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} toward completing the de-novo assembly and gene annotation of a variety of mouse strains, provide a unique opportunity to get an in-depth picture of the evolution and variation of these closely related mammalian species.

Mice have 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 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, the makeup of the present human population parallels, at generation divergence levels, the evolution of recent *Mus Musculus* strains including inbred laboratory mouse strains \cite{17284675} (Figure 1). The mouse strains under investigation possess 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}. 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 17 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 and function \cite{10692568,11160906,12034841,14616058}. Pseudogenes are previously coding DNA sequences that contain disabling mutations rendering them unable to produce a fully functional protein. There are different classes of pseudogenes based on their

Formatted: Heading 3
Formatted: Font:Not Bold
Deleted: The recent
Deleted:
Deleted: the
Deleted: the
Deleted: would aid in deciphering genome evolution and diversity in human population.
Moved down [1]: In this paper we describe the first pseudogene annotation and analysis of 17 widely-used inbred mouse strains alongside the reference mouse genome.
Deleted: Despite obvious discrepancies between humans and mice: e.g mice are small, have a short life span and high metabolic rate, the two species share a large number of similarities in their genetic makeup, particularly in tumor and disease development, making mice ideal model organisms for the study of human diseases \cite{14978070}. Understanding the genesis and functional impact of the genetic makeup of these
Deleted: These strains
Deleted: Also the
Deleted: Specifically
Moved (insertion) [1]
Deleted: and mouse
Deleted:

creation mechanism: processed pseudogenes – formed through a retrotransposition process, duplicated pseudogenes – formed during 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 additional 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. The composition of these different pseudogene classes across the mouse strains provides insight into changes in selective pressures and genome remodeling forces.

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 thus an extreme case of LOF, where the mutations result in the complete inactivation of the gene. There is a fine line between a loss-of-function variant that is increasing in a population and a pseudogene that is only partially fixed in that population, also known as polymorphic pseudogene. 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, sometimes, in the right conditions, the inactivation of a protein via pseudogenization of 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 prominent example of a LOF event creating an advantageous phenotype is the accumulation of loss-of-function mutations in the proprotein convertase subtilisin/kexin type 9 (PCSK9) gene. When expressed, the PCSK9 protein binds to the low-density lipoprotein (LDL) receptor leading to its degradation and a reduced cellular uptake of plasma LDL \cite{18631360}. Enrichment of plasma LDL cholesterol is often associated with an increased risk of atherosclerosis. 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. This finding has inspired the creation of PCSK9 inhibitors as a 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.

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 in the mouse strains – 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 presents an opportunity to investigate the processes underlying pseudogene biogenesis and activity to an extent previously not possible. Insights gained from the comparative analysis of pseudogenes across the strains can be validated via either existing functional datasets or future experimentation. Comparison to the primate lineage and human population is an exciting 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

Deleted: defined as

Deleted: when

Deleted: has

Deleted: This is the case for the myosin gene 16 pseudogenization, which has been suggested to be related to the acquisition of human-specific phenotypes in the primate lineage \cite{25887751,16464126}. Another known

Deleted: the

Deleted: thus a

Deleted: From a functional genomics perspective there is a fine line between a loss-

Deleted: -function variant that is increasing in a population and a pseudogene that is only partially fixed in that population, also known as polymorphic pseudogene. The process that gives rise to these pseudogene types

Deleted: a great

Deleted: more

Deleted: the organism's essential gene pool. Additionally, LOF mutations are important for identifying

Deleted: components in

Deleted: \cite{ecoliwiki}. LOF events give an

Deleted: selection

Deleted: the

Deleted: While most of these areas cannot be easily studied in human, mouse provides as excellent platform for LOF analysis. Moreover, the short

Deleted: the distance

Deleted: current

and associated pseudogene annotations for the 18 strains a valuable resource for population studies.

Results

1. Annotation

We first 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. Novel pseudogene annotations for a set of mouse strains with recently assembled high quality genome sequences are also presented and analyzed.

1.1 Reference genome pseudogene annotations

Using a rigorous manual curation process as previously described in the GENCODE annotation resource \cite{22951037,25157146}, we identified almost 10,000 pseudogenes in the mouse reference genome. 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 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} and the updated reference mouse protein coding annotations. PseudoPipe is a comprehensive annotation pipeline focused on identifying and characterizing pseudogenes based on their biotypes as either processed or duplicated.

The compuational pipeline identified approximately 22,000 pseudogenes of which 18,000 are present in assembled chromosomes and 4,000 are annotated in chromosome patches and scaffold DNA. 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. Reference genome pseudogene annotation in mouse and human

Organism	Pseudopipe			Manual	Overlap Manual vs Pseudopipe (%)
	Chromosome	Patches	Total		
Mouse	18,649	4,037	22,686	10,524	8,786 (83.5)
Human	15,978	2,098	18,067	14,650	13,177 (89.9)

In human we used a combination of automatic and manual curation to refine the reference pseudogene annotation to a set of 14,650 [[CSDS +200 unitary]] pseudogenes. The updated set contains considerable improvements in the identification of unitary pseudogenes, as well as better characterization of pseudogenes of previously unknown biotype (see Table XXX). 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: an outgroup – formed by two independent mouse

Deleted: 17
Deleted: an incredible
Deleted: Finally, we are also able to take advantage of the use of the mouse as a model organism, to paint a comprehensive picture of the functional changes that occur in the genome during embryonic development and find answers to questions like are pseudogenes the result of somatic or germline LOF events? By extension we will be able to shed light on the analogous human processes that are of particular interest to the biomedical and pharmaceutical industry.
Deleted: **General considerations** ... [1]
Formatted: Heading 3
Formatted: Font:Not Bold
Deleted:
Deleted: annotation of
Deleted: Given the similarities between the human and mouse genomes, and the fact that in human we have identified over 14,000 pseudogenes, the
Deleted: .
Deleted: took advantage of the updated protein coding annotation, and
Deleted: to identify ~
Deleted: the ones
Deleted: Thus the
Deleted: with respect to
Deleted: total
Deleted:
Formatted: Font:Bold
Formatted Table
Formatted: Right: -0.07"
Formatted: Indent: Left: -0.07", Right: -0.07"
Formatted: Font:Bold
Formatted: Font:12 pt
Formatted: Font:Bold
Formatted: Font:(Default) Arial
Formatted: Font color: R,G,B (38,38,38)
Deleted: PseudoPipe is a comprehensive pseudogene annotation pipeline focused on identifying and characterizing them based on their biotypes as either processed or duplicated. More than half of the annotations are processed pseudogenes, with a smaller fraction of duplicated pseudogenes (Figure 1).
Deleted: a

species, *Mus Caroli* and *Mus Pahari*; wild strains – covering two subspecies (*Mus Spretus* - SPRET and *Mus Castaneus* - CAST) and two musculus strains (*Mus Musculus Musculus* - PWK and *Mus Musculus Domesticus* - WSB), and a set of laboratory strains. A detailed summary of the genome composition for each strain is presented in \cite{MousePaper}.

We developed an annotation workflow for identifying pseudogenes in the 18 mouse strains, leveraging our automatic pipeline PseudoPipe and a set of manually curated pseudogenes from the mouse reference genome (GENCODE M8) lifted over onto each individual strain genotype. Each identified pseudogene is provided with details about the transcript biotype, genomic location, structure, sequence disablements, and a confidence level reflecting the annotation process. Complementarily, the lift over of manual annotations expands the available biotypes by including inactivated immunoglobulin and polymorphic pseudogenes.

A detailed overview of pseudogene annotation statistics including the number of pseudogenes, their confidence levels, and related biotypes is shown in Figure 1 (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. This reflects the evolutionary distance between each strain and the reference genome.

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 identification pipeline represents a low sensitivity 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).

Currently, around 30% of pseudogenes in each strain are defined as high confidence annotations (Level 1), 10% Level 2, and 60% Level 3. 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.

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 (~XX%) of the annotations are processed pseudogenes, while a smaller fraction (~XX%) are duplicated pseudogenes. A small set of pseudogenes requires further analysis of their formation mechanism in order to assign the correct biotype.

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. 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

Deleted: each strain

Deleted: , by

Deleted: , as well as

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

Deleted: The

Deleted: Additionally, the pseudogene complements reflect

Deleted: just

Deleted: curated by

Deleted: Thus, we

Deleted: follows

Deleted: },

Deleted: mouse

Deleted: also

Deleted: shows

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. Thus 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. In order to get an overview of the unitary pseudogene complement in each strain, we lifted over the reference annotation and were able to identify on average 15 unitary pseudogenes per strain relative to the reference. However, the short evolutionary distance between most the strains means this value is an underestimate of the number of unitary pseudogenes that we expect to find relative to another species. One way to get a more realistic assessment of the size of the unitary pseudogene complement in the mouse strains is to look at the unitary annotation in the human genome relative to mouse. Given the fact that in humans there are over 200 unitary pseudogenes we expect to see a comparable number of unitary pseudogenes in mouse.

We developed a specialized workflow to identify unitary pseudogenes given two comparable genomes. Using this pipeline, we annotated 237 unitary pseudogenes in human with respect to mouse and 210 unitary pseudogenes in mouse with respect to human (See table XXX). As expected a large number of the newly identified human unitary pseudogenes are characterized as GPCRs, olfactory receptors, and vomeronasal receptor proteins present in the mouse chemosensory organ, reflecting the loss of function in these genes during the primate lineage evolution. We also observed the pseudogenization of a number of genes related to the evolution of immune system in humans. In particular, we found 5 new pseudogenes related to the Toll-like receptor gene 11 (TLR11), a key player in defense against fungal and 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 leucine rich repeat protein, related to the evolution of the immune system in primates \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.

2. Genome Evolution & Plasticity

Leveraging the pseudogene annotations, we explore the differences between the 17 mouse strains by looking at the genome remodeling processes that shaped the evolutionary history of their pseudogene complements.

2.1 Phylogeny

It has long been held that pseudogenes evolve with little or no selective constraints \cite{10833048}, and that the mutation rate in pseudogenes reflects the underlying genome substitution pattern \cite{11752196}, making them ideal elements for inferring and 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).

Deleted: .
Deleted: thus
Deleted: mouse strain
Deleted: in each
Deleted: .
Deleted: real
Deleted: .
Deleted: in mouse
Deleted: -
Deleted: -functionality
Deleted: human
Deleted: commonly
Deleted: human
Deleted: associated
Deleted: futile
Deleted: commonly
Formatted: Heading 4
Deleted:
Deleted: annotation
Deleted: about

This pseudogene-based tree correctly identifies and clusters the strains into three classes: outgroup, wild, and laboratory strains.

Next 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 an independent, strain specific evolutionary pattern. The deviations of the pseudogene trees from the known lineage pattern reflect roles played by pseudogenes during the strains evolution.

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.

2.2 Conservation

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 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. 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. 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.

2.3 Transposable elements

Mammalian genomes are known for their variety and large number of transposable elements (TE or mobile elements). 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.

We investigate the evolution of the processed pseudogene complements in the human and mouse lineages by looking at the enrichment of TE families in the two species on an evolutionary time scale (Fig XXX). Both human and mouse genomes are dominated by three types of TEs, namely short interspersed nuclear elements (SINEs), long interspersed nuclear

Deleted: The

Deleted: noticed

Deleted: the

Deleted: Of these, we

Deleted: pseudogene types

Deleted: contest

Formatted: Not Highlight

Deleted: that are

Deleted:

Deleted: lineage

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.

In humans we observe a smooth distribution of TEs, with a single peak hinting at the burst of retrotransposition events, that occurred 40 MYA at the dawn of primate lineage and created the majority of human processed pseudogenes. By contrast in mouse we found the TE distribution is defined by multiple successive peaks suggesting a highly active set of transposable elements. The TE activity results in a continuous renewal of the processed pseudogene pool. This contrasting behavior is also reflected in the large difference in the number of active LINE/L1s between human and mouse (100 vs 3,000s).

2.4 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 17 mouse strains and the reference genome for one-to-one pseudogene orthologs in each pair (Fig XXX). We observed that on average more than 97.7% of loci were conserved across the laboratory strains while 96.7% of loci were 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 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 (Fig XXX).

2.5 Pseudogene paralogs

To the extent that pseudogenes resulting from retrotransposition processes are, by their mechanism of creation, not constrained to the localization of their parent genes, the large proportion of processed pseudogenes in the mouse lineage shaped the genomic neighborhood of each strain, competing with successful duplications and retrotranspositions resulting in functional paralogs of their parent genes.

In order to understand the ratio of successful copies to disabled copies of genes, we compared the number of pseudogenes with the number of functional paralogs for each parent gene based on the mechanisms of formation (retrotransposition and duplication) (Fig XXX). The gene duplication process can result in either the creation of a functional copy or a disabled one - a pseudogene. Looking at the two possible outcomes, we observed a direct correlation between the number of duplicated pseudogenes and the number of duplicated paralogs per gene, with the ratio of the two being tilted towards the creation of functional elements.

By contrast, processed pseudogenes are formed through a retrotransposition process. As such, the expression level, as well as the number of protein coding gene copies in the genome will have an impact on the pseudogene genesis. To this end we looked at transcription in protein coding genes grouped into two classes: parents of pseudogenes and non-parent genes. We observed that genes associated with pseudogenes consistently show a higher level of transcription compared to their non pseudogene related counterparts (see Sup Fig XX).

Deleted: Thus, younger

Deleted: The violin plots show

Deleted: 40MYA

Deleted: noticed that

Deleted: can be accounted for by

Deleted:

Deleted:

Deleted:

Deleted: regular

Deleted: show

Deleted:

Next we tested the hypothesis that a highly expressed gene, characterized by a high number of mRNAs, will be associated with an increase in the number of pseudogene homologs. However, a large number of mRNAs can be achieved in a number of ways: e.g. from a protein coding gene with few copies in the genome but high expression level, multiple copies of a protein coding gene with low individual expression level, etc. Thus, there is no direct positive correlation between genes with high number of paralogs and high number of pseudogenes. Moreover, similar to the human counterpart, the mouse pseudogene complement exhibits an inverse proportional evolution of the number of processed pseudogenes relative to the number of paralogs per gene.

Deleted: low number of

3. Biological relevance

Formatted: Font:Not Bold

Formatted: Heading 4

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

Deleted: the

Deleted: RNAseq

3.1 Gene ontology & pseudogene family analysis

We integrated the pseudogene annotations with gene ontology (GO) terms in order to address one of the key questions surrounding pseudogenes: what is their biological significance? For this we calculated the enrichment of GO terms across the strains. We observed that the pseudogene complement of the majority of strains share the same biological processes, molecular function and cellular components (Fig XXX). 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.

Deleted: annotation

Deleted: data

Deleted: , hinting at the shared evolutionary history between the various mouse strains (Fig XXX).

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 XXX. We observed two possible types of pseudogene-phenotype associations. First, the pseudogenization process is linked with the emergence of an advantageous phenotype. This is the case for *Mus Spretus*, where we see an enrichment of pseudogenes related to tumor repressor genes and apoptosis pathways genes. Second, we find pseudogenes reflecting a deleterious phenotype. 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 is enriched in Cytochrome c Oxidase (COX) subunit VIa pseudogenes, and it has been previously reported that disabling mutations in COX are cause for neurodegeneration \cite{17435251}.

Formatted: Font:Italic

Deleted:

3.2 Gene essentiality

We observed an enrichment of essential genes among pseudogene parent genes across all 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. 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

Deleted: and the likelihood that a gene has produced pseudogenes.

Moved (insertion) [2]

nonessential gene set with Ensembl identifiers contained 4,736 genes compared to 3,263 essential genes. In general, the essential genes are more highly transcribed than nonessential genes, which suggests that they are more likely to generate processed pseudogenes.

The number of paralogs associated with essential and nonessential genes was evaluated to provide insight into the possible role of gene duplication in the enrichment of essential genes amongst the parent gene 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 seen for essential and nonessential genes with at least one paralog. Such genes in the two groups have an average of 6.2 and 6.7 paralogs per gene respectively. The slight depletion of genes with paralogs in the experimentally determined essential gene set is likely due to the reliance on single gene knockouts to determine essentiality, which would miss genes with an essential role and a functional paralog.

3.3 Pseudogene Transcription

We leveraged the available RNA-seq data from the Mouse Genome Project to study pseudogene biology as reflected by their transcription potential. Previous pan tissue analysis pointed towards a uniform level of pseudogene transcription in both the mouse and human genomes, with 15% of the total pseudogenes showing a residual level of transcription. By contrast, the proportion of pseudogenes showing tissue specific transcription varies in both human and mouse. In this project we analyzed the pseudogene transcription in brain tissue for human and mouse. Both species show a consistent transcription level in brain, with 5% of the total pseudogene complement being transcribed (see Sup Fig XX). We also identified xxx% transcribed pseudogenes that show a discordant expression pattern with respect to their parent genes. Similar to the previously observed pattern in humans and other model organisms, pseudogene transcription in mouse strains shows higher tissue and strain specificity compared to the protein coding counterpart (see Sup Fig XX). Also, pseudogenes with strain specific transcription were more common than those with conserved 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 (see Sup Fig XX).

Utilizing rich functional datasets available for the mouse enabled us to further investigate the processes underlying pseudogene creation. An embryogenesis RNA-seq time course provided an opportunity to investigate parent gene expression during early development \cite{27309802}. We investigated parent gene expression over a series of developmental stages ranging from metaphase II oocytes to the inner cell mass. At every stage the average expression of parent genes exceeds that of non-parent genes. Furthermore, plots of gene expression vs. the number of pseudogenes for each parent gene reveal that parent genes associated with a large number of pseudogenes have low expression in early development. In later developmental stages however, these parent genes begin to exhibit higher levels of transcription, which suggests the increased likelihood of highly expressed housekeeping genes producing pseudogenes.

Deleted:

Deleted:

Moved up [2]: . Evaluating the parent gene for each pseudogene present in the mouse strains reveals essential genes are approximately three times more abundant amongst parent genes.

Deleted: We observed an enrichment of essential genes among pseudogene parent genes across all mouse strains

Deleted: Genes in the essential gene set exhibit higher levels of expression at multiple time points during mouse embryonic development. This suggests that higher expression of these genes during early development might lead to additional retrotransposition events resulting in new pseudogenes.

Deleted:

Deleted:

4. Mouse pseudogene resource

We created a pseudogene resource that organizes all of the pseudogenes across the 17 mouse strains and reference genome, as well as associated phenotypic information in a MySQL database (Fig XXX). The database contains three general types of information: details about the annotation of each pseudogene, 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.

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

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

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

Discussion

We describe the annotation and comparative analysis of the first draft of the pseudogene complements in the mouse reference genome and 17 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. 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.

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 also in line with that observed in humans. The higher proportion of processed pseudogenes is in agreement with previous findings that retrotransposition is the primary mechanism for pseudogene creation \cite{22951037}. 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

Formatted: Font:Not Bold
Formatted: Heading 4
Moved down [3]: 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.
Moved (insertion) [3]
Formatted: Font:Not Bold
Formatted: Heading 3
Deleted: complement
Deleted: the in house
Deleted: complement
Deleted: genome
Deleted: the
Deleted: pseudogenes
Deleted: the ones
Deleted: accuracy
Deleted: Given the similarities between the human and mouse genomes we expect that
Deleted: pseudogenes
Deleted: the mouse genome
Deleted: match closely the one reported for human

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.

Integrating the annotations from the 18 strains we obtained a pan genome mouse pseudogene set composed of over 45,000 unique entries. This pan genome set contains three types of pseudogenes: universally conserved, multi-strain, and strain specific, accounting for 6, XX, and YY% 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. Investigating the location of pseudogenes conserved between strains suggests multiple large scale genomic rearrangements in the mouse lineage. This is especially striking in the case of *Mus Pahari* as has been recently noted elsewhere \cite{https://doi.org/10.1101/088435}.

The pan-genome pseudogene sets also illustrate how the activity of retrotransposons has contributed to pseudogene creation and changing genomic content 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 continual activity of transposable elements. 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.

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. For example, the enrichment of vomeronasal receptor unitary pseudogenes in human with respect to mouse highlights the loss of certain olfactory functions in humans. 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.

Consequently, an 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 an enrichment for housekeeping functions is 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. Analysis of recurrent gene ontology terms similarly supports the enrichment of pseudogenes for important biological processes with terms including RNA processing and metabolic processes. Additionally, utilizing the pan-genome pseudogene set to identify strain specific functional annotations can generate 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}. *Mus Spretus* is associated with the strain specific death family, whose association with apoptosis indicates a potential mechanism for the strain's tumor resistant

Deleted: In order to annotate pseudogenes in mouse strains, we used as input a consensus set of protein coding genes between each strain and the reference genome. The size of the pseudogene complement follows closely the number of conserved protein coding genes between each strain and the reference genome. Also, we found that the relative ratio of processed to duplicated pseudogenes is preserved across all strains.

Deleted: annotation in

Deleted: The

Deleted: and multi-strain,

Deleted: respectively

Deleted: . By comparing the pseudogene complements across 6 million years

Deleted: evolution we obtain

Deleted: global

Deleted: shaped

Deleted: As such, we constructed phylogenetic trees using pseudogene sequences and we were able to associate strain phenotypes with strain specific pseudogenes.

Deleted: genome

Deleted: ,

Deleted: continued renewing

Deleted: constant

phenotype. 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.

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. 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.

Methods

Pseudogene Annotation Pipeline

The lack of available high level protein coding and peptide annotations in the 17 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.

The two individual annotation sets (PseudoPipe and liftover of manually curated elements) are merged 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 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.

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-

Deleted: ● Top pseudogene families are matching closely the human counterparts .
1.
Formatted: Font:Not Bold
Formatted: Heading 3
Deleted: 1.
Formatted: No underline
Deleted: annoatations
Deleted: 2.
Formatted: Heading 4
Formatted: No underline

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.

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.

EXTRA

In particular, Spretus specific pseudogenes are enriched in apoptosis related genes and are characterized by the DEATH superfamily. This result is in concordance with the previous reports describing the strain specific tumor resistant phenotype as a result of the highly active apoptotic pathway and enrichment in tumor repressor genes \cite{19129501}. The blind albino mouse strain (BALB), a well studied line in a variety of neurodegenerative disorders (with 100% of subjects developing sever brain lesions \cite{JAX}), is characterized by pseudogenes associated with Cytochrome c Oxidase (COX) subunit VIa protein family. The phenotype link is particularly interesting given that COX mutations have been shown to cause neurodegeneration \cite{17435251}. Another example is the strain specific enrichment in defensin associated pseudogenes for the New Zealand obese mouse (NZO) – a mouse line known for expressing severe obesity phenotype. Defensins are small peptides involved in the organisms' protection against pathogens by regulating the in inflammatory defense against microbial invasion \cite{19855381} with recent studies highlighting the role played by defensin in controlling the inflammation resulted from metabolic abnormalities in obesity and type 2 diabetes \cite{25991648} and even showcasing it's potential as markers of obesity \cite{26929193}. A full list of pseudogene family related strain specific phenotypes is available in Supplemental Material.

Deleted: 3.

Formatted: No underline

Formatted: Heading 4

Deleted: [[CSDS be completed]]

Formatted: Heading 3

Formatted: Font:Not Bold

Deleted:

Deleted:

General considerations

Pseudogenes are DNA sequences that contain disabling mutations rendering them unable to produce a fully functional protein. There are different types of pseudogenes: processed pseudogenes – formed through a retrotransposition process, duplicated pseudogenes – formed during a gene duplication event, and unitary pseudogenes – formed by the inactivation of a functional gene. From a functional perspective, the pseudogenes can also be classified into three categories: dead-on-arrival – elements that are non functional and it is expected that in time they will 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.

In this paper we analyze the evolution and function of the pseudogene complement in the mouse lineage, with a particular focus on contrasting and comparing the unitary pseudogenes in human and mouse.