Guillaume ACHAZ
MNHN, UMR7205 ISYEB, CP 50
Atelier de BioInformatique
45, rue Buffon
75005 Paris

To: Nicolas Galtier
Recommender for PCI Evol Biol

January, 24th, 2019

Dear recommender,

We would like to thank you and the reviewers for the positive and constructive feedback regarding our manuscript entitled: *The quiescent X, the replicative Y and the Autosomes*, that we have submitted to PCI Evol Biol. In this revised version, we have addressed the requested suggestions. Mainly, we worked on clarifying the text, added and discussed the references and submitted our scripts (basic parsing in *awk* language) to a public platform.

Please find below a point-by-point response to all the comments. To facilitate the reading, our responses are in blue and citations in the text are colored in purple.

We hope that we have met all the concerns in this revised version.

Sincerely yours,

Guillaume ACHAZ,
on behalf of all authors

Decision

by Nicolas Galtier, 2018-11-08 09:04
Manuscript: https://doi.org/10.1101/351288

Minor revision requested

Achaz et al. report a simple but highly meaningful observation: the ratio of indels to point mutations, and of deletions over insertions, differ between X, Y and autosomes in humans. Why is this meaningful? Because (1) the results are fully consistent with the hypothesis that indels are more frequent than point mutations in quiescent oocytes (hence the X>autosomes>Y ranking), and (2) this very pattern has been experimentally demonstrated to occur in yeast. So we appear to be here learning about an important and general aspect of the mutation process. The two reviewers agree that this is an important result. They provide a number of useful suggestions, which should help improve the manuscript further. In particular, the authors should cite and take into account the last publications in this rapidly moving field (I take this opportunity to insert my and PCI Evol Biol's apologies about the slowness of the process), and make sure they provide a reliable, long-term public distribution of their source code.

> We would like to thank the recommender for his positive comments. We have addressed the suggestions from the reviewers as explained below.

Additional requirements of the managing board
We ask you to carefully verify that your manuscript complies with the following requirements (indicated in the 'How does it work?' section and in the code of conduct) and to modify your manuscript accordingly:
-Data must be available to readers after recommendation, either in the text or through an open data repository such as Zenodo, Dryad or some other institutional repository. Data must be reusable, thus metadata or accompanying text must carefully describe the data.
-Details on quantitative analyses (e.g., data treatment and statistical scripts in R, bioinformatic pipeline scripts, etc.) and details concerning simulations (scripts, codes) must be available to readers in the text, as appendices, or through an open data repository, such as Zenodo, Dryad or some other institutional repository. The scripts or codes must be carefully described so that they can be reused.
-Details on experimental procedures must be available to readers in the text or as appendices.
-Authors must have no financial conflict of interest relating to the article. The article must contain a "Conflict of interest disclosure" paragraph before the reference section containing this sentence: "The authors of this preprint declare that they have no financial conflict of interest with the content of this article."
This disclosure may be completed by a sentence indicating that some of the authors are PCI recommenders: "XY is one of the PCI Evol Biol recommenders."

> All the script files are now available at http://doi.org/10.5281/zenodo.2551441. Please note that we only used publicly available data for this manuscript (see Script accessibility section in the methods). We further added a "Conflict of Interest Disclosure" section as requested.

Reviews

Reviewed by Marc Robinson-Rechavi, 2018-09-27 09:11

In this manuscript, Achaz et al present an interesting extension to sex chromosomes of their previous observation of quiescent patterns of mutation in yeasts. In yeasts, they have previously reported that mutations biased towards indels accumulate during quiescence. In human, they now report that the X chromosome, which spends more time in quiescent oocytes, has a similar indel-biased mutation pattern, whereas the Y has the opposite pattern. This is especially interesting in the context of the recent results on mutational rates and patterns in humans.

> We would like to thank the reviewer for his constructive comments on our study.

Major comments:

The Y chromosome is enriched in low complexity regions, which complicate read mapping and mutation calls. The assertion p. 4 that "sequencing errors would affect all chromosomes equally" is at best an hypothesis to be tested. Indeed, it is contradicted by the statement in the Methods that accessible sites represent "on average 90% of the chromosome size, with the exception of the Y where it is 18% of the chromosome". I recommend that the authors take this issue into account both in the analysis and in the discussion.

> There could well be more SSRs in the Y chromosome but we here report a deficit of indels mutation in the Y chromosome. From the literature, it seems that the Y chromosome has a similar density of SSRs than the other chromosomes (Subramanian et al., Genome Biol, 2003, 4(2):R13). Furthermore, the density in transposons seems similar to the X chromosome, about 51% (Gu et al, Gene 259 (2000) 81–88). We only assume here that for the part that was sequenced, the “accessible genome”, there is not a strong difference in the sequencing errors among the chromosomes.
> We are not entirely sure how to handle this problem in the analysis but we have added a sentence acknowledging this potential issue. The text now reads: “For instance, sequencing errors or calling bias would likely affect all the “accessible” regions of all chromosomes equally and are very likely absent in mutations detected with a frequency higher than 0.01.”

All over the manuscript, the term "indels" is used, but in the discussion the authors specify that the pattern is driven by deletions. If that is the case, then I recommend specify "deletions" everywhere in the manuscript where it is possible to make that call.

> We have reported in the yeast experiments that both insertions and deletions are enriched in quiescent cells. Although the balance is in favor of deletions, both occur. We thus decided to keep “indels” in the present manuscript. Furthermore, 2/3 of the indel variants that we have counted have no ancestral state inference. Therefore, we do not know whether there are insertions or deletions.

Be careful of distinction between frequently occurring mutation, and mutation which has increased to high frequency in the population: "Similarly, frequent mutations that

are more likely to get fixed in humans (Minor Allele Frequency > 0.01) also exhibit a higher fraction of indels."

> Indeed, we meant here mutations that have reached the 0.01 frequency in the population. The text on p.3 has been revised and now reads: "Similarly, the mutations that have reached a 0.01 frequency in the population and are thus more likely to get fixed in humans also exhibit a higher fraction of indels (Fig. 2b).".

Whether the "low proportion of indels" is "due to the high density of coding sequences" could be easily tested by comparing patterns within and outside of coding sequences.

> If we restrict the pattern to exons, we no longer observe a difference for the fraction of indels for every chromosome (A, Y and Y have 0.075 on average of indels but with a larger variance, as counts are 10 times lower -- 10^6 SNVs and 8 10^4 indels --, once restricted to MAC>=3). Mutations segregating in exons are not only affected by the mutational process but also filtered by the selection that is imposed on the coding sequence. Thus, we expected indels to be less frequent in exons than genome wide (we observed 0.075 vs 0.086) and more homogeneous among chromosomes. Out of the 20 indels in the mitochondrial genome, none are in exons. We are not convinced that this observation sheds light on the pattern of quiescence that we are trying to characterize. However, we changed "due to the high density of coding sequences" into "due to the high density of functional sequences" because the mitochondrial genome hosts many non-coding RNAs.

p. 3 "likely because they contain many "slightly deleterious" alleles (15–17)." The references cited do not support this claim.

> Yes, sorry for the confusion. References were referring here to the slightly deleterious alleles, not to their associated pattern. The text now reads: "This result supports the view arguing that SNVs are efficiently removed by purifying selection only in the long run (20) likely because they contain many mutations with a small negative fitness impact, the so-called "slightly deleterious" alleles (21–23)."

p. 4 references 7 and 19 do not analyze indels, so the relation to the results of this manuscript should be clarified.

> We have rephrased this section to make the connection clearer and to include the missing references (see last paragraph of the Discussion on p4). The revised version now reads: "The pattern we report here is in line with recent observations reported for the cause of human mutations. First, the positive correlation between the maternal age and the number of maternally inherited de novo mutations (7) clearly demonstrates that non-replicating mutations accumulate in oocytes. Second, conversion (10) and recombination (9) rates are higher in females than in males; furthermore, the number of recombination events (8) or double-strand-break related mutations (22,23) increases with the mother's age, demonstrating that DNA breaks occur at a high rate in oocytes and accumulate during quiescence. These breaks are very likely the cause of the non-replicating indels.".

I don't understand the last sentence of the Discussion: why assume that the pattern observed should be advantageous?

> Sorry for the confusion, we propose this as a hypothesis to be tested, not as a truth. We have rephrased the last sentence of our manuscript that now reads: "It now remains to be investigated whether the differential contribution of males and females to genome evolution, especially in species with slow development, may have been selected for and whether it relates to the origin of anisogamy.".

For calling indels, multiple sequence alignments would be more accurate than pairwise alignments, and should be preferred.

> Yes, the reviewer is right. However, we only analyzed variants called by the 1,000-genome project and previously aligned human-chimpanzee genomes. Thus, we are not sure of what we could do in this regard.

The paragraph "Statistical significance" in the Methods is an odd mix of methods, results and interpretation.

> We have now rephrased the paragraph that now only reports statistical tests: "We tested all the differences using $\chi 2$ homogeneity tests from counts reported in Supp Table 1 and 2. Autosomes counts were pooled. The differential of indels vs SNVs is highly significant in humans: $\log10(P\chi 2) = -1018$, df=2, as well as in the human-chimpanzee comparison: $\log10(P\chi 2) = -2096$; df=2. The difference in deletion vs insertion is also highly significant: $\log10(P\chi 2) = -114$, df=1."

End of Methods, the assertion "close to the observed I vector" is not very clear; what is "close"?

> We now report the vector of the differences.

Additional literature which I recommend citing and discussing: Makova et al (Genome Res. 2004. 14: 567-573 10.1101/gr.1971104) report male bias in indels in rat and mouse. Jónsson et al (Nature 549, 519–522 2017) report little relation of indel patterns with age or sex of parents; stronger paternal slope with age for indels (Sup Table 9); and strongest effect of father, most significant of father-age (sup table 11). Makova & Hardison (2015 Nature Reviews Genetics 16, 213–223) report that indels are influenced by DNA environment, including chromatin, in correlation with other mutational patterns. Are there differences between X, Y, and autosomes? What is the influence of being in spermatogenesis (see preprint Xia et al)? Xia et al (preprint https://www.biorxiv.org/content/early/2018/03/14/282129) report very important results on the role of transcription in testes for mutational rates and patterns: "Widespread transcriptional scanning in testes modulates gene evolution rates". Gao et al (https://www.biorxiv.org/content/early/2018/05/22/327098) report that the view that "germline point mutations stem primarily from DNA replication errors" should be "called into question". Finally, Thomas et al (Current Biol 28, 2018, 3193-3197.e5) have recently reported a relation between longevity and mutation rates in owl monkeys, relative to human and chimpanzee.

We thank the reviewer for all these suggestions. All the references plus two extra ones have been included in the introduction and discussed in the 3 paragraphs of the discussion.

Minor comment:
The term gonosomes is not widely used in the literature on this topic; it would be clearer to use "sex chromosomes".

Indeed, replacement done.

Reviewed by Robert Lanfear, 2018-09-27 09:17

This paper uses a beautifully simple approach to generalise a fundamentally important finding about molecular evolution from yeasts to primates. The authors had previously observed that quiescent yeast cells accumulate indels and SNVs in roughly equal proportion, but that replicating yeast cells accumulate proportionally more SNVs. Using the nature of the mammalian germline as a natural experiment, the authors very convincingly show that exactly the same pattern exists in human genomes, and that it also occurs when comparing human and chimp genomes. The broadening of the scope of the pattern from yeasts to primates suggests that, perhaps, this pattern is something that we might see conserved across much of the tree of life. The paper is written clearly, simply, and concisely. The result is both very convincing and very exciting.

We would like to thank the reviewer for his positive comments.

The only major comment I have is that I would like to see the code used for the analyses archived somewhere with a DOI – this would assist others in replicating or building on the work presented here, and would also assist readers and reviewers in assessing the details of the way that the analyses were performed. Suitable venues for the code would be Data Dryad, FigShare, or Zenodo. Perhaps the most preferable way to archive the code is to first upload it to GitHub, and then mint a DOI for the github repository using Zenodo. Instructions for this can be found here: https://guides.github.com/activities/citable-code/. Generally though, as long as the code has a DOI, and the DOI is presented in the manuscript, that is all that is required.

All the script files are now available at http://doi.org/10.5281/zenodo.2551441.

Minor comments

1. I suggest using ‘sex chromosomes’ in place of ‘gonosomes’. Both are correct of course, but my suggestion is based on the observation that evolutionary biology papers more commonly refer to the x and y as ‘sex chromosomes’.

   Indeed, replacement done.

2. Similarly, I suggest replacing ‘neo-mutations’ with ‘de-novo mutations’, because the latter is more commonly used.

   Done.

3. I found this sentence hard to understand: “Interestingly, at the divergence level, the degenerating Y chromosome accumulates more indels and SNVs than any other chromosome (Fig. 1), a pattern that was reported previously (12,13).”. After looking

at figure 1, I see that by 'at the divergence level' you mean when comparing humans and chimps. I think it would be useful to clarify this. One suggestion would be to rephrase in terms of substitutions, something like "Interestingly, when comparing humans and chimpanzees, the degenerating Y chromosome has accumulated more indel and single-nucleotide substitutions…"

We think the divergence word is also meaningful. Thus, following the reviewer's suggestion, we have modified the text that now reads "Interestingly, at the divergence level while comparing humans and chimpanzees, the degenerating Y chromosome has accumulated more indels and SNVs than any other chromosome (Fig. 1), a pattern that was reported previously (18,19). All differences are highly significant (see Materials and Methods) since counts are typically on the order of millions."

4. "striking simple" should be "strikingly simple"

   Done.

5. "We also noticed the negative correlation between deletions and chromosome size but did not further investigate it yet." – why not just report the statistics of the correlation? This seems like a simple result to report.

   We now report the correlation (rho2=0.58) that is indeed quite strong.
   We also noticed a negative correlation between the deletion/insertion ratio and chromosome size (spearman ρ^2=0.58; P=5.8 10^{-5}), but this has not been further investigated yet.

6. I do not find this argument quite as accurate as I think it could be: "This result supports the view arguing that SNVs are more efficiently removed by purifying selection in the long run (14) likely because they contain many "slightly deleterious" alleles (15–17).". I think the lack of accuracy stems from the lack of comparison between indel and SNV fitness effects. For indels to be removed more by purifying selection, it is necessary and sufficient to assume that they have, on average, more negative selection coefficients than SNVs. So, a statement only about the selection coefficients of indels doesn't establish this.

   We are not entirely sure about what the reviewer means. We wanted to convey that indels have either a strong negative effect (and are removed by selection) or are neutral (and then can reach a higher frequency). This is only a hypothesis to explain why higher frequency variants are enriched in indels. Low frequency indels could also be more difficult to genotype or call. The text has been changed accordingly and now reads: "This result supports the view arguing that SNVs are efficiently removed by purifying selection only in the long run (20), likely because they contain many mutations with a small negative fitness impact, the so-called "slightly deleterious" alleles (21–23). Alternatively, one could imagine that indels are not equally efficiently called at all frequencies, thus leading to fewer indels at very low frequencies.".

7. "Because of the haploidy" should be "because of haploidy"

Done.

8. “in complete line” should be “completely in line”

Done.