Peer Community In Evolutionary Biology

Combining molecular information on chromatin organisation with eQTLs and evolutionary conservation provides strong candidates for the evolution of gene regulation in mammalian brains

Marc Robinson-Rechavi *based on peer reviews by* ***Marc Robinson-Rechavi*** *and* ***Charles Danko***

Francisco J. Novo (2017) Evolutionary analysis of candidate non-coding elements regulating neurodevelopmental genes in vertebrates. Missing preprint_server, ver. Missing article_version, peer-reviewed and recommended by Peer Community in Evolutionary Biology. https://doi.org/10.1101/150482

Submitted: 29 June 2017, Recommended: 06 October 2017

Cite this recommendation as:
Robinson-Rechavi, M. (2017) Combining molecular information on chromatin organisation with eQTLs and evolutionary conservation provides strong candidates for the evolution of gene regulation in mammalian brains. *Peer Community in Evolutionary Biology*, 100035. https://doi.org/10.24072/pci.evolbiol.100035

Published: 06 October 2017

In this manuscript [1], Francisco J. Novo proposes candidate non-coding genomic elements regulating neurodevelopmental genes. What is very nice about this study is the way in which public molecular data, including physical interaction data, is used to leverage recent advances in our understanding to molecular mechanisms of gene regulation in an evolutionary context. More specifically, evolutionarily conserved non coding sequences are combined with enhancers from the FANTOM5 project, DNAse hypersensitive sites, chromatin segmentation, ChIP-seq of transcription factors and of p300, gene expression and eQTLs from GTEx, and physical interactions from several Hi-C datasets. The candidate regulatory regions thus identified are linked to candidate regulated genes, and the author shows their potential implication in brain development. While the results are focused on a small number of genes, this allows to verify features of these candidates in great detail. This study shows how functional genomics is increasingly allowing us to fulfill the promises of Evo-Devo: understanding the molecular mechanisms of conservation and differences in morphology.

References:

[1] Novo, FJ. 2017. Evolutionary analysis of candidate non-coding elements regulating neurodevelopmental genes in vertebrates. bioRxiv, 150482, ver. 4 of Sept 29th, 2017. doi: [10.1101/150482](https://doi.org/10.1101/150482)

Reviews

Evaluation round #2

Reviewed by Charles Danko, 22 September 2017

Francisco Novo appears to have made changes in his manuscript that address comments raised during my first review. I am happy to recommend his manuscript, which I believe will be of interest to reviewers in the field.

Reviewed by Marc Robinson-Rechavi, 28 September 2017

The revised manuscript has taken all remarks into account. Notably, the revised title, abstract and discussion are much clearer and reflect better the results.

Evaluation round #1

DOI or URL of the preprint: **10.1101/150482**
Version of the preprint: 2

Authors' reply, 31 August 2017

Dear Dr. Robinson-Rechavi,

I would like to thank you and Dr. Danko for your comments and suggestions in the reviews of this manuscript. In the following paragraphs I answer those points and explain how I have incorporated them in the new version of the manuscript that has been uploaded to bioRxiv. I hope this new version can be considered suitable for recommendation by PCI Evol Biol.

1. As a general remark about the extent and goals of this work, it is important to understand that it represents an attempt to gain functional knowledge about conserved putative neurodevelopmental regulatory elements by accruing information from a variety of computational and experimental datasets. This strategy ranks non-coding elements according to the likelihood that they behave as regulatory elements of particular gene(s) in specific tissues. It is only partially true that the selection of these regions was "largely based" on the frequency of HiC contacts: without additional data supporting the functionality of a region in a tissue (histone marks, chromatin accessibility, TF binding, functional variants in the vicinity, etc), they would not be selected. The trickiest point is the assignment of an enhancer to one of the genes in its vicinity in specific tissues, and HiC contact frequency was chiefly used to make such assignments; as Dr. Danko rightly points out, this rationale does not always hold, but in the absence of other data it remains the best way to infer enhancer-gene pairs. In fact, Fulco et al. found that quantitative measures of chromatin state and chromosome conformation are strongly predictive of enhancer functionality, correctly ranking 6 out of 7 distal MYC enhancers in their study. I have mentioned this in the revised version of the manuscript (both in Results and Discussion). Out of potential dozens of regions that show the required features, I have settled only on those which can be assigned a potential functional role with high confidence. This is the reason why only 8 regions have been selected for

evolutionary analysis; this low number might look disappointing, but on the other hand it ensures that those regions are very strong candidates. As I mention in the Discussion, validation of these predictions will require complex functional studies to show that the deletion or mutation of these sequences changes the expression of their putative target genes in specific brain areas and developmental stages and, furthermore, alters neurodevelopmental pathways. Such studies should be ideally performed in model animals representative of several vertebrate lineages. This is a huge task that will take years to complete, so any effort at prioritizing the regions/genes to analyse could be of great help. The aim of this work was to characterize a set of regions with high likelihood of behaving as enhancers of neurodevelopmental genes in vertebrates, so that other researchers who have the technology to validate them can do so with minimal waste of time and resources. I believe that such goal has been accomplished as far as available datasets permit.

2. Dr. Robinson-Rechavi rightly points out that I depict evolution as an anagenetic process. I took it for granted that potential readers would understand that when I refer to "earlier then lamprey", for instance, I am referring to ancestral species living before the lineage leading to present-day lampreys split from the vertebrate tree. However, I understand that such language might lead to confusion and have corrected the manuscript accordingly.

3. Dr. Robinson-Rechavi suggests that blastn might miss some distant orthologs. Since we are dealing with non-coding regions, there is no obvious alternative to this approach. I have used an E-value cut-off of 10-6 for blastn, which is the standard procedure. I have now mentioned in Methods that this might miss some very divergent orthologous sequences.

4. Dr. Robinson-Rechavi is surprised by the omission of references to teleost-specific whole-genome duplication. Although I have mentioned in several places that teleosts show additional copies of some BREs, I decided not to go into that issue because it did not affect the main results and conclusions of the work and it could distract readers from the main message. However, I have now added a few lines in Results (BRE1) explaining that the extra copy of that region seen in zebrafish, fugu, tetraodon and stickleback is in keeping with the fact that teleost fish genomes have undergone one additional whole-genome duplication (on top of the two WGD common to all other vertebrates), and added a recent reference on the subject.

5. As for other remarks by Dr. Robinson-Rechavi, the chance of finding three random genes in a particular order and orientation is of 1 in 48 (0.021). In the case mentioned, in fact, there are four genes (including RBMS1), so the random probability of this specific arrangements is 1/384 (0.003). I have added this to the manuscript. I have also rephrased the allusion to the classical view of promoter-enhancer interactions to make it sound less aggressive.

6. Dr. Danko raises an interesting point about the background signal of virtual 4C plots and the fact that nearby regions tend to show high contact frequencies. I had already taken this into account when selecting putative enhancer-gene associations, giving more weight to distant peaks than to nearby peaks and doing "reverse" 4C plots (fixing the anchoring point either on the enhancer or the promoter to see if the interaction is seen in both cases; compare Figure 2 and supplementary Figure S1, for instance). I have made this clearer in the revised version (in Methods).

7. Dr. Danko asks whether the ncRNAs overlapping some of the BREs might represent enhancer RNAs (eRNAs). Most active enhancers are known to produce bidirectional short eRNAs, but they are unlikely to be identical to the ncRNAs annotated in Genecode since these are usually longer and undergo splicing (which is not a feature of eRNAs; see https://doi.org/10.1146/annurev-genet-110711-155459 for a recent review on the subject). Other studies have suggested that some enhancers act as promoters of ncRNAs (or viceversa), but this is a complex issue still unresolved and I did not want to dwell too much into it so as not to distract readers from the main message.

8. As mentioned in #1 above, the functional validation of these conserved elements will be difficult and time-consuming, especially if it is going to be done across most vertebrate lineages. I have tried to gather all published functional information about these enhancers in other species (mostly mouse, chicken and zebrafish), but very little is known about the genes they regulate in specific brain regions during development. Following Dr. Danko's recommendation, I have changed the title and toned down the claims of functional causality.

I would like to thank you again for taking the time to read critically this manuscript. I am sure that it has improved substantially following your comments and recommendations.

Best regards,

Decision by Marc Robinson-Rechavi, posted 02 August 2017

Revse

Dear Francisco Novo,

Thank you for submitting your manuscript for recommendation at PCI Evol Biol. We are aware that this is a very new concept, and we appreciate that you are giving it a chance. The process is also new for us, so please do not hesitate to give us feedback, our common aim must be to make the best science possible available.

As you will see, the expert reviewer I invited and myself found your approach interesting, but also that there were problems with your interpretation of the data. Thus I am not proposing at present a public recommendation of your manuscript. But I hope that the two reviews will be helpful for you to improve the work and the manuscript, to re-request a recommendation at PCI Evol Biol or to submit directly an improved manuscript to a classical journal.

Best regards Marc

Reviewed by Marc Robinson-Rechavi, 02 August 2017

In this manuscript, FJ Novo used genome-wide "epigenetic" marks (histone modifications, DNA methylation, chromatin accessibility, transcription factor binding) with chromatin contacts and gene expression data, to detect putative regulatory elements in the human brain. The evolution of these elements was then studied by comparative genomics.

I am very sympathetic to the aims of this paper, and the starting point of integrating functional genomics in one species with comparative genomics is sound. But I was disappointed both by the results and by the writing. I recommend to the author

I was disappointed that all the functional genomics integration led to the study of only 3 genes. Moroever, while correlative evidence is sufficient to discuss large scale patterns, I expect stronger evidence than that presented on page 8 to specifically infer the function of a regulatory element. Especially given the "manual inspection" step, which means that the analysis cannot be reproduced and is inherently subjective. Page 10, the link with educational attainment is interesting, but it should be noted that such complex phenotypes, like size or life expectancy, can be affected by an extremely high number of pathways. Thus this does not necessarily imply a role in the brain, in itself.

The manuscript systematically represents evolution as a progress from "lamprey or earlier species" to fishes, to "chicken onwards", which is erroneous. These are all present day species, which have evolved for the same time. We do not have evidence of functional genomics of the ancestral "earlier" species. It is possible and interesting to infer some of their characteristics from comparative data in a phylogenomic framework, but that is not done here.

"BRE1 is a vertebrate innovation appearing in Gnathostomes": since homology was determined by Blastn, it is possible that other species have an ortholog, but which is too divergent for detection. For protein sequences, it is not unusual that Blastp fails to detect true orthologs, which are detected by psi-Blast.

”We observed that coelacanth, spotted gar and elephant shark have orthologs for TANK, PSDM14 and TBR1 in the same order and orientation than mammals”: how does this compare to an expectation from 3 random genes?

It is surprising that the manuscript discusses a duplication in teleostei fishes (pp 11-12) without mentioning the teleost fish genome duplication, and the enrichment in transcription factors and in brain expressed genes in the retention of genes.

”The classical and largely outdated view of promoter-enhancer interactions suggested that a regulatory element would most likely regulate the activity of the closest gene”: reference needed, or you risk attacking a straw man.

Reviewed by Charles Danko, 30 July 2017

The manuscript by Francisco Novo, Identification and evolutionary analysis of eight non-coding genomic elements regulating neurodevelopmental genes, describes a detailed evolutionary analysis of candidate non-coding regulatory elements. Eight regulatory elements were selected based on their proximity to three genes – TBR1, EMX2, and LMO4 – which encode transcription factors likely to play roles in nervous system development. The bulk of the study describes an analysis of publicly available genomic data to identify the location of regulatory elements, combined with an effort to characterize the evolutionary origin of these candidate enhancers using a number of sequence based analyses. Overall this study is well done and will be of substantial interest to researchers in the field.

Comments:

(1) The candidate enhancers selected for detailed analysis were largely chosen based on the frequency of contacts in Hi-C data collected from human fetal brain. Novo makes the assumption that these regulatory elements, which bear the marks associated with enhancers and form loop interactions with the target genes of interest, regulate the transcription of these genes. Although there is mounting evidence to support the notion that these enhancers are more likely to regulate expression of the candidate genes (see especially Fulco CP et. al. (2016) Science, PMID# 27708057), there are undoubtedly exceptions to this assumption and no direct functional validation is available for most of the regulatory elements in the present study. The manuscript would benefit from toning down the language that implies a causal relationship between candidate enhancers and the genes of interest (including in the title). I also think that some discussion on the limitations of Hi-C data for this task, mostly noting that it is not a direct functional validation of enhancer activity, would also be useful.

(2) Special care should be taken when interpreting contact frequencies in the Virtual 4C plots that are nearby the anchor points (shown in Figures 2, 4 and Supplementary Fig. S6), especially near EMX2 (Fig. 4). Hi-C data, and indeed all chromosome confirmation and capture data, has a high signal in nearby regions that lie along the “diagonal” of a Hi-C heatmap. This is often interpreted by many authors as “background”. The Y-axis reads “Hi-C read value”, which I take to mean the un-normalized contact frequencies between two loci – it would be useful for readers to make it clear if normalization was applied to correct for the decay as a function of distance that is commonly found in Hi-C contact frequencies. In either case, it is possible that these contacts are biologically relevant, but this limitation should be considered carefully, and noted in the text, when interpreting the biological function of these putative loop interactions.

(3) Many enhancers in mammals recruit RNA polymerase II, which transcribes short, unstable non-coding RNAs (Kim et. al. (2010) Nature, PMID# 20393465). Could the poorly characterized non-coding RNAs overlapping several of the BREs reflect transcription of enhancer-templated RNAs transcribed from the enhancer itself?!

(4) The author tracks the evolutionary origin of DNA sequences that are identified as candidate enhancers using experiments in either human or mouse. Many enhancers that have an orthologous sequence in the genome of another species are not conserved at the functional level (see a variety of work by Duncan Odom’s lab, as well as others). While Novo is careful throughout the manuscript not to imply that DNA sequence conservation reflects functional conservation, adding an explicit note to the text that there is a major disconnect between conservation at these two levels would be useful for readers.

In addition, several of the enhancers described herein are conserved at the DNA sequence level in both human and mouse. In these cases a direct comparison between publicly available data in human and mouse may help to sort this out.

(5) Fig. 4 would be easier to read if the position of BREs near EMX2 were included.