# Editor

*by Fernando Racimo, 24 Jun 2022 11:40*

**Minor revisions needed**

Thank you for your patience. Your preprint has now been seen by three reviewers, who all generally think this is a very well-designed and executed study. They provide a list of comments that should be reasonable to tackle, in large part involving stylistic edits and textual clarifications, but no major criticism, so I believe this can lead to a recommendation after they are addressed. I am looking forward to receiving your revised preprint.

>>> We'd like to thank you and the three reviewers for your time on our manuscript and your positive appreciation. We respond point by point to the reviewer's comments in the following. The tracked changes documents (main + supp. info) are attached to the present letter.

# Reviews

**Reviewed by Michael Westbury, 27 May 2022 09:25**

Fraïsse et al. have submitted a really nice manuscript about adaptive introgression in sea squirts in the English Channel. It shows how invasive species may actually help native species in some respect which is positive considering how many invasive species there are around the world. It confirms previous findings using smaller datasets but at a more robust scale through the inclusion of phased whole genomes. Overall I have very little to criticise, the analyses seem suitable for the purposes and results robust as a result.

>>> Thanks for your enthusiasm!

I only have a few specific comments listed below:

40: Persist for long what? durations?

>>> We rephrased it as "... persist for long periods during species divergence".

126: It would be nice to add a short summary of the sequencing results, coverage etc and then send the reader to Table S1 for more details.

>>> We followed your suggestion and added a short summary of the sequencing results at the start of the Results section: "A total of 48 whole genomes were sequenced with an average of 41M reads per individual (Table S1), including 22 *C. intestinalis* (three were excluded due to poor sequencing), 15 *C. robusta*, 6 interspecific hybrids and 5 *C. roulei*. An additional 4 *C. edwardsi* individuals were sequenced to be used as an outgroup, with an average of 88M reads per individual (Table S1). Reads were aligned against the *C. robusta* reference genome (GCA_009617815.1). Differences in the mapping quality were observed between species in agreement with their genetic distance to the reference (Table S1). On average, 80% of the reads mapped in proper pair in *C. robusta*, 60% in *C. intestinalis*, 59% in *C. roulei*, 68% in the interspecific hybrids and 44% in the outgroup *C. edwarsi*. The average depth was broadly similar among species, ranging from 18X to 26X".

137: We cannot tell apart the intra and interspecific hybrids in the figure. I suggest different colours or shapes in the figure to make it clearer.

>>> Thanks for the suggestion. We now use different shapes to distinguish the intraspecific (black star) and interspecific (black circle) hybrids in Figure 1.

164: There is another chromosome in figure S4B that has a higher value than chr5. Is there an explanation for that? Maybe this is chromosome length which you mentioned later but wasn't super clear.

>>> Yes, it has to do with the correlation between chromosome length and *fd*: the chromosome with a higher averaged admixture proportion than chr5 is chr13, the shortest chromosome. The important point is that chr5 does not follow this correlation. We now clarify this point: "Furthermore, the averaged per-chromosome admixture proportion was weakly negatively correlated with chromosome length (**Figure S4B**), a known proxy for the recombination rate (Kaback 1996). Such correlation is consistent with higher recombination rates (shorter chromosomes) producing weaker barriers to introgression (Martin and Jiggins 2017). However, chromosome 5 was a clear outlier (i.e., it has a higher *fd* value than expected given its length)."

168: How do you know it is weakly correlated? Just visual inspection or a regression line?

>>> We calculated the spearman's rank correlation coefficient ($\rho$=-0.21, non-significant). It is now indicated in the legend of **Figure S4**.

344: Where could this natural hybridisation have occurred? Since the species are now found in the Pacific and Atlantic oceans, it seems like a long dispersal.

>>> This has puzzled us as well: it is unclear where the place of the past hybridisation may have occurred, given the current species distribution. We can nevertheless provide hypotheses that we now present in the Discussion section: "The past introgression between *C. robusta* and *C. intestinalis* is puzzling given natural transoceanic migration was impossible during glacial periods. The signal of introgression we detected might come from a ghost (extinct or unsampled) lineage (Tricou, Tannier, and de Vienne 2022) related to *C. robusta* that colonized the Altlantic at the previous interglacial and came into contact with *C. intestinalis* during the last glacial maximum. Indeed cryptic lineages are often found in the genus *Ciona* (Zhan et al. 2010, Mastrototaro et al. 2020) that may prove better candidates for a 30 Ky old introgression event".

389: Italics are missing.

>>> Corrected - thanks.

457-458: How was species determined a priori? Does this mean they are easy to tell apart morphologically?

>>> The two species are quite similar morphologically, but they present diagnostic features (Sato, Satoh and Bishop 2012; Brunetti et al. 2015): for instance, *C. intestinalis* has a soft often pale and translucent tunic; while *C. robusta* has tubercules on the tunic near syphons. Mature individuals also differ by the pigmentation of *vas deferens* and *vas deferens papillae*. However, cryptic lineages are also observed sometimes. For this reason, we double-checked the species type using mtDNA genotyping. We added a sentence to clarify this: "Species were identified first by using morphological criteria (Sato, Satoh and Bishop 2012; Brunetti et al. 2015). Morphological species identification was further validated using a diagnostic mitochondrial locus (mtCOI, following Nydam and Harrison 2007)".

483: Supplementary scripts: Was there any adapter trimming/PE read merging? Anything processing prior to mapping.

>>> We controlled the quality of the reads using FastQC, and as their quality was good, we did not apply any processing to the reads before the mapping. This is now indicated in the text: “After quality control with FastQC v0.11.2, reads were aligned to the *C. robusta* reference genome using BWA-mem v0.7.5a …”

501: I assume this is to test for reference bias? Would be good to mention that.

>>> Yes, we computed the VAF to test for reference bias and to correct wrongly called heterozygous genotypes. We now clarify this point: “We then introduced a step of genotype verification (and correction where required) to check for reference bias and miscalling”.

517: Is this mutation value known for sea squirts or estimated?

>>> It was estimated for sea squirts in Tsagkogeorga et al. (2012). This is now explicitly stated in the text: “All trios were phased given parents and offspring genotype likelihoods, setting a *de novo* mutation prior to 1e-8 /bp/year (estimated for sea squirts in Tsagkogeorga et al. (2012)).”

544: The tools/software used in this section are lacking.

>>> We used Simon Martin’s tutorial to compute the D and *fd* statistics. We added the link to the tutorial at the end of the “Detection of introgression with summary statistics” section.

588: How was the log-ratio test run?

>>> The log-ratio test is based on the comparison of the likelihood of two models: a neutral model (null hypothesis) and a selected model (alternative hypothesis), which represents either a selective sweep (SweepFinder) or an introgression sweep (VolcanoFinder). The respective test statistics are implemented in SweepFinder (DeGiorgio et al. 2016) and VolcanoFinder (Setter et al. 2020). We reformulated the sentence in our manuscript to clarify that the log-ratio tests are directly implemented in the two methods: “Chromosomes were scanned with the two methods applying a log-ratio test for selection at test sites spaced by 1Kb”.

**Reviewed by Andrew Foote, 24 Jun 2022 11:09**

Fraïsse et al. present the results of an elegant study which provides strong evidence for recent introgression due to shifts from allopatric to partially sympatric distributions in *Ciona* sea squirts. The work builds upon previous work done using sparse markers, much of it from the same research group. The use of phased whole genomes in this study both confirms the hypotheses of, and provides a significant increase in resolution over, the previous work done using RAD-seq and other markers. The care and attention to detail throughout are really appreciated. As is the authors embracing of making all aspects of the study open access. A good example is the provision of the scripts via links (I will likely be using some of these myself in the future, so thank you on behalf of our research community).

>>> Thanks a lot for your nice appreciation.

One thing I lost track of was understanding which of the many datasets were used in which study and why this was. So an additional table in the supplementary materials listing the datasets, the analyses they were included in, key characteristics distinguishing each dataset, and the rationale for the choice of the dataset for each analysis would be helpful.

>>> Thanks for the suggestion. We added a Supplementary Table (S5) that summarizes the features of each dataset.

As the authors highlight, the local recombination rate is a likely cause of variation along the genome in the introgression rates (outside of the hotspot of Chr5). Given the extensive analyses done by the authors, it was a little surprising to see that recombination rates were not estimated to confirm this. Local recombination estimates could also provide support for the SFS-model-based approach in addressing whether short and long introgressed tracts reflect different introgression events (which is alluded to later in the demographic modelling section based on the recombination rate provided by Duret). But perhaps this could be included in the next draft, or is it in a forthcoming study by Duret?

>>> We fully agree that obtaining estimates of local recombination rates is crucial in population genomics. Laurent Duret's lab is currently working on a population-based estimation of the local recombination rate (4*Ne*.*r*) in *C. intestinalis* and *C. robusta* with LDhat using our data. However, the relatively low number of phased genomes per species makes the estimates locally imprecise. Furthermore, one would ideally estimate "*r*" from a linkage map instead of "4*Ne*.*r*", as this latter may be affected by linked selection locally in the genome. Unfortunately, we do not yet have a fine-scale linkage map in *C. intestinalis* to address the critical points you raise.

Regarding the introgression hotspot on chromosome 5, the presence of tandem repeats and the region of missing data reminded me of the outlier region in the comparison of hooded and carrion crows (Poelstra et al. 2014). From memory, that study found that much of the functional variation associated with the colour polymorphism between the two crow species was in this region, and that the low recombination rate due to long flanking repeat regions caused this to segregate between the two species. That study suggested that although functional, it may not be adaptive, but rather just a consequence of the local genomic architecture. Could reduced recombination be enough to explain the patterns in the introgression hotspot without evoking adaptation (I appreciate the authors have run a number of tests for selective sweeps, but I am just trying to play devil's advocate).

>>> Thanks for sharing your thoughts on this. As reflected in our Discussion, we considered alternative hypotheses that could explain the introgression hotspot. But in the end, we were quite convinced that some form of selection should have played a role in producing this pattern. We agree that duplicated repeats may be an underestimated way to trigger arrested recombination locally in the genomes (in line with Poelstra et al.'s study). On the one hand, as reduced recombination may facilitate genetic divergence to build up, one would expect the genomic region to have accumulated more differences than in the rest of the genome, which is incongruent with the observation of the introgression hotspot. On the other hand, reduced recombination is expected to generate long haplotype blocks, as we observed. We added a few words in the Discussion section to acknowledge your point: "Observing long haplotypes at intermediate frequency could thus be explained with purely neutral processes, especially if the hotspot corresponds to a region of reduced recombination (duplicated repeats may be an underestimated way to arrest recombination locally in the genomes, e.g. Kim et al. 2022)". To conclude, we are aware of the limitations in our study to firmly demonstrate the role and type of selection acting, so we look forward to future work to deepen our understanding of this striking genomic pattern.

The discussion is quite lengthy and covers some of the same ground as the introduction. It takes a while to get into the discussion of the new results. For example, lines 327-337 cover the findings of previous studies, which are also summarised in the introduction. Some summary of the work that led to the present study is of course justified, but could this be condensed?

>>> Thanks for your suggestion. We shortened the start of the Discussion; including by removing the text on lines 327-337.

Figure 2A. The shading of allele frequency runs 'low' to 'high'. Can the range be specified as actual frequencies in the figure legend?

>>> The frequency range is now specified in the legend of Figure 2A.

My comments are just very minor points that may not need any revision. The work is extremely thorough. I came away inspired by this study and excited to take some of the findings and apply them to my own work. I feel this would be an exceptional contribution to any journal that published evolutionary biology research. My congratulations to the authors.

>>> Thanks! We feel honored that our work has inspired you!

**Reviewed by Erin Calfee, 27 May 2022 07:14**

The authors show strong evidence of recent adaptive introgression from C. robusta into C. intestinalis (absent in C. roulei) on chromosome 5, against a background of low genomewide admixture. This case study of adaptive introgression is particularly interesting given the relatively high divergence between these sea squirt species and their human-mediated secondary contact. The evidence for long introgressed haplotypes surrounding a “missing data region” where C. robusta has excess copy number is particularly striking, and the authors have identified a promising candidate gene in this region for future work. The authors use small but strategic geographic samples and whole genome sequencing phased by trios to reach their conclusions. Congratulations on a fine paper! I have only a few suggestions for a stronger manuscript.

>>> Many thanks for your positive assessment!

The SweepFinder results for C. robusta (line 273) combined with the star-like phylogeny including some C. intestinalis haplotypes are key pieces of evidence that this locus was selected in C. robusta and then subsequently introgressed into C. intestinalis. Consider making these a combined main figure. The neighbour-joining tree could be presented as a simplified version of S8 (e.g. labelling the tips only by species/colour, not individual sample IDs). I could not find the SweepFinder results for C. robusta showing positive selection on chr 5, which should at least go into the supplement.

>>> We followed your suggestion by adding a panel (F) in Figure 4 that represents a simplified version of the star-like phylogeny shown in Figure S8. The SweepFinder results for *C. robusta* chr5 are now presented on a new panel (B) of Figure S6.

Please give the datasets more informative names, e.g. ‘phased SNP set’ for Dataset #1, ‘all parental SNPs’ for Dataset #2, and ‘ancestry informative SNPs’ for Dataset #3. You can still keep the numbers that correspond to the reference table at the end of the supplement; it’s just hard to keep track of in the main text when the datasets are only labelled by number.

>>> We followed your idea and gave a more descriptive label to the datasets (except in the M&M as they are described in an explicit way). We also added a Supplementary Table (S5) that summarizes the features of each dataset.

Please acknowledge the uncertainty around your 75 years estimate for the date of introgression (methods ~line 230). While a point estimate is useful, there are still many unknowns. Rapid rises in frequency due to selection can create longer tracts than neutral models. Additionally, the r used is an unpublished estimate of the average recombination rate genomewide. Local recombination could be much lower around the hotspot.

>>> Thanks for your suggestions. We now acknowledge both sources of uncertainties in the text: “Note that this point estimate for the date of introgression has to be considered carefully as several factors can produce uncertainty around it. For example, a rapid rise in frequency due to selection at the hotspot can create longer tracts than expected under neutral models. Additionally, we used the genome-wide recombination rate for *r*, while the local recombination could be lower around the hotspot. Finally, some introgressed tracts could be a bit longer than measured due to small regions lacking sufficient ancestry signal (Figure S7)”.

Figure 1: Please provide separate legend entries to distinguish intraspecific and interspecific lab hybrids visually and specify the cross. The legend should clearly indicate the number of individuals analyzed (not the number sampled). The figure description says “the F1s were considered as supplementary individuals in the PCA”, but the methods describe a regular PCA analysis with all 45 individuals treated the same. Please clarify.

>>> Thanks for your suggestions. Figure 1 has been modified to distinguish the intraspecific (black star) from the interspecific (black circle) hybrids. The number of individuals has been corrected to match the number analysed (n=45) and not the number sampled (n=48). The three individuals excluded from the analysis are indicated in Supp. Table 1. Regarding the usage of F1s as supplementary individuals in the PCA, this is an error - thanks for spotting it. We removed the sentence referring to it from the legend.

Figure 2: It’s hard to see the red arrows and how many there are. It could be more effective to colour the portion of each bar in the histogram that corresponds to tracts on chr 5.

>>> Thanks for the idea. We modified panels B and C, and their legend.

Figure 4E: If space allows, it would be clearer to label ‘8% SNPs iHS outliers’.

>>> Labels modified.

Figure 5: Really nice figure!

>>> Thanks!