nature portfolio

Peer Review File

Accessing medically relevant complex regions with a pangenome graph of 20 near-complete Japanese haplotypes

Corresponding Author: Dr Yoshihiko Suzuki

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

Version 0:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
In this study, Suzuki and colleagues present their production and analysis of near-complete, high-quality, phased genome assemblies of 10 Japanese male individuals. The authors used Nanopore ultra-long reads, PacBio HiFi reads, and Omni-C long-range reads to generate these data, focusing their analyses on medically relevant complex regions containing segmental duplications. They provide general statistics describing the extent of structural variation in these regions and provide more details for the SMN and KIR loci. While they demonstrate an improvement in assembly quality in these regions compared to other pangenome datasets, the analysis of these regions remains somewhat limited.

Overall, the authors identify genes with notable differences in copy number between the Japanese population and the T2T-CHM13 reference, as well as East Asian haplotypes. They also report the discovery of new haplotypes in the SMN and KIR regions. However, given the inherent instability of these regions, the identification of additional haplotypes with further sequencing is to be expected.

This study contributes to ongoing efforts to generate and analyse near-complete genome assemblies from individuals of different populations. While the dataset represents a valuable resource, certain analyses could be further refined.

Specific comments

Page 3, line 99
Could the authors provide the auN (area under the Nx curve) values for these haplotypes?

Pages 3-4, lines 105-113
The authors state: "The assembly size per chromosome was consistent with the T2T-CHM13 reference genome, with the exception of the acrocentric chromosomes (chr13, 14, 15, 21, 22) and chromosome Y." However, based on Figure 1c, I observe that chromosomes 9 and 16 also show a reduced assembly size compared to the T2T-CHM13 reference. Additionally, while most acrocentric chromosomes are indeed shorter than those in the T2T-CHM13 assembly, there are some exceptions.

The authors attribute the larger variance in chromosome size of chromosomes 1 and 9 to "large structural rearrangements of HSat2/3 array length variations and inversions in centromeres, as previously reported." However, inversions do not affect chromosome size, as they alter only the sequence orientation without changing its length. The mention of segmental duplications (Figure 1f) is also unclear, as these are distributed across all chromosomes, not just chromosomes 1 and 9.

To clarify this point, I suggest revising the text and figure citations accordingly. Additionally, it might be helpful to include a plot displaying the total HSat2/3 length across chromosomes and haplotypes.

Page 4 - Segmental duplications
Could the authors provide additional details on Japanese-specific segmental duplications? Specifically:

Where are these duplications located?

Which genes are found in these regions?

Is copy number variation observed across different haplotypes?

Page 4, line 130 and Supplementary Figure 1
The plots display data for structural variants (SVs) up to 10 kbp in size. Have the authors identified SVs larger than 10 kbp?

Page 5, line 146
The authors report the number of shared and Japanese-specific SVs. To provide a more comprehensive view, I suggest also reporting the total number of base pairs covered by these SVs. This would help account for sequence matches and provide insight into the impact of SV size.

Pages 5 and 18-19
The criteria for gene copy number analysis are not entirely clear. Specifically:

How was gene copy number calculated?

How were the 26 "exceptional genes" (shown in Supplementary Figure 2) selected?

Additionally, I recommend including all "exceptional genes" in Figure 2a and all genes with copy number differences between JPT and EAS in Figure 2b, rather than showing only a representative subset. Some genes, such as USP17L11 and certain chrX genes, may not be truly exceptional, as the T2T-CHM13 copy number falls within the JPT and EAS distribution. Similarly, for USP17L11, a copy number difference of 1 between JPT and EAS is relatively minor given its average copy number of ~22. I suggest revisiting the criteria to define exceptional genes.

Page 7, line 234
The authors use a threshold of 3 SNVs per region, but these regions vary greatly in size (from 150 kbp to 6.7 Mbp). I suggest using a frequency-based threshold (e.g., SNVs per kb) rather than an absolute SNV count, to account for differences in region length.

Page 8, lines 256-258
The authors suggest that complex regions are better assembled in Japanese genomes. Could this be due to lower structural complexity, or are other factors at play? One way to explore this is by comparing the size of these regions in the HPRC year-1 dataset and the Japanese genomes.

Page 10 - Gene conversion
In Figure 4a, the color of the dots (blue/red) represents the direction of gene conversion, but some dots do not fully match the donor/acceptor profile in the graph below. Could the authors clarify this apparent discrepancy?

In Figure 3c, six haplotypes have two SMN2 copies and zero SMN1, suggesting a SMN1 → SMN2 gene conversion event. However, the authors do not report any gene conversion between SMN genes in their analysis. Could they clarify this apparent discrepancy?

Page 10, lines 349-353
This section would be more appropriately placed in the Discussion. Similarly, there are other parts in the Results that would be better suited for the Discussion section.

Page 19, lines 625-628
A difference between two paralogous copies would be better described as a paralogous sequence variant rather than a polymorphism.

(Remarks on code availability)

Reviewer #2

(Remarks to the Author)
The authors have generated near-complete haplotypes from 10 Japanese individuals. This completeness in medically relevant regions can help understand disease-associations and predispositions in a better way. While I appreciate the overall work and acknowledge the usefulness of these haplotypes as an informative data resource, I think that the analyses and findings presented in the manuscript need further elaboration and refinement.

Major comments:

1). The authors have compared their results to HPRC year 1 and Chinese pangenome. They mentioned in the Discussion about HGSVC's new assemblies for 65 individuals in the context of SMN genes, however, for the rest of the comparisons

these assemblies have not been included.

2). Mentioning the HGSVC assemblies, the authors state ‘Among the 130 haplotypes generated from worldwide individuals, they found only two haplotypes with three copies of SMN genes, while we found two such haplotypes among 20 Japanese haplotypes.’ I am not sure what they aim to convey with that statement. If it is to convey that the respective haplotype is more common in the Japanese population, it needs to be supported by statistical testing particularly since the number of individuals is so different.

3). The authors talk about overlaps of the SDs, SNVs and SVs with other data sets, however, it is not clear to me how this overlap was computed. As the uniqueness of these events , particularly SVs, relies heavily on the used matching criteria, it needs to be clearly stated in the manuscript.

4). There have been multiple mentions of ‘recurrent’ events, like SDs, and SNVs. e.g. it is stated that 74% of the SNVs are recurrent in the Japanese population. But it is not clear how the recurrence of those events was determined? As recurrence estimates normally require large cohorts, it should be explained clearly how did the authors check for it in their cohort of 10 samples and how confident are these findings?

5). There are a lot of statements in the manuscript that need a quantitative support. For example:

a). ‘MANY segmental duplications were resolved without gaps.’ on line 98.

b). ‘Although accurate annotation of paralogous genes in complex regions is difficult, MOST of such genes belong to gene families in segmentally duplicated complex regions’ in line 176.

c). ‘The average size of each region among the complete Japanese haplotypes was MOSTLY consistent with T2T-CHM13, while we identified considerable size deviations in several haplotypes indicating large-scale rearrangements’ in line 241.

Instead of saying many/most etc., quantitative assessments should be provided.

d). ‘Based on the copy number analysis and the literature’ in line 213. I didn’t understand which analysis or literature was being referred to.

e). ‘For instance, in this study we observed that the copy numbers of the TSPY3 and TSPY4 genes were zero or one in some East Asian haplotypes in the HPRC Year-1 dataset, compared to 8–14 and 2–6, respectively, in our Japanese haplotypes’. This statement in line 390 is confusing unless one looks at Figure 2b.

Minor Comments:
1). In Figure 3a, the arrow colors in the highlighted haplotypes are not consistent with the legend.

2). In Figure 4, the x-axis label is not center-aligned. Additionally, warm vs cold vs darker color distinction in the caption is adding unnecessary complexity. Since it's just 4 colors, I would recommend mentioning what each of them corresponds to in the figure legend. Additionally, instead of saying left-bottom triangle and right-top triangle I would suggest just saying below and above the diagonal.

3). Line 228, Typo ‘detest’.

4). Line 141, it should me made clear what do authors mean by ‘non-private SVs’.

5). Line 397, it should be clearly stated how a ‘minor’ haplotype is quantitatively defined.

6). Supplementary Figure 1 caption typo ‘insersion’.

(Remarks on code availability)

Version 1:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
The authors have addressed all my comments. I have no additional remark.

(Remarks on code availability)

Reviewer #2

(Remarks to the Author)
I appreciate the authors' efforts in addressing my previous comments. After reviewing the responses and the revised manuscript, I have a few additional comments and questions. I therefore recommend another round of revision.

1) Firstly, I would reiterate my previous comment about avoiding any unclear or ambiguous statements. e.g. in the revised manuscript on line 449 it is stated "For example, they reported 64% completeness in the SMN region, compared to our 85%, using slightly different but broadly comparable evaluation criteria". As a reader, I am curious to know what the exact criteria is.
2) Regarding the comparison of SDs, the authors have mentioned that they used bedtools intersect to compute the overlap but I couldn't find the details about how the SDs were identified in each Japanese haplotype before liftover. The authors have now mentioned the total base pairs but it is not clear to me how much of the 30.2 Mbp (unique region) corresponds to SDs where no overlap was observed and how much of it comes from the cases where only a certain fraction of the region overlaps because I wouldn't consider the sequence belonging to the latter category necessarily as "unique". I would suggest that each step of this analysis should be explicitly mentioned.
Line 136 says "While the majority of SDs identified in the Japanese haplotypes (105.3 Mbp, 80%) were also present in non-Japanese East Asian (EAS) haplotypes in the HPRC Year-1 dataset, 25.8 Mbp (20%) were specific to Japanese haplotypes". I find this sentence confusing. Does this mean that the SD comparison has been performed only across the non-Japanese EAS haplotypes in the HPRC cohort? If yes, why were the rest of the samples excluded? If not, the sentence should be rephrased to make it explicit.
Since the other comparisons presented in the paper have been performed against both HPRC and Chinese pangenomes, I would suggest the authors do the same for SD comparison. Additionally, I don't consider a comparison only with CHM13 as insightful as you are basically comparing to a single haplotype. So, I would suggest restructuring this section and mentioning a collective comparison including HPRC and Chinese pangenome.
Lastly, line 138 states "While many of these Japanese-specific SDs (18.7 Mbp, 72%) were located in centromeric regions". Given the highly repetitive nature of centromeres, I don't think that only sequence overlap is enough to determine the uniqueness of an SD inside a centromere. I would suggest either to use further sanity checks for these cases to make sure that these are not stemming from assembly errors etc. or exclude the centromeric regions from this comparison.
3) Line 147, page 5 "Two SNVs or SVs were considered identical if they shared the same genomic location and alternative allele sequence". For SVs, this criteria is too strict and expected to overestimate the unique SVs (more detailed response below).
Line 148, "Each method yielded comparable count of identified variants", seeing Supplementary figure 2, I won't say they are comparable. The distribution pattern is comparable but the numbers are not. Minigraph-cactus is clearly calling more SVs as compared to pggb. Was it stratified how many SVs are common and how many are unique to each method and how many of them are potentially false-positives? What was the rationale behind using the pggb graphs for further comparisons? I would recommend making all of this explicit in the text.
Line 159 "Among 37 k SVs (19.6 Mbp in total) detected in two or more haplotypes and thus considered reliable, 15k (41%, 7.7 Mbp) were shared with the HPRC graph, primarily with EAS haplotypes, while 59%, 11.9 Mbp were specific to the Japanese haplotypes (Fig. 1h). Note that the total number of SV bases may be overestimated, as SVs differing by even a single nucleotide were counted as distinct, despite our efforts to perform base-level normalization of nodes in the genome graph (Methods). Nonetheless, these results suggest a considerable proportion of population-specific SVs." 59% is an extremely high number keeping in view studies stratifying population specific SVs (e.g. Ebert et al. 2021). The authors have acknowledged that this is an over estimation but this can be corrected using a better comparison criteria. I would like to see how the numbers look like if you use one of the conventionally used SV comparison tools (e.g. Truvari) for this purpose.
4) For saying that something is "putatively recurrent" one still needs to provide some phylogenetic evidence showing that the event occurred independently over the course of evolution or a reference to other studies which provide such evidence for these events. Presence of an allele in more than two haplotypes doesn't indicate recurrence in any way. Therefore, I would recommend either backing the recurrence claims with proper phylogenetic support or replacing the word with something like "shared" (e.g.).
5) I would suggest moving the part explaining the comparison of KIR region with HGSVC in the Discussion to the Results section where the comparison to HPRC and Chinese pangenome is mentioned. Same for the SMN region. All the results should go together so that the reader can clearly understand what the novel insights are. The overall flow of the Discussion section can also be improved by this because currently two different paragraphs are talking about both the KIR and SMN regions with a paragraph about chromosome Y in between. This hinders the reader's ability to understand the message that is being conveyed.

Overall, I suggest that the authors refine and streamline the comparisons with other cohorts, particularly regarding population-specific SDs and SVs, to avoid potential overestimations. All methodological details should be clearly described in the text. Additionally, any claims lacking quantitative support should be revised or removed. I also recommend restructuring the manuscript, especially the Results and Discussion sections, to improve clarity.

(Remarks on code availability)

Version 2:

Reviewer comments:

Reviewer #2

(Remarks to the Author)
I appreciate the efforts from the authors in addressing my comments and questions. While I am satisfied with the responses and the respective revisions in the manuscript w.r.t most of my comments, I am still not convinced about the population specific SV analysis and the reported percentage of unique SVs.
In my previous review, I referred to the Science 2021 study 'Haplotype-resolved diverse human genomes and integrated analysis of structural variation' by Ebert et al. to provide an example of the scale at which population specific SVs are expected to arise. Figure 6C of this paper provides some estimates for these numbers based on the cohort used in this study. If I'm not mistaken there is only one Japanese individual in their cohort. So, if we want to have a very rough estimate, from the numbers in panel C, for the Japanese samples we can expect ~1000/25k i.e. around 4% population-specific SVs given we expect 25k SVs on average in a sample. Even if we give some margin of error in this estimation and account for varying sample sizes, the percentage reported by the authors i.e. 42% is extremely high and needs to be investigated.
My assumption is that it is still an SV comparison issue which might be fixed by using a stricter Truvari setting. One option is to start with the same parameter settings as used in the HPRC Nature 2023 paper for benchmarking variants. If the percentage stays this high even after using stricter comparison settings, further validation and quality control analyses e.g. stratification of the type and genomic location of these variants needs to be performed to support the claim that 42% of the SVs carried by these 10 individuals indeed have never been reported before. Reporting only a number is not sufficient. I would suggest the authors revisit this comparison to avoid any over-estimations.

Version 3:

Reviewer comments:

Reviewer #2

(Remarks to the Author)
I appreciate the effort of the authors in addressing my concerns, however, I am still not convinced that this number is not an overestimation. The authors are basically claiming that 42% of the structural variation seen in a group of 10 individuals from a human population is unique. The reference to the Ebert et al 2021 paper was just to provide a comparison from a study where such estimation has previously been provided for the Japanese population. I totally agree that the haplotype-resolved assemblies we have today provide much deeper insights into the genetic diversity as compared to short reads, however, that doesn't explain the rise to 42%. I cannot wrap my head around how it is even possible to see unique structural variation at such a scale.

I agree with the authors that in the HPRC Nature 2023 paper the same truvari parameters were used for SV merging, however, to my knowledge, this callset was not used to make any statements about population-specific uniqueness. Also, the SV comparisons in the HPRC 2023 paper were based on SV alleles not SV sites. That's why in my previous review I suggested using stricter settings for the estimation the authors are trying to make and performing further validation if the percentage still stays this high.

If we look into the other studies authors referred to for comparison: In Gao et al, the percentage of CPC-specific SVs is around 17% and most of them were located at the centromeric and telomeric regions of chromosomes. They also clearly state that more than half of the CPC-specific variants were singletons or doubletons. In comparison, the authors are reporting 42% while not even considering singletons. In the more recent Nasser et al 2025 study, around 22% SVs were classified as population-specific. So I am curious to know where the range of 30%-50% that the authors mentioned is coming from. It should also be noted that in Nasser et al. 80% reciprocal overlap and sequence identity was used for finding unique SVs as compared to 95% used by the authors which is extremely flexible and prone to overestimate the unique SVs as I have pointed out before as well.

Additionally, I would reiterate that if the authors insist that this percentage of 42% is not an overestimation, further stratification is required e.g. by variant type and more importantly by genomic location. Another experiment that would help get a better picture and assess the 'uniqueness' of these SVs is a visual depiction of the SV sequence identity with other cohorts that the authors are comparing to i.e. the Chinese pangenome and the HPRC cohort. One way can be to align the SV sequence to the respective pangenome graphs for each unique SV and plotting the resulting sequence identities together.

Overall, I would again urge the authors to revisit this analysis and be extremely careful in making statements like finding 'a considerable proportion of population-specific SVs'.

Version 5:

Reviewer comments:

Reviewer #2

(Remarks to the Author)
The authors have now made it explicit that instead of using the conventional way of comparison i.e. comparison to all SVs, they were comparing to only Japanese SVs which explains the inflated numbers. However, the revised statement reporting the SV stats in the manuscript is extremely confusing.

In the previous version of the manuscript the authors reported

“Among 44 k SVs (10.1 Mbp in total) found in Japanese haplotypes, 26 k (58%, 5.8 Mbp) were shared with the HPRC graph, primarily with EAS haplotypes, while 18 k (42%, 4.3 Mbp) were unique to our Japanese haplotypes (Fig. 1h), suggesting a considerable proportion of population-specific SVs”

In the latest version they state:

“Among 44 k SVs (10.1 Mbp in total) found in Japanese haplotypes, 26 k (15% of all SVs, 5.8 Mbp) were shared with the HPRC graph, primarily with EAS haplotypes, while 18 k (11%, 4.3 Mbp) were unique to our Japanese haplotypes”

First of all my question is if 15% is based on all SVs, what is 11% based on? Does it also use the number of all SVs in the HPRC graph as the denominator? If yes, what is the rationale for this and what does this value tell us about the callset the authors have generated? If it is w.r.t 44k, how is the number of SVs and the covered bps not changing? The concern I have been trying to raise repeatedly is about the message being conveyed, which is that almost half of the SVs found in this Japanese cohort have not been seen before. Using a different denominator simply to change the percentage of unique SVs from 42% to 11%, while keeping the SV count and covered base pairs exactly the same, does not address that concern. Because they still make up the same fraction as what the authors had reported before. For reporting the percentages, using all SVs as the denominator in my opinion is unnecessary here, firstly because it is making the sentence confusing by adding a different scale and secondly because when it comes to QC of a callset based on such a small cohort, the more important question to answer is how many of the found variants are already known and how many are unique. To be even more clear, the shared vs unique comparison should definitely be done using ALL SVs as the authors claim to have done now but the results should be reported w.r.t the new callset (as was being done before). So the denominator should still be the number of SVs the authors have found i.e. 44k (or the respective bps). The same comment applies to the now changed numbers about SNPs. It is completely unintuitive and confusing to use different scales to compute the same quantity. If the authors really want to report them, they need to be stated separately with explicit reasoning.

The stratification of unique SVs by genomic location and the results from the alignment exercise are very helpful and clearly indicate that most of this apparent uniqueness is either occurring in complex regions or stemming from representation differences which makes the claim of being "unique" doubtful. The take home message here should be that for these cases it is difficult to distinguish actual uniqueness from assembly artifacts or representation differences however the statement “These results indicate that most of these unique SVs are derived from repetitive regions” is not conveying that and should be revised.

Overall, I would strongly recommend the authors make another careful pass through the manuscript, be consistent in their comparisons and report the observations in a coherent and understandable way for the reader.

Version 6:

Reviewer comments:

Reviewer #2

(Remarks to the Author)
The current revision adequately addresses the concerns raised in my previous reviews, and I have no further remarks.

Response to the Reviewers

We sincerely appreciate the incisive comments and suggestions of the reviewers, which have greatly contributed to improving our manuscript. In response to their feedback, we have carefully re-analyzed the data and revised the text. In the point-by-point responses below, the reviewers' comments are shown in black, and our responses are in blue. All revised sentences in the manuscript are highlighted in yellow, and are also shown in this response letter with page and line numbers.

Reviewer #1:

Comment: In this study, Suzuki and colleagues present their production and analysis of near-complete, high-quality, phased genome assemblies of 10 Japanese male individuals. The authors used Nanopore ultra-long reads, PacBio HiFi reads, and Omni-C long-range reads to generate these data, focusing their analyses on medically relevant complex regions containing segmental duplications. They provide general statistics describing the extent of structural variation in these regions and provide more details for the SMN and KIR loci. While they demonstrate an improvement in assembly quality in these regions compared to other pangenome datasets, the analysis of these regions remains somewhat limited.

Overall, the authors identify genes with notable differences in copy number between the Japanese population and the T2T-CHM13 reference, as well as East Asian haplotypes. They also report the discovery of new haplotypes in the SMN and KIR regions. However, given the inherent instability of these regions, the identification of additional haplotypes with further sequencing is to be expected.

This study contributes to ongoing efforts to generate and analyse near-complete genome assemblies from individuals of different populations. While the dataset represents a valuable resource, certain analyses could be further refined.

Answer: Thank you very much for the constructive comments. We have revised the manuscript following all suggestions, including clarification and expansion of analyses related to complex regions, satellite arrays, segmental duplications, and gene conversion.

Comment: Specific comments

Page 3, line 99

Could the authors provide the auN (area under the Nx curve) values for these haplotypes?

Answer: We calculated auN values for the haplotypes and included them in Supplementary Table 4. In addition, we have added the following sentence to the main text.

Page 3, Lines 101–102:

"(...), and the mean contig auN (area under the Nx curve) value per haplotype was 127.8 Mbp (**Fig. 1a**)."

Comment: Pages 3-4, lines 105-113

The authors state: "The assembly size per chromosome was consistent with the T2T-CHM13 reference genome, with the exception of the acrocentric chromosomes (chr13, 14, 15, 21, 22) and chromosome Y." However, based on Figure 1c, I observe that chromosomes 9 and 16 also show a reduced assembly size compared to the T2T-CHM13 reference. Additionally, while most acrocentric chromosomes are indeed shorter than those in the T2T-CHM13 assembly, there are some exceptions.

The authors attribute the larger variance in chromosome size of chromosomes 1 and 9 to "large structural rearrangements of HSat2/3 array length variations and inversions in centromeres, as previously reported." However, inversions do not affect chromosome size, as they alter only the sequence orientation without changing its length. The mention of segmental duplications (Figure 1f) is also unclear, as these are distributed across all chromosomes, not just chromosomes 1 and 9.

To clarify this point, I suggest revising the text and figure citations accordingly. Additionally, it might be helpful to include a plot displaying the total HSat2/3 length across chromosomes and haplotypes.

Answer: We thank the reviewer for the thoughtful comments. In response, we have revised the relevant sentences in the manuscript to more accurately describe chromosome size variation and its relationship to satellite array lengths, based on the reviewer's observations and additional quantitative metrics (Page 4, Lines 107–120). Specifically, we have clarified that chromosomes 9 and 16, in addition to certain acrocentric chromosomes and chromosome Y, also showed reduced assembly sizes relative to T2T-CHM13. Following the reviewer's suggestion, we have removed misleading references to inversions and segmental duplications, and restricted our interpretation on assembly size variability to satellite array size variability. We have also added the total array sizes of HSat1/2/3 across chromosomes and haplotypes in Figure 1d. Furthermore, we have included new Supplementary Figure 1a–c to provide more detailed statistics of chromosome size standard deviations, z-scores of

chromosome sizes in T2T-CHM13 relative to Japanese haplotypes, and regression slopes quantifying how differences in satellite array length account for chromosome size differences between Japanese haplotypes and T2T-CHM13. Corresponding descriptions have been added to the Results and Methods sections as well as figure legends.

Page 4, Lines 107–120:

"The assembly size per chromosome was generally consistent with the T2T-CHM13 reference genome, with several exceptions (**Fig. 1c**). Chromosomes containing large HSat2 and HSat3 arrays—chromosomes 1, 9, 16, and Y as well as the acrocentric chromosomes (13, 14, 15, 21, and 22)—exhibited greater assembly size variability, with standard deviations exceeding 2 Mbp across Japanese assemblies (**Fig. 1d**, **Supplementary Fig. 1a**). Among these chromosomes with large size variability, we observed reduced assembly sizes more than one standard deviation than the T2T-CHM13 assembly size for chromosomes 9, 13, 15, 16, and Y in Japanese haplotypes (**Supplementary Fig. 1b**). For chromosomes 1, 9, 16, and Y, the differences in total chromosome size compared to T2T-CHM13 were largely explained by differences in the lengths of alpha satellite and HSat1/2/3 arrays[37,38], as indicated by regression slopes exceeding 0.8 between total chromosome size differences and array length differences (**Supplementary Fig. 1c**). Note that both the acrocentric chromosomes and chromosome Y showed relatively low T2T assembly rates below 30% (**Supplementary Table 4**), implying that assembly difficulties also contribute to the assembly size variability observed."

Page 20, Lines 689–692 (Methods):

"Annotations of centromeric alpha satellite and HSat1 arrays were derived from the output of RepeatMasker (http://www.repeatmasker.org, v4.1.5). Centromeric HSat2 and HSat3 sequences were difficult to distinguish and were therefore annotated using the Assembly_HSat2and3_v3.pl script (https://github.com/altemose/chm13_hsat, v3)."

Page 29, Lines 947–949 (Figure 1):

"**d**, Total array size distribution of large centromeric satellites (alpha satellites and HSat1/2/3) per chromosome. For HSat1/2/3, only chromosomes that contain >1 Mbp of these arrays in T2T-CHM13 are shown."

Comment: Page 4 - Segmental duplications

Could the authors provide additional details on Japanese-specific segmental duplications? Specifically:

Where are these duplications located?

Which genes are found in these regions?

Is copy number variation observed across different haplotypes?

Answer: We analyzed the locations and genes contained within the Japanese-specific segmental duplications identified from the 20 Japanese haplotypes. While 72% of these duplications were located in centromeric regions, we identified 28 protein coding genes that showed copy number variation within these regions, as well as 9 additional genes for which no copy number variation was detected. We have added explanatory sentences to the main text (Page 4, Lines 138–141), Supplementary Figure 1d to show the genome-wide distribution of the Japanese-specific segmental duplications, and Supplementary Table 6 to list the protein coding genes found in these regions.

Page 4, Lines 138–141:

"While many of these Japanese-specific SDs (18.7 Mbp, 72%) were located in centromeric regions (**Supplementary Fig. 1d**), we also identified 28 protein coding genes that exhibited copy number variation among the Japanese haplotypes (**Supplementary Table 6**)."

Comment: Page 4, line 130 and Supplementary Figure 1

The plots display data for structural variants (SVs) up to 10 kbp in size. Have the authors identified SVs larger than 10 kbp?

Answer: We have indeed identified SVs larger than 10 kbp, with the largest insertion and deletion sizes detected being 97 kbp and 99 kbp using minigraph-cactus, and 97 kbp and 59 kbp using pggb, respectively. We have added the distributions of SVs greater than 10 kbp to Supplementary Figures 2b and 2c, which were originally presented as Supplementary Figure 1.

Comment: Page 5, line 146

The authors report the number of shared and Japanese-specific SVs. To provide a more comprehensive view, I suggest also reporting the total number of base pairs covered by these SVs. This would help account for sequence matches and provide insight into the impact of SV size.

Answer: We have added to the main text the total number of bases corresponding to the shared and Japanese-specific SVs.

Page 5, Lines 159–162:

"Among 37 k SVs (19.6 Mbp in total) detected in two or more haplotypes and thus considered reliable,

15 k (41%, 7.7 Mbp) were shared with the HPRC graph, primarily with EAS haplotypes, while 22 k (59%, 11.9 Mbp) were specific to the Japanese haplotypes (**Fig. 1h**)."

Comment: Pages 5 and 18-19

The criteria for gene copy number analysis are not entirely clear. Specifically:

How was gene copy number calculated?

How were the 26 "exceptional genes" (shown in Supplementary Figure 2) selected?

Additionally, I recommend including all "exceptional genes" in Figure 2a and all genes with copy number differences between JPT and EAS in Figure 2b, rather than showing only a representative subset. Some genes, such as USP17L11 and certain chrX genes, may not be truly exceptional, as the T2T-CHM13 copy number falls within the JPT and EAS distribution. Similarly, for USP17L11, a copy number difference of 1 between JPT and EAS is relatively minor given its average copy number of ~22. I suggest revisiting the criteria to define exceptional genes.

Answer: We have added a sentence describing how gene copy number was calculated for each haplotype (Page 6, Lines 189–192), which was based on liftover of GENCODE Comprehensive annotations (including genes in alternate loci of complex regions) to each haplotype. We have also revised the criteria for selecting genes shown in Figure 2a and 2b. For panel a, we now require that the T2T-CHM13 copy number fall outside the 95% confidence interval of the copy number distribution across Japanese haplotypes. For panel b, we now require a minimum copy number difference between JPT and EAS haplotypes greater than the standard deviations of these datasets. As a result of these changes, *USP17L11* has been excluded under the revised criteria. Additionally, we now include all genes meeting the revised criteria in Figure 2a and 2b, instead of showing only a representative subset.

Page 6, Lines 189–192:

"We calculated the copy number for each gene by lifting over the GENCODE Comprehensive annotation[45], which includes genes in alternate loci of complex regions, to each of the haplotypes."

Page 6, Lines 196–197:

"We investigated genes whose copy numbers were consistently different from T2T-CHM13 in all the Japanese haplotypes, and there existed 23 such genes (**Fig. 2a**)."

Page 31, Lines 964–971 (Figure 2):

"**a**, Copy number distributions of genes whose copy numbers differ from T2T-CHM13 (black) in all

Japanese (JPT) haplotypes (blue), along with those in the HPRC Year-1 EAS haplotypes (red). To highlight deviations from T2T-CHM13, genes whose copy numbers in T2T-CHM13 fell within the 95% confidence interval of the mean copy number in the JPT and EAS distributions were excluded. **b**, Copy number distributions of genes with a copy number difference of one or more between JPT and EAS haplotypes. To emphasize these differences, genes with copy number differences smaller than the standard deviations of the JPT and EAS distributions were excluded."

Comment: Page 7, line 234

The authors use a threshold of 3 SNVs per region, but these regions vary greatly in size (from 150 kbp to 6.7 Mbp). I suggest using a frequency-based threshold (e.g., SNVs per kb) rather than an absolute SNV count, to account for differences in region length.

Answer: We have revised the threshold from an absolute SNV count to a frequency-based criterion of up to one SNV per 100 kbp (corresponding to QV >= 50), in order to account for variation in region length and to better reflect the quality of the assembled haplotypes. The corresponding sentence in the main text has been updated accordingly (Page 8, Lines 255–256). As a result of this change, the overall haplotype reconstruction rate slightly increased from 90.5% to 91.2%, and this number has been updated in the manuscript. We have also made minor adjustments to Figure 2e and 2f, and Supplementary Figures 4 and 6 to reflect this revision.

Page 8, Lines 255–256:

"We accepted up to one SNV per 100 kbp (i.e. QV >=50) to account for systematic errors in sequencing and variant detection, (...)"

Comment: Page 8, lines 256-258

The authors suggest that complex regions are better assembled in Japanese genomes. Could this be due to lower structural complexity, or are other factors at play? One way to explore this is by comparing the size of these regions in the HPRC year-1 dataset and the Japanese genomes.

Answer: We compared the size distributions of the complex regions between the HPRC Year-1 assemblies and our Japanese assemblies, focusing only on regions without sequence gaps. The resulting distributions are shown in Supplementary Figure 5. No consistent size differences were observed between the two datasets. This indicates that the improved contiguity reported in our Japanese assemblies is not due to reduced structural complexity, but rather reflects advancements in sequencing

technologies, particularly ONT ultra-long reads, and assembly algorithms. We have added a corresponding explanation to the manuscript (Pages 8-9, Lines 281–287).

Pages 8-9, Lines 281–287:

"The sizes of the assembled sequences without gaps in these complex regions were comparable between HPRC Year-1 and our assemblies, with no overall tendency toward shorter sequences in Japanese haplotypes (**Supplementary Fig. 5**). This suggests that these complex regions were successfully assembled in our samples not because of reduced intrinsic sequence complexity but due to advancements in sequencing technologies, particularly ONT UL reads, and assembly algorithms."

Comment: Page 10 - Gene conversion

In Figure 4a, the color of the dots (blue/red) represents the direction of gene conversion, but some dots do not fully match the donor/acceptor profile in the graph below. Could the authors clarify this apparent discrepancy?

Answer: We have revised the plot and legend of Figure 4 to clarify the meaning of the four colors representing the direction and sequence type of gene conversion. We have added a schematic figure illustrating how gene conversion events are depicted in the plots, as the new Figure 4a (Page 35, Lines 1011–1016). In the bar plot below each dot plot, the donor profile corresponds to the vertical aggregation of colored dots in the dot plot, while the acceptor profile corresponds to the horizontal aggregation. This better explains how the colored markers relate to the donor and acceptor site frequencies shown in the accompanying bar plot. We have also added this clarifying sentence to the figure legend (Page 35, Lines 1025–1026).

Page 35, Lines 1011–1016 (Figure 4):

"**a**, Schematic description of the two-dimensional representation of putative interlocus gene conversion events. In the self dot plot of the T2T-CHM13 genome sequence, off-diagonal lines are segmental duplications or other types of repeats. Putative gene conversion events found in each of the 20 Japanese haplotypes were mapped onto T2T-CHM13 and combined into the single dot plot. Each colored dot indicates a gene conversion event from a donor site ("d") to an acceptor site ("a"), as indicated by green arrows."

Page 35, Lines 1025–1026 (Figure 4):

"The donor profile corresponds to the vertical aggregation of colored dots in the dot plot, while the acceptor profile corresponds to the horizontal aggregation."

Comment: In Figure 3c, six haplotypes have two SMN2 copies and zero SMN1, suggesting a SMN1 → SMN2 gene conversion event. However, the authors do not report any gene conversion between SMN genes in their analysis. Could they clarify this apparent discrepancy?

Answer: We are grateful to the reviewer for carefully examining Figure 3c and for pointing out this important observation. We found that *SMN* gene annotations of these six haplotypes were mistakenly drawn in the plot due to a bug in the code for plotting Figure 3c. Specifically, for the *SMN* region we intended to adopt specialized annotations based on a specific paralogous SNV (c.840C>T) rather than automatic annotations by liftover of GENCODE annotations, which are often not optimal for discriminating subtle differences between *SMN1* and *SMN2*. However, in the plotting code for Figure 3c, annotations of only haplotypes with three *SMN* copies were updated to the specialized annotations due to the bug, and the automatic liftover annotations were drawn for the other haplotypes with one or two *SMN* copies. We have now thoroughly reviewed the code and text, and confirmed that this issue was specific to the plot of the *SMN* region and that the main text is not affected by this error since we did not mention haplotypes with one or two *SMN* copies in the manuscript. We have corrected the code (available on Zenodo at https://doi.org/10.5281/zenodo.14160240), Figure 3c, and the same gene annotation plot of the *SMN* region in Supplementary Figure 6 in the revised manuscript.

Comment: Page 10, lines 349-353

This section would be more appropriately placed in the Discussion. Similarly, there are other parts in the Results that would be better suited for the Discussion section.

Answer: We have moved the relevant sentences from Results to Discussion (Page 12, Lines 418–424). In addition, other sentences containing biological speculations have also been relocated from Results to Discussion for better contextual placement (Page 12, Lines 424–428; Pages 12–13, Lines 430–434).

Page 12, Lines 418–428:

"Several studies have suggested that variations in the *SMN1* and *SMN2* genes alone cannot fully account for the severity of SMA symptoms, and that the broader genomic environment including nearby (pseudo-)genes such as *NAIP*, *SERF1*, *GTF2H2*, *OCLN*, and *GUSBP* may play a role in explaining the discrepancy[24,58,62]. Our complete *SMN* haplotypes and findings on putatively biased gene conversions among these genes would help facilitate a better understanding of associations between sequence composition and SMA symptoms, particularly in a Japanese population. Note that, the absence of haplotypes with three-copy *SMN* genes in previous pangenome projects likely reflects

assembly challenges rather than their true absence. This suggests that highly complex haplotypes may be under-represented in current pangenomes and emphasizes the importance of complete assemblies of complex genomic regions."

Pages 12–13, Lines 430–434:

"These observations on putatively biased gene conversion events suggest a potentially non-random evolutionary process in Japanese haplotypes within these highly complex segmental duplications, and will be further validated as more complete haplotypes from diverse global populations become available in the near future."

Comment: Page 19, lines 625-628

A difference between two paralogous copies would be better described as a paralogous sequence variant rather than a polymorphism.

Answer: We have revised the word "polymorphism" to "variant" accordingly (Page 20, Line 668). In addition, we have made the same change at two other locations in the manuscript (Page 10, Line 332; Page 14, Line 468) for consistency and accuracy.

Reviewer #2:

Comment: The authors have generated near-complete haplotypes from 10 Japanese individuals. This completeness in medically relevant regions can help understand disease-associations and predispositions in a better way. While I appreciate the overall work and acknowledge the usefulness of these haplotypes as an informative data resource, I think that the analyses and findings presented in the manuscript need further elaboration and refinement.

Answer: Thank you very much for the valuable feedback. We have revised the manuscript to clarify and expand our analyses and descriptions in accordance with the reviewer's suggestions.

Comment: Major comments:

1). The authors have compared their results to HPRC year 1 and Chinese pangenome. They mentioned in the Discussion about HGSVC's new assemblies for 65 individuals in the context of SMN genes, however, for the rest of the comparisons these assemblies have not been included.

Answer: We have now included a comparison of the complete haplotype reconstruction rates between HGSVC's new assemblies for 65 individuals and our Japanese assemblies, using the *SMN* region as a representative case (Page 13, Lines 448–451). We have also incorporated a comparison of haplotypes in the *KIR* region (Page 13, Lines 452–458). Specifically, we generated a complete list of *KIR* gene annotations, as well as classification results of either A-haplotypes (relatively common) or B-haplotypes (more diverse), for each of the 127 HGSVC haplotypes in which *KIR* genes were contained in a single contig. We have provided this information as Supplementary Table 9 in the revised manuscript. We compared these haplotypes with those from our Japanese assemblies and found that only one of the four B-haplotypes identified in our dataset was also present in HGSVC3 assemblies, whereas the remaining three B-haplotypes, including the novel one highlighted in Figure 3c, were absent. We have revised the main text accordingly and added sentences describing this additional analysis.

Page 13, Lines 448–458:

"Their assemblies achieved a level of haplotype completeness comparable to ours. For example, they reported 64% completeness in the *SMN* region, compared to our 85%, using slightly different but broadly comparable evaluation criteria, both of which estimated the completeness of the HPRC Year-1 pangenome as approximately 20%. Among 127 haplotypes of HGSVC3 in which *KIR* genes were contained in a single contig, 73 (57.5%) were A-haplotypes and 54 (42.5%) were B-haplotypes

(**Supplementary Table 9**), compared to 16 (80%) A-haplotypes and 4 (20%) B-haplotypes among our Japanese haplotypes. Although one B-haplotype identified in our Japanese pangenome was also present in the HGSVC3 assemblies, the novel B-haplotype with two tandem copies of the *KIR2DL5B*, *KIR2DS5*, and *KIR2DS1* genes (**Fig. 3a,b**), as well as the remaining two B-haplotypes, was absent from the HGSVC3 dataset."

Comment: 2). Mentioning the HGSVC assemblies, the authors state 'Among the 130 haplotypes generated from worldwide individuals, they found only two haplotypes with three copies of SMN genes, while we found two such haplotypes among 20 Japanese haplotypes.' I am not sure what they aim to convey with that statement. If it is to convey that the respective haplotype is more common in the Japanese population, it needs to be supported by statistical testing particularly since the number of individuals is so different.

Answer: We agree that the original sentence was ambiguous and also misleading as it compared a single population (Japanese) and a globally diverse population set (HGSVC3). To address this concern, we have revised the text to instead emphasize the broader importance of expanding (near-)complete assemblies across diverse populations to better capture the full spectrum of genomic diversity (Page 13, Lines 458–465).

Page 13, Lines 458–465:

"Additionally, just as we observed two haplotypes with three copies of *SMN* genes among the 20 Japanese haplotypes (**Fig. 3c,d**), they also found two such haplotypes from one African and one admixed American among the 130 haplotypes. Considering that we did not observe any assembled haplotypes with three *SMN* gene copies in the HPRC Year-1 and Chinese pangenomes, these observations highlight the importance of expanding (near-)complete assemblies across diverse populations, including well-characterized populations, to accurately uncover hidden genomic diversity in medically important complex regions."

Comment: 3). The authors talk about overlaps of the SDs, SNVs and SVs with other data sets, however, it is not clear to me how this overlap was computed. As the uniqueness of these events , particularly SVs, relies heavily on the used matching criteria, it needs to be clearly stated in the manuscript.

Answer: We thank the reviewer for pointing out the importance of clarifying the criteria used to determine overlaps of SDs, SNVs, and SVs. For segmental duplications (SDs), we lifted over the detected SD regions from each Japanese haplotype to the T2T-CHM13 reference genome. Overlaps

of SDs between Japanese haplotypes and the SD annotation of T2T-CHM13 at base-level resolution were computed using the "bedtools intersect" command. Overlaps of SDs between Japanese haplotypes and East Asian haplotypes were calculated by decomposing the SDs into sub-regions at base-level resolution based on the number of haplotypes containing each sub-region using the "bedtools multiinter" command. For SNVs and SVs, we used a multi-sample vcf file generated from the pangenome graph using T2T-CHM13 as the reference. We have provided detailed commands and tool versions in the Methods section (Page 19, Lines 641–647). Two SNVs or SVs were considered identical only if they shared the same genomic location and alternative allele sequence. We have revised the relevant sentences in the manuscript to clearly describe these criteria (Page 4, Lines 126–132; Page 5, Lines 145–148; Page 5, Lines 162–166). The scripts used for these analyses are also available on Zenodo at https://doi.org/10.5281/zenodo.14160240.

Page 4, Lines 126–132:

"Segmental duplications (SDs) were identified in each Japanese haplotype and lifted over to the T2T-CHM13 genome, which covered a total of 131.1 Mbp (**Fig. 1f**). These SD regions were compared with the SD annotation of T2T-CHM13 itself using the "`bedtools intersect`" command, and 100.9 Mbp (77%) overlapped with T2T-CHM13's SD regions, (...). The SD regions were further decomposed into sub-regions at base-level resolution based on the number of haplotypes containing each sub-region using "`bedtools multiinter`"."

Page 5, Lines 145–148:

"SNVs and SVs, defined as variants of >50 bp, for each haplotype were identified from the pangenome graph by detecting branching nodes that diverge from the reference genome path of either T2T-CHM13 or GRCh38. Two SNVs or SVs were considered identical if they shared the same genomic location and alternative allele sequence."

Page 5, Lines 162–166:

"Note that the total number of SV bases may be overestimated, as SVs differing by even a single nucleotide were counted as distinct, despite our efforts to perform base-level normalization of nodes in the pangenome graph (**Methods**). Nonetheless, these results suggest a considerable proportion of population-specific SVs."

Page 19, Lines 641–647 (Methods):

"Variants in each haplotype with respect to T2T-CHM13 (or GRCh38) were called from the graph using vg (v1.57.0)[75], vcfbub (v0.1.1)[76], and vcflib (v1.0.9)[77]. Specifically, we first generated a vcf file containing all variants inferred from the pangenome graph using the command "`vg deconstruct -`

P CHM13 -e -a *pggb_graph*.gfa". Then, the resulting variants were normalized with "vcfbub -l 0 -a 100000 --input *pggb_graph*.vcf.gz | vcfwave -I 1000". The variants were further normalized and sorted using bcftools (v1.18)[78]."

Comment: 4). There have been multiple mentions of 'recurrent' events, like SDs, and SNVs. e.g. it is stated that 74% of the SNVs are recurrent in the Japanese population. But it is not clear how the recurrence of those events was determined? As recurrence estimates normally require large cohorts, it should be explained clearly how did the authors check for it in their cohort of 10 samples and how confident are these findings?

Answer: We used the term "recurrent" in the original manuscript to refer to SDs or variants that were observed in two or more haplotypes among the 20 Japanese haplotypes analyzed. We agree that demonstrating true recurrence at the population level would require a much larger cohort. To avoid potential misunderstanding, we have clarified this in the manuscript by adding a sentence that highlights the limitation of inferring recurrence from our current sample size (Page 4, Lines 134–136). In addition, we have revised the word "recurrent" to "detected in two or more" and "putatively recurrent" at two other locations of the manuscript (Page 5, Lines 155–157; Page 5, Line 160).

Page 4, Lines 134–136:

Within the Japanese haplotypes, 104.1 Mbp (79%) of the SDs were found in two or more haplotypes, suggesting that these SDs could be common in the Japanese population. It should be noted, however, that confirming their recurrence at a population scale will require a larger cohort study.

Comment: 5). There are a lot of statements in the manuscript that need a quantitative support. For example:

a). 'MANY segmental duplications were resolved without gaps.' on line 98.

b). 'Although accurate annotation of paralogous genes in complex regions is difficult, MOST of such genes belong to gene families in segmentally duplicated complex regions' in line 176.

c). 'The average size of each region among the complete Japanese haplotypes was MOSTLY consistent with T2T-CHM13, while we identified considerable size deviations in several haplotypes indicating large-scale rearrangements' in line 241.

Instead of saying many/most etc., quantitative assessments should be provided.

Answer: We thank the reviewer for pointing out the need for quantitative support in several statements.

We have revised the ambiguous expressions and replaced vague terms such as "many" and "most" with specific values, as follows.

Page 3, Lines 98–100:

"(...), an average of 63.7 Mbp of segmental duplication regions per haplotype were resolved without gaps."

Page 6, Lines 198–200:

"(...), all of those 23 genes were either gene families in segmental duplications or genes within alternate loci of complex regions whose names start with "ENSG", (...)"

Page 8, Lines 261–266:

"For 24 (80%) of the complex regions, the size in T2T-CHM13 was not an outlier in the distribution of the complete Japanese haplotypes, indicating overall consistency between Japanese haplotypes and T2T-CHM13 (**Fig. 2e**). In contrast, we identified considerable size deviations in the remaining 6 regions: the *NOTCH2-NOTCH2NLR-SRGAP2*, *FCGR*, *DEFB-FAM90A*, *OPN1LW-OPN1MW*, *TSPY*, and AZFc regions (**Fig. 2e**)."

Comment: d). 'Based on the copy number analysis and the literature' in line 213. I didn't understand which analysis or literature was being referred to.

Answer: We apologize for the lack of clarity in the original sentence. To address this, we have added references to clarify which prior studies were being referred to (Page 7, Lines 232–233). We have also made explicit the connection to our copy number analysis (Page 7, Lines 237–239). In addition, we have included reference numbers for each complex region in Supplementary Table 8.

Page 7, Lines 232–233:

"Based on the literature analyzing challenging complex genomic regions in the assemblies of T2T-CHM13, GRCh38, and previous pangenome efforts[7,9,16–18,46,48–50], (...)"

Page 7, Lines 237–239:

"These include genes such as *FAM90A*, *ELOA3*, *DEFB*, *SPDYE*, *KIR*, *GAGE*, *TSPY* and *RBMY*, which were noted in our copy number analysis."

Comment: e). 'For instance, in this study we observed that the copy numbers of the TSPY3 and TSPY4

genes were zero or one in some East Asian haplotypes in the HPRC Year-1 dataset, compared to 8–14 and 2–6, respectively, in our Japanese haplotypes'. This statement in line 390 is confusing unless one looks at Figure 2b.

Answer: Thank you for pointing this out. We have added a reference to Figure 2b in the sentence (Page 12, Line 406).

Comment: Minor Comments:

1). In Figure 3a, the arrow colors in the highlighted haplotypes are not consistent with the legend.

Answer: We have revised the legend of Figure 3 to clarify the definitions of two types of *KIR* genes shown (Page 33, Lines 988–990). We also modified Figure 3a to eliminate overlap between highlighting colors of haplotypes and colors of gene arrows. Specifically, instead of using color highlights, we now use black frames to indicate the focal haplotypes for better visual clarity.

Page 33, Lines 988–990 (Figure 3):

"*KIR* genes characterizing *KIR* B-haplotypes (denoted as "*KIR* (B)") are shown in purple, whereas the remaining *KIR* genes ("*KIR* (A)") are shown in green."

Comment: 2). In Figure 4, the x-axis label is not center-aligned. Additionally, warm vs cold vs darker color distinction in the caption is adding unnecessary complexity. Since it's just 4 colors, I would recommend mentioning what each of them corresponds to in the figure legend. Additionally, instead of saying left-bottom triangle and right-top triangle I would suggest just saying below and above the diagonal.

Answer: Thank you for the helpful suggestions. We have center-aligned the x-axis labels and revised the legend of Figure 4 to explicitly describe the meaning of each of the four colors used in the plot (Page 35, Lines 1016–1020). To simplify the explanation, we have removed the use of the terms "warm", "cold", "darker", and "lighter" to describe the four colors. Additionally, we have replaced the terms "left-bottom triangle" and "right-top triangle" with "below the diagonal" and "above the diagonal", respectively, for clarity.

Page 35, Lines 1016–1020 (Figure 4):

"Specifically, gene conversions occurring in the 5'-end to 3'-end direction are plotted below the diagonal, with dark blue for protein coding genes and light blue for other sequences. Those in the 3'-end to 5'-end direction are plotted above the diagonal, with red for protein coding genes and yellow

for other sequences."

Comment: 3). Line 228, Typo ‘detest’.

Answer: Thank you for finding this out. We have corrected the typo to "detect" (Page 7, Line 249).

Comment: 4). Line 141, it should me made clear what do authors mean by ‘non-private SVs’.

Answer: To improve clarity, we have revised the wording to explicitly state what is meant by “non-private SVs.

Page 5, Line 160:

"SVs (19.6 Mbp in total) detected in two or more haplotypes"

Comment: 5). Line 397, it should be clearly stated how a ‘minor’ haplotype is quantitatively defined.

Answer: We have clarified the meaning of "minor" haplotypes by adding the corresponding population frequencies to the relevant sentences as follows.

Page 12, Lines 411–417:

"In the *KIR* region, we fully assembled one non-canonical B-haplotype carrying a previously unreported combination of genes: two copies of the *KIR2DL5B, KIR2DS5*, and *KIR2DS1* genes. B-haplotypes in the *KIR* regions have been reported to exist only in ~25% of Japanese haplotypes[21], and our findings highlight the utility of long-read sequencing in uncovering novel B-haplotypes even in control samples. In the *SMN* region, we obtained two complete haplotypes carrying two copies of the *SMN2* gene, which is a copy number profile reported in 2.4% of individuals in the East Asian population[23]."

Comment: 6). Supplementary Figure 1 caption typo ‘insersion’.

Answer: Thank you for finding this out. We have corrected the typo and other typos in the captions of Supplementary Figures.

Response to the Reviewers

We sincerely appreciate the continued and valuable feedback from the reviewers. In response, we have carefully addressed all remaining concerns, re-analyzed the relevant data, and revised the text accordingly. In the point-by-point responses below, the reviewers' comments are shown in black, and our responses are in blue. Revised sentences in the manuscript are highlighted in yellow, and are also shown in this response letter with page and line numbers.

Reviewer #1:

Comment: The authors have addressed all my comments. I have no additional remark.

Answer: We thank the reviewer for acknowledging our revisions. We appreciate the reviewer's time and efforts in reviewing our manuscript.

Reviewer #2:

Comment: I appreciate the authors' efforts in addressing my previous comments. After reviewing the responses and the revised manuscript, I have a few additional comments and questions. I therefore recommend another round of revision.

Answer: We are grateful for the careful review and insightful comments. We appreciate the reviewer's time and efforts in reviewing our manuscript. Following the reviewer's suggestions, we have expanded our analyses and revised the manuscript.

Comment: 1) Firstly, I would reiterate my previous comment about avoiding any unclear or ambiguous statements. e.g. in the revised manuscript on line 449 it is stated "For example, they reported 64% completeness in the SMN region, compared to our 85%, using slightly different but broadly comparable evaluation criteria". As a reader, I am curious to know what the exact criteria is.

Answer: For our Japanese haplotypes, the exact criteria is described in the manuscript (Page 8, Lines

260–266, 269–270) as follows:

> "Structural quality is the percentage of regions where sequencing coverage dropped below three, calculated from a pileup of ONT ultra-long reads of the same sample. Base quality is the number of SNVs detected from the read pileup. (...) We accepted up to one SNV per 100 kbp (i.e. QV >=50) to account for systematic errors in sequencing and variant detection, while no sequence gaps and no coverage drops were accepted as the best category, as shown in Fig. 2c, d."
>
> "We then defined "complete haplotypes" as haplotypes that belong to the best category in all the three criteria above (sequence contiguity, structural quality, and base quality)."

For the HGSVC3 assemblies, the criteria are detailed in Methods and Supplementary Information of their paper. They applied multiple orthogonal tools, including Flagger, Merqury, NucFreq, and Inspector, to annotate potential assembly errors. Flagger detects structural errors based on read depth, Merqury and NucFreq identify erroneous bases using k-mers and read-to-assembly mappings, and Inspector detects both structural and base errors through read alignment.

In the original sentence, our intention was to emphasize that although the approaches differ, both studies share the same fundamental goal of evaluating assemblies in terms of structural and base-level accuracy. We have revised the sentence in the manuscript to clarify these differences and to avoid ambiguity (Pages 10–11, Lines 351–359).

Pages 10–11, Lines 351–359:

"For instance, in the *SMN* region they reported 64% completeness, while we estimated 85% for our Japanese haplotypes using our definition of completeness based on sequence contiguity, structural quality, and base quality. HGSVC3 evaluated assembly quality by integrating multiple orthogonal tools including Flagger[7], Merqury[36], NucFreq[64], and Inspector[65] that collectively assess both structural and base-level errors, requiring that at least two tools consider the assembly reliable (see HGSVC3 Supplementary Methods). Both studies share the fundamental goal of evaluating haplotype assemblies in terms of base-level accuracy and structural integrity. Consistent with this, both studies estimated the completeness of the HPRC Year-1 pangenome as approximately 20%."

Comment: 2) Regarding the comparison of SDs, the authors have mentioned that they used bedtools intersect to compute the overlap but I couldn't find the details about how the SDs were identified in each Japanese haplotype before liftover. The authors have now mentioned the total base pairs but it is not clear to me how much of the 30.2 Mbp (unique region) corresponds to SDs where no overlap was observed and how much of it comes from the cases where only a certain fraction of the region overlaps

because I wouldn't consider the sequence belonging to the latter category necessarily as "unique". I would suggest that each step of this analysis should be explicitly mentioned.

Answer: We have added a description of how SDs were identified to the main text (Page 4, Lines 126–127). We have also restructured the Methods section and made a separate section on SD analysis with detailed descriptions (Page 19, Lines 651–663). In addition, following the reviewer's suggestion, we now require unique SDs to be completely non-overlapping with those of other haplotypes. This change, together with the exclusion of centromeric regions described below, reduced the size of unique SDs compared to T2T-CHM13 from 30.2 Mbp to 2.1 Mbp. We have revised the text (Page 4, Lines 131–133) and updated other relevant values in the manuscript.

Page 4, Lines 126–127:

"Segmental duplications (SDs) were identified using Biser[42] after masking repeats with RepeatMasker (http://www.repeatmasker.org) in each Japanese haplotype, and (...)"

Page 4, Lines 131–133:

"This resulted in a total of 46.3 Mbp of SD regions, of which 44.2 Mbp (95%) showed complete or partial overlap with SDs annotated in T2T-CHM13, while the remaining 2.1 Mbp (5%) did not."

Page 19, Lines 651–663 (Methods):

"**Segmental duplication analysis**

Segmental duplications (SDs) were identified from each haplotype using Biser (v1.4)[42] after masking repeats with RepeatMasker using parameters "`-s -xsmall -e ncbi -species human`". Only SDs longer than 1 kbp and with sequence similarity greater than 90% were retained. The resulting SDs were lifted over to T2T-CHM13 coordinates for comparison with other haplotypes using minimap2 (v2.24)[48], and merged across haplotypes using "`bedtools merge`" for the Japanese, HPRC Year-1 (EAS and non-EAS), and Chinese pangenomes. SDs found in centromeric regions were excluded due to their highly repetitive nature. Comparison of the lifted-over Japanese SD regions with SDs in T2T-CHM13 and other pangenomes was performed using the "`bedtools intersect -v`" command, which returns completely non-overlapping SDs. The number of haplotypes covering the SD regions was calculated using the "`bedtools multiinter`" command, which takes SD regions of multiple haplotypes and reports the sub-regions shared by the same set of haplotypes."

Comment: Line 136 says "While the majority of SDs identified in the Japanese haplotypes (105.3 Mbp, 80%) were also present in non-Japanese East Asian (EAS) haplotypes in the HPRC Year-1 dataset, 25.8 Mbp (20%) were specific to Japanese haplotypes". I find this sentence confusing. Does this mean

that the SD comparison has been performed only across the non-Japanese EAS haplotypes in the HPRC cohort? If yes, why were the rest of the samples excluded? If not, the sentence should be rephrased to make it explicit.

Since the other comparisons presented in the paper have been performed against both HPRC and Chinese pangenomes, I would suggest the authors do the same for SD comparison. Additionally, I don't consider a comparison only with CHM13 as insightful as you are basically comparing to a single haplotype. So, I would suggest restructuring this section and mentioning a collective comparison including HPRC and Chinese pangenome.

Answer: Since the original comparison was limited to non-Japanese East Asian haplotypes of the HPRC pangenome as the reviewer pointed out, we have additionally compared the Japanese haplotypes with those of the Chinese pangenome as well as non-East-Asian haplotypes of the HPRC pangenome. East Asian haplotypes from the HPRC pangenome completely or partially overlapped with 91% of the Japanese SDs, Chinese haplotypes overlapped with an additional 5%, and non-East-Asian haplotypes contributed an additional 0.1%. While the comparison was performed using the coordinates of T2T-CHM13, the actual comparison were conducted not only with CHM13 but also with HPRC and Chinese pangenomes. We have revised the paragraph accordingly in the manuscript (Pages 4–5, Lines 136–144).

Pages 4–5, Lines 136–144:

"The majority of SDs identified in the Japanese haplotypes (42.3 Mbp, 91%) completely or partially overlapped with SD regions in the non-Japanese East Asian (EAS) haplotypes of the HPRC Year-1 pangenome, and an additional 2.2 Mbp (5%) overlapped with those in the haplotypes of the Chinese pangenome. Among the remaining non-overlapping, Japanese-specific SD regions (1.8 Mbp, 4%), only 54 kbp (0.1%) overlapped with SDs in non-EAS haplotypes of HPRC Year-1 pangenome. Although 154 out of the 155 Japanese-specific SD regions did not contain protein coding genes, one region on chromosome 19 present in a single haplotype contained 4 protein coding genes, *MAP3K10*, *TTC9B*, *CCNP*, and *AKT2* (**Supplementary Table 6**)."

Comment: Lastly, line 138 states "While many of these Japanese-specific SDs (18.7 Mbp, 72%) were located in centromeric regions". Given the highly repetitive nature of centromeres, I don't think that only sequence overlap is enough to determine the uniqueness of an SD inside a centromere. I would suggest either to use further sanity checks for these cases to make sure that these are not stemming from assembly errors etc. or exclude the centromeric regions from this comparison.

Answer: We have excluded centromere regions from the entire SD analysis, which reduced the total

size of Japanese SDs in the CHM13 coordinate system from 131.1 Mbp to 46.3 Mbp. Accordingly, we have revised the text (Page 4, Lines 129–131) and updated other relevant values in the manuscript. We have also excluded SDs in centromeres from Figure 1f, and removed Supplementary Figure 1d, which showed enrichment of Japanese-specific SDs in centromeres and is no longer necessary.

Page 4, Lines 129–131:

"We merged the lifted-over Japanese SD regions and excluded those within centromeric satellite regions, as these regions are difficult to compare across different haplotypes. This resulted in a total of 46.3 Mbp of SD regions, (...)"

Comment: 3) Line 147, page 5 "Two SNVs or SVs were considered identical if they shared the same genomic location and alternative allele sequence". For SVs, this criteria is too strict and expected to overestimate the unique SVs (more detailed response below).

Answer: As suggested by the reviewer, we have revised our analysis of SVs by adopting Truvari for comparison, which is expected to be more robust than our previous strict definition. Details of this revision are described in the following responses below.

Comment: Line 148, "Each method yielded comparable count of identified variants", seeing Supplementary figure 2, I won't say they are comparable. The distribution pattern is comparable but the numbers are not. Minigraph-cactus is clearly calling more SVs as compared to pggb. Was it stratified how many SVs are common and how many are unique to each method and how many of them are potentially false-positives? What was the rationale behind using the pggb graphs for further comparisons? I would recommend making all of this explicit in the text.

Answer: We have revised the sentence to clarify that the number of SVs differs between minigraph-cactus and pggb, although the distribution pattern is comparable. Minigraph-cactus called more SVs particularly in centromeric regions (26 k by minigraph-cactus vs. 14 k by pggb), which accounted for 96% of the difference in SV counts between the two methods. We have also analyzed corresponding SVs between the two methods using Truvari, which showed an overall 74% F1 score. We have updated the SV length distributions of the two methods (Supplementary Figure 2b,c) in which the histograms are now stratified into common SVs (with counterparts), unique SVs (without counterparts), and centromeric SVs. For further comparisons, we used the pggb graph containing fewer unique SVs to reduce potential overestimation. Additionally, we have now excluded variants within centromeres from the comparisons. These revisions and additional analyses have been incorporated into the manuscript

(Page 5, Lines 152–160).

Page 5, Lines 152–160:

"Each method yielded a comparable pattern of identified SVs, with peaks in the SV length distribution corresponding to insertions and deletions of the *Alu* element of ~340 bp and LINE elements of ~6 kbp (**Supplementary Fig. 2**), although the total number of SVs differed between minigraph-cactus (127 k) and pggb (115 k). Notably, 96% of the difference in SV counts between the two methods (12 k) was attributable to centromeric regions, where minigraph-cactus reported more SVs (26 k) than pggb (14 k). Outside centromeres, SVs identified by the two methods showed a 74% F1 score using Truvari[47]. For subsequent analyses, we used the pggb graph containing fewer unique SVs, and excluded variants located within centromeres."

Page 20, Lines 684–686 (Methods):

"The identified SVs of the Japanese haplotypes were compared between minigraph-cactus and pggb with Truvari (v5.3.0)[47] using the command "`truvari bench -r 1000 --sizemax 100000`"."

Comment: Line 159 "Among 37 k SVs (19.6 Mbp in total) detected in two or more haplotypes and thus considered reliable, 15k (41%, 7.7 Mbp) were shared with the HPRC graph, primarily with EAS haplotypes, while 59%, 11.9 Mbp were specific to the Japanese haplotypes (Fig. 1h). Note that the total number of SV bases may be overestimated, as SVs differing by even a single nucleotide were counted as distinct, despite our efforts to perform base-level normalization of nodes in the genome graph (Methods). Nonetheless, these results suggest a considerable proportion of population-specific SVs." 59% is an extremely high number keeping in view studies stratifying population specific SVs (e.g. Ebert et al. 2021). The authors have acknowledged that this is an over estimation but this can be corrected using a better comparison criteria. I would like to see how the numbers look like if you use one of the conventionally used SV comparison tools (e.g. Truvari) for this purpose.

Answer: We used Truvari for a better comparison of SVs, which reduced the percentage of Japanese-specific SVs from 59% to 42%. We have now considered SVs detected in two or more haplotypes within either Japanese or others, while previously we considered only SVs shared among Japanese haplotypes. We have revised the text (Page 5, Lines 160–161; Page 5, Lines 172–176) and updated all relevant values in the manuscript.

Page 5, Lines 160–161:

"Additionally, we merged similar SVs across all haplotypes compared using Truvari."

Page 5, Lines 172–176:

"We focused on SVs detected in two or more haplotypes within either Japanese or HPRC haplotypes and thus considered reliable. Among 44 k SVs (10.1 Mbp in total) found in Japanese haplotypes, 26 k (58%, 5.8 Mbp) were shared with the HPRC graph, primarily with EAS haplotypes, while 18 k (42%, 4.3 Mbp) were unique to our Japanese haplotypes (**Fig. 1h**), (...)"

Page 20, Lines 686–688 (Methods):

"For the comparison involving other haplotypes, similar SVs were merged with the command "`truvari collapse -S 100000`"."

Comment: 4) For saying that something is "putatively recurrent" one still needs to provide some phylogenetic evidence showing that the event occurred independently over the course of evolution or a reference to other studies which provide such evidence for these events. Presence of an allele in more than two haplotypes doesn't indicate recurrence in any way. Therefore, I would recommend either backing the recurrence claims with proper phylogenetic support or replacing the word with something like "shared" (e.g.).

Answer: We have rephrased the word "recurrent" with "shared" (Page 5, Line 169).

Comment: 5) I would suggest moving the part explaining the comparison of KIR region with HGSVC in the Discussion to the Results section where the comparison to HPRC and Chinese pangenome is mentioned. Same for the SMN region. All the results should go together so that the reader can clearly understand what the novel insights are. The overall flow of the Discussion section can also be improved by this because currently two different paragraphs are talking about both the KIR and SMN regions with a paragraph about chromosome Y in between. This hinders the reader's ability to understand the message that is being conveyed.

Answer: We have moved the part explaining the comparison of KIR and SMN regions with HGSVC from the Discussion to the Results section (Pages 10–11, Lines 349–368). We have also restructured paragraphs in the Discussion section accordingly. Specifically, we have moved the paragraph on chromosome Y (Page 12, Lines 423–429), and merged the redundant sentences on KIR and SMN in two different paragraphs into a single paragraph (Page 13, Lines 454–461).

Comment: Overall, I suggest that the authors refine and streamline the comparisons with other cohorts,

particularly regarding population-specific SDs and SVs, to avoid potential overestimations. All methodological details should be clearly described in the text. Additionally, any claims lacking quantitative support should be revised or removed. I also recommend restructuring the manuscript, especially the Results and Discussion sections, to improve clarity.

Answer: We believe that we have addressed all the suggestions. Thank you very much again for the thorough review, which we believe greatly improved the manuscript.

Response to the Reviewers

We sincerely appreciate the continued and valuable feedback from one of the reviewers. In response, we have carefully addressed all remaining concerns, re-analyzed the relevant data, and revised the text accordingly. In the point-by-point responses below, the reviewer's comments are shown in black, and our responses are in blue. Revised sentences in the manuscript are highlighted in yellow, and are also shown in this response letter with page and line numbers.

Reviewer #2:

Comment: I appreciate the efforts from the authors in addressing my comments and questions. While I am satisfied with the responses and the respective revisions in the manuscript w.r.t most of my comments, I am still not convinced about the population specific SV analysis and the reported percentage of unique SVs.

Answer: We sincerely thank the reviewer for carefully revisiting this point and for emphasizing the need to clarify the apparent discrepancy with the estimates reported previously. We agree that the proportion of population-specific SVs requires careful explanation, and we have therefore re-examined this concern in detail below.

Comment: In my previous review, I referred to the Science 2021 study 'Haplotype-resolved diverse human genomes and integrated analysis of structural variation' by Ebert et al. to provide an example of the scale at which population specific SVs are expected to arise. Figure 6C of this paper provides some estimates for these numbers based on the cohort used in this study. If I'm not mistaken there is only one Japanese individual in their cohort. So, if we want to have a very rough estimate, from the numbers in panel C, for the Japanese samples we can expect ~1000/25k i.e. around 4% population-specific SVs given we expect 25k SVs on average in a sample. Even if we give some margin of error in this estimation and account for varying sample sizes, the percentage reported by the authors i.e. 42% is extremely high and needs to be investigated.

My assumption is that it is still an SV comparison issue which might be fixed by using a stricter Truvari setting. One option is to start with the same parameter settings as used in the HPRC Nature 2023 paper for benchmarking variants. If the percentage stays this high even after using stricter comparison settings, further validation and quality control analyses e.g. stratification of the type and genomic

location of these variants needs to be performed to support the claim that 42% of the SVs carried by these 10 individuals indeed have never been reported before. Reporting only a number is not sufficient. I would suggest the authors revisit this comparison to avoid any over-estimations.

Answer: We appreciate this important point and the opportunity to clarify sources of the observed difference. As the reviewer noted, Ebert et al., Science 2021 included one Japanese individual in their long-read assemblies. However, the estimates shown in their Figure 6C are NOT derived directly from these long-read assemblies, but rather from short-read SV genotyping across 3,202 individuals using an SV catalogue constructed from the assemblies. Because distinguishing complex SV polymorphisms is particularly difficult with short reads, this analysis necessarily involved extensive SV filtering and merging steps. In particular, their pipeline employed an aggressive merging strategy based on 50% reciprocal overlap (RO), which collapses nearby SVs into a single event solely based on their size overlap, irrespective of their sequence content. This approach substantially reduces the number of population-specific SVs. To further evaluate this point, we repeated our analysis using settings that mimics the 50% RO merging strategy used in Ebert et al., Science 2021. Under this condition, the number of Japanese-specific SVs decreased from ~13 k (42% of total Japanese SVs) to ~3 k (8%), which is consistent with the order of magnitude reported in Ebert et al., Science 2021.

By contrast, our analysis directly compared SV sequences derived from long-read assemblies. Since the entire sequences of individual SVs were given, we collapsed nearby SVs based on both sequence similarity and size overlap. Indeed, we followed the parameter settings for merging similar SVs in long-read assemblies that were used in the HPRC Nature 2023 paper (Liao et al., Nature 2023) suggested by the reviewer. Specifically, both studies applied the "truvari collapse" command with its default thresholds of >95% sequence similarity and >95% size overlap to merge similar SVs in long-read assemblies. This base-pair level comparison represents a higher-resolution approach adopted in recent long-read pangenome studies (Gao et al., Nature 2023; Nassir et al., Nature Communications 2025). Such approaches yield a higher proportion (30–50%) of population-specific SVs, as they retain distinct SVs that cannot be distinguished through short-read-based genotyping. In this context, the observed difference between our approach (long-read assemblies) and Ebert et al., Science 2021 (short-read genotyping) reflects methodological differences rather than overestimation of our approach.

We believe this contrast is informative, and we have added these results together with the corresponding Supplementary Figure 2d to the revised manuscript (Page 5, Lines 161–162; Pages 5–6, Lines 178–185). We hope that these additional analyses address the concern and reassure that our reported numbers are consistent with standard long-read assembly-based SV analysis.

Page 5, Lines 161–162:

"Additionally, we merged similar SVs across all haplotypes using Truvari with default parameters, which merge nearby SVs with >95% sequence similarity and >95% size overlap."

Pages 5–6, Lines 178–185:

"Since we reconstructed high-quality sequences of individual SVs from long-read assemblies, we collapsed similar SVs with >95% sequence similarity as well as >95% size overlap. In contrast, when we applied the 50% reciprocal overlap method, which is based solely on >50% size overlap without considering sequence similarity and is often used for large-scale short-read SV genotyping[4,48,49], the proportion of shared SVs increased from 58% to 92% (35 k) and that of Japanese-specific SVs decreased from 42% to 8% (3 k) (**Supplementary Fig. 2d**), implying the importance of long-read assemblies for fully uncovering the sequence diversity of SVs."

Page 20, Lines 696–697 (Methods):

"SV merging with the 50% reciprocal size overlap was performed using the command "`truvari collapse -r 1000 -p 0 -P 0.5`"."

Response to the Reviewers

We sincerely appreciate the continued and valuable feedback from one of the reviewers. In response, we have carefully addressed all remaining concerns, re-analyzed the relevant data, and revised the text accordingly. In the point-by-point responses below, the reviewer's comments are shown in black, and our responses are in blue. Revised sentences in the manuscript are highlighted in yellow, and are also shown in this response letter with page and line numbers.

Reviewer #2:

Comment: I appreciate the effort of the authors in addressing my concerns, however, I am still not convinced that this number is not an overestimation. The authors are basically claiming that 42% of the structural variation seen in a group of 10 individuals from a human population is unique. The reference to the Ebert et al 2021 paper was just to provide a comparison from a study where such estimation has previously been provided for the Japanese population. I totally agree that the haplotype-resolved assemblies we have today provide much deeper insights into the genetic diversity as compared to short reads, however, that doesn't explain the rise to 42%. I cannot wrap my head around how it is even possible to see unique structural variation at such a scale.

I agree with the authors that in the HPRC Nature 2023 paper the same truvari parameters were used for SV merging, however, to my knowledge, this callset was not used to make any statements about population-specific uniqueness. Also, the SV comparisons in the HPRC 2023 paper were based on SV alleles not SV sites. That's why in my previous review I suggested using stricter settings for the estimation the authors are trying to make and performing further validation if the percentage still stays this high.

If we look into the other studies authors referred to for comparison: In Gao et al, the percentage of CPC-specific SVs is around 17% and most of them were located at the centromeric and telomeric regions of chromosomes. They also clearly state that more than half of the CPC-specific variants were singletons or doubletons. In comparison, the authors are reporting 42% while not even considering singletons. In the more recent Nasser et al 2025 study, around 22% SVs were classified as population-specific. So I am curious to know where the range of 30%-50% that the authors mentioned is coming from. It should also be noted that in Nasser et al. 80% reciprocal overlap and sequence identity was used for finding unique SVs as compared to 95% used by the authors which is extremely flexible and prone to overestimate the unique SVs as I have pointed out before as well.

Additionally, I would reiterate that if the authors insist that this percentage of 42% is not an overestimation, further stratification is required e.g. by variant type and more importantly by genomic location. Another experiment that would help get a better picture and assess the 'uniqueness' of these SVs is a visual depiction of the SV sequence identity with other cohorts that the authors are comparing to i.e. the Chinese pangenome and the HPRC cohort. One way can be to align the SV sequence to the respective pangenome graphs for each unique SV and plotting the resulting sequence identities together.

Overall, I would again urge the authors to revisit this analysis and be extremely careful in making statements like finding 'a considerable proportion of population-specific SVs'.

Answer: We thank the reviewer for the careful comments. The value "42%" in the original manuscript was the proportion calculated using only the Japanese SVs rather than all SVs as the denominator, which inflated the estimate. Regarding previous studies, the "30%–50%" values, compared to "17%" and "22%", were calculated in the same manner. We recalculated the proportion over all SVs, which reduced the value from 42% to 11%. In the revised manuscript, we have updated all relevant values, and also removed the exaggerated expression.

Page 5, Lines 175–178:

"Among 44 k SVs (10.1 Mbp in total) found in Japanese haplotypes, 26 k (15% of all SVs, 5.8 Mbp) were shared with the HPRC graph, primarily with EAS haplotypes, while 18 k (11%, 4.3 Mbp) were unique to our Japanese haplotypes (Fig. 1h)~~, suggesting a considerable proportion of population-specific SVs~~."

Response to the Reviewers

We sincerely appreciate the continued and valuable feedback from one of the reviewers. In response, we have carefully addressed all remaining concerns, re-analyzed the relevant data, and revised the text accordingly. In the point-by-point responses below, the reviewer's comments are shown in black, and our responses are in blue. Revised sentences in the manuscript are highlighted in yellow, and are also shown in this response letter with page and line numbers.

Reviewer #2:

Comment: I appreciate the effort of the authors in addressing my concerns, however, I am still not convinced that this number is not an overestimation. The authors are basically claiming that 42% of the structural variation seen in a group of 10 individuals from a human population is unique. The reference to the Ebert et al 2021 paper was just to provide a comparison from a study where such estimation has previously been provided for the Japanese population. I totally agree that the haplotype-resolved assemblies we have today provide much deeper insights into the genetic diversity as compared to short reads, however, that doesn't explain the rise to 42%. I cannot wrap my head around how it is even possible to see unique structural variation at such a scale.

Answer: We are grateful to the reviewer for the thorough review and careful comments. We have re-examined the proportion and contents of Japanese-specific SVs. Regarding the proportion, our original calculation method differed from that used in previous long-read pangenome studies, namely proportion of Japanese-specific SVs to only the Japanese SVs rather than all SVs as the denominator, which led to an apparent inflation. After recalculating the proportion of Japanese-specific SVs in all SVs, it has been reduced to 11%, which we believe is a more appropriate value. In addition, we have investigated the genomic locations and sequence contents of the Japanese-specific SVs. In the following responses, we describe these points in detail.

Comment: I agree with the authors that in the HPRC Nature 2023 paper the same truvari parameters were used for SV merging, however, to my knowledge, this callset was not used to make any statements about population-specific uniqueness. Also, the SV comparisons in the HPRC 2023 paper were based on SV alleles not SV sites. That's why in my previous review I suggested using stricter settings for the estimation the authors are trying to make and performing further validation if the percentage still stays

this high.

Answer: In the previous revision, we mentioned the HPRC Nature 2023 paper (Liao et al., Nature 2023) as suggested by the reviewer, and added the result using stricter settings (50% reciprocal size overlap) to the manuscript, which greatly reduced the proportion of Japanese-specific SVs from 42% to 8%. Moreover, in the next answer below, these numbers will become further smaller, specifically 42% to 11% and 8% to 4%, respectively.

Comment: If we look into the other studies authors referred to for comparison: In Gao et al, the percentage of CPC-specific SVs is around 17% and most of them were located at the centromeric and telomeric regions of chromosomes. They also clearly state that more than half of the CPC-specific variants were singletons or doubletons. In comparison, the authors are reporting 42% while not even considering singletons. In the more recent Nasser et al 2025 study, around 22% SVs were classified as population-specific. So I am curious to know where the range of 30%-50% that the authors mentioned is coming from. It should also be noted that in Nasser et al. 80% reciprocal overlap and sequence identity was used for finding unique SVs as compared to 95% used by the authors which is extremely flexible and prone to overestimate the unique SVs as I have pointed out before as well.

Answer: The value "42%" in the original manuscript was the proportion calculated using only the Japanese SVs rather than all SVs as the denominator, which inflated the value. Regarding previous studies, the "30%–50%" values we mentioned in the previous revision, compared to "17%" and "22%", were calculated using only SVs from the specific population rather than all SVs as the denominator. We recalculated the proportion of Japanese-specific SVs over all SVs, which reduced the value from 42% to 11%. Similarly, the proportion with stricter settings (50% reciprocal overlap) was reduced from 8% to 4%. We have replaced these values with the new ones in the revised manuscript. We also agree that the choice of thresholds for finding unique SVs is quite arbitrary, and we have added a cautionary note regarding this point in the revised manuscript.

Page 5, Lines 175–178:

"Among 44 k SVs (10.1 Mbp in total) found in Japanese haplotypes, 26 k (15% of all SVs, 5.8 Mbp) were shared with the HPRC graph, primarily with EAS haplotypes, while 18 k (11%, 4.3 Mbp) were unique to our Japanese haplotypes (**Fig. 1h**)."

Page 6, Lines 188–190:

"(...), we collapsed similar SVs with >95% sequence similarity as well as >95% size overlap, while the choice of thresholds is to some extent arbitrary. In contrast, when we applied the 50% reciprocal

overlap method, which is based solely on >50% size overlap without considering sequence similarity and is often used for large-scale short-read SV genotyping, the proportion of shared SVs increased from 15% to 48% (35 k) and that of Japanese-specific SVs decreased from 11% to 4% (3 k) (**Supplementary Fig. 3c**), suggesting the importance of long-read assemblies for fully uncovering the sequence diversity of SVs."

Comment: Additionally, I would reiterate that if the authors insist that this percentage of 42% is not an overestimation, further stratification is required e.g. by variant type and more importantly by genomic location. Another experiment that would help get a better picture and assess the 'uniqueness' of these SVs is a visual depiction of the SV sequence identity with other cohorts that the authors are comparing to i.e. the Chinese pangenome and the HPRC cohort. One way can be to align the SV sequence to the respective pangenome graphs for each unique SV and plotting the resulting sequence identities together.

Answer: While the percentage has now become 11%, we also stratified the 18 k Japanese-specific SVs by genomic location and variant type. We found that 18% (3 k) of the Japanese-specific SVs were located in subtelomeric regions, which comprise only 1.5% of the genome, and 39% (7 k) were located in segmental duplications plus 100-kbp flanking regions, which occupies only 6% of the genome. In total, more than half (57%) of the Japanese-specific SVs were found from subtelomeric regions and segmental duplications, which have been almost inaccessible with short reads. A new figure (Supplementary Figure 3a) shows the positional distribution of Japanese-specific SVs in these and other regions, indicating clear enrichment near chromosome ends and a largely uniform distribution elsewhere. Moreover, of the remaining 43% of Japanese-specific SVs, 91% contained tandem repeat sequences, whose variation also tends to be difficult to distinguish with short reads. We further aligned sequences inserted by Japanese-specific SVs to the pangenome graph of the Chinese and HPRC haplotypes. We confirmed that most alignments exhibited high sequence similarity, as expected for SVs that typically arise from duplications of existing sequences, whereas 15% showed <95% sequence similarity. A new figure (Supplementary Figure 3b) shows the distribution of these sequence similarities. We have added these results to the revised manuscript.

Pages 5–6, Lines 178–187:

"Among the Japanese-specific SVs, 18% (3 k) were located in subtelomeric regions, which comprise only 1.5% of the genome, and 39% (7 k) were located in segmental duplications plus 100-kbp flanking regions, which occupies only 6% of the genome (**Supplementary Fig. 3a**). Moreover, of the remaining 43% of the Japanese-specific SVs, 91% contained tandem repeat sequences, which is also difficult to

resolve with short reads. We further aligned sequences inserted by Japanese-specific SVs to the pangenome graph of the Chinese and HPRC haplotypes and confirmed that most alignments exhibited high sequence similarity, as expected for SVs that typically arise from duplications of existing sequences, whereas 15% showed <95% sequence similarity (**Supplementary Fig. 3b**). These results indicate that most of these unique SVs are derived from repetitive regions."

Page 21, Lines 707–716 (Methods):

"For the stratification of SVs, subtelomeric regions on T2T-CHM13 were defined as the 1-Mbp segments at each chromosome end, and regions surrounding segmental duplications were defined using annotated segmental duplications (https://github.com/marbl/CHM13) with 100-kbp flanking regions. Tandem repeats within SV sequences were annotated with Tandem Repeats Finder using the command "`trf 2 7 7 80 10 50 2000`". Inserted sequences by SVs unique to the Japanese-haplotypes were aligned to the pangenome graph of the Chinese and HPRC Year-1 haplotypes (https://pog.fudan.edu.cn/cpc) using GraphAligner with the default settings. The sequence identity of each inserted sequence with the graph was calculated as the number of matches divided by its sequence length."

Comment: Overall, I would again urge the authors to revisit this analysis and be extremely careful in making statements like finding 'a considerable proportion of population-specific SVs'.

Answer: In the revised manuscript, we have updated the values mentioned above and all other relevant values, and also removed the exaggerated expression.

Page 5, Lines 177–178:

"(...), while 18 k (11%, 4.3 Mbp) were unique to our Japanese haplotypes (**Fig. 1h**)~~, suggesting a considerable proportion of population-specific SVs~~."

Response to the Reviewers

We sincerely appreciate the continued and valuable feedback from one of the reviewers. In response, we have carefully revised the text accordingly. In the point-by-point responses below, the reviewer's comments are shown in black, and our responses are in blue. Revised sentences in the manuscript are highlighted in yellow, and are also shown in this response letter with page and line numbers.

Reviewer #2:

Comment: The authors have now made it explicit that instead of using the conventional way of comparison i.e. comparison to all SVs, they were comparing to only Japanese SVs which explains the inflated numbers. However, the revised statement reporting the SV stats in the manuscript is extremely confusing.

In the previous version of the manuscript the authors reported

"Among 44 k SVs (10.1 Mbp in total) found in Japanese haplotypes, 26 k (58%, 5.8 Mbp) were shared with the HPRC graph, primarily with EAS haplotypes, while 18 k (42%, 4.3 Mbp) were unique to our Japanese haplotypes (Fig. 1h), suggesting a considerable proportion of population-specific SVs"

In the latest version they state:

"Among 44 k SVs (10.1 Mbp in total) found in Japanese haplotypes, 26 k (15% of all SVs, 5.8 Mbp) were shared with the HPRC graph, primarily with EAS haplotypes, while 18 k (11%, 4.3 Mbp) were unique to our Japanese haplotypes"

First of all my question is if 15% is based on all SVs, what is 11% based on? Does it also use the number of all SVs in the HPRC graph as the denominator? If yes, what is the rationale for this and what does this value tell us about the callset the authors have generated? If it is w.r.t 44k, how is the number of SVs and the covered bps not changing? The concern I have been trying to raise repeatedly is about the message being conveyed, which is that almost half of the SVs found in this Japanese cohort have not been seen before. Using a different denominator simply to change the percentage of unique SVs from 42% to 11%, while keeping the SV count and covered base pairs exactly the same, does not address that concern. Because they still make up the same fraction as what the authors had reported before. For reporting the percentages, using all SVs as the denominator in my opinion is unnecessary here, firstly because it is making the sentence confusing by adding a different scale and secondly because when it comes to QC of a callset based on such a small cohort, the more important question to answer

is how many of the found variants are already known and how many are unique. To be even more clear, the shared vs unique comparison should definitely be done using ALL SVs as the authors claim to have done now but the results should be reported w.r.t the new callset (as was being done before). So the denominator should still be the number of SVs the authors have found i.e. 44k (or the respective bps). The same comment applies to the now changed numbers about SNPs. It is completely unintuitive and confusing to use different scales to compute the same quantity. If the authors really want to report them, they need to be stated separately with explicit reasoning.

Answer: We thank the reviewer for this important comment. We agree that the choice of denominator was confusing with respect to assessing the fraction of unique SVs within the Japanese cohort. Following the reviewer's recommendation, we have therefore reverted to reporting shared and unique proportions with respect to **the Japanese callset itself**, using the total number of SVs identified in the Japanese haplotypes as the denominator. The same changes have been applied to the SNV analysis.

Furthermore, to address the reviewer's core concern that the fraction of SVs observed only in Japanese haplotypes was too large, we applied stricter criteria for merging similar SVs, following Nassir et al. (2025), specifically >80% sequence overlap rather than >95% used previously. This resulted in a decrease in the percentage of SVs observed only in the Japanese haplotypes from 42% to 24%. We note that Nassir et al. reported the percentage of Arab-specific SVs as 22% by using **the union of all SVs across the HPRC and Arab callsets** as the denominator, but this value becomes 55% when recalculated using **only the Arab callset itself** as the denominator as we have done here. Therefore, the 24% observed in our Japanese cohort is much lower than the fraction of population-specific SVs reported in Nassir et al. We have further revised the wording and the message of the sentences related to SVs, as detailed in the next answer to the comment.

Additionally, in accordance with the editor's suggestion, we now report exact counts for both numerators and denominators when reporting percentages to improve clarity. The revised sentences are provided at the end of this document.

Comment: The stratification of unique SVs by genomic location and the results from the alignment exercise are very helpful and clearly indicate that most of this apparent uniqueness is either occurring in complex regions or stemming from representation differences which makes the claim of being "unique" doubtful. The take home message here should be that for these cases it is difficult to distinguish actual uniqueness from assembly artifacts or representation differences however the statement “These results indicate that most of these unique SVs are derived from repetitive regions” is not conveying that and should be revised.

Answer: We agree with the reviewer's interpretation. In the updated manuscript, we have substantially restructured this section to address the concern directly. Rather than characterizing SVs not observed in the HPRC haplotypes as "Japanese-specific", we now immediately note that the majority of these SVs observed only in the Japanese haplotypes were located in complex genomic regions where it is difficult to distinguish actual uniqueness from assembly artifacts or representation differences. The revised text presents the stratification by genomic location as the primary characterization of these SVs and then notes that the majority of their inserted sequences exhibit high similarity to sequences already present in existing pangenome graphs, without implying that they represent truly unique variants. We have also removed words and sentences that conveyed an impression of uniqueness or specificity of the detected SVs.

Comment: Overall, I would strongly recommend the authors make another careful pass through the manuscript, be consistent in their comparisons and report the observations in a coherent and understandable way for the reader.

Answer: We believe that these revisions substantially improve the consistency, clarity, and interpretability of the manuscript and directly address the reviewer's concerns. The revised text is presented below. We have also updated the relevant figure and supplementary figures accordingly.

Pages 5, Lines 166–176 (SNVs):

"We next compared SNVs identified in the Japanese graph with those called from the HPRC Year-1 graph using the same T2T-CHM13 reference genome. In the Japanese pggb graph, we identified 6,659,402 SNVs relative to T2T-CHM13. Of these, 5,637,761 SNVs (85%) were shared with the HPRC Year-1 graph, primarily with the eight non-Japanese EAS haplotypes (4,590,109 SNVs, 69%), whereas 1,021,641 SNVs (15%) were not observed in the HPRC Year-1 haplotypes and were therefore classified as Japanese-specific in this comparison (**Fig. 1g**). Among the 6,659,402 Japanese SNVs, 5,064,963 SNVs (76%) were present in two or more Japanese haplotypes, suggesting their commonality in the Japanese population. Of these 5,064,963 SNVs observed in multiple Japanese haplotypes, 4,817,785 SNVs (95%) were also found in haplotypes of other populations within the HPRC dataset, whereas only 247,178 SNVs (5%) were found specific to the Japanese haplotypes."

Pages 5–6, Lines 177–195 (SVs):

"For SV analysis, we first merged similar SVs across all haplotypes using Truvari with 80% sequence

overlap, and then retained only SVs detected in two or more haplotypes within either Japanese or HPRC haplotypes. Among 46,173 SVs (11.0 Mbp in total) identified in the Japanese haplotypes, 35,168 SVs (76%, 8.0 Mbp) were also observed in the HPRC Year-1 graph, primarily in EAS haplotypes, whereas 11,005 SVs (24%, 3.0 Mbp) were not observed in the HPRC Year-1 haplotypes (**Fig. 1h**). However, the majority of these 11,005 SVs were located in complex genomic regions, where it is difficult to distinguish actual uniqueness from assembly artifacts or representation differences. Specifically, among these 11,005 SVs, 2,064 SVs (19%) were located in subtelomeric regions, which comprise only 1.5% of the genome, and 4,520 SVs (41%) were located in segmental duplications plus 100-kbp flanking regions, which occupies only 6% of the genome (**Supplementary Fig. 3a**). Moreover, of the remaining 4,421 (40%) SVs not in subtelomeric regions nor segmental duplications, 4,040 SVs (91%) contained tandem repeat sequences.

To further characterize these 11,005 SVs (5,437 insertions and 5,568 deletions) observed only in the Japanese haplotypes, we aligned their inserted sequences to the pangenome graph of the Chinese and HPRC haplotypes and confirmed that 4,666 (86%) of the 5,437 insertions exhibited high (>95%) sequence similarity, as expected for SVs that typically arise from duplications of existing sequences, whereas 771 (14%) showed <95% sequence similarity (**Supplementary Fig. 3b**)."

Pages 20, Lines 699–700 (Methods):

"For the comparison involving other haplotypes, similar SVs were merged using the command "`truvari collapse -r 1000 -p 0.8 -P 0.0 -O 0.8 -s 50 -S 100000`", following a previous study[85]."

Response to the Reviewers

We sincerely appreciate the continued and valuable feedback from one of the reviewers. In the point-by-point responses below, the reviewer's comments are shown in black, and our responses are in blue.

Reviewer #2:

Comment: The current revision adequately addresses the concerns raised in my previous reviews, and I have no further remarks.

Answer: We thank the reviewer for their continued engagement with our manuscript across multiple rounds of review and for confirming that all concerns have been adequately addressed.