Dra. Kateryna Makova
Editorial Board Member
PCI Evolutionary Biology

Dear Dra. Makova,

First, I would like to apologize for all the time that takes us to send the revised version of our manuscript. We appreciate the comments from both reviewers. We feel that after this round of suggestions, our manuscript has been improved. We hope you agree that we have adequately addressed all of the reviewers' issues and find that our manuscript is now suitable for a recommendation in PCI Evolutionary Biology.

Review for "Evolution of the DAN gene family in vertebrates"
In this manuscript by Opazo et al., the authors use homology searches to identify genes from the DAN gene family (Differential screening-selected gene Aberrant in Neuroblastoma) across chordate lineages. The phylogenetic relationships of these genes were inferred and the toplogy of the resulting tree was used to describe the evolutionary history of the gene family.

Interestingly, the authors identify a new family member related to the Gremlin genes, which they dub Grem3. Next, in the Gnathostome lineage, the authors show evidence for five genes being present in its MRCA that are also widely retained across its descendents (e.g. the major Gnathosome lineages listed in figure 4). These genes include Grem1, Grem2, SOST, SOSTDC1, and NBL1. The authors also identify 3 gene family members that they conclude are likely in the gnathostome ancestor, but have experienced loss in some of the ancestors: Grem3, Cer1, and DAND5.

Over all, the manuscript is well-written and lays out its case fairly well. And for the most part, I find the major arguments to be reasonable. However, there are a lot of areas that I feel would benefit from feedback described here.

Major comments

1. The methods are insufficiently detailed to permit the work to be repeated
The authors do not define the pool of sequences from which query and subject sequences are drawn. The specific implementation of blast and its version isn't cited. The filtering criteria used to determine whether hits are retained or discarded are not documented.

We thank the reviewer for the comment. In response to it, we rewrote this section making emphasis that the genomic pieces were identified based on conserved synteny. We also added more details about software and database versions used. We now state the version of the blast program and of the ensembl database we used at the time we curated the existing annotation of the DAN genes. We used a combination of sequences to curate the existing annotation, the general criterion was to use sequences from species that share a common ancestor more recently in time to the species of which the genomic piece (subject sequence) is being analyzed. For example, we used the chicken sequences to curate sauropsid sequences. It is important to mention that all sequences used in our study will be available to anyone.

2. The nature of the multiple alignment wasn't described. How much of the genes were alignable at the greatest divergences? In the introduction, the authors claim that there is "low inter-parallog conservation", indicating that the alignment may not be reliable in many regions. What was aligned? Nucleotides or amino acids (I assume amino acids)?

   We think that our alignment is reliable, i.e. possesses valuable phylogenetic information. Proof of this is that we recovered the monophyly of all paralogs described for the gene family, further, we also recovered the monophyly of the group of paralogs that we already know share a common ancestor more recently in time among them than with any other family member (e.g. SOST and SOSTDC1; GREM1, GREM2 and GREM3). As the reviewer mentioned, we aligned amino acid sequences. We emphasize this in the current version of the text, we included a description of the alignment.

3. The results are fairly sparse on details
   For example, display items aren't thoroughly described. The captions are very terse. For example, there appears to be a convention in the synteny plots where the absence of a bar indicates the absence of the gene (ag CER1 in Spotted Gar in Figure 2B). However, in Figures 5 and 6, dotted lines apparently indicate missing DAN genes but missing bars for flanking genes means that the gene isn't in the syntenic region.

   The reviewer brings up an excellent point. Following this suggestion, we added information to all figure legends.

4. What is the scale in Figure 1? A bar with the number "0.7" is included. The caption doesn't elaborate.

   The bar denotes the number of amino acid substitutions per site. This information was added to the figure caption.

5. I'm accustomed to bootstrap support to be reported in Numerator/Denominator or explicitly in %. The numbers corresponding to bootstrap support in Figure 1 are just bare integers.

   This is an interesting point. We typically include node support values in this way to save space in the figure. The figure legend indicates what each of the support values is. We believe it is understandable in this way.

6. The authors often point out disagreements with the literature, which is commendable. However, little effort is made to reconcile these observed disagreements. I'd feel better if the authors would discuss the discrepancies they point out.
   Examples:
   "Although the study of Walsh et al. (2010) supports..., two other studies report alternative topologies."Nolan et al. (2014) recovered NBL1 as sister... However, in support of our study Avsian-Kretchmer et al. (2004) recovered NBL1 as sister to the GREM lineages."
   "However, in contrast to Petillon et al. (2013), we did not find..."

   Again, the reviewer's concerns bring an excellent point. In most of the cases the differences are because the phylogenetic trees reported are built with a very limited

taxonomic sampling, in some cases using only the human paralogs (e.g. Avsian-Kretchmer et al. 2004). In other cases, reported phylogenetic trees do not include sufficient details about how they were inferred (e.g. Walsh et al. 2014). We feel it is unfair to question so much studies done with less data when that was all the available data at the time. We mention this just to provide a rationale for the motivation of our study. Basically, the reason we did this study is exactly that: not much attention has been paid to the sister group relationships among DAN paralogs, and the few studies on the matter did not reach a consensus. To improve our work we added a paragraph explaining this situation at the end of the gene phylogeny section.

7. The claim of "recovering monophyly" is confusing to me.
   "Our results recovered monophyly of all DAN gene family members"
   My parsing of this statement in the abstract (and others like it throughout the manuscript) is probably not what the authors intended. To me, this sounds like "we confirmed that, as a group, all DAN genes are monophyletic".

   Good point. We have now rephrase this sentence to read as follows:

   "Our phylogenetic analyses recovered the monophyly of each member of the genes in the DAN gene family with strong support (Fig. 1).

8. This doesn't make sense in an analysis where the recovery of a gene from EnsEMBL is viewed as conferring DAN membership on that gene.

   This is a good point. Orthology and paralogy assignments from ensembl are very good first passes, but there are many cases in which paralogous genes are annotated as orthologs because of reciprocal losses, and other cases when distant paralogs are assigned to independent gene families despite their genetic affinities. In addition, in many cases gene annotations reflect presumed functional similarities rather than evolutionary relationships, such as the myoglobin and hemoglobin genes from cyclostomes and gnathostomes. Thus just retrieving the sequence from a database is again not enough to confer a membership on a specific gene lineage of the family. Because of this, our phylogenetic analyses played a fundamental role in our understanding of homologous relationships.

9. So, by definition, every gene in the analysis is DAN, and with no non-DAN genes for contrast, no determination about monophyly can be made.

   We agree with this comment, in our study we are not testing the monophyly of the DAN gene family, consequently we have removed these statements from the manuscript.

10. While I can't confidently interpolate what the authors actually meant, perhaps the following is closer to the authors' meaning: "For each member of the gene family (e.g. CER1, SOST, SOSTDC1, DAND5, NBL1, GREM1, GREM2, and a new member, GREM5), the group of species sequences corresponding to each gene is monophyletic."Even this formulation is a bit confusing to me, as the monophyly seems to be how the authors would assign a particular sequence in a particular species to particular family member.

We agree with the reviewer that trying to explain in a non standard way could induce confusion to the reader, consequently, we would like to maintain the evolutionary jargon we are actually using in the text.

11. And in any event, this gets a bit muddied when there is gene duplication. *What* is monophyly when for some taxa, there are duplicates, and others, there aren't? Is "recovery of monophyly" a result as implied by the authors? Or rather is it part of how the authors are classifying the sequences into family members like CER1, etc.? Perhaps this "recovery of monophyly" could be reconciled if the authors inferred the full duplication history with synteny for every species they examined and then layered the phylogenetic analysis of the gene family on top of that. But, as far as I can tell, this was not the strategy the authors followed in most cases.

    We agree with this comment, what we are doing is to classify sequences into family members.

12. Finally, DAND5 doesn't appear to offer strong support for monophyly given that lack of support for placing the Coelacanth as sister to the other DAND5 genes. The strong synteny argument doesn't change this assessment, as it could be a brute fact that the Coelacanth sequence is simultaneously the DAND5 ortholog and there is no strong evidence of monophyly with the remaining DAND5 orthologs.

    We agree with this argument, and it was the reason why we retrieve synteny information. Thus, given that our synteny conservation analysis points out in the same direction as our maximum likelihood best tree, we propose that the most likely scenario is that the coelacanth sequence is the ortholog of all other sequences in the clade.

13. One comment relating to paralogy confused me.
    "The fourth clade corresponds to the NBL1 gene, the founding member of the DAN gene family, and was recovered as monophyletic with strong support (pink clade; Fig. 1)."
    This way of discussing paralogy (ie "founding member") seems clumsy to me. Barring clear mechanistic reasons to assign one paralog the label "founder" or "parent" (e.g. the template for the RNA in retrogenes or the copy maintaining the ancestral structure in a chimeric duplicate), immediately after duplication, the copies are provisionally assumed to be redundant. And as such, it would only be confusing to label one member the "founding member". The authors even discuss this in relation to the putative redundancy between DAND5 and CER1.

    The founding member is not a label assigned based on our analyses, it is called this way because it was the first member to be discovered. To avoid confusion we removed that statement from the text.

14. The discussion of cancer on pages 14 and 15 isn't well-integrated into the rest of the manuscript. The reference to RPRM and p53 in particular seems like it could be better

incorporated into the narrative of the manuscript. Personally, I'd recommend dropping it, but a smoother integration could also work.

We agree with the author, accordingly we removed this paragraph.

15. In a manuscript like this one, I would like to see more in depth discussion of sources of error. The task the authors set before themselves is quite ambitious and requires marshaling a lot of data from many genes across many different taxa. These taxa were sequenced by different groups, at different times, with different technology, exhibit different levels of contiguity and likely accuracy and completeness, etc. Sources of error can include errors in multiple alignment, misannotation of the genes, and evolution in gene structure, all of which can lead to aligned non-homologous residues. Moreover, low assembly or annotation completeness can lead to missing genes.

We agree with the reviewer, there are sources of error, as in all types of analyses. In this study, we tried to minimize the primary source of error, the sequences, by curating all annotations of the sequences. It was a lot of work, but we believe it was worth it. To highlight what the reviewer is mentioning we included more details in the methods section.

Minor comments

1. Why use the common name "elephant fish" when there is an "elephant fish" in both Actinopterygii and Chondrichthyes? Perhaps "elephant shark" would be better?

   We agree with this comment, accordingly we changed the name

2. Why didn't the authors use Rhincodon typus (whale shark: https://www.ncbi.nlm.nih.gov/assembly/GCF_001642345.1/) in the analysis? It has a Genbank annotation and appears to be more contiguous than the elephant shark. Also, since this manuscript was posted, there is now a much better Chondrichthyes genome (Pristis pectinata): https://www.ncbi.nlm.nih.gov/assembly/?term=Pristis+pectinata Perhaps either of these two could be valuable in the analysis.

   The reason is because when we performed this study these species were not available. Given that we are most interested in the phylogenetic representativeness of the species, and not in the species itself we believe is not worth it to perform all analyses again.

3. Typo of DAND5: DADN5

   Fixed

4. Perhaps a labeled, high-level phylogeny would be useful in orienting the readers. One like Figure 1 in this would be a great service to the reader: http://dx.doi.org/10.1016/j.cub.2017.02.029

We think it is a good idea. We included a figure showing the main groups of deuterostomes as a supplementary figure.

5. Is Urochordate / Urochordata still in common use?

We think so.

Reviewed by anonymous reviewer, 2019-12-10 03:55

This is a nice reconstruction of the evolution of a complex gene family, the DAN gene family. The authors show strong supporting evidence for the monophyly of 5 major groups and the inter-group relationships among them. While it is useful to see the information about this gene family all together, the novelty of this study is unclear as the authors often refer to previous literature that shows comparable, albeit partial, results.

We think that the main novelty is that our study is solely devoted to understanding the duplicative history of the DAN gene family. In the literature, almost all studies in which you see a phylogenetic tree of the DAN gene family, it represents a "secondary product" of the study, in which the taxonomic and paralog sampling could be importantly improved.

Minor comments:

1. the authors should provide more information about the alignments produced (length, % gaps).

We included in the method section a statement in which we described our alignment.

2. the authors used an evaluation of likelihood scores to determine convergence of the bayesian phylogenetic reconstruction. Although I generally agree with the authors that this method should produce accurate results, most researchers rely on the estimation of ESS values to determine convergence. It would be useful to know how the ESS values correlate with the number of generations required to reach an asymptotic trend in likelihood scores.

Following the reviewer's suggestion, we estimated our EES value using the software Tracer v1.7.1. Our result shows that the EES value of our MrBayes analyses (216.5) was above the recommended one (200). We have included the corresponding statement in the DNA data and phylogenetic analyses section.

3. At the end of the page with the section entitled "Definition of ancestral gene repertoires" the authors state that the "lack of DAND5 in the elephant fish is an artifact of the current genome assembly". Please provide an explanation for this statement.

It is difficult to explain why we did not find a gene in a genome that is still work in progress. To avoid confusion we decided to remove that sentence.

4.figure 2 and 3: what is the meaning of the double lines associated to some genes?

We are not sure what double lines are the reviewer referring to, as far as we see, there are no double lines in figures 2 and 3.

Also, the grey lines represent intervening genes but no information is provided on how large these intervening sections of DNA may be. Depending on the size, they could be affecting the definition of synteny so more information is necessary to support the conclusions based on synteny.

We understand the reviewer's concern. We believe that more in-depth analyses of the synteny pattern would be appropriate if there is a disagreement with the phylogenetic reconstruction. However, given that both analyses are pointing in the same direction, we think it is not necessary.