Report on

Random genetic drift sets an upper limit on mRNA splicing accuracy in metazoans

by Florian Bénitière, Anamaria Necsulea, Laurent Duret

The authors analyze transcriptome sequencing data from 53 metazoan species to evaluate the hypothesis that genetic drift explains the positive correlation between genome-wide alternative splicing rate and organismal complexity (drift-barrier hypothesis). This hypothesis is a (neutral or null) alternative to the (adaptive) explanation that alternative splicing contributes to the evolution of complex organisms. I am very supportive of this idea to evaluate the accordance of this observed correlation with a neutral evolutionary model.

The drift-barrier hypothesis bases on the assumption that many detected alternative splices are splicing errors. These errors are less efficiently purged in species with small effective population sizes, thus explaining the negative correlation of the alternative splicing rate with proxies of N_e. To evaluate this hypothesis, first proxies for the effective population size (N_e) of all the species investigated are defined: body size, longevity and dN/dS. Indeed, based on the available data, the authors find that the alternative splicing rate negatively correlates with proxies of N_e, which is consistent with their suggested hypothesis (Fig. 2).

To further validate their claim, the authors differentiate between functional and non-functional splicing variants. The expectation from their suggested hypothesis is that functional splicing variants, them being under selective pressure, should be enriched among abundant splice variants, whereas non-functional variants should be enriched among rare splicing variants. Similarly, increasing the effective population size should decrease the alternative splice rate for non-functional splices and increase the alternative splice rate for functional splices, which is precisely what the authors find (Fig. 3). The authors even move on to further support the drift-barrier hypothesis by two additional tests: The selection strength on splice sites should increase for increasing population sizes (Fig. 4) and the abundance of rare splice variants should decrease with increasing levels of gene expression (Fig. 5). The findings are, again, largely consistent with the drift-barrier hypothesis.

Overall, this is a very convincing assessment of the drift-barrier hypothesis to explain different levels of alternative splicing across metazoans. The manuscript is well written. I particularly liked the introduction and the careful language, i.e., to not jump to conclusions too quickly. The manuscript is also well structured, which helps to follow the line of arguments. There are a few passages though that, in my opinion, need some clarification (more details in my list of comments below). Also, the mathematical model and Fig. 6 do not add value to the manuscript in my opinion. They should be removed or at least moved to the appendix (more details on this below). Nevertheless, I think that this is a very well done and scientifically sound and thorough analysis of a neutral hypothesis to explain the variation of alternative splice rates across organisms, which merits publication. A list of comments, suggestions and questions follows.

Comments

1. Line 124: The definition of RANS is unclear to me. Why is $N3$ divided by 2? I am sure there is a simple reason that escapes my attention. I suggest to add a short explanation (either here or in the Methods section – line 416).

2. Line 125: '... at least 10 reads." → I guess this refers to the sum of $N1 + N2 + N3$, is that correct? I suggest to clarify this.

3. Lines 125-135: The phrasing could be more streamlined in my opinion to avoid ambiguity. For example, I think that to describe a splice variant only one word should be used consistently (at the moment isoform and transcript are used interchangeably?). Also, again to avoid confusion, I suggest to use *minor splice variant* instead of *splice variant* for RAS$\leq$ 0.5 and *major splice variant* instead of *intron* for RAS$>$ 0.5 (is this actually correctly interpreted?) – alternatively minor and major intron would also be fine, but just writing *intron* for splice variants with RAS$>$ 0.5 is unfortunate. As it is, two different terms (splice variant and intron) refer to related concepts (larger or smaller RAS values), which should also be reflected in the words used in my opinion. At least I had problems to remember the definitions and it is sometimes difficult to figure out whether intron refers to any intron or an abundant splice variant.

4. Fig. 2B,C: I suggest to use the same y-axis scale in the two plots.

5. Lines 159ff.: I think it would strengthen this test of robustness substantially if data from an invertebrate would be added – of course only if feasible. Alternatively, I suggest to emphasize again at the end of the paragraph that all seven species are vertebrates.

6. Line 191: I was confused by the definition of MIRA. Is there a mistake in the denominator? Should it be $N1$ *minor intron*? (see also line 426)

7. Fig. 3B-D: These panels are not referenced in the main text (or just later in the discussion). I suggest to either move them to the Appendix or, better, to comment on them in the main text close to the figure. I think they make a good case for the drift-barrier hypothesis, which should also be mentioned (earlier) in the main text.

8. Line 225: 'significant' → I personally try to avoid using the word 'significant' if it does not relate to a statistical test. Here, I think it belongs to a statistical test, but then also the p-value should be given. (This also refers to other places in the manuscript.)

9. Line 228: This is actually a strong argument against the adaptive hypothesis (large alternative splicing rate in complex organisms) and I suggest to spell this out explicitly.

10. Line 259: I would write 'proxies of N_e values' for the sake of precision.

11. Lines 267ff.: I suggest to add ranges of values throughout this paragraph.

12. Line 273: I suggest to cite the Bush et al. paper already in the introduction where the drift-barrier hypothesis is introduced because the idea is put forward in this paper (e.g. their Section 4). In general, I think that this paper would merit to receive some more credit for the drift-barrier hypothesis idea earlier in the paper. Essentially, the manuscript by the authors is exactly doing the suggested comparative analysis across multiple species to assess the roles of genetic drift and selection on the alternative splicing rate.

13. Line 278: I suggest to replace 'the others' with the respective precise term (I guess rare SVs with MIRA$<$ 5%).

14. Lines 287ff.: I suggest to move some bits from this paragraph to the results closer to the referenced Figure.

15. **Fig. 6 (and model)**: I am not convinced of the added value of the model because it is a purely statistical association of parameter values that the authors already describe verbally. If there would be a *true* evolutionary model, in the sense that a population is simulated over multiple generations and results derived from these stochastic simulation, I agree that this would be an interesting proof of concept. However, as the model is set up, it is not very helpful. The key message is that for smaller effective population sizes the error rate can add to the proportion of introns with high alternative splicing rate. The authors acknoweldge this in the legend of Fig. 6: "... abundant SVs (AS $> 5\%$) correspond to a mixture of functional and spurious variants, whose relative proportion depend on N_e." This overlap, however, is not an emergent property of a simulation, but an a priori parameter choice (the mean of the gamma distribution varies for different effective population sizes), so the 'results' in the plots are just reflecting modeling assumptions, rather than results from repeated stochastic simulations of populations with varying effective population sizes. The model therefore is not a proof-of-concept. To make this a proper model, the same distributions (error rate and functional propensity) need to be used and then populations be simulated with varying population sizes. The results of such a simulation would then confirm that the drift-barrier hypothesis can indeed explain the observed correlation between population size and alternative splice rate. Moreover, panels C-F are summary statistics derived from panel A that could also be listed in a table instead of separate figures. I suggest to remove the model and the figure from the manuscript.

16. Line 337/338: 'nearly all species ...' $\rightarrow$ do the exceptions of the observation have something in common so that one can speculate as to why theses species do not follow the general pattern?

17. Line 429: I was a bit confused about the definition of the per-gene AS rate. As the formula is set up, it looks like the probability of having no splice variants is averaged over all introns of the gene, is that correct? If this is correct, I was wondering why the authors use the average over all introns of a gene, even though the information about each intron is available? In that case the formula would translate to

$$1 - \prod_{j=1}^{Ni} \left(1 - \frac{N2_k}{N2_k + N1_k}\right),$$

where $N1_k$ and $N2_k$ are the number of reads corresponding to the precise excision of the k-th intron, and the number of splice variants at the k-th intron of the a gene that has Ni major introns in total. I think this would be the more accurate way of measuring the per-gene alternative splicing rate.

18. Line 435: Is there some justification for the chosen maximum distance of 30 bp or is this value chosen arbitrarily?

19. Line 491: Commas are misplaced in the number of SNPs.