Report on revision of

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

by Florian Bénitière, Anamaria Necsulea, Laurent Duret

The authors have addressed some of my comments. Unfortunately, they seem to have missed the last page of my initial review. All of those points remain valid. I restate them at the end of this review. Importantly, I think that the model is not helpful because the *results* can be directly computed from the assumptions. It is a (straightforward) statistical association between model parameters but not an evolutionary model.

Besides these missed points, the manuscript has been revised to increase readability. The authors have done a good job to improve the explanation of the variables they use. However, I think that this presentation is still difficult to follow (lines 124ff.). A table with each variable name and its respective definition might help to quickly look up the variable names, instead of searching for their first appearance in the text (Figs. 2A and 2D already help a lot though!). In particular the distinction between abundant and rare splice variants can be confusing if one forgets that both are minor introns, or at least can be. Also, I was confused by the definition of the AS rate of introns (line 157): Is it the same as 1-RAS? I think there is a difference, but that is not clear from the formulas. I suggest to place Figs. 2A and D together with a table and all the formulas so that the dependencies between all those variables become clear and are easily comparable. This will help readers to focus on the actual scientific question rather than to always search for definitions of variables (which unfortunately is the case for me), which makes the manuscript hard to read.

Overall, I still think that this is a well-designed study to address the question if genetic drift explains the positive correlation between genome-wide alternative splicing and organism complexity. Once my remaining points, in particular my criticism of the model, are addressed, I would support the publication of this manuscript.

Repetition of points from my first review that were missed – line numbers and notation refer to the first submission

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

2. **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 acknowledge 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.

3. 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?

4. 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.

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

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