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.

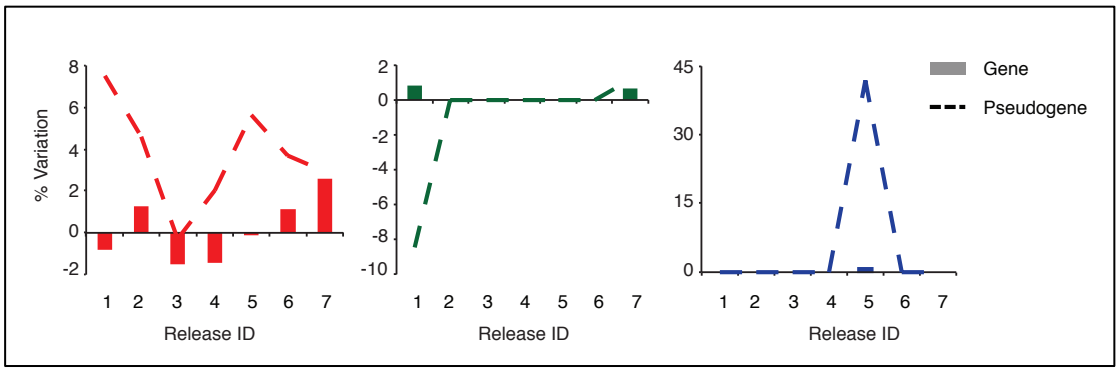
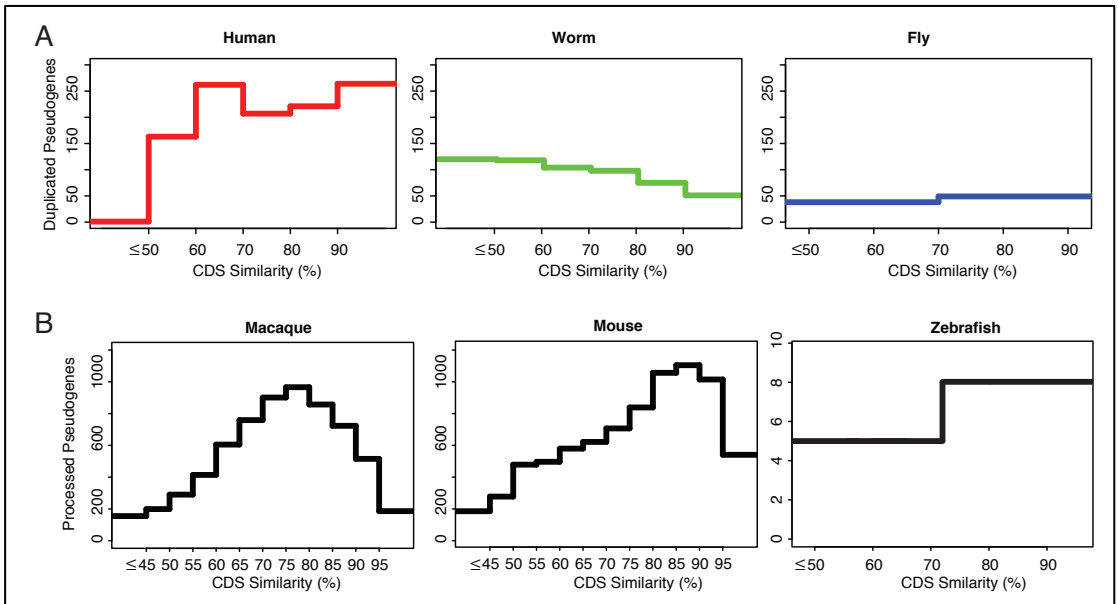


Fig. S2. Shadow figure for Fig. 1 (A) Distribution of duplicated pseudogenes in human, worm, and fly as function of age (sequence similarity to parents). (B) Distribution of processed pseudogenes in macaque, mouse, and zebrafish as function of age. (C) Disablements frequency of macaque, mouse and zebrafish.



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.

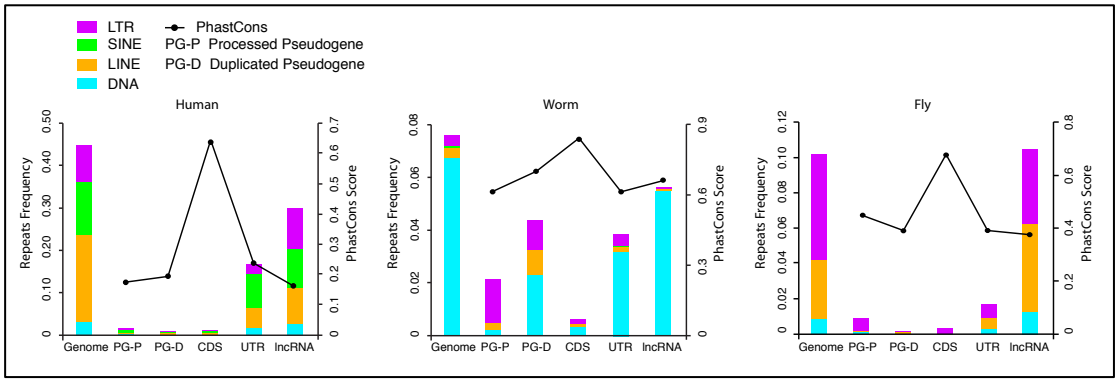


Fig. S4. Distribution of disablements in pseudogenes as function of type and pseudogene age.

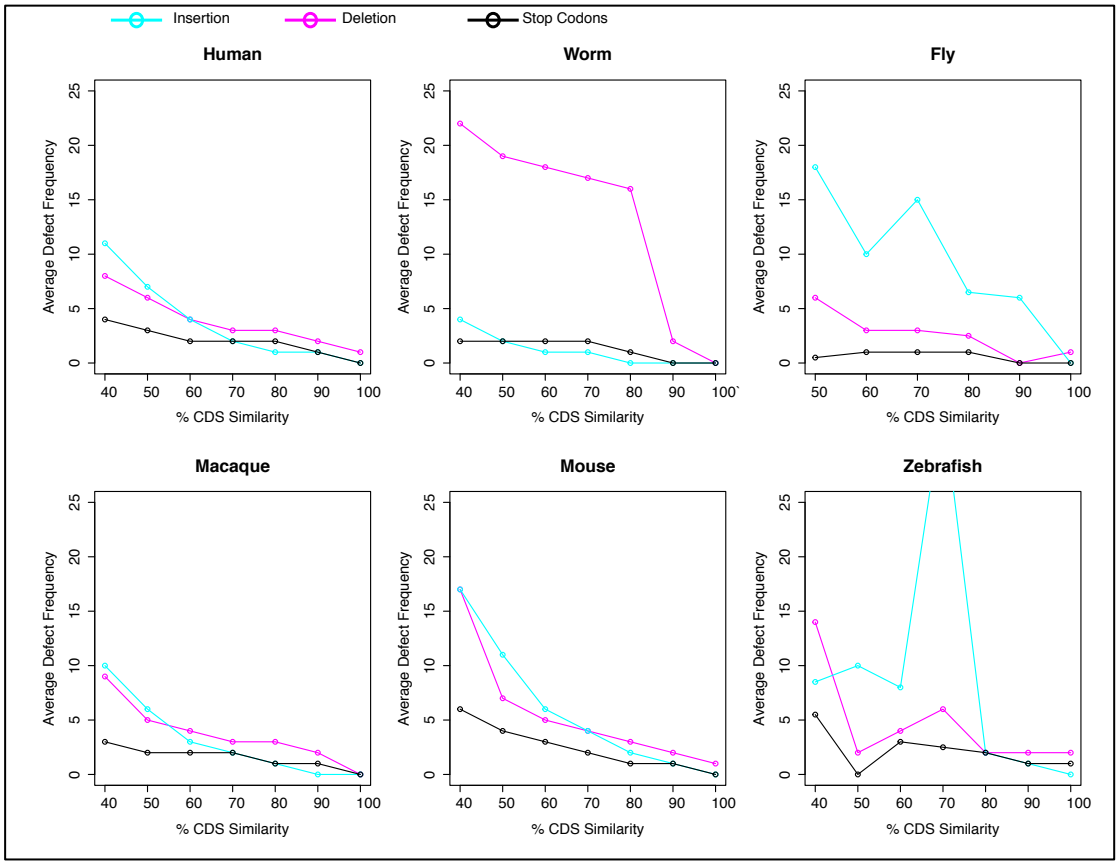
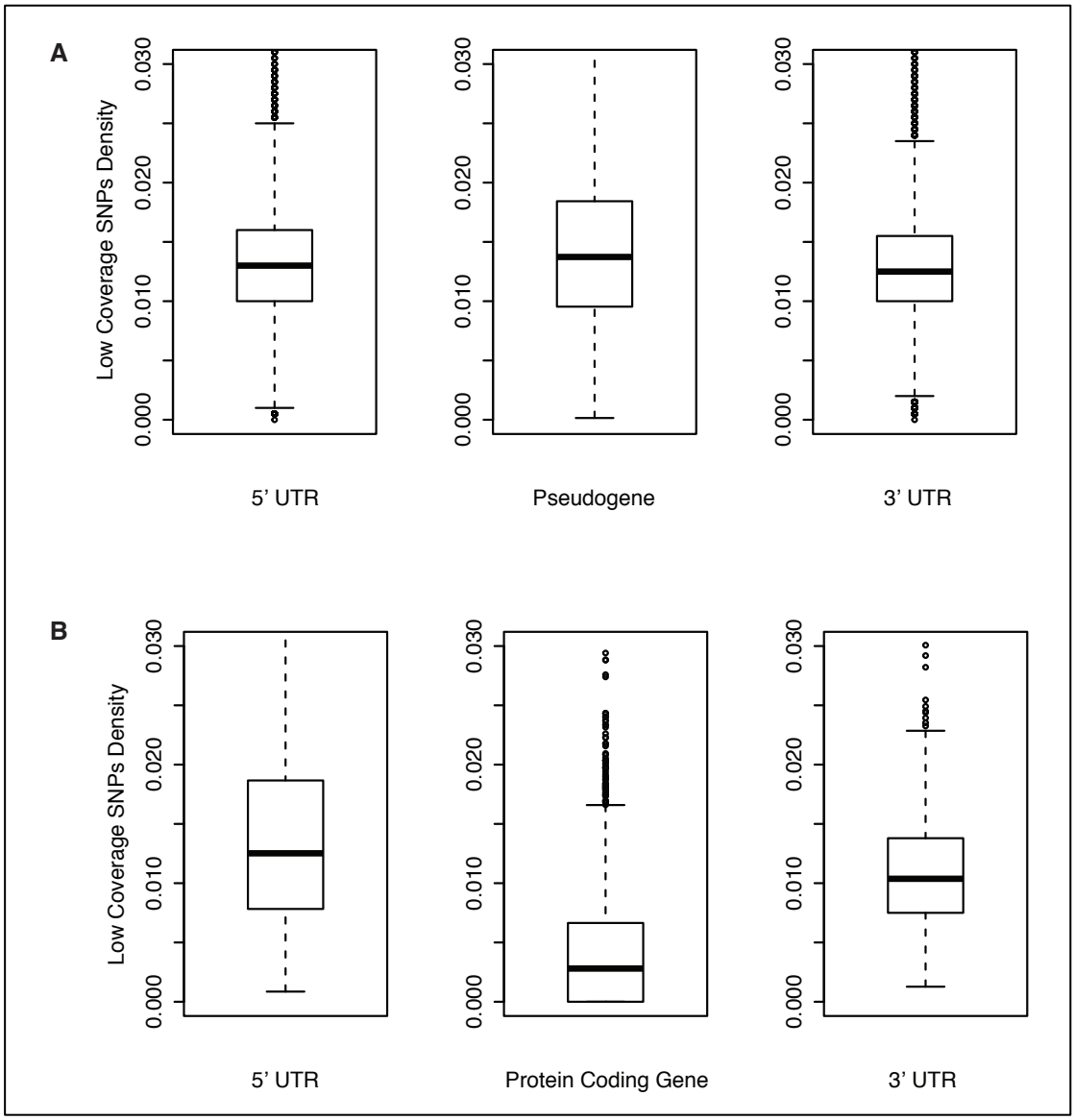


Fig. S5. Density distribution of human low coverage SNPs in (A) pseudogenes vs exonic 3' and 5' UTRs as compared to (B) protein coding genes.



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

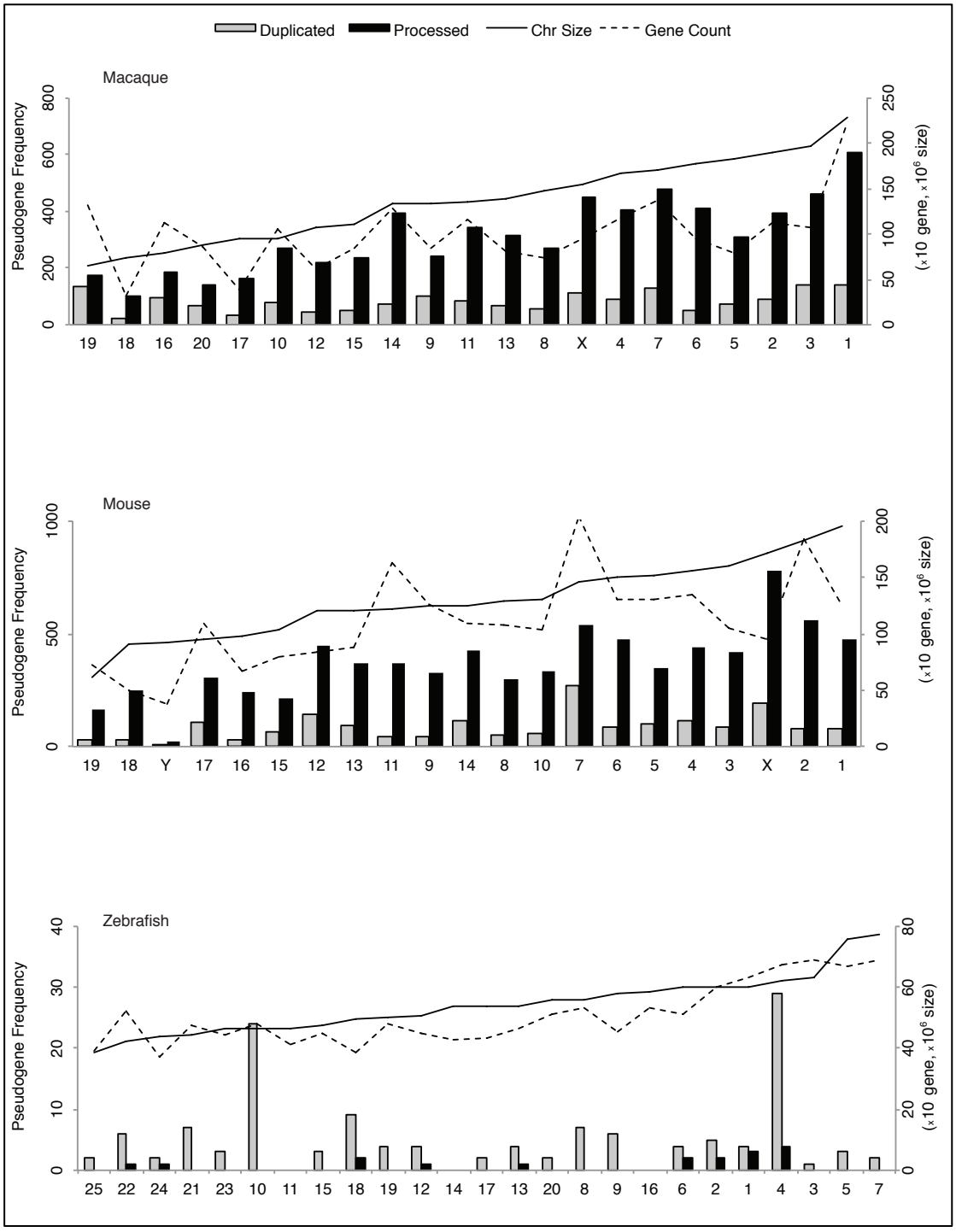
Table S1.2. Worm pseudogene localisation.

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).

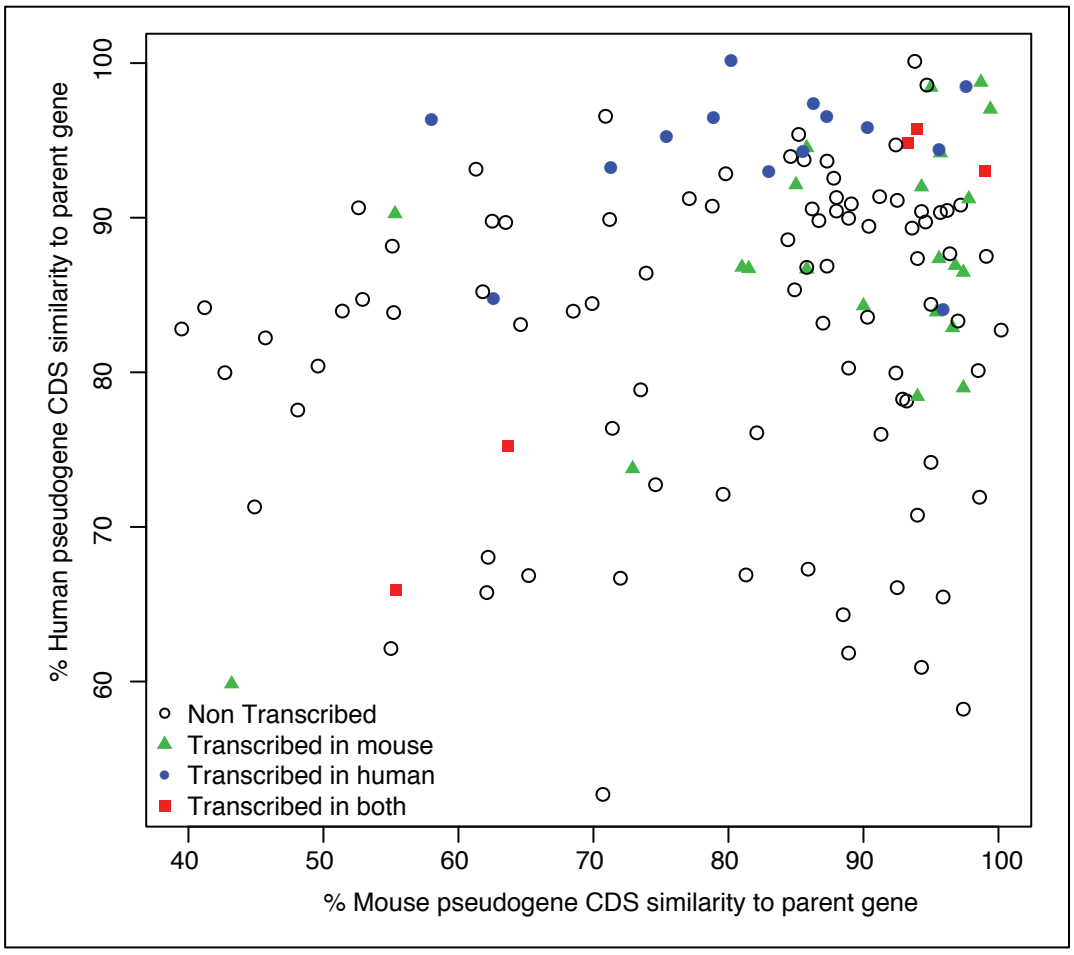
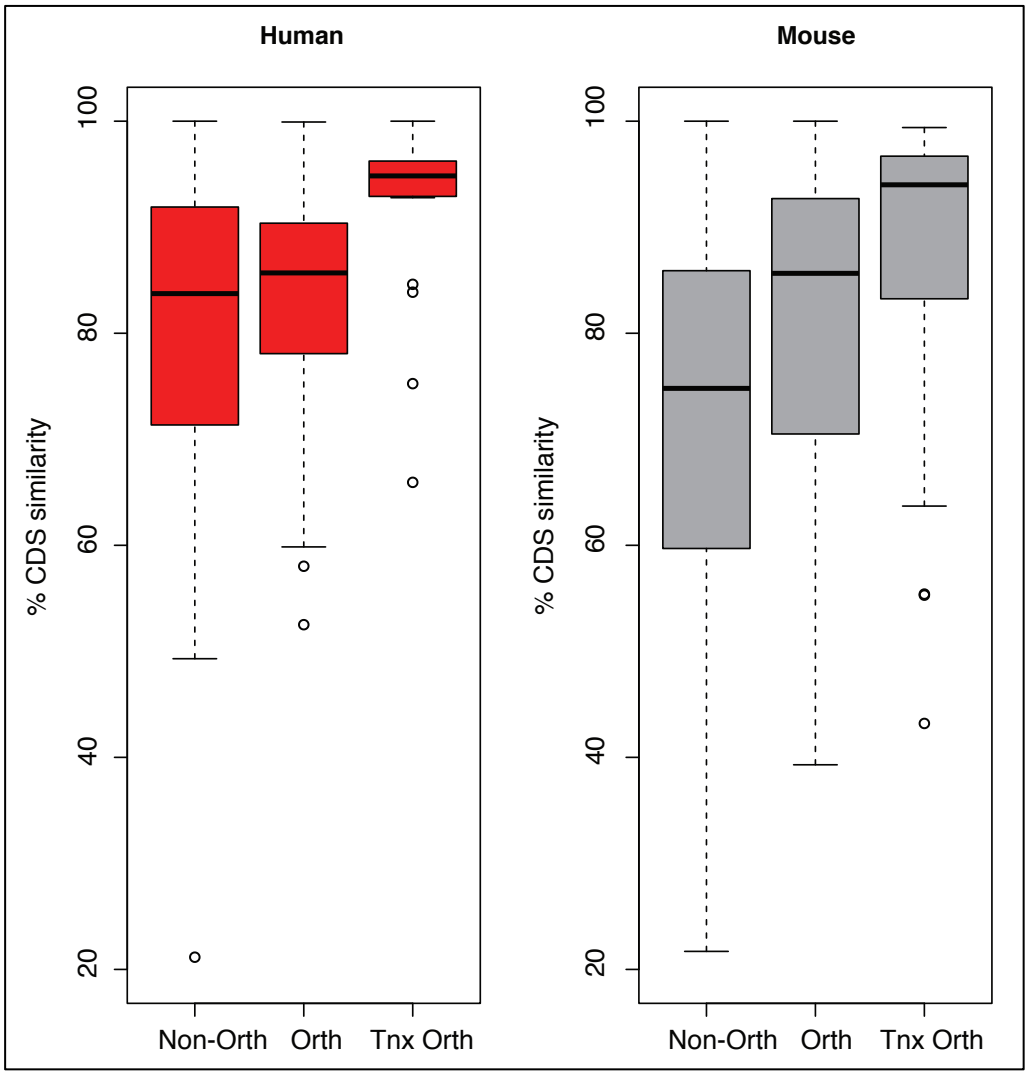


Fig. S8. Sequence conservation of human and mouse pseudogenes. Non-Orth = non orthologous human-mouse pseudogenes, Orth = orthologous human-mouse pseudogenes, Tnx Orth = transcribed orthologous human-mouse pseudogenes.



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.

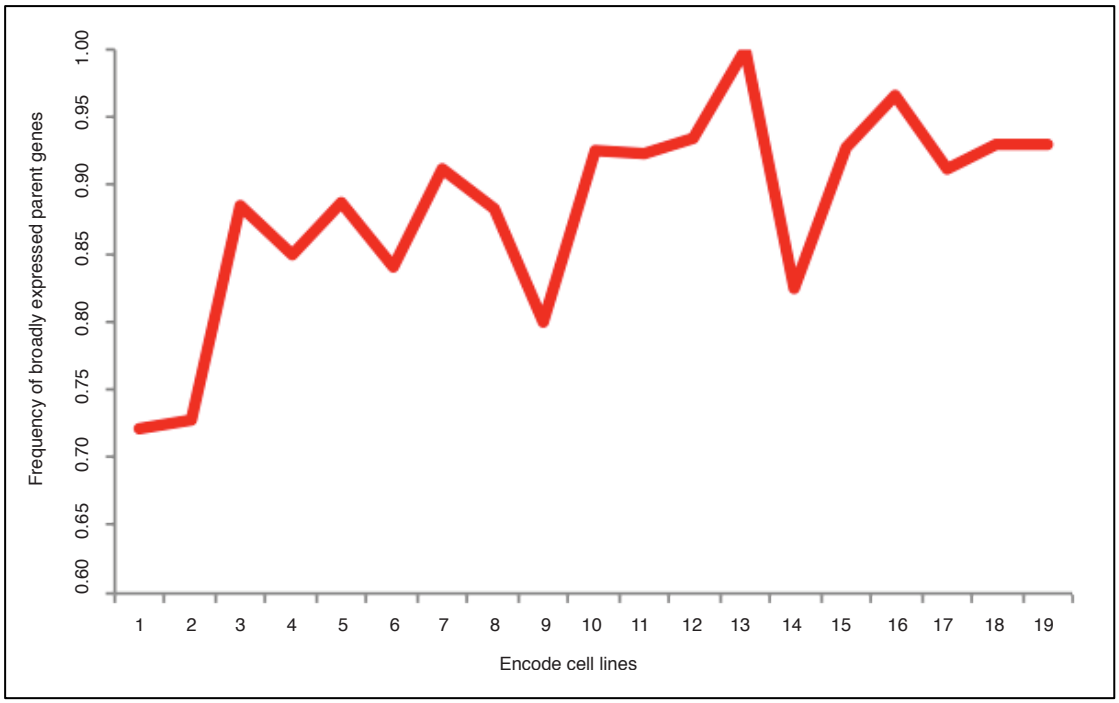


Fig. S10. Tissue specificity of transcribed pseudogenes. In human, the majority of transcribed pseudogenes are expressed in only one or a few cell lines, however a fraction of pseudogenes are universally transcribed. In worm, most pseudogenes are transcribed in only a few development stages. In fly,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

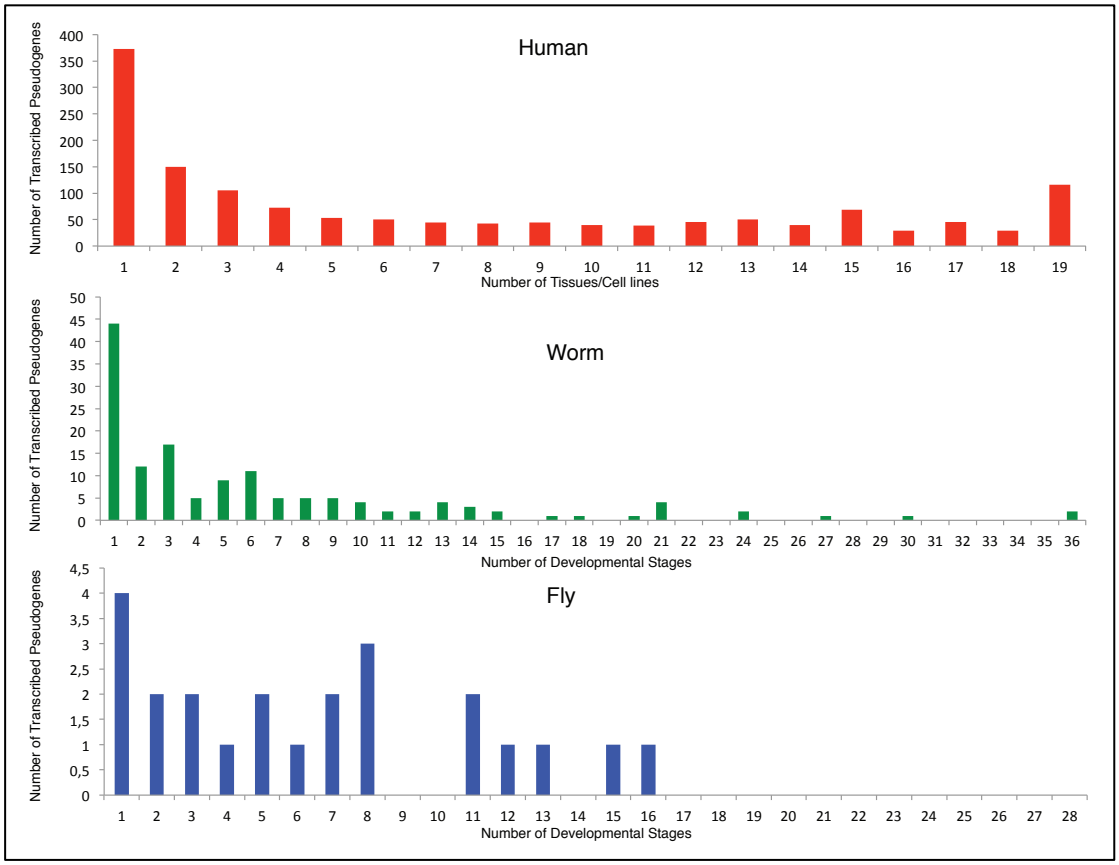


Fig. S11. Derived allele frequency for human pseudogenes. The pseudogenes are differentiated based on their activity levels: (A) transcribed vs. non-transcribed pseudogenes; and (B) highly-active, partially-active, and dead pseudogenes.

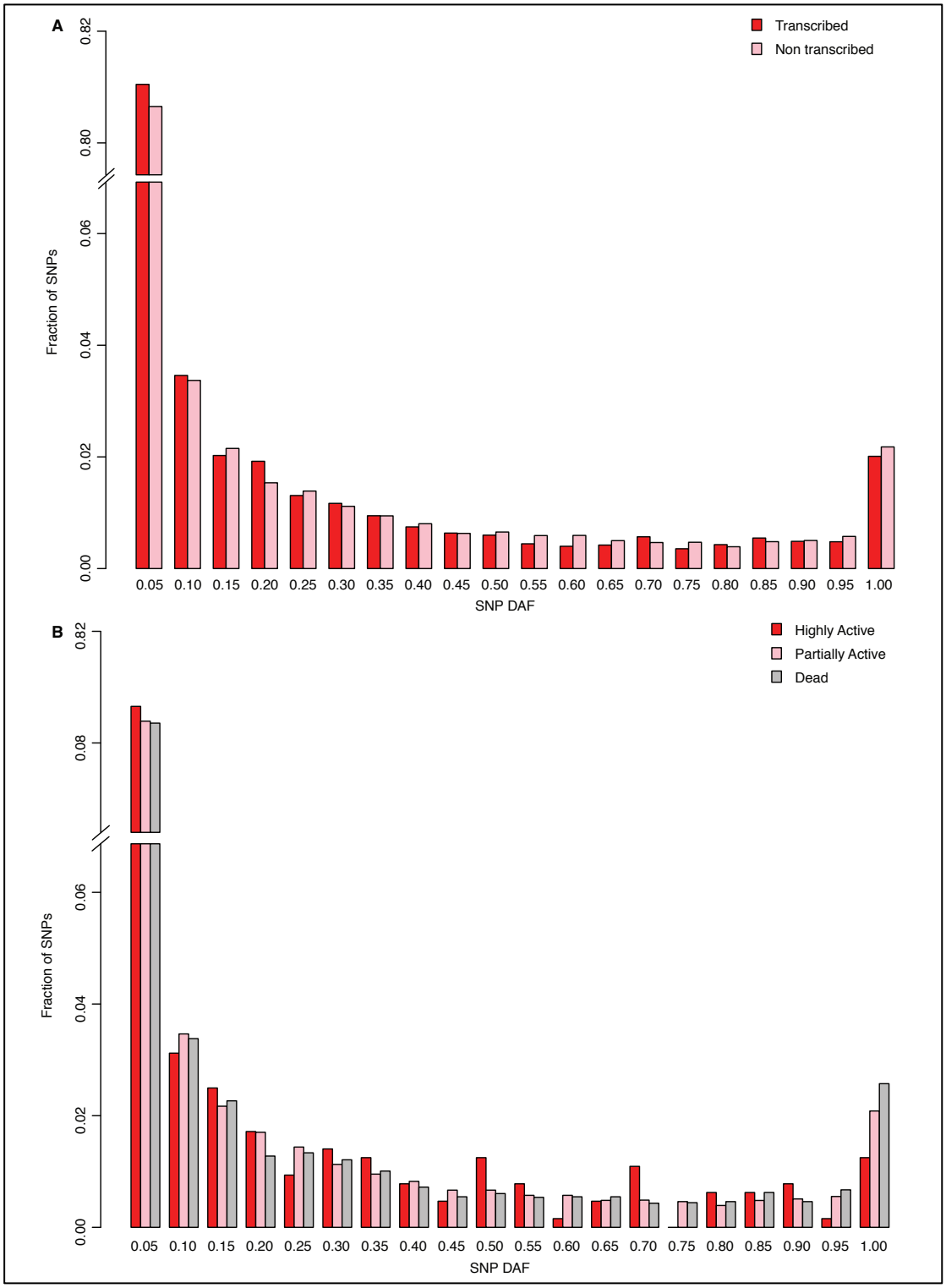


Table S3. Shadow table for Fig. 4D. Pseudogene translation candidates in human.

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

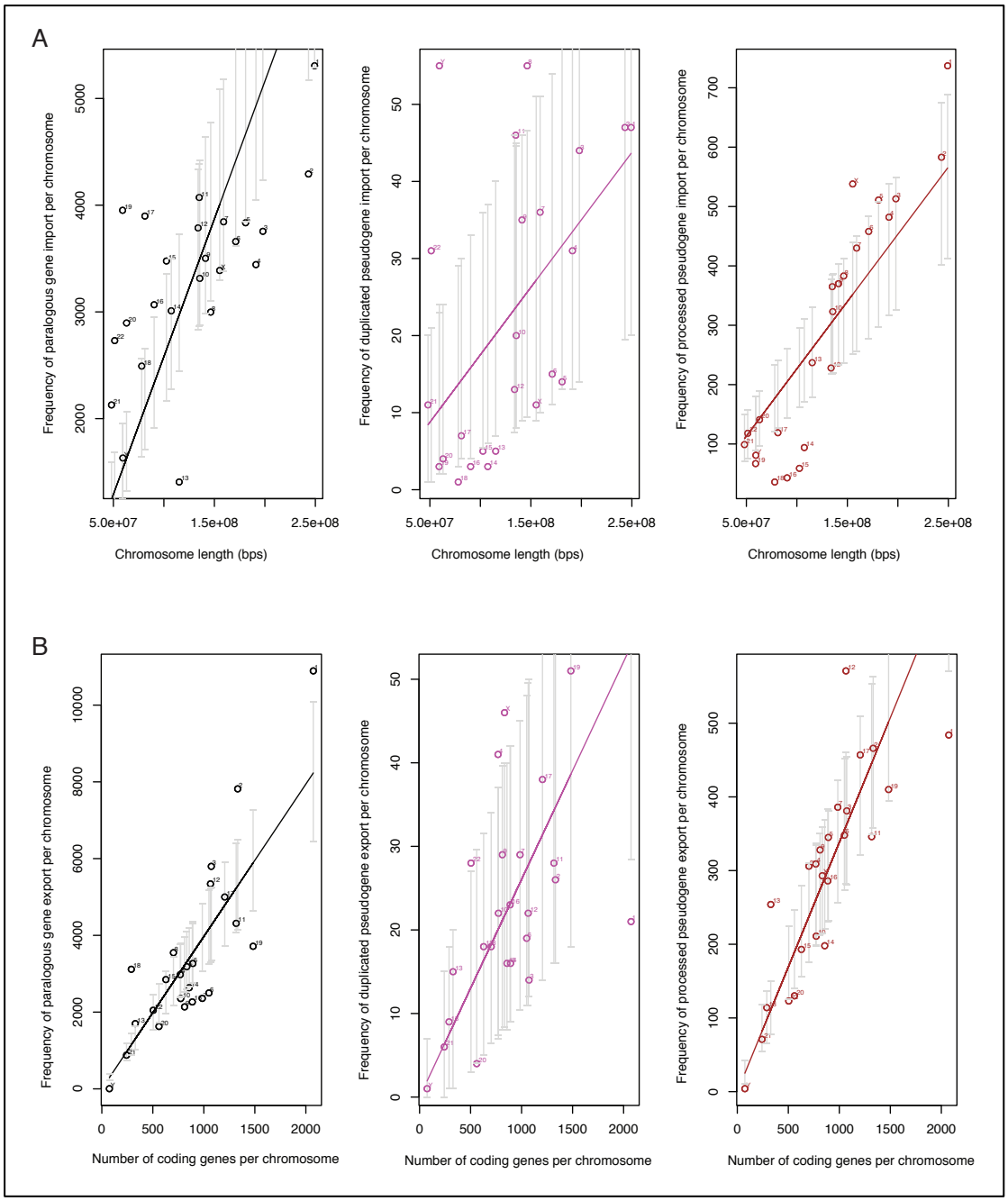


Fig. S12.2. Worm

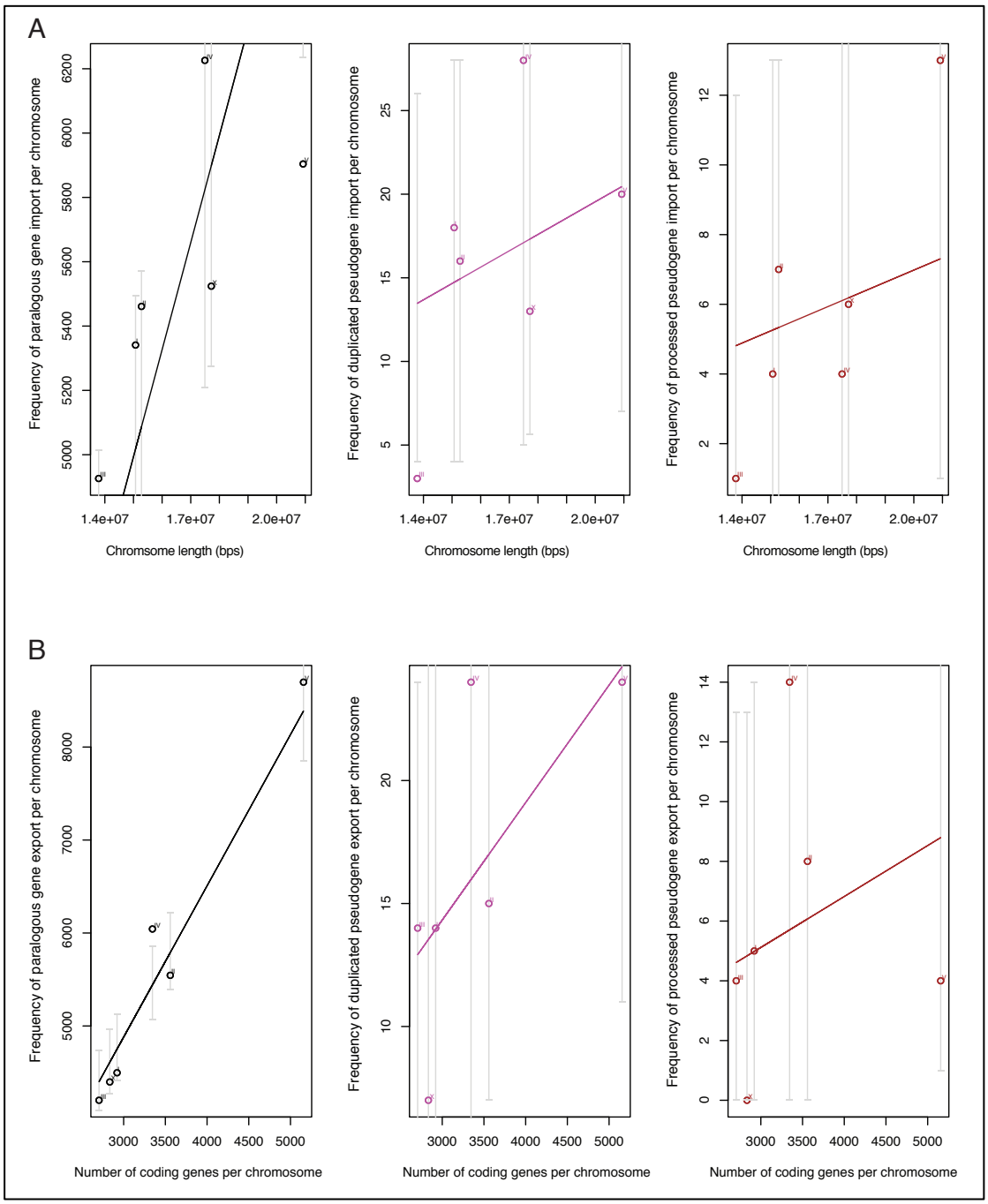
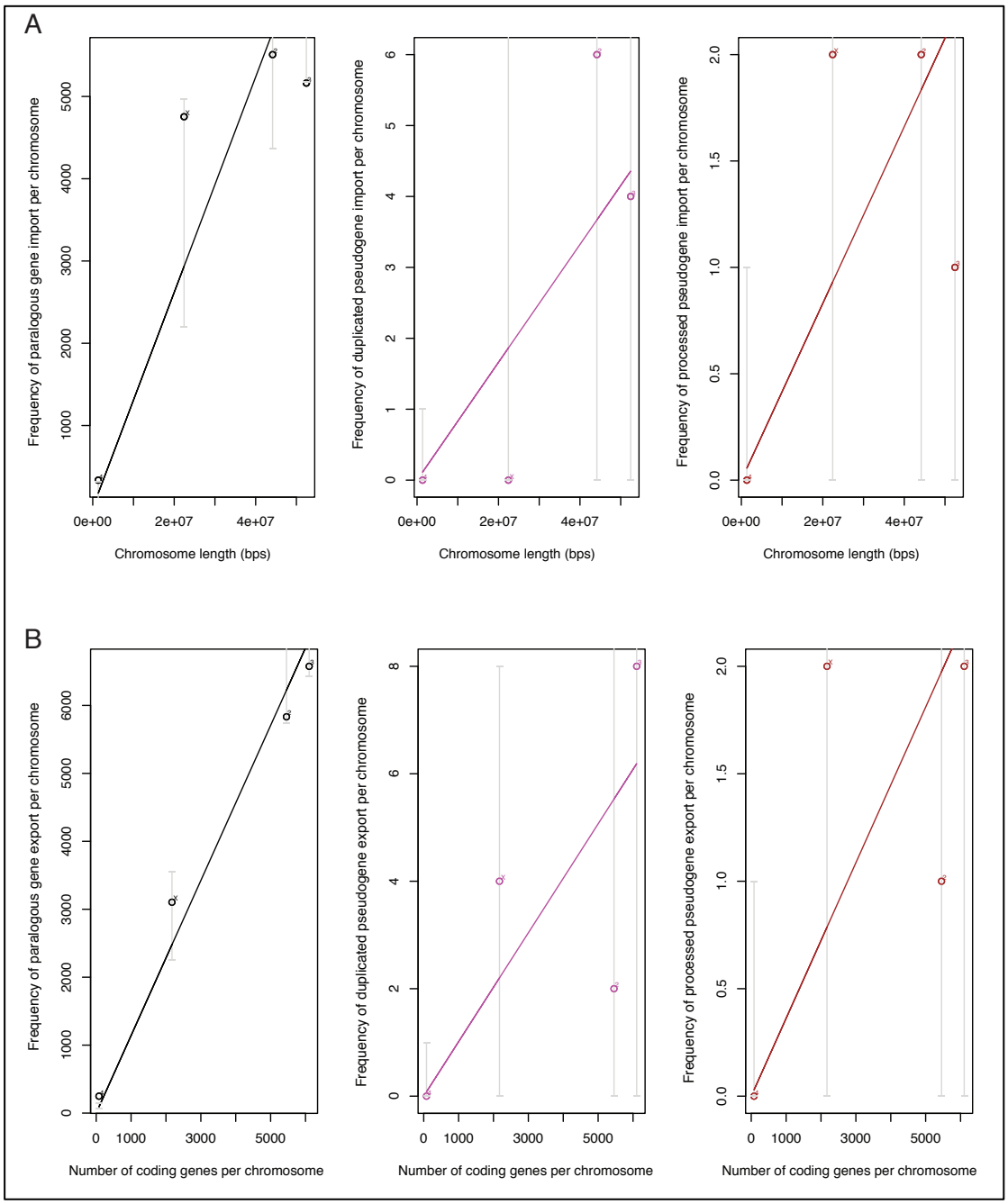


Fig. S12.3. Fly


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