nature portfolio

Peer Review File

The 1000 Chinese Pangenome empowers medical and population genetics

Corresponding Author: Professor Jian Yang

Any redactions in this file are there to maintain patient confidentiality, the confidentiality of unpublished data, or to remove third-party material.

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

Version 0:

Reviewer comments:

Referee #1

(Remarks to the Author)
The authors generate a low coverage long read data set of more than a thousand genomes from China. This is a valuable effort for the genomics community as such data sets are needed both for common disease association studies and as a reference resource to filter variants observed in rare disease genomes. At the same time, the present version of the manuscript does not provide enough details on the quality of the resulting genome reconstructions and I am providing specific comments and suggestions about this below, among other comments.

Major comments

- The abstract and title advertise "diploid genome assemblies of 1,116 Chinese individuals". Typically the term "genome assembly" is used to refer to de novo assembly without the use of a reference genome, but this is not what was done here. I suggest to use a different term to avoid confusion. The term "Pangenome-Informed Genome Assembly" is more clear, but only used further down in the main text.

- More information on the characteristics of the PacBio sequencing data are needed. The "CCS --all" mode used for data generation is unusual and discontinued (according to the PacBio website https://ccs.how/faq/mode-all.html). That is not necessarily a problem, but more details such as error rate and read length distributions would be important to include for such a non-standard data set.

- Good to see that the authors have created a website, which hosts the graphs and variant sets. No individual level data appears to be available from that website, i.e. the VCF files do not have genotypes and the graphs do not have paths corresponding to the assemblies. In "Data availability" the authors say that "The genome assembly data generated in this study can be accessed upon request at the National Genomics Data Center (https://ngdc.cncb.ac.cn).", but do not give an accession number. I went to that website and was not able to easily find the data set. Could the authors clarify how exactly access can be requested? Is access possible worldwide or is it restricted to applicants from within China?

- The whole paper hinges on the assemblies (with PIGA) and the resulting call set, but in its present form, the evaluation of their quality is not sufficient and not reproducible. Some questions/suggestions:

1) At least a small number of benchmark samples (e.g. the three samples used in this study) should be made publicly available, including the input reads.

2) Could you quantify the added value of the long reads? From short reads alone, one can expect a QV value >40 even using a standard pipeline (see Logsdon et al., bioRxiv, Figure 3b; https://www.biorxiv.org/content/10.1101/2024.09.24.614721v1).

3) The design rationale is somewhat similar to the 1000 Genomes Phase 3 data set (Sudmant et al., Nature, 2015), where low coverage data across many samples was used to discovery SVs. This is a good strategy to maximize the yield on rare

SV alleles, but there is risk of an inflation with singletons. Figure 2b hints at an excess of singletons, but that is difficult to judge from this plots. I suggest to provide a log-log plot similar to Extended Data Figure 2 of Sudmant et al. (2015). In such a visualization, a linear relation would be the expected behavior, which makes it easier to assess (would also be good for Figure 1g, i.e. make the x-axis logarithmic, too). Typically, STR/VNTRs behave quite differently from other classes of SV and it might be helpful to create stratified versions for some of the plots.

4) Flagger annotations are produced but not used much. For the benchmark samples, how many variants come from regions flagged in this process and could the authors use this to estimate how many of the singleton variants come from spurious calls?

- The HLA analysis lacks validation and QC. Lines 309-314: The number of fields in an allele is defined by the allele nomenclature and heavily depends on the gene itself. So this has little to do with new assemblies. Probably, simply selecting random gene alleles will produce similar results.

- It would be important to put the imputation results into context by comparing to imputation using other reference panels and by contrasting imputation performance of different panels for different populations.

Minor comments

- Line 24: "The pangenome..." are you referring to one specific pangenome or rather pangenome-based methods more broadly? In case of the latter, you should rephrase this sentence.

- Lines 29-31: "... we constructed a pangenome comprising 628.8 million base pairs of sequences absent from the current references GRCh38 and CHM13, including 29.3 million base pairs of functional genic and regulatory elements": I am not convinced that these quantifications in Mbp are very meaningful as they will be quite strongly affected by alignment parameters. For this reason, it is probably not the most informative statistic for the Abstract. This also applies to Figure 1a.

- Line 38: What does "pan-variant" mean? Either omit or define it.

- Line 63: "First, rare variants, which often exhibit high pathogenicity and play critical roles in inherited diseases" You are citing Kryukov et al. 2007 here, which is about rare missense mutations. I haven't gone back and read the whole study, but it seems questionable whether it supports the claim that rare variants in general often "exhibit high pathogenicity".

- Line 81: "capture the full spectrum of genetic variations within the Chinese population (Supplementary Figure 1)" That figure is a PCA plot, so that does not support the statement that the "full spectrum of genetic variations" is captured. I wasn't able to find any information on how this PCA is done, i.e. how the set of input variants is defined and how exactly the PCA is run.

- Line 158: "we mapped the haplotype coverage of pangenome segments to CHM13". I think "haplotype coverage" not a widely used term, so it is best to rewrite this sentence. Perphaps the authors meant "we mapped the haplotype-covered pangenome segments to CHM13".

- Line 192-195: Could you indicate how many of the variants nested within reference SV sites occur in tandem repeat contexts? That is, for how many is the reference SV a TR expansion/contraction?

- Figure 2f is unclear to me. Is it saying that there is <10 SNVs and indels per sample added? That cannot be correct. Or is it the mean number of alleles per variant site? In that case it'd be good to show it in a way that one can verify that indeed almost all SNPs are (and remain) biallelic. In any case, please amend the caption to make it more clear what is shown.

- Line 203: could the authors please show average counts for TRs and non-TRs?

- Line 209: rephrase "biallelic alleles"

- Lines 210-212: That is impossible to judge from the figure, please provide a version with x-axis in log scale (see major comment above).

- Line 213: What was the fraction of SVs in Hardy-Weinberg equilibrium before merging? Can the authors put 86% into context? Also please provide the details of the HWE test done in the Methods.

- Line 225/226 and Supp Fig 15: How exactly was that edit distance computed as (to my knowledge) Vamos does not provide the exact allele sequence. That should be described in Methods.

- Line 236: It would be good to mention that the number of variants is relative to the GRCh38 reference genome.

- Line 245: The authors write about elevated proportion of rare SVs, so it is better to also mention the fraction of non-gene altering rare SVs.

- Line 277: judging by Fig.3d, there is only one sample with expanded TR, is Fig.3e based on one sample only?

- Fig.3c: very hard to see if any samples went out of the normal range (for example for DAB1 and FGF14, mentioned on line 264).

- Fig.3d: left and right subplots have different y-axis. Currently this is very misleading, should either use the same y-axis, or clearly highlight the difference.

- Fig.3f ideally should have color legend. Additionally, why are some parts purple? If it is a combination of exons and introns, is it possible to make the plot clearer?

- Fig.3g: blue haplotypes look identical. If there are any differences, it's impossible to notice with human eyes.

- Line 295: show the number of affected haplotypes (86 and 1?). Possibly, select shorter names/abbreviations for the haplotypes.

- Line 395: how were HLA alleles imputed?

(Remarks on code availability)
Thank you for providing the code under MIT license. It is essentially a collection of bash scripts (and some Python and R). Its main purpose appears to be as documentation what was done, not so much as a workflow to be run by others. I have not attempted to run it, which appears somewhat tedious as a user would need to infer all the assumptions about command-line parameters and sometimes hard-coded file names by reading the scripts. It would be very helpful if the authors could at a minimum provide the exact environments with all the dependencies (e.g. in a container or as a Conda environment), a test data set, and exact instructions how to to run that test data set.

Referee #2

(Remarks to the Author)
I co-reviewed this manuscript with one of the reviewers who provided the listed reports.

(Remarks on code availability)

Referee #3

(Remarks to the Author)
Wang et al developed PIGA, a reference-guided assembly pipeline. Using PIGA, they produced the reference-guided assembly of 1116 Chinese individuals with high-coverage short reads and low-coverage long reads. They analyzed the assemblies with pangenome approaches and found many variants in the cohort. This is a huge effort and PIGA is methodologically interesting. However, a key question is whether the assembly quality is high enough for typical pangenome analysis as it is known that reference-guided assembly has limitations. Critical observations are as follows:

1) The manuscript does not describe the PIGA pipeline in detail. The first several steps up to draft assembly are akin to hapdup (Kolmogorov et al, 2023). However, started with a de novo assembly with long reads at high coverage, hapdup can produce a better backbone and hapdup can't compete with mainstream assemblers. In my understanding, the authors construct a pangenome graph with all ~1000 draft assemblies and use the graph to improve the assembly (is it correct?). If a rare SV is missed in all draft assemblies, it would not be reconstructed in the final assembly.

2) Fig 2c and Fig S22 suggest the PIGA assemblies are burdened with misassemblies, confirming 1). Also, Inspector and Flagger might underestimate misassemblies given low-coverage long reads – the actual error rate might be higher. We cannot use these assemblies in segmental duplications or satellites where de novo assembly shows real power over alignment-based methods.

3) Liftoff has not been shown to have enough power to annotate HLA genes. In Fig S19c/d, HLA-DRB3/4/5 are shown to have high LD. This is incorrect as HLA-DRB3/4/5 can't co-exist on one haplotype.

4) A few smaller issues. (a) the assemblies with low-coverage long reads do not have chromosome-scale phasing. (b) an error rate of 1/40000 in one assembly could mean 1/20 error rate across 2000 haplotypes. This will overestimate the number of variants. (c) in the mean time, QV does not tell us how many heterozygous SNPs are missing. This will underestimate the number of variants.

5) Overall, PIGA assemblies probably lag far behind real de novo assemblies produced by several other consortia in quality. The analysis might not be much better than read-to-reference alignment.

(Remarks on code availability)
I briefly looked the documentation. It is a long and sophisticated pipeline calling many tools but might not be easy for external users to use.

Referee #4

(Remarks to the Author)
In the submitted manuscript, Wang et al. use pangenomic approaches to analyze an immense newly generated dataset of genome sequences from a Chinese population. Specifically, the data comprised high-coverage PacBio HiFi long-read, Illumina short-read, and Hi-C data from 55 samples combined with modest coverage PacBio long-read and Illumina short-read data from an additional 1,099 samples (with 10 samples overlapping between the strategies). This allowed them to achieve a nearly complete view of genetic variation within their sample, including technically challenging variation such as tandem repeats (TRs), other classes of structural variation (SVs), as well as additional variation nested within SV alleles (e.g., SNPs encompassed by an SV). By combining with RNA-seq data they generated from the same sample, they also used their genotype data to perform eQTL mapping, revealing gene expression associations not only with SNPs but also other classes of variants. Finally, they built an imputation server that allows users to upload and impute missing genotype data based on their new reference panel.

This work represents a technical milestone, as I am unaware of any other project that has produced this number of human genome assemblies based on this scale and quality of data, nor has any other project applied human pangenome approaches on this scale. The large sample size allowed the authors to better capture rare variation that is abundant in human genomes. The ability to resolve nested variation, TRs, and HLA haplotypes were relatively novel aspects of the work, as was the analysis of gene expression cis-heritability, partitioned by variant class. Overall, however, even these sections lacked motivating context from the literature, so it was difficult to determine whether any truly novel biological insights were obtained. As such, the primary value of the study is as a genomic resource, but this value is somewhat undercut by the lack of availability of the raw sequencing data or individual-level genotype calls. The construction of an imputation server partially addresses this latter issue for one common application - though see below for a critique about how the advantages of this panel remain unclear. In addition, data privacy concerns preclude many researchers from uploading data to an external server.

The writing was relatively clear, though I have noted a few exceptions.

Please see my specific critiques and questions that I hope the authors will address:

Major comments:

1. The abstract (line 25) motivates the work by saying that previous smaller studies do not "fully capture" genetic diversity and that the current study "compiled a comprehensive variant catalog for the Chinese population" (line 232). Both statements should be moderated, as the goal to fully capture genetic variation is infeasible (any new sample will add some variation that is private to that individual or their close relatives) and the current study, while more comprehensive given the sample size and technologies involved, does not represent all of the genetic diversity in the broader population.

2. In addition to the point above, there should be some consideration of the ancestry composition of the sample population—for example whether it includes representation of the diverse ethnic groups that comprise the broader Chinese population. Such analysis/description should, however, be careful not to conflate the biological concept of genetic ancestry and socially or politically defined population labels (see e.g., https://www.nationalacademies.org/our-work/use-of-race-ethnicity-and-ancestry-as-population-descriptors-in-genomics-research).

3. The Methods description of PIGA (the approach developed to integrate the modest coverage data into the pangenome) was greatly lacking in detail. All of the six steps should be described in detail in the methods, and decisions about the use of various software tools and parameters should be justified. In addition to these additions to the Methods text, Supplementary Figure 3 should be integrated with the workflow diagram that is currently posted in the PIGA GitHub repository, and the updated Supplementary Figure should be referenced on line 92.

4. The authors state on lines 82-89 that 10 samples were sequenced by both high- and modest-coverage sequencing strategies, but only 3 samples (line 95) were chosen as benchmark samples. Why were only 3 samples used, and why these 3 specifically?

5. In many cases, it was unclear to me which findings were truly enabled by the pangenome approaches employed and which findings could have been obtained through conventional analyses based on linear representations of the reference genome.
a. For example, the section starting on line 238 describes various mutations in clinically relevant genes, such as SVs and TR expansions within genes in the OMIM and COSMIC databases. What proportion of these were undetectable based on short-read data alone and/or long-read data but conventional linear reference-based analysis?
b. As another example, line 308 states that "compared to short-read data, the 1KCP assemblies enabled us to resolve the HLA alleles at the four-field resolution". But what is the basis for this statement? Was the same analysis attempted with short-read data, and if so, what were the results?
c. As a third example, the imputation results (starting line 399) contrasted between two different genotyping arrays -- which I found a bit tangential -- but did not contrast between this new pangenome-enabled reference panel and conventional reference panels (either built from the same samples or other common imputation panels such as 1000 Genomes Project, TopMed, HRC, etc.).
d. As a fourth example, can the GSTM1 and GSTM2 deletions be identified with short-read data alone and/or long-read data but conventional linear reference-based analysis?

6. The use of SEI and Enformer for regulatory annotation was interesting, though these should be carefully described as regulatory predictions rather than "regulatory elements" (see line 31), and it should also be noted that these are specific to the blood—a limitation that could be expanded upon in the Discussion. More details are needed about the benchmarking approach and what 3 "benchmark samples" were used. Moreover, it seems like SEI and Enformer were compared based on their ability to predict ATAC-seq peaks (which showed high accuracy), but then applied for more granular predictions of various regulatory elements (promoters, enhancers, CTCF sites, etc., line 140). What is the justification for this approach, and what should be expected for the accuracy of the predictions?

7. Many of the eQTL results focused on cases where an SV, TR, or nested variant (together termed "complex variants") was the lead eQTL. However, as far as I could tell, these results were not based on standard fine-mapping approaches such as CAVIAR, SuSiE, etc. so it is unclear how confident we should be that a given complex variant is causal. I recommend that a fine-mapping approach such as SuSiE that allows for the possibility of multiple causal variants for a given gene be applied, followed by an analysis to determine in how many cases a complex variant has the highest posterior inclusion probability and the extent to which these are enriched compared to their background proportions.

8. The imputation methods also require much more detailed description. One major question I have is whether the imputation method employed leverages the pangenome graph or whether variant calls are ultimately collapsed back to a linear representation and used as a reference panel with a conventional imputation tool. While such an approach may be justified, I am wondering if all of the nested variants can actually be represented in a biallelic form while preserving information about their nestedness or if there is information loss when converting from the pangenome into the decomposed biallelic representation. If there is no information loss, then what is the value of the pangenome approach in the first place?

9. Line 284: I noted that immunoglobulin genes were enriched in complex structural haplotypes. Given the occurrence of somatic recombination and hypermutation in blood immune cells and the fact that the source material for this study was blood, I am wondering how the authors ruled out potential artifacts of somatic variation.

10. The entropy-based measurement of LD at the HLA locus was interesting, motivated by the inability of conventional metrics (r2, d', etc.) to deal with multiallelic variants or consider 3+ variants at a time. I understand that in a large sample and using long-read technology, it is possible to better resolve the entire sequence of HLA haplotypes, including within non-coding regions (i.e., the fourth field in the HLA allele naming scheme). However, I am wondering about the basis of the statement that "four-field alleles better capture the LD structure of HLA genes" (line 318), which appears to be based on the higher ε value. How can such a statement be made without knowing the ground truth haplotype structure? In addition, the statements about the measured levels of LD partitioned between LD groups seemed circular (lines 321-324). Are LD groups not themselves defined based on levels of LD?

Minor comments:

1. Line 197: I was not sure how to interpret the statement that 56% of SV sites are "multi-allelic". Does this just mean that SVs often possess nested variation that is apparent in a sample of this size? Is a nested variant within an SV enough to make it multiallelic? If so, contrasting the number of multiallelic SNVs, indels, and SVs in the Results and Fig. 2f feels a bit trivial/tautological because it simply reflects the fact that SNPs cannot contain nested variation and indels are smaller than SVs by definition.

2. Line 368 (and throughout): the statement that MAD1L1 "was regulated by" copy number and sequence motif is too strong, and should instead be phrased as association. This issue applies to essentially all of the association mapping results, which should not be interpreted or reported as causal relationships (see Major Comment about fine-mapping).

3. The authors state on line 436 that previously released human pangenomes typically included only a few dozen individuals, but a brief search of the literature suggests this is probably an overstatement, and should be made more specific regarding the technologies involved.
a. Chinese pangenome – 36 samples in Phase I (Gao et al., Nature, 2023) → goal is at least 500
b. Human Pangenome Reference Consortium – 47 samples (Liao et al,, Nature 2023) → goal is 700
c. African Pangenome – 910 samples (Sherman et al., Nature 2019)
d. Denmark Pangenome – 150 samples (Maretty et al., Nature 2017)

4. The Methods section would benefit from additional details:
Was RNA integrity checked after RNA extraction (e.g. Agilent Tapestation) prior to RNA-seq library preparation (Methods)?
How was the number of PEER factors chosen for eQTL analysis (Methods)?
How were medically relevant genes (line 240) defined?

5. Line 163: the word "non-reference" can be removed, as it is redundant with the phrase "absent from the GRCh38 and CHM13 references"

6. Figure 1a - Why is the HiFiasm NG50 smaller than that of PIGA, given that the former are higher coverage?

7. Figure 1f - What does "rank of epigenome profiles" mean?

8. Figure 2c - Why is the size of non-singleton non-reference sequence not directly related to the proportion relative to the pangenome? For example, I don't understand what it means that introns account for the largest proportion of annotated non-

singleton non-reference sequence but a small proportion. Are you referring to the possible point that 70 Mb is small relative to the total amount of intronic sequence that does exist in the reference? If so, this should be clarified for readers.

9. What is the size threshold that defines a structural variant? This should be noted at first mention of SVs and indels in the Results section and/or in the Methods.

10. Fig. 2g - What is a "raw" versus "merged" SV? This was never mentioned in the text.

11. Fig. 3 - The definition of a "rare" SV should be noted in the figure legend.

12. Fig. 3c - I assume that the dark grey bars extend infinitely to the right. If so, I am wondering if there is a way to indicate this visually.

13. Fig. 3d - The barplot with height of 1 is a strange way to represent these data. Perhaps simply show the left histogram with both the outlier threshold and the one outlier sample annotated as vertical lines?

14. Fig. 3j - Is there a pattern that this plot is meant to convey? I understand that the pangenome enables high resolution genotyping of HLA, which is impressive, and that this then enables measurement of LD between alleles at different loci, but I'm not sure if there is anything particularly interesting about the observed structure.

(Remarks on code availability)
I looked through the code briefly. It is a set of numbered analysis scripts, which I think meets the minimum standard for transparency, but is not designed to generalize and be used in anyone else's study. For example, each script refers to numerous inputs and produces numerous that are not described in any form of documentation. The READMEs describe the basic roles of each analysis script.

Referee #5

(Remarks to the Author)
I co-reviewed this manuscript with one of the reviewers who provided the listed reports.

(Remarks on code availability)

Version 1:

Reviewer comments:

Referee #1

(Remarks to the Author)
I appreciate the extensive revision done by the authors and the paper is now much improved. In particular, the QC of the call set is much better and more comprehensive information on the quality is provided to the readers.

Here are two minor comments for the authors to consider:

- Page 27: "Four-field alleles of HLA genes were called from assemblies using HiFiHLA". HiFiHLA is designed for HiFi data, not assemblies, is it a typo? Please also mention which data type HiFiHLA was run on (page 11, line 341). Also consider running Immuannot for detecting HLA alleles from assemblies.
- Not all tools used by PIGA are cited, for example, GLNexus, Longshot, Beagle4, merfin, etc. I suggest to double-check and make sure to include all appropriate citations.

(Remarks on code availability)
Better documentation for the Snakemake pipeline is still needed. Specifically, there is no test dataset; no unified installation of dependencies (e.g. conda package or environment). In the tutorial, config contents are not explained and there are no links to dependencies.

Referee #2

(Remarks to the Author)
I co-reviewed this manuscript with one of the reviewers who provided the listed reports.

(Remarks on code availability)

Referee #3

(Remarks to the Author)
The authors have put a lot of efforts in the revision and addressed/clarified several of my concerns. I really appreciate that. I still have additional follow-ups.

1) [previous point 1] The authors now describe PIGA in detail. This is a careful and sophisticated pipeline involving many tools for alignment, phasing, assembly and pangenome. I am very impressed. On the other hand, the PIGA pipeline is complicated and specialized for their own data. It will be challenging for others to use PIGA. Furthermore, the authors argued that their procedure is cheaper. This was true when they did sequencing with PacBio Sequel IIe, but might not be the case now with Revio+SPRQ (~16X cheaper than Sequal IIe). Given the power and logistics, most groups would choose straight long-read sequencing over the authors' hybrid procedure.

2) [previous point 2 and 3] The authors want to emphasize on HLA haplotypes, but their results suggest noticeable errors. In response, Fig. R22 shows 1.7% error rate involving DRB3/4/5. These three genes are on distinct haplotypes. If there are switch errors around them, there will be other types of errors, making the haplotype structure questionable. Furthermore, the best short-read HLA genotypers like OptiType can achieve ~99% 2-field accuracy for HLA-A, a lot higher than the ~92% 3-field consistency shown in Fig. S32b. Also, HLA-G is one of the easier genes to type. I am surprised by its low consistency. I understand the authors use HLA-HD to type more genes, but if the inconsistency is caused by HLA-HD, they should change the short-read HLA typer. In contrast, resolving MHC with standard long-read assembly will be nearly error-free and more trustworthy as is shown in the HGSVC paper.

3) [previous point 4] The authors responded that their variant calling error rate is close to 1/60kb. This means there will be 3e9/60kb=50k false variants per individual. If errors are not shared, there will be 50k * 1116 = 50 million false variants in total, more than the number of called variants. I don't think the reality is this bad because many errors are shared, but it would be good to validate the calls with orthogonal data (most QC strategies in the manuscript, albeit sophisticated, still rely on the same datasets and are thus not orthogonal). Targeted sequencing of randomly selected variants will be nice. The authors may also compare their calls to gnomAD SNPs. Ideally, most common gnomAD SNPs (say MAF>10% among Eastern Asians) should be captured by the PIGA SNPs; most PIGA SNPs, except in difficult regions, should be found by gnomAD given the larger sample size of gnomAD. Large deviation from this expectation would indicate potential errors.

4) [new comment] The small variant VCF on the 1KCP website is incomplete. It only contains the first ~15Mb of chr1.

(Remarks on code availability)
I had a look the github repo in the first round of the review. This repo is offline now. Please bring it back.

Referee #4

(Remarks to the Author)
Overall, the revised manuscript is greatly improved. The authors have rigorously and comprehensively addressed my comments, as well as those of the other reviewers. In response to my comments, the new eQTL fine-mapping and enrichment analyses are especially nice, as are the new analyses demonstrating the performance improvements for imputation, stratified by variant type.

I have a few remaining minor comments and questions about the revised manuscript:

1. Line 62: Suggest revising “which generally exhibit higher pathogenicity than common variants” to “are more likely to be pathogenic than common variants” or “are enriched for pathogenic effects compared to common variants.”

2. Line 134: Is “read-based” the appropriate term here? Since both methods use sequencing reads, perhaps “conventional single-reference alignment-based” would be more precise, as your method uses a pangenome-guided assembly approach.

3. Paragraph on line 193: What is meant by “genomic element”? Is the definition broader than putative functional elements? Some clarification would be helpful.

4. Line 270: Suggest revising “strong purifying selection” to “signatures of purifying selection” or “evidence of purifying selection.”

5. Paragraph starting on line 336 (and specifically the last sentence starting on line 356): Most of the revisions to the section on LD are appreciated, but the last sentence still reads as potentially circular. Could “LD groups” be renamed to “LD blocks”? Then you could state that, beyond these blocks defined by strong LD, there is modest LD that extends over a longer range, and report the extent in kilobases.

6. Line 368: Can the number of samples with RNA-seq data be reported?

7. Line 412: What is meant by “unit”? Can you describe this in terms of allelic fold change? For example, see https://www.nature.com/articles/s41467-024-44710-8

8. Figure S2B (middle left): “Phred” is misspelled as “Phread” in the y-axis label.

9. Figure S37: “1KCP Hifiasm + PIGA using the whole 1KCP assembly panel (N=1,116) captures 4.3-fold more rare SVs and detects 3.1-fold more TR alleles than a smaller panel using only Hifiasm assemblies (N=55).” Is this a fair comparison given that the whole 1KCP panel contains so many more samples than the Hifiasm assemblies?

10. Figure 1C: I find Figure 1C a bit confusing. It seems to show a large percentage of k-mer and structural errors with PIGA. How is this consistent with the statement that “PIGA assemblies provided more accurate sequences for SVs and TRs compared to read-based approaches (Figure S10D and Figure S27A)”? Is Figure 1C showing results before PIGA filtering?

11. Figure S6A: The y-axis shows rank of each reference type across 6 alignment metrics. Could this plot be made more informative by including the actual metric values, in addition to the rank?

(Remarks on code availability)
I was unable to run the PIGA pipeline in Snakemake after downloading from GitHub. Specific issues I encountered:

1. The PIGA tutorial workflow execution should be: snakemake -s Snakefile --cores 64 --configfile config/config.yaml --configfile config/tools.yaml

2. The provided config.yaml file returns an error in Snakemake: “snakemake.exceptions.WorkflowError: Config file is not valid JSON or YAML. In case of YAML, make sure to not mix whitespace and tab indentation.” The trailing comma should be removed from this line: topmed_east_asian: “test/variant_panel/{chr}/TOPMed_WGS_freeze.8.east_asian.merge.{chr}.filter.snps.liftover_chm13.shapeit4.vcf.gz”,

3. After correcting config.yaml, Snakemake returns a KeyError in ‘tools’ at Snakefile line 14. The tools.yaml file appears to contain a sample identifier and not a list of tools.

Referee #5

(Remarks to the Author)
I co-reviewed this manuscript with one of the reviewers who provided the listed reports.

(Remarks on code availability)

Version 2:

Reviewer comments:

Referee #1

(Remarks to the Author)
We thank the authors for their diligent work. There is one issue left, however: I was not able to download the test data the authors provide with their PIGA software because yanglab.westlake.edu.cn was not reachable, neither from my university network nor from home. Either the server is down or there is some geoblocking in place that does not allow me to access this server from my country. In any case, I suggest the authors deposit the test data in a DOI-minting repository such as Zenodo. In this way the data remain accessible even when URLs change (the download script depends on many different URLs, so sooner or later this will break).

(Remarks on code availability)
We thank the authors for adding a script to download test data and the yaml environment. The download script does not work for this reviewer, however (see comments above).

Referee #2

(Remarks to the Author)
I co-reviewed this manuscript with one of the reviewers who provided the listed reports.

(Remarks on code availability)

Referee #3

(Remarks to the Author)
I thank the authors for the revision. I still have two minor comments:

1) [previous comment 2] The accuracy of mainstream HLA genotypers approaches ~99% on HLA-A. The authors can only reach 92%. At this high error rate, LD-based analysis involving HLA-A (Line 349–358) will be problematic. Generally, the

authors probably don't have enough power to accurately generate Fig. 3j due to phase switches with HiFi only. Population phasing may not be reliable either due to MHC diversity. The right analysis should rely on long-range phased assemblies like those from HPRC and HGSVC. It is probably better to leave out this part in my opinion.

2) [previous comment 3] The authors now provide SNP calls. On chr10, I counted 1275k 1KCP SNPs matching gnomAD v4.1 SNPs at both positions and alleles. The transition/transversion ratio (ts/tv) of these SNPs is 2.35. 437k 1KCP SNPs are not found in gnomAD. Their ts/tv is only 1.36. This suggests ~30% false discovery rate among SNPs not found in gnomAD (formula: 3(x-y)/[(2x-1)(y+1)], where x is the true ts/tv, y is the observed ts/tv and errors have a ts/tv of 0.5). The overall chr10 error rate is thus ~7.6% (=30%*437/(437+1275)) – that is 3.5 million wrong SNPs genome wide. Although such an estimate is very crude and makes multiple assumptions, the real error rate is probably around this level. I am okay with the error rate, but I think the authors should show an error estimate in the main text to let readers know. It is worth noting that the expected ts/tv increases with sample size due to CpG mutations. For example, the ts/tv of singleton SNPs overlapping with gnomAD reaches 2.84. ts/tv=1.96 does not imply high SNP accuracy across 1,116 samples.

(Remarks on code availability)
The github repo looks well organized, though I have not run the pipeline.

Referee #4

(Remarks to the Author)
Thank you to the authors for thoroughly addressing most of my previous comments and improving the generalizability of their software workflow. However, please see below for some remaining issues with the latter.

(Remarks on code availability)
While the workflow is now modular at the level of execution, all software dependencies are currently bundled into a single, large Conda environment. As a result, users must install the full suite of software even if they wish to run only a subset of the pipeline.

Although I was able to successfully download the provided test dataset, I was unfortunately unable to create and activate the PIGA Conda environment in order to test the workflow. Specifically, the GitHub repository URL listed in the README appears to be incorrect. Based on my attempts, the correct repository is https://github.com/hahahafeifeifei/PIGA rather than https://github.com/JianYang-Lab/PIGA. In addition, attempts to create the Conda environment either stalled or failed during dependency resolution. Examination of the log file suggests that these issues may stem from incompatible versions of Python, WhatsHap, and Snakemake.

These issues may be related to the use of a single Conda environment, with tools invoked directly within Snakemake rule shells. To address these issues, the authors may wish to consider Snakemake's native support for rule-specific Conda environments using the conda: directive within individual rules. The Snakemake documentation on automatic deployment of software dependencies provides a helpful overview of this approach (https://snakemake.readthedocs.io/en/v8.0.0/tutorial/additional_features.html).

Referee #5

(Remarks to the Author)
I co-reviewed this manuscript with one of the reviewers who provided the listed reports.

(Remarks on code availability)

Responses to Reviewers' Comments

We thank the reviewers for their constructive comments, which have helped us significantly improve our manuscript. We have addressed all the reviewers' comments point-by-point below (our responses are in blue) in this document and made the corresponding changes in the manuscript files (highlighted in yellow). Below is a summary of the main changes.

Detailed description of the PIGA workflow

To improve the clarity of our manuscript, we have added a detailed step-by-step description of the PIGA workflow in the **Supplementary Note** and **Supplementary Figures 3-10**. The key components of the PIGA pipeline are summarized as follows:

1. **Comprehensive SNV detection and phasing**: We leveraged hybrid sequencing data to create a comprehensive SNV callset, which improved the ability to detect SNVs in repetitive regions compared to using short-read data alone. These SNVs were then phased into haplotypes using long-read information, providing accurate phasing blocks that are essential for diploid genome assembly.
2. **Personalized reference genome construction**: We generated a personalized reference for each individual by modifying the reference genome with homozygous variants genotyped from an external graph-based pangenome. For our dataset, this personalized reference improved read alignment and enabled more long reads to be partitioned into haplotypes, compared to using a standard reference or a *de novo* squashed assembly.
3. **Reference-guided draft diploid genome assembly**: Using the personalized reference as a guide, we generated draft diploid assemblies for each individual. This approach demonstrated a significant improvement in assembly completeness compared to the de novo assembly using partitioned long reads. Multiple rounds of polishing with both long and short reads were then conducted, which continuously improved the quality of the draft assembly. Although these generated draft assemblies still contained incomplete regions, they provided a wealth of haplotype-resolved sequences across most of the genome and served as a crucial foundation for generating final diploid assemblies.
4. **Pangenome construction and simplification**: A population-scale pangenome was constructed using the reference genomes, high-coverage Hifiasm assemblies, and the PIGA draft assemblies. To mitigate pangenome errors introduced during assembly or pangenome alignment, we trained a multi-layer perceptron classifier to distinguish between true variants and erroneous nodes/edges, based on a rich set of pangenome-derived features. This machine learning-based filtering outperformed simpler methods based solely on allele count, showing strong performance even in distinguishing singleton variants and

errors. The resulting filtered pangenome was significantly smaller, enabling efficient downstream analyses.

5. **Final diploid assembly generation**: We inferred diploid paths for each individual through the constructed pangenome by integrating multiple sources of information, including short reads, long reads, draft assembly haplotypes, and SNV haplotypes. The final diploid assembly was then reconstructed from these paths, with additional refinement steps to correct SV-discordant regions and clip k-mer erroneous regions.

In summary, the PIGA workflow provides a computational framework that employs a pangenome graph to perform population-scale, joint diploid genome assembly. It complements the current paradigm of high-coverage individual assembly (e.g., Hifiasm) and provides the possibility of fully utilizing cohort-level hybrid sequencing data, even at low or modest coverage. Despite being an initial attempt at this type of joint diploid assembly analysis, PIGA yields diploid assemblies that demonstrate a superior ability to detect genetic variants compared to other state-of-the-art methods.

Quality control for the 1KCP dataset

We acknowledge the reviewers' concern regarding dataset quality. To address this, we performed multiple quality control (QC) procedures to ensure the reliability of the final 1KCP dataset, as detailed below:

1. **Region-level QC**: Ideally, the PIGA workflow can fully embed genetic variants into an intermediate pangenome and reliably reconstruct haplotypes across all genomic regions. However, in practice, highly repetitive regions are still challenging to resolve for the current PIGA workflow because noisy non-HiFi long reads have limited power to resolve highly repetitive sequences, and such regions are often clipped during pangenome construction due to their extremely complex graph structures. In such cases, the PIGA workflow may aggressively borrow sequences of these regions from complete reference haplotypes (e.g., CHM13) to reconstruct the final assembly, even when they do not match the individual's actual haplotype, leading to local sequence discrepancies. While leveraging more accurate long reads and improving genome assembly and pangenome construction methods may address this issue in the future, we implemented a region-level QC strategy to filter these unreliable regions in this study. The procedure included the following steps:

i) We identified unreliable regions in the PIGA assemblies of three benchmark samples using Flagger. On average, 119.6 Mb of unreliable regions were detected per haplotype.

ii) These regions were lifted over to both GRCh38 and CHM13 references using minimap2 for alignment and rustybam for coordinate transformation, resulting in an average of 65.4 Mb (GRCh38) and 128.9 Mb (CHM13) of lifted unreliable regions per haplotype.

iii) Reference regions marked as unreliable in more than two haplotypes were defined as common unreliable regions. This yielded 53.7 Mb of such common unreliable regions in GRCh38, which were used in subsequent QC steps.

2. Gene-level QC: For gene-level analyses, we excluded genes with > 5% of their length overlapping the common unreliable regions. This filtering step removed 194 protein-coding genes and 355 noncoding genes. Consequently, the number of polymorphic gene clusters identified by Pangene was reduced from 743 to 735.

3. Pangenome QC: As pangenomes built from a large number of assemblies can accumulate significant errors, we developed a k-mer-based filtering method to remove non-reference nodes and edges in the pangenome that were unsupported by read k-mers. The procedure consisted of the following steps:

i) For each individual, we used GenomeScope2 to estimate a reliable k-mer coverage range, defining the lower bound as the 0.05th percentile of the single-copy heterozygous coverage distribution and the upper bound as the 99.5th percentile of the single-copy homozygous coverage distribution.

ii) A node or edge was considered supported in a haplotype if the proportion of its associated k-mers with coverage between the lower and upper bounds exceeded 20% of all associated k-mers with coverage below the upper bound.

iii) Nodes or edges supported by < 50% of their carrying haplotypes were filtered out from the pangenome.

iv) Graph tips introduced by this filtering were further removed to eliminate residual artifacts.

This QC reduced the size of non-reference sequences in the pangenome from 628.8 Mb to 405.3 Mb, of which 81.3% of the removed non-reference sequences were singletons

4. Pangenome variants QC: As a consequence of pangenome QC, all pangenome variants (small variants, SVs, and nested variants) derived from filtered nodes/edges were removed. Additionally, we filtered variants with > 20% of their length overlapping the common unreliable regions. As a result, the number of small variant sites decreased from 38.0 million to 35.4 million. The number of SV sites decreased from 280,503 to 110,530. The number of nested variant sites decreased from 1.81 million to 0.86 million.

5. Tandem Repeat (TR) QC: We filtered TR loci whose inferred sequences showed an average identity of < 80% to the corresponding assembled sequences across haplotypes. Additionally, TRs with > 20% of their length overlapping the common unreliable regions were removed. As a result, the number of

length-polymorphic TR sites decreased from 339,635 to 333,882. The number of motif-polymorphic TR sites decreased from 493,296 to 485,575.

6. **HLA allele QC**: We filtered HLA allele typing results with < 99% identity or < 95% coverage to their best-matching reference alleles. HLA genes with > 20% filtered alleles were excluded from downstream analysis. As a result, the number of HLA genes decreased from 39 to 36. The number of HLA alleles decreased from 1,431 to 1,348.

These multi-level QC processes lead to a more reliable 1KCP dataset for subsequent analyses. We have added a detailed description of these steps to the Methods section and updated all relevant results accordingly in the revised manuscript.

Referee #1 (Remarks to the Author):

The authors generate a low coverage long read data set of more than a thousand genomes from China. This is a valuable effort for the genomics community as such data sets are needed both for common disease association studies and as a reference resource to filter variants observed in rare disease genomes. At the same time, the present version of the manuscript does not provide enough details on the quality of the resulting genome reconstructions and I am providing specific comments and suggestions about this below, among other comments.

Re: We thank the reviewer for acknowledging our work as a valuable resource for future clinical and population genetics studies. We also appreciate the constructive comments and suggestions the reviewer provided, which have helped us improve the quality and clarity of our manuscript. Guided by the reviewer's comments, we have now performed additional analyses and thoroughly revised the manuscript to provide a comprehensive quality assessment. We hope the added details, which we address point-by-point below, now fully clarify these aspects of our work.

Major comments

- The abstract and title advertise "diploid genome assemblies of 1,116 Chinese individuals". Typically the term "genome assembly" is used to refer to de novo assembly without the use of a reference genome, but this is not what was done here. I suggest to use a different term to avoid confusion. The term "Pangenome-Informed Genome Assembly" is more clear, but only used further down in the main text.

Re: We agree that the term “pangenome-informed genome assembly” is more precise and avoids potential confusion with *de novo* assembly. The reviewer's comment prompted us to reconsider how to best

describe our resource, which is composed of two types of assemblies: (1) *de novo* assemblies for a subset of samples generated with Hifiasm, and (2) pangenome-informed assemblies for the larger cohort generated with our PIGA workflow.

To address this complexity and accurately reflect our work, we have taken the following steps:

(1) We have revised the title to "Diploid genomes of 1,116 Chinese individuals empower medical and population genetics" to avoid confusion.

(2) We now describe the composition of the 1,116 diploid assemblies in the Abstract and Introduction sections, and explain the distinction between *de novo* assembly and pangenome-informed assembly in the Results section. We only use a general term like "diploid assemblies" in broad summaries after the distinction has been established.

Updated text (lines 26-28): "Here, we generated 1,116 diploid genome assemblies, including 55 *de novo* assemblies and 1,061 pangenome-informed assemblies, with an average size of 2.98 Gb and a mean quality value (QV) of 46, as part of the 1000 Chinese Pangenome (1KCP) project."

Updated text (lines 69-72): "To address these challenges, we generated 1,116 diploid genome assemblies, including 55 *de novo* assemblies and 1,061 pangenome-informed assemblies, as part of the 1000 Chinese Pangenome (1KCP) project, constructed a functionally annotated pangenome, and provided an extensive catalog of diverse genetic variants."

Updated text (lines 96-100): "Unlike *de novo* assembly methods that use data from single individuals, the PIGA workflow employs a pangenome-based framework to integrate sequence information across the entire cohort for joint diploid genome assembly. We generated de novo assemblies for the high-coverage samples using Hifiasm, and pangenome-informed assemblies for the modest-coverage samples using PIGA."

- More information on the characteristics of the PacBio sequencing data are needed. The "CCS --all" mode used for data generation is unusual and discontinued (according to the PacBio website https://ccs.how/faq/mode-all.html). That is not necessarily a problem, but more details such as error rate and read length distributions would be important to include for such a non-standard data set.

Re: The "CCS --all" mode produces a representative read for each zero-mode waveguide (ZMW), either by generating a consensus sequence when sufficient subreads are available or by selecting the median-length subread otherwise (**Figure R1a** below). This process yields two types of reads: HiFi reads and non-HiFi reads. Using sample 1KCP-00001 as a representative example, we show that the non-HiFi reads

exhibit lower accuracy (median: 90.2%) but a higher read length N50 (28.6 kb) compared to the HiFi reads (median accuracy: 99.6%; length N50: 23.6 kb) (**Figure R1b** below).

While standard pipelines typically use HiFi reads only, we showed that incorporating additional non-HiFi reads led to a 2.7-fold increase in total sequencing coverage (**Figure R1d** below). Read alignment analysis of 10 randomly selected 1KCP samples further demonstrated that using all PacBio reads substantially reduced the proportion of uncovered genomic regions from an average of ~14% with HiFi-only data to ~4% (**Figure R1c** below). Moreover, incorporating all PacBio reads for genotyping with WhatsHap significantly improved the SNV detection performance compared to the initial HiFi-only calling with DeepVariant (**Figure R2** below).

Collectively, these findings indicate that non-HiFi reads, despite their lower accuracy, provide valuable information for genomic analysis, particularly in modest- or low-coverage settings. This observation also motivated the development of the PIGA workflow, which is tailored to fully utilize such population-scale hybrid sequencing datasets.

We have revised the manuscript to provide a more detailed characterization of the PacBio sequencing data generated using the "CCS --all" mode.
Updated figures: **Supplementary Figures 2** and **4a** (corresponding to **Figures R1** and **R2** below).
Updated text (lines 608-613): “With the --all option, CCS generates one representative read for each zero-mode waveguide (ZMW), either by generating a consensus sequence when sufficient subreads are available or by selecting the median-length subread otherwise (Supplementary Figure 2a). This process yields two types of reads: HiFi reads and non-HiFi reads. While non-HiFi reads have lower accuracy than HiFi reads, their inclusion resulted in a 2.7-fold increase in total sequencing coverage compared to using HiFi reads alone (Supplementary Figure 2b).”

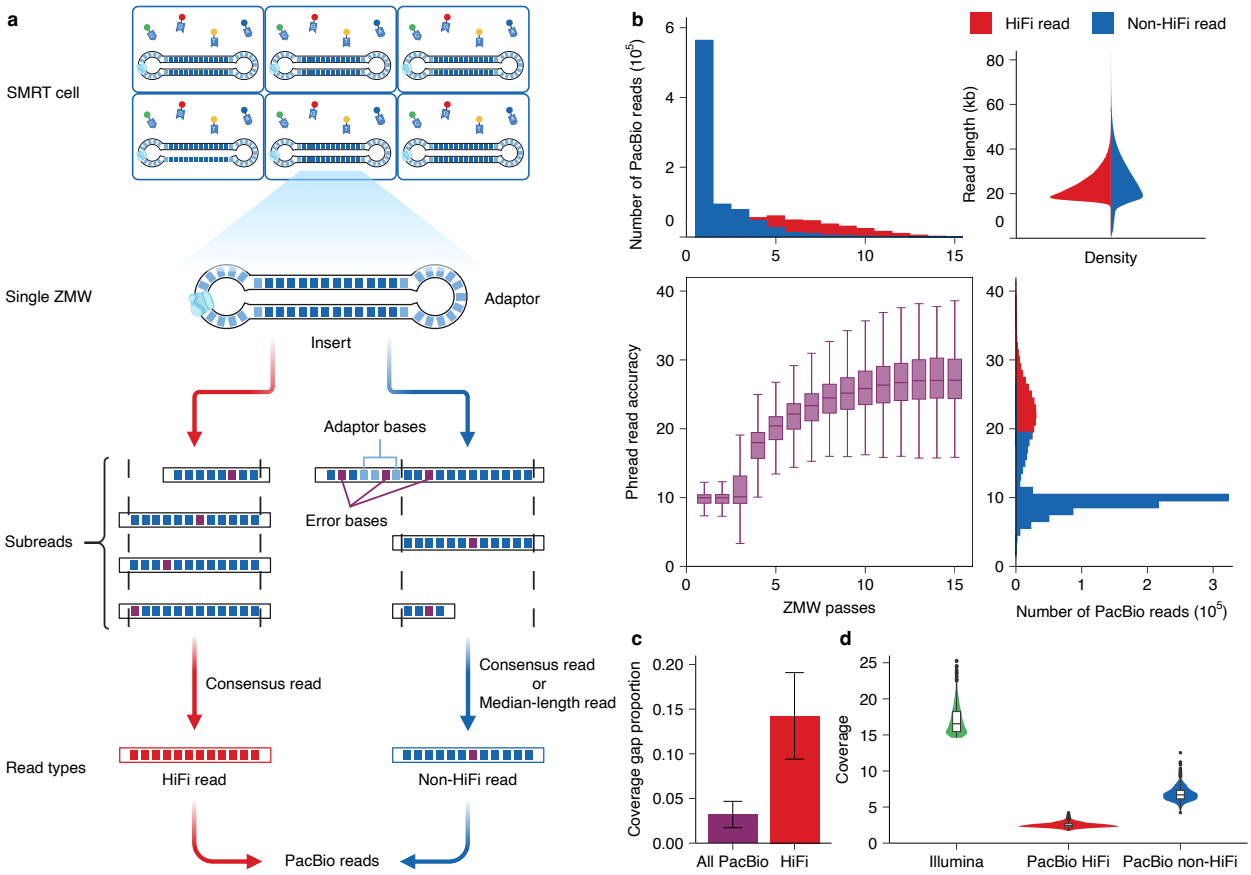

Figure R1. Read characteristics of the 1KCP modest-coverage dataset. a, Schematic illustration of the read generation process under the "CCS --all" mode. HiFi reads are generated as high-accuracy consensus sequences from multiple subreads, whereas non-HiFi reads typically originate from a consensus or median-length read with lower accuracy. **b**, Characteristics of PacBio reads from a representative sample, 1KCP-00001. Shown are distributions for the number of ZMW sequencing passes (top left), Phred read accuracy (bottom right), and read length (top right), stratified by HiFi and non-HiFi types. The relationship between ZMW sequencing passes and Phred read accuracy is shown (bottom left), with error bars representing 95% confidence intervals (CIs). The relationship between ZMW passes and read accuracy is also shown, with error bars indicating the 95% CIs (bottom left). **c**, Proportion of genomic regions with zero read coverage (gaps) using either all PacBio reads or only HiFi reads. Error bars represent 95% CIs across 10 randomly selected 1KCP samples. **d**, Coverage distribution of Illumina, PacBio HiFi, and PacBio non-HiFi reads in the 1KCP modest-coverage dataset.

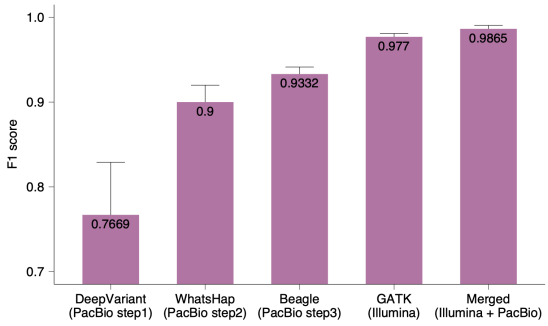

Figure R2. SNV calling performance at progressive stages of the PIGA SNV detection pipeline. F1 scores were evaluated using ten benchmark samples. The first three bars show the performance at successive steps of the PacBio-based pipeline: (1) initial calling with DeepVariant, (2) genotyping with WhatsHap, and (3) LD-based refinement with Beagle. For comparison, the performance of the GATK pipeline using only Illumina short-read data is shown. The final bar (Merged) represents the results after integrating variant calls from both PacBio and Illumina datasets. Error bars indicate 95% confidence intervals.

- Good to see that the authors have created a website, which hosts the graphs and variant sets. No individual level data appears to be available from that website, i.e. the VCF files do not have genotypes and the graphs do not have paths corresponding to the assemblies. In "Data availability" the authors say that "The genome assembly data generated in this study can be accessed upon request at the National Genomics Data Center (https://ngdc.cncb.ac.cn).", but do not give an accession number. I went to that website and was not able to easily find the data set. Could the authors clarify how exactly access can be requested? Is access possible worldwide or is it restricted to applicants from within China?

Re: We thank the reviewer for this feedback and apologize for the initial lack of clarity regarding data access. We are committed to sharing the 1KCP dataset with the global research community to the greatest extent permitted by national regulations. All individual-level data, including raw sequencing reads, genome assemblies, and variant genotypes, have been deposited in the repositories of the National Genomics Data Center (NGDC), specifically in the Genome Sequence Archive for Human (GSA-Human: https://ngdc.cncb.ac.cn/gsa-human), Genome Warehouse (GWH: https://ngdc.cncb.ac.cn/gwh), and Genome Variation Map (GVM: https://ngdc.cncb.ac.cn/gvm). Data for the 10 overlapping benchmark samples have been made publicly available under BioProject accession number PRJCA039156 (https://ngdc.cncb.ac.cn/bioproject/browse/PRJCA039156). The full cohort data will be accessible to all international researchers (BioProject accession number: PRJCA036466) through a standard NGDC data access application process upon the provisional acceptance of this paper for publication.

We have revised the Data Availability section in the manuscript to improve clarity regarding data access.
Updated text (lines 981-988): "The summary-level data, including the 1KCP pangenome graph, pangenome annotations, variant sites, and eQTL summary statistics, are available at the 1KCP data portal (https://yanglab.westlake.edu.cn/1kcp). The individual-level data, including raw sequencing reads, genome assemblies, and variant genotypes, are available in the repositories of the National Genomics Data Center (NGDC), specifically in the Genome Sequence Archive for Human (GSA-Human: https://ngdc.cncb.ac.cn/gsa-human), Genome Warehouse (GWH: https://ngdc.cncb.ac.cn/gwh), and Genome Variation Map (GVM: https://ngdc.cncb.ac.cn/gvm), under BioProject accession numbers PRJCA036466 and PRJCA039156."

- The whole paper hinges on the assemblies (with PIGA) and the resulting call set, but in its present form, the evaluation of their quality is not sufficient and not reproducible. Some questions/suggestions:

1) At least a small number of benchmark samples (e.g. the three samples used in this study) should be made publicly available, including the input reads.

Re: In the revised manuscript, we clarify that the "benchmark samples" refer to the ten overlapping samples from both high- and modest-coverage datasets, which were used to evaluate individual steps of the PIGA workflow. A specific subset of three samples, for which we generated assemblies using both PIGA and Hifiasm, was used for downstream evaluations.
Updated text (lines 100-102): "Ten overlapping samples were used to benchmark individual steps of the PIGA workflow, and a subset of three samples with assemblies from both methods was used for downstream evaluation."

We have now made the raw sequencing reads and genome assemblies of these ten benchmark samples publicly available through the National Genomics Data Center (NGDC) under BioProject accession number PRJCA039156 (https://ngdc.cncb.ac.cn/bioproject/browse/PRJCA039156).

2) Could you quantify the added value of the long reads? From short reads alone, one can expect a QV value >40 even using a standard pipeline (see Logsdon et al., bioRxiv, Figure 3b; https://www.biorxiv.org/content/10.1101/2024.09.24.614721v1).

Re: We agree with the reviewer that short-read-only methods can generate personal consensus sequences with high accuracy (e.g., QV > 40). However, the key value of incorporating long reads in the PIGA workflow is to generate population-scale draft diploid assemblies for the entire cohort (as detailed in the Supplementary Note). Although these draft assemblies are not fully complete, alignment results show that 96.45% of 10-kb windows in the CHM13 reference genome are covered by more than 50% of the assemblies, indicating that they provide extensive haplotype-resolved sequences across the majority of the genome (**Figure R3** below). These assemblies are then used to construct a large-scale, internal pangenome, which substantially increases the accessible haplotype diversity for diploid path inference, compared to the HGSVC3 pipeline, which relies on a smaller, external pangenome panel. Additionally, long-read data were further leveraged in three forms: SNV haplotypes, draft assemblies, and raw reads, providing complementary evidence alongside short-read data for diploid path inference.

The resulting increase in haplotype diversity enhances the detection of rare variants and alleles across individuals in the population, which is a central aim of our study. Specifically, we show that the full 1KCP assembly panel (N = 1,116) captures 4.3-fold more rare SVs (AF < 0.01) and detects 3.1-fold more TR alleles than a smaller panel using only Hifiasm assemblies (N = 55) (**Figure R4** below).

Beyond population-scale improvements, the benefits of incorporating long reads are also evident at the individual level. To demonstrate this, we followed the HGSVC3 pipeline to generate personal consensus sequences using an external pangenome consisting of two references (CHM13 and GRCh38) and 52 1KCP Hifiasm assemblies (excluding the three benchmark samples). We then compared the Dipcall-derived variant callsets from the PIGA assemblies against those from the consensus sequences generated with the HGSVC3 pipeline in three benchmark samples. Several state-of-the-art read-based variant detection methods were also included for comparison. Benchmarking results show that PIGA assemblies consistently outperform all other methods in detecting SNVs, indels, and SVs, highlighting the advantage of fully leveraging long reads in the PIGA workflow for comprehensive variant discovery (**Figure R5** below).

We have incorporated these results into the revised manuscript.
Updated figures: **Supplementary Figures 7c**, **37**, and **10** (corresponding to **Figures R3-R5** below).
Updated text (lines 492-498): “This large-scale pangenome greatly enhances the representation of genetic variants, particularly rare and low-frequency variants (Supplementary Figure 37a). The improvement is reflected in the larger size of non-reference sequences, the increased number of variant sites, and the higher allele counts observed across these sites. Notably, TR sites exhibited extraordinary

diversity, accounting for a substantial proportion of multiallelic variants, underscoring the necessity of representing and imputing TRs using a large-scale haplotype panel (Supplementary Figure 37b)."

Updated text (Supplementary Note, lines 131-134): "Applying this pipeline to the entire cohort of 1,116 individuals, we generated draft assemblies with sizes ranging from 1.9 to 2.6 Gb. Although the completeness of these assemblies is limited, 96.45% of 10-kb windows in the CHM13 reference genome were covered by contigs from more than 50% of haplotypes (Supplementary Figure 7c)."

Updated text (Supplementary Notes, lines 226-237): "We also followed the HGSVC3 pipeline to construct personal consensus sequences and generate a variant callset using Dipcall. Specifically, the PanGenie callset was generated by genotyping 12,992,029 variants from a Minigraph-Cactus pangenome comprising two references (CHM13 and GRCh38) and 52 1KCP Hifiasm assemblies (excluding 3 benchmark samples), and then filtered using a trained SVM model. In addition, 34,761,465 GATK-specific variant sites not present in the PanGenie callset were incorporated into the merged callset. Genotypes with a quality score below 10 were set to missing, and the merged callset was phased using SHAPEIT5. Personal consensus sequences were generated by modifying the CHM13 reference with each individual's final variant callset using BCFtool consensus.

Benchmarking results demonstrated that the PIGA assemblies consistently outperformed all other methods in detecting SNVs, indels, and SVs, underscoring the advantages of the PIGA assembly workflow for comprehensive variant discovery (Supplementary Figure 10)."

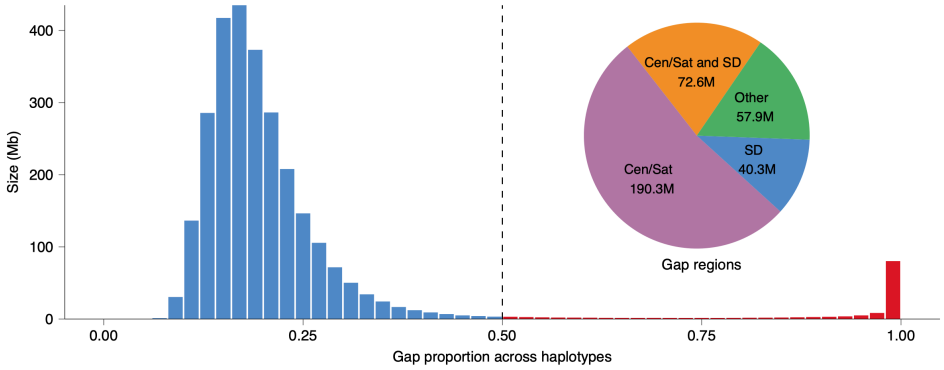

Figure R3. Alignment coverage of the 1KCP draft assemblies against the CHM13 reference. The histogram shows the distribution of gap proportions, defined as the fraction of haplotypes that fail to cover 10-kb windows of the CHM13 reference. The pie chart characterizes the genomic composition of these gap regions (gap proportion > 0.5).

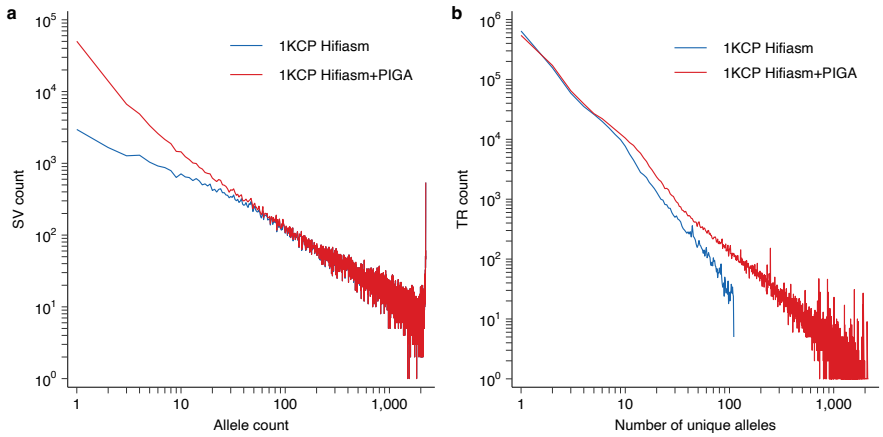

Figure R4. Comparison of variant discovery between the 1KCP full assembly panel (N = 1,116) and the Hifiasm-only assembly panel (N = 55). a, Number of SVs discovered in the two assembly panels, stratified by allele count in the full assembly panel. **b**, Number of TR sites stratified by the number of unique alleles per site in the two assembly panels.

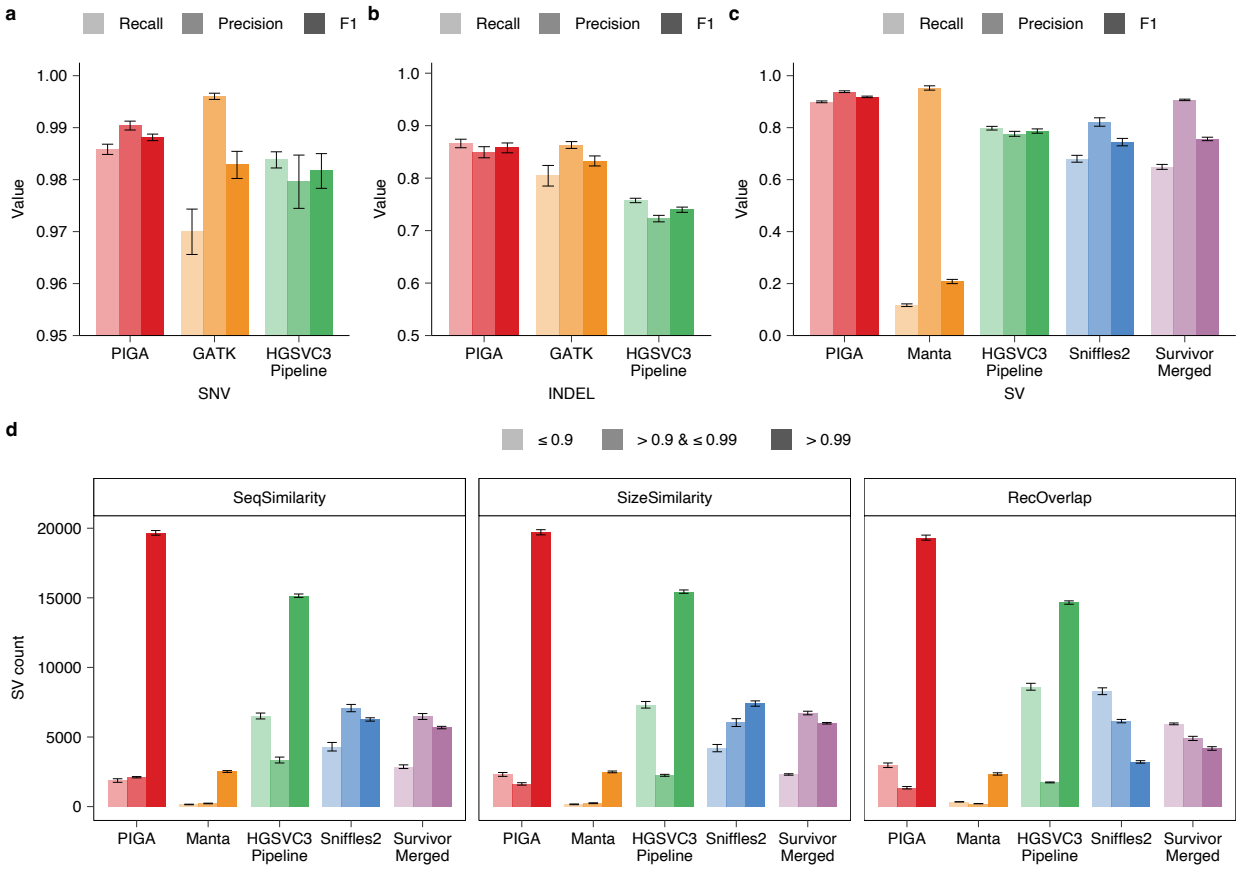

Figure R5. Benchmarking of variant detection performance across different methods in three benchmark samples. a, SNV detection performance evaluated using hap.py. Error bars indicate 95%

confidence intervals. **b**, Indel detection performance evaluated using hap.py. Error bars indicate 95% confidence intervals. c, SV detection performance evaluated using Truvari bench (--pick multi --sizemin 50). Error bars indicate 95% confidence intervals. d, Number of SVs stratified by sequence accuracy levels. Accuracy is evaluated using three metrics: sequence similarity (left), size similarity (middle), and reciprocal overlap percentages (right). Error bars indicate 95% confidence intervals.

3) The design rationale is somewhat similar to the 1000 Genomes Phase 3 data set (Sudmant et al., Nature, 2015), where low coverage data across many samples was used to discovery SVs. This is a good strategy to maximize the yield on rare SV alleles, but there is risk of an inflation with singletons. Figure 2b hints at an excess of singletons, but that is difficult to judge from this plots. I suggest to provide a log-log plot similar to Extended Data Figure 2 of Sudmant et al. (2015). In such a visualization, a linear relation would be the expected behavior, which makes it easier to assess (would also be good for Figure 1g, i.e. make the x-axis logarithmic, too). Typically, STR/VNTRs behave quite differently from other classes of SV and it might be helpful to create stratified versions for some of the plots.

Re: We thank the reviewer for this insightful comment. We agree that log-log plots provide an intuitive way to evaluate variant callsets, especially for identifying potential inflation of singletons in large-scale variant discovery.

To address this, we first examined the log-log distribution of SNVs as a reference, given their abundance and high calling accuracy. Our SNV callset exhibited a similar allele frequency spectrum to that of the 1000G East Asian SNV callset, with both showing non-linear patterns on a log-log scale (**Figure R6a** below). We acknowledge that, under idealized conditions with constant mutation rates and effective population sizes across generations, a linear relationship would be expected. However, due to the complexity of real population histories, we consider a smooth curve on a log-log plot to be a more robust indicator of callset quality.

We next generated a log-log plot for our SV callset. Consistent with the reviewer's observation, we detected an excess of singleton SVs, particularly evident as an elevated left-tail of the curve (**Figure R6b** below). Although our benchmarking showed that PIGA assemblies achieve high SV detection accuracy (accuracy: 0.978), we recognized that even low per-sample error rates can lead to substantial cumulative noise in large cohorts. This underscores the importance of systematic quality control to ensure the reliability of the final dataset. To mitigate this, we developed a k-mer-based filtering method to remove non-reference nodes and edges in the pangenome that are unsupported by read k-mers (as described at the

beginning of this document). This filtering reduced the total size of non-reference sequences in the pangenome from 628.8 Mb to 405.3 Mb (**Figure R7** below). We further filtered variants with > 20% of their length overlapping the common unreliable regions, defined using Flagger annotations. Together, these QC steps reduced the SV allele count from 376,495 to 164,216, with 82.1% of the excluded SVs being singletons. Importantly, the post-QC SV callset displays a much smoother curve on a log-log plot, indicating effective removal of error-driven singletons (**Figure R6b** below). Furthermore, we observed substantial reductions in singleton SV counts for both Hifiasm (from 111.5 to 49.4 singletons per individual on average) and PIGA (from 205.7 to 44.5 singletons per individual on average) assemblies, suggesting that singleton inflation is a common issue for both assembly types.

To support further evaluation, we followed the reviewer's suggestion and replotted Figure 2g using a log scale on the x-axis, and additionally replaced the y-axis from variant proportion to variant count to avoid potential bias from differences in total variant counts across callsets (**Figure R8** below). The result shows that the merged SV callset exhibits more consistent trends with the SNV callset compared to the unmerged SV callset. Additionally, we generated a stratified log-log plot to distinguish SVs by repeat composition (**Figure R9** below). The curve for TR SVs did not show a markedly different pattern compared to other SV categories.

These QC procedures are now described in detail in the Methods section, and the relevant results have been updated throughout the revised manuscript.
Updated figures: **Figures 2b**, **2g**, and **Supplementary Figure 24** (corresponding to **Figures R7b**, **R8**, and **R9** below).
Updated text (lines 170-171): "The refined subgraphs were then normalized, clipped, filtered by read k-mers, and merged into the integrated pangenome."
Updated text (lines 755-764): "For quality control, we developed a k-mer-based filtering method to remove non-reference nodes and edges in the pangenome that are unsupported by read k-mers. The procedure included the following steps: (1) For each individual, we used GenomeScope2 to estimate a reliable k-mer coverage range, defining the lower bound as the 0.05th percentile of the single-copy heterozygous coverage distribution and the upper bound as the 99.5th percentile of the single-copy homozygous coverage distribution. (2) A node or edge was considered supported in a haplotype if the proportion of its associated k-mers with coverage between the lower and upper bounds exceeded 20% of all associated k-mers with coverage below the upper bound. (3) Nodes or edges supported by < 50% of their carrying haplotypes were filtered out from the pangenome. (4) Graph tips introduced by this filtering were further removed to eliminate residual artifacts."

Updated text (lines 805-808): "For quality control, we filtered out variants with > 20% of their length overlapping common unreliable regions, treating the positions of nested variants as those of their parent variant sites. For TRs, we additionally removed loci whose inferred sequences exhibited an average identity of < 80% to the corresponding assembled sequences across haplotypes."

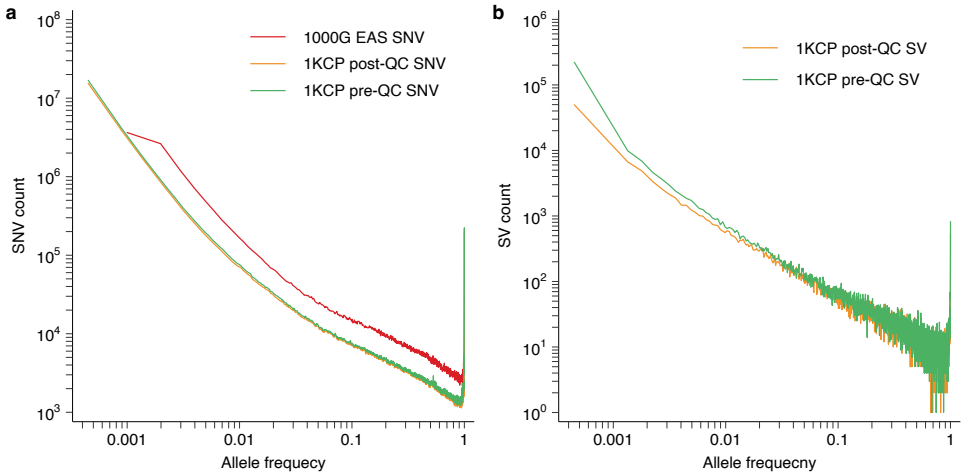

Figure R6. Allele frequency spectrum across different variant callsets. a, Allele frequency spectrum of SNV callsets. **b**, Allele frequency spectrum of SV callsets.

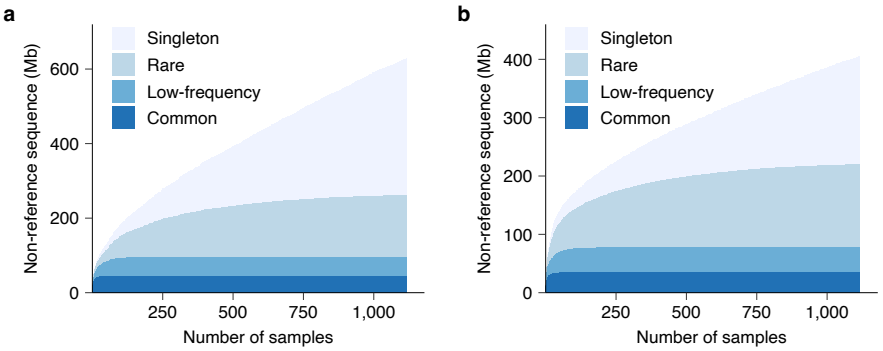

Figure R7. Cumulative size of non-reference sequences in the 1KCP pangenomes as the sample size increases. a, Unfiltered pangenome. **b**, Filtered pangenome.

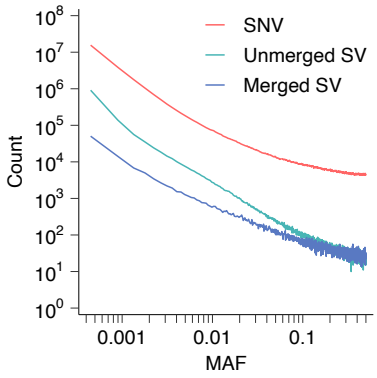

Figure R8. MAF spectrum of SNVs, unmerged SVs, and merged SVs.

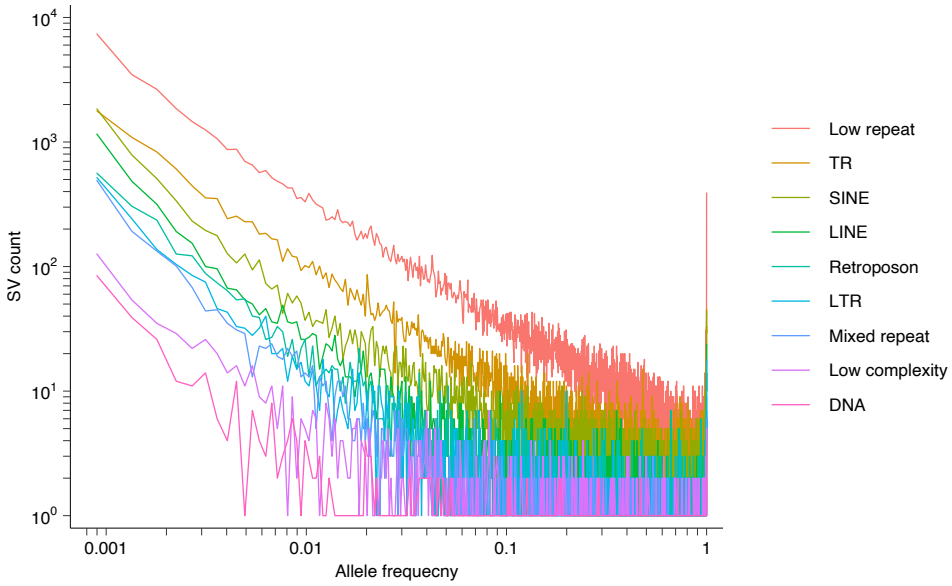

Figure R9. Allele frequency spectrum of SVs, stratified by repeat type. Repeat types include low repeat (SVs with <50% repeat content), mixed repeat (SVs with >50% content of multiple repeat types), and specific repeat classes (SVs with >50% content of single repeat types).

4) Flagger annotations are produced but not used much. For the benchmark samples, how many variants come from regions flagged in this process and could the authors use this to estimate how many of the singleton variants come from spurious calls?

Re: We have performed a more in-depth analysis leveraging the Flagger annotations. Specifically, we examined the distribution of errors in the PIGA assemblies of the three benchmark samples and found that 39% of k-mer errors and 48% of structural errors occurred within unreliable regions identified by Flagger

(**Figure R10** below). This observation prompted us to define 53.7 Mb of common unreliable regions in the GRCh38 reference shared across individuals, which were used in quality control steps to improve the reliability of the 1KCP dataset (as described at the beginning of this document).

Following the reviewer's suggestion, we estimated the false discovery rate (FDR) of SVs. We used SV calls derived from the pangenome constructed using both PIGA and Hifiasm assemblies of the three benchmark samples, together with the GRCh38 reference and other 1KCP assemblies. A given SV in the PIGA assemblies was considered a true positive if validated by at least one of the following pieces of evidence: (1) a direct match to the corresponding Hifiasm assembly, (2) a match identified using Truvari bench and refine, or (3) read-based validation using high-coverage HiFi data aligned with GraphAligner and genotyped using VG call.

Based on this strategy, we estimated the FDR of SVs stratified by minor allele frequency (**Figure R11** below). Before quality control, 77.3% of singleton SVs were estimated to be false positives. After applying our k-mer- and Flagger-based filters, this proportion was reduced to 26%. The overall FDR for the post-QC SV callset was estimated at 2.6%. Furthermore, we found that, on average, 12% of SVs located in Flagger-designated unreliable regions could not be validated by supporting evidence in each individual after QC, and about one-third of these unvalidated SVs (corresponding to 4% of all SVs in unreliable regions) were singletons.

We have incorporated these results into the revised manuscript.
Updated figure: **Figure 1c** and **Supplementary Figure 25** (corresponding to **Figures R10 and R11b** below).
Updated text (lines 116-122): "We further assessed the regional reliability using Flagger, finding that 39% of k-mer errors and 48% of structural errors in PIGA assemblies occurred within unreliable regions, mostly located in satellites and segmental duplication (Figure 1c and Supplementary Figure 11). Based on these results, we used Flagger annotations from benchmark samples to define 53.7 Mb of common unreliable regions in the GRCh38 reference, which were used in downstream quality control steps to improve the reliability of the 1KCP dataset (Methods)."
Updated text (lines 236-238): "Based on three benchmark samples, we estimated a false discovery rate (FDR) of 2.6% for the SV callset (Methods and Supplementary Figure 25)."
Updated text (lines 688-692): "The unreliable regions (including collapsed, duplicated, and error regions) in the PIGA assemblies were lifted over to both GRCh38 and CHM13 references, using minimap2 (-cxasm20) for alignment and rustybam (https://github.com/mrvollger/rustybam) for coordinate

transformation. Reference regions marked as unreliable in more than two haplotypes were defined as common unreliable regions."

Updated text (lines 812-818): "The FDR of the SV callset was estimated based on SV calls derived from the pangenome constructed using both PIGA and Hifiasm assemblies of the three benchmark samples, together with the GRCh38 reference and other 1KCP assemblies. A given SV in the PIGA assemblies was considered a true positive if validated by at least one of the following: (1) a direct match to the corresponding Hifiasm assembly, (2) a match identified using truvari bench and refine, or (3) read-based validation using high-coverage HiFi data aligned with GraphAligner and genotyped using VG call."

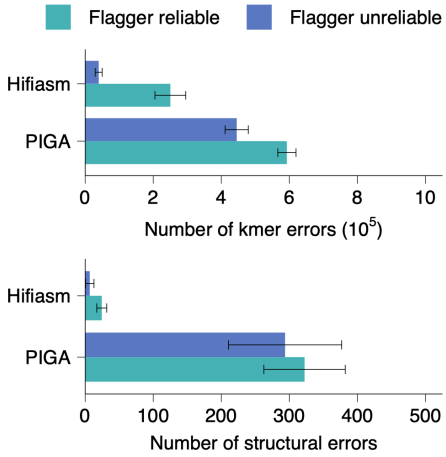

Figure R10. The number of k-mer errors and structural errors in Hifiasm and PIGA assemblies for the three benchmark samples, stratified by regional reliability as estimated by Flagger. Reliable regions correspond to haploid regions, while unreliable regions include collapsed, duplicated, and error-prone regions. Error bars represent 95% confidence intervals.

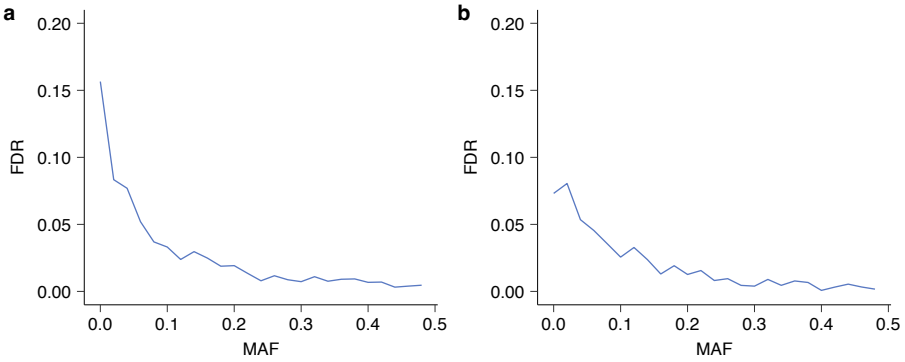

Figure R11. Evaluation of false discovery rate (FDR) of the 1KCP SV callsets across 25 MAF bins, estimated using three benchmark samples. a, Pre-QC SV callset. **b**, Post-QC SV callset.

- The HLA analysis lacks validation and QC. Lines 309-314: The number of fields in an allele is defined by the allele nomenclature and heavily depends on the gene itself. So this has little to do with new assemblies. Probably, simply selecting random gene alleles will produce similar results.

Re: To address the reviewer's concern about the reliability of the HLA analysis, we have added additional quality control and validation steps.

First, we applied a QC filter to retain only allele typing results with > 99% sequence identity and > 95% alignment coverage to their best-matching reference alleles. We further excluded three HLA genes (HLA-U, HLA-K, and HLA-W) in which > 20% of alleles failed these criteria. After filtering, the number of identified HLA alleles decreased from 1,432 to 1,348.

Second, we validated our HLA typing results using two complementary approaches. (1) In the three benchmark samples, we observed a four-field allele concordance rate of 96.5% (167/173) between PIGA assemblies and Hifiasm assemblies (**Figure R12a** below). (2) For the entire cohort, we performed independent HLA typing from short reads using HLA-HD, which produces three-field resolution calls. The comparison showed that the assembly-based typing achieved an average three-field allele concordance of 97.3% across genes relative to the HLA-HD results (**Figure R12b** below). These results confirm the reliability of our HLA typing.

We acknowledge that our current HLA typing relies on matching individual allele sequences to the most similar alleles in the IMGT database. While this approach is inherently limited by the scope of known reference alleles, it remains the standard paradigm widely used in read-based HLA typing studies to characterize HLA allele diversity in populations. Although our assemblies have the potential to identify novel alleles, we did not attempt novel allele discovery in this study, as such efforts require a more stringent validation framework (e.g., PCR-based confirmation) to ensure error-free allele reporting.

We have incorporated these QC and validation results into the revised manuscript.
Updated figure: **Supplementary Figure 32** (corresponding to **Figures R12** below).
Updated text (lines 341-343): “The typing results showed a mean three-field concordance rate of 97.3% on 22 shared genes with the short-read typing results obtained using HLA-HD (Supplementary Figure 32).”
Updated text (lines 891-894): “For quality control, we applied a filter to retain only allele typing results with > 99% sequence identity and > 95% alignment coverage to their best-matching reference alleles. We

further excluded three HLA genes (HLA-U, HLA-K, and HLA-W) in which > 20% of alleles failed these criteria. We also performed independent HLA typing from short reads using HLA-HD to validate the HiFiHLA typing results."

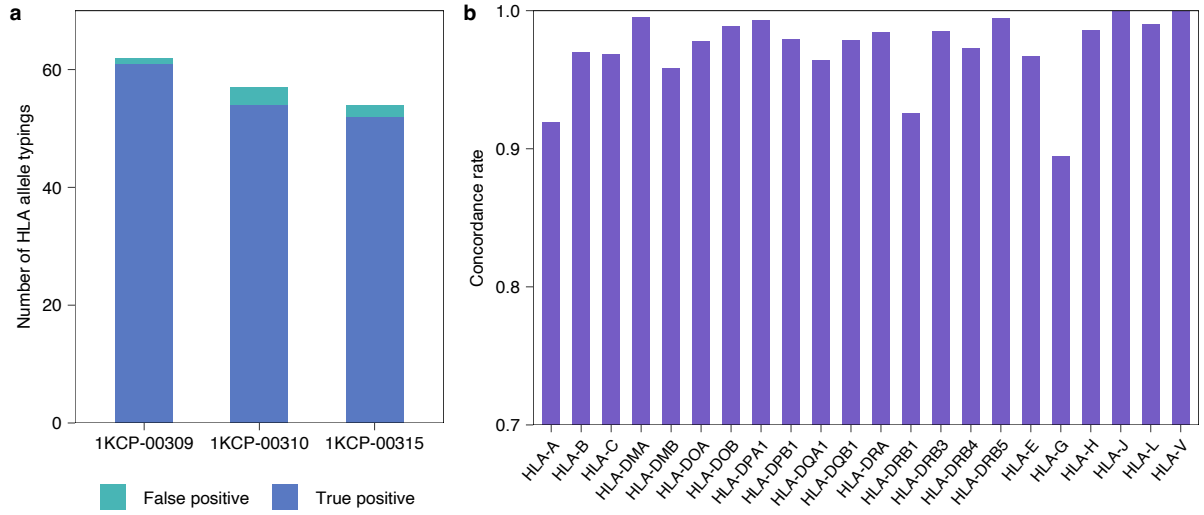

Figure R12. Evaluation of HLA typing accuracy. a, Number of validated four-field alleles in PIGA assembly-based HLA typing compared with Hifiasm assemblies in three benchmark samples. **b**, Three-field allele concordance rates of assembly-based typing results compared with short-read-based typing using HLA-HD on shared HLA genes.

- It would be important to put the imputation results into context by comparing to imputation using other reference panels and by contrasting imputation performance of different panels for different populations.

Re: As the 1KCP panel is specifically developed for imputing East Asian samples rather than being a multi-ancestry panel, we have now benchmarked its performance against existing panels targeting specific variant types, using 70 external East Asian samples from HPRC, HGSVC3, and CPC with high-quality assemblies.

For SVs, we compared imputation performance against the 1000G-ONT SV panel. The 1KCP panel achieved a higher mean r^2 across MAF ranges and imputed 26% more SVs with $r^2 > 0.8$ (**Figure R13a-b** below). For TRs, the 1KCP panel achieved a higher mean F1 score than the EnsembleTR panel (0.56 vs 0.46) (**Figure R13c** below), and uniquely enabled imputation of TR motif composition. For HLA alleles, while the 1KCP panel showed a slightly lower mean r^2 for shared alleles than the Four-digit multi-ethnic HLA v2 panel (0.89 vs 0.93), likely reflecting the difference in panel sample sizes (1,116 vs 20,349), it

enabled imputation of 2.7-fold more alleles with $r^2 > 0.8$ (**Figure R13d-e** below). These additional alleles were largely from low-resolution (one- and two-field) alleles of non-classical HLA genes and high-resolution (three- and four-field) alleles across all HLA genes. Finally, for nested variants within non-reference sequences of the pangenome, the 1KCP panel was the only resource capable of imputation.

We have incorporated these results into the revised manuscript.

Updated figure: **Supplementary Figure 36** (corresponding to **Figure R13** below)

Updated text (lines 460-472): "To benchmark the 1KCP panel against existing panels targeting specific variant types, we further evaluated imputation performance on 70 external East Asian samples with high-quality assemblies (Methods and Supplementary Table 34). For SVs, we compared imputation performance against the 1000G-ONT SV panel. The 1KCP panel achieved a higher mean r^2 across MAF ranges and imputed 26% more SVs with $r^2 > 0.8$ (Supplementary Figure 36a-b). For TRs, the 1KCP panel achieved a higher mean F1 score than the EnsembleTR panel (0.56 vs 0.46) (Supplementary Figure 36c), and uniquely enabled imputation of TR motif composition. For HLA alleles, while the 1KCP panel showed a slightly lower mean r^2 for shared alleles than the Four-digit multi-ethnic HLA v2 panel (0.89 vs 0.93), likely reflecting the difference in panel sample sizes (1,116 vs 20,349), it enabled imputation of 2.7-fold more alleles with $r^2 > 0.8$ (Supplementary Figure 36d-e). These newly imputed alleles were largely from low-resolution (one- and two-field) alleles of non-classical HLA genes and high-resolution (three- and four-field) alleles across all HLA genes. Finally, for nested variants within non-reference sequences of the pangenome, the 1KCP panel was the only resource capable of imputation."

Updated text (lines 968-978): "For comparison with the existing panels, we used 70 East Asian samples from CPC, HPRC, and HGSVC3 to construct the truth callsets and performed imputation benchmarking. The SV callset was generated using PAV and further merged using Jasmine. The HLA callset was generated using HifiHLA. Since the 1000G-ONT SV panel contains only SNV sites restricted to the UKB array, imputation for SV comparison was performed using the UKB pseudo-array genotypes as input. For TR and HLA comparisons, we used the Omni 2.5 pseudo-array genotypes as input. Imputations using the 1000G-ONT SV panel and the EnsembleTR panel were conducted with Minimac4. Imputation with the Four-digit multi-ethnic HLA v2 panels was performed on the Michigan online imputation server. For SV benchmarking, we used truvari bench (--pick ac -p 0.7 -P 0.7 -s 50) to match SVs across different callsets. For TR benchmarking, we assessed the performance of imputed TR alleles using Truvari bench (-s 5) and refine (-u) against the PAV callset, restricted to GIAB-defined TR regions."

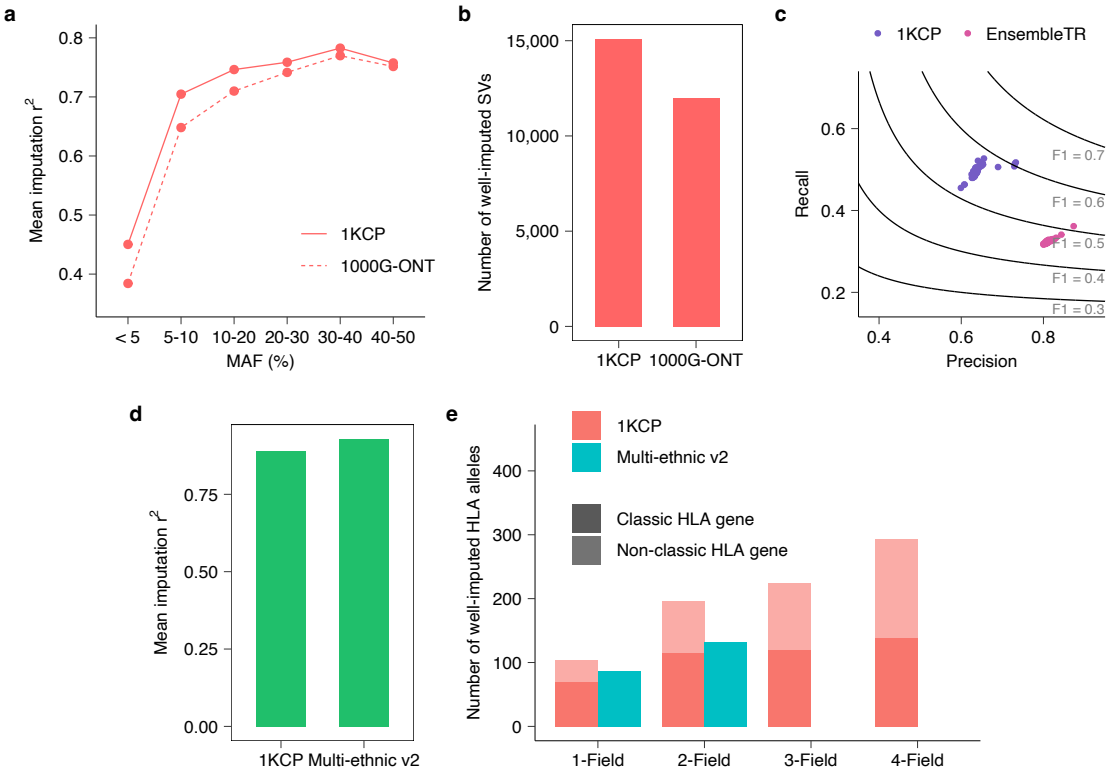

Figure R13. Comparison of imputation performance between the 1KCP reference panel and existing panels. a, Imputation performance (r^2) of the 1KCP and 1000G-ONT panels for shared SVs stratified by MAF bins. **b**, Numbers of well-imputed SVs ($r^2 > 0.8$) across the 1KCP and 1000G-ONT SV panels. **c**, Benchmarking of imputed TR alleles from the 1KCP and EnsembleTR panels, evaluated with Truvari bench (-s 5) and refine (-u), restricted to GIAB-defined TR regions. The black curves represent constant F1 scores from 0.3 to 0.7. **d**, Imputation performance (r^2) of the 1KCP and Four-digit multi-ethnic HLA v2 panels for shared HLA alleles. **i**, Numbers of well-imputed HLA alleles ($r^2 > 0.8$) from the 1KCP and Four-digit multi-ethnic HLA panels stratified by allele resolution levels and gene classes.

Minor comments

- Line 24: "The pangenome..." are you referring to one specific pangenome or rather pangenome-based methods more broadly? In case of the latter, you should rephrase this sentence.

Re: We thank the reviewer for pointing out this ambiguity. Our intention was to refer to pangenomes as a general class of reference resources, rather than to a single specific pangenome. To clarify this, we have revised the sentence in the abstract.

Updated text (line 24): “Pangenomes are revolutionizing our ability to resolve genomic regions with complex variations.”

- Lines 29-31: "... we constructed a pangenome comprising 628.8 million base pairs of sequences absent from the current references GRCh38 and CHM13, including 29.3 million base pairs of functional genic and regulatory elements": I am not convinced that these quantifications in Mbp are very meaningful as they will be quite strongly affected by alignment parameters. For this reason, it is probably not the most informative statistic for the Abstract. This also applies to Figure 1a.

Re: We agree with the reviewer that estimates of non-reference sequence size can be influenced by alignment parameters. Nevertheless, this metric is widely reported in current pangenome studies and provides a relatively meaningful measure of the additional sequence represented in the pangenome. We implemented a k-mer-based pangenome QC method to filter false-positive nodes and edges (as described at the beginning of this document), which improved the reliability of non-reference sequence size estimation. Additionally, we constructed an integrated pangenome including the 1KCP samples alongside 44 HPRC and 58 CPC samples, which serve as positive controls for evaluating the robustness of our pangenome construction pipeline. The resulting non-reference sequence sizes for CPC (193.2 Mb) and HPRC (204.2 Mb) assemblies are comparable to the value reported in their respective studies (Liao et al. 2023 Nature; Gao et al. 2023 Nature) (**Figure R14** below). Taken together, these analyses support the reliability of the non-reference sequence sizes reported in our study.

We have updated the relevant results in the revised manuscript:
Updated text (lines 184-188): “A total of 405.3 Mb of non-reference sequences were identified from the 1KCP assemblies, of which 277.5 Mb were not detected in CPC and HPRC assemblies (Figure 2a). Stratifying non-reference sequences by their frequency revealed 34.6 Mb as common (frequency > 0.05), 43.3 Mb as low-frequency (frequency > 0.01 and ≤ 0.05), 142 Mb as rare (frequency ≤ 0.01), and 185.4 Mb as singletons (unique to a single haplotype) (Figure 2b).”

Updated figure: **Figure 2a** (corresponding to **Figure R14** below)

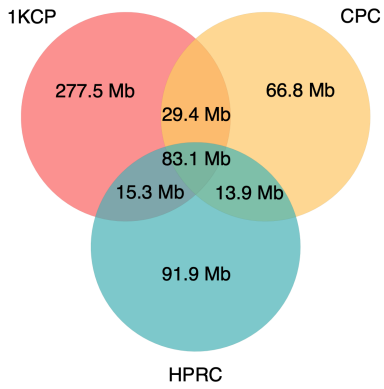

Figure R14. Comparison of non-reference sequence sizes among the 1KCP, CPC, and HPRC assemblies.

References

1. Liao, W.-W. et al. A draft human pangenome reference. ***Nature*** 617, 312-324 (2023).

2. Gao, Y. et al. A pangenome reference of 36 Chinese populations. ***Nature*** 619, 112-121 (2023).

- Line 38: What does "pan-variant" mean? Either omit or define it.

Re: We introduced “pan-variant” to describe a unified analytical framework that incorporates all major classes of genetic variation into downstream analyses, such as eQTL mapping and genotype imputation. This framework emphasizes the distinction from previous variant-specific analyses (e.g., SNP-based GWAS) and has the potential to provide more comprehensive insights into the roles of diverse variant types in shaping phenotypic variation.

To improve clarity, we have revised the Abstract to be more self-explanatory and introduced a formal definition of “pan-variant” in the Results section, where the concept is first described in detail.

Updated text (lines 35-40): “Coupled with the 1KCP gene expression data, we conducted pan-variant expression quantitative trait loci (eQTL) mapping to analyze diverse variant types, thereby identifying 3,256 eQTLs involving complex variants (SVs, TRs, and nested variants) and elucidating their regulatory complexity. Finally, we developed the 1KCP pan-variant imputation reference panel, providing multi-type genetic markers to enhance the resolution of future association studies.”

Updated text (lines 364-367): “This broad dataset enables the incorporation of all major classes of genetic variation into a unified analytical framework, referred to as pan-variant analysis. Such analyses have the potential to improve resolution over previous association studies that focused on only one or a

few variant types, providing a more comprehensive view of how diverse variants contribute to phenotypic variation."

- Line 63: "First, rare variants, which often exhibit high pathogenicity and play critical roles in inherited diseases" You are citing Kryukov et al. 2007 here, which is about rare missense mutations. I haven't gone back and read the whole study, but it seems questionable whether it supports the claim that rare variants in general often "exhibit high pathogenicity".

Re: We thank the reviewer for pointing this out. We agree that our original phrasing, "often exhibit high pathogenicity", was too strong when referring to rare variants in general, and that the citation of Kryukov et al. was not the most appropriate for such a broad claim. To address this, we have toned down the claim and replaced the citation with more suitable references (Kircher et al. 2014 Nature Genetics; Fiziev et al. 2023 Science; Cipriani et al. 2025 Nature) that discuss the role of rare variants more broadly.
Update text (lines 62-63): "First, rare variants, which generally exhibit higher pathogenicity than common variants (Kircher et al. 2014 Nature Genetics) and play critical roles in inherited diseases (Fiziev et al. 2023 Science; Cipriani et al. 2025 Nature), are underrepresented in small-scale pangenomes."

References

1. Kircher, M. et al. A general framework for estimating the relative pathogenicity of human genetic variants. ***Nature Genetics*** 46, 310-315 (2014).
2. Fiziev, P. P. et al. Rare penetrant mutations confer severe risk of common diseases. ***Science*** 380, eabo1131 (2023).
3. Cipriani, V. et al. Rare disease gene association discovery in the 100,000 Genomes Project. ***Nature***, 1-9 (2025).

- Line 81: "capture the full spectrum of genetic variations within the Chinese population (Supplementary Figure 1)" That figure is a PCA plot, so that does not support the statement that the "full spectrum of genetic variations" is captured. I wasn't able to find any information on how this PCA is done, i.e. how the set of input variants is defined and how exactly the PCA is run.

Re: Our original intent was to use the PCA plot to show the ancestry composition of the 1KCP samples. To clarify this, we have revised the Results section and added a description of the PCA in the Methods section.

Updated text (lines 81-85): “The 1KCP project enrolled 1,379 participants, the majority of whom clustered with Han Chinese reference populations based on principal component analysis (PCA) (Methods and Supplementary Figure 1). Both short-read and long-read whole-genome sequencing (WGS) were performed on a subset of 1,144 samples, aiming to establish a large-scale diploid assembly panel and an extensive catalog of genetic variation for medical and population genetics applications.”

Updated text (lines 646-656): “**Principal component analysis**

To place the 1KCP samples in a global genetic ancestry context, we performed a Principal Component Analysis (PCA) using SNV genotypes from the 1KCP samples combined with unrelated samples from the 1000 Genomes Project (1000G) and the Human Genome Diversity Project (HGDP). For the 1KCP samples, short-read data were aligned using BWA-MEM v0.7.17, and SNV genotypes were called using the GATK v4.2.6.1 workflow. Genotypes at overlapping SNV sites across the three datasets were merged to create two analytical sets: (i) a global set with all samples, and (ii) an East Asian-only set. For each set, we retained biallelic SNVs with minor allele frequency (MAF) > 0.05, Hardy-Weinberg Equilibrium (HWE) p-value > 1×10^{-6}, and genotype missingness < 10%. The remaining SNVs were further pruned to remove sites in LD using PLINK2 (--indep-pairwise 50 5 0.5). PCA was then conducted on this pruned dataset using PLINK2.”

- Line 158: "we mapped the haplotype coverage of pangenome segments to CHM13". I think "haplotype coverage" not a widely used term, so it is best to rewrite this sentence. Perphaps the authors meant "we mapped the haplotype-covered pangenome segments to CHM13".

Re: To improve clarity, we have revised the sentence to more accurately describe our analysis.

Updated text (lines 176-178): “To assess the quality of the 1KCP pangenome, we projected the number of haplotypes covering each pangenome segment onto its corresponding coordinates on the CHM13 genome.”

- Line 192-195: Could you indicate how many of the variants nested within reference SV sites occur in tandem repeat contexts? That is, for how many is the reference SV a TR expansion/contraction?

Re: Following the reviewer’s suggestion, we found that 30.5% (0.230 M/0.756 M) of nested variant sites within reference SV sites occur in tandem repeat contexts. This result has been incorporated into the revised manuscript.

Updated text (lines 216-217): “Among the nested variants within reference SV sites, 30.5% were located in TRs, likely reflecting TR expansion or contraction events.”

- Figure 2f is unclear to me. Is it saying that there is <10 SNVs and indels per sample added? That cannot be correct. Or is it the mean number of alleles per variant site? In that case it'd be good to show it in a way that one can verify that indeed almost all SNPs are (and remain) biallelic. In any case, please amend the caption to make it more clear what is shown.

Re: We thank the reviewer for pointing out the ambiguity in Figure 2f. This figure was intended to illustrate the mean observed number of alleles per variant site as the sample size increases. To improve clarity, we have revised the y-axis title accordingly. Furthermore, as suggested by Reviewer #4 (Minor comment 1), we have removed SNVs and indels from this figure and now focus on characterizing the allelic complexity of SVs. The updated figure has been included in the revised manuscript.

Update figure: **Figure 2f** (corresponding to **Figure R15** below)

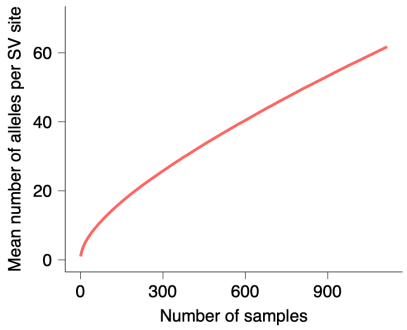

Figure R15. The mean observed number of unmerged alleles per SV site as the sample size increases.

- Line 203: could the authors please show average counts for TRs and non-TRs?

Re: We have estimated that the average distinct allele count is 99.5 for non-singleton TR SVs and 64 for non-TR SVs. These results have been incorporated into the revised manuscript.
Updated text (lines 223-226): "TR SV sites contain significantly more distinct alleles than non-TR SV sites, with an average of 99.5 versus 64 alleles per non-singleton site (Wilcoxon rank-sum test, $P < 1\times10^{-100}$), reflecting the high multiallelic complexity of TRs."

- Line 209: rephrase "biallelic alleles"

Re: We have revised the phrase "biallelic alleles" to "constituent biallelic variants".
Updated text (lines 230-232): "After decomposing the multiallelic sites into multiple biallelic variants, we obtained the 1KCP SV callset with 376,495 merged SV alleles (Methods)."

- Lines 210-212: That is impossible to judge from the figure, please provide a version with x-axis in log scale (see major comment above).

Re: We have replotted Figure 2g with the x-axis in log scale (**Figure R8** above).

- Line 213: What was the fraction of SVs in Hardy-Weinberg equilibrium before merging? Can the authors put 86% into context? Also please provide the details of the HWE test done in the Methods.

Re: We have now calculated the HWE pass rate for the unmerged SV set, which is 88%, largely consistent with the 87% observed for the merged SV set. We have incorporated this result and a detailed description of the HWE test procedure into the revised manuscript.
Updated text (lines 235-236): "Moreover, 87% of the merged SV alleles with MAF > 0.01 were in Hardy–Weinberg equilibrium (HWE), comparable to the 88% observed among unmerged SVs."
Updated text (lines 811-812): "For evaluating the SV callset, HWE p-values were calculated for SV alleles using the fill-tags plugin in BCFtools v1.20. SVs with HWE p-values > 1×10^{-6} were considered to be in HWE."

- Line 225/226 and Supp Fig 15: How exactly was that edit distance computed as (to my knowledge) Vamos does not provide the exact allele sequence. That should be described in Methods.

Re: We apologize for the previous ambiguity regarding the calculation of edit distance. In our analysis, the edit distance refers to the difference between repeat unit representations of paired TR alleles reported by vamos at corresponding TR sites, which is computed using the edlib. Alleles from different callsets were paired by minimizing the total edit distance across all possible allele pairs. We have clarified this approach in the revised manuscript.
Updated text (lines 248-252): "Benefiting from the high accuracy of assemblies, 98.2% of TR loci showed high concordance (≤ 1 edit distance in repeat unit representations of TR alleles) between Hifiasm and PIGA assemblies, whereas only 75.3% showed high concordance with Hifiasm assemblies when directly genotyped from long reads (Methods and Supplementary Figure 27)."

Updated text (lines 824-827): "The PIGA and long-read TR callsets from three benchmark samples were compared against the Hifiasm TR callset by using edlib to calculate edit distances between repeat unit representations of paired TR alleles at corresponding TR sites. Alleles from different callsets were paired to minimize the total edit distance across all possible allele pairs."

- Line 236: It would be good to mention that the number of variants is relative to the GRCh38 reference genome.

Re: We have clarified in the revised manuscript that the reported variant counts are relative to the GRCh38 reference genome.
Updated text (lines 259-262): From this catalog, we estimated that, relative to the GRCh38 reference genome, a typical Chinese diploid genome contains, on average, 5.09 million small variants, 23,617 SVs, 0.35 million nested variants, 174,318 TR length variants, and 496,621 TR motif variants (Supplementary Table 14).

- Line 245: The authors write about elevated proportion of rare SVs, so it is better to also mention the fraction of non-gene altering rare SVs.

Re: We now report the fraction of intergenic rare SVs in the revised manuscript.
Updated text (lines 269-272): "These exon-overlapping SVs, capable of altering gene transcripts, exhibit strong purifying selection, as reflected by their higher proportion of rare SVs (AF ≤ 0.01) compared with intergenic SVs (58.8%) and SVs in other genomic regions (Figure 3a and Supplementary Table 16)."

- Line 277: judging by Fig.3d, there is only one sample with expanded TR, is Fig.3e based on one sample only?

Re: We have clarified in the text and figure legend that the TR expansion is observed in a single individual.
Updated text (lines 303-305): "Among these events, a CGG expansion in the 5' UTR of *GIPC1*, carried by one individual, exhibited allele-specific hypermethylation, consistent with characteristics of folate-sensitive fragile sites and warranting further investigation (Figures 3d–e)."
Updated text (lines 1056-1057): "Methylation rates around the *GIPC1* TR site for expanded and unexpanded alleles in the individual carrying the expansion."

- Fig.3c: very hard to see if any samples went out of the normal range (for example for DAB1 and FGF14, mentioned on line 264).

Re: We have updated Figure 3c to additionally display circles representing samples with TR alleles exceeding the pathogenic threshold (**Figure R16** below).
Updated figure: **Figure 3c** (corresponding to **Figure R16** below)

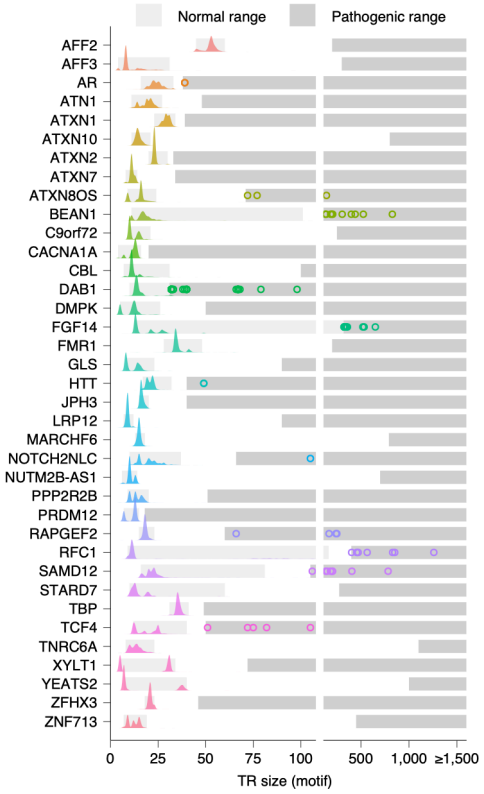

Figure R16. Size distribution of 37 disease-associated genic TR loci. Light grey boxes denote the normal TR size range observed in 99% of individuals from the 1KCP dataset, while dark grey boxes denote the pathogenic TR size range derived from the STRchive database. Circles indicate TR alleles with sizes exceeding the defined TR pathogenic threshold.

- Fig.3d: left and right subplots have different y-axis. Currently this is very misleading, should either use the same y-axis, or clearly highlight the difference.

Re: We have revised Figure 3d to include a rug plot and to apply the same y-axis range to both the left and right subplots (**Figure R17** below).
Updated figure: **Figure 3d** (corresponding to **Figure R17** below)

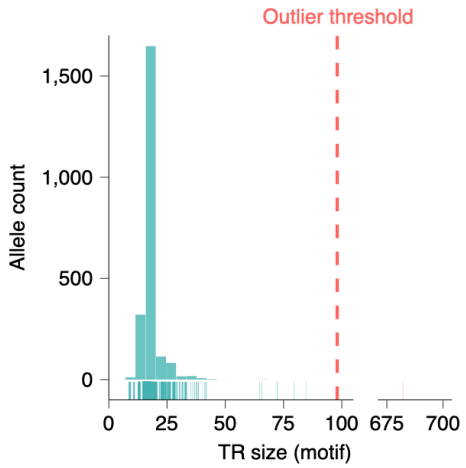

Figure R17. Size distribution of the *GIPC1* TR site. The vertical green ticks represent the sizes of normal TR alleles, and the red tick indicates the size of the outlier allele. The vertical red dashed line marks the outlier threshold determined by the DBSCAN algorithm.

- Fig.3f ideally should have color legend. Additionally, why are some parts purple? If it is a combination of exons and introns, is it possible to make the plot clearer?

Re: We thank the reviewer for the suggestion. A color legend has now been added to Figure 3f to improve clarity (**Figure R18** below).
Updated figure: **Figure 3f** (corresponding to **Figure R18** below)

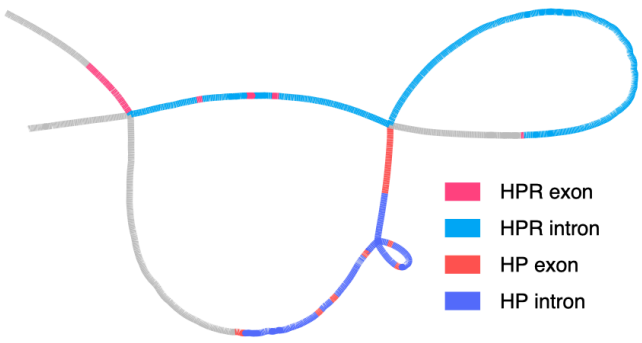

Figure R18. The subgraph of the *HP* gene cluster within the 1KCP pangenome. Colored segments indicate regions traversed by *HP* and *HPR*.

- Fig.3g: blue haplotypes look identical. If there are any differences, it's impossible to notice with human eyes.

Re: The issue may have been due to the visualization of the Bandage plot. We have now redrawn the plot, which resolves this issue (**Figure R19** below).

Updated figure: **Figure 3g** (corresponding to **Figure R19** below)

Figure R19. Structural haplotype paths of the *HP* gene cluster. Colored segments indicate regions traversed by each haplotype.

- Line 295: show the number of affected haplotypes (86 and 1?). Possibly, select shorter names/abbreviations for the haplotypes.

Re: Following the reviewer's suggestion, we have added the counts of affected haplotypes in the revised manuscript. Regarding the haplotype names, we decided to retain the original nomenclature to clearly indicate the status of the *HP* gene cluster.

Updated text (lines 321-326): "A 1.7 kb deletion (AF = 0.2766, allele count = 617) spanning exons 5 and 6 of *HP* distinguished the HP1-HPR haplotype from the reference haplotype HP2-HPR and is associated with low-density lipoprotein (LDL) and total cholesterol levels. Two additional deletions were identified: a 27 kb deletion (AF = 0.0385; allele count = 86) generating the HPdel-HPRdel haplotype, and an 11 kb deletion (AF = 0.0004; allele count = 1) generating the HP2-HPRdel haplotype."

- Line 395: how were HLA alleles imputed?

Re: HLA alleles were imputed using the same approach as other variants and represented in biallelic form in the VCF file. Variant representations of HLA alleles were derived from typing results, with the variant positions defined at the center of the corresponding HLA genes. We have revised the Methods section to describe this process in detail.

Updated text (lines 950-958): "We constructed the 1KCP imputation reference panel based on the GRCh38 linear reference, incorporating a comprehensive range of variant types in biallelic form, including SNVs, indels, SVs, nested variants, TR variants, and HLA alleles. For HLA alleles, the typing results were converted into variant representations, with the variant positions defined at the center of the corresponding HLA genes. Variants with a missing rate > 10%, HWE p-value < 1e-6, or allele count < 2 were excluded from the reference panel.

To evaluate imputation performance, we performed leave-one-out imputation using Minimac4 v4.1.6, where each of the 55 high-coverage samples was sequentially excluded as the imputation target sample."

Referee #1 (Remarks on code availability):

Thank you for providing the code under MIT license. It is essentially a collection of bash scripts (and some Python and R). Its main purpose appears to be as documentation what was done, not so much as a workflow to be run by others. I have not attempted to run it, which appears somewhat tedious as a user would need to infer all the assumptions about command-line parameters and sometimes hard-coded file names by reading the scripts. It would be very helpful if the authors could at a minimum provide the exact environments with all the dependencies (e.g. in a container or as a Conda environment), a test data set, and exact instructions how to to run that test data set.

Re: Following the reviewer's suggestion, we have reorganized the analysis code into a cleaner structure and converted the bash scripts into a Snakemake pipeline. We now provide the required Conda environments along with instructions for running the pipeline. We have also provided detailed descriptions of the required input data formats in the README and JSON configuration files.

Referee #2 (Remarks to the Author):

I co-reviewed this manuscript with one of the reviewers who provided the listed reports.

Re: We thank both Reviewers #1 and #2 for their constructive comments.

Referee #3 (Remarks to the Author):

Wang et al developed PIGA, a reference-guided assembly pipeline. Using PIGA, they produced the reference-guided assembly of 1116 Chinese individuals with high-coverage short reads and low-coverage long reads. They analyzed the assemblies with pangenome approaches and found many variants in the cohort. This is a huge effort and PIGA is methodologically interesting. However, a key question is whether the assembly quality is high enough for typical pangenome analysis as it is known that reference-guided assembly has limitations. Critical observations are as follows:

Re: We thank the reviewer for the constructive comments and for recognizing both the scale of this work and our PIGA workflow.

To mitigate the limitations of reference-guided assembly, our PIGA workflow incorporates several key improvements. Specifically, we first leverage an external small-scale pangenome to construct a personalized reference for each sample, which helps reduce reference bias during draft diploid genome assembly. In the second stage, we integrate both the external assemblies and PIGA draft assemblies to build an intermediate, large-scale pangenome, which further enhances the ability to resolve sequences of individuals, particularly in divergent genomic regions. As suggested by Reviewer #1, we now refer to these assemblies as pangenome-informed assemblies to more accurately reflect the methodological strategy employed.

Below, we address the reviewer's specific concerns in detail and provide supporting analyses and clarifications.

1) The manuscript does not describe the PIGA pipeline in detail. The first several steps up to draft assembly are akin to hapdup (Kolmogorov et al, 2023). However, started with a de novo assembly with long reads at high coverage, hapdup can produce a better backbone and hapdup can't compete with mainstream assemblers. In my understanding, the authors construct a pangenome graph with all ~1000 draft assemblies and use the graph to improve the assembly (is it correct?). If a rare SV is missed in all draft assemblies, it would not be reconstructed in the final assembly.

Re: We thank the reviewer for this comment. Following the reviewer's suggestions, we have now added a detailed step-by-step description of the PIGA workflow in the **Supplementary Note** and **Supplementary Figures 3-10**. The key components of the PIGA pipeline are summarized as follows:

1. **Comprehensive SNV detection and phasing**: We leveraged hybrid sequencing data to create a comprehensive SNV callset, which improved the ability to detect SNVs in repetitive regions compared to using short-read data alone. These SNVs were then phased into haplotypes using long-read information, providing accurate phasing blocks that are essential for diploid genome assembly.
2. **Personalized reference genome construction**: We generated a personalized reference for each individual by modifying the reference genome with homozygous variants genotyped from an external graph-based pangenome. For our modest-coverage dataset, this personalized reference improved read alignment and enabled more long reads to be partitioned into haplotypes, compared to using a standard reference or a *de novo* squashed assembly.
3. **Reference-guided draft diploid genome assembly**: Using the personalized reference as a guide, we generated draft diploid assemblies for each individual. This approach demonstrated a significant improvement in assembly completeness compared to the *de novo* assembly using partitioned long reads. Multiple rounds of polishing with both long and short reads were then conducted, which continuously improved the quality of the draft assembly. Although these generated draft assemblies still contained incomplete regions, they provided a wealth of haplotype-resolved sequences across most of the genome and serve as a crucial foundation for generating final diploid assemblies.
4. **Pangenome construction and simplification**: A population-scale pangenome was constructed using the reference genomes, high-coverage Hifiasm assemblies, and the PIGA draft assemblies. To mitigate pangenome errors introduced during assembly or pangenome alignment, we trained a multi-layer perceptron classifier to distinguish between true variants and erroneous nodes/edges, based on a rich set of pangenome-derived features. This machine learning-based filtering outperforms simpler methods based solely on allele count, showing strong performance even in distinguishing singleton variants and errors. The resulting filtered pangenome is significantly smaller, enabling efficient downstream analyses.
5. **Final diploid assembly generation**: We infer diploid paths for each individual through the constructed pangenome by integrating multiple layers of information, including short reads, long reads, draft assembly haplotypes, and SNV haplotypes. The final diploid assembly is then reconstructed from these paths, with additional refinement steps to correct SV-discordant regions and clip k-mer erroneous regions.

The reviewer is correct that the initial draft assembly step in PIGA is similar to hapdup. However, in our pipeline, we used a personalized reference as the backbone for generating draft diploid assemblies, which was better suited to the characteristics of our dataset. This strategy improved read alignment and enabled more long reads to be partitioned into haplotypes, compared to using a standard reference (GRCh38 and CHM13) or a *de novo* squashed assembly (wtdbg2 and Flye) (**Figure R20** below).

This clarification has been added to the revised manuscript.

Updated text (Supplementary Note, lines 68-76**)**: "Previous methods for generating diploid genome assemblies from low-accuracy long reads often rely on either a standard reference genome or a *de novo* squashed assembly to guide read alignment and haplotype partitioning. However, standard reference genomes poorly represent individual-specific structural variation, especially in highly divergent regions, while squashed assemblies are often collapsed in repetitive regions. To address these limitations, we developed a strategy to construct a personalized reference for each individual by modifying the reference genome with homozygous variants genotyped from an external graph-based pangenome. This personalized reference better represents individual genomic structures than reference genomes and offers superior contiguity and completeness compared to squashed assemblies."

Updated text (Supplementary Note, lines 85-98**)**: "To objectively assess the quality of the personalized CHM13 reference, we compared its read alignment performance against four other reference types (original CHM13, original GRCh38, Flye squashed assembly, and wtdbg2 squashed assembly) using 10 benchmark samples. Across all read alignment metrics, the personalized CHM13 consistently outperformed all alternatives, while the squashed assemblies showed the weakest performance (Supplementary Figure 6a). These results demonstrate that the personalized CHM13 serves as a more effective reference for our modest-coverage dataset.

Next, we leveraged this personalized reference to enhance read partitioning. We lifted over phased SNV haplotypes from the original CHM13 to the personalized CHM13 using GATK LiftoverVcf. In regions where the lift-over failed, typically corresponding to structural divergence from the reference, we recalled heterozygous SNVs using DeepVariant, and phased them using WhatsHap to complete the haplotypes. Incorporating SNV haplotypes from the personalized reference enabled, on average, 1.6% more long reads to be assigned to haplotypes compared to using the original CHM13, ultimately allowing for the successful partitioning of approximately 80% of long reads per individual (Supplementary Figure 6b)."

We acknowledge that performing graph phasing-based diploid genome assembly with noisy non-HiFi long reads (accuracy ~ 90%) remains challenging. Recent advances, such as the ONT mode of Hifiasm (Cheng et al. 2025 bioRxiv), effectively address this issue and substantially improve the assembly quality of ONT data. Although our current PIGA workflow was developed in a different context, the joint assembly paradigm could be further adapted to incorporate Hifiasm to enhance assembly performance in low- and modest-coverage settings. We have noted this methodological advancement and the potential future improvement of PIGA in the Discussion section.

Updated text (lines 545-547): “Moreover, leveraging more accurate long reads and incorporating more advanced genome assembly (Cheng et al. 2025 bioRxiv) and pangenome construction methods into the PIGA workflow may help address this challenge.”

Regarding the reviewer’s question on how the large-scale draft assemblies are used, we indeed construct an intermediate pangenome graph from all draft assemblies and other external assemblies, which serves as the foundation for generating the final diploid assemblies. This strategy substantially improves assembly quality, as evidenced by the increase in QV from approximately 35 in the draft assemblies (**Figure R21** below) to 46 in the final assemblies.

Regarding the rare variant, we agree with the reviewer that rare variants absent from all draft assemblies cannot be reconstructed in the final assemblies of a given individual. Nevertheless, at the population level, incorporating PIGA assemblies enabled the detection of 4.3-fold more rare SVs (AF < 0.01) compared to using only high-coverage assemblies (**Figure R4** below). This suggests that while certain rare variants may be missed at the individual level, the PIGA framework substantially enhances the detection of rare variation across the cohort. We have added the relevant results and included a discussion of this limitation of the current PIGA workflow in the manuscript.

Updated figures: **Supplementary Figures 6** and **7b** (corresponding to **Figures R20** and **R21** below)

Updated text (lines 536-539): “Third, although the PIGA workflow fully utilizes population-scale hybrid sequencing data to identify a substantial number of genetic variants at the population level, the modest-coverage design employed may lead to some rare variants being missed at the individual level.”

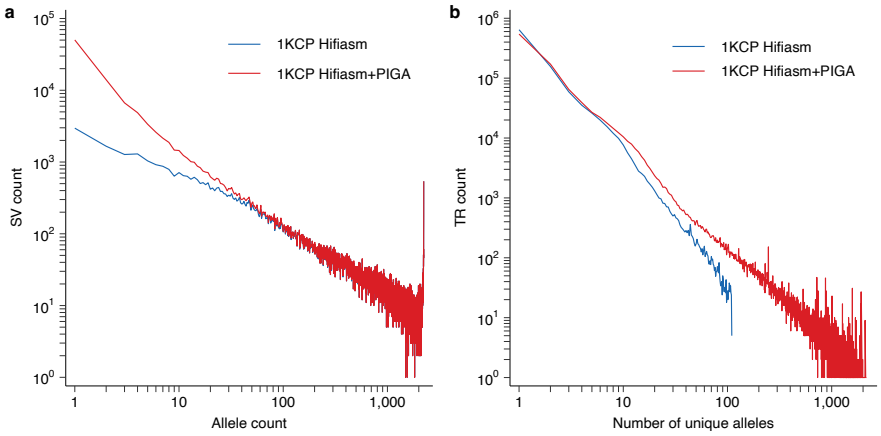

Figure R4. Comparison of variant discovery between the 1KCP full assembly panel (N = 1,116) and the Hifiasm-only assembly panel (N = 55). a, Number of SVs discovered in the two assembly panels, stratified by allele count in the full assembly panel. **b**, Number of TR sites stratified by the number of

unique alleles per site in the two assembly panels. (This figure is also included in our response to Reviewer #1.)

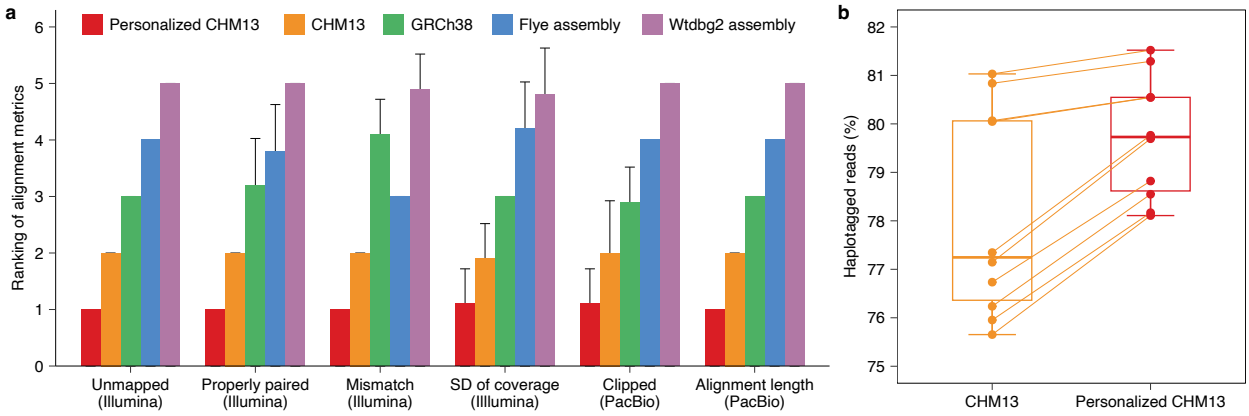

Figure R20. Evaluation of the PIGA personalized reference for read alignment and haplotype partitioning in ten benchmarking samples. a, Comparison of alignment metrics across five reference types. Metrics include the proportion of unmapped Illumina (short) reads, proportion of properly paired Illumina reads, mismatch rate of Illumina reads, standard deviation of Illumina read coverage, proportion of clipped PacBio (long) reads, and average alignment length of PacBio reads. Error bars indicate 95% confidence intervals. **b**, Proportion of haplotagged PacBio reads using the original CHM13 versus the personalized CHM13 reference. Each connected pair of points represents results from the same sample.

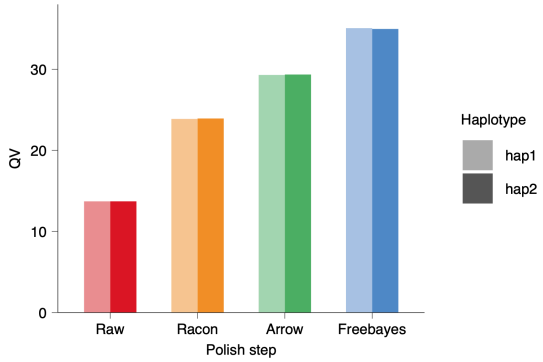

Figure R21. Evaluation of the PIGA reference-guided draft diploid assembly pipeline. Quality values (QV) of draft assemblies across polishing steps, using the individual 1KCP-00309 as an example.

Reference

1. Cheng, H. et al. Efficient near telomere-to-telomere assembly of Nanopore Simplex reads. ***bioRxiv***, 2025.2004. 2014.648685 (2025).

2) Fig 2c and Fig S22 suggest the PIGA assemblies are burdened with misassemblies, confirming 1). Also, Inspector and Flagger might underestimate misassemblies given low-coverage long reads – the actual error rate might be higher. We cannot use these assemblies in segmental duplications or satellites where de novo assembly shows real power over alignment-based methods.

Re: We would like to clarify that Flagger was already applied using high-coverage HiFi data. To more accurately assess structural errors, we re-analyzed the PIGA assemblies of three benchmark samples using Inspector with high-coverage HiFi data, and observed an average of 617 structural errors per sample.

Due to the lack of high-coverage data to assess structural errors across population-scale datasets, we removed original Figure 2c from the manuscript and replaced it with a new analysis (**Figure R10** below). This new figure shows that 39% of k-mer errors and 48% of structural errors occur within Flagger unreliable regions, which are predominantly located in highly repetitive regions, including segmental duplications and satellites. These regions are challenging for the current PIGA workflow because noisy non-HiFi long reads have limited power to resolve highly repetitive sequences, and such regions are often clipped during pangenome construction due to their extremely complex graph structures. Therefore, the PIGA workflow may aggressively borrow sequences of these regions from complete reference haplotypes (e.g., CHM13) to reconstruct the final assembly, even when they do not match the individual's actual haplotype. To mitigate this limitation, we derived common unreliable regions in the GRCh38 reference that were shared across benchmark samples. Then we implemented a new k-mer-based pangenome QC and region-level QC steps to improve the reliability of our dataset (as described at the beginning of this document). As a result, the total number of SV alleles decreased from 376,495 to 164,216, and the FDR decreased from 4.3% to 2.6%. The estimated false discovery rate (FDR) for all SVs dropped from 4.5% to 2.6%, and notably, the FDR for singleton SVs decreased from 77.3% to 26% (**Figure R11** below).

We acknowledge that segmental duplications and satellites represent the most difficult regions for genome assembly and remain a current limitation of the PIGA workflow. Nevertheless, PIGA assemblies can still reconstruct diploid sequences with high accuracy across the majority of the genome. This enables more accurate characterization of TRs and SVs, detection of nested variants within non-reference sequences, and high-resolution HLA typing at a population scale, as demonstrated in our study. These advances allow the construction of an extensive variant catalog and imputation reference panel tailored to Chinese and East Asian populations, supporting future research in population and clinical genetics.

We have incorporated these results into the revised manuscript and described the limitations of the PIGA assembly in these highly repetitive regions in the Discussion section.

Updated text (lines 114-122): “Structural accuracy was evaluated using Inspector on three benchmark samples, revealing an average of 31 structural errors per Hifiasm assembly and 616 per PIGA assembly. We further assessed regional reliability using Flagger, finding that 39% of k-mer errors and 48% of structural errors in PIGA assemblies occurred within unreliable regions, which were mostly located in satellites and segmental duplications (Figure 1c and Supplementary Figure 12). Based on these results, we used Flagger annotations from the benchmark samples to define 53.7 Mb of common unreliable regions in the GRCh38 reference, which were used in downstream quality control steps to improve the reliability of the 1KCP dataset (Methods).”

Updated text (lines 685-692): “Structural errors were detected using high-coverage HiFi reads with Inspector v1.0.1. Additionally, unreliable assembly regions in both PIGA and Hifiasm assemblies of three benchmark samples were identified using high-coverage HiFi reads with Flagger v0.2.0. The unreliable regions (including collapsed, duplicated, and error regions) in the PIGA assemblies were lifted over to both GRCh38 and CHM13 references, using minimap2 (-cxasm20) for alignment and rustybam (https://github.com/mrvollger/rustybam) for coordinate transformation. Reference regions marked as unreliable in more than two haplotypes were defined as common unreliable regions.”

Updated text (lines 539-551): “Finally, highly repetitive regions, such as segmental duplications and satellites, remain challenging to resolve with the current PIGA workflow. The reliance on noisy non-HiFi reads limits resolution in these regions, and such sequences are often clipped from the PIGA intermediate pangenome due to their extremely complex graph structure. Therefore, the current PIGA workflow may aggressively borrow sequences of these regions from complete reference haplotypes to reconstruct the final assembly. To mitigate potential artifacts, we identified common unreliable regions using Flagger annotations and excluded the affected genes and variants from downstream analyses. Moreover, leveraging more accurate long reads and incorporating more advanced genome assembly and pangenome construction methods into the PIGA workflow may help address this challenge. In addition, Telomere-to-telomere (T2T) assemblies have demonstrated the ability to fully resolve these regions. Continued advancements in sequencing technology and scalable T2T assembly algorithms hold great promise for resolving complete genomes across human populations and broadly capturing the genetic diversity within the most intricate regions of the human genome.”

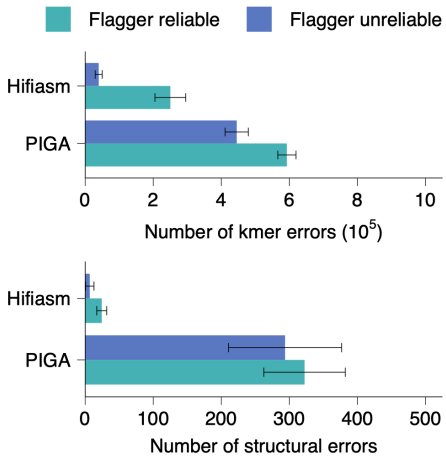

Figure R10. The number of k-mer errors and structural errors in Hifiasm and PIGA assemblies for the three benchmark samples, stratified by regional reliability as estimated by Flagger. Reliable regions correspond to haploid regions, while unreliable regions include collapsed, duplicated, and error-prone regions. Error bars represent 95% confidence intervals. (This figure is also included in our response to Reviewer #1.)

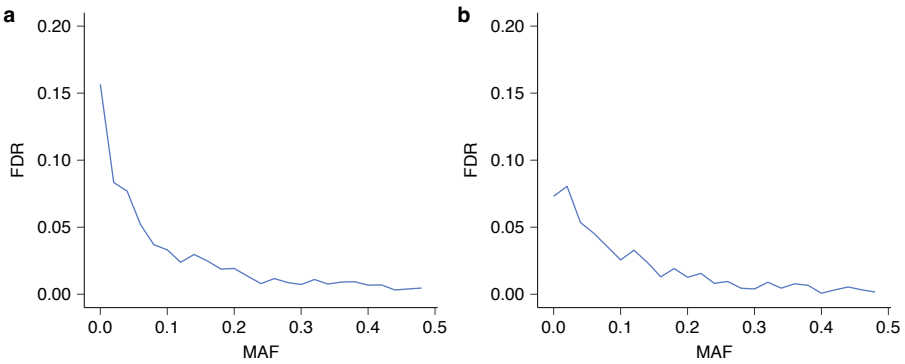

Figure R11. Evaluation of false discovery rate (FDR) of the 1KCP SV callsets across MAF bins, estimated using three benchmark samples. a, Pre-QC SV callset. **b**, Post-QC SV callset. (This figure is also included in our response to Reviewer #1.)

3) Liftoff has not been shown to have enough power to annotate HLA genes. In Fig S19c/d, HLA-DRB3/4/5 are shown to have high LD. This is incorrect as HLA-DRB3/4/5 can't co-exist on one haplotype.

Re: We apologize for the confusion in the original manuscript regarding the software used and the interpretation of the LD results. In this study, HLA typing was performed using HiFiHLA, and we have explicitly stated this in the revised Results section.
Updated text (line 341): "Using HiFiHLA, we identified HLA alleles for 36 HLA genes."

Regarding the reviewer's point about LD, we note that LD measures the non-random association of alleles between two genes. Thus, the absence of one gene (e.g., HLA-DRB3) being correlated with the presence of another gene (HLA-DRB4/HLA-DRB5) can still result in high ε values. We re-examined the typing results for HLA-DRB3/4/5 and found that 98.3% (1,519/1,546) of haplotypes contained only one of these genes, while 1.7% (27/1,546) contained two, which we attribute to potential switch errors (**Figure R22** below). We have also clarified the interpretation of LD in the revised Results section.
Updated text (lines 348-350): "Given the extensive allele diversity and complex haplotype structure of HLA genes, we used an entropy-based LD measurement index (ε) to evaluate the multiallelic LD between HLA genes, which measures the non-random association of multiple alleles across two loci."

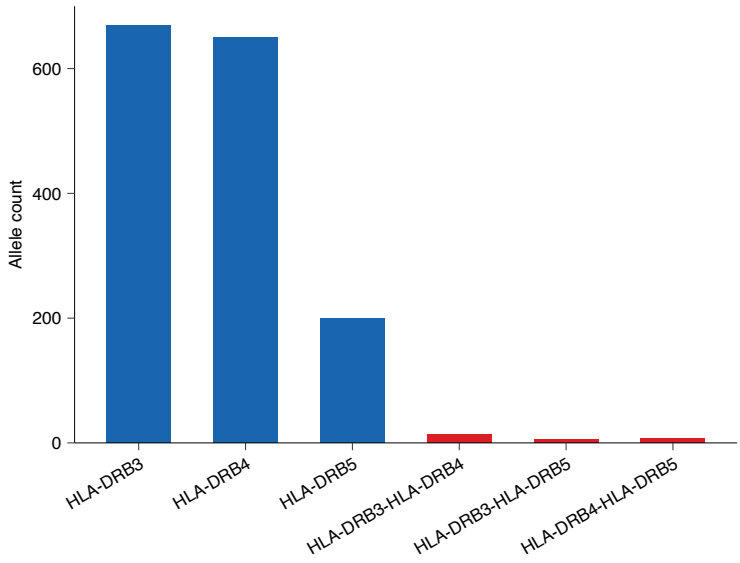

Figure R22. HLA typing results for HLA-DRB3, HLA-DRB4, and HLA-DRB5. Most haplotypes (98.3%, blue bar) contain only one of these genes, while a small proportion (1.7%, red bar) contains two, which likely reflects switch errors.

4) A few smaller issues. (a) the assemblies with low-coverage long reads do not have chromosome-scale phasing. (b) an error rate of 1/40000 in one assembly could mean 1/20 error rate across 2000 haplotypes.

This will overestimate the number of variants. (c) in the mean time, QV does not tell us how many heterozygous SNPs are missing. This will underestimate the number of variants.

Re: We agree with the reviewer that the current PIGA assemblies provide only local phasing rather than chromosome-scale phasing. However, the analyses presented in this study, including gene cluster haplotype, HLA allele, and imputation, primarily rely on local phasing. We have clarified this limitation in the revised manuscript.
Updated text (lines 534-536): "Second, the PIGA assemblies achieve accurate local phasing but lack chromosome-level phasing, which could be improved by integrating additional phasing data, such as trio-based or Hi-C sequencing."

We acknowledge that our QV estimation of PIGA assemblies was based on modest-coverage short-read data, which could treat some missed real k-mers as false positives, thus slightly overestimating the error rate. To assess the impact of coverage, we compared QV estimates using both modest- and high-coverage short-read data in three benchmark samples. We observed that the average QV of PIGA assemblies increased from 46.15 (modest coverage) to 47.89 (high coverage), suggesting that the true error rate may be closer to 1 in 60,000 rather than 1 in 40,000 (**Figure R23** below). The QV estimation results are now revised in the manuscript.
Updated Figure: **Supplementary Figure 11** (Corresponding to **Figure R23** below)
Updated text (lines 111-114): "Although QV values were slightly underestimated based on the modest-coverage short-read dataset (Supplementary Figure 11 and Supplementary Table 4), the PIGA assemblies were estimated to have an average QV of 46, which is lower than the Hifiasm assemblies (average QV: 54), but still indicative of a low error rate."

Additionally, we implemented a new k-mer-based pangenome QC and region-level QC steps to reduce potential errors (as described at the beginning of this document). These efforts reduced the number of small variant alleles from 54.9M to 45.7M, and SV alleles from 376,495 to 164,216. The total variant-affected length (sum of variant lengths) decreased from 318.6 Mb to 228.2 Mb, which is close to the estimated length of error-affected regions (~100 Mb).

We also agree that QV does not reflect missing heterozygous variants. To evaluate this, we checked the benchmarking result of small variant detection in three benchmark samples. We found that, on average, 2.08% of heterozygous SNVs and 15.22% of heterozygous indels were missed, compared to lower

missing rates for homozygous variants (0.05% for SNVs and 10.7% for indels). (**Figure R24** below). These results have been incorporated into the revised manuscript.

Updated figure: **Supplementary Figure 15** (corresponding to **Figures R24** below)

Updated text (lines 132-133): "We also observed lower recall for heterozygous variants compared with homozygous variants (Supplementary Figure 15)."

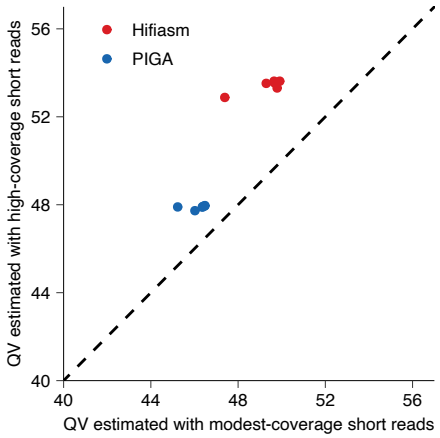

Figure R23. QV estimates for PIGA and Hifiasm assemblies of three benchmark samples, based on both modest-coverage and high-coverage short-read datasets.

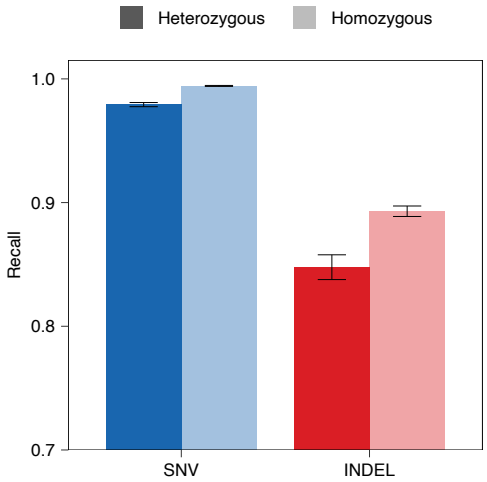

Figure R24. Recall of small variant detection using PIGA assemblies for three benchmark samples, stratified by heterozygous and homozygous variants. Error bars indicate 95% confidence intervals.

5) Overall, PIGA assemblies probably lag far behind real de novo assemblies produced by several other consortia in quality. The analysis might not be much better than read-to-reference alignment.

Re: We fully acknowledge recent efforts of other consortia (e.g., T2T, HGSVC, and HPRC Consortium) that use multiple sequencing technologies with high coverage to generate complete or nearly complete genome assemblies with exceptionally high quality. These efforts represent the current gold standard in genome assembly.

However, our study addresses a different but critical challenge: how to generate diploid assemblies and build an extensive variant catalog for a large-scale cohort in a cost-effective strategy. In such contexts, high-coverage long-read data are often not available, and assemblies produced from low- or modest-coverage datasets using standard *de novo* assembly methods tend to be incomplete and error-prone, as also observed in our draft assemblies. This limitation motivated us to develop the PIGA workflow, which fully leverages the population-scale hybrid sequencing dataset and performs joint genome assembly to improve the quality of diploid assemblies across a large-scale cohort.

Despite an initial attempt at this type of joint diploid assembly analysis, we have demonstrated that PIGA assemblies substantially improve over initial draft assemblies, achieving a significant increase in both accuracy and completeness. Additionally, PIGA assemblies significantly outperform read-based variant calling in detecting various variant types, including SNVs, indels, SVs, and TRs (**Figure R5a-c** and **Figure R25b** below). In particular, PIGA assemblies provide more accurate sequences for SVs and TRs compared to long-read-based approaches (**Figure R5d** and **Figure R25a** below). These demonstrate the value of our assembly-based approach for comprehensive variant discovery.

These results have been incorporated into the revised manuscript.
Updated figure: **Supplementary Figure 27** (corresponding to **Figure R25** below)
Updated text (lines 133-134): "Nonetheless, PIGA assemblies outperformed read-based variant calling across multiple variant classes (Supplementary Figure 10)."
Updated text (lines 248-251): "Benefiting from the high accuracy of assemblies, 98.2% of TR loci showed high concordance (≤ 1 edit distance in repeat unit representations of TR alleles) between Hifiasm and PIGA assemblies, whereas only 75.3% showed high concordance with Hifiasm assemblies when directly called from long reads (Methods and Supplementary Figure 27)."
Updated text (lines 823-829): "For evaluating the TR callset, we additionally called TRs from short-read data using GangSTR v2.5 and from long-read data using vamos. The PIGA and long-read TR callsets from three benchmark samples were compared against the Hifiasm TR callset by using edlib to calculate edit distances between repeat unit representations of paired TR alleles at corresponding TR sites. Alleles from different callsets were paired to minimize the total edit distance across all possible allele pairs.

Furthermore, TR alleles from these callsets were normalized using vcfwave v1.0.14 and benchmarked using Truvari bench (-s 5) and refine (-u) against the Hifiasm Dipcall callset, restricted to GIAB-defined TR regions."

Finally, we believe our resource occupies a unique and valuable niche in the landscape of human genetics. While T2T assemblies provide a perfect representation for an individual genome, and short-read/array provide genotypes for millions of individuals, our work provides a powerful intermediate resource to bridge the gap between them. In particular, our large-scale assemblies provide an extensive variant catalog, especially for including SVs, TRs, nested variants, and HLA alleles, and serve as a valuable resource and imputation reference panel to support broader population and clinical genetics research (**Figure R13** below).

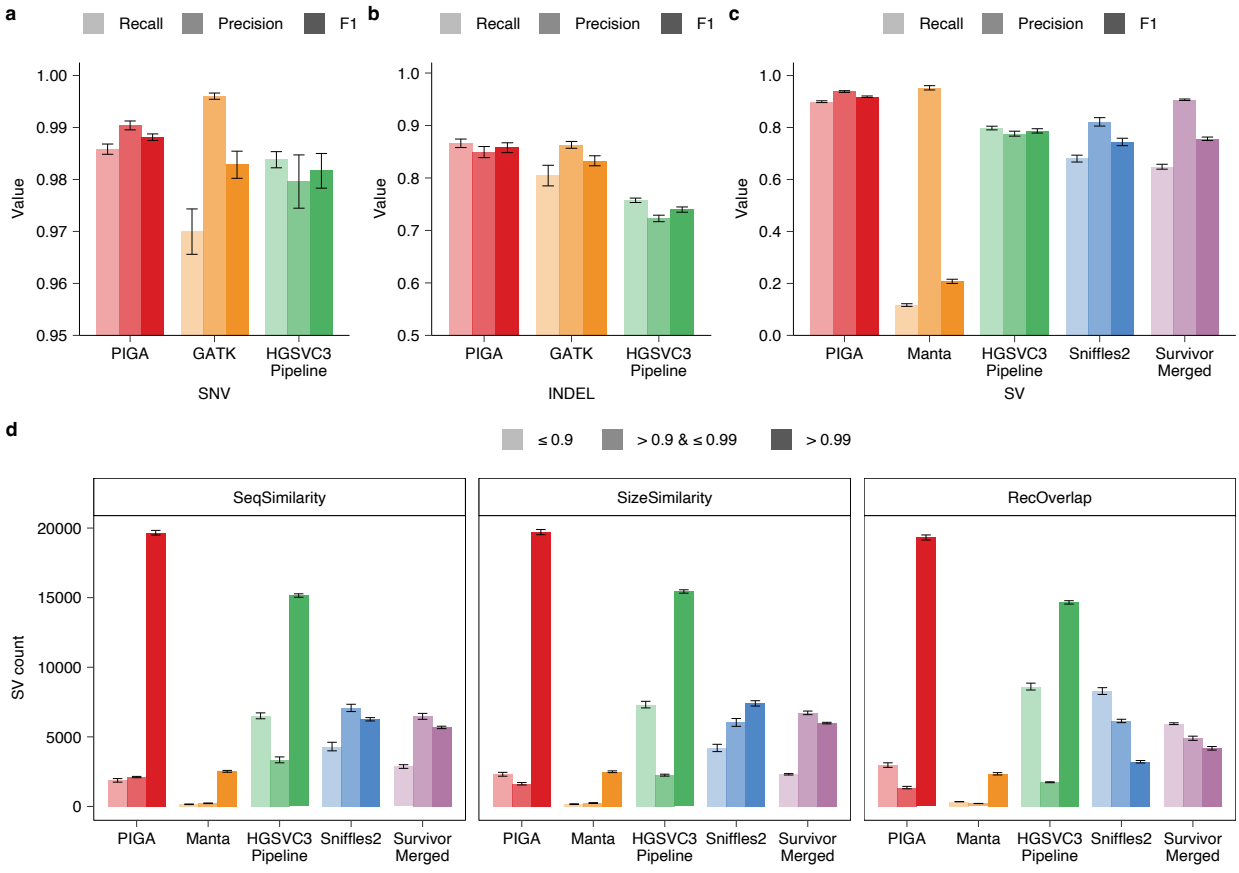

Figure R5. Benchmarking of variant detection performance across different methods in three benchmark samples. a, SNV detection performance evaluated using hap.py. Error bars indicate 95% confidence intervals. **b**, Indel detection performance evaluated using hap.py. Error bars indicate 95% confidence intervals. c, SV detection performance evaluated using Truvari bench (--pick multi --sizemin

50). Error bars indicate 95% confidence intervals. d, Number of SVs stratified by sequence accuracy levels. Accuracy is evaluated using three metrics: sequence similarity (left), size similarity (middle), and reciprocal overlap percentages (right). Error bars indicate 95% confidence intervals. (This figure is also included in our response to Reviewer #1.)

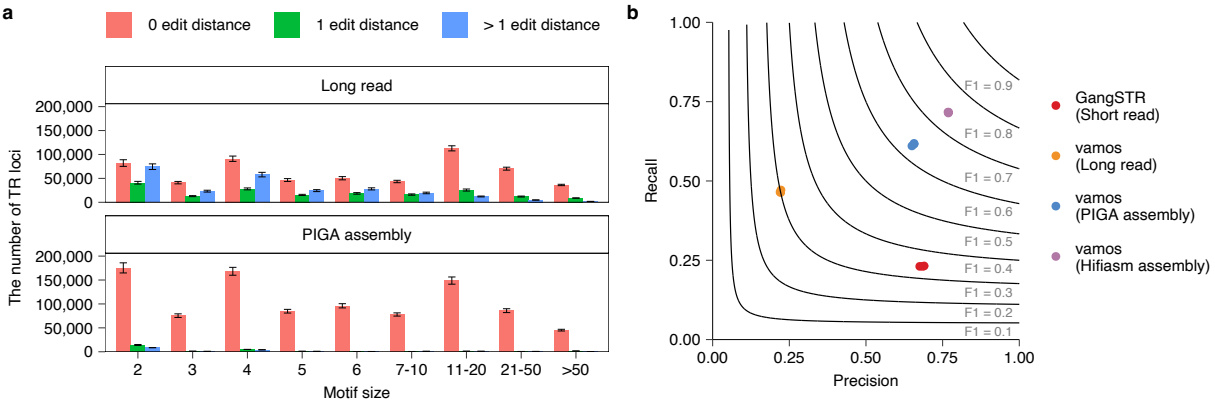

Figure R25. **Benchmark of TR calling performance across different methods in three benchmark samples.** a, Number of TR loci characterized by vamos, stratified by motif size and edit distance. Edit distance refers to the differences in motif unit representations of TR alleles between the query callsets (Long-read and PIGA assembly) and the truth callset (Hifiasm assembly). Error bars indicate the 95% confidence intervals. **b**, TR benchmarking performance using Truvari bench (-s 5) and refine (-u), restricted to GIAB-defined TR regions. The black curves represent constant F1 scores from 0.1 to 0.9.

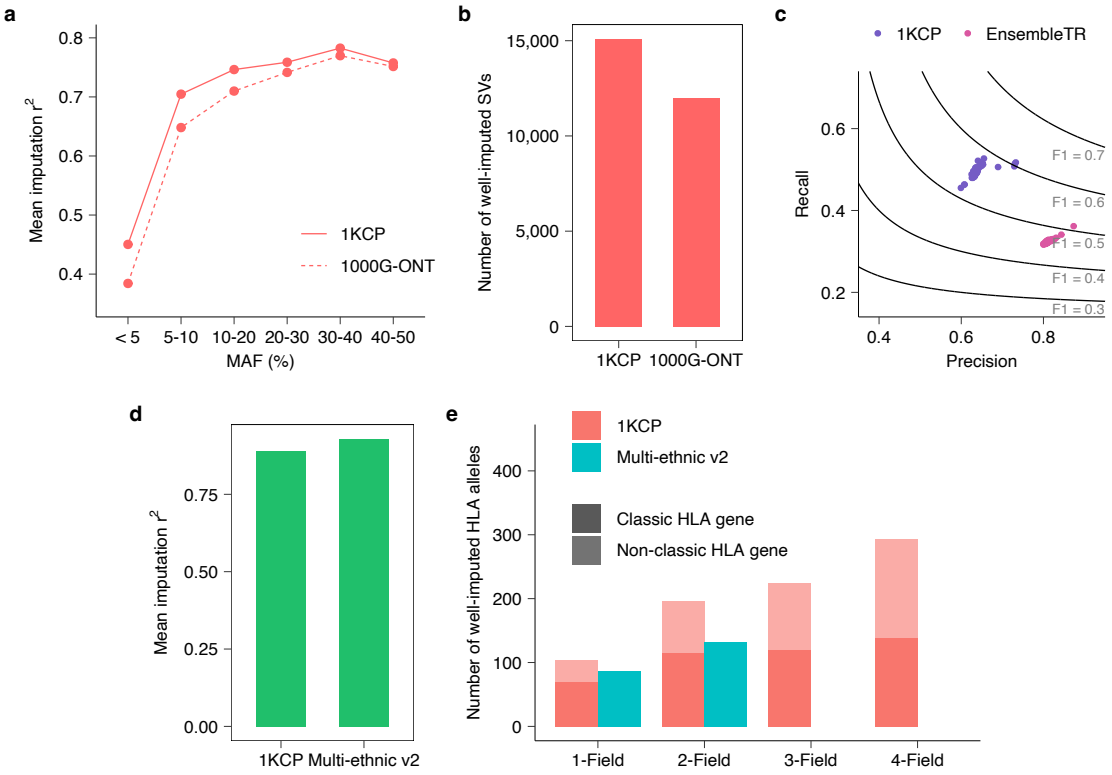

Figure R13. Comparison of imputation performance between the 1KCP reference panel and existing panels. a, Imputation performance (r^2) of the 1KCP and 1000G-ONT panels for shared SVs stratified by MAF bins. **b**, Numbers of well-imputed SVs ($r^2 > 0.8$) across the 1KCP and 1000G-ONT SV panels. **c**, Benchmarking of imputed TR alleles from the 1KCP and EnsembleTR panels, evaluated with Truvari bench (-s 5) and refine (-u), restricted to GIAB-defined TR regions. The black curves represent constant F1 scores from 0.3 to 0.7. **d**, Imputation performance (r^2) of the 1KCP and Four-digit multi-ethnic HLA v2 panels for shared HLA alleles. **i**, Numbers of well-imputed HLA alleles ($r^2 > 0.8$) from the 1KCP and Four-digit multi-ethnic HLA panels stratified by allele resolution levels and gene classes. (This figure is also included in our response to Reviewer #1.)

Referee #3 (Remarks on code availability):

I briefly looked the documentation. It is a long and sophisticated pipeline calling many tools but might not be easy for external users to use.

Re: We thank the reviewer for the comment. We agree that the PIGA workflow is complex in its current form, as it was initially designed for the specific hybrid sequencing data in the 1KCP dataset, which required tailored optimizations. To address this, we are now developing a simplified and more user-

friendly version that will be compatible with ONT or PacBio HiFi-only data. This new version will incorporate Hifiasm to directly generate *de novo* draft assemblies, thereby reducing the complexity of the original draft assembly process and further mitigating the reference bias issue. We are also exploring improved graph construction and simplification strategies to enhance both usability and performance of the PIGA workflow, which we plan to report in a separate publication.

Furthermore, we have made significant improvements to the analysis scripts for the 1KCP project. We have converted the original bash scripts into a more modular and reproducible Snakemake pipeline. We now provide the required Conda environments along with instructions for running the pipeline. We have also provided detailed descriptions of the required input data formats in the README and JSON configuration files. We believe these changes make the analysis pipeline more accessible to external users.

Referee #4 (Remarks to the Author):

In the submitted manuscript, Wang et al. use pangenomic approaches to analyze an immense newly generated dataset of genome sequences from a Chinese population. Specifically, the data comprised high-coverage PacBio HiFi long-read, Illumina short-read, and Hi-C data from 55 samples combined with modest coverage PacBio long-read and Illumina short-read data from an additional 1,099 samples (with 10 samples overlapping between the strategies). This allowed them to achieve a nearly complete view of genetic variation within their sample, including technically challenging variation such as tandem repeats (TRs), other classes of structural variation (SVs), as well as additional variation nested within SV alleles (e.g., SNPs encompassed by an SV). By combining with RNA-seq data they generated from the same sample, they also used their genotype data to perform eQTL mapping, revealing gene expression associations not only with SNPs but also other classes of variants. Finally, they built an imputation server that allows users to upload and impute missing genotype data based on their new reference panel.

This work represents a technical milestone, as I am unaware of any other project that has produced this number of human genome assemblies based on this scale and quality of data, nor has any other project applied human pangenome approaches on this scale. The large sample size allowed the authors to better capture rare variation that is abundant in human genomes. The ability to resolve nested variation, TRs, and HLA haplotypes were relatively novel aspects of the work, as was the analysis of gene expression cis-heritability, partitioned by variant class. Overall, however, even these sections lacked motivating context from the literature, so it was difficult to determine whether any truly novel biological insights were

obtained. As such, the primary value of the study is as a genomic resource, but this value is somewhat undercut by the lack of availability of the raw sequencing data or individual-level genotype calls. The construction of an imputation server partially addresses this latter issue for one common application - though see below for a critique about how the advantages of this panel remain unclear. In addition, data privacy concerns preclude many researchers from uploading data to an external server.
The writing was relatively clear, though I have noted a few exceptions.
Please see my specific critiques and questions that I hope the authors will address:

Re: We thank the reviewer for the thoughtful summary of our work, the positive remarks, and the constructive comments.

To address the reviewer's concerns about the lack of motivating context, we have revised the manuscript and added additional references to more clearly place our analyses within the existing literature and to highlight the novel aspects of our study. These include:

1. Functional annotation of non-reference sequences

Updated text (lines 193-195): "While previous studies have also identified non-reference sequences from pangenomes (Liao et al. 2023 Nature; Gao et al. 2023 Nature), their functional potential remains poorly explored. To address this, we developed a path-guided pangenome annotation method to identify genomic elements within the 1KCP pangenome."

2. Detection of nested variants

Updated text (lines 213-216): "Of these nested variants, 88.2% were embedded within reference SV sites, highlighting the substantial genetic diversity encoded as nested variants within SVs, which is typically overlooked in previous coarse-grained SV analyses (Collins et al. 2020 Nature; Schloissnig et al. 2025 Nature; Logsdon et al. 2025 Nature)."

3. High-resolution HLA allele typing

Updated text (lines 338-340): "Compared to previous studies using short-read data (Luo et al. 2021 Nature Genetics; Hirata et al. 2019 Nature Genetics), which were limited to resolving HLA alleles at a maximum of three-field resolution, the 1KCP assemblies provide an extended phasing range across both coding and noncoding regions of HLA genes, enabling four-field resolution of HLA alleles."

4. Pan-variant eQTL analysis

Updated text (lines 364-367): "This broad dataset enables the incorporation of all major classes of genetic variation into a unified analytical framework, referred to as pan-variant analysis. Such analyses have the potential to improve resolution over previous association studies that focused on only one or a few variant types (GTEx Consortium 2020 Science; Chiang et al. 2017 Nature Genetics; Cui et al. 2025 Nature

Genetics), providing a more comprehensive view of how diverse variants contribute to phenotypic variation."

Regarding data availability, we are committed to sharing the 1KCP dataset with the global research community to the greatest extent permitted by national regulations. All individual-level data, including raw sequencing reads, genome assemblies, and variant genotypes, are available in the repositories of the National Genomics Data Center (NGDC), specifically in the Genome Sequence Archive for Human (GSA-Human: https://ngdc.cncb.ac.cn/gsa-human), Genome Warehouse (GWH: https://ngdc.cncb.ac.cn/gwh), and Genome Variation Map (GVM: https://ngdc.cncb.ac.cn/gvm). Data for the 10 overlapping benchmark samples have been made publicly available under BioProject accession number PRJCA039156 (https://ngdc.cncb.ac.cn/bioproject/browse/PRJCA039156). The full cohort data will be accessible to all researchers (BioProject accession number: PRJCA036466) through a standard NGDC data access application process upon the provisional acceptance of this paper for publication.

Regarding the imputation panel, we have added comparisons with existing panels and demonstrated the advantages of our assembly-based panel. To protect the data privacy of researchers, all uploaded data submitted to the imputation server is automatically deleted after one week.

These revisions, which we address point-by-point below, improve both the clarity and the impact of our study.

References

1. Liao, W.-W. et al. A draft human pangenome reference. ***Nature*** 617, 312-324 (2023).
2. Gao, Y. et al. A pangenome reference of 36 Chinese populations. ***Nature*** 619, 112-121 (2023).
3. Collins, R. L. et al. A structural variation reference for medical and population genetics. ***Nature*** 581, 444-451 (2020).
4. Schloissnig, S. et al. Structural variation in 1,019 diverse humans based on long-read sequencing. ***Nature***, 1-11 (2025).
5. Logsdon, G. A. et al. Complex genetic variation in nearly complete human genomes. ***Nature***, 1-12 (2025).
6. Luo, Y. et al. A high-resolution HLA reference panel capturing global population diversity enables multi-ancestry fine-mapping in HIV host response. ***Nature genetics*** 53, 1504-1516 (2021).
7. Hirata, J. et al. Genetic and phenotypic landscape of the major histocompatibilty complex region in the Japanese population. ***Nature genetics*** 51, 470-480 (2019).

8. GTEx Consortium. The GTEx Consortium atlas of genetic regulatory effects across human tissues. ***Science*** 369, 1318-1330 (2020).
9. Chiang, C. et al. The impact of structural variation on human gene expression. ***Nature genetics*** 49, 692-699 (2017).
10. Cui, Y. et al. Multi-omic quantitative trait loci link tandem repeat size variation to gene regulation in human brain. ***Nature genetics*** 57, 369-378 (2025).

Major comments:

1. The abstract (line 25) motivates the work by saying that previous smaller studies do not "fully capture" genetic diversity and that the current study "compiled a comprehensive variant catalog for the Chinese population" (line 232). Both statements should be moderated, as the goal to fully capture genetic variation is infeasible (any new sample will add some variation that is private to that individual or their close relatives) and the current study, while more comprehensive given the sample size and technologies involved, does not represent all of the genetic diversity in the broader population.

Re: We agree that phrases such as “fully capture” and “a comprehensive catalog for the Chinese population” may overstate the scope of our study, as no single study can capture the entirety of a population’s genetic diversity. To address this, we have revised the relevant statements to more accurately reflect the scope of our work.

Updated text (lines 25-26): “However, existing human pangenomes, constrained by small sample sizes, provide limited utility for medical and population genetics applications.”

Updated text (lines 83-85): “Both short-read and long-read whole-genome sequencing (WGS) were performed on a subset of 1,144 samples, aiming to establish a large-scale diploid assembly panel and an extensive catalog of genetic variation for medical and population genetics applications.”

Updated text (lines 258-259): “In summary, we have compiled an extensive variant catalog for the 1KCP cohort, encompassing multiple variant types and a broad allele frequency spectrum.”

We have also performed a careful review of the entire manuscript to ensure that other similar claims are appropriately moderated. We believe the text is now more precise and scientifically rigorous.

2. In addition to the point above, there should be some consideration of the ancestry composition of the sample population—for example whether it includes representation of the diverse ethnic groups that comprise the broader Chinese population. Such analysis/description should, however, be careful not to conflate the biological concept of genetic ancestry and socially or politically defined population labels

(see e.g., https://www.nationalacademies.org/our-work/use-of-race-ethnicity-and-ancestry-as-population-descriptors-in-genomics-research).

Re: We thank the reviewer for this helpful comment. We clarify that self-reported ancestry information was not explicitly collected from participants in the 1KCP cohort. We have now stated this limitation in the Methods section.
Updated text (lines 572-573): "Self-reported ancestry information was not explicitly collected and was therefore not available for this study."

Therefore, to characterize the ancestry composition of our samples, we relied on genetic analysis. Specifically, we updated our PCA results by incorporating additional samples from the 1000 Genomes Project and the Human Genome Diversity Project, thereby placing the 1KCP samples within a more comprehensive genetic context. The PCA results show that the majority of participants cluster with Han Chinese reference populations, while a small subset is positioned closer to other East Asian populations, suggesting underlying genetic heterogeneity within the cohort (**Figure R26** below). We have now described the ancestry composition of the 1KCP cohort in the Results section.
Updated figure: **Supplementary Figure 1** (corresponding to **Figure R26** below)
Updated text (lines 81-82): "The 1KCP project enrolled 1,379 participants, the majority of whom clustered with Han Chinese reference populations based on principal component analysis (PCA) (Methods and Supplementary Figure 1)."
Updated text (lines 646-656): "**Principal component analysis**
To place the 1KCP samples in a global genetic ancestry context, we performed a Principal Component Analysis (PCA) using SNV genotypes from the 1KCP samples combined with unrelated samples from the 1000 Genomes Project (1000G) and the Human Genome Diversity Project (HGDP). For the 1KCP samples, short-read data were aligned using BWA-MEM v0.7.17, and SNV genotypes were called using the GATK v4.2.6.1 workflow. Genotypes at overlapping SNV sites across the three datasets were merged to create two analytical sets: (i) a global set with all samples, and (ii) an East Asian-only set. For each set, we retained biallelic SNVs with minor allele frequency (MAF) > 0.05, Hardy-Weinberg Equilibrium (HWE) p-value > 1×10^{-6}, and genotype missingness < 10%. The remaining SNVs were further pruned to remove sites in LD using PLINK2 (--indep-pairwise 50 5 0.5). PCA was then conducted on this pruned dataset using PLINK2."

Based on this genetic characterization, we have moderated our language throughout the manuscript to use more precise terms, such as “the 1KCP cohort”, instead of making broad claims about “the Chinese population”. We believe these revisions more accurately reflect the scope of our dataset.

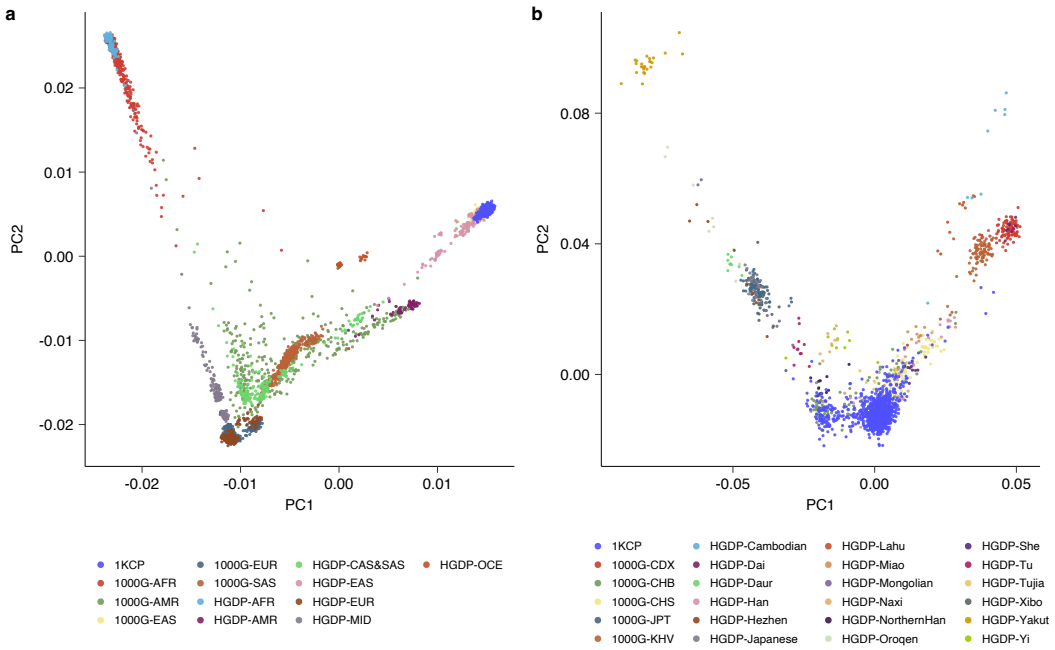

Figure R26. Principal Component Analysis (PCA) based on short-read WGS genotypes. a, PCA of the 1KCP samples alongside geographically diverse samples from the 1000 Genomes Project and Human Genome Diversity Project. **b**, PCA of the 1KCP samples alongside East Asian samples from the 1000 Genomes Project and Human Genome Diversity Project.

3. The Methods description of PIGA (the approach developed to integrate the modest coverage data into the pangenome) was greatly lacking in detail. All of the six steps should be described in detail in the methods, and decisions about the use of various software tools and parameters should be justified. In addition to these additions to the Methods text, Supplementary Figure 3 should be integrated with the workflow diagram that is currently posted in the PIGA GitHub repository, and the updated Supplementary Figure should be referenced on line 92.

Re: Following the reviewer’s suggestions, we have now added a detailed step-by-step description of the PIGA workflow in the **Supplementary Note** and **Supplementary Figures 3-10**. The key components of the PIGA pipeline are summarized as follows:

1. Comprehensive SNV detection and phasing: We leveraged hybrid sequencing data to create a comprehensive SNV callset, which improved the ability to detect SNVs in repetitive regions compared to

using short-read data alone. These SNVs were then phased into haplotypes using long-read information, providing accurate phasing blocks that are essential for diploid genome assembly.

2. **Personalized reference genome construction**: We generated a personalized reference for each individual by modifying the reference genome with homozygous variants genotyped from an external graph-based pangenome. For our modest-coverage dataset, this personalized reference improved read alignment and enabled more long reads to be partitioned into haplotypes, compared to using a standard reference or a *de novo* squashed assembly.
3. **Reference-guided draft diploid genome assembly**: Using the personalized reference as a guide, we generated draft diploid assemblies for each individual. This approach demonstrated a significant improvement in assembly completeness compared to the *de novo* assembly using partitioned long reads. Multiple rounds of polishing with both long and short reads were then conducted, which continuously improved the quality of the draft assembly. Although these generated draft assemblies still contained incomplete regions, they provided a wealth of haplotype-resolved sequences across most of the genome and serve as a crucial foundation for generating final diploid assemblies.
4. **Pangenome construction and simplification**: A population-scale pangenome was constructed using the reference genomes, high-coverage Hifiasm assemblies, and the PIGA draft assemblies. To mitigate pangenome errors introduced during assembly or pangenome alignment, we trained a multi-layer perceptron classifier to distinguish between true variants and erroneous nodes/edges, based on a rich set of pangenome-derived features. This machine learning-based filtering outperforms simpler methods based solely on allele count, showing strong performance even in distinguishing singleton variants and errors. The resulting filtered pangenome is significantly smaller, enabling efficient downstream analyses.
5. **Final diploid assembly generation**: We infer diploid paths for each individual through the constructed pangenome by integrating multiple layers of information, including short reads, long reads, draft assembly haplotypes, and SNV haplotypes. The final diploid assembly is then reconstructed from these paths, with additional refinement steps to correct SV-discordant regions and clip k-mer erroneous regions.

We have also updated Supplementary Figure 3 and referenced it in the Results section.

Updated text (lines 93-96): “To fully utilize this population-scale hybrid sequencing dataset with a large fraction of modest-coverage samples, we developed the Pangenome-Informed Genome Assembly (PIGA) workflow (Methods, Supplementary Note, and Supplementary Figures 3-10).”

4. The authors state on lines 82-89 that 10 samples were sequenced by both high- and modest-coverage sequencing strategies, but only 3 samples (line 95) were chosen as benchmark samples. Why were only 3 samples used, and why these 3 specifically?

Re: We now clarify in the revised manuscript that the "benchmark samples" refer to the ten overlapping samples from both high- and modest-coverage datasets, which were used to evaluate individual steps of the PIGA workflow (as described in Supplementary Note). Within the PIGA workflow, we trained a multi-layer perceptron (MLP) classifier to distinguish between true variants and erroneous nodes/edges in the pangenome, based on a rich set of pangenome-derived features (**Figure R27** below). Of the 10 overlapping samples, 7 were used to label true variants and erroneous nodes/edges for model training, and the remaining 3 were reserved as independent benchmark samples for downstream evaluation. These three samples were selected to represent both sexes (one male and two females), which was the only criterion for their selection.

We have revised the Results section to clearly define the benchmark samples and described the machine learning-based filtering process in the Supplementary Note.
Updated figure: **Supplementary Figure 8** (corresponding to **Figure R27** below)
Updated text (lines 100-102): "Ten overlapping samples were used to benchmark individual steps of the PIGA workflow, and a subset of three samples with assemblies from both methods was used for downstream evaluation."
Updated text (Supplementary Note, lines 156-172): "To mitigate pangenome errors introduced during assembly or alignment, we implemented a machine learning-based pangenome filtering strategy. This method trains a Multi-Layer Perceptron (MLP) classifier to distinguish between true variants and erroneous nodes/edges, based on a rich set of pangenome-derived features (Supplementary Figure 8), including: allele count, number of heterozygous and homozygous carriers, whether the node/edge is covered by Minigraph paths, whether flanking sequences are homopolymers, node length, and graphlet-based descriptors that capture the local structure around the node/edge. The predicted error nodes/edges are then filtered out, followed by removing graph tips using VG clip. After simplification, we injected the previously generated SNV haplotypes into the pangenome as additional SNV paths, enabling integration of haplotype information for downstream diploid path inference. Finally, all subgraphs were merged into the final pangenome graph.

We applied this pipeline to construct a population-scale pangenome from 2,248 haploid assemblies, including CHM13, GRCh38, 2,142 1KCP modest-coverage draft assemblies (including assemblies of 10 samples overlapping with the high-coverage dataset: 7 for model training and 3 for

downstream benchmarking) and 104 1KCP high-coverage Hifiasm assemblies (excluding the 3 benchmarking samples).”

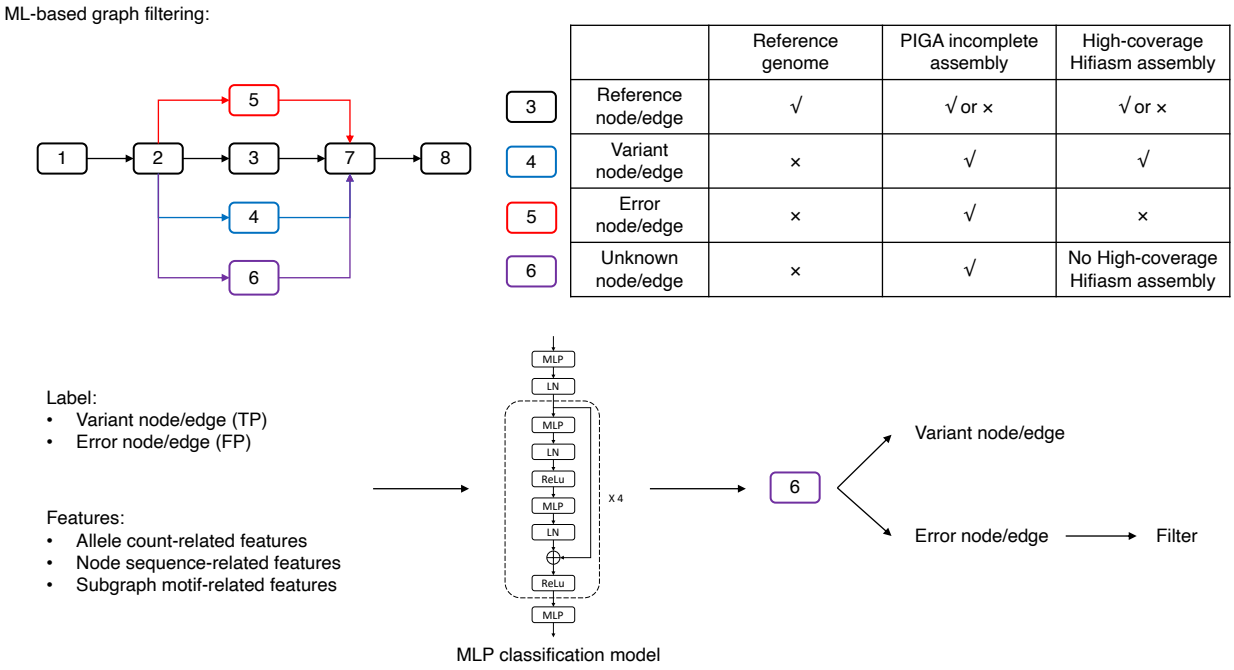

Figure R27. Framework of the PIGA machine learning-based pangenome filtering. Overview of the workflow used to identify and remove error nodes/edges from the pangenome graph. Ground-truth labels for training are assigned by comparing node/edge presence across three sources: the reference genome, PIGA draft assemblies, and high-coverage Hifiasm assemblies. This classification defines each node/edge as reference, variant, error, or unknown. A rich set of features, categorized into allele counts, node sequences, and local graph topology, are extracted for each node/edge and used to train a MLP classifier. The trained model is then applied to classify unknown nodes/edges as either variant or error.

5. In many cases, it was unclear to me which findings were truly enabled by the pangenome approaches employed and which findings could have been obtained through conventional analyses based on linear representations of the reference genome.

Re: In our study, we refer to the broad-sense pangenome as the collection of diploid assemblies from a population, as introduced in the Introduction section. While graph pangenome-based analyses are an essential component, several parts of our work still rely on linear reference-based approaches (e.g., TR and HLA analyses). The primary goal of our work is to demonstrate the value of a cohort-level assembly panel.

First, our benchmarking analysis revealed that our PIGA assemblies significantly outperformed read-based variant calling across various variant types, including SNVs, indels, SVs, and TRs (**Figure R5a-c** and **Figure R25b** below). Moreover, the assemblies provided more accurate sequences for SVs and TRs compared to read-based approaches (**Figure R5d** and **Figure R25a** below), demonstrating their ability to support fine-grained genomic analyses, including resolving sequence composition and multiallelic complexity of SVs and TRs, detecting and functionally annotating nested variants, and performing high-resolution HLA typing. Together, these advantages of the assemblies enabled the construction of an extensive variant catalog for a large cohort.

The graph pangenome analysis can be considered a specific advantage of assemblies, as it provides a unified coordinate system to integrate assemblies across the cohort. This enables effective joint analysis of SVs by alleviating inconsistencies in SV breakpoints across individuals that often arise in linear reference-based analyses, and facilitates investigation of non-reference sequences, detection of nested variants, and visualization of multiallelic SV sites.

In summary, these complementary advantages form the basis of our study, and the following responses address the reviewer's specific questions.

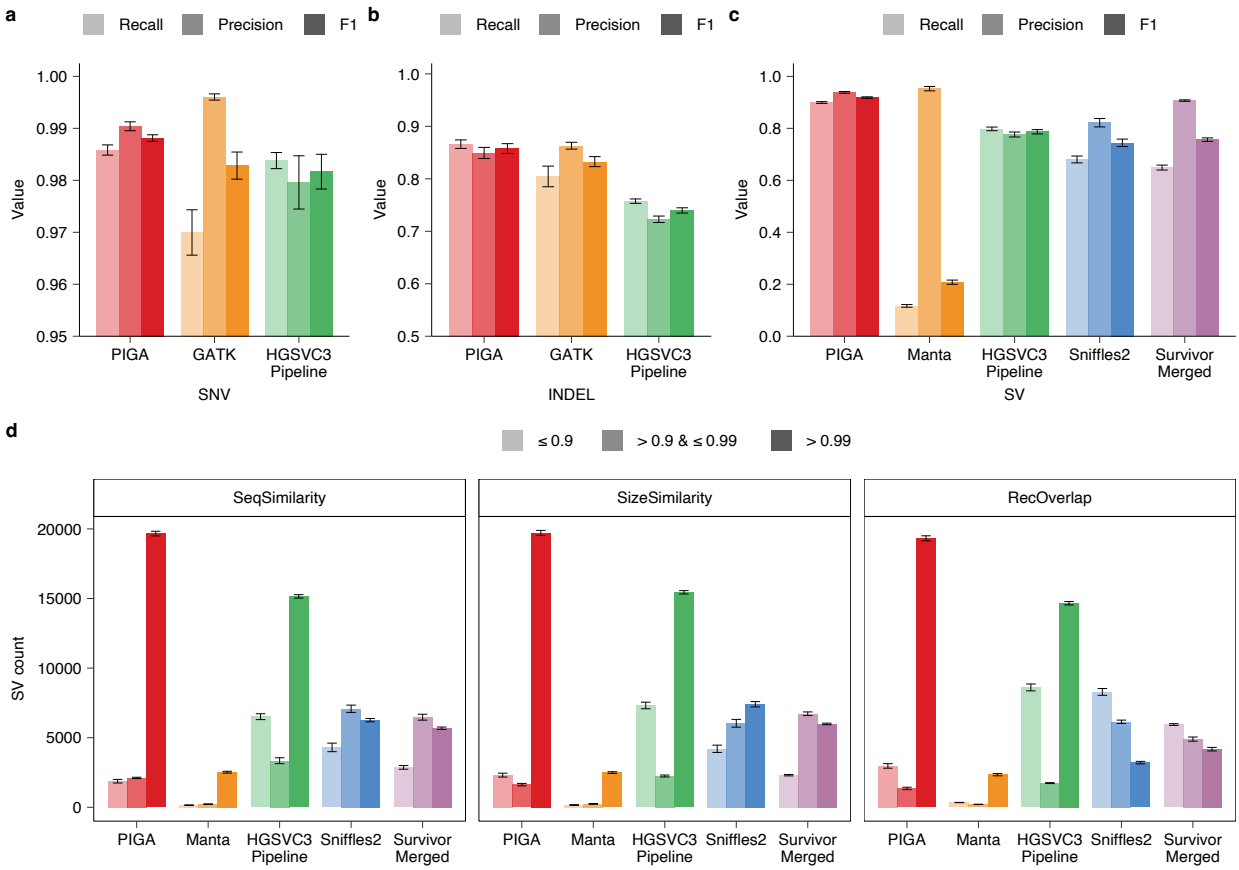

Figure R5. Benchmarking of variant detection performance across different methods in three benchmark samples. a, SNV detection performance evaluated using hap.py. Error bars indicate 95% confidence intervals. **b**, Indel detection performance evaluated using hap.py. Error bars indicate 95% confidence intervals. c, SV detection performance evaluated using Truvari bench (--pick multi --sizemin 50). Error bars indicate 95% confidence intervals. d, Number of SVs stratified by sequence accuracy levels. Accuracy is evaluated using three metrics: sequence similarity (left), size similarity (middle), and reciprocal overlap percentages (right). Error bars indicate 95% confidence intervals. (This figure is also included in our response to Reviewer #1.)

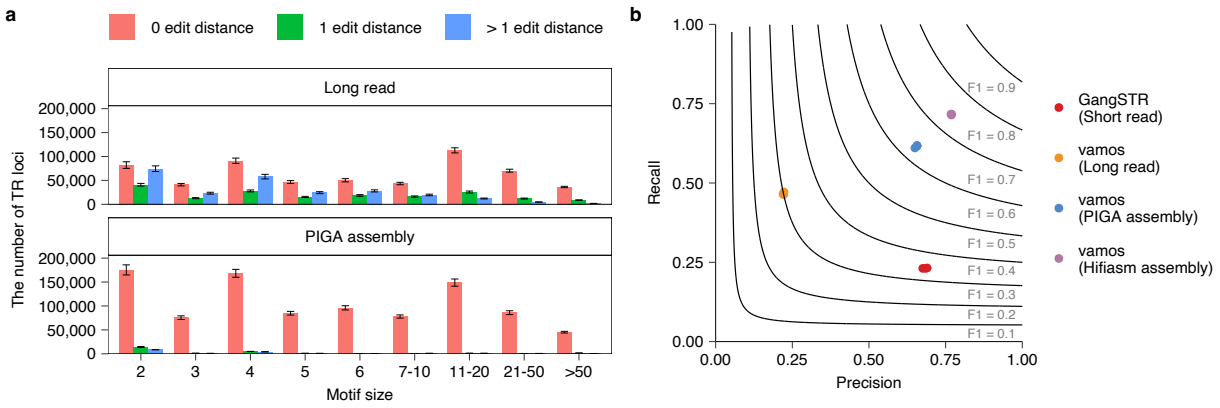

Figure R25. **Benchmark of TR calling performance across different methods in three benchmark samples.** a, Number of TR loci characterized by vamos, stratified by motif size and edit distance. Edit distance refers to the differences in motif unit representations of TR alleles between the query callsets (Long-read and PIGA assembly) and the truth callset (Hifiasm assembly). Error bars indicate the 95% confidence intervals. **b**, TR benchmarking performance using Truvari bench (-s 5) and refine (-u), restricted to GIAB-defined TR regions. The black curves represent constant F1 scores from 0.1 to 0.9. (This figure is also included in our response to Reviewer #2.)

5a. For example, the section starting on line 238 describes various mutations in clinically relevant genes, such as SVs and TR expansions within genes in the OMIM and COSMIC databases. What proportion of these were undetectable based on short-read data alone and/or long-read data but conventional linear reference-based analysis?

Re: Following the reviewer's suggestion, we generated SV and TR callsets using short-read and long-read data based on a linear reference genome for comparison. For SVs, the long-read callset was jointly called using Sniffles2, and the short-read callset was generated using Manta and merged with SURVIVOR. For TR expansions, the long-read callset was generated using vamos, and the short-read callset was generated using GangSTR.

We found that among 5,239 exon-overlapping SVs, 53.2% (2,788/5,239) were detected in the long-read Sniffles2 callset and 26.4% (1,383/5,239) in the short-read Manta callset (Table R1 below). Taken together, 61.8% (3,237/5,239) of these SVs were detected by either method. For TR expansions, of 124 exon-overlapping events identified by the DBSCAN approach, 75 were detected in the long-read vamos callset and 1 in the short-read GangSTR callset. In total, 75 exon-overlapping TR expansions were detected by either method.

We have incorporated these results into the revised manuscript.

Updated Table: **Supplementary Table 15** (Corresponding to **Table R1** below)

Updated text (lines 818-821): "The SV callset was also compared against other SV callsets, including public callsets, as well as the 1KCP long-read Sniffles2 callset (jointly called using Sniffles2) and short-read Manta callset (generated with Manta and merged using SURVIVOR) (Supplementary Table 15), using Jasmine v1.1.5 (--allow_intrasample)."

Updated text (lines 823-824): "For evaluating the TR callset, we additionally called TRs from short-read data using GangSTR v2.5 and from long-read data using vamos."

Table R1. Comparison of exon-overlapping SVs and TR expansions with the 1KCP read-based callsets

Variant type	Count	Count (proportion) in the long-read callset	Count (proportion) in the short-read callset	Count (proportion) in either callset
SV	5239	1383 (26.4%)	2788 (53.2%)	3237 (61.8%)
TR expansion	124	1 (0.8%)	75 (60.5%)	75 (60.5%)

5b. As another example, line 308 states that "compared to short-read data, the 1KCP assemblies enabled us to resolve the HLA alleles at the four-field resolution". But what is the basis for this statement? Was the same analysis attempted with short-read data, and if so, what were the results?

Re: We have added a more detailed statement regarding the advantage of assemblies for HLA typing in the revised manuscript.

Updated text (lines 338-340): "Compared to previous studies using short-read data, which were limited to resolving HLA alleles at a maximum of three-field resolution, the 1KCP assemblies provide an extended phasing range across both coding and noncoding regions of HLA genes, enabling four-field resolution of HLA alleles."

We have also performed HLA typing using short reads with HLA-LD. While assembly-based and short-read-based typing results show high concordance (97.3% across genes) at the three-field resolution, only the assembly-based approach enables HLA typing at four-field resolution (**Figure R12** below). We have incorporated these results into the revised manuscript.

Updated text (lines 341-343): "The typing results showed a mean three-field concordance rate of 97.3% on 22 shared genes with the short-read typing results obtained using HLA-HD (Supplementary Figure 32)."

Updated text (lines 893-894): "We also performed independent HLA typing from short reads using HLA-HD to validate the HiFiHLA typing results."

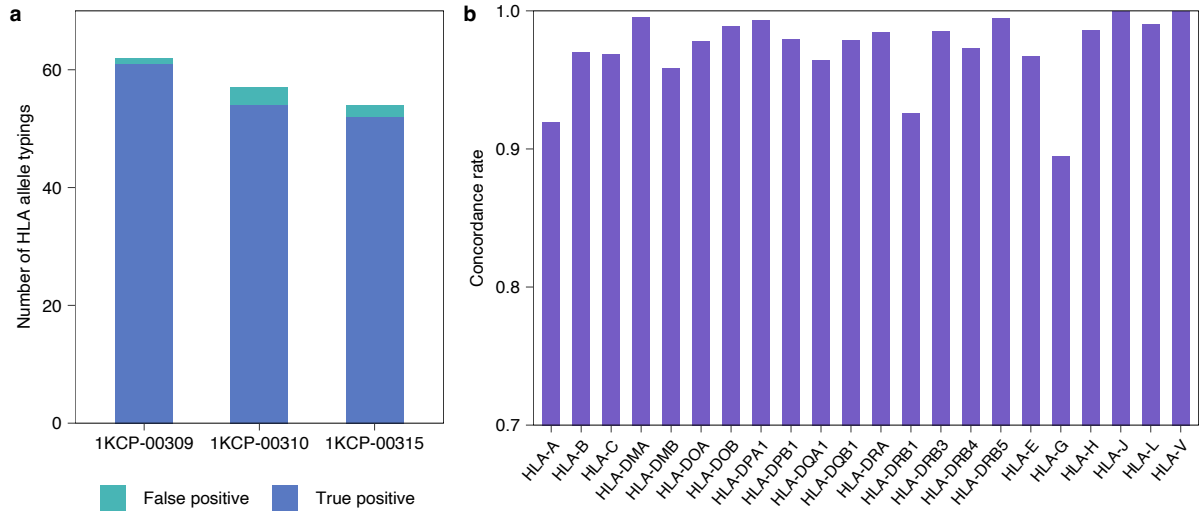

Figure R12. Evaluation of HLA typing accuracy. a, Number of validated four-field alleles in PIGA assembly-based HLA typing compared with Hifiasm assemblies in three benchmark samples. **b**, Three-field allele concordance rates of assembly-based typing results compared with short-read-based typing using HLA-HD on 22 shared HLA genes. (This figure is also included in our response to Reviewer #1.)

5c. As a third example, the imputation results (starting line 399) contrasted between two different genotyping arrays -- which I found a bit tangential -- but did not contrast between this new pangenome-enabled reference panel and conventional reference panels (either built from the same samples or other common imputation panels such as 1000 Genomes Project, TopMed, HRC, etc.).

Re: We thank the reviewer for this helpful suggestion. To simplify the benchmarking design and avoid confusion, we have revised the manuscript to remove the ASA array results from the leave-one-out imputation analysis. This section now focuses on comprehensively evaluating imputation performance across all variant types (**Figure R28** below).

Updated figure: **Figure 5** (Corresponding to **Figure R28** below)

Updated text (lines 441-445): "To comprehensively evaluate the imputation performance of the 1KCP reference panel across diverse variant types, we conducted imputation for each of the 55 internal high-

coverage samples individually, using a reference panel that excluded the imputed sample (Methods). We extracted SNV genotypes from genome assemblies at sites matching those on the Omni2.5 arrays to create pseudo-array input data."
Updated text (lines 959-960): "The SNV sites used for imputation were restricted to those accessible by the Infinium Omni2.5 array or short-read WGS."

In addition, we have now benchmarked the performance of the 1KCP panel against existing imputation panels targeting specific variant types, using 70 external East Asian samples from HPRC, HGSVC3, and CPC with high-quality assemblies. For SVs, we compared imputation performance against the 1000G-ONT SV panel. The 1KCP panel achieved a higher mean r^2 across MAF ranges and imputed 26% more SVs with $r^2 > 0.8$ (**Figure R13a-b** below). For TRs, the 1KCP panel achieved a higher mean F1 score than the EnsembleTR panel (0.56 vs 0.46) (**Figure R13c** below), and uniquely enabled imputation of TR motif composition. For HLA alleles, while the 1KCP panel showed a slightly lower mean r^2 for shared alleles than the Four-digit multi-ethnic HLA v2 panel (0.89 vs 0.93), likely reflecting the difference in panel sample sizes (1,116 vs 20,349), it enabled imputation of 2.7-fold more alleles with $r^2 > 0.8$ (**Figure R13d-e** below). These additional alleles were largely from low-resolution (one- and two-field) alleles of non-classical HLA genes and high-resolution (three- and four-field) alleles across all HLA genes. Finally, for nested variants within non-reference sequences of the pangenome, the 1KCP panel was the only resource capable of imputation.

We have incorporated these results into the revised manuscript.
Updated text (lines 460-472): "To benchmark the 1KCP panel against existing panels targeting specific variant types, we further evaluated imputation performance on 70 external East Asian samples with high-quality assemblies (Methods and Supplementary Table 34). For SVs, we compared imputation performance against the 1000G-ONT SV panel. The 1KCP panel achieved a higher mean r^2 across MAF ranges and imputed 26% more SVs with $r^2 > 0.8$ (Supplementary Figure 36a-b). For TRs, the 1KCP panel achieved a higher mean F1 score than the EnsembleTR panel (0.56 vs 0.46) (Supplementary Figure 36c), and uniquely enabled imputation of TR motif composition. For HLA alleles, while the 1KCP panel showed a slightly lower mean r^2 for shared alleles than the Four-digit multi-ethnic HLA v2 panel (0.89 vs 0.93), likely reflecting the difference in panel sample sizes (1,116 vs 20,349), it enabled imputation of 2.7-fold more alleles with $r^2 > 0.8$ (Supplementary Figure 36d-e). These newly imputed alleles were largely from low-resolution (one- and two-field) alleles of non-classical HLA genes and high-resolution (three- and four-field) alleles across all HLA genes. Finally, for nested variants within non-reference sequences of the pangenome, the 1KCP panel was the only resource capable of imputation."

Updated text (lines 968-78): "For comparison with the existing panels, we used 70 East Asian samples from CPC, HPRC, and HGSVC3 to construct the truth callsets and performed imputation benchmarking. The SV callset was generated using PAV and further merged using Jasmine. The HLA callset was generated using HifiHLA. Since the 1000G-ONT SV panel contains only SNV sites restricted to the UKB array, imputation for SV comparison was performed using the UKB pseudo-array genotypes as input. For TR and HLA comparisons, we used the Omni 2.5 pseudo-array genotypes as input. Imputations using the 1000G-ONT SV panel and the EnsembleTR panel were conducted with Minimac4. Imputation with the Four-digit multi-ethnic HLA v2 panels was performed on the Michigan online imputation server. For SV benchmarking, we used truvari bench (--pick ac -p 0.7 -P 0.7 -s 50) to match SVs across different callsets. For TR benchmarking, we assessed the performance of imputed TR alleles using Truvari bench (-s 5) and refine (-u) against the PAV callset, restricted to GIAB-defined TR regions."

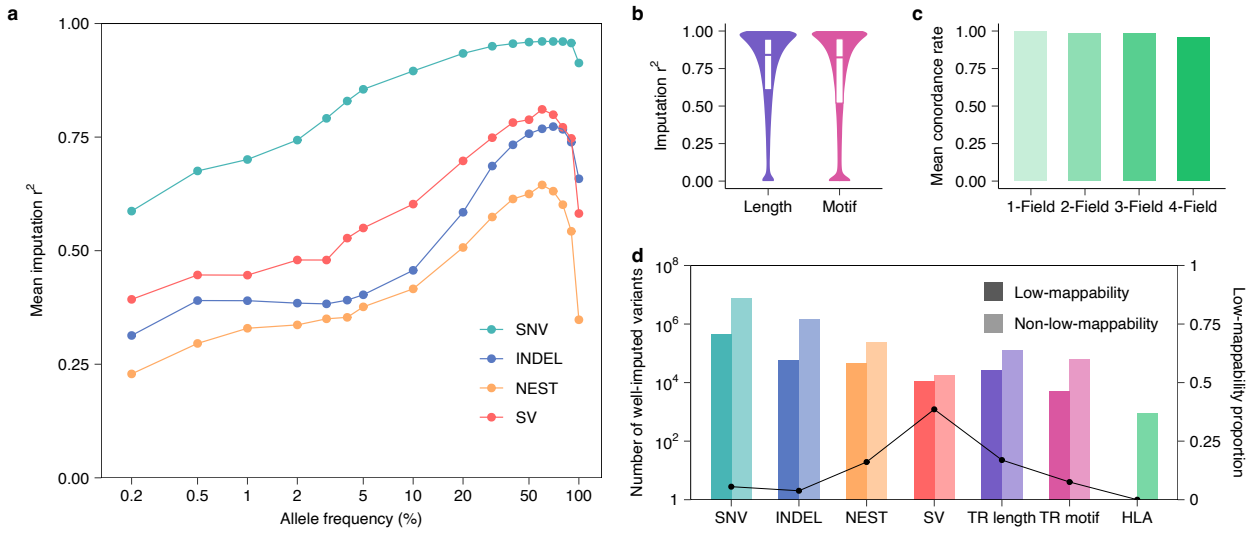

Figure R28. 1KCP imputation panel. **a**, Imputation performance (r^2) for different variant types stratified by AF bins. **b**, Imputation performance (r^2) for TR length and motif variants. **c**, Imputation allele concordance for HLA genes at different resolutions. **d**, Columns indicate the number of well-imputed variants ($r^2 > 0.8$) stratified by regional mappability. Dots indicate the proportion of well-imputed variants in low-mappability regions.

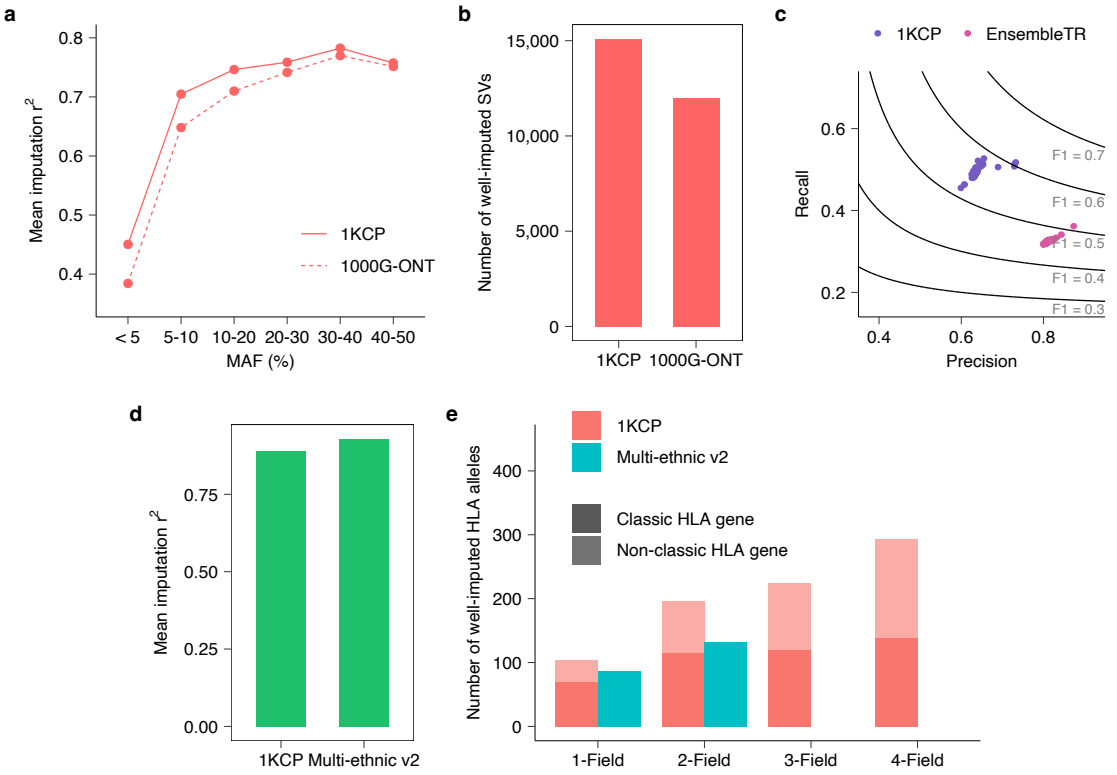

Figure R13. Comparison of imputation performance between the 1KCP reference panel and existing panels. a, Imputation performance (r^2) of the 1KCP and 1000G-ONT panels for shared SVs stratified by MAF bins. **b**, Numbers of well-imputed SVs ($r^2 > 0.8$) across the 1KCP and 1000G-ONT SV panels. **c**, Benchmarking of imputed TR alleles from the 1KCP and EnsembleTR panels, evaluated with Truvari bench (-s 5) and refine (-u), restricted to GIAB-defined TR regions. The black curves represent constant F1 scores from 0.3 to 0.7. **d**, Imputation performance (r^2) of the 1KCP and Four-digit multi-ethnic HLA v2 panels for shared HLA alleles. **i**, Numbers of well-imputed HLA alleles ($r^2 > 0.8$) from the 1KCP and Four-digit multi-ethnic HLA panels stratified by allele resolution levels and gene classes. (This figure is also included in our response to Reviewer #1.)

5d. As a fourth example, can the GSTM1 and GSTM2 deletions be identified with short-read data alone and/or long-read data but conventional linear reference-based analysis?

Re: In the revised manuscript, we have clarified that the deletion spans *GSTM1* but not *GSTM2*, and we provide fine-mapping evidence supporting its role as the most likely causal eVariant for both genes.

Updated text (lines 422-424): "For example, we identified an 18 kb deletion spanning the *GSTM1*, which is the most likely causal eVariant for both *GSTM1* and *GSTM2* based on fine-mapping ($PIP_{SuSiE\text{-}GSTM1} = 1$, $PIP_{SuSiE\text{-}GSTM2} = 1$)."

Regarding detection using read data, we generated SV callsets using short-read data (Manta) and long-read data (Sniffles2). Although both approaches could detect the deletion in the *GSTM1* locus, the allele counts in these callsets were substantially lower compared to the PIGA callset, indicating the PIGA SV callset better reflects the characteristics of SVs from a population genetics perspective (**Table R2** below). As a result, the association p-values of the *GSTM1* deletion with expression of *GSTM1* and *GSTM2* were weaker in these callsets compared to the PIGA callset.

Table R2. Comparison of eQTL signals for the *GSTM1* deletion across callsets.

Callset	Chrom	Start	End	Allele count	P_{GSTM1}	P_{GSTM2}
PIGA assembly	chr1	109683912	109702358	1679	1.1e-235	1.1e-220
Short reads (Manta)	chr1	109687622	109711967	53	1.1e-5	2.1e-5
Long reads (Sniffles)	chr1	109683689	109702133	15	0.51	1.9e-2

6. The use of SEI and Enformer for regulatory annotation was interesting, though these should be carefully described as regulatory predictions rather than "regulatory elements" (see line 31), and it should also be noted that these are specific to the blood—a limitation that could be expanded upon in the Discussion. More details are needed about the benchmarking approach and what 3 "benchmark samples" were used. Moreover, it seems like SEI and Enformer were compared based on their ability to predict ATAC-seq peaks (which showed high accuracy), but then applied for more granular predictions of various regulatory elements (promoters, enhancers, CTCF sites, etc., line 140). What is the justification for this approach, and what should be expected for the accuracy of the predictions?

Re: We agree that the term “regulatory elements” should be described more rigorously, and we have replaced it with “predicted regulatory elements” throughout the manuscript.

Updated text (lines 28-31): “Based on these assemblies, we constructed a pangenome comprising 405.3 million base pairs of sequences absent from the current references GRCh38 and CHM13, including 26.2 million base pairs of functional genic and predicted regulatory elements.

Updated text (lines 158-159): “On average, 59.54% of each genome was predicted as epigenomic elements, such as promoters, enhancers, and CTCF binding sites.”

Updated text (lines 200-202): “Additionally, we identified 26.2 Mb of non-reference sequences containing functional genic and predicted regulatory elements (excluding introns and heterochromatin), such as enhancers, promoters, and transcription factor (TF) binding sites.”

Updated text (lines 503-505): “This approach identified 26.2 Mb of functional genic and predicted regulatory elements within the non-reference sequences, which were further validated by the enrichment of lead eNested variants in the pan-variant eQTL analysis.”

We also acknowledge that the epigenome predictions generated in this analysis were derived from a broad tissue context, not limited to blood. For benchmarking, we used blood ATAC-seq data from the same three benchmark samples as in our previous analysis to compare the performance of SEI and Enformer on both Hifiasm and PIGA assemblies. We selected these samples because evaluating predictions for non-reference sequences requires matched samples between the assemblies and ATAC-seq experiments. We have revised the Methods section to describe this process in detail.

Updated text (lines 728-734): “To annotate epigenomic elements, we compared the performance of SEI and Enformer for epigenome prediction in 128 bp steps on both PIGA and Hifiasm assemblies from three benchmark samples. Non-reference sequence regions in each assembly were identified using PAV and defined as insertion records ≥ 50 bp. ATAC-seq peaks were identified on the assemblies of the matched samples using MASC3 v3.0.0 with an FDR < 0.01 and served as ground truth for model evaluation. Predicted regions were labeled as true positive if they overlapped with the ground truth peaks and false positive otherwise. The AUROC for each predicted track was calculated to evaluate the performance of SEI and Enformer.”

Regarding using ATAC-seq data for benchmarking, we agree with the reviewer that high accuracy in ATAC-seq peak prediction does not necessarily guarantee equivalent accuracy for more granular regulatory annotations in a broad tissue context. Our rationale for using blood ATAC-seq peak performance as an indicator is twofold:

1. Both SEI and Enformer are multi-task deep learning models trained to predict a wide range of epigenomic profiles simultaneously. Thus, performance on blood chromatin accessibility predictions can partially reflect overall model performance.
2. Chromatin accessibility is a shared upstream feature of many regulatory elements (Klemm et al. 2019 Nature Reviews Genetics), and the SEI predicted regulatory element classes are derived from its predicted epigenomic profiles. Therefore, accurate accessibility predictions provide a reasonable proxy for assessing model competence in predicting regulatory elements.

In the Results section, we have clarified this rationale and the selection of SEI:

Updated text (lines 149-158): “Since chromatin accessibility is a fundamental feature of many regulatory elements (Klemm et al. 2019 Nature Reviews Genetics), we used blood ATAC-seq data from three benchmark samples to compare the prediction performance of two state-of-the-art models, SEI and

Enformer, on both Hifiasm and PIGA assemblies (Methods). Both models accurately predicted blood-related epigenomic profiles when compared to real ATAC-seq data, with SEI achieving superior performance (average area under the receiver operating characteristic (AUROC) for the best-predicted track: 0.9802) and demonstrating better generalization capability to non-reference sequences in individual assemblies (Figure 1f and Supplementary Table 8). Using these blood-specific results as an indicator of model performance, we selected SEI to predict broader cross-tissue epigenomic profiles and annotate epigenomic elements based on its defined sequence classes."

To further assess the accuracy of predicted epigenomic elements, we integrated assembly annotations into the pangenome graph and measured prediction concordance among haplotypes. Predicted epigenomic elements showed lower concordance than other annotation types (**Figure R29** below). Therefore, we retained only annotations with > 80% concordance to improve the reliability of the predicted regulatory elements.

We have expanded the Discussion section to note the limitations of the current epigenome predictions.
Updated figure: **Supplementary Figure 22b** (corresponding to **Figure R29** below)
Updated text (lines 507-515): "While the SEI model demonstrated strong generalization ability in predicting blood-related epigenomic profiles and showed potential for elucidating the functional roles of non-reference sequences, its current application still reveals limitations in predicting concordant epigenomic elements across different assemblies (Supplementary Figure 22b). More advanced genomic language models may help address this limitation and enable more accurate annotation of epigenomic elements in non-reference sequences. Furthermore, our benchmarking relied solely on blood-specific ATAC-seq data, which may not fully reflect model performance across diverse tissues and regulatory contexts. Expanding benchmarking to include multiple tissue types and experimental assays will be essential to validate and refine these predictions."

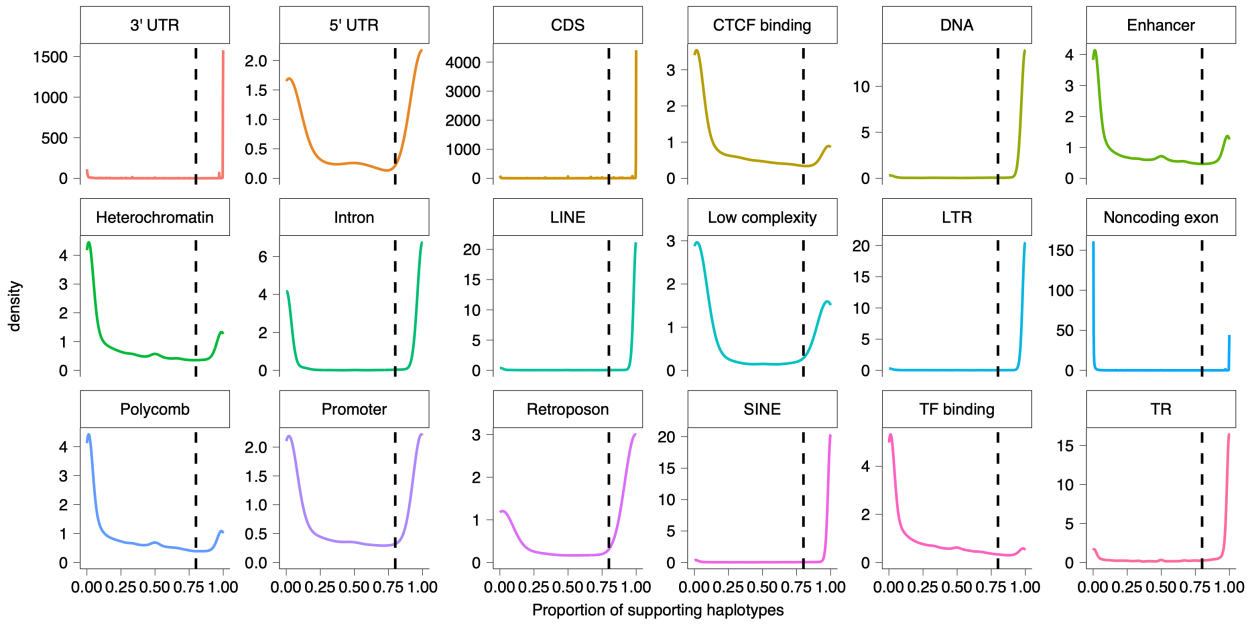

Figure R29. Pangenome annotation metrics. Distribution of supporting haplotype proportions for annotated genomic elements. Each panel represents a specific genomic element category.

Reference

1. Klemm, S. L., Shipony, Z. & Greenleaf, W. J. Chromatin accessibility and the regulatory epigenome. ***Nature Reviews Genetics*** 20, 207-220 (2019).

7. Many of the eQTL results focused on cases where an SV, TR, or nested variant (together termed "complex variants") was the lead eQTL. However, as far as I could tell, these results were not based on standard fine-mapping approaches such as CAVIAR, SuSiE, etc. so it is unclear how confident we should be that a given complex variant is causal. I recommend that a fine-mapping approach such as SuSiE that allows for the possibility of multiple causal variants for a given gene be applied, followed by an analysis to determine in how many cases a complex variant has the highest posterior inclusion probability and the extent to which these are enriched compared to their background proportions.

Re: Following the reviewer's suggestion, we performed fine-mapping analysis using the SuSiE algorithm implemented in TensorQTL. This yielded 17,563 credible sets, of which 1,364 contained complex variants with the highest posterior inclusion probability (PIP). We also evaluated enrichment of eVariants with PIP > 0.9 across variant types. TRs exhibited the strongest enrichment among variant types (fold enrichment = 2.96), as estimated from 100,000 resampling iterations (**Table R3** below).

We have incorporated these results into the revised manuscript.

Updated table: **Supplementary Table 27** (corresponding to **Table R3**)

Updated text (lines 381-385): "Fine-mapping analysis yielded 17,563 credible sets, of which 1,364 contained complex variants with the highest posterior inclusion probability (PIP) (Supplementary Table 26). Among the eVariants with PIP > 0.9, TRs exhibited the strongest enrichment across variant types (fold enrichment = 2.96; Supplementary Table 27)."

Updated text (lines 918-919): "Fine-mapping analysis was performed using the SuSiE algorithm implemented in tensorQTL to identify credible sets of eQTL signals."

Update text (lines 931-934): "To assess the enrichment of eVariants with PIP > 0.9 across variant types, we performed 100,000 resampling iterations. For each resampling, an equal number of variants with matching distance to TSS as the eVariants with PIP > 0.9 were randomly selected to generate the null distribution. Fold-enrichment and empirical p-values were then calculated for each genomic element based on the resampling results."

Table R3. Enrichment of fine-mapped eVariant with PIP > 0.9.

Type	Count	Fold enrichment	95% CI lower	95% CI upper	Empirical *P* value
SNV	2864	1.38	1.34	1.42	1.0×10^{-5}
indel	384	0.33	0.31	0.35	1.0
SV	18	1.07	0.72	2.00	0.42
Nest variant	15	0.06	0.05	0.06	1.0
TR	372	2.96	2.51	3.54	1.0×10^{-5}

8. The imputation methods also require much more detailed description. One major question I have is whether the imputation method employed leverages the pangenome graph or whether variant calls are ultimately collapsed back to a linear representation and used as a reference panel with a conventional imputation tool. While such an approach may be justified, I am wondering if all of the nested variants can actually be represented in a biallelic form while preserving information about their nestedness or if there is information loss when converting from the pangenome into the decomposed biallelic representation. If there is no information loss, then what is the value of the pangenome approach in the first place?

Re: Our current imputation pipeline was performed using the conventional imputation tool Minimac4 on a linear reference representation, with all variants decomposed into biallelic form. The primary reason for this choice is the lack of mature and scalable imputation tools that can directly operate on a pangenome

graph with complex multi-allelic variants. We have revised the Methods section to describe this process in detail.

Updated text (lines 950-958): "We constructed the 1KCP imputation reference panel based on the GRCh38 linear reference, incorporating a comprehensive range of variant types in biallelic form, including SNVs, indels, SVs, nested variants, TR variants, and HLA alleles. For HLA alleles, the typing results were converted into variant representations, with the variant positions defined at the center of the corresponding HLA genes. Variants with a missing rate > 10%, HWE p-value < 1×10^{-6}, or allele count < 2 were excluded from the reference panel.

To evaluate imputation performance, we performed leave-one-out imputation using Minimac4 v4.1.6, where each of the 55 high-coverage samples was sequentially excluded as the imputation target sample."

While the nested relationships of nested variants were preserved in the annotation metadata, we acknowledge that converting them into biallelic form inevitably leads to information loss. This limitation is consistent with current practice in genetic analyses, where loci containing multiple alleles (e.g., both an indel and SNPs) are typically analyzed in a biallelic framework, losing the ability to directly compare effect sizes among alternative alleles. We have revised the Discussion section to emphasize the limitations.

Updated text (lines 521-523): "However, current genetic analyses typically model these complex loci in a biallelic framework on a linear reference representation, which underscores the need for advanced methods to jointly model the phenotypic effects of all alleles within multiallelic SV loci."

The value of the pangenome approach in this context lies in its ability to align non-reference sequences among different haplotypes and discover nested variants absent from the linear reference. Without the pangenome graph, such variants would remain undetected, as seen in previous coarse-grained SV studies. In our eQTL analysis, nested variants harbored top association signals approximately five times more than SVs, suggesting that they may be functionally important but overlooked in previous analyses. We now highlight the advantage of the pangenome for discovering nested variants in the Results section.

Updated text (lines 210-212): "In addition to these variant sites located in the reference genome (referred to as reference variants), the pangenome offers a unique advantage in resolving variants nested within non-reference sequences (referred to as nested variants) through pangenome alignment (Figure 2d)."

9. Line 284: I noted that immunoglobulin genes were enriched in complex structural haplotypes. Given the occurrence of somatic recombination and hypermutation in blood immune cells and the fact that the

source material for this study was blood, I am wondering how the authors ruled out potential artifacts of somatic variation.

Re: We agree with the reviewer that somatic recombination and hypermutation events could, in principle, be incorporated into an assembly. However, as each specific somatic recombination event typically occurs in only a small fraction of blood cells, and our assemblies are generated from the consensus of multiple reads derived from many cells, such events are expected to be filtered out as noise during the assembly process.

To further evaluate this possibility, we examined potential V(D)J recombination signals in our assemblies, focusing on the immunoglobulin and T cell receptor gene clusters. Using the pangene-derived gene-gene connection graph, we identified 351 polymorphic gene-gene connections involving V, D, and J genes in these regions, including 344 within the same gene type (V-V or J-J) and 7 between different gene types (V-J) (**Figure R30a** below). We then examined these 7 V-J connections in detail and found that all were separated by > 2 kb within the haplotype assemblies, which is inconsistent with somatic V(D)J recombination events that generate compact V-J junctions (**Figure R30b** below). These results suggest that the observed patterns are unlikely to be artifacts of somatic recombination.

We have incorporated these results into the revised manuscript.
Updated figure: **Supplementary Figure 31** (corresponding to **Figure R30** below)
Updated text (lines 313-315): "Further analysis suggests that these enrichments are unlikely to be attributable to somatic V(D)J recombination (Supplementary Figure 31)."

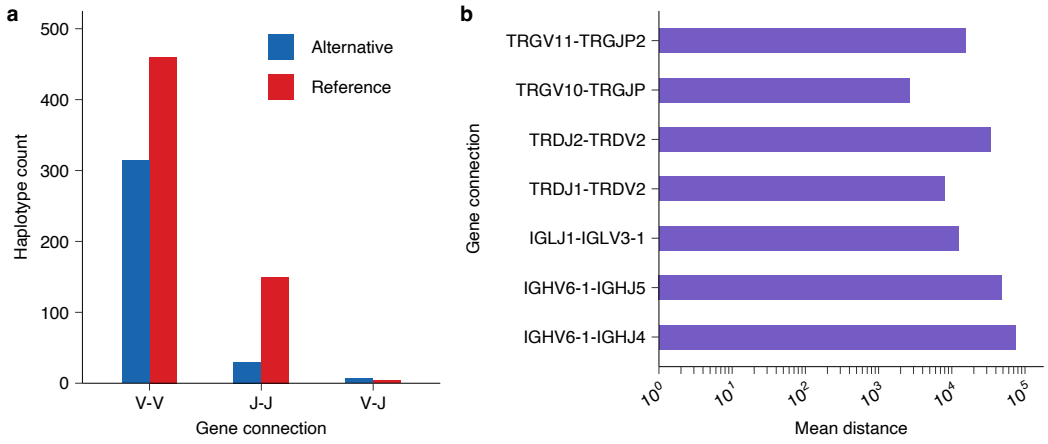

Figure R30. Gene-gene connection analysis among V, D, and J genes within immunoglobulin and T cell receptor gene clusters. a, Number of gene-gene connections involving V, D, and J genes in the

pangene graph. Only seven alternative junctions between different gene types (V-J) were detected. **b**, Mean genomic distances between polymorphic connected V-J gene pairs within the corresponding haplotype assemblies, all exceeding 2 kb, indicating that these patterns are unlikely to result from somatic V(D)J recombination

10. The entropy-based measurement of LD at the HLA locus was interesting, motivated by the inability of conventional metrics (r2, d', etc.) to deal with multiallelic variants or consider 3+ variants at a time. I understand that in a large sample and using long-read technology, it is possible to better resolve the entire sequence of HLA haplotypes, including within non-coding regions (i.e., the fourth field in the HLA allele naming scheme). However, I am wondering about the basis of the statement that "four-field alleles better capture the LD structure of HLA genes" (line 318), which appears to be based on the higher ε value. How can such a statement be made without knowing the ground truth haplotype structure? In addition, the statements about the measured levels of LD partitioned between LD groups seemed circular (lines 321-324). Are LD groups not themselves defined based on levels of LD?

Re: We agree that our original wording ("better capture") was imprecise in the absence of ground truth haplotype structure. As the ε index quantifies the non-random association of alleles between two genes in a multiallelic context, higher ε values indicate reduced randomness in allele combinations, that is, knowing the allele at one gene provides more predictive information about the allele at another gene. We have therefore rephrased the statement to clarify that higher allelic resolution decreases combination randomness and uncovers finer-scale LD patterns that are masked at lower resolution.
Updated text (lines 351-354): "Values of ε calculated from four-field alleles were generally higher than those from three-field alleles (0.0373 vs. 0.0245 on average), indicating that the higher allelic resolution reduces the randomness of allele combinations and reveals a finer-scale LD structure that is masked at lower resolution. (Supplementary Figure 33b-c)."

Regarding the definition of LD groups, the reviewer is correct that our original description could appear circular. Our intention was not to restate the same result but to emphasize the existence of long-range haplotypes in the population that may not be fully captured by gene-level LD. We have therefore rephrased the statement to focus on describing these long-range haplotypes
Updated text (lines 356-359): "When assessing allele-wise LD for HLA alleles with MAF > 0.01, we found that although strong LD relationships ($r^2 > 0.8$) were predominantly within the same LD groups, 34.4% of modest LD relationships (r^2: 0.4-0.8) occurred between alleles not in the same LD group,

revealing the existence of long-range HLA haplotypes in the 1KCP cohort (Figure 3j and Supplementary Table 25)."

Minor comments:

1. Line 197: I was not sure how to interpret the statement that 56% of SV sites are "multi-allelic". Does this just mean that SVs often possess nested variation that is apparent in a sample of this size? Is a nested variant within an SV enough to make it multiallelic? If so, contrasting the number of multiallelic SNVs, indels, and SVs in the Results and Fig. 2f feels a bit trivial/tautological because it simply reflects the fact that SNPs cannot contain nested variation and indels are smaller than SVs by definition.

Re: We acknowledge that the definition of a multi-allelic SV can be flexible, depending on the resolution used to group similar SV alleles. Graph pangenome-based methods offer a unique opportunity to resolve SV allelic diversity at base-level resolution, which is challenging for linear reference–based approaches because their detected SVs often have irregular breakpoints and are difficult to analyze jointly across the cohort.

While we agree that the fine-scale multi-allelic nature of SVs often arises from nested variants, we believe that the observed SV allelic heterogeneity across a large cohort is meaningful and worth reporting. This observation also underscores the importance of the SV merging procedure described in the following section.

We further agree that the direct comparisons with SNVs and indels are less informative in this context. Therefore, we have removed this comparison from both the text and figure (**Figure R15** below), and now focus on characterizing the allelic complexity of SVs.
Updated text (lines 219-220): "We further explored the complexity of SV sites and found that 80.3% of SV sites are multiallelic, with an average of 61.7 distinct alleles observed per site."

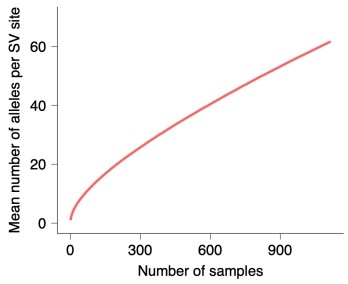

Figure R15. The mean observed number of unmerged alleles per SV site as the sample size increases. (This figure is also included in our response to Reviewer #1.)

2. Line 368 (and throughout): the statement that MAD1L1 "was regulated by" copy number and sequence motif is too strong, and should instead be phrased as association. This issue applies to essentially all of the association mapping results, which should not be interpreted or reported **as causal relationships** (see Major Comment about fine-mapping).

Re: We agree that the original phrasing implying causality was too strong. In the revised manuscript, we have replaced such statements throughout with more appropriate wording.
Updated text (lines 377-379): "Our results showed that complex variants explained an average of 12.6% of the cis-heritability for genes with cis-heritability > 0.05 (n = 7,341), highlighting their contribution to variation in gene expression (Supplementary Figure 22)."
Updated text (lines 407-414): "Among the 1,934 TR sites with lead eVariants, 305 contained multiple unique lead eTR variants, suggesting that the same TR locus may regulate gene expression through different mechanisms. For example, we observed that the expression level of *MAD1L1* was associated with both the copy number of a TR site and its 28 bp motif (CCCAGGCCTCACCCTCCTCTTCCTCCCC). Specifically, each additional TR copy was associated with a 0.126-unit increase in MAD1L1 expression, while an additional motif copy was associated with a 0.245-unit increase (Figure 4d)."
Updated text (lines 418-420): "To address this, we conducted eQTL-GWAS colocalization analysis using SMR and coloc, leveraging the 1KCP eQTLs dataset with a more complete variant profile to prioritize variants potentially underlying GWAS loci."

3. The authors state on line 436 that previously released human pangenomes typically included only a few dozen individuals, but a brief search of the literature suggests this is probably an overstatement, and should be made more specific regarding the technologies involved.
a. Chinese pangenome – 36 samples in Phase I (Gao et al., Nature, 2023) → goal is at least 500
b. Human Pangenome Reference Consortium – 47 samples (Liao et al,, Nature 2023) → goal is 700
c. African Pangenome – 910 samples (Sherman et al., Nature 2019)
d. Denmark Pangenome – 150 samples (Maretty et al., Nature 2017)

Re: We have revised the statement to be more precise and now refer specifically to the "human pangenome constructed from diploid assemblies".

Updated text (lines 490-492): "The 1KCP pangenome, comprising over 1,000 individuals, represents a significant increase in sample size compared to previously released human pangenomes constructed from diploid assemblies, which typically included only a few dozen individuals."

4. The Methods section would benefit from additional details:
Was RNA integrity checked after RNA extraction (e.g. Agilent Tapestation) prior to RNA-seq library preparation (Methods)?
How was the number of PEER factors chosen for eQTL analysis (Methods)?
How were medically relevant genes (line 240) defined?

Re: We have added the requested methodological details in the revised manuscript.

RNA integrity assessment
Updated text (lines 636-637): "The extracted RNA was assessed using an Agilent Bioanalyzer 2100, and only samples with a total RNA amount >1 µg and an RNA Integrity Number (RIN) > 6.5 were used for library preparation."

Choice of PEER Factors
Updated text (lines 908-909): "Following the GTEx Consortium guidelines, the number of PEER factors was selected based on sample size."

Definition of medically relevant genes (MRGs)
Updated text (lines 843-848): "**Definition of MRG sets**
Given that different medical genetics databases curate gene lists with varying scopes and purposes, we defined MRGs specifically for each downstream analysis. For gene-altering SV analysis, MRGs were defined as genes cataloged in the OMIM and COSMIC databases. For TR expansion analysis, MRGs were defined as known disease-associated TR genes cataloged in the STRchive database. For HLA allele analysis, MRGs were defined as genes cataloged in the IPD-IMGT/HLA database."

5. Line 163: the word "non-reference" can be removed, as it is redundant with the phrase "absent from the GRCh38 and CHM13 references"

Re: We have removed the term "non-reference" from the sentence.

Updated text (lines 183-184): “Next, we quantified the sequences in the 1KCP pangenome that are absent from the GRCh38 and CHM13 references.”

6. Figure 1a - Why is the HiFiasm NG50 smaller than that of PIGA, given that the former are higher coverage?

Re: The observed difference in NG50 between PIGA and HiFiasm assemblies is primarily due to the current limitations of the PIGA workflow in resolving highly repetitive regions, such as segmental duplications and satellites. These regions are challenging to assemble because noisy non-HiFi long reads have limited power to resolve highly repetitive sequences, and such regions are often clipped during pangenome construction due to their extremely complex graph structures. In such cases, the PIGA workflow may aggressively borrow sequences of these regions from complete reference haplotypes (e.g., CHM13) to reconstruct the final assembly, even when they do not match the individual's actual haplotype. As a result, these regions may appear in PIGA assemblies, contributing to a higher NG50 but also introducing more errors (**Figure R10** below). In contrast, HiFiasm assemblies correctly introduce breaks in regions that cannot be unambiguously resolved, leading to a lower NG50 but a higher QV. To mitigate potential artifacts, we defined 53.7 Mb of common unreliable regions in the GRCh38 reference shared across individuals using flagger annotations, which were used in quality control steps to improve the reliability of the 1KCP dataset (as described at the beginning of this document).

We have clarified this limitation in the Discussion section.
Update text (lines 539-547): “Finally, highly repetitive regions, such as segmental duplications and satellites, remain challenging to resolve with the current PIGA workflow. The reliance on noisy non-HiFi reads limits resolution in these regions, and such sequences are often clipped from the PIGA intermediate pangenome due to their extremely complex graph structure. Therefore, the current PIGA workflow may aggressively borrow sequences of these regions from complete reference haplotypes to reconstruct the final assembly. To mitigate potential artifacts, we identified common unreliable regions using Flagger annotations and excluded the affected genes and variants from downstream analyses. Moreover, leveraging more accurate long reads and incorporating more advanced genome assembly and pangenome construction methods into the PIGA workflow may help address this challenge.”

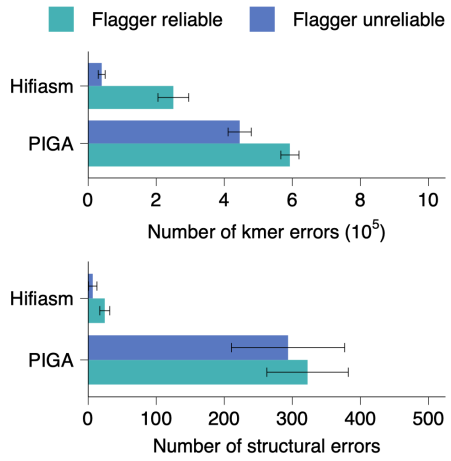

Figure R10. The number of k-mer errors and structural errors in Hifiasm and PIGA assemblies for the three benchmark samples, stratified by regional reliability as estimated by Flagger. Reliable regions correspond to haploid regions, while unreliable regions include collapsed, duplicated, and error-prone regions. Error bars represent 95% confidence intervals. (This figure is also included in our response to Reviewer #1.)

7. Figure 1f - What does "rank of epigenome profiles" mean?

Re: In Figure 1f, the "rank of epigenome profiles" refers to the ranking of predicted epigenomic tracks based on their mean AUROC values computed against blood ATAC-seq data.

We have clarified this description in the figure and figure legend to avoid ambiguity.
Updated figure: **Figure 1f** (corresponding to **Figure R31** below)

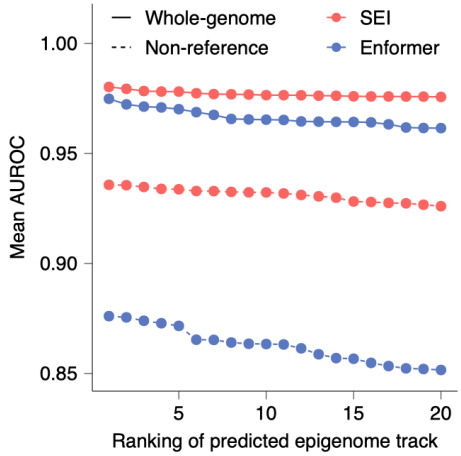

Figure R31. Epigenome track prediction performance of two deep-learning models, SEI and Enformer. For each predicted epigenome track, the mean AUROC was computed against blood ATAC-seq data separately for the whole genome and non-reference sequences across assemblies of three benchmark samples. Dots represent the top 20 predicted tracks, ranked by mean AUROC from highest to lowest.

8. Figure 2c - Why is the size of non-singleton non-reference sequence not directly related to the proportion relative to the pangenome? For example, I don't understand what it means that introns account for the largest proportion of annotated non-singleton non-reference sequence but a small proportion. Are you referring to the possible point that 70 Mb is small relative to the total amount of intronic sequence that does exist in the reference? If so, this should be clarified for readers.

Re: We apologize for any confusion caused by the original figure. The reviewer is correct that the "non-reference proportion" refers to the proportion of each genomic element relative to the total size of that element in the pangenome.

We have clarified this point in the revised figure legend.
Updated text (lines 1030-1033): "Columns indicate the size of genomic elements in non-singleton non-reference sequences, while dots indicate their proportion relative to the total size of the corresponding elements in the pangenome. The dashed line indicates the proportion of non-singleton non-reference sequences relative to the total size of all non-singleton sequences in the pangenome."

9. What is the size threshold that defines a structural variant? This should be noted at first mention of SVs and indels in the Results section and/or in the Methods.

Re: In our study, SVs are defined as genomic variants ≥ 50 bp, while small variants, including SNVs and indels, are defined as variants < 50 bp.

We have clarified this definition in the Results section.
Updated text (lines 208-209): "Using GRCh38 as the reference genome, we detected 35.4 million small variant sites (< 50 bp) and 110,530 SV sites (≥ 50 bp) in the 1KCP pangenome (Supplementary Table 11)."

10. Fig. 2g - What is a "raw" versus "merged" SV? This was never mentioned in the text.

Re: In Figure 2g, we have replaced the term “raw SV” with “unmerged SV” to clarify the distinction (**Figure R8** below). Here, unmerged SV refers to the SV callset before merging similar alleles. Accordingly, the text has been updated to use “unmerged SV” instead of “original SV” throughout the manuscript to maintain consistency.

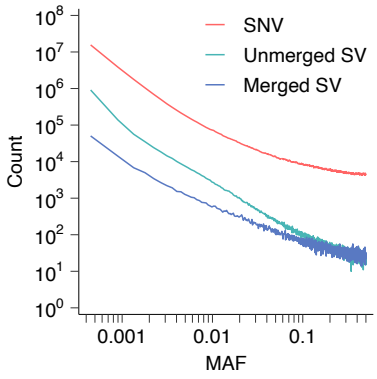

Figure R8. MAF spectrum of SNVs, unmerged SVs, and merged SVs. (This figure is also included in our response to Reviewer #1.)

11. Fig. 3 - The definition of a "rare" SV should be noted in the figure legend.

Re: We have clarified in the updated Figure 3a legend that rare SVs are defined as SVs with AF ≤ 0.01.
Updated text (lines 1044-1045): “Proportion of rare SVs (AF ≤ 0.01) across functional genomic regions. Error bars indicate 95% confidence intervals (Wilson score).”

12. Fig. 3c - I assume that the dark grey bars extend infinitely to the right. If so, I am wondering if there is a way to indicate this visually.

Re: We have updated the y-axis label to use “≥1,500” instead of “1,500” as the last tick mark (**Figure R16** below).
Updated figure: **Figure 3c** (corresponding to **Figure R16** below)

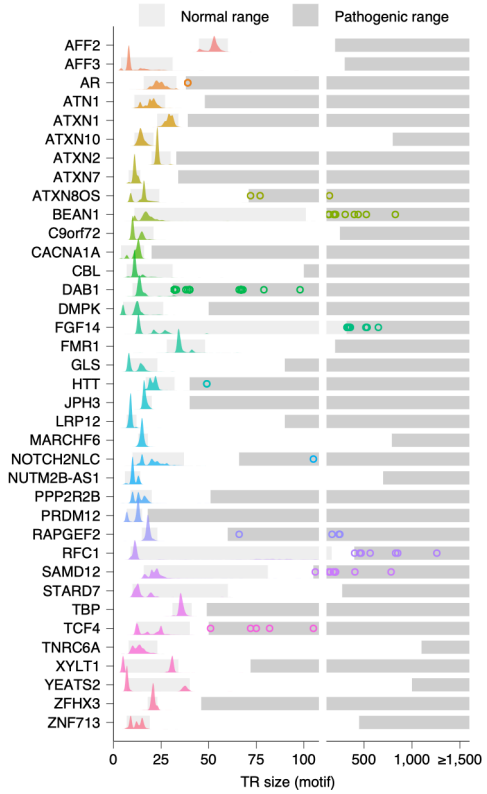

Figure R16. Size distribution of 37 disease-associated genic TR loci. Light grey boxes denote the normal TR size range observed in 99% of individuals from the 1KCP dataset, while dark grey boxes denote the pathogenic TR size range derived from the STRchive database. Circles indicate TR alleles with sizes exceeding the defined TR pathogenic threshold. (This figure is also included in our response to Reviewer #1.)

13. Fig. 3d - The barplot with height of 1 is a strange way to represent these data. Perhaps simply show the left histogram with both the outlier threshold and the one outlier sample annotated as vertical lines?

Re: We have revised Figure 3d to include a rug plot and to apply the same y-axis range to both the left and right subplots (**Figure R17** below).
Updated figure: **Figure 3d** (corresponding to **Figure R17** below)

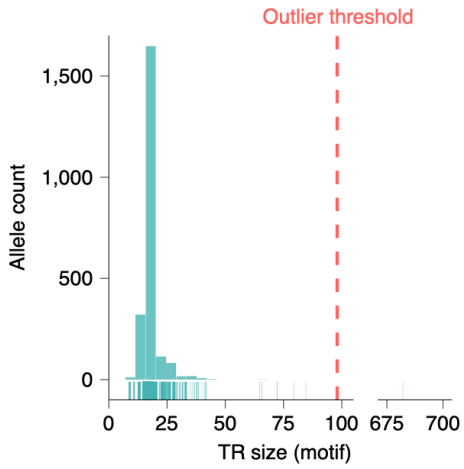

Figure R17. Size distribution of the *GIPC1* TR site. The vertical green ticks represent the sizes of normal TR alleles, and the red tick indicates the size of the outlier allele. The vertical red dashed line marks the outlier threshold determined by the DBSCAN algorithm. (This figure is also included in our response to Reviewer #1.)

14. Fig. 3j - Is there a pattern that this plot is meant to convey? I understand that the pangenome enables high resolution genotyping of HLA, which is impressive, and that this then enables measurement of LD between alleles at different loci, but I'm not sure if there is anything particularly interesting about the observed structure.

Re: The aim of Figure 3j is to illustrate the LD patterns of HLA alleles within the same LD group versus alleles not in the same LD group. To clarify this, we have updated the figure to use edge colors to distinguish LD relationships within an LD group from those between alleles not in the same LD group (**Figure R32** below).
Update figure: **Figure 3j** (corresponding to **Figure R32** below)

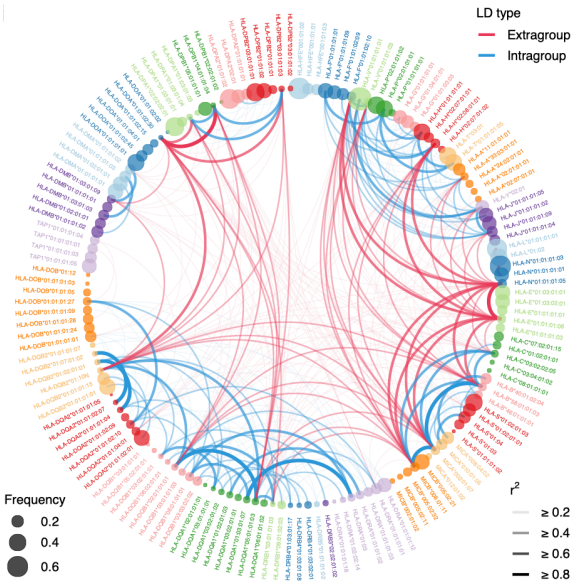

Figure R32. Fine-scale LD relationships between alleles of 36 HLA genes at four-field resolution. Only allele nodes with MAF > 0.05 and positive LD edges (r > 0) are shown.

Referee #4 (Remarks on code availability):

I looked through the code briefly. It is a set of numbered analysis scripts, which I think meets the minimum standard for transparency, but is not designed to generalize and be used in anyone else's study. For example, each script refers to numerous inputs and produces numerous that are not described in any form of documentation. The READMEs describe the basic roles of each analysis script.

Re: We have reorganized the analysis code into a cleaner structure and converted the bash scripts into a Snakemake pipeline. We now provide the required Conda environments along with instructions for running the pipeline. We also provided detailed descriptions of the required input data formats in the README and JSON configuration files.

Referee #5 (Remarks to the Author):

I co-reviewed this manuscript with one of the reviewers who provided the listed reports.

Re: We thank both Reviewers #4 and #5 for their constructive comments.

Responses to Reviewers' Comments

We thank the reviewers for their valuable comments on the previous and current versions of our manuscript. We have addressed the specific points raised in this second revision below (our responses are in blue) and highlighted the corresponding changes in the manuscript files in yellow.

Referee #1 (Remarks to the Author):

I appreciate the extensive revision done by the authors and the paper is now much improved. In particular, the QC of the call set is much better and more comprehensive information on the quality is provided to the readers.

Re: We thank the reviewer for the positive remarks.

Here are two minor comments for the authors to consider:

- Page 27: "Four-field alleles of HLA genes were called from assemblies using HiFiHLA". HiFiHLA is designed for HiFi data, not assemblies, is it a typo? Please also mention which data type HiFiHLA was run on (page 11, line 341). Also consider running Immuannot for detecting HLA alleles from assemblies.

Re: HiFiHLA indeed supports HLA allele calling directly from assembly contigs via the 'call-contigs' option, which provides functionality similar to Immuannot. We have revised the manuscript to clarify this.

Updated text (line 1118): "Four-field alleles of HLA genes were called from assemblies using HiFiHLA (https://github.com/PacificBiosciences/hifihla) v0.3.1 with the 'call-contigs' option."

- Not all tools used by PIGA are cited, for example, GLNexus, Longshot, Beagle4, merfin, etc. I suggest to double-check and make sure to include all appropriate citations.

Re: We have carefully rechecked the manuscript and ensured that all tools used in the PIGA workflow are now appropriately cited in the main text.

Updated text (lines 900-901): "We employed the PIGA workflow, an integrated framework incorporating multiple tools[1-33], to generate diploid genome assemblies using population-scale hybrid sequencing data."

References:

1. Rhie, A., Walenz, B. P., Koren, S. & Phillippy, A. M. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome biology* **21**, 1-27 (2020).
2. Garrison, E. *et al.* Building pangenome graphs. *Nature Methods*, 1-5 (2024).
3. Li, H., Feng, X. & Chu, C. The design and construction of reference pangenome graphs with minigraph. *Genome biology* **21**, 1-19 (2020).
4. Garrison, E. & Guarracino, A. Unbiased pangenome graphs. *Bioinformatics* **39**, btac743 (2023).
5. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv preprint arXiv:1303.3997* (2013).
6. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome research* **20**, 1297-1303 (2010).
7. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094-3100 (2018).
8. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841-842 (2010).
9. Danecek, P. *et al.* Twelve years of SAMtools and BCFtools. *Gigascience* **10**, giab008 (2021).
10. Poplin, R. *et al.* A universal SNP and small-indel variant caller using deep neural networks. *Nature biotechnology* **36**, 983-987 (2018).
11. Yun, T. *et al.* Accurate, scalable cohort variant calls using DeepVariant and GLnexus. *Bioinformatics* **36**, 5582-5589 (2020).
12. Martin, M. *et al.* WhatsHap: fast and accurate read-based phasing. *BioRxiv*, 085050 (2016).
13. Browning, S. R. & Browning, B. L. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *The American Journal of Human Genetics* **81**, 1084-1097 (2007).
14. Formenti, G. *et al.* Merfin: improved variant filtering, assembly evaluation and polishing via k-mer validation. *Nature methods* **19**, 696-704 (2022).
15. Delaneau, O., Zagury, J.-F., Robinson, M. R., Marchini, J. L. & Dermitzakis, E. T. Accurate, scalable and integrative haplotype estimation. *Nature communications* **10**, 5436 (2019).
16. Rautiainen, M. & Marschall, T. GraphAligner: rapid and versatile sequence-to-graph alignment. *Genome biology* **21**, 253 (2020).
17. Garrison, E. *et al.* Variation graph toolkit improves read mapping by representing genetic variation in the reference. *Nature biotechnology* **36**, 875-879 (2018).
18. Hickey, G. *et al.* Genotyping structural variants in pangenome graphs using the vg toolkit. *Genome biology* **21**, 1-17 (2020).
19. Shen, W., Sipos, B. & Zhao, L. SeqKit2: A Swiss army knife for sequence and alignment processing. *Imeta* **3**, e191 (2024).

20. Xiao, C.-L. *et al.* MECAT: fast mapping, error correction, and de novo assembly for single-molecule sequencing reads. *nature methods* **14**, 1072-1074 (2017).
21. Ruan, J. & Li, H. Fast and accurate long-read assembly with wtdbg2. *Nature methods* **17**, 155-158 (2020).
22. Vaser, R., Sović, I., Nagarajan, N. & Šikić, M. Fast and accurate de novo genome assembly from long uncorrected reads. *Genome research* **27**, 737-746 (2017).
23. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. journal* **17**, 10-12 (2011).
24. Luo, X., Kang, X. & Schönhuth, A. Phasebook: haplotype-aware de novo assembly of diploid genomes from long reads. *Genome biology* **22**, 1-26 (2021).
25. Smolka, M. *et al.* Detection of mosaic and population-level structural variants with Sniffles2. *Nature biotechnology* **42**, 1571-1580 (2024).
26. Kirsche, M. *et al.* Jasmine and Iris: population-scale structural variant comparison and analysis. *Nature methods* **20**, 408-417 (2023).
27. Garrison, E. & Marth, G. Haplotype-based variant detection from short-read sequencing. *arXiv preprint arXiv:1207.3907* (2012).
28. Sirén, J. *et al.* Personalized pangenome references. *Nature Methods* **21**, 2017-2023 (2024).
29. Ebler, J. *et al.* Pangenome-based genome inference allows efficient and accurate genotyping across a wide spectrum of variant classes. *Nature genetics* **54**, 518-525 (2022).
30. Holt, J. M. *et al.* HiPhase: jointly phasing small, structural, and tandem repeat variants from HiFi sequencing. *Bioinformatics* **40**, btae042 (2024).
31. Shafin, K. *et al.* Haplotype-aware variant calling with PEPPER-Margin-DeepVariant enables high accuracy in nanopore long-reads. *Nature Methods* **18**, 1322-1332 (2021).
32. Jiang, T. *et al.* Long-read-based human genomic structural variation detection with cuteSV. *Genome biology* **21**, 1-24 (2020).
33. Sirén, J. *et al.* Pangenomics enables genotyping of known structural variants in 5202 diverse genomes. *Science* **374**, abg8871 (2021).

Referee #1 (Remarks on code availability):

Better documentation for the Snakemake pipeline is still needed. Specifically, there is no test dataset; no unified installation of dependencies (e.g. conda package or environment). In the tutorial, config contents are not explained and there are no links to dependencies.

Re: Following the reviewer's suggestion, we have substantially improved the documentation of the Snakemake pipeline for the PIGA workflow. Specifically, we have:

1. Added a test dataset, which can be downloaded using the provided script (download.sh).
2. Provided a unified Conda environment (environment.yaml) that includes required dependencies.
3. Expanded the tutorial, which now includes detailed explanations of parameters in the configuration files.
4. Added explicit links to all dependencies.

Referee #2 (Remarks to the Author):

I co-reviewed this manuscript with one of the reviewers who provided the listed reports.

Re: We thank both Reviewers #1 and #2 for their constructive comments.

Referee #3 (Remarks to the Author):

The authors have put a lot of efforts in the revision and addressed/clarified several of my concerns. I really appreciate that. I still have additional follow-ups.

Re: We thank the reviewer for the positive remark. We provide detailed responses to the remaining follow-up comments below.

1) [previous point 1] The authors now describe PIGA in detail. This is a careful and sophisticated pipeline involving many tools for alignment, phasing, assembly and pangenome. I am very impressed. On the other hand, the PIGA pipeline is complicated and specialized for their own data. It will be challenging for others to use PIGA. Furthermore, the authors argued that their procedure is cheaper. This was true when they did sequencing with PacBio Sequel IIe, but might not be the case now with Revio+SPRQ (~16X cheaper than Sequal IIe). Given the power and logistics, most groups would choose straight long-read sequencing over the authors' hybrid procedure.

Re: We thank the reviewer for the positive remark and the additional suggestion. To improve the usability of the PIGA workflow for external users, we have modularized the workflow so that each major module can be executed independently. In addition, we now provide a unified Conda environment (environment.yaml) that includes required dependencies for user-friendly installation, add a test dataset for streamlined testing, and expand the tutorial with more detailed instructions. The core PIGA modules, including pangenome construction/simplification and final diploid assembly reconstruction, can directly

take Hifiasm draft assemblies as input, enabling users with different sequencing designs to leverage the joint-assembly framework.

Regarding cost considerations, we agree that recent advances in long-read sequencing have substantially reduced sequencing costs. However, long-read sequencing remains considerably more expensive than short-read sequencing, and many large population cohorts have already generated extensive short-read datasets. As a result, hybrid sequencing continues to be a practical strategy for population-scale studies, as adopted by the 1000 Genomes project, UK BioBank, and All of Us. Even in future large-scale long-read-only cohorts (e.g., 10,000 to 100,000+ individuals), trade-offs between per-sample coverage and overall sample size remain important considerations. In such contexts, PIGA's joint-assembly framework retains its value by improving assembly quality under lower-coverage scenarios.

We have added a discussion in the revised manuscript.
Updated text (lines 492-495): "This large-scale pangenome benefits from the joint-assembly framework of the PIGA workflow, which fully leverages modest-coverage hybrid sequencing data and can be applied to other large-scale projects employing either hybrid or long-read-only sequencing strategies with low to modest coverage."

2) [previous point 2 and 3] The authors want to emphasize on HLA haplotypes, but their results suggest noticeable errors. In response, Fig. R22 shows 1.7% error rate involving DRB3/4/5. These three genes are on distinct haplotypes. If there are switch errors around them, there will be other types of errors, making the haplotype structure questionable. Furthermore, the best short-read HLA genotypers like OptiType can achieve ~99% 2-field accuracy for HLA-A, a lot higher than the ~92% 3-field consistency shown in Fig. S32b. Also, HLA-G is one of the easier genes to type. I am surprised by its low consistency. I understand the authors use HLA-HD to type more genes, but if the inconsistency is caused by HLA-HD, they should change the short-read HLA typer. In contrast, resolving MHC with standard long-read assembly will be nearly error-free and more trustworthy as is shown in the HGSVC paper.

Re: We agree with the reviewer that switch errors may occur within the DRB3/4/5 region. However, these events affect only a small proportion of haplotypes (1.7%), leaving the remaining 98.3% with biologically consistent linkage structures.

As the reviewer notes, we originally used HLA-HD because it supports typing a broader set of HLA genes at three-field resolution. Following the reviewer's suggestion, we additionally benchmarked our

assembly-based typing results against OptiType for the Class I genes. The two-field concordance is 92.16% for HLA-A, 98.18% for HLA-B, and 97.53% for HLA-C (**Table R1** below, corresponding to **Supplementary Table 24** in the manuscript). Regarding HLA-G, we attempted validation but found no alternative short-read tool capable of typing it.

We acknowledge that HLA typing remains challenging at a small number of loci in a subset of samples. However, multiple lines of evidence indicate that the majority of haplotypes are reliably reconstructed and provide valuable population-level information:

1. The majority (19/22) of HLA genes show > 95% three-field concordance with HLA-HD typing results (**Figure R1a** below, corresponding to **Supplementary Figure 26b** in the manuscript).
2. Gene-wise LD block patterns are well defined and biologically consistent, which are unlikely to be driven by typing errors (**Figure R1b** below, corresponding to **Extended Data Figure 7c** in the manuscript).
3. HLA imputation performance demonstrates the overall haplotype quality of our panel. In the leave-one-out imputation evaluation, we observed an average four-field allele concordance of 0.96 and an average three-field allele concordance of 0.98, indicating generally high haplotype quality across HLA genes. In the comparison with the four-digit multi-ethnic HLA reference panel, our panel, although approximately 18× smaller, shows comparable imputation accuracy for classical HLA genes and enables the imputation of 2.7-fold more HLA alleles with high confidence ($r^2 > 0.8$), including non-classical genes and high-resolution alleles that are not captured by existing panels (**Figure R1c** below, corresponding to **Extended Data Figure 9e** in the manuscript).

We have revised the manuscript to incorporate the result and provide additional clarity.

Updated Table: **Supplementary Table 24** (corresponding to **Table R1** below)

Updated text (lines 452-453): “At the gene level, allele concordance reached 0.96 for four-field alleles and 0.98 for three-field alleles on average, indicating generally high haplotype quality across HLA genes.”

Table R1. Two-field allele concordance rate of HLA genes validated using Optitype

Gene	Concordance
HLA-A	92.16%
HLA-B	98.18%
HLA-C	97.53%

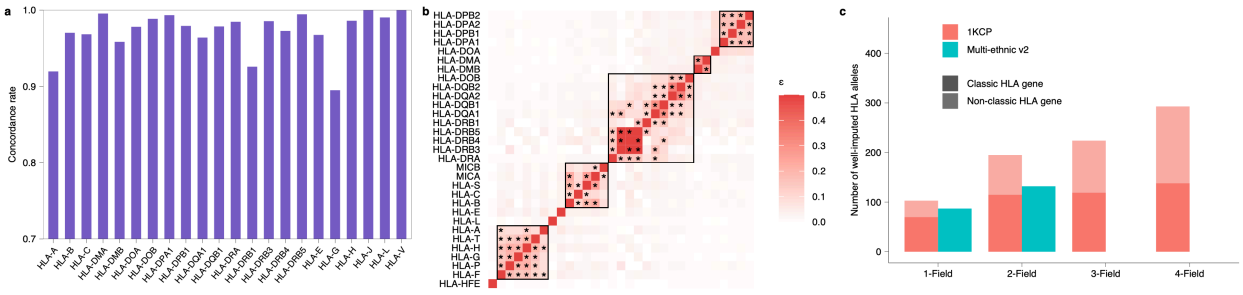

Figure R1. HLA gene validation. a, Three-field allele concordance rates of assembly-based typing results compared with short-read-based typing using HLA-HD on 22 shared HLA genes. **b**, Entropy-based LD measurement index (ε) between HLA genes, calculated based on four-field alleles. Black borders indicate defined LD groups. **c**, Numbers of well-imputed HLA alleles ($r^2 > 0.8$) from the 1KCP and Four-digit multi-ethnic HLA panels stratified by allele resolution levels and gene classes.

3) [previous point 4] The authors responded that their variant calling error rate is close to 1/60kb. This means there will be 3e9/60kb=50k false variants per individual. If errors are not shared, there will be 50k * 1116 = 50 million false variants in total, more than the number of called variants. I don't think the reality is this bad because many errors are shared, but it would be good to validate the calls with orthogonal data (most QC strategies in the manuscript, albeit sophisticated, still rely on the same datasets and are thus not orthogonal). Targeted sequencing of randomly selected variants will be nice. The authors may also compare their calls to gnomAD SNPs. Ideally, most common gnomAD SNPs (say MAF>10% among Eastern Asians) should be captured by the PIGA SNPs; most PIGA SNPs, except in difficult regions, should be found by gnomAD given the larger sample size of gnomAD. Large deviation from this expectation would indicate potential errors.

Re: We agree with the reviewer that, in a large-scale dataset, a substantial number of errors may accumulate. Therefore, we implemented kmer-based pangenome QC and region-level QC steps to reduce potential errors and improve the reliability of our callset, which reduced the number of small-variant alleles from 54.9M to 45.7M. We then evaluated SNV quality using multiple metrics:

1. Ts/Tv Ratio. After QC, the Ts/Tv ratio increased from 1.83 to 1.96, approaching the expected value of ~2 and indicating effective removal of spurious calls.
2. Allele frequency spectrum. The post-QC allele-frequency distribution closely matches that of the 1KGP EAS population and shows no evidence of singleton inflation, which would be expected under high false-positive rates (**Figure R2a** below).

3. Benchmarking FDR. Using the Hifiasm assemblies of three benchmark samples as truth sets, the estimated FDR of post-QC SNVs is 1.77% (**Figure R2b** below, corresponding to **Supplementary Figure 22b** in the manuscript).

Together, these results show that the final SNV set has a low error rate and is not dominated by false positives.

Following the reviewer's suggestion, we then compared our SNVs with gnomAD v3. Given the large size of the full gnomAD callset, we performed this comparison on chromosome 20 as a representative region. Among gnomAD SNVs with AF > 0.1 in the EAS population, 94% were captured in our callset, demonstrating high sensitivity for common variants. For SNVs in our callset, the overall validation rate against gnomAD is 71.87%, and it increases with higher MAF (**Figure R2c** below). Considering the low estimated FDR above and the limited number of EAS samples in gnomAD (N≈5,200), we believe that a large proportion of SNVs absent from gnomAD represent true variants rather than errors.

We have revised the manuscript to provide additional clarity.

Updated text (lines 1037-1038): "The quality control steps effectively reduced the error rate of our callset (Supplementary Figure 22)."

Updated figure: **Supplementary Figure 22b** (corresponding to **Figure R2b** below)

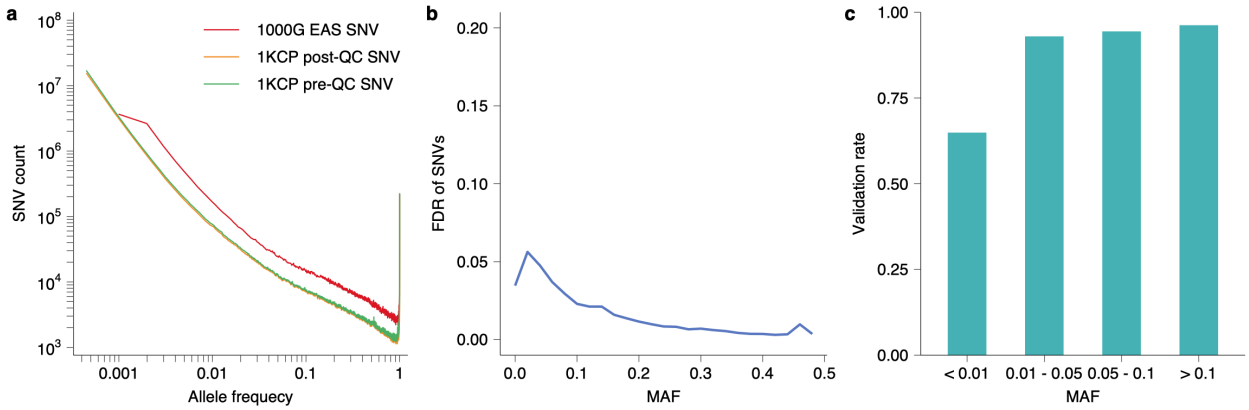

Figure R2. SNV validation. a, Allele frequency spectrum of SNV callsets. **b**, FDR of SNVs across 25 MAF bins, estimated using three benchmark samples. **c**, Validation rate of SNVs against gnomAD v3 across MAF categories.

4) [new comment] The small variant VCF on the 1KCP website is incomplete. It only contains the first

~15Mb of chr1.

Re: We thank the reviewer for pointing this out. We have now updated the files on the website, and the complete VCF files are available.

Referee #3 (Remarks on code availability):
I had a look the github repo in the first round of the review. This repo is offline now. Please bring it back.

Re: We have substantially updated the GitHub repositories, and they are now online again.

Referee #4 (Remarks to the Author):
Overall, the revised manuscript is greatly improved. The authors have rigorously and comprehensively addressed my comments, as well as those of the other reviewers. In response to my comments, the new eQTL fine-mapping and enrichment analyses are especially nice, as are the new analyses demonstrating the performance improvements for imputation, stratified by variant type.

Re: We thank the reviewer for the positive feedback. We appreciate the constructive comments, which significantly improved the quality of our manuscript.

I have a few remaining minor comments and questions about the revised manuscript:
1. Line 62: Suggest revising "which generally exhibit higher pathogenicity than common variants" to "are more likely to be pathogenic than common variants" or "are enriched for pathogenic effects compared to common variants."

Re: Following the reviewer's suggestion, we have revised the sentence in the manuscript.
Updated text (lines 62-63): "First, rare variants, which are more likely to be pathogenic than common variants and play critical roles in inherited diseases, are underrepresented in small-scale pangenomes."

2. Line 134: Is "read-based" the appropriate term here? Since both methods use sequencing reads, perhaps "conventional single-reference alignment-based" would be more precise, as your method uses a pangenome-guided assembly approach.

Re: To more precisely distinguish our approach from methods that directly rely on read alignment results, we have revised the term to "read-alignment-based."

Updated text (lines 135-136): “Nonetheless, PIGA assemblies outperformed read-alignment-based methods in variant calling across multiple variant classes.”

3. Paragraph on line 193: What is meant by “genomic element”? Is the definition broader than putative functional elements? Some clarification would be helpful.

Re: We have clarified the definition of "genomic elements" in this paragraph to explicitly include repeats, genes, and epigenomic elements, consistent with our earlier definition in the assembly annotation section.
Updated text (lines 195-197): “To address this, we developed a path-guided pangenome annotation method to identify genomic elements, including repeats, genes, and epigenomic elements, within the 1KCP pangenome.”

4. Line 270: Suggest revising “strong purifying selection” to “signatures of purifying selection” or “evidence of purifying selection.”

Re: We have rephrased “strong purifying selection” as “evidence of purifying selection.”
Updated text (lines 269-272): “These exon-overlapping SVs, capable of altering gene transcripts, exhibit evidence of purifying selection, as reflected by their higher proportion of rare SVs (AF ≤ 0.01) compared with intergenic SVs (58.8%) and SVs in other genomic regions.”

5. Paragraph starting on line 336 (and specifically the last sentence starting on line 356): Most of the revisions to the section on LD are appreciated, but the last sentence still reads as potentially circular. Could “LD groups” be renamed to “LD blocks”? Then you could state that, beyond these blocks defined by strong LD, there is modest LD that extends over a longer range, and report the extent in kilobases.

Re: Following the reviewer’s suggestion, we have replaced “LD groups” with “LD blocks” and revised the corresponding sentences.
Updated text (lines 354-358): “Based on gene-wise LD, we identified five distinct LD blocks for HLA genes ($\varepsilon > 0.10$), including 2 HLA class I blocks and 3 HLA class II blocks. Beyond these gene-wise LD blocks, allele-wise LD (MAF > 0.01) revealed modest long-range correlations ($r^2 = 0.4–0.8$) that extend across block boundaries, indicating the existence of long-range HLA haplotypes in the 1KCP cohort (Figure 3j and Supplementary Table 25).”

6. Line 368: Can the number of samples with RNA-seq data be reported?

Re: The number of RNA-seq samples (n = 1,101) used for eQTL analysis has been added to the revised manuscript.
Updated text (lines 367-370): "To demonstrate the value of pan-variant association studies, we used the expression levels of 21,372 genes (13,860 protein-coding genes and 7,512 noncoding genes) quantified by RNA-seq in the 1KCP cohort (n = 1,101) as example phenotypes and investigated their genetic regulation by different variant types."

7. Line 412: What is meant by "unit"? Can you describe this in terms of allelic fold change? For example, see https://www.nature.com/articles/s41467-024-44710-8

Re: We thank the reviewer for the suggestion. We have clarified in the text that "unit" refers to the standard deviations of the normalized expression. We have maintained this metric rather than allelic fold change because our analysis relies on a linear model, whereas fold change interpretation typically assumes an exponential relationship between TR copy number and expression.
Updated text (lines 410-412): "Specifically, each additional TR copy was associated with a 0.126 standard deviation increase in normalized MAD1L1 expression, while an additional motif copy was associated with a 0.245 standard deviation increase (Figure 4d)."

8. Figure S2B (middle left): "Phred" is misspelled as "Phread" in the y-axis label.

Re: The plot has been corrected in the revised manuscript.
Updated figure: **Supplementary Figure 2b**

9. Figure S37: "1KCP Hifiasm + PIGA using the whole 1KCP assembly panel (N=1,116) captures 4.3-fold more rare SVs and detects 3.1-fold more TR alleles than a smaller panel using only Hifiasm assemblies (N=55)." Is this a fair comparison given that the whole 1KCP panel contains so many more samples than the Hifiasm assemblies?

Re: The reviewer is correct that the sample sizes differ substantially. However, this comparison is intentional and central to our study's premise. It is designed to demonstrate the added value of a population-scale assembly panel (enabled by the PIGA workflow) relative to a smaller Hifiasm-only panel (n = 55), which is comparable in size to previous pangenomes. We have revised the manuscript to clarify.

Updated text (lines 495-497): "This large-scale pangenome demonstrates its value by greatly enhancing the representation of genetic variants, particularly rare and low-frequency variants."

10. Figure 1C: I find Figure 1C a bit confusing. It seems to show a large percentage of k-mer and structural errors with PIGA. How is this consistent with the statement that "PIGA assemblies provided more accurate sequences for SVs and TRs compared to read-based approaches (Figure S10D and Figure S27A)"? Is Figure 1C showing results before PIGA filtering?

Re: Figure 1C shows the assembly errors after PIGA's ML-based filtering but before the downstream variant callset QC. Although PIGA assemblies exhibit more errors than Hifiasm assemblies, the error rate remains low, and our subsequent QC steps further reduced residual errors in the final callset. Importantly, the mean number of structural errors detected by Inspector (616 per sample) is consistent with the mean number of false-positive SVs estimated in benchmarking using Truvari (488 per sample), corresponding to a high precision of 97.8%. In contrast, SV calls obtained directly from read-alignment-based methods, such as Sniffles, exhibit substantially lower precision (82.17%) with many more false-positive SVs. These results support our conclusion that PIGA assemblies provide more accurate variant calls than read-alignment-based approaches. We have clarified this point in the manuscript.
Updated text (lines 128-129): "The number of false-positive SVs closely matches the structural error counts detected by Inspector, further supporting the high accuracy of PIGA SV calls."

11. Figure S6A: The y-axis shows rank of each reference type across 6 alignment metrics. Could this plot be made more informative by including the actual metric values, in addition to the rank?

Re: We have added a plot that shows the actual alignment-metric values (**Figure R3** below).
Updated figure: **Supplementary Figure 6a** (corresponding to **Figure R3** below)

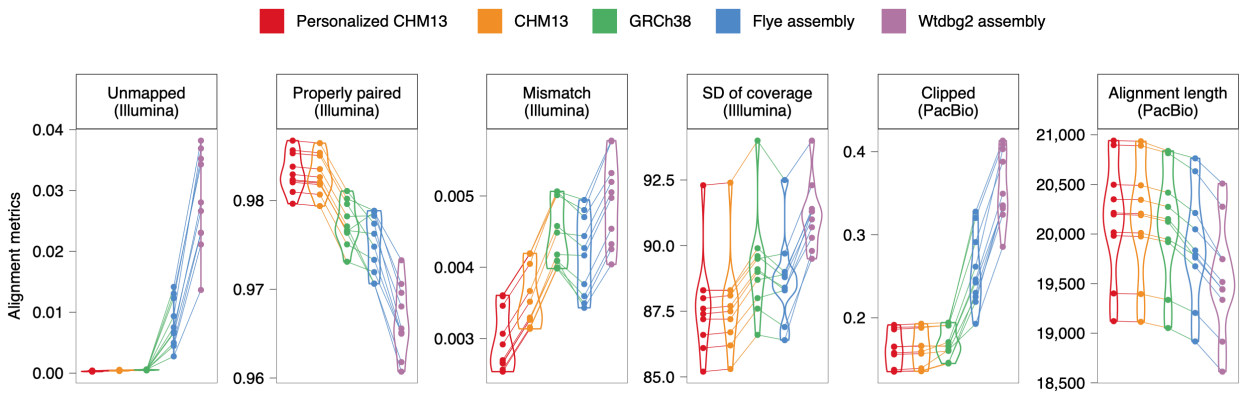

Figure R3. Comparison of alignment metrics across five reference types. Metrics include the proportion of unmapped Illumina reads, proportion of properly paired Illumina reads, mismatch rate of Illumina reads, standard deviation of Illumina read coverage, proportion of clipped PacBio reads, and average alignment length of PacBio reads

Referee #4 (Remarks on code availability):

I was unable to run the PIGA pipeline in Snakemake after downloading from GitHub. Specific issues I encountered:

1. The PIGA tutorial workflow execution should be: snakemake -s Snakefile --cores 64 --configfile config/config.yaml --configfile config/tools.yaml
2. The provided config.yaml file returns an error in Snakemake: "snakemake.exceptions.WorkflowError: Config file is not valid JSON or YAML. In case of YAML, make sure to not mix whitespace and tab indentation." The trailing comma should be removed from this line: topmed_east_asian: "test/variant_panel/{chr}/TOPMed_WGS_freeze.8.east_asian.merge.{chr}.filter.snps.liftover_chm13.shapeit4.vcf.gz",
3. After correcting config.yaml, Snakemake returns a KeyError in 'tools' at Snakefile line 14. The tools.yaml file appears to contain a sample identifier and not a list of tools.

Re: We apologize for the errors in the documentation. To address these issues, we have substantially improved the PIGA workflow. The updated version includes: error fixes in the tutorial and configuration files, a dataset for testing, and a unified Conda environment to streamline installation.

Referee #5 (Remarks to the Author):

I co-reviewed this manuscript with one of the reviewers who provided the listed reports.

Re: We thank both Reviewers #4 and #5 for their constructive comments.

Responses to Reviewers' Comments

We thank the reviewers for their constructive comments, which have helped us significantly improve our manuscript. We have addressed all the reviewers' comments point-by-point below (our responses are in blue).

Referee #1 (Remarks to the Author):

We thank the authors for their diligent work. There is one issue left, however: I was not able to download the test data the authors provide with their PIGA software because yanglab.westlake.edu.cn was not reachable, neither from my university network nor from home. Either the server is down or there is some geoblocking in place that does not allow me to access this server from my country. In any case, I suggest the authors deposit the test data in a DOI-minting repository such as Zenodo. In this way the data remain accessible even when URLs change (the download script depends on many different URLs, so sooner or later this will break).

Re: We sincerely apologize for the inconvenience caused by the temporary downtime of our laboratory server during the review process. We attempted to upload the dataset to Zenodo but were restricted by its file size limits (50 GB), as our test dataset exceeds this capacity. To ensure stable and long-term accessibility, we have compressed the test dataset into a single archive and hosted it on a stable cloud storage service. The download script has been updated accordingly.

Referee #1 (Remarks on code availability):

We thank the authors for adding a script to download test data and the YAML environment. The download script does not work for this reviewer, however (see comments above).

Re: We thank the reviewer for this feedback. In response, we have consolidated all test data into a single compressed archive and hosted it on a stable cloud storage service. We have revised the download script accordingly. These changes improve the robustness of data sharing and simplify the testing procedure.

Referee #2 (Remarks to the Author):

I co-reviewed this manuscript with one of the reviewers who provided the listed reports.

Re: We thank both Reviewers #1 and #2 for their constructive comments.

Referee #3 (Remarks to the Author):

I thank the authors for the revision. I still have two minor comments:

Re: We appreciate the reviewer's acknowledgement of our revision efforts. We have addressed the two remaining comments below.

1) [previous comment 2] The accuracy of mainstream HLA genotypers approaches ~99% on HLA-A. The authors can only reach 92%. At this high error rate, LD-based analysis involving HLA-A (Line 349–358) will be problematic. Generally, the authors probably don't have enough power to accurately generate Fig. 3j due to phase switches with HiFi only. Population phasing may not be reliable either due to MHC diversity. The right analysis should rely on long-range phased assemblies like those from HPRC and HGSVC. It is probably better to leave out this part in my opinion.

Re: We thank the reviewer for the suggestion. To ensure that our LD analyses are robust to genotyping errors, we have excluded three HLA genes (including HLA-A) that showed concordance below 95% with HLA-HD from all LD analyses.
Updated text: "HLA genes with concordance < 95% were excluded from LD analyses."

We acknowledge the reviewer's point regarding the limitations in long-range LD inference. While our LD analyses are based on unphased, genotype-based LD estimation (via PLINK2) and do not rely on individual-level haplotypes, we agree that these results should be interpreted cautiously. Therefore, we have removed related statements and Figure 3j from the manuscript text to avoid overinterpretation.

2) [previous comment 3] The authors now provide SNP calls. On chr10, I counted 1275k 1KCP SNPs matching gnomAD v4.1 SNPs at both positions and alleles. The transition/transversion ratio (ts/tv) of these SNPs is 2.35. 437k 1KCP SNPs are not found in gnomAD. Their ts/tv is only 1.36. This suggests ~30% false discovery rate among SNPs not found in gnomAD (formula: $3(x-y)/[(2x-1)(y+1)]$, where x is the true ts/tv, y is the observed ts/tv and errors have a ts/tv of 0.5). The overall chr10 error rate is thus ~7.6% (=30%*437/(437+1275)) – that is 3.5 million wrong SNPs genome wide. Although such an estimate is very crude and makes multiple assumptions, the real error rate is probably around this level. I

am okay with the error rate, but I think the authors should show an error estimate in the main text to let readers know. It is worth noting that the expected ts/tv increases with sample size due to CpG mutations. For example, the ts/tv of singleton SNPs overlapping with gnomAD reaches 2.84. ts/tv=1.96 does not imply high SNP accuracy across 1,116 samples.

Re: We thank the reviewer for raising this point and providing the calculation formula. Following this suggestion, we reassessed SNV accuracy using genome-wide SNVs against gnomAD v4.1.

In the 1KCP dataset, 25.4 million SNVs were found in gnomAD with a Ti/Tv ratio of 2.27. And 9.1 million SNPs were absent in gnomAD with a Ti/Tv ratio of 1.34. Using the reviewer's formula, this corresponds to an estimated FDR of 33.7% among novel SNVs and an overall FDR of 8.9%.

To contextualize this estimate, we performed the same analysis on the ChinaMAP dataset (Cao et al., 2020), which was generated using high-coverage (~40×) short-read sequencing of 10,588 individuals. In ChinaMAP, 63.8 million SNVs were found in gnomAD with a Ti/Tv ratio of 2.99. And 72.9 million SNPs were absent in gnomAD with a Ti/Tv ratio of 1.41. This corresponds to an estimated FDR of 39.5% among novel SNVs and an overall FDR of 21.1%. This comparison suggests that lower Ti/Tv ratios in novel SNVs reflect a general characteristic of large population-scale variant discovery.

Following the reviewer's suggestion, we have now included this Ti/Tv-based FDR estimation in the main text.

Updated text: "Using GRCh38 as the reference genome, we detected 35.4 million small variant sites (< 50 bp; see Methods for an estimated SNV FDR of ~8.9% using the Ti/Tv ratio) and 110,530 SV sites (≥ 50 bp) in the 1KCP pangenome (Supplementary Table 11)."

Updated text: "We employed a transition-to-transversion (Ti/Tv) ratio-based approach to estimate the FDR of the SNV callset. Using the gnomAD v4.1 database as a reference, SNVs were classified as known (present in gnomAD) or novel (absent from gnomAD). Let x denote the Ti/Tv ratio of known SNVs, representing the expected biological signal, and y denote the Ti/Tv ratio of novel SNVs, reflecting a mixture of true rare variants and random errors (assuming a random error Ti/Tv ratio of 0.5). The FDR among novel SNVs was calculated as $FDR_{novel} = 3(x - y) / [(2x - 1)(y + 1)]$. In the 1KCP callset, we identified 25.4 million known SNVs (x = 2.27) and 9.1 million novel SNVs (y = 1.34). Substituting these values yields an estimated FDR of 33.7% for novel SNVs. The overall genome-wide FDR was then calculated by weighting the novel FDR by the proportion of novel SNVs in the total SNV set, resulting in an estimated overall FDR of 8.9%."

Referee #3 (Remarks on code availability):

The github repo looks well organized, though I have not run the pipeline.

Re: We thank the reviewer for the positive assessment of the repository organization.

Referee #4 (Remarks to the Author):

Thank you to the authors for thoroughly addressing most of my previous comments and improving the generalizability of their software workflow. However, please see below for some remaining issues with the latter.

Re: We thank the reviewer for the time and contribution to the evaluation of this manuscript.

Referee #4 (Remarks on code availability):

While the workflow is now modular at the level of execution, all software dependencies are currently bundled into a single, large Conda environment. As a result, users must install the full suite of software even if they wish to run only a subset of the pipeline.

Although I was able to successfully download the provided test dataset, I was unfortunately unable to create and activate the PIGA Conda environment in order to test the workflow. Specifically, the GitHub repository URL listed in the README appears to be incorrect. Based on my attempts, the correct repository is https://github.com/hahahafeifeifei/PIGA rather than https://github.com/JianYang-Lab/PIGA. In addition, attempts to create the Conda environment either stalled or failed during dependency resolution. Examination of the log file suggests that these issues may stem from incompatible versions of Python, WhatsHap, and Snakemake.

These issues may be related to the use of a single Conda environment, with tools invoked directly within Snakemake rule shells. To address these issues, the authors may wish to consider Snakemake's native support for rule-specific Conda environments using the conda: directive within individual rules. The Snakemake documentation on automatic deployment of software dependencies provides a helpful

overview of this approach (https://snakemake.readthedocs.io/en/v8.0.0/tutorial/additional_features.html).

Re: We thank the reviewer for the suggestion. While we acknowledge the benefits of rule-specific environments, we opted to maintain a single unified Conda environment to facilitate deployment and ensure compatibility with computational clusters that have restricted internet access. To address the installation issues, we have pinned the Snakemake version to v7 and resolved the dependency conflicts (including those with Python and WhatsHap). Although the unified environment is large, it now can be efficiently installed using Mamba.

We have also moved the repository to the original URL (https://github.com/JianYang-Lab/PIGA). Furthermore, we have updated the dataset and compressed it into a single archive. We have verified that the updated workflow now runs successfully across multiple server environments.

Referee #5 (Remarks to the Author):

I co-reviewed this manuscript with one of the reviewers who provided the listed reports.

Re: We thank both Reviewers #4 and #5 for their constructive comments.