Article

The 1000 Chinese Pangenome empowers medical and population genetics

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

Received: 25 February 2025

Accepted: 17 February 2026

Published online: 1 April 2026

Open access

Check for updates

Yifei Wang[1,2,10], Zhongqu Duan[1,2,10], Dan Chen[1,2,3], Dandan Shi[1,3,4], Yi Ding[1,2], Zhibin Wang[3,5], Baoqing Li[6], Zhiyi Wang[6,7], Minmin Guo[1,2], Wen Yang[1,2], Junren Hou[1,2], Wenhao Chen[1,2], Yazhou Guo[1,2], Wenjie Wei[1,2], Yujie Cao[1,2], Xiwei Sun[1,2], Weiyang Bai[1,2], Mingdong Lu[6], Ting Qi[1,2,8], Xian Shen[9] ✉ & Jian Yang[1,2] ✉

Pangenomes are revolutionizing our ability to resolve genomic regions with complex variations[1]. However, existing human pangenomes[2,3], constrained by small sample sizes, provide limited utility for medical and population genetic applications. Here we generated 1,116 diploid genome assemblies (55 de novo and 1,061 pangenome-informed) with an average size of 2.98 Gb and a mean quality value of 46 as part of the 1000 Chinese Pangenome (1KCP) project. On the basis of 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. We catalogued a full spectrum of genetic variation, including 35.4 million small variants, 110,530 structural variants (SVs), 485,575 tandem repeats (TRs) and 0.86 million nested variants embedded in non-reference sequences. This extensive dataset enabled detailed characterization of multiscale genic variations relevant to medical genetics, including gene-altering SVs, TR expansions, gene cluster variations and HLA gene haplotypes. Coupled with the 1KCP gene expression data, we conducted pan-variant expression quantitative trait locus (eQTL) mapping to analyse diverse variant types. We identified 3,256 eQTLs involving complex variants (SVs, TRs and nested variants) and elucidated their regulatory complexity. Finally, we developed a 1KCP pan-variant imputation reference panel, which provides multitype genetic markers to enhance the resolution of future association studies. This resource advances our understanding of complex variants and their functional implications to provide new insights into human health.

Since the release of the first human reference genome[4], understanding the diversity of the human genome has become a fundamental task with far-reaching implications for human health and biology. The substantial reduction in the cost of short-read sequencing has enabled the sequencing of millions of human genomes, which has revealed a wealth of genetic variation essential for advancing genomic research and associated applications[5–8]. However, although short-read sequencing is effective for detecting small variants, such as single-nucleotide variants (SNVs) and small insertions and deletions (indels), it has limitations in identifying larger and more complex variants, such as SVs and TRs[9,10]. Recent advancements in long-read sequencing technology[11] and assembly algorithms[12] have enabled the generation of high-quality diploid genome assemblies, which provide a more comprehensive view of complex variants and enhance our understanding of their formation mechanisms and functional consequences[13]. To effectively integrate the full spectrum of genetic variants in populations into a unified framework, researchers have turned to the concept of the pangenome[14].

A pangenome refers to the collection of genome sequences in a population and is typically constructed from the diploid assemblies of multiple individuals[1]. Recent efforts by the Human Pangenome Reference Consortium (HPRC)[2] and the Chinese Pangenome Consortium (CPC)[3] have demonstrated the potential of pangenomes in resolving structurally divergent regions of the human genome. However, the relatively small sample sizes of current human pangenomes pose several challenges to their broader applications. First, rare variants, which are more likely to be pathogenic than common variants[15] and have pivotal roles in inherited diseases[16,17], are underrepresented in small-scale pangenomes. Second, the limited sample sizes hamper accurate estimations of allele frequencies (AFs), which is crucial for conducting association analyses and clinical diagnostics. Third, highly repetitive genomic regions, such as TRs[18], exhibit increased mutation

[1]New Cornerstone Science Laboratory, School of Life Sciences, Westlake University, Hangzhou, China. [2]Westlake Laboratory of Life Sciences and Biomedicine, Hangzhou, China. [3]Institute for Advanced Research, Wenzhou Medical University, Wenzhou, China. [4]Chinese Institutes for Medical Research, Capital Medical University, Beijing, China. [5]School of Basic Medical Sciences, Wenzhou Medical University, Wenzhou, China. [6]The Second Affiliated Hospital and Yuying Children's Hospital of Wenzhou Medical University, Zhejiang, China. [7]Wenzhou Key Laboratory of Precision General Practice and Health Management, Wenzhou, China. [8]CAS Key Laboratory of Computational Biology, Shanghai Institute of Nutrition and Health, University of Chinese Academy of Sciences, Chinese Academy of Sciences, Shanghai, China. [9]Zhejiang Key Laboratory of Intelligent Cancer Biomarker Discovery and Translation, The First Affiliated Hospital of Wenzhou Medical University, Wenzhou, China. [10]These authors contributed equally: Yifei Wang, Zhongqu Duan. ✉e-mail: 13968888872@163.com; jian.yang@westlake.edu.cn

rates, which makes their sequence diversity difficult to broadly resolve in small populations.

To address these challenges, here we generate 1,116 diploid genome assemblies (55 de novo and 1,061 pangenome-informed) as part of the 1KCP project. We also construct a functionally annotated pangenome and provide an extensive catalogue of diverse genetic variants. Using this extensive dataset, we describe the medical relevance of genic variations at multiple resolutions and demonstrate the regulatory complexity of complex variants through eQTL analyses. Finally, we build a 1KCP imputation reference panel and develop a 1KCP data portal (https://yanglab.westlake.edu.cn/1kcp) with user-friendly tools for online browsing and imputation. Together, our results highlight the potential utility of the 1KCP dataset in human genetic research and associated applications.

Overview of the 1,116 diploid assemblies

The 1KCP project enrolled 1,379 participants, most of whom clustered with Han Chinese reference populations based on principal component analysis (PCA) (Methods and Supplementary Fig. 1). Both short-read and long-read whole-genome sequencing (WGS) were performed on a subset of 1,144 samples. We aimed to establish a large-scale diploid assembly panel and an extensive catalogue of genetic variation for medical and population genetic applications. Initially, 55 samples were sequenced with high-coverage Illumina short-read WGS (around 34.4 times coverage), PacBio HiFi long-read WGS (around 37.3 times coverage) and Hi-C (about 70.7 times coverage), referred to as the high-coverage dataset (Supplementary Table 1). This high-coverage sequencing strategy has been routinely used to produce high-quality diploid assemblies[2,3]. To scale the analysis to a larger cohort, we adopted a cost-effective sequencing strategy for 1,099 additional samples, including 10 overlapping with the high-coverage dataset. These samples were sequenced with modest-coverage Illumina short-read WGS (about 17.3 times coverage) and PacBio long-read WGS (around 9.3 times coverage, including 2.5 times HiFi reads), referred to as the modest-coverage dataset (Extended Data Fig. 1 and Supplementary Table 2). 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[19] (Methods, Extended Data Fig. 2, Supplementary Note 1 and Supplementary Figs. 2–7). Unlike de novo assembly methods that use data from single individuals, the PIGA workflow uses 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[12] and pangenome-informed assemblies for the modest-coverage samples using PIGA. 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.

After quality control, 1,116 diploid assemblies were generated (excluding PIGA assemblies for the benchmark samples), which consisted of 55 hifiasm assemblies and 1,061 PIGA assemblies (Supplementary Table 3). These diploid assemblies encompassed a total of 2,232 haploid genomes, with an average size of 2.98 Gb. A clear size distinction was observed between X chromosome-containing haplotypes (2.98–3.11 Gb) and Y chromosome-containing haplotypes (2.86–2.95 Gb) (Fig. 1a). Assembly continuity was evaluated using NG50, a metric defined as the contig length at which 50% of the genome is contained in contigs of this length or longer, with most assemblies exceeding 40 Mb. We next assessed assembly base-level accuracy (that is, the quality value (QV)) using Merqury[20] by comparing assembly *k*-mers against short-read *k*-mers. Although QVs were slightly underestimated based on the modest-coverage dataset (Supplementary Table 4 and Supplementary Fig. 8), the PIGA assemblies were estimated to have an average QV of 46, which is lower than the hifiasm assemblies (average QV of 54), but still indicative of a low error rate (Fig. 1b). Structural accuracy was evaluated using Inspector[21] on the three benchmark samples, which revealed an average of 31 structural errors per hifiasm assembly and 616 per PIGA assembly. We further assessed the regional reliability using Flagger, which revealed that 39% of *k*-mer errors and 48% of structural errors in PIGA assemblies occurred in unreliable regions, which were mostly located in satellites and segmental duplications (Fig. 1c and Supplementary Fig. 9). On the basis of 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).

We further assessed the ability of PIGA assemblies to capture genetic variation and to reconstruct haplotype sequences (Methods). PIGA variant calls demonstrated high concordance with the ground truth for both SNVs (average F1 score of 0.9881) and SVs (average F1 score of 0.9669) (Fig. 1d). The number of false-positive SVs closely matched the structural error counts detected by Inspector, a result that further supported the high accuracy of PIGA SV calls (Supplementary Table 5). Among SVs, 93.6% exhibited >95% sequence similarity to the benchmark SVs, which highlighted the accuracy of PIGA assemblies even in polymorphic SV regions (Supplementary Fig. 10). Compared with SNVs and SVs, indels showed a relatively lower concordance (average F1 score of 0.8576), with 90% of the inconsistencies occurring in homopolymer or TR regions (Supplementary Fig. 11). We also observed lower recall for heterozygous variants compared with homozygous variants (Supplementary Fig. 12). Nonetheless, PIGA assemblies outperformed methods based on read alignments for variant calling across multiple variant classes (Extended Data Fig. 3). When benchmarked with the hifiasm haplotypes, PIGA haplotypes exhibited a switch error rate of 0.50% and a correctly phased block N50 (defined as the block length such that 50% of the total phased sequence is contained in blocks of this size or larger) of 618 kb on average, which demonstrated the reliability of locally phased haplotypes (Supplementary Fig. 13).

We further annotated genomic elements across all the 1KCP assemblies, including repeats, genes and epigenomic elements. On average, 53.05% of each haploid assembly comprised repeat elements, such as long interspersed nuclear elements (LINEs), short interspersed nuclear elements (SINEs) and TRs (Supplementary Table 6). Genes were annotated using a combination of GENCODE transcript alignments and individual transcriptome assemblies (Methods). On average, 98.71% of autosomal protein-coding genes and 98.19% of autosomal non-coding genes from GENCODE were annotated per assembly (Fig. 1e and Supplementary Table 7). Furthermore, we identified an average of 77.9 protein-coding genes and 1,593.3 non-coding genes per assembly absent from the reference genome, a finding that reflected the genetic variations between individual assemblies and the reference genome (Supplementary Fig. 14). For epigenomic annotation, we used the SEI deep-learning model[22] to predict personal epigenome profiles from assembly sequences, as it has demonstrated enhanced performance and generalization capability in benchmarking against assay for transposase-accessible chromatin using sequencing (ATAC-seq) data from blood samples (Fig. 1f, Methods, Supplementary Note 2 and Supplementary Table 8). On average, 59.54% of each genome was predicted as epigenomic elements, such as promoters, enhancers and CCCTC-binding factor (CTCF)-binding sites (Supplementary Table 9).

Pangenome construction and annotation

To jointly represent and analyse the large-scale genome assemblies in a unified coordinate system, we constructed an integrated pangenome that incorporated the 1KCP, HPRC and CPC datasets, alongside the reference genomes GRCh38 and CHM13 (ref. 23) (Methods and Supplementary Table 10). From this integrated pangenome, the 1KCP pangenome was generated by extracting 2,232 1KCP analysis assemblies and 2 reference genomes. With a total size of 3.74 Gb, the 1KCP pangenome effectively collapsed redundant sequences across assemblies into non-redundant

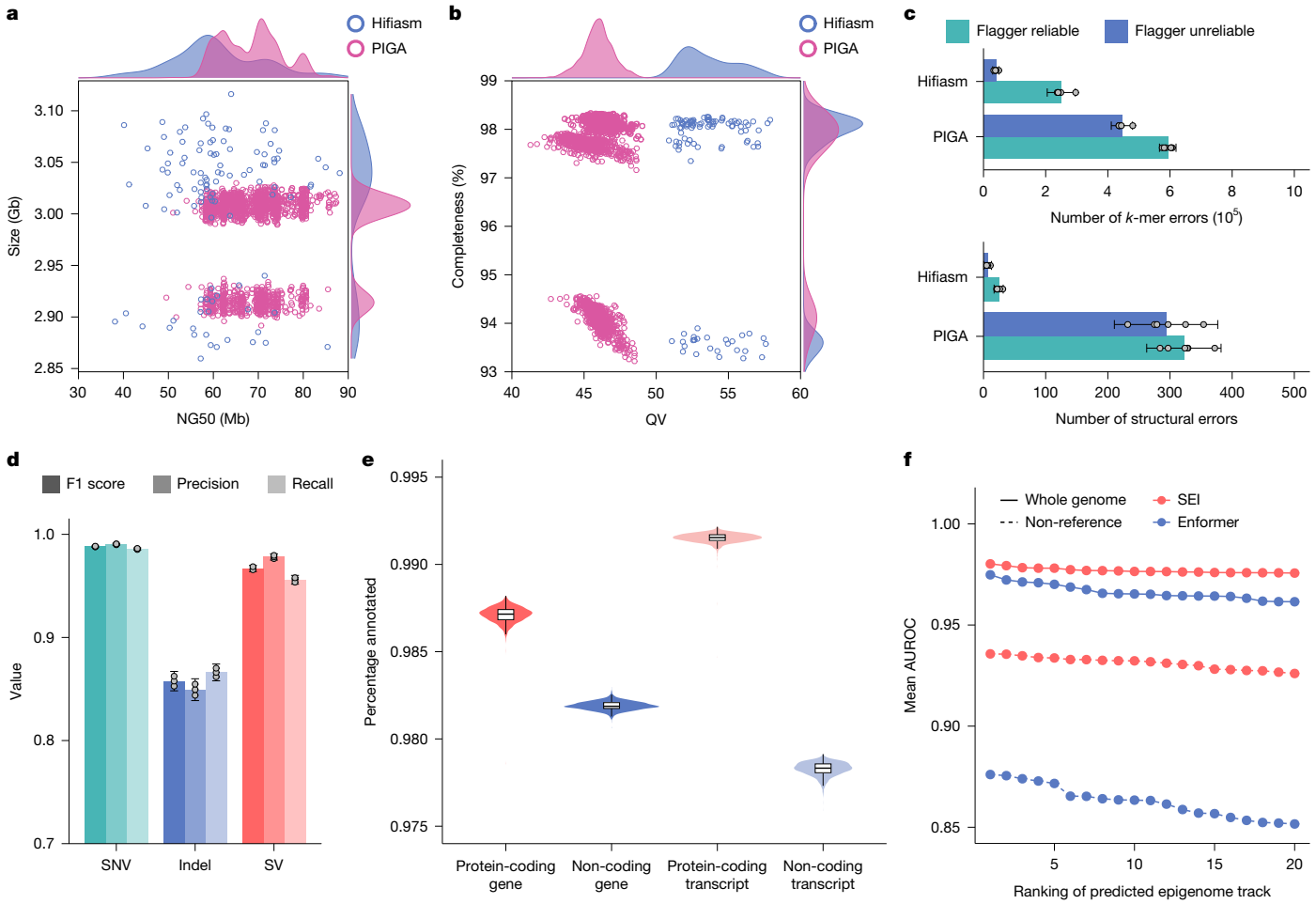


Fig. 1 | Evaluation and annotation of the 1,116 diploid genome assemblies. **a**, Distribution of assembly size and contig NG50 for hifiasm and PIGA assemblies. **b**, Distribution of *k*-mer completeness and QV measured using Merqury for hifiasm and PIGA assemblies. **c**, The number of *k*-mer errors and structural errors in hifiasm and PIGA haplotype assemblies ($n = 6$) for the three benchmark samples, stratified by regional reliability as estimated by Flagger. Reliable regions correspond to haploid regions, whereas unreliable regions include collapsed, duplicated and error regions. Each dot represents a haplotype assembly, bars indicate the mean value and error bars denote 95% confidence intervals. **d**, Variant calling performance of PIGA assemblies on three benchmark samples for SNVs, indels and SVs. Each dot represents an individual sample, bars indicate the mean value and error bars denote 95% confidence intervals. **e**, The percentage of GENCODE autosomal genes and transcripts annotated in the 1KCP assemblies ($n = 1,116$). 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. **f**, Epigenome track prediction performance of two deep-learning models: SEI and Enformer. For each predicted epigenome track, the mean area under the receiver operating characteristic curve (AUROC) was computed against blood ATAC-seq data separately for the entire 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.

segment representations while maintaining uniform coverage across the genome (Supplementary Note 3 and Supplementary Fig. 17).

Next, we quantified the sequences in the 1KCP pangenome that are absent from GRCh38 and CHM13. 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 the CPC and HPRC assemblies (Fig. 2a). Stratifying non-reference sequences by their frequency revealed 34.6 Mb as common (frequency of >0.05), 43.3 Mb as low frequency (frequency of >0.01 and ≤0.05), 142 Mb as rare (frequency of ≤0.01) and 185.4 Mb as singletons (unique to a single haplotype) (Fig. 2b). When analysing the growth of detected non-reference sequences with increasing sample size, we observed that 95% of common and low-frequency sequences could be detected with as few as 77 samples, whereas 95% of rare sequences required 776 samples. These results highlight the extensive non-reference sequences captured by the 1KCP pangenome.

Although previous studies[2,3] have also identified non-reference sequences from pangenomes, their functional potential remains poorly explored. To address this limitation, we developed a path-guided pangenome annotation method to identify genomic elements, including repeats, genes and epigenomic elements, in the 1KCP pangenome. This approach integrates assembly annotations into the pangenome using haplotype paths as a guide and generates concordant annotations for pangenome segments (Methods and Supplementary Fig. 18). Using this method, we annotated 146.6 Mb of non-reference sequences as genomic elements. Notably, TRs were disproportionately enriched in non-reference regions, a finding consistent with their high mutation rate[18], which probably drives the formation of non-reference sequences (Fig. 2c). We also 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-binding sites. These non-reference functional elements may have important biological implications and warrant further investigation.

A complex and extensive variant catalogue

In a pangenome graph, genetic variant sites are represented as bubbles[24], which can be identified using graph decomposition methods[2].

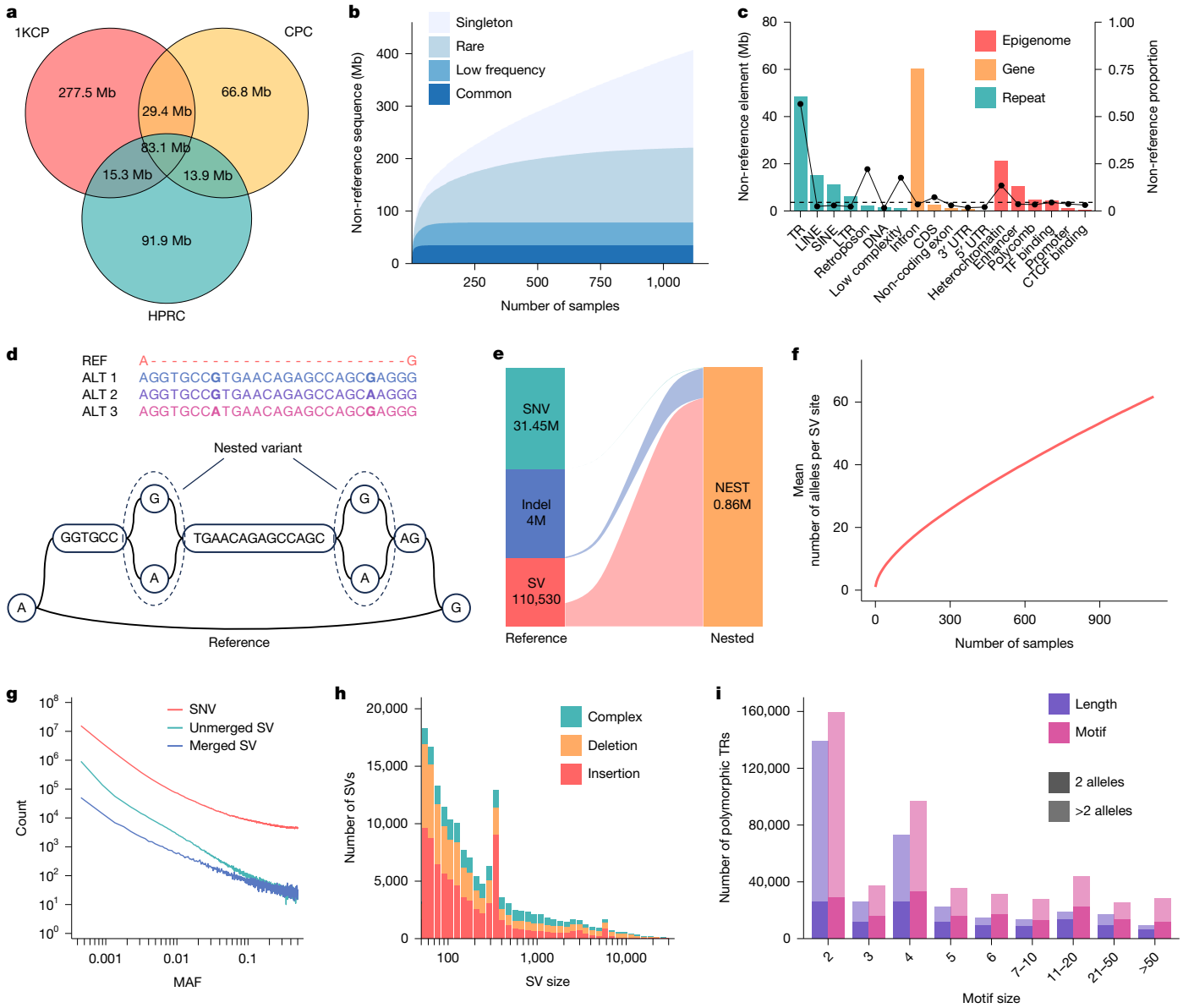


Fig. 2 | The 1KCP pangenome and variant catalogue. a, Comparison of non-reference sequence sizes among the 1KCP, CPC and HPRC assemblies. **b**, Cumulative size of non-reference sequences as the sample size increases. Non-reference sequences were stratified into four frequency classes: common (frequency >0.05), low frequency (frequency >0.01 and ≤0.05), rare (frequency ≤0.01) and singleton (unique to a single haplotype). **c**, Columns indicate the size of genomic elements in non-singleton non-reference sequences, whereas 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. **d**, Schematic of nested variants. The variant site contains three alternative alleles (ALT 1 to ALT 3) with insertion sequences compared with the reference (REF). Differences between these insertions (bases highlighted in bold) define two nested variants, which appear as bubbles in the pangenome. **e**, The number of reference and nested variant sites across variant types, with relative proportions $\log_{10}$-normalized. Flows illustrate the relationship between reference and nested variant sites. **f**, The mean observed number of unmerged alleles per SV site as the sample size increases. **g**, MAF spectrum of SNVs, raw SVs and merged SVs with MAF > 0.001. **h**, Size distribution of different SV types, including complex SVs, deletions and insertions. **i**, The number of polymorphic TR sites defined by length or motif variation, stratified by allele count. CDS, coding sequence; LTR, long terminal repeat; NEST, nested variant; TF, transcription factor.

Using GRCh38 as the reference genome, we detected 35.4 million small variant sites (<50 bp; see Methods for an estimated SNV false discovery rate (FDR) of around 8.9% using the transition-to-transversion (Ti/Tv) ratio)) and 110,530 SV sites (≥50 bp) in the 1KCP pangenome (Supplementary Table 11). In addition to these variant sites located in the reference genome (referred to as reference variants), the pangenome offers a distinct advantage in resolving variants nested in non-reference sequences (referred to as nested variants) through pangenome alignment (Fig. 2d). By identifying bubbles not traversed by reference paths, we resolved 0.86 million nested variant sites (Methods). Of these nested variants, 88.2% were embedded in reference SV sites. This result highlights the substantial genetic diversity encoded as nested variants in SVs, which is typically overlooked in previous coarse-grained SV analyses[25–27] (Fig. 2e and Supplementary Table 12). Among the nested variants in reference SV sites, 30.5% were located in TRs, which probably reflect TR expansion or contraction events.

We further explored the complexity of SV sites and found that 80.3% of SV sites are multiallelic, which complicates AF estimations of SVs for population-level analyses (Fig. 2f, Supplementary Note 4 and Supplementary Fig. 20). To address this issue, we used Truvari[28]

to merge similar alleles in the same SV sites, which reduced the mean allele number from 61.7 to 5.5. After decomposing the multiallelic sites into constituent biallelic variants, we obtained the 1KCP SV callset with 164,216 merged SV alleles (Methods). Compared with the unmerged SV alleles, the merged SV alleles exhibited a minor allele frequency (MAF) spectrum more similar to that of SNVs. This result demonstrates that SV merging better reflects the characteristics of SVs from the perspective of population genetics (Fig. 2g, Supplementary Table 13 and Supplementary Fig. 21). Moreover, 87% of the merged SV alleles with MAF > 0.01 were in Hardy–Weinberg equilibrium (HWE), a value comparable to the 88% observed among unmerged SVs. On the basis of the three benchmark samples, we estimated a FDR of 2.6% for the SV callset (Methods and Supplementary Fig. 22). This SV callset captured diverse SV types and sizes, including 77,456 insertions, 56,897 deletions and 29,863 complex SVs, with sizes ranging from 50 bp to 100 kb (Fig. 2h). The SV size distribution revealed two prominent peaks, which corresponded to the active transposable elements Alu (300 bp) and LINE1 (6,000 bp). Comparisons with existing cohort-based SV callsets (GnomAD[25], 1000G-ONT[26] and HGSVC3 (ref. 27)) revealed that 33.3% of SVs in the 1KCP callset are previously unreported (Extended Data Fig. 4a). Among these unreported SVs, 83.5% have an AF ≤ 0.01, a result that underscores the value of the 1KCP callset in capturing rare SVs (Extended Data Fig. 4b).

Given the extreme diversity and complexity of the TR sites, we further characterized alleles of 1.03 million candidate TR sites in the 1KCP assemblies on the basis of their repeat unit representations using vamos[29]. 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. By contrast, only 75.3% showed high concordance with hifiasm assemblies when directly called from long reads (Methods and Extended Data Fig. 5). In the 1KCP cohort, we identified 333,882 polymorphic TR sites with length variations. When further considering motif variation, this number increased to 485,575, with the proportion of multiallelic sites rising from 20.7% to 30.7% (Fig. 2i). These TR sites contained 1.47 million polymorphic motifs, 4.4 times the number of length-polymorphic TR sites. This finding highlights the substantial contribution of motif composition to TR variation.

In summary, we compiled an extensive variant catalogue for the 1KCP cohort that encompasses multiple variant types and a broad AF spectrum. From this catalogue, 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).

Medical relevance of multiscale variants

Leveraging the 1KCP dataset, we explored the gene structure and haplotype diversity in the cohort at various resolutions. We focused on medically relevant genes (MRGs), which are crucial for understanding disease and clinical interpretation.

We first identified 5,239 SVs in the exons of 3,326 protein-coding genes. These exon-overlapping SVs, capable of altering gene transcripts, exhibited evidence of purifying selection, as reflected by their higher proportion of rare SVs (AF ≤ 0.01) than of intergenic SVs (58.8%) and SVs in other genomic regions (Fig. 3a and Supplementary Table 16). On average, each individual carried 450 gene-altering SVs, including 9 rare ones (Supplementary Table 17). The deleteriousness of these gene-altering SVs depended on the evolutionary constraint of their affected genes, which was quantified by the loss-of-function observed/expected upper bound fraction (LOEUF)[30] (Spearman's correlation between the proportion of rare SVs in exons and the LOEUF scores of corresponding genes, $\rho = -0.85$, $P = 7.1 \times 10^{-6}$; Fig. 3b). Notably, we identified 1,013 SVs in the exons of 623 MRGs catalogued in the Online Mendelian Inheritance in Man (OMIM)[31] and COSMIC[32] databases, with an exceptionally high rare SV proportion of 74.6%. These included previously reported pathogenic SVs, such as a 340-bp Alu insertion in *PALB2*, which was found in two 1KCP individuals and associated with hereditary breast cancer (ClinVar: VCV001071843)[33], and a 100-bp deletion in *SLC34A3*, which was found in two 1KCP individuals and linked to hypophosphataemic bone disease (ClinVar: VCV000001433)[34] (Supplementary Fig. 23). These findings indicate the pathogenic potential of rare gene-altering SVs and emphasize the value of incorporating more comprehensive variant catalogues to improve clinical diagnostics.

TR expansion represents a pathogenic form of TR sites that is characterized by an abnormally long TR allele and is known to cause multiple disorders when occurring in certain MRGs. Using the 1KCP TR callset, we constructed a TR size 'null distribution' for each TR site, which enabled the identification of TR expansion events. We observed that most known disease-associated genic TR sites had normal size ranges (within 99% of 1KCP individuals) below previously defined pathogenic size thresholds[35]. Exceptions were for TRs in *DAB1* and *FGF14*, which exhibited frequent expansions of benign motifs (Fig. 3c and Extended Data Fig. 6a,b). This result is consistent with previous studies showing that the pathogenicity of TR expansions depends on both length and motif[36,37]. Moreover, we confirmed motifs previously associated with expanded TR alleles at the *RFC1* locus, including ACGGG, AAAGG and AAGGG (Extended Data Fig. 6c). To extend the scope of our investigation, we performed an unbiased TR expansion screen using the DBSCAN algorithm[38], which led us to identify 2,427 TR expansions, of which 124 were located in exons (Methods and Supplementary Table 18). On average, each individual carried 3.79 TR expansions, with 0.15 located in exons (Supplementary Table 17). Notably, we identified a TR expansion in *C11orf80* that induces the rare fragile site FRA11A, which is associated with increased genomic instability[39] (Supplementary Fig. 24). This result suggests that the detected TR expansions may explain the molecular basis of cytogenetically identified fragile sites. Further exploration revealed 9 GC-rich or AT-rich genic TR expansions located in 3 out of the 17 remaining unmapped rare fragile sites (Supplementary Table 19). Among these events, a CGG expansion in the 5′ untranslated region (UTR) of *GIPC1*, carried by one individual, exhibited allele-specific hypermethylation. This result is consistent with characteristics of folate-sensitive fragile sites and warrants further investigation[40] (Fig. 3d,e).

We then broadened our analysis from single genes to gene clusters, which often exhibit complex structural haplotypes that involve multiple SVs. Using pangene[41], we identified 735 gene clusters containing gene variations, including gene duplication, deletion and rearrangement (Supplementary Table 20). These gene clusters encompassed 2,038 genes, which are significantly enriched in blood and immune-related gene ontology (GO) categories, including immunoglobulin complex (GO:0019814, fold enrichment = 5.6, $P = 2.8 \times 10^{-39}$), T cell receptor complex (GO:0042101, fold enrichment = 6.6, $P = 7.3 \times 10^{-46}$) and blood microparticle (GO:0072562, fold enrichment = 2.8, $P = 8.9 \times 10^{-8}$) (Supplementary Table 21). Further analysis revealed that these enrichments are unlikely to be attributable to somatic V(D)J recombination (Supplementary Fig. 25). The high structural diversity of blood and immune-related gene clusters reflects their specific role in shaping human traits and enabling human adaptation to diverse pathogens throughout evolution. A prominent example is the *HP* gene cluster, comprising the *HP* (haptoglobin) and *HPR* (haptoglobin-related) genes, which are duplicated from a single ancestral gene and known for their role in binding free haemoglobin during acute-phase responses[42]. Based on the exonic SVs, we resolved four structural haplotypes of the *HP* gene cluster in the 1KCP pangenome, including three common haplotypes (AF > 0.01) and one rare haplotype (AF ≤ 0.01) (Fig. 3f,g). 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

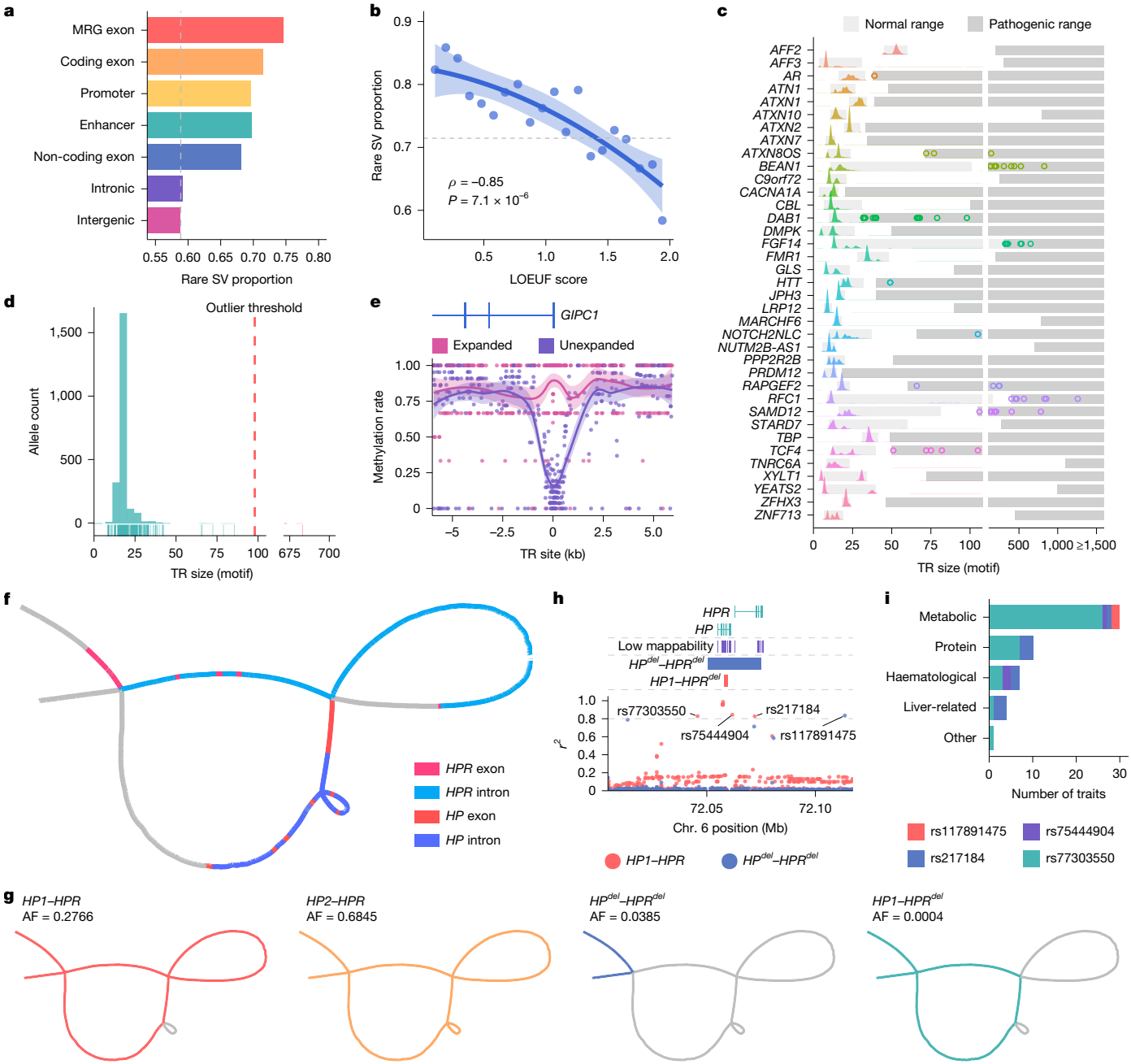


Fig. 3 | Multiscale genic variation landscape. a, Proportion of rare SVs (AF ≤ 0.01) across functional genomic regions. **b**, Relationship between the proportion of rare SVs located in exons and the LOEUF scores of the corresponding genes. Genes are binned into 20 bins on the basis of their LOEUF scores. Correlation was assessed using the two-sided Spearman's correlation test. The solid line shows a second-order polynomial regression fit, and the shaded area indicates the 95% confidence interval. The dashed line indicates the proportion of rare SVs in all gene exons. **c**, 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, whereas dark grey boxes denote the pathogenic TR size range derived from the STRchive[35] database. Circles indicate TR alleles with sizes exceeding the defined TR pathogenic threshold. **d**, 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. **e**, Methylation rates around the *GIPC1* TR site for expanded and unexpanded alleles in the individual carrying the expansion. Solid lines indicate the smoothed (LOESS) methylation rates, and the shaded area indicates the 95% confidence interval. **f**, The subgraph of the *HP* gene cluster in the 1KCP pangenome. Coloured segments indicate regions traversed by *HP* and *HPR*. **g**, Structural haplotype paths of the *HP* gene cluster. Coloured segments indicate regions traversed by each haplotype. **h**, The LD relationships between the structural haplotypes of the *HP* gene cluster and nearby SNVs. **i**, The number of traits associated with SNVs in strong LD with structural haplotypes of the *HP* gene cluster.

total cholesterol levels[43]. Two additional deletions were identified: a 27-kb deletion (AF = 0.0385, allele count = 86) generating the HP^{del}–HPR^{del} haplotype; and an 11-kb deletion (AF = 0.0004, allele count = 1) generating the *HP2*–HPR^{del} haplotype. In the 1KCP cohort, the *HP1*–*HPR* deletion was in strong linkage disequilibrium (LD) ($r^2 > 0.8$) with 3 SNVs associated with 47 traits across multiple categories (Fig. 3h,i and Supplementary Table 22). This result suggests that there are broader implications of the *HP* gene for human health than previously recognized. Although several SNVs showed stronger LD with the *HP1*–*HPR* deletion than the trait-associated SNVs, they are located in low-mappability

regions for short reads, which made them unreliable for genotyping in previous genome-wide association studies (GWASs). Moreover, the HP^{del}–HPR^{del} deletion showed strong LD with rs117891475, a SNV specific to East Asian individuals in the 1000 Genomes dataset[5], which has been associated with LDL and total cholesterol levels in a GWAS of East Asian individuals[44] (Supplementary Table 23).

Finally, we examined the haplotype structure of HLA genes in the major histocompatibility complex (MHC) region, which is recognized for its extraordinary genetic diversity and associations with numerous diseases[45]. Compared with previous studies[46,47] that used 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 non-coding regions of HLA genes, which enabled four-field resolution of HLA alleles. Using HiFiHLA, we identified HLA alleles for 36 genes. 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[48] (Supplementary Table 24 and Supplementary Fig. 26). In total, we identified 1,348 four-field alleles, more than double the number of identified three-field alleles, which highlighted the substantial non-coding diversity of HLA genes. Classical HLA genes, such as *HLA-A*, *HLA-B* and *HLA-C*, exhibited a higher number of four-field alleles than non-classical HLA genes (103.6 compared with 18.5 alleles on average). However, exceptions occurred for cases in which the non-classical HLA genes *MICA* and *MICB* showed greater four-field allele diversity than the classical HLA gene *HLA-DPA1* (Extended Data Fig. 7a,b and Supplementary Table 25). 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[49]. Values of ε calculated from four-field alleles were generally higher than those from three-field alleles (0.0385 compared with 0.0235 on average), which indicated that the higher allelic resolution reduces the randomness of allele combinations and reveals a finer-scale LD structure that is masked at lower resolution. (Extended Data Fig. 7c,d and Supplementary Table 26). Based on gene-wise LD, we identified five distinct LD blocks for HLA genes ($\varepsilon > 0.10$), including two HLA class I blocks and three HLA class II blocks.

Pan-variant eQTL analysis

The 1KCP dataset provides an extensive variant catalogue that contains not only small variants (SNVs and indels) but also complex variants (SVs, TRs and nested variants) that were poorly detected in previous short-read datasets. 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[50–52] that focused on only one or a few variant types. Consequently, a more comprehensive view of how diverse variants contribute to phenotypic variation can be obtained. 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 non-coding genes) quantified by RNA sequencing (RNA-seq) in the 1KCP cohort ($n = 1{,}101$) as example phenotypes and investigated their genetic regulation by different variant types (Methods).

We first assessed the contribution of different variant types to *cis*-heritability (the variance explained by all genetic variants in a 1 Mb window around the transcription start site (TSS)) of gene expression phenotypes using a multicomponent GREML model in GCTA[53–55]. Complex variants explained an average of 12.6% of the *cis*-heritability for genes with *cis*-heritability of >0.05 ($n = 7{,}341$), which highlighted their contribution to variation in gene expression (Extended Data Fig. 8). To further explore the regulatory impact of complex variants, we performed *cis*-eQTL mapping and identified 15,722 eGenes significantly associated with 2.68 million unique eVariants (FDR < 5%). Fine-mapping analysis produced 17,563 credible sets, of which 1,364 contained complex variants with the highest posterior inclusion probability (PIP) (Supplementary Table 27). Among the eVariants with PIP > 0.9, TRs exhibited the strongest enrichment across variant types (fold enrichment = 2.96; Supplementary Table 28). Stepwise conditional analysis revealed 32,154 jointly associated eVariants, including 3,256 that were lead complex eVariants (Fig. 4a and Supplementary Table 29).

Although only 115 lead eVariants were SVs, an additional 1,744 lead eVariants were located in SV sites, including 556 SNVs, 696 indels and 492 nested variants. In total, we identified 1,591 SV sites that contain lead eVariants, of which 162 had multiple unique lead eVariants. This finding indicates the complexity of SV-mediated gene regulation (Supplementary Table 30). For instance, we identified a SV site with inserted transposable elements and enhancers that contained eight nested variants undetectable using the linear reference GRCh38 (Fig. 4b). Among these nested variants, two located in enhancers were lead eVariants for *ADAM10* ($P = 5.6 \times 10^{-112}$) and *MINDY2* ($P = 2.5 \times 10^{-69}$). Another example involved a repetitive SV site that contained two lead eNested variants for *SNORA71B* ($P = 4.4 \times 10^{-21}$) and ENSG00000275401 ($P = 4.4 \times 10^{-21}$; Supplementary Fig. 27). These results highlight the regulatory potential of nested variants in gene expression. Using the 1KCP pangenome annotation, we further characterized the functional properties of lead eNested variants based on their pangenome contexts (Methods). These variants were enriched in genic regions (5′ UTRs, 3′ UTRs and non-coding exons) and regulatory elements (promoters and enhancers), which underscores the functional importance of non-reference sequences (Fig. 4c and Supplementary Table 31). Moreover, SINEs and long terminal repeats exhibited modest enrichment, which highlights their unique regulatory roles compared with other repeat elements.

We identified 2,482 lead eVariants as TR variants, including 1,047 eTR length variants and 1,435 eTR motif variants. Although 82.4% (860 out of 1,047) of eQTLs with lead eTR length variants were also significantly detected by TR motif variants, only 42.6% (612 out of 1,435) of eQTLs with lead eTR motif variants were identified by TR length variants. This result emphasizes the unique genetic variation captured by TR motifs. Among the 1,934 TR sites with lead eVariants, 305 contained multiple unique lead eTR variants, which suggests 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 (CCCAGGCCTCACCCTCCTCTTC CTCCCC). Specifically, each additional TR copy was associated with a 0.126 standard deviation increase in normalized *MAD1L1* expression, whereas an additional motif copy was associated with a 0.245 standard deviation increase (Fig. 4d). Overall, this TR locus explained 81% of *MAD1L1* expression variation (Supplementary Table 32).

Although we demonstrated the contribution of complex variants to phenotypic variation in gene expression, most GWASs incorporate only SNVs and indels, which limits their ability to uncover the phenotypic impact of complex variants on complex traits. Therefore, we conducted eQTL–GWAS colocalization analysis using SMR[56] and coloc[57], leveraging the 1KCP eQTL dataset with a more complete variant profile to prioritize variants that potentially underlie GWAS loci. Using GWAS data of 215 traits from BioBank Japan[58], we identified 1,563 eQTL–GWAS colocalization signals, of which 119 signals involved complex variants as lead eVariants (Supplementary Table 33). For example, we identified an 18-kb deletion spanning *GSTM1*, which is the most likely causal eVariant for both *GSTM1* and *GSTM2* based on fine-mapping ($PIP_{SuSiE\text{-}GSTM1} = 1$, $PIP_{SuSiE\text{-}GSTM2} = 1$). These two eQTLs colocalized with a GWAS locus for platelet count ($PP4_{coloc\text{-}GSTM1} = 0.90$, $PP4_{coloc\text{-}GSTM2} = 0.90$, $P_{SMR\text{-}GSTM1} = 4.2 \times 10^{-9}$, $P_{SMR\text{-}GSTM2} = 4.06 \times 10^{-9}$) (Fig. 4e). This result suggests that this deletion might affect platelet count by reducing *GSTM1* and *GSTM2* expression.

Overall, these results demonstrate that pan-variant eQTL analysis reveals previously undetected complex variants that regulate gene expression and provide new insights into the genetic basis of complex traits.

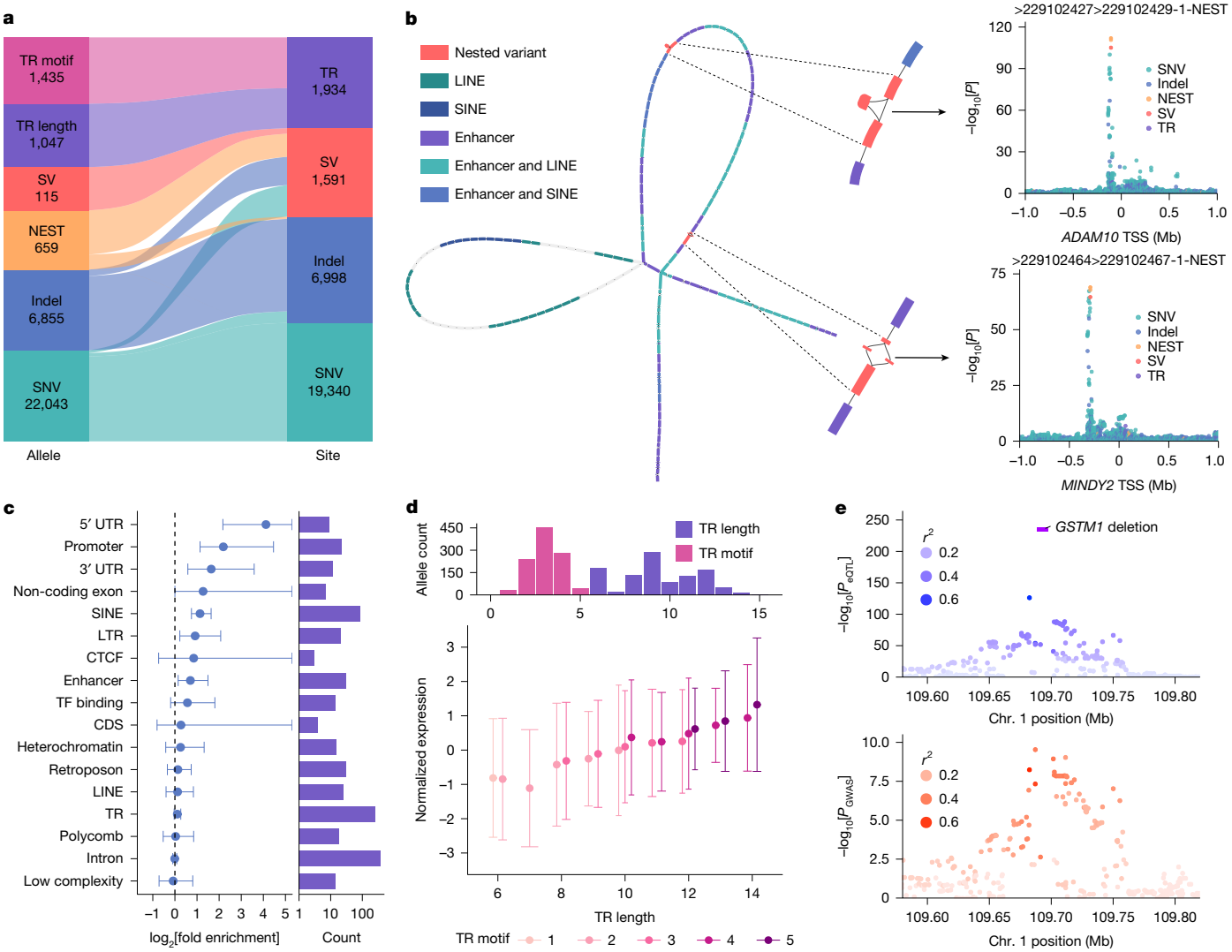


Fig. 4 | Pan-variant eQTL analysis. a, The number of lead eVariants and their corresponding variant sites across variant types, with relative proportions $\log_{10}$-normalized. Flows illustrate the relationship between variant alleles and variant sites. **b**, The subgraph of a SV site and its two lead eNested variants. Coloured segments indicate annotated genomic elements. The locus plots show the eQTLs associated with the expression level of *ADAM10* and *MINDY2*, with the text at the top indicating the lead eVariant name. eQTL nominal *P* values were computed using two-sided *t*-tests from linear regression models (tensorQTL). **c**, The number and fold enrichment of lead eNested variants in genomic elements of non-reference sequences. Dots represent mean values and error bars indicate 95% confidence intervals, estimated from 100,000 resampling iterations. **d**, The size distribution of the total length of a TR locus and the length of its 28-bp motif CCCAGGCCTCACCCTCCTCTTCCTCCCC, and their association with the normalized expression level of *MAD1L1* estimated using RNA-seq data ($n = 1{,}011$). Dots represent mean values and error bars indicate 95% confidence intervals. **e**, Colocalization between an eQTL signal for *GSTM1* and a GWAS signal for platelet count, which is probably driven by an 18-kb *GSTM1* deletion. eQTL nominal *P* values were computed using two-sided *t*-tests from linear regression models (tensorQTL). CDS, coding sequence; LTR, long terminal repeat; NEST, nested variant; TF, transcription factor.

1KCP pan-variant imputation panel

Leveraging the 1KCP assemblies, which reconstruct individual haplotypes with diverse variant types, we developed the 1KCP pan-variant imputation reference panel. This panel encompasses an extensive set of genetic markers, including 26.3 million small variants, 100,989 SVs, 1.48 million nested variants, 162,474 TR length variants, 357,416 TR motif variants and 1,871 HLA alleles (from one-field to four-field resolution). To facilitate pan-variant analyses in future genetic association studies, we developed the 1KCP data portal with an online imputation tool.

To comprehensively evaluate the imputation performance of the 1KCP reference panel, we conducted leave-one-out imputation for the 55 high-coverage samples. We achieved high accuracy across diverse variant types, including SNVs, SVs, nested variants, TRs and HLA alleles (Fig. 5, Methods, Supplementary Note 5 and Supplementary Table 34). Benchmarking against existing panels on external samples demonstrated that the 1KCP panel provided enhanced or comparable accuracy for SVs, TRs and HLA alleles while also enabling the imputation of nested variants, TR motif compositions and high-resolution (three-field and four-field) HLA alleles (Methods, Extended Data Fig. 9, Supplementary Note 6 and Supplementary Table 35). Taken together, these results highlight the value of the 1KCP imputation reference panel in enabling access to previously inaccessible variants, which will expand the scope of future genetic studies.

Discussion

In this study, we generated a large-scale diploid assembly panel of 1,116 Chinese individuals and constructed the 1KCP pangenome, which captured the extensive genetic diversity. We provided an extensive catalogue of genetic variants, including small variants, SVs, TRs and

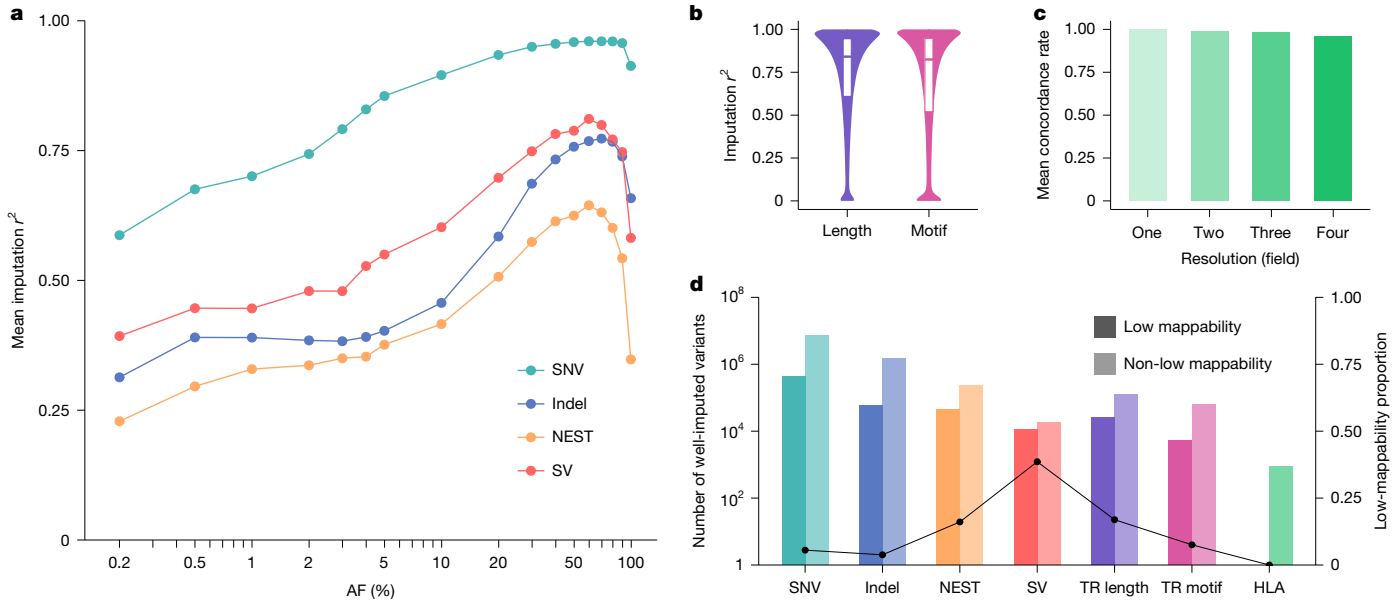


Fig. 5 | 1KCP imputation panel. a, Imputation performance (r^2) for different variant types stratified by AF bins. **b**, Imputation performance (r^2) for TR length and motif variants. In the box plots, the centre line represents the median and the box limits indicate the 25th and 75th percentiles. **c**, 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. NEST, nested variant.

nested variants, encompassing both common and rare alleles, all of which can be browsed through user-friendly tools at the 1KCP data portal. Leveraging this dataset, we characterized medically relevant genic variations across multiple scales, from gene-altering SVs and TR expansions to gene cluster variations and HLA gene haplotypes. Furthermore, we conducted pan-variant eQTL analyses, which uncovered the regulatory complexity mediated by complex variants and their potential contributions to complex traits. Finally, we developed the 1KCP imputation reference panel, which incorporates a full spectrum of genetic variants to enhance the resolution of genetic-association studies at a pan-variant level. Collectively, these efforts demonstrate the potential impact of the 1KCP project in advancing human genetics and precision medicine.

The 1KCP pangenome, comprising over 1,000 individuals, represents a substantial increase in sample size compared with previously released human pangenomes constructed from diploid assemblies, which typically included only a few dozen individuals. 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 that use either hybrid or long-read-only sequencing strategies with modest or low coverage. This large-scale pangenome demonstrates its value by greatly enhancing the representation of genetic variants, particularly rare and low-frequency variants (Supplementary Fig. 28a). 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, which accounted for a substantial proportion of multiallelic variants, a finding that underscores the necessity of representing and imputing TRs using a large-scale haplotype panel (Supplementary Fig. 28b). Moreover, the 1KCP pangenome revealed substantial non-reference sequences with functional genic and regulatory elements. However, we acknowledge the current limitations of deep-learning models in consistently predicting epigenomic profiles across assemblies and tissue contexts (Supplementary Note 7). Finally, the 1KCP variant catalogue highlights the functional complexity of complex variants, which highlights the necessity of developing advanced multiallelic frameworks to fully capture their phenotypic effects (Supplementary Note 8).

We acknowledge certain limitations inherent to the current PIGA workflow and sequencing strategies, particularly regarding indel errors, chromosome-level phasing and the resolution of highly repetitive regions (Supplementary Note 9). Despite these limitations, the 1KCP dataset provides a valuable resource for advancing genetic research and clinical applications. For clinical diagnosis, the 1KCP variant catalogue serves as a reference for assessing the functional impact of SVs and TRs to aid the identification of pathogenic variants that underlie inherited diseases. For genetic-association studies, we developed the 1KCP pan-variant imputation reference panel to enhance the resolution of future analyses, particularly for complex variants, to enable deeper insights into their contributions to human traits and diseases. We also built the 1KCP data portal with user-friendly tools for querying, visualization and imputation to enhance the accessibility and utility of the 1KCP dataset. This work demonstrates the potential of assembly-based analysis to bridge gaps in previous short-read-based studies. As the costs of long-read sequencing continue to fall, we anticipate that this strategy will be increasingly adopted in large cohorts to advance precision medicine and benefit human health[59].

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-026-10315-y.

1. Sherman, R. M. & Salzberg, S. L. Pan-genomics in the human genome era. *Nat. Rev. Genet.* **21**, 243–254 (2020).
2. Liao, W.-W. et al. A draft human pangenome reference. *Nature* **617**, 312–324 (2023).
3. Gao, Y. et al. A pangenome reference of 36 Chinese populations. *Nature* **619**, 112–121 (2023).
4. International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
5. Byrska-Bishop, M. et al. High-coverage whole-genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios. *Cell* **185**, 3426–3440 (2022).
6. Taliun, D. et al. Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *Nature* **590**, 290–299 (2021).

7. Halldorsson, B. V. et al. The sequences of 150,119 genomes in the UK Biobank. *Nature* **607**, 732–740 (2022).
8. The All of Us Research Program Genomics Investigators. Genomic data in the All of Us research program. *Nature* **627**, 340–346 (2024).
9. Chaisson, M. J. et al. Multi-platform discovery of haplotype-resolved structural variation in human genomes. *Nat. Commun.* **10**, 1784 (2019).
10. Chaisson, M. J., Sulovari, A., Valdmanis, P. N., Miller, D. E. & Eichler, E. E. Advances in the discovery and analyses of human tandem repeats. *Emerg. Top. Life Sci.* **7**, 361–381 (2023).
11. Wenger, A. M. et al. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nat. Biotechnol.* **37**, 1155–1162 (2019).
12. Cheng, H., Concepcion, G. T., Feng, X., Zhang, H. & Li, H. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat. Methods* **18**, 170–175 (2021).
13. Ebert, P. et al. Haplotype-resolved diverse human genomes and integrated analysis of structural variation. *Science* **372**, eabf7117 (2021).
14. Eizenga, J. M. et al. Pangenome graphs. *Annu. Rev. Genomics Hum. Genet.* **21**, 139–162 (2020).
15. Kircher, M. et al. A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.* **46**, 310–315 (2014).
16. Fiziev, P. P. et al. Rare penetrant mutations confer severe risk of common diseases. *Science* **380**, eabo1131 (2023).
17. Cipriani, V. et al. Rare disease gene association discovery in the 100,000 Genomes Project. *Nature* https://doi.org/10.1038/s41586-025-08623-w (2025).
18. Porubsky, D. et al. Human de novo mutation rates from a four-generation pedigree reference. *Nature* **643**, 427–436 (2025).
19. Wang, Y., Ding, Y., Duan, Z. & Jian, Y. Pangenome-informed genome assembly (PIGA). *Github* https://github.com/JianYang-Lab/PIGA (2025).
20. Rhie, A., Walenz, B. P., Koren, S. & Phillippy, A. M. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biol.* **21**, 245 (2020).
21. Chen, Y., Zhang, Y., Wang, A. Y., Gao, M. & Chong, Z. Accurate long-read de novo assembly evaluation with Inspector. *Genome Biol.* **22**, 312 (2021).
22. Chen, K. M., Wong, A. K., Troyanskaya, O. G. & Zhou, J. A sequence-based global map of regulatory activity for deciphering human genetics. *Nat. Genet.* **54**, 940–949 (2022).
23. Nurk, S. et al. The complete sequence of a human genome. *Science* **376**, 44–53 (2022).
24. Paten, B. et al. Superbubbles, ultrabubbles, and cacti. *J. Comput. Biol.* **25**, 649–663 (2018).
25. Collins, R. L. et al. A structural variation reference for medical and population genetics. *Nature* **581**, 444–451 (2020).
26. Schloissnig, S. et al. Structural variation in 1,019 diverse humans based on long-read sequencing. *Nature* **644**, 442–452 (2025).
27. Logsdon, G. A. et al. Complex genetic variation in nearly complete human genomes. *Nature* **644**, 430–441 (2025).
28. English, A. C., Menon, V. K., Gibbs, R. A., Metcalf, G. A. & Sedlazeck, F. J. Truvari: refined structural variant comparison preserves allelic diversity. *Genome Biol.* **23**, 271 (2022).
29. Ren, J., Gu, B. & Chaisson, M. J. vamos: variable-number tandem repeats annotation using efficient motif sets. *Genome Biol.* **24**, 175 (2023).
30. Karczewski, K. J. et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).
31. Amberger, J. S., Bocchini, C. A., Schiettecatte, F., Scott, A. F. & Hamosh, A. OMIM.org: Online Mendelian Inheritance in Man (OMIM®), an online catalog of human genes and genetic disorders. *Nucleic Acids Res.* **43**, D789–D798 (2015).
32. Bamford, S. et al. The COSMIC (Catalogue of Somatic Mutations in Cancer) database and website. *Br. J. Cancer* **91**, 355–358 (2004).
33. Rahman, N. et al. *PALB2*, which encodes a BRCA2-interacting protein, is a breast cancer susceptibility gene. *Nat. Genet.* **39**, 165–167 (2007).
34. Ichikawa, S. et al. Intronic deletions in the *SLC34A3* gene cause hereditary hypophosphatemic rickets with hypercalciuria. *J. Clin. Endocrinol. Metab.* **91**, 4022–4027 (2006).
35. Hiatt, L. et al. STRchive: a dynamic resource detailing population-level and locus-specific insights at tandem repeat disease loci. *Genome Med.* **17**, 29 (2025).
36. Seixas, A. I. et al. A pentanucleotide ATTTC repeat insertion in the non-coding region of DAB1, mapping to SCA37, causes spinocerebellar ataxia. *Am. J. Hum. Genet.* **101**, 87–103 (2017).
37. Mohren, L. et al. Identification and characterisation of pathogenic and non-pathogenic *FGF14* repeat expansions. *Nat. Commun.* **15**, 7665 (2024).
38. Trost, B. et al. Genome-wide detection of tandem DNA repeats that are expanded in autism. *Nature* **586**, 80–86 (2020).
39. Debacker, K. et al. The molecular basis of the folate-sensitive fragile site *FRA11A* at 11q13. *Cytogenet. Genome Res.* **119**, 9–14 (2007).
40. Mirceta, M., Shum, N., Schmidt, M. H. & Pearson, C. E. Fragile sites, chromosomal lesions, tandem repeats, and disease. *Front. Genet.* **13**, 985975 (2022).
41. Li, H., Marin, M. & Farhat, M. R. Exploring gene content with pangene graphs. *Bioinformatics* **40**, btae456 (2024).
42. di Masi, A. et al. Haptoglobin: from hemoglobin scavenging to human health. *Mol. Aspects Med.* **73**, 100851 (2020).
43. Boettger, L. M. et al. Recurring exon deletions in the HP (haptoglobin) gene contribute to lower blood cholesterol levels. *Nat. Genet.* **48**, 359–366 (2016).
44. Lee, C.-J. et al. Phenome-wide analysis of Taiwan Biobank reveals novel glycemia-related loci and genetic risks for diabetes. *Commun. Biol.* **5**, 1175 (2022).
45. Dendrou, C. A., Petersen, J., Rossjohn, J. & Fugger, L. HLA variation and disease. *Nat. Rev. Immunol.* **18**, 325–339 (2018).
46. Luo, Y. et al. A high-resolution HLA reference panel capturing global population diversity enables multi-ancestry fine-mapping in HIV host response. *Nat. Genet.* **53**, 1504–1516 (2021).
47. Hirata, J. et al. Genetic and phenotypic landscape of the major histocompatibilty complex region in the Japanese population. *Nat. Genet.* **51**, 470–480 (2019).
48. Kawaguchi, S., Higasa, K., Shimizu, M., Yamada, R. & Matsuda, F. HLA-HD: an accurate HLA typing algorithm for next-generation sequencing data. *Hum. Mut.* **38**, 788–797 (2017).
49. Okada, Y. eLD: entropy-based linkage disequilibrium index between multiallelic sites. *Hum. Genome Var.* **5**, 29 (2018).
50. GTEx Consortium. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* **369**, 1318–1330 (2020).
51. Chiang, C. et al. The impact of structural variation on human gene expression. *Nat. Genet.* **49**, 692–699 (2017).
52. Cui, Y. et al. Multi-omic quantitative trait loci link tandem repeat size variation to gene regulation in human brain. *Nat. Genet.* **57**, 369–378 (2025).
53. Yang, J., Lee, S. H., Goddard, M. E. & Visscher, P. M. GCTA: a tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.* **88**, 76–82 (2011).
54. Yang, J. et al. Common SNPs explain a large proportion of the heritability for human height. *Nat. Genet.* **42**, 565–569 (2010).
55. Yang, J. et al. Genome partitioning of genetic variation for complex traits using common SNPs. *Nat. Genet.* **43**, 519–525 (2011).
56. Zhu, Z. et al. Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nat. Genet.* **48**, 481–487 (2016).
57. Wallace, C. Eliciting priors and relaxing the single causal variant assumption in colocalisation analyses. *PLoS Genet.* **16**, e1008720 (2020).
58. Nagai, A. et al. Overview of the BioBank Japan Project: study design and profile. *J. Epidemiol.* **27**, S2–S8 (2017).
59. Porubsky, D. & Eichler, E. E. A 25-year odyssey of genomic technology advances and structural variant discovery. *Cell* **187**, 1024–1037 (2024).

Methods

Sample collection

The first phase of the 1KCP project enrolled 1,379 participants aged 18–76 years (737 men and 642 women) from the Health Examination Center of the Second Affiliated Hospital of Wenzhou Medical University. Peripheral blood samples were collected for library preparation and sequencing. All samples underwent Illumina short-read WGS and RNA-seq. Among these participants, 55 samples were selected for high-coverage PacBio HiFi long-read WGS and Hi-C sequencing, and 1,099 were selected for modest-coverage PacBio long-read WGS. Self-reported ancestry information was not explicitly collected and was therefore not available for this study. Informed consent was obtained from all participants. The study was approved by the Ethics Committee of Westlake University (approval no. 20201127YJ001) and the Ethics Committee of Wenzhou Medical University (approval no. 2019-107).

Short-read WGS

gDNA was extracted from peripheral blood using a gDNA extraction kit and a HF24 nucleic acid purification instrument (Concert). DNA concentration was measured using a Qubit 4.0 fluorometer (Thermo Fisher). A total of 1 μg of DNA was treated with ribonuclease A (Takara) and purified using CleanNGS-R beads (Vdobiotech). Fragmentation, end-repair and A-tailing were performed using 5× WGS fragmentation mix and 10× fragmentation buffer (Enzymatics), followed by ligation with TruSeq index adaptors using T4 DNA ligase and 5× rapid ligation buffer (Enzymatics). Purification, size selection and PCR amplification produced a sequencing library with a target size of 450–550 bp and normalized to 3 ng μl^{-1}. Short-read WGS was performed on an Illumina NovaSeq-6000 platform.

High-coverage long-read WGS

High-quality DNA samples (absorbance of $A_{260}/A_{280} = 1.8–2.0$, $A_{260}/A_{230} > 2.0$, concentration of ≥100 ng μl^{-1} and fragment size of ≥40 kb) were purified with Ampure PB beads and fragmented using a Megaruptor2 system (Diagenode). Libraries were prepared using a SMRTbell Prep kit 2.0 with steps similar to modest-coverage sequencing, targeting a library size of 15 kb. Long-read WGS was performed on a PacBio Sequel IIe platform using 8M SMRT Cell and Sequencing plates (v.2.0), with a pre-extension time of 4 h and a video time of 30 h. Primary CCS reads (accuracy >0.88) were generated using CCS (v.6.2.0; https://github.com/PacificBiosciences/ccs) with the --all option. Subreads were aligned to their corresponding primary CCS reads using actc (v.0.60; https://github.com/PacificBiosciences/actc). Finally, HiFi reads (accuracy >0.99) were generated using DeepConsensus (v.1.2.0)[60].

Modest-coverage long-read WGS

DNA quality was assessed using a Nanodrop spectrophotometer (Thermo Fisher), a Qubit 4.0 fluorometer (Thermo Fisher) and Mapper XA Chiller system (Bio-Rad). Samples meeting quality criteria ($A_{260}/A_{280} = 1.8–2.0$, $A_{260}/A_{230} > 2.0$, concentration of ≥70 ng μl^{-1} and fragment size of ≥40 kb) were purified with Ampure PB beads and fragmented using an Eppendorf centrifuge. Libraries were prepared using a SMRTbell Prep kit 2.0 (PacBio), which involved DNA damage repair, end-repair, adapter ligation and size selection. The libraries were then bound with Sequencing Primer v.3 and Bind Polymerase 2.0, followed by purification with 1.2× AMPure PB beads. Long-read WGS was performed on a PacBio Sequel IIe platform using 8M SMRT Cell and Sequencing plates v.2.0, with a pre-extension time of 4 h and a video time of 30 h. 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 (Extended Data Fig. 1a). This process produces two types of read: HiFi reads and non-HiFi reads. Although non-HiFi reads have lower accuracy than HiFi reads, their inclusion resulted in a 2.7-fold increase in total sequencing coverage compared with using HiFi reads alone (Extended Data Fig. 1b).

Hi-C sequencing

Peripheral blood samples were treated with 4 ml of 37% formaldehyde solution (Sigma-Aldrich) for crosslinking, followed by centrifugation to remove the supernatant. The pellets were washed twice with DPBS (Solarbio) and treated with a red blood cell lysis solution (Miltenyi) to remove erythrocytes. After additional washes with DPBS, the samples were resuspended in deionized water. gDNA was released from the suspension using 10% SDS and 10% Triton-X100 (Sigma), followed by digestion with the MboI enzyme (NEB). The resulting DNA fragments were end-repaired using Klenow fragment (Enzymatics) with Biotin-16-dCTP (Trilink) and ligated to adjacent fragments with T4 DNA ligase (Enzymatics). The ligated product was de-crosslinked using proteinase K (Qiagen) and 10% SDS and purified with SPRIselect beads (Beckman Coulter). DNA concentration was quantified using a Qubit 3.0 fluorometer (Thermo Fisher). Approximately 500 ng of purified DNA was subjected to fragmentation, end repair and A-tailing using 5× WGS fragmentation mix and 10× fragmentation buffer (Enzymatics). The fragmented DNA was ligated to Truseq index adaptors and purified using CleanNGS-R beads (Vdobiotech). Biotin-labelled target fragments were captured with M270 magnetic beads (Invitrogen) and amplified using KAPA HiFi HotStart ReadyMix (Kapa Biosystems). The PCR product was purified and normalized to 3 ng μl^{-1}, producing a sequencing library with a target size range of 500–600 bp. Hi-C sequencing was performed on an Illumina NovaSeq-6000 platform.

RNA-seq

Peripheral blood samples were treated with a red blood cell lysis solution (Tiangen) to remove erythrocytes. RNA was then extracted from the leukocyte precipitate using a PAXgene Blood RNA kit (Qiagen). The extracted RNA was assessed using an Agilent Bioanalyzer 2100, and only samples with a total RNA amount >1 μg and a RNA integrity number of >6.5 were used for library preparation. rRNA was depleted from the total RNA using a Ribo-Zero Plus rRNA Depletion kit (Illumina), leaving mRNA and long non-coding RNA intact for downstream processing. The sequencing library was prepared using a Hieff NGS Ultima Dual-mode mRNA Library Prep kit (Yeasen), which involved cDNA synthesis, end repair, A-tailing, adaptor ligation, purification, size selection and PCR amplification. The resulting PCR product was normalized to a concentration of 3 ng μl^{-1}, producing a library with a target insert size of 450 bp. RNA-seq was performed on an Illumina NovaSeq-6000 platform.

PCA

To place the 1KCP samples in a global genetic ancestry context, we performed a PCA using SNV genotypes from the 1KCP samples combined with unrelated samples from the 1000 Genomes Project (1000G) and the Human Genome Diversity Project[61]. For the 1KCP samples, short-read data were aligned using BWA-MEM (v.0.7.17)[62], and SNV genotypes were called using the GATK (v.4.2.6.1) workflow[63]. Genotypes at overlapping SNV sites across the three datasets were merged to create two analytical sets: (1) a global set with all samples, and (2) an East Asian-only set. For each set, we retained biallelic SNVs with MAF > 0.05, HWE $P > 1 \times 10^{-6}$ and genotype missing rate <10%. The remaining SNVs were further pruned to remove sites in LD using PLINK2 (ref. 64) (--indep-pairwise 50 5 0.5). PCA was then conducted on this pruned dataset using PLINK2.

Diploid genome assembly using the high-coverage dataset

High-coverage HiFi assemblies were generated using hifiasm (v.0.15.3)[12], incorporating Hi-C data for haplotype phasing[65]. To improve mitochondrial genome assembly, we removed contigs aligned to mitochondrial

DNA using minimap2 (v.2.24)[66] (-cxasm5) and reassembled the mitochondrial genome with Unicycler (v.0.5.0)[67].

To clean raw assemblies, PacBio adapter sequences were identified using minimap2 (-cxsr -f5000 -N2000 -secondary=yes) and masked with BEDTools (v.2.30.0)[68]. Non-human sequences (bacterial, viral, fungal and archaeal sequences) were detected by aligning to the NCBI RefSeq database using blastn (v.2.13.0)[69] (-task megablast -word_size 28 -evalue 0.0001 -perc_identity 98.0) and removed.

Diploid genome assembly using the modest-coverage dataset

We used the PIGA workflow[19], an integrated framework incorporating multiple tools[20,62,63,66,68,70–97], to generate diploid genome assemblies using population-scale hybrid sequencing data. A detailed step-by-step description of the PIGA workflow is provided in Supplementary Note 1. In brief, this workflow involves the following steps:

1) Performing population-level SNV calling by integrating long-read and short-read sequencing data.
2) Performing SNV haplotype phasing by leveraging long-read and population information.
3) Generating personalized references by modifying the reference genome with homozygous variants genotyped from the external pangenome and partitioning long reads into haplotypes.
4) Producing draft diploid assemblies based on partitioned long reads.
5) Constructing the pangenome using draft diploid assemblies and simplifying it.
6) Reconstructing the final diploid assemblies by inferring the diploid paths from the constructed pangenome.

Assembly evaluation

The quality of the 1KCP assemblies was evaluated using multiple approaches. We assessed assembly completeness and QV using Illumina short-read *k*-mers. A 21-mer database was built using Meryl (v.1.4.1), and completeness and QV of the assemblies were estimated using Merqury (v.1.3)[20]. Structural errors were detected using high-coverage HiFi reads with Inspector (v.1.0.1)[21]. Moreover, unreliable assembly regions in both PIGA and hifiasm assemblies of the three benchmark samples were identified using high-coverage HiFi reads with Flagger (v.0.2.0)[2]. 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 (v.0.1.34; https://github.com/mrvollger/rustybam) for coordinate transformation. Reference regions marked as unreliable in more than two haplotypes were defined as common unreliable regions.

Variant calling and haplotype phasing performance of the PIGA assemblies were evaluated using the three benchmark samples. Evaluation callsets were generated from PIGA assemblies with Dipcall (v.0.3)[98]. Benchmark small variant callsets were constructed by combining variant calls obtained independently using GATK, DeepVariant (v.1.4.0)[71] and Dipcall from high-coverage short-read, long-read and hifiasm assemblies. Variant genotypes supported by at least two methods were considered true. Small variant calling performance was evaluated by comparing the evaluation callset to the benchmark callset in Dipcall-confident regions using hap.py (v.0.3.14)[99] with the vcfeval engine. SV calling performance was evaluated by treating Dipcall SV calls from hifiasm assemblies as the benchmark set. The evaluation callset was compared with the benchmark set in Dipcall-confident regions using Truvari (v.4.2.2)[28] bench (--pick ac --sizemin 50), with harmonization performed using Truvari refine (--recount -u -U). Haplotype phasing performance was evaluated by comparing SNV haplotypes from Dipcall datasets of PIGA and hifiasm assemblies using WhatsHap (v.1.4)[73].

Assembly annotation

We developed an integrative assembly annotation pipeline to identify repeats, genes and epigenomic elements in the 1KCP assemblies. For repeat annotation, we used Tandem Repeats Finder (v.4.09)[100] (parameters: 2 7 7 80 10 50 15 -l 25 -h) to identify TRs and used RepeatMasker (v.4.1.1)[101] with RMBlast to annotate other types of repeat elements, such as LINEs, SINEs and long terminal repeats.

Gene annotation was carried out in multiple steps. First, we lifted the GENCODE (v.40)[102] annotation from GRCh38 to individual assemblies with a 95% identity threshold using Liftoff (v.1.6.3)[103] (-sc 0.95 -copies -polish). We then aligned RNA-seq reads to the assemblies using HISAT2 (v.2,2.1)[104], and performed transcriptome assembly with StringTie (v.2.2.1)[105], using Liftoff annotations as guidance. StringTie annotations were filtered to exclude single-exon transcripts or those with transcripts per million (TPM) < 0.5. The filtered StringTie annotations were compared with the Liftoff annotations using gffcompare (v.0.12.9)[106] to identify novel transcripts, including intronic, opposite-strand and non-overlapping transcripts. These novel transcripts were further classified as coding or non-coding using CPC2 (v.1.0.1)[107]. For novel coding transcripts, gene sequences were predicted with GeneMarkS-T (v.5.1)[108] and further validated using InterProScan (v.5.66)[109] with several databases (for example, CDD, Gene3D, PANTHER, Pfam and PIRSF). Novel non-coding transcripts, specifically long non-coding RNAs, were annotated if their length exceeded 200 bp and had no putative coding sequence longer than 100 bp. Finally, the novel coding and non-coding genes were merged with Liftoff annotations to create the final annotations.

To annotate epigenomic elements, we compared the performance of SEI and Enformer (v.0.8.8) for epigenome prediction in 128-bp steps on both PIGA and hifiasm assemblies from the three benchmark samples. Non-reference sequence regions in each assembly were identified using PAV[13] and defined as insertion records of ≥50 bp. ATAC-seq peaks were identified on the assemblies of the matched samples using MASC3 (v.3.0.0)[110] with a FDR < 0.01 and served as the ground truth for model evaluation. Predicted regions were labelled as true positive if they overlapped with the ground truth peaks and as false positive otherwise. The AUROC for each predicted track was calculated to evaluate the performance of SEI and Enformer. For large-scale annotation, SEI predictions were performed in 300-bp steps, balancing speed and accuracy (Supplementary Fig. 15). The SEI sequence classes were used to annotate epigenomic elements across all the 1KCP assemblies.

Pangenome construction

Given the computational challenges of integrating such a large dataset, we developed a pangenome construction pipeline that combines features from Minigraph-Cactus[111] and PGGB[91] while implementing subgraph parallelization to enhance efficiency (Supplementary Fig. 16). We first constructed the pangenome containing only SVs using Minigraph (v.0.21)[89], which iteratively integrates assemblies into the pangenome. The order of integration was determined on the basis of the *k*-mer-based distance of assemblies to the CHM13 reference, estimated using Mash (v.2.3)[112]. We then realigned the assemblies to the Minigraph pangenome to obtain assembly-to-graph alignments, which were filtered using gaffilter (https://github.com/ComparativeGenomicsToolkit/cactus-gfa-tools) (-r 5.0 -m 0.25 -q 5 -b 250000 -o 0 -i 0.5). To accelerate subsequent steps, we divided the pangenome into approximately 20 Mb subgraphs by removing edges in border regions not part of the bubbles, using the API of odgi (v.0.8.6)[113]. Alignments and assembly sequences were split accordingly for each subgraph. We then induced the base-level pangenomes for subgraphs using seqwish (v.0.7.4)[90] and refined them through three rounds of smoothxg (v.0.7.9)[91] with partial order alignment parameters for 0.1% divergence (-p 1,19,39,3,81,1) and target lengths of 1,400, 1,800 and 2,200. The pangenome was then normalized with GFAffix (v.0.1.5; https://github.com/marschall-lab/GFAffix) to collapse shared affixes. Complex regions of the pangenome were filtered by clipping assembly regions masked by dna-brnn (v.0.1; https://github.com/lh3/dna-nn) or unaligned by the Minigraph path, with a size greater than 100 kb.

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 GenomeScope (v.2.0.1)[114] 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 artefacts.

The integrated pangenome, which encompasses not only the 1KCP assemblies but also publicly available genome assemblies (HPRC and CPC) and reference genomes (GRCh38 and CHM13), was used for quality assessment. For further analysis, we created a 1KCP pangenome that contains only the 1KCP assemblies and reference genomes. To enhance the annotation resolution of the pangenome, segments were divided into lengths of ≤20 bp, as most functional elements exceed this size (Supplementary Fig. 19a).

Path-guided pangenome annotation

We developed a path-guided pangenome annotation method to detect genomic elements in pangenome segments. First, we mapped the pangenome segments to the assembly coordinates guided by the assembly paths. We then compared the overlap between segments and assembly annotations in each assembly. Segments with >80% overlap with specific genomic elements were annotated. The annotation results from all assemblies were integrated into the final pangenome annotation. To ensure robust annotations, we only annotated non-singleton segments and retained annotations supported by >80% of haplotypes (Supplementary Fig. 19b).

Variant detection

We identified reference variant sites by detecting allele traversals in pangenome bubbles using VG (v.1.59.0)[78] deconstruct. Variant sites with maximum allele sizes >100 kb or redundant representations were filtered using vcfbub (v.0.1.0; https://github.com/pangenome/vcfbub) (-a 100000 -l 0). The retained variant sites were then classified on the basis of the size difference between the longest and shortest alleles: SNVs (size difference = 0), indels (size difference ≥ 1 and <50) and SVs (size difference ≥ 50).

To simplify the representations of multiallelic reference variant sites, we decomposed them into constituent biallelic variants by clipping shared affixes between reference and alternative alleles using a modified script of PanGenie[97]. Small variant alleles were identified by directly decomposing the original variant sites. For detecting SV alleles, similar alleles (≥70% size and sequence similarity) from original SV sites were first merged using Truvari collapse (-k common -r 0 -p 0.7 -P 0.7) and then decomposed.

To identify nested variant sites, we discovered additional bubbles in a reference-free manner using VG snarls[24]. Variant sites not traversed by the reference allele were classified as nested variant sites. The alleles of nested variants were further resolved by extracting substrings of allele traversals from reference variant sites based on the flanking segments.

We characterized TR sites by analysing their motif sequences using vamos (v.2.1.5)[29] with the v.2.1 e0.1 motif set and the STRchive[35] disease-associated TR set. Single-base motifs were excluded, as they are commonly associated with long-read sequencing errors. The individual-level TR calls were then integrated into the population-level callset using tryvamos combineVCF.

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. The quality control steps reduced the error rate of our callset (Supplementary Fig. 22). For TRs, we also removed loci for which inferred sequences exhibited an average identity of <80% to the corresponding assembled sequences across haplotypes.

Variant assessment

We used a Ti/Tv ratio-based approach to estimate the FDR of the SNV callset. Using the gnomAD (v.4.1) 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, thereby 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 $\mathrm{FDR}_{\mathrm{novel}} = 3(x - y)/[(2x - 1)(y + 1)]$. 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 produced 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, which resulted in an estimated overall FDR of 8.9%.

For evaluating the SV callset, HWE *P* values were calculated for SV alleles using the fill-tags plugin in BCFtools (v.1.20)[70]. SVs with HWE $P > 1 \times 10^{-6}$ were considered to be in HWE. 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 (v.1.0.20)[77] and genotyped using VG call[79]. 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 (v.2.6.2)[86]) and the 1KCP short-read callset (generated with Manta (v.1.6.0)[115] and merged using SURVIVOR (v.1.0.6)[116]) (Supplementary Table 15), using Jasmine (v.1.1.5)[87] (--allow_intrasample).

For evaluating the TR callset, we also called TRs from short-read data using GangSTR (v.2.5)[117] and from long-read data using vamos. The PIGA and long-read TR callsets from the three benchmark samples were compared against the hifiasm TR callset by using edlib[118] 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 (v.1.0.14)[119] and benchmarked using Truvari bench (-s 5) and refine (-u) against the hifiasm Dipcall callset, which was restricted to GIAB-defined TR regions[120].

Variant annotation

For SV sites, the longest allele was considered the representative allele, and its sequence properties were characterized on the basis of the pangenome annotations. The proportion of each repeat type was calculated according to the cumulative size of segments annotated with that repeat. SV sites with more than 50% of their length annotated as a specific repeat were classified as specific repeats. SV sites with more than 50% annotated as multiple repeat types, but without a dominant repeat annotation, were classified as mixed repeats. SV sites with less than 50% annotated as repeats were labelled as non-repeats.

For nested variant sites, annotation was performed on the basis of their context in the pangenome. If a flanking segment was part of a specific genomic element, the nested variant site was considered to be located in that element.

Definition of MRG sets

Given that different medical genetic 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 catalogued in the OMIM[31] and COSMIC[32] databases. For TR expansion analysis, MRGs were defined as known disease-associated TR genes catalogued in the STRchive[35] database. For HLA allele analysis, MRGs were defined as genes catalogued in the IPD-IMGT/HLA[121] database.

TR expansion detection

As TR expansion events can be regarded as outliers in the TR size distribution in the population, we used the non-parametric outlier detection algorithm DBSCAN to identify TR expansions. Following the settings of a previous study[38], we used the $\log_2$ of the number of haplotypes as the minimum number of points (minPts) and the mode size of a TR site as the maximum distance (ε). TR alleles with sizes exceeding the longest clustered TR alleles were identified as TR expansion events.

Methylation rate estimation

To analyse the methylation status of expanded TR alleles, HiFi reads from samples with expanded alleles were mapped to the GRCh38 reference genome. Methylation status was estimated using MeLoDe[122], and haplotypes of HiFi reads were assigned with WhatsHap. Methylation rates for each CpG site around the TR sites were separately calculated for each haplotype.

Gene cluster variation detection

To identify gene cluster variations, we aligned the protein sequences of GENCODE canonical isoforms to the assemblies using miniport (v.0.12)[123] (--outs=0.97 --no-cs -Iu). Next, a gene graph was constructed using pangene (v.1.1)[41], and gene cluster variations were detected by identifying bubbles in the graph using the pangene.js call.

GO enrichment analysis

GO enrichment analysis was performed using the clusterProfiler (v.4.12.6)[124] on a given gene list. Redundant enriched GO terms were filtered using the GOSemSim (v.2.30.2)[125] with a similarity cut-off value of 0.7.

Pangenome visualization of complex loci

For the visualization of complex loci, subgraphs of regions of interest were generated based on the GRCh38 coordinates using odgi extract (-d 50000). These subgraphs were visualized using Bandage (v.0.9.0)[126], whereby segments were coloured according to pangenome annotations.

Comparison with the GWAS catalogue

Trait associations from the GWAS catalogue[127], downloaded on 15 October 2024, were linked to 258,641 unique variants. SV alleles in strong LD ($r^2 > 0.8$) with these trait-associated variants were identified using PLINK2 (v.2.0.0) for LD calculation.

HLA gene analysis

Four-field alleles of HLA genes were called from assemblies using HiFiHLA (v.0.3.1; https://github.com/PacificBiosciences/hifihla) with the 'call-contigs' option, and one-field to three-field alleles were derived from the four-field alleles. The 39 typed HLA genes were categorized into 8 classical HLA genes (*HLA-A*, *HLA-B*, *HLA-C*, *HLA-DRB1*, *HLA-DQA1*, *HLA-DQB1*, *HLA-DPA1* and *HLA-DPB1*) and 31 non-classical HLA genes.

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 did not pass these criteria. We also performed independent HLA typing from short reads using HLA-HD (v.1.7.1)[48] to validate the HiFiHLA typing results. HLA genes with concordance <95% were excluded for LD analysis.

To assess the haplotype structure of these HLA genes, we used eLD (v.1.0)[49] to calculate the entropy-based LD measurement index (ε), which reflects the multiallelic LD relationships between the HLA genes. The analysis was performed separately using three-field and four-field HLA alleles with AF > 0.05.

eQTL analysis

To quantify gene expression levels, RNA-seq reads were aligned to GRCh38 using STAR (v.2.7.9)[128] with GENCODE (v.40) annotation. Gene read counts and TPM values were generated with RNA-SeQC (v.2.4.2)[129], and samples with fewer than 10 million reads were excluded. Gene expression outlier samples were identified by calculating the average correlation of each sample with all others and labelling those with notably lower correlations[130] (median absolute deviations of <0.85). Genes with >0.1 TPM and ≥6 reads in at least 20% of samples were retained. Expression levels were TMM-normalized using edgeR (v.3.22.5)[131] and inverse-normal transformed. To account for the hidden factors influencing expression variation, probabilistic estimation of expression residuals[132] (PEER) was calculated. Following the GTEx Consortium guidelines[50], the number of PEER factors was selected based on sample size. A total of 60 PEER factors, along with sample age and sex, were used to adjust gene expression levels.

Non-TR variants with MAF > 0.01 and TRs with heterozygosity >0.98, both having a HWE $P > 1 \times 10^{-6}$, were retained. We included 8.03 million SNVs, 4.17 million indels, 56,242 SVs, 0.89 million nested variants with MAF > 1% and 0.46 million multiallelic TR variants (0.15 million TR length variants and 0.31 million TR motif variants) for the subsequent analysis. After excluding samples with genetic relatedness >0.05, 1,101 samples were included in the following analysis. *Cis*-eQTL mapping (within 1 Mb upstream and downstream of the TSS) was performed using tensorQTL (v.1.0.9)[133], with 20 genotype principal components (PCs) and assembly methods included as covariates. Significant eGenes were identified at a 5% FDR using adaptive permutation mode with 10,000 permutations[134]. Nominal *P* values were then computed to detect significant gene–variant associations. Fine-mapping analysis was performed using the SuSiE algorithm implemented in tensorQTL to identify credible sets of eQTL signals. Stepwise conditional regression, implemented in tensorQTL, was then used to identify independent eQTL signals for each eGene.

Heritability estimation

We used GCTA (v.1.94.1)[53–55] GREML to estimate the *cis*-heritability of gene expression, partitioned by variant type. Genetic relationship matrices (GRMs) for non-TR variant types (SNV, indel, SV and nested variant) were constructed using PLINK2. A GRM for TRs was constructed using dgrm (https://github.com/PeixiongYuan/dgrm). These five GRMs were fitted into a GREML multicomponent model, with the 20 genotype PCs and assembly methods as covariates. The parameter --reml-no-constrain was used to obtain unbiased heritability estimates.

Enrichment of eVariants

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.

To assess the enrichment of eNested variants in genomic elements, we performed 100,000 resampling iterations. For each resampling, an equal number of nested variants with matching MAF and distance to TSS as the eNested variants 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.

Colocalization analysis

We leveraged eQTLs of 15,699 significant eGenes in the 1KCP dataset and GWAS data for 215 traits from BioBank Japan to perform the eQTL–GWAS colocalization analysis. Genetic variants common to both eQTL and GWAS datasets were selected for further analysis. Colocalization analyses were conducted using the coloc (v.5.2.3)[57] (in ABF mode) and SMR (v.1.3.1)[56] with HEIDI test. eQTL–GWAS signal pairs were considered colocalized if they met either of the following conditions: (1) coloc PP4 > 0.9 or (2) $P_{SMR} < 3.18 \times 10^{-6}$ and $P_{HEIDI} > 0.05$.

Imputation panel construction and assessment

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 centre of the corresponding HLA genes. Variants with a missing rate >10%, HWE $P < 1 \times 10^{-6}$ or allele count <2 were excluded from the reference panel.

To evaluate imputation performance, we performed leave-one-out imputation using Minimac4 (v.4.1.6)[135], whereby each of the 55 high-coverage samples was sequentially excluded as the imputation target sample. The SNV sites used for imputation were restricted to those accessible by the Infinium Omni2.5 array or short-read WGS. For non-TR variants, the imputation accuracy was assessed by calculating the squared correlation between the true genotypes and the imputed dosages across individuals. For TRs, the imputed length and motif copy numbers at multiallelic loci were derived as dosage-weighted sums. Specifically, the length copy number was calculated as the sum of genotype dosage multiplied by the repeat unit count for each allele, whereas the motif copy number was calculated as the sum of genotype dosage multiplied by the count of specific motifs for each allele. The squared correlation was calculated by comparing the true and imputed TR copy numbers across individuals.

For comparison with the existing panels, we used 70 East Asian samples from CPC, HPRC and HGSVC3 to construct the truth callsets and then performed imputation benchmarking. The SV callset was generated using PAV and further merged using Jasmine. The HLA callset was generated using HiFiHLA. As 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 multiethnic HLA v.2 panels was performed on the Michigan online imputation server[135]. 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.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The release of the 1KCP data was approved by the Human Genetic Resources Administration of China (permission numbers 2025BAT00888 and 2025BAT00720). 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: https://ngdc.cncb.ac.cn), 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 following public datasets were used in this study: GRCh38 genome (https://gdc.cancer.gov/about-data/gdc-data-processing/gdc-reference-files); CHM13 genome (https://github.com/marbl/CHM13); HPRC v1 assemblies (https://github.com/human-pangenomics/HPP_Year1_Assemblies); CPC assemblies (https://github.com/Shuhua-Group/Chinese-Pangenome-Consortium-Phase-I); GENCODE annotation (https://www.gencodegenes.org/human); NCBI RefSeq database (https://ftp.ncbi.nlm.nih.gov/refseq/release); 1000G high-coverage Illumina variants (https://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/1000G_2504_high_coverage/working/20220422_3202_phased_SNV_INDEL_SV); 1000G-ONT SVs (https://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/1KG_ONT_VIENNA/release/v1.0); GnomAD variants (http://www.gnomad-sg.org); HGSVC3 SVs (https://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/HGSVC3/release/Variant_Calls/1.0); OMIM database (https://www.omim.org); COSMIC database (https://cancer.sanger.ac.uk/cosmic); STRchive database (https://strchive.org/resources); IPD-IMGT/HLA database (https://www.ebi.ac.uk/ipd/imgt/hla); GWAS catalogue (https://www.ebi.ac.uk/gwas); BioBank Japan GWAS (https://humandbs.dbcls.jp/en/hum0197-v23); and EnsembleTR panel (https://github.com/gymrek-lab/EnsembleTR).

Code availability

The analysis scripts used in this study are available at GitHub (https://github.com/JianYang-Lab/1KCP-analysis) and Zenodo (https://zenodo.org/records/18659653)[136]. The PIGA workflow is available at GitHub (https://github.com/JianYang-Lab/PIGA) and Zenodo (https://zenodo.org/records/18659653)[136].

60. Baid, G. et al. DeepConsensus improves the accuracy of sequences with a gap-aware sequence transformer. *Nat. Biotechnol.* **41**, 232–238 (2023).
61. Bergström, A. et al. Insights into human genetic variation and population history from 929 diverse genomes. *Science* **367**, eaay5012 (2020).
62. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. Preprint at https://doi.org/10.48550/arXiv.1303.3997 (2013).
63. McKenna, A. et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
64. Chang, C. C. et al. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**, s13742-015-0047-8 (2015).
65. Cheng, H. et al. Haplotype-resolved assembly of diploid genomes without parental data. *Nat. Biotechnol.* **40**, 1332–1335 (2022).
66. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
67. Wick, R. R., Judd, L. M., Gorrie, C. L. & Holt, K. E. Unicycler: resolving bacterial genome assemblies from short and long sequencing reads. *PLoS Comput. Biol.* **13**, e1005595 (2017).
68. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
69. Madden, T. The BLAST sequence analysis tool. *NCBI Handbook* **2**, 425–436 (2013).
70. Danecek, P. et al. Twelve years of SAMtools and BCFtools. *Gigascience* **10**, giab008 (2021).
71. Poplin, R. et al. A universal SNP and small-indel variant caller using deep neural networks. *Nat. Biotechnol.* **36**, 983–987 (2018).
72. Yun, T. et al. Accurate, scalable cohort variant calls using DeepVariant and GLnexus. *Bioinformatics* **36**, 5582–5589 (2020).
73. Martin, M. et al. WhatsHap: fast and accurate read-based phasing. Preprint at *bioRxiv* https://doi.org/10.1101/085050 (2016).
74. 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. *Am. J. Hum. Genet.* **81**, 1084–1097 (2007).
75. Formenti, G. et al. Merfin: improved variant filtering, assembly evaluation and polishing via k-mer validation. *Nat. Methods* **19**, 696–704 (2022).
76. Delaneau, O., Zagury, J.-F., Robinson, M. R., Marchini, J. L. & Dermitzakis, E. T. Accurate, scalable and integrative haplotype estimation. *Nat. Commun.* **10**, 5436 (2019).
77. Rautiainen, M. & Marschall, T. GraphAligner: rapid and versatile sequence-to-graph alignment. *Genome Biol.* **21**, 253 (2020).

78. Garrison, E. et al. Variation graph toolkit improves read mapping by representing genetic variation in the reference. *Nat. Biotechnol.* **36**, 875–879 (2018).
79. Hickey, G. et al. Genotyping structural variants in pangenome graphs using the vg toolkit. *Genome Biol.* **21**, 35 (2020).
80. Shen, W., Sipos, B. & Zhao, L. SeqKit2: a Swiss army knife for sequence and alignment processing. *iMeta* **3**, e191 (2024).
81. Xiao, C.-L. et al. MECAT: fast mapping, error correction, and de novo assembly for single-molecule sequencing reads. *Nat. Methods* **14**, 1072–1074 (2017).
82. Ruan, J. & Li, H. Fast and accurate long-read assembly with wtdbg2. *Nat. Methods* **17**, 155–158 (2020).
83. Vaser, R., Sović, I., Nagarajan, N. & Šikić, M. Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Res.* **27**, 737–746 (2017).
84. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnetJ.* **17**, 10–12 (2011).
85. Luo, X., Kang, X. & Schönhuth, A. Phasebook: haplotype-aware de novo assembly of diploid genomes from long reads. *Genome Biol.* **22**, 299 (2021).
86. Smolka, M. et al. Detection of mosaic and population-level structural variants with Sniffles2. *Nat. Biotechnol.* **42**, 1571–1580 (2024).
87. Kirsche, M. et al. Jasmine and Iris: population-scale structural variant comparison and analysis. *Nat. Methods* **20**, 408–417 (2023).
88. Garrison, E. & Marth, G. Haplotype-based variant detection from short-read sequencing. Preprint at https://doi.org/10.48550/arXiv.1207.3907 (2012).
89. Li, H., Feng, X. & Chu, C. The design and construction of reference pangenome graphs with minigraph. *Genome Biol.* **21**, 265 (2020).
90. Garrison, E. & Guarracino, A. Unbiased pangenome graphs. *Bioinformatics* **39**, btac743 (2023).
91. Garrison, E. et al. Building pangenome graphs. *Nat. Methods* **21**, 2008–2012 (2024).
92. Sirén, J. et al. Personalized pangenome references. *Nat. Methods* **21**, 2017–2023 (2024).
93. Holt, J. M. et al. HiPhase: jointly phasing small, structural, and tandem repeat variants from HiFi sequencing. *Bioinformatics* **40**, btae042 (2024).
94. Shafin, K. et al. Haplotype-aware variant calling with PEPPER-Margin-DeepVariant enables high accuracy in nanopore long-reads. *Nat. Methods* **18**, 1322–1332 (2021).
95. Jiang, T. et al. Long-read-based human genomic structural variation detection with cuteSV. *Genome Biol.* **21**, 189 (2020).
96. Sirén, J. et al. Pangenomics enables genotyping of known structural variants in 5202 diverse genomes. *Science* **374**, abg8871 (2021).
97. Ebler, J. et al. Pangenome-based genome inference allows efficient and accurate genotyping across a wide spectrum of variant classes. *Nat. Genet.* **54**, 518–525 (2022).
98. Li, H. et al. A synthetic-diploid benchmark for accurate variant-calling evaluation. *Nat. Methods* **15**, 595–597 (2018).
99. Krusche, P. et al. Best practices for benchmarking germline small-variant calls in human genomes. *Nat. Biotechnol.* **37**, 555–560 (2019).
100. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* **27**, 573–580 (1999).
101. Tarailo-Graovac, M. & Chen, N. Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr. Protoc. Bioinformatics* **25**, 4.10.1–14.10.14 (2009).
102. Frankish, A. et al. GENCODE 2021. *Nucleic Acids Res.* **49**, D916–D923 (2021).
103. Shumate, A. & Salzberg, S. L. Liftoff: accurate mapping of gene annotations. *Bioinformatics* **37**, 1639–1643 (2021).
104. Kim, D., Paggi, J. M., Park, C., Bennett, C. & Salzberg, S. L. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat. Biotechnol.* **37**, 907–915 (2019).
105. Pertea, M. et al. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.* **33**, 290–295 (2015).
106. Pertea, G. & Pertea, M. GFF utilities: GffRead and GffCompare. *F1000Research* **9**, 304 (2020).
107. Kang, Y.-J. et al. CPC2: a fast and accurate coding potential calculator based on sequence intrinsic features. *Nucleic Acids Res.* **45**, W12–W16 (2017).
108. Tang, S., Lomsadze, A. & Borodovsky, M. Identification of protein coding regions in RNA transcripts. *Nucleic Acids Res.* **43**, e78 (2015).
109. Jones, P. et al. InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**, 1236–1240 (2014).
110. Zhang, Y. et al. Model-based analysis of ChIP–seq (MACS). *Genome Biol.* **9**, R137 (2008).
111. Hickey, G. et al. Pangenome graph construction from genome alignments with Minigraph-Cactus. *Nat. Biotechnol.* **42**, 663–673 (2024).
112. Ondov, B. D. et al. Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biol.* **17**, 132 (2016).
113. Guarracino, A., Heumos, S., Nahnsen, S., Prins, P. & Garrison, E. ODGI: understanding pangenome graphs. *Bioinformatics* **38**, 3319–3326 (2022).
114. Ranallo-Benavidez, T. R., Jaron, K. S. & Schatz, M. C. GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. *Nat. Commun.* **11**, 1432 (2020).
115. Chen, X. et al. Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinformatics* **32**, 1220–1222 (2016).
116. Jeffares, D. C. et al. Transient structural variations have strong effects on quantitative traits and reproductive isolation in fission yeast. *Nat. Commun.* **8**, 14061 (2017).
117. Mousavi, N., Shleizer-Burko, S., Yanicky, R. & Gymrek, M. Profiling the genome-wide landscape of tandem repeat expansions. *Nucleic Acids Res.* **47**, e90 (2019).
118. Šošić, M. & Šikić, M. Edlib: a C/C++ library for fast, exact sequence alignment using edit distance. *Bioinformatics* **33**, 1394–1395 (2017).
119. Garrison, E., Kronenberg, Z. N., Dawson, E. T., Pedersen, B. S. & Prins, P. A spectrum of free software tools for processing the VCF variant call format: vcflib, bio-vcf, cyvcf2, hts-nim and slivar. *PLoS Comput. Biol.* **18**, e1009123 (2022).
120. English, A. C. et al. Analysis and benchmarking of small and large genomic variants across tandem repeats. *Nat. Biotechnol.* **43**, 431–442 (2025).
121. Barker, D. J. et al. The ipd-imgt/hla database. *Nucleic Acids Res.* **51**, D1053–D1060 (2023).
122. Cao, Y. et al. MeLoDe: detect DNA methylation from low-depth long-read sequencing data. *Github* https://jianyanglab.github.io/MeLoDe (2025).
123. Li, H. Protein-to-genome alignment with miniprot. *Bioinformatics* **39**, btad014 (2023).
124. Yu, G., Wang, L.-G., Han, Y. & He, Q.-Y. clusterProfiler: an R package for comparing biological themes among gene clusters. *Omics* **16**, 284–287 (2012).
125. Yu, G. et al. GOSemSim: an R package for measuring semantic similarity among GO terms and gene products. *Bioinformatics* **26**, 976–978 (2010).
126. Wick, R. R., Schultz, M. B., Zobel, J. & Holt, K. E. Bandage: interactive visualization of de novo genome assemblies. *Bioinformatics* **31**, 3350–3352 (2015).
127. Sollis, E. et al. The NHGRI-EBI GWAS Catalog: knowledgebase and deposition resource. *Nucleic Acids Res.* **51**, D977–D985 (2023).
128. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
129. Graubert, A., Aguet, F., Ravi, A., Ardlie, K. G. & Getz, G. RNA-SeQC 2: efficient RNA-seq quality control and quantification for large cohorts. *Bioinformatics* **37**, 3048–3050 (2021).
130. Wright, F. A. et al. Heritability and genomics of gene expression in peripheral blood. *Nat. Genet.* **46**, 430–437 (2014).
131. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2010).
132. Stegle, O., Parts, L., Piipari, M., Winn, J. & Durbin, R. Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. *Nat. Protoc.* **7**, 500–507 (2012).
133. Taylor-Weiner, A. et al. Scaling computational genomics to millions of individuals with GPUs. *Genome Biol.* **20**, 228 (2019).
134. Ongen, H., Buil, A., Brown, A. A., Dermitzakis, E. T. & Delaneau, O. Fast and efficient QTL mapper for thousands of molecular phenotypes. *Bioinformatics* **32**, 1479–1485 (2016).
135. Das, S. et al. Next-generation genotype imputation service and methods. *Nat. Genet.* **48**, 1284–1287 (2016).
136. Wang, Y. Code for the 1000 Chinese Pangenome (1KCP) analysis. *Zenodo* https://doi.org/10.5281/zenodo.18659653 (2026).

Acknowledgements We thank S. Liu, F. Cheng, R. Ma, T. Xu, L. Song, L. Chen, P. Yuan and M. Zheng for discussions; and staff at the High-Performance Computing Center at Westlake University for computational support. This research was supported by the National Natural Science Foundation of China (U23A20165, 32100493 and 32595482), the National Key R&D Program of China (2024YFC3405800 and 2024YFC3405802), the 'Pioneer & Leading Goose' R&D Program (2024SSYS0032 and 2025C01092) and the New Cornerstone Science Foundation.

Author contributions J.Y. conceptualized and supervised the study. Y.W., J.Y. and Z.D. designed the experiment. Y.W. and Y.D. developed the analytical methods with input from Z.D. and J.Y. Y.W., Z.D., T.Q., W.C., Y.D., Y.G., Y.C., X. Sun and W.B. performed the data analyses. D.C., D.S., X. Shen, J.Y., M.L., Zhibin Wang, Z.D., T.Q., B.L. and Zhiyi Wang contributed to sample collection. Y.W., W.W., M.G., W.Y. and J.H. developed the data portal. Y.W. and J.Y. wrote the manuscript. All authors reviewed and approved the final manuscript.

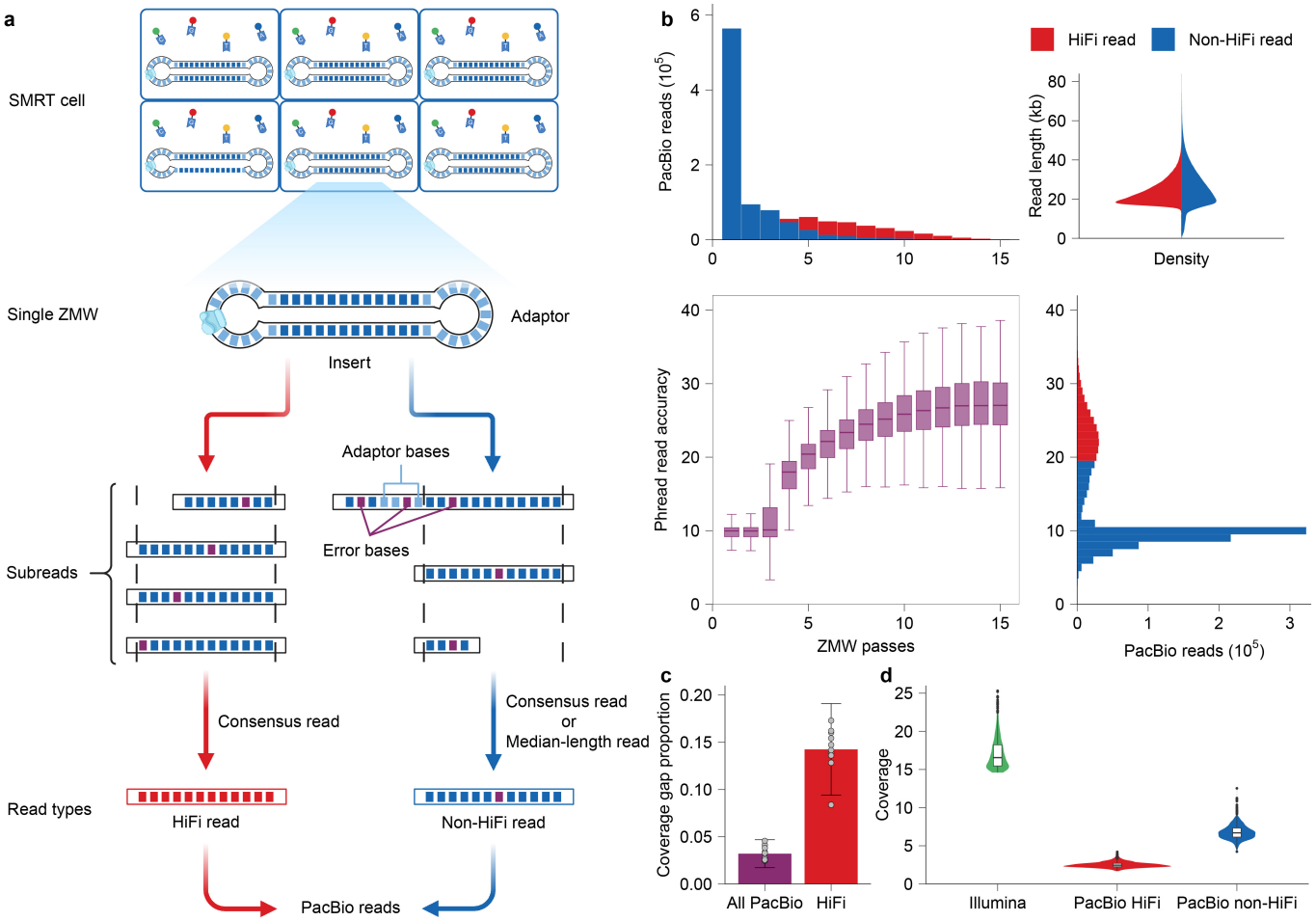
Competing interests The authors declare no competing interests.

Additional information

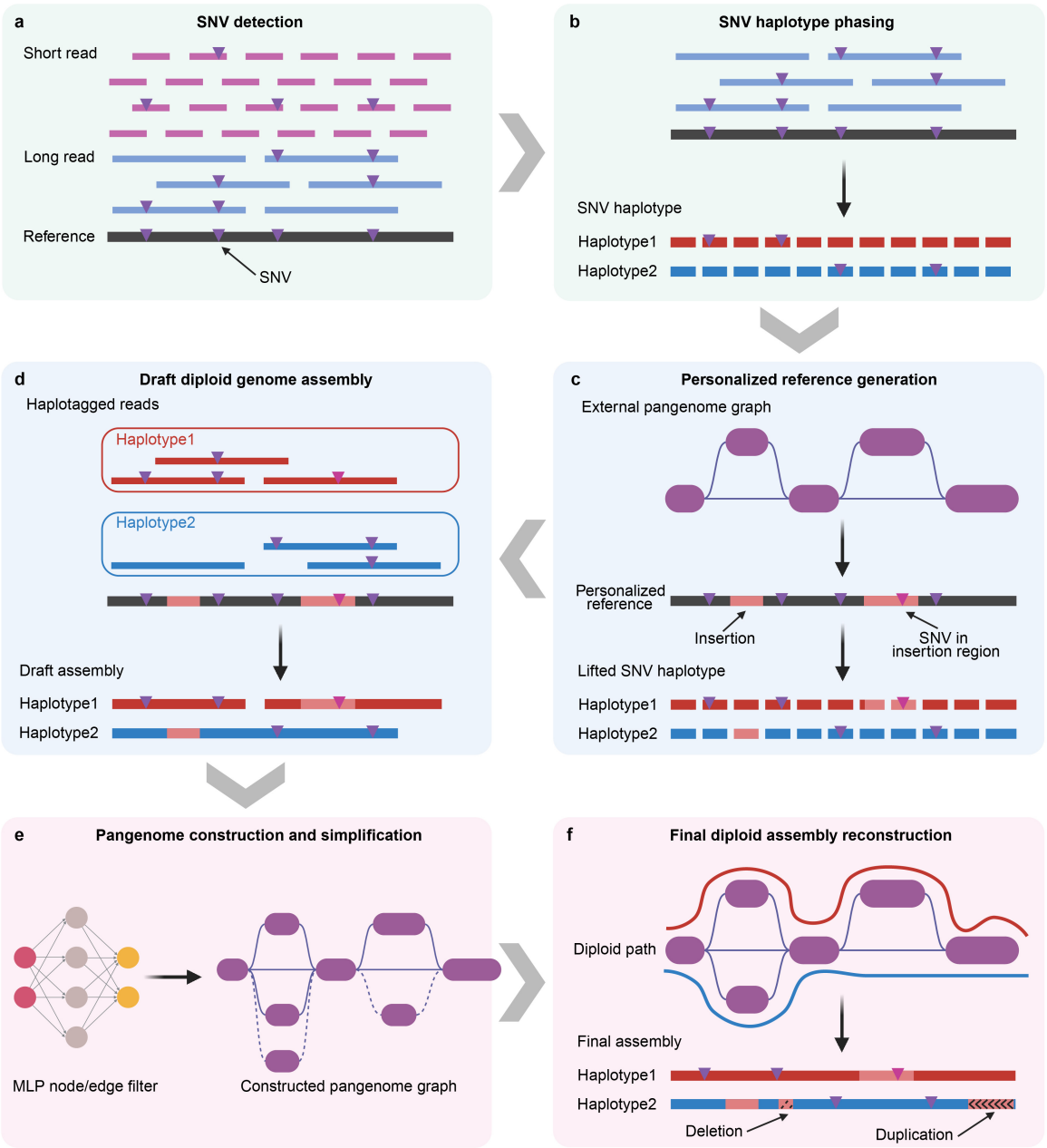
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41586-026-10315-y.

Correspondence and requests for materials should be addressed to Xian Shen or Jian Yang.

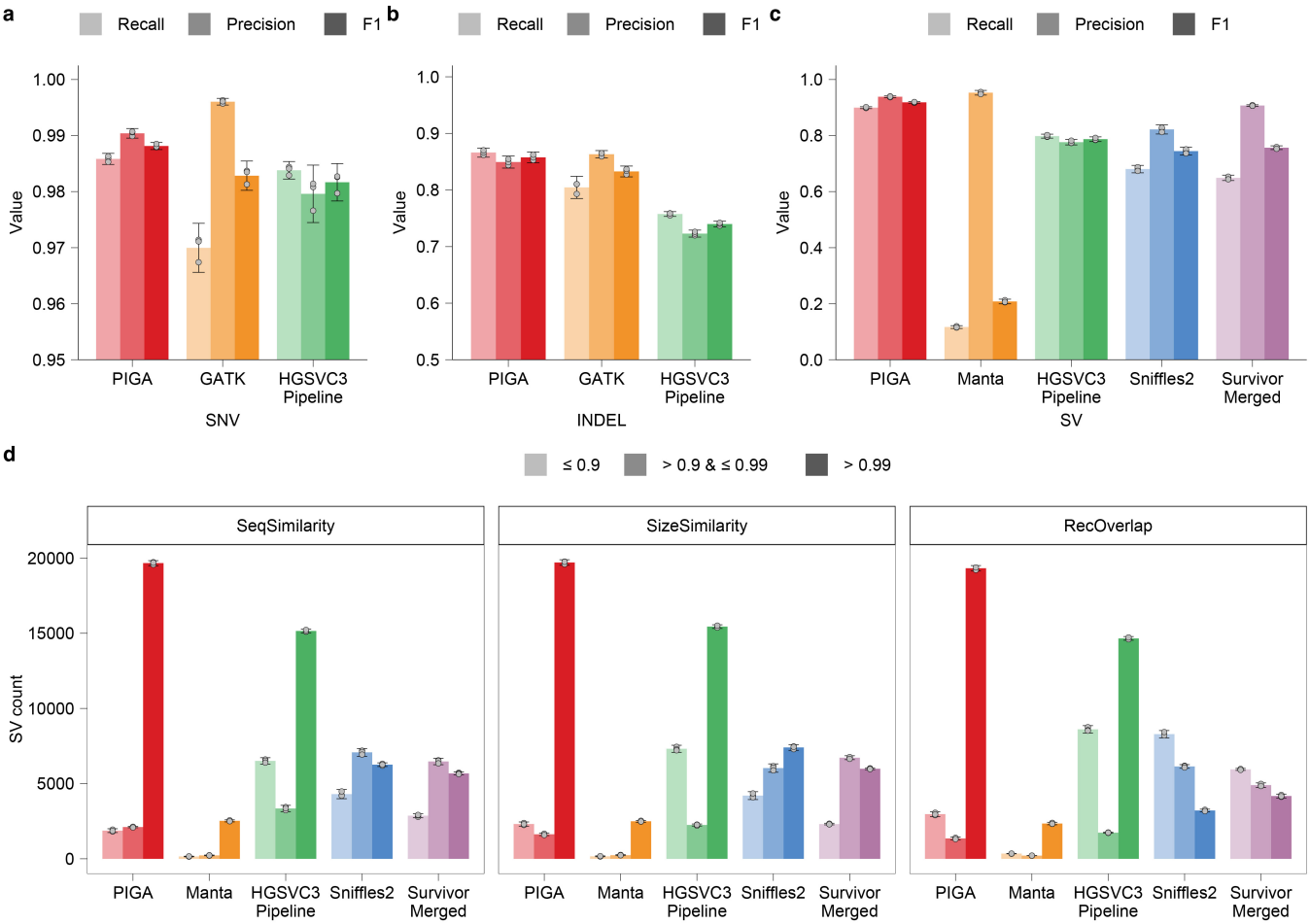
Peer review information *Nature* thanks the anonymous reviewers for their contribution to the peer review of this work. Peer reviewer reports are available.



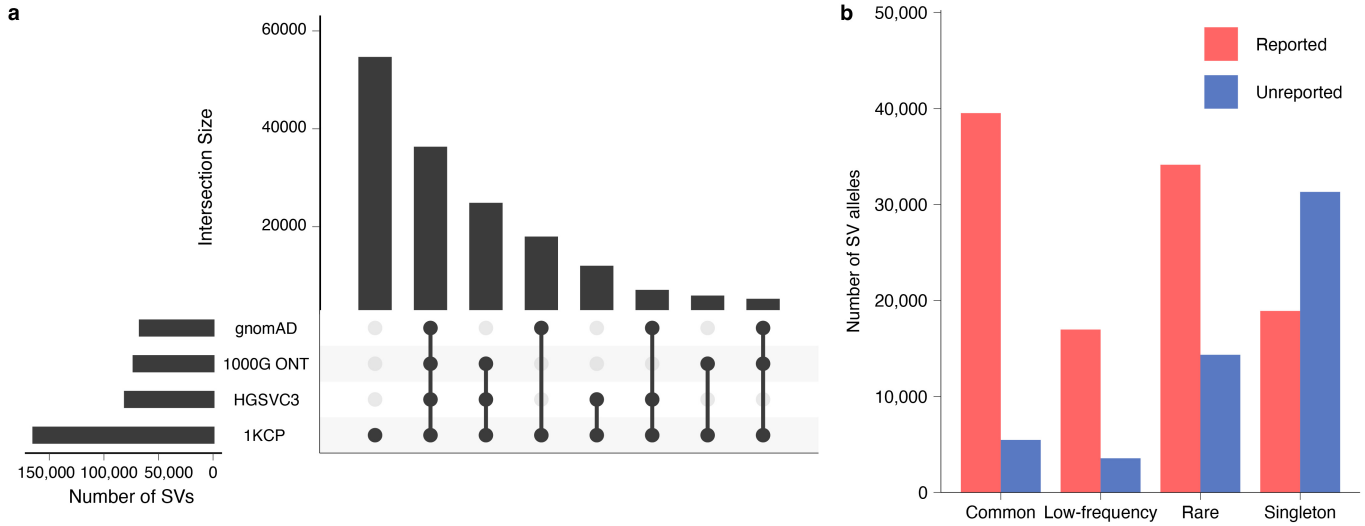
Extended Data Fig. 1 | Read characteristics of the 1KCP modest-coverage dataset. a, Schematic illustration of the read generation process under the "CCS–all" mode. HiFi reads are generated as high-accuracy consensus sequences from multiple subreads, whereas non-HiFi reads typically originate from a consensus or median-length read with lower accuracy. **b**, Characteristics of PacBio reads (n = 1,077,675) 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). 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**, Proportion of genomic regions with zero read coverage (gaps) on randomly selected 1KCP samples (n = 10), using either all PacBio reads or only HiFi reads. Each dot represents an individual sample, bars indicate the mean value, and error bars denote 95% confidence intervals. **d**, Coverage distribution of Illumina, PacBio HiFi, and PacBio non-HiFi reads in the 1KCP modest-coverage dataset (n = 1,099). 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. Illustrations in **a** were created using BioRender; Yang, J. https://www.biorender.com/bkemfes (2026).



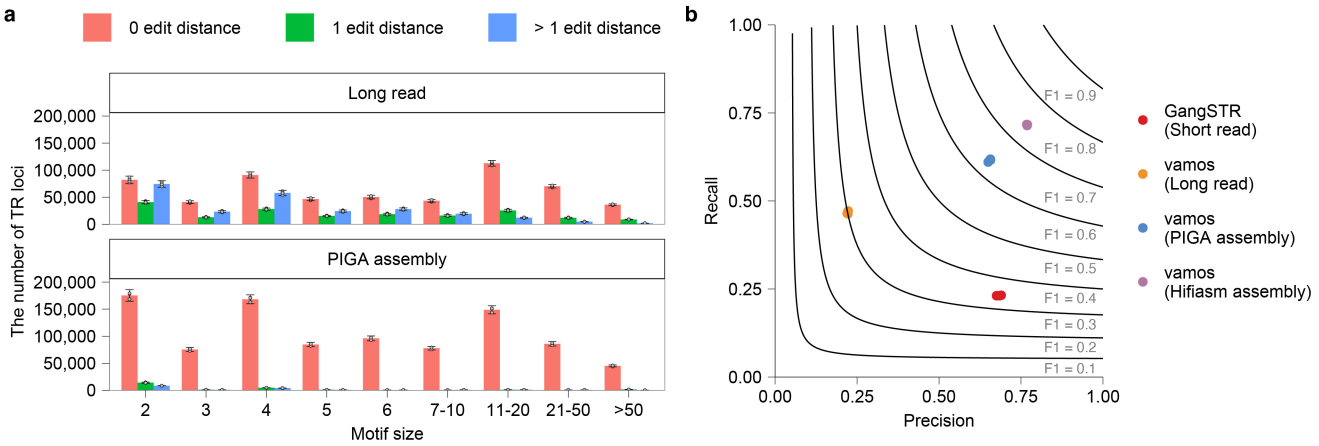
Extended Data Fig. 2 | The Pangenome-Informed Genome Assembly (PIGA) workflow. The PIGA workflow consists of several main steps: **a**, Population-level SNV calling by integrating long-read and short-read sequencing data. **b**, SNV haplotype phasing by leveraging long-read and population information. **c**, Generating personalized references by modifying the reference genome with homozygous variants genotyped from the external pangenome. **d**, Partitioning long reads into haplotypes and producing draft diploid assemblies. **e**, Constructing the pangenome using draft diploid assemblies and simplifying its structurally complex or error-prone regions. **f**, Reconstructing final diploid assemblies by inferring diploid paths from the constructed pangenome. Figure created using BioRender; Yang, J. https://biorender.com/wuy8fxl (2026).



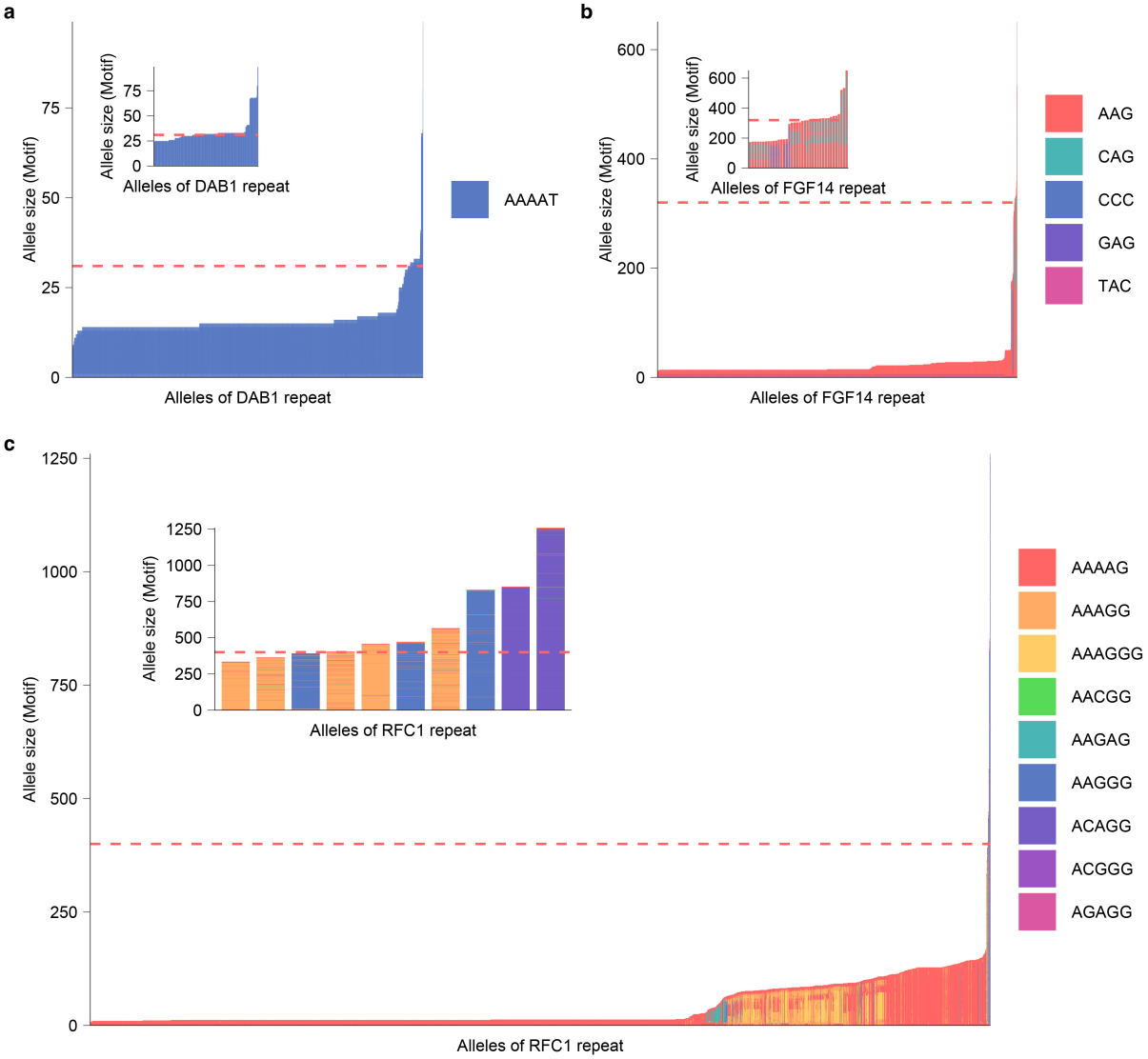
Extended Data Fig. 3 | Benchmarking of variant detection performance across different methods in three benchmark samples. a, SNV detection performance on three benchmark samples evaluated using hap.py[99]. Each dot represents an individual sample, bars indicate the mean value, and error bars denote 95% confidence intervals. **b**, Indel detection performance evaluated using hap.py. Each dot represents an individual sample, bars indicate the mean value, and error bars denote 95% confidence intervals. **c**, SV detection performance on three benchmark samples evaluated using Truvari[28] bench (–pick multi–sizemin 50). Each dot represents an individual sample, bars indicate the mean value, and error bars denote 95% confidence intervals. **d**, Number of SVs in three benchmark stratified by sequence accuracy levels. Accuracy is evaluated using three metrics: sequence similarity (left), size similarity (middle), and reciprocal overlap percentages (right). Each dot represents an individual sample, bars indicate the mean value, and error bars denote 95% confidence intervals.



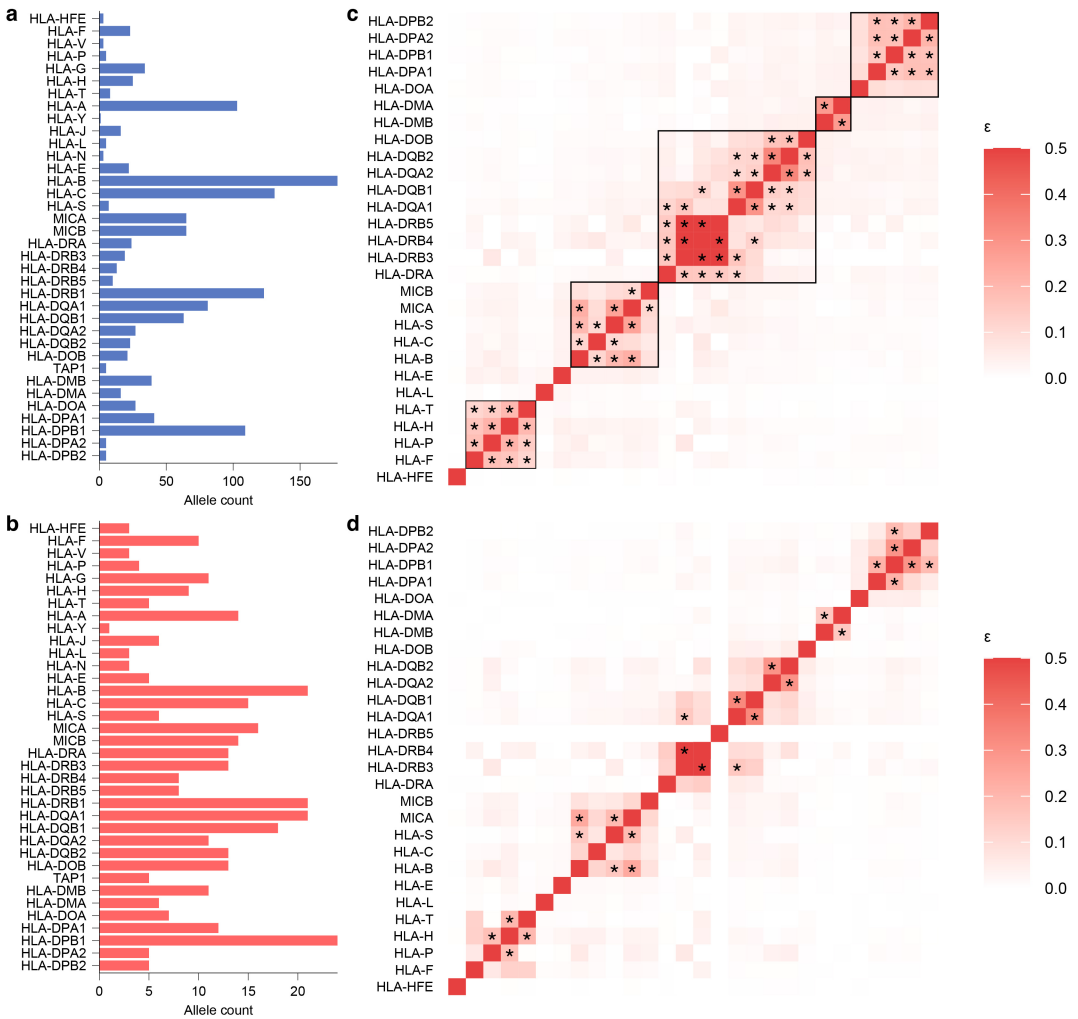
Extended Data Fig. 4 | Comparison of the 1KCP SV callset with public SV callsets. a, Overlap between the 1KCP SV callset and public SV callsets, including gnomAD, 1000G ONT, and HGSVC3. The intersection size represents the number of shared SV alleles among callsets. **b**, Number of SV alleles stratified by allele frequency, including common (AF > 0.05), low-frequency (AF > 0.01 and ≤ 0.05), rare (AF ≤ 0.01), and singleton, further categorized as reported (red) or unreported (blue) in public callsets.



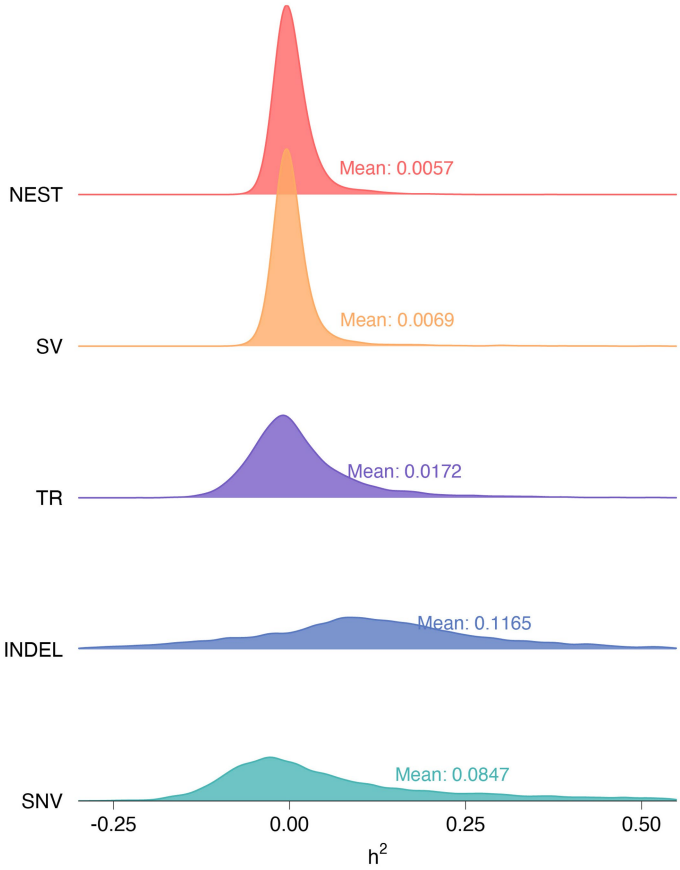
Extended Data Fig. 5 | Benchmark of TR calling performance across different methods in three benchmark samples. a, Number of TR loci in haplotypes ($n = 6$) of three benchmark samples characterized by vamos, stratified by motif size and edit distance. Edit distance refers to the differences in motif unit representations of TR alleles between the query callsets (Long-read and PIGA assembly) and the truth callset (hifiasm assembly). Each dot represents a haplotype, bars indicate the mean value, and error bars denote 95% confidence intervals. **b**, TR benchmarking performance using Truvari bench (-s 5) and refine (-u), restricted to GIAB-defined TR regions[120]. The black curves represent constant F1 scores from 0.1 to 0.9.



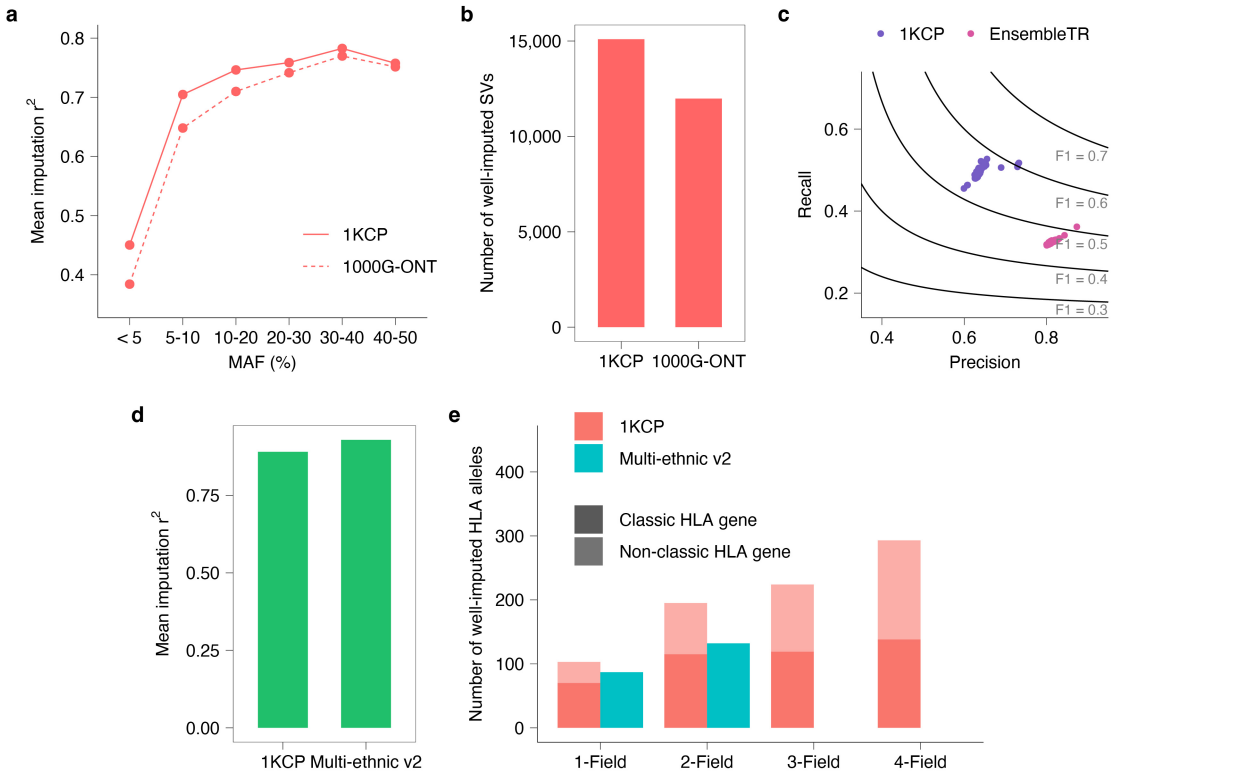
Extended Data Fig. 6 | Disease-related tandem repeat sites. The motif composition and allele sizes of known disease-related TR sites exhibiting expansion events: **a**, TR site in the *DAB1* gene. **b**, TR site in the *FGF14* gene. **c**, TR sites in the *RFC1* gene. Different colors represent distinct TR motifs. The horizontal red dashed line indicates the previously defined pathogenic size threshold. Insets in each panel provide zoomed-in views of expanded TR alleles.



Extended Data Fig. 7 | The allele map of HLA genes. a, The number of four-field alleles across different HLA genes in the 1KCP dataset. **b**, The number of four-field alleles with allele frequency >0.01 across different HLA genes. **c**, Entropy-based LD measurement index (ε) between HLA genes, calculated based on four-field alleles. Black borders indicate defined LD groups. **d**, Entropy-based LD measurement index (ε) between HLA genes, calculated based on 3-field alleles.



Extended Data Fig. 8 | Distribution of cis-heritability of gene expression across variant types. Cis-heritability of gene expression explained by different variant types, including SNVs, indels, SVs, TRs, and nested variants.



Extended Data Fig. 9 | 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. **e**, 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.

nature portfolio

Corresponding author(s): Jian Yang

Last updated by author(s): Jan 24, 2026

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	No data-collection software was used.
Data analysis	The analysis scripts used in this study: https://github.com/JianYang-Lab/1KCP-analysis PIGA: https://github.com/JianYang-Lab/PIGA CCS (v6.2.0): https://github.com/PacificBiosciences/ccs actc (v0.6.0): https://github.com/PacificBiosciences/actc DeepConsensus (v1.2.0): https://github.com/google/deepconsensus Hifiasm (v0.15.3): https://github.com/chhylp123/hifiasm minimap2 (v2.24): https://github.com/lh3/minimap2 Unicycler (v0.5.0): https://github.com/rrwick/Unicycler BEDTools (v2.30.0): https://github.com/arq5x/bedtools2 blastn (v2.13.0): https://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST Meryl (v1.4.1): https://github.com/marbl/meryl Merqury (v1.3): https://github.com/marbl/merqury Inspector (v1.0.1): https://github.com/Maggi-Chen/Inspector Flagger (v0.2.0): https://github.com/mobinasri/flagger rustybam (v0.1.34): https://github.com/mrvollger/rustybam Dipcall (v0.3): https://github.com/lh3/dipcall BWA (v0.7.17): https://github.com/lh3/bwa GATK (v4.2.6.1): https://github.com/broadinstitute/gatk DeepVariant (v1.4.0): https://github.com/google/deepvariant

hap.py (v0.3.14): https://github.com/Illumina/hap.py
Truvari (v4.2.2): https://github.com/ACEnglish/truvari
WhatsHap (v1.4): https://github.com/whatshap/whatshap
Tandem Repeat Finder (v4.09): https://tandem.bu.edu/trf/trf.html
RepeatMasker (v4.1.1): https://www.repeatmasker.org/
Liftoff (v1.6.3): https://github.com/agshumate/Liftoff
HISAT2 (v2.2.1): https://github.com/DaehwanKimLab/hisat2
StringTie (v2.2.1): https://github.com/gpertea/stringtie
gffcompare (v0.12.9): https://github.com/gpertea/gffcompare
CPC2 (v1.0.1): https://github.com/gao-lab/CPC2_standalone
GeneMarkS-T (v5.1): https://genemark.bme.gatech.edu/GeneMark
InterProScan (v5.66): https://www.ebi.ac.uk/interpro
SEI: https://github.com/FunctionLab/sei-framework
Enformer (v0.8.8): https://github.com/lucidrains/enformer-pytorch
MASC3 (v3.0.0): https://github.com/macs3-project/MACS
Minigraph (v0.21): https://github.com/lh3/minigraph
Mash (v2.3): https://github.com/marbl/Mash
gaffilter: https://github.com/ComparativeGenomicsToolkit/cactus-gfa-tools
odgi (v0.8.6): https://github.com/pangenome/odgi
seqwish (v0.7.4): https://github.com/ekg/seqwish
smoothxg (v0.7.9): https://github.com/pangenome/smoothxg
GFAffix (v0.1.5): https://github.com/marschall-lab/GFAffix
GenomeScope (v2.0.1): https://github.com/tbenavi1/genomescope2.0
dna-brnn (v0.1): https://github.com/lh3/dna-nn
VG (v1.59.0): https://github.com/vgteam/vg
vcfbub (v0.1.0): https://github.com/pangenome/vcfbub
vamos (v2.1.5): https://github.com/ChaissonLab/vamos
BCFtools (v1.20): https://github.com/samtools/bcftools
GraphAligner (v1.0.20): https://github.com/maickrau/GraphAligner
Sniffles2 (v2.6.2): https://github.com/fritzsedlazeck/Sniffles
SURVIVOR (v1.0.6): https://github.com/fritzsedlazeck/SURVIVOR
Jasmine (v1.1.5): https://github.com/mkirsche/Jasmine
Manta (v1.6.0): https://github.com/Illumina/manta
GangSTR (v2.5): https://github.com/gymreklab/GangSTR
vcfwave (v1.0.14): https://github.com/vcflib/vcflib
edlib (v1.2.7): https://github.com/Martinsos/edlib
MeLoDe: https://github.com/JavenCao/MeLoDe
miniprot (v0.12): https://github.com/lh3/miniprot
pangene (v1.1): https://github.com/lh3/pangene
clusterProfiler (v4.12.6): https://github.com/YuLab-SMU/clusterProfiler
GOSemsim (v2.30.2): https://github.com/YuLab-SMU/GOSemSim
Bandage (v0.9.0): https://github.com/rrwick/Bandage
PLINK2 (v2.0.0): https://www.cog-genomics.org/plink/2.0
HiFiHLA (v0.3.1): https://github.com/PacificBiosciences/hifihla
HLA-HD (v1.7.1): https://w3.genome.med.kyoto-u.ac.jp/HLA-HD
eLD (v1.0): https://www.sg.med.osaka-u.ac.jp/tools.html
STAR (v2.7.9): https://github.com/alexdobin/STAR
RNA-SeQC (v2.4.2): https://github.com/getzlab/rnaseqc
edgeR (v3.22.5): https://bioconductor.org/packages/release/bioc/html/edgeR.html
PEER: https://github.com/PMBio/peer
tensorQTL (v1.0.9): https://github.com/broadinstitute/tensorqtl
GCTA (v1.94.1): https://yanglab.westlake.edu.cn/software/gcta
dgrm: https://github.com/PeixiongYuan/dgrm
coloc (v5.2.3): https://github.com/chr1swallace/coloc
SMR (v1.3.1): https://yanglab.westlake.edu.cn/software/smr
Minimac4 (v4.1.6): https://genome.sph.umich.edu/wiki/Minimac4

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

The release of the 1KCP data is approved by the Human Genetic Resources Administration of China (permission numbers 2025BAT00888 and 2025BAT00720). 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. Public datasets used in this study include: GRCh38 genome (https://gdc.cancer.gov/about-data/gdc-data-

processing/gdc-reference-files); CHM13 genome (https://github.com/marbl/CHM13); HPRC v1 assemblies (https://github.com/human-pangenomics/HPP_Yead_Assemblies); CPC assemblies (https://github.com/Shuhua-Group/Chinese-Pangenome-Consortium-Phase-I); GENCODE annotation (https://www.gencodegenes.org/human); NCBI RefSeq database (https://ftp.ncbi.nlm.nih.gov/refseq/release); 1000G high-coverage Illumina variants (https://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/1000G_2504_high_coverage/working/20220422_3202_phased_SNV_INDEL_SV); 1000G ONT SVs (https://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/1KG_ONT_VIENNA/release/v1.0); GnomAD variants (http://www.gnomad-sg.org); HGSVC3 SVs (https://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/HGSVC3/release/Variant_Calls/1.0); OMIM database (https://www.omim.org); COSMIC database (https://cancer.sanger.ac.uk/cosmic); STRchive database (https://strchive.org/resources); IPD-IMGT/HLA database (https://www.ebi.ac.uk/ipd/imgt/hla); GWAS catalog (https://www.ebi.ac.uk/gwas); BioBank Japan GWAS (https://humandbs.dbcls.jp/en/hum0197-v23); EnsembleTR panel (https://github.com/gymrek-lab/EnsembleTR).

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	Sex was recorded by self-reporting during sample collection and confirmed by calculating the coverage depth ratio of X and Y chromosomes from WGS data. The final cohort comprised 737 males and 642 females. Sex is considered a biological variable in this study. No sex-based exclusion criteria were applied.
Reporting on race, ethnicity, or other socially relevant groupings	The first phase of the 1KCP project enrolled 1,379 Chinese participants in Wenzhou, China.
Population characteristics	The cohort consists of 1,379 individuals ranging in age from 18 to 76 years. The cohort represents a general population sample derived from individuals undergoing routine health examinations, rather than a disease-specific case-control cohort; therefore, participants were not selected based on past or current diagnostic categories. All samples underwent Illumina short-read WGS and RNA-seq. Among these participants, 55 samples were selected for high-coverage PacBio HiFi long-read WGS and Hi-C sequencing, and 1,099 were selected for modest-coverage PacBio long-read WGS.
Recruitment	Participants were recruited from volunteers undergoing routine health examinations at the Health Examination Center of the Second Affiliated Hospital of Wenzhou Medical University. While there is a potential self-selection bias towards individuals with higher health awareness, germline genomic architecture is independent of such behavioral traits, and thus the results regarding variant detection remain robust.
Ethics oversight	Informed consent was obtained from all participants. The study was approved by the Ethics Committee of Westlake University (approval no. 20201127YJ001) and the Ethics Committee of Wenzhou Medical University (approval no. 2019-107).

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	In this study, we analyzed 1,144 Chinese individuals from the 1KCP project who had both short-read and long-read whole-genome sequencing (WGS) data. The sample size was determined to capture the vast majority of common and low-frequency variants, and the cohort represents the largest long-read datasets for the Chinese population to date.
Data exclusions	During data preprocessing, 28 samples were excluded due to mismatched short-read and long-read WGS data.
Replication	We selected 10 samples subjected to both modest-coverage and high-coverage PacBio HiFi long-read WGS. Among these, 3 samples were utilized to compare the assembly quality between PIGA and Hifiasm assemblies. The variant detection performance demonstrated that the vast majority of variants were reproducibly identified across different assemblies of the same samples. Variants in extremely repetitive regions could not be reliably resolved across different assemblies due to their complex graph structure and the inherent noise of the long-read data used in the study.
Randomization	Randomization is not required for this study because all experiments were conducted under identical conditions.
Blinding	Blinding is not required for this study because all experiments were conducted under identical conditions.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Plants

Seed stocks

Report on the source of all seed stocks or other plant material used. If applicable, state the seed stock centre and catalogue number. If plant specimens were collected from the field, describe the collection location, date and sampling procedures.

Novel plant genotypes

Describe the methods by which all novel plant genotypes were produced. This includes those generated by transgenic approaches, gene editing, chemical/radiation-based mutagenesis and hybridization. For transgenic lines, describe the transformation method, the number of independent lines analyzed and the generation upon which experiments were performed. For gene-edited lines, describe the editor used, the endogenous sequence targeted for editing, the targeting guide RNA sequence (if applicable) and how the editor was applied.

Authentication

Describe any authentication procedures for each seed stock used or novel genotype generated. Describe any experiments used to assess the effect of a mutation and, where applicable, how potential secondary effects (e.g. second site T-DNA insertions, mosiacism, off-target gene editing) were examined.