nature portfolio

Peer Review File

The impacts of reproductive systems on grapevine genome and breeding

Corresponding Author: Professor Yongfeng Zhou

This file contains all reviewer reports in order by version, followed by all author rebuttals in order by version.

Version 0:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
This manuscript reassembles the haplotype-resolved genome of a world's renowned wine grape species and analyzes an important and interesting question, with statistical methods and tools that are widely accepted. However, I have some major concerns that need to be clarified, particularly about the sampling strategy and the genetic background of these samples used in this study.
1. I would like to know how these clone populations were selected, such as the 18 PN samples. The accumulation of variations through different reproductive modes, such as crossing, cloning, and selfing, is different. For cloning reproduction, do these samples have accurate records of the time they were separated from the parent plant? The longer the time since separation, the more likely mutations have accumulated. Alternatively, what is the somatic mutation rate in grapes, and can this mutation rate be estimated using these clone samples, or vice versa? If these questions cannot be answered, then there is missing information, and it would be difficult to assess the representativeness of the 18 clone samples. This information should be detailed in the Materials and Methods section.

2. For the selfing population, it would be highly informative to compare the differences between the samples chosen from different selfing generations, even if only a few generations out of all 9 generations are available. Whether we can completely rule out the possibility that the heterozygous sites in PN40024 are randomly retained.

3. In each GO enrichment analysis Figure, please state the number of genes included in the analyzed gene set. Due to the limitations of the statistical method used, I believe that if the number of genes is less than 20, it would result in a significant bias.

4. Fig. 2A selects some key genes, and the method used to determine the function of these genes should be detailed in the Materials and Methods section. Providing the phylogenetic relationship of these genes with the source genes described in the literature in the supplementary material is necessary to help readers better understand the functional annotation of these genes.

5. To distinguish somatic variations from germline ones, the authors identified variations unique to PN population. I don't fully understand the underlying assumptions of this strategy. Why wouldn't somatic mutations found in the clone population also be present in the wild population? Are 10 EU + 10 ME representative enough? If I include more wild populations, will the results change?

6. Line 283, the authors claim that "Similar patterns were also found on structure variation and deleterious SNPs in each group. (Fig. 4C-D) ". What exactly is this pattern? Is “the heterozygosity of PN, SG, and HE is higher than wild grapes”? Is it statistically significant in Fig. 4C-D to claim that?

7. The “protein phosphorylation pathway” was under selection in PN genome (Fig. 2B) and also keep heterozygous in PN40024 and HE but homozygous in wild grapes (Fig. 5B). I am curious if there could be any connection, either biologically or computationally.

8. The data is presented twice in the same sentence (lines 281 and 283).

9. What do the different colors represent in Fig. S11?

10. Line 39. The “cossing” should be “crossing”.

Reviewer #2

(Remarks to the Author)
This work of Wang et al provides new data leading to a chromosome-scale assembly of an important grapevine cultivar and to an inventory of somatic variations in its clonally propagated plant material. In the version in which it is currently presented, the manuscript is missing a clearly defined biological question.

Title, abstract, and a large part of the Introduction section anticipate that this manuscript is going to focus on the different modes of propagation in grapevine (vegetative, selfing, outcrossing) and their effect on genome properties and variation in the cultivated germplasm. However, a large part of the results and discussion sections deal with the genome assembly of a single genotype (Pinot Noir) and its comparison with the genome assembly of one descendant of Pinot Noir (PN40024, an experimental genotype forcibly obtained from reiterated selfing starting from an offspring of Pinot Noir). As for the parts of the manuscript that deal more specifically with the topics anticipated in the introduction, the authors move from observations in a few cultivated genotypes and draw generalizations for the crop germplasm.

The manuscript has elements of potential interest for the grapevine scientific community that require, however, both a deeper and a wider investigation to yield robust and conclusive evidence, but also several aspects of weakness that should be more carefully considered (see below for specific points).

Specific points:

Line 68: transformation from dioecy (obligate outcrossing) to monoecy (potential self-pollination and cleistogamy).

Lines 69-70: history of outcrossing and clonal propagation of highly heterozygous genotypes.

Lines 113-120 are inappropriately placed in the Results Section.

Line 128: it should be made more explicit at first citation and in figure legends that the acronym PN_T2T refers to the T2T genome assembly of PN40024, an individual that was obtained by repeated rounds of selfing, starting from Helfensteiner.

Lines 146-155: the 4-Mbp inversion on chromosome 19 should be experimentally validated using a PCR-based assay. This assay could also be used for assessing the origin and frequency in the grapevine germplasm of the haplotype carrying the inverted region.

In the paragraph “Introgression and selection shaped the Pinot Noir genome”, the authors explained that they “used resequencing data from 38 samples (Table S2), including 18 Pinot Noir (PN) clones, 10 EU wild grapes, and 10 ME wild grapes, to explore the introgression and selection signals in PN”. It is unclear how they conducted this analysis to identify introgressed regions (I assume from the EU wild grape population) into domesticated lineages deriving from ME wild grape population (then leading to Pinot Noir), selective sweeps and candidate genes under selection, using a group of Pinot Noir clonal replicates with software that normally applies to comparisons among populations. As the resulting candidate genes (deriving from adaptive introgression ? from selection of ME ancestral diversity ?) are then described in detail and discussed extensively in Lines 162-204, their identification should be supported by stronger evidence or accompanied by a more detailed description of the analytical procedure.

Line 206: what is Pinot Noir (PN) population? The group of Pinot Noir bud sports, the offspring Helfensteiner and PN40024, as described in Lines 555-556 of the Methods section?

Paragraph “Somatic mutations in Pinot Noir clonal population” and Lines 553-565

It makes little sense using the reference genome assembly of a different genotype (derived from selfing of Helfensteiner) to realign short-reads of Pinot Noir bud sports and call somatic variants, when a haplotype-resolved T2T assembly of Pinot Noir has been generated in the present paper. A preprint that has recently appeared in bioRxiv shows the power of the use of a haplotype-resolved T2T assembly of the very same diploid genotype in the identification of somatic variants among Malbec bud sports (https://doi.org/10.1101/2023.11.30.569420). Regardless of the reference used, experimental validation of a set of candidate somatic variants needs to be provided in order to estimate false discovery rate.

Line 254: wording would suggest that the authors used 4 independent lines obtained from selfing of Pinot Noir. Is it correct ? Or are they clonal replicates of PN40024 ? PN40024 is a single seedling obtained from several rounds of selfing starting from selfing of Helfensteiner.

Paragraph “Effect of various mating systems on grape genomes”.

The authors used “groups” of cultivated varieties that were formed by one single genotype each, and three “groups” of these consisted of a parent-offspring trio. The forth “group” consisted of one single experimental highly inbred line (PN40024),

showing severe inbreeding depression, that has no parallel in the grapevine cultivated germplasm. While π is largely independent from the number of sampled chromosomes and among-groups comparisons would make sense even using one diploid individual per group, data and inferences reported in Lines 279-294 refer to individual genomes and should be used with caution for inferring generalization on the species and on crop germplasm, as those discussed in Lines 377-434.

Paragraph “Conserved heterozygous regions in cultivated grapes”.
Lines 296-300. These sentences need rephrasing.

Lines 303-305. Why intriguingly ? Heterozygous regions in PN40024 must be heterozygous in Helfensteiner by definition, as PN40024 derived from selfing of Helfensteiner.
Lines 333-334. Again. Why SVs and hSNPs should differ in heterozygous regions between Helfensteiner and PN40024, if both PN40024 homologs are identical by descent with respect to Helfensteiner ?

Paragraph “Close leakage of deleterious variants in repulsion phases”

I understand that this paragraph deals with the possible reasons (i.e. linked heterozygous deleterious variants in repulsion on the two homologs), other than chance alone, explaining why residual heterozygosity was found in specific chromosome blocks of PN40024. Figs. S13-S14 do not seem to suggest that residual heterozygous regions in PN40024 contain more deleterious variants than other heterozygous regions in Helfensteinter, which later became homozygous in the descendant upon selfing. In supporting their claim, the authors also should provide evidence that what is shown in Fig. S15 is not simply the result of incorrect variant phasing across heterozygous regions, because hap1 and hap2 are both identical by descent between Helfensteiner and PN40024 and do not appear as such in the graphical representation of Fig. S15.

Lines 385-387: This is a generalization that is inferred from observations in only three individuals.

Methodology for the assembly of the Pinot Noir genome in the Methods Section

Lines 481-489: in the generation of the HiFi-based assembly, Hi-C reads were repeatedly used at the stage of contig generation (using Hifiasm) and scaffolding (using Juicer) as well as for other intermediate steps (also in combination with RagTag-based contig ordering) that are not clearly described.

Lines 495-500: the procedure used for integrating and merging the HiFi-based partial assembly and the ONT-based partial assembly into a haplotype-resolved T2T final assembly needs a more clear and detailed explanation. Adding Supplementary Tables that report for each chromosome contig-level and scaffold-level metrics obtained for the HiFi-based assembly and for the ONT-based assembly would help follow procedure and the results described in the main text.

Line 538: structural annotation should read structural variation

References to the additional homologs illustrated in Fig. 3D should be added.

Reviewer #3

(Remarks to the Author)
Wang et al's manuscript, "Various mating systems shape grapevine genome evolution" uses genomic resources to investigate evolutionary processes in grapes. Overall, the authors seem to aim to "To investigate the genomic effects of mating systems during clonal, crossing, and selfing" (L351; first line of Discussion, and this phrase is consistent with the message of the Abstract). It is noteworthy that previous studies have addressed these same questions; e.g., the effect of clonal growth on somatic mutations and the effect of self-fertilization on genetic diversity, so the ideas presented in the manuscript are not very novel, although the data are new and useful. (Frustratingly, the authors fail to point out that a large body of work has addressed the effect of mating system and clonality on genetic diversity and plant fitness, which makes the current manuscript appear more novel than it really is. With respect to clonal growth, see this review from 2015; much other work examines the effect of selfing - https://www.pnas.org/doi/abs/10.1073/pnas.1501712112) The authors do present data that suggest an increase in somatic mutations curing clonal growth (which is consistent with previous work see link provided above), and these data are interesting. The authors also present results that suggest that some genome regions remain exceptionally heterozygous and suggest that this heterozygosity is maintained by deleterious alleles in repulsion phase. This latter result might be exciting, but I maintain healthy skepticism as it would require very strong selection (e.g., akin to recessive lethal effects) to maintain the patterns of genetic diversity documented here (assuming I understood the paper). Overall, the authors present a new genome with some analysis, and provide some results that are potentially exciting. However, my sense is that the ideas tested here (and some of the main results) are not as novel as the manuscript leads the reader to believe.

GOALS (L93-110):
1) Look at Pinot Noir genome... no real goal stated other than "look at intricacies";
2) Look at processes to generate Pniot Noir
3) Examine structural variants and deleterious variants and their changes
4) Why do some regions remain heterozygous in PN40024 -

OUTPUT 1 (L121 - 155) Genome of (heterozygous) Pinot Noir - Does this meet their goal 1? They highlight an inversion.

OUTPUT 2 (L157 - 203) - Introgression analysis of Pinot Noir Genome - IS THIS GOAL 2?
- This section provides evidence for selection on a few genes, but says little (if anything?) about the role for introgression in the evolution of the Pinot Noir genome, despite the title of this section (L156) indicating that introgressions is a process of interest. It is unclear to me whether the authors even examined for evidence of introgression - if they did, how much evidence for introgression did they detect (and in which direction between regions)?.

SOMATIC MUTATIONS IN PINOT NOIR CLONAL POPULATION (L204)
- The authors state, "To distinguish somatic variations from germline ones, we identified variations unique to PN". Is is unclear to me how this approach can reliably detect somatic mutations because it assumes that the mutations that have been identified in the small number of samples of European and Middle Eastern samples (20 genomes, total) will include *all* possible mutations that could have been present in the ancestor of the Pinot Noir clonal population. i.e., with only 20 EU and ME samples the authors have sampled only 40 alleles from the wild populations (2 alleles for each of the 20 samples, with 10 from EU and 10 from ME). Therefore, the authors are unlikely to identify variants at relatively low frequency in wild populations that can be mistaken as somatic mutations. Notably, the authors use patterns Fig 3B to support arguments that 18 mutations arose somatically. However, it is unclear to me how the authors distinguish between whether these mutations arose somatically *after* clonal propagation began vs. were simply unique to Pinot Noir and present in the genotype that gave rise to the clonal genotype *before* clonal propagation began. The authors may be correct in their logic, but they need to clarify their presentation to eliminate this alternative explanation to their data. For example, it would be useful to include a figure that illustrated the evolutionary process that the authors imagine that yields the pattern observed in their data.

L251: EFFECT OF MATING SYSTEMS ON GRAPE GENOMES - GOAL 3(?)
L265 - the authors state, "Due to nine generations of selfing, PN40024 was not counted as parent-offspring relationship with the other three cultivars." What evidence do the authors have to conclude that it was selfing, per se, that caused the analysis to fail to indicate that PN40024 was not counted as an offspring in the analysis? Selfing, per se, should only reduce genetic diversity by increasing homozygosity - therefore, it is unclear to me how the "Kinship" analysis would be so affected by increased homozygosity. The authors need to either further justify their assertion or consider either (1) that PN40024 does not derive from Helfensteiner (although this is unlikely) or (2) there are issues with the Kinship analysis, more generally, which should make the reader question *any* conclusions from this analysis.

L275: When comparing genetic diversity of inbred vs. outcrossed varieties, the authors conclude "the selfing PN40024 exhibited the lowest values for both diversity and heterozygosity when compared to all other groups". This is not a surprise, as the impact of selfing on genetic diversity is understood well in theoretical terms and is well documented empirically. Therefore, while this result adds to current data on the effect of self-fertilization on genetic diversity, it does not expand our understanding.

L280: The authors also find that the genetic diversity of colonal propagated lines exceeds that of outbred populations; "the heterozygosity of PN, SG, and HE is higher than wild grapes, ME, and EU". They authors conclude that clonal propogation increases genomic diversity and this this occurs via (in part, at least) deleterious mutations. These data are very useful, and are consistent with previous findings (see a review in 2015 here: https://www.pnas.org/doi/epdf/10.1073/pnas.1501712112) that also suggest that clonal reproduction increases opportunity for somatic mutation. The authors need to highlight that previous work also points to the accumulation of deleterious, somatic mutations during clonal growth; if the authors wish to claim that their results are somehow novel, they need to clarify which aspects are indeed novel.

CONSERVED HETEROZYGOUS REGIOUS IN CULTIVATED GRAPES (GOAL 4?)
L300: The authors "identified 379 alleles that remain heterozygous in all grape samples (wild and cultivated)." The authors further remark, "Remarkably, we identified a total of 59,944 alleles that retained their heterozygous status within PN40024. Intriguingly, these heterozygous genes exhibit consistency between PN40024 and HE, and they also remain in heterozygous states in at least one PN and one SG sample." In reading this, my first reaction was that this result can arise due to errors in genome assembly; the authors examine read coverage to address this possibility (L318 - 327 - this is good) and remove several cases where this explanation likely is true. The authors conclude that heterozygosity is maintained due to deleterious mutations in repulsion-phase (also termed pseudo-overdominance; see Charlesworth, Deborah, and John H. Willis. "The genetics of inbreeding depression." Nature reviews genetics 10.11 (2009): 783-796.) [[Please note that what I say next relies on my understanding of the author's phrase, "alleles that remain heterozygous in all grape samples", as this implies to me that all individuals that were genotyped were heterozygous at the focal loci - if I mis-understood the author's meaning, then my following comments need to be adjusted.]] If the authors' explanation is true, that multiple loci remain heterozygous in **all** genotypes individuals, then this implies *very* strong natural selection in order to maintain heterozygosity in all sampled individuals. Personally, I am surprised by (and therefore have healthy skepticism for) this finding and suggestion; I do find it interesting, but encourage authors to emphasize the requirement for very strong selection to maintain this pattern of genetic variation, and I encourage the authors to point out that this finding does require further investigation to confirm the suggestions made in the current manuscript. I can illustrate my skepticism mathematically. Take, for example, a single locus that remains highly heterozygous due to very strong selection: in this case, self-fertilization (which is how PN40024 reproduces) will lead to homozygous offspring at this locus for 50% of its offspring and many of them will experience a large fitness cost due to selection at this single locus. Now, the authors have implied that multiple loci remain heterozugous due to strong selection - if this is correct, then the fitness cost I just described for a single locus will multiple across other genome regions. Overall, we expect that most offspring that derive from self-pollination of PN40024 to have low fitness if the authors' explanation for maintaining heterozygosity is correct (and if I've understood their manuscript correctly); is this seen empirically? If not, then my sense is that the authors need to either (i) clarify their presentation (because I may have mis-understood aspects I highlight above), or (ii) amend or clarify their arguments / re-evaluate their

findings. Are genome assembly or read-mapping errors possible reasons for the observed results?

Version 1:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
I have read the revised MS, which has been improved a lot based on the comments and suggestions of reviewers including mine.

1. I suggest not using exaggerated expressions, like I propose the following modifications： change “We found very interestingly that the mating systems, namely, crossing, clonal, and selfing have dramatic effects on grape genomic variation and breeding based on comparative genomics and population genetics.” to “We found that the mating systems, namely, crossing, clonal, and selfing have significant effects on grape genomic variation and breeding based on comparative genomics and population genetics.”

Reviewer #2

(Remarks to the Author)
The revised version R1 is mostly text editing, with two exceptions: the requested wet-lab validation of a large inversion in Pinot Noir and a new analysis that focused on the seemingly forced maintenance of the heterozygous state at the flower sex locus upon selfing of a F/H2 individual, which could theoretically become H2/H2 homozygous at any selfing generation, yet maintaining an hermaphrodite phenotype for the next round of selfing, but it never did so until the S9 generation.

In my previous remarks, I pointed out that a deeper and a wider investigation to yield robust and conclusive evidence on the effects of mating systems on genetic variation was required to support the authors’ claims. From my point of view, there is not additional data in the revised version that makes the overall manuscript robust enough for being accepted for publication in a high impact journal such as Nature Communications. Most of the remarks were not addressed convincingly or exhaustively.

While some inaccuracy in the terms used in the original version of the manuscript for referring to populations, reproduction systems, mutation types seemed to be at that stage simply a matter of semantics, it seems now from their revised text and from their replies to the reviewers’ remarks that it actually reflects some misconception.

An illustrative yet not exhaustive list includes: vegetative propagation is not a mating system, a group of phenotypically divergent clonal lines is not a population, heterozygosity and chimerism of somatic mutations are not treated separately, chimerism is entirely ignored in the analysis.

Some misconception has not only theoretical implications but also important consequences on the adopted methods of analysis.
As an example, the authors state in their response to previous remarks and therefore assume in their analyses “that somatic variants, they are genetic mutations that occur in the body cells instead of the germ cells, which cannot be passed on to the next generation through sexual reproduction”, which is not true. Novel somatic variation can be sexually transmitted if the mutation has occurred in somatic cells belonging to the cell layer of the shoot apical meristem that is going to generate the germ line.
The authors also state and therefore assume that “to detect the somatic variations, we considered these mutations to be specific to the PN population and not shared by all PN individuals. The variants shared by all PN individuals are counted as germ-line mutations as these 18 individuals were collected worldwide (Supplementary Table 3), and not possible the clones of the same PN sample.” The condition mentioned by the authors to be verified in the group of the 18 stocks that were derived from the historical vegetative propagation of the original seedling of Pinot Noir is not sufficient for distinguishing somatic mutations and germline mutations. Part of the somatic mutations might also be private to Pinot Noir, yet shared by all analyzed stocks, if they occurred before clonal lineages with distinctive phenotypic variation were intentionally selected by viticulturists and nurseries and if the model of clonal lineage evolution was tree-shaped instead of star-shaped.

Below is a detailed list of critical points that were not completely fixed or have newly emerged with the R1 changes.

Line 32 and elsewhere: clonal propagation is not a mating system. Mating systems always involve crossing and if this occurs between unrelated individuals it is defined as outcrossing, between related individuals is defined as inbreeding and with itself it is defined as selfing. Clonal propagation is a mean to propagate the same genotype and does not involve mating between a male and female gamete, so it cannot be considered as a type of mating system.

Lines 33-35: There is no statistics in the results to support the claim that “The mating systems have more significant effects than traditionally focused factors such as genetic drift and natural/artificial selection.”

Line 37: “cloning increases genomic heterozygosity”. In terms of relative numbers, somatic mutations lead to a minimal increase of heterozygosity cultivar-wise, as shown by this manuscript and by previous research. Furtherly considering that a

substantial part of this novel heterozygosity is expected to be chimerical within each individual, its population-wise contribution through subsequent sexual reproduction is even less impactful than cultivar-wise. Overall, this sentence is misleading for novel point mutations and it does not consider the opposite effect of somatic deletions. It has been shown exactly in Pinot Noir that vegetative propagation leads to local loss of heterozygosity in the case of nonchimeric deletions (i.e. white-skinned bud sport) and local reduction of heterozygosity in the case of chimeric somatic deletions (i.e. gray-skinned bud sport).

Lines 99-100: “Second, we embarked on an exploration of the processes that underly clonal diversity among Pinot Noir clones, including the identification of somatic variants and genomic regions that may contribute to phenotypic traits.” Actually, this paper only reports the identification of putative somatic SNPs and the list of affected genes. Association with phenotypes is not reported experimetally.

Lines 109-114: This knowledge was well-established prior to the present work.

Lines 115-126: In the revised version and in the response to reviewers’ comments letter, the authors explain that the new Pinot Noir assembly in the latest version of the manuscript is built by omitting the step of integrating the HiFi-based assembly and the ONT-based assembly, compared to the original version of the manuscript. In spite of this, the statistics of the assembly, in particular chromosome pseudomolecule sizes have not changed between versions. Overall, this adds new uncertainty and concern about how the assembly procedure was performed and how carefully it is described, rather than easing previous concerns.

Lines 139-140: I acknowledge that the authors have experimentally validated the inversion with a wet-lab assay, as requested.

Lines 141-147: How gene families and the boundaries between gene families were defined ? Do the numbers refer to the present or absence of entire gene families or individual gene family members ? The closing comment is inappropriate: what the authors observed might be due to pre-existing presence/absence variation between the Pinot Noir and Schiava Grossa haplotypes that were donated to Helfensteiner and then to PN40024.

Lines 148-191: My previous remark was not address either in the revised text or in the response letter. I do understand how SweeD and ABBA-BABA work, but how can this software do the job for which they were designed if fed with two populations that might be approximated to natural populations and another population that is not a population but a group of clonal replicates with no germline ancestral variation ? The group of 18 Pinot Noir clones contains only novel somatic variation and collectively captures only one or two ancestral haplotypes over each genomic region. Both software deal with allele frequencies within and between populations. In particular for the statistically supported detection of selective sweeps, it is totally unclear to me how this was applied to the specific case of this manuscript in which allele frequency in Pinot Noir that may only be 0, 0.5, 1 for germline SNPs, while somatic SNPs in Pinot Noir have by definition allele frequency = 0 in the ancestral populations.
See my previous comment on the original submission in detail. In the paragraph “Introgression and selection shaped the Pinot Noir genome”, the authors explained that they “used resequencing data from 38 samples (Table S2), including 18 Pinot Noir (PN) clones, 10 EU wild grapes, and 10 ME wild grapes, to explore the introgression and selection signals in PN”. It is unclear how they conducted this analysis to identify introgressed regions, selective sweeps and candidate genes under selection, using a group of Pinot Noir clonal replicates with software that normally applies to comparisons among populations.

For supporting the authors’ claim that most of Pinot Noir private SNPs are somatic mutations (approximately 50,000 out of 70,291), the plot in Fig.3B should report on the x-axis the read count ratio supporting the ancestral and the derived allele. The direction of the mutation can be defined using the ME and EU populations the define the ancestral state. In the present form, this distribution is only showing graphically a subsampling bias or suggesting that many of the putative somatic SNPs may be false variant sites.

Lines 211-212: As most of the putative somatic SNPs were predicted to be private to individual clones, this may be either indicative of a star-shaped model of clonal differentiation or, more likely, suggestive of a high incidence of false variant sites. In the absence of an experimental wet-lab validation of at least a random subset of candidate somatic SNPs to assess the false discovery rate, this part of the manuscript reporting on somatic SNP identification is weakly supported.

Lines 215-216: The fact that putative somatic SNPs are more frequently associated with intergenic regions is another indication that many of them might be false variant sites associated with partial read misalignments in repetitive regions.

Lines 199-201: Here and in general across this section, it seems that the issue of the chimerical state of somatic mutations has not even been considered in the analyses, which may provide alternative explanations to the data compared to those given by authors.

Lines 222-244: only germline structural variation in the cultivar Pinot Noir is reported here. If the authors detected structural variation among clones, it should be reported in this paragraph. If not, they should state that they did not detect structural variation among clones.

Lines 255-260: This part has already been shown and better explained by Velt et al 2023 10.1093/g3journal/jkad067

Line 263: The He symbol in population genetics is used for expected heterozygosity under Hardy Weinberg equilibrium while the heterozigosity measured here is rather observed heterozygosity. So it would be better to list it as either Ho or just H.

Figure 4B: PN is erroneously labelled by HE in two instances.

Lines 266-272: These interpretations are questionable and not supported by the data presented in this manuscript. If one considers the number of somatic SNPs identified e.g. among the Pinot Noir clones, it is clear that in comparison to the total number of germinal SNPs that number cannot justify such a large difference in He. This difference could instead be attributed to the mating system because cultivated varieties are the result of crosses that often combine different genetic materials while wild populations may suffer from a certain level of inbreeding. It is commonly observed and generally accepted that heterozygosity is higher in grape cultivars than in the present-day individuals of the wild progenitor, but this is commonly explained by the fact that heterozygosity is exceptionally low in wild population due to small population effective size and geographical isolation leading to forced inbreeding rather than heterozygosity is exceptionally high in the cultivars due to other reasons.

Line 276: It is rather obvious that the dSNPs are seen less frequently in heterozygous condition in PN40024 where >90% of the genome is in homozygous state. So the comparison should not be with highly heterozygous accessions such as PN or SG or HE but rather it should be evaluated in comparison with the frequency of heterozygosity of non deleterious or neutral SNPs in PN40024

Line 277: This should not be indicated as recessive because there is no information on the effect of the deleterious alleles in terms of recessive or dominant, so it should be indicated as homozygous.

Lines 279-280: this could be simply explained by the fact that selfing generations of Helfensteiner leading to PN40024 and PN40024 itself are experimental genotypes that were not exposed to natural selection and were, on the contrary, nurtured with extreme care for their survival and reproduction. This is an extreme artificial situation that has no parallel in grapevine evolution and viticulture. Artificial selection during the generation of PN40024 has only targeted hermaphroditism in order to proceed with the next round of selfing.

Lines 289-296: Putative somatic mutations in Schiava Grossa, Helfensteiner and PN20024 are first mentioned here. If the procedure of identification was the same as in Pinot Noir, the numbers are not comparable among cultivars as they depend on the number of individuals that are compared within each cultivar.

Line 295: truncated sentence

Lines 329-339: This section is not very relevant because there is no quantitative analysis of the local recombination rates in this manuscript.

Lines 374-386. This part is an interesting addition with respect to the original R1 version. In the end and overall, this is the only really exciting part of the manuscript. However, as it occurred to other parts, it only scratches the surface of the biological issue and would require extended data.

Lines 376-377: If I am not wrong, H1 derived from a recombination between F and M haplotypes. H2 has then derived from a recombination between H1 and F, shortening the M derived region in the H haplotype.

Lines 435-438: Any conclusion and comment on the effects of domestication and improvement on diversity by comparing populations of the wild ancestor against single individual genomes of cultivated varieties makes no sense.

Lines 509-534: The entire paragraph “Lessons and implications for grapevine genomic breeding” contains general considerations and personal interpretations that are inappropriate in this context.

Supplementary Table 4. Taxonomic assignment in the column “species” is not accurate.

The issue with former Figure 5 panel F and Supplementary Figure S15 has been simply removed, and not solved.

Reviewer #4

(Remarks to the Author)
This is a genome paper for Pinot noir. The primary data contribution is a phased outbred whole genome assembly of a Pinot Noir plant. The authors present many bioinformatic analyzes of the two chromosome sets (PN1 and PN2) in relation to the “reference genome” – a chromosome level assembly on one highly inbred line of this species (PN40024_T2T) – plus lots of resequencing data from other grapes. The latter is (apparently) Illumina data published previously, although this is not entirely clear from the paper. I was not a referee on the first submission and was asked to comment on mating system aspects and particularly on the claims (lines 103-104 and elsewhere) about residual heterozygosity in the 9th generation inbred line.

It is plausible for a locus to be heterozygous after 9 generations of selfing. With no selection, the probability is only 0.002, but with preferential preservation of heterozygotes during line formation, the probability increases. For example, if heterozygotes

are twice as likely to survive as homozygotes, the probability increases to 0.075, which is still not a large number, but if numerous loci across the genome are experiencing this selection, a few might remain after 9 generations. For the loci on chromosomes 2 and 7 (Fig 5), the authors have the correct data signal – many heterozygous SNPs within a localized 'block' combined with the observation of normal read depths when Illumina short reads are mapped back onto the genome. The authors absolutely need to be much more explicit on one key point: How many heterozygous blocks do they believe are real residual heterozygosity? On lines 484-485, they say "lots of heterozygous regions are real and can't be ignore" – what number corresponds to "lots"? For how many regions can the authors make the picture they have for chroms 2 and 7 (Fig 5a)? The heterotic selection argument (that heterozygote advantage is responsible for residual heterozygosity) becomes less and less plausible as you add more and more loci.

Of course, the most compelling relevant data would be to directly measure the fitness effects of alternative genotypes. If you generate and genotype a few hundred selfed progeny from the line, differential fitness should be estimable because fitness differences must be large (else the line would not be heterozygous). The GERP stuff (line 364) is fine, but very indirect. Apparently, information about genotype to phenotype/fitness is available for the SDR locus on Chrom 2 (lines 374 onward). There are no citations here, but it seems that one homozygote (f/f) cannot self (has fitness = 0 during inbred line formation) because it makes only female plants. It is not clear what is going on with the alternative homozygote (H2/H2) – are these plants hermaphrodites or not? These are the essential details to make the residual heterozygosity interesting.

For the broader mating system arguments, I have no doubt that the frequencies of outcrossing/selfing/clonal propagation were important for grape genomes. However, the text requires revision because it does not make sense to say "The mating systems have more significant effects than traditionally focused factors such as genetic drift and natural/artificial selection" (lines 33-35, similar on lines 418-419 and elsewhere). Mating system does not change allele frequencies on its own. It alters the way that genetic drift, gene flow, and natural/artificial selection act on variation. There are numerous good reviews to guide these arguments – see papers by Stephen Wright and references therein.

Other comments:

Line 148: I could not make sense of the arguments on introgression. For most biologists, introgression is an explanation for shared variation between populations or species—often, it is posited as an alternative to shared ancestral variation. Here, we need a much better sense of what the different lineages actually are, and what alleles moved between lineages, and how we can be certain that these variants were not present in the common ancestor of the (now) distinct lineages. Are EU and PN (lines 179-180) sister lineages currently isolated from each other?

Line 202 onward: The general audience will need an explanation of how the authors are using the term "somatic mutation." Plants do not have a segregated germline, so all mutations start out as somatic. Are the somatic mutations identified here those that have never been transmitted through gametes? If so, the main paper could use a sentence or two to explain how that is established from sequence data.

Lines 273-274: "Then we compared the genetic burden in each group by identifying deleterious SNPs (dSNPs)…" The very important term "deleterious SNPs" comes out of nowhere without any explanation or citation. What does this mean? Does it mean that the SNP has two alleles, one that is good and one that is always bad? If so, how is this determined (the paper has no fitness assays that I can see). There are bioinformatic classifications of alternative alleles (e.g. synonymous versus non-synonymous) but these do not identify the fitness effects of particular loci. They make groups where the distribution of selection coefficients is likely be different by some (always unknown) amount between groups. This can be useful (e.g. if the fraction of non-syn SNPs is much higher in selfing species than outcrossing), but the reader will need much more explanation to make any sense of the downstream analyses on 'dSNPs'.

Lines 321-323: "… after nine generations of selfing to generate PN40024, most of the genome was homozygous, except for a 59,944 heterozygous SNPs."
What fraction of 60k SNPs are within the heterozygous blocks that the authors believe to be residual heterozygosity? After only 9 generations, real het blocks should be quite large (e.g. Chrom 2 and 7 in Fig 5). After excluding all the SNPs that are false hets due to mapping, how many loci are there in this inbred line? Perhaps the authors just need to explain the results in Supplementary Table 10.

Version 2:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
I have no more comments.

Reviewer #2

(Remarks to the Author)
In the newest version, the addition of an extra focus in one individual on the effect of experimentally-induced multiple rounds of selfing on the flower sex locus provides substantially novel and relevant biological information in comparison to the rest of

the manuscript. I appreciated this extended part.

The other methodological concerns that I expressed in my previous rounds of reviewing remain unaddressed in this new version. In particular, those regarding the issues related to the identification of reliable sequence variation among clonal lines, to the identification of introgressed regions in Pinot Noir, and to the consequences brought about by selfing.

As to the entire parts of the manuscript regarding the impacts of vegetative propagation on the genetic composition of crop germplasm, all of this is elaborated upon a critical set of variant sites. As explained by the authors in one point of their rebuttal letter, it is commonplace to acknowledge that from 1% to 5% SNP calls in reference-based analyses using short-read alignments are likely to be false variant sites. If they called 3,915,920 SNPs in their analysis of different grapevine genotypes (Line 176), it follows that they consider a reasonable number between 39,000 and 195,000 variant sites to be possible errors. If they had analysed 18 clonal lines that are genetically identical with no sequence variation, under the same experimental and analytical conditions as above, in the absence of sequence variation in the biological samples they would have therefore expected to call a similar absolute number of false variant sites due to the experimental procedure alone, which are to be considered possible errors. Among 18 Pinot Noir clones that might have biological sequence variation, the authors called a number of variant sites comprised between 20,328 and 70,291, which largely overlap with the absolute numbers of false variant sites that may result from the experimental procedure alone, as above. This situation is nearly identical for SVs. I do not believe that all of these variant sites among Pinot Noir clonal lineages are false, but in the absence of an experimental validation and in the absence of an independent assessment of false discovery rate in this experiment, I cannot take this dataset on trust either. While up to 200 K false SNPs out of 4 M true SNPs have no significant impacts on the results of population genetics analyses, the absolute number of called SNPs among clonal lineages compared to the expected absolute number of possible errors might lead to what the authors called in the point I mentioned above “horrible incorrect studies” or conclusions.

As to the entire parts of the manuscript regarding the impacts of selfing on the genetic composition of crop germplasm, I do not have doubts on the data extracted from the genome composition of the PN40024 line. However, part of the observations of the characteristics of the genome in a single individual might be due to the effect of chance alone and drawing generalizations on the consequences of selfing at population-wise level is unsound.

Reviewer #4

(Remarks to the Author)
The authors have improved their paper in response to my previous comments. My review focused on the issue of residual heterozygosity within a putatively highly inbred line. They have resolved the real heterozygosity to five loci that cover about 4% of the genome. Given strong selection against homozygotes in certain portions of the genome, this number is reasonable even after 9 generations of selfing. The authors did not collect additional data to explore the nature of heterotic selection, but that may be beyond the scope of this work. The overall narrative still needs a great deal of work, but I am satisfied with their treatment of the heterozygosity issue.

Version 3:

Reviewer comments:

Reviewer #2

(Remarks to the Author)
Regarding the specific question I am asked, whether my remaining concerns have been addressed in the last round of revision of the manuscript, the answer is no.

Version 4:

Reviewer comments:

Reviewer #2

(Remarks to the Author)
The manuscript has been improved to an adequate degree. The results and discussion sections include sufficient explanations about the limitations of this study.

Authors response to reviewers' comments

Reviewer #1 (Remarks to the Author):

This manuscript reassembles the haplotype-resolved genome of a world's renowned wine grape species and analyzes an important and interesting question, with statistical methods and tools that are widely accepted. However, I have some major concerns that need to be clarified, particularly about the sampling strategy and the genetic background of these samples used in this study.

Response: We are grateful for your positive views and constructive comments for our manuscript. Following your advice, we have clarified the information of the sampled used in this study. We also addressed all minor concerns, please read the manuscript with major revisions highlighted for details. We hope the revised version will be more suitable for publication on NC.

1. I would like to know how these clone populations were selected, such as the 18 PN samples. The accumulation of variations through different reproductive modes, such as crossing, cloning, and selfing, is different. For cloning reproduction, do these samples have accurate records of the time they were separated from the parent plant? The longer the time since separation, the more likely mutations have accumulated. Alternatively, what is the somatic mutation rate in grapes, and can this mutation rate be estimated using these clone samples, or vice versa? If these questions cannot be answered, then there is missing information, and it would be difficult to assess the representativeness of the 18 clone samples. This information should be detailed in the Materials and Methods section.

Response: Thank you for concerning this issue. We have revised the manuscript based on your advice. As shown in Supplementary Table 3, we included information about the separated time from the parent plant and resource for the clone population samples in the revised version. The geographically widespread distribution enhances the representativeness of these 18 clone samples, albeit at the expense of missing records regarding the separated time from the parent plant in some samples.

2. For the selfing population, it would be highly informative to compare the differences between the samples chosen from different selfing generations, even if only a few generations out of all 9 generations are available. Whether we can completely rule out the possibility that the heterozygous sites in PN40024 are randomly retained.

Response: Thanks for your suggestion. PN40024 was generated before 2007 and was first used for the first grape genome assembly (Jaillon et al. 2007 Nature). This 9-generation selfing material has been spread to and cultivated in many labs for further studies. However, there are no other materials available out of 9 generations. In this study, we focused on the large heterozygous fragments, and there is only a 50% possibility of keeping heterozygous in each generation. We think there is an extremely low probability for them to retain heterozygous randomly after 9 selfing generations because there is a 50% chance to turn to homozygous during selfing except undergo

strong selection to maintain this pattern. In addition, almost all the retained heterozygous sites (90.1%) in PN40024 are included in five heterozygous blocks instead of randomly distributed on the grape genome (5.2% of the PN40024 genome). Thus, it is unlikely these heterozygous sites are randomly retained. Thanks for the suggestion from the reviewer, we also added some discussion about this in the manuscript Lines 482-508.

Reference:

Jaillon, O. et al. The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. Nature 449, 463–467; 10.1038/nature06148 (2007).

3. In each GO enrichment analysis Figure, please state the number of genes included in the analyzed gene set. Due to the limitations of the statistical method used, I believe that if the number of genes is less than 20, it would result in a significant bias.

Response: Thanks for the comment. We have included the number of genes in each GO term, updated the figures (Fig.2B, Fig. 3D, and Supplementary Figure 15) and data sheet (Supplementary Data). Upon examining the number of genes enriched in these pathways, we found that some pathways had fewer than 20 enriched genes. However, we also found that most of the Gene Ontology (GO) databases inherently contain a limited number of genes. Therefore, in most cases, the available data could serve as a reference.

4. Fig. 2A selects some key genes, and the method used to determine the function of these genes should be detailed in the Materials and Methods section. Providing the phylogenetic relationship of these genes with the source genes described in the literature in the supplementary material is necessary to help readers better understand the functional annotation of these genes.

Response: We appreciate your suggestions. We have provided the phylogenetic trees of the candidate genes with the source gene and other homologous genes in Supplementary Figures 6 and 7 according to your suggestion.

4. To distinguish somatic variations from germline ones, the authors identified variations unique to PN population. I don't fully understand the underlying assumptions of this strategy. Why wouldn't somatic mutations found in the clone population also be present in the wild population? Are 10 EU + 10 ME representative enough? If I include more wild populations, will the results change?

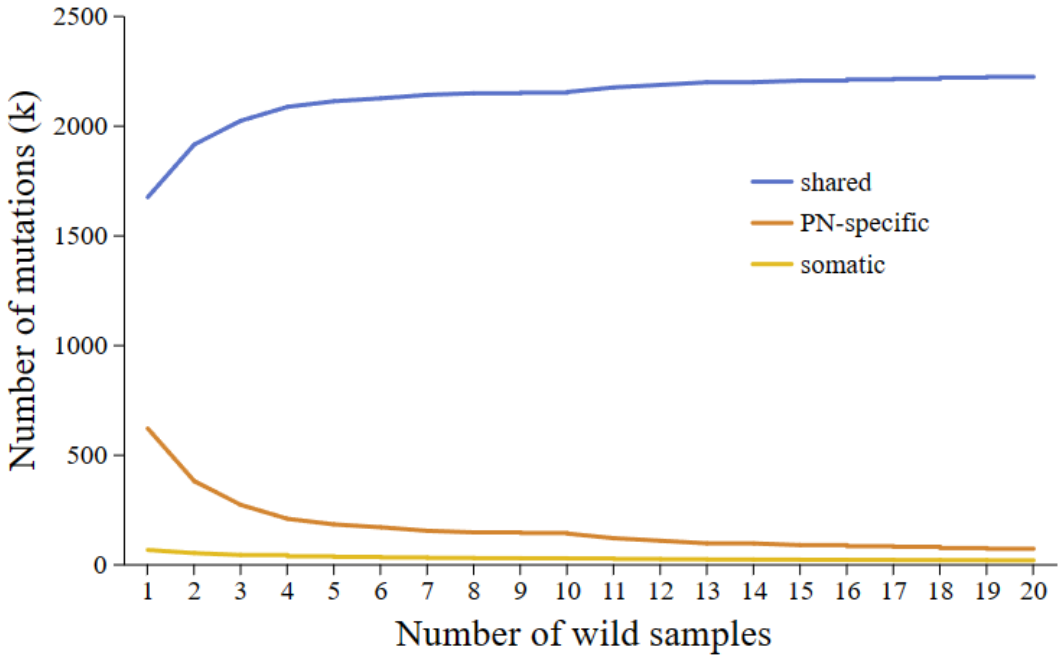
Response: We apologize for the unclear description. According to the definition of somatic variants, they are genetic mutations that occur in the body cells instead of the germ cells, which cannot be passed on to the next generation through sexual reproduction. Thus, to detect the somatic variations, we considered these mutations to be specific to the PN population and not shared by all PN individuals. The variants shared by all PN individuals are counted as germ-line mutations as these 18 individuals were collected worldwide (Supplementary Table 3), and not possible the clones of the same PN sample. As to the somatic mutation in PN presented in the wild population,

this could happen when PN lost the variation in wild grapes, and then mutation happened in the same site with the same type (A, T, C, or G) in PN as wild grapes. However, the possibility is extremely low. We can make a rough estimation, the size of the grape genome is around 500M, the total variations in the wild population is around 2.2M, and there are four types of nucleotide (ATCG), so a new mutation is the same as the old one is:

$$2.2/500/3=0.15\%$$

Thus, we think the effect of this situation is negligible.

The number of wild populations used could affect the results of somatic variants detection but is limited. We use wild grapes to detect PN-specific mutations and somatic mutations are counted as the mutations of PN-specific mutations only existed in part of the PN samples. If the mutations originated from wild grapes, they should be shared by all PN samples as they were clonal propagation. Of course, we didn't consider the small probability event, such as back mutation, as mentioned above, which has a very limited impact on the result. We estimated the effect of different numbers of wild grapes used. As we can see (Response Figure 1), when the number of wild grapes increased the number of somatic variants detected tended to be stable quickly. More wild grapes included can but in a very limited impact on the result. Thus, we think the result using 20 wild grapes in this study is credible.



Response Figure 1. Effects of the number of wild grapes used on mutation detection. The blue line indicates the number of the mutations shared by PN and wild grapes; the orange line indicates the number of PN-specific mutations; the yellow line indicates the number of predicted somatic mutations.

5. Line 283, the authors claim that "Similar patterns were also found on structure variation and deleterious SNPs in each group. (Fig. 4C-D) ". What exactly is this pattern?

Is “the heterozygosity of PN, SG, and HE is higher than wild grapes”? Is it statistically significant in Fig. 4C-D to claim that?

Response: We apologize for the unclear description. We have revised these sentences (Lines 275-285) and made a statistical analysis according to the reviewer’s comment which shown on Figure 4D&4E. We found that different mating systems could affect the accumulation of deleterious SNPs and SVs, especially on the variants in heterozygous states. We have made a detailed discussion in Discussion section (Lines 468-479).

6. The “protein phosphorylation pathway” was under selection in PN genome (Fig. 2B) and also keep heterozygous in PN40024 and HE but homozygous in wild grapes (Fig. 5B). I am curious if there could be any connection, either biologically or computationally.

Response: We greatly appreciate your suggestions. However, we redid the GO analyses with another tool to confirm the reliability of our results. It seems there are many difference between the new results and the old one, and we have updated the manuscript with the new results. The previous analysis was based on *Arabidopsis* protein database, which is different from grapes. In the revised manuscript, we constructed the database using grape genome annotation first, and conducted GO analysis using clusterProfiler (Wu, T. et al., 2021 Innovation (Camb)) as shown in Material and Methods section (Lines 718-728). We think this updated results is more reliable.

Reference:

Wu, T. et al. clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. Innovation (Cambridge (Mass.)) 2, 100141; 10.1016/j.xinn.2021.100141 (2021)

8. The data is presented twice in the same sentence (lines 281 and 283).

Response: Thanks you. We have corrected it.

9. What do the different colors represent in Fig. S11?

Response: We apologize for the unclear presentation of the figures. We have corrected this figure. The gray color indicates the sites with 0/0 genotype, the red color indicates the sites with 0/1 genotype and the blue indicates the sites with 1/1 genotype. It is Supplementary Figure 14 in the revised manuscript.

10. Line 39. The “cossing” should be “crossing”.

Response: Thank you. We have corrected it.

Reviewer #2 (Remarks to the Author):

This work of Wang et al provides new data leading to a chromosome-scale assembly of an important grapevine cultivar and to an inventory of somatic variations in its clonally propagated plant material. In the version in which it is currently presented, the manuscript is missing a clearly defined biological question.

Response: Thank you for your time and constructive comments. We have weakened the parts of genome assembly and paid more attention to the effect of the mating system on grapes to make our research focus on the effects of mating systems on grape genome and breeding. In previous studies, the effects of outcrossing, cloning, and selfing have been discussed. However, all of them were independent research. Here, we tried to observe this in a continuous process. Fortunately, we found the grape materials including various mating systems in a lineage, which could help us further understand the interaction effect on the grape genome between different mating systems that can't be studied in an independent system. We think this is significant for future grape breeding. We had specified the research questions in the Introduction part. We hope the revised version is more suitable for publication on NC.

Title, abstract, and a large part of the Introduction section anticipate that this manuscript is going to focus on the different modes of propagation in grapevine (vegetative, selfing, outcrossing) and their effect on genome properties and variation in the cultivated germplasm. However, a large part of the results and discussion sections deal with the genome assembly of a single genotype (Pinot Noir) and its comparison with the genome assembly of one descendant of Pinot Noir (PN40024, an experimental genotype forcibly obtained from reiterated selfing starting from an offspring of Pinot Noir). As for the parts of the manuscript that deal more specifically with the topics anticipated in the introduction, the authors move from observations in a few cultivated genotypes and draw generalizations for the crop germplasm.
Response: Thank you for this suggestion. We modified the manuscript and significantly shortened the genome assembly part, and paid more attention to the effect of the mating system on grapes. The mating system transition might have more significant effects on genomic variation than previously focused evolutionary factors, such as genetic drift and selection. According to the reviewer, we have modified the results and discussion to give cautious conclusions from our results. Although these results are drawn from grape data, we think they also provide valuable information on other crops.

The manuscript has elements of potential interest for the grapevine scientific community that require, however, both a deeper and a wider investigation to yield robust and conclusive evidence, but also several aspects of weakness that should be more carefully considered (see below for specific points).
Response: Thanks for the comment. We have revised the manuscript to yield a deeper and a wider influence of this study. We also added more analyses and modified the manuscript to give cautious conclusions from our results.

Specific points:

Line 68: transformation from dioecy (obligate outcrossing) to monoecy (potential self-pollination and cleistogamy).
Response: Thanks. We have modified the sentence.

Lines 69-70: history of outcrossing and clonal propagation of highly heterozygous genotypes.
Response: Thanks. We have modified the sentence.

Lines 113-120 are inappropriately placed in the Results Section.
Response: Thanks for the comment. We rewrote this paragraph (Lines 109-114) and moved some of the content into the introduction (Lines 82-84)

Line 128: it should be made more explicit at first citation and in figure legends that the acronym PN_T2T refers to the T2T genome assembly of PN40024, an individual that was obtained by repeated rounds of selfing, starting from Helfensteiner.
Response: Thanks for your suggestion. We have modified these sentences (Lines 122-124).

Lines 146-155: the 4-Mbp inversion on chromosome 19 should be experimentally validated using a PCR-based assay. This assay could also be used for assessing the origin and frequency in the grapevine germplasm of the haplotype carrying the inverted region.
Response: Thank you for the suggestion. We have validated the 4-Mbp inversion using PCR according to your advice (Supplementary Figure 4). We also examined the frequency of this inversion and found it only existed in heterozygous Pinot Noir, but lost in PN40024. In the future, with more available haplotype resolved assemblies for grapes, we could conduct a population genomic analyses of this INV and other ones.

In the paragraph “Introgression and selection shaped the Pinot Noir genome”, the authors explained that they “used resequencing data from 38 samples (Table S2), including 18 Pinot Noir (PN) clones, 10 EU wild grapes, and 10 ME wild grapes, to explore the introgression and selection signals in PN”. It is unclear how they conducted this analysis to identify introgressed regions (I assume from the EU wild grape population) into domesticated lineages deriving from ME wild grape population (then leading to Pinot Noir), selective sweeps and candidate genes under selection, using a group of Pinot Noir clonal replicates with software that normally applies to comparisons among populations. As the resulting candidate genes (deriving from adaptive introgression ? from selection of ME ancestral diversity ?) are then described in detail and discussed extensively in Lines 162-204, their identification should be supported by stronger evidence or accompanied by a more detailed description of the analytical procedure.
Response: We apologize for the unclear description. We have modified the sentences in the Results (Lines 157-191) and Material and Methods (Lines 664-665) section. For introgression detection, we used fd analysis which is based on the classical ABBA method. In this analysis, it used 4 pops (P1, P2, P3 and outgroup). It can detect the introgression between P2 and P3 using P1 and outgroup as control. In our study, we used ME as P1, PN as P2 and EU as P3. For selection analysis, we make CLR analysis on the data of the PN group using “Sweed”, a widely used software for selection region

detection.

Line 206: what is Pinot Noir (PN) population? The group of Pinot Noir bud sports, the offspring Helfensteiner and PN40024, as described in Lines 555-556 of the Methods section?

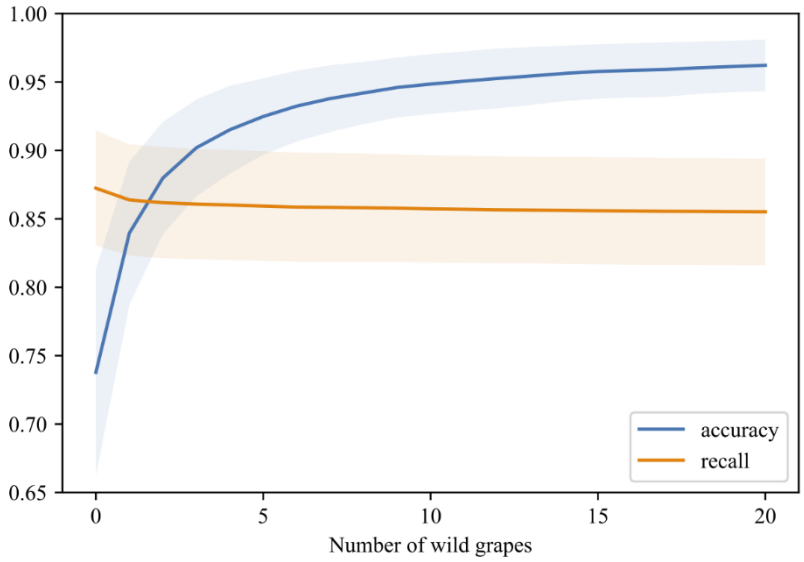
Response: We apologize for any confusion. PN population is 18 clonal Pinot Noir samples that we download or sequenced by ourselves (Supplementary Table 3 & 4). Sorry, we make a mistake in the description of the PN40024 in the Material and Methods section. PN40024 is not Pinot Noir and is not included in the PN group. We modified this sentence in Lines 634-635 of this updated manuscript and also presented the grouping information in Supplementary Table 3 & 4.

Paragraph "Somatic mutations in Pinot Noir clonal population" and Lines 553-565

It makes little sense using the reference genome assembly of a different genotype (derived from selfing of Helfensteiner) to realign short-reads of Pinot Noir bud sports and call somatic variants, when a haplotype-resolved T2T assembly of Pinot Noir has been generated in the present paper. A preprint that has recently appeared in bioRxiv shows the power of the use of a haplotype-resolved T2T assembly of the very same diploid genotype in the identification of somatic variants among Malbec bud sports (https://doi.org/10.1101/2023.11.30.569420). Regardless of the reference used, experimental validation of a set of candidate somatic variants needs to be provided in order to estimate false discovery rate.

Response: We appreciate your constructive comments. We have read this paper and cited it in our manuscript. We agree with you that it might be a better idea to use hap1 and hap2 assemblies of Pinot Noir as reference genome in somatic mutation detection. However, to compare our results with previous population genetic studies, we chose to use PN40024_T2T assembly as a reference genome for somatic variants identification since it has better genome annotations.

Based on our previous investigations, most of the somatic mutations are heterozygous. It is impossible to distinguish the somatic mutation with germline mutation. So we take the comparative population genetics approach using the wild grapes (ME and EU) as a control. It turns out to be very efficient showed in our forward simulations (SLIM). As shown in the response Figure 2, using wild grapes as a control could decrease the false positive of our detection and the accuracy of our somatic mutation detection is more than 95%, and the recall rate is more than 85%.



Response Figure 2. Estimation of the detection efficiency using simulated data by SLIM.

Line 254: wording would suggest that the authors used 4 independent lines obtained from selfing of Pinot Noir. Is it correct? Or are they clonal replicates of PN40024? PN40024 is a single seedling obtained from several rounds of selfing starting from selfing of Helfensteiner.

Response: Thanks for your comment. PN40024 is a single seedling from nine rounds of selfing starting from Helfensteiner. All the 4 PN40024 lines are clones. Sorry for the confusion due to our unclear description. We have modified the sentence (Line 246).

Paragraph “Effect of various mating systems on grape genomes”.

The authors used “groups” of cultivated varieties that were formed by one single genotype each, and three “groups” of these consisted of a parent-offspring trio. The forth “group” consisted of one single experimental highly inbred line (PN40024), showing severe inbreeding depression, that has no parallel in the grapevine cultivated germplasm. While π is largely independent from the number of sampled chromosomes and among-groups comparisons would make sense even using one diploid individual per group, data and inferences reported in Lines 279-294 refer to individual genomes and should be used with caution for inferring generalization on the species and on crop germplasm, as those discussed in Lines 377-434.

Response: Thank you for your insightful comments. We have rewrote the whole Discussion section, drawn most of the conclusion on the cultivar we focus and make general discussion with caution.

Paragraph "Conserved heterozygous regions in cultivated grapes".
Lines 296-300. These sentences need rephrasing.
Response: Thanks. We have modified the sentence (Lines 299-303).

Lines 303-305. Why intriguingly ? Heterozygous regions in PN40024 must be heterozygous in Helfensteiner by definition, as PN40024 derived from selfing of Helfensteiner.
Response: Thank you. We have modified the sentence (Lines 306-307).

Lines 333-334. Again. Why SVs and hSNPs should differ in heterozygous regions between Helfensteiner and PN40024, if both PN40024 homologs are identical by descent with respect to Helfensteiner ?
Response: Thanks for your suggestions. We understand that all the heterozygous variants in heterozygous regions should be also heterozygous in Helfensteiner. However, it different from the comment above involved Lines 303-305, we assumed that the number of these variants (SVs, hSNPs) should be less than Helfensteiner due to the selfing process, as shown in other regions, the heterozygous hSNPs and SVs almost disappeared. However, we observed that these variants are almost no change in PN40024 compared to Helfensteiner. That's why we said "interestingly". Sorry for our unclear description. We have modified the sentence to make it clearer (Lines 354-356).

Paragraph "Close leakage of deleterious variants in repulsion phases"

I understand that this paragraph deals with the possible reasons (i.e. linked heterozygous deleterious variants in repulsion on the two homologs), other than chance alone, explaining why residual heterozygosity was found in specific chromosome blocks of PN40024. Figs. S13-S14 do not seem to suggest that residual heterozygous regions in PN40024 contain more deleterious variants than other heterozygous regions in Helfensteinter, which later became homozygous in the descendant upon selfing. In supporting their claim, the authors also should provide evidence that what is shown in Fig. S15 is not simply the result of incorrect variant phasing across heterozygous regions, because hap1 and hap2 are both identical by descent between Helfensteiner and PN40024 and do not appear as such in the graphical representation of Fig. S15.
Response: We appreciate your suggestion. We used BEAGLE to phase our VCF data. The result of BEAGLE has been verified as reliable and widely used in other studies. In the revised manuscript, we made more analyses and restructured the latter part of the results (Lines 299-383). We used the population specific marker to make a discussion on recombination on the PN40024 genome during selfing, and focused on the heterozygous region on Chr2 which includes SDR regions to explain what we find in this study. We hope this revised manuscript presents more persuasive results.

Lines 385-387: This is a generalization that is inferred from observations in only three individuals.
Response: Thank you for the comment. We have modified the sentence and the whole Discussion section.

Methodology for the assembly of the Pinot Noir genome in the Methods Section

Lines 481-489: in the generation of the HiFi-based assembly, Hi-C reads were repeatedly used at the stage of contig generation (using Hifiasm) and scaffolding (using Juicer) as well as for other intermediate steps (also in combination with RagTag-based contig ordering) that are not clearly described.
Response: We apologize for the previous unclear description. HiC reads are used in two places in the assembly process, one at HiFiasm to assemble the raw HiFi data to the contig level, and the other is to divide the contig segments to the chromosome using HiC to the scaffold level. In HiFiasm, the HiC data are simply used for typing into two sets of haplotypes, and in the subsequent Juicer stage, the HiC data is further utilized in depth for contig segmentation and assembly to the scaffold level. Meanwhile, we used RagTag just to perform a simple chromosome localization using the reference genome PN40024_T2T.

Lines 495-500: the procedure used for integrating and merging the HiFi-based partial assembly and the ONT-based partial assembly into a haplotype-resolved T2T final assembly needs a more clear and detailed explanation. Adding Supplementary Tables that report for each chromosome contig-level and scaffold-level metrics obtained for the HiFi-based assembly and for the ONT-based assembly would help follow procedure and the results described in the main text.
Response: Thank you for your suggestion We are sorry for the unclearness. We actually assmebled the haplotype resolved genome using a combination of HiFi and HiC reads. The ONT reads were noisy with low quality, so that we didn't include it in the revision. We have modified the methods part (Lines 567-570) and the differences in N50 values for contig and scaffold levels are presented in Supplementary Table S1

Line 538: structural annotation should read structural variation
Response: Thanks. We have corrected it.

References to the additional homologs illustrated in Fig. 3D should be added.
Response: Thank you for your comment. We presented the detailed results of GO analysis in Supplementary Data.

Reviewer #3 (Remarks to the Author):

Wang et al's manuscript, "Various mating systems shape grapevine genome evolution" uses genomic resources to investigate evolutionary processes in grapes. Overall, the

authors seem to aim to "To investigate the genomic effects of mating systems during clonal, crossing, and selfing" (L351; first line of Discussion, and this phrase is consistent with the message of the Abstract). It is noteworthy that previous studies have addressed these same questions; e.g., the effect of clonal growth on somatic mutations and the effect of self-fertilization on genetic diversity, so the ideas presented in the manuscript are not very novel, although the data are new and useful. (Frustratingly, the authors fail to point out that a large body of work has addressed the effect of mating system and clonality on genetic diversity and plant fitness, which makes the current manuscript appear more novel than it really is. With respect to clonal growth, see this review from 2015; much other work examines the effect of selfing - https://www.pnas.org/doi/abs/10.1073/pnas.1501712112) The authors do present data that suggest an increase in somatic mutations curing clonal growth (which is consistent with previous work see link provided above), and these data are interesting. The authors also present results that suggest that some genome regions remain exceptionally heterozygous and suggest that this heterozygosity is maintained by deleterious alleles in repulsion phase. This latter result might be exciting, but I maintain healthy skepticism as it would require very strong selection (e.g., akin to recessive lethal effects) to maintain the patterns of genetic diversity documented here (assuming I understood the paper). Overall, the authors present a new genome with some analysis, and provide some results that are potentially exciting. However, my sense is that the ideas tested here (and some of the main results) are not as novel as the manuscript leads the reader to believe.

Response: We are grateful for your positive views and constructive comments for our manuscript. We have read the related papers and cited them in appropriate places. In previous work, some research could have been done on clonal and selfing population independently. However, in this study, we discussed the crossing, clonal, and selfing in one family. We observed the effects of different mating systems in a successive process. In the revised manuscript, we reshaped the introduction and discussion parts and added some new analyses, such as discussing recombination based on specific markers, and restructured the latter part of results and the whole discussion section. We hope this new version is more suitable for publication on NC.

GOALS (L93-110):
1) Look at Pinot Noir genome... no real goal stated other than "look at intricacies";
2) Look at processes to generate Pniot Noir
3) Examine structural variants and deleterious variants and their changes
4) Why do some regions remain heterozygous in PN40024 -

OUTPUT 1 (L121 - 155) Genome of (heterozygous) Pinot Noir - Does this meet their goal 1? They highlight an inversion.

Response: Thanks for your comment. The haplotype-resolved chromosomal Pinot Noir genome was assembled to compare it with PN40024_T2T (its offspring associated with crossing and inbreeding generations), which was used to detect the effects on mating systems on genome structure and contents. According to your suggestion and other

reviewers, we shortened this part and put more attention on the results of the effect of the mating system. We also removed the verification of inversion to supplementary files. We added the results of the comparative genomics to detect the effects of mating systems on genome structure and contents in clonal grape (PN) and selfing grape (PN40024).

OUTPUT 2 (L157 - 203) - Introgression analysis of Pinot Noir Genome - IS THIS GOAL 2?

- This section provides evidence for selection on a few genes, but says little (if anything?) about the role for introgression in the evolution of the Pinot Noir genome, despite the title of this section (L156) indicating that introgressions is a process of interest. It is unclear to me whether the authors even examined for evidence of introgression - if they did, how much evidence for introgression did they detect (and in which direction between regions)?.

Response: We appreciate your input. We modified the sentences in this part (Lines 149-191). There are many studies showed that a strong introgression in Wine grapes from EU wild grapes. But most of them are performed on population level that involved many cultivars. In this study, we tried to confirm this introgression signal based on Pinot Noir population to make a deeper understand of this famous wine cultivar. We used f_d statistic method, one of the most popular methods for introgression detection, to evaluated the gene flow between PN and EU grapes. As shown in Fig. 2C & 2D, the PN genome has affected by EU significantly. However, we didn't pay much attention on this part, as we focused on the effect of mating system on the PN lineage grapes.

SOMATIC MUTATIONS IN PINOT NOIR CLONAL POPULATION (L204)

- The authors state, "To distinguish somatic variations from germline ones, we identified variations unique to PN". Is is unclear to me how this approach can reliably detect somatic mutations because it assumes that the mutations that have been identified in the small number of samples of European and Middle Eastern samples (20 genomes, total) will include *all* possible mutations that could have been present in the ancestor of the Pinot Noir clonal population. i.e., with only 20 EU and ME samples the authors have sampled only 40 alleles from the wild populations (2 alleles for each of the 20 samples, with 10 from EU and 10 from ME). Therefore, the authors are unlikely to identify variants at relatively low frequency in wild populations that can be mistaken as somatic mutations. Notably, the authors use patterns Fig 3B to support arguments that 18 mutations arose somatically. However, it is unclear to me how the authors distinguish between whether these mutations arose somatically *after* clonal propagation began vs. were simply unique to Pinot Noir and present in the genotype that gave rise to the clonal genotype *before* clonal propagation began. The authors may be correct in their logic, but they need to clarify their presentation to eliminate this alternative explanation to their data. For example, it would be useful to include a figure that illustrated the evolutionary process that the authors imagine that yields the pattern observed in their data.

Response: We greatly appreciate your suggestions. For the first question, the number

of wild populations used could affect the results of somatic variants detection. Thus, we estimated the effect of different numbers of wild grapes used. As shown in Response Figure 1, when the number of wild grapes increased the number of somatic variants detected tended to be stable. More wild grapes included can but in a very limited impact on the result. Thus, we think the result using 20 wild grapes in this study is credible.

For the second question, to detect the somatic variations, we considered these mutations to be specific to the PN population and not shared by all PN individuals as they are genetic mutations that occur in the body cells instead of the germ cells. The variants shared by all PN individuals are counted as germ-line mutations as these 18 individuals were collected from worldwide, and not possible the clones of the same one PN sample. We are sorry not to describe clearly that the PN-specific SNPs shown in Fig. 3B are not only somatic mutations and also include germ-lime mutations. As shown in Fig. 3C, the mutations only shared by less than 18 PN samples were counted as somatic mutations. We have modified these sentences to make it more clearly (Lines 193-210)

We also used SLIM to simulated data to estimate the accuracy of our methods. As shown in response Figure 2, using wild grapes as a control could decrease the false positive of our detection and the accuracy of our somatic mutation detection is more than 95%, and the recall rate is more than 85%.

L251: EFFECT OF MATING SYSTEMS ON GRAPE GENOMES - GOAL 3(?)
L265 - the authors state, "Due to nine generations of selfing, PN40024 was not counted as parent-offspring relationship with the other three cultivars." What evidence do the authors have to conclude that it was selfing, per se, that caused the analysis to fail to indicate that PN40024 was not counted as an offspring in the analysis? Selfing, per se, should only reduce genetic diversity by increasing homozygosity - therefore, it is unclear to me how the "Kinship" analysis would be so affected by increased homozygosity. The authors need to either further justify their assertion or consider either (1) that PN40024 does not derive from Helfensteiner (although this is unlikely) or (2) there are issues with the Kinship analysis, more generally, which should make the reader question *any* conclusions from this analysis.
Response: Thanks for the comment. The relationship between PN, SG, HE and PN40024 is already known. The reason we make kinship here is to confirm that the samples we used are true to type and present some characteristics between these 4 grapes based on Kinship. The standard for judging the relationship between grapes based on Kinship has been done in previous studies (Roach M et al., PloS genetics, 2018; Jazmín, R et al, 2018, Nature Plants) and has been proven reliable in other studies. We have revised the sentence to make it clearer (Lines 255-260).

Reference:
Roach, M. J. et al. Population sequencing reveals clonal diversity and ancestral inbreeding in the grapevine cultivar Chardonnay. PLoS genetics 14, e1007807;

10.1371/journal.pgen.1007807 (2018).
Ramos-Madrigal, J. et al. Palaeogenomic insights into the origins of French grapevine diversity. Nature plants 5, 595–603; 10.1038/s41477-019-0437-5 (2019).

L275: When comparing genetic diversity of inbred vs. outcrossed varieties, the authors conclude "the selfing PN40024 exhibited the lowest values for both diversity and heterozygosity when compared to all other groups". This is not a surprise, as the impact of selfing on genetic diversity is understood well in theoretical terms and is well documented empirically. Therefore, while this result adds to current data on the effect of self-fertilization on genetic diversity, it does not expand our understanding.
Response: Thanks for your comment. We have made more analyses on other aspects of slefing PN40024. We restructured the latter part of the results (Lines 261-383). We used the population specific marker to make a discussion on recombination on the PN40024 genome during selfing, and focused on the heterozygous region on Chr2 which includes SDR regions to explain what we find in this study. We hope this revised manuscript presents more persuasive results. We have also rewrote the discussion to make a deeper and a wider presentation of the significant effects of various mating systems on crop genome and breeding.

L280: The authors also find that the genetic diversity of colonal propagated lines exceeds that of outbred populations; "the heterozygosity of PN, SG, and HE is higher than wild grapes, ME, and EU". They authors conclude that clonal propogation increases genomic diversity and this this occurs via (in part, at least) deleterious mutations. These data are very useful, and are consistent with previous findings (see a review in 2015 here: https://www.pnas.org/doi/epdf/10.1073/pnas.1501712112) that also suggest that clonal reproduction increases opportunity for somatic mutation. The authors need to highlight that previous work also points to the accumulation of deleterious, somatic mutations during clonal growth; if the authors wish to claim that their results are somehow novel, they need to clarify which aspects are indeed novel.
Response: Thank you for your thoughtful recommendation. We have carefully cited the references you provided. We also modified the latter part of the results to provide a comprehensive view of the manuscript.

CONSERVED HETEROZYGOUS REGIOUS IN CULTIVATED GRAPES (GOAL 4?)
L300: The authors "identified 379 alleles that remain heterozygous in all grape samples (wild and cultivated)." The authors further remark, "Remarkably, we identified a total of 59,944 alleles that retained their heterozygous status within PN40024. Intriguingly, these heterozygous genes exhibit consistency between PN40024 and HE, and they also remain in heterozygous states in at least one PN and one SG sample." In reading this, my first reaction was that this result can arise due to errors in genome assembly; the authors examine read coverage to address this possibility (L318 - 327 - this is good) and remove several cases where this explanation likely is true. The authors conclude that heterozygosity is maintained due to deleterious mutations in repulsion-phase (also termed pseudo-overdominance; see Charlesworth, Deborah, and John H. Willis. "The

genetics of inbreeding depression." Nature reviews genetics 10.11 (2009): 783-796.) [[Please note that what I say next relies on my understanding of the author's phrase, "alleles that remain heterozygous in all grape samples", as this implies to me that all individuals that were genotyped were heterozygous at the focal loci - if I mis-understood the author's meaning, then my following comments need to be adjusted.]] If the authors' explanation is true, that multiple loci remain heterozygous in **all** genotypes individuals, then this implies *very* strong natural selection in order to maintain heterozygosity in all sampled individuals. Personally, I am surprised by (and therefore have healthy skepticism for) this finding and suggestion; I do find it interesting, but encourage authors to emphasize the requirement for very strong selection to maintain this pattern of genetic variation, and I encourage the authors to point out that this finding does require further investigation to confirm the suggestions made in the current manuscript. I can illustrate my skepticism mathematically. Take, for example, a single locus that remains highly heterozygous due to very strong selection: in this case, self-fertilization (which is how PN40024 reproduces) will lead to homozygous offspring at this locus for 50% of its offspring and many of them will experience a large fitness cost due to selection at this single locus. Now, the authors have implied that multiple loci remain heterozugous due to strong selection - if this is correct, then the fitness cost I just described for a single locus will multiple across other genome regions. Overall, we expect that most offspring that derive from self-pollination of PN40024 to have low fitness if the authors' explanation for maintaining heterozygosity is correct (and if I've understood their manuscript correctly); is this seen empirically? If not, then my sense is that the authors need to either (i) clarify their presentation (because I may have mis-understood aspects I highlight above), or (ii) amend or clarify their arguments / re-evaluate their findings. Are genome assembly or read-mapping errors possible reasons for the observed results?

Response: Thank you for this great question. First of all, we has reevaluated our analyses that our results based on population samples (n ≥ 4) were solid and reliable. We do see a strong inbreeding depression in PN40024, it barely reproductive, since reproductive were selected during each sdlfing generations. In the five heterozygous regions, we also detected genes enriched in the biological functions including reproduction and other important biological process for plant life. We also added the literature you suggested and reshaped the main text, please see the revised manuscript for details.

Reviewer #1 (Remarks to the Author):

I have read the revised MS, which has been improved a lot based on the comments and suggestions of reviewers including mine.

1. I suggest not using exaggerated expressions, like I propose the following modifications： change “We found very interestingly that the mating systems, namely, crossing, clonal, and selfing have dramatic effects on grape genomic variation and breeding based on comparative genomics and population genetics.” to “We found that the mating systems, namely, crossing, clonal, and selfing have significant effects on grape genomic variation and breeding based on comparative genomics and population genetics.”

Response: We are grateful for your positive views and constructive comments for our manuscript. We have scoured the manuscript to remove superlatives.

Reviewer #2 (Remarks to the Author):

The revised version R1 is mostly text editing, with two exceptions: the requested wet-lab validation of a large inversion in Pinot Noir and a new analysis that focused on the seemingly forced maintenance of the heterozygous state at the flower sex locus upon selfing of a F/H2 individual, which could theoretically become H2/H2 homozygous at any selfing generation, yet maintaining an hermaphrodite phenotype for the next round of selfing, but it never did so until the S9 generation.

In my previous remarks, I pointed out that a deeper and a wider investigation to yield robust and conclusive evidence on the effects of mating systems on genetic variation was required to support the authors’ claims. From my point of view, there is not additional data in the revised version that makes the overall manuscript robust enough for being accepted for publication in a high impact journal such as Nature Communications. Most of the remarks were not addressed convincingly or exhaustively.

Response: We are grateful for the insightful feedback. Based on your comments, we have made the following revisions: 1) We included additional grape cultivars related to Pinot Noir and expanded the sample sizes for previously analyzed cultivars; 2) We clarified the definitions of key terms, including “somatic mutation” and “reproductive system”; 3) We reanalyzed the selective sweep using the PBS analysis instead of the

CLR analysis; 4) We performed further analysis on the sex determination region within the heterozygous blocks of selfing PN40024, which further clarified the Hill-Robertson effect. We identified a candidate lethal recessive gene that may maintain the SDR region in a heterozygous state in PN40024; 5) We enhanced the presentation of the Results and Discussion sections. We have provided detailed explanations in response to each of your comments, we believe these revisions have substantially strengthened the manuscript.

While some inaccuracy in the terms used in the original version of the manuscript for referring to populations, reproduction systems, mutation types seemed to be at that stage simply a matter of semantics, it seems now from their revised text and from their replies to the reviewers' remarks that it actually reflects some misconception.

Response: We completely agree that some of the terms we used were too loose, and we have revised them following your suggestions to avoid misunderstanding.

An illustrative yet not exhaustive list includes: vegetative propagation is not a mating system, a group of phenotypically divergent clonal lines is not a population, heterozygosity and chimerism of somatic mutations are not treated separately, chimerism is entirely ignored in the analysis.

Response: Thank you for this comment. Mating system was, indeed, an case where the vocabulary used was inappropriately and did not match the intent. We made concerted efforts to correct the terminology throughout. We address chimerism and other aspects of this comment below.

Some misconception has not only theoretical implications but also important consequences on the adopted methods of analysis.
As an example, the authors state in their response to previous remarks and therefore assume in their analyses "that somatic variants, they are genetic mutations that occur in the body cells instead of the germ cells, which cannot be passed on to the next generation through sexual reproduction", which is not true. Novel somatic variation can be sexually transmitted if the mutation has occurred in somatic cells belonging to the cell layer of the shoot apical meristem that is going to generate the germ line.

Response: We apologize that the previous description was unclear; clearly there are problems with the quoted statement. We are, however, puzzled by the emphasis on chimerism. This comment refers specifically to the part of the paper that identifies mutations specific to individual Pinot Noir clones. These mutations can, indeed, be chimeric; that we readily admit. But chimeras are usually considered a subset of somatic

mutations and treated as such. We did not set out to identify chimeras, nor can our sampling (based on whole leaf tissue) reliably distinguish the subset of chimeric mutations from other somatic mutations. We have inserted some language about chimerism and our inability to distinguish chimeric mutations from other types (e.g., heterozygous somatic mutations that are fixed throughout cells and tissue layers) in the main text (Discussion). Please do note that previous work on somatic mutations in plants (including grapes) also do not discriminate between chimeric and other types of somatic mutations. Some of the confusion likely came from our use of the terms 'germline' and 'somatic', which were indeed unclear. For example, to avoid misunderstanding, we changed the "germline mutation" to "fixed mutations" (Lines 193-195).

The authors also state and therefore assume that "to detect the somatic variations, we considered these mutations to be specific to the PN population and not shared by all PN individuals. The variants shared by all PN individuals are counted as germ-line mutations as these 18 individuals were collected worldwide (Supplementary Table 3), and not possible the clones of the same PN sample." The condition mentioned by the authors to be verified in the group of the 18 stocks that were derived from the historical vegetative propagation of the original seedling of Pinot Noir is not sufficient for distinguishing somatic mutations and germline mutations. Part of the somatic mutations might also be private to Pinot Noir, yet shared by all analyzed stocks, if they occurred before clonal lineages with distinctive phenotypic variation were intentionally selected by viticulturists and nurseries and if the model of clonal lineage evolution was tree-shaped instead of star-shaped.

Response: Thanks for the comment. The reviewer is technically correct, but the somatic mutations we analyze and discuss have not been fixed across Pinot Noir (Lines 193-195). We do recognize that the mutations fixed in PN (that is shared among PN lines but that differ from other non-PN samples) likely represent mutations prior to the divergence of PN clones. We mention that class of mutation separately and do not analyze them as somatic mutations. Interestingly, the clones we have analyzed appear to follow a star phylogeny (Figure 2C), which is consistent with our inference that most PN-specific variants are either private to one PN sample or shared by all PN samples.

Below is a detailed list of critical points that were not completely fixed or have newly emerged with the R1 changes.

Line 32 and elsewhere: clonal propagation is not a mating system. Mating systems always involve crossing and if this occurs between unrelated individuals it is defined as outcrossing, between related individuals is defined as inbreeding and with itself it is defined as selfing. Clonal propagation is a mean to propagate the same genotype and

does not involve mating between a male and female gamete, so it cannot be considered as a type of mating system.

Response: We agree and have modified the text accordingly from “mating system” to “propagation system”

Lines 33-35: There is no statistics in the results to support the claim that “The mating systems have more significant effects than traditionally focused factors such as genetic drift and natural/artificial selection.”

Response: Thanks for your comment. We deleted this sentence.

Line 37: “cloning increases genomic heterozygosity”. In terms of relative numbers, somatic mutations lead to a minimal increase of heterozygosity cultivar-wise, as shown by this manuscript and by previous research. Furtherly considering that a substantial part of this novel heterozygosity is expected to be chimerical within each individual, its population-wise contribution through subsequent sexual reproduction is even less impactful than cultivar-wise. Overall, this sentence is misleading for novel point mutations and it does not consider the opposite effect of somatic deletions. It has been shown exactly in Pinot Noir that vegetative propagation leads to local loss of heterozygosity in the case of nonchimeric deletions (i.e. white-skinned bud sport) and local reduction of heterozygosity in the case of chimeric somatic deletions (i.e. gray-skinned bud sport).

Response: Thanks for your comment. We deleted this sentence and have modified the related sentences in results (Lines 252-254). As for deletions, other work in grapes on a genome-wide scale has shown that deletion of structural variants increases heterozygosity because it affects only one haplotype (e.g., Zhou et al. 2019 Nature Plants; Vondras et al. 2019 BMC Genomics). We appreciate the example of a local loss of heterozygosity, but this local example flies against genomic trends that are now well substantiated.

Reference:

Zhou, Y. et al. The population genetics of structural variants in grapevine domestication. Nature plants 5, 965–979; 10.1038/s41477-019-0507-8 (2019).

Vondras, A. M. et al. The genomic diversification of grapevine clones. BMC genomics 20, 972; 10.1186/s12864-019-6211-2 (2019).

Lines 99-100: “Second, we embarked on an exploration of the processes that underly clonal diversity among Pinot Noir clones, including the identification of somatic variants and genomic regions that may contribute to phenotypic traits.” Actually, this paper only reports the identification of putative somatic SNPs and the list of affected genes. Association with phenotypes is not reported experimetally.

Response: Thanks for your comment. We have modified the sentence (Lines 96-98).

Lines 109-114: This knowledge was well-established prior to the present work.

Response: Thanks for your comment. We removed this part and Figures 1A and 1B in this new version.

Lines 115-126: In the revised version and in the response to reviewers’ comments letter, the authors explain that the new Pinot Noir assembly in the latest version of the manuscript is built by omitting the step of integrating the HiFi-based assembly and the ONT-based assembly, compared to the original version of the manuscript. In spite of this, the statistics of the assembly, in particular chromosome pseudomolecule sizes have not changed between versions. Overall, this adds new uncertainty and concern about how the assembly procedure was performed and how carefully it is described, rather than easing previous concerns.

Response: Thanks for your comment. In the original revised versions, the assembly was not modified and thus there is no change on the statistics of the assemblies. As we responded in the previous “response to reviewers’ comments letter”, we only used HiFi and Hi-C reads for assembling. There is a mistake in the methods of original manuscript, so we modified it in the R1 version.

Lines 139-140: I acknowledge that the authors have experimentally validated the inversion with a wet-lab assay, as requested.

Response: Thank you for help us with improving our manuscript.

Lines 141-147: How gene families and the boundaries between gene families were defined ? Do the numbers refer to the present or absence of entire gene families or individual gene family members? The closing comment is inappropriate: what the authors observed might be due to pre-existing presence/absence variation between the Pinot Noir and Schiava Grossa haplotypes that were donated to Helfensteiner and then to PN40024.

Response: Thanks for the comment. The gene families in our manuscripts were indentified by the Markov clustering algorithm using a reported strategy (Lines 572-577). Basically, the algorithm will cluster the genes that are similar to each other and distinct from genes in other groups based on sequence similarity. In our manuscript, "present/absence" reflects where there are genes within (or not within) a specific gene family. We also showed the "present" gene families by red gradients to indicate the number of genes in each family. For last comment, PN40024 inherited one haplotype of Pinot Noir, but it has far fewer gene families compared to both haplotypes of Pinot Noir. This indicates either that the other parent of PN40024 had a low number of gene familes or that some gene families were lost during selfing, which would be consistent with other studies of selfing lineages. We have modified the sentences to make them clearer (Lines 135-139).

Reference:

Liu Y, Du H, Li P, et al. Pan-Genome of Wild and Cultivated Soybeans. Cell. 2020;182(1):162-176.e13. doi:10.1016/j.cell.2020.05.023

Lines 148-191: My previous remark was not address either in the revised text or in the response letter. I do understand how SweeD and ABBA-BABA work, but how can this software do the job for which they were designed if fed with two populations that might be approximated to natural populations and another population that is not a population but a group of clonal replicates with no germline ancestral variation ? The group of 18 Pinot Noir clones contains only novel somatic variation and collectively captures only one or two ancestral haplotypes over each genomic region. Both software deal with allele frequencies within and between populations. In particular for the statistically supported detection of selective sweeps, it is totally unclear to me how this was applied to the specific case of this manuscript in which allele frequency in Pinot Noir that may only be 0, 0.5, 1 for germline SNPs, while somatic SNPs in Pinot Noir have by definition allele frequency = 0 in the ancestral populations.

See my previous comment on the original submission in detail. In the paragraph "Introgression and selection shaped the Pinot Noir genome", the authors explained that they "used resequencing data from 38 samples (Table S2), including 18 Pinot Noir (PN) clones, 10 EU wild grapes, and 10 ME wild grapes, to explore the introgression and selection signals in PN". It is unclear how they conducted this analysis to identify introgressed regions, selective sweeps and candidate genes under selection, using a group of Pinot Noir clonal replicates with software that normally applies to comparisons among populations.

Response: Thanks for the comment. We modified the work to apply the population

branch statistic to detect selected regions using the PBScan software (Hämälä, T. & Savolainen, O., 2019) (Lines 149-157). This method searches for regions that have higher divergence than expected, based on comparison to two groups, in this case two wild grapes (EU and ME) by comparing sequence similarity (Dxy).

I appreciate the reviewer's insights regarding ABBA-BABA; however, I believe there may be some misunderstandings about it for a couple of reason. First, it is a method not explicitly rooted in population genetics but rather phylogenetics. It examines shared alleles rather than allele frequencies. Thus, as long as it considers groups that appropriately distinct in the right phylogenetic order, it is valid. Second, and more practically, the method has been applied to groups of selfers, like collections of Arabidopsis and rice in well-cited and occasionally highly-cited manuscripts. These collections of selfers are of course not identical to clonal lineages, but they are similar in spirit to old clonal lineages. We continue to see no problems applying these tests to our data.

Reference:

Hämälä, T. & Savolainen, O. Genomic Patterns of Local Adaptation under Gene Flow in Arabidopsis lyrata. Molecular biology and evolution 36, 2557–2571; 10.1093/molbev/msz149 (2019).

Martin, S. H., Davey, J. W. & Jiggins, C. D. Evaluating the use of ABBA-BABA statistics to locate introgressed loci. Molecular biology and evolution 32, 244–257; 10.1093/molbev/msu269 (2015)

For supporting the authors' claim that most of Pinot Noir private SNPs are somatic mutations (approximately 50,000 out of 70,291), the plot in Fig.3B should report on the x-axis the read count ratio supporting the ancestral and the derived allele. The direction of the mutation can be defined using the ME and EU populations the define the ancestral state. In the present form, this distribution is only showing graphically a subsampling bias or suggesting that many of the putative somatic SNPs may be false variant sites.

Response: Thanks for the suggestion. We need to be clear: we did not claim that most of Pinot Noir private SNPs are somatic mutations. In Figure 3B, we observed a higher middle peak in PN-specific SNPs compared to all SNPs in PN. We hypothesized that many of them might be somatic mutations which in heterozygous state that contribute to this phenomenon. In Figure 3C, we presented that approximately 30% of PN-specific mutations are exclude to 1-17 PN samples that are putative somatic mutations.

As to false variant sites, we used GTX which is based on the Haplotype Caller of GATK for SNP calling. GATK is wildly used for variants detection and has been verified have a good precision on SNP calling. The Mendelian error rates of the data generated by GATK is ranged from 1% ~ 5% (Lin et al. 2022 Sci Rep.; Betschart et al. 2022 Sci Rep.). In this study, the samples we used had average mapping coverages of >20x, and we filtered the original VCF files by vcftools using command "--GQ 20" to keep the reliable variants. Thus, the effect of false variant sites is very limitation. It is vanishingly improbable that the ~30% of PN-specific SNPs are false variant sites. Either that or most SNP-based studies in all species are horrible incorrect.

Reference:

Lin, YL *et al.* Comparison of GATK and DeepVariant by trio sequencing. *Sci Rep.* 12, 1809 (2022).

Betschart RO *et al.* Comparison of calling pipelines for whole genome sequencing: an empirical study demonstrating the importance of mapping and alignment. *Sci Rep.* 12(1):21502 (2022).

Lines 211-212: As most of the putative somatic SNPs were predicted to be private to individual clones, this may be either indicative of a star-shaped model of clonal differentiation or, more likely, suggestive of a high incidence of false variant sites. In the absence of an experimental wet-lab validation of at least a random subset of candidate somatic SNPs to assess the false discovery rate, this part of the manuscript reporting on somatic SNP identification is weakly supported.

Response: Thanks for your comment. As mentioned above, the methods for SNP detection are widely used and reliable. With the coverage of our samples, the number of false variant sites is expected to be limited. We don't think it is necessary to conduct experiment for further validation. Please note note that several other studies have been published without such wet-lab verification (Monroe et al. 2022 Nature; Roach et al. 2018 PLoS Genetics; Vondras et al., 2019 BMC genomics). Generally, this request for wet-lab verification does not reflect the current state of the field.

Reference:

Monroe, J.G. *et al.* Mutation bias reflects natural selection in *Arabidopsis thaliana*. *Natur*e 602, 101–105 (2022)

Roach MJ *et al.* Pretorius IS et al. Population sequencing reveals clonal diversity and ancestral inbreeding in the grapevine cultivar Chardonnay. *PLoS Genetics* 14(11):1-24 (2018).

Vondras, A. M. *et al.* The genomic diversification of grapevine clones. *BMC genomics* 20, 972; 10.1186/s12864-019-6211-2 (2019)

Lines 215-216: The fact that putative somatic SNPs are more frequently associated with intergenic regions is another indication that many of them might be false variant sites associated with partial read misalignments in repetitive regions.

Response: Thanks for your comment. The increase in intergenic SNPs by mismapping is one interpretation, but we think it unlikely. First, we are mapping to the PN40024_T2T reference, so mismatches should be limited in this carefully resolved genome. Second, previous work documenting somatic mutations in grapes found a similar pattern (Vondras et al., 2019 BMC genomics), with elevated somatic mutations in intergenic and non-genic regions. Third, similar patterns have been found in mutation accumulation lines of Arabidopsis (Monroe et al. 2022 Nature), suggesting it is a general phenomenon. Fourth, we have used standard methods, such as keeping high-quality SNPs with GQ (genotype quality) > 20, that are well accepted in community. Finally, it strikes us as highly unlikely that multiple reads were misaligned at one region for one sample but not for other closely-related samples.

Reference:

Vondras, A. M. *et al.* The genomic diversification of grapevine clones. *BMC genomics* 20, 972; 10.1186/s12864-019-6211-2 (2019)

Monroe, J.G. *et al.* Mutation bias reflects natural selection in *Arabidopsis thaliana*. *Natur*e 602, 101–105 (2022)

Lines 199-201: Here and in general across this section, it seems that the issue of the chimerical state of somatic mutations has not even been considered in the analyses, which may provide alternative explanations to the data compared to those given by authors.

Response: Thank you for your comment, which (again) addresses the issue of chimerism. See our comment above.

Lines 222-244: only germline structural variation in the cultivar Pinot Noir is reported here. If the authors detected structural variation among clones, it should be reported in this paragraph. If not, they should state that they did not detect structural variation among clones.

Response: Thanks for your comment. In Lines 209-229 and Figure 3C, we presented the fixed and somatic SVs. We have tried to make this also more obvious in the text.

Lines 255-260: This part has already been shown and better explained by Velt et al 2023 10.1093/g3journal/jkad067.

Response: Thanks for your comment. We had cited this paper here in previous version. We didn't intend to clarify the relationship between these cultivars again. We want to use Kinship to confirm the true to type of these samples we used, and to let the reader to have a better understand of the relationship of them.

Line 263: The He symbol in population genetics is used for expected heterozygosity under Hardy Weinberg equilibrium while the heterozigosity measured here is rather observed heterozygosity. So it would be better to list it as either Ho or just H.

Response: We have corrected it.

Figure 4B: PN is erroneously labelled by HE in two instances.

Response: Thanks. We have corrected it.

Lines 266-272: These interpretations are questionable and not supported by the data presented in this manuscript. If one considers the number of somatic SNPs identified e.g. among the Pinot Noir clones, it is clear that in comparison to the total number of germinal SNPs that number cannot justify such a large difference in He. This difference could instead be attributed to the mating system because cultivated varieties are the result of crosses that often combine different genetic materials while wild populations may suffer from a certain level of inbreeding. It is commonly observed and generally accepted that heterozygosity is higher in grape cultivars than in the present-day individuals of the wild progenitor, but this is commonly explained by the fact that heterozygosity is exceptionally low in wild population due to small population effective size and geographical isolation leading to forced inbreeding rather than heterozygosity is exceptionally high in the cultivars due to other reasons.

Response: Thanks for the comment. We agree with you that the high heterozygosity in HE than in its wild progenitors could be due to small Ne in the wild population. However, in our previous studies and simulations, we had revealed that genomic heterozygosity is increasing along the clonal generations due to the accumulation of somatic mutations in heterozygous state (Zhou et al. 2019, Xiao et al. 2023). We modified the sentences in line 248-254.

Reference:

Zhou, Y. et al. The population genetics of structural variants in grapevine domestication. Nature plants 5, 965–979; 10.1038/s41477-019-0507-8 (2019).

Xiao, H. et al. Adaptive and maladaptive introgression in grapevine domestication. Proceedings of the National Academy of Sciences of the United States of America 120, e2222041120; 10.1073/pnas.2222041120 (2023).

Line 276: It is rather obvious that the dSNPs are seen less frequently in heterozygous condition in PN40024 where >90% of the genome is in homozygous state. So the comparison should not be with highly heterozygous accessions such as PN or SG or HE but rather it should be evaluated in comparison with the frequency of heterozygosity of non deleterious or neutral SNPs in PN40024

Response: Thanks for the comment. Here we intended to compare the heterozygous dSNPs in different propagation system. The heterozygous dSNPs are less frequently, this is a characteristic of this selfing cultivar.

Line 277: This should not be indicated as recessive because there is no information on the effect of the deleterious alleles in terms of recessive or dominant, so it should be indicated as homozygous.

Response: Thank you for your suggestion. We are using standard nomenclature from the population genetics literature that often refers to the dSNPs in a homozygous state as the 'recessive load'.

Reference:

Nigenda-Morales, S.F., Lin, M., Nuñez-Valencia, P.G. et al. The genomic footprint of whaling and isolation in fin whale populations. Nat Commun 14, 5465 (2023).

Zhou, Y. et al. The population genetics of structural variants in grapevine domestication. Nature plants 5, 965–979; 10.1038/s41477-019-0507-8 (2019).

Lines 279-280: this could be simply explained by the fact that selfing generations of Helfensteiner leading to PN40024 and PN40024 itself are experimental genotypes that were not exposed to natural selection and were, on the contrary, nurtured with extreme care for their survival and reproduction. This is an extreme artificial situation that has no parallel in grapevine evolution and viticulture. Artificial selection during the generation of PN40024 has only targeted hermaphroditism in order to proceed with the next round of selfing.

Response: Thanks for the comment. We have studied PN40024 simply because it is such

a rare opportunity to investigate how selfing has shaped the genome. Such studies are rare even for more-easily selfed species (but see, for example, Roessler et al Nature Plants for maize).

Reference:

Roessler, K. et al. The genome-wide dynamics of purging during selfing in maize. Nature plants 5, 980–990; 10.1038/s41477-019-0508-7 (2019).

Lines 289-296: Putative somatic mutations in Schiava Grossa, Helfensteiner and PN20024 are first mentioned here. If the procedure of identification was the same as in Pinot Noir, the numbers are not comparable among cultivars as they depend on the number of individuals that are compared within each cultivar.

Response: Thanks for the comment. We agree and have removed the related results here.

Line 295: truncated sentence

Response: Thanks for the comment. We corrected it.

Lines 329-339: This section is not very relevant because there is no quantitative analysis of the local recombination rates in this manuscript.

Response: Thanks for the comment. Based on specific marker, we can have a glimpse on recombination between two haplotypes originated from PN and SG in PN40024. Although the resolution is not high, we can observe the recombination on PN40024 genomes. More work is required in the future to further understand of the recombination, but we can't ignore what we have discovered.

Lines 374-386. This part is an interesting addition with respect to the original R1 version. In the end and overall, this is the only really exciting part of the manuscript. However, as it occurred to other parts, it only scratches the surface of the biological issue and would require extended data.

Response: Thanks for the comment. We have done more work in this new version. We found a candidate gene *VviVI2* that is probably recessive lethal on H2 haplotype of sex determination region. We had modified the texts (Lines 337-359) and Figure 5.

Lines 376-377: If I am not wrong, H1 derived from a recombination between F and M haplotypes. H2 has then derived from a recombination between H1 and F, shortening the

M derived region in the H haplotype.

Response: Thanks for the comment. It is not completely clear how the H haplotype originated. H1 and H2 are the most commonly haplotypes observed in domesticated grapes, and it seems H2 is not derived from H1 according to previous study. The origination of H haplotype didn't affect the results here.

Lines 435-438: Any conclusion and comment on the effects of domestication and improvement on diversity by comparing populations of the wild ancestor against single individual genomes of cultivated varieties makes no sense.

Response: Thanks for the comment. We had removed these sentences here.

Lines 509-534: The entire paragraph "Lessons and implications for grapevine genomic breeding" contains general considerations and personal interpretations that are inappropriate in this context.

Response: Thank you for your suggestion. We had shortened this section to focus on reproductive systems.

Supplementary Table 4. Taxonomic assignment in the column "species" is not accurate.

Response: Thanks for the comment. We corrected it.

The issue with former Figure 5 panel F and Supplementary Figure S15 has been simply removed, and not solved.

Response: Thanks for the comment. As responded in previous response letter. We rearrange the latter part of the results and did new analyses. We used the population specific marker to make a discussion on recombination on the PN40024 genome during selfing, and focused on the heterozygous region on Chr2 which includes SDR regions to explain what we find in this study. We think the new results could be better to explain what we found here than the original one. Thus, we removed the old one.

Reviewer #4 (Remarks to the Author):

This is a genome paper for Pinot noir. The primary data contribution is a phased outbred whole genome assembly of a Pinot Noir plant. The authors present many bioinformatic analyzes of the two chromosome sets (PN1 and PN2) in relation to the "reference genome" – a chromosome level assembly on one highly inbred line of this species (PN40024_T2T) – plus lots of resequencing data from other grapes. The latter is (apparently) Illumina data published previously, although this is not entirely clear from the

paper. I was not a referee on the first submission and was asked to comment on mating system aspects and particularly on the claims (lines 103-104 and elsewhere) about residual heterozygosity in the 9th generation inbred line.

Response: We are grateful for your time and constructive comments for our manuscript. Nine of the clonal Pinot Noir samples were resequenced for this study. Additional samples were publicly available, as mentioned on Lines 589-595 and Supplementary Table 4.

It is plausible for a locus to be heterozygous after 9 generations of selfing. With no selection, the probability is only 0.002, but with preferential preservation of heterozygotes during line formation, the probability increases. For example, if heterozygotes are twice as likely to survive as homozygotes, the probability increases to 0.075, which is still not a large number, but if numerous loci across the genome are experiencing this selection, a few might remain after 9 generations. For the loci on chromosomes 2 and 7 (Fig 5), the authors have the correct data signal – many heterozygous SNPs within a localized 'block' combined with the observation of normal read depths when Illumina short reads are mapped back onto the genome. The authors absolutely need to be much more explicit on one key point: How many heterozygous blocks do they believe are real residual heterozygosity? On lines 484-485, they say "lots of heterozygous regions are real and can't be ignore" – what number corresponds to "lots"? For how many regions can the authors make the picture they have for chroms 2 and 7 (Fig 5a)? The heterotic selection argument (that heterozygote advantage is responsible for residual heterozygosity) becomes less and less plausible as you add more and more loci.

Response: Thanks for the comment. Not only for Chromosome 2 and 7 in Fig. 5a, we showed the specific marker from PN and SG in all chromosomes of HE and PN40024 in Supplementary Figures 14 and 15. We observed many heterozygous sites throughout the PN40024 genome, and most of them were cluster together. As described in methods (Lines 747-750), we detected six large blocks (Supplementary Table 8). By compare the reads coverage between these blocks and whole genome, we removed one blocks that might be the result of erroneous assembly or structural disparities with the reference genome (Supplementary Table 9), and we identified five blocks on Chromosomes 2, 4, 7, 10 and 11 that could be real. That's why we said lots of heterozygous regions are real and can't be ignore. We had modified the sentences in Lines 443-447 to make it clearer.

Of course, the most compelling relevant data would be to directly measure the fitness effects of alternative genotypes. If you generate and genotype a few hundred selfed progeny from the line, differential fitness should be estimable because fitness differences must be large (else the line would not be heterozygous). The GERP stuff (line 364) is fine,

but very indirect. Apparently, information about genotype to phenotype/fitness is available for the SDR locus on Chrom 2 (lines 374 onward). There are no citations here, but it seems that one homozygote (f/f) cannot self (has fitness = 0 during inbred line formation) because it makes only female plants. It is not clear what is going on with the alternative homozygote (H2/H2) – are these plants hermaphrodites or not? These are the essential details to make the residual heterozygosity interesting.

Response: Thanks for the suggestion. We have citation in other places. In the revised manuscript we also added citation here. It was reported that, the f/f genotype plant is female as a recessive male sterile gene located on f haplotype. All the domesticated grapes are hermaphrodites with genotype mainly classified to F/H1, F/H2, and H1/H1. H2/H2 is rarely observed. In this new version, we did more analyses, and provide a candidate gene, *VviVI2*, that is probably recessive lethal located on H2 haplotype that inhibited the genome becoming homozygous H2/H2 as observed in PN40024. (Lines 336-358)

For the broader mating system arguments, I have no doubt that the frequencies of outcrossing/selfing/clonal propagation were important for grape genomes. However, the text requires revision because it does not make sense to say “The mating systems have more significant effects than traditionally focused factors such as genetic drift and natural/artificial selection” (lines 33-35, similar on lines 418-419 and elsewhere). Mating system does not change allele frequencies on its own. It alters the way that genetic drift, gene flow, and natural/artificial selection act on variation. There are numerous good reviews to guide these arguments – see papers by Stephen Wright and references therein.

Response: We appreciate your suggestions. We removed the sentence on Lines 33-35, and rewrote the sentences of Lines 418-419 (Lines 394-396 in the new version).

Other comments:

Line 148: I could not make sense of the arguments on introgression. For most biologists, introgression is an explanation for shared variation between populations or species—often, it is posited as an alternative to shared ancestral variation. Here, we need a much better sense of what the different lineages actually are, and what alleles moved between lineages, and how we can be certain that these variants were not present in the common ancestor of the (now) distinct lineages. Are EU and PN (lines 179-180) sister lineages currently isolated from each other?

Response: Thanks for comment. It is common review that domesticated grape is originated from Middle East and Caucasus region, and the wild grape (*vinifera* subsp.

Sylvestris, here we called ME) in this region is its wild relatives. In this manuscript we tried to detect the introgression signal from European wild grapes (*vinifera* subsp. *Sylvestris*, here we called EU) to domesticated grape cultivar PN. EU and ME are belonging to the same subspecies, there are several papers talk about the introgression between wild grapes EU and domesticated wine grapes (e.g., Magris et al. Nature communications; Freitas et al. Science advances; Xiao et al. PNAS). Thus, in the fd statistic, we used ME as the sister lineage of PN, to verify whether there is a significant higher shared variants between EU and PN compared to that of EU and ME. However, it is also likely that the inferred introgression events occurred prior to the diversificiation of PN clonal lineages. We have modified the sentence to make it clear (Lines 158-172).

Reference:

Magris G. et al. The genomes of 204 Vitis vinifera accessions reveal the origin of European wine grapes. *Nat Commun.* 2021 Dec 21;12(1):7240. doi: 10.1038/s41467-021-27487-y.

Freitas, S. et al. Pervasive hybridization with local wild relatives in Western European grapevine varieties. *Science advances* 7, eabi8584; 10.1126/sciadv.abi8584 (2021).

Xiao, H. et al. Adaptive and maladaptive introgression in grapevine domestication. *Proceedings of the National Academy of Sciences of the United States of America* 120, e2222041120; 10.1073/pnas.2222041120 (2023).

Line 202 onward: The general audience will need an explanation of how the authors are using the term “somatic mutation.” Plants do not have a segregated germline, so all mutations start out as somatic. Are the somatic mutations identified here those that have never been transmitted through gametes? If so, the main paper could use a sentence or two to explain how that is established from sequence data.

Response: We appreciate your suggestions. We have modified these sentences (Lines 178-229).

Lines 273-274: “Then we compared the genetic burden in each group by identifying deleterious SNPs (dSNPs)…” The very important term “deleterious SNPs” comes out of nowhere without any explanation or citation. What does this mean? Does it mean that the SNP has two alleles, one that is good and one that is always bad? If so, how is this determined (the paper has no fitness assays that I can see). There are bioinformatic

classifications of alternative alleles (e.g. synonymous versus non-synonymous) but these do not identify the fitness effects of particular loci. They make groups where the distribution of selection coefficients is likely be different by some (always unknown) amount between groups. This can be useful (e.g. if the fraction of non-syn SNPs is much higher in selfing species than outcrossing), but the reader will need much more explanation to make any sense of the downstream analyses on 'dSNPs'.

Response: Thanks for your comment. We used SIFT (Vaser et.al) to detect the deleterious SNPs. This software is wildly used to evaluate the effect of derived variant. We have modified the sentence to make it clear (Lines 255-257).

Reference:

Vaser, R., Adusumalli, S., Leng, S. N., Sikic, M. & Ng, P. C. SIFT missense predictions for genomes. Nature protocols 11, 1–9; 10.1038/nprot.2015.123 (2016).

Lines 321-323: "… after nine generations of selfing to generate PN40024, most of the genome was homozygous, except for a 59,944 heterozygous SNPs."
What fraction of 60k SNPs are within the heterozygous blocks that the authors believe to be residual heterozygosity? After only 9 generations, real het blocks should be quite large (e.g. Chrom 2 and 7 in Fig 5). After excluding all the SNPs that are false hets due to mapping, how many loci are there in this inbred line? Perhaps the authors just need to explain the results in Supplementary Table 10.

Response: Thanks for your comment. We detected six large blocks by the methods described in methods (Lines 679-682). And only one of them probably false heterozygous blocks. The remained five heterogynous blocks occupied ~4.3% of the genome. After excluding the false heterozygous block, ~93% (SNPs) and ~90% (SVs) heterozygous sites are in the heterozygous blocks. We have modified the manuscript to make it clear (Line 312-315).

Reviewer #1 (Remarks to the Author):

I have no more comments.

Response: Thank you very much for your time, patience, and constructive comments! The manuscript has been greatly improved after revision following your suggestions.

Reviewer #2 (Remarks to the Author):

In the newest version, the addition of an extra focus in one individual on the effect of experimentally-induced multiple rounds of selfing on the flower sex locus provides substantially novel and relevant biological information in comparison to the rest of the manuscript. I appreciated this extended part.

Response: We appreciate your time and constructive comments on our manuscript. We had revised the manuscript and we hope the current version is more suitable for publication.

The other methodological concerns that I expressed in my previous rounds of reviewing remain unaddressed in this new version. In particular, those regarding the issues related to the identification of reliable sequence variation among clonal lines, to the identification of introgressed regions in Pinot Noir, and to the consequences brought about by selfing.

Response: Thanks for your comments. In this new version, we have revised these sections to enhance clarity and rigor. 1) Regarding introgression, the fd-test examines shared alleles rather than allele frequencies. Thus, it is an efficient way to detect the fragments in clonal PN that originated from EU wild grapes. Since these signals have not been detected in our previous population genetic study on introgression and grapevine domestication (Xiao et al. 2023, PNAS). It is likely that this introgression

occurred during the formation of Pinot Noir or in its earlier ancestors. The results we present here reflect the influence of European wild grapes on Pinot Noir. We have revised the manuscript to address and clarify this point (Lines 172-176).

We have also revised the manuscript regarding 2) the identification of sequence variation among clonal lines (Lines 421-428); and 3) the consequences of selfing (Lines 496-499). Please also see the responses in below.

Reference:

Xiao, H. et al. Adaptive and maladaptive introgression in grapevine domestication. Proceedings of the National Academy of Sciences of the United States of America 120, e2222041120; 10.1073/pnas.2222041120 (2023).

As to the entire parts of the manuscript regarding the impacts of vegetative propagation on the genetic composition of crop germplasm, all of this is elaborated upon a critical set of variant sites. As explained by the authors in one point of their rebuttal letter, it is commonplace to acknowledge that from 1% to 5% SNP calls in reference-based analyses using short-read alignments are likely to be false variant sites. If they called 3,915,920 SNPs in their analysis of different grapevine genotypes (Line 176), it follows that they consider a reasonable number between 39,000 and 195,000 variant sites to be possible errors. If they had analysed 18 clonal lines that are genetically identical with no sequence variation, under the same experimental and analytical conditions as above, in the absence of sequence variation in the biological samples they would have therefore expected to call a similar absolute number of false variant sites due to the experimental procedure alone, which are to be considered possible errors. Among 18 Pinot Noir clones that might have biological sequence variation, the authors called a number of variant sites comprised between 20,328 and 70,291, which largely overlap with the absolute numbers of false variant sites that may result from the experimental procedure alone, as above. This situation is nearly identical for SVs. I do not believe that all of these variant sites among Pinot Noir clonal lineages are false, but in the absence of an

experimental validation and in the absence of an independent assessment of false discovery rate in this experiment, I cannot take this dataset on trust either. While up to 200 K false SNPs out of 4 M true SNPs have no significant impacts on the results of population genetics analyses, the absolute number of called SNPs among clonal lineages compared to the expected absolute number of possible errors might lead to what the authors called in the point I mentioned above “horrible incorrect studies” or conclusions.

Response: We appreciate the points you've raised concerning the interpretation of variant sites. We do agree with you that there is a problem in genotyping in resequencing data, especially for those samples with highly heterozygous genomes, such as grapevine. However, in our study, we used the most updated sequencing and bioinformatic methods. First, we used the complete T2T reference genome PN40024_T2T for read mapping; Second, we used high coverage whole genome sequencing (WGS) data from PN clones (⩾20X, for nine samples ⩾40X); Third, the original VCF file was filtered using strict criteria to remove low quality variant sites; Fourth, we applied the comparative methods to detect unique variants in PN clones. This is much better than previous studies (Monroe et al. 2022, Nature; Roach et al. 2018, PLoS Genetics; Vondras et al., 2019, BMC genomics) that mixed somatic and germline mutations. Fifth, our focus on the overall characteristics of somatic variations rather than on specific variant sites. As the reviewer noted, these errors would not have significant impacts on the results of population genetics analyses. When specifically considering a particular somatic variation, we believe it is necessary to conduct further validations due to the presence of errors, but this is beyond the scope of this study. We added these issues to the discussion part (Lines 421-428).

Reference:

Monroe, J.G. et al. Mutation bias reflects natural selection in Arabidopsis thaliana. Nature 602, 101–105 (2022)

Roach MJ et al. Pretorius IS et al. Population sequencing reveals clonal diversity and

ancestral inbreeding in the grapevine cultivar Chardonnay. PLoS Genetics 14(11):1-24 (2018).

Vondras, A. M. et al. The genomic diversification of grapevine clones. BMC genomics 20, 972; 10.1186/s12864-019-6211-2 (2019)

As to the entire parts of the manuscript regarding the impacts of selfing on the genetic composition of crop germplasm, I do not have doubts on the data extracted from the genome composition of the PN40024 line. However, part of the observations of the characteristics of the genome in a single individual might be due to the effect of chance alone and drawing generalizations on the consequences of selfing at population-wise level is unsound.

Response: Thank you for your insightful comments. We understand your concerns regarding the observations made from a single individual. It is the only sample observed in grapevine, however, several studies have been done in other species, and they have presented similar results, such as in potato and maize, which we discussed in our study. In our revised manuscript, we have compared our observations to those in potato and maize, reflecting a general pattern in genetics associated with the transitions of reproductive systems. We have revised the manuscript to address and clarify this point (Lines 496-499).

Reference:

Roessler, K. et al. The genome-wide dynamics of purging during selfing in maize. Nature plants 5, 980–990; 10.1038/s41477-019-0508-7 (2019).

Zhang, C. et al. The genetic basis of inbreeding depression in potato. Nature genetics 51, 374–378; 10.1038/s41588-018-0319-1 (2019).

Zhang, C. et al. Genome design of hybrid potato. Cell 184, 3873-3883.e12; 10.1016/j.cell.2021.06.006 (2021)

Reviewer #4 (Remarks to the Author):

The authors have improved their paper in response to my previous comments. My review focused on the issue of residual heterozygosity within a putatively highly inbred line. They have resolved the real heterozygosity to five loci that cover about 4% of the genome. Given strong selection against homozygotes in certain portions of the genome, this number is reasonable even after 9 generations of selfing. The authors did not collect additional data to explore the nature of heterotic selection, but that may be beyond the scope of this work. The overall narrative still needs a great deal of work, but I am satisfied with their treatment of the heterozygosity issue.

Response: Thank you for your thoughtful review and for acknowledging the improvements we made in response to your previous comments. We acknowledge your feedback about the overall narrative and we have refined it to enhance clarity and cohesion in our presentation. Thank you once again for your constructive feedback.

We are grateful for editor and all reviewers' constructive comments. We also acknowledge Reviewers #1 and #3 for their approval of our revisions. Following we are readdressing Reviewer #2' comments.

Reviewer #2 Comments on manuscript:

Regarding the specific question I am asked, whether my remaining concerns have been addressed in the last round of revision of the manuscript, the answer is no.

Since the reviewer was not satisfied with our previous revision, we have revised the manuscript again in accordance with your comments on manuscript NCOMMS-23-52462B:

The other methodological concerns that I expressed in my previous rounds of reviewing remain unaddressed in this new version. In particular, those regarding the issues related to the identification of reliable sequence variation among clonal lines, to the identification of introgressed regions in Pinot Noir, and to the consequences brought about by selfing.

Response: Thanks for your comments. In this new version, we did more analyses and revised the related sections to enhance clarity and rigor.

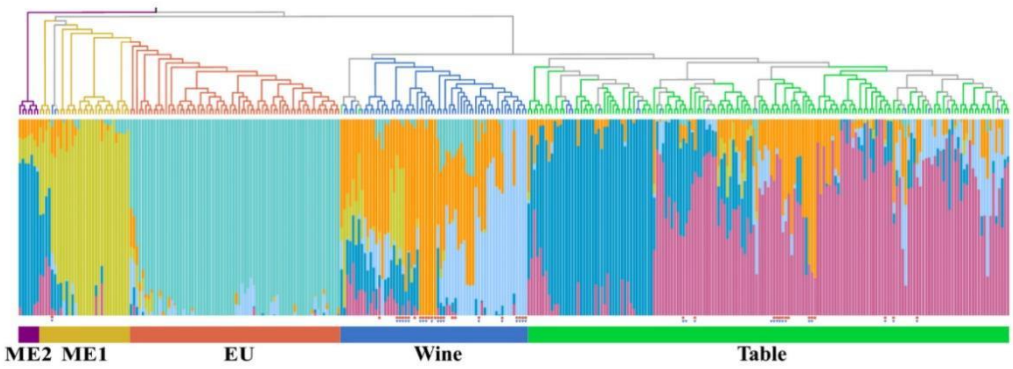
Regarding introgression, we used fd statistic to detect the introgression. According to the article (Simon et al. 2014, Mol Biol Evol), the fd is calculated using the equation:

$$f_d = \frac{S(P_1, P_2, P_3, O)}{S(P_1, P_D, P_D, O)}$$

In this equation, there are four groups P_1, P_2, and P_3 and outgroup O, with the relationship (((P1,P2),P3),O). P_D is the population (either P_2 or P_3) that has the higher frequency of the derived allele The fd is to detect the introgression between P_2 and P_3. S indicted the difference between sums of ABBAs and BABAs. ABBAs are sites at which the derived allele B is shared between the nonsister taxa P_2 and P_3, whereas P_1 carries the ancestral allele, as defined by the outgroup. Similarly, BABAs are sites at which the derived allele is shared between P_1 and P_3, whereas P_2 carries the ancestral

allele. From this equation, we can see that it focuses on shared alleles rather than allele frequencies. In brief, it checked whether there are excessive shared alleles between P_2 and P_3 than between P_1 and P_2. Although the fd-test is primarily used in sexual populations, we think it is an efficient way to detect the fragments in clonal PN that originated from EU wild grapes, as it assesses the shared alleles between PN and EU within the observed window, rather than the allele frequencies in each group. We also added these sentences in Lines 166-170 to explain to readers why we used the fd statistic. To validate the introgression region we identified, we generated a phylogenetic tree based on the gene *COL5* for wild grapes and Pinot Noir, including an additional 41 wine grape varieties. As shown in Figure for Review 1B, in the putative introgressed region, the wild grapes ME and EU split into two clades; however, unlike the tree constructed from whole genome SNPs (Figure for review 1A), the wine grapes do not cluster together to form an independent clade. Some of them, including Pinot Noir, are located within the EU clade. This indicates that the genomes of some grapevines, including Pinot Noir, were introgressed by EU wild grapes in this region. We have modified the related content accordingly (Lines 179-184, Supplementary Figure 7)

A



B

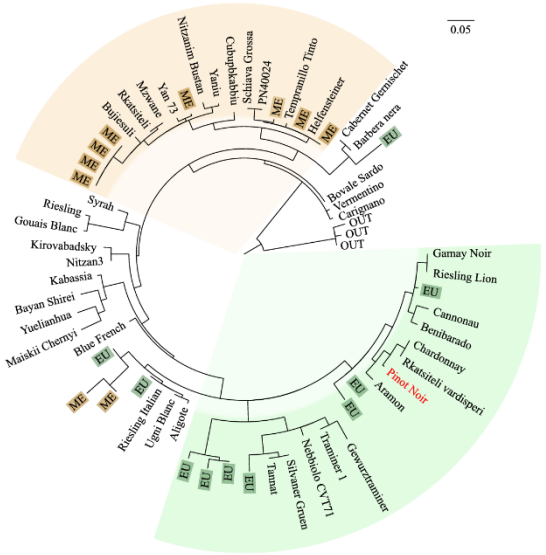


Figure for review 1: (A) Phylogenetic tree of 339 grapes based on whole genome SNPs (Xiao et al., PNAS, 2023); (B) Phylogenetic tree of 20 kb region where *COL5* is located.

We have also revised the manuscript regarding 2) the identification of sequence variation among clonal lines; and 3) the consequences of selfing. Please also see the responses in below.

Reference:

Xiao, H. et al. Adaptive and maladaptive introgression in grapevine domestication. Proceedings of the National Academy of Sciences of the United States of America 120, e2222041120; 10.1073/pnas.2222041120 (2023).

Martin, S. H., Davey, J. W. & Jiggins, C. D. Evaluating the use of ABBA-BABA statistics to locate introgressed loci. Molecular biology and evolution 32, 244–257;

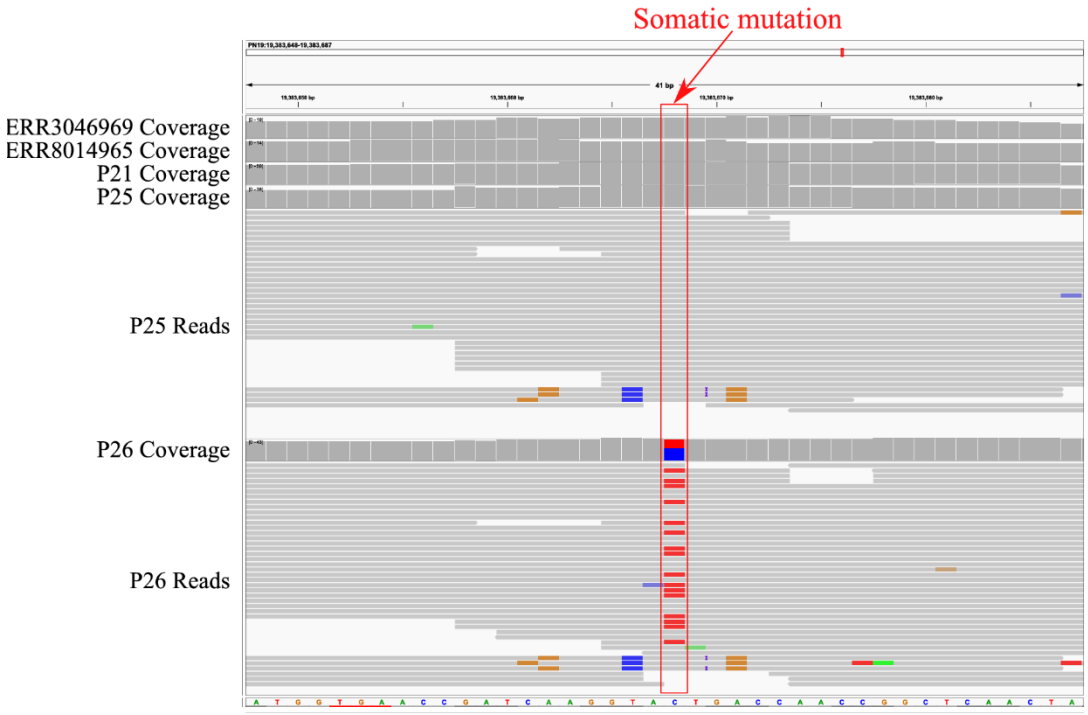
10.1093/molbev/msu269 (2015)

As to the entire parts of the manuscript regarding the impacts of vegetative propagation on the genetic composition of crop germplasm, all of this is elaborated upon a critical set of variant sites. As explained by the authors in one point of their rebuttal letter, it is commonplace to acknowledge that from 1% to 5% SNP calls in reference-based analyses using short-read alignments are likely to be false variant sites. If they called 3,915,920 SNPs in their analysis of different grapevine genotypes (Line 176), it follows that they consider a reasonable number between 39,000 and 195,000 variant sites to be possible errors. If they had analysed 18 clonal lines that are genetically identical with no sequence variation, under the same experimental and analytical conditions as above, in the absence of sequence variation in the biological samples they would have therefore expected to call a similar absolute number of false variant sites due to the experimental procedure alone, which are to be considered possible errors. Among 18 Pinot Noir clones that might have biological sequence variation, the authors called a number of variant sites comprised between 20,328 and 70,291, which largely overlap with the absolute numbers of false variant sites that may result from the experimental procedure alone, as above. This situation is nearly identical for SVs. I do not believe that all these variant sites among Pinot Noir clonal lineages are false, but in the absence of an experimental validation and in the absence of an independent assessment of false discovery rate in this experiment, I cannot take this dataset on trust either. While up to 200 K false SNPs out of 4 M true SNPs have no significant impacts on the results of population genetics analyses, the absolute number of called SNPs among clonal lineages compared to the expected absolute number of possible errors might lead to what the authors called in the point I mentioned above “horrible incorrect studies” or conclusions.

Response: First of all, the methods we are using are standard genomic tools widely used in detecting somatic mutations in plants (Monroe et al., 2022, Nature; Roach et al., 2018, PLoS Genetics; Vondras et al., 2019, BMC genomics). Based on previous studies, we used deep sequencing data (>30X) and comparative population genomic methods,

which is more reliable than previous studies. With 30X coverage, we believe most of our genotypes are correct. At the same time, this study focused on general genome-wide patterns of somatic mutations instead of specific mutation, so that the minimal potential errors don't affect the conclusions in this study. To show the reliability of our data, we presented ten random somatic mutation sites by examining the reads in IGV (Lines 214-217, Supplementary Figure 9). As shown in Figure for Review 2, the reads with predicted somatic mutations are observed exclusively in one sample and not in any of the other samples. These reads were generated through PCR prior to sequencing and subsequently sequenced using second-generation sequencing. These results indicate that most of the identified somatic mutations are reliable.

Figures for review 2. IGV visualization to view identified somatic mutations. The



coverage of reads that match and differ from the reference genome of five PN clones is shown, along with detailed reads for one sample without somatic mutation and one sample with somatic mutation. We included the IGV figures in supplementary Figure S9.

Reference:

Monroe, J.G. et al. Mutation bias reflects natural selection in Arabidopsis thaliana. Nature 602, 101–105 (2022)

Roach MJ et al. Pretorius IS et al. Population sequencing reveals clonal diversity and ancestral inbreeding in the grapevine cultivar Chardonnay. PLoS Genetics 14(11):1-24 (2018).

Vondras, A. M. et al. The genomic diversification of grapevine clones. BMC genomics 20, 972; 10.1186/s12864-019-6211-2 (2019)

As to the entire parts of the manuscript regarding the impacts of selfing on the genetic composition of crop germplasm, I do not have doubts on the data extracted from the genome composition of the PN40024 line. However, part of the observations of the characteristics of the genome in a single individual might be due to the effect of chance alone and drawing generalizations on the consequences of selfing at population-wise level is unsound.

Response: Although the consequences are only based on PN40024, several studies have been done in other crops, and they have presented similar results, such as in potato (Zhang, C. et al., 2021, Cell) and maize (Roessler, K. et al.,2019, Nature Plants), which we discussed in our study. We have modified the related content to limit the results and discussion in PN40024 (Line 506-512).

Reference:

Roessler, K. et al. The genome-wide dynamics of purging during selfing in maize. Nature plants 5, 980–990; 10.1038/s41477-019-0508-7 (2019).

Zhang, C. et al. Genome design of hybrid potato. Cell 184, 3873-3883.e12; 10.1016/j.cell.2021.06.006 (2021)

We are grateful for editor and all reviewers' constructive comments and their approval of our revisions.

Reviewer #2 (Remarks to the Author):

The manuscript has been improved to an adequate degree. The results and discussion sections include sufficient explanations about the limitations of this study.

Response: Thank you very much for your time, patience, and constructive comments! The manuscript has been greatly improved after revision following your suggestions.