

Beyond the linear genome: A population-scale Chinese pangenome for human genetics

Lulu Shi,^{1,2} Wenjian Li,^{1,2} Wenfeng Yi,^{1,2} and Yang Wu^{1,*}

¹Institute of Rare Diseases, West China Hospital, Sichuan University, Chengdu 610000, China

²These authors contributed equally

*Correspondence: yang.wu@wchscu.edu.cn

Received: May 18, 2026; Accepted: September 11, 2026; Published Online: September 14, 2026; <https://doi.org/10.59717/j.xinn-life.2026.100256>

© 2026 The Author(s). This is an open access article under the CC BY license (<https://creativecommons.org/licenses/by/4.0/>).

Citation: Shi L., Li W., Yi W., et al. (2026). Beyond the linear genome: A population-scale Chinese pangenome for human genetics. *The Innovation Life* 4:100256.

For more than two decades, the linear human reference genome has provided the central framework for modern genetics, from variant discovery to clinical interpretation. However, a single reference sequence cannot fully capture the structural variation, repetitive regions, and population-specific diversity present in human genomes. Complete-genome projects and pangenome studies have addressed complementary aspects of this problem, with graph-based pangenomes reducing reference bias and improving access to complex loci.^{1–3} Parallel work in crops and other species has

established the same broader principle: pangenomes can expose presence–absence variation, structural haplotypes and trait-associated alleles obscured by a single reference.⁴ The Human Pangenome Reference Consortium (HPRC) demonstrated that the value of high-quality phased assemblies for graph-based analysis,³ and the Chinese Pangenome Consortium (CPC) expanded population representation through generating 116 phased Chinese assemblies. Against this background, the 1000 Chinese Pangenome Project (1KCP) marks a substantial advance in both scale and analytical scope.⁵ By

Beyond the linear genome: A population-scale Chinese pangenome for human genetics

From reference bias to population-scale variant discovery and precision medicine

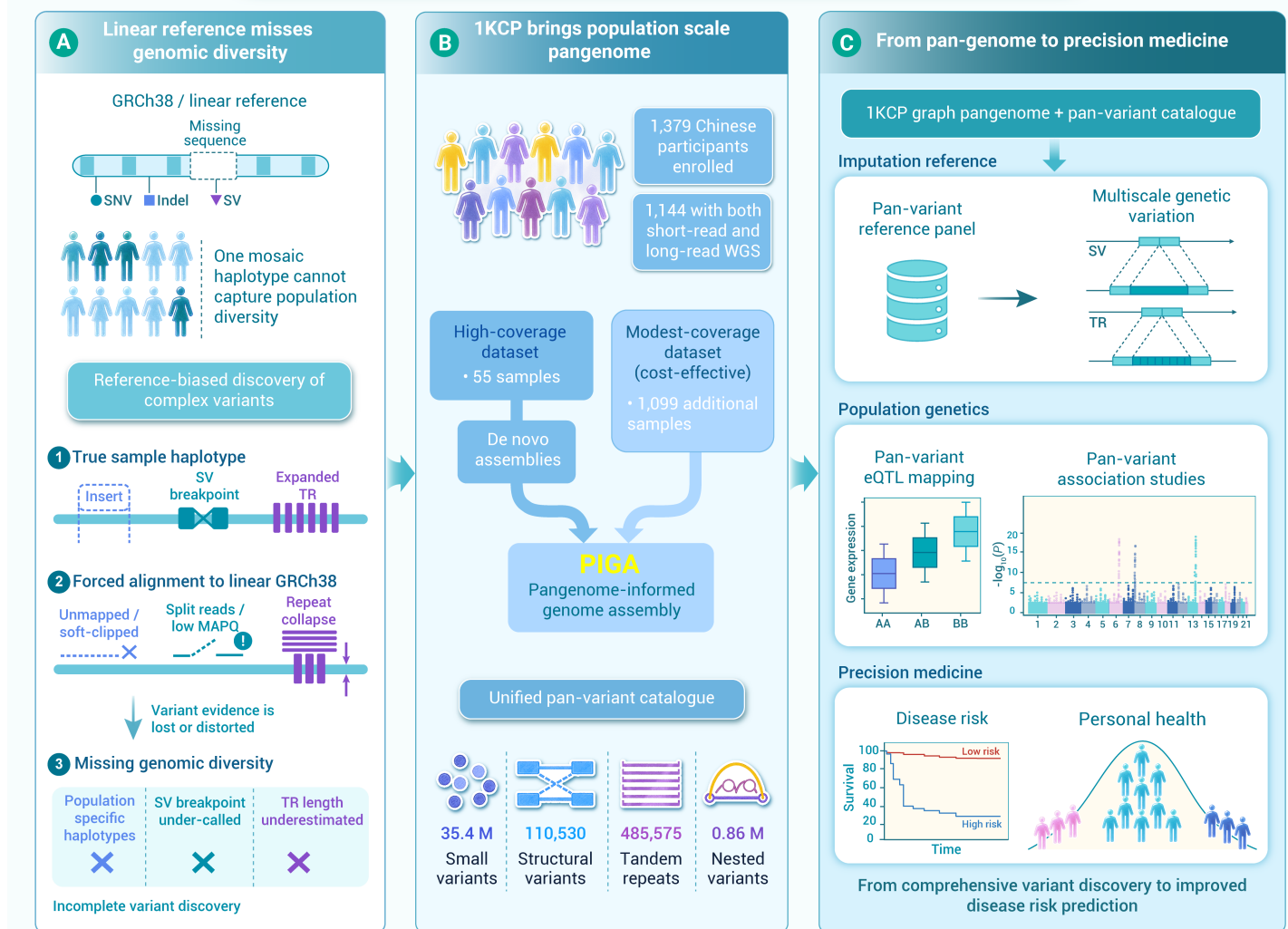


Figure 1. Population-scale Chinese pangenome uncovers genetic diversity beyond the linear reference The 1KCP dataset integrates 1,116 high-quality diploid assemblies from 1,379 individuals, capturing 405.3 Mb of previously missing sequences, including 26.2 Mb of functional genetic and regulatory elements absent from GRCh38 and CHM13. Using a graph-based pangenome framework, it represents diverse variant types (35.4 million small variants, 110,530 structural variants, 485,575 tandem repeats, and 0.86 million nested variants) enabling a unified pan-variant catalogue. This resource facilitates downstream analyses including multiscale genetic variation, HLA haplotypes, gene-cluster variation, pan-variant eQTL mapping, association studies, and precision medicine, exemplifying the transition from linear references to population-informed graph genomes for comprehensive variant discovery and translational genomics. Some graphic elements created in BioRender. wenjian, L. (2026) <https://BioRender.com/furlzml>.

generating 1,116 diploid assemblies, 1KCP extends earlier human pangenomes from reference construction towards population-scale analysis. Its significance lies not simply in recovering additional non-reference sequence, an expected consequence of sampling more genomes, but in assigning frequencies to complex alleles and linking them to functional analysis, imputation and medical genetics (Figure 1).

This scale was enabled by the Pangenome-Informed Genome Assembly (PIGA) workflow (<https://github.com/JianYang-Lab/PIGA>). Earlier haplotype-resolved assemblies and pangenome-based genotyping established the value of sequence-resolved structural alleles, whereas 1KCP combined 55 de novo assemblies with 1,061 pangenome-informed assemblies generated from modest-coverage sequencing. PIGA integrates population-level phasing, personalized reference construction and graph-path inference to reconstruct diploid genomes economically. Its variant calls showed average F1 scores of 0.9881 for single-nucleotide variants and 0.9669 for structural variants, although indels, highly repetitive regions and chromosome-scale phase remained less reliable. Conventional gap closing relies on additional reads or long-range information from the same individual and remains difficult when repeats exceed read length or homologous sequences collapse. Population-informed reconstruction may instead assist reconstruction across under-covered intervals by borrowing recurrent haplotype paths from the cohort. This advantage is greatest for common, graph-represented haplotypes whereas private alleles, satellites and segmental duplications remain challenging. PIGA should therefore be viewed not as a replacement for high-coverage de novo assembly, but as a scalable complement whose regional uncertainty must be retained in downstream analyses.

Population scale sampling changes variant discovery in three ways. First, repeated sampling stabilizes allele-frequency estimates and distinguishes common from rare and private alleles. Second, broad sampling improves rare-variant discovery: 1KCP found that common and low-frequency non-reference sequences approached saturation in relatively small panels, whereas identifying rare sequences required hundreds of genomes. Third, observing many homologous haplotypes reveals multiallelic and nested architectures that are difficult to define in smaller studies. The 1KCP catalogue includes 35.4 million small variants, 110,530 structural variants (SVs), 485,575 tandem repeats (TRs) and 0.86 million nested variants within large SVs. This multi-layered variant landscape underscores the complexity of human genomes. Particularly notable is the representation of nested variants, which capture fine-scale differences embedded within larger structural changes. Such variants are difficult to resolve reliably using traditional approaches and linear reference frameworks, yet they represent a significant component of genetic diversity. The finding that 80% of SV sites were multiallelic also exposes a statistical bottleneck. Although graph-based pangenomes can preserve alternative alleles and their haplotypic context, representation alone does not solve association testing. In 1KCP, related alleles were merged and the multi-allelic sites were then decomposed into constituent biallelic variants for population analyses. This strategy improves compatibility with existing GWAS frameworks but can flatten relationships among alternative alleles at the same locus. Future analyses will require allele-specific, locus-level or haplotype- and graph-path-based tests that model correlated alternatives and genotype uncertainty. Thus, the next methodological step is not merely to place multiallelic variation in a graph, but to develop calibrated association models that use its structure.

The medical value of this richer representation is illustrated at several resolutions. The study identified 1,013 exon-overlapping SVs across 623 medically relevant genes and used population distributions to recognize outlier TR expansions, including a *GIPC1* expansion associated with allele-specific hypermethylation. At the *HP* locus, structurally distinct haplotypes were linked to multiple GWAS signals, suggesting that some associations detected using short-read data may tag unresolved SVs or nested alleles. Long-read phasing across HLA likewise revealed fine-scale coding and non-coding haplotype diversity. These examples show why population pangenomes are more than variant catalogues: they provide the allele-frequency and haplotype context needed to distinguish recurrent, rare and potentially pathogenic configura-

tions.

A further advance is the integration of complex variation into functional and cohort-scale analysis. In the pan-variant eQTL analysis, complex variants contributed an average of 12.6% of cis-heritability among genes with estimated cis-heritability above 0.05, and 3,256 eQTLs involved SVs, TRs or nested variants. The accompanying imputation panel then projected these variant classes and HLA alleles into short-read and array-based cohorts. It increased the number of well-imputed SVs by 26% compared with the 1000G-ONT panel and also improved TR and HLA imputation.⁵ Together, these advances create a practical bridge between assembly-based discovery and existing biobanks, enabling complex variants to be incorporated into association studies and evaluated for potential clinical relevance.

The central question is no longer whether pangenomes improve genomic representation, but how they should be integrated into routine analysis. Graph-native mapping and genotyping may be most useful at structurally diverse loci, whereas clinical reporting and longitudinal cohorts still depend on stable genomic coordinates. A practical near-term model may therefore combine graph-based discovery with explicit projection onto linear reporting systems. Assembly quality must also be assessed locally rather than only through genome-wide averages. Errors in repeats, segmental duplications or phase blocks can bias allele-frequency and GWAS estimates, nominate the wrong regulatory-coding relationship for functional validation or alter the clinical interpretation of compound heterozygosity. Region- and allele-level confidence should therefore be propagated into association, functional and clinical analyses instead of being hidden behind a single assembly-quality threshold.

Population pangenomes and personal diploid assemblies should likewise be viewed as complementary. Population graphs provide shared coordinates, frequency priors and imputation scaffolds, whereas personal assemblies can resolve private alleles and individual-specific haplotypes. As complete diploid genomes become increasingly accessible, human genetics may move from testing isolated reference-relative variants towards analysing sequence-defined alleles, local haplotypes and graph paths. Realizing this transition will require interoperable representations, multiallelic statistical models, functional validation across tissues and clinically interpretable confidence metrics. The long-term impact of 1KCP will therefore depend less on how many additional paths are added to the graph than on how reliably those paths can be connected to function and phenotype.

REFERENCES

- Gao Y., Yang X., Chen H., et al. (2023). A pangenome reference of 36 Chinese populations. *Nature* **619**:112–121. DOI:10.1038/s41586-023-06173-7
- Miga K. H. and Wang T. (2021) The need for a human pangenome reference sequence. *Annu. Rev. Genomics Hum. Genet.* **22**:81–102. DOI:10.1146/annurev-genom-120120-081921
- Liao W.-W., Asri M., Ebler J., et al. (2023). A draft human pangenome reference. *Nature* **617**:312–324. DOI:10.1038/s41586-023-05896-x
- Jayakodi M., Shim H. and Mascher M. (2025). What are we learning from plant pangenomes. *Annu. Rev. Plant Biol.* **76**:663–686. DOI:10.1146/annurev-arplant-090823-015358
- Wang Y., Duan Z., Chen D., et al. (2026). The 1000 Chinese Pangenome empowers medical and population genetics. *Nature* **654**:121–130. DOI:10.1038/s41586-026-10315-y

FUNDING AND ACKNOWLEDGMENTS

This research was supported by the Fundamental Research Funds for the Central Universities, the 1·3·5 project for disciplines of excellence, West China Hospital, Sichuan University (ZYXC24006). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

AUTHOR CONTRIBUTIONS

Y.W. conceived the project. L.S., W.L., W.Y. and Y.W. wrote the manuscript. L.S. and W.L. designed and drew Figure in the manuscript. L.S., W.L., W.Y., and Y.W. edited the manuscript. All authors contributed to the manuscript and approved the final version.

DECLARATION OF INTERESTS

The authors declare no competing interests.