Dear Dr. Andrew Cosgrove,

We thank you and the reviewers for their suggestions and comments. We have addressed their comments below and have added a few sentences in the discussion of our manuscript. The answers below are marked in blue to easier distinguish them from the reviewer comments. We have this time not highlighted the changes in the main text, which only are focused on the discussion. Furthermore, we followed the editorial suggestions and requirements as laid out by you.

## Reviewer #1:

The updated manuscript addressed most of my previous comments. However, I have some further suggestions.

1. My generalizability concern hasn’t been fully addressed. The section “Matching TVs between haplotypes” established that “Strict merge” maintains a good balance of removing redundancy and avoiding missing SVs. However, in a new cohort or new dataset, it is not clear whether the “Strict merge” is still recommended. Or if not, how to set the thresholds.

In performing the computational performance experiments, we used short-read sequencing sourced from the same individuals used which have long-read assemblies and two separate short-read SV callers. The resulting variants were merged with each of the tested callers and results presented in Fig S7. This figure shows that the default “Strict” parameters of Truvari collapse are generalizable across SV results from multiple tools and sequencing technologies.

2. SV here seems to mainly refer to deletion and tandem duplication, and inversion and translocations are ignored. But should be made clear of exactly what types of SV are considered.

We had highlighted in the previous submission: “While Truvari can process any SV type except unresolved breakends (BNDs), we focus here on only insertions and deletion.” and further “ It’s important to note that Truvari is currently most useful for ‘resolved’ SVs (i.e. DEL, INS, INV, and DUP). What we have not addressed in this manuscript are the challenges of multi-technology or unharmonized pipeline based SV comparison.” We think this addressed these concerns.

3. For the bar-chart of the number of SVs of each type, it would be helpful to include the counts both before merging.

We have updated the text to include the number of SVs having overlaps with genes.

4. The overview figure of Truvari only includes the methodology part, but it would be more comprehensive if input and output, especially different modes of Truvari (bench, collapse, annotation), are included in the overview figure.

The manuscript is focused on the merging of SV and the comparison. Thus the figure only focuses on this aspect of Truvari. The other (bench, collapse, annotation) do not rise to a significance that we are describing in detail. We have updated the text to include links to the user documentation on the github for these functionalities.

5. The principle of HWE and ExcHet is too specific to population genetics rather than general math metrics, and thus it should be better explained here for their definitions and assumptions.

We have updated the text to include more details.

6. In the “Benchmarking pVCFs” part, exact merge results from BCFtools are used to evaluate other methods. According to this part, BCFtools is expected to have has high sensitivity but does not have a good performance in removing redundant SVs. However, Table 1 doesn’t have a proper metric to evaluate redundancy-removing performance, which is misleading in the way that BCFtools is almost the best method.

Figure 3 and Figure 4 highlight the redundancy of bcftools. Table 1 shows the benchmark consequences when tools are over merging and thus produce lower performing results. While the bcftools and Truvari merge shows a high accuracy in Table 1. This is also because Table1 is based on a ~200 SV benchmark from GIAB.

7. The sentence is hard to parse: “The ratio of how many true-positive (TP) GIAB SVs are lost and how many potentially redundant calls are removed compared to the Exact merge is 1:790 for Strict and 1:141 for Loose”. There are 6 numbers, TP and FP for 3 merging methods, but it's it’s unclear from the phrasing how the ratios are computed.

We have updated the text for clarity.

## **Reviewer #2:**

I’d like to thank the authors for their thoughtful responses to my comments. While they did not incorporate all my suggestions, most were intended as additions to strengthen or augment the study rather than necessities and their corresponding edits have greatly improved the manuscript.

There is one minor point I would like the authors to still address:

> 7. Population-level metrics including HWE and excess heterozygosity are used, but the dataset is an admixture of ancestries. The authors need to show that the assumptions of these metrics still hold or to control for ancestry in their analysis.

> Here we are using HWE as a comparison method between different merging approaches based on the same input SV call set. Thus showing an impact of HWE not based on ethnicities or other external parameters. If all methods were to perform equally we would expect the HWE to be the same, but clearly they are not.

I may not have been clear in the initial comment about this. The manuscript states that fewer violating calls indicate a higher quality merge. However, the dataset being used is an admixture which breaks the well-mixed population assumption of both HWE and excess heterozygosity, and so it is not clear how accurately they reflect merging quality.

For example, one could imagine a scenario where a particular ancestry is highly homozygous for a particular variant that is not present in the other subpopulations. In this case, the variant would appear to be violating HWE and a tool that changes correct hom-var to het or hom-ref, while reducing HWE violations, would in fact be incorrect. Is there any evidence, perhaps from SNV calls, that HWE assumptions hold reasonably well in this dataset that the authors can cite?

We thank the reviewer for their assistance in improving our manuscript. We appreciate this concern from the reviewer, and we broadly agree that care must be taken when interpreting the results of Hardy-Weinberg tests on diverse and structured populations. However, we still believe that the tests are useful for flagging potential spurious genotypes, as implemented in genotyping best practice pipelines (e.g., GATK) as a standard method of quality control.

As noted in the hypothetical example provided by the reviewer, population structure tends to increase homozygosity—a phenomenon termed the "Wahlund effect”. However, allele frequency differences among human populations tend to be subtle. A plot (attached) of allele vs. genotype frequencies of randomly selected SNPs computed on the entire diverse cohort of 1000 Genomes samples demonstrates close adherence to Hardy Weinberg expectations (dotted lines), with a slight Wahlund effect and several clear outliers that are likely best explained as genotyping errors.

Meanwhile, admixture between previously separated populations may temporarily generate excess heterozygosity, but this signature quickly erodes in the following generation. Notably, our sample is almost entirely composed of individuals from the 1000 Genomes Project, who were specifically selected on the basis that all four grandparents came from the same location or ethnic group—mitigating the impact of very recent admixture. https://www.internationalgenome.org/sites/1000genomes.org/files/docs/Informed%20Consent%20Form%20Template.pdf

In light of these points, the threshold for calling a variant as “ExcHet” is likely slightly over-conservative, as the general effect of population structure in this sample is to decrease rather than increase heterozygosity compared to Hardy-Weinberg expectations. Meanwhile, the two-tailed Hardy Weinberg test (HWE) is likely slightly, but not strongly, under-conservative. In either case, for a sample of this size (n = 38), only very strong deviations achieve the threshold of statistical significance and are much more likely to be driven by technical as opposed to demographic factors.

We now briefly note these caveats of interpretation in the Results section when these Hardy-Weinberg metrics are first introduced.

In case Genome Biology’s reviewer response system allows images, we have uploaded the figure related to the above text to our paper’s repository https://github.com/ACEnglish/TruvariData/tree/develop/manuscript. Note that in that same link we have also added figures relevant to the mendelian consistency discussion from the previous round of reviews.