Materials and Methods

Pseudogene Annotation

The pseudogene annotation was conducted using a combination of manual annotation and in silico pipelines. The annotation files are available online at www.pseudogenes.org/psi3.

(a) Manual Annotation

We manually annotated the human pseudogenes on the basis of their homology to protein data from the UniProt database. The protein data was aligned to the individual bacterial artificial chromosome (BAC) clones that make up the reference genome sequence using BLAST \cite{9254694}. We created gene models based on these alignments using the ZMAP annotation interface and the Otterlace annotation system \cite{15123593}. The alignments were navigated using the Blixem alignment viewer \cite{7922687}. We used visual inspection of the dot-plot output from the Dotter tool \cite{7922687} 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 - 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.

Fly pseudogenes were annotated in a similar way to human with two notable differences demanded by the creation method of the two pseudogene sets; human pseudogenes being identified *de novo* (though informed by the PseudPpipe set \cite{16574694}) and fly pseudogenes being annotated in the presence of existing pseudogene sets from PseudoPipe and FlyBase \cite{flybase}. Firstly, 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 sequence was used to investigate a pseudogene it was aligned to the genome using Exonerate \cite{15713233} before being assessed using Blixem and Dotter.

Worm pseudogenes were annotated following a similar mechanism: using a combination of automated (PseudoPipe) and manual annotation (WormBase \cite{wormbase}). The PseudoPipe pseudogene set was intersected with the manually annotated one. All the pseudogenes passing the threshold of 80% sequence overlap between the two data sets were selected as part of the high confidence data set. Further we manually validated the biotype annotation.

(b) Automatic Annotation

PseudoPipe is an automatic pseudogene annotation tool that uses protein homology data to identify pseudogenes. PseudoPipe uses a six-frame translational BLAST to search all the know protein sequences from Ensembl. The 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 the parental gene structure and conversely “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 as “unprocessed” (“duplicated”) and 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 as “unprocessed” (“duplicated”) while 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 tie-breaker in calling pseudogene biotype.

(b) Timeline

The differences in the dynamics of genome evolution make it difficult to directly estimate the pseudogene age. We used the sequence similarity to the parent gene as an indicator of pseudogene age. Thus young pseudogenes were defined by a high sequence similarity to the parents, while the older, more diverged pseudogenes were characterized by a smaller percent of sequence similarity to parents. Given the large differences in the number of pseudogenes in the three organisms it was difficult to bin them consistently. Thus we divided the pseudogenes based on their sequence similarity to parents in 11, 11 and 2 bins for human, worm, and fly respectively. Consequently in each human and worm bin there were on average 10% of the total number of pseudogenes. Due to the low numbers of pseudogenes in fly we chose only 2 bins each containing on average 50 pseudogenes.

(c) Repeats

JJL-XXX

(d) Disablements

Using PseudoPipe, we identified three types of pseudogene disablements: insertions, deletions, and stop codons, by comparing the pseudogene and protein-coding parent gene sequences. We calculated the average defect density per pseudogene per megabase for each organism.

(e) Selection

HUM.

Using the 1000 Genomes Project Phase 1 data we calculated the frequency of low coverage SNPs in the pseudogene exons. As a proxy of the genomic average we used the frequency of low coverage SNPs in the upstream and downstream UTR exons of the pseudogenes. Overall the pseudogenes have a similar SNPs frequency as the genomic average.

Next we calculated the derived allele frequency (DAF) for each pseudogene. Overall the 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). The length of the telomeric/centromeric region was defined as 15% of the total chromosome length. In the case of acrocentric chromosomes we defined the centromeric region around the geometrical middle of the chromosome. As such each chromosome has 2 telomeric regions (one at the 5' and one at the 3' end), 2 centromeric regions (upstream and downstream of the chromosome center) and 2 body regions spanning in total 30%, 30% and respectively 40% of the total chromosome length. 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 two regions for the entire genome. We used a binomial test to evaluate the statistical significance of the difference in the pseudogene frequency between the telomeric and the centromeric regions.

(b) Recombination

WC-XXX

(c) Co-localization tendency

We evaluated the pseudogene tendency to reside on the same chromosome as their parent gene using a 2-by-2 contingency table "A" (Fig SXXX), with elements $A_{i,j}$, where $i,j \in \{1,2\}$:

- $A_{1,1}$ - 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 the 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 the pseudogene exchange between different chromosomes, excluding the co-localizing pseudogenes-parent pairs. We used a Poisson regression model to detect chromosomes with a significant pseudogene exchange characteristic.

We hypothesized 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. The 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 a significant pseudogene exchanger.

Orthologs, Paralogs & Families

(a) Orthologs

We defined orthologus pseudogenes if they were located in syntenic regions and their respective parent genes were orthologous. We obtained the human-mouse synteny information from the USCS Genome Browser human HG19 and mouse MM9. The parent protein coding gene orthology information was downloaded from the Ensembl website. The human-worm-fly protein coding gene orthologs set was obtained combining the MIT prepared orthologous gene list \cite{mod14} with the one obtained from the Ensembl. We obtained about 28,000 orthologous gene triplets of which 1,935 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 downloaded the list of protein coding gene paralogs to the pseudogene parent genes from the Ensembl website.

(b) Family Membership

We grouped all the pseudogenes in families according to their parents’ membership to a family in the Pfam database \cite{18957444,22127870}. 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 the 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 checked the RPKM values of each pseudogene annotation as described below. Then within the list, we also identified pseudogenes with discordant expression patterns with their parent genes, using the PseudoSeq pipeline.

- *RPKM*

We identified 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 a unique mapping. (ii) We filtered out pseudogene regions with mappability lower than 1. (iii) We also discarded the pseudogene regions shorter than 100 bp after mappability filtering. (iv) We computed the RPKM value on 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 \cite{17568002,22951037}, 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 similar distributions. We applied a quantile normalization on the pooled “matched compendium” data for worm and fly, using human as a reference. This forces the protein coding genes (but not the pseudogenes) to follow as similar distribution in the three organisms. (As a control, we also performed the normalization on non-coding transcription instead of protein coding genes and obtained consistent results.) (vii) We used the protein coding gene normalization 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.

- *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 \cite{22951037}. The pipeline maps RNA-Seq reads to reference genome in conjunction with a splice junction library using Bowtie \cite{19261174} and RSEQtools \cite{21134889}. The signal tracks of the reads mapped to each pseudogene and its parent are generated across all the samples. Using this pipeline we analysed the pseudogene-parent correlated expression pattern. We found that a pseudogene may exhibit either a concordant or a discordant expression pattern with respect to its parent.

(b) Additional Activity Features

We defined 2kb upstream of the 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 \cite{modEncodeDataSite}. 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 the active chromatin in pseudogenes using chromatin segmentation for human (Segway \cite{22426492}) and fly pseudogenes (9 State-Chromatin Segmentation \cite{segmodencode}), and the 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 the model previously described \cite{22951037}. We compared the distribution of active and repressive marks in protein coding genes. On average the ratio of the frequency of the active to repressive chromatin marks for protein coding genes is 5. Based on this analysis we developed a model for labeling pseudogenes with active chromatin. If the ratio of the frequency of the active to repressive chromatin state marks was equal or larger than 3 we called 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 the 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 is larger than 2 in either of the cell line, we label the pseudogene with an active chromatin tag. The active chromatin marks are Pro (promoter), Enh (enhancer) and Tnx (transcription). The repressive marks are 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 the of the histone marks around the pseudogene body to the coding gene signal. If the signals are comparable we label the pseudogene with an active chromatin mark.

(b) Upstream Sequence Analysis

JJL-XXX

“Functional” Pseudogene Candidates

(a) Pseudogene-parent Coexpression

We calculated the Spearman correlations of gene expression levels (RKPM values in RNA-Seq) across stages or cell lines between pseudogenes and parent genes for studying their co-expression relationships. 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 a 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 \cite{UniProt}. (iii) We used Peppy to map the target peptides against raw MS spectra (available from \cite{22278370}) under the default search settings \cite{23614390}. 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 the 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 greater than 2. We used BLASTP \cite{9254694} to compare the sequence similarity between the pseudogene peptides and ones of their parent protein.