

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 Rutenber-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}

^aProgram in Computational Biology and Bioinformatics and ^bDepartment of Molecular Biophysics and Biochemistry, Yale University, New Haven, CT 06520; ^cWellcome Trust Sanger Institute, Cambridge CB10 1SA, United Kingdom; ^dCenter for Biomolecular Science and Engineering, University of California, Santa Cruz, CA 95064; and ^eDepartment 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 (*Caenorhabditis 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>.

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

²To 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. 1A).

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. 1A).

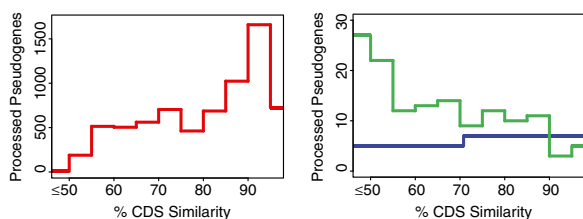
All of the data resulting from the annotation and comparative analysis are collected into a comprehensive online pseudogene resource: psicube.pseudogene.org.

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. 1A).

A

Organism	Total Pseudogenes	Biotype Distribution	ENCODE Functional Genomics Data	Completed Manual Annotation
		Processed Duplicated		
Human	12,358	8908 2266	✓	✓
Worm	911	159 566	✓	✓
Fly	145	16 109	✓	✓
Zebrafish	229	21 177	✓	✓
Macaque	11,136	6570 1726	×	×
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. 1B). 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. 1C 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. 2A). 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. 2B). 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. 2B 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

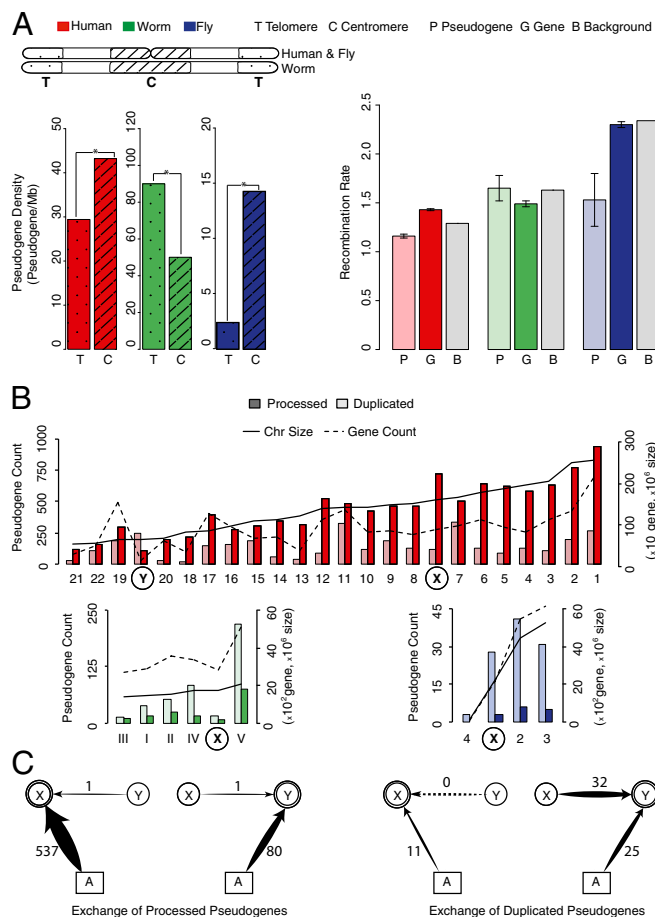


Fig. 2. Localization and mobility. (A, Left) The relative chromosomal localization preference for pseudogenes in human, worm, and fly. (Right) Average recombination rates for pseudogenes, protein-coding genes, and genomic background. (B) Distributions of processed and duplicated pseudogenes across chromosomes,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

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. 3A 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 (I). Also the number of pseudogenes associated with 1-1-1 protein-coding orthologs differs greatly across species. As an example (Fig. 3B), 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. 3C). 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. 3D). 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. 4A), 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

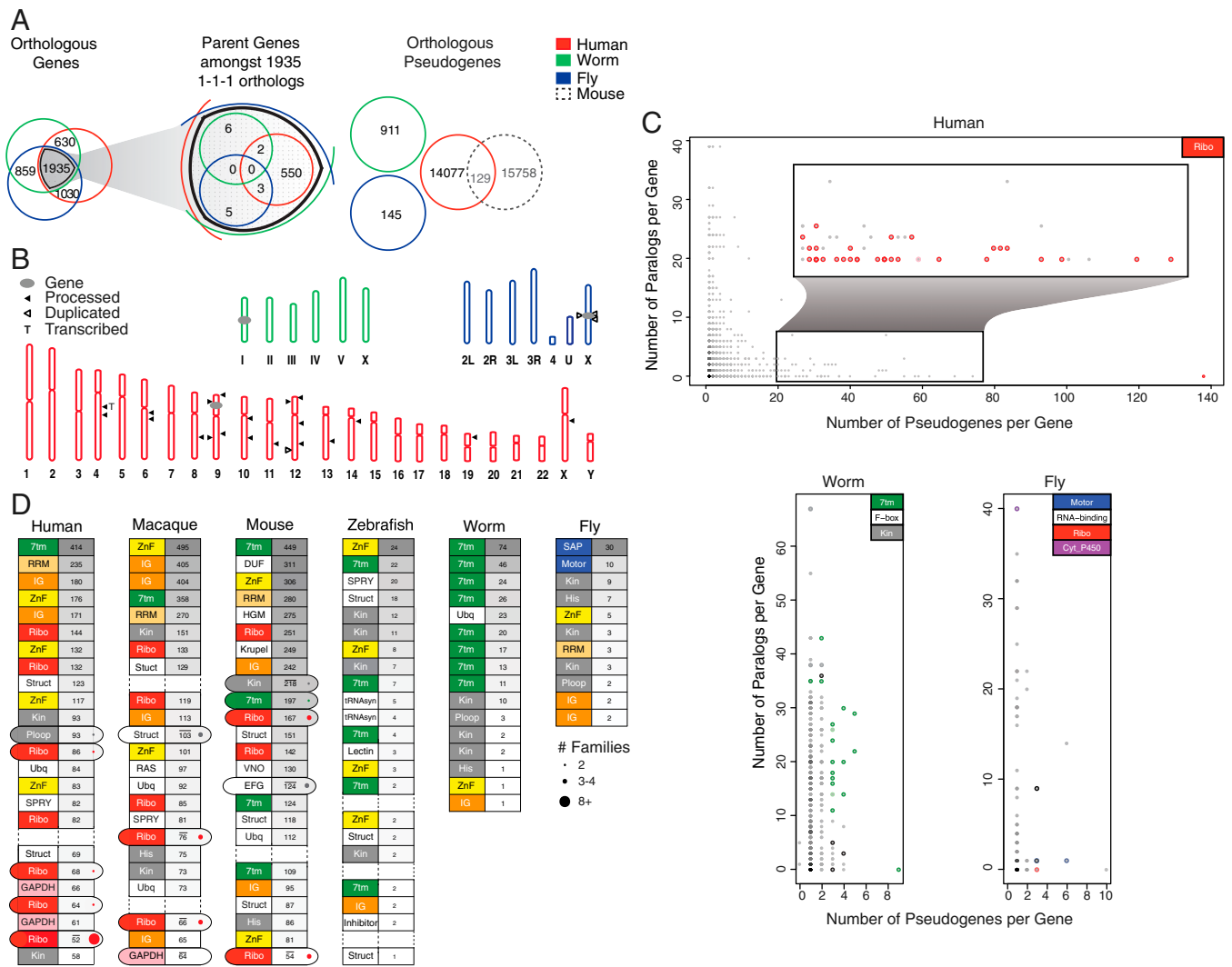


Fig. 3. Orthologs, paralogs, and families. (A) Venn diagrams showing the total number of orthologous genes and pseudogenes, in human, worm, and fly. (Right) Pseudogene orthologs between human and mouse. (B) Per chromosome distribution of RpS6 pseudogenes in human, worm, and fly. (C) Comparative distribution 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. 4B). 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. 4C).

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-PI* (8).

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

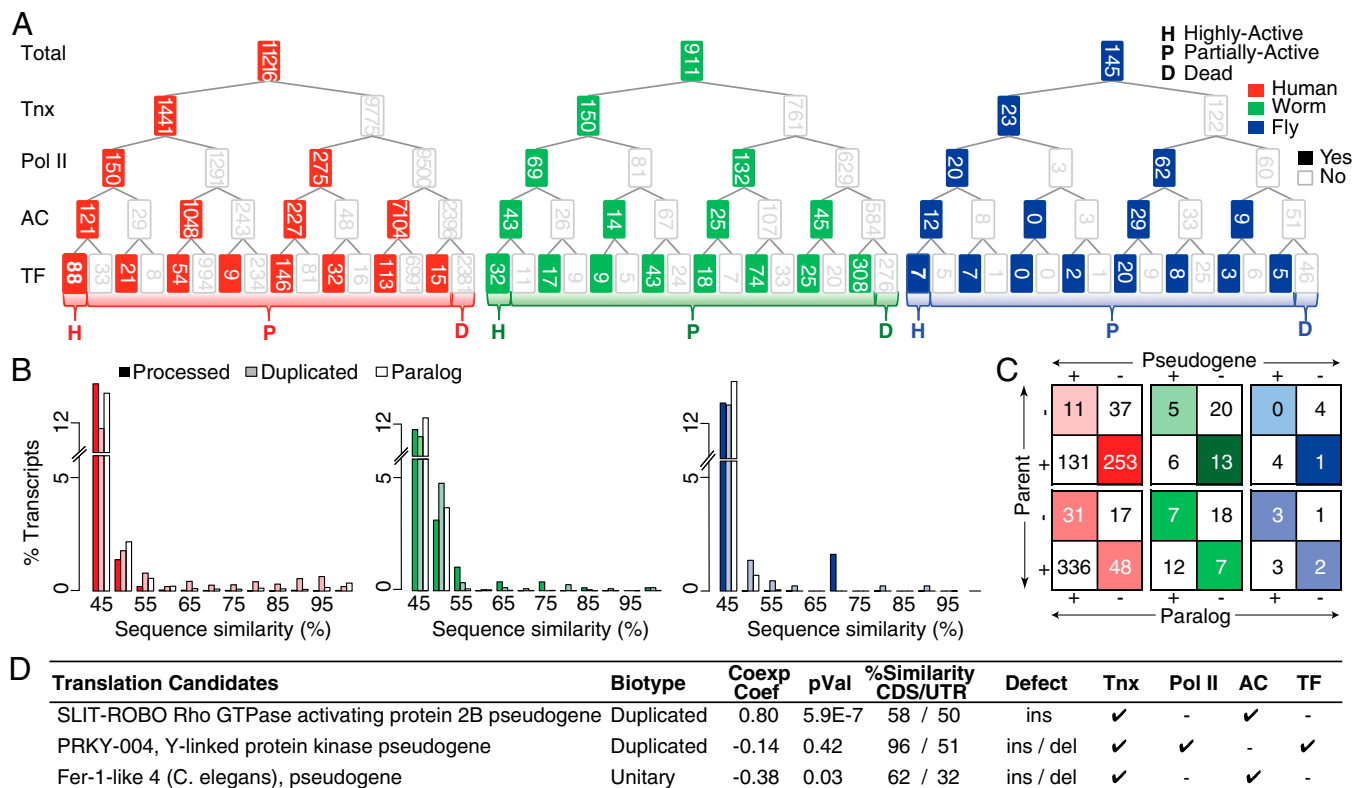


Fig. 4. Pseudogene activity. (A) Distribution of pseudogenes as a function of various activity features: transcription (Txn), active chromatin (AC), and presence of active Pol II and TF binding sites in the upstream region. (B) Conservation of the upstream sequences in processed and duplicated pseudogenes 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. 4D 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*.

- Zheng D, et al. (2007) Pseudogenes in the ENCODE regions: Consensus annotation, analysis of transcription, and evolution. *Genome Res* 17(6):839–851.
- Zhang Z, et al. (2006) PseudoPipe: An automated pseudogene identification pipeline. *Bioinformatics* 22(12):1437–1439.
- 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.
- Pei B, et al. (2012) The GENCODE pseudogene resource. *Genome Biol* 13(9):R51.
- Harrison PM, Zheng D, Zhang Z, Carriero N, Gerstein M (2005) Transcribed processed pseudogenes in the human genome: An intermediate form of expressed retro-sequence lacking protein-coding ability. *Nucleic Acids Res* 33(8):2374–2383.
- 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.
- Iskow RC, et al. (2012) Regulatory element copy number differences shape primate expression profiles. *Proc Natl Acad Sci USA* 109(31):12656–12661.
- Poliseno L, et al. (2010) A coding-independent function of gene and pseudogene mRNAs regulates tumour biology. *Nature* 465(7301):1033–1038.
- Muro EM, Mah N, Andrade-Navarro MA (2011) Functional evidence of post-transcriptional regulation by pseudogenes. *Biochimie* 93(11):1916–1921.
- Petrov DA, Hartl DL (2000) Pseudogene evolution and natural selection for a compact genome. *J Hered* 91(3):221–227.
- Ophir R, Graur D (1997) Patterns and rates of indel evolution in processed pseudogenes from humans and murids. *Gene* 205(1–2):191–202.
- 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.
- Karro JE, et al. (2007) Pseudogene.org: A comprehensive database and comparison platform for pseudogene annotation. *Nucleic Acids Res* 35(Database issue):D55–D60.
- 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.
- 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.
- Howe K, et al. (2013) The zebrafish reference genome sequence and its relationship to the human genome. *Nature* 496(7446):498–503.
- Fairbanks DJ, Maughan PJ (2006) Evolution of the NANOG pseudogene family in the human and chimpanzee genomes. *BMC Evol Biol* 6:12.
- 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.
- Harrison PM, Gerstein M (2002) Studying genomes through the aeons: Protein families, pseudogenes and proteome evolution. *J Mol Biol* 318(5):1155–1174.
- Balasubramanian S, et al. (2009) Comparative analysis of processed ribosomal protein pseudogenes in four mammalian genomes. *Genome Biol* 10(1):R2.
- Gerstein MB, et al. (2014) Comparative analysis of the transcriptome across distant species. *Nature*, 10.1038/nature13424.
- Boyle AP, et al. (2014) Comparative analysis of regulatory information and circuits across distant species. *Nature*, 10.1038/nature13668.
- Mutimer H, Deacon N, Crowe S, Sonza S (1998) Pitfalls of processed pseudogenes in RT-PCR. *Biotechniques* 24(4):585–588.
- Garbay B, Boue-Grabot E, Garret M (1996) Processed pseudogenes interfere with reverse transcriptase-polymerase chain reaction controls. *Anal Biochem* 237(1):157–159.
- Torrents D, Suyama M, Zdobnov E, Bork P (2003) A genome-wide survey of human pseudogenes. *Genome Res* 13(12):2559–2567.
- 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.
- Ding W, Lin L, Chen B, Dai J (2006) L1 elements, processed pseudogenes and retrogenes in mammalian genomes. *IUBMB Life* 58(12):677–685.
- Yang H-P, Barbash DA (2008) Abundant and species-specific DINE-1 transposable elements in 12 *Drosophila* genomes. *Genome Biol* 9(2):R39.
- Andersen EC, et al. (2012) Chromosome-scale selective sweeps shape *Caenorhabditis elegans* genomic diversity. *Nat Genet* 44(3):285–290.
- Barnes TM, Kohara Y, Coulson A, Hekimi S (1995) Meiotic recombination, noncoding DNA and genomic organization in *Caenorhabditis elegans*. *Genetics* 141(1):159–179.
- Hillier LW, et al. (2003) The DNA sequence of human chromosome 7. *Nature* 424(6945):157–164.
- Glusman G, Yanai I, Rubin I, Lancet D (2001) The complete human olfactory sub-genome. *Genome Res* 11(5):685–702.
- 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.
- Jensen-Seaman MI, et al. (2004) Comparative recombination rates in the rat, mouse, and human genomes. *Genome Res* 14(4):528–538.
- Emerson JJ, Kaessmann H, Betrán E, Long M (2004) Extensive gene traffic on the mammalian X chromosome. *Science* 303(5657):537–540.
- Castillo-Davis CI, Hartl DL (2002) Genome evolution and developmental constraint in *Caenorhabditis elegans*. *Mol Biol Evol* 19(5):728–735.
- Thomas JH, Robertson HM (2008) The *Caenorhabditis* chemoreceptor gene families. *BMC Biol* 6:42.
- Ishii K, et al. (2006) Characteristics and clustering of human ribosomal protein genes. *BMC Genomics* 7:37.
- Pan D, Zhang L (2009) Burst of young retrogenes and independent retrogene formation in mammals. *PLoS ONE* 4(3):e5040.
- 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.
- Zhao S, et al. (2004) Human, mouse, and rat genome large-scale rearrangements: Stability versus speciation. *Genome Res* 14(10A):1851–1860.
- Waterston RH, et al.; Mouse Genome Sequencing Consortium (2002) Initial sequencing and comparative analysis of the mouse genome. *Nature* 420(6915):520–562.
- 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.
- Lynch M, Conery JS (2003) The origins of genome complexity. *Science* 302(5649):1401–1404.
- Luque T, Marfany G, González-Duarte R (1997) Characterization and molecular analysis of *Adh* retrosequences in species of the *Drosophila obscura* group. *Mol Biol Evol* 14(12):1316–1325.
- Heard E, Disteché CM (2006) Dosage compensation in mammals: Fine-tuning the expression of the X chromosome. *Genes Dev* 20(14):1848–1867.
- Wong A, et al. (2004) Diverse fates of paralogs following segmental duplication of telomeric genes. *Genomics* 84(2):239–247.
- Piehler AP, et al. (2008) The human ABC transporter pseudogene family: Evidence for transcription and gene-pseudogene interference. *BMC Genomics* 9:165.
- Tam OH, et al. (2008) Pseudogene-derived small interfering RNAs regulate gene expression in mouse oocytes. *Nature* 453(7194):534–538.
- Rapicavoli NA, et al. (2013) A mammalian pseudogene lncRNA at the interface of inflammation and anti-inflammatory therapeutics. *eLife* 2:e00762.
- Snyder M, Gerstein M (2003) Genomics. Defining genes in the genomics era. *Science* 300(5617):258–260.
- Sasidharan R, Gerstein M (2008) Genomics: Protein fossils live on as RNA. *Nature* 453(7196):729–731.

13366 | www.pnas.org/cgi/doi/10.1073/pnas.1407293111

Sisu et al.