Introduction

The mouse is one of the most widely studied model organisms \cite{17173058}, with the field of mouse genetics counting more than a century of studies towards the understanding of mammalian physiology and development \cite{12586691,12702670}. The 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.

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 put a direct line between the two species, the make up of the present human population parallels, at generation divergence levels, the evolution of the recent *Mus Musculus* strains including the inbred laboratory mouse strains \cite{17284675} (Figure 1). 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 mouse strains would aid in deciphering genome evolution and diversity in human population.

Deleted: Moreover, despite

In this paper we describe the first pseudogene annotation and analysis of 17 widely-used inbred mouse strains alongside the reference mouse genome. These strains possess differences in their genetic makeup that manifest in an array of phenotypes, ranging from coat/eye color to predisposition for various disease \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, allowing each strain the potential to offer a unique reaction to an acquired mutation \cite{19710643}. Also the use of inbred mice minimizes a number of problems raised by the genetic variation between animals \cite{11528054}.

Moved (insertion) [1]

To uncover the key genome remodeling processes that governed the mouse strain evolution, we focus our analysis on the study of pseudogene complements, while also highlighting their key shared features with the human genome. Specifically, we provide the latest updates on the pseudogene annotation for both the human and mouse genomes, with a particular emphasis on the identification of unitary pseudogenes with respect to each organism.

Deleted: these organisms'

Deleted: mouse strain pseudogenes

Often regarded as genomic relics, pseudogenes provide an excellent perspective on genome evolution and function \cite{10692568,11160906,12034841,14616058}. Moreover, pseudogenes play an important role in functional analysis as they can be regarded as markers for loss and gain of function events. In recent years, loss of function (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. This is the case for the myosin gene (MYH16) pseudogenization that has been suggested to be

related to the acquisition of human-specific phenotypes in the primate linage \cite{25887751,16464126}. Another known 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 the PCSK9 result in lower plasma LDL levels, and thus a reduced risk of heart diseases.

From a functional genomics perspective there is a fine line between a loss of function that is increasing in a population and a pseudogene that is only partially fixed in that population, also known as polymorphic pseudogene. The 17 mouse strains are an excellent platform for revealing significant information about the occurrence and evolution of LOF events in a well understood model organism. The process that gives rise to these pseudogene types is especially interesting, because it has a great potential to tell us more about the organism's essential gene pool. Additionally, we are able to take advantage of the use of 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.

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 dead-on-arrival – these are elements that are non functional and it is expected that in time they will be eliminated from the genome, partially active – these are pseudogenes that exhibit residual biochemical activity, and exaptive pseudogenes – elements that 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 linage, with a particular focus on contrasting and comparing the unitary pseudogenes in human and mouse.

Results

1. Annotation

We present the latest annotation of the mouse reference genome as part of the GENCODE project, as well as updates on the human pseudogene reference set with a particular emphasis on unitary pseudogenes.

1.1 Reference genome

Using a rigorous manual annotation process as previously described in the GENCODE annotation resource \cite{22951037,25157146}, we identified almost 10000 pseudogenes in the mouse reference genome. Given the similarities between the human and mouse genomes, and the fact that in human we have identified over 14000 pseudogenes, the number of

Deleted: (

Deleted:)

Moved up [1]: 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, allowing each strain the potential to offer a unique reaction to an acquired mutation \cite{19710643}. Also the use of inbred mice minimizes a number of problems raised by the genetic variation between animals \cite{11528054}.

Deleted: The process that gives rise to these new pseudogenes is especially interesting, because it has a great potential to tell us more about the organism's essential gene pool and from a human perspective is of particular interest to the pharmaceutical industry. ... [1]

Deleted: Thus, the

Deleted: , and by extension shed light on human loss of function

manually annotated pseudogenes in the mouse lineage is likely to be an underestimate of the real size of the mouse pseudogene complement. Thus, taking advantage of the updated protein coding annotation, we used the in house annotation pipeline Pseudopipe \cite{16574694} to identify over 14000 pseudogenes in the mouse reference genome, of which more than 50% are shared with the manually curated set. PseudoPipe is a comprehensive pseudogene annotation pipeline focused on identifying and pseudogenes 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, Sup Table XXX).

In human we used a combination of automatic and manual curation to refine the reference pseudogene annotation to a set of 14650 [[CSDS +200 unitary]] pseudogenes. The updated set contains considerable improvements in the identification of unitary pseudogenes, as well as a better characterization of pseudogenes of previously unknown biotype (Sup Table XXX).

1.2 Mouse strains

The Mouse Genome Project sequenced and assembled genomes for 17 mouse strains, and developed a draft annotation of the strains' protein coding genes \cite{MousePaper}. The strains are organized into 3 classes: an outgroup – formed by two independent mouse 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 each strain genome composition is presented in \cite{MousePaper}.

We developed an annotation workflow for identifying pseudogenes in the 17 mouse strains, by leveraging our automatic pipeline PseudoPipe, as well as a set of manually curated pseudogenes from the mouse reference genome (GENCODE M8) lifted over onto each individual strain. Complementarily, the lift over of manual annotation expands the available biotypes by including inactivated immunoglobulin and polymorphic pseudogenes.

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 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 the out group species. The difference in the pseudogene complete size follows closely the variation in the number of conserved protein coding genes between each strain and the reference genome. Additionally, the pseudogene complements reflect the evolutionary distance between each strain and the reference genome.

However, the annotated pseudogenes are just a lower bound indication of the total number of pseudogenes in each strain, with the size of reference genome pseudogene complement curated by the automatic identification pipeline representing a low sensitivity upper bound. Thus we expect that the final number reflecting the true size of the pseudogene complement in the mouse lineage to be comparable to the number of pseudogene in human genome (e.g. ~14000).

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

Deleted: were able to annotate

Deleted: It is important to note that

Deleted: annotated

Deleted: men

Deleted: , with Pahari and Caroli having the lowest number of annotated pseudogenes. However, this is not a reflection of the total number of pseudogenes that are present in these two strains, but rather an indication of the lower number of conserved protein coding transcripts with respect to the reference mouse genome. We expect the number of pseudogenes in the outgroup species to increase with improvement in their respective protein coding annotations

Deleted: predictions

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

Deleted: predictions

The pseudogene biotype distribution closely follows the reference genome and is consistent with the biotype distributions observed in other mammalian genomes (e.g. Human \cite{22951037}, 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.

Deleted: predictions

Examining the pseudogene length distribution, we observed that on average pseudogenes are 782 bp long compared to an average size XXX for their parents, suggesting that sequence truncations were common during the pseudogene genesis process. We also identified a number of truncated pseudogenes in each of the strains by comparing the conservation of the 3' and 5' pseudogenic regions to their respective parent sequence.

The distribution of mouse pseudogene disablements follows closely 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. Also the proportion of pseudogene defects shows 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 of loss of function events and changes in selective pressures resulting in the fixation 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 thus require a large degree of attention during the annotation process.

In order to get an overview of the mouse strain unitary pseudogene complement, we lifted over the reference annotation and were able to identify on average 15 unitary pseudogenes in each strain. However, this value is an underestimate of the real number of unitary pseudogenes that we expect to find. 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 human-primates unitary annotation. Given the fact that in humans there are over 200 unitary pseudogenes with respect to primates, and the divergence scale between humans and primates matches that of the reference mouse and the outgroup species, we expect to see a similar number of unitary pseudogenes in the reference mouse (and lab strains) with respect to the outgroup.

For this 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, 210 unitary pseudogenes in mouse with respect to human and on average XXX unitary pseudogenes in each of the mouse classes with respect to the reference (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 chemosensory organ in mouse, reflecting the loss of functionality in these genes during the human linage evolution. We also observed the pseudogenization of a number of genes commonly related to the evolution of immune system in human. In particular we found 5 new pseudogenes associated 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 it's presence became futile during evolution. We also observed the pseudogenization of a leucine rich repeat protein, commonly 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.

[[CSDS]]

Deleted: to finish

2. Genome Evolution & Plasticity

Leveraging the pseudogene annotation, 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 about 3000 pseudogenes that are conserved across all strains (see Fig XXX). The 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 noticed 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 the 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, and NOD 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 (New Zealand obese mouse) and NOD (non-obese diabetic mouse) 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. Of these, we found almost 3,000 ancestral pseudogenes that are preserved across all strains. A detailed summary of the other pseudogene types 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 from shared evolutionary history. A pair-wise analysis of the 3 classes of strains (Fig XXX) shows that the laboratory strains share a larger number of pseudogenes with the outgroup species than with the wild strains, despite being evolutionarily closer to the latter. This anomaly is potentially related to the diversity of the 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 that are 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 lineage by looking at the enrichment of TE families in the two species on an evolutionary time scale lineage (Fig XXX). Both human and mouse 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 mRAN transcripts. We analysed the LINE, SINE and ERV repeat content in the human and mouse processed pseudogene complement. We define the evolutionary time scale by using the pseudogene sequence similarity to the parent gene as a proxy for age. Thus, younger pseudogenes are having a higher degree of sequence similarity to the parent, while older pseudogenes are showing a more diverged sequence.

As expected the human processed pseudogenes are defined by a single burst of retrotransposition events, that occurred 40MYA at the dawn of primate lineage \cite{}. By contrast in mouse we noticed that the three TE classes are associated with multiple successive bursts of pseudogene generation resulting in a continuous renewal of the processed pseudogene pool. This contrasting behavior can be accounted by the large difference in the number of active L1s between human and mouse (100 vs 3000s). This can be best observed by comparing the TE families in the three strain groups.

[[CSDS]]

2.4 Genome remodeling processes

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

Deleted:

Deleted: d

Deleted: TE families and

Deleted: genesis of pseudogenes

Deleted: (Fig XXX)

Deleted: distribution

Deleted: various

Deleted: as a function of age.

Deleted: We identified the latest

Deleted: as the main progenitor of processed pseudogenes in the human genome occurring

Deleted: with human, where the TEs became silent following the last event, the

Deleted: families exhibit

Deleted: retrotransposition

Deleted: TO ADD

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). Starting with the premise that a gene duplication can have two equally probable 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 pseudogene 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. Explicitly, a protein coding gene with a high expression level and low number of copies will yield a high number of mRNAs per copy, and thus will have a higher probability of producing a defective one. As a result, this protein coding gene will have a large number of associated processed pseudogenes. By contrast a highly expressed protein coding gene with multiple copies, will have a lower number of mRNA products per copy, and thus we expect a lower incidence of associated processed pseudogenes. 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: By contrast, there is no such expectation when the pseudogenes are the result of retrotransposition. As such, similar to the human counterpart, the mouse pseudogene complement exhibits an independent

3. Biological relevance

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

Deleted: Commonly

Deleted: have

Deleted: thought to present no biological significance. However, our

Deleted: highlight a large spectrum of biochemical activity associated with

Deleted: .

Deleted: characterize

Deleted: relevance

3.1 Gene ontology & pseudogene family analysis

We integrated the pseudogene annotation with gene ontology (GO) data 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, hinting at the shared evolutionary history between the various mouse strains (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 mouse pseudogenes top

Deleted: . However, we also identified a number of strain specific processes that relate to strain specific phenotypes (Table XXX, Fig XXX). ... [2]

families are related to highly expressed proteins such as GAPDH, Ribosomal proteins and Zinc fingers.

However, a closer look suggests that pseudogene repertoire also reflects the 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 of Spretus, where we see an enrichment of pseudogenes related to tumor repressor genes and apoptosis pathways genes [[CSDS2PM: can you please check the list of these pseudogenes form the DEATH clan so we can get their parent expression in Spretus and comment on expression pattern for those genes]]. Second, we find pseudogenes reflecting a deleterious phenotype. For example, this is the case of the blind albino mouse strain (BALB), a representative line for neurodegenerative disorders (100% of subjects developing sever brain lesions \cite{JAX}). BALB is enriched in with Cytochrome c Oxidase (COX) subunit VIa pseudogenes, and it has been previously reported that disabling mutations in COX are cause for neurodegeneration \cite{17435251}.

Deleted: with their associated

3.2 Gene essentiality

[[CSDS2PM can you please add a bit more intro to gene essentiality?]]

We observed an enrichment of essential genes among pseudogene parent genes across all mouse strains. 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 with 3,263 essential genes. Evaluating the parent gene for each pseudogene present in the mouse strains reveals essential genes are approximately three times more abundant amongst parent genes. 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: also

[[CSDS2PM TO extend]]

— do essential genes have duplicated copies (in order to guarantee the organism fitness and survival) or are they unique thus, their disabling resulting in the organism's death

— are the pseudogenes associated with essential genes because the are duplicated for conservation reason or for expression reasons?

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 mouse and human genomes, with 15% of the total pseudogenes showing a residual level of transcription at a FPKM greater than 2. By contrast, tissue specific transcription varies largely in both human and mouse from 2% to 6% in liver and testis respectively. In this project we focus our analysis on single tissue pseudogene expression. In brain on average 5% of pseudogenes show evidence of transcription. This result is consistent with the fraction of pseudogenes that are transcriptionally active in human brain (see Sup Fig XX). We also identified xxx% transcribed pseudogenes that show a discordant expression pattern with respect to their parent genes.

Deleted: biological activity

Deleted: a

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 the one observed across the entire dataset. Moreover, for strain specific pseudogenes, the fraction of transcribed elements varies across the strains (see Sup Fig XX).

Next we looked at transcription in protein coding genes since their RNA transcript is the check point in both protein expression and processed pseudogene formation through retrotransposition. We observed that genes associated with pseudogenes show a consistently a higher level of transcription compared to their non pseudogenes related counterparts (see Sup Fig XX). However, no significant correlation was observed between the levels of transcription in pseudogenes and their corresponding functional homolog.

When evaluating pseudogene transcription across both the laboratory and wild strains the 393 pseudogenes with transcription in all assayed strains was lower than the number of strain specific transcribed pseudogenes for all but one strain. This contrasts with pseudogene conservation in which case the number of shared pseudogenes is greater than that of all laboratory strains and two wild strains (CAST and WSB). However, when shared pseudogene transcription is evaluated within the context of either the laboratory strains or wild strains slightly different patterns emerge. Amongst the wild strains strain specific transcription is greater than cross strain transcription for each strain. Within the laboratory strains the number pseudogenes with cross-strain transcription was greater than the number of pseudogenes with strain specific expression for 4 of the 10 strains.

To investigate the pseudogene formation milestones, we analyzed the expression of protein coding genes during mouse fetal development. We observed a clear relationship between the expression level of parent genes and the number of associated processed pseudogenes. In particular …. [[PM can you please fill in the analysis here.]]

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). 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. 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. 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 complement in the mouse reference genome and 17 related strains. The surveyed set was created employing both manual curation and computational pipelines and consists of 9379 pseudogenes in the reference genome. Given the similarities between the human and mouse genomes we expect that the total number of pseudogenes in the mouse genome to match closely the one reported for human (currently standing at 14650).

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. As such the size of each reported pseudogene complement set decreases slightly with the increase in the evolutionary distance between the strain and the reference genome. However, we found that the relative ratio of processed to duplicated pseudogenes is preserved across all strains.

Integrating the annotation in the 18 strains we obtained a pan genome mouse pseudogene set composed of over 45000 unique entries. The pan genome set contains three types of pseudogenes: conserved, strain specific, and multi-strain, accounting for 6, XX, and respectively YY% of the elements. By comparing the pseudogene complements across the 6 million years of evolution we obtain a global picture of genome remodeling processes that shaped the mouse lineage. As such, we constructed phylogenetic trees using pseudogene sequences and we were able to associate strain phenotypes with strain specific pseudogenes.

Sequence analysis reveals that while the majority of human genome pseudogenes have been obtained relatively recently through a single burst of retrotransposition, the mouse lineage shows a continued renewing of the pseudogene pool through the constant activity of transposable elements.

- Top pseudogene families are matching closely the human counterparts
- While human TE activity became silent after the retrotransposition burst, TE are still active in mouse strains
- Similar to human, pseudogene prolific genes are not enriched in paralogs and vice versa

- Pseudogene localisation suggests multiple large scale genomic rearrangements between the out group - wild strains and the reference (lab strains) mouse genome
- A significant proportion of pseudogenes show signs of transcriptional activity

Methods

1. 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 annoatations (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.

Deleted: predictions

Deleted: predictions

2. 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.

3. Data integration & pangenome pseudogene generation

[[CSDS be completed]]

EXTRA

Deleted: . ... [3]

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.

Page 2: [1] Deleted	Sisu, Cristina	12/03/2017 19:56

The process that gives rise to these new pseudogenes is especially interesting, because it has a great potential to tell us more about the organism's essential gene pool and from a human perspective is of particular interest to the pharmaceutical industry.

The creation of all the mouse lines analyzed in this study has been extensively documented.

Page 7: [2] Deleted	Sisu, Cristina	12/03/2017 19:56

. However, we also identified a number of strain specific processes that relate to strain specific phenotypes (Table XXX, Fig XXX).

[[CSDS2PM: can you please look at the universal GO terms that define all pgenes and also at the the strain specific terms and add them to a table]]

Page 11: [3] Deleted	Sisu, Cristina	12/03/2017 19:56