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 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 F_{ST}, 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.

Major Comments

1. 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.
2. 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 V_{ST} 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.
3. I am concerned that the V_{ST} 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).
4. I wonder if the authors have a good explanation for why only the highly CN-differentiated genes (high V_{ST}) 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 V_{ST}. 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 V_{ST}). 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-V_{ST} genes also reveal the expected population structure.

5. 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.
6. 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).
7. 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.
8. 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.
9. 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.
10. 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.

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*".

Abstract

- See 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 per se neutral.

Results

- [l.128] **Q:** Did you use ADMIXTURE (as stated in the Methods) or STRUCTURE (as stated here)?
- [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?
- [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)?
- [l.253] **S:** I wonder if it would be better to break the paragraph here, but to omit the break in l.247.
- [l.272] **S:** I suggest a paragraph break here.
- [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.

- [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.
- [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.

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

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.
- [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?
- [l.485] **R:** Please explain what V_{ST} is, or refer to the Supplementary Text.
- [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.

- [l.511] **C/S:** I felt that a transition was missing between the previous sentence and this sentence here ("LFMM . . . "). Please insert one.
- [l.513] **Q/R:** Did you implement LFMM2? If so, please provide a reference to the code.

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?
- 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.
  - [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".
- 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.
  - [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.
  - [l.788] **Q:** Did you use ADMIXTURE or STRUCTURE (it says "Structure analysis" here, but in l.474 you state you used ADMIXTURE)?
- 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.
  - [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).
- Fig. 5:
  - [l.821–822] **R:** The title sentence to me sounds to interpretative. I think the authors should chose more descriptive, neutral wording.
  - [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.

Tables

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

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.

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.
- 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.
- 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).
- 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").
- Figure S6:
  - What is meant by "CN value(s)"? Please define / write out once at least.
- Figure S7:
  - **C/R**: The description in the caption of panel C is unclear. Do you mean Do you mean "The number of genes associated with a response to . . . "? Please fix.
- 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.
- Figure S11:
  - **C/R:** The caption misses a title sentence. Please fix.
  - **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.
- Figure S12:
  - **R:** Please add a title sentence to the caption and describe what panel A shows.
- Figure S13:
  - **R:** Please revise the title sentence to be more informative.
- Table S3:
  - **C:** This table seems fully redundant to Figure S2 and could in my view be omitted.
- Table S4:
  - **C:** This table seems fully redundant to Fig. 1E and could in my view be omitted.
- 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.