Introduction

The mouse is one of the most widely studied model organisms \cite{}, with the field of mouse genetics counting more than a century of studies towards the understanding of mammalian physiology and development \cite{}. The recent advances of the Mouse Genome Project in 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.

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. Understanding the genesis and functional impact of the genetic makeup of these mouse strains would set the tone in deciphering the genome evolution and diversity in human population.

Wtf FUCK!!

In this paper we described the pseudogene annotation and analysis of 17 key mouse strains alongside the reference mouse genome. The strains display a variety of phenotypes, ranging from coat/eye color, to differences in their genetic makeup \cite{}. To uncover the key genome remodeling processes that governed these organisms' evolution, we focus our analysis on the study of mouse strain pseudogene complements, while also underlining their key shared features with the human genome. Specifically, we highlight the latest updates on the pseudogene annotation in both human and mouse 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 viewpoint on genome evolution and function. Moreover, pseudogenes play key roles in functional analysis as they can be regarded as markers for loss and gain of function events. In recent years, loss of function (LOF) has become one of the hot topics in genomics. In general, the loss of gene function would be detrimental to the organism's fitness levels, however, sometimes, in the right conditions, the removal of a protein through an inactivation process (pseudogenization) of its gene, can also be advantageous. The relaxation of the selection constraints on that gene would favor the accumulation of disabling mutations, eventually resulting in the pseudogene fixation in that organism. This is the case in the myosin gene (MYH16) pseudogenization that has been suggested to be related to the acquisition of human-specific phenotypes in the primate linage\cite{}.

From a functional genomics perspective there is a fine line that defines a loss of function that is increasing in a population from a pseudogene that is only partially fixed in that population, also known as polymorphic pseudogene. The process that gives rise to these new pseudogenes is particularly 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 homozygous mouse strains can reveal significant information about the apparition and evolution of LOF events, thus also shedding light on the human LOF events. The large and detailed records available regarding the creation of each of the present laboratory mouse strains as well as the vast array of genomic data, make the 17 strains an excellent platform for LOF study.

MORE

Results

Deleted: with

Deleted: . Particularly

Deleted: genomes

Deleted: focused

Deleted: . ... [1]

Deleted: connecting to

Deleted: within

Deleted: complement

Deleted:

Deleted: view point

Deleted: the

Deleted: trending

Deleted: both functional and evolutionary

Deleted: an intact

Deleted: it's

Deleted: on

Deleted:

Deleted: of

Deleted: acquiring

Deleted: human

Deleted:

Deleted: of LOF events

Deleted: of telling

Deleted: , that,

Deleted: also very useful for

Deleted:

Deleted: completely

Deleted: , have a great potential for revealing a lot

Deleted: and by comparison, shed

Deleted: in human genome. Thus given

Deleted: mouse strains' highly structured

Deleted: mechanism and

Deleted: available

Deleted: it makes them

1. Annotation

Deleted:

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.

Deleted: In this study we

Deleted: GENOCDE

Deleted:

1.1 Reference genome

Taking advantage of the updated protein coding annotation from Ensembl 87, we used the in house annotation pipeline Pseudopipe to identify over 22000 pseudogenes in the mouse reference genome, of which 14095 are present in the assembled chromosomes, while the remaining annotations are identified in scaffold and DNA patch sequences. More than half of the assembled genome annotations are processed pseudogenes, with a smaller fraction of duplicated pseudogenes (Tables XXX). The in house pipeline annotations share a 86% overlap with the manually annotated set \cite{}.

METH

Furthermore, using a combination of automatic and manual curation we refined the human 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 (Tables XXX).

Deleted:

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 a pseudogene annotation workflow leveraging our in house automatic annotation pipeline PseudoPipe \cite{}, as well as the lift over set of manually curated pseudogenes from the mouse reference genome (GENCODE M8) to each individual strain. PseudoPipe is a comprehensive pseudogene annotation pipeline focused on identifying three pseudogene biotypes: processed, duplicated, and unitary. Complementarily, the lift over of manual annotation expands the available biotypes by including inactivated immunoglobulin and polymorphic pseudogenes.

Each identified element provides details with respect to the pseudogene transcript biotype, genomic location and structure and is characterized by a confidence level reflecting the annotation process.

A detailed summary of the number of pseudogenes, their confidence levels and related biotypes is shown in figure XX (Sup Table XX). On average we were able to annotate 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. It is important to note that the annotated pseudogene complement size follows the evolutionary distance between each strain and the reference genome, 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 regarding the conserved number of protein coding transcripts with respect to the reference mouse genome. We expect their numbers to increase with the improvement in their respective protein coding annotations.

Deleted: *Mus*

Deleted: *Mus*

Deleted:

Currently, around 30% of pseudogenes in each strain are defined as high level predictions (Level 1), 10% Level 2, and 60% Level 3. With improvements in both the annotation of the mouse reference genome as well as refinement of the strain assemblies and annotation, we expect that the number of high confidence predictions will increase, matching the fraction observed in the human genome.

Deleted:

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{}, chimp \cite{}, macaque \cite{}). As such, the bulk (~XX%) of the predictions are processed pseudogenes, while a smaller fraction (~XX%) are duplicated pseudogenes. A small fraction of pseudogenes requires further analysis of their formation mechanism in order to be assigned the correct biotype.

Moreover, examining the pseudogene length distribution we observed that on average pseudogenes are 782 bp long compared to the average size of their parents of XXX 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 mouse pseudogene disablements distribution follows closely the previously observed distribution in the reference genome as well as in 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 pseudogene are the result of a complex interplay of loss of function events and changes in the selection pressure resulting in the fixation in a species of an inactive element. 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.

Deleted: in

Deleted:

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 annotating them.

Deleted: larger

Deleted: when

Deleted: a fast

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 a gross 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 unitary pseudogene complement in the mouse strains is to look at the human-primates unitary annotation. Given the fact that in human there are over 200 unitary pseudogenes with respect to primates, and the divergence scale between human 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.

Deleted:

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 leucine reach repeat protein, commonly related to the evolution of the immune system in primates \cite{}. By contrast

Deleted: give

Deleted: related

Deleted: GPCR coupled receptors

Deleted: that lost their

Deleted: identified

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 to add mouse strain Unitary example related to variation in phenotypes]]

2. Evolution

Leveraging the pseudogene annotation, we explore the differences between the 17 mouse strain by looking at the evolutionary history of their pseudogene complements.

2.1 Phylogeny

It has long been held that pseudogenes evolve with little or no selective constraints at all, and that the mutation rate in pseudogenes reflects the underlying genome substitution pattern, 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 conserved across all strains (see Fig XXX). The pseudo-based tree correctly identifies and clusters the strains into three classes: outgroup, wild, and laboratory strains.

Next we grouped the ancestral 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. By comparing the resulting trees to the protein-coding one, we noticed that some display an independent, strain specific evolutionary pattern. The deviations of the pseudogene trees from the known lineage pattern reflect key roles played by pseudogenes during the strains evolution.

For example, the olfactory receptor 987 pseudogene tree, while maintaining the Mus 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 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 ancestral gene in Spretus and PWK wild strains. The rest of the strains, including Caroli and Castaneus, 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 link observed between obesity and olfactory receptor regulation \cite{}, given the fact that the two strains display a common obesity 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 the 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 outgroup, wild strains, or 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 TE

[[CSDS TO ADD]]

Deleted: [[CSDS to add intro]]

Deleted: constraint

Deleted: out group

Deleted:

Deleted: .), and

Deleted: individual pseudo-

Deleted: systems. Comparing

Deleted: sub-systems

Deleted: and

Deleted: deviation

Deleted: individual

Deleted: reflects the

Deleted: the

Deleted: in their respective strain

Deleted: , with striking divergences

Deleted: in the

Deleted: indicate

Deleted: to

Deleted: change

Deleted: differences between

Deleted:

- Use repeat masker datasets for reference + the download data for strains
- divide the the elements in LINE, SINE, DNA and LTRs
- distribution of TE in pseudogenes,
- map strain evolution in the burst of TE
  - *group by strain*
  - *group by pseudogene conservation:*
    - *Ancient - conserved group (3K)*
    - *Conserved 1-1 outgroup (254)*
    - *Conserved 1-1 wild strain (10)*
    - *Conserved 1-1 lab strain (63)*
- compare with human families and test for shared with primates TE fam
- —> refer to the TE analysis in main paper
- while TE families got silenced in humans and primates after the last retrotransposition burst, TE are still active in Mouse resulting in multiple pseudogene genesis bursts
- TE plots conservation

Deleted:

Deleted:

Deleted:

Deleted:

Deleted:

Deleted:

Deleted:

3. Genome plasticity

Deleted:

[[CSDS to add intro]]

3.1 Genome remodeling processes

The large proportion of strain and class specific pseudogenes, as well as the presence of active transposable elements families, points 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 Mus Caroli loci were conserved in the reference genome, while Mus 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).

3.2 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.

Deleted:

In order to understand the distribution of successful to disabled copies of genes, we compared the number of pseudogenes with the number of functional paralogs for each parent gene for both mechanism of formation (retrotransposition and duplication) (Fig XXX). Starting at the premise that a gene duplication can have two similarly 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 significantly towards the creation of functional elements. 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 evolution of the number of processed pseudogenes relative to the number of paralogs per gene.

Deleted:

[[CSDS point to transcription and plot the highly transcribed parents vs paralogs vs pgenes]]

4. Biological relevance

[[CSDS to add intro]]

4.1 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 relevance? 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. However, we also identified a number of strain specific processes that relate to strain specific phenotypes (Table XXXX).

[[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]]

Deleted:

Moreover, the GO terms that universally characterize the pseudogene complements in all the mouse strains are closely reproduced in the family and clan 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 families are related to highly expressed proteins such as GAPDH, Ribosomal proteins and Zinc fingers.

Deleted:

However, a closer look suggests that pseudogene repertoire also reflects the individual strain specific phenotypes. For example, Mus Spretus specific pseudogenes are characterized at a strain specific level by the protein clan DEATH that reflects the strain enrichment in apoptosis genes and explains the previously observed peculiar tumor resistant phenotype (http://www.pnas.org/content/106/3/859.full).

Deleted:

MORE

4.2 Gene essentiality

We also 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 (http://www.nature.com/nature/journal/v537/n7621/full/nature19356.html). 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.

[[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

Deleted:

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

4.3 Pseudogene Transcription

We leveraged the available RNA-seq data from the Mouse Genome Project to study the pseudogene biological activity as reflected by their transcription potential. In brain samples, overall, 25% of pseudogenes exhibited some residual level of transcription, with ~5% of them having an FPKM greater than 2. Moreover, 1% of transcribed pseudogenes show a tissue specific expression pattern. We also identified xxx% transcribed pseudogenes that show a discordant expression pattern with respect to their parent genes. Similar to the previous observed pattern in human and other model organisms, the pseudogene transcription in mouse strains shows a 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.

Consistent

Deleted:

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.

5. Mouse pseudogene resource

Deleted: [[CSDS to be completed]] ... [2]

We created a pseudogene resource that organizes all of the pseudogenes across the 17 mouse strains and reference genome, and associated phenotypic information in a MySQL database. 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.

HUM.

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

Deleted:

[[CSDS TO DO]]

- Completed the first draft of pseudogene annotation in 18 mouse strains
- On average 20% of pseudogenes are strain specific and 20% are ancestral pseudogenes, being conserved in all the strains
- 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

Deleted:

1. Pseudogene Annotation Pipeline

Deleted:

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 predictions (using 1 bp minimum overlap) and extending the predicted boundaries to ensure the full annotation of the pseudogene transcript. As such, Level 1 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 predictions, while Level 3 pseudogenes are predicted solely using the automation identification pipeline.

2. Unitary Pseudogene Annotation Pipeline

Deleted:

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

Deleted:

[[CSDS be completed]]

Page 1: [1] Deleted **Sisu, Cristina** **2/4/17 3:51:00 AM**

.

In this paper we focus on the annotation and comparative analysis of pseudogenes in the available mouse strains

Page 7: [2] Deleted **Sisu, Cristina** **2/4/17 3:51:00 AM**

[[CSDS to be completed]]