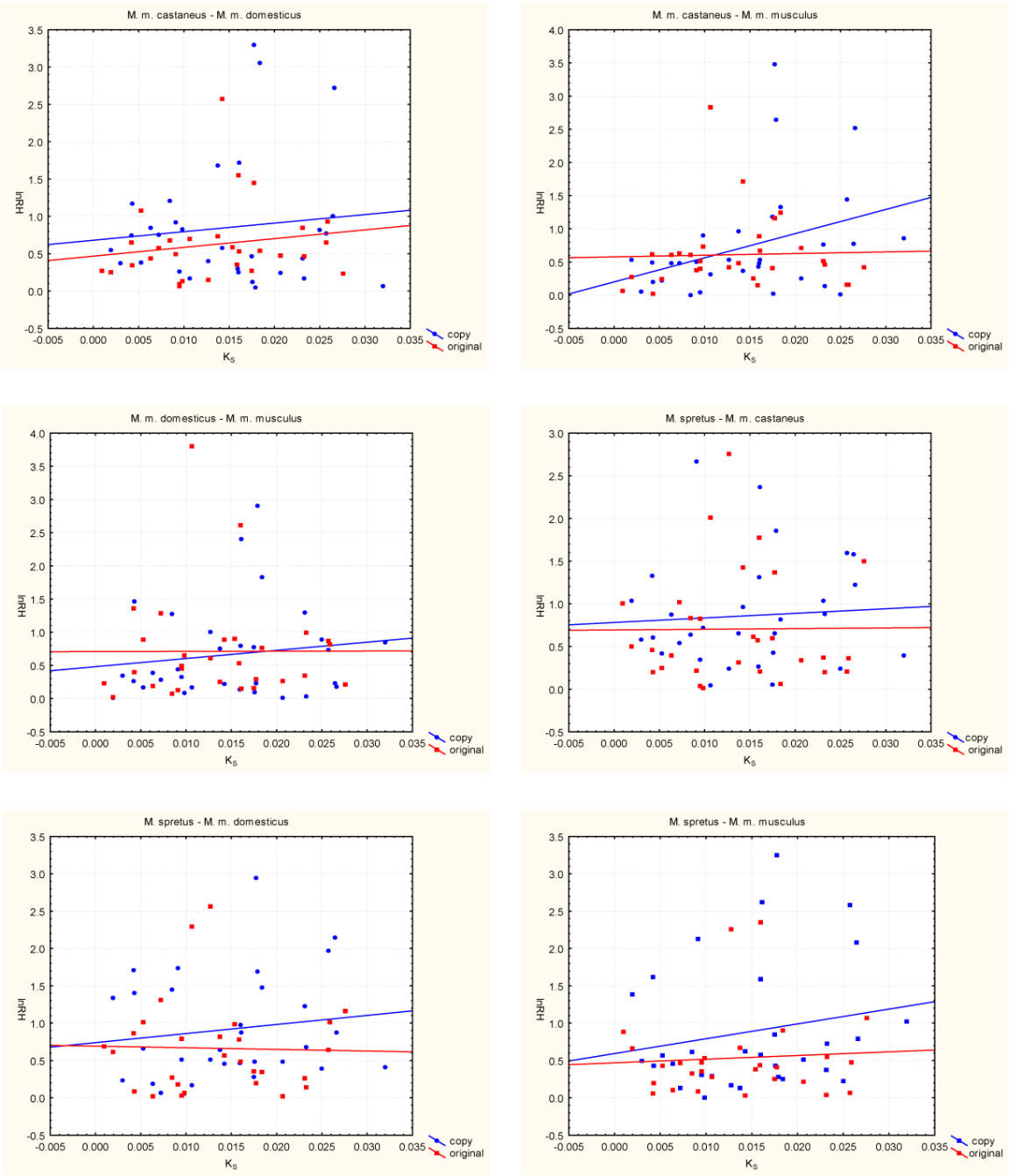


Figure 9: Scatterplots of *lnRH* values against K_S for all (sub)species comparisons

Two observations can me made from these plots: First, loci with a high *lnRH* (>|1.96|) are indeed not present in the range of $K_S < 0.005$, where very recent duplicates would be expected. Second, there seems to be a slight difference between original and copy loci. While a linear correlation is significant in only one comparison, and there only in the copy set of loci, the slope of the regression line is consistently higher for the copy genes in all plots (see also Table 10).

Table 10: Linear correlation between *lnRH* and KS of datasets divided by origial and copy genes

Comparison	Copy/original	r^2	r	p
M. spretus - M. m. castaneus	copy	0.0045	0.673	0.719
	original	0.0001	0.0087	0.9642
M. spretus - M. m. domesticus	copy	0.0188	0.1372	0.4618
	original	0.0007	-0.0261	0.8952
M. spretus - M. m. musculus	copy	0.035	0.187	0.3139
	original	0.0042	0.065	0.7425
M. m. castaneus - M. m. domesticus	copy	0.0118	0.1087	0.5604
	original	0.0279	0.167	0.3864
M. m. castaneus - M. m. musculus	copy	0.1269	0.3562	0.0492*
	original	0.0011	0.0326	0.8668
M. m. domesticus - M. m. musculus	copy	0.0191	0.1381	0.4589
	original	0	0.0027	0.9889

r = correlation coefficient, significant p values are marked with a *

2.4 Discussion

The screening of duplicated genes with linked neutrally evolving microsatellite loci in this study attempts to gain insight into the number and patterns of positive selection on these genes. The focus on relatively young duplicates allows drawing conclusions of the role of these genes in speciation processes. The first step is the assembly of a dataset of duplicated genes in the mouse genome, and the selection of young duplicates.

2.4.1 Duplicate genes dataset

The distribution of K_S in the set of genes extracted from the database shows a pattern found previously in other studies, and one that is reproduced in other species. There is an excess of very young duplicates that have not yet diverged, either to gain new functions or toward pseudogenization. A second peak in the distribution is caused by a large number of duplicates with a K_S of about two that are retained in the mouse genome. It is not known why this is the case, it has been suggested that a whole genome duplication event may be responsible (Gu, Wang et al. 2002; Cotton and Page 2005). Alternatively, a wave of retrotransposition could cause a similar pattern, as was found in primates, although at a much younger rate (Marques, Dupanloup et al. 2005). While those two possibilities concern the generation of duplicate genes, a higher conservation during the timeframe in question would yield the same results, even if the rate of duplicate generation would not change. Both these possibilities could be resolved by categorizing the genes by their mode of duplication, as can be most easily achieved by separating retrotransposed from the other duplicates. If the age distribution would be found the same for genes duplicated by different mechanisms, it would be evidence toward the higher rate of retention, rather than a higher rate of duplicate generation.

The age estimated here for this duplication wave is only half that found by the two previous studies of Gu et al. and Cotton and Page. The dataset used herein cannot be used to analyze duplication age with the methods of those authors, as they require data from several species whose phylogeny includes speciation events of known age. Thus this study is limited to an age estimation based on the rate of synonymous

substitutions as suggested by Lynch and Conery (2000). However, this method has several drawbacks: It requires an unbiased sample of duplicates; this may not be given in this study because only duplicates were selected that are not part of a larger gene family, which may cause the set to include more young duplicates. Another problem, especially for the estimation of the duplication peak, is that the estimation from K_S suffers a large inaccuracy due to saturation of substitutions when $K_S > 1$ (Long and Thornton 2001). These problems may explain the large difference in time estimates.

The vast majority of the young duplicate genes singled out for further study lacks introns in one copy, and has therefore been duplicated by retrotransposition. There are no numbers for the frequency of retrotransposition as opposed to other mechanisms available in the literature for comparison, especially because in this study I only analyzed the youngest age class. However, it has been estimated that 1% of human DNA has been retrotransposed to new locations, and retrotransposons are very abundant in all mammal genomes (Pickeral, Makalowski et al. 2000), so it seems not unlikely that the resulting intronless genes are so abundant.

The detection of selection at duplicated genes was performed by evaluating the variability of microsatellite loci located in close proximity. The dataset was examined for various properties to exclude possible bias.
To widen the range of available loci close to the target genes, microsatellites of different repeat unit length were included. Different mutation rates of the repeat types could influence their variability, as slow mutating loci would not recover as fast from sweep events as fast mutation ones, and thus have a lower heterozygosity. However an ANOVA analysis of the heterozygosity shows it is not biased toward any repeat unit class.

The heterozygosity is one of the population genetic parameters that can give hints to population history, and is relevant when looking evaluating the significance of selective sweeps.
When compared among the four subspecies, *M. m. castaneus* shows the highest heterozygosity. As this subspecies includes the most ancestral population, the mice from India, this result fits the expectation. *M. spretus* has the least variable microsatellites, which may be caused by a late colonization of Spain from Africa and

a population bottleneck. An additional explanation for the reduced heterozygosity may be a technical one: All microsatellite loci were searched in the *M. m. domesticus* genome based on their length, as long microsatellites are more informative. This locus selection can however lead to an ascertainment bias, as most microsatellite have fewer repeat units than the 9 to 25 units in the dataset used here. As the loci mutate independently in either species or population after the split, selectively choosing large microsatellites from one species will reduce the mean size of these loci in the other species (Ellegren 1995; Zhu, Queller et al. 2000). This in turn would lower the estimate for heterozygosity, as shorter microsatellites are less polymorphic.
Nonetheless, any of these differences are not significant as estimated with an ANOVA and may not impact sweep detection.
The fact that the observed heterozygosity is always lower than the expected one has previously been explained with the mating structure of the house mouse (Ihle, Ravaoarimanana et al. 2006), which breeds in demes in which frequent inbreeding may occur (Berry and Bronson 1992).

An allele sharing tree provides an overview over the relationships of the different populations that were typed. All subspecies are well separated, and as expected *M. spretus* is most distanced from the other species. All populations are also clearly separated from one another, with only one exception: The Taiwan population is part of the cluster of *M. m. castaneus* individuals from India, while clearly separated on an extended branch.

On the population level the presumably more ancient populations also have higher heterozygosity, except for the Chicago population whose value lies in between the French and German *M. m. domesticus*.

2.4.2 Microsatellite screen

The *lnRH* statistic was also calculated on data from (sub)species as well as population level data. For the (sub)species comparisons, the statistical prerequisite of normally distributed data is only met in one species pair, *M. spretus – M. m. castaneus*. A dataset of neutrally evolving loci to compare against was not available for most of the subspecies studied here. So, while the loci listed for the other comparisons cannot be

called significant according to the *lnRH* statistic as published (Kauer, Dieringer et al. 2003), they are nonetheless interesting candidates. The large difference in heterozygosity between the loci identified provides ample grounds for the assumption of non neutral evolution in the less variable subspecies.

Another caveat in the *lnRH* method as used here is that it statistically involves multiple tests, as all loci are tested against each other. This results in a Type I error especially for very large datasets. These false positive data could be avoided by use of statistical procedures such as the Bonferroni correction, however these methods are very conservative, and would in turn introduce a large Type II error in the dataset (Schlötterer and Dieringer 2005). If applied to the dataset, four and one loci in the subspecies and population comparisons respecitvely remain significant (see Table 5, Table 8). However, as a screening approach aims for a maximum number of candidates, Bonferroni correction was omitted in this study.

The highest number of loci with strongly reduced heterozygosity is found in *M. spretus*, where it is often reduced to zero. While this could mean that the selection pressure on these loci is very high, there is no reason to assume that it should be noticeably higher in *M. spretus* than the other species over the range of significant loci. The demographic effects outlined before may also play an important role, as bottlenecks in the colonization history can skew the results (Haddrill, Thornton et al. 2005).

The functional analysis of the genes of interest is hampered by the little information that is often available on the genes, especially the copies, which are often annotated only based on automatic computer analysis or EST matching.

What immediately catches the eye is that the list contains four genes that encode proteins that localize to the mitochondrion. Two of these are closely related, the NADH dehydrogenases *Ndufs3* and *Ndufs5*. Both belong to the large protein complex located on the inner mitochondrial membrane named ‘mitochondrial respiratory chain complex I’ and are involved in electron transport. *Ndufs3* is a candidate sweep locus in *M. m. musculus* as well as in *M. spretus*, *Ndufs5* also in *M. m. musculus*.

Mterf is a transcription terminator factor active in the mitochondrion, which is active in transcription from the mtDNA, and also has functions in mtDNA synthesis (Hyvarinen, Pohjoismaki et al. 2007). *Mrps33* is involved one step further, as a ribosomal protein it has a function in protein synthesis, but also in the mitochondrion. BC003885 also has a function connected to ribosomes, but on the biogenesis side.

The last annotated protein, *Ddx3x*, is a DNA helicase involved in DNA transport in and out of the nucleus.
The copies outnumber the original genes in the list, but this result is not significant. There may be reason to expect more copies than originals to be under selection, as retrotransposed genes will have to acquire a promoter from another gene after being retrotransposed. It may thus be more likely that the copy would be positively selected in this case, as it provides a novel expression regulation. If the selected difference between the copies is confined to the functional sequence, neither of the copies is more likely to be positively selected. The data presented here, however, does not allow a final conclusion to be drawn.

The *lnRH* statistic was also calculated for comparisons within subspecies, with the intent on comparing an ancestral and derived population. The *lnRH* values of two of the four comparisons are normally distributed, and two fall only slightly below the 0.05 significance level for the Shapiro-Wilk test of normality. The Indian mouse population does not have any locus with strongly reduced heterozygosity when compared to the mice from Taiwan, and the reciprocal comparison yields only one. This may reflect that these two populations are not as well separated from each other, as was also seen in the microsatellite sharing tree.

Microsatellite loci can become significant in an *lnRH* comparison even if they are present in only very few alleles, which makes it worth to look at the allele distribution. Indeed five loci are only represented by three or two alleles, which reduces the likelihood that the loci have undergone a selective sweep.
Another very consistent patter concerns the loci with many alleles: The ‘sweep allele’ seems to be confined to the lower end of the size range in almost all cases. This is explained by the mutation characteristics of microsatellites: The mutation rate is inversely correlated to the number of repeats, and the more stable alleles are the smaller ones.

The list of sweep candidate genes on the population level also contains mitochondrion associated genes: the aforementioned *Ndufs5* and *Mrps36*, and additionally the gene *Tfb2m*, a transcription factor that regulates mitochondrial gene expression, and an nameless gene that is the duplicate of *Mterf*. *Mterf* is a mitochondrial transcriptional

termination factor and significant in the *M. m. castaneus – M. m. musculus* subspecies comparison, but on the population level its duplicate shows significance within *M. m. domesticus*, between the Chicago and German populations.

As mentioned before, the evolutionary forces that act on duplicated genes change over time after the birth of the gene copy. To find out how the age of the duplication relates to the possible positive selection indicated by a high *lnRH* value, those two variables were plotted against each other. The K_S between duplicates is used here to infer age, which may not be correct in all cases (see next chapter).
No significant linear correlation was found between the two parameters for either original genes or the copies. Considering the models for gene duplication described earlier, it is not expected that a newly copied gene or its original counterpart is under positive selection right away, with the possible exception of the case that a higher expression level through two copies is positively selected. Indeed there are no genes with a high *lnRH* in the class of gene pairs with K_S smaller than 0.01.
Also, this study does not test whether the gene copies are functional, which would require additional experiments which test for transcription and translation of the genes in a way that can differentiate between copies. Experiments regarding the first step, transcription, have been performed in our group (Bilkovski 2006). The expression profiles of four gene pairs were assayed using the Pyrosequencing technique. In two of there pairs, no changes in expression level could be observed between gene copies. The two other gene copies though have tissue specific elevated expression levels as compared to their ancestors. One of the two is *Ndufs5*, which was found to be under selection in this study, its expression is high in the lungs and spleen.
While expression is no proof that a gene is functional, it is a prerequisite, and especially tissue specific expression may be a basis for adaptation. To further pursue the question of functionality, the genes would have to be analyzed on the protein level.

2.4.3 Conclusions

In this study I attempt to find duplicated genes that underwent a selective sweep to infer positive selection. Genes among the young duplicates that come under positive selection in one mouse subspecies or population are likely to contribute to the

speciation process, if they gain a new function only in one population and degrade in the other.

Such candidates for selective sweeps were found in all comparisons conducted with the criteria used. However, this screen can only provide candidates which show the greatest potential of positive selection among the loci tested. More experiments must be performed on these candidates to reach more definite conclusions for two reasons. The limited dataset analyzed here does not provide the statistical power to show selective sweeps with high reliability, especially because a comparison dataset of neutrally evolving loci is not available. Also, it was not tested if all of the genes in the study are active and expressed, or present in all species or populations.

Nonetheless this screen provides a good basis for further study of duplicated genes in a speciation context in the house mouse model system.

3 Estimation of duplication age

3.1 Introduction

The composition of the list of genes to be analyzed in the previous chapter relies on the number of synonymous changes per synonymous site, K_S, to estimate young duplicates. It is assumed that this value will correlate with the age of the duplication event, and increase in time dependent on the mutation rate. Another way to estimate the time at which the duplication event has taken place is to find out which mouse species has both gene copies. If some species do not have both, but more recent species do, the duplication event can be placed between two nodes of the species tree.
In this study I will examine the relationship of K_S and the duplication age estimated from absence and presence in different mouse species.

3.2 Materials and methods

3.2.1 Animal material

All experiments are based on genomic DNA of several mouse species: *M. caroli, M. famulus, M. macedonicus, M. specilegus, M. spretus, M. m. castaneus, M. m. molossinus, M. m. domesticus* and *M. m. musculus*. In addition, a C57Bl6 inbred strain animal was used as a positive control.
The three *M. musculus* subspecies are represented with three individuals each from different populations: *M. m. domesticus* with mice of the Germany, France and Chicago populations; *M. m. musculus* with mice from Kazakhstan and Czech Republic; *M. m. castaneus* with mice from India and Taiwan. Details about these populations are given in 2.2.2.1. *M. spretus* was also represented with three individuals, for all other species only one sample was available.
Genomic DNA of *M. caroli, M. famulus, M. macedonicus, M. spicilegus* was provided by the GPIA laboratory in Montpellier.

3.2.2 Selection of genes

Duplicate gene pairs to test were selected form the dataset described in 2.3.1. They were picked based on their K_S values, to include young and older duplicates, 20 genes

from the microsatellite screening dataset were tested, and 10 additional genes with a K_S ranging from 0.16 to 1.38. All genes are retrotransposed, as it is necessary that no or little flanking sequence is copied from the original locus.

3.2.3 PCR assay for absence or presence of genes

To asses the time of origin of a duplication a PCR strategy was devised to detect whether a duplicated gene was present in the different mouse species. This strategy is based on the fact that retrotransposed genes are inserted randomly in the genome and the flanking sequence differs from the sequence surrounding the original gene.

Figure 10 shows a schematic overview of the PCR primer design. Three primers are needed to obtain a result that includes a control if no duplicated gene is present in any species tested. As shown, one primer is placed in the region upstream of the gene, and another one in the gene itself. A successful amplification with this primer set yields a product of the size x as denoted in the figure. This product will not be amplified if the gene is not present; however, a third primer is added to the reaction, labeled rev K+ in Figure 10. This primer binds in the noncoding genomic sequence downstream of the putative duplicate, which results in a product of the length y_1 plus y_2, should the duplicate be absent. Primers were designed such that the amplicon generated by primers fwd and rev has a different size to the added y_1 and y_2 sizes. Theoretically, primers fwd and rev K+ may give a PCR product that includes the whole gene, but in practice this is unlikely due to the large size of the genes.

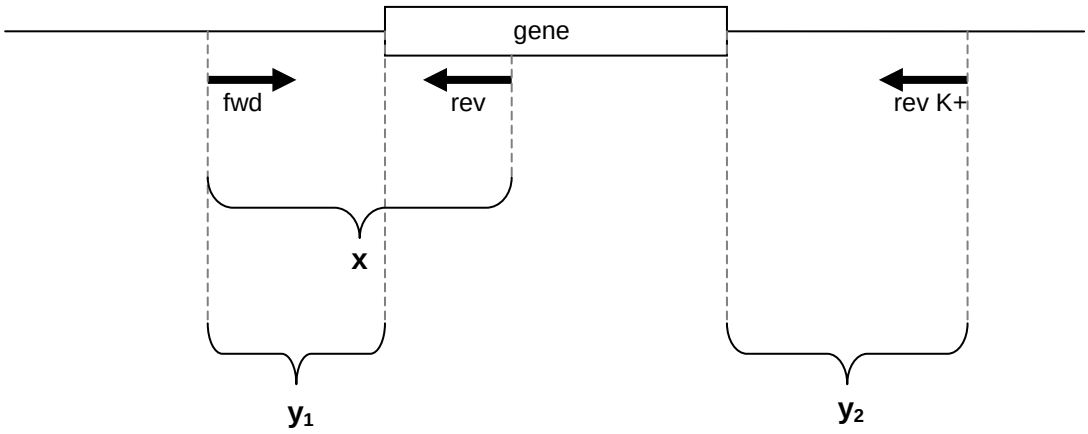


Figure 10: PCR strategy to detect absence or presence of genes. Arrows denote primer binding regions.

For a subset of the primer sets, the forward primer was labeled with a fluorescent dye, fam or hex, to enable detection of the PCR products on a capillary sequencer. All primers are listed in supplement 2.
All PCR reactions were carried out with the Qiagen (Hilden, Germany) Multiplex PCR Kit according to the instructions, primers were ordered from Metabion (Martinsried, Germany).

3.2.4 Analysis of PCR assay

PCR products were either analysed on an agarose gel, or were diluted and analyzed on a ABI 3730 capillary sequencer. In the latter case the resulting data files were analyzed with the GeneMarker program (SoftGenetics, State College PA, USA).
The PCR results for each sample were classified into four categories: Those that yielded no product (no result), those with a product of size x (present), those with a product of size y_1 plus y_2 (absent), and inconclusive results (inconclusive). The latter category includes all samples in which both products were detected, or only products that differed from both the expected amplicon sizes.
The gene was classified as present in a given species or subspecies when two of the three individual DNA samples available for the three *M. musculus* subspecies were assigned to the 'present' category. The same scheme was followed for the absence of the gene copy.
To estimate the time of the duplication event, the duplication was mapped on the phylogenetic tree calculated by Guenet et. al. (2003).

3.3 Results

Not all genes tested yielded results that could be used to estimate the age of the duplication. Eleven sets of diagnostic PCR results were precise enough to be analyzed, the remainder did not yield any product or product of the wrong sizes in some or most of the species. If it was not possible for a number of the ancestral species to infer absence or presence, the dataset was discarded. Seven of the duplicates that could be analyzed were young genes from the microsatellite dataset, and four were of those with higher K_S added to the analysis. The results for nine genes are shown in Figure 11 as colored labels on the phylogenetic tree.

The duplicates vary very much in their distribution among the species. Two are present only in *M. m. domesticus*, while others could be detected in all species. Two genes are not depicted in Figure 11, Ensembl genes ENSMUSG00000055936 and ENSMUSG0000006270, they were also found in all species.

Table 11: K_S value and age as estimated from absence and presence in the species

Gene	K_S	est. age (mio years)
ENSMUSG00000044111	0.00912	0.6
ENSMUSG00000046153	0.0106	2.5
ENSMUSG00000060552	0.016	>3
ENSMUSG00000027081	0.01614	0.7
ENSMUSG00000043062	0.0179	0.6
ENSMUSG00000047509	0.01836	2
ENSMUSG00000037359	0.02311	0.6
ENSMUSG00000047168	0.16782	3
ENSMUSG00000049635	0.20561	2
ENSMUSG00000006270	1.261	>3
ENSMUSG00000055936	0.615	>3

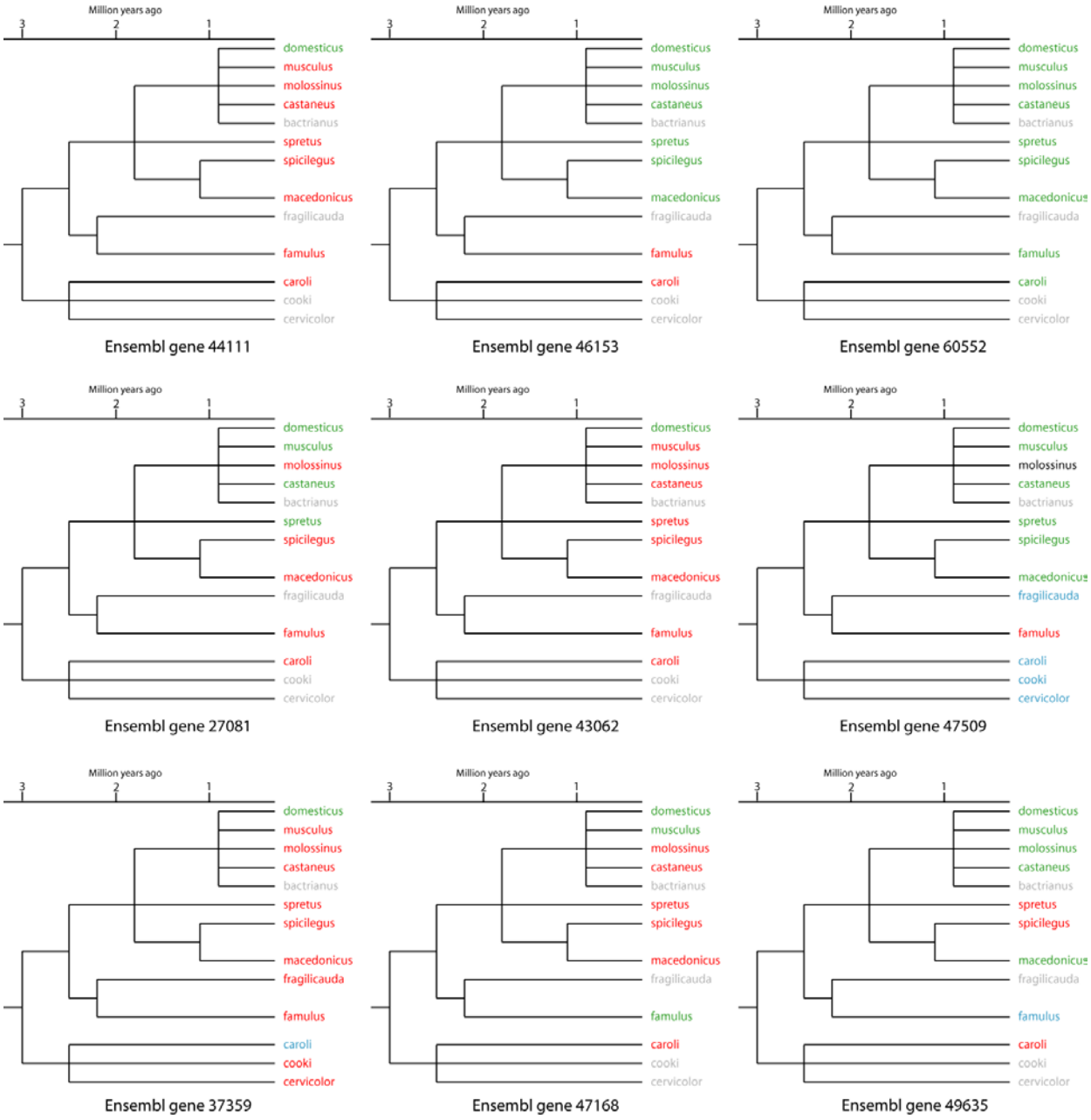


Figure 11: Absence or presence of retrotransposed genes in mouse species. The gene is present in species colored green, absent in those colored red. Blue signifies inconclusive data, black no data. Grey species were not available for testing.

In Figure 12 the data is plotted in a scatter plot, two genes (ENSMUSG00000055936 and ENSMUSG0000006270) with K_S values higher than 0.6 are not included. The data points form three distinct groups; four genes of low K_S cluster around 0.5 million years age, three of the same K_S range are older by estimated age, and two genes with the much higher K_S are an estimated three and two million years old.

The genes ENSMUSG00000055936 and ENSMUSG0000006270 have the highest K_S values of the loci tested here, and are also found in all species, so the estimated age is higher than 3 million years.

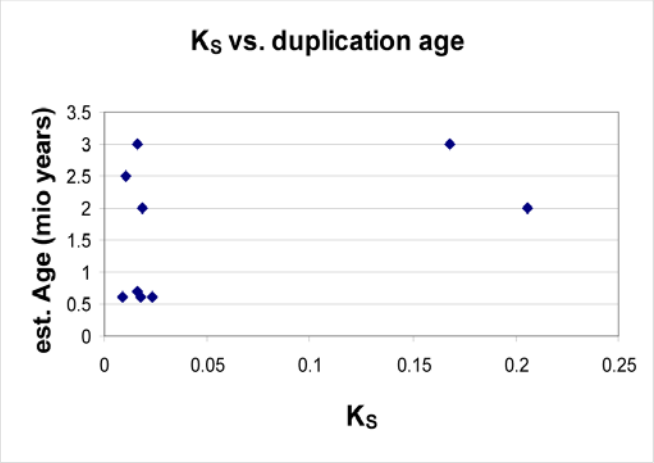


Figure 12: Plot of K_S against the estimated duplication age for nine genes.

Due to the low number of data points, and the difficulty in making judgments of the precision of the estimated duplication age, no statistical analysis was performed. It is nonetheless evident that three loci deviate from the expected linear correlation of age with K_S.

Some properties of those three genes are listed in Table 11. The name of the original genes are given, as the copies are not annotated with name. In all cases the original and retrotransposed genes are found on different chromosomes.

Table 12: Details of three 'outlier' genes

Gene	Name or org.	Chr. location org / copy	Size of gene (bp)
ENSMUSG00000060552	Ddx3x	X / 6	168
ENSMUSG00000046153	Ndufs3	2 / 11	792
ENSMUSG00000047509	Lztfl1	9 / 19	900

3.4 Discussion

The PCR assay method itself does not have a high rate of success. It is problematic because two of the primers bind in non coding sequence, and they have to bind in several different species for which no sequence data is available.

The age estimates performed by placing the origin of a duplication between two nodes of the phylogenetic tree is not a precise method. The range of time between two nodes of the tree is large, and correct dating also depends on the accuracy of the divergence times calculated when constructing the phylogenetic tree. However, realistic time estimates in years are not necessary to compare the K_S and duplication time relationship, as long as the values among lineages are consistent.

Even without tests for statistical significance three gene copies stand out from the set tested. While they have the same K_S as other genes in the set, they are present in many more or all species tested, and are thus old duplications. The expectation is that the synonymous sites should evolve linearly with time depending on the mutation rate, these genes do not fit that pattern.

One reason for this could be gene conversion, an event during DNA repair that leads to DNA transfer from one locus to another homologous locus. The requirement for gene conversion seems to be a very high sequence similarity over a stretch of about 200 bp in mammals, and the region affected is rarely larger than 1 kb in mammals (Chen, Cooper et al. 2007). Thus, the genomic size of the genes' coding regions would allow for gene conversion. However, the retrotransposed gene copies studied here are not on homologous chromosomes. The vast majority of reported cases of gene conversion happen either between homologous chromosomes, or are intrachromosomal (Chen, Cooper et al. 2007). Interchromosomal gene conversion has been found only rarely, one example is the human von Willebrand factor gene (Gupta, Adamtziki et al. 2005). Even though rare, this type of gene conversion could explain the data.

The percentage of genes that show the unusual pattern can not be estimated from the results here, as the method most likely introduces a bias towards gene copies that have not changed much in their history: If the reverse primer, which has been designed to fit sequence from *M. m. domesticus*, does not bind in the gene in more ancestral species, the data can not be used and is discarded.

To confirm these results and potentially make out characteristics shared by the genes that lie outside of the expected correlation more data is needed. But whatever the mechanism is, the fact that K_S does not necessarily reflect the age of a duplication is important to regard in surveys that target young duplicates specifically, or try to make inferences from patterns or trends involving K_S.

4 Analysis of *Dnahc8*

4.1 Introduction

Genes that cause hybrid incompatibility between different species have been termed 'speciation genes'. Only few of such genes are know, which makes it necessary to analyze more such examples in order to answer basic questions about speciation. One of these questions is whether the genetic incompatibilities are based on coding changes, or changes in expression.

To answer this question for one hybrid incompatibility gene, *Dnahc8*, sequence based tests for positive selection and tests of expression in different species are conducted in this study.

4.2 Materials and Methods

4.2.1 Animal material

Live mice of the species *M. caroli*, *M. macedonicus* and *M. spicilegus* were obtained from the GPIA laboratory at the University of Montepellier II. They are derived from wild animals and bred under laboratory conditions for several generations. The *M. castaneus* animal originates from Taiwan, also bred in the lab for several generations from mice provided by A. Yu, Taiwan University. The *M. m. musculus* mouse sequenced is of a population caught in Vienna. The population sample of *M. m. domesticus* mice consists of animals from western Germany (see Table 15), bred for one generation in the lab, with the exception of D3, D4, D6, TP 5.1 and TP17_2, which were taken directly from the wild. Care was taken to sample different demes when catching the mice, as described in (Ihle, Ravaoarimanana et al. 2006).

The populations sample of *M. spretus* consists of mice trapped in the wild in central Spain, as described previously (2.2).

4.2.2 Sample preparation and Sequencing

The *Dnahc8* gene was sequenced in various mouse species and population samples. For the different mouse species as well as the *M. m. domesticus* population sample the whole gene or large parts thereof were sequenced, for the *M. spretus* population sample single exons were chosen for sequencing. Due to the large size and complex exon structure of the gene, in the first two cases the sequencing template was cDNA produced from RNA extracted from testis tissue.

4.2.3 RNA and cDNA preparation

Mice were killed through gassing with carbon dioxide in a closed container. The animals were then dissected with minimal time delays and the organs frozen in liquid nitrogen immediately after removal.

Total RNA extraction was performed using Trizol reagent (Invitrogen, Carlsbad, USA) and LiCl precipitation after the manufacturer's recommendations. RNA samples were dissolved in water and stored at -80ºC.

For RNA to be used as a real time PCR template, a DNA digestion step was carried out, using the product DNA-free (Ambion, Austin, USA) as per the manufacturer's manual. The RNA was controlled for degradation by analysis with an Agilent BioAnalyzer 2100 and the corresponding RNA Nano chip and kit (Agilent Technologies, Santa Clara, USA).

Between 1 and 2 µg total RNA were transcribed into cDNA with the enzyme and reagents supplied in Invitrogen's Superscript III kit, following the accompanying protocol. Random hexamer primers were chosen for the reactions (Fermentas, Vilnius, Lithuania).

4.2.3.1 PCR amplification and sequencing

For all PCR reactions the Multiplex PCR kit (Qiagen, Hilden, Germany) was employed. Cycling conditions were choosen to represent the manufacturers recommendations and standard protocols (Sambrook and Russell 2001).

To facilitate sequencing of the *Dnahc8* gene with a minimum of PCR reactions, the gene sequence was subdivided into 5 regions of about 3 kb to be amplified. Sufficient primers were designed to bind within those fragments to enable gapless sequencing of the PCR products. All primer sequences can be found in supplement 3. The sequencing reaction was performed using the BigDye Terminator kit 3.1 (Applied Biosystems, Foster City, USA) according to the instructions provided, and then run on an ABI 3700 DNA sequencer. If any one sequencing reaction failed, another PCR was performed using appropriate primers to generate a smaller product for repeated sequencing.

For all sequence editing and assembly the Codon Code Aligner software package (CodonCode Corp., Dedham, USA) was used.

Table 13: Sequence data for different species

Dnahc8 single individuals sequenced from cDNA

Species	Individual	Origin	bp sequenced	% of gene
M. m. musculus	SPB 17.2	Vienna	14197	100
M. m. castaneus	tai 5.2b	Taiwan	14197	100
M. macedonicus	XBS strain	Bulgaria	12487	88
M. spicilegus	ZRU strain	Ukraine	12509	88
M. spretus	3.4a	Spain	14197	100
M. caroli	KTK strain	Thailand	12997	91

As shown in Table 13 the *Dnahc8* gene could not be sequenced fully in all species this was attempted for. Three species lack data, although this missing sequence is localized only at the 5' and 3' ends of the transcript, there are no gaps in the sequence obtained.

Table 14: Sequence data from a *M. spretus* population sample

Population sample sequencing in *M. spretus*

Fragment	No. of individuals	bp sequenced	bp coding sequence
Exon 2	48	354	354
Exon 50/51	48	424	282
Exon 65	12	227	227
Exon 78	48	293	195
total	-	1298	1058
tsc2	48	269	0

Table 14 lists the sequence data obtained from the DNA based sequencing of the *M. spretus* population sample. The Exon 50/51 fragment contains two exons and one

intron in one PCR product that was sequenced in one run. As a control, noncoding sequence from another region of chromosome 17 was sequenced, an intron of the *tsc2* gene.

All single individual *M. m. domesticus* mice for which a large proportion of the gene was sequenced is shown in Table 15. These sequence data do contain gaps, for some parts sequence could not be obtained due to problems in the PCR or sequencing reactions.

Table 15: Sequence data from a *M. domesticus* population sample

Population sample sequencing in *M. m. domesticus*			
Individual	origin	bp sequenced	% of gene
TP3a_2	Germany	10797	76
TP4a	Germany	11155	79
TP5.1	Germany	10767	76
TP7/10_B	Germany	10869	77
TP17_2	Germany	11239	79
D3	Germany	11047	78
D5	Germany	11211	79
D6	Germany	11087	78

4.2.4 Polymorphism analysis

Sequence diversity measures that were determined include π, which is based on the average pairwise difference between sequences (Tajima 1983) and θ_W, which is based on the number of segregating sites (Watterson 1975). The Tajima's *D* statistic measures the difference between π and θ_W; negative values indicate an excess of rare mutations, a pattern consistent with a selective sweep, and positive values indicate an excess of high frequency mutations, consistent with balancing selection (Tajima 1989).

These data (π, θ_w and Tajima's D) were calculated using the DnaSP program in version 4.20 (Rozas and Rozas 1999). To generate input files, the sequences were aligned with the muscle alignment program (Edgar 2004) at standard settings. In the case of the *M. spretus* exon sequences, all sequences were trimmed to restrict them to the coding parts. All sequences were converted to random haplotypes, by generating two copies, each with the corresponding base of a heterozygote base in the sequence.

For all analysis of divergence between *M. spretus* and *M. m. domesticus* the published *Dnahc8* sequence (NCBI Refseq NM_013811) was used to represent the latter species.

4.2.5 Phylogenetic tree

A phylogenetic tree was calculated using Bayesian inference as implemented in the program *MrBayes* (Ronquist and Huelsenbeck 2003). The dataset described in Table 13 was clipped to the shortest sequence in order to make the set homogenous in length. MrBayes was run using a GTR model, and a rate variation gamma distribution with a proportion of invariable sites; the default values for the priors were not changed.

4.2.6 Detection of positive selection

4.2.6.1 McDonald Kreitman Test

The McDonald Kreitman test was performed on the *M. m. domesticus* population sample sequence data (Table 13). This test contrasts synonymous and nonsynonymous substitutions within and between species (McDonald and Kreitman 1991) to detect traces of positive selection. Under the assumption that mutations are neutral, the ratio of the synonymous and nonsynonymous changes should remain constant over time. This is the null hypothesis, it is tested with a chi square test if the classes of polymorphisms are independent from another.
The second species used for the test was *M. spretus* (individual 3.4a, bred for several generations in the lab from a wild mouse). The test was computed with the DnaSP program.

4.2.6.2 K_A/K_S based methods of detecting positive selection

The gene sequences from the various species studied were analyzed with respect to the synonymous (K_A) and nonsynonymous (K_S) substitution rates and the comparison among those rates. In order to analyze sequence data in this respect, alignments must be correctly representing the codon structure of the sequences. To achieve this, gene

sequences were translated to amino acid sequences in the correct reading frame through DnaSP. The alignment was calculated by muscle, and then applied to the DNA sequence data by the program tranalign, part of the EMBOSS package (Rice, Longden et al. 2000).

All K_A/K_S calculations were performed using the yn00 program that is part of the PAML package and implements the method of Yang and Nielsen (Yang and Nielsen 2000).

Two maximum likelihood (ML) methods were employed to analyze all the sequences: The Method of Yang and Nielsen works by comparing two codon substitution models, one that is neutral and does not allow sites to have values for ω (= K_A/K_S) greater than one, and a selection model that does not restrict ω for a class of sites (Yang 1997; Yang, Nielsen et al. 2000).

Models M1 and M2, and M7 and M8, respectively, were compared as implemented in the PAML 4 program (Yang 2007). M1 allows 2 ω site classes with $\omega_0 < 1$ estimated from the data or $\omega_1 = 1$, while M2 allows an additional ω value to be estimated from the data which may be >1. M7 fits ω to 10 site classes between 0 and 1 following a beta distribution, whereas M8 adds an additional site class which may be >1 to be estimated from the data. As these models are nested, the significance of the comparison can be evaluated using a likelihood ratio test.

Furthermore, single lineages were tested for signatures of selection using branch models. Three different methods were used: A free ratios model, comparisons of model M0 to runs of the same model with foreground and background branches, and the branch-site model A. The free ratios model allows all branches to have a separate ω, it is compared to model M0 (Yang and Nielsen 1998). The degrees of freedom in the χ^2 test used to compare these models depends on the number of branches tested, and was 10 in this case. The branch model was used in a way that allows two groups of branches to have a separate ω each, and can be compared to the same model with an invariable ω. Model M0 was used for this analysis, and a likelihood ratio test with one degree of freedom was applied.

The 'updated' branch-site model A was utilized as described in (Zhang, Lu et al. 2005). Here, two foreground and background models are compared, but in the null model one site class is fixed to ω = 1. As recommended in the PAML documentation, a χ^2 test with one degree of freedom was used to find significance.

4.2.7 Quantitative real time PCR and analysis

Quantitative real time PCR (q-rtPCR) is a method to quantify cDNA through measurement of the DNA amount during the cycles of a PCR reaction.

The SYBR Green method of rtPCR in the form of the QuantiFast SYBR Green PCR Kit (Qiagen, Hilden, Germany) was chosen to assay transcript abundance. Here, DNA amount is measured by the fluorescent marker SYBR Green, which intercalates into the double helix.

Reactions of 10 µl total volume were set up as described in the accompanying manual, with primer concentrations of a final 1 pmol/ul. The template cDNA was diluted 20fold before use in the reaction. The primers for *Dnahc8* were designed to lie in regions without differences between the species assayed (see supplement for sequences). The regions amplified are shown in Table 16. The glyceraldehyd-3-phosphate dehydrogenase gene (*Gapdh)* served as an endogenous control; the primers used here amplify a 158 bp fragment. Primer sequences are supplied in supplement 5.

Table 16: Fragments amplified in rtPCR

fragment name	length (bp)	position in gene (bp)
hst2 6.7 2	131	1514
affy2	131	13762
5.1r/5.2f	140	11766

All reactions were performed in triplicate and the values averaged across triplicates. In the case that the standard deviation for any triplicate set was above 0.3, one outlier value was removed to get a standard deviation lower that threshold. If the was not the case, the data point was omitted. The average C_T value of the endogenous control was subtracted from the corresponding C_T value of the *Dnahc8* probes to yield the ΔC_T value.

For a visual overview, all expression level data were grouped into four categories (see results) according to the following ΔC_T value ranges: <2 strong, 2-10 medium, 10-14 weak, and >14 absent.

Statistical tests such as ANOVA and Tukey's HSD post hoc test were calculated with ΔC_T as the dempendent variable. The calculations were performed with the Statistica software (StatSoft Inc, USA).

4.3 Results

4.3.1 Phylogenetic tree

To characterize the evolutionary dynamics of *Dnahc8* within the genus *Mus*, the entire coding region was sequenced for several mouse species: *Mus mus musculus, M. m. castaneus, M. macedonicus, M. spicilegus, M. spretus,* and *M. caroli*. Published sequence data from *M. m. domesticus* were also included in the analysis (Fossella, Samant et al. 2000).

To assess the relationship of the sequences between species, a phylogenetic tree was constructed using Bayesian inference. The resulting tree (Figure 13) shows the same topology as expected for the species involved as shown previously (Figure 3). This is also true for the relationship of *M. spretus* to *M. caroli*, if rat sequence is included as an outgroup in the tree calculation (data not shown).

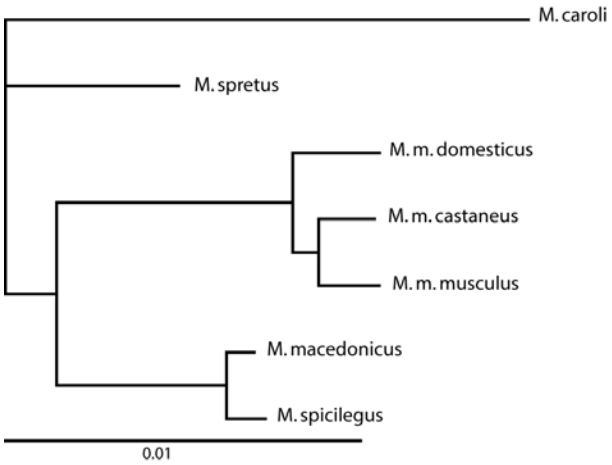


Figure 13: Phylogenetic tree generated with MrBayes using the *Dnahc8* sequence data.

4.3.2 Nucleotide polymorphism

To assess variability in *Dnahc8*, nucleotide polymorphism data were collected for a *M. m. domesticus* and a *M. spretus* population. Table 17 shows the summary data for 5 concatenated exon sequences of *Dnahc8* and a non coding control region (*tsc2*) from *M. spretus* for comparison. Two values of nucleotide diversity are reported: π which is based on average pairwise difference between sequences and θ_W, which is based on the number of segregating sites (Watterson 1975; Tajima 1983). Compared

to the control region, the nucleotide diversity value π is high in *Dnahc8*, but both values are based on only a few segregating sites in the sequences. There are few nonsynonymous polymorphisms; only one was present in the 831 bp region analyzed. Tajima's D statistic is calculated to detect deviations from neutral evolution; the value calculated here for *Dnahc8* approaches significance, but the power of the test is reduced by the low number of synonymous segregating sites. The high value for D in both fragments sequenced suggests that there has been a decrease in population size, or balancing selection acting on both genes. Comparable population data is not available for other loci in *M. spretus*, thus it is not possible to speculate whether high values of *D* are the result of demography or selection.

Table 17: Polymorphism data for *M. spretus*. π and θ_W are shown for synonymous and nonsynonymous sites. Concatenated data from five exons (see Table 14).

					π per site		θ_W per site		Tajima's D	
n (alleles)	base-pairs	segregating sites	syn. poly-morphisms	repl. poly-morphisms	syn	nonsyn	syn	nonsyn	syn	
Dnahc8										
92	831	4	3	1	0.00589	0.00026	0.00292	0.00031	1.8582	0.1 > p > 0.05
tsc2 intron										
92	269	1	1	na	0.00181		0.0073		1.7216	p > 0.1

Table 18 shows an analysis of a subset of sequence data that is available for both *M. spretus* and *M. m. domesticus*. Both measures of nucleotide diversity are higher in *M. m. domesticus* than in *M. spretus*: π is about twice as high and θ_W is 4.5 times higher. While this subset of sequence data enables a direct comparison between the two species, it reduces the number of segregating sites on which calculations are based even more. Tajima's D is slightly negative in *M. m. domesticus* as opposed to the positive value in *M. spretus*, although neither value is significantly different from zero.

Table 18: Comparison of polymorphism data from *M. spretus* and *M. m. domesticus*. Concatenated data from *Dnahc8* exons 50, 51, 66 and 78 (see Table 14).

						π per site		θ_W (per site)		Tajima's D	
	Chromosomes	base-pairs	segregating sites	synonymous changes	replacement changes	syn	nonsyn	syn	nonsyn	syn	
M. domesticus	16	702	4	4	0	0.00672	0	0.00748	0	-0.31696	p > 0.1
M. spretus	24	702	1	1	0	0.00325	0	0.00167	0	1.59613	p > 0.1

Table 19: Polymorphism data for *M. m. domesticus.* Π and θ are shown for synonymous and nonsynonymous sites.

					π per site		θ_W (per site)		Tajima's D	
Chromosomes	basepairs	segregating sites	synonymous changes	replacement changes	syn	nonsyn	syn	nonsyn	syn	
16	9725	65	41	24	0.00579	0.0083	0.00566	0.00096	0.0999	p > 0.1

For *M. m. domesticus*, more sequence data are available than for *M. spretus*, including a larger region of *Dnahc8* sequenced in this study and sequence data from other loci from the literature.

Table 19 shows polymorphism data for a large part of the *Dnahc8* in a population sample of *M. m. domesticus*. Nucleotide diversity values, both π and θ_W, are similar to those observed for *M. spretus* (Table 17). These statistics were measured previously in populations of mice from the same areas in Germany and also France for several autosomal loci (6 genes, 3135 total basepairs) with values very much higher than the ones reportet here: 0.00126 and 0.0026 respectively (Baines and Harr 2007). However, these data were gathered from non coding sequence, a comparison even with synonymous variation may not be applicable.
Tajima's D statistic is slightly negative and not significantly different from zero.

4.3.3 Divergence in functional regions

The main focus of this study is the comparison of full sequences of *Dnahc8* between *M. m. domesticus* and *M. spretus*. The identity between the *Dnahc8* sequences between species is 98.6 % on the nucleotide level and 98.4 % on the amino acid level. The gene sequence in *M. spretus* is missing two complete codons, making the protein

lack serine 1036 and threonine 1061 of the *M. m. domesticus* ortholog. Table 20 lists divergence data for the full gene sequence as well as only the annotated functional domains. The overall nucleotide divergence (K) is 0.0138, and the K_A/K_S ratio is low (0.2671), consistent with purifying selection.

Dnahc8 contains several domains that can be found in other proteins of its family, shown in Figure 14. At the N-terminus lies a proline rich domain 34 amino acids in size which has not been functionally characterized. The dynein heavy chain N-terminal region 1 is thought to enable dimerization among dyneins, and thus may be important for the correct function of the protein (Habura, Tikhonenko et al. 1999; King 2000). The second N-terminal domain does not have any known function, while the ATPase domain is a member of the large superfamily of AAA+ proteins. These are ATPases which can give the energy gained by ATP hydrolysis off in the form of mechanical energy, and in the case of dynein facilitate the motor function of the protein. The heavy chain is common to all dynein proteins and can bind microtubules and also contains ATPase activity.

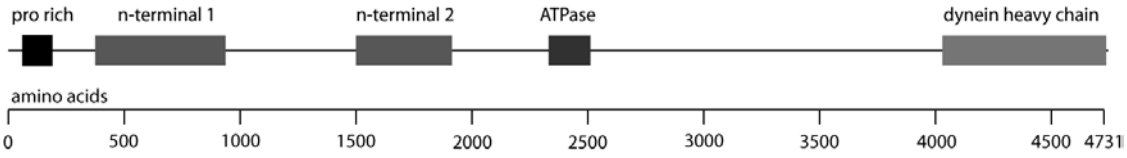


Figure 14: Annotated protein domains in the *Dnahc8* gene.

Divergence data were compared among protein domains (Table 20Table 20). Both amino acid deletions in *M. spretus* do not lie within an annotated protein domain, but in a region of low hydrophobicity. While the divergence and K_A/K_S ratio are similar in the concatenated domain regions, single domains differ somewhat. The ATPase region and also the heavy chain domain have lower K_A/K_S ratios, which could be expected for such widespread and conserved parts of the protein. The N-terminal domain 1 stands out with a higher K_A/K_S ratio, which suggests there is less constraint on that part of the protein.

Table 20: Divergence (K) data between *M. m. domesticus* and *M. spretus* in annotated functional domains of *Dnahc8*. Data is shown for single domains as well as for all domains concatenated.

	basepairs	variable sites	K	syn. changes	nonsyn. changes	K_A	K_S	K_A/K_S
whole gene	14190	196	0.0138	118	78	0.0078	0.0293	0.2671
proline rich	132	1	0.00758	0	1	0.0221	0	na
n-terminal domain 1	1680	28	0.01667	13	15	0.0125	0.0286	0.437
n-terminal domain 2	1236	17	0.01375	10	7	0.0084	0.0259	0.3239
ATPase	432	4	0.00926	4	0	0	0.039	0
dynein heavy chain	2088	29	0.01389	23	6	0.0041	0.039	0.1046
total domains	5568	79	0.01418	50	29	0.0072	0.0333	0.2175

4.3.4 Tests for positive selection

To test for positive selection on *Dnahc8,* codon based maximum likelihood methods were applied to the dataset including sequence data from various mouse species. Here the implementation of the method in the program PAML was used.

For the site specific tests, models M1a and M2a, as well as M7 and M8 were compared and tested for significant differences with a χ^2 test (Table 21). No evidence for positive selection could be found, the lnL values of the corresponding models are almost equal.

Table 21: Results of site model comparisons in PAML.

length (codons)	ω overall	Model	lnL	p
4733	0.14793	M1a	-22452.834	1
		M2a	-22452.834	
		M7	-22452.829	1
		M8	-22452.828	

While no sites could be shown to have traces of selection among the group of mouse species, single branches may evolve under positive selection as compared to others. Branch models available in PAML are able to test this. The free ratio model estimates ω values for all branches in the tree, as shown in Figure 15. In the area of interest, the branch connecting *M. caroli* to the other species and the one at the base of *M. m. castaneus* and *M. m. musculus* have a higher ratio than most others. Whether those differences in ω are significant can be determined by comparing different models as follows.

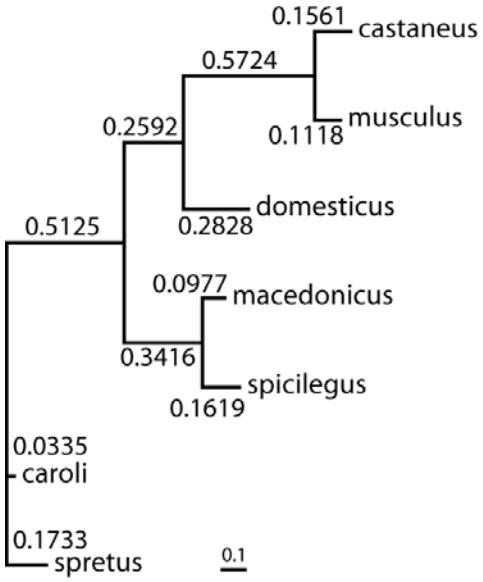


Figure 15: Species tree with ω values obtained in a PAML free ratio model. Branch length correspond to ω, also given in labels.

The first test in Table 22 is the free ratio model. It is compared to the standard model M0, which assumes a fixed ω over all branches. The difference is not significant, and it cannot be concluded that the lineages have different ω ratios. As big differences between lineages are required to get significance in the free ratio test, single branches or parts of the tree were also tested separately. Figure 16 shows a graphic overview of the comparisons performed, the corresponding data is given in Table 22. The main interest in this study lies on the question if the *M. m. domesticus* or *M. spretus* lineages evolve fast or under positive selection. To clarify this issue, PAML models with the branches leading to *M. m. domesticus* as the foreground branches were run and compared to the model that assumes only one ω. The same test was applied to the branch leading to *M. spretus*; all branches tested are labeled in the tree in Figure 16. The likelihood ratio test for branches A,B and C is highly significant, as it is for branches A and B when tested separately, which confirms that ω is significantly different from all other branches. The foreground ω for the combined three branches is 0.3057, branch A has the highest ω (0.5333) when tested separately in accordance with the high ω obtained for this branch in the free ratio model. The ω for branch D leading to *M. spretus* is not significantly different from the rest of the tree.

Although the site model did not detect selection at any sites, it is possible that adaptive evolution occurred only during some time and only in single lineages. The branch-sites model tests this, and was run for the branches A, B and C. The results in

Table 23 show that likelihood ratio test is significant, but the ω in the foreground in the 2a and 2b classes is below one.

Table 22: Results of branch model comparisons in PAML.

	ω / selection parameters	background ω	foreground ω	lnL	p
M 0	0.14793	-	-	-22461.7	0.9955
free ratio	-	-	-	-22423.8	
M 0	0.14793	-	-	-22461.7	**<0.0001**
branches ABC, M 0	-	0.1075	0.3057	-22450.4	
M 0	0.14793	-	-	-22461.7	**0.0061**
branch A, M 0	-	0.1387	0.5333	-22457.12	
M 0	0.14793	-	-	-22461.7	**0.0034**
branch B, M 0	-	0.1291	0.2726	-22457.5	
M 0	0.14793	-	-	-22461.7	0.0583
branch C, M 0	-	0.1406	0.2941	-22459.1	
M 0	0.14793	-	-	-22461.7	0.6029
branch D, M 0	-	0.1452	0.1703	-22461.6	

Note: Significant p values printed in bold.

Table 23: Results of branch-sites model comparison. Site class ω not shown for null model.

	site class	0	1	2a	2b	lnL	p
sites branch model A, branches ABC	**proportion**	0.30257	0.39934	0.12850	0.16959		
	background ω	0.95189	1	0.95189	1	-24477.1	
	foreground ω	0.95189	1	0.95189	0.95189		<0.0001
sites branch model A1, branches ABC, null model						-23911.2	

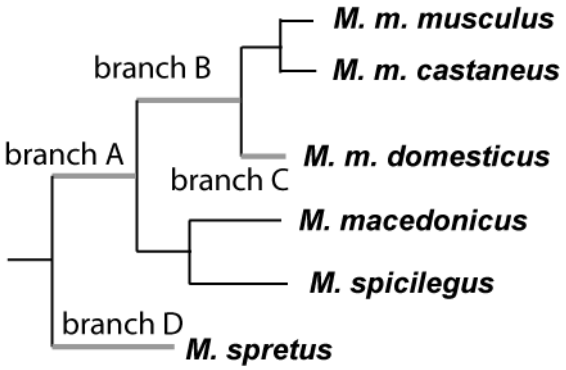


Figure 16: Illustration of the foreground branches used in PAML models.

With population data available it is possible to calculate the McDonald-Kreitman test (McDonald and Kreitman 1991) as a second approach to detect adaptive evolution. This test compares the number of nonsynonymous and synonymous sites within and between species, and the results for a comparison of a *M. m. domesticus* population sample with *M. spretus* as the between species data are listed in Table 24. The test suggests positive selection if the ratio of nonsynonymous to synonymous substitutions is significantly lower within species than between, but this is not the case for the data analyzed here.

Table 24: Results of McDonald-Kreitman test.

	Divergent sites	Polymorphic sites
Nonsynonymous	30	31
Synonymous	66	43
Ratio	0.45	0.72
Fishers exact test p	0.1968	

Notes: Population sample *M. m. domesticus*, n=20; compared with *M. spretus*, n=1. basepairs in test: 11217.

4.3.5 Expression

Expression levels of *Dnahc8* were determined in eight tissues for seven mouse species. Expression results, using the Δct measure, are reported in Figure 17. The *Dnahc8* transcript is most abundant in testis tissue in all species tested. The apparent strong expression value in *M. m. musculus* spleen is caused by an anomalous value in that tissue for the *Gapdh* housekeeping gene used for normalization, the ct value for *Dnahc8* falls into the same range as the other spleen data (see supplementary data).

For all species, expression is significantly higher in testis than in all other tissues ($p < 0.05$, ANOVA and Tukey HSD test).

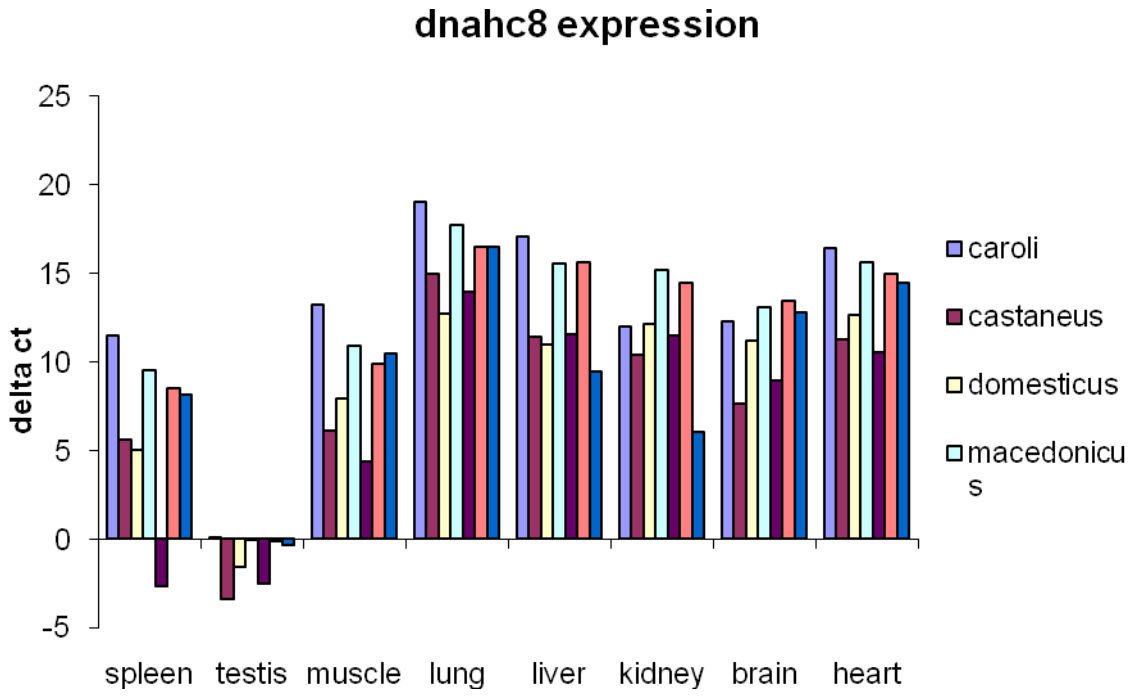


Figure 17: rtPCR for *Dnahc8* in different tissues. Shown are Δct values, a higher value indicates lower expression.

Table 25 reports the expression results, separated into four categories and color coded. The species are sorted according to their phylogenetic relationship (Figure 13). The changes in expression are consistent in some tissues; the expression in spleen and muscle is 'medium' in all three subspecies of the *M. musculus* group, but weak in all other species. *M. macedonicus* and *M. spicilegus* have the same expression profile and are closely related; together with *M. spretus* and *M. caroli* they don't express *Dnahc8* more than weakly in any tissue other than testis.

Table 25: Tabular representation of rtPCR results, grouped in 4 categories.

	Liver	Testis	Kidney	Brain	Spleen	Heart	Muscle	Lungs
musculus	weak	strong	weak		strong	weak	medium	weak
castaneus	weak	strong	weak	medium	medium	weak	medium	absent
domesticus	medium	strong	weak	weak	medium	weak	medium	weak
spicilegus	absent	strong	absent	weak	weak	absent	medium	absent
macedonicus	absent	strong	absent	weak	weak	absent	weak	absent
spretus	weak	strong	medium	weak	weak	absent	weak	absent
caroli	absent	strong	weak	weak	weak	absent	weak	absent

In *M. spretus*, as in the other species, *Dnahc8* expression seems to be mostly testis specific, but not as high as in *M. m. domesticus.* For further clarification of this comparison, several individuals from both species were tested in a further experiment; these data are shown in Figure 18.

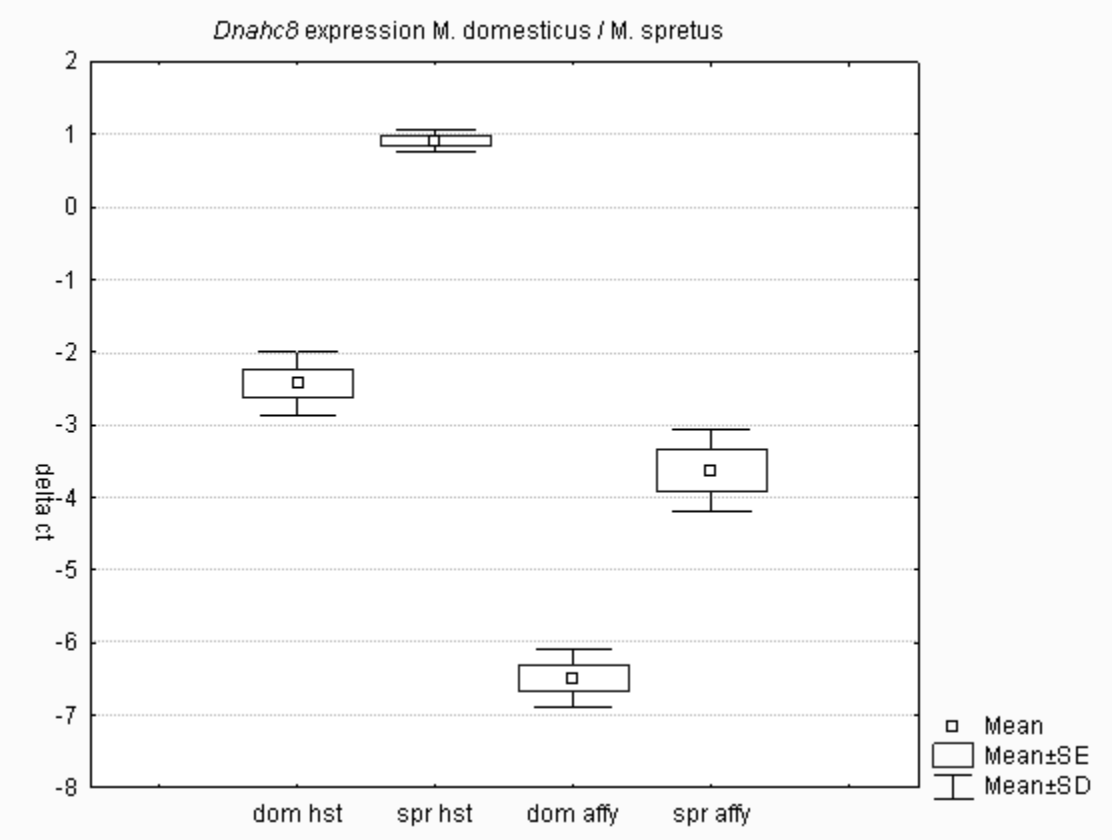


Figure 18: *Dnahc8* expression in *M. m. domesticus* and *M. spretus* as Δct values. Data for two primer pairs are shown, hst and affy. Sample size n = 4 for spr, n = 5 for dom data points, respectively. Mann-Whitney U tests for both comparisons dom hst vs. spr hst and dom affy vs. spr affy p = 0.01430, respectively.

Two different primer pairs were used to measure transcript abundance. Detected expression level differs between the two primer pairs tested, even in the same individual; the ΔC values obtained with the hst primer pair are about 4 cycles lower in both groups of individuals, indicating a much higher expression level than inferred from the affy primers. Nonetheless, the differences between species are highly significant for both primer sets used, and average about 3 ΔC_T steps, corresponding to an eight fold higher expression in *M. m. domesticus*.

Discussion

In this study, I analyzed sequence variation and expression of *Dnahc8*, a potential 'speciation gene' with a well-characterized role in sperm motility, in a range of mouse (sub)species. To find out if coding changes might contribute to the incompatibility this sequence data was used to test for evidence of posititve selection. Additionally, expression levels in different mouse species were assayed to investigate the alternative possibility that regulative changes play an important part in the hybrid sterility.

The phylogenetic tree that was constructed gives some information on the evolutionary history of *Dnahc8*. The topology of the tree is consistent with published phylogenetic relationships between mouse species (Lundrigan, Jansa et al. 2002). This indicates that the gene has not taken an evolutionary path that deviates very much from the species tree, as might be expected, for example, if there had been extensive introgression between species.

4.3.6 No evidence for selection on *Dnahc8* from population genetic data

We compared patterns of polymorphism between *Dnahc8* and other loci to investigate evidence for selection; drastically reduced variation or a significant Tajima's D might indicate a recent selective sweep

The interpretation of the *M. spretus* polymorphism data is hampered by the lack of comparison data from other loci for this species. The single locus sequenced for comparison, *Tsc2*, is not very long, and may also be under selection. Also for this reason, the high Tajima's D, which approaches significance, can not be classified as being caused by balancing selection, because a population size decrease is an alternative explanation for this result. Although sequence data from additional loci are required to assess genome-wide patterns, the similarly high Tajima's D value for *Tsc2* hints that a demographic effect is likely. (Tajima 1989).

A more comprehensive data set was analyzed for *M. m. domesticus*; the entire coding sequence of *Dnahc8* was determined here for 10 individuals and compared to published sequence data for 7 loci from the same populations (Baines and Harr 2007). Nucleotide diversity in *Dnahc8* is about two times as high as for the 7 loci described

by Baines and Harr. A reason may be that the population I sequenced consists of mice from two populations analyzed separately in the previous study. All data presented here come from this combined set of individuals, thus diversity may be expected to be higher. Tajima's D is close to zero, thus there is no evidence for positive selection from these population genetic data. One caveat in this approach is that the Baines and Harr data is based on noncoding sequence, which is compared to data from synonymous sites of coding region of *Dnahc8*, which may not be data of equal evolutionary constraint.

4.3.7 Amino acid divergence in known functional regions of dnahc8

To evaluate whether incompatibility could be caused by amino acid sequence divergence we compared divergence and K_A/K_S among annotated functional regions.

In *M. spretus,* two amino acids of dnahc8 are deleted that are present in all other species sequenced. In a protein as large as dnahc8, the absence of a few amino acids may not have any impact on folding or function, especially as the deletions are not located in an annotated domain. The deletions are in a hydrophobic region, which suggests they lie within the protein structure, and may not contact interaction partners of the dynein chain. But as there is no detailed functional information available for this region of dnahc8, any inference of the potential functional consequences of these deletions remains speculative.

Overall, the K_A/K_S ratio between *M. m. domesticus* and *M. spretus* is rather low, consistent with purifying selection on dnahc8. Comparing divergence among functional domains provides a more detailed picture of selection acting on dnahc8.

Both the N-terminal domains have a ratio that is higher than the values for the overall gene sequence or any of the other domains, which indicates that they are less constrained. Unfortunately not much is known about the function of these domains, except that N-terminal domain 1 may be involved in dimerization, and is described as a flexible protein domain (Habura, Tikhonenko et al. 1999; King 2000). It is possible that the interaction is not so optimized as to require a very strict amino acid sequence, or that the interaction partner changed over time. Nothing is known of the function of the second N-terminal domain.

The ATPase that is part of *Dnahc8* belongs to the AAA+ family of protein domains. This family occurs in a wide variety of proteins in animals (Patel and Latterich 1998;

Frickey and Lupas 2004). It provides an integral part of the dynein function, mechanical movement through ATP hydrolysis, and is strongly conserved in dynein heavy chain proteins (Chapelin, Duriez et al. 1997). Thus, it is not surprising there are no amino acid replacements between *M. spretus* and *M. m. domesticus* in this highly conserved domain. Similarly, there is a low K_A/K_S ratio for the heavy chain domain, which also contains ATPase activity.

4.3.8 Tests for positive selection based on nonsynonymous vs. synonymous substitution rates

Two additional methods were applied to test for evidence of positive selection on dnahc8. Codon based maximum likelihood models were applied to test for positive selection acting on specific amino acid sites or in specific lineages. No evidence for positive selection was found using models allowing variation among sites. However, these comparisons test for repeated selection on codons in the whole species tree. As the sterility phenotype shows when *M. spretus* alleles are present in a *M. m. domesticus* background, selection could be taking place in only one of the involved lineages. The free ratio model indicated that the at least one of the branches leading to *M. m. domesticus* has a relatively high ω value. Additional model comparisons isolating these branches confirmed that the ratio for these branches is significantly different from the remaining tree. When all branches leading to *M. m. domesticus* are tested against the remainder of the tree, the average ω is higher than the background, but a value of 0.3057 does not indicate selection, but may provide some evidence for relaxed constraint as compared to *M. spretus*.

The branch-sites model provides the possibility to find selection on only single lineages. It is significant difference with the branches leading to *M. m. domesticus* as foreground, but the ω is not higher than one, and the test does not provide evidence for positive selection.

4.3.9 Expression

Genetic incompatibility can not only be cause by changes in the amino acid sequence of a gene product, but also in differences in gene regulation and expression between

species. To test for such differences, expression was measured and compared among species.

It has previously been reported that *Dnahc8* is expressed in a testis specific manner in *M. m. domesticus* (Fossella, Samant et al. 2000). This result could be confirmed not only for *M. m. domesticus*, but also for all other species of mouse tested. There are no striking differences in expression among the other species tested, but the differences that are present seem to group according to the evolutionary tree of the mouse species. Whether the expression of *Dnahc8* in the organs other than testis is functional in any way cannot be determined, but is not unlikely that a gene which is expressed in one tissue so much stronger has no essential function in the others.

The finding here that *Dnahc8* is expressed in *M. spretus* testis contradicts the results from Fossella et. al., which report little or no *Dnahc8* expression in *M. spretus* testis based on northern blots. Here, expression of *Dnahc8* in *M. spretus* testis was confirmed in a second rtPCR experiment with multiple individuals. One primer pair was chosen to bind in the area that served as a probe binding region in the northern blot experiments reported by Fossella et. al. to obtain comparable results. This experimental design rules out transcription effects such as alternative splicing as explanations for the contradicting results. Although there is evidence for *Dnahc8* expression in *M. spretus* testis tissue, there remains a significant difference in expression strength when compared to *M. m. domesticus*. Thus, differences in dnahc8 expression between *M. spretus* and *M. m. domesticus* may be related to the sterility phenotype.

Although results from both probes used here indicate a similar expression difference between *M. m. domesticus* and *M. spretus*, they show very different amplification levels in this experiment. This difference could stem from several causes. Primer binding, and thus amplification, may not be of the same efficiency for both pairs. Also, the two regions that served as templates are located far from each other, with the ‘hst’ region being located towards the 5’ end of the transcript. Throughout the effort to sequence the gene from cDNA, I noticed difficulties amplifying or sequencing the 5’ end of the transcript. It is possible that this part is degraded faster than the rest, or is not transcribed to DNA as efficiently during the cDNA synthesis reaction. If that were the case it would explain that the ‘hst’ rtPCR reports a lower amount of transcript than the ‘affy’ region that is located closer to the 3’ end. Despite these technical issues, both probes indicate a consistent difference in transcript

amounts between *M. spretus* and *M. m. domesticus*. This difference in expression was independently confirmed by microarray experiments (data not shown).

4.3.10 Conclusions with respect to speciation

The incompatibility of the *M. spretus Dnahc8* gene on the *M. m. domesticus* background that results in hybrid sterility may have its cause at one of two different levels: amino acid sequence divergence resulting in a change in protein function, or regulatory divergence resulting in a change in protein expression.

Protein function could be affected in two aspects. For one, the *M. spretus* allele could have lost its function altogether. This option seems very unlikely, because in this case one would expect the sequence to acquire mutations, such as nonsense mutations, that indicate that the protein is not under selection anymore. Furthermore, it has previously been shown that the phenotype caused by the *M. spretus* allele of *Dnahc8* in the *M. m. domesticus* background does not resemble a complete loss of axonemal dynein, but rather points to a disturbed flagellar developmental process (Phillips, Pilder et al. 1993; Pilder, Olds-Clarke et al. 1993). The other possibility is that interactions with other proteins are disrupted by incompatible amino acid substitutions. The fact that the N-terminal domain 1, which is implicated in protein interactions, seems to be under more relaxed constraint that other parts of the protein may be evidence for this alternative. However, the amino acid sequence identity between both species is quite high, and no significant evidence for selection was found. Of course, such an incompatibility cannot be excluded, as it is possible that only few or even a single change in sequence could cause a large change in function (Hughes 2007).

The regulation of expression is the second mechanism that could potentially cause the incompatibility. I found large quantitative differences in *Dnahc8* expression between *M. m. domesticus* and *M. spretus*. The six-fold lower expression in *M. spretus* compared to *M. m. domesticus* may be enough to prevent correct assembly of the sperm tail in hybrids. Also, the time window of *Dnahc8* expression is very specific during sperm development (Samant, Ogunkua et al. 2002); strong changes in quantitative expression, may reflect differences in temporal regulation which could lead to deleterious effects.

Lower levels of expression of *Dnahc8* in *M. spretus* suggest this protein is not as essential as in other mouse species, and it has thus been down regulated in expression, for example due to pleiotropic effects, or another gene that has taken over its function. A possible reason for this may be a different mating system: *M. spretus* has recently been shown to have some behavioral characteristics of a monogamous species (Cassaing and Isaac 2007). Sperm motility is know to be an important factor in sperm competition (Burness, Casselman et al. 2004; Gage, Macfarlane et al. 2004), so in a monogamous system the selective pressure on this trait may be reduced.
A third characteristic in which *Dnahc8* may differ between the two species is alternative splicing. *Dnahc8* consists of a very large number of exons, thus a large number of different splice variants is possible, and indeed different mRNA products have been found in the 129sv laboratory mouse strain (Samant, Ogunkua et al. 2002).

In summary, it seems more likely that the hybrid sterility caused by *Dnahc8* is based on regulation of the gene rather than changes in the amino acid sequence in light of the data gathered here. To show this conclusively, however, further experiments that can test the mechanism in detail are needed.

5 References

Ardlie, K. G. and L. M. Silver (1996). "Low frequency of mouse t haplotypes in wild populations is not explained by modifiers of meiotic drive." Genetics **144**(4): 1787-97.

Baines, J. F. and B. Harr (2007). "Reduced X-linked diversity in derived populations of house mice." Genetics **175**(4): 1911-21.

Barbash, D. A., P. Awadalla, et al. (2004). "Functional divergence caused by ancient positive selection of a Drosophila hybrid incompatibility locus." PLoS Biol **2**(6): e142.

Barbash, D. A., J. Roote, et al. (2000). "The Drosophila melanogaster hybrid male rescue gene causes inviability in male and female species hybrids." Genetics **154**(4): 1747-71.

Barbash, D. A., D. F. Siino, et al. (2003). "A rapidly evolving MYB-related protein causes species isolation in Drosophila." Proc Natl Acad Sci U S A **100**(9): 5302-7.

Bateson, W. and G. Mendel (1909). Mendel's principles of heredity. Cambridge [Eng.], At the University Press.

Beisswanger, S. and W. Stephan (2008). "Evidence that strong positive selection drives neofunctionalization in the tandemly duplicated polyhomeotic genes in Drosophila." Proc Natl Acad Sci U S A **105**(14): 5447-52.

Benson, G. (1999). "Tandem repeats finder: a program to analyze DNA sequences." Nucleic Acids Res **27**(2): 573-80.

Bergthorsson, U., D. I. Andersson, et al. (2007). "Ohno's dilemma: evolution of new genes under continuous selection." Proc Natl Acad Sci U S A **104**(43): 17004-9.

Berry, R. J. and F. H. Bronson (1992). "Life history and bioeconomy of the house mouse." Biol Rev Camb Philos Soc **67**(4): 519-50.

Bilkovski, R. (2006). Feincharakterisierung von duplizierten Genen in Mus musculus. Institut fuer Genetik, Universitaet zu Koeln. **Diplomarbeit**.

Bonhomme, F., S. Martin, et al. (1978). "[Hybridization between Mus musculus L. and Mus spretus Lataste under laboratory conditions (author's transl)]." Experientia **34**(9): 1140-1.

Boursot, P., J. C. Auffray, et al. (1993). "The evolution of house mice." Annual Review of Ecology and Systematics **24**: 119-152.

Burness, G., S. Casselman, et al. (2004). "Sperm swimming speed and energetics vary with sperm competition risk in bluegill (Lepomis macrochirus)." Behavioral Ecology and Sociobiology **56**(1): 65-70.

Cairns, J., J. Overbaugh, et al. (1988). "The origin of mutants." Nature **335**(6186): 142-5.

Cassaing, J. and F. Isaac (2007). "Pair bonding in the wild mouse Mus spretus: inference on the mating system." C R Biol **330**(11): 828-36.

Chapelin, C., B. Duriez, et al. (1997). "Isolation of several human axonemal dynein heavy chain genes: genomic structure of the catalytic site, phylogenetic analysis and chromosomal assignment." FEBS Lett **412**(2): 325-30.

Chen, J. M., D. N. Cooper, et al. (2007). "Gene conversion: mechanisms, evolution and human disease." Nat Rev Genet **8**(10): 762-75.

Cheung, J., M. D. Wilson, et al. (2003). "Recent segmental and gene duplications in the mouse genome." Genome Biol **4**(8): R47.

Christoffels, A., E. G. Koh, et al. (2004). "Fugu genome analysis provides evidence for a whole-genome duplication early during the evolution of ray-finned fishes." Mol Biol Evol **21**(6): 1146-51.

Cotton, J. A. and R. D. Page (2005). "Rates and patterns of gene duplication and loss in the human genome." Proc Biol Sci **272**(1560): 277-83.

Coyne, J. A. and B. Charlesworth (1986). "Location of an X-linked factor causing sterility in male hybrids of Drosophila simulans and D. mauritiana." Heredity **57 (Pt 2)**: 243-6.

Coyne, J. A. and H. A. Orr (1998). "The evolutionary genetics of speciation." Philos Trans R Soc Lond B Biol Sci **353**(1366): 287-305.

Coyne, J. A., H. A. Orr, et al. (1988). "Do We Need a New Species Concept." Systematic Zoology **37**(2): 190-200.

Crandall, K. A., C. R. Kelsey, et al. (1999). "Parallel evolution of drug resistance in HIV: failure of nonsynonymous/synonymous substitution rate ratio to detect selection." Mol Biol Evol **16**(3): 372-82.

Cucchi, T., J. D. Vigne, et al. (2005). "First occurrence of the house mouse (Mus musculus domesticus Schwarz & Schwarz, 1943) in the Western Mediterranean: A zooarchaeological revision of subfossil occurrences." Biological Journal of the Linnean Society **84**(3): 429-445.

Darwin, C. (1859). On the origin of species by means of natural selection, or preservation of favoured races in the struggle for life. London, John Murray, Albemarle Street.

Dehal, P. and J. L. Boore (2005). "Two rounds of whole genome duplication in the ancestral vertebrate." PLoS Biol **3**(10): e314.

Dieringer, D. and C. Schlötterer (2003). "MICROSATELLITE ANALYSER (MSA): A platform independent analysis tool for large microsatellite data sets." Molecular Ecology Notes **3**(1): 167-169.

Din, W., R. Anand, et al. (1996). "Origin and radiation of the house mouse: Clues from nuclear genes." Journal of Evolutionary Biology **9**(5): 519-539.

Dobzhansky, T. G. (1937). Genetics and the origin of species. New York,, Columbia Univ. Press.

Edgar, R. C. (2004). "MUSCLE: multiple sequence alignment with high accuracy and high throughput." Nucleic Acids Res **32**(5): 1792-7.

Ellegren, H. (1995). "Mutation rates at porcine microsatellite loci." Mamm Genome **6**(5): 376-7.

Ellegren, H. (2004). "Microsatellites: simple sequences with complex evolution." Nat Rev Genet **5**(6): 435-45.

Ellegren, H. and J. Parsch (2007). "The evolution of sex-biased genes and sex-biased gene expression." Nat Rev Genet **8**(9): 689-98.

Enright, A. J., S. Van Dongen, et al. (2002). "An efficient algorithm for large-scale detection of protein families." Nucleic Acids Res **30**(7): 1575-84.

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

Felsenstein, J. (1989). "Mathematics vs. Evolution: Mathematical Evolutionary Theory." Science **246**(4932): 941-942.

Force, A., M. Lynch, et al. (1999). "Preservation of duplicate genes by complementary, degenerative mutations." Genetics **151**(4): 1531-1545.

Forejt, J. (1996). "Hybrid sterility in the mouse." Trends Genet **12**(10): 412-7.

Forejt, J., V. Vincek, et al. (1991). "Genetic mapping of the t-complex region on mouse chromosome 17 including the Hybrid sterility-1 gene." Mamm Genome **1**(2): 84-91.

Fossella, J., S. A. Samant, et al. (2000). "An axonemal dynein at the Hybrid Sterility 6 locus: implications for t haplotype-specific male sterility and the evolution of species barriers." Mamm Genome **11**(1): 8-15.

Frickey, T. and A. N. Lupas (2004). "Phylogenetic analysis of AAA proteins." J Struct Biol **146**(1-2): 2-10.

Gage, M. J., C. P. Macfarlane, et al. (2004). "Spermatozoal traits and sperm competition in Atlantic salmon: relative sperm velocity is the primary determinant of fertilization success." Curr Biol **14**(1): 44-7.

Goldstein, D. B. and A. G. Clark (1995). "Microsatellite variation in North American populations of Drosophila melanogaster." Nucleic Acids Res **23**(19): 3882-6.

Gomez, A., C. Wellbrock, et al. (2001). "Ligand-independent dimerization and activation of the oncogenic Xmrk receptor by two mutations in the extracellular domain." J Biol Chem **276**(5): 3333-40.

Gu, X., Y. Wang, et al. (2002). "Age distribution of human gene families shows significant roles of both large- and small-scale duplications in vertebrate evolution." Nature Genetics **31**(2): 205-209.

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

Guenet, J. L., C. Nagamine, et al. (1990). "Hst-3: an X-linked hybrid sterility gene." Genet Res **56**(2-3): 163-5.

Gupta, P. K., E. Adamtziki, et al. (2005). "Gene conversions are a common cause of von Willebrand disease." Br J Haematol **130**(5): 752-8.

Habura, A., I. Tikhonenko, et al. (1999). "Interaction mapping of a dynein heavy chain. Identification of dimerization and intermediate-chain binding domains." J Biol Chem **274**(22): 15447-53.

Haddrill, P. R., K. R. Thornton, et al. (2005). "Multilocus patterns of nucleotide variability and the demographic and selection history of Drosophila melanogaster populations." Genome Res **15**(6): 790-9.

Haldane, J. B. S. (1933). "The part played by recurrent mutation in evolution." Am. Nat. **67**: 5-19.

Hammer, M. F., J. Schimenti, et al. (1989). "Evolution of mouse chromosome 17 and the origin of inversions associated with t haplotypes." Proc Natl Acad Sci U S A **86**(9): 3261-5.

Hammer, M. F. and L. M. Silver (1993). "Phylogenetic analysis of the alpha-globin pseudogene-4 (Hba-ps4) locus in the house mouse species complex reveals a stepwise evolution of t haplotypes." Mol Biol Evol **10**(5): 971-1001.

Harrison, P. M., H. Hegyi, et al. (2002). "Molecular fossils in the human genome: identification and analysis of the pseudogenes in chromosomes 21 and 22." Genome Res **12**(2): 272-80.

Hendrickson, H., E. S. Slechta, et al. (2002). "Amplification-mutagenesis: evidence that "directed" adaptive mutation and general hypermutability result from growth with a selected gene amplification." Proc Natl Acad Sci U S A **99**(4): 2164-9.

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

Hughes, A. L. (2007). "Looking for Darwin in all the wrong places: the misguided quest for positive selection at the nucleotide sequence level." Heredity **99**(4): 364-73.

Hyvarinen, A. K., J. L. Pohjoismaki, et al. (2007). "The mitochondrial transcription termination factor mTERF modulates replication pausing in human mitochondrial DNA." Nucleic Acids Res **35**(19): 6458-74.

Ihle, S., I. Ravaoarimanana, et al. (2006). "An analysis of signatures of selective sweeps in natural populations of the house mouse." Mol Biol Evol **23**(4): 790-7.

Jaillon, O., J. M. Aury, et al. (2004). "Genome duplication in the teleost fish Tetraodon nigroviridis reveals the early vertebrate proto-karyotype." Nature **431**(7011): 946-57.

Kauer, M. O., D. Dieringer, et al. (2003). "A microsatellite variability screen for positive selection associated with the "out of Africa" habitat expansion of Drosophila melanogaster." Genetics **165**(3): 1137-48.

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

Kellis, M., B. W. Birren, et al. (2004). "Proof and evolutionary analysis of ancient genome duplication in the yeast Saccharomyces cerevisiae." Nature **428**(6983): 617-24.

Kimura, M. (1983). The neutral theory of molecular evolution. Cambridge [Cambridgeshire] ; New York, Cambridge University Press.

King, S. M. (2000). "AAA domains and organization of the dynein motor unit." J Cell Sci **113 (Pt 14)**: 2521-6.

Li, W. H. (1993). "Unbiased estimation of the rates of synonymous and nonsynonymous substitution." J Mol Evol **36**(1): 96-9.

Li, W. H., Z. Gu, et al. (2001). "Evolutionary analyses of the human genome." Nature **409**(6822): 847-9.

Li, W. H., C. I. Wu, et al. (1985). "A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes." Mol Biol Evol **2**(2): 150-74.

Long, M. and K. Thornton (2001). "Gene duplication and evolution." Science **293**(5535): 1551.

Lundrigan, B. L., S. A. Jansa, et al. (2002). "Phylogenetic relationships in the genus mus, based on paternally, maternally, and biparentally inherited characters." Syst Biol **51**(3): 410-31.

Lynch, M. (2002). "Genomics. Gene duplication and evolution." Science **297**(5583): 945-7.

Lynch, M. and J. S. Conery (2000). "The evolutionary fate and consequences of duplicate genes." Science **290**(5494): 1151-1155.

Lynch, M. and J. S. Conery (2000). "The evolutionary fate and consequences of duplicate genes." Science **290**(5494): 1151-5.

Lynch, M. and J. S. Conery (2003). "The evolutionary demography of duplicate genes." J Struct Funct Genomics **3**(1-4): 35-44.

Lynch, M., M. O'Hely, et al. (2001). "The probability of preservation of a newly arisen gene duplicate." Genetics **159**(4): 1789-804.

Lynch, M. and B. Walsh (1998). Genetics and analysis of quantitative traits. Sunderland, Ma., Sinauer.

Marques, A. C., I. Dupanloup, et al. (2005). "Emergence of young human genes after a burst of retroposition in primates." PLoS Biol **3**(11): e357.

Masly, J. P., C. D. Jones, et al. (2006). "Gene transposition as a cause of hybrid sterility in Drosophila." Science **313**(5792): 1448-50.

Masterson, J. (1994). "Stomatal Size in Fossil Plants: Evidence for Polyploidy in Majority of Angiosperms." Science **264**(5157): 421-424.

Matsuda, Y., T. Hirobe, et al. (1991). "Genetic basis of X-Y chromosome dissociation and male sterility in interspecific hybrids." Proc Natl Acad Sci U S A **88**(11): 4850-4.

Mayr, E. (1942). Systematics and the origin of species, from the viewpoint of a zoologist. New York,, Columbia Univ. Press.

McDonald, J. H. and M. Kreitman (1991). "Adaptive protein evolution at the Adh locus in Drosophila." Nature **351**(6328): 652-4.

Michalak, P. and M. A. Noor (2004). "Association of misexpression with sterility in hybrids of Drosophila simulansand D. mauritiana." J Mol Evol **59**(2): 277-82.

Muller, H. J. (1935). "The origination of chromatin deficiencies as minute deletions subject to insertion elsewhere." Genetics **17**: 237-252.

Muller, H. J. (1942). "Isolating mechanisms, evolution, and temperature." Biol. Symp. **6**: 71-125.

Nachman, M. W. (1997). "Patterns of DNA variability at X-linked loci in Mus domesticus." Genetics **147**(3): 1303-1316.

Nei, M. (1978). "Estimation of Average Heterozygosity and Genetic Distance from a Small Number of Individuals." Genetics **89**(3): 583-590.

Nei, M. and T. Gojobori (1986). "Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions." Mol Biol Evol **3**(5): 418-26.

Nowak, M. A., M. C. Boerlijst, et al. (1997). "Evolution of genetic redundancy." Nature **388**(6638): 167-71.

Ohno, S. (1970). Evolution by gene duplication. Berlin, New York, Springer-Verlag.

Olds-Clarke, P. and L. R. Johnson (1993). "t haplotypes in the mouse compromise sperm flagellar function." Dev Biol **155**(1): 14-25.

Orr, H. A. and S. Irving (2000). "Genetic analysis of the hybrid male rescue locus of Drosophila." Genetics **155**(1): 225-31.

Orr, H. A., J. P. Masly, et al. (2004). "Speciation genes." Curr Opin Genet Dev **14**(6): 675-9.

Orr, H. A. and D. C. Presgraves (2000). "Speciation by postzygotic isolation: forces, genes and molecules." Bioessays **22**(12): 1085-94.

Pamilo, P. and N. O. Bianchi (1993). "Evolution of the Zfx and Zfy genes: rates and interdependence between the genes." Mol Biol Evol **10**(2): 271-81.

Patel, S. and M. Latterich (1998). "The AAA team: related ATPases with diverse functions." Trends Cell Biol **8**(2): 65-71.

Perez, D. E. and C. I. Wu (1995). "Further characterization of the Odysseus locus of hybrid sterility in Drosophila: one gene is not enough." Genetics **140**(1): 201-6.

Phillips, D. M., S. H. Pilder, et al. (1993). "Factors that may regulate assembly of the mammalian sperm tail deduced from a mouse t complex mutation." Biol Reprod **49**(6): 1347-52.

Pickeral, O. K., W. Makalowski, et al. (2000). "Frequent human genomic DNA transduction driven by LINE-1 retrotransposition." Genome Res **10**(4): 411-5.

Pilder, S. H., M. F. Hammer, et al. (1991). "A novel mouse chromosome 17 hybrid sterility locus: implications for the origin of t haplotypes." Genetics **129**(1): 237-46.

Pilder, S. H., P. Olds-Clarke, et al. (1993). "Hybrid sterility-6: a mouse t complex locus controlling sperm flagellar assembly and movement." Dev Biol **159**(2): 631-42.

Presgraves, D. C., L. Balagopalan, et al. (2003). "Adaptive evolution drives divergence of a hybrid inviability gene between two species of Drosophila." Nature **423**(6941): 715-9.

Presgraves, D. C. and W. Stephan (2007). "Pervasive adaptive evolution among interactors of the Drosophila hybrid inviability gene, Nup96." Mol Biol Evol **24**(1): 306-14.

Rice, P., I. Longden, et al. (2000). "EMBOSS: the European Molecular Biology Open Software Suite." Trends Genet **16**(6): 276-7.

Ronquist, F. and J. P. Huelsenbeck (2003). "MrBayes 3: Bayesian phylogenetic inference under mixed models." Bioinformatics **19**(12): 1572-4.

Rosenberg, H. F. (1995). "Recombinant human eosinophil cationic protein. Ribonuclease activity is not essential for cytotoxicity." J Biol Chem **270**(14): 7876-81.

Rousset, F. O. (2008). "genepop'007: a complete re-implementation of the genepop software for Windows and Linux." Molecular Ecology Resources **8**(1): 103-106.

Rozas, J. and R. Rozas (1999). "DnaSP version 3: An integrated program for molecular population genetics and molecular evolution analysis." Bioinformatics **15**(2): 174-175.

Rozen, S. and H. Skaletsky (2000). "Primer3 on the WWW for general users and for biologist programmers." Methods Mol Biol **132**: 365-86.

Samant, S. A., J. Fossella, et al. (1999). "Mapping and cloning recombinant breakpoints demarcating the hybrid sterility 6-specific sperm tail assembly defect." Mamm Genome **10**(2): 88-94.

Samant, S. A., O. Ogunkua, et al. (2002). "The T complex distorter 2 candidate gene, Dnahc8, encodes at least two testis-specific axonemal dynein heavy chains that differ extensively at their amino and carboxyl termini." Dev Biol **250**(1): 24-43.

Sambrook, J. and D. W. Russell (2001). Molecular cloning : a laboratory manual. Cold Spring Harbor, N.Y., Cold Spring Harbor Laboratory Press.

Samonte, R. V. and E. E. Eichler (2002). "Segmental duplications and the evolution of the primate genome." Nat Rev Genet **3**(1): 65-72.

Scannell, D. R., K. P. Byrne, et al. (2006). "Multiple rounds of speciation associated with reciprocal gene loss in polyploid yeasts." Nature **440**(7082): 341-5.

Schartl, A., N. Dimitrijevic, et al. (1994). "Evolutionary origin and molecular biology of the melanoma-inducing oncogene of Xiphophorus." Pigment Cell Res **7**(6): 428-32.

Schlötterer, C. (2002). "A microsatellite-based multilocus screen for the identification of local selective sweeps." Genetics **160**(2): 753-763.

Schlötterer, C. and D. Dieringer (2005). A microsatellite-based multilocus screen for the identification of local selective sweeps based on microsatellite gene diversity. Selective sweep. D. Nurminsky. Boston, Kluwer Academic Publishers**:** pp. 55-64.

Seoighe, C. and K. H. Wolfe (1999). "Updated map of duplicated regions in the yeast genome." Gene **238**(1): 253-61.

Silver, L. M. and K. Artzt (1981). "Recombination suppression of mouse t-haplotypes due to chromatin mismatching." Nature **290**(5801): 68-70.

Slatkin, M. (1995). "Hitchhiking and associative overdominance at a microsatellite locus." Mol Biol Evol **12**(3): 473-80.

Smith, J. M. and J. Haigh (1974). "Hitch-Hiking Effect of a Favorable Gene." Genetical Research **23**(1): 23-35.

Smith, J. M. and J. Haigh (1974). "The hitch-hiking effect of a favourable gene." Genet Res **23**(1): 23-35.

Storchova, R., S. Gregorova, et al. (2004). "Genetic analysis of X-linked hybrid sterility in the house mouse." Mamm Genome **15**(7): 515-24.

Sun, S. (2003). Functional analysis of the hybrid male sterility gene Odysseus in Drosophila**:** x, 99 leaves.

Sun, S., C. T. Ting, et al. (2004). "The normal function of a speciation gene, Odysseus, and its hybrid sterility effect." Science **305**(5680): 81-3.

Tajima, F. (1983). "Evolutionary relationship of DNA sequences in finite populations." Genetics **105**(2): 437-60.

Tajima, F. (1989). "Statistical method for testing the neutral mutation hypothesis by DNA polymorphism." Genetics **123**(3): 585-595.

Tajima, F. (1989). "Statistical method for testing the neutral mutation hypothesis by DNA polymorphism." Genetics **123**(3): 585-95.

Tamura, K., J. Dudley, et al. (2007). "MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0." Mol Biol Evol **24**(8): 1596-9.

Ting, C. T., S. C. Tsaur, et al. (1998). "A rapidly evolving homeobox at the site of a hybrid sterility gene." Science **282**(5393): 1501-4.

Trachtulec, Z., O. Mihola, et al. (2005). "Positional cloning of the Hybrid sterility 1 gene: fine genetic mapping and evaluation of two candidate genes." Biological Journal of the Linnean Society **84**(3): 637-641.

Vyskocilova, M., Z. Trachtulec, et al. (2005). "Does geography matter in hybrid sterility in house mice?" Biological Journal of the Linnean Society **84**(3): 663-674.

Wade, C. M. and M. J. Daly (2005). "Genetic variation in laboratory mice." Nat Genet **37**(11): 1175-80.

Wade, C. M., E. J. Kulbokas, 3rd, et al. (2002). "The mosaic structure of variation in the laboratory mouse genome." Nature **420**(6915): 574-8.

Waterston, R. H., K. Lindblad-Toh, et al. (2002). "Initial sequencing and comparative analysis of the mouse genome." Nature **420**(6915): 520-562.

Watterson, G. A. (1975). "On the number of segregating sites in genetical models without recombination." Theor Popul Biol **7**(2): 256-76.

Wittbrodt, J., D. Adam, et al. (1989). "Novel putative receptor tyrosine kinase encoded by the melanoma-inducing Tu locus in Xiphophorus." Nature **341**(6241): 415-21.

Wolfe, K. H., P. M. Sharp, et al. (1989). "Mutation rates differ among regions of the mammalian genome." Nature **337**(6204): 283-5.

Wolfe, K. H. and D. C. Shields (1997). "Molecular evidence for an ancient duplication of the entire yeast genome." Nature **387**(6634): 708-713.

Wu, C. I. and C. T. Ting (2004). "Genes and speciation." Nat Rev Genet **5**(2): 114-22.

Yang, Z. (1997). "PAML: A program package for phylogenetic analysis by maximum likelihood." Computer Applications in the Biosciences **13**(5): 555-556.

Yang, Z. (2007). "PAML 4: phylogenetic analysis by maximum likelihood." Mol Biol Evol **24**(8): 1586-91.

Yang, Z. and R. Nielsen (1998). "Synonymous and nonsynonymous rate variation in nuclear genes of mammals." J Mol Evol **46**(4): 409-18.

Yang, Z. and R. Nielsen (2000). "Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models." Mol Biol Evol **17**(1): 32-43.

Yang, Z., R. Nielsen, et al. (2000). "Codon-substitution models for heterogeneous selection pressure at amino acid sites." Genetics **155**(1): 431-49.

Zhang, J., H. F. Rosenberg, et al. (1998). "Positive Darwinian selection after gene duplication in primate ribonuclease genes." Proc Natl Acad Sci U S A **95**(7): 3708-13.

Zhang, L., H. H. Lu, et al. (2005). "Patterns of segmental duplication in the human genome." Mol Biol Evol **22**(1): 135-41.

Zhu, Y., D. C. Queller, et al. (2000). "A phylogenetic perspective on sequence evolution in microsatellite loci." J Mol Evol **50**(4): 324-38.

6 Supplement

- Supplement 1: Primer sequences for microsatellite loci
- Supplement 2: Primer sequences for the detection of absence or presence of duplicates
- Supplement 3: PCR and sequencing primers for *Dnahc8*
- Supplement 4: Primer sequences for *Dnahc8* population sequencing in *M. spretus*
- Supplement 5: Primer sequences for q-rtPCR of *Dnahc8*
- Supplement 6: Quantitative real time PCR data

Supplement 1: Primer sequences for microsatellite loci

Gene	Repeat unit	Repeat no.	Primer fwd		Primer rev		amplicon length
ENSMUSG00000005510	CAC	15	5510 fwd F	ATTGCTCGAAGCAAAGAACC	5510 rev F	GAGAGGCGCACAGGAAATAG	401
ENSMUSG00000020163	GAG	10.7	20163 fwd H	TGCCCTGTCAGAGTGTGTTC	20163 rev H	CTGTGCTACACTCGCTCTGG	212
ENSMUSG00000021243	TTG	12.7	21243 fwd T	TGTGGTTTGAGTCCTTTTTGG	21243 rev T	CCTCCTAGCTCCAGGGAGAG	183
ENSMUSG00000021520	TG	24.5	21520-2 fwd T	GCAGGAAGACAGAGGTGAGG	21520-2 rev	GCTACTCAAATCGTCCCAGC	440
ENSMUSG00000021741	GA	22	21741-2 fwd T	AAACAACGTTCTGGCATTTG	21741-2 rev	GCCACAGGCTTTCTGTAAGG	159
ENSMUSG00000021953	ATTT	10	21953 fwd T	GAGAGCCATCAAAAGGCAAG	21953 rev T	TGCAATCCTATGCCCTCTTC	191
ENSMUSG00000022427	AC	16.5	22427 fwd H	GCTAGAGATGTGTGGCGACC	22427 rev H	TCTGTTTCTGGGGCTTGAAC	235
ENSMUSG00000022992	GA	14.5	22992 fwd H	TAGAGCGTGAAGAAGAGGGC	22992 rev H	AAGACTGACCAACATTCACGG	228
ENSMUSG00000024121	ACAT	13	24121 fwd T	AACAATTGGCCTAGTGGCAG	24121 rev T	CTTCTCTTTATCCCCAGCCC	237
ENSMUSG00000024726	TTTA	13.5	24726 fwd T	GCTTGAGTTTGGTCCTCAGC	24726 rev T	CTGTTCCTCTCTGAGCCCTG	239
ENSMUSG00000025156	GTTT	9.3	25156-2 fwd F	TTCCACACAATGAGTGCTGAG	25156-2 rev F	ACTCCATTCTCGCTGGTGTC	305
ENSMUSG00000025245	AAAT	9.5	25245 fwd T	ATGCCACTTGGACATCATGG	25245 rev T	TGAAAGCTGGTGTTGTGAGC	399
ENSMUSG00000026492	AG	19	26492 fwd H	CCCTCTTCCTGCACAGAATC	26492 rev H	AGCATTATTCAAAGGCGTGG	416
ENSMUSG00000026844	GA	22	26844-2 fwd F	CCAGGGAATACCCATGTTTG	26844-2 rev	AGGTCACCTGTCCAGTGAGG	369
ENSMUSG00000027081	TC	12	27081 fwd T	GGTTCCCAGTTGGTGATCTG	27081 rev T	GGGATGGTTTACAGTGCAGG	230
ENSMUSG00000029633	CA	12	29633-2 fwd T	TGGGAAGATGGGTAGTCCTG	29633-2 rev	TGAAGCAGGAGGACATTGTG	386
ENSMUSG00000029918	AC	11.5	29918 fwd H	ATGTCTTGGAACCAACTCCC	29918 rev H	TATGCCTTCCTGGGACACTC	427
ENSMUSG00000032889	TG	16.5	32889 fwd F	TCCACTTGAGGCACTGTCAC	31146 rev F	ATTATGAATGGGGGTAGGGC	241
ENSMUSG00000034102	ATAG	14.5	34102 fwd F	ATAAAGCCTCATGCCCACAG	32889 rev F	AGCCAACATGTAGGGACTGG	396
ENSMUSG00000034499	GT	17	34499 fwd H	CTCCGTTCTCTCATTGGAGC	34499 rev H	TGAGGTCAATCCCTGGAATC	160
ENSMUSG00000035253	AAAC	11	35253 fwd H	ACAGAGACCAAAGCCCAGTC	35253 rev H	TGGTTTACAAAAGCAGGCAG	231
ENSMUSG00000035692	TG	17.5	35692-2 fwd H	TCAGTGTTTGGAAATGACTTGC	35692-2 rev H	GCCTAGTTGGCCATCACTG	247
ENSMUSG00000035796	AC	20.5	35796 fwd H	TCCCTGGATCACACTTGTTG	35796 rev H	TCAAAGTGTGGGTGTGTTGG	444
ENSMUSG00000036427	TG	20.5	36427 fwd T	AGATGCTTTCCATGTCTGGC	36427 rev T	TCCGAATAACTGAAAGAGAGCC	244
ENSMUSG00000037359	GTTT	11	37359 fwd F	TGGTTTACAAAAGCAGGCAG	37359 rev F	ACAGAGACCAAAGCCCAGTC	231
ENSMUSG00000038323	TG	11.5	38323 fwd H	TTATCCAGGTCTGGGCTGTC	38323 rev H	GGAGAGAAGTTCAAGGCCAG	442
ENSMUSG00000039078	TATC	17.8	39078 fwd T	TGAGATGGGATCTTGCTGTG	39078 rev T	TGTGCAAAGGCGTAAGTTTG	196
ENSMUSG00000040429	TTTG	9.8	40429 fwd T	TGTCCATCTCCTTACCAGCC	40429 rev T	AGAGAATGTTCAAGCCCACC	184
ENSMUSG00000040621	TG	21.5	40621 fwd H	CAGCCCTTTCTCGTTTCTTG	40621 rev H	ATGTAGCCCGTGCAGAACTC	184
ENSMUSG00000041274	TG	19	41274 fwd T	GTGAGAGGTGCAGAAGAGGC	41274 rev T	GAGGAATGCCAGACAGGAAG	180
ENSMUSG00000041418	AGAT	15.3	41418 fwd T	TGTCTACCAGGAAAAGCCAAG	41418 rev T	TATTGGCCTGGAGTAAGCAC	366
ENSMUSG00000043062	GT	16	43062 fwd F	GAACAAGGGGATGAATGGTC	43062 rev F	GCCACACCCACACCTTTTAC	435
ENSMUSG00000043192	AGT	19.3	43192 fwd T	TGAATGACTTCCCCACAATG	43192 rev T	TTTCAACTCTGTGGAGGTGG	353
ENSMUSG000000	TAC	13.7	44111 fwd	ATGGGGGACTTTTGGG	44111 rev	TTTCCTAATTTCTCACTTT	177

44111			H	ATAG	H	CTGAGG	
ENSMUSG00000044792	GT	15.5	44792 fwd H	TGCTCTCCTGGACAACAGTG	44792 rev H	CCAAAGAGCAAGACAAAGCC	387
ENSMUSG00000044795	AT	10.5	44795 fwd F	GAAGCTCCTGCTCTGACCTG	44795 rev F	GAGAGCGCCCTCAGTAATTG	413
ENSMUSG00000045122	AGAC	11.5	45122-2 fwd T	CTCCCTTCTGAGGTGACTGC	45122-2 rev	TGGCCTGTCCTGGAACTTAC	354
ENSMUSG00000045609	TAGA	10.8	45609 fwd H	GCCTTCAGGAAGGTACCAAG	45609 rev H	ATTGTGAAAGTCAGGGCCAC	422
ENSMUSG00000046153	CA	18	46153 fwd H	CCCTGTATTTAATGCCCAGC	46153 rev H	GCCAGCCTAAAACTGGTGAG	213
ENSMUSG00000046687	TG	12	46687-2 fwd F	ATATGGTGTGTGGCCCTGTC	46687-2 rev	GGCTCAATGGCTAAGAGCAC	242
ENSMUSG00000046836	TG	11	46836 fwd H	TTTACATTCCGAGCAGGGAG	46836 rev H	TTCTGCAAGTTGAGCTGTGG	200
ENSMUSG00000046849	GGAA	12	46849 fwd H	GACAAGAAGGAGAGGGGGAG	46849 rev H	ATCTGGCCACTTTGCTGAAG	398
ENSMUSG00000047016	AC	20	47016 fwd F	CAATTGTGTGGGAAGCAAAG	47016 rev F	TGTTGGGAAGATAACTGCTGG	197
ENSMUSG00000047509	TCC	13.7	47509 fwd T	TGTCCATCACAGCGCTCTAC	47509 rev T	GAAAAGCAGCCTACGTCACC	236
ENSMUSG00000047776	AGAT	16.5	47776 fwd H	GTCTAGCACATGTGGGGCTC	47776 rev H	CCCTAGAGATCCCAAGACCC	419
ENSMUSG00000048989	CT	13.5	48989 fwd F	GTGGTCTGAAAGAGAACGGC	48989 rev F	ATTTGGACTCTGTGGGCTTG	370
ENSMUSG00000049916	AC	19	49916-2 fwd T	GCATGTACAAGGCATAGGGC	49916-2 rev	AATCTGTCATTCCTTCCCCC	187
ENSMUSG00000050043	AC	19.5	50043 fwd T	TCTCATGTCTCTGACCTCCC	50043 rev T	CCTCATCATGCCAATCACAC	434
ENSMUSG00000050383	AAAC	11	50383 fwd F	TGCCAATATTTAAGGAGGGC	50383 rev F	TTACATCCACGGAGACCCTG	218
ENSMUSG00000051116	TG	18.5	51116 fwd T	GGTATTGCTGGACCTTCTGG	51116 rev T	AACCCCATCTCCTCACTTCC	224
ENSMUSG00000053178	AC	19	53178 fwd F	GAGGATCCTGATGTTCTGGG	53178 rev F	AGGATGGCATCTTCCTTGC	377
ENSMUSG00000053740	TG	10	53740 fwd T	CTTCTCAACGAAAGCCCAAG	53740 rev T	TTGCTGCTCAAAGGTAAGGG	443
ENSMUSG00000056369	ATTT	11	56369 fwd T	TTATTGGCCCTCTGTTGGAC	56369 rev T	TACCTTGAATGCACTGGCAC	410
ENSMUSG00000056663	TGTC	10.5	56663 fwd H	TCTCACTGTGGCACTTTTGC	56663 rev H	GCTCAACCACCTGTAGCCTC	235
ENSMUSG00000056815	TGT	10	56815 fwd F	TCTCTGTAAGAAGCGCCTGC	56815 rev F	TATGGCAGTTGCTCACCTTG	195
ENSMUSG00000056836	ACA	9	56836 fwd F	TTCTGGCCATCATAAGGCTC	56836 rev F	AAAGTCCCAAGTTCTGCTGG	155
ENSMUSG00000057202	AC	15	57202 fwd T	TTTCCATGCTGATTGTGGTG	57202 rev T	TGCTTTTAACTCAGGGGTGG	434
ENSMUSG00000057744	AAAC	10.8	57744 fwd H	TGGTGCTTGCTTTTAATCCC	57744 rev H	GGAATGCCAGCTCTTCAGAC	434
ENSMUSG00000057762	GAAA	17	57762 fwd F	AATGGCTCACAGAGTCAGGG	57762 rev F	TATTTCCAGCCTTTGGCAAC	384
ENSMUSG00000058676	AT	12	58676 fwd H	GTCCTTCCCCATTCTCCTTC	58676 rev H	TGCCACGTGGTGATTAGAAC	385
ENSMUSG00000058805	AAAC	10.8	58805 fwd H	CTCTGAGTTCAAAGCCAGCC	58805 rev H	TTCTCACGTGGTGTGAATCC	240
ENSMUSG00000059070	CA	15.5	59070-2 fwd H	GGAAAAGGTTTAGGGCCATC	59070-2 rev	TATCTCTGCACATTGCCTGC	242
ENSMUSG00000059422	ATAC	11.5	59422 fwd T	AACAAACCCTCACTGCCAAC	59422 rev T	CAAGAGCATGAGAAGGAGGC	423
ENSMUSG00000059792	AGAC	20.8	59792 fwd T	CCAGATGTGGACCATTAGCC	59792 rev T	GGTGACAGATGGTAGGGTGG	406
ENSMUSG00000060552	ACAT	9.3	60552 fwd F	AAGGGCCTCTGGCTCTCTAC	60552 rev F	TCCTTTTCCCAAGGAATCAG	372
ENSMUSG00000060628	CA	11.5	60628 fwd F	CTGCCCTGACTTCAAACCTC	60628 rev F	AACTGGATCCAAATTCAATTCC	248
ENSMUSG00000061518	TG	10.5	61518 fwd F	TGGAGCATGTGAGTCTGGTC	61518 rev F	TATGAAGAAGATGGCTCCCG	399
ENSMUSG00000062884	AC	11.5	62884 fwd T	AAGCAAAACTGGACTCGCTC	62884 rev T	CAGGTTATTCCCATGGTTTTG	222

Supplement 2: Primer for the absence or presence detection of duplicate genes

gene	fwd primer		rev primer		rev K+		presence	absence
ENSMUSG00000047016	ap-47016-fwd	TTAAGGACGGGAAGCGAAC	ap-47016-int	CCACTGTAGCTGGCTGCATA	ap-47016-rev	GACAAGTCCCCAAACAGATGA	202	283
ENSMUSG00000056369	ap-56369-fwd	AGACCAGGGAGGGGAAACTA	ap-56369-int	AACACGTCCTGGTCTCCAAG	ap-56369-rev	GGAAAGACAGAGGGAGGAAGA	333	460
ENSMUSG00000052825	ap-52825-fwd	CGCTGATGCAAGAGCCTAGT	ap-52825-int	GCGTCCACGTCTTTTTCTTC	ap-52825-rev	GAAATGGTCAGGGCACACTT	287	343
ENSMUSG00000045122	ap-45122-fwd	AAAGGCTCCCCTGCTTTTAG	ap-45122-int	ATTCGGGGTTGTTCTTGATG	ap-45122-rev	CAGCACCCAGTTGAAGAAAA	309	378
ENSMUSG00000021741	ap-21741-fwd	CCCATTGGACTTGCAAACTT	ap-21741-int	CCACTTTCGAAAACCTTCCA	ap-21741-rev	GGAATATTTTCACCATGCCACT	264	324
ENSMUSG00000044111	ap-44111-fwd	TTGCTCCAGTAAGCACTTGG	ap-44111-int	CTGGCCACTTCACCAAAGAT	ap-44111-rev	GCAGGCTGTGAGATACACTCC	255	335
ENSMUSG00000060389	ap-60389-fwd	ATGCCCGATTAAATGCAAAT	ap-60389-int	TGTGGCTTCACTACCTGCAC	ap-60389-rev	CAAATCAGTGAATCCAAGAAGG	250	326
ENSMUSG00000046687	ap-46687-fwd	TCTAGGGAAAGCCGTTCTGA	ap-46687-int	TCCCTGGAGTCCGTGTAGTC	ap-46687-rev	GCTCCTCCCCATCCTTTTT	207	306
ENSMUSG00000047509	ap-47509-fwd	CCTTACGGGAGGGGTATGAT	ap-47509-int	GGTGAAAGTCTCCTCCACCA	ap-47509-rev	CTGCACTGAATGTGGATTGC	203	345
ENSMUSG00000037359	ap-37359-fwd	AATGGCCTGGGATTAAAGGT	ap-37359-int	CCAGTAGCCTGCTCCTCATC	ap-37359-rev	TGCAGTCATTGAAAACCTCCT	223	282
ENSMUSG00000046849	46849-fwd	AACACAAATAAATCAGTAGCCTTCC	46849-rev	TCAGACTGAGCATTCGCTTC	46849-K+	AATTCTGCCAATCCATGAGC	153	448
ENSMUSG00000046153	46153-fwd	CCTGTGTGAGACAGCAACAG	46153-rev	AAGAGCCCACGACAACAGAC	46153-K+	CAGAGCTGCTGACAACGAAG	160	529
ENSMUSG00000059422	59422-fwd	CACCCACCCATCTGTACCTG	59422-rev	GAGAACTGCACCTCCACGTC	59422-K+	TGCAGACTCGCAATAGCTTC	214	320
ENSMUSG00000035796	35796-fwd	CAGGCAGATTTCTGAGTTCG	35796-rev	CACCAACCTGCCCTAGTAGC	35796-K+	CCTGAGCGGGAGAAGTTTAG	175	328
ENSMUSG00000060552	60552-fwd	TTTTCTGGAGTACAGCTTCTTGAG	60552-rev	CAGAATTGACACCTTGTGCAG	60552-K+	CTCACTTCCCTGGAGTCTGG	240	395
ENSMUSG00000027081	27081-fwd	TTGCCTTGGTTATGGTGTTC	27081-rev	AGCAATCAGAGGCGCAAG	27081-K+	TGTTGCTGGGCTACAGAGTG	196	285
ENSMUSG00000056836	56836-fwd	GAACAATTGCCGGAGAGTTC	56836-rev	CTTTACTGCATTGCCACCAC	56836-K+	GGAATGCACATTGTGGAAAA	154	194
ENSMUSG00000020163	20163-fwd	TGGGATATGATGGGAGAAGG	20163-rev	ATGTAAGGCACCCAGTCCAG	20163-K+	CCTCCCTATGTGTGCATGTG	167	364
ENSMUSG00000043062	43062-fwd	TGAACAAATAAGGCGTCTGC	43062-rev	GCGTTCTTGTAGGGCTGTTC	43062-K+	CCCCATGTTGTTCTCACTTG	170	344
ENSMUSG00000047168	47168-fwd	TACTGGGCTTGCTTCTGACC	47168-rev	GGCGTTGGTATCTGCATAGG	47168-K+	GCCACATCAGACGGAGAAAT	169	119
ENSMUSG00000049635	49635-fwd	AACCAACCAACCACTTCTGC	49635-rev	CTGCTTGGCCTCTTTCTCTG	49635-K+	CAGCCTCCTGTGCCATAAAT	197	561
ENSMUSG00000021001	21001-fwd	ATGATGCTGAGAGGAACTGG	21001-rev	TCTTCATCTGGCTCCGATTC	21001-K+	CATGGACAGGACAAAGACGA	150	303
ENSMUSG00000055936	55936-fwd	CAGCCTGAAGGTGTATCAGC	55936-rev	CTGGGACTTTCGTGGTCTTC	55936-K+	CTTCATCCTCCTCACCGAAA	179	536
ENSMUSG00000036258	36258-fwd	AGGAGGAGGAGGAGGAGAGG	36258-rev	TGGAAGAGGTGCCAGAGC	36258-K+	CCACCCTCACTGGATGTGTT	250	350
ENSMUSG00000020424	20424-fwd	CGCCTCCTGCCTAGGTCTC	20424-rev	GCAGGAAGAGCAGCTTGATG	20424-K+	GGACCAGGGACTTGTTCTGA	152	134
ENSMUSG00000006270	06270-fwd	GAGCTTAGGTGCTGCGTTG	06270-rev	CGGACGTCCATTTTGTCTG	06270-K+	AGAGCAAAACAGCCCGAGTA	151	321
ENSMUSG00000057777	57777-fwd	GAGTGAAATTGGGGCTTCAG	57777-rev	GTAAACCAGCTTGGCCTGAG	57777-K+	GTGACAATCACGCAAGCAAC	159	508
ENSMUSG00000032058	32058-fwd	AGCCGGCTGATTGGATTTAG	32058-rev	GAGCGAATCGTCTCCATCTC	32058-K+	CATCCTCCTGCCTCAGTCTC	195	350
ENSMUSG00000049026	49026-fwd	AGGGGCACATATCAGTCTGG	49026-rev	AAACTCCATGGCCTTTTTCC	49026-K+	GCGAACAGCTAAGGGATGAG	186	243

Supplement 3: PCR and sequencing primers for *Dnahc8*

PCR and sequencing primers

Name	Sequence	PCR product size	Positon in transcript
1 fwd	ATCTTACTCCACAACCCCGC	2916	17
1 rev	TCCAAAACGGACTCAACCTC		2913
2 fwd	CCCAAAATGAAGAAGGTGGA	2940	2827
2 rev	CTTTGGCCATAACGGGAGTA		5747
3 fwd	ACCATTCAGCCACATCTTCC	2875	5629
3 rev	TGAAGTCTCCAGGCTTGTCC		8485
4 fwd	ATCAACGAATGGGGAGATCA	3054	8410
4 rev	CCTTGTCACCCACTTTCACC		11444
5 fwd	CCACAAATATTTCCGCACAC	3091	11305
5 rev	AGCACCTATCAGGCAGGAGA		14376

Sequencing primers

Name	Sequence	Position in transcript	Remarks
1.1 rev	CCAACCCAAGCTTTTCTCCTAG	641	
1.2 fwd	GCAAGATTTAGAGAAGCAAGGG	570	
1.2 rev	GAAGGTGGACAATATCGACTTCTC	1171	
1.3 fwd	AGCTTTAAACCAGTCCAAGCAG	1048	
1.3 rev	TGTCACCTTGATGAATAATGACG	1677	
1.4 fwd	GTGTCAGCCCCTCTATAACTACG	1558	
1.4 rev	CCTGCTGGGAGGACAAGAT	2180	
1.5 fwd	CATCTACCAGGGCATTAAGAAG	2035	
1.5 rev	CGTCCAGCTTCATCTTTATCA	2637	
1.6 fwd	ACAGAGCCTGGTTCAAGGAG	2535	
2.1 rev	GGAAACGATGTCCTCTGACC	3450	
2.2 fwd	ATGGAGGTCGACACCAATGA	3287	
2.2 rev	TCTCAGTCAGGGAAGGGTTG	3907	
2.3 fwd	AAGCTGGTGCTCCTCCTGT	3773	
2.3 rev	ACCTTCACCGACTCAAGCAG	4412	
2.4 fwd	GAGTCCGAAGGCGTTGAC	4304	
2.4 rev	TATCCGATTCCACATCGAAC	4939	
2.5 fwd	CAGGCGTTTCTGGACCTC	4811	
2.5 rev	CTGATCACGTTGGGGTTCTC	5432	
3.1 rev	GTGCATCCCAGGAACTCGT	6252	
3.2 fwd	GACTTCGAGTGGCTGAAACAG	3158	
3.2 rev	GCGCCAAGATAACATTTTCAA	6795	
3.3 fwd	TCCTAACCATGAACCCTGGAT	6660	
3.3 rev	CCTTTGGATTCATTCGCATC	7304	
3.4 fwd	AGCGGTTCTGGGAAGACG	7211	
3.4 rev	CAGCTTGTGGAGGTGGTC	7853	
3.5 fwd	AGTCTCTCAATCTCCTGGAAGG	7791	
4.1 rev	AGTGGGCTCCGGCATATCT	9043	
4.2 fwd	GGTTTATAACTCCGGATGACGA	8919	
4.2 rev	TATCAGGCCTTGGGTGATTT	9555	
4.3 fwd	GATGAGGCGTTCCTGGAATA	9473	
4.3 rev	ACCGCCAGTTCCTTCTCCT	10095	
4.4 fwd	ATCAATGAACAAGCGGAACG	10001	
4.4 rev	GCAAGTGTCCAGGAAAGCAG	10631	
4.5 fwd	GACACAATCAACGAGGAAACTG	10529	

4.5 rev	GCAGAACATCGCCTACGAGT	10963	
4.6 fwd	AATGGACCTGCTCAATGATG	10831	
5.1 rev	CCATGGACTGGTCAAATAACTTC	11962	
5.2 fwd	CAATACAGCTCAGGAGGAGTTC	11845	
5.2 rev	TTTAGGATAACTGGCTCCGTGT	12489	
5.3 fwd	AGACGCCCCAGAGGAAGAG	12343	
5.3 rev	ACGGTGGAGTGCAGAAATG	12984	
5.4 fwd	AAGGACATTCGCTGGGATCA	12904	
5.4 rev	ATCAAGCGAGCTTTCACCTC	13541	
5.5 fwd	AAGAGAGCGGTGGTGGTGT	13443	
5.5 rev	GAAGCTGCGTGAAGAGAACC	14092	
5.6 fwd	TCAGCTTTGGAAGAGGGTGT	13741	
1 fwd b	CGTCTGAGTATCTTACTCCACAACC	9	alternative to 1 fwd
1 fwd C	ATGGAGTCTGAGGAAGGCAA	101	alternative to 1 fwd
1.2 fwd B	ATCCCAAACCTCCAGGAGAC	533	alternative to 1.2 fwd
1.2 rev B	GGCACTTCTCCTTCCTGGTA	761	alternative to 1.2 rev
3.4 rev B	GTGCTCCCAGTCACCGTAGT	8025	alternative to 3.4 rev
KTH 5.5 fwd	CGCGAGGCCATTGTCTACAG	13472	fits *M. caroli* sequence
KTH 5.5 rev	CTTCATTGTGGATGGTCACA	13948	fits *M. caroli* sequence

Supplement 4: Primer sequences for *Dnahc8* population sequencing in *M. spretus*

***M. spretus* population sequencing**	
Exon 2	710 bp
dnahc8pop-2-fwd	ATGGAGTCTGAGGAAGGCAA
dnahc8pop-2-rev	GACGGGCAATCCAAAATTAG
Exon 50/51	630 bp
dnahc8pop-51-52-fwd	ATATGAGACCTCTCTGGTACGG
dnahc8pop-51-52-rev	CCTGTAATATCGGCCTCCAG
Exon 66	373 bp
dnahc8pop-66B-fwd	ACCCCTTTCTCCATCTGACC
dnahc8pop-66B-rev	CTCCAGTTCCAGCAGCTACC
Exon 79	519 bp
dnahc8pop-79-fwd	CTGGTGGACGACGAGTCTCT
dnahc8pop-79-rev	GGCCATGGACTGGTCAAATA

Supplement 5: Primer sequences for q-rtPCR

Primerset	Primer forward	Sequence forward	Primer reverse	Sequence reverse	Product size	Position on transcript
5.2 fwd / 5.1 rev	5.2 fwd	CAATACAGCTCAGGAGGAGTTC	5.1 rev	CCATGGACTGGTCAAATAACTTC	140	11845
affy 2	rt affy 2 fwd	GAGGCCTAACGTGTTCTGGA	rt affy 2 rev	GCAGGACTTCATTGTGGATG	131	13843
hst6.7 2	hst6.7 2 fwd	GCACACGGGATACAGAACCT	hst6.7 2 rev	GCTTTGCAGGCTGTTACCAT	131	1595

Supplement 6: q-rtPCR data

species	tissue	delta ct
M. caroli	spleen	11.5052
M. caroli	Testis	0.110946
M. caroli	muscle	13.2431
M. caroli	lung	19.02203
M. caroli	liver	17.07246
M. caroli	kidney	12.00723
M. caroli	brain	12.30637
M. caroli	heart	16.40481
M. m. castaneus	spleen	5.570852
M. m. castaneus	Testis	-3.42652
M. m. castaneus	muscle	6.122243
M. m. castaneus	lung	14.94272
M. m. castaneus	liver	11.41355
M. m. castaneus	kidney	10.40619
M. m. castaneus	brain	7.617178
M. m. castaneus	heart	11.2407
M. m. domesticus	spleen	4.992377
M. m. domesticus	Testis	-1.58354
M. m. domesticus	muscle	7.963969
M. m. domesticus	lung	12.75827
M. m. domesticus	liver	11.01173
M. m. domesticus	kidney	12.13095
M. m. domesticus	brain	11.20375
M. m. domesticus	heart	12.65582
M. macedonicus	spleen	9.496099
M. macedonicus	Testis	-0.06297
M. macedonicus	muscle	10.92125
M. macedonicus	lung	17.73057
M. macedonicus	liver	15.53936
M. macedonicus	kidney	15.22774
M. macedonicus	brain	13.10955
M. macedonicus	heart	15.62266
M. m. musculus	spleen	-2.64279
M. m. musculus	Testis	-2.54107
M. m. musculus	muscle	4.368729
M. m. musculus	lung	13.93138
M. m. musculus	liver	11.55081
M. m. musculus	kidney	11.4616
M. m. musculus	brain	8.918091
M. m. musculus	heart	10.55628
M. spicilegus	spleen	8.49556
M. spicilegus	Testis	-0.16851
M. spicilegus	muscle	9.892725
M. spicilegus	lung	16.49241
M. spicilegus	liver	15.66141
M. spicilegus	kidney	14.48556

M. spicilegus	brain	13.4639
M. spicilegus	heart	14.97264
M. spretus	spleen	8.149102
M. spretus	Testis	-0.32368
M. spretus	muscle	10.45092
M. spretus	lung	16.47874
M. spretus	liver	9.448132
M. spretus	kidney	6.077386
M. spretus	brain	12.77396
M. spretus	heart	14.44263

6.1 Digital supplement

- Chapter 1
  - List of all duplicates with K_S, K_A, and K_A/K_S
  - List of all young duplicates with detail information
  - Microsatellite data
    - Microsatellite Analyzer input and ouput files for (sub)species and population data
    - Raw microsatellite data, .rsd files

- Chapter 3
  - Sequences *Dnahc8*
    - Full gene sequences
      - Assemblies, Codon Code aligner format
      - Sequences, fasta format
    - Population samples
      - *M. domesticus* whole gene
        - Assemblies, Codon Code aligner format
        - Sequences and alignments, fasta format
      - *M. spretus*
        - Assemblies, Codon Code aligner format
        - Sequences and alignments, fasta format
  - Quantitative real time PCR raw data

Erklärung

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

Till Bayer

Köln, den 5.12.2008