# Response to reviewers for “Comparison of Pseudogenes Across 3 Phyla”

# Response Letter

### -- Ref1 – General Remarks --

|  |  |
| --- | --- |
| ReviewerComment | “Comparative analysis of pseudogenes across three phyla” presents impressive extension of the ENCODE project into important area of pseudogenes. The authors mapped pseudogenes in human, worm and fly genomes (as well as macaque, mouse, and zebrafish) and studied patterns of pseudogenes structural organization and distribution across the genomes. Results of this comprehensive analysis indicated significant lineage specific differences in pseudogene dominant structures and distribution patterns; these differences were attributed to differences in the driving forces directing the lineages evolution. The manuscript is very well written.The results are of significant importance for genomics and molecular evolution. This work is certainly of interest for PNAS as it belongs to category of top research advances. |
| AuthorResponse | We thank the reviewer for the constructive comments. We address them in a point by point manner below. |
| Excerpt FromRevised Manuscript |  |

### -- Ref1.1 – Pseudogene Annotation Variance --

|  |  |
| --- | --- |
| ReviewerComment | The authors give (Fig. 1) specific numbers of pseudogenes that were characterized in each of the three genomes: 12,358 in human, 911 in worm, 145 in fly. Since there could be a margin of error – it would be useful to discuss this margin in Discussion. |
| AuthorResponse | The pseudogene annotation accuracy is always dependent on the accuracy of the coding gene annotation. As such looking at variations between annotation releases we can see significant differences between the total number of pseudogenes (see Fig. S1). However, given that the pseudogene sets described here are based on the finished manual annotations of human, worm, and fly genomes, we do not expect to see any statistically significant differences in the future genome releases. Also all three genome pseudogene datasets (worm, fly, and human) used the same methodology, quality of evidence and analysis pipelines for manual annotation to remove inconsistency between the datasets. Also in our previous paper on an incomplete version of the human annotation [1] we discussed thoroughly the difficulties encountered in the pseudogene annotation as well as potential sources of error. We have added to the discussion a mention of this earlier work to give the reader a reference point in understanding the potential uncertainties in the total number of pseudogenes. |
| Excerpt FromRevised Manuscript | We report a multi-organism comparison of pseudogenes leveraging the finished annotations of the genomes of human, worm, and fly. Given that these are finished annotations we do not expect to see any significant changes in the total number of pseudogenes in the future. (For a detailed discussion of the variance in gene and pseudogene counts over draft annotation releases see Fig. S1, Supplementary Information in [4,26]). |

### -- Ref1.2 – Pseudogene Numbers --

|  |  |
| --- | --- |
| ReviewerComment | For instance, the distribution of worm processed pseudogenes by similarity to parent genes reaches max at 50% - chosen as the cut-off (Fig 1B); it is natural to expect dozens of processed pseudogenes with similarity at 40% still detectable level, but obviously they are not included in 911. |
| AuthorResponse | Regarding Fig. 1B, we thank the reviewer for pointing out our misleading labels on the “x”-axis. Given that the number of pseudogenes with a sequence similarity to parents less than 50% is very low, we aggregated all these cases in the 50% sequence similarity bin. We updated all the figures to explicitly show this fact.  |
| Excerpt FromRevised Manuscript | Fig. 1B and Fig. S2 update. |

### -- Ref1.3 – Duplicated Pseudogenes Timeline --

|  |  |
| --- | --- |
| ReviewerComment | The timeline is analyzed & discussed only for processed pseudogenes. The duplicated pseudogenes – more than 2200 of them in human – have not been analyzed in terms of timeline (at least the data were not shown). Are there clear reasons for this void? |
| AuthorResponse | We included a figure in the supplementary material showing the distribution of pseudogenes at various ages. We also addressed the variation of duplicated pseudogenes as function of time in the main text.  |
| Excerpt FromRevised Manuscript | We observe that the distribution of duplicated pseudogenes shows little variation with age (Fig. S2). However, the creation of processed pseudogenes varies very much over time (Fig. 1B) |

### -- Ref1.4 – Pseudogene Biotype Classification --

|  |  |
| --- | --- |
| ReviewerComment | In Classification section (p.5) the pseudogenes are classified in three groups – and then the ones of the third type – unitary pseudogenes – are not mentioned any more. Any comments why? |
| AuthorResponse | There are 151 unitary pseudogenes in the human genome, comprising about 1% of the total number of pseudogenes. These pseudogenes were included in all the analysis that did not require a biotype differentiation or the presence of a parent gene. As such, chromosomal distribution of unitary pseudogenes is not statistically significant due to the low number of unitary pseudogenes compared to processed or duplicated ones. Similarly, upstream sequence comparison between unitary pseudogenes and parent genes is not possible due to the fact that unitary pseudogenes do not have functioning counterparts in same organism.  |
| Excerpt FromRevised Manuscript |  |

### -- Ref1.5 – Pseudogene Annotation --

|  |  |
| --- | --- |
| ReviewerComment | Definition of a human gene model as pseudogene relies on several clearly specified features (Suppl. Methods p.1) However, in some functional human genes read-through stop codons code for celenocystein; also some known genes with “recoding” have programmed frameshifts; these exceptions makes sense to mention. |
| AuthorResponse | Yes, these exceptions are taken into account in the manual annotation process. If a model has evidence (transcriptional, functional, publication) showing that the locus represents a protein-coding gene with structural/functional divergence from its parent (paralog), it is discarded from the pseudogene dataset. In addition we run selenoprofiles supplied by Roderic Guigo's group, which characterises the selenoproteome and provides accurate predictions of selenoprotein genes [2].We have added a small amount of text to the supplementary methods to make clear that the HAVANA manual annotation pipeline does check for the read-through stop codons coding for selenocysteine and for known genes with programmed frameshifts. |
| Excerpt FromRevised Manuscript | We note that the manual annotation pipeline checks the possibility than any putative pseudogene might instead be a protein-coding gene. If any putative pseudogene locus has transcriptional, functional or publication evidence to support coding potential, including selenocysteine incorporation, stop-codon read-through and programmed frameshift events, it is excluded from the set of pseudogene transcripts.  |

### -- Ref1.6 – Functional Pseudogenes --

|  |  |
| --- | --- |
| ReviewerComment | The concept of “functional” pseudogenes – p.8 (transcribed or even translated) is controversial, almost oxymoron-like concept. Should the concept of a gene be further generalized – a revision that the corresponding author already initiated in a previous ENCODE publication? |
| AuthorResponse | We agree with the reviewer that the concept of “functional” pseudogene is counterintuitive, and we do realize the necessity to revisit the definition of the notions of pseudogene and gene. We have written a number of reviews on the definition of gene and pseudogene [3, 4, 5] and we also put in a call for future perspective on this matter. However we believe that the present paper is not the most suitable place to address such a complex issue. In this manuscript we refine the pseudogene dataset to a list of so called “functional” candidates, that comprises of a number of interesting pseudogenes that have characteristics, commonly attributed to functional coding genes: e.g. have transcription/translation evidence, are located in regions of active chromatin, and/or have active upstream sequences. We consider that these interesting scenarios are worth of further analyses that can potential result in reclassifying them in other annotation categories (e.g. coding gene, ncRNA, etc.). We added a sentence in the text to define the term “functional”-candidate. |
| Excerpt FromRevised Manuscript | **Results****[…]** **“Functional” Pseudogene Candidates** Finally, combining the annotation, functional genomics and evolutionary data, we refine the active pseudogene group to a set of “functional” candidates. This term refers to a pseudogene that possesses numerous signs of activity, commonly attributed to canonical coding genes (e.g. transcription, translation, active chromatin, etc.).**Discussion****[…]**Hence we suggest that some pseudogenes may play active roles in genome biology and warrant further experimental investigation. We realize the notion of “functional pseudogene” is, in a sense, an oxymoron. However, here we focus only on tabulating and enumerating these potential functional candidates. In light of recent advances in functional genomics and genome biology, it may be useful to revisit the definition of gene and pseudogene to better and more accurately describe these entities. |

### -- Ref1.7 – Pseudogene Localization --

|  |  |
| --- | --- |
| ReviewerComment | In Table S1 results for human Chromosomes 12, 14 and 19 – appear to be contra intuitive (significance – True, when P-value is 1.00) |
| AuthorResponse | We thank the reviewer for noticing our mistake. We corrected the wrong p-values and added a more detailed explanation to the binomial test in the supplementary information.  |
| Excerpt FromRevised Manuscript | We used a two hypotheses binomial test to evaluate the statistical significance of the difference in the pseudogene frequency between the telomeric and the centromeric regions (Table S1). The first hypothesis is that the pseudogenes are equally distributed at the centromeric and telomeric regions. The second hypothesis describes the observed distribution of pseudogenes at in the centromeric and telomeric regions. As such, there are two options: “\*” – the centromere has more pseudogenes than the telomere; and “#” – the telomere has more pseudogenes than the centromere. The significance threshold p-value was set to 0.05.[…] Updated Table S1.1 |

### -- Ref1.8 – Typos --

|  |  |
| --- | --- |
| ReviewerComment | – Fig S3 – has typo in the labelsp.7 of Suppl. Methods – typo in RPKM |
| AuthorResponse | We thank the reviewer for pointing out these mistakes. The typos have been corrected in the revised version. |
| Excerpt FromRevised Manuscript |  |

### -- Ref2 – General Remarks --

|  |  |
| --- | --- |
| ReviewerComment | This paper describes a comparative analysis of pseudogenes in several organisms, with emphasis on human, fly, and worm, in light of ENCODE functional genomics data. The analysis is accompanied with a very useful online pseudogene resource, psicube.pseudogenes.org. Interestingly, the authors find that mammalian pseudogenes are more frequent in genes that are highly transcribed, suggesting retrotransposition resulting processed pseudogenes as the main source, while worm and fly pseudogenes are mostly due to duplication. Another interesting analysis involves the categorization of human pseudogenes as "active" and "very active", in light of ENCODE data, and the identification of a subset of human pseudogenes with conserved promoter activity and sequence with respect to the parent genes, suggesting some functional role for these pseudogenes. Overall, the paper is well written, the analysis is thorough and interesting, the online resource is very useful, and this is a paper of great value to the genomics community.  |
| AuthorResponse | We thank the reviewer for the comments.  |
| Excerpt FromRevised Manuscript |  |

**References**

1. Pei, B. et al. (2012) The GENCODE pseudogene resource. *Genome Biol* **13**:R51.2. Mariotti, M. & Guigó, R. (2010) Selenoprofiles: profile-based scanning of eukaryotic genome sequences for selenoprotein genes. *Bioinformatics* **26**:2656-63.3. Zheng, D. & Gerstein, M. B. (2007) The ambiguous boundary between genes and pseudogenes: the dead rise up, or do they? *Trends Genet* **23**:219-24.4. Snyder, M. & Gerstein, M. (2003) Genomics. Defining genes in the genomics era. *Science* **300**:258-60.5. Sasidharan, R. & Gerstein, M. (2008) Genomics: protein fossils live on as RNA. *Nature* **453**:729--731.