Pseudogenes in the mouse lineage: transcriptional activity and strain-specific history

Cristina Sisu*[1,2,3], Paul Muir*[1], Adam Frankish[4], Ian Fiddes[5], Mark Diekhans[5], David Thybert[4,6] Duncan T. Odom[7,8], and Paul Flicek[4,9], Thomas Keane[4], Mark Gerstein[1,2,10]

Pseudogenes are ideal markers of genome remodelling. In turn, the mouse is an ideal platform for studying them, particularly with the availability of transcriptional time course data during development (just completed in phase 3 of ENCODE) and the sequencing of 18 strains (completed by the Mouse Genome Project). Here we present a comprehensive genome-wide annotation of the pseudogenes in the mouse reference genome and associated strains. We compiled this by combining manual curation of over 10,000 pseudogenes with results from automatic annotation pipelines. Also, by comparing human and mouse, we annotated 217 new unitary pseudogenes in human and 237 unitary pseudogenes in mouse. (We make our annotation available through a resource website mouse.pseudogene.org.) The overall mouse pseudogene repertoire (in the reference and strains) is similar to human in terms of overall size, biotype distribution (~80% processed, 20% duplicated) and top family composition (with many GAPDH and ribosomal pseudogenes). However, notable differences arise in the age distribution of pseudogenes with multiple retro-transpositional bursts in mouse evolutionary history and only a single one in human. Furthermore, in each strain ~20% of the pseudogenes are unique, reflecting strain-specific functions and evolution – e.g. the pseudogenization of taste receptors is clearly linked to a change in the diet of the NZO strain. Finally, we find ~15% of the pseudogenes are transcribed, a fraction similar to human. Furthermore, we show that processed pseudogenes are commonly associated with highly transcribed genes. While this can be observed through all of mouse development, the relationship is strongest not at the early embryo stages but later on, after depletion of maternal RNA.

Introduction

The mouse is one of the most widely studied model organisms \cite{17173058}, with the field of mouse genetics counting for more than a century of studies towards the understanding of mammalian physiology and development \cite{12586691,12702670}. Recent advances of the Mouse Genome Project \cite{22772437,21921910} towards completing the de-novo assembly and gene annotation of a variety of mouse strains, provide a unique opportunity to get an in-depth picture of the evolution and variation of these closely related mammalian species.

Mice have frequently been used as a model organism for the study of human diseases due to their experimental tractability and similarities in their genetic makeup \cite{14978070}. This has been achieved through the development of mouse models of specific diseases or the creation of knockout mice to recapitulate the phenotype associated with a loss of function mutation observed in humans. The advent of high-throughput sequencing has led to the emergence of population and comparative genomics as new windows into the relationship between genotype and phenotype amongst the human population. Current efforts to catalog genetic variation amongst closely related mouse strains extend this paradigm.

Since their divergence around 90 million years ago (MYA) \cite{12651866,12466850,11214318,11214319, 17021158, 26589719}, the human and mouse lineages followed a comparable evolutionary pattern \cite{17284675}. While it is hard to make a direct comparison between the two species, there is a large range of divergence in the mouse population, with some even approaching the human-chimp divergence levels in terms of the number of intervening generations \cite{17284675} (Figure 1A). The mouse strains under investigation have differences in their genetic makeup that manifest in an array of phenotypes, ranging from coat/eye color to predisposition for various diseases \cite{21921910}. Moreover, the creation of these strains has been extensively documented. Following a well-characterized inbreeding process for 20 sequential generations, the inbred mice are homozygous at all loci and show a high level of consistency at genomic and phenotypic levels \cite{JAX}. The repeated inbreeding resulted in substantial differences between the mouse strains, giving each strain the potential to offer a unique reaction to an acquired mutation

Deleted: been

Comment [G1]: Inserted: been

Comment [G3]: Deleted:been

Deleted:

Comment [G2]: Inserted: -

Deleted:

Comment [G4]: Inserted: -

Deleted:

\cite{19710643}. The use of inbred mice also minimizes a number of problems raised by the genetic variation between animals \cite{11528054}.

To uncover key genome remodeling processes that governed mouse strain evolution, we focus our analysis on the study of mouse pseudogene complements, while also highlighting their shared features with the human genome. In particular, here we describe the first pseudogene annotation and analysis of 18 widely-used inbred mouse strains alongside the reference mouse genome. Additionally, we provide the latest updates on the pseudogene annotation for both the mouse and human reference genomes, with a particular emphasis on the identification of unitary pseudogenes with respect to each organism.

Often regarded as genomic relics, pseudogenes provide an excellent perspective on genome evolution \cite{10692568,11160906,12034841,14616058}. Pseudogenes are DNA sequences that contain disabling mutations rendering them unable to produce a fully functional protein. Different classes of pseudogenes are distinguished based on their creation mechanism: processed pseudogenes – formed through a retrotransposition process, duplicated pseudogenes – formed through a gene duplication event, and unitary pseudogenes – formed by the inactivation of a functional gene. From a functional perspective, pseudogenes can be classified into three categories: (1) dead-on-arrival – elements that are nonfunctional and are expected, in time, to be eliminated from the genome; (2) partially active – pseudogenes that exhibit residual biochemical activity; and (3) exapted pseudogenes – commonly represented by transcribed pseudogenes, are elements that have acquired new functions and can interfere with the regulation and activity of protein-coding genes. Finally, there are polymorphic pseudogenes – elements that are present in the population as both functional and nonfunctional alleles \cite{20210993}. Such pseudogenes represent disablements that have occurred on a much more recent timescale. They are loss of function mutations that are not fixed in the human population and still subject to evolutionary pressures.

Moreover, pseudogenes reflect changes in selective pressures and genome remodeling forces. Duplicated pseudogenes can reveal the history of gene duplication, one of the key mechanisms of establishing new gene functions \cite{14671323}. While the majority of duplicated gene copies are dead-on-arrival \cite{27690225}, successful paralogs can acquire new functions during a phase of rapid, almost neutral evolution \cite{17053091}, a process known as neofunctionalization \cite{Ono1970}. Furthermore, duplicated pseudogenes can also arise at a later point from functional duplicated gene copies when these are subjected to negative selection due to an increased gene dosage \cite{11864370,25197576}. Processed pseudogenes inform on the evolution of gene expression as well as the history of transposable element activity, while unitary pseudogenes are indicative of gene families that died out by acquiring loss of function mutations that became fixed in the population. Thus, pseudogenes can play an important role in evolutionary analysis as they can be regarded as markers for loss of function events.

A loss-of-function (LOF) event is a mutation that results in a modified gene product that lacks the molecular function of the wild-type gene \cite{JAX2}. Unitary pseudogenes are an extreme case of LOF, where mutations that result in complete inactivation of the gene are fixed in the population. In recent years, LOF mutations have become a key research topic in genomics. In general, loss of a functional gene is detrimental to an organism's fitness. However, there are also numerous examples showcasing evolutionary advantages for the accumulation and fixation of LOF mutations resulting in formation of new pseudogenes. As such pseudogenes can reflect either advantageous or deleterious phenotypes. For example, the pseudogenization of proprotein convertase subtilisin/kexin type 9 (PCSK9) in human evolution is commonly associated with a reduced risk of heart diseases by lowering the plasma low-density lipoprotein (LDL) levels. This is achieved by preventing the expression PCSK9 protein and its subsequent binding to and degradation of cellular LDL receptor \cite{18631360}. By contrast, its gain of function mutations resulting in the expression of PCSK9 are commonly associated with an enrichment in plasma LDL cholesterol and an increased risk of atherosclerosis for the affected individuals. This finding has inspired the creation of PCSK9 inhibitors as treatment for high cholesterol and highlights the potential for the investigation of pseudogenes to shed light on biological processes of interest to the biomedical and pharmaceutical industry \cite{24958078}.

Comment [G5]: Inserted: -

Deleted: this paper

Comment [G6]: Inserted: e

Comment [G7]: Inserted: r, h

Comment [G8]: Inserted: cul

Comment [G9]: Inserted: par

Comment [G10]: Inserted: mouse

Comment [G11]: Deleted:h

Comment [G12]: Deleted:s p

Comment [G13]: Deleted:p

Comment [G14]: Inserted: (1)

Deleted: ,

Comment [G15]: Inserted: (3)

Comment [G16]: Inserted: ;

Comment [G17]: Inserted: ; (2)

Comment [G19]: Deleted:,

Comment [G20]: Deleted:,

Deleted: ,

Deleted:

Comment [G18]: Inserted: -

Formatted: Highlight

Deleted: the

Formatted: Highlight

Comment [G21]: Deleted:the

Comment [G22]: Deleted:the

Deleted: the

Formatted: Highlight

Deleted:

Deleted: organism's

Deleted: , however

Comment [G23]: Inserted: ,

Comment [G24]: Inserted: H

Comment [G25]: Inserted: .

Comment [G26]: Inserted: -

Comment [G27]: Deleted:,

Comment [G28]: Deleted:h

Deleted: ,

Comment [G29]: Deleted:,

Taken together the well-defined evolutionary relationships between the mouse strains and the wealth of associated functional data from the recently completed ENCODE 3 project present an opportunity to investigate the processes underlying pseudogene biogenesis and activity to an extent previously not possible. In particular, we can explore the creation of pseudogenes during early embryo development. We leverage timecourse RNAseq data to investigate whether pseudogenization occurs in the gametes or earlier in development in a germline precursor. Also, comparison to the primate lineage and human population is a possibility as the evolutionary distance between some of the mouse strains parallels the human-chimp divergence as well as distances between the modern-day human populations in terms of generations, making the collection of high-quality genomes and associated pseudogene annotations for the 18 strains a valuable resource for population studies.

Results

1. Annotation

We present the latest pseudogene annotations for the mouse reference genome as part of the GENCODE project, as well as updates on the human pseudogene reference set. Leveraging the recently assembled high-quality genome sequences for the mouse strains we introduce the first draft annotation of the pseudogene complement in the 18 strains.

1.1 Reference genome

Using a combination of rigorous manual curation \cite{22951037,25157146} and automatic identification \cite{16574694} we were able to annotate a comprehensive set of pseudogenes for the mouse reference genome (Table 1, S1). However, pseudogene assignments are highly dependent on the quality of the protein coding annotation. Thus, the current manually curated set provides a high quality lower bound with respect to the true number of pseudogenes in the mouse genome, while the [[union of]] automatic annotation informs on the upper limit of the pseudogene complement size. In agreement with our previous work \cite{22951037,25157146} there is a considerable overlap, of over 83%, between the manual and automatic annotation sets.

Similarly, for human we [[have??]] a similar workflow to refine the reference pseudogene annotation to a high-quality set of 14,650 pseudogenes. The updated set contains considerable improvements in the characterization of pseudogenes of previously unknown biotype (Table S2). In both the human and mouse reference genomes more than half of the annotations are processed pseudogenes, with a smaller fraction of duplicated pseudogenes (Table S2).

1.2 Mouse strains

The Mouse Genome Project has sequenced and assembled genomes for 18 mouse strains, and developed a draft annotation of the strains' protein coding genes \cite{MousePaper}. The strains are broadly organized into 3 classes (Table 2): the outgroup strains – formed by two independent mouse species, *Mus Caroli* and *Mus Pahari*; wild strains – covering two subspecies (*Mus Spretus* and *Mus Castaneus*) and two musculus strains (*Mus Musculus Musculus* and *Mus Musculus Domesticus*), and a set of laboratory strains. A detailed summary of the genome composition for each strain is presented in \cite{MousePaper}.

We developed an annotation workflow for identifying pseudogenes in the 18 mouse strains leveraging our in-house automatic pipelines and the set of manually curated pseudogenes from the mouse reference genome lifted over onto each individual strain (Figure 1C). The combined pseudogene identification process gives rise to three levels reflecting the annotation quality. Each identified pseudogene is provided with details about the transcript biotype, genomic location, structure, sequence disablements, and a confidence level reflecting the annotation process. A detailed overview of pseudogene annotation statistics including the number of pseudogenes, their confidence levels, and related biotypes is shown in Figure 1B.

Currently, around 30% of pseudogenes in each strain are defined as high confidence Level 1 annotations, being identified through both automatic curation and manual lift over, 10% Level 2 characterized only using the lift over process, and 60% Level 3 resulted solely from the automatic

Deleted: are able to
Comment [G30]: Inserted: n
Comment [G31]: Inserted: c
Comment [G34]: Deleted:are
Comment [G35]: Deleted:ble to
Deleted:
Deleted:
Comment [G32]: Inserted: -
Comment [G33]: Inserted: -
Deleted:
Comment [G36]: Inserted: -
Comment [G37]: Inserted: [[union of]]
Comment [G38]: Inserted: [[have??]]
Deleted: the
Deleted:
Deleted: pipeline
Comment [G39]: Inserted: s
Comment [G40]: Inserted: -
Comment [G41]: Inserted: our
Comment [G42]: Deleted:the

annotation pipeline. The pseudogene biotype distribution across the strains closely follows the reference genome and is consistent with the biotype distributions observed in other mammalian genomes (e.g. Human \cite{22951037} and macaque \cite{25157146}). As such, the bulk (~80%) of the annotations are processed pseudogenes, while a smaller fraction (~15%) are duplicated. Finally, the distribution of pseudogene disablements follows the previously observed distributions in the mouse reference genome and other mammals, with stop codons being the most frequent defect per base pair followed by deletions and insertions (Figure S1). As expected, older pseudogenes show an enrichment in the number of disablements compared with the parental gene sequence. The proportion of pseudogene defects exhibits a linear inverse correlation with the pseudogene age, expressed as the sequence similarity between the pseudogene and the parent gene.

1.3 Unitary pseudogenes

Unitary pseudogenes are the result of a complex interplay between loss-of-function events and changes in selective pressures resulting in the fixation of an inactive element in a species. The importance of unitary pseudogenes resides not only in their ability to mark loss-of-function events but also in their potential to highlight changes in the genome evolution. Due to their formation mechanism as a result of gene inactivation, the identification of unitary pseudogenes is highly dependent on the quality of the reference genome protein coding annotation and requires a large degree of attention during the annotation process.

These pseudogenes are defined relative to the functional protein-coding elements in another species. Using a combination of multi-sequence alignments, and a specialized unitary pseudogene annotation workflow (Figure 1C), we identified 218 new unitary pseudogenes in human and 237 unitary pseudogenes in mouse (Table S3). In human, a large proportion of unitary pseudogenes are related to the chemosensory system (e.g. GPCRs, olfactory and vomeronasal receptor proteins) and have functional homologs in mouse, reflecting the loss of function in these genes during the primate lineage evolution.

Moreover, we observed the pseudogenization of a number of innate immune response related genes in humans such as Toll-like receptor gene 11 and leucine rich repeat protein genes hinting at potentially advantageous LOF/pseudogenization events in human lineage evolution \cite{22724060}. By contrast, the majority of mouse unitary pseudogenes with respect to human, are associated with structural Zinc finger domains, Kruppel associated box proteins, and immunoglobulin V-set proteins (Table S4).

Moreover, to get an overview of the unitary pseudogenes in each strain, we lifted over the reference pseudogene annotation and checked their conservation as pseudogenes or functional genes in each strain. We identified on average 200 unitary pseudogenes. [[wrt ??]] However, the short evolutionary distance between most strains means this value is an overestimate of the number of unitary pseudogenes. One way to get a more realistic assessment of the size of the unitary pseudogene complement is to look at the unitary annotation in the human genome relative to chimp.

2. Conservation and divergence in pseudogene complements

To decipher the evolutionary history of the mouse strains we created a *pangenome* pseudogene dataset containing 49,262 unique entries relating the pseudogenes across strains. We found 2,925 ancestral pseudogenes that are preserved across all strains. A detailed summary of the other subsets of pseudogenes is shown in Figure 2A,B. On average, each strain contains between 1,000 and 3,000 strain-specific pseudogenes. The proportion of pseudogenes conserved only in the outgroup, the wild strains, or the lab strains is considerably smaller, suggesting that the bulk of the pseudogenes in each strain are derived during the shared evolutionary history.

Next, we took advantage of pseudogenes' ability to evolve with little or no selective constraints \cite{10833048}, and compared the mutational processes across the mouse strains. To this end, we built a phylogenetic tree based on approximately 3,000 pseudogenes that are conserved across all strains (Figure 2C). This pseudogene-based tree follows closely the protein coding genes tree and correctly identifies and clusters the strains into three classes: outgroup, wild, and laboratory strains.

Deleted: pseudogenes

Comment [G43]: Deleted: pseudogenes

Deleted: ,

Deleted: ,

Comment [G44]: Deleted:,

Comment [G45]: Deleted:,

Deleted:

Deleted:

Comment [G46]: Inserted: -

Comment [G47]: Inserted: -

Deleted: 1C

Comment [G48]: Inserted: [[wrt ??]]

Comment [G49]: Inserted: T

Comment [G51]: Deleted:In order t

Deleted: In order to

Deleted:

Comment [G50]: Inserted: -

Comment [G52]: Inserted: ,

Deleted: 2C

Furthermore, we grouped the conserved pseudogenes into subgroups based on their parent gene families (e.g. olfactory receptors, CDK, Ribosomal proteins, etc.) and constructed pseudogene phylogenetic trees for each of these subgroups (Figure 2C). By comparing the resulting trees to the protein-coding one, we found that they display different patterns, reflecting different rates of evolution and evolutionary histories for different gene families. For example, the olfactory receptor tree, shows discrepancies in both the divergence order as well as in the degree of conservation of the ancestral sequence (as reflected by the branch length), with notable differences observed for NZO, and NOD laboratory strains. These two strains are known for exhibiting a diabetic phenotype. The result suggests that changes in diet can affect chemoreceptors and consequently their evolution \cite{25943692}.

3. Genome Evolution & Plasticity

Leveraging the pseudogene annotations, we explore the differences between the mouse strains by looking at the genome-remodelling processes that shaped the evolutionary history of their pseudogene complements.

3.1 Pseudogene Genesis

Taking advantage of the available functional genomics and evolutionary data we can study pseudogene Genesis on a unique scale: during embryo development at one extreme and the mouse lineage at the other.

Given that processed pseudogenes are formed through the retrotransposition of the parent mRNAs, we hypothesized that there is a direct correlation between the parent gene expression level and the number of processed pseudogenes. Moreover, as pseudogenes are inherited, the genesis of new elements should occur in the germline. To this end, we used an embryogenesis RNA-seq time course to test our assumptions during early development \cite{27309802}. We calculated the parent gene expression for a series of developmental stages ranging from metaphase II oocytes to the inner cell mass. At every stage, the average expression level of parent genes is higher than the one observed for regular protein-coding genes. However, genes associated with large pseudogene families show low transcription levels during very early development, with high expression levels achieved only during later stages. This can be related to the fact that during very early development, maternal RNA accounts for the largest proportion of embryonic RNA, with only a smaller fraction resulting from the actual gene transcription. We evaluated the correlation between the number of pseudogenes associated with a gene and its expression level at different developmental time-points. This correlation improves as we move forward through the developmental stages suggesting that pseudogenes are most likely generated by highly expressed housekeeping genes.

We further tested the correlation between high expression levels and a large number of associated pseudogenes by looking at RNA-seq samples from adult mouse brain. Similar to our previous observations, the pseudogene parent genes show a statistically significant increase in average expression levels compared to non-pseudogene generating protein-coding genes (Figure SF3).

Next, we looked at the degree to which the number of pseudogenes is related to the number of copies or functional paralogs of the parent gene (Figure 3A). For duplicated pseudogenes, we observe a weak correlation between the number of paralogs and the number of pseudogenes of a particular parent gene. This result suggests that a highly-duplicated protein family will tend to give rise to more disabled copies than a less duplicated family. [[This is consistent with the assumption that]] if we assume that each duplication process can potentially give rise to either a pseudogene of a functional gene.

By contrast, for processed pseudogenes, we observed a weak inverse correlation. This result implies that in the case of large protein families we can expect to see a lower level of transcription for each family member, with high mRNA abundance being achieved from multiple duplicated copies of the gene rather than increasing the expression of a single unit. Therefore, there is a weak correlation between the number of paralogs of the parent and the potential gene expression level of the parent genes, and thus we observe a smaller number of associated pseudogenes.

3.2 Transposable elements

Comment [G53]: Inserted: for different gene families
Comment [G54]: Inserted: rates of evolution and
Deleted: .
Formatted: Highlight
Comment [G55]: Inserted: -
Comment [G56]: Deleted:l
Deleted: remodelling
Deleted: are able to
Deleted: the
Deleted: genesis
Comment [G57]: Inserted: G
Comment [G58]: Inserted: n
Comment [G59]: Inserted: c
Comment [G60]: Deleted:are
Comment [G61]: Deleted:ble to
Comment [G62]: Deleted: the
Comment [G63]: Deleted:g
Deleted:
Comment [G64]: Inserted: -
Comment [G65]: Inserted: ,
Comment [G66]: Inserted: a
Comment [G67]: Inserted: -
Deleted:
Deleted: .
Comment [G68]: Inserted: . [[This is consistent with the assumption that]]
Comment [G69]: Deleted:,
Comment [G70]: Inserted: ,
Comment [G71]: Inserted: the
Comment [G72]: Inserted: ,

To the extent that majority of mouse and human pseudogenes are the result of retrotransposition processes mediated by transposable elements (TE), we investigated the genome mobile element content in the two species on an evolutionary time scale (Figure 3B).

TEs are sequences of DNA characterized by their ability to integrate themselves at new loci within the genome. TEs are commonly classified into two classes: DNA transposons and retrotransposons, with the latter being responsible for the formation of processed pseudogenes and retrogenes. Both human and mouse genomes are dominated by three types of TEs, namely short interspersed nuclear elements (SINEs), long interspersed nuclear elements (LINEs) and the endogenous retrovirus (ERV) superfamily. LINE-1 elements (L1) have been shown to mobilize Alu's, small nuclear RNAs and mRNA transcripts. We analyzed the LINE, SINE and ERV content in the human and mouse processed pseudogene complements. We define the evolutionary time scale by using the pseudogene sequence similarity to the parent gene as a proxy for age. Younger pseudogenes have a higher degree of sequence similarity to the parent, while older pseudogenes show a more diverged sequence.

In human, we observe a smooth age distribution of L1 flanked processed pseudogenes, with a single peak (at 92.5% sequence similarity to parents) hinting at the burst of retrotransposition events, that occurred 40 MYA at the dawn of primate lineage and created the majority of human pseudogene content. By contrast in mouse, we found the L1 derived pseudogene age distribution is defined by two successive peaks at 92.5 and 97% sequence similarity to parents. Also by contrast to human where the density of L1 associated pseudogenes shows a steep decrease for young pseudogenes following the peak at 92.5, the density of mouse pseudogenes remains at high levels in the interval of 97 to 100% sequence similarity to parents. This observation suggests the presence of highly active transposable elements in mouse. The TE activity results in a continuous renewal of the processed pseudogene pool. This behavior is also reflected in the large difference in the number of active LINE/L1s between human and mouse, with just over 100 for human \cite{12682288} vs. 3,000 for mouse \cite{ 11591644}.

3.3 Genome remodeling

The large proportion of strain and class specific pseudogenes, as well as the presence of active TE families, point towards multiple genomic rearrangements in mouse genome evolution. To this end, we examined the conservation of pseudogene genomic loci between each of the mouse strains and the reference genome for one-to-one pseudogene orthologs in each pair (Figure 4A,B). We observed that, on average, more than 97.7% of loci are conserved across the laboratory strains, and 96.7% of loci are conserved with respect to the wild strains. By contrast, only 87% of Caroli loci were conserved in the reference genome, while Pahari showed only 10% conservation. The significant drop in the number of conserved pseudogene loci between the reference genome and outgroup strains is in accord with the observed major karyotype-scale differences and large genomic rearrangements exhibited by Caroli and Pahari \cite{doi.org/10.1101/088435}. The proportion of un-conserved loci follows a logarithmic curve that matches closely the divergent evolutionary time scale of the mouse strains suggesting a uniform rate of genome remodeling processes across the murine taxa (Figure 4C).

4. Biological relevance[[better head??]]

The role of pseudogenes in genome biology has long been debated, however, recent studies \cite{25157146} have highlighted the fact the pseudogenes can reflect the evolution of genome function and activity. Here we address the biological relevance of pseudogene activity leveraging data from gene ontology, protein families, and RNA-seq experiments.

Comment [G73]: Inserted: .
Deleted:)
Deleted: analysed
Comment [G74]: Inserted: z
Comment [G75]: Deleted:s
Comment [G76]: Inserted: age
Comment [G77]: Inserted: e ag
Comment [G78]: Inserted: ,
Comment [G79]: Inserted: .
Comment [G80]: Inserted: ,
Comment [G81]: Inserted: ,
Comment [G82]: Inserted: ,
Comment [G83]: Inserted: ,
Comment [G84]: Inserted: ,
Comment [G85]: Inserted: [[better head??]]
Deleted: ,
Comment [G86]: Inserted: ;
Comment [G88]: Deleted:,
Comment [G87]: Inserted: ,

4.1 Gene ontology & pseudogene family analysis

We integrated the annotations with gene ontology (GO) data in order to characterize the functions associated with pseudogene generation. For this, we calculated the enrichment of GO terms across the strains. We observed that the majority of top biological processes, molecular function, and cellular component GO terms are shared across the strains (Figure 5A). Moreover, the GO terms that universally characterize the pseudogene complements in all the mouse strains are closely reproduced in the family classification of pseudogenes. The top pseudogene family 7-Transmembrane encompasses the chemoreceptors GPCR proteins, reflecting the mouse genome enrichment in olfactory receptors. Similar to the human and primate counterparts, the top families seen in mouse pseudogenes are related to highly expressed proteins such as GAPDH, ribosomal proteins and Zinc fingers (Figure 5B).[[not sure znf highly expr]]

However, a closer look suggests that the pseudogene repertoire also reflects particular strain-specific phenotypes. A detailed list of the strain specific and strain enriched pseudogenes families, strain-specific phenotypes, and strain-specific molecular and cellular GO-defined processes is shown in Table 3. There are two possible types of pseudogene-phenotype associations. [[expl??]]First, the pseudogenization process is linked with the emergence of an advantageous phenotype, as in the case of *Mus Spretus* genome, where we see an enrichment of pseudogenes related to tumor repressor genes and apoptosis pathways genes. Second, as expected, we find that the majority of pseudogenes reflect a deleterious phenotype. A known example is the pseudogenization of Cytochrome c Oxidase subunit VIa through accumulation of LOF mutations in the blind albino mouse strain, that is commonly linked with neurodegeneration \cite{17435251} and is characteristic for the observed brain lesions in the affected mice \cite{JAX}.

4.2 Gene essentiality

We observed an enrichment of essential genes among pseudogene parent genes in the mouse strains. Evaluating the parent gene for each pseudogene present in the mouse strains reveals essential genes are approximately three times more abundant amongst parent genes. In general, the essential genes are more highly transcribed than nonessential ones, and thus can potentially be associated with a higher propensity of generating processed pseudogenes. We evaluated the probability that a gene is essential given its transcription level and parent gene status (see Methods) and found that pseudogene parents are ~20% more likely to be essential genes compared to regular protein-coding genes.

We also analysed the number of paralogs associated with essential and nonessential genes to get an insight into the possible role of gene duplication in the enrichment of essential genes amongst the parent genes set. In the reference mouse, 19.4% of nonessential genes and 25.9% of essential genes lack paralogs. Meanwhile, there isn't a large difference in the average number of paralogs per genes, 6.2 for essential compared to 6.7 for nonessential genes. The slight depletion of genes with paralogs in the experimentally determined essential gene set is likely to be linked to the experimental set up relying on single gene knockouts to determine essentiality, which would miss genes with an essential role and a functional paralog.

4.3 Pseudogene Transcription

We leveraged the available RNA-seq data from the Mouse Genome Project and ENCODE 3 to study pseudogene biology as reflected by their transcription potential. Pseudogene transcription is thought to either relate to the exaptive functionality of pseudogenes or be a residual leftover from their existence as genes. For both the human and the mouse reference genomes, we detected that about 15% of pseudogenes were transcribed across a variety of tissues, a result similar to previous pan tissue analyses (Figure 6A,B). Due to data availability for the 18 mouse strains, we restricted our tissue analysis to the brain. Similar to the previously observed pattern in human and other model organisms, pseudogene transcription in mouse strains shows higher tissue and strain specificity compared to the protein coding

Comment [G89]: Inserted: ,
Comment [G90]: Inserted: ,
Deleted: Figure
Comment [G91]: Inserted: ,
Comment [G92]: Inserted: [[not sure znf highly expr]]
Deleted:).
Deleted: individual
Deleted:
Deleted:
Deleted:
Comment [G93]: Inserted: [[expl??]]
Comment [G94]: Inserted: -
Comment [G95]: Inserted: -
Comment [G96]: Inserted: -
Comment [G97]: Inserted: ar
Comment [G98]: Inserted: c
Comment [G99]: Inserted: part
Comment [G100]: Deleted:indiv
Comment [G101]: Deleted:d
Comment [G102]: Deleted:a
Deleted: genes
Comment [G103]: Inserted: -
Comment [G104]: Inserted: ~
Comment [G105]: Inserted: o
Comment [G106]: Deleted:ge
Deleted:
Deleted: isn’t
Comment [G107]: Inserted: ,
Deleted: This
Comment [G108]: Inserted: on
Comment [G109]: Inserted: Pseudogene transcript
Comment [G113]: Deleted:Th
Comment [G114]: Deleted:s
Comment [G110]: Inserted: o
Comment [G115]: Deleted:wi
Comment [G116]: Deleted:h
Deleted: with

counterpart (Supplementary Figure SF4). Also, pseudogenes with strain-specific transcription were more common than those with cross-strain transcription.

The pseudogenes conserved across all strains show a uniform level of transcription. However, the proportion of transcribed pseudogenes in the brain is half (2.5%) that observed across the entire dataset. Moreover, for strain-specific pseudogenes, the fraction of transcribed elements varies across the strains (Supplementary Figure SF4).

5. Mouse pseudogene resource

We created a pseudogene resource that organizes all of the pseudogenes across the available mouse strains and the reference genome, as well as associated phenotypic information in a MySQL database (Figure 6). All the available data is also provided as flat files for ease of manipulation. The database contains three types of information: details about the annotation, comparisons of the pseudogenes across strains, and phenotypic information associated with the pseudogenes and the corresponding mouse strains. Each pseudogene is given a unique universal identifier as well as a strain-specific ID in order to facilitate both the comparison of specific pseudogenes across strains and collective differences in pseudogene content between strains. To facilitate a direct comparison between human and mouse we also provide orthology links between each mouse entry and the corresponding human counterpart. A flat file for each pseudogene annotation containing all pertinent associated information will also be available. Queries on specific pseudogenes will return the relevant flat file.

Pseudogene annotation information encompasses the genomic context of each pseudogene, its parent gene and transcript Ensembl IDs, the level of confidence in the pseudogene as a function of agreement between manual and automated annotation pipelines, and the pseudogene biotype.

Information on the cross-strain comparison of pseudogenes is derived from the liftover of pseudogene annotations from one strain to another and subsequent intersection with that strain's native annotations. This enables pairwise comparisons of pseudogenes between the various mouse strains and the investigation of differences between multiple strains of interest. The database provides both liftover annotations and information about intersections between the liftover and native annotations.

Links between the annotated pseudogenes, their parent genes, and relevant functional and phenotypic information help inform biological relevance. In the database, the Ensembl ID associated with each parent gene is linked to the appropriate MGI gene symbol, which serves as a common identifier to connect to the phenotypic information. These datasets include information on gene essentiality, Pfam families, GO terms, and transcriptional activity. Furthermore, paralogy and homology information provide links between human biology and the well-characterized mouse strain collection.

Discussion

We describe the annotation and comparative analysis of the updated pseudogene complement in the mouse reference genome and the first draft of the pseudogene complements in 18 related strains. By combining manual curation and an automatic annotation pipeline, we were able to obtain a comprehensive view of the pseudogene content in genomes throughout the mouse lineage. The overlap between manually curated pseudogene sets and those identified using computational methods is over 80% reflecting the high sensitivity of the computational detection methods.

A high-level comparison of pseudogene statistics for each of the strains highlights shared properties of pseudogene biogenesis. Each of the strains exhibits a consistent ratio of processed to duplicated pseudogenes, which is in line with previous observations in human. The higher proportion of processed pseudogenes is in agreement with earlier findings that retrotransposition is the primary mechanism for pseudogene creation in numerous mammalian species \cite{22951037}.

Integrating the annotations from the mouse strains we obtained a pan-genome pseudogene set composed of over 45,000 unique entries. This set contains three types of pseudogenes: universally conserved, multi-strain, and strain specific, accounting for 6, 23, and 71% of the elements respectively. Comparative analysis of the pseudogenes in the pan-genome set provides a picture of the genome remodeling processes that have occurred in the mouse lineage. The lack of conservation of pseudogenes'

Comment [G111]: Inserted: the
Deleted:
Comment [G112]: Inserted: -
Comment [G117]: Inserted: the
Deleted: of
Deleted:
Comment [G118]: Inserted: -
Comment [G119]: Deleted:of
Deleted:
Comment [G120]: Inserted: -
Comment [G121]: Inserted: the
Deleted: In order to
Comment [G122]: Inserted: T
Comment [G123]: Deleted:In order t
Comment [G124]: Inserted: -
Deleted:
Comment [G125]: Inserted: ,
Deleted: exhibit
Comment [G126]: Inserted: s
Deleted:
Comment [G127]: Inserted: -
Deleted: pseudogenes'

chromosomal location between strains hints at multiple large-scale genomic rearrangements in the mouse lineage. This is especially striking in the case of *Mus Pahari* as has been recently reported by large-scale chromosomal imagining and karyotype analysis \cite{https://doi.org/10.1101/088435}.

Examination of the pseudogene complement reveals retrotransposon activity, how it contributed to pseudogene creation, and how it shaped the genomic environment of each strain over time. Sequence analysis reveals that while the majority of human pseudogenes have been obtained relatively recently through a single burst of retrotransposition \cite{22951037}, the mouse lineage shows a sustained renewal of the pseudogene pool through continuous transposable element activity. Looking closely at the sequence context of the processed pseudogenes indicates that the various retrotransposons exhibit differential contributions to the pseudogene set over time.

Analysis of pseudogenes and their parent genes can provide a window into changing functional constraints and selective pressures. Unitary pseudogenes are markers of loss of function mutations that have become fixed in the population. Here we annotated over 200 new unitary pseudogenes in mouse and a similar number in human. We found that the enrichment of vomeronasal receptor unitary pseudogenes in human with respect to mouse highlights the loss of certain olfactory functions in humans. Moreover, the unitary analysis is especially interesting because it provides us with key moments in the evolution of gene function by marking the loss and gain of function events. A known example of fixed LOF mutation in a human with respect to the mouse is the pseudogenization of Cyp2G1 gene (Figure 7A). Here the human gene acquired a C-T mutation resulting in a stop codon in the middle of a coding exon resulting the gene disablement and thus the creation of a unitary pseudogene. By contrast, in Caroli we observed a A-G gain of function mutation for the NCR3 gene that is pseudogenized in all the other mouse strains including the reference, reverting the initial TGA stop to a tryptophan codon (Figure 7B).

Since a processed pseudogene's likelihood of creation is related to its parent's expression level, they can act as a record of their parent gene's expression level and perhaps provide insight into the past importance of their parent gene. The link between the creation of processed pseudogenes and parent genes associated with key biological functions is further supported by an enrichment of parent genes amongst mouse essential genes. Meanwhile, duplicated pseudogenes record duplication events that shaped both the genome environment and function during the organism's evolution. Furthermore, the wealth of functional genomics assays available for the experimentally relevant mouse strains presents an opportunity to investigate both the activity of parent genes as well as pseudogene genesis. As expected parent genes have higher levels of expression relative to non-parents both during embryo development as well as in adult tissue. Moreover, time series expression analysis during embryo development suggest that most pseudogene creation is commonly related to the high expression levels of housekeeping genes.

The analysis of the functional annotations enriched amongst parent genes highlights key biological processes across the mouse lineage. We utilized both gene ontology terms and Pfam families to annotate parent gene function. Looking at Pfam families overrepresented amongst conserved pseudogenes we see an enrichment for housekeeping functions as illustrated by the presence of GapDH, ribosomal protein families, and zinc finger nucleases. These top Pfam families amongst the mouse pseudogenes closely match those seen in the human. Studying recurrent gene ontology terms supports the enrichment of pseudogenes for important biological processes with top GO terms including RNA processing and metabolic processes. Additionally, using the pan-genome pseudogene set to identify strain-specific functional annotations can suggest hypotheses as to what cellular processes and genes might underpin phenotypic differences between the mouse strains. PWK is associated with strain-specific GO terms for melanocyte-stimulating hormone receptor activity and melanoblast proliferation, which may play a role in the strain's patchwork coat color \cite{10385914}. NZO, an obesity-prone mouse strain, is characterized by a specific enrichment in defensin associated pseudogenes. Defensins are small peptides involved in controlling the inflammation resulted from metabolic abnormalities in obesity and type 2 diabetes \cite{25991648}, and more recently described as potential markers of obesity \cite{26929193}. Taken together the functional analysis of pseudogenes provides an opportunity to understand better the selective pressures that have shaped an organism's genomic content and phenotype.

Deleted:
Comment [G128]: Inserted: -
Comment [G129]: Inserted: -
Deleted:
Comment [G132]: Deleted:that
Deleted: that
Formatted: Highlight
Formatted: Highlight
Comment [G130]: Inserted: the
Formatted: Highlight
Comment [G131]: Inserted: the
Formatted: Highlight
Formatted: Highlight
Deleted: matches
Deleted: set
Deleted:
Deleted:
Comment [G133]: Inserted: -
Comment [G134]: Inserted: -
Comment [G137]: Deleted:es
Comment [G138]: Deleted: set
Deleted:
Comment [G135]: Inserted: -
Deleted: better
Comment [G136]: Inserted: better
Comment [G139]: Deleted:better
Deleted: organism's

Meanwhile, looking at pseudogene expression across the strains we observe evidence of both pseudogenes with broadly conserved transcription as well as some with strain-specific expression. As additional RNA-seq datasets for multiple tissues for each strain become available future work can investigate both pan-strain and pan tissue expression patterns.

This comprehensive annotation and analysis of pseudogenes across 18 mouse strains has provided support for conserved aspects of pseudogene biogenesis while also expanding our understanding of pseudogene evolution and activity. Integration of the pseudogene annotations with existing knowledge bases including Pfam and the gene ontology have provided insight into the biological functions associated with pseudogenes and their parent genes. The well-defined relationships between the strains aided evolutionary analysis of the pseudogene complements. The experimental and functional genomics datasets associated with these well-studied strains shed light on the transcriptional activity of pseudogenes and offer promise for future studies.

Deleted:

Comment [G140]: Inserted: -

Comment [G141]: Inserted: -

Deleted:

Tables

Table 1. Reference genome pseudogene annotation in mouse and human.

Organism	Manual	Pseudopipe*	Manual Overlap (%)
Mouse	10,524	18,649	8,786 (83.5)
Human	14,650	15,978	13,177 (89.9)

*Chromosomal assembled DNA only

Formatted Table

Table 2. Mouse strains description and nomenclature.

Strain ID	Description	Class
Pahari	PAHARI/EiJ – Mus Pahari	Outgroup
Caroli	CAROLI/EiJ – Mus Caroli	
Spret	SPRET/EiJ – Mus Spretus	Wild strains
PWK	PWK/J – Mus Musculus Musculus	
Cast	CAST/EiJ – Must Castaneus	
WSB	WSB/J – Mus Musculus Domesticus	
NOD	NOD/ShiLtJ – Non-obese Diabetic	Lab Strains
C57BL	C57BL/6NJ – Black 6N	
NZO	NZO/HlLtJ – New Zealand Obese	
AKR	AKR/J	
BALB	BALB/cJ	
A	A/J	
CBA	CBA/J	
C3H	C3H/HeJ	
DBA	DBA/2J	
LP	LP/J	
FVB	FVB/NJ	
129S1	129S1/SvImJ	

Formatted Table

Table S1. Reference genome pseudogene annotation in mouse and human.

Organism	Manual	PseudoPipe			
		Autosomes	Sex Chromosomes	Others*	Total
Mouse	10,524	14,084	4,565	4,162	22,811
Human	14,650	14,644	1,325	2,098	18,067

*Includes patches, scaffolds, and unassembled DNA.

Formatted Table

Table S2. Human and mouse pseudogene annotation summary.

	Human (v25)	Mouse (M12)
Total GENCODE	**14,650**	**10,524**
processed pseudogenes	10,725	7,486
unprocessed pseudogenes	3,400	2,625
unitary pseudogenes	214	34
polymorphic pseudogenes	51	77
ambiguous pseudogenes	21	99
Total PseudoPipe	**18,067**	**22,811**
processed pseudogenes	8,739	10,516
unprocessed pseudogenes	3,118	2,201
ambiguous pseudogenes	6,198	10,094

Formatted Table

Table S3. Unitary pseudogenes in human and mouse. (see SupTable_S3_Unitary.xlsx)

Table S4. Pseudogene family and clan characterization. (see SupTable_S4_Family.xlsx)

Methods

Datasets

Mouse reference genome is based on the Mus Musculus strain C57BL/6J strain. The mouse reference annotation is based on GENCODE vM12.

The human reference genome annotation is based on GENCODE v25.

The 16 laboratory and wild strains (Table 2) assemblies and strain specific annotations were obtained from the Mouse Genome Project \cite{MainMousePaper}. The laboratory strain C57BL/6NJ is a subline of the reference strain \cite{JAX} and is used here as the laboratory strain reference.

The two outgroup mouse species, Mus Caroli and Mus Pahari were sequenced, assembled, and annotated by the Flicek lab.

Human – Mouse Lineage Comparison

Human – primate lineage divergence and generation times were obtained from \cite{22891323}. The divergence times for the wild and laboratory strains were obtained from \cite{24608277,7284675,25038446}. While data for two outgroup species speciation was obtained from \cite{Flicek}. The generation time for all the mice was estimated from \cite{Jax}.

Pseudogene Annotation

Reference genome annotation

We manually curated almost 10,000 pseudogenes in the mouse reference genome (GENCODE M12) using a workflow as previously described \cite{22951037,25157146}.

The number of manually annotated pseudogenes in the mouse lineage is likely an underestimate of the true size of the mouse pseudogene complement given the similarities between the human and mouse genomes, and the fact that in human we have manually identified over 14,000 pseudogenes. Thus, to get a more accurate idea of the number of pseudogenes in the mouse genome, we use the in house annotation pipeline PseudoPipe \cite{16574694}. PseudoPipe is a comprehensive annotation pipeline focused on identifying and characterizing pseudogenes based on their biotypes as either processed or duplicated. The computational pipeline identifies approximately 22,000 pseudogenes of which about 14,000 are present in autosomal chromosomes (a number comparable with the one observed previously in human (Table S1)).

Mouse strain annotation

We used as input the conserved protein coding genes between each mouse strain and the reference genome. The number of shared transcripts follows an evolutionary trend with more distant strains having a smaller number of common protein coding genes with the reference genome compared with more closely related laboratory strains. PseudoPipe was run with the strain conserved protein set as shown in Figure 1C. Next, we used HAL tools package \cite{ 23505295} to lift over the manually annotated pseudogenes from the mouse reference genome onto each strain using the UCSC multi strain sequence alignments. We merged the two annotation set using BEDTools \cite{ 25199790} with 1bp minimum overlap requirement. We extended each overlap predicted boundaries to ensure full annotation of the pseudogene transcript.

Unitary Pseudogene Annotation Pipeline

We modified PseudoPipe to allow cross-strains and cross species protein coding inputs. We annotated cross-organism pseudogenes as shown in Fgure 1C. “Functional organism” is defined as the genome providing the protein coding information and thus containing a working copy of the element of interest. “Non-functional” organism as the genome analysed for unitary pseudogene presence. The resulting data set was subjected to a number of filters:

- removal of previously known pseudogenes,

- removal of pseudogenes with parents that have orthologs in the annotated specie,
- removal of pseudogenes that overlap with annotated protein coding and ncRNAs loci,
- removal of pseudogenes shorter than 200 bp.

Conservation and divergence in pseudogene complemenets

Pangenome data set generation

We performed an all against all liftover of pseudogene annotation using HAL tools package and the UCSC multi strain sequence alignment. Each liftover was intersected with the know strain annotation and all the entries that matched protein coding or ncRNAs were removed. The resulting set is further filtered for:

- conservation of pseudogene identity,
- conservation of parent gene identity,
- conservation of pseudogene locus,
- conservation of pseudogene biotype,
- conservation of pseudogene length,
- conservation of pseudogene structure.

All the filtered binary mappings were integrated in a master set. The common entries were merged into a unique pangeome pseudogene reference. We obtained 49,262 pangenome pseudogenes. A number of 1,158 pangenome entries are multi matching across strains.

Phylogenetic analysis

Sequences of the 1,460 randomly selected conserved pseudogenes in the 18 mouse strains were extracted and assembled in a strain specific contig. The multi-sequence alignment of the 18 contigs was obtained using MUSCLE aligner \cite{15318951} under standard conditions. Similarly, the parent protein coding genes of the 1,460 pseudogenes were assembled into a strain specific sequence and aligned using MUSCLE. The tree was generated using Tamura-Nei genetic distance model and neighbouring-joining tree build method with Pahari as outgroup using GENEIOUS 10.2 software package.

Genome evolution and plasticity

Parent gene expression analysis
RNAseq adult mouse brain data was obtained from ENCODE3. We estimated the pseudogene parent protein coding genes expression levels using a workflow involving the following steps: filtering the protein coding genes for uniquely mappable regions longer than 100bp, mapping reads using TopHat \cite{23618408}, selecting high quality mapped reads, calculating the expression FPKM levels using Cufflinks \cite{ 22383036}.

PM – can you please add a similar short description for the developmental stages ?

Transposable elements
TE in human and mouse reference genomes were identified using RepeatMasker 3.2.8 (http://repeatmasker.org).

Gene essentiality enrichment analysis

Lists of essential and nonessential genes were compiled using data from the MGI database and recent work from the International Mouse Phenotyping Consortium \cite{27626380}. The nonessential gene set with Ensembl identifiers contained 4,736 genes compared to 3,263 essential genes.

In order to evaluate the impact of parent gene status on the probability of a gene being essential while controlling for transcription we fit a linear probability model and a probit model for the probability that a gene is essential given its transcription level and parent gene status.

Pseudogene transcription

We estimated the pseudogene transcription levels for the mouse reference in 18 adult tissues using a protocol previously described \cite{25157146} using RNAseq ENCODE data.

Pseudogene transcription levels in the mouse strains were calculated in a similar manner using Mouse Genome Project RNAseq data from adult brain.