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

Cristina Sisu*, Paul Muir*, Adam Frankish, Ian Fiddes, Mark Diekhans, [[Outgroup Strains People ToBeAdded]], Thomas Keane, Mark Gerstein

Pseudogenes are ideal markers of genome remodelling. In turn, the mouse is an ideal platform for studying them, particularly with the availability of the transcriptional time course during development (just completed in phase 3 of ENCODE) and the sequencing of 17 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 327 new unitary pseudogenes in human with respect to mouse and 210 unitary pseudogenes in mouse with respect to human. (We make our annotation available through a resource website mouse.pseudogene.org.) The overall mouse pseudogene repertoire (in the reference and strains) is similar to human in terms of overall size, biotype distribution (80% processed, 20% duplicated) and top family composition (with many GAPDH and ribosomal pseudogenes). However, notable differences arise in the age distribution of pseudogenes with multiple retro-transpositional bursts in mouse evolutionary history and only a single one in human. Furthermore, in each strain ~20% of the pseudogenes are unique, reflecting strain-specific functions and evolution – e.g. the pseudogenization of taste receptors is clearly linked to a change in the diet of the NZO strain. Finally, we find ~15% of the pseudogenes are transcribed, a fraction similar to human. Furthermore, we show that processed pseudogenes are commonly associated with highly transcribed genes. While this can be observed through all of mouse development, the relationship is strongest not at the early embryo stages but later on, after depletion of maternal RNA.

Introduction

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

Deleted: toward

Mice have been frequently used as a model organism for the study of human diseases since the two species share a large number of 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.

Deleted: into the well characterized and experimentally tractable mouse model organism

Since their divergence about 65 to 110 million years ago (MYA) \cite{12651866,12466850,11214318,11214319}, the human and mouse lineages followed a comparable evolutionary pattern \cite{17284675}. While it is hard to make a direct comparison between the two species, there is a large range of divergence in the mouse population, with

Deleted: makeup of the present human

Deleted: parallels, at generation

some even approaching the human chimp divergence levels \cite{17284675} (Figure 1). The mouse strains under investigation show differences in their genetic makeup that manifest in an array of phenotypes, ranging from coat/eye color to predisposition for various diseases \cite{21921910}. Moreover, the creation of these strains has been extensively documented. Following a well characterized inbreeding process for 20 sequential generations, the inbred mice are homozygous at all loci and show a high level of consistency at genomic and phenotypic levels \cite{JAX}. The repeated inbreeding resulted in substantial differences between the mouse strains, giving each strain the potential to offer a unique reaction to an acquired mutation \cite{19710643}. The use of inbred mice also minimizes a number of problems raised by the genetic variation between animals \cite{11528054}.

Deleted: , the evolution of recent *Mus Musculus* strains including inbred laboratory mouse strains

Deleted: possess

Deleted: Understanding the genesis and functional impact of the genetic variation of these mouse strains would aid in deciphering genome evolution and diversity in human population

To uncover the key genome remodeling processes that governed mouse strain evolution, we focus our analysis on the study of pseudogene complements, while also highlighting their key 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.

Deleted: 17

Often regarded as genomic relics, pseudogenes provide an excellent perspective on genome evolution and function \cite{10692568,11160906,12034841,14616058}. Pseudogenes are DNA sequences that contain disabling mutations rendering them unable to produce a fully functional protein. There are different classes of pseudogenes based on their creation mechanism: processed pseudogenes – formed through a retrotransposition process, duplicated pseudogenes – formed through a gene duplication event, and unitary pseudogenes – formed by the inactivation of a functional gene. From a functional perspective, pseudogenes can also 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 – commonly represented by transcribed pseudogenes, are elements that have acquired new functions and can interfere with the regulation and activity of protein coding genes. Finally, there are polymorphic pseudogenes – elements that contain simultaneously both functional and non-functional alleles in the population \cite{20210993}. MORE

Deleted: during

Deleted: additional

Deleted: The composition of these different pseudogene classes across the mouse strains provides insight into changes in selective pressures and genome remodeling forces.

Moreover, pseudogenes reflect the changes in selective pressures and genome remodeling forces. Specifically, the duplicated dead-on-arrival pseudogenes, tell a story of gene duplication, one of the key mechanisms of establishing new gene functions \cite{14671323}, and are indicative of neo-functionalization processes in ~~the~~ organism evolution \cite{27690225,Oto1970}. Furthermore, duplicated pseudogenes resulting from functional duplicated paralogs by accumulation of disabling mutations at a later stage reveal information about genes whose products are sensitive to dosage effects \cite{24907529,27690225} leading to selection for disabling mutations in duplicated copies. The 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 the functional analysis as they can be regarded as markers for loss and gain of function events.

?

Deleted: Moreover, pseudogenes can play an important role in functional analysis as they can be regarded as markers for loss and gain of function events.

A loss-of-function (LOF) event is a mutation that results in a modified gene product that lacks the molecular function of the wild type gene \cite{JAX2}. Pseudogenes are an extreme case

GNO -

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

DJ?

Functional analysis of the different types of pseudogenes is especially interesting, because it has the potential to tell us about key biological processes associated with highly-transcribed genes (in the case of processed pseudogenes), and past loss of function variants that have become fixed in the population (in the case of unitary pseudogenes). Both of these cases provide insight into selective pressures and gene death – essential features in understanding genome function and evolution.

Taken together the well-defined evolutionary relationships between the mouse strains and the wealth of associated functional data from the recently completed ENCODE 3 project present an opportunity to investigate the processes underlying pseudogene biogenesis and activity to an extent previously not possible. In particular, we are able to explore the pseudogene complement during early embryo development, on a scale currently unavailable in human. We leverage time-course RNAseq data to question whether pseudogenization occurs in the gametes or earlier in development in a germline precursor. Also, comparison to the primate lineage and human population is a possibility as the evolutionary distance between some of the mouse strains parallels the human-chimp divergence as well as distances between the modern day human populations, 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 in 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

Deleted: Thus there is a fine line between a loss-of-function variant that is increasing in a population and a special type of pseudogene that is only partially fixed in that population, also known as polymorphic pseudogene.

Deleted: . However, sometimes, in the right conditions

Deleted: inactivation of a protein via

Deleted: its gene, can also be advantageous. The relaxation of the selection constraints on such a gene would favor the accumulation of disabling mutations, eventually resulting in fixation of that pseudogene in the organism. ... [1]

Deleted: gene. When expressed, the PCSK9 protein binds to the

Deleted: receptor leading to its

Deleted: and a reduced

Deleted: uptake of plasma

Deleted: Enrichment of

Deleted: is often associated with

Deleted: . By contrast, the accumulation of loss-of-function mutations and subsequent pseudogenization of PCSK9 result in lower plasma LDL levels, and reduced risk of heart diseases.

Deleted: a

Deleted: .

Deleted:

Deleted: an exciting

Deleted:

Deleted: 17

Deleted: process as previously described in the GENCODE annotation resource

Deleted: },

Deleted: identified almost 10,000

Deleted: in

Deleted: . The

set provides a high quality lower bound with respect to the true number of pseudogenes in the mouse genome, while the automatic annotation informs on the upper limit of the pseudogene complement size. In agreement with our previous work \cite{22951037,25157146} there is a considerable overlap, of over 83%, between the manual and automatic annotation sets.

Similarly, for

human we used a combination of automatic and manual curation 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 (Supplementary Table S2). In both the human and mouse reference genomes more than half of the annotations are processed pseudogenes, with a smaller fraction of duplicated pseudogenes (Figure 1).

1.2 Mouse strains

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

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

Currently, around 30% of pseudogenes in each strain are defined as high confidence Level 1 annotations, being identified through both automatic curation and manual lift over, 10% Level 2 characterized only using the lift over process, and 60% Level 3 resulted solely from the automatic annotation pipeline. The pseudogene biotype distribution across the strains closely follows the reference genome and is consistent with the biotype distributions observed in other mammalian genomes (e.g. Human \cite{22951037} and macaque \cite{25157146}). As such, the bulk (~80%) of the annotations are processed pseudogenes, while a smaller fraction (~15%) are duplicated 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 selective pressures resulting in the fixation of an inactive element in a species.

Deleted: manually annotated

Deleted: lineage is likely an underestimate of the true size of the mouse

Deleted: given the similarities

Deleted: human and mouse genomes,

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

Moved down [1]: Thus, to get a more accurate idea of the number of pseudogenes in the mouse genome, we used 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.

Moved down [2]: The computational pipeline identifies approximately 22,000 pseudogenes of which 14,000 are present in autosomal chromosomes.

Deleted: These numbers are comparable to those seen in human (Table XXX). This automatic annotation provides an upper bound on the number of pseudogenes present.
Organism ... [2]

Moved down [3]: . Reference genome pseudogene annotation in mouse and human.

Deleted: **Organism** ... [3]

Deleted: see SupTable XXX).

Deleted: : an

Deleted: - SPRET

Deleted: - CAST

Deleted: - PWK

Deleted: - WSB

Deleted: our

Deleted: PseudoPipe

Deleted: a

Deleted: (GENCODE M8)

Deleted: genotype

Deleted: Complementarily, the lift over of manual annotations expands the available biotypes by including inactivated immunoglobulin and polymorphic pseudogenes.

Deleted: (Sup Table XX). On average we identified over 12,000 pseudogenes in each laboratory strain, over 11,000 pseudogenes in each of the wild strains, and just over 10,000 pseudogenes for each of the out group species. In order to annotate pseudogenes in the different mouse strains, we used as input a conse ... [4]

Moved down [4]: However, the manually annotate ... [5]

Deleted: upper bound. We expect the true size of ... [6]

Deleted: (Level 1),

Deleted: . With improvements in the annotation of ... [7]

Deleted: XX

Deleted: XX

Deleted: A small set of pseudogenes requires furth ... [8]

Deleted: .

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

These pseudogenes are defined relative to the functional protein coding elements in another species. Using a combination of multi sequence alignments, PhyloCSF scores, and a specialized unitary pseudogene annotation workflow (Figure 1), we identified 227 new unitary pseudogenes in human with respect to mouse and 210 unitary pseudogenes in mouse with respect to human (Supplementary Table S3). In human, a large proportion of unitary pseudogenes are characterized as disabled GPCRs, olfactory receptors, and vomeronasal receptor proteins with functional homologs in mouse, reflecting the loss of function in these genes during the primate lineage evolution.

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

Moreover, to get an overview of the unitary pseudogenes in each strain, we lifted over the reference pseudogene annotation and checked their conservation as pseudogenes or functional genes in each strain. We identified on average XX unitary pseudogenes. However, the short evolutionary distance between most strains means this value is an underestimate of the number of unitary pseudogenes. One way to get a more realistic assessment of the size of the unitary pseudogene complement is to look at the unitary annotation in the human genome relative to chimp. Given the fact that in humans there are XX unitary pseudogenes we expect to see a comparable number of unitary pseudogenes between the mouse strains.

2. Conservation and divergence in pseudogene complements

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

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

Furthermore, we grouped the conserved pseudogenes into subgroups based on their parent gene families (e.g. olfactory receptors, CDK, cytochrome C oxidase, etc.), or associated

Deleted: as well as information from

Deleted: we identified 102 new unitary pseudogenes in human (see Sup Table XXX). Of these, a large number are olfactory receptor pseudogenes with functional counterparts in mouse. ... [9]

Deleted: to identify unitary pseudogenes given two comparable genomes. Using this pipeline, we annotated additionally 225

Deleted: see Sup

Deleted: XXX). As expected

Deleted: number of the newly identified human

Deleted: present

Deleted: the

Deleted: chemosensory organ

Deleted: We also

Deleted: genes

Deleted: to the evolution of immune system

Deleted: . In particular, we found 5 new pseudogenes related to the

Deleted: (TLR11), a key player in defense against fungal

Deleted: bacterial infection, and activator of innate immunity. The lack of functional TLR11 in the human genome suggests that its functions might have been replaced by other immunity genes and thus its presence became expendable during evolution. We also observed the pseudogenization of a

Deleted: , related to the evolution of the immune system in primates

Deleted: The pseudogenization of these immune response genes hints at the ancestral disablement of sialic acid-recognized signaling receptors (siglecs) genes that provided early human ancestors with resistance against a number of bacterial pathogens shaping thus the course of human evolution \cite{ 22665810}.

Deleted:

Deleted: pseudogene complement

Deleted: were able to identify

Deleted: 15

Deleted: per strain relative to the reference.

Deleted: the

Deleted: that we expect to find relative to another species

Deleted: mouse

Deleted: over 200

Deleted: in

Moved (insertion) [5]

Moved (insertion) [6]

phenotype (e.g. rough coat, colour, diabetes, etc.) and constructed pseudogene phylogenetic trees for each of these subgroups (Figure 2, SF2). By comparing the resulting trees to the protein-coding one, we found that they display different patterns, reflecting different evolutionary histories. For example, the olfactory receptor 987 pseudogene tree, presents discrepancies both in 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. The rest of the strains, show little or no sequence variation at all compared to the common ancestor. This result hints at the previously observed link between obesity, metabolic diseases, and olfactory receptors \cite{25943692}, given the fact that the NZO, and NOD strains display a common diabetes specific phenotype.

3. Genome Evolution & Plasticity

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

F.1D

We further tested the correlation between high expression levels and large number of associated pseudogenes by looking at RNA-seq samples from adult mouse brain. Similar to our previous observations, the pseudogene parent genes show a statistically significant increase in average expression levels compared to non-pseudogene generating protein coding genes (Supplementary 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 3). 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

Deleted: 17

Deleted:

Deleted: 2

Deleted: that is currently unavailable in human

Deleted:

Deleted:

Deleted:

Deleted: assumption

Deleted: exceeds that of non-parent

Deleted: Furthermore

Deleted: and higher

Deleted: on

Deleted: Moreover,

Deleted: strong

Deleted: high gene expression levels and large

Deleted: associated

Deleted: observed in later

Deleted: however suggests the increased likelihood of

Deleted: producing pseudogenes.

Deleted: our hypothesis

Deleted: vs non-parent genes transcription in brain tissue for the 18 mouse strains. The results matched the ones observed earlier, with parent genes showing

Deleted: see Sup Fig XX

Deleted: Fig XXX

Deleted: pseudogene

Deleted: see there is

Deleted: results is can be explained by the fact

Deleted:

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 of a functional gene.

Deleted:

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.

Deleted: suggests

Deleted:

3.2 Transposable elements

Deleted: 2

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

Deleted: Fig XXX

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 in the human and mouse processed pseudogene complements. We define the evolutionary time scale by using the pseudogene sequence similarity to the parent gene as a proxy for age. Younger pseudogenes have a higher degree of sequence similarity to the parent, while older pseudogenes show a more diverged sequence.

Deleted:

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. 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 parents. Also by contrast to human where the density of L1 associated pseudogenes shows a steep decrease for young pseudogenes following the peak at 92.5, the density of mouse pseudogenes remains at high levels in the interval of 97 to 100% sequence similarity to parents. This observation suggests the presence of highly active transposable elements in mouse. The TE activity results in a continuous renewal of the processed pseudogene pool. This behavior is also reflected in the large difference in the number of active LINE/L1s between human and mouse, with just over 100 for human \cite{12682288} vs 3,000 for mouse \cite{ 11591644}.

Deleted: humans

Deleted: pseudogene

Deleted: processed pseudogenes.

Deleted: respective

Deleted: % sequence similarity to parents

Deleted: observations

Deleted: (100 vs 3,000s).

3.3 Genome remodeling

Deleted: 2

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 4). We observed that on average more than 97.7% of loci are conserved

Deleted: 17

Deleted: Fig XXX

Deleted: were

across the laboratory strains and 96.7% of loci are conserved with respect to the wild strains. By contrast only 87% of Caroli loci were conserved in the reference genome, while Pahari showed only 10% conservation. The significant drop in the number of conserved pseudogene loci between the reference genome and outgroup strains is in accord with the observed major karyotype-scale differences and large genomic rearrangements exhibited by Caroli and Pahari \cite{doi.org/10.1101/088435}. The proportion of un-conserved loci follows a logarithmic curve that matches closely the divergent evolutionary time scale of the mouse strains suggesting a uniform rate of genome remodeling processes across the murine taxa (Figure 4).

4. 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 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 (Fig 5). Moreover, the GO terms that universally characterize the pseudogene complements in all the mouse strains are closely reproduced in the family classification of pseudogenes. The top pseudogene family 7-Transmembrane encompasses the chemoreceptors GPCR proteins reflecting the mouse genome enrichment in olfactory receptors. Similar to the human and primate counterparts, the top families seen in mouse pseudogenes are related to highly expressed proteins such as GAPDH, ribosomal proteins and Zinc fingers.

However, 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. There are two possible types of pseudogene-phenotype associations. First, the pseudogenization process is linked with the emergence of an advantageous phenotype, as in the case of *Mus Spretus* genome, where we see an enrichment of pseudogenes related to tumor repressor genes and apoptosis pathways genes. Second, as expected, we find that the majority of pseudogenes reflect a deleterious phenotype. A known example is the pseudogenization of Cytochrome c Oxidase subunit VIa through accumulation of LOF mutations in the blind albino mouse strain, that is commonly linked with neurodegeneration \cite{17435251} and is characteristic for the observed brain lesions in the affected mice \cite{JAX}.

4.2 Gene essentiality

We observed an enrichment of essential genes among pseudogene parent genes in the mouse strains. Evaluating the parent gene for each pseudogene present in the mouse strains reveals essential genes are approximately three times more abundant amongst parent genes. In general, the essential genes are more highly transcribed than nonessential genes, and thus can potentially be associated with a higher propensity of generating processed pseudogenes.

Deleted: while

Deleted: were

Deleted: Fig XXX

Deleted: 3

Moved up [5]: . **Conservation and divergence in pseudogene complements**
In order to decipher the evolutionary history of the mouse strains we created a pangenome pseudogene dataset containing 49,262 unique entries relating the pseudogenes across strains.

Deleted: We found almost 3,000 ancestral pseudogenes that are preserved across all strains. A detailed summary of the other subsets of shared pseudogenes is shown in Table XXX. On average each strain contains 3,000 strain specific pseudogenes.

Moved up [6]: The proportion of pseudogenes conserved only in the outgroup, the wild strains, or the lab strains is considerably smaller, suggesting that the bulk of the pseudogenes in each strain are derived during the shared evolutionary history.

Deleted: A pair-wise analysis of the 3 classes of strains (Fig XXX) shows that the outgroup strains share a large number of pseudogenes with the laboratory strains than with the wild strains, despite being evolutionarily closer to the latter. This anomaly is potentially related to the diversity of mouse wild strains but also to the slightly lower quality of genome assembly available for this class of mice. By contrast, pairwise analysis within each class points to a uniform distribution of shared pseudogenes, reflecting the close evolutionary history between the strains of each ... [10]

Deleted: contribute to

Deleted: pseudogene

Deleted: terms

Deleted: address one of

Deleted: key questions surrounding pseudogenes: what is their biological significance?

Deleted: pseudogene complement of the

Deleted: strains share the same

Deleted: components

Deleted: XXX

Deleted: XXX. We observed

Deleted: . This is

Deleted: for

Deleted: reflecting

Deleted: This can be seen in the blind albino mou ... [11]

Deleted: enriched in

Deleted: (COX)

Deleted: pseudogenes, and it has been previousl ... [12]

Deleted: COX are cause for

Deleted: across all

Moved down [7]: Lists of essential and nonessen ... [13]

Moved down [8]: The nonessential gene set with ... [14]

Deleted: which suggests that they are more likely ... [15]

We evaluated the probability that a gene is essential given its transcription level and parent gene status (see Methods) and found that pseudogene parents are 20% more likely to be essential genes compared to regular protein coding genes.

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

4.3 Pseudogene Transcription

We leveraged the available RNA-seq data from the Mouse Genome Project and ENCODE 3 to study pseudogene biology as reflected by their transcription potential. This is thought to either relate to the exaptive functionality of pseudogenes or be a residual leftover from their existence as genes. For both the human and the mouse reference genomes, we detected that about 15% of pseudogenes were transcribed across a variety of tissues, a result similar with previous pan tissue analyses. Due to data availability for the 18 mouse strains, we restricted our tissue analysis to brain. Similar to the previously observed pattern in human and other model organisms, pseudogene transcription in mouse strains shows higher tissue and strain specificity compared to the protein coding counterpart (Supplementary Figure SF4). Also, pseudogenes with strain specific transcription were more common than those with cross-strain transcription.

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

5. Mouse pseudogene resource

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

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

Deleted: was evaluated

Deleted: provide

Deleted: gene

Deleted: seen

Deleted: and

Deleted: with at least one paralog. Such genes in the two groups have an average of 6.2 and 6.7 paralogs per gene respectively.

Deleted: due

Deleted: reliance

Deleted: actual

Deleted: pseudogenes

Deleted: them

Deleted: analysis

Deleted: tissue.

Moved down [9]: Both human and mouse show a consistent transcription level in brain, with 5% of the total pseudogene complement being transcribed (see Sup Fig XX).

Deleted: We also identified xxx% transcribed pseudogenes that show a discordant expression pattern with respect to their parent genes.

Deleted: humans

Deleted: see Sup Fig XX).

Deleted: conserved

Deleted: see Sup Fig XX

Deleted: 17

Deleted: Fig XXX).

Deleted: general

Deleted: of each pseudogene

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

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

Discussion

We describe the annotation and comparative analysis of the first draft of the pseudogene complements in the mouse reference genome and 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 combined 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 sale chromosomal imagining and karyotype analysis \cite{https://doi.org/10.1101/088435}.

Examination of pseudogene complement informs of retrotransposons 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 the continuous transposable elements activity. Looking closely at the sequence context of the processed pseudogenes reveals that the various retrotransposons exhibit differential contributions to the pseudogene set over time as well.

Deleted: 17

Deleted: This high confidence set comprises 30% of the total population. We expect the number of annotations in this set to grow as manual annotation catches up with the automated pipelines.

Deleted:

Deleted: also

Deleted: that observed

Deleted: humans

Deleted: previous

Deleted: Furthermore, the size of the pseudogene complements are generally similar, although a slight decrease in strains more divergent from the reference mouse is observed. This trend is likely due to the smaller number of conserved protein coding genes between these strains and the reference genome, which are used for pseudogene identification. Future improvements in the protein coding gene sets for the more evolutionarily distant species will help refine these strains' pseudogene annotations.

Deleted: 18

Deleted: mouse

Deleted: pan genome

Deleted: XX

Deleted: YY

Deleted: Investigating the location

Deleted: conserved

Deleted: suggests

Deleted: noted elsewhere

Deleted: The pan-genome

Deleted: sets also illustrate how the activity

Deleted: has

Deleted: changing

Deleted: content

Deleted: continual activity of

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 with respect to human and a similar number in human with respect to mouse. We found that the enrichment of vomeronasal receptor unitary pseudogenes in human with respect to mouse highlights the loss of certain olfactory functions in humans. Moreover the variety of unitary pseudogenes provides a valuable source for LOF analysis giving an indication of both advantageous and deleterious phenotype in both the mouse and human lineage.

I.

Deleted: For example,

Meanwhile, since a processed pseudogene's likelihood of creation is proportional 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. By contrast, 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 house-keeping genes.

? ✓

Consequently, 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 for the mouse pseudogenes closely matches those seen in the human set. Studying the 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 suggests 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.

Deleted: an

Deleted: is

Deleted: Analysis of

Deleted: similarly

Deleted: utilizing

Deleted: generate

Deleted: *Mus Spretus* is associated with the strain specific death family, whose association with apoptosis indicates a potential mechanism for the strain's tumor resistant 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.

Deleted: The wealth of functional genomics assays available for the experimentally relevant mouse strains presents an opportunity to investigate the both the activity of parent genes as well as potential activity of pseudogenes. As expected parent genes have higher levels of expression relative to non-parents.

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.

Moved (insertion) [3]

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

*Chromosomal assembled DNA only

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/6J – Black 6	
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	

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

Organism	Manual	Pseudopipe			
		Autosomes	Sex Chromosomes	Others*	Total
Mouse	10,524	14,084	4,565	4,037	22,686
Human	14,650	14,644	1,325	2,098	18,067

*Includes patches, scaffolds, and unassembled DNA.

Table S2. Human GENCODE pseudogene annotation summary.

Pseudogenes total	14650
processed pseudogenes	10725
unprocessed pseudogenes	3400
unitary pseudogenes	214
polymorphic pseudogenes	51
unknown pseudogenes	21

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

Methods

Pseudogene Annotation Pipeline

Reference genome annotation

We manually curated almost 10,000 pseudogenes in the mouse reference genome using a workflow 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 used 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 14,000 are present in autosomal chromosomes.

These numbers are comparable to those seen in human (Table S1).

Mouse strain annotation

The lack of available high quality protein coding and peptide annotations in the 18 mouse strains created a bottleneck in the pseudogene identification process. This was resolved by generating protein input sets that are shared between the 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. As pseudogene annotation is highly depended on availability and quality of the protein coding annotation, we expect that the differences in the pseudogene complements to follow closely the ones in the protein coding annotation.

As such, using PseudoPipe on average we identified over 11,000 pseudogenes in each laboratory strain, over 10,000 pseudogenes in each of the wild strains, and just over 9,000 pseudogenes for each of the out group species. As expected, the difference in the pseudogene complement size closely follows the variation in the number of conserved protein coding genes between each strain and the reference genome reflecting the evolutionary distance between each strain and the reference genome.

Next, we lifted over manually annotated pseudogenes from the mouse reference genome onto each strain and were able to identify on average over 6,000 pseudogenes in the laboratory strains, over 5,000 pseudogenes in each of the wild strains and over 4,000 pseudogenes in each of the out group species.

Finally, we integrated the two annotation sets (PseudoPipe and liftover of manually curated elements) are merged them to produce the final pseudogene complement set. The merging process was conducted by overlapping the annotations (using 1 bp minimum overlap) and extending the predicted boundaries to ensure the full annotation of the pseudogene transcript. A Level 1 designation indicates a high confidence prediction, with the annotated pseudogene being validated by both automatic and manual curation processes, Level 2 pseudogenes are

Moved (insertion) [10]

Moved (insertion) [1]

Moved (insertion) [2]

Moved up [10]: Pseudogene Annotation Pipeline

Deleted: level

Deleted: 17

Moved (insertion) [11]

Moved up [11]: This was resolved by generating protein input sets that are shared between the strain and the reference genome. The number of shared transcripts

Moved down [12]: 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.

Moved (insertion) [12]

Deleted: individual

identified only through the manual lift-over of the GENCODE reference genome annotations, while Level 3 pseudogenes are predicted solely using the automation identification pipeline.

However, the manually annotated pseudogenes are a lower bound of the total number of pseudogenes in each strain. Meanwhile, the size of reference genome pseudogene complement identified using the automatic annotation pipeline represents a low sensitivity upper bound. We expect the true size of the pseudogene complement in the mouse strains to be comparable to the number of pseudogenes in reference genome.

Moved (insertion) [4]

Deleted:

Unitary Pseudogene Annotation Pipeline

We adapted PseudoPipe to work as part of a strict curation workflow that can be used both in identifying cross-strain and cross species unitary pseudogenes. A schematic is shown in Figure 1. In summary, we define the “functional” organism as the genome providing the protein coding information and thus containing a working copy of the element of interest, and the “non-functional” organism as the genome analysed for pseudogenic presence, containing a disabled copy of the gene. In order to make sure that false positives are eliminated, we introduced a number of filtering steps for removing all cross-species pseudogenes or pseudogenes with orthologous parent genes in the two organisms.

Deleted:

Data integration & pangenome pseudogene generation

Utilizing the pseudogene annotations for each strain and liftover mappings between the different strains under investigation we generated sets of pseudogenes shared amongst different subsets of strains. The pseudogene annotations from one strain are lifted over onto the genomic coordinates of each of the other strains. Pseudogenes conserved between each binary combination of strains are identified by looking for the intersection of the lifted over pseudogene annotations and the native pseudogene annotations. 90% reciprocal overlap between two annotations is required to identify them as conserved. In order to remove false positives the conserved pseudogenes are filtered for pseudogene identity, parent identity, genomic location, size, biotype, and structure conservation. The sets of binary conserved pseudogenes are merged into a master set from which unique entries are filtered and extracted.

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.

Moved (insertion) [7]

Moved (insertion) [8]

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.

Cross strain pseudogene transcription

Both human and mouse show a consistent transcription level in brain, with 5% of the total pseudogene complement being transcribed (see Sup Fig XX).

Moved (insertion) [9]

Page 3: [1] Deleted **Microsoft Office User** **5/28/17 1:48:00 PM**

its gene, can also be advantageous. The relaxation of the selection constraints on such a gene would favor the accumulation of disabling mutations, eventually resulting in fixation of that pseudogene in the organism.

A well knonw example of a LOF event creating an advantageous phenotype is the accumulation of loss-of-function mutations in the

Page 4: [2] Deleted **Microsoft Office User** **5/28/17 1:48:00 PM**

These numbers are comparable to those seen in human (Table XXX). This automatic annotation provides an upper bound on the number of pseudogenes present.

Table XXX

Page 4: [3] Deleted **Microsoft Office User** **5/28/17 1:48:00 PM**

Organism	**Pseudopipe**			**Manual**	**Overlap Manual vs Pseudopipe (%)**
	Autosomes	Others*	Total		
Mouse	14,084	8,602	22,686	10,524	8,786 (83.5)
Human	14,644	3,423	18,067	14,650	13,177 (89.9)

*Includes sex chromosomes, patches, scaffolds, and unassembled DNA.

In

Page 4: [4] Deleted **Microsoft Office User** **5/28/17 1:48:00 PM**

(Sup Table XX). On average we identified over 12,000 pseudogenes in each laboratory strain, over 11,000 pseudogenes in each of the wild strains, and just over 10,000 pseudogenes for each of the out group species. In order to annotate pseudogenes in the different mouse strains, we used as input a consensus set of protein coding genes between each strain and the reference genome. Consequently, the difference in the pseudogene complement size closely follows the variation in the number of conserved protein coding genes between each strain and the reference genome reflecting the evolutionary distance between each strain and the reference genome.

Page 4: [5] Moved to page 15 (Move #4) **Microsoft Office User** **5/28/17 1:48:00 PM**

However, the manually annotated pseudogenes are a lower bound of the total number of pseudogenes in each strain. Meanwhile, the size of reference genome pseudogene complement identified using the automatic annotation pipeline represents a low sensitivity

Page 4: [6] Deleted **Microsoft Office User** **5/28/17 1:48:00 PM**

upper bound. We expect the true size of the pseudogene complement in the mouse lineage to be comparable to the number of pseudogenes in human genome (e.g. ~14,000).

Page 4: [7] Deleted **Microsoft Office User** **5/28/17 1:48:00 PM**

. With improvements in the annotation of the mouse reference genome as well as refinement of the strain assemblies and annotation, we expect the number of high confidence annotations will increase, matching the fraction observed in the human genome.

Page 4: [8] Deleted **Microsoft Office User** **5/28/17 1:48:00 PM**

A small set of pseudogenes requires further analysis of their formation mechanism in order to assign the correct biotype.

Page 5: [9] Deleted **Microsoft Office User** **5/28/17 1:48:00 PM**

we identified 102 new unitary pseudogenes in human (see Sup Table XXX). Of these, a large number are olfactory receptor pseudogenes with functional counterparts in mouse.

Next, we developed

Page 8: [10] Deleted **Microsoft Office User** **5/28/17 1:48:00 PM**

A pair-wise analysis of the 3 classes of strains (Fig XXX) shows that the outgroup strains share a large number of pseudogenes with the laboratory strains than with the wild strains, despite being evolutionarily closer to the latter. This anomaly is potentially related to the diversity of mouse wild strains but also to the slightly lower quality of genome assembly available for this class of mice. By contrast, pairwise analysis within each class points to a uniform distribution of shared pseudogenes, reflecting the close evolutionary history between the strains of each class.

Next we took advantage of pseudogenes ability to evolve with little or no selective constraints \cite{10833048}, to analyze and compare comparing 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 (see Fig XXX). This pseudogene-based tree correctly identifies and clusters the strains into three classes: outgroup, wild, and laboratory strains.

We grouped the conserved pseudogenes into subgroups based on their parents' protein families (e.g. olfactory receptors, CDK, leucine rich repeats, cytochrome C oxidase, etc.), and phenotypic characterization (e.g. rough coat, colour, diabetes, etc.). We constructed pseudogene phylogenetic trees for each of these subgroups (see Fig XXX, Sup Fig XXX). By comparing the resulting trees to the protein-coding one, we found that they display a different evolution pattern, some sequences evolving faster and some slower than the corresponding protein coding genes.

For example, the olfactory receptor 987 pseudogene tree, while maintaining Pahari as an outgroup species, presents a completely different evolutionary history for the 17 strains both in divergence order as well as in the degree of conservation of the ancestral sequence (as reflected by the branch length). In particular, we observed striking sequence changes in 129S1, NZO (New Zealand obese mouse), and NOD (non-obese diabetic mouse) laboratory strains, and smaller differences with respect to the common ancestor gene in SPRET and PWK wild strains. The rest of the strains, including Caroli and CAST, show little or no sequence variation at all compared to the common ancestor. The large number of changes observed in the olfactory receptor sequences in NZO and NOD hint towards the previously link observed between obesity, metabolic diseases, and olfactory receptors \cite{25943692}, given the fact that the two strains display a common diabetic prone phenotype.

Page 8: [11] Deleted **Microsoft Office User** **5/28/17 1:48:00 PM**

This can be seen in the blind albino mouse strain (BALB), a representative line for neurodegenerative disorders (100% of subjects developing severe brain lesions \cite{JAX}). BALB

Page 8: [12] Deleted **Microsoft Office User** **5/28/17 1:48:00 PM**

pseudogenes, and it has been previously reported that disabling

Page 8: [13] Moved to page 15 (Move #7) **Microsoft Office User** **5/28/17 1:48:00 PM**

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}.

Page 8: [14] Moved to page 15 (Move #8) **Microsoft Office User** **5/28/17 1:48:00 PM**

The nonessential gene set with Ensembl identifiers contained 4,736 genes compared to 3,263 essential genes.

Page 8: [15] Deleted **Microsoft Office User** **5/28/17 1:48:00 PM**

which suggests that they are more likely to generate