Dear Editor,

We thank you for the opportunity to resubmit our manuscript entitled "Copy number variation shapes structural genomic diversity underpinning with ecological adaptation in the wild tomato *Solanum chilense* (MBE-23-1089)”. We thank both reviewers, who provided many valuable comments and contributed to improving the manuscript. We especially thank Dr Aeschbacher for his precise and useful comments. We have now revised the study along the lines suggested by the reviewers. In short, we have improved the descriptions of the methods and the results to clarify the strengths and limitations of our approach. We also revised the discussion to explicitly highlight the limitations of our study and the importance of the results. Both reviewers acknowledge that the dataset is interesting, and the questions addressed are relevant to the readership of MBE. We also carefully edited the manuscript for the English.

We hope you find the manuscript much improved.

Sincerely,

On behalf of the authors,

Aurélien Tellier

Reviewer: 1  
  
Comments to the Author  
This paper examines patterns of genome-wide Copy Number Variations (CNVs) in a collection of 35 wild tomato genotypes and identifies links between CNVs and potential environmental adaptations to more stressful habitats. The authors uncover some particularly intriguing copy number changes in genes involved in photosynthesis and flowering time, which they had not previously identified with SNPs alone.  
  
While I did think there were some potentially interesting findings in this paper that could offer insight into the role of structural variants in adaptive evolution, I did have some questions/concerns about the methods being used to call the CNVs and to test for evidence of selection, which I detail below.

REPLY: We are very grateful for your careful comments on our manuscript, which are essential to improving it. We have made changes to your comments on a case-by-case basis, and our response to your comments is below.  
  
1. It was not clear from the beginning of the paper that the results were all based on short-read sequencing data; in fact, I didn’t realize this until I got to the first paragraph of the Materials and Methods section at line 458. Since using short-read data (instead of whole genomes or long reads) can very much make a difference in terms of what CNVs can be identified, I think this needs to be set out clearly from the beginning.

REPLY: We clarified it in this revision's abstract and last paragraph of the Introduction.

2. While the authors do state that the data were processed in the same way for a previously published study (line 461), I would still like to see some details here related to average read depth statistics, since again this can affect structural variant calling.

REPLY: Yes, the read depth is an important factor for the structural variant calling. The sequencing and read mapping statistics can be found in **Dataset S1**, including average read depth.

3. In the section called ‘Identification and genotyping of CNVs’ the authors describe using several SV calling tools (LUMPY, Wham, DELLY, and Manta), but I do not see any information describing what parameters were used for each tool. Also, it seems the final CNV set was generated by merging the results of all of the tools, but does this mean that all of the SVs called by each tool were merged into one inclusive data set? Or were only the SVs called by all tools (or more than one tool) kept? Was there any filtering done on the called SVs? The way that this paragraph reads suggests that the raw SV calls from each tool were all saved and put into SVtyper, which, if true, would cause me some concern about false positives.

REPLY: We included the full methodological details in the previous manuscript in the supplementary text. We have placed the methodological details in the current revision of the manuscript. However, some default parameters are not all mentioned due to manuscript length. We've prompted the reader to check out our original command lines (script) used in this study in each section of the methods on our Gitlab: <https://gitlab.lrz.de/population_genetics/s_chilense_cnv>.

We also filtered the CNVs by parameters, such as size bigger than 50 bp and CNV must be called with at least two methods. Some details, including CNV calling, merging and genotyping, can be obtained in our Gitlab. Finally, only CNVs that met all filtering criteria were inputted into the SVtyper for genotyping.

4. In the same section, 2nd paragraph, the authors describe a simulation pipeline to help validate their SV calls, but I did not see the results of this test anywhere (maybe I missed them).

REPLY: We described the results of simulations in lines 115 to 121. The results of the simulations can also be found in **Table S3**.

5. In the section called ‘Quantification of gene copy number,’ a completely different set of methods is described for identifying CNVs. I don’t understand why one pipeline was used to find CNVs in genes, while the previous pipeline was used to find all CNVs. Wouldn’t the gene CNV results be contained within the LUMPY/DELLY/Manta results from before? And if the CNV calls do not overlap, isn’t that an issue?

REPLY: These two strategies serve different purposes. LUMPY/DELLY/Manta/Wham were used to call the whole-genome CNVs for genotyping and analysis of population structure. However, these tools do not output the copy number (CN) of genomic regions. Therefore, we used Control-FREEC to quantify the gene CN. Control-FREEC offers advantages in calculating gene CN, as it automatically computes and normalizes segment copy numbers. We can only calculate the VST to identify the CN-differentiated genes when we get the gene CN data. We also examined the overlap between CNV datasets from LUMPY/DELLY/Manta/Wham and Control-FREEC. Importantly, all duplicated and deleted genes detected by Control-FREEC are also found in the CNV dataset detected by LUMPY/DELLY/Manta/Wham (lines 143-145).

1. I am not very familiar with the VST statistic used to find CNVs with signatures of divergence, but if it is very similar to FST, then it could be prone to some of the issues that FST has in low diversity regions.  Namely, that FST can have a tendency to become inflated when diversity is low. This could be of particular concern in this system, as the authors do mention that some populations underwent expansion/colonization events which would result in lower diversity in those populations. The usual way to deal with this is to use Dxy along with FST, so I would recommend the authors use that statistic as well.

REPLY: Thank you for this reminder. The VST is functionally similar to the FST for the estimation of population differentiation. FST statistics is based on allele frequency differences between populations, and VST is based on the copy number (CN) differences between populations. Indeed, within-population levels of variation influence FST, and Dxy can be freed from dependence on within-population variation. Unfortunately, to date, it is not possible to apply the Dxy formula to gene copy number (CN) datasets. In addition, the VST statistics are a relatively recent invention to complement FST, designed to address the new insights that CNV may play an important role in species evolution. As far as we are aware, there are no formal descriptions published of the shortcomings of the VST and no alternative methods have been proposed so far. Consequently, current studies using gene CN to assess the level of population differentiation (including ours) are limited to using the VST statistics only. That said, we have added a few lines in the last paragraph of the discussion to clarify that the VST and FST share shortcomings.

In addition, unlike point mutations, which also have high levels of diversity among individuals within populations, gene CN diversity among individuals within the same population is weaker than point mutations (see the dataset of gene CN in our gitlab: <https://gitlab.lrz.de/population_genetics/s_chilense_cnv/-/blob/main/CN_cutoff95.xlsx?ref_type=heads>). So, we suggest the influence of within-population levels of variation on the VST statistic based on gene CN to be rather weak.

We did not find any other statistics to measure population differentiation based on gene CN data. As a supplementary statistic, we also calculated the Dxy statistic using SNP data in this manuscript, and Dxy values show high correlations with VST (see Fig. 2C; Fig. S4; Table S6).

7. I was a bit confused by exactly how the CAFÉ analysis of gene expansions and contractions was done. In my experience with CAFÉ, it is used to identify gene families that have undergone expansion/contraction, but the way the authors described the results it sounded like they were talking about rates of expansion/contraction of particular gene copies. I think this needs to be clarified in the sections of the manuscript where the expansion/contraction rates are described.  Also, when I have used CAFÉ, I thought it required an orthology analysis to group genes into families, so those methods should be described here as well.

REPLY: Yes, we totally agree with your understanding of the CAFE analysis. We apologize for the lack of a careful description of the CAFE analysis method used in our manuscript. Indeed, in gene family expansion and contraction analysis, an orthology analysis needs to be performed to group genes into different gene families. Then, the number of genes in the different gene families is provided to CAFE to assess changes in the number of genes in these gene families across evolutionary branches. In this study, we provided directly the number of gene copy number (CN) for each gene to CAFE to assess changes in the number of gene CN across evolutionary branches. Thus, we treat each gene as a gene family and each copy of that gene as a member of that family (gene). Therefore, a homology analysis is unnecessary because all gene copies are homologous. We have also discussed the feasibility with the developer of CAFE during our analysis, and he confirmed that this procedure is identical in the calculation method and principle to the classic gene family analysis.

We provide a script for CAFE analysis used in this study on our Gitlab: <https://gitlab.lrz.de/population_genetics/s_chilense_cnv/-/blob/main/run_cafe.sh>.

8. For the association analyses between CNVs and climate variables, my understanding was that only the CNVs were used. However, I think a more robust analysis would be to look at associations between climate variables and ALL CNVs, and then see how many significant associations overlapped with the highly differentiated CNVs. This would help validate the VST outlier analysis results and could find additional associations between CNVs and environmental variables.

REPLY: We are very grateful for your suggestions. Here, we performed the association analyses between gene copy number (CN) and climate variables, not CNVs. Yes, we used the CN profiles of highly CN-differentiated genes for association analysis with climate variables, not all genes. The first reason is that when we use all the genes for association analysis with climatic variables, about 43% of the highly CN-differentiated genes were significantly associated with climatic variables. But we also got a large number of false-positive candidate genes; the CN of these false-positive genes varied only slightly across populations. This may be due to the limitations of our sample size (we only have seven different populations). Second, when we performed RDA analyses using all gene CN, we did not obtain any significantly associated climate variables (Fig. S13A). This suggests that some genes with small CN changes across populations can greatly dilute the resolution of association analysis between gene CN and climatic variables. This is why we obtained a large number of false positive candidate genes when we looked at associations between climate variables and all gene CN. However, when using the CN with highly differentiated genes, we obtained meaningful RDA results (Fig. 5A). And six overrepresented climatic variables in RDA model can explain 52.11% of the variance in gene CN. This ratio is also close to the ratio of overlapping genes (43%) between the CN-differentiated genes and the candidate genes when we perform LFMM analysis using all genes. In summary, in order to obtain more accurate candidate gene associated with climatic variables, we selected genes with highly CN-differentiation identified by VST to perform LFMM analysis. We added this discussion point on lines 480-483.

Review of Manuscript MBE-23-1089 by Wei *et al.* “Copy number variations shape genomic structural diversity underpinning ecological adaptation in the wild tomato *Solanum chilense*”

Simon Aeschbacher

7 February 2024

**Summary**

Wei et al. explore the extent, nature, evolution, and adaptive role of copy number variation (CNV) in the wild tomato species *S. chilense* using whole-genome re-sequencing data derived from seven populations from three geographic areas representing three contrasting environments. I consider the purpose of this study highly relevant and of high interest. The authors detected 212,207 in an ensemble approach based on four different software tools. The authors validated this approach using simulated data and showing that their approach reduced the false positive rate relative to any single calling tool. The authors analysed population differentiation from CNV using PCA and ADMIXTURE to find that the southern coastal populations are strongly differentiated from the other two groups (central, southern highland), and strongly differentiated amongst each other. The authors interpret the concordance of these two results as indicating that CNV at large is driven by demographic history and recent colonisation of the two southern areas. To explore how CNV varies among populations and might be related to gene function, the authors first identified genes with particularly high differentiation in copy number among the populations using a relative measure of between-population variation analogous to FST, and then performed a gene ontology enrichment analysis for the highly CN-differentiated genes. The authors identified 3,539 (2,192) genes (very) highly differentiated in CN among the populations. The southern highland and southern coastal populations differed in the relative extents of duplication vs. deletion at these genes. The authors fond the CN-differentiated genes to be functionally associated with abiotic stress (drought, cold, heat, and light) and pathways involved in the regulation of flowering time (sensitivity to photoperiod, vernalisation). The authors take these results as evidence that selective pressures linked to divergent habitat is manifested in CNV, and that adaptive changes in CN might have facilitated the colonisation of extreme habitats in the south of the contemporary species range. The authors also explored the dynamics of CN evolution and found overall trends of CN contraction in central and southern coastal populations, and CN expansion in southern highland populations. To further explore potential evidence for adaptive divergence at CNV, the authors performed genome-environment association (GEA) analyses using redundancy analysis to identify a set of climate variables associated with CN variation, followed by latent factor mixed modelling to identify gene sets associated with individual climate variables. The GEA revealed six climate variables representing variation in temperature, solar radiation, and potential evapotranspiration to be highly associated with CN differentiation. The authors identified 312 CN-differentiated genes highly associated with these six climatic variables. A subset of 34 CN-differentiated genes was found to be associated with both annual temperature range as well as annual mean solar radiation, and patterns of CN divergence among the populations at these genes suggested an overrepresentation of duplications in populations at high elevations (one in the center, two in the southern part of the species range). The authors conclude that the patterns of CNV they found among the populations studied, as well as the inferred dynamics of CN (expansion, contraction), are driven both by the demographic history (spatial expansion and bottlenecks) as well as spatially divergent selection promoting local adaptation to the different habitats. The authors therefore suggest that genomic studies of adaptive divergence in natural populations should take into account structural genomic variation as a potential source of causal or linked variation informative about ecological adaptation. I found this manuscript to be of high scientific relevance. The methods and approaches seem to have been carefully chosen and well implemented. I detected no major flaws. However, the authors seem to be overly confident in an adaptive evolutionary explanation of the observed patterns of CNV. I suggest the authors change their wording to be more descriptive and more balanced at some places in the text (see Minor Comments). Unfortunately, the manuscript suffers from many issues with language and writing that should have been addressed before a first submission. These issues together make for a major issue. For this reason, and in spite of the scientific quality, I need to suggest major revisions.

REPLY: Thank you for thoroughly parsing our study and recognizing its scientific relevance. We are also grateful for such professional and detailed advice, which is essential to improving our manuscript. We apologize for the writing issues that exist in the manuscript. In this revision, we have addressed all the issues you raised and answered all the questions to the best of our knowledge. We also checked the annotated PDF which you provided and made one by one changes according to your markup. We thank you once again for your efforts in enhancing this manuscript. We respond to each of your comments below.

Major Comments

In the abstract and the Discussion, the authors need to make an effort to better differentiate between the generic context of the current research on CNV as opposed to the specific context of the study. In the abstract, the motivation of the current specific study from the generic context remains unclear. In the Discussion, it sometimes remains unclear when the authors refer to previous work on (wild) tomatoes and when they refer to work on CNV more generally. The authors should distinguish between the scope of the study and its study system vs. the broader context. This includes specifying what taxa were the subject of studies that are cited. See Minor Comments for specific comments.

REPLY: We apologize for the unclear description of this information. We have made changes based on your comments and annotations in the PDF file.

For the identification and calling of CNV as well as the quantification of CN differentiation among populations at individual genes, the authors used multiple approaches. In the first case, they devised an ensemble calling methods; in the second case, the authors showed results based on both of the implemented types of VST and gene sets obtained with two different significance thresholds. The multitude of approaches and sets of results presented and discussed is a bit overwhelming and limits the clarity of the text. The authors should make an effort to more strictly differentiate in terms of complexity and level of detail between the main text and the Supplementary Text. The main text should be streamlined to feature only the absolutely necessary level of complexity; the Supplementary Text can give the details. As of now, the main message is confounded by methodological details and decisions. Addressing this point will also resolve the current issue of a high degree of redundancy between the main-text Methods and the Supplementary Text.

REPLY: We apologize for not clearly describing each analysis step and causing confusion. Yes, in the first case, we called genome-wide CNVs using an ensemble calling method. This result demonstrated the general distribution and characterization of CNVs on the genomes of different populations. In the second case, we first quantified the values of copy number for each gene based on the two different approaches, we then calculated VST statistics to identify genes with high CN differentiation. We have clarified the different analyses in this revision and fully described the objectives of the different analyses.

I am concerned that the VST outlier analysis is inflated due to multiple testing. I know it is hard to come up with the “correct” way of addressing this issue because CNVs might be partially linked. However, I think the authors should at least acknowledge the fact that they did not correct for multiple testing (e.g. as a limitation to be stated in the Discussion).

REPLY: Thank you for reminding us of this limitation. We have added this limitation at the end of the Discussion (lines 471 to 473).

I wonder if the authors have a good explanation for why only the highly CN-differentiated genes (high VST) show strong population structure in the PCA, whereas the PCA on genome-wide CNV seems to show strong population structure without any further a priori restriction on high VST. To me, this contrast might suggest something that seems to be partially misaligned with what the authors conclude: that most genes are under a strong constraint to maintain CN stable across geographic areas, and only a small proportion of genes are free to differentiate in CN number (those with high VST). On the other hand, for the whole genome, selective constraints on maintaining stable CN is much relaxed on average, and so CN is free to evolve neutrally, i.e. differentiate by mutation and genetic drift among the geographic areas. The authors, on the other hand, seem to be determined to focus on the highly CN differentiated genes. I can understand this focus, but I wonder if the authors could address why only high-VST genes also reveal the expected population structure.

REPLY: We are sorry for this confusion. There is no doubt that the PCA based on highly (and very highly) CN-differentiated genes (high VST; Fig. 3A and Fig. S7C) should exhibit a strong population structure because these genes have been screened with significantly differentiated CN profiles across populations. However, for the PCA based on all genes CN (Fig. S7A), the PCA did not exhibit such a strong population structure. The population structure revealed by the PCA is indeed based on past demographic events. This is because the gene set contains a large number of normal genes, i.e., genes with no CNVs. So, this is different from the PCA based on genome-wide CNVs. These normal genes have the same CN (CN=2), and likely decrease the resolution of the PCA.

The interpretation of the evolution of CN along the population tree seems to be biased to an adaptation perspective. Looking at Fig. 4b, I think these patterns could also be explained by neutral evolution (genetic drift associated with expansion and bottlenecks). The authors should provide a more balanced explanation. I am fine with the authors interpretation regarding the highly rapidly expanding/contracting genes, but my concern relates to the interpretation prior to the restriction to rapidly expanding/contracting genes.

REPLY: Sorry, we, of course, agree that the effect of neutral evolutionary forces such as drift and past demographic events drive population CNV patterns. We clarified this point and added this explicitly in the discussion on lines 412-415.

The authors use very confident wording when describing, explaining, and speculating about their results on CNV being involved in local adaptation. Given the associative nature of the analyses and some arbitrary choices that need to be made as part of such analyses, the authors should switch to a more tentative wording. I made some suggestions (see Minor Comments and annotated PDFs).

REPLY: Thank you for this word of caution. This reminder is very necessary and improves the rigour of the manuscript. We have changed our wording according to Minor Comments and annotated PDFs.

The terms “population”, “accession”, and “individual” seem to have been confounded at several places. I can see that population and accession may be used exchangeably, but the confusion between individuals and populations needs to be fixed.

REPLY: Corrected.

There seems to be a generic confusion between the terms “variation” (which, in my view, has no plural in this context, but is sometimes used by the authors in the plural form) and “variant(s)”. I suggest that the authors differentiate carefully between “variant(s)” and “variation” whenever these words occur. “Variation” is the overarching term, and “(a) variant(s)”are/is the individual constituent(s) of this variation. The issue is tricky in so far as I understand the need for abbreviating “copy number variation” as well as “copy number variant(s)”. The authors might want to introduce and use “CNV” for the former and “CNVs“ for the latter (if the latter is in plural, i.e. refers to multiple variants). This leaves the singular “copy number variant”unabbreviated, but I think this can be tolerated for the following reasons: i) “[copy number] variant” occurs [40] 3 times, whereas “[copy number] variants” occurs [63] 9 times in the main text; ii) the 40 occurrences of “CNV” also include instances in which “copy number variation” is meant, and in these instances “CNV” could still be used; iii) the authors seems to use “copy number variations” / “CNVs” in several cases where I think they should actually be using “copy number variation” / “CNV”, and, again, in these cases “CNV” will remain. So, I think the lack of an abbreviation for the singular form “copy number variant” can be tolerated, and the authors should make an effort to fix this point.

REPLY: Sorry, we didn't notice this problem before. Thank you for giving such a detailed explanation and valid advice. We checked and fixed this problem in the new manuscript.

The authors repeatedly visualise the output of PCAs with a 3D plot (Fig. 2A; Figure S3A; Figure S6B, C, D; Figure S10B, C). I find it hard to read and interpret 3D point clouds and I think the authors should dissect the 3D plots into two to three 2D projections, depending on how many are needed to illustrate the main patterns.

REPLY: Corrected. We have dissected all 3D PCA plots into three 2D projections.

I am very concerned about the many minor issues with language and writing, which hamper the clarity, precision, and brevity at many positions, and which to me amount to a major issue in total. It would have taken me too much effort to transcribe and list all the minor issues I annotated while reading the manuscript and the two supplementary files. In the Minor Comments section below, I therefore only picked and stated some comments and questions. For the great majority of my suggestions w.r.t. to writing and language, please see the annotated PDF files I attached to this review. Beyond that, two recurrent issues are the following:

The majority of the manuscript is written in present tense, which to me sounded unnatural if not incorrect given that the authors often write about what they did, not about what happens at the time of reading. I suggest the authors consistently use past tense when describing what they did and what they found.

• There is an arbitrary mix of active and passive voice. I suggest the authors homogenise the language with respect to this point to either use active or passive voice more consistently when they describe what they did.

REPLY: We are sorry for these minor issues with language and writing, and thank you for your patience in marking up all the changes needed for us in the annotated PDF. We have checked and revised all the writing according to your notes you provided in the annotated PDF. In particular, we have also carefully revised the tense problems and voice problems that exist in the manuscript.

Minor Comments

**C**: comment; **Q**: question; **S**: suggestion; **R**: request.

**Title**

• **C**: I find “underpinning” misaligned with the purpose of the study and the evidence provided, and would suggest “associated with” instead. To me, it should say “copy number variation” (not “. . . variations”), and I have the impression that “structural genomic diversity” is more commonly used for what the authors seem to mean by “genomic structural diversity”. S: Overall, I suggest to rephrase the title to “Copy number variation shapes structural genomic diversity associated with ecological adaptation in the wild tomato Solanum chilense”.

REPLY: Corrected.

**Abstract**

• See annotated PDF.

REPLY: We have corrected the abstract according to your suggestions in the annotated PDF.

**Introduction**

• [l.33] S: I would omit “chromosomal rearrangements” or specify them as “neutral chromosomal rearrangements” because there would be no consensus on whether rearrangements are perseneutral.

REPLY: Corrected.

**Results**

• [l.128] Q: Did you use ADMIXTURE (as stated in the Methods) or STRUCTURE (as stated here)?

REPLY: We used ADMIXTURE to perform population structure analyses. We have changed “STRUCTURE” to “ADMIXTURE”.

• [l.173] Q: Is there a good reason for why the authors performed the functional enrichment analysis on the set of 3,539 highly CN-differentiated genes, and not the 2,192 very highly CN-differentiated genes?

REPLY: The 2,192 very highly CN-differentiated genes are also contained in the set of 3,539 highly CN-differentiated genes. So, all enriched GO terms based on the 2,192 very highly CN-differentiated genes are also observed in the GO terms enriched by 3,539 highly CN-differentiated genes. In addition, from the PCA results, 3,539 highly CN-differentiated genes showed the same sample clusters with PCA using the 2,192 very highly CN-differentiated genes (Fig. 3A; Fig. S7C). Therefore, the 3,539 high CN-differentiated genes reflected the consistent patterns of CN-differentiation as observed with the 2,192 very highly CN-differentiated genes.

• [l.237–239] Q: I wonder about the interpretation of the result here, i.e. about the speculation that a high CN expansion rate along an internal branch in the population tree could be driven by a high rate of expansion along one of the subtending terminal branches. Does CAFE v4.2.1 not separate the rates on internal branches from the rates on terminal branches? If it does, how can the rate along the internal branch be inflated because of a high rate along just one of the subtending terminal branches (leaves)?

REPLY: Yes, your understanding is correct. CAFE separates the rates on internal branches from the rates on terminal branches. We should compare the results at the same level of branches. We have changed the description for this result.

• [l.253] S: I wonder if it would be better to break the paragraph here, but to omit the break in l.247.

REPLY: Yes, we have adopted this recommendation.

• [l.272] S: I suggest a paragraph break here.

REPLY: Corrected.

• [l.329–330] C/S: Fig. 5B suggests that Bio7 and ann\_Rmean are correlated, and so it may not be too surprising that the sets of genes associated with these two variables overlap. I think you should mention this point.

REPLY: Corrected. We have mentioned this point in the manuscript (lines 346).

• [l.340-342] C/S: It was not clear to me what test the result stated here was based on. Could you please clarify in the text? An analogous comment applies to the statement in l. 343–344.

REPLY: Here, we mentioned the results based on the analysis of GO enrichment. We have clarified it in the manuscript (lines 357 and 361).

• [l.344–349] C/R: This part is on the side of a Discussion paragraph, with quite some interpretation and speculation mixed into what should be a summary of the results. I think the authors should use a more neutral wording for points (ii) and (iii), i.e. language that is more descriptive of the results, and less on the side of interpreting these results as showing certain evidence of locally adaptive divergence.

REPLY: We have rephrased this part (lines 365 to 368).

**Discussion**

• Generic comment to the Discussion: Please more clearly differentiate between the scope of the study and its study system vs. the broader context. Specify what taxa were the subject of studies you cite.

REPLY: Corrected. We clearly state the study system and taxa we cited.

• [l.351–353] C: The scope of this statement remains unclear. Does the sentence refer to plants in general, to wild tomatoes, to *S. chilense*? I also think the authors should start the Discussion with a concise reminder of the purpose of the study.

REPLY: We have referred to *S. chilense* in this sentence. We also added the purpose of the study at the start of the Discussion (line 370).

• [l.369–370] C/S: I found this phrasing overloaded. How about: “We identified patterns of gene CN variation that likely represent footprints of adaptive divergence.”

REPLY: Corrected. We have adopted the more concise writing you provided (line 390 to 391).

• [l.381–383] R: Please specify the taxa for which the result reviewed here was found. The references given do not seem to be specific to *S. chilense*, but to include other taxa.

REPLY: Yes, the references given include also other taxa. We have rephrased this sentence (lines 400 to 403).

• [l.399–402] C: I wonder if it is necessary to mention the anthocyanin pathway in cultivated tomato in the way it is. Either omit this statement or make clear how it relates to your finding.

REPLY: We have rephrased this part. In plants, anthocyanin accumulation can improve the biotic and abiotic stresses tolerance. Anthocyanin accumulation helps in eliminating Reactive Oxygen Species (ROS) molecules and protects the DNA damage under UV radiation. The SH populations (high altitude) of *S. chilense* showed high correlation with Solar radiation (ann\_Rmean) and previous ecological niche study suggested that *S. chilense* populations are expanding to the habitats of high altitude (Wei, et al. 2023b). Therefore, we inferred that the gene CNVs that occurred in the anthocyanin accumulation pathway may be important for adaptation in high-altitude populations of *S.chilense.* (lines 424 to 430)

• [l.408–410] C/R: This sentence illustrates the problem with the generic use of present tense: The statement made here sounds like a general statement because you use present tense. But the statement is meant to be specific to the study. Please rephrase to use past tense and to make clear that you are referring to your own result in this study.

REPLY: We have rephrased this sentence using past tense.

• [l.424] C/R: This is the first time it is stated that these genes are chloroplast genes found in the nuclear genome, so the reader may need more than just a clause to appreciate the fact. Please expand to increase clarity.

REPLY: Yes, we have expanded this point (lines 452 to 455).

• [l.436–438] Q/S: Do you mean that the tests with simulations showed that your approach was likely conservative? If so, please rephrase to make this point more clear. Also, it would make sense to repeat that the simulations simulated short-read data.

REPLY: We have rephrased this part and emphasized that the simulations are based on short reads (lines 466 to 468).

• [l.440–441] C/S: How does this sentence relate to the study? I do not classify GEA as a selective sweep method. I also do not think the authors used ‘several’ methods, but just RDA coupled with LFMM. Please rephrase to a more precise statement.

REPLY: We removed this sentence that was not relevant to this study.

• [l.444–447] S: Please expand a bit on how, not only that, the new method will help.

REPLY: We have added a short sentence to explain explicitly how the new method may help (lines 476 to 479).

**Materials and Methods**

• Generic comment: There is currently a considerable degree of redundancy between the Methods in the main text and those in the Supplementary Text. When revising the manuscript, I suggest that you make an effort to reduce this redundancy.

REPLY: We have removed the Supplementary Text in this revision. So, we described all the details of the method in this manuscript.

• [l.483] Q: I did not understand why there is a factor of two in the formula for the copy number. Could you please state that in the text?

REPLY: A factor of 2 originates from the diploidy of *S. chilense*. We have added this description in the Methods (line 556). Please also see the citation:

*Rinker DC, Specian NK, Zhao S, Gibbons JG. 2019. Polar bear evolution is marked by rapid changes in gene copy number in response to dietary shift. Proceedings of the National Academy of Sciences 116:13446-13451.*

• [l.485] R: Please explain what VST is, or refer to the Supplementary Text.

REPLY: We have added the details about VST (lines 559 to 564).

• [l.499–500] Q: Please justify why you forced the tree to be ultrametric. Is it justified to assume a constant molecular clock? Does the rate of change of CN depend on *N*e or does *N*e cancel? I ask because branches in TreeMix trees scale with *N*e, and if one forces the tree to be ultrametric, one loses the scaling by *N*e.

REPLY: We thank you for raising this issue, which had not been considered before. We emphasize ultrametric trees because CAFE analysis needs to force phylogenetic trees to be ultrametric. Otherwise, CAFE cannot run.

So far, the relationship between changes in gene CN and *N*e is unclear. Drift, demographic effects and shifts in selection regime are shown to result in changes in CN distributions in humans. For some candidate genes, neutral factors are not sufficient to explain the observed patterns, but methods to test the expected distribution of gene CN are yet incomplete:

<https://www.biorxiv.org/content/10.1101/2023.08.14.553171v1.full>.

The effect of *N*e on gene CN changes could not be fully investigated here, nor definitive conclusions drawn.

• [l.511] C/S: I felt that a transition was missing between the previous sentence and this sentence here (“LFMM . . . ”). Please insert one.

REPLY: We have inserted a sentence of transition.

• [l.513] Q/R: Did you implement LFMM2? If so, please provide a reference to the code.

REPLY: We provided a reference for LFMM2, and we also provided the R script we used in this study: https://gitlab.lrz.de/population\_genetics/s\_chilense\_cnv/-/blob/main/lfmm.R.

**Figures**

• Generic comment: I find it difficult to read and interpret bar plots in which bars belonging to different categories are stacked (e.g. Fig. 1B, E; Figure S6A; Figure S13B, C). Did the authors consider alternatives, e.g. clustering bars side by side, or plotting individual bar plots for each category on the same horizontal line so that it is easier to compare values for a given category across different classes on the x-axis?

REPLY: We appreciate the improvement in the readability as well as the aesthetics of the figures. We have redesigned the plots based on your suggestions.

• Fig. 1:

– R: The map in panel A is too small. Consider enlarging that panel and reorganising the other panels. S: Panel B seems fully redundant with Table S1, so I think the authors could drop Table S1.

REPLY: We have enlarged the map (Fig. 1A), and removed Table S1.

– [l.781–782] R: There seems to be a confusion between accessions and individuals. As far as I understood, population is equivalent to accession in this case, but what is “accessions” here should read “individuals”.

REPLY: This reminder is very important for our manuscript. We have checked and corrected the full text.

• Fig. 2:

– R: The colour scheme in the ADMIXTURE plot does not seem to correspond well with the colour scheme of the PCA. For instance, for K = 7, it is confusing to see the purple plots below the blue SC\_LA4107. Also, the shades of the respective colours do not match (e.g., the green in the ADMIXTURE plot seems too bright compared to the green of SC\_LA2932; neither the light green nor the dark yellow in the ADMIXTURE plot seem to unambiguously match the olive green of C\_LA2931. Please adjust the colour scheme of the ADMIXTURE plot so that it matches the one of the PCA in panel A.

REPLY: Corrected.

– [l.788] R: There is apparently again a confusion between accessions and individuals. According to the Introduction, there are 35 individuals from seven populations (accessions), not 35 accessions.

REPLY: Corrected.

– [l.788] Q: Did you use ADMIXTURE or STRUCTURE (it says “Structure analysis” here,

but in l.474 you state you used ADMIXTURE)?

REPLY: We used ADMIXTURE. Here, when we refer to “Structure analysis”, we mean “Population structure analysis”. So, we have changed “Structure analysis” to “Population structure analysis”.

• Fig. 4:

– C/S: The legend in panel B was unclear to me. Specifically, what is meant by “proportion of CN [copy number] lost” and “proportion of CN gained”? Do you mean “proportion of genes with CN loss” and “proportion of genes with CN gain”, respectively? Please rephrase.

REPLY: We are sorry for this unclear legend. Here, “proportion of CN lost” and “proportion of CN gained” are the ratio of the number of CN losses or gains to the total number of CN changes for all genes. The formula is calculated as “proportion of CN losses/gains = the number of CN losses/gains / (the number of CN losses + the number of CN gains)”. We added this formula to the caption.

– [l.806–813] S: In the caption, I suggest to improve the wording when you write about gene copy number gain and loss (see specific suggestions in the annotated PDF).

REPLY: Corrected.

• Fig. 5:

– [l.821–822] R: The title sentence to me sounds to interpretative. I think the authors should chose more descriptive, neutral wording.

REPLY: Corrected.

– [l.824] Q: Could you please check if this is precisely the meaning of the vector lengths? Is it not that the projection of a vector onto an ordination axis shows the correlation with that axis? Please rephrase if necessary.

REPLY: Corrected.

**Tables**

• Table 1: See minor fixes in the annotated PDF.

REPLY: Corrected.

**Supplementary Text**

• Please refer to the annotated PDF. I did not make detailed suggestions for the entire text because I was a bit overwhelmed by the density of writing and language issues. Please revise the entire text carefully before submitting a revision.

REPLY: Since the supplementary text contains only details of the method and there is no more redundancy with the main manuscript. We have therefore described all method details in the main manuscript and omitted the supplementary text.

**Supplementary Figures and Tables**

• Generic comment: The captions of the supplementary figures and tables seem incomplete and rudimentary at times. Please add full title sentences and also make full sentences in the remainder of the captions.

REPLY: Corrected.

• Figure S2:

– There seems to be again a confusion between “accession” and “individual”. In the label of the x-axis, it should say “individuals“ in my view.

REPLY: Corrected.

• Figure S3:

– R: See my comment to Fig. 2 w.r.t. the colour scheme of the ADMIXTURE plot in panel C (it does not match the colours assigned to the accessions and used in panel A).

REPLY: Corrected.

• Figure S5:

– Q/R: Do the dots in the figure represent genomic windows or genes? Please adjust the

caption if necessary (it currently mentions “genes”).

REPLY: The dots in the Fig. S5 represent genes, not genomic windows. We first quantified the copy number (CN) based on the sliding windows, and then calculate the CN of each gene based on the CN values of the sliding windows. We finally calculate the VST statistics for each gene.

• Figure S6:

– What is meant by “CN value(s)”? Please define / write out once at least.

REPLY: Corrected.

• Figure S7:

– C/R: The description in the caption of panel C is unclear. Do you mean “The number of genes associated with a response to . . . ”? Please fix.

REPLY: Corrected, now in Figure S8.

• Figure S9:

– Q/R: Am I right in thinking that this figure is reproduced from Wei et al. (2023a)? If so, I

do not think it is necessary (nor appropriate) to reproduce the figure here.

REPLY: This is a completely new figure that has not appeared in previous studies in Wei et al. (2023a). We just used previous transcriptome data to check the changes in expression of these 11 drought-responsive genes under water deficit conditions.

• Figure S11:

– C/R: The caption misses a title sentence. Please fix.

REPLY: Corrected.

– R: The second part of the caption needs revision. The wording used to describe the RDA plot is unclear. See detailed comments in the annotated PDF.

REPLY: Corrected.

• Figure S12:

– R: Please add a title sentence to the caption and describe what panel A shows.

REPLY: Corrected.

• Figure S13:

– R: Please revise the title sentence to be more informative.

REPLY: Corrected.

• Table S3:

– C: This table seems fully redundant to Figure S2 and could in my view be omitted.

REPLY: Corrected.

• Table S4:

– C: This table seems fully redundant to Fig. 1E and could in my view be omitted.

REPLY: Corrected.

• Table S5:

– C: I do not think the footnote is necessary given the column labels in the table and the

details given in the Methods and Supplementary Text.

REPLY: Corrected.