

On the way to diploidization and unexpected ploidy in *Sporobolus* section *Spartina* (Poaceae) mesopolyploids

Corresponding Author: Professor Malika Ainouche

This file contains all reviewer reports in order by version, followed by all author rebuttals in order by version.

Version 0:

Reviewer comments:

Reviewer #1

(Remarks to the Author)

What are the noteworthy results?

A(answer): I think there are two relatively noteworthy results: (a) new ploidy level inferred for both species sequenced, (b) explaining how the chromosome difference originated (+1 or -1 chromosome pair).

Will the work be of significance to the field and related fields? How does it compare to the established literature? If the work is not original, please provide relevant references.

A: this work is similar to stories appearing now more and more frequently (some published by Nat Comm). the results are basically technology-drive (which is okay) - meaning that "the right ploidy" of these species and many other plant species with polyploidizations followed by diploidization was not known before the era of chromosome-level whole-genome assemblies. If I evaluate the present submission in comparison with other papers published recently and the impact of Nat Comm, I found the reviewed work as reporting an incremental knowledge gain.

As for the one-chromosome difference between the two *Spartina* species: would not be more parsimonious to consider descending dysploidy from $n = 31$ to $n=30$. I admit that I did not have enough time to analyze supplementary files and the claimed chromosomal rearrangement from $n=30$ to $n=31$ is not sufficiently explained/displayed in the main text. This rearrangement requires an extra centromere...

Does the work support the conclusions and claims, or is additional evidence needed?

A: yes. the work supports the conclusions.

Are there any flaws in the data analysis, interpretation and conclusions? Do these prohibit publication or require revision?

A: n/a.

Is the methodology sound? Does the work meet the expected standards in your field?

A: I would say "yes".

Is there enough detail provided in the methods for the work to be reproduced?

A: most likely yes.

Reviewer #2

(Remarks to the Author)

The manuscript by Salmon et al. addresses important evolutionary processes by using some of the most invasive species worldwide as a case study. As a field experimental ecologist, I suggest finding an expert to review the mat&mets in detail. Having said that, the approach seems correct, and the quality of the data and presentation is great. The research was conducted by a brilliant international team whose members have led the molecular and genetic research on *Spartina* species over the past decade worldwide. The central question framing the study (i.e., "[what are the] consequences of hybridization and WGD at various evolutionary time scales, in neo- and meso-polyploidy context?") is attractive and interesting, their results are novel, sound and convincing, and their reasoning is provoking, and it provides an unexpected turn in the existing evolutionary knowledge about the relevance of cyclical Whole Genome Duplication in the *Spartina* clade and its implications for biological invasions in general. Although the writing needs some improvement (particularly in the final paragraph), I think this research will likely trigger interesting debates and inspire/define future research.

I only have a few minor suggestions to make:

line 69- I suggest also mentioning reference #46 here, as it registers the spread of this same species across the Atlantic

shores of South America. Together with the China case, they represent the largest known cases of geographic invasion worldwide.

Lines 79-85. Merge small paragraphs (made of 1-2 sentences) into one to improve clarity.

Line 89. Please add a reference to support "as previously thought."

Lines 165-176. Please merge the first two paragraphs of the discussion section into one short(er) paragraph.

Line 273. Add here the reference #14 to support the statement about sterile hybrids

Line 279. The summary-paragraph includes several speculative statements and arguments. I suggest focusing more on your results and evidence presented, avoiding speculation around hypotheses or predictions. Also, I suggest avoiding convoluted rhetorical (e.g., "The fundamental understanding of...") and get straight to the point. Get rid of irrelevant statements like "This study opens" In line 284.

Line 289. Change "allows", for "enables" or "makes possible" to do more accurately important processes..."

Also in this section, I suggest highlighting the relevance of improving our understanding of the evolutionary history of hybrids to environmental management decision-making and legislation when/where hybrids are involved (see: Adam P. 2015. The shock of the new. Australasian P. Conserv.).

Line 296. The Plant Material section is correct, but the authors need to mention where the vouchers of the specimens used were deposited and under what ID code. If there are digital photo-records of those vouchers I suggest including them as suppl. info.

Reviewer #3

(Remarks to the Author)

The manuscript under review provides a new concept of the genome evolution history in *Sporobolus* sect. *Spartina*, a well-studied example of recent allopolyploid speciation. The authors have made a genome assembly for *S. maritimus*, the parental species to the allopolyploid *S. anglicus*, and have compared it with the other parental species *S. alterniflorus* whose genome was published earlier this year. Using genome collinearity and ortholog divergence analyses, they found traces of two shared whole-genome duplications that shaped these genomes, rendering them octoploid and not hexaploid as thought before. This analysis also clarified the chromosomal rearrangements that occurred between the two species since the last genome duplication; the functions of genes in rearranged regions suggested their possible role in the attenuation of the cell division machinery and ecological separation of the two species.

At the conceptual level, the manuscript communicates a solid piece of research with clear and valuable results that build upon the previous research in this system. Although the fourfold nature of "hexaploid" *S. alterniflorus* seems to have been hypothesized in the reference 41 earlier this year as shown by the synteny patterns and general appearance of figure 1C there, I did not find it being stated explicitly in that article so the claimed novelty of this result holds. The obtained results are well-interpreted, but some logical transitions can be polished further:

It is hard to tie the discovery of four "subgenomes" (first mentioned in line 126) with the hypothesized autopolyploidy (lines 208-210). The argumentation for this scenario does not seem convincing: firstly, if we assume one $2n=30$ lineage having existed, why could there not be two? Secondly, absence of structural variants between the subgenomes could also be seen as possible support for allopolyploidy because it means that cytological diploidization was possible without structural barriers for chromosome pairing (which should imply some sequence divergence). It seems hard to say if it was allo- or autopolyploidy from the provided results; I suggest either making the conclusions on lines 206-211 less specific or providing references to described cases of diploidized paleo-autopolyploids without structural differences between the "subgenomes". The repeated assumption of "rapid" separation of *S. alterniflorus* and *S. maritimus* following WGD2 (lines 145-146, then 212-215) does not seem convincingly supported by the available data, as all inferred datings are based on very rough estimates of mutation rates and are considerably uncertain. It seems likely that the WGD2 is shared between the two species from the genome collinearity and similar paralog divergence levels, but I think we cannot say confidently how close speciation was to this event from the provided results.

The research relies on state-of-the-art methods of comparative genomics that concordantly support the proposed main hypothesis. However, some choices and actions can be specified in more detail to ensure the reproducibility of the analyses:

Importantly, even though the study involved extensive computational work with numerous tools and some "home-made Python scripts" (lines 366, 372), the authors do not provide ways to examine the actual commands run on the data. The provided parameter values might be enough to reproduce most of the steps after some trial and error, but giving access to the ready-made code would save a lot of time for everyone curious about the study's methodology.

The claimed assembly of *S. maritimus* "at chromosome-scale" (line 96) might be seen as a miscommunication, as 19 out of 30 pseudochromosomes consist of several large scaffolds grouped together using only evidence from another species. This is not a flawless approach, as it might lead to reference bias stemming from the impossibility to detect translocations of regions corresponding to the scaffolds. Given this and the fact that some scaffold join sites stand out from the general centromere-to-telomere gradient of gene and repeat densities (e.g. the join of scaffolds 04 and 56 in Figure 1), I recommend to try validating these joins by predicting centromere and telomere positions in silico, e.g. using TRASH

(<https://github.com/vlothe/TRASH>) in case the centromeres are repeat-based and telomeric-identifier

(<https://github.com/tolkit/telomeric-identifier>, specifically `tidk search -s TTAGGG`). More than one centromeric repeat array in any chromosome and/or non-terminal positions of telomeric repeat occurrence peaks would suggest mis-scaffolding that has to be corrected.

The purpose of rerunning the k-mer analysis using a smaller k value as explained in lines 320-321 is not immediately clear.

The motivation for this run and the importance of the obtained result should be explained in more detail.

The GO terms enrichment analysis is not described in sufficient detail, there is only reference to the R package on line 376 and Supplementary Data 1 to guess the rest. What does “enrichment at the whole-genome scale” (line 650, Figure 4A left) mean? What was used as background data in this analysis? How less significant are the remaining tens and hundreds of significant GO terms from Figure 4A that are not in Figure 4B? All of this would be good to specify in the methods section. In addition, I could not find the “Supplementary table S2” referred to on lines 617 and 654, I guess it is Supplementary Data 1 now.

Although Supplementary Figure 2A is said to demonstrate “homogenous coverage among scaffolds” (line 106), it has at least two scaffolds (68 and 69) that clearly have below-average short read mapping depth. It also seems that they are missing from Figure 1. What are they, how big are they? It would be easier to understand in a picture that shows differences in scaffold sizes and GC content along with the coverage, like done by <https://github.com/DRL/blobtools>. Alternatively, one can quickly rule out the possibility of non-plant contamination in the assembly using the easy-taxonomy workflow of <https://github.com/soedinglab/MMseqs2> and considering the phylum assignment of contigs.

The text and figures act well together, communicating the research made, and yet there is room for further improvement here as well:

Figure 1 is communicating a lot of information and therefore needs additional guidance for the readers. The plot needs a legend allowing for quick deciphering of the track indices and scales for C-E heatmaps and the F barplot. The distinction between red and grey synteny tracks as well as (never explained explicitly) blue, green and purple homology groups can be also put into that legend. Additionally, the subgenomes A and B can be shown in the plot as boxes or shaded background portions of the circular plot.

The divergence of the subgenomes revealed by Ks analysis is mentioned in the manuscript a few times, but none of the main text figures demonstrate it as clearly as any panel of Supplementary Figure 5, with panel 5C being probably the most informative. I suggest including this panel into a main text figure, e.g. as an additional panel for Figure 2. It would be also better to have a legend for colors in these dotplots.

The ecological differences between the two *Sporobolus* species summarized in lines 235–245 will be easier to understand if the rich descriptions of both species provided there are boiled down to the actual differences. In addition, outcrossing in *S. alterniflorus* does not seem to contradict “self-sterility” in *S. maritimus*.

Minor text issues suggesting a thorough read-through by co-authors is needed:

L90 ca.

L102 Gb

L104 with 1.59 Gb “of” contigs

L120 The satellite repeat ... appeared mostly distributed

L144 MYA

L168 “...illuminated our understanding ...” sounds a bit too dramatic

L196 making the genomes... appear as meso-octoploids

L213 divergence time

None of the abovementioned imperfections of this manuscript undermine the significance of the research, and I think this article fits the interests of Nature Communications' readership well.

Reviewer #4

(Remarks to the Author)

I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

Version 1:

Reviewer comments:

Reviewer #2

(Remarks to the Author)

Dear Editor-in-Chief,

I've read the new corrected version of this manuscript, and I believe all issues raised by me and the other reviewers have been properly addressed. Considering the thorough revision and improvements made by the authors in this new version, I am glad to support its acceptance for publication. I think this work will make a great contribution and trigger further interesting debate in the field.

Reviewer #4

(Remarks to the Author)

I thank the authors for the effort put into further improving and polishing the manuscript. I am happy with a less decisive interpretation of the origins of ancient polyploidy and the timeline of diversification. The analysis of tandem repeats with TRASH seems to have supported some of the scaffold joins made in the *S. maritimus* assembly, which is great. I list the remaining concerns below in the order of decreasing importance.

I do not see any improvement in code availability except for the statement pointing to the parameter values and software citations provided earlier in the Methods section. Therefore, I am very much looking forward to having a link to the promised code repository at later stages of the publication preparation, preferably before the issue is out.

The additional explanations provided were mostly helpful, but I would like to request a few further clarifications in the enrichment part. As I understand it now from the response letter, the union of whole-genome GO terms from *S. maritimus* and *S. alterniflorus* together was used as background data for the enrichment analyses (a) of each assembly as a whole and (b) in the rearranged regions of both assemblies. Four enrichment analyses overall, all use the same background data, right? This makes sense, but the wording in the methods still does not communicate it very clearly. The results of analysis (a) seem to stay unused in the article except in lines 165–166 (“...majority of GO terms are shared”) where the figure is not even referenced; they are also missing from the table in Supplementary Data 1. Would it not then be easier to not mention them and just keep the report of enrichment in the rearranged regions? Therefore, I suggest removing the left part of panel 4A and words “and whole genome” from line 405 as these results do not seem informative. Instead, it would be good to indicate the enrichment significance (corrected p-values, “p-adjusted” in supplementary table) next to each table row in 4B. The two last sentences added in lines 406–409 are a basic explanation of the enrichment analysis and can be omitted. Instead, it would be good to add more technical details about the clusterProfiler, i.e. what databases were used for functional annotation (do they have to be cited?) and what were the run parameters. It would be best to have the R code published along with all other code.

Concerning the visualization suggestions, there have probably been a few miscommunications in my last review. I anticipated having additional visual guidance added as a graphical legend, not just further elaboration of the textual caption. Ideally, Figure 1 should be understandable without the caption. My suggestions on this way were (a) short textual descriptions instead of letters A–F in the plot (e.g. “*S. alterniflorus* chr.”, “*S. maritimus*:\nchr.”, “Repeats”, “Copia”, “Gypsy”, “Genes”), (b) a graphical legend next to the plot that explains what blue, purple and green colors in the outer track as well as red and grey color in the innermost link track mean and (c) adding boxes encircling subgenomes A and B of both assemblies under the circos plot. I would be happy if the authors addressed these initial suggestions.

Reading the manuscript a few more times, I noticed another minor inconsistency. The range of WGD2 age is given at different precision in different places of the manuscript (“2–6 MYA”: lines 43,92,193,219,248,303,653,693; “2.6–6.8 MYA”: lines 151,225; “2.08–6.23 MYA”: line 143) – which one is used when interpreting the results? The estimate for “2.6–6.8 MYA” seems to be applied jointly to the WGD and the divergence of two species (line 152) – does it not contradict other dating estimates? Is it possible to use a single range throughout the discussion?

Among the minor corrections suggested earlier, a few points were missed by the authors:

L92,120,653,693: the abbreviation of “circa” is correctly typed as “ca.” or “c.”, not “c.a.”

L105: the suggestion was “1.59 Gb of contigs”, not “1.59 of Gb contigs”

L129–130: The full sentence corrected as I suggest would look like

The 370-bp satellite repeat, more abundant in *S. alterniflorus* than in *S. maritimus* (Supplementary Table 1), appeared mostly distributed in distal chromosomal regions (Supplementary Fig. 5).

instead of

The distribution of the 370-bp satellite repeat, more abundant in *S. alterniflorus* than in *S. maritimus* (Supplementary Table 1), appeared mostly distributed in distal chromosomal regions (Supplementary Fig. 5).

L146: I suggested changing “MYs” to “MYA” only in this particular sentence, not in the entire document. The newly placed “MYA” on lines 90 and 210 have to be corrected back to “MY” or “MYs”.

I suggest incorporating these truly minor corrections before publishing the article, putting a special emphasis on making the code publicly available as early as possible. It does not have to be a production-level software, just the actual commands run on actual data would be already good enough.

Reviewer #5

(Remarks to the Author)

I joined during the second round of review with a special task to comment on the authors’ responses to the previous Reviewer #1’s suggestions. I can see that the previous Reviewer #1 questioned the novelty of the authors’ findings and asked a critical question regarding the evolutionary scenario of chromosome number variation in $n=31$ and $n=30$ genomes.

First, regarding the novelty of this work, the authors didn’t respond to Reviewer #1’s comment. In particular, the reviewer mentioned that there were similar cases reported recently and the current study represents an incremental knowledge gain. However, it seems that the authors were not aware of similar studies in other plants.

Second, as for the karyotype evolution, I don't think the comment was properly addressed by the authors. Reviewer #1 was asking whether +1 or -1 chromosome pairs is the more parsimonious scenario, which is a very important conclusion in the current study. In my opinion, Reviewer #1's comment is concerning on the reasoning process. The authors only indicated some words in their discussion, which would not help to strengthen their argument. Although I can see some comparisons in the figures, those are built on the existing hypothesis, not reasoning. In addition, the authors replied that "... it would have involved more chromosomal restructuring events, and would have resulted in more structural differences between the two species". I can't agree with this opinion because there is no information from the ancestral genome (which is currently unknown).

Third, regarding the new centromere, I think Reviewer #1 was still questioning on the proposed scenario of karyotype evolution. Although the authors mentioned their efforts to search for centromeric and telomeric repeats, this only corroborated the completeness and/or correctness of the assembly. In case of the birth of new chromosomes, the current discussion might benefit from the fate of 'old' centromeres in the n=30 genome. For example, were 30 out of the 31 centromeres in n=31 genome inherited from the n=30 genome?

Version 2:

Reviewer comments:

Reviewer #4

(Remarks to the Author)

I thank the authors for another round of corrections that improved the manuscript. I am largely happy with the introduced changes — the GO analysis methodology became clearer, and Figure 1 is more readable than before.

With the additional information provided in Figure 4, it seems that the only significantly enriched GO terms are those from Biological Processes and Molecular Function private to the restructured regions of *S. maritimus* — is that correct? Is it then necessary to mention the non-significant GO terms, especially given that they are not discussed in the manuscript? Maybe Figure 5 can be then reduced to the Venn diagrams with callouts to significant *S. maritimus* BP and MF terms? It was not obvious in previous versions of the figure because the authors do not provide the adjusted p-values for any GO terms mentioned.

Unfortunately, I have to say that I am still unhappy with the code availability. Imagine a person wants to redraw the Figure 1 — what would they need to do? At the current state of the manuscript, doing this would involve quite a number of wheel reinventions, since some scripts that are mentioned or implied in the Methods are not shared. What I dream of is a code repository that will bring that person from actual raw data generated for this article (or intermediate files if the authors would like to share them in that repo) to the actual results and figures presented in the manuscript. Sharing the GO code and `WGD\_Tracker` is a good step on this way, even though it is obvious that the latter pipeline is raw and not ready to be easily run by outsiders for a number of reasons (SIF files are missing, no easy way to install dependencies etc.) — I accept the authors' promise to improve and publish WGD_Tracker properly.

I recognize that I might be annoying with the point of code availability, as this is going there and back a few times already. Therefore, if the authors think that this request is unreasonable and making a repository with (1) scripts tracking ALL data manipulations used for presented results and (2, preferably) an easy way to recreate the run environment (a Singularity container?) is overly hard, they should feel free to defend this stance in the response letter. I will not be fighting over it to death, but I genuinely do not understand why such a repository cannot be made in the course of manuscript preparation given that the scripts are presumably somewhere around.

Having said that, I should reiterate that I like the presented story as a whole and do not want to hold it more than it is needed to make the results transparent and (easily) reproducible. I thank the authors for their patience and understanding.

P.S. one tiny correction in the newly introduced text — on line 454, "are" should be replaced by "is".

(Remarks on code availability)

The results of the paper cannot be reproduced with the given code repository. The examples section contains an unrelated dataset; the code cannot be run as the Singularity container files are missing and the paths are kept as in the original (and inaccessible) run environment; no easy way to install the dependencies (e.g. container/environment file) is provided. The repository does not cover all data manipulations done for the manuscript.

Reviewer #5

(Remarks to the Author)

The authors fully addressed my previous questions. I have no further comments.

(Remarks on code availability)

The page is not accessible yet.

Version 3:

Reviewer comments:

Reviewer #4

(Remarks to the Author)

I thank the authors for their patience with me through several rounds of review and for considering all of my comments to the manuscript. I do not see any important flaws in the manuscript, and it can be published as is. As always, there is room for improvement, and my (minute) suggestions are given below.

With the new round of clarifications, I have finally understood the rationale behind the GO analysis being presented as is, and I agree with keeping Figure 4 in its current form. As I understand it now, the authors also consider non-enriched GO categories (the ones without asterisk next to them in Figure 4) and compare the full GO term spectra in the restructured regions of the two species to justify their functional divergence. Indeed, with this kind of analysis, displaying non-significant terms is appropriate, and background data is not so important. However, it can be hardly called an "enrichment analysis", as (to my knowledge) this term implies looking at enriched categories and considering other differences insignificant.

If the authors would like to make this analysis more understandable for broader scientific community including myself, I suggest to clarify the analysis and specify the results a bit more. My suggestion for replacement of the sentence on lines 171–175 is: "Compared to the whole genome where the majority of the GO terms are shared, the restructured regions demonstrate functional divergence between the two species (Fig. 4). In addition, the restructured region of *S. maritimus** appeared to be enriched in several biological functions of adaptive interest such as reproduction (pollen tube development), defense (chalcone biosynthetic process) or stress tolerance (lipid biosynthetic process)."

Regarding the code availability, I agree that the newly updated repository allows for reproducing most of the important results discussed in the manuscript. I have successfully run all of the analysis steps using the provided code and config files. In a better world, scientists would publish all of the code they used for producing figures etc., much like they do publish all laboratory protocols even when they are not describing new methods. However, after a look at other papers published in Nature Communications I agree that this manuscript deserves being accepted as is.

Congratulations on the well-done research and happy holidays!

(Remarks on code availability)

The instructions for installing the environment are complete and easy to follow, they work for Linux but not on MacOS. The code runs perfectly upon installation and allows to reproduce the source datasets used for making a portion of the figures in the manuscript.

permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

In cases where reviewers are anonymous, credit should be given to 'Anonymous Referee' and the source.

The images or other third party material in this Peer Review File are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

To view a copy of this license, visit <https://creativecommons.org/licenses/by/4.0/>

RESPONSES TO THE REVIEWER'S COMMENTS

We appreciated the valuable and constructive comments of the reviewers, and made changes in the revised version of the manuscript accordingly. Please see our point by point responses below.

REVIEWER COMMENTS

Reviewer #1 (Remarks to the Author):

What are the noteworthy results?

A(answer): I think there are two relatively noteworthy results: (a) new ploidy level inferred for both species sequenced, (b) explaining how the chromosome difference originated (+1 or -1 chromosome pair).

Will the work be of significance to the field and related fields? How does it compare to the established literature? If the work is not original, please provide relevant references.

A: this work is similar to stories appearing now more and more frequently (some published by Nat Comm). the results are basically technology-drive (which is okay) - meaning that "the right ploidy" of these species and many other plant species with polyploidizations followed by diploidization was not known before the era of chromosome-level whole-genome assemblies. If I evaluate the present submission in comparison with other papers published recently and the impact of Nat Comm, I found the reviewed work as reporting an incremental knowledge gain.

As for the one-chromosome difference between the two *Spartina* species: would not be more parsimonious to consider descending dysploidy from $n = 31$ to $n=30$.

RESPONSE: In this study we show that the two species share two WGD events; the first one led to the evolution from $n=10$ to $n=20$; this was followed by a base chromosome number reduction to $n=15$ as evidenced by the 15 homoeologs we detect as duplicated (by WGD2) in both species. This strongly suggests that these genomes evolved from $n=20$ to $n = 15$ (duplicated into $n=30$), and that $n=31$ evolved from $n=30$ following the chromosomal restructuring described in this paper. Considering an alternative hypothesis (descending dysploidy from $n=31$ to $n=30$), is less parsimonious as it would have involved more chromosomal restructuring events, and would have resulted in more structural differences between the two species. This is now indicated in the Discussion of the revised manuscript (Page 7, Lines 234-240).

I admit that I did not have enough time to analyze supplementary files and the claimed chromosomal rearrangement from $n=30$ to $n=31$ is not sufficiently explained/displayed in the main text. This rearrangement requires an extra centromere...

RESPONSE: As suggested, we performed an *in silico* centromere and telomere detection (see similar comment from reviewer 3 and reply). Our results indicate that this extra chromosome originated from a recombination/rearrangement between 2 chromosomes leading to its excision. We rephrased the explanation (Page 5, Lines 161-163; Fig. 1 legend, Page 19, Lines 669-670).

Does the work support the conclusions and claims, or is additional evidence needed?

A: yes. the work supports the conclusions.

Are there any flaws in the data analysis, interpretation and conclusions? Do these prohibit publication or require revision?

A: n/a.

Is the methodology sound? Does the work meet the expected standards in your field?

A: I would say "yes".

Is there enough detail provided in the methods for the work to be reproduced?

A: most likely yes.

Reviewer #2 (Remarks to the Author):

The manuscript by Salmon et al. addresses important evolutionary processes by using some of the most invasive species worldwide as a case study. As a field experimental ecologist, I suggest finding an expert to review the methods in detail. Having said that, the approach seems correct, and the quality of the data and presentation is great. The research was conducted by a brilliant international team whose members have led the molecular and genetic research on *Spartina* species over the past decade worldwide. The central question framing the study (i.e., “[what are the] consequences of hybridization and WGD at various evolutionary time scales, in neo- and meso-polyploidy context?”) is attractive and interesting, their results are novel, sound and convincing, and their reasoning is provoking, and it provides an unexpected turn in the existing evolutionary knowledge about the relevance of cyclical Whole Genome Duplication in the *Spartina* clade and its implications for biological invasions in general. Although the writing needs some improvement (particularly in the final paragraph), I think this research will likely trigger interesting debates and inspire/define future research.

I only have a few minor suggestions to make:

line 69- I suggest also mentioning reference #46 here, as it registers the spread of this same species across the Atlantic shores of South America. Together with the China case, they represent the largest known cases of geographic invasion worldwide.

RESPONSE: We agree to highlight the intercontinental scale of this species spread, and cited the reference 46 (updated as #26) in the main text (Page 2, Lines 69-70).

Lines 79-85. Merge small paragraphs (made of 1-2 sentences) into one to improve clarity.

RESPONSE: We revised accordingly (Page 3, Lines 80-87).

Line 89. Please add a reference to support “as previously thought.”

RESPONSE: References are now provided (Page 3, Line 91).

Lines 165-176. Please merge the first two paragraphs of the discussion section into one short(er) paragraph.

RESPONSE: We combined the two short paragraphs (Page 5, Lines 172-183).

Line 273. Add here the reference #14 to support the statement about sterile hybrids

RESPONSE: We added the reference 14 accordingly (Page 8, Lines 291).

Line 279. The summary-paragraph includes several speculative statements and arguments. I suggest focusing more on your results and evidence presented, avoiding speculation around hypotheses or predictions. Also, I suggest avoiding convolute rhetorical (e.g., “The fundamental understanding of...”) and get straight to the point. Get rid of irrelevant statements like “This study opens” In line 284.

RESPONSE: We rephrased the paragraph accordingly (Page 8, Lines 300-301 and Line 306).

Line 289. Change “allows”, for “enables” or “makes possible” to do more accurately important processes...”

RESPONSE: These changes have been made (Page 9, Line 309).

Also in this section, I suggest highlighting the relevance of improving our understanding of the evolutionary history of hybrids to environmental management decision-making and legislation when/where hybrids are involved (see: Adam P. 2015. The shock of the new. Australasian P. Conserv.).

RESPONSE: We highlighted this point as suggested (Page 8, Lines 300-301; Page 9 Lines 311-316).

Line 296. The Plant Material section is correct, but the authors need to mention where the vouchers of the specimens used were deposited and under what ID code. If there are digital photo-records of those vouchers I suggest including them as suppl. info.

RESPONSE: The sample S13 that selected to assemble reference genome of *S. alterniflorus* is growing in the greenhouse at Fudan university, and its voucher ID is USSA13 (Page 9, Line

326). *Sporobolus maritimus* voucher ID is RENO15692 (Page 9, Line 323).

Reviewer #3 (Remarks to the Author):

The manuscript under review provides a new concept of the genome evolution history in *Sporobolus* sect. *Spartina*, a well-studied example of recent allopolyploid speciation. The authors have made a genome assembly for *S. maritimus*, the parental species to the allopolyploid *S. anglicus*, and have compared it with the other parental species *S. alterniflorus* whose genome was published earlier this year. Using genome collinearity and ortholog divergence analyses, they found traces of two shared whole-genome duplications that shaped these genomes, rendering them octoploid and not hexaploid as thought before. This analysis also clarified the chromosomal rearrangements that occurred between the two species since the last genome duplication; the functions of genes in rearranged regions suggested their possible role in the attenuation of the cell division machinery and ecological separation of the two species.

At the conceptual level, the manuscript communicates a solid piece of research with clear and valuable results that build upon the previous research in this system. Although the fourfold nature of “hexaploid” *S. alterniflorus* seems to have been hypothesized in the reference 41 earlier this year as shown by the synteny patterns and general appearance of figure 1C there, I did not find it being stated explicitly in that article so the claimed novelty of this result holds. The obtained results are well-interpreted, but some logical transitions can be polished further:

It is hard to tie the discovery of four “subgenomes” (first mentioned in line 126) with the hypothesized autopolyploidy (lines 208–210). The argumentation for this scenario does not seem convincing: firstly, if we assume one $2n=30$ lineage having existed, why could there not be two? Secondly, absence of structural variants between the subgenomes could also be seen as possible support for allopolyploidy because it means that cytological diploidization was possible without structural barriers for chromosome pairing (which should imply some sequence divergence).

It seems hard to say if it was allo- or autopolyploidy from the provided results; I suggest either making the conclusions on lines 206–211 less specific or providing references to described cases of diploidized paleo-autopolyploids without structural differences between the “subgenomes”.

RESPONSE: We thank the reviewer for this comment; ascertaining auto or allopolyploidy in paleo (or even meso-) polyploids is particularly difficult, regarding the various potential evolutionary fates of these genomes. We have modified the text accordingly (Page 6, Lines 215–217; Lines 220–224).

The repeated assumption of “rapid” separation of *S. alterniflorus* and *S. maritimus* following WGD2 (lines 145–146, then 212–215) does not seem convincingly supported by the available data, as all inferred datings are based on very rough estimates of mutation rates and are considerably uncertain. It seems likely that the WGD2 is shared between the two species from the genome collinearity and similar paralog divergence levels, but I think we cannot say confidently how close speciation was to this event from the provided results.

RESPONSE: We agree that molecular dating provides rough estimations, and we modulated the assumption of “rapid speciation” following WGD2 in the corresponding text (Page 6, Lines 226–227); we should however point out that similar “time windows” are detected for both events.

The research relies on state-of-the-art methods of comparative genomics that concordantly support the proposed main hypothesis. However, some choices and actions can be specified in more detail to ensure the reproducibility of the analyses:

Importantly, even though the study involved extensive computational work with numerous tools and some “home-made Python scripts” (lines 366, 372), the authors do not provide ways to examine the actual commands run on the data. The provided parameter values might be enough to reproduce most of the steps after some trial and error, but giving access to the ready-made code would save a lot of time for everyone curious about the study’s methodology.

RESPONSE: Code for data cleaning and analysis associated with the current submission is available as cited in the Material and Methods section. The home made subcodes were developed for formatting input-output files for the cited tools. We are planning to create a repository on github, but the full pipeline is not yet ready-made for the community.

The claimed assembly of *S. maritimus* “at chromosome-scale” (line 96) might be seen as a miscommunication, as 19 out of 30 pseudochromosomes consist of several large scaffolds grouped together using only evidence from another species. This is not a flawless approach, as it might lead to reference bias stemming from the impossibility to detect translocations of regions corresponding to the scaffolds. Given this and the fact that some scaffold join sites stand out from the general centromere-to-telomere gradient of gene and repeat densities (e.g. the join of scaffolds 04 and 56 in Figure 1), I recommend to try validating these joins by predicting centromere and telomere positions *in silico*, e.g. using TRASH (<https://github.com/vlothech/TRASH>) in case the centromeres are repeat-based and telomeric-identifier (<https://github.com/tolkitt/telomeric-identifier>, specifically tidk search -s TTTAGGG). More than one centromeric repeat array in any chromosome and/or non-terminal positions of telomeric repeat occurrence peaks would suggest mis-scaffolding that has to be corrected.

RESPONSE: We corrected the “chromosome-scale” statement in the manuscript (Page 3, Line 98-99). We added to the manuscript this sentence (Page 11, Lines 400-403) to raise that we cannot exclude other restructuring in these regions: "It is important to note that since the *S. maritimus* pseudo-chromosomes were constructed based on the *S. alterniflorus* genome assemblies, we cannot exclude the possibility that recombination occurs at scaffold breaks, and subsequently need to be cautious when examining these regions.."

We thank reviewers for their recommendations for validating these joins by centromere and telomere *in silico* detection. As suggested, We performed data analysis to identify centromeres (with TRASH - not that straightforward to install and run...(Page 10, Lines 369-371)) and telomeres (with TIDK). We identified 21 telomeres from the 31 chromosomes in the *S. alterniflorus* genome and 12 telomeres from the 69 chromosomes (data not shown). The tandem repeat annotation carried out with TRASH allowed the detection of a High Order Repeat (HOR) of 147bp shared by the two species within *S. alterniflorus* chromosomes and *S. maritimus* scaffolds. A second HOR of 169bp maps at chromosomes and scaffolds distal regions. As suspected thanks to the synteny analysis, longer *S. maritimus* scaffolds exhibit internal 147bp and distal 169bp signals, while shorter scaffolds bear either both or single or no signal(s). Interestingly, some scaffolds appear to assemble single chromosome arms (from a distal/telomeric region to putative centromeric repeats). We described newly produced results (Page 4, Lines 121-127); added a supplemental Figure describing the location of HORs for both species (Supplemental Figure 4, Page 27); and discuss them (Page 8, Lines 296-298). Notably we highlight the finding that *S. alterniflorus* chromosome 31 (resulting from ancestral chromosomal restructuring) appears depauperate in the 147 bp putative centromeric repeat.

The purpose of rerunning the k-mer analysis using a smaller k value as explained in lines 320–321 is not immediately clear. The motivation for this run and the importance of the obtained result should be explained in more detail.

RESPONSE: Using GenomeScope, we tested different k-mer values to reach a genome size fitting with flow cytometry estimations, the longer k-mers are, the larger the genome size estimation. We added a sentence in the manuscript (Pages 9-10, Lines 343-347) to precise the aim of this approach.

The GO terms enrichment analysis is not described in sufficient detail, there is only reference to the R package on line 376 and Supplementary Data 1 to guess the rest. What does “enrichment at the whole-genome scale” (line 650, Figure 4A left) mean? What was used as background data in this analysis? How less significant are the remaining tens and hundreds of significant GO terms from Figure 4A that are not in Figure 4B? All of this would be good to specify in the methods section.

RESPONSE: We added some details about GO terms enrichment analysis (Page 11, Lines 405-409). "enrichment at the whole-genome scale" refers to conducting a GO term enrichment analysis on all genes in the genomes of *S. alterniflorus* and *S. maritimus*, respectively. The purpose is to use the results of the GO analysis at the whole-genome level as a reference to see if there is any specificity in the GO enrichment of the recombined fragments. We selected some representative GO terms to be displayed in Figure 4B; if we were to include too many, the table would become very crowded.

In addition, I could not find the “Supplementary table S2” referred to on lines 617 and 654, I guess it is Supplementary Data 1 now.

RESPONSE: Thank you for noticing this MS-formatting issue, edited ([Page 18, Line 661](#); [Page 22, Line 704](#))

Although Supplementary Figure 2A is said to demonstrate “homogenous coverage among scaffolds” (line 106), it has at least two scaffolds (68 and 69) that clearly have below-average short read mapping depth. It also seems that they are missing from Figure 1. What are they, how big are they? It would be easier to understand in a picture that shows differences in scaffold sizes and GC content along with the coverage, like done by <https://github.com/DRL/blobtools>. Alternatively, one can quickly rule out the possibility of non-plant contamination in the assembly using the easy-taxonomy workflow of <https://github.com/soedinglab/MMseqs2> and considering the phylum assignment of contigs.

RESPONSE: We thank the reviewer for spotting this error. It was a problem with the limits of the y-axis. The two scaffolds, 68 and 69, contain very few windows, which explains the higher variability in coverage. However, scaffold 69 is highly repetitive and contains few genes, which is compatible with higher coverage. This is now evident after increasing the limit of the y-axis. We updated the Supplementary Figure 2A and added points for scaffolds that contain fewer than 15 windows of 50Kb."

The text and figures act well together, communicating the research made, and yet there is room for further improvement here as well:

Figure 1 is communicating a lot of information and therefore needs additional guidance for the readers. The plot needs a legend allowing for quick deciphering of the track indices and scales for C-E heatmaps and the F barplot. The distinction between red and grey syntenic tracks as well as (never explained explicitly) blue, green and purple homology groups can be also put into that legend. Additionally, the subgenomes A and B can be shown in the plot as boxes or shaded background portions of the circular plot.

The divergence of the subgenomes revealed by Ks analysis is mentioned in the manuscript a few times, but none of the main text figures demonstrate it as clearly as any panel of Supplementary Figure 5, with panel 5C being probably the most informative. I suggest including this panel into a main text figure, e.g. as an additional panel for Figure 2. It would be also better to have a legend for colors in these dotplots.

RESPONSE: We updated Figures 1 & 2 as suggested.

The ecological differences between the two *Sporobolus* species summarized in lines 235–245 will be easier to understand if the rich descriptions of both species provided there are boiled down to the actual differences. In addition, outcrossing in *S. alterniflorus* does not seem to contradict “self-sterility” in *S. maritimus*.

RESPONSE: In this paragraph, we focused on the actual differences between species; we removed the “in contrast” words in the revised version to avoid confusion regarding breeding systems .

Minor text issues suggesting a thorough read-through by co-authors is needed:

L90 ca.

L102 Gb

L104 with 1.59 Gb “of” contigs

L120 The satellite repeat ... appeared mostly distributed

L144 MYA

L168 “...illuminated our understanding ...” sounds a bit too dramatic

L196 making the genomes... appear as meso-octoploids

L213 divergence time

RESPONSE:All these sentences were edited, as suggested

None of the abovementioned imperfections of this manuscript undermine the significance of the research, and I think this article fits the interests of Nature Communications' readership well.

Reviewer #4 (Remarks to the Author):

I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

Thanks again for all the comments, which greatly improved the manuscript.

We hope this improved manuscript is now suitable for publication in *Nature Communications*.

Thank you for your consideration.

Sincerely,

Malika AINOUCHE on behalf co-authors

RESPONSES TO THE REVIEWER'S COMMENTS

Please see below our point-by-point responses to reviewers. The requested changes have been made in this new revised version (apparent in the resubmitted manuscript) as following:

Reviewer #2 (Remarks to the Author):

Dear Editor-in-Chief,

I've read the new corrected version of this manuscript, and I believe all issues raised by me and the other reviewers have been properly addressed. Considering the thorough revision and improvements made by the authors in this new version, I am glad to support its acceptance for publication. I think this work will make a great contribution and trigger further interesting debate in the field.

ANSWER: Reviewer # 2 is thanked for the encouragement and earlier valuable comments that helped improve the initial version of this manuscript.

Reviewer #4 (Remarks to the Author):

I thank the authors for the effort put into further improving and polishing the manuscript. I am happy with a less decisive interpretation of the origins of ancient polyploidy and the timeline of diversification. The analysis of tandem repeats with TRASH seems to have supported some of the scaffold joins made in the *S. maritimus* assembly, which is great. I list the remaining concerns below in the order of decreasing importance.

I do not see any improvement in code availability except for the statement pointing to the parameter values and software citations provided earlier in the Methods section. Therefore, I am very much looking forward to having a link to the promised code repository at later stages of the publication preparation, preferably before the issue is out.

ANSWER: We are providing the requested code availability (WGD_Tracker pipeline (Link: https://github.com/MorganeMilin/WGD_Tracker) as indicated in sections Materials and Methods (lines 451-455) and Code Availability (496) that will be accessible at the later stage of publication. An archive of this pipeline was uploaded with this submission, containing a short description of the pipeline, the full set of codes, how to install and run it, including a demo analysis.

The additional explanations provided were mostly helpful, but I would like to request a few further clarifications in the enrichment part. As I understand it now from the response letter, the union of whole-genome GO terms from *S. maritimus* and *S. alterniflorus* together was used as background data for the enrichment analyses (a) of each assembly as a whole and (b) in the rearranged regions of both assemblies. Four enrichment analyses overall, all use the same background data, right? This makes sense, but the wording in the methods still does not communicate it very clearly.

ANSWER: We thank the reviewer for suggesting to clarify the text. Yes, we did four enrichment analyses for the two species. For each of the two species, we performed GO enrichments for the whole genome and rearranged region, respectively. We did the revision accordingly (Page 12, Lines 459-463).

The results of analysis (a) seem to stay unused in the article except in lines 165–166 (“...majority of GO terms are shared”) where the figure is not even referenced; they are also missing from the table in Supplementary Data 1. Would it not then be easier to not mention them and just keep the report of enrichment in the rearranged regions? Therefore, I suggest removing the left part of panel 4A and words “and whole genome” from line 405 as these results do not seem informative. Instead, it would be good to indicate the enrichment significance (corrected p-values, “p-adjusted” in supplementary table) next to each table row in 4B. The two last sentences added in lines 406–409 are a basic explanation of the enrichment analysis and can be omitted. Instead, it would be good to add more technical details about the clusterProfiler, *i.e.*, what databases were used for functional annotation (do they have to be cited?) and what were the run parameters. It would be best to have the R code published along with all other code.

ANSWER: We have removed the left part of panel 4A (see details in the new Figure 4A) and deleted the last two sentences. In brief, functional annotation of the protein-coding genes was performed based on the database TrEMBL, eggNOG and InterPro. Then, the annotated GO terms of the two species were applied to perform GO enrichment analyses using clusterProfiler (Wu et al., 2021). Per this comment, we rephrased the related content in the revised version of our manuscript (Page 12, Lines 459-463). We have submitted the codes to github (https://github.com/tytrhr/GO_enrichment.git).

Concerning the visualization suggestions, there have probably been a few miscommunications in my last review. I anticipated having additional visual guidance added as a graphical legend, not just further elaboration of the textual caption. Ideally, Figure 1 should be understandable without the caption. My suggestions on this way were (a) short textual descriptions instead of letters A–F in the plot (e.g. “*S. alterniflorus* chr.”, “*S. maritimus*:\nchr.”, “Repeats”, “Copia”, “Gypsy”, “Genes”), (b) a graphical legend next to the plot that explains what blue, purple and green colors in the outer track as well as red and grey color in the innermost link track mean and (c) adding boxes encircling subgenomes A and B of both assemblies under the circos plot. I would be happy if the authors addressed these initial suggestions.

ANSWER: We thank the reviewer for these suggestions and we have made the corresponding changes in the upgraded Figure 1.

Reading the manuscript a few more times, I noticed another minor inconsistency. The range of WGD2 age is given at different precision in different places of the manuscript (“2–6 MYA”: lines 43,92,193,219,248,303,653,693; “2.6–6.8 MYA”: lines 151,225; “2.08–6.23 MYA”: line 143) – which one is used when interpreting the results? The estimate for “2.6–6.8 MYA” seems to be applied jointly to the WGD and the divergence of two species (line 152) – does it not contradict other dating estimates? Is it possible to use a single range throughout the discussion?

ANSWER: We homogenized the time range accordingly: The text in the results has been clarified (lines 42, 43; 142-146; 153-154; 198-199; 237; 243; 300; 356-357; 739; 774) to distinguish more clearly the different values provided by our analyses. The text has also been homogenized to retain only the following values: 2.1-6.2 MYs corresponding to the estimate of WGD2 event (These ranges were determined by the lowest and highest values encountered in *S. maritimus* and/or *S. alterniflorus*), and 2.6-6.8 MYs corresponding to the divergence between the two *Sporobolus* species.

Among the minor corrections suggested earlier, a few points were missed by the authors:

L92,120,653,693: the abbreviation of “circa” is correctly typed as “ca.” or “c.”, not “c.a.”

L105: the suggestion was “1.59 Gb of contigs”, not “1.59 of Gb contigs”

L129–130: The full sentence corrected as I suggest would look like

*The 370-bp satellite repeat, more abundant in *S. alterniflorus* than in *S. maritimus* (Supplementary Table 1), appeared mostly distributed in distal chromosomal regions (Supplementary Fig. 5).*

instead of

The distribution of the 370-bp satellite repeat, more abundant in *S. alterniflorus* than in *S. maritimus* (Supplementary Table 1), appeared mostly distributed in distal chromosomal regions (Supplementary Fig. 5).

L146: I suggested changing “MYs” to “MYA” only in this particular sentence, not in the entire document. The newly placed “MYA” on lines 90 and 210 have to be corrected back to “MY” or “MYs”.

I suggest incorporating these truly minor corrections before publishing the article, putting a special emphasis on making the code publicly available as early as possible. It does not have to be a production-level software, just the actual commands run on actual data would be already good enough.

ANSWER: We thank the reviewer for taking time to list these helpful suggestions, and the corresponding corrections have now been performed. We also produced the “WGD-tracker” pipeline that will be available at the time of publication as requested above (https://github.com/MorganeMilin/WGD_Tracker).

Reviewer #5 (Remarks to the Author):

I joined during the second round of review with a special task to comment on the authors’ responses to the previous Reviewer #1’s suggestions. I can see that the previous Reviewer #1 questioned the novelty of the authors’ findings and asked a critical question regarding the evolutionary scenario of chromosome number variation in $n=31$ and $n=30$ genomes.

First, regarding the novelty of this work, the authors didn’t respond to Reviewer #1’s comment. In particular, the reviewer mentioned that there were similar cases reported recently and the current study represents an incremental knowledge gain. However, it seems that the authors were not aware

of similar studies in other plants.

ANSWER: We now listed several studies reporting diploidization through dysploidy in different plant families, including grasses (lines 218-225). Moreover, we highlighted the knowledge gap in *Sporobolus* and the contribution of our study (lines 225-229) as following “In *Sporobolus*, although the various basic chromosome numbers reported in the literature (Peterson et al. 2014) suggest chromosomal restructuring events, this is the first time to our knowledge, descending dysploidy is documented at the genomic scale in this genus containing worldwide invasive species and most particularly in section *Spartina* where it seems to have occurred between WGD1 and WGD2, thus reshuffling the ancestral genome”.

Second, as for the karyotype evolution, I don't think the comment was properly addressed by the authors. Reviewer #1 was asking whether +1 or -1 chromosome pairs is the more parsimonious scenario, which is a very important conclusion in the current study. In my opinion, Reviewer #1's comment is concerning on the reasoning process. The authors only indicated some words in their discussion, which would not help to strengthen their argument. Although I can see some comparisons in the figures, those are built on the existing hypothesis, not reasoning. In addition, the authors replied that “... it would have involved more chromosomal restructuring events, and would have resulted in more structural differences between the two species”. I can't agree with this opinion because there is no information from the ancestral genome (which is currently unknown).

ANSWER: We now present the reasoning and discuss in more detail (lines 252-291) alternatively the “+1 hypothesis” (n=30 ancestral) and the “-1 hypothesis” (n=31 ancestral) with possible corresponding scenario. Moreover, we provided a new Supplementary Figure 10, to illustrate such scenario. Indeed, no information is available from the ancestral genome before divergence of *S. maritimus* and *S. alterniflorus* and no extant sister lineage (with n=30 or n=31) is available that could have (as outgroup) provided hypotheses regarding directionality of the chromosome number difference between the analyzed genomes. However, according to our results, the most parsimonious scenario more readily supports the “+1” hypothesis as discussed. We hope to have improved the clarity of our reasoning in this revised version.

Third, regarding the new centromere, I think Reviewer #1 was still questioning on the proposed scenario of karyotype evolution. Although the authors mentioned their efforts to search for centromeric and telomeric repeats, this only corroborated the completeness and/or correctness of the assembly. In case of the birth of new chromosomes, the current discussion might benefit from the fate of ‘old’ centromeres in the n=30 genome. For example, were 30 out of the 31 centromeres in n=31 genome inherited from the n=30 genome?

ANSWER: We now provide a new Supplementary Figure 9 displaying the distribution of putative centromeric and telomeric repeats as well as a satellite sequence that was found particularly abundant in *S. alterniflorus*, focusing on the restructured regions between *S. alterniflorus* and *S. maritimus*, presented lines 166-169 and discussed lines 262-268. From these comparisons, we can see the conservation of putative centromeric sequences between these species, except for the “additional” chromosome 31 in *S. alterniflorus* which appears depauperate

of putative centromeric repeats. However, we noticed the abundance of the 370-bp satellite repeat on this chromosome. We would like to stress out that it seems difficult from the available data to determine whether “30 out of the 31 centromeres in n=31 genome were inherited from the n=30 genome”, according to (1) the “putative” nature of the examined centromeric repeats and (2) the time elapsed since maritimus-alterniflorus divergence where rapid evolution of satellite repeats (including ancestral centromere inactivation or birth) could have taken place, which is also illustrated by the observed striking difference of the 370-bp repeat amounts between these species.

We thank all reviewers for their time and valuable comments during this process.

ANSWER TO REVIEWER COMMENTS

Reviewer #4 (Remarks to the Author):

I thank the authors for another round of corrections that improved the manuscript. I am largely happy with the introduced changes — the GO analysis methodology became clearer, and Figure 1 is more readable than before.

ANSWER: We thank the reviewer for valuable comments that helped to improve the initial version of this manuscript.

With the additional information provided in Figure 4, it seems that the only significantly enriched GO terms are those from Biological Processes and Molecular Function private to the restructured regions of *S. maritimus* — is that correct? Is it then necessary to mention the non-significant GO terms, especially given that they are not discussed in the manuscript? Maybe Figure 5 (reviewer means figure 4?) can be then reduced to the Venn diagrams with callouts to significant *S. maritimus* BP and MF terms? It was not obvious in previous versions of the figure because the authors do not provide the adjusted p-values for any GO terms mentioned.

ANSWER: The information for Figure 4 was provided according to comments from previous reviews. Please allow us to provide more clarification here.

As now mentioned in the Figure 4 caption, adjusted p-values are available in Supplementary Data 1.

- i) The restructured region only contains a small part of protein-coding genes (~1.5% of the total). It is reasonable that non-significant GO terms were identified in the restructured region;**
- ii) Our aim was to show functional differences in the restructured regions between the two species. As expected, our results identified significant GO terms that are specific to one of the two species. It clearly indicates functional divergence of the restructured region between the two species.**

Unfortunately, I have to say that I am still unhappy with the code availability. Imagine a person wants to redraw the Figure 1 — what would they need to do? At the current state of the manuscript, doing this would involve quite a number of wheel reinventions, since some scripts that are mentioned or implied in the Methods are not shared. What I dream of is a code repository that will bring that person from actual raw data generated for this article (or intermediate files if the authors would like to share them in that repo) to the actual results and figures presented in the manuscript. Sharing the GO code and `WGD\_Tracker` is a good step on this way, even though it is obvious that the latter pipeline is raw and not ready to be easily run by outsiders for a number of reasons (SIF files are missing, no easy way to install dependencies etc.) — I accept the authors' promise to improve and publish WGD_Tracker properly.

I recognize that I might be annoying with the point of code availability, as this is going there and back a few times already. Therefore, if the authors think that this request is unreasonable and making a repository with (1) scripts tracking ALL data manipulations used for presented results and (2, preferably) an easy way to recreate the run environment (a Singularity container?) is overly hard, they should feel free to defend this stance in the response letter. I will not be fighting over it to death, but I genuinely do not understand why such a repository cannot be made in the course of manuscript preparation given that the scripts are presumably somewhere around.

Having said that, I should reiterate that I like the presented story as a whole and do not want to hold it more than it is needed to make the results transparent and (easily) reproducible. I thank the authors for their patience and understanding.

ANSWER: Please see the answer below.

P.S. one tiny correction in the newly introduced text — on line 454, "are" should be replaced by "is".

ANSWER: We corrected this accordingly.

Reviewer #4 (Remarks on code availability):

The results of the paper cannot be reproduced with the given code repository. The examples section contains an unrelated dataset; the code cannot be run as the Singularity container files are missing and the paths are kept as in the original (and inaccessible) run environment; no easy way to install the dependencies (e.g. container/environment file) is provided. The repository does not cover all data manipulations done for the manuscript.

ANSWER: The “READ ME” file is now updated, explaining the whole installation process, including codes installing all dependencies (using conda and singularity). As requested by Nature Com Code availability policy, we previously provided a “DEMO” dataset using another plant genome (*Arabidopsis thaliana*), demonstrating that this pipeline was usable for any genome of interest. We are now providing a new DEMO dataset from our *Sporobolus* genomes analyses, corresponding to the intra-genomic analysis using homeogroups 3 and 4 of *S. maritimus*. A detailed Manual is also provided, describing all the optional settings developed for each pipeline.

More globally, we would like to stress out that all the methodologies used to produce our results are described in the Material and Methods sections including all the bioinformatic approaches used for genome assembly, gene and transposon annotations, parameters for homology searches by BLAST alignments, their filtering as well as the parameter settings in the provided WGD_Tracker pipeline (that will be available upon publication) and GO-term enrichment analyses. The availability of genome data (assemblies and annotations) should allow the user to perform all sets of analyses presented in this paper. For instance, the information regarding the program (Circos) and parameters used to produce Figure 1 from the available data are provided (we added information in the Figure legend about TE, gene densities calculations by 100kb-windows).

By doing this, we take into account the importance of sharing methods and code availability with the scientific community. We would like to mention, however, that this paper is not dedicated to the publication of a bioinformatic pipeline, although we make all efforts to implement the process followed in this study.

Reviewer #5 (Remarks to the Author):

The authors fully addressed my previous questions. I have no further comments.

ANSWER: We would like to thank Reviewer #5 for helpful comments which improved our manuscript.

Reviewer #5 (Remarks on code availability):

The page is not accessible yet.

ANSWER: This would be available upon publication.

ANSWER TO REVIEWER COMMENTS

REVIEWERS' COMMENTS

Reviewer #4 (Remarks to the Author):

I thank the authors for their patience with me through several rounds of review and for considering all of my comments to the manuscript. I do not see any important flaws in the manuscript, and it can be published as is. As always, there is room for improvement, and my (minute) suggestions are given below.

—

ANSWER: We thank the reviewer for valuable comments all along this reviewing process that helped to improve this manuscript.

With the new round of clarifications, I have finally understood the rationale behind the GO analysis being presented as is, and I agree with keeping Figure 4 in its current form. As I understand it now, the authors also consider non-enriched GO categories (the ones without asterisk next to them in Figure 4) and compare the full GO term spectra in the restructured regions of the two species to justify their functional divergence. Indeed, with this kind of analysis, displaying non-significant terms is appropriate, and background data is not so important. However, it can be hardly called an "enrichment analysis", as (to my knowledge) this term implies looking at _enriched_ categories and considering other differences insignificant.

If the authors would like to make this analysis more understandable for broader scientific community including myself, I suggest to clarify the analysis and specify the results a bit more. My suggestion for replacement of the sentence on lines 171–175 is: "Compared to the whole genome where the majority of the GO terms are shared, the restructured regions demonstrate functional divergence between the two species (Fig. 4). In addition, the restructured region of *S. maritimus** appeared to be enriched in several biological functions of adaptive interest such as reproduction (pollen tube development), defense (chalcone biosynthetic process) or stress tolerance (lipid biosynthetic process)."

ANSWER: We agree and edited accordingly the suggested paragraph (now lines 166-171) and detailed statistical analyses in the Method section (lines 458-462).

Regarding the code availability, I agree that the newly updated repository allows for reproducing most of the important results discussed in the manuscript. I have successfully run all of the analysis steps using the provided code and config files. In a better world, scientists would publish all of the code they used for producing figures etc., much like they do publish all laboratory protocols even when they are not describing new methods. However, after a look at other papers published in Nature Communications I agree that this manuscript deserves being accepted as is.

Congratulations on the well-done research and happy holidays!

ANSWER: Thank you for your time!

Reviewer #4 (Remarks on code availability):

The instructions for installing the environment are complete and easy to follow, they work for Linux but not on MacOS. The code runs perfectly upon installation and allows to reproduce the source datasets used for making a portion of the figures in the manuscript.

ANSWER: Yes the pipeline was developed for Linux distributions.