Comparative Analysis of Pseudogenes: History Trumps Conservation…

[[m22: hist trumps cons – CSDS still working on the title]]

Abstract

In this study, we present a comprehensive pseudogene resource highlighting the completed pseudogene annotation in human and three key model :organisms, worm, fly and zebrafish. We also introduce the first analysis of the draft mouse pseudogene annotation. We find that even more than the protein coding genes, the pseudogene complement has a strong lineage specificity reflecting the different genome remodeling processes that marked the organisms' evolution.

To this end we find that the human pseudogene complement is marked by a single large event, the retrotranspositional burst that occurred at the dawn of the primate lineage. The uniformity in the pseudogene distribution and variety across human and mouse reflects a highly correlated evolutionary history. By contrast the fly pseudogenes are the product of its large population size and thus are governed by numerous disabling deletions. The worm pseudogene complement is dominated by large duplication events associated with particular environmental responses gene families. A similar pattern of multiple tandem duplications combined with high recombination rates depleted the pseudogene complement in zebrafish. Despite a diverse distribution of pseudogenes, we find a consistent inter-chromosomal traffic for the sex chromosomes across all organisms.

Finally we identify a large spectrum of biochemical activity for the pseudogenes in each organism ranging from "highly active" to "dead". The distribution of these activity levels is consistent across all four species implying a uniform degradation mechanism of functional elements. Furthermore while we see a consistent trend in terms of the fall off of the pseudogene' upstream histone marks activity relative to the coding genes, we find a small population of pseudogenes with a highly conserved upstream sequence hinting at potential regulatory roles. We complete our analysis ranking the pseudogenes based on their activity features and pinpointing 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}. Presumed to evolve with little selection constraints \cite{10833048}, the 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

Cristina Sisu 4/2/14 14:44
Deleted: [[m28: say what we find not generally what we do]]

Cristina Sisu 4/2/14 14:44
Deleted: of pseudogenes

Cristina Sisu 4/2/14 14:44
Deleted: pseudogene annotation in the

Cristina Sisu 4/2/14 14:44
Deleted: genome. We create a detailed map of the pseudogene complement for each organism integrating the annotation with functional genomics and evolutionary data.

Cristina Sisu 4/2/14 14:44
Deleted: is

Cristina Sisu 4/2/14 14:44
Deleted: specific

Cristina Sisu 4/2/14 14:44
Deleted: evolutionary history of the studied species. [[m28: much more so than genes]]Also, specie specific traits are revealed by differences in the disablement accumulation in the pseudogene sequence. Similarly, the distribution of pseudogene families is organism dependent. By contrast, studying pseudogene localization and mobilit ... [1]

Cristina Sisu 4/2/14 14:44
Deleted: characteristics of the sex chromosomes pseudogene compler ... [2]

Cristina Sisu 4/2/14 14:44
Deleted: burst of retrotransposition events

Cristina Sisu 4/2/14 14:44
Deleted: when the bulk of human

Cristina Sisu 4/2/14 14:44
Deleted: were created is actually a general mammalian characteristic. [... [3]

Cristina Sisu 4/2/14 14:44
Deleted: .

Cristina Sisu 4/2/14 14:44
Deleted: pseudogene

Cristina Sisu 4/2/14 14:44
Deleted: as well as a consistent involvement

Cristina Sisu 4/2/14 14:44
Deleted: in organism biology.

Cristina Sisu 4/2/14 14:44
Deleted: .[[m28: more preservation of seq than of activlit than para]]

Cristina Sisu 4/2/14 14:44
Deleted: [[m28: who are the referees and make sure we cite them right]]

\cite{17099229,22951037,11160906,12560500,15860774,12083509,16925835}. Earlier comparative studies across various organisms have focused on a specific family or class of pseudogenes \cite{15289807,16469101,12417195,16680195,19835609,12034841,19123937,23555032}. The absence of a solid finished annotation and the potential of mis-mapping of functional genomics data had previously rendered a detailed comparison of the pseudogene complement in various organisms rather inaccurate. The availability of the complete genome annotation of the human, worm, fly, and zebrafish genomes allows us for the first time to embark on a uniform and comprehensive comparison of pseudogenes across these organisms. We use the pseudogenes as fingerprints of genome history to find key steps in these species evolution.

While they all share common regulatory and transcriptional principles, these organisms could not have been more different. In order to better understand the implications of our results for the human genome we introduce and analyse the first draft of mouse pseudogenes.

The pseudogene prevalence, non-standardized annotation reflected by large fluctuations from one release to another, as well as the 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 even more important in analysing pseudogenes with potential biological activity since it reduces the false discovery rate and the potential of mis-annotation.

Our analysis shows that the pseudogene repertoire is lineage specific and has important implications for the genome evolution. Integrating the extensive functional genomics, proteomics and evolutionary data available allowed us to uncover the differences in the pseudogenes evolution. Also, despite large differences in the pseudogene content, the fraction of pseudogenes with residual biochemical activity is similar in all four organisms.

Results

The Pseudogene Annotation Resource.

In this study, we present the completed pseudogene annotation in human, worm, fly and zebrafish. 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 scrutiny with computational predictions. 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. 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 (Table XXX).

Cristina Sisu 4/2/14 14:44
Deleted: At

Cristina Sisu 4/2/14 14:44
Deleted: first glance, these results hint at large differences in the pseudogene distribution and functionality between organisms.[[m28: shots in the foot]][[m28: can really compare betw organisms w o

Cristina Sisu 4/2/14 14:44
Deleted: b/c you don't if something missing isnt]]

Cristina Sisu 4/2/14 14:44
Deleted: the pseudogene complement

Cristina Sisu 4/2/14 14:44
Deleted: four

Cristina Sisu 4/2/14 14:44
Deleted: .

Cristina Sisu 4/2/14 14:44
Deleted: In this paper we describe the first study focused on analysing and contrasting the pseudogene complement in human, worm, fly, and zebrafish.

Cristina Sisu 4/2/14 14:44
Deleted: the four

Cristina Sisu 4/2/14 14:44
Deleted: Thus we use the

Cristina Sisu 4/2/14 14:44
Deleted: as fingerprints of genome history to find key steps in the four species evolution.

Cristina Sisu 4/2/14 14:44
Deleted:

Cristina Sisu 4/2/14 14:44
Deleted: their parents (

Cristina Sisu 4/2/14 14:44
Deleted: from which they originated) raised

Cristina Sisu 4/2/14 14:44
Deleted: Thus, using a uniform annotation procedure we curated a highly accurate and exhaustive pseudogene set for each of the four organisms.

Cristina Sisu 4/2/14 14:44
Deleted: [[m22: rel. of pgenes to overall genome remodelling - processing , dupl.]] .

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). In contrast with the model organisms, 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.

Cristina Sisu 4/2/14 14:44
Deleted: in order to better understand the implications of our results for study of the human genome, we included in the analysis two mammalian species:

Cristina Sisu 4/2/14 14:44
Deleted: 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). 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. 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. In fly we observed a constant, if rather low, ratio of processed to duplicated pseudogenes.

c) Genesis

Further we studied the complex process of pseudogene genesis. 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

Finally 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, fly and zebrafish (Fig SXXX). While this might hint at a potential selection bias, the analysis of derived allele frequency shows that at the population level, the human pseudogenes have no statistical significant enrichment over the genomic average. 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.

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 chromosomes length with a slight enrichment towards the centromere (Fig XXX). This distribution is even more pronounced in fly and zebrafish. By contrast, in worm, the majority of pseudogenes are near the telomeres, regions characterized by high recombination rates and rapid gene evolution \cite{8536965}.. To this end we analysed the recombination rate for each species. In accord with our observations, the human, fly and zebrafish pseudogenes show enrichment for regions of low recombination. However in worm we observed a rather uniform recombination rate for genes and pseudogenes, consequence of recent selective sweeps that pruned its genome.

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 result relates to the fact that chromosome 11 is enriched in olfactory receptors, which tend to be highly duplicated sequences\cite{11337468}, and chromosome 7 is enriched for genome duplication events \cite{12853948}.

Cristina Sisu 4/2/14 14:44
Deleted: Given the fact that the majority of human pseudogenes are of recent descent, we

Cristina Sisu 4/2/14 14:44
Deleted: their

Cristina Sisu 4/2/14 14:44
Deleted: . ... [4]

Cristina Sisu 4/2/14 14:44
Deleted: However, worm,

Cristina Sisu 4/2/14 14:44
Deleted: ,

Cristina Sisu 4/2/14 14:44
Deleted: showed a skewed distribution of pseudogenes. Surprisingly

Cristina Sisu 4/2/14 14:44
Deleted: }. 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

(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 studied organisms 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 XXX). 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 speces. Surprisingly the majority of the orthologous pseudogenes (127) are processed and have a high sequence similarity to their parents (Fig SXXX).

Next, we analysed the pseudogenes originating from ~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 XXX). Also we observed that the number of pseudogenes associated with protein coding orthologs, differs greatly across species. As an example (Fig XXX) 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 or zebrafish.

(b) Paralogs & Families

Next we compared the overall distribution pattern of pseudogenes and paralogs per parent gene (Fig XXX). The distribution of pseudogenes per gene is highly uneven. 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. As a result, a large fraction of pseudogenes are associated with a few highly expressed gene families. Surprisingly there is little overlap between gene families 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.

Cristina Sisu 4/2/14 14:44
Deleted: four

Cristina Sisu 4/2/14 14:44
Deleted: [[m28: there

Cristina Sisu 4/2/14 14:44
Deleted: 2000 1-1-1 orthologs … there is some conservation…. there orloghogs

Cristina Sisu 4/2/14 14:44
Deleted: mouse… this is the basel… [5]

Cristina Sisu 4/2/14 14:44
Deleted: .

Cristina Sisu 4/2/14 14:44
Deleted: ~1% p[[m28:

Cristina Sisu 4/2/14 14:44
Deleted:]]

Cristina Sisu 4/2/14 14:44
Deleted: .

Cristina Sisu 4/2/14 14:44
Deleted: studied

Cristina Sisu 4/2/14 14:44
Deleted: orthologous protein coding genes in human, worm, and fly. We analysed

Cristina Sisu 4/2/14 14:44
Deleted: (Table XXX)

Cristina Sisu 4/2/14 14:44
Deleted: SXXX). The

Cristina Sisu 4/2/14 14:44
Deleted: a

Cristina Sisu 4/2/14 14:44
Deleted: ortholog

Cristina Sisu 4/2/14 14:44
Deleted: the three

Cristina Sisu 4/2/14 14:44
Deleted: [[m28::

Cristina Sisu 4/2/14 14:44
Deleted:]]distributions

Cristina Sisu 4/2/14 14:44
Deleted: four

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 (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 a highly duplicated 7TM as its 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 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 and studying their diversity in human, worm, fly, and zebrafish.

(a) Transcription

Analysing RNA-Seq data we found 1,441, 143, 23, 31, and 878 potentially transcribed pseudogenes in human, worm, fly, zebrafish and mouse respectively (Fig XXX). 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 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 XXX), 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". Even more, 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

Cristina Sisu 4/2/14 14:44
Deleted: It is interesting to note that

Cristina Sisu 4/2/14 14:44
Deleted: well this

Cristina Sisu 4/2/14 14:44
Deleted: the studied

Cristina Sisu 4/2/14 14:44
Deleted: . ... [6]

Cristina Sisu 4/2/14 14:44
Deleted: 31

Cristina Sisu 4/2/14 14:44
Deleted: and

Cristina Sisu 4/2/14 14:44
Deleted: 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).

RACHEL

Cristina Sisu 4/2/14 14:44
Deleted:).[[m28: mouse txn is consistent]]

Cristina Sisu 4/2/14 14:44
Deleted: [[m28: 1 / 6 txn cons & 1/ 5 dead consis]] . ... [7]

Cristina Sisu 4/2/14 14:44
Deleted: in each organisms.

Cristina Sisu 4/2/14 14:44
Deleted: %),

between these two, having only a few of the classic indicators of activity. We labelled these pseudogenes as "partially active". Surprisingly the distribution of pseudogenes for the three activity levels is consistent across all studied species.

(c) Translation

Following this analysis we studied the translation potential of transcribed human pseudogenes in four ENCODE cell lines using mass spectrometry data. From over 14000 pseudogenes we identified only 20, 18, 14, and 19 translated candidates in the four cell lines respectively. The low number of translation candidates (<1%) is indicative of the high quality of our annotation and gives us confidence in their authenticity. Three pseudogenes exhibit high confidence translation evidence (Table YZ1). Their corresponding peptides have little or no sequence similarity with any protein products of known coding genes or variants suggesting that these pseudogenes may use open reading frames different from their parents'. 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). 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.

(d) Upstream sequence similarity

The pseudogene activity is strongly 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 XXXIDEN_parent_PSSDpgene_human). 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.

(e) Upstream sequence activity

To complete our analysis we examined the preservation of promoter activity in pseudogenes compared to their parents'. 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 XXXGRIDplots).

"Functional" Pseudogene Candidates

Finally we attempted to pinpoint potentially "functional" pseudogenes. By definition, pseudogenes are considered non-functional genomic elements. However, an increasing number

Cristina Sisu 4/2/14 14:44
Deleted: . ... [8]

Cristina Sisu 4/2/14 14:44
Deleted: *activity*

Cristina Sisu 4/2/14 14:44
Deleted: of pseudogenes

Cristina Sisu 4/2/14 14:44
Deleted: their

Cristina Sisu 4/2/14 14:44
Moved down [1]: 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.

Cristina Sisu 4/2/14 14:44
Deleted: sequence of pseudogenes in human, worm, and fly.

Cristina Sisu 4/2/14 14:44
Deleted: upstream regions of the

Cristina Sisu 4/2/14 14:44
Deleted: .

Cristina Sisu 4/2/14 14:44
Deleted: 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).

Cristina Sisu 4/2/14 14:44
Deleted: *(e) Upstream sequence similarity* . ... [9]

Cristina Sisu 4/2/14 14:44
Deleted: upstream

Cristina Sisu 4/2/14 14:44
Moved (insertion) [1]

of studies report biologically active pseudogenes performing regulatory roles through their RNA products \cite{21816204,18405356,20577206,18404147}.

By combining the annotation, functional genomics and evolutionary data we identified a set of "functional" candidates. Due to data availability we restricted our analysis to human and worm. We calculated the coexpression correlation coefficient between each pseudogene and their parent using the RNA-seq data (Fig SXXX). To identify a list of potentially functional pseudogenes, we divided the pseudogenes in various categories based on their sequence conservation, 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) including the known regulatory pseudogene PTEN-P1.

With the example of PTEN-P1 in mind, we also investigated the human pseudogenes of 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 experimental validation of pseudogene activity and are suggestive that other pseudogenes may play functional roles.

Discussion

We report the first uniform multi organisms' pseudogene comparison leveraging on the finished annotations of the human, worm, fly, and zebrafish genomes and the draft mouse genome. 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.

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 X chromosome. 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 intrinsic characteristic of mammals. In contrast, worm, fly, and zebrafish pseudogene complements tell a story of numerous duplication events. This became apparent in the worm genome due to the

Cristina Sisu 4/2/14 14:44
Deleted: fully annotated genomes of

Cristina Sisu 4/2/14 14:44
Deleted: , as a [[m28: most]]population,

Cristina Sisu 4/2/14 14:44
Deleted: [[m28::

Cristina Sisu 4/2/14 14:44
Deleted: a history of

Cristina Sisu 4/2/14 14:44
Deleted: specfici

Cristina Sisu 4/2/14 14:44
Deleted:]]For example, given the fact

Cristina Sisu 4/2/14 14:44
Deleted: the bulk of human

Cristina Sisu 4/2/14 14:44
Deleted: was created rather recently,

Cristina Sisu 4/2/14 14:44
Deleted: observed

Cristina Sisu 4/2/14 14:44
Deleted: differences in the number and

Cristina Sisu 4/2/14 14:44
Deleted: between human

Cristina Sisu 4/2/14 14:44
Deleted: three model organisms. The large size

Cristina Sisu 4/2/14 14:44
Deleted: the pseudogene complement in the human genome as well as the preference for processed pseudogenes, can be traced back 40 MYa to a burst of retrotransposition events. Surprisingly the frequency and biotype distribution of human

Cristina Sisu 4/2/14 14:44
Deleted: is shared by macaque and mouse as well. While

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 and the zebrafish genomes, we observe tandem duplication events \cite{22702965}. However, the high deletion rate resulted in a depletion of the pseudogene complement in the two organisms and consequently we see a segregation of the remaining pseudogenes to areas of low recombination. This may also reflect the fly large effective population size \cite{12572619,9501496,14631042} and the strong selection it's intergenic sequence is under \cite{12572619,1806330,9402741}.
Despite a diverse distribution of pseudogenes, we find a consistent inter-chromosomal traffic for the sex chromosomes across all organisms'.

Finally, we identify a large spectrum of biochemical activity (as defined by transcription, active chromatin, POL2 and transcription factors) for the pseudogenes ranging from "highly active" to "dead". 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.

Furthermore we relate the activity of pseudogenes to the conservation of their upstream region. Comparing the pseudogenes and functioning paralogs, surprisingly we find a shift in the distribution of upstream region conservation, with many pseudogenes being even more conserved than paralogs. 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 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 repeats driven genesis, may contribute to the species divergence due to their high organisms specificity.

Our functional analysis shows that pseudogenes importance bypasses their common use as false positives in coding annotation, suggesting that pseudogenes may play active roles in the genome biology.

MORE

Cristina Sisu 4/2/14 14:44
Moved (insertion) [2]

Cristina Sisu 4/2/14 14:44
Deleted: The distribution of pseudogenes in the three model organisms is indicative of their respective specie evolution. The scarcity of pseudogenes in the fly genome is reflective of its high rate of DNA loss \cite{12572619,1806330,9402741}, an intrinsic characteristic of its large effective population size \cite{12572619,9501496,14631042}.Consequence of a high deletion rate, the fly pseudogenes are enriched in disabling deletions. By contrast, a rather small population size and recent genome wide selective sweeps \cite{22286215} resulted in a fixation of non-adaptive mutations \cite{17637734} in the worm genome. As such worm has a larger pseudogene complement than fly. Also the high rate of recombination in worm, led to the removal of deletion rich pseudogenes and as such we found an enrichment of disabling insertions in the pseudogenic regions. In zebrafish numerous intra-chromosomal and tandem duplications \cite{22702965} and high recombination rates depleted the pseudogene complement. ... [10]

Cristina Sisu 4/2/14 14:44
Deleted: repeat element

Cristina Sisu 4/2/14 14:44
Deleted: . 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).

Cristina Sisu 4/2/14 14:44
Deleted: 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 reflecting perhaps a consistent degradation rate for active genomi ... [11]

Cristina Sisu 4/2/14 14:44
Moved up [2]:

Cristina Sisu 4/2/14 14:44
Deleted: , our

Cristina Sisu 4/2/14 14:46
Deleted:

Cristina Sisu 4/2/14 14:45
Deleted: g

Cristina Sisu 4/2/14 14:44
Deleted: interesting regulatory rol ... [12]