

Regional diversity meets global representation in large-scale human pangenomics

Kailin Chen^{1,2} and Zhaoyuan Fang^{1,2,*}

¹Center for Biomedical Systems and Informatics, Zhejiang University-University of Edinburgh Institute, Zhejiang University, Hangzhou 310000, China

²Edinburgh Medical School, College of Medicine and Veterinary Medicine, The University of Edinburgh, Edinburgh EH8 9YL, United Kingdom

*Correspondence: zhaoyuanfang@intl.zju.edu.cn (Z. F.); fangzhaoyuan@sibs.ac.cn (Z. F.)

Received: April 28, 2026; Accepted: September 3, 2026; Published Online: September 4, 2026; <https://doi.org/10.59717/j.xinn-life.2026.100334>

© 2027 The Author(s). This is an open access article under the CC BY license (<https://creativecommons.org/licenses/by/4.0/>).

Citation: Chen K. and Fang Z. (2027). Regional diversity meets global representation in large-scale human pangenomics. *The Innovation Life* 5:100334.

Human pangenomics is moving beyond the single linear reference paradigm towards population-scale representations that better capture complex genomic variation. Many current human pangenomes include both global and regional initiatives, but their limited sample sizes constrain the representation of population diversity. Recently, the establishment of the 1000 Chinese Pangenome (1KCP) project represents a major milestone in large-scale regional pangenomics. By combining high-coverage de novo assemblies with pangenome-informed assembly of modest-coverage samples, 1KCP expanded the resource to 1,116 diploid genome assemblies while substantially reducing sequencing costs relative to uniformly high-coverage de novo assembly. This large-scale framework not only broadens the catalog of non-reference sequences and genetic variants, but also enables complex variants, including structural variants and tandem repeats, to be linked to gene regulation and human phenotypes, revealing classes of variation incompletely captured by conventional small-variant-centered analyses. Nevertheless, some challenges remain, including reduced assembly accuracy in repetitive and structurally complex genomic regions when using modest-coverage data. Broader and more balanced sampling will also be needed to better represent population diversity. Building on these insights, we propose a hierarchical framework for the next-generation global pangenome, where large-scale regional pangenomes guide diversity-informed selection of representative individuals for high-quality genome assembly, while global allele frequencies are integrated with the resulting pangenome graph.

The standard linear reference human genome (currently GRCh38), offering a unified coordinate system, has witnessed the rapid expansion of human genomics for over twenty years. One further milestone was the telomere-to-telomere (T2T) CHM13 genome assembly that essentially resolved genomic regions remaining incomplete in GRCh38, providing a near-complete sequence.¹ Moreover, the Human Pangenome Reference Consortium (HPRC) draft extended the concept of a single reference genome to a population-scale pangenome graph for human species diversity.² Almost concurrently, the Chinese Pangenome Consortium (CPC) presented the first Chinese population pangenome, marking an important milestone in regional pangenomics.³ Since then, human pangenomics has continued to evolve rapidly. In a recent *Nature* study,⁴ a research team at Westlake University led by Professor Jian Yang presented a large-scale Chinese pangenome as part of the 1000 Chinese Pangenome (1KCP) project, highlighting a cost-effective and generalizable approach for characterizing the extensive landscape of population genomic variations and enabling widespread downstream pan-variant analyses. In this Perspective, we discuss the significance of 1KCP as a template for large-scale regional pangenomics and its implications for building a next-generation global human pangenome.

Over the past decades, our understanding of human genomes has largely been shaped by a single reference genome and accompanied by single-nucleotide or short variations revealed by short-read sequencing. Such a small-variant-centered analytical framework has enabled standardized and robust detection of single-nucleotide variants (SNVs) and small insertions/deletions (indels) at a population scale and relatively low cost, supporting large-cohort genetic studies and associated applications. However, it is less effective at resolving complex and larger variants,⁵ such as tandem repeats (TRs) and structural variants (SVs), because short reads mapped to a single linear reference often fail to accurately identify repetitive or structurally divergent regions. Advances in long-read sequencing, such as

PacBio and Oxford Nanopore technologies, have enabled haplotype-resolved assembly of diploid genomes, providing the data for pangenome construction that compactly incorporates larger variants and improves variation discovery. We can distinguish two types of human pangenomes (Figure 1A): one is the global pangenome such as the HPRC draft² and the near-complete HGVS assemblies,⁵ the other is the regional pangenome focusing on a specific population, which has been actively developed, including the CPC,³ the United-Arab-Emirates-based Arab Pangenome Reference (UPR),⁶ and pangenome graphs representing Japanese and Saudi Arabian populations.⁷ Both strategies are equally important and complementary, with the global ones providing broad cross-population representation and the regional ones capturing finer population-specific diversity. However, existing human pangenomes remain constrained by limited sample sizes (47 for HPRC, 65 for HGVS, 58 for CPC, and 53 for UPR), hampering the representation of a full spectrum of variants,⁵ especially rare variants that are more likely to be pathogenic, as well as the accurate estimation of allele frequencies that are critical in subsequent genetic applications.

In this recent study, Yang and his team developed a smart hybrid strategy to boost up the scale to 1,116 diploid genomes, significantly reducing sequencing costs relative to uniformly high-coverage de novo assembly. They accomplished high-coverage long- and short-read sequencing of 55 samples and modest-coverage long- and short-read sequencing of 1,099 samples, with 10 in overlap for training and evaluation. For the small high-coverage dataset, they generated de novo assemblies through haplotype phasing assisted by Hi-C data. For the large modest-coverage dataset, they designed and employed a sophisticated pangenome-informed genome assembly (PIGA) workflow. Briefly, after SNV calling and phasing, draft diploid genome assemblies were obtained for each individual as guided by a personalized reference, which was constructed through applying to CHM13 the homozygous variants called against a preliminary pangenome based on the high-coverage dataset. These large-scale draft assemblies were integrated into a pangenome graph, which was further simplified by training a neural network classifier on the overlapping dual-coverage samples to remove erroneous nodes or edges. Incorporating all information available, the final diploid assemblies were inferred from the pangenome graph. Overall, PIGA achieved a high validation concordance for both SNVs and SVs, with average F1 scores of 0.9881 and 0.9669, respectively. Leveraging all those 1,116 and public (HPRC and CPC) genome assemblies, as well as CHM13 and GRCh38, the 1KCP pangenome was finally created alongside an integrated pangenome, demonstrating a convincing and cost-effective approach towards large-scale pangenomics.

The resulting 1KCP pangenome substantially expands the current human genome resources, capturing 405.3 Mb sequences absent from GRCh38 and CHM13 (more than half also missed in CPC and HPRC), including 26.2 Mb annotated as either genic or cis-regulatory (such as enhancers and promoters). Beyond the sheer addition of non-reference sequences, the 1KCP pangenome supports the extraction of an extensive population-scale variant catalogue: 35.4 million small variants and 110 thousand SVs against the GRCh38 reference, 0.86 million variants nested in the non-reference sequences, as well as 0.49 million TR polymorphisms involving repeat copy numbers or motif compositions. Notably, the 1KCP pangenome systematically connects complex variations to functions and phenotypes. Over 1,000 SVs overlap exonic regions of pathogenic genes indexed in OMIM and COSMIC, and 735 polymorphic gene clusters are enriched in immune and blood functions, including the haptoglobin linked to lipid traits and the HLA

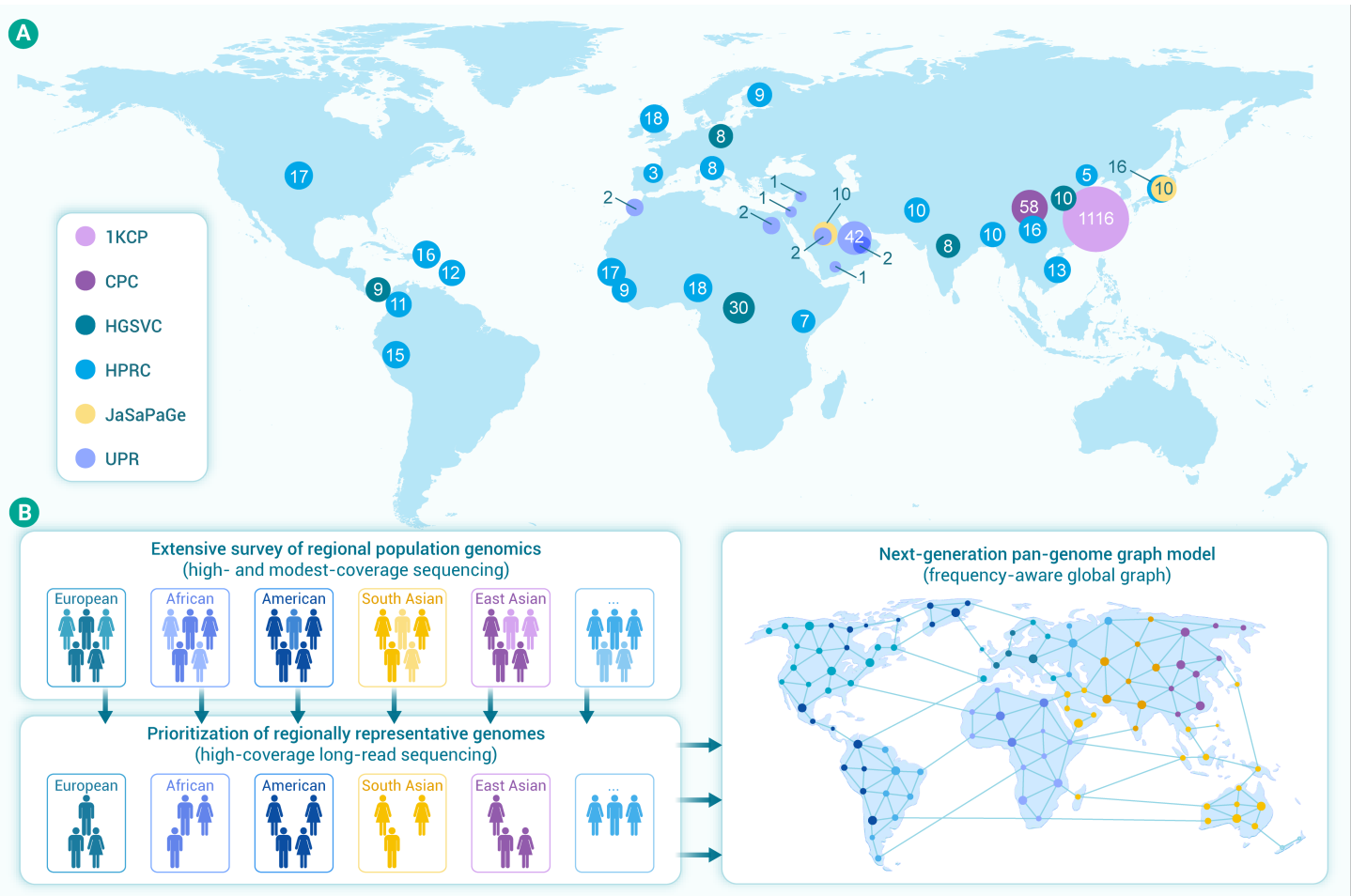


Figure 1. Current pangenomes and the next-generation global pangenome model (A) Current pangenome initiatives. Each circle represents a regional sample set (with a number labeling the sample size) included in an initiative denoted by different colors. (B) The next-generation global pangenome model. The key steps include an extensive survey of regional genomes, prioritization of regionally representative genomes, and construction of a frequency-aware global pangenome. In the right panel, each node represents a regional variant and node size represents its frequency. Note: This is a conceptual illustration for demonstrative purposes and does not represent actual data. Figure partially created in BioRender. Fang, Z. (2026) <https://BioRender.com/6oeInx8>

region now resolved at four-field allelic resolution. A unified pan-variant expression quantitative trait locus (eQTL) analysis further suggests that complex variants collectively explain about 12.6% of cis-heritability for gene expression, highlighting a substantial contribution from variant classes beyond conventional SNVs and indels. Among those complex variant eQTLs, TR polymorphisms account for the majority (76%), and notably, pangenome-derived motif composition variants capture 57% of TR eQTL signals undetectable by conventional copy-number-based analysis alone. Further colocalization analysis reveals potentially causal complex variants for phenotypic traits, exemplified by an 18-kb GSTM1 deletion in platelet count variation. Together, these findings raise an interesting possibility that the missing heritability problem in traits might be partially rooted in missing complex variants, a gap large-scale pangenome-empowered analyses are now in a position to address, although this cis-heritability differs from the total heritability of traits that are largely attributed to common SNVs.⁸

Nevertheless, scaling a regional pangenome does not inherently guarantee the comprehensive capture of population diversity. Regional initiatives, including the Chinese initiatives, need to mitigate the masking effect of the dominant majority to adequately represent the ethnic minorities, geographically isolated groups, and historically admixed populations. Achieving this inclusivity requires a paradigm shift in study design, transitioning from mere numerical scaling to a framework underpinned by balanced sampling, proactive community engagement, transparent informed consent, and privacy-preserving data governance and sharing.

Meanwhile, PIGA does not mean to replace traditional strategies. Traditional approaches prioritize single-sample assembly quality and variant accuracy, but the cost scales up and limits population-scale application. In

contrast, PIGA focuses more on population-level variant recall rate as a cost-effective expansion, with a potential trade-off in assembly accuracy and resolution. Thus, PIGA provides an alternative strategy in such scenarios. It is also worth noting that, although PIGA uses high-coverage samples to guide modest-coverage assemblies, difficult regions such as repeats or complex structural variants may still be less accurately resolved than in high-coverage *de novo* assemblies. In addition, neural-network-based filtering may also introduce training-set-dependent biases, warranting further evaluation of its performance for rare or population-specific variants. Hence, its application to highly diverse or underrepresented populations will require further evaluation, especially when the resulting pangenome is intended for fine-scale variant interpretation or clinical applications.

Finally, we believe it is time to propose a next-generation global pangenome model (Figure 1B). Current global pangenome initiatives, often biased towards very few better-covered populations, are largely limited by a poor coverage of region-specific diversity.⁹ To address such representational limitations, the next-generation model may take a bottom-up strategy. First, it extensively surveys region-specific pangenomes in a large-scale manner throughout populations over the world, and the PIGA approach could be useful at this stage. Second, it identifies the most representative individuals in each regional population, which are prioritized for high-coverage long-read sequencing and *de novo* assembly,⁵ contributing modest-scale high-quality haplotypes for constructing an initial global graph model. Unlike the traditional pangenomes that capture both the presence and frequency of variants, this initial global pangenome built from regionally representative assemblies is inherently qualitative rather than quantitative, in that the global variant frequencies are not encoded. To transform the initial graph into a frequency-

aware model, the global allele frequencies should be additionally encoded, most naturally as graph node weights (Figure 1B). Therefore, the future picture of human pangenomics is likely to be hierarchical: the next-generation global pangenome graph maintains only the most representative sequences and variants (frequency-aware), whereas the regional pangenomes aim to be as inclusive as possible for local variants. A related emerging hierarchical logic is the concept of super-pangenomes in plant genomics that extends species-level pangenomes into genus-wide frameworks and similarly illustrates that genomic diversity is better represented through nested reference resources.¹⁰ By prioritizing coordinated sequencing protocols, unified coordinate systems, and standardized variant nomenclatures, this hierarchical roadmap may significantly mitigate current barriers to cross-consortium data integration. Nonetheless, realizing the full potential of this model will require sustained community-driven efforts to accurately represent complex multi-allelic and nested variants, as well as seamlessly bridging the gap between graph-native variation and standardized clinical and functional annotations.

REFERENCES

1. Nurk S., Koren S., Rhie A., et al. (2022). The complete sequence of a human genome. *Science* **376**:44–53. DOI:10.1126/science.abj6987
2. 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
3. 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
4. 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
5. Logsdon G.A., Ebert P., Audano P.A., et al. (2025). Complex genetic variation in nearly complete human genomes. *Nature* **644**:430–441. DOI:10.1038/s41586-025-09140-6
6. Nassir N., Almarri M.A., Kumail M., et al. (2025). A draft UAE-based Arab pangenome reference. *Nat. Commun.* **16**:6747. DOI:10.1038/s41467-025-61645-w
7. Kulmanov M., Ashouri S., Liu Y., et al. (2025). Phased genome assemblies and pangenome graphs of human populations of Japan and Saudi Arabia. *Sci. Data* **12**:1316. DOI:10.1038/s41597-025-05652-y
8. Yang J., Benyamin B., McEvoy B.P., et al. (2010). Common SNPs explain a large proportion of the heritability for human height. *Nat. Genet.* **42**:565–569. DOI:10.1038/ng.608
9. Nassir N., Almarri M.A., Akter H., et al. (2025). Advancing clinical genomics with Middle Eastern and South Asian pangenomes. *Nat. Med.* **31**:725–727. DOI:10.1038/s41591-025-03544-7
10. Sun H., Zhang J., Liao S., et al. (2026). Population-level super-pangenome reveals genome evolution and empowers precision breeding in watermelon. *Nat. Genet.* **58**:1176–1185. DOI:10.1038/s41588-026-02598-8

FUNDING AND ACKNOWLEDGMENTS

This work was supported by the National Natural Science Foundation of China (82273093, T2341024), the Huadong Medicine Joint Funds of the Zhejiang Provincial Natural Science Foundation of China (LHDMZ23H160006), and the Medical Interdisciplinary Innovation Program 2024, Zhejiang University School of Medicine. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. The authors would like to thank Prof. Jian Yang for his careful reading and valuable comments on a draft version of this perspective.

AUTHOR CONTRIBUTIONS

All authors contributed to the manuscript and approved the final version.

DECLARATION OF INTERESTS

The authors declare no competing interests.

ETHICAL STATEMENT AND PATIENT CONSENT

Not Applicable.