Reply to recommender

Version of manuscript reviewed:
https://www.biorxiv.org/content/10.1101/2022.04.12.488043v2

We thank the recommender and all three peer reviewers for their encouraging comments and suggestions for improvement. We have revised our manuscript accordingly and respond to each point raised in the reviews below (blue text).

Line numbers refer to the PDF version with tracked changes.

Dear authors,

Thank you for submitting this interesting piece of work to PCI Genomics. I very much enjoyed reading this study and as you have probably seen, all three Reviewers were very positive about it too. They also highlighted some minor aspects that could be improved. In particular, please check the possible mislabeling in Figure 2 and consider the confirmation of species IDs using molecular data, which is an important aspect and should be easy to do with the new RNAseq data.

From my side, I have two additional minor comments:

Lines 79-81. I am not sure I fully understand this sentence, could you please clarify what is meant by the frequencies of stop codons within conserved protein domains falling "within the range of observed for coding codons in organisms with known genetic codes"?

Rephrased this sentence to: "Among karyorelicts, all three canonical stop codons (UAA, UAG, UGA) were observed in conserved protein domains, with frequencies between 0.08-2.9%, which fell within the range of codon frequencies observed for unambiguous sense codons in other ciliates where the genetic code is known…" (lines 83-86)

Figure 4C. Could you please name the X axis?

Labeled horizontal axis for the first row of panels too.

Provided an appropriate response to our comments, I would be more than happy to recommend this study. If possible, please enclose a point-by-point response to all the comments.

Sincerely,

Iker Irisarri

Reviews

Reviewed by Vittorio Boscaro, 07 Jun 2022 18:50

The paper by Seah, Singh, & Swart reports an already known but intriguing phenomenon in a large number of other ciliate species, which are good representative of two classes of phylogenetic (and hence evolutionary) interest. The manuscript is very straightforward, the scope is a bit narrow but with strongly supported conclusions, and the figures and text are clear. I also think the authors do a good job at explaining context-dependent codon usage to readers outside the field, and at describing the evolutionary framework they are found in. I have a single main comment, and a series of very minor suggestions that the authors might wish to consider.

Also, while the structure of the manuscript is sound and well thought out, there are occasional sentences that are a bit harder on the reader. This is partly because some of the essential concepts in this work have intrinsically confusing names (e.g. actually terminating stop codons vs. coding stop codons). Some examples are given below, but one suggestion that might fix many issues would be to spell out the subject of sentences whenever feasible, minimizing the instanced of “this” “these”, “it” etc.

Thanks for pointing out this issue. In addition to the specific instances mentioned in other comments below, we have changed some other instances where the referent is potentially unclear:

Line 32: “The outcome during translation..”

Line 106: “This frequency…”

Line 284: “context-dependent sense/stop codons confer…”

MAIN COMMENT

Ideally, the authors should provide a bit more data/info on the way they IDed the ciliates. All I could find was “ciliate cells were identified by morphology under a dissection microscope” (lines 288-289). The authors only go as far down as genera with their assignments, so in a pinch this might do, since genera are usually easy to identify in both heterotrichs and karyorelicts. However, there is some room for uncertainty, and results on imprecisely assigned specimens has long-reach consequences, even if it does not impact the conclusions of this paper.

The authors should have access to the 18S rRNA gene sequences of the specimens they isolated. I would suggest to deposit these separately, and explicitly state that they were used to confirm the morphological assignments. At a minimum, the authors should use BLASTN similarities. Preferably, they should build a small phylogenetic tree by adding their sequences to reference heterotrichs and karyorelicts (e.g., taken from the PR2/EukRef database). Please note I do not suggest that the authors bog down their result section with this. A few sentences in the methods and a tree as a Supplementary Figure would suffice.

Thanks for this suggestion. We agree that including an explicit 18S rRNA based identification would give a better baseline for future interpretation of our results and reuse of our data.

We extracted 18S rRNA sequences from the transcriptome assemblies and included them in a phylogenetic tree with reference sequences from the PR2 database, as suggested. These confirm the preliminary identifications based on morphology, as well as previous observations that many trachelocercid genera as well as the loxodid genus *Remanella* are probably non-monophyletic.

Added tree figure as Figure S3.

Described methods under “RNA-seq library quality control and transcriptome assembly” (lines 350-362), and briefly reported in Supplementary Text (“Confirmation of phylogenetic identity with 18S rRNA sequences”).

Deposited alignment and tree file in Edmond at https://doi.org/10.17617/3.QLWR38, and 18S rRNA sequences in ENA as accessions OX095806-OX095846 (release pending).

MINOR SUGGESTIONS

Lines 23, 59, and 216: While Karyorelictea are certainly fascinating, and they are indeed globally distributed and probably under-sampled, it is a bit of a stretch to call them “abundant” – most ciliates are ecologically important for one reason or another, but only rarely in terms of their number and biomass. Maybe highlight a different trait, or provide a citation somewhere in the text about their underappreciated abundance?

We wished to convey that karyorelicts are commonly encountered in marine coastal environments, more so than the relative dearth of attention paid to them would suggest, so “abundant” may not have been the most appropriate word. Karyorelicts have been observed to be locally abundant (e.g. >1000 per cm^{-3} sediment at peak abundance, doi:10.1007/BF00016241), but it is admittedly difficult to generalize this statement without being caught in a truism (any species can be locally abundant given a sufficiently small spatial scale and specific locality).

Changed “abundant” to “common” (line 232) or “commonly encountered” (line 62), and to “diverse” in the abstract (line 23).

Line 42: “alveolate” and “ciliate” refer to different taxonomic ranks, so I would not use them in the same sentence in opposition. I suggest saying “dinoflagellate” for *Amoebophrya*.

Changed to “marine parasitic dinoflagellate *Amoebophrya*”. (line 43)

Lines 60-62: It is not essential and there is no obvious causal connection with the genetic code, but maybe the authors could also mention here that karyorelicts also differ from all other ciliates in their macro/micronuclear cycle pattern.

Described non-dividing macronuclei as a distinctive feature of karyorelicts.

Lines 70-71: please change to “… 25 transcriptome assemblies (15 [of which] previously published) were used to…” for symmetry.

Changed to “25 transcriptome assemblies (of which 15 were previously published)”. (line 73)

Lines 89-90: an example of a sentence where “these” is a bit ambiguous.

Replaced “these” with “frequencies of the UGA codon in karyorelicts” (line 94)

Line 97: The reference to Figure 4D is out of order compared to all other figures.

Moved Figure 4D to Figure 1B, updated captions and references in text.

Line 100: provide a citation for BUSCO and the Alveolata marker set you used.

Added citation here too (was already cited in Methods).

Lines 147-150: I don’t understand this sentence. I guess that my main issue is understanding how the second part logically follows the first?

Changed this sentence to:

“Permitting both UGA+UAA as stops in karyorelicts resulted in a higher variance in 3’-UTR lengths compared to permitting only UGA. Although this was contrary to criterion (ii) above, we judged that this metric was not as useful in deciding whether UAA was also a stop codon, because the difference was small, and transcripts with putative UAA stops were relatively few” (lines 159-163)

Line 181: “and” instead of “while”, maybe?

Replaced “while” with “and” as suggested. (line 196)

Lines 212-214: feel free to mention that Wilbertomorphidae are also monotypic and, to my knowledge, only observed once. It is absolutely understandable that the sort of data needed for this paper is missing for this family.

Changed to: “... the monotypic Wilbertomorphidae, … and which has to our knowledge only been reported once.” (lines 228-230)

Lines 260-278: no comments here, I just want to give credit to the authors for highlighting how facile adaptationist speculations can be – and clarifying that one would need evidence to claim there is an adaptive value in any trait in the first place. Very weak adaptive explanations have been proposed for other genomic processes in ciliates in the past.

Thanks for this. We agree that adaptationist explanations are often proposed uncritically, and should be weighed against a suitable null model.

FIGURE 1: “Trachelocercidae” and “sp.” should not be in italics. Also, please move the two *Blepharisma* entries next to each other.

Changed “Trachelocercidae” and “sp.” to non-italicized.

Also changed labels at bottom of figure to non-italicized.

The two *Blepharisma* entries are not adjacent because we wished to group the libraries representing genomic CDS data together, separately from those from transcriptome assemblies.

FIGURE 2: I would suggest not using italics in certain labels, such as “In-frame UGA” or in the Library source box.

Changed these labels to non-italicized.

FIGURE 3: in here, genera should be italicized. Also, shouldn’t the codons related to each column be shown somehow?

Italicized genera names; added codon sequences to figure.

FIGURE 6: the layout of the right side of this figure suggests, at first glance, that *Remanella* and *Kentrophoros* are karyorelicts, while *Trachelocerca* and *Anigsteinia* are heterotrichs…

Added colored labels to the micrograph panels to distinguish karyorelicts vs. heterotrichs.

Also notice that elsewhere in the paper you only mention about having collected Trachelocercidae, not specifically *Trachelocerca*, which is a trickier claim (see my main comment, but in this case probably not even the 18S would suffice, since genera in Trachelocercidae are probably non-monophyletic).

Thanks for catching this. Changed label to “Trachelocercidae sp.”

Reviewed by anonymous reviewer, 06 Jun 2022 15:58

Seah et al. assessed the genetic codes of two ciliate groups (karyorelicts and heterotrichs) using existing and newly generated genomic/transcriptomic data, and show that karyorelicts use an ambiguous stop/sense codons. This study should be of broad interest to geneticists and protistologists, and will be helpful for genome annotations of ciliates (as well as other eukaryotes). I think the manuscript is well written, the analyses are well-designed and appropriately performed, and all code and raw data are publicly available. I congratulate the authors on a very nice manuscript!

I only have a couple of minor suggestions that the authors might want to consider for improving their manuscript.

L 89-90: "Nonetheless, *these* were all still...". What do the authors mean by "these"?

Replaced “these” with “frequencies of the UGA codon in karyorelicts” (line 94)

Ambiguous usage of “these” and similar wording was also pointed out by reviewer 1, and we have made changes elsewhere (see responses above).

L 96-97: It would be interesting to know if the percentage of transcripts with in-frame UGAs is impacted by genome completeness. Do the authors investigate this?

We observed that the fraction of in-frame UGAs in karyorelicts varies between different families or genera (Figure 1B, formerly figure 4D). The differences observed in the fraction of in-frame UGAs between different taxa was greater than the variation observed in completeness scores for the newly sequenced libraries from this study (Figure 2A, library names in bold), which were all around 15-20%. Therefore, we think that it is unlikely that genome/transcriptome completeness has a substantial effect on the fraction of in-frame UGAs observed.

Given the variation in in-frame UGA prevalence between taxa, to investigate a relationship to genome completeness, we would also need to have more libraries from the same taxon, sequenced at different depths or with different library preparation kits. At the moment, we do not think that there are sufficient data to investigate this question, nor that it affects the conclusions we draw from our results.

Figure 3: I found this figure to be a difficult to read. First, "codons with frequencies less than 0.02 are highlighted in red". I did not see this at first, and I wonder if this can be made more obvious somehow. Second, I think it would be helpful to have an axis at the bottom with the indicating the codon under consideration.

Increased size of the red highlight to make it more obvious.

Added codon sequences to the figure as suggested.

Reviewed by anonymous reviewer, 03 Jun 2022 15:55

After generating ten new single-cell RNA-seq libraries, Kwee Boon Seah and colleagues performed an in-depth computational analysis to infer the genetic code of a number of karyorelict and heterotrich ciliates. In continuity with Swart et al 2016, this work expands our knowledge about alternative nuclear genetic codes and provides additional evidence about the existence of a context-dependent ambiguous genetic code in karyorelicts ciliates. While lacking some direct experimental evidence (see below) and mechanistic insights, the genomic analysis is carefully conducted and the results are compelling and convincingly discussed. Overall the paper reads very well and I believe it could be of interest for a broad scientific audience.

However, I have some minor comments that should be addressed:

- Line 109-110: For the sake of clarity, I would add “(i.e., UAA and UAG)” after “ which were comparable to frequencies of the known stop codons”.

Changed to: “the known stop codons in *Blepharisma* (UAA, UAG) and *Stentor* (UAA, UAG, UGA)” (line 119)

- I would suggest merging Figure 4D with Figure 1. This would facilitate the reading of the text.

Moved Figure 4D to Figure 1B.

- Figure 2: In the small legend box, below panels B and C, “Karyorelictea transcriptome” should be highlighted in blue and “Heterotrichea transcriptome” in green.

Thanks for catching this error. Swapped color labels to correct taxon.

- Figure 3: The individual codon sequences should be included in the Weblogos plots (similar to Swart et al. 2016 - Figure 1B). Furthermore, the size of the codon frequency values should be increased.

Codon sequences added to figure; increased label font size for frequency values.

- Line 182-185: I would be curious to know what is the estimated percentage of transcripts with putative UAG stop codons? Is the UAG codon depleted before the stop codon in those transcripts? Is there enough signal to answer these questions?

We did not observe any obvious depletion in coding-UAGs before putative UAA and UGA stops in karyorelicts, unlike coding-UAAs and coding UGAs, which are depleted before both putative true stops, regardless of whether they are UAA or UGA. On the contrary, coding-UAGs actually had higher frequencies in karyorelict CDSs compared to 3’-UTR (Figure 5, green bars), exactly the opposite pattern expected for an ambiguous stop/sense codon.

We did not estimate the percentage of transcripts with putative UAG stops, because most of them would be false positives because of the frequency of coding-UAGs close to the 3’-end of CDSs. Coding UAGs are common in karyorelicts (codon frequencies close to the median frequencies for each library, see Figure 1).

The initial indication that UAA could also be an ambiguous stop was that when UAA was also considered a possible stop codon, there was an increase in the number of transcripts where an in-frame putative stop could be identified (Figure 4A; 5402 vs. 5022 transcripts if only UGA were permitted as stop). However, allowing UAG as a stop codon in addition to UAA and UGA did not result in any further increase in the number of transcripts with a putative stop (5403 transcripts).

- Line 195-196: I feel that this sentence should be toned down. The RNA-seq data provide a robust base to infer genetic codes but additional direct experimental evidence (e.g., Ribosome profiling or MS data) would be needed in order to confirm the computational predictions. Furthermore, in Swart et al 2016 the Ribo-seq and MS analysis were performed on the heterotrich Condylostoma and not on the karyoelict Parduczia sp. I would also recommend discussing possible complementary experimental approaches that would make the authors’ claim stronger and could provide some more mechanistic insights into the proposed context-dependent stop/sense codon model.

Changed the first line of Discussion (line 210) to: “We have found evidence that …”

Discussed applicable experimental approaches (lines 275-282): “To verify our predictions that UGA is the main stop codon and UAA a lower-frequency alternative stop, ribosome profiling and mass spectrometry detection of peptide fragments corresponding to the expected 3’-ends of coding sequences, e.g. as performed on *Condylostoma*, are most applicable experimental methods. If a karyorelict species can be developed into a laboratory model amenable to genetic transformation, manipulation of the 3’-UTR length and sequence would allow us to test the ‘backstop’ hypothesis more directly and tease apart the factors contributing to translation termination in these organisms.”