nature portfolio

https://doi.org/10.1038/s41586-026-10315-y

Supplementary information

The 1000 Chinese Pangenome empowers medical and population genetics

In the format provided by the authors and unedited

The 1000 Chinese Pangenome empowers medical and population genetics

Wang *et al.*

Supplementary Notes

Supplementary Note 1: Pangenome-Informed Genome Assembly (PIGA) workflow

Overview of PIGA workflow

The PIGA workflow begins with population-level single-nucleotide variant (SNV) detection and haplotype phasing (Extended Data Figures 2a and 2b). For each individual, we generate a personalized reference by incorporating homozygous variants identified from an external pangenome into the original reference. Phased SNV haplotypes are lifted over from the original reference to the personalized reference (Extended Data Figure 2c). In regions where this lift-over fails, additional SNVs are directly detected and phased within the personalized reference to complete the SNV haplotypes. Leveraging these phased heterozygous SNVs, we assign long reads to haplotypes and assemble them into draft diploid assemblies (Extended Data Figure 2d). To unify sequence information across all draft assemblies, we construct a graph-based pangenome and simplify its structurally complex or error-prone regions (Extended Data Figure 2e). The final diploid assemblies are reconstructed for each individual by inferring the diploid path from the refined graph pangenome (Extended Data Figure 2f).

SNV detection

To generate a comprehensive SNV callset, we developed a population-scale SNV detection pipeline that integrates both PacBio long-read and Illumina short-read sequencing data. SNVs were independently identified from each data type and then combined into a merged callset.

For the Illumina short-read data, reads were aligned to the CHM13 reference genome using BWA-MEM[62] v0.7.17. Subsequently, we performed joint SNV calling across all samples following the GATK[63] v4.2.6.1 best-practices workflow. For the PacBio long-read data, the reads were aligned to the CHM13 reference using pbmm2 v1.9.0 (https://github.com/PacificBiosciences/pbmm2). Candidate SNV sites were first detected using HiFi reads with Deepvariant[71] v1.4.0 and GLnexus[72] v1.4.1. To enhance accuracy, we then genotyped these sites using all PacBio reads (both HiFi and non-HiFi reads) with WhatsHap[73] v1.4. The resulting genotypes were further refined based on linkage disequilibrium using Beagle[74] v4.0.

To consolidate the two callsets, we compared SNVs from both sources. Unique SNVs from either callset were directly incorporated into the merged callset. For SNVs present in both callsets, we used the ratio of Hardy–Weinberg equilibrium (HWE) p-values to guide genotype selection. Specifically, genotypes from the PacBio callset were retained if their p-value was more than 1,000 times greater than that of the Illumina callset; otherwise, the Illumina genotype was used. Finally, the entire merged callset underwent an additional k-mer-based filtering with Merfin[75] v1.1 to ensure high accuracy.

We benchmarked the SNV calling pipeline using ten 1KCP samples with both modest and high-coverage data, which were also used for evaluating downstream analyses. The performance of the PacBio callset improved progressively through the pipeline stages, with the most significant gain observed after genotyping with WhatsHap (average F1 score increased from 0.77 to 0.90), highlighting the value of leveraging all PacBio reads beyond HiFi reads alone (Supplementary Figure 2a). Ultimately, our integrative approach yielded a merged callset (average F1 score: 0.987) that outperformed the Illumina-only callset (average F1 score: 0.977). When stratifying SNVs into easy and difficult regions based on GIAB[136] annotation, the merged callset showed higher recall than the Illumina callset in both easy and difficult regions without compromising precision (Supplementary Figure 2b). This improvement in recall was particularly pronounced in difficult regions, underscoring the strength of long reads in resolving repetitive and complex genomic regions.

SNV haplotype phasing

We then phased the detected SNVs into haplotypes using SHAPEIT4[76] v4.2.2, incorporating prior haplotype information from long-read phasing blocks generated with WhatsHap. For comparison, phasing was also performed without using long-read data. Phasing accuracy was assessed against ground-truth haplotypes derived from the Hifiasm assemblies of benchmark samples.

The integration of long-read information markedly improved phasing accuracy, reducing the average switch error rate from 0.37% to 0.20%. Notably, the improvement was especially pronounced for ultra-rare variants (Minor Allele Frequency < 0.1%), where the average switch error rate dropped from 17.97% to 5.51% (Supplementary Figure 3), underscoring the value of long-read data for accurately phasing rare variants.

Personalized reference generation

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.

For this analysis, we used the Minigraph-Cactus[111] pangenome, which integrates CHM13, GRCh38, and 45 Hifiasm diploid assemblies from the 1KCP cohort (excluding 10 benchmark samples). Long reads and short reads were aligned separately to the pangenome using GraphAligner[77] v1.0.13 and VG Giraffe[78,96] v1.59.0, respectively. Variant genotypes were called using VG call[79], and homozygous variants were applied to CHM13 using BCFtools[70] consensus v1.12, yielding a personalized CHM13 reference for each 1KCP modest-coverage sample.

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[137] squashed assembly, and wtdbg2[82] 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 Figures 4a and 4b). 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 4c).

Reference-guided draft diploid genome assembly

Based on the personalized reference and partitioned long reads, we developed a reference-guided genome assembly pipeline to generate draft diploid assemblies. This pipeline is designed to mitigate the incompleteness and assembly collapse issues that often arise in *de novo* assembly under modest- or low-coverage conditions.

Our pipeline proceeds as follows: First, all long and short reads were aligned to the personalized reference using minimap2[66] v2.24 and BWA-MEM, respectively. Genomic intervals with sufficient haplotype-

partitioned long-read coverage (>3x) were then identified and divided into subintervals of up to 50 kb. For each subinterval, we extracted the corresponding long and short reads using SAMtools[70] v1.15.1. PacBio subreads were trimmed using MECAT2[70], and only the median-length subread per ZMW was retained for local assembly using wtdbg2, with parameters optimized for low-coverage conditions (-S 1 --edge-min 1 --rescue-low-cov-edges --aln-dovetail -1). The resulting contigs were polished with two rounds of Racon[83] v1.5.0 and one round of gcpp v2.0.2 (https://github.com/PacificBiosciences/gcpp) Arrow, followed by quality filtering to retain only contigs with a Phred quality score > 15. These subinterval contigs were then merged into interval contigs using the super-read module of Phasebook[85]. We corrected structural errors in the interval contigs using long reads with Sniffles2[86] v2.0.7 and Jasmine[87] v1.1.5, and refined base errors using short reads with freebayes[88] v1.3.7. To further reduce switch errors, we aligned the contigs back to the personalized reference using minimap2 and compared their haplotypes against the original SNV haplotypes. Contig segments falling within 50-kb windows with more than two switch errors were clipped. Finally, remaining PacBio adaptor sequences within contigs were detected using minimap2 and masked using BEDtools[68].

We first validated this assembly pipeline on the 10 benchmark samples, yielding draft assemblies with a mean size of 2.07 Gbp (Supplementary Figure 5a). In contrast, standard de novo assemblies generated directly by wtdbg2 using partitioned long reads showed significantly lower completeness with a mean size of 1.70 Gbp, confirming the advantage of our reference-guided approach in modest-coverage settings. Base-level accuracy, evaluated using Merqury[20], showed that all draft assemblies achieved quality values (QV) above 30, highlighting the effectiveness of iterative polishing with both long and short reads (Supplementary Figure 5b).

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 5c). Inspection of the remaining underrepresented regions revealed that 84% are located within segmental duplications or centromeric regions, which remain challenging to resolve using noisy long reads due to their highly repetitive structure. These results demonstrate that while gaps persist in complex regions, these population-scale draft diploid assemblies provide a wealth of haplotype-resolved sequences for most of the genome, which provide fundamental information for generating final diploid assemblies.

Graph construction and simplification

To enable the sharing of sequence information across population-scale genome assemblies, we employed a graph pangenome as a unified framework. Given the large number of draft assemblies with relatively low accuracy, we developed a parallelized pipeline for efficient graph construction and simplification.

The pipeline begins by constructing an SV pangenome using Minigraph[89] v0.21-r606, followed by realigning input contigs to this graph. To enable parallel processing, the genome-wide graph is partitioned into smaller subgraphs, and all sequences and alignments are segmented accordingly. For each subgraph, a base-level pangenome is induced by seqwish[90] v0.7.9, then refined through two rounds of smoothxg[91] v0.7.9 with partial order alignment (POA) parameters for 5% divergence (-p 1,4,6,2,26,1) and target lengths of 1400 and 2200. The refined pangenome was then normalized with GFAffix (https://github.com/marschall-lab/GFAffix) v0.1.4 to collapse shared affixes.

To simplify graph complexity arising from highly repetitive regions, we clipped contiguous segments exceeding 100 kb from pangenome assembly paths that were either masked by dna-brnn (https://github.com/lh3/dna-nn) or left unaligned by the Minigraph paths. 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 6), 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). The initial SV pangenome had a size of 3.53 Gb and was split into 637 subgraphs, each spanning an approximately 5-Mb region of the CHM13 reference. After base-level induction and normalization, the graph size expanded to 22.85 Gb, underscoring the necessity of further simplification (Supplementary Figure 7a).

To train the filtering model, we labeled variant and error nodes/edges using the 7 overlapping samples with both Hifiasm and PIGA draft assemblies. The trained MLP model demonstrated high accuracy, achieving an Area Under the ROC Curve (AUROC) of 0.923 for nodes and 0.931 for edges, outperforming simpler filtering methods based on allele count alone (Supplementary Figure 7b). Stratifying nodes/edges by allele frequency further demonstrated the superior power of the machine learning model even for distinguishing singletons (Supplementary Figure 7c). Using a conservative threshold corresponding to a true positive rate of 0.9, we retained 90% of true variant nodes/edges, while filtering out 78.2% of error nodes and 80.7% of error edges, ultimately reducing the graph size from 15.73 Gb to 4.15 Gb. The final pangenome graph had a total size of 3.92 Gb, supporting efficient downstream analysis.

Final diploid assembly generation

Leveraging the constructed pangenome as prior knowledge, we developed an integrative pipeline to reconstruct a final diploid assembly for each individual. This pipeline infers an individual's diploid paths through the pangenome by incorporating multiple layers of information, including short reads, long reads, draft assemblies, and SNV haplotypes.

For each individual, we first created a personalized pangenome by sampling 17 candidate haplotype paths from the full pangenome based on read k-mer similarity using VG haplotypes[92], and augmenting it with the individual's draft assembly and SNV haplotype paths. The resulting personalized pangenome was then normalized using VG unchop and GFAffix.

Using the highest-scoring sampled path as the haploid reference, we identified candidate genetic variant sites from the pangenome using VG deconstruct. Genotypes at these sites were then resolved by integrating data from three sources: PanGenie[97] v3.1.0, SNV haplotypes, and draft assembly haplotypes. Genotypes with a PanGenie quality score > 100 were accepted directly. For lower-quality PanGenie calls (GQ < 100), we used SNV haplotypes if available, followed by draft assembly haplotypes; otherwise, we defaulted to the PanGenie genotype. These variants were then phased into consensus haplotypes[138] by combining multiple sources of phasing information, including SNV haplotypes, draft assembly haplotypes, long-read alignments (HiPhase[93] v1.4.0 and Margin[94] v2.3), and short-read kmer (Pangenie). The diploid assembly was then reconstructed by augmenting the haploid reference with these phased variants using BCFtools consensus.

To correct structural discordances, draft assemblies were aligned to their corresponding newly generated diploid assembly using minimap2 and SecPhase (https://github.com/mobinasri/secphase) v0.4.3. Structural variants were identified with paftools.js and validated using long-read alignments via cuteSV[95] v2.1.0. Regions surrounding the validated SVs (±2 × variant size) were patched using corresponding segments from the draft assemblies. Finally, we used Merqury to identify erroneous k-mer regions (defined as ±2 kb around kmer errors) and clipped any such region larger than 10 kb from the final assemblies.

Evaluation of variant detection based on PIGA assembly

To evaluate the performance of variant detection using our PIGA assemblies, we generated a variant callset from the PIGA diploid assemblies using Dipcall[98] and compared it against several state-of-the-art methods. For read-based approaches, we included variant callsets from GATK (SNVs and indels from short reads), Manta[115] (SVs from short reads), Sniffles2 (SVs from long reads), and Survivor[116] (SVs supported by at least two of the following callers: Sniffles2, CuteSV, Delly[139], and Manta using both short and long reads).

We also followed the HGSVC3[140] 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[141]. Personal consensus sequences were generated by modifying the CHM13 reference with each individual's final variant callset using BCFtools 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 (Extended Data Figure 3).

Supplementary Note 2: Evaluation of epigenomic prediction models

For epigenomic annotation, we used deep-learning models to predict personalized epigenome profiles from assembly sequences. Since chromatin accessibility is a fundamental feature of many regulatory elements[142], we used blood ATAC-seq data from three benchmark samples to compare the prediction performance of two state-of-the-art models, SEI[22] and Enformer[143], 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.

Supplementary Note 3: Quality assessment of the 1KCP pangenome

With a total size of 3.74 Gb, the 1KCP pangenome was larger than previous human pangenomes[2,3], yet remained close to the typical 3 Gb size of a haploid human genome. 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. The coverage distribution was approximately uniform across the genome, with 96.23% of unclipped autosomal regions covered by > 95% of haplotypes (Supplementary Figure 17). This result indicates that the pangenome effectively collapsed redundant sequences across assemblies into non-redundant segment representations.

Supplementary Note 4: Complexity of multiallelic SVs

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. The prevalence of multiallelic SVs is primarily driven by the sample size of the callset (Figure 2f). Specifically, the number of SV alleles detected in the full 1KCP callset is five times greater than in a smaller callset with 100 samples. Additionally, the number of alleles per SV site correlates with its size and sequence properties (Supplementary Figure 20). 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.

Supplementary Note 5: Internal validation of imputation performance

To comprehensively evaluate the imputation performance of the 1KCP reference panel across diverse variant types, we first 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. The imputation accuracy was assessed using the squared Pearson correlation coefficient (r^2) between imputation dosages and true genotypes. Among the variant types with biallelic representations

(SNVs, indels, SVs, and nested variants), SNVs exhibited the highest imputation accuracy, followed by SVs, indels, and nested variants (Figure 5a). For common variants (MAF > 0.05), the mean r^2 was 0.72 for SVs and 0.54 for nested variants, while for rare and low-frequency variants (MAF ≤ 0.05), the mean r^2 was 0.47 for SVs and 0.32 for nested variants (Supplementary Table 34). For TRs, we merged multiallelic dosages and achieved a mean r^2 of 0.73 for TR length variants and 0.69 for TR motif variants (Figure 5b). HLA alleles showed high imputation accuracy, with a mean r^2 of 0.87 for four-field alleles and 0.91 for three-field alleles. 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 (Figure 5c). In total, we imputed 9.9 million variants with $r^2 > 0.8$, including 0.58 million located in low-mappability regions that are typically challenging to genotype with short reads (Figure 5d). Additionally, we demonstrated that short-read WGS data can be used to impute complex variants and HLA alleles with higher accuracy than array data, showing the critical role of marker density in achieving high imputation accuracy (Supplementary Table 34).

Supplementary Note 6: Comparison with existing imputation panels

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 35). For SVs, we compared imputation performance against the 1000G-ONT SV panel[144]. The 1KCP panel achieved a higher mean r^2 across MAF ranges and imputed 26% more SVs with $r^2 > 0.8$ (Extended Data Figure 9a-b). For TRs, the 1KCP panel achieved a higher mean F1 score than the EnsembleTR panel[145] (0.56 vs 0.46) (Extended Data Figure 9c), 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[46] (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 $r^2 > 0.8$ (Extended Data Figure 9d-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.

Supplementary Note 7: Functional annotation of non-reference sequences and model limitations

From the 1KCP pangenome, we identified 405.3 Mb of non-reference sequences absent from current reference genomes, with their functional properties yet to be explored. To investigate these non-reference sequences, we developed a path-guided annotation method to transfer functional annotations from assemblies to the pangenome. 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. These results confirm the existence of functional non-reference sequences within the population that may contribute to phenotypic variations. 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 19b). More advanced genomic language models[146] 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.

Supplementary Note 8: Implications of complex variant modeling

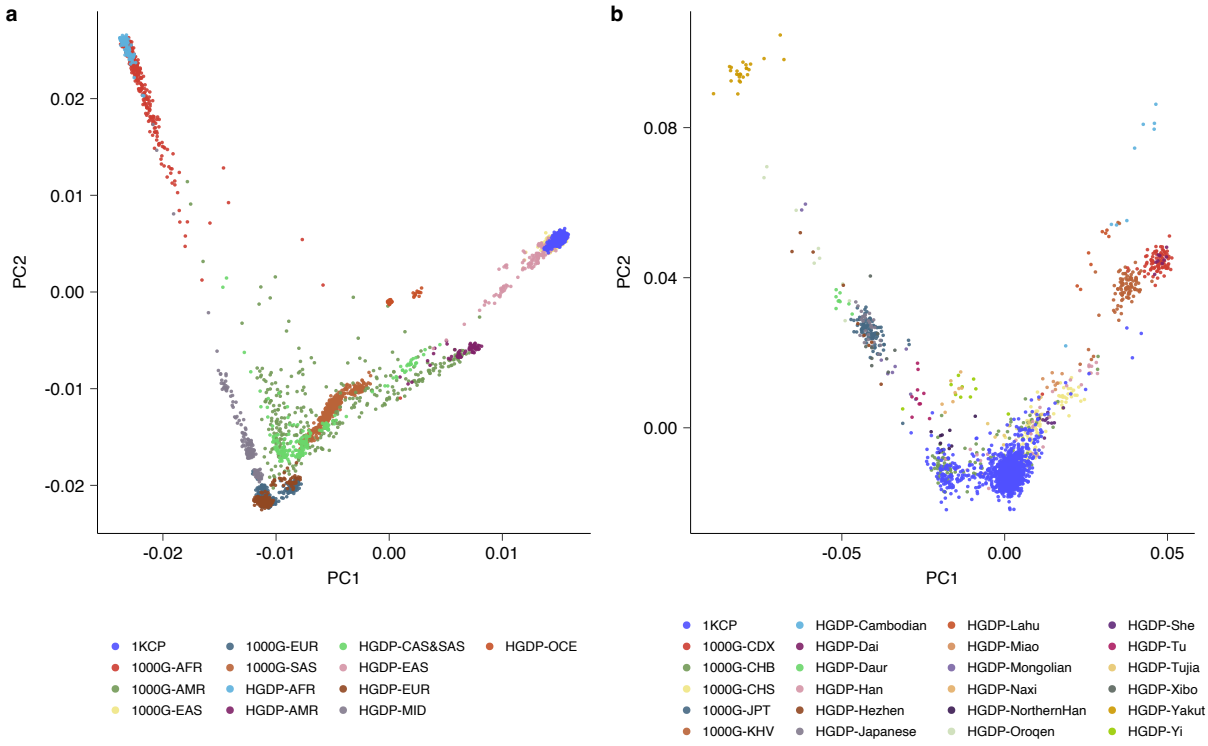
Using the 1KCP assemblies, we built an extensive variant catalog containing both small and complex variants, showcasing the unique strength of diploid assembly in capturing the complexity of genetic variations. Unlike small variants, complex variants can influence phenotype in more complicated ways. For instance, SVs may directly alter gene structures or regulatory elements while introducing numerous nested variants within their sequences, resulting in fine-tuned functional impacts. 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. TRs provide a representative example; their effects are typically modeled based on repeat length. This modeling approach outperforms treating TR alleles as biallelic variants and unlocks the analytical power of TR sites. Our findings further reveal that many TRs regulate gene expression through specific sequence motifs, emphasizing the importance of considering both repeat length and motif composition when assessing TR functionality.

Supplementary Note 9: Limitations of the study

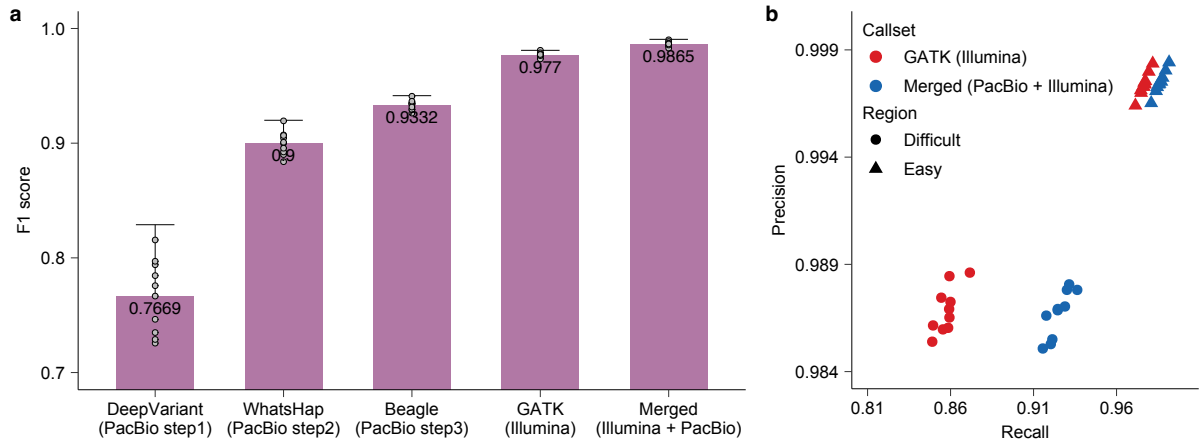
While the 1KCP assemblies have provided valuable insights into the diversity of complex variants and their functional implications, certain limitations remain. First, the PIGA assemblies exhibit a moderate level of indel errors, potentially due to the use of PacBio non-HiFi reads, which have lower sequencing accuracy compared to HiFi reads. A similar issue is observed in diploid assemblies generated using Oxford Nanopore Technology (ONT) sequencing data[147]. 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. 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 sequencing data used may lead to some rare variants being missed at the individual level. 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[148] 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[149]. Continued advancements in sequencing technology and scalable T2T assembly algorithms[150,151] 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.

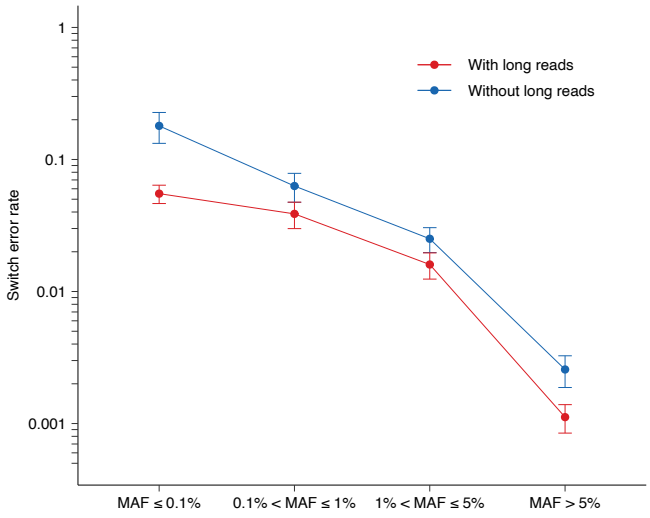
Supplementary Figures



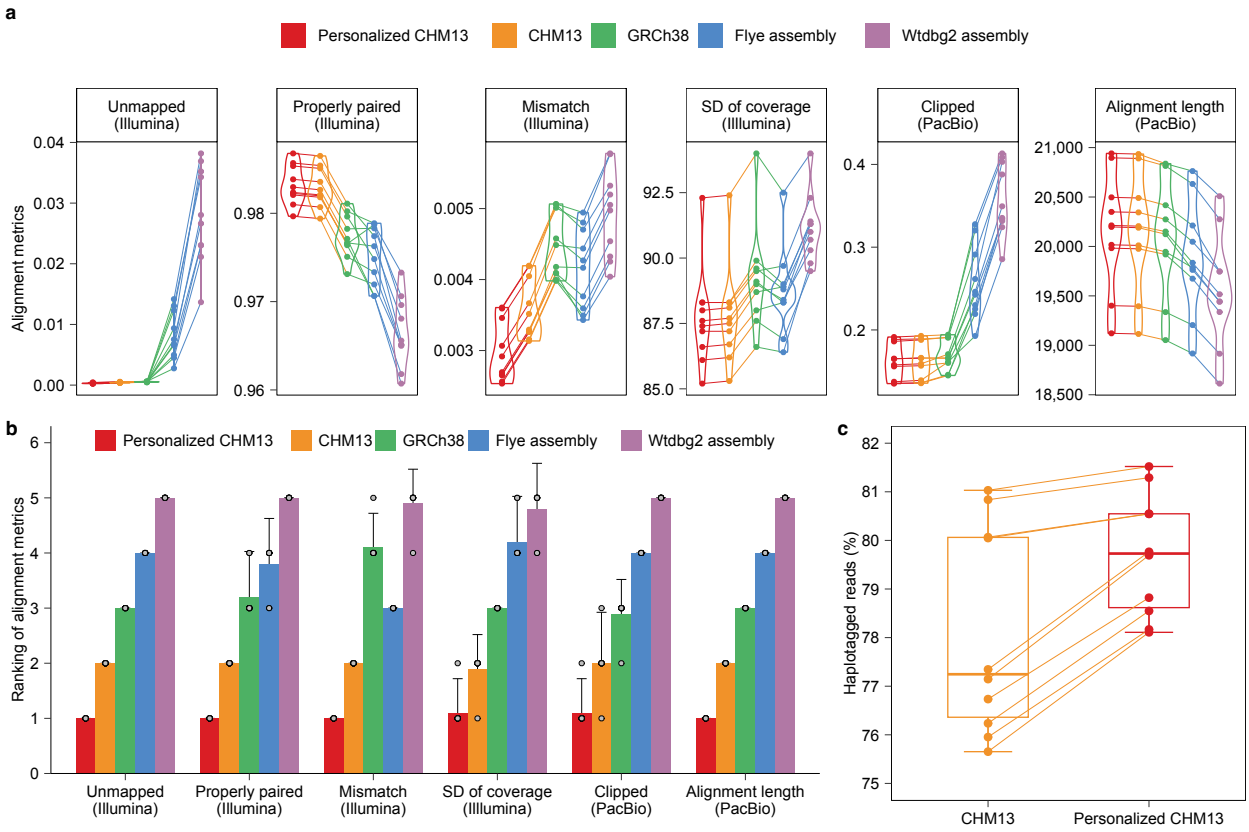
Supplementary Figure 1. Principal Component Analysis (PCA) based on short-read WGS genotypes. **a**, PCA of the 1KCP samples alongside geographically diverse samples from the 1000 Genomes Project and Human Genome Diversity Project. **b**, PCA of the 1KCP samples alongside East Asian samples from the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



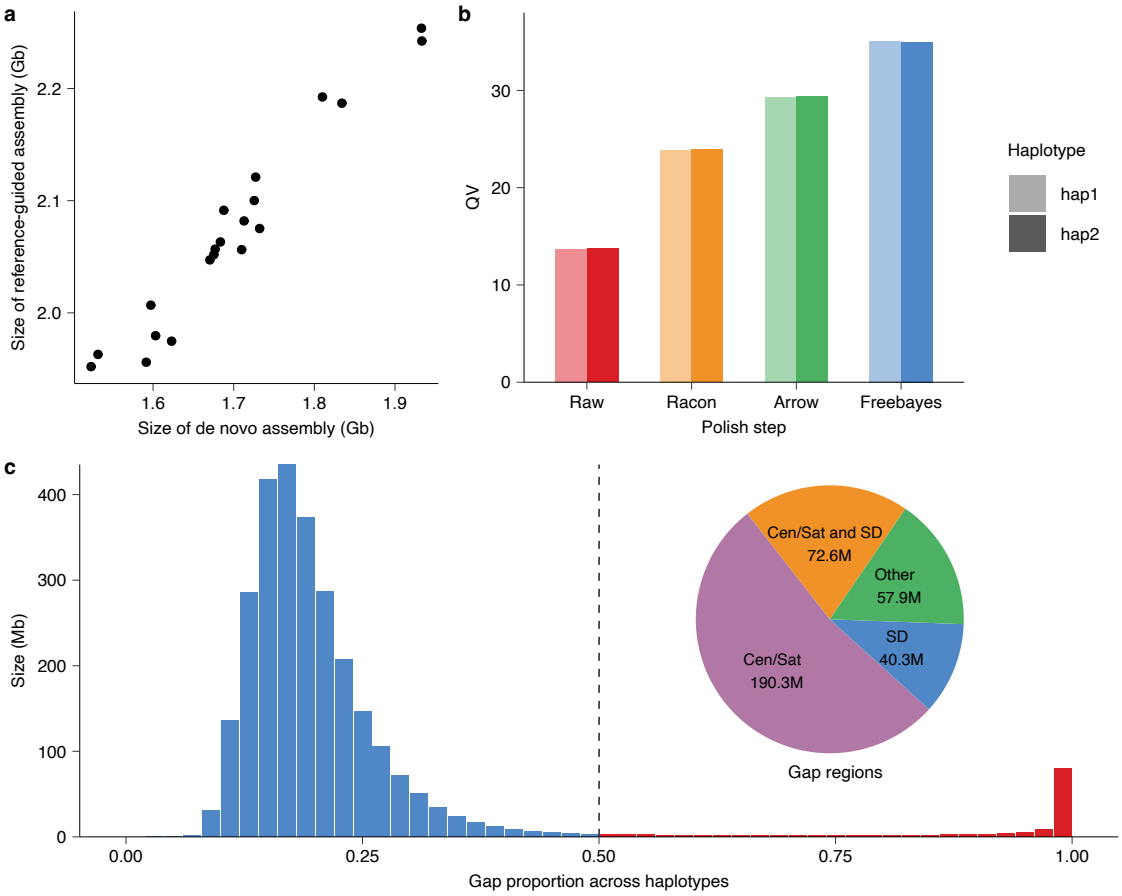
Supplementary Figure 2. Performance benchmarking of the PIGA SNV detection pipeline in ten benchmarking samples. a, F1 scores for SNV calling at progressive stages of the pipeline in ten benchmarking samples. Each dot represents an individual sample, bars indicate the mean value, and error bars denote 95% confidence intervals. **b**, Precision-recall plot of the merged callset and the Illumina callset, stratified by easy and difficult genomic regions.



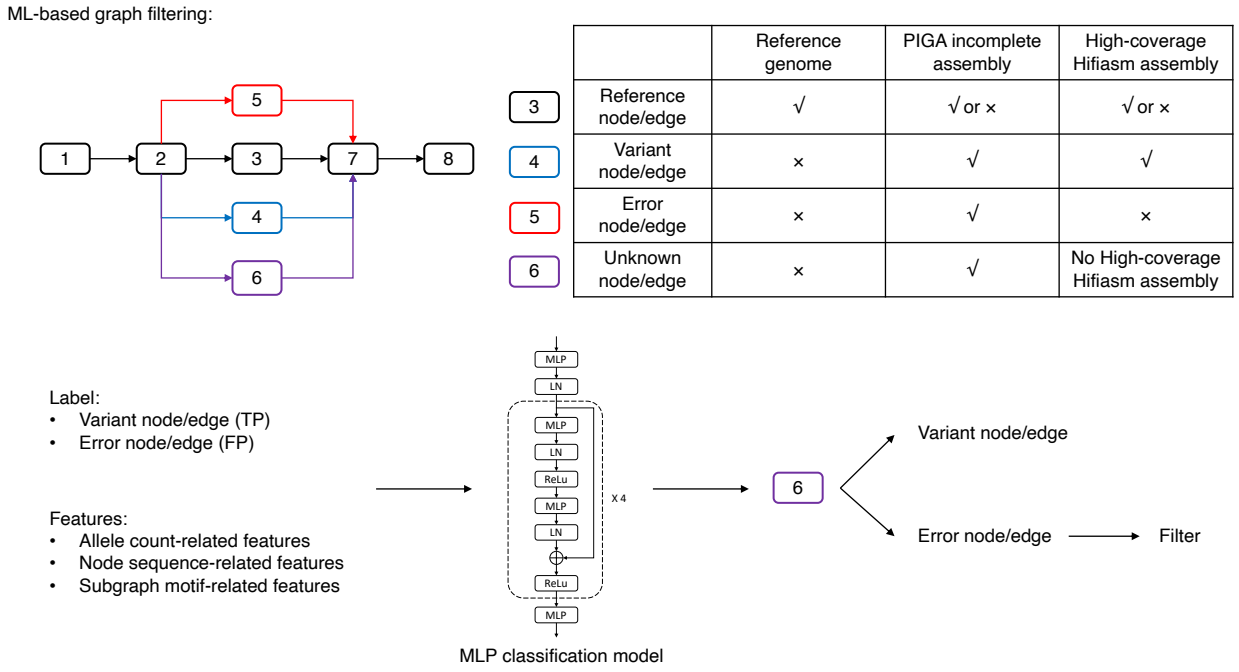
Supplementary Figure 3. Performance benchmarking of the PIGA SNV phasing. Switch error rates for SNV phasing in ten benchmarking samples with and without long-read information, stratified by minor allele frequency. Dots represent mean values, and error bars indicate 95% confidence intervals.



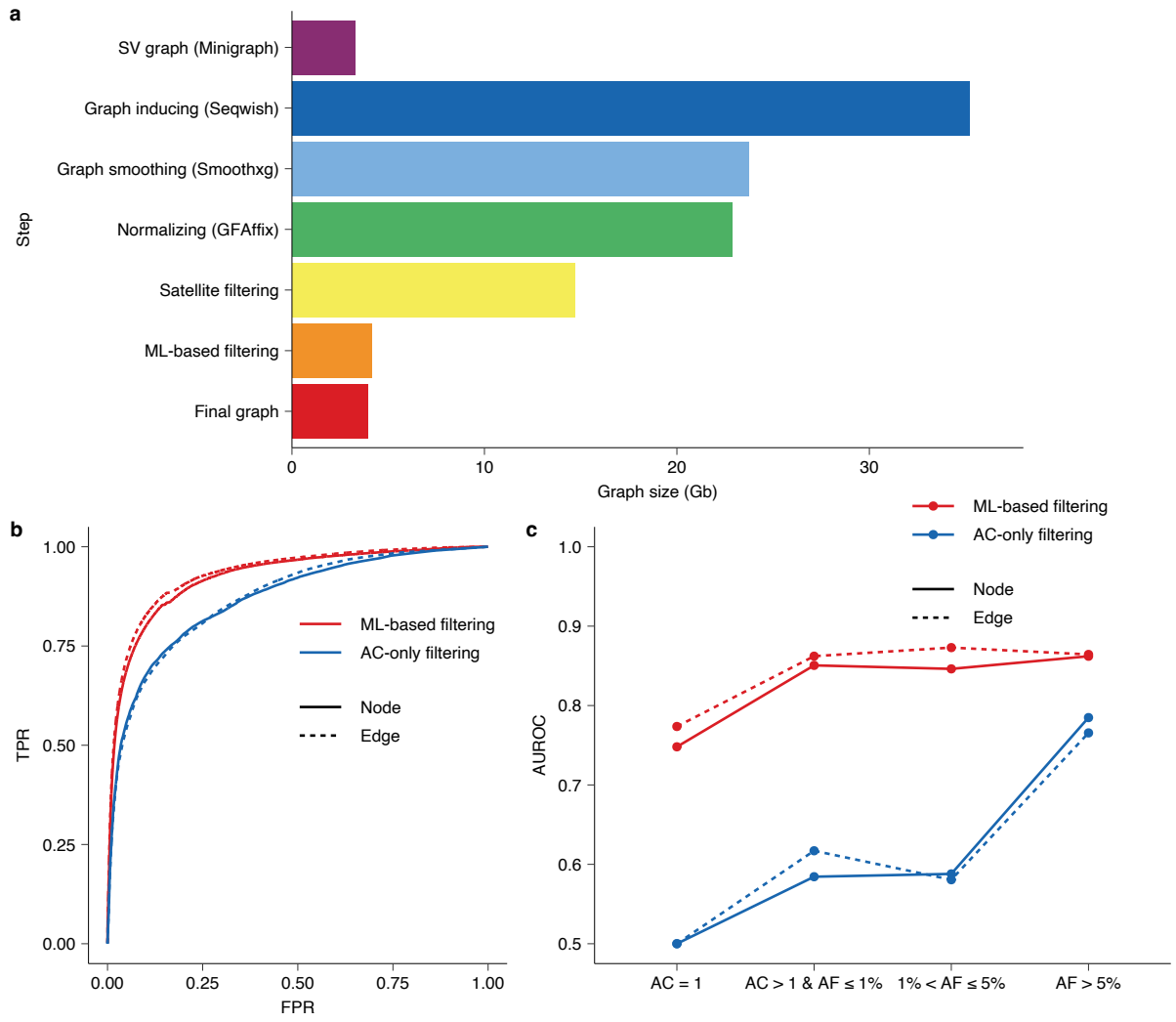
Supplementary Figure 4. Evaluation of the PIGA personalized reference for read alignment and haplotype partitioning in ten benchmark samples. a, Comparison of alignment metrics across five reference types. Metrics include the proportion of unmapped Illumina reads, proportion of properly paired Illumina reads, mismatch rate of Illumina reads, standard deviation of Illumina read coverage, proportion of clipped PacBio reads, and average alignment length of PacBio reads. **b**, Ranking of alignment metrics across five reference types in ten benchmark samples. Each dot represents an individual sample, bars indicate the mean value, and error bars denote 95% confidence intervals. **c**, 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. In the box plots, the centre line represents the median; the box limits indicate the 25th and 75th percentiles; and the whiskers extend to 1.5× the interquartile range from the box limits.



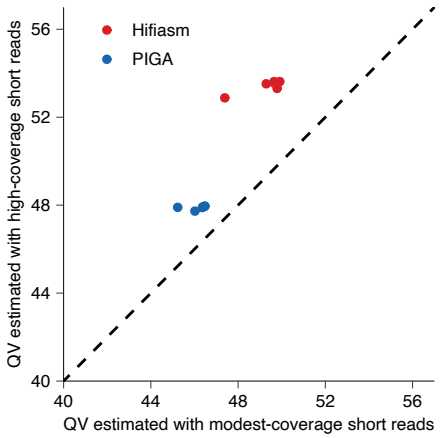
Supplementary Figure 5. Evaluation of the PIGA reference-guided draft diploid assembly pipeline. **a**, Comparison of assembly sizes between the reference-guided and de novo assemblies in ten benchmark samples. **b**, Quality values (QV) of draft assemblies across polishing steps, using the individual 1KCP-00309 as an example. **c**, 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).



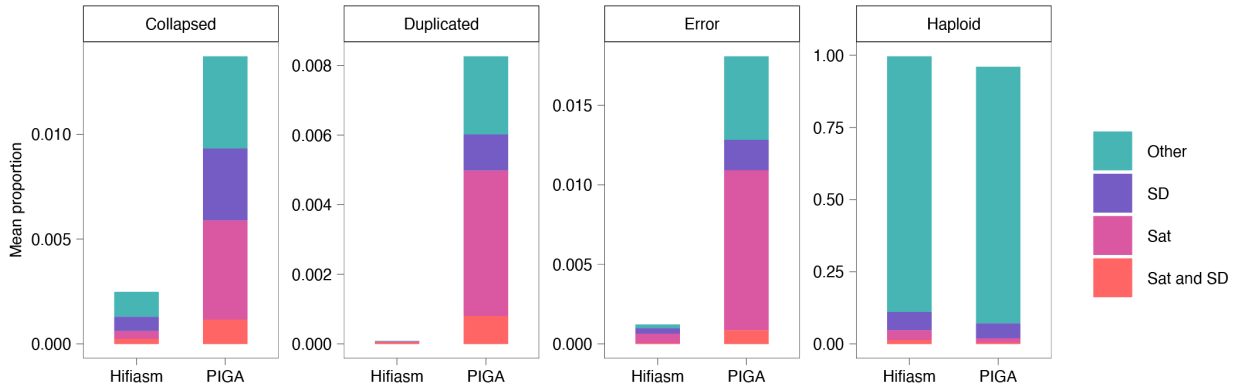
Supplementary Figure 6. Framework of the PIGA machine learning-based pangenome filtering. Overview of the workflow used to identify and remove error nodes/edges from the pangenome graph. Ground-truth labels for training are assigned by comparing node/edge presence across three sources: the reference 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.



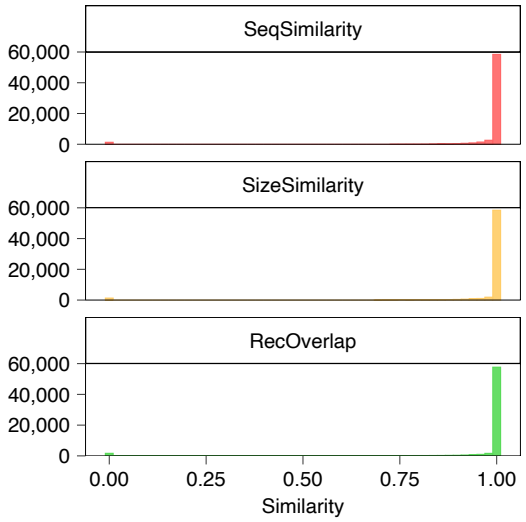
Supplementary Figure 7. Performance evaluation of the PIGA pangenome construction and simplification pipeline. a, Graph size of the pangenome at different stages of the pipeline. **b**, ROC curves evaluating the classification performance on unseen nodes/edges from three benchmark samples. Comparison is shown between the machine learning (ML)–based filtering and the allele count (AC)–only filtering methods, for both nodes and edges. **c**, AUROC values for ML-based and AC-only filtering to classify variants and errors within different allele frequency bins.



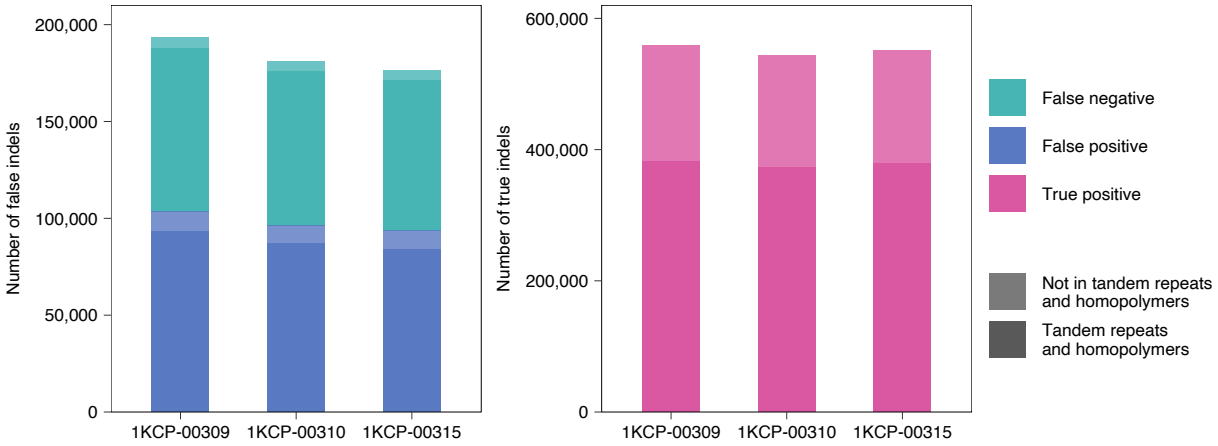
Supplementary Figure 8. QV estimates for PIGA and Hifiasm assemblies of three benchmark samples, based on both modest-coverage and high-coverage short-read datasets.



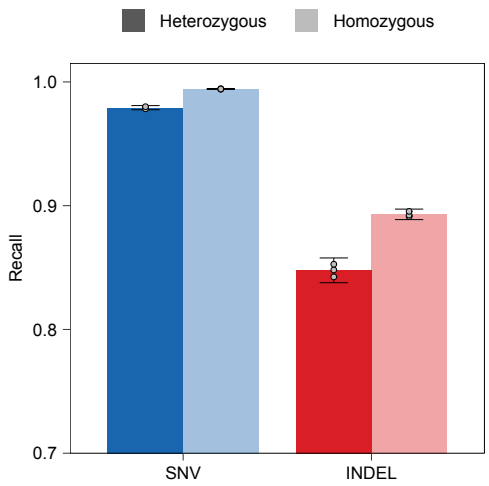
Supplementary Figure 9. Regional reliability of three 1KCP benchmark assemblies. Reliability classes were determined using Flagger with high-coverage HiFi reads. Regions are categorized as segmental duplications (SD), satellites (Sat), both satellites and segmental duplications (Sat and SD), or other regions.



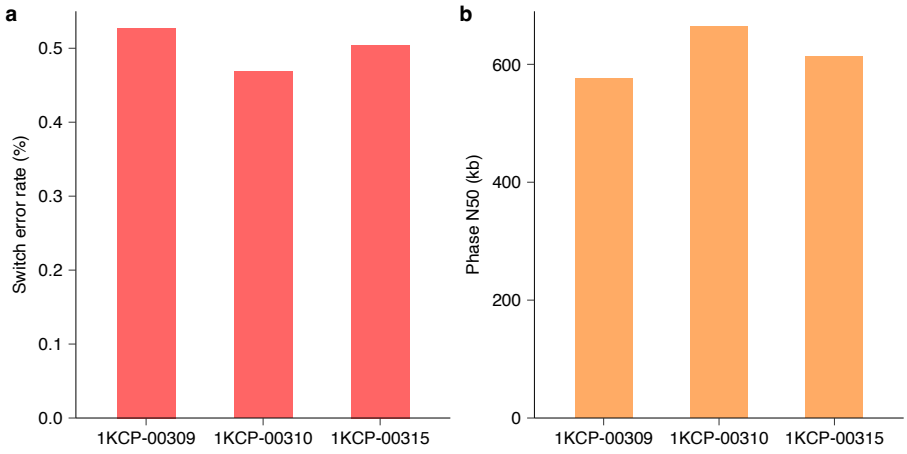
Supplementary Figure 10. Benchmarking SV accuracy in three benchmark PIGA assemblies. **a**, Distribution of SV sequence similarities between the evaluation callset and the benchmark callset. **b**, Distribution of SV size similarities between the evaluation callset and the benchmark callset. **c**, Distribution of SV reciprocal overlap percentages between the evaluation callset and the benchmark callset.



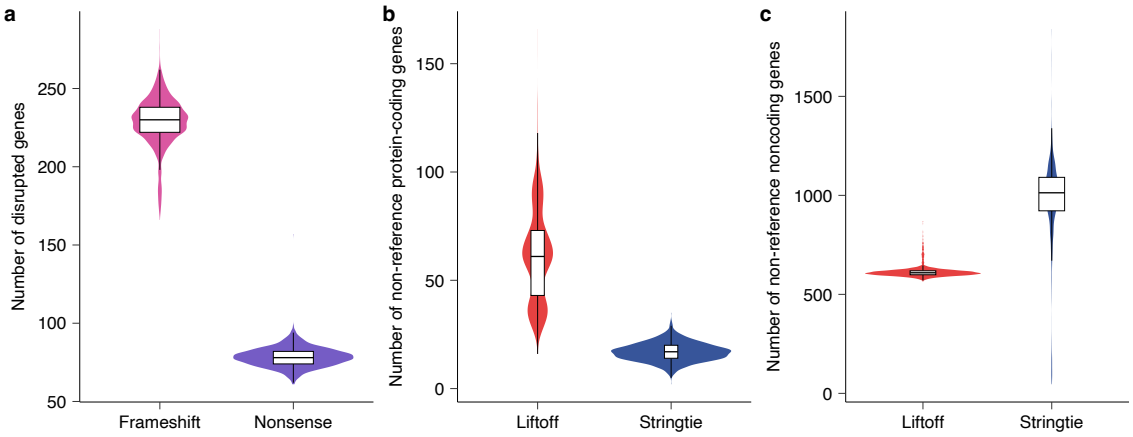
Supplementary Figure 11. Benchmarking indel detection performance in three benchmark PIGA assemblies. The numbers of false positive, false negative, and true positive indels stratified by their presence within or outside tandem repeat and homopolymer regions.



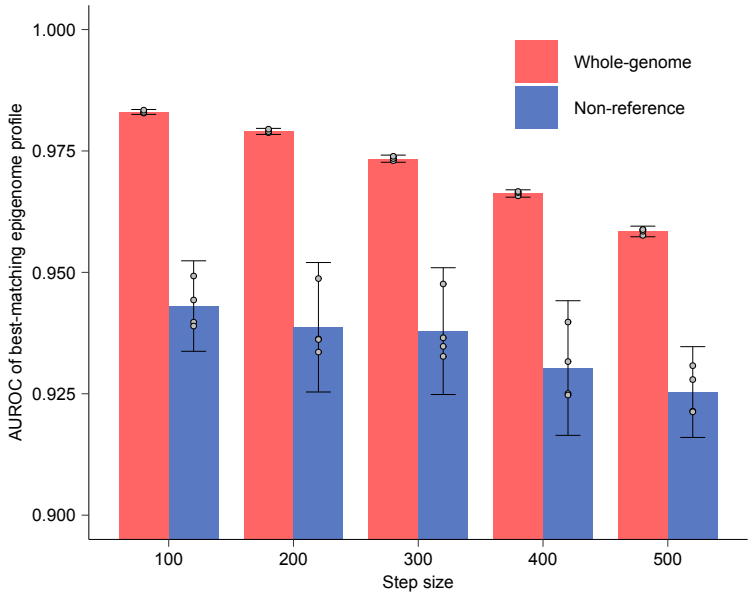
Supplementary Figure 12. Recall of small variant detection using PIGA assemblies for three benchmark samples, stratified by heterozygous and homozygous variants. Each dot represents an individual sample, bars indicate the mean value, and error bars denote 95% confidence intervals.



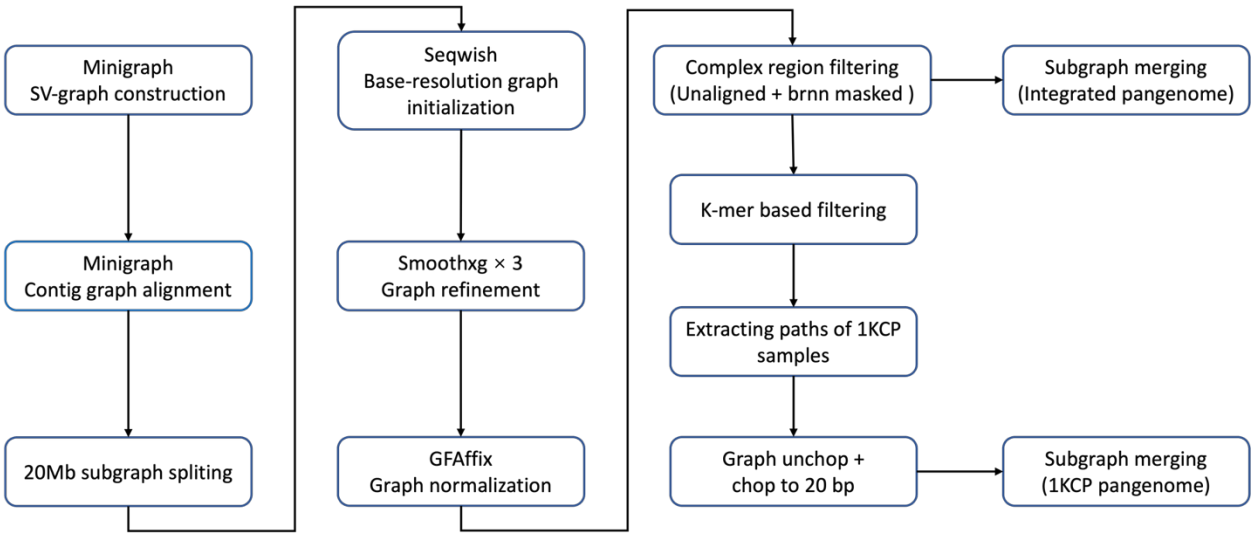
Supplementary Figure 13. Haplotype phasing performance of three benchmark PIGA assemblies. **a**, Switch error rate of PIGA assemblies compared to Hifiasm assemblies. **b**, N50 size of correctly phased blocks in PIGA assemblies compared to Hifiasm assemblies.



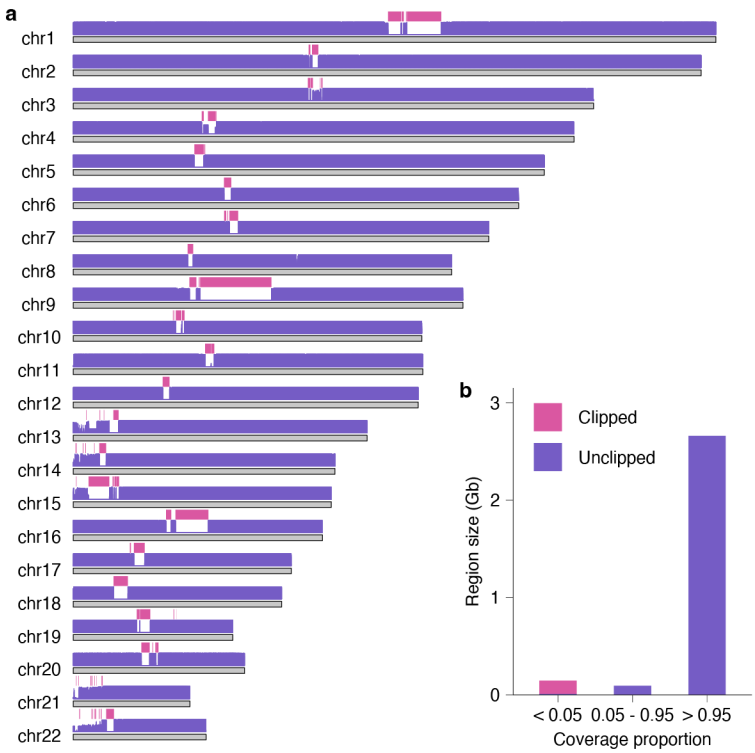
Supplementary Figure 14. Gene annotation metrics for the 1KCP assemblies. **a**, Distribution of the number of disrupted protein-coding genes caused by frameshift or nonsense mutations. In the box plots, the centre line represents the median; the box limits indicate the 25th and 75th percentiles; and the whiskers extend to 1.5× the interquartile range from the box limits. **b**, Distribution of the number of non-reference protein-coding genes identified using Liftoff and StringTie. In the box plots, the centre line represents the median; the box limits indicate the 25th and 75th percentiles; and the whiskers extend to 1.5× the interquartile range from the box limits. **c**, Distribution of the number of non-reference noncoding genes identified using Liftoff and StringTie. In the box plots, the centre line represents the median; the box limits indicate the 25th and 75th percentiles; and the whiskers extend to 1.5× the interquartile range from the box limits.



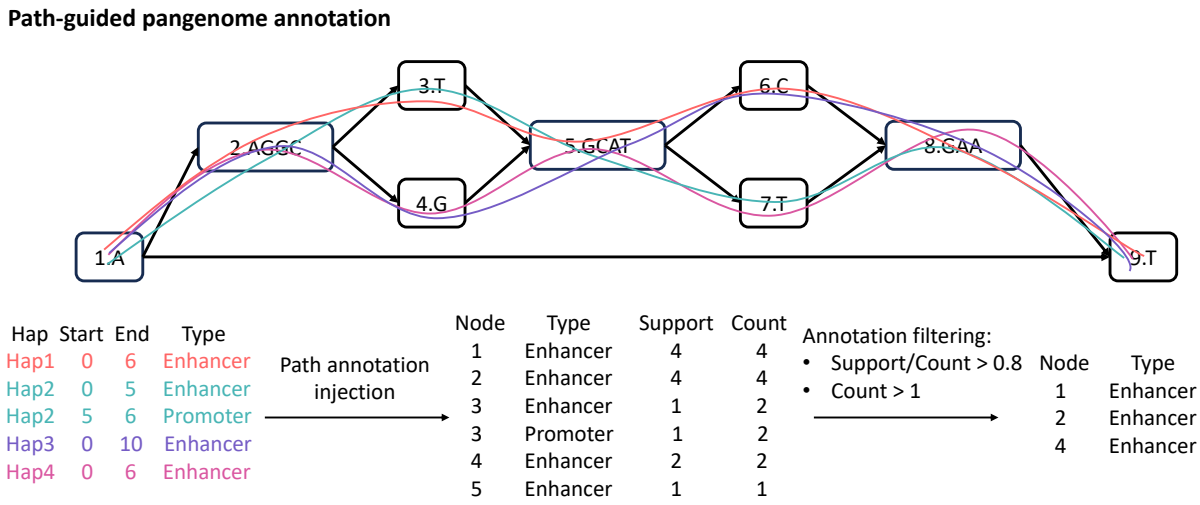
Supplementary Figure 15. Epigenome annotation performance of SEI. Epigenome annotation performance of SEI evaluated on whole-genome (red bars) and non-reference sequences (blue bars) for PIGA and Hifiasm haplotype assemblies (n = 4) of the benchmark sample 1KCP-00309, stratified by annotation step sizes. The performance is measured by the average area under the receiver operating characteristic (AUROC). Each dot represents a haplotype assembly, bars indicate the mean value, and error bars denote 95% confidence intervals.



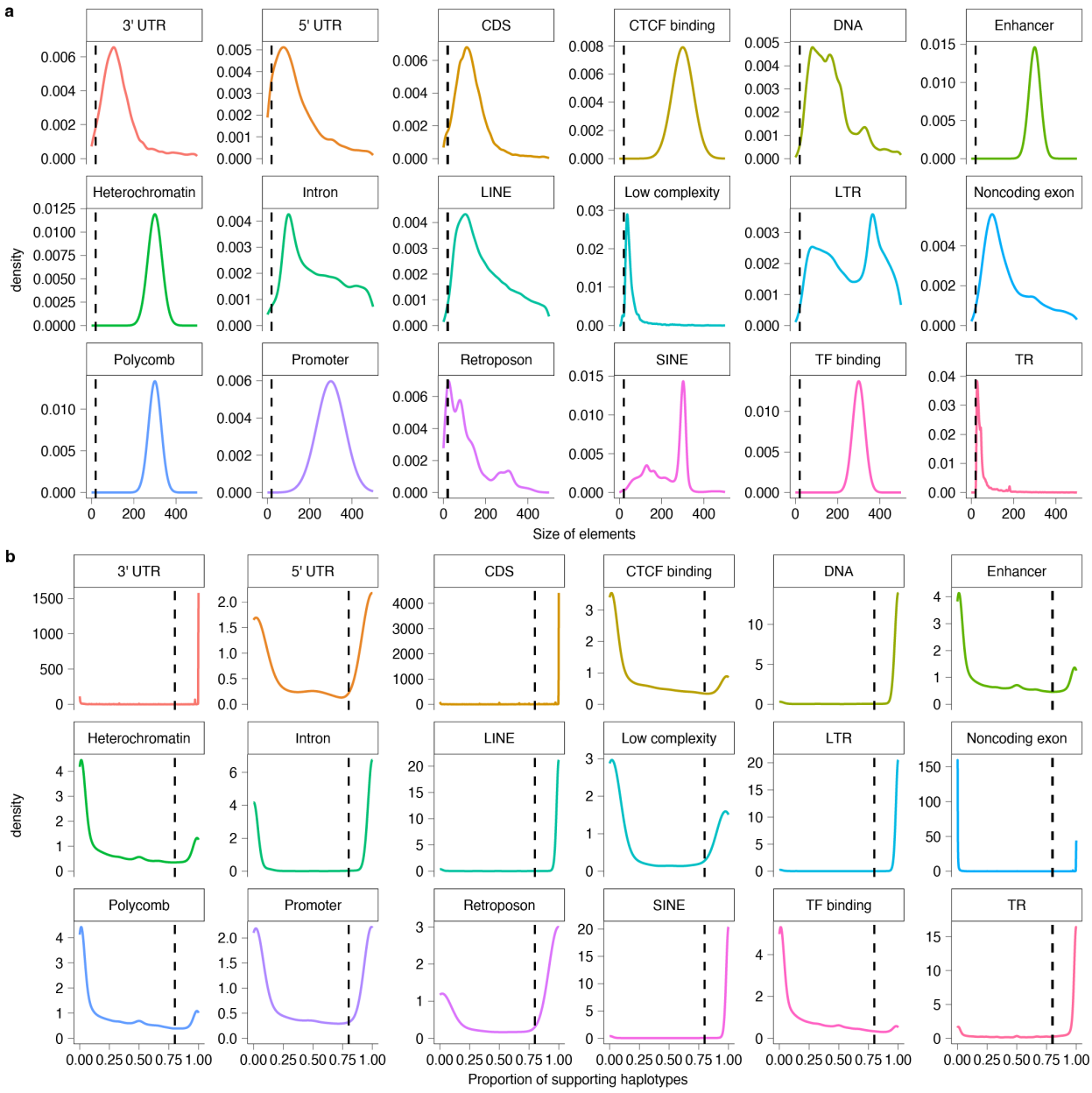
Supplementary Figure 16. Overview of the pangenome construction pipeline. The Minigraph-based pangenome was initially constructed and divided into 171 subgraphs, each approximately 20 Mb in size. Subsequent steps were conducted in parallel. The base-level pangenome was generated using Seqwish and refined with Smoothxg. The resulting subgraphs were normalized using GFAffix, filtered by read k-mers, and clipped to form the integrated pangenome. From this, a pangenome containing only 1KCP and reference genomes was extracted to create the 1KCP pangenome. Finally, the 1KCP pangenome was further segmented into 20 bp fragments for annotation.



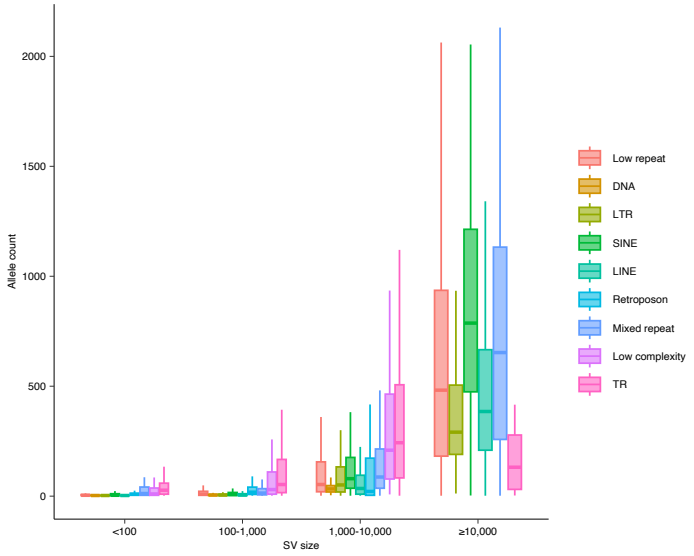
Supplementary Figure 17. Haplotype coverage of pangenome segments mapped to CHM13 coordinates. **a**, Haplotype coverage across CHM13 genomic regions. Purple bars represent haplotype coverage, while pink bars indicate clipped regions. **b**, Distribution of region sizes with varying haplotype coverage proportions, stratified by clipping status.



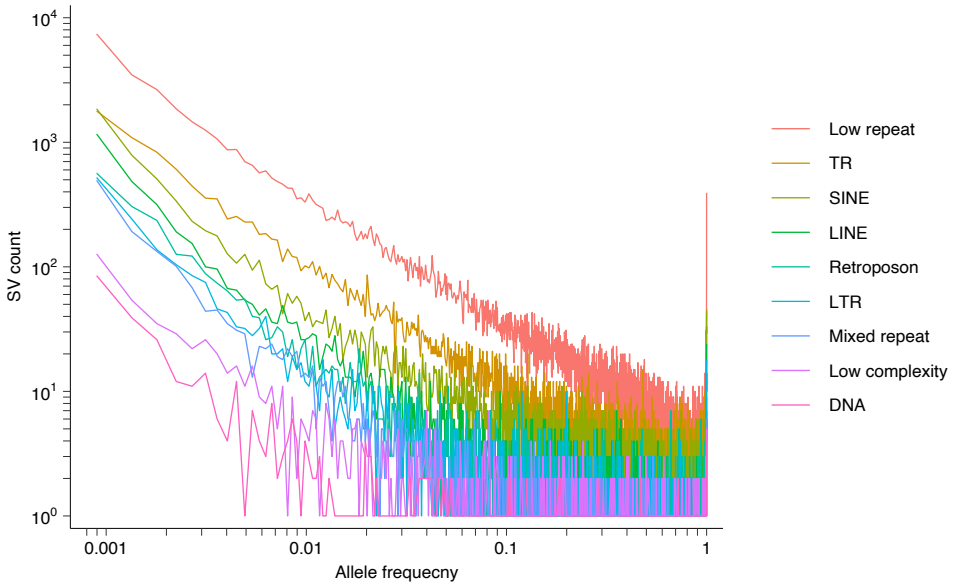
Supplementary Figure 18. Path-guided pangenome annotation example. Three haplotype paths traverse and annotate distinct segments of the pangenome, using assembly-based annotations. These annotations from multiple assemblies are integrated and filtered to generate a consensus pangenome annotation, ensuring accuracy and consistency.



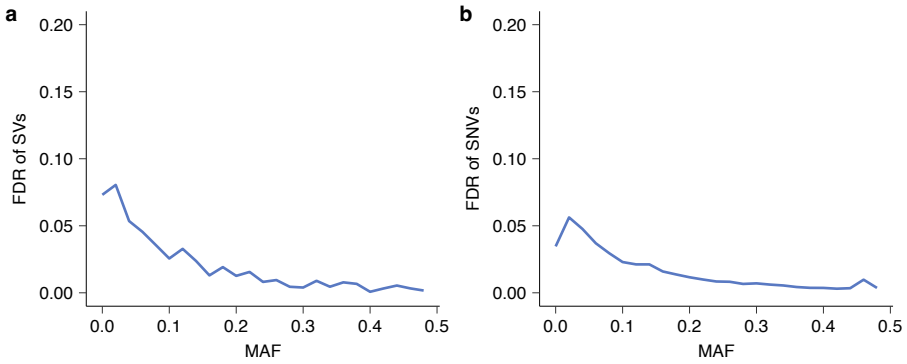
Supplementary Figure 19. Pangenome annotation metrics. **a**, Size distribution of annotated genomic elements within the 1KCP-00309 sample. Each panel represents a specific genomic element category. **b**, Distribution of supporting haplotype proportions for annotated genomic elements. Each panel represents a specific genomic element category.



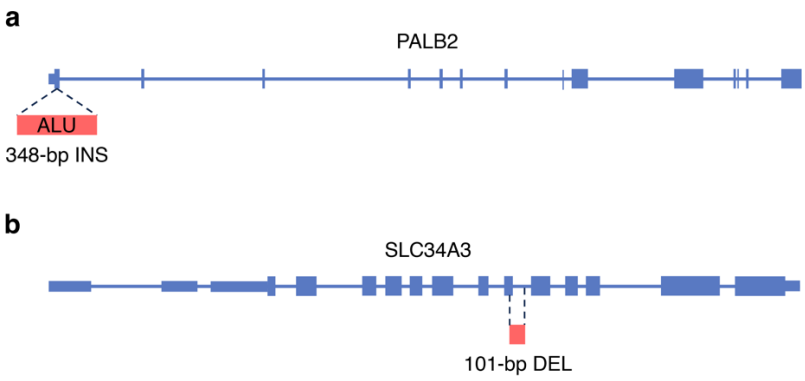
Supplementary Figure 20. Allele counts of SV sites, stratified by repeat type and SV size. Repeat types include low repeat (SVs with <50% repeat content), mixed repeat (SVs with >50% content of multiple repeat types), and specific repeat classes (SVs with >50% content of single repeat types). In the box plots, the centre line represents the median; the box limits indicate the 25th and 75th percentiles; and the whiskers extend to 1.5× the interquartile range from the box limits.



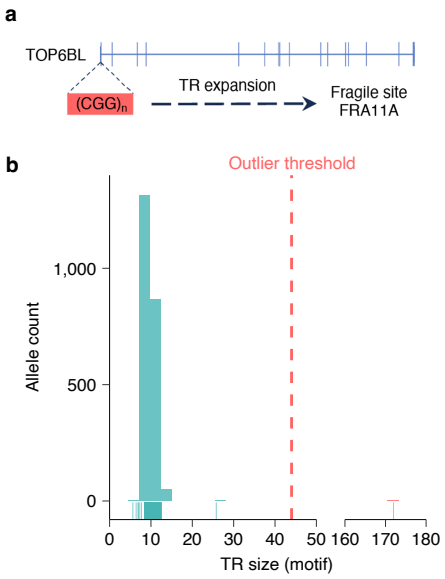
Supplementary Figure 21. Allele frequency spectrum of SVs, stratified by repeat type. Repeat types include low repeat (SVs with <50% repeat content), mixed repeat (SVs with >50% content of multiple repeat types), and specific repeat classes (SVs with >50% content of single repeat types).



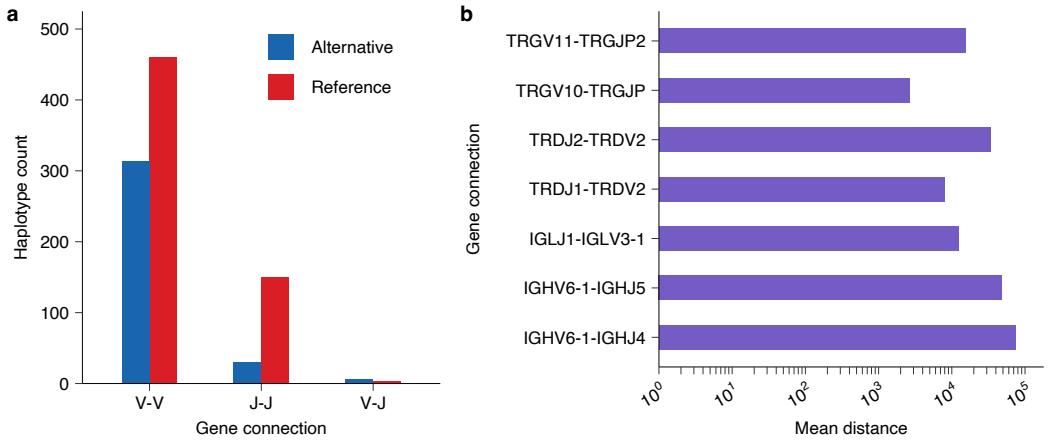
Supplementary Figure 22. False discovery rate (FDR) evaluation of the 1KCP callsets. a, FDR of SVs across 25 MAF bins, estimated using three benchmark samples. **b**, FDR of SNVs across 25 MAF bins, estimated using three benchmark samples.



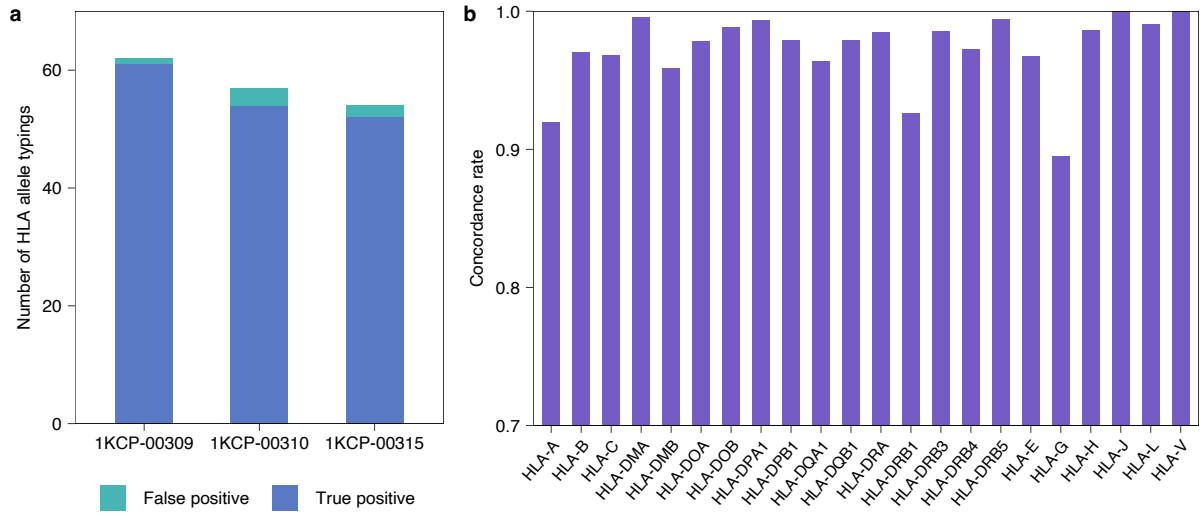
Supplementary Figure 23. Disease-related SVs identified in the 1KCP SV callset. **a**, A 348-bp ALU insertion identified in the *PALB2* gene. **b**, A 101-bp deletion located in the *SLC34A3* gene.



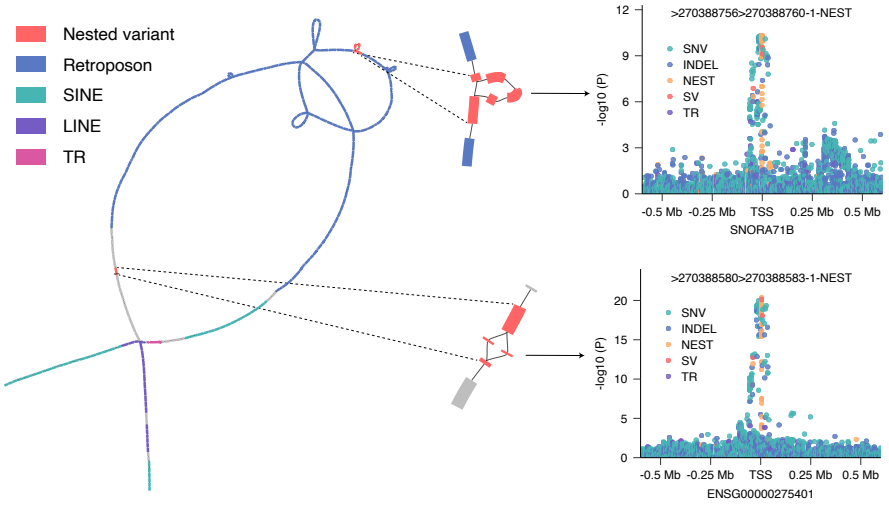
Supplementary Figure 24. Illustration of *TOP6BL* TR expansion. **a**, Genomic location of the TR within the *TOP6BL* gene, which is linked to the rare fragile site FRA11A. **b**, Size distribution of the *TOP6BL* TR in the 1KCP dataset. The vertical green ticks represent the sizes of normal TR alleles, 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.



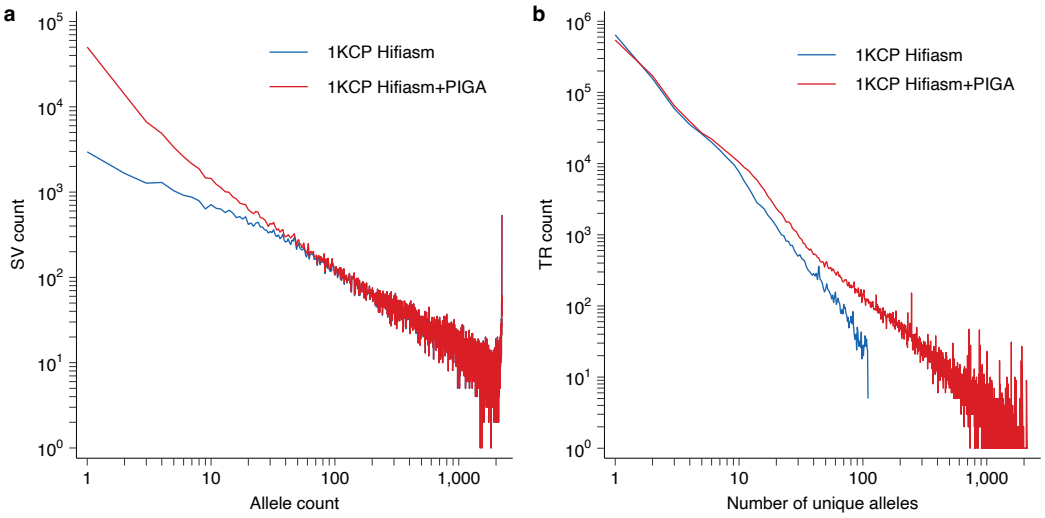
Supplementary Figure 25. Gene–gene connection analysis among V, D, and J genes within immunoglobulin and T cell receptor gene clusters. a, Number of gene–gene connections involving V, D, and J genes in the pangene graph. Only seven alternative junctions between different gene types (V–J) were detected. **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



Supplementary Figure 26. 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.



Supplementary Figure 27. The eNested variant examples. The subgraph of a 3.6kb insertion site and its two lead eNested variants. The locus plots show the eQTL signals associated with the expression level of *SNORA71B* and ENSG00000275401.



Supplementary Figure 28. 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.

136 Dwarshuis, N. *et al.* The GIAB genomic stratifications resource for human reference genomes. *Nature communications* **15**, 9029 (2024).

137 Kolmogorov, M., Yuan, J., Lin, Y. & Pevzner, P. A. Assembly of long, error-prone reads using repeat graphs. *Nature biotechnology* **37**, 540-546 (2019).

138 Al Bkhetan, Z., Zobel, J., Kowalczyk, A., Verspoor, K. & Goudey, B. Exploring effective approaches for haplotype block phasing. *BMC bioinformatics* **20**, 540 (2019).

139 Rausch, T. *et al.* DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics* **28**, i333-i339 (2012).

140 Logsdon, G. A. *et al.* Complex genetic variation in nearly complete human genomes. *bioRxiv* (2024).

141 Hofmeister, R. J., Ribeiro, D. M., Rubinacci, S. & Delaneau, O. Accurate rare variant phasing of whole-genome and whole-exome sequencing data in the UK Biobank. *Nature genetics* **55**, 1243-1249 (2023).

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

143 Avsec, Ž. *et al.* Effective gene expression prediction from sequence by integrating long-range interactions. *Nature methods* **18**, 1196-1203 (2021).

144 Noyvert, B. *et al.* Imputation of structural variants using a multi-ancestry long-read sequencing panel enables identification of disease associations. *medRxiv*, 2023.2012. 2020.23300308 (2023).

145 Ziaei Jam, H. *et al.* A deep population reference panel of tandem repeat variation. *Nature Communications* **14**, 6711 (2023).

146 Dalla-Torre, H. *et al.* Nucleotide Transformer: building and evaluating robust foundation models for human genomics. *Nature Methods*, 1-11 (2024).

147 Kolmogorov, M. *et al.* Scalable Nanopore sequencing of human genomes provides a comprehensive view of haplotype-resolved variation and methylation. *Nature Methods* **20**, 1483-1492 (2023).

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

149 Logsdon, G. A. *et al.* The variation and evolution of complete human centromeres. *Nature* **629**, 136-145 (2024).

150 Cheng, H., Asri, M., Lucas, J., Koren, S. & Li, H. Scalable telomere-to-telomere assembly for diploid and polyploid genomes with double graph. *Nature Methods*, 1-4 (2024).

151 Rautiainen, M. *et al.* Telomere-to-telomere assembly of diploid chromosomes with Verkko. *Nature Biotechnology* **41**, 1474-1482 (2023).