Evolution der Genexpression am Beispiel zweier Subspezies der Hausmaus und populationsgenetische Analyse eines jungen Gens

Inaugural – Dissertation

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

vorgelegt von
Fabian Staubach
aus Bergisch Gladbach
Köln, 2009

Berichterstatter:	**Prof. Dr. Diethard Tautz**
	Prof. Dr. Thomas Wiehe
Tag der letzten mündlichen Prüfung:	**25.05.2009**

Danksagung

Mein besonderer Dank gilt Prof. Dr. Diethard Tautz, der mir die vorliegende Arbeit ermöglichte. Durch seine Kreativität, Diskussionsbereitschaft und seinen Blick für das Wesentliche, aber auch durch seine Menschlichkeit war er maßgeblich für den Erfolg dieser Arbeit und mir stets ein wichtiges Vorbild. Ich bedanke mich besonders für das unerschütterliche Vertrauen in mich und meine Arbeit.
Ich bedanke mich bei Prof. Dr. Thomas Wiehe für die Übernahme des zweiten Gutachtens. Ich danke ihm außerdem insbesondere für eine ausführliche Diskussion über die Varianz von Pi und Theta. Seine Begeisterung, mathematische Probleme kreativ und systematisch zu lösen, ist ansteckend und war mir sehr hilfreich.
Ich bedanke mich herzlich bei Prof. Dr. Siegfried Roth für die Übernahme des Prüfungsvorsitzes und seine schnelle und freundliche Zusage.
Dr. Tobias Heinen danke ich für die Übernahme des Beisitzes und manche Lehrstunde im Labor. Außerdem danke ich ihm für kölsches Liedgut und seine Freundschaft. Dr. Jochen Wolf hat mir den Eintritt in die Welt der Statistik erleichtert und unzählige Stunden seiner Zeit geopfert, um sich mit mir den Kopf über meine Daten zu zerbrechen. Dafür möchte ich ihm herzlich danken. Dr. Till Bayer danke ich für das Korrekturlesen und so manches kleine Perlskript, er ist ein zuverlässiger Freund. Ich danke Dr. Meike Teschke und Dr. Kathryn Stemshorn für ihr offenes Ohr und Schokolade. Dr. Bettina Harr danke ich für ihre Diskussionsbereitschaft und ihre kritischen Kommentare, die meine Arbeit mit vorwärts gebracht haben. Ich danke Dr. Chris Voolstra für Tipps, Tricks und seine Wohnung. Ich danke Prof. Bernhard Haubold und Prof. Göran Kauermann für ihre Ratschläge in der Datenauswertung. Dr. Arne Nolte danke ich für einen Haufen Dickdorsche. Dr. Leslie Turner und Dr. Guy Reeves danke ich für ihre offene Art, auch den wissenschaftlichen Gedankenaustausch zu pflegen. Ich danke dem Kochclub, insbesondere einer unerschütterlichen Christine Pfeifle, die sich mit mir durch Kasachstan geboxt hat. In diesem Zusammenhang danke ich auch Katya Shabanova, die den Kontakt nach Kasachstan hergestellt hat. Ich danke den gesamten Abteilungen Evolutionsgenetik, Evolutionsökologie und unseren noch frischen Theoretikern am MPI in Plön für eine angenehme Arbeitsatmosphäre. Ich freue mich an dieser Stelle besonders die Mitglieder meines Labors herausheben zu dürfen, allen voran Silke Carstensen, die neben viel technischer Hilfe für mich auch ein Quäntchen Kölsch Bloot hat. Aber auch Jan von Rönn und Elke Bustorf trugen wesentlich zu einer herzlichen und offenen Atmosphäre bei.
Meinem Bruder Daniel danke ich für nächtliche Chats am Rande des Abgrunds und meinem Bruder Simon für die großen Lebensweisheiten. Beiden danke ich, dass sie so ausgezeichnete Brüder sind.
Ich bedanke mich bei meinen Eltern bei meiner geliebten Freundin Sybil für ihre Unterstützung.

Zusammenfassung

Seit mehr als 50 Jahren wird vermutet, dass die Evolution der Genregulation eine wichtige Rolle für die Evolution der Organismen spielt. Allein Unterschiede in proteinkodierenden Regionen der Gene erscheinen unzureichend, um die Vielfalt des Lebens umfassend zu erklären. In jüngerer Vergangenheit werden immer mehr Beispiele für regulatorische Evolution gefunden. Wie wichtig regulatorische Evolution im Verhältnis zur Evolution kodierender Bereiche ist, und ob sie neutral evolviert oder in erster Linie durch die Wirkung positiver Selektion geformt wird, ist Gegenstand lebhafter Debatten. In der vorliegenden Doktorarbeit wird anhand zweier Subspezies der Hausmaus untersucht, ob die Genregulation der Erwartung eines neutralen Evolutionsmodells und damit dem Prinzip der molekularen Uhr folgt. Die neutrale Erwartung wäre, dass die Divergenz der Genexpression von der Mutationsrate, die zu Veränderungen der Genexpression führt, abhängt. Um die Mutationsrate abzuschätzen, wird in der vorliegenden Studie die Variabilität innerhalb einer Subspezies der Hausmaus herangezogen (Genexpressionspolymorphismus). Es wird eine Methode entwickelt, um die gemessene Varianz der Genexpression in den Subspezies der Hausmaus soweit von technischen Messeffekten zu befreien, dass die echte, biologische Variabilität zugänglich wird. In diesem Zusammenhang wird auch eine neue Methode zur Normalisierung von Quantitative Real Time PCR-Experimenten entwickelt, die von der Verwendung einzelner Referenzgene unabhängig ist und so erhöhte Zuverlässigkeit bietet. Mit diesen neuen Instrumenten wird die Genexpression von 24 Genen, die in einem vorangegangenen Microarrayexperiment als zwischen den Subspezies differentiell exprimiert klassifiziert wurden, untersucht. Es wird gezeigt, dass die Divergenz der Genexpression tatsächlich vom Expressionspolymorphismus abhängt und damit ein neutrales Modell der Genexpressionsevolution gestützt. Eine Korrelation des Sequenzpolymorphismus stromaufwärts der untersuchten Gene mit dem Polymorphismus der Genexpression legt ähnliche Evolutionsmechanismen von DNA-Sequenz und Genexpression nahe. Additivität und Kontinuität werden als Grundlagen der Evolution der Genexpression bestätigt. Es werden keine Hinweise auf ein vermehrtes Auftreten positiver Selektion unter den zwischen den Subspezies unterschiedlich exprimierten Genen gefunden. Interessanterweise ist die Mehrzahl der

Expressionsunterschiede zwischen den Subspezies gewebespezifisch (zehn von zwölf).

Die genetische Variabilität der Stromaufwärtsregion ist für eines der Gene (Poldi) in der östlichen Hausmaus stark reduziert. Weiterführende, in dieser Studie erhobene populationsgenetische Daten zeigen, dass die genomische Region, welche Poldi enthält, die Signatur eines rezenten Selektionsereignisses trägt. Mögliche Ursachen für das Selektionsereignis konnten identifiziert werden. Interessanterweise ist Poldi ein Orphan Gen: Trotz ausgeprägter Syntänie zu Ratte und Mensch entsteht in diesen Spezies kein homologes Transkript. In einer früheren Arbeit konnte das erste Auftreten des Transkripts im Genus *Mus* auf einen Zeitraum vor etwa zwei Millionen Jahren datiert werden. In dieser Arbeit wird anhand von Sequenzdaten verschiedener Mausspezies eine Mutationen am 5’ Ende des ersten Exons identifiziert, die mit dem ersten Auftreten des Transkripts korreliert. Eine unabhängige Mutation an einer spleißrelevanten Position am 3’ Ende des ersten Exons legt einen sekundären Verlust des Transkripts in *Mus spicilegus* nahe.

Abstract

Already more than 50 years ago a role for the evolution of gene regulation in the evolution of organisms has been proposed. Differences in protein coding regions alone seem insufficient to explain the diversity of life. In the recent past examples of regulatory evolution are accumulating. The relative importance of regulatory evolution as compared to protein changes and the question whether gene expression evolves neutrally or under a selective regime is a matter of ongoing discussion. In the study at hand we approach these questions using two subspecies of the house mouse as a model. We test whether the evolution of gene expression follows the predictions of a neutral model, mainly a molecular clock model. The expectation is that divergence in gene expression depends on the mutation rate. Variability within a subspecies (gene expression polymorphism) is used to estimate the mutation rate. To assess gene expression polymorphism two methods are developed in this study: A new standardization method for qRT-PCR, which provides independence from single reference genes and a method to purify the measured expression variance from technical effects in a way, that allows the biological variance between individuals to be estimated. These new methods are applied to qRT-PCR data of 24 genes, which have been proven to be differentially expressed between the two subspecies in a previous microarray experiment. Indeed I can show, that the divergence of gene expression depends on gene expression polymorphism, compatible with the neutral expectation. A correlation between upstream sequence polymorphism and gene expression polymorphism supports a similar model for the evolution of sequence and gene expression and indicates additivity and continuity of small effects. No evidence for more frequent selection on the differentially expressed loci can be detected on the sequence level. Interestingly ten out of twelve expression differences detected by qRT-PCR are tissue specific.

The upstream region of one of the investigated genes (Poldi) shows a strong reduction in genetic variability in the eastern house mouse compared to the western house mouse. Additional population genetic data of the whole genomic region surrounding Poldi indicates recent positive selection acting on this locus. Putative mutations underlying the selective event can be indentified. Interestingly Poldi turns out to be an orphan gene. Despite distinct synteny to rat and human no homologous transcript exists in these species. In a previous study the first appearance of the transcript within

Mus was dated about 2mya. Due to sequence analysis throughout the whole genus I was able to identify mutations correlating with the presence of the transcript.

1 Einleitung

Im Rahmen dieser Dissertation werden zwei interessante, schon lange bestehende aber dennoch äußerst aktuelle Fragen der Evolutionsbiologie wissenschaftlich bearbeitet. Der erste Teil zielt darauf, die Relevanz neutraler und selektiver Effekte für den evolutiven Wandel der Genexpression abzuschätzen. Im zweiten Teil wird die Frage nach Art und Weise der Entstehung neuer Gene an einem Beispiel erläutert. Eines der Gene aus dem ersten Teil wird aufgrund auffallender populationsgenetischer Daten näher untersucht. Es stellt sich heraus, dass dieses Gen in einer der untersuchten Subspezies in einem ausgedehnten Tal geringer genetischer Variabilität liegt, wie es der Signatur eines rezenten Selektionsereignisses (Selective Sweep) entspricht. Faszinierend ist die Tatsache, dass es sich als Gen ohne Homologe und Paraloge außerhalb des Genus *Mus* um ein echtes Orphan Gen handelt. In der vorliegenden Arbeit werden Einblicke in die *de novo* Genese dieses Gens aus nichtkodierender DNA gewährt.

1.1 Evolution der Genexpression

Bereits zur Mitte des vergangenen Jahrhunderts wurde dem Wandel der Genexpression eine Rolle für die Evolution der Organismen eingeräumt. Grula (Grula 2008) zufolge nahmen Waddington und Goldschmidt bereits in den 1940er Jahren Untersuchungen zur Mitwirkung der Genregulation an evolutiven Prozessen auf. Einflußreiche Veröffentlichungen von Britten und Davidson (1969; 1971), King und Wilson (1975) und Wilson, Maxson und Sarich (1974) nährten den Gedanken, dass Unterschiede in proteinkodierenden Bereichen nicht ausreichen, um die Unterschiede der untersuchten Spezies und damit die Evolution der Organismen umfassend zu erklären. In jüngster Vergangenheit verdichten sich die Beispiele regulatorischer Evolution. So spielt regulatorische Evolution eine Rolle in der Pigmentierung der Flügel (Gompel et al. 2005), des Abdomens (Prud'homme et al. 2006; Jeong et al. 2008; Williams et al. 2008), sowie dem Auftreten der Trichome (McGregor et al. 2007) in *Drosophila*. In Mais geht die mit der Domestikation verbundene Apikaldominanz auf Änderungen der Genexpression von *tb1* zurück (Doebley, Stec und Hubbard 1997; Hubbard et al. 2002; Clark et al. 2006). Ein anderes prominentes Beispiel ist die Beckenstruktur von Stichlingen. In marinen Formen sind die

Beckenknochen zu einem Körperschutz ausgebildet, während diese Struktur in Süßwasserformen im Vergleich stark reduziert (Shapiro et al. 2004) ist. Diese morphologischen Unterschiede sind wahrscheinlich auf unterschiedliche Regulation von *pitx* zurückzuführen. Auch in Primaten (Khaitovich, Paabo und Weiss 2005; Blekhman et al. 2008; Chaix et al. 2008), Hefe (Townsend, Cavalieri und Hartl 2003) und nicht zuletzt in der Maus (Harr et al. 2006; Ihle et al. 2006; Voolstra et al. 2007) gibt es Hinweise auf das Wirken regulatorischer Evolution. Diese Beispiele stimulieren die Debatte um die Frage, in welchem Ausmaß regulatorische Evolution maßgeblich für die Evolution der Organismen ist, in welchem Verhältnis sie zur Evolution kodierender Sequenzen steht und ob sie neutral oder unter positiver Selektion evolviert (Tautz 2000; Khaitovich et al. 2004; Khaitovich, Paabo und Weiss 2005; Lemos et al. 2005; Hoekstra und Coyne 2007; Wray 2007; Carroll 2008). Der erste Teil dieser Arbeit (Kapitel 2) widmet sich der Frage, ob die Evolution der Genexpression einem neutralen Evolutionsmodell entspricht und leistet einen Beitrag zur aktuellen Debatte.

1.2 Die Entstehung neuer Gene

Genduplikation wurde bereits in den 1930er Jahren als ein möglicher Mechanismus für die Entstehung neuer Gene postuliert (Haldane 1932; Muller 1935). Seit Ohnos „Evolution by Gene Duplication“ (Ohno 1970) gilt Genduplikation als wichtigster Mechanismus für die Entstehung neuer Gene. Entsteht eine Genkopie, führt dies häufig aufgrund funktionaler Redundanz zu einer Reduktion der selektiven Zwänge auf Kopie und Original. Von diesen Zwängen befreit, kann ein dupliziertes Gen frei evolvieren, was häufig zum Verlust der Funktion und damit zur Entstehung eines Pseudogenes führt, aber auch Subfunktionalisation (Force et al. 1999; Lynch et al. 2001) und Neofunktionalisation nach sich ziehen kann. Dass die Mitglieder von Genfamilien wie z.B. den Hox-Genen durch Genduplikation entstanden sind, ist offensichtlich. Aber auch abgewandelte Duplikationsmechanismen, wie Retrotransposition und die Verschmelzung von Genen oder deren Teile (Genfusion, Genfission, Exonshuffling) spielen eine Rolle in der Entstehung neuer Gene (Long et al. 2003). Häufig treten diese Mechanismen in Kombination auf. So geht die Entstehung des ersten jungen Gens, dessen Entstehungsgeschichte genauer nachvollzogen wurde (Long und Langley 1993) auf eine Kombination klassischer

Genduplikation und Retrotransposition zurück. Chimäre Genstrukturen sind häufig die Folge (Wang et al. 2002; Jones, Custer und Begun 2005).

In jüngster Vergangenheit verdichten sich jedoch die Hinweise, dass die *de novo* Genese von Genen, das heißt die Entstehung von Genen aus nichtkodierender Sequenz eine größere Rolle für die Entstehung neuer Gene spielt als bislang angenommen. David Begun und Koautoren (2007) konnten elf potentiell *de novo* entstandene Gene in Drosophila indentifizieren. In einer genomweiten Suche von Zhou und Koautoren (2008) sind sogar fast 12% aller neuen Gene Resultate einer *de novo* Genese. Damit rückt die *de novo* Genese als wichtiger Entstehungsmechanismus von Gründergenen funktional und strukturell verwandter Genfamilien im Sinne eines Phylostratums (Domazet-Loso, Brajkovic und Tautz 2007) in den Vordergrund.

Junge Gene stehen oft unter dem Einfluß positiver Selektion (Begun 1997; Nurminsky et al. 1998; Johnson et al. 2001; Enard et al. 2002; Maston und Ruvolo 2002; Wang et al. 2002). Sowohl das erste Auftreten, als auch die folgende Feinjustierung ziehen wahrscheinlich Selektionsereignisse nach sich, bevor negative Selektion zur Wahrung der neuen Funktion in den Vordergrund tritt (Domazet-Loso und Tautz 2003; Jones, Custer und Begun 2005). Anhand dieses Merkmals wurde im zweiten Teil dieser Arbeit Poldi identifiziert. Im ersten Teil dieser Arbeit fällt die Stromaufwärtsregion von Poldi durch eine deutliche Reduktion der genetischen Variabilität in einer der untersuchten Hausmaussubspezies auf. Das Auftreten einer adaptiven Variante eines Gens führt dazu, dass andere konkurrierende Varianten innerhalb kurzer Zeit getilgt werden und die genetische Variabilität ausgelöscht wird. Man spricht von einem Selective Sweep. Da angrenzende Bereiche auf dem Chromosom auf dem sich diese vorteilhafte Variante befindet physikalisch gekoppelt sind, wird auch deren Variabilität reduziert (Genetic Hitchhiking, (Maynard Smith und Haigh 1974; Fay und Wu 2000)). Dieser Effekt ist indikativ für positive Selektion und wird in verschiedenen Nachweisverfahren genutzt (Kim und Stephan 2002; Kauer, Dieringer und Schlotterer 2003; Wiehe et al. 2007; Teschke et al. 2008). Nach einem Selective Sweep steigt die genetische Variabilität durch Neumutation wieder an und strebt dem Mutations-Drift-Equilibrium entgegen. Diese Neumutationen führen zu einer Verschiebung des Allelfrequenzspektrums, die ebenfalls detektierbar ist (Tajima 1989).

Unter diesen Gesichtspunkten wird Poldi und die Poldi umgebende genomische Region ausführlich populationsgenetisch untersucht. Exemplarisch wird die *de novo*

Genese eines jungen Gens unter dem Regime positiver Selektion in Säugetieren herausgearbeitet.

1.3 Die Maus als Modell

Anhand zweier Populationen der Hausmaus wird die Evolution der Genexpression in Kapitel 2 untersucht. Proben einer der Populationen wurden in der Köln-Bonner Bucht gesammelt und gehören zur Subspezies *Mus musculus domestius*. Die Exemplare der anderen Population wurden in Studenec, Tresov, Rousek, Poszdatin, und Reijtar in Tschechien gesammelt und gehören zu *Mus musculus musculus*. Zwei weitere Population werden in die populationsgenetische Analyse der Poldi Region miteinbezogen. Die eine stammt aus dem französischen Zentralmassiv (*M. m. domesticus*), die andere aus Almaty und Umgebung in Kasachstan (*M. m. musculus*). Um die Entstehungsgeschichte des Poldi Gens nachvollziehen zu können werden *Mus castaneus*, *Mus spretus*, *Mus spicilegus*, *Mus macedonicus*, *Mus cypriacus*, *Mus famulus*, *Mus caroli*, *Apodemus flavicollis* und *Rattus norvegicus* analysiert (siehe auch 4.1.1).

Die Maus bietet als Modellsystem für evolutionsbiologische Studien eine Vielfalt verfügbarer molekularbiologischer Methodik und Daten. Sie ist das bestuntersuchte Säugetier und dient im Kontext der biomedizischen Forschung als Modell für den Menschen. Bereits 2002 wurde eine vollständige Genomsequenz veröffentlicht (Waterston et al. 2002). Über acht Millionen SNPs wurden in Labormäusen typsiert und stehen als genetische Marker zur Verfügung (Frazer et al. 2007). Ein Microarray zur Typisierung von mehr als 600000 SNPs und etwa 900000 CNVs wird gerade in einer Kooperation des Jaxlabs mit Affymetrix entwickelt. Mit der Möglichkeit des Gene Targetings, das heißt gezielt Gene auszuschalten und zu verändern (Thomas und Capecchi 1987), bietet sie die ideale Basis, um die genaue Wirkung evolutionsrelevanter Mutationen zu analysieren. Für populationsgenetische Studien bietet ihre ausführlich studierte Ökologie und Stammesgeschichte viele Vorteile.

Die Trennung der Linien von Ratte und Maus kann etwa zehn bis zwölf Millionen Jahre zurück datiert werden (Guenet und Bonhomme 2003; Chevret, Veyrunes und Britton-Davidian 2005). Die Abspaltung von *Apodemus* fand ca. eine Million Jahre später statt (siehe auch Abb. 24). *Mus caroli* ist in Südostasien beheimatet. Das Verbreitungsgebiet reicht vom Ryukyu Archipel (Japan), über Taiwan, Hainan und

Südchina bis nach Malaysia, Sumatra, Java und Flores (Corbet und Hill 1992). Die Hausmaus und *M. caroli* divergieren seit etwa drei Millionen Jahren. Durch künstliche Besamung können noch gemeinsame Nachkommen erzeugt werden. Diese sind aber unfruchtbar (West, Frels und Chapman 1978). *Mus famulus* tritt endemisch in Sympatrie mit *Mus musculus* in den Nilgiri Bergen Südindiens auf (Chevret, Jenkins und Catzeflis 2003). Die Divergenz zur *Mus musculus* Gruppe beträgt etwa zweieinhalb Millionen Jahre. *Mus spretus*, *Mus macedonicus*, und *Mus cypriacus* sind mediterrane Spezies. Ihre Stammesgeschichte hat sich vor etwa eineinhalb Millionen Jahren von der Stammesgeschichte der Hausmaus getrennt. Sie treten sympatrisch mit *M. m. domesticus* auf, sind aber nicht kommensal (Boursot et al. 1993). *M. spretus* ist die westliche mediterrane Kurzschwanzmaus. Sie tritt in Nordafrika bis nach Südfrankreich auf, mit der höchsten genetischen Diversität in Nordafrika. Die Diversität nimmt nach Norden ab, was auf den Ursprung von *M. spretus* in Nordafrika hinweist. Hybride mit *M. m. domesticus* treten gelegentlich auf und können trotz Sterilität der F1 Männchen zu einem limitierten Genfluss führen (Orth et al. 2002). *M. macedonicus* ist die östliche Hausmaus. Ihr Verbreitungsgebiet reicht vom Kaukasus bis zum Balkan. Sie tritt außerdem im Nahen Osten, Klein Asien und dem südlichen Balkan auf (Macholan et al. 2007). *M. cypriacus* ist endemisch auf Zypern und nah mit *M. macedonicus* verwandt (Cucchi et al. 2006). *Mus spicilegus* ist nordöstlich von *M. macedonicus* in Bulgarien, Moldawien, Ungarn, Österreich und der Ukraine verbreitet, legt ein ausgeprägtes Sammelverhalten an den Tag und baut Erdhügel. Sie interagiert sympatrisch hauptsächlich mit *M. m. musculus* und kann mit dieser, wie auch *M. spretus* und *M. macedonicus*, zumindest unter Laborbedingungen Hybride hervorbringen.

Die Hausmaus *M. musculus* bildet einen Subspezieskomplex bestehend aus *M. m. musculus*, *M. m. domesticus* und *M. m. castaneus*. *M. m. musculus* und *M. m. castaneus* hybridisieren in Japan zu *M. m. molossinus*. Diese Subspezies sind die Quelle der Labormausstämme (Frazer et al. 2007; Yang et al. 2007), wobei *M. m. domesticus* mit mehr als zwei Dritteln den größten Anteil an Haplotypen zu den klassischen Laborstämmen beigetragen hat. Die drei Unterarten sind wahrscheinlich vor weniger als einer Million Jahre auf dem indischen Subkontinent entstanden, wo die größte genetische Variabilität herrscht (Boursot et al. 1996). Aller Wahrscheinlichkeit nach hat die Hausmaus von Indien ausgehend ihr heutiges Verbreitungsgebiet besiedelt. Während *M. m. musculus* über eine nördliche Rute nach

Europa vorgedrungen ist, hat *M. m. domesticus* eine südliche Route eingeschlagen. Dort wo keine geographische Trennung durch das Kaspische- oder das Schwarze Meer vorliegt, formen die beiden Subspezies eine enge Hybridzone (Abb. 1). Das erste Auftreten der Hausmaus in Westeuropa kann bei kritischer Betrachtung der fossilen Funde (Cucchi, Vigne und Auffray 2005), erst vor etwa 2.000 - 3.000 Jahren mit ausreichender Sicherheit belegt werden, während im Nahen Osten schon vor etwa 10.500 - 14.000 Jahren Hausmäuse existiert haben. Als sicher gilt, dass Europa ausgehend vom Fruchtbaren Halbmond erst nach der letzten Eiszeit, der Würm, vor etwa 10.000 Jahren von *M. m. domesticus* kolonisiert wurde. Vermutlich entwickelte sich der Kommensalismus der Hausmaus im Fruchtbaren Halbmond, in dem schon sehr früh Ackerbau und Viehzucht betrieben wurden. Von hier aus folgte die kommesale *M. m. domesticus* dem Menschen nach Europa, konnte aber die nördlichen Bereiche erst besiedeln, als auch hier Ackerbau und Vorratshaltung Einzug hielten (Rajabi-Maham, Orth und Bonhomme 2008). Hausmäuse leben hauptsächlich kommensal, auch wenn sekundär wilde Populationen von *M. m. musculus* und *M. m. domesticus* in den milden mediterranen Klimaten bekannt sind (Sage 1981).

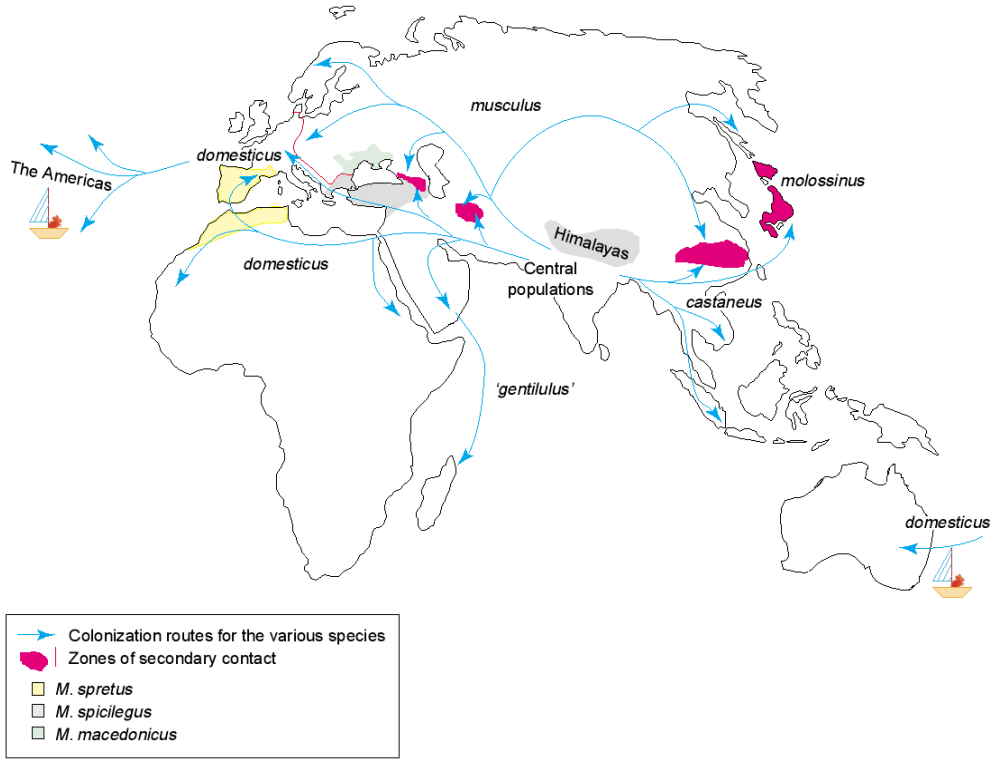


Abb. 1 Kolonisationswege und Verbreitung der Hausmaus und nahverwandter Spezies (Guenet und Bonhomme 2003)

In sehr dichten kommensalen Populationen können bis zu 70 Mäuse pro Quadratmeter leben. Werden hohe Populationsdichten erreicht, bilden die Mäuse Sippen, die aus einem dominanten Männchen mit meist mehreren Weibchen und Nachkommen bestehen. Die Größe der Territorien beträgt dann nicht mehr als 2 m² (Gray, Jensen und Hurst 2000).

Nur ein sehr kleiner Teil kommensal lebender Mäuse entfernt sich weiter als 25 m von seinem Geburtsort. Diese Tiere sind hauptsächlich junge Männchen (Singleton 1983; Pocock, Hauffe und Searle 2005). In wilden Populationen mit Populationsdichten von 150 Tieren pro Hektar und darunter sind Mäuse deutlich weniger territorial, auch wenn eine ähnliche Familienstruktur vermutet wird (Fitzgerald, Karl und Moller 1981).

Die kommensale Lebensweise führt zu einem intensiven, passiven Transport der Mäuse, welcher insbesondere mit dem Beginn der interkontinentalen Seefahrt im 15. und 16. Jahrhundert eine starke Ausbreitung der westeuropäischen Hausmäuse über den gesamten Globus zur Folge hat. Die Kolonisation Amerikas, des tropischen Afrikas und Australiens mit Hausmäusen ist vornehmlich durch die Verschiffung der westeuropäischen Hausmaus *M. m. domesticus* (Berry 1991) erfolgt.

Die in dieser Studie bearbeitete kasachische Population der Unterart *M. m. musculus* liegt zumindest geographisch nahe am Ursprung der Hausmaus und könnte eher dem anzestralen Status der Subspezies entsprechen. Die tschechische Population ist eher als abgeleitet zu bewerten. Die Populationen dieser Studie, die *M. m. domesticus* repräsentieren bieten ein ähnliches Bild: Während die französiche Hausmaus mitochondrialen Daten zufolge stark abgeleitet ist, bestätigen Untersuchungen mitochondrialer DNA in der deutschen Population anzestrale Allele und eine hohe Diversität (Rajabi-Maham, Orth und Bonhomme 2008). Die untersuchten Populationen gelten als distinkt (Ihle et al. 2006; Rajabi-Maham, Orth und Bonhomme 2008).

Ökologie, Stammesgeschichte und Verhalten der Maus werden nun seit über 30 Jahren intensiv erforscht. Dieses Hintergrundwissen macht sie zusammen mit den verfügbaren Daten und Methoden zum idealen Modellorganismus für evolutionsbiologische Studien.

1.4 Ziele

Im ersten Teil dieser Dissertation soll geprüft werden, ob die Evolution der Genexpression dem neutralen Modell entsprechend einer molekularen Uhr folgt. Hieraus ergibt sich die zu prüfende Hypothese: Der Polymorphismus ist ein maßgeblicher Faktor für die Divergenz der Genexpression. Neue Methoden sind nötig, um ein Maß für den Genexpressionspolymorphismus zu finden und die biologische Variabilität von technischen Effekten zu trennen. Diese Methodik gilt es zu entwickeln. Zusätzlich sollen Sequenzdaten der untersuchten Gene auf ihren evolutiven Modus geprüft werden (neutral oder selektiert).

Im zweiten Teil soll die Entstehungsgeschichte des Poldi Transkripts aus zuvor untranskribierter DNA anhand von Sequenzdaten mehrerer Spezies des Genus *Mus* untersucht werden. Die genomische Region in der sich Poldi befindet, soll weiterhin umfassend populationsgenetisch analysiert werden, um festzustellen, ob es Hinweise auf ein Selektionsereignis gibt, das kennzeichnend für ein junges Gen wäre. Mögliche Ursachen für ein Selektionsereignis sollen identifiziert werden.

2 Unterschiede der Genexpression zwischen natürlichen Populationen der Hausmaus unterstützen ein vorrangig neutrales Modell des evolutiven Wandels

2.1 Einleitung

Die relative Rolle von Veränderungen der Genexpression im Verhältnis zur Änderung von Proteinen ist Gegenstand lebhafter Debatten (Tautz 2000; Lemos et al. 2005; Hoekstra und Coyne 2007; Wray 2007; Carroll 2008). Obwohl unser Wissen über intra- und interspezifische Variation und Veränderung der Genregulation mit zunehmender Geschwindigkeit wächst, ist noch immer unklar, welche Rolle dies für Adaptationsprozesse spielt. Schätzungen des Anteils der Gene, die unter positiver oder negativer Selektion stehen, schwanken stark (Rifkin, Kim und White 2003; Yanai, Graur und Ophir 2004; Lemos et al. 2005). An diesen Schwankungen ist erkennbar, dass ein umfassendes Modell der Evolution der Genexpression erst noch gefunden werden muss. Khaitovich et al. (2004; 2005) schlagen basierend auf Microarraydaten in Primaten ein vorrangig neutrales Modell der Transkriptomevolution vor. Auf Basis ähnlicher Experimente glauben Blekhman et al. (2008) Evidenz für positive Selektion in der menschlichen Linie gefunden zu haben. In einer Studie über Expressionsunterschiede innerhalb und zwischen *Drosophila* Spezies finden Wittkopp et al. (2008) eine Häufung cisregulatorischer Veränderungen zwischen den Spezies, was gegen ein strikt neutrales Modell regulatorischer Evolution spräche.

Um die populationsgenetischen Mechanismen, die der Evolution der Genexpression zugrunde liegen nachvollziehen zu können, ist ein integrativer Ansatz nötig, in welchem Expressionsniveaus und deren Divergenz mit Sequenzpolymorphismusdaten abgeglichen werden. Brown und Feder (2005) fanden für eine Reihe von Genen, die sie zwischen verschiedenen *Drosophila melanogaster* Stämmen verglichen haben, keine Korrelation zwischen diesen Parametern. Holloway et al. (2007) und Lawniczak et al. (2008) hingegen analysierten vorhandene genomweite Daten von *Drosophila simulans* und fanden, dass Sequenzpolymorphismus in *cis* tatsächlich eine wichtige Determinante der Expressionsvariation sein könnte und damit Objekt adaptiver Veränderung.

In diesem Teil der vorliegenden Arbeit befinden sich zwei Populationen der Hausmaus im Fokus. Die deutsche Population repräsentiert die westliche Subspezies *M. m. domesticus* und die tschechische Population repräsentiert die östliche Subspezies *M. m. musculus*. Mäuse beider Populationen wurden wild gefangen und in dieser Studie untersucht. Eine Gruppe von Genen, die zuvor in einem Microarrayexperiment als zwischen diesen Populationen differentiell exprimiert identifiziert wurde (Voolstra et al. 2007), wurde mittels quantitative Real Time PCR (qRT-PCR) genauer untersucht. Drei Gewebe wurden in die Studie eingeschlossen (Gehirn, Testis und Leber/Niere) und Expressionspolymorphismus mit Expressionsdivergenz verglichen. Zusätzlich wurden Sequenzpolymorphismusdaten einer Vielzahl von Individuen jeder der Populationen für die Stromaufwärtsregionen aller untersuchten Gene erhoben. Diese Daten ermöglichen es, umfassende Fragen nach der Neutralität oder Adaptivität genregulatorischer Veränderungen, einer möglichen Korrelation mit Sequenzpolymorphismus und dem Effekt gewebespezifischer Veränderungen zu beantworten.

2.2 Methoden zur Ermittlung der Expressionsniveaus

Die Quantifizierung von Expressionsniveaus mittels qRT-PCR kann von zwei Hauptfaktoren beeinträchtigt werden. Der erste betrifft Polymorphismen in den Primer- und Sondenbindestellen der Expressionsassays. Der zweite betrifft die polymorphe Expression des Referenz Gens, der so genannten endogenen Kontrolle, die benutzt wird, um das Assay zu kalibrieren. Da eine Nichtbeachtung dieser Schwierigkeiten die Ergebnisse signifikant verändert hätte, wurden im Folgenden Strategien zur Vermeidung und Korrektur entwickelt.

2.2.1 Kontrolle der Assaybinderegionen

Um mögliche Polymorphismen in den für die Assaybindung relevanten Bereichen detektieren zu können, wurden diese sequenziert. Es wurden Polymorphismen und Spleißvarianten in 15 von 39 getesteten Assays entdeckt (Tabelle 1).

Taqman assays	assay ID	MGI gene symbol	status
Included	Mm01217369_m1	1110017D15Rik	no polymorphism detected
	Mm01172741_g1	1700125F08Rik	no polymorphism detected
	Mm00432248_m1	Cacng2	no polymorphism detected
	Mm00436443_m1	Ccl25	no polymorphism detected
	Mm00432437_m1	Cdk5	no polymorphism detected
	Mm00558327_s1	Etd	no polymorphism detected
	Mm00468389_m1	Etv2	no polymorphism detected
	Mm00514956_m1	Flot2	no polymorphism detected
	Mm00516235_m1	Gpc6	no polymorphism detected
	Mm00468869_m1	Hif1a	no polymorphism detected
	Mm00498065_m1	Kcnd2	no polymorphism detected
	Mm01300291_m1	Krt2-17	no polymorphism detected
	Mm00450997_m1	Mir16	no polymorphism detected
	Mm01298523_m1	Nf1	no polymorphism detected
	Mm01290707_g1	Tomm40l	no polymorphism detected
	Mm00450900_m1	PanX1	no polymorphism detected
	Mm01192227_m1	Ppt1	no polymorphism detected
	Mm00453021_m1	Rab4b	no polymorphism detected
	Mm00503581_gH	Rarres2	no polymorphism detected
	Mm00803317_m1	Rgs16	no polymorphism detected
	Mm00491014_m1	Scamp5	no polymorphism detected
	Mm01282622_m1	Sv2c	no polymorphism detected
	Mm00843984_s1	Tcte3	no polymorphism detected
	Mm01168596_m1	Tmem24	no polymorphism detected
Excluded	Mm01217598_g1	4833411C07Rik	fixed difference probe
	Mm00661819_m1	AI604832	polymorphic probe domesticus
	Mm00731639_m1	Crisp1	shared polymorphism F primer
	Mm01174266_m1	Dscaml1	polymorphic F primer musculus
	Mm00834825_g1	Edf1	fixed difference R primer
	Mm00784689_s1	MGC118210;Xmr;Xmr	multicopy gene
	Mm00435145_m1	Nkx2-9	polymorphic probe domesticus
	Mm00439358_m1	Nr4a1	fixed difference probe
	Mm00510343_m1	Ppil3	fixed difference F primer
	Mm00499682_m1	Ppp1r11	duplicated F primer binding region domesticus
	Mm00839568_m1	Spt1	polymorphic probe domesticus
	Mm01352176_m1	Tmem16k	polymorphic R primer musculus
	Mm02017439_g1	Tmsb10	polymorphic probe domesticus
	Mm00840578_g1	Tnfrsf13c	different splice variants
	Mm00441325_m1	Sema3f	shared polymorphism R primer

Tabelle 1 Zusammenfassung der Ergebnisse der cDNA Sequenzierung von Primer- und Sondenbindestelle der in der qRT-PCR verwendeten Assays. Assays, die an polymorphe Sequenzen binden, wurden von der weiteren Analyse ausgeschlossen. Die Statusspalte enthält einen kurzen Hinweis, aus welchem Grund die Assays von der weiteren Analyse ausgeschlossen wurden.

Manche dieser Polymorphismen korrelieren nachweislich mit dem gemessenen Expressionsniveau (Abb. 2) und würden daher zu einem ungültigen Ergebnis führen. Deshalb wurde die Analyse auf 24 Gene, in denen keine Bindungsrelevanten Polymorphismen gefunden wurden, beschränkt.

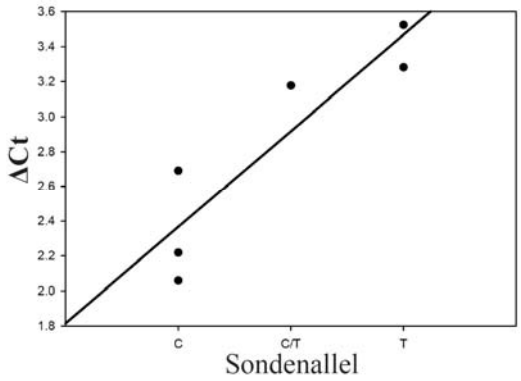


Abb. 2 Beispiel eines SNPs in der Sondenbinderegion des Genexpressionsassays Mm00661819_m
Die Sonde bindet an eine Region, die einen SNP in *M. m. domesticus* enthält. Die gemessenen Expressionsniveaus der sechs Individuen korrelieren stark mit dem allelischen Status dieses SNPs ($r^2 > 0,9$; $p < 0,05$; Pearson's Product Moment Correlation).

2.2.2 Kalibrierung des endogenen Kontrollgens *hprt*

Als interner Standard wurden drei technische Replikate jeder Maus und jedes Gewebes auf jeder qRT-PCR Platte mit einem Assay für *hprt* (hypoxanthin phosphoribosyltransferase) unter gleichen Bedingungen wie die übrigen Proben prozessiert. Im Gegensatz zu den Erwartungen für ein Housekeeping Gen, wies hprt neben erheblichen Unterschieden der Expression zwischen Individuen auch systematische Unterschiede des Expressionsniveaus, sowie der Varianz zwischen den Populationen auf und konnte deshalb nicht direkt für die Normalisierung verwendet werden (Tabelle 2, Abb. 3). Zwei sorgfältig auszuführende Schritte waren notwendig, um diese Unterschiede zu detektieren: Zuerst wurde die in der cDNA-Synthese eingesetzte RNA-Menge mittels Messungen auf einem Agilent 2100 BioAnalyzer eingestellt. Daraufhin wurde die in der qRT-PCR verwendete cDNA mit einer Fluoreszenzmethode nach Libus und Storchova (Libus und Storchova 2006) auf einem Nanodrop 3300 Fluorometer gemessen und auf gleiche Mengen eingestellt.

Tissue	Subspecies	Mean Ct	Std. Deviation	Wilcoxon W Test p-value	Levene's Test p-value
Brain	domesticus	27.12	0.20	0.94	0.34
	musculus	27.16	0.26		
Liver/kidney	domesticus	28.75	0.19	0.004	0.04
	musculus	29.62	0.44		
Testis	domesticus	29.36	0.51	0.002	0.77
	musculus	30.48	0.50		

Tabelle 2 Die hprt Expression unterscheidet sich in Varianz und Höhe zwischen den Subspezies

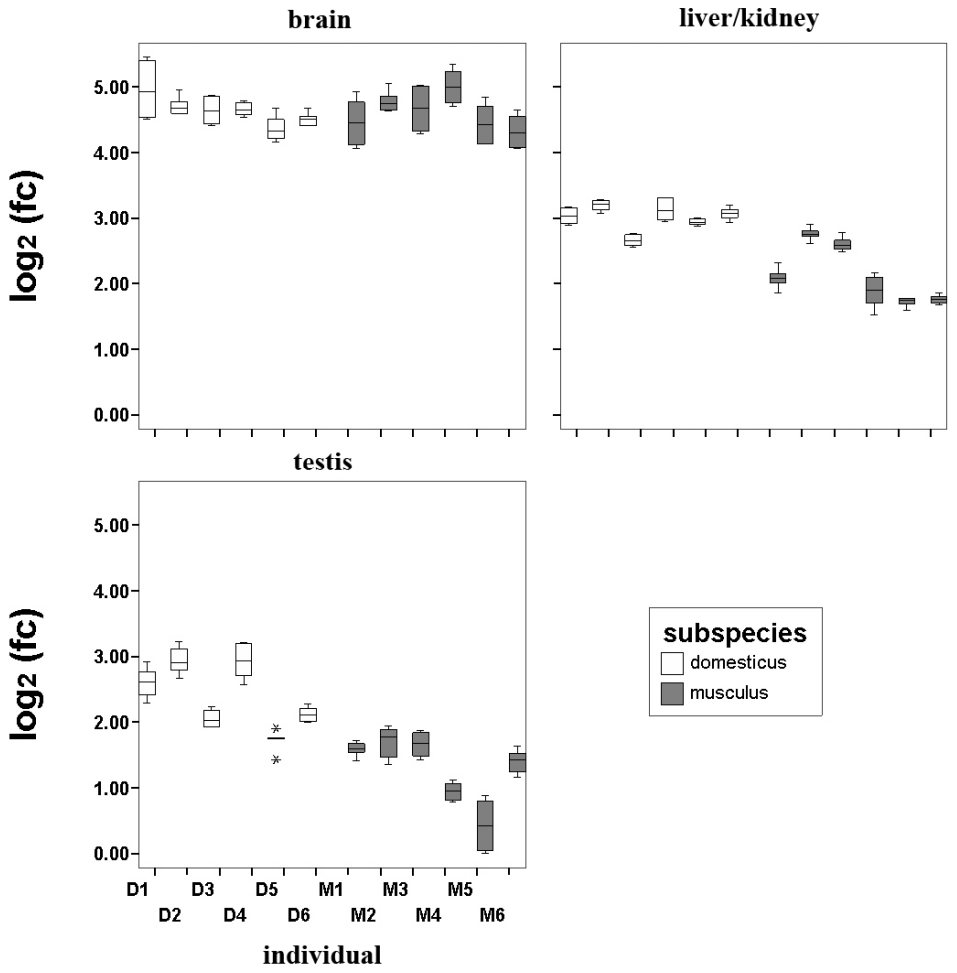


Abb. 3 Hprt Expression in den einzelnen Mäusen, basierend auf absoluter Quantifizierung von RNA und cDNA ,sortiert nach Gewebe
Log_2 Foldchanges relativ zum kleinsten gemessenen Wert sind auf der y-Achse. Die Fehlerbalken entsprechen den 95% CI sechs technischer Replikate.

Die Auswahl der endogenen Kontrolle ist ein generelles Problem in qRT-PCR-Experimenten (Bustin und Nolan 2004; Dheda et al. 2005). Daher wurde ein Verfahren zur Berechnung eines Korrekturfaktors (Cor) entwickelt. Aus den Expressionsdaten, die für ein bestimmtes Individuum für alle Gene in einem bestimmten Gewebe relativ zu *hprt* vorliegen, lässt sich ein individueller Korrekturfaktor ermitteln. Dieser korrigiert dann für die spezielle *hprt* Expression dieser Maus im entsprechenden Gewebe. Es handelt sich also um eine Normalisierung über alle Gene im Datensatz. *Hprt* dient dann nur noch als Vermittler zwischen den Platten. Da diese Gene zufällig gewählt sind und demnach voraussichtlich nicht direkt interagieren, ist keine systematische Hoch- oder Herunterregulierung zu erwarten.

Auch wenn einzelne dieser Gene spezifische Änderungen zeigen, sollte dies nur einen kleinen Einfluss auf den Median haben.

Um diese Korrekturfaktoren zu berechnen, wurden zuerst die Konzentrationen der Zielgene ([G]) relativ zu hprt berechnet: $[G] = 2^{-\Delta Ct}$. Um jedem Gen denselben Einfluss auf den Korrekturfaktor einzuräumen, muss [G] des Individuums durch den Mittelwert von [G] dieses Gens in diesem Gewebe geteilt werden. Dies führt zu einem Expressionsniveau des Zielgens in jeder Maus relativ zu einem Mittelwert von 1. Nun wurde der Median des relativ zu *hprt* gemessenen Expressionsniveaus für jedes Individuum über alle Gene in jedem Gewebe separat berechnet. Die resultierenden 36 Korrekturfaktoren (einer pro Individuum pro Gewebe) wurden nun zur Korrektur der ΔCts verwendet. Alle weiteren Analysen wurden mit diesen korrigierten ΔCts durchgeführt. Die Varianz der Korrekturfaktoren ist im Bootstrapping klein und rechtfertigt die Anwendung der Methode (S 6). Auch aus den in Abb. 3 dargestellten durch cDNA Normalisierung gewonnenen *hprt* Konzentrationen lässt sich ein Korrekturfaktor berechnen. Dieser korreliert stark mit dem Korrekturfaktor (Cor) über alle Gene ($r^2 = 0{,}71$; Abb. 4), was die Ermittlung von Cor unabhängig validiert. Der Messung des Korrekturfaktors über alle Gene wurde hier der Vorzug vor der Bestimmung mittels normalisierter cDNA gegeben, da sie auf einer deutlich größeren Zahl unabhängiger Messungen beruht.

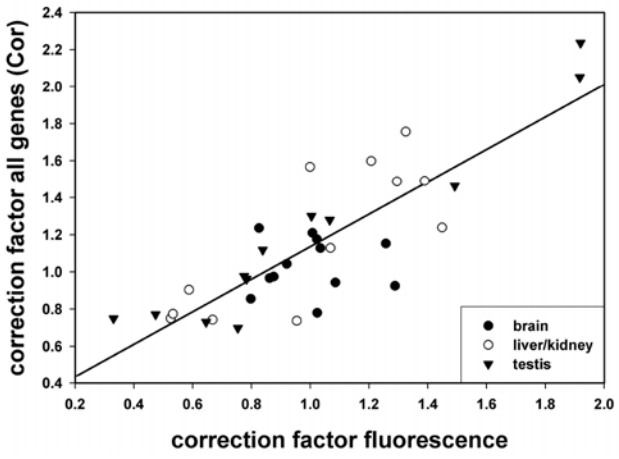


Abb. 4 Der Korrekturfaktor für unterschiedliche *hprt* Expression aus der Fluoreszenznomalisierung nach Libus (2006, x-Achse) korreliert stark ($r^2 = 0{,}71$, $p = 1{,}53 \times 10^{-10}$, Pearson's Correlation) mit dem Korrekturfaktor über alle Gene (Cor, y-Achse). Jeder Datenpunkt repräsentiert die in den beiden Verfahren unabhängig voneinander gewonnenen Korrekturfaktoren für je eine Maus in einem Gewebe.

2.3 Methode zur Ermittlung der biologischen Expressionsvariabilität innerhalb einer Population (Expressionspolymorphismus)

Das Bestimmen der biologischen Expressionsvarianz in natürlichen Populationen ist schwierig, da der Messung eine technische Varianz innewohnt, die nicht konstant ist (Heteroskedastizität), sondern von der Höhe des gemessenen Expressionsniveaus abhängt. Es handelt sich dabei um ein Problem, das häufig dann eine Rolle spielt, wenn Messmethoden über mehrere Größenordnungen hinweg angewendet werden, wie z.B. bei Microarraydaten (Tusher, Tibshirani und Chu 2001; Manda, Walls und Gilthorpe 2007) oder in diesem Fall in der qRT-PCR. Daher ist es notwendig, die in den Daten enthaltene biologische Variabilität zwischen den Individuen einer Population (Expressionspolymorphismus) von der technisch bedingten Varianz zu reinigen, um diese realistisch abschätzen zu können.

Um die technisch bedingte Varianz ermitteln zu können, wurde jede Messung des Zielgens wie auch der endogenen Kontrolle in Triplikaten durchgeführt. Die gemessene Standardabweichung der drei technischen Replikate wächst mit dem gemessenen Ct (Abb. 5A), also der kleiner werdenden Konzentration des Zielgens. Da der qRT-PCR eine Verdopplung der in der Reaktion enthaltenen DNA mit jedem Zyklus also ein exponentielles Wachstum zugrunde liegt, wachsen auch die Fehler (z.B. Pipettierfehler) exponentiell. Für die qRT-PCR-Messungen in Abb. 5B wurden mittels *in vitro* Transkription künstlich erzeugte, definierte Mengen RNA eingesetzt (4.1.8). Die Standardabweichungen der technischen Replikate verhalten sich hier offenbar wie in Abb. 5A. Dies zeigt, dass das exponentielle Wachstum der Fehler tatsächlich in der Methode begründet liegt und nicht etwa einen Effekt der natürlichen Transkriptionsmaschinerie darstellt.

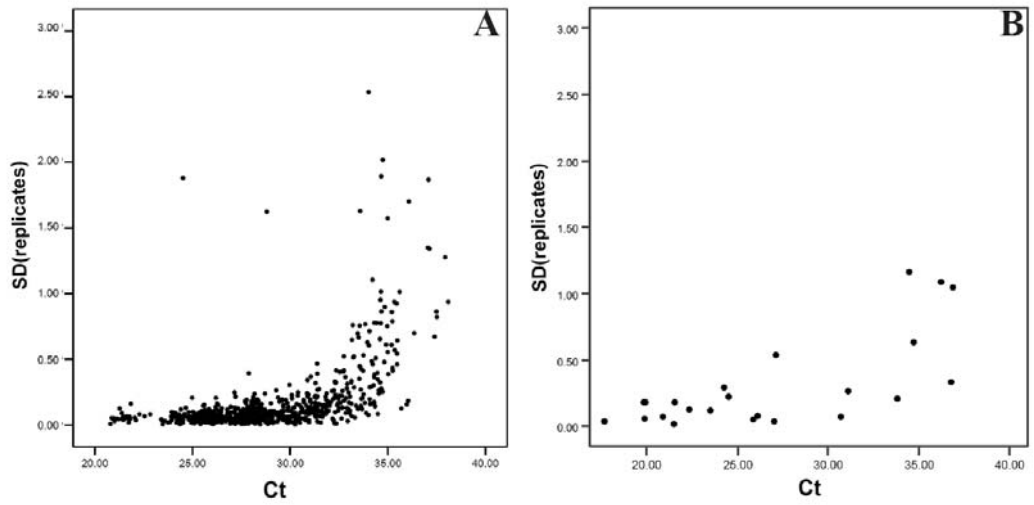


Abb. 5 Die Standardabweichung der technischen Replikate wächst mit dem gemessenen Ct
A: Standardabweichung der technischen Replikate in Abhängigkeit des Cts aller Gene, Gewebe und Individuen der vorliegenden Studie B: Messungen basierend auf definierten Verdünnungen einer *in vitro* transkribierten RNA.

Die Abhängigkeit der Varianz der technischen Replikate vom gemessenen Expressionsniveau führt dazu, dass auch die Standardabweichung der Expressionsniveaus innerhalb der Population (SD_{pop}) von ΔCt abhängen (Abb. 7A). Die technische Varianz der Messungen für die einzelnen Mäuse ist demnach maßgeblich für die Standardabweichung der ΔCts in der Population (SD_{pop}).

Die Varianz der ΔCt innerhalb einer Population enthält demnach neben der biologischen Varianz (unterschiedliche Expression des Zielgens zwischen Individuen einer Population) noch die technische Varianz der Expressionsmessungen der Individuen (Fehler der technischen Replikate). Schätzt man nun die technische Varianz der Individualmessungen ab, kann man auch den Einfluss dieser technischen Varianz auf SD_{pop} abschätzen.

Um den Gesamtfehler der Individualmessung zu ermitteln wurde die Gaußsche Fehlerfortpflanzung angewendet. Da ΔCt die Differenz der Cts von Zielgen und endogener Kontrolle ist, pflanzt sich der Fehler der Messung des Zielgens (SE(Ct(G))), der endogenen Kontrolle (SE(Ct(EC))) und des Korrekturfaktors (SE(Cor)) der endogenen Kontrolle (siehe 2.2.2) im technischen Fehler von ΔCt (SE(ΔCt)) fort:

$$SE(\Delta Ct) = \sqrt{SE(Ct(G))^2 + (SE(Ct(EC))^2 + (\frac{SE(Cor)}{Cor \cdot \ln(2)})^2}$$

SE(Ct(G)) und SE(Ct(EC)) wurden als Standardfehler der technischen Replikate des Zielgens und der endogenen Kontrolle ermittelt. Der Standardfehler des Korrekturfaktors (SE(Cor)) wurde durch 100000faches Bootstrapping über alle Gene ermittelt (siehe Anhang S 6), da SE(Cor) ein Verhältnis und daher nicht normalverteilt ist.

Um den Einfluss der technischen Fehler der Einzelmessungen auf die Standardabweichung in einer Population abzuschätzen (SD_{pop}), wurde der Fehler der Einzelmessungen (SE(ΔCt)) für jedes Gen und Gewebe über die Population gemittelt ($\overline{SE(\Delta Ct)}$). Beide Größen wurden dann logarithmiert, um sie einer Normalverteilung anzunähern (siehe Anhang S 7) und für die Regressionsanalyse vorzubereiten (Abb. 6). Der mittlere Standardfehler der einzelnen Expressionsmessungen in einer Population ist ein signifikanter Einflussfaktor auf die Standardabweichung in der Population ($r^2 = 0{,}387$; $p = 3{,}3 \times 10^{-14}$), und erklärt nahezu 40% der Varianz.

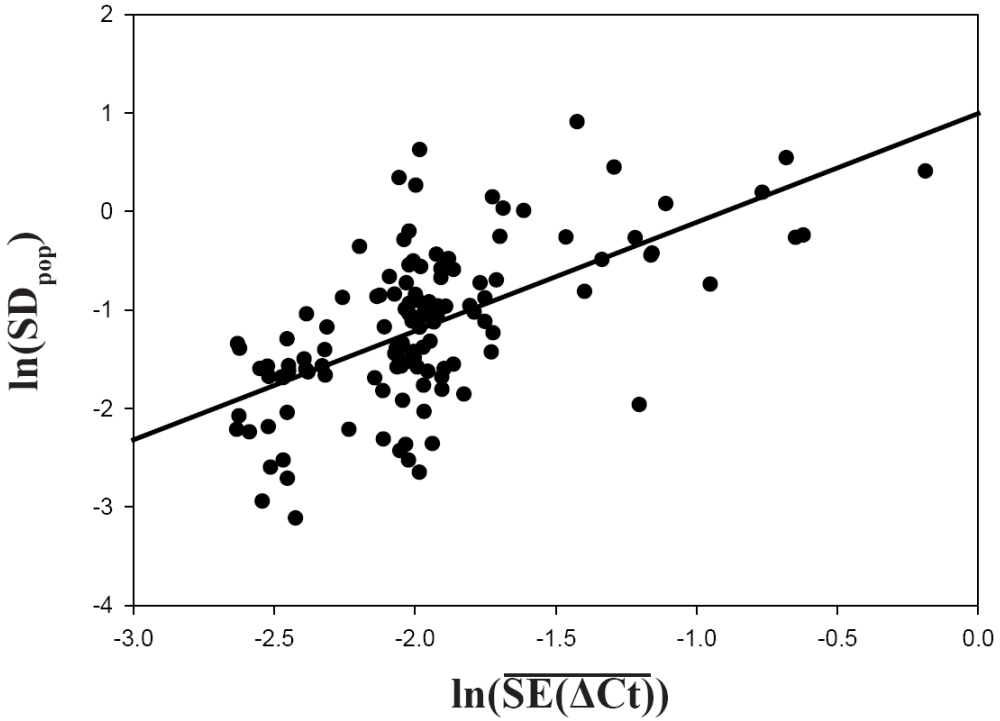


Abb. 6 Der mittlere Messfehler der Einzelmessungen $\overline{SE(\Delta Ct)}$ beeinflusst die Standardabweichung der ΔCt in der Population (SD_{pop}). Beide Größen wurden vor der Regression logarithmiert (r^2=0.387,p=3.3x10-14; Pearson's Product Moment Correlation).

Der Teil der Varianz, der nicht durch die technischen Messfehler vorhergesagt werden kann, ist wahrscheinlich biologisch, d.h. er entstammt der natürlichenVarianz zwischen den Individuen einer Population. Demnach sind die Residuen der Regression in Abb. 6 ein Maß für den Expressionspolymorphismus, der von einem technischem Anteil auf diese Weise befreit wurde.

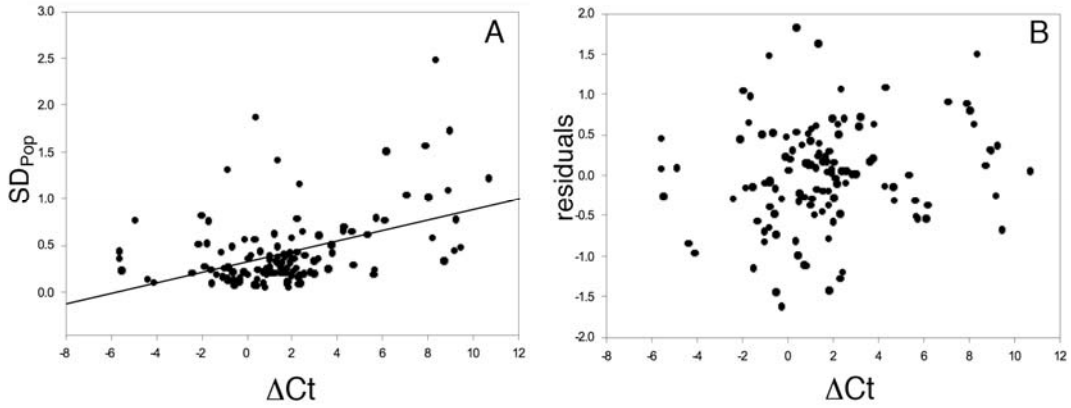


Abb. 7 A: Abhängigkeit der Maße für Expressionspolymorphismus vom Expressionsniveau vor und nach Berücksichtigung des Effekts technischer Messfehler.
A: Die Standardabweichung der ΔCt in der Population (SD_{pop}) enthält technische und biologische Varianz und hängt vom Expressionsniveau (ΔCt) in der Population ab ($p < 3x10^{-7}$ und $r^2 = 0.2$, Pearson's Product Moment Correlation).
B: Die Residuen der Regression $\ln(SD_{pop})$ gegen $\ln(\overline{SE(\Delta Ct)})$ sind unabhängig vom Expressionsniveau ($p > 0.1$ und $r^2 < 0.04$, Pearson's Product Moment Correlation) und werden in der vorliegenden Studie als Maß für den Expressionspolymorphismus verwendet.

Die Methode wird von der Tatsache validiert, dass diese Residuen nun nicht mehr vom gemessenen Expressionsniveau abhängen. Die aus dem technischen Fehler stammende Heteroskedastizität (Abb. 5, Abb. 7A) ist nicht mehr vorhanden (Abb. 7B), die Residuen sind normalverteilt (p=0.44, Shapiro-Wilks-Test for Normality). Daher werden diese Residuen im weiteren Verlauf als Expressionspolymorphismus betrachtet. Die Residuen und damit der Expressionspolymorphismus können auch negative Werte annehmen, was bedeutet, dass ein Gen in einer bestimmten Population eine niedrigere Varianz als der Durchschnitt der Gene mit gleich hohem technischem Fehler aufweist.

2.4 Ergebnisse

2.4.1 Signifikante Expressionsunterschiede

Anhand der für unterschiedliche Expression der endogenen Kontrolle korrigierten Expressionsniveaus ist es nun möglich festzustellen, welche Gene zwischen den Subspezies *M. musculus* und *M. domesticus* differentiell exprimiert sind. Zwölf der getesteten Gene zeigten einen signifikanten Unterschied in mindestens einem Gewebe (Abb. 8). Da die Unterschiede nicht anhand eines einfachen Foldchange-Schwellenwertes ermittelt wurden, können auch kleine Unterschiede signifikant sein, wenn die Varianz innerhalb der Populationen klein ist. Da ΔCt ein logarithmisches Maß und nicht normalverteilt ist, wurde der Wilcoxon-W-Test verwendet, um Expressionsunterschiede zu detektieren. Dem multiplen Testen wurde durch Verwendung der FDR (False Discovery Rate), (Storey und Tibshirani 2003) Rechnung getragen. Hierzu wurde die R-Bibliothek Q-VALUE mit Bootstrapping und „robust method“ Option für limitierte Stichprobengrößen gewählt (S 3).

Die hier untersuchten Gene stammen aus einer Vorauswahl von Kandidatengenen, die eine signifikant unterschiedliche Expression in Microarrayexperimenten aufwiesen (4.1.4). Die Tatsache, dass nur die Hälfte der Gene mittels qRT-PCR bestätigt werden konnte, kann generellen Problemen der Microarrayhybridisierung zugeordnet werden (Pozhitkov, Tautz und Noble 2007). Hinzu kommt die durchgeführte Kontrolle auf Sequenzpolymorphismus in den für die qRT-PCR relevanten Bereichen, die für Microarrays so nicht durchführbar ist (2.2.1). Es ist nicht ungewöhnlich, dass nur ein Teil der in Microarrayexperimenten als differentiell exprimiert identifizierten Gene durch andere Methoden bestätigt werden kann.

Obwohl die Gene aus einer Vorauswahl stammen, repräsentieren sie eine Zufallsauswahl hinsichtlich der Genfunktion. In diesem Zusammenhang ist es überraschend, dass neun der zwölf unterschiedlich exprimierten Gene den Expressionsunterschied in nur einem der drei untersuchten Gewebe aufwiesen. Zwei Gene wurden nur in einem Gewebe exprimiert und zeigen eine Veränderung in die gleiche Richtung. Ein Gen war nur im Hoden exprimiert (Tabelle 3). Die meisten analysierten Gene sind aber in allen drei Geweben exprimiert, wenn auch auf unterschiedlichen Niveaus, was auf die Wirkung gewebespezifischer Enhancer hindeutet.

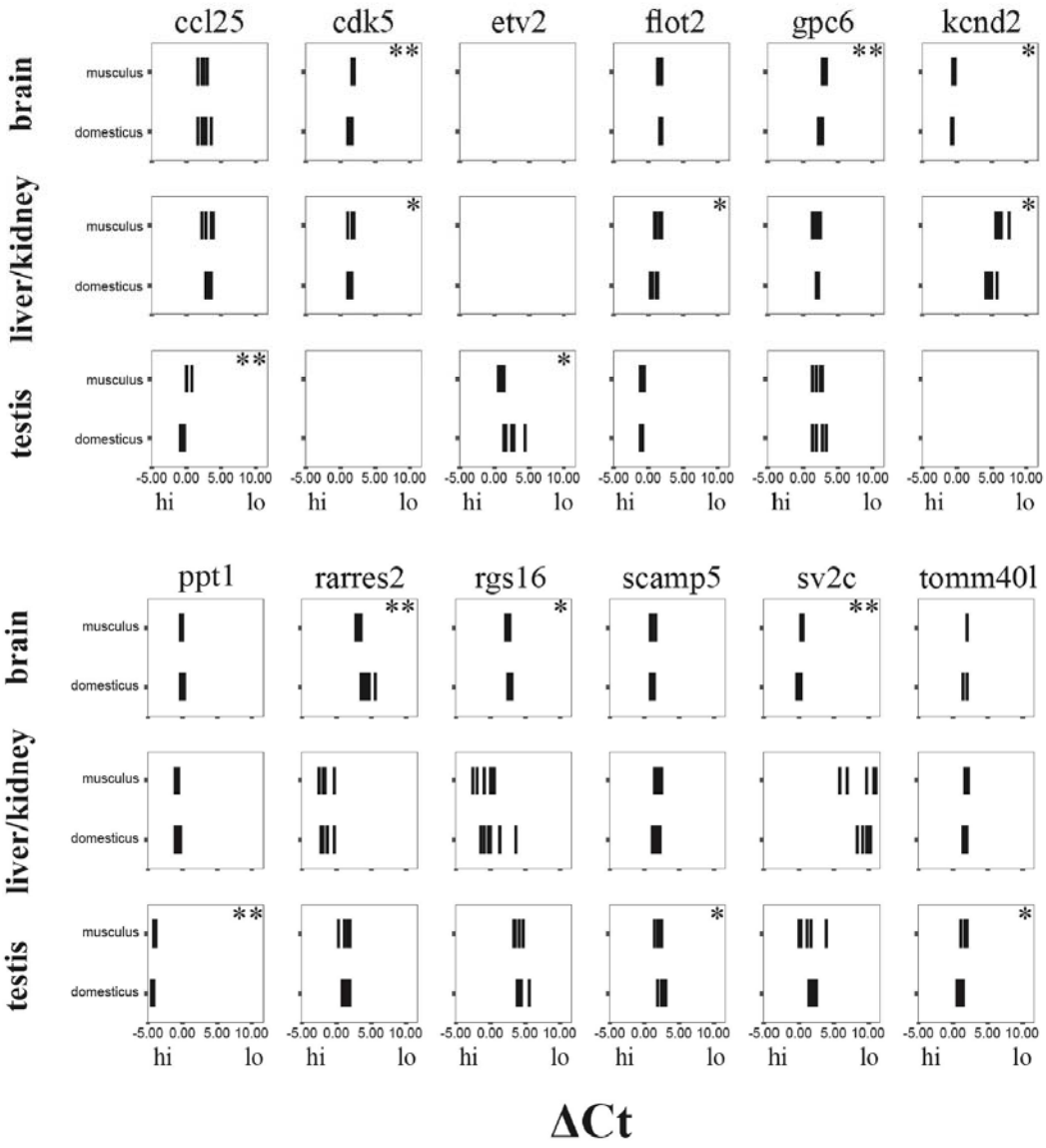


Abb. 8 Expressionsniveaus der zwischen den Subspezies signifikant unterschiedlich exprimierten Gene. Die Sterne markieren signifikant unterschiedliche Wertepaare. ** = p < 0,01; * = p < 0,05; Wilcoxon-W-Test). Jede Vertikale repräsentiert das Expressionsniveau einer einzelnen Maus in ΔCt. Kleine ΔCts stehen für hohe Genexpression (hi), große ΔCts stehen für niedrige Genexpression (lo).

MGI Gene Symbol			Gene expression						Genetic diversity	
Acc. Num.	Chr	Subsp	Brain		Liver/kidney		Testis		θ	D
			Expr	Res	Expr	Res	Expr	Res		
1110017D15Rik	4	dom	0.02	-0.51	0.002	-0.26	30.319	0.08	0.52	1.48
AK003742.1		mus	0.021	-0.32	0.002	0.31	45.277	-0.27	2.44	0.73
1700125F08Rik	10	dom	n.e.	n.e.	n.e.	n.e.	1.066	0.47	5.37	0.96
AK007277		mus	n.e.	n.e.	n.e.	n.e.	0.883	0.3	1.03	0.25
Cacng2	15	dom	1.448	-1.45	n.e.	n.e.	0.019	-0.55	0.91	1.78
NM_007583		mus	1.448	-0.74	n.e.	n.e.	0.014	-0.38	1.19	1.97
Ccl25	8	dom	0.182	0.69	0.127	0.02	1.513**	-0.49	1.08	-1.45
NM_009138		mus	0.212	0.5	0.109	0.72	0.943**	0.2	1.37	0.71
Cdk5	5	dom	0.390**	0.39	0.448*	-0.5	n.e.	n.e.	0	n.c.
NM_007668		mus	0.286**	-0.38	0.352*	-0.46	n.e.	n.e.	0.27	0.96
Etd	X	dom	0.004	0.88	n.e.	n.e.	0.78	0.53	0.26	-0.76
NM_175147.2		mus	0.002	0.3	n.e.	n.e.	0.603	-0.28	0.99	1.64
Etv2	7	dom	n.e.	n.e.	n.e.	n.e.	0.199*	1.06	1.37	-0.99
NM_007959		mus	n.e.	n.e.	n.e.	n.e.	0.495*	0.13	0.67	-0.77
Flot2	11	dom	0.294	-0.8	0.659*	0.37	2.077	-0.83	0.29	0.57
NM_008028.1		mus	0.321	0.23	0.386*	0.26	1.991	-0.11	1.48	0.63
Gpc6	14	dom	0.173**	-0.1	0.254	-0.59	0.215	0.63	0.45	0.96
NM_011821		mus	0.125**	0	0.266	0.02	0.255	0.69	4.99	-1.05
Hif1a	12	dom	0.602	0.14	2.074	-0.7	2.538	-0.57	2.21	1.35
NM_010431		mus	0.543	0.51	1.741	-0.1	2.205	0.5	1.8	1.25
Kcnd2	6	dom	1.739*	-0.4	0.040*	-0.16	n.e.	n.e.	1.97	0.84
NM_019697		mus	1.488*	-0.17	0.015*	-0.54	n.e.	n.e.	1.12	1.57
Krt2-17	15	dom	0.003	0.63	n.e.	n.e.	n.e.	n.e.	3.29	0.23
NM_010668		mus	0.002	0.11	n.e.	n.e.	n.e.	n.e.	0	n.c.
Mir16	7	dom	0.733	-1	2.869	-1.16	4.329	0.44	0.4	0.57
NM_019580		mus	0.721	-0.33	2.964	-0.15	5.319	-0.29	1.99	-0.87
Nf1	11	dom	0.567	-1.13	0.503	-0.38	3.62	-0.16	0.98	1.25
NM_010897.1		mus	0.54	0.11	0.482	-0.3	3.311	0.65	0	n.c.
PanX1	9	dom	0.072	0.62	0.025	-0.01	0.233	-0.04	3.94	1.58
NM_019482		mus	0.082	0.16	0.038	-0.32	0.221	-0.11	1.32	-0.18
Ppt1	4	dom	1.081	0.22	1.737	-0.07	20.933**	-0.85	1.28	2.43
NM_008917.1		mus	1.208	-0.3	1.795	-0.66	17.291**	-0.97	0.82	-1.5
Rab4b	7	dom	0.188	-1.21	0.203	-1.29	1.211	-1.63	0.75	0.6
NM_029391.1		mus	0.203	-0.48	0.244	-0.29	1.579	0.52	0.75	-0.89
Rarres2	6	dom	0.051**	1.07	3.183	0.97	0.393	0.23	3.04	1.58
NM_027852.1		mus	0.114**	0.6	3.963	1.04	0.423	0.6	2.1	1.91
Rgs16	1	dom	0.143*	0.01	0.768	1.82	0.051	-0.14	1.55	2.38
NM_011267.1		mus	0.194*	0.04	1.807	1.48	0.075	0.2	2.51	-1.31
Scamp5	9	dom	0.55	0.15	0.321	0.24	0.168*	0.04	4.07	-1.73
NM_020270		mus	0.498	0.42	0.28	0.3	0.245*	0.15	0.5	1.61
Sv2c	13	dom	0.974**	0.05	0.002	0.36	0.264	0.06	3	1.77
AK173092.1		mus	0.788**	-0.82	0.003	1.49	0.392	1.62	2.09	0.3
Tcte3	17	dom	0.001	0.05	0.004	0.8	48.593	0.08	1.06	1.51
NM_011560.2		mus	0.001	-0.68	0.007	0.9	48.673	0.45	1.52	-1.21
Tmem24	9	dom	0.42	0.09	0.604	-1.12	0.413	-0.18	0.36	-1.14
NM_027909.1		mus	0.493	0.57	0.707	-0.23	0.305	0.16	0.75	1.4
Tomm40l	1	dom	0.295	0.04	0.344	-0.2	0.498*	0.13	1.37	0.14
AK186544.1		mus	0.281	-1.43	0.293	-0.2	0.340*	0.16	0.82	0.24

Tabelle 3
Zusammenfassung der Expressions- und Sequenzpolymorphismusdaten aller untersuchten Gene. Expr: lineare Expressionsniveaus basierend auf ΔCt; Res: Residuen der Varianzregression (siehe 2.3); θ: Watterson's θ per site (Watterson 1975) x10³; D: Tajima's D (Tajima 1989). Die Sterne repräsentieren signifikant unterschiedliche Expression zwischen den Subspezies (= p < 0,01; * = p < 0,05).**

2.4.2 Genetische Variabilität

Um die genetische Variabilität innerhalb der untersuchten Populationen messen zu können und mögliche Anzeichen natürlicher Selektion zu detektieren, wurde bis zu 1 kb der Stromaufwärtsregion der 24 Gene dieser Studie sequenziert. In der Stromaufwärtsregion von Genen sind regulatorische Elemente zu erwarten, welche die basale Transkriptionsmaschinerie steuern. Eine Übersicht der sequenzierten Bereiche befindet sich in Kapitel 4.5, Anzahl der sequenzierten Chromosomen und die Länge der Sequenzen gehen aus Tabelle S 5 hervor.

Auf Basis dieser Sequenzdaten wurden Watterson's θ (Watterson 1975) und Tajima's D (Tajima 1989) berechnet (Tabelle 3). Keines der Ds ist signifikant negativ, d.h. es gibt in den vorliegenden Daten keinen Beleg für die Wirkung positiver Selektion auf einen potentiellen Promotor in nächster Nähe der differentiell exprimierten Gene. Andererseits ist die statistische Aussagekraft (Power) von Tajima's D-Test in unserem Datensatz aufgrund einer verhältnismäßig kleinen Zahl variabler Positionen klein (Simonsen, Churchill und Aquadro 1995).

Wenn Selective Sweeps in *cis* für die Änderungen der Expressionsniveaus verantwortlich wären, würde man erwarten, dass eine Reduktion der genetischen Variabilität in einer der beiden Subspezies an Expressionsunterschiede gekoppelt wäre. Um zu untersuchen, ob dies der Fall sein könnte, wurde das Verhältnis der θ der beiden Populationen unabhängig für jede untersuchte Region berechnet. Der Betrag des natürlichen Logarithmus dieses Verhältnisses gibt dann an, ob ein großer Unterschied der genetischen Variabilität zwischen den Populationen besteht, unabhängig davon, welche Population im Zähler und welche im Nenner steht, ähnlich der lnRH oder lnRV Statistik (Kauer, Dieringer und Schlotterer 2003). Das Mittel dieser Beträge wurde zwischen den signifikant unterschiedlich exprimierten Genen einerseits und den gleich exprimierten Genen andererseits verglichen. Es besteht kein Unterschied zwischen den beiden Gruppen ($p = 0.63$, Wilcoxon-W-Test). Daher wurde keine Evidenz dafür gefunden, dass Expressionsunterschiede mit einer Reduktion der genetischen Variabilität einhergehen. Demnach spielen Selective

Sweeps zumindest keine erhebliche Rolle für die Verwirklichung von Expressionsunterschieden.

2.4.3 Korrelation von Sequenz- und Expressionspolymorphismus

Eines der Ziele der vorliegenden Arbeit war herauszufinden, ob und in welchem Umfang der Polymorphismus auf Sequenzebene die Expressionsvarianz in den untersuchten Populationen beeinflusst. Sowohl Expressionsniveaus, als auch deren Varianzen unterscheiden sich voraussichtlicht zwischen Subspezies, Genen und Geweben. Daher wurden diese Faktoren im Rahmen einer linearen Modellierung untersucht (Tabelle 4).

model	k	AICc	ΔAICc	wAICc	BIC
θ	3	233.858	0	0.322	242.119
Subspecies + θ	4	234.154	0.296	0.278	245.099
Subspecies * θ	5	236.254	2.396	0.097	249.847
Tissue * θ	7	237.249	3.391	0.059	256.026
Tissue + Subspecies + θ + Tissue:θ	8	237.663	3.805	0.048	258.973
Tissue + θ	5	237.159	3.301	0.062	250.752
Tissue + Subspecies + θ	6	237.507	3.649	0.052	253.711
Tissue + Subspecies + θ + Subspecies:θ + Tissue:θ	9	239.834	5.976	0.016	263.638
Tissue + Subspecies + θ + Subspecies:θ	7	239.669	5.811	0.018	258.446
Tissue + Subspecies + θ + Tissue:Subspecies + Tissue:θ	10	240.786	6.928	0.01	267.042
Tissue + Subspecies + θ + Tissue:Subspecies	8	240.448	6.59	0.012	261.758
Tissue + Subspecies + θ + Subspecies:θ + Tissue:θ + Tissue:Subspecies	11	242.984	9.126	0.003	271.650
Subspecies	3	241.105	7.248	0.009	249.366
Nullmodel	2	241.13	7.272	0.008	246.672
Tissue * Subspecies * θ	13	246.803	12.945	0	280.157
Tissue + Subspecies	5	244.187	10.329	0.002	257.780
Tissue	4	244.161	10.304	0.002	255.106
Tissue * Subspecies	7	247.061	13.203	0	265.838

Tabelle 4 Vergleich der Modelle zur Erklärung des Expressionspolymorphismus. Die Modelle sind nach AICc geordnet. Das AICc erfasst die durch das Modell erklärte Varianz („goodness of fit"), bestraft aber zusätzliche Parameter. Ein kleines AICc spricht für hohe Aussagekraft bei gleichzeitiger Berücksichtigung der Parsimonie. k: Anzahl der Parameter; AICc: Akaike's Information Criterion korrigiert für den Stichprobenumfang; wAICc: Akaike Weights korrigiert für den Stichprobenumfang (vergleichbar mit r^2); BIC: Bayesian Information Criterion oder Schwarz Criterion (bestraft zusätzliche Parameter härter); + steht für Additivität; : steht für Interaktion; * steht für Additivität und Interaktion.

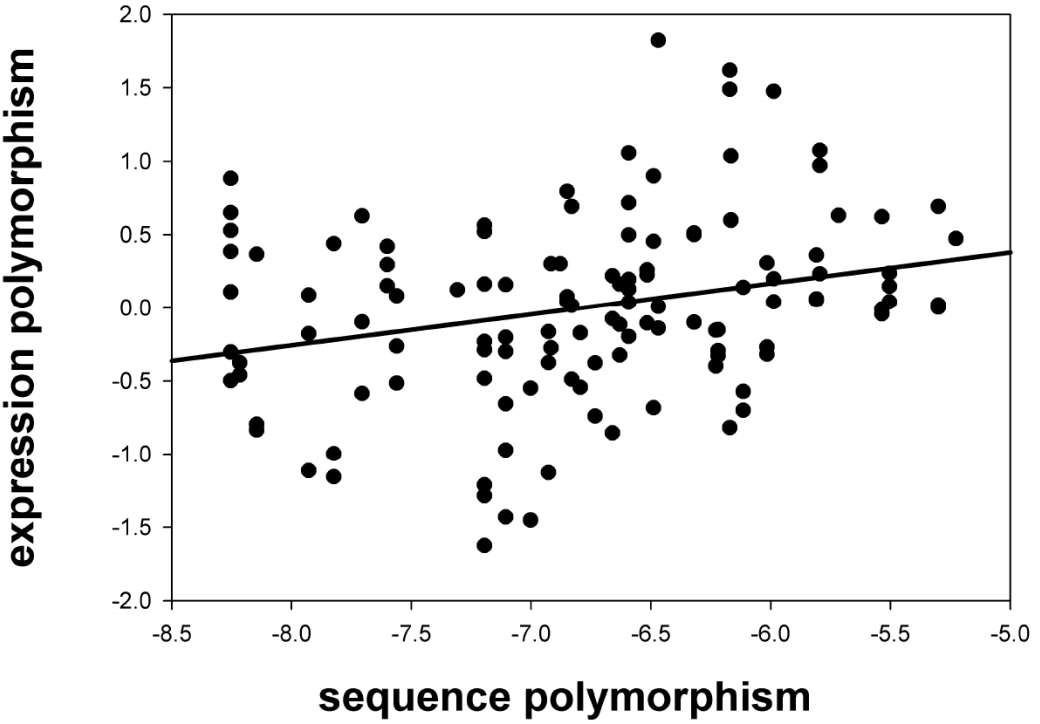


Abb. 9 Expressionspolymorphismus (y-Achse, Residuen) korreliert mit Sequenzpolymorphismus (x-Achse, lnθ). Jeder Datenpunkt repräsentiert den Expressionspolymorphismus in einem bestimmten Gewebe in einer der beiden Populationen ($r^2 = 0{,}073$; $p < 0{,}01$). lnθ basiert auf der Stromaufwärtsregion des selben Gens in der selben Population.

Es wurden 18 Modelle unter Verwendung aller möglichen Kombinationen der Faktoren „subspecies", „tissue" und „lnθ" untersucht. Watterson's θ wurde vor der Analyse logarithmiert, um die Verteilung der θ deutlich an die Normalverteilung anzunähern (S 7). Akaike's corrected Information Criterion (AICc); (Sugiura 1978) wurde verwendet, um das aussagekräftigste Modell zu identifizieren. Eine lineare Abhängigkeit des Expressionspolymorphismus von lnθ ist das bevorzugte Modell, gefolgt von einem Modell, welches „subspecies" als Faktor mit einbezieht (Tabelle 4). In einem Backward Model Selection Ansatz (nichtsignifikante Faktoren werden einer nach dem anderen Entfernt) wurde „subspecies" jedoch als nicht signifikanter Faktor identifiziert ($p = 0{,}18$; ANOVA zwischen dem nur auf lnθ basierenden Modell und dem additiven Modell (lnθ + „subspecies")). Dies spricht deutlich für das auf lnθ basierende Modell als die wahrscheinlichste Erklärung für die Daten unter Berücksichtigung des Parsimonieprinzips.

Beschreibt man diesen so gewonnenen Zusammenhang in einer einfachen Regression erhält man eine hochsignifikante Korrelation zwischen lnθ und dem Expressionspolymorphismus (Abb. 9; $p < 0{,}01$; $r^2 = 0{,}073$).

Demnach scheinen Gene mit einer hohen Sequenzvarianz (lnθ) in der Stromaufwärtsregion im Mittel auch eine höhere biologische Expressionsvarianz innerhalb der selben Population zu haben.

2.4.4 Korrelation von Expressionspolymorphismus und Expressionsdivergenz

Um herauszufinden, ob höherer Expressionspolymorphismus größere Expressionsdivergenz nach sich zieht, wurde für jedes Gen und Gewebe die Expressionsdivergenz als absolute Differenz der Expressionsniveaus zwischen den Subspezies berechnet und durch deren Mittelwert geteilt (siehe 4.1.5).

Die Genexpressionsdivergenz korreliert über die drei untersuchten Gewebe deutlich und hochsignifikant mit der mittleren biologischen Varianz der Genexpression der beiden Populationen ($p < 2x10^{-4}$; $r^2 = 0{,}21$; Abb. 10A). Die Gewebeidentität ist hierbei weder im Vergleich der Modelle (Tabelle 5) noch in der Backward Selection (ANOVA, $p = 0{,}75$) ein signifikanter Faktor und kann daher als unabhängig betrachtet werden.

model	k	AICc	deltaAICc	wAICc	BIC
Polymorphism	3	-13.339	0	0.868	-7.371
Polymorphism+Tissue	5	-9.294	4.045	0.115	0.270
Polymorphism*Tissue	7	-5.483	7.856	0.017	7.333
Tissue	4	2.791	16.13	0	10.597

Tabelle 5 Vergleich linearer Modelle zur Bestimmung der relevanten Faktoren in der Expressionsdivergenz. Abkürzungen wie Tabelle 4.

Obwohl die Expressionsdivergenz insgesamt vom Expressionspolymorphismus abhängt, gibt es Gene mit hoher Expressionsdivergenz trotz geringen Polymorphismus in den Populationen (Abb. 10A). Ein solches Muster könnte die Folge eines rezenten Selektionsereignisses sein, das die Expressionsvarianz stark reduziert hat. Daher wurde überprüft, ob Selective Sweeps in unserem Datensatz eine Rolle spielen könnten. Wäre dies der Fall, würde man eine Tendenz zu reduzierter Expressionsvarianz in der Linie erwarten, die in ihrer Expression divergiert. Eine größere Differenz der Expressionsvarianz zwischen den Populationen für diesen Locus wäre die Folge.

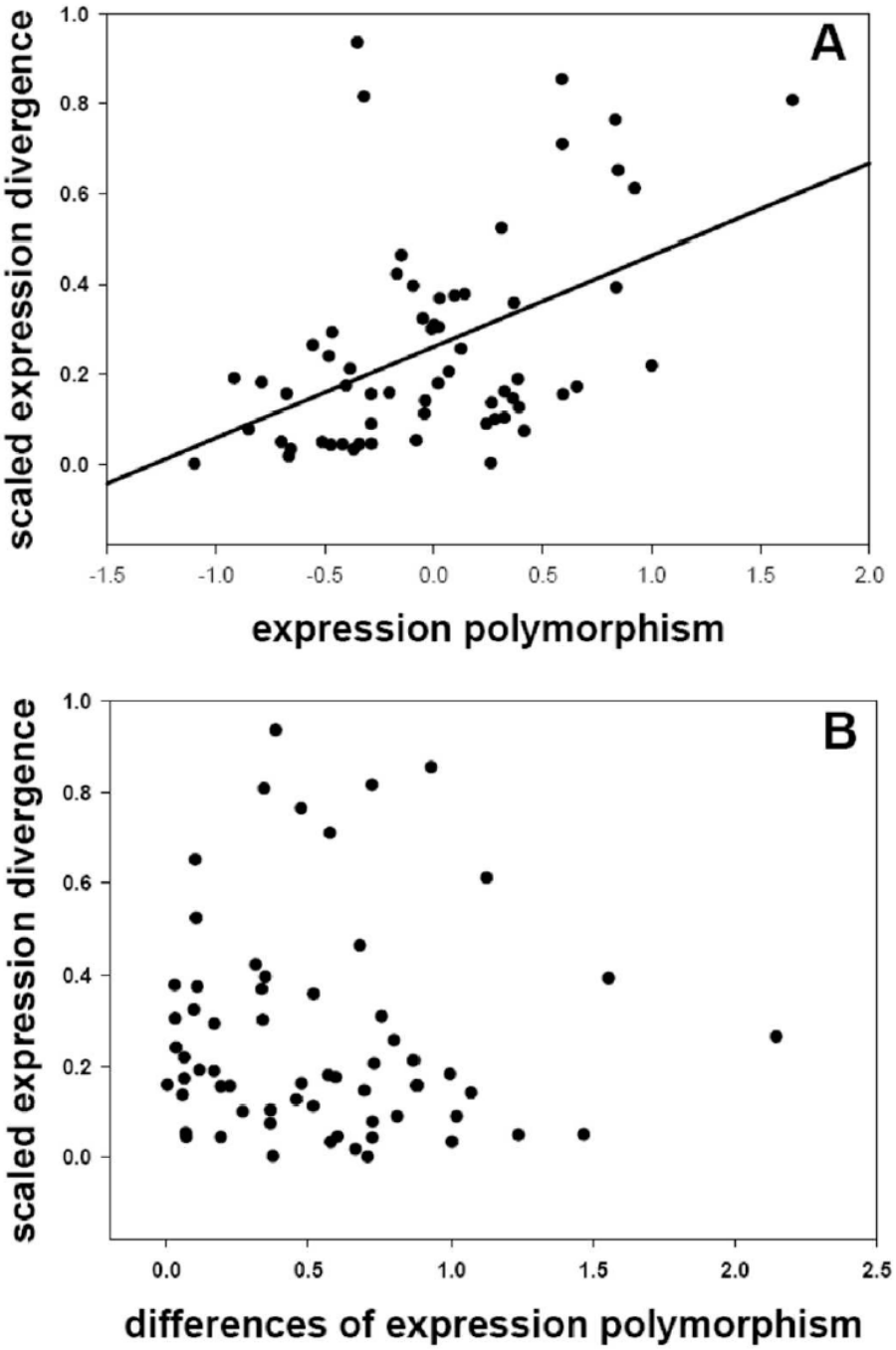


Abb. 10 Zusammenhänge zwischen Polymorphismus und Divergenz. (A) Die Expressionsdivergenz zwischen den Populationen korreliert mit dem Expressionspolymorphismus (Residuen) innerhalb der Populationen. Jeder Datenpunkt repräsentiert die Divergenz eines bestimmten Gens in einem bestimmten Gewebe ($r^2 = 0,21$; $p < 2x10^{-4}$; Pearson's Product Moment Correlation). (B) Die Expressionsdivergenz zwischen den Populationen korreliert nicht mit den Differenzen des Expressionspolymorphismus (Differenz der Residuen) zwischen den Populationen ($r^2 = 0,040$; $p = 0,759$, Pearson's Product Moment Correlation).

Abb. 10B zeigt, dass es einen solchen Zusammenhang nicht gibt – die Differenz der Expressionsvarianzen zwischen den Populationen korreliert nicht mit der Expressionsdivergenz. Obwohl dieses Ergebnis positive Selektion in einigen Fällen nicht ausschließt, kann kein genereller Trend zu positiver Selektion vorliegen.

2.5 Diskussion

Das vorliegende Kapitel zielte darauf, die Zusammenhänge zwischen Expressions- und Sequenzvarianz sowie der Evolution von Expressionsunterschieden genauer zu analysieren. Khaitovich et al. (2004; 2005) nehmen anhand ihrer Microarraydaten an, dass die neutrale Evolutionstheorie auf diese Zusammenhänge angewendet werden könnte. Ihre zentrale Beobachtung war, dass die Varianz der Expressionsniveaus in Menschen mit der Divergenz der Expression zwischen den Primaten korreliert. Es deutet in dieselbe Richtung, dass in dieser Arbeit eine Korrelation der Varianz innerhalb der Populationen mit der Divergenz der Subspezies gefunden wird . Dieses Resultat stützt die Annahme der Additivität: Kleine zufällige Änderungen der Expressionsniveaus können im Laufe der Zeit zu einem größeren Unterschied akkumulieren. Demnach muss es auch einen erheblichen Grad an Kontinuität und Additivität der Allele geben, die Einfluss auf die Genexpression ausüben. Odom et al. (2007) haben den Wandel der Bindestellen von vier Transkriptionsfaktoren an deren Zielgenen zwischen Mensch und Maus untersucht und gefunden, dass 40-90% Veränderungen unterliegen. Dieser Sachverhalt stimmt ebenfalls mit einem Modell kontinuierlicher kleiner Änderungen überein. Andererseits kann Genexpression aufgrund funktionaler Zwänge nicht *ad infinitum* divergieren und gleichzeitig ihre Funktion behalten (Whitehead und Crawford 2006b; Bedford und Hartl 2009).

Kontinuität und Additivität kleiner Effekte lässt vermuten, dass eine Vielzahl von Mutationen involviert ist, wenn sich Expressionsniveaus ändern. Diese These wird von der in der vorliegenden Arbeit beschriebenen Korrelation zwischen Sequenzvariabilität und Expressionsvariabilität gestützt. Einerseits wurden die Sequenzdaten von potentiellen basalen Promotorregionen generiert, die einen direkten Einfluss auf die Expressionsvariabilität in der Population haben könnten, auch wenn es Hinweise auf die Beteilung weiter entfernter Enhancer gibt (siehe unten). Andererseits repräsentieren die Sequenzdaten wahrscheinlich den mittleren

Sequenzpolymorphismus der gesamten Genregion, da das LD in wilden Populationen der Hausmaus durchschnittlich 20kb beträgt (Laurie et al. 2007).

In einer ähnlichen Studie in *Drosophila simulans* haben Lawniczak et al. (2008) ebenfalls eine Korrelation zwischen Sequenzpolymorphismus und Expressionsvariabilität entdeckt. Sie unterscheiden jedoch zwischen unterschiedlichen Genregionen in ihrer Analyse und finden diese Korrelation hauptsächlich in transkribierten Regionen, nicht jedoch in der Stromaufwärtsregion, was unseren Ergebnissen gewissermaßen widerspräche. Die statistischen Ansätze sind jedoch nicht vollständig vergleichbar. Erstens wurden in der Studie von Lawniczak et al. „Light Shotgun" Sequenzdaten verwendet, um den Sequenzpolymorphismus zu ermitteln, was mit kleinerer Coverage pro analysierter Region und daher mit kleinerer statistischer Aussagekraft einhergeht. Zweitens verwenden sie ein Maß für den Expressionspolymorphismus, das weniger spezifisch sein könnte. Sie nehmen an, dass die p-Werte von ANOVAs zwischen den Microarrayexperimenten den Expressionspolymorphismus repräsentieren. Es bleibt in Ihrer Studie jedoch unklar, zu welchem Grad p-Werte natürliche Expressionsvarianz repräsentieren, da sie die technische Varianz der Arrayhybridisierung enthalten. Die Verwendung von qRT-PCR in Kombination mit der Sequenzierung von Sonden- und Primerbindestellen, um Sequenzpolymorphismus auszuschließen, die Trennung von technischer und biologischer Varianz und die große Stichprobe sequenzierter Individuen führt anscheinend zu einer höheren statistischen Aussagekraft in der vorliegenden Arbeit. Ähnliche Argumente treffen auf die Studie von Brown und Feder (2005) zu, in der ebenfalls keine Korrelation zwischen Sequenzpolymorphismus der Stromaufwärtsregion und der Expressionsvariabilität zwischen verschiedenen *Drosophila melanogaster* Stämmen gefunden wurde.

Im Gegensatz zu den *Drosophila* Studien konnte in der vorliegenden Arbeit auch die Gewebespezifität von Veränderungen der Genexpression untersucht werden. Interessanterweise zeigen der größte Teil der unterschiedlich exprimierten Gene (neun von zwölf) differentielle Expression in nur einem der untersuchten Gewebe. Dies legt nahe, dass gewebespezifische Enhancer unabhängig voneinander evolvieren können. Dieses Ergebnis bestätigt das Resultat einer ähnlichen Studie von Blekhman et al. (2008) über gewebespezifische Expressionsdivergenz zwischen drei Primaten (Makaken, Schimpansen und Menschen). Die Autoren finden außerdem eine größere

Anzahl spezifischer Veränderungen im Menschen, verglichen mit den anderen beiden Linien und interpretieren dies als ein Zeichen positiver Selektion.

Nichtsdestotrotz würde man auch unter einem neutralen Modell Fluktuationen zwischen den Linien erwarten und gegeben, dass beide, sowohl die Zahl der betrachteten Linien, als auch der Umfang der Unterschiede klein ist, ist positive Selektion nicht notwendigerweise am beobachteten Muster beteiligt. Es wäre notwendig zu prüfen, ob ein neutrales Modell ebenfalls in der Lage wäre, die Daten der genannten Studie zu erklären.

Es ist eine intensive Debatte um die Rolle von Genexpressionsveränderungen in Entwicklungsprozessen im Gange (Hoekstra und Coyne 2007). Obwohl es überzeugende Evidenz für mehrere solcher Fälle gibt (zusammengefasst in (Carroll 2008)), sind meistens Speziesvergleiche betroffen, die fernab einer evolutiven Distanz sind, in welcher der Prozess, der zu der Änderung geführt hat, noch ermittelt werde kann. So hat z.B. der Fall eines ausführlich beschriebenen regulatorischen Elements, das die Expression von Eigenschaften steuert, die für den Geschlechtsdimorphismus in *Drosophila* verantwortlich sind, eine Stammesgeschichte von mindestens 30 Millionen Jahren (Williams et al. 2008). Beachtet man Generationszeit und Evolutionsraten, entspräche dies einer Divergenz, welche vergleichbar mit der zwischen Fischen und Säugetieren ist. In diesem Zusammenhang über den Effekt von Sexual Selection auf die Evolution von Enhancern zu spekulieren, erscheint unangebracht. Jeder Versuch die Rolle der Genexpression in der Kreation evolutionärer Neuheiten besser zu verstehen, bedarf der Studie näher verwandter Spezies oder Subspezies. Eine detaillierte Studie eines Enhancerelements, das die Pigmentation zwischen nah verwandten *Drosophila* Spezies beeinflusst (Jeong et al. 2008), liefert keine Evidenz für positive Selektion und ist kompatibel mit einem neutralen Divergenzmodell.

Die Tatsache, dass keine Anzeichen positiver Selektion auf die untersuchten Gene, die einen Expressionsunterschied aufweisen, detektierbar sind, legt nahe, dass positive Selektion zumindest kein gravierender Faktor in der Erzeugung dieser Unterschiede sein kann. Dennoch schließen weder die vorliegenden Daten noch andere Populationsanalysen (Whitehead und Crawford 2006a) Fälle von positiver Selektion auf differentielle Genexpression aus. Es wurde von Harr et al. (2006) sogar ein Fall beschrieben, in dem differentielle Promotornutzung mit einem Selective Sweep in der Genregion korreliert. Dennoch weist die weitere Analyse darauf hin, dass auch

nichtsynonyme Mutationen als Auslöser für den Selective Sweep in Frage kommen (Heinen 2008).

Es ist eine generelle Einschränkung von Studien der Genexpression auf mikroevolutiver Ebene, dass nur Unterschiede im Expressionsniveau adulter Tiere gemessen werden, nicht jedoch Veränderungen, die eine Rolle für Entwicklungsprozesse spielen. Dies könnten neue Expressionsdomänen in verschiedenen Regionen des Embryos sein oder Regulation zu verschiedenen Zeitpunkten der Entwicklung. Es bedarf solcher Studien, bevor generelle Schlüsse über die Rolle adaptiver regulatorischer Veränderungen gezogen werden können.

3 Analyse des Poldi Locus

Poldi bzw. der EST 1700125f08Rik (Poldi ist derzeit noch ein vorläufiger Arbeitsname) wurde in einem Microarrayexperiment von Voolstra (2007) als zwischen *M. m. domesticus* und *M. m. musculus* differentiell exprimiert identifiziert. Dies rückte das Gen bereits im ersten Teil dieser Arbeit neben anderen Genen in den Fokus der Aufmerksamkeit. Im Rahmen dieser ersten Analyse konnten Hinweise darauf gefunden werden, dass die genetische Variabilität im Stromaufwärtsbereich des Poldi-Gens in *M. m. musculus* im Vergleich zu *M. m. domesticus* nur ein Fünftel beträgt (Tabelle 3). Ein adaptives Allel könnte die Variabilität der Region in *M. m. musculus*, durch Verdrängen anderer Varianten, reduziert haben (Selective Sweep).

Interessanterweise lassen fehlende homologe Transkripte in Mensch und Ratte bei konservierter Syntänie eine sehr junge Phylogenie des Gens vermuten. Poldi könnte ein evolutionäres Novum (Orphan-Gen) sein.

Um die Stammesgeschichte des Locus sowie die Möglichkeit eines Selective Sweep genauer zu studieren, werden der Poldi Locus und die dort annotierten Transkripte in diesem Kapitel analysiert, beginnend mit einer Übersicht über den Locus und seine beiden Transkripte.

3.1 Der Poldi Locus und seine annotierten Transkripte

Die mRNA Annotation am Poldi Locus ist in Abb. 11 wiedergegeben. Zwei gegenläufige Transkripte sind auf Chromosom 10, basierend auf ESTs, annotiert. Die zu Poldi (1700125f08Rik) gehörenden ESTs stammen ausschließlich aus dem Hoden und umfassen drei Exons, während die zum gegenläufigen Transkript gehörenden ESTs aus dem visuellen Kortex stammen und bis zu fünf Exons umfassen (AK158810). Abb. 18A zeigt, dass aber auch im Testis Bereiche transkribiert sind, die teilweise auf dem EST AK158810 lägen.

Poldi besitzt zwei potentielle ORFs im dritten Exon von denen der in Transkriptionsrichtung zweite als kodierend annotiert ist und das 128 Aminosäuren lange, hypothetische Protein EDL32162 (GenBank) kodiert. Der vordere ORF würde für 106 Aminosäuren kodieren.

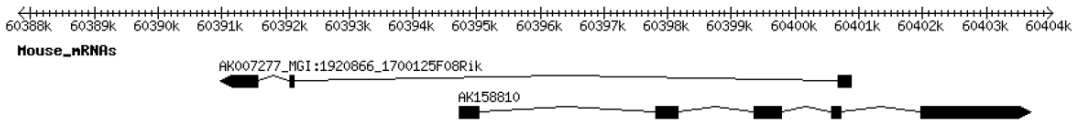


Abb. 11 Übersicht der beiden annotierten Transkripte am Poldi Locus
Die obere Leiste gibt die Position auf Chromosom 10 in Kilobasen an. Erstellt mittels http://gbrowse.informatics.jax.org/ .

Der EST AK158810 trägt einen potentiellen ORF auf dem dritten Exon (91aa) und einen zweiten ORF (110aa), welcher Teile des vierten und fünften Exons umfasst. Keiner dieser ORFs ist als kodierend annotiert.

Beide potentiellen AK158810 ORFs liegen im hinteren Bereich des Transkripts. Es existiert jedoch nur ein einzelner EST, der außer dem ersten und dem zweiten Exon noch weitere Exons enthält. Dieser ist AK158810. Eine Translation wäre also unwahrscheinlich für die beiden weiteren annotierten ESTs (BY273633 und BY278135). Zweitens ist das dritte ATG Startcodon des potentiellen ORFs auf Exon 3, also des vorderen und kürzeren dieser ORFs. Der hintere ORF beginnt demnach noch einige ATGs später.

Auch den potentiellen Poldi ORFs gehen Stromaufwärts gelegene ATGs voran.

Nach dem „scanning model“ (Kozak 1978) ist das erste ATG eines Transkripts das Startcodon. Von einer Translation der beiden Transkripte und der Entstehung funktionaler Proteine ist auf dieser Grundlage nicht auszugehen.

3.2 Populationsgenetische Struktur der Poldi-Region in vier natürlichen Populationen der Hausmaus

Ein Selective Sweep hinterlässt in der Population, in der er stattgefunden hat, eine populationsgenetisch erfassbare Signatur. Da die Selektion eines vorteilhaften Allels nicht nur den Anstieg der Frequenz des vorteilhaften Allels selbst, sondern auch der physikalisch gekoppelten Region zur Folge hat (Genetic Hitchhiking (Maynard Smith und Haigh 1974)), ist diese Signatur als ausgedehntes Tal reduzierter Variabilität identifizierbar. Im ersten Teil dieser Arbeit (Tabelle 3) konnte bereits eine stark unterschiedliche genetische Variabilität zwischen der deutschen zu *M. m. domesticus* gehörenden Population und der tschechischen zu *M. m. musculus* gehörenden Population detektiert werden (θ_{Ger}/S = 0,00103; θ_{Cze}/S =0,00537). Um diesen ersten

Hinweis auf ein Selektionsereignis zu prüfen ist es erforderlich die Poldi umgebende genomische Region populationsgenetisch zu untersuchen. Nur so kann geklärt werden, ob ein größeres Tal reduzierter Variabilität vorliegt, das der Signatur eines Selective Sweep entspräche, oder ob es sich um eine zufällige, punktuelle Schwankung handelt.

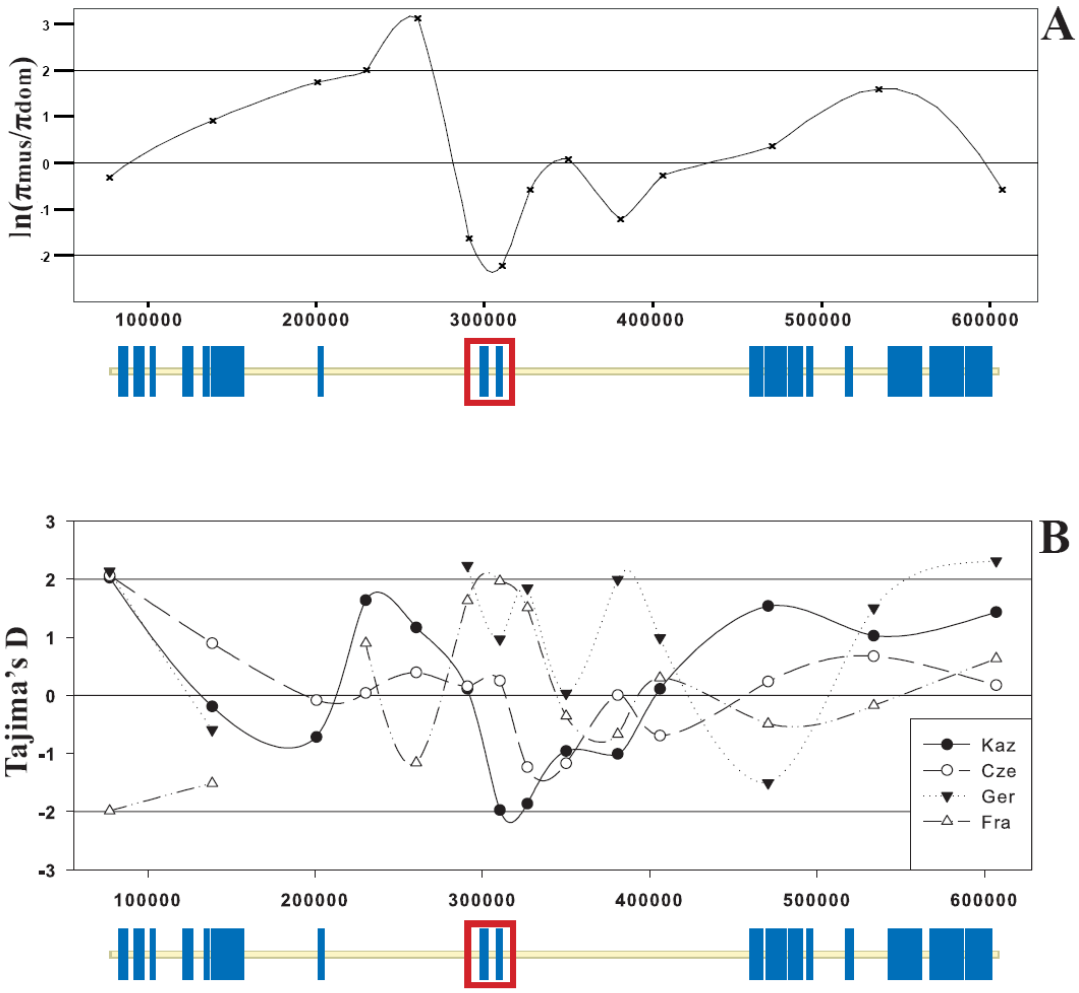


Abb. 12 Genetische Variabilität in der Poldi Region
A: Verhältnis der genetischen Variabilität zwischen den untersuchten Subspezies. Hierfür wurden die Nucleotide Diversity Per Site (π) der *M. m. musculus* Populationen (Kaz,Cze) und der *M. m. domesticus* Populationen (Ger,Fra) gemittelt. Das Verhältnis des mittleren Subspezies π wurde logarithmiert und aufgetragen. Jeder Datenpunkt entspricht einem sequenzierten Bereich. Die x-Achse markiert Distanzen in Basenpaaren. Unterhalb des Graphen befindet sich eine schematische Darstellung der Region. Die zwei vertikalen blauen Linien in der Mitte (rot umrahmt) markieren die Position des ersten und letzten Poldi Exons. Die übrigen blauen Linien gehören zu den benachbarten Genen Unc5b (links) und Pcbd1/Sgpl1 (rechts).
B: Tajima's Ds für die Poldi Region. Jeder Datenpunkt repräsentiert das D für ein Sequenzfragment in der entsprechenden Population. Die x-Achse markiert Distanzen in Basenpaaren. Drei Loci in der deutschen Population (Ger) und ein Locus in der französischen (Fra) wurden ausgenommen, da sie keine variablen Stellen (SNPs) enthielten (S 1).

Zu diesem Zweck wurden 14 Fragmente in einer 500 kb Region in regelmäßigen Abständen zu Poldi sequenziert. Jeweils elf Individuen einer kasachischen (Kaz), tschechischen (Cze), deutschen (Ger) und französischen (Fra) Hausmauspopulation wurden analysiert. Die kasachische und die tschechische Population gehören der Subspezies *M. m. musculus* an, während die deutsche und französische Population *M. m. domesticus* angehört. Durchschnittlich 550bp pro Fragment (S 1) wurden populationsgenetisch ausgewertet.

In Abb. 12A ist eine deutliche Reduktion der genetischen Variabilität in *M. m. musculus* in direkter Umgebung des Poldi Gens erkennbar. Diese Reduktion bildet ein Variabilitätstal in *M. m. musculus* von etwa 100 Kilobasen. Die genetische Variabilität in *M. m. musculus* nimmt dann mit der Annäherung an die benachbarten Gene wieder zu. Ein Einfluss benachbarter Gene auf die Reduktion der Variabilität wird damit unwahrscheinlich.

Trägt man das Verhältnis der Variabilitäten der einzelnen Populationen gegen die Position auf (S 2), bestätigt sich das Bild aus Abb. 12A für die Vergleiche zwischen den Subspezies. Sobald Populationen der gleichen Subspezies miteinander verglichen werden, verschwindet das Muster zugunsten eines zufälligen Hintergrundrauschens. Demnach handelt es sich offenbar um einen subspeziesspezifischen Effekt.

Beide *M. m. musculus* Populationen weisen ein niedriges Tajima's D nahe des Poldi Locus auf (Abb. 12B). Zwei Sequenzabschnitte in nächster Nähe des Poldi Gens zeigen eine Verschiebung des Frequenzspektrums hin zu seltenen Allelen in der kasachischen Population, die nicht mit einem neutralen Evolutionsmodell vereinbar ist ($D = -1{,}97$, $p < 0.05$ und $D = -1{,}86$, $p < 0{,}05$), sondern auf positive oder negative Selektion hinweist.

Zusammen mit dem ausgeprägten Variabilitätstal, das der typischen Signatur eines Selective Sweep entspricht, sind die signifikant niedrigen Ds am ehesten als Folge positiver Selektion zu deuten.

Für drei Loci in der deutschen Population und einen in der tschechischen konnten in Ermangelung variabler Stellen keine Tajima's Ds ermittelt werden (S 1). Die angrenzenden Ds weisen nicht auf positive Selektion hin ($D = -0{,}59$ und $D = 2{,}23$ in der deutschen Population, $D = -1{,}51$ und $D = 0{,}9$ in der französischen Mauspopulation).

Sowohl π als auch D haben in *M. m. musculus* ihr Minimum exakt an der Position des Poldi Transkripts. Da sich im Variabilitätstal und in einer Umgebung von nahezu

300kb des Poldi Locus kein anderes Gen befindet, ist die Rückführung eines Selektionsereignisses auf Poldi möglich.

Das scheinbar stärkere Selektionssignal in der kasachischen Population im Vergleich zur tschechischen Population könnte auf die starke Abhängigkeit der Sensibilität (Power) des D-Tests von der Anzahl variabler Stellen sein (Simonsen, Churchill und Aquadro 1995). Die kasachische Population weist am Fragment des D-Minimums (D = -1,97) fünf variable Stellen auf, während die tschechische nur drei variable Stellen zu verzeichnen hat.

Kleine ZnS Werte (Kelly 1997) von 0,10 (Kaz) und 0,04 (Cze) geben Anlass zur Annahme, dass dem vorliegenden Variabilitätsmuster eine Sternphylogenie in *M. m. musculus* zugrunde liegt, d.h. die vorhandenen Haplotypen stammen vermutlich von einem Haplotypen ab, der durch einen Selective Sweep fixiert wurde. Dies und die Reduktion der Variabilität in beiden *M. m. musculus* Populationen begünstigt eine Interpretation des populationsgenetischen Musters am Poldi Locus als subspeziesspezifisches Selektionsereignis.

Gleichzeitig wird es durch die weiträumige Abwesenheit anderer Gene als des Poldi unwahrscheinlich, dass das Maximum in Abb. 12A, welches eine Reduktion der Variabilität in *M. m. domesticus* widerspiegelt, eine Folge von Selektion ist.

Vergleicht man die genetische Variabilität der 24 über das Mausgenom verteilten Loci aus dem ersten Teil dieser Arbeit, stellt man fest, dass weder π ($p = 0,42$, Wilcoxon W-Test) noch θ ($p = 0,77$, Wilcoxon W-Test) noch Tajima's D ($p = 0,28$, Wilcoxon W-Test) sich zwischen den Subspezies unterscheiden. Es gibt also keinen Anhaltspunkt für unterschiedliche demographische Szenarien zwischen deutscher *M. m. domesticus* und tschechischer *M. m. musculus* Population. Die Analyse sieben unabhängiger Loci von Harr (unveröffentlicht) in den vier Populationen, die auch Objekt der vorliegenden Studie sind, gibt ebenfalls keine Anhaltspunkte für genomweit unterschiedliche Populationsparameter, wie sie z.B. durch Bottlenecks verursacht würden.

3.3 Potentielle Ursachen eines Selektionsereignisses am Poldi Locus

Ursachen eines Selective Sweep verhelfen dem Träger zu einem Selektionsvorteil. Ohne die Ausprägung einer vorteilhaften Eigenschaft kann natürliche Selektion nicht

wirksam werden. Ein solches vererbbares phänotypisches Merkmal hat einen genetischen Ursprung, der in der Sweep-Region, dem Tal reduzierter Variabilität, zu finden ist (Maynard Smith und Haigh 1974; Kim und Stephan 2002).

In diesem Absatz sollen potentielle Ursachen eines Selective Sweeps in der Poldi Region gefunden werden. In Frage kommen Mutationen, die in *M. m. musculus* durch den Sweep fixiert sind und gleichzeitig funktionale Konsequenzen haben könnten.

Zu diesem Zweck wurden Sequenzdaten der Exons beider Transkripte (1700125f08Rik und Ak158810) in den vier Populationen dieser Studie erhoben (elf kasachische, 17 tschechische, 17 deutsche und elf französische Mäuse) und hinsichtlich funktionaler Mutationen untersucht. Hierunter fallen insbesondere Mutationen, die Einfluss auf die Transkriptstruktur haben, also Spleißerkennungsstellen und Mutationen, die in den potentiellen ORFs liegen. Nichtsynonyme Basensubstitutionen ändern das entstehende Protein und können so funktional sein.

Ob eine Änderung der Transkriptstruktur zwischen den Subspezies vorliegt wurde mittels PCR und Gelelektrophorese sowie Sequenzierung auf cDNA-Basis überprüft.

Neben Proteinsequenz und Transkriptstruktur kann auch die Regulation von Poldi einem adaptiven Vorteil zugrunde liegen. Adaptive cisregulatorische Elemente können ebenfalls ein Variabilitätstal verursachen. Poldi (1700125f08Rik) wurde im ersten Teil dieser Arbeit bereits analysiert, weil Voolstra (2007) es in einer Microarrayanalyse als unterschiedlich exprimiert zwischen *M. m. musculus* und *M. m. domesticus* identifiziert hat. Die RT-PCR Daten aus Teil 1 werden hier genauer betrachtet.

Um zu unterscheiden, ob ein fixiertes Allel durch Neumutation entstanden ist oder der bereits vorhandenen Variation in der Population - der Standing Variation - entstammt, muss festgestellt werden, ob das Sweep-Allel anzestral oder abgeleitet ist. Daher wurden Sequenzdaten von phylogenetisch entfernteren Spezies (*Mus musculus castaneus* (CAS)*, Mus cypriacus* (CYP)*, Mus macedonicus* (MAC)*, Mus spicilegus* (SPI)*, Mus spretus* (SPR)*, Mus famulus* (FAM) und *Mus caroli* (CAR)) generiert (Datenträger: Sequenzanhang.doc) und eine Konsensussequenz erstellt, die den anzestralen Status der Allele repräsentiert.

3.3.1 Poldi Transkription

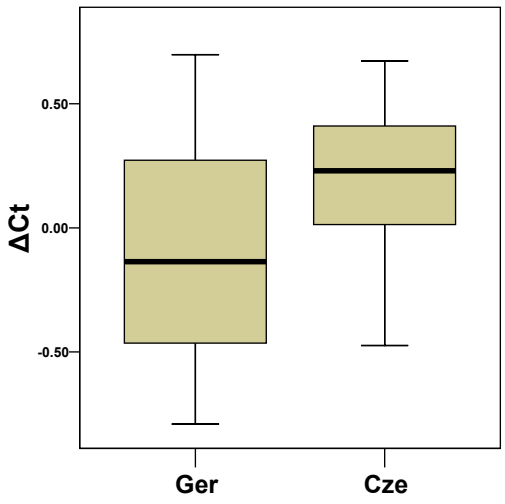


Abb. 13 Poldi Genexpression im Hoden
ΔCt Ermittelt in sechs tschechischen (Cze) und sechs deutschen (Ger) Individuen. Fehlerbalken : 95%CI.

In Hirn, Leber und Niere ist Poldi in den zwölf Individuen dieser Studie nicht exprimiert, sondern ausschließlich im Hoden. Die ESTquellen, der GNF Expression Atlas 2 (http://genome.ucsc.edu/) und der ESTProfileViewer belegen ebenfalls die Testisspezifität (http://www.ncbi.nlm.nih.gov/UniGene/ESTProfileViewer.cgi?uglist=Mm.159038) der Poldi Expression.

Der von Voolstra (2007) gefundene Expressionsunterschied konnte mittels qRT-PCR nicht verifiziert werden. Die Mediane unterscheiden sich nicht signifikant ($p = 0.49$, Wilcoxon W-Test; Abb. 13), offenbar auf Grund der hohen Varianz. Deutet sich ein niedrigeres Expressionsniveau in der tschechischen Population (höherer ΔCt) an, so reicht die statistische Aussagekraft nicht dieses aufzulösen.

Liegt kein Expressionsunterschied vor, ist ein Selektionsvorteil durch Regulation der Poldi Transkription dennoch nicht auszuschließen. So könnte ein bestimmtes Expressionsniveau von Vorteil sein, das auch in der Nicht-Sweep-Population mit gewisser Frequenz auftritt. Wird ein solches Expressionsniveau aus der Standing Variation durch eine Änderung der äußeren Umstände adaptiv, würde man eine reduzierte Variabilität der Expression mit dem adaptiven Expressionsniveau als Mittel in der Sweep-Population vermuten. Auch dies entspricht nicht den Daten ($p = 0{,}22$, Levene-Test). Ein adaptiver Vorteil einer bestimmten Transkriptmenge in adulten Mäusen wird damit als Ursache des Sweeps unwahrscheinlich.

Räumliche Expressionsunterschiede von Poldi innerhalb des Hodens kommen anscheinend auch nicht in Frage, da Tobias Heinen in einem *in situ* Experiment (unveröffentlicht) die Expression sowohl in *M. m. musculus* als auch in *M. m. domesticus* ausschließlich in postmeiotischen Spermatiden nachweisen konnte. Zeitliche Regulation, z.B. zu bestimmten Phasen der Embryonalentwicklung kann

jedoch als Ursachen der Selektionssignatur am Poldi-Locus nicht ausgeschlossen werden.

3.3.2 Transkriptstruktur des Poldi Gens

Um festzustellen, ob eine Veränderung der Transkriptstruktur zwischen den Subspezies vorliegt, die den Selective Sweep hervorgerufen haben könnte, wurde cDNA der Individuen aus dem qRT-PCR-Experiment sequenziert (sechs Deutsche, sechs Tschechische). Die Primer wurden im ersten und dritten Exon platziert, so dass PCR-Produkte entstehen, welche die Exongrenzen überspannen und so Aufschluss über die Transkriptstruktur geben.

Außerdem wurden die Exons und ihre flankierende Sequenz auf der Basis genomischer DNA in kasachischen, tschechischen, deutschen und französischen Mäusen mit dem Ziel untersucht, spleißrelevante Mutationen identifizieren zu können. Hierzu wurden die Exon-Intron-Grenzen sowie etwaige Verzweigungspunkte der während des Spleißvorgangs entstehenden Lassostruktur entsprechend Abb. 14 untersucht.

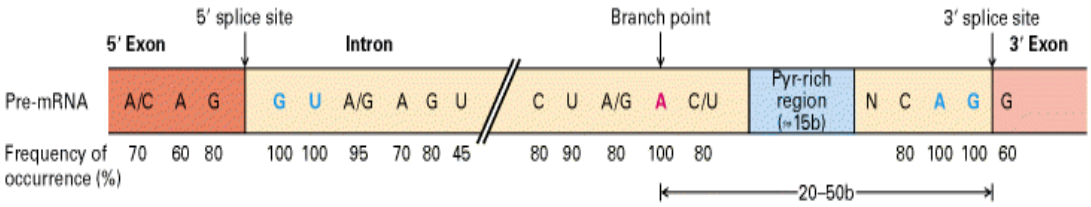


Abb. 14 Schematische Übersicht der untersuchten Spleißrelevanten Stellen
Aus Molecular Cell Biology (Lodish et al. 1999)

Stellen, die für das Spleißen von Bedeutung sind, sind zwischen den Populationen konserviert. Ein einziger Polymorphismus am Beginn des zweiten Exons segregiert niedrigfrequent auf zwei von 22 französischen Chromosomen. Die vollständige Sequenzinformation befindet sich auf dem anhängenden Datenträger (Sequenzanhang.doc, Poldi Exon2).

Die cDNA Sequenzierung zeigt differentielles Spleißen des zweiten Exons. An den Exongrenzen beginnen sich die Nukleotidsequenzen der jeweils anderen Exons zu überlagern. So folgen auf die 3’ Exongrenze des ersten Exons gleichzeitig Sequenzen die dem zweiten und dem dritten Exon zuzuordnen sind. Das Verhältnis der Höhen

der Maxima ist zwischen den Individuen konstant. Dies deutet darauf hin, dass auch das Verhältnis der Spleißprodukte konstant ist.

Eine unterschiedliche Transkriptstruktur, sowie die vermehrte Produktion eines bestimmten Spleißproduktes, kommen daher als Auslöser eines selektiven Ereignisses nicht in Frage.

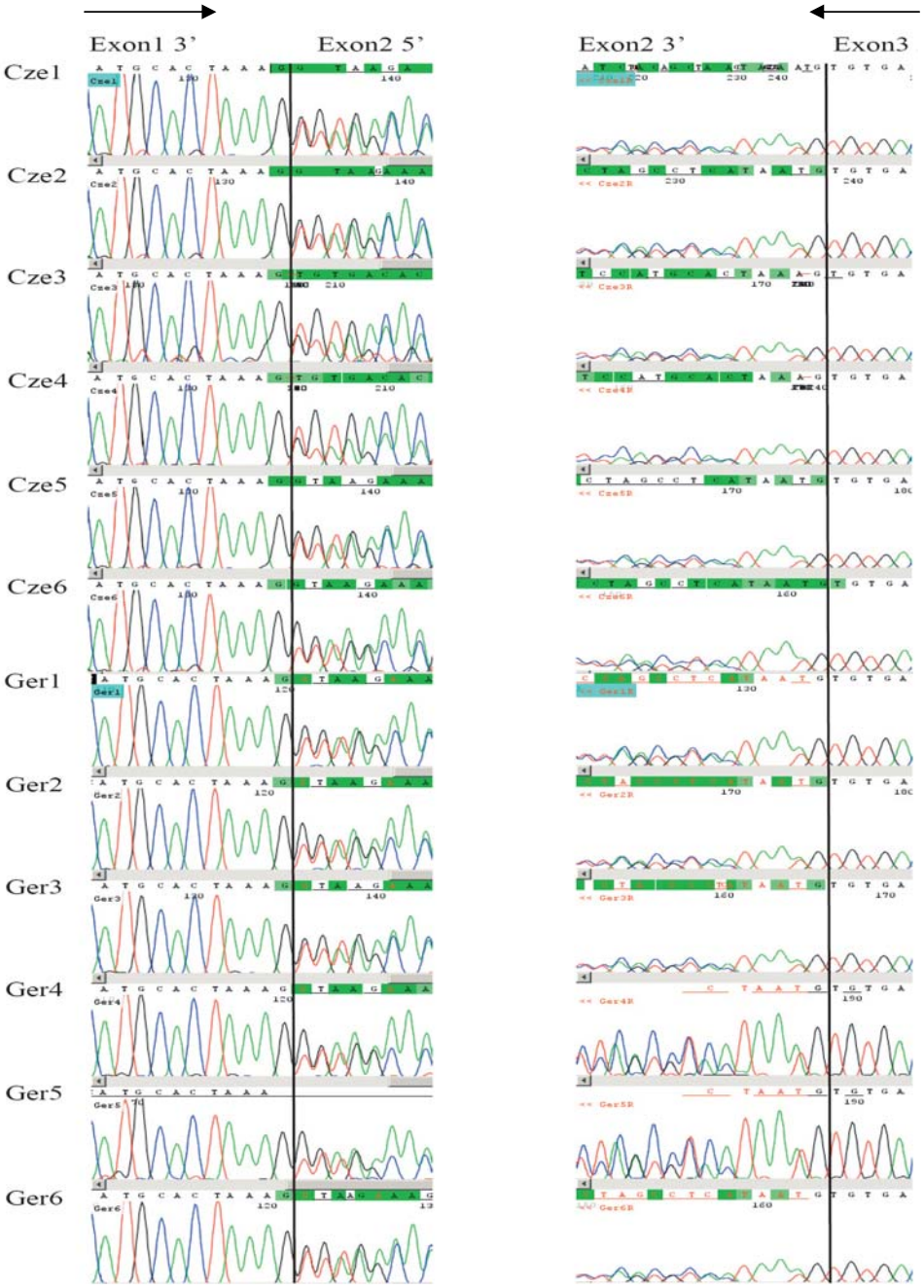


Abb. 15 Elektropherogramme der Poldi cDNA-SequenzenLinks: Exongrenze am 3' Ende des ersten Exons, rechts: Exongrenze am 5' Ende des dritten Exons. Die Pfeile markieren die Leserichtung. Die sich überlagernden Sequenzen belegen differentielles Spleißen des zweiten Exons in allen Tieren gleichermaßen.

3.3.3 Kodierende Sequenz des Poldi Gens

Ob Poldi Transkripte translatiert werden ist unklar. 5' der Startcodons der beiden oben besprochenen ORFs befinden sich bereits drei ATGs im Poldi Transkript (ein ATG im 3' Bereich des ersten Exons (Datenträger: Sequenzanhang.doc, Poldi Exon1 Pos. 236) und zwei weitere ATGs auf dem zweiten Exon (Datenträger: Sequenzanhang.doc, Poldi Exon2 Pos. 44 und 97)), die aber für sehr kleine ORFs kodieren. Solche Stromaufwärts ATGs und ORFs sind zwar auch bei kodierenden Genen bekannt (Crowe, Wang und Rothnagel 2006) und können eine Rolle in der Regulation des Gens spielen (Meijer und Thomas 2002), die Translationsinitiation nach dem „scanning model" (Kozak 1978), in dem das Ribosom ausgehend vom 5'Cap die RNA entlang wandert und die Translation am ersten ATG beginnt, ist damit aber ausgeschlossen. Hinzu kommt, dass keines der im Transkript enthaltenen ATGs in eine Translationsinitiationskonsensussequenz (Nakagawa et al. 2008) eingebettet ist oder deutliche Ähnlichkeit zu einer solchen aufweist.

Mehrere unterschiedliche Varianten der beiden möglicherweise kodierenden ORFs auf 1700125f08Rik segregieren in den Populationen dieser Studie (Abb. 16). In den *M. m. musculus* Populationen weist ORF1 (Abb. 16A) keine fixierten und gleichzeitig abgeleiteten Aminosäuren auf. Das Selektionsereignis wurde demnach nicht von einer Neumutation in diesem ORF ausgelöst. Die Positionen 21, 53 und 63 zeigen fixierte Aminosäuren, die in den *M. m. domesticus* Populationen segregieren, in *M. m. musculus* jedoch fixiert sind. Da die anzestrale Variante fixiert ist, käme nur die Selektion einer bereits vorhandenen Variante durch Änderung der äußeren Umstände als Auslöser des Selektionsereignisses in Frage. Dies entspräche einem Selective Sweep aus der Standing Variation.

ORF2 (Abb. 16B) enthält eine fixierte und gleichzeitig abgeleitete Aminosäure an Position 63. Im Gegensatz zum Glycin (G) der *M. m. domesticus* Populationen und der Außengruppen, befindet sich an dieser Position in *M. m. musculus* Glutaminsäure (E). Dieser Austausch kann durchaus funktionale Konsequenzen für das entstehende Protein haben. Glycin ist die kleinste Aminosäure und außerdem unpolar, und hydrophob. Es wurde in *M. m. musculus* durch die deutlich größere, saure und negativ geladene Glutaminsäure ersetzt.

Es fällt auf, dass Leserastermutationen des zweiten ORF in allen Populationen segregieren. Eine in den beiden *M. m. domesticus* Populationen segregierende

Insertion von vier Basenpaaren (Datenträger: Sequenzanhang.doc, Poldi Exon 3) führt zu einem kürzeren Leseraster von 108 aa des potentiellen Proteins. In den *M. m. musculus* Populationen tritt die 4 bp Insertion nicht auf, stattdessen segregiert eine Deletion von einem Basenpaar. Diese Deletion führt zu einem Verlust des Stoppcodons und würde das Ablösen des Ribosoms während der Transkriptionstermination stören.

Zwar ist es möglich, dass das Protein auch als ternärer Komplex mit dem Ribosom aktiv sein könnte und seine Funktion bewahrt. Dies bleibt jedoch spekulativ. Weil eine demnach nicht funktionale Variante in den *M. m. musculus* Populationen mit nicht zu vernachlässigender Frequenz auftritt, ist eher davon auszugehen, dass Veränderungen des zweiten ORFs nicht die Ursache eines Selektionsereignisses darstellen.

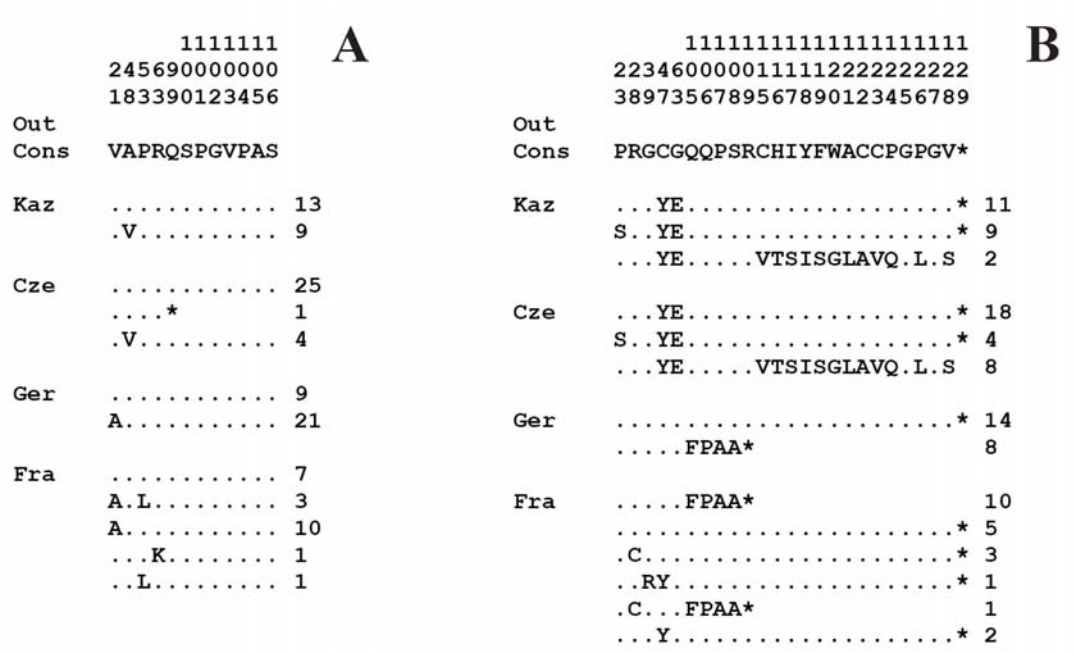


Abb. 16 Variable Aminosäuren der beiden hypothetischen Poldi Proteine
Über der Sequenz sind die Positionen der Aminosäuren relativ zum Startcodon angegeben. Die Sequenz OutCons gibt das Protein an, das vom Konsensus der Außengruppen des Genus *Mus* kodiert würde (anzestrale Sequenz). * = Stopcodon. Hinter der jeweiligen Sequenz ist die Häufigkeit des potentiellen Proteins auf der Basis von Haplotypen in den untersuchten Populationen angegeben.
A: Protein des ORF1 B: Protein des ORF2

3.3.4 Struktur der Poldi-RNA

Poldi könnte auch als strukturelle RNA aktiv sein. In diesem Fall kommen auch fixierte Mutationen innerhalb des Transkripts aber außerhalb der ORFs als Auslöser eines Selektionsereignisses in Frage. Im Transkript befinden sich insgesamt drei

zwischen den Subspezies fixierte Unterschiede. Zwei befinden sich in Exon1 an Position 180 und 217 (Datenträger: Sequenzanhang.doc, Poldi Exon1). Ein Weiterer in Exon3 an Position 432 (Sequenzanhang.doc, Poldi Exon3). Diese Mutation wurde schon im vorangegangenen Abschnitt diskutiert und könnte auch für eine Aminosäuresubstitution verantwortlich sein. Jede dieser drei Stellen zeigt das abgeleitete Allel in *M. m. musculus*. Selektion auf eine dieser Mutationen wäre also mit einem Modell der Selektion auf eine Neumutation kompatibel. Es haben jedoch nur die beiden 3' gelegenen Mutationen einen Einfluß auf die *in silico* vorhergesagte Sekundärstruktur der entstehenden RNA (Abb. 17). Diese beiden Mutationen könnten die Ursache eines Selektionereignisses in *M. m. musculus* sein, falls Poldi als strukturelle RNA wirkt.

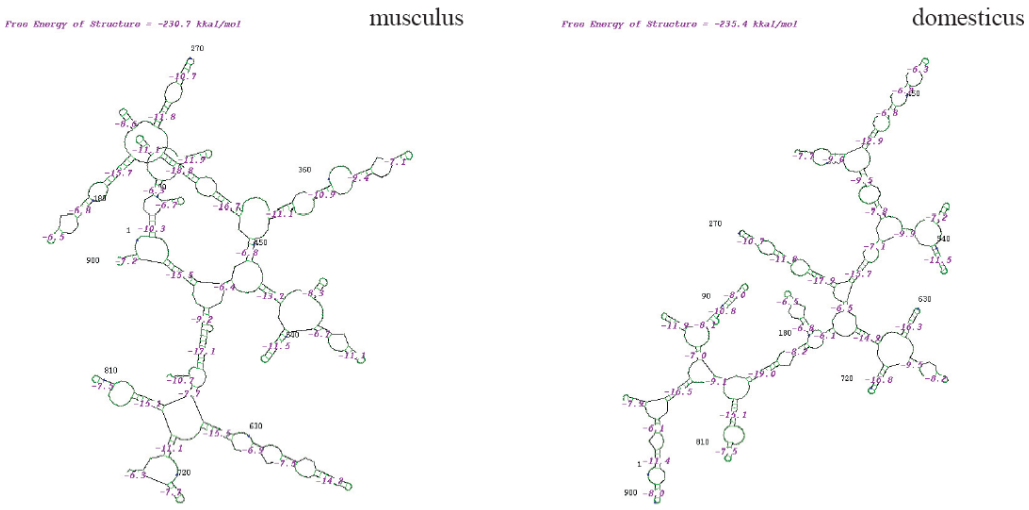


Abb. 17 RNA-Sekundärstruktur des Poldi Gens
Die aus dem *musculus* Haplotypen (links) generierte Sekundärstruktur unterscheidet sich von der von einem *domesticus* Haplotypen abgeleiteten Sekundärstruktur (rechts). Erstellt mittels GeneBee RNA secondary structure prediction.
(http://www.genebee.msu.su/services/rna2_reduced.html)

3.3.5 AK158810

Die fünf Exons des AK158810 Transkripts wurden wie oben beschrieben in den vier Populationen dieser Studie und den Außengruppen sequenziert und eine Analyse der 5' Transkriptstruktur auf cDNA-Basis durchgeführt.

Die Primer für die Analyse auf cDNA-Basis wurden so konzipiert, dass ein 378 bp langes Fragment, das die Exongrenze zwischen erstem und zweitem Exon überspannt,

amplifiziert werden sollte. In der Tat wurde aber eine Vielzahl Fragmente unterschiedlicher Größe amplifiziert (Abb. 18).

Abbildung Abb. 18A belegt die Existenz multipler 5' Transkripte in den Hoden, die keinem der bekannten ESTs (378bp Fragment) entsprechen. Es existieren nur in einigen Individuen aber in beiden untersuchten Populationen Transkripte. Etwa zehn verschiedene Transkriptvarianten können in den Hoden unterschieden werden. Darunter sowohl kurze zwischen 200 bp und 400 bp, als auch sehr lange Varianten über 1500 bp (schwache Banden deutsches Individuum ganz links).

Sowohl in den sechs tschechischen *M. m. musculus* (Abb. 18B links) als auch den sechs deutschen *M. m. domesticus* (Abb. 18B rechts) sind Transkripte im Hirn vorhanden. Aus dem Hirn stammen auch die annotierten ESTs. Die wenigsten dieser Transkripte entsprechen der erwarteten Länge von 378 bp. Einige sind etwas länger als 400 bp. Eine kürzere Variante von unter 200 bp scheint in allen Individuen vorhanden. Es lassen sich wieder etwa zehn verschiedene Varianten unterscheiden, die aber nicht den in Abb. 18A identifizierten entsprechen müssen.

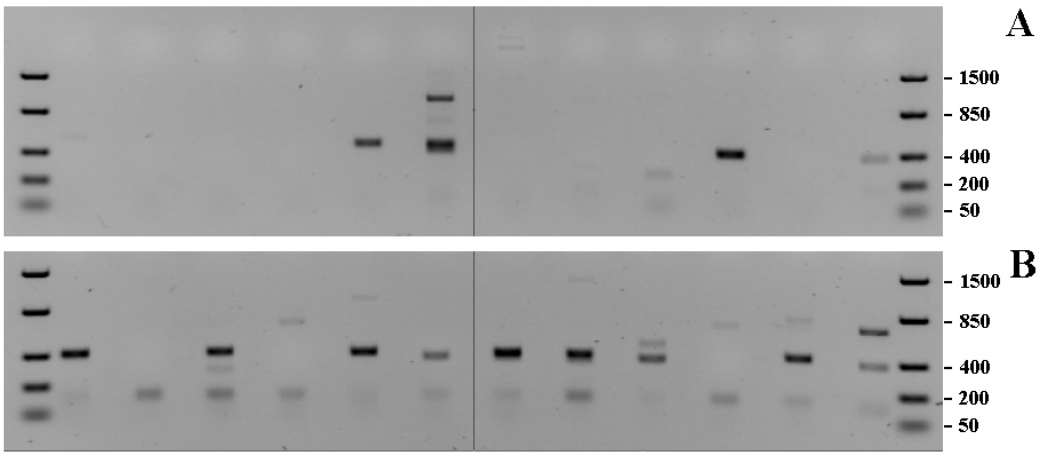


Abb. 18 Gelelektrophorese der AK158810 cDNA-PCR
Die Primer wurden so gewählt, dass die Exongrenzen des ersten und zweiten Exons in der PCR überschritten werden. Erwartet wurde ein Produkt von 378bp Länge. Links: sechs Individuen der tschechischen Population, rechts: sechs deutsche Individuen.
A: PCR-Produkte mit Hoden-cDNA als Ausgangsmaterial
B: PCR-Produkte mit Hirn-cDNA als Ausgangsmaterial

Mit dem Ziel zu überprüfen, ob es sich bei den multiplen Transkripten um PCR-Artefakte handeln könnte und um die Herkunft der zusätzlich transkribierten Bereiche aus dem Genom zu klären, wurden PCRs auf cDNA-Basis aus dem Hoden

sequenziert. Es konnten Teile der 5’ Bereiche von AK158810 aus beiden untersuchten Subspezies sequenziert werden (Abb. 19).

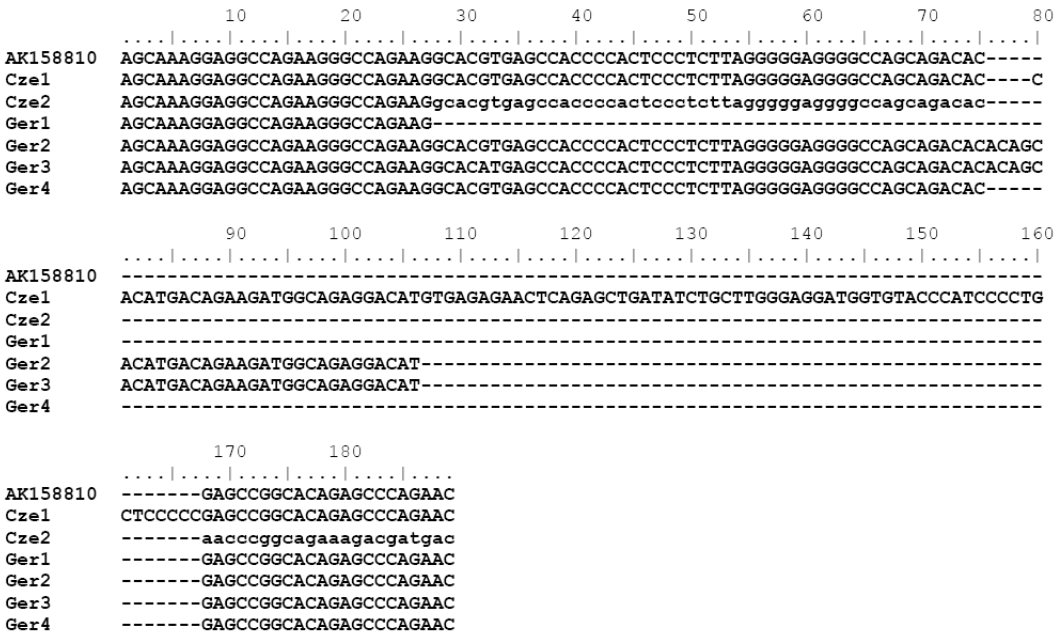


Abb. 19 cDNA Sequenzen der Exon1 und Exon2 übergreifenden PCR-Produkte
Die Ergebnisse der Sequenzierung sind mit dem annotierten EST AK158810 aligniert (oberste Zeile).
Vier verschiedene Längenvarianten können unterschieden werden. Cze1, Ger1 und Ger4 sind verschieden. In Ger 2 und Ger 3 ist die vierte Variante realisiert. Cze 2 hat Transkripte dreier Varianten die sich teilweise überlagern (Kleinbuchstaben).

Vier verschiedene Varianten konnten anhand der Basensequenz identifiziert werden. Nur die Variante in Tier Ger4 entspricht der Annotation der ESTs. Ger1 zeigt ein 47 bp kürzeres Exon1. Sowohl die längste Variante, realisiert in Cze1, als auch die Variante aus Ger2 und Ger3 enthalten zusätzliche Basen. Diese Basenabfolge kann an einen Bereich zwischen Exon1 und Exon2 aligniert werden (Abb. 20). Die PCR-Produkte unterschiedlicher Länge aus
Abb. 19 sind demnach keine PCR-Artefakte oder aus anderen Teilen des Genoms stammende, unspezifische Produkte. Stattdessen handelt es sich um transkribierte Bereiche unterschiedlicher Länge am Locus.
Die Elektropherogramme der aus Cze2 gewonnenen Sequenz weisen ab dem klein gedruckten Bereich zwei sich überlagernde Sequenzen auf. An der annotierten Exongrenze kommen weitere Übergelagerungen der Sequenz hinzu. Diese

Überlagerungen sind als cDNAs verschiedener Länge interpretierbar. In der cDNA von Cze2 sind mindestens drei verschieden lange Transkriptvarianten repräsentiert.

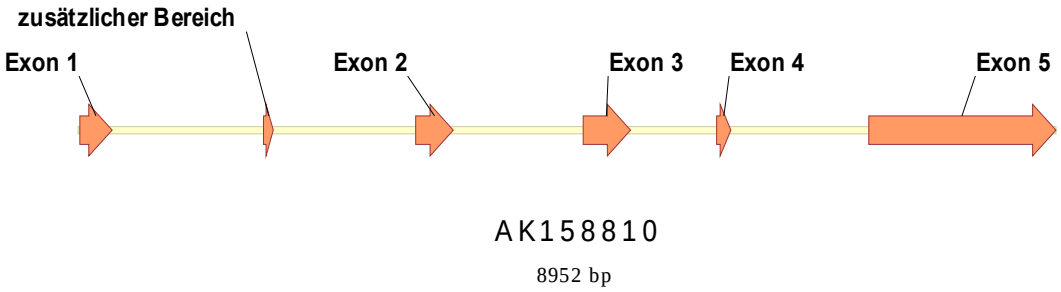


Abb. 20 Übersicht des Transkripts AK158810 inklusive des nicht annotierten, aber dennoch in verschiedenen Varianten transkribierten Bereichs (zusätzlicher Bereich)

Die Sequenzanalyse auf genomischer DNA hinsichtlich Polymorphismen, die den unterschiedlichen Spleißvarianten zugrunde liegen könnten zeigt, dass die Spleißerkennungsstellen konserviert sind (Abb. 21)

In der Abbildung fällt auf, dass der zusätzlich zu Exon1 und Exon2 transkribierte Bereich weder 5' noch 3' den Spleißkonsensussequenzen entspricht. Das differentielle Spleißen scheint entweder äußerst komplex und ungewöhnlich gesteuert zu sein (ohne die genomweit hochkonservierten Konsensussequenzen) oder es ist ein Zufallsprodukt.

Tobias Heinen konnte in einem Northernblot auf Hoden-RNA mit einer AK158810 spezifischen Sonde in *M. cypriacus*, *M. macedonicus* und *M. spicilegus* einen Schmier detektieren (Abb. 22). Dieser Schmier entspräche einer Mixtur vieler verschiedenlanger, unspezifischer Transkripte am Locus. Aufgrund der Konservierung in den oben aufgeführten Außengruppen (Datenträger: Sequenzanhang.doc, AK158810 Außengruppensequezen), wäre dieser Schmier auch in der *M. musculus* Gruppe zu erwarten. Da bereits im 5' Bereich von AK158810 viele verschiedene Transkriptvarianten in *M. musculus* identifiziert werden konnten (s.o.), ist zu erwarten, dass mit zunehmender Länge in 3' Richtung zusätzliche Varianten hinzukommen, die zusammengenommen den im Northernblot sichtbaren Schmier bilden. Entweder wurden für den Northernblot Individuen der *musculus* Gruppe verwendet, die kein Transkript exprimieren, wie manche Individuen in Abb. 18, oder das Transkript ist sehr schwach exprimiert.

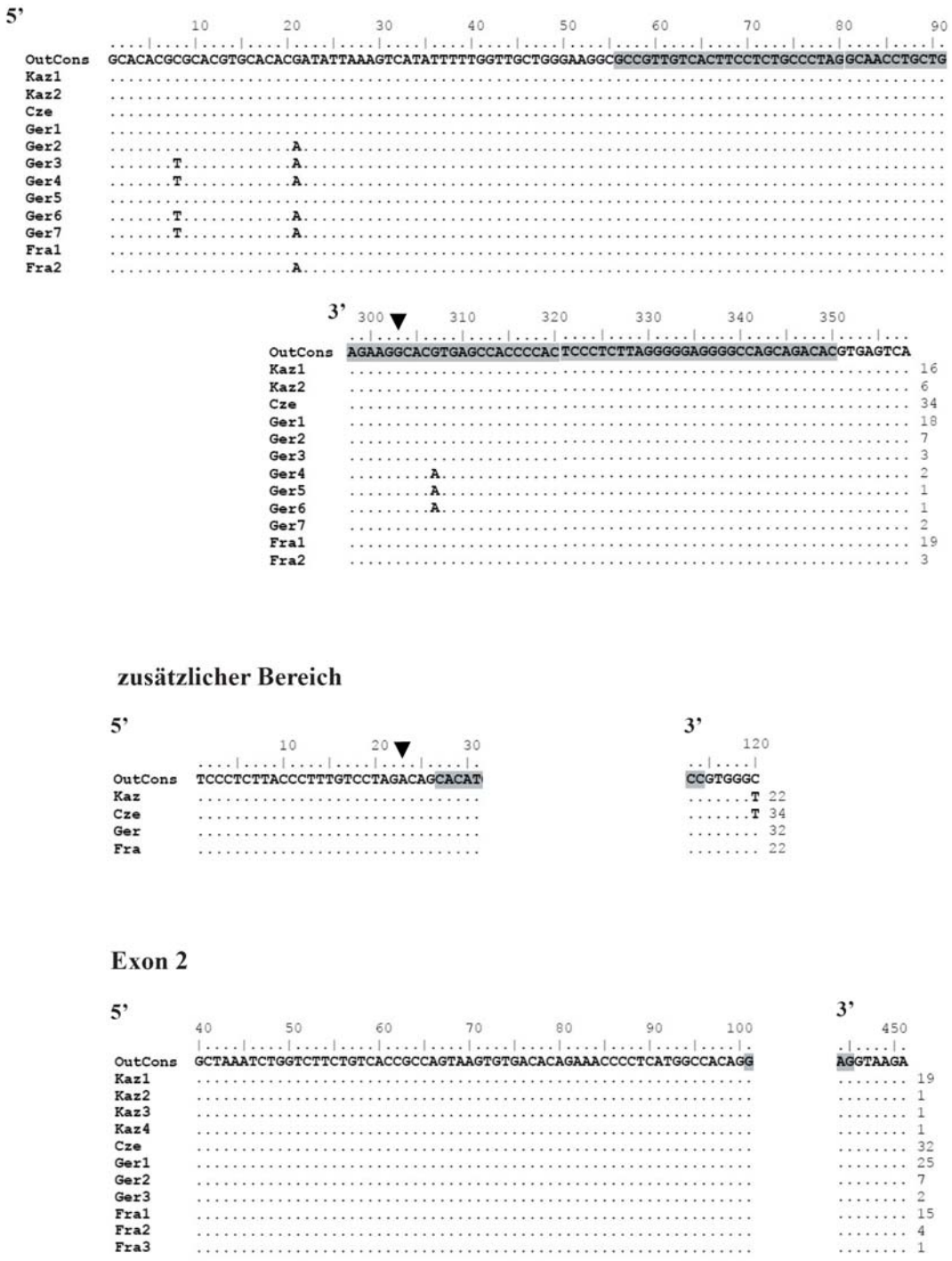


Abb. 21 Spleißrelevante Stellen im 5' Bereich von AK158810 und Stromaufwärtsbereich Exonsequenzen und eine Variante des zusätzlich transkribierten Bereichs sind grau markiert. In zusätzlichen Transkriptvarianten bricht das erste Exon früher ab oder der zusätzlich transkribierte Bereich beginnt früher (schwarze Pfeile).

Ein Indiz für ein äußerst niedriges Expressionsniveau ist die ungewöhnlich lange Belichtungszeit, die nötig war, um im Northernblot ein Signal zu detektieren (persönliche Kommunikation mit Heinen). Für ein niedriges Expressionsniveau spricht ebenfalls die kleine Anzahl (drei) annotierter ESTs die AK158810 ganz oder teilweise entsprechen. Die Detektion der verschiedenen 5' Transkriptvarianten (Abb. 18) widerspricht einem niedrigen Expressionsniveau nicht, da mittels PCR selbst kleinste Transkriptmengen aufgespürt werden können und eine hohe Anzahl von Zyklen (40) für die Analyse verwendet wurde.

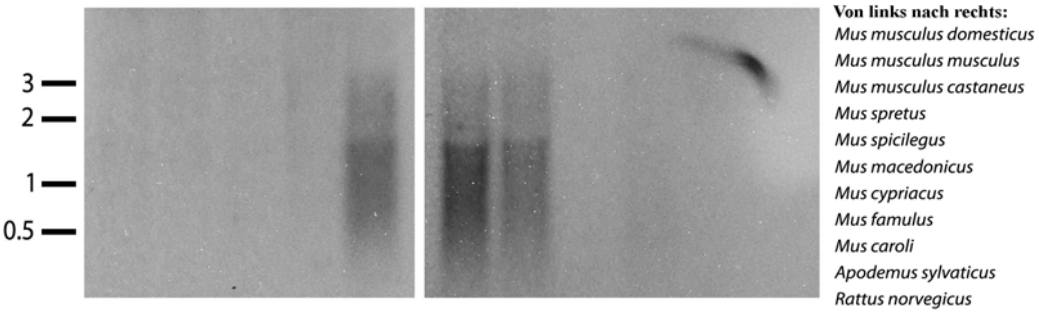


Abb. 22 AK158810 Northernblot von Tobias Heinen (unpubliziert)
Die Größenskala links markiert Größen in Kilobasen.

AK158810 produziert vermutlich kein funktionales Protein. Die kurzen ORFs (91aa, 110 aa) sind auf einem seltenen EST. Erst das dritte ATG wäre Startcodon des kürzeren der beiden ORFs. Das Startcodon des längeren ORFs läge noch weiter hinten. Keines der ATGs ist in eine Translationsinitiationskonsensussequenz eingebettet. Keiner der beiden ORFs würde für ein bekanntes Protein kodieren (blastx der „Non-redundant protein sequence" Datenbank auf http://blast.ncbi.nlm.nih.gov/).
Des Weiteren konnte bereits im 5' Bereich von AK158810 eine derartige Vielzahl von Varianten, die nicht auf alternative Spleißerkennungsstellen zurückzuführen sind, detektiert werden, dass entweder von hochkomplexer Regulation oder unspezifischer Transkription ausgegangen werden muss. Ein Northernblot offenbart einen unspezifischen Schmier in engverwandten Spezies. Zusätzlich sprechen der Northernblot und die wenigen ESTs in den Datenbanken für ein niedriges Expressionsniveau, bei einer durchschnittlichen Transkriptlänge von 1kb. Diese unstrukturierte, und äußerst niedrige Transkription wäre auch für funktionale, nichtkodierende RNAs ungewöhnlich.

In der Stromaufwärtsregion und nahe dem annotierten Transkriptionsstart sind wie auch bei Poldi keine klassischen Promotorelemente vorhanden. Weder TATA-Box, noch BRE-Element, noch ein Down Stream Promoter Element (DPE) sind vorhanden, nicht in der Datenbanksequenz, nicht in den *musculus* Populationen und auch nicht in den Außengruppen des Genus.

Damit fehlen AK158810 viele Merkmale, die für funktionale Gene typisch sind. Unspezifische Transkription scheint unter den gegebenen Umständen am ehesten geeignet die gefundenen ESTs zu erklären. Mit der Anwendung genomweiter Tilingarrays wurde klar, dass ein großer Teil des Säugetiergenoms transkribiert ist. So sind bis zu 93% des menschlichen Genoms transkribiert (Birney et al. 2007). Da jedoch nur etwa 5% des Genoms die Spuren negativer Selektion aufweist, wird ein beträchtlicher Teil dessen nicht funktionalen Elementen zugeordnet, sondern z.B. Lecks transkriptionaler Aktivität in transkribierten Regionen oder einem transkriptionalen Hintergrundrauschen.

Handelt es sich um unspezifische und damit nonfunktionale Transkription, ist auch eine Rückführung des Selektionsereignisses auf AK158810 oder einen der kürzeren, überlappenden ESTs nicht möglich.

3.4 Phylogenie des Poldi Gens

Es besteht Syntänie zwischen Maus, Ratte und Mensch für die genomische Region in der sich Poldi befindet. Die flankierenden Gene liegen auf dem menschlichen Chromosom 10 und sind ebenfalls PCBD1 und UNCB5. Auch auf dem Rattenchromosom 20 konnten die Homologen dieser Gene als nächste Nachbarn identifiziert werden.

Die Konservierung der homologen Sequenzen zwischen Maus, Ratte und Mensch ermöglicht ein umfassendes Alignment. Es befinden sich keine großen Lücken in diesem Alignment (Abb. 23). Ein etwas stärkeres Konservierungsmuster des dritten Exons könnte vermuten lassen, dass einer der ORFs konserviert sei. Dies wäre ein Hinweis auf eine Funktionalität des Gens in anderen Säugern. Ein Alignment der potentiellen ORFs zeigt aber, dass diese keinesfalls konserviert sind. Die Proteine, die entstünden, gehen aus Abb. 26 hervor und wären grundverschieden. Ein Überblick der Region (Abb. 23), zeigt auch, dass die Konservierung keinesfalls homogen ist,

sondern Schwankungen unterliegt. Etwas stärker konservierte Bereiche sind auch außerhalb von Exons nicht ungewöhnlich.

Trotz Konservierung und Zuordenbarkeit ist weder in der Ratte noch im Menschen ein Trankskript der zu Poldi homologen Region bekannt. Auch sind keinerlei Proteine bekannt, die dem potentiellen Produkt eines der beiden ORFs entsprächen (blastx der NCBI „Non-redundant protein sequence“ Datenbank).

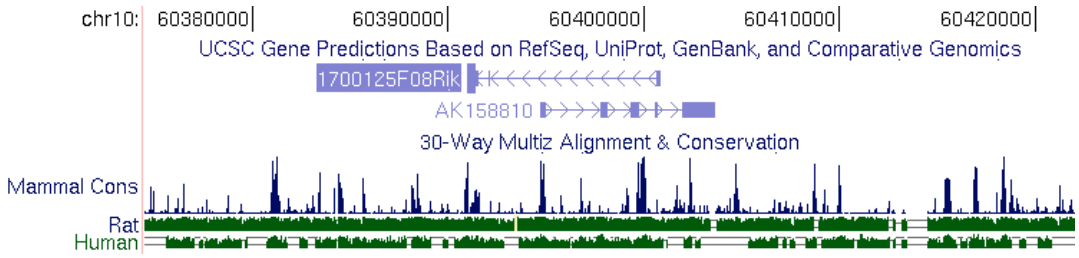


Abb. 23 Übersicht der Konservierung in der Poldi Region
Die Exons der beiden Transkripte weisen im Alignment mit dem Konsensus von 30 Säugetieren (Mammal Cons, dunkelblau) keine deutlicheren Maxima als die umliegende Region auf. Das Alignment mit Ratte und Mensch (Rat, Human, grün) weist keine größeren Lücken auf. Quelle: http://genome.ucsc.edu/

Das Fehlen eines Transkripts in Ratte und Mensch bei gleichzeitiger Syntänie weist entweder darauf hin, dass Poldi und seine mögliche Funktion in den verwandten Spezies verloren ging, oder dass es sich um ein evolutionär junges Gen handelt, das spezifisch für eine bestimmte taxonomische Einheit ist, ein so genanntes Orphan-Gen. Interessanterweise sind von Poldi keine Paraloge bekannt. Die als häufigster Mechanismus der Entstehung neuer Gene angesehene Genduplikation oder ein Exonshuffling, sowie eine Lokalisation der transkriptionellen Einheit durch Retrotransposition fallen daher als Erklärung für die Existenz des Poldi Gens aus. Stattdessen erscheint eine *de novo* Entstehung des Gens aus nichtkodierender Sequenz möglich.

Ein phylogenetischer Northernblot von Heinen (2008) weist auf ein erstes Auftreten des Poldi Transkripts vor etwa 2 Millionen Jahren hin. Die Phylogenie des Genus *Mus* ist in Abb. 24 beschrieben. Spezies, die das Transkript exprimieren sind mit einem roten Pfeil gekennzeichnet.

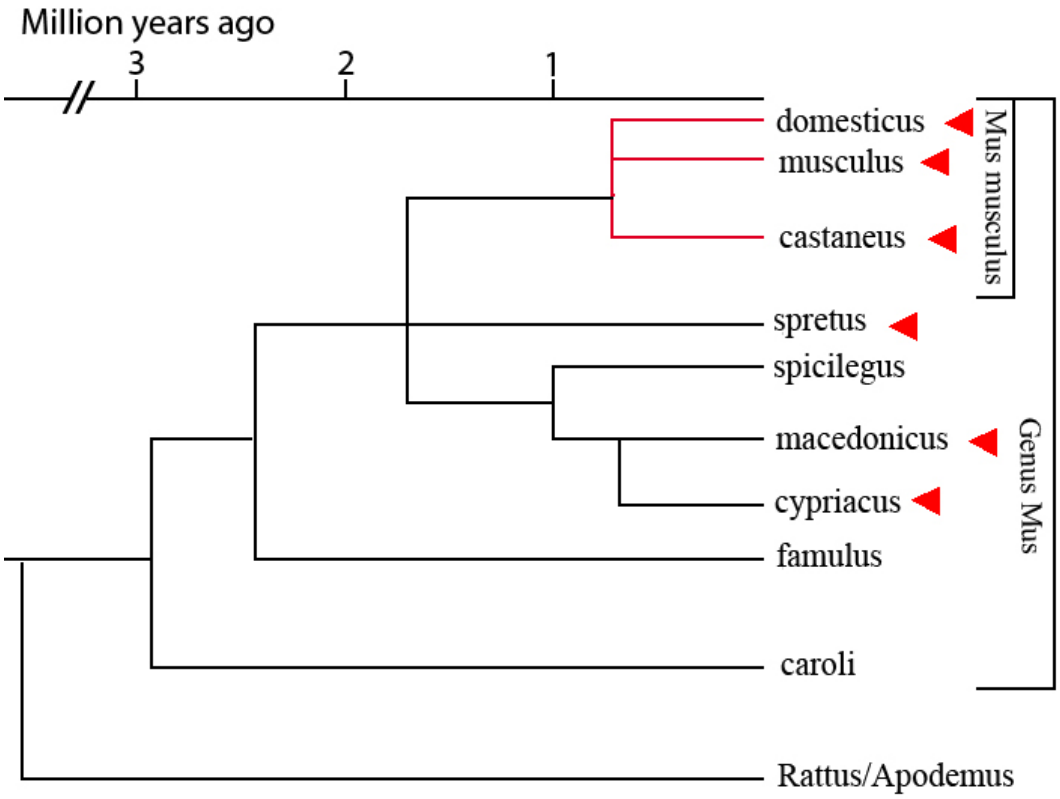


Abb. 24 Phylogenie des Genus *Mus* mit den Außengruppen *Rattus* und *Apodemus* modifiziert nach (Guenet und Bonhomme 2003) und (Cucchi et al. 2006)
Poldi transkribierende Spezies nach Heinen (2008) sind mit einem roten Pfeil markiert.

Parsimonie spricht für eine Entstehung von Poldi nach der Abspaltung von *M .famulus* aber vor der Abspaltung von *M. spretus*, *M. spicilegus*, *M. macedonicus* und *M. cypriacus* von der *M. musculus* Untergruppe. Eine Neuentstehung vor 2 Millionen Jahren und ein sekundärer Verlust des Poldi Transkripts in *M. spicilegus* wären als Ereignisse für das Auftreten des Transkripts ausreichend. Ein vierfacher unabhängiger Verlust hätte dem Blot zufolge in *M. famulus*, *M. caroli*, *Rattus* und *Apodemus* stattfinden müssen. Bezieht man die EST-Datenbank des NCBI (dbEST) mit ein, wären unabhängige Verluste in allen enthaltenen Spezies dieser Liste hinzuzufügen, da das Transkript in anderen Spezies unbekannt ist. Zumindest das Transkriptom des Menschen ist derart intensiv untersucht, dass hier von einem Fehlen des Transkripts ausgegangen werden muss. Demnach stehen mindestens fünf unabhängige Verlustereignisse gegen eine *de novo* Entstehung und einen sekundären Verlust.

Hat Poldi seine Funktion mehrfach unabhängig verloren oder wurde eine alte Funktion reaktiviert, würde man erwarten, dass die Spuren negativer Selektion

sichtbar wären, solange keine Mutationssättigung eingetreten ist. Funktionale Einheiten wie Exons und insbesondere ORFs sollten ein stärkeres Konservierungssignal zeigen. Hierfür lässt sich aber keine Evidenz finden. Es gibt zwar einige Konservierungsmaxima (Abb. 23), aber diese scheinen zufällig verteilt. Exons und ORFs sind nicht stärker konserviert als Intron- und extragenische Bereiche. In Ratte und Mensch entstünden weder vergleichbare Proteine (Abb. 26), noch ist die Transkriptstruktur intakt. Dies spricht zusammen mit dem Fehlen stärkerer Konservierung der Exons gegen eine Wiederauferstehung eines einst funktionalen Poldi in *Mus* oder den unabhängigen Verlust in anderen Taxa.

Mit dem Auftreten des Transkripts korreliert eine mehrere Basenpaare umfassende Mutation 5', nahe des Transkriptionsstarts im ersten Exon (Abb. 25A) und eine unabhängige Mutation in *M. spicilegus* am 3'-Ende des ersten Exons (Abb. 25B). Es gibt keine weiteren Unterschiede zwischen den analysierten Spezies, die einen potentiellen Einfluss auf die Transkriptstruktur haben könnten und mit dem Auftreten des Transkripts zusammenfallen (Datenträger: Sequenzanhang.doc, AK158810 Exon 1 Außengruppen). Die sieben Basenpaare umfassende Mutation 5' liegt nahe dem Transkriptstionsstart und könnte die Transkriptionsinitiation beeinflussen oder in posttranskriptioneller Regulation eine Rolle spielen. Die Transversion der letzten 3'-Base zerstört in *M. spicilegus* den Spleißkonsensus und könnte ausreichen, die Entstehung eines gespleißten und funktionalen Poldi Transkripts zu verhindern oder stark zu vermindern. Die von den 5' Unterschieden unabhängige Punktmutation in *M. spicilegus* unterstützt das oben entworfene Szenario größtmöglicher Parsimonie für die Präsenz von Poldi im Genus *Mus*. Die Funktionalität könnte durch die größere 5' Mutation erworben und in *M. spicilegus* durch die unabhängige Mutation der Spleißstelle wieder verloren gegangen sein. Dies bleibt dennoch Mutmaßung, solange funktionale Evidenz für die Wirksamkeit dieser Mutationen fehlt.

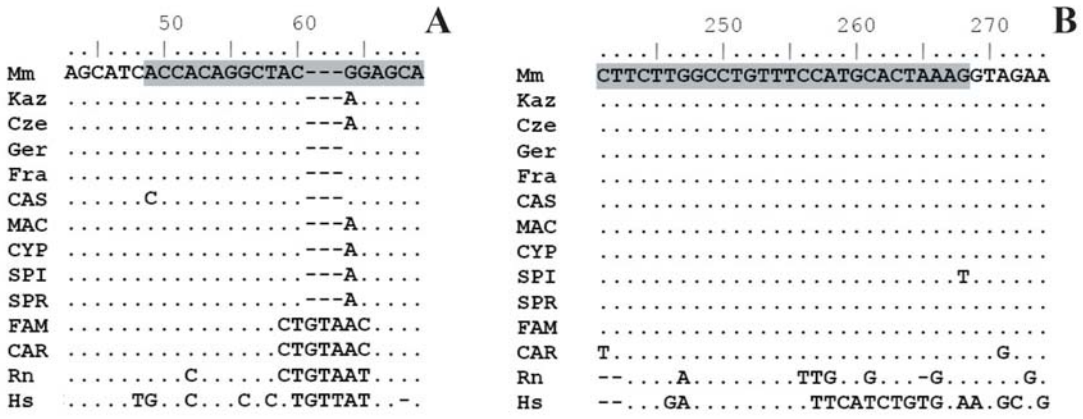


Abb. 25 Mit dem Auftreten des Poldi Transkripts korrelierende Polymorphismen
Die Exonsequenz ist grau markiert. Alle Sequenzen sind mit der annotierten *M. musculus* Datenbanksequenz aligniert (Mm).
A: 5' Ende des ersten Exons. Spezies, welche die kürzere Variante tragen transkribieren Poldi mit Ausnahme von *M. spicilegus*.
B: 3' Ende des ersten Exons. Mus spicilegus trägt an Position 268 eine G>T Transversion. Die kanonische Spleißerkennungssequenz geht dadurch verloren.

Überraschend ist die Tatsache, dass in *M. famulus* die beiden potentiellen ORFs des Poldi Gens vorhanden sind (Abb. 26), obwohl kein Transkript existiert. Die Entstehung der ORFs wäre demnach der Transkription vorangegangen. In *M. caroli* ist zumindest der erste potentielle ORF vorhanden, während der zweite, falls er ein Protein produzierte, ein durch eine Leserastermutation zum C-Terminus hin abgewandeltes Protein entstünde. In Ratte und Mensch können weder vergleichbare Proteine entstehen, noch ist die Transkriptstruktur intakt, hier sind weder ORF noch Transkript vorhanden.

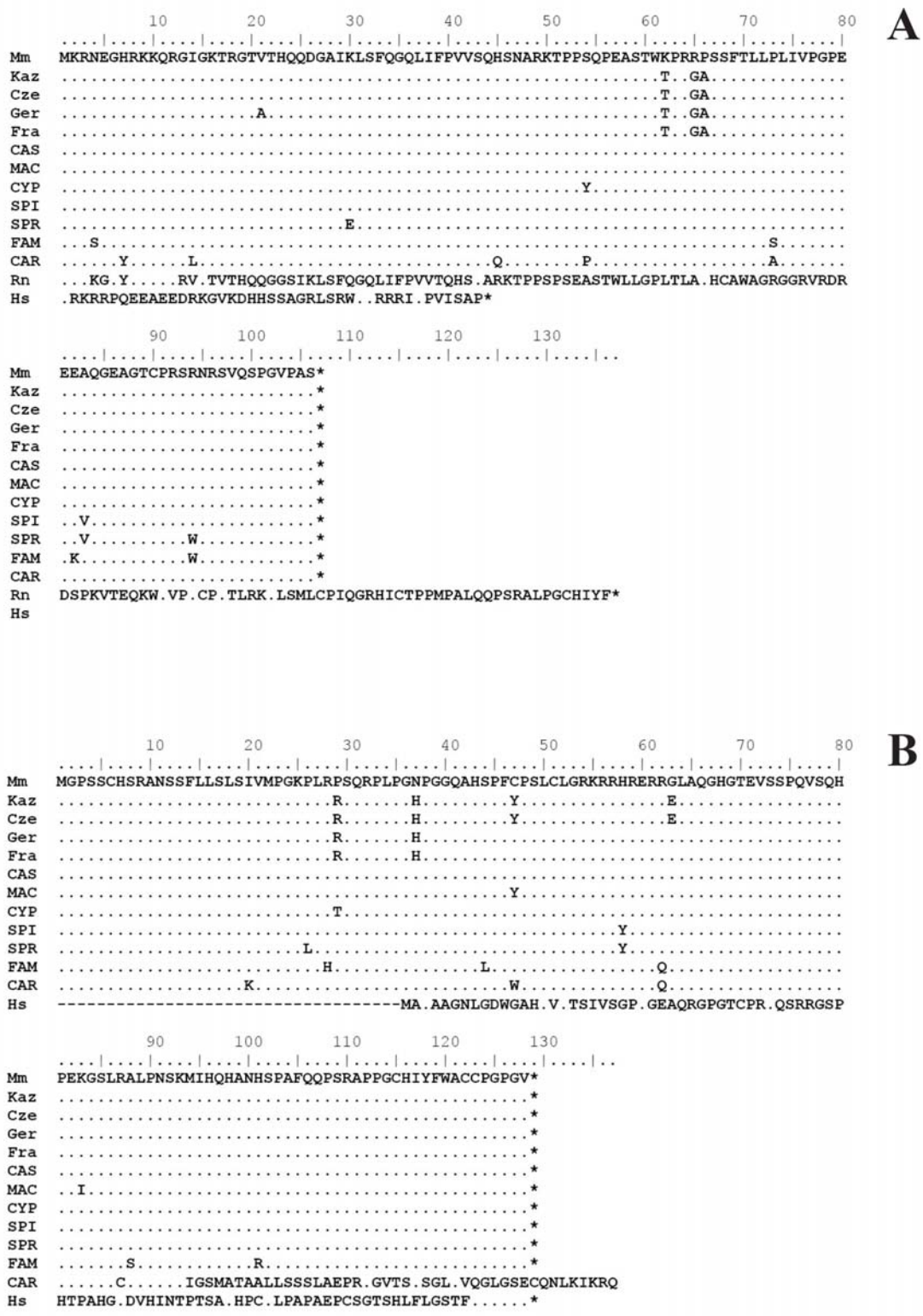


Abb. 26 Potentielle Poldi-Proteine in den Außengruppen
A: ORF1 B:ORF2

3.5 Zusammenfassung

Ein Tal reduzierter Variabilität in *M. m. musculus* und signifikante Verschiebungen des Allelfrequenzspektrums am Poldi Locus entsprechen der Signatur eines Selective Sweep. Im Zentrum des Variabilitätstals befinden sich zwei auf ESTs basierende Transkripte (Poldi/1700125f08Rik und AK158810). Der EST AK158810 ist wahrscheinlich ein Resultat unspezifischer Transkription und trägt keine Funktion. In einem Abschnitt von fast 300kb sind keine weiteren funktionalen Einheiten annotiert. Damit ist das Selektionsereignis offenbar auf Poldi zurückzuführen. Als mögliche Ursache des Selective Sweeps wurden Mutationen in den potentiellen ORFs von Poldi identifiziert, welche funktionale Konsequenzen haben könnten und in *M. m. musculus* fixiert sind. Ein cisregulatorischer Auslöser des Selektionsereignisses kann nicht ausgeschlossen werden.

Die Abwesenheit von Poldi Homologen und Paralogen bei gleichzeitiger Syntänie der Region zu Ratte und Mensch sprechen für eine *de novo* Entstehung des Poldi Gens aus nichtkodierender DNA. Da die Exons des Gens innerhalb der Säugetiere kein von der Umgebung abweichendes Konservierungsmuster aufweisen, ist eine Wiedererweckung eines ehemals funktionalen Gens oder der unabhängige Verlust äußerst unwahrscheinlich. Heinen (2008) konnte das erste Auftreten des Transkripts auf einen Zeitpunkt vor etwa 2 Millionen Jahre datieren. In der vorliegenden Arbeit konnte gezeigt werden, dass zwei potentiell funktionale Mutationen im ersten Exon mit dem Auftreten des Transkripts korrelieren. Eine größere, sieben Basenpaare umfassende Mutation am 5’ Ende des ersten Exons korreliert mit dem Auftauchen des Transkripts im Genus *Mus*, während eine Basensubstitution am 5’ Ende des ersten Exons mit einem sekundären Verlust in *M. spicilegus* korreliert. Interessanterweise könnte die Entstehung der ORFs der Transkription vorangegangen sein. Dabei bleibt fraglich, ob tatsächlich ein Protein kodiert wird.

3.6 Diskussion

Die Neuentstehung von Genen aus nichtkodierender DNA wurde lange Zeit als nahezu unmöglich angesehen (Ohno 1970; Jacob 1977). Es galt das Dogma der Entstehung neuer Gene durch Genduplikation. Noch 2003 spricht Long (Long et al. 2003) der *de novo* Genese von Genen aus nichtkodierender DNA nur eine untergeordnete Rolle zu. Allenfalls wird eine Rekrutierung zusätzlicher Sequenz zu bereits bestehenden Genen diskutiert. Stattdessen gelten Genduplikation, Exon Shuffling, Genfusion, Genfission, horizontaler Gentransfer und Retrotransposition als die zentralen Mechanismen der Entstehung neuer Gene. Diesen Mechanismen ist gemein, dass sie auf existierende Gene zurückgreifen. Jedes neu entstehende Gen wäre damit ausschließlich eine Kombination bereits vorhandener Muster.

In der jüngsten Vergangenheit konnten jedoch einige *de novo* entstandene Gene in *Drosophila* (Levine et al. 2006; Begun et al. 2007; Chen et al. 2007) und in Hefe (Cai et al. 2008) identifiziert werden. Zhou und Koautoren (Zhou et al. 2008) gehen anhand ihres Vergleichs von cDNA- und Genomdatenbanken verschiedener *Drosophila* Spezies von einem Anteil *de novo* entstandener Gene unter allen neuen Genen von fast 12% aus. Dies übertrifft sogar den Anteil der durch Retrotransposition entstandenen Gene (10%) in ihrer Studie. Damit wäre die *de novo* Genese von weitaus größerer Bedeutung, als bislang angenommen.

Zwei Dinge hat die Mehrzahl der neu entstandenen Gene gemein: Erstens weisen überraschend viele testisspezifische Regulation auf (Bai, Chan und Xu 2003; Begun et al. 2007; Chen et al. 2007; Metta und Schlotterer 2008; Zhou et al. 2008). Zweitens stehen junge Gene unter dem Einfluss positiver Selektion (Begun 1997; Nurminsky et al. 1998; Johnson et al. 2001; Enard et al. 2002; Maston und Ruvolo 2002; Wang et al. 2002).

Die zweite Gemeinsamkeit ist leicht zu verstehen. Das initiale Auftreten vorteilhafter Genfunktion zieht wahrscheinlich einen Selective Sweep nach sich. Das neu entstandene Gen ist noch weit von seiner optimalen Funktion entfernt und bietet viele Möglichkeiten der Vervollkommnung und Feinjustierung. Hieraus ergibt sich eine Entstehungsgeschichte unter dem Regime positiver Selektion, die dann in negative Selektion zur Aufrechterhaltung der neugewonnenen Funktion münden kann (Domazet-Loso und Tautz 2003; Jones, Custer und Begun 2005).

Für das gehäufte Auftreten neuer Gene mit testisspezifischer Expression gibt es mehrere mögliche Faktoren, die aber zu diesem Zeitpunkt spekulativ bleiben. So könnte die einfache Konstruktion testisspezifischer Promotoren eine initiale Transkription in den Testes begünstigen. Dies könnte auch für Poldi eine besondere Rolle spielen, da keine klassischen Promotorelemente 5' des Poldi Gens identifiziert werden konnten. Gleiches gilt für die Chromatinstruktur nach der Meiose (Caron et al. 2005). Reduzierte Pleiotropie und beschleunigte Evolution testisspezifischer Gene könnten ebenfalls eine Rolle spielen (Zhang et al. 2007). Interessant ist, dass mit dem gehäuften Auftreten neuer Gene im Testis auch Gene gehäuft auftreten, die eine potentielle Rolle in reproduktiver Isolation spielen könnten und die Speziation vorantreiben.

Poldi ist ebenfalls testisspezifisch exprimiert. Die Region weist ein Tal reduzierter Variabilität auf, wie es Genetic Hitchhiking unter positiver Selektion erwarten lässt (Maynard Smith und Haigh 1974). Zusätzliche Verschiebungen des Frequenzspektrums hin zu seltenen Mutationen entsprechen der Signatur natürlicher Selektion. Damit ist Poldi ein typisches junges Gen.

Poldi ist meines Wissens zu diesem Zeitpunkt das erste *de novo* entstandene Gen, welches in Säugern beschrieben wurde. Mit einem Alter von etwa zwei Millionen Jahren zählt es zu den jüngsten bekannten Genen (Long et al. 2003). Das Auftreten neuer Gene während einer Divergenz von zwei Millionen Jahren verdeutlicht, dass neue Gene auch in der Abspaltung der Hominiden von den Pongiden, deren Divergenz mehr als doppelt so lang ist (Hobolth et al. 2007), eine größere Rolle als bisher angenommen spielen könnten. Die Detektion von 72 neuen Genen in *D. melanogaster* seit der Abspaltung von *D. simulans* und *D. sechellia* vor 5,4 Millionen Jahren (Zhou et al. 2008) stützt diese These. Die Entdeckung, dass bis zu 93% des menschlichen Genoms transkribiert sind (Birney et al. 2007), lässt vermuten, dass sich noch viele unentdeckte neue Gene im menschlichen Genom und damit auch in den Genomen anderer Säuger befinden. Diese werden bis jetzt nicht als funktional erkannt, da die Zuordnung von Funktion über konservierte Muster stattfindet. *De novo* entstandene Gene müssen nicht konservierten Genmustern entsprechen und sind definitionsgemäß Orphan Gene. Dadurch lassen sie sich auch nicht durch Homologie aufspüren. Weil neue Gene oft kurze ORFs kodieren oder auch nichtkodierende RNAs sein können, wird die automatisierte Suche noch erschwert (Martignetti und Brosius 1993b; Martignetti und Brosius 1993a; Levine et al. 2006; Metta und

Schlotterer 2008). Ein Kriterium zu finden, nach dem sich ein *de novo* entstandenes Gen von der es umgebenden DNA unterscheidet, ist schwer. Am ehesten ließe sich ein Nachweis der Funktion des Gens für den Organismus hier anführen.

Es gibt mehrere unabhängige Hinweise auf eine Funktion von Poldi. So konnte in dieser Arbeit gezeigt werden, dass Poldi unter natürlicher Selektion steht. Außerdem konnte Heinen (2008) zeigen, dass ein Knock Out des Poldi Gens in der Maus Einfluss auf die Spermienmotilität hat. Poldi könnte auch in ein regulatorisches Netzwerk eingebunden sein. So konnte Heinen neben einigen Genen, die einen kleinen aber signifikanten Expressionsunterschied zeigen, ein Gen identifizieren, welches in Poldi Knock Out Mäusen 15fach hochreguliert ist. Poldi ist anscheinend vorwiegend in postmeiotischen Spermatiden exprimiert und könnte daher hier seine Funktion ausüben.

Da die Strategie mittels homologer Rekombination eine Poldi Knock Out Maus zu erzeugen so gewählt wurde, dass Poldi und AK158810 ausgeschaltet wurden, ist ein direkter Rückschluss auf die Funktion von Poldi, ausgehend von diesen Daten, nicht zulässig. Spermienphänotyp und Selektion könnten auch AK158810 zugeschrieben werden. Jedoch wurde in der vorliegenden Arbeit gezeigt, dass AK158810 viele Strukturen, die für ein funktionales Gen typisch sind, entbehrt. Weder Poldi noch AK158810 besitzen klassische Promotorelemente (BRE, TATA, Initiator, DPE) aber im Gegensatz zu Poldi ist AK158810 äußerst niedrig exprimiert. Während ein Genprodukt von AK158810 im Northernblot kaum detektierbar ist, wird mit einer Poldi spezifischen Sonde eine klare Bande detektiert (Heinen 2008). Mittels qRT-PCR wurde ein Poldi Expressionsniveau im Hoden aller untersuchten Mäuse festgestellt, welches mit dem anderer aktiver Gene vergleichbar ist. Hingegen zeigt die Gelelektrophorese der cDNA-PCR, dass das AK158810 Genprodukt in den Hoden der Mäuse beider Populationen nur sporadisch zu finden ist, d.h. nur in sieben von zwölf untersuchten Mäusen, ohne in einer der Populationen deutlich häufiger aufzutreten. Die Funktion eines derart niedrig exprimierten Gens mit sporadischer Expression ist äußerst fraglich. Daher ist der Spermienphänotyp der Knock Out Mäuse und damit die Genfunktion wahrscheinlich auf Poldi zurückzuführen.

Da es einen überlappenden Bereich der gegenläufigen Transkripte von Poldi und AK158810 gibt, könnte man vermuten, dass eine Wechselwirkung auf RNA-Ebene stattfände. Vergleicht man das Poldi Expressionsniveau in den Hoden zwischen den Tieren, für die AK158810 Genprodukt in der cDNA-PCR nachweisbar ist und jenen

für die dies nicht zutrifft, stellt man keinen Unterschied fest (p = 0,94, Wilcoxon W-Test). Zwischen dem Auftreten des Schmiers im AK158810 Northern Blot (*M. spicilegus*, *M. macedonicus* und *M. cypriacus*, Abb. 22) und dem Auftreten des Poldi Transkripts (*M. musculus*, *M. spretus*, *M. macedonicus* und *M. cypriacus*, Abb. 24) ist auch kein Zusammenhang erkennbar. Für eine Wechselwirkung auf RNA-Ebene gibt es daher keinen Anhaltspunkt.

Eine Funktion von AK158810 im Hirn ist durch sporadische und niedrige Expression nicht völlig ausgeschlossen, denn im Hirn gibt es in allen Mäusen, die im cDNA-PCR-Experiment verwendet wurden, AK158810 Genprodukt. Gäbe es eine Funktion im Hirn, könnte auch der Selective Sweep von einer im Hirn wirksamen, vorteilhaften Variante ausgelöst worden sein. Gegen eine Funktion spricht aber die auch im Hirn, wie auch im Hoden auftretende, anscheinend zufällige Transkriptstruktur von AK158810. In der Gelelektrophorese der cDNA-PCR konnten bereits im 5' Bereich des Transkripts etwa zehn unterschiedliche Varianten sowohl im Hoden, als auch im Hirn identifiziert werden. Die Sequenzierung dieser PCR-Produkte ergab, dass scheinbar beliebige Bereiche in das Transkript eingegliedert oder annotierte Exons verkürzt werden. Dies geschieht unabhängig von bekannten spleißrelevanten Stellen.

Poldi könnte auch als nichtkodierende RNA wirksam sein. Ist dies der Fall, sind Regulationsmechanismen, die auf Hybridisierung und damit auf Sequenzhomologie beruhen unwahrscheinlich (RNAi, posttranskriptionelles Gene Silencing), da keine ähnlichen RNAs gefunden werden konnten.

Mit dem Auftreten des Poldi Transkripts in der Mausphylogenie korreliert eine 7 bp umfassende Mutation am 5' Ende des ersten Exons. Eine spleißrelevante unabhängige Mutation am 3' Ende des ersten Exons könnte mit einem sekundären Verlust der Poldi Transkription in *M. spicilegus* zusammenhängen. Hierbei wird davon ausgegangen, dass das im Northernblot verwendete Individuum den transkriptionellen Status der Spezies repräsentiert. In der Tat könnte die Expression des Poldi Gens in den untersuchten Spezies polymorph und das Auftreten von Poldi im Genus auf diese Art nicht zu datieren sein. Fest steht, dass alle untersuchten zu *M. musculus* gehörenden Individuen das Transkript tragen und eine Transkription in Mensch und Ratte aufgrund fehlender Datenbankeinträge unwahrscheinlich ist. Demnach stünde nur die Datierung des Auftretens zwischen *Rattus* und *M. musculus* in Frage, nicht jedoch die Hinweise auf *de novo* Genese und Selektion.

Vollzieht man die Entstehung des Poldi Gens anhand der Daten schrittweise nach, stellt man fest, dass in allen untersuchten Spezies ein Polyadenylierungssignal am 3' Ende des dritten Exons vorhanden ist (5' AAUAAA 3', Datenträger: Sequenzanhang.doc, Poldi Außengruppen Exon3, Position 675 - 681). Im Menschen konnten keine weiteren strukturellen Merkmale von Poldi detektiert werden. Spleißstellen homolog zu denen der Maus können nicht identifiziert werden. Die dem Poldi Transkript homologen Sequenzen enthalten lange Homopolymere (Datenträger: Sequenzanhang.doc, Poldi Exon1 Außengruppen Position 166-185). Die Ratte unterscheidet sich am 5' Ende des ersten Exons in 7 bp von den Mäusen, die Poldi transkribieren. Es handelt sich dabei um den gleichen Sequenzunterschied, in dem sich auch die Mäuse, die Poldi nicht transkribieren von denen, die ein Transkript bilden, unterscheiden. Außerdem trägt die Ratte eine Basensubstitution von T nach C der zweiten 5' Base des zweiten Introns (Datenträger: Sequenzanhang.doc, Außengruppensequenzen Exon2 Position 102). Diese Position gilt als relevant für Spleißvorgänge. Die Variante in der Ratte könnte die Entstehung des Transkripts verhindern. Die beiden ORFs sind jedoch weder in Ratte noch in Mensch konserviert. Mehrere Leserastermutationen würden zu völlig anderen Proteinen führen.

In *M. caroli* ist der vordere ORF bereits im Leseraster des *M. musculus* ORF. In *M. famulus* sind beide potentiellen ORFs im Leseraster vorhanden. Zwischen *M. famulus*, die das Transkript nicht exprimiert und den Mäusen, die es exprimieren, bleibt dann der 7 bp 5' Unterschied als potentiell relevant. Dieser Unterschied wäre ein starker Kandidat für funktionale Experimente, wie z.B. Promotorassays in Zellkultur. Die Generierung zweier sich nur in dieser Mutation unterscheidenden Mäuse wäre eine Alternative.

Eine regulatorische Mutation, die in dieser Studie nicht erfasst wurde, könnte ebenfalls durchaus für An- und Abwesenheit des Transkripts in den hier untersuchten Spezies verantwortlich sein. Die Rückführung auf die 5' Mutation bleibt ohne funktionale Analyse spekulativ. Dennoch bleibt festzuhalten, dass die ORFs im Gegensatz zu der Studie von Cai und Koautoren (Cai et al. 2008) bereits vor dem Auftreten des Transkripts in der Stammesgeschichte des Genus *Mus* intakt vorhanden sind. Da ohne Transkription ein potentielles Protein gar nicht wirksam werden könnte, würde dies bedeuten, dass neutrale Mutationen zur Entstehung des ORFs stattgefunden haben müssen, die dann erst mit der Transkription des Gens ihre Wirkung entfalten konnten. Dass neutrale Mutationen neuen funktionalen Mutationen

vorangehen, beschreiben auch Ortlund und Koautoren (Ortlund et al. 2007) am Beispiel der Bindungsaffinität des Glukokortikoidrezeptors und ist nicht ungewöhnlich.
Ob und welcher der ORFs tatsächlich aktiv ist, bleibt zu diesem Zeitpunkt leider offen, da bislang kein Protein identifiziert werden konnte.
Damit ist es auch schwierig den Auslöser des Selective Sweeps klar zu definieren. Da ein Poldi-Expressionsunterschied zwischen den Subspezies nicht ausgeschlossen werden kann, kommt eine regulatorische vorteilhafte Variante in Frage. Des Weiteren könnten sich noch unidentifizierte funktionale Einheiten neben Poldi in der Region befinden, z.B. eine in trans wirkende regulatorische Einheit. Wird AK158810 aufgrund seines niedrigen Expressionsniveaus und der unklaren Transkriptstruktur als Ursache ausgeschlossen, bleiben neben den oben genannten Möglichkeiten noch Unterschiede in beiden ORFs. Weil der zweite ORF in der Sweep-Population nicht funktional fixiert ist, sondern eine das Stoppcodon zerstörende Leserastermutation 3' in der kasachischen und der tschechischen Population segregieren, kommt für den Auslöser am ehesten ein vorteilhaftes Valin an Position 21 des ersten ORFs (Abb. 16) in Frage. Da Valin hier dem anzestralen Allel entspricht und es in *M. domesticus* segregiert, wäre von Selektion aus der Standing Variation auszugehen. Das verdrängte Alanin ist ebenfalls unpolar, aliphatisch und ähnlich groß. Der Unterschied wäre eher in den Bereich der Feinjustierung einzuordnen. Ist Poldi als strukturelle RNA aktiv, kommen zwei weitere Mutationen als Ursache des Selektionsereignisses in Frage. Beide Mutationen weisen in *M. m. musculus* das abgeleitete Allel auf. Daher entspräche Selektion auf eine dieser Varianten einem Selektionsmodell auf Neumutation. Beide Mutationen haben Einfluß auf die *in silico* vorhergesagte Sekundärstruktur von Poldi. Interessanterweise wäre die in *M. m. musculus* (Sweep-Population) segregierende 1bp Deletion am 3' Ende von Exon3 für die Sekundärstruktur irrelevant und folglich neutral, wenn Poldi als strukturelle RNA wirken würde. Dies könnte ein Hinweis darauf sein, dass die Funktion von Poldi tatsächlich im Bereich der strukturellen RNAs liegt.

Poldi repräsentiert das erste *de novo* entstandene Gen in Säugern. Es ist mit zwei Millionen Jahren ein sehr junges Gen und steht wie viele andere junge Gene unter dem Einfluss positiver Selektion. Die Poldi Region scheint generell transkriptionell aktiv. So gibt es eine Vielzahl unspezifischer Transkripte am Locus, obwohl keine

klassischen Promotorsequenzen vorhanden sind. Eines dieser Transkripte (Poldi) scheint eine Funktion erworben und unter den Einfluss positiver Selektion geraten zu sein. So kann aus transkriptionellem Hintergrundrauschen eine völlig neue Funktion entstehen.

Die *de novo* Entstehung von Genen ist aufgrund der überschaubaren Anzahl bekannter Fälle ein wenig untersuchter Mechanismus. Sie könnte jedoch für die Evolution der Organismen von bemerkenswerter Bedeutung sein, da sie im Gegensatz zu Genduplikation nicht auf bekannte Muster zurückgreift, sondern gänzlich neue Eigenschaften entstehen lassen kann. Selbst bei lange divergierenden Genduplikaten, die als neofunktional charakterisiert sind, sind die Genfuktionen noch ähnlich (Rodriguez-Trelles, Tarrio und Ayala 2003; Zahn et al. 2005). Zwar tritt zu Beginn einer Neofunktionalisation nach einer Genduplikation häufig positive Selektion auf, die eine schnelle Divergenz antreiben könnte, doch bald danach ist mit negativer Selektion zu rechnen (Domazet-Loso und Tautz 2003; Jones, Custer und Begun 2005). Die Funktion des vorteilhaften Duplikats wird aufrechterhalten, eine weitere Divergenz wird gebremst. *De novo* Genese ist unabhängig von Duplikation und ermöglicht dadurch dem evolvierenden Organismus neue Gene und Funktionen zu gewinnen. Unter den Gründergenen von Genfamilien (Domazet-Loso, Brajkovic und Tautz 2007) könnte ein erheblicher Anteil *de novo* entstanden sein.

4 Anhang

4.1 Material und Methoden

4.1.1 Mausproben

M. m. domesticus Proben wurden im Zentralmassiv in Frankreich und in der Köln Bonner Bucht gesammelt. Individuen von *M. m. musculus* stammen aus Tschechien und Kasachstan, wie in (Ihle et al. 2006) und (Voolstra et al. 2007) beschrieben. Um sicher zu stellen, dass die gefangenen Mäuse aus verschiedenen Sippen stammen, wurden die einzelnen Fangplätze mindestens 1 km voneinander entfernt gewählt. Tiere für die qRT-PCR wurden vor der RNA Extraktion in Lebendfallen gefangen und für 3-5 Tage unter kontrollierten Bedingungen im Labor gehalten. Sechs Männchen jeder Subspezies (tschechische und deutsche Population) wurden für die Genexpressionsanalyse ausgewählt. Diese Männchen waren ähnlichen Körpergewichts, ähnlicher Größe und demnach auch vergleichbaren Alters. Elf zusätzliche Individuen dieser Populationen wurden für die Erhebung der Sequenzpolymorphismusdaten sequenziert. Für die Charakterisierung von Poldi und der umliegenden Region wurden noch je elf zusätzliche Individuen der kasachischen und französischen Population sequenziert. Die verwendete *M. m. castaneus* Probe gehört zum CIM-Stamm (Rottscheidt und Harr 2007) und wurde von Ruth Rottscheidt und Bettina Harr zur Verfügung gestellt. Die *M. spretus* Probe wurde von Christian Voolstra zur Verfügung gestellt. *M. famulus*, *M. macedonicus*, *M. cypriacus*, *M. spicilegus* und *M. caroli* entstammen der Sammlung von Francois Bonhomme am Institut des Sciences de l'Evolution de Montpellier. *Apodemus flavicollis* wurde in Plön in Schleswig-Holstein gefangen.

4.1.2 DNA- und RNA-Extraktion

DNA wurde aus ethanolgelagertem oder frischem Gewebe extrahiert. In 7ml HOM-Puffer (80mM EDTA, 100mM Tris und 1% SDS) und 40µl Proteinase K (0,2 mg/ml) wurden Gewebeproben von etwa 300 mg gelöst und nach Inkubation über Nacht wurde 1 g Salz zugefügt. Anschließend wurden die Proben 10 min auf Eis inkubiert. Auf einen Waschschritt mit 5ml Chloroform folgte die Zentrifugation für 1 h bei 4000 g. Die DNA wurde aus der oberen Phase mit Ethanol präzipitiert. Das Pellet wurde in TE-Puffer (10 mM Tris, 1 mM EDTA) gelöst. Für die RNA Extraktion wurden Gewebeproben des Gehirns, der Hoden und der Leber/Niere mechanisch in

TRIzol (Invitrogen, Carlsbad, CA) homogenisiert. Die RNA wurde den Vorgaben des Herstellers entsprechend extrahiert und sofort in der cDNA Synthese verwendet oder für die Aufbewahrung in DEPC-H_2O resuspendiert, in 4 M LiCl präzipitiert und auf -80°C eingefroren.

4.1.3 cDNA Synthese

Für die cDNA Synthese wurde Thermoskript RT (Invitrogen) mit zufälligen Hexameren (Fermentas K1612) als Primer eingesetzt (1-5µg RNA, 200ng Hexamer-Primer, 10mM dNTP mix, 5x Synthesepuffer, 0,1M DTT, RnaseOUT (40U/µl), ThermoScript (15U/µl, DEPC Wasser)). Vor dem Hinzufügen des Enzyms zur Reaktion wurden DTT, Primer, Puffer und dNTPs mit der RNA für 5 Minuten bei 65°C inkubiert. Nach 10 Minuten bei 25°C wurde die Reaktion bei 50°C für 50 Minuten durchgeführt. Die Reaktion wurde durch erhitzen auf 85°C für 5 Minuten beendet.

4.1.4 Quantitative real-time PCR (qRT-PCR)

Insgesamt 39 Kandidatengene für Expressionsänderungen zwischen den Subspezies in mindestens einem Gewebe wurde zufällig aus einem vorangegangenen Microarrayexperiment ausgewählt (Voolstra et al. 2007). Taqman® Gene Expression Assays (ABI) wurden für die qRT-PCR Analyse auf einem ABI 7900HT verwendet. Sonden- und Primerbindestellen aller Taqman® Assays wurden von cDNA amplifiziert, sequenziert und sorgfältig auf Polymorphismus geprüft. Assays, die an polymorphe Sequenz binden wurde von weiteren Analysen ausgenommen (siehe 2.2.1). Die qRT-PCR wurde entsprechend den Empfehlungen des Herstellers unter Verwendung einer 1:10 Verdünnung der cDNA durchgeführt. Drei technische Replikate jeder Reaktion wurden im gleichen Lauf und gleichen Bedingungen gemessen. *Hprt* (Hypoxanthin Phosphoribosyltransferase) wurde als endogene Kontrolle verwendet (siehe 2.2.2).

4.1.5 Berechnung der relativen Expressionsdivergenz (ED)

Für die Berechnung der Expressionsdivergenz (ED) wurden die Populationsmittel der ΔCts in Konzentrationen umgewandelt ($2^{-\Delta Ct}$), um eine lineare Größe zu erhalten. Anschließend wurde der Betrag der Differenz der mittleren Expressionsniveaus der beiden Populationen durch das Mittel beider Populationen geteilt:

$ED = \frac{|E(musculus) - E(domesticus)|}{\overline{E_{musdom}}}$. Die Divergenz wird also relativ zum Mittelwert der Expression des Gens in beiden Populationen gemessen.

4.1.6 PCR, Sequenzierung und Sequenzanalyse

Das Quiagen Multiplex Kit (Qiagen, Hilden) wurde entsprechend der Empfehlung des Herstellers verwendet, um die 1kb Stromaufwärtsregionen der im ersten Teil dieser Studie untersuchten Gene zu sequenzieren. Einer der beiden Primer wurde, wenn möglich, in der 5' UTR des analysierten Gens platziert (siehe Anhang 4.5 und **S 12**). Auch für die Amplifikation aller zu Poldi und AK158810 gehörender Exons, der cDNA und der flankierenden Regionen wurde das Quiagen Kit verwendet. Die resultierenden PCR-Produkte wurden auf einem ABI 3730 DNA-Analyzer (Applied Biosystems) in beide Richtungen sequenziert. ABI Big Dye terminator mix wurde entsprechend dem Herstellerprotokoll benutzt. Die so erhaltenen Sequenzen wurden mit dem Codoncodealigner v2.02 (CodonCode Corporation) analysiert. Watterson`s θ, Tajima's D und π wurden mittels DnaSP v4.0 (Rozas et al. 2003) berechnet. Die Haplotypen wurden durch PHASE (Stephens, Smith und Donnelly 2001) ermittelt, unter den in der DnaSP v4.50 Implementation vorgegebenen Standardeinstellungen.

4.1.7 Lineare Modellierung

Modellierung und Regressionsanalyse wurde unter Verwendung des R Statistiksoftwarepakets v2.6.2 durchgeführt. Für die Modellierung wurden Watterson's θ (Watterson 1975) logarithmiert, um es der Normalverteilung anzunähern (S 7). Zuvor wurden θ von Null (drei von 48 Fällen) durch das kleinste θ im Datensatz (0,00026) ersetzt. AICc (Akaike's Information Criterion corrected) und BIC (Bayesian Information Criterion) wurden wie in (Sugiura 1978) und (Schwarz 1978) beschrieben berechnet und zur Auswahl des besten Modells bei den gegebenen Daten herangezogen. AICc korrigiert AIC (Akaike 1974) für kleine Stichproben im Vergleich zur Zahl der Parameter und konvergiert für große Stichproben gegen AIC.

4.1.8 Klonierung und *in vitro* Transkription

Um zu prüfen, ob der exponentiell wachsende Fehler der Ct rein technischen Ursprungs ist, wurde ein Bereich auf cDNA mittels PCR amplifiziert, welcher die Binderegion des Taqman® Assays für das Gen *ppt1* enthält. Das PCR-Produkt wurde anschließend in den pGEM®-T Easy vector (Promega, Madison, WI) entsprechend

den Empfehlungen des Herstellers mittels Ligation eingefügt und in TOP10® (Invitrogen, Carlsbad, CA) vermehrt. Nach einer Colony-PCR wurde zur Kontrolle, ob das PCR-Produkt in der richtigen Orientierung inseriert ist, das Insert sequenziert. Das Plasmid wurde aus den Bakterien präpariert (Qiaprep Spin Miniprep Kit, Qiagen, Hilden) linearisiert und in der *in vitro* Transkription eingesetzt. 1µg Plasmid DNA, 2µl Transkriptionspuffer, 2µl (40U) T7 Polymerase (Roche, Basel) wurde für die *in vitro* Transkription verwendet. Die Reaktion wurde auf 20µl aufgefüllt und 2h bei 37°C inkubiert. Um die Plasmid DNA zu entfernen, wurde anschließend 1µl (2U) Turbo® DNAse (Ambion, Austin, TX) zugefügt und 15min bei 37°C inkubiert. Die erfolgreiche *in vitro* Transkription wurde im BioAnalyzer 2100 (Agilent Technologies, Santa Clara, CA, S 4) verifiziert und dann wie die natürlichen Proben (siehe 4.1.3 und 4.1.4), jedoch in einer Verdünnungsreihe, prozessiert.

4.2 Referenzen

Akaike, H. 1974. A new look at the statistical model identification. IEEE Trans. Automat. Control **19**:716–723.

Bai, X., E. D. Chan und X. Xu. 2003. The protein of a new gene, Tctex4, interacts with protein kinase CK2beta subunit and is highly expressed in mouse testis. Biochem Biophys Res Commun **307**:86-91.

Bedford, T.und D. L. Hartl. 2009. Optimization of gene expression by natural selection. Proc Natl Acad Sci U S A **106**:1133-1138.

Begun, D. J. 1997. Origin and evolution of a new gene descended from alcohol dehydrogenase in Drosophila. Genetics **145**:375-382.

Begun, D. J., H. A. Lindfors, A. D. Kern und C. D. Jones. 2007. Evidence for de novo evolution of testis-expressed genes in the Drosophila yakuba/Drosophila erecta clade. Genetics **176**:1131-1137.

Berry, R. 1991. House mouse *Mus domesticus*. Pp. 239– 247 *in* G. C. S. Harris, Hrsg. The handbook of British mammals.

Birney, E.J. A., P. J. de Jong und das ENCODE pilot project 2007. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature **447**:799-816.

Blekhman, R., A. Oshlack, A. E. Chabot, G. K. Smyth und Y. Gilad. 2008. Gene Regulation in Primates Evolves under Tissue-Specific Selection Pressures. PLoS Genet **4**:e1000271.

Boursot, P., J. C. Auffray, J. Britton-Davidian und F. Bonhomme. 1993. The Evolution of house mice. Annual Review of Ecology and Systematics **24**:119-152.

Boursot, P., Din, W., Anand, R., Darviche, D., Dod, B., von Deimling, F., Talwar, G. P. und F. Bonhomme. 1996. Origin and radiation of the house mouse: mitochondrial DNA phylogeny. J. Evol. Biol. **9**:391-415.

Britten, R. J. und E. H. Davidson. 1969. Gene regulation for higher cells: a theory. Science **165**:349-357.

Britten, R. J. undE. H. Davidson. 1971. Repetitive and non-repetitive DNA sequences and a speculation on the origins of evolutionary novelty. Q Rev Biol **46**:111-138.

Brown, R. P. und M. E. Feder. 2005. Reverse transcriptional profiling: non-correspondence of transcript level variation and proximal promoter polymorphism. BMC Genomics **6**:110.

Bustin, S. A. und T. Nolan. 2004. Pitfalls of quantitative real-time reverse-transcription polymerase chain reaction. J Biomol Tech **15**:155-166.

Cai, J., R. Zhao, H. Jiang und W. Wang. 2008. De novo origination of a new protein-coding gene in Saccharomyces cerevisiae. Genetics **179**:487-496.

Caron, C., J. Govin, S. Rousseaux und S. Khochbin. 2005. How to pack the genome for a safe trip. Prog Mol Subcell Biol **38**:65-89.

Carroll, S. B. 2008. Evo-devo and an expanding evolutionary synthesis: a genetic theory of morphological evolution. Cell **134**:25-36.

Chaix, R., M. Somel, D. P. Kreil, P. Khaitovich und G. Lunter. 2008. Evolution of Primate Gene Expression: Drift and Corrective Sweeps? Genetics **180**:1379-1389.

Chen, S.-T., H.-C. Cheng, D. A. Barbash und H.-P. Yang. 2007. Evolution of hydra, a Recently Evolved Testis-Expressed Gene with Nine Alternative First Exons in Drosophila melanogaster. PLoS Genetics **3**:e107.

Chevret, P., P. Jenkins und F. Catzeflis. 2003. Evolutionary systematics of the Indian mouse *Mus famulus* Bonhote, 1898: molecular (DNA/DNA hybridization and 12S rRNA sequences) and morphological evidence. Zool. J. Linn. Soc. **137**:385-401.

Chevret, P., F. Veyrunes und J. Britton-Davidian. 2005. Molecular phylogeny of the genus *Mus* (Rodentia: Murinae) based on mitochondrial and nuclear data. Biol. J. Linn. Soc. **84**:417-427.

Clark, R. M., T. N. Wagler, P. Quijada und J. Doebley. 2006. A distant upstream enhancer at the maize domestication gene tb1 has pleiotropic effects on plant and inflorescent architecture. Nat Genet **38**:594-597.

Corbet, G. B. und J. E. Hill. 1992. The mammals of the Indomalayan Region : a systematic review. Oxford University Press, Oxford.

Crowe, M. L., X. Q. Wang und J. A. Rothnagel. 2006. Evidence for conservation and selection of upstream open reading frames suggests probable encoding of bioactive peptides. BMC Genomics **7**:16.

Cucchi, T., A. Orth, J. C. Auffray, S. Renaud, L. Fabre, J. Catalan, E. Hadjisterkotis, F. Bonhomme und J. D. Vigne. 2006. A new endemic species of the subgenus *Mus* (Rodentia, Mammalia) on the Island of Cyprus. Zootaxa **1241**:1-36.

Cucchi, T., J. Vigne und J. Auffray. 2005. First occurrence of the house mouse *Mus musculus domesticus* Schwarz & Schwarz, 1943) in the Western Mediterranean: a zooarchaeological revision of subfossil occurrences. Biol. J. Linn. Soc. **84:**429-445.

Dheda, K., J. F. Huggett, J. S. Chang, L. U. Kim, S. A. Bustin, M. A. Johnson, G. A. W. Rook und A. Zumla. 2005. The implications of using an inappropriate reference gene for real-time reverse transcription PCR data normalization. Analytical Biochemistry **344**:141-143.

Doebley, J., A. Stec und L. Hubbard. 1997. The evolution of apical dominance in maize. Nature **386**:485-488.

Domazet-Loso, T., J. Brajkovic und D. Tautz. 2007. A phylostratigraphy approach to uncover the genomic history of major adaptations in metazoan lineages. Trends Genet **23**:533-539.
Domazet-Loso, T. und D. Tautz. 2003. An evolutionary analysis of orphan genes in Drosophila. Genome Res **13**:2213-2219.
Enard, W., M. Przeworski, S. E. Fisher, C. S. Lai, V. Wiebe, T. Kitano, A. P. Monaco und S. Paabo. 2002. Molecular evolution of FOXP2, a gene involved in speech and language. Nature **418**:869-872.
Fay, J. C. und C. I. Wu. 2000. Hitchhiking under positive Darwinian selection. Genetics **155**:1405-1413.
Fitzgerald, B. M., B. J. Karl und H. Moller. 1981. Spatial Organization and Ecology of a Sparse Population of House Mice (Mus musculus) in a New Zealand Forest. The Journal of Animal Ecology **50**:489-518.
Force, A., M. Lynch, F. B. Pickett, A. Amores, Y. L. Yan und J. Postlethwait. 1999. Preservation of duplicate genes by complementary, degenerative mutations. Genetics **151**:1531-1545.
Frazer, K. A., E. Eskin, H. M. Kang, M. A. Bogue, D. A. Hinds, E. J. Beilharz, R. V. Gupta, J. Montgomery, M. M. Morenzoni, G. B. Nilsen, C. L. Pethiyagoda, L. L. Stuve, F. M. Johnson, M. J. Daly, C. M. Wade und D. R. Cox. 2007. A sequence-based variation map of 8.27 million SNPs in inbred mouse strains. Nature **448**:1050-1053.
Gompel, N., B. Prud'homme, P. J. Wittkopp, V. A. Kassner und S. B. Carroll. 2005. Chance caught on the wing: cis-regulatory evolution and the origin of pigment patterns in Drosophila. Nature **433**:481-487.
Gray, S. J., S. P. Jensen und J. L. Hurst. 2000. Structural complexity of territories: preference, use of space and defence in commensal house mice, Mus domesticus. Anim Behav **60**:765-772.
Grula, J. W. 2008. Gene regulation in evolution: a history. Science **322**:1633.
Guenet, J. L. und F. Bonhomme. 2003. Wild mice: an ever-increasing contribution to a popular mammalian model. Trends Genet **19**:24-31.
Haldane, J. B. S. 1932. The causes of evolution. Harper & Brothers, Hrsg., New York
Harr, B., C. Voolstra, T. J. Heinen, J. F. Baines, R. Rottscheidt, S. Ihle, W. Muller, F. Bonhomme, und D. Tautz. 2006. A change of expression in the conserved signaling gene MKK7 is associated with a selective sweep in the western house mouse Mus musculus domesticus. J Evol Biol **19**:1486-1496.
Heinen, T. J. 2008. Characterization of genes involved in recent adaptations. Dissertation, Universität Köln
Hobolth, A., O. F. Christensen, T. Mailund und M. H. Schierup. 2007. Genomic relationships and speciation times of human, chimpanzee, and gorilla inferred from a coalescent hidden Markov model. PLoS Genet **3**:e7.
Hoekstra, H. E. und J. A. Coyne. 2007. The locus of evolution: evo devo and the genetics of adaptation. Evolution **61**:995-1016.
Holloway, A. K., M. K. Lawniczak, J. G. Mezey, D. J. Begun und C. D. Jones. 2007. Adaptive gene expression divergence inferred from population genomics. PLoS Genet **3**:e187.
Hubbard, L., P. McSteen, J. Doebley und S. Hake. 2002. Expression patterns and mutant phenotype of teosinte branched1 correlate with growth suppression in maize and teosinte. Genetics **162**:1927-1935.

Ihle, S., I. Ravaoarimanana, M. Thomas und D. Tautz. 2006. An analysis of signatures of selective sweeps in natural populations of the house mouse. Mol Biol Evol **23**:790-797.
Jacob, F. 1977. Evolution and tinkering. Science **196**:1161-1166.
Jeong, S., M. Rebeiz, P. Andolfatto, T. Werner, J. True und S. B. Carroll. 2008. The evolution of gene regulation underlies a morphological difference between two Drosophila sister species. Cell **132**:783-793.
Johnson, M. E., L. Viggiano, J. A. Bailey, M. Abdul-Rauf, G. Goodwin, M. Rocchi und E. E. Eichler. 2001. Positive selection of a gene family during the emergence of humans and African apes. Nature **413**:514-519.
Jones, C. D., A. W. Custer und D. J. Begun. 2005. Origin and evolution of a chimeric fusion gene in Drosophila subobscura, D. madeirensis and D. guanche. Genetics **170**:207-219.
Kauer, M. O., D. Dieringer und C. Schlotterer. 2003. A microsatellite variability screen for positive selection associated with the "out of Africa" habitat expansion of Drosophila melanogaster. Genetics **165**:1137-1148.
Kelly, J. K. 1997. A test of neutrality based on interlocus associations. Genetics **146**:1197-1206.
Khaitovich, P., S. Paabo und G. Weiss. 2005. Toward a neutral evolutionary model of gene expression. Genetics **170**:929-939.
Khaitovich, P., G. Weiss, M. Lachmann, I. Hellmann, W. Enard, B. Muetzel, U. Wirkner, W. Ansorge und S. Paabo. 2004. A neutral model of transcriptome evolution. PLoS Biol **2**:e132.
Kim, Y. und W. Stephan. 2002. Detecting a local signature of genetic hitchhiking along a recombining chromosome. Genetics **160**:765-777.
King, M. C. und A. C. Wilson. 1975. Evolution at two levels in humans and chimpanzees. Science **188**:107-116.
Kozak, M. 1978. How do eucaryotic ribosomes select initiation regions in messenger RNA? Cell **15**:1109-1123.
Laurie, C. C., D. A. Nickerson, A. D. Anderson, B. S. Weir, R. J. Livingston, M. D. Dean, K. L. Smith, E. E. Schadt und M. W. Nachman. 2007. Linkage disequilibrium in wild mice. PLoS Genet **3**:e144.
Lawniczak, M. K., A. K. Holloway, D. J. Begun und C. D. Jones. 2008. Genomic analysis of the relationship between gene expression variation and DNA polymorphism in Drosophila simulans. Genome Biol **9**:R125.
Lemos, B., C. D. Meiklejohn, M. Caceres und D. L. Hartl. 2005. Rates of divergence in gene expression profiles of primates, mice, and flies: stabilizing selection and variability among functional categories. Evolution **59**:126-137.
Levine, M. T., C. D. Jones, A. D. Kern, H. A. Lindfors und D. J. Begun. 2006. Novel genes derived from noncoding DNA in Drosophila melanogaster are frequently X-linked and exhibit testis-biased expression. Proc Natl Acad Sci U S A **103**:9935-9939.
Libus, J. und H. Storchova. 2006. Quantification of cDNA generated by reverse transcription of total RNA provides a simple alternative tool for quantitative RT-PCR normalization. Biotechniques **41**:156-164
Lodish, H., A. Berk, S. L. Zipursky, P. Matsudaira, D. Baltimore und J. E. Darnell. 1999. Molecular Cell Biology. W.H.Freeman and Company, Hrsg., New York.
Long, M., E. Betran, K. Thornton und W. Wang. 2003. The origin of new genes: glimpses from the young and old. Nat Rev Genet **4**:865-875.

Long, M. und C. H. Langley. 1993. Natural selection and the origin of jingwei, a chimeric processed functional gene in Drosophila. Science **260**:91-95.
Lynch, M., M. O'Hely, B. Walsh und A. Force. 2001. The probability of preservation of a newly arisen gene duplicate. Genetics **159**:1789-1804.
Macholan, M., M. Vyskocilova, F. Bonhomme, B. Krystufek, A. Orth und V. Vohralik. 2007. Genetic variation and phylogeography of free-living mouse species (genus Mus) in the Balkans and the Middle East. Mol Ecol **16**:4774-4788.
Manda, S. O., R. E. Walls und M. S. Gilthorpe. 2007. A full Bayesian hierarchical mixture model for the variance of gene differential expression. BMC Bioinformatics **8**:124.
Martignetti, J. A. und J. Brosius. 1993a. BC200 RNA: a neural RNA polymerase III product encoded by a monomeric Alu element. Proc Natl Acad Sci U S A **90**:11563-11567.
Martignetti, J. A. und J. Brosius. 1993b. Neural BC1 RNA as an evolutionary marker: guinea pig remains a rodent. Proc Natl Acad Sci U S A **90**:9698-9702.
Maston, G. A. und M. Ruvolo. 2002. Chorionic gonadotropin has a recent origin within primates and an evolutionary history of selection. Mol Biol Evol **19**:320-335.
McGregor, A. P., V. Orgogozo, I. Delon, J. Zanet, D. G. Srinivasan, F. Payre und D. L. Stern. 2007. Morphological evolution through multiple cis-regulatory mutations at a single gene. Nature **448**:587-590.
Meijer, H. A. und A. A. Thomas. 2002. Control of eukaryotic protein synthesis by upstream open reading frames in the 5'-untranslated region of an mRNA. Biochem J **367**:1-11.
Metta, M. und C. Schlotterer. 2008. Male-biased genes are overrepresented among novel Drosophila pseudoobscura sex-biased genes. BMC Evol Biol **8**:182.
Muller, H. J. 1935. The origination of chromatin deficiencies as minute deletions subject to insertion elsewhere. Genetica **17**:237-252.
Nakagawa, S., Y. Niimura, T. Gojobori, H. Tanaka und K. Miura. 2008. Diversity of preferred nucleotide sequences around the translation initiation codon in eukaryote genomes. Nucleic Acids Res **36**:861-871.
Nurminsky, D. I., M. V. Nurminskaya, D. De Aguiar und D. L. Hartl. 1998. Selective sweep of a newly evolved sperm-specific gene in Drosophila. Nature **396**:572-575.
Odom, D. T., R. D. Dowell, E. S. Jacobsen, W. Gordon, T. W. Danford, K. D. MacIsaac, P. A. Rolfe, C. M. Conboy, D. K. Gifford und E. Fraenkel. 2007. Tissue-specific transcriptional regulation has diverged significantly between human and mouse. Nat Genet **39**:730-732.
Ohno, S. 1970. Evolution by gene duplication. Springer, New York.
Orth, A., K. Belkhir, J. Britton-Davidian, P. Boursot, T. Benazzou und F. Bonhomme. 2002. Natural hybridization between 2 sympatric species of mice, Mus musculus domesticus L. and Mus spretus Lataste. C R Biol **325**:89-97.
Ortlund, E. A., J. T. Bridgham, M. R. Redinbo und J. W. Thornton. 2007. Crystal structure of an ancient protein: evolution by conformational epistasis. Science **317**:1544-1548.
Pocock, M., H. Hauffe und J. Searle. 2005. Dispersal in house mice. Biol. J. Linn. Soc. **84**:565-583.

Pozhitkov, A. E., D. Tautz und P. A. Noble. 2007. Oligonucleotide microarrays: widely applied--poorly understood. Brief Funct Genomic Proteomic **6**:141-148.
Prud'homme, B., N. Gompel, A. Rokas, V. A. Kassner, T. M. Williams, S. D. Yeh, J. R. True und S. B. Carroll. 2006. Repeated morphological evolution through cis-regulatory changes in a pleiotropic gene. Nature **440**:1050-1053.
Rajabi-Maham, H., A. Orth und F. Bonhomme. 2008. Phylogeography and postglacial expansion of Mus musculus domesticus inferred from mitochondrial DNA coalescent, from Iran to Europe. Mol Ecol **17**:627-641.
Rifkin, S. A., J. Kim und K. P. White. 2003. Evolution of gene expression in the Drosophila melanogaster subgroup. Nat Genet **33**:138-144.
Rodriguez-Trelles, F., R. Tarrio und F. J. Ayala. 2003. Convergent neofunctionalization by positive Darwinian selection after ancient recurrent duplications of the xanthine dehydrogenase gene. Proc Natl Acad Sci U S A **100**:13413-13417.
Rottscheidt, R. und B. Harr. 2007. Extensive additivity of gene expression differentiates subspecies of the house mouse. Genetics **177**:1553-1567.
Rozas, J., J. C. Sanchez-DelBarrio, X. Messeguer und R. Rozas. 2003. DnaSP, DNA polymorphism analyses by the coalescent and other methods. Bioinformatics **19**:2496-2497.
Sage, R. 1981. Wild mice. The mouse in biomedical research **1**:39-90.
Schwarz, G. 1978. Estimating the Dimension of a Model. The Annals of Statistics **6**:461-464.
Shapiro, M. D., M. E. Marks, C. L. Peichel, B. K. Blackman, K. S. Nereng, B. Jonsson, D. Schluter und D. M. Kingsley. 2004. Genetic and developmental basis of evolutionary pelvic reduction in threespine sticklebacks. Nature **428**:717-723.
Simonsen, K. L., G. A. Churchill und C. F. Aquadro. 1995. Properties of statistical tests of neutrality for DNA polymorphism data. Genetics **141**:413-429.
Singleton, G. R. 1983. The Social and Genetic Structure of a Natural Colony of House Mice, Mus musculus, at Healesville WildlifeSanctuary. Austral. J. of Zool. **31**:155-166.
Maynard Smith, J. und J. Haigh. 1974. The hitch-hiking effect of a favourable gene. Genet Res **23**:23-35.
Stephens, M., N. J. Smith und P. Donnelly. 2001. A new statistical method for haplotype reconstruction from population data. Am J Hum Genet **68**:978-989.
Storey, J. D. und R. Tibshirani. 2003. Statistical significance for genomewide studies. Proc Natl Acad Sci U S A **100**:9440-9445.
Sugiura, N. 1978. Further analysis of the data by Akaike' s information criterion and the finite corrections. Communications in Statistics-Theory and Methods **7**:13-26.
Tajima, F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics **123**:585-595.
Tautz, D. 2000. Evolution of transcriptional regulation. Curr Opin Genet Dev **10**:575-579.
Teschke, M., O. Mukabayire, T. Wiehe und D. Tautz. 2008. Identification of selective sweeps in closely related populations of the house mouse based on microsatellite scans. Genetics **180**:1537-1545.
Thomas, K. R. und M. R. Capecchi. 1987. Site-directed mutagenesis by gene targeting in mouse embryo-derived stem cells. Cell **51**:503-512.

Townsend, J. P., D. Cavalieri und D. L. Hartl. 2003. Population genetic variation in genome-wide gene expression. Mol Biol Evol **20**:955-963.
Tusher, V. G., R. Tibshirani und G. Chu. 2001. Significance analysis of microarrays applied to the ionizing radiation response. Proc Natl Acad Sci U S A **98**:5116-5121.
Voolstra, C., D. Tautz, P. Farbrother, L. Eichinger und B. Harr. 2007. Contrasting evolution of expression differences in the testis between species and subspecies of the house mouse. Genome Res **17**:42-49.
Wang, W., F. G. Brunet, E. Nevo und M. Long. 2002. Origin of sphinx, a young chimeric RNA gene in Drosophila melanogaster. Proc Natl Acad Sci U S A **99**:4448-4453.
Waterston, R.H.K., E. S. Lander und das mouse genome sequencing consortium 2002. Initial sequencing and comparative analysis of the mouse genome. Nature **420**:520-562.
Watterson, G. 1975. On the number of segregating sites in genetical models without recombination. Theoretical Population Biology **7**:256-276.
West, J. D., W. I. Frels und V. M. Chapman. 1978. Mus musculus x Mus caroli hybrids: mouse mules. J Hered **69**:321-326.
Whitehead, A. und D. L. Crawford. 2006a. Neutral and adaptive variation in gene expression. Proc Natl Acad Sci U S A **103**:5425-5430.
Whitehead, A. und D. L. Crawford. 2006b. Variation within and among species in gene expression: raw material for evolution. Mol Ecol **15**:1197-1211.
Wiehe, T., V. Nolte, D. Zivkovic und C. Schlotterer. 2007. Identification of selective sweeps using a dynamically adjusted number of linked microsatellites. Genetics **175**:207-218.
Williams, T. M., J. E. Selegue, T. Werner, N. Gompel, A. Kopp und S. B. Carroll. 2008. The regulation and evolution of a genetic switch controlling sexually dimorphic traits in Drosophila. Cell **134**:610-623.
Wilson, A. C., L. R. Maxson und V. M. Sarich. 1974. Two types of molecular evolution. Evidence from studies of interspecific hybridization. Proc Natl Acad Sci U S A **71**:2843-2847.
Wittkopp, P. J., B. K. Haerum und A. G. Clark. 2008. Regulatory changes underlying expression differences within and between Drosophila species. Nat Genet **40**:346-350.
Wray, G. A. 2007. The evolutionary significance of cis-regulatory mutations. Nat Rev Genet **8**:206-216.
Yanai, I., D. Graur und R. Ophir. 2004. Incongruent expression profiles between human and mouse orthologous genes suggest widespread neutral evolution of transcription control. Omics **8**:15-24.
Yang, H., T. A. Bell, G. A. Churchill und F. Pardo-Manuel de Villena. 2007. On the subspecific origin of the laboratory mouse. Nat Genet **39**:1100-1107.
Zahn, L. M., H. Kong, J. H. Leebens-Mack, S. Kim, P. S. Soltis, L. L. Landherr, D. E. Soltis, C. W. Depamphilis und H. Ma. 2005. The evolution of the SEPALLATA subfamily of MADS-box genes: a preangiosperm origin with multiple duplications throughout angiosperm history. Genetics **169**:2209-2223.
Zhang, Y., D. Sturgill, M. Parisi, S. Kumar und B. Oliver. 2007. Constraint and turnover in sex-biased gene expression in the genus Drosophila. Nature **450**:233-237.

Zhou, Q., G. Zhang, Y. Zhang, S. Xu, R. Zhao, Z. Zhan, X. Li, Y. Ding, S. Yang und W. Wang. 2008. On the origin of new genes in Drosophila. Genome Res **18**:1446-1455.

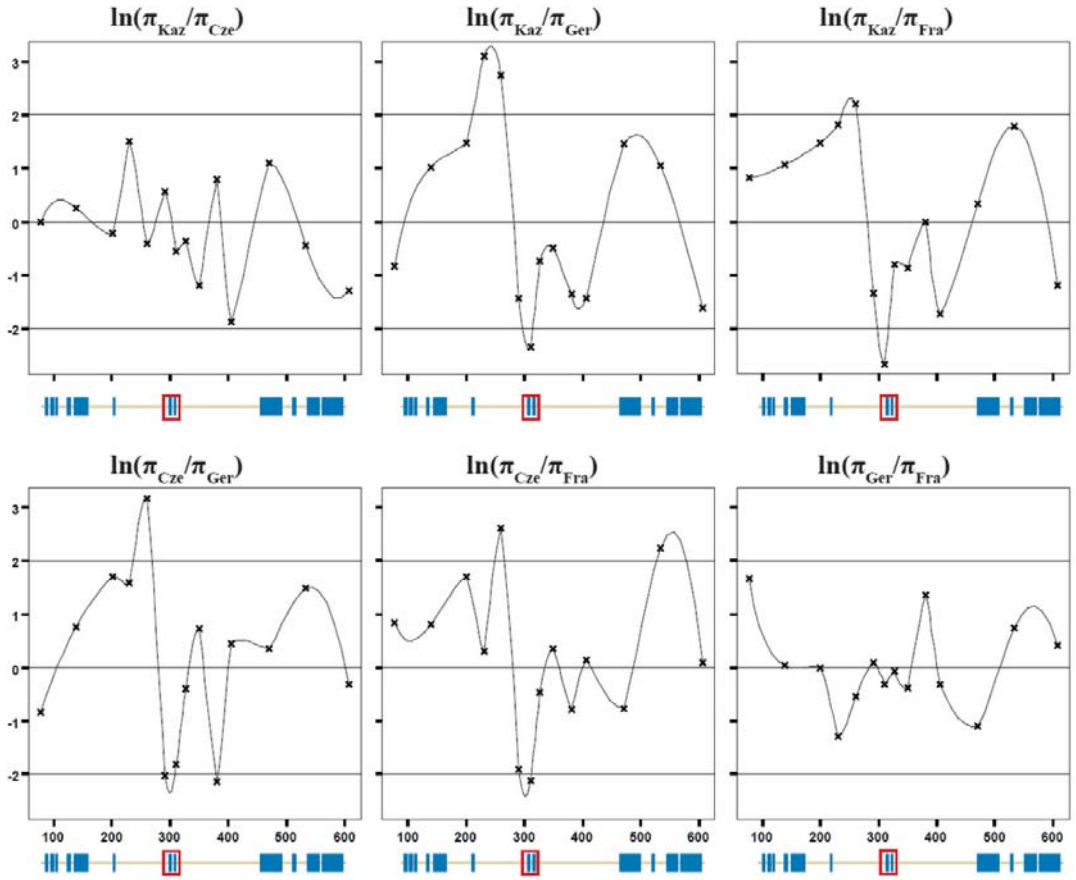
4.3 Weiterführende Tabellen und Abbildungen

S 1 Tabelle der populationsgenetischen Daten in der Poldi Region. Die Stromaufwärtsregion von Poldi liegt bei Position 310483.

Pop	Position	#Chromo-somes	bp	π in %	S	θ/Länge	Tajima's D
AL	77305	22	573	0.181	2	0.00096	2.02478
AL	138639	18	616	0.087	2	0.00094	-0.19106
AL	200858	18	657	0.098	3	0.00133	-0.71573
AL	230367	22	513	0.49	6	0.00321	1.63658
AL	260533	22	241	0.345	2	0.00228	1.1667
AL	291129	22	561	0.103	2	0.00098	0.11197
AL	310483	20	752	0.066	5	0.00187	-1.97429
AL	326970	22	625	0.175	9	0.00395	-1.86487
AL	350121	22	602	0.12	4	0.00182	-0.95805
AL	380921	20	457	0.532	12	0.0074	-1.00912
AL	406192	22	607	0.095	2	0.0009	0.11197
AL	470828	18	502	0.189	2	0.00116	1.5371
AL	533888	22	608	0.246	4	0.0018	1.02513
AL	607121	20	396	0.128	1	0.00071	1.43024
CR	77305	22	573	0.183	2	0.00096	2.06053
CR	138639	22	616	0.067	1	0.00045	0.89527
CR	200858	22	657	0.121	3	0.00125	-0.08306
CR	230367	22	519	0.108	2	0.00106	0.04046
CR	260533	22	241	0.519	4	0.00455	0.39361
CR	291129	16	561	0.058	1	0.00054	0.15575
CR	310483	28	752	0.113	3	0.00103	0.25201
CR	326970	22	625	0.249	9	0.00395	-1.23603
CR	350121	22	602	0.397	13	0.00592	-1.16856
CR	380921	22	457	0.241	4	0.0024	0.00584
CR	406192	22	607	0.623	17	0.00768	-0.69334
CR	470828	22	501	0.062	1	0.00055	0.23682
CR	533888	22	608	0.382	7	0.00316	0.67398
CR	607121	18	396	0.464	6	0.00441	0.174
D	77305	20	573	0.418	5	0.00246	2.13938
D	138639	20	616	0.031	1	0.00046	-0.59155
D	200858	14	657	0	0	0	0
D	230367	19	519	0	0	0	0
D	260533	20	241	0	0	0	0
D	291129	20	561	0.434	5	0.00251	2.23079
D	310483	30	752	0.69	15	0.00537	0.96493
D	326970	18	625	0.368	5	0.00233	1.84421
D	350121	18	602	0.195	4	0.00193	0.03489
D	380921	16	457	2.068	21	0.01385	1.9917
D	406192	14	607	0.396	6	0.00311	0.98915
D	470828	18	502	0.044	2	0.00116	-1.50776
D	533888	18	608	0.086	1	0.00048	1.50518
D	607121	18	396	0.635	5	0.00367	2.31385
MC	77305	22	573	0.079	5	0.00239	-1.98725
MC	138639	22	616	0.03	2	0.00089	-1.51481
MC	200858	22	657	0	0	0	0

Pop	Position	#Chromo-somes	bp	π in %	S	θ/Länge	Tajima's D
MC	230367	22	519	0.08	1	0.00053	0.89527
MC	260533	22	241	0.038	1	0.00114	-1.1624
MC	291129	18	561	0.393	5	0.00259	1.63074
MC	310483	18	752	0.937	16	0.00619	1.96488
MC	326970	22	625	0.391	6	0.00263	1.50617
MC	350121	22	602	0.283	7	0.00319	-0.35767
MC	380921	22	457	0.532	11	0.0066	-0.67042
MC	406192	22	607	0.54	11	0.00497	0.2976
MC	470828	22	502	0.133	3	0.00164	-0.49124
MC	533888	22	608	0.041	1	0.00045	-0.17472
MC	607121	22	396	0.42	5	0.00346	0.63001

S 2 Vergleiche der genetischen Variabilität am Poldi Locus in den einzelnen Populationen. x-Achse: Position in kb. Die zwei vertikalen Linien bei Position 300kb (rot umrahmt) markieren das erste und das letzte Exon des Poldi Transkripts (1700125f08Rik). Die entfernteren vertikalen Linien markieren Exons benachbarter Gene. Wurde in einem Fragment kein Polymorphismus entdeckt, wurde π für die Berechnung von $\ln(\pi_{pop1/}\ \pi_{pop2)}$ durch den niedrigsten Wert im Datensatz ersetzt (0,00022).

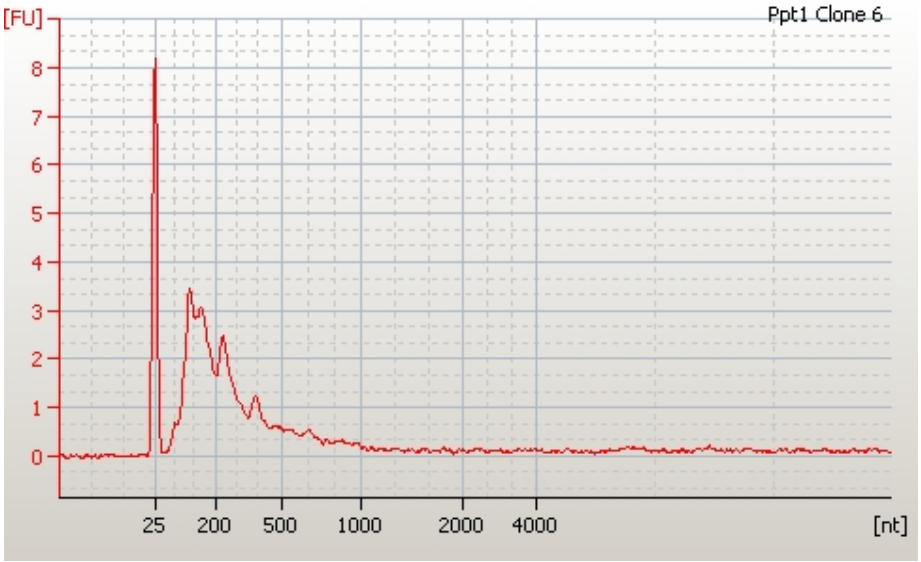


S 3 Tabelle Kontrolle für multiples Testen mittels FDR und q-values (Storey and Tibshirani 2003). Signifikante Expressionsunterschiede zwischen den Subspezies sind fett gedruckt.

MGI gene symbol	Tissue	p-value	q-value
1110017D15Rik	brain	0.631	0.238
1110017D15Rik	liver/kidney	0.522	0.222
1110017D15Rik	testis	0.078	0.071
1700125F08Rik	testis	0.423	0.197
Cacng2	brain	0.631	0.238
Cacng2	testis	0.631	0.238
Ccl25	brain	0.522	0.222
Ccl25	liver/kidney	0.262	0.142
Ccl25	**testis**	**0.004**	**0.020**
Cdk5	**brain**	**0.004**	**0.020**
Cdk5	**liver/kidney**	**0.020**	**0.031**
Etd	brain	0.423	0.197
Etd	testis	0.128	0.094
Etv2	**testis**	**0.025**	**0.036**
Flot2	brain	0.337	0.173
Flot2	**liver/kidney**	**0.016**	**0.027**
Flot2	testis	0.873	0.291
Gpc6	**brain**	**0.004**	**0.020**
Gpc6	liver/kidney	1.000	0.323
Gpc6	testis	0.522	0.222
Hif1a	brain	0.262	0.142
Hif1a	liver/kidney	0.078	0.071
Hif1a	testis	0.262	0.142
Kcnd2	**brain**	**0.010**	**0.023**
Kcnd2	**liver/kidney**	**0.016**	**0.027**
Krt2-17	brain	0.109	0.084
Mir16	brain	0.749	0.263
Mir16	liver/kidney	0.748	0.263
Mir16	testis	0.262	0.142
Nf1	brain	0.150	0.103
Nf1	liver/kidney	1.000	0.323
Nf1	testis	0.810	0.279
PanX1	brain	0.522	0.222
PanX1	liver/kidney	0.055	0.058
PanX1	testis	0.749	0.263
Ppt1	brain	0.200	0.125
Ppt1	liver/kidney	0.631	0.238
Ppt1	**testis**	**0.006**	**0.022**
Rab4b	brain	0.337	0.173
Rab4b	liver/kidney	0.010	0.023
Rab4b	testis	0.109	0.084
Rarres2	**brain**	**0.006**	**0.022**
Rarres2	liver/kidney	0.262	0.142
Rarres2	testis	0.873	0.291
Rgs16	**brain**	**0.010**	**0.023**

MGI gene symbol	Tissue	p-value	q-value
Rgs16	liver/kidney	0.423	0.197
Rgs16	testis	0.109	0.084
Scamp5	brain	0.200	0.125
Scamp5	liver/kidney	0.423	0.197
Scamp5	**testis**	**0.037**	**0.050**
Sv2c	**brain**	**0.004**	**0.020**
Sv2c	liver/kidney	0.631	0.238
Sv2c	testis	0.200	0.125
Tcte3	brain	0.055	0.058
Tcte3	liver/kidney	0.055	0.058
Tcte3	testis	0.631	0.238
Tmem24	brain	0.109	0.084
Tmem24	liver/kidney	0.078	0.071
Tmem24	testis	0.055	0.058
Tomm40l	brain	0.749	0.263
Tomm40l	liver/kidney	0.150	0.103
Tomm40l	**testis**	**0.016**	**0.027**

S 4 Ergebnis der *in vitro* Transkription (Agilent 2100 Bioanalyzer). Eine RNA der Zielgröße wurde erfolgreich transkribiert. Konzentration: 2,6µg/µl

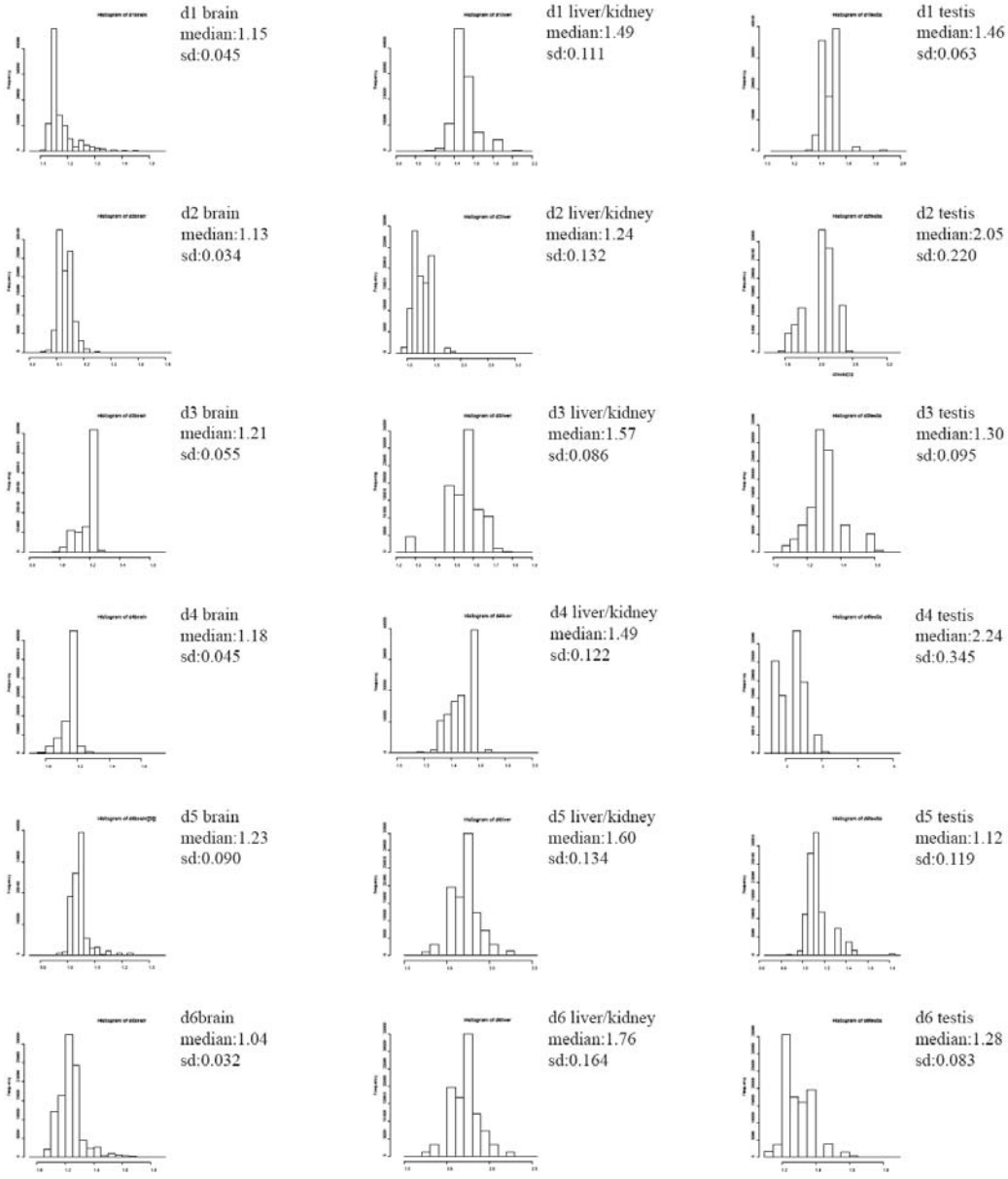


S 5 Liste des Umfangs der im Rahmen der Analyse der Stromaufwärtsregionen erhobenen Daten. Die Länge der analysierten Sequenz (bp) und die Anzahl der analysierten Chromosomen (#Chromosomes) ist für die beiden untersuchten Subspezies für jedes Gen angegeben.

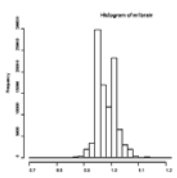
MGI Gene Symbol	Subspecies	bp	#Chromosomes
1110017D15Rik	domesticus	968	30
AK003742.1	musculus	968	26
1700125F08Rik	domesticus	752	30
AK007277	musculus	752	28
Cacng2	domesticus	846	28
NM_007583	musculus	846	30
Ccl25	domesticus	920	32
NM_009138	musculus	920	14
Cdk5	domesticus	894	28
NM_007668	musculus	894	34
Etd	domesticus	984	30
NM_175147.2	musculus	984	34
Etv2	domesticus	753	30
NM_007959	musculus	753	30
Flot2	domesticus	841	34
NM_008028.1	musculus	841	32
Gpc6	domesticus	539	34
NM_011821	musculus	539	34
Hif1a	domesticus	594	26
NM_010431	musculus	594	24
Kcnd2	domesticus	870	34
NM_019697	musculus	870	34
Krt2-17	domesticus	999	30
NM_010668	musculus	999	34
Mir16	domesticus	614	34
NM_019580	musculus	614	34
Nf1	domesticus	498	34
NM_010897.1	musculus	499	34
PanX1	domesticus	869	34
NM_019482	musculus	830	22
Ppt1	domesticus	593	30
NM_008917.1	musculus	593	34
Rab4b	domesticus	716	24
NM_029391.1	musculus	716	24
Rarres2	domesticus	997	30
NM_027852.1	musculus	997	26
Rgs16	domesticus	487	30
NM_011267.1	musculus	487	34
Scamp5	domesticus	488	32
NM_020270	musculus	488	34
Sv2c	domesticus	770	28
AK173092.1	musculus	770	24
Tcte3	domesticus	483	28
NM_011560.2	musculus	483	34
Tmem24	domesticus	689	32
NM_027909.1	musculus	686	28
Tomm40l	domesticus	618	20
AK186544.1	musculus	618	30

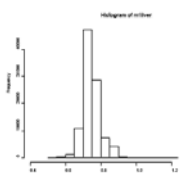
S 6 Histogramme des 100000fachen Bootstrapping des Korrekturfaktors (Cor) über alle Gene für jede Maus (domesticus = d1-d6 musuclus = m1-m6) und jedes Gewebe der Studie. Die Standardabweichungen (sd) sind klein (Mittelwert der Standardabweichung: 0.082)

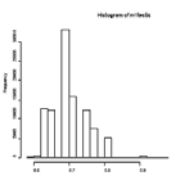
domesticus

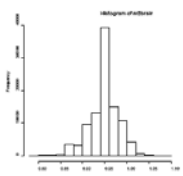


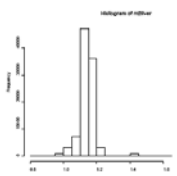
musculus

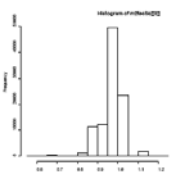

m1 brain
median:0.97
sd:0.035

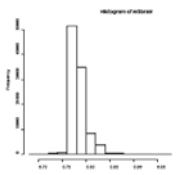

m1 liver/kidney
median:0.74
sd:0.054

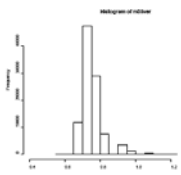

m1 testis
median:0.70
sd:0.052

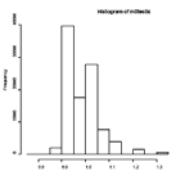

m2 brain
median:0.94
sd:0.033

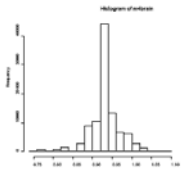

m2 liver/kidney
median:1.13
sd:0.054

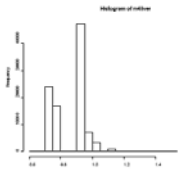

m2 testis
median:0.98
sd:0.059

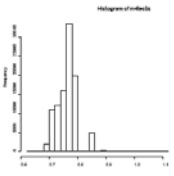

m3 brain
median:0.78
sd:0.013

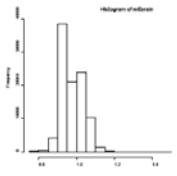

m3 liver/kidney
median:0.74
sd:0.058

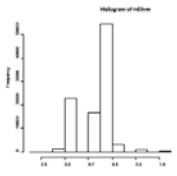

m3 testis
median:0.96
sd:0.065

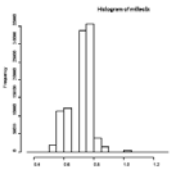

m4 brain
median:0.92
sd:0.037

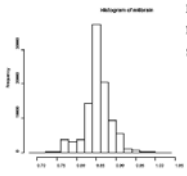

m4 liver/kidney
median:0.90
sd:0.091

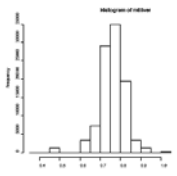

m4 testis
median:0.77
sd:0.032

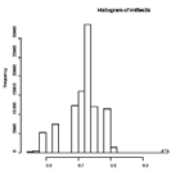

m5 brain
median:0.97
sd:0.062


m5 liver/kidney
median:0.75
sd:0.069

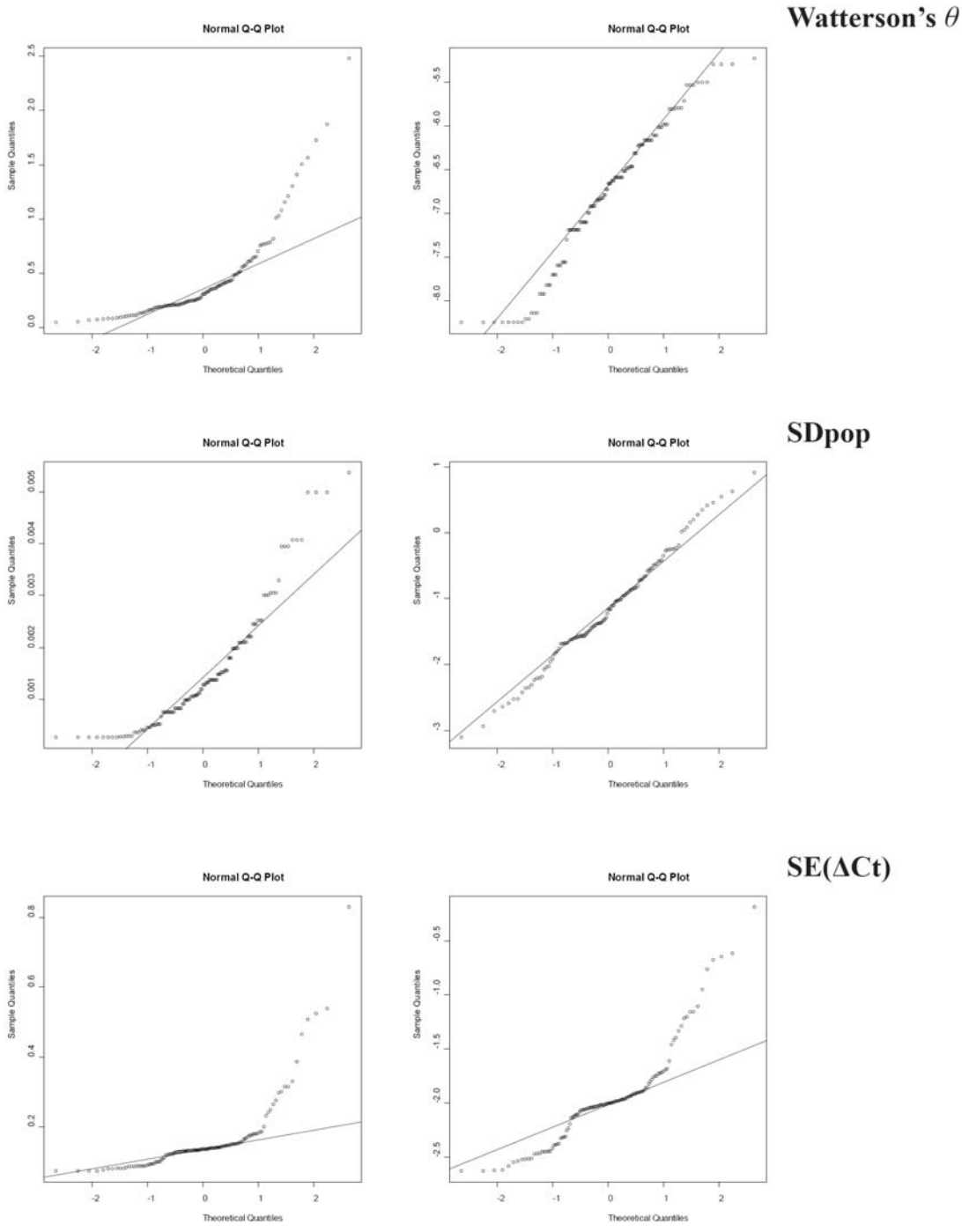

m5 testis
median:0.75
sd:0.086


m6 brain
median:0.86
sd:0.033


m6 liver/kidney
median:0.77
sd:0.070


m6 testis
median:0.73
sd:0.051

S 7 QQ-Plots von Watterson's θ, SD_{pop} und SE(ΔCt) vor (links) und nach Logarithmierung(rechts). Die Logarithmierung bewirkt in allen drei Fällen eine Annäherung an die Normalverteilung. Der Unterschied ist für SE(ΔCt) zwar nicht so deutlich, aber die Teststatistik (W) ist auch hier für den logarithmierten Datensatz größer (0,88 statt 0,61) und damit näher an der Normalverteilung.



4.4 Primerlisten

S 8 Primer zur Erfassung des Variabilitätstals

Position	*F*	*R*
200858	GGGGAAAGCCTGAGCTCTGGCC	GGCAGACTCGGGTATTTAAAG
230367	CTACGCGTCAAGACCACCTAGG	CTTTGGCAGCCCAGCTCTGTCC
260533	CCAGGAAGTGACCCGGTGGTGG	TCCAGGAGGAGCAGGTGCGTGC
291129	TCTAGCATCCCCCTTCTCTGGG	GGATTAGAAGTGCGCACCGCCA
310483	1700125F08Rik Tabelle S 12	
326970	CAACCAGAGTAGCATTCTAAGC	GGTGTTTGCTGTGCAGGCA
350121	CCTGCAGACCTCTTAGCTTCGC	GGTGGGGAATCAGGGTCAGTCC
380921	TGCCTCGGCTTATTAGCCATCC	CATTTGCAGTGGTGACCCCTCG
406192	ACAAACAGTCCAAAGCCACTGG	CCTAGACAGCCCCAAACCAATA
138639	GGCACAAGGGAAGCTGAGCTGG	TAGAGAGCTCTGGACTGTTCCC
77305	CAGCTAATGAGCCCCAGAGAGC	CCTGTTCTGGGTGTGACTCCA
470828	CTACCACAAAGGAAGGGAGCGGC	TTCCTCAGGCCCCGCTAAAAGG
533888	GCACATATGAGGCAAGGGGAGC	AGGTAATTGTGAGAATGGCTG
607121	TCCACTTCCCAAATAACTCACG	CTGCCAAGAGTGGCCTTGAGC

S 9 cDNA Primer

	F	*R*
POLDI	AACACAGAGGGCAGAAGTCC	GGGCTCTGCTAGGCTGCTGG
AK158810	TTCCTCTGCCCTAGGCAACC	GTCTTTGCAGCTGTGACACC

S 10 Primer Poldi Exons

	F	*R*
EXON1	ACTACAGTTCCTCATCCAATGC	ACTTTTATGAATCTCGATGTGG
	1700125F08Rik Tabelle S 12	
EXON2	TGTTGCCTCCTATGTGAGCTGG	TGCAGACGGACAGGGCAGCCAC
	TCCCCCATGAGGAGTCTTCCAG	CCTAATCAGGGTGTCAGGGAC
EXON3	GGGCTTGTCAGGGTCAGAGGCC	CCCTCTCATTCTTTCTCCACAC
	GCCATCAAGTTGTCATTCCAGG	CCTGGAATGACAACTTGATGGC

S 11 Primer Ak158810 Exons

	F	*R*
EXON1	CCTCTAACTTCCAGGGCTTAGG	TGTATGTCCCCACCACAGAGCG
	AACCCTGCTCTGCAAGCTGTCAG	CAGAAGAACTCAGCTCTGAAGC
	ATCTGAGTGGGCACATCCTGCC	GACAGCACTTCTCAGAACACGC
EXON zusatz	GAGCGGGAGCTTCCATCACAGG	CGAGCTGGGAGGAAGGGACAGG
	AGGGTTAGGCTAGCTTGATTGC	GGAAGGAGACTGACCACTGACC
EXON2	TCCGCTCAGCAAGGCCCCAGGC	GATAATCAAAGACCTTTTGGGC
	TCAGGGTCAGAACAGTCAGGGG	CCTGGCACCTAGGATGCGCCCC
	CCCAGGCACTGGGTGTCACAGC	GATCCGTACCTCAAATCAATGC
EXON3	CAGCCACTTGTGGACCTTGGCG	AGGCCTGGTGGCTGCTTGGTCC
	CCATTTCCAAGCATGGTCCTGCC	CTCTCTCCATCGCCTTTGATGG
	TAGCAGTGGGGATGGAGAGGGC	GCAGGTGCTTGTGTATGACTCC
EXON4	TGTACACGCAGGCTTGGCAGCC	CACCAGACAGATTCCTGGTGGC
EXON5	GCTGCAGCTTCCTGGATGCTGG	CCTCAGGCATTGCTATTCCATC
	AAAAATTGCCCCAAAATGGTGG	GCTGGGATTGGAGACATACACC
	GTTAAGTCAGTTTCGTATTTCC	GTAGACTTGCACTGGAATCTCC
	GCTAAGCCCTTGCACAGCCTCC	AGATGGCTCAGCAGTTAAGAGC
	TGAGCCACCATGTGGTTGCTGG	AACTGCATTCTCCTACCCCAGC
	TCTTCAGACACACCAGAAGAGG	AACTGCATTCTCCTACCCCAGC
	ATGTAAGTATACTGTTTCTGTC	AACTGCATTCTCCTACCCCAGC
	TTATATAACTGGCATCACAGCG	AACTGCATTCTCCTACCCCAGC
	GTAGACTTGCACTGGAATCTCC	TTCATAATTGAGTAAGACCTGG
	TCCTCTTGCTGCAGCTTCCTGG	ACCCTCTAGGTTCCTCCACCG

S 12 Liste der Primer, die für die Sequenzierung der Stromaufwärtsregionen verwendet wurden

MGI Gene Symbol /Acc. Num.	PCR primers		sequencing primers	
direction	*F*	*R*	*F*	*R*
1110017D15Rik	CATTCTGCAGGGTCTCTTCCCC	CCTCAGCAGCTCCCTGTTCTCC		
AK003742.1	GTGCTGAAATTTGAAGCCAGGC	CACACTCGTGCACGCGTGCACG		
1700125F08Rik	GCCACAAACCTACGTGTGGTGG	CTGTACACGCAGGCTTGGCAGC	TTCCACCATTAGGGTTCCACTG	GTCGTGACTGCTTGGCAAACCT
AK007277				
Cacng2	CTGGAATTTACACCCAAGGAAC	AAACCTCTCTACTATGAGCA	ATCGCCAGCTACGCCTTCCTCCCA	GTGTTTCTTCCAGTTCAGCCT
NM_007583	TCAATCTCATTTATGAATGAC	CCCACCTACTGCGTCGCCGTGGT		
Ccl25	GGCCAGGACAGAGCAAGAGAGC	ATGCCTTTCTGGTCCTGAGAGC		
NM_009138	ACTGCAGGGTGGGGCTCTGACT	CATCTACCTCCAGTAGTACCAG		
Cdk5	TGAAAGACCCTTGCGTGTCATC	AGCGATCAACTCCAGGGACTCG	CGCTGGAGTCCGGTGGGTTTCG	GACCTCACAGGGACCAGCTGAC
NM_007668				
Etd	GAGGGAAATCCCAGGAAGGATG	CTCTGCAGCATTTCTCCACAAC		
NM_175147.2	GAAGCAGGTATTTCTATTCTCA	TGCATCAGCTGAAAGACTTCCT		
Etv2	TGTCACGACTTCTCGGACCCAG	TCCCAAGCCACAGCAGCTTACC	GCAACTTGACCCAGGCTGCGAC	AGTTCGTGGCTCACCTCTGGCA
NM_007959	AAGCCAAGGTTCGACAAGACTT	CAGGAAGAGGGATTTCCGGCCA		
	GATGTTATTTTGATTATGGTTG	AGCCAGGGGGGTTTCAGCCCCAT		
Flot2	GCGCCCACTGGCTTGGCTGGTG	TTACAGTGTCTCTTAGGAAGTA		
NM_008028.1	ACCTCGGAACCTCTTGTCATGT	AGCCCCGCCCGCCTGCGCCCCTCT		
Gpc6	TCCCTGGCTTTGTGTTAGGTAC	AGCTCGCCTATCCAGTGTGCG		
NM_011821	CTCCAACGATTTTCTACCGGAG	CAGCCCGGATCCAAGAAGGCAT		
Hif1a	ACTGGAAACTCGGGCGGGATGG	GAGGGAAAAGCCGAGGGGTGCG		
NM_010431	CCAGCATAGCCGGTGAGCAGTC	GCCAACTTTGCGTTCTGGCAGC		
Kcnd2	CTGGGATCTGGCTGCTCGGGAG	GCAACAGGCATCCACCCAATGG	AGTACAGGCGGCCAGCGGACTC	AGGGAAAGTCAACCAGTCAGTG
NM_019697				
Krt2-17	ACCGGAACCTTGACCTGGACAG	TCCACATCCAGGGACAGCTTGG	GAAGCTGTGTCTCCAACCACTG	GAAGCTGTGTCTCCAACCACTG
NM_010668	GAGCTTCACTAACGGGGTCACA	ACCTGCTTCTTCACATGTGATA		
Mir16	TCTGGAAGAGCAGCCAGTACTC	GCAAGGGCTGTTTCCACCAGGA		
NM_019580	AATGCAGACATGTGTCCTGGTG	CAAGTCCGTTTTCAGGGTTCTG		
Nf1	GAAGCCCCATCGACTGCGT	AGCGAGTCCTCCTGGAGGTGAC		
NM_010897.1	GCAGCAGGCCCCTTCCCTCTCG	CGGCGAGCCGAGCGGCTGAGGA		
PanX1	ACTACTGCGAGACCCAACCGAG	CATTTACGGCAAACGGCCTTGG	CAGCACTCCATAGCCATCTGGA	CTGCACAGCCAGCAACCAGCAC
NM_019482	TGCCTTTCGAGGAACAGACAGG	TGCCCAGCCCTTTGCCCTTTCC		
	GAACAGAACCCGGATTGCACCC	GCTCTTCGCTACAGCTGCCCGC		
Ppt1	TTCATATGTCGCTCTCTACAAG	GGACGACGCCATCTTAGCAATC	CCTTCCCCAGTCCCGACACTGA	GTGAGAAATTAGGTAAGGTCCT
NM_008917.1				
Rab4b	ATACCCGGGATACAACAGTGAAC	CTGTACGTTCTATGCGCTTCTC	GTTGTTGGTATGTACCCAAGCT	TTAGCTTCCAGTACTGCACCCT
NM_029391.1	TGTAACGTCAGTGGCTACTAGG	AGCTCCTTTCCACTTCTGGACC		
	AGCCCTTTGCTCCAAGGTAAGG	GCCACTTCCGCCTTCACCGCCC		
Rgs16	CCTTACTGCATAGAATTCCATA	ATGAGAGACCTTAGAGACTCCA		
NM_011267.1	TGCTCCCTGGGGAAGTCCAG	GGTCAGAACAGGTACCATCCCA		
Scamp5	AACTGGCTTTGAACTTGCACTG	GTAGTGCCAGCAACTAGTGCCT		
NM_020270	TTCCCGGTACATCCCACAGGTC	GCCCTGCCTGCTGCAACCACTC		
Sv2c	CAGAGCTGGCCCATATGGCTCA	CAAACTGGACTCAGAGCAGGTG		
AK173092.1	TCACAATCATAATCACTCCT	CAGCAGCTCGGCTGCAACAGCA		
	GGTGAGTTGGATGAGGCTAAGG	ATAATGTACACTGTGCACCAGG		
Tcte3	TGAATATTTCTCCTCTGATTTA	TGGAACTGCGGTATAATAAAGG		
NM_011560.2	AATCAGAGATACTTTACAGGGA	GCCTACGCCGCGTGCAGCCGCG		
Tmem24	AAAAGGGTACATCCAGAGGTTG	TTACTCCTTGACTGAAGCTAGCT		
NM_027909.1	TTCTCGCCGCCGTTGACATTAG	CGGCACCGGCTCCGGGAGGGCT		
Tomm40l	GTCGTTGAGACAGGGTTTCTCT	CTTCCATATGGTTCAATGCTAG	ATAGGCAAAATGCTCCCTCTAG	GAGTTTGTGGGTGGCCCTTTAG
AK186544.1	TGAGCCACTATGTGATTGCTGG	GCAGGCCCCAGGGCTTCAGTCC		

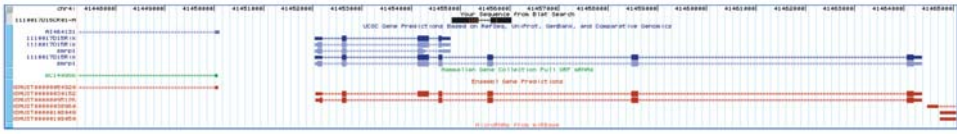
S 13 Primer Polymorphismuskontrolle der Assaybinderegion

MGI Gene Symbol	*F*	*R*
1110017D15RIK	CGCCTACGAGCGCGAGGTAGT	ACTAGGAAGGGTAGACAGGACAG
1700125F08RIK	TCAGACCTGGATTCTTGCCCTGG	TCTGCTTCTTCCTGTGACCTTC
	AGAAAGAACGATGTGGGTGAAG	CAGGAAAGATGAGTTGGCCCTG
	GATTCCTGGTGGCACTCTCTAG	CAGCCTCCACTGCTGCTATTGC
4833411C07RIK	TACGCTCTTCTGTCTTTGGGGA	GACAATTAAGATTGACTTCCAC
		CAGATGTACCTCGATTAACCTG
AI604832	TATTGGCGTATGGCAGTTGGCT	GCTTCATTTGCTGTTGCTTGAG
Cacng2	GCTGTTTGATCGAGGTGTTCA	CTCGAGGCCCTCACGGCCCGGAG
Ccl25	CAACCTACGTGCTGTGAGATTC	GGGCATCATCACCATCCTGGGA
Cdk5	CTGCAACGCCAGGGCCGGGAGT	GTGACCACCTCAGCAGAGTAGC
Crisp1	TTATACTCAGGTTGTTTGGAAC	TAAGCCTAACTAGAGAGTATGAGC
Dscaml1	GCCAGGAACCAATCCAGTGTCC	GCGGCAGATACCAGGCTCTGAG
Edf1	GCCGCGCCGGGTCTCGAGCAG	TTCTCATTGATTTTCGTGGCCA
	GGGACTGTTAACTTTCTTCCTA	CTCCTGACAGGTCCTCAGGGTA
	CTGCCAGAGCCGCCGGACGGAT	TTCCCCCGGAGCTTGAGGCCGA
Etd	CACCCAGAGCCAAGGCTTTCTC	AGTTGCAATGAAGACTACTTTA
Etv2	CTCGTGGTCGCACCCTCCAGCT	CGCCCCACAGCCGGGCCACCTC
	CCAGGCGGAGCCCGTTGCTC	CGGGCCACCTCTTTGGGGTCGC
Flot2	GTGGTCTCAGGAGGCTGTTGT	TGCATTCAGCTTCCCGGATGCC
Gpc6	ACACAGCAAAGCCAGATACCTG	CTGTAGAGTCTCTGCAGGGCCA
Hif1a	TACATGGGGTTAACTCAGTTTG	CGTGCAGTGAAGCACCTTCCA
Kcnd2	CGAAGGGCACAGAAGAAAGCCAG	GATGCTGCACAGAAACTCGTCC
Krt2-17	TCATTGACAAGGTGCGATTCCT	CTGCTGTAGCTGGGACAGTTCCGC
Mir16	AGTCATCTATAAAATGAGACAAAC	GCTGTCAGTGATGTAGCTGGAC
Nf1	TCAGTGGTTAGCCAGCGCTTCC	CCAACAGCTTTATGATCCCTGT
Nkx2-9	ATGGCCACCTCTGGACGCCTCG	TCGCCCTCGCTTCAGCTTGTAG
Nr4a1	CTGGCATACCGATCTAAACCCG	TCAGGCAGTTTGCCCAGCAGAC
PanX1	GCTGTTTGCCATACTCCTGTAC	GATGCTGCACAGAAACTCGTCC
Ppil3	CATGTGAGAATTTCTTGGCTC	TACATCATTAAGAGGTCTGTAT
Ppp1r11	AGAGCATCCTGGCCTCTACTCC	CATAAATACAGCAGCATTTCGA
Ppt1	CATGATGGAGGATGTGGAGAAC	ACTCATTGACACACCTCTCTTGA
Rab4b	GCGGTGAAGGCGGAAGTGGCGG	TCCGCGTCACCGACCGAAACCGC
	GAAGTGGAAAGGAGCTGAGGCT	GGAGTCCTGTTTGAACTTATTC
	TTTGTATTTATACTTTTGTGT	GCCCTGCCCCGTATCCGGGCAG
Rbm9	AATCAGTAAACCAGGCATTCCT	CTTGCTATAGAGTCCTCTACTA
Rgs16	TGAACAGTAAAAATGGGGTGGC	AGCTGGTGACTTGAGGAAGCG
Scamp5	TTCCCACCATTGCCCAAATTCA	ATGGGCCGAAACCAGCAGACGT
Sema3	CCTCTTCAGGCTCTGTGTTCCG	ACTGTCCCGCGGTCTGTGCCCA
Spt1	AAACTTCTGGAACTGCTGATTC	AGTTAGCAATGAGAGAGAGGGA
Sv2c	CAATGCTAGGTGGCTCTATGGT	TCGTGTGTCAGGCAGGCACAG
	GGCTCCATGGTGCTCTCCGGGAT	GGGAATGGCTTTTGTCACGTCA
	TTAACGATGCTAGGAGGCTCCA	GCTTCAGAGTTTTAATGTATAG
Tcte3	GCAGATATTAAAGGACAGTCTT	GTATATATCATTTATTTTGAAGGA
Tmem16k	GAACGTACCTGGGAACCTTTGA	CAAGAGTGCATGCTCTACGGCC
Tmem24	CAGCGGAGCTACATTATGAGCA	ATGGGCACCTTGGAAACACCAG
Tmsb10	TACCATCAAGGGCATGATTAGG	AGTGCCCGGGTTTCACAGTGCA
	CTCTTGCTGCAGCAACGAGAGT	GGCAAACCGGTGAAATTTGGCA
Tnfrsf13c	AAGCAGCCTGGAGCCTGGGACAGCTCTG	TGAGTTCTTGCACACTGCCTCTA
Tomm40l	CAGCCATTTCCAGGTGGCTCAT	CCAGTGGACAGAGTACTTCCCA
Xmr	CTTATGGAAGTACAGAATCCAG	CTGGATTCTGTACTTCCATAAG

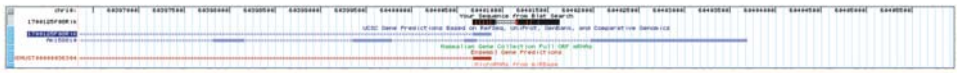
4.5 Übersicht der sequenzierten Stromaufwärtsbereiche

Die Sequenzierten Bereiche wurden mittels BLAT identifiziert (http://genome.ucsc.edu/cgi-bin/hgBlat) und durch den UCSC Genome Browser als schwarze Balken im Kontext der umgebenden Gene visualisiert.

1110017D15Rik(Smrp1)

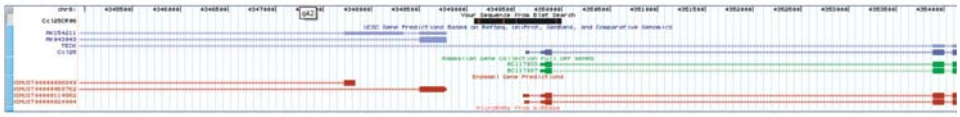


1700125f08



Cacng2

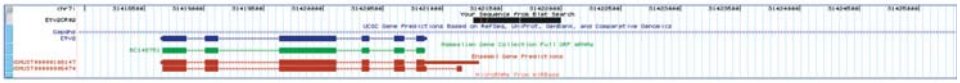
Ccl25



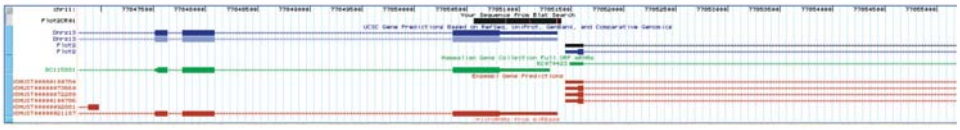
Cdk5

Etd

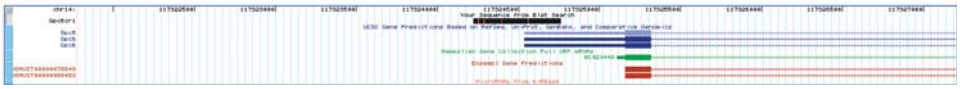
Etv2



Flot2



Gpc6



Hif1a

Kcnd2

Krt2

Mir16

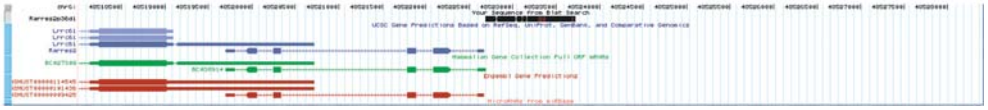
Nf1

PanX1

Ppt1

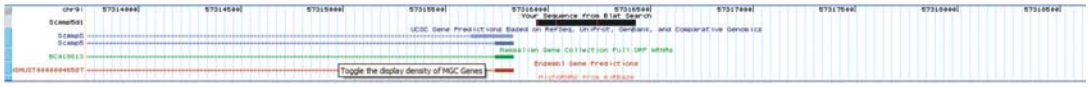
Rab4b

Rarres2

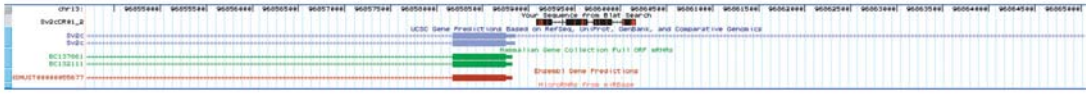


Rgs16

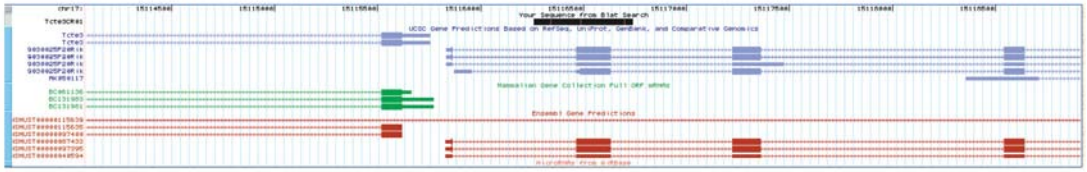
Scamp5



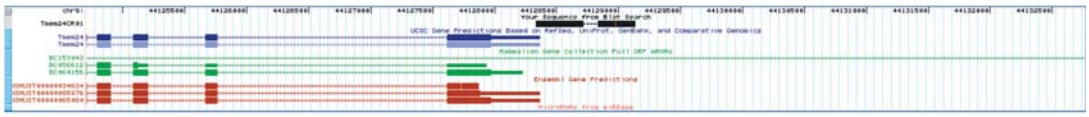
Sv2c



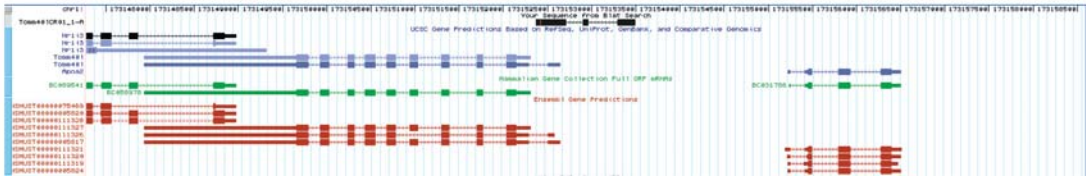
Tcte3



Tmem24



Tomm40l



4.6 Digitaler Anhang (Datenträger)

Sequenzdateien der Stromaufwärtsbereiche aus Kapitel 2 im .fasta Format im Ordner Stromaufwärts

Sequenzdateien der Umgebung des Poldi Gens im .fasta Format im Ordner PoldiUmgebung

Worddokument mit den alignierten Sequenzen der Exons und Spleißrelevanter Stellen (grau markiert) des Poldi Gens und von AK158810 in der Datei Sequenzanhang.doc

4.7 Erklärung

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

Köln, den

Fabian Staubach

4.8 Lebenslauf

Name:	Fabian Staubach
Anschrift:	Gilbachstr. 27 50672 Köln
Geburtstag, Geburtsort:	12.06.1981 in Bergisch Gladbach
Staatsangehörigkeit:	deutsch

Schulbildung:

1987 – 1991:	Kath. Grundschule Am Portzenacker, Köln
1991 – 2000:	Rheingymnasium, Köln

Studium:

2000 – 2005:	Studium der Biologie mit Abschlussziel Diplom. Schwerpunkte: Genetik, Entwicklungsbiologie und Physik. Diplomarbeit in der Arbeitsguppe von Professor Dr. Diethard Tautz Thema der Diplomarbeit: "Detektion von Genen unter dem Einfluß von Selektion inWildpopulationen der Hausmaus *Mus musculus* auf der Grundlage von DNA Polymorphismus- und Genexpressionsdaten"
2006:	Beginn der Promotion bei Professor Dr. Diethard Tautz am Institut für Genetik der Universität Köln
seit 2007:	Fortsetzung der Promotion am Max-Planck-Insitut für Evolutionsbiologie in Plön, Schleswig-Holstein
Mai 2009:	Voraussichtlicher Abschluß der Promotion an der Universität Köln

Köln, den Fabian Staubach