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], Paul Flicek[4,9], Thomas Keane[4], Mark Gerstein[1,2,10]

[-1] SEE WORD DOC. [NEW UNITARY + NZO TASTE]

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 can be linked to a change in the diet. Additionally we show that processed pseudogenes are commonly associated with highly transcribed genes. Finally, we find that ~15% of the pseudogenes are transcribed, a fraction similar to human, and that pseudogenes exhibit greater tissue and strain specificity compared to their protein coding counterparts.

( )

+ txn

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 organisms.

Mice have been frequently 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 parallel 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 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 \cite{10615122}. Following a well characterized inbreeding process for at least 20 sequential generations, the inbred mice are homozygous at nearly all loci and show a high level of consistency at genomic and phenotypic levels \cite{JAX}. This helps minimize a number of problems raised by the genetic variation between research animals \cite{11528054}. The repeated inbreeding has

Style Definition: Normal: Font:(Default) Times New Roman, 12 pt, Font color: Auto, English (US), Line spacing: single, Border:Top: (No border), Bottom: (No border), Left: (No border), Right: (No border), Between : (No border)

Style Definition: Heading 1: Font:(Default) Arial, Font color: Black, English (UK), Line spacing: multiple 1.15 li, Border:Top: (No border), Bottom: (No border), Left: (No border), Right: (No border), Between : (No border)

Style Definition: Heading 2: Font:(Default) Arial, Font color: Black, English (UK), Line spacing: multiple 1.15 li, Border:Top: (No border), Bottom: (No border), Left: (No border), Right: (No border), Between : (No border)

Style Definition ... [9]

Style Definition ... [8]

Style Definition ... [7]

Style Definition ... [6]

Style Definition ... [5]

Style Definition ... [4]

Style Definition ... [3]

Style Definition ... [2]

Style Definition ... [1]

Deleted: is

Deleted: clearly

Deleted:

Deleted: of the NZO strain

Deleted: Finally,

Deleted: Investigating we find ~15% of the transcript ... [10]

Deleted: are transcribed, a fraction similar to human. ... [11]

Moved down [2]: find ~15% of the pseudogenes are ... [12]

Deleted: Furthermore

Deleted: Additionally,

Moved (insertion) [2]

Deleted: the

Deleted: ancestors

Deleted: .show that processed pseudogenes are comm ... [13]

Deleted: While this can be observed through all of m ... [14]

Deleted: species

Deleted: comparable

Deleted: parallel

Deleted: the

Deleted:

Deleted:

Deleted: .

Formatted: Font:11 pt, Not Highlight

Formatted: Font:11 pt, Not Highlight

Moved (insertion) [3]

Deleted: e use of inbred mice also

Deleted:

Deleted: s

also resulted in substantial differences between the mouse strains, giving each strain the potential to offer a unique reaction to an acquired mutations \cite{19710643}.

To uncover key genome remodeling processes that governed mouse strain evolution, we focus our analysis on the study of pseudogene complements, while also highlighting their shared features with the human genome. In this paper 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 subsequent disablement of one of the duplicates, and unitary pseudogenes – formed by the inactivation of a functional gene. Additionally, pseudogenes that are present in a population as both functional and nonfunctional alleles are termed polymorphic pseudogenes \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 \cite{20210993}. From a functional perspective, pseudogenes can be classified into three categories: dead-on-arrival – elements that are nonfunctional and are expected, in time, to be eliminated from the genome, partially active – pseudogenes that exhibit residual biochemical activity, and exapted pseudogenes – elements that have acquired new functions and can interfere with the regulation and activity of protein coding genes.

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 for establishing new gene functions \cite{14671323}. While the majority of the duplicated gene copies are eventually pseudogenized \cite{27690225}, successfully retained paralogs can acquire new functions \cite{17053091}, a process known as neofunctionalization \cite{Ono1970}. Furthermore, duplicated pseudogenes can help explore the role of gene dosage in the inactivation or preservation of duplicate genes \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 ancestral gene \cite{JAX2}. Unitary pseudogenes are an extreme case of LOF, where mutations that result in complete inactivation of a 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 numerous examples showcasing evolutionary advantages for the accumulation and fixation of LOF mutations resulting in formation of new pseudogenes. 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 receptors \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 \cite{15677715}. 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}.

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

Deleted: mutation

Moved up [3]: The use of inbred mice also minimizes a number of problems raised by the genetic variation between animals \cite{11528054}.

Formatted: Font:11 pt, Not Highlight

Deleted: T

Deleted: here are polymorphic

Deleted: pseudogenes – elements

Deleted: the

Formatted: Font:11 pt, Not Highlight

Moved (insertion) [1]

Deleted: Finally, t

Deleted: .

Formatted: Font:11 pt

Formatted: Font:11 pt, Not Highlight

Deleted:

Deleted: commonly represented by transcribed pseudogenes,

Deleted: are

Moved up [1]: Finally, there are polymorphic pseudogenes – elements that are present in the population as both functional and nonfunctional alleles \cite{20210993}.

Deleted: 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.

Deleted: of

Deleted: dead-on-arrival

Deleted:

Deleted: the

Deleted: successful

Deleted: during a phase of rapid, almost neutral evolution

Deleted: Furthermore,

Deleted: 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

Formatted: Border:Top: (No border), Bottom: (No border), Left: (No border), Right: (No border), Between : (No border)

Deleted: wild type

Deleted: the

Deleted: also

Deleted: As such pseudogenes can reflect either advantageous or deleterious phenotypes.

Deleted: receptor

Deleted: .

Formatted: Font:11 pt, Font color: Black possible. Leveraging mouse developmental timecourse RNAseq data, we explore whether pseudogene creation occurs primarily 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 pipelines 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.

For human, we used 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 the majority of the annotations are processed pseudogenes, with a smaller fraction of duplicated pseudogenes (Table S2).

3-1

1.2 Mouse strains

The Mouse Genome Project has sequenced and assembled genomes for 12 laboratory and 4 wild mice, and developed a draft annotation of each organisms' protein coding genes \cite{MousePaper}. The Flicek lab has sequenced and assembled the genome of two distant Mus species: *Mus Caroli* and *Mus Pahari* \cite{Flicek2017}. Collectively the 18 strains provide a unique overview of mouse evolution. 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 12 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 the in house automatic pipeline PseudoPipe and the set of manually curated pseudogenes from the mouse reference genome lifted over onto each individual strain (Figure 1C). This combined pseudogene identification process gives rise to three confidence levels reflecting the annotation quality. Each identified pseudogene is associated with details about its transcript biotype, genomic location, structure, sequence disablements, and confidence level. A detailed overview of pseudogene annotation statistics including the number of pseudogenes, their confidence levels, and 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% are Level 2 annotations characterized only using the lift over process, and 60% are Level 3 annotations identified solely by the automatic annotation pipeline. The pseudogene biotype distribution across the strains closely follows the reference genome and is consistent with the biotype distributions observed in other

Deleted: In particular,

Deleted: are

Deleted: able to

Deleted: the creation of pseudogenes during early embryo development. We leverage timecourse RNAseq data to investigate

Deleted: izationpseudogenization

Deleted: [[CSDS to update table S1]]

Deleted: more than half

Deleted:

Formatted: Font:11 pt, Not Highlight

Deleted: 8

Deleted: mouse

Deleted: strains

Deleted: the strains'

Deleted: ,

Deleted: and annotated

Deleted: https://doi.org/10.1101/158659}.

Formatted: Font:11 pt, Not Highlight

Deleted: mouse

Deleted: eThe

Deleted: provided

Deleted: the

Deleted: a

Deleted: reflecting the annotation process

Deleted: related

Deleted: resulted

Deleted: from

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 pseudogenes. 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 evolutionary 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 selective pressures guiding genome evolution. Due to their formation 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) in human and mouse, we identified 218 and respectively 237 new unitary pseudogenes (Table S3). In human, a large proportion of unitary pseudogenes are related to the chemosensory system (e.g. GPCRs, olfactory and vomeronasal receptor proteins) which have functional homologs in mouse, reflecting the loss of function in these genes during the primate lineage evolution.

4-1 SAME AS ABS.

HOW MANY THERE ALREADY?

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}.[[CSDS to add a note about AF caveat that immune genes can be hard to distinguish and classify]] 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).

To get an overview of the unitary pseudogenes in each strain with respect to the reference genome, we used a similar workflow as above, except in this case, using the mouse reference strain specific peptides as input. The resulting pseudogene calls were intersected with the lift over of the reference strain specific protein coding genes in order to validate the conservation of location and loss of function of the latter. On average we obtained around 20 unitary pseudogenes in each strain with larger number observed for wild strains (Table S5). However, the fast rate of evolution among the mouse strains, as well as the highly specific generation of the laboratory strains, suggests that the number of unitary pseudogenes could be higher, reflecting the strain specific phenotypes. Thus, a 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 other primates \cite{20210993}, as previous studies suggest that protein gene loss rate is similar in both mouse and primate lineages \cite{23153069}. 4-2

WRT TO ALL OTHER STRAINS!

2. Conservation and divergence in pseudogene complements

In order to investigate the evolutionary history of pseudogenes in 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 were created during the shared evolutionary history.

Next, we took advantage of pseudogenes' ability to evolve with little or no selective constraints \cite{10833048}, and compared 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

Deleted: selective

Deleted: the

Deleted: mechanism

Deleted: ,

Deleted: in human and 237 unitary pseudogenes in mouse

Deleted: and

Formatted: Font:11 pt, Highlight

Deleted: Moreover, to

Deleted: laboratory strain

Deleted: , 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.

Deleted: short

Deleted: between

Deleted: evolutionary distance between most strains means this value is an overestimate of the number of unitary pseudogenes

Deleted: One

Formatted: Font:11 pt

Deleted: chimp

Formatted: Font:11 pt, Not Highlight

Formatted: Font:11 pt

Formatted: Font:11 pt

Formatted: Font:11 pt, Not Highlight

Formatted: Font:11 pt

Deleted: [[xxxx]]

Formatted: Font:11 pt, Highlight

Deleted: decipher

Deleted:

Deleted: the

Deleted: a

Deleted: derived

Deleted: the

(Figure 2C). This pseudogene-based tree follows closely the tree constructed from protein coding genes and correctly identifies and clusters the mice into three classes: outgroup, wild, and laboratory strains.

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 tree, we found that they display different patterns, reflecting different evolutionary histories. 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 two distinct diabetic phenotypes. Despite this difference the result suggests that in both cases changes in diet can affect chemoreceptors (key players in both phenotypes) 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 are able to study the 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 \cite{10810090}. Moreover, as pseudogenes are inherited, the genesis of new elements occurs 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 that observed for non-parent 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. 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 the 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, if we assume that each duplication process can potentially give rise to either a pseudogene or 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 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 (Figure SF4).

Deleted: protein coding genes

Deleted: strains

Deleted: one

Formatted: Font:11 pt

Deleted: a

Formatted: Font:11 pt, Not Highlight

Formatted: Font:11 pt

Formatted: Font:11 pt, Not Highlight

Deleted: T

Formatted: Font:11 pt, Not Highlight

Deleted: suggests

Formatted: Font:11 pt

Formatted: Font:11 pt

Formatted: Font:11 pt, Not Highlight

Formatted: Font:11 pt

Formatted: Font:11 pt, Not Highlight

Formatted: Font:11 pt

Formatted: Font:11 pt, Not Highlight

Deleted: s

Formatted: Font:11 pt

Deleted: [[xxxx]]

Deleted: }.

Deleted: .

Deleted: should

Deleted: occur

Deleted: the one

Deleted: for regular

Deleted: regular

Deleted: 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.

Deleted: large

Deleted: fof

Formatted: Font:11 pt

Formatted: Font:11 pt

3.2 Transposable elements

Since the majority of mouse and human pseudogenes are the result of retrotransposition processes mediated by transposable elements (TE), we investigated the genomic mobile element content in human and mouse on an evolutionary time scale (Figure 3B)

Deleted: To the extent that

Deleted: egenome

Deleted: the two species

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 analysed the LINE, SINE and ERV content flanking pseudogenes in human and mouse. 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.

6-1

WHY FLANKING?

Deleted: the

Deleted: processed pseudogene complements

FIG 6-2

FLANK IN FIG.

THOUGHT IT WAS JUST PROC PGENES IN FIG

In human, we observe a smooth 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 \cite{14611660,15261647}. By contrast in mouse we found the L1 derived pseudogene distribution is defined by two successive peaks at 92.5% and 97% sequence similarity to parent genes. Also, in contrast to human where the density of L1 associated pseudogenes shows a steep decrease amongst young pseudogenes following the peak at 92.5% similarity, the density of L1 flanked mouse pseudogenes remains at a high level in the interval between 97% to 100% sequence similarity to parents. This observation suggests the presence of active transposable elements in mouse, which results in a continuous renewal of the processed pseudogene pool. This is also reflected in the large difference in the number of active LINE/L1s between human and mouse, with just over 100 in human \cite{12682288} compared to 3,000 in mouse \cite{ 11591644}.

Deleted:

Deleted: .

Deleted: sparents

Deleted: b

Deleted: y by

Deleted: for

Deleted: s

Deleted: of

Deleted: highly

Deleted: . The TE activity

Deleted: behavior

Deleted: for

Deleted: vs

Deleted: for

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 agreement with the observed major karyotype-scale differences and large genomic rearrangements exhibited by Caroli and Pahari \cite{Kolmogorov2017, Flicek2017}. 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).

Deleted: accord

Deleted: doi.org/10.1101/088435

4. Functional analysis

Deleted: Biological relevance

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 the gene ontology, protein families and RNA-seq experiments.

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). We also evaluated GO term enrichment amongst parent genes of both processed and duplicated pseudogenes across the mouse strains.

1. 7-1 COULD SAY BETTER

Clustering of GO terms based on semantic similarity revealed enrichment for [redacted] terms related to ribosomal functions, cell cycle, translation and RNA processing, and ubiquitination for processed pseudogenes. Amongst duplicated pseudogenes we observed enrichment for apoptosis, sensory and smell processes, and immune functions (Figure 5B). Additionally, 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 Pfam family observed amongst the pseudogenes, 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 and duplicated proteins such as GAPDH and ribosomal proteins, and regulatory protein families such as the Zinc fingers (Figure 5C).

A closer look suggests that the pseudogene repertoire also reflects individual 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. The pseudogene - phenotype relationship can be viewed from different perspectives. First, pseudogenes reflect duplication events that are linked with the emergence of an advantageous phenotype. This is observed in the *Mus Spretus* genome, where we see an enrichment of duplicated tumor repressor and apoptosis pathways genes \cite{19129501} and correspondingly as increase in the number of associated pseudogenes. Second, we find pseudogenes reflecting the death of a gene family. As such we observe an increase in the number of pseudogenes associated deleterious phenotypes. 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}. However, a detailed analysis of the pseudogene repertoire suggests that there are more ways to describe the pseudogene – phenotype association, in particular looking at the apparition of advantageous phenotype through the pseudogenization process \cite{25887751}.

1. 7-2

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 genes \cite{26472758}, and thus might 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.

7-3 STILL ?

We also analysed the number of paralogs associated with our essential and nonessential gene sets 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. This is in agreement with previous work showing non-essential genes are more likely than essential to be duplicated successfully \cite{23675306}.

CONTRADICT? PREV. SENT. 7-4

4.3 Pseudogene Transcription

We leveraged RNA-seq data from the Mouse Genome Project and ENCODE 3 to study pseudogene biology as reflected by their transcriptional activity. This is thought to either relate to the exaptive functionality of pseudogenes or be a residual leftover from their existence as genes. In both the human and 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 restricted data availability in the mouse strains, we focused our transcriptional analysis to a single tissue – adult brain from wild and laboratory strains. Overall pseudogenes with strain specific transcription were more common than those with cross-strain transcription (Figure 6C,D). Moreover, the proportion of pseudogenes conserved across all strains that are transcribed is constant (~2.5%) across the wild and laboratory strains (Figure 6D). By contrast, the fraction of transcribed strain specific pseudogenes varies across the strains from 1.5% to 4% (Figure 6D).

NEW ¶

5. Mouse pseudogene resource

7-5

Deleted: Moreover…dditionallyMoreover… 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 Pfam family observed amongst the pseudogene …family…family ... [15]

Deleted: and duplicated

Deleted: ,…ribosomal proteins, ... [16]

Deleted: regulatory protein families such as the

Deleted: B…. 5B). ... [17]

Deleted: However, a

Formatted: Font:11 pt, Not Highlight

Deleted: pseudogene - phenotype relationship can be viewed from different perspectives. First, re pseudogenes reflect duplication events that are linked with the

Formatted: Font:11 pt, Not Highlight

Deleted: are two possible types of pseudogene-phenotype associations. First, the pseudogenization process is linked with the …mergence of an advantageous phenotype. This is observed in the, as in the case of…*Mus Spretus* genome, where we see an enrichment of duplicated pseudogenes related to …umor repressor genes ... [18]

Formatted ... [19]

Deleted: and correspondingly as increase in the number of associated pseudogenes. …nd correspondingly as increase in the number of associated pseudogenes. Second, w ... [20]

Deleted: as expected, we find that the majority of ... [21]

Formatted: Font:11 pt

Deleted: However, a detailed analysis of the pseudog ... [22]

Deleted: ,…and thus can potentially ... [23]

Deleted: sgenes…to get an insight into the possible r ... [24]

Deleted: the available …NA-seq data from the Mous ... [25]

Formatted: Font:11 pt

Deleted: For…both the human and the ... [26]

Deleted: we detected that about…e detected that abo ... [27]

Deleted: with

Deleted: for the 18 mouse strains…n the mouse strai ... [28]

Formatted ... [29]

Deleted:

Formatted: Font:11 pt, Not Highlight

Deleted: T

Formatted ... [30]

Deleted: show a unifor

Formatted ... [31]

Deleted: m level of transcription.

Formatted ... [32]

Deleted: However, the proportion of transcribed ... [33]

Formatted ... [34]

Deleted: Supplementary

Formatted: Font:11 pt, Not Highlight

Deleted: SF4

Formatted: Font:11 pt

We created a pseudogene resource that organizes all of the pseudogenes across the available mouse strains and reference genome, as well as associated phenotypic information, in a MySQL database (Figure 6). All the available data are also provided as flat files for ease of manipulation. Queries on specific pseudogenes will return the relevant pseudogene annotation flat file containing all pertinent associated information. The database contains three general types of information: details about the annotations, 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. In order to assist direct comparison between human and mouse we also provide orthology links between mouse entries and the corresponding human counterparts.

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 onto 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.

SPEC ?

8-1

Discussion

We report the updated and refined pseudogene annotation in the mouse and human reference genomes, and describe the curation and comparative analysis of the first draft of pseudogene complements in 18 related mouse strains. By combining manual curation and computational annotation, 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 method.

Comparable to our previous observations in human, worm, and fly, the pseudogene complement in mouse strains, reflects an organism specific evolution, highlighting pseudogenes as ideal markers of genome remodelling processes. However, despite the strain dependent evolution, the pseudogenes share a number of similarities, in particular regarding their biogenesis and diversity. As such we noticed a uniform ratio of processed to duplicated pseudogene of 4 to 1 in all of the strains result consistent with previous observations in human. The higher proportion of processed pseudogenes accounting for 80% of the total, is in agreement with earlier findings that suggest retrotransposition as the primary mechanism for pseudogene creation in numerous mammalian species \cite{22951037}. Moreover, examining the retrotransposon activity, and in particular the L1 content, we observed that while the majority of human pseudogenes have been formed relatively recently through a single burst of retrotransposition \cite{22951037}, the mouse lineage shows a sustained renewal of the pseudogene pool through successive retrotransposition bursts. The sequence context of the processed pseudogenes indicates that the various retrotransposons exhibit differential contributions to the pseudogene set over time.

Since a pseudogene's likelihood of creation is related to its parent's functional role and 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 events that shaped both the

Deleted: is

Deleted: annotation

Deleted: facilitate

Deleted: a

Deleted: each

Deleted: yentry

Deleted: counterpart

Deleted: 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.

Deleted: to

Deleted: describe the annotation and comparative analysis of the

Deleted: complement

Deleted: genome and the

Deleted: the

Deleted: an automatic

Deleted: pipeline

Deleted: methods

Deleted: A high-level comparison

Deleted: pseudogene statistics for each

Deleted: the strains highlights shared properties of

Deleted: biogenesis. Each

Deleted: exhibit a

Deleted: ratio of processed to duplicated pseudogenes, which is in line

Deleted: is

Deleted: ... [35]

Deleted: 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' 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{Flicek2017,Kolmogorov2017} ... [36]

Deleted: how it contributed to pseudogene creation, and how it shaped the genomic environment of each strain over time. Sequence analysis reveals

Deleted: obtained

Deleted: continuous transposable element activity. Looking closely at

Moved (insertion) [4]

Formatted: Font:11 pt, Not Highlight

Moved (insertion) [5]

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 we observed that 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.

To better understand the evolutionary and functional relationship between the pseudogenes in the 18 strains we constructed a pan-genome pseudogene set as the union of all individual strain complements and resulting in over 45,000 unique entries. The pan-genome pseudogene repertoire distinguishes three types of pseudogenes: universally conserved (present in all of the 18 studied strains), multi-strain (present in at least 2 of the strains), and strain specific (unique to a specific strain), accounting for 6, 23, and 71% of the elements respectively. Despite the large number of strain specific pseudogenes in the pangenome set, these account for only 25% of the total pseudogenes in any particular strain, a comparable proportion to the universally conserved pseudogenes present in each strain. Moreover, the pseudogene cross strain relationship identified from the pangenome allows us to have a closer look at their evolution by studying the conservation of their chromosomal location. In particular, we observed a stark contrast between the high level of genomic loci retention shared by the laboratory strains and the lack of conservation noticed when looking at the outgroup species. These results hint at multiple large scale genomic rearrangements in the mouse lineage. This is especially noticeable in the case of *Mus Pahari* as has been recently reported by large scale chromosomal imagining and karyotype analysis \cite{Flicek2017, Kolmogorov2017}.

9-1

IN RESULTS ?

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 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, 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 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).

9-2

REWORD

Finally, 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 matches those seen in the human set. 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 better understand the selective pressures that have shaped an organism's genomic content and phenotype.

GAPDH

GAPDH

9-3

Moved (insertion) [6]

Moved up [4]: 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.

Deleted: Meanwhile, duplicated pseudogenes record duplication

Moved up [5]: 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.

Deleted: Since a processed pseudogene's likelihood of creation is related to its parent's

Deleted: As expected

Moved up [6]: 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.

Deleted:

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.

In summary, 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.

Tables

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

Organism	**Manual**	**PseudoPipe***	**RetroFinder**	**Union PseudoPipe**	**Manual Overlap**
				Retrofinder	**PseudoPipe (%)**
Mouse	10,524	18,649	18,467	26,093	8,786 (83.5)
Human	14,650	15,978	15,474	22,396	13,177 (89.9)

*Chromosomal assembled DNA only

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: Font:11 pt, Bold

Deleted: 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. ... [37]

Deleted: Tables
Manual Overlap (%) ... [38]

Formatted: Font:11 pt

Formatted: Font:11 pt

Formatted Table

Formatted: Font:11 pt

Formatted Table

Deleted: ... [39]

Formatted Table

Formatted: Font:11 pt

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

	PseudoPipe (PP)			RetroFinder (RF)	PP-RF overlap
	Autosomes	Sex Chr.	Others*		
Mouse	14,084	4,565	4,162	18,467	10,522
Human	14,644	1,325	2,098	15,474	9,057

*Includes patches, scaffolds, and unassembled DNA.

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

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)

Deleted: Organism

Deleted: Manual

Deleted Cells

Inserted Cells

Inserted Cells

Deleted: Chromosomes

Deleted: 10,524

Deleted: 14,650

Inserted Cells

Inserted Cells

Figure 1. A – Human vs mouse lineage comparison. Abbreviations *MYA* – million years ago, λ – laboratory strain. B – Summary of mouse strains pseudogene annotation. *Level 1* are pseudogenes identified by automatic pipelines and lift over of manual annotation from the reference genome; *Level 2* are pseudogenes identified only through the lift over of manually annotated cases from the reference genome; *Level 3* are pseudogenes identified only by the automatic annotation pipeline. C (top) – pseudogene annotation workflow for mouse strains. C (bottom) – unitary pseudogene annotation pipeline.

Figure 2. A – Summary of pseudogene distribution in the pangenome mouse strain dataset. B – Venn diagram of evolutionarily conserved and group specific pseudogenes. The number in bracket is indicative pseudogenes that are unique to each group. C – phylogenetic trees for parents of evolutionarily conserved pseudogenes, evolutionary conserved pseudogenes, and pfam families of evolutionarily conserved pseudogenes: 7 trans membrane proteins (7TM), ribosomal proteins (Ribosome), cyclin-dependent kinases (CDK), olfactory receptor proteins (Olfr)

Deleted: a

Figure 3. A – Relationship between the number of pseudogenes and functional paralogs for a given parent gene (left – duplicated pseudogenes, right – processed pseudogenes). Fitting lines show a vague correlation between the number of functional vs disabled copies of a gene, with a linear fit for duplicated pseudogenes and a negative logarithmic fit for processed pseudogene. The gray area is the standard deviation. B – Distribution of L1 flanked pseudogenes (y-axis) as function of age (x-axis). The pseudogene age is approximated as DNA sequence similarity to the parent gene.

Figure 4. A – CIRCOS-like plots showing the conservation of the pseudogene genomic loci between each mouse strain and the laboratory reference strain C57BL/6NJ. Gray-lines indicate a change of the genomic locus between the two strains; black-lines indicate the conservation of the pseudogene locus. B – the numbers of pseudogenes that preserved or changed their loci between each strain and the laboratory reference strain. C – Strain speciation times as function of percentage of conserved pseudogene loci between each strain and the laboratory reference, fitted by an invers logarithmic curve.

Figure 5. A – Distribution of top 300 GO biological processes terms across the mouse strains. B – Heatmap illustrating enrichment of GO biological processes terms across the mouse strains for the parent genes of processed and duplicated pseudogenes. GO terms (rows) are clustered by semantic similarity. C – Summary of the top 24 Pfam pseudogene families in each mouse strain.

Deleted: ….

Figure 6. A – Cross tissue pseudogene transcription in the mouse reference genome. x-axis indicates the number of tissue a pseudogene in transcribed in. B – Distribution of pseudogene transcription in 18 adult mouse tissues. C – Number of transcribed pseudogenes in brain tissue for each wild and laboratory mouse strain. D (top) – number of transcribed pseudogenes that are conserved across all the strains. D (bottom) –number of transcribed strain specific pseudogenes in each mouse strain.

Figure 7. A – Cyp2G1 LOF in human. B – NCR3 gain of function mutation in Mus Caroli as compared to the reference genome and the other mouse strains.

Methods

Datasets

Mouse reference genome is based on the Mus Musculus strain C57BL/6J strain. The mouse reference annotation is based on GENCODE vM12/Ensembl 87.

The human reference genome annotation is based on GENCODE v25/Ensembl 87.

The 16 laboratory and wild strains (Table 2) assemblies and strain specific annotations were obtained from the Mouse Genome Project \cite{MousePaper}. 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 (Table 2), Mus Caroli and Mus Pahari were sequenced, assembled, and annotated in the protein coding domain by the Flicek lab \cite{Flicek2017}.

Deleted: ,

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}. The data for two outgroup species divergence times was obtained from \cite{Flicek2017}. The generation time for all the mice was estimated from \cite{JAX}.

Deleted: While

Deleted: speciation

Pseudogene Annotation

Reference genome annotation

We manually curated 10,524 pseudogenes in the mouse reference genome (GENCODE M12) and 14,650 pseudogenes in the human reference genome (GENCODE v25), using a workflow previously described in \cite{22951037,25157146}. The manual annotation is based on the on sequence homology to protein data from UniProt database \cite{25157146} and the protocol is summarised in Figure S1.

Deleted: almost

Deleted: 000

Deleted: as

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. Thus, to get a more accurate idea of the number of pseudogenes in the mouse genome, we used a combination of two automatic annotation pipelines: PseudoPipe \cite{16574694} and RetroFinder \cite{18842134}. PseudoPipe is a comprehensive annotation pipeline focused on identifying and characterizing pseudogenes based on their biotypes as either processed or duplicated. The automatic annotation workflow using PseudoPipe is summarised in Figure 1 and has been previously described in detail in \cite{16574694,22951037,25157146}. Pseudopipe identifies 22,811 mouse pseudogenes of which 14,084 are present in autosomal chromosomes (a number comparable with the one observed previously in human (Table S1)). RetroFinder is computational annotation pipeline focused on identifying retrotransposed genes and pseudogenes. Using RetroFinder we were able to annotate 18,467 and respectively 15,474 processed pseudogenes in mouse and human. There is a good overlap between the two identification pipeline with respect to the number of processed pseudogenes present in both organisms (Table S1).

Deleted: , and the fact that in human we have manually identified over 14,000 pseudogenes.

Deleted: use the in house

Deleted: pipeline

Deleted: computational pipeline

Deleted: approximately

Deleted: 000

Deleted: about

Deleted: 000

Mouse strain annotation

The mouse strain pseudogene annotation workflow is summarised in Figure 1C. The protein coding input set contains 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.

Deleted: We used as input

Deleted:

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 such as 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, and removal of pseudogenes shorter than 100 bp. The filtered PseudoPipe set was intersected with the lift-over of the protein coding annotation from the “functional organism using BEDTools \cite{ 25199790}with a 1bp overlap minimum required. The intersection set was further refined flagging protein coding genes that have functional relatives (paralogs) in the non-“functional organism”. The remaining matches were subjected to manual inspection of the alignment.

Deleted: The resulting data set was subjected to a number of filters:

Conservation and divergence in pseudogene complements

Deleted: <#>removal of previously known pseudo ... [40]

Deleted: <#>**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 Ensembl ID, where available (used for Level 1 and 2 pseudogenes), conservation of parent gene identity, conservation of pseudogene locus, conservation of pseudogene biotype, conservation of pseudogene length, and conservation of pseudogene structure.

Deleted: The resulting set is further filtered for:

Next we integrated all filtered binary mappings in a master pan-strain set. The common entries were collapsed into a unique pangeome pseudogene reference. We obtained 49,262 pangenome pseudogenes. 1,158 pangenome entries are multi matching across strains.

Deleted: <#>conservation of pseudogene identity, ... [41]

Formatted: Line spacing: multiple 1.15 li

Deleted: <#>were integrated

Deleted: <#>merged

Deleted: <#>A number of

Phylogenetic analysis

Sequences of the 1,460 randomly selected conserved pseudogenes in the 18 mouse strains accounting for approximately 50% of the total number of conserved pseudogenes, 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 sequences of 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

Deleted: Datasets ... [42]

Genome mappability maps

We created mappabilty maps for the mouse reference genome and the 18 mouse strains using the GEM library \cite{GEM}. The workflow is composed of indexing the genome using gem-indexer, followed by creation of the map using a window of 75 nucleotides under the following conditions -m 0.02 -T 2.

Formatted: Normal, Space Before: 0 pt, After: 0 pt

Formatted: Font:Not Bold, Font color: Auto

Formatted: Normal, Left, Space Before: 0 pt, After: 0 pt

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 with a quality score higher than 30, and calculating the expression FPKM levels using Cufflinks \cite{ 22383036}. Transcriptional activity of pseudogene parent genes during early embryonic development was investigated using RNAseq data as processed and described in \cite{27309802}.

Deleted: ,

Deleted: ... [43]

Transposable elements analysis

TE in human and mouse reference genomes were informed from RepeatMasker libraries Repbase 21.11 and using RepeatMasker 3.2.8 (http://repeatmasker.org). We extracted all the four major groups of repeats SINE, LINE, LTR and DNA and identified all the processed pseudogenes associated with L1 elements. Next we binned the L1 annotated pseudogenes into age groups based on their sequence similarity to the parent gene, with younger elements exhibiting a higher sequence similarity while older elements show a large sequence divergence when compared to the functional gene counterparts.

Deleted: identified

Deleted: .

Deleted: Transposable elements ... [44]

Deleted: http://repeatmasker.org).

Deleted:

Formatted: Left, Line spacing: multiple 1.15 li

Gene ontology and Pfam analysis

Linking of gene ontology terms to the pseudogene parent genes was conducted using the R package biomaRt \cite{16082012, 19617889}. Visualization of shared and distinct GO term sets amongst the strains was done using the R package UpSetR \cite{26356912}. Enrichment of GO terms amongst the pseudogene parent genes and clustering of mouse strains based on similar enrichment profiles was performed using the goSTAG software package \cite{28413437}. Semantic clustering of the GO terms was done with the OntologyX packages \cite{28062448}. Parent genes were labelled with both strain and biotype information in order to better evaluate differences in the pseudogene complements based on their mechanism of creation.

Analysis of the Pfam representation in the pseudogene complements was performed as described in \cite{ 18957444}.

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 ENCODE3 data. The pseudogene sequences were filtered for uniquely mapable exon regions longer than 100 bp. Next the RNAseq raw data was mapped using TopHat and the mapped reads were filtered for quality scores higher than 30. The resulting alignments were quantified using Cufflinks. A pseudogene was considered transcribed if it had an FPKM larger than 3.3 in accord with previous studies \cite{25157146}.

Deleted: ENCODE data.

RNAseq data from mouse adult brain was obtained from the Mouse Genome project for 12 laboratory and 4 wild strains. Next we created mappability maps for each of the 16 mouse strains genomes and selected only the pseudogene exons in uniquely mappable regions and longer than 100bp for further transcription analysis. The pseudogene transcription levels in mouse strains were estimated using a similar workflow as described above. The transcription cut off level was set to 1.

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

Deleted: 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. ... [45]

Page 1: [1] Style Definition **Muir, Paul** **21/07/2017 11:41:00**

Comment Text: Font:(Default) Arial, Font color: Black, English (UK), Border:Top: (No border), Bottom: (No border), Left: (No border), Right: (No border), Between : (No border)

Page 1: [2] Style Definition **Muir, Paul** **21/07/2017 11:41:00**

Balloon Text: Font color: Black, English (UK), Border:Top: (No border), Bottom: (No border), Left: (No border), Right: (No border), Between : (No border)

Page 1: [3] Style Definition **Muir, Paul** **21/07/2017 11:41:00**

List Paragraph: Font:(Default) Arial, 11 pt, Font color: Black, English (UK), Line spacing: multiple 1.15 li, Border:Top: (No border), Bottom: (No border), Left: (No border), Right: (No border), Between : (No border)

Page 1: [4] Style Definition **Muir, Paul** **21/07/2017 11:41:00**

Subtitle: Font:(Default) Arial, English (UK), Line spacing: multiple 1.15 li, Border:Top: (No border), Bottom: (No border), Left: (No border), Right: (No border), Between : (No border)

Page 1: [5] Style Definition **Muir, Paul** **21/07/2017 11:41:00**

Title: Font:(Default) Arial, Font color: Black, English (UK), Line spacing: multiple 1.15 li, Border:Top: (No border), Bottom: (No border), Left: (No border), Right: (No border), Between : (No border)

Page 1: [6] Style Definition **Muir, Paul** **21/07/2017 11:41:00**

Heading 6: Font:(Default) Arial, 11 pt, English (UK), Line spacing: multiple 1.15 li, Border:Top: (No border), Bottom: (No border), Left: (No border), Right: (No border), Between : (No border)

Page 1: [7] Style Definition **Muir, Paul** **21/07/2017 11:41:00**

Heading 5: Font:(Default) Arial, 11 pt, English (UK), Line spacing: multiple 1.15 li, Border:Top: (No border), Bottom: (No border), Left: (No border), Right: (No border), Between : (No border)

Page 1: [8] Style Definition **Muir, Paul** **21/07/2017 11:41:00**

Heading 4: Font:(Default) Arial, English (UK), Line spacing: multiple 1.15 li, Border:Top: (No border), Bottom: (No border), Left: (No border), Right: (No border), Between : (No border)

Page 1: [9] Style Definition **Muir, Paul** **21/07/2017 11:41:00**

Heading 3: Font:(Default) Arial, English (UK), Line spacing: multiple 1.15 li, Border:Top: (No border), Bottom: (No border), Left: (No border), Right: (No border), Between : (No border)

Page 1: [10] Deleted **Microsoft Office User** **24/07/2017 12:17:00**

Investigating we find ~15% of the transcriptional activity of pseudogenes and their parent genes w

Page 1: [11] Deleted **Muir, Paul** **21/07/2017 11:41:00**

are transcribed, a fraction similar to human. Furthermore, w

Page 1: [12] Moved to page 1 (Move #2) **Muir, Paul** **19/07/2017 13:13:00**

find ~15% of the pseudogenes are transcribed, a fraction similar to human.

Page 1: [13] Deleted **Muir, Paul** **19/07/2017 13:16:00**

.show that processed pseudogenes are commonly associated with highly transcribed genes.

Page 1: [14] Deleted **Muir, Paul** **19/07/2017 11:15:00**

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.[MP1]

Page 7: [15] Deleted Muir, Paul 17/07/2017 18:56:00

Moreover

Page 7: [15] Deleted Muir, Paul 17/07/2017 18:56:00

Moreover

Page 7: [15] Deleted Muir, Paul 17/07/2017 18:56:00

Moreover

Page 7: [15] Deleted Muir, Paul 17/07/2017 18:56:00

Moreover

Page 7: [15] Deleted Muir, Paul 17/07/2017 18:56:00

Moreover

Page 7: [16] Deleted Microsoft Office User 17/07/2017 19:37:00

,

Page 7: [16] Deleted Microsoft Office User 17/07/2017 19:37:00

,

Page 7: [17] Deleted Muir, Paul 21/07/2017 10:59:00

B

Page 7: [17] Deleted Muir, Paul 21/07/2017 10:59:00

B

Page 7: [18] Deleted Microsoft Office User 17/07/2017 19:48:00

are two possible types of pseudogene-phenotype associations. First, the pseudogenization process is linked with the

Page 7: [18] Deleted Microsoft Office User 17/07/2017 19:48:00

are two possible types of pseudogene-phenotype associations. First, the pseudogenization process is linked with the

Page 7: [18] Deleted Microsoft Office User 17/07/2017 19:48:00

are two possible types of pseudogene-phenotype associations. First, the pseudogenization process is linked with the

Page 7: [18] Deleted Microsoft Office User 17/07/2017 19:48:00

are two possible types of pseudogene-phenotype associations. First, the pseudogenization process is linked with the

Page 7: [19] Formatted Cristina Sisu 24/07/2017 23:21:00

Font:11 pt

Page 7: [19] Formatted Cristina Sisu 24/07/2017 23:21:00

Font:11 pt

Page 7: [20] Deleted Cristina Sisu 24/07/2017 12:37:00

and correspondingly as increase in the number of associated pseudogenes.

Page 7: [20] Deleted Cristina Sisu 24/07/2017 12:37:00

and correspondingly as increase in the number of associated pseudogenes.

Page 7: [21] Deleted Microsoft Office User 17/07/2017 19:55:00

as expected, we find that the majority of pseudogenes reflect a

Page 7: [22] Deleted Cristina Sisu 24/07/2017 12:36:00

However, a detailed analysis of the pseudogene repertoire suggests that there are more ways to describe the pseudogene – phenotype association, in particular looking at the apparition of advantageous phenotype through the pseudogenization process.

Page 7: [23] Deleted Muir, Paul 21/07/2017 11:41:00

,

Page 7: [23] Deleted Muir, Paul 21/07/2017 11:41:00

,

Page 7: [24] Deleted Muir, Paul 20/07/2017 17:04:00

sgenes

Page 7: [24] Deleted Muir, Paul 20/07/2017 17:04:00

sgenes

Page 7: [25] Deleted Muir, Paul 19/07/2017 09:50:00

the available

Page 7: [25] Deleted Muir, Paul 19/07/2017 09:50:00

the available

Page 7: [26] Deleted Muir, Paul 19/07/2017 09:55:00

For

Page 7: [26] Deleted Muir, Paul 19/07/2017 09:55:00

For

Page 7: [27] Deleted Cristina Sisu 24/07/2017 23:27:00

we detected that about

Page 7: [27] Deleted Cristina Sisu 24/07/2017 23:27:00

we detected that about

Page 7: [28] Deleted Cristina Sisu 24/07/2017 23:29:00

for the 18 mouse strains

Page 7: [28] Deleted Cristina Sisu 24/07/2017 23:29:00

for the 18 mouse strains

Page 7: [28] Deleted Cristina Sisu 24/07/2017 23:29:00

for the 18 mouse strains

Page 7: [28] Deleted Cristina Sisu 24/07/2017 23:29:00

for the 18 mouse strains

Page 7: [29] Formatted Cristina Sisu 24/07/2017 23:49:00

Font:11 pt

Page 7: [29] Formatted Cristina Sisu 24/07/2017 23:49:00

Font:11 pt

Page 7: [29] Formatted Cristina Sisu 24/07/2017 23:49:00

Font:11 pt

Page 7: [29] Formatted Cristina Sisu 24/07/2017 23:49:00

Font:11 pt

Page 7: [29] Formatted Cristina Sisu 24/07/2017 23:49:00

Font:11 pt

Page 7: [29] Formatted Cristina Sisu 24/07/2017 23:49:00

Font:11 pt

Page 7: [29] Formatted Cristina Sisu 24/07/2017 23:49:00

Font:11 pt

Page 7: [30] Formatted Cristina Sisu 24/07/2017 23:49:00

Font:11 pt

Page 7: [30] Formatted Cristina Sisu 24/07/2017 23:49:00

Font:11 pt

Page 7: [30] Formatted Cristina Sisu 24/07/2017 23:49:00

Font:11 pt

Page 7: [30] Formatted Cristina Sisu 24/07/2017 23:49:00

Font:11 pt

Page 7: [31] Formatted Cristina Sisu 24/07/2017 23:49:00

Font:11 pt

Page 7: [31] Formatted Cristina Sisu 24/07/2017 23:49:00

Font:11 pt

Page 7: [32] Formatted Cristina Sisu 24/07/2017 23:49:00

Font:11 pt

Page 7: [32] Formatted Cristina Sisu 24/07/2017 23:49:00

Font:11 pt

Page 7: [33] Deleted Cristina Sisu 24/07/2017 23:43:00

However, the proportion of transcribed pseudogenes in brain (2.5%) is half (2.5%) of that observed across the entire dataset. Moreover, for strain specific pseudogenes, the fraction of transcribed elements varies

Page 7: [34] Formatted Cristina Sisu 24/07/2017 23:49:00

Font:11 pt

Page 7: [34] Formatted Cristina Sisu 24/07/2017 23:49:00

Font:11 pt

Page 7: [34] Formatted Cristina Sisu 24/07/2017 23:49:00

Font:11 pt

Page 8: [35] Deleted CSDS 21/07/2017 14:42:00

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

Page 8: [36] Deleted **CSDS** **21/07/2017 14:42:00**

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' 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{Flicek2017,Kolmogorov2017}

Examination of the pseudogene complement reveals

Page 10: [37] Deleted **Microsoft Office User** **21/07/2017 15:17:00**

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 exhibit 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' 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 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, 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 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 matches those seen in the human set. 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 better understand the selective pressures that have shaped an organism's genomic content and phenotype.

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.

Page 10: [38] Deleted Muir, Paul 21/07/2017 11:43:00

Tables

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

Pseudopipe*	Manual Overlap (%)
18,649	8,786 (83.5)
15,978	13,177 (89.9)

*Chromosomal assembled DNA only

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	

Column Break

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.

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

Page 10: [39] Deleted **Microsoft Office User** **21/07/2017 15:18:00**

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)

Page 14: [40] Deleted **CSDS** **21/07/2017 14:42:00**

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.

Page 14: [41] Deleted **CSDS** **21/07/2017 14:42:00**

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

Page 14: [42] Deleted **Microsoft Office User** **21/07/2017 15:21:00**

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.

Page 14: [43] Deleted	Muir, Paul	19/07/2017 11:42:00

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

Page 15: [44] Deleted	Microsoft Office User	21/07/2017 15:25:00

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

Page 15: [45] Deleted	Microsoft Office User	21/07/2017 15:25:00

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.