

High-quality Population-specific Haplotype-resolved Reference Panel in the Genomic and Pangenomic Eras

Qingxin Yang  (杨青鑫)^{1,2,3,#}, Yuntao Sun  (孙韵韬)^{1,2,4}, Shuhan Duan  (段淑涵)^{1,2,5},
Shengjie Nie  (聂胜洁)³, Chao Liu  (刘超)⁶, Hong Deng  (邓虹)^{3,*},
Mengge Wang  (王萌鸽)^{1,7,8,9,#,*}, Guanglin He  (何光林)^{1,2,#,*}

¹Department of Oto-Rhino-Laryngology & Institute of Rare Diseases, West China Hospital of Sichuan University, Sichuan University, Chengdu 610000, China

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

³School of Forensic Medicine, Kunming Medical University, Kunming 650500, China

⁴West China School of Basic Science & Forensic Medicine, Sichuan University, Chengdu 610000, China

⁵School of Basic Medical Sciences, North Sichuan Medical College, Nanchong 637100, China

⁶Anti-Drug Technology Center of Guangdong Province, Guangzhou 510230, China

⁷Faculty of Forensic Medicine, Zhongshan School of Medicine, Sun Yat-sen University, Guangzhou 510275, China

⁸Department of Forensic Medicine, College of Basic Medicine, Chongqing Medical University, Chongqing 400331, China

⁹Human Genetics and Forensic Genomics Research Institute, College of Basic Medicine, Chongqing Medical University, Chongqing 400331, China

*Corresponding authors: denghong@kmmu.edu.cn (Deng H), menggewang@wchscu.cn (Wang M), guanglinhe@wchscu.edu.cn (He G).

#Equal contribution.

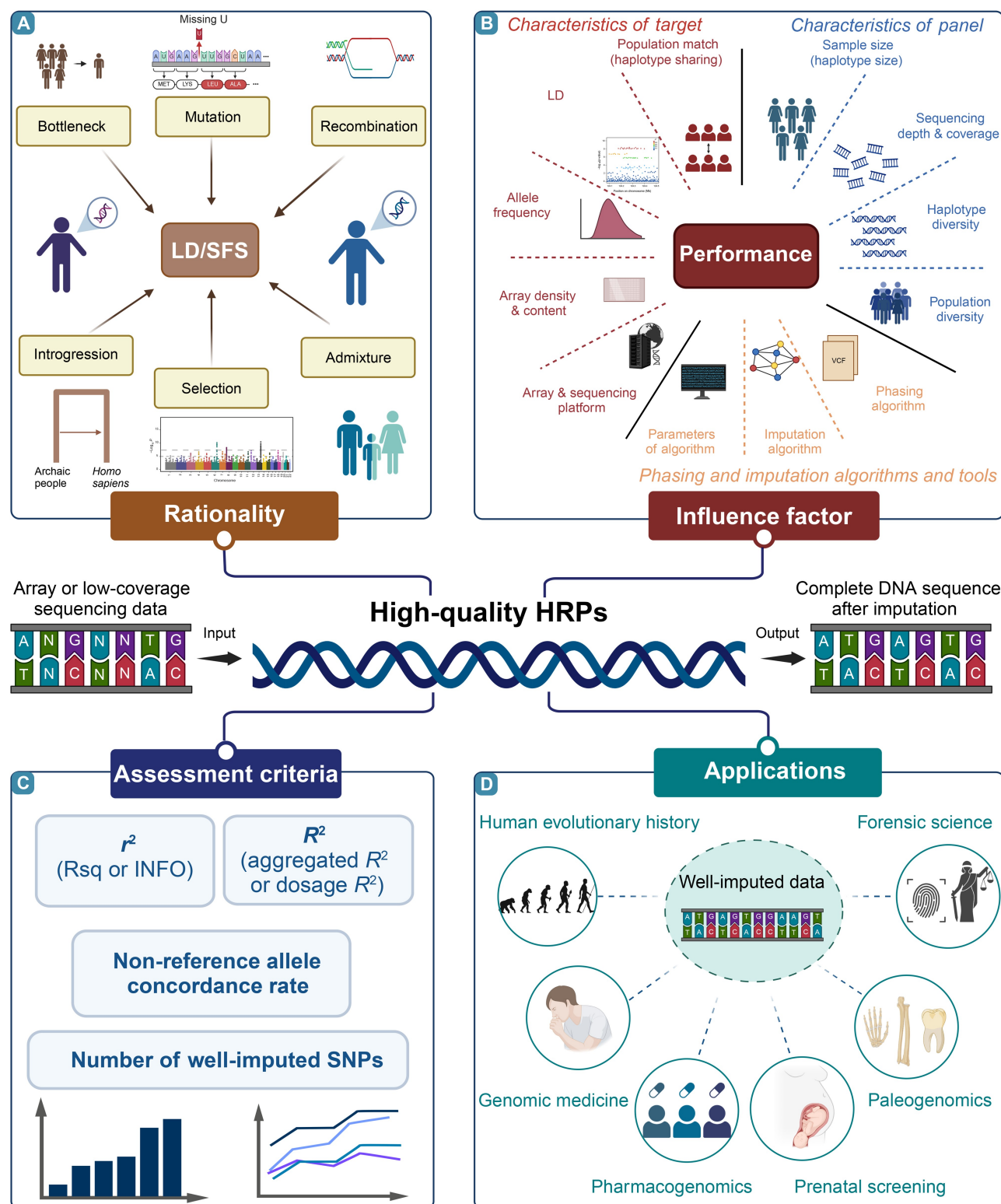
Handling Editor: Minxian Wang

Abstract

Large-scale international and regional human genomic and pangenomic resources derived from population-scale biobanks and ancient DNA sequences have provided significant insights into human evolution and the genetic determinants of complex diseases and traits. Despite these advances, challenges persist in optimizing the integration of phasing tools, merging haplotype reference panels (HRPs), developing imputation algorithms, and fully exploiting the diverse applications of post-imputation data. This review comprehensively summarizes the advancements, applications, limitations, and future directions of HRPs in human genomics research. Recent progress in the reconstruction of HRPs, based on over 830,000 human whole-genome sequences, has been synthesized, highlighting the broad spectrum of human genetic diversity captured. Additionally, we recapitulate advancements in 56 HRPs for global and regional populations. The evaluation of imputation accuracy indicated that Beagle and Glimpse are the most effective tools for phasing and imputing data from genotyping arrays and low-coverage sequencing, respectively. A critical strategy for selecting an appropriate HRP involves matching the population background of target groups with HRP reference populations and considering multi-ancestry or homogeneous genetic structures. The necessity of a single, integrative, high-quality HRP that captures haplotype structures and genetic diversity across various genetic variation types from globally representative populations is emphasized to support both modern and ancient genomic research and advance human precision medicine.

Key words: Haplotype reference panel; Imputation accuracy; Genotype imputation; Population genomics; Genomic medicine.

Graphical abstract



Introduction

High-coverage whole-genome sequencing (WGS) has long been regarded as the gold standard for genotyping and identifying single-nucleotide polymorphisms (SNPs), insertions and deletions (InDels), and other structural variations (SVs) [1–6].

However, its high cost presents a significant barrier to population-scale genomic studies involving large cohorts. Consequently, much of the current research on modern humans depends on microarrays or low-depth WGS, which complicates genotype calling and restricts the ability to

capture global genetic diversity [7]. Similar challenges have been observed in studies involving ancient DNA (aDNA) [8] and human leukocyte antigen (HLA) genes [9,10]. The established theoretical frameworks suggest that individuals within genetically close populations share population-specific long haplotype segments inherited from common ancestors. Accordingly, population-specific haplotype reference panels (HRPs), which consist of DNA sequences characterized by linkage disequilibrium (LD) and are tailored to document patterns of human genetic diversity, have been developed to facilitate the imputation of common, low-frequency, and rare genetic variations not directly genotyped in medical genome research or missing variants in aDNA studies.

Genetic variant information can be predicted or inferred through genotype imputation based on the documented patterns of haplotype structure in HRP [11]. The accuracy of imputation depends mainly on the genetic architecture of the target populations (multi-ancestry heterogeneous background or genetically homogeneous structures) [12–14], algorithms such as the hidden Markov model (HMM) or deep learning employed by phasing and imputation tools [12–14], and properties of HRP, such as marker density, the allele frequency spectrum, and the composition of reference populations [15–17]. Most imputation tools, including IMPUTE [18] and Markov Chain Haplotyping algorithm (MaCH) [19], leverage the HMM algorithm; however, updates to these tools have primarily targeted improvements in computational efficiency or memory requirements. In contrast, advances in HRP, designed to capture the full breadth of human genetic diversity across distinct modern populations at scale, have significantly enhanced the quality of HRP and imputation performance [20]. As a result, numerous HRP have been constructed based on global genomic projects, such as the 1000 Genomes Project (1KGP), the Haplotype Reference Consortium (HRC) program, and Trans-Omics for Precision Medicine (TOPMed) [1–3]. More recently, population-specific HRP, tailored to regional populations such as Han Chinese from the NYUWa Genome resource, Westlake BioBank for Chinese (WBBC), the Chinese Academy of Sciences Precision Medicine Initiative (CASPMI) project, and China Metabolic Analytics Project (ChinaMAP), Korean HRP from the Northeast Asian Reference Database (NARD), or meta-Asian HRP from the South and East Asian reference Database (SEAD), have improved the genetic discovery of these target populations [4,14,21–23]. Following imputation, low-coverage WGS or array-based databases enable the capture of more low-frequency or rare variations, increasing marker density, enhancing the statistical power of large-scale genome-wide association studies (GWASs) or meta-based work across different cohorts, and offering a cost-effective genotyping approach for downstream analyses, including polygenic risk score (PRS) estimation, genetic genealogy reconstruction, and demographic genetic history inference [24,25].

The imputation process introduces systematic errors that are difficult to avoid, particularly when analyzing rare and low-frequency variants [26]. As the minor allele frequency (MAF) decreases, the imputation error rate increases [27]. Furthermore, imputation performance reliant on LD is shaped by haplotype patterns and variant spectra, which vary across genetically distinct populations [4]. Recent genetic studies have demonstrated that demographic events, including severe human bottlenecks occurring during migration out

of Africa, along with biological adaptations driven by environmental extremes, pathogen exposure, dietary changes [28], and other evolutionary forces, such as mutation, migration, admixture, introgression, and recombination, significantly reshaped patterns of genetic diversity, allele frequency spectra, and LD patterns (Figure 1A). Template switching rates of theta values in the multi-ancestry HRP also influence the imputation accuracy and imputation quality metrics of minor ancestry [29]. Populations with closer genetic affinities tend to share longer haplotypes and similar variant spectra, consistent with coalescence theory, which facilitates the acquisition of more precise haplotype information. However, HRP, primarily established from European-specific haplotype data, have proven inadequate for genotype imputation in African, Asian, and other underrepresented indigenous populations [30,31]. This European-centric bias has led to inaccuracies in interpreting population-specific genetic foundations and conducting medical genetic studies focused on complex phenotypes in genetically diverse non-European populations [22,32,33]. Evidence indicates that reducing imputation error rates for low-coverage WGS and low-density array data can be achieved through factors such as larger sample sizes, deeper sequencing depths, better population representation, and closer genetic matching between reference and target populations (Figure 1B) [20,34,35].

Thus, constructing an integrative HRP encompassing worldwide population-specific haplotypes with diverse patterns of genetic diversity remains critical for advancing both population and medical genetics research. The foundational concepts and evaluation of HRP have been extensively reviewed in prior studies [36,37]. Earlier reviews provide detailed insights into the interplay between imputation advancements and GWASs, particularly regarding common and rare diseases [11,36,38]. Over the past eight years, human genomics research and computational genomics have undergone transformative and explosive advancements driven by innovations in multi-ancestry and population-specific HRP. These developments have significantly advanced both basic research and clinical applications [1,14,25,39,40]. This review synthesizes recent advances in HRP and genotype imputation tools, providing an overview of five key areas: (1) the progress and geographical and ethnic distribution patterns of populations in global human genome projects and corresponding HRP, with a focus on large-scale multi-ancestry or population-specific high-quality HRP; (2) the current status and efficiency of state-of-the-art phasing and imputation tools in the era of large-scale genomics; (3) performance assessments of publicly available HRP and the necessity for high-quality merged HRP (Figure 1C); (4) applications of imputed datasets in human genetics, genome science, genome medicine, and forensic science (Figure 1D); and (5) challenges in HRP applications and future directions in the context of the telomere-to-telomere (T2T)/pangenome-based paradigm shift.

Advances in worldwide human genome projects and corresponding HRP in the past two decades

The draft human genome sequence was completed two decades ago, marking a pivotal advancement supported by the Genome Reference Consortium's publication of human reference genomes. Innovations in sequencing and computational

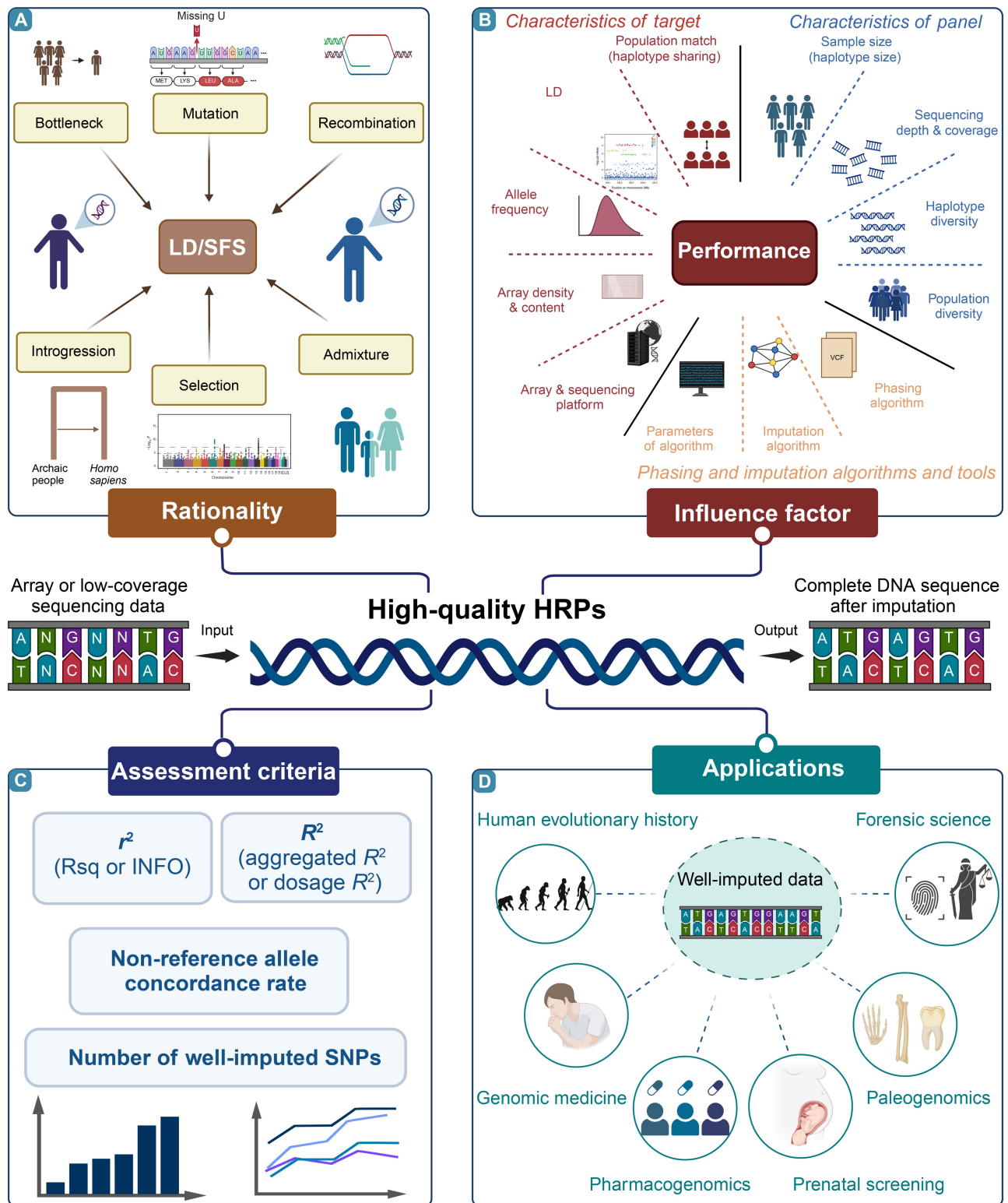


Figure 1 Summary of the HRP

A. Differences in LD and SFS patterns among genetically distinct populations constitute the fundamental reason for the need for population-specific HRPs in precision medicine. These patterns are affected by multiple factors, including bottlenecks, mutations, admixture, introgression, recombination, and selection. **B.** Factors influencing the efficiency of HRPs include the following three aspects: the characteristics of the HRP (sample size, sequencing depth and coverage, haplotype diversity, and population diversity), the type and parameters of phasing and imputation tools, and the characteristics of the target population to be imputed (array or sequencing platform, array density and content, allele frequencies, LD patterns between loci, and the matching of the target population with the populations in the HRP). **C.** Metrics for measuring imputation quality include the statistics generated by imputation tools (r^2 or INFO score), the squared Pearson's correlation coefficients (aggregated R^2 or dosage R^2) between the true genotypes and imputed genotype dosages, non-reference allele concordance rate, and the number of well-imputed SNPs. **D.** Well-imputed data are applied in genomics studies of human evolutionary history, genomic medicine, pharmacogenomics, prenatal screening, paleogenomics, and forensic science. HRP, haplotype reference panel; LD, linkage disequilibrium; SFS, site frequency spectrum; SNP, single-nucleotide polymorphism.

techniques have accelerated the transition from single-genome studies to large-scale population genomic projects. In 2005, the International HapMap Consortium released the first haplotype map of the human genome, followed by the first HRP [20,41], marking a significant milestone (Figure 2A). Subsequent phases, including the pilot, phase 1, and phase 3 of the 1KGP, as well as the expanded 1KGP leveraging high-depth WGS technology, have provided valuable insights into the genomic diversity across major continental populations (Figure 2B) [3,42–44]. These initiatives have uncovered previously uncharacterized gaps in missing sequences and diversity, which were attributed to the limitations of array-based genotyping in earlier studies, and have also identified extensive human genetic variations, such as SNPs, InDels, and SVs. This progress heralded a new era of comprehensive human DNA sequencing and research. The availability of data from these genomic projects has enabled integrative analyses of human genome variations, contributing to the development of numerous large-size, multi-ancestry, high-quality HRPs or regional population-specific HRPs. These efforts have significantly advanced our understanding of the genomic architecture and genetic determinants of clinical diseases and complex physical traits (Figure 2C–F; Table 1, Table S1).

Recent advancements in high-quality HRP innovations have emerged at a rapid pace. In Europe, the HRC program and subsequent European genomic initiatives, including the Genome of the Netherlands Project (GoNL), 10,000 UK Genome Sequences (UK10K), Genomics England (GEL), and other genomic studies of Estonians, Sardinians, Cretans, Ashkenazi Jewish, Icelanders, and Latvians, have yielded high-quality HRPs and extensive genetic datasets [39,45–58]. These large-scale human genomic resources have facilitated high-quality HRP construction, fine-scale genetic analyses, and significant advancements in precision medicine across diverse European populations (Figure 2E). Genetic analysis revealed a striking European bias in human genome research, with 86% of GWASs conducted in populations of European ancestry, resulting in an overrepresentation of European populations in HRPs [28,33,59]. Efforts to develop genomic projects focused on Asian populations have aimed to address the underrepresentation and selection bias of non-European groups in human genome research and to illuminate novel genetic determinants of East Asian-specific diseases and health traits. Over 30 initiatives focused on East Asians, Southeast Asians, or single-ethnic groups such as Chinese Mongolians and Tibetans have recently been launched, significantly contributing to this effort (Figure 2C) [4,12,13,22,23,32,60–82]. Northern Americans, historically dominated by migrants from other continents with complex mixed ancestry, exhibit diverse genetic backgrounds. The peopling of the Americas is traced back to at least the late Paleolithic period [83,84]. Native American populations possess 14% to 38% ancestry related to MA-1, a 24,000-year-old individual from Mal'ta in south-central Siberia, with the remainder of their ancestry deriving from East Asians [83,84]. Latin American populations comprise both admixed and predominantly indigenous subpopulations, many of which remain genetically uncharacterized. Genomic initiatives in Northern America, such as the TOPMed program, the Consortium on Asthma among African-ancestry Populations in the Americas (CAAPA), the African Americans reference panel (AFAM), the Anabaptist Genome Reference Panel (AGRP), and others, have aimed to

diversify genetic collections across these groups (Figure 2F) [1,40,85–90]. Africa, recognized as the cradle of modern humans, harbors immense genetic and linguistic diversity, including Afro-Asiatic, Nilo-Saharan, Khoisan, and Niger-Congo language groups. However, few cohort studies have been conducted to explore the genetic architecture and disease susceptibility in African populations, with notable exceptions including the African Genome Variation Project and Uganda Genome Resource [31,91]. For Pacific islanders, only one HRP specific to Samoan haplotype genotypes has been developed, while public reference panels lack sufficient samples from Oceania. This underrepresentation exacerbates the health disparities faced by Pacific islanders [92]. Collectively, these genomic projects and the reference panels developed in parallel underscore global efforts to advance equitable healthcare and precision medicine. Seventeen of these panels are accessible through publicly available online imputation platforms, encompassing samples from European, Asian, African, and indigenous Oceanian ancestries (Table S1).

In addition to enhancing diversity by including genetically distinct continental populations, imputation accuracy was found to improve with an increased number of haplotypes and samples in the HRPs, particularly for rare and low-frequency variants [11,25,38]. Additionally, closer genetic proximity between the target population and the populations in HRPs corresponded to higher imputation accuracy, as demonstrated in these validation studies [25,58,93,94]. For example, when the TOPMed HRP was used, the mean imputation r -squared (r^2) for European variants was 0.93, compared to 0.62 for Papua New Guineans [25]. For populations with excess African ancestry, both the TOPMed and AGRP HRPs demonstrated superior imputation accuracy [95]. Meta-imputation utilizing HRPs from the TOPMed and the expanded 1KGP yielded the best results for the Pakistani population [96]. The inclusion of diverse reference panels has been shown to increase imputation accuracy for rare variants significantly [38]. Analysis of sample size and ancestry composition in high-quality HRPs (sequencing depth $\geq 30\times$, sample size > 1000) revealed that individuals of European ancestry accounted for 60.7%. In fine-scale regional or isolated populations, the selection bias or Han bias has hindered the equitable representation of ethnolinguistically diverse ethnic minorities within regional population genomic cohorts (Figure 3). Asian HRPs with unspecified sequencing depths were excluded from this analysis (Table S1). While large-scale, population-specific initiatives are essential for unlocking the full potential of genome sequencing, non-European populations remain underrepresented in genomic studies owing to the high cost of sequencing. Consequently, the development of high-quality, population-specific HRPs may offer a more effective strategy for advancing genomic research in developing countries. Given the limited integration of population genetic backgrounds in current molecular anthropological and medical studies, a combined approach is recommended. This approach would involve merging high-quality sequencing data from genetically diverse cohorts, such as the 1KGP and Human Genome Diversity Project (HGDP), or integrating population-specific reference panels with multi-ethnic resources, such as the TOPMed and SEAD reference panels. Such efforts would facilitate the creation of a comprehensive, multi-ancestry, high-quality HRP that better represents underrepresented populations or those with

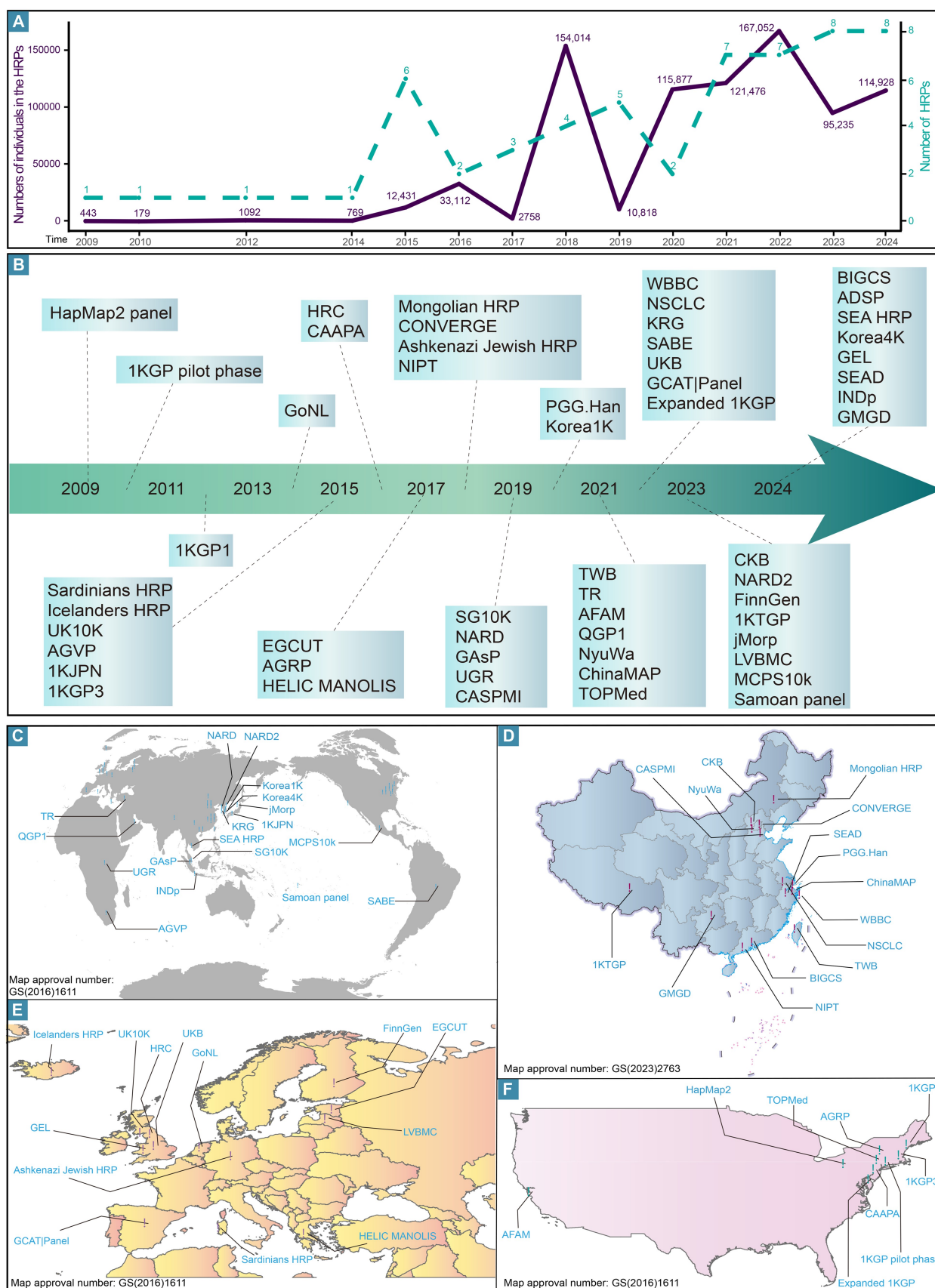


Figure 2 Geographical distribution and timeline of published HRP's

A. Line charts showing the number of HRP's published annually (green dashed line) and their sample sizes (purple solid line). **B.** Timeline of published genome projects. **C.-F.** HRP's in different regions around the world (C), in China (D), in Europe (E), and in America (F). The annotations on the map represent only the general geographical location, and see Table S1 for details. HapMap, International HapMap Project; 1KGP, 1000 Genomes Project;

mixed ancestral origins, including European, American, Oceanian, Singaporean, and South African populations.

Optimal combination of phasing and imputation tools

Haplotype phasing, including methods such as family-based phasing, physical phasing, and LD-based phasing, serves as a cornerstone in statistical and population genetics. This approach facilitates the identification of genetic variant combinations on individual chromosomes and underpins the construction of reference panels. Family-based phasing utilizes kinship information, including trios or multigenerational family data, to allocate alleles to haplotypes with high accuracy based on Mendelian inheritance patterns. However, its efficacy relies on the availability of comprehensive family datasets. Physical phasing employs experimental approaches, such as long-read and single-molecule sequencing technologies (e.g., PacBio and Nanopore) and next-generation sequencing, to determine haplotypes directly, often leveraging trios or physically dissected chromosomal markers. LD-based phasing infers haplotypes via LD patterns derived from population-level data. Tools such as SHAPEIT [97] and Beagle [98] have been developed to facilitate haplotype phasing in individuals without family data. However, the accuracy of these methods can be affected by population-specific LD structures.

In the current well-designed statistical phasing strategy applied to array-based data, whole-genome data [whole-exome sequencing (WES) and WGS] primarily rely on haplotype diversity patterns driven by LD within populations or utilize coalescent theory to identify shared identity-by-descent (IBD) segments among samples [99]. The most widely used phasing tools include Eagle [17], SHAPEIT [97], and Beagle [98] (Table S2). Phasing performance is typically assessed using trio samples to establish the true haplotypes of offspring, which are then compared with those inferred by phasing algorithms. Phasing accuracy was evaluated through the switch error rate (SER), defined as the fraction of consecutive heterozygous genotypes that are incorrectly phased. In an analysis of array data from 500 European trio samples, the reported SERs were as follows: SHAPEIT4 (0.117%), Beagle5 (0.125%), and Eagle2 (0.178%) [100]. Updates to SHAPEIT5 and Beagle5.4 have further improved the phasing accuracy for rare variants. In evaluations with European trio samples, both methods demonstrated low SERs (< 0.2%) when considering Axiom array loci alone. However, for rare

variants in WGS data (with minor allele counts between 11 and 20 or allele frequencies less than 0.01), their SERs were 4.36% and 8.76%, respectively, whereas Eagle2 did not significantly improve the phasing performance for rare variants. Enhanced phasing accuracy for rare variants has also contributed to improved genotype imputation accuracy within corresponding high-resolution panels [97]. An independent assessment by Cole et al. examined the phasing performance of SHAPEIT5.1.1 and Beagle5.4 on genetically distinct populations (Table S2). Overall, Beagle demonstrated a lower SER than that SHAPEIT did, achieving extremely low SERs ranging from 0.012% to 0.028% for array data from individuals of European, Ashkenazi Jewish, African-admixed, and American-admixed ancestry. Conversely, average SERs for East Asian and South Asian individuals were reported to be 0.36% (standard deviation 0.17%) and 0.50% (standard deviation 0.34%), respectively. In WGS data, both SHAPEIT and Beagle displayed strong performance within European samples (SER = 0.021%) but exhibited suboptimal results for East Asian samples (SER = 3.49%) [101]. Under comparable conditions, phasing with Beagle yielded superior imputation results, and phasing with a reference panel outperformed reference-free approaches [102].

The imputation tools, developed to optimize uncertain genotype likelihoods and bridge gaps in sparsely mapped reads, were designed to maximize the utility of HRP. Their application significantly improved the accuracy and statistical power of low-coverage WGS data following genotype imputation [39,103]. Constantly updated HRPs, incorporating more individuals from different ancestral populations and newly reported variants, facilitate iterative improvements to these tools, enhancing the accuracy of rare variant detection while minimizing computational time and memory requirements. Various genotyping and sequencing methods, including SNP tagging approaches for genotype-wide SNPs and low-coverage WGS, are commonly employed. Key tools in this domain include IMPUTE, Beagle, and Minimac (Table S2) [18,98,104]. Validation studies indicated that, under consistent conditions, the performance of these imputation tools was broadly comparable, with R^2 differences of no more than 0.01. Beagle demonstrated superior performance, followed by IMPUTE, with Minimac ranking last [102]. Additionally, QUILT and Glimpse, which utilize Gibbs sampling and HMM, were designed explicitly for low-pass WGS data [105,106]. For low-coverage aDNA and non-invasive prenatal testing data, Glimpse slightly outperformed QUILT and Beagle [107,108]. Moreover, both versions of Glimpse

Figure 2 Continued

GoNL, Genome of the Netherlands Project; UK10K, 10,000 UK Genome Sequences; AGVP, African Genome Variation Project; 1KJPN, a reference panel of 1070 Japanese individuals; HRC, Haplotype Reference Consortium (release 1.1); CAAPA, Consortium on Asthma among African-ancestry Populations in the Americas; HELIC MANOLIS, HELIC Pomak collection and the MANOLIS Cohorts; HELIC, Hellenic Isolated Cohorts; MANOLIS, Mylopotamos; EGCUT, Estonian Biobank of the Estonian Genome Center, University of Tartu; AGRP, Anabaptist Genome Reference Panel; NIPT, Non-invasive prenatal testing in China; CONVERGE, the China, Oxford and Virginia Commonwealth University Experimental Research on Genetic Epidemiology; SG10K, whole-genome-sequence 10,000 Singaporeans; NARD, Northeast Asian Reference Database; GAsP, GenomeAsia 100K Project pilot phase; UGR, Uganda Genome Resource; CASPMI, the Chinese Academy of Sciences Precision Medicine Initiative project; PGG.Han, Han Chinese Genome Initiative; Korea1K, the Korean Genome Project; TWB, Taiwan Biobank; TR, Turkish Variome; TOPMed, Trans-Omics for Precision Medicine; QGP1, Qatar Genome Program Phase 1; NyuWa, the NyuWa Genome resource; ChinaMAP, China Metabolic Analytics Project; AFAM, African Americans reference panel; WBBC, Westlake BioBank for Chinese pilot project; GCAT|Panel, GCAT|Genomes for Life Cohort; NSCLC, Non-small cell lung cancer in China; KRG, the Korean Reference Genome project; SABE, The Health, Well-being and Aging Study; CKB, China Kadoorie Biobank; UKB, UK Biobank; FinnGen, FinnGen project; 1KTGP, 1000 Tibetans Genomes Project; jMorp, The Tohoku Medical Megabank project; MCPS10k, the Mexico City Prospective Study; Samoan panel, Samoan-specific genotype reference panels; LVBMC, the Latvian population-specific reference panel; BIGCS, Born in Guangzhou Cohort Study; ADSP, the Alzheimer's Disease Sequencing Project; SEA HRP, the Southeast Asian Specific Reference Panel; Korea4K, the Korean Genome Project; GEL, the Genomics England dataset; SEAD, the South and East Asian Reference Database reference panel; INDp, the Indonesian panel; GMGD, Guizhou Multi-Ethnic Genome Database.

Table 1 Key characteristics and comparisons of all HRP's

Name	Published time	Sample size	Ancestry	Depth	Total number of variants	Phasing algorithm	Imputation algorithm	Assessment criterion	Target population	Imputation performance measurement
HapMap2 panel	2009	443	Multi-ethnic	/	513,008	/	MaCH	r^2	/	/
1KGP pilot phase	2010	179	Multi-ethnic	3.6×	16,224,519	/	IMPUTE	r^2	/	/
1KGP1	2012	1092	Multi-ethnic	5.1×	28,975,367	SHAPEIT2	Beagle	R^2	/	SHAPEIT2 <i>vs.</i> Thunder <i>vs.</i> Beagle
GoNL	2014	769	Netherlands	13×	21,524,538	BEAGLE	IMPUTE2	R^2	Dutch	GoNL + 1KGP <i>vs.</i> GoNL + 1KGP1
Sardinians HRP	2015	2120	Sardinians	20×	21,131,214	/	/	r^2	Sardinians	European <i>vs.</i> GoNL <i>vs.</i> 1KGP1 <i>vs.</i> 1KGP1 European
Icelanders HRP	2015	2636	Iceland	7×	45,492,035	SHAPEIT2	Beagle	R^2	/	Sardinians HRP <i>vs.</i> 1KGP phase3 <i>vs.</i> 1KGP phase1
UK10K	2015	3781	European	4×	29,809,603	SHAPEIT2	IMPUTE2	r^2	UK	UK10K + 1KGP <i>vs.</i> UK10K <i>vs.</i> 1KGP
AGVP	2015	320	African	32.4×	24,574,727	SHAPEIT2	/	r^2	/	/
1KJPN	2015	1070	Japanese	7.4×	49,143,605	SHAPEIT2	IMPUTE2	R^2	Japanese	1KJPN + 1KGP <i>vs.</i> 1KJPN <i>vs.</i> 1KGP <i>vs.</i> 1KGP JPT
1KGP3	2015	2504	Multi-ethnic	4×	17,600,000	MaCH	Minimac	R^2	Multi-ethnic	1KGP phase3 <i>vs.</i> 1KGP phase1
HRC	2016	32,470	European	4×	39,635,008	SHAPEIT3	IMPUTE2	R^2	CEU	HRC <i>vs.</i> UK10K <i>vs.</i> 1KGP3
CAAPA	2016	642	African	35×	41,163,897	/	/	/	/	/
HELIC MANOLIS	2017	249	Cretan	4×	9,554,503	SHAPEIT2	IMPUTE2	r^2	/	/
EGCUT	2017	2244	Estonian	30×	16,536,512	SHAPEIT2	IMPUTE2	r^2 /number of well-imputed SNVs	/	/
AGRP	2017	265	Amish and Mennonite	30×	1,081,253	SHAPEIT2	IMPUTE2	R^2	AGRP	AGRP + 1KGP <i>vs.</i> HRC <i>vs.</i> AGRP <i>vs.</i> 1KGP
NIPT	2018	141,431	Chinese	0.06×	9,040,000	/	STITCH	r^2 /number of well-imputed SNVs	/	/
Mongolian HRP	2018	175	Mongolian	21.8×	16,526,134	SHAPEIT2	IMPUTE2	R^2	Mongolian	Mongolian HRP + 1KGP3 <i>vs.</i> Mongolian HRP <i>vs.</i> 1KGP3 East Asian + Mongolian HRP <i>vs.</i> 1KGP3
AJ HRP	2018	738	Ashkenazi Jewish	30×	26,400,000	SHAPEIT2	IMPUTE2	R^2 /genotype discordance	Ashkenazi Jewish	AJ HRP + 1KGP3 <i>vs.</i> AJ HRP <i>vs.</i> HRC <i>vs.</i> 1KGP3 <i>vs.</i> UK10K + 1KGP <i>vs.</i> UK10K <i>vs.</i> 1KGP
CONVERGE	2018	11,670	Chinese	1.7×	24,114,249	/	/	/	/	/
SG10K	2019	4810	Multi-ethnic	13.7×	98,273,706	Eagle2	Beagle4	r^2	Multi-ethnic	SG10K + 1KGP3 <i>vs.</i> SG10K <i>vs.</i> 1KGP3

(continued)

Table 1 (continued)

Name	Published time	Sample size	Ancestry	Depth	Total number of variants	Phasing algorithm	Imputation algorithm	Assessment criterion	Target population	Imputation performance measurement
NARD	2019	1779	Multi-ethnic	10×–20×	44,444,122	SHAPEIT3	Minimac3	R ² /genotype discordance	Korean	NARD + 1KGP3 <i>vs.</i> NARD <i>vs.</i> 1KGP3 <i>vs.</i> HRC
GASeP	2019	1654	Asian	30×	21,494,814	SHAPEIT2	Eagle2	/	/	/
UGR	2019	1978	African	4×	46,000,000	SHAPEIT2	IMPUTE2	/	/	/
CASPMI	2019	597	Chinese	25×–35×	/	/	/	/	/	/
PGG.Han	2020	114,783	Chinese	30×–80×	8,056,973	/	/	/	/	/
Korea1K	2020	1094	Korean	31×	38,800,000	SHAPEIT2	Minimac3	R ²	Korean	Korea1K + 1KGP3 <i>vs.</i> Korea1K <i>vs.</i> 1KGP3
TWB	2021	1445	Chinese	/	/	SHAPEIT2	SHAPEIT2	R ² /concordance	Chinese	TWB + 1KGP East Asian <i>vs.</i> TWB <i>vs.</i> 1KGP East Asian
TR	2021	773	Turkish	34×	45,981,721	SHAPEIT2	IMPUTE2	R ²	Balkan	TR + 1KGP3 <i>vs.</i> TR <i>vs.</i> 1KGP3
TOPMed	2021	97,256	Multi-ethnic	30×	308,107,085	Eagle2	Minimac4	R ²	UKB	QGP <i>vs.</i> 1KGP3 <i>vs.</i> HRC <i>vs.</i> CAAPA
QGP1	2021	6218	Qatari	30×	68,107,887	Eagle2	Minimac3	R ² /number of well-imputed SNVs/ <i>r</i> ²	Qatari	<i>vs.</i> HapMap2
NyuWa	2021	2902	Chinese	26.2×	19,256,267	SHAPEIT4	Minimac4	R ²	Multi-ethnic	NyuWa + 1KGP3 <i>vs.</i> NyuWa <i>vs.</i> TOPMed <i>vs.</i> GASeP <i>vs.</i> 1KGP3 <i>vs.</i> HRC
ChinaMAP	2021	10,588	Chinese	40.8×	59,010,860	/	/	/	/	/
AFAM	2021	2294	Sub-Saharan African	15×	52,500,000	SHAPEIT4	Minimac4	R ²	African Americans	TOPMed <i>vs.</i> AFAM <i>vs.</i> 1KGP3 <i>vs.</i> HRC
WBBC	2022	4535	Chinese	13.9×	81,498,995	SHAPEIT2	Minimac4	<i>r</i> ² /number of well-imputed SNVs/ non-reference genotype concordance rate	Multi-ethnic	WBBC + 1KGP East Asian <i>vs.</i> WBBC + 1KGP <i>vs.</i> WBBC <i>vs.</i> 1KGP3
UKB	2022	149,960	Multi-ethnic	32.5×	643,747,446	/	IMPUTE2	/	/	<i>vs.</i> CONVERGE
NSCLC	2022	6004	Chinese	30×	100,565,590	SHAPEIT4	Minimac4	<i>r</i> ² /number of well-imputed SNVs	Chinese	NSCLC <i>vs.</i> 1KGP3
GCATPanel	2022	690	European	30×	35,431,441	SHAPEIT4	IMPUTE2	<i>r</i> ²	Multi-ethnic	/
Expanded 1KGP	2022	3202	Multi-ethnic	34×	70,768,225	SHAPEIT2	IMPUTE2	R ²	Multi-ethnic	/
KRG pilot	2022	1490	Korean	29.0×	13,637,761	SHAPEIT4	Minimac4	R ² /number of well-imputed SNVs	Multi-ethnic	Differences across target populations
SABE	2022	1171	Brazilians	38.6×	/	SHAPEIT2	IMPUTE2	<i>r</i> ²	Brazilians	SABE + 1KGP3 <i>vs.</i> SABE <i>vs.</i> 1KGP3
CKB	2023	9964	Chinese	15.41×	129,743,542	Beagle5.2	Beagle5.2	R ² /number of well-imputed SNVs/ precision/sensitivity	Chinese	ChinaMAP <i>vs.</i> CKB <i>vs.</i> TOPMed <i>vs.</i> NyuWa <i>vs.</i> extended 1KGP
NARD2	2023	14,393	Multi-ethnic	/	/	Beagle5.0	Minimac4	R ² /number of well-imputed SNVs	KOR	NARD2 <i>vs.</i> NARD <i>vs.</i> TOPMed
FinnGen	2023	3775	Finnish	30×	/	Beagle4.1	Beagle4.1	<i>r</i> ²	/	/

(continued)

Table 1 (continued)

Name	Published time	Sample size	Ancestry	Depth	Total number of variants	Phasing algorithm	Imputation algorithm	Assessment criterion	Target population	Imputation performance measurement
1KTGP	2023	1064	Chinese	11.8×	28,200,000	SHAPEIT2	IMPUTE2	/	/	/
jMorp	2023	54,302	Japanese	/	/	/	/	/	/	/
MCPS10k	2023	9950	Mexican	30×	/	SHAPEIT4	IMPUTE5	R^2	MCPS individuals	MCPS 10k vs. TOPMed
Samoan panel	2023	1285	Samoan	/	/	Eagle2	Minimac4	r^2 /number of well-imputed SNVs	Samoans	1KGP3 + Samoan panel vs. TOPMed vs. 1KGP3
LVBMC	2023	502	Latvian	35.7×	/	Eagle2	Beagle4.1	Number of well-imputed SNVs	Latvians	/
BIGCS	2024	2245	Chinese	6.63×	47,239,473	Beagle4.0	Minimac3	R^2	Chinese	BIGCS + TOPMed vs. BIGCS vs. TOPMed vs. 1KGP3 vs. GAsP
ADSP	2024	16,564	Multi-ethnic	/	54,000,000	SHAPEIT4	Minimac3	R^2/r^2	Multi-ethnic	Differences across target populations
SEA HRP	2024	2550	Southeast Asian	/	113,851,450	Beagle5 vs. SHAPEIT4 and IMPUTE5		r^2 /number of well-imputed SNVs/ non-reference concordance rate	Orang Asli	SEA HRP vs. 1KGP3
Korea4K GEL	2024	3614	Korean	27.75×	26,210,741	SHAPEIT2	Minimac3	R^2	Korean	Korea4K vs. Korea1K
	2024	78,195	European	30×	342,573,817	SHAPEIT4	IMPUTE5	r^2	Multi-ethnic	Differences across target populations
SEAD	2024	11,067	Multi-ethnic	/	80,367,720	SHAPEIT2	Minimac4	r^2 /number of well-imputed SNVs/ non-reference concordance rate	Chinese	SEAD vs. SG10K vs. WBBC vs. 1KGP vs. GAsP
INDp	2024	217	Indonesian	30×	10,144,296	SHAPEIT2	IMPUTE2	r^2 /non-reference concordance rate	West Javanese	INDp + 1KGP East Asian vs. INDp vs. 1KGP East Asian
GMGD	2024	476	Chinese	5.5×	16,336,982	Beagle5.2	Beagle5.2	R^2	/	/

Note: R^2 , the squared Pearson's correlation coefficient (aggregated R^2 or dosage R^2) between the true genotypes and imputed genotype dosages; r^2 , the score generated by software without the true genotype (Rsq and INFO). In the "imputation performance measurement" column, we list the reference panels in order of their performance in the target population, from best to worst. The complete details can be found in Table S1. HRP, haplotype reference panel; MaCH, Markov Chain Haplotype algorithm; HapMap, International HapMap Project; 1KGP, 1000 Genomes Project; GoNL, Genome of the Netherlands Project; UK10K, 10,000 UK Genome Sequences; AGVP, African Genome Variation Project; 1KJPN, a reference panel of 1070 Japanese individuals; HRC, Haplotype Reference Consortium (release 1.1); CAAPA, Consortium on Asthma among African-ancestry Populations in the Americas; HELIC MANOLIS, HELIC Pomak collection and the MANOLIS Cohorts; HELIC, Hellenic Isolated Cohorts; MANOLIS, Myeloplasma; EGGUT, Estonian Biobank of the Estonian Genome Center, University of Tartu; AGRP, Anabaptist Genome Reference Panel; NIP1, Non-invasive prenatal testing in China; CONVERGE, the China, Oxford and Virginia Commonwealth University Experimental Research on Genetic Epidemiology; SG10K, whole-genome-sequence 10,000 Singaporeans; NARD, Northeast Asian Reference Database; GAsP, GenomeAsia 100K Project pilot phase; UGR, Uganda Genome Resource; CASPMI, the Chinese Academy of Sciences Precision Medicine Initiative project; PGG-Han, Han Chinese Genome Initiative; Korea1K, the Korean Genome Project; TWB, Taiwan Biobank; TR, Turkish Variome; TOPMed, Trans-Omics for Precision Medicine; QGP1, Qatar Genome Program Phase 1; NyuWa, the NyuWa Genome resource; ChinaMAP, China Metabolic Analytics Project; AFAM, African Americans reference panel; WBBC, Westlake Biobank for Chinese pilot project; GCAT|Panel, GCAT|Genomes for Life Cohort; NSCLC, Non-small cell lung cancer in China; KRKG, the Korean Reference Genome project; SABE, The Health, Well-being and Aging Study; CKB, China Kadoorie Biobank; UKB, UK Biobank; FinnGen, FinnGen project; 1KTGP, 1000 Tibetans Genomes Project; jMorp, The Tohoku Medical Megabank project; MCPS10k, the Mexico City Prospective Study; Samoan panel, Samoan-specific genotype reference panels; LVBMIC, the Latvian population-specific reference panel; BIGCS, Born in Guangzhou Cohort Study; ADSP, the Alzheimer's Disease Sequencing Project; SEA HRP, the Southeast Asian Specific Reference Panel; Korea4K, the Korean Genome Project; GEL, the Genomes England dataset; SEAD, the South and East Asian Reference Database reference panel; INDp, the Indonesian panel; GMGD, Guizhou Multi-Ethnic Genome Database; JPT, Japanese in Tokyo, Japan; CEU, Utah Residents with Northern and Western European Ancestry; SNV, single-nucleotide variant.

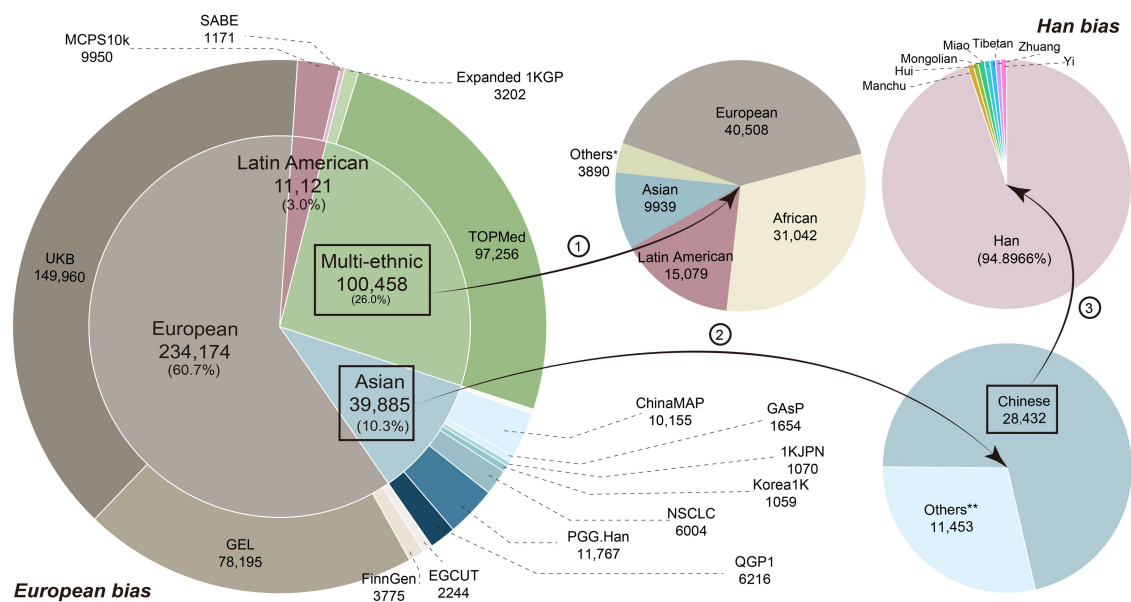


Figure 3 Number of samples and proportion of ancestry in the high-quality HRPs

The high-quality HRPs have a sequencing depth of $\geq 30\times$ and include more than 1000 individuals. In the left pie chart, the inner ring represents the distribution of samples by ancestral composition, whereas the outer ring reflects the distribution of HRPs. Importantly, the presence of duplicate individuals in the integrated reference panel was not excluded. Others* includes Samoan, Native American, multiple groups, or unknown ancestries. Others** includes Indian, Malaysian, South Korean, Pakistani, Mongolian, Papuan New Guinean, Indonesian, Philippine, Japanese, and Russian.

exhibited similar accuracy levels for ancient genomic data [109,110], with version 1 showing superior accuracy for data with $0.1\times$ coverage and variants with a $MAF > 0.02$. In conclusion, Beagle is recommended for phasing and imputation of array data using a unified HRP, while Glimpse remains the preferred tool for low-coverage WGS data in modern and ancient populations.

Performance and disparities of diverse HRPs

The size and number of high-quality HRPs have increased with advancements in large-scale human genomic cohorts, necessitating the development of a comprehensive reference panel for global or regional population use. This is essential for the medical and population genetics communities. We systematically summarized the effects of multiple factors, such as the target population, reference panel composition, and computational strategies, on imputation accuracy (Figure 1). Recently, several HRPs have been validated as highly effective for genetically similar populations. However, limitations in publicly available resources and challenges in selecting appropriate panels for specific target populations have impeded their broader application. The genotype imputation accuracy was evaluated via two main metrics: r^2 , derived without reference to true genotypes, and aggregated R^2 (also known as dosage R^2), which reflects the squared Pearson's correlation coefficients between imputed dosage and true genotypes [111]. Aggregated R^2 values were obtained by grouping markers according to the MAF. In addition to standard indices, concordance between predicted and observed genotypes, high- r^2 variants, and the density or coverage of high-quality imputed markers were incorporated as key metrics for evaluating the validation performance of the custom panels [25]. Bai et al. designed hundreds of customized reference panels with varying haplotype sizes and

diversity to investigate the relationship between imputation accuracy and panel composition [37]. It was demonstrated that simulated reference panels with differing diversity yielded varying imputation performance in Han Chinese and European populations. For Han Chinese, imputation accuracy plateaued when haplotype diversity within the reference panels was limited. An intriguing explanation was proposed via this work, attributing higher “diversity acceptability” in Western Eurasians, which was inconsistent with contributions from three ancestral sources: European hunter-gatherers, Near East farmers, and Steppe Yamnaya herders [112]. This finding indicates that ancient complex demographic events have an obvious influence on imputation performance. This systematic evaluation underscores the need for large-scale, high-quality reference panels representing underrepresented populations, such as Han Chinese. To address this gap, the Han Chinese-specific WBBC panel was developed and integrated with multi-ancestry sources to form the SEAD panel, enhancing imputation accuracy for Han Chinese and broader Asian populations [14,63].

Cahoon et al. evaluated the imputation performance of the large-scale, multi-ancestry, state-of-the-art TOPMed reference panel, which demonstrated higher simulation accuracy in European populations but did not yield similar improvements for underrepresented non-European populations. The r^2 estimates were shown to correlate with genetic distance to European populations and were overestimated in non-European populations [25]. Shi et al. further highlighted that estimated template switching rates in the meta-ancestry reference panel may contribute to inflated r^2 values [29]. However, this had a limited impact on evaluating the imputation performance of population-specific panels. Comparative analyses have been conducted via genomics-based reference panels from various population cohorts. Notable examples include TOPMed and AGRP in Africans [95], ChinaMAP

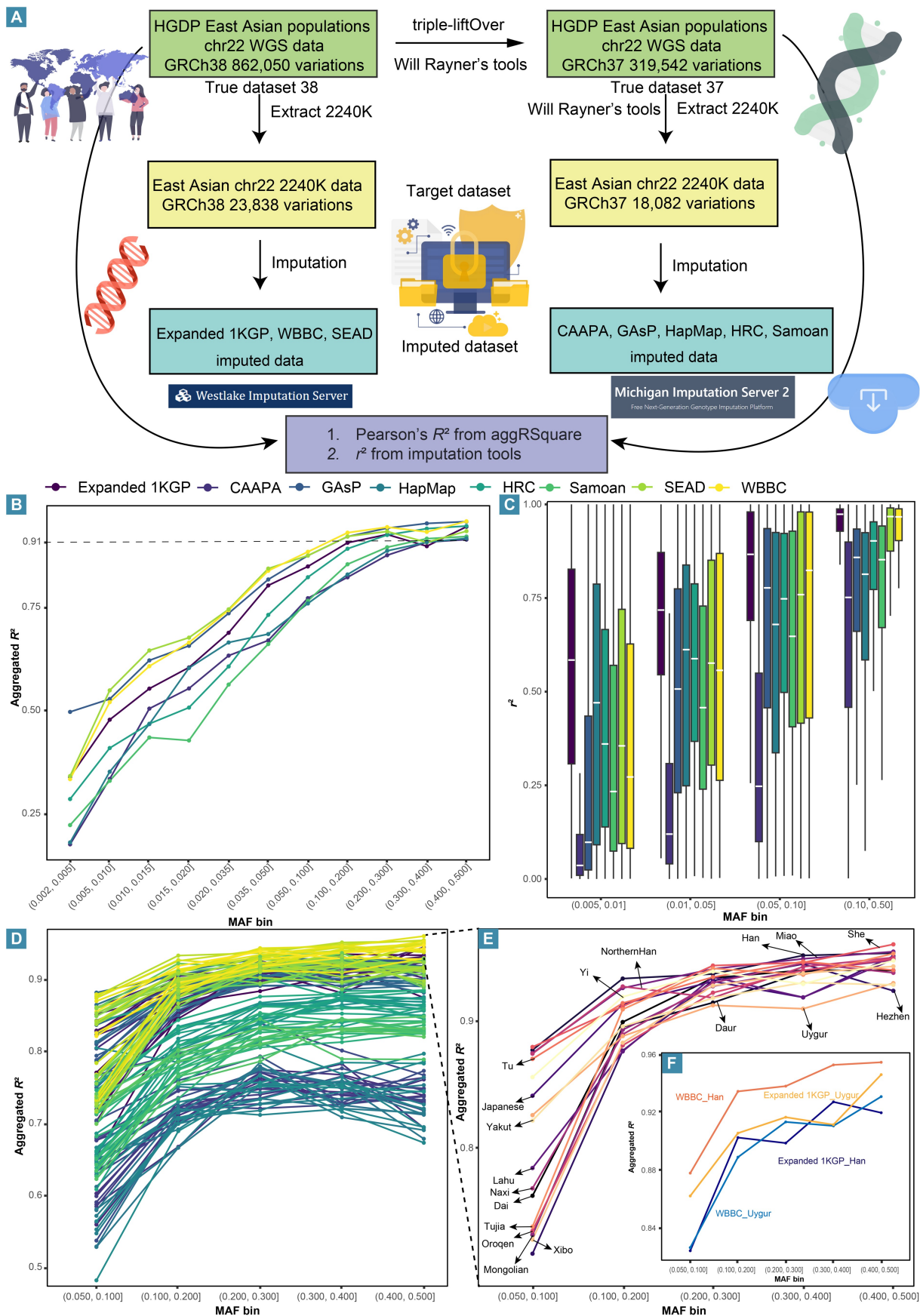


Figure 4 Evaluation of the performance of different reference panels
A. Workflow of the evaluation process. **B.** Line chart showing the aggregated R^2 values of the 8 HRP across different MAF bins. Different colored lines represent different HRPs. A higher aggregated R^2 indicates greater similarity between the imputed data and the true data. The dashed line indicates the threshold for high-quality imputation (aggregated $R^2 = 0.91$). **C.** Box plot showing the r^2 values of the 8 HRPs across different MAF bins. Different colored

and WBBC for Han Chinese [13,63], and NARD for Northeast Asians [22]. These findings highlight the importance of incorporating diverse genomic resources to improve imputation precision and health equity.

We addressed this issue by imputing Chinese genomic diversity using all available high-quality HRP. Our validation work underscored the importance of using population-matched reference panels for accurate genomic imputation. The imputation performance of publicly available HRPs was assessed via WGS data from 224 East Asian samples in the HGDP as the ground truth dataset. Quality control was conducted using VCFtools with parameters including “--maf 0.002”, “--max-missing 0.95”, “--minQ 30”, and “--hwe 1e-10”. Inconsistencies in reference genome versions among different HRPs were resolved using triple-liftOver to convert genomic coordinates to hg19 [113], and allele switches or strand flips were corrected by comparing allelic data against HRC data via Will Rayner’s tools. Imputation was halted if more than 10% of allele switches were detected during quality control (Figure 4A). Seventeen HRPs were accessed via imputation websites (Table S1), with imputed data obtained from eight usable HRPs after excluding HGDP samples already represented in the HRPs (Table 2).

The squared Pearson’s correlation coefficients (aggregated R^2) between the true genotypes and imputed genotype dosages were calculated to assess imputation accuracy. The results demonstrated a positive correlation between imputation accuracy and increasing MAF. For variants with MAFs ranging from 0.002 to 0.005, the GAsP HRP exhibited the highest accuracy (aggregated $R^2 = 0.497$). For variants with MAFs between 0.01 and 0.05, panels containing larger Asian sample sizes performed better, with SEAD achieving the highest accuracy (Figure 4B). Overall, the SEAD HRP was found to be the most suitable for the East Asian population in the HGDP. When r^2 statistics were evaluated, the expanded 1KGP HRP showed overall advantages (Figure 4C), underscoring the potential for imputation tool metrics to overestimate accuracy and lead to erroneous conclusions [25,29]. Benchmarking should thus avoid relying solely on r^2 or aggregated R^2 values. Further evaluation of imputation performance across 18 East Asian populations indicated that the WBBC panel exhibited the highest overall performance, although imputation accuracy varied across populations (Figure 4D). The WBBC panel showed better performance for Han populations but poorer results for Uyghur population (Figure 4E), likely due to the predominance of Han samples in the panel, emphasizing the importance of population matching in enhancing imputation accuracy. In contrast, the Uyghur population achieved better accuracy with the more diverse expanded 1KGP panel (Figure 4F), reflecting its mixed Eastern and Western ancestry, which was underrepresented in the WBBC HRP [114]. Overall, target populations that closely match those in the HRPs demonstrate superior imputation accuracy, particularly when the target population is homogeneous, or its genetic diversity is well-represented in large multi-ancestry reference panels.

Benefits and applications of genotype imputation

Genomic medicine and statistical genetics

Complex disorders related to Mendelian diseases and non-communicable diseases substantially contribute to the health-care burden, primarily driven by a combination of polygenic genetic architecture and diverse environmental risk factors. GWASs have elucidated the complex relationships between genotype and phenotype [115], identifying numerous SNPs linked to common complex diseases and traits; however, these variants explain only a fraction of the observed genetic variance or heritability. Rare SNP mutations or SVs may contribute to the missing heritability. Early applications of phasing and imputation innovations focused on GWAS-based medical discoveries, aiming to dissect heritability by increasing SNP density, enhancing genomic coverage in meta-analyses, and fine-mapping causal variants [11,36,38]. Low-coverage WGS and microarray-based genotyping have proven to be cost-effective and reliable approaches when combined with genotype imputation based on population-specific HRPs.

Recent large-scale GWASs or multi-ancestry GWAS meta-analyses via genotype imputation have demonstrated that combining new genomic resources with imputation strategies significantly expands the pool of available genetic variants, enhancing association signals, facilitating the identification of causal variants, and enabling the meta-analysis of multiple cohorts [5,11,14,39,40,97,110]. Rigorous quality control practices for SNP data in GWASs have been established to ensure data accuracy and reliability [116,117]. The UK Biobank (UKB) cohort, encompassing approximately 500,000 individuals, represents a crucial resource for well-powered GWASs on various quantitative traits [118], including body mass index [119], type 2 diabetes mellitus [120], and major depressive disorder [121]. Notably, Gaynor et al. highlighted that single-variant and gene-based association analyses using WES combined with imputed array data yield signal detection rates within 1% of those achieved with WGS data [122]. Yang et al. conducted single-variant testing and variant-set analysis on East Asian-based GWAS for hip and femoral neck bone mineral density traits, identifying the SEAD-imputed rare variant rs60103302 near *SNTG1* as associated with hip bone mineral density. Similarly, the GEL-imputed UKB GWAS revealed numerous rare trait-associated variants [39]. These newly identified causal variants offer valuable insights into the genetic architecture of human diseases or traits, allowing for the stratification of individuals at elevated risk for specific diseases. Polygenic risk score reconstruction can be performed to estimate genetic predispositions, incorporating individual variability in relevant quantitative traits [123]. This knowledge has contributed to improved patient outcomes through early detection, prevention, and targeted treatment strategies.

Population genetics

The reconstruction of genetic architecture in ethnolinguistically diverse groups in population genetics has proven crucial

Figure 4 Continued

boxes represent different HRPs. The r^2 values are derived from the imputation tool’s evaluation without the true genotype. **D.** Imputation performance of 18 East Asian populations from HGDP across 8 HRPs. Each line represents a population, and lines with the same color correspond to the same HRP. **E.** Imputation performance of 18 East Asian populations from HGDP using the WBBC HRP. Different colored lines represent different populations. **F.** Imputation performance of the Han and Uyghur populations using the WBBC HRP (population-specific) and the expanded 1KGP HRP (multi-ancestry). WGS, whole-genome sequencing; HGDP, Human Genome Diversity Project; MAF, minor allele frequency.

Table 2 Basic information of publicly available HRP and the number of variants obtained after imputing chromosome 22

Reference panel	Sample size	Reference genome	Ancestry	Number of imputed sites
Expanded 1KGP	3202	GRCh38	Multi-ethnic	680,128
CAAPA	642	GRCh37	African American	381,379
GAsP	1654	GRCh37	Asian	288,682
HRCr1.1	32,470	GRCh37	European	524,544
Samoan	1285	GRCh37	Oceanian	222,128
HapMap2	443	GRCh37	Multi-ethnic	33,805
WBBC	4535	GRCh38	Chinese	494,305
SEAD	11,067	GRCh38	Asian	1,228,248

Note: Expanded 1KGP, CAAPA, GAsP, HRCr1.1, Samoan, and HapMap2 were imputed at <https://imputationserver.sph.umich.edu/#!pages/home>, which uses Eagle2.4 and Minimac4. WBBC and SEAD were imputed at <https://imputationserver.westlake.edu.cn/index.html> via SHAPEIT2 and Minimac4.

for understanding human origins, evolution, migration, and admixture history [124]. However, limited access to resources and financial constraints have historically restricted sampling and genotyping efforts [125]. The application of HRP and imputation techniques has mitigated these limitations, facilitating the collection of more extensive genomic diversity data. Population-specific HRP-based genotype imputation has enabled the accurate reconstruction of missing genotype data, enhancing dataset completeness and providing higher-resolution data for downstream analyses of admixture modeling, biological adaptation, archaic introgression, and medical relevance interpretation [126]. Moreover, imputation increases marker density in integrative genomic datasets, thereby enabling more statistically robust analyses of mutation, recombination, natural selection, genetic drift, and gene flow. This process transforms individual genotype data into shared haplotypes derived from multiple reference populations, reducing or correcting false positives or negatives caused by population stratification [127]. As a result, fine-scale reconstructions of population evolutionary processes are possible, elucidating demographic histories among diverse populations and establishing a foundational framework for precision medicine systems [128]. Borda et al. evaluated haplotype discrepancies between genotyping-only and imputed datasets by calculating IBD segments. Their findings demonstrated that using exclusively imputed data for IBD analysis did not introduce bias for segments exceeding four centimorgans. Leveraging imputed data, this work provided a detailed depiction of fine-scale population structure, recent gene flow, and long-distance migration across Latin America [129]. Overall, the availability of high-quality genomic resources has dramatically advanced the study of human evolutionary history, facilitating the use of sophisticated algorithms and tools to analyze high-density DNA sequence variation and offering deeper insights into complex evolutionary processes [130].

Pharmacogenomics

Precision medicine aims to understand individualized disease progression and treatment responses. Genetic variations may influence susceptibility to specific diseases, either increasing or reducing risk, and they also affect responses to certain medications, a field known as pharmacogenomics [131]. Pharmacogenomics explores how genetic variation impacts individual drug responses, particularly in relation to absorption, distribution, metabolism, and excretion (ADME) processes. Given the increasing costs and slow pace of new drug discovery, there has been growing interest in drug repurposing, the practice of adapting existing drugs to treat both

common and rare diseases [132]. Studies have indicated that genes identified through GWAS as being associated with disease traits are more likely to encode druggable proteins than other genomic regions [133]. Genotype imputation, utilizing population-specific HRP, has been applied to estimate missing genomic data and predict individual drug responses, such as the *HLA-B**15:02 variant related to carbamazepine responses and other genes related to clopidogrel, peginterferon, and warfarin reactions [74]. This approach facilitates the development and optimization of personalized drug therapies, identifying potential drug targets and advancing the understanding of drug mechanisms [134].

Prenatal screening

The integration of ultra-low-depth sequencing data from non-invasive prenatal testing (NIPT) with genotype imputation leveraging population-specific HRP has increased the coverage, resolution, and overall comprehensiveness of genotype data for prenatal screening [60,135]. Improvements in genotype imputation performance have been observed with increasing NIPT sequencing depth and the expansion of reference panels [107]. Additionally, WGS on a fetus has been applied to identify potential conditions that may manifest during infancy or childhood, with the objectives of prevention, treatment, or preparation for the child’s arrival [136]. Imputation has proven valuable in genetic prediction and counseling, enriching the understanding of familial genetic histories, predicting disease risk, and facilitating personalized recommendations. Moreover, more precise association studies of complex diseases have been conducted for prenatal genetic diagnosis. Sequencing of fetal DNA allows parents to assess potential health risks, thereby supporting informed decisions regarding pregnancy continuation or termination [137]. The potential transition of genomic sequencing from a specialized test to a broadly accessible healthcare resource will necessitate the collection of high-quality outcome data from large cohorts and sustained efforts in monitoring screening program effectiveness [5]. Achieving these aims in an evidence-based, equitable, and sustainable manner remains essential for safeguarding the well-being of newborns.

Paleogenomics

The analysis of aDNA extracted from fossils and ancient hominin remains has significantly reshaped human genetic history. However, accurate genotype determination from ancient genomes remains challenging due to limited sequencing depth caused by DNA degradation and microbial contamination [138,139]. Consequently, pseudo-diploids are often employed in population genetic and medical studies,

leveraging genomic variations rather than true diploids. Despite these difficulties, evidence has shown that imputation using HRP constructed from modern humans with similar ancestry represents a reliable method for enhancing aDNA studies within the diploid-based research paradigm, even for populations with coverage depths as low as $0.5\times$ [8]. Martiniano et al. first used imputation and haplotype-based methods in aDNA research, imputing genome-wide diploid genotypes from 14 Middle Neolithic to Middle Bronze Age individuals from Portugal, which identified close relationships between local hunter-gatherers and later Iberian Neolithic populations [140]. Eske et al. analyzed over 5000 imputed ancient genomes from western Eurasia, uncovering genetic changes driven by admixture among ancient steppe pastoralists, agriculturalists, and hunter-gatherers [141]. Their findings highlighted a heightened genetic predisposition to multiple sclerosis, introduced by steppe pastoralists [142], possibly as a selective adaptation to livestock-borne pathogens. The height disparities observed between northern and southern Europeans were linked to varying levels of steppe ancestry. Neolithic farmer ancestry was enriched with risk alleles for emotional traits, whereas Western hunter-gatherer lineages presented a higher prevalence of alleles linked to diabetes and Alzheimer's disease. Additionally, alleles for lactase persistence emerged in Europe approximately 6000 years ago [143]. Ringbauer et al. developed the ancIBD tool, further advancing the use of imputed aDNA in ancient human history reconstruction [144]. The complete diploid ancient genome offers critical insights into the origins and spatiotemporal evolution of human diseases and traits, elucidating the influence of migration, admixture, and natural selection on disease emergence and development. These findings provide new perspectives on disease evolution and potential therapeutic strategies. With the exponential increase in sequenced ancient genomes, a greater representation of diverse ancestries and time periods is anticipated, offering a unique opportunity to expand HRP via ancient genomes and standardize imputation methodologies.

Forensic science and forensic investigative genetic genealogy

Degraded forensic samples often hinder the generation of high-quality, high-coverage genomes required for forensic applications [145]. Genotype imputation utilizing population-specific HRP has been widely applied in forensic genetic genealogy. In specific criminal cases, degraded DNA biomaterials may limit data availability. The imputation of missing loci enhances the comprehensiveness of genotype profiles, improving parentage testing, individual identification, forensic phenotype prediction, biogeographic ancestry inference, and phylogenetic reconstruction. This method enhances the accuracy of DNA comparisons, aiding in the identification and familial relationship determination of criminal suspects [146]. A notable example underscoring this approach's impact is the Golden State Killer case. Despite raising privacy concerns and sparking intense debate within the academic community [147,148], forensic genealogy has demonstrated significant potential in resolving cold cases and bringing perpetrators to justice [149]. Its utility extends to identifying missing persons, locating relatives, and identifying human remains.

Population-specific, high-quality HRP have shown tremendous value across diverse fields, including human

evolutionary studies, disease genomics, pharmacogenomics, prenatal screening, paleogenomics, and forensic science. Continual updates to reference panels and imputation algorithms remain critical to maintaining the precision of imputed genotypes and advancing both research and clinical applications.

Challenges and perspectives

Inclusion of underrepresented ethnolinguistically diverse populations in the HRP

The application of high-quality HRP presents significant opportunities and challenges in the evolutionary genomics and pangenomics eras (Figure 5). As precision medicine, population genetics, and evolutionary biology advance, HRP have become critical tools for improving genotype imputation, ancestry inference, and disease association studies. However, a key challenge lies in the underrepresentation of ethnolinguistically diverse populations in existing human genomic cohorts and the summarized panels. Large-scale, medical-driven human genomics data have primarily been collected from participants in metropolitan areas, often failing to capture the diversity of anthropologically informed local populations [25,39]. Consequently, current panels exhibit biases toward data from a limited number of ancestries, reducing imputation accuracy in underrepresented groups [25]. This shortfall hinders the comprehensive characterization of global genetic diversity, exacerbating health disparities and undermining equitable and personalized healthcare initiatives.

Historically, genomic studies have focused predominantly on populations of European or North American ancestry, perpetuating so-called European bias [150,151]. Our findings demonstrate that 60.7% of high-quality HRP derived from European descendants inadequately represent non-European genetic variants, resulting in imprecise imputation. To address this issue, regional genome projects have been launched to construct population-specific HRP, increasing diversity to some extent. However, the lack of internationally standardized sequencing and data-processing guidelines has introduced inconsistencies, complicating joint analyses and masking true signals. The Global Alliance for Genomics and Health has sought to address these challenges by promoting collaboration and interoperability, aiming to develop a standardized framework for capturing human genetic diversity.

High-quality genomic infrastructure and bioinformatics facilitating data sharing

Generating high-quality, population-specific HRP requires large-scale, high-resolution sequencing data, which remains costly and resource-intensive. The need for robust data curation, quality control, and significant computational resources further complicates the development and maintenance of these panels. Key foundational elements include biorepositories and computing infrastructure, funding strategies, capacity building, global consortia cooperation, and stakeholder will from research and funders. Data sharing and equity have been longstanding points of debate in genomic research and reference panel merging and optimization, as the availability of genomic data plays a pivotal role in advancing precision medicine and personalized treatments [152]. Currently, the common formats for sharing raw genomic data can be categorized into two types: intra-federation

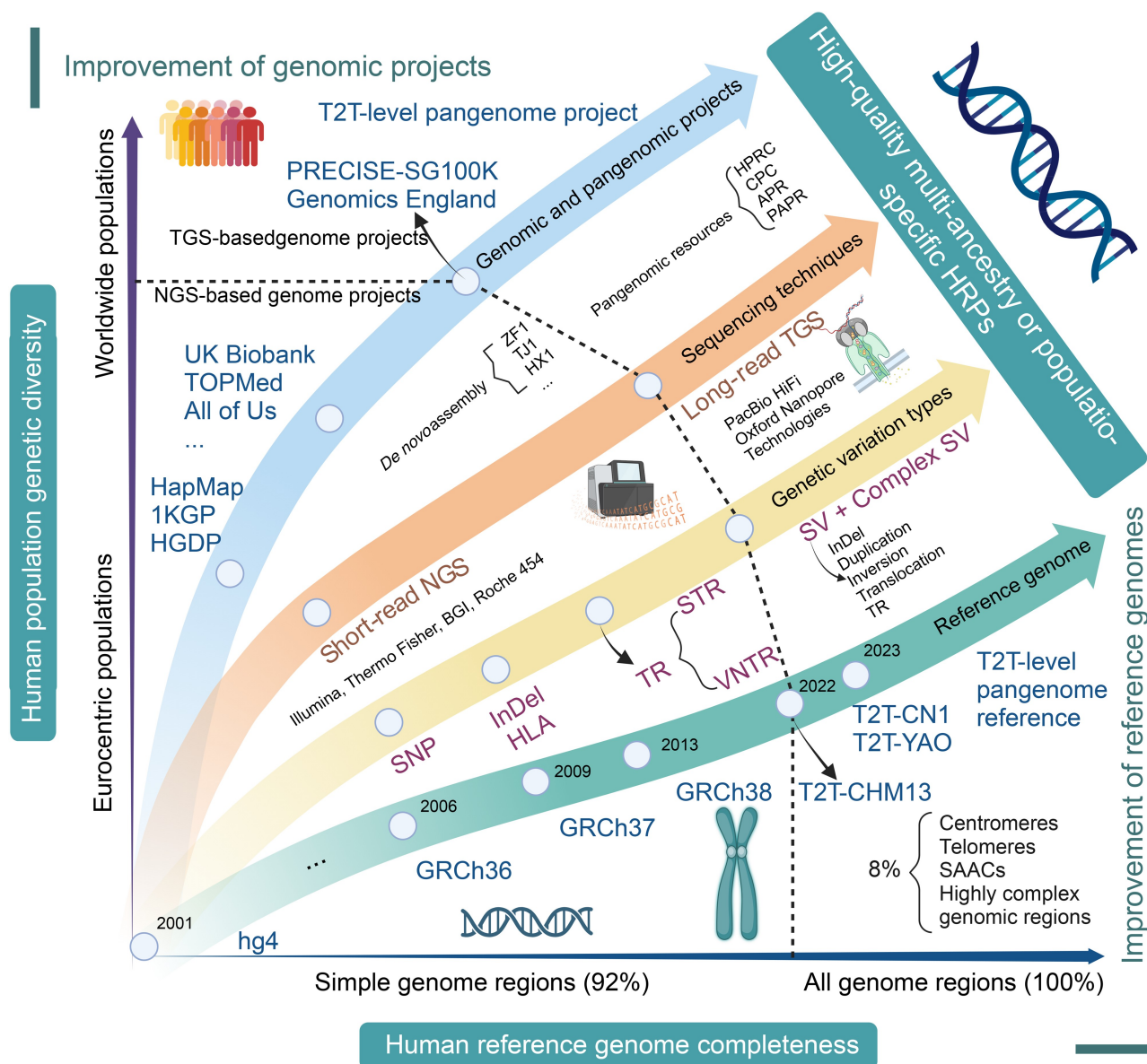


Figure 5 Development and enhancement of the human reference genome, genome projects, and HRP in the genomic and pangenomic eras

We summarized the “2-3-4-5” rule to present all the past, present, and future of the HRP. “2” denotes the two dimensions: human reference genome completeness and human population genetic diversity. The X-axis represents the improvement in the completeness of the human reference genome, from 92% to 100%, while the Y-axis represents the enhancement of human population diversity in genomic research, transitioning from being centered on European populations to including all populations. “3” highlights three key goals of the reference genome, human genomic projects, and HRP: the enhancement of human reference panels, including the creation of a perfect T2T-level pangenome reference; the improvement of genomic projects; and the establishment of high-quality multi-ancestry or population-specific HRP. The four (“4”) arrows in different colors represent the development of genomic and pangenomic projects, the advancement of sequencing techniques, the increase in the types of genetic variations discovered and analyzed, and the updates to human reference genome versions. The five (“5”) major blocks represent the transition from genomic projects based on NGS technologies to those based on TGS technologies; the shift from single-sample *de novo* assembly to the creation of multi-sample human pangenomic resources; the evolution from short-read NGS platforms to long-read TGS platforms; the expansion from simple SNPs to complex SVs; and the progression from reference genomes with gaps to complete, gap-free reference genomes. Ultimately, by integrating these advancements, we can construct a multi-ancestral or population-specific high-quality HRP. T2T, telomere-to-telomere; ZF1, *de novo* assembly of a Tibetan genome; TJ1, *de novo* assembly of a Tujia genome; HX1, *de novo* assembly of a Han Chinese genome; HPRC, Human Pangenome Reference Consortium; CPC, Chinese Pangenome Consortium; APR, Arab Pangenome Reference; PAPP, Pacific Ancestry Pangenome Reference; NGS, next-generation sequencing; TGS, third-generation sequencing; InDel, insertion and deletion; HLA, human leukocyte antigen; TR, tandem repeat; STR, short tandem repeat; VNTR, variable number of tandem repeats; SV, structural variation; SAAC, short arm of acrocentric chromosome.

sharing and cloud-based sharing. For example, data from the UKB have attracted global scientific interest, with its sharing through cloud-based platforms, significantly advancing genomic research in the UK. Similarly, initiatives such as TOPMed and All of Us have expanded data sharing among their members. The International Hundred Thousand Plus

Cohort Consortium has brought together over 100 cohorts from 43 countries, encompassing more than 50 million participants [153].

However, unrestricted data sharing is not universally appropriate. In 2016, representatives from academia, industry, funding agencies, and scholarly publishers established the

FAIR Data Principles, emphasizing data that are findable, accessible, interoperable, and reusable [154]. In 2019, the Global Indigenous Data Alliance introduced the CARE Principles, focusing on collective benefit, authority to control, responsibility, and ethics in indigenous data governance [155]. Genomic data sharing must adhere to both FAIR and CARE principles to maximize value while addressing ethical, legal, and social implications. Advanced algorithms are essential to enable secure sharing, evaluation, and selection of genomic data. For example, the Recombine and Share Haplotypes method generates synthetic HRP by simulating hypothetical descendants of reference panel samples after a user-defined number of meiosis events [156]. Meta-imputation offers another approach, leveraging multiple imputation servers based on different reference panels to improve the accuracy of target sample imputation [114]. These strategies partially integrate the benefits of multiple HRPs while addressing privacy concerns.

Although numerous HRPs have been generated, only a limited number are publicly accessible (Figure 2). Enhanced data accessibility could facilitate reciprocal imputation approaches, integrating the strengths of multiple HRPs. Achieving a balance between data utility and privacy preservation requires collaboration among the technical, regulatory, and ethical communities. Experts in computer security, genetics, computer science, ethics, and privacy law must work together to establish policies that support efficient, privacy-respecting genomic data sharing [157].

T2T-level HRPs and multi-variant integration

The advent of large-scale genomic datasets [158,159], the completion of the T2T human genome sequence [160,161], and the development of human pangenome projects [162,163] have underscored the importance of comprehensively capturing human genetic diversity and elucidating the functional roles of population-specific sequences and variations (Figure 5). The T2T reference genome and pangenome-based third-generation sequencing paradigm have provided geneticists unprecedented opportunities to develop high-quality HRPs with multiple ancestries and genetic variant types. In the T2T and pangenomic eras, integrating complex SVs and rare haplotypes into HRPs introduces additional complexity, requiring sophisticated algorithms and analytical approaches. Technological advancements, particularly in long-read sequencing and phased genome assembly, have enabled significant progress in refining HRPs. Improved methodologies for resolving complex haplotypes and rare alleles are essential for enhancing panel specificity and accuracy. Collaborative efforts across global consortia and ongoing algorithmic innovations will be key to addressing existing limitations.

Empirical studies on the impact of rare variants on complex traits and diseases remain limited, largely due to challenges in genotype imputation for rare variants, compounded by European bias, small sample sizes, and inadequate sequencing depth [38]. Previous HRPs predominantly included SNPs but rarely incorporated InDels or SVs, largely due to challenges in high-quality SV calling and phasing. This absence has severely constrained downstream analyses, as complex variations undetected by SNPs likely account for a significant portion of unexplained heritability [164]. To address these limitations, the expanded 1KGP has improved HRPs by integrating high-confidence InDels and SVs [42].

However, short-read WGS methods are limited in detecting SVs in highly repetitive genomic regions compared with long-read sequencing approaches. Recent efforts, such as SNP/short tandem repeat (STR) combined panels, have enabled genome-wide STR imputation [165,166]. However, current imputation tools remain optimized solely for SNPs, highlighting the need for further improvements to accommodate other variant types.

The future objective for HRPs is to establish a pangenome HRP based on the T2T assembly. Such advancements will enable large-scale cohort studies in precision medicine, encompassing targeted prevention, treatment, and diagnosis. High-quality HRPs hold transformative potential across diverse fields, including paleogenomics, forensic science, pharmacogenomics, and clinical diagnostics. Expanding and diversifying these panels while maintaining stringent quality standards is essential to maximize their impact. Addressing these challenges will foster more equitable research outcomes, drive biomedical innovation, and deepen our understanding of human genetic diversity. Despite these challenges, haplotype-resolved reference panels offer significant opportunities to advance precision medicine, enhance the detection of disease-associated variants, and provide deeper insights into population histories. Future efforts should prioritize emerging technologies, such as long-read sequencing and artificial intelligence-driven bioinformatics tools, to overcome existing limitations. Collaboration across scientific, medical, and indigenous communities will be crucial to ensuring the quality, inclusivity, and ethical application of HRPs in the genomic and pangenomic landscapes.

The trade-off between multi-ancestry integrative HRPs and population-specific HRPs

Multi-ancestry integrative HRPs incorporate high-quality genomic data from individuals of diverse genetic backgrounds, often spanning multiple continents or genetically distinct subgroups within a continent, as exemplified by the state-of-the-art TOPMed reference panel [1]. TOPMed-based imputed datasets identify more variant sites and high-impact consequence variants than those generated from other panel-imputed datasets, highlighting the potential of multi-ancestral reference panels to provide more comprehensive genotypic information in the context of mixed and heterogeneous target populations with different ancestral sources. In contrast, population-specific panels focus on genomic data from relatively homogeneous populations with well-characterized genetic backgrounds, such as the NyuWa, ChinaMAP, and WBBC panels, which are optimized for imputation within Han Chinese populations [4,21,63].

Several key factors, including haplotype diversity, panel size, genotype imputation accuracy, preservation of the LD structure, privacy and data sharing considerations, adaptability, and hardware requirements for big data processing, influence the choice between these panel types. Multi-ancestry panels capture a broader spectrum of genetic variation, making them crucial for studies involving globally diverse populations. By integrating data from multiple populations, these panels offer a more comprehensive view of haplotype diversity. However, population-specific panels, which focus on a single group, typically achieve higher resolution and more precise genetic insights for that population. They present greater genotype imputation accuracy, especially for rare variants, as they are optimized to reflect unique genetic variants

and LD patterns. In contrast, multi-ancestry panels, while accommodating a range of heterogeneous populations, may suffer from reduced accuracy for specific groups due to the need to balance genetic variation across heterogeneous datasets. These panels also face challenges in preserving LD structure, as LD patterns differ significantly between populations. Population-specific panels, by concentrating on a single group, better maintain the characteristic LD structure, thereby increasing the resolution of population-specific analyses.

From a privacy perspective, multi-ancestry panels offer enhanced protection by blending haplotypes across diverse populations, reducing the risk of directly linking genotypes to phenotypes and facilitating broader data sharing while safeguarding individual privacy. In contrast, population-specific panels may require additional privacy measures due to their targeted design. Despite this, population-specific panels provide fine-grained insights into genetic studies within a specific population, making them ideal for tailored research. Multi-ancestry panels, on the other hand, are better suited for cross-population analyses and global studies, offering broader applicability. Both multi-ancestry and population-specific HRP present unique advantages and challenges. The selection of an appropriate panel should align with the imputed objectives, the genetic characteristics of the target population, and privacy requirements. A flexible approach, leveraging one or both panel types based on research needs, may maximize the utility of HRP across diverse genetic studies.

CRedit author statement

Qingxin Yang: Formal analysis, Methodology, Writing – original draft, Visualization, Writing – review & editing. **Yuntao Sun:** Visualization. **Shuhan Duan:** Investigation. **Shengjie Nie:** Project administration. **Chao Liu:** Project administration, Supervision. **Hong Deng:** Funding acquisition, Writing – review & editing. **Mengge Wang:** Funding acquisition, Supervision, Writing – review & editing. **Guanglin He:** Conceptualization, Writing – review & editing, Resources. All authors have read and approved the final manuscript.

Competing interests

The authors have declared no competing interests.

Supplementary material

Supplementary material is available at *Genomics, Proteomics & Bioinformatics* online (<https://doi.org/10.1093/gpbjnl/qzaf022>).

Acknowledgments

This study was supported by the National Natural Science Foundation of China (Grant Nos. 82402203 and 82202078), the Major Project of the National Social Science Foundation of China (Grant No. 23&ZD203), the Open Project of the Key Laboratory of Forensic Genetics of the Ministry of Public Security (Grant Nos. 2022FGKFKT05 and 2024FGKFKT02), the Center for Archaeological Science of Sichuan University (Grant Nos. 23SASA01 and 24SASB03), the 1.3.5 Project for Disciplines of Excellence, West China Hospital, Sichuan University (Grant No. ZYJC20002), and

the Sichuan Science and Technology Program (Grant No. 2024NSFSC1518), China. We acknowledge Grammarly (<https://app.grammarly.com/>) for its invaluable contribution to refining the language and enhancing the readability of this manuscript and BioRender (<https://app.biorender.com/>) for its assistance in creating the figures. Thanks to the China National Supercomputing Center in Chengdu for providing storage for sequencing data and computational resources.

ORCID

0009-0007-5099-6925 (Qingxin Yang)
0009-0007-8419-1420 (Yuntao Sun)
0009-0005-7214-581X (Shuhan Duan)
0000-0001-6414-6621 (Shengjie Nie)
0000-0001-5633-3929 (Chao Liu)
0009-0003-9889-6465 (Hong Deng)
0000-0002-3673-1855 (Mengge Wang)
0000-0002-6614-5267 (Guanglin He)

References

- [1] Taliun D, Harris DN, Kessler MD, Carlson J, Szpiech ZA, Torres R, et al. Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *Nature* 2021;590:290–9.
- [2] McCarthy S, Das S, Kretschmar W, Delaneau O, Wood AR, Teumer A, et al. A reference panel of 64,976 haplotypes for genotype imputation. *Nat Genet* 2016;48:1279–83.
- [3] The 1000 Genomes Project Consortium, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, et al. A global reference for human genetic variation. *Nature* 2015;526:68–74.
- [4] Zhang P, Luo H, Li Y, Wang Y, Wang J, Zheng Y, et al. NyuWa Genome resource: a deep whole-genome sequencing-based variation profile and reference panel for the Chinese population. *Cell Rep* 2021;37:110017.
- [5] Stark Z, Scott RH. Genomic newborn screening for rare diseases. *Nat Rev Genet* 2023;24:755–66.
- [6] He G, Yao H, Duan S, Luo L, Sun Q, Tang R, et al. 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* 2025;68:914–33.
- [7] Li X, Wang M, Su H, Duan S, Sun Y, Chen H, et al. Evolutionary history and biological adaptation of Han Chinese people on the Mongolian Plateau. *hLife* 2024;2:296–313.
- [8] Sousa da Mota B, Rubinacci S, Cruz Dávalos DI, Amorim CEG, Sikora M, Johannsen NN, et al. Imputation of ancient human genomes. *Nat Commun* 2023;14:3660.
- [9] Cook S, Choi W, Lim H, Luo Y, Kim K, Jia X, et al. Accurate imputation of human leukocyte antigens with CookHLA. *Nat Commun* 2021;12:1264.
- [10] Sakaue S, Gurajala S, Curtis M, Luo Y, Choi W, Ishigaki K, et al. Tutorial: a statistical genetics guide to identifying HLA alleles driving complex disease. *Nat Protoc* 2023;18:2625–41.
- [11] Marchini J, Howie B. Genotype imputation for genome-wide association studies. *Nat Rev Genet* 2010;11:499–511.
- [12] Choi J, Kim S, Kim J, Son HY, Yoo SK, Kim CU, et al. A whole-genome reference panel of 14,393 individuals for East Asian populations accelerates discovery of rare functional variants. *Sci Adv* 2023;9:ead6319.
- [13] Cao Y, Li L, Xu M, Feng Z, Sun X, Lu J, et al. The ChinaMAP analytics of deep whole genome sequences in 10,588 individuals. *Cell Res* 2020;30:717–31.
- [14] Yang MY, Zhong JD, Li X, Tian G, Bai WY, Fang YH, et al. SEAD reference panel with 22,134 haplotypes boosts rare variant imputation and genome-wide association analysis in Asian populations. *Nat Commun* 2024;15:10839.

- [15] Loh PR, Palamara PF, Price AL. Fast and accurate long-range phasing in a UK Biobank cohort. *Nat Genet* 2016;48:811–6.
- [16] O’Connell J, Sharp K, Shrine N, Wain L, Hall I, Tobin M, et al. Haplotype estimation for biobank-scale data sets. *Nat Genet* 2016;48:817–20.
- [17] Loh PR, Danecek P, Palamara PF, Fuchsberger C, Reshef YA, Finucane HK, et al. Reference-based phasing using the Haplotype Reference Consortium panel. *Nat Genet* 2016;48:1443–8.
- [18] Das S, Forer L, Schönherr S, Sidore C, Locke AE, Kwong A, et al. Next-generation genotype imputation service and methods. *Nat Genet* 2016;48:1284–7.
- [19] Li Y, Willer CJ, Ding J, Scheet P, Abecasis GR. MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genet Epidemiol* 2010;34:816–34.
- [20] Huang L, Li Y, Singleton AB, Hardy JA, Abecasis G, Rosenberg NA, et al. Genotype-imputation accuracy across worldwide human populations. *Am J Hum Genet* 2009;84:235–50.
- [21] Li L, Huang P, Sun X, Wang S, Xu M, Liu S, et al. The ChinaMAP reference panel for the accurate genotype imputation in Chinese populations. *Cell Res* 2021;31:1308–10.
- [22] Yoo SK, Kim CU, Kim HL, Kim S, Shin JY, Kim N, et al. NARD: whole-genome reference panel of 1779 Northeast Asians improves imputation accuracy of rare and low-frequency variants. *Genome Med* 2019;11:64.
- [23] Du Z, Ma L, Qu H, Chen W, Zhang B, Lu X, et al. Whole genome analyses of Chinese population and *de novo* assembly of a northern Han genome. *Genomics Proteomics Bioinformatics* 2019;17:229–47.
- [24] Pasaniuc B, Rohland N, McLaren PJ, Garimella K, Zaitlen N, Li H, et al. Extremely low-coverage sequencing and imputation increases power for genome-wide association studies. *Nat Genet* 2012;44:631–5.
- [25] Cahoon JL, Rui X, Tang E, Simons C, Langie J, Chen M, et al. Imputation accuracy across global human populations. *Am J Hum Genet* 2024;111:979–89.
- [26] Si Y, Vanderwerff B, Zöllner S. Why are rare variants hard to impute? Coalescent models reveal theoretical limits in existing algorithms. *Genetics* 2021;217:iyab011.
- [27] The International HapMap Consortium, Frazer KA, Ballinger DG, Cox DR, Hinds DA, Stuve LL, et al. A second generation human haplotype map of over 3.1 million SNPs. *Nature* 2007;449:851–61.
- [28] He G, Wang M, Luo L, Sun Q, Yuan H, Lv H, et al. Population genomics of Central Asian peoples unveil ancient Trans-Eurasian genetic admixture and cultural exchanges. *hLife* 2024;2:554–62.
- [29] Shi M, Tanikawa C, Munter HM, Akiyama M, Koyama S, Tomizuka K, et al. Genotype imputation accuracy and the quality metrics of the minor ancestry in multi-ancestry reference panels. *Brief Bioinform* 2023;25:bbad509.
- [30] Lin Y, Liu L, Yang S, Li Y, Lin D, Zhang X, et al. Genotype imputation for Han Chinese population using Haplotype Reference Consortium as reference. *Hum Genet* 2018;137:431–6.
- [31] Gurdasani D, Carstensen T, Tekola-Ayele F, Pagani L, Tachmazidou I, Hatzikotoulas K, et al. The African Genome Variation Project shapes medical genetics in Africa. *Nature* 2015;517:327–32.
- [32] Zheng W, He Y, Guo Y, Yue T, Zhang H, Li J, et al. Large-scale genome sequencing redefines the genetic footprints of high-altitude adaptation in Tibetans. *Genome Biol* 2023;24:73.
- [33] Popejoy AB, Fullerton SM. Genomics is failing on diversity. *Nature* 2016;538:161–4.
- [34] Jostins L, Morley KI, Barrett JC. Imputation of low-frequency variants using the HapMap3 benefits from large, diverse reference sets. *Eur J Hum Genet* 2011;19:662–6.
- [35] Bomba L, Walter K, Soranzo N. The impact of rare and low-frequency genetic variants in common disease. *Genome Biol* 2017;18:77.
- [36] Das S, Abecasis GR, Browning BL. Genotype imputation from large reference panels. *Annu Rev Genomics Hum Genet* 2018;19:73–96.
- [37] Bai WY, Zhu XW, Cong PK, Zhang XJ, Richards JB, Zheng HF. Genotype imputation and reference panel: a systematic evaluation on haplotype size and diversity. *Brief Bioinform* 2019;21:bbz108.
- [38] Hoffmann TJ, Witte JS. Strategies for imputing and analyzing rare variants in association studies. *Trends Genet* 2015;31:556–63.
- [39] Shi S, Rubinacci S, Hu S, Moutsianas L, Stuckey A, Need AC, et al. A Genomics England haplotype reference panel and imputation of UK Biobank. *Nat Genet* 2024;56:1800–3.
- [40] All of Us Research Program Genomics Investigators. Genomic data in the All of Us Research Program. *Nature* 2024;627:340–6.
- [41] The International HapMap Consortium. A haplotype map of the human genome. *Nature* 2005;437:1299–320.
- [42] Byrka-Bishop M, Evani US, Zhao X, Basile AO, Abel HJ, Regier AA, et al. High-coverage whole-genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios. *Cell* 2022;185:3426–40.e19.
- [43] The 1000 Genomes Project Consortium, Abecasis GR, Altshuler D, Auton A, Brooks LD, Durbin RM, et al. A map of human genome variation from population-scale sequencing. *Nature* 2010;467:1061–73.
- [44] Delaneau O, Marchini J, The 1000 Genomes Project Consortium. Integrating sequence and array data to create an improved 1000 Genomes Project haplotype reference panel. *Nat Commun* 2014;5:3934.
- [45] Xue Y, Mezzavilla M, Haber M, McCarthy S, Chen Y, Narasimhan V, et al. Enrichment of low-frequency functional variants revealed by whole-genome sequencing of multiple isolated European populations. *Nat Commun* 2017;8:15927.
- [46] Genome of the Netherlands Consortium. Whole-genome sequence variation, population structure and demographic history of the Dutch population. *Nat Genet* 2014;46:818–25.
- [47] UK10K Consortium, Walter K, Min JL, Huang J, Crooks L, Memari Y, et al. The UK10K project identifies rare variants in health and disease. *Nature* 2015;526:82–90.
- [48] Sidore C, Busonero F, Maschio A, Porcu E, Naitza S, Zoledziwska M, et al. Genome sequencing elucidates Sardinian genetic architecture and augments association analyses for lipid and blood inflammatory markers. *Nat Genet* 2015;47:1272–81.
- [49] Gudbjartsson DF, Helgason H, Gudjonsson SA, Zink F, Oddsson A, Gylfason A, et al. Large-scale whole-genome sequencing of the Icelandic population. *Nat Genet* 2015;47:435–44.
- [50] Pagani L, Lawson DJ, Jagoda E, Mörseburg A, Eriksson A, Mitt M, et al. Genomic analyses inform on migration events during the peopling of Eurasia. *Nature* 2016;538:238–42.
- [51] Southam L, Gilly A, Süveges D, Farmaki AE, Schwartzentruber J, Tachmazidou I, et al. Whole genome sequencing and imputation in isolated populations identify genetic associations with medically-relevant complex traits. *Nat Commun* 2017;8:15606.
- [52] Mitt M, Kals M, Pärn K, Gabriel SB, Lander ES, Palotie A, et al. Improved imputation accuracy of rare and low-frequency variants using population-specific high-coverage WGS-based imputation reference panel. *Eur J Hum Genet* 2017;25:869–76.
- [53] Lencz T, Yu J, Palmer C, Carmi S, Ben-Avraham D, Barzilai N, et al. High-depth whole genome sequencing of an Ashkenazi Jewish reference panel: enhancing sensitivity, accuracy, and imputation. *Hum Genet* 2018;137:343–55.
- [54] Valls-Margarit J, Galván-Femenía I, Matías-Sánchez D, Blay N, Puiggròs M, Carreras A, et al. GCAT|Panel, a comprehensive structural variant haplotype map of the Iberian population from high-coverage whole-genome sequencing. *Nucleic Acids Res* 2022;50:2464–79.
- [55] Halldorsson BV, Eggertsson HP, Moore KHS, Hauswedell H, Eiriksson O, Ulfarsson MO, et al. The sequences of 150,119 genomes in the UK Biobank. *Nature* 2022;607:732–40.

- [56] Banasik K, Møller PL, Techlo TR, Holm PC, Walters GB, Ingason A, et al. DanMAC5: a browser of aggregated sequence variants from 8,671 whole genome sequenced Danish individuals. *BMC Genom Data* 2023;24:30.
- [57] Kurki MI, Karjalainen J, Palta P, Sipilä TP, Kristiansson K, Donner KM, et al. FinnGen provides genetic insights from a well-phenotyped isolated population. *Nature* 2023; 613:508–18.
- [58] Reščenko R, Brīvība M, Atava I, Rovīte V, Pečulis R, Silamīkēlis I, et al. Whole-genome sequencing of 502 individuals from Latvia: the first step towards a population-specific reference of genetic variation. *Int J Mol Sci* 2023;24:15345.
- [59] Sirugo G, Williams SM, Tishkoff SA. The missing diversity in human genetic studies. *Cell* 2019;177:26–31.
- [60] Liu S, Huang S, Chen F, Zhao L, Yuan Y, Francis SS, et al. Genomic analyses from non-invasive prenatal testing reveal genetic associations, patterns of viral infections, and Chinese population history. *Cell* 2018;175:347–59.e14.
- [61] Iglesias AI, Mishra A, Vitart V, Bykhovskaya Y, Höhn R, Springelkamp H, et al. Cross-ancestry genome-wide association analysis of corneal thickness strengthens link between complex and Mendelian eye diseases. *Nat Commun* 2018;9:1864.
- [62] Wei CY, Yang JH, Yeh EC, Tsai MF, Kao HJ, Lo CZ, et al. Genetic profiles of 103,106 individuals in the Taiwan Biobank provide insights into the health and history of Han Chinese. *NPJ Genom Med* 2021;6:10.
- [63] Cong PK, Bai WY, Li JC, Yang MY, Khederzadeh S, Gai SR, et al. Genomic analyses of 10,376 individuals in the Westlake BioBank for Chinese (WBBC) pilot project. *Nat Commun* 2022; 13:2939.
- [64] Wang C, Dai J, Qin N, Fan J, Ma H, Chen C, et al. Analyses of rare predisposing variants of lung cancer in 6,004 whole genomes in Chinese. *Cancer Cell* 2022;40:1223–39.e6.
- [65] Gao Y, Zhang C, Yuan L, Ling Y, Wang X, Liu C, et al. PGG. Han: the Han Chinese genome database and analysis platform. *Nucleic Acids Res* 2020;48:D971–6.
- [66] Jeon S, Bhak Y, Choi Y, Jeon Y, Kim S, Jang J, et al. Korean Genome Project: 1094 Korean personal genomes with clinical information. *Sci Adv* 2020;6:eaaz7835.
- [67] Kars ME, Başak AN, Onat OE, Bilguvar K, Choi J, Itan Y, et al. The genetic structure of the Turkish population reveals high levels of variation and admixture. *Proc Natl Acad Sci U S A* 2021; 118:e2026076118.
- [68] Razali RM, Rodriguez-Flores J, Ghorbani M, Naeem H, Aamer W, Aliyev E, et al. Thousands of Qatari genomes inform human migration history and improve imputation of Arab haplotypes. *Nat Commun* 2021;12:5929.
- [69] Jain A, Bhojar RC, Pandhare K, Mishra A, Sharma D, Imran M, et al. IndiGenomes: a comprehensive resource of genetic variants from over 1000 Indian genomes. *Nucleic Acids Res* 2021; 49:D1225–32.
- [70] Nagasaki M, Yasuda J, Katsuoka F, Nariai N, Kojima K, Kawai Y, et al. Rare variant discovery by deep whole-genome sequencing of 1,070 Japanese individuals. *Nat Commun* 2015;6:8018.
- [71] Okada Y, Momozawa Y, Sakaue S, Kanai M, Ishigaki K, Akiyama M, et al. Deep whole-genome sequencing reveals recent selection signatures linked to evolution and disease risk of Japanese. *Nat Commun* 2018;9:1631.
- [72] Tadaka S, Katsuoka F, Ueki M, Kojima K, Makino S, Saito S, et al. 3.5KJPNv2: an allele frequency panel of 3552 Japanese individuals including the X chromosome. *Hum Genome Var* 2019; 6:28.
- [73] Wu D, Dou J, Chai X, Bellis C, Wilm A, Shih CC, et al. Large-scale whole-genome sequencing of three diverse Asian populations in Singapore. *Cell* 2019;179:736–49.e15.
- [74] GenomeAsia 100K Consortium. The GenomeAsia 100K Project enables genetic discoveries across Asia. *Nature* 2019; 576:106–11.
- [75] Cengnata A, Deng L, Yap WS, Lim LR, Leong CO, Xu S, et al. A genotype imputation reference panel specific for native Southeast Asian populations. *NPJ Genom Med* 2024;9:47.
- [76] Hwang MY, Choi NH, Won HH, Kim BJ, Kim YJ. Analyzing the Korean reference genome with meta-imputation increased the imputation accuracy and spectrum of rare variants in the Korean population. *Front Genet* 2022;13:1008646.
- [77] Jeon S, Choi H, Jeon Y, Choi WH, Choi H, An K, et al. Korea4K: whole genome sequences of 4,157 Koreans with 107 phenotypes derived from extensive health check-ups. *Gigascience* 2024;13:giae014.
- [78] Tadaka S, Kawashima J, Hishinuma E, Saito S, Okamura Y, Otsuki A, et al. jMorp: Japanese Multi-Omics Reference Panel update report 2023. *Nucleic Acids Res* 2024;52:D622–32.
- [79] Ardiansyah E, Riza AL, Dian S, Ganiem AR, Alisjahbana B, Setiabudiawan TP, et al. Sequencing whole genomes of the West Javanese population in Indonesia reveals novel variants and improves imputation accuracy. *Front Genet* 2025;15:1492602.
- [80] He Y, Lei C, Wan C, Zeng S, Zhang T, Luo F, et al. A comprehensive whole genome database of ethnic minority populations. *Sci Rep* 2024;14:13954.
- [81] Yu C, Lan X, Tao Y, Guo Y, Sun D, Qian P, et al. A high-resolution haplotype-resolved reference panel constructed from the China Kadoorie Biobank study. *Nucleic Acids Res* 2023; 51:11770–82.
- [82] Huang S, Liu S, Huang M, He JR, Wang C, Wang T, et al. The Born in Guangzhou Cohort Study enables generational genetic discoveries. *Nature* 2024;626:565–73.
- [83] Skoglund P, Mallick S, Bortolini MC, Chennagiri N, Hünemeier T, Petzl-Erler ML, et al. Genetic evidence for two founding populations of the Americas. *Nature* 2015;525:104–8.
- [84] Raghavan M, Skoglund P, Graf KE, Metspalu M, Albrechtsen A, Moltke I, et al. Upper Palaeolithic Siberian genome reveals dual ancestry of Native Americans. *Nature* 2014;505:87–91.
- [85] Mathias RA, Taub MA, Gignoux CR, Fu W, Musharoff S, O'Connor TD, et al. A continuum of admixture in the Western Hemisphere revealed by the African Diaspora genome. *Nat Commun* 2016;7:12522.
- [86] O'Connell J, Yun T, Moreno M, Li H, Litterman N, Kolesnikov A, et al. A population-specific reference panel for improved genotype imputation in African Americans. *Commun Biol* 2021; 4:1269.
- [87] Hou L, Kember RL, Roach JC, O'Connell JR, Craig DW, Bucan M, et al. A population-specific reference panel empowers genetic studies of Anabaptist populations. *Sci Rep* 2017;7:6079.
- [88] Cheng PL, Wang H, Dombroski BA, Farrell JJ, Horng I, Chung T, et al. A specialized reference panel with structural variants integration for improving genotype imputation in Alzheimer disease and related dementias. *HGG Adv* 2025;6:100487.
- [89] Naslavsky MS, Scliar MO, Yamamoto GL, Wang JYT, Zverinova S, Karp T, et al. Whole-genome sequencing of 1,171 elderly admixed individuals from Brazil. *Nat Commun* 2022; 13:1004.
- [90] Ziyatdinov A, Torres J, Alegre-Díaz J, Backman J, Mbachou J, Turner M, et al. Genotyping, sequencing and analysis of 140,000 adults from Mexico City. *Nature* 2023;622:784–93.
- [91] Gurdasani D, Carstensen T, Fatumo S, Chen G, Franklin CS, Prado-Martinez J, et al. Uganda genome resource enables insights into population history and genomic discovery in Africa. *Cell* 2019;179:984–1002.e36.
- [92] Carlson JC, Krishnan M, Liu S, Anderson KJ, Zhang JZ, Yapp TJ, et al. Improving imputation quality in Samoans through the integration of population-specific sequences into existing reference panels. *medRxiv* 2023;23297835.
- [93] Herzig AF, Velo-Suarez L, FrEx Consortium, FranceGenRef Consortium, Dina C, Redon R, et al. How local reference panels improve imputation in French populations. *Sci Rep* 2024;14:370.
- [94] Deng T, Zhang P, Garrick D, Gao H, Wang L, Zhao F. Comparison of genotype imputation for SNP array and low-

- coverage whole-genome sequencing data. *Front Genet* 2022; 12:704118.
- [95] Sengupta D, Botha G, Meintjes A, Mbiyavanga M, AWI-Gen Study, H3Africa Consortium, et al. Performance and accuracy evaluation of reference panels for genotype imputation in sub-Saharan African populations. *Cell Genom* 2023;3:100332.
- [96] Xu J, Liu D, Hassan A, Genovese G, Cote AC, Fennessy B, et al. Evaluation of imputation performance of multiple reference panels in a Pakistani population. *HGG Adv* 2025;6:100395.
- [97] Hofmeister RJ, Ribeiro DM, Rubinacci S, Delaneau O. Accurate rare variant phasing of whole-genome and whole-exome sequencing data in the UK Biobank. *Nat Genet* 2023;55:1243–9.
- [98] Browning BL, Tian X, Zhou Y, Browning SR. Fast two-stage phasing of large-scale sequence data. *Am J Hum Genet* 2021; 108:1880–90.
- [99] Browning SR, Browning BL. Haplotype phasing: existing methods and new developments. *Nat Rev Genet* 2011;12:703–14.
- [100] Delaneau O, Zagury JF, Robinson MR, Marchini JL, Dermizakis ET. Accurate, scalable and integrative haplotype estimation. *Nat Commun* 2019;10:5436.
- [101] Williams CM, O'Connell J, Jewett E, Freyman WA, 23andMe Research Team, Gignoux CR, et al. Phasing millions of samples achieves near perfect accuracy, enabling parent-of-origin analyses. *HGG Adv* 2026;7:100526.
- [102] De Marino A, Mahmoud AA, Bose M, Bircan KO, Terpolovsky A, Bamunusinghe V, et al. A comparative analysis of current phasing and imputation software. *PLoS One* 2022; 17:e0260177.
- [103] Huang J, Howie B, McCarthy S, Memari Y, Walter K, Min JL, et al. Improved imputation of low-frequency and rare variants using the UK10K haplotype reference panel. *Nat Commun* 2015;6:8111.
- [104] Rubinacci S, Delaneau O, Marchini J. Genotype imputation using the Positional Burrows Wheeler Transform. *PLoS Genet* 2020;16:e1009049.
- [105] Davies RW, Kucka M, Su D, Shi S, Flanagan M, Cunliffe CM, et al. Rapid genotype imputation from sequence with reference panels. *Nat Genet* 2021;53:1104–11.
- [106] Rubinacci S, Ribeiro DM, Hofmeister RJ, Delaneau O. Efficient phasing and imputation of low-coverage sequencing data using large reference panels. *Nat Genet* 2021;53:120–6.
- [107] Liu S, Liu Y, Gu Y, Lin X, Zhu H, Liu H, et al. Utilizing non-invasive prenatal test sequencing data for human genetic investigation. *Cell Genom* 2024;4:100669.
- [108] Cox SL, Moos H, Stock JT, Shbat A, Bitarello BD, Nicklisch N, et al. Predicting skeletal stature using ancient DNA. *Am J Biol Anthropol* 2022;177:162–74.
- [109] Çubukcu H, Kiling GM. Evaluation of genotype imputation using Glimpse tools on low coverage ancient DNA. *Mamm Genome* 2024;35:461–73.
- [110] Rubinacci S, Hofmeister RJ, Sousa da Mota B, Delaneau O. Imputation of low-coverage sequencing data from 150,119 UK Biobank genomes. *Nat Genet* 2023;55:1088–90.
- [111] Yu K, Das S, LeFaive J, Kwong A, Pleinness J, Forer L, et al. Meta-imputation: an efficient method to combine genotype data after imputation with multiple reference panels. *Am J Hum Genet* 2022;109:1007–15.
- [112] Lazaridis I, Patterson N, Mittnik A, Renaud G, Mallick S, Kirsanow K, et al. Ancient human genomes suggest three ancestral populations for present-day Europeans. *Nature* 2014; 513:409–13.
- [113] Sheng X, Xia L, Cahoon JL, Conti DV, Haiman CA, Kachuri L, et al. Inverted genomic regions between reference genome builds in humans impact imputation accuracy and decrease the power of association testing. *HGG Adv* 2023;4:100159.
- [114] Pan Y, Zhang C, Lu Y, Ning Z, Lu D, Gao Y, et al. Genomic diversity and post-admixture adaptation in the Uyghurs. *Natl Sci Rev* 2022;9:nwab124.
- [115] Tam V, Patel N, Turcotte M, Bossé Y, Paré G, Meyre D. Benefits and limitations of genome-wide association studies. *Nat Rev Genet* 2019;20:467–84.
- [116] Anderson CA, Pettersson FH, Clarke GM, Cardon LR, Morris AP, Zondervan KT. Data quality control in genetic case-control association studies. *Nat Protoc* 2010;5:1564–73.
- [117] Winkler TW, Day FR, Croteau-Chonka DC, Wood AR, Locke AE, Mägi R, et al. Quality control and conduct of genome-wide association meta-analyses. *Nat Protoc* 2014;9:1192–212.
- [118] Bycroft C, Freeman C, Petkova D, Band G, Elliott LT, Sharp K, et al. The UK Biobank resource with deep phenotyping and genomic data. *Nature* 2018;562:203–9.
- [119] Yengo L, Sidorenko J, Kemper KE, Zheng Z, Wood AR, Weedon MN, et al. Meta-analysis of genome-wide association studies for height and body mass index in ~700,000 individuals of European ancestry. *Hum Mol Genet* 2018;27:3641–9.
- [120] Zhao W, Rasheed A, Tikkanen E, Lee JJ, Butterworth AS, Howson JMM, et al. Identification of new susceptibility loci for type 2 diabetes and shared etiological pathways with coronary heart disease. *Nat Genet* 2017;49:1450–7.
- [121] Hyde CL, Nagle MW, Tian C, Chen X, Paciga SA, Wendland JR, et al. Identification of 15 genetic loci associated with risk of major depression in individuals of European descent. *Nat Genet* 2016;48:1031–6.
- [122] Gaynor SM, Joseph T, Bai X, Zou Y, Boutkov B, Maxwell EK, et al. Yield of genetic association signals from genomes, exomes and imputation in the UK Biobank. *Nat Genet* 2024; 56:2345–51.
- [123] Kachuri L, Chatterjee N, Hirbo J, Schaid DJ, Martin I, Kullo IJ, et al. Principles and methods for transferring polygenic risk scores across global populations. *Nat Rev Genet* 2024;25:8–25.
- [124] Lewontin RC. Population genetics. *Annu Rev Genet* 1973;7:1–17.
- [125] Bradburd GS, Ralph PL. Spatial population genetics: it's about time. *Annu Rev Ecol Evol Syst* 2019;50:427–49.
- [126] Nagar SD, Conley AB, Jordan IK. Population structure and pharmacogenomic risk stratification in the United States. *BMC Biol* 2020;18:140.
- [127] Almarri MA, Bergström A, Prado-Martinez J, Yang F, Fu B, Dunham AS, et al. Population structure, stratification, and introgression of human structural variation. *Cell* 2020;182: 189–99.e15.
- [128] Aneli S, Birolo G, Matullo G. Twenty years of the Human Genome Diversity Project. *Hum Popul Genet Genom* 2022; 2:0005.
- [129] Borda V, Loesch DP, Guo B, Laboulaye R, Veliz-Otani D, French JN, et al. Genetics of Latin American Diversity Project: insights into population genetics and association studies in admixed groups in the Americas. *Cell Genom* 2024;4:100692.
- [130] Schraiber JG, Akey JM. Methods and models for unravelling human evolutionary history. *Nat Rev Genet* 2015;16:727–40.
- [131] Barbarino JM, Whirl-Carrillo M, Altman RB, Klein TE. PharmGKB: a worldwide resource for pharmacogenomic information. *Wiley Interdiscip Rev Syst Biol Med* 2018;10:e1417.
- [132] Reay WR, Cairns MJ. Advancing the use of genome-wide association studies for drug repurposing. *Nat Rev Genet* 2021; 22:658–71.
- [133] Sanseau P, Agarwal P, Barnes MR, Pastinen T, Richards JB, Cardon LR, et al. Use of genome-wide association studies for drug repositioning. *Nat Biotechnol* 2012;30:317–20.
- [134] Zhou K, Pedersen HK, Dawed AY, Pearson ER. Pharmacogenomics in diabetes mellitus: insights into drug action and drug discovery. *Nat Rev Endocrinol* 2016;12:337–46.
- [135] Norton ME. Noninvasive prenatal testing to analyze the fetal genome. *Proc Natl Acad Sci U S A* 2016;113:14173–5.
- [136] Haidar H, Le Clerc-Blain J, Vanstone M, Laberge AM, Bibeau G, Ghulmiyyah L, et al. A qualitative study of women and partners from Lebanon and Quebec regarding an expanded scope of non-invasive prenatal testing. *BMC Pregnancy Childbirth* 2021;21:54.

- [137] Chan KCA, Jiang P, Sun K, Cheng YKY, Tong YK, Cheng SH, et al. Second generation noninvasive fetal genome analysis reveals *de novo* mutations, single-base parental inheritance, and preferred DNA ends. *Proc Natl Acad Sci U S A* 2016;113:E8159–68.
- [138] Hui R, D'Atanasio E, Cassidy LM, Scheib CL, Kivisild T. Evaluating genotype imputation pipeline for ultra-low coverage ancient genomes. *Sci Rep* 2020;10:18542.
- [139] Ausmees K, Sanchez-Quinto F, Jakobsson M, Nettelblad C. An empirical evaluation of genotype imputation of ancient DNA. *G3 (Bethesda)* 2022;12:jkac089.
- [140] Martiniano R, Cassidy LM, Ó'Maoldúin R, McLaughlin R, Silva NM, Manco L, et al. The population genomics of archaeological transition in west Iberia: investigation of ancient substructure using imputation and haplotype-based methods. *PLoS Genet* 2017;13:e1006852.
- [141] Allentoft ME, Sikora M, Refoyo-Martínez A, Irving-Pease EK, Fischer A, Barrie W, et al. Population genomics of post-glacial western Eurasia. *Nature* 2024;625:301–11.
- [142] Barrie W, Yang Y, Irving-Pease EK, Attfield KE, Scorrano G, Jensen LT, et al. Elevated genetic risk for multiple sclerosis emerged in steppe pastoralist populations. *Nature* 2024;625:321–8.
- [143] Irving-Pease EK, Refoyo-Martínez A, Barrie W, Ingason A, Pearson A, Fischer A, et al. The selection landscape and genetic legacy of ancient Eurasians. *Nature* 2024;625:312–20.
- [144] Ringbauer H, Huang Y, Akbari A, Mallick S, Olalde I, Patterson N, et al. Accurate detection of identity-by-descent segments in human ancient DNA. *Nat Genet* 2024;56:143–51.
- [145] Wang M, Chen H, Luo L, Huang Y, Duan S, Yuan H, et al. Forensic investigative genetic genealogy: expanding pedigree tracing and genetic inquiry in the genomic era. *J Genet Genomics* 2025;52:460–72.
- [146] Erlich Y, Shor T, Pe'er I, Carmi S. Identity inference of genomic data using long-range familial searches. *Science* 2018;362:690–4.
- [147] Ram N, Murphy EE, Suter SM. Regulating forensic genetic genealogy. *Science* 2021;373:1444–6.
- [148] May T. Sociogenetic risks — ancestry DNA testing, third-party identity, and protection of privacy. *N Engl J Med* 2018;379:410–2.
- [149] Dowdeswell TL. Forensic genetic genealogy: a profile of cases solved. *Forensic Sci Int Genet* 2022;58:102679.
- [150] Bentley AR, Callier S, Rotimi CN. Diversity and inclusion in genomic research: why the uneven progress? *J Community Genet* 2017;8:255–66.
- [151] Need AC, Goldstein DB. Next generation disparities in human genomics: concerns and remedies. *Trends Genet* 2009;25:489–94.
- [152] Jones KM, Cook-Deegan R, Rotimi CN, Callier SL, Bentley AR, Stevens H, et al. Complicated legacies: the human genome at 20. *Science* 2021;371:564–9.
- [153] Manolio TA, Goodhand P, Ginsburg G. The International Hundred Thousand Plus Cohort Consortium: integrating large-scale cohorts to address global scientific challenges. *Lancet Digit Health* 2020;2:e567–8.
- [154] Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* 2016;3:160018.
- [155] Carroll SR, Herczog E, Hudson M, Russell K, Stall S. Operationalizing the CARE and FAIR principles for indigenous data futures. *Sci Data* 2021;8:108.
- [156] Cavinato T, Rubinacci S, Malaspinas AS, Delaneau O. A resampling-based approach to share reference panels. *Nat Comput Sci* 2024;4:360–6.
- [157] Wang S, Jiang X, Singh S, Marmor R, Bonomi L, Fox D, et al. Genome privacy: challenges, technical approaches to mitigate risk, and ethical considerations in the United States. *Ann N Y Acad Sci* 2017;1387:73–83.
- [158] Ouzhuluobu, He Y, Lou H, Cui C, Deng L, Gao Y, et al. *De novo* assembly of a Tibetan genome and identification of novel structural variants associated with high-altitude adaptation. *Natl Sci Rev* 2020;7:391–402.
- [159] Lou H, Gao Y, Xie B, Wang Y, Zhang H, Shi M, et al. Haplotype-resolved *de novo* assembly of a Tujia genome suggests the necessity for high-quality population-specific genome references. *Cell Syst* 2022;13:321–33.e6.
- [160] Nurk S, Koren S, Rhie A, Rautiainen M, Bizakadze AV, Mikheenko A, et al. The complete sequence of a human genome. *Science* 2022;376:44–53.
- [161] Chao KH, Zimin AV, Pertea M, Salzberg SL. The first gapless, reference-quality, fully annotated genome from a Southern Han Chinese individual. *G3 (Bethesda)* 2023;13:jkac321.
- [162] Wang T, Antonacci-Fulton L, Howe K, Lawson HA, Lucas JK, Phillippy AM, et al. The Human Pangenome Project: a global resource to map genomic diversity. *Nature* 2022;604:437–46.
- [163] Gao Y, Yang X, Chen H, Tan X, Yang Z, Deng L, et al. A pangenome reference of 36 Chinese populations. *Nature* 2023;619:112–21.
- [164] Hannan AJ. Tandem repeats mediating genetic plasticity in health and disease. *Nat Rev Genet* 2018;19:286–98.
- [165] Saini S, Mitra I, Mousavi N, Fotsing SF, Gymrek M. A reference haplotype panel for genome-wide imputation of short tandem repeats. *Nat Commun* 2018;9:4397.
- [166] Ziaei Jam H, Li Y, DeVito R, Mousavi N, Ma N, Lujumba I, et al. A deep population reference panel of tandem repeat variation. *Nat Commun* 2023;14:6711.