Comparative analysis of pseudogenes across three phyla

Cristina Sisu[a,b,1], Baikang Pei[a,1], Jing Leng[a,1], Adam Frankish[c,1], Yan Zhang[a,1], Suganthi Balasubramanian[b], Rachel Harte[d], Daifeng Wang[a], Michael Rutenberg-Schoenberg[a], Wyatt Clark[a], Mark Diekhans[d], Joel Rozowsky[b], Tim Hubbard[c], Jennifer Harrow[c], and Mark B. Gerstein[a,b,e,2]

[a]Program in Computational Biology and Bioinformatics and [b]Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, CT 06520; [c]Wellcome Trust Sanger Institute, Cambridge CB10 1SA, United Kingdom; [d]Center for Biomolecular Science and Engineering, University of California, Santa Cruz, CA 95064; and [e]Department of Computer Science, Yale University, New Haven, CT 06511

Edited* by Robert H. Waterston, University of Washington, Seattle, WA, and approved July 18, 2014 (received for review April 21, 2014)

Pseudogenes are degraded fossil copies of genes. Here, we report a comparison of pseudogenes spanning three phyla, leveraging the completed annotations of the human, worm, and fly genomes, which we make available as an online resource. We find that pseudogenes are lineage specific, much more so than protein-coding genes, reflecting the different remodeling processes marking each organism's genome evolution. The majority of human pseudogenes are processed, resulting from a retrotranspositional burst at the dawn of the primate lineage. This burst can be seen in the largely uniform distribution of pseudogenes across the genome, their preservation in areas with low recombination rates, and their preponderance in highly expressed gene families. In contrast, worm and fly pseudogenes tell a story of numerous duplication events. In worm, these duplications have been preserved through selective sweeps, so we see a large number of pseudogenes associated with highly duplicated families such as chemoreceptors. However, in fly, the large effective population size and high deletion rate resulted in a depletion of the pseudogene complement. Despite large variations between these species, we also find notable similarities. Overall, we identify a broad spectrum of biochemical activity for pseudogenes, with the majority in each organism exhibiting varying degrees of partial activity. In particular, we identify a consistent amount of transcription (~15%) across all species, suggesting a uniform degradation process. Also, we see a uniform decay of pseudogene promoter activity relative to their coding counterparts and identify a number of pseudogenes with conserved upstream sequences and activity, hinting at potential regulatory roles.

genome annotation | functional genomics | transcriptomics

Often referred to as "genomic fossils" (1–3), pseudogenes are defined as disabled copies of protein-coding genes. However, some have been found to be transcribed (4–7) and play important regulatory roles (8, 9). Presumed to evolve with little selective constraints (10), pseudogenes are of great value in estimating the rate of spontaneous mutation and hence provide insight into genome evolution (11, 12).

Previously, pseudogenes have been characterized within individual genomes (1, 4, 13–16). Pseudogene assignments are dependent on reliable and stable protein-coding annotations of their "parents" within the organism. Earlier nonstandardized annotations resulted in fluctuations of pseudogene assignments from one database release to another (*SI Appendix*, Fig. S1). As such, the absence of a comprehensive annotation and the potential of mis-mapping of functional genomics data had restricted former comparisons of the pseudogene complement in various organisms to specific families or classes of pseudogenes (17–20). The availability of complete genome annotations of human (*Homo sapiens*), worm (*Caenorhabidis elegans*), and fly (*Drosophila melanogaster*) on stable reference assemblies, allows us, for the first time to our knowledge, to embark on a uniform and comprehensive cross-species comparison. Moreover, we are able to elucidate functional aspects of pseudogenes leveraging the rich diversity of the functional genomics data from the Encyclopedia of DNA Elements (ENCODE) consortium.

Although they all share common regulatory and transcriptional principles (21, 22), the human, worm, and fly are members of different phyla. To complement our comparison of these distant organisms and provide an intraphylum context, we extend our analysis to include three select chordates. We study the zebrafish (*Danio rerio*), mouse (*Mus musculus*), and macaque (*Macaca mulata*) pseudogenes, taking advantage of the variety of functional genomics data available for mouse and the manual genomic annotation of zebrafish.

The prevalence of pseudogenes, as well as their high sequence similarity to coding genes, raises various issues in experiments designed to probe protein-coding regions (23, 24). The finished annotation highlighted in this study is useful for reducing false discoveries and mis-annotations. It also gives us the opportunity to correctly identify and analyze pseudogenes with potential biological activity.

Results

The Pseudogene Resource. In this study, we present completed pseudogene annotations in human, worm, and fly, as part of the ENCODE project. Pseudogene annotation is a difficult and

Significance

Pseudogenes have long been considered nonfunctional elements. However, recent studies have shown they can potentially regulate the expression of protein-coding genes. Capitalizing on available functional-genomics data and the finished annotation of human, worm, and fly, we compared the pseudogene complements across the three phyla. We found that in contrast to protein-coding genes, pseudogenes are highly lineage specific, reflecting genome history more so than the conservation of essential biological functions. Specifically, the human pseudogene complement reflects a massive burst of retrotranspositional activity at the dawn of the primates, whereas the worm's and fly's repertoire reflects a history of deactivated duplications. However, we also observe that pseudogenes across the three phyla have a consistent level of partial activity, with ~15% being transcribed.

Author contributions: C.S., B.P., J.H., and M.B.G. designed research; C.S., B.P., and J.L. performed research; C.S., B.P., J.L., A.F., Y.Z., S.B., R.H., D.W., M.R.-S., W.C., M.D., J.R., T.H., and J.H. analyzed data; and C.S., B.P., J.L., A.F., and M.B.G. wrote the paper.

The authors declare no conflict of interest.

*This Direct Submission article had a prearranged editor.

Freely available online through the PNAS open access option.

Data deposition: All data associated with this paper has been deposited in a publicly accessible database at http://psicube.pseudogene.org.

[1]C.S., B.P., J.L., A.F., and Y.Z. contributed equally to this work.

[2]To whom correspondence should be addressed. Email: mark@gersteinlab.org.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1407293111/-/DCSupplemental.

complex process. Sequence decay at pseudogene loci makes it challenging to identify authentic pseudogenes and accurately define their boundaries (4). Therefore, we use a hybrid approach, combining manual annotation with computational pipelines to identify pseudogenes. Although providing high accuracy, the manual process is slow and may overlook highly mutated or truncated pseudogenes with weak homology to their parents. Conversely, computational pipelines are fast and provide an unbiased annotation of pseudogenes but are also prone to errors due to mis-annotation of parent gene loci. Thus, using a uniform annotation procedure, we curate a highly accurate and exhaustive pseudogene set for each organism.

Comparing the different organisms, the pseudogene distribution does not follow relative genome size or gene counts. For example, the human genome has about 50-fold more pseudogenes than zebrafish, 100-fold more than fly, but only 15-fold more than worm (Fig. 1*A*).

Given the large evolutionary distance between the model organisms and human, we use the macaque and mouse as a mammalian pseudogene baseline. We estimate the pseudogene content in the two organisms using an in-house computational annotation pipeline [PseudoPipe (2)]. As expected, the two mammals show similar pseudogene content to human (Fig. 1*A*).

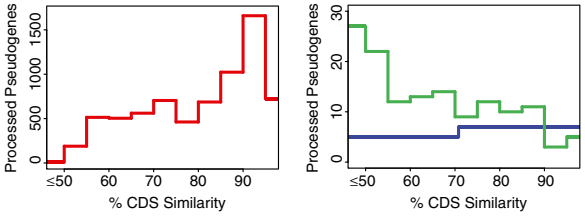
All of the data resulting from the annotation and comparative analysis are collected into a comprehensive online pseudogene resource: psicube.pseudogene.org.

Classification and Evolution. ***Classification.*** Based on their mechanism of formation (18), pseudogenes can be classified into several categories: duplicated (unprocessed), processed (resulting from retrotransposition), and unitary (unprocessed pseudogenes with an active ortholog in another species). We find that processed pseudogenes are the dominant biotype in mammals, whereas worm, fly, and zebrafish genomes are enriched for duplicated pseudogenes (Fig. 1*A*).

A

Organism	Total Pseudogenes	Biotype Distrbution (Processed / Duplicated)	ENCODE Functional Genomics Data	Completed Manual Annotation
Human	12,358	8908 / 2266	✔	✔
Worm	911	159 / 566	✔	✔
Fly	145	16 / 109	✔	✔
Zebrafish	229	41 / 177	✔	✔
Macaque	11,136	8570 / 1725	✗	✗
Mouse	13,169	7811 / 1827	✔	✗

B



C

Organism	Defect / Pseudogene x MB		
	Insertion	Deletion	Stop
Human	4.4	4.9	2.4
Worm	**25.8**	7.45	2.5
Fly	7.9	**12.7**	1.1

Fig. 1. Annotation, classification, and evolution. (*A*) Pseudogene annotation and ENCODE functional data availability. (*B*) Distribution of processed pseudogenes as a function of pseudogene age (sequence similarity to parent genes) for human (*Left*) and worm and fly (*Right*). (*C*) Pseudogene disablement variation and density.

Timeline. Next, we study pseudogene evolution. We infer pseudogene age using sequence similarity to the parent gene and assess the abundance of pseudogenes of different ages. We observe that the distribution of duplicated pseudogenes shows little variation with age (*SI Appendix*, Fig. S2). However, the creation of processed pseudogenes varies very much over time (Fig. 1*B*). In human, the peak of processed pseudogenes (at high sequence similarity) corresponds to the burst of retrotransposition events (20, 25, 26). Likewise, macaque and mouse show a stepwise increase in the number of processed pseudogenes at similar time points (*SI Appendix*, Fig. S2). By contrast, in worm, we see a higher proportion of older processed pseudogenes compared with younger ones. In fly and zebrafish, we find a small constant number of processed pseudogenes across all age groups.

Repeats. Repeat elements play an important role in transposition events and thus in the creation of pseudogenes (27, 28). To this end, we examine the transposable element content of various annotated features in the genome, namely coding sequences (CDSs), UTRs, long noncoding RNAs (lncRNAs), and pseudogenes (*SI Appendix*, Fig. S3). In general, pseudogenes show a lower transposable element content than UTRs and lncRNAs and even the genomic average. In the case of processed pseudogenes, this is consistent with the fact that, although repeats are required for their genesis, they are not reinserted at the pseudogene loci themselves. Similarly, the transposable element content in the CDS is low, indicating a strong purifying selection pressure in these regions. By contrast, the lncRNAs and UTRs show a high transposable element content and low conservation in all three species.

Disablements and selection. Pseudogenes are believed to evolve neutrally; hence, they accumulate mutations and indels. We analyze the variety and kinds of disablements as markers of pseudogene evolution. Based on their origins, we distinguish three types of disablements: insertions, deletions, and stop codons (Fig. 1*C* and *SI Appendix*, Fig. S2). We observe a lower disablement density in human pseudogene sequences compared with the worm and fly (*SI Appendix*, Fig. S4). The average number of indels is constant in human and is twice the number of stop codons. However, the fly and worm genomes show a preference for deletions and insertions, respectively.

Further, we study the selection in human pseudogenes by analyzing the frequency of rare SNPs. At population level, we do not find any statistically significant enrichment in pseudogenes for these SNPs over the genomic average (*SI Appendix*, Fig. S5).

Localization and Mobility. Given the fact that the majority of pseudogenes are not under strong selective pressure, we expect to find them in regions of low recombination rates. To this end, we analyze the recombination rate at pseudogene loci for each species (Fig. 2*A*). We find that the human and fly pseudogenes are enriched in regions of low recombination and thus are preferentially located near the centromere and on the sex chromosomes. However, for worm pseudogenes, we observe a somewhat similar recombination rate to that of genes, a possible consequence of recent selective sweeps (29). As such, the pseudogenes are relatively enriched near the telomeres, regions usually characterized by high recombination rates and rapid gene evolution (30).

Looking at the distribution of pseudogenes, we find, as expected, a strong correspondence between the number of duplicated pseudogenes and protein-coding gene density in worm and fly (Fig. 2*B*). By contrast, in human, the number of processed pseudogenes is proportional to the chromosome length but is less correlated to the number of protein-coding genes, suggesting the existence of interchromosomal transfers (Fig. 2*B* and *SI Appendix*, Fig. S6). However, duplicated pseudogenes are commonly found on the same chromosome as their parent genes. This coresidence is notable for human chromosomes 7 and 11, due to their enrichment in genome duplication events (31) and duplicated olfactory receptors, respectively (32). The colocalization is also significant for sex chromosomes (human Y, fly X), where, as a consequence of low recombination rates the pseudogenes cannot

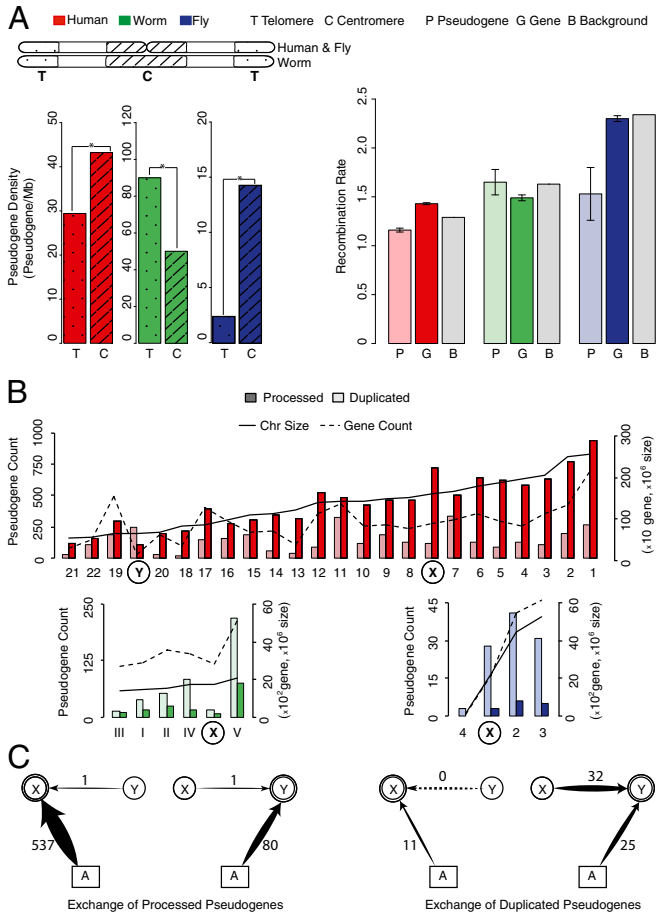


Fig. 2. Localization and mobility. (*A*, *Left*) The relative chromosomal localization preference for pseudogenes in human, worm, and fly. (*Right*) Average recombination rates for pseudogenes, protein-coding genes, and genomic background. (*B*) Distributions of processed and duplicated pseudogenes across chromosomes, sorted by length. (*C*) Pseudogene exchange between sex chromosomes and autosomes in humans.

be "crossed out" (33, 34). Further, in human, we observe a large accumulation of imported processed pseudogenes on X (35) (pseudogenes on X with parents on other chromosomes) and an enrichment of duplicated pseudogenes on Y with apparent parent genes on the X chromosome (Fig. 2*C*).

Orthologs, Paralogs, and families. We compare the lineage specificity of pseudogenes by analyzing their families and orthologs.

Orthologs. Numerous protein-coding genes have preserved orthologs even for such distant organisms as the human, worm, and fly; in particular, there are ∼2,000 1-1-1 human-worm-fly ortholog triplets (*Materials and Methods*). However, there are no pseudogene orthologs preserved across all three species (Fig. 3*A* and *SI Appendix*, Table S2). In contrast, we are able to identify orthologous pairs for closer relatives such as human and mouse. We find that only 129 (∼1%) of the human pseudogenes have mouse orthologs. The majority of these (127) are processed and have high sequence similarity to their parents. Also ∼20% of the orthologous pseudogenes are transcribed in both organisms (*SI Appendix*, Figs. S7 and S8).

Next, analyzing ∼2,000 1-1-1 human-worm-fly orthologs, we find that not one of the triplets have associated pseudogenes in all three organisms (l). Also the number of pseudogenes associated with 1-1-1 protein-coding orthologs differs greatly across species. As an example (Fig. 3*B*), ribosomal protein S6 has 25 (mostly processed) pseudogenes spread randomly across the human genome, three duplicated pseudogenes clustered near the parent gene in fly, and no corresponding pseudogenes in worm.

Paralogs and families. We compare the distribution pattern of pseudogenes per parent gene (Fig. 3*C*). In human, despite the fact that pseudogenes are almost as numerous as protein-coding genes (4), only 25% of genes have a pseudogene counterpart. Consequently, the distribution of pseudogenes per gene is highly uneven. As a control, we looked at the distribution of paralogs per parent gene. Across all species, there is little overlap between genes with a large number of paralogs and those with a large pseudogene complement. At the extreme, we find a number of genes that are enriched in pseudogenes and depleted in paralogs and vice versa, a trend common across all organisms.

Family analysis allows for a larger pattern to emerge (Fig. 3*D*). The relative ranks of the gene families with the most pseudogenes are organism specific. In fly, amyloid P component serum (SAP) and kinesin motor domain protein families are dominant. The top pseudogene families in worm are the seven-transmembrane domain receptor (7TM) proteins, perhaps reflecting the family's rapid evolution (36) and the large number of duplication events in nematode genome history (37). Interestingly, even though processed pseudogenes are dominant in human, the human genome shares 7TM as its top family, an indication of the duplication and divergence of the olfactory receptors.

Collectively, as expected, the ribosomal proteins are the dominant families in human, comprising almost 20% of the total pseudogenes. These abundantly expressed genes are indicative of the general burst of retrotransposition events (38–40). Analysis of top mouse and macaque families shows that this pattern is common across mammalian genomes.

Finally, despite the lineage specificity of the top pseudogene families, we find a number of highly duplicated families common to all organisms: kinases, histones, and P-loop NTPases, reflecting perhaps the essential role that these genes play in the species evolution.

Activity. Next we directed our investigation toward identifying potentially active pseudogenes by looking for signs of biochemical activity.

Transcription. Analyzing RNA-Seq data, we find 1,441, 143, and 23 potentially transcribed pseudogenes in human, worm, and fly, respectively. We also identify 31 transcribed pseudogenes in zebrafish and 878 in mouse. These numbers represent a fairly uniform fraction (∼15%) of the total pseudogene complement in each organism. Among transcribed pseudogenes, ∼13% in human and ∼30% in worm and fly have a discordant transcription pattern with their parent genes over multiple samples. Also, a large fraction of pseudogenes are associated with a few highly expressed gene families, e.g., the ribosomal proteins in human.

The parent genes of broadly expressed pseudogenes tend to be broadly expressed as well (*SI Appendix*, Fig. S9), but the reciprocal statement is not valid. Specifically, only 5.1%, 0.69%, and 4.6% of the total number of pseudogenes are broadly expressed in human, worm, and fly, respectively. However, in general, transcribed pseudogenes show higher tissue specificity than protein-coding genes (*SI Appendix*, Fig. S10).

Activity features. Next we examine a number of additional markers of biochemical activity, including the presence of active transcription factors (TFs) and RNA polymerase II (Pol II) binding sites in the upstream sequence and proximal regions of "active chromatin" for each pseudogene. We integrated the transcriptional information with additional functional data to create a comprehensive map of pseudogene activity (Fig. 4*A*), grouping them into different categories. At one extreme, we find a group of dead pseudogenes, with no indicators of activity. Contrary to the actual definition of pseudogenes ("dead genomic elements"), this group comprises only ∼20% of the total pseudogenes. On the other extreme, some, albeit very few, pseudogenes (<5%) are transcribed and simultaneously exhibit all other activity features, despite the presence of disruptive mutations. We label these pseudogenes as highly active. Also, in human, we find that the transcribed

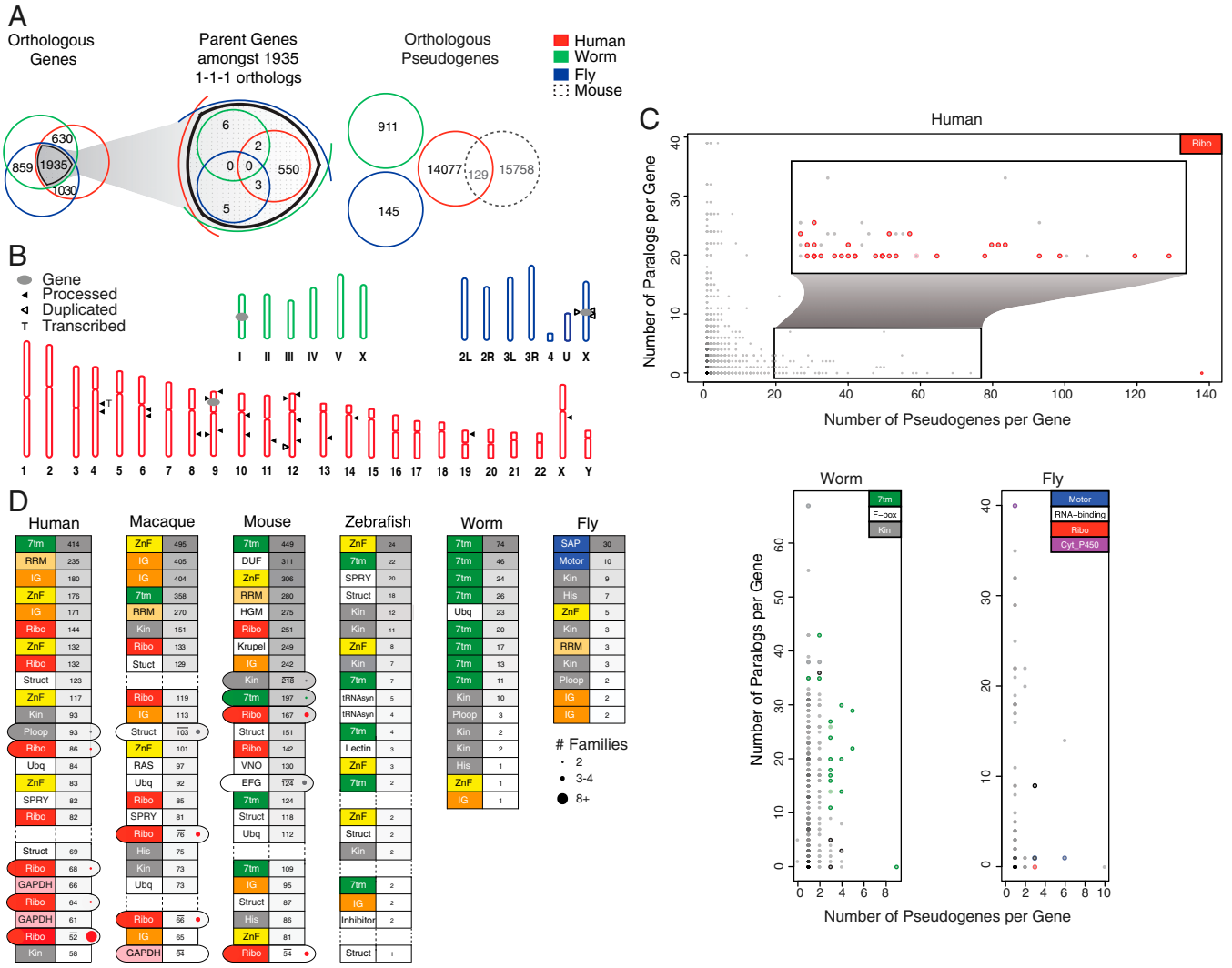


Fig. 3. Orthologs, paralogs, and families. (*A*) Venn diagrams showing the total number of orthologous genes and pseudogenes, in human, worm, and fly. (*Right*) Pseudogene orthologs between human and mouse. (*B*) Per chromosome distribution of RpS6 pseudogenes in human, worm, and fly. (*C*) Comparative distribution of pseudogene and paralogs per gene. (*D*) Top pseudogene families that give rise to 25% of the total number of pseudogenes in each organism (*Left*, family type; *Right*, number of pseudogenes). Oval rows indicate the collapse of two or more consecutive families of the same type. 7tm, G protein-coupled receptors; His, histone; IG, Ig; Kin, kinase; Ploop, P-loop NTPase proteins; Ribo, ribosomal proteins; RRM, RNA recognition motifs; Struct, structural protein; ZnF, Zinc finger proteins (TF); Ubq, ubiquitination proteins; Motor, kinesin motor domain proteins; SAP, SAP domain proteins.

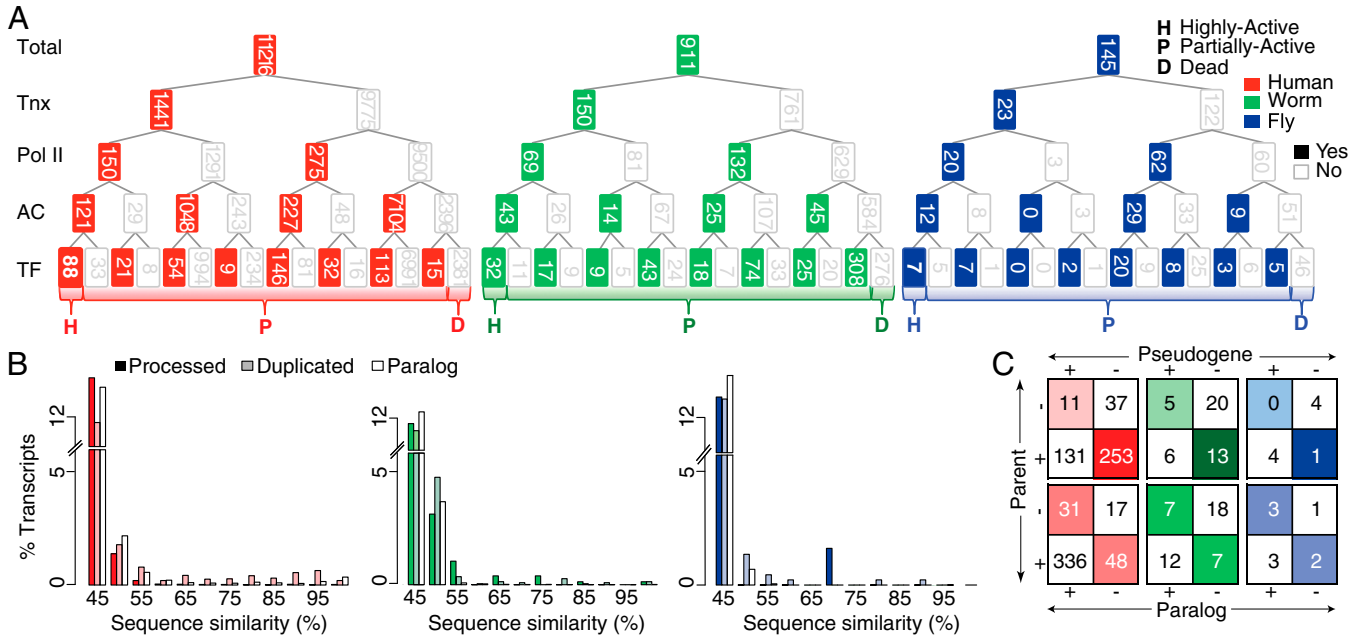
pseudogenes in general, and the highly active pseudogenes in particular, are enriched in rare alleles, indicating that they are under stronger negative selection than the other, less active pseudogenes (*SI Appendix*, Fig. S11). However, the majority of pseudogenes (~75%) are intermediate between these two, having only a few of the classic indicators of activity. We label these as partially active. The distribution of pseudogenes for the three activity levels is consistent across all studied species.

Upstream sequence similarity and promoter activity. Pseudogene activity is connected to the upstream regulatory region. We examine the sequence divergence in the proximal (within 2 kb of the 5′ end) upstream region of pseudogenes (i.e., their promoters) using the promoter regions of parent–gene paralogs as a control.

Contrary to expectations, a small fraction of duplicated pseudogenes exhibits highly conserved upstream regions, even more so than paralogs, compared with the parent genes (Fig. 4*B*). These pseudogenes may be recent duplicated loci that have diverged little from their parents. Interestingly, we find a number of duplicated pseudogene–parent pairs with high upstream similarity despite low coding sequence identity, suggesting that the upstream regions may have been especially conserved via purifying selection. These scenarios could lead to a coordinated expression pattern between the transcriptional products regulated by these promoter regions. To this end, we analyze the ChIP-seq data of H3K27Ac, an important marker in defining active promoters and enhancers. The comparison is focused on protein-coding genes with only one pseudogene but no paralogs, and those with one pseudogene and one paralog. We note that, in general, although the pseudogenes have highly conserved promoter regions, the activity is less preserved compared with their protein-coding gene counterparts (Fig. 4*C*).

Functional Pseudogene Candidates. Finally, combining the annotation, functional genomics, and evolutionary data, we refine the active pseudogene group to a set of functional candidates. This term refers to a pseudogene that possesses numerous signs of activity, commonly attributed to canonical coding genes (e.g., transcription, translation, and active chromatin). This list focuses on the regulatory potential of pseudogenes and includes the known regulatory cancer pseudogene *PTEN-P1* (8).

For this set, using MS data, we study the translation potential of transcribed human pseudogenes in four ENCODE cell lines. We



D

Translation Candidates	Biotype	Coexp Coef	pVal	%Similarity CDS/UTR	Defect	Tnx	Pol II	AC	TF
SLIT-ROBO Rho GTPase activating protein 2B pseudogene	Duplicated	0.80	5.9E-7	58 / 50	ins	✔	-	✔	-
PRKY-004, Y-linked protein kinase pseudogene	Duplicated	-0.14	0.42	96 / 51	ins / del	✔	✔	-	✔
Fer-1-like 4 (C. elegans), pseudogene	Unitary	-0.38	0.03	62 / 32	ins / del	✔	-	✔	-

Fig. 4. Pseudogene activity. (*A*) Distribution of pseudogenes as a function of various activity features: transcription (Tnx), active chromatin (AC), and presence of active Pol II and TF binding sites in the upstream region. (*B*) Conservation of the upstream sequences in processed and duplicated pseudogenes compared with paralogs. (*C*) Conservation of an upstream sequence activity mark (H3K27Ac) in pseudogene-parent pairs vs. parent-paralogs. +, active H3K27Ac; –, inactivity. We find that the majority of parent–paralog pairs have coordinated H3K27Ac activity (larger diagonal values) as opposed to parent–pseudogene pairs (larger off-diagonal values). (*D*) Functional pseudogene candidates with translation evidence.

find three pseudogenes with high translation evidence (Fig. 4*D* and *SI Appendix*, Table S3). The low number of candidate translated pseudogenes is indicative of the high quality of our annotation. Interestingly, one of the candidates (*chromosome Y-linked protein kinases pseudogene*) shows numerous activity features and a low coexpression correlation to its parent, suggesting that it is under a different regulatory pattern than its parent gene.

Discussion

We report a multiorganism comparison of pseudogenes leveraging the finished annotations of the genomes of human, worm, and fly. Given that these are high-quality annotations, we do not expect to see any significant changes in the total number of pseudogenes in the future. (For a detailed discussion of the variance in gene and pseudogene counts over draft annotation releases, see *SI Appendix*, Fig. S1 and the supplementary information in refs. 4 and 21.) Unlike protein-coding genes, which are essential to the correct development and function of the organism and thus are under strong selective pressure, the majority of pseudogenes evolve neutrally, making them an ideal proxy for the study of genome evolution.

Overall, our results show that the pseudogene complement is lineage specific, reflecting the different genome remodeling processes characterizing each organism's evolution. There are essentially no orthologous pseudogenes between these distant organisms, and we only see an overlap at the protein family level, where a few large, highly duplicated families (e.g., kinases) give rise to a large number of pseudogenes in all of the studied species.

We find that the mammalian pseudogene complement is marked by a large event, a retrotranspositional burst that occurred ~40 Mya, at the dawn of the primate lineage (25, 39, 40). This burst can be clearly seen in the largely uniform distribution of pseudogenes across the chromosomes and their slight accumulation increase in areas with low recombination rates, e.g., the sex chromosomes and the centromere regions. It also resulted in a preponderance of pseudogenes associated with highly transcribed genes such as those in pathways of central metabolism and the ribosomal proteins. Although the burst of retrotransposition events happened after the human/mouse speciation (~75 Mya) (41, 42), the high occurrence of processed pseudogenes in the mouse genome suggests that this event occurred on a much larger scale, and it may be a more general mammalian characteristic. In contrast, the worm and fly pseudogene complements tell a story of numerous duplication events. This scenario is apparent in the worm genome due to the fact that a large number of pseudogenes are associated with highly duplicated gene families such as the chemoreceptors. Moreover, due to recent selective sweeps, many of these pseudogenes, which otherwise would have been purged by recombination, have been preserved on the chromosome arms. In the fly genome, a large population size (43, 44) combined with a strong selection in the intergenic sequence (43, 45) and a high deletion rate have resulted in a depletion of the pseudogene complement. Consequently, we see segregation of the remaining pseudogenes to areas of low recombination.

The apparent duplicated pseudogene exchange between the X and Y chromosomes in human is a consequence of the numerous gene loss events in Y's evolutionary history (46). As such, the majority of "X-exported" duplicated pseudogenes on Y are likely degenerated copies that subsequently accumulated deleterious mutations (47).

Finally, we identify a large spectrum of biochemical activity (as defined by transcription, active chromatin, and Pol II and TF binding) for pseudogenes ranging from highly active to dead. The

majority of pseudogenes (~75%) are found between these two extremes, exhibiting various proportions of residual activity. In particular, we identify a consistent amount of transcription (~15%) in each organism. The distribution of these activity levels is consistent across all species implying a uniform rate of degradation.

We relate the activity of pseudogenes to the conservation of their upstream regions. Comparing pseudogenes and functional paralogs, we find that many pseudogenes have more conserved upstream sequences than is typical for paralogs. Further, we identify a number of pseudogenes with highly conserved upstream regions relative to their parent genes. However, this conservation is not always preserved in terms of upstream activity (as defined by histone marks). In this case, pseudogenes are less active than their coding counterparts, reflecting the functional degradation of these regions. The small subset of pseudogenes with conserved promoters both in sequence and activity hints at potential regulatory roles.

We complete our analysis by ranking pseudogenes based on their activity features and by pinpointing potentially functional candidates. The regulatory roles of several pseudogenes through their RNA products have been previously demonstrated (8, 9, 48–50). Hence, we suggest that some pseudogenes may play active roles in genome biology and warrant further experimental investigation. We realize the notion of functional pseudogene is, in a sense, an oxymoron. However, here we focus only on tabulating and enumerating these potential functional candidates. In light of recent advances in functional genomics and genome biology, it may be useful to revisit the definition of gene and pseudogene to better and more accurately describe these entities (6, 51, 52).

Materials and Methods

We present the annotation and analysis of the pseudogene complement in human, worm, and fly, leveraging functional genomics data available from the ENCODE and modENCODE consortia. The human pseudogene annotation is based on the GENCODE 10 release. For worm and fly, we curated pseudogene annotation sets extending beyond WormBase WS220 and FlyBase 5.45. A detailed description of the materials and methods is available in the *SI Appendix*.

1. Zheng D, et al. (2007) Pseudogenes in the ENCODE regions: Consensus annotation, analysis of transcription, and evolution. *Genome Res* 17(6):839–851.
2. Zhang Z, et al. (2006) PseudoPipe: An automated pseudogene identification pipeline. *Bioinformatics* 22(12):1437–1439.
3. Harrison PM, et al. (2002) Molecular fossils in the human genome: Identification and analysis of the pseudogenes in chromosomes 21 and 22. *Genome Res* 12(2):272–280.
4. Pei B, et al. (2012) The GENCODE pseudogene resource. *Genome Biol* 13(9):R51.
5. Harrison PM, Zheng D, Zhang Z, Carriero N, Gerstein M (2005) Transcribed processed pseudogenes in the human genome: An intermediate form of expressed retrosequence lacking protein-coding ability. *Nucleic Acids Res* 33(8):2374–2383.
6. Zheng D, Gerstein MB (2007) The ambiguous boundary between genes and pseudogenes: The dead rise up, or do they? *Trends Genet* 23(5):219–224.
7. Iskow RC, et al. (2012) Regulatory element copy number differences shape primate expression profiles. *Proc Natl Acad Sci USA* 109(31):12656–12661.
8. Poliseno L, et al. (2010) A coding-independent function of gene and pseudogene mRNAs regulates tumour biology. *Nature* 465(7301):1033–1038.
9. Muro EM, Mah N, Andrade-Navarro MA (2011) Functional evidence of post-transcriptional regulation by pseudogenes. *Biochimie* 93(11):1916–1921.
10. Petrov DA, Hartl DL (2000) Pseudogene evolution and natural selection for a compact genome. *J Hered* 91(3):221–227.
11. Ophir R, Graur D (1997) Patterns and rates of indel evolution in processed pseudogenes from humans and murids. *Gene* 205(1-2):191–202.
12. Balasubramanian S, et al. (2002) SNPs on human chromosomes 21 and 22 — analysis in terms of protein features and pseudogenes. *Pharmacogenomics* 3(3):393–402.
13. Karro JE, et al. (2007) Pseudogene.org: A comprehensive database and comparison platform for pseudogene annotation. *Nucleic Acids Res* 35(Database issue):D55–D60.
14. Harrison PM, Echols N, Gerstein MB (2001) Digging for dead genes: An analysis of the characteristics of the pseudogene population in the Caenorhabditis elegans genome. *Nucleic Acids Res* 29(3):818–830.
15. Harrison PM, Milburn D, Zhang Z, Bertone P, Gerstein M (2003) Identification of pseudogenes in the Drosophila melanogaster genome. *Nucleic Acids Res* 31(3): 1033–1037.
16. Howe K, et al. (2013) The zebrafish reference genome sequence and its relationship to the human genome. *Nature* 496(7446):498–503.
17. Fairbanks DJ, Maughan PJ (2006) Evolution of the NANOG pseudogene family in the human and chimpanzee genomes. *BMC Evol Biol* 6:12.
18. Echols N, et al. (2002) Comprehensive analysis of amino acid and nucleotide composition in eukaryotic genomes, comparing genes and pseudogenes. *Nucleic Acids Res* 30(11):2515–2523.
19. Harrison PM, Gerstein M (2002) Studying genomes through the aeons: Protein families, pseudogenes and proteome evolution. *J Mol Biol* 318(5):1155–1174.
20. Balasubramanian S, et al. (2009) Comparative analysis of processed ribosomal protein pseudogenes in four mammalian genomes. *Genome Biol* 10(1):R2.
21. Gerstein MB, et al. (2014) Comparative analysis of the transcriptome across distant species. *Nature*, 10.1038/nature13424.
22. Boyle AP, et al. (2014) Comparative analysis of regulatory information and circuits across distant species. *Nature*, 10.1038/nature13668.
23. Mutimer H, Deacon N, Crowe S, Sonza S (1998) Pitfalls of processed pseudogenes in RT-PCR. *Biotechniques* 24(4):585–588.
24. Garbay B, Boue-Grabot E, Garret M (1996) Processed pseudogenes interfere with reverse transcriptase-polymerase chain reaction controls. *Anal Biochem* 237(1): 157–159.
25. Torrents D, Suyama M, Zdobnov E, Bork P (2003) A genome-wide survey of human pseudogenes. *Genome Res* 13(12):2559–2567.
26. Zhang ZD, Cayting P, Weinstock G, Gerstein M (2008) Analysis of nuclear receptor pseudogenes in vertebrates: How the silent tell their stories. *Mol Biol Evol* 25(1): 131–143.
27. Ding W, Lin L, Chen B, Dai J (2006) L1 elements, processed pseudogenes and retrogenes in mammalian genomes. *IUBMB Life* 58(12):677–685.
28. Yang H-P, Barbash DA (2008) Abundant and species-specific DINE-1 transposable elements in 12 Drosophila genomes. *Genome Biol* 9(2):R39.
29. Andersen EC, et al. (2012) Chromosome-scale selective sweeps shape Caenorhabditis elegans genomic diversity. *Nat Genet* 44(3):285–290.
30. Barnes TM, Kohara Y, Coulson A, Hekimi S (1995) Meiotic recombination, noncoding DNA and genomic organization in Caenorhabditis elegans. *Genetics* 141(1):159–179.
31. Hillier LW, et al. (2003) The DNA sequence of human chromosome 7. *Nature* 424(6945):157–164.
32. Glusman G, Yanai I, Rubin I, Lancet D (2001) The complete human olfactory subgenome. *Genome Res* 11(5):685–702.
33. Wilson ACC, Sunnucks P, Bedo DG, Barker JSF (2006) Microsatellites reveal male recombination and neo-sex chromosome formation in Scaptodrosophila hibisci (Drosophilidae). *Genet Res* 87(1):33–43.
34. Jensen-Seaman MI, et al. (2004) Comparative recombination rates in the rat, mouse, and human genomes. *Genome Res* 14(4):528–538.
35. Emerson JJ, Kaessmann H, Betrán E, Long M (2004) Extensive gene traffic on the mammalian X chromosome. *Science* 303(5657):537–540.
36. Castillo-Davis CI, Hartl DL (2002) Genome evolution and developmental constraint in Caenorhabditis elegans. *Mol Biol Evol* 19(5):728–735.
37. Thomas JH, Robertson HM (2008) The Caenorhabditis chemoreceptor gene families. *BMC Biol* 6:42.
38. Ishii K, et al. (2006) Characteristics and clustering of human ribosomal protein genes. *BMC Genomics* 7:37.
39. Pan D, Zhang L (2009) Burst of young retrogenes and independent retrogene formation in mammals. *PLoS ONE* 4(3):e5040.
40. Marques AC, Dupanloup I, Vinckenbosch N, Reymond A, Kaessmann H (2005) Emergence of young human genes after a burst of retroposition in primates. *PLoS Biol* 3(11):e357.
41. Zhao S, et al. (2004) Human, mouse, and rat genome large-scale rearrangements: Stability versus speciation. *Genome Res* 14(10A):1851–1860.
42. Waterston RH, et al.; Mouse Genome Sequencing Consortium (2002) Initial sequencing and comparative analysis of the mouse genome. *Nature* 420(6915):520–562.
43. Petrov DA, Chao YC, Stephenson EC, Hartl DL (1998) Pseudogene evolution in Drosophila suggests a high rate of DNA loss. *Mol Biol Evol* 15(11):1562–1567.
44. Lynch M, Conery JS (2003) The origins of genome complexity. *Science* 302(5649): 1401–1404.
45. Luque T, Marfany G, Gonzàlez-Duarte R (1997) Characterization and molecular analysis of Adh retrosequences in species of the Drosophila obscura group. *Mol Biol Evol* 14(12):1316–1325.
46. Heard E, Disteche CM (2006) Dosage compensation in mammals: Fine-tuning the expression of the X chromosome. *Genes Dev* 20(14):1848–1867.
47. Wong A, et al. (2004) Diverse fates of paralogs following segmental duplication of telomeric genes. *Genomics* 84(2):239–247.
48. Piehler AP, et al. (2008) The human ABC transporter pseudogene family: Evidence for transcription and gene-pseudogene interference. *BMC Genomics* 9:165.
49. Tam OH, et al. (2008) Pseudogene-derived small interfering RNAs regulate gene expression in mouse oocytes. *Nature* 453(7194):534–538.
50. Rapicavoli NA, et al. (2013) A mammalian pseudogene lncRNA at the interface of inflammation and anti-inflammatory therapeutics. *eLife* 2:e00762.
51. Snyder M, Gerstein M (2003) Genomics. Defining genes in the genomics era. *Science* 300(5617):258–260.
52. Sasidharan R, Gerstein M (2008) Genomics: Protein fossils live on as RNA. *Nature* 453(7196):729–731.

Materials and Methods

Data Sets

In this paper, we present the annotation and analysis of the pseudogene complement in human, worm, and fly, leveraging functional genomics data available from the ENCODE and modENCODE consortia. The human pseudogene annotation is based on the GENCODE 10 release. For worm and fly, we curated pseudogene annotation sets extending beyond WormBase WS220 and FlyBase 5.45. For mouse and macaque we used PseudoPipe automated pseudogene assignments based on the Ensembl 72 genome annotations. For zebrafish, we used pseudogene assignments from the Vega 53/Ensembl 73 manual annotation.

Pseudogene Annotation

Pseudogene annotation was conducted using a combination of manual annotation and in silico pipelines. The annotation files are available online at psicube.pseudogene.org for GRCh37 and GRCh38 (upon release).

(a) Manual Annotation

We manually annotated human pseudogenes on the basis of their homology to protein data from the UniProt database. The protein data were aligned to the individual bacterial artificial chromosome (BAC) clones that make up the reference genome sequence using BLAST [1]. We created gene models based on these alignments using the ZMAP annotation interface and the Otterlace annotation system [2]. Alignments were navigated using the Blixem alignment viewer [3]. We used visual inspection of the dot-plot output from the Dotter tool [3] to resolve any alignment with the genomic sequence that was unclear in, or absent from, Blixem. We defined a model as *pseudogene* if it possessed one or more of the following characteristics, unless there was evidence (transcriptional, functional, publication) showing that the locus represented a protein-coding gene with structural/functional divergence from its parent (paralog): (i) a premature stop codon relative to parent CDS, which could be introduced by nonsense or frame-shift mutation; (ii) a frame-shift in a functional domain - even where the length of the resulting CDS was similar to that of the parent CDS; (iii) a truncation of the 5' or 3' end of the CDS relative to the parent CDS; (iv) a deletion of an internal portion of the CDS relative to the parent CDS. Pseudogene loci lacking disabling mutations were annotated as "*ambiguous pseudogene"* when they lacked locus-specific transcriptional evidence. We note that the manual annotation pipeline checks the possibility than any putative pseudogene might instead be a protein-coding gene. If any putative pseudogene locus has transcriptional, functional or publication evidence to support coding potential, including selenocysteine incorporation, stop-codon read-through and programmed frameshift events, it is excluded from the set of pseudogene transcripts.

Fly pseudogenes were annotated in a similar way to human with two notable differences. First, while UniProt proteins were used to support the pseudogene annotation, we also used the CDS sequences of the parent gene loci predicted by PseudoPipe and/or FlyBase to build pseudogenes. Where the parent CDS was not clear, homologs of the pseudogene sequence were identified using BLAST. Secondly, where a parent CDS was used to investigate a

pseudogene it was aligned to the genome using Exonerate [4] before being assessed using Blixem and Dotter.

Worm pseudogenes were annotated in a similar fashion using a combination of automated (PseudoPipe) and manual annotation (WormBase [5]). The PseudoPipe pseudogene set was intersected with the manually annotated one. All pseudogenes passing a threshold of 80% sequence overlap between the two data sets were selected as part of the high confidence data set. Further, we manually validated biotype annotations.

(b) Automatic Annotation

PseudoPipe is an automatic pseudogene annotation tool that uses protein homology data to identify pseudogenes. PseudoPipe uses six-frame translational BLAST to search all known protein sequences from Ensembl. Pseudogene disablements were determined through sequence alignments to functional genes. The pseudogene parents (functional gene paralogs) were identified on the basis of sequence similarity.

Classification & Evolution

(a) Classification

Pseudogenes were classified as “processed” if they have lost their parental gene structures. Conversely, we classified pseudogenes as “unprocessed”/ “duplicated” if they retained the same exon-intron structure as their parent loci. In ambiguous cases we used other features to resolve the provenance of the pseudogene. Where the pseudogene represented a fragment of the parent, and the homology ended precisely at a splice junction the pseudogene was called “unprocessed” (“duplicated”). Conversely, where the fragment contained the fusion of two or more exons the pseudogene was called “processed”. If the parent had a single exon CDS, the presence of parent gene structure in the 5' UTR region (identified by alignment of mRNA and EST evidence) allowed the pseudogene to be called “unprocessed”/“duplicated”. Meanwhile, the presence of a pseudopoly(A) signal (the position of the parent poly(A) signal at the pseudogene locus) followed by a tract of A-rich sequence in the genome (indicating the insertion site of the polyadenylated parental mRNA) indicated a “processed” pseudogene. If there was no other evidence available to resolve the route by which the pseudogene was created, we used the position of the pseudogene relative to its parent. As such “processed” pseudogenes are reinserted into the genome with an approximately random distribution while “unprocessed”/“duplicated” pseudogenes tend to be more closely associated with the parent locus. Parsimony therefore suggests that pseudogenes that lie near to the parent locus are more likely to have arisen via a gene-duplication event than retrotransposition, and this was used as a tie-breaker in defining the pseudogene biotype.

(b) Timeline

Differences in the dynamics of genome evolution make it difficult to directly estimate pseudogene ages. We used sequence similarity to parent genes as an indicator of pseudogene age. Thus, young pseudogenes were defined by a high sequence similarity to their parents, while older, more diverged pseudogenes were characterised by a lower percent sequence similarity to parents. Next we binned pseudogenes by age. Given the large differences in the number of pseudogenes in the studied organisms, it was difficult to bin them consistently. Thus, we divided pseudogenes based on sequence similarities to their parents in 11, 11, 2 and 2 bins for mammals (human, macaque, and mouse), worm, fly and zebrafish respectively

(Fig. 1B, SI Appendix; S2). Consequently, in each mammal and worm bin there were on average 10% of the total number of pseudogenes. Due to the low numbers of pseudogenes in fly and zebrafish we chose a smaller number of bins, each containing on average 50% of the total number of pseudogenes.

(c) Repeats

We extracted genomic features such as CDS, UTR, and lncRNA for the human, worm and fly genome, leveraging existing available annotations (GENCODE 10, WormBase WS220 and FlyBase 5.45). We defined the genomic background as all un-gapped bases in the respective genomes. We used the repeat annotation for each genome from the UCSC Genome Browser, and extracted four major repeat classes: DNA, LINE (Long Interspersed Nuclear Elements), SINE (Short Interspersed Nuclear Elements) and LTR (Long terminal repeats). The repeat content for each annotation class or genome background was counted as the percentage of total nucleotides overlapping each of the repeat classes. Next, we analysed the sequence conservation using the PhastCons scores from the UCSC Genome Browser. For human, we used primate conservation scores; for worm, we used scores from alignments of *C. elegans* with 6 other worm strains; while for fly, we used scores from alignments of *D. melanogaster* with 14 different insects. For each annotation class or genome background, we calculated the average per nucleotide PhastCons score (SI Appendix, Fig. S3).

(d) Disablements

Analysing the sequence alignment between pseudogenes and parent genes obtained from PseudoPipe we identified three types of pseudogene disablements: insertions, deletions, and stop codons. We calculated the average defect density per pseudogene per megabase for each organism.

(e) Selection

Using the 1000 Genomes Project Phase 1 data we calculated the frequency of low coverage SNPs in pseudogene exons. As a proxy of the genomic average, we used the frequency of human low coverage SNPs in the upstream and downstream UTR exons of the pseudogenes. Overall, pseudogenes have a SNP frequency similar to the genomic average (SI Appendix; Fig. S5).

Next we calculated the derived allele frequency (DAF) for each pseudogene (SI Appendix Fig. S11). Overall, pseudogenes are enriched in rare alleles (DAF < 0.05).

Localization & Mobility

(a) Chromosomal localization

We defined three chromosomal regions: the telomere (T), the body, and the centromere (C). We defined two telomeric regions: one at the 5' end and one at the 3' end, each representing 15% of the chromosome length. The centromeric region was defined as the middle 30% of the chromosome, by length, while the remaining 40% (2x20% between the inner ends of the telomeres and the respective edges of the chromosome centre) was labelled as the chromosome body. In the case of acrocentric chromosomes, we defined the centromeric region around the geometrical middle of the chromosome. We calculated the pseudogene frequency in the telomeric and centromeric regions for each chromosome in human, worm

and fly. Based on these values, we calculated the average pseudogene frequency in the two regions for the entire genome (Fig. 2A). We used a two hypotheses binomial test to evaluate the statistical significance of the difference in the pseudogene frequency between the telomeric and the centromeric regions (SI Appendix; Table S1). The first hypothesis is that the pseudogenes are equally distributed at the centromeric and telomeric regions. The second hypothesis describes the observed distribution of pseudogenes in the centromeric and telomeric regions. As such, there are two options: "*" – the centromere has more pseudogenes than the telomere; and "#" – the telomere has more pseudogenes than the centromere. The significance threshold p-value was set to 0.05.

(b) Recombination

We obtained recombination rate estimates for human, worm, and fly, from Nato *et al.* (2011) [6], Rockman and Kruglyak (2009)[7] and Comeron *et al.* (2012) [8] respectively. We applied a simple linear interpolation from these datasets to obtain recombination rates for each nucleotide. We used the Tajima's D and Achaz's Y values from Andersen *et al.* (2012) [9]. In order to replicate results from their publication, we used a local polynomial regression smoothing for all data-points, before applying linear interpolation to obtain recombination rates for all nucleotides in the genome.

Due to the fact that recombination rates can differ within genes, we calculated the average recombination rates for pseudogenes by averaging their recombination rates across the length of each element, and then averaging this value for all pseudogenes. Error bars represent standard errors (Fig. 2A).

(c) Co-localisation tendency

We evaluated the tendency of pseudogenes to reside on the same chromosome as their parent genes using a 2-by-2 contingency table "A" (SI Appendix; Table S4), with elements $A_{i,j}$, where $i,j = \{1,2\}$:

- $A_{1,1}$ - the frequency of both the pseudogene and its parent residing on this chromosome;
- $A_{1,2}$ is the frequency of only the pseudogene residing on this chromosome;
- $A_{2,1}$ is the frequency of only the parent gene residing on this chromosome; and
- $A_{2,2}$ is the frequency of neither of the pseudogene or its parent residing on this chromosome.

We used Fischer's exact test to analyse whether pseudogenes and their parents tend to reside on the same chromosome. Using the Bonferroni correction, the significance threshold was set to $0.05/n$, where n is the total number of tested chromosomes in this species.

(d) Pseudogene mobility

We inspected pseudogene exchange between different chromosomes, excluding the pseudogenes that reside on the same chromosome as their parents. We used a Poisson regression model to detect chromosomes that display significant pseudogene exchange.

We hypothesised that on a chromosome, the pseudogene export / import frequency follows a Poisson distribution with the mean and variance proportional to the number of coding genes /

the chromosome size, respectively. Poisson regression was used to fit the pseudogene exchange frequency to the number of protein-coding genes / chromosome length. Any chromosome outside of the 95% prediction interval was considered to exhibit significant pseudogene exchanger (SI Appendix; Fig. S12).

Orthologs, Paralogs & Families

(a) Orthologs

We defined pseudogenes as orthologous if they were located in syntenic regions and their respective parent genes were orthologous. We obtained human-mouse synteny information from the UCSC Genome Browser chain alignments files for human HG19 and mouse MM9. Parent protein-coding gene orthology information was downloaded from the Ensembl website. The human-worm-fly orthologous protein-coding gene set was obtained by combining the MIT prepared orthologous gene list [10] with that obtained from the Ensembl. This totalled about 28,000 orthologous gene triplets of which 1935 were in a 1-1-1 relationship.

The lists of orthologous genes and pseudogenes can be found in the Associated Data Files.

(b) Paralogs

We obtained the protein-coding gene paralogs of all pseudogene parent genes from the Ensembl website.

(b) Family Membership

We grouped all pseudogenes into families according to their parents' membership to a family in the Pfam database [11, 12]. We ranked the families based on the number of corresponding pseudogenes. We grouped the top families containing 25% of the total number of pseudogenes in each organism based on their biological relationship.

Pseudogene Activity

We defined pseudogene activity based on four features: transcription potential, presence of Polymerase II (Pol II) and Transcription Factor (TF) binding sites in the upstream region of the pseudogenes, and chromatin accessibility.

(a) Transcription

In order to determine the list of potentially transcribed pseudogenes, we determined the RPKM (Reads Per Kilobase per Million mapped reads) values of each pseudogene in human, worm and fly. Among the transcribed pseudogenes, we also identified those with discordant expression patterns with their parent genes, using the PseudoSeq pipeline. Methods are described below.

- *RPKM*

We quantified the transcriptional activity for each pseudogene annotation using the following workflow. (i) For each nucleotide we calculated a mappability index as *1/m,* where *m* is the number of matches found in the genome for the 75 bp sequence starting at that nucleotide position allowing up to 2 mismatches. A mappability index of 1 indicates unique mapping. (ii) We filtered out pseudogene regions with mappability lower than 1. (iii) We discarded any

pseudogene regions shorter than 100 bp. (iv) We computed RPKM values for all unique pseudogene regions. (v) We set the human pseudogene RPKM selection threshold at 2. This value was chosen in agreement with previously published results [13, 14], which imply that on average 15% of human pseudogenes are transcribed. (vi) We evaluated the pseudogene RPKM selection threshold in worm and fly following the assumption that the transcription of protein-coding genes in human, worm and fly has a similar distribution. We applied quantile normalisation on the pooled "matched compendium" data for worm and fly, using human as a reference. This forces the transcription of protein-coding genes (but not the pseudogenes) to follow a similar distribution across the three organisms. (As a control, we also performed the normalisation on non-coding transcription instead of protein-coding genes and obtained consistent results.) (vii) We used the protein-coding gene normalisation to evaluate the RPKM selection threshold in worm and fly, obtaining 5.7 and 10.9 respectively. (viii) We used the calculated RPKM thresholds to obtain a list of transcribed pseudogenes in worm and fly respectively. For mouse, we used a similar approach following steps (i) to (vii) and obtained a RPKM selection threshold of 3.28. As a result we identified 878 transcribed pseudogenes in mouse.

- *PseudoSeq Pipeline*

PseudoSeq is a computational pipeline that makes use of RNA-Seq data from multiple tissues or developmental stages to compare the transcription of pseudogenes and their parents [14]. The pipeline maps RNA-Seq reads to the reference genome in conjunction with a splice junction library using Bowtie [15] and RSEQtools [16]. The signal tracks of the reads mapped to each pseudogene and its parent are generated across all the samples. Using this pipeline, we analysed pseudogene-parent expression correlation patterns. We found that pseudogenes may exhibit either concordant or discordant expression patterns with respect to their parents.

(b) Additional Activity Features

We defined the 2kb upstream of each pseudogene start site as the upstream region. We studied this region for the presence of Pol II and TF binding sites. The coordinates for Pol II and TFs were obtained from [17]. We annotated a pseudogene as Pol II active if at least 50% of the length of the Pol II binding site was included within the upstream region. Similarly, we annotated a pseudogene as TF active if at least 3 different TFs have at least 50% of their binding site within 2kb of the pseudogene start site.

Next, we analysed active chromatin in pseudogenes using chromatin segmentation for human (Segway [18]) and fly pseudogenes (9 State-Chromatin Segmentation [19]), and histone marks for worm pseudogenes. We analysed the distribution of the chromatin states along the pseudogene body. We annotated the human pseudogenes with an active chromatin label using a previously described model [14]. We compared the distribution of active and repressive marks in protein-coding genes. On average the ratio of the frequency of active to repressive chromatin marks for protein-coding genes is 5. Based on this analysis we developed a model for labelling pseudogenes with active chromatin. If the ratio of the frequency of active to repressive chromatin state marks was greater than or equal to 3, we labelled the pseudogene as having an active chromatin. The Segway active chromatin marks are GS (gene start), e/GM (enhancer, gene middle), GE (gene end), TSS (transcription start site). The Segway repressive chromatin marks are C (CTCF), R (repressive), F (FAIRE signal), L (low signal) and D (dead).

For fly, we looked at chromatin segmentation in 2 cell lines, S2 and BG3. If the ratio of the frequency of active chromatin marks to the frequency of repressed marks was larger than 2 in either of the cell lines, we labelled the pseudogene with an active chromatin tag. There are three active chromatin marks: Pro (promoter), Enh (enhancer) and Tnx (transcription); and three repressive marks: Rep (repressive), Het (heterochromatin) and Low (low signal).

Finally, we looked at the chromatin signatures of H3K4me3 and H3K4me1 in worm pseudogenes. We compared the signal intensities of these histone marks around the pseudogene body to coding gene signals. If the signals were of similar intensities, we labelled the pseudogene as having active chromatin.

(b) Upstream Sequence Analysis

We examined upstream proximal regions within 2kb of the annotated start sites for all pseudogenes, parent genes and paralogs.

We calculated the sequence similarity of the upstream regions between pseudogenes and parents, and between paralogs and parents using ClustalW2.1 [20]. For this process, we used the default settings of this alignment tool. The fraction of identical total nucleotides was calculated as the sequence similarity.

For the study of upstream sequence activity, we used H3K27Ac histone mark ChIP-Seq data [21] (uniformly processed signals with fold change calculated against control). The comparison is focused on protein-coding gene–pseudogene, 1-1 pairs where the parent gene does not have a corresponding gene paralog, and protein-coding gene–paralog 1-1 pairs where the protein-coding gene has one pseudogene.

In human, we analysed data from 15 cell lines: Dnd41, Gm12878, H1hesc, Helas3, Hepg2, Hmec, Hsmm, Hsmmt, Huvec, K562, Nha, Nhdfad, Nhek, Nhlf, Osteobl; in worm, we used data from three developmental stages (EE, L3, AD) while in fly we studied the EL and L3 developmental stages. For each upstream region, the normalised signal from each experiment was aggregated and averaged over the 2kb sequence. Using a threshold value of 1, we labelled regions as active if their signal values were higher than the set threshold in all the experiments considered. We labelled regions as inactive if their signal values were less than the defined threshold in all the experiments studied. For the parent-pseudogene-paralog trio set in Fig. 4C, the number of trios belonging to each of the four scenarios were counted.

“Functional” Pseudogene Candidates

(a) Pseudogene-parent Coexpression

To study pseudogene-parent co-expression patterns, we calculated Spearman correlations of expression levels (RPKM values in RNA-Seq) across different developmental stages or cell lines. In worm and fly, we used gene expression data across embryonic developmental stages (33 stages in worm, 30 stages in fly). In human, we used gene expression data across 19 human ENCODE cell lines.

(b) Translation

We used a proteo-genomic search to identify translated pseudogenes. (i) We generated putative peptides using 3-frame translation of annotated pseudogenes. (ii) We built a target peptide sequence database by merging the putative peptide and the complete human proteome datasets [22]. (iii) We used Peppy to map the target peptides against raw MS spectra

(available from [23]) under the default search settings [24]. The peptide identification false discovery rate was set lower than 0.01 using a target-decoy method. (iv) We refined the peptide-spectra matches by eliminating all peptides matching known proteins or variants (according to UniProt). Also we retained only the unique peptides identified at least twice in our analysed cell lines. (v) We annotated a pseudogene as putatively translated if it had two or more unique peptide matches.

The putatively translated pseudogenes were evaluated in terms of RNA expression (RPKM value) in the corresponding ENCODE human cell lines. We labelled the pseudogene translation candidates as highly confident if they had a RPKM value greater than 2. We used BLASTP [1] to compare sequence similarity between the pseudogene peptides and those originating from their parent protein.

References

1. Altschul SF, et al. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 25:3389-3402.
2. Searle SMJ, Gilbert J, Iyer V, Clamp M (2004) The otter annotation system. *Genome Res* 14:963-70.
3. Sonnhammer EL, Durbin R (1994) A workbench for large-scale sequence homology analysis. *Comput Appl Biosci* 10:301-307.
4. Slater GSC, Birney E (2005) Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* 6:31.
5. http://www.wormbase.org Last accessed on February, 24th 2014.
6. Nato AQ, Buyske S, Matise TC. The Rutgers Map: A third-generation combined linkage-physical map of the human genome. *In Preparation*.
7. Rockman MV & Kruglyak L (2009) Recombinational landscape and population genomics of Caenorhabditis elegans. *PLoS Genet* 5:e1000419.
8. Comeron JM, Ratnappan R, Bailin S (2012) The many landscapes of recombination in Drosophila melanogaster. *PLoS Genet* 8:e1002905.
9. Andersen EC, et al. (2012) Chromosome-scale selective sweeps shape Caenorhabditis elegans genomic diversity. *Nat Genet* 44:285-290.
10. Wu J, Bansal A, Rasmussen MD, Kellis M. Orthology identification and validation across human, mouse, fly, worm, yeast. *In preparation*.
11. Lam HYK, et al. (2009) Pseudofam: the pseudogene families database. *Nucleic Acids Res* 37:D738-43.
12. Punta M, et al. (2012) The Pfam protein families database. *Nucleic Acids Res* 40:D290-301.
13. Zheng D, et al. (2007) Pseudogenes in the ENCODE regions: consensus annotation, analysis of transcription, and evolution. *Genome Res* 17:839-851.
14. Pei B, et al. (2012) The GENCODE pseudogene resource. *Genome Biol* 13:R51.
15. Langmead B, Trapnell C, Pop M, Salzberg SL (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* 10:R25.
16. Habegger L, et al. (2011) RSEQtools: a modular framework to analyze RNA-Seq data using compact, anonymized data summaries. *Bioinformatics* 27:281-283.
17. http://data.modencode.org, Last accessed on April, 9th 2013.
18. Hoffman MM, et al. (2012) Unsupervised pattern discovery in human chromatin structure through genomic segmentation. *Nat Methods* 9:473-476.
19. http://compbio.med.harvard.edu/chromatin/ChromatinStates/ Last accessed on April, 9th 2013.
20. Sievers F, et al. (2011) Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol Syst Biol* 7:539.
21. Ho J, et al. (2014) modENCODE and ENCODE resources for analysis of metazoan chromatin organization. *Nature*, 10.1038/nature13415.
22. The UniProt Consortium (2014) Activities at the Universal Protein Resource (UniProt). *Nucleic Acids Res* 42:D191-D198.
23. Geiger T, Wehner A, Schaab C, Cox J, Mann M (2012) Comparative proteomic analysis of eleven common cell lines reveals ubiquitous but varying expression of most proteins. *Mol Cell Proteomics* 11:M111.014050.
24. Risk BA, Spitzer WJ, Giddings MC (2013) Peppy: proteogenomic search software. *J Proteome Res* 12:3019-3025.

Supplementary Figures and Tables

Annotation, Classification & Evolution

Fig. S1. Pseudogene annotation. The total number of pseudogenes annotated in human (red), worm (green), and fly (blue) respectively varies significantly from one release to another compared to the protein coding gene annotation.

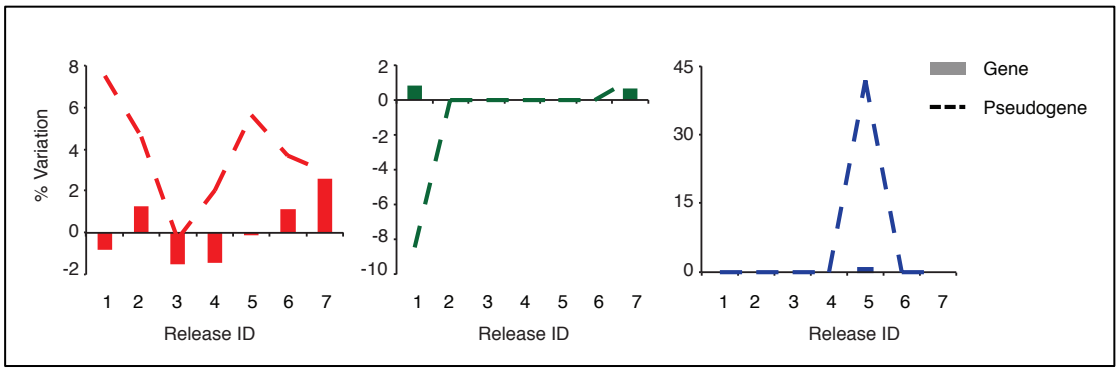
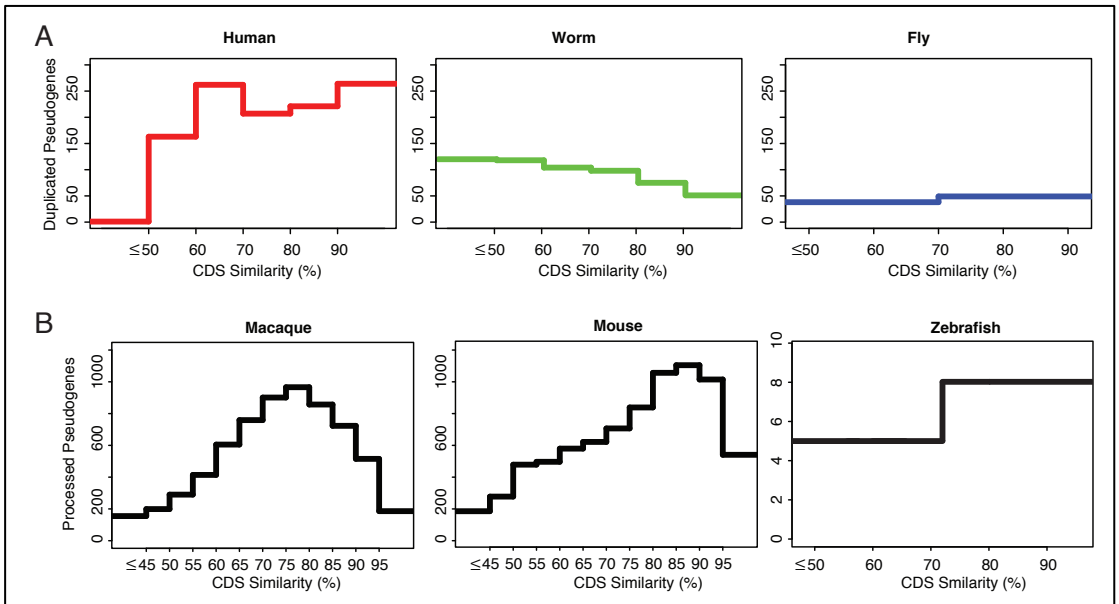


Fig. S2. Shadow figure for Fig. 1 (A) Distribution of duplicated pseudogenes in human, worm, and fly as function of age (sequence similarity to parents). (B) Distribution of processed pseudogenes in macaque, mouse, and zebrafish as function of age. (C) Disablements frequency of macaque, mouse and zebrafish.



C

Organisms	Defect / Pseudogene x MB		
	Insertion	Deletion	Stop Codon
Macaque	5.1	5.7	2.5
Mouse	5.6	7.7	2.6
Zebrafish	5.1	6.7	1.7

Fig. S3. Repeats distribution in human, worm, and fly.

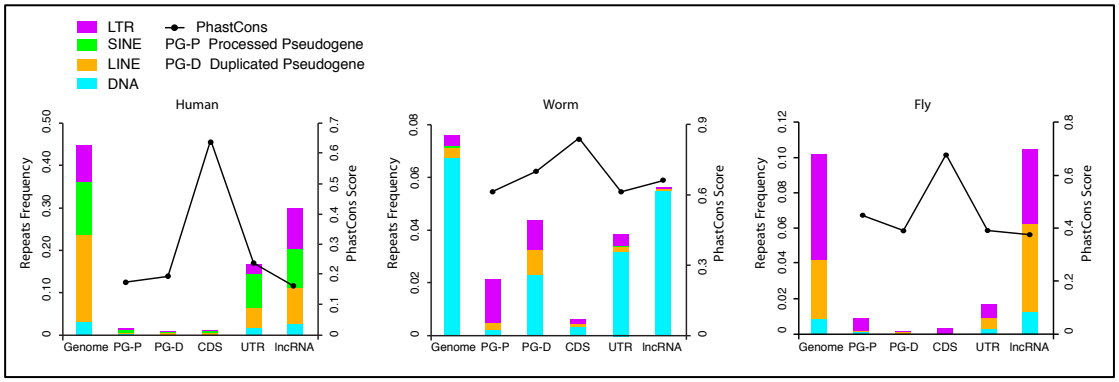


Fig. S4. Distribution of disablements in pseudogenes as function of type and pseudogene age.

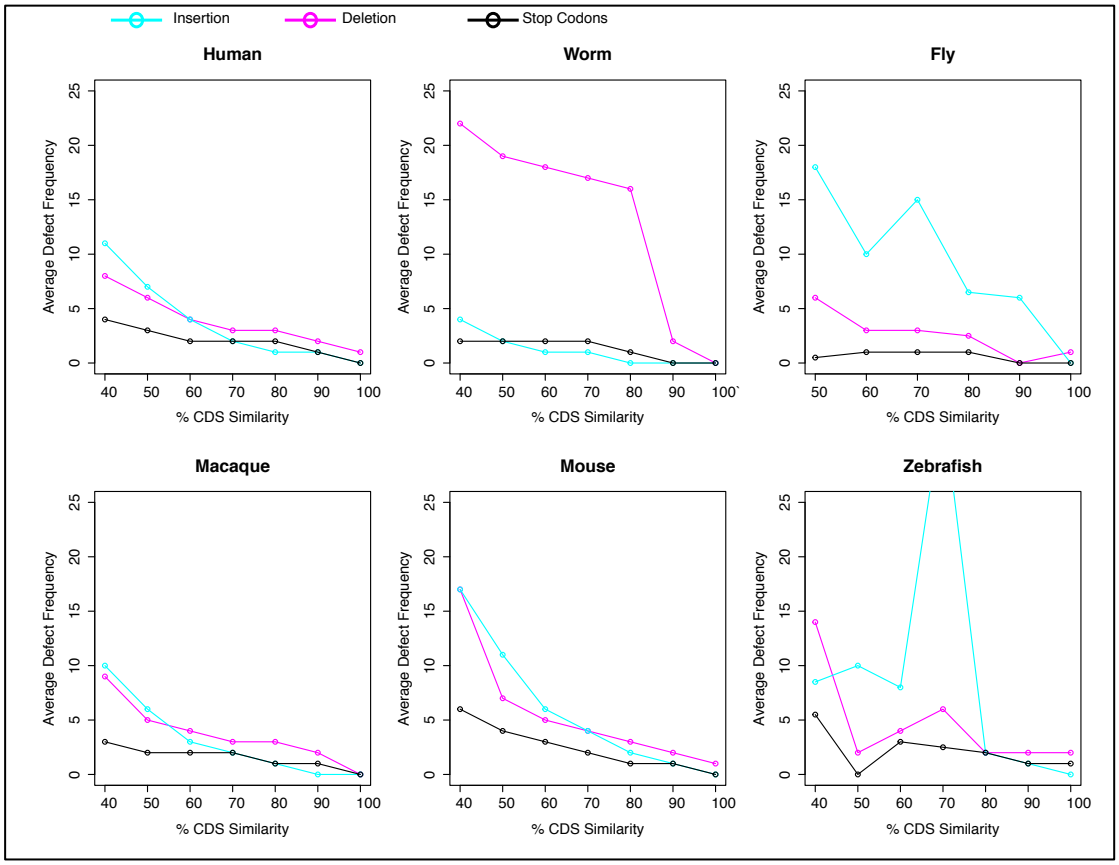
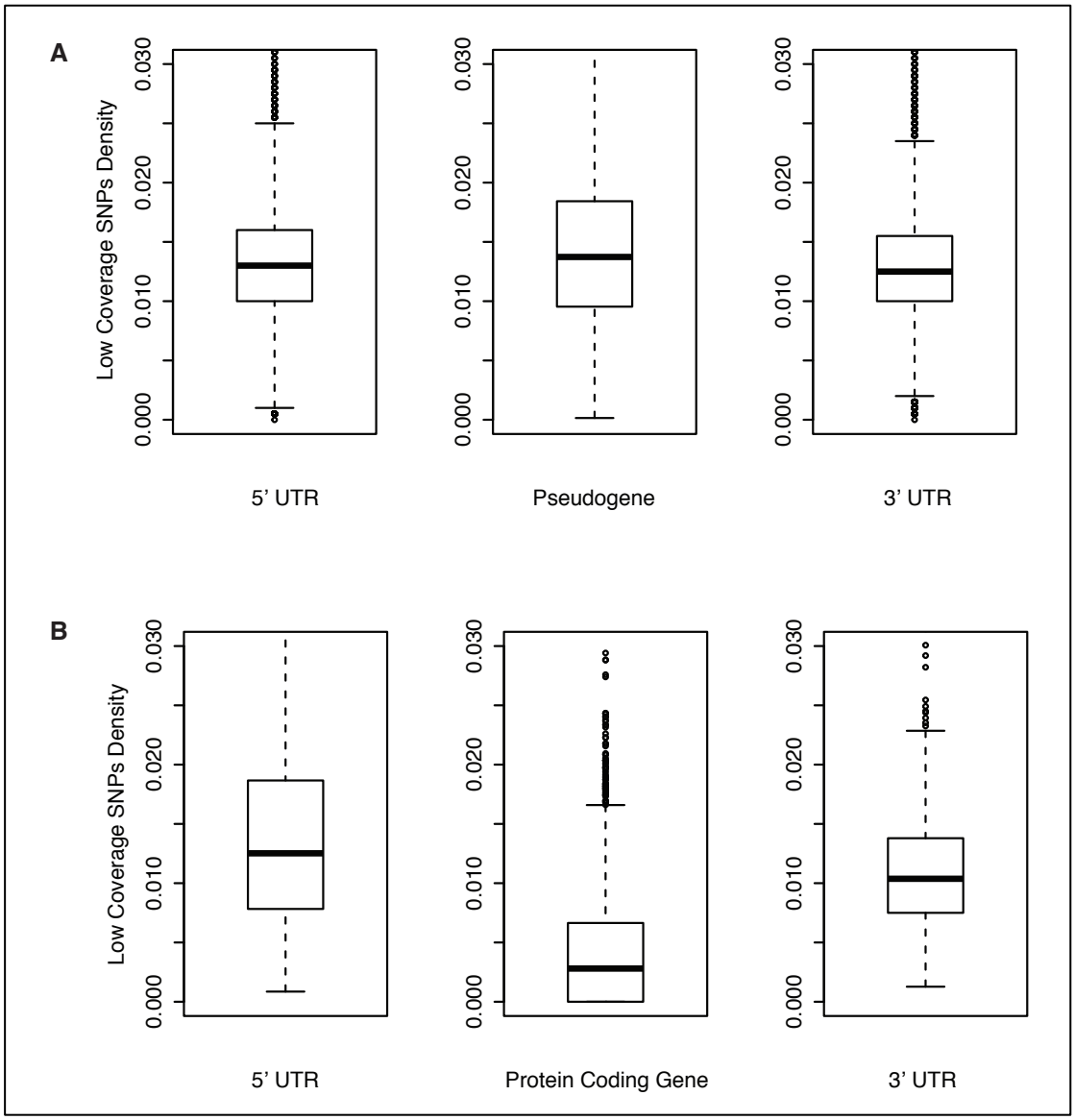


Fig. S5. Density distribution of human low coverage SNPs in (A) pseudogenes vs exonic 3' and 5' UTRs as compared to (B) protein coding genes.



Pseudogene localisation

Table S1. Pseudogene localisation. The significance of the pseudogene enrichment depending on the chromosomal localisation is assessed using a binomial test. # indicates that the chromosome telomeric regions are enriched in the number of pseudogenes compared to the centromeric regions. * indicates that there are significantly more pseudogenes in the middle of the chromosome compared to the end.

Table S1.1. Human pseudogene localisation

Chromosome	Telomere	Centromere	p-value	Significant?
1	359	326	9.03E-01	FALSE
2	186	357	9.47E-14	*TRUE
3	201	182	8.47E-01	FALSE
4	213	218	4.24E-01	FALSE
5	197	217	1.75E-01	FALSE
6	171	176	4.15E-01	FALSE
7	195	359	1.52E-12	*TRUE
8	183	176	6.64E-01	FALSE
9	139	330	2.81E-19	*TRUE
10	101	187	2.27E-07	*TRUE
11	225	298	8.08E-04	*TRUE
12	176	94	3.41E-07	#TRUE
13	33	102	1.08E-09	*TRUE
14	113	27	5.19E-14	#TRUE
15	13	20	1.48E-01	FALSE
16	28	16	4.81E-02	#TRUE
17	33	119	6.51E-13	*TRUE
18	10	13	3.39E-01	FALSE
19	61	24	3.69E-05	#TRUE
20	45	82	6.54E-04	*TRUE
21	18	46	3.09E-04	*TRUE
22	19	97	4.32E-14	*TRUE
X	183	300	5.69E-08	*TRUE
Y	67	161	2.08E-10	*TRUE
Whole Genome	2,969	3,927	4.02E-31	*TRUE

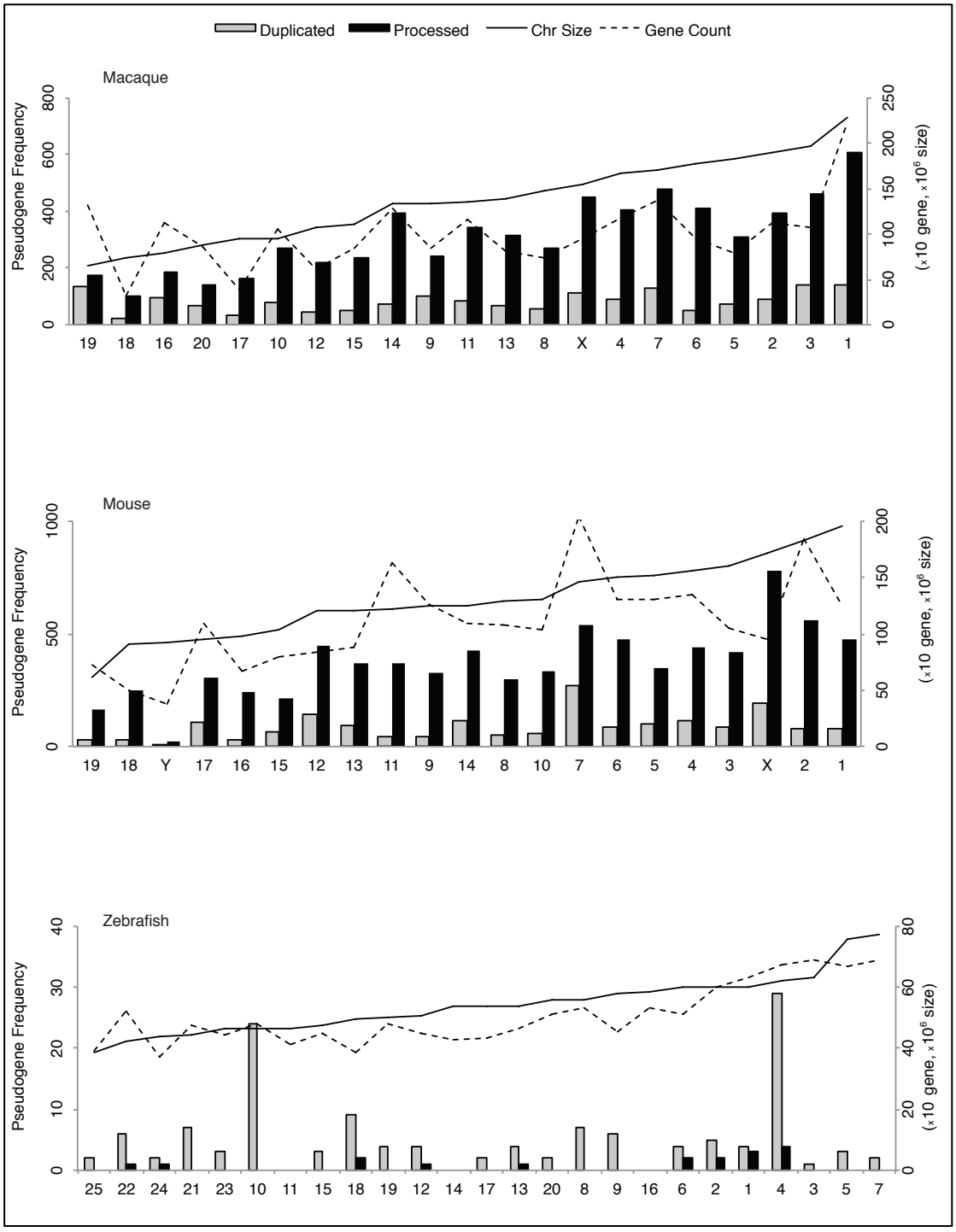
Table S1.2. Worm pseudogene localisation.

Chromosome	Telomere	Centromere	p-value	Significant?
I	36	10	7.82E-05	#TRUE
II	38	26	8.43E-02	FALSE
III	16	2	6.56E-04	#TRUE
IV	61	37	9.85E-03	#TRUE
V	120	74	5.89E-04	#TRUE
X	15	6	3.92E-02	#TRUE
Whole genome	286	155	2.25E-10	#TRUE

Table S1.3. Fly pseudogene localisation.

Chromosome	Telomere	Centromere	p-value	Significant?
2L	1	19	2.00E-05	*TRUE
2R	1	7	3.52E-02	*TRUE
3L	1	5	1.09E-01	FALSE
3R	4	7	2.74E-01	FALSE
X	2	12	6.47E-03	*TRUE
Whole Genome	9	50	2.63E-08	*TRUE

Fig. S6. Shadow figure for Fig. 2B. Distribution of pseudogenes per chromosome in macaque, mouse, and zebrafish. The chromosomes are sorted by length.



Orthologs, Paralogs and Family

Table S2. Pseudogenes associated with 1-1-1 orthologous genes in human, worm, and fly.

Organisms	Parent Genes	Pseudogenes
Human	560	2,145
Worm	8	8
Fly	8	15

Fig. S7. Human-Mouse orthologous pseudogenes distribution as function of pseudogene age, and activity (transcribed/ not transcribed).

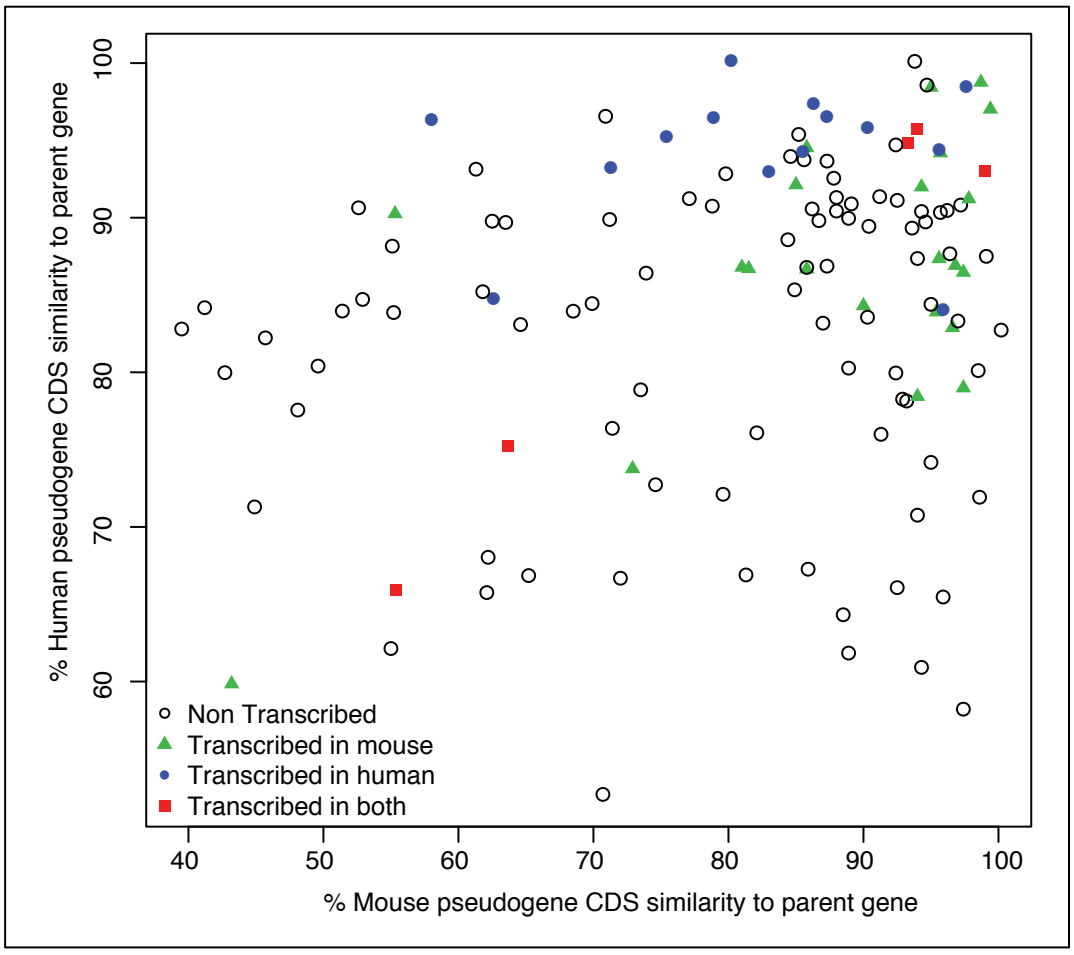
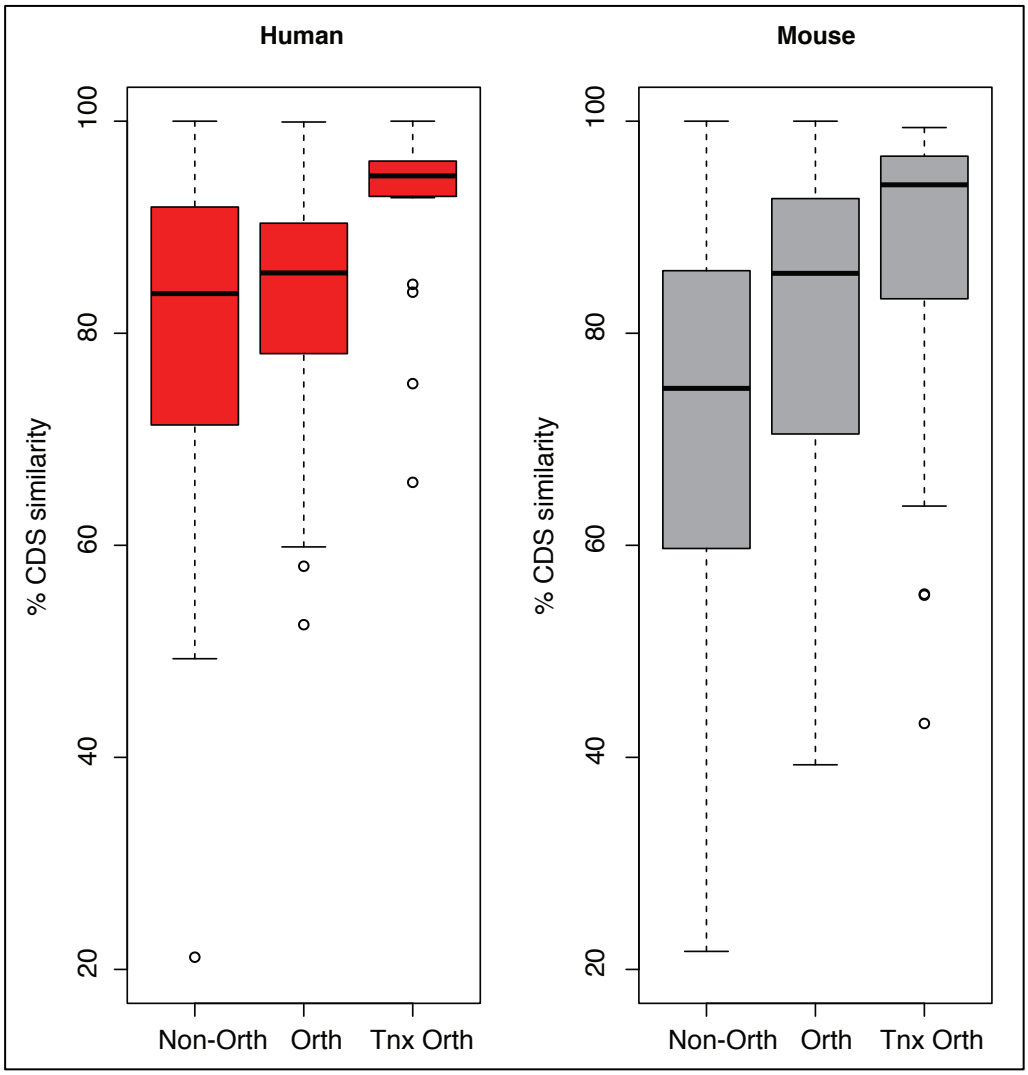


Fig. S8. Sequence conservation of human and mouse pseudogenes. Non-Orth = non orthologous human-mouse pseudogenes, Orth = orthologous human-mouse pseudogenes, Tnx Orth = transcribed orthologous human-mouse pseudogenes.



Activity & Function

Fig. S9. Broadly expressed parents of transcribed human pseudogenes. Transcribed human pseudogenes are binned based on the number of cell lines in which they are transcribed, and the fraction of broadly expressed parents over all the parents is calculated for each bin. The fraction increases, following the numbers of cell lines in which pseudogenes are transcribed.

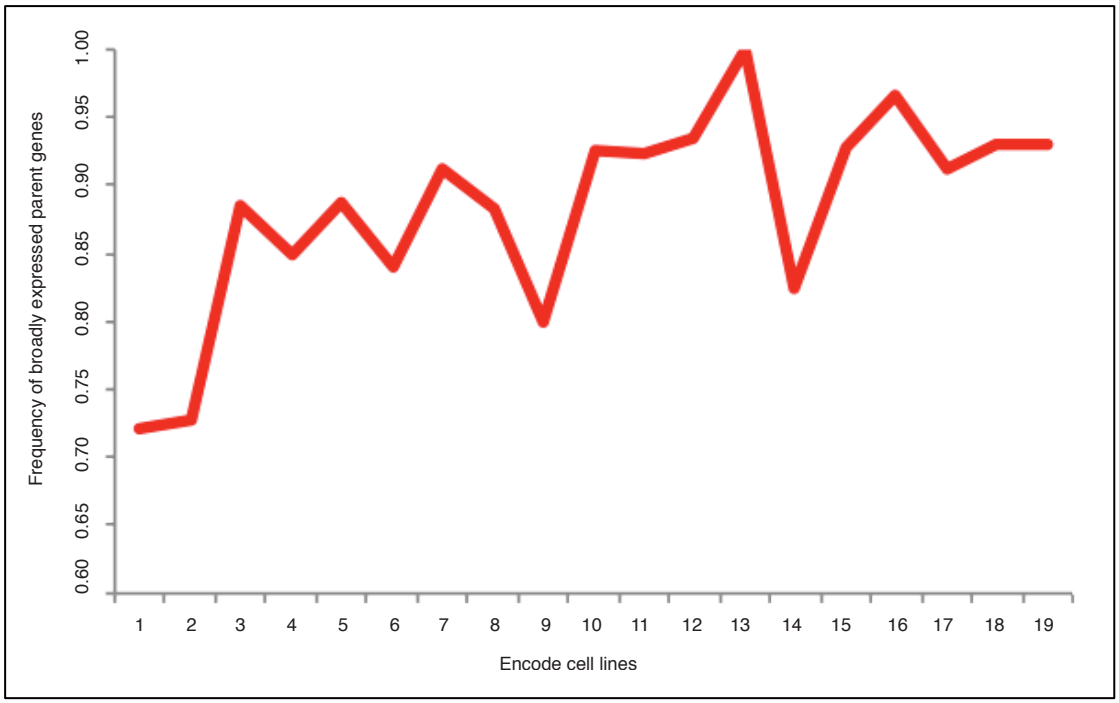


Fig. S10. Tissue specificity of transcribed pseudogenes. In human, the majority of transcribed pseudogenes are expressed in only one or a few cell lines, however a fraction of pseudogenes are universally transcribed. In worm, most pseudogenes are transcribed in only a few development stages. In fly, the specificity pattern is more evenly distributed than for human and worm.

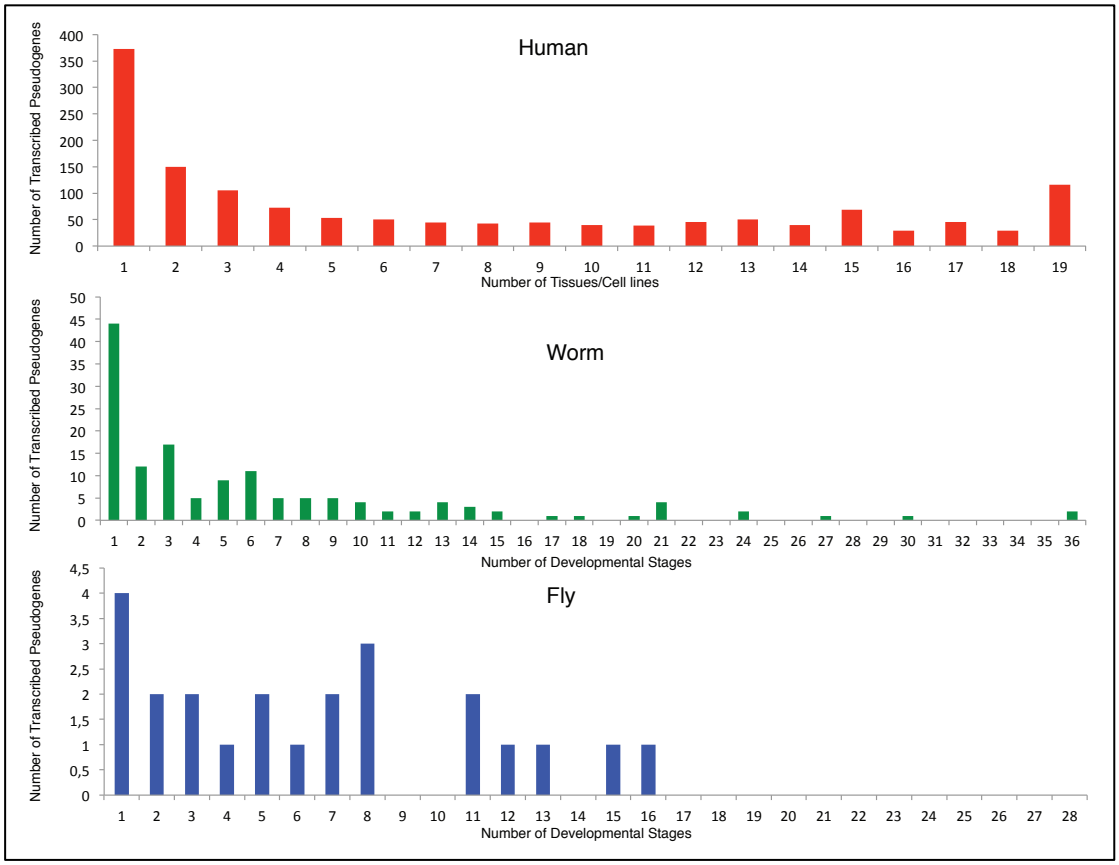


Fig. S11. Derived allele frequency for human pseudogenes. The pseudogenes are differentiated based on their activity levels: (A) transcribed vs. non-transcribed pseudogenes; and (B) highly-active, partially-active, and dead pseudogenes.

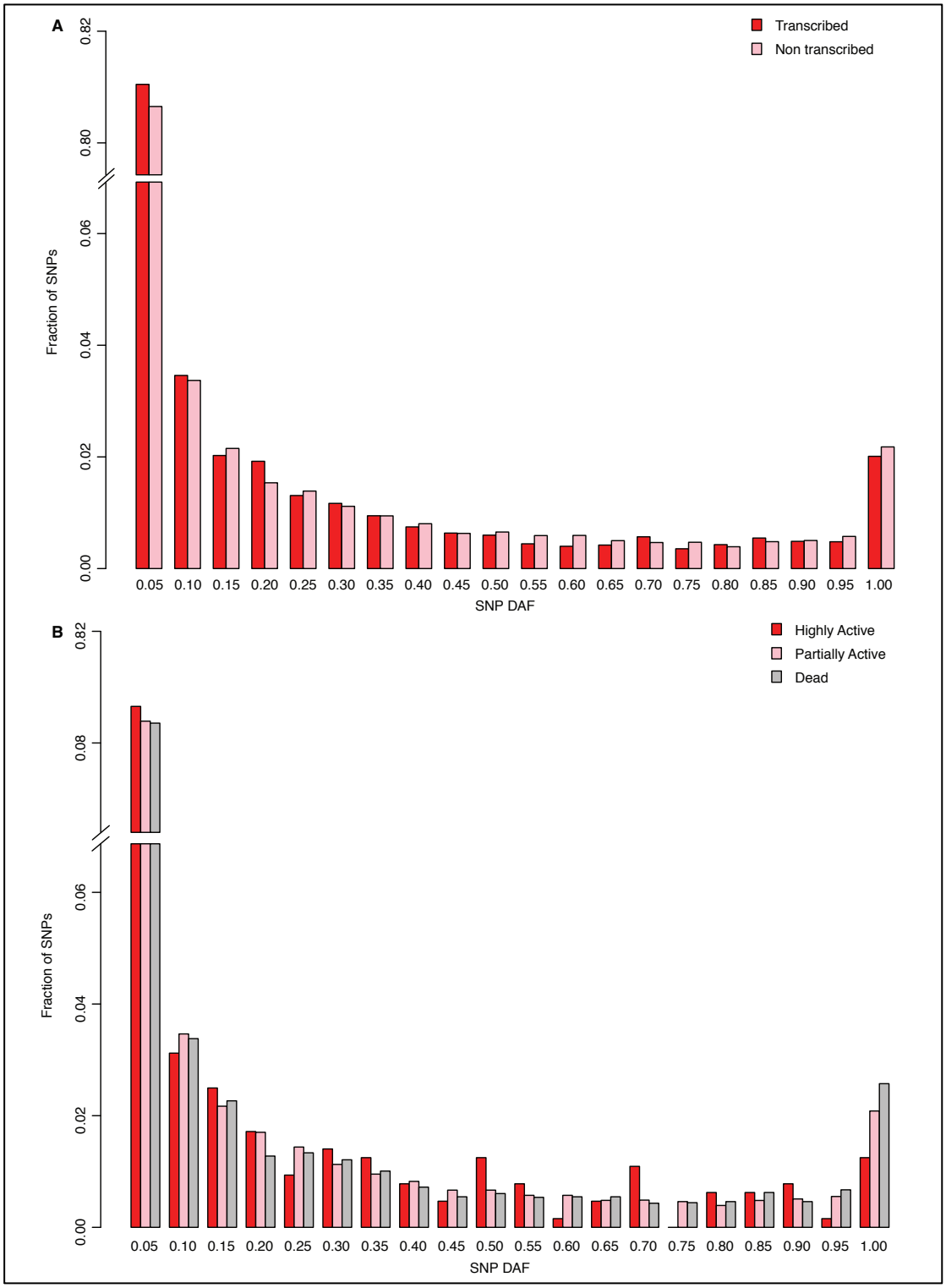


Table S3. Shadow table for Fig. 4D. Pseudogene translation candidates in human.

Translation Candidates	Parent Gene	Coex Coef	pVal	% Similarity CDS / UTR	Defect	Tnx	Pol II	AC	TF
SLIT-ROBO Rho GTPase activating protein 2B pseudogene (ENST00000491897)	ENST00000414359	0.80	5.9E-7	0.58 / 0.50	ins	✔	-	✔	-
PRKY-004, Y-linked protein kinase pseudogene (ENST00000533551)	ENST00000262848	-0.14	0.42	0.96 / 0.51	ins / del	✔	✔	-	✔
FER1L4-010, Fer-1-like 4 (C. elegans), pseudogene (ENST00000431615)	-	-0.38	0.03	0.62 / 0.32	ins / del	✔	-	✔	-

Pseudogene Mobility

Table S4. Contingency tables showing exchanges between the sex chromosomes and the pool of autosomal chromosomes. The diagonal values indicate the self-contribution of duplicated pseudogenes on the respective chromosomes. The values in the yellow coloured cells indicate the exchange between sex chromosomes and the pool of autosomes, while the values in the brown coloured cells refer to the exchange between the X and Y, chromosomes.

Table S4.1. Contingency table for human duplicated pseudogenes. Fisher's Exact Test (two-sided) p-value < 2.2e-16.

Pseudogene Location	Parent Gene Location		
	Autosome	X	Y
Autosomes	1092	14	1
X	11	42	0
Y	25	32	84

Table S4.2. Contingency table of human processed pseudogenes. Fisher's Exact Test (two-sided) p-value = 2.357e-6.

Pseudogene Location	Parent Gene Location		
	Autosome	X	Y
Autosomes	6611	292	3
X	537	39	1
Y	80	1	3

Table S4.3. Contingency table of worm duplicated pseudogenes. Fisher's Exact Test (two-sided) p-value = 0.0005386.

Pseudogene Location	Parent Gene Location	
	Autosome	X
Autosomes	391	7
X	13	4

Table S4.4. Contingency table of worm processed pseudogenes. Fisher's Exact Test (two-sided) p-value = 0.002919.

Pseudogene Location	Parent Gene Location	
	Autosome	X
Autosomes	131	0
X	6	2

Table S4.5. Contingency table of fly duplicated pseudogenes. Fisher's Exact Test (two-sided) p-value < 2.2e-16.

Pseudogene Location	Parent Gene Location		
	Autosome	X	Y
Autosomes	49	4	-
X	0	28	-
Y	4	0	-

Table S4.6 Contingency table of fly processed pseudogenes. Fisher's Exact Test (two-sided) p-value = 1. Note: Due to the low number of processed pseudogenes in fly, the colocalisation test is not statistically significant.

Pseudogene Location	Parent Gene Location	
	Autosome	X
Autosomes	7	2
X	2	1

Fig. S12. Detection of importer/exporter chromosomes (excluding colocalizing pseudogene-parent pairs and paralog-parent pairs). Detection of (A) importer and (B) exporter chromosomes for: paralogs (left), duplicated (middle), and processed (PSSD) pseudogenes (right). The think diagonal line is the Poisson regression fitting line. The grey vertical lines show the 95% prediction interval for each chromosome. If a point is above the corresponding prediction interval, the chromosome is considered a strong importer (in A) or exporter (in B). If a point is below the corresponding prediction interval, the chromosome is considered a weak importer (in A) or exporter (in B).

Fig. S12.1. Human

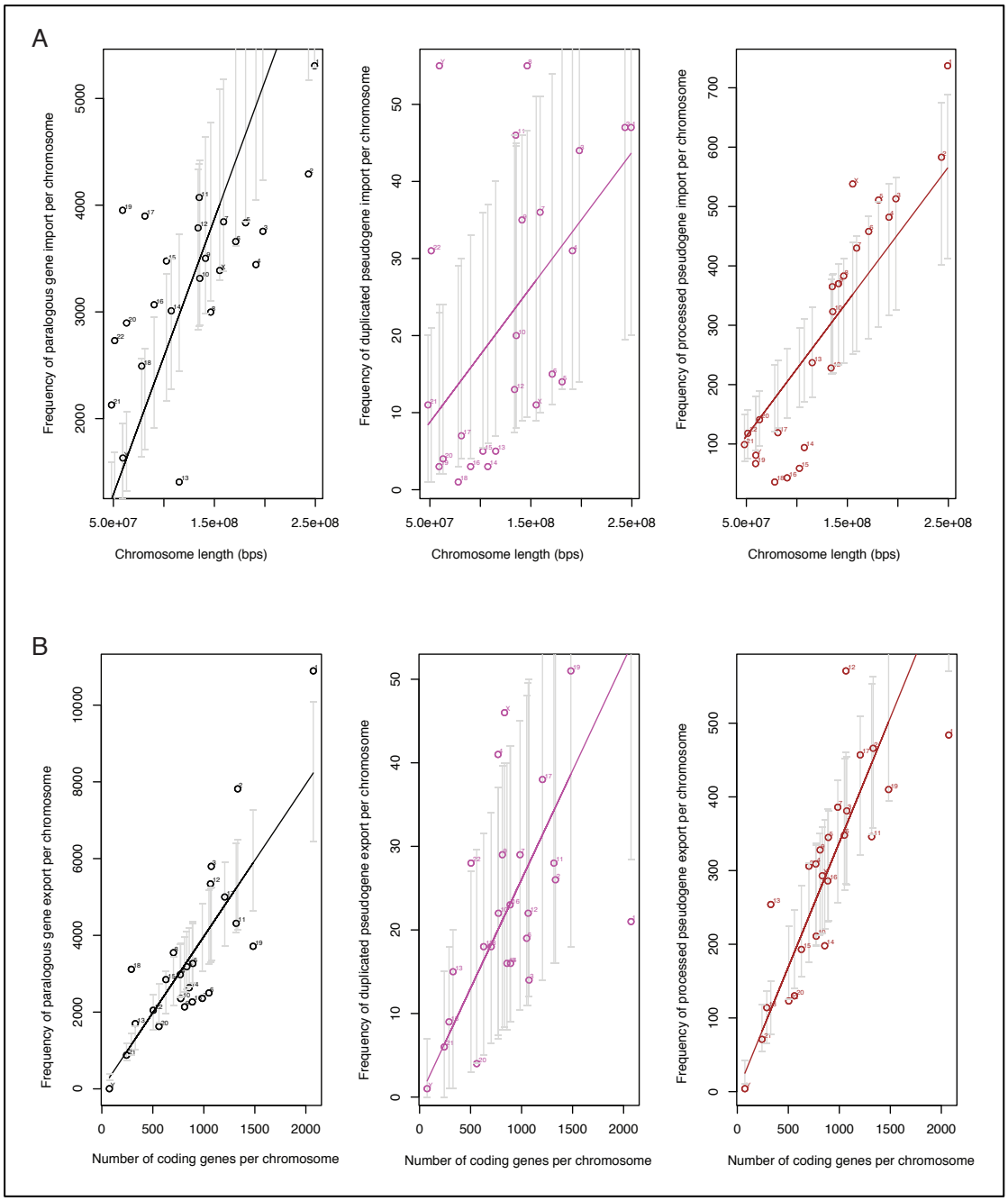


Fig. S12.2. Worm

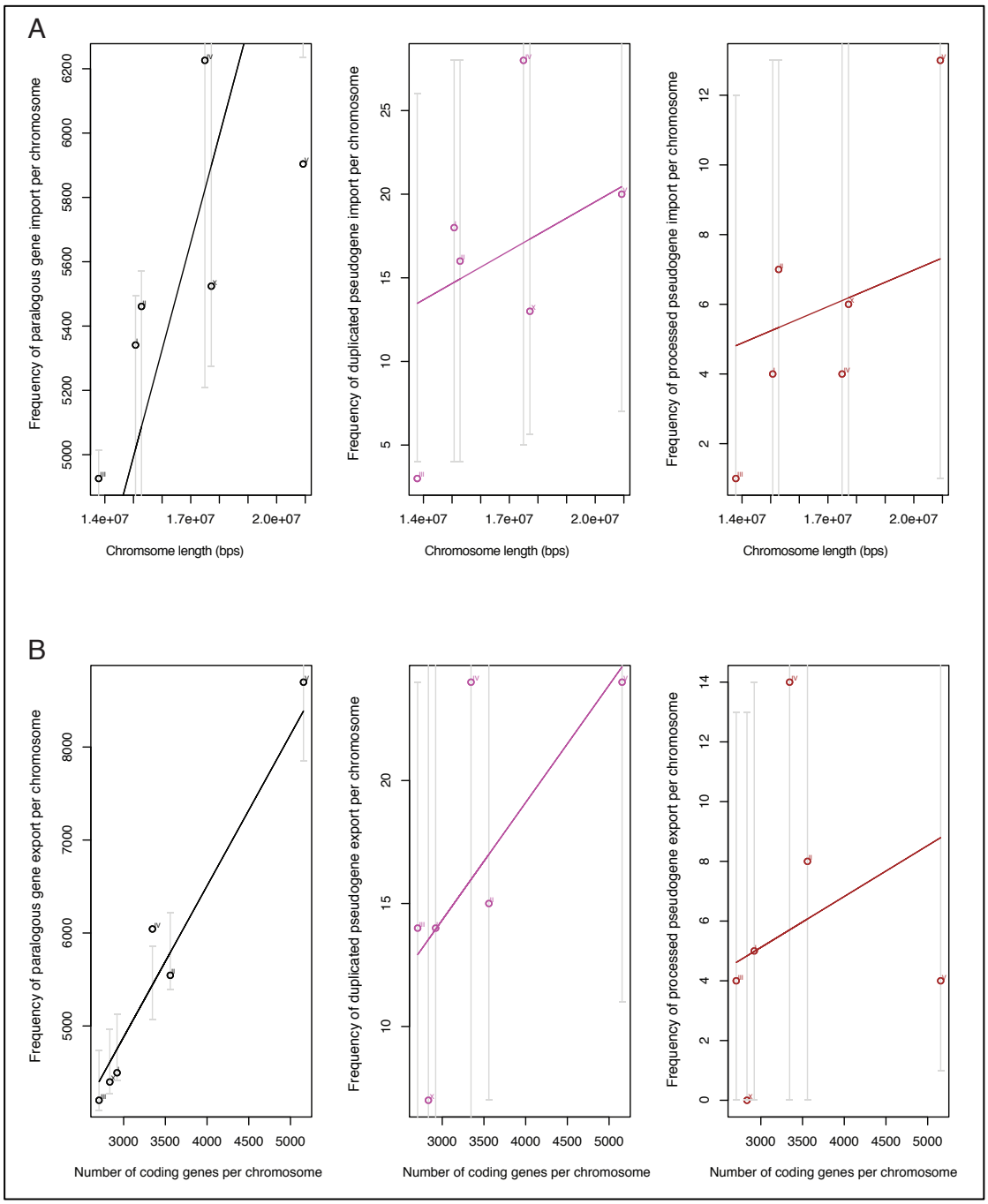
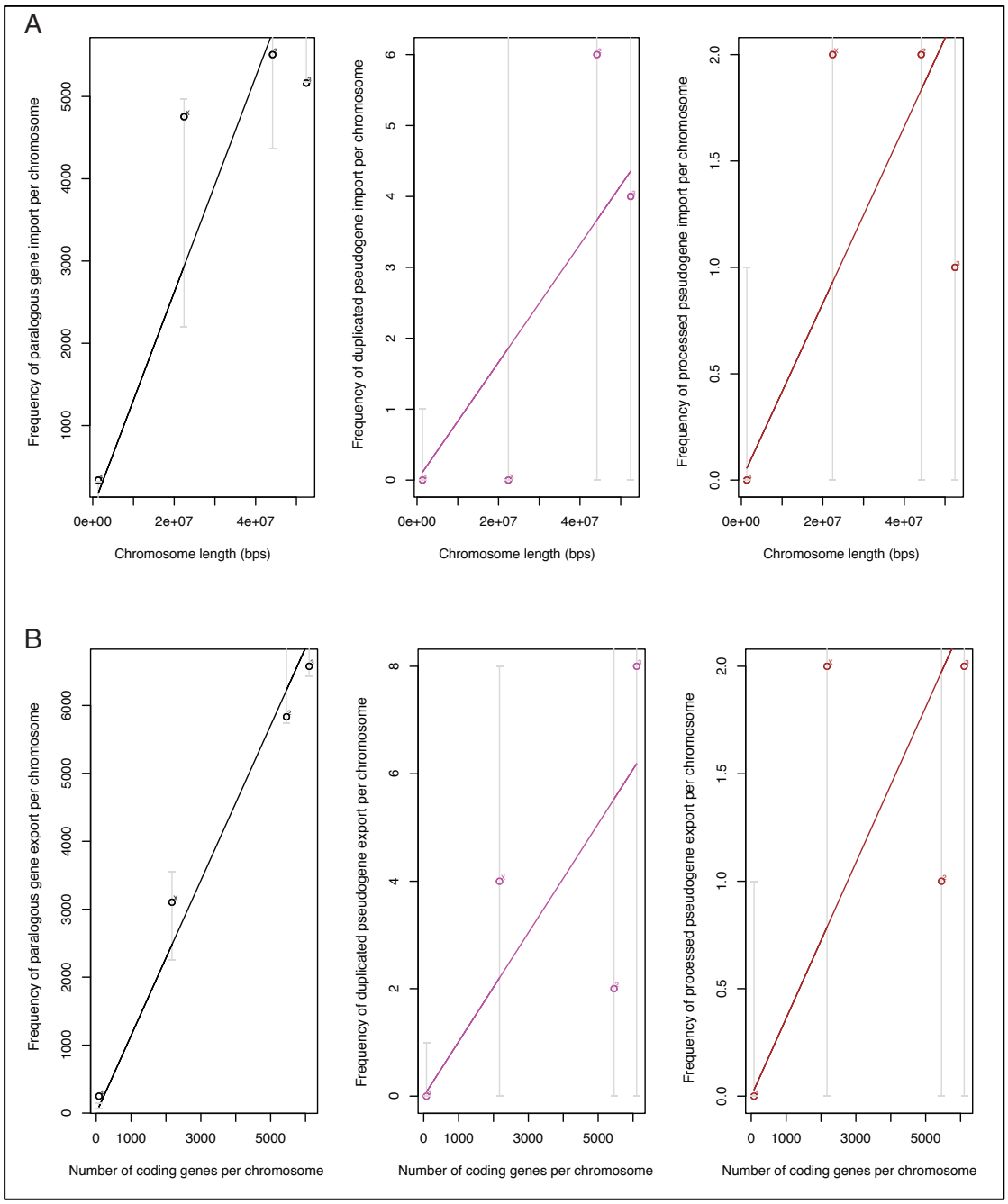


Fig. S12.3. Fly


A
Frequency of paralogous gene import per chromosome
Chromosome length (bps)
Frequency of duplicated pseudogene import per chromosome
Chromosome length (bps)
Frequency of processed pseudogene import per chromosome
Chromosome length (bps)
B
Frequency of paralogous gene export per chromosome
Number of coding genes per chromosome
Frequency of duplicated pseudogene export per chromosome
Number of coding genes per chromosome
Frequency of processed pseudogene export per chromosome
Number of coding genes per chromosome