Comparative analysis of pseudogenes across three phyla

[[MG18: need name for pgene resource]] pseudoba, pseudocompara, pseudophyla, psi^3, pseudo^3, pseudocube (logo = a cube with top human and sides worm/fly)]]

Abstract

Pseudogenes are fossil copies of genes reflecting the genome history. Here, we report a ~~uniform~~ multi organisms pseudogene comparison across three different phyla leveraging on the completed annotations of the human, worm and fly, which we make available as an online resource. We introduce zebrafish, mouse, and macaque pseudogene annotations to ~~give depth to our comparison~~ and provide context for an intra phylum comparison.

We find that much more than the protein coding genes, the pseudogene complement has a strong lineage specificity reflecting the different genome remodelling processes that marked each organism's evolution.

In human we see a majority of processed pseudogenes, reflecting that the mammalian pseudogene complement is governed by a single large event, the retrotranspositional burst that occurred at the dawn of the primate lineage. This can be clearly seen in the uniform distribution of pseudogenes across the chromosomes, their accumulation in areas with low recombination levels (e.g. the sex chromosome) and their preponderance in highly expressed gene families (e.g. the ribosomal proteins).

In contrast, worm and fly pseudogene complements tell a story of numerous duplication events. In worm these duplications have been preserved through selective sweeps and consequently we see a large number of pseudogenes associated with highly duplicated gene families such as the chemoreceptors. However, in fly, the large population size and high deletion rate resulted in a depletion of the pseudogene complement.

Despite large variations between the species, we also find some notable similarities. We identify a large spectrum of biochemical activity for the pseudogenes in each organism, with the majority of them exhibiting various proportions of partial activity. In particular, we identify a consistent amount of transcription (~15%) across all species implying a uniform degradation mechanism of functional elements. Also see a uniform decay of the pseudogene promoters activity relative to ~~the one~~ of their coding counterparts and we identify a small population of pseudogenes with highly conserved upstream sequences and activity hinting at potential regulatory roles.

Cristina Sisu 27/2/14 18:08
Deleted: residual [[m25: intermediate or partial]]

Introduction

Often referred to as "genomic fossils" \cite{17568002,16574694}, pseudogenes are defined as

disabled copies of protein-coding genes. However, some can be transcribed \cite{22951037,17382428} and play important regulatory roles \cite{20577206,21816204}. Presumed to evolve with little selection constraints \cite{10833048}, pseudogenes are of great value in estimating the rate of spontaneous mutation and hence provide insight into the genome evolution \cite{2499684,9461394}.

Previously, pseudogenes have been characterized within individual genomes \cite{17099229,22951037,11160906,12560500,15860774} Earlier non-standardized annotations were characterized by large fluctuations from one release to another. As such, the absence of a finished annotation and the potential of mis-mapping of functional genomics data had restricted former comparisons of the pseudogene complement in various organisms to a specific family or class of pseudogenes \cite{15289607,16469101,12417195,19835609,12034841,12083509,19123937}. The availability of the complete genome annotation of the human, worm, and fly, allows us for the first time to embark on a uniform and comprehensive cross-species comparison of pseudogenes levering of the rich diversity of the ENCODE data.

While they all share common regulatory and transcriptional principles \cite{mod1,mod2}, these organisms could not have been more different. In order to understand the intra-phylum variations in vertebrates, we bring in the zebrafish, mouse, and macaque, taking advantage of the variety of functional genomics data available for mouse and the complete genomic annotation of the zebrafish.

The pseudogene prevalence, as well as their high sequence similarity to coding genes rose numerous and difficult problems in experiments directed at protein coding regions. The finished annotation highlighted in this study is relevant in reducing the false discovery rate and mis-annotation, and it also gives us the opportunity to correctly identify and analyse pseudogenes with potential biological activity.

Cristina Sisu 1/3/14 22:10
Deleted: not only

Results

The Pseudogene Annotation Resource

In this study, we present the completed pseudogene annotation in human, worm, and fly as part of the ENCODE project. The pseudogene annotation is a difficult and complex process. The sequence decay at pseudogene loci makes it challenging to identify authentic pseudogenes and accurately define their boundaries \cite{22951037}. To this end we used a hybrid approach, combining manual annotation with computational predictions. While providing high accuracy, the manual process is slow and may overlook highly mutated or truncated pseudogenes with weak homologies to their parents. Complementary, 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 curated a highly accurate and exhaustive pseudogene set for each organism.

Comparing the different organisms, the pseudogene distribution does not follow the relative

genome size or gene counts, e.g. 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 used macaque and mouse as a mammalian pseudogene baseline. We estimated the pseudogene content in the two organisms using the in house computational annotation pipeline (PseudoPipe). As expected, the two mammals show a similar pseudogene content to human (Fig 1A).

All the data resulting from the annotation and comparative analysis was collected into a comprehensive online pseudogene resource: www.pseudogenes.org/psi3.

psi cube

Cristina Sisu 27/2/14 18:09
Deleted: ψ^3. [[m25: psi cube]]

Classification & Evolution

(a) Classification

Based on their mechanism of formation \cite{12034841}, pseudogenes are classified into several categories: duplicated, processed (resulting from retrotransposition) and unitary. For this analysis we focused solely on the duplicated and processed pseudogenes. We found that processed pseudogenes are the dominant biotype in mammals, whereas worm, fly and zebrafish genomes are enriched in duplicated pseudogenes (Fig 1A).

(b) Timeline

Next we looked at the pseudogene evolution. We inferred the pseudogene age using its sequence similarity to the parent gene as timescale, and assessed the abundance of processed pseudogenes at different ages (Fig 1B). In human, the prominent peak of processed pseudogenes content, at high sequence similarity, corresponds to the burst of retrotransposition events. Likewise macaque and mouse show a step-wise increase in the number of processed pseudogenes at similar time points (Fig SXXX). By contrast, in worm, we see a higher proportion of older processed pseudogenes compared to younger ones. In fly and zebrafish we observed a constant and small content of processed pseudogenes.

c) Repeats

Repeat elements play an important role in the transposition events and thus in the creation of pseudogenes \cite{17424906,18291035}. To this end, we examined the repeat content of various annotated features in the genome namely CDS, UTR, lncRNA and pseudogenes (Fig SXXXREPEAT). In general, pseudogenes show a lower repeat content than UTR, lncRNA, and even the genomic average. In the case of processed pseudogenes, this result is consistent with the fact that although repeats are required for their genesis, they are not re-inserted at the pseudogene loci themselves. Similarly, the repeat content in the CDS is low, indicating a strong purifying selection pressure in these regions. By contrast the lncRNAs and UTRs showed a high repeat content and low conservation in all four species

(d) Disablements & Selection

We analysed the variety and propensity of disablements as markers of the pseudogene evolution. We observed a lower disablements density in the human pseudogene sequences,

compared to worm and fly (Fig SXXX). Based on their origins, we distinguished three types of disablements: insertions, deletions, and stop codons (Fig 1C). The average number of indels is constant across all the mammals and is twice the number of stop codons. However, the fly and worm genomes show a preference for deletions and insertions respectively.

Further we looked at the selection in human pseudogenes analysing the derived allele frequency. At the population level, we did not find any statistical significant enrichment for the human pseudogenes over the genomic average. A similar pattern was observed when separating the pseudogenes based on their biotype.

Localization & Mobility

Given the fact that the majority of pseudogenes are not under strong selection pressure, we looked at the pseudogene localization and anticipated finding them in regions of low recombination rates. To this end, we analysed the recombination rate at pseudogene loci for each species (Fig 2A). We found that the human and fly pseudogenes are enriched in regions of low recombination and thus are preferentially located near the centromere and in particular on the sex chromosomes (Fig 2A). However, in worm we observed a rather uniform recombination rate for genes and pseudogenes, a possible consequence of recent selective sweeps that pruned its genome. As such, the pseudogenes are preferentially found near the telomeres, regions characterized by high recombination rates and rapid gene evolution \cite{8536965}.

Looking at the distribution of pseudogenes, we found, as expected, a strong correspondence between the duplicated pseudogenes and protein coding genes density in worm and fly. However in human, the processed pseudogene distribution follows closely the chromosome size and it is only weakly correlated with the protein coding genes content suggesting the existence of pseudogene inter-chromosomal transfers. By contrast the duplicated pseudogenes are commonly found on the same chromosome as their parent genes. This co-residence is notable for human chromosomes 7 and 11, due to their enrichment in genome duplication events \cite{12853948} and olfactory receptors respectively \cite{11337468}.The co-localization is also significant for the sex chromosomes (human Y, fly X), where, consequence of a low recombination rate \cite{16545149,1875027,15059993}, the pseudogenes cannot be “crossed out”. As a result in human, we observed a large accumulation of imported processed pseudogenes on X \cite{14739461} and an enrichment of duplicated pseudogenes on Y with apparent parent genes on the X chromosome.

Orthologs, Paralogs & Families

We compared the lineage specificity of pseudogenes by analysing their families and orthologs.

(a) Orthologs

While numerous protein-coding genes are conserved even for distant relatives, there are no pseudogene orthologs across all species (Fig 3A). However, we were able to identify orthologous pairs for closer relatives such as human and mouse. We found that only 129 (~1%) of the human pseudogenes have mouse orthologs, setting thus a base line for pseudogene orthology between human and other species. The majority of the orthologous pseudogenes

(127) are processed and have a high sequence similarity to their parents (Fig SXXX).

Next, we analysed ~2000 1-1-1 human-worm-fly orthologous protein-coding genes and observed that not one of the triplets has associated pseudogenes in all three organisms (Fig 3A). Also the number of pseudogenes associated with 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.

(b) Paralogs & Families

We compared the distribution pattern of pseudogenes per parent gene (Fig 3C). In human, despite the fact that the pseudogenes are almost as numerous as the protein coding genes \cite{22951037}, only 25% of the genes have a pseudogene counterpart. Consequently the distribution of pseudogenes per gene is highly uneven. As a control we introduce 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 found a number of genes that are enriched in pseudogenes and depleted in paralogs, and vice-versa, a trend common across all organisms.

Family analysis allowed for a bigger pattern to emerge (Fig 3D). As expected, the ribosomal proteins are the dominant families in human. These abundantly expressed genes are indicative of the general burst of retrotransposition events \cite{16504170}. Analysis of mouse and macaque top families shows that this pattern is common across mammalian genomes. However, top families relative rank is organism specific. The top pseudogene families in worm are the 7 Transmembrane (7TM) proteins, perhaps reflecting the family rapid evolution \cite{11961106} and the many duplications events in nematode genome history \cite{19289596,18837995}. Interestingly, even though dominated by processed pseudogenes, the human genome shares 7TM as its top family, as evidence of the duplication and divergence of the olfactory receptors. In fly, SAP and MOTOR families are dominant.

Finally, despite the lineage specificity of the pseudogene top families, we found a number of large duplicated families common to all organisms namely – kinases, histone and P-loop NTPase, reflecting perhaps the essential role these genes play in the species evolution.

Activity

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

(a) Transcription

Analysing RNA-Seq data we found 1,441, 143, 23 potentially transcribed pseudogenes in human, worm, and fly respectively. We also identified 31 transcribed pseudogenes in zebrafish and 878 in mouse. This represents a fairly uniform fraction (~15%) of the total pseudogene complement in each organism. Within transcribed pseudogenes, ~13% in human and ~30% in worm, and fly, have a discordant transcription pattern with their parent genes over multiple samples (Fig SXXX). Also, a large fraction of pseudogenes are associated with a few highly

expressed gene families, for example, the ribosomal proteins in human.

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

(b) Activity features

Next we examined a number of additional markers of biochemical activity, including the presence of active transcription factors and RNA Polymerase 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 identified a group of "dead" pseudogenes – with no indicators of activity. Contrary to the actual definition of pseudogenes ("dead genomic elements"), this group comprised 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 labelled these pseudogenes as "highly active". Also, in humans, we found that the transcribed 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. However, the majority of pseudogenes (~75%) are intermediate between these two, having only a few of the classic indicators of activity. We labelled these pseudogenes as "partially active". The distribution of pseudogenes for the three activity levels is consistent across all studied species.

(c) Upstream sequence similarity and promoter activity

The pseudogene activity is connected to the regulatory upstream region. To this end we examined the divergence of pseudogene promotors in the proximal (within 2kb of the 5' end) upstream region. As a control we used the parent gene paralogs promotor regions.

Contrary to expectations, a small fraction of duplicated pseudogenes exhibited highly conserved upstream and "coding" regions, even more than paralogs do when compared to the parent genes (Fig 4B). These pseudogenes may be recent duplicated loci that have diverged little from their parents. Interestingly, we found a number of duplicated pseudogene-parent pairs with high upstream similarity despite low "coding" sequence identity, suggesting that the upstream regions may have been 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 analysed the ChIP-seq data of H3K27ac, an important marker in defining active promoters and enhancers. We focused our study on protein coding genes with only one pseudogene but no paralogs, and those with one pseudogene and one paralog. We observed that in general, while the pseudogenes have highly conserved promoter regions, the activity is less preserved when compared to their protein coding gene counterparts (Fig 4C).

"Functional" Pseudogene Candidates

Finally we refined the active pseudogene list and combining the annotation, functional genomics

and evolutionary data.Focusing on the regulatory potential, we identified a set of “functional” candidates (active and with a significant parent/pseudogene coexpression correlation coefficient) including the known regulatory cancer pseudogene PTEN-P1.

Next, using mass spectrometry data, we studied the translation potential of transcribed human pseudogenes in four ENCODE cell lines. From over 14000 pseudogenes we identified three pseudogenes with high translation evidence (Fig 4D). The low number candidate translated pseudogenes is indicative of the high quality of our annotation. Interestingly, one of the candidates (ENST00000533551) showed 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 uniform multi organisms’ pseudogene comparison leveraging on the finished annotations of the human, worm, and fly. Unlike the protein coding genes, which are essential to the correct development and function of the organism and thus are under strong negative selection, the majority of pseudogenes evolve neutrally, making them an ideal proxy for the study of genome evolution.

Cristina Sisu 27/2/14 18:09
Deleted: ,

Cristina Sisu 27/2/14 18:09
Deleted: , [[m25: keep only human, worm and fly]] and zebrafish genomes and the draft mouse and macaque genome

Overall our results show that the pseudogene complement, even more than its coding counterpart, is strongly lineage specific reflecting the different genome remodelling processes that marked the organisms' evolution. There are essentially no orthologous pseudogenes between the distant organisms and we only see an overlap at the protein family level, where are few large, highly duplicated families (e.g. kinases) tend to give rise to numerous pseudogenes in all the studied species.

We find that the mammalian pseudogene complement is marked by a single large event, the retrotranspositional burst that occurred approximately 40 million years ago, at the dawn of the primate lineage. This can be clearly seen in the uniform distribution of pseudogenes across the chromosomes and their slight accumulation increase in areas with low recombination levels, e.g. the sex chromosomes and the centromere regions. It also resulted in a preponderance of pseudogenes associated with highly transcribed proteins such as those in pathways of central metabolism and the ribosomal proteins. Also, while the burst of retrotransposition events happened after the human/mouse speciation (~90 MYa), the high occurrence of processed pseudogenes in the mouse genome suggests that this event occurred on a much larger scale and it can be regarded as an general mammalian characteristic. In contrast, worm and fly pseudogene complements tell a story of numerous duplication events. This became 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 \cite{22286215}, 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 \cite{12572619,9501496,14631042} combined with a strong selection in the intergenic sequence \cite{12572619,1806330,9402741} and a high deletion rate resulted in a depletion of the pseudogene complement and consequently we see a segregation of the

remaining pseudogenes to areas of low recombination.

The apparent duplicated pseudogene exchange between the X and Y, chromosomes is potentially a consequence of the numerous gene loss events in Y's evolutionary history \cite{16847345}. As such the majority of "X exported" duplicated pseudogene on Y are "degenerated paralogs", products of gene duplications, that subsequently accumulated deleterious mutations \cite{15233989}.

Finally we identify a large spectrum of biochemical activity (as defined by transcription, active chromatin, Pol II and transcription factors) for the 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 degradation mechanism.

We relate the activity of pseudogenes to the conservation of their upstream region. Comparing the pseudogenes and functioning paralogs, we find that many pseudogenes have more conserved upstream sequences than paralogs do. Even more, we identify a number of pseudogenes with highly conserved upstream regions relative to their parent gene. However, this conservation is not always preserved in the terms of upstream activity (as defined by histone marks). In this case the 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 ranking the pseudogenes based on their activity features and pinpoint potentially functional candidates. The regulatory roles of several pseudogenes through their RNA products have been previously demonstrated \cite{21816204,18405356,20577206,18404147}. Hence we suggest that pseudogenes may play active roles in the genome biology and warrant further experimental validation.

Figure Captions

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

Figure 2: Localization and mobility. (A) (left) The relative chromosomal localization preference for pseudogenes in human, worm, and fly. (right) Average recombination rate for pseudogenes, protein coding genes and genomic background. (B) Distribution of pseudogene per chromosome as function of biotype. The chromosomes are sorted by length. (C) Sex chromosome pseudogene and parent gene paralog exchange in human.

Figure 3: Orthologs, paralogs and family. (A) Venn diagrams showing the total number of orthologous genes and pseudogenes in human, worm, and fly. (far right) Intra phylum pseudogene orthologs for 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 totaling 25% of the total number of pseudogenes in each organism. Family type legend: GAPDH – Glyceraldehyde 3-phosphate dehydrogenase, 7tm – GPCR, His – Histone, IG – Imminoglobulin, Kin – Kinase, Ploop – P-loop NTPase proteins, Ribo – Ribosomal proteins, RRM – RNA-recognition motifs, Struct – Structural protein, ZnF– Zinc finger proteins (TF), Ubiq – Ubiquitination proteins, Motor – Motor proteins, SAP – SAP domain proteins.

Figure 4: Pseudogene activity. (A) Distribution of pseudogenes as function of various activity features: transcription (Tnx), active chromatin (AC) and presence of active Pol II and TF binding sites in the upstream region. (B) Conservation of the upstream sequences in processed and duplicated pseudogenes as compared to paralogs. (C) Conservation of the upstream sequence activity marks (H3K27Ac) in pseudogene-parent pairs versus parent-paralogs. (D) Functional pseudogene candidates.