nature research

Peer Review Information

Journal: Nature Methods
Manuscript Title: Jasmine and Iris: Population-scale structural variant comparison and analysis
Corresponding author name(s): Michael Schatz

Editorial Notes:

Reviewer Comments & Decisions:

Decision Letter, initial version:

Date: 30th Jun 21 22:09:29
Last Sent: 30th Jun 21 22:09:29
Triggered By: Lin Tang
From: Lin.tang@nature.com
To: mschatz@cs.jhu.edu
CC: methods@us.nature.com;ziqian.li@nature.com
Subject: Decision on Nature Methods submission NMETH-A46161
Message: 30th Jun 2021

Dear Dr. Schatz,

Your Article, "Jasmine: Population-scale structural variant comparison and analysis", has now been seen by 2 reviewers. As you will see from their comments below, although the reviewers find your work of potential interest, they have raised a number of concerns. We are interested in the possibility of publishing your paper in Nature Methods, but would like to consider your response to these concerns before we reach a final decision on publication.

We therefore invite you to revise your manuscript to fully address all the concerns raised by our reviewers with additional analyses and other revisions.

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

When revising your paper:

* include a point-by-point response to the reviewers and to any editorial suggestions

* please underline/highlight any additions to the text or areas with other significant changes to facilitate review of the revised manuscript

* address the points listed described below to conform to our open science requirements

* ensure it complies with our general format requirements as set out in our guide to authors at www.nature.com/naturemethods

* resubmit all the necessary files electronically by using the link below to access your home page

[REDACTED]

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

We hope to receive your revised paper within eight weeks. We are very aware of the difficulties caused by the COVID-19 pandemic to the community. If you cannot send it within this time, please let us know. In this event, we will still be happy to reconsider your paper at a later date so long as nothing similar has been accepted for publication at Nature Methods or published elsewhere.

OPEN SCIENCE REQUIREMENTS

REPORTING SUMMARY AND EDITORIAL POLICY CHECKLISTS
When revising your manuscript, please update your reporting summary and editorial policy checklists.

Reporting summary: https://www.nature.com/documents/nr-reporting-summary.zip
Editorial policy checklist: https://www.nature.com/documents/nr-editorial-policy-checklist.zip

If your paper includes custom software, we also ask you to complete a supplemental reporting summary.

Software supplement: https://www.nature.com/documents/nr-software-policy.pdf

Please submit these with your revised manuscript. They will be available to reviewers to aid in their evaluation if the paper is re-reviewed. If you have any questions about the checklist, please see http://www.nature.com/authors/policies/availability.html or contact me.

Please note that these forms are dynamic 'smart pdfs' and must therefore be downloaded and completed in Adobe Reader. We will then flatten them for ease of use by the reviewers. If you would like to reference the guidance text as you complete the template, please access these flattened versions at http://www.nature.com/authors/policies/availability.html.

DATA AVAILABILITY
We strongly encourage you to deposit all new data associated with the paper in a persistent repository where they can be freely and enduringly accessed. We recommend submitting the data to discipline-specific and community-recognized repositories; a list of repositories is provided here: http://www.nature.com/sdata/policies/repositories

All novel DNA and RNA sequencing data, protein sequences, genetic polymorphisms, linked genotype and phenotype data, gene expression data, macromolecular structures, and proteomics data must be deposited in a publicly accessible database, and accession codes and associated hyperlinks must be provided in the "Data Availability" section.

Refer to our data policies here: https://www.nature.com/nature-research/editorial-policies/reporting-standards#availability-of-data

To further increase transparency, we encourage you to provide, in tabular form, the data underlying the graphical representations used in your figures. This is in addition to our data-deposition policy for specific types of experiments and large datasets. For readers, the source data will be made accessible directly from the figure legend. Spreadsheets can be submitted in .xls, .xlsx or .csv formats. Only one (1) file per figure is permitted: thus if there is a multi-paneled figure the source data for each panel should be clearly labeled in the csv/Excel file; alternately the data for a figure can be included in multiple, clearly labeled sheets in an Excel file. File sizes of up to 30 MB are permitted. When submitting source data files with your manuscript please select the Source Data file type and use the Title field in the File Description tab to indicate which figure the source data pertains to.

Please include a "Data availability" subsection in the Online Methods. This section should inform readers about the availability of the data used to support the conclusions of your study, including accession codes to public repositories, references to source data that may be published alongside the paper, unique identifiers such as URLs to data repository entries, or data set DOIs, and any other statement about data availability. At a minimum, you should include the following statement: "The data that support the findings of this study are available from the corresponding author upon request", describing which data is available upon request and mentioning any restrictions on availability. If DOIs are provided, please include these in the Reference list (authors, title, publisher (repository name), identifier, year). For more guidance on how to write this section please see: http://www.nature.com/authors/policies/data/data-availability-statements-data-citations.pdf

CODE AVAILABILITY
Please include a "Code Availability" subsection in the Online Methods which details

how your custom code is made available. Only in rare cases (where code is not central to the main conclusions of the paper) is the statement "available upon request" allowed (and reasons should be specified).

We request that you deposit code in a DOI-minting repository such as Zenodo, Gigantum or Code Ocean and cite the DOI in the Reference list. We also request that you use code versioning and provide a license.

For more information on our code sharing policy and requirements, please see: https://www.nature.com/nature-research/editorial-policies/reporting-standards#availability-of-computer-code

MATERIALS AVAILABILITY
As a condition of publication in Nature Methods, authors are required to make unique materials promptly available to others without undue qualifications.

Authors reporting new chemical compounds must provide chemical structure, synthesis and characterization details. Authors reporting mutant strains and cell lines are strongly encouraged to use established public repositories.

More details about our materials availability policy can be found at https://www.nature.com/nature-portfolio/editorial-policies/reporting-standards#availability-of-materials

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

Please do not hesitate to contact me if you have any questions or would like to discuss these revisions further. We look forward to seeing the revised manuscript and thank you for the opportunity to consider your work.

Sincerely,

Lin

Lin Tang, PhD
Senior Editor
Nature Methods

Reviewers' Comments:

Reviewer #1:
Remarks to the Author:
The authors have developed an SV merging tool Jasmine, alongside a long-read SV refining pipeline Iris, in order to perform accurate SV comparison and merging across studies, and at population scale. They present an evaluation of their method which focuses on the reduction of discordant SVs in parent-offspring analyses compared to a number of other approaches, alongside the characteristics of SV calling experiment results on three long-read technologies, i.e. ONT, PacBio CLR, and PacBio HiFi. Finally, they genotype a merged SV set generated using 31 long-read sequenced individuals of diverse ancestry, on 444 short-read sequenced individuals using Paragraph, a third-party genotyper, to investigate SV-gene expression associations.

Main comments:
1. Iris is a pipeline that consists of the application of Sniffles on a BAM file, using the alternate allele supporting reads given by Sniffles, running Racon to polish the inserted sequence or the refining of the breakpoints, aligning it back to the reference genome using minimap2 and parsing the CIGAR alignment string to give the updated SV. Is this correct?
2. Regarding the sentence: "Jasmine improves upon other SV merging methods by representing variants as points in space based on their breakpoints and lengths and constructing a graph of SV proximity, where edges represent pairs of SVs with a small Euclidean distance between them": This is an almost identical representation to what is done in SV merging in https://www.nature.com/articles/s41588-021-00865-4 . This approach, however, is presented in this manuscript as a key strategy differing from all other SV merging approaches.
3. Besides the similar representation of SVs and the distances between them, the application of using a minimum spanning tree using a modified version of Kruskal's algorithm also appears to be analogous in purpose to the approach employed in the study by (Beyter et al., 2021) which uses a clique finding heuristic, although is a different algorithm, designed to answer a different question. It would be useful to point out the benefits of the proposed approach.
4. The authors compare their approach to five existing methods listed as (Shi et al. 2021, Jeffares et al. 2017, Ebert et al. 2021, Larson et al. 2019, and Beyter et al. 2021), however it is unclear which tool name belongs to which work in this comparison. For example, the tool "svimmer" shown in Figure 2F is a tool used in https://www.nature.com/articles/s41467-019-13341-9 (Eggertsson et al. 2019).
5. It is also unclear whether the SV merging approach employed in the cited study (Beyter et al. 2021), which uses a similar graph based SV merging approach, is compared against or not.
6. Besides regions where the mapping of long-reads are ambiguous, a potentially challenging yet frequent task would be merging SVs within tandem repeat regions due to their abundance. It is true that the application of Iris can be beneficial in such regions, but it would be informative to evaluate performance within and outside of such more difficult regions to highlight where the challenge in SV merging lies.
7. Although the study aims to perform a population-scale SV comparison and analysis, the study uses at most 31 long-read samples. While this is understandable due to the lack of larger publicly available long-read datasets, typically population-scale studies can involve hundreds or thousands of individuals. It could be useful to have insights on how Jasmine will perform in the existence of such large scale data, perhaps using simulated datasets and SVs.

8. While the finding regarding a deletion in SEMA5A affecting the expression of the gene is interesting, based on Supp Fig. 22, the deletions appears to be around 36 basepairs. SV studies have thus far used a variant size of 50 basepairs or more. As a result previous SV studies or databases may lack this variant simply due to not satisfying a size cutoff threshold rather than previous methodical limitations. The authors also state that this deletion is difficult to detect using short-reads alone, but do not provide a reason or evidence for this claim.
9. Overall, I believe the text for the main manuscript and the number of main figures/panels used can be significantly reduced and more emphasis on the method, highlighting its advantages, can be included.

Minor comments:
10. What do you mean by "harmonized" callset in the abstract? And how many of the 205,192 SVs would be of higher confidence/quality?
11. In the sentence: "... in complex regions to be accurately identified by short reads, which make up the majority of existing genomic sequencing data". Please see https://www.sciencedirect.com/science/article/abs/pii/S0002929721000987 , which mentions that about 9,7% of the current GRCh38 reference is defined by segmental duplications (SD) and simple repeats (SR). While it is true that the majority of SVs detected using long-reads according to the study given above are from long-reads alone (25,000 lrWGS SVs vs. 11,000 srWGS SVs), I disagree that the majority of the human DNA is composed of complex regions -- if that is what the authors express.
12. Why did the authors decide to include variants between size 20-50 (Fig 2D) or 30-50 (the entirety of the manuscript) as SVs given that SVs have typically focused on variants larger than or equal to 50 nucleotides in size (Chaisson et al. 2019, Audano et al. 2019, gnomAD-SV, … etc.)
13. The sentence: "Jasmine avoids merging variants of the same type which correspond to unique breakpoint adjacencies, which is particularly important when resolving complex nested SVs" is unclear. Although the Supp Fig 21 elaborates, the problem is yet difficult to understand.
14. Although Jasmine successfully decreases the denovo candidate SVs from 2194 SVs down to 404 SVs in the analyzed sample, this number is still impractically high to perform manual analyses (less than 10 per generation on average), particularly in the presence of multiple parent-offspring trios. As a result, detecting denovo SVs may be considered a different problem than SV merging.
15. In the results presented in Fig. 2, it would be beneficial to know what number of discordant SVs are filtered by which approach, as several filters are applied to called SVs before the application of Jasmine. A table showing the removed number (and percentage given total number of SVs) of discordant SVs may be clearer to understand. For example, how many of the 2194 discordant SVs can be removed using strict yet simple filters?
16. Please clarify the explanation and interpretation of Figure 2F as I find it to be the most important panel in the figure. For example, what does it mean "... but corresponding to unique breakpoint adjacencies (mixed strand)" ?
17. I cannot see the increased number of variants around size 6-7 kb in any of the provided plots.
18. There exists, however, a peak around 500-750 in almost all SV size plots in the paper, and no explanation is given about it. This peak does not exist in other SV datasets such as (Audano et al. 2019, Beyter et al. 2021) which uses long-reads or

gnomAD-SV which uses short-reads. Is this an artifact, or is this expected?
19. The behavior of Sniffles shown in Fig 2E is strange as a seemingly more restrictive parameter provides an increase in the total number of variants discovered (while reducing discordant SVs). Do you think this is simply a mis-adjusted default parameter in Sniffles?
20. In the text discussing Fig. 3, no explanations made for HiFi-only SVs, whereas similar explanations are available for ONT-only or CLR-only SVs. This main figure can be a supplementary figure instead.
21. While the usage of three different sequencing technologies/data naturally provides a more manageable (for manual examination) list of denovo SVs, performing a 3-fold increased sequencing at scale may not be a realistic endeavour for the detection of denovo SVs.
22. As the authors manually examine the set of denovo SV candidates for plausibility, they mainly check for the reliability of mapping of the long reads, which can be done programmatically as well. Again, this may be considered a different problem than what Jasmine is designed for.
23. I do not find panel Fig 4B useful. What is the message/purpose intended in this panel?
24. Again, I believe it is important to have an insight on the percent of denovo SV candidates that can be filtered with simple albeit strict filters, and the cases where Jasmine has a clear advantage with the algorithm it employs.
25. In Figure 5B, it may be more useful to give the breakpoint range as percentage of length of the final merged SV, rather than the actual breakpoint range. I also do not see much benefit or extra information in the panels 5D through 5G, they could reside in supplementary if desired. What is shown in panel 5D x-axis is rather "number of carriers" rather than allele frequency.
26. While Jasmine does not merge SVs in an intra-sample method, it may yet be important to do so in the presence of multiple SV discovery algorithms applied to a single individual.
27. Again, the 6-7kb peak in Figure 5F is not present as mentioned.
28. A similar analysis of genotyping SVs detected using long reads in short read samples have also been done in https://www.biorxiv.org/content/10.1101/2020.12.04.412486v2 using Giraffe (Siren et al., 2021), a recent software. Was there a particular reason for the use of Paragraph in genotyping short reads?
29. Again, a similar SV-eQTL association analysis has been done by Siren et al. (https://www.biorxiv.org/content/10.1101/2020.12.04.412486v2) Have you compared your eQTL association results to those presented there? What would constitute a novelty in your results?
30. The discussion section is too verbose and can be shortened significantly, particularly the second paragraph.
31. It is a good idea to polish the insertion sequences and refine the breakpoints using Racon, but other methods such as performing a multiple sequence alignment or a haplotype resolved assembly can also achieve useful results for the task of refining variants/sequences. Were there particular reasons for the employment of Racon within Iris?
32. To my understanding the Jasmine pipeline uses polished SVs as input. As one can also use it without such SV polishing, it would be interesting to see the SV merging results without the SV polishing to establish the effect of the Iris pipeline into Jasmine.
33. In discordant SV analysis, while the merging goes through using all called SVs,

what is the reason for applying various filters when assessing the total number of variants? Filtering erroneous SV calls can also reduce an erroneous SV merging.
34. In the section regarding Double Thresholding, I do not see the motivation in properly detecting/discovering a variant in all of the samples in which they are present. While it is true that a variant may be confidently discovered in only a subset of the samples, it can be genotyped in all carriers, with appropriate genotyping likelihoods, i.e. uncertainties, which is more suitable for downstream approaches such as imputation, association, … etc. SVjedi (https://github.com/llecompte/SVJedi), and LRcaller (https://github.com/DecodeGenetics/LRcaller) are example tools for long-read SV genotyping.
35. The authors indicate that out of the 205K merged SVs, 138K of them were left unfiltered. It would be beneficial to compare this list to other publicly available SV lists such as gnomAD-SV to assess the benefit of genotyping SVs detected using long-read sequencing samples on short-read sequencing samples, over using short-reads alone. Using lenient comparison approaches (due to the application of different sequencing technologies) on variants with a size of at least 50 bp (in order to have comparable size thresholds) will be the most meaningful.

Reviewer #2:
Remarks to the Author:
Kirsche et al. share two algorithms to improve SV calling accuracy and merging, Iris for refining SV breakpoints, and Jasmine for merging SVs. Iris polishes SV sequences with Racon to include all reads in the SV representation. The paper focuses on Jasmine, a merging algorithm using a novel algorithm framed as a spanning tree problem to link SVs across samples and merge intra-sample variants. With optimizations, such as double thresholding, they are able to produce a merge less susceptible to noise than other merging approaches. The authors go on to show that Mendelian error is reduced. Using Jasmine, they merge 31 SV callsets and show that the callset characteristics follow previous results and harbors potentially functional variants, such as an SV in SEMA5.

Methods to properly address SV merging is lacking. In past years, reciprocal overlap was mostly adequate for short-read SVs, but the overlap approach under-performs in modern SV callsets. New methods are needed, and Jasmine is potentially an important advancement. Large consortia, such as the HGSVC and HPRC, are generating long-read callsets, and a growing number of smaller projects are as well. The presentation of Jasmine is timely.

I had no trouble installing and running Jasmine through Conda. I did run a comparison of Jasmine with default parameters against SV-Pop (merging parameters "nr:szro=50:offset=200", same used for HGSVC) against HG00733 (PBSV, HiFi). I found Mendelian errors of 700 and 710 for Jasmine and SV-Pop outside centromeres, respectively (116 and 125 outside all tandem repeats). Although these are not quite the large gains I was hoping to see from Jasmine compared to SV-Pop, I still think it's a valuable approach with more potential than current methods.

I have several major concerns with this manuscript:

1) SV sizes and counts are not adding up through the manuscript, and it's unclear what your size threshold for defining an SV is. Most commonly, it's 50 bp., but I believe you are using 30 bp (from online methods). SV counts will differ with most other published callsets that define SVs at 50 bp, which makes it hard to compare. Fig 2C suggests 30 bp, Fig 2D suggests 20 bp. In Fig 1A, there are 26,591 SVs in the child (14257 + 5260 + 4880 + 2194), which is what I would expect if the SVs are >= 50 bp. However, compared to the abstract, 205,192 merged from 31 samples makes no sense when Ebert 2021 just published half as many for the same number of samples. Is the SV size definition consistent throughout the paper, or is there another explanation for these differences? Please clarify and make it clear at the beginning what your SV size is.

2) A large fraction of callset and merging errors occur in tandem repeats because aligners and callers do not get them right no matter how long or high-quality reads and contigs are. I think it's important to report key findings on all variants (post-QC) as well as variants outside tandem repeats (with some supporting figures at least in supplement if possible).

3) Lastly, HPRC data used for population studies is publicly available but unpublished. You must check with the PIs before publishing it and making sure data and assembly production receives due credit.

Minor comments:

* Italicize gene names (SEMA5A).

* "While discovering significantly more SVs (Figure 2C)." Significantly more SVs than what, and where's the test of significance?

* Fig 2A & 2B. "number of samples called" (in legend). Do you mean the number of SVs called? The title says "Mendelian Discordance", but that is only the last column.

* Figure 2C suggests a high false-call rate. Previous long-read papers report around 15k insertions and 10k deletions (50 bp and greater), but Fig 2C shows 16k insertions and 14k deletions in the same size (seems too small if SVs are 30+ bp).

* I am not sure why the numbers between 2B and 2C don't match. If I add up the bins in 2B that include the son (36k), it doesn't match 2C with (46k) or without (31k) the 30-50 bp size range. What accounts for this difference?

* Fig 2C is cited in the text as "discovering significantly more SVs", but 2C is just a size distribution not a comparison, so it doesn't support this claim.

* Fig 2D: Double-thresholding is not explained at all before citing Fig 2D. leaving the reader wondering what it is. Provide a brief explanation and cite methods for details. It's also not clear to me what "rescued from absence" means, if you are merging calls, no call should be absent. In the methods, please provide an explanation how SVs are assigned to each category that appears in the legend.

* Fig 2E: Again, please make it clear in the manuscript what SV size threshold is used. It's difficult to interpret the middle panel without knowing. A brief explanation for those less with Sniffles might help if you are using it to illustrate a main point. This should probably be a supplementary figure.

* Fig 2E: "The effects of the sniffles m ax_dist". Capitalize "Sniffles".

* Fig 2F: Explain "unique breakpoint adjacencies (mixed strand)" in the text, it only makes sense after looking at Sup Fig 21, but it's part of a main figure. Is this only a Sniffles anomaly, as Sup Fig 21 suggests? If so, that should probably also be clarified. Text citation is misplaced, should also be moved to the next sentence where it is introduced.

* "with a slight enrichment of insertions, shown in previous studies to be caused by missing sequence in the human reference genome." This accounts for only a small fraction, there are probably just more insertions because deletions are more often deleterious (Cooper 2011, Nat Gen).

* "There is also an increased number of variants around sizes of 300bp and 6-7kbp, corresponding to SINE and LINE elements respectively." The 6 kbp peak isn't shown anywhere because the bins are too large in Fig 3C. You could add a supplementary histogram with smaller bins to show it. These peaks are commonly used as an indication of callset quality, so I think it's worth adding.

* Fig 3D: The TRA bin ended up in "10k+" (no TRA label).

* Supplementary figures 11-13: Add legend (bar color)

* Supplementary figures are referenced out of order.

* In "De Novo Variant Discovery", "a 43-fold reduction in candidates from using prior methods (Figure 2a)", the reduction is part method, but can also be attributed to using three sets of rather expensive sequence data, so it's not all methods as the statement suggests. The wording strikes me as disingenuous.

* Fig 4A caption: "both of the examples in (a) and (b)". Do you mean parts c and d?

* Fig 4B: I don't think this is a main figure. It's hard to see where the points are in 3D space. Consider making it 2D with bubble size representing one of the dimensions or move it all to supplement.

* The HiFi discordance rate doesn't make sense to me and I wonder if it can be attributed to using Sniffles. PBSV was designed for HiFi variant calling, and I suspect the de novo rate would come down.

* In "Population SV Inference", Supplementary figures 17 and 18 should be 18 and 19.

* Supplementary figure 19 could be improved by replacing "Population Allele Frequencies" in the title of every panel with the actual tool name.

* "These data suggest that many of the SVs that are only visible through genotyping long-read-based variant calls have large effects on gene expression and thus are potentially functionally relevant.". True, but it also suggests noise introduced by genotyping.

* Refer to Nature guidelines with revisions, multipart figure labels mix case between figure and legend (OK for review).

Author Rebuttal to Initial comments

Reviewer #1:

Remarks to the Author:

The authors have developed an SV merging tool Jasmine, alongside a long-read SV refining pipeline Iris, in order to perform accurate SV comparison and merging across studies, and at population scale. They present an evaluation of their method which focuses on the reduction of discordant SVs in parent-offspring analyses compared to a number of other approaches, alongside the characteristics of SV calling experiment results on three long-read technologies, i.e. ONT, PacBio CLR, and PacBio HiFi. Finally, they genotype a merged SV set generated using 31 long-read sequenced individuals of diverse ancestry, on 444 short-read sequenced individuals using Paragraph, a third-party genotyper, to investigate SV-gene expression associations.

Thank you for your thoughtful review.

Main comments:

1. Iris is a pipeline that consists of the application of Sniffles on a BAM file, using the alternate allele supporting reads given by Sniffles, running Racon to polish the inserted sequence or the refining of the breakpoints, aligning it back to the reference genome using minimap2 and parsing the CIGAR alignment string to give the updated SV. Is this correct?

Yes, your understanding is correct. We updated the text to be clearer: “The first new method, Iris, refines variant calls by using racon to polish the variant sequence from reads supporting the alternate allele and realigning this polished sequence to the reference with minimap2.”

2. Regarding the sentence: “Jasmine improves upon other SV merging methods by representing variants as points in space based on their breakpoints and lengths and constructing a graph of SV proximity, where edges represent pairs of SVs with a small Euclidean distance between them”: This is an almost identical representation to what is done in SV merging in https://www.nature.com/articles/s41588-021-00865-4 . This approach, however, is presented in this manuscript as a key strategy differing from all other SV merging approaches.

We updated the text to better clarify that the novelty in Jasmine’s methods is in how it processes the graph, not by its use of an SV graph: “Jasmine represents variants as points in space based on their breakpoints and lengths and constructs a graph of SV proximity, where edges represent pairs of SVs with a small Euclidean distance between them. It improves upon other methods by globally considering the entire graph to prioritize merging nearby variants.”

Also see below for the results of our benchmarking of this method.

3. Besides the similar representation of SVs and the distances between them, the application of using a minimum spanning tree using a modified version of Kruskal’s algorithm also appears to be analogous in purpose to the approach employed in the study by (Beyter et al., 2021) which uses a clique finding heuristic, although is a different algorithm, designed to answer a different question. It would be useful to point out the benefits of the proposed approach.

Thanks for pointing this out. We added a few sentences to the text about this to clarify the advantages. One of the main advantages of this approach is that computing a minimum spanning tree is substantially faster than clique-finding (asymptotically polynomial time instead of exponential time). Consequently, Jasmine does not require the use of pre-processing heuristics to reduce the number of variant pairs being considered, resulting in merges based on the nearest breakpoints possible. Another benefit of this approach is the fine control over breakpoint distance that it enables. Finally, cliques-finding based approaches may be too restrictive when a subset of variants form a highly connected, but not fully connected, subgraph.

4. The authors compare their approach to five existing methods listed as (Shi et al. 2021, Jeffares et al. 2017, Ebert et al. 2021, Larson et al. 2019, and Beyter et al. 2021), however it is unclear which tool name belongs to which work in this comparison. For example, the tool "svimmer" shown in Figure 2F is a tool used in https://www.nature.com/articles/s41467-019-13341-9 (Eggertsson et al. 2019).

When we made our initial submission, we read the Beyter et. al publication (at the time a preprint) and since it made no mention of sv-merger we tried to determine which merging software it used by looking for merging software on the authors' Github account. This led us to falsely ascribe svimmer to that paper, which we fixed in our revisions when we added comparisons against sv-merger. The sentence now reads "Furthermore, we compared Jasmine to six existing methods for SV comparison between samples (Figure 2f): dbsvmerge (Shi et al. 2021), SURVIVOR (Jeffares et al. 2017), svpop (Ebert et al. 2021), svtools (Larson et al. 2019), sv-merger (Beyter et al. 2021), and svimmer (Eggertsson et al. 2019)." to explicitly clarify which tool came from which prior publication. In the revision we also compare to sv-merger (see next query).

5. It is also unclear whether the SV merging approach employed in the cited study (Beyter et al. 2021), which uses a similar graph based SV merging approach, is compared against or not.

We did not compare against this approach in our original submission. For our revision, we compared our approach against the source code here https://github.com/dbeyter/sv-merger following the instructions in the manuscript. We updated Figure 2 to include the discordance results on the HG002 trio, which shows more than a 3-fold rate of discordant SVs compared to Jasmine.

We also attempted to merge our larger cohort level callsets with this software, running the included preprocessing script and parallelizing across both chromosomes and SV types. Even so, it did not terminate after 72 hours, so we did not include it in that analysis. The manuscript has been updated to reflect this and include more details about how we ran it.

6. Besides regions where the mapping of long-reads are ambiguous, a potentially challenging yet frequent task would be merging SVs within tandem repeat regions due to their abundance. It is true that the application of Iris can be beneficial in such regions, but it would be informative to evaluate

performance within and outside of such more difficult regions to highlight where the challenge in SV merging lies.

In the revision, we expanded our analysis of the Mendelian discordance in the HG002 trio to separately analyze SVs which are contained in tandem repeat regions. Overall, we found a similar discordance rate to the overall callset. We have updated the text to reflect these findings:

"To address the large number of discordant variants, our optimized pipeline offers a number of improvements which reduce the rate of Mendelian discordance by more than a factor of five with <1% (404/47,326 = 0.009) of merged SVs being discordant (Figure 2b). We also evaluated the discordance rate among SVs contained in tandem repeats and found a similar discordance rate of 0.007 (209/28,339)."

7. Although the study aims to perform a population-scale SV comparison and analysis, the study uses at most 31 long-read samples. While this is understandable due to the lack of larger publicly available long-read datasets, typically population-scale studies can involve hundreds or thousands of individuals. It could be useful to have insights on how Jasmine will perform in the existence of such large scale data, perhaps using simulated datasets and SVs.

Thanks for the suggestion! We simulated per-sample callsets based on the 1000 Genomes Phase 3 Structural Variant Calls and measured Jasmine's ability to scale up to the 2,504 samples included in that dataset. Across all 2,504 samples, the peak RAM was less than 12 GB, and the analysis was completed in under 9 hours. The newly added Supplementary Figure 18 shows these results.

8. While the finding regarding a deletion in SEMA5A affecting the expression of the gene is interesting, based on Supp Fig. 22, the deletions appears to be around 36 basepairs. SV studies have thus far used a variant size of 50 basepairs or more. As a result previous SV studies or databases may lack this variant simply due to not satisfying a size cutoff threshold rather than previous methodical limitations. The authors also state that this deletion is difficult to detect using short-reads alone, but do not provide a reason or evidence for this claim.

As several studies have reported that indel variants between 30 bp and 50 bp are also difficult to discover with short reads, we wanted to highlight this use in addition to canonical SV calls that are at least 50bp in size. However, to address this issue and the concerns of the other reviewer we have replaced this example with another insertion SV (3,143 bp in size) that serves as an SV-eQTL for CSF2RB and that we previously reported in the supplemental results.

To show the difficulty of identifying this variant, we first checked that it does not appear in the 1000 Genomes Project SV callset. We next checked the GTEx WGS data and found that it was not present in the GTEx indel callset or the GTEx SV callset. To further validate this variant, we then used the PARAGRAPH genotyping tool to genotype this SV with GTEx WGS alignment data, and we confirmed that the SV allele distributions within the GTEx population and the 1KGP population largely agree with each other. We believe the GTEx callset is an ideal benchmark for this analysis due to its high sample size, reasonably deep sequencing depth, and a highly confident set of short read SV calls curated with state-of-the-art pipelines.

To further demonstrate the importance of recovering this variant, we performed SV-eQTL analysis with the GTEx gene expression data from matching tissue in EBV transformed lymphocytes, and were able to produce similarly significant results compared to the results found in 1KGP data albeit with a lower sample size. Further causality analysis of the variant using CAVIAR shows that the SV in CSF2RB has the highest CAVIAR posterior score.

9. Overall, I believe the text for the main manuscript and the number of main figures/panels used can be significantly reduced and more emphasis on the method, highlighting its advantages, can be included.

In addressing earlier comments, we added text which further highlights the methodological advantages of Jasmine. We also made several revisions to the text to improve the clarity of the manuscript.

Minor comments:

10. What do you mean by “harmonized” callset in the abstract? And how many of the 205,192 SVs would be of higher confidence/quality?

We changed “harmonized” to “unified” for clarity - the intent was to communicate that our callset represents a union of all the SV calls in any of the 31 samples. In terms of confidence, this is post-filtering, so all SVs represent high-confidence calls.

11. In the sentence: “... in complex regions to be accurately identified by short reads, which make up the majority of existing genomic sequencing data”. Please see https://www.sciencedirect.com/science/article/abs/pii/S0002929721000987 , which mentions that about 9,7% of the current GRCh38 reference is defined by segmental duplications (SD) and simple repeats (SR). While it is true that the majority of SVs detected using long-reads according to the study given above are from long-reads alone (25,000 lrWGS SVs vs. 11,000 srWGS SVs), I disagree that the majority of the human DNA is composed of complex regions -- if that is what the authors express.

We changed the sentence to clarify what we meant. Our intent was that short reads make up the majority of existing genomic sequencing data, not that complex regions make up the majority of genomic sequencing data.

12. Why did the authors decide to include variants between size 20-50 (Fig 2D) or 30-50 (the entirety of the manuscript) as SVs given that SVs have typically focused on variants larger than or equal to 50 nucleotides in size (Chaisson et al. 2019, Audano et al. 2019, gnomAD-SV, … etc.)

The consideration of variants with length below 30 was only for avoiding thresholding effects (e.g., a true variant of length 30 that is called as a variant of length 29 in one or more samples due to sequencing or alignment errors). Variants below 30bp after Iris polishing or Jasmine merging can be removed as needed.

As for the use of 30bp as a threshold instead of 50bp, these variants have been shown to be an important class of variation e.g. (Alonge et al, Cell, 2020), yet benchmarking results in small indel calling have shown that indels in the size range 30-50 are difficult to map and detect with short-read small variant methods e.g. (Zook et al, Nature Biotechnology, 2014; Narzisi et al, Nature Methods, 2014; Sirén et al, Science, 2021). However, to ensure fair comparisons against other tools such as svpop which non-optionally filter out calls below 50bp, we also performed our allele frequency analysis separately on only the SV calls with length at least 50bp. We clarified these results in the text, and replaced several examples in the main text with larger variants (≥50bp) to emphasize this point.

13. The sentence: “Jasmine avoids merging variants of the same type which correspond to unique breakpoint adjacencies, which is particularly important when resolving complex nested SVs” is unclear. Although the Supp Fig 21 elaborates, the problem is yet difficult to understand.

We added more detail to the text and Supplementary Figure 21 (now Supplementary Figure 2) caption to help clarify this further.

14. Although Jasmine successfully decreases the denovo candidate SVs from 2194 SVs down to 404 SVs in the analyzed sample, this number is still impractically high to perform manual analyses (less than 10 per generation on average), particularly in the presence of multiple parent-offspring trios. As a result, detecting denovo SVs may be considered a different problem than SV merging.

While it is a different problem, having more accurate SV calling/merging enables the discovery of *de novo* SVs through the trio merging approach we describe. However, it is a fair point that 404 is still a large number of SVs to inspect manually, so we reworded the results section to clarify that this is just one step towards the routine identification of de novo SVs.

15. In the results presented in Fig. 2, it would be beneficial to know what number of discordant SVs are filtered by which approach, as several filters are applied to called SVs before the application of Jasmine. A table showing the removed number (and percentage given total number of SVs) of discordant SVs may be clearer to understand. For example, how many of the 2194 discordant SVs can be removed using strict yet simple filters?

We added a note to the text which explains the filtering (which is the same procedure for all merging software), and more details are described in the Evaluating Mendelian Discordance section of the Methods.

16. Please clarify the explanation and interpretation of Figure 2F as I find it to be the most important panel in the figure. For example, what does it mean “... but corresponding to unique breakpoint adjacencies (mixed strand)” ?

We added more detail to the text and Supplementary Figure 2 caption to help clarify this further. One example of this is partially detected inversions where there is read support for the alternate sequence at one end of the inverted sequence but not at the other end. Correctly identifying these partial inversions, or similarly partial translocations, is important for resolving complex rearrangements made of up nested SVs.

17. I cannot see the increased number of variants around size 6-7 kb in any of the provided plots.

We added density line plots to the supplement which show this peak for the HG002 HiFi analysis, the cross-technology HG002 analysis, and the cohort-level analysis.

18. There exists, however, a peak around 500-750 in almost all SV size plots in the paper, and no explanation is given about it. This peak does not exist in other SV datasets such as (Audano et al. 2019, Beyter et al. 2021) which uses long-reads or gnomAD-SV which uses short-reads. Is this an artifact, or is this expected?

We investigated this, and many of these variants are in centromeres (Supplementary Figure 20), and are results of the poor sequence representation in GRCh38. Notably, in our companion work within the Telomere-to-Telomere Consortium (Aganezov *et al*, bioRxiv, 2021; journal version under review) we have shown this peak is largely absent when applying these same methods using the T2T-CHM13 genome as the reference, so we conclude that this peak observed with GRCh38 supports the need for higher quality reference genomes when calling SVs. We added a note about this result and this work in the discussion.

19. The behavior of Sniffles shown in Fig 2E is strange as a seemingly more restrictive parameter provides an increase in the total number of variants discovered (while reducing discordant SVs). Do you think this is simply a mis-adjusted default parameter in Sniffles?

The default parameter was determined to perform well on older, higher-error sequencing data (circa 2015-2017) and help to detect variants despite the high error rate. But in doing so, nearby SV calls are often falsely collapsed into a single variant. We previously found and reported this issue with our work

with adaptive sequencing with Oxford Nanopore sequencing (Kovaka et al, Nature Biotechnology, 2020) although this is the first genome-wide analysis of this problem. We added a note to the text to clarify this point.

20. In the text discussing Fig. 3, no explanations made for HiFi-only SVs, whereas similar explanations are available for ONT-only or CLR-only SVs. This main figure can be a supplementary figure instead.

We focus on the SVs in the other technologies because they are higher-error than HiFi reads, thus making it more likely for calls unique to those technologies to be artifacts of sequencing errors rather than true SVs which are uniquely detectable by the technologies. We added a note to the text to clarify this reason. On the other hand, we do discuss the HiFi-only SVs in terms of their genomic context, because their high accuracy enables SVs to be detected in repetitive regions where reads derived from other technologies are less able to align.

21. While the usage of three different sequencing technologies/data naturally provides a more manageable (for manual examination) list of denovo SVs, performing a 3-fold increased sequencing at scale may not be a realistic endeavour for the detection of denovo SVs.

The purpose of this analysis was not so much to recommend performing sequencing on all three technologies, but to illustrate 1) that there are strong *de novo* SV candidates supported across multiple different technologies, and 2) that combining even just two technologies drastically reduces the number of candidate *de novo* SVs. We updated the text to emphasize these points more.

22. As the authors manually examine the set of denovo SV candidates for plausibility, they mainly check for the reliability of mapping of the long reads, which can be done programmatically as well. Again, this may be considered a different problem than what Jasmine is designed for.

Thank you for the suggestion! Since we are looking at a single trio with a small number of *de novo* candidates, we opted to manually inspect the IGV candidates. However, as long-read trio datasets become more abundant, we hope to develop methods for programmatically performing this analysis.

23. I do not find panel Fig 4B useful. What is the message/purpose intended in this panel?

The purpose of this panel is to show *de novo* candidates in terms of how they could be filtered by simple filters on values such as length and breakpoint precision from the SV caller. This was an effort to provide some insight into the question below (24), but we modified the caption of this figure to further clarify the message.

24. Again, I believe it is important to have an insight on the percent of denovo SV candidates that can be filtered with simple albeit strict filters, and the cases where Jasmine has a clear advantage with the algorithm it employs.

While we attempted to provide some insight here (see question 23 above), we are hesitant to provide recommendations for particular filters that could be used since we are only considering a single trio and there is a high risk of overfitting.

25. In Figure 5B, it may be more useful to give the breakpoint range as percentage of length of the final merged SV, rather than the actual breakpoint range.

Thanks for the suggestion! We added this as Supplementary Figure 8.

I also do not see much benefit or extra information in the panels 5D through 5G, they could reside in supplementary if desired.

Since we are providing the SV callset as a panel, we thought it important to highlight the characteristics of the SVs so that readers can easily understand the data they are working with.

What is shown in panel 5D x-axis is rather “number of carriers” rather than allele frequency.

Thank you for pointing this out; we changed the axis label as suggested.

26. While Jasmine does not merge SVs in an intra-sample method, it may yet be important to do so in the presence of multiple SV discovery algorithms applied to a single individual.

We agree with your point, and in fact Jasmine has a command line option "--allow_intrasample" which enables variants in the same VCF to be merged together. In the case of multiple SV discovery algorithms for a single individual, this feature is not needed as separate VCFs are treated by Jasmine as separate "samples" and merging between them would be allowed. We added a note to the manuscript to clarify these points.

27. Again, the 6-7kb peak in Figure 5F is not present as mentioned.

We added density line plots to the supplement which show this peak for the HG002 HiFi analysis, the cross-technology HG002 analysis, and the cohort-level analysis.

28. A similar analysis of genotyping SVs detected using long reads in short read samples have also been done in https://www.biorxiv.org/content/10.1101/2020.12.04.412486v2 using Giraffe (Siren et al., 2021), a recent software. Was there a particular reason for the use of Paragraph in genotyping short reads?

It is important to highlight that Jasmine and Giraffe serve very different goals. The goal of Jasmine is to unify variant calls from multiple individuals while Giraffe is a pangenome short-read mapping algorithm. Giraffe itself relies on precomputing a unified set of variant calls, but the approach used by Siren et al is highly specialized for this paper and is not available as a separate standalone tool. Consequently, we cannot directly compare Jasmine to their method, although we expect Jasmine to provide superior results since this was not the main focus of the Giraffe paper.

As for genotyping, the results reported by Giraffe and their earlier work on the vg genotyper (Hickey et al, Genome Biology, 2020) offer similar accuracy to Paragraph, but these methods require more computation time and more complexity to determine the genotypes, so we opted to use Paragraph. We have also released our SV panel so that users can run other existing or future methods on their own datasets.

29. Again, a similar SV-eQTL association analysis has been done by Siren et al. (https://www.biorxiv.org/content/10.1101/2020.12.04.412486v2) Have you compared your eQTL association results to those presented there? What would constitute a novelty in your results?

We checked the three SV-eQTLS we highlighted (the insertion associated with *CSF2RB* expression, the deletion associated with *LRGUK* expression, and the insertion associated with *CAMKMT* expression) for presence in the dataset provided by Siren et al., and found that while they also detected two of the three SVs as part of their callset, they did not report any of the same gene expression associations.

We also note that for the purpose of our manuscript, the main contribution is a novel method for SV calling and processing, and while we used this method to perform an initial look at SVs associated with gene expression changes, we are excited to leverage our methods for more extensive novel SV-eQTL analyses in the future.

30. The discussion section is too verbose and can be shortened significantly, particularly the second paragraph.

We have updated the text to make this section less verbose.

31. It is a good idea to polish the insertion sequences and refine the breakpoints using Racon, but other methods such as performing a multiple sequence alignment or a haplotype resolved assembly can also achieve useful results for the task of refining variants/sequences. Were there particular reasons for the employment of Racon within Iris?

While there are other polishing methods out there, we opted to use racon because it is well-established and it has been shown to perform well on both ONT and CLR data, which enables Iris to be more broadly used.

32. To my understanding the Jasmine pipeline uses polished SVs as input. As one can also use it without such SV polishing, it would be interesting to see the SV merging results without the SV polishing to establish the effect of the Iris pipeline into Jasmine.

Since our goal is to produce the most accurate SV calls possible, we recommend always running the polishing. While we expect the more accurate calls to slightly improve merging, we are less concerned with this effect of running Iris and instead interested in how it improves the accuracy of the individual calls, which we benchmarked.

33. In discordant SV analysis, while the merging goes through using all called SVs, what is the reason for applying various filters when assessing the total number of variants? Filtering erroneous SV calls can also reduce an erroneous SV merging.

The purpose of keeping variants with weaker evidence through the merging step is that some of these variants may correspond to true variants, and because of threshold effects they would be missed if using a stricter filter pre-merging. We filter them out after merging only if they were not merged with a higher-evidence call from another sample, allowing the presence of a variant in one sample to strengthen the case for it being present in other samples. We added more detail about double thresholding to the text.

34. In the section regarding Double Thresholding, I do not see the motivation in properly detecting/discovering a variant in all of the samples in which they are present. While it is true that a variant may be confidently discovered in only a subset of the samples, it can be genotyped in all carriers, with appropriate genotyping likelihoods, i.e. uncertainties, which is more suitable for downstream approaches such as imputation, association, … etc. SVjedi (https://github.com/llecompte/SVJedi), and LRcaller (https://github.com/DecodeGenetics/LRcaller) are example tools for long-read SV genotyping.

Our double thresholding approach is analogous to joint genotype calling using gVCF files that is now well established for single nucleotide variants: the initial screen finds candidates, which are then refined by comparison to other samples. While genotyping all SVs called in each individual is another option, it is more efficient to perform a single pass on the data and detect evidence for all SVs in all samples at the same time. In addition, this approach is more robust to noise (e.g., calls from high-error technologies) to enable nearby SVs to be merged rather than genotyping a particular call.

35. The authors indicate that out of the 205K merged SVs, 138K of them were left unfiltered. It would be beneficial to compare this list to other publicly available SV lists such as gnomAD-SV to assess the benefit of genotyping SVs detected using long-read sequencing samples on short-read sequencing samples, over using short-reads alone. Using lenient comparison approaches (due to the application of different sequencing technologies) on variants with a size of at least 50 bp (in order to have comparable size thresholds) will be the most meaningful.

The publicly available variant records of gnomAD-SV are collapsed across thousands of individuals, and lack the resolution required for an effective comparison or matching between variants from our callset and gnomAD call set. Consequently, it is not feasible to acquire the data and satisfactorily perform a large-scale analysis as proposed.

Instead, we refer the reviewer (and readers) to several studies that have reported the increased sensitivity and precision of long read SV calling over short reads e.g. (Sedlazeck et al, Nature Methods, 2018; Chaisson et al, Nature Communication, 2019; Audano et al, Cell, 2019; etc). Furthermore, the PARAGRAPH manuscript (Chen et al, Genome Biology, 2019) has already performed a detailed analysis of the advantages and accuracy of genotyping long read SV calls within short read datasets so it is not necessary to repeat that analysis here.

~~~~~~

Reviewer #2:

Remarks to the Author:

Kirsche et al. share two algorithms to improve SV calling accuracy and merging, Iris for refining SV breakpoints, and Jasmine for merging SVs. Iris polishes SV sequences with Racon to include all reads in the SV representation. The paper focuses on Jasmine, a merging algorithm using a novel algorithm framed as a spanning tree problem to link SVs across samples and merge intra-sample variants. With optimizations, such as double thresholding, they are able to produce a merge less susceptible to noise than other merging approaches. The authors go on to show that Mendelian error is reduced. Using
~~~~~~

Jasmine, they merge 31 SV callsets and show that the callset characteristics follow previous results and harbors potentially functional variants, such as an SV in SEMA5.

Methods to properly address SV merging is lacking. In past years, reciprocal overlap was mostly adequate for short-read SVs, but the overlap approach under-performs in modern SV callsets. New methods are needed, and Jasmine is potentially an important advancement. Large consortia, such as the HGSVC and HPRC, are generating long-read callsets, and a growing number of smaller projects are as well. The presentation of Jasmine is timely.

I had no trouble installing and running Jasmine through Conda. I did run a comparison of Jasmine with default parameters against SV-Pop (merging parameters "nr:szro=50:offset=200", same used for HGSVC) against HG00733 (PBSV, HiFi). I found Mendelian errors of 700 and 710 for Jasmine and SV-Pop outside centromeres, respectively (116 and 125 outside all tandem repeats). Although these are not quite the large gains I was hoping to see from Jasmine compared to SV-Pop, I still think it's a valuable approach with more potential than current methods.

Thank you for your comments, especially your efforts to install and run the software yourself.

I have several major concerns with this manuscript:

1) SV sizes and counts are not adding up through the manuscript, and it's unclear what your size threshold for defining an SV is. Most commonly, it's 50 bp., but I believe you are using 30 bp (from online methods). SV counts will differ with most other published callsets that define SVs at 50 bp, which makes it hard to compare. Fig 2C suggests 30 bp, Fig 2D suggests 20 bp. In Fig 1A, there are 26,591 SVs in the child (14257 + 5260 + 4880 + 2194), which is what I would expect if the SVs are >= 50 bp. However, compared to the abstract, 205,192 merged from 31 samples makes no sense when Ebert 2021 just published half as many for the same number of samples. Is the SV size definition consistent throughout the paper, or is there another explanation for these differences? Please clarify and make it clear at the beginning what your SV size is.

We use a definition of 30bp throughout the manuscript as these variants have been shown to be an important class of variation e.g. (Alonge et al, Cell, 2020), yet benchmarking results in small indel calling

have shown that indels in the size range 30-50 are difficult to map and detect with short-read small variant methods e.g. (Zook et al, Nature Biotechnology, 2014; Narzisi et al, Nature Methods, 2014; Sirén et al, Science, 2021). We added a note at the beginning that we define them this way. We also added a note to the Figure 2d caption to clarify that SVs below 20bp are shown only to illustrate double thresholding and that they are only kept if they get merged with another variant which exceeds the 30bp threshold.

2) A large fraction of callset and merging errors occur in tandem repeats because aligners and callers do not get them right no matter how long or high-quality reads and contigs are. I think it's important to report key findings on all variants (post-QC) as well as variants outside tandem repeats (with some supporting figures at least in supplement if possible).

Thank you for this suggestion. We evaluated the accuracy within and outside of tandem repeat regions by stratifying the Mendelian discordance analysis by the UCSC track of tandem repeats. We added a note to the results in the text showing that we see similar discordance results inside and outside of tandem repeat regions.

3) Lastly, HPRC data used for population studies is publicly available but unpublished. You must check with the PIs before publishing it and making sure data and assembly production receives due credit.

For this analysis, we are using the “HPRC+” datasets (also called HPRC_PLUS) which were contributed to the HPRC but do not fall under the same embargo policy as the more recent HPRC data (https://github.com/human-pangenomics/HPP_Year1_Data_Freeze_v1.0). The HPRC+ samples can be used without restrictions (e.g. https://twitter.com/aphillippy/status/1441083435049652225). We are appreciative of this contribution, and added a sentence to the acknowledgements. We have also clarified this in the supplemental table where we listed the data sources.

Minor comments:

4) Italicize gene names (SEMA5A).

Updated in the text

5) "While discovering significantly more SVs (Figure 2C)." Significantly more SVs than what, and where's the test of significance?

We removed the word significantly from the text

6) Fig 2A & 2B. "number of samples called" (in legend). Do you mean the number of SVs called? The title says "Mendelian Discordance", but that is only the last column.

Yes, that should be SVs instead of samples. We updated the text accordingly.

7) Figure 2C suggests a high false-call rate. Previous long-read papers report around 15k insertions and 10k deletions (50 bp and greater), but Fig 2C shows 16k insertions and 14k deletions in the same size (seems too small if SVs are 30+ bp).

We have clarified the figure panel title to explain that these are the SVs across the entire trio, so include variants from HG002, HG003, and HG004. In addition, we expect to discover more SVs through the combination of HiFi data (many existing long-read studies relied on higher-error sequencing technologies), as well as methodological improvements such as double thresholding which enable the discovery of SVs based on weaker evidence when there is strong evidence for the SV in a different individual in the trio.

8) I am not sure why the numbers between 2B and 2C don't match. If I add up the bins in 2B that include the son (36k), it doesn't match 2C with (46k) or without (31k) the 30-50 bp size range. What accounts for this difference?

The text was incorrect here and labeled the SVs in that panel as those in HG002, but it is actually the SVs across the entire trio. We apologize for the confusion, and updated the text and figure panel title to reflect that.

9) Fig 2C is cited in the text as "discovering significantly more SVs", but 2C is just a size distribution not a comparison, so it doesn't support this claim.

We updated the text to place the figure call-out in a more appropriate location.

10) Fig 2D: Double-thresholding is not explained at all before citing Fig 2D. leaving the reader wondering what it is. Provide a brief explanation and cite methods for details. It's also not clear to me what "rescued from absence" means, if you are merging calls, no call should be absent. In the methods, please provide an explanation how SVs are assigned to each category that appears in the legend.

We added a brief explanation to the text and cited the section of the online methods. We also expanded the online methods section on double thresholding to include detailed descriptions of the different categories.

11) Fig 2E: Again, please make it clear in the manuscript what SV size threshold is used. It's difficult to interpret the middle panel without knowing. A brief explanation for those less with Sniffles might help if you are using it to illustrate a main point. This should probably be a supplementary figure.

We added a note to the beginning of the manuscript to clarify that the length threshold is 30bp here and elsewhere.

12) Fig 2E: "The effects of the sniffles m ax_dist". Capitalize "Sniffles".

Changed in text

13) Fig 2F: Explain "unique breakpoint adjacencies (mixed strand)" in the text, it only makes sense after looking at Sup Fig 21, but it's part of a main figure. Is this only a Sniffles anomaly, as Sup Fig 21 suggests?

If so, that should probably also be clarified. Text citation is misplaced, should also be moved to the next sentence where it is introduced.

We reworded the text and added more detail there and in the supplementary figure caption to make it more clear. We also moved the text citation to this panel.

14) "with a slight enrichment of insertions, shown in previous studies to be caused by missing sequence in the human reference genome." This accounts for only a small fraction, there are probably just more insertions because deletions are more often deleterious (Cooper 2011, Nat Gen).

Thanks for pointing that out! We updated the text to include this.

15) "There is also an increased number of variants around sizes of 300bp and 6-7kbp, corresponding to SINE and LINE elements respectively." The 6 kbp peak isn't shown anywhere because the bins are too large in Fig 3C. You could add a supplementary histogram with smaller bins to show it. These peaks are commonly used as an indication of callset quality, so I think it's worth adding.

We added density line plots to the supplement which show this peak for the HG002 HiFi analysis, the cross-technology HG002 analysis, and the cohort-level analysis.

16) Fig 3D: The TRA bin ended up in "10k+" (no TRA label).

Fixed

17) Supplementary figures 11-13: Add legend (bar color)

Added legends to all three of them

18) Supplementary figures are referenced out of order.

We have updated the order.

19) In "De Novo Variant Discovery", "a 43-fold reduction in candidates from using prior methods (Figure 2a)", the reduction is part method, but can also be attributed to using three sets of rather expensive sequence data, so it's not all methods as the statement suggests. The wording strikes me as disingenuous.

We have updated the text to also attribute this to the use of multiple technologies. Also see our response to reviewer #1 above.

20) Fig 4A caption: "both of the examples in (a) and (b)". Do you mean parts c and d?

Fixed in text

21) Fig 4B: I don't think this is a main figure. It's hard to see where the points are in 3D space. Consider making it 2D with bubble size representing one of the dimensions or move it all to supplement.

Thanks for the feedback! We updated the figure to be a 2-D scatter plot with one of the dimensions as size.

22) The HiFi discordance rate doesn't make sense to me and I wonder if it can be attributed to using Sniffles. PBSV was designed for HiFi variant calling, and I suspect the de novo rate would come down.

Previous studies of HiFi sequence analysis, including the original HiFi paper (Wenger et al, Nature Biotechnology, 2019) report that Sniffles and pbsv have similar precision and recall when working with HiFi reads. While Jasmine works with pbsv as well as Sniffles, we opted to use a caller that could perform well across multiple data types for cross-technology analysis.

23) In "Population SV Inference", Supplementary figures 17 and 18 should be 18 and 19.

Fixed in text

24) Supplementary figure 19 could be improved by replacing "Population Allele Frequencies" in the title of every panel with the actual tool name.

Updated this and Supplementary Figure 20 (now Supplementary Figures 11 and 12)

25* "These data suggest that many of the SVs that are only visible through genotyping long-read-based variant calls have large effects on gene expression and thus are potentially functionally relevant.". True, but it also suggests noise introduced by genotyping.

The Paragraph paper (Chen et al, Genome Biology, 2019) has a detailed analysis of the accuracy of genotyping SVs detected by long reads in short-read datasets. Furthermore, to ensure that the top hits we highlighted were not genotyping noise, we added a new analysis to validate them in the independent GTEx cohort. We look forward to larger long-read datasets in which short-read genotyping is not necessary, and we expect Jasmine will remain a critical component in those analyses.

26) Refer to Nature guidelines with revisions, multipart figure labels mix case between figure and legend (OK for review).

We updated the labels appropriately.

Decision Letter, first revision:

Date: 23rd Mar 22 21:43:20

Last Sent: 23rd Mar 22 21:43:20
Triggered By: Lin Tang
From: Lin.tang@nature.com
To: mschatz@cs.jhu.edu
CC: methods@us.nature.com;ziqian.li@nature.com
Subject: Decision on Nature Methods submission NMETH-A46161A
Message: 22nd Mar 2022

Dear Dr Schatz,

We very much apologize again for the delay on the decision for your Article, "Jasmine: Population-scale structural variant comparison and analysis". The paper has now been seen by 2 reviewers. As you will see from their comments below, our reviewers have still raised a number of concerns. We are interested in the possibility of publishing your paper in Nature Methods, but would like to consider your response to these concerns before we reach a final decision on publication.

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

When revising your paper:

* include a point-by-point response to the reviewers and to any editorial suggestions

* please underline/highlight any additions to the text or areas with other significant changes to facilitate review of the revised manuscript

* address the points listed described below to conform to our open science requirements

* ensure it complies with our general format requirements as set out in our guide to authors at www.nature.com/naturemethods

* resubmit all the necessary files electronically by using the link below to access your home page

[REDACTED]

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

We hope to receive your revised paper within eight weeks. We are very aware of

the difficulties caused by the COVID pandemic to the community. If you cannot send it within this time, please let us know. In this event, we will still be happy to reconsider your paper at a later date so long as nothing similar has been accepted for publication at Nature Methods or published elsewhere.

OPEN SCIENCE REQUIREMENTS

REPORTING SUMMARY AND EDITORIAL POLICY CHECKLISTS
When revising your manuscript, please update your reporting summary and editorial policy checklists.

Reporting summary: https://www.nature.com/documents/nr-reporting-summary.zip
Editorial policy checklist: https://www.nature.com/documents/nr-editorial-policy-checklist.zip

If your paper includes custom software, we also ask you to complete a supplemental reporting summary.

Software supplement: https://www.nature.com/documents/nr-software-policy.pdf

Please submit these with your revised manuscript. They will be available to reviewers to aid in their evaluation if the paper is re-reviewed. If you have any questions about the checklist, please see http://www.nature.com/authors/policies/availability.html or contact me.

Please note that these forms are dynamic 'smart pdfs' and must therefore be downloaded and completed in Adobe Reader. We will then flatten them for ease of use by the reviewers. If you would like to reference the guidance text as you complete the template, please access these flattened versions at http://www.nature.com/authors/policies/availability.html.

DATA AVAILABILITY
We strongly encourage you to deposit all new data associated with the paper in a persistent repository where they can be freely and enduringly accessed. We recommend submitting the data to discipline-specific and community-recognized repositories; a list of repositories is provided here: http://www.nature.com/sdata/policies/repositories

All novel DNA and RNA sequencing data, protein sequences, genetic polymorphisms, linked genotype and phenotype data, gene expression data, macromolecular structures, and proteomics data must be deposited in a publicly accessible database, and accession codes and associated hyperlinks must be provided in the "Data Availability" section.

Refer to our data policies here: https://www.nature.com/nature-research/editorial-policies/reporting-standards#availability-of-data

To further increase transparency, we encourage you to provide, in tabular form, the data underlying the graphical representations used in your figures. This is in addition to our data-deposition policy for specific types of experiments and large

datasets. For readers, the source data will be made accessible directly from the figure legend. Spreadsheets can be submitted in .xls, .xlsx or .csv formats. Only one (1) file per figure is permitted: thus if there is a multi-paneled figure the source data for each panel should be clearly labeled in the csv/Excel file; alternately the data for a figure can be included in multiple, clearly labeled sheets in an Excel file. File sizes of up to 30 MB are permitted. When submitting source data files with your manuscript please select the Source Data file type and use the Title field in the File Description tab to indicate which figure the source data pertains to.

Please include a "Data availability" subsection in the Online Methods. This section should inform readers about the availability of the data used to support the conclusions of your study, including accession codes to public repositories, references to source data that may be published alongside the paper, unique identifiers such as URLs to data repository entries, or data set DOIs, and any other statement about data availability. At a minimum, you should include the following statement: "The data that support the findings of this study are available from the corresponding author upon request", describing which data is available upon request and mentioning any restrictions on availability. If DOIs are provided, please include these in the Reference list (authors, title, publisher (repository name), identifier, year). For more guidance on how to write this section please see: http://www.nature.com/authors/policies/data/data-availability-statements-data-citations.pdf

CODE AVAILABILITY
Please include a "Code Availability" subsection in the Online Methods which details how your custom code is made available. Only in rare cases (where code is not central to the main conclusions of the paper) is the statement "available upon request" allowed (and reasons should be specified).

We request that you deposit code in a DOI-minting repository such as Zenodo, Gigantum or Code Ocean and cite the DOI in the Reference list. We also request that you use code versioning and provide a license.

For more information on our code sharing policy and requirements, please see: https://www.nature.com/nature-research/editorial-policies/reporting-standards#availability-of-computer-code

MATERIALS AVAILABILITY
As a condition of publication in Nature Methods, authors are required to make unique materials promptly available to others without undue qualifications.

Authors reporting new chemical compounds must provide chemical structure, synthesis and characterization details. Authors reporting mutant strains and cell lines are strongly encouraged to use established public repositories.

More details about our materials availability policy can be found at https://www.nature.com/nature-portfolio/editorial-policies/reporting-standards#availability-of-materials

ORCID

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

Please do not hesitate to contact me if you have any questions or would like to discuss these revisions further. We look forward to seeing the revised manuscript and thank you for the opportunity to consider your work.

Sincerely,

Lin

Lin Tang, PhD
Senior Editor
Nature Methods

Reviewers' Comments:

Reviewer #1:
Remarks to the Author:
Comments on the rebuttal (main):

(prev. Comment 1) Please cite racon and minimap2 in the given sentence "Iris, refines variant calls by using racon to polish the variant sequence from reads supporting the alternate allele and realigning this polished sequence to the reference with minimap2".

(prev. Comment 2) I find it difficult to understand how Jasmine improves over other methods by "globally considering the entire graph to prioritize merging variants". I believe the conceptual advantages of Jasmine over other approaches/algorithms should be explained more clearly.

(prev. Comment 3) While a clique finding problems, such as the Corrupted Cliques Problem is NP-hard, there exist polynomial heuristics to approximately solve it, such as the Cluster Affinity Search Technique (Ben-Dor et al. 1999), as applied in the study compared against – as a result, I disagree with the initial advantage given. However, careful implementation can achieve better runtime. I do not understand what the authors mean by "Jasmine does not require the use of pre-processing heuristics to reduce the number of variant pairs being considered, resulting in merges based on the nearest breakpoints possible". Similarly, what is meant by "fine control over breakpoint distance"? Again, a number of the methods compared against also provide breakpoint distance or overlap percent parameters.

I agree with the last point of the authors regarding "clique-finding based approaches may be too restrictive when a subset of variants form a highly connected, but not fully connected subgraph" – it is quite plausible that when given fuzzy SV breakpoints with a restrictive set of clustering parameters, a clique based approach may yield to more than a single merged SV.

(prev. Comment 4) The tools compared are clarified and corrected.

(prev. Comment 6) I am glad to see that the SV discordance results remained comparable within tandem repeats as to the total rate.

*It would be also informative to see the effect of Iris on Jasmine, i.e. have Jasmine run on data that have not gone through the Iris pipeline, as this would highlight what is most important, i.e. providing accurate SV breakpoints or the merging algorithm used.

(prev. Comment 7) The newly added scalability analysis is useful.

(prev. Comment 8) The newly added SV-eQTL example given in the updated manuscript (3,143 bp insertion in NCF4) and its association with the expression levels for the gene CSF2RB is already reported in Audano et al., 2019 (https://pubmed.ncbi.nlm.nih.gov/30661756/ , Table S3) therefore cannot be referred to as a new discovery. The authors should compare their SV-eQTL results to Audano et al. 2019, and other previously published results on SV-eQTLs.

(prev. Comment 9) I would like to reiterate that higher emphasis on the methodological advances could be added into the main text, such as the most crucial points included in the section "Jasmine methods" and the moving of several figures/panels to the Supplementary, especially Fig 4 and Fig 5.

Comments on the rebuttal (minor):
(prev. Comment 14) I could not find where you clarified that using Jasmine is one step towards the routine identification of de novo SVs. The section "De Novo Variant Discovery" under the results section still mentions that the suggested approach can identify known and yet unknown de novo SVs without further detail.

Reviewer #2:
Remarks to the Author:
I am extremely conflicted about this manuscript. On the positive side, I do not doubt Jasmine's ability to perform as advertised for most bioinformaticians. Current merging tools are still using variations of reciprocal overlap, incremental improvements on reciprocal overlap, or are a mess of parameters and cutoffs that don't generalize well. They are also affected by merge order. The closest alternative according to this work is SV-Pop, but it was never ported to a framework for the general bioinformatics community, it relies on incremental improvements on reciprocal-overlap, and it is affected by merge order arbitrarily picking a representative variant from the first sample. Jasmine has the ability to transcend these limitations and provide a useful mature tool that is easy to install and run. This is critical to improve results and reproducibility, and it is broadly

applicable to the field.

On the negative side, I do not believe this manuscript is mature enough for publication. While I don't doubt Jasmine's performance because I have (briefly) run it, the manuscript itself does not deliver a cogent story. I have found this callset confusing and problematic, and still do.

1) Differing choices for the SV size cutoff in the field is unfortunate, and it makes it difficult to compare across studies. Historically, the SV cutoff has been placed and moved as technologies have changed, but over the last decade, it has been 50 bp, which is consistently used by SV consortia and major efforts including GnomAD, CCDG, HGSVC, the 1000 Genomes Project, and Sniffles itself [32461652, 32460305, 33632895, 26432246, 29713083, 31729472]. You are correct that Illumina has sensitivity problems for smaller variants, but I think you need to separate 30-49 in your analyses instead of acknowledging the 50 bp precedence and redefining it to 30 bp for this manuscript. This is needed to help readers match the numbers in your manuscript with what is already widely published.

2) When you remove tandem repeats (TR), Mendelian error rate should go down, not up (0.9% all, 0.7% in TRs, 1.0% outside TRs - "(404-209) / (47326 - 28339)"). Tandem repeats are very difficult to compare in callsets, GIAB often has to ommit them [32541955], they cause assembly collapses [26442640, 31584084], and the true de novo rate is higher in TRs [31659027]. I cannot see any explanation for why Mendelian error would go down in TRs unless Jasmine was over-merging in these loci. TRs lead to clusters of indels and SVs, and if allowed to match arbitrarily, would inflate the numerator in the Mendelian error calculations. This is must also be exacerbated by counting 30-49 bp variants as SVs.

3) Even with the SV size cutoff clarification, the callsets make it difficult to understand how Jasmine would generalize to other samples. -- Specifically: i) I think the manuscript does a poor job separating confident variant calls from noise, for example, it appears the authors left in centromeres and large repeats. ii) I find the reliance on Winnowmap and Sniffles for all technologies an odd choice, for example, Winnowmap is not commonly used HiFi as far as I can tell after a quick literature search. iii) It doesn't look like any corrections were made for callable regions across technologies or samples, for example, loci with adequate read depth for variant detection across all callsets. In Fig 3d-f, SV discordance goes up in HiFi compared to other technologies even though it is widely known that HiFi produces better SV calls [31406327, 33319909, 33632895]. -- The most likely explanation for these observations seems to be more reads mapping to repetitive loci with no corrections made for this. If you are including SVs from regions that are not callable from other technologies, then it obscures any point you are making about merging (i.e. merging against nothing is vacuously unique). As far as I can tell, it could also be Winnowmap and Sniffles, or Iris might be polishing small errors into the callsets, but these seem less likely. Rerunning each callset with current best practices would be advisable, but I think proper QC would be acceptable since the main purpose of this paper is merging and not SV discovery. These same problems show up again in the Child-only analysis.

4) More minor than 1-3 above, but I still find the point about double-thresholding and varying max_dist concerning. This actually makes it look to the reader like

Jasmine is a merging tool for Sniffles. For example, if I merge variants from assemblies, as is commonly done [33632895, 32686750, 33288905, 10.1101/664623], how would Jasmine perform? It might be helpful to quantify the gain and make it clear that this would generalize outside of Sniffles. The whole point about max_dist is completely lost to me, and as written, the manuscript says that Jasmine will "mitigate (threshold effects...), improved variant calling parameters". Why would you want to mitigate improved variant calling parameters? The point about max_dist is also unclear. The first time I read this, it took me a while to see what you were showing here, and the second time makes less sense. I think it would be good to dedicated more space in the manuscript to make these points clear or move it to supplement.

Because of these issues, I do not think the manuscript is ready for publication. I hope the authors will make significant improvements in a timely manner and that the journal will either allow an additional round of revisions or accept a resubmission.

Author Rebuttal, first revision:

Reviewers' Comments:

Reviewer #1:

Remarks to the Author:

Comments on the rebuttal (main):

1. Please cite racon and minimap2 in the given sentence “Iris, refines variant calls by using racon to polish the variant sequence from reads supporting the alternate allele and realigning this polished sequence to the reference with minimap2”.

Thank you for the suggestion. We added the citations to the text.

2. I find it difficult to understand how Jasmine improves over other methods by “globally considering the entire graph to prioritize merging variants”. I believe the conceptual advantages of Jasmine over other approaches/algorithms should be explained more clearly.

Thanks for the suggestion! We clarified this point by adding the sentence “This is in contrast to prior methods, which often perform sub-optimal merging because they utilize heuristics to consider

smaller subgraphs of the variant proximity graph." to that part of the manuscript and also added edits to the methods descriptions for additional clarity throughout.

3.

a. While a clique finding problems, such as the Corrupted Cliques Problem is NP-hard, there exist polynomial heuristics to approximately solve it, such as the Cluster Affinity Search Technique (Ben-Dor et al. 1999), as applied in the study compared against – as a result, I disagree with the initial advantage given. However, careful implementation can achieve better runtime. I do not understand what the authors mean by "Jasmine does not require the use of pre-processing heuristics to reduce the number of variant pairs being considered, resulting in merges based on the nearest breakpoints possible".

To clarify these points from our last reviewer response, an *exact* algorithm for clique-finding is necessarily exponential in runtime, and while algorithms such as the CAST algorithm exist to find an approximate solution faster, they do not always reconstruct all of the cliques. This was found in even the simulated data in the experiments from Ben-Dor et al., particularly in high-noise settings. Additionally, the method we compare against, sv-merger, which uses the CAST algorithm, requires the use of a pre-clustering step to reduce the graph size and speed up the algorithm. This requires careful tuning, and may result in the removal of valid edges from the graph, which will lead to sub-optimal merging.

b. Similarly, what is meant by "fine control over breakpoint distance"? Again, a number of the methods compared against also provide breakpoint distance or overlap percent parameters.

While clique-based methods also allow for the addition of breakpoint distance parameters, the results from the minimum spanning tree approach are more interpretable. For example, merging with Jasmine, any pairs of SVs in different samples which are within the distance threshold will always be merged (unless one or both of them has a nearer neighbor in the other sample), regardless of the variants in other samples. On the other hand, when using clique-based approaches the reason for variants remaining unmerged is more complex to interpret. For example, even when two SVs across different samples are very close together, the addition of an SV in a third sample which is within the distance threshold of only one of those variants can cause the original pair to not be merged. These scenarios become even more complex as the number of samples increases.

c. I agree with the last point of the authors regarding "clique-finding based approaches may be too restrictive when a subset of variants form a highly connected, but not fully connected subgraph" – it is quite plausible that when given fuzzy SV breakpoints with a restrictive set of clustering parameters, a clique based approach may yield to more than a single merged SV.

Thanks for your feedback.

4. The tools compared are clarified and corrected.

Thanks for your feedback.

6. I am glad to see that the SV discordance results remained comparable within tandem repeats as to the total rate.

While the discordance rate within TRs was similar to the overall discordance rate, we did find through our analyses for this latest revision that the discordance rate in HiFi data for HG002 among variants within 500bp of TRs was higher (.0096 within these regions; .0046 outside of them; .0087 overall). While this is still less than 1% it does demonstrate the difficulty of calling SVs in those regions, and so we added a few sentences to the results and discussion about this. We also define regions at least 500bp outside of TRs as "high-confidence" and separately provide a confidence-filtered callset for the 31-sample cohort.

*It would be also informative to see the effect of Iris on Jasmine, i.e. have Jasmine run on data that have not gone through the Iris pipeline, as this would highlight what is most important, i.e. providing accurate SV breakpoints or the merging algorithm used.

We evaluated this on the HG002 trio with HiFi-derived variant calls and showed that the discordance was reduced from 1.02% to 0.85% through the addition of Iris refinement. While this is a relatively small improvement, the main benefit of Iris refinement is in the more accurate breakpoint/sequences in the variant calls rather than its impact on merging.

7. The newly added scalability analysis is useful.

Thanks for your feedback.

8. The newly added SV-eQTL example given in the updated manuscript (3,143 bp insertion in NCF4) and its association with the expression levels for the gene CSF2RB is already reported in Audano et al., 2019 (https://pubmed.ncbi.nlm.nih.gov/30661756/ , Table S3) therefore cannot be referred to as a new discovery. The authors should compare their SV-eQTL results to Audano et al. 2019, and other previously published results on SV-eQTLs.

Thank you for pointing this out. As this result was previously reported, we have reframed our results with the 1000 Genomes cohort as a validation study to show how our method can identify SV-eQTLs that were previously known and moved our previous figure 6 to become supplementary figure 21. We have additionally performed new analysis applying our SV-eQTL framework using Jasmine and Paragraph to the entire GTEx collection spanning 873 individuals with as many as 48 non-diseased tissues. From this analysis, we find over 111,000 significant eGenes across 48 tissues. Notably, we conservatively estimate the number of cases where an SV-eQTL is the top eQTL to be 10,436, which is over 2,000 more examples than previously reported even when using a stricter FDR threshold. We then describe in more detail a few notable examples, including SV-eQTLs in HACL1, DDTL, and ASMTL that were not previously reported but we find the p-value, t-statistic, and CAVIAR posterior probability of causality to be substantially stronger with the SVs than any flanking SNPs. Interestingly, the DDTL SV-eQTL is putatively causal in 36 tissues and the tissue log p-values distribution is significantly higher (p-value= 1.1e-8, one-sided Wilcoxon rank sum test) than the top SNP associations in the same tissues. These new results are presented in the revised Figure 6 along with several new supplementary figures (S. Fig. 24-27)

9. I would like to reiterate that higher emphasis on the methodological advances could be added into the main text, such as the most crucial points included in the section “Jasmine methods” and the moving of several figures/panels to the Supplementary, especially Fig 4 and Fig 5.

Thanks for the suggestion. We added more detail and clarification about the methods to the main text, and moved some of the optimizations which are unrelated to SV merging to the supplement. We would be open to reordering/restructuring the main figures if the editor deems this is necessary.

Comments on the rebuttal (minor):

14. I could not find where you clarified that using Jasmine is one step towards the routine identification of de novo SVs. The section "De Novo Variant Discovery" under the results section still mentions that the suggested approach can identify known and yet unknown de novo SVs without further detail.

In our previous edits we had added a clarification about this to the end of the "Reduced Mendelian Discordance in an Ashkenazim Trio" section: The resulting reduction in Mendelian discordant variants is an important step towards the rapid identification of *de novo* variants, as it is typically necessary to screen all discordant variants manually when searching for true *de novo* variants." To make this more clear, we added a similar sentence at the end of the "De Novo Variant Discovery" section.

Reviewer #2:

Remarks to the Author:

I am extremely conflicted about this manuscript. On the positive side, I do not doubt Jasmine's ability to perform as advertised for most bioinformaticians. Current merging tools are still using variations of reciprocal overlap, incremental improvements on reciprocal overlap, or are a mess of parameters and cutoffs that don't generalize well. They are also affected by merge order. The closest alternative according to this work is SV-Pop, but it was never ported to a framework for the general bioinformatics community, it relies on incremental improvements on reciprocal-overlap, and it is affected by merge order arbitrarily picking a representative variant from the first sample. Jasmine has the ability to transcend these limitations and provide a useful mature tool that is easy to install and run. This is critical to improve results and reproducibility, and it is broadly applicable to the field.

Thanks for your feedback. We also believe Jasmine will be a mature tool to transcend the issues you have identified! We also hope our revisions will address your other concerns so that our manuscript will be ready for publication.

On the negative side, I do not believe this manuscript is mature enough for publication. While I don't doubt Jasmine's performance because I have (briefly) run it, the manuscript itself does not deliver a cogent story. I have found this callset confusing and problematic, and still do.

1) Differing choices for the SV size cutoff in the field is unfortunate, and it makes it difficult to compare across studies. Historically, the SV cutoff has been placed and moved as technologies have changed, but over the last decade, it has been 50 bp, which is consistently used by SV consortia and major efforts including GnomAD, CCDG, HGSVC, the 1000 Genomes Project, and Sniffles itself [32461652, 32460305, 33632895, 26432246, 29713083, 31729472]. You are correct that Illumina has sensitivity problems for smaller variants, but I think you need to separate 30-49 in your analyses instead of acknowledging the 50 bp precedence and redefining it to 30 bp for this manuscript. This is needed to help readers match the numbers in your manuscript with what is already widely published.

We restructured our analysis and updated all applicable main and supplementary figures to primarily show results on SVs (adding additional supplementary figures with the combined SV+indel results as necessary), as well as numeric results in the text. The figure changes are distributed through most main figures and supplementary figures (~40 panels) that didn't previously describe the variants by their size.

Overall, the resulting numbers are much more comparable to counts from existing SV studies. For example, in our cross-technology analysis (Figure 3), we find a total of 24,596 SVs identified with HiFi in HG002, which is in line with prior estimates: on average 22,755 per person from (Audano et al, Cell, 2019) and on average 24,653 SVs per person from (Ebert et al, Science, 2021). In addition, our 31-sample cohort SV analysis yielded 122,813 SVs, which is comparable to the 107,590 count which was detected from long reads across 32 human individuals in (Ebert et al., 2021). We also updated the text where we introduce 30-49bp indels to refer to this class of variants as separate from SVs, and adopt the classical 50bp definition of SV throughout the manuscript.

2) When you remove tandem repeats (TR), Mendelian error rate should go down, not up (0.9% all, 0.7% in TRs, 1.0% outside TRs - "(404-209) / (47326 - 28339)"). Tandem repeats are very difficult

to compare in callsets, GIAB often has to ommit them [32541955], they cause assembly collapses [26442640, 31584084], and the true de novo rate is higher in TRs [31659027]. I cannot see any explanation for why Mendelian error would go down in TRs unless Jasmine was over-merging in these loci. TRs lead to clusters of indels and SVs, and if allowed to match arbitrarily, would inflate the numerator in the Mendelian error calculations. This is must also be exacerbated by counting 30-49 bp variants as SVs.

Thanks for pointing this out! We investigated this and found that while the discordance rate of SVs within TRs was similar to overall, there was an increased discordance rate among SVs within 500bp of TRs (.0096 within these regions; .0046 outside of them). We added a few sentences about this to the results and discussion section, and adapted a filter for variants at least 500bp outside of TRs as our high-confidence filter for the cohort variant panel.

3) Even with the SV size cutoff clarification, the callsets make it difficult to understand how Jasmine would generalize to other samples. -- Specifically:

a. i) I think the manuscript does a poor job separating confident variant calls from noise, for example, it appears the authors left in centromeres and large repeats.

In our trio analysis, we found a significant reduction in discordance (<0.5%) among SVs which are at least 500bp away from tandem repeats, and so used this to define high-confidence regions. We added a few sentences to the discordance analysis about this and defined high-confidence regions at that point, and then in our cohort analysis we filtered to variants falling in these regions provided this set of 22,132 SVs and 13,615 indels as a part of our data release.

b. ii) I find the reliance on Winnowmap and Sniffles for all technologies an odd choice, for example, Winnowmap is not commonly used HiFi as far as I can tell after a quick literature search.

We opted to use Winnowmap even for HiFi data because our work as part of the T2T Consortium in variant calling with respect GRCh38 and CHM13 (Nurk et al. 2022) showed that it outperforms other aligners due to its prioritization of unique and low-frequency k-mers, especially in repetitive regions of the genome where other aligners struggle. The Winnowmap v2 paper (Jain et al. 2022) corresponding to the version we used for our analysis was also published in Nature Methods on April 1 and shows that it performs well across

technologies. In addition, since there may be cases where a user would want to utilize different alignment software, our pipeline enables the use of ngmlr or minimap2 for alignment in place of Winnowmap. We updated the text to refer to the Winnowmap2 publication in place of the previous Winnowmap one.

c. iii) It doesn't look like any corrections were made for callable regions across technologies or samples, for example, loci with adequate read depth for variant detection across all callsets. In Fig 3d-f, SV discordance goes up in HiFi compared to other technologies even though it is widely known that HiFi produces better SV calls [31406327, 33319909, 33632895]. -- The most likely explanation for these observations seems to be more reads mapping to repetitive loci with no corrections made for this. If you are including SVs from regions that are not callable from other technologies, then it obscures any point you are making about merging (i.e. merging against nothing is vacuously unique). As far as I can tell, it could also be Winnowmap and Sniffles, or Iris might be polishing small errors into the callsets, but these seem less likely.

In our previous analysis, we did not look at the discordance in each individual technology and were only looking at variant calls in HG002 rather than the entire trio (which we looked at in our discordance analysis for Figure 2) - the purpose of Figures 3d and 3f was to look at the variants identified uniquely by different technologies, and so we expect the distributions of SVs there to be a combination of technology-specific errors and SVs in regions where other technologies are less well-equipped to detect SVs (e.g., regions with near-exact repeats for HiFi reads). Our intersection with repeats and other genomic elements characterizes where these technology-specific differences occur.

To add additional clarity to the manuscript, we added a few sentences about the discordance in the trio for each technology and found the HiFi and ONT have similarly low discordance (.0087), while CLR data produces much higher discordance (.0161).

d. Rerunning each callset with current best practices would be advisable, but I think proper QC would be acceptable since the main purpose of this paper is merging and not SV discovery. These same problems show up again in the Child-only analysis.

Thanks for your feedback. As you state the main purpose of this paper is merging not SV discovery. We hope the edits we made above will clarify this point and these findings.

4) More minor than 1-3 above, but I still find the point about double-thresholding and varying max_dist concerning. This actually makes it look to the reader like Jasmine is a merging tool for Sniffles. For example, if I merge variants from assemblies, as is commonly done [33632895, 32686750, 33288905, 10.1101/664623], how would Jasmine perform? It might be helpful to quantify the gain and make it clear that this would generalize outside of Sniffles. The whole point about max_dist is completely lost to me, and as written, the manuscript says that Jasmine will "mitigate (threshold effects...), improved variant calling parameters". Why would you want to mitigate improved variant calling parameters? The point about max_dist is also unclear. The first time I read this, it took me a while to see what you were showing here, and the second time makes less sense. I think it would be good to dedicated more space in the manuscript to make these points clear or move it to supplement.

Thanks for the suggestion! It is a good point that the main focus of our contribution is the merging method independent of the particular variant caller, and so we updated the manuscript to better place emphasis on this, including moving the max_dist and double thresholding optimizations to the supplement as well as adding additional detail and clarification to the description of our merging method. Additionally, Jasmine can be used with other SV callers, such as the pbsv caller that we previously tested with an earlier version of Jasmine in (Aganezov et al, 2020, Genome Research)

Because of these issues, I do not think the manuscript is ready for publication. I hope the authors will make significant improvements in a timely manner and that the journal will either allow an additional round of revisions or accept a resubmission.

Thank you again for your feedback. We hope our new submission will satisfy your concerns.

Decision Letter, second revision:

Date: 22nd Sep 22 22:54:40
Last Sent: 22nd Sep 22 22:54:40

nature research

Triggered By: Lin Tang
From: Lin.tang@nature.com
To: mschatz@cs.jhu.edu
CC: methods@us.nature.com
Subject: AIP Decision on Manuscript NMETH-A46161B
Message: Our ref: NMETH-A46161B

22nd Sep 2022

Dear Dr. Schatz,

Thank you for submitting your revised manuscript "Jasmine: Population-scale structural variant comparison and analysis" (NMETH-A46161B). It has now been seen by Reviewer 2 and their comments are below (Reviewer 1 recently told us that they were not able to re-review). The reviewers find that the paper has improved in revision, and therefore we'll be happy in principle to publish it in Nature Methods, pending minor revisions to satisfy the referees' final requests and to comply with our editorial and formatting guidelines.

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

TRANSPARENT PEER REVIEW
Nature Methods offers a transparent peer review option for new original research manuscripts submitted from 17th February 2021. We encourage increased transparency in peer review by publishing the reviewer comments, author rebuttal letters and editorial decision letters if the authors agree. Such peer review material is made available as a supplementary peer review file. **Please state in the cover letter 'I wish to participate in transparent peer review' if you want to opt in, or 'I do not wish to participate in transparent peer review' if you don't.** Failure to state your preference will result in delays in accepting your manuscript for publication.
Please note: we allow redactions to authors' rebuttal and reviewer comments in the interest of confidentiality. If you are concerned about the release of confidential data, please let us know specifically what information you would like to have removed. Please note that we cannot incorporate redactions for any other reasons. Reviewer names will be published in the peer review files if the reviewer signed the comments to authors, or if reviewers explicitly agree to release their name. For more information, please refer to our FAQ page.

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

Sincerely,

Lin Tang, PhD
Senior Editor

Nature Methods

ORCID
IMPORTANT: Non-corresponding authors do not have to link their ORCIDs but are encouraged to do so. Please note that it will not be possible to add/modify ORCIDs at proof. Thus, please let your co-authors know that if they wish to have their ORCID added to the paper they must follow the procedure described in the following link prior to acceptance: https://www.springernature.com/gp/researchers/orcid/orcid-for-nature-research

Reviewer #2 (Remarks to the Author):

Overall, the callset makes more sense to me. Thank you for making the small changes that helped clarify the numbers and where they came from. I still believe the cross-technology section (ONT, CLR, & HiFi) could have benefit from restricting comparisons to callable regions, although I don't believe it's strictly necessary since you have shown where most of those differences are coming from.

The SV eQTL results seem to fit what has been published before, and I don't see any apparent problems with the additions, although it is not my expertise.

For the record, the extra mode in Figure 5b from SV-Pop is from a bug in the merging pipeline that allowed long-distance merges if the SV sizes were similar. This was found and fixed before the Jasmine paper was submitted, but it was present when SV-Pop was run for this paper and is fair game.

Minor comments

Lines 132-139: I think defining "high-confidence" based on relative TR location is ambiguous. For example, "high-confidence" is used on line 212 describing noise in CLR and on line 568 describing a filter. I think adding 500 bp to TR loci and calling that a TR region might be more clear. Merging errors near TRs is a useful observation, and I'm glad to see it in the paper.

Line 840 (Fig 3 legend): Should "across" in the legend title be capitalized for consistency?

Figure 3: Main text says 18,778 variants are common across all 3 technologies, and the Fig 3a Venn intersect agrees, but summing across Fig 3c is 311 variants short. The SVs in Figs 3d-f also don't seem to add up. Double-check numbers to make sure they are accurate.

Figure 4b: Add space to "Breakpoint Variance".

Supplementary Figure 9: Erroneous "Text Editor" label in figure.

Supplementary Figure 18: "filtering out SVs with lengths below 50bp" is a little odd since all SVs are 50+ bp. Maybe add parenthesis to clarify (e.g. "filtering out SVs (< 50bp)") or just remove it since this is already defined in the main.

Line 215: Don't need to re-define indels here, it's already done.

Fig 6 d,e: Suggest adding the chromosome name to the horizontal axis to complete the chromosomal coordinate. Same of similar supplementary figures. The chromosome name is only mentioned in the main text, not the figure or figure legend.

Supplementary & main: I think Sniffles, Racon, and Winnowmap should be capitalized (that's how they were published). Looks like minimap2 was published with a lower-case name. "svpop" is "SV-Pop".

Ensure font/spacing changes (e.g. lines 604-608) don't make it into the final supplement.

- Peter Audano

Author Rebuttal, second revision:

Reviewer #2:

Remarks to the Author:

Overall, the callset makes more sense to me. Thank you for making the small changes that helped clarify the numbers and where they came from. I still believe the cross-technology section (ONT, CLR, & HiFi) could have benefit from restricting comparisons to callable regions, although I don't believe it's strictly necessary since you have shown where most of those differences are coming from.

Thank you for your comments. We agree this will be an important point for future research to further evaluate the error models of the different technologies, but as you point out we hope we have satisfactorily identified where most of the differences come from.

The SV eQTL results seem to fit what has been published before, and I don't see any apparent problems with the additions, although it is not my expertise.

Thank you for your support.

For the record, the extra mode in Figure 5b from SV-Pop is from a bug in the merging pipeline that allowed long-distance merges if the SV sizes were similar. This was found and fixed before the Jasmine paper was submitted, but it was present when SV-Pop was run for this paper and is fair game.

Thank you for confirming it is “fair game”.

Minor comments

Lines 132-139: I think defining "high-confidence" based on relative TR location is ambiguous. For example, "high-confidence" is used on line 212 describing noise in CLR and on line 568 describing a filter. I think adding 500 bp to TR loci and calling that a TR region might be more clear. Merging errors near TRs is a useful observation, and I'm glad to see it in the paper.

On your suggestion we have renamed these as “TR regions” and “non-TR regions” to specifically refer to these regions.

Line 840 (Fig 3 legend): Should "across" in the legend title be capitalized for consistency?

We have revised all of the figures for consistency and other small updates.

Figure 3: Main text says 18,778 variants are common across all 3 technologies, and the Fig 3a Venn intersect agrees, but summing across Fig 3c is 311 variants short. The SVs in Figs 3d-f also don't seem to add up. Double-check numbers to make sure they are accurate.

Thanks for pointing this out. We realized in some places we were inconsistently using >50bp (exclusive of 50bp variants) as the threshold for SVs while in other places we used ≥50bp (inclusive of 50bp variants). We have refreshed the text and figures to use ≥50bp consistently throughout.

Figure 4b: Add space to "Breakpoint Variance".

We have updated the text.

Supplementary Figure 9: Erroneous "Text Editor" label in figure.

Thanks for spotting this. Not sure how this was added but we have confirmed it is fixed in our resubmission.

Supplementary Figure 18: "filtering out SVs with lengths below 50bp" is a little odd since all SVs are 50+ bp. Maybe add parenthesis to clarify (e.g. "filtering out SVs (< 50bp)") or just remove it since this is already defined in the main.

We have updated the caption as suggested.

Line 215: Don't need to re-define indels here, it's already done.

We have removed this as suggested.

Fig 6 d,e: Suggest adding the chromosome name to the horizontal axis to complete the chromosomal coordinate. Same of similar supplementary figures. The chromosome name is only mentioned in the main text, not the figure or figure legend.

We added the chromosome name as suggested.

Supplementary & main: I think Sniffles, Racon, and Winnowmap should be capitalized (that's how they were published). Looks like minimap2 was published with a lower-case name. "svpop" is "SV-Pop".

We have updated this as suggested.

Ensure font/spacing changes (e.g. lines 604-608) don't make it into the final supplement.

We have adjusted this as suggested.

Final Decision Letter:

Dear Dr Schatz,

I am very pleased to inform you that your Article, "Jasmine and Iris: Population-scale structural variant comparison and analysis", has now been accepted for publication in Nature Methods. Your paper is tentatively scheduled for publication in our January print issue, and will be published online prior to that. The received and accepted dates will be 27th May 2021 and 15th Dec 2022. This note is intended to let you know what to expect from us over the next month or so, and to let you know where to address any further questions.

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

Once your paper is typeset, you will receive an email with a link to choose the appropriate publishing options for your paper and our Author Services team will be in touch regarding any additional information that may be required.

Please note that Nature Methods is a Transformative Journal (TJ). Authors may publish their research with us through the traditional subscription access route or make their paper immediately open access through payment of an article-processing charge (APC). Authors will not be required to make a final decision about access to their article until it has been accepted. Find out more about Transformative Journals

Authors may need to take specific actions to achieve <a

href="https://www.springernature.com/gp/open-research/funding/policy-compliance-faqs"> compliance with funder and institutional open access mandates. If your research is supported by a funder that requires immediate open access (e.g. according to Plan S principles) then you should select the gold OA route, and we will direct you to the compliant route where possible. For authors selecting the subscription publication route, the journal's standard licensing terms will need to be accepted, including self-archiving policies. Those licensing terms will supersede any other terms that the author or any third party may assert apply to any version of the manuscript.

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

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

Your paper will now be copyedited to ensure that it conforms to Nature Methods style. Once proofs are generated, they will be sent to you electronically and you will be asked to send a corrected version within 24 hours. It is extremely important that you let us know now whether you will be difficult to contact over the next month. If this is the case, we ask that you send us the contact information (email, phone and fax) of someone who will be able to check the proofs and deal with any last-minute problems.

If, when you receive your proof, you cannot meet the deadline, please inform us at rjsproduction@springernature.com immediately.

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

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

Once your paper has been scheduled for online publication, the Nature press office will be in touch to confirm the details.

Content is published online weekly on Mondays and Thursdays, and the embargo is set at 16:00 London time (GMT)/11:00 am US Eastern time (EST) on the day of publication. If you need to know the exact publication date or when the news embargo will be lifted, please contact our press office after you have submitted your proof corrections. Now is the time to inform your Public Relations or Press Office about your paper, as they might be interested in promoting its publication. This will allow them time to prepare an accurate and satisfactory press release. Include your manuscript tracking number NMETH-A46161C and the name of the journal, which they will need when they contact our office.

About one week before your paper is published online, we shall be distributing a press release to news organizations worldwide, which may include details of your work. We are happy for your institution or

funding agency to prepare its own press release, but it must mention the embargo date and Nature Methods. Our Press Office will contact you closer to the time of publication, but if you or your Press Office have any inquiries in the meantime, please contact press@nature.com.

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

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

Nature Portfolio journals encourage authors to share their step-by-step experimental protocols on a protocol sharing platform of their choice. Nature Portfolio 's Protocol Exchange is a free-to-use and open resource for protocols; protocols deposited in Protocol Exchange are citable and can be linked from the published article. More details can found at www.nature.com/protocolexchange/about.

Please note that you and any of your coauthors will be able to order reprints and single copies of the issue containing your article through Nature Portfolio 's reprint website, which is located at http://www.nature.com/reprints/author-reprints.html. If there are any questions about reprints please send an email to author-reprints@nature.com and someone will assist you.

Please feel free to contact me if you have questions about any of these points. Thank you very much again for publishing your paper with Nature Methods. We hope you have a very pleasant holiday season!

Best regards,

Lin Tang, PhD
Senior Editor
Nature Methods