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 and integrated the 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 biochemical activity for the pseudogenes in each organism. 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.

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 comprehensive 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). 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 (genes from which they originated) can cause problems in experiments directed at protein coding regions. However the pseudogene annotation is a difficult and complex process. The sequence decay at pseudogene loci makes it challenging to rightly identify authentic pseudogenes and accurately define their boundaries \cite{22951037}. To this end we used a hybrid approach, combining manual annotation with

computational pipelines. While providing high accuracy, the manual annotation 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. The finished annotation is even more important in analysing pseudogenes with potential biological activity since it reduces the false discovery rate and the potential of mis-annotation. Furthermore integrating the extensive functional genomics, proteomics and evolutionary data available allowed us to uncover the differences in the pseudogenization process. Also, our analysis shows that the pseudogene repertoire is lineage specific and has important implications for the genome evolution. Even more the fraction of pseudogenes with residual biochemical activity is similar in all four organisms.

Results

Annotation Resource

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 50-fold more pseudogenes than zebrafish, 100-fold more than fly but only 15-fold more than worm (Table XXX). The scarcity of fly pseudogenes can be explained by the high rate of DNA loss \cite{12572619,1806330,9402741}, an intrinsic characteristic of its large effective population size \cite{12572619,9501496,14631042}. 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 annotation pipeline (PseudoPipe). As expected, the two mammals show a similar pseudogene content to human (Table XXX). All the data resulting from the annotation and comparative analysis of the four species was collected into a comprehensive pseudogene resource and is available through *pseudogenes.org*.

Classification, Genesis & 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 SXXX). The preference for processed pseudogenes in human, 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 occurred

on a much larger scale and it can be regarded as an intrinsic characteristic of mammals. The enrichment of duplicated pseudogenes in worm and fly can be related to relatively high gene duplication rates \cite{11861885,11230161,21295484,19622155,19289596}. 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}. Overall, analysing the genome annotation we find a significant difference in the pseudogene complement of the four organisms.

(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 fraction of processed pseudogenes at different ages (Fig SXXX). In human, the prominent peak of processed pseudogenes fraction, at high sequence similarity, corresponds to the burst of retrotransposition events, 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 at similar time points. By contrast, in zebrafish and worm, the majority of older pseudogenes are processed whereas younger ones are mostly duplicated. The constant, if rather low, ratio of processed to duplicated pseudogenes in fly genome is the result of numerous duplication events combined with a high deletion rate.

c) Genesis

Further we studied the complex process of pseudogene genesis. Repeat elements play an important role in the retrotransposition 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

Finally we analysed the variety and propensity of disablements as markers of the pseudogene evolution. Given the fact that the majority of human pseudogenes are of recent descent, we observed a lower disablements density in their sequences, compared to worm, fly and zebrafish (Fig SXXX). Based on their origins, we distinguished three types of disablements: insertions, deletions, and stop codons (Table XXX). 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. 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 relative insertion abundance in

worm is the by-product of its small genome size. Worm's compact genome supports a higher frequency of small insertions \cite{15295601} while favouring large deletions over shorter ones \cite{12911038}.. Consequently we found an enrichment of insertions as pseudogene disablements.

Localization & Mobility

Next we 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. The weaker correlation between the number of pseudogenes and protein coding genes per chromosome (Fig SXXX) suggests the existence of pseudogene inter-chromosomal transfers. By contrast in worm and fly we see a strong correlation between the two, while in zebrafish there is no correlation at all. To this end we analysed the relative position of the pseudogenes within a chromosome and their inter-chromosomal mobility

(b) Localization

In human, we observed a uniform distribution of pseudogenes across the chromosome length (Fig XXX). However, worm, fly, and zebrafish showed a skewed distribution of pseudogenes. In worm, the majority of pseudogenes are near the telomeres, regions characterized by a high number of recombination events and rapid gene evolution \cite{8536965}. By contrast, fly pseudogenes are preferentially located near the centromeres, consistent with a high deletion rate in the telomeric regions due to the large effective population size.

Further we looked at pseudogene tendency to reside on the same chromosome as their parent genes. As expected, the 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 is especially significant for human Y, and fly X chromosome. This result is indicative of the low recombination rate of the sex chromosomes \cite{16545149,1875027,15059993}, the duplicated pseudogenes therefore cannot be "crossed out". The colocalization of duplicated pseudogenes and parent genes is also statistically significant (FigYZ 1 (B)) for human autosomal chromosomes 7 and 11. This results relates to the fact that chromosome 11 is enriched in olfactory receptors \cite{11337468} while chromosome 7 is enriched for genome duplication events \cite{12853948}.

(c) Mobility

Next we studied 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 all 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), 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.

Orthologs, Paralogs & Families

Further, we compared the lineage specificity of pseudogenes in the four organisms by analysing their families and orthologs.

(a) Orthologs

In this study we focused on three sets of orthologs: human-worm-fly, human-zebrafish, and human-mouse (Table XXX). Overall there are no pseudogene orthologs across all organisms and only a small number are found at a pair-wise level.
First, we analysed ~2000 1-1-1 human-worm-fly orthologous protein-coding genes (Table XXX). We observed that not one of the triplets 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 25 (mostly processed) pseudogenes spread randomly across the human genome, three duplicated pseudogenes clustered near the parent gene in fly and none in worm.
In order to get a better understanding of human pseudogene evolution and specificity we looked at closer relatives examining the human-mouse orthologs. We found that only 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).

(b) Paralogs & Families

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 pseudogene counterpart, and a large fraction of pseudogenes are associated with a few highly expressed gene families. Surprisingly there is little overlap between large gene and large pseudogene families. At the extreme we found a number of genes that are enriched in pseudogenes are depleted in paralogs, and vice-versa, a trend common across all four organisms.
Pfam analysis allowed for a bigger pattern to emerge. As expected, the ribosomal proteins are the dominant families across human, macaque and mouse (Fig XXX). These abundantly expressed genes are indicative of the general burst of retrotransposition events \cite{16504170}. However, while the top families are shared among mammals their relative rank is organism specific. The top pseudogene families in worm are the 7 Transmembrane proteins, perhaps reflecting the family rapid evolution \cite{11961106} and the many duplications events in nematode genome history \cite{19289596,18837995}. It is interesting to note that the human genome shares as well this top family, as evidence of 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 lineage specificity of the pseudogene top families, 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.

Activity

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

(a) Transcription

Analysing RNA-Seq data we 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, a subset of these (~13% in human and ~30% in worm and fly), have a discordant transcription pattern with their parent genes over multiple samples (Fig SXXX). Also the parent genes of broadly expressed pseudogenes tend to be broadly expressed as well (Fig SXXX), but the reciprocal statement is not valid. However, in general pseudogenes are less broadly transcribed than their coding counterparts, being expressed in only a single cell line or developmental stage (Fig SXXX). Specifically, only 5.1%, 0.69%, and 4.6% are broadly expressed in human, worm, and fly, respectively (Table 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 XXX), grouping them into different categories. At one extreme, we clustered “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 in each organisms. On the other extreme, some, albeit very few, pseudogenes (<5%) are both transcribed and simultaneously exhibit all other activity features, despite the presence of disruptive mutations. We labelled these pseudogenes as “highly active”. 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, having only a few of the classic indicators of activity. We labelled these pseudogenes as “partially active”.

(c) Translation

Following this analysis we studied 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. The low number of translation candidates (<1%) is indicative of the annotation

quality and gives us confidence that they are potentially real translated entities. Evidence of translation was obtained with high confidence for three pseudogenes (Table YZ1). Even though the sequence similarity between the pseudogenes and their parents ranges between 50 to 90%, the corresponding pseudogene peptides have little or no sequence similarity with any protein products of known coding genes or variants. This discordance is related perhaps to the difference in reading frames between the translated products. The three candidates have numerous disablements and are only extreme cases of active pseudogenes. To study their full potential we analysed them in the context of activity and evolutionary data (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 the other two pseudogenes, coupled with various activity marks hint at potential regulatory roles.

(d) Upstream sequence similarity

To complete our activity analysis of pseudogenes we examined the similarity in proximal (within 2kb of the 5' end) upstream regions of pseudogenes and their parents. First, we compared the similarity in the upstream sequence of pseudogenes and parents with that of paralogs and parents (Fig XXXIDEN_up2k_human_v2). The processed pseudogenes upstream sequence similarity matches the genomic average. The majority of duplicated pseudogenes show high level of similarity in both upstream and "coding" regions (Fig XXXIDEN_parent_PSSDpgene_human). These pseudogenes may be recent duplicated loci that have diverged little from their parents. However, there is also a number of interesting 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 upstream regions. In human, the paralog-parent upstream similarity shows a comparable trend to the duplicated pseudogene one (Fig XXXIDEN_parent_paralog_human) with only a few examples of pairs with high upstream but low coding sequence similarity.

(e) Upstream sequence activity

To further our analysis we studied the pseudogene upstream sequence regulatory activity. To this end, we examined the ChIP-seq data of H3K27ac, an important marker in defining genomic functionality. We focused our analysis on protein coding genes with only one pseudogene but no paralogs, and those with one pseudogene and one paralog. We 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)

(f) Selection

Finally, we examined the selection in human pseudogenes studying the derived allele frequency. 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 RNA products \cite{21816204,18405356,20577206,18404147}. 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. We observed that cancer genes are significantly more likely than other genes to generate pseudogenes. Among the 325 cancer pseudogenes, 48 are transcribed and three, including PTEN-P1, are "highly-active".. These findings warrant further study of pseudogene activity and are suggestive that other pseudogenes may play active role in various diseases.

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. 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).

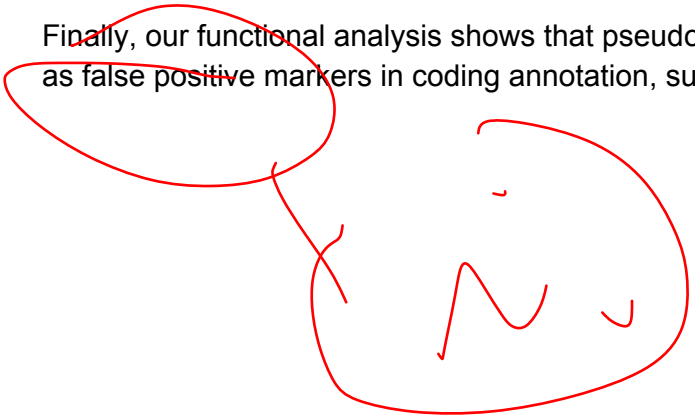
By comparing the pseudogene sequence with genes, UTRs and lncRNAs in terms of repeat content we found that repeat elements may underline the origin of some species-specific regulatory activities and even phenotypes \cite{XXX}. The regulatory function of several pseudogenes and lncRNAs have been previously 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).

By examining the parallel evolution of orthologous pseudogene across all the species we were able to identify a small 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 species.

We completed our analysis by looking at a variety of pseudogene activity features. Using RNA-seq data we found that the fraction of transcribed pseudogene is fairly consistent across all organisms (~15%). Even more, contrary to previous assumptions, the majority of pseudogene (~80%) shows various signs of biochemical activity.. While the pseudogene biochemical activity in their upstream regions is less consistent with their parents’ and overall the pseudogene upstream regions diverge much faster, we found a subgroup of duplicated pseudogenes with highly conserved upstream regions hinting at potential regulatory roles for these particular elements.

Finally, our functional analysis shows that pseudogene importance bypasses their common use as false positive markers in coding annotation, suggesting interesting regulatory roles.



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]]