Summary

In the first review of Benitiere et al., I and the other two reviewers expressed our interest and belief in the value of the work performed whilst raising a number of points that could, or should be, addressed before we were happy to recommend the manuscript to PCI Evol Biol. Most of these points were relatively minor and related to either the explanation of methods and statistics or the discussion of previously published observations.

Although there are many ways in which the modified manuscript could be improved, both to strengthen the evidence for their thesis and to improve the explanation of their methods, I am largely satisfied with their revised manuscript.

However, whether the manuscript can be recommended depends on editorial policy regarding how thoroughly the computational methods need to be described and how easy the authors need to make it for others to reproduce their work. The authors have made available all the code they wrote (at zenodo.org) in order to perform the analyses and prepare the figures. However, in the original submission they provided only a minimal description of the purpose of individual scripts and how they connect to each other and the input data.

With this second revision they have provided a much improved `README.md` file and additional supplementary material. The updated README.md file and the addition of the data in `data/per_species/` helps greatly in clarifying the methods used by the authors. However, rather little emphasis has been placed on how the data used to make the figures has been generated. That is, there is no description of the pipelines used to convert sequence data to estimates of alternative splicing, and how these connect to the data present in the `data/per_species` files.

It is certainly possible to discern the process used from reading the scripts in the `pipelines/data generator` directory; however, this is time consuming and is certainly one factor that has delayed this review.

There are however, a number of things that I have not been able to work out:

1. Where is the code used to select the 53 out of 69 species finally analysed? I was suprised that *Danio rerio* was not included in the analyses. The only reason that I can think of, would be that the teleost genome duplication somehow makes the unambiguous identification of BUSCO orthologues difficult, but I cannot find any support for this in the supplementary data. (I note that the data tables normally use the species present in the phylogenetic tree to only analyse the selected species.)
2. Many of the analyses make use of, or depend on, one or more Excel files. In particular I am curious as to `Fichiers-data/metazoa_69species.xls` as it seems that this ought to have been included.
3. `Data5_supp.R` makes use of a file, `polymorphism/by_minor_intron.tab`. I've not found any trace of any such file apart from in `overlap_detection.py`,

which I suspect might create such a table for each species.

Although I believe that sufficient material exists to reproduce the analyses, this would require at least some editing of the scripts (esp. of hard-coded absolute paths) and the recreation of some of the input files from some of the output files. I do not think this is a completely unreasonable requirement as it is not trivial to produce a turn-key solution to reproducing an analysis as complex that peformed here. It would however, be much easier given a more thorough description of the pipelines used. Whether that should be a requirement I don't know, but it would certainly have made my review faster had it been available.

Other comments related to changes:

Line 61: Did you mean 'AS-NMD' rather than 'AN-NMD'?

Line 178-179: 'variation in AS rate among organs in each species is limited compared to differences in AS rate among species',

is better written as, 'variation in AS rate among organs in each species is limited compared to differences between species'

Figure 6: It might be an idea to swap panels A and B as B is mentioned first in the text. As it stands it is somewhat confusing to the reader.

Line 355: 'Fig 3A' should be 'Fig 4' ?

Detailed response

Quoting in this section

My comments to the authors responses follow below using the following notation:

> > My initial comments

> The authors response

My response to their response

Comments and responses

> I note that there is a large discrepancy between their title and the concluding statement of their abstract:
>
> *All these observations are consistent with the hypothesis that variation in AS rates across metazoans reflects the limits set by drift on the capacity of selection to prevent gene expression errors.*
>
> I think that the tone of the latter is more appropriate, and that the title over-states the certainty of the conclusions that can be drawn from the work. This is not because of any obvious weaknesses, but because it is inherently a difficult question to answer conclusively.

> We agree that in the end, we just propose a model (as always), and of course, a short title cannot give all the nuances that can be developed in the text. But we think it is important that the title gives a clear statement of our main conclusion.

I do not have that a strong opinion on this matter and consider it more of a decision for the editor. I would certainly hope that those interested will read at least to the end of the abstract.

But If the authors like it short, may I suggest:

"[Random] Genetic drift limits mRNA splicing accuracy [in metazoans]"

(Where words in [] are optional).

> In particular, Chen et al. (2014) claimed to have excluded an explanation based on Ne. Benitiere et al. do cite Chen, but they do not provide any reason as to the diference in the conclusions reached. There can be a large number of reasons, but the conclusions are incompatible and for Benitiere to be correct Chen must be wrong and this needs to be addressed directly.

> Chen et al (2014) measured the rate of alternative splicing across 47 eukaryotic species. They observed a strong positive correlation between the AS ...

I am very satisfied with the explanation and feels that it fills an obvious hole.

> I am also concerned that more recent work using long-read sequencing technology (Leung et al. Cell Reports, 2021, 10.1016/j.celrep.2021.110022) does not seem to show more AS in humans compared to mice (if anything the opposite was observed). This contrasts with several studies based on short read sequencing and again I feel that these discrepancies ought to be discussed.

> We agree with the referee that using long-read RNA-seq data would likely improve our estimates of AS rates. However, this type of data is not yet publicly available for enough species, in contrast with short-read RNA-seq data, which is abundant in public databases. We now discuss this point in our manuscript (line 379).
>
> Regarding the differences in AS rates between human and mouse, we would like to point out that the manuscript by Leung et al. did not aim to quantitatively compare human and mouse brain transcriptomes. The data they generated is indeed not directly comparable between the two species: this dataset includes considerably more Iso-seq reads for mouse (5.66 million) than for human (3.30 million). The number of analyzed individuals is also higher for mouse (12) than for human (7). Thus, it is possible that the sequencing depth, which is still a limiting factor for long-read transcriptome sequencing, could affect the authors' estimates of AS rates.

Although it is true that both the sequencing depth and the number of replicates was higher for mouse than human in Leung et al., figure S2 seems to argue that the depth of sequencing for both species was sufficient to detect the majority of splice variants. Figure S3, on the other hand, suggests a higher variability in RNA quality for the human samples than for the mouse ones, so it is possible that the sample preparation has affected the results. But this is likely to be an issue for any study involving human or other large animals where it is difficult to obtain very fresh samples.

Similarly, it is questionable as to whether any sample of mouse and human brain can be considered equivalent without a detailed description of the dissection procedure, and that is something that I've not found for any of the papers comparing splicing rates with human and other species. Certainly I do not think that the samples used in Barbosa-Morais et al. can be considered directly comparable. In fact they used publicly available data for primates (single-ended) and compared that to data (paired-ended) produced in house from a range of species. I also note that Mazin et al., do not really observe a higher rate of AS in primates; their highest observed rate is again in humans, but the rate observed in macaques is similar to that in rats and mice. Interestingly all the studies do seem to agree on chicken having the lowest splicing rate.

I don't know why apparently different results were observed by Leung et al., but it is of course possible that the library preparation means that not exactly the same thing is being estimated (eg. if for example there is some selection for full length transcripts).

My concern here is that there is much about alternative splicing that has become more or less accepted knowledge, but which to my mind has not been adequately demonstrated and that it might be good acknowledge the extent of uncertainty that still exists.

> > I think that the weakest point of Benitiere et al. is related to the composition of the data that they have used. They seem to be aware of this, but consider that it could only lead to an under-estimate of the affect of drift on AS. I am not completely convinced by this, and am concerned that the data is likely to comprise sequences from a range of technologies that can influence their observations. Unfortunately, there is a good chance that the different sequencing technologies will not be uniformly distributed between species owing to the fact that analyses of non-model organisms is likely to have been carried out at later dates and thus with more up to date technologies.

> Among the 3496 RNAseq dataset that we analyzed, 3463 (99%) were sequenced with Illumina. The sequencing technologies are therefore very homogenous across taxa. We added a sentence (line 108) to mention this point. We controlled for sequencing depth, which should be the main technical factor affecting AS detection. It should also be noted that the main results (Fig. 3A) were confirmed when using a subset of species for which the exact same protocol was used to prepare RNAseq data from seven vertebrate species (Fig. 3B).

Not only have there been several versions of Illumina sequencers, but there are also many different ways in which sequencing libraries can be produced. I cannot explain how these differences could result in differences in AS rates, but I have experience of how very minor differences in protocol can result in observable differences in the sequence data. As such I think that this is a potential confounding factor even if I do not expect that it is likely to have skewed the data sufficiently to affect the conclusions.

In addition, in this revision the authors have included a table, `data/Data10_supp.tab`, that provides details about the sequencing runs used in their analyses, and there is at least nothing obvious in that would indicate problems with the data composition.

The data presented in figure 3B is indeed striking and does support the conclusions of the paper; however the difference is really only between primates and others, and with one outlier (Gallus). Excluding Gallus (reasonable as it is not a mammal) that leaves open the possibility that the differences are not related to Ne, but some confounding factor, potentially related to the sampling procedure. Again, it is notable (but not surprising) that the authors of the cited work did not themselves perform the dissection of the primate tissues used and it is entirely possible that this may have affected the resulting observations.

> I think that the work would benefit from including analyses from more carefully collated data sets where care is taken to make sure that the underlying tech-nologies are equivalent. Ideally this would be done from species that differ in Ne but which are otherwise similar (eg. marine and fresh-water teleosts). There is also transcriptome data and estimates of Ne in asellid isopods (Lefebure et al., Genome Research 2017, http://www.genome.org/cgi/doi/10.1101/gr.212589.116), who argue that smaller Ne leads to larger genomes as a consequence of less effective selection. If Benitiere et al. are correct, there should also be an increase in the amount of low-frequency splicing events in species with lower Ne.

> We agree with the referee: it would be interesting to extend the analysis by comparing closely related species with contrasted effective population sizes, to limit potential sources of variation that we might have overlooked. We did analyze the asellid isopods dataset (we were co-author of this 2017 study): unfortunately, the RNAseq sequencing depth is not suffcient to quantify AS accurately, and furthermore, a reference genome assembly is lacking for most of these species. It would be worth investigating whether appropriate data (reference genome + deep RNAseq data) are available for other clades (e.g. marine vs fresh-water teleosts or endemic insular vs mainland passerine birds). However, this would considerably delay the publication of our results (it took us two years to collect the data presented here). We believe that the results reported here are already suffcient to support solid and original conclusions.

I did not intend to suggest that this would be necessary for publication; but that it would provide better proof of the thesis. I am well aware of the amount of work that has already gone into the manuscript and feel that is reasonable to be published in it's current state. I hope that the observations published here can be used to argue for such a study to be carried out.

> > The methods section of the main manuscript does a reasonable job of explaining what was done, but is unable to provide sufficient detail to describe how the analyses were carried out. This additional detail is provided from an external source (zenodo.org) which provides a large number of data files and scripts. However I've not been able to find a description of the overall pipeline. For example, there are individual R scripts that generate the different figures which is nice; however, these scripts read data from files of processed data, and worse the locations of these files are sometimes outside of the data archive itself.

> We provided in supplementary figure (Supplementary Fig. 10) a description of the pipeline used to process the data.

This is useful, but it provides information of what was done, rather than how. What I am concerned with are the programs, scripts and specific options used. It is possible to extract this information from the archive, but it is not easy (see more details later on).

> We also added information regarding the computing resources that are required to process these datasets. (line 461)

This is useful information and I'm happy to see it included.

> > What is worse is that I am unable to find tables of the original data sources; they may well be there, but to my mind I should not need to go looking for them as they (eg. identifiers for all of the SRA data, genome assemblies and annotations) are fundamental to the description of the materials used. Hopefully the authors need only provide a more detailed README.md file to address these issues.

> The identifiers of SRA data, genome assemblies and annotations are provided on the zenodo archive, in the file data/Data1_supp.tab. We added a sentence in the 'Data and code availability' section (line 599) to mention this point, and to give a brief description of the main content of this archive. As suggested by the referee we extensively completed the README.md file.

The changes mentioned did make it easier to work out the process followed, but it was more time-consuming than necessary (esp. for the review process). My main concern regarding the methods, is the process whereby sequence data from a large number of sources was selected and then converted to estimates of alternative splicing. The process followed is found in the "pipelines/SV pipeline" directory as indicated in the README.md file.

This directory contains a number of shell, perl, python and R scripts in addition to a Snakefile, but does not have much description as to the order of execution of the scripts. Most of the scripts are (not surprisingly) called by snakemake reading dependancy information from the included Snakefile. However, the snakemake command is itself called by the launch.sh command. The launch.sh script is an interactive script that takes 5 arguments which are used to define directory structures and to determine whether to run the data_source_generator.R script and provides arguments to snakemake.

One of the arguments to launch.sh is the name of an Excel file that it appears should contain SRA identifiers for a number of species (one per worksheet?) that are used in the downstream analyses. This Excel file (or files) is not provided but it would seem possible to create it from the data provided in data/Data1_supp.tab. However, it would certainly be more convenient for this Excel data to be included as it would make it far simpler to recreate and or modify the analyses carried out. It would also be sufficient to include a description of the requirements of these files (i.e. the columns that should be present).

The remaining procedure is defined by 38 Snakefile rules that either make use of the scripts in the same directory or call relatively well known programs (eg. hisat2). In some cases the versions of the programs run (eg. hisat2) are clear, but in other cases I am unable to find these (eg. kalisto). There are also programs run (eg, parallel-fastq-dump) where I've not found the source of the program used specified. I suspect that this pipeline will produce (amongst other) the sets of files provided in data/per_species, but it is not trivial to infer as to how this happens given the number of rules and scripts (and since I have no personal experience with snakemake).

The scripts themselves are generally well written, easy to read and as far as I can tell do what they are supposed to do (I've only looked at a subset for obvious reasons). It might be an idea to use a single language for comments even if Google translate works well enough. It also seems that not all the analyses run by the pipeline were used in the final manuscript (eg. kallisto). It would be useful to have some indication of the subset required for the analysis presented.

I'm mostly satisfied that it would be possible to recreate the analyses carried out in this manuscript, but the authors could certainly make it easier by documenting the process more clearly. It strikes me that as the pipeline is formally defined within the Snakefile, that there ought to be some application that can extract this into a visual representation similar to the manner in which is done for relational database structures, and that some such thing might be helpful in this case (as the `Snakefile` defines a dependancy graph).

> As far as I can tell the statistics are reasonably chosen; however, I cannot confirm that they have been correctly carried out. But in any case I am not overly concerned about the details of the statistical tests as these do not matter as much as the nature of the data upon which they were applied. That is, I am much more concerned about what unknown factors may affect the analyses in a non-random manner. In this case there may be issues that relate to the sequencing technologies used as well as the choice of species and individual samples that could affect the validity of the conclusions. Unfortunately, although they provide a list of species analyzed I have not found more detailed descriptions of the individual samples from which sequencing data was obtained. These details should be included in order to be able to address the validity of the analyses.

> We now include in the zenodo repository a table (data/data10_supp.tab) providing information on the samples used. Most of the samples come from Illumina platform (3463) and also PACBIO (4), ION_TORRENT (2), ABI_SOLID (15), L454 (4) and BGISEQ (8).

This is very useful and, at least from an initial inspection it would appear to me to argue for the soundness of the data set, as I do not see any obvious correlations between methods used and the parameters derived.