Evolutionary analyses of orphan genes in mouse lineages in the context of *de novo* gene birth

Dissertation

zur Erlangung des Doktorgrades

der Mathematisch-Naturwissenschaftlichen Fakultät

der Christian-Albrechts-Universität zu Kiel

vorgelegt von

Rafik Tarek Neme Garrido

Plön, April, 2013

Erstgutachter: Prof. Dr. Diethard Tautz

Zweitgutachter: Prof. Dr. Thomas C. G. Bosch

Tag der mündlichen Prüfung: 07.07.2014

Zum Druck genehmigt: 07.07.2014

gez. Prof. Dr. Wolfgang Duschl (Dekan)

Contents

Summary of the thesis

Gene birth is the process through which new genes appear. For a long time it was argued that the natural way of generating new genes was from copies of existing genes, and the possibility of *de novo* gene emergence was neglected. However, recent evidence has forced to reconsider old models and *de novo* gene birth gained recognition as a widespread phenomenon. *De novo* gene birth is the process by which a non-genic sequence is able to gain gene-like features through few mutations.

The following work is a compilation of analyses that seek to highlight the importance and prevalence of *de novo* gene birth in genomes, suggesting that this is a process that is present at all times and which becomes very relevant upon ecological shifts.

In the first chapter, I showed through phylostratigraphic analyses that new genes are substantially simpler than older, a trend which was consistent for several features and organisms, and suggestive of a frequent emergence of new genes through non-duplicative processes. In addition to this, I detected a strong association between gene birth and high transcriptional activity and chromosomal proximity. As part of this work, I was also able to use phylostratigraphy to evaluate a different model of gene birth, overprinting of alternative reading frames.

In the following chapters of this dissertation, I made use of high-throughput sequencing of transcriptomes and genomes to ask questions about the origin and change of genes at closer time divergences than ever before, ranging from nearly 3000 years to 10 million years of divergence. I was able to detect the theoretically predicted effects of short time scale comparisons on the rate of protein evolution. Also, I contribute evidence that genes of different ages show different selective constraints even after only a few thousand years of divergence.

Finally, in the last part of this thesis I evaluated the role of transcription in gene birth dynamics. Transcription seems to be a predominant feature of genomes, as most of the genome showed some level of transcription. In terms of *de novo* gene birth, I was able to identify 663 candidate loci from presence and absence of transcription. Analyses of these candidate loci indicated that gains are rather stable, meaning that subsequent losses were rarely found. In agreement with previous studies, I confirmed the role of testis as a driver of new genes.

These results indicate that transcription is not a limiting factor in the emergence of new genes, and that our knowledge about the key regulatory elements of transcription and their turnover is still limited to explain why new genes seem to arise at a higher rate than they decay.

Zusammenfassung der Dissertation

Gen-Geburt ist der Prozess, durch den neue Gene entstehen können. Für lange Zeit wurde gedacht, dass neue Gene nur aus Duplizierung und Anpassung von Kopien entstehen können. Die Möglichkeit *„de novo“*-Gene zu generieren wurde vernachlässigt. Allerdings haben die neuesten Beweise uns alle gezwungen, alte Modelle zu überdenken. Der „*de novo“*-Geburt der Gene hat mehr und mehr Anerkennung als weit verbreitetes Phänomen bekommen. „*De novo“*-Geburt der Gene ist der Prozess neue Gene zu generieren, durch wenige Mutationen von Sequenzen die vorher keine Geninformation hatten.

Die vorliegende Arbeit ist eine Zusammenstellung von Analysen, die die Bedeutung und Verbreitung von „*de novo“* Gen-Geburt in Genomen hervorheben. Damit will ich zeigen, dass dies ein Prozess ist, der zu allen Zeiten vorhanden und relevant ist, vor allem wenn ökologische Änderungen vorkommen.

Im ersten Kapitel habe ich durch phylostratigraphische Analysen gezeigt, dass neue Gene wesentlich einfacher als ältere sind. Das ist eine Tendenz, die für verschiedene Eigenschaften und Organismen konsistent ist. Das ist andeutend einer häufigen Entstehung neuer Gene durch Prozesse die nicht mit Gen-Duplikation zu tun haben. Weiterhin habe ich endeckt eine starke Verbindung zwischen Gen-Geburt, Transkriptionsaktivität und Chromosomen Nähe. In dieser Arbeit, war ich auch in der Lage Phylostratigraphie zu verwenden, um ein weiteres Modell der Gen-Geburt auszuwerten: „Aufdruck“ zwischen alternativen Leserahmen.

In den folgenden Kapiteln dieser Arbeit, versuche ich mit Transkriptom- und Genomsequezierung, Fragen zu beantworten über die Entstehung und Veränderung von Genen bei Divergenzen, die näherer als üblich in der Zeit liegen. Diese Divergenzen entsprechen zeiten zwischen ca. 3000 Jahre biz zu 10 Millionen Jahre. Damit war ich in der Lage, die theoretischen vorausberechneten Auswirkungen des kurzen Zeitskala auf der Evolution von Proteinen zu erkennen. Weiterhin habe ich gezeigt, dass Gene der verschiedenen Altersstufen unterschiedliche selektive Einschränkungen zeigen, schon nach einige tausend Jahre Divergenz.

Schließlich, im letzten Teil dieser Dissertation habe ich versucht die Rolle der Transkription in der Gen-Geburt Dynamik auszuwerten. Transkription scheint ein wesentliches Merkmal von Genomen zu sein, da die meisten des Genoms eine gewisse Transkription zeigen. In Bezug auf „*de novo*“Gen-Geburt, war ich in der Lage, 663 Kandidaten-Loci mit An- und Abwesenheit zu identifizieren. Analysen dieser Kandidaten-Loci zeigen, dass Transkriptionsgewinne sehr stabil

sind. Das bedeutet, dass spätere Verluste wurden nur selten gefunden. In Übereinstimmung mit anderen Studien habe ich die Rolle des Hodens als Betrieb für neue Gene bestätigt.

Diese Ergebnisse zeigen, dass die Transkription kein limitierender Faktor bei der Entstehung neuer Gene ist und dass unser Wissen über die Regulationselemente der Transkription und ihren Umsatz noch begrenzt ist, um zu erklären, warum neue Gene mit einer höheren Rate scheinen als sie zerfallen können.

Triumphs as well as failures of nature's past experiments appear to be contained in our genome

- Susumu Ohno, 1972

Acknowledgements

I am grateful to Prof. Dr. Diethard Tautz for the opportunity to be part of his department, for allowing me to dive in in this wonderful topic. I thank him for his support, patience and fruitful discussions, for being open to new ideas, and especially for granting me independence to develop these projects in my own way and at my own pace.

I thank Prof. Dr. Thomas Bosch and Prof. Dr. Arne Traulsen for being part of my Thesis Committee during the last years, and for discussions and guidance.

I am also grateful to Tomislav Domazet-Loso for allowing me to use his Phylostrat server in Croatia, and Robert Bakaric for his continuous help and very insightful discussions, some of which might or might not have been accompanied by drinks. I also thank Arne Nolte, Henrik Krehenwinkel, Till Czypionka, Philipp Rausch, Jun Wang, Frank Chan, Alexander Pozhitkov, Zeljka Pezer, Natascha Hasenkamp, Miriam Linnenbrink, Jarek Bryk, Ania Lorenc and Freddy Chain for suggestions, critics, hallway discussions and in general for helping me with all sorts of problems. I thank Bettina Harr and Meike Teschke for long distance support with their data and for allowing me to access their samples.

I am grateful to all those good friends and colleages which I fail to mention here but which have contributed to great times in Plön and in Kiel.

A large part of the work here presented would not have been possible without the support from Derk Wachsmuth, Werner Wegner and Herbert Kiesewetter from our IT department. I am deeply grateful for the moral and lab-related support from Barbara Kleinhenz, Heike Harre, Nicole Thomsen, Elke Blohm-Sievers, Sarah Lemke, Heinke Buhtz and Ellen McConnell.

I am also grateful to our Mouse Team in Plön, for taking such good care of the mice, and Christine Pfeifle for helping me coordinate all aspects related with my biological samples. I also thank Anja Schunke and Dörthe Thiele for helping us acquire *Mus mattheyi* specimens.

I thank Christian Becker, Janine Altmüller and their staff at the Cologne Center for Genomics for their help with my sequencing projects.

I thank the International Max Planck Research School for Evolutionary Biology, especially Kerstin Mehnert, for financial and moral support, and enrichment of my last experience as a student.

I thank Thomas Wiehe, David Karlin, Erich Bornberg-Bauer and Joanna Masel for welcoming me in their respective institutes as a guest, for a very pleasant time and nice discussions.

I thank Sebastian Meyer, Maria Thieser, Puspendu Sardar and Chen Ming for their efforts and contribution to our work together, and especially for their patience towards me as their supervisor.

I thank Weini Huang for her time, patience and interesting discussions.

I am among the very few people I know which have the fortune of having more than one family. I thank my father Rafik for his neverending support and confidence. I thank my mother Monica, her husband Hugo and my brother and sister, Lamia and Chaid, for their support and encourangement. I also thank my german parents, Silvia and Jürgen Noltze, and the ever growing Noltze family: Martin, Stephan, Nikki, Max, Tirza, Jonathan and Janosch; for welcoming me in their home, for making me feel part of the family and for granting me so many moments of joy among them.

I also thank Luisa Fernanda Pallares, Henrik Krehenwinkel, Joshka Kaufmann, Chen Ming and Weini Huang for commenting on early drafts and helping me improve this text.

Finally, I would like to thank Luisa Fernanda Pallares for being there for me and with me, for her continuous support, for helping me find time for vacations, and most of all for stimulating my mind and cheering me up through this process.

General introduction

Diversity seems to be an intrinsic property of life present at every possible hierarchical level (Mayr, 1982). The study of the diversity of genes and their functions has recurrently lead to the question of how new genes appear. For a very long time it has been argued that nature acts as a tinkerer regarding new gene generation (Jacob, 1977), and therefore all modern genes are derived forms from other genes through duplication (Ohno, 1970). Over the last two decades this perception has been challenged, leading to alternative models of gene emergence. Currently, there are at least three scenarios that can explain discrete increases in gene diversity, i.e. how completely new genes are born.

One of these scenarios is known as *de novo* gene birth, and it involves the generation of a gene through mutations from a non-genic sequence. This dissertation is an approximation to the emergence of new genes through non-duplicative mechanisms, mainly considering *de novo* gene birth.

De novo gene formation is a process that we are just beginning to understand, and that has been almost dogmatically neglected in favor of the idea that all genes come from other genes. The following is a brief recollection of the history behind the ideas and models of how new genes appear.

A brief historic perspective on the concepts of gene birth

Gene duplication is the main source of new genes

All molecular functions were thought for a long time to have evolved from a limited set of ancestral functional sequences (Chothia, 1992) and expanded continuously by gene duplication (Ohno, 1970). Innovation at this level was thought to follow a path much similar to those of new species, in a Darwinian sense. In a similar fashion that all modern organisms can be traced back to ancestral organisms, all genes should derive from other ancestral genes. The evolutionary process that best describes these ideas is known as gene duplication, and it is arguably one of the best understood phenomena in molecular evolution (Hahn, 2009; Innan and Kondrashov, 2010; Kaessmann et al., 2009).

Susumu Ohno is usually credited as the pioneer of gene duplication. However, ideas of duplications contributing to gene repertoires are almost as old as the field of genetics (reviewed comprehensively in Taylor and Raes, 2004).

During the early 20th century, plant and Drosophila cytogeneticists proposed correlations between morphological adaptations and the changes in chromosomal number, and imagined the potential advantage of increases of sets of genes, as opposed to the changes derived from variation involving whole chromosomes (Bridges, 1935).

Hermann Muller, through experiments in *Drosophila*, noticed that short chromosomal duplications (now commonly known as segmental duplications) were not always deleterious, and suggested that short chromosomal duplications could lead increase in the number of genes (Muller, 1935).

J.B.S. Haldane suggested that duplicated genes could be source of novelty. He proposed that duplications could have adaptive potential, given that duplicated genes would be more robust to mutations due to redundancy from both copies. At the same time, the accumulation of such mutations would allow duplicates to explore new sequences and functions much easier than non-duplicated genes (Haldane, 1932).

In 1951, some 20 years prior to Ohno, S.G. Stephens hypothesized that duplications have the potential for swift innovation, but also noticed that functional duplicates might be limited by their existing functions in their potential to completely innovate (Stephens, 1951). Stephens also mentioned the possibility of *de novo* gene birth, but unfortunately his ideas were too early to have found any experimental endorsement.

In 1970 Susumu Ohno compiled in his book “Evolution by gene duplication” comprehensive and convincing cases highlighting the evolutionary relevance of gene duplicates. In his postulates genes are discrete entities which are able to grow in number by errors in replication, which result in partial or complete copies of a gene, several genes or even whole genomes (Ohno, 1970).

Having one or more duplicated genes might be detrimental if the concentration of the gene product is tightly regulated, or if the excess of a gene (or a part of the gene) prompts competition between other interacting genes. However, there are biochemical reactions and pathways in which this balance is not relevant or is not altered, as in the case of whole genome duplications. Ohno argued that duplicated genes have three possible adaptive fates: (1) copies are maintained (‘more of the same’) if having more available concentration of a gene product is advantageous; (2) each copy evolves different (and maybe even complementary) versions of the ancestral gene (‘transformation into isozymes’) if balancing selection would profit from a wide array of similar genes or if the ancestral gene would have dual functions under adaptive

conflict; or (3) one of the copies evolves a complete new function ('creation of a new gene') by initial neutral degeneration and further exploration of the sequence space under adaptive conditions.

Subsequent research has identified that main mechanisms leading to gene duplication (reviewed in Zhang, 2003) are duplicative transpositions (DNA transposition) (Huang et al., 2012), retrotranspositions (transposition through RNA intermediates) (Kaessmann et al., 2009), segmental duplications (tandem and scattered through the genome) (Marques-Bonet et al., 2009) and whole-genome duplications (also known as polyploidization) (Wolfe and Shields, 1997). From here on, whenever gene duplication is mentioned, for the sake of inclusion, it should be understood as a generic concept that includes all these processes that result in increased number of parts of genes, complete genes or groups of genes (as reviewed in Long et al., 2003).

These ideas were assimilated by the community along with the steady accumulation of evidence. The majority of sequence information gathered prior to the genomics era indicated that most genes had paralogs (Chothia, 1992), which is direct evidence of gene duplication events. This led to the development of population genetics models to explain the dynamics than can lead to fixation of duplicates under different scenarios (Hahn, 2009; Innan and Kondrashov, 2010). Many of these models have even been tested and confirmed experimentally (Näsvall et al., 2012).

From a general perspective, all these models have the common underlying idea that the generation of a new gene would always require a genic template. This assumes that genes are highly organized structures which can only degenerate or specialize. This has led to more than 40 years of an almost dogmatic view of gene birth focused on duplication of existing genes, with the alternative idea of *de novo* evolution from random sequences being overly neglected on the basis of its presumed improbability.

Orphan genes and the genomics era

The first large scale sequencing efforts, which were less focused on specific genes than before and more on long stretches of unknown DNA (Oliver et al., 1992), yielded interesting results in terms of hidden catalogs of genes (Dujon, 1996). A great proportion of the predicted open reading frames (ORFs) lacked any homology to other sequences available at that time, and therefore could not be classified into gene families. These genes were named orphan ORFs

(Dujon, 1996) or ORFans (Fischer and Eisenberg, 1999) because of their lack of evident 'family ties' to other genes. Initially considered prediction artifacts from annotation software or artifacts due to low phylogenetic sampling, orphans were neglected which would lead to few experimental projects to understand them (Fischer, 1999).

However, as many more genomes became available, it was not too long until it was realized that the number of orphan genes increases linearly with each new genome sequenced, while the number of genes common to most species quickly reaches saturation (Wilson et al., 2005). This provided strong support to the notion that orphan genes are biologically relevant. Orphan genes are lineage- or taxon-specific, indicating that these genes appeared after a given split in the phylogenetic history of a lineage (Khalturin et al., 2009; Tautz and Domazet- Loso, 2011).

Initial explanations were borrowed from duplication models, and assumed orphan genes would derive from fast evolving duplicates: a combination of neutral and adaptive processes could follow after a duplication event, generating an elevated number of mutations upon one of the duplicates in a short time-span, and resulting into one of the copies having lost all detectable homology to other members of its gene family (Domazet-Loso and Tautz, 2003; Tautz and Domazet- Loso, 2011) (Figure 1). The dynamics of the neutral and adaptive processes could be coupled to the exploration of new functional parts of the sequence space, and once the function became relevant for the organism, the new orphan gene would slow down its rate of mutation and would be only present in those organisms which descend from the lineage that underwent this specific process (Domazet- Loso et al., 2007).

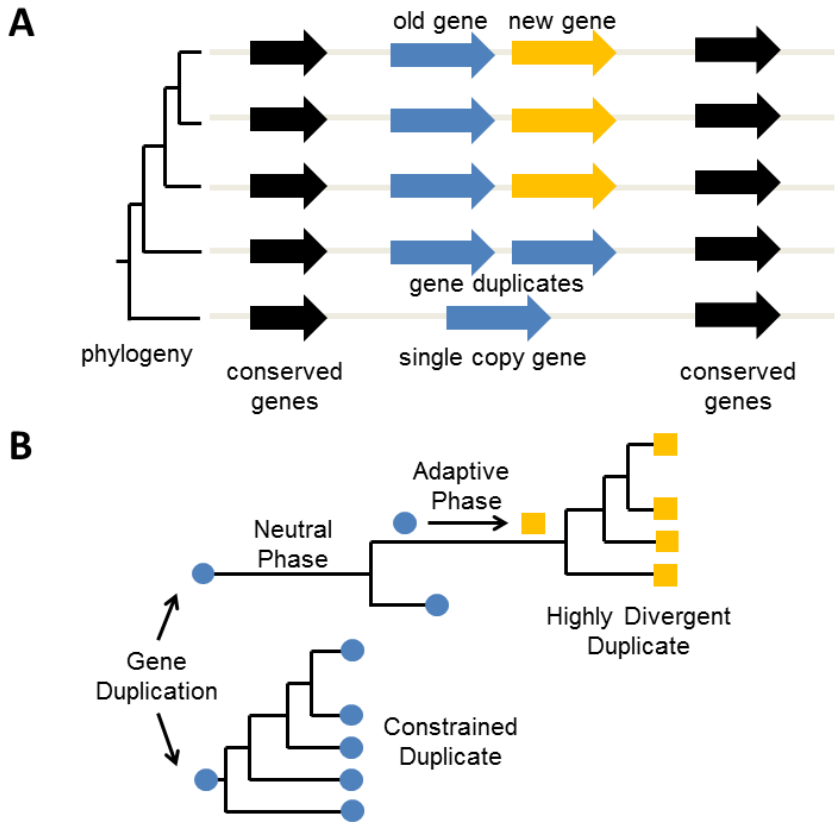

Figure 1. Duplication and divergence can lead to completely new genes (modified from Tautz and Domazet- Loso, 2011).

A. Comparative genomics setup to identify highly divergent duplicates. Genomic alignments, guided by well-supported phylogenetic information, are used to define the ancestral and derived states, i.e. phylogenetic divergence before and after the gain of a new gene upon duplication. Conserved collinear genes (in black) can be used to identify synteny, to ensure that no large scale rearrangements have occurred. B. Duplication with fast divergence model (from Domazet-Loso and Tautz, 2003), in which a duplicated copy is kept constrained, while the other copy drifts, and eventually undergoes an adaptive phase, yielding a different protein. The combination of both drift and adaptive phases should account for a loss of similarity in the derived duplicate.

Phylostratigraphy and the continuous emergence of new genes

The idea of new genes appearing relatively fast and slowing down as they became functional led to the formal development of phylostratigraphy, a method of estimating the phylogenetic age of a gene (Domazet- Loso et al., 2007).

Phylostratigraphy relies on having a well-described phylogeny for a focal organism and a large number of reference sequences for representative outgroups. This information is used to give each gene a phylogenetic age, according to their lineage-specific character. For example, in the mouse, eukaryotic-specific genes are older than vertebrates-specific genes, and these in turn

are older than rodent-specific genes. Each divergence represents an epoch, or phylostratum, at which novelties have appeared and accumulated (Domazet- Loso et al., 2007). In theory, at each divergence time should be possible to identify new genes, but in practice this is determined by the phylogenetic coverage of each divergence.

Following the assumptions of the phylostratigraphic approach, organisms are able to accumulate duplicates at steady rates, and new genes are predominantly associated periods of adaptive transitions, such as speciation or radiation events (Tautz and Domazet- Loso, 2011). The phylostratigraphic approach has been successful in linking genes to large scale evolutionary questions such as the origin of genes associated with multicellularity and cancer (Domazet- Loso and Tautz, 2010a), genetic diseases (Domazet- Loso and Tautz, 2008), germ layers (Domazet- Loso et al., 2007) and the correlation between ontogenetic gene-expression and phylogenetic origin (Domazet- Loso and Tautz, 2010b).

Interestingly, these analyses always yield high number of orphans after the last divergence, compared to relatively low numbers from divergences immediately preceding. The last divergence usually contains species-specific orphans (Tautz and Domazet- Loso, 2011). This recurrent phenomenon has been interpreted as high levels of gene birth followed by high levels of losses, as most new genes would fail to find a suitable function that enables their future conservation (Palmieri et al., 2014; Tautz et al., 2013).

Not all genes come from other genes

Nevertheless, phylostratigraphy is unable to distinguish different types of origin of genes. Any innovation acquired by any mechanism resulting in a new gene (except maybe those of horizontal gene transfer, which would be misplaced as older events) can be detected using similar approaches (Toll-Riera et al., 2009; Wissler et al., 2013).

Many authors have pointed out that across many levels of organization nature works mainly by tinkering; copying and modifying functional components to match the necessities an ever changing environment poses (Bornberg-Bauer et al., 2010; Bridgham et al., 2010; Di Roberto and Peisajovich, 2014).

One of the main proponents of the ‘nature as tinkerer’ idea, Francois Jacob, stated in 1977 that the generation of a functional protein by random association of nucleotides was “practically impossible” (Jacob, 1977). This is a common argument in favor of duplication-like mechanisms for generation of new genes over other possibilities, based on the high complexity of a

functional protein and on the seeming lack of information contained a random sequence of a similar length.

But this argument is misleading for a number of reasons. The first being that duplications do indeed occur at a higher frequency than the number of mutations needed for the generation of a new gene, but most duplications will not actively contribute to protein sequence and fold innovation, as they are self-limited in their exploratory properties (Ohno, 1984). The second reason is that only a few regulatory changes are needed for the stable expression and translation of any nucleotide sequence, and while that does not always constitute a functional protein, these sequences have far more potential to become a completely new gene that any given duplicate (Carvunis et al., 2012; Heinen et al., 2009; Wilson and Masel, 2011). The third reason is not theoretical, but practical, namely the general observation of new genes through non-duplicative processes appearing in almost all organisms in which this has been queried (Begun et al., 2007; Cai et al., 2008; Carvunis et al., 2012; Chen et al., 2007; Delaye et al., 2008; Donoghue et al., 2011; Guerzoni and McLysaght, 2011; Heinen et al., 2009; Khalturin et al., 2008; Knowles and McLysaght, 2009; Levine et al., 2006; Li et al., 2010b; Neme and Tautz, 2013; Reinhardt et al., 2013; Sabath et al., 2012; Toll-Riera et al., 2009; Yang and Huang, 2011). Nonetheless, the argument by Jacob and many others after him has prevailed for a long time (Lander, 2011).

Many genome-scale studies have provided lists of candidate genes which have been used to understand general properties of new genes, but also to dissect their functions and origins (Cai et al., 2008; Chen et al., 2007; Heinen et al., 2009; Reinhardt et al., 2013). Comparative genomics and functional studies have shown that duplication-divergence could be a conservative way of generating new proteins, albeit not the only one. Analyses from *Drosophila* (Begun et al., 2007; Levine et al., 2006; Palmieri et al., 2014; Zhou et al., 2008), yeast (Cai et al., 2008; Carvunis et al., 2012; Li et al., 2010a), mouse (Heinen et al., 2009; Neme and Tautz, 2013), *Plasmodium* (Yang and Huang, 2011), plants (Donoghue et al., 2011) and primates (Knowles and McLysaght, 2009; Li et al., 2010b; Toll-Riera et al., 2009; Wu et al., 2011) indicate that *de novo* emergence of new genes is not only likely buy rather frequent. *De novo* – in this context, “from scratch” – is the acquisition of a gene from a region which previously lacked any genic information (Figure 2). In this model, non-genic regions of the genome are able to progressively gain motifs that enable stable transcription and association with ribosomes, thus forming protogenes: entities generated as a byproduct of widespread transcription and

translation in a genome (Siepel, 2009), which can act as precursors of new genes from non-genic regions (Carvunis et al., 2012).

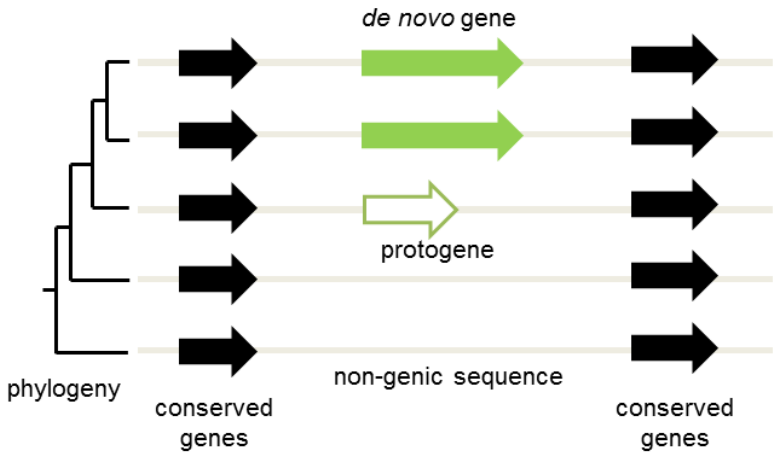

Figure 2. *De novo* gene emergence from an intergenic region (modified from Tautz et al., 2013).

Illustration of a comparative genomics setup used to infer *de novo* gene birth. Genomic alignments in a phylogenetic context aid in the identification of ancestral non-genic sequences with orthologous genes or protogenes present in sister taxa.

Considering gene birth from molecular and evolutionary perspectives

The concept of the gene is central to biology and is strongly dependent on the available tools to study genes, genomes and their phenotypes. A gene is often described as a segment of DNA composed of overlapping regulatory and transcribed regions which result in a functional product, but also as the fundamental unit of heredity, which is responsible for the transfer of traits from parent to offspring (Gerstein et al., 2007).

The molecular concept corresponds to the state-of-the-art techniques for the analyses of functional elements in genomes (ENCODE Project Consortium, 2011; ENCODE Project Consortium et al., 2007; Gerstein et al., 2007) while the hereditary one is almost the original definition of a gene that started with the field of genetics. Both definitions reflect more than a century of evolution of biological thought, as well as the technological leaps across this time. While it is not uncommon to see these two concepts together, there is usually a sharp distinction between their molecular and evolutionary implications.

One issue with both definitions is that the same name is used for the individual molecular entity and the large-scale evolutionary entity. At the molecular level, each individual gene is composed of nucleotides, exists in a genome, is transcribed upon stimuli into messenger RNA (mRNA), is

usually subjected to modifications such as splicing, and in most cases is further translated into a protein product. The products derived from that specific locus perform a specific task, which influence the performance of an organism in its environment. On the other side, at the evolutionary level, the gene is a large-scale entity – a statistical conglomerate – composed of all those individual genes in each of their individual carriers and the impact their common molecular task has across multiple scales of organization on the fitness of the species.

Our current knowledge about genome architecture indicates that most of the sequence patterns responsible for the transcription, translation, splicing and stability of a messenger RNA (mRNA) in the cell are rather short motifs (Beaudoing et al., 2000; Brooks et al., 2011; D'haeseleer, 2006; Stanke et al., 2008). Random combinations of such sites can be expected at any given point in a genome, and local sequence patterns can be found to match those of seemingly complex regulatory regions (Heinen et al., 2009). Provided a sufficiently large genome, at each time there should be sequences which exist as protogenes, and sequences which are a few mutational steps away from protogenes.

Interestingly, until the beginning of the genomic era, the definitions of a gene assumed that a gene is a discrete entity, as opposed to more contemporary definitions which also consider the context in which a gene is embedded (Gerstein et al., 2007). The discrete view imposes a limitation since it is indeed very unlikely that a region that does not contain any useful genic information suddenly gains transcription, translation and function in one step.

Genome-centered definitions of the gene have a major advantage regarding our conceptual understanding of how genes can appear *de novo*. Genic potential can be assigned to every region in the genome according to complementary pieces of information contained within each region (Stanke et al., 2008). In this respect, one can imagine a dynamic continuum of genic potential, with mutations inducing shifts of such potential over time, and some regions being able to develop into *de novo* genes (Heinen et al., 2009; Reinhardt et al., 2013).

Overprinting: true innovation from existing genes

The way proteins are translated from nucleotide sequences presents the opportunity for more than one reading frames, i.e. different peptide sequences within a single nucleotide sequence. This property has been recognized for a long time, and has ever since played a role in many instances of protein sequence prediction (Doolittle, 1986).

In most cases, only one of the possible frames is functional, but it is a common feature of organisms with small genome sizes to have overlapping or overprinted reading frames (Chung et al., 2007; Krakauer, 2000; Liang and Landweber, 2006; Rancurel et al., 2009; Vanderperre et al., 2013).

Alternative reading frames can be exposed by mutations, e.g. frameshifts, which could lead to deleterious phenotypes most likely derived from loss of function via an early stop codon. These events are usually targeted by RNA surveillance mechanisms (Hu and Ng, 2012). It is less known if the alternative frame on its own would be able to induce a negative effect (i.e. toxicity).

Susumu Ohno also recognized the duplication of genes might not always represent true innovation, but that the overprinting of reading frames would be able to generate completely new sequences, as the multiple reading frames in a protein do not resemble one another (Ohno, 1984). To this day, there is a substantial body of evidence that new genes in viruses and bacteria can be generated by overprinting of reading frames (Delaye et al., 2008; Sabath et al., 2012). There are also well-known examples from eukaryotes (Chung et al., 2007; Nekrutenko et al., 2005; Sherr, 2006).

In the case of gene birth through overprinting (Figure 3), alternative reading frames act as random peptides, available as cellular byproducts of a given transcribed gene. The examples of alternative reading frames which have acquired a function are strong evidence to suggest that random sequences can indeed become functional.

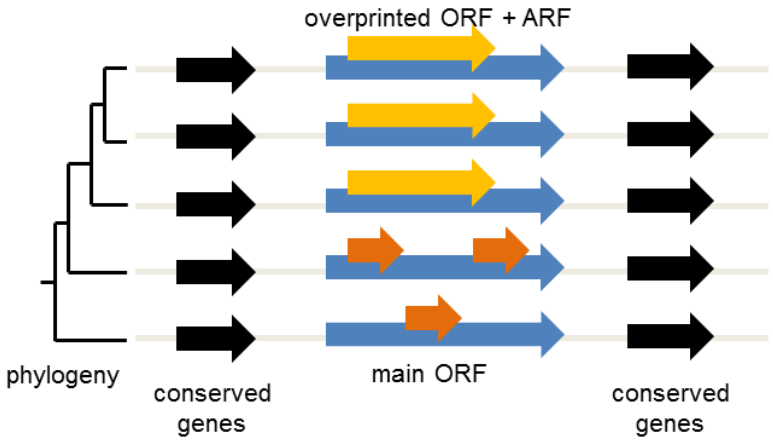

Figure 3. Overprinting of alternative reading frames can generate protein innovation

Comparitive genomics setup used to infer the evolutionary history of overprinted genes. Genomic alignments in a phylogenetic context are used to identify ortholog groups. Information from reading frames is evaluated to identify when an alternative reading frame (ARF) was first present. In the example, the ancestral contains the main open reading frame (ORF) and several scattered ARFs. Over time, these ARFs appear and disappear stochastically. Once an ARF emerges and has adaptive potential, it is possible for the two reading frames to coexist. Due to the properties of the genetic code, both frames should be completely unrelated in terms of protein sequence.

The life cycle of genes

As mentioned above, there are three distinct models of how new genes can appear, based mainly on comparative genomics and case-studies: duplication-divergence (Figure 1), *de novo* (Figure 2) and overprinting (Figure 3). It is likely that all three scenarios occur simultaneously in genomes, although formal comparisons to understand their relative impact have not been reported yet.

In addition to these mechanisms of new gene generation, it is possible to imagine a conceptual life cycle for genes (Figure 4), by further expanding on the metaphor of new gene emergence as gene birth. In this cycle, genes are born as protogenes from non-genic sequences (Carvunis et al., 2012), reproduce by gene duplication-like mechanisms (segmental or whole genome duplications, retrotranspositions, gene fusion or fission, horizontal gene transfer) (Long et al., 2003; Ohno, 1970; Zhang, 2003), and die when they become pseudogenes (Demuth and Hahn, 2009), which further decay into new non-genic sequences.

Pseudogenes are defective relatives of genes (Vanin, 1985) that arise when the selective pressures that maintain a gene are removed. Pseudogenes also arise frequently from gene duplications, when the selective pressures on the new copies are relaxed (Wagner, 1998). In

most cases pseudogenes are not functional due to loss of transcription or translation, which further promote their decay.

It is expected that many protogenes do not develop fully as genes, therefore decaying before even becoming functional. Likewise, there are known examples in which pseudogenized genes can become functional again (Bekpen et al., 2009).

It is important to highlight that this cycle can be further partitioned between a stochastic stage, which includes the emergence of protogenes on one side and the decay of pseudogenes on the opposite; and an adaptive stage, which deals with the acquisition of new functions by random sequences and function maintenance in a genome. The selective pressures from the environment are crucial in the preservation of a function, and it can be assumed that once a selective pressure is removed, the associated gene or genes can decay (Near et al., 2006).

Furthermore, this cycle is a description of the possible dynamics between different conceptual entities existing in genomes. This does not mean that is a process that every gene undergoes.

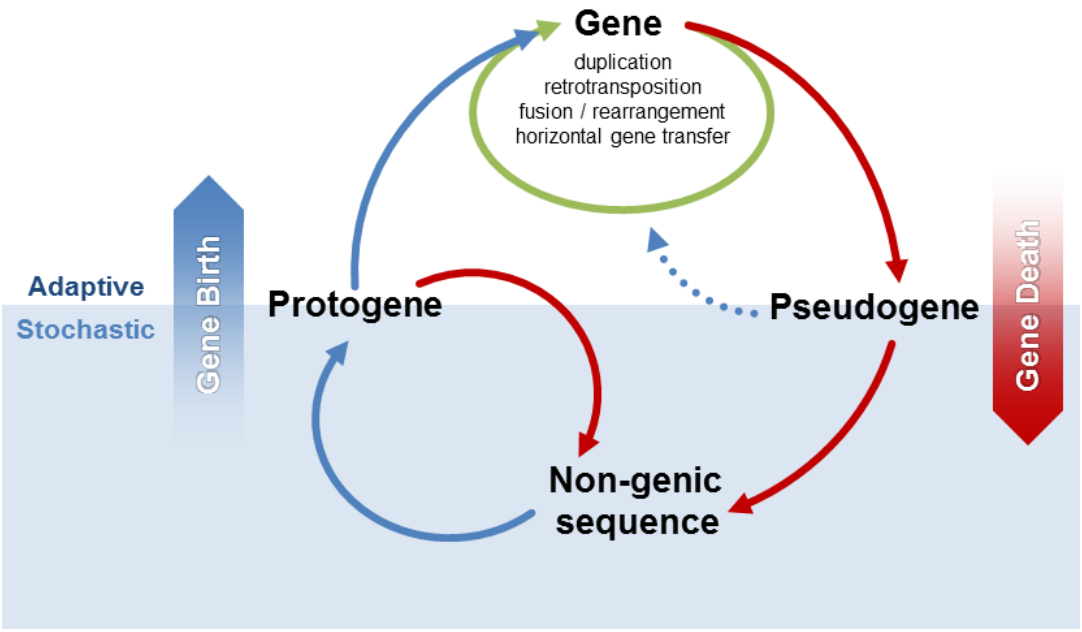

Figure 4. The life cycle of genes (from Neme and Tautz, 2014).

Blue arrows represent transitions which lead, either partially or completely, to a new gene, and are therefore dubbed processes of gene birth. Red arrows represent the loss of features which result in the degradation of the genic potential of a sequence. Green arrows represent the processes which increase the gene repertoire from existing genes. Raw material for genes is stochastically generated as protogenes (Carvunis et al., 2012), entities that have gene-like properties (i.e. stable expression or translation), but may still lack a proper function. Once a protogene is able to perform a function that has an adaptive advantage, it will become fixed in a population. Gene loss through pseudogenization can lead to the death of a gene in a lineage, when the selective pressure upon it is released (Near et al., 2006).

This will also be the case for protogenes which have not fully developed into genes or for very young genes (Carvunis et al., 2012; Palmieri et al., 2014).

Overview

The work presented in this thesis is a genome-scale effort towards understanding the dynamics of gene birth. My foremost aim was to understand general features underlying gene birth, which could be used to further enrich the current ideas towards more robust generalizations.

In the first chapter, I present phylostratigraphic analyses of four vertebrate genomes: mouse, human, zebrafish and stickleback. Here I reveal several trends associated with the emergence of new genes. For instance, younger genes exhibit less complexity in genomic features, such as lengths or number of exons and domains, and many new genes show significant association to bidirectional promoters. In addition to this, I present how overprinting can be detected using phylostratigraphic methods, and show examples of recent gains of overprinting events in the mouse and other related taxa.

The second chapter is a general introduction to the datasets generated during the course of this thesis, which were used for the analyses described in the third and fourth chapters. I generated comprehensive sets of testis, liver and brain transcriptomes from closely related mouse populations (German, French and Iranian *Mus musculus domesticus*; Czech and Austrian *Mus musculus musculus*), subspecies (*M. m. domesticus*, *M. m. musculus* and *M. m. castaneus*) and species (*M. spicilegus*, *M. spretus*, *M. mattheyi* and *Apodemus uralensis*), which cover time divergences between 3,000 to 10 million years. In addition to this, I also generated whole genome sequences from *M. spicilegus*, *M. mattheyi* and *Apodemus uralensis*, which serve as support for the transcriptomic analyses. These datasets constitute also very powerful reference material for future research of other processes related to short evolutionary time dynamics of mammalian biology.

In the third chapter, combining information derived from polymorphisms in closely related species with annotations from the reference mouse genome, I evaluated how the evolutionary rate of a gene is influenced by its phylogenetic age and how younger genes exhibit greater reading frame instability than older genes, consistent with a model of frequent loss at the protein-coding level.

In the fourth chapter, I assessed how genome-wide transcription changes over the sampled phylogeny, and show that gains at the transcription level are much more common than losses.

Here I also developed an algorithm to focally detect *de novo* gene gains, revealing that *de novo* transcripts are gained at a steady rate over time, and are predominantly transcribed in the testis. Taken together, these results indicate that pervasive transcription along the genome is able to steadily provide material for the generation of new genes.

Chapter 1: Phylogenetic patterns of emergence of new genes support a model of frequent *de novo* evolution

Introduction

The hallmark of the signature of a new gene (or orphan gene) is that it arises at some time within the evolutionary lineage towards an extant organism and has no similarity with genes in organisms that have split before this time (Domazet-Loso and Tautz, 2003; Khalturin et al., 2009; Tautz and Domazet- Loso, 2011). This distinguishes orphan genes from genes that arise through full or partial duplication processes to form paralogous genes or gene families (Kaessmann, 2010; Zhang, 2003). It has been proposed that orphan genes are likely to play a major role in lineage specific adaptations (Cai and Petrov, 2010; Domazet-Loso and Tautz, 2003; Khalturin et al., 2009; Tautz and Domazet- Loso, 2011) and thus contribute to evolutionary innovations. There are two major models of how orphan genes can arise (Tautz and Domazet- Loso, 2011). The first is the duplication-divergence model, which assumes that they emerge through an initial duplication of other genes, but this is followed by rapid divergence, such that all similarity to the parent gene is lost (Domazet-Loso and Tautz, 2003). The alternative is the *de novo* evolution model, which assumes that genes can directly arise out of non-coding DNA (Siepel, 2009). Although this second possibility seemed initially rather unlikely, such genes have been found in *Drosophila* (Begun et al., 2007; Levine et al., 2006; Zhou et al., 2008), yeast (Cai et al., 2008; Li et al., 2010a), mouse (Heinen et al., 2009), Plasmodium (Yang and Huang, 2011) plants (Donoghue et al., 2011) and humans (Knowles and McLysaght, 2009; Li et al., 2010b; Wu et al., 2011). In fact, there is now increasing evidence that *de novo* evolution may be rather frequent. Studies in yeast have suggested that a large number of transcripts without annotation are actively transcribed and translated (Carvunis et al., 2012; Wilson and Masel, 2011) and that such transcripts could be a source for *de novo* gene emergence (called "proto-genes") (Carvunis et al., 2012; Siepel, 2009).

We have developed phylostratigraphy as a method that identifies the genes that have arisen at each stage of a series of phylogenetically relevant splitting events (Domazet- Loso et al., 2007). This allows to systematically study the characteristics of such genes over time (Domazet- Loso and Tautz, 2008, 2010a, 2010b; Quint et al., 2012). Using this approach we found that gene emergence rates are particularly high in the youngest lineages, implying a very active process of *de novo* evolution, since the times considered for these youngest lineages are too short for the duplication-divergence model to apply (Tautz and Domazet- Loso, 2011). This is in agreement with the proto-gene concept, where non-coding transcripts are considered as

possible sources of new genes (Carvunis et al., 2012; Wilson and Masel, 2011). However, a study of emergence trends across the whole phylogeny is still missing.

In the present paper we use the mouse as a focal species, which has a particularly well annotated genome. We show that it is indeed possible to derive distinctive patterns for gene emergence, which appear to be generally in accordance with a *de novo* evolution model. As a special case of *de novo* evolution, we revisit the possibility that existing genes have developed an independent second reading frame. Evolution of new genes within such double reading frame arrangements have been known since some time (Keese and Gibbs, 1992; Ohno, 1984) (called "overprinting" by (Keese and Gibbs, 1992)). They have been well studied in viruses (Rancurel et al., 2009; Sabath et al., 2012), but several examples are also known from eukaryotes and have been studied in detail for some genes (Klemke et al., 2001; Nekrutenko et al., 2005; Sherr, 2006). Chung et al. (Chung et al., 2007) provided a first systematic approach to identify such alternative reading frames (ARFs) in mammals and suggested 40 candidate genes which appeared to use ARFs. We find here that it is indeed possible to retrieve even among annotated genes additional cases of overprinting, where the alternative reading frame maps to a different phylostratum than the original reading frame. This suggests that existing genes may readily become templates for *de novo* evolution of new gene functions within them, further supporting the notion that *de novo* evolution of gene functions are possible.

Results

The duplication-divergence versus the *de novo* evolution model for orphan gene emergence make some different predictions with respect to gene emergence over time, for example on length distributions and exon distributions, as detailed below. Apart of looking for such differential predictions, it is also of interest to assess general patterns, such as orphan gene distribution across the genome, as well as the emergence of associated promotors. Below, we describe first how we assign the genes to different age classes and then use this assignment to study gene emergence trends and patterns.

Phylostratigraphy of mouse genes

The phylostratigraphic approach was used to estimate the time of emergence of each of 20,775 annotated protein coding loci in the mouse genome (Figure 1.1). Twenty phylogenetic classes or phylostrata were defined according to consensus phylogenetic relationships between groups with enough available protein sequence information. The first phylostratum (ps1) represents the basis of all cellular life, i.e. the oldest genes, while the last phylostratum (ps20) represents the

lineage leading to mouse since the split from rat. blastp was used to assign for each mouse gene its presumptive origin within this phylostratigraphy. For this we use an e-value cutoff of < 10^{-3}, which has previously been found to provide an optimal compromise between sensitivity and accuracy (Alba and Castresana, 2007; Domazet-Loso and Tautz, 2003). The results of the assignment to the respective phylostrata are listed in Appendix A and summarized in Figure 1.1.

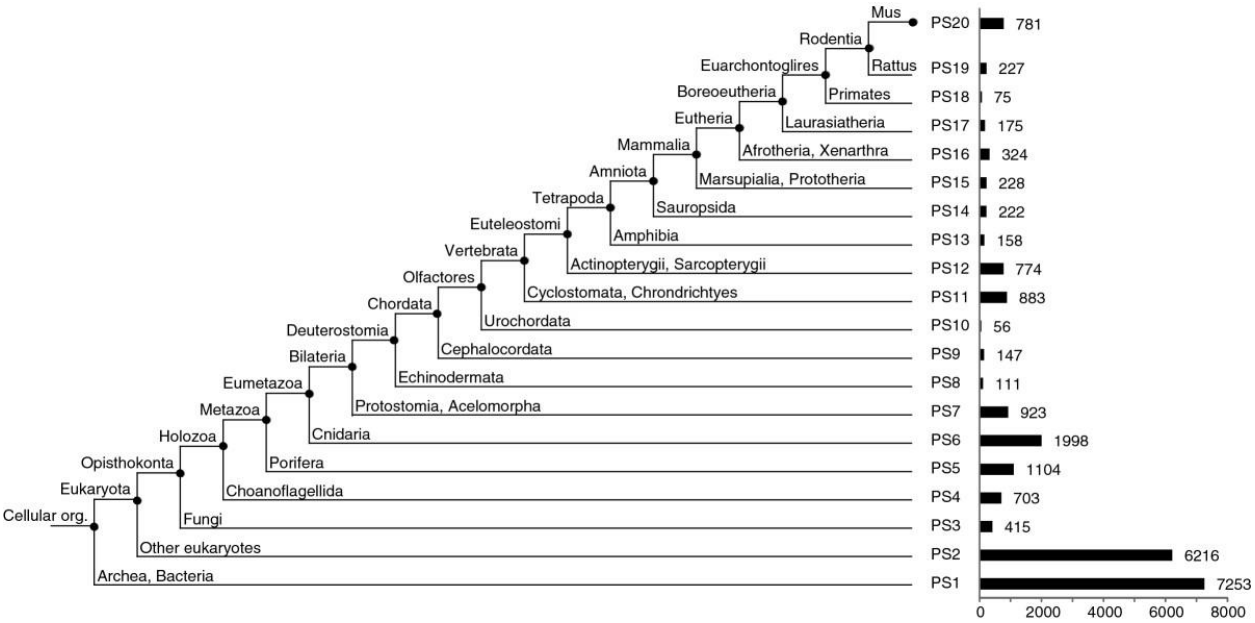

Figure 1.1. Phylostratigraphy of the mouse genome.

Each phylostratum corresponds to a node in the phylogenetic tree of the species. Representative outgroups are named under each node. The bar graph to the right represents the number of annotated protein-coding genes mapped to the respective phylostratum at a blastp threshold of e < 10^{-3}.

Approximately 60% of the annotated protein coding genes in the mouse genome originate from prokaryotic and basal eukaryotic ancestors (ps1-2). The rest of the genes have emerged later in the phylogenetic history, with peaks correlating to large scale biological transitions. For example, the peak around ps6 represents the single-cell to multicellular organism transition (Domazet- Loso and Tautz, 2010a) and the peak around ps11-12 represents the invertebrate to vertebrate transition. Another peak is evident at ps20, representing all genes that have evolved since the rat/mouse split. Although this may partly be ascribed to annotation problems within the youngest group of genes (Siepel, 2009) many of them are likely to represent *de novo* evolved genes, since mouse and rat are so close to each other that any duplicated gene would easily be traceable, even if it would evolve with the rate of a non-functional pseudogene.

Genomic features across ages

We used the phylostratigraphic assignment of the genes to assess the emergence trends over time for several relevant gene features (Figure 1.2). Some of the gene features were selected to allow to distinguish the duplication-divergence model from the *de novo* model.

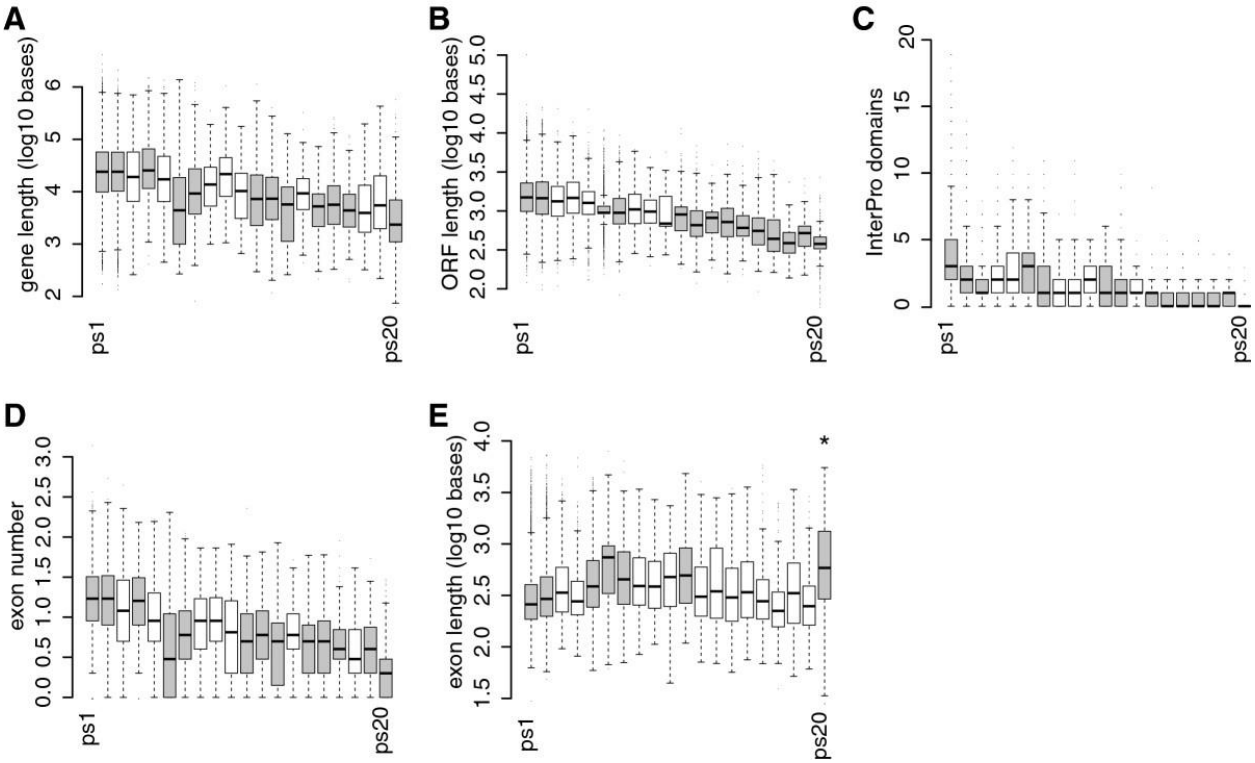

Figure 1.2. Features of genes for different phylogenetic age groups in the mouse.

A. Gene length distributions (includes exons and introns); B. ORF length distributions; C. Median number of InterPro domains per gene; D. Median exon numbers per gene; E. Median exon lengths per gene. Box-whisker plots around median values (bars) with quartile ranges and outliers as dots. Significant ($p < 0.01$) distribution differences were found for ps20 (marked with *) in E (t-test). Gray bars indicate phylostrata with non-randomly distributed values for each variable, based on permutations (n = 1,000) and Kolmogorov-Smirnov tests ($p<0.01$).

Table 1.1. Features of genes are strongly correlated with age.

Spearman´s ρ rank correlation coefficients across phylostrata calculated for the means of the respective distributions (compare Figures 1.2 and 1.3). All are significant at $p < 0.01$.

	mouse	human	stickleback	zebrafish
gene length	- 0.88	- 0.90	- 0.82	- 0.93
ORF length	- 0.98	- 0.96	- 0.98	- 0.97
domain number	- 0.94	- 0.91	- 0.72	- 0.90
exon number	- 0.93	- 0.96	- 0.94	- 0.94

With respect to gene length, the *de novo* model would predict that younger genes should be shorter than older genes, since it is unlikely that complex protein sequences emerge *de novo*. Rather one would expect that they could increase in size over evolutionary time. In the duplication-divergence model one would not expect a length dependence over time, since long and short genes should be equally likely subject to duplication at any time level. The results show, however, a strong length-dependence over time, both with respect to gene length (Figure 1.2A) as well as open reading frame length (Figure 1.2B). The Spearman rank correlations across the 20 phylostrata are very high (Table 1.1) suggesting an almost continuous trend over time. Such trends for gene length distributions had also previously been noted in analyses using fewer age classes (Lipman et al., 2002; Wolf et al., 2009).

A differential prediction can also be made for the expected correlation with protein domain emergence. *De novo* evolved proteins will initially have no domains which are shared with other genes, while duplicated genes would tend to retain domains of their parental genes (Chothia and Gough, 2009). Hence, the *de novo* evolution would predict domain gain over time, while no distinct pattern is expected for the duplication-divergence model. Again we find indeed a strong time-dependence with a continuous trend for domain emergence (Figure 1.2C; Table 1.1), supporting the *de novo* model.

De novo emerged genes should also have initially fewer exons, but could be expected to accumulate additional ones over time. In the duplication-divergence model, on the other hand, one would not expect a time dependency of exon numbers, since this mechanism should work the same at every time horizon. However, we find a strong trend of exon gain over time (Figure 1.2D; Table 1.1), supporting the *de novo* model.

Average exon length, on the other hand, shows no clear age-dependence (Figure 1.2E). Only the youngest genes (ps20) have significantly longer exons (Figure 1.2E) suggesting a fast secondary acquisition of introns after gene emergence, or gene fusion effects (Buljan et al., 2010).

To assess whether these patterns constitute general trends that can be observed in other lineages as well, we have also analyzed them for humans, stickleback and zebrafish lineages. Humans were included since the genome is equally well annotated as the mouse genome, the fish species represent another vertebrate lineage split more than 400 million years ago. Analysis of these three genomes confirms indeed almost all trends with similarly high correlation coefficients (Figure 1.3; Table 1.1). Gene length, ORF length, domain numbers and exon numbers show all a clear time-dependence. Only one comparison, namely the significantly

longer exons in the youngest genes was not confirmed for the two fish genomes. However, for these genomes this may in part be due to a bias against annotating genes that have no homologs in other genomes. Note that the shared trends can only partly be ascribed to the shared early history of vertebrates. The fish versus mammal lineages have had 800 million years of independent evolution, during which the trends seen in the genes shared between the lineages could have been subject to changes, unless they were robust.

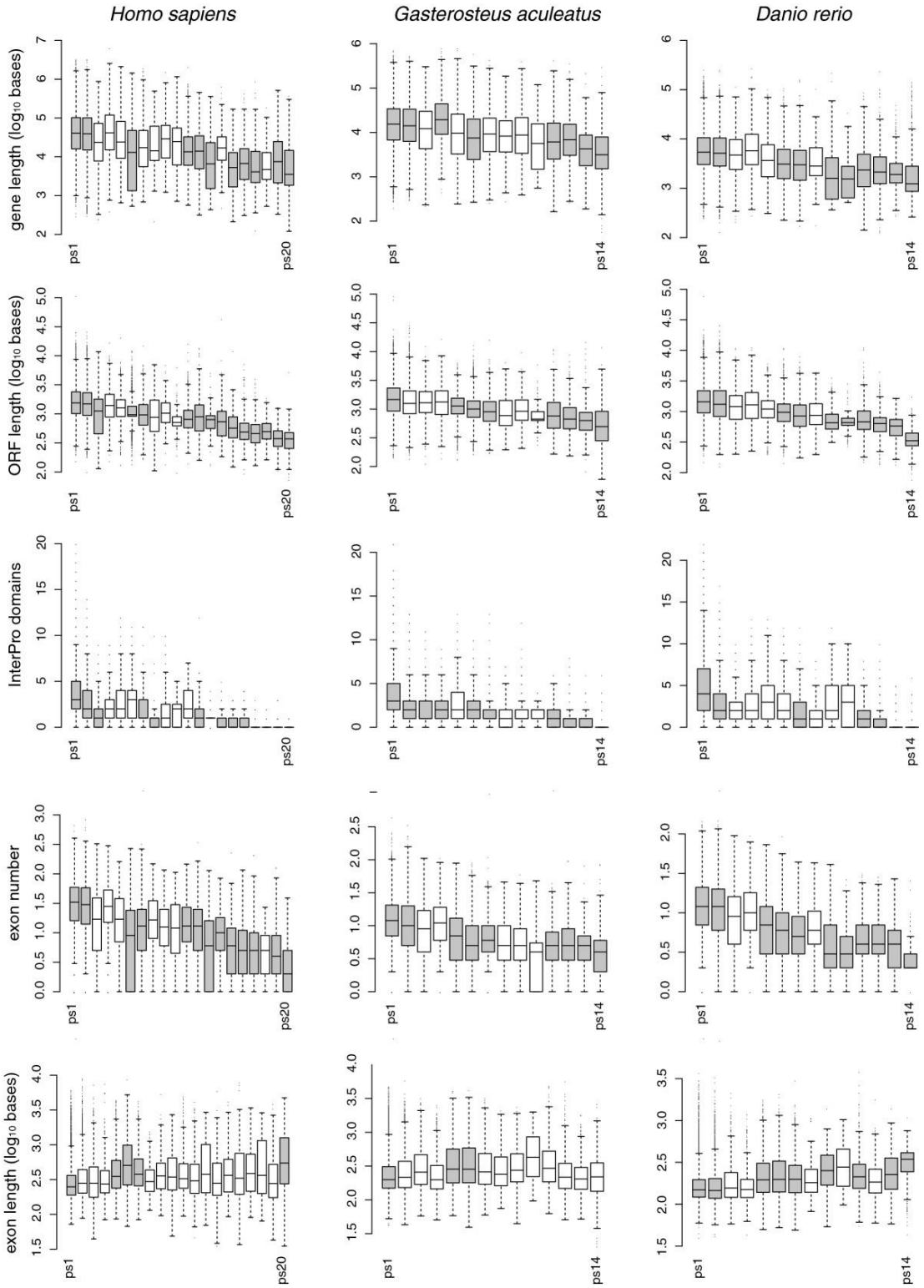

Figure 1.3. Trend comparisons in additional genomes.

Same analysis as shown in Figure 1.2, but for humans (*Homo sapiens*), stickleback (*Gasterosteus aculeatus*) and zebrafish (*Danio rerio*). Note that the fish phylostratigraphy has only 14 phylostrata in total so far, whereby ps1-12 are shared with the mammal genomes. Statistical annotations as in Figure 1.2.

Chromosomal distribution

Gene emergence appears to be randomly scattered across all chromosomes (Kolmogorov-Smirnov test, 10,000 permutations), with exception of a few clusters (Figure 1.4A). However, most of these represent a single locally expanded gene family, with one interesting exception on chromosome 14. This is a block of about 5 Mb located at the centromeric end of the chromosome (Figure 1.4B). This cluster has already been described as a complex region including a gene family involved in regulating synaptic activity in mouse (Tu et al., 2007). Our analysis suggests that it is indeed a region with a high rate of gene birth, composed of sets of genes that have arisen at different times. But, apart of this special region, there is currently no indication for a localized generation of new genes. Hence, although the *de novo* and the duplication-divergence model are both compatible with this pattern, one could have expected for a duplication model that more local clusters could have become apparent.

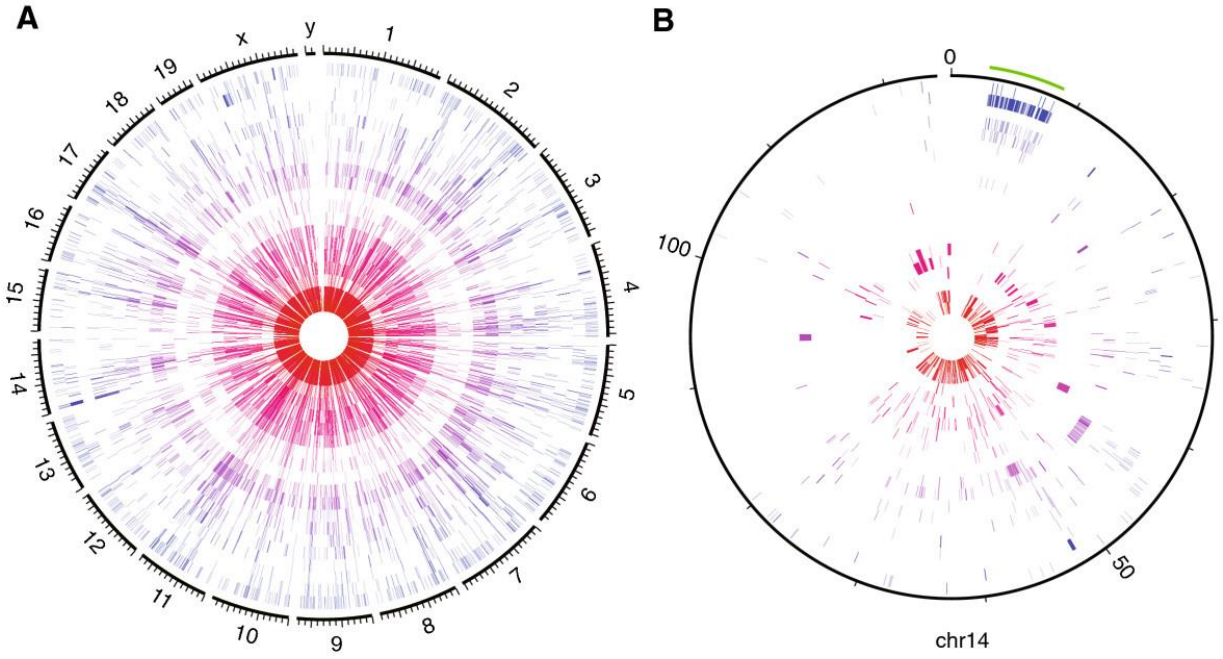

Figure 1.4. Circos plots of chromosomes and phylogenetic age of their genes.

Clockwise orientation, tick marks every 10 Mb. Genes become younger towards the outer part of the circle (represented by hues of red to blue). A. Whole genome representation. B. Chromosome 14. Green mark indicates a local cluster of young genes spanning several phylostrata.

Association with transcriptionally active sites

Transcriptionally active regions can be identified by specific marks, such as CpG islands, histone methylation (H3K4me3) peaks or DNAseI sensitivity hotspots. We find that genes in

ps1-3 (representing origin of cellular organisms, eukaryotes and opisthokonts, respectively) have a significant excess of genes associated with these regions (Figure 1.5A), in line with their predominantly general cellular functions. Another over-representation peak occurs at ps8 (evolution of chordates), which is of yet unclear significance.

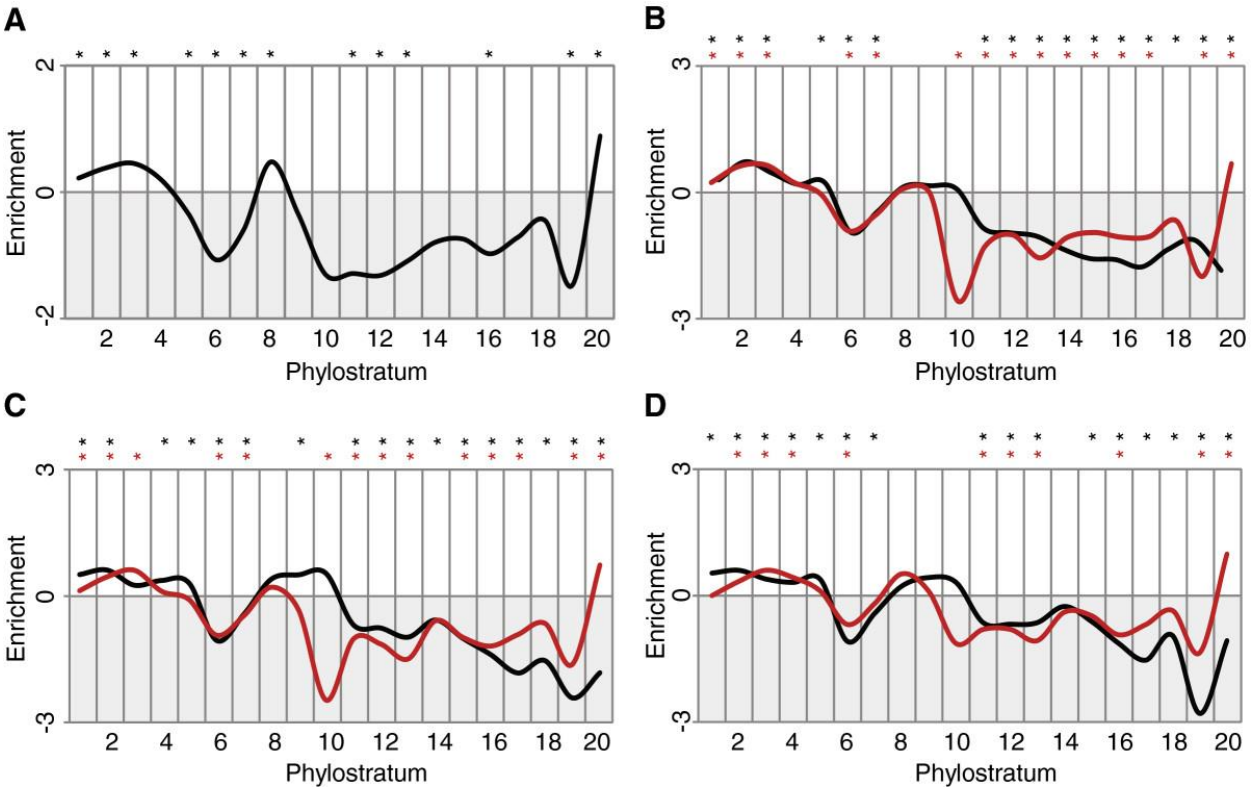

Figure 1.5. Association of transcription marks by phylostratum.

Log-odds of gene counts as enrichment. A) combination of three transcriptional hallmarks: CpG islands, H3K4me3 peaks and DNAseI sensitivity hotspots. B) to D) profiles for single transcription marks, separately for unidirectional promotors (black lines) and bidirectional promotors (red lines). B) CpG islands, C) H3K4me3 peaks, D) DNAseI sensitivity hotspots. * Hypergeometric test, FDR corrected, $p<0.01$.

With respect to the *de novo* model, it is particularly interesting to ask whether the most recently evolved genes are associated with such marks, since this could imply that they tend to make use of existing promotors upon their emergence. We find indeed a significant over-representation of transcriptional marks for genes that have emerged in ps20 (Figure 1.5A). This would suggest that the transcription of *de novo* evolved genes is initially often dependent on the proximity to an existing transcriptionally active region. Intriguingly, however, the ps19 genes show a significant under-representation with respect to the association of these three marks. This would suggest that new genes acquire rather quickly own regulatory elements, independent of the standard marks.

To explore this pattern further, we analyzed each of the three marks separately and further distinguished between unidirectional and bidirectional promotors (Figure 1.5 B-C). The latter are the most evident candidates of cases where newly evolved genes take advantage of an existing regulatory region. We find that bidirectional promotors are indeed consistently over-represented in genes from ps20 for all three marks.

Testis expressed genes

Testis is known to have the largest number of tissue-specifically expressed genes, many of which are newly evolved genes (Kaessmann, 2010). It has therefore been suggested that new genes arise predominantly first in the context of testis expression, before acquiring roles in other tissues - the "out of testis hypothesis" (Kaessmann, 2010).

When plotting the over- and under-representation profiles specifically for testis expressed genes, we find a significant enrichment for testis genes mostly from ps15 onwards (Figure 1.6). But there is no significant peak at ps20 as one would have expected under the "out of testis" hypothesis. On the other hand, it should be noted that we are looking here at protein-coding genes only, while many newly emerged testis expressed genes may initially have been non-coding and have evolved a functional ORF only later on (Tautz and Domazet- Loso, 2011). This hypothesis is in line with the peak seen in ps19, which represents the time frame within which functional ORFs could have evolved.

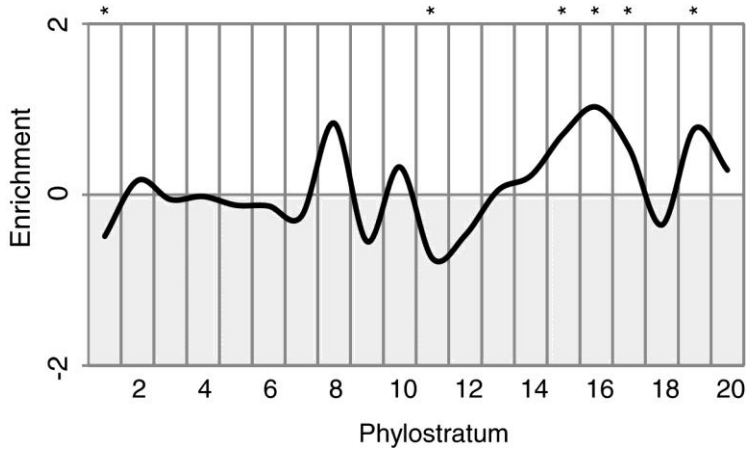

Figure 1.6. Phylogenetic profile of genes expressed in mouse testes.

Log-odds of expressed genes as enrichment for each phylostratum. *Hypergeometric test, FDR corrected, $p<0.01$.

Alternative reading frames

De novo evolution of genes could also occur within the context of an existing gene, for example through the emergence of an alternative exon that changes the reading frame or by making use of a different start codon which would lead to the translation of an alternative reading frame (Keese and Gibbs, 1992; Ohno, 1984). We used the phylostratigraphy approach to assess the age of the ORFs of genes with two annotated reading frames and find that they can indeed be significantly different, indicating a secondary evolution of a new gene within an existing gene. We can find 13 such genes among the current Ensembl annotated reading frames, only two of which were previously identified as overprinted genes (Table 1.2). We discuss here three further examples representing three general patterns (Figure 1.7).

Table 1.2. List of overprinted genes detected via a phylostratigraphic approach based on annotated ORFs in Ensembl.

ENSMUS IDs				Phylostratum		Comment
Gene	Newer protein	Older protein	Gene name	Overprint	Original	
G00000029642	P00000106186	P00000058355	Polr1d	5	2	Same start as main gene, but acquired additional exons
G00000030970	P00000127123	P00000033269	Ctbp2	12	1	Same start as main gene, but acquired an additional internal exon
G00000035504	P00000100994	P00000100995	Reep6	17	2	New initiation codon creates second reading frame
G00000089756	P00000104646	P00000104577	Gm8898	18	2	Same start, but new splice variant; paralog of Gm4723
G00000078898	P00000104676	P00000104675	Gm4723	18	2	Same start, but new splice variant; paralog of Gm8898
G00000038227	P00000133896	P00000046939	Hoxa9	18	2	New starting exon initiates a separate reading frame
G00000067786	P00000134415	P00000085836	Nnat	18	16	Same start, alternative splicing leads to new reading frames
G00000044405	P00000105110	P00000051732	Adig	20	16	Same start as main gene, but acquired an additional internal exon
G00000025144	P00000101761	P00000026137	Stra13	20	2	Gain of alternative second exon induces a shift from the older frame

G00000033720	P00000109417	P00000041872	Sfxn5	20	2	Alternative first exon and last exons, common second exon
G00000063235	P00000107087	P00000077036	Ptpmt1	20	1	Alternative transcription start site and start codon
G00000044303	P00000030237	P00000061847	Cdkn2a[a]	16	1	New starting exon initiates a separate reading frame. Also known as Arf, Pctr1, MTS1, Ink4a
G00000027523	P00000104716	P00000085184	Gnas[b]	18	2	New initiation codon creates second reading frame. Also known as Nesp, GPSA

[a]has previously been described, see (Chung et al., 2007; Sherr, 2006).

[b]has previously been described, see (Chung et al., 2007; Klemke et al., 2001; Nekrutenko et al., 2005).

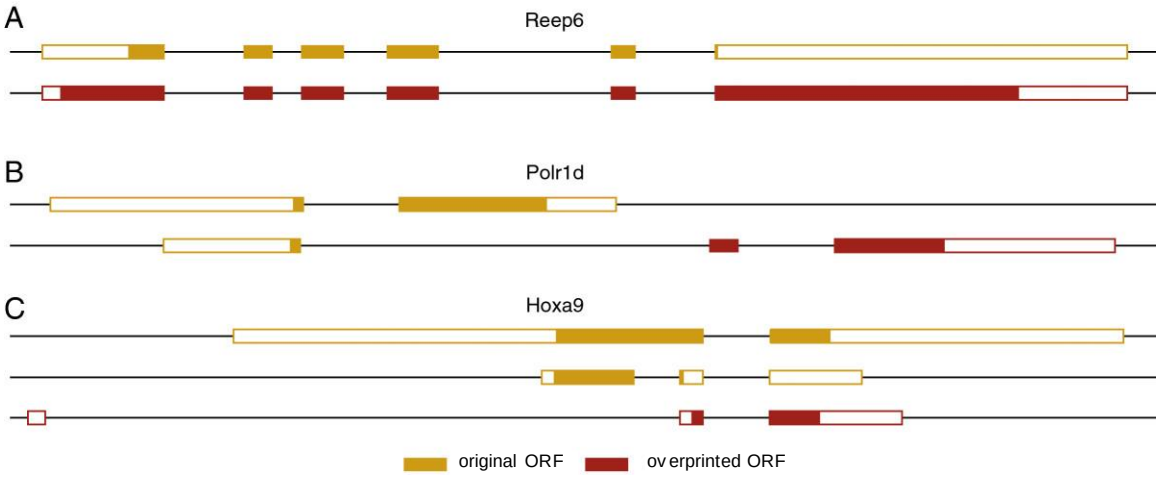

Figure 1.7 Examples for overprinting of genes.

Gene structures of four genes are shown, whereby the exons (boxes) are drawn to scale, while the introns (lines) are not to scale. Open boxes are non-coding, filled boxes represent the reading frames. Ancestral gene versions are yellow, derived ones are purple. The figure is based on annotations and graphics from Ensembl, whereby only the relevant splice variants are shown. A: Reep6 (ENSMUST00000030237 and ENSMUST00000060501); B: Polr1d (ENSMUST00000154641 and ENSMUST00000114425); C: Hoxa9 (ENSMUST00000048680, ENSMUST00000110557 and ENSMUST00000050970).

The first example is the gene Reep6, where an additional start codon has evolved in the first exon, which initiates a new reading frame, overlapping the ancestral one (Figure 1.7A). The older product of Reep6 maps to ps2, the newer one to ps17, i.e. it appears to have acquired a new function at the boreoeutherian divergence. Interestingly, when looking at the gene trees of

these proteins, one can see a clear acceleration of divergence rates in conjunction with the emergence of the second reading frame for Reep6, but not for its nearest paralog Reep5, which has not developed the second reading frame (Figure 1.8). Such acceleration is a hallmark of an adaptive phase and was also found in viruses (Sabath et al., 2012).

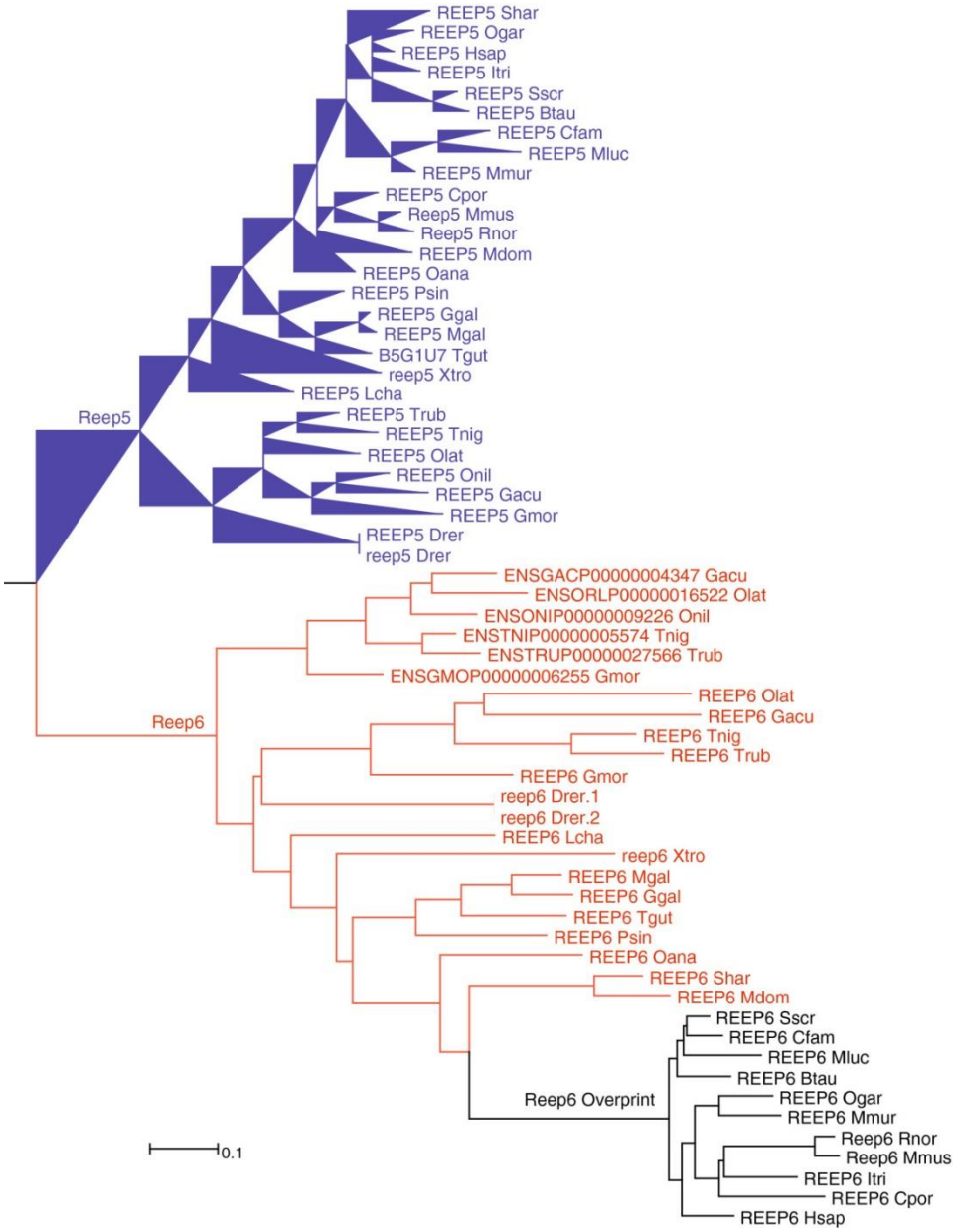

Figure 1.8. Phylogenetic tree for Reep5 (blue labels) and Reep6 (red/black labels).

Both genes are present in euteleostomes (ps12), and belong to a larger family of eukaryotic genes (ps2). The Reep6 locus in mouse codes for two proteins of different age. The older protein (Reep6) was mapped to ps2 (eukaryotes), the newer protein (Reep6 overprint) to ps17 (boreoeutherians). Note the enhanced substitution rate at the basis of this subtree (in black), as represented by branch length. Modified gene tree from Ensembl record ENSGT00550000074535.

The second example for overprinting is Polr1d, a subunit of RNA polymerase I and III, which has acquired two additional exons at the end of the ancestral gene. Alternative splicing leads thus to a new protein that shares only the start codon and a few initial amino acids with the ancestral gene (Figure 1.7B). The ancestral protein maps to ps2, the derived one to ps5, i.e. this arrangement with two protein products from the same gene region is highly conserved.

The third example is Hoxa9, one of the canonical Hox genes involved in anterior-posterior patterning. In this case, the ancestral gene has first acquired an additional intron that leads to a truncated version of a protein, an arrangement that is conserved between birds to mammals (Dintilhac et al., 2004) (ps14). On top of this, an additional 5´-exon, driven by a new promotor, has evolved within the Euarchontoglires (ps18). This splices to the acceptor of the new intron and creates thus a new reading frame (Figure 1.7C). Interestingly, this reading frame covers the homeobox and is conserved between primates and rodents.

Discussion

The trends described above provide new insights into the modes of gene emergence over time. For the two models, *de novo* evolution versus duplication-divergence, it seems that *de novo* evolution is better compatible with these trends. But before coming to the interpretations, we should first like to discuss the technical aspects of our approach.

We rely generally on blastp searches for classifying the genes to phylostrata. There have been extensive simulation efforts that have shown that this is an adequate procedure (Alba and Castresana, 2007). However, if one would add manual curation, including the use of a combination of different search algorithms, one would indeed classify a number of genes to older phylostrata. On the other hand, we are focusing here on general trends, not on absolute numbers. Given that most of these trends are robust, both with respect to statistical testing, as well as for confirming them for the much less well annotated fish genomes, we consider the possible misclassification problem as small.

We relate our analysis only to the currently annotated Ensembl reading frames, although these are in a constant flux, due to curation and further refinement of annotation procedures. In fact, it has already been noted that the currently available annotations underestimate the number of orphan genes, since finding a homologue for a gene is one accessory criterion for annotation. This affects mostly the genes from ps20, which are under-represented (Begun et al., 2007; Tautz and Domazet-Lošo, 2011), although they are the best candidates for ongoing *de novo*

evolution. Hence, although some noise is expected in the data and the assignment fidelity, it would be very unlikely that a systematic artifact causes the trends observed.

De novo evolution versus duplication-divergence

The *de novo* emergence of a gene out of non-coding DNA requires only some form of transcription, as well as simple signals that define its start and its end and possibly splice sites, as well as some open reading frame (Siepel, 2009; Tautz and Domazet- Loso, 2011). Since all of these signals are rather short, they are expected to occur frequently even in random sequences. Genes emerging from such random combination of signals have been called proto-genes (Carvunis et al., 2012; Siepel, 2009) and analysis of ribosome association profiles in yeast has suggested that they are abundantly translated (Carvunis et al., 2012; Wilson and Masel, 2011). Accordingly, they could easily serve as a continuous source of short genes that are ready to become recruited to functional pathways and can then become more complex over time. Hence, new genes that arise according to this model would initially be short, have few introns and domains and would often be associated with existing regulatory elements. These are indeed the overall trends that we observe.

The duplication-divergence model, on the other hand, seems much less compatible with these trends. Under this model, one would expect that the new gene should inherit the gene structure from the parental gene. Since long and short genes should equally often be the source of new genes, and since duplications should happen similarly at all time horizons, one would not expect to see the dependence between age and length features.

Domain number is also highly correlated with age, with younger genes having far fewer domains. This is not a simple effect of the similarity searches that we have used, since the domain annotation in Interpro is based on a combination of a variety of different procedures that go beyond blastp matches (Hunter et al., 2011). Hence, this observation confirms that not only new genes, but also new domains can arise over time (Moore and Bornberg-Bauer, 2012; Pal and Guda, 2006). On the other hand, only half of the genes contain known domains (Chothia and Gough, 2009), i.e. having a domain is not a prerequisite of protein function. In fact, many proteins are known to be intrinsically unstructured (Dyson and Wright, 2005; Schlessinger et al., 2011; Tompa and Kovacs, 2010).

Regulatory evolution

It is still unclear how a new gene can acquire its regulatory elements. One possibility is that there are many cryptic transcriptional initiation sites around the genome. Indeed, it appears that

most of the genome becomes transcribed at some time (Carninci, 2010; Clark et al., 2011). However, much of this may be co-transcription or spurious initiation. Moreover, to allow a transcript to become functional (i.e. to become subject to positive selection), it requires some form of stable and heritable regulation. We have therefore evaluated the possibility that new genes make use of existing promotors. It is known that RNA polymerase II promotors have a general tendency for divergent transcription within the nucleosome-free region associated with most promotors (Seila et al., 2009; Tautz, 2008). We find indeed an enrichment of general signatures of active promotors in association with the most recently evolved genes (ps20). This is mostly due to bidirectional promotors, where the general tendency of RNA PolII for bidirectional transcription may have become extended to form a new transcript. Intriguingly, the next phylostratum (ps19) shows an under-representation of genes among bidirectional promotors, which would suggest that a new gene that has become functional could rather quickly gain its own independent promotor elements.

Overprinting

Another way of making use of an existing promotor is to develop an alternative reading frame within an existing gene. This can be caused by the acquisition of an alternative splicing, whereby the original start codon is retained (e.g. in Polr1d). Alternatively, a separate start codon becomes used that initiates a different reading frame (e.g. Reep6). This has long been thought to be very unlikely, mostly because of the common notion that in eukaryotes only the first AUG serves as a start codon in an mRNA. However, polycistronic mRNAs are known to occur in eukaryotes as well (Tautz, 2008), i.e. the use of additional start codons from the same transcript is not without precedence. The third possibility to initiate an alternative reading frame within an existing gene is a new upstream exon, driven by a new promotor, combined with alternative splicing. This has apparently happened in the case of the Hoxa9 gene. This is also the mechanism that was found for the previously well-studied example of overprinting in the Cdkn2a gene (Sherr, 2006). This raises of course the question of how the new promotor for the new upstream exon has evolved. However, it has been shown that there is a widespread presence of long-range regulatory activities in the mouse genome, which can act on inserted promotors (Ruf et al., 2011). Thus, it seems indeed rather conceivable that random mutations in such potentially active regions might suffice to create a new regulated initiation site.

We expect that it should be possible to detect many more cases of overprinting, if one does not only search annotated reading frames, as we have done here. For example, Chung et al. (Chung et al., 2007) have identified 40 candidates for overprinting in humans using a

probabilistic search strategy. With the much better genome sampling that we have nowadays, it should be possible to refine the searches even further.

Our search has specifically focused on cases where the overprinted reading frame has emerged later than the original one. Two of the previously well-studied genes fall into this class and we have recovered them. Such secondarily evolved proteins are the ones that give the strongest support for a *de novo* evolution mechanism, since alternative reading frames of long existing genes can be considered as almost random sequences. Hence, the fact that new proteins can arise out of them is a strong argument for the reality of *de novo* evolution (Chung et al., 2007; Keese and Gibbs, 1992; Ohno, 1984).

Conclusion

The phylostratigraphy-based analysis of trends associated with gene emergence in the mouse genome is well compatible with a frequent *de novo* emergence of orphan genes. This seems to be in contrast to previous assessments, which found only a small fraction of cases of *de novo* evolution (Ekman and Elofsson, 2010; Toll-Riera et al., 2009; Zhou et al., 2008). However, it is necessary to emphasize that this depends very much on the criteria that were used. These early studies were still constrained by the assumption that *de novo* evolution must be rare and the criteria were therefore tuned to be very restrictive to be sure that only the best-supported cases were included. In addition, it has initially been unclear whether any new gene that includes part of a transposable element should be classified in a separate class(Toll-Riera et al., 2009), since strictly speaking it contains at least partly a duplicated sequence. On the other hand, if the transposable element fragment does not contribute its reading frame to the new gene, we would now consider it as a *de novo* gene, given that we find also overprinting in other existing genes. We should also reiterate that our analysis here is strictly based on genes that were annotated as protein coding, whereby the criteria for annotation of genes are still rather restrictive and tend not to consider short open reading frames, although these may be functional as well (Tautz, 2008). Further, all non-coding RNAs are still excluded from this analysis, although the emergence of new *de novo* genes may be characterized by a phase where it acts as non-coding RNA first (Cai et al., 2008; Heinen et al., 2009). Hence, we conclude that we are only at the beginning to understand the true impact of *de novo* gene evolution on shaping the genome and emergence of new gene functions.

Methods

Phylostratigraphy

The phylostratigraphic procedure (Domazet- Loso et al., 2007) is a blastp-based sorting of all protein sequences of an organism according to their phylogenetic emergence. The procedure uses the annotated genes of the focal organism and compares them to all available annotated and non-annotated genome data to infer the first time of emergence of a given gene. Accordingly, all available proteins from protein coding loci in the version 66 of Ensembl (Flicek et al., 2011) for *Mus musculus* (obtained through BioMart (Kinsella et al., 2011)) were queried against the *nr* database from NCBI using an e-value threshold of $< 10^{-3}$, which has been shown to be optimal for such an analysis (Alba and Castresana, 2007; Domazet-Loso and Tautz, 2003). For phylostratum 12, given the low number of protein sequences for outgroups (Cyclostomata/Chondrichthyes), EST and Trace data were included in a tblastn query (translated nucleotide comparison), using an e-value threshold of $<10^{-15}$. The computation of the phylostratigraphic maps was performed on the Phylostrat server of the IRB Institute, Zagreb, Croatia. Twenty phylogenetic age classes, i.e. phylostrata, were defined based on consensus phylogenetic relationships (Figure 1). The age of a locus was assigned taking into account the oldest detectable similarity of any of its protein products. This approach is targeted to the detection of orphan genes, as it neglects events of exon shuffling or gene fusion as genomic novelties.

Gene structure analyses

Structural gene features were obtained from version 66 of Ensembl through BioMart for mouse (*Mus musculus*), and from version 68 for human (*Homo sapiens*), zebrafish (*Danio rerio*) and stickleback (*Gasterosteus aculeatus*). Domain information from Interpro (Hunter et al., 2011) was also obtained through BioMart, and the number of different entries per gene was used as a proxy to the number of domains. Phylostratigraphic analyses were tested with hypergeometric statistics for discrete features and correlations were calculated for continuous features. A combination of permutations (n=10,000) and Kolmogorov-Smirnov tests was used to assess the significance of each phylostratum per variable. Kolmogorov-Smirnov tests were also applied to distance distributions. Other statistical tests were perfomed using R version 2.15.1 (R Core Team, 2012) and PASW version 18.0.0 (SPSS Inc, 2009). Circular plots for the mouse genome were done with Circos (Krzywinski et al., 2009).

Transcription associated regions

Regions of high transcriptional activity from basal promotors were defined as those containing any of these three features: presence of CpG islands, H3K4me3 peaks or DNAseI sensitivity hotspots. These features allow broad range recognition of potential and actual sites with enhanced transcriptional activity. All datasets were taken from the UCSC Genome Browser (Fujita et al., 2011; Kent et al., 2002) through the Table Browser tool (Karolchik et al., 2004). Datasets for H3K4me3 ChIP-seq (Mouse ENCODE Consortium et al., 2012) were obtained from the available tracks from *Histone Modification by ChIp-seq* at ENCODE/LICR (Ludwig Institute for Cancer Research). Available tissue data at the time of the study include bone marrow, cortex, cerebellum, heart, kidney, liver, lung, mouse embryonic fibroblasts and spleen (all from 8 week old mice). Only peak data were used. Datasets for DNAseI sensitivity assays were obtained from the *DNAseI Hypersensitivity by Digital DNAseI* from ENCODE/University of Washington tracks (Mouse ENCODE Consortium et al., 2012). Only hotspots information was used and only tracks corresponding to C57BL/6 mice. Genes were considered to be associated to these marks if the transcription start site was found at a distance of 1,250 bases or less from the mark, accounting for potential offsets in annotations and allowing the assumption that transcriptional activity might affect more drastically those regions in a short range. Analyses of overlap between regions were performed with the BEDtools suite (Quinlan and Hall, 2010). Phylostratigraphic enrichment was calculated as log-odds and tested using hypergeometric statistics and FDR correction.

Expression data for testis

Mouse microarray expression data from (Zhang et al., 2004) were obtained from the authors' website (http://hugheslab.ccbr.utoronto.ca/supplementary-data/Zhang/). This study was selected because of the wide spectrum of tissues considered, which allow for an unbiased measure of expression for a large set of genes. Given that the study was performed using a draft of the mouse genome, the probes were re-annotated using Blat (Kent, 2002a) to match the phylostratigraphic map of the mouse. Ambiguous and poorly matching probes were discarded from the analyses.

Secondary reading frames

This screen was devised to find annotated candidates for emergence of new genes within existing genes based on annotated products. All complete open reading frames corresponding to the same genomic location (ENSMUSG) were considered as candidates, if the minimum and

maximum age values differed by at least 2 phylostrata (to avoid screening borderline classifications between phylostrata). Within each genomic location, ORFs were aligned at the nucleotide and protein level using global (needle) (Rice et al., 2000) and local alignments (blastn and blastp, database size adjusted to emulate nr-sized searches) (Altschul et al., 1997). The oldest product was used as reference, and any products with younger phylostrata values were used as query. In the case of multiple older products, comparisons were made against all possible products from the oldest phylostratum. Non-matching protein alignments coming from matching nucleotide alignments were considered as genes with alternative reading frames. These were screened manually in Geneious (version 5.6.5) to identify conservation patterns of start and stop codons in other species. Additionally, using the Compara platform from Ensembl (Vilella et al., 2008), phylogenetic trees for selected candidates were analyzed.

Acknowledgements

We thank Tomislav Domazet-Lošo for providing access to the Phylostrat server at the IRB in Zagreb, Croatia; Robert Bakaric for development and support of the Phylostrat server and Sebastian Meyer for work on preliminary tests regarding mouse overprinting. RN is member of the International Max-Planck Research School (IMPRS) for Evolutionary Biology.

This study was published online the 21st of February 2013 in BMC Genomics. I was responsible for the data collection and analyses. The design of the analyses, interpretation and manuscript were done together with Prof. Diethard Tautz. The section related to analyses using testes data was originally made for my Master Thesis (Neme Garrido, 2011) prior to publication. All other analyses were performed during the course of my doctoral studies.

BMC Genomics 2013, 14:117

doi:10.1186/1471-2164-14-117

Chapter 2: Sequencing of genomes and transcriptomes of closely related mouse species

Introduction

A further goal of my work was to explore the properties of transcription at a genome- and phylogeny-wide level and to inspect how much of the detectable transcription is conserved across a time scale that spans between 3,000 years and 10 million years of divergence. To achieve this, the genomes and three tissue transcriptomes of mouse populations and species with increasing evolutionary distance to the reference strain were sequenced. This is described in this chapter. The results from the analyses of these datasets in the context of gene birth are presented and discussed in Chapters 3 and 4.

Using wild mice to understand gene birth at the transcriptome level

Addressing questions about how genes emerge requires large quantities of detailed sequence information. In order to identify if a given gene has arisen *de novo*, it is not only necessary to detect it in a given species, but also to detect its absence in other organisms, which translates into needs of high-quality information for many closely related species. In addition to this, a well annotated reference genome is required to be able to identify the candidates for *de novo* evolution. For this reason, this can best be addressed using classic model organisms, like yeast, *Arabidopsis*, *Drosophila* or mice as reference, for which ample genetic tools and resources exist.

In my project, the study of gene birth in mouse profits from one of the best assembled and most feature-rich genomes available for a species: the reference genome of the C57BL/6 line. The molecular cell biology and molecular genetics of mice are among the best understood regarding mammals, and to this date, mice are the closest organisms to humans with a comparable amount of biological resources for which genetic information is available, and this allows the experimental validation of predictions based on genomic information.

It is likely that the dynamics governing the emergence of new genes in mouse, given its molecular, cellular and organismal complexity, are similar to those in other mammals, humans included.

Furthermore, the knowledge about closely related populations, subspecies, and species of wild mice with known phylogeographic distributions enables fine-scaled phylogenomic analyses, suited to identify and understand how new genes appear for the first time in a genome.

Phylogeographic distribution of the samples

I included ten taxa, ranging from diverging populations through sister genera (Figure 2.1). The youngest divergence point sampled, at about 3,000 years, corresponds to the split between two European populations of *Mus musculus domesticus* (Cucchi et al., 2005); one from France and one from Germany. These European populations in turn have diverged from the ancestral *M. m. domesticus* between 12,000 and 30,000 years ago (Cucchi et al., 2005). The European *M. m. domesticus* are also the closest relatives of the reference genome, the C57BL/6J strain. Although the exact divergence time between them is uncertain, it is known that the C57BL/6J strain derives from American *M. m. domesticus* (Goios et al., 2007) and that the progression of *M. m. domesticus* into the Americas followed the history of human colonization (Guénet and Bonhomme, 2003; Jones et al., 2012).

I also included two populations of *Mus musculus musculus*; one from Austria, near the *M. m. domesticus* / *M. m. musculus* hybrid zone, and another from Kazakhstan. These two populations are supposed to have a longer divergence between them than any of the *M. m. domesticus* populations, but more accurate estimates are not available. For this reason I have set the divergence for analyses at around 10,000 years as an approximate estimate . *M. m. domesticus* has diverged from *M. m. musculus* and *Mus musculus castaneus* about 0.4 to 0.5 million years ago, with a subsequent divergence, not long after, between *M. m. musculus* and *M. m. castaneus* (Suzuki et al., 2013).

In addition to this dense group of subspecies, I have included *Mus spicilegus* (estimated divergence of 1.2 million years); *Mus spretus* (estimated divergence of 1.7 million years) (Suzuki et al., 2013); *Mus matteyii* (subgenus *Nannomys*), the North African miniature mouse (estimated divergence of 6.6 million years) (Catzeflis and Denys, 1992; Lecompte et al., 2008), and *Apodemus uralensis*, the ural field mouse (estimated divergence of 10.6 million years) (Lecompte et al., 2008).

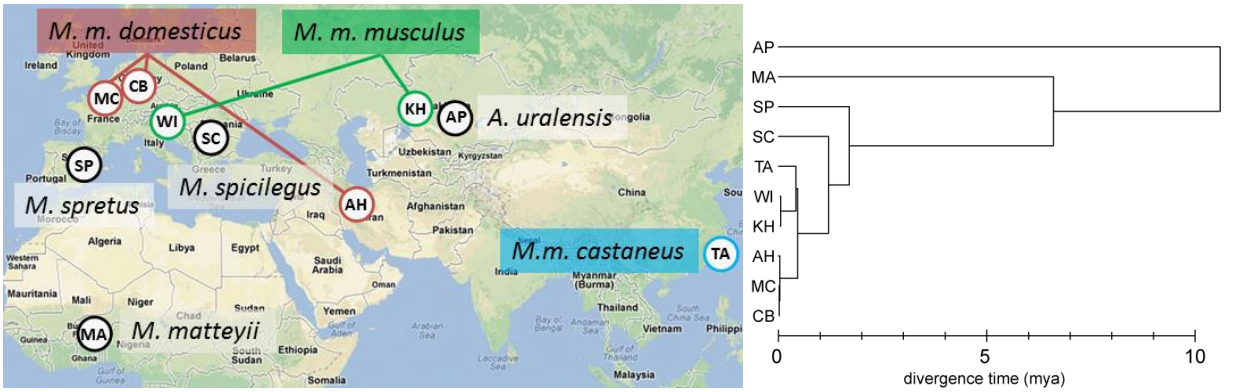

Figure 2.1. Approximate geographic origin of the samples used for transcriptome sequencing (left), and estimated divergence time between them (right).

Red circles indicate populations of *Mus musculus domesticus*: AH from Iran, CB from Germany and MC from France; green circles indicate populations of *Mus musculus musculus*: KH from Kazakhstan and WI from Austria; the blue circle indicates a population of *Mus musculus castaneus* from Taiwan (TA). Outgroup species of *Mus musculus* indicated in black: *Mus spicilegus* (SC), *Mus spretus* (SP), *Mus mattheyi* (MA) and *Apodemus uralensis* (AP). Map modified from Google Maps. Divergence estimates based on fossil, mitochondrial, nuclear estimates from different sources: *M. m. domesticus* and *M. m. musculus* (Cucchi et al., 2005); *Mus musculus* subspecies, *Mus spicilegus* and *Mus spretus* (Suzuki et al., 2013); and on *Mus mattheyi* and *Apodemus uralensis* (Lecompte et al., 2008).

Until very recently, a common strategy for detecting *de novo* genes was based on comparative genomics setups from publicly available data, which usually contain a handful of annotated genomes (Capra et al., 2012; Domazet- Loso et al., 2007; Donoghue et al., 2011; Guerzoni and McLysaght, 2011; Neme and Tautz, 2013; Toll-Riera et al., 2009; Wu and Zhang; Wu et al., 2011; Yang and Huang, 2011). The annotation information generally derives from expression data and bioinformatics predictions.

Many predicted genes lack evidence of transcription in public databases (See Appendix B). In addition to this, many existing genes are not annotated due to biases in the data collection processes, which arise from the different levels of interest and funding available for each model species. This leads to a generalized abundance of low quality detection of *de novo* genes. Furthermore, the lack of understanding about *de novo* genes has led to biases in public databases, where well-known examples of *de novo* genes have been dismissed from specific releases on the basis of irregular features (Wu et al., 2011).

My data are aimed to directly contribute to evidence-based analyses of gene birth across timescales not evaluated before. An even sampling and treatment of the data enables comparisons of previously unavailable orthologs (Chapter 3), and strongly reduces biases seen

in previous analyses therefore enabling quantitative analyses, e.g. the rates at which *de novo* genes are gained (Chapter 4).

Methods

Biological material

Mice of different ages were sacrificed by CO_2 asphyxiation followed by cervical dislocation. Mice were then dissected and tissues were snap-frozen within 5 minutes post-mortem. The tissues collected were liver (front view: front left lobe), both testis and whole brain including brain stem.

For the outbred populations, Iran (AH), France (M), and Germany (CB) for *Mus musculus domesticus*, and Austria (WI) and Kazakhstan (KH) for *Mus musculus musculus*, eight individuals each were sampled. For inbred groups, *Mus musculus castaneus* (TA), *Mus spretus* (SP), *Mus spicilegus* (SC), *Mus mattheyi* (MA) and *Apodemus uralensis* (AP), four individuals each were sampled. All mice were obtained from the mouse collection at the Max Planck Institute for Evolutionary Biology.

Transcriptome sequencing

The sampled tissues of each group were used for RNA extraction with the RNAeasy Kit and pooled at equimolar concentrations. Quality of the RNA was measured with BioAnalyzer chips, for the individual samples and pools, and samples with RIN values below 7.5 were discarded. The pools were subsequently submitted to the Cologne Center for Genomics (CCG) for further processing and sequencing. The sequencing of the samples was performed using a polyA tail purification step, followed by cDNA synthesis, Illumina library preparation, and sequencing with an Illumina HiSeq 2000 sequencer. Each transcriptome sample was sequenced in approximately one third of a HiSeq2000 lane.

Genome sequencing

One individual from each of *M. spicilegus*, *M. mattheyi*, and *Apodemus uralensis* were selected for genome sequencing. The same liver tissue samples used for transcriptome sequencing were used, with the exception of *M. mattheyi*, since the remaining tissue sample was too small. For this I used a different individual. DNA extraction was performed using a standard salt extraction protocol (Appendix), and the samples were sent to the Cologne Center for Genomics for sequencing on an Illumina HiSeq2000 sequencer. After library preparation, the three genome

samples were pooled together and run in a whole IlluminaHiSeq 2000 flow cell (8 lanes, approximately 2.6 lanes per sample).

The genome from the strain SPRET/EiJ derived from *Mus spretus* is currently available as short reads (Keane et al., 2011; Yalcin et al., 2012), and was downloaded from the European Nucleotide Archive (ENA) from the accessions ERS076388 and ERS138732.

Raw data processing

All raw data files were trimmed for adaptors and quality using Trimmomatic (Lohse et al., 2012). The quality trimming was performed basewise, removing bases below quality score of 20 (Q20), and keeping reads whose average quality was of at least Q30. Reads whose trimmed length was shorter than 40 bases were excluded from further analyses, and pairs missing one member because of poor quality were also removed from any further analyses.

Transcriptome read mapping, annotation and quantification

The reconstruction of transcriptomes using high-throughput sequencing data is not trivial when comparing information across different species to a single reference genome. This is due to the fact that most of the tools designed for such tasks do not work in a phylogenetically aware context. For this reason, any approximation which deals with fractional data (i.e. any high-throughput sequencing setup available to this date) is limited to the detection abilities of the software of choice.

Transcriptome sequencing reads were aligned against the mm10 version of the mouse reference genome from UCSC (Fujita et al., 2011) using NextGenMap (Sedlazeck et al., 2013) and TopHat2 (Kim et al., 2013). TopHat2 was used to obtain read sequence alignments, which were further processed with cufflinks (Roberts et al., 2011) to obtain an annotation file for each tissue in each species. This annotation file contains models for expressed transcripts, with splicing information when available. The outputs of each species and tissue were merged using cuffmerge, and a final annotation file was generated from the combination of all species/tissues.

The TopHat2-cufflinks gene models were used to quantify read counts, but instead of using the same alignments used to generate the models, I used NextGenMap alignments. NextGenMap is able to map across longer phylogenetic distances (Sedlazeck et al., 2013), but is unable to define spliced reads, therefore it is inadequate for gene model reconstruction, but very powerful for expression detection. Reads which were ambiguously or poorly mapped were removed from the analyses.

The expression information was extracted using the featureCounts facility from the subreads alignment suite (Liao et al., 2013); counting fragments (instead of reads) across the exonic regions of the predicted models. The expression information was normalized across samples by giving each library a weight factor based on the sample median versus the mean median across all samples, similar to the one used in DESeq (Anders and Huber, 2010).

Total transcriptome quantification against the genome from NextGenMap alignments was performed using the bedtools (Quinlan and Hall, 2010) suite to identify the genome-wide basewise coverage across all tissues and taxa. Expression was called even for regions with only one read, as long as it was uniquely mapping. Regions between spliced reads from TopHat2 junctions were also included in the coverage statistics.

Expression information was also obtained for each tissue. Any transcript having more than 10 fragments was considered present in a given tissue (see below). Given the nature of the pooled data, the expression information was only used to infer general patterns of expression and by no means should be considered the result of differential gene expression analyses.

Basic statistics about the mapping of transcriptomic reads can be found in Appendix D.

Genome read mapping

Genomic reads from *Apodemus uralensis*, *M. mattheyi*, *M. spretus* and *M. spicilegus* were aligned against the mm10 mouse reference genome using NextGenMap (Sedlazeck et al., 2013) restricting alignments to uniquely mapping regions with samtools (Li et al., 2009). The output of the alignments was used to establish whether each exon was detectable or not at the genome level. This serves as an overall mappability control since there are both absent regions as well as regions of poor mappability (e.g. recent duplicates or highly conserved paralogs). One of the disadvantages of using programs which allow alignments in a phylogenetically-aware context is that the detection of recent paralogs becomes more difficult. For this reasons, I considered the candidates detected in this screen as either single locus, or in the eventual case of potential duplications, as well resolved paralogs from other loci. A minimum coverage of 10x was used to define the region as present and mappable (the genome average for each species was between 26x and 32x).

Available resources

All data resources used will be made available upon request for reviewers on a private server. All data, raw and processed, will be freely available upon the publication of this work.

During this project Illumina sequencing reads were generated. The reads are available in FASTQ format, both as raw and quality filtered reads. Read alignment files are available in BAM format, and the specific mouse reference sequence is available in FASTA format. Genome and transcriptome coverage are available in BED and WIG format. BED files contain information about covered regions and WIG files contain base-wise coverage. Annotation files are available in GTF and BED format.

Chapter 3: Differential selective constrains across phylogenetic ages and their impact on the turnover of protein-coding genes.

Introduction

One of the defining characteristics of orphan genes is their lack of traceable homology to other genes in closely related species (Domazet-Loso and Tautz, 2003; Wilson et al., 2005), and it is also this lineage-specific behavior that operationally defines orphans in our current datasets as genes which could not be successfully matched to collections in other organisms.

In practice, this property limits close-scale analyses, since the closest related well-annotated species are usually within the range of hundreds of million years apart (Albà and Castresana, 2005), and because orphans at terminal branches usually lack orthologous sequences for detailed comparative analyses.

It is evident that the properties that govern gene birth and change across time can be different for short and long time-scales (Carvunis et al., 2012; Neme and Tautz, 2013; Palmieri et al., 2014; Tautz and Domazet- Loso, 2011; Zhao et al., 2014), and there is a growing body of evidence that suggests that new genes are easier lost than older genes (Zhao et al., 2014).

To further expand our understanding of the dynamics of gene birth, I address two complementary questions using phylostratigraphic analyses: (1) how is the behavior of evolutionary rates across phylogenetic ages when comparing organisms with increasing divergence times and (2) how the polymorphisms that influence open reading frame (ORF) stability correlate with gene age.

This was addressed by identifying orthologous sequences between the mouse reference and the taxa for which transcriptome information was previously generated (Chapter 2). I focused on the variation at short evolutionary times, between 3,000 years and 10.6 million years, across closely related mouse species, subspecies and populations.

Methods

For a description of the high-throughput data generation and processing see Chapter 2.

Transcriptome assembly

Quality-filtered transcriptome reads for each taxon were merged into a single input, discarding tissue information, and assembled *de novo* with the Trinity platform (Haas et al., 2013).

Generation of ortholog pairs and rate analyses

I was able to identify ortholog pairs for most of the genes from the Ensembl 72 annotation version of the mouse genome, and to generate codon alignments for those pairs using the gKaKs pipeline (Zhang et al., 2013). Alignments were performed exon-wise, keeping transcript structure intact. In the case of multi-transcript genes, one single representative sequence was chosen. The set of genes was filtered to contain only genes for which phylostratigraphic information was available (Neme and Tautz, 2013). Evolutionary rate was calculated as dN/dS ratios using codeml (Yang, 2007), and further filtering was applied to pairs yielding rate values greater than 2 and/or containing less than 50 codons to increase stringency as suggested by Toll-Riera (Toll-Riera et al., 2012). A total of 17,532 ortholog-pairs were obtained, with 12,132 (69.2%) being present and valid across all sampled taxa (Table 3.1).

Table 3.1. Orthologs detected between the mouse reference (Ensemble 72) and each of the assembled transcriptomes.

Taxon	Orthologs detected	%
MC	15902	90.7%
AH	15952	91.0%
CB	15953	91.0%
TA	16334	93.2%
KH	16336	93.2%
WI	16358	93.3%
SP	16726	95.4%
SC	16730	95.4%
MA	16941	96.6%
AP	17529	100.0%
Common	12132	69.2%
Total	17532	-

Overlapping genes

Overlap information was derived from Ensembl 72 annotations, using the genomic coordinates of protein-coding genes and the bedtools suit (Quinlan and Hall, 2010) to determine overlaps between them.

Reading frame polymorphism detection and annotation

Polymorphisms were called from quality-filtered, uniquely-mapping transcriptome read alignments (Chapter 2) using samtools mpileup (Li et al., 2009) and vcftools (Danecek et al., 2011). A maximum basewise coverage of 8,000 was allowed, and a minimum coverage of 16. Only exonic, coding regions were screened. Polymorphisms were annotated using the snpEff platform (Cingolani et al., 2012) using a publicly available annotation database corresponding to Ensembl 72 (Flicek et al., 2013).

Statistical analyses

Statistical analyses were performed using basic functions in R (R Core Team, 2012). Phylostratigraphic enrichment was tested using a hypergeometric test. Multiple testing corrections were applied using the Benjamini-Hochberg method, as implemented in R.

Results

Rate differences between genes of different ages

A genome-wide scan of dN/dS ratios was performed as proxy to detect long term selection acting on genes (Yang, 2007). Pairwise calculations were performed between each focal taxon and the mouse reference strain, and the correlation between phylogenetic age and average rate per phylostratum was assessed.

For the five most divergent taxa, namely *Apodemus uralensis*, *Mus matteyii*, *Mus spretus*, *Mus spicilegus* and *Mus musculus castaneus*, there is a significant difference (Mann-Whitney's U, p-value <0.01) between the dN/dS ratios of mouse orphan genes, derived from phylostratum 20, and all other available proteins (Figure 3.1), with mouse orphans having higher dN/dS ratios. This difference becomes less pronounced with decreasing time divergence between the compared taxa, and is non-significant for the *Mus musculus domesticus* and *Mus musculus musculus* populations sampled (not shown).

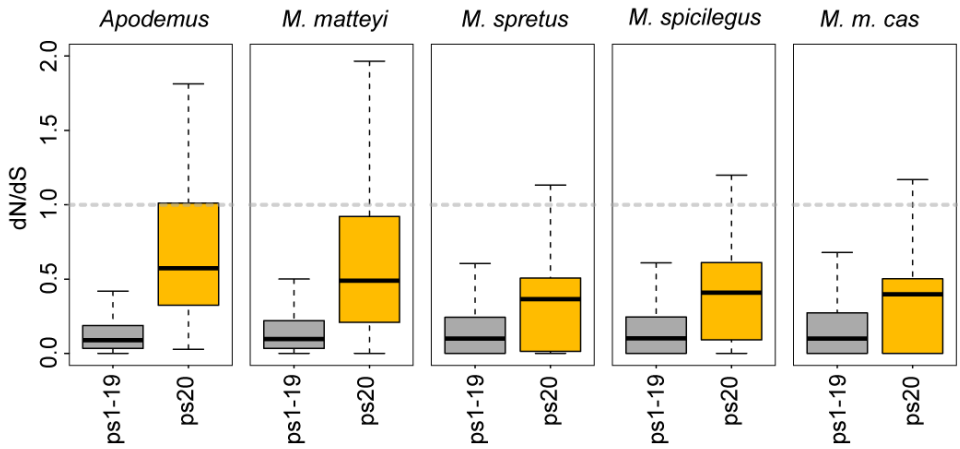

Figure 3.1. Mouse orphan genes have higher rates of evolution than older genes.

Box plot of dN/dS ratios for mouse orphan genes (phylostratum 20; ps20) are significantly higher (Mann-Whitney's U, p-value <0.01) than the rest of the genes in the mouse (phylostrata 1 – 19). Whiskers indicate interquartile ranges. Gray dashed line indicates dN/dS = 1. The shown taxa cover divergences between 0.4 and 10.6 million years (see Chapter 2). Given the differences in sample sizes, the differences between groups were confirmed by permutation tests (10,000 permutations, Mann-Whitney's test). Comparisons between more closely related taxa are not significant (not shown).

Seen in a phylostratigraphic context (Figure 3.2), older genes tend to have on average lower dN/dS values, while younger genes tend to have higher dN/dS values. Again, the effect of this trend is most evident for the longest-diverging pairs, i.e. *Apodemus uralensis* – *Mus musculus* or *Mus mattheyii* – *Mus musculus*, and becomes less evident with decreasing phylogenetic distance to the reference, until it becomes non-significant for the three *Mus musculus domesticus* populations, i.e. the closest wild relatives of the laboratory strain.

For genes that have arisen before the Boreoetherian divergence, approximately 100 million years ago (ps1 – ps16), the rate-age correlation becomes stronger, and highly significant even for the closest populations. This can be understood as the difference in the selective constraints on these genes which is even visible for divergences as short as those between the European and American *M. m. domesticus*.

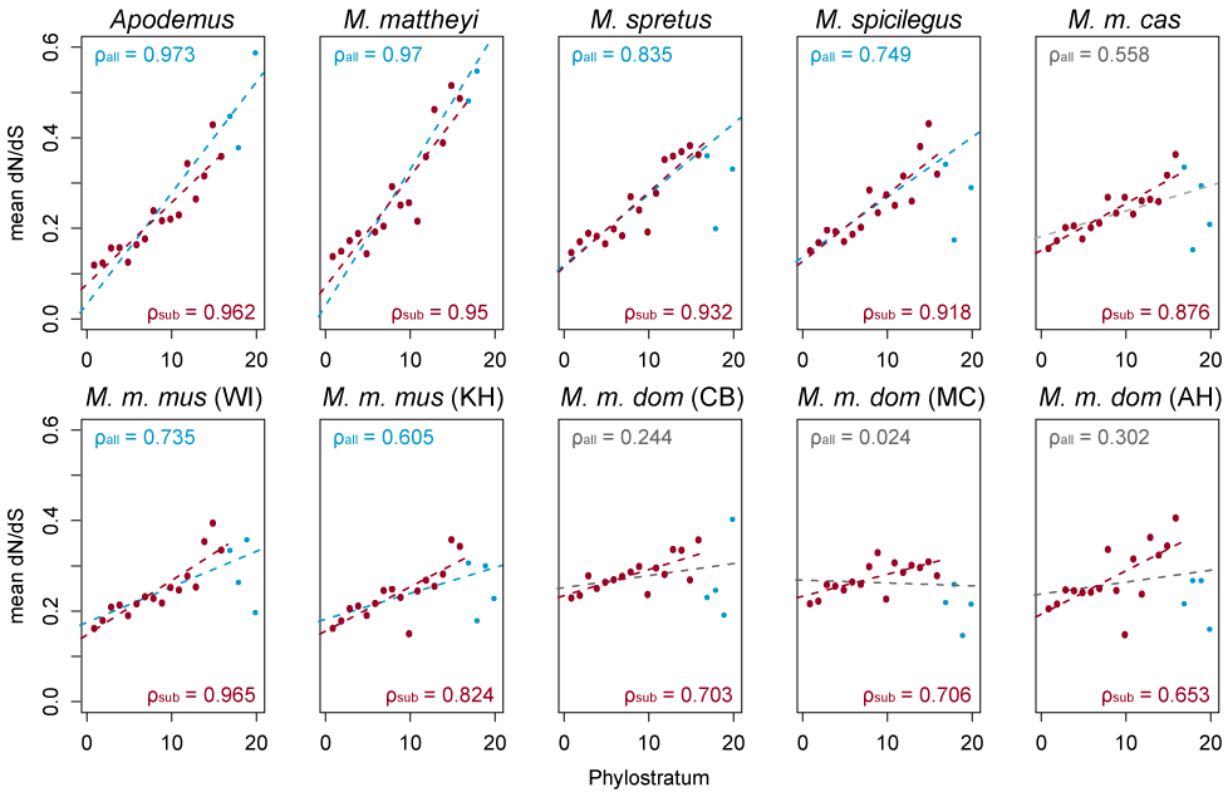

Figure 3.2. The evolutionary rate is strongly correlated with phylogenetic age but the strength of the correlation depends on the divergence time between the compared species.

Dots represent mean dN/dS value per phylostraum for each taxon. Red dots indicate phylostrata 1 – 16 (origin of life - Eutheria); blue dots indicate phylostrata 17 – 20 (Boreutheria – mouse). Blue dashed line and blue text at the upper left corner indicates the correlation between phylostratum and mean dN/dS covering all phylostrata (red and blue dots). Red dashed line and red text at the lower right corner indicates the correlation between phylostratum and mean dN/dS containing only the first 16 phylostrata (red dots). Red and blue text indicate significant Spearman rank correlations ($p < 0.01$, FDR corrected). Grey text indicates non-significant correlations.

Overlapping genes are an unlikely source of bias

The tests of selection which depend on the number of substitutions at synonymous and non-synonymous sites tend to have biased estimates whenever the locus includes more than one reading frame, either on sense or antisense orientation (Rancurel et al., 2009; Sabath et al., 2012). Synonymous sites are assumed to accumulate mutations mostly neutrally and the non-synonymous sites according to the selective constraints on the protein. However, when two overlapping reading frames exist, allowed mutations which would result in synonymous changes for one reading frame could lead to deleterious mutations for the other, limiting the mutational trajectory at the locus. In order to test whether potential overlaps might bias the previously shown trends, I categorized the protein-coding genes between genes with known overlaps and genes without.

As expected, the two classes of genes have significantly different synonymous substitutions (Wilcoxon rank sum test, p<0.01) in each of the taxa sampled, confirming that there is an overall bias from the overlaps in the computation of the evolutionary rates (Figure 3.3).

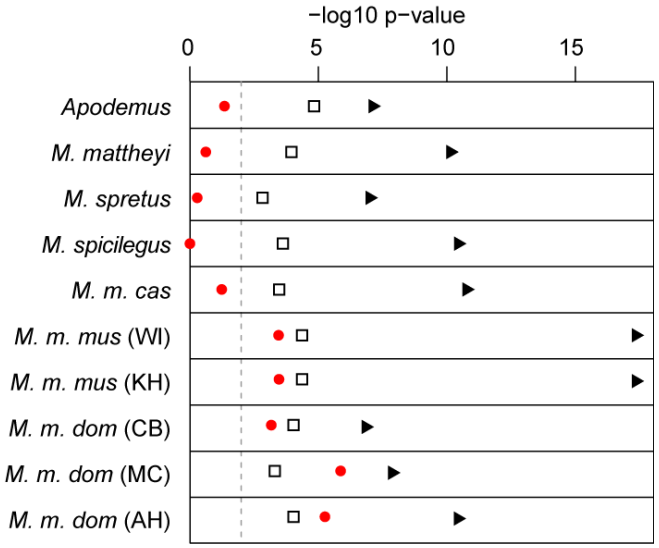

Figure 3.3. Overlapping genes produce distorted synonymous substitution rates when compared to non-overlapping genes.

Summary of FDR-corrected p-values from Wilcoxon rank sum tests comparing evolutionary rates (open squares), non-synonymous substitutions (red circles) and synonymous substitutions (black triangles) between overlapping and non-overlapping genes, indicated as –log10 p-value. Dashed line indicates a corrected p-value of 0.01.

In a phylostratigraphic context, only the synonymous substitution rates of genes in the oldest two phylostrata are significantly different between overlapping and non-overlapping genes, and this only has a detectable impact at the overall rate of evolution in the oldest phylostrata among the populations of *Mus musculus musculus* (Figure 3.4). From this I infer that the biases generated by overlapping genes are unlikely to play a role in the trends described above.

However, it is possible that those genes which bear stronger selective pressure experience a much marked constraint on the onset of an overprinting event (Sabath et al., 2012). Complementary to this, younger genes which would have less constraints could have more favorable conditions for the acquisition of new reading frames.

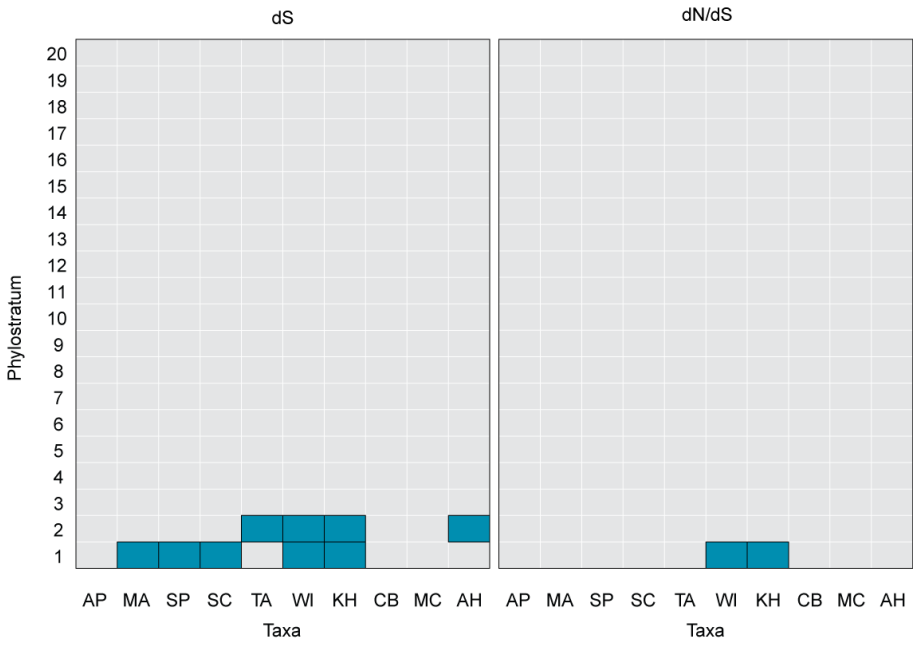

Figure 3.4. The distortion of substitutions due to in overlapping genes compared to non-overlapping genes is mostly limited to synonymous substitutions in very old genes.

Summary of p-values from Wilcoxon rank sum tests comparing synonymous substitutions (left) and evolutionary rates (right) between overlapping and non-overlapping genes for each phylostratum. Blue cells indicate significant differences between overlapping and non-overlapping genes at the given phylostratum and for a given species comparison (p-value <0.01, FDR corrected). Non-synonymous substitutions are not presented because all comparisons were non-significant.

Impact of reading frame polymorphisms across phylogenetic time

In the first chapter, I presented a strong correlation between the phylogenetic age of a gene and structural properties of genes, such as lengths, the number of exons, introns and protein domains. In order to further expand our notions of how genes change over time, I annotated polymorphisms obtained from transcriptome reconstructions from wild mice.

The polymorphism annotations were done at the protein level, taking into account codon sequences represented in the sequencing read alignments. Since the transcriptomes sequenced include only three distinct tissues (liver, brain and testis) and include only eight unrelated individuals of different ages, it is possible that some genes could remain undetected. For this reason, polymorphic genes, i.e. genes having at least one specific type of polymorphism, were normalized per phylostratum by the number of genes for which at least one synonymous substitution was detected. Hence, analyses shown hereafter are based on reliably expressed orthologs for which polymorphisms of interest as well as synonymous substitutions were available.

I explored the five most relevant types of polymorphisms which could have an impact on the presence or absence of an open reading frame: frameshift mutations and gains and losses of start and stop codons.

It would be expected that if no relevant enrichment was present across genes of different ages, that older genes would accumulate more such mutations, provided the probability of mutation is length dependent and given that older genes have longer reading frames (Neme and Tautz, 2013).

However, younger genes show a recurrent enrichment in all five types of mutations (Figures 3.4 and 3.5), but the patterns at the population, species, and genus level are slightly different between the types of polymorphism.

For example, frame shift mutations are most common in younger genes at the population and species level (Figure 3.5A), while older genes tend to be more enriched in frame shifts for the most distant species (*M. mattheyi* and *A. uralensis*).

Stop codon gains are more frequent than losses (Figure 3.5B-C), and both are consistently enriched in young genes. This effect seems to become stronger as the divergence between species increases. A similar pattern is observed for the loss of start codons (Figure 3.5E). The gain of alternative start codons seems to be a frequent feature of young genes on close phylogenetic divergences, but becomes widespread across the genome as divergence increases (Figure 3.5D).

Furthermore, the frequency of polymorphisms per gene seems to be very high for the youngest genes, and shows significant correlations with age at the most distant divergences (Figure 3.6).

These trends are consistent with the prediction that reading frames for young genes in the genome are rather unstable and very dynamic, being able to accumulate mutations that disrupt existing reading frames much faster than older genes.

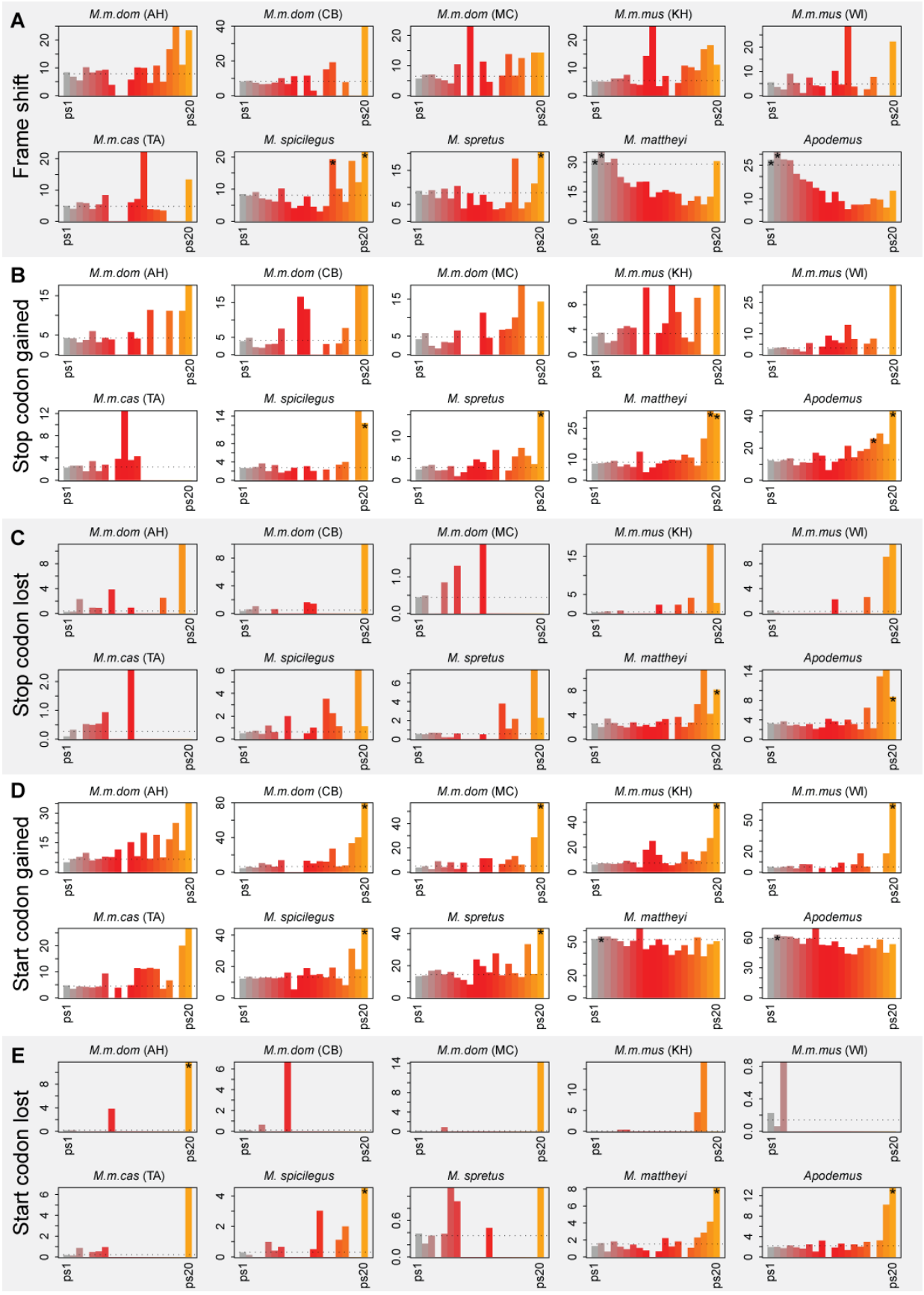
A
Frame shift
B
Stop codon gained
C
Stop codon lost
D
Start codon gained
E
Start codon lost
M.m.dom (AH)
M.m.dom (CB)
M.m.dom (MC)
M.m.mus (KH)
M.m.mus (WI)
M.m.cas (TA)
M. spicilegus
M. spretus
M. mattheyi
Apodemus
ps1
ps20

Figure 3.5. Young genes tend to accumulate more frame disrupting mutations.

X-axis represents phylostratum and y-axis represents the percentage of genes with polymorphisms occurring per phylostratum versus the total number of genes with synonymous substitutions per phylostratum. A. Frame shift mutations. B. Premature stop codon gains. C. Loss of stop codons. D. Gain of alternative start codons. E. Loss of start codons. The dotted line indicates the genome-wide ratio. Asterisks highlight the phylostrata with significant enrichment (hypergeometric test, $p < 0.01$, FDR corrected).

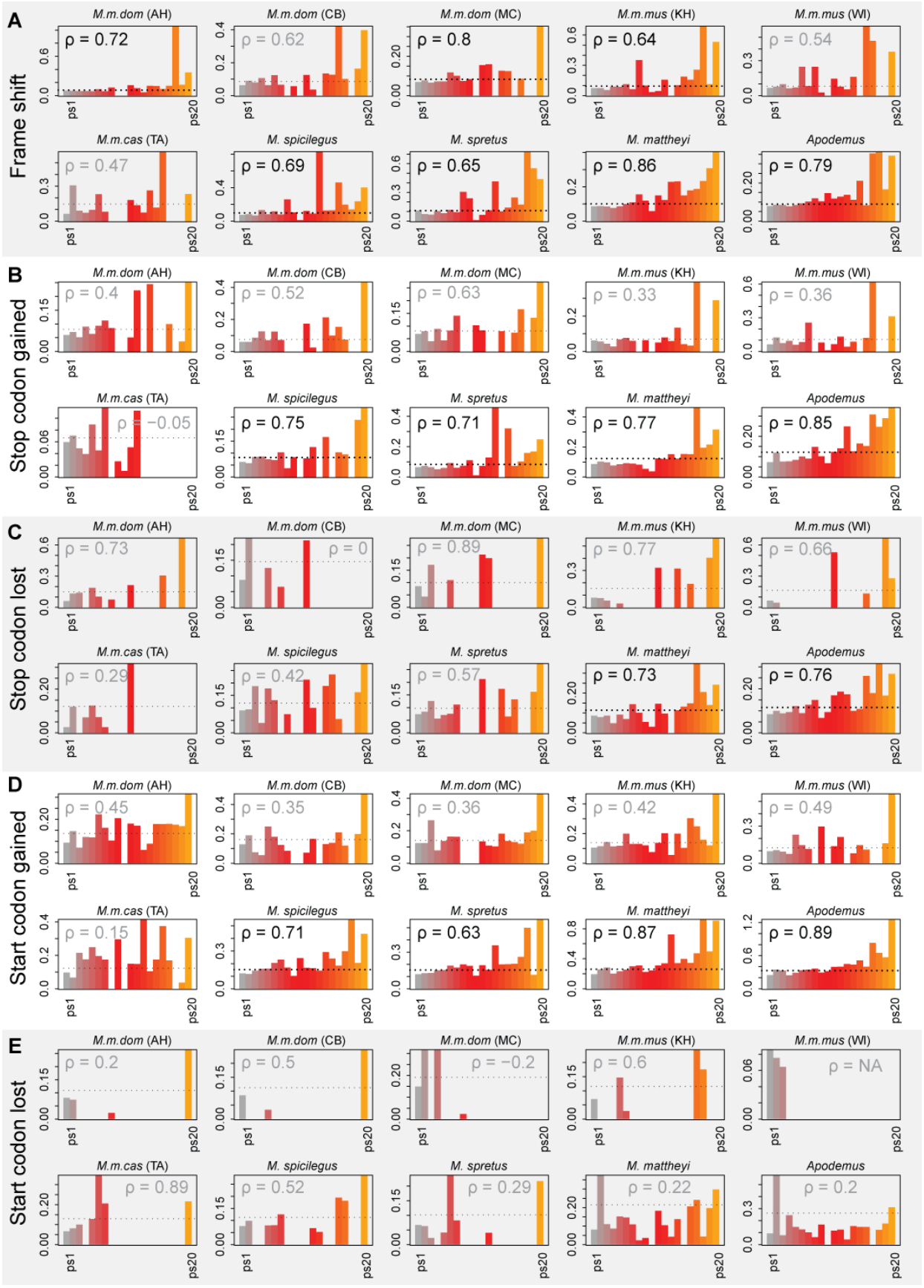
A
Frame shift
M.m.dom (AH)
ρ = 0.72
M.m.dom (CB)
ρ = 0.62
M.m.dom (MC)
ρ = 0.8
M.m.mus (KH)
ρ = 0.64
M.m.mus (WI)
ρ = 0.54
M.m.cas (TA)
ρ = 0.47
M. spicilegus
ρ = 0.69
M. spretus
ρ = 0.65
M. mattheyi
ρ = 0.86
Apodemus
ρ = 0.79
B
Stop codon gained
M.m.dom (AH)
ρ = 0.4
M.m.dom (CB)
ρ = 0.52
M.m.dom (MC)
ρ = 0.63
M.m.mus (KH)
ρ = 0.33
M.m.mus (WI)
ρ = 0.36
M.m.cas (TA)
ρ = −0.05
M. spicilegus
ρ = 0.75
M. spretus
ρ = 0.71
M. mattheyi
ρ = 0.77
Apodemus
ρ = 0.85
C
Stop codon lost
M.m.dom (AH)
ρ = 0.73
M.m.dom (CB)
ρ = 0
M.m.dom (MC)
ρ = 0.89
M.m.mus (KH)
ρ = 0.77
M.m.mus (WI)
ρ = 0.66
M.m.cas (TA)
ρ = 0.29
M. spicilegus
ρ = 0.42
M. spretus
ρ = 0.57
M. mattheyi
ρ = 0.73
Apodemus
ρ = 0.76
D
Start codon gained
M.m.dom (AH)
ρ = 0.45
M.m.dom (CB)
ρ = 0.35
M.m.dom (MC)
ρ = 0.36
M.m.mus (KH)
ρ = 0.42
M.m.mus (WI)
ρ = 0.49
M.m.cas (TA)
ρ = 0.15
M. spicilegus
ρ = 0.71
M. spretus
ρ = 0.63
M. mattheyi
ρ = 0.87
Apodemus
ρ = 0.89
E
Start codon lost
M.m.dom (AH)
ρ = 0.2
M.m.dom (CB)
ρ = 0.5
M.m.dom (MC)
ρ = −0.2
M.m.mus (KH)
ρ = 0.6
M.m.mus (WI)
ρ = NA
M.m.cas (TA)
ρ = 0.89
M. spicilegus
ρ = 0.52
M. spretus
ρ = 0.29
M. mattheyi
ρ = 0.22
Apodemus
ρ = 0.2
ps1
ps20

Figure 3.6. Younger genes tend to accumulate more frame disrupting mutations per unit of coding length.

X-axis represents phylostratum and y-axis represents mean number of polymorphisms occurring per 100 bases of coding length per phylostratum. A. Frame shift mutations. B. Premature stop codon gains. C. Loss of stop codons. D. Gain of alternative start codons. E. Loss of start codons. The dotted line indicates the genome-wide mean value. Inset text indicates the correlation coefficient (Spearman's ρ), in black print when the correlation is significant ($p < 0.01$, FDR corrected) and gray print when non-significant.

Discussion

I present here two arguments to the idea that recently acquired genes in the mouse lineages face less strong selective constraints than other older genes, and that this results in a higher chance of an open reading frame distortion, which can be indicative of early decay at the protein-coding level.

New genes have significantly higher dN/dS ratios than older genes. This has been associated before with fast evolution, which could be directly correlated with a constantly changing environment, especially in the context of reproductive isolation (Haerty et al., 2007; Jagadeeshan and Singh, 2005), development (Domazet-Loso and Tautz, 2003) or transcription factor interaction dynamics (Stefflova et al., 2013). This holds true for new genes which quickly become essential (Chen et al., 2010; Reinhardt et al., 2013). However, the number of genes for which complete functional characterization is still small, thereby hampering the predictions for genome-wide analyses.

An alternative to the hypothesis of fast evolution induced by a constantly changing environment, the evolution after *de novo* gene birth could be driven by increasing gains of selective constraints (Albà and Castresana, 2005) due to optimization of the protein during a progressive integration into cellular networks (Abrusán, 2013). This process would see the transition from large numbers of nearly-neutral protogenes to a smaller number of adaptively relevant new genes. These would then be able to persist for longer times in the genome, getting increasingly integrated into the cellular networks.

Newly-acquired genes are more likely to be lost at the protein level (Palmieri et al., 2014) than older genes. Furthermore, the total number and strength of protein-protein interactions are positively correlated with age (Abrusán, 2013), as well structural stability and robustness to mutations (Abrusán, 2013; Toll-Riera et al., 2012), supporting the inference that genes would increase in selective constraints as they become older. These are further arguments to the

notion that most genes start as simple entities, and increase in complexity over time (Neme and Tautz, 2013; Toll-Riera et al., 2012). The results presented here are also consistent with differential selective constraints between genes of different ages (Figure 3.2), which would result in new genes being more prone to loss of protein-coding potential (Figures 3.4 and 3.5).

A negative correlation between rate and age has been previously described, based on less fine-grained age measurements and using species with much longer divergences (e.g. human and mouse) (Albà and Castresana, 2005; Toll-Riera et al., 2012).

It has been theoretically derived from population genetics models that dN/dS ratios are sensitive to the divergence between species, and that at very short divergences the estimation can become noisy (Kryazhimskiy and Plotkin, 2008; Peterson and Masel, 2009). It has been hypothesized that such noise comes from ancestral polymorphisms shared by the compared taxa or from slightly deleterious mutations (Peterson and Masel, 2009).

Consequently, for closely related lineages dN/dS values can yield results which could be difficult to compare and interpret with those obtained from deeply divergent lineages (Mugal et al., 2014; Peterson and Masel, 2009; Rocha et al., 2006). However, the correlation between age and rate is evident for the comparisons including more divergent lineages. This signal is not completely lost at short divergences. Analyses excluding very young genes (phylostrata 17 – 20; below 100 million years) yield the expected correlation, even for the comparisons including European *Mus musculus domesticus*.

This can be considered yet another argument in favor of the differences in selective constraints of genes of different ages, as it has been shown that dN/dS can successfully detect negative selection between closely related lineages, provided that the selection intensity is large enough (Kryazhimskiy and Plotkin, 2008).

Nevertheless, this still remains an issue to be addressed in the future through more detailed and specific theoretical information, since the current framework notes the limitations between comparing sequences from the same population and comparing samples from deeply divergent species (Kryazhimskiy and Plotkin, 2008; Mugal et al., 2014; Rocha et al., 2006). In my case, the closest comparisons fall within an inconvenient intermediate problem: On the one side, the reference genome clearly does not come from any of the *Mus musculus* populations sampled, so one can be sure that the dynamics observed are not due to randomly picking two highly related individuals; on the other side, the divergence between those lineages is still very small

so that it is possible that low selection coefficients and neutrality are virtually indistinguishable in terms of accumulated mutations.

Conclusion

The results presented here highlight the continuum between the features of old and young genes, consistent with the notion that completely new genes appear, and have different levels of complexity compared to already existing genes. At a given time point, different selective constrains should operate on genes of different age, probably because of their differential functional integration, and this is evidenced as slower rates of evolution for older genes and higher likelihood of becoming lost for younger genes.

Acknowledgements

I thank Chen Ming for her helpful assistance with the snpEff annotation pipeline and Chengjun Zhang for support with the gKaKs pipeline.

Chapter 4: A transcriptomics approach to the gain and loss of *de novo* genes in mouse lineages

Introduction

How is a gene made?

The model of *de novo* gene birth through protogenes states that functional genes can be generated stochastically from non-genic regions of the genome (Carvunis et al., 2012). The fundamental underlying assumption is that at the sequence level, genes can be explained as a collection of small motifs responsible for the stable presence of a product (Figure 4.1) (Tautz and Domazet- Loso, 2011). Following this assumption, at a given point in time, there should be a number of small motifs scattered across the genome, with a fraction of those sites potentially arranged to produce stably transcribed, and even translated, protogenes. These protogenes can eventually become subjects of selection, provided their products have an impact on the organismal fitness (Carvunis et al., 2012; Reinhardt et al., 2013).

This model does not explicitly require that these motifs appear at the same time, but rather that they coexist during a given window of time (Heinen et al., 2009; Tautz and Domazet- Loso, 2011). This means that these sites can appear and disappear at different rates, most likely associated with the length and complexity of the motif. Sites with potentially relevant small motifs in regions lacking complete genes have been dubbed cryptic functional sites (Tautz and Domazet- Loso, 2011) since many of them will be present before their effect is visible. The effect only makes sense in combination with other motifs, e.g. the peptide in an open reading frame can only be synthesized from a properly transcribed and processed messenger RNA.

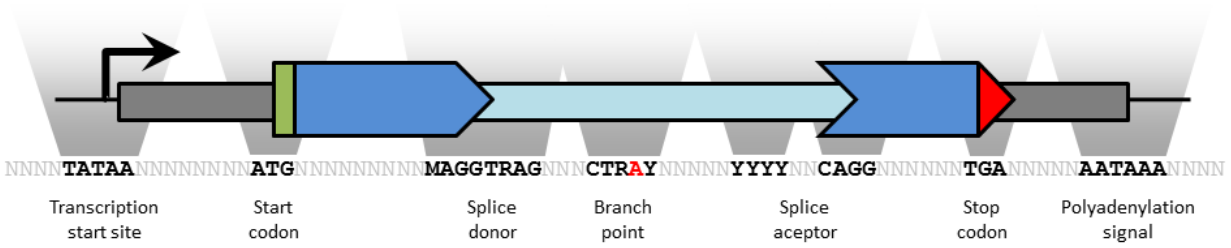

Figure 4.1. Simplified scheme of the gene as a collection of small motifs.

Depiction of motifs playing a role for the generation of transcripts and proteins. Transcription start sites (black arrow) can be summarized as the sequence patterns which lead to the association of RNA polymerases (mainly RNApol II in eukaryotes), responsible for the production of primary transcripts. The start and stop codons (green and red regions) encompass the exonic region of the transcript (blue) able to produce a protein after successful association with ribosomes. The splice donor, splice acceptor and branch points are required for efficient splicing of introns (light blue region). Gray regions represent untranslated portions of the exons. A polyadenlylation signal is necessary for the stability of a transcript. Furthermore, the stability of the transcript also depends on RNA surveillance mechanisms which control the products of splicing and translation. Several of these motifs are not restricted to a unique sequence that fulfills its purpose, e.g. stop-codons or polyadenylation signals. Also, I indicate motifs associated with splicing and translation, but neither is an absolute requirement for a functional gene. For example, long intergenic non-coding RNAs (lincRNAs) can be functional without producing proteins, and single exon transcripts can also produce proteins.

As a consequence, the frequency of available protogenes at a given point in time could be interpreted as a function of the length of the genome, the complexity of each cryptic site, together with the probability of co-occurrence of those cryptic sites. Over evolutionary scales, there would be a number of regions which are only a few mutations away from completing the requirements for becoming a protogene.

Once a protogene exists, its fate will be determined by the interplay between what it can effectively do at the molecular scale and the adaptive potential of that function at an organismal level. So far, the functional value of random peptides is an open subject, whether they can stably fold or exist as intrinsically disordered, and how their presence can effectively contribute to the fitness of the organism. However, one can initially assume that the earliest phase of gene emergence through protogenes is mostly neutral, being the product of either stochastic associations of the molecular machinery responsible for the production of transcripts and peptides, or spontaneous combination of cryptic sites.

The early phase of new gene emergence

Genome-wide evolutionary analyses of the origin of genes along a phylogeny, also known as phylostratigraphic analyses, have contributed to the idea that the processes and rates of gene gain are very dynamic, with periods of increased gene gain surrounding relevant functional and morphological transitions in the phylogenetic history (Domazet- Loso and Tautz, 2010a; Domazet- Loso et al., 2007; Khalturin et al., 2009; Neme and Tautz, 2013; Tautz and Domazet-Loso, 2011; Toll-Riera et al., 2009; Wissler et al., 2013). This process relies on the completeness of the gene repertoires in each of the outgroups to generate accurate predictions, and therefore the most recently acquired genes we have been able to identify through this method can be up to several million years old. In the case of the house mouse, the most recently acquired genes are those absent in the rat or other rodents (Neme and Tautz, 2013), a divergence estimated to have taken place between 8 to 50 million years ago (Benton and Donoghue, 2007; Pereira and Baker, 2006; Pyron, 2010; Robins et al., 2010). Currently, the repository TimeTree.org, which uses combined information from multiple sources, estimates the divergence at 25.5 million years ago (Hedges et al., 2006; Kumar and Hedges, 2011).

It has been previously shown that the estimated rate of gene gain at the youngest divergence is very high, usually the highest detected along the whole phylogeny (Tautz and Domazet- Loso, 2011; Wissler et al., 2013). Conversely, the gains during the immediately preceding divergences are usually low, indicating that most of the genes belonging to the most recent divergence will not be retained in future divergences (Palmieri et al., 2014). A dynamic equilibrium is thought to be established between gene gain and loss, in which many new genes are continuously generated, but only a fraction is able to acquire functions and stay over long periods of time.

It is possible to speculate that most new genes will not be able to gain a function before the decay of their gene-like properties occurs (Palmieri et al., 2014), unless a major shift in the fitness landscape allows new random peptides to contribute in the slightest way to the organismal fitness. This is one interpretation of the correlation observed at macroevolutionary scales between gene gain events and major ecological shifts (Colbourne et al., 2011; Khalturin et al., 2009; Tautz and Domazet- Loso, 2011).

As mentioned in the previous chapter, many mouse-specific genes show signs of reading frame instability, and a few already show signatures of selection (purifying and positive) when evaluated in divergences as recent as 10.6 and 6.6 million years old (Lecompte et al., 2008), indicating that at these timescales we can observe both processes of gene birth and early decay.

Pervasive transcription and junk-DNA as raw material for new genes

One of the “pandora’s boxes” opened by the advances in genome analysis technologies was the finding that more of the genome was transcribed than apparently needed to generate protein-coding genes and most known types of RNA genes (Berretta and Morillon, 2009; Jensen et al., 2013; Kapranov et al., 2007). This was initially suggested based on data from tiling arrays, which could show the transcriptional properties of genomic regions by using arrays of probes of contiguous location in the genome, or tiles, and hybridizing them with RNA (or cDNA) (Bertone et al., 2004). However, the largest estimates came with the development of high-throughput transcriptome sequencing, which suggested that not less than 93% of a genome could have transcription of some sort (either regulated or spurious) when assessed across multiple organs, tissues and cell types (Clark et al., 2011; ENCODE Project Consortium et al., 2007; Kapranov et al., 2007).

Among the discoveries that followed this phenomenon were the long, often polyadenlyated, and spliced macroRNAs (to differentiate from microRNAs), now known as long intergenic non-coding RNAs, or lincRNAs (Marques and Ponting, 2009). Many examples show that these lincRNAs are functional in almost all molecular processes and under selective constraints (Kung et al., 2013; Managadze et al., 2011; Necsulea et al., 2014), but this still remains a largely understudied subject and the question of whether all of the detected transcription serves a purpose is still unsolved (van Bakel et al., 2011).

After the final publication of the ENCODE project, it was proposed that over 80% of the human genome could be functional based on combined information from transcriptomes and other types of genome-wide scans such as histone-methylation patterns through chromatin immunoprecipitation and sequencing (ChIP-seq) (ENCODE Project Consortium, 2012). This has sparked some discussion because the biochemical activity of these regions is not necessarily indicative of a functional association, and much of that pervasive activity could be pushed into the realm of the junk-DNA (Graur et al., 2013). The term junk-DNA encompasses those regions of the genome which have no detectable function, and which do not contribute in a detectable way to the overall fitness of the organism (i.e. have no selective advantage) yet evolve according to neutral processes and remain present in the genome for large portions of its evolutionary history (Ohno, 1972).

In the context of gene birth, “transcribed junk-DNA” or “junk-RNA” sequences are key players, with advantage in terms of their potential to become genes compared to the rest of the genome, so much that these can be considered some type of protogenes.

Stable transcription seems to be the most necessary requirement of protogenes, since according to our working definition a gene can be either coding or non-coding, but it cannot be non-transcribed. For this reason, it is important to explore and understand the dynamics of gain and loss at the transcriptional level.

The present study is an approximation using comparable levels of transcriptome sampling along a phylogeny, without any assumptions about the structure of genes, other than the patterns that can be analyzed directly from genome-wide expression. Through this, I expect to expand our current knowledge of transcriptome divergence in mouse lineages, the role of the emergence of new transcripts in the acquisition of new genes, and provide a suitable framework for the quantification of the rates of *de novo* transcript emergence.

Methods

For a description of the high-throughput data generation and processing see Chapter 2.

Transcriptome presence/absence matrix and mapping of gains and losses

The expression information was binarized into a presence/absence matrix following these criteria: A presence is counted if the normalized fragment count is larger than 50, and an absence is counted if the normalized fragment count has less than 10 reads. If the count is intermediate, the region is discarded on the basis of uncertainty. The lower end of the criteria was derived from the assumption that the noise in read counts is Poisson distributed, which was confirmed by repeated sampling from non-overlapping, non-genic regions of the genome. A rate parameter of 4 reads was approximated, resulting on a minimum of 10 reads per region to be significantly different from the expected noise at a p-value of 0.01. To increase stringency and to control for synteny, only regions which were mappable based on genomic reads information were considered.

Transcripts showing presence-absence variation along the tree were kept as candidates, and explored manually using the IGV browser (Thorvaldsdóttir et al., 2013). Candidates shorter than 300 bases were excluded from downstream analyses. Absence of expression in *Apodemus* was selected as mandatory for a candidate to pass the final cutoff.

Manually curated candidates were used for presence-absence pattern analyses. The transcriptome matrix was analyzed assuming maximum parsimony with Gloome (Cohen and Pupko, 2011; Cohen et al., 2010), giving equal weights for gains and losses of characters, and using a phylogenetic tree which describes approximate divergence times between the sampled taxa (Figure 2.1). The resulting gain and loss patterns were explored with the ape package (Paradis et al., 2004). The rates of gain and loss were fitted using the R base function lm(), and he confidence intervals were identified with the function predict() (R Core Team, 2012), plots were generated with base packages and ggplot2 (Wickham, 2009).

For known genes, the annotation was obtained from Ensembl 74 (Flicek et al., 2013). Functional association analyses were performed using the DAVID online platform (Huang et al., 2009a, 2009b).

Results

How much of the mouse genome has evidence of transcription?

Using maximum parsimony in combination with presence and absence of transcription across the genome, it was possible to classify regions as having lineage specific transcription, lineage specific absence of transcription, or with patterns of recurrent gain or loss. I decided to focus on the mappable regions of the reference genome, using information from *de novo* genome sequencing as an empiric measure of mappability. Those regions for which genomic and transcriptomic reads can be reliably mapped are considered as present. The percentages are shown scaled to the estimated genome size of the mouse reference sequence (Table 4.1).

These estimates are done directly from the sequencing read information, and no assumptions were made about the genic nature of the detected transcripts. According to Ensembl (version 74), the house mouse has 39,179 genes; including 22,740 protein-coding genes, 5,945 pseudogenes, 1,795 lincRNAs, 2,010 microRNAs, and 6,689 of other biotypes. This annotation has an exonic coverage of 3.28%, and a total transcriptional coverage of 38.62% from the reference genome.

Across all taxa sampled here, the average base-wise transcriptome coverage per species of the reference genome is 52% (testis 41%, brain 37%, liver 21%), with 1.1% being the average portion with specific expression for single taxa (testis 1.3%, brain 1.1%, liver 1.0%) and 10.7% of the genome being lineage-specific expression (testis 13%, brain 11.2%, liver 9.5%).

Interestingly, 81% of the available mouse genome was found to have expression when combining all the species. The remaining 19% of the mouse genome had no detectable expression in any species. This can be considered a phylogeny-wide coverage estimate. 25% of the coverage (testis 14%, brain 19%, liver 9%) is present across all taxa, i.e. has been invariably conserved over the last 10 million years.

Closely related species of *Mus musculus* show the highest coverage, compared to species with increasing phylogenetic distance (Table 4.1), while the most distant species have the highest species-specific coverage (Table 4.2).

Table 4.1. Transcriptome coverage statistics by taxon and tissue

tissue	*M. m. dom* (Iran)	*M. m. dom* (Fra)	*M. m. dom* (Ger)	*M. m. mus* (Kaz)	*M. m. mus* (Aus)	*M. m. cas* (Tai)	*M. spc*	*M. spr*	*M. mat*	*Apo*	Conserved	Total
all	57%	56%	56%	56%	55%	55%	53%	53%	42%	38%	25%	81%
testis	47%	45%	44%	45%	44%	44%	41%	43%	29%	25%	14%	73%
liver	22%	22%	22%	21%	22%	22%	22%	22%	19%	17%	9%	38%
brain	38%	39%	40%	40%	40%	39%	39%	38%	32%	28%	19%	58%

Table 4.2. Coverage specific to each taxon and tissue

tissue	*M. m. dom* (Iran)	*M. m. dom* (Fra)	*M. m. dom* (Ger)	*M. m. mus* (Kaz)	*M. m. mus* (Aus)	*M. m. cas* (Tai)	*M. spc*	*M. spr*	*M. mat*	*Apo*	Exclusive
all	1.0%	0.9%	0.9%	0.9%	0.8%	0.9%	1.1%	1.2%	1.1%	2.0%	10.7%
testis	1.4%	1.0%	1.0%	1.1%	1.0%	1.1%	1.5%	1.7%	1.3%	1.8%	13.0%
liver	0.7%	0.7%	0.7%	0.6%	0.6%	0.7%	0.8%	1.0%	1.4%	2.3%	9.5%
brain	0.7%	0.9%	0.9%	1.0%	0.9%	1.0%	1.1%	1.0%	1.2%	2.7%	11.2%

Genome-wide transcription: gain and loss dynamics

According to maximum parsimony estimates, approximately 55% of the common genome of the analyzed species has experienced transcriptome presence/absence changes in the last 10 million years (Figure 4.2). This also indicates that the coverage of regions with loss of transcription is at least threefold smaller than those regions which show gain of transcription (Figure 4.2 B-C).

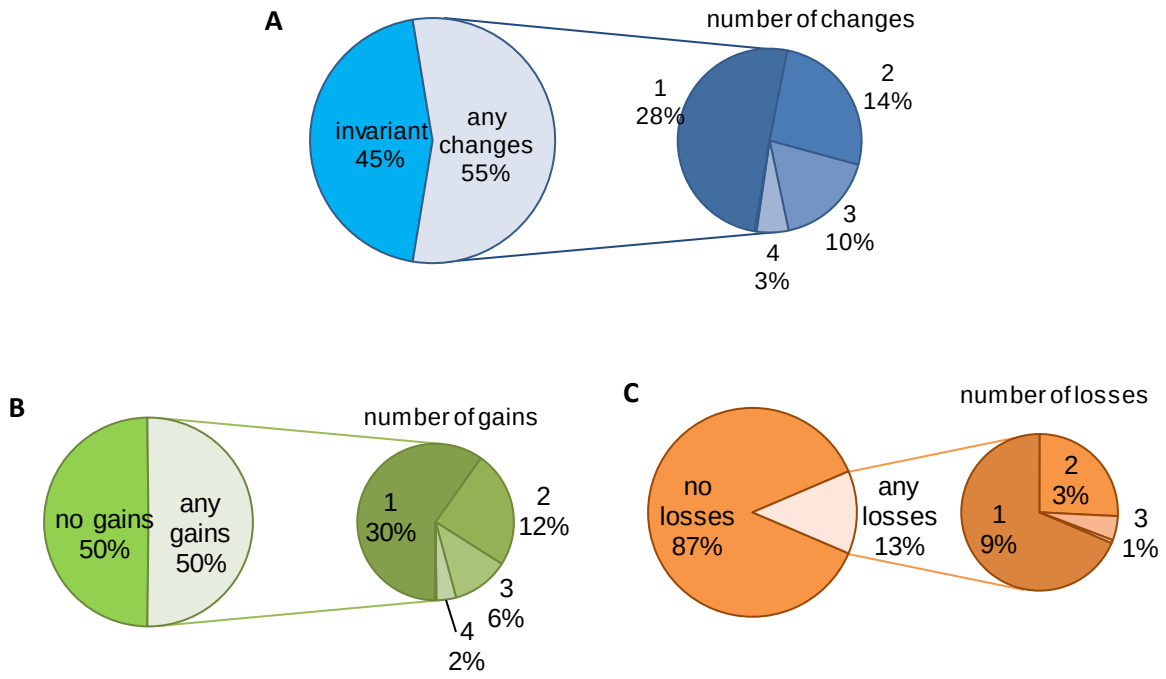

Figure 4.2. The mouse genome has undergone more gains of expression than losses over the last 10 million years.

Maximum parsimony quantification of changes (A), gains (B) and losses (C) based on the transcribed genome along a 10.6 million year phylogeny. Changes, gains and losses can be further divided into regions which show multiple events (represented to the left of each chart). The majority of gains and losses are single events, and the proportion of the genome which has gained expression is at least three fold larger than the percentage of the genome which has lost expression at some point.

Phylogenetic patterns in genome-wide transcription

Pairwise analyses of transcription coverage indicate a strong phylogenetic association, in which the coverage shared by two groups is correlated with the phylogenetic distance between them (Table 4.3). A neighbor-joining reconstruction of the pairwise comparisons returns a tree with a topology similar to the one described for the species.

Table 4.3. Transcriptome coverage correlates with phylogenetic signals.

Percentage of transcribed genome shared by any two taxa. The diagonal describes the percentage of coverage in each individual taxon. The standard deviation for the comparisons was calculated using equidistant taxa. *M. spc*: *Mus spicilegus*; *M. spr*: *Mus spretus*; *M. mat*: *Mus mattheyi*; *Apo*: *Apodemus uralensis*.

	***M. m. dom* (Iran)**	***M. m. dom* (Fra)**	***M. m. dom* (Ger)**	***M. m. mus* (Kaz)**	***M. m. mus* (Aus)**	***M. m. cas* (Tai)**	***M. spc***	***M. spr***	***M. mat***	***Apo***
M. m. dom (Iran)	57%	50%	50%	49%	48%	48%	46%	47%	37%	32%
M. m. dom (Fra)	-	56%	50%	48%	48%	48%	46%	46%	37%	32%
M. m. dom (Ger)	-	-	56%	48%	48%	48%	46%	46%	37%	32%
M. m. mus (Kaz)	-	-	-	56%	49%	48%	46%	46%	37%	32%
M. m. mus (Aus)	-	-	-	-	55%	48%	46%	46%	37%	31%
M. m. cas (Tai)	-	-	-	-	-	55%	46%	46%	37%	31%
M. spc	-	-	-	-	-	-	53%	45%	36%	31%
M. spr	-	-	-	-	-	-	-	53%	36%	31%
M. mat	-	-	-	-	-	-	-	-	42%	29%
Apo	-	-	-	-	-	-		-	-	38%
Std. dev.			0.040%	0.144%	0.528%	0.170%	0.230%	0.382%	1.715%	0.975%

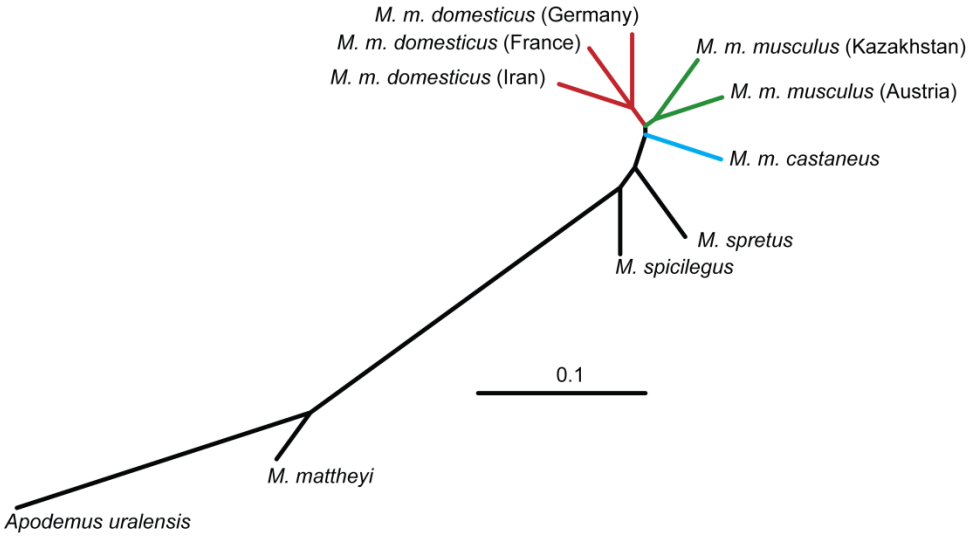

Figure 4.3. Transcriptome coverage correlates with phylogenetic signals.

Neighbor-joining representation of the matrix in Table 4.3, yields a topology that describes the phylogenetic relationships between the sampled taxa (Figure 2.1). The only exception can be seen for the *M. spicilegus* / *M. spretus* divergences, as *M. spicilegus* is assumed to have diverged more recently from the house mouse than *M. spretus*.

How much of the genome is transcribed in a lineage specific way?

Lineage-specific expression was computed as gains which seem to have occurred for the first time during each branch along the phylogeny (Table 4.4 and Figure 4.4). Again, I find consistent asymmetries in the proportion of gained and lost transcriptome coverage at this divergence scale, in which gains are much more frequent than losses, both as lineage specific and as non-lineage specific events.

Table 4.4. Distribution of lineage specific gains and losses of coverage.

Lineage specific gains (percentage to the left) and losses (percentage to the right) of transcriptomic coverage at each phylogenetic divergence.

<table>
<tr><td colspan="2">M. m. domesticus
Iran
0.96% / 0.07%</td><td rowspan="3">M. m.domesticus
0.31% / 0.01%</td><td rowspan="6">Mus musculus
0.22% / 0.01%</td><td rowspan="7">M. musculus
+
M. spicilegus
0.47% / 0.02%</td><td rowspan="8">M. musculus
+
M. spicilegus
+
M. spretus
0.51% /0.06%</td><td rowspan="9">Mus
4.56% / -</td></tr>
<tr><td>M. m. domesticus
France
0.86% / 0.07%</td><td rowspan="2">M. m. domesticus
Europe
0.25% / 0.02%</td></tr>
<tr><td>M. m. domesticus
Germany
0.85% / 0.06%</td></tr>
<tr><td>M. m. musculus
Kazakhstan
0.88% / 0.07%</td><td rowspan="2">M. m. musculus
0.35% / 0.02%</td><td rowspan="3">M. m. musculus
+
M. m. castaneus
0.16% / 0.01%</td></tr>
<tr><td>M. m. musculus
Austria
0.80% / 0.06%</td></tr>
<tr><td colspan="2">M. m. castaneus
0.94% / 0.09%</td></tr>
<tr><td colspan="4">Mus spicilegus
1.12% / 0.17%</td></tr>
<tr><td colspan="5">Mus spretus
1.23% / 0.14%</td></tr>
<tr><td colspan="6">Mus mattheyi
1.11% / 2.08%</td></tr>
<tr><td colspan="7">Apodemus
1.97 % / -</td></tr>
</table>

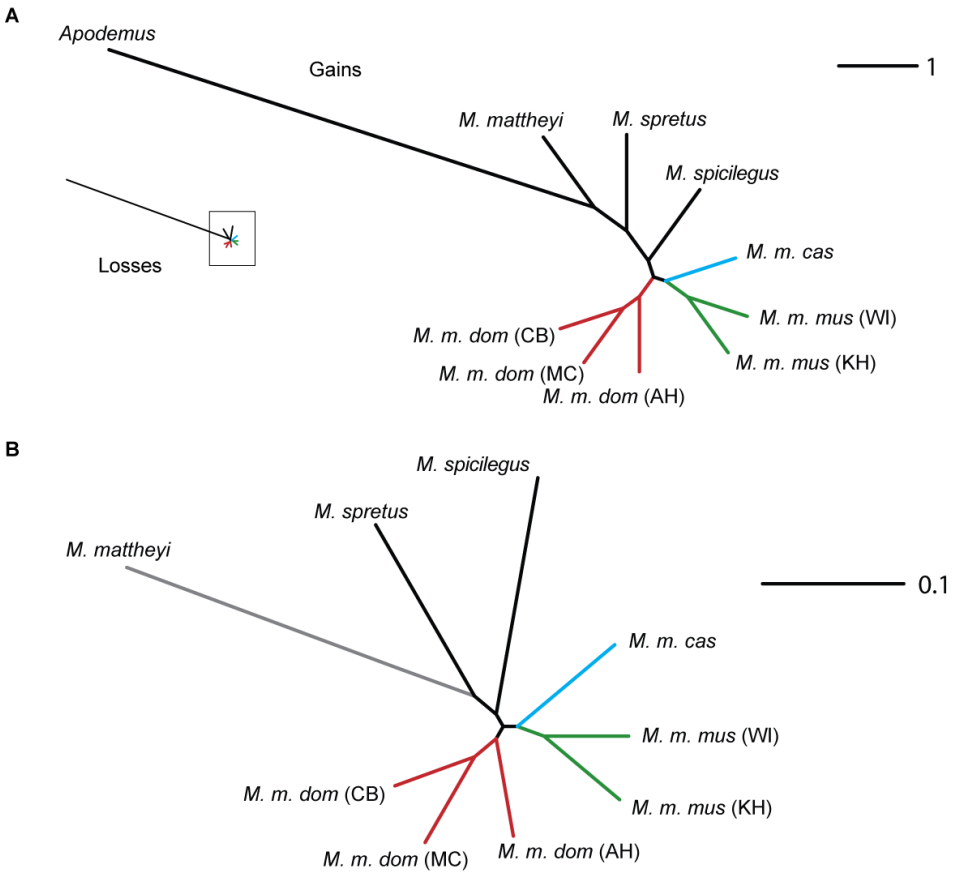

Figure 4.4. Lineage-specific gains of transcription greatly offset lineage-specific losses.

Tree representation of the lineage specific gains and losses of precentages of coverage across species shown in Table 4.4. A. Gains and losses, with losses represented by the small tree. Scale bar indicates 1% of total genomic coverage. B. Detail of losses, enclosed by the gray rectangle in A. Losses were not determined for *Apodemus*, and *M. matteyi* branch extends beyond the detail. Scale bar represents 0.1% of total genomic coverage.

In addition to the lineage-specific character of an event, it is possible to observe convergences, as transcription can be gained or lost multiple times along a phylogeny. Furthermore, a given lineage-specific event can be stable, meaning that it has been maintained as it is in all extant species derived from that lineage; or can be unstable, meaning that after an initial event, it has been reversed at least once (e.g. a lineage-specific gain that has been later lost in some of the derived taxa).

Multiple events are particularly difficult to assess through parsimony, since they can indicate convergences, as mentioned before, or single ancestral gains with many later losses. In order to evaluate how the occurrence and stability of events are represented in the transcriptional landscape, I compared lineage-specific events (only one occurrence) to non-lineage specific events (multiple occurrences) and the combined amount of genomic coverage they represent,

and stable events (lineage-specific gains or losses which do not show reversal in later divergences (Figure 4.5).

Some of these categories can be represented by different phylogenetic patterns out of the total presence/absence potential patterns, e.g. there are many combinations that can yield a lineage-specific gain. To address this, I normalized by the total number of phylogenetic categories that can be lineage or non-lineage specific, single or multiple event, and stable or unstable. This enables the discrimination between patterns which are very numerous, but with small contribution from each, from those which are less frequent and could have larger individual contributions.

A clear example can be observed in the case of lineage-specificity in stable gains and losses: non lineage-specific events have almost the same combined coverage that lineage-specific events have (Figure 4.5A). However, upon normalization the contribution of lineage-specific events becomes much more relevant (Figure 4.5B), since only a few patterns can show stable lineage-specific gains and each these have more coverage than stable-non lineage-specific gains.

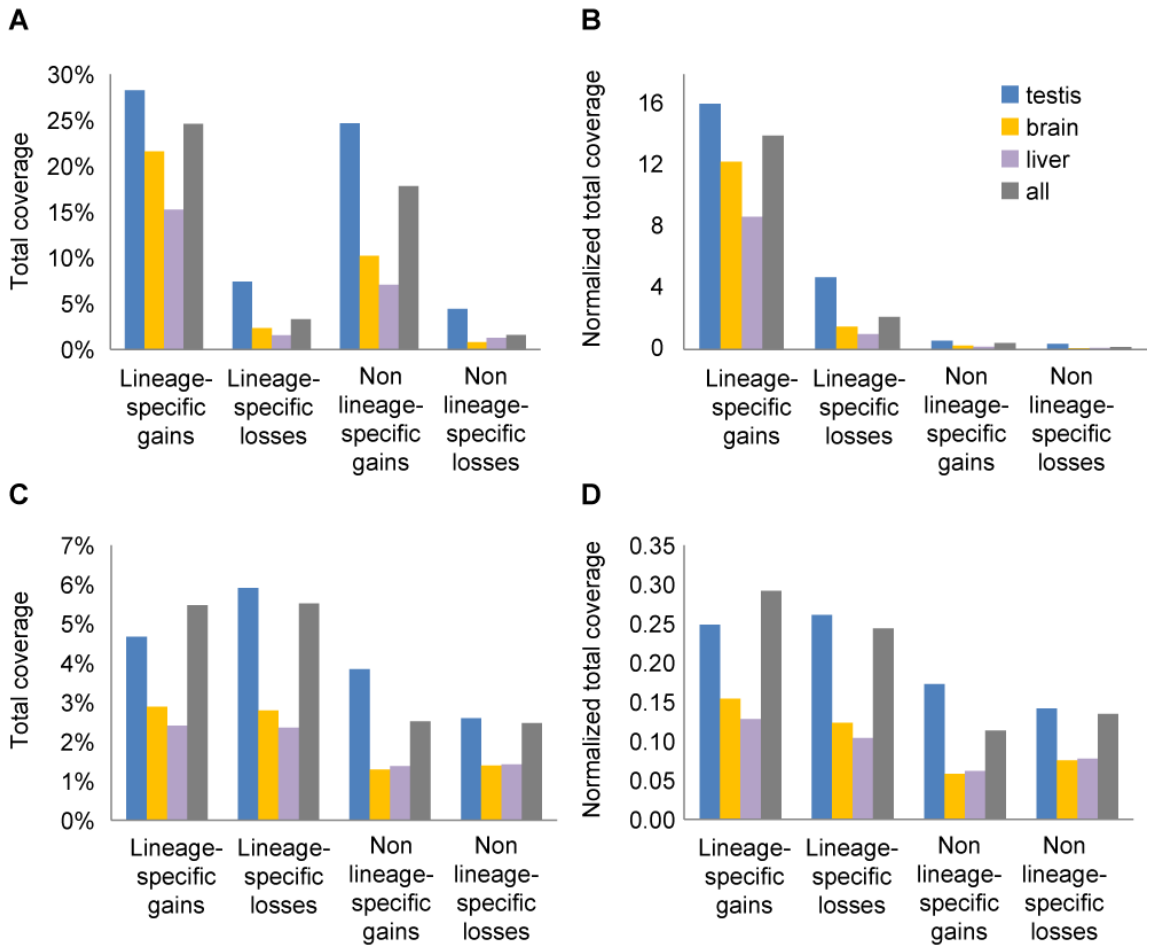

Figure 4.5. Most of the transcriptional coverage is represented by lineage specific and stably maintained gains over evolutionary time.

A. Combined coverage of the reference genome for stable lineage- and non-lineage specific gains and losses along the phylogeny. B. Normalized coverage of stable gains and losses by the amount of phylogenetic presence/absence patterns which fit each category. C. Combined coverage of the reference genome for unstable lineage- and non-lineage specific gains and losses along the phylogeny. D. Normalized coverage of unstable gains and losses by the amount of phylogenetic presence/absence patterns which fit each category.

Identification of cases of *de novo* transcripts

The analyses presented so far describe the bulk of transcription, regardless of their genic information, and assuming that there are regions of the genome which are not genes but are transcribed nonetheless. In order to determine if the general transcriptional dynamics also apply to gene-like entities, I decided to detect and quantify loci which show gene-like expression patterns and which have differences in presence or absence across the mouse phylogeny.

For this I combined genome and transcriptome information with splice junction detection and expression calling across the previously described phylogeny. This enabled the detection of focal cases of transcripts which have appeared *de novo* in the mouse phylogeny. While the previous analyses were done using base-wise coverage of the genome, the following analyses were done in a transcript-wise way, focusing on *bona fide* presence and absence across the phylogeny. This implies that only transcripts with expression beyond noise levels could be considered, and that absences were considered only in loci for which genomic reads, but not transcriptomic reads, were detected.

The algorithm for detection of candidates (see Methods) was able to recover 2,220 loci showing variation in presence/absence of expression. At this point, manual inspection and evaluation of the candidates was performed taking into consideration surrounding expression levels, noise-like behavior which might have escaped the automated screening and removing cases where uncertainty of *de novo* emergence would not allow a proper comparison.

This was also necessary because the transcript reconstruction step uses a different mapping algorithm (bowtie2; see Methods) and the expression calling is done using a more sensitive mapper (NextGenMapper; see Methods). The discordance between mapping programs was a common source of problems, and cases where no agreement was found between them were initially discarded. Most of these cases were part of the predictions for *M. mattheyi*. Very frequently the expression patterns exclusive of this species could not be reliably merged into gene-like models. For this reason, I refrained from using the data corresponding to *M. mattheyi* gains for quantification, as it is potentially underestimated compared to the ingroups. Furthermore, gene models in the proximity of the 3' end of an existing transcript or shorter than 300 nucleotides long were discarded. Models overlapping with other models or known genes were discarded, unless the splicing information would give a clear idea that the orientation of the transcript and exon boundaries were different from the known model.

After curation, I retained 663 candidates for which the expression patterns, gene models and phylogenetic distribution were consistent with at least one lineage-specific gain within the sampled taxa (Figure 4.6). A maximum parsimony mapping of gain and loss events for the 663 candidates along the phylogeny suggests that 763 events best describe the observed patterns, corresponding to 753 gains and 10 losses. There were 581 single independent gains, 148 double independent gains and 24 triple independent gains. All detected losses are single events.

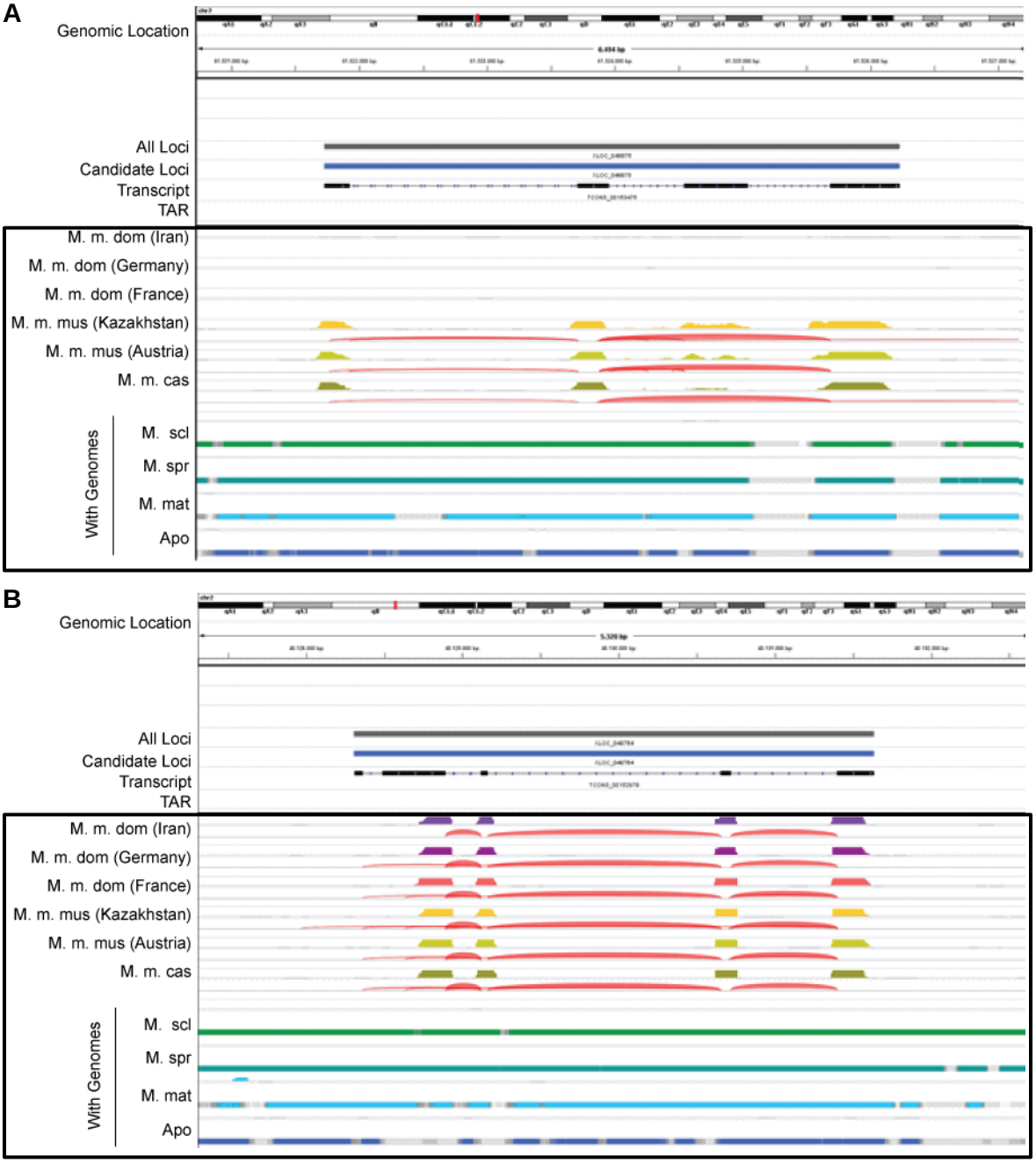

Figure 4.6. Two examples of *de novo* candidates based on expression and splicing information.

Top to bottom (both plots): Band ideogram corresponding to chromosome 2 in the mose. Scale and location of the locus in the mm10 version of the genome. Gray bar represents the genomic region covered by the transcript including putative introns and exons. Blue regions (below) indicate computational candidates. Transcript indicates the transcript structure as generated with TopHat2 (see Methods). TAR indicates the presence of transcriptionally active elements (as introduced in Chapter 1). Black rectangle indicates transcriptomes per taxon, in which expression is observed as discontinuous coverage and splicing is evident as red ribbons. Thick lines for the last four taxa indicate genomic read coverage. Below 10 reads per nucleotide (average 28) the region is gray. Absent regions are white. Genomic reads indicate the presence of the region and help define the absence of expression. A. Transcript gained at the divergence between *M. m. castaneus* and *M. m. musculus* from *M. m. domesticus*. B. Transcript gained at the early divergence between *M. musculus* and other *Mus* species.

This maximum parsimony reconstruction was performed under the assumption that transcriptional gains and losses at this stage are equally likely. Usual methods for the estimation of new gene gains at the protein level usually rely on Dollo's parsimony (Albert, 2006). This is equivalent to say that it is unlikely that the same protein arose in two different lineages independently, and that a most likely explanation is that it was gained before the divergence of the two lineages. In other words, proteins are more likely to be lost than to be gained.

However, the same cannot be assumed for the gain or loss of transcripts. Transcription of a region could in theory be achieved through many different mutations, and therefore there are many scenarios in which multiple gains would be plausible.

Table 4.5. Summary of loci with evidence of recently acquired and lost transcriptional activity per branch.

Branch	Time (mya)	Single gains	Two gains	Three gains	Single losses
***M. m. dom* (Germany)**	0.003	5	1	0	1
***M. m. dom* (France)**	0.003	8	2	1	0
***M. m. dom* (European)**	0.01	6	0	0	0
***M. m. dom* (Iran)**	0.03	15	9	1	0
M. m. dom	0.43	23	15	4	0
***M. m. mus* (Kazakhstan)**	0.01	13	3	0	0
***M. m. mus* (Austria)**	0.01	3	2	0	0
M. m. mus	0.267	27	4	3	0
M. m. cas	0.418	42	8	4	1
M. m. cas* + *M. m. mus	0.042	15	9	0	2
Mus musculus	0.74	31	21	0	2
M. spicilegus	1.2	105	20	3	4
M. mus* + *M. spicilegus	0.5	16	1	0	0
M. spretus	1.7	96	43	6	0
M. musculus* + *M. spicilegus* + *M. spretus	4.9	86	0	0	0
M. mattheyi	6.6	28	10	2	0
Genus *Mus*	4	62	0	0	0
Apodemus	10.6	-	-	-	-

Quantification of gain rates for curated genes

Assuming that the observed lineage specific transcribed loci could be the result of steady accumulation, I tested a linear model to approximate a rate at which genes are accumulated per million years as unit of time (Figure 4.7). This results in an estimated rate of accumulation of genes of 90 loci per million years (±13; 95% confidence interval, r^2=0.906) (Figure 4.7D).

However, at closer divergences it is possible to observe an increased rate of gains, of nearly 15 genes every 20,000 years (±13; 95% confidence interval, r^2=0.718) (Figure 4.7E). This scales up to 495 loci per million years (± 347 loci; 95% confidence interval), suggesting a nearly 5-fold acceleration for divergences shorter than 50,000 years.

Interestingly, the intercept of the linear model is larger than zero, which suggests that at zero divergence time there are available transcripts, i.e. there is some expected level of polymorphism in a given population.

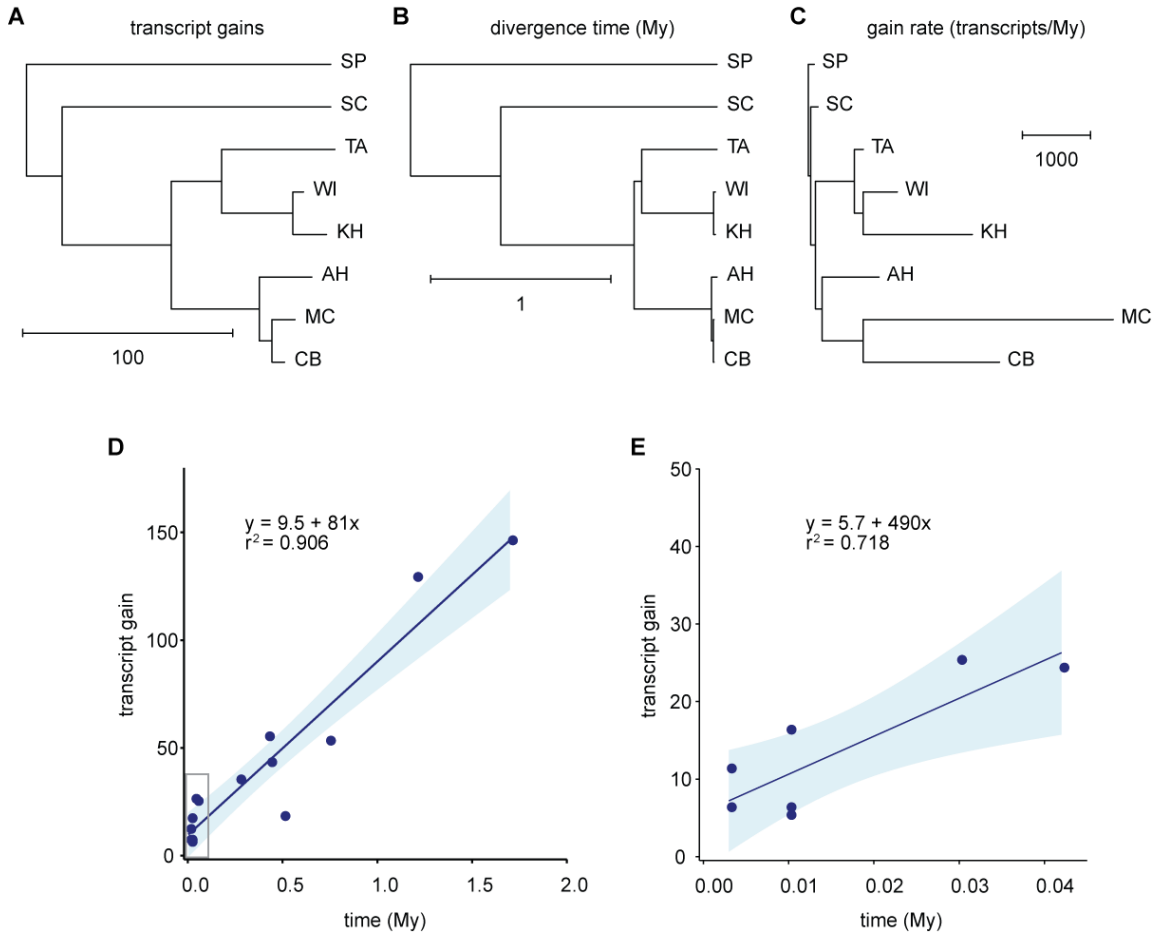

Figure 4.7. Linear estimation of the rates of *de novo* transcription gains along the mouse phylogeny.

A. Transcript gains per branch. B. Divergence times since the split between the house mouse and *M. spretus*. C. Rate of transcript gain per million years, illustrated by branch. SP: *M. spretus*; SC: *M. spicilegus*; TA: *M. m. castaneus*; WI: *M. m. musculus* from Austria; KH: *M. m. musculus* from Kazakhstan; AH: *M. m. domesticus* from Iran; MC: *M. m. domesticus* from France; CB: *M. m. domesticus* from Germany D. Linear estimation across 1.7 million years of divergence. E. Linear estimation for divergences shorter than 50,000 years, namely among the populations of *Mus musculus musculus* and *Mus musculus domesticus*, and corresponding to the gray square in E. Solid blue lines indicate the linear estimate; blue shading indicates the 95% confidence interval around the estimate.

What are the dynamics of transcription loss in known genes?

Given that the loss events detected in recently gained loci were almost negligible compared to the gains, I decided to detect losses of expression from known genes. This follows the rationale that transcription can indeed be completely lost, and that by looking at the annotated genic transcriptome one can reliably detect those events.

The overlap between genes annotated in Ensembl 74 and the recently gained loci is small, hence the majority of the identified loci for gains of transcription were located in regions where no genes were previously reported. 87% of the loci do not overlap with any Ensembl gene and 7.5% overlap only partially and clearly have different exon structures. Only 5.5% corresponds to annotated genes.

One key assumption is that known genes which were not recovered by the pipeline that detects *de novo* gene candidates were present before the divergence between the genera *Apodemus* and *Mus*. This means that all absences can be interpreted as losses.

666 out of 38,561 genes in the Ensembl 74 version were found to have losses in the phylogeny. The identified genes could be further split into gene types (Ensembl biotypes): 262 are protein coding genes, 63 are long intergenic non-coding RNAs (lincRNAs), 218 are pseudogenes, and 123 belong to other categories (snoRNAs, snRNAs, antisense RNA, miRNAs, processed transcripts, rRNAs, among others).

There was no significant phylostratigraphic enrichment among those loci, as one would predict that young, lineage-specific genes would be easily lost. However, functional enrichment tests show significant overrepresentation of several groups like olfactory receptors, major urinary proteins, secreted proteins and cytokines. Other groups with marginal overrepresentation are serine protease inhibitors, vomeronasal receptors, KRAB domain-containing proteins and recurrent transcripts from cDNA screens (Appendix C).

The estimated rate at which transcription is lost for genes is approximately 33 loci per million years (±16; 95% confidence interval, r^2=0.914) (Figure 4.8A). Interestingly, different types of genes yield different rates of loss (Fig 4.6B-E).

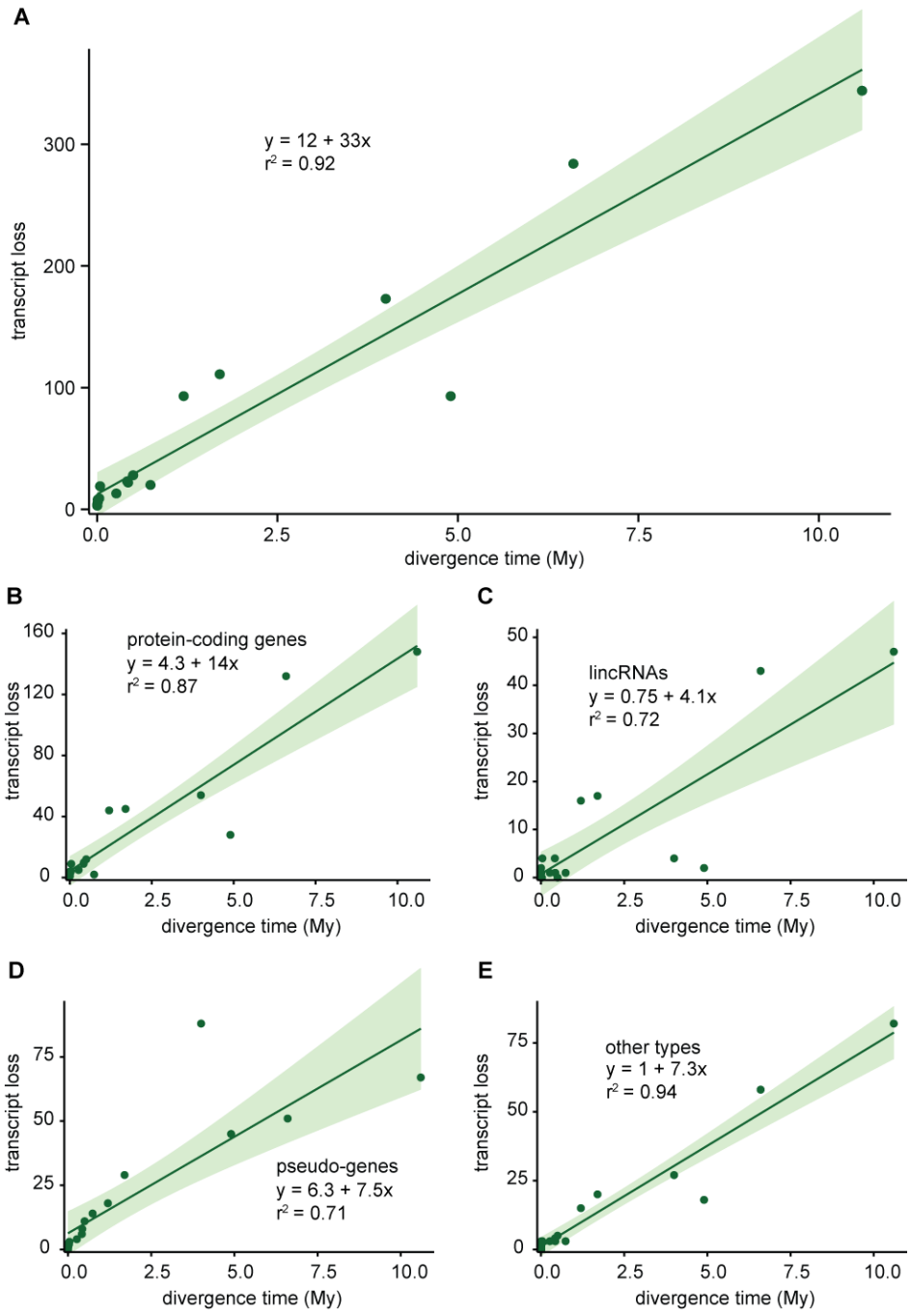

Figure 4.8. Linear estimation of the loss of transcription for known genes shows that different types of genes lose their transcription at different rates.

A. All genes which show loss of expression along the phylogeny, regardless of their type. B-E. Estimation of transcription loss for protein-coding genes (B), lincRNAs (C), pseudo-genes (D), and other types (E) according to annotations from Ensembl 74. The quantification of events is based on maximum parsimony, accounting for losses only and assuming that these genes were transcribed before the split between *Mus* and *Apodemus*, approximately 10.6 million years ago. Solid green lines indicate the linear estimate; green shading indicates the 95% confidence interval around the estimate.

Where are new genes expressed?

In terms of tissue specificity, 69% of the new genes are tissue specific, with 66% of those being testis specific, 1.5% liver specific and 1.5% brain specific. 7.5% of all detected loci have expression across all three tissues. However, 69.5% of all genes are expressed to some extent in the testis, while brain expresses up to 22% and liver up to 9% (Figure 4.9).

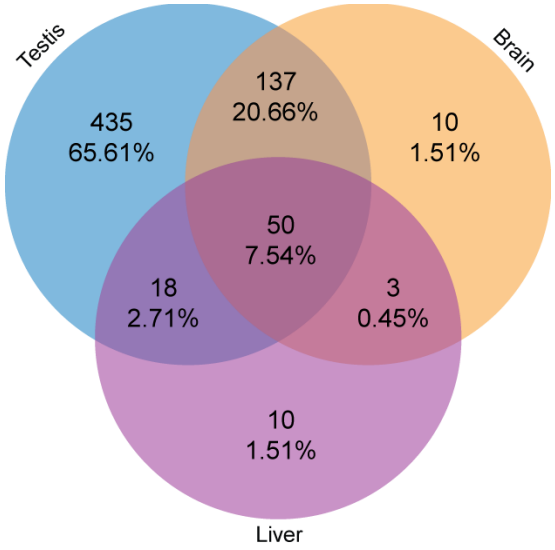

Figure 4.9. *De novo* genes are expressed mostly in the testis.

New transcripts above noise level in three sampled tissues, including all taxa. A large proportion is also expressed simultaneously in testis and brain.

In testis, the levels of expression for new loci are correlated (rho=0.96, p-val=0.002, 7 age classes) with the estimated phylogenetic age, such that the most recently acquired genes have low expression values, but still higher than most liver- and brain- expressed transcripts, and progressively older genes have much higher expression (Figure 4.10).

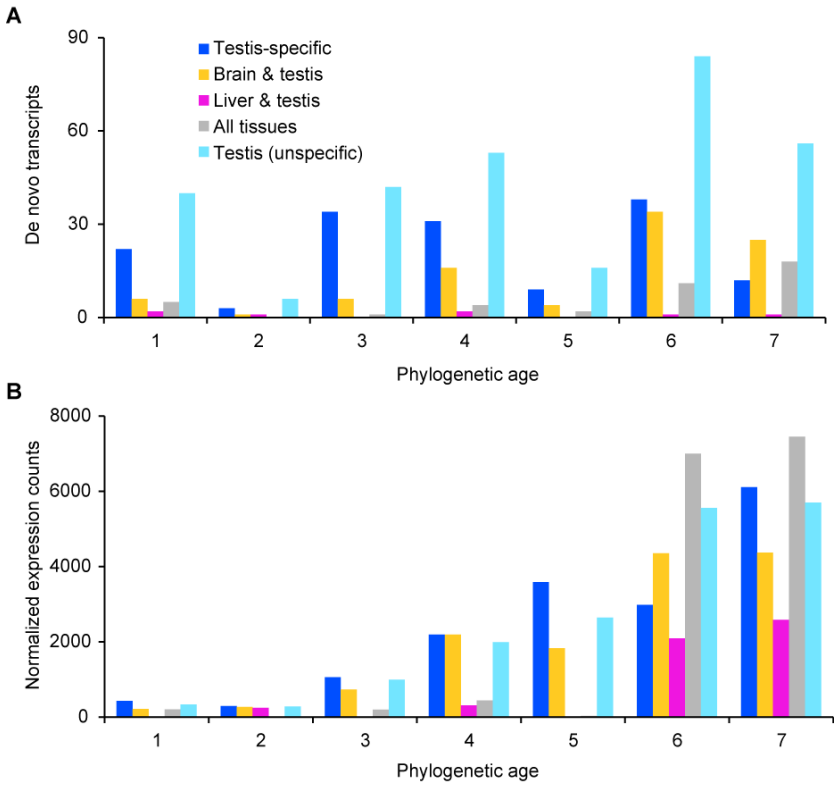

Figure 4.10. Testes express the most new transcripts, and the expression levels of new transcripts are correlated with their phylogenetic age.

A. Number of genes across short-scale phylogenetic ages according to their expression in different combinations of sampled tissues. B. Mean normalized expression across short-scale phylogenetic ages according to their expression in different combinations of sampled tissues. The phylogenetic age is estimated according to the least recent expression gain at each locus. A phylogenetic age of 1 represents the most recently gained transcripts and an age of 7 represents the least recently gained. Only genes for which a reliable phylogenetic age classification could be built were included, i.e. *M. spretus*- and *M. spicilegus*-specific loci are absent here.

Discussion

This study is an approximation to the role of transcript emergence in the generatiuon of new genes and functions. Current models of *de novo* gene birth expect an intermediary between non-coding intergenic sequences and protogenes. The results presented here indicate that the transcriptional dynamics over short evolutionary times are able to provide ample raw material for new genes to emerge.

Pervasive transcription can provide material for new genes

The debate about pervasive transcription of the genome has been rather a matter of definitions (van Bakel et al., 2011; Berretta and Morillon, 2009; Clark et al., 2011; Kapranov et al., 2007).

Among the strongest opponents of pervasive transcription, van Bakel and colleagues (2010, 2011) do not refute the actual widespread transcription of the genome. However, they argue that the abundance and stability of those “dark-matter” transcripts is so low that their relevance is negligible in comparison to well-defined genes (van Bakel et al., 2011).

I argue here that in the context of evolutionary innovation through *de novo* gene birth this is indeed relevant.

The definition of function related to pervasive transcription is still far from complete, and it is possible that the two common examples of pervasive transcription belong to two completely different classes, both at the molecular and evolutionary scale: On the one side we have long intergenic non-coding RNAs, which are known to function in many processes by interacting with various cellular components as regulators (Chodroff et al., 2010; Kung et al., 2013; Necsulea et al., 2014); on the other side we have RNA species which seem to be byproducts of other processes, like CUTs (Neil et al., 2009; Xu et al., 2009) and PROMPTs (Ntini et al., 2013; Preker et al., 2011), whose functions are less clear. It is also unclear if the role these transcripts play derives from the transcription event itself (Batada and Hurst, 2007; Ebisuya et al., 2008; Wang et al., 2011), or if they bear sequence properties which allow them to have more specific functions.

The first class refers to ‘traditional’ RNA genes, which fit right next to ribosomal RNAs, transfer RNAs, microRNAs, spliceosomal RNAs, and many more, while the latter still falls within the realm of transcribed sequences without known function. In terms of sequences which could theoretically contribute to the emergence of new genes, both classes have properties that make them suitable, i.e. stable or frequent transcription. These two classes are only mentioned here to illustrate the point that pervasive transcription is most likely a sparse collection of unrelated phenomena. This is analog to the definition of non-coding RNAs by their absence of protein-coding potential instead of functional attributes.

The approaches I describe here make no initial classification between classes of transcripts, and assume that having stable transcription over long periods of time gives a region enough potential to develop a gene-like structure, and even genes.

The mapped reference genome enables the inclusion of syntenic relationships between species in the comparison, and obviates the need of a high resolution genome for each of the taxa analyzed. For this reason, the results of this study are valid for the uniquely mappable portion of the reference genome; hence I assume that the processes here described could be

underestimated, by not accounting for those genomic regions which have evolved beyond our detection possibilities.

In the context of gene birth, an initial function is not required, and one could even suggest that the fact that a pre-existing transcript has a given function, e.g. antisense regulator, could affect the emergence of a new function (Pavesi et al., 2013). However, it is of particular relevance for how long a region can maintain its transcription actively and stably. At the molecular scale, transcription is achieved through recruitment of RNA polymerases by signals in the core promoters, and is stabilized by many different interacting transcription factors (Alberts et al., 2002; Lodish et al., 2007). These signals could also be generated in both strands from bidirectional promotors (Seila et al., 2008).

Furthermore, transcript stability largely depends on the presence of a polyadenylation signal (Dreyfus and Régnier, 2002; Ntini et al., 2013). However, the overall stability is the result of complex interactions between the transcriptional machinery, splicing machinery, the cleavage/polyadenylation multi-protein complex and RNA surveillance mechanisms (Dreyfus and Régnier, 2002; Fang et al., 2013; Wilusz et al., 2001). Given the nature of the transcriptome sample preparation, most of the here detected transcripts are likely to be polyadenylated.

At the evolutionary scale, less is known about these processes. Based on high-throughput transcriptome sequencing data, it has been suggested that the human genome is transcribed in at least 80% of its length (Clark et al., 2011; ENCODE Project Consortium, 2012; ENCODE Project Consortium et al., 2007; Hangauer et al., 2013), and already from the first large full-length cDNA sequencing efforts in the mouse, it was stated that at least 63% of the genome is transcribed (Carninci et al., 2005; Katayama et al., 2005; Okazaki et al., 2002). Currently, it is known that many lincRNAs are deeply conserved (Necsulea et al., 2014), but that they are also subjected to high turnover over different evolutionary scales (Kutter et al., 2012; Managadze et al., 2011).

Here, I show that in a single tissue from a single species (representing several individuals and including intronic regions) it is possible to detect up to 47% of the genome being transcribed. Furthermore, the addition of only two other tissues allows the detection of transcripts corresponding to almost two thirds of the genome. This means that at a given time point in the evolution of a genome, there is at least two thirds of the genome available as transcripts, consistent with current estimates in humans.

For the sake of including all genomic regions subjected to transcription, I include intronic regions together with exons. The process of splicing is not always absolutely efficient, and often results in transcripts with spliced and unspliced variants (Tilgner et al., 2012). In addition to this, introns can give rise to new exons and functional genes are known to exist within exon boundaries (Sorek et al., 2004).

The comparisons including more closely related taxa indicate that a large proportion has the potential to become transcribed. Approximately 80% of the mouse genome is either transcriptionally active now, has been transcriptionally active in the past 10 million years or has evolved transcriptional activity in closely related lineages. Approximately 25% seems to have been constantly transcribed in the same period of time.

Asymmetry in gains and losses of transcription

It is evident from the results that a much larger proportion of the genome shows patterns of lineage specific gains of transcription. Interestingly, transcription does not seem to follow the same gain-loss balance observed at the gene level, and more specifically the balance of gain-loss of protein-coding genes.

I suggest here that the evolutionary half-lives of newly emerged transcripts are relatively long, thus providing a plausible opportunity for the exploration of new functions as transcripts (Heinen et al., 2009; Tautz and Domazet- Loso, 2011) or by associating with ribosomes and producing peptides (Cai et al., 2008; Carvunis et al., 2012; Wilson and Masel, 2011), thus eventually becoming non-coding or coding genes.

It is possible to hypothesize that transcripts with very low levels of expression are selectively neutral. For example, it has been shown that the toxic effect induced by repetitive transcripts is highly dependent on expression levels (Nalavade et al., 2013). It has been previously suggested that transcriptional noise could be a main driver of gene birth, providing the possibility to test the genic potential of genes under different conditions simultaneously (Polev, 2012).

The asymmetry in the dynamics of transcription gain and loss cannot be explained if one assumes that the transition between active and inactive states is governed by a few causal mutations that have equal probabilities of occurrence. Hence, a factor or process that contributes to the offset in patterns between transcription gains and losses is still missing.

It is paradoxical that transcription is gained at a much faster rate than it is lost, because this would eventually result in a fully transcribed genome. Assuming that pervasive transcription is true and most of the genome is transcribed (van Bakel et al., 2010, 2011; Clark et al., 2011; ENCODE Project Consortium, 2012; Hangauer et al., 2013), it would be possible to speculate that the steady state of the genome is to be actively transcribed. This leads to the idea that all changes we observe here as gains are shifts in the expression levels.

Taking into account this asymmetric behavior, it might be even necessary to reconsider in the future the parsimony criteria to assume that gains of transcription are more likely to be observed than losses.

From transcribed protogenes to *de novo* genes

It is expected that any organism, at any given time point, expresses new genes which are not present in closely related taxa (Tautz and Domazet- Loso, 2011; Wilson et al., 2005). These genes are in a dynamic equilibrium, through which protogenes are constantly explored (Carvunis et al., 2012) and removed from the genome (Palmieri et al., 2014).

From this dynamic equilibrium, it can be also expected that few of these protogenes become fixed in a population and retained in a genome as genes (Zhao et al., 2014). This is one of the explanations of why we observe lineage-specific genes, i.e. orphan genes, at each divergence along the phylogenetic history of an organism (Tautz and Domazet- Loso, 2011). However, the acquisition of new genes in large amounts has been mostly associated to functional or ecological shifts, such that the fitness of the existing genes departs from the optimum, favoring the retention of new genes (Colbourne et al., 2011; Khalturin et al., 2009; Tautz and Domazet-Loso, 2011).

The analyzed repertoire of newly gained and lost transcripts reveals a previously unsuspected property of *de novo* genes, probably derived from pervasive transcription dynamics. Contrary to the expectation from protein-coding dynamics, the results indicate that most newly arisen transcripts are not quickly lost.

It is important to highlight that the analyses described here represent quantification at the transcriptional stage, and therefore a more accurate notation of the candidate loci identified would be “transcribed protogenes”. I make the distinction from more general protogenes, which would be any kind of sequence with some gene-like features, and from protein-coding protogenes, for which open reading frames or even ribosomal association information would be

available. The protogenes I describe have evidence of transcription beyond noise levels, but their evidence of translation is yet to be discovered.

My current genome- and phylogeny-wide estimate for gains of stable transcripts, i.e. those having detectable expression levels and some degree of splicing, lies at 89 loci per million years (±14; 95% confidence interval), in what seems to follow a linear behavior, during the first 10 million years of divergence. Conversely, estimates based on older genes suggest that 45 loci are lost per million years (±16; 95% confidence interval), on what also seems to follow a linear behavior over the past 10 million years. Interestingly, the most affected protein-coding genes are genes with a relatively complex duplication history, such as olfactory receptors and mouse urinary proteins.

It is possible that the loss rates are slightly inflated due to mappability issues derived from regions with history of recent duplication. The mapping program of choice is able to successfully detect regions of high divergence compared to more standard tools (Sedlazeck et al., 2013). However, recent or highly conserved paralogs could in principle become intermingled. I have tried to correct for this (see methods), but given the overrepresentation of known multi-paralog genes (Appendix C) in loss of transcription of protein-coding genes (Figure 4.8), one should be cautious in the future exploration of this trend. In any case, I can imagine that this effect would potentially result in an overestimation of the loss rates, while gain rates would hardly be influenced.

Further assuming that these are stable rates, it is worth noticing that the total amount of losses (regardless of gene type) is not more than half of the observed gains at the transcriptional level. This partially rules out the hypothesis that transcription is the main player in the gain/loss balance. I suggest that in terms of transcription, there is enough available material at all time horizons to generate new genes from stable transcripts.

In concordance with the results from the previous chapter, which showed the instability of recently arisen protein coding genes, it is possible to suggest that the balance between gain and loss of new protein-coding genes is achieved at the reading frame level (see Chapter 2). This could explain why the genome does not fill up constantly with genes, or at least with protein-coding genes. However, under these assumptions, genomes would indeed tend to fill up with RNA genes or transcribed protogenes. It is a forthcoming challenge to understand if this is true, and in the likely scenario that it is not true, what mechanisms explain the observed patterns.

Differences in expression levels

It is little what I can infer about the functionality of these *de novo* transcripts from my analyses. Among the new transcripts detected for the mouse lineages, the expression levels of older transcripts are higher than more recently acquired genes. One possible interpretation is that very new transcripts are expressed at low levels, and with time their expression increases as they become functional and integrated.

In a previous work, I found a similar pattern using microarray data for several mouse tissues. Tissue-specific genes with an estimated origin after the divergence of Mammalia from Marsupialia show a positive correlation between phylogenetic age and expression (Neme Garrido, 2011). Furthermore, it has been shown in *Drosophila* that among orphan genes, those which have been acquired earlier in the phylogenetic history show higher expression levels than younger genes (Palmieri et al., 2014).

These trends are suggestive of expression-level maturation, which starts from low expression and increases as the transcript is required in higher amounts. This might be insufficient to explain the genome-wide patterns of transcriptional gains, but in the context of protogenes, it might indeed be associated with transitions between non-functional entities to functional genes (Carvunis et al., 2012; Palmieri et al., 2014).

Testis as a niche for new genes

The first examples of *de novo* genes were found to be overrepresented in testis (Begun et al., 2007; Chen et al., 2007; Heinen et al., 2009; Levine et al., 2006), and it has been hypothesized that an “out-of-testis” mechanism could be described in the context of gene birth (Kaessmann, 2010). In the first chapter, I addressed this issue through genomic phylostratigraphy of annotated protein coding genes, and found that the most recently evolved genes in this class were not significantly enriched in the testis transcriptomes, while slightly older genes do have enriched testis expression (Neme and Tautz, 2013).

It has been suggested that testis might have a more favorable environment for the generation of new genes. This is due to a particular convergence of factors, like alternative machinery for transcriptional regulation (Kleene, 2001), open chromatin features (Kimmins and Sassone-Corsi, 2005; Kleene, 2001), and general simpler promoters (Kleene, 2005) which results in higher pervasive transcription, by enabling transcription of regions which in other tissues or cell types would not be usually transcribed. Furthermore, a variety of selective pressures act on the

testis at multiple levels, such as sperm competition, sexual selection and reproductive isolation, among others (Kaessmann, 2010).

Consistent with this, I find that the per-species proportion of the genome transcribed in the testis is on average 12% higher than in the brain, and double than in the liver. In terms of conservation of expression, testis has the lowest relative conserved portion, with 19% of the total transcriptome (14% of the genome) being transcribed across all taxa, compared to 33% (19% of the genome) in the brain, and 23% (9% of the genome) in the liver. Similarly, the proportion of the genome which is specifically transcribed in each of the *M. m. domesticus* and *M. m. musculus* populations, i.e. the most recently diverging sampled taxa, is almost double for testis than liver or brain. All of this is consistent with the hypothesis that testis has a higher potential for the development of transcribed protogenes than other tissues.

Also, the large majority of the *de novo* transcripts are testis-specific (69%), and almost all of the loci expressed in other tissues are also expressed to some extent in the testis (96%). There seems to be an increase of the expression levels as genes become older, as it has been also shown in *Drosophila* (Palmieri et al., 2014). The youngest genes have lower expression levels, and those expression levels seem to become higher as the age of genes increases, possibly linked to a functional association. It is possible that this process might be also related to the way a gene is born in the context of one tissue, and expands its expression not only in amount of expressed transcript, but also towards other tissues.

Conclusion

These analyses constitute a first approximation to the cross-species comparison of pervasive transcription. The conservation and turnover of genome-wide transcription seems to support the notion of a highly transcriptionally active genome in mammals.

To this date, this is the densest catalogue of *de novo* gene candidates in terms of phylogenetic coverage and time scales sampled for any organism and the first evidence-based large-scale analysis of *de novo* genes in the mouse. Furthermore, I have detected candidates for population-specific transcripts in the mouse.

I have managed to approximate a rate at which transcripts are stably gained in a genome, and find that they are not as easily lost as one would have assumed from an equilibrium model of gain and loss. Accordingly, I postulate that an as yet undiscovered process or factor is required to explain this discrepancy. But from the perspective of potential for *de novo* gene evolution,

there seems to be no limitation to the amount of available precursors at any given time point for a new gene to emerge.

Concluding remarks

The ever increasing complexity of life can also be observed at the gene level. New genes appear frequently and at all divergences through multiple mechanisms, some of which include the generation of completely new entities from previously non-genic precursors.

A major contributing factor to this increase of complexity is the seemingly large availability of transcribed material over time, providing random sequences upon which selection can operate to generate new functions or improve existing ones. Much of the genome is transcribed and a large proportion of that transcriptional activity seems to be stable over long periods of time. One can hypothesize that this gives enough time for non-functional transcripts to become associated with ribosomes and produce a high number of protogenes. Protogenes, as it has been previously shown, have a high turnover at the protein level, but the dynamics at the transcriptional level had never been inquired. The results of my analyses indicate that protogenes can be accumulated steadily at all divergences in the form of transcripts. This reinforces the idea that *de novo* gene birth is not a rare phenomenon, but rather widespread and frequent. If one considers the wide transcriptional availability of the genome, and distribution of small motifs that confer gene-like properties to a random transcript, the emergence of genes should be a dependent on the functional potential of random peptides.

As part of the life cycle of genes which has been previously described, it can also be stated that genes are constantly changing over time. I observed that genes tend to increase in complexity, namely they become longer, contain more domains and exons. This can be considered the outcome of many generations of tinkering and improvement of functions, but the possibility that this is the result of passive dynamics cannot be neglected.

Furthermore, older genes have stronger negative selection acting on them and tend to accumulate fewer mutations that could have a negative impact on the reading frame. I argue that this is most likely the result of a slow but steady integration of nearly-neutral entities that eventually become locked into restrictive cellular environments. Nevertheless, these results are based on generalized trends, and individual cases remain to be studied.

Perspectives

De novo gene birth and overprinting of reading frames are the two mechanisms which are able to generate true innovation at the protein level in relatively few steps. However, this statement relies in a distribution of functionally relevant structures which is until now unknown. Understanding how random peptides are able to contribute to the fitness landscape of an organism is probably the most relevant question at present. It is known that many new genes are associated with ecological shifts and speciation processes. For this reason, *de novo* genes are an important system to determine the causal links between innovation, protein folds and general structure, the emergence of functions in coding and non-coding sequences and how the interactions with the environment are able to shape the gene repertoires of living organisms.

More individual-based analyses of *de novo* genes are also needed to understand the population dynamics underlying the emergence and fixation of new genes. Furthermore, mathematical models including a well-defined population genetics perspective as well as information coming from the molecular genetics and genomics evidence are needed to move forward and generate accurate predictions about the forces and strength of the evolutionary forces that surround gene birth and death processes.

References

Abrusán, G. (2013). Integration of New Genes into Cellular Networks, and Their Structural Maturation. Genetics *195*, 1407–1417.

Alba, M.M., and Castresana, J. (2007). On homology searches by protein Blast and the characterization of the age of genes. BMC Evol Biol *7*, 53.

Albà, M.M., and Castresana, J. (2005). Inverse relationship between evolutionary rate and age of mammalian genes. Mol. Biol. Evol. *22*, 598–606.

Albert, V.A. (2006). Parsimony, Phylogeny, and Genomics (Oxford University Press).

Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., and Walter, P. (2002). Molecular Biology of the Cell.

Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. *25*, 3389–3402.

Anders, S., and Huber, W. (2010). Differential expression analysis for sequence count data. Genome Biol. *11*, R106.

Van Bakel, H., Nislow, C., Blencowe, B.J., and Hughes, T.R. (2010). Most “dark matter” transcripts are associated with known genes. PLoS Biol. *8*, e1000371.

Van Bakel, H., Nislow, C., Blencowe, B.J., and Hughes, T.R. (2011). Response to “The Reality of Pervasive Transcription.”PLoS Biol *9*, e1001102.

Batada, N.N., and Hurst, L.D. (2007). Evolution of chromosome organization driven by selection for reduced gene expression noise. Nat. Genet. *39*, 945–949.

Beaudoing, E., Freier, S., Wyatt, J.R., Claverie, J.-M., and Gautheret, D. (2000). Patterns of Variant Polyadenylation Signal Usage in Human Genes. Genome Res. *10*, 1001–1010.

Begun, D.J., Lindfors, H.A., Kern, A.D., and Jones, C.D. (2007). Evidence for *de novo* evolution of testis-expressed genes in the Drosophila yakuba/Drosophila erecta clade. Genetics genetics.106.069245.

Bekpen, C., Marques-Bonet, T., Alkan, C., Antonacci, F., Leogrande, M.B., Ventura, M., Kidd, J.M., Siswara, P., Howard, J.C., and Eichler, E.E. (2009). Death and resurrection of the human IRGM gene. PLoS Genet. *5*, e1000403.

Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J., and Wheeler,, D.L. (2004). GenBank: update. Nucleic Acids Res. *32*, 23D–26.

Benton, M.J., and Donoghue, P.C.J. (2007). Paleontological evidence to date the tree of life. Mol. Biol. Evol. *24*, 26–53.

Berretta, J., and Morillon, A. (2009). Pervasive transcription constitutes a new level of eukaryotic genome regulation. EMBO Rep. *10*, 973–982.

Bertone, P., Stolc, V., Royce, T.E., Rozowsky, J.S., Urban, A.E., Zhu, X., Rinn, J.L., Tongprasit, W., Samanta, M., Weissman, S., et al. (2004). Global Identification of Human Transcribed Sequences with Genome Tiling Arrays. Science *306*, 2242–2246.

Bornberg-Bauer, E., Huylmans, A.-K., and Sikosek, T. (2010). How do new proteins arise? Curr. Opin. Struct. Biol. *20*, 390–396.

Brawand, D., Soumillon, M., Necsulea, A., Julien, P., Csárdi, G., Harrigan, P., Weier, M., Liechti, A., Aximu-Petri, A., Kircher, M., et al. (2011). The evolution of gene expression levels in mammalian organs. Nature *478*, 343–348.

Bridges, C.B. (1935). SALIVARY CHROMOSOME MAPS With a Key to the Banding of the Chromosomes of Drosophila Melanogaster. J. Hered. *26*, 60–64.

Bridgham, J.T., Eick, G.N., Larroux, C., Deshpande, K., Harms, M.J., Gauthier, M.E.A., Ortlund, E.A., Degnan, B.M., and Thornton, J.W. (2010). Protein evolution by molecular tinkering: diversification of the nuclear receptor superfamily from a ligand-dependent ancestor. PLoS Biol. *8*.

Brooks, A.N., Aspden, J.L., Podgornaia, A.I., Rio, D.C., and Brenner, S.E. (2011). Identification and experimental validation of splicing regulatory elements in Drosophila melanogaster reveals functionally conserved splicing enhancers in metazoans. RNA *17*, 1884–1894.

Buljan, M., Frankish, A., and Bateman, A. (2010). Quantifying the mechanisms of domain gain in animal proteins. Genome Biol. *11*, R74.

Cai, J.J., and Petrov, D.A. (2010). Relaxed purifying selection and possibly high rate of adaptation in primate lineage-specific genes. Genome Biol Evol *2*, 393 – 409.

Cai, J., Zhao, R., Jiang, H., and Wang, W. (2008). De Novo Origination of a New Protein-Coding Gene in Saccharomyces cerevisiae. Genetics *179*, 487–496.

Capra, J.A., Williams, A.G., and Pollard, K.S. (2012). ProteinHistorian: Tools for the Comparative Analysis of Eukaryote Protein Origin. PLoS Comput Biol *8*, e1002567.

Carninci, P. (2010). RNA dust: where are the genes. DNA Res *17*, 51 – 59.

Carninci, P., Kasukawa, T., Katayama, S., Gough, J., Frith, M.C., Maeda, N., Oyama, R., Ravasi, T., Lenhard, B., Wells, C., et al. (2005). The transcriptional landscape of the mammalian genome. Science *309*, 1559–1563.

Carvunis, A.R., Rolland, T., Wapinski, I., Calderwood, M.A., and Yildirim, M.A. (2012). Proto-genes and *de novo* gene birth. Nature *487*, 370 – 374.

Catzeflis, F.M., and Denys, C. (1992). The African Nannomys (Muridae): An early offshoot from the Mus lineage - evidence from scnDNA hybridization experiments and compared morphology. Isr. J. Zool. *38*, 219–231.

Chen, S., Zhang, Y.E., and Long, M. (2010). New Genes in Drosophila Quickly Become Essential. Science *330*, 1682–1685.

Chen, S.-T., Cheng, H.-C., Barbash, D.A., and Yang, H.-P. (2007). Evolution of hydra, a recently evolved testis-expressed gene with nine alternative first exons in Drosophila melanogaster. PLoS Genet. *3*, e107.

Chodroff, R.A., Goodstadt, L., Sirey, T.M., Oliver, P.L., Davies, K.E., Green, E.D., Molnár, Z., and Ponting, C.P. (2010). Long noncoding RNA genes: conservation of sequence and brain expression among diverse amniotes. Genome Biol. *11*, R72.

Chothia, C. (1992). One thousand families for the molecular biologist. Nature *357*, 543–544.

Chothia, C., and Gough, J. (2009). Genomic and structural aspects of protein evolution. Biochem J *419*, 15 – 28.

Chung, W.Y., Wadhawan, S., Szklarczyk, R., Pond, S.K., and Nekrutenko, A. (2007). A first look at ARFome: Dual-coding genes in mammalian Genomes. PLoS Comp Biol *3*, 855 – 861.

Cingolani, P., Platts, A., Wang, L.L., Coon, M., Nguyen, T., Wang, L., Land, S.J., Lu, X., and Ruden, D.M. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. Fly (Austin) *6*, 80–92.

Clark, M.B., Amaral, P.P., Schlesinger, F.J., Dinger, M.E., and Taft, R.J. (2011). The reality of pervasive transcription. PLoS Biol *9*, e1000625.

Cohen, O., and Pupko, T. (2011). Inference of Gain and Loss Events from Phyletic Patterns Using Stochastic Mapping and Maximum Parsimony—A Simulation Study. Genome Biol. Evol. *3*, 1265–1275.

Cohen, O., Ashkenazy, H., Belinky, F., Huchon, D., and Pupko, T. (2010). GLOOME: gain loss mapping engine. Bioinformatics *26*, 2914–2915.

Colbourne, J.K., Pfrender, M.E., Gilbert, D., Thomas, W.K., Tucker, A., Oakley, T.H., Tokishita, S., Aerts, A., Arnold, G.J., Basu, M.K., et al. (2011). The Ecoresponsive Genome of Daphnia pulex. Science *331*, 555 –561.

Cucchi, T., Vigne, J.-D., and Auffray, J.-C. (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.

Czypionka, T., Cheng, J., Pozhitkov, A., and Nolte, A.W. (2012). Transcriptome changes after genome-wide admixture in invasive sculpins (Cottus). Mol. Ecol. *21*, 4797–4810.

D'haeseleer, P. (2006). What are DNA sequence motifs? Nat. Biotechnol. *24*, 423–425.

Danecek, P., Auton, A., Abecasis, G., Albers, C.A., Banks, E., DePristo, M.A., Handsaker, R.E., Lunter, G., Marth, G.T., Sherry, S.T., et al. (2011). The variant call format and VCFtools. Bioinformatics *27*, 2156–2158.

Delaye, L., DeLuna, A., Lazcano, A., and Becerra, A. (2008). The origin of a novel gene through overprinting in Escherichia coli. BMC Evol. Biol. *8*, 31.

Demuth, J.P., and Hahn, M.W. (2009). The life and death of gene families. BioEssays News Rev. Mol. Cell. Dev. Biol. *31*, 29–39.

Dintilhac, A., Bihan, R., Guerrier, D., Deschamps, S., and Pellerin, I. (2004). A conserved non-homeodomain Hoxa9 isoform interacting with CBP is co-expressed with the “typical” Hoxa9 protein during embryogenesis. Gene Expr Patterns *4*, 215 – 222.

Domazet- Loso, T., and Tautz, D. (2008). An ancient evolutionary origin of genes associated with human genetic diseases. Mol Biol Evol *25*, 2699 – 2707.

Domazet- Loso, T., and Tautz, D. (2010a). Phylostratigraphic tracking of cancer genes suggests a link to the emergence of multicellularity in metazoa. BMC Biol *8*, 66.

Domazet- Loso, T., and Tautz, D. (2010b). A phylogenetically based transcriptome age index mirrors ontogenetic divergence patterns. Nature *468*, 815 – 818.

Domazet- Loso, T., Brajkovic, J., and Tautz, D. (2007). A phylostratigraphy approach to uncover the genomic history of major adaptations in metazoan lineages. Trends Genet *23*, 533 – 539.

Domazet-Loso, T., and Tautz, D. (2003). An evolutionary analysis of orphan genes in Drosophila. Genome Res. *13*, 2213–2219.

Donoghue, M.T., Keshavaiah, C., Swamidatta, S.H., and Spillane, C. (2011). Evolutionary origins of brassicaceae specific genes in Arabidopsis thaliana. BMC Evol Biol *11*, 47.

Doolittle, R.F. (1986). Of Urfs and Orfs: A Primer on how to Analyze Derived Amino Acid Sequences (University Science Books).

Dreyfus, M., and Régnier, P. (2002). The poly(A) tail of mRNAs: bodyguard in eukaryotes, scavenger in bacteria. Cell *111*, 611–613.

Dujon, B. (1996). The yeast genome project: what did we learn? Trends Genet. TIG *12*, 263–270.

Dyson, H.J., and Wright, P.E. (2005). Intrinsically unstructured proteins and their functions. Nat Rev Mol Cell Biol *6*, 197 – 208.

Ebisuya, M., Yamamoto, T., Nakajima, M., and Nishida, E. (2008). Ripples from neighbouring transcription. Nat Cell Biol *10*, 1106–1113.

Ekman, D., and Elofsson, A. (2010). Identifying and quantifying orphan protein sequences in fungi. J Mol Biol *396*, 396 – 405.

ENCODE Project Consortium (2011). A user’s guide to the encyclopedia of DNA elements (ENCODE). PLoS Biol. *9*, e1001046.

ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. Nature *489*, 57–74.

ENCODE Project Consortium, Birney, E., Stamatoyannopoulos, J.A., Dutta, A., Guigó, R., Gingeras, T.R., Margulies, E.H., Weng, Z., Snyder, M., Dermitzakis, E.T., et al. (2007).

Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature *447*, 799–816.

Fang, Y., Bateman, J.F., Mercer, J.F., and Lamandé, S.R. (2013). Nonsense-mediated mRNA decay of collagen -emerging complexity in RNA surveillance mechanisms. J. Cell Sci. *126*, 2551–2560.

Fischer, D. (1999). Rational structural genomics: affirmative action for ORFans and the growth in our structural knowledge. Protein Eng. *12*, 1029–1030.

Fischer, D., and Eisenberg, D. (1999). Finding families for genomic ORFans. Bioinforma. Oxf. Engl. *15*, 759–762.

Flicek, P., Amode, M.R., Barrell, D., Beal, K., and Brent, S. (2011). Ensembl 2011. Nucleic Acids Res *39*, D800–D806.

Flicek, P., Amode, M.R., Barrell, D., Beal, K., Billis, K., Brent, S., Carvalho-Silva, D., Clapham, P., Coates, G., Fitzgerald, S., et al. (2013). Ensembl 2014. Nucleic Acids Res. *42*, D749–D755.

Fujita, P.A., Rhead, B., Zweig, A.S., Hinrichs, A.S., and Karolchik, D. (2011). The UCSC Genome Browser database: update 2011. Nucleic Acids Res *39*, D876–D882.

Gerstein, M.B., Bruce, C., Rozowsky, J.S., Zheng, D., Du, J., Korbel, J.O., Emanuelsson, O., Zhang, Z.D., Weissman, S., and Snyder, M. (2007). What is a gene, post-ENCODE? History and updated definition. Genome Res. *17*, 669–681.

Goios, A., Pereira, L., Bogue, M., Macaulay, V., and Amorim, A. (2007). mtDNA phylogeny and evolution of laboratory mouse strains. Genome Res. *17*, 293–298.

Graur, D., Zheng, Y., Price, N., Azevedo, R.B.R., Zufall, R.A., and Elhaik, E. (2013). On the Immortality of Television Sets: “Function” in the Human Genome According to the Evolution-Free Gospel of ENCODE. Genome Biol. Evol. *5*, 578–590.

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

Guerzoni, D., and McLysaght, A. (2011). De Novo Origins of Human Genes. PLoS Genet *7*, e1002381.

Haas, B.J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P.D., Bowden, J., Couger, M.B., Eccles, D., Li, B., Lieber, M., et al. (2013). *De novo* transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nat. Protoc. *8*, 1494–1512.

Haerty, W., Jagadeeshan, S., Kulathinal, R.J., Wong, A., Ram, K.R., Sirot, L.K., Levesque, L., Artieri, C.G., Wolfner, M.F., Civetta, A., et al. (2007). Evolution in the Fast Lane: Rapidly Evolving Sex-Related Genes in Drosophila. Genetics *177*, 1321–1335.

Hahn, M.W. (2009). Distinguishing among evolutionary models for the maintenance of gene duplicates. J. Hered. *100*, 605–617.

Haldane, J.B.S. (1932). The causes of evolution (Princeton, N.J.: Princeton University Press).

Hangauer, M.J., Vaughn, I.W., and McManus, M.T. (2013). Pervasive Transcription of the Human Genome Produces Thousands of Previously Unidentified Long Intergenic Noncoding RNAs. PLoS Genet *9*, e1003569.

Harr, B., and Turner, L.M. (2010). Genome-wide analysis of alternative splicing evolution among Mus subspecies. Mol. Ecol. *19 Suppl 1*, 228–239.

Hedges, S.B., Dudley, J., and Kumar, S. (2006). TimeTree: a public knowledge-base of divergence times among organisms. Bioinforma. Oxf. Engl. *22*, 2971–2972.

Heinen, T.J.A.J., Staubach, F., Häming, D., and Tautz, D. (2009). Emergence of a new gene from an intergenic region. Curr. Biol. CB *19*, 1527–1531.

Hu, J., and Ng, P.C. (2012). Predicting the effects of frameshifting indels. Genome Biol. *13*, R9.

Huang, C.R.L., Burns, K.H., and Boeke, J.D. (2012). Active transposition in genomes. Annu. Rev. Genet. *46*, 651–675.

Huang, D.W., Sherman, B.T., and Lempicki, R.A. (2009a). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat. Protoc. *4*, 44–57.

Huang, D.W., Sherman, B.T., and Lempicki, R.A. (2009b). Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res. *37*, 1–13.

Hunter, S., Jones, P., Mitchell, A., Apweiler, R., and Attwood, T.K. (2011). InterPro in 2011: new developments in the family and domain prediction database. Nucleic Acids Res *40*, D306–D312.

Innan, H., and Kondrashov, F. (2010). The evolution of gene duplications: classifying and distinguishing between models. Nat. Rev. Genet. *11*, 97–108.

Jacob, F. (1977). Evolution and tinkering. Science *196*, 1161–1166.

Jagadeeshan, S., and Singh, R.S. (2005). Rapidly evolving genes of Drosophila: differing levels of selective pressure in testis, ovary, and head tissues between sibling species. Mol. Biol. Evol. *22*, 1793–1801.

Jensen, T.H., Jacquier, A., and Libri, D. (2013). Dealing with Pervasive Transcription. Mol. Cell *52*, 473–484.

Jones, E.P., Skirnisson, K., McGovern, T.H., Gilbert, M.T.P., Willerslev, E., and Searle, J.B. (2012). Fellow travellers: a concordance of colonization patterns between mice and men in the North Atlantic region. BMC Evol. Biol. *12*, 35.

Kaessmann, H. (2010). Origins, evolution, and phenotypic impact of new genes. Genome Res *20*, 1313 – 1326.

Kaessmann, H., Vinckenbosch, N., and Long, M. (2009). RNA-based gene duplication: mechanistic and evolutionary insights. Nat Rev Genet *10*, 19–31.

Kapranov, P., Cheng, J., Dike, S., Nix, D.A., Duttagupta, R., Willingham, A.T., Stadler, P.F., Hertel, J., Hackermüller, J., Hofacker, I.L., et al. (2007). RNA Maps Reveal New RNA Classes and a Possible Function for Pervasive Transcription. Science *316*, 1484–1488.

Karolchik, D., Hinrichs, A.S., Furey, T.S., Roskin, K.M., and Sugnet, C.W. (2004). The UCSC Table Browser data retrieval tool. Nucleic Acids Res *32*, D493–D496.

Katayama, S., Tomaru, Y., Kasukawa, T., Waki, K., Nakanishi, M., Nakamura, M., Nishida, H., Yap, C.C., Suzuki, M., Kawai, J., et al. (2005). Antisense transcription in the mammalian transcriptome. Science *309*, 1564–1566.

Keane, T.M., Goodstadt, L., Danecek, P., White, M.A., Wong, K., Yalcin, B., Heger, A., Agam, A., Slater, G., Goodson, M., et al. (2011). Mouse genomic variation and its effect on phenotypes and gene regulation. Nature *477*, 289–294.

Keese, P.K., and Gibbs, A. (1992). Origins of genes - big-bang or continuous creation. Proc Natl Acad Sci USA *89*, 9489 – 9493.

Kent, W.J. (2002a). BLAT--the BLAST-like alignment tool. Genome Res. *12*, 656–664.

Kent, W.J. (2002b). BLAT--The BLAST-Like Alignment Tool. Genome Res. *12*, 656–664.

Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M., and Haussler, D. (2002). The human genome browser at UCSC. Genome Res. *12*, 996–1006.

Khalturin, K., Anton-Erxleben, F., Sassmann, S., Wittlieb, J., Hemmrich, G., and Bosch, T.C.G. (2008). A Novel Gene Family Controls Species-Specific Morphological Traits in Hydra. PLoS Biol *6*, e278.

Khalturin, K., Hemmrich, G., Fraune, S., Augustin, R., and Bosch, T.C.G. (2009). More than just orphans: are taxonomically-restricted genes important in evolution? Trends Genet. *25*, 404–413.

Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., and Salzberg, S.L. (2013). TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol. *14*, R36.

Kimmins, S., and Sassone-Corsi, P. (2005). Chromatin remodelling and epigenetic features of germ cells. Nature *434*, 583–589.

Kinsella, R.J., Kahari, A., Haider, S., Zamora, J., and Proctor, G. (2011). Ensembl BioMarts: a hub for data retrieval across taxonomic space. Database bar030.

Kleene, K.C. (2001). A possible meiotic function of the peculiar patterns of gene expression in mammalian spermatogenic cells. Mech. Dev. *106*, 3–23.

Kleene, K.C. (2005). Sexual selection, genetic conflict, selfish genes, and the atypical patterns of gene expression in spermatogenic cells. Dev. Biol. *277*, 16–26.

Klemke, M., Kehlenbach, R.H., and Huttner, W.B. (2001). Two overlapping reading frames in a single exon encode interacting proteins - a novel way of gene usage. EMBO J *20*, 3849 – 3860.

Knowles, D.G., and McLysaght, A. (2009). Recent *de novo* origin of human protein-coding genes. Genome Res *19*, 1752 – 1759.

Krakauer, D.C. (2000). Stability and evolution of overlapping genes. Evol. Int. J. Org. Evol. *54*, 731–739.

Kryazhimskiy, S., and Plotkin, J.B. (2008). The Population Genetics of dN/dS. PLoS Genet *4*, e1000304.

Krzywinski, M., Schein, J., Birol, İ., Connors, J., Gascoyne, R., Horsman, D., Jones, S.J., and Marra, M.A. (2009). Circos: An information aesthetic for comparative genomics. Genome Res. *19*, 1639–1645.

Kumar, S., and Hedges, S.B. (2011). TimeTree2: species divergence times on the iPhone. Bioinforma. Oxf. Engl. *27*, 2023–2024.

Kung, J.T.Y., Colognori, D., and Lee, J.T. (2013). Long Noncoding RNAs: Past, Present, and Future. Genetics *193*, 651–669.

Kutter, C., Watt, S., Stefflova, K., Wilson, M.D., Goncalves, A., Ponting, C.P., Odom, D.T., and Marques, A.C. (2012). Rapid Turnover of Long Noncoding RNAs and the Evolution of Gene Expression. PLoS Genet *8*, e1002841.

Lander, E.S. (2011). Initial impact of the sequencing of the human genome. Nature *470*, 187–197.

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S.L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. *10*, R25.

Lecompte, E., Aplin, K., Denys, C., Catzeflis, F., Chades, M., and Chevret, P. (2008). Phylogeny and biogeography of African Murinae based on mitochondrial and nuclear gene sequences, with a new tribal classification of the subfamily. BMC Evol. Biol. *8*, 199.

Levine, M.T., Jones, C.D., Kern, A.D., Lindfors, H.A., and Begun, D.J. (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.

Li, C.-Y., Zhang, Y., Wang Zhang, Y., Cao, C., and Zhang, P.W. (2010a). A human-specific *De novo* protein-coding gene associated with human brain functions. PLoS Comput Biol *6*, e1000734.

Li, D., Dong, Y., Jiang, Y., Jiang, H.F., and Cai, J. (2010b). A *de novo* originated gene depresses budding yeast mating pathway and is repressed by the protein encoded by its antisense strand. Cell Res *20*, 408 – 420.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and 1000 Genome Project Data Processing Subgroup (2009). The Sequence Alignment/Map format and SAMtools. Bioinforma. Oxf. Engl. *25*, 2078–2079.

Liang, H., and Landweber, L.F. (2006). A genome-wide study of dual coding regions in human alternatively spliced genes. Genome Res. *16*, 190–196.

Liao, Y., Smyth, G.K., and Shi, W. (2013). featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinforma. Oxf. Engl.

Lipman, D.J., Souvorov, A., Koonin, E.V., Panchenko, A.R., and Tatusova, T.A. (2002). The relationship of protein conservation and sequence length. BMC Evol Biol *2*, 20.

Lodish, H., Berk, A., Kaiser, C.A., Krieger, M., Scott, M.P., Bretscher, A., Ploegh, H., and Matsudaira, P. (2007). Molecular Cell Biology (W. H. Freeman).

Lohse, M., Bolger, A.M., Nagel, A., Fernie, A.R., Lunn, J.E., Stitt, M., and Usadel, B. (2012). RobiNA: a user-friendly, integrated software solution for RNA-Seq-based transcriptomics. Nucleic Acids Res. *40*, W622–627.

Long, M., Betran, E., Thornton, K., and Wang, W. (2003). The origin of new genes: glimpses from the young and old. Nat Rev Genet *4*, 865–875.

Managadze, D., Rogozin, I.B., Chernikova, D., Shabalina, S.A., and Koonin, E.V. (2011). Negative correlation between expression level and evolutionary rate of long intergenic noncoding RNAs. Genome Biol. Evol. *3*, 1390–1404.

Marques, A.C., and Ponting, C.P. (2009). Catalogues of mammalian long noncoding RNAs: modest conservation and incompleteness. Genome Biol. *10*, R124.

Marques-Bonet, T., Girirajan, S., and Eichler, E.E. (2009). The origins and impact of primate segmental duplications. Trends Genet. TIG *25*, 443–454.

Mayr, E. (1982). The growth of biological thought: diversity, evolution, and inheritance (Cambridge, Mass.: Belknap Press).

Moore, A.D., and Bornberg-Bauer, E. (2012). The dynamics and evolutionary potential of domain loss and emergence. Mol Biol Evol *29*, 787 – 796.

Mouse ENCODE Consortium, Stamatoyannopoulos, J., Snyder, M., Hardison, R., Ren, B., Gingeras, T., Gilbert, D., Groudine, M., Bender, M., Kaul, R., et al. (2012). An encyclopedia of mouse DNA elements (Mouse ENCODE). Genome Biol. *13*, 418.

Mugal, C.F., Wolf, J.B.W., and Kaj, I. (2014). Why time matters: codon evolution and the temporal dynamics of dN/dS. Mol. Biol. Evol. *31*, 212–231.

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

Nalavade, R., Griesche, N., Ryan, D.P., Hildebrand, S., and Krauß, S. (2013). Mechanisms of RNA-induced toxicity in CAG repeat disorders. Cell Death Dis. *4*, e752.

Näsvall, J., Sun, L., Roth, J.R., and Andersson, D.I. (2012). Real-Time Evolution of New Genes by Innovation, Amplification, and Divergence. Science *338*, 384–387.

Near, T.J., Parker, S.K., and Detrich, H.W., 3rd (2006). A genomic fossil reveals key steps in hemoglobin loss by the antarctic icefishes. Mol. Biol. Evol. *23*, 2008–2016.

Necsulea, A., Soumillon, M., Warnefors, M., Liechti, A., Daish, T., Zeller, U., Baker, J.C., Grützner, F., and Kaessmann, H. (2014). The evolution of lncRNA repertoires and expression patterns in tetrapods. Nature *505*, 635–640.

Neil, H., Malabat, C., d' Aubenton-Carafa, Y., Xu, Z., Steinmetz, L.M., and Jacquier, A. (2009). Widespread bidirectional promoters are the major source of cryptic transcripts in yeast. Nature *457*, 1038–1042.

Nekrutenko, A., Wadhawan, S., Goetting-Minesky, P., and Makova, K.D. (2005). Oscillating Evolution of a Mammalian Locus with Overlapping Reading Frames: An XLαs/ALEX Relay. PLoS Genet *1*, e18.

Neme, R., and Tautz, D. (2013). Phylogenetic patterns of emergence of new genes support a model of frequent *de novo* evolution. BMC Genomics *14*, 117.

Neme, R., and Tautz, D. (2014). Evolution: Dynamics of De Novo Gene Emergence. Curr. Biol. *24*, R238–R240.

Neme Garrido, R.T. (2011). Phylostratigraphic analyses of mouse tissue transcriptomes and comparative genomics of orphan genes. Master Thesis. Georg August University.

Ntini, E., Järvelin, A.I., Bornholdt, J., Chen, Y., Boyd, M., Jørgensen, M., Andersson, R., Hoof, I., Schein, A., Andersen, P.R., et al. (2013). Polyadenylation site-induced decay of upstream transcripts enforces promoter directionality. Nat. Struct. Mol. Biol. *20*, 923–928.

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

Ohno, S. (1972). So much "junk" DNA in our genome. Brookhaven Symp. Biol. *23*, 366–370.

Ohno, S. (1984). Birth of a unique enzyme from an alternative reading frame of the preexisted, internally repetitious coding sequence. Proc. Natl. Acad. Sci. U. S. A. *81*, 2421–2425.

Okazaki, Y., Furuno, M., Kasukawa, T., Adachi, J., Bono, H., Kondo, S., Nikaido, I., Osato, N., Saito, R., Suzuki, H., et al. (2002). Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. Nature *420*, 563–573.

Oliver, S.G., van der Aart, Q.J., Agostoni-Carbone, M.L., Aigle, M., Alberghina, L., Alexandraki, D., Antoine, G., Anwar, R., Ballesta, J.P., and Benit, P. (1992). The complete DNA sequence of yeast chromosome III. Nature *357*, 38–46.

Pal, L.R., and Guda, C. (2006). Tracing the origin of functional and conserved domains in the human proteome: implications for protein evolution at the modular level. BMC Evol Biol *6*, 91.

Palmieri, N., Kosiol, C., and Schlötterer, C. (2014). The life cycle of Drosophila orphan genes. eLife *3*.

Paradis, E., Claude, J., and Strimmer, K. (2004). APE: Analyses of Phylogenetics and Evolution in R language. Bioinformatics *20*, 289–290.

Pavesi, A., Magiorkinis, G., and Karlin, D.G. (2013). Viral proteins originated *de novo* by overprinting can be identified by codon usage: application to the “gene nursery” of Deltaretroviruses. PLoS Comput. Biol. *9*, e1003162.

Pereira, S.L., and Baker, A.J. (2006). A mitogenomic timescale for birds detects variable phylogenetic rates of molecular evolution and refutes the standard molecular clock. Mol. Biol. Evol. *23*, 1731–1740.

Peterson, G.I., and Masel, J. (2009). Quantitative prediction of molecular clock and ka/ks at short timescales. Mol. Biol. Evol. *26*, 2595–2603.

Polev, D. (2012). Transcriptional noise as a driver of gene evolution. J. Theor. Biol. *293*, 27–33.

Pozhitkov, A.E., Noble, P.A., Bryk, J., and Tautz, D. (2014). A revised design for microarray experiments to account for experimental noise and uncertainty of probe response. PloS One *9*, e91295.

Preker, P., Almvig, K., Christensen, M.S., Valen, E., Mapendano, C.K., Sandelin, A., and Jensen, T.H. (2011). PROMoter uPstream Transcripts share characteristics with mRNAs and are produced upstream of all three major types of mammalian promoters. Nucleic Acids Res.

Pyron, R.A. (2010). A Likelihood Method for Assessing Molecular Divergence Time Estimates and the Placement of Fossil Calibrations. Syst. Biol. *59*, 185–194.

Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841–842.

Quint, M., Drost, H.G., Gabel, A., Ullrich, K.K., Bonn, M., and Grosse, I. (2012). A transcriptomic hourglass in plant embryogenesis. Nature *490*, 98 – 101.

R Core Team (2012). R: A language and environment for statistical computing (R Foundation for Statistical Computing, Vienna, Austria).

Rancurel, C., Khosravi, M., Dunker, A.K., Romero, P.R., and Karlin, D. (2009). Overlapping genes produce proteins with unusual sequence properties and offer insight into *De novo* protein creation. J Virol *83*, 10719 – 10736.

Reinhardt, J.A., Wanjiru, B.M., Brant, A.T., Saelao, P., Begun, D.J., and Jones, C.D. (2013). De Novo ORFs in Drosophila Are Important to Organismal Fitness and Evolved Rapidly from Previously Non-coding Sequences. PLoS Genet *9*, e1003860.

Rice, P., Longden, I., and Bleasby, A. (2000). EMBOSS: the European Molecular Biology Open Software Suite. Trends Genet. TIG *16*, 276–277.

Di Roberto, R.B., and Peisajovich, S.G. (2014). The role of domain shuffling in the evolution of signaling networks. J. Exp. Zoolog. B Mol. Dev. Evol. *322*, 65–72.

Roberts, A., Pimentel, H., Trapnell, C., and Pachter, L. (2011). Identification of novel transcripts in annotated genomes using RNA-Seq. Bioinformatics *27*, 2325–2329.

Robins, J.H., McLenachan, P.A., Phillips, M.J., McComish, B.J., Matisoo-Smith, E., and Ross, H.A. (2010). Evolutionary relationships and divergence times among the native rats of Australia. BMC Evol. Biol. *10*, 375.

Rocha, E.P.C., Smith, J.M., Hurst, L.D., Holden, M.T.G., Cooper, J.E., Smith, N.H., and Feil, E.J. (2006). Comparisons of dN/dS are time dependent for closely related bacterial genomes. J. Theor. Biol. *239*, 226–235.

Ruf, S., Symmons, O., Uslu, V.V., Dolle, D., and Hot, C. (2011). Large-scale analysis of the regulatory architecture of the mouse genome with a transposon-associated sensor. Nat Genet *43*, 379 – 341.

Sabath, N., Wagner, A., and Karlin, D. (2012). Evolution of viral proteins originated *De novo* by overprinting. Mol Biol Evol *29*, 3767 – 3780.

Schlessinger, A., Schaefer, C., Vicedo, E., Schmidberger, M., and Punta, M. (2011). Protein disorder - a breakthrough invention of evolution. Curr Opin Struct Biol *21*, 412 – 418.

Sedlazeck, F.J., Rescheneder, P., and von Haeseler, A. (2013). NextGenMap: fast and accurate read mapping in highly polymorphic genomes. Bioinforma. Oxf. Engl. *29*, 2790–2791.

Seila, A.C., Calabrese, J.M., Levine, S.S., Yeo, G.W., Rahl, P.B., Flynn, R.A., Young, R.A., and Sharp, P.A. (2008). Divergent transcription from active promotors. Science *322*, 1849 – 1851.

Seila, A.C., Core, L.J., Lis, J.T., and Sharp, P.A. (2009). Divergent transcription: a new feature of active promotors. Cell Cycle *8*, 2557 – 2564.

Sherr, C.J. (2006). Divorcing ARF and p53: an unsettled case. Nat Rev Cancer *6*, 663 – 673.

Siepel, A. (2009). Darwinian alchemy: Human genes from noncoding DNA. Genome Res. *19*, 1693–1695.

Sorek, R., Lev-Maor, G., Reznik, M., Dagan, T., Belinky, F., Graur, D., and Ast, G. (2004). Minimal conditions for exonization of intronic sequences: 5' splice site formation in alu exons. Mol. Cell *14*, 221–231.

SPSS Inc (2009). PASW Statistics for Windows (Chicago: SPSS Inc.).

Stanke, M., Diekhans, M., Baertsch, R., and Haussler, D. (2008). Using native and syntenically mapped cDNA alignments to improve *de novo* gene finding. Bioinforma. Oxf. Engl. *24*, 637–644.

Stefflova, K., Thybert, D., Wilson, M.D., Streeter, I., Aleksic, J., Karagianni, P., Brazma, A., Adams, D.J., Talianidis, I., Marioni, J.C., et al. (2013). Cooperativity and Rapid Evolution of Cobound Transcription Factors in Closely Related Mammals. Cell *154*, 530–540.

Stephens, S.G. (1951). Possible Significance of Duplication in Evolution. In Advances in Genetics, (Elsevier), pp. 247–265.

Suzuki, H., Nunome, M., Kinoshita, G., Aplin, K.P., Vogel, P., Kryukov, A.P., Jin, M.-L., Han, S.-H., Maryanto, I., Tsuchiya, K., et al. (2013). Evolutionary and dispersal history of Eurasian

house mice Mus musculus clarified by more extensive geographic sampling of mitochondrial DNA. Heredity *111*, 375–390.

Tautz, D. (2008). Polycistronic peptide coding genes in eukaryotes - how widespread are they? Brief Funct Gen Proteom *8*, 68 – 74.

Tautz, D., and Domazet- Loso, T. (2011). The evolutionary origin of orphan genes. Nat Rev Genet *12*, 692 – 702.

Tautz, D., Neme, R., and Domazet-Lošo, T. (2013). Evolutionary Origin of Orphan Genes. In eLS, John Wiley & Sons, Ltd, ed. (Chichester, UK: John Wiley & Sons, Ltd),.

Taylor, J.S., and Raes, J. (2004). Duplication and divergence: the evolution of new genes and old ideas. Annu. Rev. Genet. *38*, 615–643.

Thorvaldsdóttir, H., Robinson, J.T., and Mesirov, J.P. (2013). Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Brief. Bioinform. *14*, 178–192.

Tilgner, H., Knowles, D.G., Johnson, R., Davis, C.A., Chakrabortty, S., Djebali, S., Curado, J., Snyder, M., Gingeras, T.R., and Guigó, R. (2012). Deep sequencing of subcellular RNA fractions shows splicing to be predominantly co-transcriptional in the human genome but inefficient for lncRNAs. Genome Res. *22*, 1616–1625.

Toll-Riera, M., Bosch, N., Bellora, N., Castelo, R., and Armengol, L. (2009). Origin of primate orphan genes: a comparative genomics approach. Mol Biol Evol *26*, 603 – 612.

Toll-Riera, M., Bostick, D., Albà, M.M., and Plotkin, J.B. (2012). Structure and age jointly influence rates of protein evolution. PLoS Comput. Biol. *8*, e1002542.

Tompa, P., and Kovacs, D. (2010). Intrinsically disordered chaperones in plants and animals. Biochem. Cell Biol.-Biochim. Biol. Cell. *88*, 167 – 174.

Tu, S., Shin, Y., Zago, W.M., States, B.A., Eroshkin, A., Lipton, S.A., Tong, G.G., and Nakanishi, N. (2007). Takusan: a large gene family that regulates synaptic activity. Neuron *55*, 69–85.

Vanderperre, B., Lucier, J.-F., Bissonnette, C., Motard, J., Tremblay, G., Vanderperre, S., Wisztorski, M., Salzet, M., Boisvert, F.-M., and Roucou, X. (2013). Direct Detection of Alternative Open Reading Frames Translation Products in Human Significantly Expands the Proteome. PLoS ONE *8*, e70698.

Vanin, E.F. (1985). Processed Pseudogenes: Characteristics and Evolution. Annu. Rev. Genet. *19*, 253–272.

Vilella, A.J., Severin, J., Ureta-Vidal, A., Heng, L., Durbin, R., and Birney, E. (2008). EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates. Genome Res. *19*, 327–335.

Wagner, A. (1998). The fate of duplicated genes: loss or new function? BioEssays News Rev. Mol. Cell. Dev. Biol. *20*, 785–788.

Wang, G.-Z., Lercher, M.J., and Hurst, L.D. (2011). Transcriptional coupling of neighbouring genes and gene expression noise: evidence that gene orientation and non-coding transcripts are modulators of noise. Genome Biol. Evol.

Wickham, H. (2009). Ggplot2 elegant graphics for data analysis (Dordrecht; New York: Springer).

Wilson, B.A., and Masel, J. (2011). Putatively Noncoding Transcripts Show Extensive Association with Ribosomes. Genome Biol. Evol.

Wilson, G.A., Bertrand, N., Patel, Y., Hughes, J.B., Feil, E.J., and Field, D. (2005). Orphans as taxonomically restricted and ecologically important genes. Microbiology *151*, 2499–2501.

Wilusz, C.J., Wormington, M., and Peltz, S.W. (2001). The cap-to-tail guide to mRNA turnover. Nat. Rev. Mol. Cell Biol. *2*, 237–246.

Wissler, L., Gadau, J., Simola, D.F., Helmkampf, M., and Bornberg-Bauer, E. (2013). Mechanisms and dynamics of orphan gene emergence in insect genomes. Genome Biol. Evol. *5*, 439–455.

Wolf, Y.I., Novichkov, P.S., Karev, G.P., Koonin, E.V., and Lipman, D.J. (2009). The universal distribution of evolutionary rates of genes and distinct characteristics of eukaryotic genes of different apparent ages. Proc Natl Acad Sci USA *106*, 7273 – 7280.

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

Wu, D.-D., and Zhang, Y.-P. Evolution and Function of De Novo Originated Genes. Mol. Phylogenet. Evol.

Wu, C., Macleod, I., and Su, A.I. (2013). BioGPS and MyGene.info: organizing online, gene-centric information. Nucleic Acids Res. *41*, D561–565.

Wu, D.-D., Irwin, D.M., and Zhang, Y.-P. (2011). De Novo Origin of Human Protein-Coding Genes. PLoS Genet *7*, e1002379.

Xu, Z., Wei, W., Gagneur, J., Perocchi, F., Clauder-Münster, S., Camblong, J., Guffanti, E., Stutz, F., Huber, W., and Steinmetz, L.M. (2009). Bidirectional promoters generate pervasive transcription in yeast. Nature *457*, 1033–1037.

Yalcin, B., Adams, D.J., Flint, J., and Keane, T.M. (2012). Next-generation sequencing of experimental mouse strains. Mamm. Genome *23*, 490–498.

Yang, Z. (2007). PAML 4: Phylogenetic Analysis by Maximum Likelihood. Mol. Biol. Evol. *24*, 1586–1591.

Yang, Z.F., and Huang, J.L. (2011). *De novo* origin of new genes with introns in Plasmodium vivax. FEBS Lett *585*, 641 – 644.

Zhang, J. (2003). Evolution by gene duplication: an update. Trends Ecol. Evol. *18*, 292–298.

Zhang, C., Wang, J., Long, M., and Fan, C. (2013). gKaKs: the pipeline for genome-level Ka/Ks calculation. Bioinformatics *29*, 645–646.

Zhang, W., Morris, Q., Chang, R., Shai, O., Bakowski, M., Mitsakakis, N., Mohammad, N., Robinson, M., Zirngibl, R., Somogyi, E., et al. (2004). The functional landscape of mouse gene expression. J. Biol. *3*, 21.

Zhao, L., Saelao, P., Jones, C.D., and Begun, D.J. (2014). Origin and Spread of de Novo Genes in Drosophila melanogaster Populations. Science 1248286.

Zhou, Q., Zhang, G.J., Zhang, Y., Xu, S.Y., and Zhao, R.P. (2008). On the origin of new genes in Drosophila. Genome Res *18*, 1446 – 1455.

Chapter contributions

Chapter 1

This study was designed from preliminary results and ideas obtained during my Master's Thesis. I downloaded and analyzed the data from the specified repositories. No new data were generated for the purposes of this project. The interpretation of the results and manuscript were done together with Prof. Tautz. Sebastian Meyer was involved during the early development of the overprinting analyses, although the published results derived from a portion of the analyses were exclusively performed by me. The analyses of phylostratigraphic distribution of testes-expressed genes were derived from initial analyses started during my Master Thesis (Neme Garrido, 2011).

Chapters 2 – 4

This study was designed from analyses performed by me, but discussed together with Prof. Tautz. It involved the generation of new sequences. However, during the initial phase of the project I used also available RNA-Seq data from wild mouse species and subspecies (Harr and Turner, 2010). The analyses performed with these data are not included given the superior quality of the most recent data.

The animals were sacrificed and dissected by me from the stock collection at the Max Planck Institute for Evolutionary Biology, with the support of Christine Pfeifle and members of her team responsible for the collection.

RNA and DNA extractions were performed by Nicole Thomsen. The sequencing scheme was coordinated together with Dr. Janine Altmüller from the Cologne Center for Genomics (CCG), and Christian Becker was responsible for the execution of the sequencing. The CCG provided the data in the form of sequence reads.

I carried out all the mentioned analyses and data processing steps regarding downstream analyses. The only exception was the snpEff pipeline, for which Chen Ming kindly processed some of the samples at my request.

The microarray analyses that complemented the orphan gene curation procedure (see Appendix B) were performed following the protocol established by Alexander Pozhitkov in the Department of Evolutionary Genetics. I was responsible for the sacrifice, dissection of tissues and RNA extractions. Sarah Lemke provided help with RNA extractions. Labeling, hybridization and scans of samples were performed by Elke Blohm-Sievers. Data processing and

downstream analyses were carried out by me, with support from Dr. Till Czypionka and Dr. Alexander Pozhitkov.

In addition to the data analyses regarding the Chapters 1 through 4, I was also involved, together with Prof. Tautz, in the production of two reviews that integrate current ideas about gene birth from a wide variety of model systems and case-studies (Neme and Tautz, 2014; Tautz et al., 2013). A great part of those ideas is included in the general introduction of this thesis, and can be considered the result of discussions with Prof. Tautz during my time in his group.

Appendices

Appendix A. Phylostratigraphic maps

Phylostratigraphic maps including gene ages at the Ensemble Gene level for human, zebrafish, stickleback and mouse are available at:

http://www.biomedcentral.com/content/supplementary/1471-2164-14-117-s1.xlsx

Appendix B. Curation data from orphan genes

This orphan gene curation procedure is based on data available prior to the start of the genome and transcriptome sequencing projects (November 2012).

Mouse orphan genes were defined as those classified as annotated protein-coding genes in phylostratum 20 from the first chapter (Neme and Tautz, 2013). In order to obtain the most reliable orphan genes, a curation process was performed taking into consideration different sources, representing the wide variety of available technologies and experimental setups:

Full-length ESTs from the GenBank (Benson et al., 2004), as mapped to the mouse genome (mm9) and as displayed in the UCSC Genome Browser (Kent, 2002b), were considered sufficient if two or more different libraries (different organ, sex, tissue, developmental stage) had transcripts matching orphans.

RNA-Seq reads for six different tissues from the mouse (Brawand et al., 2011) (GEO accession GSE30352) were mapped onto the genome using bowtie (Langmead et al., 2009), and absolute expression counts were obtained for each orphan gene with samtools (Li et al., 2009). Genes were considered to be expressed if any predicted exon contained at least 15 uniquely

mapped reads, however if no other source had evidence, only genes with at least 100 reads were considered.

Agilent microarray evidence from 55 mouse tissues (Zhang et al., 2004) were obtained from the authors' website (http://hugheslab.ccbr.utoronto.ca/). Since the annotation scheme is based on an outdated version of the mouse genome, the protein sequences were re-annotated according to version 66 of Ensembl for the mouse (BLAT, tile size 2, 98% identity). Binary expression values based on expression above the 99% of the intensities of the negative controls were used.

Data from the Affymetrix Mouse MOE430 Gene Atlas were obtained from the bioGPS website (Wu et al., 2013) (GEO accession GSE10246). Replicate experiments of the same tissue and probes covering the same transcript were averaged.

Presence/absence determination array. Custom arrays (SurePrint G3 Custom GE, 1x1M) from Agilent Technologies were designed to cover all Ensembl transcripts available for version 64. Each transcript had two different probes on the array, and each probe was spotted eight times. Seven independent single-color microarray experiments were performed, each with a different final concentration of labeled RNA (LowInput QuickAmp Labeling Kit, Agilent Technologies). RNA was extracted using the TRIzol method (as indicated by the manufacturer) from different tissues (brain, testes, lung, heart, liver, spleen, kidney, muscle, bone) and three different ages (1 month, 8 months, 20 months) from three different animals per age (C57BL/6J-Rj, JANVIER). Labeling was performed in independent aliquots which were subsequently pooled. The concentration ranges were 0.125x, 0.25x, 0.5x, 1x, 2x, 4x and 8x. Data analyses were performed upon the raw intensity values, due to the fact that the lower and higher concentrations tend to be highly modified by the signal processing algorithms. Concentration dependent changes (following linear or log-linear behavior) were recorded for each probe based on the average signal from spots. Probes were classified as responsive (intensity as a function of the concentration) and non-responsive (signal changes independent of concentration). We assumed that intensity changes of responsive probes were due to the actual presence of the target transcript in the sample. With this approach we overcome the problem of detecting low expression transcripts in high-throughput experiments, and differentiating weak signals from absent targets. This array experiment was done in agreement with the protocols devised by Alex Pozhitkov (Czypionka et al., 2012; Pozhitkov et al., 2014).

Automatic classification proceeded for genes with two or more sources of evidence (listing each EST library as an independent source of evidence). The remaining genes were evaluated by hand whenever any of the sources had evidence.

Further curation proceeded by removing any potential non-orphan which had ComparaOrtholog information in any vertebrate. For the remaining genes, homology searches were performed against collections of rat and primate ESTs (megablast, 1e-10) (Altschul et al., 1997).

From 781 phylostratum 20 genes, 526 orphan genes with more than one source of evidence, or with manually curated evidence in cases of single sources. From these 526, there was no evidence of cryptic homology in human or rat for 438 genes. I concluded that these 438 are mouse orphans of the highest quality.

Relevant information about the features of these genes and their sources of evidence is available at:

Appendix C. Functional annotation clusters based on known genes with loss of expression

This table contains the four most significant clusters based on literature, shared domains and Gene Ontology terms among other classifiers. The full table can be downloaded from http://www.evolbio.mpg.de/~rneme/CAU/dissertation_RafikNeme/DAVID_functional_annotation.xlsx

Annotation Cluster 1	**Enrichment Score: 15.679433906271832**									
Category	**Term**		**Count**	**%**	**PValue**	**List Total**	**Pop Hits**	**Pop Total**	**Fold Enrichment**	**FDR**
PUBMED_ID		11802173	36	9.302326	2.54E-17	195	1273	41427	6.007904	4.40E-14
PUBMED_ID		11875048	36	9.302326	2.78E-16	195	1378	41427	5.550117	5.77E-13
PUBMED_ID		14611657	36	9.302326	1.30E-15	195	1442	41427	5.303787	2.31E-12
Notes	Olfactory receptors are common elements among the literature titles found.									
Annotation Cluster 2	**Enrichment Score: 11.493611015192476**									
Category	**Term**		**Count**	**%**	**PValue**	**List Total**	**Pop Hits**	**Pop Total**	**Fold Enrichment**	**FDR**
PUBMED_ID		18815613	8	2.067183	5.24E-15	195	10	41427	169.9569	9.05E-12
PUBMED_ID		18064011	8	2.067183	1.48E-13	195	14	41427	121.3978	2.56E-10
INTERPRO	IPR002971:Mus/Rat 1 allergen		8	2.067183	1.60E-12	146	12	17763	81.10959	2.01E-09
INTERPRO	IPR002345:Lipocalin		8	2.067183	8.56E-08	146	46	17763	21.15902	1.08E-04
Notes	Pseudogenized MUPs (major urinary proteins) with pheromone potential									
Annotation Cluster 3	**Enrichment Score: 7.80675705856884**									
Category	**Term**		**Count**	**%**	**PValue**	**List Total**	**Pop Hits**	**Pop Total**	**Fold Enrichment**	**FDR**

GOTERM_BP_ALL	GO:0007608~sensory perception of smell	32	8.268734	1.46E-10	110	1117	14219	3.703166	2.17E-07
GOTERM_BP_ALL	GO:0007606~sensory perception of chemical stimulus	32	8.268734	7.33E-10	110	1192	14219	3.470165	1.08E-06
GOTERM_BP_ALL	GO:0007600~sensory perception	32	8.268734	3.53E-08	110	1402	14219	2.950383	5.22E-05
GOTERM_BP_ALL	GO:0050890~cognition	32	8.268734	1.22E-07	110	1480	14219	2.794889	1.81E-04
GOTERM_BP_ALL	GO:0050877~neurological system process	32	8.268734	2.01E-06	110	1681	14219	2.4607	0.002972
Annotation Cluster 4	**Enrichment Score: 5.717239145049761**								
Category	**Term**	**Count**	**%**	**PValue**	**List Total**	**Pop Hits**	**Pop Total**	**Fold Enrichment**	**FDR**
SP_PIR_KEYWORDS	receptor	43	11.11111	8.32E-09	124	2465	17854	2.511686	9.78E-06
GOTERM_MF_ALL	GO:0004872~receptor activity	43	11.11111	2.53E-06	124	2606	15404	2.049773	0.003116
GOTERM_MF_ALL	GO:0004871~signal transducer activity	43	11.11111	2.53E-05	124	2851	15404	1.873627	0.031119
GOTERM_MF_ALL	GO:0060089~molecular transducer activity	43	11.11111	2.53E-05	124	2851	15404	1.873627	0.031119

Appendix D. Transcriptome information and statistics

	Geographical Origin	Breeding Scheme	Pooled individuals	Sequenced Reads (million reads)	Mapping Efficiency (including multireads)*	Uniquely mapping reads*
Mus musculus domesticus	Iran	Wild derived outbred	8	371.7	94.4%	45%
	France	Wild derived outbred	8	393.3	98.7%	44%
	Germany	Wild derived outbred	8	397.9	98.7%	44%
Mus musculus musculus	Kazakhstan	Wild derived outbred	8	371.9	95.4%	47%
	Austria	Wild derived outbred	8	378.6	99.9%	46%
Mus musculus castaneus	Taiwan	Wild derived inbred	4	367.5	98.5%	46%
Mus spretus	Spain	Wild derived inbred	4	379.0	99.2%	45%
Mus spicilegus	Ukraine	Wild derived inbred	4	383.0	96.8%	45%
Mus (Nannomys) mattheyii	Ivory Coast	Commercially derived inbred	4	366.2	84.5%	48%
Apodemus uralensis	Kazakhstan	Wild derived outbred	4	386.5	79.0%	44%

* Reads mapped to UCSC mm10 mouse genome. Regions marked as N's are not included in the calculations

Curriculum Vitae

Name	Rafik Tarek Neme Garrido
Date and place of birth	24.12.1985 / Bogotá, Colombia
Nationality	Colombian
Place of residence	Rautenbergstraße 55, 24306 Plön, Germany
Civil status	Single

Education

February 1989 – November 2002	Primary, Middle and High School German School Barranquilla Barranquilla, Colombia
October 2001 – January 2002	School Exchange Kaiserin-Auguste-Viktoria Gymnasium Celle, Germany
January 2005 – June 2009	Bachelor of Sciences in Biology National University of Colombia Bogota, Colombia
September 2009 – March 2011	Master of Sciences in Molecular Biology Georg-August-University of Göttingen International Max Planck Research School (IMPRS) Molecular Biology Göttingen, Germany
October 2010 – March 2011	External Master's Thesis Max Planck Institute for Evolutionary Biology Plön, Germany
Since April 2011	Doctoral Studies Max Planck Institute for Evolutionary Biology Department of Evolutionary Genetics IMPRS for Evolutionary Biology Christian-Albrecht-University of Kiel Plön, Germany

Stipends / Awards

2009 – 2010	Stipend of the Excellence Foundation for the Promotion of the Max Planck Society
2013	100 Colombianos 2013, Colombia Country Brand (Marca País Colombia) and Fusionarte Foundation

Affidavit

Hiermit erkläre ich, dass die vorliegende Arbeit

nach Inhalt und Form meine eigene ist, abgesehen von der Beratung durch meinen Betreuer Prof. Diethard Tautz

an keiner anderen Stelle im Rahmen eines Prüfungsverfahrens vorgelegen hat, noch nicht veröffentlicht ist und auch nicht zur Veröffentlichung eingereicht wurde

unter Einhaltung der Regeln guter wissenschaftlicher Praxis der Deutschen Forschungsgemeinschaft entstanden ist.

Plön, den 7.8.2014

[Rafik Tarek Neme Garrido]