Bénitière et al “Random genetic drift sets an upper limit on mRNA splicing accuracy in metazoans”

Lynch’s drift barrier hypothesis postulates that, because the efficiency of purifying selection is inversely related to population size, mutation rates should be expected to show a similar correlation. Under this view (https://pubmed.ncbi.nlm.nih.gov/20594608/), selection against “mutators” (genetic variants responsible for higher mutation rates) is driven mainly by selection against the increase in mutation load they induce. As per the first sentence, this selection is expected to be more efficient in larger populations, explaining their typically lower mutation rates. This same perspective (due ultimately to Kimura and especially to Ohta) has been used to explain variation in levels of genetic variability of all sorts observed in nature, and represents a critical benchmark relative to which inferences of positive selection should be made.

The present contribution applies this line of thinking to splice site variation. Eukaryotic mRNAs often contain introns that are removed (spliced out) before being translated. Moreover, mRNA sequence data makes clear that many mRNAs exist in more than one splice isoform within cells. Previous work has demonstrated that the number of such alternatively spliced mRNAs is positively correlated with organismal complexity (defined as cell type heterogeneity), suggesting the possibility that greater levels of diversity are required to support said complexity.

On the other hand, it has also long been understood that organismal complexity is inversely correlated with population size (see also https://pubmed.ncbi.nlm.nih.gov/14631042/), suggesting the opposite interpretation of the data. Namely, because the efficiency of purifying selection is negatively correlated with population size, we might on first principles predict greater mRNA splicing diversity in smaller populations (as is also seen in their mutation rates). In this reading of the data, that those species happen also to exhibit greater cell-type heterogeneity would be a coincidence, rather than the selective driver of mRNA splicing diversity.

And indeed, the present contribution demonstrates a strong, negative correlation between three proxies for population size and splicing variability using large, published datasets from both insects and mammals. In both groups of species, putatively smaller populations exhibit higher splicing variability, consistent with the Kimura/Ohta/Lynch drift barrier hypothesis.

This is a very important finding, because it suggests a more appropriate “null hypothesis” under which to test adaptive explanations for splice variability, and I generally regard this work quite favorably for that reason. It will be appreciated both by individuals interested in mRNA splice function and evolution, as well as those more broadly interested in general principles of evolutionary and population genetics.

At present however, I find the manuscript hard to follow in its technical details, which risks minimizing its impact. For the second camp of anticipated readers (which includes me), the complexity of the biology requires more hand-holding to allow easy comprehension of the paper’s result. I enumerate points of confusion below, but before that, I also note two points of contact with the literature that seem to be missing at present.

First, the present work brings to mind the drift barrier's impact on transcription error rate (see https://pubmed.ncbi.nlm.nih.gov/26884158/). This phenomenon seems very closely related to the present work, yet those authors find quite a different pattern than that described here. I would therefore be interested in hearing the present authors' views. And second (more esoterically), I am reminded of much older work on intron phase (e.g., https://pubmed.ncbi.nlm.nih.gov/8618928/), and wonder whether there are any interesting correlations between this intron attribute and splice variability.

1. Line 2: this is trivial, but because of the diversity of definitions (and opinions) of biological complexity (and its evolution), I recommend that the authors explicitly state theirs, e.g., "...noticed that the complexity of organisms (i.e., the number of distinct cell types) correlates positively with…" One would hate to lose readers on such a "partisan" point.

2. Lines 33-34: one or two more sentences on the variability of intron splicing efficiency in nature would be most welcome, both with respect to spliceosome and splice signal "quality." I myself have no previous knowledge of how these facts are understood. Adding some details here will help all readers to imagine the mechanisms of splicing accuracy that are putatively under purifying selection..

3. Major vs minor isoforms. If I understand the biology correctly, isoform abundances fall into two entirely disjunct classes, as illustrated clearly by figure S2. If that's correct, I would encourage the authors to make that point before the material that begins at line 65. As it stands, I read the sentence that begins "This pattern is mainly driven by low-abundance isoforms…" as casting the situation as a two-dimensional problem, with mRNAs having a spectrum of isoforms, varying both in number and frequency.

   Indeed, clarifying that point much earlier might also improve comprehension of the material beginning on line 42. How do major/minor isoforms correlate with the 1% of isoforms that produce detectable amounts of protein? How do we know constraints are weaker on the protein products of minor isoforms? And how might minor isoforms be involved in gene regulation? To be clear, I have little doubt that the facts as stated are correct, and that they reflect the authors' deep understanding of the biology. Moreover, that understanding is likely shared by most readers in the mRNA splicing community. But as noted above, this work also has exciting implications for evolutionary biology, and much of that readership would appreciate more information on the basic facts.

   Relatedly, is it true that figure S2 is a histogram of AS frequencies across all species in the final dataset? So each gray dot is a frequency bin for some species, with lines connecting bins within species? And then the D mel and human bins and lines are highlighted? Esp because of what I regard as the centrality of this figure for the basic biology (does it perhaps thus deserve promotion to the main text?), I encourage the authors to explain its construction more clearly. Finally, I don't see the yellow trace for Apis that the legend promises.

4. The size and shape of the dataset. If I understand correctly, N = 978 is the number of single-copy orthologs across metazoans, ≈80% of which could be unambiguously identified. What does “unambiguously defined” mean? Is this a reflection of incomplete annotation, or something else? Please explain. And how many orthologous introns are there among those genes before and after you apply your >N=10 reads filter? (Trivial point, but recycling the symbol N risks confusion. When I first read line 100, I thought you were down to just N=10 genes. And as just noted, I would have liked to know the number of surviving introns, rather than only the percent surviving.) Finally, similar to my question about perhaps promoting Fig S1, I had a hard time following lines 86-96 without ready access to the figure. My thought would be to try rewriting this paragraph so as to simply cite the punchline of Fig S1, directing the reader to the supplement to learn more. Or, fully unpack all the details here. The present “hybrid” approach seems suboptimal.

5. Figure 1A. I was surprised to see a single phylogeny for insects and mammals. We know they exhibit reciprocal monophyly, a fact which in any case has no implications for this study. More seriously, I worry about the positive correlation in 1B, which seems to be driven entirely by the non-insects. (Equivalently, I can discern little to no trend among points corresponding to organisms whose body length is below 5 cm.) Similarly, removing non-insects from 1C seems to change its message.

   I would encourage the authors to explore the story that emerges by the independent analysis of insects and mammals. That might be the more appropriate framing of the data.

6. Line 115: that three poorly correlated measures provide noisy estimates of a fourth could be construed by some readers as wishful thinking. I’m not saying it’s not so, but especially since the fourth quantity is the fulcrum of the whole study, I feel the authors owe the reader a much stronger explanation here.

7. I found the section beginning line 116 exceedingly difficult to follow. N1, N2, RANS, RAS? Fig 2A is excellent, but my head is swimming nevertheless! It might help to use more informative variable names, but perhaps more importantly, I encourage the authors to add more English prose to illustrate how these quantities each work.

   Unfortunately, this is as far as I was able to get with the manuscript. My inability to internalize these key statistics left me unable to push further.

To summarize, I am very excited about the overarching thesis of this work, and its implications for two communities of readers: AS works and evolutionary geneticists. However as presented, it fails to reach this member of one of the second of those communities. I very much hope the authors will attempt a revision that includes enough intellectual and conceptual hand-holds to help my community appreciate their work. If successful, I have no reason to believe that this paper won’t be an important contribution to my field.