Introduction

Mouse is one of the most widely used model organisms \cite{}, with the field of mouse genetics counting for more than a century of studies towards the understanding of mammalian physiology and development \cite{}. The recent advancements 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 mammalians.

Despite obvious discrepancies between the human and mouse: e.g mice are small, with short life span and high metabolic rate, the two species share a large number of similarities in their genetic makeup, and in particular in tumor and disease development, making mice ideal model organisms for the study of human diseases. Understanding the genesis and impact of the genetic makeup of the mouse strains would set the tone in deciphering the genome evolution and diversity in human population.

In this paper we described the pseudogene annotation and analysis of 17 key mouse genomes 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 the organisms' evolution, we directed our analysis on the study of mouse strain pseudogene complements.

Often regarded as genomic relics, the pseudogenes provide an excellent view point on the genome evolution and function. Moreover, the pseudogenes play key roles in functional analysis as they can be regarded as markers for loss and gain of function events. In recent years, the loss of function (LOF) has become one of the trending topics in both functional and evolutionary genomics. Given the strains' creation mechanism and the vast array of available genomic data, it makes them an excellent platform for LOF study.

In this paper we focus on the annotation and comparative analysis of pseudogenes in the available mouse strains, pin-pointing to key shared features within the human complement.

The Mouse Genome Project sequenced and assembled genomes for 17 mouse strains, as well as providing a draft annotation of the strains' protein coding genes \cite{MousePaper}. The strains are organized into 3 classes: the outgroup – formed by two independent mouse species, *Mus Caroli* and *Mus Pahari*; the 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 the laboratory strains. A detailed summary of each strain genome composition is presented in \cite{MousePaper}.

Results

1. Annotation and Summary

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 focusing on identifying three

pseudogene biotypes: processed, duplicated, and unitary. Complementary, the lift over of manual annotation expands the available biotype by including inactivated immunoglobulin and polymorphic pseudogenes.

Each prediction gives information 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 12000 pseudogenes in each laboratory strain, over 11000 pseudogenes in each of the wild strains, and just over 10000 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 *Mus Pahari* and *Mus Caroli* having the lowest number of annotated pseudogenes. This however is not a reflection of the total number of pseudogenes that are presented in these two strains, as we expect their numbers to increase with the improvement in their respective protein coding annotation, but rather an indication regarding the conserved number of protein coding transcripts with respect to the reference mouse genome.

Currently, around 30% of pseudogenes in each strain are defined as high level predictions (Level 1), 10% Level 2, and 60% Level 3. With the improvement 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 to increase, matching the fraction observed in the human genome.

The pseudogene biotype distribution follows closely the reference genome being consistent with the biotype distributions observed in other mammalian genomes (e.g. Human \cite{}, chimp \cite{}, macaque \cite{}). As such, the bulk (~XX%) of predictions are processed, while a smaller fraction (~XX%) are duplicated pseudogenes. A small fraction of pseudogenes requires further analyses in their formation mechanism in order to be assigned the correct biotype.

Moreover, examining the distribution of the pseudogene length we observed that on average the 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 by comparison to the paternal 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.

While identifying pseudogene based on their formation mechanism is relatively straight forward, a larger degree of attention has to be given in annotating unitary pseudogenes. The importance of unitary pseudogenes resides in their ability to mark off function events.

On average we annotated about 15 unitary pseudogenes in each strain by liftover from the reference genome and XXX unitary pseudogenes using the in house pipeline across human and mouse reference genome. 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 the one from 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.

[[CSDS to add Human - mOuse Unitary + examples + LOF]]

2. Evolution

2.1 Phylogeny

It has long been held that pseudogenes evolve with little or no selective constraint at all, the mutation rate in pseudogenes reflects the underlying genome substitution pattern, making them ideal elements for inferring and comparing the mutational process across the mouse strains. To this end we build a phylogenetic tree based on about 3000 conserved pseudogenes across all the strains (see Fig XXX). The pseudo-based tree correctly identifies and clusters the strains in the three classes: out group, wild, and laboratory strains. Next we selected at random a number of pseudogenes and constructed individual pseudo-trees. The deviation of the individual trees from the known lineage pattern reflects the key roles played by the pseudogenes in their respective strain evolution.

For example, Olfactory receptor 987 pseudogene tree, while maintaining the Mus Pahari as an outgroup species, presents a completely different evolutionary history for the 17 strains, striking divergences being observed in 129S1, NZO and NOD laboratory strains, and smaller in the Spretus and PWK wild strains. The rest of the strains including Caroli and Castaneus indicate little to no change at all compared to the common ancestor. The observed differences between the 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 mouse strains we created a pangenome pseudogene dataset containing 49262 unique entries relating the pseudogenes across strains. Of these, we found almost 3000 ancestral pseudogenes that are preserved across all the strains. A detailed summary of the other pseudogenes types is shown in table XXX. On average each strain boasts 3000 strain specific pseudogenes. The proportion of pseudogenes conserved only in outgroup, wild strains or lab strains is considerably smaller, suggesting that the bulk of pseudogenes in each strain is formed through 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

them being evolutionarily closer to the latter. This anomaly can be potentially related to the diversity of the mouse wild strain 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]]

- 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

3. Genome plasticity

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 genomics rearrangements in the mouse genome evolution. To this end we examined the conservation of the 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).

[[CSDS TO ADD examples & discussion of the observed jumps]]

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 pseudogene in 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 to failed copy genes, we compared the number of pseudogenes with the number of functional paralogs for each parent gene. Similar to the human counterpart, the mouse pseudogene complement exhibits an anti correlation between the number of processed pseudogenes and the number of paralogs per gene. By contrast, when pseudogenes are created through a duplication process, there is a more uniform distribution of functional to non functional elements per parent gene, with the ratio however being tilted significantly towards the creation of functional elements.

4. Biological relevance

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 the pseudogenes: what is their biological relevance? For this we calculated the enrichment of the 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 murine strains. However, we also identified a number of strain specific processes that relate to the strain specific phenotypes.

Moreover the GO term enrichment was closely reproduced in the family and clan classification of pseudogenes. As expected, matching the human and primate counterparts, murine pseudogenes top families are GAPDH, 7-Transmembrane proteins, Ribosomal proteins, Ribosomal receptor proteins and Zinc finger. However a closer look at the pseudogene clan distribution highlight a number of strain specific pseudogenes. For examples, Mus Spretus pseudogenes are clustered in a strain specific DEATH clan that reflects the strain enrichment in apopotosis genes and explain the previously observed peculiar tumor resistant phenotype (http://www.pnas.org/content/106/3/859.full).

4.3 Gene essentiality

An enrichment of essential genes among pseudogene parent genes was observed 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 contained 4,736 genes compared with 3,263 essential genes. Evaluating the parent genes 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.

4.4 Pseudogene Transcription

[[CSDS + PM TO ADD]]

5. Mouse pseudogene resource

We created a pseudogene resource that organizes all of the pseudogenes across the 17 mouse strains and associated phenotypic information in a MySQL database. Each pseudogene is given both 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 mouse strains.

Pseudogene annotation information encompasses the genomic context of each pseudogene, its parent gene and transcript IDs, 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 enables 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 their 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 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

[[CSDS TO ADD]]

- 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

1. Pseudogene Annotation Pipeline

The lack of available high level protein coding and peptide annotation in the 17 mouse strains, created a bottleneck in the pseudogenes identification process, that was by-passed by generating protein input sets that are shared between the strain and the reference genome. The summary of shared transcripts follows the evolutionary trend with more distant strains having a smaller number of common protein coding genes with the reference genome compared with more recent laboratory strains.

The two individual annotation sets (PseudoPipe and Lift-over) are merged to produce the final pseudogene complement set. The merge 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 level prediction, with the annotated pseudogene being validated by both automatic and manual curation processes, Level 2 pseudogenes are characterized 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

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

3. Data integration & pangenome pseudogene generation

[[CSDS TO ADD]]