

Super pangenome of Vitis empowers identification of downy mildew resistance genes for grapevine improvement

Corresponding Author: Professor Wenxiu Ye

This manuscript has been previously reviewed at another journal. This document only contains information relating to versions considered at Nature Genetics.

Version 0:

Decision Letter:

2nd May 2024

Dear Professor Ye,

Your Article, "Super pangenome of Vitis empowers identification of downy mildew resistance genes for grapevine improvement" 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 substantially 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 identified important aspects of the analysis methods and interpretation of results that need to be substantially improved. We particularly ask that you perform additional analyses to improve the robustness of the association analyses, and address all referee comments as thoroughly as possible with appropriate revisions. Please do not hesitate to get in touch if you would like to discuss these issues further.

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](https://www.nature.com/nature-research/editorial-policies/image-integrity).

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

Link 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:

In this manuscript, the authors report the construction of a pangenome of 72 global grapevines. The data analysis is rigorous, and this large-scale genome analysis provides a deep understanding of the genomic structure of grapes. The report on pangenome construction is considered a significant achievement.

In addition to the construction of the pan-genome, authors also reported about identification of DW resistance gene through eQTL analysis. The part indicated as the utilization of the constructed pangenome, as described in the title. However, there are significant doubts regarding the analysis methods and interpretation of results concerning DM resistance. One concern is the lack of consideration for population stratification despite analyzing four genetically distinct populations. Furthermore, in the results demonstrating differences in allele types and gene expression levels, no significant difference in expression levels was observed between the pathogen-infected (Pv) and the control (Mock). This suggests that the gene expression under investigation is not induced by infection with the inoculated pathogen. Therefore, the credibility of the description regarding the association between SVs, including those related to downy mildew resistance, and traits is considered very low. Consequently, additional analyses to address these doubts are necessary, or the description regarding the association with traits should be removed. In the latter case, the authors should add a different perspective on what novel biological insights they have discovered using the constructed Super-Pangenome.

Other comments

L 173 SNP-based phylogenetic tree of the 591 accessions suggested that our 71 grapevine accessions were a good representation of genetic diversity in the Vitis genus

I could not read the fact from Extended Fig.3A, because I could not see the locations of the 71 accessions in the three.

Please draw a figure to distinguish between the 71 accessions and the others. Please also explain the meaning for the four colors in the figure legend.

L262 : Muscadine, North American and East Asian grapevines were overall more DM-resistant. Although there is significant difference in East Asian and European, East Asian seems less DM-resistant than Muscadine and North American. It is better to describe like as 'Muscadine exhibited the highest resistance to DM, followed by North American and then East Asian in descending order of resistance. '

L264: which suggests that this subfamily plays a pivotal role in conferring resistance to DM. For instance, while North American demonstrates higher resistance to DM compared to East Asian, the number of NLRs is slightly greater or equal in East Asian. Therefore, it is deemed inappropriate to infer NLR involvement in DM resistance based on the results of Fig 4G and 4H, although previous study reported that DM resistance loci Rpv1 and Rpv3 are TIR-NBARC-LRR genes.

L295-L298

The authors described that SVs of the VsSEC14 gene are associated with disease resistance. However, evidence for this association could not be observed from Figures 5H, 5I, and Extended Figures 9 and 10. Firstly, there appears to be no significant differences in gene expression levels between Mock and PV, suggesting that the gene was not expressed upon infection with Pv. Secondly, while the figures showing the sporangium number of SV allele types suggest a potential association between SV allele types and DM resistance, the authors should consider about existence of a population stratification in the materials.

As shown in Fig2D, there are numerous genetic variations on the genome that characterize the four populations, Muscadine, North America, East Asia, and European populations. Therefore, when there are differences in traits between populations, there arises a problem where genomic variations unrelated to the trait can be detected as if they were associated with it. Hence, to demonstrate that SVs on VsSEC14 are associated with DM resistance, it is necessary to present analytical results that consider the issue of population stratification.

L 309: Does '535,122,987 bases' represent the total length of the SV? If so, please specify that this number is the total length of the SV.

L312: Please show the number of SVs genotyped for the 113 accessions. I think all the 16,830,980 nodes (i.e. SVs) did not genotyped for the 113 accessions, because some were excluded because of low quality.

eQTL analysis: Authors used for the eQTL analysis, and it can consider population stratification in populations. However, there is no description about the parameters, and I could not find the evidence whether the authors considered population stratification in the analysis. If the authors consider it, please add the description. If not, please analyze with consideration of population stratification.

L321 and L322

I hold the same suspicions for the descriptions in L295-L298. Firstly, there seems to be no difference in TPM expression levels between Mock and Pv in Fig 6C, especially in the 1/1 allele type where the gene expression levels in Mock appear to be higher instead. This suggests that the gene expression was not induced by Pv infection but occurs due to other factors, and it might indicate that expression is suppressed in the 0/0 allele type. Similarly, the results in 6D also cannot deny the possibility of detecting false-positive associations due to population stratification.

Same as rge description about L323-323 (Fig 6E, 6F).

L660 Please describe the original word of the abbreviation, XG and GBWT

Materials and Methods

Please add the descriptions about a method for estimation of the heterozygous rate

Figure legend.

1C. Bubble plots showing the heterozygous segments between two haplotypes for 19 chromosomes. The fig is not shown. Please confirm.

Fig.2A. Please explain what each circle represents. For example, is one circle drawn for each accession?

Fig. 5I. No purple dots shown. Is the green dots represents East Asian?

Fig 6G. Please add the description about position of SV on the genome.

Fig 6H. Relative expression level: What is the expression level of?

Reviewer #2:

Remarks to the Author:

The manuscript presents pangenome of grape vine, including 72 haplotype resolved assemblies. This is clearly a huge body of work and will become an important resource for the community. The methods used are mostly state of the art.

I have several queries:

1. How confident are the authors in 'correctness' of the haplotype resolved assemblies produced by hifiasm? I am

particularly referring to potential presence of haplotype switches. Was the possibility of haplotype switch errors investigated?

2. Results from soybean suggest combination of giraffe and vg call may not be the best for SV genotyping from short reads (<https://link.springer.com/article/10.1186/s12915-022-01255-w>, Fig S26- Fig S28). What was the genotyping fail (most genotypes for the variant missing) rate? In the methods it was stated that Beagle was used for SV genotype imputation. What was used as input for Beagle exactly? Just SVs or SVs and SNPs? In our experience with graph based genotyping (from other crops) SVs either genotype correctly or completely fail to genotype. Was Beagle helpful here? What was the proportion of missing genotypes before and after imputation? Was there estimation of expected imputation error rate? I could not see any analysis of LD between SNPs, SVs and SNPs and SVs. I think it would be good to have especially since LD was used for imputation.

3. When performing eQTL analysis with SVs it is common to also include SNPs (<https://pubmed.ncbi.nlm.nih.gov/38479835/>, <https://elifesciences.org/reviewed-preprints/86324v2>, <https://www.nature.com/articles/s41477-019-0577-7>) to show that the putative causative SV has the strongest association (variant with strongest association is not necessarily causative, but it is indicative). In this case it looks like SNPs were not included? How many of the 79 significant SV association signals were also identified in a joint SV and SNP analysis? For how many SV had the strongest association?

Reviewer #3:

None

Version 1:

Decision Letter:

Our ref: NG-A64935R

29th Aug 2024

Dear Dr. Ye,

Thank you for submitting your revised manuscript "Super pangenome of Vitis empowers identification of downy mildew resistance genes for grapevine improvement" (NG-A64935R). 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
www.nature.com/ng

Reviewer #1 (Remarks to the Author):

The authors have adequately addressed and corrected the issues raised in the original manuscript. Therefore, I recommend the acceptance of this manuscript.

Reviewer #2 (Remarks to the Author):

The authors provided comprehensive answers to my questions and performed additional analyses. I am satisfied with their responses.

Open Access This Peer Review File is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

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

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

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

Reviewers' Comments:

Reviewer #1:

Remarks to the Author:

In this manuscript, the authors report the construction of a pangenome of 72 global grapevines. The data analysis is rigorous, and this large-scale genome analysis provides a deep understanding of the genomic structure of grapes. The report on pangenome construction is considered a significant achievement.

RESPONSE: Thank you very much for your positive comments on our work!

In addition to the construction of the pan-genome, authors also reported about identification of DW resistance gene through eQTL analysis. The part indicated as the utilization of the constructed pangenome, as described in the title. However, there are significant doubts regarding the analysis methods and interpretation of results concerning DM resistance. One concern is the lack of consideration for population stratification despite analyzing four genetically distinct populations. Furthermore, in the results demonstrating differences in allele types and gene expression levels, no significant difference in expression levels was observed between the pathogen-infected (Pv) and the control (Mock). This suggests that the gene expression under investigation is not induced by infection with the inoculated pathogen. Therefore, the credibility of the description regarding the association between SVs, including those related to downy mildew resistance, and traits is considered very low. Consequently, additional analyses to address these doubts are necessary, or the description regarding the association with traits should be removed. In the latter case, the authors should add a different perspective on what novel biological insights they have discovered using the constructed Super-Pangenome.

RESPONSE: Thank you for the great suggestion. We agree with you that population stratification is important to our association analysis. In fact, we have indeed considered population stratification in our previous SV-eQTL analysis of all 113 grapevine accessions by setting the covariates of hidden confounding factors and principal components from PCA analysis, although we did not integrate a factor for the distinct populations of grapevines. Per your suggestion, we have re-conducted our SV-eQTL analysis by taking population stratification into full consideration, and introduced a covariate that considers the different populations in addition to the other existing covariates. As a result, the updated SV-eQTL analysis identified 63 significant association signals that were also captured in the

79 significant SV-eQTLs found by our previous analysis, suggesting the population stratification improves the robustness of our eQTL analysis. Most importantly, the most significant eQTL from our previous analysis that represents the *VvLHT8* gene remained the most significant associated locus in the new eQTLs, suggesting this eQTL containing the *VvLHT8* gene is indeed a *bona fide* signal associated with downy mildew resistance in grapevine. In addition, this eQTL was also identified in a SNP-eQTL analysis requested by Reviewer2. We have revised the manuscript accordingly, updated the Figure 6B and revised the eQTL analysis methods as follows:

“The expressed genes (TPM >1) and SV genotypes were used in eQTL analysis as previously described⁴³, taking the population stratification into consideration, which identified 63 significant association signals ($P < 5.90 \times 10^{-5}$) (**Fig. 6B, Supplementary Table 11, 12**).” in Results;

“The resultant vcf files of SV genotyping and SNP calling were filtered using vcftools with parameter settings “--max-missing 1 --minDP 5 --min-alleles 2 --max-alleles 2 --maf 0.05”. PCA was performed to infer population structure using R package rMVP. Genes with a mean TPM value larger than 0.1 were contained. The top 20 hidden and confounding factors in expression data were inferred using the probabilistic estimation of expression residuals (PEER) method. The first 20 factors in PEER results, the first ten principal components in PCA analysis and population stratification were used as covariates. The eQTL mapping was conducted using MatrixEQTL¹⁰⁸.” in Materials and Methods.

Regarding the expression level, it is known that changes of expression levels induced by pathogen are not necessarily essential for the function of the genes of interest given that they can function at certain expression level. Secondly, since the manuscript only presented gene expression at a single time point, 24 hours post inoculation, it is difficult to deduce the overall expression pattern of the genes in response to pathogen infection. For example, our data (**Extended Fig. 12**) by RT-qPCR revealed that the expression level of *VvLHT8* did not change significantly in the early phase but increased in the later phase of infection. We added the results as **Extended Fig. 12** and the following sentence to Results: “Furthermore, the expression level of *VvLHT8* was dynamically regulated at different time points by *P. viticola* infection (**Extended Fig. 12**).”.

Other comments

L173 SNP-based phylogenetic tree of the 591 accessions suggested that our 71 grapevine accessions were a good representation of genetic diversity in the Vitis genus

I could not read the fact from Extended Fig.3A, because I could not see the locations of the 71 accessions in the three. Please draw a figure to distinguish between the 71 accessions and the others.

Please also explain the meaning for the four colors in the figure legend.

RESPONSE: Thank you for the comment. We apologize for the labeling issue of figures and lack of explanation on the colors. We have redrawn Extended Fig. 3A with dots next to the names of the 71 accessions in the phylogenetic tree. We also updated the figure legend as follows:

“The branch colors represent different population/group of accessions. Black, outgroup (*Parthenocissus tricuspidata*); Purple, Muscadine grapes; Blue, North American accessions; Green, East Asian accessions; Red, European accessions.”.

L262: Muscadine, North American and East Asian grapevines were overall more DM-resistant. Although there is significant difference in East Asian and European, East Asian seems less DM-resistant than Muscadine and North American. It is better to describe like as ‘Muscadine exhibited the highest resistance to DM, followed by North American and then East Asian in descending order of resistance.’

RESPONSE: We appreciate the reviewer comments. According to the comments, we revised the sentences as follows in Results:

“Muscadine exhibited the highest resistance to DM, followed by North American and then East Asian in descending order of resistance (**Fig. 4G**)”.

L264: which suggests that this subfamily plays a pivotal role in conferring resistance to DM. For instance, while North American demonstrates higher resistance to DM compared to East Asian, the number of NLRs is slightly greater or equal in East Asian. Therefore, it is deemed inappropriate to infer NLR involvement in DM resistance based on the results of Fig 4G and 4H, although previous study reported that DM resistance loci Rpv1 and Rpv3 are TIR-NBARC-LRR genes.

RESPONSE: We agree that the number of NLRs cannot fully explain the resistance mechanism and other factors such as the expression level of NLRs and the presence of different types of the well-known Pattern Recognition Receptors are also critical as summarized in a recent Cell review paper by Jonathan D.G. Jones, Brian J. Staskawicz, and Jeffery L. Dangl. Thus, we tone down the interpretation of the NLR analysis by adding the following sentence to Results:

“Nevertheless, in addition to the number of NLRs, other factors such as the expression levels of each NLR and cell surface Pattern Recognition Receptors are also critical for the DM resistance³⁶.”.

L295-L298

The authors described that SVs of the VsSEC14 gene are associated with disease resistance. However, evidence for this association could not be observed from Figures 5H, 5I, and Extended Figures 9 and 10. Firstly, there appears to be no significant differences in gene expression levels between Mock and PV, suggesting that the gene was not expressed upon infection with Pv. Secondly, while the figures showing the sporangium number of SV allele types suggest a potential association between SV allele types and DM resistance, the authors should consider about existence of a population stratification in the materials. As shown in Fug2D, there are numerous genetic variations on the genome that characterize the four populations, Muscadine, North America, East Asia, and European populations. Therefore, when there are differences in traits between populations, there arises a problem where genomic variations unrelated to the trait can be detected as if they were associated with it. Hence, to demonstrate that SVs on VsSEC14 are associated with DM resistance, it is necessary to present analytical results that consider the issue of population stratification.

RESPONSE: Thank you for the comment. We believe there is misunderstanding of the results regarding VvSEC14 gene. In fact, this gene was not identified by the eQTL analysis that was not described until the next result section. Instead, the gene was found affected by SVs around the previously reported molecular marker for downy mildew resistance in grapevine (Di Gaspero, et al., 2012). We then found the marker locus indeed harbored SVs among different populations of grapevines (Figure 5H). Therefore, this is simply a confirmation of previously identified molecular marker using our detected SVs in a large collection of samples. By doing this analysis, we found the North American, Asian and European grapes had distinctive genotypes regarding the SV located on and near the VvSEC14 gene (Figure 5H), and the phenotypes (disease score and gene expression) showed clear correlation with the genotype (Figure 5I). Therefore, we believe population stratification analysis was not necessary for this part, but should be considered in the SV-eQTL analysis section later on. We hope this clarifies your confusion about this gene.

Regarding the expression difference between Mock and Pv infection, we have explained in our response to your comments above. During the infection course, expression of disease-relevant genes such as VvLHT8 is time-dependent and tends to fluctuate over time. Moreover, it is known that changes

of expression levels induced by pathogen are not necessarily essential for the function of the genes of interest given that they can function at certain expression level.

L309: Does '535,122,987 bases' represent the total length of the SV? If so, please specify that this number is the total length of the SV.

RESPONSE: Thanks for your question. No, the number is not the length of SV. The number represents the total sequence length of the graph-based pangenome, which integrated the genomic sequences from the reference genome and 67 *Euvtis* genomes. The SV information was stored in the pangenome graph containing 16,830,980 nodes and 16,946,222 edges.

L312: Please show the number of SVs genotyped for the 113 accessions. I think all the 16,830,980 nodes (i.e. SVs) did not genotyped for the 113 accessions, because some were excluded because of low quality.

RESPONSE: Thank you for the suggestion. We have added the information as Supplementary Table 11 to show the number of SVs genotyped for the 113 accessions. Indeed, not all of the SVs were genotyped from all 113 accessions.

eQTL analysis: Authors used for the eQTL analysis, and it can consider population stratification in populations. However, there is no description about the parameters, and I could not find the evidence whether the authors considered population stratification in the analysis. If the authors consider it, please add the description. If not, please analyze with consideration of population stratification.

RESPONSE: Thank you for the suggestion that improves our manuscript. As explained above, we have considered population structure in our previous SV-eQTL analysis that included a principal component analysis of all 113 grapevine accessions. However, we did not consider the four distinctive populations as a covariate. Therefore, per your suggestion, we have re-conducted our SV-eQTL analysis by taking population stratification into consideration, and introduce an additional covariate that considers the four different populations. As a result, the new SV-eQTL analysis identified 63

significant association signals that were also captured in the 79 significant SV-eQTLs found by our previous analysis. Most importantly, the most significant eQTL from our previous analysis that represent the *VvLHT8* gene remained the most significant one in the new eQTLs, suggesting this QTL and the *VvLHT8* gene is indeed a *bona fide* signal for downy mildew resistance in grapevine. We have revised the manuscript accordingly, updated the Figure 6B and revised the eQTL analysis methods as follows:

“The expressed genes (TPM >1) and SV genotypes were used in eQTL analysis as previously described⁴³, taking the population stratification into consideration, which identified 63 significant association signals ($P < 5.90 \times 10^{-5}$) (**Fig. 6B, Supplementary Table 11, 12**).” in Results;

“The resultant vcf files of SV genotyping and SNP calling were filtered using vcftools with parameter settings “--max-missing 1 --minDP 5 --min-alleles 2 --max-alleles 2 --maf 0.05”. PCA was performed to infer population structure using R package rMVP. Genes with a mean TPM value larger than 0.1 were contained. The top 20 hidden and confounding factors in expression data were inferred using the probabilistic estimation of expression residuals (PEER) method. The first 20 factors in PEER results, the first ten principal components in PCA analysis and population stratification were used as covariates. The eQTL mapping was conducted using MatrixEQTL¹⁰⁸.” in Materials and Methods.

Thank you again for the great suggestion to improve our manuscript!

L321 and L322

I hold the same suspicions for the descriptions in L295-L298. Firstly, there seems to be no difference in TPM expression levels between Mock and Pv in Fig 6C, especially in the 1/1 allele type where the gene expression levels in Mock appear to be higher instead. This suggests that the gene expression was not induced by Pv infection but occurs due to other factors, and it might indicate that expression is suppressed in the 0/0 allele type. Similarly, the results in 6D also cannot deny the possibility of detecting false-positive associations due to population stratification. Same as rge description about L323-323 (Fig 6E, 6F).

RESPONSE: Thank you for the question. As explained above, per your suggestion, we have re-conducted our SV-eQTL analysis by taking population stratification into full consideration, and introduced a covariate that considers the different populations in addition to the other existing covariates. As a result, the most significant eQTL from our previous analysis that represent the

VvLHT8 gene remained the most significant associated locus in the new eQTLs, suggesting this eQTL containing the *VvLHT8* gene is indeed a *bona fide* signal associated with downy mildew resistance in grapevine. We have revised the manuscript accordingly, updated the Figure 6B and revised the eQTL analysis methods.

Regarding the expression level, our data by RT-qPCR revealed that the expression level of *VvLHT8* did not change significantly in the early phase but increased in the later phase of infection. We added the results as **Extended Fig. 12** and the following sentence to Results:

“Furthermore, the expression level of *VvLHT8* was dynamically regulated at different time points by *P. viticola* infection (**Extended Fig. 12**).”.

L660 Please describe the original word of the abbreviation, XG and GBWT

RESPONSE: Thanks for your suggestion. Both terms were defined by the vg software developers (<https://github.com/vgteam/vg/wiki/Index-Types>). The XG is a space-efficient immutable graph. It is typically used in mappers and other applications that do not modify the graph. The GBWT (Graph Burrows–Wheeler transform) is an FM-index that stores similar threads (lightweight paths) space-efficiently. In most applications, it is used for storing (real or artificial) haplotypes.

We have revised the manuscript text accordingly, updated the description of XG and GBWT in Materials and Methods as follows:

“The preliminary graph was indexed in XG and GBWT by using *vg index* with the '-L' parameter to retain alternative allele paths (<https://github.com/vgteam/vg/wiki/Index-Types>). The XG is a space-efficient immutable graph typically used in mappers and other applications that do not modify the graph. The GBWT (Graph Burrows–Wheeler transform) is an FM-index that stores similar threads (lightweight paths) space-efficiently and is used for storing (real or artificial) haplotypes.”.

Materials and Methods

Please add the descriptions about a method for estimation of the heterozygous rate

RESPONSE: Thanks for your comment. We have described the method in Material and Methods as follows:

“For genome survey to estimate genome sizes and heterozygosity of 72 grapevine accessions, Illumina paired-end reads were used in kmer frequency analysis with Jellyfish v2.3.0⁵⁸ (k-mer=21) and

Genomescope v1.0⁵⁹.”.

Figure legend.

1C. Bubble plots showing the heterozygous segments between two haplotypes for 19 chromosomes. The fig is not shown. Please confirm.

RESPONSE: We appreciate your comments and apologize for the mistake. We have revised the figure legend.

Fig.2A. Please explain what each circle represents. For example, is one circle drawn for each accession?

RESPONSE: Thank you for the comment. We added the following sentences to Figure legend: “Each circle represents an accession.”.

Fig. 5I. No purple dots shown. Is the green dots represents East Asian?

RESPONSE: Thank you for the question. In fact, all purple dots (Muscadine accessions) were plotted but overlayed by other color dots and they have the genotype of 0/0 with sporangium number of 0. The green dots represent East Asian grapes. We revised the sentence in Figure legend as follows: “Purple dots, Muscadine accessions; Blue dots, North American accessions; Red dots, European accessions; Green dots, East Asian accessions. All purple dots are overlayed by other color dots and have the genotype of 0/0 with sporangium number of 0.”.

Fig 6G. Please add the description about position of SV on the genome.

RESPONSE: SV is found in the chromosome 3 at 2,489,994 bp position. We added the information to Figure legend as follows:

“The primers target the SV at location of Chr03:2489994.”.

Fig 6H. Relative expression level: What is the expression level of?

RESPONSE: We appreciate the reviewer comments. We improved the Y axis title as “Relative expression level of *VvLHT8*” in the revised figure and revised the figure legends accordingly.

Reviewer #2:

Remarks to the Author:

The manuscript presents pangenome of grape vine, including 72 haplotype resolved assemblies. This is clearly a huge body of work and will become an important resource for the community. The methods used are mostly state of the art.

RESPONSE: Thank you very much for your positive comments on our work!

I have several queries:

1. How confident are the authors in ‘correctness’ of the haplotype resolved assemblies produced by hifiasm? I am particularly referring to potential presence of haplotype switches. Was the possibility of haplotype switch errors investigated?

RESPONSE: Thank you for the questions regarding the haplotype-resolved assembly. The short answer for this question is that we are confident about the quality of our haplotype-resolved assemblies for grapevines produced by hifiasm using both PacBio HiFi reads and Hi-C reads given the success of the pipeline in haplotype-phased assembly of human genomes and several plant genomes, although we could not reliably evaluate the switching errors since we do not have parental lines for any of our 72 grapevine accessions.

To date, a fully-resolved haplotype assembly for diploid and polyploid organisms remains one of the most difficult tasks for genome assembly which is under active research. For human genome, this process is typically assisted by the availability of sequencing data from both parents of the child under genome assembly, also known as the trio-binning method which used short reads from parents to partition the long-reads from child, and perform haplotype-phased assemblies. However, trio-binning remains unable to resolve regions heterozygous in all three samples of the trio and will leave such regions unphased, although with trio-binning, 'switch errors' can be reliably calculated using the

paternal and maternal lines serving as accurate references. By contrast, the situation for plants is rather different where parental materials are often inaccessible for crop plants such as grapevines, particularly for the wild species. Thus, fully-phased haplotype genomes are rare in plants unless the trio sequencing data are available (in this case it is still challenging to get a completely correct haplotype phasing). Similarly, evaluation of switching errors in plants is unreliable without the trio sequencing data, which is probably why many studies of haplotype-phased plant genome assemblies do not report the statistics.

The 72 grapevine accessions in this study don't have accessible parental lines and thus we cannot perform trio-binning for haplotype-phasing. As far as we know, no genomic assembly algorithms can achieve fully phased plant genomes without relying on the parental information. The most advanced tool that approaches this challenge is an improved version of hifiasm recently developed by Cheng *et al.* (Nature Biotechnology 2022), called hifiasm (Hi-C), that can perform chromosome-scale haplotype-phasing assembly in a diploid organism without trio data using HiFi long-reads and Hi-C reads, relying on the signals from chromatin interactions provided by Hi-C data. Using this latest hifiasm (Hi-C) algorithm, we have performed the haplotype phasing of our 72 grapevine accessions. Again, due to the lack of trio information, we could not calculate the exact switching error rate from the haplotype-resolved grapevine genomes for any of the 72 grapevine accessions.

That said, we believe our haplotype-phased grapevine genomes using the hifiasm (Hi-C) algorithm are high quality given the success of the algorithm performance on the genome assembly of human and several other plant and animal species. For example, the developers of hifiasm (Hi-C) applied the algorithm to the haplotype-resolved assemblies for humans, birds, European badger, and sterlet using only HiFi and Hi-C data. Their results show that the phasing accuracy of haplotype-resolved assemblies in human using HiFi and Hi-C data is almost comparable to that with trio sequencing data with minimal phasing errors (Figure 1 in Cheng, et al., 2022). In addition, hifiasm (Hi-C) has also been used to generate several high-quality haplotype-resolved assemblies in various plants without trio data (Han, et al., 2023; Li, et al., 2024; Xia, et al., 2024).

In short, we believe that haplotype-resolved assemblies for the 72 grapevine accessions using the latest hifiasm (Hi-C) algorithm are high-quality, although a lack of trio information for grapevines prevents us from a reliable evaluation of the phasing errors. With continuous progress in bioinformatic algorithms and maybe availability of additional data types in the future, we hope that haplotype genomes can be fully resolved for these accessions.

References for this response:

Cheng, H., Jarvis, E.D., Fedrigo, O. et al. Haplotype-resolved assembly of diploid genomes without parental data. *Nat Biotechnol* 40, 1332–1335 (2022). <https://doi.org/10.1038/s41587-022-01261-x>.

Han X., Zhang Y., Zhang Q., et al. Two haplotype-resolved, gap-free genome assemblies for *Actinidia latifolia* and *Actinidia chinensis* shed light on the regulatory mechanisms of vitamin C and sucrose metabolism in kiwifruit. *Mol Plant*. 16, 452-470 (2023). <https://doi.org/10.1016/j.molp.2022.12.022>.

Li, W., Chu, C., Li, H. *et al.* Near-gapless and haplotype-resolved apple genomes provide insights into the genetic basis of rootstock-induced dwarfing. *Nat Genet* 56, 505–516 (2024). <https://doi.org/10.1038/s41588-024-01657-2>.

Xia, Z., Fan, W., Liu, D., et al. Haplotype-resolved Chromosomal-Level Genome Assembly Reveals Regulatory Variations in Mulberry Fruit Anthocyanin Content. *Hort. Res.*, uhae120 (2024). <https://doi.org/10.1093/hr/uhae120>.

2. Results from soybean suggest combination of giraffe and vg call may not be the best for SV genotyping from short reads (<https://link.springer.com/article/10.1186/s12915-022-01255-w>, Fig S26- Fig S28). What was the genotyping fail (most genotypes for the variant missing) rate?

RESPONSE: Thank you for the comment. We agree that development of bioinformatic software for graph-pangenome SV genotyping using short reads is an active research area. There are currently no universally accepted best pipeline for doing this as the graph pangenome concept is developed until recently and algorithm development remains at infant stage, although several pipelines have been reported such as *vg-giraffe* (Jouni Sirén, et al., 2021) and Paragraph (Chen, et al., 2019). That said, we found combination of *vg* and *giraffe* has been widely adopted and successfully used in multiple published studies in human pangenome projects (Wang, et al., 2022), vertebrate pangenome (Li, et al., 2023) and plant pangenome projects (Li, et al., 2023). Therefore, we performed the pangenome and genotyping analysis using *vg* and *giraffe* considering it as the popular choice for such practices.

Based on the graph pangenome we built for *Vitis* genus, we used *vg* to perform genotyping of SVs in our collection of 113 grapevine samples from NGS short-reads. In our study, we used 54,784 insertion variants and 62,652 deletion variants (117,436 in total) as known structural variants, and then genotyped these structural variants using next-generation sequencing (NGS) data. The process involves mapping the NGS reads against an indexed pangenome graph using *vg index*, followed by SV genotyping using *vg giraffe*. The genotyping results obtained indeed did not include all known structural variations in all grape samples, which means some SVs were missing. Among the 113 samples, 93.36%±1.25% of structural variation could be successfully genotyped. In addition, about 6.64%±1.39% of structural variants were identified as missing in various numbers of samples. The

failed (missing) SV genotyping was presumably caused by the insufficient coverage of NGS short reads at the particular SV loci, which is unsurprising given the limitation of NGS reads compared to the long-read sequencing reads that were used for genome assembly and SV detection process. Most importantly, we only used the successfully genotyped SVs to conduct the downstream eQTL analysis for grapevine. Therefore, our discovery of significantly associated SV-eQTLs for downy mildew resistance was not affected by the missing SV genotyping in certain grapevine samples.

References for this response:

Chen, S., Krusche, P., Dolzhenko, E. et al. Paragraph: a graph-based structural variant genotyper for short-read sequence data. *Genome Biol* 20, 291 (2019). <https://doi.org/10.1186/s13059-019-1909-7>.

Li, Ran, et al. A sheep pangenome reveals the spectrum of structural variations and their effects on tail phenotypes. *Genome Res* 33.3 (2023): 463-477. <https://doi.org/10.1101/gr.277372.122>.

Li, N., He, Q., Wang, J. et al. Super-pangenome analyses highlight genomic diversity and structural variation across wild and cultivated tomato species. *Nat Genet* 55, 852–860 (2023). <https://doi.org/10.1038/s41588-023-01340-y>.

Jouni Sirén et al. Pangenomics enables genotyping of known structural variants in 5202 diverse genomes. *Science* 374,abg8871(2021). <https://doi.org/10.1126/science.abg887>.

Wang, T., Antonacci-Fulton, L., Howe, K. et al. The Human Pangenome Project: a global resource to map genomic diversity. *Nature* 604, 437–446 (2022). <https://doi.org/10.1038/s41586-022-04601-8>.

In the methods it was stated that Beagle was used for SV genotype imputation. What was used as input for Beagle exactly? Just SVs or SVs and SNPs? In our experience with graph based genotyping (from other crops) SVs either genotype correctly or completely fail to genotype. Was Beagle helpful here? What was the proportion of missing genotypes before and after imputation? Was there estimation of expected imputation error rate? I could not see any analysis of LD between SNPs, SVs and SNPs and SVs. I think it would be good to have especially since LD was used for imputation.

RESPONSE: Thank you for the comments and questions. In our previous submission, the genotyping of SVs was conducted using *vg call* and the imputation was performed by Beagle which only took SVs as input. However, we did not consider the linkage disequilibrium (LD) among variants before we did the imputation. As suggested by you, we calculated the R^2 value to estimate the LDs for SNPs, SVs, as well as between SNPs and SVs, and found overall low LDs for the genetic variants in grapevine

(Figure R1). Under such circumstances, we believe performing imputation using SNPs or SVs or both will not work well and therefore we revised our analysis accordingly by not doing imputation to the SV or SNP genotyping results. In fact, we noted that the low LDs for grapevine have been reported in previous studies of grapevines (Liang, et al., *Nature Communications* 2019), which is consistent with our analysis of grapevine population that includes various wild and cultivated *Vitis* species. Therefore, we decided not to perform imputation for our grape samples and used the genotyped SVs directly for downstream analysis. We found the results from association studies with or without imputation are overall consistent, suggesting that doing imputation or not has little effects on our SV-eQTL findings. We have revised the sentences in the Materials and Methods as follows:

“Then we mapped resequencing data (50x coverage Illumina paired-end reads) of 113 grape accessions (Supplementary Table 4) using *vg giraffe* pipeline against the pangenome graph indexed by *vg index*. Low-quality alignments with a mapping quality and base quality both below five were excluded. Finally, SV genotyping was conducted using *vg call* with default parameters.

eQTL analysis

The resultant vcf files of SV genotyping and SNP calling were filtered using *vcftools* with parameter settings “--max-missing 1 --minDP 5 --min-alleles 2 --max-alleles 2 --maf 0.05”.

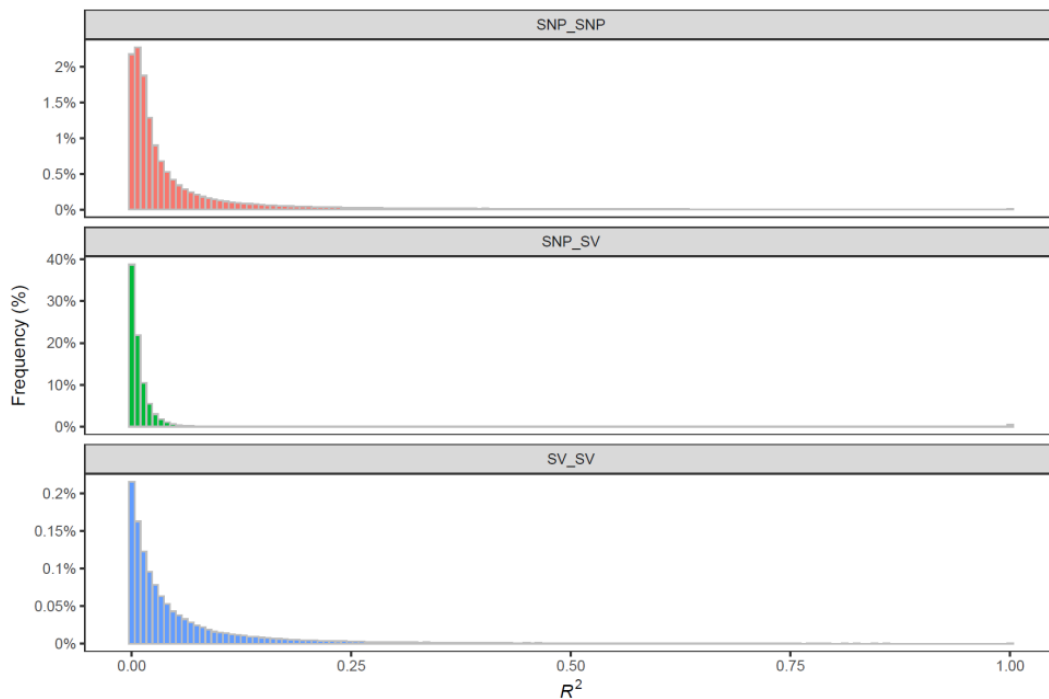


Figure R1. The distribution of R^2 for the LD analysis between SNPs (top), SNPs and SVs (middle), and SVs (bottom).

References for this response:

Liang, Z., Duan, S., Sheng, J. et al. Whole-genome resequencing of 472 *Vitis* accessions for grapevine diversity and demographic history analyses. *Nat Commun* 10, 1190 (2019). <https://doi.org/10.1038/s41467-019-09135-8>

3. When performing eQTL analysis with SVs it is common to also include SNPs (<https://pubmed.ncbi.nlm.nih.gov/38479835/>, <https://elifesciences.org/reviewed-preprints/86324v2>, <https://www.nature.com/articles/s41477-019-0577-7>) to show that the putative causative SV has the strongest association (variant with strongest association is not necessarily causative, but it is indicative). In this case it looks like SNPs were not included? How many of the 79 significant SV association signals were also identified in a joint SV and SNP analysis? For how many SV had the strongest association?

RESPONSE: Thank you for the suggestion. In fact, we did not conduct SNP-based GWAS analysis in the manuscript since our main interest is in structural variants that are functionally more disruptive to gene functions and crop traits. SNP-based GWAS analysis for grapevines has been extensively performed in several previous studies by other research groups. However, SV-based GWAS and eQTL analysis have been lacking or at least not performed on a *Vitis* genus super-pangenome scale. Therefore, for this study, we were more interested in the effects of structural variants on key traits of grapevines such as disease resistance. Per your suggestion, in addition to the SV-eQTL analysis, we also added analysis of SNP-eQTL in the revised manuscript. Using VHP-T2T hap1 as the reference genome, the SNPs of 113 grapevine samples was detected by read-mapping and SNP calling using the well-established "BWA + samtools +picard+ bcftools" pipeline, and the SNP dataset was filtered according to the minimum allele frequency of 0.05 and the deletion rate of 0. For SVs, we conducted SV genotypes based on the constructed *Vitis* pangenome graph using the NGS reads of 113 samples. We then performed eQTL analysis using the genotyped SNPs and SVs and gene expression levels of 113 samples during downy mildew infection. With a significance threshold of 1.45×10^{-7} calculated by Bonferroni correction (FDR), we identified a total of 1,809 significantly associated SNP-eQTLs (Extended Figure 11). As for SVs, we identified a total of 63 SV-eQTLs under the FDR threshold of 5.9×10^{-5} calculated by Bonferroni correction. We then examined the eQTLs that are shared between SNPs and SV association, and considered those as jointly detected by both SNP and SV if the distance between SV and SNP is smaller than 2 kb or if the genes associated with the SNP/SV loci are the same. As a result, 17 eQTLs were jointly detected by SNPs and SVs, which included the most significantly

associated SV-eQTL that corresponds to the gene *VvLHT8*.

We added the SNP-eQTL results as Extended Figure 11A and Supplementary Table 13, and the following sentences to Results:

“In addition, SNP-eQTL analysis was also performed, which identified a total of 1,809 significantly associated signals ($P < 1.45 \times 10^{-7}$) (Extended Fig. 11A, Supplementary Table 13). Among the signals, 17 signals were identified in both SV-eQTL and SNP-eQTL analysis (Supplementary Table 12, 13). Given the more disruptive effects of SVs on gene function and crop traits, we focused on the 63 significant SV-eQTLs for downstream analysis.”.

We also revised the sentences in Materials and Methods as follows:

“The resultant vcf files of SV genotyping and SNP calling were filtered using vcftools with parameter settings "--max-missing 1 --minDP 5 --min-alleles 2 --max-alleles 2 --maf 0.05". PCA was performed to infer population structure using R package rMVP. Genes with a mean TPM value larger than 0.1 were contained. The top 20 hidden and confounding factors in expression data were inferred using the probabilistic estimation of expression residuals (PEER) method. The first 20 factors in PEER results, the first ten principal components in PCA analysis and population stratification were used as covariates. The eQTL mapping was conducted using MatrixEQTL¹⁰⁸. The resultant P-value were adjusted using the FDR method. The significance thresholds of SV ($P < 5.90 \times 10^{-5}$) and SNP ($P < 1.45 \times 10^{-7}$) were calculated using the Genetic type 1 Error Calculator¹⁰⁹.”.