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

Cristina Sisu*[1,2,3], Paul Muir*[4,5], Adam Frankish[6], Ian Fiddes[7], Mark Diekhans[7], David Thybert[6,8] Duncan T. Odom[9,10], Paul Flicek[6,10], Thomas Keane[6], Tim Hubbard[11], Jennifer Harrow[12], Mark Gerstein[1,2,5,13]

1 Program in Computational Biology and Bioinformatics, Yale University, New Haven, CT 06520, USA;
2 Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, CT 06520, USA;
3 Department of Biosciences, Brunel University London, London UB8 3PH, UK;
4 Department of Molecular, Cellular & Developmental Biology, Yale University, New Haven, CT, 06520, USA.
5 Systems Biology Institute, Yale University, West Haven, CT, 06516, USA.
6 European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, Cambridge, CB10 1SD, UK;
7 Center for Biomolecular Science and Engineering, University of California, Santa Cruz, CA 95064, USA;
8 Earlham Institute, Norwich research Park, Norwich, NR4 7UH, UK;
9 University of Cambridge, Cancer Research UK Cambridge Institute, Robinson Way, Cambridge CB2 0RE, UK;
10 Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SA, UK;
11 Department of Medical & Molecular Genetics, King's College London, London SE1 9RT, UK;
12 Illumina Cambridge Ltd, Chesterford Research Park, Little Chesterford, Saffron Walden CB10 1XL, UK.
13 Department of Computer Science, Yale University, New Haven, CT 06511, USA.
* These two authors contributed equally.

Pseudogenes are ideal markers of genome remodeling. In turn, the mouse is an ideal platform for studying them, particularly with the availability of the sequencing data from 18 strains (completed by the Mouse Genome Project). Here we present a new online pseudogene resource focused on the analysis of the mouse reference genome and 18 associated strains. We performed a comprehensive genome-wide annotation of the pseudogenes in the mouse reference genome and associated strains by combining manual curation of over 10,000 pseudogenes with results from automatic annotation pipelines. By comparing the human and mouse, we annotated 271 unitary pseudogenes in mouse, and 431 unitaries in human. The overall mouse pseudogene repertoire (in the reference and strains) is similar to human in terms of overall size, biotype distribution (~80% processed/~20% duplicated) and top family composition (with many GAPDH and ribosomal pseudogenes). However, notable differences arise in the pseudogene age distribution, with multiple retro-transpositional bursts in mouse evolutionary history and only one in human. Furthermore, in each strain ~20% of the pseudogenes are unique, reflecting strain-specific functions and evolution. Additionally, we find that ~15% of the pseudogenes are transcribed, a fraction similar to that for human and that pseudogene transcription exhibits greater tissue and strain specificity compared to their protein coding counterparts. Finally, we show that processed pseudogenes tend to be derived from highly transcribed parent genes.

Deleted: transcriptional data during development (just completed in phase 3 of ENCODE) and

Deleted: of

Deleted: . We compiled this from

Deleted: Also, by

Deleted: We make all our annotation available through mouse.pseudogene.org.

Deleted: ; for instance, the differences observed in the evolution of taste receptors associated pseudogenes in the NZO mice can be related to their change in the diet.

Introduction

The mouse is one of the most widely studied model organisms [1], with the field of mouse genetics counting for more than a century of studies towards the understanding of mammalian physiology and development [2, 3]. Recent advances of the Mouse Genome Project [4, 5] towards completing the de-novo assembly and gene annotation of a variety of mouse strains, provide a unique opportunity to get an in-depth picture of the evolution and variation of these closely related mammalian organisms.

Mice have frequently been used as a model organism for the study of human diseases due to their experimental tractability and similarities in their genetic makeup [6]. This has been achieved through the development of mouse models of specific diseases and the creation of knockout mice to recapitulate the phenotype associated with a loss-of-function mutation observed in humans. The advent of high throughput sequencing has led to the emergence of population and comparative genomics as new windows into the relationship between genotype and phenotype amongst the human population. Current efforts to catalog genetic variation amongst closely related mouse strains extend this paradigm.

Deleted:

Deleted:

Since their divergence around 90 million years ago (MYA) [7, 8, 9, 10, 11, 12], the human and mouse lineages followed a parallel evolutionary pattern [13]. While it is hard to make a direct comparison between the two species, there is a large range of divergence in the mouse lineage, with some approaching human-chimp divergence levels in terms of the number of intervening generations [13] (**Figure 1A**). The mouse strains under investigation have differences in their genetic makeup that manifest in an array of phenotypes, ranging from coat/eye color to predisposition for various diseases [5]. Moreover, the creation of these strains has been extensively documented [14]. Following a well characterized inbreeding process for at least 20 sequential generations, the inbred mice are homozygous at nearly all loci and show a high level of consistency at genomic and phenotypic levels [15]. This helps minimize a number of problems raised by the genetic variation between research animals [16]. The repeated inbreeding has also resulted in substantial differences between the mouse strains, giving each strain the potential to offer a unique reaction to an acquired mutation [17].

Deleted: population

Deleted:

To uncover key genome remodeling processes that governed mouse strain evolution, we focus our analysis on the study of the pseudogene complements of each strain, while also highlighting their shared features with the human genome. In this resource paper, we describe the first pseudogene annotation and analysis of 18 widely-used inbred mouse strains alongside the reference mouse genome. Additionally, we provide the latest updates on the pseudogene annotation for both the mouse and human reference genomes, with a particular emphasis on the identification of unitary pseudogenes with respect to each organism.

Often regarded as genomic relics, pseudogenes provide an excellent perspective on genome evolution [18, 19, 20, 21]. Pseudogenes are DNA sequences that contain disabling mutations rendering them unable to produce a fully functional protein. Different classes of pseudogenes are distinguished based on their creation mechanism: processed pseudogenes – formed through a retrotransposition process, duplicated pseudogenes – formed through a gene duplication event and subsequent disablement of one of the duplicates, and unitary pseudogenes – formed when functional genes acquire disabling mutations resulting in the inactivation of the original coding loci. Unitary pseudogenes are also characterized by the presence of a functional gene. Additionally, pseudogenes that are present in a population as both functional and nonfunctional alleles are termed polymorphic [22]. Such pseudogenes represent disablements that have occurred on a much more recent timescale. They are loss-of-function (LOF) mutations that are not fixed in the population and still subject to evolutionary pressures [22]. From a functional perspective, pseudogenes can be classified into three categories: dead-on-arrival – elements that are nonfunctional and are expected, in time, to be eliminated from the genome, partially active –

Deleted: and

Deleted: . There is also a third class of

Deleted: , called unitary. These pseudogenes are

Deleted: by the

Deleted: on the same locus in other species.

Deleted: pseudogenes

Deleted: human

pseudogenes that exhibit residual biochemical activity, and exapted pseudogenes – elements that have acquired new functions and can interfere with the regulation and activity of protein coding genes.

Moreover, pseudogenes reflect changes in selective pressures and genome remodeling forces. Duplicated pseudogenes can reveal the history of gene duplication, one of the key mechanisms for establishing new gene functions [23]. While the majority of the duplicated gene copies are eventually pseudogenized [24], successfully retained paralogs can acquire new functions [25], a process known as neofunctionalization [26]. Furthermore, duplicated pseudogenes can help explore the role of gene dosage in the inactivation or preservation of duplicate genes [27, 28]. Processed pseudogenes inform on the evolution of gene expression as well as the history of transposable element activity, while unitary pseudogenes are indicative of gene families that died out. Thus, pseudogenes can play an important role in evolutionary analysis as they can be regarded as markers of LOF events.

A loss-of-function event is a mutation that results in a modified gene product that lacks the molecular function of the ancestral gene [29]. Unitary pseudogenes are an extreme case of LOF, where mutations that result in complete inactivation of a gene are fixed in the population. In recent years, LOF mutations have become a key research topic in genomics. In general, loss of a functional gene is detrimental to an organism's fitness. However, there are numerous examples showcasing evolutionary advantages for the accumulation and fixation of LOF mutations resulting in the formation of new pseudogenes. For example, the pseudogenization of proprotein convertase subtilisin/kexin type 9 (PCSK9) in human evolution is commonly associated with a reduced risk of heart disease by lowering the plasma low-density lipoprotein (LDL) levels. This is achieved by preventing the expression PCSK9 protein and its subsequent binding to and degradation of cellular LDL receptors [30]. By contrast, gain of function mutations resulting in the expression of PCSK9 are commonly associated with an enrichment in plasma LDL cholesterol and an increased risk of atherosclerosis for the affected individuals [31]. This finding has inspired the creation of PCSK9 inhibitors as treatment for high cholesterol, and highlights the potential for the investigation of pseudogenes to shed light on biological processes of interest to the biomedical and pharmaceutical industry [32].

Taken together the well-defined evolutionary relationships between the mouse strains and the wealth of associated functional data from the ENCODE project present an opportunity to investigate the processes underlying pseudogene biogenesis and activity to an extent previously not possible. Leveraging mouse developmental time-course RNAseq data, we explore whether pseudogene creation occurs primarily in the gametes or earlier in development in a germline precursor. Also, comparison to the primate lineage and human population is a possibility as the evolutionary distance between some of the mouse strains parallels the human-chimp divergence as well as distances between the modern day human populations in terms of generations, making the collection of high quality genomes and associated pseudogene annotations for the 18 strains a valuable resource for both population studies and the broader mouse genetics research community.

Results

1. Annotation

We present the latest pseudogene annotations for the mouse reference genome as part of the GENCODE project, as well as updates on the human pseudogene reference set. Leveraging the recently assembled high-quality genome sequences of 18 mouse strains, we introduce the first draft annotation of the pseudogene complement in these genomes.

1.1 Reference genome

Deleted:

Deleted: by acquiring loss of function mutations that became fixed in the population.

Deleted: for loss

Deleted: function

Deleted:

Deleted: (LOF)

Deleted: s

Deleted:

Deleted:

Deleted: recently completed

Deleted: 3

Deleted: for the

Deleted: the 18 strains

Using a combination of rigorous manual curation [33, 34] and automatic identification [35] we were able to annotate a comprehensive set of pseudogenes for the mouse reference genome (**Table** S1A&B). However, pseudogene assignments are highly dependent on the quality of the protein coding annotation. Thus, the current manually curated set provides a high quality lower bound with respect to the true number of pseudogenes in the mouse genome, while the union of automatic annotation pipelines informs on the upper limit of the pseudogene complement size. In agreement with our previous work [33, 34] there is a considerable overlap, of over 83%, between the manual and automatic annotation sets.

Deleted: C57BL/6J

Deleted: 1, S1

Deleted:

For human, we used a similar workflow to refine the reference pseudogene annotation to a high-quality set of 14,650 pseudogenes. The updated set contains considerable improvements in the characterization of pseudogenes of previously unknown biotype (**Table** S1C). In both the human and mouse reference genomes the majority of the annotations are processed pseudogenes, with a smaller fraction of duplicated pseudogenes (**Table** S1C).

Deleted: S2

Deleted: S2

1.2 Mouse strains

The Mouse Genome Project has sequenced and assembled genomes for 12 laboratory, and 4 wild-derived mice, and developed a draft annotation of each organisms' protein coding genes [36]. Another two distant Mus species, *Mus Caroli* and *Mus Pahari*, were also sequenced and assembled [37]. Collectively the 18 strains provide a unique overview of mouse evolution. The strains are broadly organized into 3 classes (**Table** S2): the outgroup strains – formed by two independent mouse species, *Mus Caroli* and *Mus Pahari*; wild strains – covering two subspecies *Mus Spretus* and three musculus strains (*Mus Musculus Musculus* and *Mus Musculus Domesticus*), and a set of 12 laboratory strains. A detailed summary of the genome composition for each strain is presented in [36].

Deleted: 2

Deleted: -derived inbred

Deleted: the

Deleted: (

Deleted: ,

Deleted: *Mus Castaneus*) and two

Deleted: *Castaneus, Mus*

Deleted: *Musculus,*

Deleted: inbred

We developed an annotation workflow for identifying pseudogenes in the 18 mouse strains leveraging the in house automatic pipeline PseudoPipe and the set of manually curated pseudogenes from the mouse reference genome lifted over onto each individual strain (**Figure 1B**). This combined pseudogene identification process gives rise to three confidence levels reflecting the annotation quality. Each identified pseudogene is associated with details about its transcript biotype, genomic location, structure, sequence disablements, and confidence level. A detailed overview of pseudogene annotation statistics including the number of pseudogenes, their confidence levels, and biotypes is shown in **Figure 1C**. The observed reduction in the number of pseudogenes in the distant species is correlated to the decrease in the number of conserved protein coding genes (between the analyzed strain and the reference mouse genome) used as input in the annotation workflow (**Figure SF1A&B**). However, based on close relationship between the mouse reference strain C57BL/6J and its related laboratory inbred strain counterpart C57BL/6NJ, we are able to estimate the total number of pseudogenes in each of the 18 mouse genomes (**Table S3**). The results suggest that all of the studied strains have pseudogene complements of similar size. The difference between the number of annotated pseudogenes and the expected total can be overcome by improving the protein coding annotation in each of the studied strains.

S-1

REW

Currently, around 30% of pseudogenes in each strain are defined as high confidence Level 1 annotations, being identified through both automatic curation and manual lift over, 10% are Level 2 annotations characterized only using the lift over process, and 60% are Level 3 annotations identified solely by the automatic annotation pipeline. The pseudogene biotype distribution across the strains closely follows the reference genome and is consistent with the biotype distributions observed in other mammalian genomes (e.g. human [33] and macaque [34]). As such, the bulk (~80%) of the annotations are processed pseudogenes, while a smaller fraction (~15%) are duplicated pseudogenes. Finally, the density of pseudogene disablements follows the previously observed distributions in the mouse

Deleted: distribution

reference genome and other mammals, with stop codons being the most frequent defect per base pair followed by deletions and insertions (**Figure SF2**). As expected, older pseudogenes show an enrichment in the number of disablements compared with the parental gene sequence. The proportion of pseudogene defects exhibits a linear inverse correlation with the pseudogene age, expressed as the sequence similarity between the pseudogene and the parent gene.

Deleted: 1

1.3 Unitary pseudogenes

Unitary pseudogenes are the result of a complex interplay between LOF events and changes in evolutionary pressures resulting in the fixation of an inactive element in a species. The importance of unitary pseudogenes resides not only in their ability to mark LOF events, but also in their potential to highlight changes in the selective pressures guiding genome evolution. Due to their formation as a result of gene inactivation the identification of unitary pseudogenes is highly dependent on the quality of the reference genome protein coding annotation, and requires a large degree of attention during the annotation process.

Deleted: loss-of-function

Deleted: loss-of-function

These pseudogenes are defined relative to the functional protein coding elements in another species. Using a combination of multi sequence alignments, manual curation, and a specialized unitary pseudogene annotation workflow (**Figure 1B**) in human and mouse, we identified 217 and respectively 237 new unitary pseudogenes. These results bring the total number of unitary pseudogenes in mouse to 271 and raise the size of unitary pseudogene class in human to 431 entries (**Table** S4). This is a considerable increase compared to previous GENCODE releases and can be largely attributed to the improvements in the mouse genome annotation and assembly. In human, a large proportion of unitary pseudogenes are related to the chemosensory system (e.g. GPCRs, olfactory receptor proteins) which have functional homologs in mouse, reflecting the loss of function in these genes during the primate lineage evolution. One such example if the Cyp2G1 unitary pseudogene in human, which has functional counterparts in mouse, rabbits and several primates (**Figure 2A**). Here the human gene acquired a C-T mutation resulting in a stop codon in the middle of a coding exon resulting in gene disablement and thus the creation of a unitary pseudogene.

Deleted: S3

Deleted: and vomeronasal

Deleted: the

Moreover, we observed the pseudogenization of a number of innate immune response related genes in humans such as Toll-like receptor gene 11 and leucine rich repeat protein genes hinting at potentially advantageous LOF/pseudogenization events in human lineage evolution [38]. 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 (**Table** S5).

Deleted: S4

The draft nature of the mouse strains' annotation and assembly makes it difficult to identify unitary pseudogenes in them. To get an overview of the unitary pseudogenes in each strain we used the mouse reference genome as the required canonical organism and followed a similar workflow as described above, except in this case, we used the mouse reference strain specific peptides as input. The resulting pseudogene calls were intersected with the lift over of the reference strain specific protein coding genes in order to validate the conservation of location and loss of function of the latter. On average, we found around 20 unitary pseudogenes in each strain with larger numbers observed for wild-derived inbred strains (**Table** S6). The unitary pseudogenes are distinguished from other strain specific pseudogenes by the fact that they do not have a functional homolog (parent gene) in the same organism. Moreover, the fast rate of evolution among the mouse strains, as well as the highly specific generation of the laboratory strains, suggests that the number of unitary pseudogenes could be considerably higher, reflecting the strain specific phenotypes. A way to get a realistic assessment of the size of the unitary pseudogene complement is to look at the unitary annotation in the human genome relative to other primates [22], as previous studies suggest that protein gene loss rate is similar in both mouse and primate lineages [39]. As such we expect that the total number of unitary pseudogenes with respect to the

Deleted: S5

reference in each strain will be dependent on the evolutionary distance between the two and will be comparable to the number of human specific unitary pseudogenes with functional homologs in chimp that is estimated to be 403 [40].

Similarly, future improvements in the strain annotation will allow us to annotate unitary pseudogenes in the reference with respect to the mouse strains. These elements will not only highlight LOF events in the reference, but also fixation of gain-of-function mutations in divergent strains and species as we found in the case of NCR3 gene in Caroli (**Figure 2B**). Here, we observed an A-G gain of function mutation for the NCR3 gene that is pseudogenized in all the other mouse strains including the reference, reverting the initial TGA stop to a tryptophan codon.

2. Conservation and divergence in pseudogene complements

In order to investigate the evolutionary history of pseudogenes in the mouse strains, we created a *pangenome* pseudogene dataset containing 49,262 unique entries relating the pseudogenes across strains. We found 2,925 ancestral pseudogenes that are preserved across all strains. A detailed summary of the other subsets of pseudogenes is shown in **Figure 3A,B**. On average, each strain contains between 1,000 and 3,000 pseudogenes that are not directly associated with any pseudogenes in the other strains based on the imposed ortholog selection criteria (see **Methods**). By relaxing these constraints, we are able to estimate the minimum number of strain-specific pseudogenes. To this end we were able to identify on average 293 unique elements in each analyzed genome. This is however only a lower bound estimate. Moreover, the proportion of pseudogenes conserved only in the outgroup, the wild-derived strains, or the lab strains is considerably smaller, suggesting that the bulk of the pseudogenes in each strain was created during the shared evolutionary history.

Next, we took advantage of pseudogenes' ability to evolve with little or no selective constraints [41], and compared mutational processes across the mouse strains. To this end, we built a phylogenetic tree based on approximately 3,000 pseudogenes that are conserved across all strains (**Figure 3C**). This pseudogene-based tree follows closely the tree constructed from protein coding genes and correctly identifies and clusters the mice into three classes: outgroup, wild, and laboratory strains. In constructing these trees, we concatenated the gene sequences in the same order in all the strains, thus overriding any potential bias induced by the strains' mosaicism, and focusing only on the sequence alterations

Deleted: a

Deleted:

Deleted:

Deleted: The

Deleted: sequences selected from the

Deleted: -derived

Deleted: ... [1]

Deleted: subgroups (**Figure 3C**). By comparing

Deleted: resulting trees to

Deleted: protein-coding tree, we found that they display different patterns, reflecting different evolutionary histories. For example, the olfactory receptor tree, shows discrepancies in both the divergence

Deleted: as well as

Deleted: degree of conservation of the ancestral

Deleted: (as reflected by the branch length), with notable differences observed for NZO, and NOD laboratory strains. These two strains are known for exhibiting two distinct diabetic phenotypes. Despite this difference, the result suggests that in both cases changes in diet can affect chemoreceptors (key players in both phenotypes) and consequently their evolution [42].

3. Genome Evolution & Plasticity

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

3.1 Pseudogene Genesis

Taking advantage of the available functional genomics and evolutionary data we can study the pseudogene genesis on a unique scale: during embryo development at one extreme and the mouse lineage at the other.

Given that processed pseudogenes are formed through the retrotransposition of the parent mRNAs, we hypothesized that there is a direct correlation between the parent gene expression level and the number of processed pseudogenes [43]. Moreover, as pseudogenes are inherited, the genesis of new elements occurs in the germline. To this end, we used an embryogenesis RNA-seq time course dataset to test our assumptions during early development [44]. We calculated the parent gene expression for a series of developmental stages ranging from metaphase II oocytes to the inner cell mass. At every stage, the average expression level of parent genes is higher than that observed for non-parent protein coding genes. However, genes associated with large pseudogene families show low transcription levels during very early development, with high expression levels achieved only during later stages. We evaluated the correlation between the number of pseudogenes associated with a gene and its expression level at different developmental time-points. This correlation improves as we move forward through the developmental stages suggesting that pseudogenes are most likely generated by highly expressed housekeeping genes.

We further tested the correlation between high expression levels and the number of associated pseudogenes by looking at RNA-seq samples from adult mouse brain. Similar to our previous observations, the pseudogene parent genes show a statistically significant increase in average expression levels compared to non-pseudogene generating protein coding genes (**Figure SF3**).

Deleted: 2

Next, we looked at the degree to which the number of pseudogenes is related to the number of copies or functional paralogs of the parent gene (**Figure 4A**). For duplicated pseudogenes, we observe a weak correlation between the number of paralogs and the number of pseudogenes of a particular parent gene. This result suggests that a highly-duplicated protein family will tend to give rise to more disabled copies than a less duplicated family, if we assume that each duplication process can potentially give rise to either a pseudogene or a functional gene.

By contrast, for processed pseudogenes we observed a weak inverse correlation. This result implies that in the case of large protein families we can expect to see a lower level of transcription for each family member, with high mRNA abundance being achieved from multiple duplicated copies of a gene rather than increasing the expression of a single unit. Therefore, there is a weak correlation between the number of paralogs of the parent and the potential gene expression level of the parent genes, and thus we observe a smaller number of associated pseudogenes (**Figure SF4**).

Deleted:

Deleted: 3

3.2 Transposable elements

Since the majority of mouse and human pseudogenes are the result of retrotransposition processes mediated by transposable elements (TE), we investigated the genomic mobile element content in human and mouse as well as the generation of processed pseudogenes on an evolutionary time scale (**Figure 4B**).

Deleted:)

TEs are sequences of DNA characterized by their ability to integrate themselves at new loci within the genome. TEs are commonly classified into two classes: DNA transposons and retrotransposons, with the latter being responsible for the formation of processed pseudogenes and retrogenes. Both human and mouse genomes are dominated by three types of TEs, namely short interspersed nuclear elements, long interspersed nuclear elements (LINEs) and endogenous retrovirus superfamily. LINE-1 elements (L1) have been shown to mobilize Alu's, small nuclear RNAs and mRNA transcripts. We analyzed the L1 retroposed processed pseudogenes in human and mouse. We define the evolutionary time scale by using the pseudogene sequence similarity to the parent gene as a proxy for age. Younger pseudogenes have a higher degree of sequence similarity to the parent, while older pseudogenes show a more diverged sequence.

Deleted: (SINEs),
Deleted: the
Deleted: (ERV)
Deleted: s

9-1

In human, we observe a smooth distribution of processed pseudogenes, with a single peak (at 92.5% sequence similarity to parents) hinting at the burst of retrotransposition events, that occurred 40 MYA at the dawn of primate lineage and created the majority of human pseudogene content [45, 46]. By contrast in mouse, we found that the processed pseudogene distribution is defined by two successive peaks at 92.5% and 97% sequence similarity to parent genes. Also, in contrast to human where the density of processed pseudogenes shows a steep decrease amongst young pseudogenes following the peak at 92.5% similarity, the density of mouse processed pseudogenes remains at a high level in the interval 97% to 100% sequence similarity to parents. A close examination of the young pseudogene density suggests a reduction in the number of new pseudogenes being created. This is most likely a consequence of the stringent criteria used in calling pseudogenes at high sequence similarity to parents and showcases the difficulty in annotating recently disabled/dead genes due to their high similarity to functional protein coding counterparts. Overall these results suggest the presence of active transposable elements in mouse which results in a continuous renewal of the processed pseudogene pool. This is also reflected in the large difference in the number of active LINE/L1s between human and mouse, with just over 100 in human [47] compared to 3,000 in mouse [48].

9+2

AGE DIST 2

REL TO FEW IN STRAINS

Deleted: between
Deleted: observation suggests

3.3 Genome remodeling

The large proportion of strain and class specific pseudogenes, as well as the presence of active TE families, point towards multiple genomic rearrangements in mouse genome evolution. To this end, we examined the conservation of pseudogene genomic loci between each of the mouse strains and the reference genome for one-to-one pseudogene orthologs (**Figure 5A,B**). We observed that on average more than 97.7% of loci are conserved across the laboratory strains, and 96.7% of loci are conserved with respect to the wild-derived strains. By contrast, only 87% of Caroli loci were conserved in the reference genome, while Pahari showed only 10% conservation. The significant drop in the number of conserved pseudogene loci between the reference genome and outgroup strains is in agreement with the observed major karyotype-scale differences and large genomic rearrangements exhibited by Caroli and Pahari [37, 49]. The proportion of un-conserved loci follows a logarithmic curve that matches closely the divergent evolutionary time scale of the mouse strains suggesting a uniform rate of genome remodeling processes across the murine taxa (**Figure 5C**).

Deleted: in each pair

4. Functional analysis

The role of pseudogenes in genome biology has long been debated. However, recent studies [34] have highlighted the fact the pseudogenes can reflect the evolution of genome function and activity. Here we address the biological relevance of pseudogene activity leveraging data from the gene ontology, protein families, and RNA-seq experiments.

4.1 Gene ontology & pseudogene family analysis

We integrated the annotations with gene ontology (GO) data in order to characterize the functions associated with pseudogene generation. For this, we calculated the enrichment of GO terms across the strains. We observed that the majority of top biological processes, molecular function, and cellular component GO terms are shared across the strains (**Figure 6A**). We also evaluated GO term enrichment amongst parent genes for both processed and duplicated pseudogenes across the mouse strains. Enriched GO terms were clustered based on semantic similarity and the strains were clustered based on GO term enrichment profile similarity. The resultant heatmap (**Figure 6B**) enables the identification of both related terms with conserved enrichment across all strains as well as blocks of terms that exhibit conservation within a single or a few closely related strains. Conserved enrichment for GO terms related to ribosomal functions, cell cycle, translation and RNA processing, and ubiquitination was observed for processed pseudogenes. Amongst duplicated pseudogenes, we observed enrichment for apoptosis, sensory and smell processes, and immune functions. Additionally, the GO terms that universally characterize the pseudogene complements in all the mouse strains are closely related to the family classification of pseudogenes. The top pseudogene family is 7-Transmembrane. This Pfam family encompasses the chemoreceptors GPCR proteins reflecting the enrichment in olfactory receptors in the mouse. Similar to the human and primate counterparts, many top families in mouse pseudogenes are related to highly expressed and duplicated proteins such as GAPDH and ribosomal proteins, and regulatory protein families such as the Zinc fingers (**Figure 6C**).

Deleted: clustered

Deleted:

A closer look suggests that the pseudogene repertoire also reflects individual strain-specific phenotypes (Table S5). The pseudogene phenotype relationship can be viewed from different perspectives. First, pseudogenes reflect duplication events linked with the emergence of an advantageous phenotype. This is observed in the *Mus Spretus* genome, where we see as an enrichment of duplicated tumor repressor and apoptosis pathways genes [50] and correspondingly an increase in the number of associated pseudogenes. Second, we find pseudogenes reflecting the death of a gene family. As such we observe an increase in the number of pseudogenes associated deleterious phenotypes. A known example is the pseudogenization of Cytochrome c Oxidase subunit VIa through accumulation of LOF mutations in the blind albino mouse strain, that is commonly linked with neurodegeneration [51] and is characteristic for the observed brain lesions in the affected mice [15]. However, a detailed analysis of the pseudogene repertoire suggests that there are more ways to describe the pseudogene phenotype association, in particular looking at the emergence of advantageous phenotypes through the pseudogenization process [52].

Deleted: . 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

Deleted: 3.

Deleted: -

Deleted:

Deleted: –

Deleted:

4.2 Gene essentiality

We observed an enrichment of essential genes among pseudogene parent genes in the mouse strains. Evaluating the parent gene for each pseudogene present in the mouse strains reveals essential genes are approximately three times more abundant amongst parent genes (**Table** S7A). In general, the essential genes are more highly transcribed than nonessential genes [53], and thus might be associated with a higher propensity of generating processed pseudogenes. However, one potential confounder is the gene expression level which is associated with both more processed pseudogenes and more essential genes. Thus, we evaluated the probability that a gene is essential by controlling for its transcription level and parent gene status (see Methods), and found that pseudogene parents are still 20% more likely to be essential genes compared to regular protein coding genes (**Table** S7B).

Deleted: 3

Deleted:)

Deleted: S6

We also analyzed the number of paralogs associated with our essential and nonessential gene sets to get an insight into the possible role of gene duplication in the enrichment of essential genes amongst the parent genes set. In the reference mouse 80.6% of nonessential genes and 74.1% of essential genes have paralogs. This is in agreement with previous work showing non-essential genes are more likely than essential to be duplicated successfully [54].

Deleted: s

Deleted:

4.3 Pseudogene Transcription

We leveraged RNA-seq data from the Mouse Genome Project and ENCODE to study pseudogene biology as reflected by their transcriptional activity. This is thought to either relate to the exaptive functionality of pseudogenes or be a residual leftover from their existence as genes. In both the human and mouse reference genomes, we detected that about 15% of pseudogenes were transcribed across a variety of tissues, a result similar to previous pan tissue analyses (**Figure 7A,B**).

Due to restricted data availability in the mouse strains, we focused our transcriptional analysis to a single tissue – adult brain from wild-derived and laboratory strains. Overall pseudogenes with strain specific transcription were more common than those with cross-strain transcription (**Figure 7C,D**). Moreover, the proportion of pseudogenes conserved across all strains that are transcribed is constant (~2.5%) across the wild-derived and laboratory strains (**Figure 7D**). By contrast, the fraction of transcribed strain specific pseudogenes varies across the strains from 1.5% to 4% (**Figure 7D**).

5. Mouse pseudogene resource

We created a comprehensive resource that organizes all of the pseudogenes across the available mouse strains and the reference genome, as well as associated phenotypic information, in a database that is available online at mouse.pseudogenes.org. The database contains information regarding strain and cross-strain annotation, pseudogene family and phenotypic data, as well as expression data. All the available data are provided as flat files for ease of manipulation. Queries on specific pseudogenes will return the relevant pseudogene annotation containing all pertinent associated information. The pseudogenes are annotated with 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. This enables pairwise comparisons of pseudogenes between the various mouse strains and the investigation of differences between multiple strains of interest.

Deleted: pseudogene

Discussion

We report the updated and refined pseudogene annotation in the mouse and human reference genomes, and describe the curation and comparative analysis of the first draft of pseudogene complements in 18 related mouse strains. By combining manual and computational annotations we were able to obtain a comprehensive view of the pseudogene content in genomes throughout the mouse lineage. The overlap between manually curated pseudogenes and those identified using computational methods is over 80% reflecting the high sensitivity of the computational detection method.

Deleted: curation

Deleted: pseudogene sets

Comparable to our previous observations in human, worm, and fly, the pseudogene complement in mouse strains, reflects an organism specific evolution, highlighting pseudogenes as ideal markers of genome remodelling processes. However, despite the strain dependent evolution, the pseudogenes share a number of similarities, in particular regarding their biogenesis and diversity. As such we noticed a uniform ratio of processed to duplicated pseudogene of 4 to 1 in all of the strains, a result consistent with previous observations in human. The higher proportion of processed pseudogenes accounting for ~80% of the total, is in agreement with earlier findings that suggest retrotransposition as the primary mechanism for pseudogene creation in numerous mammalian species [33]. Moreover, examining the retrotransposon activity, and in particular the L1 content, we observed that while the majority of human pseudogenes have been formed relatively recently through a single burst of retrotransposition [33], the mouse lineage shows a sustained renewal of the pseudogene pool through successive bursts. The sequence context of the processed pseudogenes indicates that the various retrotransposons exhibit differential contributions to the pseudogene set over time.

Deleted: retrotranspositional

Since a pseudogene's likelihood of creation is related to its parent's functional role and expression level, they can act as a record of their parent gene's expression level and perhaps provide insight into the past importance of their parent gene. The link between the creation of processed pseudogenes and parent genes associated with key biological functions is further supported by an enrichment of parent genes amongst mouse essential genes. Meanwhile, duplicated pseudogenes record events that shaped both the genome environment and function during the organism's evolution. Furthermore, the wealth of functional genomics assays available for the experimentally relevant mouse strains presents an opportunity to investigate both the activity of parent genes as well as pseudogene genesis. As expected we observed that parent genes have higher levels of expression relative to non-parents both during embryo development as well as in adult tissue. Moreover, time series expression analysis during embryo development suggest that most pseudogene creation is commonly related to the high expression levels of housekeeping genes (**Figure SF5**).

Deleted: 4

To better understand the evolutionary and functional relationship between the pseudogenes in the 18 strains we constructed a pan-genome pseudogene set as the union of all individual strain complements, resulting in over 45,000 unique entries. The pan-genome pseudogene repertoire distinguishes three types of pseudogenes: universally conserved (present in all 18 strains), multi-strain (present in at least 2 strains), and strain specific (unique to a specific strain and without an associated ortholog in another strain), accounting for 6, 23, and 71% of the elements respectively. Despite the large number of pseudogenes without an associated orthologue in the pangenome set, these account for only 25% of the total pseudogenes in any particular strain, a comparable proportion to the universally conserved pseudogenes present in each strain. Moreover, the pseudogene cross strain relationship allows us to have a closer look at their evolution by studying the conservation of their chromosomal location. In particular, we observed a stark contrast between the high level of genomic loci retention shared by the laboratory strains and the lack of conservation noticed when looking at the outgroup species. These results hint at multiple large scale genomic rearrangements in the mouse lineage. This is especially noticeable in the case of *Mus Pahari* as has been recently reported by large scale chromosomal imagining and karyotype analysis [37, 49].

Deleted: and

Deleted: of the

Deleted: studied

Deleted: of the

Deleted: strain specific

Deleted: identified from the pangenome

Analysis of pseudogenes and their parent genes can provide a window into changing functional constraints and selective pressures. Unitary pseudogenes are markers of loss of function mutations that that have become fixed in the population. Here we annotated over 200 new unitary pseudogenes in mouse and a similar number in human. We found that the enrichment of vomeronasal receptor unitary pseudogenes in human with respect to mouse highlights the loss of certain olfactory functions in humans. Unitary analysis is especially interesting because it provides us with key moments in the evolution of gene function by marking the loss and gain of function events. In particular, we note the pseudogenization of the human Cyp2G1 gene while its mouse counterpart is still functional, and the disablement reversal mutation that lead to a functional NCR3 gene in Caroli while the mouse reference and the other laboratory and wild-derived strains show the presence of a pseudogene on the same locus.

Deleted: Moreover, unitary

Taking advantage of the availability of information-rich resources such as Gene Ontology and Pfam, we looked to functionally characterize the pseudogenes. For this we annotated pseudogenes and parent genes in each strain with GO terms and Pfam families. We observed an enrichment in housekeeping functions associated with conserved pseudogenes as illustrated by the presence of GAPDH, ribosomal proteins, and zinc finger nucleases as top Pfam families amongst the mouse pseudogenes. The top mouse pseudogene families closely match those seen in human. The GO enrichment analysis supports the above results, with top terms including RNA processing and metabolic processes. Additionally, we used the pan-genome pseudogene set to identify strain specific functional annotations and suggest hypotheses as to what cellular processes and genes might underpin phenotypic differences between the

Deleted: in the mouse lineage.

Deleted: Moreover, the

mouse strains. For example, we observed that PWK is associated with strain specific GO terms for melanocyte-stimulating hormone receptor activity and melanoblast proliferation, which may play a role in the strain's patchwork coat color [55]. NZO, an obesity prone mouse strain, is characterized by a specific enrichment in defensin associated pseudogenes. Defensins are small peptides involved in controlling the inflammation resulted from metabolic abnormalities in obesity and type 2 diabetes [56], and more recently described as potential markers of obesity [57]. Taken together the functional analysis of pseudogenes provides an opportunity to better understand the selective pressures that have shaped an organism's genomic content and phenotype.

Meanwhile, looking at pseudogene expression across the strains we observed evidence of both pseudogenes with broadly conserved transcription as well as some with strain specific expression. As additional RNA-seq datasets for multiple tissues for each strain become available future work can investigate both pan strain and pan tissue expression patterns.

In summary, this comprehensive annotation and analysis of pseudogenes across 18 mouse strains has provided support for conserved aspects of pseudogene biogenesis while also expanding our understanding of pseudogene evolution and activity. Integration of the pseudogene annotations with existing knowledge bases including Pfam and the gene ontology have provided insight into the biological functions associated with pseudogenes and their parent genes. The well-defined relationships between the strains aided evolutionary analysis of the pseudogene complements. Taken together, annotation of pseudogenes across a range of extensively used laboratory mouse strains and their integration into a comprehensive database with evolutionary and functional genomics data provides a useful resource for the broader research community. Additionally, the experimental and functional genomics datasets associated with these well-studied strains shed light on the transcriptional activity of pseudogenes and offer promise for future studies.

Deleted: The

Materials and Methods

Code and data availability

Moved (insertion) [1]

The pseudogene annotation pipeline is freely available at http://pseudogene.org/pseudopipe. All supplementary data is available at http://mouse.pseudogene.org/Supplement/

Moved (insertion) [2]

Datasets

Mouse reference genome is based on the Mus Musculus strain C57BL/6J strain. The mouse reference annotation is based on GENCODE vM12/Ensembl 87.

The human reference genome annotation is based on GENCODE v25/Ensembl 87.

The 16 laboratory and wild-derived inbred strains (**Table S2**) assemblies and strain specific annotations were obtained from the Mouse Genome Project [36] (http://www.sanger.ac.uk/science/data/mouse-genomes-project, last accessed on 21.08.2017). The laboratory strain C57BL/6NJ is a subline of the reference strain [15]. There is a high sequence and evolutionary similarity between the reference genome single inbred strain C57BL/6J and the laboratory inbred mouse strain C57BL/6NJ. For the purpose of this study and in order to facilitate a reliable comparison across all the studied mouse genomes, we used the laboratory inbred strain C57BL/6NJ as a reference point.

The two outgroup mouse species (**Table S2**). Mus Caroli and Mus Pahari were sequenced, assembled, and annotated in the protein-coding domain by ref. [37].

Moved (insertion) [3]

Human – Mouse Lineage Comparison

Human – primate lineage divergence and generation times were obtained from [58]. The divergence times for the wild-derived and laboratory strains were obtained from [59, 60, 61]. The data for two outgroup species divergence times was obtained from [37]. The generation time for all the mice was estimated from [15].

Moved (insertion) [4]

Pseudogene Annotation

Reference genome annotation

We manually curated 10,524 pseudogenes in the mouse reference genome (GENCODE M12) and 14,650 pseudogenes in the human reference genome (GENCODE v25), using a workflow previously described in [33, 34]. The manual annotation is based on the sequence homology to protein data from UniProt database [34] and the protocol is summarised in **Figure SF6**.

Moved (insertion) [5]

The number of manually annotated pseudogenes in the mouse lineage is likely an underestimate of the true size of the mouse pseudogene complement given the similarities between the human and mouse genomes. Thus, to get a more accurate idea of the number of pseudogenes in the mouse genome, we used a combination of two automatic annotation pipelines: PseudoPipe [35] and RetroFinder [62]. PseudoPipe is a comprehensive annotation pipeline focused on identifying and characterizing pseudogenes based on their biotypes as either processed or duplicated. The automatic annotation workflow using PseudoPipe is summarised in **Figure 1B** and has been previously described in detail in [33, 34, 35]. Pseudopipe identifies 22,811 mouse pseudogenes of which 14,084 are present in autosomal chromosomes (a number comparable with the one observed previously in human (**Table S1**)). RetroFinder is computational annotation pipeline focused on identifying retrotransposed genes and pseudogenes. Using RetroFinder we were able to annotate 18,467 and respectively 15,474 processed pseudogenes in mouse and human. There is a good overlap between the two identification pipelines with respect to the number of processed pseudogenes present in both organisms (**Table S1**).

Mouse strain annotation

The mouse strain pseudogene annotation workflow is summarised in **Figure 1B**. The protein coding input set contains the conserved protein coding genes between each mouse 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. PseudoPipe was run with the strain conserved protein set as shown in **Figure 1B**. Next, we used HAL tools package [63] to lift over the manually annotated pseudogenes from the mouse reference genome onto each strain using the UCSC multi strain sequence alignments. We merged the two annotation sets using BEDTools [64] with 1bp minimum overlap requirement. We extended each overlap predicted boundaries to ensure full annotation of the pseudogene transcript. Finally, we did a manual inspection of the resultant annotation set in order to eliminate all potential false positives (e.g. pseudogene calls larger than 5Kb, smaller than 100bp with poor protein coding gene query similarity and coverage).

Moved (insertion) [6]

Moved (insertion) [7]

To estimate the total number of pseudogenes in each strain we set two hypotheses:

1. Given the close evolutionary relationship between the mouse reference strain C57BL/6J and the laboratory reference strain C57BL/6NJ, we expect that given the same genome assembly quality and protein coding annotation, the two strains will exhibit the same number of pseudogenes.
2. The pseudogene generation is linear across all the strains.

14-1

SIMIL,

Thus, we relate the reduction in the number of input protein coding gene sequences with the total number of annotated pseudogenes. As such, using 82.7% of the protein input in the C57BL/6NJ will result in the annotation of only 79.3% of the total pseudogenes. Following this ratio, we estimated the total number of pseudogenes in each strain using the formula:

$$EstimateTotal_{strain} = \frac{PseudoPipeOutput}{\%PCTconserved_{strain}} \cdot \frac{\%PCTconserved_{C57BL/6NJ}}{\%PG_{C57BL/6NJ}}$$

Where %PCTconserved is the % of Protein Coding Transcripts that are conserved between the strain and the mouse reference, and %PG is the % of the identified pseudogenes in the C57BL/6NJ with respect to the total number of pseudogenes annotated in the reference genome.

Unitary Pseudogene Annotation Pipeline

Moved (insertion) [8]

We modified PseudoPipe to allow cross-strains and cross species protein coding inputs. We annotated cross-organism pseudogenes as shown in **Figure** 1B. "Functional organism" is defined as the genome providing the protein coding information and thus containing a working copy of the element of interest. "Non-functional" organism is the genome queried for unitary pseudogene presence. The resulting data set was subjected to a number of filters such as removal of previously known pseudogenes, removal of pseudogenes with parents that have orthologs in the annotated specie, removal of pseudogenes that overlap with annotated protein coding and ncRNAs loci, and removal of pseudogenes shorter than 100 bp. The filtered PseudoPipe set was intersected with the lift-over of the protein coding annotation from the "functional organism using BEDTools [64] with a 1bp overlap minimum required. The intersection set was further refined flagging protein coding genes that have functional relatives (paralogs) in the "non-functional" organism. The remaining matches were subjected to manual inspection of the alignment.

Moved (insertion) [9]

Moved (insertion) [10]

Moved (insertion) [11]

Conservation and divergence in pseudogene complements

Pangenome data set generation

We performed an all against all liftover of pseudogene annotation using HAL tools package and the UCSC multi strain sequence alignment. Each liftover was intersected with the know strain annotation and all the entries that matched protein coding or ncRNAs were removed. The resulting set is further filtered for conservation of pseudogene Ensembl ID, where available (used for Level 1 and 2 pseudogenes), conservation of parent gene identity, conservation of pseudogene locus (overlap of 90% or higher), conservation of pseudogene biotype, conservation of pseudogene length, and conservation of pseudogene structure.

Next, we integrated all filtered binary mappings in a master pan-strain set. The common entries were collapsed into a unique pangenome pseudogene reference. We obtained 49,262 pangenome pseudogenes. 1,158 pangenome entries are multi matching across strains.

Moved (insertion) [12]

Moved (insertion) [13]

To estimate the number of strain specific pseudogenes, we relaxed the cut-off level in the conservation of pseudogene locus and sequence overlap (see **Figure SF7**). The lower the threshold, the larger number of called orthologs and consequently a smaller number of strain specific pseudogenes. The minimum number of expected strain specific pseudogenes in the current dataset was calculated under the hypothesis that a strain specific pseudogene will have 0% sequence overlap with any annotated elements in any of the other strains. Thus, there are a minimum of 295 strain unique pseudogenes on average in any of the 18 mouse genomes.

Phylogenetic analysis

Moved (insertion) [14]

Sequences of the 1,460 pseudogenes were randomly selected out of the total of 2925 conserved pseudogenes in the 18 mouse strains accounting for approximately 50% of the total number of conserved pseudogenes. For each of the 18 mouse genomes, the extracted sequences were concatenated into strain-specific contig (supergene). The order of the pseudogene sequences was kept the same in all 18 contigs. Preserving the same order of pseudogenes or protein coding genes across all strains eliminates any potential bias resulting from the laboratory strain mosaicism, as the relative location of a gene is not considered when creating the trees. Thus, the resulting phylogeny depends only on the sequence evolution. The 18 supergenes were subjected to a multi-sequence alignment using MUSCLE aligner [65] under standard conditions. Similarly, the sequences of parent protein coding genes of the 1,460 pseudogenes were assembled into a strain specific sequence and aligned using MUSCLE. The tree was generated using Tamura-Nei genetic distance model and neighbouring-joining tree build method with Pahari as outgroup using GENEIOUS 10.2 software package [66].

Moved (insertion) [15]

Genome evolution and plasticity

Genome mappability maps

We created mappabilty maps for the mouse reference genome and the 18 mouse strains using the GEM library [67]. The workflow is composed of indexing the genome using gem-indexer, followed by creation of the map using a window of 75 nucleotides under the following conditions -m 0.02 -T 2.

Parent gene expression analysis

RNAseq mouse tissue data was obtained from ENCODE. The complete list of experiments used is available in **Table S8**. We estimated the pseudogene parent protein coding genes expression levels using a workflow involving the following steps: filtering the protein coding genes for uniquely mappable regions longer than 100bp, mapping reads using TopHat2 [68], selecting high quality mapped reads with a quality score higher than 30 using samtools [69], and calculating the expression FPKM levels using Cufflinks [70]. Transcriptional activity of pseudogene parent genes during early embryonic development was investigated using RNAseq data as processed and described in [44]. Raw sequencing data and processed data containing FPKM levels at each embryonic stage are available on the SRA under Series GSE66582.

Moved (insertion) [16]

Transposable elements analysis

TE in human and mouse reference genomes were informed from RepeatMasker libraries Repbase 21.11 and using RepeatMasker 3.2.8 [71]. We extracted all the four major groups of repeats SINE, LINE, LTR and DNA and identified all the processed pseudogenes associated with L1 elements. Next, we binned the L1 annotated pseudogenes into age groups based on their sequence similarity to the parent gene, with younger elements exhibiting a higher sequence similarity while older elements show a large sequence divergence when compared to the functional gene counterparts.

Gene ontology and Pfam analysis

Linking of gene ontology terms to the pseudogene parent genes was conducted using the R package biomaRt [72, 73]. Visualization of shared and distinct GO term sets amongst the strains was done using the R package UpSetR [74]. Enrichment of GO terms amongst the pseudogene parent genes and clustering of mouse strains based on similar enrichment profiles was performed using the goSTAG software package [75]. Semantic clustering of the GO terms was done with the OntologyX packages [76]. Parent genes were labelled with both strain and biotype information in order to better evaluate differences in the pseudogene complements based on their mechanism of creation.

Analysis of the Pfam representation in the pseudogene complements was performed as previously described in [77] and focused on associating the pseudogene with the protein family of its parent gene.

Gene essentiality enrichment analysis

Moved (insertion) [17]

Lists of essential and nonessential genes were compiled using data from the MGI database and recent work from the International Mouse Phenotyping Consortium [78]. The nonessential gene set with Ensembl identifiers contained 4,736 genes compared to 3,263 essential genes.

In order to evaluate the impact of parent gene status on the probability of a gene being essential while controlling for transcription we fit a linear probability model and a probit model for the probability that a gene is essential given its transcription level and parent gene status using the StatsModels package in Python. The linear probability model fits an ordinary least squares regression of gene essentiality on parent gene status and transcription level. While the linear probability model generally estimates relationships well close to the mean of the independent variables, it often loses explanatory power at low and high values of these variables. Because of this deficiency, we looked also at the probit model, which is similar to the linear probability model but instead fits the data to a cumulative Gaussian distribution. Around the mean values, we find that parent gene status increases the probability of essentiality by around 20% in both models.

Pseudogene transcription

We estimated the pseudogene transcription levels for the mouse reference in 18 adult tissues following a similar protocol to the one described earlier for calculating the expression of protein coding genes, a method that we have successfully used in the past [34] using RNAseq ENCODE data (**Table** S8). The pseudogene sequences were filtered for uniquely mappable exon regions longer than 100 bp. Next the RNAseq raw data was mapped using TopHat and the mapped reads were filtered for quality scores higher than 30. The resulting alignments were quantified using Cufflinks. A pseudogene was considered transcribed if it had an FPKM larger than 3.3 in accord with previous studies [34].

Moved (insertion) [18]

RNAseq data from mouse adult brain was obtained from the Mouse Genome project for 12 laboratory and 4 wild-derived strains (ftp://ftp-mouse.sanger.ac.uk/REL-1509-Assembly-RNA-Seq), sanger experiment, last accessed on 21.08.2017). Next, we created mappability maps for each of the 16 mouse strains genomes and selected only the pseudogene exons in uniquely mappable regions and longer than 100bp for further transcription analysis. The pseudogene transcription levels in mouse strains were estimated using a similar workflow as described above. The transcription cut off level was set to 1.

Moved (insertion) [19]

Mouse pseudogene resource

All the annotation data produced in the analysis is collected and available online through mouse.pseudogene.org. Pseudogene annotation information encompasses the genomic context of each pseudogene, its parent gene and transcript Ensembl IDs, the corresponding mouse reference pseudogene Ensembl ID, the level of confidence in the pseudogene as a function of agreement between manual and automated annotation pipelines, and the pseudogene biotype.

Moved (insertion) [20]

Information on the cross-strain comparison of pseudogenes is derived from the liftover of pseudogene annotations from one strain onto another and subsequent intersection with that strain's native annotations. The database provides liftover annotations and information about intersections between the liftover and native annotations. Furthermore, homology information provide links between the well-characterized mouse strain collection.

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.

Acknowledgements

This project was supported by the Wellcome Trust (grant numbers WT108749/Z/15/Z, WT098051, WT202878/Z/16/Z and WT202878/B/16/Z), Cancer Research UK (20412), the European Research Council (615584), and the European Molecular Biology Laboratory. The research leading to these results has received funding from the European Union's Seventh Framework Programme (FP7/2007-2013) under grant agreement HEALTH-F4-2010-241504 (EURATRANS).

References:

1. Peters, L.L. *et al*. The mouse as a model for human biology: a resource guide for complex trait analysis. *Nat Rev Genet* **8**, 58-69 (2007).
2. Paigen, K. One hundred years of mouse genetics: an intellectual history. I. The classical period (1902-1980). *Genetics* **163**, 1-7 (2003).
3. Paigen, K. One hundred years of mouse genetics: an intellectual history. II. The molecular revolution (1981-2002). *Genetics* **163**, 1227-35 (2003).
4. Yalcin, B, Adams, D. J, Flint, J, Keane, T. M. Next-generation sequencing of experimental mouse strains. *Mamm Genome* **23**, 490-8 (2012).
5. Keane, T. M. *et al.* Mouse genomic variation and its effect on phenotypes and gene regulation. *Nature* **477**, 289-94 (2011).
6. Mestas, J. & Hughes, C. C. W. Of mice and not men: differences between mouse and human immunology. *J Immunol* **172**, 2731-8 (2004).
7. Emes, R.D, Goodstadt, L, Winter, E.E, Ponting, C.P. Comparison of the genomes of human and mouse lays the foundation of genome zoology. *Hum Mol Genet* **12**, 701-9 (2003).
8. Mouse Genome Sequencing Consortium *et al.* Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**, 520-62 (2002).
9. Madsen, O. *et al.* Parallel adaptive radiations in two major clades of placental mammals. *Nature* **409**, 610-4 (2001).
10. Murphy, W. J. *et al.* Molecular phylogenetics and the origins of placental mammals. *Nature* **409**, 614-8 (2001).
11. Hedges, S. B, Dudley, J, Kumar, S. TimeTree: a public knowledge-base of divergence times among organisms. *Bioinformatics* **22**, 2971-2 (2006).
12. Monaco, G, van Dam, S, Casal Novo Ribeiro, J. L, Larbi, A, de Magalhães, J. P. A comparison of human and mouse gene co-expression networks reveals conservation and divergence at the tissue, pathway and disease levels. *BMC Evol Biol* **15**, 259 (2015).
13. Goios, A, Pereira, L, Bogue, M, Macaulay, V, Amorim, A. mtDNA phylogeny and evolution of laboratory mouse strains. *Genome Res* **17**, 293-8 (2007).
14. Beck, J. A. *et al.* Genealogies of mouse inbred strains. *Nat Genet* **24**, 23-5 (2000).
15. http://www.informatics.jax.org/mgihome/other/homepage_IntroMouse.shtml. Last accessed on February, 24th 2017.
16. Richardson, A. *et al.* Use of Transgenic Mice in Aging Research. *ILAR J* **38**, 125-136 (1997).
17. Editorial, Troublesome variability in mouse studies. *Nat Neurosci* **12**, 1075 (2009).
18. Mighell, A.J, Smith, N. R, Robinson, P.A, Markham, A.F. Vertebrate pseudogenes. *FEBS Lett* **468**, 109-14 (2000).
19. Harrison, P.M, Echols, N, Gerstein, M.B. Digging for dead genes: an analysis of the characteristics of the pseudogene population in the Caenorhabditis elegans genome. *Nucleic Acids Res* **29**, 818-830 (2001).

20. Echols, N. *et al.* Comprehensive analysis of amino acid and nucleotide composition in eukaryotic genomes, comparing genes and pseudogenes. *Nucleic Acids Res* **30**, 2515-2523 (2002).
21. Balakirev, E.S. & Ayala, F.J. Pseudogenes: are they "junk" or functional DNA? *Annu Rev Genet* **37**, 123-51 (2003).
22. Zhang, Z.D, Frankish, A, Hunt, T, Harrow, J, Gerstein, M. Identification and analysis of unitary pseudogenes: historic and contemporary gene losses in humans and other primates. *Genome Biol* **11**, R26 (2010).
23. Moore, R.C. & Purugganan, M.D. The early stages of duplicate gene evolution. *Proc Natl Acad Sci U S A* **100**, 15682-7 (2003).
24. Kuang, M.C, Hutchins, P.D, Russell, J.D, Coon, J.J, Hittinger, C.T. Ongoing resolution of duplicate gene functions shapes the diversification of a metabolic network. *Elife* **5**, (2016).
25. Shakhnovich, B.E. & Koonin, E.V. Origins and impact of constraints in evolution of gene families. *Genome Res* **16**, 1529-36 (2006).
26. Ohno, S. Evolution by Gene Duplication. *Springer, New York*, (1970).
27. Kondrashov, F.A, Rogozin, I.B, Wolf, Y.I, Koonin, E.V. Selection in the evolution of gene duplications. *Genome Biol* **3**, RESEARCH0008 (2002).
28. Rogozin, I.B. Complexity of gene expression evolution after duplication: protein dosage rebalancing. *Genet Res Int* **2014**, 516508 (2014).
29. http://www.informatics.jax.org/silver/chapters/1-3.shtml Last accessed on February, 24th 2017.
30. Cameron, J. *et al.* Investigations on the evolutionary conservation of PCSK9 reveal a functionally important protrusion. *FEBS J* **275**, 4121-33 (2008).
31. Maxwell, K.N, Fisher, E.A, Breslow, J.L. Overexpression of PCSK9 accelerates the degradation of the LDLR in a post-endoplasmic reticulum compartment. *Proc Natl Acad Sci U S A* **102**, 2069-74 (2005).
32. Dadu, R.T. & Ballantyne, C.M. Lipid lowering with PCSK9 inhibitors. *Nat Rev Cardiol* **11**, 563-75 (2014).
33. Pei, B. *et al.*The GENCODE pseudogene resource. *Genome Biol* **13**, R51 (2012).
34. Sisu, C. *et al.* Comparative analysis of pseudogenes across three phyla. *Proc Natl Acad Sci U S A* **111**, 13361-6 (2014).
35. Zhang, Z. *et al.* PseudoPipe: an automated pseudogene identification pipeline. *Bioinformatics* **22**, 1437-9 (2006).
36. Lilue, J. Multiple laboratory mouse reference genomes define strain specific haplotypes and novel functional loci. *BioRxiv* https://doi.org/10.1101/235838.
37. Thybert, D. *et al.* Repeat associated mechanisms of genome evolution and function revealed by the Mus caroli and Mus pahari genomes. *BioRxiv* https://doi.org/10.1101/158659.
38. Bryant, C.E. & Monie, T.P. Mice, men and the relatives: cross-species studies underpin innate immunity. *Open Biol* **2**, 120015 (2012).
39. Marques, A.C. *et al.* Evidence for conserved post-transcriptional roles of unitary pseudogenes and for frequent bifunctionality of mRNAs. *Genome Biol* **13**, R102 (2012).
40. Gilad, Y, Man, O, Glusman, G. A comparison of the human and chimpanzee olfactory receptor gene repertoires. *Genome Res* **15**, 224-30 (2005).
41. Petrov, D.A. & Hartl, D.L. Pseudogene evolution and natural selection for a compact genome. *J Hered* **91**, 221-7 (2000).
42. Mariman, E. C. M. *et al.*Olfactory receptor genes cooperate with protocadherin genes in human extreme obesity. *Genes Nutr* **10**, 465 (2015).
43. Gonçalves, I, Duret, L, Mouchiroud, D. Nature and structure of human genes that generate retropseudogenes. *Genome Res* **10**, 672-8 (2000).
44. Wu, J. *et al.* The landscape of accessible chromatin in mammalian preimplantation embryos. *Nature* **534**, 652-7 (2016).
45. Ohshima, K. *et al.* Whole-genome screening indicates a possible burst of formation of processed pseudogenes and Alu repeats by particular L1 subfamilies in ancestral primates. *Genome Biol* **4**, R74 (2003).
46. Zhang, Z. & Gerstein, M. Large-scale analysis of pseudogenes in the human genome. *Curr Opin Genet Dev* **14**, 328-335 (2004).
47. Brouha, B. *et al.* Hot L1s account for the bulk of retrotransposition in the human population. *Proc*

Natl Acad Sci U S A **100**, 5280-5 (2003).
48. Goodier, J.L, Ostertag, E. M, Du, K, Kazazian, Jr, H. H. A novel active L1 retrotransposon subfamily in the mouse. *Genome Res* **11**, 1677-85 (2001).
49. Kolmogorov, M. *et al.* Chromosome assembly of large and complex genomes using multiple references. *BioRxiv* https://doi.org/10.1101/088435.
50. Klein, G. Toward a genetics of cancer resistance. *Proc Natl Acad Sci U S A* **106**, 859-63 (2009).
51. Liu, W. *et al.* Mutations in cytochrome c oxidase subunit VIa cause neurodegeneration and motor dysfunction in Drosophila. *Genetics* **176**, 937-46 (2007).
52. Oh, H. J, Choi, D, Goh, C. J, Hahn, Y. Loss of gene function and evolution of human phenotypes. *BMB Rep* **48**, 373-9 (2015).
53. Wang, T. *et al.* Identification and characterization of essential genes in the human genome. *Science* **350**, 1096-101 (2015).
54. Woods, S. *et al.* Duplication and retention biases of essential and non-essential genes revealed by systematic knockdown analyses. *PLoS Genet* **9**, e1003330 (2013).
55. Aubin-Houzelstein, G. & Panthier, J.J. The patchwork mouse phenotype: implication for melanocyte replacement in the hair follicle. *Pigment Cell Res* **12**, 181-6 (1999).
56. Oh, Y.T, Tran, D, Buchanan, T.A, Selsted, M.E, Youn, J.H. θ-Defensin RTD-1 improves insulin action and normalizes plasma glucose and FFA levels in diet-induced obese rats. *Am J Physiol Endocrinol Metab* **309**, E154-60 (2015).
57. Prats-Puig, A. *et al.* α-Defensins and bacterial/permeability-increasing protein as new markers of childhood obesity. *Pediatr Obes*, (2016).
58. Langergraber, K.E. *et al.* Generation times in wild chimpanzees and gorillas suggest earlier divergence times in great ape and human evolution. *Proc Natl Acad Sci U S A* **109**, 15716-21 (2012).
59. Vicens, A, Lüke, L, Roldan, E.R.S. Proteins involved in motility and sperm-egg interaction evolve more rapidly in mouse spermatozoa. *PLoS One* **9**, e91302 (2014).
60. Hayward, R. & Swanton, H. Congenital cardiovascular abnormalities. *Br J Hosp Med* **26**, 211-2, 217, 219-20 (1981).
61. Zheng, J. *et al.* mtDNA sequence, phylogeny and evolution of laboratory mice. *Mitochondrion* **17**, 126-31 (2014).
62. Baertsch, R, Diekhans, M, Kent, W. J, Haussler, D, Brosius, J. Retrocopy contributions to the evolution of the human genome. *BMC Genomics* **9**, 466 (2008).
63. Hickey, G, Paten, B, Earl, D, Zerbino, D, Haussler, D. HAL: a hierarchical format for storing and analyzing multiple genome alignments. *Bioinformatics* **29**, 1341-2 (2013).
64. Quinlan, A.R. BEDTools: The Swiss-Army Tool for Genome Feature Analysis. *Curr Protoc Bioinformatics* **47**, 11.12.1-34 (2014).
65. Edgar, R.C. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* **5**, 113 (2004).
66. http://www.geneious.com Last accessed on August, 21 2017.
67. Guo, Y, Mahony, S, Gifford, D.K. High resolution genome wide binding event finding and motif discovery reveals transcription factor spatial binding constraints. *PLoS Comput Biol* **8**, e1002638 (2012).
68. Kim, D. *et al.* TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol* **14**, R36 (2013).
69. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-9 (2009).
70. Trapnell, C. *et al.* Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat Protoc* **7**, 562-78 (2012).
71. http://repeatmasker.org Last accessed on August, 21 2017.
72. Durinck, S. *et al.* BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. *Bioinformatics* **21**, 3439-40 (2005).
73. Durinck, S, Spellman, P. T, Birney, E, Huber, W. Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nat Protoc* **4**, 1184-91 (2009).
74. Lex, A, Gehlenborg, N, Strobelt, H, Vuillemot, R, Pfister, H. UpSet: Visualization of Intersecting Sets. *IEEE Trans Vis Comput Graph* **20**, 1983-92 (2014).
75. Bennett, B.D. & Bushel, P. R. goSTAG: gene ontology subtrees to tag and annotate genes within

a set. *Source Code Biol Med* **12**, 6 (2017).
76. Greene, D, Richardson, S, Turro, E. ontologyX: a suite of R packages for working with ontological data. *Bioinformatics* **33**, 1104-1106 (2017).
77. Lam, H.Y.K. *et al.* Pseudofam: the pseudogene families database. *Nucleic Acids Res* **37**, D738-43 (2009).
78. Dickinson, M.E. *et al.* High-throughput discovery of novel developmental phenotypes. *Nature* **537**, 508-514 (2016).

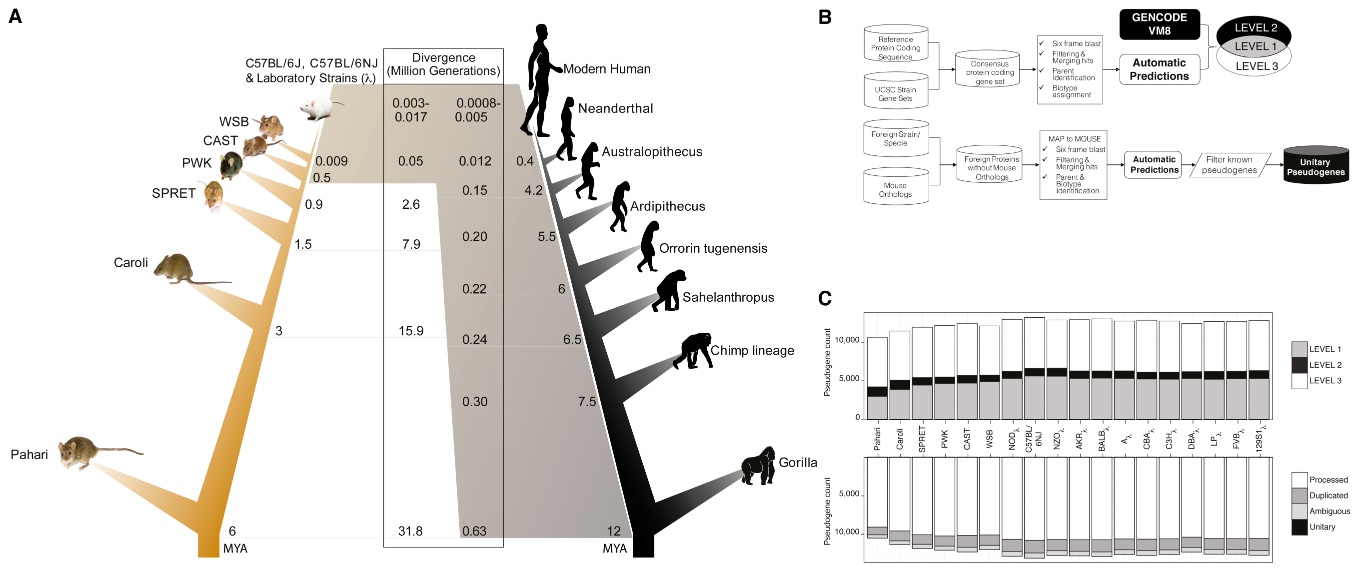

Figure 1. A – Human vs mouse lineage comparison. *MYA* – million years ago, λ – laboratory strain. B (top) – pseudogene annotation workflow for mouse strains. B (bottom) – unitary pseudogene annotation pipeline. C – Summary of mouse strains pseudogene annotation. *Level 1* are pseudogenes identified by automatic pipelines and lift over of manual annotation from the reference genome; *Level 2* are pseudogenes identified only through the lift over of manually annotated cases from the reference genome; *Level 3* are pseudogenes identified only by the automatic annotation pipeline.

Deleted: Column Break
Manual
PseudoPipe*
RetroFinder*
Union
ntersection
Mouse
... [2]

Deleted: Abbreviations

Deleted: B

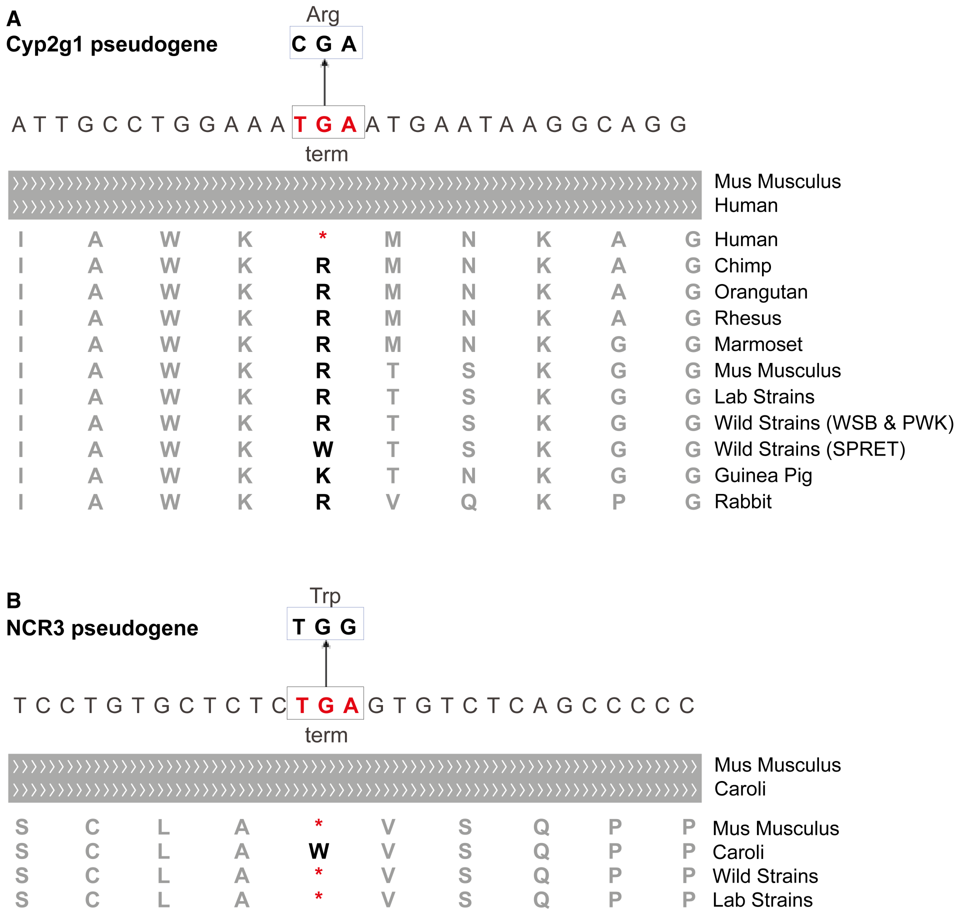

Figure 2 A – Cyp2G1 LOF in human. B – NCR3 gain of function mutation in Mus Caroli as compared to the reference genome and the other mouse strains.

Deleted: C (top) – pseudogene annotation workflow for mouse strains. C (bottom) – unitary pseudogene annotation pipeline. ... [3]

Deleted: ... [4]

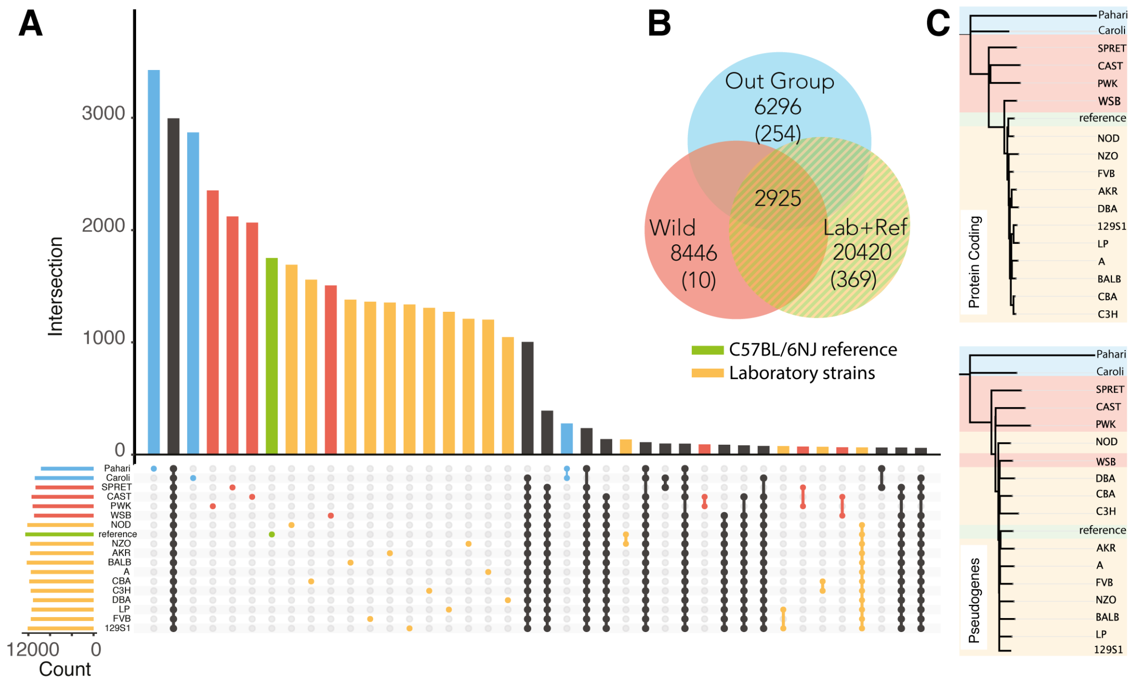

Figure 3. A – Summary of pseudogene distribution in the pangenome mouse strain dataset. The different group of mouse strains are highlighted by colours: blue relates to outgroup mice (Pahari and Caroli), red corresponds to wild-derived mice (SPRET, CAST, WSB, PWK), yellow indicates the laboratory inbred strains as listed in Table S2, and green highlights the laboratory inbred "reference" strain C57BL/6NJ. B – Venn diagram of evolutionarily conserved and group specific pseudogenes. The number in brackets is indicative of pseudogenes that are unique to each group. C – Phylogenetic trees for parents of evolutionarily conserved pseudogenes and evolutionary conserved pseudogenes.

Deleted: p

Deleted: ,

Deleted: , and pfam families of evolutionarily conserved pseudogenes: 7 trans membrane proteins (7TM), ribosomal proteins (Ribosome), cyclin-dependent kinases (CDK), olfactory receptor proteins (Olfr).

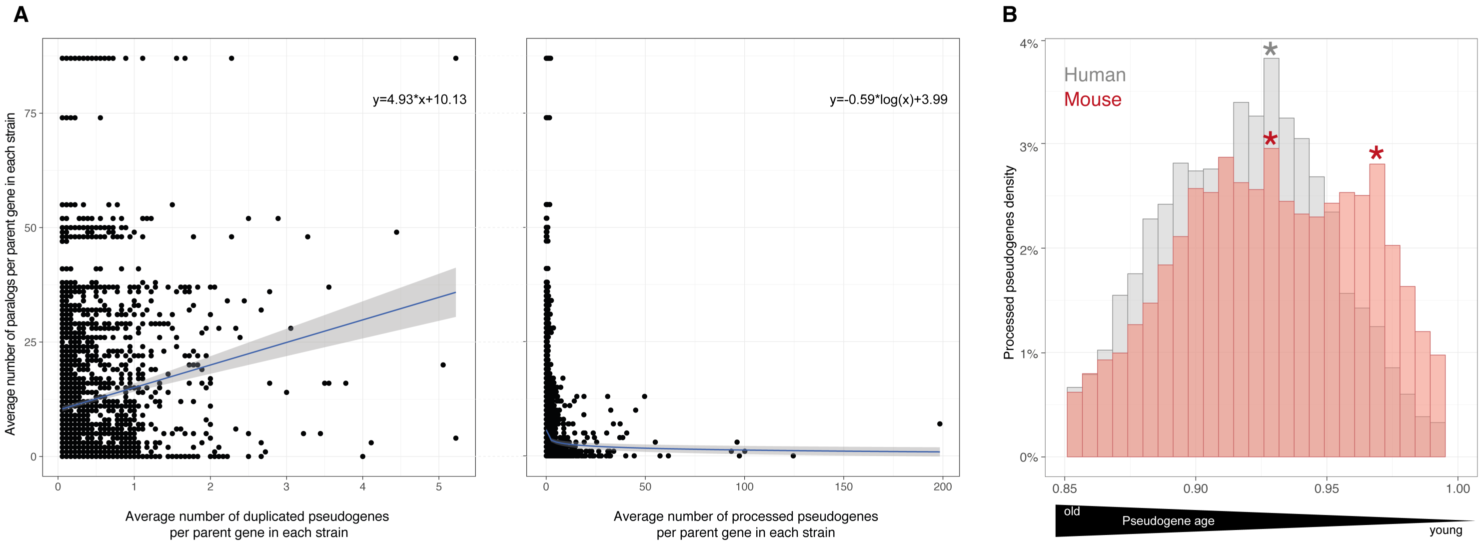

Figure 4. A – Relationship between the number of pseudogenes and functional paralogs for a given parent gene (left – duplicated pseudogenes, right – processed pseudogenes). Fitting lines show a vague correlation between the number of functional vs disabled copies of a gene, with a linear fit for duplicated pseudogenes and a negative logarithmic fit for processed pseudogene. The gray area is the standard deviation. B – Distribution of L1 flanked pseudogenes (y-axis) as function of age (x-axis). The pseudogene age is approximated as DNA sequence similarity to the parent gene.

Deleted: ... [5]

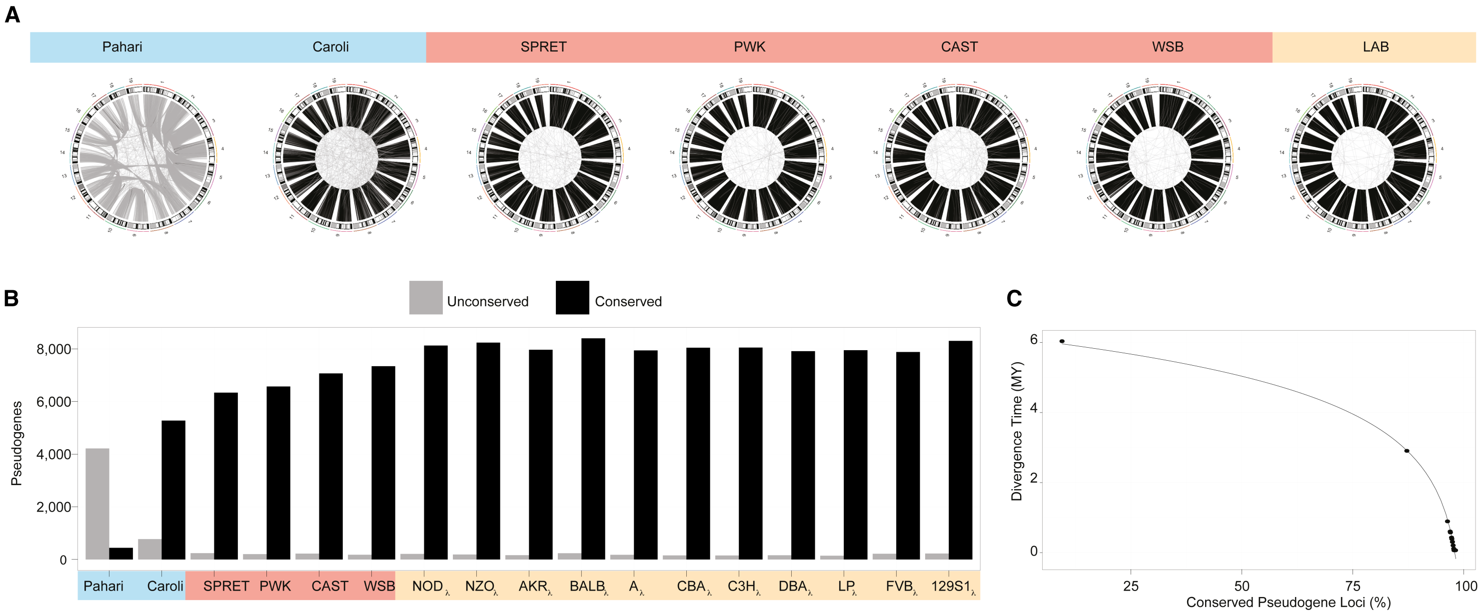

Figure 5. A – CIRCOS-like plots showing the conservation of the pseudogene genomic loci between each mouse strain and the laboratory reference strain C57BL/6NJ. Gray-lines indicate a change of the genomic locus between the two strains and connect two different genomic locations (e.g. a pseudogene located on chr7 in C57BL/6NJ and chr1 in Pahari). Black-lines indicate the conservation of the pseudogene locus. B – The numbers of pseudogenes that are preserved or have changed their loci between each strain and the laboratory reference strain. B – The numbers of pseudogenes that preserved or changed their loci between each strain and the laboratory reference strain. C – Strain speciation times as function of percentage of conserved pseudogene loci between each strain and the laboratory reference, fitted by an inverse logarithmic curve.

Deleted: ... [6]

Deleted: ; black

Deleted: B – the

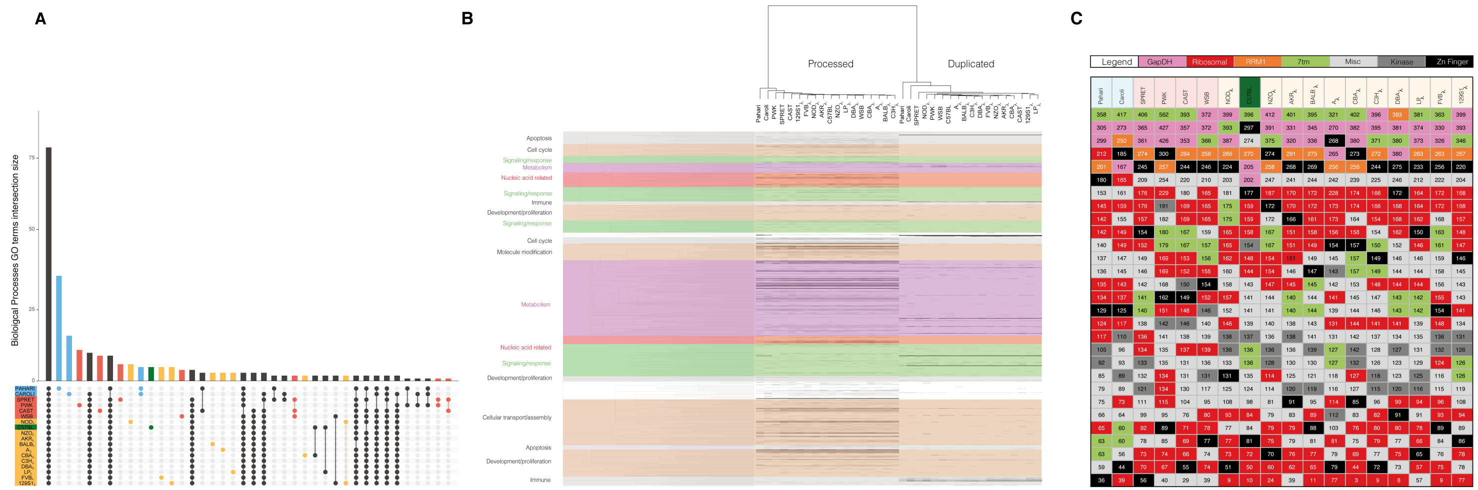

Figure 6. A – Distribution of enriched GO biological processes terms across the mouse strains. B – Heatmap illustrating enrichment of GO biological processes terms across the mouse strains for the parent genes of processed and duplicated pseudogenes. GO terms (rows) are clustered by semantic similarity (colour). Each line in the heat map indicates the presence of a pseudogene. C – Summary of the top 24 Pfam pseudogene families in each mouse strain.

Deleted: ... [7]

Deleted: .

Deleted:

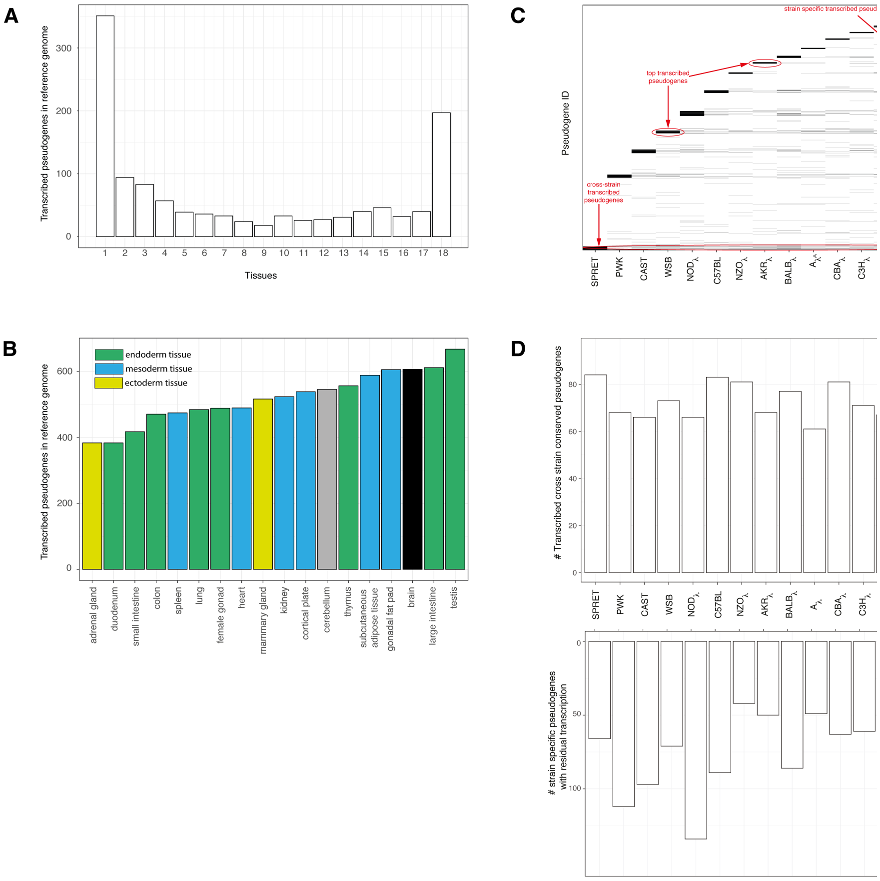

Figure 7. A – Cross tissue pseudogene transcription in the mouse reference genome. x-axis indicates the number of tissues a pseudogene is transcribed in. B – Distribution of pseudogene transcription in 18 adult mouse tissues. C – Number of transcribed pseudogenes in brain tissue for each wild-derived and laboratory mouse strain. D (top) – number of transcribed pseudogenes that are conserved across all the strains. D (bottom) – number of transcribed strain specific pseudogenes in each mouse strain.

The pseudogene annotation pipeline is freely available at http://pseudogene.org/pseudopipe . All supplementary data is available at http://mouse.pseudogene.org/Supplement/ . The 16 laboratory and wild strains (Table 2) assemblies and strain specific annotations were obtained from the Mouse Genome Project [36] (http://www.sanger.ac.uk/science/data/mouse-genomes-project, last accessed on 21.08.2017). The laboratory strain C57BL/6NJ is a subline of the reference strain [15] and is used here as the laboratory strain reference.

The two outgroup mouse species (Table 2RNAseq data from mouse adult brain was obtained from the Mouse Genome project for 12 laboratory, and 4 wild strains (ftp://ftp-mouse.sanger.ac.uk/REL-1509-Assembly-RNA-Seq), sanger experiment, last accessed on 21.08.2017). Nextmouse.pseudogene.org

- **Deleted:** in
- **Deleted:** ... [8]
- **Moved up [1]: Code and data availability**
- **Moved up [2]:** ... [9]
- **Moved up [3]:**), Mus Caroli and Mus Pahari were sequenced, assembled, and annotated in the protein-coding domain by ref. [37]. ... [10]
- **Deleted:** The divergence times for the wild
- **Moved up [4]:** and laboratory strains were obtained from [59, 60, 61]. The data for two outgroup species divergence times was obtained from [37]. The generation time for all the mice was estimated from [15]. ... [11]
- **Deleted: SF5**.
- **Moved up [5]:** ... [12]
- **Deleted: 1C**
- **Moved up [6]:** . The protein coding input set contains the conserved protein coding genes between each mouse 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. PseudoPipe was run with the strain conserved protein set as shown in **Figure**
- **Deleted: 1C**
- **Moved up [7]:** . Next, we used HAL tools package [63] to lift over the manually annotated pseudogenes from the mouse reference genome onto each strain using the U ... [13]
- **Deleted:**
- **Moved up [8]:** Unitary Pseudogene Annotation Pipel ... [14]
- **Deleted: 1C**
- **Moved up [9]:** . "Functional organism" is defined as ... [15]
- **Deleted:** "Non-functional" organism as the genome a ... [16]
- **Moved up [10]:** The resulting data set was subjected ... [17]
- **Deleted:** with a 1bp overlap minimum required. The ... [18]
- **Moved up [11]:** The remaining matches were subjec ... [19]
- **Deleted:** , conservation of pseudogene biotype, conse ... [20]
- **Moved up [12]:** Next, we integrated all filtered binar ... [21]
- **Deleted:** pangeome pseudogene reference.
- **Moved up [13]:** We obtained 49,262 pangenome ... [22]
- **Moved up [14]:** Phylogenetic analysis ... [23]
- **Moved up [15]:** using MUSCLE aligner [65] under ... [25]
- **Deleted:** , were extracted and assembled in a strain sp ... [24]
- **Deleted:** in available in **Table S7**
- **Moved up [16]:** . We estimated the pseudogene parer ... [26]
- **Moved up [17]: Gene essentiality enrichment analy** ... [28]
- **Deleted: S7**
- **Moved up [18]:**). The pseudogene sequences were f ... [29]
- **Deleted:** the associating the pseudogene with the prot ... [27]
- **Moved up [19]:** we created mappability maps for ea ... [30]
- **Moved up [20]:** Pseudogene annotation information ... [31]
- **Deleted:** Column Break ... [32]

Page 7: [1] Deleted	Paul Muir	5/13/18 7:54:00 PM

Furthermore, we grouped the conserved pseudogenes into subgroups based on their parent gene families (e.g. olfactory receptors, Ribosomal proteins, etc.) and constructed pseudogene phylogenetic trees for each of

Page 22: [2] Deleted	Paul Muir	5/13/18 7:54:00 PM

Column Break

Table 1. Reference genome pseudogene annotation in mouse and human.

Organism	Manual curation (M)	PseudoPipe* (PP)	RetroFinder* (RF)	Union PP&RF	Intersection M&PP (%)
Mouse	10,524	18,649	18,467	26,093	8,786 (83.5)
Human	14,650	15,978	15,474	22,396	13,177 (89.9)

*Chromosomal assembled DNA only

Table 2. Mouse strains description and nomenclature.

Strain ID	Description	Class
Pahari	PAHARI/EiJ – Mus Pahari	Outgroup
Caroli	CAROLI/EiJ – Mus Caroli	
SPRET	SPRET/EiJ – Mus Spretus	Wild strains
PWK	PWK/J – Mus Musculus Musculus	
CAST	CAST/EiJ – Must Castaneus	
WSB	WSB/J – Mus Musculus Domesticus	
NOD_{λ}	NOD/ShiLtJ – Mus Musculus Non-obese Diabetic	Lab Strains
C57BL	C57BL/6NJ – Mus Musculus Black 6N	
NZO_{λ}	NZO/HlLtJ – Mus Musculus New Zealand Obese	
AKR_{λ}	AKR/J – Mus Musculus	
$BALB_{\lambda}$	BALB/cJ – Mus Musculus	
A_{λ}	A/J – Mus Musculus	
CBA_{λ}	CBA/J – Mus Musculus	
$C3H_{\lambda}$	C3H/HeJ – Mus Musculus	
DBA_{λ}	DBA/2J – Mus Musculus	
LP_{λ}	LP/J – Mus Musculus	
FVB_{λ}	FVB/NJ – Mus Musculus	
$129S1_{\lambda}$	129S1/SvImJ – Mus Musculus	

Table 3. Enrichment of pseudogene parent gene class in essential genes.

Pseudogenes	Genes	Essential	Nonessential	Odds Ratio	p-Value
Total	Parent	1162	1061	1.93	$7.7*10^{-39}$
	Non-Parent	2050	3620		
Processed	Parent	1034	869	2.08	$2.3*10^{-43}$
	Non-Parent	2178	3812		
Duplicated	Parent	334	349	1.44	$6.0*10^{-6}$
	Non-Parent	2878	4332		

Page 23: [3] Deleted	Paul Muir	5/13/18 7:54:00 PM

C (top) – pseudogene annotation workflow for mouse strains. C (bottom) – unitary pseudogene annotation pipeline.

Page 23: [4] Deleted	Paul Muir	5/13/18 7:54:00 PM

Page 25: [5] Deleted	Paul Muir	5/13/18 7:54:00 PM

Page 26: [6] Deleted	Paul Muir	5/13/18 7:54:00 PM

Page 27: [7] Deleted	Paul Muir	5/13/18 7:54:00 PM

Page 1: [8] Deleted	Paul Muir	5/13/18 7:54:00 PM

Column Break

Page 1: [9] Moved to page 13 (Move #2)	Paul Muir	5/13/18 7:54:00 PM

Datasets

Mouse reference genome is based on the Mus Musculus strain C57BL/6J strain. The mouse reference annotation is based on GENCODE vM12/Ensembl 87.

The human reference genome annotation is based on GENCODE v25/Ensembl 87.

Page 1: [10] Moved to page 13 (Move #3)	Paul Muir	5/13/18 7:54:00 PM

), Mus Caroli and Mus Pahari were sequenced, assembled, and annotated in the protein-coding domain by ref. [37].

Human – Mouse Lineage Comparison

Human – primate lineage divergence and generation times were obtained from [58].

Page 1: [11] Moved to page 14 (Move #4)	Paul Muir	5/13/18 7:54:00 PM

and laboratory strains were obtained from [59, 60, 61]. The data for two outgroup species divergence times was obtained from [37]. The generation time for all the mice was estimated from [15].

Pseudogene Annotation

Reference genome annotation

We manually curated 10,524 pseudogenes in the mouse reference genome (GENCODE M12) and 14,650 pseudogenes in the human reference genome (GENCODE v25), using a workflow previously described in [33, 34]. The manual annotation is based on the sequence homology to protein data from UniProt database [34] and the protocol is summarised in **Figure**

Page 1: [12] Moved to page 14 (Move #5)	Paul Muir	5/13/18 7:54:00 PM

The number of manually annotated pseudogenes in the mouse lineage is likely an underestimate of the true size of the mouse pseudogene complement given the similarities between the human and mouse genomes. Thus, to get a more accurate idea of the number of pseudogenes in the mouse genome, we used a combination of two automatic annotation pipelines: PseudoPipe [35] and RetroFinder [62]. PseudoPipe is a comprehensive annotation pipeline focused on identifying and characterizing pseudogenes based on their biotypes as either processed or duplicated. The automatic annotation workflow using PseudoPipe is summarised in **Figure 1B** and has been previously described in detail in [33, 34, 35]. Pseudopipe identifies 22,811 mouse pseudogenes of which 14,084 are present in autosomal chromosomes (a number comparable with the one observed previously in human (**Table S1**)). RetroFinder is computational annotation pipeline focused on identifying retrotransposed genes and pseudogenes. Using RetroFinder we were able to annotate 18,467 and respectively 15,474 processed pseudogenes in mouse and human. There is a good overlap between the two identification pipelines with respect to the number of processed pseudogenes present in both organisms (**Table S1**).

Mouse strain annotation

The mouse strain pseudogene annotation workflow is summarised in **Figure**

Page 1: [13] Moved to page 14 (Move #7)	Paul Muir	5/13/18 7:54:00 PM

. Next, we used HAL tools package [63] to lift over the manually annotated pseudogenes from the mouse reference genome onto each strain using the UCSC multi strain sequence alignments. We merged the two annotation sets using BEDTools [64] with 1bp minimum overlap requirement. We extended each overlap predicted boundaries to ensure full annotation of the pseudogene transcript.

Page 1: [14] Moved to page 15 (Move #8)	Paul Muir	5/13/18 7:54:00 PM

Unitary Pseudogene Annotation Pipeline

We modified PseudoPipe to allow cross-strains and cross species protein coding inputs. We annotated cross-organism pseudogenes as shown in **Figure**

Page 1: [15] Moved to page 15 (Move #9)	Paul Muir	5/13/18 7:54:00 PM

. "Functional organism" is defined as the genome providing the protein coding information and thus containing a working copy of the element of interest.

Page 1: [16] Deleted	Paul Muir	5/13/18 7:54:00 PM

"Non-functional" organism as the genome analysed for unitary pseudogene presence.

Page 1: [17] Moved to page 15 (Move #10)	Paul Muir	5/13/18 7:54:00 PM

The resulting data set was subjected to a number of filters such as removal of previously known pseudogenes, removal of pseudogenes with parents that have orthologs in the annotated specie, removal of pseudogenes that overlap with annotated protein coding and ncRNAs loci, and removal of pseudogenes shorter than 100 bp. The filtered PseudoPipe set was intersected with the lift-over of the protein coding annotation from the "functional organism using BEDTools [64]

Page 1: [18] Deleted	**Paul Muir**	**5/13/18 7:54:00 PM**

with a 1bp overlap minimum required. The intersection set was further refined flagging protein coding genes that have functional relatives (paralogs) in the non-"functional organism".

Page 1: [19] Moved to page 15 (Move #11)	**Paul Muir**	**5/13/18 7:54:00 PM**

The remaining matches were subjected to manual inspection of the alignment.

Conservation and divergence in pseudogene complements

Pangenome data set generation

We performed an all against all liftover of pseudogene annotation using HAL tools package and the UCSC multi strain sequence alignment. Each liftover was intersected with the know strain annotation and all the entries that matched protein coding or ncRNAs were removed. The resulting set is further filtered for conservation of pseudogene Ensembl ID, where available (used for Level 1 and 2 pseudogenes), conservation of parent gene identity, conservation of pseudogene locus

Page 1: [20] Deleted	**Paul Muir**	**5/13/18 7:54:00 PM**

, conservation of pseudogene biotype, conservation of pseudogene length, and conservation of pseudogene structure.

Page 1: [21] Moved to page 15 (Move #12)	**Paul Muir**	**5/13/18 7:54:00 PM**

Next, we integrated all filtered binary mappings in a master pan-strain set. The common entries were collapsed into a unique

Page 1: [22] Moved to page 15 (Move #13)	**Paul Muir**	**5/13/18 7:54:00 PM**

We obtained 49,262 pangenome pseudogenes. 1,158 pangenome entries are multi matching across strains.

Page 1: [23] Moved to page 15 (Move #14)	**Paul Muir**	**5/13/18 7:54:00 PM**

Phylogenetic analysis

Sequences of the 1,460 pseudogenes were randomly selected out of the total of 2925 conserved pseudogenes in the 18 mouse strains accounting for approximately 50% of the total number of conserved pseudogenes

Page 1: [24] Deleted	**Paul Muir**	**5/13/18 7:54:00 PM**

, were extracted and assembled in a strain specific contig. The multi-sequence alignment of the 18 contigs was obtained

Page 1: [25] Moved to page 16 (Move #15)	**Paul Muir**	**5/13/18 7:54:00 PM**

using MUSCLE aligner [65] under standard conditions. Similarly, the sequences of parent protein coding genes of the 1,460 pseudogenes were assembled into a strain specific sequence and aligned using MUSCLE. The tree was generated using Tamura-Nei genetic distance model and neighbouring-joining tree build method with Pahari as outgroup using GENEIOUS 10.2 software package [66].

Genome evolution and plasticity

Genome mappability maps

We created mappabilty maps for the mouse reference genome and the 18 mouse strains using the GEM library [67]. The workflow is composed of indexing the genome using gem-indexer, followed by creation of the map using a window of 75 nucleotides under the following conditions -m 0.02 -T 2.

Parent gene expression analysis

RNAseq mouse tissue data was obtained from ENCODE. The complete list of experiments used

Page 1: [26] Moved to page 16 (Move #16)	Paul Muir	5/13/18 7:54:00 PM

. We estimated the pseudogene parent protein coding genes expression levels using a workflow involving the following steps: filtering the protein coding genes for uniquely mappable regions longer than 100bp, mapping reads using TopHat2 [68], selecting high quality mapped reads with a quality score higher than 30 using samtools [69], and calculating the expression FPKM levels using Cufflinks [70]. Transcriptional activity of pseudogene parent genes during early embryonic development was investigated using RNAseq data as processed and described in [44]. Raw sequencing data and processed data containing FPKM levels at each embryonic stage are available on the SRA under Series GSE66582.

Transposable elements analysis

TE in human and mouse reference genomes were informed from RepeatMasker libraries Repbase 21.11 and using RepeatMasker 3.2.8 [71]. We extracted all the four major groups of repeats SINE, LINE, LTR and DNA and identified all the processed pseudogenes associated with L1 elements. Next, we binned the L1 annotated pseudogenes into age groups based on their sequence similarity to the parent gene, with younger elements exhibiting a higher sequence similarity while older elements show a large sequence divergence when compared to the functional gene counterparts.

Gene ontology and Pfam analysis

Linking of gene ontology terms to the pseudogene parent genes was conducted using the R package biomaRt [72, 73]. Visualization of shared and distinct GO term sets amongst the strains was done using the R package UpSetR [74]. Enrichment of GO terms amongst the pseudogene parent genes and clustering of mouse strains based on similar enrichment profiles was performed using the goSTAG software package [75]. Semantic clustering of the GO terms was done with the OntologyX packages [76]. Parent genes were labelled with both strain and biotype information in order to better evaluate differences in the pseudogene complements based on their mechanism of creation.

Analysis of the Pfam representation in the pseudogene complements was performed as previously described in [77] and focused on

Page 1: [27] Deleted	Paul Muir	5/13/18 7:54:00 PM

the associating the pseudogene with the protein family of its parent gene.

Page 1: [28] Moved to page 17 (Move #17)	Paul Muir	5/13/18 7:54:00 PM

Gene essentiality enrichment analysis

Lists of essential and nonessential genes were compiled using data from the MGI database and recent work from the International Mouse Phenotyping Consortium [78]. The nonessential gene set with Ensembl identifiers contained 4,736 genes compared to 3,263 essential genes.

In order to evaluate the impact of parent gene status on the probability of a gene being essential while controlling for transcription we fit a linear probability model and a probit model for the probability that a gene is essential given its transcription level and parent gene status using the StatsModels package in Python. The linear probability model fits an ordinary least squares regression of gene essentiality on parent gene status and transcription level. While the linear probability model generally estimates relationships well close to the mean of the independent variables, it often loses explanatory power at low and high values of these variables. Because of this deficiency, we looked also at the probit model, which is similar to the linear probability model but instead fits the data to a cumulative Gaussian distribution. Around the mean values, we find that parent gene status increases the probability of essentiality by around 20% in both models.

Pseudogene transcription

We estimated the pseudogene transcription levels for the mouse reference in 18 adult tissues following a similar protocol to the one described earlier for calculating the expression of protein coding genes, a method that we have successfully used in the past [34] using RNAseq ENCODE data (**Table**

Page 1: [29] Moved to page 17 (Move #18)	Paul Muir	5/13/18 7:54:00 PM

). The pseudogene sequences were filtered for uniquely mappable exon regions longer than 100 bp. Next the RNAseq raw data was mapped using TopHat and the mapped reads were filtered for quality scores higher than 30. The resulting alignments were quantified using Cufflinks. A pseudogene was considered transcribed if it had an FPKM larger than 3.3 in accord with previous studies [34].

Page 1: [30] Moved to page 17 (Move #19)	Paul Muir	5/13/18 7:54:00 PM

we created mappability maps for each of the 16 mouse strains genomes and selected only the pseudogene exons in uniquely mappable regions and longer than 100bp for further transcription analysis. The pseudogene transcription levels in mouse strains were estimated using a similar workflow as described above. The transcription cut off level was set to 1.

Mouse pseudogene resource

All the annotation data produced in the analysis is collected and available online through

Page 1: [31] Moved to page 17 (Move #20)	Paul Muir	5/13/18 7:54:00 PM

Pseudogene annotation information encompasses the genomic context of each pseudogene, its parent gene and transcript Ensembl IDs, the corresponding mouse reference pseudogene Ensembl ID, 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 onto another and subsequent intersection with that strain's native annotations. The database provides liftover annotations and information about intersections between the liftover and native annotations. Furthermore, homology information provide links between the well-characterized mouse strain collection.

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.

Page 1: [32] Deleted	Paul Muir	5/13/18 7:54:00 PM

Column Break

1. Peters, L. L., Robledo, R. F., Bult, C. J., Churchill, G. A., Paigen, B. J., & Svenson, K. L. The mouse as a model for human biology: a resource guide for complex trait analysis. *Nat Rev Genet* **8**, 58-69 (2007).
2. Paigen, K. One hundred years of mouse genetics: an intellectual history. I. The classical period (1902-1980). *Genetics* **163**, 1-7 (2003).
3. Paigen, K. One hundred years of mouse genetics: an intellectual history. II. The molecular revolution (1981-2002). *Genetics* **163**, 1227-35 (2003).
4. Yalcin, B., Adams, D. J., Flint, J., & Keane, T. M. Next-generation sequencing of experimental mouse strains. *Mamm Genome* **23**, 490-8 (2012).
5. Keane, T. M., Goodstadt, L., Danecek, P., White, M. A., Wong, K., Yalcin, B., Heger, A., Agam, A., Slater, G., Goodson, M., Furlotte, N. A., Eskin, E., Nellåker, C., Whitley, H., Cleak, J., Janowitz, D., Hernandez-Pliego, P., Edwards, A., Belgard, T. G., Oliver, P. L., McIntyre, R. E., Bhomra, A., Nicod, J., Gan, X., Yuan, W., van der Weyden, L., Steward, C. A., Bala, S., Stalker, J., Mott, R., Durbin, R., Jackson, I. J., Czechanski, A., Guerra-Assunção, J. A., Donahue, L. R., Reinholdt, L. G., Payseur, B. A., Ponting, C. P., Birney, E., Flint, J., & Adams, D. J. Mouse genomic variation and its effect on phenotypes and gene regulation. *Nature* **477**, 289-94 (2011).
6. Mestas, J. & Hughes, C. C. W. Of mice and not men: differences between mouse and human immunology. *J Immunol* **172**, 2731-8 (2004).
7. Emes, R. D., Goodstadt, L., Winter, E. E., & Ponting, C. P. Comparison of the genomes of human and mouse lays the foundation of genome zoology. *Hum Mol Genet* **12**, 701-9 (2003).
8. Mouse Genome Sequencing Consortium, Waterston, R. H., Lindblad-Toh, K., Birney, E., Rogers, J., Abril, J. F., Agarwal, P., Agarwala, R., Ainscough, R., Alexandersson, M., An, P., Antonarakis, S. E., Attwood, J., Baertsch, R., Bailey, J., Barlow, K., Beck, S., Berry, E., Birren, B., Bloom, T., Bork, P., Botcherby, M., Bray, N., Brent, M. R., Brown, D. G., Brown, S. D., Bult, C., Burton, J., Butler, J., Campbell, R. D., Carninci, P., Cawley, S., Chiaromonte, F., Chinwalla, A. T., Church, D. M., Clamp, M., Clee, C., Collins, F. S., Cook, L. L., Copley, R. R., Coulson, A., Couronne, O., Cuff, J., Curwen, V., Cutts, T., Daly, M., David, R., Davies, J., Delehaunty, K. D., Deri, J., Dermitzakis, E. T., Dewey, C., Dickens, N. J., Diekhans, M., Dodge, S., Dubchak, I., Dunn, D. M., Eddy, S. R., Elnitski, L., Emes, R. D., Eswara, P., Eyras, E., Felsenfeld, A., Fewell, G. A., Flicek, P., Foley, K., Frankel, W. N., Fulton, L. A., Fulton, R. S., Furey, T. S., Gage, D., Gibbs, R. A., Glusman, G., Gnerre, S., Goldman, N., Goodstadt, L., Grafham, D., Graves, T. A., Green, E. D., Gregory, S., Guigó, R., Guyer, M., Hardison, R. C., Haussler, D., Hayashizaki, Y., Hillier, L. W., Hinrichs, A., Hlavina, W., Holzer, T., Hsu, F., Hua, A., Hubbard, T., Hunt, A., Jackson, I., Jaffe, D. B., Johnson, L. S., Jones, M., Jones, T. A., Joy, A., Kamal, M., Karlsson, E. K., Karolchik, D., Kasprzyk, A., Kawai, J., Keibler, E., Kells, C., Kent, W. J., Kirby, A., Kolbe, D. L., Korf, I., Kucherlapati, R. S., Kulbokas, E. J., Kulp, D., Landers, T., Leger, J. P., Leonard, S., Letunic, I., Levine, R., Li, J., Li, M., Lloyd, C., Lucas, S., Ma, B., Maglott, D. R., Mardis, E. R., Matthews, L., Mauceli, E., Mayer, J. H., McCarthy, M., McCombie, W. R., McLaren, S., McLay, K., McPherson, J. D., Meldrim, J., Meredith, B.,

Mesirov, J. P., Miller, W., Miner, T. L., Mongin, E., Montgomery, K. T., Morgan, M., Mott, R., Mullikin, J. C., Muzny, D. M., Nash, W. E., Nelson, J. O., Nhan, M. N., Nicol, R., Ning, Z., Nusbaum, C., O'Connor, M. J., Okazaki, Y., Oliver, K., Overton-Larty, E., Pachter, L., Parra, G., Pepin, K. H., Peterson, J., Pevzner, P., Plumb, R., Pohl, C. S., Poliakov, A., Ponce, T. C., Ponting, C. P., Potter, S., Quail, M., Reymond, A., Roe, B. A., Roskin, K. M., Rubin, E. M., Rust, A. G., Santos, R., Sapojnikov, V., Schultz, B., Schultz, J., Schwartz, M. S., Schwartz, S., Scott, C., Seaman, S., Searle, S., Sharpe, T., Sheridan, A., Shownkeen, R., Sims, S., Singer, J. B., Slater, G., Smit, A., Smith, D. R., Spencer, B., Stabenau, A., Stange-Thomann, N., Sugnet, C., Suyama, M., Tesler, G., Thompson, J., Torrents, D., Trevaskis, E., Tromp, J., Ucla, C., Ureta-Vidal, A., Vinson, J. P., Von Niederhausern, A. C., Wade, C. M., Wall, M., Weber, R. J., Weiss, R. B., Wendl, M. C., West, A. P., Wetterstrand, K., Wheeler, R., Whelan, S., Wierzbowski, J., Willey, D., Williams, S., Wilson, R. K., Winter, E., Worley, K. C., Wyman, D., Yang, S., Yang, S.-P., Zdobnov, E. M., Zody, M. C., & Lander, E. S. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**, 520-62 (2002).

9. Madsen, O., Scally, M., Douady, C. J., Kao, D. J., DeBry, R. W., Adkins, R., Amrine, H. M., Stanhope, M. J., de Jong, W. W., & Springer, M. S. Parallel adaptive radiations in two major clades of placental mammals. *Nature* **409**, 610-4 (2001).
10. Murphy, W. J., Eizirik, E., Johnson, W. E., Zhang, Y. P., Ryder, O. A., & O'Brien, S. J. Molecular phylogenetics and the origins of placental mammals. *Nature* **409**, 614-8 (2001).
11. Hedges, S. B., Dudley, J., & Kumar, S. TimeTree: a public knowledge-base of divergence times among organisms. *Bioinformatics* **22**, 2971-2 (2006).
12. Monaco, G., van Dam, S., Casal Novo Ribeiro, J. L., Larbi, A., & de Magalhães, J. P. A comparison of human and mouse gene co-expression networks reveals conservation and divergence at the tissue, pathway and disease levels. *BMC Evol Biol* **15**, 259 (2015).
13. Goios, A., Pereira, L., Bogue, M., Macaulay, V., & Amorim, A. mtDNA phylogeny and evolution of laboratory mouse strains. *Genome Res* **17**, 293-8 (2007).
14. Beck, J. A., Lloyd, S., Hafezparast, M., Lennon-Pierce, M., Eppig, J. T., Festing, M. F., & Fisher, E. M. Genealogies of mouse inbred strains. *Nat Genet* **24**, 23-5 (2000).
15. http://www.informatics.jax.org/mgihome/other/homepage_IntroMouse.shtml. Last accessed on February, 24th 2017.
16. Richardson, A., Heydari, A. R., Morgan, W. W., Nelson, J. F., Sharp, Z. D., & Walter, C. A. Use of Transgenic Mice in Aging Research. *ILAR J* **38**, 125-136 (1997).
17. Troublesome variability in mouse studies. *Nat Neurosci* **12**, 1075 (2009).
18. Mighell, A. J., Smith, N. R., Robinson, P. A., & Markham, A. F. Vertebrate pseudogenes. *FEBS Lett* **468**, 109-14 (2000).
19. Harrison, P. M., Echols, N., & Gerstein, M. B. Digging for dead genes: an analysis of the characteristics of the pseudogene population in the Caenorhabditis elegans genome. *Nucleic Acids Res* **29**, 818--830 (2001).
20. Echols, N., Harrison, P., Balasubramanian, S., Luscombe, N. M., Bertone, P., Zhang, Z., & Gerstein, M. Comprehensive analysis of amino acid and nucleotide composition in eukaryotic genomes, comparing genes and pseudogenes. *Nucleic Acids Res* **30**, 2515--2523 (2002).
21. Balakirev, E. S. & Ayala, F. J. Pseudogenes: are they "junk" or functional DNA? *Annu Rev Genet* **37**, 123-51 (2003).
22. Zhang, Z. D., Frankish, A., Hunt, T., Harrow, J., & Gerstein, M. Identification and analysis of unitary pseudogenes: historic and contemporary gene losses in humans and other primates. *Genome Biol* **11**, R26 (2010).
23. Moore, R. C. & Purugganan, M. D. The early stages of duplicate gene evolution. *Proc Natl Acad Sci U S A* **100**, 15682-7 (2003).
24. Kuang, M. C., Hutchins, P. D., Russell, J. D., Coon, J. J., & Hittinger, C. T. Ongoing resolution of duplicate gene functions shapes the diversification of a metabolic network. *Elife* **5**, (2016).
25. Shakhnovich, B. E. & Koonin, E. V. Origins and impact of constraints in evolution of gene families. *Genome Res* **16**, 1529-36 (2006).

26. Ohno, S. Evolution by Gene Duplication. *Springer, New York* , (1970).
27. Kondrashov, F. A., Rogozin, I. B., Wolf, Y. I., & Koonin, E. V. Selection in the evolution of gene duplications. *Genome Biol* **3**, RESEARCH0008 (2002).
28. Rogozin, I. B. Complexity of gene expression evolution after duplication: protein dosage rebalancing. *Genet Res Int* **2014**, 516508 (2014).
29. http://www.informatics.jax.org/silver/chapters/1-3.shtml Last accessed on February, 24th 2017.
30. Cameron, J., Holla, Ø. L., Berge, K. E., Kulseth, M. A., Ranheim, T., Leren, T. P., & Laerdahl, J. K. Investigations on the evolutionary conservation of PCSK9 reveal a functionally important protrusion. *FEBS J* **275**, 4121-33 (2008).
31. Maxwell, K. N., Fisher, E. A., & Breslow, J. L. Overexpression of PCSK9 accelerates the degradation of the LDLR in a post-endoplasmic reticulum compartment. *Proc Natl Acad Sci U S A* **102**, 2069-74 (2005).
32. Dadu, R. T. & Ballantyne, C. M. Lipid lowering with PCSK9 inhibitors. *Nat Rev Cardiol* **11**, 563-75 (2014).
33. Pei, B., Sisu, C., Frankish, A., Howald, C., Habegger, L., Mu, X. J., Harte, R., Balasubramanian, S., Tanzer, A., Diekhans, M., Reymond, A., Hubbard, T. J., Harrow, J., & Gerstein, M. B. The GENCODE pseudogene resource. *Genome Biol* **13**, R51 (2012).
34. Sisu, C., Pei, B., Leng, J., Frankish, A., Zhang, Y., Balasubramanian, S., Harte, R., Wang, D., Rutenberg-Schoenberg, M., Clark, W., Diekhans, M., Rozowsky, J., Hubbard, T., Harrow, J., & Gerstein, M. B. Comparative analysis of pseudogenes across three phyla. *Proc Natl Acad Sci U S A* **111**, 13361-6 (2014).
35. Zhang, Z., Carriero, N., Zheng, D., Karro, J., Harrison, P. M., & Gerstein, M. PseudoPipe: an automated pseudogene identification pipeline. *Bioinformatics* **22**, 1437-9 (2006).
36. The Mouse Genomes Project Consortium. http://www.sanger.ac.uk/science/data/mouse-genomes-project. Last accessed on August, 21 2017
37. Thybert, D., Roller, M., Navarro, F. C. P., Fiddes, I., Streeter, I., Feig, C., Martin-Galvez, D., Kolmogorov, M., Janoušek, V., Akanni, W., Aken, B., Aldridge, S., Chakrapani, V., Clarke, L., Cummins, C., Doran, A., Dunn, M., Goodstadt, L., Howe, K., Howell, M., Josselin, A.-A., Karn, R. C., Lauketis, C. M., Jingtao, L., Martin, F., Muffato, M., Quail, M., Sisu, C., Stanke, M., Stefflova, K., Oosterhout, C. V., Veyrunes, F., Ward, B., Yang, F., Yazdanifar, G., Zadissa, A., Adams, D., Brazma, A., Gerstein, M., Paten, B., Pham, S., Keane, T., Odom, D. T., & Flicek, P. Repeat associated mechanisms of genome evolution and function revealed by the Mus caroli and Mus pahari genomes. *BioRxiv* https://doi.org/10.1101/158659.
38. Bryant, C. E. & Monie, T. P. Mice, men and the relatives: cross-species studies underpin innate immunity. *Open Biol* **2**, 120015 (2012).
39. Marques, A. C., Tan, J., Lee, S., Kong, L., Heger, A., & Ponting, C. P. Evidence for conserved post-transcriptional roles of unitary pseudogenes and for frequent bifunctionality of mRNAs. *Genome Biol* **13**, R102 (2012).
40. Gilad, Y., Man, O., & Glusman, G. A comparison of the human and chimpanzee olfactory receptor gene repertoires. *Genome Res* **15**, 224-30 (2005).
41. Petrov, D. A. & Hartl, D. L. Pseudogene evolution and natural selection for a compact genome. *J Hered* **91**, 221-7 (2000).
42. Mariman, E. C. M., Szklarczyk, R., Bouwman, F. G., Aller, E. E. J. G., van Baak, M. A., & Wang, P. Olfactory receptor genes cooperate with protocadherin genes in human extreme obesity. *Genes Nutr* **10**, 465 (2015).
43. Gonçalves, I., Duret, L., & Mouchiroud, D. Nature and structure of human genes that generate retropseudogenes. *Genome Res* **10**, 672-8 (2000).
44. Wu, J., Huang, B., Chen, H., Yin, Q., Liu, Y., Xiang, Y., Zhang, B., Liu, B., Wang, Q., Xia, W., Li, W., Li, Y., Ma, J., Peng, X., Zheng, H., Ming, J., Zhang, W., Zhang, J., Tian, G., Xu, F., Chang, Z., Na, J., Yang, X., & Xie, W. The landscape of accessible chromatin in mammalian preimplantation embryos. *Nature* **534**, 652-7 (2016).
45. Ohshima, K., Hattori, M., Yada, T., Gojobori, T., Sakaki, Y., & Okada, N. Whole-genome screening indicates a possible burst of formation of processed pseudogenes and Alu repeats by particular L1

subfamilies in ancestral primates. *Genome Biol* **4**, R74 (2003).
46. Zhang, Z. & Gerstein, M. Large-scale analysis of pseudogenes in the human genome. *Curr Opin Genet Dev* **14**, 328--335 (2004).
47. Brouha, B., Schustak, J., Badge, R. M., Lutz-Prigge, S., Farley, A. H., Moran, J. V., & Kazazian, Jr, H. H. Hot L1s account for the bulk of retrotransposition in the human population. *Proc Natl Acad Sci U S A* **100**, 5280-5 (2003).
48. Goodier, J. L., Ostertag, E. M., Du, K., & Kazazian, Jr, H. H. A novel active L1 retrotransposon subfamily in the mouse. *Genome Res* **11**, 1677-85 (2001).
49. Kolmogorov, M., Armstrong, J., Raney, B. J., Streeter, I., Dunn, M., Yang, F., Odom, D., Flicek, P., Keane, T., Thybert, D., Paten, B., & Pham, S. Chromosome assembly of large and complex genomes using multiple references. *BioRxiv* https://doi.org/10.1101/088435.
50. Klein, G. Toward a genetics of cancer resistance. *Proc Natl Acad Sci U S A* **106**, 859-63 (2009).
51. Liu, W., Gnanasambandam, R., Benjamin, J., Kaur, G., Getman, P. B., Siegel, A. J., Shortridge, R. D., & Singh, S. Mutations in cytochrome c oxidase subunit VIa cause neurodegeneration and motor dysfunction in Drosophila. *Genetics* **176**, 937-46 (2007).
52. Oh, H. J., Choi, D., Goh, C. J., & Hahn, Y. Loss of gene function and evolution of human phenotypes. *BMB Rep* **48**, 373-9 (2015).
53. Wang, T., Birsoy, K., Hughes, N. W., Krupczak, K. M., Post, Y., Wei, J. J., Lander, E. S., & Sabatini, D. M. Identification and characterization of essential genes in the human genome. *Science* **350**, 1096-101 (2015).
54. Woods, S., Coghlan, A., Rivers, D., Warnecke, T., Jeffries, S. J., Kwon, T., Rogers, A., Hurst, L. D., & Ahringer, J. Duplication and retention biases of essential and non-essential genes revealed by systematic knockdown analyses. *PLoS Genet* **9**, e1003330 (2013).
55. Aubin-Houzelstein, G. & Panthier, J. J. The patchwork mouse phenotype: implication for melanocyte replacement in the hair follicle. *Pigment Cell Res* **12**, 181-6 (1999).
56. Oh, Y. T., Tran, D., Buchanan, T. A., Selsted, M. E., & Youn, J. H. θ-Defensin RTD-1 improves insulin action and normalizes plasma glucose and FFA levels in diet-induced obese rats. *Am J Physiol Endocrinol Metab* **309**, E154-60 (2015).
57. Prats-Puig, A., Gispert-Saüch, M., Carreras-Badosa, G., Osiniri, I., Soriano-Rodríguez, P., Planella-Colomer, M., de Zegher, F., Ibánez, L., Bassols, J., & López-Bermejo, A. α-Defensins and bacterial/permeability-increasing protein as new markers of childhood obesity. *Pediatr Obes* , (2016).
58. Langergraber, K. E., Prüfer, K., Rowney, C., Boesch, C., Crockford, C., Fawcett, K., Inoue, E., Inoue-Muruyama, M., Mitani, J. C., Muller, M. N., Robbins, M. M., Schubert, G., Stoinski, T. S., Viola, B., Watts, D., Wittig, R. M., Wrangham, R. W., Zuberbühler, K., Pääbo, S., & Vigilant, L. Generation times in wild chimpanzees and gorillas suggest earlier divergence times in great ape and human evolution. *Proc Natl Acad Sci U S A* **109**, 15716-21 (2012).
59. Vicens, A., Lüke, L., & Roldan, E. R. S. Proteins involved in motility and sperm-egg interaction evolve more rapidly in mouse spermatozoa. *PLoS One* **9**, e91302 (2014).
60. Hayward, R. & Swanton, H. Congenital cardiovascular abnormalities. *Br J Hosp Med* **26**, 211-2, 217, 219-20 (1981).
61. Zheng, J., Chen, Y., Deng, F., Huang, R., Petersen, F., Ibrahim, S., & Yu, X. mtDNA sequence, phylogeny and evolution of laboratory mice. *Mitochondrion* **17**, 126-31 (2014).
62. Baertsch, R., Diekhans, M., Kent, W. J., Haussler, D., & Brosius, J. Retrocopy contributions to the evolution of the human genome. *BMC Genomics* **9**, 466 (2008).
63. Hickey, G., Paten, B., Earl, D., Zerbino, D., & Haussler, D. HAL: a hierarchical format for storing and analyzing multiple genome alignments. *Bioinformatics* **29**, 1341-2 (2013).
64. Quinlan, A. R. BEDTools: The Swiss-Army Tool for Genome Feature Analysis. *Curr Protoc Bioinformatics* **47**, 11.12.1-34 (2014).
65. Edgar, R. C. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* **5**, 113 (2004).
66. http://www.geneious.com Last accessed on August, 21 2017.

67. Guo, Y., Mahony, S., & Gifford, D. K. High resolution genome wide binding event finding and motif discovery reveals transcription factor spatial binding constraints. *PLoS Comput Biol* **8**, e1002638 (2012).
68. Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., & Salzberg, S. L. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol* **14**, R36 (2013).
69. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., & 1000 Genome Project Data Processing Subgroup The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-9 (2009).
70. Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D. R., Pimentel, H., Salzberg, S. L., Rinn, J. L., & Pachter, L. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat Protoc* **7**, 562-78 (2012).
71. http://repeatmasker.org Last accessed on August, 21 2017.
72. Durinck, S., Moreau, Y., Kasprzyk, A., Davis, S., De Moor, B., Brazma, A., & Huber, W. BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. *Bioinformatics* **21**, 3439-40 (2005).
73. Durinck, S., Spellman, P. T., Birney, E., & Huber, W. Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nat Protoc* **4**, 1184-91 (2009).
74. Lex, A., Gehlenborg, N., Strobelt, H., Vuillemot, R., & Pfister, H. UpSet: Visualization of Intersecting Sets. *IEEE Trans Vis Comput Graph* **20**, 1983-92 (2014).
75. Bennett, B. D. & Bushel, P. R. goSTAG: gene ontology subtrees to tag and annotate genes within a set. *Source Code Biol Med* **12**, 6 (2017).
76. Greene, D., Richardson, S., & Turro, E. ontologyX: a suite of R packages for working with ontological data. *Bioinformatics* **33**, 1104-1106 (2017).
77. Lam, H. Y. K., Khurana, E., Fang, G., Cayting, P., Carriero, N., Cheung, K.-H., & Gerstein, M. B. Pseudofam: the pseudogene families database. *Nucleic Acids Res* **37**, D738-43 (2009).
78. Dickinson, M. E., Flenniken, A. M., Ji, X., Teboul, L., Wong, M. D., White, J. K., Meehan, T. F., Weninger, W. J., Westerberg, H., Adissu, H., Baker, C. N., Bower, L., Brown, J. M., Caddle, L. B., Chiani, F., Clary, D., Cleak, J., Daly, M. J., Denegre, J. M., Doe, B., Dolan, M. E., Edie, S. M., Fuchs, H., Gailus-Durner, V., Galli, A., Gambadoro, A., Gallegos, J., Guo, S., Horner, N. R., Hsu, C.-W., Johnson, S. J., Kalaga, S., Keith, L. C., Lanoue, L., Lawson, T. N., Lek, M., Mark, M., Marschall, S., Mason, J., McElwee, M. L., Newbigging, S., Nutter, L. M. J., Peterson, K. A., Ramirez-Solis, R., Rowland, D. J., Ryder, E., Samocha, K. E., Seavitt, J. R., Selloum, M., Szoke-Kovacs, Z., Tamura, M., Trainor, A. G., Tudose, I., Wakana, S., Warren, J., Wendling, O., West, D. B., Wong, L., Yoshiki, A., International Mouse Phenotyping Consortium, Jackson Laboratory, Infrastructure Nationale PHENOMIN, Institut Clinique de la Souris (ICS), Charles River Laboratories, MRC Harwell, Toronto Centre for Phenogenomics, Wellcome Trust Sanger Institute, RIKEN BioResource Center, MacArthur, D. G., Tocchini-Valentini, G. P., Gao, X., Flicek, P., Bradley, A., Skarnes, W. C., Justice, M. J., Parkinson, H. E., Moore, M., Wells, S., Braun, R. E., Svenson, K. L., de Angelis, M. H., Herault, Y., Mohun, T., Mallon, A.-M., Henkelman, R. M., Brown, S. D. M., Adams, D. J., Lloyd, K. C. K., McKerlie, C., Beaudet, A. L., Bućan, M., & Murray, S. A. High-throughput discovery of novel developmental phenotypes. *Nature* **537**, 508-514 (2016).