

Genome-driven Chinese precision medicine: Biobank-scale genomic research as a new paradigm

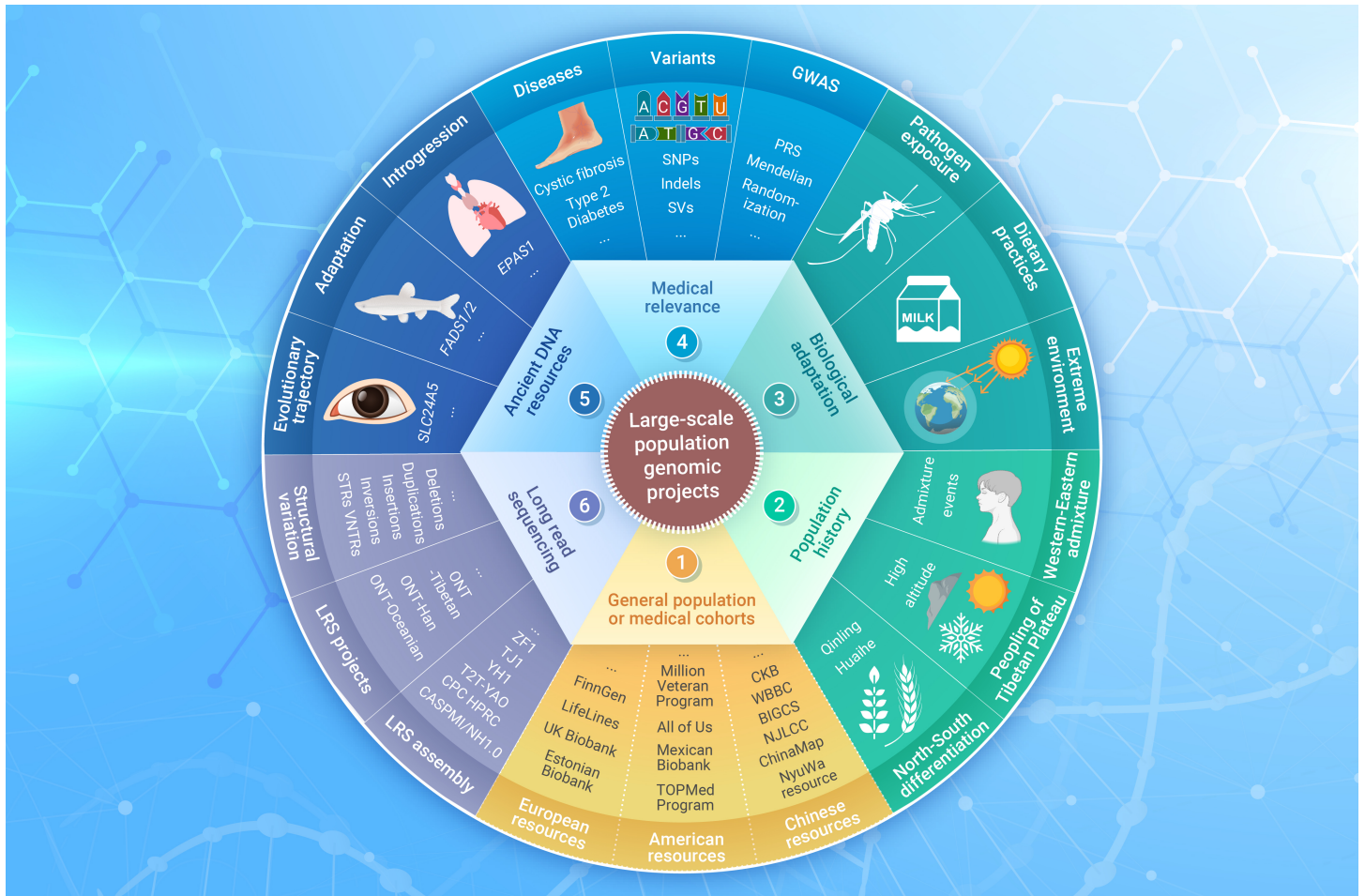
Mengge Wang,^{1,2,3,6,*} Shuhan Duan,^{1,2,4,5,6} Xiangping Li,^{1,2} Junbao Yang,^{4,5,*} Huijun Yuan,^{1,2,*} Chao Liu,^{3,*} and Guanglin He^{1,2,3,5,6,*}

*Correspondence: Menggewang2021@163.com (M.W.); yuanhj301@163.com (H.Y.); liuchaogzf@163.com (C.L.); yjb3589@vip.163.com (J.Y.); guanglinhescu@163.com (G.H.)

Received: September 25, 2024; Accepted: March 27, 2025; Published Online: April 7, 2025; <https://doi.org/10.59717/j.xinn-life.2025.100131>

© 2025 The Author(s). This is an open access article under the CC BY license (<https://creativecommons.org/licenses/by/4.0/>).

GRAPHICAL ABSTRACT



PUBLIC SUMMARY

- Integrating population and medical genomics has advanced genome-driven precision medicine.
- Large-scale genomic datasets enhance disease prediction, treatment, and human diversity insights.
- Chinese genomic studies boost PRS accuracy, pharmacogenomics, and precise disease subtyping.
- This study emphasizes extensive sampling, high-throughput sequencing, and representative genomic databases.

Genome-driven Chinese precision medicine: Biobank-scale genomic research as a new paradigm

Mengge Wang,^{1,2,3,6,*} Shuhan Duan,^{1,2,4,5,6} Xiangping Li,^{1,2} Junbao Yang,^{4,5,*} Huijun Yuan,^{1,2,*} Chao Liu,^{3,*} and Guanglin He^{1,2,3,5,6,*}

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

²Center for Archaeological Science, Sichuan University, Chengdu 610000, China

³Anti-Drug Technology Center of Guangdong Province, Guangzhou 510230, China

⁴Institute of Basic Medicine and Forensic Medicine, North Sichuan Medical College, Nanchong 637100, China

⁵Research Center for Genomic Medicine, North Sichuan Medical College, Nanchong 637100, China

⁶These authors contributed equally

*Correspondence: Menggewang2021@163.com (M.W.); yuanhj301@163.com (H.Y.); liuchaogzf@163.com (C.L.); yjb3589@vip.163.com (J.Y.); guanglinhesu@163.com (G.H.)

Received: September 25, 2024; Accepted: March 27, 2025; Published Online: April 7, 2025; <https://doi.org/10.59717/j.xinn-life.2025.100131>

© 2025 The Author(s). This is an open access article under the CC BY license (<https://creativecommons.org/licenses/by/4.0/>).

Citation: Wang M., Duan S., Li X., et al. (2025). Genome-driven Chinese precision medicine: Biobank-scale genomic research as a new paradigm. *The Innovation Life* 3:100131.

Large-scale genomic resources from biobank sequencing projects are crucial for understanding the interplay between environmental and genetic factors in human disease and health traits, as well as for reconstructing human evolutionary history. We summarize recent advances in genomic cohorts and highlight opportunities for non-Eurocentric populations from a multidisciplinary perspective. Initiatives like the UK100K, All of Us, and TOPMed precision medicine programs have shifted research paradigms from problem-derived to data-driven approaches, enhancing our understanding of the genetic architecture of diseases in Europeans and their descendants. However, biases persist, such as the Han bias in Chinese genomic projects focused on medical and anthropological purposes. These biases contribute to global genomic inequalities and disparities in Chinese precision medicine. Evolutionary genomic studies of modern and ancient genomes provide new insights into the history and adaptive trajectories of critical human genetic mutations. These findings underscore the importance of personal genome medicine tailored to ethnolinguistically and genetically diverse populations. This strategy is vital for assessing disease burden and understanding the etiology of human disease. Our work emphasizes the need to include underrepresented genomic diversity to create a comprehensive catalog of human genetic variations and understand their biological implications.

INTRODUCTION

The Human Genome Project (HGP), completed in 2003, generated the first comprehensive human genome sequence, providing foundational insights into human genetics.^{1,2} Over the past two decades, advancements in genomics have expanded our understanding of human genetic diversity and the mechanisms underlying human variants. Key initiatives such as the Human Haplotype Map Project (HapMap),³ the 1000 Genomes Project (1KGP),^{4,5} the Simons Genome Diversity Project (SGDP),⁶ and the Human Pangenome Project (HPP)⁷ have been instrumental in this progress. Recent large-scale medical genomic programs, including the UK Biobank,^{8,9} Trans-Omics for Precision Medicine (TOPMed),¹⁰ All of Us Research Program,¹¹ and Genome Aggregation Database (gnomAD),¹² have integrated deep phenotypes with molecular, clinical, and genomic data, enhancing our ability to analyze multi-omics data (Figure 1). Biobank-based genomic studies have rapidly expanded, focusing on regional population diversity. Africa, with its vast genetic and linguistic diversity, has launched significant initiatives, such as the Three Million African Genomes (3MAGs),¹³ Harnessing Data Science for Health Discovery and Innovation in Africa (DS-I Africa), Human Heredity and Health in Africa (H3Africa),¹⁴ the Nigerian 100K Genomic Project,¹⁵ and Uganda Genomic Resources.¹⁶ Non-African and non-Eurocentric studies have provided insights into human divergence, admixture, and medical characteristics of underrepresented populations from Asia (GenomeAsia100K),¹⁷ Oceania (Australian Genomics Health Futures Mission),¹⁸ and America (Mexican Biobank).^{19,20} These resources from diverse populations have unveiled new genotype-phenotype associations, elucidated causal relationships between variants and phenotypic differences, and highlighted the complex evolutionary processes shaping genetic diversity.

Genetic variant genotyping and the exploration of genotype-phenotype

associations have historically relied on microarray genotyping, which detects most of the common variants in early population genomic databases but often overlooks rare variants and population-specific mutations, particularly from underrepresented groups. Whole-genome sequencing (WGS) has emerged as an efficient tool for evaluating rare genetic variants and advancing new sequencing technologies. High-depth, large-scale WGS research allows direct assessment of genetic variants, especially rare population-specific variants linked to various human diseases. It also facilitates the integration of population and clinical genetic information, becoming a cornerstone of precision medicine. WGS is applied in Mendelian disease diagnosis, disease risk prediction, and personalized treatment. Launched in 2015, the Precision Medicine Initiative (PMI) has made precision medicine a global trend, ushering in a new era of human genomic medicine. This approach aims to understand disease mechanisms, transform healthcare, and identify therapeutic interventions using advanced technologies to enhance personalized disease detection and treatment. Vast amounts of genomic data and precision medicine strategies for equitable human health marked this era. Large-scale population cohort studies are crucial for revealing disease etiology and pathogenesis, improving patient prognosis, and reducing disease burden. These methods are preferred for studying genotype-phenotype associations and the causal relationships between exposures and disease outcomes. Several regional, medical-driven genome sequencing projects have been launched in Western Eurasian and American countries, such as the UK100K, Estonian Biobank, Genomes of Icelanders, FinnGen, All of Us, and TOPMed programs.^{21–26} The precision medicine boom in Europe and America has spurred the creation of genomic resources in Asia, although these efforts face limitations, including the underrepresentation of regional ethnic minorities and the identification of prospective disease relevance and biomarkers.

The underrepresentation and global disparities of Asian ancestries in human genetic datasets have hindered our understanding of genetic diversity, reducing medical applicability for many Asians.¹⁷ Recent genomic studies targeting Asians through anthropologically informed sampling and large-scale biobank programs have elucidated the genetic basis of complex diseases and the demographic history of Japanese, Korean, Singaporean, and Chinese populations (Figures 1–2).^{27–33} However, Chinese genomic studies remain underrepresented, perpetuating health disparities.^{34,35} China, a multiethnic nation with vast cultural, linguistic, ecological, and genetic diversity, is significantly affected by this gap.^{36–38} Ancient DNA research and archaeological findings indicate that extensive population migrations and complex admixture scenarios have contributed to mosaic patterns of genomic diversity from the Paleolithic to modern times.^{36,38–40} Admixture modeling and linguistic phylogenetic analyses support the farming-and-language dispersal hypothesis in East Asia.^{41,42} This hypothesis posits that various population expansion events disseminated language families, including Hmong-Mien, Tai-Kadai, Austroasiatic, Austronesian, Sino-Tibetan, and Trans-Eurasian, across different regions.^{38,43–45} Despite these findings, large-scale reference genomic datasets for Chinese populations suitable for anthropological research and precision medicine are lacking. Few genetic studies have focused on ethnically and geographically diverse Chinese populations, impacting the comprehensive understanding necessary for precision medicine.^{25,46–48} This underrepresentation in genomic projects limits the global

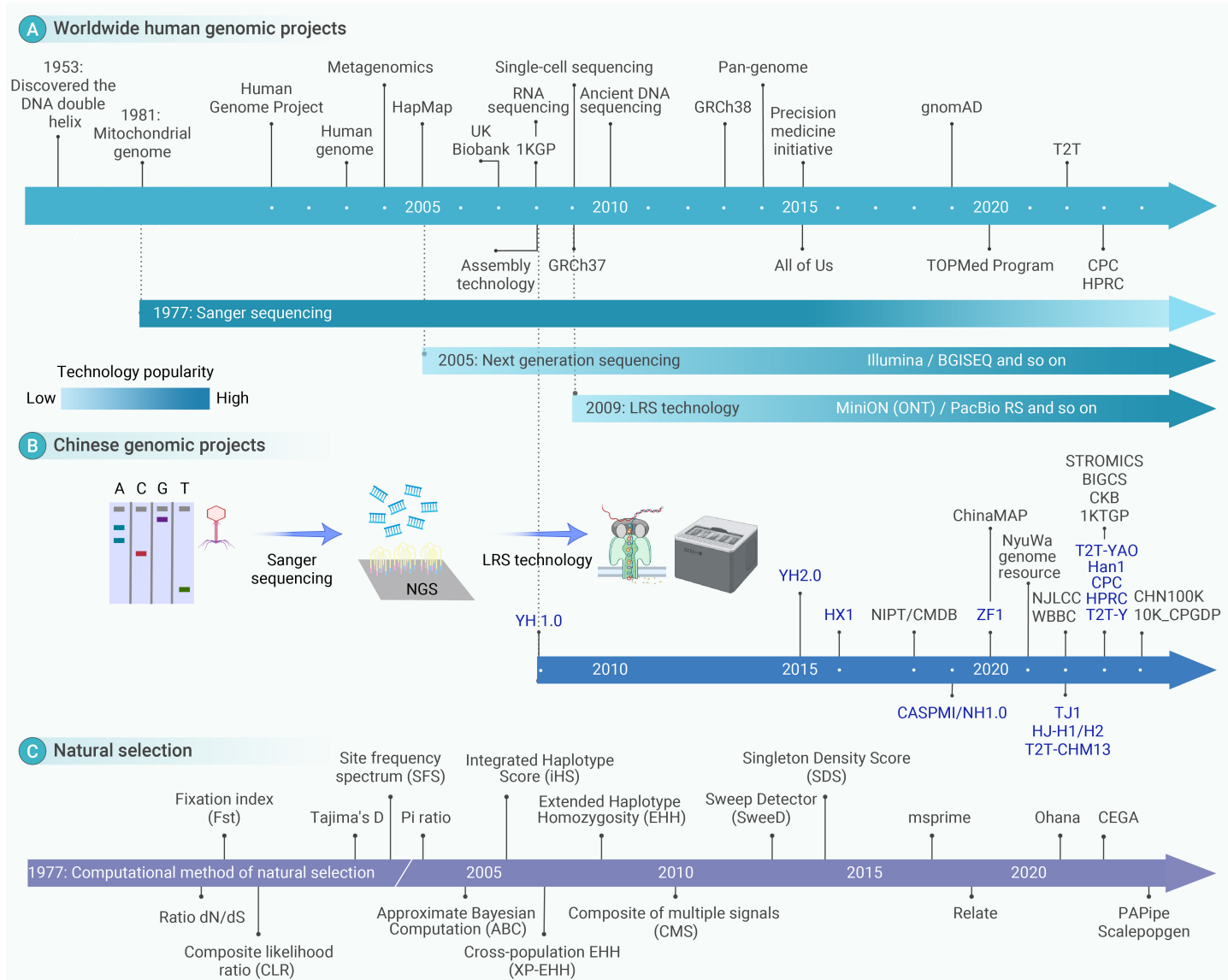


Figure 1. Timeline of key events in human genomic research, sequencing, and technological advances The top timeline highlights major human genome projects over the past two decades. (A) The middle timeline focuses on Chinese projects utilizing second- (NGS) and third-generation sequencing (TGS) (B) The bottom timeline illustrates the process of signature identification via various methods (C) The gradient bar indicates the popularity of each sequencing technology: first-generation (Sanger sequencing), second-generation (NGS, e.g., BGISEQ-T7 and Illumina), and long-read sequencing (LRS), including third- and fourth-generation innovations such as Oxford Nanopore Technologies and Pacific Biosciences). Essential events related to sequencing and computational innovations, paradigm shifts, and changes in research scale from archaic, ancient, and modern humans are highlighted. Abbreviations: HapMap: Human Haplotype Map Project; 1KGP: 1000 Genomes Project; GRCh37/38: Genome Reference Consortium 37/38; gnomAD: Genome Aggregation Database; TOPMed: Trans-Omics for Precision Medicine; T2T: Telomere-to-Telomere; CPC: Chinese Pangenome Consortium; HPRC: Human Pangenome Reference Consortium; NGS: Next-generation Sequencing; LRS: Long-Read Sequencing; YH1.0/2.0: YanHuang 1.0/2.0; CASPMI: Chinese Academy of Sciences Precision Medicine Initiative; 10K_CPGDP: 10K Chinese People Genomic Diversity Project; CONVERGE: China, Oxford, and Virginia Commonwealth University Exper; NH1.0: Northern Han 1.0; T2T-CHM13: Telomere-to-Telomere CHM13; NJLCC: Nanjing Lung Cancer Cohort; ChinaMAP: China Metabolic Analytics Project; CKB: China Kadoorie Biobank; WBBC: Westlake BioBank for Chinese; BIGCS: Born in Guangzhou Cohort Study; 1KTGP: 1000 Tibetan Genomes Project; NIPT: Non-invasive prenatal testing; CHN100K: China 100K.

recognition of East Asian genetic diversity, impeding the analysis of complex diseases and the identification of mutations linked to rare Mendelian diseases.⁴⁹⁻⁵¹ Population genetic evidence suggests that clinically significant variants are prevalent in underrepresented populations.^{8,19,20,42,52-54} Therefore, large-scale studies on Chinese populations are essential to elucidate genetic variation patterns and the genetic basis of complex diseases and traits.^{31,55,56}

We comprehensively summarized genomic initiatives and their applications from multiple perspectives. We first detailed advances in human genomic research, emphasizing the extensive Chinese genomic resources transitioning from array-based precision medicine genomics (Figures 2-3) and outlined the general pipelines for genomic discovery in large-scale sequencing projects, including variation identification, population evolutionary history reconstruction, medical relevance exploration, and haplotype reference panel reconstruction (Figure 3). We then discussed the influence of large-scale Chinese genomic projects and complex evolutionary events on

human disease and health and highlighted the adaptive biological signatures identified under evolutionary pressures from extreme environments, dietary shifts, and pathogen exposure (Figure 4). Additionally, we described innovations from ancient and archaic genomes, highlighting their impact on modern human disease and health through archaic introgression and admixture (Figure 5). We finally presented the role of long-read sequencing (LRS) technology in revealing complex genetic variations and their applications (Figure 5) and discussed future directions to overcome the under-representation of Chinese genomic initiatives, including large-scale genomics-phenomics database reconstruction, data sharing, and multiomics transitions from genomics to phenomics. We emphasized the urgent need for a high-quality, large-scale genomics-phenomics database to characterize genetic diversity, offering a comprehensive genome blueprint for Chinese populations and insights into associations between human diseases and population-specific or rare genetic variants.

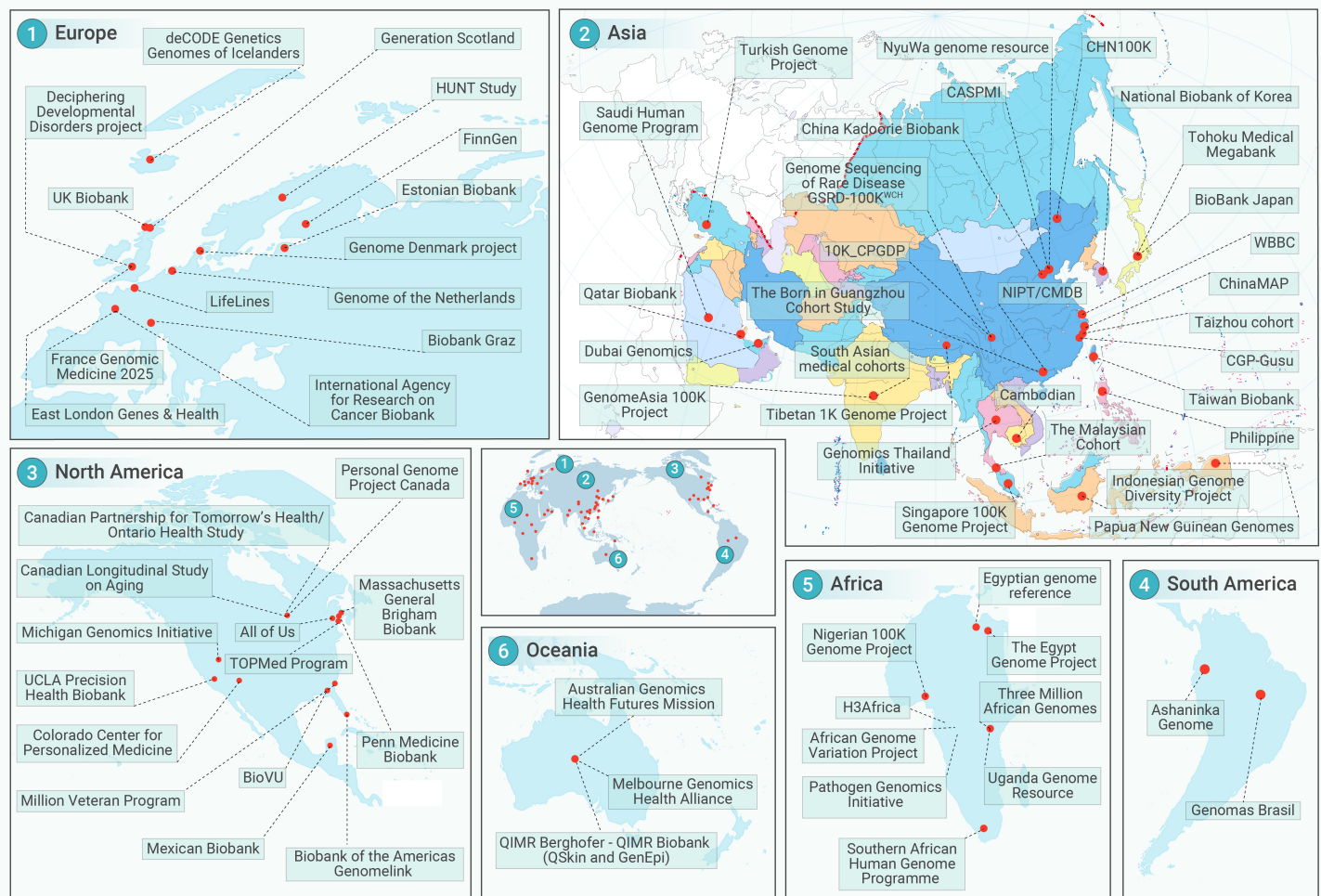


Figure 2. The geographical distribution of large-scale worldwide human genomic projects or regional biobanks The famous large-scale continental population cohort genomic research is highlighted. Abbreviations: HUNT study: The Trøndelag Health Study; BioVU: Vanderbilt DNA Databank. The worldwide and Asian maps were officially approved with the number GS(2016)1611 and GS(2020) 4393 (<http://bzdt.ch.mnr.gov.cn/>). The annotations on the map represent only the general geographical location.

PROGRESS IN LARGE-SCALE POPULATION GENOMIC PROJECTS

Addressing European bias and disparities in Chinese genomic resources

Advances in sequencing technologies, from first-generation Sanger sequencing to next-generation sequencing and then to LRS platforms such as PacBio HiFi and Oxford Nanopore Technologies (ONT), have propelled genomic research in human medicine and population genetics. The HGP marked a significant milestone by providing the first human reference genome. Following this, the HapMap project established a comprehensive database of genetic variation patterns, laying the groundwork for human health and disease studies. The 1KGP furthered this effort by reconstructing the genomes of 2,504 individuals from 26 global populations via low-resolution sequencing, thereby elucidating common human genetic variants and their implications for disease studies.⁵ The expanded 1KGP, with 3,202 high-coverage whole-genome sequences, revealed additional variants, enhancing our genetic understanding.⁴ Several large-scale cohort studies are also underway (Figures 1-2). For example, Choudhury et al. reported 426 high-depth African genomes from 50 ethnolinguistic groups, revealing numerous novel variants and improving our understanding of African genomic diversity and genetic diseases.¹⁴ The UK100K project identified significant trait associations for rare variants, advancing our knowledge of the links between genetic variation and phenotypic diversity.⁹ The "All of Us" initiative, with data from more than 413,000 participants, aims to personalize medical prevention, treatment, and care, whereas Australian Genomics explores genomic testing in healthcare.^{11,53} However, these projects have focused predominantly on populations of European ancestry, leading to a significant underrepresentation of genetic diversity in Asia. Addressing this Eurocentric bias is crucial for a more comprehensive understanding of global genomic diversity.³⁴

The GAsP100K project aims to address the underrepresentation of non-European individuals in human genetic research. The sample included 1,739 individuals from 219 populations across 64 Asian countries. This project seeks to identify rare disease-associated genes and improve health outcomes for Asians.¹⁷ Wu et al. reported preliminary findings from the SG10K pilot study, which involved 4,810 Singaporean Chinese, Malay, and Indian individuals with an average sequencing depth of 13.73x. This initiative, which is part of the Singapore National Precision Medicine Program, fosters research on genetic traits and complex diseases, providing a valuable genomic resource for Asians.^{29,30} The Tohoku Medical Megabank Organization sequenced nearly 4,000 Japanese individuals, creating population-specific allele frequency and haplotype reference panels that enhance the understanding of East Asian genetic diversity.^{27,57,58} The Korean Genome Project sequenced 1,094 high-coverage genomes, identified millions of Korean-specific variants and produced a reference panel that improves imputation accuracy, preserving genotypic data from previous studies.²⁸ However, genomic studies in East Asia have revealed significant population coverage imbalances, particularly the underrepresentation of Chinese populations and ethnolinguistically distinct minority ethnic groups.

Imbalance of genetic diversity within ethno-linguistically diverse Chinese populations

The emergence of biobank-scale genomic research. The Chinese Human Genome Diversity Project (CHGDP), initiated 26 years ago, aimed to study human evolution via microsatellites, identifying population stratifications between northern and southern East Asians.^{59,60} Subsequent large-scale studies using phylogenetic informative single nucleotide polymorphisms (SNPs) from Y chromosomes and mitochondrial DNA revealed that modern

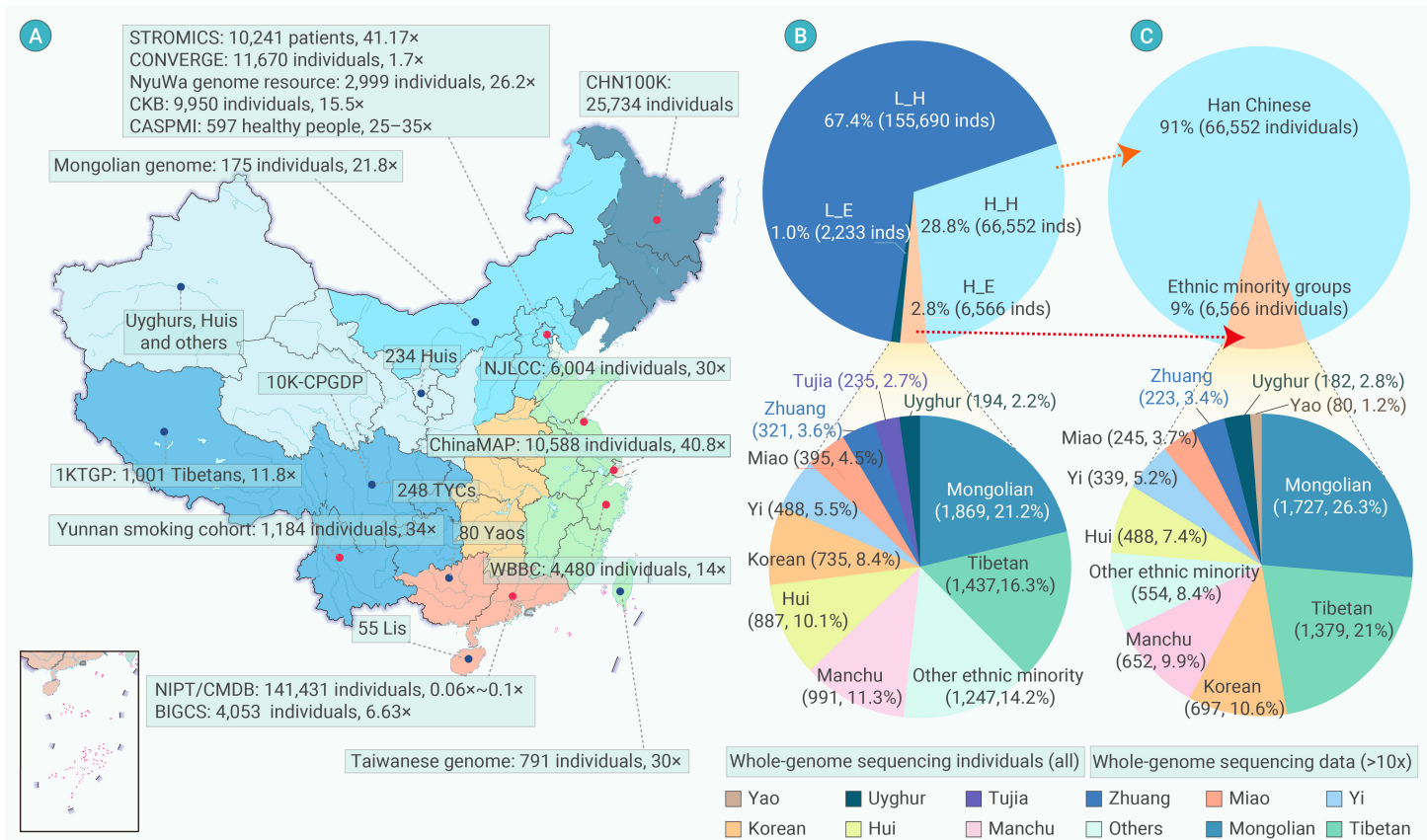


Figure 3. Overview of Chinese genomic projects (A) Locations of institutes and populations targeted by Chinese genomic programs. The red points represent medical-driven cohorts focused on the genetic architectures of human diseases, whereas the blue dots indicate anthropologically informed research on population evolutionary history. The base maps were officially approved with the number GS(2022)4309 (<http://bzdt.ch.mnr.gov.cn/>). The annotations on the map represent only the general geographical location. (B) The ethnic composition of all whole-genome sequenced Chinese individuals, with the upper chart showing two subgroups (Han Chinese and ethnic minorities) and the lower chart including 8,799 ethnic minority individuals. (C) The ethnic composition of individuals whose sequencing depth was greater than 10. Abbreviations: L_H: Han Chinese populations with low-coverage sequencing data; L_E: Ethnic minority with low-coverage sequencing data; H_H: Han Chinese populations with high-coverage sequencing data; H_E: Ethnic minority with high-coverage sequencing data.

East Asians originated from Africa. The genetic landscape of contemporary Chinese populations was shaped by the complex migration and admixture of ancient populations from North China, which was associated with Yangshao culture, and South China, which was related to rice farmers.^{41,61–67} The HUGO Pan-Asian project, in 2009, reported a large-scale survey of autosomal variations from linguistically diverse Asians, showing a strong correlation between ancestral components and linguistic affiliations,⁶⁸ which were recently confirmed via large-scale Y-chromosome and mitochondrial whole genome sequences.^{65,69} In the same year, Chen et al. genotyped 350K SNPs from more than 6,000 Han Chinese individuals across ten provinces, whereas Xu et al. collected high-density SNP data from more than 1,700 Han Chinese individuals from 26 regions in China to study genetic origins and population substructures.^{70,71} The admixture modeling results indicated that the observed north-south genetic clines among geographically diverse Han Chinese individuals aligned with historical migration patterns.

Sequencing and array-based genotyping have become cost-effective techniques for obtaining genomic resources. Rapid advancements in sequencing, bioinformatics, and data computing have enabled large-scale population and medical genomic projects to explore genomic variation at a fine scale, identify population-specific genetic variations and their associations with complex phenotypic traits, reveal the genetic architectures of Mendelian diseases, and develop screening panels and diagnostic strategies, accelerating Chinese precision medicine initiatives. In 2016, Chinese national genomic initiatives launched general population cohorts (GPCs) of more than a million individuals, focusing on populations from southwestern, southern, northeastern, northwestern, and eastern China.³³ We summarize the findings from the “proof-of-concept” phase of these projects.

The CONVERGE (China, Oxford, and Virginia Commonwealth University Experimental Research on Genetic Epidemiology) consortium recruited

11,670 Han individuals and conducted low-coverage sequencing (1.7×) of these samples, providing China's first large-scale WGS resource.⁷² The Han Chinese Genome Initiative comprehensively captured the genetic diversity of geographically diverse Han populations. The PGG.Han database was subsequently developed to archive whole-genome sequences from the CONVERGE project and high-density genome-wide SNP data from 114,783 Han Chinese individuals.⁷³ The Non-invasive Prenatal Testing (NIPT) project analyzed WGS data from 141,431 Chinese women at an average depth of <0.1× using computational innovations to estimate the allele frequency of common variants, reported fine-scale genetic substructures of Chinese populations and identified new causal variants for maternal phenotypes.⁷⁴ The Chinese Millionome Database (CMDB) contains population-specific genomic data and an allele frequency spectrum, providing a critical reference for allele frequency and linkage disequilibrium (LD) patterns for statistical algorithm development and polygenic risk score (PRS) modeling for Chinese people. These initial Chinese genomic resources have promoted precision medicine and accelerated the application of sequencing technology in congenital disabilities, cancers, and infectious diseases.^{74,75}

High-quality population-scale genomic datasets. Low-coverage WGS projects often lack the resolution to fully identify genetic variants, predominantly low-frequency, rare, or singleton variants in complex genomic regions. Consequently, several high-quality Chinese precision genomic projects have been launched. For instance, the Chinese Deafness Genetics Consortium (CDGC) initiated the Genome Sequencing of Rare Diseases in 100,000 people in China (GSRD-100K^{WCH}) to enhance precision prevention and diagnosis of Mendelian-inherited deafness and other rare diseases.⁷⁶ To understand the genetic determinants of complex diseases such as metabolic disorders, cancers, and stroke, high-quality, high-depth, medical-driven WGS projects have been conducted (Figure 3).^{31,55,56,66,69,77–81}

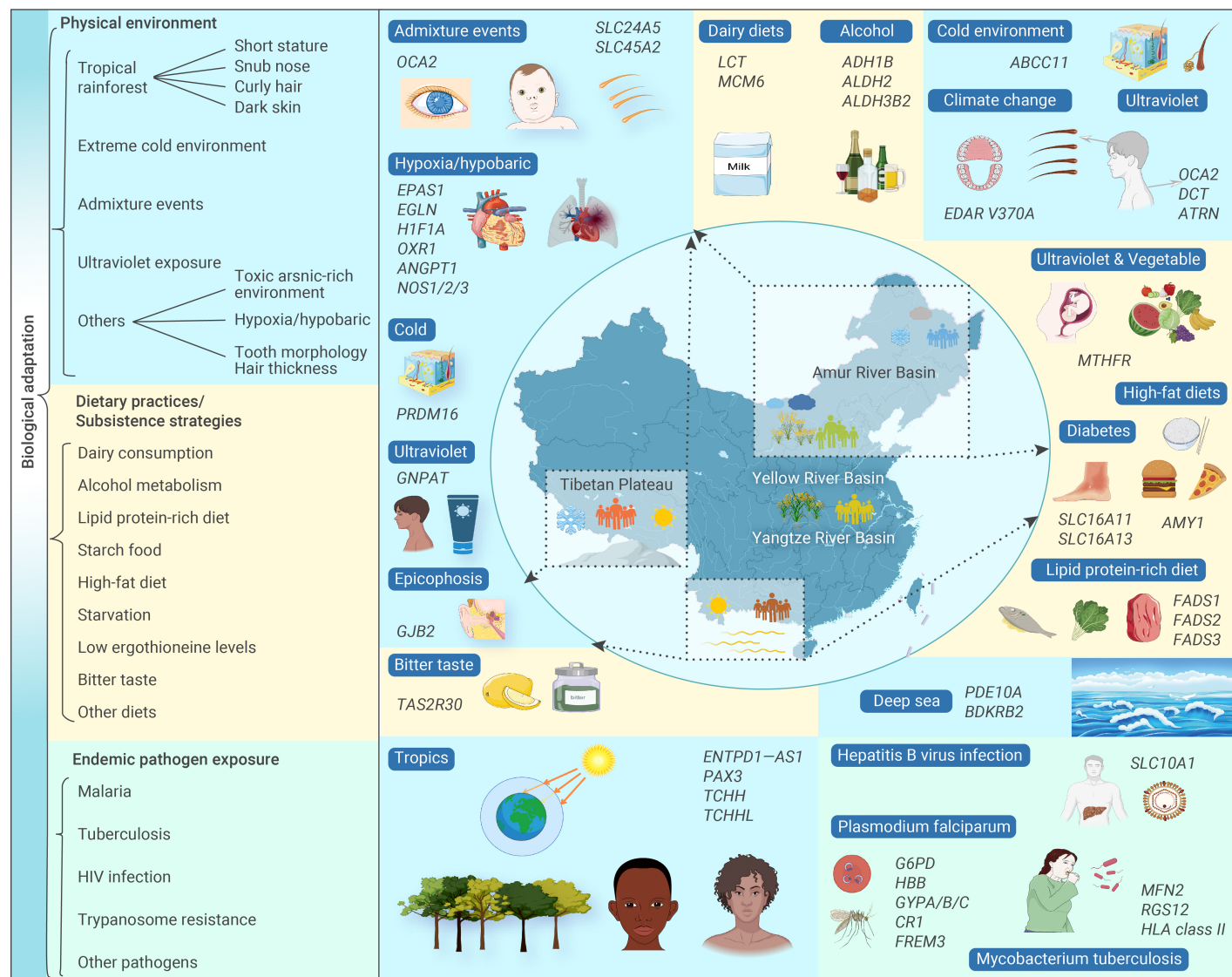


Figure 4. Evolutionary forces shaping biological adaptations in Chinese populations and neighbors The left panel outlines three major evolutionary forces (extreme environments, dietary practices, and endemic pathogen exposures) that impact biological adaptations, including specific influencing factors and associated genes. The right panel illustrates the geographical distribution of populations in East Asia, indicating positions in highlands and lowlands and directions from south to north and west to east. The base maps were officially approved with the number GS(2022)4309 (<http://bzdt.ch.mnr.gov.cn/>). The annotations on the map represent only the general geographical location.

The China Metabolic Analytics Project (ChinaMAP) collected 10,588 high-coverage WGS datasets (40.8×) to create a precision medical system for the Chinese population. This project reported a high-quality genetic variation dataset, explored the population structures of different Chinese regions, and identified medically relevant variations.³¹ The Nanjing Lung Cancer Cohort (NJLCC), which is part of the GuSu project by the Jiangsu government and Nanjing Medical University, involved 6,004 high-coverage genomes (30×) and provided comprehensive WGS data for non-small cell lung cancer (NSCLC).⁵⁶ This project broadened our understanding of the genetic etiology of lung cancer and improved precision medicine for lung cancer patients. The Third China National Stroke Registry (CNSR-III) involved 10,914 patients with acute ischemic stroke or transient ischemic attack, and high-depth sequencing (41.17×) was used to reveal the genetic mechanisms of stroke and identify potential drug targets for secondary prevention.⁸² The China Kadoorie Biobank (CKB), encompassing over 512,000 individuals from 10 regions, provided new insights into the genetic basis of complex diseases and the relationships between risk factors and disease outcomes through GWASs and high-depth WGS.^{79,80}

The Westlake BioBank for Chinese (WBBC) pilot project reported genomic data for 4,535 individuals (14× depth) and 5,841 high-density genotyping individuals. This panel improved the imputation quality for low- and rare-

frequency variants (MAF < 5%) in Chinese populations.⁵⁵ The NyuWa genome resource reports the genomic landscapes of SNPs, complex structural variations (SVs), and adaptive biological signatures from high-coverage genomes (26.2×) of 2,999 Chinese individuals, aiming to accelerate genetic and precision medicine research in China.⁷⁷ The Taiwan Biobank, comprising over 150,000 individuals, provides a rich biomedical research database for genetic and epidemiological studies, especially for AN-speaking populations on Taiwan Island.⁸³ Prospective birth cohorts in Europe and China have also been launched. The Born in Guangzhou Cohort Study (BIGCS) investigated the genetic history of South Chinese populations and new GWAS signatures associated with maternal and infant phenotypes.⁵² The Chinese Academy of Sciences Precision Medicine Initiative (CASPMI) focused on the phenotypic dynamics of specific worker populations, revealing genetic architectures related to vitamin B12, circulating folate, and lipid metabolism.⁸⁴ The Taizhou Longitudinal Study and HuaBiao Project reported high-coverage exon data to illuminate environmental and genetic risk factors for non-communicable diseases.⁸⁵

The anthropologically informed sampling of underrepresented populations is crucial for obtaining a comprehensive catalog of human genetic variation. Significant efforts have focused on Turkic-speaking people, Sino-Tibetan-speaking Hui, Tibetan, Sherpa, and Deng people, Tai-Kadai-speaking Li, and

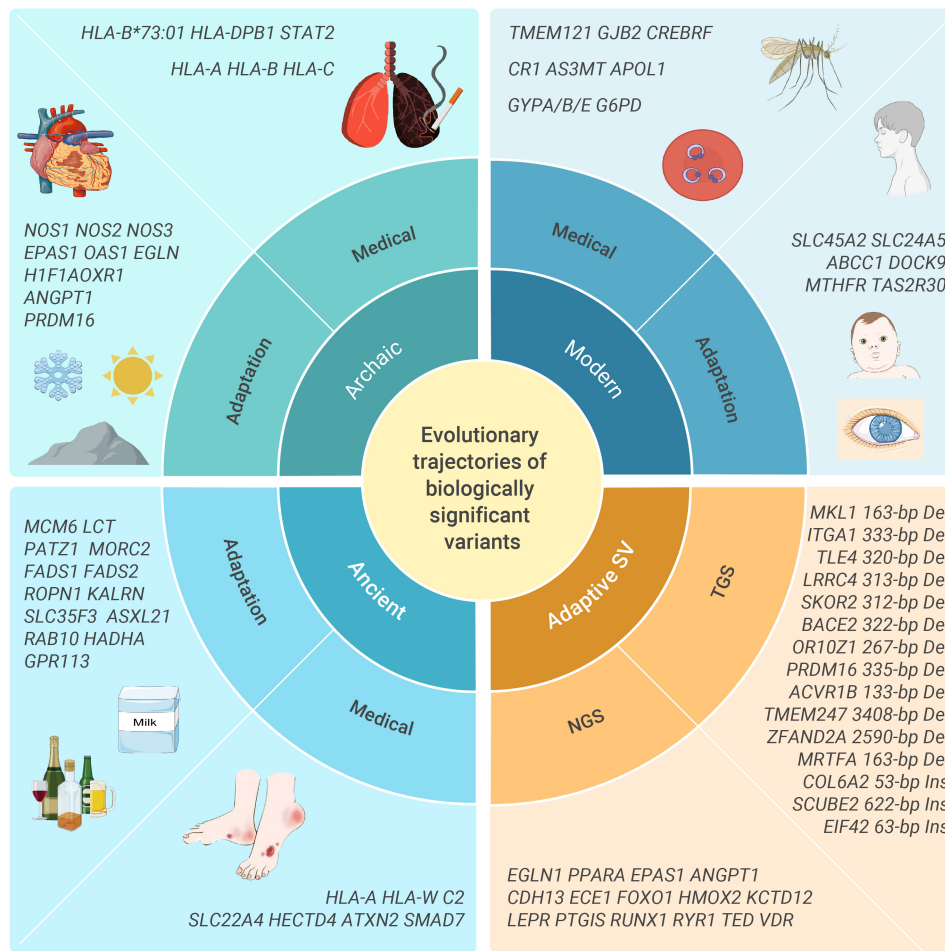


Figure 5. Representative genes associated with biological adaptation or medical consequences were identified from modern human adaptation, archaic introgression, and ancient human admixture. The left and upper right colors indicate the populations: archaic, ancient, or modern, each subdivided into adaptation-related or medically essential genes. The lower right section displays adaptive structural variations (SVs) inferred from NGS and TGS techniques.

Chinese into four subgroups: North Han, Central Han, South Han, and Lingnan Han.⁵⁵ Genetic evidence from ChinaMAP further reveals different categories and suggests that the Han Chinese population can be divided into seven clusters: Northwest Han, North Han, East Han, Central Han, Southeast Han, South Han, and Lingnan Han.³¹

Moreover, anthropologically informed genomic projects have reported fine-scale population substructures among ethno-linguistically different Chinese populations. The trans-Eurasian people from the crossroads between Siberia and North China possess Tungusic-dominant ancestry from the Amur River Basin and Turkic-like western Eurasian admixture profiles in Northwest China.^{87,95,96} The Tibeto-Burman-speaking Tibetan people, a long-settled highland population on the Tibetan Plateau, have been extensively studied for their settlement, admixture, and biological adaptation history.^{90,97,98} The evidence indicates that the major ancestral component of the Tibetan people originated in North China, showing significant genetic differentiation from the Han Chinese people and diverse adaptive land-

scapes among geographically distinct groups.^{38,99} Population stratification among linguistically different Tai-Kadai, Hmong-Mien, Austroasiatic, and Austronesian people in South China and their differentiation from southern Han Chinese people have also been demonstrated.^{86,89,100} Overall, these findings refine the understanding of genetic structure and diversity, enabling a more precise reconstruction of population evolutionary history and advancing applications in precision medicine. However, the underrepresentation of geographically and ethnically diverse populations remains a critical limitation, complicating the accurate interpretation of population clustering patterns.^{52,101} Despite significant advancements in genomic research, which have elucidated the genetic substructures and diversity of Han Chinese and other East Asian populations, further efforts are needed to achieve a more comprehensive representation and integrate these insights into evolutionary and medical research.

Fruitful achievements in large-scale Chinese precision medicine genomic projects

Evolutionary history reconstruction and fine-scale genetic substructure exploration. Han Chinese, the world's largest population with more than 1.4 billion individuals, exhibits extensive genetic diversity. Previous studies using low-coverage sampling, low-density SNP data, and qualitative methods such as fixation indices, principal component analysis, and ADMIXTURE analysis have identified genetic differences among geographically diverse Han Chinese individuals. However, the lack of comprehensive research using high-depth WGS and nationwide sampling has limited our understanding of the intricate genetic diversity of East Asia. To address these gaps, researchers have launched numerous Chinese genomic projects. Notably, the PGG project has provided a comprehensive depiction of the genetic diversity and population stratification of geographically diverse Han Chinese individuals. This study identified six major Han subgroups, revealing greater divergence in the southern subgroup than in the northern subgroup.⁷³ The WBBC pilot project further revealed that genetic differentiation within Han Chinese people follows a north-to-south cline shaped by geographical barriers such as the Qinling-Huaihe River Line and Nanling Mountains, categorizing Han

scapes among geographically distinct groups.^{38,99} Population stratification among linguistically different Tai-Kadai, Hmong-Mien, Austroasiatic, and Austronesian people in South China and their differentiation from southern Han Chinese people have also been demonstrated.^{86,89,100} Overall, these findings refine the understanding of genetic structure and diversity, enabling a more precise reconstruction of population evolutionary history and advancing applications in precision medicine. However, the underrepresentation of geographically and ethnically diverse populations remains a critical limitation, complicating the accurate interpretation of population clustering patterns.^{52,101} Despite significant advancements in genomic research, which have elucidated the genetic substructures and diversity of Han Chinese and other East Asian populations, further efforts are needed to achieve a more comprehensive representation and integrate these insights into evolutionary and medical research.

Construction of population-specific high-quality haplotype-resolved reference panels. Array-based genotyping or low-coverage WGS captures only a fraction of human genetic variation and is limited in identifying rare or population-specific variants.⁸¹ High-coverage WGS can reveal many genetic susceptibility variants but at a significant cost. Genome-wide genotype imputation has emerged as a cost-effective solution for predicting missing genotypes based on known haplotype information in a given reference panel. This approach increases the coverage of genetic markers, improves the efficiency of detecting new causal associations, and enhances the success rate of screening trait-related causative loci.⁸⁰ The development of population-specific haplotype reference panels is crucial, as is the leveraging of known haplotype linkage information from large-scale cohort genomic sources to increase imputation accuracy. Combining low-depth WGS/array genotyping with a highly effective genotype imputation strategy optimizes data utilization and cost savings. However, constructing robust haplotype reference panels unrestricted by sample size, sequencing depth, or specific population composition is challenging.

Haplotype reference panels based on the 1KGP, Haplotype Reference Consortium (HRC), UK10K, and TOPMed have been developed over the past two decades alongside global or regional genomic projects.^{8,12,17,81} Despite their utility, unbalanced ancestral diversity limits the imputation coverage of genomic variants in Chinese populations. The 1KGP-based haplotype reference panel is widely used for genotype imputation.⁴ However, owing to low sequencing coverage and limited population coverage, its imputation accuracy is constrained by rare/low-frequency variants, which are often population-specific. The HRC, UK10K, and TOPMed panels, while significantly improving imputation efficiency for Europeans, are not applicable to East Asians due to the underrepresentation of East Asian ancestry.¹⁰² The European bias in existing panels has hindered phasing and imputation in Chinese populations, limiting the understanding of their genetic architecture and disease etiology. Establishing a high-quality, population-specific reference panel is essential and requires a high degree of ancestry matching and coverage of genetic diversity for successful genotype imputation.

Efforts have been made to construct more refined population-specific reference panels to increase imputation accuracy in Chinese populations. The CONVERGE program published the first Chinese-specific haplotype reference panel, including 11,670 genomes, identifying approximately 22 million high-quality SNPs based on the innovations of bioinformatics.^{32,103,104} PGG.Han developed a Han-Chinese-specific reference panel to optimize genotype imputation, aiding in the evolutionary history and precision of Chinese medicine.⁷³ Compared with the 1KGP, the WBBC-based reference panel incorporated more than fifteen times more Chinese samples, improving the imputation quality for rare and low-frequency variants.⁵⁵ Combining WBBC with other panels, such as 1KGP and CONVERGE, resulted in better performance for the Chinese population. The updated SEAD integrated 22,134 haplotypes from the WBBC, SG10K, and GenomeAsia pilot projects, aiding in identifying associations between rare variants and hip bone mineral density.⁸¹

Other efforts to create population-specific high-quality reference panels for whole-genome sequences have increased the accuracy of imputing common and rare genetic variants, including singletons.¹⁰⁵ The NyuWa genome resource, for example, introduced a Chinese-specific haplotype reference panel that notably reduced imputation error rates for Han Chinese individuals by 30% to 51%.⁷⁷ Despite these advancements, the imputation performance of population-specific panels for complex diseases and specific populations, such as Tibetans, requires further enhancement to reduce health disparities.⁹⁰ The 1KTGP offers a population-specific genome reference panel to better understand refined genetic architecture, redefine natural selection footprints, and explore disease susceptibility associations among Tibetans. Similarly, a Mongolian-specific high-quality haplotype reference panel reflecting population stratifications from northern China has aided in inferring fine-scale demographic history and understanding the genetic basis of traits among Trans-Eurasian speakers.¹⁰⁶

Large-scale medical cohorts focusing on particular diseases have also developed interactive databases that integrate genomic and multi-omics big data along with corresponding high-quality reference panels. These databases play crucial roles in the prevention, diagnosis, and development of new drugs for critical chronic diseases, tumors, and genetic disorders, such as CKB,^{80,107} ChinaMAP,³¹ BIGCS,⁵² and NJLCC.⁵⁶ However, these databases and reference panels have limitations in representing genetic diversity. For example, the PGG.Han database and the NyuWa database predominantly focus on Han Chinese populations, whereas the ChinaMAP database mainly covers the central and northern Chinese populations despite including eight ethnic groups from China. Geneticists aim to conduct ongoing genomic programs to develop more efficient integrated population-specific haplotype databases to advance individualized precision health management, disease diagnosis, and treatment.⁸¹ A crucial aspect of this effort involves establishing an integrated genomic database with standardized analysis and benchmarking pipelines to facilitate the translation of genomic findings into clinical practice.

Medical relevance and genetic discovery. Large-scale human genomic projects are crucial for identifying genetic variations across diverse populations, including SNPs, insertion/deletion polymorphisms (InDels), SVs, and functional elements (promoters/enhancers/miRNA binding sites). These

projects enhance our understanding of the genetic basis and inheritance of human traits and diseases. Genetic variants can be categorized by non-reference allele frequency (AF): singletons (AC=1), doubletons (AC=2), very rare variants (AC>2 and AF≤0.1%), rare variants (0.1%<AF≤1%), low-frequency variants (1%<AF≤5%), and common variants (AF>5%). Efforts to capture genetic diversity have enriched the map of human variants and improved health outcomes, reducing health disparities.¹⁰⁸ The PGG.Han project identified ~25,057K variants from 114,783 Han Chinese individuals. ChinaMAP found 68.3 million novel variants from 10,588 high-depth WGS genomes, offering vast genetic resources for studying metabolic characteristics and disorders in Chinese populations. The NJLCC discovered new common, low-frequency, and rare variants associated with lung cancer risk in 2,984 NSCLC patients and 3,020 healthy individuals. The NyuWa resource, which analyzes 2,999 whole-genome sequences of Chinese diabetic patients and controls, detected 25 million variants, including 22,504 loss-of-function (LOF) variants, aiding in understanding chronic diseases in Chinese populations. Low-frequency and common variants play significant roles in health and disease, whereas rare variants remain poorly understood, particularly in the diagnosis of rare genetic diseases. Studies suggest that identifying low-frequency or rare variants improves with increased sequencing depth (>50×) and better imputation tools.

Accurate functional prediction of genetic variants identified by WGS is vital for understanding their biological consequences. Functional annotation via tools such as Variant Effect Predictors is necessary to predict physiological and molecular impacts. Annotation databases and tools enhance our understanding of genomic variation patterns and the functional impact of genetic variants on phenomics. The dbSNP database contains abundant SNPs, short InDels, and known human genomic variants. The Exome Aggregation Consortium (ExAC) integrates whole exome sequencing data from various ethnic groups, providing distribution frequency data to evaluate variant rarity and disease associations. The gnomAD database, an updated ExAC version, is the most considerable genomic database documenting high-confidence variant information from hundreds of thousands of individuals.¹² ClinVar, a public genetic resource, collects detailed variant information, including variant position, allele frequency, classification, and disease association, offering clinical annotations valuable for diagnosing genetic diseases and personalized medicine. These databases support large-scale genomic projects and precision medicine, contributing to a better understanding of human genetic complexity and diversity.

Genetic variants have varying effects on the biological adaptability of different populations. LOF variants can significantly impact phenotypes by disrupting normal biological processes or causing disease. Monogenic LOF variants often result in Mendelian diseases, following inheritance patterns such as dominant, recessive, and X/Y-linked inheritance.^{109,110} A notable example is cystic fibrosis (CF), which is caused by mutations in the *CFTR* gene, leading to damage to the respiratory and digestive systems.^{111,112} This mutation is prevalent among Europeans (1 in 2000–3000 births), rare among African Americans (1 in 17,000 births), and extremely rare in China (1 in 64,000–128,434 births). The clinical symptoms of CF vary between populations due to differences in genetic backgrounds, pathogenic variants, and mutation frequencies. To improve the global understanding of disease outcomes related to genetic variants, the American College of Medical Genetics and Genomics has established standards and guidelines for interpreting Mendelian disease variations, proposing 28 diagnostic criteria and five diagnostic conclusions.¹¹³ Genetic variants are categorized as pathogenic, likely pathogenic, of uncertain significance, limited penetrance, protective, risk factors, likely benign, or benign. These classifications aid in assessing clinical significance and enhancing diagnostic accuracy. Genomic databases also reveal patterns of genetic variants related to pharmacogenomics and other clinically essential traits.

Genotypic associations with disease phenotypes. Genomic resources have been effectively applied in GWASs and Mendelian randomization studies to explore associations between diseases and genetic variants or between exposures and outcomes.⁵² By scanning disease/trait-related genetic markers (typically SNPs) in large cohorts of diseased and healthy individuals, GWASs can be used to identify genetic variants and elucidate the genetic basis of specific traits or diseases. When a locus is significantly

associated with a disease, its surrounding sites are further investigated to uncover the underlying causes and mechanisms. Although the impact of rare variants on common diseases remains underexplored, GWASs have identified thousands of genetic variants linked to health and disease, increasing the number and scope of studies in China. Traditional GWASs primarily detect common and low-frequency variants via genotype imputation on microarray data with non-population-specific haplotype reference panels.⁸¹ However, potential trait-associated variants poorly imputed by public reference panels are missing. Consequently, several projects target specific populations to capture population-specific genetic variants and develop tailored reference panels, which are crucial for fine-mapping GWAS signals and identifying target genes.⁵⁴ Complex diseases such as type 2 diabetes, cancer, and hypertension often result from polygenic and environmental interactions rather than from single-gene effects. Genetic susceptibility highlights the role of genetic factors in disease onset and progression, with some variants influencing gene function, expression, and regulation, thereby affecting disease risk.^{114,115} The most common diseases involve polygenic inheritance, where cumulative effects of multiple variants significantly increase disease risk. Genetic susceptibility varies across environments and populations due to genetic background differences. Therefore, genetic variants alone do not fully explain disease risk; instead, they must be integrated with environmental and lifestyle factors for comprehensive disease risk assessment. Understanding the genetic and molecular basis of disease requires considering polygenic variants alongside non-genetic factors, encompassing rare pathogenic and common variants.¹¹⁶

BIGCS-based GWAS identified 11 loci linked to adult and infant traits, such as a missense variant in *SLC10A1* increasing cholestasis risk during pregnancy and *APOE* related to cholesterol levels.⁵² Other associations include *SOAT2* with low-density lipoprotein and *APOE* with high-density lipoprotein in infant cord blood. *APOA5* is related to triglyceride levels, and *CELSR2* is related to low-density lipoprotein in maternal peripheral blood.⁵² NIPT-related GWAS revealed 16 novel associations, including height-related *UBQLN2-LINC01420* and BMI-associated *PLD5-LINC01347*.⁷⁴ The STROMICS reported 56 independent genetic and stroke-related phenotype associations.⁷⁸ PRS, influenced by allele frequencies, LD patterns, and environmental factors, is valuable in genetic research and precision medicine, offering insights into genetic contributions to disease and enabling accurate risk assessments and treatment recommendations. Many pathogenic variants with strong effects exist in specific regions and ethnic minorities and are closely linked to complex genetic disease factors. Identifying genetic variants affecting traits and disease susceptibility is crucial for understanding individual genetic differences and disease mechanisms and advancing genetic research, drug development, and personalized medicine. Advances in sequencing technologies and new genetic discoveries have increased our ability to identify novel associations between genetic variants and specific traits or diseases.

Biological adaptations within East Asians

The human journey began in Africa approximately 200,000 years ago, with migrations spreading across diverse environments worldwide.¹¹⁷ From tropical rainforests with intense UV exposure to frigid Arctic zones, humans have adapted to various ecological conditions.^{118,119} These adaptations were influenced by large-scale migrations, admixture, and demographic changes, shaping subsistence strategies and dietary habits. Initially, hunting-gathering prevailed and later shifted to agriculture and pastoralism. This transition led to distinct dietary structures, with nomadic groups in western Eurasia consuming meat and dairy, ancient northern East Asians favoring millet, and ancient southern East Asians relying on rice. Such lifestyle shifts also triggered genetic adaptations, such as lactose tolerance, in populations with an animal domestication history.^{120,121} Additionally, challenging living conditions and pathogens, such as tuberculosis and malaria, play crucial roles in local human adaptations.¹²² For instance, humid tropical rainforests in Southeast Asia and sub-Saharan Africa facilitated mosquito-borne diseases such as malaria. Similarly, *Trypanosoma brucei rhodesiense*, which causes chronic kidney disease in sub-Saharan Africa, has high *APOL1* gene variant frequencies.^{123,124} This interplay of environments, diets, and pathogens has shaped distinct gene pools, driving adaptive evolution and natural selection

(Figure 4).^{125,126} Population-specific adaptive genetic variants influence corresponding phenotypes and complex traits, posing significant challenges to human health.^{126,127}

Biological adaptations shaped by extreme environments. Research on the biological adaptations of East Asians has focused primarily on populations at different altitudes and along the North-south axis. The Tibetan Plateau, known as the "Roof of the World," is one of the world's most unique landscapes. Its distinctive topography and meteorology create extreme high-altitude environments characterized by hypobaric hypoxia, frigid temperatures, and high levels of ultraviolet (UV) radiation. These conditions have historically discouraged permanent human habitation. However, indigenous populations have adapted to these harsh conditions since the late Pleistocene and are undergoing positive selection processes that result in unique phenotypes and traits for high-altitude adaptation.¹²⁸

Extensive research has provided compelling evidence of genetic adaptation in Tibetans based on profound advances via computational tools or deep learning methods (Figure 1C).^{104,129} Given the lower oxygen pressure at high altitudes, critical genes associated with the hypoxia-inducible transcription factor (HIF) pathway, such as *EPAS1* (encoding endothelial PAS domain protein 1), *EGLN1* (encoding egl-9 family hypoxia-inducible factor 1), and *HIF1A* (encoding hypoxia-inducible factor-1), have emerged as central players in this adaptation process.^{126,130,131} Additional functional variants in genes linked to hypobaric and hypoxic adaptation have also undergone positive natural selection. These genes include *OXR1* (encoding oxidation resistance gene 1), which protects against DNA damage and oxidative stress; *ANGPT1* (encoding angiotensinogen 1), which is relevant to vascular development and angiogenesis; and nitric oxide synthase (*NOS1/2/3*), which is crucial for regulating oxygen delivery. Notably, a nonsynonymous variant in *NOS1*, rs549340789, promoted hypoxic adaptation in Sherpas.⁹⁹ Highland Tibetans have undergone significant genetic adaptations to high UV radiation, resulting in distinctive skin pigmentation traits. An enhancer mutation of rs75356281 within the *GNPAT* gene is notably linked to darker skin pigmentation and increased melanin synthesis and occurs more frequently in Tibetans than in other populations.⁸⁸ Typically, proximity to the equator correlates with increased UV exposure, leading to darker skin as an adaptive response.¹³² However, the genetic basis of skin pigmentation in Tibetans differs markedly from that of other populations in low-latitude regions with extreme UV radiation, highlighting the unique environmental adaptations of Tibetans.

Similarly, the genetic mechanisms underlying hypoxia tolerance in marine Bajau nomads are distinct. Known as the 'Sea Nomads', the Bajau are indigenous Southeast Asians who live a subsistence lifestyle involving breath-hold diving.¹³³ Extensive natural selection has reshaped some phenotypes to increase their deep-sea survival. Unlike terrestrial populations, Bajau populations present enlarged spleens and an increased number of oxygenated red blood cells due to genetic variants in the *PDE10A* and *BDKRB2* genes. Another notable adaptation involves the combination of "low-temperature" and "high-altitude" or "high-latitude" environments, where glaciers are common, and many peaks remain snow-covered year-round. Research has highlighted a 335-bp deletion at 1p36.32 that facilitates Tibetan highlanders' local adaptation by downregulating *PRDM16* (PR/SET domain 16) expression.¹³⁴ *PRDM16* is crucial for converting white adipose tissue to brown adipose tissue as part of the thermogenic program and may also protect liver functions under extremely cold conditions.¹³⁵

Adaptation to dietary changes. China, with its vast area and two original agricultural centers, experienced a significant Neolithic transition from hunter-gathering to farming and animal domestication. The coastal regions of East Asia have developed specific subsistence strategies for fisher-hunter-gatherers, likely with high-animal-based diets. Significant differences in dietary habits exist between ancient and modern populations, ranging from north to south, inland to coastal, and highlands to lowlands. Human evolutionary history and geographical factors strongly influence genetic adaptation patterns.⁴¹ The shift from hunting-gathering to agriculture created diverse natural selection pressures, shaping local phenotypes. The earliest significant event in agricultural development occurred in the Fertile Crescent in Southwest Asia, where agricultural populations migrated northward to Europe and eastward to the Indian subcontinent approximately 10,500 years ago,

accelerating agricultural expansion. Ancient admixture and selection influenced modern European adaptation patterns.¹³⁶ Significant changes in human dietary structure resulted from these shifts in subsistence strategies. Additionally, cultural differences and environmental factors influence agricultural development and expansion in China and East Asia.

Genes involved in alcohol metabolism, such as *ADH1B*, *ALDH2*, and *ALDH3B2*, have undergone significant natural selection in East Asian populations, including Japanese and Korean populations.⁷⁴ Since the Neolithic era and the advent of agriculture, fermented foods and alcohol have influenced human health, contributing to conditions such as cardiovascular disease, stroke, and cancer. Northern Chinese populations, including ethnic minorities such as Tibetan, Mongolian, and Yi, typically exhibit a greater ability to metabolize alcohol than southern Chinese and Han populations do. This difference is linked to agricultural practices and rice domestication. In contrast, lactose intolerance is prevalent among Chinese people and other non-pastoralist Asians and West Africans due to the absence of dairy in their traditional diets. However, populations in Europe, North America, Central Asia, and parts of Africa have developed lactase persistence through natural selection and cultural adaptation over the past 10,000 years.¹³⁷ This trait is a prime example of convergent evolution involving genetic and cultural factors. The *LCT* and *MCM6* genes are crucial for lactase persistence, with the C/T-13910 SNP being a key regulator of *LCT* transcription in Europeans. In China, the frequency of lactase persistence-related SNPs remains low, except in pastoralist groups, which are influenced by agricultural dietary patterns that favor plant-based diets. Populations that consume more meat or marine products, which are scarce in plant-based diets, have higher levels of long-chain polyunsaturated fatty acids (LC-PUFAs), which are essential for cell and neuronal membrane development. Positive natural selection and local adaptation have shaped the *FADS1* and *FADS2* genes in different populations.¹³⁸ Ancestral *FADS* haplotypes are prevalent in regions where crops such as rice are grown, whereas derived haplotypes are common in hunting and fishing communities. Unique positive selection signals linked to specific diets and agricultural expansion pressures since the Neolithic period have been identified, such as the *TAS2R30* gene in the Dai people of Yunnan Province, who prefer bitter foods.¹³⁹

Adaptation to pathogen exposure. Pathogen exposure has significantly driven local human adaptation and evolutionary selection. Southeast Asia, located at the equator, features unique landscapes, such as tropical rainforests with high temperatures and humidity, which are ideal for parasites and other pathogens. Malaria, a common infectious disease in this region, is caused by mosquito-borne protozoan parasites. Severe malaria, which is especially dangerous to children and pregnant women, is a major cause of mortality in tropical areas such as sub-Saharan Africa.¹⁴⁰ Interestingly, populations in malaria-prone regions present lower proportions of severe cases, indicating the presence of genetic variants and adaptive signatures that confer malaria resistance. These include variants in genes such as *G6PD*, *HBB*, *GYPAB/C*, *CR1*, and *FREM3*. Historical and epidemiological data also highlight South China as a malaria hotspot. Residents in these pathogen-endemic areas developed genetic resistance to malaria. However, these protective variants also lead to blood disorders, such as sickle cell anemia, *G6PD* deficiency, and thalassemia, due to balancing selection.

Tuberculosis, caused by *Mycobacterium tuberculosis*, is prevalent in East Asia, particularly among Western Han Chinese people. It poses a severe threat to global public health. Originating in Africa approximately 40,000 years ago, *M. tuberculosis* spread worldwide with human migration and admixture.¹⁴¹ Large-scale GWASs have identified susceptibility variants in immune-related genes. Studies on Han Chinese individuals have revealed variants in the *MFN2*, *RGS12*, and *HLA* class II genes.¹⁴² Host-pathogen coevolution and selective pressures have led to adaptive variants associated with tuberculosis, notably in Africans but also in East Asians, primarily in the human leukocyte antigen (HLA) region. Leprosy, caused by *Mycobacterium leprae*, is another chronic infectious disease affecting the Chinese population.¹⁴³ GWASs have shown that genetic susceptibility to leprosy varies among populations. In Chinese populations, loci related to leprosy susceptibility are often located near immune-related genes, reflecting adaptation to pathogenic environments and positive selection through evolutionary events such as admixture, migration, and drift. Additionally, AIDS, HBV, and other

infectious diseases are prevalent in East Asia, highlighting the need to understand genetic bases and reconstruct evolutionary trajectories to improve human health.

Ultraviolet radiation-related adaptations. The *SLC24A5/2* gene is crucial for determining skin pigmentation diversity, with the derived allele frequency nearly fixed in Europeans due to positive selection.¹³⁶ This allele also drove convergent Eurasian evolution during out-of-Africa dispersal. Yang et al. reported that the *OCA2* gene and its His615Arg variant are prevalent in East Asians, especially northern Han individuals, indicating the role of *SLC24A5* in recent East Asian evolution.¹⁴⁴ Zhang et al. reported that the *OCA2*-His615Arg variant first appeared in ancient Liangdao, an early Neolithic population in southern coastal China.¹⁴⁵ Positive selection for this variant led to lighter skin in northern Han Chinese individuals due to reduced UV radiation at high latitudes. In addition to *OCA2*, other pigmentation genes (*DCT* and *ATRN*) also present natural signatures in East Asians. Studies have elucidated the genetic mechanisms underlying morphological traits in Southeast Asian aboriginal populations, particularly hunter-gatherers in tropical rainforests.^{118,119,144} These populations have adapted to local humid, hot environments and limited food resources, resulting in traits such as dark skin, short stature, wide or snub noses, and curly hair. Missense variants in genes such as *Trichohyalin* and *Trichohyalin 1* (*TCHH* and *TCHH1*) contribute to the curly hair phenotype in Cambodian aborigines, aiding in heat dissipation through sweat evaporation. Additionally, variants in *PAX3* and *ENTPD1-AS1* are involved in nose morphology and short stature, respectively.¹¹⁹ These examples illustrate how populations adapt to their unique environmental challenges.

Prior/post-Admixture Adaptation. Population dynamics and distinct demographic histories result in unique patterns of biological adaptation signals across regions. The Uyghur population, a representative admixed group with European-like and East Asian-like ancestry in Northwest China, exemplifies this.⁴² Extensive genetic evidence has revealed their complex ancestral origins and evolutionary history, revealing four major ancestral components: South Asian, European, Siberian, and East Asian.^{39,87,146,147} Uyghurs have undergone two significant admixture events, resulting in unique natural adaptive signals and variations due to long-term isolation and recent gene flow. These candidate genes are relevant to facial morphology, skin pigmentation, and metabolism.¹⁴⁷ Compared with Han Chinese individuals, Uyghurs present distinct facial features and lighter skin, which can be attributed to admixture from genetically divergent ancestors. Ancient genomes indicate that changes in ancestral composition influenced phenotypes among populations in the Tianshan Mountains (TSM). Populations in the western and northern TSM regions exhibited light hair and skin pigmentation from the late Bronze Age to the Iron Age, with blue eyes appearing in the Iron Age due to early Eurasian contact.^{146,148} However, over time, the phenotypes shifted to darker eyes and hair with increasing East Asian ancestral components. GWASs and model animal experiments have shown that genes such as *SLC24A5* and *OCA2* influence skin pigmentation in West Eurasians and East Asians, respectively. The intron variant rs11636073 in *SLC24A1* is linked to pigmentation in Europeans, and natural selection signatures in the *SLC24A5*-*SLC12A1* region among Uyghurs promote light skin color. Additionally, rs12913832 in *HERC2* affects eye pigmentation in Europeans, whereas rs1800414 and rs74653330 in *OCA2* are associated with skin color in East Asians. Three SNPs in the *HERC2*-*OCA2* region also influence eye pigmentation in Uyghurs. In summary, admixture-enabled selection and variations in the *SLC24A5* and *OCA2* genes have contributed to the diversity of skin pigmentation in the Uyghur population.

Significant evidence has revealed the human adaptive allele *EDARV370A* in the *EDAR* gene, which is undergoing robust positive selection in East Asians and Native Americans but is deficient among Africans and Europeans.⁴¹ Mao et al. traced the evolutionary trajectory of *EDARV370A*, revealing that this mutation first appeared in East Asia before the Last Glacial Maximum (LGM) and became widespread after the LGM. This was demonstrated through several large temporal-scale genomes, including 40,000-year-old Tianyuan individuals and 25 ancient genomes from the Amur River Basin. The derived allele is significantly associated with incisor shoveling, sweat gland activity, and hair thickness/straightness, reflecting the pleiotropic nature of *EDARV370A*. These traits are typical features and signals of natural selection

in East Asians at different times in human evolution.⁴¹ Owing to its pleiotropic nature, the selective pressure on 370A remains unclear, and more genetic evidence is needed. In Uyghurs, influenced by gene flow from Europeans, the *EDARV370A* variant plays a vital role in early embryonic development, affecting facial characteristics and shaping the genetic landscape. The frequency of derived alleles aligns with the geographical distribution of admixture proportions among Uyghurs, making *EDAR* an admixture-representative gene in these populations.

Genomic resources and the discovery of archaic and ancient populations

Influence of archaic introgression on genomic medicine. Non-African anatomically modern humans (AMHs) interbred with archaic hominins, such as Neanderthals and Denisovans, on the Eurasian continent, leading to genetic mixing. Among non-Africans, 1–4% of their genomes can be traced back to Neanderthals, and 0–5% can be traced to Denisovans.¹⁴⁹ Admixture modeling suggested that a significant introgression event occurred between the common ancestors of non-Africans and Neanderthals. Subsequent admixture with the basal Eurasian lineage diluted the Neanderthal variants in western Eurasians, leading to a gradient of Neanderthal ancestry from west to east.¹⁵⁰ Denisovan introgression is more complex, involving at least four events between different Denisovan lineages and modern humans. One event affected the common lineages of East Asians and Oceanians, whereas three separate events impacted Oceanians, East Asians, and Southeast Asians.¹⁵¹ The greatest proportions of Denisovan introgression are found in the Philippine Negritos Ayta and Agata people and in the Oceanian Melanesians from Papua New Guinea and Bougainville Island.¹⁵² A small proportion of archaic introgression sequences are present in modern genomes, with some genomic regions completely devoid of archaic variants, indicating that purifying selection was the primary evolutionary force. Statistical algorithms based on shared haplotypes, match rates, and identical-by-descent patterns have identified numerous genes or variants inherited from archaic humans.¹⁵³ GWAS-related annotations have linked these introgressed variants to traits such as skin tone, hair color, height, and various disease phenotypes in modern humans.¹⁵⁰ The high-altitude adaptation of Tibetans, associated with the Denisovan-derived *EPAS1* gene, exemplifies rapid adaptation to extreme environments through archaic genomic material. Additionally, genes inherited from Neanderthals and Denisovans have been linked to functions in innate and adaptive immunity (*TNFAIP3*), metabolism (*FADS1*, *SUMF1*, *PLPP1*, *LIPA*, *CH25H*, *TBX15*, and *WARS2*), neuronal development, and dermatological or pigmentation phenotypes.¹⁴⁹ The epigenetic status of these archaic genes also affects modern phenotypes, as indicated by the methylation status of the Denisovan-related *HOXD9* gene, which influences limb development.

Ancient DNA served as the direct window for a glimpse into human demographic history. Ancient DNA from more than 10,000 AMHs has provided unprecedented insight into the genomic diversity patterns of different populations across time and space.¹⁵⁴ Studies on ancient Eurocentric DNA have shown that early geographically diverse Paleolithic hunter-gatherers mixed with incoming Anatolian Neolithic farmers and Bronze Age steppe pastoralists formed the primary gene pool of modern Europeans.¹⁵⁵ Ancient DNA findings can support previous findings derived from artistic depictions, historical documents, and phenotypic changes in contemporary populations. Lazaridis et al. reported that pre-pottery and pottery Neolithic DNA from Mesopotamia revealed a multisource admixture involving hunter-gatherers from the Levant, the Caucasus, and Anatolia, contributing to the formation of pre-pottery Neolithic Mesopotamians.¹⁵⁶ The patterns observed in the pre-pottery and pottery Anatolian people suggest that at least two pulses of gene flow from the Levant and Mesopotamia, combined with local Epipaleolithic sources, shaped Neolithic Anatolian farmers.¹⁵⁶ This northwestward migration of Anatolian ancestry significantly influenced Neolithic farmers in Europe, whereas the southward migration of Proto-Indo-European-speaking Yamnaya influenced Bronze Age ancestry.¹⁵⁷ Post-Bronze Age populations in southern Europe and western Eurasia underwent complex differentiation and admixture events.¹⁵⁸ Olalde et al. analyzed Balkan Iron Age-related ancestry during the Roman Empire and reconstructed internal migratory patterns, revealing that Slavic gene flow contributed to the Balkan gene pool.¹⁵⁹ Allentoft et al. reconstructed large-scale human migrations and population

turnover among farmers, pastoralists, and local hunter-gatherers in post-glacial western Eurasia, detailing the population history from the Black Sea to the Baltic Sea.¹⁶⁰ In northwestern Europe, including Denmark, population dynamics have experienced multiple demographic events, including turnover and admixture, under varying cultural influences.¹⁵⁵ High-density sampling of different Europeans over time has bridged the genetic gaps between culturally distinct Europeans.

Non-Eurocentric ancient DNA resources from Africa have revealed significant demographic changes involving three highly divergent source populations: the central African rainforest hunter-gatherers and the deeply diverged eastern and southern African lineages.¹⁶¹ Ancient DNA from Oceania has shown the demographic impacts of Proto-Austronesian-related population expansion on the genomic formation of the initial Lapita people and subsequent population turnover or replacement.¹⁶² Recent fine-scale regional studies revealed at least five gene flow events involving three East Asian sources, one Papuan source, and one Polynesian source, highlighting the migration patterns of Micronesian seafarers.¹⁵⁶ Extensive research has elucidated the formation of Native American populations, including divergence, dispersal, and admixture events with regional northern and southern Native American lineages.^{163,164} Recent regional DNA data have provided a fine-scale resolution of population continuity and admixture along the Santa Barbara mainland coast and the northern Channel Islands of California.¹⁶⁵

With the rapid advancement of ancient DNA techniques in East Asia, over 700 Paleolithic and historic genomes have been analyzed to uncover the genetic legacy of ancient populations from the Yellow and Yangtze River Basins and Holocene-to-Historic Trans-Eurasian migrations.⁴¹ The Bronze and Iron Age genomes from the Tarim and Dzungarian Basins revealed that Western Eurasian Steppe Yamnaya-related ancestry and BMAC-related Iran farmer ancestry influenced the admixture patterns in ancient Xinjiang and Hexi Corridor populations.^{39,166,167} Early local lineages around the Tianshan Mountains were genetically linked to the ancient northern Eurasian lineage near Lake Baikal.¹⁶⁶ Ancient Northeast Asian ancestry predominated in the postglacial Amur River region, Russia's Far East DevilsGate, and Mongolian Plateau hunter-gatherers, who maintained long-term genetic stability, contributed to the formation of Native Americans, and shared close genetic ties with early millet farmers.³⁶ Early hunter-gatherers from Shandong Province and Fujian Paleolithic Qihe and Liangdao populations exhibited genetic differentiation from the early Neolithic period, with subsequent directional gene flow from ancient northern and southern East Asians shaping population admixture along with shifts in cultural diffusion and subsistence strategies.⁴⁰ Genetic interactions among millet farmers from different regions indicate that changes in subsistence strategies influence population movements and agricultural patterns.⁴⁵ Ancient highland Tibetan genomes resulted primarily from mixing Neolithic millet farmers with local hunter-gatherers, with distinct substructures emerging among geographically diverse highlanders.¹⁶⁸ Multiple southward migrations of Chinese farmers along coastal and inland routes impacted the gene pool of Southeast Asia's post-Hòabinhian populations, aligning with the spread of proto-languages from South China.⁴⁴ Overall, the complex admixture and migration history of Paleolithic and Neolithic populations support the hypothesis that ancient population expansions disseminated the languages spoken by modern East Asians, which is consistent with farming and language dispersal theory.³⁸

Ancient DNA serves as a powerful tool for understanding human disease. Ancient genomic resources provide a crucial strategy for synthesizing human evolutionary history and the coevolution of hosts and pathogens with personal genomic medicine. This synthesis can promote the integration of evolutionary and precision medicine in clinical decision-making. Evolutionary innovations in biomedicine have created substrates for human disease phenotypes. Generally, constraints, trade-offs, mismatches, and conflicts during demographic evolutionary events can influence human traits and diseases. The evolutionary trade-off theory in age-related diseases sheds light on the rules for new life generation and trends toward death, serving as a paradigm of macroevolutionary imprints. Pathogen exposure can modulate ancient innate and young adaptive immune systems through self-/nonself-recognition, activating responses to pathogens. This hypothesis for human adaptation and disease generation after genetic adaptation to pathogens is summarized as antagonistic pleiotropy.

The deep evolutionary past and recent human evolution have contributed to the emergence of human trait- and disease-related variations. Ancient DNA enables the reconstruction of detailed evolutionary histories, including migration, expansion, admixture, introgression, and population bottlenecks. This approach enhances our understanding of the evolutionary trajectory of the genetic architecture of human diseases, particularly infectious and immune diseases. The shift from Paleolithic subsistence strategies to Neolithic farming practices facilitated increased contact with animals, increased population densities in urban settings, accelerated pathogen transmission due to greater human mobility, and altered human-environment interactions. An evolutionary genomic approach focused on defense genes in post-Neolithic Europeans has shed light on the mechanisms underlying disparities in infectious and inflammatory disease risk in modern Western Eurasians. Kerner et al. reported that directional selection of immune genes in post-Bronze Age populations enhanced defense against infectious diseases through genetic adaptation but increased susceptibility to autoimmune disorders in post-Neolithic Europe.¹⁶⁹ This phenomenon of antagonistic pleiotropy, where genetic adaptations to one aspect may increase the risk for another, also plays a role in the spread of genetic disease risks through population migration. Barrie et al. reported that the elevated risk for multiple sclerosis, originating from Yamnaya pastoralists, was disseminated among modern northern Europeans, with a high proportion of Steppe ancestry.¹⁷⁰ Additionally, pathway analysis suggested greater genetic susceptibility to rheumatoid arthritis in European hunter-gatherer populations. These findings indicate that admixture and selection influence genetic susceptibility to human diseases. Trait-related genetic adaptive signatures in genes such as *LCT*, *FADS*, and *SLC24A5* were identified in ancient Europeans through ancestry-specific statistical strategies, reflecting earlier selection events.¹³⁶ Moreover, immune genes have continually evolved in response to exposure to different pathogens, including the strong selective pressure exerted by the bacterium *Yersinia pestis*.¹⁷¹

Ancient DNA can shed light on the genetic architectures of Mendelian diseases by identifying pathogenic mutations, examining physical features through anthropological analyses and reconstructing complex social organizations. Brace et al. discovered the Ashkenazi-associated genetic basis of hereditary diseases in a community before the 12th century, attributed to a population founder event.¹⁷² Ancient DNA data also trace the genetic origins of complex trait variability in contemporary Europeans. Marnetto et al. identified differentiated ancestral genomic components, illuminating the associations between anthropometric, pigmentation, and metabolic phenotypes and the genotypes of farmers, hunter-gatherers, and pastoralists. Admixture-induced adaptation promoted rapid agricultural adaptations in mixed Neolithic European agricultural groups through local Mesolithic hunter-gatherers and Anatolian farmers. In this context, Neolithic farmer-related ancestry contributed haplotype fragments of the pigmentation-associated gene *SLC24A5*, whereas Mesolithic local hunter-gatherer ancestry contributed to the major histocompatibility complex loci.¹⁷³ There is a long way to go in exploring the genetic origins and spread of common and rare diseases in ancient Chinese genomes. To date, these sources have focused on highland adaptive signatures and other anthropometric-related genetic adaptations.

High-quality genomic resources and genetic discovery with new sequencing technique innovations

Genomic resources of structural variations. SVs are large rearranged genomic segments with significant impacts on human disease, evolution, and genomic mutational processes. The variation spectrum includes small InDels, variable numbers of repetitive sequences (short tandem repeats, STRs, and variable number tandem repeats, VNTRs), and various genomic structural alterations, such as insertions (INs), deletions (DELs), duplications (DUPs), inversions (INVs), copy number variations (CNVs), translocations, and mobile elements. These have garnered significant attention in human genetics, particularly with large-scale national biobanks. Comprehensive characterizations of human SVs in the 1KGP and HGDP have provided critical atlases and catalogs for ethnolinguistically diverse populations worldwide.^{4,174} In the 1KGP, 3,202 samples revealed 173,366 loci, including 90,259 DELs, 28,242 DUPs, 673 CNVs, 49,693 INs, 920 INVs, 3,568 complex SVs, and 11 interchromosomal translocations. HGDP identified 126,018 variants,

many of which were population-specific, population-stratified, or archaic introgressed. The gnomAD project has established the largest SV reference database for medical and population genetic research, reporting 433,371 SVs and identifying differential selection effects on coding and noncoding SVs.¹⁷⁵ These resources enhance our understanding of clinically significant incidental findings of SVs.

High-quality genomes in the LRS era. The Human Genome Structural Variation Consortium has significantly advanced SV discovery through various genomic technologies, including G-banding cytogenetics, chromosomal microarrays, short-read sequencing (SRS), LRS, phased genome assembly strategies, and artificial intelligence for genomes.^{104,151,176} SRS is primarily used for genotyping targeted SV sets, whereas LRS focuses on assembly-based SV calling.¹⁵¹ A comprehensive catalog of 99,604 common human SVs from 15 individuals worldwide was created via LRS, enabling highly accurate genotyping of known SVs from short reads via advanced algorithms and effectively patching GRCh38 as an alternative allele.¹⁷⁷ A fully haplotype-resolved genome assembly of 32 genetically representative individuals, utilizing PacBio single-molecule real-time (SMRT) sequencing and single-cell template strand sequencing, produced a graph-based representation of 107,590 SVs. This resource aids in reconstructing human ancestry and exploring population-structured and adaptive SVs, such as the 4.0-kbp deletion of the *LCT* gene in Europeans, while also facilitating reliable graph-based genotyping of 50,340 SVs from short reads.¹⁷⁸ The pangenomic tool Giraffe enables the genotyping of 167,000 newly discovered SVs through a pangenomic approach to short-read mapping, allowing researchers to investigate the characteristics, mutations, and structures of SVs and their correlations with gene expression, splicing, and disease in large-scale SV resources.¹⁷⁹ In Icelanders, 133,886 reliably genotyped SV alleles were identified based on LRS, revealing associations between a rare deletion in *PCSK9* and LDL cholesterol levels, as well as a multiallelic SV in *ACAN* linked to height. These findings demonstrate the power of combining advanced sequencing technologies and computational methods to enhance our understanding of SVs and their impact on human health and disease.¹⁸⁰

Population-specific SV resources reveal unique genomic diversity. Recent studies on population-scale whole-genome LRSs have revealed numerous previously unreported variations. For instance, in Indigenous Australians, 62% of 136,797 large InDels, 159,912 SVs, and 156 CNVs were newly identified.⁵³ Comprehensive characterization of STR sequences further revealed unique diversity patterns and gene expansion trends. Despite limited LRS-based research in East Asians, recent Chinese SV resources have provided significant insights into adaptation, phenotypes, and diseases. LRS-based SV data from 405 Han Chinese individuals were analyzed to explore associations between SVs and 68 phenotypic and clinical measurements. This analysis revealed 132,312 nonredundant SV annotations and 1,929 LOF variants affecting the coding sequences of 1,681 genes. Geographic differences were noted, with north-south differentiated SVs linked to immune functions. Additionally, rare deletions in the *HBA1/HBA2/HBB* genes were correlated with anemia susceptibility. A comparative study of 15 highland Tibetan and 10 lowland Han Chinese individuals identified 38,216 unique SVs, 27% of which were novel. Further pangenomic genotyping of 276 additional genomes revealed several adaptive introgressed SV candidates, 69 Tibetan-Han stratified SVs, and 80 candidate adaptive genes, including *TMEM247*, *OR10Z1*, and *MRTFA*.¹³⁴ An expanded LRS-based study confirmed that Tibetan-specific SVs are involved in regulating pulmonary function, the hypoxia response, and energy metabolism. Notably, a Tibetan-specific deletion in *EPAS1* downregulated its expression by disrupting a super-enhancer.¹⁸¹

ONT-based LRS projects exhibit relatively high error rates, and GRCh38-based mapping and genotyping have limitations in identifying SVs in complex genomic regions. Recently, multiple LRS and scaffolding sequencing techniques have been combined to reconstruct population-specific haplotype-resolved reference sequences for Chinese populations and medical research, including pilot studies on Tibetan, Tujia, and Han Chinese populations.¹⁸²⁻¹⁸⁴ The high-resolution Tibetan reference genome of ZF1 revealed that archaic hominids shared non-reference sequences (1.32%-1.53%) contributing to better lung function, such as the 662-bp insertion in *SCUBE2* influencing the FEV1/FVC ratio and the Tibetan-Han-stratified 163-bp intronic deletion in the *MKL1* gene controlling lower systolic pulmonary arterial pressure.¹⁸² Haplo-

type-resolved de novo assembly of TJ1 from lowland East Asian regions further identified Tujia-specific SVs and population-stratified SVs of *LCT* and *UBXN8*.¹⁸³ Despite advancements in identifying SVs, some complex regions, such as telomeres and centromeres, remain elusive. The telomere-to-telomere (T2T) references of T2T-CHM13, CN1, and T2T-YAO provided complete human reference sequences, offering a population-specific reference for genomics to avoid reference bias that overlooks critical genetic discoveries.^{185,186} Recent graph-based human pangenomic projects from the Human Pangenome Reference Consortium (HPRC) and the Chinese Pangenome Consortium (CPC) established an ideal research paradigm for fully characterizing and understanding the complete catalog of human SVs and their biological functions.^{91,187}

PERSPECTIVE

We provide a comprehensive overview of Chinese genomic initiatives addressing the underrepresentation of Chinese populations in global genomic research and the prevailing Han bias in East Asian regional projects. Our study expands on recent advancements in genetic diversity characterization, population history reconstruction, medical relevance interpretation, haplotype reference panel construction, and genotype-phenotype association analyses to elucidate the genetic architecture of human diseases. Additionally, we examined complex human evolutionary events, including migration, admixture, introgression, and adaptation, and their impacts on genetic diversity, disease susceptibility, and health outcomes in Chinese populations. We further highlighted technological innovations in SV characterization and their contributions to resolving missing heritability in human diseases. These advances underscore the progress of Chinese genomic research, strengthening its methodological foundations, expanding its analytical capabilities, and enhancing its global recognition.¹⁸⁸ Emerging discoveries provide deeper insights into human evolutionary history and its implications for disease.

Biobank-scale genomic datasets offer a transformative opportunity to resolve undiagnosed cases by refining the genetic architecture of diseases in the precision medicine era. Advances in Chinese precision medicine strategies include disease classification, risk prediction through the adaptation of European-based polygenic risk score models adjusted for Chinese-specific genetic architectures, pharmacogenomics-driven personalized drug responses informed by a fine-scale pharmacogenomic variant spectrum, and the integration of gene-environment interactions to refine disease interventions. AI-driven frameworks further enable ancestry-informed disease prediction and diagnosis.¹⁷⁶ Large-scale genomic datasets from biobank initiatives establish a foundation for precision medicine in Chinese populations by improving risk prediction, optimizing drug therapies, elucidating disease mechanisms, and facilitating ancestry-informed interventions. Expanding these efforts is critical for translating genomic research into clinical applications, ultimately enhancing healthcare outcomes and reducing disease burdens in China. However, overcoming the continued underrepresentation of Chinese genomes in human genomic research requires large-scale, national initiatives akin to All of Us and the UK Biobank. Future advancements in sequencing technologies, computational innovations, and AI-driven precision medicine will be essential for establishing comprehensive population-based genomic projects and fully integrating genomic insights into clinical practice.

Broaden genomic diversity from ethno-linguistically diverse and under-represented populations

Genomic studies have focused predominantly on populations of European ancestry, leading to an underrepresentation of ethno-linguistically diverse East Asian populations. This disparity hampers the global understanding of genetic diversity and limits the identification of genetic variants linked to specific diseases and traits. The scarcity of genomic projects representing diverse populations obstructs the comprehensive exploration of population genetic structures, demographic histories, and biological adaptation signals within Chinese populations. Therefore, expanding the genetic diversity characterization of Chinese populations is crucial. This effort will enhance the understanding of human ancestry, improve health equity, enrich genomic resources, and facilitate the application of genetic findings. Genomic cohort studies must also prioritize respecting individuals and communities. Ethical, legal, and social implications (ELSI) teams from global and regional projects

should offer professional guidance for modern and ancient human genome research.¹⁸⁹

Data availability and sharing principles

Another concern is how to share data while protecting individuals' privacy. Advances in sequencing technologies have made high-depth WGS practical for identifying diverse variants and providing comprehensive genetic information. Large-scale reference panels with high imputation accuracy are crucial for disease-associated genetic studies, as they capture most trait-associated variants and enable precision medicine and personalized treatment. GWASs reveal genes or variants linked to disease risk and complex traits, forming the foundation for precision medicine, personalized treatment, and early disease diagnosis. Significant progress in data sharing has been made through initiatives such as the UK Biobank and All of Us programs. It is recommended that raw data from other projects be submitted to public databases, such as the Genome Sequence Archive and the European Nucleotide Archive. In addition, the establishment of a genome alliance can strengthen project cooperation and communication and increase the reusability of data. Implementing FAIR (findable, accessible, interoperable, reusable) principles, as announced by the HPRC, is essential in all human genomic research.¹⁹⁰⁻¹⁹²

Integrative evolutionary and functional genomics

The reconstruction of evolutionary history can illuminate how evolutionary forces influence the occurrence of human disease-related variations. The comprehensive characterization of genetic variant origins, patterns, and biological consequences benefits from integrating genomics with transcriptomic, metabolomic, proteomic, epigenetic, and microbiome data, such as the comprehensive identification and characterization of the highland adaptations of *MCUR1*.¹⁹³ Multi-omics data analysis is essential for understanding the functions of new or introgressed variants in biological adaptation or malfunctioning traits. Published Chinese genomic projects have overlooked mainly the combined exploration of genomics and phenomics for personalized population health. The Chinese integrative genomics approach, which incorporates genomic and other multiomics data from ethnically diverse Chinese populations, represents a new era of genomic discovery, addressing fundamental human genomic questions.

New sequence technology innovations

Innovations in genome sequencing, from the initial to the current long-read (third or fourth) advances, have profoundly transformed genomic research for humans, animals, and plants. ONT-based population genomic cohorts have revealed insights into human SVs and their roles in disease and other traits. However, there is still a scarcity of high-quality, haplotype-resolved genomic databases for global populations. The ONT-based Han Chinese SV landscape helps open new windows for exploration of the genetic architectures contributed via SV, such as the phenotype diversity and disease susceptibility caused via the specific SV in *GSDMD* and *WWP2*.¹⁹⁴ Recently, the CPC and the HPRC have published comprehensive human pangenome references, capturing the genetic diversity of various ethno-linguistically diverse continental and Chinese populations.^{91,187} These population-scale pangenome resources provide new opportunities to explore previously obscure regions of human genomes, shedding light on human origins and the genetic underpinnings of traits and diseases.

CONCLUSION

Genome research focused on underrepresented Chinese populations has yielded significant insights into fundamental biological principles and the mechanisms of human disease. These studies have provided new perspectives on population genetics, medical genomics, and evolutionary genomics, illuminating essential interactions between genes, the environment, and genetic etiology. The development of large-scale, multiethnic genomic projects has led to promising breakthroughs in understanding the human genome. These advancements will enhance disease diagnosis, risk prediction, personalized treatment, and biotechnological innovations. Further remarkable research results are expected to deepen our knowledge and lead to continued progress in these areas.

REFERENCES

- Lander E. S., Linton L. M., Birren B., et al. (2001). Initial sequencing and analysis of the human genome. *Nature* **409**:860–921. DOI:10.1038/35057062
- Venter J. C., Adams M. D., Myers E. W., et al. (2001). The sequence of the human genome. *Science* **291**:1304–1351. DOI:10.1126/science.1058040
- International HapMap C. (2003). The International HapMap Project. *Nature* **426**:789–796. DOI:10.1038/nature02168
- Byrsk-Bishop M., Evani U. S., Zhao X., et al. (2022). High-coverage whole-genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios. *Cell* **185**:3426–3440 e3419. DOI:10.1016/j.cell.2022.08.004.
- Genomes Project C., Auton A., Brooks L. D., et al. (2015). A global reference for human genetic variation. *Nature* **526**:68–74. DOI:10.1038/nature15393
- Mallick S., Li H., Lipson M., et al. (2016). The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. *Nature* **538**:201–206. DOI:10.1038/nature18964
- Wang T., Antonacci-Fulton L., Howe K., et al. (2022). The Human Pangenome Project: a global resource to map genomic diversity. *Nature* **604**:437–446. DOI:10.1038/s41586-022-04601-8
- Hallardsson B. V., Eggertsson H. P., Moore K. H. S., et al. (2022). The sequences of 150,119 genomes in the UK Biobank. *Nature* **607**:732–740. DOI:10.1038/s41586-022-04965-x
- Turnbull C., Scott R. H., Thomas E., et al. (2018). The 100 000 Genomes Project: bringing whole genome sequencing to the NHS. *BMJ (Clinical research ed.)* **361**:k1687. DOI:10.1136/bmj.k1687
- Taliun D., Harris D. N., Kessler M. D., et al. (2021). Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *Nature* **590**:290–299. DOI:10.1038/s41586-021-03205-y
- All of Us Research Program Genomics I. (2024). Genomic data in the All of Us Research Program. *Nature* **627**:340–346. DOI:10.1038/s41586-023-06957-x
- Chen S., Francioli L. C., Goodrich J. K., et al. (2024). A genomic mutational constraint map using variation in 76,156 human genomes. *Nature* **625**:92–100. DOI:10.1038/s41586-023-06045-0
- Wonkam A. (2021). Sequence three million genomes across Africa. *Nature* **590**:209–211. DOI:10.1038/d41586-021-00313-7
- Choudhury A., Aron S., Botigue L. R., et al. (2020). High-depth African genomes inform human migration and health. *Nature* **586**:741–748. DOI:10.1038/s41586-020-2859-7
- Fatumo S., Yakubu A., Oyedele O., et al. (2022). Promoting the genomic revolution in Africa through the Nigerian 100K Genome Project. *Nat. Genet.* **54**:531–536. DOI:10.1038/s41586-022-01071-6
- Fatumo S., Mugisha J., Soremekun O. S., et al. (2022). Uganda Genome Resource: A rich research database for genomic studies of communicable and non-communicable diseases in Africa. *Cell Genom.* **2**:100209. DOI:10.1016/j.xgen.2022.100209.
- GenomeAsia K. C. (2019). The GenomeAsia 100K Project enables genetic discoveries across Asia. *Nature* **576**:106–111. DOI:10.1038/s41586-019-1793-z
- Stark Z., Boughtwood T., Phillips P., et al. (2019). Australian genomics: A federated model for integrating genomics into healthcare. *Am. J. Hum. Genet.* **105**:7–14. DOI:10.1016/j.ajhg.2019.06.003
- Ziyatdinov A., Torres J., Alegre-Diaz J., et al. (2023). Genotyping, sequencing and analysis of 140,000 adults from Mexico City. *Nature* **622**:784–793. DOI:10.1038/s41586-023-06595-3
- Sohail M., Palma-Martinez M. J., Chong A. Y., et al. (2023). Mexican Biobank advances population and medical genomics of diverse ancestries. *Nature* **622**:775–783. DOI:10.1038/s41586-023-06560-0
- Leitsalu L. and Metspalu A. (2017). From biobanking to precision medicine: The Estonian experience. In *Genomic and precision medicine* (Elsevier), pp: 119–129.
- Metspalu A., Kohler F., Laschinski G., et al. (2004). [The Estonian Genome Project in the context of European genome research]. *Dtsch Med Wochenschr* **129 Suppl** **1**:S25–28. DOI:10.1055/s-2004-824840
- Leitsalu L., Haller T., Esko T., et al. (2015). Cohort Profile: Estonian Biobank of the Estonian Genome Center, University of Tartu. *Int J Epidemiol* **44**:1137–1147. DOI:10.1093/ije/dyt268
- Gudbjartsson D. F., Helgason H., Gudjonsson S. A., et al. (2015). Large-scale whole-genome sequencing of the Icelandic population. *Nat. Genet.* **47**:435–444. DOI:10.1038/ng.3247
- Kurki M. I., Karjalainen J., Palta P., et al. (2023). FinnGen provides genetic insights from a well-phenotyped isolated population. *Nature* **613**:508–518. DOI:10.1038/s41586-022-05473-8
- Brody J. A., Morrison A. C., Bis J. C., et al. (2017). Analysis commons, a team approach to discovery in a big-data environment for genetic epidemiology. *Nat. Genet.* **49**:1560–1563. DOI:10.1038/ng.3968
- Nagasaki M., Yasuda J., Katsuoka F., et al. (2015). Rare variant discovery by deep whole-genome sequencing of 1,070 Japanese individuals. *Nat. Commun.* **6**:8018. DOI:10.1038/ncomms9018
- Jeon S., Bhak Y., Choi Y., et al. (2020). Korean Genome Project: 1094 Korean personal genomes with clinical information. *Sci. Adv.* **6**:eaaz7835. DOI:10.1126/sciadv.aaz7835
- Wu D., Dou J., Chai X., et al. (2019). Large-Scale Whole-Genome Sequencing of Three Diverse Asian Populations in Singapore. *Cell* **179**:736–749 e715. DOI:10.1016/j.cell.2019.09.019.
- Wong E., Bertin N., Hebrard M., et al. (2023). The Singapore National Precision Medicine Strategy. *Nat. Genet.* **55**:178–186. DOI:10.1038/s41588-022-01274-x
- Cao Y., Li L., Xu M., et al. (2020). The ChinaMAP analytics of deep whole genome sequences in 10,588 individuals. *Cell Res.* **30**:717–731. DOI:10.1038/s41422-020-0322-9
- Chiang C. W. K., Mangul S., Robles C., et al. (2018). A comprehensive map of genetic variation in the world's largest ethnic group-Han Chinese. *Mol. Biol. Evol.* **35**:2736–2750. DOI:10.1093/molbev/msy170
- Jiang T., Guo H., Liu Y., et al. (2024). A comprehensive genetic variant reference for the Chinese population. *Sci. Bull. (Beijing)* **69**:3820–3825. DOI:10.1016/j.scib.2024.06.017
- Sirugo G., Williams S. M. and Tishkoff S. A. (2019). The missing diversity in human genetic studies. *Cell* **177**:26–31. DOI:10.1016/j.cell.2019.02.048
- Popejoy A. B. and Fullerton S. M. (2016). Genomics is failing on diversity. *Nature* **538**:161–164. DOI:10.1038/538161a
- Mao X., Zhang H., Qiao S., et al. (2021). The deep population history of northern East Asia from the Late Pleistocene to the Holocene. *Cell* **184**:3256–3266 e3213. DOI:10.1016/j.cell.2021.04.040.
- Wang T., Wang W., Xie G., et al. (2021). Human population history at the crossroads of East and Southeast Asia since 11,000 years ago. *Cell* **184**:3829–3841.e3821. DOI:10.1016/j.cell.2021.05.018.
- Wang C. C., Yeh H. Y., Popov A. N., et al. (2021). Genomic insights into the formation of human populations in East Asia. *Nature* **591**:413–419. DOI:10.1038/s41586-021-03336-2
- Kumar V., Wang W., Zhang J., et al. (2022). Bronze and Iron Age population movements underlie Xinjiang population history. *Science* **376**:62–69. DOI:10.1126/science.abk1534
- Yang M. A., Fan X., Sun B., et al. (2020). Ancient DNA indicates human population shifts and admixture in northern and southern China. *Science* **369**:282–288. DOI:10.1126/science.aba0909
- Liu Y., Mao X., Krause J., et al. (2021). Insights into human history from the first decade of ancient human genomics. *Science* **373**:1479–1484. DOI:10.1126/science.abi8202
- He G., Wang M., Luo L., et al. (2024). Population genomics of Central Asian peoples unveil ancient Trans-Eurasian genetic admixture and cultural exchanges. *hLife* **2**:554–562. DOI:10.1016/j.hlife.2024.06.006
- Lipson M., Cheronet O., Mallick S., et al. (2018). Ancient genomes document multiple waves of migration in Southeast Asian prehistory. *Science* **361**:92–95. DOI:10.1126/science.aat3188
- McColl H., Racimo F., Vinner L., et al. (2018). The prehistoric peopling of Southeast Asia. *Science* **361**:88–92. DOI:10.1126/science.aat3628
- Ning C., Li T., Wang K., et al. (2020). Ancient genomes from northern China suggest links between subsistence changes and human migration. *Nat. Commun.* **11**:2700. DOI:10.1038/s41467-020-16557-2
- Lim E. T., Wurtz P., Havulinna A. S., et al. (2014). Distribution and medical impact of loss-of-function variants in the Finnish founder population. *PLoS Genet.* **10**:e1004494. DOI:10.1371/journal.pgen.1004494
- Xue Y., Mezzavilla M., Haber M., et al. (2017). Enrichment of low-frequency functional variants revealed by whole-genome sequencing of multiple isolated European populations. *Nat. Commun.* **8**:15927. DOI:10.1038/ncomms15927
- Zuk O., Schaffner S. F., Samocha K., et al. (2014). Searching for missing heritability: designing rare variant association studies. *Proc. Natl. Acad. Sci. USA* **111**:E455–464. DOI:10.1073/pnas.1322563111
- Genovesse G., Tonna S. J., Knob A. U., et al. (2010). A risk allele for focal segmental glomerulosclerosis in African Americans is located within a region containing APOL1 and MYH9. *Kidney Int.* **78**:698–704. DOI:10.1038/ki.2010.251
- Rotimi C. N., Bentley A. R., Doumatey A. P., et al. (2017). The genomic landscape of African populations in health and disease. *Hum. Mol. Genet.* **26**:R225–R236. DOI:10.1093/hmg/ddx253
- Cohen J., Pertsemidis A., Kotowski I. K., et al. (2005). Low LDL cholesterol in individuals of African descent resulting from frequent nonsense mutations in PCSK9. *Nat. Genet.* **37**:161–165. DOI:10.1038/ng1509
- Huang S., Liu S., Huang M., et al. (2024). The Born in Guangzhou Cohort Study enables generational genetic discoveries. *Nature* **626**:565–573. DOI:10.1038/s41586-023-06988-4
- Reis A. L. M., Rapadas M., Hammond J. M., et al. (2023). The landscape of genomic structural variation in Indigenous Australians. *Nature* **624**:602–610. DOI:10.1038/s41586-023-06842-7
- He G., Yao H., Duan S., et al. (2025). Pilot work of the 10K Chinese People Genomic Diversity Project along the Silk Road suggests a complex east-west admixture landscape and biological adaptations. *Sci. China Life Sci.* DOI:10.1007/s11427-024-2748-4.
- Cong P. K., Bai W. Y., Li J. C., et al. (2022). Genomic analyses of 10,376 individuals in the Westlake BioBank for Chinese (WBBC) pilot project. *Nat. Commun.* **13**:2939. DOI:10.1038/s41467-022-30526-x

56. Wang C., Dai J., Qin N., et al. (2022). Analyses of rare predisposing variants of lung cancer in 6,004 whole genomes in Chinese. *Cancer Cell* **40**:1223-1239.e1226. DOI:10.1016/j.ccell.2022.08.013.
57. Tadaka S., Katsuoka F., Ueki M., et al. (2019). 3.5KJPNv2: an allele frequency panel of 3552 Japanese individuals including the X chromosome. *Hum. Genome Var.* **6**:28. DOI:10.1038/s41439-019-0059-5.
58. Liu X., Koyama S., Tomizuka K., et al. (2024). Decoding triancestral origins, archaic introgression, and natural selection in the Japanese population by whole-genome sequencing. *Sci. Adv.* **10**:ead18419. DOI:10.1126/sciadv.adi8419
59. Cavalli-Sforza L. L. (1998). The Chinese human genome diversity project. *Proc. Natl. Acad. Sci. USA* **95**:11501-11503. DOI:10.1073/pnas.95.20.11501
60. Chu J. Y., Huang W., Kuang S. Q., et al. (1998). Genetic relationship of populations in China. *Proc. Natl. Acad. Sci. USA* **95**:11763-11768. DOI:10.1073/pnas.95.20.11763
61. Ke Y., Su B., Song X., et al. (2001). African origin of modern humans in East Asia: a tale of 12,000 Y chromosomes. *Science* **292**:1151-1153. DOI:10.1126/science.1060011
62. Wen B., Li H., Lu D., et al. (2004). Genetic evidence supports demic diffusion of Han culture. *Nature* **431**:302-305. DOI:10.1038/nature02878
63. Wang M., Sun Q., Feng Y., et al. (2024). Paleolithic divergence and multiple Neolithic expansions of ancestral nomadic emperor-related paternal lineages. *J. Genet. Genomics*. DOI:10.1016/j.jgg.2024.11.012.
64. Wang M., Liu Y., Luo L., et al. (2024). Genomic insights into Neolithic founding paternal lineages around the Qinghai-Xizang Plateau using integrated YanHuang resource. *iScience* **27**:111456. DOI:10.1016/j.isci.2024.111456
65. Wang M., Huang Y., Liu K., et al. (2024). Multiple human population movements and cultural dispersal events shaped the landscape of Chinese paternal heritage. *Mol. Biol. Evol.* **41**. DOI:10.1093/molbev/msae122.
66. Wang M., Chen H., Luo L., et al. (2024). Forensic investigative genetic genealogy: expanding pedigree tracing and genetic inquiry in the genomic era. *J. Genet. Genomics*. DOI:10.1016/j.jgg.2024.06.016.
67. Sun Y., Wang M., Sun Q., et al. (2024). Distinguished biological adaptation architecture aggravated population differentiation of Tibeto-Burman-speaking people. *J. Genet. Genomics* **51**:517-530. DOI:10.1016/j.jgg.2023.10.002
68. Consortium H. P.-A. S., Abdulla M. A., Ahmed I., et al. (2009). Mapping human genetic diversity in Asia. *Science* **326**:1541-1545. DOI:10.1126/science.1177074
69. Luo L., Wang M., Liu Y., et al. (2025). Sequencing and characterizing human mitochondrial genomes in the biobank-based genomic research paradigm. *Sci. China Life Sci.* DOI:10.1007/s11427-024-2736-7.
70. Chen J., Zheng H., Bei J. X., et al. (2009). Genetic structure of the Han Chinese population revealed by genome-wide SNP variation. *Am. J. Hum. Genet.* **85**:775-785. DOI:10.1016/j.ajhg.2009.10.016
71. Xu S., Yin X., Li S., et al. (2009). Genomic dissection of population substructure of Han Chinese and its implication in association studies. *Am. J. Hum. Genet.* **85**:762-774. DOI:10.1016/j.ajhg.2009.10.015
72. Cai N., Bigdeli T. B., Kretschmar W. W., et al. (2017). 11,670 whole-genome sequences representative of the Han Chinese population from the CONVERGE project. *Sci. Data* **4**:170011. DOI:10.1038/sdata.2017.11
73. Gao Y., Zhang C., Yuan L., et al. (2020). PGG Han: The Han Chinese genome database and analysis platform. *Nucleic Acids Res.* **48**:D971-D976. DOI:10.1093/nar/gkz829
74. Liu S., Huang S., Chen F., et al. (2018). Genomic analyses from non-invasive prenatal testing reveal genetic associations, patterns of viral infections, and Chinese population history. *Cell* **175**:347-359.e314. DOI:10.1016/j.cell.2018.08.016.
75. Li Z., Jiang X., Fang M., et al. (2023). CMDb: the comprehensive population genome variation database of China. *Nucleic Acids Res.* **51**:D890-D895. DOI:10.1093/nar/gkac638
76. Liu S., Zhong M., Huang Y., et al. (2023). Quantitative thresholds for variant enrichment in 13,845 cases: Improving pathogenicity classification in genetic hearing loss. *Genome Med.* **15**:116. DOI:10.1186/s13073-023-01271-7
77. Zhang P., Luo H., Li Y., et al. (2021). NyuWa Genome resource: A deep whole-genome sequencing-based variation profile and reference panel for the Chinese population. *Cell Rep.* **37**:110017. DOI:10.1016/j.celrep.2021.110017
78. Cheng S., Xu Z., Bian S., et al. (2023). The STROMICS genome study: deep whole-genome sequencing and analysis of 10K Chinese patients with ischemic stroke reveal complex genetic and phenotypic interplay. *Cell Discov.* **9**:75. DOI:10.1038/s41421-023-00582-8
79. Yu C., Lan X., Tao Y., et al. (2023). A high-resolution haplotype-resolved Reference panel constructed from the China Kadoorie Biobank Study. *Nucleic Acids Res.* **51**:11770-11782. DOI:10.1093/nar/gkad779
80. Walters R. G., Millwood I. Y., Lin K., et al. (2023). Genotyping and population characteristics of the China Kadoorie Biobank. *Cell Genom.* **3**:100361. DOI:10.1016/j.xgen.2023.100361
81. Yang Q., Sun Y., Duan S., et al. (2025). High-quality Population-specific Haplotype-resolved Reference Panel in the Genomic and Pangenomic Eras. *Genomics, Proteomics Bioinf.* Online ahead of print. DOI:10.1093/gpbjnl/qzaf022
82. Cheng S., Xu Z., Liu Y., et al. (2021). Whole genome sequencing of 10K patients with acute ischaemic stroke or transient ischaemic attack: design, methods and baseline patient characteristics. *Stroke Vasc. Neurol.* **6**:291-297. DOI:10.1136/svn-2020-000664
83. Feng Y. A., Chen C. Y., Chen T. T., et al. (2022). Taiwan Biobank: A rich biomedical research database of the Taiwanese population. *Cell Genom.* **2**:100197. DOI:10.1016/j.xgen.2022.100197
84. Du Z. L., Ma L., Qu H. Z., et al. (2019). Whole Genome Analyses of Chinese Population and Assembly of A Northern Han Genome. *Genom. Proteom. Bioinf.* **17**:229-247. DOI:10.1016/j.gpb.2019.07.002
85. Hao M., Pu W., Li Y., et al. (2021). The HuaBiao project: whole-exome sequencing of 5000 Han Chinese individuals. *J. Genet. Genomics.* **48**:1032-1035. DOI:10.1016/j.jgg.2021.07.013
86. Peng M. S., Quang H. H., Dang K. P., et al. (2010). Tracing the Austronesian footprint in Mainland Southeast Asia: a perspective from mitochondrial DNA. *Mol. Biol. Evol.* **27**:2417-2430. DOI:10.1093/molbev/msq131
87. Pan Y., Zhang C., Lu Y., et al. (2022). Genomic diversity and post-admixture adaptation in the Uyghurs. *Natl. Sci. Rev.* **9**:nwab124. DOI:10.1093/nsr/nwab124
88. Yang Z., Bai C., Pu Y., et al. (2022). Genetic adaptation of skin pigmentation in highland Tibetans. *Proc. Natl. Acad. Sci. USA* **119**:e2200421119. DOI:10.1073/pnas.2200421119
89. He G., Wang P., Chen J., et al. (2024). Differentiated genomic footprints suggest isolation and long-distance migration of Hmong-Mien populations. *BMC Biol.* **22**:18. DOI:10.1186/s12915-024-01828-x
90. Zheng W., He Y., Guo Y., et al. (2023). Large-scale genome sequencing redefines the genetic footprints of high-altitude adaptation in Tibetans. *Genome Biol.* **24**:73. DOI:10.1186/s13059-023-02912-1
91. 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
92. He G., Yao H., Sun Q., et al. (2023). Whole-genome sequencing of ethnolinguistic diverse northwestern Chinese Hexi Corridor people from the 10K_CPGDP project suggested the differentiated East-West genetic admixture along the Silk Road and their biological adaptations. *bioRxiv*:2023.2002.2026.530053. DOI:10.1101/2023.02.26.530053
93. Wang M., Huang Y., Liu K., et al. (2024). Multiple Human Population Movements and Cultural Dispersal Events Shaped the Landscape of Chinese Paternal Heritage. *Mol. Biol. Evol.* **41**:2023.2008.2028.555114. DOI:10.1093/molbev/msae122.
94. Wang Z., Wang M., Liu K., et al. (2023). Paternal genomic resources from the YanHuang cohort suggested a Weakly-Differentiated Multi-source Admixture model for the formation of Han's founding ancestral lineages. *bioRxiv*:2023.2011.2008.566335. DOI:10.1101/2023.11.08.566335
95. He G. L., Wang M. G., Zou X., et al. (2023). Extensive ethnolinguistic diversity at the crossroads of North China and South Siberia reflects multiple sources of genetic diversity. *J. Syst. Evol.* **61**:230-250. DOI:10.1111/jse.12827
96. Li X., Wang M., Su H., et al. (2024). Evolutionary history and biological adaptation of Han Chinese people on the Mongolian Plateau. *hLife* **2**:296-313. DOI:10.1016/j.hlife.2024.04.005
97. Sun Y., Wang M., Sun Q., et al. (2023). Distinguished biological adaptation architecture aggravated population differentiation of Tibeto-Burman-speaking people. *J. Genet. Genomics.* **51**:517-530DOI:10.1016/j.jgg.2023.10.002
98. Lu D., Lou H., Yuan K., et al. (2016). Ancestral origins and genetic history of Tibetan Highlanders. *Am. J. Hum. Genet.* **99**:580-594. DOI:10.1016/j.ajhg.2016.07.002
99. Zhang C., Lu Y., Feng Q., et al. (2017). Differentiated demographic histories and local adaptations between Sherpas and Tibetans. *Genome Biol.* **18**:115. DOI:10.1186/s13059-017-1242-y
100. Chen H., Lin R., Lu Y., et al. (2022). Tracing Bai-Yue ancestry in Aboriginal Li people on Hainan Island. *Mol. Biol. Evol.* **39**:msac210. DOI:10.1093/molbev/msac210
101. Yang C., Zhang X., Yan S., et al. (2024). Large-scale lexical and genetic alignment supports a hybrid model of Han Chinese demic and cultural diffusions. *Nat. Hum. Behav.* **8**:1163-1176. DOI:10.1038/s41562-024-01886-9
102. Loh P. R., Danecek P., Palamara P. F., et al. (2016). Reference-based phasing using the Haplotype Reference Consortium panel. *Nat. Genet.* **48**:1443-1448. DOI:10.1038/ng.3679
103. consortium C. (2015). Sparse whole-genome sequencing identifies two loci for major depressive disorder. *Nature* **523**:588-591. DOI:10.1038/nature14659
104. Yang Y., Xu S., Hong Y., et al. (2024). Computational modeling for medical data: From data collection to knowledge discovery. *The Innovation Life* **2**:100079. DOI:10.59717/j.xinn-life.2024.100079
105. Rubinacci S., Hofmeister R. J., Sousa da Mota B., et al. (2023). Imputation of low-coverage sequencing data from 150,119 UK Biobank genomes. *Nat. Genet.* **55**:1088-1090. DOI:10.1038/s41588-023-01438-3
106. Bai H., Guo X., Narisu N., et al. (2018). Whole-genome sequencing of 175 Mongolians uncovers population-specific genetic architecture and gene flow throughout North and East Asia. *Nat. Genet.* **50**:1696-1704. DOI:10.1038/s41588-018-0250-5
107. Gao M., Lv J., Yu C., et al. (2020). Metabolically healthy obesity, transition to unhealthy metabolic status, and vascular disease in Chinese adults: A cohort study. *PLoS Med.* **17**:e1003351. DOI:10.1371/journal.pmed.1003351
108. Martin A. R., Kanai M., Kamatani Y., et al. (2019). Clinical use of current polygenic risk scores may exacerbate health disparities. *Nat. Genet.* **51**:584-591. DOI:10.1038/s41588-019-0379-x

109. Biffen R. H. (1905). Mendel's laws of inheritance and wheat breeding. *J. Agric. Sci.* **1**:4–48. DOI:10.1017/S0021859600000137
110. Weldon W. F. R. (1902). Mendel's laws of alternative inheritance in peas. *Biometrika* **1**:228–233.
111. Stewart C. and Pepper M. S. (2017). Cystic Fibrosis in the African Diaspora. *Ann. Am. Thorac. Soc.* **14**:1–7. DOI:10.1513/AnnalsATS.201606-481FR
112. Padoa C., Goldman A., Jenkins T., et al. (1999). Cystic fibrosis carrier frequencies in populations of African origin. *J. Med. Genet.* **36**:41. DOI:10.1136/jmg.36.1.41
113. Richards S., Aziz N., Bale S., et al. (2015). Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet. Med.* **17**:405–424. DOI:10.1038/gim.2015.30
114. Hill A. V. (2006). Aspects of genetic susceptibility to human infectious diseases. *Annu. Rev. Genet.* **40**:469–486. DOI:10.1146/annurev.genet.40.110405.090546
115. Chapman S. J. and Hill A. V. (2012). Human genetic susceptibility to infectious disease. *Nat. Rev. Genet.* **13**:175–188. DOI:10.1038/nrg3114
116. Crouch D. J. M. and Bodmer W. F. (2020). Polygenic inheritance, GWAS, polygenic risk scores, and the search for functional variants. *Proc. Natl. Acad. Sci. USA* **117**:18924–18933. DOI:10.1073/pnas.2005634117
117. Bergstrom A., Stringer C., Hajdinjak M., et al. (2021). Origins of modern human ancestry. *Nature* **590**:229–237. DOI:10.1038/s41586-021-03244-5
118. Deng L., Pan Y., Wang Y., et al. (2022). Genetic connections and convergent evolution of tropical indigenous peoples in Asia. *Mol. Biol. Evol.* **39**:msab361. DOI:10.1093/molbev/msab361
119. Zhang X., Liu Q., Zhang H., et al. (2022). The distinct morphological phenotypes of Southeast Asian aborigines are shaped by novel mechanisms for adaptation to tropical rainforests. *Natl. Sci. Rev.* **9**:nwab072. DOI:10.1093/nsr/nwab072
120. Segurel L., Guarino-Vignon P., Marchi N., et al. (2020). Why and when was lactase persistence selected for? Insights from Central Asian herders and ancient DNA. *PLoS Biol.* **18**:e3000742. DOI:10.1371/journal.pbio.3000742
121. Evershed R. P., Davey Smith G., Roffet-Salque M., et al. (2022). Dairying, diseases and the evolution of lactase persistence in Europe. *Nature* **608**:336–345. DOI:10.1038/s41586-022-05010-7
122. Duan S., Wang M., Wang Z., et al. (2024). Malaria resistance-related biological adaptation and complex evolutionary footprints inferred from one integrative Tai-Kadai-related genomic resource. *Heliyon* **10**:e29235. DOI:10.1016/j.heliyon.2024.e29235
123. Genovese G., Friedman D. J., Ross M. D., et al. (2010). Association of trypanolytic ApoL1 variants with kidney disease in African Americans. *Science* **329**:841–845. DOI:10.1126/science.1193032
124. Ko W. Y., Rajan P., Gomez F., et al. (2013). Identifying Darwinian selection acting on different human APOL1 variants among diverse African populations. *Am. J. Hum. Genet.* **93**:54–66. DOI:10.1016/j.ajhg.2013.05.014
125. Larsen C. S. (2023). The past 12,000 years of behavior, adaptation, population, and evolution shaped who we are today. *Proc. Natl. Acad. Sci. USA* **120**:e2209613120. DOI:10.1073/pnas.2209613120
126. Fan S., Hansen M. E., Lo Y., et al. (2016). Going global by adapting local: A review of recent human adaptation. *Science* **354**:54–59. DOI:10.1126/science.aaf5098
127. Benton M. L., Abraham A., LaBella A. L., et al. (2021). The influence of evolutionary history on human health and disease. *Nat. Rev. Genet.* **22**:269–283. DOI:10.1038/s41576-020-00305-9
128. Deng L., Zhang C., Yuan K., et al. (2019). Prioritizing natural-selection signals from the deep-sequencing genomic data suggests multi-variant adaptation in Tibetan highlanders. *Natl. Sci. Rev.* **6**:1201–1222. DOI:10.1093/nsr/nwz108
129. Ma X.-K., Yu Y., Huang T., et al. (2024). Bioinformatics software development: Principles and future directions. *The Innovation Life* **2**:100083. DOI:10.59717/j.xinn-life.2024.100083
130. Huerta-Sanchez E., Jin X., Asan, et al. (2014). Altitude adaptation in Tibetans caused by introgression of Denisovan-like DNA. *Nature* **512**:194–197. DOI:10.1038/nature13408
131. Xiang K., Ouzhuluobu, Peng Y., et al. (2013). Identification of a Tibetan-specific mutation in the hypoxic gene EGLN1 and its contribution to high-altitude adaptation. *Mol. Biol. Evol.* **30**:1889–1898. DOI:10.1093/molbev/mst090
132. Jablonski N. G. and Chaplin G. (2010). Colloquium paper: Human skin pigmentation as an adaptation to UV radiation. *Proc. Natl. Acad. Sci. USA* **107** Suppl 2:8962–8968. DOI:10.1073/pnas.0914628107
133. Ilardo M. A., Moltke I., Korneliussen T. S., et al. (2018). Physiological and genetic adaptations to diving in sea nomads. *Cell* **173**:569–580.e515. DOI:10.1016/j.cell.2018.03.054
134. Quan C., Li Y., Liu X., et al. (2021). Characterization of structural variation in Tibetans reveals new evidence of high-altitude adaptation and introgression. *Genome. Biol.* **22**:159. DOI:10.1186/s13059-021-02382-3
135. Cohen P., Levy J. D., Zhang Y., et al. (2014). Ablation of PRDM16 and beige adipose causes metabolic dysfunction and a subcutaneous to visceral fat switch. *Cell* **156**:304–316. DOI:10.1016/j.cell.2013.12.021
136. Irving-Pease E. K., Refoyo-Martinez A., Barrie W., et al. (2024). The selection landscape and genetic legacy of ancient Eurasians. *Nature* **625**:312–320. DOI:10.1038/s41586-023-06705-1
137. Ma X., Lu Y., Stoneking M., et al. (2025). Neanderthal adaptive introgression shaped LCT enhancer region diversity without linking to lactase persistence in East Asian populations. *Proc. Natl. Acad. Sci. USA* **122**:e2404393122. DOI:10.1073/pnas.2404393122
138. Su H., Wang M., Li X., et al. (2024). Population genetic admixture and evolutionary history in the Shandong Peninsula inferred from integrative modern and ancient genomic resources. *BMC Genom.* **25**:611. DOI:10.1186/s12864-024-10514-9
139. Yang Z., Chen H., Lu Y., et al. (2022). Genetic evidence of tri-genealogy hypothesis on the origin of ethnic minorities in Yunnan. *BMC Biol.* **20**:166. DOI:10.1186/s12915-022-01367-3
140. Kwiatkowski D. P. (2005). How malaria has affected the human genome and what human genetics can teach us about malaria. *Am. J. Hum. Genet.* **77**:171–192. DOI:10.1086/432519
141. Wirth T., Hildebrand F., Allix-Beguec C., et al. (2008). Origin, spread and demography of the Mycobacterium tuberculosis complex. *PLoS Pathog.* **4**:e1000160. DOI:10.1371/journal.ppat.1000160
142. Qi H., Zhang Y. B., Sun L., et al. (2017). Discovery of susceptibility loci associated with tuberculosis in Han Chinese. *Hum. Mol. Genet.* **26**:4752–4763. DOI:10.1093/hmg/ddx365
143. Mi Z., Liu H. and Zhang F. (2024). Advances in the pathogenic, genetic and immunological studies of leprosy. *hLife* **2**:6–17. DOI:10.1016/j.hlife.2023.10.003
144. Yang Z., Zhong H., Chen J., et al. (2016). A Genetic Mechanism for Convergent Skin Lightening during Recent Human Evolution. *Mol. Biol. Evol.* **33**:1177–1187. DOI:10.1093/molbev/msw003
145. Zhang X., Ji X., Li C., et al. (2022). A Late Pleistocene human genome from Southwest China. *Curr. Biol.* **32**:3095–3109.e3095. DOI:10.1016/j.cub.2022.06.016
146. Feng Q., Lu Y., Ni X., et al. (2017). Genetic history of Xinjiang's Uyghurs suggests bronze age multiple-way contacts in Eurasia. *Mol. Biol. Evol.* **34**:2572–2582. DOI:10.1093/molbev/msx177
147. Pan Y., Wen J., Ning Z., et al. (2023). Comparative genomic and transcriptomic analyses reveal the impacts of genetic admixture in Kazaks, Uyghurs, and Huis. *Mol. Biol. Evol.* **40**:msad054. DOI:10.1093/molbev/msad054
148. Keyser C., Bouakaze C., Crubezy E., et al. (2009). Ancient DNA provides new insights into the history of south Siberian Kurgan people. *Hum. Genet.* **126**:395–410. DOI:10.1007/s00439-009-0683-0
149. Peyregne S., Slon V. and Kelso J. (2024). More than a decade of genetic research on the Denisovans. *Nat. Rev. Genet.* **25**:83–103. DOI:10.1038/s41576-023-00643-4
150. Wei X., Robles C. R., Pazokitoroudi A., et al. (2023). The lingering effects of Neanderthal introgression on human complex traits. *eLife* **12**:e80757. DOI:10.7554/eLife.80757
151. Porubsky D. and Eichler E. E. (2024). A 25-year odyssey of genomic technology advances and structural variant discovery. *Cell* **187**:1024–1037. DOI:10.1016/j.cell.2024.01.002
152. Larena M., McKenna J., Sanchez-Quinto F., et al. (2021). Philippine Ayta possess the highest level of Denisovan ancestry in the world. *Curr. Biol.* **31**:4219–4230.e4210. DOI:10.1016/j.cub.2021.07.022
153. Chen L., Wolf A. B., Fu W., et al. (2020). Identifying and Interpreting Apparent Neanderthal Ancestry in African Individuals. *Cell* **180**:677–687.e616. DOI:10.1016/j.cell.2020.01.012
154. Mallick S., Micco A., Mah M., et al. (2024). The Allen Ancient DNA Resource (AADR) a curated compendium of ancient human genomes. *Sci. Data* **11**:182. DOI:10.1038/s41597-024-03031-7
155. Allentoft M. E., Sikora M., Fischer A., et al. (2024). 100 ancient genomes show repeated population turnovers in Neolithic Denmark. *Nature* **625**:329–337. DOI:10.1038/s41586-023-06862-3
156. Liu Y. C., Hunter-Anderson R., Cheronet O., et al. (2022). Ancient DNA reveals five streams of migration into Micronesia and matrilocality in early Pacific seafarers. *Science* **377**:72–79. DOI:10.1126/science.abm6536
157. Lazaridis I., Alpaslan-Roodenberg S., Acar A., et al. (2022). The genetic history of the Southern Arc: A bridge between West Asia and Europe. *Science* **377**:eabm4247. DOI:10.1126/science.abm4247
158. Lazaridis I., Alpaslan-Roodenberg S., Acar A., et al. (2022). A genetic probe into the ancient and medieval history of Southern Europe and West Asia. *Science* **377**:940–951. DOI:10.1126/science.abq0755
159. Olalde I., Carrion P., Mikic I., et al. (2023). A genetic history of the Balkans from Roman frontier to Slavic migrations. *Cell* **186**:5472–5485.e5479. DOI:10.1016/j.cell.2023.10.018
160. Allentoft M. E., Sikora M., Refoyo-Martinez A., et al. (2024). Population genomics of post-glacial western Eurasia. *Nature* **625**:301–311. DOI:10.1038/s41586-023-06865-0
161. Lipson M., Sawchuk E. A., Thompson J. C., et al. (2022). Ancient DNA and deep population structure in sub-Saharan African foragers. *Nature* **603**:290–296. DOI:10.1038/s41586-022-04430-9
162. Skoglund P., Posth C., Sirak K., et al. (2016). Genomic insights into the peopling of the Southwest Pacific. *Nature* **538**:510–513. DOI:10.1038/nature19844
163. Moreno-Mayar J. V., Vinner L., de Barros Damgaard P., et al. (2018). Early human dispersals within the Americas. *Science* **362**:eaav2621. DOI:10.1126/science.aav2621
164. Willerslev E. and Meltzer D. J. (2021). Peopling of the Americas as inferred from

- ancient genomics. *Nature* **594**:356–364. DOI:10.1038/s41586-021-03499-y
165. Nakatsuka N., Holguin B., Sedig J., et al. (2023). Genetic continuity and change among the Indigenous peoples of California. *Nature* **624**:122–129. DOI:10.1038/s41586-023-06771-5
 166. Zhang F., Ning C., Scott A., et al. (2021). The genomic origins of the Bronze Age Tarim Basin mummies. *Nature* **599**:256–261. DOI:10.1038/s41586-021-04052-7
 167. Xiong J., Wang R., Chen G., et al. (2024). Inferring the demographic history of Hexi Corridor over the past two millennia from ancient genomes. *Sci. Bull.* **69**:606–611. DOI:10.1016/j.scib.2023.12.031
 168. Wang H., Yang M. A., Wangdue S., et al. (2023). Human genetic history on the Tibetan Plateau in the past 5100 years. *Sci. Adv.* **9**:ead5582. DOI:10.1126/sciadv.ead5582
 169. Kerner G., Neehus A. L., Philippot Q., et al. (2023). Genetic adaptation to pathogens and increased risk of inflammatory disorders in post-Neolithic Europe. *Cell Genom.* **3**:100248. DOI:10.1016/j.xgen.2022.100248
 170. Barrie W., Yang Y., Irving-Pease E. K., et al. (2024). Elevated genetic risk for multiple sclerosis emerged in steppe pastoralist populations. *Nature* **625**:321–328. DOI:10.1038/s41586-023-06618-z
 171. Klunk J., Vilgaly T. P., Demeure C. E., et al. (2022). Evolution of immune genes is associated with the Black Death. *Nature* **611**:312–319. DOI:10.1038/s41586-022-05349-x
 172. Brace S., Diekmann Y., Booth T., et al. (2022). Genomes from a medieval mass burial show Ashkenazi-associated hereditary diseases pre-date the 12th century. *Curr. Biol.* **32**:4350–4359.e4356. DOI:10.1016/j.cub.2022.08.036
 173. Davy T., Ju D., Mathieson I., et al. (2023). Hunter-gatherer admixture facilitated natural selection in Neolithic European farmers. *Curr. Biol.* **33**:1365–1371.e1363. DOI:10.1016/j.cub.2023.02.049
 174. Almari M. A., Bergstrom A., Prado-Martinez J., et al. (2020). Population Structure, Stratification, and Introgression of Human Structural Variation. *Cell* **182**:189–199.e115. DOI:10.1016/j.cell.2020.05.024
 175. Collins R. L., Brand H., Karczewski K. J., et al. (2020). A structural variation reference for medical and population genetics. *Nature* **581**:444–451. DOI:10.1038/s41586-020-2287-8
 176. Luo M., Yang W., Bai L., et al. (2024). Artificial intelligence for life sciences: A comprehensive guide and future trends. *The Innovation Life* **2**:100105. DOI:10.59717/j.xinn-life.2024.100105
 177. Audano P. A., Sulovari A., Graves-Lindsay T. A., et al. (2019). Characterizing the Major Structural Variant Alleles of the Human Genome. *Cell* **176**:663–675.e619. DOI:10.1016/j.cell.2018.12.019
 178. Ebert P., Audano P. A., Zhu Q., et al. (2021). Haplotype-resolved diverse human genomes and integrated analysis of structural variation. *Science* **372**:eabf7117. DOI:10.1126/science.abf7117
 179. Siren J., Monlong J., Chang X., et al. (2021). Pangenomics enables genotyping of known structural variants in 5202 diverse genomes. *Science* **374**:abg8871. DOI:10.1126/science.abg8871
 180. Beyter D., Ingimundardottir H., Oddsson A., et al. (2021). Long-read sequencing of 3,622 Icelanders provides insight into the role of structural variants in human diseases and other traits. *Nat. Genet.* **53**:779–786. DOI:10.1038/s41588-021-00865-4
 181. Shi J., Jia Z., Sun J., et al. (2023). Structural variants involved in high-altitude adaptation detected using single-molecule long-read sequencing. *Nat. Commun.* **14**:8282. DOI:10.1038/s41467-023-44034-z
 182. Ouzhuluobu, He Y., Lou H., et al. (2020). De novo assembly of a Tibetan genome and identification of novel structural variants associated with high-altitude adaptation. *Natl. Sci. Rev.* **7**:391–402. DOI:10.1093/nsr/nwz160
 183. Lou H., Gao Y., Xie B., et al. (2022). Haplotype-resolved de novo assembly of a Tujia genome suggests the necessity for high-quality population-specific genome references. *Cell Syst.* **13**:321–333.e326. DOI:10.1016/j.cels.2022.01.006
 184. Yang X., Zhao X., Qu S., et al. (2022). Haplotype-resolved Chinese male genome assembly based on high-fidelity sequencing. *Fundam. Res.* **2**:946–953. DOI:10.1016/j.fmre.2022.02.005
 185. Yang C. T., Zhou Y., Song Y. N., et al. (2023). The complete and fully-phased diploid genome of a male Han Chinese. *Cell Res.* **33**:745–761. DOI:10.1038/s41422-023-00849-5
 186. He Y., Chu Y., Guo S., et al. (2023). T2T-YAO: A Telomere-to-telomere Assembled Diploid Reference Genome for Han Chinese. *Genomics, Proteomics Bioinform.* **21**:1085–1100. DOI:10.1016/j.gpb.2023.08.001
 187. 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
 188. Yang T., Ma F., Qian H., et al. (2024). AI-driven construction of digital cell model. *The Innovation Life* **2**:100102. DOI:10.59717/j.xinn-life.2024.100102
 189. Alpaslan-Roodenberg S., Anthony D., Babiker H., et al. (2021). Ethics of DNA research on human remains: five globally applicable guidelines. *Nature* **599**:41–46. DOI:10.1038/s41586-021-04008-x
 190. Fatumo S., Chikowore T., Choudhury A., et al. (2022). A roadmap to increase diversity in genomic studies. *Nat. Med.* **28**:243–250. DOI:10.1038/s41591-021-01672-4
 191. Madden E. B., Hindorff L. A., Bonham V. L., et al. (2024). Advancing genomics to improve health equity. *Nat. Genet.* **56**:752–757. DOI:10.1038/s41588-024-01711-z
 192. Green E. D., Gunter C., Biesecker L. G., et al. (2020). Strategic vision for improving human health at The Forefront of Genomics. *Nature* **586**:683–692. DOI:10.1038/s41586-020-2817-4
 193. Ping J., Liu X., Lu Y., et al. (2025). A highland-adaptation variant near MCUR1 reduces its transcription and attenuates erythropoiesis in Tibetans. *Cell Genom.* **5**:100782. DOI:10.1016/j.xgen.2025.100782
 194. Gong J., Sun H., Wang K., et al. (2025). Long-read sequencing of 945 Han individuals identifies structural variants associated with phenotypic diversity and disease susceptibility. *Nat. Commun.* **16**:1494. DOI:10.1038/s41467-025-56661-9

FUNDING AND ACKNOWLEDGMENTS

This study was supported by the National Natural Science Foundation of China (82402203), the Sichuan Science and Technology Program (2024NSFSC1518), the 1.3.5 Project for Disciplines of Excellence, West China Hospital, Sichuan University (ZYJC20002), the Major Project of the National Social Science Foundation of China (23&ZD203), the Open Project of the Key Laboratory of Forensic Genetics of the Ministry of Public Security (2022FGKFKT05 and 2024FGKFKT02), and the Center for Archaeological Science of Sichuan University (23SASA01 and 24SASB03). The funders had no role in the study design, data collection and analysis, or decision to publish or prepare the manuscript.

AUTHOR CONTRIBUTIONS

M.G.W. and S.H.D. contributed to collecting information from other references, discussing the results, and initiating diagrams. X.P.L. contributed to drawing the diagrams. M.G.W. corrected and proofread the paper. M.G.W., J.B.Y., H.J.Y., CL., and G.L.H. established review ideas, information collection, and writing. All authors read and approved the final manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

ETHICAL STATEMENT AND PATIENT CONSENT

Not applicable

DATA AND CODE AVAILABILITY

Not applicable