Comparative Analysis of Pseudogenes: History Trumps Conservation…

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

Abstract

W

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 each organism's evolution.

W MAM REP

At a first glance we notice significant differences in the pseudogene diversity between the studied species. As such we find that the human pseudogene complement is governed 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 the parallel evolutionary history of the two organisms. By contrast the fly pseudogenes are the product of its large population size and thus are dominated by numerous disabling deletions. The worm pseudogene complement is shaped by large duplication events associated with particular environmental response gene families. A similar pattern of multiple tandem duplications combined with high recombination rates depleted the pseudogene complement in zebrafish.

NO

Despite large variations in the pseudogene complement of the four species, we also find some notable similarities. To this end, we observe a consistent inter-chromosomal pseudogene exchange for the sex chromosomes. Also 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 species implying a uniform degradation mechanism of functional elements. Furthermore we see a consistent trend in terms of the decline in the pseudogene promoters' activity relative to their coding counterparts. However, we also find a small population of pseudogenes with highly conserved upstream sequences and activity hinting at potential regulatory roles. We complete our analysis ranking the pseudogenes based on their activity features and pinpointing potentially functional candidates.

N

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}.

Cristina Sisu 9/2/14 22:43
Deleted: the organisms'

Cristina Sisu 9/2/14 22:43
Deleted: [[move sex chr to act and see similarities]] ... [1]

Cristina Sisu 9/2/14 22:43
Deleted: marked

Cristina Sisu 9/2/14 22:43
Deleted: a highly correlated

Cristina Sisu 9/2/14 22:43
Deleted: .

Cristina Sisu 9/2/14 22:43
Deleted: governed

Cristina Sisu 9/2/14 22:43
Deleted: dominated

Cristina Sisu 9/2/14 22:43
Deleted: responses

Cristina Sisu 9/2/14 22:43
Deleted: Despite a diverse distribution of pseudogenes, we find a consistent inter-chromosomal traffic for the sex chromosomes across all organisms.

Cristina Sisu 9/2/14 22:43
Deleted: [[move sex chr to act and see similarities]] ... [2]

Cristina Sisu 9/2/14 22:43
Deleted: four

Cristina Sisu 9/2/14 22:43
Deleted: while

Cristina Sisu 9/2/14 22:43
Deleted: fall off of

Cristina Sisu 9/2/14 22:43
Deleted: pseudogene' upstream histone marks

Cristina Sisu 9/2/14 22:43
Deleted: the

Cristina Sisu 9/2/14 22:43
Deleted: genes

Cristina Sisu 9/2/14 22:43
Deleted: a

Cristina Sisu 9/2/14 22:43
Deleted: sequence

Previously, pseudogenes have been characterized within individual genomes \cite{17099229,22951037,11160906,12560500,15860774,12083509,16925835}. The absence of a finished annotation and the potential of mis-mapping of functional genomics data had previously rendered a detailed whole genome comparison of the pseudogene complement in various organisms rather inaccurate. As such, earlier studies were restricted to the comparison of a specific family or class of pseudogenes \cite{15289607,16469101,12417195,16680195, 19835609,12034841,19123937,23555032}. 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.

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. [[m6: focus on finished, mention some perspective of draft for closer species]]

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 ~~rigidly~~ 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).

PROCESS ANNOT.

Cristina Sisu 9/2/14 22:43
Deleted: Earlier comparative studies across various organisms have focused on a specific family or class of pseudogenes \cite{15289607,16469101,12417195,16680195,19835609,12034841,19123937,23555032}.

Cristina Sisu 9/2/14 22:43
Deleted: solid

Cristina Sisu 9/2/14 22:43
Deleted: [[m6: whole concept, family]]

Cristina Sisu 9/2/14 22:43
Deleted: We use the pseudogenes as fingerprints of genome history to find key steps in these species evolution.

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.

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 studied genomes (Fig XXX, SXXX). Overall we found large discrepancies between the four species. For example, in human, the pseudogene distribution follows closely the chromosome size but it is only weakly correlated with the protein coding genes frequency suggesting the existence of pseudogene inter-chromosomal transfers. By contrast in worm and fly we observed a strong correspondence DUPL between the pseudogenes and protein coding genes density, while in zebrafish we found no correlation at all. ~~To shed light on the peculiarities of pseudogene localization~~, W we analysed the recombination rate at pseudogene loci for each species (Fig XXX). We found that the human, fly and zebrafish pseudogenes are enriched in regions of low recombination and thus are preferentially located near the centromere and in particular on the sex chromosomes. However in worm we observed a rather uniform recombination rate for genes and pseudogenes, a possible consequence of recent selective sweeps that pruned its genome. As such, the majority of worm pseudogenes are found near the telomeres, regions characterized by high recombination rates and rapid gene evolution \cite{8536965}.

DISCUSS MEET PT

Cristina Sisu 9/2/14 22:43
Deleted: four genomes.

Further we looked at the pseudogene transfer between the chromosomes. While the processed pseudogenes are easily exchanged, evidence of their random distribution across the genome, the duplicated pseudogenes have low mobility, commonly residing on the same chromosome as their parent genes. This co-residence is notable for human chromosomes 7 and 11, due to their enrichment in genome duplication events \cite{12853948} and olfactory receptors respectively \cite{11337468}.The co-localization is also highly significant for the sex chromosomes (human Y, fly X), where, consequence of a low recombination rate \cite{16545149,1875027,15059993}, the pseudogenes cannot be "crossed out".

Cristina Sisu 9/2/14 22:43
Deleted: *(a) Chromosomal Distribu* ... [3]

Even more, as a result of this low recombination rate, we observed a large accumulation of "imported" processed and duplicated pseudogenes on human X and Y, chromosomes respectively. ~~While X pseudogene import has been previously reported~~ \cite{14739461}, the duplicated pseudogene import from X observed on Y can be explained regarding the pseudogenes as "degenerated 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}.

NO

Cristina Sisu 9/2/14 22:43
Deleted: 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 ~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.

Cristina Sisu 9/2/14 22:43
Deleted: the pseudogenes originating from

?

(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.

LATER

FAM

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). [[m6: parent expression vs. duplicated and processed, parent expression vs. number of processed pseudogenes – BP to do]]

KIBO

(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 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.

W JUSTIN HUM.

MG

(c) Upstream sequence similarity and promoter activity

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

Cristina Sisu 9/2/14 22:43
Deleted: *Translation* ... [4]

Cristina Sisu 9/2/14 22:43
Moved down [1]: 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.

Cristina Sisu 9/2/14 22:43
Deleted: *(d)*

selection. These scenarios could lead to a coordinated expression pattern between the transcriptional products regulated by these promoter regions.

To this end we analysed the ChIP-seq data of H3K27ac, an important marker in defining active promoters and enhancers. We focused our study on protein coding genes with only one pseudogene but no paralogs, and those with one pseudogene and one paralog. We observed that in general, while the pseudogenes have highly conserved promoter regions, the activity is less preserved when compared to their protein coding gene counterparts (Fig XXXGRIDplots).

“Functional” Pseudogene Candidates

TOOK OUR LIST & ...

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

First, using mass spectrometry data, we studied the translation potential of transcribed human pseudogenes in four ENCODE cell lines. From over 14000 pseudogenes we identified three pseudogenes with high translation evidence (Table YZ1). The low number of translation candidates is indicative of the high quality of our annotation and gives us confidence in their authenticity. 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.

Next, focusing on the regulatory potential, we curated a list of putative “functional” candidates using activity, coexpression and sequence conservation data (Table SXXX). We obtained a set of 10 “functional” candidates (highly active, with a high sequence similarity to parents and with a high parent/pseudogene coexpression correlation coefficient) including the known regulatory pseudogene PTEN-P1.

Using PTEN-P1 (a cancer pseudogene) as lead example, we found 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 some pseudogenes may play functional roles.

FIG 2.

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.

Cristina Sisu 9/2/14 22:43
Deleted: ... [5]

Cristina Sisu 9/2/14 22:43
Deleted:

Cristina Sisu 9/2/14 22:43
Deleted: 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 [[m6: rewording]] human pseudogenes (highly active, with a high sequence similarity to parents and a high coexpression correlation coefficient) including the known regulatory pseudogene PTEN-P1.

Cristina Sisu 9/2/14 22:43
Deleted: With the example of PTEN-P1 in mind, we also investigated the human pseudogenes of cancer related genes. We observed

Cristina Sisu 9/2/14 22:43
Moved (insertion) [1]

Cristina Sisu 9/2/14 22:43
Deleted: other

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, centromere regions. It also resulted in a preponderance of pseudogenes associated with highly transcribed proteins such as those in pathways of central metabolism and the ribosomal proteins. Also, while the burst of retrotransposition events happened after the human/mouse speciation (~90 MYa), the high occurrence of processed pseudogenes in the mouse genome suggests that this event occurred on a much larger scale and it can be regarded as an 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 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}.

MOST PART ACT.

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.

Cristina Sisu 9/2/14 22:43

Deleted: Despite a diverse distribution of pseudogenes, we find a consistent inter-chromosomal traffic for the sex chromosomes across all organisms'.

FN

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.

Cristina Sisu 9/2/14 22:43

Deleted:

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 suggests that pseudogenes may play active roles in the genome biology.

Future directions for mouse

- finish annotation
- check for upstream region degradation - is the conservation of the sequence human/primate specific or general for all mammals
- check if the upstream sequence activity decays at the same rate as human (most probably)
- human/mouse orhtologs (so far only 1% - look at the family)
-

Cristina Sisu 9/2/14 22:43
Deleted: shows that pseudogenes importance bypasses their common use as false positives in coding annotation, suggesting

Cristina Sisu 10/2/14 04:50
Formatted: Bulleted + Level: 1 + Aligned at: 0.63 cm + Indent at: 1.27 cm

Cristina Sisu 10/2/14 04:52
Formatted: Font color: Accent 6