Comparative analysis of pseudogenes

Abstract

[[MG(30dec)2CSDS: add something on specific fams in abs. Also, "is general involvement" right?]]

In this study, we present a comprehensive pseudogene resource highlighting the completed annotation of pseudogenes in human and three model organisms, worm, fly and zebrafish. We obtained a detailed map of the pseudogene complement for each organism integrating annotation with functional genomics and evolutionary data. Comparing the four organisms, we found that, overall, the pseudogene complement differs much more between species than protein-coding genes, reflecting more closely the genome evolution history. The pseudogene families are specific.[[add note about location, genesis, evolution]] Also we identified a large spectrum of activity for the pseudogenes in each organism. However, the distribution of their activity levels is consistent across all the studied organisms, implying the general involvement of pseudogenes in organism biology. Finally we ranked the pseudogenes based on their activity features and identified a number of potentially functional candidates.

[[MG(30dec)2CSDS: there has been no direct and rigorous comparison between pseudogenes of various model organisms : Perhaps too strong]]

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}. Previously, pseudogenes have been characterized within individual genomes \cite{17099229,22951037,11160906,12560500,15860774,12083509,16925835}. At a first glance, these individual results hint at large differences in the pseudogene distribution and functionality between organisms. However, to date, there has been no direct and rigorous comparison between pseudogenes of various model organisms. In this paper we describe the first study focused on analysing and contrasting the pseudogene complement in human, worm, fly, and zebrafish.
While the number of the protein coding genes has been stable for a long time, the number of pseudogenes showed large fluctuations from one annotation release to another (Fig XXX). The sequence decay at pseudogene loci makes it challenging to rightly identify authentic pseudogenes and accurately define their boundaries \cite{22951037}. In human, the

pseudogenes are almost as numerous as the protein coding genes \cite{22951037}. Their prevalence, as well as the similarity to their parents can cause problems in experiments directed at protein coding genes. However the pseudogene annotation is a difficult and complex process. To this end we used a hybrid approach, combining manual annotation with computational pipelines. While providing high accuracy, the manual annotation is slow and can overlook highly mutated or truncated pseudogenes with weak homologies to their parents. Complementary, the automatic pipeline (PseudoPipe) is fast and provides an unbiased annotation of pseudogenes. However computational methods are prone to error due to mis-annotation of parent gene loci. Thus the finished annotation provides a reliable set of pseudogenes and consequently allows for a statistical significant quantitative comparison. It is even more important in identifying and analysing pseudogenes with potential biological activity since it reduces the false discovery rate and the potential of mis-annotation.
The finished annotation of the genomes of the four species leads the way to a rigorous multi-way comparison of the pseudogene complement. Also the extensive functional genomics, proteomics and evolutionary data available allows us to conduct an integrative study to uncover the differences in the pseudogenization process as well as highlighting the evolution and biological activity of pseudogenes in the four organisms.
Our analysis shows that the pseudogene repertoire is lineage specific and has important implications for the genome evolution. We also highlight a similar fraction of residual, pseudogenic biochemical activity in all of the four organisms.

Results

We analysed the pseudogenes complement of four fully annotated organisms.

[[MG(30dec)2CSDS: Similarly, the larger fraction of duplicated pseudogenes in zebrafish can be accounted for by the prevalence of intra-chromosomal and tandem duplication events in its genome \cite{22702965}. --- really?]]

Annotation

Overall, the pseudogenes differ greatly between organisms, reflecting the unique evolutionary history of each of them. The pseudogene distribution does not follow the relative genome size or gene counts, e.g. the human genome has about 15-fold more pseudogenes than worm, 100-fold more than fly, and 50-fold more than zebrafish. Given the large evolutionary distance between the model organisms and human, in order to better understand the implications of our results for study of the human genome, we included in the analysis two mammalian species: macaque and mouse. We estimated the pseudogene content in the two organisms using the in house computational pipeline (PseudoPipe) accuracy against the manual annotation in human. As expected, the two mammals show a similar pseudogene content to humans (Table XXX).
Based on their mechanism of formation \cite{12034841}, pseudogenes are classified into several categories: duplicated, processed (resulting from retrotransposition), unitary and

polymorphic. For this analysis we focused solely on the duplicated and processed pseudogenes. We found that processed pseudogenes are the dominant biotype in human and other mammals, whereas worm, fly and zebrafish genomes are enriched in duplicated pseudogenes (Fig SXXX). The preference for processed pseudogenes in mammals, can be traced back 40 MYa to a burst of retrotransposition events. While this episode happened after the human/mouse speciation (~90 MYa), the high occurrence of processed pseudogenes in the mouse genome suggests that this event occured on a much larger scale and it can be regarded as an intrinsic characteristic of mammals rather than being primate specific. The enrichment of duplicated pseudogenes in worm and fly can be related to the relatively high gene duplication rate in these two genomes \cite{11861885,11230161,21295484}. Moreover, previous studies \cite{12572619,1806330,9402741} suggest that the scarcity of fly pseudogenes can be explained by the high rate of DNA loss inherent to the fly genome. This is an intrinsic characteristic of the large effective population size in fly \cite{12572619,9501496,14631042}. Similarly, the larger fraction of duplicated pseudogenes in zebrafish can be accounted for by the prevalence of intra-chromosomal and tandem duplication events in its genome \cite{22702965}.
Analysing the genome annotation we find a significant difference in the pseudogene complement of the four organisms.

<<================>>

Large Scale Analysis of Pseudogene Annotation

Next took a closer look at the distribution of pseudogenes in the four genomes.

(a) Chromosomal Distribution

First, we calculated the pseudogene frequency in each chromosome (Fig XXX). In human, we observed that in contrast to protein coding genes, the pseudogene distribution follows the chromosome size. However it has a weaker correlation to the protein coding gene count (Fig SXXX) suggesting the existence of pseudogene inter-chromosomal transfers. To this end we analysed next the relative position of the pseudogenes within a chromosome and their inter-chromosomal mobility.

(b) Localization

Given the large number of pseudogenes in the mammalian genomes, we observed a uniform distribution of pseudogenes across the chromosome length (Fig XXX), with a ratio between the pseudogene density in the telomeric and centromeric regions in human, of ~0.85. However, worm, fly, and zebrafish show a skewed distribution of pseudogenes. In worm, the majority of pseudogenes are near the telomeres, a location characterized by a high number of recombination events and rapid gene evolution \cite{8536965}. In contrast, fly pseudogenes are preferentially located near the chromosome centre, consistent with a high deletion rate in the telomeric regions due to the large effective population size.
To further our understanding, we looked at the tendency of pseudogenes to reside on the same

chromosome as their parent genes. As expected, we found that duplicated pseudogenes tend to be located on the same chromosome as their parent genes, whereas the processed pseudogenes are randomly scattered across the genome (FigYZ 1, FigYZ S1-3). The colocalization event is especially significant on the sex chromosomes, in particular for human Y, and fly X chromosome. This result is indicative of the sex chromosomes evolution and differentiation \cite{23436913}. Due to low recombination rate of the sex chromosomes \cite{16545149,1875027,15059993}, the duplicated pseudogenes cannot be “crossed out”
A more detailed analysis of the human autosomes, shows that for chromosomes 7, 11, the colocalization of duplicated pseudogenes and parents genes is also statistically significant (FigYZ 1 (B)). This results relates to the fact that chromosome 11 is enriched in olfactory receptors\cite{11337468} while chromsomes 7 is enriched for genome duplication events \cite{XXX}.

(c) Mobility

Next we looked at the pseudogene exchange between chromosomes, focusing on the sex chromosomes (FigYZ2, FigYZ S4-7). Consistent with previous reports \cite{14739461}, we observed that in human, X is an importer of processed pseudogenes. By contrast, the worm and fly genomes show a uniform pseudogene exchange between the chromosomes. Given the similarity in the genesis of duplicated pseudogenes and paralogous genes, we compared their import on the Y chromosome. While the majority of Y’s duplicated pseudogenes are imported from X (FigYZ 2 and FigYZ S4-6 regression-plots excluding self-contribution), we found only a small number of imported paralogs. This discrepancy can be explained regarding the duplicated pseudogenes as paralogs, products of gene duplications, that subsequently accumulated deleterious mutations \cite{15233989} due to the numerous gene loss events in Y’s evolutionary history \cite{16847345}. Furthermore, the pseudogene exchange between the sex chromosomes in all four organisms is significantly larger than the exchange with autosomes. These observations support the close evolutionary relationship between the X and Y, chromosomes

Pseudogene Orthologs, Paralogs & Family

Following the annotation analysis, we contrasted the lineage specificity of pseudogenes in the four organisms by studying their families and orthologs.

(a) Orthologs

In this study we focus on three sets of orthologs: human-worm-fly, human-zebrafish, and human-mouse. Overall we found that pseudogenes arise from different progenitors in the four organisms (Table XXX).
First, we analysed the human-worm-fly orthologous pseudogenes. We tested ~2000 1-1-1 orthologous protein-coding genes for pseudogene parenthood (Table XXX). We observed that there is no similarity in the pseudogene complement of orthologous genes. In fact, not one of the triplets of 1-1-1 orthologous genes has associated pseudogenes in all three species (Fig SXXX). As an example (Fig XXX) the number of *RpS6* pseudogenes varies significantly among the

analysed genomes, with human having 25 (mostly processed) pseudogenes spread randomly across the genome, fly having three duplicated pseudogenes clustered near the parent gene and worm having no *RpS6* pseudogenes at all.
Next, we looked at closer relatives in order to get a better understanding of human pseudogene evolution and specificity. To this end, we studied the human-mouse pseudogene orthologs. We observed that 1% of the human pseudogenes have mouse orthologs. Surprisingly the majority of the orthologous pseudogenes are processed and have a high sequence similarity to their parents (Fig SXXX). Also 20% of the mouse and 15% of the human orthologous pseudogenes are transcribed.

(b) Paralogs & Family

Next we compared the distributions of pseudogenes and paralogs per parent gene (Fig XXX). The distribution of pseudogenes per gene is highly uneven. Only 25% of the human genes have a have a pseudogene counterpart, and a large fraction of pseudogenes are associated with a few highly expressed gene families (e.g. *Ribosomal Protein* family in human, *Transmembrane Protein* (7TM) family in worm, and *Motor* family in fly). At the extreme we found that the top pseudogene parent genes are not simultaneously large paralog generating genes. This trend is common across all four organisms.
Pfam analysis showed that, as expected, the ribosomal proteins are the dominant families across human, macaque and mouse (Fig XXX). These abundantly expressed genes, are likely targets of retrotransposition \cite{16504170}. Even more, the distribution of ribosomal pseudogenes reflects the general burst of retrotransposition events. However, while overall the top families are shared among mammals their relative rank is organisms specific. The top pseudogene families in worm are the 7TMs perhaps reflecting the many duplications of this family in nematode evolution \cite{19289596,18837995} and the fact that this family is rapidly evolving \cite{11961106}. It is interesting to note that the human genome shares as well this top family, probably reflecting the duplication and divergence of the olfactory receptors. In fly, SAP and MOTOR families are dominant. Zinc finger is the major family type in zebrafish.
Finally, despite the organisms family specificity, we found a number of families common to all the studied organisms namely – kinases, histone and P-loop NTPase, reflecting perhaps the essential role these genes play in the species evolution.

Evolution

The next step in our analysis was to compare their evolution by looking at their age, genesis, development, and selection in the context of their species.

(a) Timeline

We inferred age of pseudogenes using the sequence similarity to parent gene as timescale. The results become most informative when combined with the fraction of processed pseudogenes at different ages (Fig SXXX). In human, a prominent peak of processed pseudogenes fraction, at high sequence similarity, corresponds to a burst of retrotransposition, at the dawn of the primate lineage when the bulk of human pseudogenes were created. Likewise macaque and mouse show a step-wise increase in the fraction of processed pseudogenes supporting the hypothesis

that this event might be a rather general mammalian trait. By contrast, in worm, older pseudogenes are preferentially processed, whereas younger ones are more likely to be duplicated. Zebrafish showed a similar distribution of pseudogenes to worm. The preponderance of duplicated pseudogenes in both worm and zebrafish relative to human might relate to numerous duplication events in the recent evolutionary history of the two organisms \cite{19622155,19289596,11230161,11861885}. Fly pseudogenes shows a constant, if rather low, ratio of processed to duplicated pseudogenes. This once again is in accord with the high deletion rate in the fly genome.

(b) Genesis

Secondly, we looked a the complex process of pseudogene genesis. Repeat elements play an important role in the retrotransposition events and thus in the creation of pseudogenes. For instance, in the human genome, the L1 retro-element, facilitates the retrotransposition and consequently the generation of processed pseudogenes \cite{XXX}. Furthermore, it has been previously shown \cite{XXX} that repeats are, potentially, a source of species specificity.
To this end, we examined the repeat content of various annotated features in the genome namely CDS, UTR, lncRNA and pseudogenes (Fig XXXREPEAT). In general, the pseudogenes show a low repeat content. In the case of processed pseudogenes, this result is consistent with the fact that although repeats are required for the pseudogene genesis, they are not re-inserted at the pseudogene loci themselves. However for the lncRNAs we observed a high repeat content and low conservation rate. Similar to the pseudogene result, we observed that the repeat content in the CDS is significantly lower than that in UTR, lncRNA, and the genomic average. The strong purifying selection pressure of these regions can explain this observation.

(c) Disablements

Next we looked at the variety and propensity of disablements as time stamps of the pseudogene evolution. Given the fact that the majority of human pseudogenes are of recent descent, as expected we observed a lower disablements density in the pseudogene sequences, compared to worm and fly (Fig SXXX). Based on their origins, we distinguished three types of disablements: insertions, deletions, and stop codons (Table XXX). The average number of insertions and deletions is constant across all the mammals and is twice the number of stop codons. However, the fly and worm genomes show a preferences for deletions and insertions respectively. In comparing worm and fly, association of the pseudogenes with indels reflects once again the organism evolutionary differences. The depletion of pseudogenes in the fly genome is reflective of its large effective population size \cite{14631042} and its prevalence for deletions. By contrast, the indels abundance in worm is primarily the by product of a largely asexual mode of reproduction. The worm has a small effective population size and its genome is prone to the accumulation of mutations/insertions \cite{17637734}. Consequently we found an enrichment of insertions as pseudogene disablements.
[[talk: extra DNA is not favourable, see less of a deletion process]]
Integrating functional data with the genome annotation of pseudogenes gave us the chance to pin-point interesting elements that might have crucial roles in the organism evolution.

Activity

Next we directed our investigation towards identifying potentially active elements by looking for signs of biochemical activity and studying their diversity in human, worm, fly, and zebrafish.

(a) Transcription

We computed an expression value based on RNA-Seq data and found 1,441, 143, 23, and XXX potentially transcribed pseudogenes in human, worm, fly and zebrafish respectively (Fig XXX). This represents a fairly uniform fraction (~15%) of the total pseudogene complement in each organism. Interestingly, we found a subset of these (~13% in human and ~30% in worm and fly) that have discordant transcription patterns with their parent genes over multiple samples (Fig SXXX). The results also indicate pseudogenes are less broadly transcribed than protein coding genes. Specifically, only 5.1%, 0.69%, and 4.6% are broadly expressed in human, worm, and fly, respectively (Table SXXX). Moreover, a substantial number of pseudogenes are expressed in only a single cell line or developmental stage (Fig SXXX). Finally, we found that the parent genes of broadly expressed pseudogenes tend to be broadly expressed as well (Fig SXXX), however the reciprocal statement is not valid.

(b) Activity features

Next we analysed for each pseudogene a number of additional markers of biochemical activity, including the presence of active TF and RNA Polymerase II (Pol II) binding sites in their upstream regions and proximal regions of "open chromatin" (as determined from histone modification data). We thus integrated the transcriptional information with this additional data to create a comprehensive map of pseudogene activity (Fig XXX), grouping pseudogenes into different categories. At one extreme, we obtained nonfunctional/"dead" pseudogenes – these elements responded negatively to all our activity indicators (are not transcribed, and lack any evidence of TF and Pol II binding, and active chromatin marks). Contrary to the actual definition of pseudogenes (“dead” genomic elements), this group comprised only ~20% of the total pseudogenes in each of the model organisms. On the other extreme, some, albeit very few, pseudogenes (88 in human, 32 in worm, 7 in fly) are both transcribed and simultaneously exhibit all other activity features (namely open chromatin, transcription factor and Pol II binding), despite the presence of mutations that disrupt the protein coding sequence. We label these pseudogenes as “highly active”. Some of the highly active pseudogenes have proof of translation without passing the protein coding gene annotation. This special set of pseudogenes requires a detailed experimental validation to assess their full biochemical activity potential. The majority of pseudogenes (~75%) are intermediate between these two, which have only a few of the classic indicators of activity. We labelled these pseudogene as “partially active”.

(c) Translation

Following this analysis we went even deeper and analysed the translation potential of transcribed human pseudogenes in four cell lines. We identified 20, 18, 14, and 19 translated pseudogene candidates in the four cell lines respectively. Evidence of translation was obtained with high confidence for three pseudogenes (Table YZ1). The unique peptide matches have not been

previously detected in any known proteins (UniProt) \cite{XXX} or variants (1000 Genomes Project dataset) \cite{XXX}. Even though the DNA sequence similarity between the pseudogenes and their parents ranges between 50 to 90%, the corresponding peptides have little or no sequence similarity with the protein products of the parent genes. This discordance is potentially due to the difference in reading frames between the translated products of the parents and the pseudogenes. We note that the three candidates have numerous disablements are only extreme cases of active pseudogenes. The fact that they make less than 1% of the total number of pseudogenes gives us confidence that they are real translated entities and not a case of mis-annotation. To study their full potential we analysed them in the context of additional activity and evolutionary data. All three pseudogenes showed various additional signs of biochemical activity. The coexpression correlation coefficient for the three pseudogenes with respect to their parent is reported in Table XXX. The low coexpression correlation coefficient for ENST00000533551, combined with its high sequence similarity to parents as well as the large number of activity features suggests that it is recently deceased, maintaining residual activity due to the recent pseudogenization event. By contrast, the relatively high coexpression correlation coefficients for ENST00000491897 and ENST00000431615 hint to potential regulatory roles.

(d) Upstream sequence

To complete our activity analysis of pseudogenes, we examined the proximal (within 2kb of the 5' end) upstream regions of pseudogenes and compared them to the analogous sequences in the parent and their paralogous genes. First, we analysed the upstream sequence similarity of pseudogenes and paralogs with that of their parents (Fig XXXIDEN_up2k_human_v2). While processed pseudogenes upstream sequence similarity to their parents matches, as expected, that of the genomic average, the duplicated pseudogenes show high level of similarity to their parent genes. Also, the majority of duplicated pseudogenes with high upstream sequence similarity also exhibit high sequence similarity to their parents in the "coding" region (Fig XXXIDEN_parent_PSSDpgene_human). These pseudogenes may be recent duplicated loci that have diverged little from their parents. However, there are also a number of interesting pseudogene-parent pairs with high upstream similarity despite low "coding" sequence identity, suggesting that the duplicated upstream regions may have been conserved via purifying selection. These scenarios could lead to a coordinated expression pattern between the parent gene and the transcriptional products regulated by these upstream regions (e.g. the duplicated pseudogene and/or other neighbouring coding genes).

In human, the paralog-parent comparison shows a similar trend to the duplicated pseudogene-parent comparison. Consequently we observed that the majority of paralog-parent pairs show high sequence similarity in both upstream and coding regions (Fig XXXIDEN_parent_paralog_human) and only a few examples exhibit high upstream but low coding sequence similarity.

To further our analysis we studied the evolution of the pseudogene upstream sequence with regard to their regulatory activities. To this end, we examined the histone modification (H3K27ac) ChIP-seq data. H3K27ac is a marker for active promoters and enhancers and thus is important in defining genomic functionality. We focused our analysis on pseudogene-parent pairs with only

one pseudogene per gene and no paralogs and pseudogene-parent-paralog triplets with one pseudogene and one paralog per gene. We examined the H3K27ac ChIP-seq signal in the upstream regions and observed that in general, the pseudogenes display a lower level of activity than the parent, while the paralogs have comparable activity to that of the protein coding gene (Fig XXXGRIDplots).

(e) Selection

Finally, we examined the selection in human pseudogenes. At the population level, the pseudogenes, as a unit, do not show any statistical significant enrichment over the genomic average. Therefore we divided them into different groups based on their activity features: transcribed vs. non-transcribed, and "highly-active" vs "partially-active", and "dead". As expected we found that the transcribed and "highly-active" pseudogenes are enriched in rare-alleles.

Function

Pseudogenes, by definition, were considered non-functional genomic elements, however, an increasing number of studies report the identification of biologically active pseudogenes performing regulatory role through their mRNA products \cite{21816204,18405356,20577206,18404147}. Finally, by combining the annotation, functional genomics and evolutionary data we annotated a set of potentially functional pseudogenes.

To this end we calculated the coexpression correlation coefficient between each pseudogene and their parent using the RNA-seq data (Fig XXX). Due to data availability we restricted this analysis to human and worm. The relationship between pseudogenes and their parental counterparts is extremely varied. In human, two thirds of the pseudogenes showed a significant level of correlated expression with their parent gene. By contrast, only half of the worm pseudogenes displayed a similar behaviour. In both organisms these pseudogene are a subset of the "highly-active" and "partially-active" pseudogene groups. Further we divided the pseudogenes in various categories based on their age, activity group and coexpression correlation coefficient (Table SXXX). We obtained a set of 10 high performance human pseudogenes (highly active, with a high sequence similarity to parents and a high coexpression correlation coefficient). Using this classification we were able to identify known regulatory pseudogene PTEN-P1 as part of the high performance group.

With the example of PTEN-P1 in mind, we also investigated the pseudogenes of other cancer-related genes annotated by the Sanger Institute. These cancer genes are significantly more likely than other genes to be parents of pseudogenes (1.5-fold, $p < 10^{-6}$). Among the 325 pseudogenes of cancer genes, 48 are transcribed and three, including PTEN-P1, are "highly-active". These findings warrant further study of pseudogene activity in cancer tissues, and are suggestive that other pseudogenes may be active in cancer.

Discussion

We report the first pseudogene comparison of the fully annotated genomes of human and three model organisms. We found that while all the species share common genic, regulatory and transcriptional principles \cite{mod1,mod2,mod3}, the pseudogene complement is organism specific reflecting their different evolutionary history. We show that the burst of retrotransposition events is a general mammalian characteristic.. Even more the pseudogene family analysis highlights the common ancestry of the olfactory elements between human and worm. Furthermore we show that differences in the disablement accumulation in the pseudogene sequence match the specific traits of each species (e.g. high deletion rate in fly genome, asexual reproduction for worm). The specificity extends even further to the physical localization and mobility of the pseudogene within and between chromosomes

Using RNA-seq data we found that the fraction of transcribed pseudogene is fairly consistent across all organisms (~15%). Even more, we found that a conserved fraction of transcribed pseudogenes show additional signs of activity. Also, contrary to previous assumptions, the majority of pseudogene (~75%) shows various signs of biochemical activity. Even more we found translation evidence for three human “partially active” pseudogenes suggesting a potentially regulatory role for these elements in specific cell lines.

We compared the pseudogene sequence with genes and lncRNAs in terms of repeat content within each specie. While the pseudogene and coding gene show overall a low repeat content, for the lncRNAs we observed a high repeat content and low conservation rate. These results suggest that repeat elements may underline the origin of some species-specific regulatory activities and even phenotypes \cite{XXX}. The regulatory function of several pseudogenes and lncRNAs have been previously demonstrated \cite{21816204,18405356,20577206,18404147}. Hence we suggest that these less conserved non-coding RNAs, with a repeat element driven genesis, may contribute to the species divergence due to their high organisms specificity. This highlights the importance of charting these “underdogs” of the ncRNAs’ world. For this purpose, we have taken the first step by annotating the pseudogenes, and prioritizing the potentially functional candidates by integrating the annotation with activity data (RNA-seq and ChIP-seq).

Next we examined the parallel evolution of pseudogene sequence looking at orthologous pseudogene across all the species. We found almost no conservation of pseudogenes between human and worm, and human and fly due to the large evolutionary distance between the three species. However we were able to identify a substantial number of orthologs between human and mouse with a high sequence similarity to the parents suggesting potentially that these elements are under selection being biologically relevant to their respective organisms.

We completed the evolutionary analysis of the pseudogene looking at the divergence of their upstream sequence. We found that the pseudogene biochemical activity in their upstream regions is less consistent with their parents’ and overall the pseudogene upstream regions diverge much faster. Further note that the upstream sequence diverges at different rates in the studied organisms.
Overall, analyzing the genome annotation we find a significant difference in the pseudogene

complement of the four organisms.

Materials and Methods

Annotation - Localization & mobility

Pseudogene distribution across chromosome length

Statistical tests of co-localization tendency

For each of the studied species, we performed the co-localization tendency analysis. We extracted duplicated and processed pseudogenes from the annotated pseudogenes, and analysed the two biotypes respectively.

Each pseudogene in a specific biotype was paired with its unique parent coding gene. For each chromosome, we generated a 2-by-2 contingency table A, whose elements are A_i,j, i=1 or 2, j=1 or 2. A_1,1 is 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. Fisher's exact test was applied to the contingency table for each chromosome, to test whether the pseudogenes and their parents tend to reside on the same chromosome. The significance threshold with Bonferroni correction was 0.05/n, where n is the total number of tested chromosomes in this species.

Statistical tests of importer/exporters

Next we inspected the material exchange between different chromosomes, excluding the co-localizing pseudogenes-parent pairs. The analysis was performed for two pseudogene biotypes respectively. We used two linear regression models to detect significant importer and exporter chromosomes.

For exporter chromosome detection, the null hypothesis is that for most of the chromosomes, the frequency of exporting parent genes (Fex_i) is proportional to the number of coding genes on the same chromosome (N_i), where i is the index of chromosome. This proportionality can be captured by a linear regression N_i ~ Fex_i. Any chromosome outside of the 95% confidence interval are considered a significant strong or weak exporter.

For importer chromosome detection, the null hypothesis is that for most of the chromosomes, the frequency of imported pseudogenes (or paralogs) (Fim_i) is proportional to the length of the chromosome (L_i). Similarly, this proportionality can be captured by a linear regression L_i ~ Fim_i. Any chromosome outside of the 95% confidence interval are detected as a significant strong or weak importer.

Annotation - Orthologs

The large difference in the speciation time between our model organisms resulted in a pair-specific definition of pseudogene orthologs. We define human – mammal pseudogene orthologs if they are syntenic and share parent gene orthology. Going further away from humans on the evolutionary scale, we restrict the orthology to pseudogene that share orthologous parents.

Activity - Translation

We constructed a workflow to identify translated pseudogenes (FigYZ S8). First, we generated putative peptides using a 3-frame translation of annotated pseudogenes. We built a target peptide sequence database by merging the putative peptide datasets with the complete human proteome \cite{UniProt}. Next, we matched the pooled mass spectrometry data to the target peptide sequence dataset using the Peppy software \cite{23614390}. We used the default search settings (note XXX) and the Peppy-generated decoy database, with peptide identification FDR < 0.01. Subsequently, any peptides matching known proteins or variants (according to UniProt) were excluded from the unique peptide list. Furthermore, only the unique peptides identified in at least two of the analysed cell lines were selected for subsequent analysis.
We annotated a pseudogene as putatively translated if it has two or more unique peptide matches, that do not match any known gene or variants. We used two high quality data sets: RNA expression (RPKM data \cite{askBP}) and protein expression (mass spectrometry spectra \cite{22278370}). For quality control and validation we used additional datasets (TableYZ S1) We used blastp algorithm \cite{XXX} to compare the sequence similarity of pseudogene peptides and their parent proteins. The 1000 Genomes Project variant data \cite{askSB} was used for further validation of novel peptides originated from pseudogenes.

Summary

*** *RESOURCE: man annotation + activity data*

We describe a comprehensive pseudogene resource highlighting the completion of the manual annotation of four model organisms: worm, fly, zebrafish and human. We integrate the manual annotation with functional genomics and evolutionary data to obtain a detailed map of the pseudogene complement of the four organisms. We aim to give an insight into the presence and role of pseudogenes in various species.

In order to understand the role of pseudogenes in different organisms we analyse them on multiple levels, from genomic localization and genesis to evolution and activity.

-- Localization and mobility

We start our study looking into the pseudogene chromosomal localization and exchange. We found that most of the human pseudogenes are uniformly distributed along the chromosomes arm with a slight enrichment towards the centromer. On the contrary, the majority of worm pseudogenes are located near the telomeres, while in fly there is a statistical significant increase in the number of pseudogene located near the centre of the chromosome.

Next we analysed the tendency of pseudogenes to co-localize on the same chromosome as their parents'. We found, as expected, that duplicated pseudogene tend to be situated on the same chromosomes as their parent gene, while processed pseudogene are randomly scattered across the genome. Differentiating between autosomes and sex chromosomes, we found that in human and fly the co-localization tendency of duplicated pseudogenes is more substantial for the latter. Studying the pseudogene exchange between chromosomes, we observed that in human the X chromosome is a significant importer of processed pseudogenes; whereas, Y is an importer of duplicated pseudogenes. Also, as expected, we observed a preferential exchange of duplicated pseudogene between the sex chromosome, while the rate of exchange with the autosomes is significantly lower.

*** *PSEUDOGENES DIFFER REFLECTING ORGANISM HISTORY*

Overall we find that the pseudogenes differ much more between organisms than protein coding genes or other non-coding elements, reflecting much more closely the genome history.

-- First - no ortholog pgenes preserved

First we study the concurrent evolution of pseudonome by analysing pseudogenes of orthologous genes. We observe that the model organisms share no similarity in the "orthologous" pseudogenes set.

-- Family

Pseudogene family analysis reveals only few similarities between the organisms. Fish,

nematode and insect genomes show an organism-specific family distribution, while mammalian genomes share the identity of the top pseudogene families, though without preserving their rank. For instance the ribosomal protein families top the charts in human, while worm and fly pseudofamily's hierarchy are lead by the chemoreceptor and the SAP protein family, respectively. While pseudogene family distribution is very much organisms specific we observed the conservation of the 7-transmembrane protein as the top family in both human and worm possibly reflecting the coevolution of olfactory and chemoreceptor in primates and nematodes. Also, we found that there is no relationship between the large gene families and the large pseudogene families in any of the studied species, as well as no relationship in terms of number of pseudogenes and the size of the gene family.

-- Age & Pseudogene disablements

Next we focus on the relative distribution of pseudogene biotypes as a function of age. We find that the human genome is enriched in processed pseudogenes while worm, fly, and zebrafish are enriched in duplicated pseudogenes.

In comparing worm & fly, association of the pseudogenes with indels reflects once again the organism evolutionary differences. The depletion of pseudogenes in the fly genome is reflective a large effective population size \cite{14631042} and its prevalence for deletions. By contrast, the indels abundance in worm is primarily the byproduct of a largely asexual mode of reproduction. The worm has a small effective population size and its genome is prone to the accumulation of mutations/insertions \cite{17637734}. Consequently we found an enrichment of insertions as pseudogene disablements.

**** CONSISTENT INTERMEDIATE ACTIVITY*

Next, we integrated functional genomics data with the pseudogene annotation in order to identify pseudogene with signs of biochemical activity. Overall, we found that ~20% of pseudogenes are transcribed. Further we tested the pseudogenes for features of genomic activity and classified them into three groups: highly active, partially active and dead. We observed a consistent distribution of activity levels in all the organisms with ~5% of pseudogenes being fully active, 20% dead and the majority (75%) showing only partial signs activity.

**** EVOLUTION*

-- Upstream sequence

In order to be able to understand the evolution of pseudogene in various species we analysed the divergence of the upstream regions. Given the similarity in the genesis of pseudogenes and paralogous genes, we found that the pseudogene upstream region biochemical activity (as exemplified by the presence of active histone marks) is not preserved relative to the parent. By contrast paralogous protein coding genes maintain a high level of activity in the upstream

regions, similar to the parent gene. However, for duplicate pseudogenes there seems to be subpopulation that has higher levels of sequence similarity in their upstream regions than what is observed in paralogs. Furthermore, we found that the pseudogene activity in their upstream regions is less consistent with the parents one, than it is observed for paralogous genes and overall the pseudogene upstream regions diverge much faster. Further we examine the degree to which they diverge relative to the actual coding sequences and how they differ from the regulatory features of active genes. We note that the upstream sequence diverges at different rates in the studied organisms.

*** CONCLUSION

Finally we identified and characterized a group of potentially functional pseudogenes. Our analysis shows that pseudogenes are a fingerprint of the organism evolution. [[TBC]]

[[Original Abstract]]

We describe a comprehensive comparison of human pseudogenes to three other fully annotated model organisms as part of the ENCODE project. We aim to give an insight into the presence and role of pseudogenes in various species. The pseudogenes are analysed at four levels: annotation, activity, evolution, and function.

First, we compared the distribution of pseudogenes with respect to their biotype, age, defects, family and paralog diversity. We note that mammalian organisms show stages of processed pseudogene enrichment indicating bursts of retrotransposition events, while insects and nematodes are depleted in the processed pseudogene complement, but are enriched in organisms specific defects and have unprocessed pseudogenes as the dominant biotype. The pseudogene family analysis indicates that while there are similarities in the top families across the mammalian species, pseudogenes are mostly organism-specific. The results reflect differences in the pseudogenization processes between the various organisms.

Secondly, we looked at pseudogene activity. We selected a number of features that are characteristic to protein coding genes and classified the pseudogenes accordingly. Thus we obtained 3 classes of pseudogenes (“active”, “zombie”, and “dead”) and we compared their variations in mammals, nematodes and insects.

Thirdly we studied the evolution of pseudogenes in different organisms. We examined the degree to which they diverge relative to the actual coding sequences and how they differ from the regulatory features of active genes. We note that the upstream sequence diverges at different rates. Next, we study pseudogenes of orthologous genes. We observe that the three organisms share no similarity in the “orthologous” pseudogenes set.

Finally we identified and characterized a group of potentially functional pseudogenes. Our analysis shows that pseudogenes are a fingerprint of the organism evolution. [[TBC]]

Comments:

In worm, the majority of pseudogenes are near the telomeres, a location characterized by a high number of recombination events and rapid gene evolution \cite{8536965}

WC: This may not be true. the telomeres and centromeres in worm see less recombination. (see PMID: 19289596)

CSDS: Well this paper seems to differ PMID 8536965 but, yep, it's older, however it's cited in a 2009 too. will look into this more.

[[WC: First off refer to PMID: 17637734. Major factors contributing to types of mutations are recombination rate and effective population size. Worm is hermaphroditic. There is some recombination, but the offspring resemble parents aside from the accumulated mutations. Consequentially, worm has a small effective population size (but not smaller than mouse and human). It accumulated junk and insertions. Fly on the other hand is the opposite, it reproduces sexually, and has a large effective population size. This enables the genome to be streamlined (or to be precise, non-adaptive junk doesn't become fixed). Refer to PMID: 14631042 for citation on popsize for these organisms]]