

Peer Review Information

Journal: Nature Genetics

Manuscript Title: Grapevine pangenome facilitates trait genetics and genomic breeding

Corresponding author name(s): Professor Yongfeng Zhou, Hua Xiao, Professor Jinggui Fang

Editorial Notes:

Transferred manuscripts	This document only contains reviewer comments, rebuttal and decision letters for versions considered at Nature Genetics.
Reviewer comments in marked-up manuscript	In their review of the initial version of this manuscript, reviewer no. 3 added their comments to the manuscript file. These comments, excluding minor textual revisions, have been copied into this Peer Review File.

Reviewer Comments & Decisions:

Decision Letter, initial version:
--

7th Sep 2023

Dear Professor Zhou,

Your Article, "Grapevine pangenome facilitates trait genetics and genomic breeding" has now been seen by 2 referees. You will see from their comments copied below that while they find your work of considerable potential interest, they have raised quite substantial concerns that must be addressed. In light of these comments, we cannot accept the manuscript for publication, but would be very interested in considering a revised version that addresses these serious concerns.

We hope you will find the referees' comments useful as you decide how to proceed. If you wish to submit a substantially revised manuscript, please bear in mind that we will be reluctant to approach the referees again in the absence of major revisions.

To guide the scope of the revisions, the editors discuss the referee reports in detail within the team with a view to identifying key priorities that should be addressed in revision. In this case, we think both referees have provided constructive reviews aimed at strengthening the analyses and improving the presentation. As you will see from these comments, both referees have concerns regarding the reliability of the GWAS analyses since the phenotypic data were collected from a single planting

season (2016). Reviewer #2 also notes that the assembled genomes are not real T2T. We particularly ask that you perform additional experiments to enhance the robustness of the GWAS results, tone down some claims about "T2T" assemblies, and address all referee comments as thoroughly as possible with appropriate revisions. In addition, please cite and discuss the recent Science 2023 paper: "Dual domestications and origin of traits in grapevine evolution" [<https://www.science.org/doi/10.1126/science.add8655>]. We hope that you will find the prioritized set of referee points to be useful when revising your study.

If you choose to revise your manuscript taking into account all reviewer and editor comments, please highlight all changes in the manuscript text file. At this stage we will need you to upload a copy of the manuscript in MS Word .docx or similar editable format.

We are committed to providing a fair and constructive peer-review process. Do not hesitate to contact us if there are specific requests from the reviewers that you believe are technically impossible or unlikely to yield a meaningful outcome.

If revising your manuscript:

*1) Include a "Response to referees" document detailing, point-by-point, how you addressed each referee comment. If no action was taken to address a point, you must provide a compelling argument. This response will be sent back to the referees along with the revised manuscript.

*2) If you have not done so already please begin to revise your manuscript so that it conforms to our Article format instructions, available [here](#). Refer also to any guidelines provided in this letter.

*3) Include a revised version of any required Reporting Summary: <https://www.nature.com/documents/nr-reporting-summary.pdf>
It will be available to referees (and, potentially, statisticians) to aid in their evaluation if the manuscript goes back for peer review.
A revised checklist is essential for re-review of the paper.

Please be aware of our [guidelines on digital image standards](#).

You may use the link below to submit your revised manuscript and related files:

[redacted]

Note: This URL links to your confidential home page and associated information about manuscripts you may have submitted, or that you are reviewing for us. If you wish to forward this email to co-authors, please delete the link to your homepage.

If you wish to submit a suitably revised manuscript we would hope to receive it within 6 months. If you cannot send it within this time, please let us know. We will be happy to consider your revision so long as nothing similar has been accepted for publication at Nature Genetics or published elsewhere. Should your manuscript be substantially delayed without notifying us in advance and your article is eventually published, the received date would be that of the revised, not the original, version.

Please do not hesitate to contact me if you have any questions or would like to discuss the required revisions further.

Nature Genetics is committed to improving transparency in authorship. As part of our efforts in this direction, we are now requesting that all authors identified as 'corresponding author' on published papers create and link their Open Researcher and Contributor Identifier (ORCID) with their account on the Manuscript Tracking System (MTS), prior to acceptance. ORCID helps the scientific community achieve unambiguous attribution of all scholarly contributions. You can create and link your ORCID from the home page of the MTS by clicking on 'Modify my Springer Nature account'. For more information please visit www.springernature.com/orcid.

Thank you for the opportunity to review your work.

Sincerely,
Wei

Wei Li, PhD
Senior Editor
Nature Genetics
New York, NY 10004, USA
www.nature.com/ng

Reviewers' Comments:

Reviewer #1:

Remarks to the Author:

A. Summary of the key results

The authors present here a very large dataset, and genomic and statistical analyses that could be of interest for grape genomics and breeding. First, using long-read DNA technologies, they generated telomere-to-telomere genomes for 9 grapes, built a pangenome and detected structural variations in it. Then they imputed (with an unclear method) the structural variations derived from the pangenome to a short-read DNA dataset of 466 grapes (of which approx 30% were already in the public domain). Most of these grapes were grown in one of the contributing institutes, ZFRI-CAAS (Zhengzhou) and were phenotyped (with unclear methods) only once in 2016. The combination of genotyping and phenotyping allowed the authors to carry out GWAS and genomic prediction approaches. This work contains a large quantity of data, which testify of a large quantity of work, and this is commendable; it could certainly be attractive and of interest for the grape scientific community, both for genomics and breeding purposes. However, the paper is somehow overwhelmed by its own size; there is a large number of subsequent steps in methodology but the paper misses the precision needed for many of them, probably for lack of space and time. I would suggest to split the paper in two, one about genomics, pangenome and structural variations, the second about phenotyping, GWAS and genomic prediction.

B. Originality and significance: if not novel, please include reference

This work goes in the general direction of modern genomics, made possible by the huge advances in sequencing technologies and bioinformatics. It is well in line with the projects of several groups that I know of, and it arrives first in the publication process. It is quite remarkable in this view. However, as I wrote above, going too fast multiply the chances to be imprecise or even straight wrong. I encourage the team to continue in this line of research, but putting more effort in methodology development and presentation.

C. Data & methodology: validity of approach, quality of data, quality of presentation

In this section, I will present the main issues that make the paper not acceptable in the present state.

1) I didn't find a list of the individual grapes used in this study (it is stated that it is presented in Supplementary Table S1 but it is not). Along the paper one can find some short names (in figures or tables) but one has to guess to which exact grape they are referring to. There are no accession codes, no geographic location of the samples, no specification if the sequence was done in-house or found on NCBI, no effort to determine if they are true-to-type varieties, both for the newly generated sequences nor for those already in the public domain. For example in Figure 1: Muscat Humburger is not known in the international databases; maybe it is the well known Muscat Hamburg? *Muscadina rotundifolia*, ok but which one? There are three assemblies at NCBI, Carlos, Noble, Dixie, plus one on a private website (Trayshed).

The question of true-to-typeness is very important. I have been comparing the genotyping of the public data available at NCBI, from the paper here : <https://www.nature.com/articles/s41467-019-09135-8> with the genotypes produced by another former paper here : <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5805323/> and I found several instances of individuals carrying the same name but genetically very different, definitely not the same variety. So the problem of publishing genome sequences when the true-to-typeness of their initial accession is not certified, is that the risk to provide the scientific community with mislabelled and erroneous data is too high. Before publishing a genome, its identity must be certified.

There are two ways, to be used jointly, to determine if one accession name is true-to-type: A) one is using ampelography (check if the morphological traits measured on this individual corresponds to the reference measures for the named grape in international databases here: <http://www.eu-vitis.de/index.php>, based on the OIV descriptors http://www.eu-vitis.de/docs/descriptors/mcpd/Descriptors_EUVitisDB_11Jan12.pdf). B) The second is based on DNA SSR fingerprint, that one can easily obtain in-house and then compare to international databases (same as above). The authors could use some time to go through this procedure and verify their accession names.

2) I couldn't find any details of the methodology or protocol used for measuring the phenotypic traits. How do the authors measured "juiciness of flesh", "particularity of flavor", "thickness of skin", "color of skin" (visual or by an optic instrument, etc.) and so on? We need to know how the traits are measured to determine the degree of precision and confidence.

3) A third point that left me very perplexed is the attested presence of interspecific hybrids within the cultivated individuals used for the GWAS. It is well known that mixing taxa risks to generate (a large number) of false positive in the GWAS. I could not find any attempt to deal or to explain the strategy

the authors used to solve this problem.

4) I didn't find the explanation of the methodology used for the imputation of the detected SVs onto the 466 short read sequences.

5) The GWAS and genomic prediction chapters are based on a very little number of phenotypic measures, carried out only on three clones per genotype, in only one year (2016). The field experiment layout is also not presented (size, homogeneity of the soil, weather conditions, watering, grafting, virus and pathogen status, presence of buffer line, and so on).

6 The legends of figures and tables are not containing sufficient information. For example, just to mention Figure 1: a) What is Nobel? b) What is PHATE1, PHATE2? e-f) why there is no VL group in e but it is in f? f) How did you calculate the deleterious burden? There is no mention of it in Material and Methods

D. Appropriate use of statistics and treatment of uncertainties

Some methods were not explained at all. For example, how did the authors calculate gene expansion (139-141)? The deleterious burden? How they imputed the SVs onto the short-read DNA sequences?

I did not understand why the authors mapped the PacBios on the Sunshine Muscat (line 478) primary "genome" (I would say "haplome") and not on the reference Pinot Noir T2T, which is already published. Maybe because the reference PN, being nearly-homozygous, is missing some regions, but this should be explained.

Which tool did the authors use to call the SVs? At one point one understands maybe SYRI (line 467) but then they say they used 3 other softwares (Sniffles, CuteSV and PBSV; line 480). Please clarify.

Some methods were not clear, for example (214-215): "Furthermore, we calculated the percent phenotypic variation explanation (PVE) of 29 agronomic traits. How? Maybe the authors mean the PVE explained by (single or all) SNPs (what, average among all SNPs, or only those significantly associated? min-max, etc.)".

The objectives of the work from line 511 to 517 are unclear.

Concerning detection of traces of selection, I am sure there are better methods than using simply F_{st} and n .

To my knowledge, calculation of IBD (line 534) requires knowledge of the pedigree. Otherwise it is IBS. Authors should precise what they did in this paragraph.

In the GWAS (line 552), it is not clear to me why the authors used as covariate the 4 principal components of the PCA instead of the output of Admixture. Please explain, or correct.

Line 557 : Unclear. It seems to me that LDAK is used for association analysis. So it should be on the previous paragraph. Otherwise, explain. Also explain the definition of heritability here. Classically, it refers to quantitative trait measures, but here also the SNPs seem to play a role, so it could be a different definition. Maybe it is predictability instead ?

Line 572 : Transcriptome and gene ontology enrichment analyses. What is the objective of this analysis? PRJNA391820 is the RNAseq of Pinot noir and Thompson. How did you use this, in comparison with what, for what objective?

E. Conclusions: robustness, validity, reliability

As stated above, the paper suffers from its ambitions. Wanting to present too many data and analysis at once, the paper has no time and space to develop each section correctly enough.

There is no plan to deal with true-to-typeness of the studied varieties, which produces flaws in both validity of the results and reliability.

F. Suggested improvements: experiments, data for possible revision

I think the relevant solution to the problems described above is to split this paper in two. The first part could be about genomics, pangenome, SVs, and it is almost ready "as is".

The second part could comprise repeating the field measures (across more years, at least one more, with a better specified protocol), and redoing the GWAS and the genomic prediction. This would take one additional year but should not be a big problem if the reward is to publish clear and robust data.

Most importantly, first of all, I would suggest to better explore the true-to-typeness and description of the studied individuals. First using ampelography and SSR fingerprinting. Then presenting a better table with accession number, geo-location, characterisation, year of planting, donated by who, collected where... etc. I mean basic passport and description data. If one individual who is presently called, "Black monukka" for example, with the SSR fingerprint reveals to be not-true-to-type, then rename it as "Black monukka false". Another example, Muscat Humburger (Figure 1) does not exist in any international database. Maybe it is Muscat Hamburg ? Muscadina rotundifolia, which accession? and so on.

G. References: appropriate credit to previous work?

OK but I am missing one important reference : <https://www.science.org/doi/10.1126/science.add8655>

H. Clarity and context: lucidity of abstract/summary, appropriateness of abstract, introduction and conclusions

Generally the quality of the english language is sometime poor, with some uncomprehensible sentences, including the very first one in the abstract.

Quite frequently, some statements are not precise or even not correct. Just for example (but there are many instances similar to this) at:

Line 62: "more genomic sites affected by SVs than SNPs". This is no true if we speak of simple counts of number of SNPs versus number of SVs.

Line 75 : only~7% of the genes are shared while ~8% are unique to each individual". Something is obviously wrong here.

All the paragraph from line 199 to 212 as well.

Finally, the discussion is too short as compared to the large amount of results presented.

Reviewer #2:

Remarks to the Author:

In this manuscript, the authors de novo assembled 18 grapevine genomes, and then performed GWAS for 29 traits and genomic prediction for some traits. The work provides valuable genomic resource for grapevine. However, some points need to be carefully addressed based on current version.

1. The authors claimed that the assembled genomes are 'T-to-T gap-free' level. Based on the description, the mean contig N50 size of the de novo assemblies were 22.95 Mb, with approximately 1.5 contigs for each chromosome, which has not reach the level of real 'T-to-T'. The description of 'T-to-T gap-free' is not accuracy, the quality is exaggerated.
2. For the gene annotation need to be carefully considered. The number of protein coding genes ranged from 29,860 to 38,594. Why the accessions have such big divergence. Based on the results, the others characters of RR are not significant different from other accessions, but it has the longest genome but with the smallest protein coding gene number. Why? What happened? This needs to be clearly investigated.
3. Based on the results for the classification of all accessions (Figure 1b and Figure S7), the classification may not be so accurate. For example, the two Table subgroups, and ME, more like to be classified based on man-made classification, do not have a standard. In addition, the population structure also does not support the classification of the accessions to 6 groups. If this kind of classification is based on the consideration of some agronomic traits, there are some evolutionary process, which need to be clarified clearly.
Also, LD need to be investigated for each group to further check the accuracy of the classification.
4. For the GWAS, an obvious weakness is that the phenotyping is only from one year. In addition, based on the GWAS results, many traits did not get significant association loci. This makes the GWAS results are not reliable.
5. For the examples of the GWAS, since this work tried to emphasize the importance or powerful of pan-genome and SV in traits discovery, more examples to clearly show how SV associate with trait divergence need to be provided.
6. For the genomic prediction, although the prediction predictive power sounds to be a higher correlation, however, still based on the same population. This need to be validated using different population.

Some errors need to be carefully checked, for example: typo of Phylognic tree (Figure S7). There are some more.

Reviewer #3:

Remarks to the Author:

The paper describes the construction of a pan genome for the *Vitis vinifera* species and its use for identifying genetic determinants of simple and complex traits using a GWAS approach.

The paper is interesting but suffers from some severe drawbacks.

Firstly the paper is written quite carelessly so that it often makes it difficult to understand exactly what is meant and/or what has been done. It requires extensive rewriting after reading by a native English speaking person.

Scientifically there are a few severe issues that undermine the validity of the conclusions, the main one being the lack of any validation of the new data presented. I will list the main ones hereunder. There is no justification for the choice of genotypes for the construction of the pan genome. Normally one would choose individuals as to maximise the breadth of diversity within the species. Here there are mostly table grapes with a very limited representation of wine grapes and many of the genotypes are quite closely related (some of them even in parent offspring relationship). The relatedness among genotypes, even though not ideal for the purpose of the paper, could have been exploited for the purpose of validating the differences observed by comparing haplotypes that are identical by descent between the pairs of genotypes. The lack of any validation of the observed differences is a major shortcoming of the paper that makes it difficult to trust many of the conclusions. The fact that the total n. of genes goes from 31148 to 37972 within *V. vinifera* is hard to believe but would become believable if properly validated.

It is not clear why the authors felt the need for adding to the new assemblies that were produced using PacBio HIFI technology that lead to highly accurate and complete haplotype phased assemblies older assemblies obtained with the old PacBio technology that are not haplotype phased and that have a much poorer quality and completeness.

It is not clear why after building the different assemblies, the pan genome graph was not built by comparing assemblies with one another but rather by mapping reads of each accession against an individual reference to detect structural variants. No validation data is presented to determine the rate of false positive and false negative detection of SV using the procedure described that uses default parameters and a minimum read depth of three. The use of tools that have been ad hoc developed for pan genome reconstruction and used for example in the recent human pan genome effort, should have been tested, such as for example Pan Genome Graph Builder (pggb, [pggb \(notion.site\)](https://github.com/PanGenome/pggb)) that makes use of wfmash or Minigraph-Cactus (<https://www.nature.com/articles/s41587-023-01793-w>). The main purpose of the paper is to show that the pan genome approach is better at identifying genetic determinants of simple and complex traits using a GWAS approach than the use of a single reference genome. This should have been experimentally proven by performing two independent experiments with the same data set: one where the reference used is the pan genome graph, the other one where the reference used is a single genotype such as for example the widely used PN40024 in its new T2T version.

The finding of many correlations among berry size related traits and berry content traits is not surprising if berry content is measured on a concentration basis and not on a per berry basis. It is well known that larger berries at veraison lead to decreased concentration of many berry metabolites, so this issue would deserve a much greater attention.

The causative mutations for fruit colour are well known and are located in chr. 2 around 14 Mpb. The involvement of the inversion is not demonstrated and the data presented here shows that there is an inversion on chr 2 in this region in only one accession (*Muscadinia rotundifolia*, see supplementary figure 3 for chromosome 2). The large region of LD around the berry color is presumably due to a very recent and intense selective sweep followed by a limited number of meiotic events that led to the production of white varieties. The involvement of the 5 Mbp inversion in the trait determination is negated by the genome assemblies presented here. No inversion is visible between white and red haplotypes. So either these assemblies are incorrect or the previous observation of the presence of the inversion is not correct).

No biological/genetic explanation is provided for the extremely low LD observed between SNPs and SVs. A proper explanation should be provided.

Other minor comments are provided in the attached PDF file.

Summarized Comments from attached PDF prepared by the Editorial Staff:

Introduction

Line 54: A very recent paper in Science advocating a double domestication of grapevine should be cited here also (<https://www.science.org/doi/10.1126/science.add8655>)

Line 75: where does this conclusion come from?

Line 91: what is meant exactly by wild grape accessions? Are these non vinifera Vitis species? Then they should not be called wild grape accessions

Results

Line 109: See comment above, this should not be called wild grape because this could be confused with Vitis vinifera subsp. sylvestris, the wild ancestor of cultivated grapes.

Line 169-173: this sentence as it is is very cryptic. there should be an explanation of what is meant here.

Line 183-185: too broad a claim for the data presented

Line 194: which ones are they?

Line 234-243: This region was shown to have undergone a severe selective sweep in the domestication step (see for example ref. 7). The results should be interpreted in light of this observation.

Line 303: The causative mutations for fruit color are well known and are located in chr. 2 around 14 Mpb. The involvement of the inversion is not demonstrated and the data presented here shows that there is an inversion on chr 2 in this region in only one accession (see supplementary figure 3 for chromosome 2).

Discussion

Line 339-343: When it comes to wine grapes the reason for the lack of breeding efforts is none of those listed here but rather the conservative attitude of the business and the complex regulatory system for quality wine appellations that encourages the continued use of traditional old varieties.

Line 349-352: These assemblies have rather poor quality and should not be confused with the HiFi assemblies

Line 357-360: I do not see exactly where this conclusion is derived from: a careful comparison with the wild grapes (Vitis vinifera subsp. sylvestris) is needed to reach such a conclusion

Line 365-369: The large region of LD around the berry color is presumably due to a very recent and intense selective sweep followed by a limited number of meiotic events that led to the production of white varieties. The involvement of the 5 Mbp inversion in the trait determination is negated by the genome assemblies presented here. No inversion is visible between white and red haplotypes.

Line 477-480: Why was the pan genome graph built by aligning reads to one specific assemblies and not by comparing assembled haplotypes as usually done?

Line 550-555: it is not explained what was the set of variants/genotypes that were used in the GWAS analysis

Author Rebuttal to Initial comments

Reviewers' Comments:

Reviewer #1: Remarks to the Author:

A. Summary of the key results: The authors present here a very large dataset, and genomic and statistical analyses that could be of interest for grape genomics and breeding. First, using long-read DNA technologies, they generated telomere-to-telomere genomes for 9 grapes, built a pangenome and detected structural variations in it. Then they imputed (with an unclear method) the structural variations derived from the pangenome to a short-read DNA dataset of 466 grapes (of which approx 30% were already in the public domain). Most of these grapes were grown in one of the contributing institutes, ZFRI-CAAS (Zhengzhou) and were phenotyped (with unclear methods) only once in 2016. The combination of genotyping and phenotyping allowed the authors to carry out GWAS and genomic prediction approaches.

RESPONSE: We are grateful for the reviewer's time, patience and constructive comments. We had phenotypic measurements for two years while only 2016 data were presented in previous version for simplification. We noticed the importance of replication in phenotyping, so we added 2017 data and redone the analyses using both the phenotypic measurements across two years (in 2016 and 2017). Based on the reviewer's comment, we updated the GWAS and GS analyses by combining two years of data and rewrote the Results and Methods. Please see details in Results L178-180, methods L535-538 and Supplementary Note 8.

This work contains a large quantity of data, which testify of a large quantity of work, and this is commendable; it could certainly be attractive and of interest for the grape scientific community, both for genomics and breeding purposes. However, the paper is somehow overwhelmed by its own size; there is a large number of subsequent steps in methodology but the paper misses the precision needed for many of them, probably for lack of space and time. I would suggest to split the paper in two, one about genomics, pangenome and structural variations, the second about phenotyping, GWAS and genomic prediction.

RESPONSE: We are grateful for the reviewer's constructive comment. In the revised manuscript, we had simplified the overall presentation and moved the detailed methods to the supplementary file. We apologize again about the lack of clarity, but we would like to keep the whole story complete in one paper, because the topics are intricately linked. The pangenome contributes to the identification of SVs that are critical for capturing missing heritability, thereby enhancing our understanding of the genetic basis of complex agronomic traits for grapevine breeding.

B. Originality and significance: if not novel, please include reference

This work goes in the general direction of modern genomics, made possible by the huge advances in sequencing technologies and bioinformatics. It is well in line with the

projects of several groups that I know of, and it arrives first in the publication process. It is quite remarkable in this view. However, as I wrote above, going too fast multiply the chances to be imprecise or even straight wrong. I encourage the team to continue in this line of research, but putting more effort in methodology development and presentation.

RESPONSE: We are grateful for the reviewer's encouragement. We have carefully revised the methods and presentations and checked the citation of relevant references. All statistical analyses and procedures have been described in details in Supplementary Methods and Supplementary Notes.

C. Data & methodology: validity of approach, quality of data, quality of presentation

In this section, I will present the main issues that make the paper not acceptable in the present state.

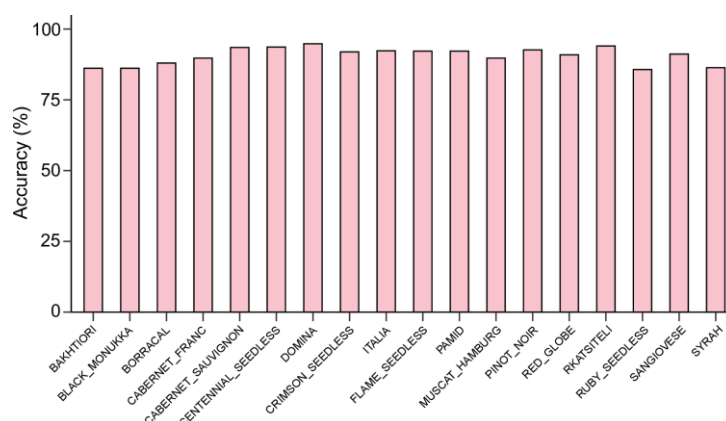
1) I didn't find a list of the individual grapes used in this study (it is stated that it is presented in Supplementary Table S1 but it is not). Along the paper one can find some short names (in figures or tables) but one has to guess to which exact grape they are referring to. There are no accession codes, no geographic location of the samples, no specification if the sequence was done in-house or found on NCBI, no effort to determine if they are true-to-type varieties, both for the newly generated sequences nor for those already in the public domain. For example in Figure 1: Muscat Humburger is not known in the international databases; maybe it is the well known Muscat Hamburg? *Muscadina rotundifolia*, ok but which one? There are three assemblies at NCBI, Carlos, Noble, Dixie, plus one on a private website (Trayshed).

RESPONSE: We apologize for the lack of clarity. We have revised Supplementary Table S6 to include short names and corresponding full cultivar names for each accession. We have also carefully identified all true-to-type accessions by genotype and morphological features. Additionally, previously published geographic locations and collection information from NCBI has been integrated into the revised Supplementary Table S6 and properly cited in the manuscript. With apologies for this typo, we revised to Muscat Hamburg. *Muscadina rotundifolia* cv. Noble was corrected in Supplementary Table S1 as well.

The question of true-to-typeness is very important. I have been comparing the genotyping of the public data available at NCBI, from the paper here : <https://www.nature.com/articles/s41467-019-09135-8> with the genotypes produced by another former paper here : <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5805323/> and I found several instances of individuals carrying the same name but genetically very different, definitely not the same variety. So the problem of publishing genome sequences when the true-to-typeness of their initial accession is not certified, is that the risk to provide the scientific community with mislabelled and erroneous data is too high. Before publishing a genome, its identity must be certified. There are two ways, to be

used jointly, to determine if one accession name is true-to-type: A) one is using ampelography (check if the morphological traits measured on this individual corresponds to the reference measures for the named grape in international databases here: <http://www.eu-vitis.de/index.php>, based on the OIV descriptors http://www.eu-vitis.de/docs/descriptors/mcpd/Descriptors_EUVitisDB_11Jan12.pdf). B) The second is based on DNA SSR fingerprint, that one can easily obtain in-house and then compare to international databases (same as above). The authors could use some time to go through this procedure and verify their accession names.

RESPONSE: Thanks for your careful checks. We are sorry for our carelessness. Based on your comments, we have made the corrections to make the word harmonized within the whole manuscript.. We have carefully checked the data downloaded from NCBI. All sequencing data utilized in our manuscript was validated against breeding historical documents from Zhengzhou Fruit Research Institute of the Chinese Academy of Agricultural Sciences, Zhengzhou, China (ZFRI-CAAS), which was described in Supplementary Note 1. Some varieties have been checked on the international database (<https://www.vivc.de/index.php>). Additionally, we performed a SNP profile analyses to confirm the true-to-type of varieties (**Supplementary Table 7**). We download SNP markers from VIVC database and identified these genotypes in our sequenced grape accessions. The genotypes of sequenced grape accessions are identical with the VIVC database records (**Response Figure 1**).



Response Figure 1. The accuracy of SNP profile in identify true-to-type of grape varieties.

2) I couldn't find any details of the methodology or protocol used for measuring the phenotypic traits. How do the authors measured "juiciness of flesh", "particularity of flavor", "thickness of skin", "color of skin" (visual or by an optic instrument, etc.) and so on? We need to know how the traits are measured to determine the degree of precision and confidence.

RESPONSE: Thanks for reviewer's kind suggestions, the measurements and criteria of 21 phenotypes were based on the standardized protocols from the International Organization of Vine and Wine (OIV), with minor adaptations made for certain traits as necessary. We performed additional analyses to precisely measure eight flavor content traits. We have presented detailed descriptions of the methods of phenotypic

measurements for all 29 agronomic traits in the Methods L538-544 and Supplementary Note 7.

3) A third point that left me very perplexed is the attested presence of interspecific hybrids within the cultivated individuals used for the GWAS. It is well known that mixing taxa risks to generate (a large number) of false positive in the GWAS. I could not find any attempt to deal or to explain the strategy the authors used to solve this problem.

RESPONSE: We are grateful for the reviewer's comment. To address the concern of false positives, we have provided information about putative hybrids and removed them for the downstream GWAS analyses. Additionally, to account for population structure, we have implemented a mixed linear model (MLM) approach to estimate heritability for each trait as described by Yu et al. (2005, Nature Genetics). This MLM model considers both fixed effects (using PCA components to capture population structure) and random effects (using a kinship matrix to model individual covariance):

$$Y = \text{SNP} + Q \text{ (or PCs)} + \text{Kinship} + e$$

The use of both PCA components and kinship in the model allows proper control for population structure and relatedness between individuals. We have added details on this approach to the Methods L560-565 and Supplementary Method 10.1.

Reference:

Yu, J., Pressoir, G., Briggs, W. et al. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nat Genet 38, 203–208 (2006). <https://doi.org/10.1038/ng1702>.

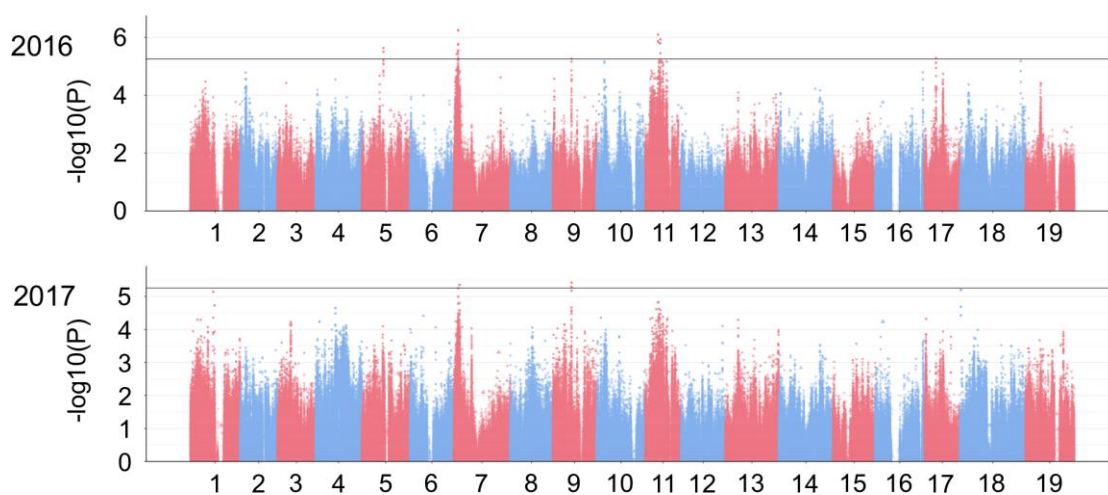
4) I didn't find the explanation of the methodology used for the imputation of the detected SVs onto the 466 short read sequences.

RESPONSE: Thank the reviewer for raising this important point. We combined graph deconstruction and assemblies' alignment to generate the pangenome SV dataset. This dataset was used for population-scale SV genotyping. We imputed the SVs onto the short-read DNA datasets of 466 grapes by PanGenie tools: (1) We filtered SVs corresponding to top-level bubbles, retaining non-top-level bubbles. (2) PanGenie-index was executed to process input variants and determine unique k-mers in the graph. We combined paired reads to genotype SV for each accession and generated a merged SV map. (3) We distinguished biallelic variants and variants nested inside graph bubbles (multiallelic variants). The biallelic bubble-based VCF was utilized for downstream GWAS analysis, while the multiallelic bubble-based VCF was used for haplotype graph classification. For details, see Supplementary Note 4.3 and Supplementary Method 6.

5) The GWAS and genomic prediction chapters are based on a very little number of phenotypic measures, carried out only on three clones per genotype, in only one year

(2016). The field experiment layout is also not presented (size, homogeneity of the soil, weather conditions, watering, grafting, virus and pathogen status, presence of buffer line, and so on).

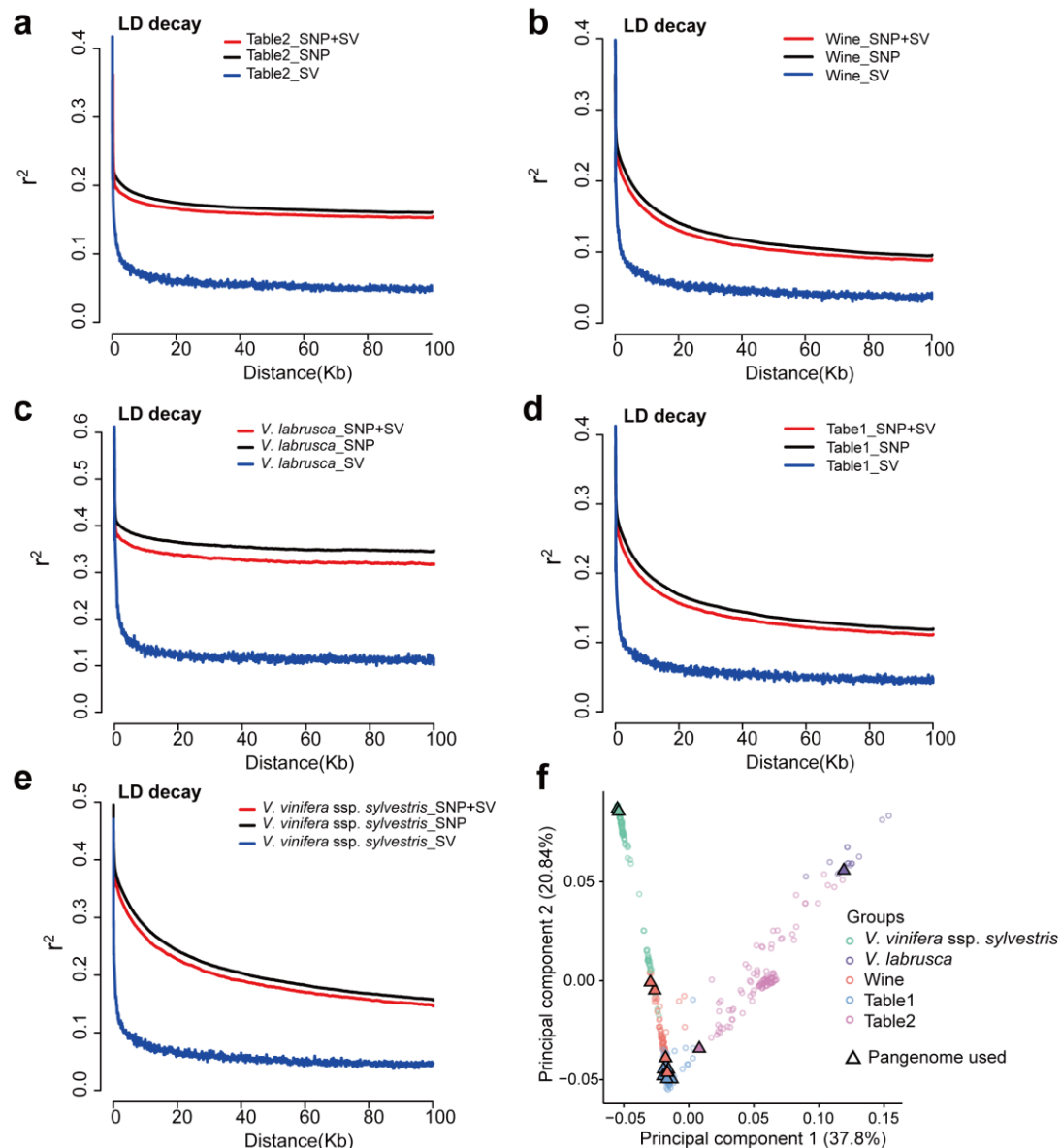
RESPONSE: Thank you for raising this issue. To strengthen the GWAS and genomic prediction analysis, we incorporated phenotype data on three clones per genotype across two years (in 2016 and 2017), for example, we performed GWAS based on the content of soluble solids, and the results between two years have good consistency (**Supplementary Figure 19**). In addition, we have provided updated details on the field experiments and cultivation conditions (**Supplementary Note 1.2**), and phenotypic measurements (**Supplementary Note 7**) in the supplementary file.



Response Figure 2. Two independently GWAS of the content of soluble solids in 324 grape accessions based on pangenome SNPs in 2016 and 2017 years.

6) The legends of figures and tables are not containing sufficient information. For example, just to mention Figure 1: a) What is Nobel? b) What is PHATE1, PHATE2? e-f) why there is no VL group in e but it is in f? f) How did you calculate the deleterious burden? There is no mention of it in Material and Methods

RESPONSE: Thanks for your comment. We have checked and revised all figure and table legends to unify abbreviations of accession names, groups, and corresponding color labels. Nobel is the variety of *Muscadinia rotundifolia*; we have moved PHATE analysis and deleterious burden to Supplementary Figure 9 and 12; and added descriptions in the legend. We have provided a principal component analysis (PCA) to illustrate the population structure of sequenced grape accessions. We have added LD analysis of *V. labrusca* grapes in Supplementary Figure 13. The calculation of deleterious burden was described in Supplementary Method 9.



Response Figure 3. LD analysis of SNPs and SVs in five grape populations and PCA analysis of SNPs. a-e, The LD decay of five grape populations (Table2, Wine, *V. labrusca*, Table1 and *V. vinifera* ssp. *sylvestris*) based on three datasets (SNP+SV, SNP and SV). f, The PCA of 466 sequenced samples, while the assemblies in pangenome construction were highlighted.

D. Appropriate use of statistics and treatment of uncertainties

Some methods were not explained at all. For example, how did the authors calculate gene expansion (139-141)? The deleterious burden? How they imputed the SVs onto the short-read DNA sequences?

RESPONSE: We are grateful for the reviewer's comment. We have expanded the descriptions of all methods and statistical analysis throughout the revised manuscript to

enhance reproducibility. The core and non-essential gene family analysis was performed among 18 grape accessions. The approach for calculating gene family expansion was described in Methods in L473-476, based on the improved gene annotations. We have added the methodology for calculating deleterious burden in Supplementary Method 9. As mentioned in response above, pangenome-based SV and PanGenie tools were used for SV genotype on 466 short-read sequences (Methods L502-508 and Supplementary Method 6).

I did not understand why the authors mapped the PacBios on the Sunshine Muscat (line 478) primary “genome” (I would say “haplome”) and not on the reference Pinot Noir T2T, which is already published. Maybe because the reference PN, being nearly-homozygous, is missing some regions, but this should be explained.

RESPONSE: Thanks for raising this import issue. . We used Shine Muscat for the reference due to it is one of the most popular table grapes. In the revised manuscript, we used the reference PNT2T instead of Shine Muscat, as PNT2T was the frequent used one. The graph pangenome can capture information on SVs and SNPs without being affected by the nearly homozygous of a single genome

Which tool did the authors use to call the SVs? At one point one understands maybe SYRI (line 467) but then they say they used 3 other softwares (Sniffles, CuteSV and PBSV; line 480). Please clarify.

RESPONSE: We apologize for the lack of clarity. We rephrased the manuscript. We imputed the structural variations onto the short-read DNA datasets of 466 grapes based on MC pangenome by PanGenie tools as suggested by Reviewer #3. In the revised manuscript, we used SyRI (v1.6.3) and Plotsr (v1.0.0) programs to perform comparative genomics analysis (Methods L468-470 and Supplementary Method 6).

Some methods were not clear, for example (214-215): “Furthermore, we calculated the percent phenotypic variation explanation (PVE) of 29 agronomic traits. How? Maybe the authors mean the PVE explained by (single or all) SNPs (what, average among all SNPs, or only those significantly associated? min-max, etc.)”.

RESPONSE: We have re-written this part according to the reviewer’s suggestion (L575-578).. We focused on phenotypic variation explanation (PVE) for significantly associated SNP or SV sites for each agronomic trait. The PVE of the most significantly associated single variant was estimated using the following formula (Teslovich, *Nature*, 2010):

$$PVE = [2 \times (\beta^2) \times MAF \times (1 - MAF)] / [2 \times (\beta^2) \times MAF(1 - MAF) + ((s.e. \times \beta)^2) \times 2 \times N \times MAF \times (1 - MAF)]$$

Where N represents the sample size, s.e. is the standard error of the effect size of genetic

variants, beta is the effect size of genetic variants, and MAF is the minor allele frequency of the target marker. We have revised Methods L575-581 and Supplementary Method 10.2.

Reference:

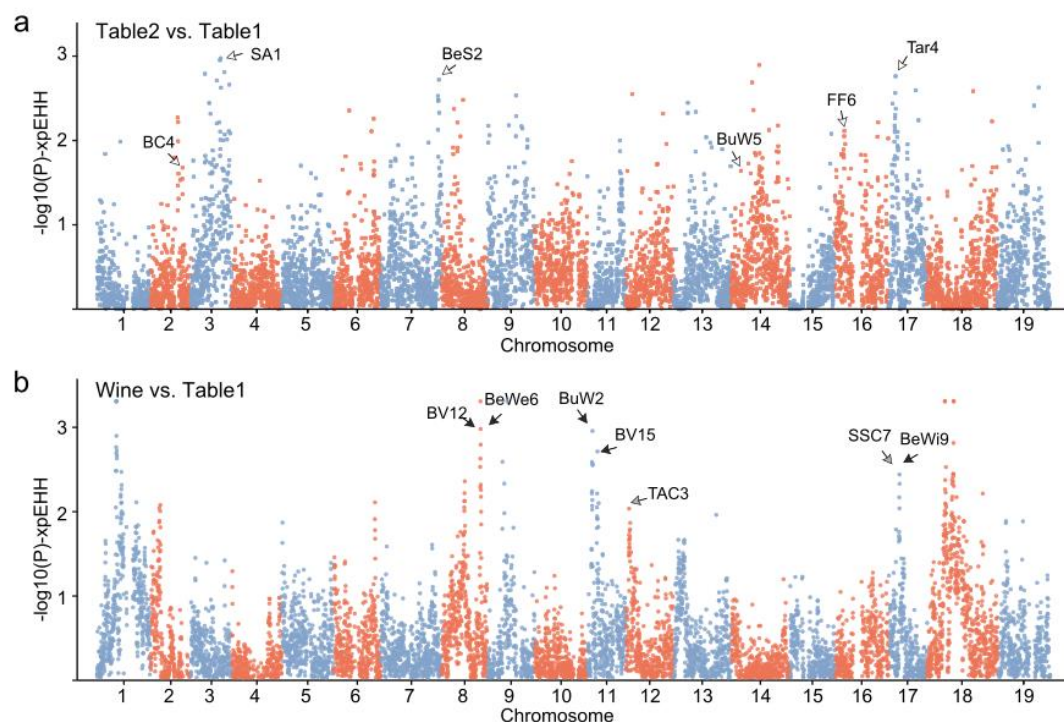
Teslovich, T.M. et al. Biological, clinical and population relevance of 95 loci for blood lipids. *Nature* 466, 707-13 (2010).

The objectives of the work from line 511 to 517 are unclear.

RESPONSE: We are grateful for the reviewer's suggestion, and this part was rewritten. We reconstructed the paragraph and make it clear. Done.

Concerning detection of traces of selection, I am sure there are better methods than using simply F_{st} and π .

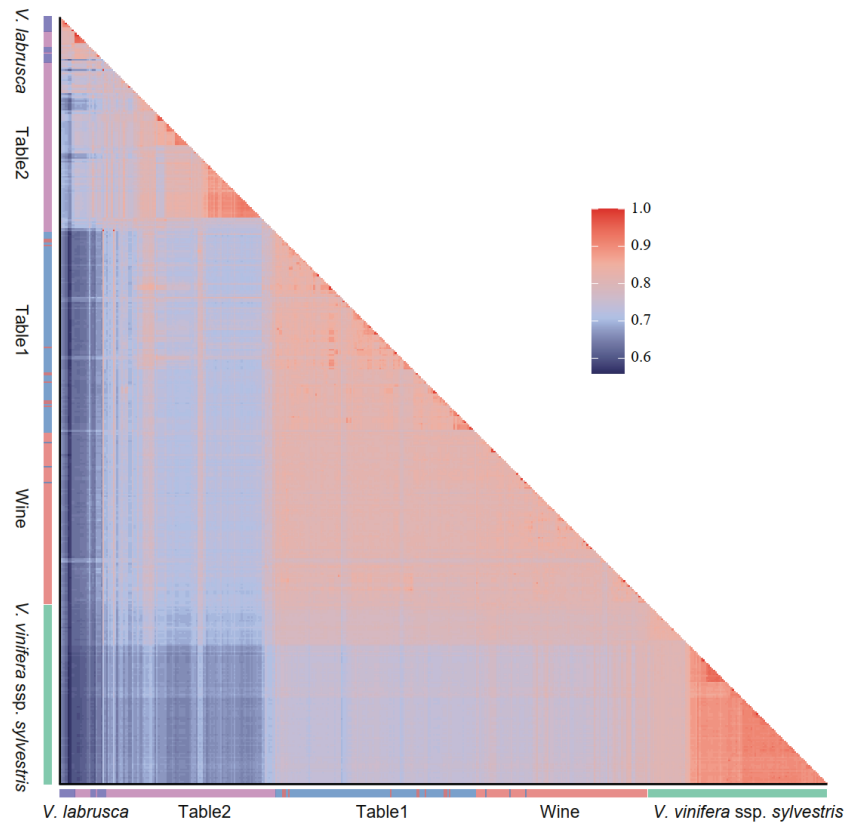
RESPONSE: Thanks for this suggestion, we have calculated the genome-wide XP-EHH statistics to present the divergent selection of two grapevine groups. We also used F_{st} and π for supporting the detection of selection. We have revised Results L256-259 and Methods L529-533.



Response Figure 4. The XP-EHH analysis among different grape populations. a, the XP-EHH genome-wide scanning between Table1 and Table2 grape populations, and six GWAS loci located in the divergent regions were highlighted. b, the XP-EHH genome-wide scanning between Table1 and Wine grape populations. Two metabolites related GWAS loci and fruit size related GWAS loci were highlighted.

To my knowledge, calculation of IBD (line 534) requires knowledge of the pedigree. Otherwise it is IBS. Authors should precise what they did in this paragraph.

RESPONSE: Thanks for pointing out this problem. We calculated IBS values using PLINK program based on the genomic variation map in **supplementary Fig. 11** (L162-164).



Response Figure 5. Identical by state (IBS) values between different individuals.

Heatmap colors indicate the distant of IBS blocks for each pairwise comparison

In the GWAS (line 552), it is not clear to me why the authors used as covariate the 4 principal components of the PCA instead of the output of Admixture. Please explain, or correct.

RESPONSE: We sincerely thank the reviewer for careful reading. We used the top four PCs from PCA as covariates because PCs are orthogonal vector of variation. The top four PCs have been shown to effectively capture population structure. In contrast, Admixture may overfit structure when there is trait associated subtle subpopulations as shown by Zhang et al. (2010, *Nature Genetics*). We have added descriptions in Method section in L523-528 and Supplementary Method 10.1.

Reference:

Zhang, Z., Ersoz, E., Lai, CQ. et al. Mixed linear model approach adapted for genome-wide association studies. *Nat Genet* 42, 355–360 (2010). <https://doi.org/10.1038/ng.546>

Line 557: Unclear. It seems to me that LDAK is used for association analysis. So it should be on the previous paragraph. Otherwise, explain. Also explain the definition of heritability here. Classically, it refers to quantitative trait measures, but here also the SNPs seem to play a role, so it could be a different definition. Maybe it is predictability instead?

RESPONSE: We apologize for unclear description. The genome SNP heritability ($G_h^2_{\text{SNP}}$) of a trait is the fraction of phenotypic variance explained by additive contributions from all pangenome SNPs. Similarly, we calculate the SV heritability ($G_h^2_{\text{SV}}$) using all pangenome SVs. The LDAK35 (v5.2) program provides a model for estimating $G_h^2_{\text{SNP}}$ and $G_h^2_{\text{SV}}$ by deriving approximate relationships between the heritability of a variant and minor allele frequency (MAF) (Speed D et al., 2017, *Nature Genetics*; Speed D et al., 2020, *Nature Genetics*). Consider the heritability model of the form:

$$E[h^2_j] = \tau_i w_j [f_j(1-f_j)]^{(1+\alpha)}$$

where w_j is the weighting for each SNP or SV (j), and f_j is its MAF. And τ are the corresponding coefficients and are estimated from the SNP or SV. The parameter α determines how the expected heritability contributed by a variant depends on its minor allele frequency (<https://dougsspeed.com/technical-details/>). To improve robustness, we inferred the $G_h^2_{\text{SNP}}$ and $G_h^2_{\text{SV}}$ based on kinship matrix generated by pangenome SNPs or SVs (for example in Wang et al., 2020, *Nature Genetics*).

We have added descriptions of genome selection in Methods L583-599 and Supplementary Note 9.

Reference:

Speed D, Cai N; UCLEB Consortium, Johnson MR, Nejentsev S, Balding DJ. Reevaluation of SNP heritability in complex human traits. *Nat Genet*. 2017;49(7):986-992. doi:10.1038/ng.3865

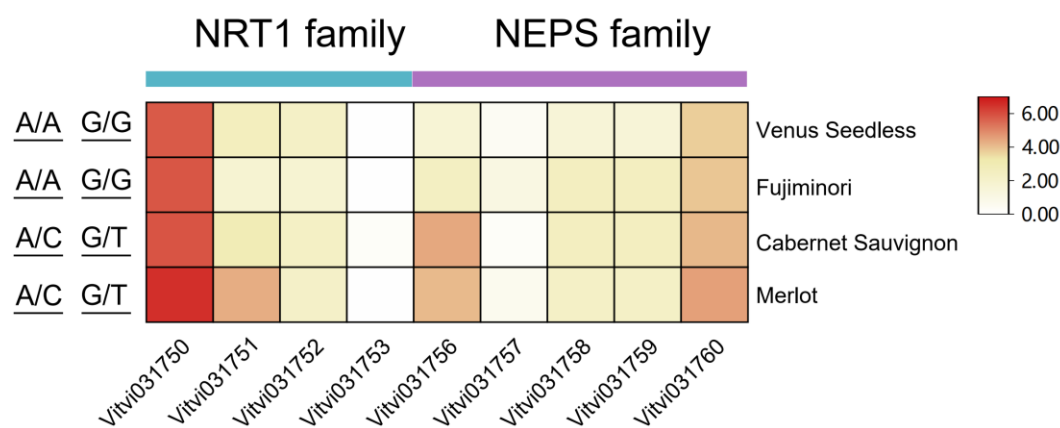
Speed D, Holmes J & Balding DJ, Evaluating and improving heritability models using summary statistics. *Nat Genet* 52, 458–462 (2020). <https://doi.org/10.1038/s41588-020-0600-y>

Wang, B., Lin, Z., Li, X. et al. Genome-wide selection and genetic improvement during modern maize breeding. *Nat Genet* 52, 565–571 (2020). <https://doi.org/10.1038/s41588-020-0616-3>

Line 572: Transcriptome and gene ontology enrichment analyses. What is the objective of this analysis? PRJNA391820 is the RNAseq of Pinot noir and Thompson. How did

you use this, in comparison with what, for what objective?

RESPONSE: We apologize for the lack of clarity. In revised manuscript, we used GO enrichment analyses in the comparative genomics of Table1 and Table2 grape groups, as well as the comparative genomics of Wine and Table1 grape groups, to investigate the putative divergent selection in different grape groups. In revised manuscript, we re-collected 4 RNA seq data from a large RNA-seq dataset (PRJNA565689) and identified gene expression. We compared gene expression in different grape accessions with different haplotypes to infer the candidate genes in GWAS analysis. We had added this information in Methods L610-617.



Response Figure 6. The expression analysis of four candidate NRT1 family members and five NEPS family members.

E. Conclusions: robustness, validity, reliability

As stated above, the paper suffers from its ambitions. Wanting to present too many data and analysis at once, the paper has no time and space to develop each section correctly enough. There is no plan to deal with true-to-typeness of the studied varieties, which produces flaws in both validity of the results and reliability.

RESPONSE: Thanks for raising this question. We revised the presentation of the manuscript. We also performed careful quality checks on all data and collected samples to ensure accurate correspondence. The revised manuscript has been improved substantially, due principally to the reviewers' constructive comments.

F. Suggested improvements: experiments, data for possible revision

I think the relevant solution to the problems described above is to split this paper in two. The first part could be about genomics, pangenome, SVs, and it is almost ready “as is”. The second part could comprise repeating the field measures (across more years, at least one more, with a better specified protocol), and redoing the GWAS and the genomic prediction. This would take one additional year but should not be a big problem if the reward is to publish clear and robust data.

RESPONSE: We thank the review for this suggestion. Our primary goal is to use grape graph pangenome to enhance the understanding of complex agronomic traits, and facilitate grapevine breeding. For our purposes, we feel that keeping the article in one would be of great benefit for understanding the data and the genetic basis of key agronomic traits. We had substantially revised the presentation of manuscript, and hope the current version is more suitable for publication.

Most importantly, first of all, I would suggest to better explore the true-to-typeness and description of the studied individuals. First using ampelography and SSR fingerprinting. Then presenting a better table with accession number, geo-location, characterisation, year of planting, donated by who, collected where... etc. I mean basic passport and description data. If one individual who is presently called, “Black monukka” for example, with the SSR fingerprint reveals to be not-true-to-type, then rename it as “Black monukka false”. Another example, Muscat Humburger (Figure 1) does not exist in any international database. Maybe it is Muscat Hamburg? *Muscadinia rotundifolia*, which accession? and so on.

RESPONSE: Thanks for your careful checks. To address concerns about confusion between data and sample names, we have prepared a very detailed introduction on the sequenced samples and downloaded WGS data, including the cultivar breeding background, collection, cultivation, and field performance. The details are presented in Supplementary Note 1. The sample names, such as “Black monukka” and “Muscat Hamburg”, have also been unified throughout the text. We used the assembly of *Muscadinia rotundifolia* as the outgroup.

G. References: appropriate credit to previous work?

OK but I am missing one important reference:
<https://www.science.org/doi/10.1126/science.add8655>

RESPONSE: Thank the reviewer for pointing this out. We added this citation in L57-58 and L346-347. We had also checked all references.

H. Clarity and context: lucidity of abstract/summary, appropriateness of abstract, introduction and conclusions

Generally the quality of the english language is sometime poor, with some uncomprehensible sentences, including the very first one in the abstract.

Quite frequently, some statements are not precise or even not correct. Just for example (but there are many instances similar to this) at:

Line 62: “more genomic sites affected by SVs than SNPs”. This is no true if we speak of simple counts of number of SNPs versus number of SVs.

RESPONSE: We have polished the English writing throughout the manuscript. We had

also rephrased the sentence. Please see L58-62.

Line 75 : only~7% of the genes are shared while ~8% are unique to each individual”. Something is obviously wrong here. All the paragraph from line 199 to 212 as well.

RESPONSE: Thank the reviewer for the comment. We have revised these sentences in L75-78.

Finally, the discussion is too short as compared to the large amount of results presented.

RESPONSE: Thank you. We have added the discussion of the recent science paper and extended the graph-based pangenome in future grape breeding in L346-347.

Reviewer #2:

Remarks to the Author:

In this manuscript, the authors de novo assembled 18 grapevine genomes, and then performed GWAS for 29 traits and genomic prediction for some traits. The work provides valuable genomic resource for grapevine. However, some points need to be carefully addressed based on current version.

RESPONSE: We sincerely appreciate the valuable comments. We had revised the manuscript accordingly and we hope the current version is more suitable for publication.

1. The authors claimed that the assembled genomes are ‘T-to-T gap-free’ level. Based on the description, the mean contig N50 size of the de novo assemblies were 22.95 Mb, with approximately 1.5 contigs for each chromosome, which has not reach the level of real ‘T-to-T’. The description of ‘T-to-T gap-free’ is not accuracy, the quality is exaggerated.

RESPONSE:

Based on reviewer’s comments, we enhanced the gapless assemblies through (1) the generation of 150-kb ONT ultra-long reads; (2) the development of a strategy combining HiFi and ONT reads to enhance assemblies; (3) quality assessment using three independent methods, including switch error quality value and hamming error. Consequently, all 18 haplotypes achieved gapless assembly. Additionally, we characterized telomere and centromere sequences. The quality of these assemblies now matches that of previously published T2T genomes, such as grape T2T genome (Shi et al., 2023, *Hortic Res.*) and maize T2T genome (Chen et al., 2023, *Nat Genet.*). Please see Supplementary Note 3 for details.

Reference:

Shi X, Cao S, Wang X et al. The complete reference genome for grapevine (*Vitis vinifera* L.) genetics and breeding. *Hortic Res.* 2023 Apr 4;10(5): uhad061. doi: 10.1093/hr/uhad061.

Chen J, Wang Z, Tan K et al. A complete telomere-to-telomere assembly of the maize genome. *Nat Genet.* 2023 Jul;55(7):1221-1231. doi: 10.1038/s41588-023-01419-6.

2. For the gene annotation need to be carefully considered. The number of protein coding genes ranged from 29,860 to 38,594. Why the accessions have such big divergence. Based on the results, the others characters of RR are not significant different from other accessions, but it has the longest genome but with the smallest protein coding gene number. Why? What happened? This needs to be clearly investigated.

RESPONSE: We think this is an excellent suggestion. We have improved gene annotations by a uniform pipeline combined RNA-seq, homologous protein and model-based prediction. We re-investigated gene expansion and the pangenome based on the improved gene annotations. After updating, the number of protein coding genes ranged from 36,093 to 37,487.

3. Based on the results for the classification of all accessions (Figure 1b and Figure S7), the classification may not be so accurate. For example, the two Table subgroups, and ME, more like to be classified based on man-made classification, do not have a standard. In addition, the population structure also does not support the classification of the accessions to 6 groups. If this kind of classification is based on the consideration of some agronomic traits, there are some evolutionary process, which need to be clarified clearly.

Also, LD need to be investigated for each group to further check the accuracy of the classification.

RESPONSE: Thank the reviewer for the suggestion. We have clarified the population classification combined information from PCA, ADMIXTURE analysis, phylogenetic tree, PHATE analysis, IBS values and incorporated with recent breeding histories. As a result, our analyses identified three major species (*V. labrusca*, *V. vinifera* ssp. *sylvestris* and *V. vinifera* ssp. *vinifera*). The modern cultivars of *V. vinifera* ssp. *vinifera* were separated into three groups: the wine-making varieties from Europe and the Middle-East (Wine), and two type of table grapes, one from Europe and Eastern Asia (Table1) and one from hybridization with the *V. labrusca* grape (Table2, *V. labrusca* × *V. vinifera*). We have revised in Results L153-162.

Based on the reviewer's comments, we performed LD based on pangenome SNPs and pangenome SVs in five grape populations. Relevant details are provided in Results section in L168-172 and Supplementary Note 6(**Response Figure 3**).

4. For the GWAS, an obvious weakness is that the phenotyping is only from one year. In addition, based on the GWAS results, many traits did not get significant association loci. This makes the GWAS results are not reliable.

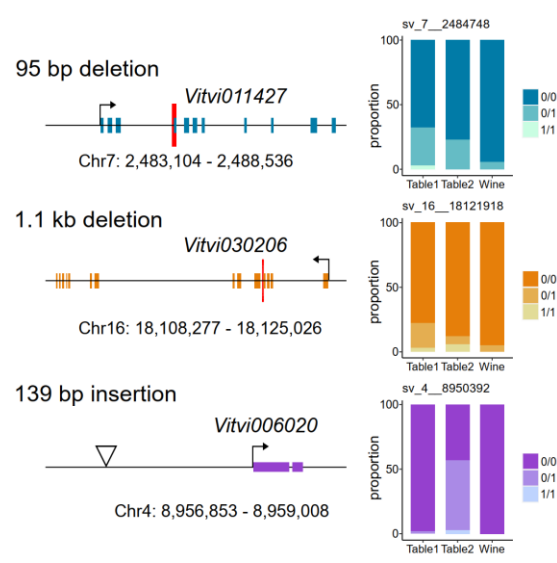
RESPONSE: Thanks for pointing out this problem. We agree with the reviewer that some traits, like fruit weight, is likely a complex quantitative trait influenced by environments and many small-effect loci, which is difficult to capture significant genetic loci. To improve the power of GWAS, we (1) added phenotypic data of 2017, and conducted GWAS analyses for each year, then combined the QTLs detected in 2016 and 2017 on 29 agronomic traits, as suggested by the reviewers. In addition, we (2) performed permutation test with 1000 repeats to increase the robustness of GWAS as followed by John, M. et. al. (2022, Bioinformatics). The combination of phenotypic data (2016 and 2017), a total of 148 QTLs were detected, please see the main text in L206-212 for details.

Reference:

John, M. et al. Efficient permutation-based genome-wide association studies for normal and skewed phenotypic distributions. *Bioinformatics* 38, ii5-ii12 (2022).

5. For the examples of the GWAS, since this work tried to emphasize the importance or powerful of pan-genome and SV in traits discovery, more examples to clearly show how SV associate with trait divergence need to be provided.

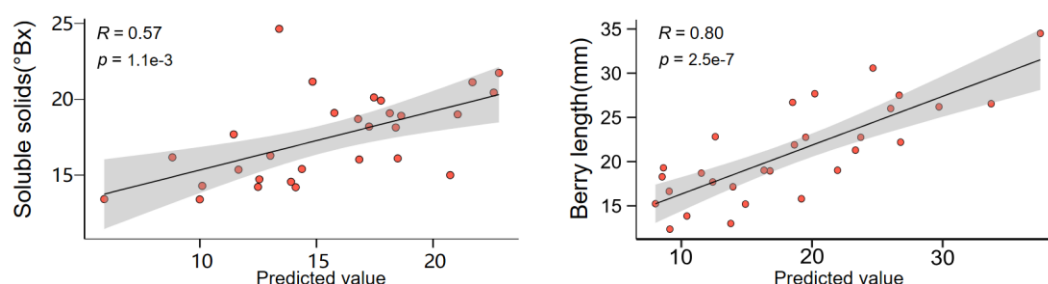
RESPONSE: Thank the reviewer for this suggestion. We added more examples in the revised manuscript in Results in L217-232 (**Revised Figure 4**).



Response Figure 7. Three significant SVs for BL1, SN6 and Suc1 GWAS loci and population frequencies.

6. For the genomic prediction, although the prediction predictive power sounds to be a higher correlation, however, still based on the same population. This need to be validated using different population.

RESPONSE: Thank you for raising this insight suggestion. We developed a cross-validation strategy to improve the genomic selection. We split the samples into training dataset and test dataset (4:1) and building a prediction model in the training dataset. The test dataset was used to indicate the predictive power. This method avoids using the same population for prediction. The PGS prediction showed an accuracy of 57.46% in the content of soluble solids (SSC) and found an accuracy of 79.53% in berry volume (BL). We have revised Results L322-329 and Methods 600-607.



Response Figure 8. The correlation analysis of prediction in the soluble solids and berry length traits.

Some errors need to be carefully checked, for example: typo of Phylogenic tree (Figure S7). There are some more.

RESPONSE: Thank you for your reminder. We have revised it and carefully check the writing throughout the revised manuscript.

Reviewer #3:

Remarks to the Author:

The paper describes the construction of a pan genome for the *Vitis vinifera* species and its use for identifying genetic determinants of simple and complex traits using a GWAS approach. The paper is interesting but suffers from some severe drawbacks.

RESPONSE: Thanks for the reviewer's constructive comments. We had revised the manuscript accordingly. The revised version has greatly improved and we hope the current version is more suitable for publication.

Firstly the paper is written quite carelessly so that it often makes it difficult to understand exactly what is meant and/or what has been done. It requires extensive rewriting after reading by a native English speaking person.

RESPONSE: Thank the reviewer for this suggestion. We have thoroughly revised the manuscript to improve the clarity of the results, logic, and conclusions. Furthermore, we have incorporated extensive suggestions on scientific writing from Dr. Brandon S. Gaut (University of California, Irvine, CA, USA) to strengthen the quality of

presentation.

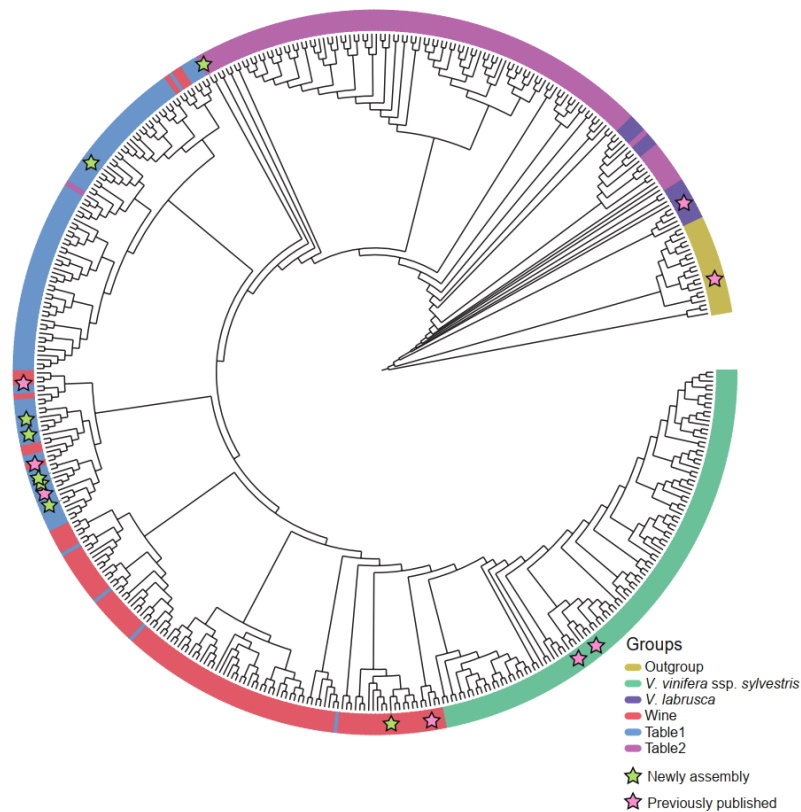
Scientifically there are a few severe issues that undermine the validity of the conclusions, the main one being the lack of any validation of the new data presented. I will list the main ones hereunder.

There is no justification for the choice of genotypes for the construction of the pan genome. Normally one would choose individuals as to maximise the breadth of diversity within the species. Here there are mostly table grapes with a very limited representation of wine grapes and many of the genotypes are quite closely related (some of them even in parent offspring relationship). The relatedness among genotypes, even though not ideal for the purpose of the paper, could have been exploited for the purpose of validating the differences observed by comparing haplotypes that are identical by descent between the pairs of genotypes. The lack of any validation of the observed differences is a major shortcoming of the paper that makes it difficult to trust many of the conclusions. The fact that the total n. of genes goes from 31148 to 37972 within *V. vinifera* is hard to believe but would become believable if properly validated.

RESPONSE: Thank the reviewer for this comment. We have given utmost attention to the validation of results and data, and also performed comprehensive further analyses.

To represent the grapevine genetic diversity, we newly assembled nine accessions based on HiFi, ONT and Hi-C technology, and combined it with eleven published assembled haplotypes, which represents most of the genetic diversity of grapevine, including six Wine grapes, six Table1 grapes, one Table2 grape and one *V. rotundifolia* grape, one *V. labrusca* grape, two *V. vinifera* ssp. *sylvestris* grape and one *V. arizonica* grape (Supplementary Table 1). We performed PCA analyses, phylogenetic tree and IBS of 466 sequenced grape accessions and highlighted the samples used for pangenome construction. We have added information of assembled samples in Supplementary Note 1.

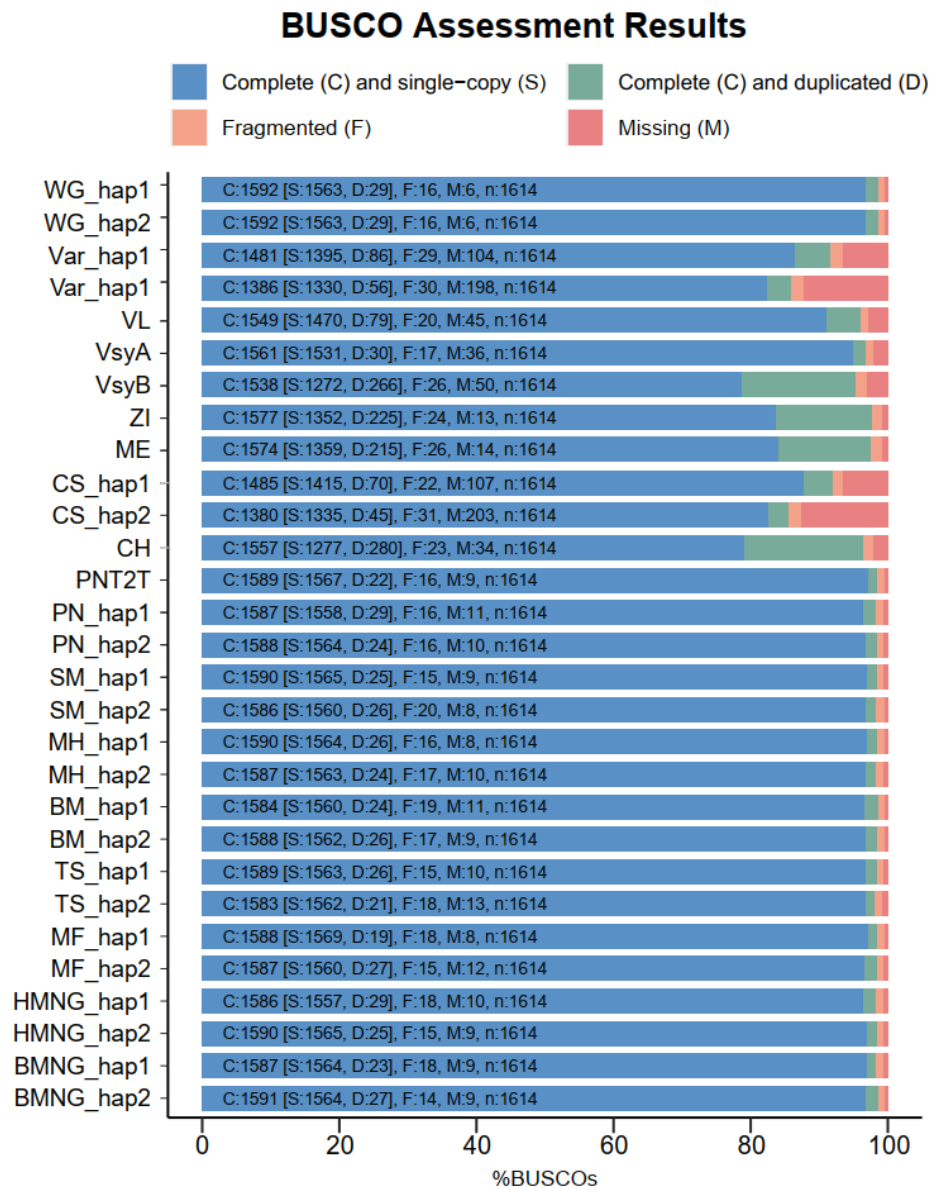
In addition, we used a uniform pipeline to improve gene annotation. As results, the number of protein coding genes ranged from 36,093 to 37,487 (L108-112). The methodology details are described in Supplementary Method 2.



Response Figure 9. The phylogenetic tree of 466 sequenced samples. The pangenome used accessions were highlighted including nine newly sequenced accessions and nine previously published accessions.

It is not clear why the authors felt the need for adding to the new assemblies that were produced using PacBio HIFI technology that lead to highly accurate and complete haplotype phased assemblies older assemblies obtained with the old PacBio technology that are not haplotype phased and that have a much poorer quality and completeness.

RESPONSE: Indeed, the newly assemblies base on HiFi were high quality and gapless, as well as haplotype-resolved, which is much better than the CLR assemblies previously generated. To improve the pangenome to represent the diversity of grapes and capture the most information of SVs in grape genomes, we collected eleven long reads-based CLR assemblies. Our validation suggested these assemblies showed contig N50 > 1 Mb. We have described in method section in L469-470.



Response Figure 10. Genomic completeness estimation of assemblies by Benchmarking universal single-copy orthologs (BUSCOs).

It is not clear why after building the different assemblies, the pan genome graph was not built by comparing assemblies with one another but rather by mapping reads of each accession against an individual reference to detect structural variants. No validation data is presented to determine the rate of false positive and false negative detection of SV using the procedure described that uses default parameters and a minimum read depth of three. The use of tools that have been ad hoc developed for pan genome reconstruction and used for example in the recent human pan genome effort, should have been tested, such as for example Pan Genome Graph Builder (pggb, [pggb](https://github.com/PanGenome/pggb) (notion.site)) that makes use of wfmask or Minigraph-Cactus (<https://www.nature.com/articles/s41587-023-01793-w>).

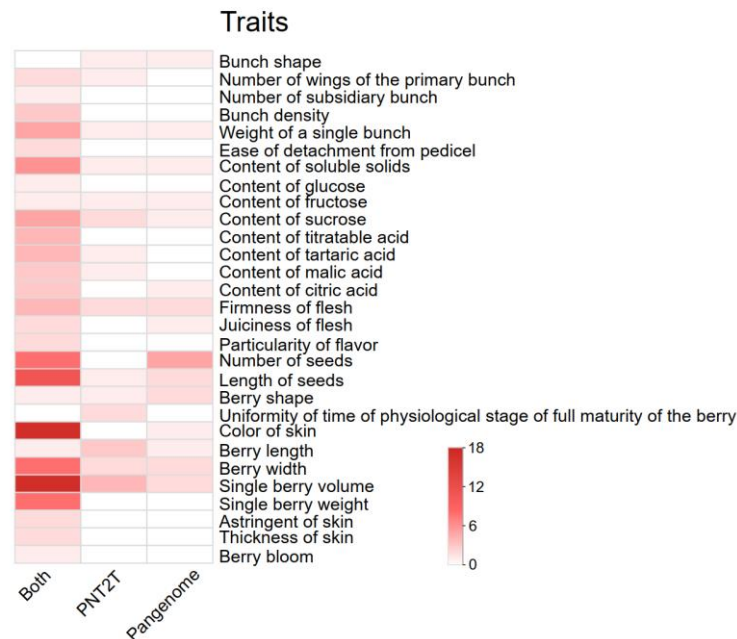
RESPONSE: We are grateful for the reviewer's comment. We had used both PGGB and MC for pangenome construction. Then, we used two complementary approaches to generate the SV map. First, we deconstructed the MC pangenome to obtain an SV map as described by Liao et al. (2023, Nature). Meanwhile, we identified SVs by aligning assemblies using SVIM-asm program (<https://github.com/eldariont/svim-asm>). To validate these pangenome SVs, we compared the SV generated by MC and SVIM-asm and obtained three validation metrics (precise, recall ratio and f1 score) using the program Truvari (<https://github.com/ACEnglish/truvari>). The high-quality pangenome SVs were filtered for the downstream analysis. We have added descriptions in the Methods L498-507, and provided detailed methodology of this process in Supplementary Method 5.

Reference:

Liao, WW., Asri, M., Ebler, J. et al. A draft human pangenome reference. Nature 617, 312–324 (2023). <https://doi.org/10.1038/s41586-023-05896-x>

The main purpose of the paper is to show that the pan genome approach is better at identifying genetic determinants of simple and complex traits using a GWAS approach than the use of a single reference genome. This should have been experimentally proven by performing two independent experiments with the same data set: one where the reference used is the pan genome graph, the other one where the reference used is a single genotype such as for example the widely used PN40024 in its new T2T version.

RESPONSE: Thank the reviewer for this constructive comment. We performed analyses to compare the SNP datasets generated using pangenome and the PN40024 (T2T) as the reference genome. We also performed GWAS independently of the SNP dataset from the PNT2T single reference genome. We found that 124 of 136 (91.2%) loci were detected by both the pangenome SNP and PNT2T reference SNP. We have added descriptions of the comparative analysis in the Results section in L226-231.



Response Figure 11. Comparison GWAS results between PNT2T based SNPs and pangenome based SNPs.

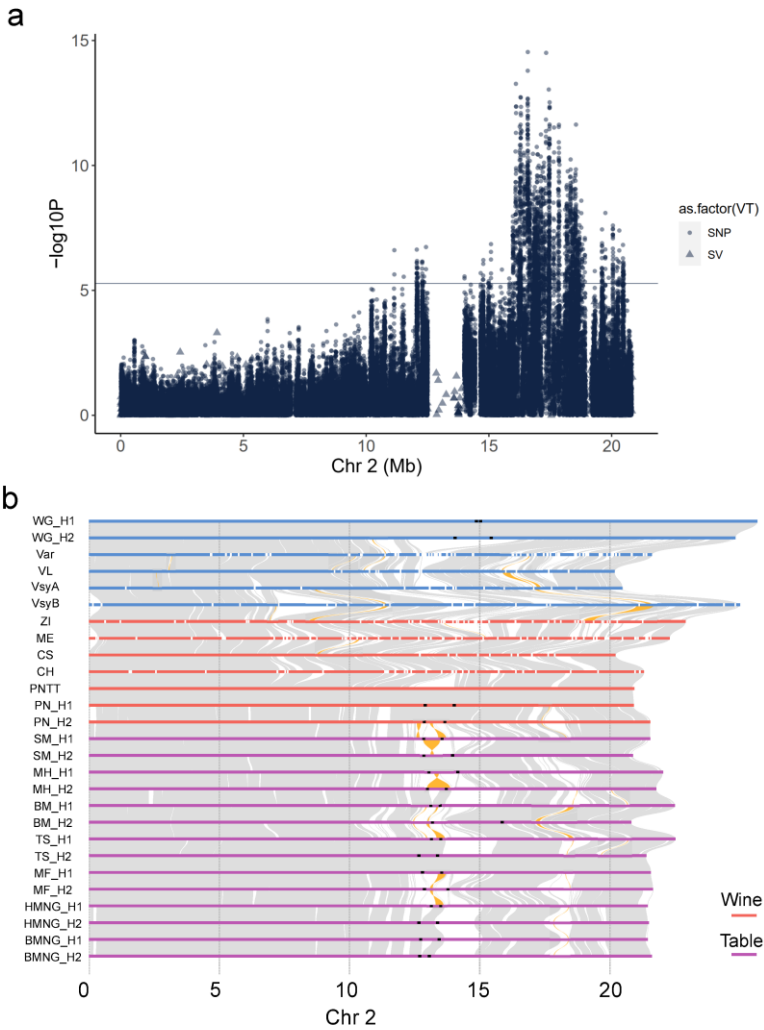
The finding of many correlations among berry size related traits and berry content traits is not surprising if berry content is measured on a concentration basis and not on a per berry basis. It is well known that larger berries at veraison lead to decreased concentration of many berry metabolites, so this issue would deserve a much greater attention.

RESPONSE: We agree with the reviewer that the berry size and berry contents should be correlated as expected. We have revised the descriptions of the berry size related traits and berry content traits in Results section in L185-190 and L192-195.

The causative mutations for fruit colour are well known and are located in chr. 2 around 14 Mpb. The involvement of the inversion is not demonstrated and the data presented here shows that there is an inversion on chr 2 in this region in only one accession (*Muscadinia rotundifolia*, see supplementary figure 3 for chromosome 2). The large region of LD around the berry color is presumably due to a very recent and intense selective sweep followed by a limited number of meiotic events that led to the production of white varieties. The involvement of the 5 Mbp inversion in the trait determination is negated by the genome assemblies presented here. No inversion is visible between white and red haplotypes. So either these assemblies are incorrect or the previous observation of the presence of the inversion is not correct).

RESPONSE: Thank the reviewer for the constructive comment. We re-examined the genomic segment of chromosome 2 that was significantly related to fruit color in pangenome GWAS analysis (**Response Figure 12**), which overlaps the well-known

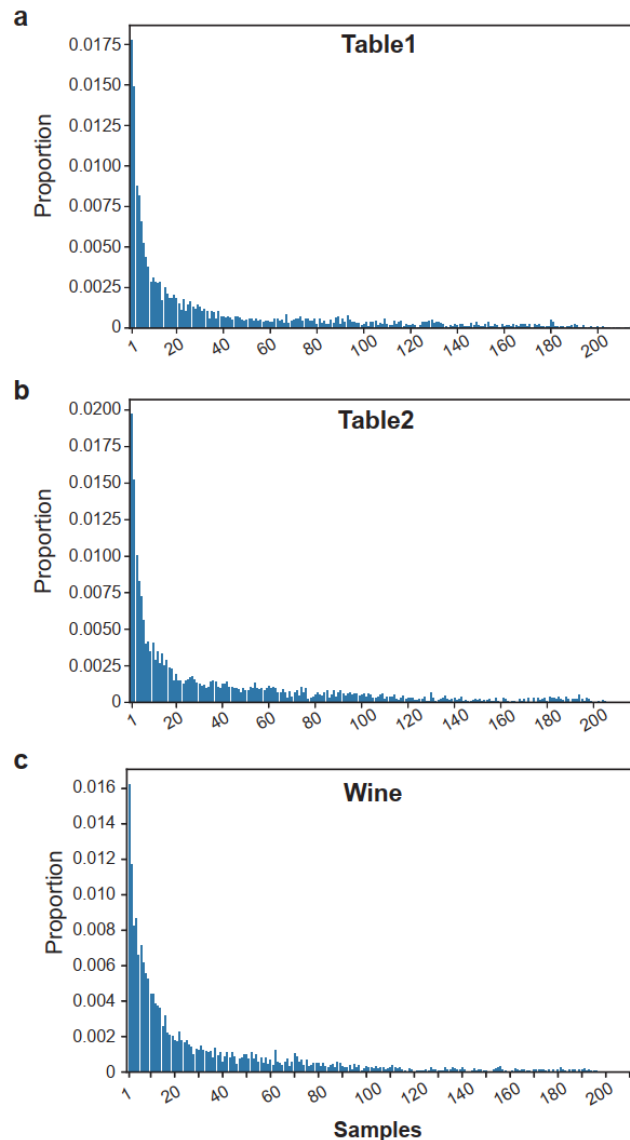
locus of chromosome 2 around 14 Mb. Our collinearity analysis revealed without 5 Mb inversion in our assemblies (**Response Figure 12**). Based on the reviewer's suggestion, we revised the description of the genetic mechanism of berry color in the L267-268.



Response Figure 12. The GWAS of Chr 2 on berry color and collinearity analysis. a, The GWAS of berry color on chromosome 2 based on pangenome SV and SNP datasets. b, The collinearity analysis of chromosome 2 based on 29 assembled haplotypes.

No biological/genetic explanation is provided for the extremely low LD observed between SNPs and SVs. A proper explanation should be provided.

RESPONSE: We thank the reviewer for pointing this out. The low LD between SNPs and SVs could due to the low population frequency under strong purifying selection of SVs (Zhou et al. 2019 Nature Plants). We had added this information in Supplementary Figs. 14 and 15.



Response Figure 13. The frequency of SV in three grape populations (Table1, Table2 and Wine).

Other minor comments are provided in the attached PDF file:

Line54. A very recent paper in Science advocating a double domestication of grapevine should be cited here also (<https://www.science.org/doi/10.1126/science.add8655>)

RESPONSE: We have checked references and cited this research in the Introduction and Discussion section.

Line75. where does this conclusion come from?

RESPONSE: Thank you for raising this issue. We have revised this sentence (L75-78).

Line91. what is meant exactly by wild grape accessions? Are these non vinifera Vitis

species? Then they should not be called wild grape accessions. And line109. See comment above, this should not be called wild grape because this could be confused with *Vitis vinifera* subsp. *sylvestris*, the wild ancestor of cultivated grapes.

RESPONSE: Thank those suggestion. The “wild grapes” is not clear and so we removed and replaced by species name (Supplementary Note 1). We have revised this sentence.

Line169-173. this sentence as it is very cryptic. there should be an explanation of what is meant here.

RESPONSE: Thank you. We have added detailed descriptions of PHATE analysis. To clear represent the genetic diversity, we calculated PCA of 466 WGS accessions and highlighted the HiFi sequencing samples. Combined PCA, ADMIXTURE and phylogeny to inferred the population structure.

Line183-185. too broad a claim for the data presented

RESPONSE: Thank you. We have revised this conclusion.

Line194. which ones are they?

RESPONSE: Thanks. The positive correlation was evaluated in 4 traits of berry size in grapes (L185-190 and 192-195).

Line234-243. This region was shown to have undergone a severe selective sweep in the domestication step (see for example ref. 7). The results should be interpreted in light of this observation.

RESPONSE: Thanks for this great suggestion. We have revised the sentences to explain the strong selective sweeps in the candidate genetic loci underlying the content of soluble solids in grapes.

Line262. Polygenic.

RESPONSE: Done.

Line303-310. The causative mutations for fruit color are well known and are located in chr. 2 around 14 Mpb. The involvement of the inversion is not demonstrated and the data presented here shows that there is an inversion on chr 2 in this region in only one accession (see supplementary figure 3 for chromosome 2).

RESPONSE: Based on the reviewer's suggestion that a recent and intense selective sweep may lead to similar transformations, we performed two genetic statistics (Genetic diversity and Tajima'D), and the phylogenetic relationship based on genomic variations. We found that these results are highly consistent with the reviewers' opinions. Therefore, we carefully revised this sentence and the conclusion.

Line325. The term breed is not used for plants or for grapevine.

RESPONSE: Revised.

Line339-343. When it comes to wine grapes the reason for the lack of breeding efforts is none of those listed here but rather the conservative attitude of the business and the complex regulatory system for quality wine appellations that encourages the continued use of traditional old varieties.

RESPONSE: Great. Actually, accelerating the innovation of grape varieties is urgent to adapt to the future planting, rapidly changed market demands, and climate changes. Although some resistance comes from the conservative attitude and supervision that encouraging traditional and old varieties, even more than 900 years, the pressure to updating varieties is more from genetic factors in the grape genome. We have revised these sentences.

Line 349-253. These assemblies have rather poor quality and should not be confused with the HiFi assemblies.

RESPONSE: We tried our best to improve this part and we have improved the variations detection in the revised manuscript. Please see L502-508.

Line357-360. I do not see exactly where this conclusion is derived from: a careful comparison with the wild grapes (*Vitis vinifera* subsp. *sylvestris*) is needed to reach such a conclusion.

RESPONSE: Thanks to the reviewer for helpful comments. We have deleted these sentences and focused on the pangenome and phased variation map in improving grape breeding. Pangenome contribute to a large of high-quality SVs and improved GWAS power for grape traits. In addition, the phased genomic variation map could be benefit for breeding superior variations with good traits combining at haplotype. We have rewritten the L341-345 in Discussion section.

Line364-369. The large region of LD around the berry color is presumably due to a very recent and intense selective sweep followed by a limited number of meiotic events that led to the production of white varieties. The involvement of the 5 Mbp inversion in the trait determination is negated by the genome assemblies presented here. No inversion is visible between white and red haplotypes.

RESPONSE: Thank you. We carefully revised this sentence and the conclusion , please see Response Figure 12 and L265.

Line394. This table is missing in the supplementary materials provided; Supplementary Table S1 contains a summary of repeated sequence annotations.

RESPONSE: We feel sorry for our carelessness. We have added Supplementary Tables 1 and 6.

Line 477-480. Why was the pan genome graph built by aligning reads to one specific assemblies and not by comparing assembled haplotypes as usually done?

RESPONSE: Thank you for raising this import issue. To represent the grapevine

diversity, we newly sequenced nine accessions and collected eleven assemblies (Supplementary Table 1). These sequences were used to construct graph based pangenome using the PGGB and MC programs. The PN40024 T2T genome was used as a reference guided the graph pangenome in MC. While the all-to-all alignments were used in PGGB to construct the graph pangenome. We have revised the L478-489.

Line 550-555. it is not explained what was the set of variants/genotypes that were used in the GWAS analysis.

RESPONSE: We sincerely appreciate the valuable comments. We independent performed GWAS based on three different datasets: single-reference based SNP dataset, pangenome based SNP dataset and pangenome based SNP+SV dataset. We compared the SNPs from single-reference for improving of pangenome at the SNP level. Then, the SV dataset was used to capture the miss heritability and investigating traits related to structural makers. We have clarified datasets used in GWAS and provide details in Supplementary Note.

Decision Letter, first revision:

13th Mar 2024

Dear Professor Zhou,

Your Article, "Grapevine pangenome facilitates trait genetics and genomic breeding" has now been seen by 2 referees. You will see from their comments below that while they find your work of interest, some important points are raised. We are interested in the possibility of publishing your study in Nature Genetics, but would like to consider your response to these concerns in the form of a revised manuscript before we make a final decision on publication.

As you will see from these comments, Reviewer 1 has identified important aspects of the presentation that need to be improved. Please address their comments as thoroughly as possible with appropriate revisions.

***Data Availability*:** As per the journal policy, all sequence data - including genome assemblies and raw reads - must be deposited to repositories that are part of the International Nucleotide Sequence Collaboration (INSDC) (<http://www.insdc.org/>). Since the NCBI data processing usually takes a long time, we suggest depositing all sequence data to NCBI as soon as possible.

We therefore invite you to revise your manuscript taking into account all reviewer and editor comments. Please highlight all changes in the manuscript text file. At this stage we will need you to upload a copy of the manuscript in MS Word .docx or similar editable format.

We are committed to providing a fair and constructive peer-review process. Do not hesitate to contact us if there are specific requests from the reviewers that you believe are technically impossible or unlikely to yield a meaningful outcome.

When revising your manuscript:

*1) Include a "Response to referees" document detailing, point-by-point, how you addressed each referee comment. If no action was taken to address a point, you must provide a compelling argument. This response will be sent back to the referees along with the revised manuscript.

*2) If you have not done so already please begin to revise your manuscript so that it conforms to our Article format instructions, available [here](#).

Refer also to any guidelines provided in this letter.

*3) Include a revised version of any required Reporting Summary:

<https://www.nature.com/documents/nr-reporting-summary.pdf>

It will be available to referees (and, potentially, statisticians) to aid in their evaluation if the manuscript goes back for peer review.

A revised checklist is essential for re-review of the paper.

Please be aware of our [guidelines on digital image standards](#).

Please use the link below to submit your revised manuscript and related files:

[redacted]

Note: This URL links to your confidential home page and associated information about manuscripts you may have submitted, or that you are reviewing for us. If you wish to forward this email to co-authors, please delete the link to your homepage.

We hope to receive your revised manuscript within four to eight weeks. If you cannot send it within this time, please let us know.

Please do not hesitate to contact me if you have any questions or would like to discuss these revisions further.

Nature Genetics is committed to improving transparency in authorship. As part of our efforts in this direction, we are now requesting that all authors identified as 'corresponding author' on published papers create and link their Open Researcher and Contributor Identifier (ORCID) with their account on the Manuscript Tracking System (MTS), prior to acceptance. ORCID helps the scientific community achieve unambiguous attribution of all scholarly contributions. You can create and link your ORCID from the home page of the MTS by clicking on 'Modify my Springer Nature account'. For more information please visit www.springernature.com/orcid.

We look forward to seeing the revised manuscript and thank you for the opportunity to review your work.

Sincerely,

Wei Li, PhD
Senior Editor
Nature Genetics
New York, NY 10004, USA
www.nature.com/ng

Reviewers' Comments:

Reviewer #1:

Remarks to the Author:

AS this is a second-round review, I will not repeat point by point all the details that I already provided before. I will use a simpler structure, here :

General remarks. The authors did a very thorough work to answer the questions of the reviewers, providing new data and analyses, and many points were correctly addressed and solved. I am satisfied with the work done for answering all the reviewers' technical and methodological questions.

The authors also answered to a more basic question, explaining why they wish to present a large-scope paper exploring all the line from pangenome to the association of variation with agronomic traits in the field. While I respect the authors' choice, I remain nevertheless with the feeling that the paper is composed of two parts of unequal quality and value for the scientific community. I will list here the points that make up this feeling :

1) The first part related to genomics is impressive in terms of quality and value for the scientific

community. I would strongly support the publishing of this part. Still, I would have liked more work spent to characterise the SVs, a) In term of annotation (are they TEs, fragments of TEs, random sequences, sparsed centromere motifs?) ; b) In which positions are they more frequent (introns, exons, intergenic regions, are they uniformly distributed or grouped in hot-spots?) ; c) In term of time of evolution (are they very recent, or old? ; can we see any active TE moving its complete form in recent times? d) Are they more abundant in the dispensable or the private genome ? and so on. I found no mention at all of the word "transposon", and "transposable elements" was only mentioned once, providing their sequence length but with no tentative to relate with SVs. In discussion, I am missing the references to some important work published on this topic, for example :
<https://bmcpplantbiol.biomedcentral.com/articles/10.1186/1471-2229-14-7>
<https://bmccgenomics.biomedcentral.com/articles/10.1186/s12864-022-08389-9>
<https://onlinelibrary.wiley.com/doi/10.1111/tpj.13274>

Summarizing this part on the pangenome, I would strongly support its publication in Nature Genetics with some additional work to characterise the SVs, and a more thorough discussion.

2) The second part related to the GWAS in my opinion has less value, as GWAS was done many times before on grape and using more complex phenotyping as compared to the one used here. For example, finding a correlation among 29 agronomic traits is nothing new (line 378) ; genomic prediction was dealt with before (but not cited here) ; adding some 12 regions captured by SVs (over 148) doesn't change much to the story ; the region determining the color was already known, and that it is determined by an SV was known as well. In this part I am more specifically missing the citation of some important previous works, for example :
<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8772106/>
<https://www.nature.com/articles/s41438-018-0089-z>
<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9070645/>
<https://pubmed.ncbi.nlm.nih.gov/35485948/>
<https://www.nature.com/articles/hortres201735>

3) The discussion in its present form, while listing some examples of SV-impacted traits, in my opinion is still weak and it is not easy to derive from it what is the main message(s) carrying the most important scientific relevance.

In summary, my feeling is that on the one hand, the authors replied correctly to the reviewers questions, and this deserves certainly to be commended. However, the two parts (genomics versus agronomics) are of unequal quality and I am a bit worried that the part with less quality will lower the overall perception of this work. Probably the choice to proceed or not with this paper in its present form is purely an editorial choice belonging to the journal.

In any case, I would suggest the authors to add more arguments into the discussion with the proper and up-to-date citation of existing public scientific papers (some suggested above).
 Minor corrections in the main paper :

Line 27 : it is not investigation that impedes breeding. It is the lack of / incomplete / curtailed [understanding, knowledge of the genetic architecture of...], etc.

Line 65: in my opinion, many SVs are newly created by recent TEs activity, and that's may be one reason why they are at low frequency.

Line 75 : this sentence misses something. I would skip "Whereas"

Line 105 : please provide a reference for "hamming error", I searched in literature and could not find a clear definition

Line 121 : as said before, be more specific than "wild grape" as this term can generate confusion. Please name the taxa explicitly

Line 123-124 : you use Shine Muscat here with the label *V. vinifera* ssp. *vinifera*, but in other tables (and to my knowledge), Shine Muscat is an interspecific hybrid

Line 126 : showed "that"

Line 126 : the authors say that "12.57% (4,645) genes were in a hemizygous state". I would like they specify which threshold they used to define that one gene is present on one haplotype (maybe a blast match of 80% or more, something like that ?), and not on the other (what threshold to decide that it does not exist there? 0%, 20% ? etc.). I did not find the explanation of the threshold used neither in the main nor in the supp data.

Line 162 : identity and not identical

Line 257: I don't understand very well the reason to compare Table1 and Table2. Table 1 and 2 have different taxonomic composition (*V. vinifera* versus *Vitis* spp hybrids) so many differences may be due to species status instead of differences in traits, even if you used structure and kinship as covariates. While taking the risk of false-positives, I am not sure this paragraph adds any solid interest (what is the basic question?), so maybe take it away.

Line 302 : the word "only" is repeated
Minor suggestions for the supplementary data

Line 97 : "The breeding periods for these varieties range from hundreds to nearly twenty years » : unclear

Line 153 : the cited article provides 4 *Vitis sylvestris*. You used two. Please specify which ones (*VsyA*, *VsyB* short names are not cited in that article).

Line 407 : identity instead of homogeneity

Reviewer #2:

Remarks to the Author:

All my questions were addressed. Thank the authors for their efforts. I do not have further questions.

Reviewer #3:

None

Author Rebuttal, first revision:

Reviewers' Comments:

Reviewer #1:

Remarks to the Author:

The authors also answered to a more basic question, explaining why they wish to present a large-scope paper exploring all the line from pangenome to the association of variation with agronomic traits in the field. While I respect the authors' choice, I remain nevertheless with the feeling that the paper is composed of two parts of unequal quality and value for the scientific community.

Response: We sincerely thank the reviewer for the respectful and generous understanding of our main aspirations of our paper. This project aims to address trait genetics and genomic breeding in a broader context through integrated exploration from pan-genomes to GWAS and GS. We believe that the integration of both components is essential for a holistic understanding of how genomic variation influences phenotypic variation. We respect that you may feel that the two parts are uneven, but we do not see them as two separate parts. To thoroughly explore relationships between genomic and phenotypic variation, it is crucial to identify SVs. Our construction of a pangenome has greatly improved the detection of SVs, which has given us unparalleled insights into their contribution to quantitative traits.

To address the reviewers' concerns, we have added additional analysis of SV in the hope of generating widespread discussion within the scientific community. At the same time, we revised the GWAS and GS sections, reduced the consensus content, and fully and appropriately cited previous work. We have revised the Results and Discussion sections to encourage new findings and facilitate in-depth discussion.

I will list here the points that make up this feeling :

- 1) The first part related to genomics is impressive in terms of quality and value for the scientific community. I would strongly support the publishing of this part. Still, I would have liked more work spent to characterise the SVs,
 - a) In term of annotation (are they TEs, fragments of TEs, random sequences, sparsed centromere motifs?);
 - b) In which positions are they more frequent (introns, exons, intergenic regions, are they uniformly distributed or grouped in hot-spots?);
 - c) In term of time of evolution (are they very recent, or old? ; can we see any active TE moving its complete form in recent times?
 - d) Are they more abundant in the dispensable or the private genome ? and so on.

I found no mention at all of the word "transposon", and "transposable elements" was only mentioned once, providing their sequence length but with no tentative to relate with SVs.

Response: We appreciate your constructive comments. In the revised manuscript, we performed additional work to characterize SVs in Grapepan v1.0, examined the overlap between SVs and transposable elements and emphasized their roles in grape evolution.

To annotate these SVs, we categorized them as containing transposable elements (TEs), TE fragments and simple repeats. We found significant overlap between SVs and

TEs in the grapevine genome. TE are major evolution force for the generation of crop genomic SVs, and these observations have been reported in other crops, such as tea, tomato, citrus and rice (Chen, 2023; Alonge, 2020; Huang,2023 Fuentes 2019). At the same time, we examined their distribution and frequency. By assessing the presence or absence of these fragments in all grape genomes, we calculated the population frequency of SVs. With these additional analyses, we elucidated the genetic signature and potential impact of SVs in the grape genome. We have revised L 156-162 in Results and L 395-400 in discussed.

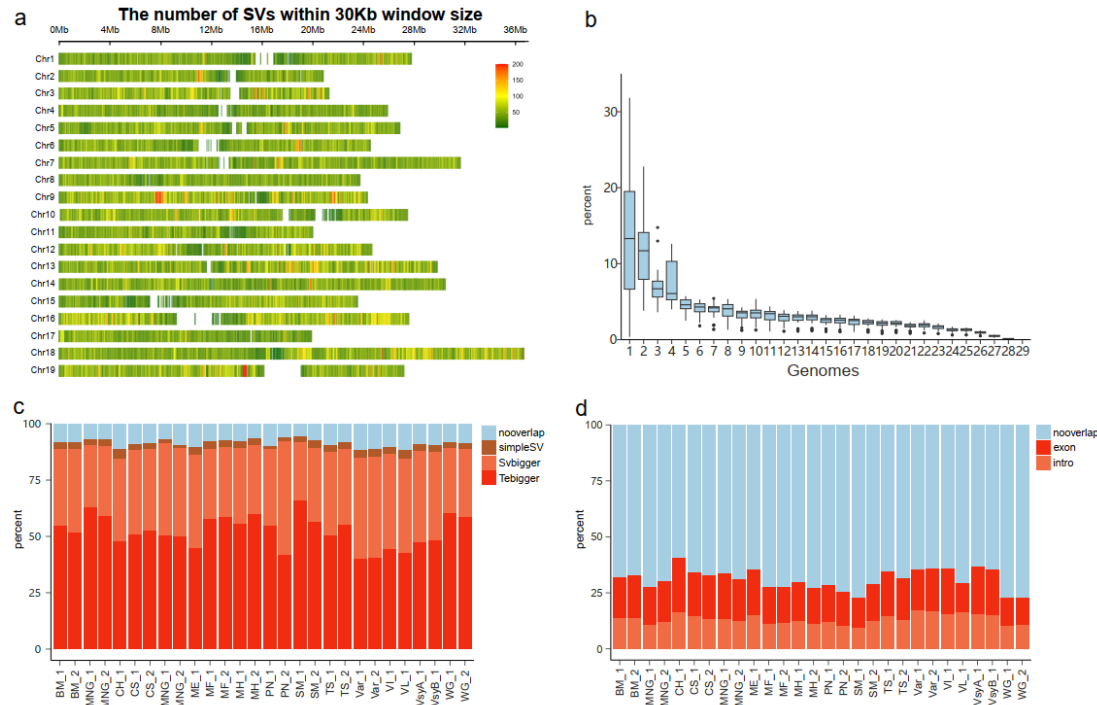
References:

Chen, S., Wang, P., Kong, W., Chai, K., Zhang, S., Yu, J., . . . Zhang, X. (2023). Gene mining and genomics-assisted breeding empowered by the pangenome of tea plant *Camellia sinensis*. *Nature Plants*, 9(12), 1986-1999. doi:10.1038/s41477-023-01565-z

Alonge, M., Wang, X., Benoit, M., Soyk, S., Pereira, L., Zhang, L., . . . Lippman, Z. B. (2020). Major Impacts of Widespread Structural Variation on Gene Expression and Crop Improvement in Tomato. *Cell*, 182(1), 145-161.e123. doi:10.1016/j.cell.2020.05.021

Huang, Y., He, J., Xu, Y., Zheng, W., Wang, S., Chen, P., . . . Xu, Q. (2023). Pangenome analysis provides insight into the evolution of the orange subfamily and a key gene for citric acid accumulation in citrus fruits. *Nature Genetics*, 55(11), 1964-1975. doi:10.1038/s41588-023-01516-6

Fuentes, R. R., Chebotarov, D., Duitama, J., Smith, S., De la Hoz, J. F., Mohiyuddin, M., . . . Alexandrov, N. (2019). Structural variants in 3000 rice genomes. *Genome Res*, 29(5), 870-880. doi:10.1101/gr.241240.118



Response Fig1. Genomic landscape of structural variations (SVs). a. The heatmap represents the distribution of SVs in the grape pangenome. The enrichment of each window (100 Kb) was highlighted with different color. b. The frequency of shared SVs

across 29 assembled genome haplotypes. Y-axis indicated the frequency of SVs detected in these assemblies. c and d. The percentage of transposable elements (TEs) (c) and genic regions (d) that spatially overlap with the SV structure in the grape pangenome. (see Supplementary Fig. 7)

In discussion, I am missing the references to some important work published on this topic, for example :

<https://bmcpplantbiol.biomedcentral.com/articles/10.1186/1471-2229-14-7>

<https://bmccgenomics.biomedcentral.com/articles/10.1186/s12864-022-08389-9>

<https://onlinelibrary.wiley.com/doi/10.1111/tpj.13274>

Summarizing this part on the pangenome, I would strongly support its publication in Nature Genetics with some additional work to characterise the SVs, and a more thorough discussion.

Response: We have revised the Discussion and incorporated previous findings to improve the quality of the discussion.

2) The second part related to the GWAS in my opinion has less value, as GWAS was done many times before on grape and using more complex phenotyping as compared to the one used here. For example, finding a correlation among 29 agronomic traits is nothing new (line 378) ; genomic prediction was dealt with before (but not cited here) ; adding some 12 regions captured by SVs (over 148) doesn't change much to the story ; the region determining the color was already known, and that it is determined by an SV was known as well.

In this part I am more specifically missing the citation of some important previous works, for example :

<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8772106/>

<https://www.nature.com/articles/s41438-018-0089-z>

<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9070645/>

<https://pubmed.ncbi.nlm.nih.gov/35485948/>

<https://www.nature.com/articles/hortres201735>

Response: Thank you for your insightful comments. We respectfully acknowledge that GWAS has indeed been conducted in grapevines previously, such as color (Guo,2019) and berry size (Zhang ,2022). We have added appropriate citations based on reviewer's comments.

Here, we would like to clarify several key aspects in which our study differs from previous studies. Our study represents the first comprehensive exploration of grapevine traits utilizing a wider range of genetic variation, especially SVs, that are often overlooked in marker-based analyses. Through analyzing the genetic correlation among traits, our aim is to evaluate the potential for selecting superior grapevine varieties possessing multiple desirable traits via genome selection. We investigated the pivotal role of SVs in elucidating the 'missing heritability' observed in quantitative genetics within the grapevine genome this is certainly a novel contribution to grapevine genetics and rare for any plant taxon. Notably, SVs significantly contribute to phenotype variation, an aspect that has been insufficiently

addressed in previous genome selection efforts. Our emphasis on fruit color, which is well known to be associated with SVs, underscores the power and accuracy of GWAS and GS based on SVs and acts as a positive control of sorts. We have found that SVs identified significant QTLs in 12 out of 29 traits. However, this effect remains conservative, primarily due to the low frequency of SVs and the limitations of existing algorithms to fully analyze the comprehensive impact of SV on the phenotype.

The reviewers' perception of insufficient value in our work may stem from unclear writing. We have re-edited the results section L 199-202 based on your feedback and streamlined the phenotype analysis. In the revised manuscript, we have adjusted the Discussion section L 365-377 to emphasize novel findings and to integrate our findings into the wider literature.

Reference:

Guo, D.-L., Zhao, H.-L., Li, Q., Zhang, G.-H., Jiang, J.-F., Liu, C.-H., & Yu, Y.-H. (2019). Genome-wide association study of berry-related traits in grape [*Vitis vinifera* L.] based on genotyping-by-sequencing markers. *Horticulture Research*, 6(1), 11. doi:10.1038/s41438-018-0089-z

Zhang, C., Cui, L., & Fang, J. (2022). Genome-wide association study of the candidate genes for grape berry shape-related traits. *BMC Plant Biol*, 22(1), 42. doi:10.1186/s12870-022-03434-x

3) The discussion in its present form, while listing some examples of SV-impacted traits, in my opinion is still weak and it is not easy to derive from it what is the main message(s) carrying the most important scientific relevance.

Response: We apologize for the previous unclear description and oversight in addressing the logical understanding of breeding traits. Our goal was a comprehensive analysis of grape breeding traits, with a special focus on traits affected by newly discovered SVs. We provided several examples to illustrate the reliability of our method and to offer an exploration of trait heritability. In the revised manuscript, we discuss the genetic correlation between SV and traits to enhance our understanding of grape breeding.

In summary, my feeling is that on the one hand, the authors replied correctly to the reviewers questions, and this deserves certainly to be commended. However, the two parts (genomics versus agronomics) are of unequal quality and I am a bit worried that the part with less quality will lower the overall perception of this work. Probably the choice to proceed or not with this paper in its present form is purely an editorial choice belonging to the journal.

Response: We greatly appreciate your suggestions; however, we believe that integrating these two aspects is central to our work. By constructing a pangenome, we identified SVs and other genomic variants that are, in turn, evaluated for their potential contributions to phenotypes. Our goal is to provide a comprehensive understanding of the genetic control of agronomic traits in grapevine, and we are committed to improving the quality of our research to meet the high standards expected by the scientific community.

In any case, I would suggest the authors to add more arguments into the discussion with the proper and up-to-date citation of existing public scientific papers (some suggested above).

Response: Thank you for your thoughtful recommendation. We have carefully cited the references you provided. Additionally, we have added other recent studies to enrich our discussion and provide a comprehensive view of the manuscript.

Minor corrections in the main paper:

Line 27 : it is not investigation that impedes breeding. It is the lack of / incomplete / curtailed [understanding, knowledge of the genetic architecture of...], etc.

Response: We fully concur with your observation. We have made revisions within the text at the appropriate location (Line 27).

Line 65: in my opinion, many SVs are newly created by recent TEs activity, and that's may be one reason why they are at low frequency.

Response: We appreciate your input. In the revised manuscript, we have elaborated on the potential role of recent TEs activity in generating SVs, as well as the characteristic of low frequency. These revisions can be found in L 70-72 of the manuscript.

Line 75: this sentence misses something. I would skip "Whereas"

Response: Based on your feedback, we have revised the sentence

Line 105: please provide a reference for "hamming error", I searched in literature and could not find a clear definition

Response: We apologize for any confusion regarding the term 'Hamming error.' In the context of genetics or sequence analysis, a Hamming error typically denotes an allele that is not present on the most supported haplotype of a contig. We have included a relevant reference to clarify this definition.

Reference:

Porubsky D, Garg S, Sanders AD, Korbel JO, Guryev V, Lansdorp PM, Marschall T. Dense and accurate whole-chromosome haplotyping of individual genomes. Nat Commun. 2017 Nov 3;8(1):1293. doi: 10.1038/s41467-017-01389-4.

Line 121: as said before, be more specific than "wild grape" as this term can generate confusion. Please name the taxa explicitly

Response: Thank you. To address this issue, we have replaced the term "wild grape" with the explicit taxonomic names throughout the manuscript.

Line 123-124: you use Shine Muscat here with the label *V. vinifera* ssp. *vinifera*, but in other tables (and to my knowledge), Shine Muscat is an interspecific hybrid

Response: Thank you for catching this inconsistency. Revised.

Line 126: showed “that”

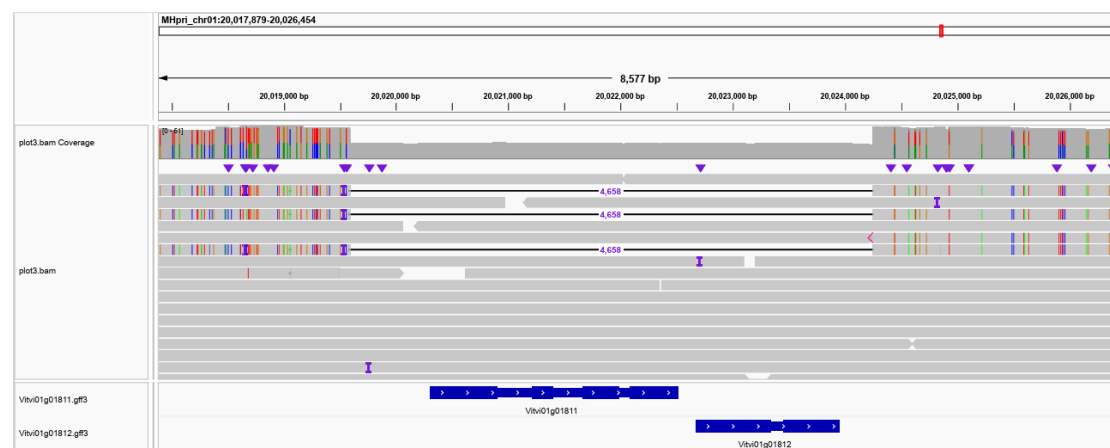
Response: Done.

Line 126: the authors say that “12.57% (4,645) genes were in a hemizygous state”. I would like they specify which threshold they used to define that one gene is present on one haplotype (maybe a blast match of 80% or more, something like that ?), and not on the other (what threshold to decide that it does not exist there? 0%, 20% ? etc.). I did not find the explanation of the threshold used neither in the main nor in the supp data.

Response: Thanks for your comment on the criteria for defining hemizygosity. In our study, we identify hemizygous genomic regions by analyzing heterozygous SVs, particularly deletions and insertions, between two homologous haplotypes. Regions where the entire gene structure (100%) is located within an insertion/deletion region were classified as hemizygous (Zhou et al., 2019, Nature Plants), as illustrated in Response Fig. 2. We have incorporated this clarification into the Supplementary Note.

Reference:

Zhou, Y., Minio, A., Massonnet, M. et al. The population genetics of structural variants in grapevine domestication. Nat. Plants 5, 965–979 (2019). <https://doi.org/10.1038/s41477-019-0507-8>



Response Fig2. The example of two hemizygous genes in the Muscat Hamburg genome.

Line 162: identity and not identical

Response: Thank you.

Line 257: I don't understand very well the reason to compare Table1 and Table2. Table 1 and 2 have different taxonomic composition (V. vinifera versus Vitis spp hybrids) so many differences may be due to species status instead of differences in traits, even if you used structure and kinship as covariates. While taking the risk of false-positives, I am not sure this paragraph adds any solid interest (what is the basic question?), so maybe take it away.

Response: We appreciate the feedback provided by the reviewers. Our study did not directly compare the data presented in Table1 and Table2 from the Genome-Wide

Association Studies (GWAS). Instead, we performed selection analysis on these two groups to investigate genome-wide differential differentiation sites. By integrating the QTLs identified through GWAS, we observed that certain key QTL sites were located within significant differentiation intervals, suggesting a potential association with selection in distinct groups. We have made revisions to L 299-301 to ensure accuracy in the description.

Line 302: the word “only” is repeated

Response: Revised.

Minor suggestions for the supplementary data

Line 97: “The breeding periods for these varieties range from hundreds to nearly twenty years » : unclear

Response: Thank you for your comments. We have removed this sentence.

Line 153: the cited article provides 4 *Vitis sylvestris*. You used two. Please specify which ones (VsyA, VsyB short names are not cited in that article).

Response: Thank you. We have clarified in the Supplementary Note that the two *Vitis sylvestris* samples used in our study: [VsyA: VvSyIDVIT3603.16 and VsyB: VvSyIDVIT3351.27].

Line 407: identity instead of homogeneity

Response: Done.

Decision Letter, second revision:

8th May 2024

Dear Dr. Zhou,

Thank you for submitting your revised manuscript "Grapevine pangenome facilitates trait genetics and genomic breeding" (NG-A63069R1). It has now been seen by the original referees and their comments are below. The reviewers find that the paper has improved in revision, and therefore we'll be happy in principle to publish it in Nature Genetics, pending minor revisions to comply with our editorial and formatting guidelines.

We are now performing detailed checks on your paper and will send you a checklist detailing our editorial and formatting requirements soon. Please do not upload the final materials and make any revisions until you receive this additional information from us.

Thank you again for your interest in Nature Genetics Please do not hesitate to contact me if you have any questions.

Sincerely,
Wei

Wei Li, PhD
Senior Editor
Nature Genetics

Reviewer #1 (Remarks to the Author):

This paper went through a couple of revisions based on suggestions from reviewers.
With this latest submission, I feel that all my previous questions were properly and patiently addressed.
Thank to the authors for their efforts.
I do not have any further questions.

Author Rebuttal, first revision:

Reviewers' Comments:

Reviewer #1:

Remarks to the Author:

This paper went through a couple of revisions based on suggestions from reviewers.
With this latest submission, I feel that all my previous questions were properly and patiently addressed.

Thank to the authors for their efforts.
I do not have any further questions.

Response: We are deeply grateful for your meticulous review and the constructive feedback you have provided throughout the revision process of our paper.

Final Decision Letter:

1st Oct 2024

Dear Dr. Zhou,

I am delighted to say that your manuscript "Grapevine pangenome facilitates trait genetics and genomic breeding" has been accepted for publication in an upcoming issue of *Nature Genetics*.

Over the next few weeks, your paper will be copyedited to ensure that it conforms to *Nature Genetics* style. Once your paper is typeset, you will receive an email with a link to choose the appropriate publishing options for your paper and our Author Services team will be in touch regarding any additional information that may be required.

After the grant of rights is completed, you will receive a link to your electronic proof via email with a request to make any corrections within 48 hours. If, when you receive your proof, you cannot meet this deadline, please inform us at rjsproduction@springernature.com immediately.

You will not receive your proofs until the publishing agreement has been received through our system.

Due to the importance of these deadlines, we ask that you please let us know now whether you will be difficult to contact over the next month. If this is the case, we ask you provide us with the contact information (email, phone and fax) of someone who will be able to check the proofs on your behalf, and who will be available to address any last-minute problems.

Your paper will be published online after we receive your corrections and will appear in print in the next available issue. You can find out your date of online publication by contacting the Nature Press Office (press@nature.com) after sending your e-proof corrections.

You may wish to make your media relations office aware of your accepted publication, in case they consider it appropriate to organize some internal or external publicity. Once your paper has been scheduled you will receive an email confirming the publication details. This is normally 3-4 working days in advance of publication. If you need additional notice of the date and time of publication, please let the production team know when you receive the proof of your article to ensure there is sufficient time to coordinate. Further information on our embargo policies can be found here: <https://www.nature.com/authors/policies/embargo.html>

Before your paper is published online, we shall be distributing a press release to news organizations worldwide, which may very well include details of your work. We are happy for your institution or funding agency to prepare its own press release, but it must mention the embargo date and *Nature Genetics*. Our Press Office may contact you closer to the time of publication, but if you or your Press Office have any enquiries in the meantime, please contact press@nature.com.

Acceptance is conditional on the data in the manuscript not being published elsewhere, or announced in the print or electronic media, until the embargo/publication date. These restrictions are not intended to deter you from presenting your data at academic meetings and conferences, but any enquiries from the media about papers not yet scheduled for publication should be referred to us.

Please note that *Nature Genetics* is a Transformative Journal (TJ). Authors may publish their research

with us through the traditional subscription access route or make their paper immediately open access through payment of an article-processing charge (APC). Authors will not be required to make a final decision about access to their article until it has been accepted. [Find out more about Transformative Journals](#)

Authors may need to take specific actions to achieve [compliance](#) with funder and institutional open access mandates. If your research is supported by a funder that requires immediate open access (e.g. according to [Plan S principles](#)) then you should select the gold OA route, and we will direct you to the compliant route where possible. For authors selecting the subscription publication route, the journal's standard licensing terms will need to be accepted, including <https://www.nature.com/nature-portfolio/editorial-policies/self-archiving-and-license-to-publish>. Those licensing terms will supersede any other terms that the author or any third party may assert apply to any version of the manuscript.

If you have any questions about our publishing options, costs, Open Access requirements, or our legal forms, please contact ASJournals@springernature.com

If you have posted a preprint on any preprint server, please ensure that the preprint details are updated with a publication reference, including the DOI and a URL to the published version of the article on the journal website.

To assist our authors in disseminating their research to the broader community, our SharedIt initiative provides you with a unique shareable link that will allow anyone (with or without a subscription) to read the published article. Recipients of the link with a subscription will also be able to download and print the PDF.

As soon as your article is published, you will receive an automated email with your shareable link.

You can now use a single sign-on for all your accounts, view the status of all your manuscript submissions and reviews, access usage statistics for your published articles and download a record of your refereeing activity for the Nature journals.

An online order form for reprints of your paper is available at <https://www.nature.com/reprints/author-reprints.html>. Please let your coauthors and your institutions' public affairs office know that they are also welcome to order reprints by this method.

If you have not already done so, we strongly recommend that you upload the step-by-step protocols used in this manuscript to protocols.io. protocols.io is an open online resource that allows researchers to share their detailed experimental know-how. All uploaded protocols are made freely available and are assigned DOIs for ease of citation. Protocols can be linked to any publications in which they are used and will be linked to from your article. You can also establish a dedicated workspace to collect all your lab Protocols. By uploading your Protocols to protocols.io, you are enabling researchers to more readily reproduce or adapt the methodology you use, as well as increasing the visibility of your protocols and papers. Upload your Protocols at <https://protocols.io>. Further information can be found at <https://www.protocols.io/help/publish-articles>.

Sincerely,
Wei

Wei Li, PhD
Senior Editor
Nature Genetics
www.nature.com/ng