Review of first revision of Manuscript MBE-24-1101 by Wei *et al.* "Copy number variation shapes structural genomic diversity associated with ecological adaptation in the wild tomato *Solanum chilense*"

Simon Aeschbacher

24 January 2025

Summary

I thank the authors for answering to and addressing the Major Issues I raised in my first review. In some cases, the authors replied with a generic comment saying that they addressed my concern, but often without saying where exactly they did so in the manuscript. This made it hard for me to double-check.

I appreciate the authors' efforts to fix the many language issues. The manuscript text reads more homogeneously and does not suffer from obvious flaws anymore. That said, the authors have now moved to the main text text parts that they previously put into a Supplementary Text. These parts concern the description of the methods related to the inference of gene CNV and overall CNV. While I had suggested that the authors clean the main text from technical details that, in my view, make the article hard to digest for the majority of the readers of Molecular Biology and Evolution, Reviewer 1 had requested that the authors describe the methods in more detail. I thus realise that we reviewers made different suggestions and the authors decided to follow Reviewer 1. I have no issue with that. I must highlight, though, that I now invested some more effort in those text parts that have now been added to the main text. As a consequence, I now have a somewhat lengthy list of Minor Issues (see subsection "Materials and Methods" below).

At multiple positions throughout the manuscript, the authors refer to scripts written and used for specific analysis steps. They provide URLs to files in their GitLab repository, but this repository cannot be freely accessed from outside the TU Munich, as it seems (see Major Issue 2 below).

I appreciate the changes made by the authors in the Results to remove parts that were too interpretative and in the Discussion to clarify which results and studies cited concern which species, as well as to motivate why finding homologs of genes involved in the anthocyanin pathway is relevant in the context of altitudinal adaptation.

I am now left with two major issues and, unfortunately, still a long-ish list of minor issues. I think it should be straightforward for the authors to address these issues and bring the manuscript in shape for publication in MBE. I have not changed my fundamentally positive opinion about the relevance of manuscript: in times where the field is eager for evidence from natural populations about the patterns and functional implications of structural variation, the manuscript makes an important and timely contribution.

Major Comments

1. I did not understand the authors' argument in the second-to-last paragraph of their response to Major Issue 6 by reviewer 1. Reviewer 1 seemed to be concerned about V_{ST} being inflated if within-population copy number variation is low. However, the authors seem to argue that because within-population copy number variation is low, and so its influence on V_{ST} would be minor. Why that? At least when drawing from the analogy to F_{ST}, low within-population variation is prone to inflating V_{ST}. In this context, I find it unsatisfactory that the authors do not explicitly define how V_{ST} is calculated. What the authors added on l. 131–133 in the Results is too vague and cannot replace a proper definition in the Materials and Methods. The authors' scripts that implement the calculation of V_{ST} are not freely accessible. I suggest that the authors add a a few lines to the paragraph on l. 548–557 to address this concern.
2. I was not able to access the script files from the GitLab repository at the LRZ center for computation. I therefore could not (spot) check the code. I consider this a limitation that should be fixed in the process of peer review if the journal wants the review to cover the supplementary code files. Multiple URLs given to specific sections within the GitLab repository are given throughout the Methods section, but none of them leads to content that is publicly accessible.

Minor Comments

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

General

- **C:** The line numbers that the authors given in their response to the reviewers' comments to not agree with those of the changes in the manuscript that these refer to. This mismatch led to quite some overhead work in the review of the revisions.

Title

- **C:** I appreciate the modification from "underpinning" to "associated".

Abstract

- No comments.

Introduction

- No comments.

Results

- [l.152–153] **R:** Replace "In Fig. S6, we showed . . . " by "In Fig. S6, we show . . . " (present tense), or, better, write "Figure S6 shows . . . "
- [l.233–234] **S:** Rephrase this sentence to "Our findings so far indicate that a considerable number of CN-differentiated genes may be involved in adaptation to local habitats".
- [l.251] **S:** Delete "the" before "opposite".
- [l.263] **C:** I find it difficult to relate the "slight decrease in the rate of CN expansion in C_LA3111 (Table S7; Fig. S11)" to what Fig. S11 actually shows. The quantities shown in that figure are the "proportion of CN lost" and the "proportion of CN gained". **S:** I ask the authors to use consistent wording; how does a rate relate to a proportion? Fig. S11 shows that the proportion of

CN gained is higher than the of CN lost in C_LA3111. How do I reconcile this with the apparent "slight decrease in the rate of CN expansion" in that population? Not only do I need to make an intellectual exercise to resolve one too many negations (decrease in expansion), but after that exercise, I find the result to be the opposite of what Fig. S11 shows. I am confused, and I think other readers will be so, too.

- [l.287] **S:** Insert "a" before "adaptive response".
- [l.315–318] **C:** This sentence is hard to read. Could the authors please streamline and improve the phrasing?
- [l.332–334] **C:** I find it difficult to make sense of the statement saying that "physiological processes" are "essential processes to drought stress". I suggest the authors rephrase this sentence to clarify what exactly they mean.
- [l.337] **S:** Replace ", because of the correlation ..." by ", which is likely a consequence of the correlation ...".
- [l.342] **S:** Insert "a" before "response".
- [l.357–359] **C:** I appreciate the changes in language that the authors implemented to make the end of this Results paragraph more descriptive and less speculative.

Discussion

- In their response to comment no. 6 by Reviewer 1, the authors state that they now added a few lines to the last paragraph of the Discussion about the shortcomings shared between F_{ST} and V_{ST}. Reviewer 1 was concerned with V_{ST} being inflated due to low within-population variation in CNV. As I stated in my Major Issue 2 above, I did not find this concern by Reviewer 1 to be addressed in the manuscript.
- [l.391] **S:** Replace "plants" by "plant species".
- [l.391–396] **C:** I appreciate the changes the authors made to clarify what species the results and studies cited were based on. These clarifications help to differentiate between the current study and its scientific context.
- [l.394] **S:** Replace "In this study, adaptive gene loss may also occur ..." by "Our study suggests that adaptive gene loss may also occur ...".
- [l.396] **S:** Replace "This confirmed the ..." by "These findings confirm the ...".
- [l.401] **S:** Replace "agreed" by "agree".
- [l.414–420] **C:** I appreciate that the authors now explain why finding homologs of genes involved in anthocyanin accumulation is relevant in the context of altitudinal adaptation.
- [l.416] **S:** Insert "the" before "anthocyanin accumulation pathway".
- [l.423] **R:** Replace "variant" by "variation".
- [l.442–444] **C:** I thank the authors for providing some more background on organellar gene transfer to the nuclear genome.
- [l.446] **S:** Replace "displayed" by "showed".
- [l.447] **S:** Replace "This is ..." by "This trend is ...".
- [l.457] **S:** Replace "meaning that" by "which means that".

Materials and Methods

- [l.492–492] **C:** The abbreviation "SV" for "structural variation" has already been introduced in the main text. The repeated definition here likely results from the fact that the authors moved some methods text from the Supporting Materials to the main text.
- [l.493–495] **S:** Perhaps rephrase to "This evaluation enumerated ... and/or duplications." to

"This study found that combining SV detection tools tends to give higher precision and recall than individual tools and that LUMPY (Layer, et al. 2014), Manta (Chen, et al. 2016), Wham (Kronenberg, et al. 2015) and DELLY (Rausch, et al. 2012) are tools with good overall performance for deletions and duplications."

- [l.495–497] **C:** This sentence does not seem to be complete. It remains unclear what the difference is between "tool" and "algorithm". It seems as if the authors use "algorithm" for what Kosugi et al. (2019) used "method", whereas Kosugi et al. (2019) used "algorithm" for the software tool. **S:** I suggest that the authors restrict their statement here to what is strictly necessary for the reader to obtain an overview of what information from the sequencing data data was used to call CV overall. Later in the text, when the application of specific tools is described, the description can state what information the respective tool relies on. Specific suggestion for here: "These tools implement different calling algorithm that jointly draw information from patterns in read pairs, split reads, read depth, and de novo assembly".
- [l.498] **S:** For consistency with l. 493, I suggest to write "LUMPY" in capital letters. This comment also applies to other occurrences of the tool name.
- [l.498–500] **S:** Rephrase ", and the split-read alignments also were extracted ..." to ", and then we extracted the split-read alignments ..." to increase consistency (active voice) and clarity (repetition of the phrasing reassures to the reader that both steps are done to extract data to be analysed, not to be excluded).
- [l.502] **S:** Rephrase "CNV calling used DELLY ... into vcf file (Danecek, et al. 2011; Danecek, et al. 2021)" to "For CNV calling with DELLY v0.7.6, we chose the default parameters and converted the output file from bcf to vcf format using bcftools v1.9 (Danecek, et al. 2011; Danecek, et al. 2021)".
- [l.504] **S:** Omit "Furthermore" and just say "We also ran ...".
- [l.506–507] **S:** Replace "as 50bp" and "as 1Mb" by "to 50 bp" and "to 1 Mb", respectively. Not that I suggest to insert spaces between the numbers and the units to be consistent with other places in the manuscript, e.g., l. 541 to l. 543. **C:** It was unclear to me what is meant by "..., and CNV types and DNA strands must match". Please clarify and improve the formulation.
- [l.509] **R:** The current phrasing does not work. Please rephrase to "We (finally) used the merged CNV set **as input** for SVTyper v0.7.0 to ...".
- [l.509–512] **S:** I think this part could be written more compactly as "We used the merged CNV set as input for SVTyper v0.7.0 to call breakpoint genotypes of the structural variants (Chiang, et al. 2015)".
- [l.512–513] **S:** The sentence did not make sense. I suggest to rephrase to "The script we used for CNV calling, merging, and breakpoint genotypic is available from ". (Note that the URL leads to a GitLab repository two which external users have no access.)
- [l.516–517] This sentence in parts seems to be a verbatim copy of the description on the tool's website. To avoid the risk of committing plagiarism, I recommend to rephrase the description. I think the copying of the description from the website also transferred a statement from the website that I found misleading: as far as I understand, what CNV-Sim does is to first simulate NGS reads using ART, and then to introduce CNV. In sum, I think that l.515 to 519 could be simplified to and written more clearly as "To assess the sensitivity and accuracy of our pipeline for CNV calling, we simulated short-read data with CNV using CNV-Sim v0.9.2 (https://github.com/NabaviLab/CNV-Sim)". Since you will ultimately provide the script, interested readers can look up the details there.
- [l.521] **S:** Rephrase "The command lines for simulations can be found on: ..." to "The script implementing these CNV simulations is available from ...". (Again, note that external users have

no access to the linked GitLab repository. Please make the repository public.)

- [l.535–537] **R:** Please consistently do or do not surround the "equals" symbols ("=") by spaces.
- [l.558–571] **C:** I appreciate that the authors inserted this paragraph to describe the procedures they used to obtain candidate genes with high CN divergence based on V_{ST}. However, I am still missing an explicit and unambiguous definition of what V_{ST} is; how is it defined in terms of more fundamental measures of CN variation? **R:** I think the authors should provide such an explicit definition.
- [l.560] **S:** Omit "then" before "independently". Omit "value" after "V_{ST}".
- [l.562] **C:** The linked GitLab repository is not freely accessible. Please make the repository public.
- [l.573] **S:** Insert "a" before "BLAST".
- [l.575] **S:** Insert "a" before "GO enrichment analysis" and "the" before "*A. thaliana*".
- [l.576–578] **R:** This sentence reads unnecessarily complicated. Please rephrase to something like "We used the Benjamini-Hochberg method (reference) to control the false discovery rate at 0.5 when we determined the enriched GO terms".
- [l.585–586] **S:** Replace "we performed analysis of gene CN expansion and contraction" by "we analysed gene CN expansion and contraction".
- [l.591–592] **R:** This sentence does not read well. Please adjust to something like "We chose a significance threshold of 0.05 when identifying genes with an excess rate of evolution (expansion or contraction) in different groups/populations".
- [l.592] **S:** Rephrase to "The code we used to analyse CN expansion and contraction can be found on . . . ". (Note that the linked GitLab repository cannot be freely accessed.)
- [l.600] **S:** Please insert "the" before RDA analysis. Replace "from the vegan package as implemented in R" by "from the R package vegan".
- [l.604] **S:** Replace "that showed VIF $>$ 10" by "with a VIF of 10 or above". Insert "the" before "RDA".
- [l.605] **S:** Replace "The R script of RDA analyses" by "The R script for the RDA analysis". (Please note that the URL leads to a GitLab repository that is not freely accessible.)
- [l.607–608] **C/S:** I did not understand what exactly the authors mean by "the dynamics of CN across populations". Should there be a "the" before "RDA"?
- [l.608] **C/S:** The authors so far used "climatic variable(s)", but here they use "climate variable". I suggest the authors unify their use of language.
- [l.610–613] **S:** Rephrase "We first performed *lfmm_ridge*" function implemented in the R library LFMM" to "We used the *lfmm_ridge*" function in the R package LFMM". Replace "matrix using a K value of four latent factors" by "matrix under the assumption of $K = 4$ latent factors".
- [l.617] **C:** The GitLab repository is not freely accessible.

Figures

- Generic:
  - **C:** I thank the authors for unstacking the bars in the bar plots in Figs. 1 and 2. This change has much improved the readability of these plots.
  - **C:** I thank the authors for resolving the confusion between the terms "individual(s)" and "population(s)".
- Fig. 1:
  - **C:** I thank the authors for enlarging the map in panel A.
  - [l.909] **S:** Change the title sentence to "Overview of copy number variation detected in 35 *S. chilense* individuals".
  - [l.910] **C/S:** I was confused about the TGRC populations being mentioned here. What is

the difference between the black dots and the coloured dots? The Methods section is missing a clear description of the sampling design. Where did the authors take how many samples from? I realise I should have noted the black dots on the map in Fig. 1A in my first review, but I did not. I also realise the reader might be able to collect the information from various places in the manuscript, including the captions of Fig. 1 and 2. However, I would encourage the authors state this information concisely early in the Methods, where most readers would expect the information.

- Fig. 2:
  - **C:** I thank the authors for aligning the colour scheme in the ADMIXTURE plot with the colour scheme in the PCA plots in Fig. 2A and 3A and the bar plot in Fig. 1E. However, I still find that the reader needs to do quite some unnecessary extra-processing to align these colours with the ones used in the map in Fig. 1A.
  - [l.920] **S:** Replace "... assuming $K = 2 - 7$ subgroups ..." by "... assuming between $K = 2$ and $K = 7$ subgroups ...".
- Fig. 3:
  - In the caption, avoid the repeated introduction of the abbreviation "CN" for "copy number".
  - [l.927] **S:** Replace "The PCA ..." by "A PCA".
  - [l.929] **Q:** Should it read "stimuli" instead of "stimulus"?
  - [l.930] **S:** Replace "The ratio of gene enrichment is equal to ..." by "The proportion of gene enrichment is defined as ...".
- Fig. 4:
  - [l.939] **S:** Omit "is" to make clear to the reader that panel A) shows the tree used for the expansion and contraction analyses.
  - [l.945–947] **C:** I realise I am unclear about what panel D shows. I see that the graph shows results for four populations, but I fail to see how it shows both the CN gains as well as the CN losses for each of these populations because there is only one point cloud and box plot for each population. Maybe I am misinterpreting the label of the y axis. The notation "gained / lost" is ambiguous, as the slash could mean mathematical division or "respectively". Could the authors please clarify?
  - [l.943–944] **C:** I thank the authors for aiming to be more clear about what "proportion of CN losses" and "proportion of CN gains" means. However, the way this is now explained in the format of a pseudo-formula and an ambiguous use of the forward dash ("/") that can mean either division or "respectively" is very unfortunate. **R:** I request that the authors spend the effort to fully spell out the definitions in a full sentence.
- Fig. 5:
  - **C:** I thank the authors for amending the caption title to be more descriptive rather than interpretative.
  - [l.954] **S:** Replace "illustrates" by "illustrating".

Tables

- Table 1:
  - **C/S:** I am unclear about what the columns "Number of CN gained" and "Number of CN lost" mean. Are these the numbers of gene copies gained (lost) summed over all expanding and contracting genes, respectively?
  - [l.897–898] **S:** Remove "that" after "shows".

Supplementary Figures and Tables

- The authors should adjust the title of the paper in the file "Supplementary Figures and Tables" so that it matches the revised title of the paper.
- Figure S3:
  - **C:** Thank you for improving the colour scheme in panel B.
- Figure S5:
  - **C:** Thank you for clarifying to me that the dots represent genes, as is stated in the caption title.
- Figure S6:
  - **C:*** Thank you for expanding "CN value(s)".
- Figure S7 (new):
  - **C:** I understand this figure is new and that the authors inserted it in response to my Major Concern asking for a clarification about the contrast in population structure between genes with high vs. low V_{ST} and the interpretation of this contrast. I appreciate the additional PCA plots shown in Figure S7. **R:** My concern, though, is that PCs 1 and 2 in panel A as well as PC 1 in panel B are strongly affected by two (three?) outlier individuals. These outliers lead to an expansion of the respective axes and a compaction of the remaining points. These scalings make it difficult to see if there is population structure in the areas of the plot that would be of most interest. The scalings also make it difficult to compare the respective plots across panels A, B, and C. I wonder if the authors could insert another supplementary figure in which they remove the outliers mentioned above. My suspicion is that the apparent contrasts between panels A, B, and C will be reduced. This speculation is based on the fact that, along those axes not affected by the outliers, the difference between the range of values across panels A, B, and C is much less pronounced than for the axes affected by the outliers. In other words, I ask the authors to convince the readers that the signal they describe is not driven by a few outliers.
  - **R:** Please remove the words "The" at the beginning of the four sentences in the caption.
- Figure S10 (previously S9):
  - **C:** Thank you for clarifying that the figure is new and not reproduced from a previous publication. **R:** However, I still have a hard time understanding your caption. My issue is the word "overlapped with". What overlaps with what here? Can the authors please rewrite this sentence so it becomes clear what the figure shows?
- Figure S12:
  - **C:** This figure seems new (as an extraction from previous Figure S10B and C). **S:** Remove "The" from the beginning of the caption, but insert "the" before "copy number". Replace "using the" by "relative to the".
- Figure S13 (previously S11):
  - **C/S:** Thanks for adding a caption title. Please remove "The" at the beginning and write "Results of the redundancy analysis ... ".
  - **C:** Thanks for improving the wording in the second part of the caption.
  - **S:** Insert "The" before "RDA model shows" in line 2 of the caption. Remove "The" before "eigenvalues" and "proportion" in lines 3, 4, 6, 7, and 8 (at the beginning of the respective panel descriptions.
- Figure S14 (previously S12):
  - **C:** Thanks for adding a title sentence to the caption and describing what the layers of the circle plot in panel A show.
  - **R:** The first sentence describing panel (A) in the caption needs to be fixed. Suggestion:

"Strength of evidence for an association between gene CN and six climatic variables based on the RDA". From the center to the margin, the results are shown for Bio7, Bio8, ann_Rmean, PETDriestQuarter, annualPET, and PETColdestQuarter."

- Figure S15 (formerly S13):
  - **R:** Remove "The" before "summary" in the caption title.
- Table S3:
  - **R:** The table caption does not inform about the meaning of the numbers given outside and inside of the parentheses. One may suspect that the numbers report the true and false positives, but this remains unclear. Please describe what the numbers mean. The authors should clarify this point in the caption and perhaps in a footnote.
- Table S4:
  - **S:** Remove "The" at the beginning of the caption.
- Table S5:
  - **S:** Remove "The" at the beginning of the caption.