RESEARCH | Article in Press

Korea10K integrates 10,239 whole genomes with multiomic and clinical health information as a Korean reference dataset

Kyungwhan An[1,2,3] · Sungwon Jeon[1,4] · Yoonsung Kwon[1,2,5] · Yookyung Choi[1,2,5] · Changhan Yoon[1,2] · Yeonsu Jeon[5] · Jihun Bhak[1,2] · Dong-Hyun Shin[1,2] · Hyoung-Jin Choi[1,2] · Hyomin Lee[1,2] · Yeo Jin Kim[1] · Eun-Seok Shin[6] · Hyojung Ryu[1,4] · Asta Blazyte[1,7] · Sangsoo Park[1,2] · Juok Cho[1,2] · Dan Bolser[8] · Soobok Joe[5] · Jin Ok Yang[5] · Jongbum Jeon[5,9] · Jong-Hwan Kim[5,9] · Jungeun Kim[10] · Dooyoung Jung[2,11] · Yun Sung Cho[12,13] · Kiyuk Chang[14,15] · Eun Ho Choo[14,15] · Eunmin Kim[14] · Sang Yeub Lee[16] · Weon Kim[17] · Min Gyu Kang[18] · Ae-Young Her[19] · Suk Chon[20] · Jeong-Taek Woo[20] · Sang Youl Rhee[20] · Siwoo Lee[21] · Hyo-Jeong Ban[22] · Hee-Jeong Jin[22] · Younghwa Baek[22] · Yong Min Ahn[23,24,25] · Sang Jin Rhee[23,24] · Min Ji Kim[26] · Sang Yeol Lee[27] · Chan-Mo Yang[27] · Se-Hoon Shim[28] · Seong-Jin Cho[29] · Shin Gyeom Kim[30] · Hyung-Tae Jung[31] · Byung-Joo Ham[32,33,34] · Yoon Young Choi[35] · Jae-Ho Cheong[36] · Seung-Ki Kim[37,38] · Ji Hoon Phi[37,38] · Seung Ah Choi[37,38] · Heon Yung Gee[39] · Sun Young Joo[39] · Jinsei Jung[40] · Wonsuk Shin[41,42] · Sang-Hyuk Lee[41,43] · Borah Kim[43,44] · Woojae Myung[24,45] · Chong Kun Cheon[46] · Dong Uk Kim[47] · Seok-Soo Byun[48,49] · Gangnam Jin[50] · Hojun Lee[5,50] · Kyun Shik Chae[51] · Chang Geun Kim[51] · Bonghee Lee[7,52] · Jaesuk Lee[52] · Kwangwoo Kim[53,54] · Semin Lee[1,2] · Neung-Hwa Park[55,56] · Haeyoung Jeong[5] · George M. Church[57,58,59] · Jong Bhak[1,2,4]

Received: 1 April 2026 / Accepted: 19 October 2026
© The Author(s) 2026

Abstract

Background Large-scale genome projects have advanced the characterization of human genetic diversity. However, the lack of deeply sequenced, multiomic resources representing the Korean population has limited systematic evaluation of population-specific variation and its functional and translational impact.

Results We present Korea10K, a population-scale multiomic resource comprising 10,239 high-depth whole genomes (mean coverage 30×) with integrated molecular and phenotypic data. Analysis of 9,000 unrelated

Kyungwhan An, Sungwon Jeon and Yoonsung Kwon contributed equally to this work.

This Article in Press is shared early to give you faster access to new research. It is citable and carries a permanent DOI. The final edited version will replace it automatically.

https://doi.org/10.1186/s13059-026-04321-6

Published online: 20 November 2026

individuals enables near-complete discovery of ultra-rare variants (allele count > 2 and allele frequency ≤ 0.0005) and supports the construction of a high-resolution, population-specific imputation panel. Koreans exhibit pronounced autosomal genetic homogeneity together with a narrow Y-chromosomal and *HLA-A* allelic spectrum and comparatively deep mitochondrial lineage diversity, consistent with long-term demographic continuity. We further identify 16.4 million variants that alter CpG dinucleotide context, revealing widespread sequence-driven modulation of the genomic CG landscape. Notably, CG-eliminating variants generate loss of methylation signal and disrupt CpG probe targets in widely used methylation arrays, introducing a systematic source of bias in epigenome-wide association studies and epigenetic clock estimation.

Conclusions Korea10K establishes a high-resolution genomic and multiomic reference for the Korean population and reveals a previously unrecognized interaction between genetic variation and epigenomic measurement. This resource provides a foundation for precision medicine and highlights the need for ancestry-aware interpretation of molecular data.

Keywords Korean Genome Project · Korea10K · Whole-genome sequencing · Population genomics · Imputation panel · Multiomics · CpG variation · Epigenomics · Epigenetic clocks

Background

Population-scale whole-genome sequencing (WGS) has enabled high-resolution investigations into global genetic diversity, population structure, and disease susceptibility. Initiatives such as the 1,000 Genomes Project (1KGP) and the Genome Aggregation Database (gnomAD) [1] have collectively established a foundation for cataloging genetic variants across global populations. More recent large-scale biobank projects, including the UK Biobank [2] and FinnGen [3], have further integrated genomic data with extensive phenotypic records, offering powerful platforms for genotype–phenotype association and precision medicine. Across Asia, a growing number of national and regional genome programs have begun to capture both local and continental genetic diversity. Notable projects include the Taiwan Biobank [4], Biobank Japan [5], and Singapore's PRECISE-SG100K project [6, 7], which collectively have advanced genomic medicine in East and Southeast Asia.

Within Korea, whole-genome sequencing projects have progressed, beginning with KoVariome (2018) with 50 individuals under Korean Personal Genome Project (KPGP) [8], Korea1K (2020) with 1,094 whole genomes under the Korean Genome Project (KGP) [9] and Korean Variant Archive 2 (KOVA2; 2022) with 5,305 genomes and exomes which partly includes KGP data [10], followed by Korea4K (2024) with 4,157 high-depth whole genomes [11], providing important insights into population-specific variation. However, these studies have been limited by sample size, sequencing depth, or incomplete integration of molecular and clinical data. As a result, a comprehensive, population-scale reference capturing the full spectrum of Korean genetic and molecular variation has remained unavailable.

Here, we present Korea10K, a large-scale genomic and multiomic resource comprising 10,239 high-depth whole genomes integrated with transcriptomic, epigenomic, and phenotypic data. Since 2015, Korea10K has expanded prior Korean Genome Project efforts (KPGP, Korea1K, and Korea4K) into a nationwide, population-scale resource comprising 10,239 high-depth whole genomes (Fig. 1A). Participants were recruited across multiple institutions, capturing broad demographic representation and phenotypic diversity. We finally collected both healthy individuals (44.9%, $N = 4{,}596$) and participants with clinically diagnosed conditions (55.1%, $N = 5{,}643$), spanning cardiovascular, neuropsychiatric, metabolic, oncologic, and congenital diseases (Fig. 1B, C). Cardiovascular disorders represent the largest subgroup (56.6%), followed by neuropsychiatric conditions, enabling disease-focused analyses at a population scale.

For future resources, Korea10K marks a major expansion of the KGP into the multiomic era. We generated multiomic profiles from blood that include 868 transcriptomes, 1,211 DNA methylomes, and 102 chromatin accessibility datasets, providing a systems-level view of the molecular consequences of genetic variation. Complementary

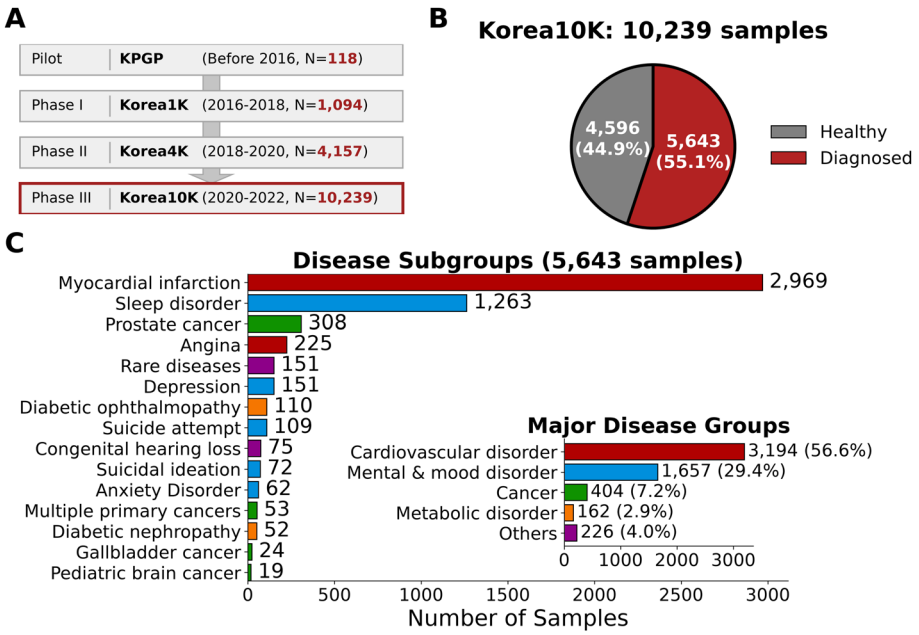


Fig. 1 Overview of the Korea10K cohort. **A** Sequential recruitment phases of the Korea10K project, expanding from a pilot study (KPGP) to large-scale national inclusion through Korea1K, Korea4K, and Korea10K, comprising 10,239 high-depth whole genomes collected in total. **B** Distribution of healthy ($N=4,596$) and clinically diagnosed ($N=5,643$) participants. **C** Composition of disease subgroups and major disease categories (y-axis) among diagnosed participants (x-axis). Cardiovascular disorders represented the largest group ($N=3,194$), followed by mental and mood disorders ($N=1,657$), cancers ($N=404$), metabolic disorders ($N=162$), and others ($N=226$). Abbreviation: KPGP = Korean Personal Genome Project

phenotypic resources include the Health Check-up (HC) dataset (6,583 participants, 336 phenotypes; Additional file 1: Table S1) and the Lifestyle Questionnaire (LQ) dataset (4,923 participants, 370 survey measures; Additional file 1: Table S2), encompassing diet, physical activity, smoking, alcohol use, sleep, and psychosocial factors (Table 1). In addition, KOREF (KOrean REFerence) represents a reference-quality multiomic resource within the Korean Genome Project, with deeply profiled longitudinal datasets spanning more than a decade (Additional file 2: Supplementary Methods).

These datasets enable systematic characterization of population-specific genetic variation and its functional consequences together with matched multiomics data, providing a foundation for integrative multiomics analyses in the Korean population.

Table 1 Overview of multiomic and phenomic data resource in Korea10K

Omics type	Sequencing type	Sequencing platform	Source	Sample size
Epigenome	Bisulfite-seq (Targeted)	NovaSeq6000, HiSeq2500	Whole blood	1,211
	ATAC-seq	NovaSeq6000	PBMC	102
Transcriptome	mRNA-seq	NovaSeq6000, HiSeq2500, BGISeq500	Whole blood	868
Clinical Records	-	-	Whole blood, Clinical data	6,583
Lifestyle Questionnaire	-	-	Survey data	4,923

Abbreviation: *PBMC* Peripheral Blood Mononuclear Cell

Results

Variant discovery and population reference imputation panel for Koreans

WGS of 10,239 Korean individuals yielded the most comprehensive catalog of genomic variation in this population to date. Analysis of 9,000 unrelated Koreans identified 61,223,704 autosomal variants, including 57,184,188 single-nucleotide variants (SNVs) and 4,039,516 short insertions and deletions (indels). Of these, 9,880,066 (16.1%) were previously unreported in dbSNP (v157), with enrichment for low-frequency alleles (Fig. 2A). Specifically, 33.6% of novel variants were classified as singletons, 14.9% as doubletons, and 3.4% as ultra-rare (AC > 2 and AF ≤ 0.0005). Across all frequencies, most of the novel variants were classified as SNV, except for the very rare ones, where 34.9% were small indels (22.1% deletions and 12.8% insertions; Additional file 2: Fig. S1).

Cumulative variant discovery curves showed a continuous increase in novel variant counts with expanding sample size (Fig. 2B), especially singletons and doubletons. Saturation analysis over 100 permutations indicated that ultra-rare variant discovery plateaued at a mean of 8,955 samples (95% CI: 8,951–8,958), while

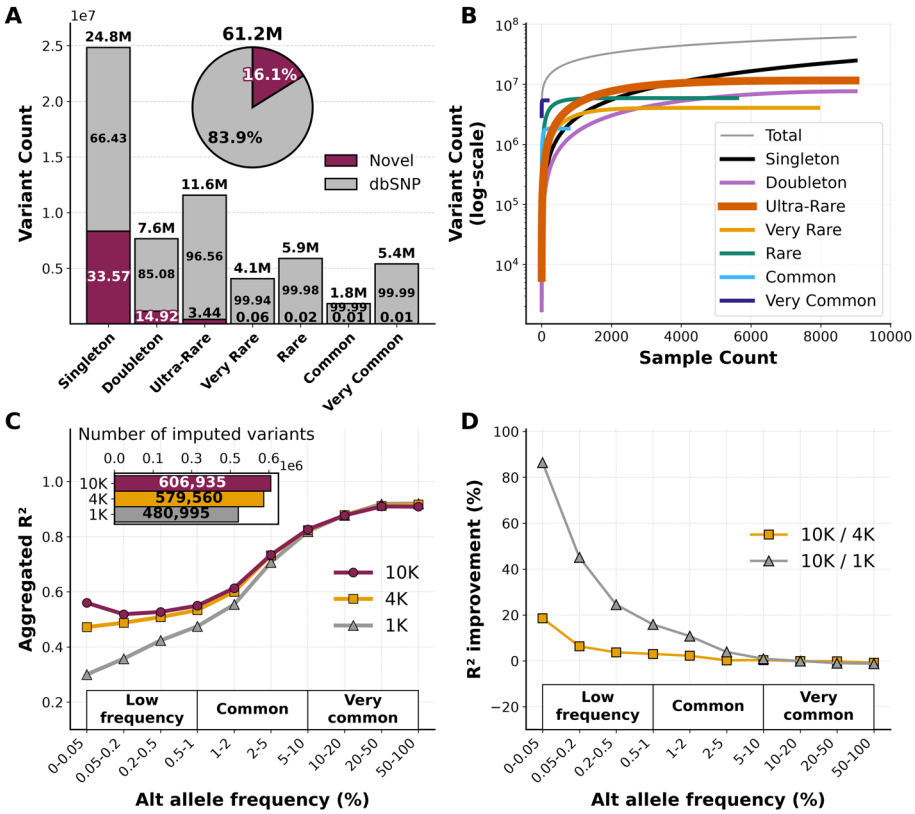


Fig. 2 Variome landscape of the Korea10K whole genomes. **A** Stacked bar charts highlighting the distribution of discovered variant counts (y-axis) by allele frequency categories (x-axis), partitioned into novel and dbSNP-reported variants. The pie chart summarizes overall variant statistics. **B** Cumulative variant discovery curve showing the number of variants identified with increasing sample size (y-axis, on a log scale) across allele frequency categories (x-axis). **C** Line plot depicting imputation accuracies (y-axis, as aggregated R^2) across alternate allele frequency bins (x-axis) comparing Korea1K, Korea4K, and Korea10K population genome reference panels. The inset horizontal bar plot shows the total number of imputed variants (x-axis) per population genome reference panel (y-axis). **D** Line plot showing R^2 improvement (y-axis, as a percentage) comparing Korea10K to Korea1K and Korea10K to Korea4K across allele frequency bins (x-axis). The allele frequency categories are defined as follows: Singleton (AC = 1), Doubleton (AC = 2), Ultra-Rare (AC > 2 and AF ≤ 0.0005), Very Rare (AF > 0.0005 and AF ≤ 0.001), Rare (AF > 0.001 and AF ≤ 0.01), Common (AF > 0.01 and AF ≤ 0.05), and Very Common (AF > 0.05). Abbreviations: Alt allele = Alternative allele, AC = allele count, AF = allele frequency

very rare (AF > 0.0005 and AF ≤ 0.001) and rare variants (AF > 0.001 and ≤ 0.01) required 7,083 samples (95% CI: 7,026–7,141) and 4,692 samples (95% CI: 4,636–4,747), respectively. Common variants (AF > 0.01 and AF ≤ 0.05) saturated at 569 samples (95% CI: 557–581) while very common variants (AF > 0.05) saturated at 109 samples (95% CI: 106–111) (Additional file 2: Fig. S2; Additional file 1: Table S3). These results underscore the importance of large-scale sequencing to capture low-frequency Korean variation.

The Korea10K population reference panel demonstrated markedly improved imputation performance compared with earlier Korean panels (Korea1K and Korea4K), particularly for low-frequency variants (AF < 1%; Fig. 2C). Relative to Korea1K, imputation accuracy increased by 15.8% to 86.4%, and by 3.0% to 18.6% compared with Korea4K in the descending order of allele frequency bins (Additional file 1: Table S4). The enhancement was most striking for ultra-rare variants, including singletons and doubletons (Fig. 2D). Furthermore, the Korea10K panel imputed 606,935 variants, representing a 26% increase over Korea1K (Fig. 2C, upper left), underscoring the value of progressive variant discovery for genotype imputation.

Population structure and genetic homogeneity of Koreans

Principal component analysis (PCA) of autosomal variation positioned Koreans as a distinct cluster within the East Asian genetic continuum, clearly separated from Han Chinese (CHB, CHS) and Japanese (JPT) populations (Fig. 3A, B). Ethnically relevant reference genomes, KOREF1 [12] and KOREF2, co-localized within this cluster, validating their suitability as representative genomic references. Admixture analysis revealed a dominant ancestry component at K = 7, which best explained global autosomal genetic structure (Fig. 3C, top panel; Additional file 2: Fig. S3). At a finer-scale with K = 9, which showed the second-lowest cross-validation error, Koreans exhibited a dual-ancestry pattern with components shared by northern Han Chinese and Japanese populations. A total of 485 Koreans, one representative per genetic cluster, were used for the ADMIXTURE analysis to limit computational burden while retaining the observed range of Korean genetic variation (Fig. 3C, bottom panel).

Analysis of 5,290 unrelated Korean Y-chromosomal genomes revealed a strong dominance of haplogroup O (N = 4,053, 76.6%), comprising both O1 (N = 2,105, 39.8%) and O2 (N = 1,948, 36.8%), followed by C (N = 847, 16.0%). This distribution is broadly consistent with patterns observed across northern East Asia (Fig. 3D), although population-specific signatures were observable, such as higher D frequencies in the Japanese (35.7%) and a stronger representation of O2 in northern Han Chinese (50.1%). Outside East Asia, Europeans are dominated by haplogroup R (58%), whereas Africans show overwhelming prevalence of haplogroup E (89.9%), illustrating broad continental divergence in male ancestries. Phylogenetic reconstruction based on pairwise Y-chromosomal genetic distances from 5,871 haplogroup-classified males showed limited resolution of major haplogroups and their subclades. The limited phylogenetic resolution reflects shallow sequence divergence among lineages due to recent demographic expansions, probably since Bronze age population explosion [13], into the Korean peninsula and the persistence of shared ancestral polymorphisms (Additional file 2: Fig. S4A).

Mitochondrial genomes in Koreans exhibited markedly greater diversity compared to the Y chromosome. Phylogenetic reconstruction of 10,166 mitochondrial assemblies revealed deeply branching maternal lineages, primarily partitioned into macrohaplogroups M and N (Additional file 2: Fig. S4B). Dominant subclades included D (N = 2,988, 33.2%), M (N = 1,315, 14.6%), and B (N = 1,285, 14.3%) in 9,000 unrelated Koreans (Fig. 3E). The absence of Western Eurasian (H and X) and African (L) haplogroups indicates limited maternal gene flow from outside East Asia (Additional file 2: Fig. S4B), whereas fine-scale structure captured by PCA of 1,247 mitochondrial variants indicates multiple maternal lineage contributions from East Asian populations (Additional file 2: Fig. S5). These findings support multiple ancient migration and admixture events involving females from neighboring populations into the Korean peninsula [14].

At the immunogenetic level, HLA class I (*HLA-A*) allele frequencies illustrate characteristic features of the Korean population. Among 9,000 unrelated Koreans (18,000 *HLA-A* alleles), four alleles – including A*24:02 (N = 4,059, 22.6%), A*02:01 (N = 3,003, 16.7%), A*33:03 (N = 2,879, 16.0%), and A*11:01 (N = 1,871, 10.4%)

https://doi.org/10.1186/s13059-026-04321-6

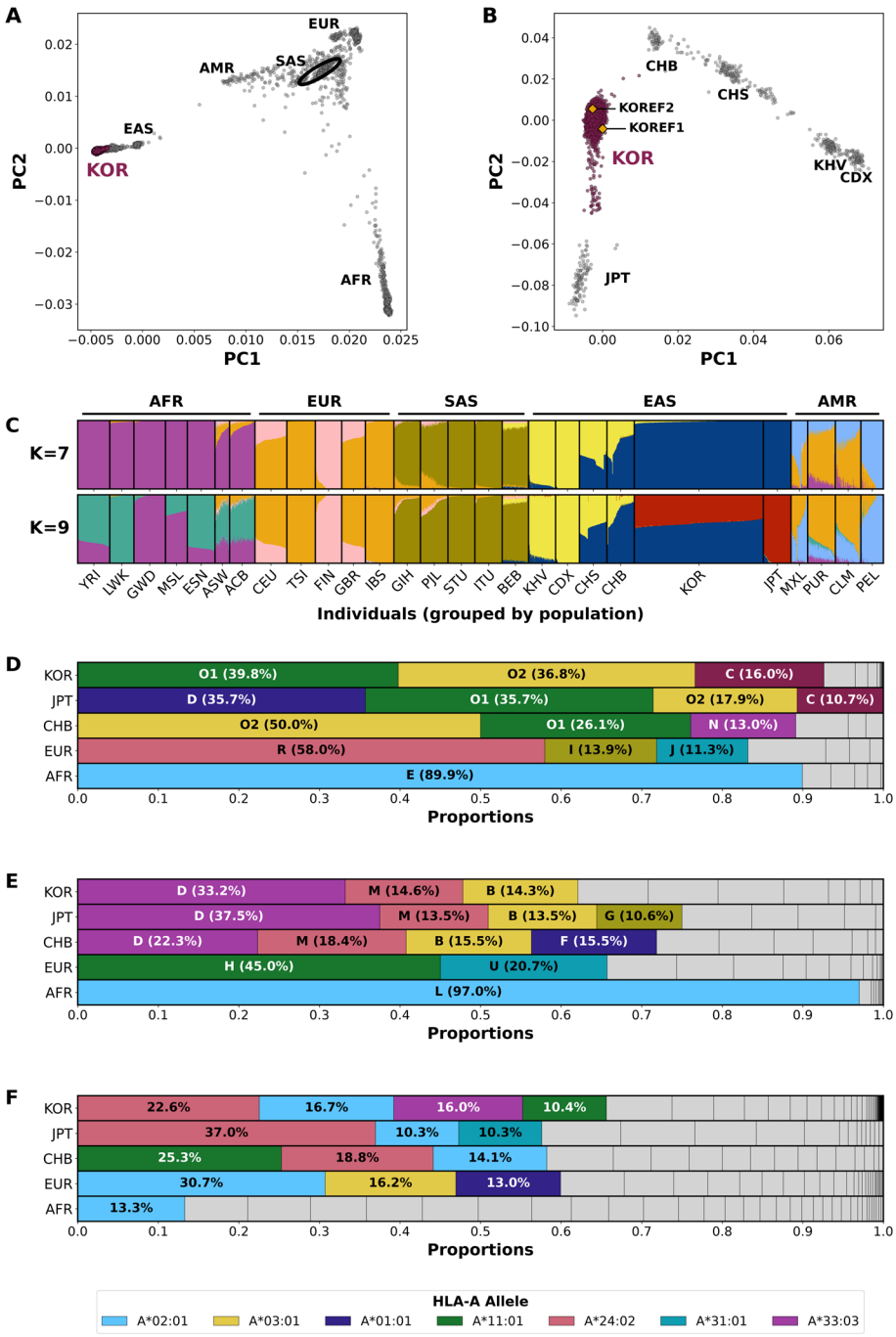


Fig. 3 Population structure and uniparental lineage composition of the Korean population. **A** Global principal component (PC) bi-plot of PC1 (x-axis) and PC2 (y-axis) based on autosomal variants from 9,000 unrelated Koreans (KOR) and 2,541 individuals from the 1000 Genomes Project (1KGP). SAS placed within AMR are open-circled with a black ellipse for distinction. **B** East Asian PC bi-plot of PC1 (x-axis) and PC2 (y-axis) showing 9,000 unrelated Koreans (KOR) and 501 samples in other East Asian populations (CHB, CHS, CDX, JPT, and KHV) of 1KGP. Koreans are displayed as closed dim red dots. The Korean References (KOREF1 and KOREF2) are denoted as orange diamonds. **C** Admixture plots at K = 7 (top panel) and K = 9 (bottom panel). The x-axis displays individuals in analysis that have been grouped by population. The y-axis denotes number of admixture clusters in analysis (i.e., K). **D** Stacked bar chart showing proportions (x-axis) of major Y-chromosomal (except O1 and O2 for scrutiny) haplogroups of 5,290 unrelated Korean males compared to other populations (y-axis). **E** Stacked bar chart showing proportions (x-axis) of major Mitochondrial (MT or mt) haplogroups of 9,000 unrelated Koreans and other populations (y-axis). **F** Stacked bar chart showing proportions (x-axis) of 18,000 *HLA-A* alleles of 9,000 unrelated Koreans compared to other populations (y-axis). Abbreviations: AFR = African, EUR = European, SAS = South Asian, EAS = East Asian, AMR = Admixed American, KHV = Vietnamese, CDX = Chinese Dai, CHS = Southern Han Chinese, CHB = Northern Han Chinese, KOR = Korean in Korea10K, JPT = Japanese in Tokyo, KOREF = Korean Reference, 1KGP = the 1000 genome project

– accounted for nearly two-thirds of all *HLA-A* alleles (Fig. 3F), indicating a narrow and coherent allelic spectrum. In comparison, Japanese showed a stronger enrichment of A*24:02 (37.0%), whereas Han Chinese exhibited a predominance of A*11:01 (25.3%); both populations displayed reduced representation of A*33:03. By contrast, Europeans and Africans exhibited markedly broader HLA polymorphism.

CG-eliminating variants systematically bias DNA methylation measurements and epigenetic clock estimates

DNA sequence variation can directly alter CpG dinucleotide context, introducing a genetically encoded layer of epigenetic variability. Profiling CG context–associated variants (CGVs; also known as, CG-SNPs or meSNPs, or CpG-altering variants) in 9,000 unrelated Koreans identified 16,388,405 variants (26.8% of 61,223,704 high-quality biallelic variants) that directly modify CpG context, including contributions from single-nucleotide variants, insertions and deletions, with distinct functional impacts comprising CG-creating, CG-eliminating, and context-shifting variants (Additional file 2: Fig. S6A, B). These variants were broadly distributed across the allele frequency spectrum (Additional file 2: Fig. S6C, E), indicating that CG context disruption is not restricted to rare alleles but persists across common variation, while genome-wide aggregation of CG differences revealed that most regions exhibit modest deviation, with a minority of windows showing more pronounced shifts (Additional file 2: Fig. S6D). We also confirmed that an actual loss of methylation signal directly in Korea10K methylation data: across 15,442 CpG sites with at least ten individuals in every genotype class, the mean beta value declined with the dosage of the CG-eliminating allele, from 0.721 in non-carriers to 0.386 in heterozygotes and 0.021 in homozygotes (Fig. 4A). The near-linear halving between successive genotypes shows that each disrupted allele is silenced independently, and that a homozygous carrier returns a constitutively unmethylated readout irrespective of biological regulation the site no longer reports methylation (Additional file 2: Fig. S7). In spike-in simulation, variant carriers present at each site's own Hardy–Weinberg frequency reduced the power to detect a true $\Delta\beta$ of 0.05 from 0.93 to 0.13 at sites of 10–20% allele frequency ($N = 400$, Bonferroni threshold); adjusting for allele dosage restored power to 0.61 at these sites and to 0.86–0.90 below 10% allele frequency, but only to 0.13–0.25 above 20%, where too few samples with an intact CpG remain (Additional file 2: Fig. S8, Additional file 2: Supplementary Methods).

To assess their impact on epigenomic assays, we intersected CG-eliminating variants with CpG probe target sites across four widely used methylation arrays: Infinium HumanMethylation27 (27 K), HumanMethylation450 (450 K), MethylationEPIC v2.0 (EPIC v2.0) and Methylation Screening Array (MSA) in Korea10K and in the five 1000 Genomes superpopulations (Fig. 4B-E). Importantly, this effect was not limited to rare variation. On EPIC v2.0, for example, 11,510 CpG sites were disrupted by common variants with allele frequency (AF) > 1% in Koreans, indicating common absence of the target CpG across individuals and representing systematic measurement failure rather than true biological hypomethylation (Fig. 4D). The two frequency regimes, however, behave quite differently across populations. The bulk of the loss is likely individual-specific: 95.5% of the EPIC v2.0 sites eliminated in Korea10K (243,232 of 254,742) were attributable to singletons or variants at AF ≤ 0.01, and the cumulative curves for Korea10K rise far to the left of every superpopulation curve and are already saturated around AF = 0.01 (Fig. 4B-E, right). This gradient is a direct consequence of sampling depth: Korea10K resolves allele frequencies some twenty-fold rarer than the 490–893 genomes available per superpopulation from the 1000 Genomes Project(1KGP). So, the deeper the reference panel, the more of this individual-level probe loss becomes visible. Population specificity was the rule rather than the exception; for example, of the 598,403 CG-eliminating variants that disrupt an EPIC v2.0 target in any cohort, 71.5% segregated in a single cohort and only 1.2% were shared by all six, with the same pattern on every array (single-cohort fraction 70.4–76.5%; Additional file 2: Fig. S9).

Consistently, all six epigenetic clocks examined contained disrupted CpG markers (Fig. 4F). In Korea10K, 12.2% of Horvath [15] (43 of 353), 19.7% of Horvath skin-and-blood [16] (77 of 391), 12.7% of Hannum [17]

https://doi.org/10.1186/s13059-026-04321-6

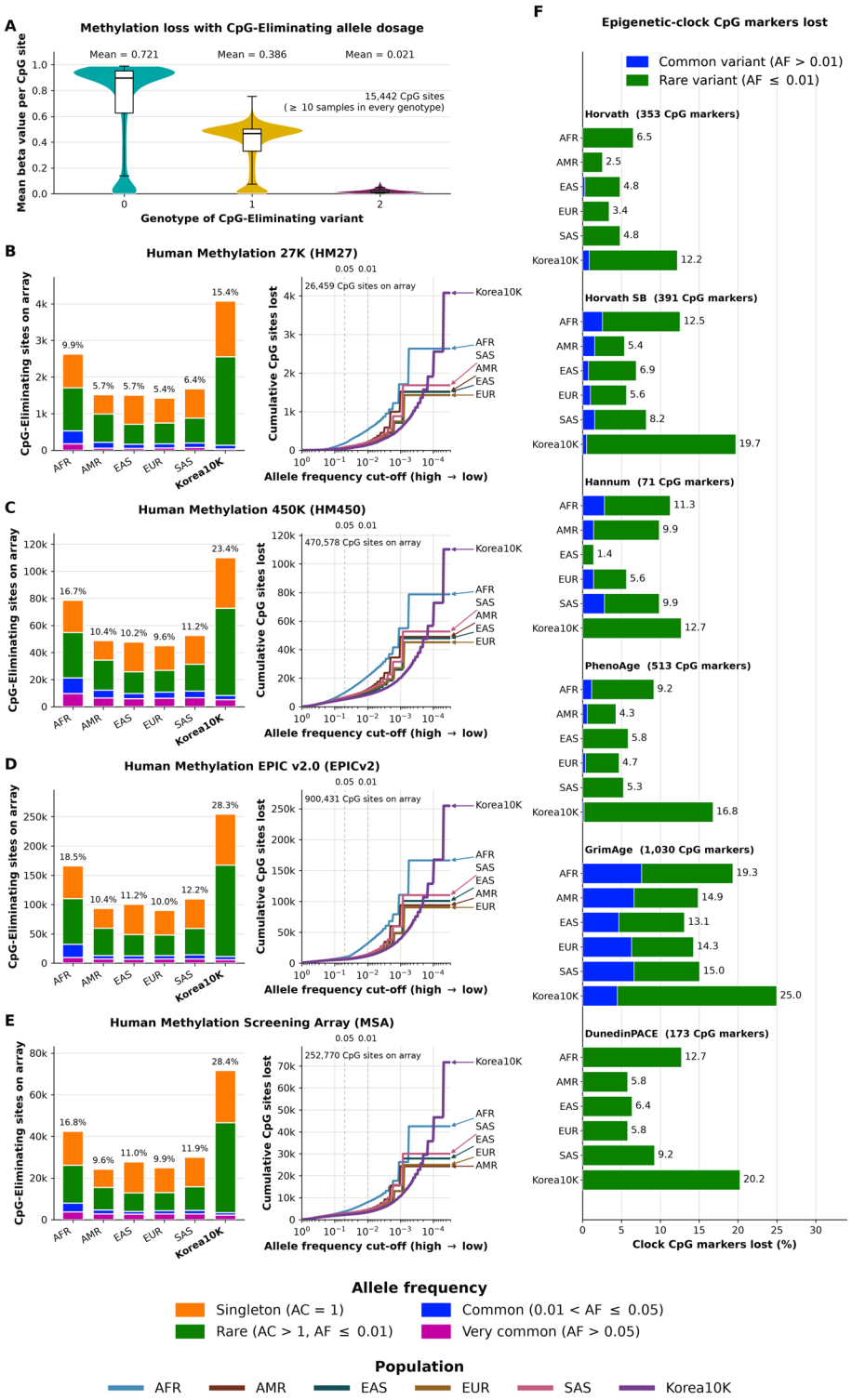


Fig. 4 CG-eliminating variants remove CpG targets from methylation arrays and from epigenetic-clock marker panels. **A** Per-CpG mean beta value in Korea10K as a function of the genotype of the disrupting CG-eliminating variant (0, 1 or 2 copies of the eliminating allele). Each observation is one CpG site; the value plotted is the beta value averaged over all individuals of that genotype at that site. Only the 15,442 CpG sites with at least ten individuals in every genotype class are included, so that the three distributions describe an identical set of sites. **B**–**E** CpG sites eliminated on four Infinium methylation arrays: **B** HumanMethylation27 (26,459 CpG sites), **C** HumanMethylation450 (470,578 sites), **D** MethylationEPIC v2.0 (900,431 sites) and **E** Methylation Screening Array (252,770 sites). Left: number of eliminated CpG sites per cohort, stacked by the allele-frequency class of the most frequent disrupting variant; the percentage above each bar is the eliminated fraction of all CpG sites on that array. Right: cumulative number of eliminated CpG sites as a function of the allele-frequency cut-off. **F** Percentage of the CpG markers of six epigenetic clocks, Horvath, Horvath skin-and-blood (SB), Hannum, PhenoAge, GrimAge, and DunedinPACE, eliminated in at least one individual of each cohort. Bars are grouped by clock and labeled by cohort and are split into markers disrupted by common (AF > 0.01) and rare (AF ≤ 0.01) variants; the number at the end of each bar is the total percentage of markers lost

(9 of 71), 16.8% of PhenoAge [18] (86 of 513), 20.2% of DunedinPACE [19] (35 of 173) and 25.0% of GrimAge [20] markers (257 of 1,030) were eliminated in at least one individual, against 1.4–19.3% in the superpopulations of 1KGP. GrimAge was the most affected clock in every cohort and the only one with a substantial common-variant component (46–78 of its lost markers exceeded AF = 0.01 across cohorts), making it the one clock whose

marker attrition is shared across a population rather than confined to individual carriers; for the remaining clocks the loss was almost entirely from rare variants (for example, all 35 DunedinPACE markers lost in Korea10K).

Discussion

There are several limitations to this study. Korea10K is cross-sectional, except for two KOREF samples. Direct assessment of biological aging, disease progression, and causal inference over time are precluded. Samples acquired via Material Transfer Agreements lack detailed phenotypic and lifestyle information. Incomplete phenotypic data reduced the effective sample size and limited integrative modeling across molecular and clinical layers. Structurally complex regions, including centromeres, sub-telomeric regions, and the repetitive portion of the Y chromosome, were left unresolved. Reliance on short-read sequencing and a linear reference genome further restricted this release to SNVs and short indels.

Future efforts should prioritize longitudinal sampling to enable direct modeling of aging and disease progression. Expansion and harmonization of clinical and lifestyle data will enhance statistical power and biological interpretability. At the genomic level, adoption of long-read sequencing [21], ethnically-relevant assemblies such as the KOREF assembly [12], and graph-based pangenomes [22] will be essential to resolve complex and longer variation and reduce reference bias. Transcriptomic and epigenomic profiling will further enable population-specific mapping of regulatory quantitative trait loci, including transcriptomic [23] and epigenomic QTLs [24, 25], facilitating integrative analysis of genetic and molecular regulation.

Despite these limitations, Korea10K substantially expands the scale of Korean population genomics. Identification of 61 million genetic variants, with nearly 10 million novel alleles, enabled near-saturation discovery of ultra-rare alleles. This comprehensive variant catalog markedly improves imputation accuracy and fine-mapping resolution, enhancing the power of genome-wide and rare-variant association studies.

Population structure analyses reaffirm that Koreans form a distinct cluster within East Asia. Y-chromosomal haplogroups indicate relatively recent male-mediated expansions associated with the peopling of the Korean Peninsula [13], whereas extensive mitochondrial diversity preserves signatures of ancient female-mediated admixture from neighboring East Asian populations [14]. In addition, the generation of 10,239 mitochondrial assemblies and alignments represents one of the largest population-scale mitogenomic resources to date, enabling high-resolution applications in forensics, anthropology, and disease research.

Immunogenetic profiles further reveal a broadly shared dominance of *HLA-A* alleles across northern East Asia, with implications for vaccine response [26], autoimmune disease susceptibility [27], and drug contraindications [28]. The population-scale HLA resource of 10,239 Koreans generated here provides a foundation for clinically relevant immunogenetic studies.

Korea10K has already facilitated the discovery of genomic and multiomic biomarkers associated with a wide range of phenotypes, including cardiovascular disorders (acute myocardial infarction) [29–32], neuropsychiatric traits (depression, anxiety, and suicidal behavior) [33, 34], infectious diseases (COVID-19) [35], and biological aging [36], providing mechanistic insights into disease etiology and progression.

Importantly, our findings highlight a critical interaction between genetic variation and epigenomic measurement, and the two regimes we observe call for different approaches. A minority of CG-eliminating variants are common ($AF > 0.01$) and therefore deplete the effective probe space of an entire population in a systematic, reproducible manner: on EPIC v2.0, this accounts for 11,510 CpG sites in Korea10K and as many as 32,188 in AFR, and among the six epigenetic clocks they are concentrated almost exclusively in GrimAge. Because they recur across individuals in a population, they may bias group-level comparisons between populations, where a probe that is informative in one cohort may be constitutively unmethylated in another, but for the same reason they are tractable, in that a static, ancestry-matched adjustment can resolve them. Carriers of CpG-eliminating variants, present in any unselected cohort at frequencies set by the variant itself, reduce power to detect a $\Delta\beta$ of 0.05 from 0.93 to 0.13 at sites whose disrupting variant is at 10 to 20% allele frequency. Because the number

of individuals retaining an intact CpG falls as $(1-p)^2$, the supply of informative samples at common disrupting variants is capped by the variant rather than by recruitment. The far larger reservoir of CpG disruption is the rare variants. These do not shift a cohort mean methylation value and will not be adjusted by a static method or by cohort-level quality control; they act one individual at a time, silently assigning a methylation beta near zero and displacing that individual's estimate. Addressing these effects requires per-sample, genotype-aware probe adjustment against a deeply sampled, ancestry-matched whole-genome reference that population-scale sequencing efforts such as Korea10K provide, and one that is unavailable for most populations at present. However, our estimates are conservative, because we counted only variants that destroy the target CpG itself and not those falling within the probe body or at the single-base-extension position, which can perturb hybridization and extension without eliminating the CpG but cannot be separated from true biological variation at the per-probe level with the 1,211 methylomes available here; the figures reported above should therefore be read as a lower bound on total probe perturbation. Together, these results establish CG-eliminating variants as a pervasive and previously underrecognized source of bias in array-based methylation profiling.

Conclusions

Korea10K establishes a high-resolution, population-specific genomic and multiomic reference for the Korean population, enabling near-complete discovery of genetic variation, which will substantially improve imputation and fine-mapping performance. By integrating genomic, molecular, and phenotypic data at population scale, this resource provides a foundation for precision medicine, population genetics, and disease association studies in East Asia. Importantly, the finding that CG-eliminating variants disrupt array probes in every population, while the specific probes affected differ between populations, highlights the need for ethnicity-aware probe design and interpretation of molecular data. Continued expansion through longitudinal phenotyping, long-read sequencing, and integrative multiomic profiling will further enhance its utility, enabling deeper insight into population-specific regulatory architecture, aging, and disease mechanisms.

Methods

Sample collection

Peripheral blood or saliva samples were collected using standardized procedures. Genomic DNA was extracted using the DNeasy Blood & Tissue Kit (Qiagen) for blood and the GeneAll Exgene Clinic SV Mini Kit for saliva, following the manufacturers' protocols. DNA quality and quantity were assessed using Qubit 3.0 Fluorometer (RRID:SCR_020311, Thermo Fisher Scientific) and Bioanalyzer 2100 (RRID:SCR_019715, Agilent) for library preparation.

Whole-genome sequencing

Whole-genome sequencing (WGS) was performed at a mean depth of 30×using MGI DNBSEQ-T7 (RRID:SCR_024847) ($N=6,311$; 61.6%), Illumina NovaSeq 6000 (RRID:SCR_016387) ($N=2,832$; 27.7%), and HiSeq X (RRID:SCR_016385) ($N=1,096$; 10.7%) (Additional file 2: Fig. S10; Additional file 1: Table S5). Early-phase sequencing employed Illumina platforms, whereas large-scale production was performed on the MGI T7 platform for high-throughput efficiency. High-molecular-weight DNA was enzymatically fragmented, and libraries were prepared using platform-specific kits: TruSeq DNA Nano (Illumina) and MGIEasy FS DNA Library Prep (BGI), followed by circularization and rolling circle amplification for DNBSEQ-T7 libraries. Libraries were sequenced as 2×150 bp (or 151 bp for Illumina) paired-end reads.

Read processing and alignment

Raw sequencing reads were trimmed with fastp (RRID:SCR_016962, v0.23.2) [37] to remove platform-specific adapters, polyG tails, and reads with low-quality bases (mean base quality below Q20). Trimmed reads were aligned to the GRCh38 reference genome using BWA-MEM (RRID:SCR_010910, v0.7.17) [38] with read group information. Alignments were coordinate-sorted and duplicates marked with Picard (RRID:SCR_006525, v2.27.5) [39]. Base quality score recalibration was performed using GATK (RRID:SCR_001876, v4.6.1.0) BaseRecalibrator and ApplyBQSR [40].

Variant calling and joint genotyping

Per-sample variant calling was performed using GATK (RRID:SCR_001876, v4.6.1.0) HaplotypeCaller in GVCF mode, adjusting ploidy for sex chromosomes. GVCFs from 10,239 samples were aggregated using GenomicsDBImport and jointly genotyped with GATK (RRID:SCR_001876, v4.6.1.0). Post-calling, variant quality score recalibration (VQSR) was applied separately to SNPs and indels using dbSNP (RRID:SCR_002338, v156), Mills & 1000 G Gold Standard, HapMap, and Omni arrays. Final VCFs were compressed, indexed, and standardized according to VCF v4.2 specifications.

Sample and variant filtering

Joint-called variant call format (VCF) files from 10,239 Korean individuals were subjected to multi-step quality control (QC) to obtain a high-quality, unrelated cohort of 9,000 individuals for downstream analyses. Autosomal and X-chromosomal (chrX) VCFs were first filtered to retain biallelic variants that passed variant quality score recalibration (VQSR). For Y-chromosomal (chrY) VCFs, the VQSR was not applied. All VCFs were converted to PLINK2 format for downstream analyses.

At the sample level, individuals with genotype missingness exceeding 10% or heterozygosity beyond ± 3 standard deviations (SD) from the mean were excluded (205 samples). Heterozygous allele balance (ABHet) was examined to identify samples with excessive allelic imbalance; individuals with mean ABHet outside ± 2 SD were removed (221 samples). Relatedness was assessed using KING (RRID:SCR_009251) from PLINK (RRID:SCR_001757, v2.0.0-a.6.13 LM) [41], and one individual from each pair with kinship coefficients > 0.125 (third-degree relatives or closer) was removed (800 samples). Population analysis using principal component analysis of 9,013 Koreans with the 1000 Genomes Project (1KGP) variome detected non-Korean outliers (13 samples; Methods: Population Structure and Admixture Analyses with EBI's 1KGP Reference).

Variant-level QC was applied to remove variants with: (i) ABHet outside 0.4–0.6, (ii) allele balance in homozygous genotypes $> 10\%$, (iii) $> 10\%$ of genotypes with zero allele depth, (iv) ExcessHet > 60, (v) call rate $< 99\%$, or (vi) Hardy–Weinberg equilibrium (HWE) $P < 10^{-6}$. After applying all filtering steps, 61,223,704 high-confidence biallelic variants were retained across 9,000 independent Korean genomes for downstream analyses. The variants were annotated using Variants Effect Predictor (VEP) with Ensembl database (RRID:SCR_007931, v115) [42].

Variant discovery saturation analysis

To assess the saturation of variant discovery, all 9,000 unrelated individuals were included in each analysis, and the cumulative number of unique variants was recorded as samples were sequentially added. Because cohort heterogeneity, including population substructure and disease-specific variant profiles, can bias saturation estimates depending on the order in which samples are included, we performed 100 independent permutations in which the sample order was randomly shuffled without replacement. For each permutation, the cumulative variant discovery curve was generated, and saturation was defined as the mean point at which all unique variants had been observed across permutations. This permutation-based approach ensures that the saturation estimate is robust to cohort

composition and sample ordering, rather than reflecting artifacts introduced by the inclusion of individuals with atypically high private variant loads or shared genetic backgrounds.

Population reference panel construction and benchmarking imputation performance

A Korean-specific population reference panel was constructed from 9,013 unrelated genomes. Autosomal variants were phased using SHAPEIT5 (v5.1.1, commit = 990ed0d) [43], and imputation was performed with Minimac3 (RRID:SCR_009292, v2.0.1) [44]. Accuracy was evaluated as the squared Pearson correlation (R^2) between imputed dosages and observed genotypes, stratified by allele frequency using the "GetCorrelation" from the TOPMed Imputation evaluation pipeline (https://github.com/sgagliano/UKB_WES_vs_TOPMed_IMP) [45]. For variants exhibiting a negative correlation, the negative sign was retained in the squared correlation value to preserve directional information. Korea10K (this study), Korea4K, and Korea1K reference panels were tested for their imputation performance against 19 unrelated Korean patients with gastric cancers utilized in the previous study [9].

Population structure and admixture analyses with EBI's 1KGP reference

A total of 3,202 high-coverage (30×) whole-genome datasets from the 1000 Genomes Project (1KGP) were retrieved [46]. After applying the same kinship-filtering criteria used for Korea10K (Methods: Sample and Variant Filtering), 2,541 unrelated individuals were retained. These samples were jointly merged with 9,013 Korean genomes. Variant harmonization was performed using reciprocal intersection of SNPs present in both datasets using PLINK (RRID:SCR_001757, v2.0.0-a.6.13 LM) "–extract-intersect", ensuring positional and allelic concordance prior to merging. The two VCFs were then merged using PLINK (RRID:SCR_001757, v1.9.0-b.7.11) "–bmerge" [47], producing a unified dataset containing 15,501,578 unfiltered autosomal variants.

Quality control was applied uniformly across all samples. Variants were excluded if they met any of the following thresholds: (i) minor allele frequency (MAF) < 1%, (ii) deviation from Hardy–Weinberg equilibrium (HWE) with $P < 10^{-6}$, and (iii) missing genotype rate > 1%. After filtering, 4,315,937 variants were retained. To reduce redundancy due to linkage disequilibrium (LD), variants were LD-pruned using a 200 kb sliding window with the parameters –indep-pairwise 200 kb 1 0.5, resulting in 850,281 variants for population structure analysis.

These variants were used for the global principal component analysis. When restricting the merge to the 501 East Asian (EAS) individuals in 1KGP, 876,547 QC-passed and LD-pruned variants were obtained and used for principal component (PC) computation.

For model-based ancestry estimation, ADMIXTURE (RRID:SCR_001263, v1.3.0) [48] was run from K = 1 to K = 17. To reduce computational burden and account for long-range LD, variants were further pruned using –indep-pairwise 500 kb 1 0.2. Identity-by-state clustering was performed using PLINK (RRID:SCR_001757, v1.9.0-b.7.11) "–cluster –K 5", producing 500 genetic clusters. Clusters containing ≤ 2 samples were removed, leaving 485 clusters. For each cluster, PCA was performed on the first five components (PC1-PC5), and the sample closest to the cluster centroid was selected as its representative. These 485 representative samples in Korea10K were merged with 2,541 unrelated 1KGP individuals, yielding 517,041 high-quality, LD-pruned variants for ADMIXTURE analysis. Ten-fold cross-validation was used to determine the optimal K, defined as the model with the lowest cross-validation error (Additional file 2: Fig. S3).

Y-chromosome haplogroup classification and phylogenetic reconstruction

For 5,975 male Korean participants, Y-chromosome variants were lifted to GRCh37 using GATK (RRID:SCR_001876, v4.6.1.0) LiftoverVcf, to matched supported genome coordinate, and then haplogroups were analyzed with yhaplo (v2.1.14.dev7 + gf8b300b) [49] based on the ISOGG 2016 phylogeny. Of these, 104 failed to detect appropriate haplogroups being assigned to A major haplogroup, leaving 5,871 individuals with

successful classification. Of 5,350 unrelated males, 60 failed. A total of 5,290 were used to calculate Y-haplogroup proportion in Korea.

Cross-population comparisons were performed using 1,244 male samples in the 1000 Genomes Project (1KGP) Phase 1 dataset.

For phylogenetic reconstruction, pairwise genetic distances were computed using VCF2Dis (RRID:SCR_022513, v1.54) [50] based on p-distance of VCF containing 5,871 haplogroup-classified males. To assess the robustness of the topology, we generated 100 bootstrap replicates, each constructed by randomly sampling 25% of SNPs. For each replicate, a Newick tree was inferred, and the resulting 100 trees were merged to obtain a final consensus tree. Consensus tree reconstruction and visualization were performed using IQ-TREE (RRID:SCR_017254, v3.0.1) [51].

Mitochondrial DNA assembly and quality control

Filtered mitochondrial reads (chrM) were extracted from whole-genome BAM files using samtools (RRID:SCR_002105, v1.18) and converted to paired-end FASTQ format. Mitochondrial genomes were then assembled using NOVOPlasty (RRID:SCR_017335, v4.3) [52], with parameters optimized for mitochondrial assembly (genome range = 12,000–21,000 bp; k-mer = 39; maximum memory = 20 GB; insert size = 300 bp; read length = 151 bp; Illumina paired-end reads). Ambiguous nucleotides (Reduce ambiguous N's) were reduced, and automatic insert size estimation was enabled.

A total of 10,239 samples were processed, from which 73 assemblies were excluded due to non-circularized sequences or multiple contigs. The final dataset comprised 10,166 high-quality circularized mitochondrial genomes. The rest of the mitochondrial genomes were generated making a consensus genome of the mitochondrial variants to the rCRS reference sequence using bcftools (RRID:SCR_005227, v1.21) "consensus" [53].

Annotation of 10,166 circularized assemblies to check their completeness was performed using MitoZ (v3.5) [54] with the following parameters: "annotate –species_name Homo sapiens –genetic_code 2 –clade Chordata".

Mitochondrial haplogroup classification and phylogenetic reconstruction

Mitochondrial haplogroups were assigned to 10,239 individuals via Haplogrep (v3.2.1) [55] with the PhyloTree Build 17.2 database (RRID:SCR_012948, phylotree- rcrs@17.2).

All 10,166 high-quality circularized mitochondrial genomes were collected together with the revised Cambridge Reference Sequence (rCRS) and the Reconstructed Sapiens Reference Sequence (RSRS) for phylogenetic analysis. To ensure consistent alignment, each sequence was rotated to match the rCRS starting point at the first 30 bp of the *CO2* gene, allowing up to four mismatches by using rotate v1.0 [56]. Multiple sequence alignment was conducted using MAFFT (RRID:SCR_011811, v7.525) [57] with the –auto parameter.

Phylogenetic tree reconstruction was performed using IQ-TREE (RRID:SCR_017254, v3.0.1). An initial model search was conducted with ModelFinder (MFP), and the best-fitting substitution model identified in this step was subsequently applied to the final analysis. The maximum-likelihood tree was inferred under the TIM3 + F + I + R6 model with 1,000 ultrafast bootstrap replicates "-B 1000" and 1,000 SH-aLRT tests "-alrt 1000". All analyses were run using 32 computational threads "-T 32".

Mitochondrial variant processing and principal component analysis

To reveal novel clusters in mitochondrial variants, 4,392 mitochondrial variants from 9,000 unrelated Korean individuals were first normalized against the rCRS reference using bcftools (RRID:SCR_005227, v1.21) "norm -f rCRS.fasta -m -any", yielding 4,990 biallelic variants. Symbolic deletion alleles (*) were removed using bcftools "view -e ALT = *", resulting in 4,830 variants. Single-nucleotide polymorphisms (SNPs) were then extracted

using bcftools "view -v SNP", leaving 4,566 SNPs, which were further filtered to retain sites with minor allele frequency (MAF) ≥ 0.01, yielding 1,247 SNPs for downstream analysis.

Filtered VCF files were converted to PLINK binary format using PLINK (RRID:SCR_001757, v2.0.0-a.6.13 LM) "–make-bed", with variant IDs standardized "–set-all-var-ids @:# –sort-vars" via intermediate PGEN conversion. Principal component analysis was performed using "–pca biallelic-var-wt".

HLA typing from whole-genome sequencing data

High-resolution human leukocyte antigen (HLA) genotyping was performed on 10,239 Korean whole-genome sequencing samples. Prior to typing, BAM files were normalized to ensure consistent alignment within the HLA region, particularly for samples generated with ALT or HLA contigs or non-standard alignment parameters, as advised by xHLA. Following the procedure implemented in get-reads-alt-unmap.sh, reads were extracted using samtools (RRID:SCR_002105, v1.13) [58] by selecting (i) unmapped reads or reads with unmapped mates, (ii) reads mapping to the canonical HLA locus on chromosome 6 (chr6:29,844,528–33,100,696), and (iii) reads mapping to all HLA-associated or ALT contigs (HLA- or chr6 alt-contigs).

Extracted reads were converted to FASTQ and re-aligned with BWA-MEM (RRID:SCR_010910, v0.7.17) using default parameters to the GRCh38 chromosome 6 primary assembly. The resulting BAM was sorted and indexed, and only the chr6:29.8–33.1 Mb region was retained for downstream analysis. HLA genotyping was conducted using xHLA (RRID:SCR_022277). HLA alleles were reported as classical four-digit resolution.

Only 9,000 unrelated individuals (therefore, 18,000 *HLA-A* alleles) were used to calculate the proportions of *HLA-A* alleles in Korean.

Assessment of methylation bias and CpG probe disruption by CG-eliminating variants

We extracted CG-eliminating variant sites from genome-wide methylation beta values from available samples (Table 1). The beta values were calculated as described in a previous publication [32]. Then, mean methylation beta values of each genotype dosage were calculated for each target site. To ensure that the three genotype classes described an identical set of sites, the analysis was restricted to the 15,442 CpG sites with at least ten individuals in every genotype class. For CpG probe disruption, we intersected the genomic coordinates of CpG probe target sites with CG-eliminating variants across four widely used methylation array platforms: Infinium HumanMethylation27 BeadChip, Infinium HumanMethylation450 BeadChip, Infinium MethylationEPIC v2.0, and Infinium Methylation Screening Array. Manifest files of the chips were downloaded from https://github.com/zhou-lab/InfiniumAnnotationData and comprised 26,459, 470,578, 900,431, and 252,770 CpG target sites for the 27 K, 450 K, EPIC v2.0, and MSA arrays, respectively.

Cross-population comparison. The same criteria were applied to the five 1000 Genomes Project super-populations (AFR, AMR, EAS, EUR and SAS) using the Phase 3 high coverage call set. For each cohort a CpG target was counted as eliminated if at least one individual carried a variant destroying that CpG, and eliminated sites were stratified by the allele frequency of the most frequent disrupting variant. Cohort sharing of disrupting variants was summarized as UpSet plots (Additional file 2: Fig. S9).

Power simulation. The effect of CG-eliminating variants on differential-methylation testing was assessed by a spike-in simulation built from the observed beta values, with carriers introduced at each site's Hardy–Weinberg frequency (Additional file 2: Fig. S8). The full procedure, including the cohort construction, the significance threshold, and the constraint on assembling intact-CpG cohorts, is given in Additional file 2: Supplementary Methods.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1186/s13059-026-04321-6.

Acknowledgements We appreciate all participants and Ulsan citizens who supported this project. We thank former members of Korean Genomics Center (KOGIC), namely Hansol Choi, Whan-Hyuk Choi, Muhyeok Kim, Chanyoung Lee, Jeongwoo Seo, Hyeonjae Lee, Seolbin An, Hyeonju Choi, Hyeonsu Cho, Jiyu Lee, Hwan Yu, Gwon Yul Jo, Changjae Kim, Yeonkyung Kim, Younghui Kang, Yeong Ju Woo, Jungae Shim, Nayeong Kim, Shinseob Yoon, Suji Hong, Ju Yeon Park, Boram Park, Sangryoul Han, Yeshin Park, Hak-Min Kim, Suan Cho, and Seung Gu Park who contributed to this project. We thank Moo-Jae Cho, Moo-Young Jung, and Jeongyeon Lee of Ulsan National Institute of Science and Technology (UNIST). We thank Du-Gyeom Kim, Cheol-Ho Song, Ki-Hyun Kim, Jung-Ik Kim, Soo-Jung Shin, Byung-Ryel Yoo, and Jeong-Sik Yang of Ulsan City for their support on this project. We thank Hwanjo Yu of Pohang Institute of Science and Technology for supporting this work. This work was supported by Biodatafarm computing infrastructure funded by the Ulsan metropolitan city government. We thank Seungwoo Jeon, Changwoo Park, and Doyung Oh for supporting the use of Biodatafarm. We thank the Korea Institute of Science and Technology Information (KISTI) for providing us with the Korea Research Environment Open NETwork (KREONET). We thank our collaborators in the National Center for Standard Reference Data (NCSRD) of the Korea Research Institute of Standards and Science (KRISS).

The biospecimens and data used in this study were provided by: the Biobanks of Ulsan University Hospital (60SA2016001-002, 60SA2016001-003, 60SA2016001-005, 60SA2016001-010-5, 60SA2017002-001, 60SA2017002-004, and 60SA2017002-006-4); Ulsan Medical Center; the Biobanks of Gyeongsang National University Hospital (KHUH 2018-11-043); the Biobanks of Kyung Hee University Hospital (2018-4, 2019-4, 2019-6, 2021-2); Chungbuk National University Hospital (18-27, 20-4, 21-3), a member of the National Biobank of Korea, which is supported by the Ministry of Health, Welfare and Family Affairs; the Biobank of Pusan National University Hospital a member of the Korea Biobank Network (19092301-08-01); the support of Borah Kim, MD, Ph, Assistant Professor, Department of Psychiatry, CHA Bundang Medical Center, CHA University (CHAMC 2018-11-004); the Biobank of Severance Hospital, Seoul, Korea. All samples derived from the National Biobank of Korea were obtained with informed consent under institutional review board-approved protocols.

We also acknowledge the use of ChatGPT (OpenAI) and Claude (Anthropic) for assistance with language editing, polishing, and improving the clarity of the manuscript.

Peer review information Alison Monroe was the primary editor of this article and managed its editorial process and peer review in collaboration with the rest of the editorial team. The peer-review history is available in the online version of this article.

Authors' contributions Conceptualization: Kyungwhan An, Sungwon Jeon, Yoonsung Kwon, Neung-Hwa Park, Haeyoung Jeong, George M. Church, Jong Bhak. Methodology: Kyungwhan An, Sungwon Jeon, Yoonsung Kwon, Yookyung Choi, Changhan Yoon, Yeonsu Jeon, Kyun Shik Chae, Chang Geun Kim, Jong Bhak. Software: Kyungwhan An, Sungwon Jeon, Yoonsung Kwon, Yookyung Choi, Changhan Yoon, Yeonsu Jeon. Formal analysis: Kyungwhan An, Sungwon Jeon, Yoonsung Kwon, Yookyung Choi, Changhan Yoon. Investigation: Kyungwhan An, Sungwon Jeon, Yoonsung Kwon, Yookyung Choi, Changhan Yoon, Yeonsu Jeon, Jihun Bhak, Dong-Hyun Shin, Hyoung-Jin Choi, Hyomin Lee, Yeo Jin Kim, Eun-Seok Shin, Hyojung Ryu, Asta Blazyte, Dan Bolser, Sangsoo Park, Juok Cho, Dooyoung Jung, Yun Sung Cho, Hojun Lee, Jong Bhak. Data curation: Hyomin Lee, Eun-Seok Shin, Kiyuk Chang, Eun Ho Choo, Eunmin Kim, Sang Yeub Lee, Weon Kim, Min Gyu Kang, Ae-Young Her, Suk Chon, Jeong-Taek Woo, Sang Youl Rhee, Siwoo Lee, Hee-jeong Jin, Younghwa Baek, Hyo-Jeong Ban, Yong Min Ahn, Sang Jin Rhee, Min Ji Kim, Sang Yeol Lee, Chan-Mo Yang, Se-Hoon Shim, Seong-Jin Cho, Shin Gyeom Kim, Hyung-Tae Jung, Byung-Joo Ham, Yoon Young Choi, Jae-Ho Cheong, Seung-Ki Kim, Ji Hoon Phi, Seung Ah Choi, Heon Yung Gee, Sun Young Joo, Jinsei Jung, Wonsuk Shin, Sang-Hyuk Lee, Borah Kim, Woojae Myung, Chong Kun Cheon, Dong Uk Kim, Seok-Soo Byun, Gangnam Jin, Neung-Hwa Park. Resources: Eun-Seok Shin, Kiyuk Chang, Eun Ho Choo, Eunmin Kim, Sang Yeub Lee, Weon Kim, Min Gyu Kang, Ae-Young Her, Suk Chon, Jeong-Taek Woo, Sang Youl Rhee, Siwoo Lee, Hee-jeong Jin, Younghwa Baek, Hyo-Jeong Ban, Yong Min Ahn, Sang Jin Rhee, Min Ji Kim, Sang Yeol Lee, Chan-Mo Yang, Se-Hoon Shim, Seong-Jin Cho, Shin Gyeom Kim, Hyung-Tae Jung, Byung-Joo Ham, Yoon Young Choi, Jae-Ho Cheong, Seung-Ki Kim, Ji Hoon Phi, Seung Ah Choi, Heon Yung Gee, Sun Young Joo, Jinsei Jung, Wonsuk Shin, Sang-Hyuk Lee, Borah Kim, Woojae Myung, Chong Kun Cheon, Dong Uk Kim, Seok-Soo Byun, Gangnam Jin, Semin Lee, Neung-Hwa Park. Visualization: Kyungwhan An, Sungwon Jeon, Yoonsung Kwon, Yookyung Choi, Jong Bhak. Writing – original draft: Kyungwhan An, Sungwon Jeon, Yoonsung Kwon, Yookyung Choi, Jong Bhak. Writing – review & editing: Asta Blazyte, Juok Cho, Kwangwoo Kim, Bonghee Lee, Jaesuk Lee. Project administration: Sungwon Jeon, Hyomin Lee, Yeo Jin Kim, Eun-Seok Shin, Soobok Joe, Jin Ok Yang, Jongbum Jeon, Jong-Hwan Kim, Jungeun Kim, Yun Sung Cho, Kyun Shik Chae, Chang Geun Kim, Semin Lee, Neung-Hwa Park, Haeyoung Jeong, George M. Church, Jong Bhak. Funding acquisition: Sungwon Jeon, Yeo Jin Kim, Hyojung Ryu, Soobok Joe, Jin Ok Yang, Jongbum Jeon, Jong-Hwan Kim, Jungeun Kim, Yun Sung Cho, Semin Lee, Haeyoung Jeong, Jong Bhak. Supervision: Haeyoung Jeong, George M. Church, Neung-Hwa Park, Jong Bhak. All authors read and approved the final manuscript.

Funding This work was supported by the U-K BRAND Research Fund (1.200108.01) of UNIST (Ulsan National Institute of Science & Technology), the Research Project Funded by Ulsan City Research Fund (1.200047.01) of UNIST, the Promotion

of Innovative Businesses for Regulation-Free Special Zones funded by the Ministry of SMEs and Startups (MSS, Korea) (1425157301 and 1425156792), the Establishment of Demonstration Infrastructure for Regulation-Free Special Zones funded by the Ministry of SMEs and Startups (MSS, Korea) (1425157253), and the Technology Innovation Program (20016225, Development and Dissemination on National Standard Reference Data) funded by the Ministry of Trade, Industry & Energy (MOTIE, Korea). This work was also supported by funding from the Korea Planning & Evaluation Institute of Industrial Technology with support from the Ministry of Trade, Industry and Energy in 2024 [RS-2024-00435468, Development and Dissemination of National Standard Technology]. This work was supported by the Korea Innovation Foundation (INNOPOLIS) grant funded by the Korean government (MSIT) (RS-2025-25449754). This research was supported by the BioBigData. Korea grant "Genomic and Other Omics Data Production and Analysis" (RS-2024-00438566), funded by four ministries of Korea: the Ministry of Health and Welfare, the Ministry of Science and ICT, the Ministry of Trade, Industry and Energy, and the Korea Disease Control and Prevention Agency.

Data availability The source codes used to generate data and calculate statistics are openly available under the MIT License on the GitHub page: https://github.com/korean-genomics-center/korea10k or https://doi.org/10.5281/zenodo.23062236 [59, 60]. The variant frequency data, HLA typing summary, and haplogroup data of Korea10K are freely available via https://doi.org/10.5281/zenodo.21767453 [61]. The data that pose no privacy concerns from this study are available under the accession number (BioProject) KAP242495 at the Korea BioData Station (K-BDS, https://kbds.re.kr) [62]. Access to individual-level full datasets requires approval from the relevant Institutional Review Board (IRB) and Data Access Committee (DAC), in accordance with international guidelines for human genomic data sharing. Access to raw sequencing and processed data can be granted upon request and is subject to data-use agreements. Researchers seeking access should contact corresponding authors and submit the following information and documentation: the principal investigator's institutional affiliation, position, and field of research; a research proposal describing the intended use of the Korea10K data; an institutional review board (IRB) approval letter or documentation of exemption for the proposed study; and a description of the types and scope of data requested. Applications are evaluated by the IRB of Ulsan National Institute of Science and Technology (UNIST). Data may be provided only for participants who have consented to secondary data sharing. The types and scope of data that can be provided, as well as their permitted use, are limited by participant consent and the approved research purpose. In accordance with applicable Korean national guidelines, access to individual-level whole-genome sequencing (WGS) data by commercial entities or institutions outside the Republic of Korea may require additional consultation with the corresponding author. Researchers from such organizations should contact the corresponding author to discuss the requirements and arrangements for data access. Upon receipt of an application, and where required, the request undergoes a data-sharing review by the UNIST IRB before a final decision on data release is made. The UNIST IRB review typically takes approximately 4–6 weeks, depending on the application and the IRB review schedule. This timeframe refers to the IRB review; the overall time required to complete the data access process may vary. Any dataset provided is subject to restriction or withdrawal in the future in response to policy updates or regulatory requirements.

The human genome reference (hg38) file used in this study can be downloaded from the GATK resource bundle [63]. High-coverage data from the 1000 Genome Project were downloaded from ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/1000G_2504_high_coverage [64]. The Korea1K data and epigenome data were presented in the authors' previous studies [9, 32] and are also available upon request to the authors directly.

Declarations

Ethics approval and consent to participate Sample collection and sequencing were approved by the Institutional Review Board (IRB) of the Ulsan National Institute of Science and Technology (UNISTIRB-15-19-A and UNISTIRB-16-13-C). The data employed in our study originates from voluntary blood or saliva donations, and we have diligently secured explicit, comprehensive consent forms from the participants prior to sample collection. These consent forms explicitly underscore the voluntary nature of their participation. Our study adheres to the ethical guidelines and regulations stipulated by the IRB. All experimental methods complied with the Declaration of Helsinki.

Consent for publication The consent forms explicitly state that participants' data may be used for research purposes and published as part of the resulting research.

Competing interests SJ is CEO, JB is founder, and HR is an employee of AgingLab Inc, Ulsan, Republic of Korea. YCho is an employee of CG Invites Co., LTD, Seoul, Republic of Korea, and Invites Genomics Co., LTD, Jeju, Republic of Korea. B.L. is CEO, and JL is an employee of nSAGE, Incheon, Republic of Korea. GC is a co-founder of Nebula Genomics, San Francisco, CA, USA, and declares conflicts of interest via fees and equity listed at www.glottatech.com, Technology Transfer section. All other authors declare no competing interests.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

1. Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alfoldi J, Wang Q, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. Nature. 2020;581:434–43.
2. Consortium UKBW-GS. Whole-genome sequencing of 490,640 UK Biobank participants. Nature. 2025;645:692–701.
3. Kurki MI, Karjalainen J, Palta P, Sipila TP, Kristiansson K, Donner KM, et al. FinnGen provides genetic insights from a well-phenotyped isolated population. Nature. 2023;613:508–18.
4. Feng YA, Chen CY, Chen TT, Kuo PH, Hsu YH, Yang HI, et al. Taiwan Biobank: a rich biomedical research database of the Taiwanese population. Cell Genom. 2022;2:100197.
5. Nagai A, Hirata M, Kamatani Y, Muto K, Matsuda K, Kiyohara Y, et al. Overview of the BioBank Japan project: study design and profile. J Epidemiol. 2017;27:S2-8.
6. 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.
7. Bellis C, Kolle G, Yong J, Hebrard M, Bertin N, Alexandrine Lin BC, Koh TH, Cheng PC, Sunmugam SDO, Heng KK, et al. National scale genomic engine for precision medicine: Singapore PRECISE-SG100K experience. bioRxiv. 2025:2025.2003.2013.642552.
8. Kim J, Weber JA, Jho S, Jang J, Jun J, Cho YS, et al. KoVariome: Korean national standard reference variome database of whole genomes with comprehensive SNV, indel, CNV, and SV analyses. Sci Rep. 2018;8:5677.
9. 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.
10. Lee J, Lee J, Jeon S, Lee J, Jang I, Yang JO, et al. A database of 5305 healthy Korean individuals reveals genetic and clinical implications for an East Asian population. Exp Mol Med. 2022;54:1862–71.
11. 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. https://doi.org/10.1093/gigascience/giae014.
12. Cho YS, Kim H, Kim HM, Jho S, Jun J, Lee YJ, et al. An ethnically relevant consensus Korean reference genome is a step towards personal reference genomes. Nat Commun. 2016;7:13637.
13. Kim SH, Kim KC, Shin DJ, Jin HJ, Kwak KD, Han MS, et al. High frequencies of Y-chromosome haplogroup O2b-SRY465 lineages in Korea: a genetic perspective on the peopling of Korea. Investig Genet. 2011;2:10.
14. Kim J, Jeon S, Choi JP, Blazyte A, Jeon Y, Kim JI, et al. The origin and composition of Korean ethnicity analyzed by ancient and present-day genome sequences. Genome Biol Evol. 2020;12:553–65.
15. Horvath S. DNA methylation age of human tissues and cell types. Genome Biol. 2013;14:R115.
16. Horvath S, Oshima J, Martin GM, Lu AT, Quach A, Cohen H, et al. Epigenetic clock for skin and blood cells applied to Hutchinson Gilford Progeria Syndrome and ex vivo studies. Aging (Albany NY). 2018;10:1758–75.
17. Hannum G, Guinney J, Zhao L, Zhang L, Hughes G, Sadda S, et al. Genome-wide methylation profiles reveal quantitative views of human aging rates. Mol Cell. 2013;49:359–67.
18. Levine ME, Lu AT, Quach A, Chen BH, Assimes TL, Bandinelli S, et al. An epigenetic biomarker of aging for lifespan and healthspan. Aging (Albany NY). 2018;10:573–91.
19. Belsky DW, Caspi A, Corcoran DL, Sugden K, Poulton R, Arseneault L, et al. DunedinPACE, a DNA methylation biomarker of the pace of aging. Elife. 2022. https://doi.org/10.7554/eLife.73420.
20. Lu AT, Quach A, Wilson JG, Reiner AP, Aviv A, Raj K, et al. DNA methylation GrimAge strongly predicts lifespan and healthspan. Aging (Albany NY). 2019;11:303–27.
21. Kim J, Park JL, Yang JO, Kim S, Joe S, Park G, et al. Highly accurate Korean draft genomes reveal structural variation highlighting human telomere evolution. Nucleic Acids Res. 2025. https://doi.org/10.1093/nar/gkae1294.
22. Liao WW, Asri M, Ebler J, Doerr D, Haukness M, Hickey G, et al. A draft human pangenome reference. Nature. 2023;617:312–24.
23. Gusev A, Ko A, Shi H, Bhatia G, Chung W, Penninx BW, et al. Integrative approaches for large-scale transcriptome-wide association studies. Nat Genet. 2016;48:245–52.
24. Gate RE, Cheng CS, Aiden AP, Siba A, Tabaka M, Lituiev D, et al. Genetic determinants of co-accessible chromatin regions in activated T cells across humans. Nat Genet. 2018;50:1140–50.

25. Banovich NE, Lan X, McVicker G, van de Geijn B, Degner JF, Blischak JD, et al. Methylation QTLs are associated with coordinated changes in transcription factor binding, histone modifications, and gene expression levels. PLoS Genet. 2014;10:e1004663.
26. Crocchiolo R, Gallina AM, Pani A, Campisi D, Cento V, Sacchi N, et al. Polymorphism of the HLA system and weak antibody response to BNT162b2 mRNA vaccine. HLA. 2022;99:183–91.
27. Kang EH, Kim JY, Takeuchi F, Kim JW, Shin K, Lee EY, et al. Associations between the HLA-A polymorphism and the clinical manifestations of Behcet's disease. Arthritis Res Ther. 2011;13:R49.
28. Li X, Zhao Z, Sun SS. Association of human leukocyte antigen variants and allopurinol-induced Stevens-Johnson syndrome and toxic epidermal necrolysis: a meta-analysis. Am J Health Syst Pharm. 2017;74:e183–92.
29. Bhak Y, Jeon Y, Jeon S, Yoon C, Kim M, Blazyte A, et al. Polygenic risk score validation using Korean genomes of 265 early-onset acute myocardial infarction patients and 636 healthy controls. PLoS ONE. 2021;16:e0246538.
30. Jeon Y, Jeon S, Choi WH, An K, Choi H, Kim BC, et al. Genome-wide analyses of early-onset acute myocardial infarction identify 29 novel loci by whole genome sequencing. Hum Genet. 2023;142:231–43.
31. Jeon Y, Jeon S, An K, Kim YJ, Kim BC, Ryu H, et al. Identification and validation of six acute myocardial infarction-associated variants, including a novel prognostic marker for cardiac mortality. Front Cardiovasc Med. 2023;10:1226971.
32. Jeon Y, Kwon Y, Kim YJ, Jeon S, Ryu H, An K, et al. Epigenetic modulation of vascular smooth muscle cell phenotype switching in early-onset acute myocardial infarction. Arterioscler Thromb Vasc Biol. 2025;45:e217–30.
33. Bhak Y, Jeong HO, Cho YS, Jeon S, Cho J, Gim JA, et al. Depression and suicide risk prediction models using blood-derived multi-omics data. Transl Psychiatry. 2019;9:262.
34. Kwon Y, Blazyte A, Jeon Y, Kim YJ, An K, Jeon S, et al. Identification of 17 novel epigenetic biomarkers associated with anxiety disorders using differential methylation analysis followed by machine learning-based validation. Clin Epigenetics. 2025;17:24.
35. Ryu H, An K, Kwon Y, Jeon Y, Jeon S, Choi H, et al. Identification of reactive CpGs and RNA expression in early COVID-19 through cis-eQTM analysis reflecting disease severity and recovery. Commun Biol. 2025;8:1174.
36. An K, Kwon Y, Bhak J, Ryu H, Jeon S, Nam D, et al. Dynamic and reversible transcriptomic age shifts induced by COVID-19 in Korean whole blood. Aging (Albany NY). 2025;17:1484–510.
37. Chen S, Zhou Y, Chen Y, Gu J. Fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics. 2018;34:i884–90.
38. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25:1754–60.
39. Picard Toolkit. https://broadinstitute.github.io/picard/.
40. Poplin R, Ruano-Rubio V, DePristo MA, Fennell TJ, Carneiro MO, Van der Auwera GA, Kling DE, Gauthier LD, Levy-Moonshine A, Roazen D, et al. Scaling accurate genetic variant discovery to tens of thousands of samples. bioRxiv. 2018:201178.
41. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience. 2015;4:7.
42. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GR, Thormann A, et al. The Ensembl variant effect predictor. Genome Biol. 2016;17:122.
43. 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.
44. Das S, Forer L, Schonherr S, Sidore C, Locke AE, Kwong A, et al. Next-generation genotype imputation service and methods. Nat Genet. 2016;48:1284–7.
45. 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.
46. Genomes Project C, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, et al. A global reference for human genetic variation. Nature. 2015;526:68–74.
47. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007;81:559–75.
48. Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. Genome Res. 2009;19:1655–64.
49. Poznik GD. Identifying Y-chromosome haplogroups in arbitrarily large samples of sequenced or genotyped men. bioRxiv. 2016:088716.
50. Xu L, He W, Tai S, Huang X, Qin M, Liao X, et al. VCF2Dis: an ultra-fast and efficient tool to calculate pairwise genetic distance and construct population phylogeny from VCF files. Gigascience. 2025. https://doi.org/10.1093/gigascience/giaf032.
51. Nguyen LT, Schmidt HA, von Haeseler A, Minh BQ. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol Biol Evol. 2015;32:268–74.
52. Dierckxsens N, Mardulyn P, Smits G. NOVOPlasty: de novo assembly of organelle genomes from whole genome data. Nucleic Acids Res. 2017;45:e18.

53. Bonfield JK, Marshall J, Danecek P, Li H, Ohan V, Whitwham A, et al. HTSlib: C library for reading/writing high-throughput sequencing data. Gigascience. 2021. https://doi.org/10.1093/gigascience/giab007.
54. Meng G, Li Y, Yang C, Liu S. MitoZ: a toolkit for animal mitochondrial genome assembly, annotation and visualization. Nucleic Acids Res. 2019;47:e63.
55. Schonherr S, Weissensteiner H, Kronenberg F, Forer L. Haplogrep 3 - an interactive haplogroup classification and analysis platform. Nucleic Acids Res. 2023;51:W263–8.
56. Durbin R, De Sanctis B, Blumer M. Rotate: a command-line program to rotate circular DNA sequences to start at a given position or string. Wellcome Open Res. 2023;8:401.
57. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol. 2013;30:772–80.
58. Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, et al. Twelve years of SAMtools and BCFtools. Gigascience. 2021. https://doi.org/10.1093/gigascience/giab008.
59. Jeon S. Korea10K: analysis and figure-generation code. 2026. Zenodo. https://doi.org/10.5281/zenodo.23062236.
60. Jeon S: Korea10K: analysis and figure-generation code. 2026. GitHub.https://github.com/korean-genomics-center/korea10k.
61. Bhak J. Korea10K allele frequency information. 2026. Datasets. https://doi.org/10.5281/zenodo.21767453.
62. Bhak J: Korea10K Genome Project. Datasets. Korea BioData Station; 2026.
63. GATK Team: Resource bundle. 2026. https://gatk.broadinstitute.org/hc/en-us/articles/360035890811-Resource-bundle.
64. The 1000 Genome Project consortium IGSR: The International Genome Sample Resource. 2025. ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/1000G_2504_high_coverage.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Authors and Affiliations

Kyungwhan An[1,2,3] · Sungwon Jeon[1,4] · Yoonsung Kwon[1,2,5] · Yookyung Choi[1,2,5] · Changhan Yoon[1,2] · Yeonsu Jeon[5] · Jihun Bhak[1,2] · Dong-Hyun Shin[1,2] · Hyoung-Jin Choi[1,2] · Hyomin Lee[1,2] · Yeo Jin Kim[1] · Eun-Seok Shin[6] · Hyojung Ryu[1,4] · Asta Blazyte[1,7] · Sangsoo Park[1,2] · Juok Cho[1,2] · Dan Bolser[8] · Soobok Joe[5] · Jin Ok Yang[5] · Jongbum Jeon[5,9] · Jong-Hwan Kim[5,9] · Jungeun Kim[10] · Dooyoung Jung[2,11] · Yun Sung Cho[12,13] · Kiyuk Chang[14,15] · Eun Ho Choo[14,15] · Eunmin Kim[14] · Sang Yeub Lee[16] · Weon Kim[17] · Min Gyu Kang[18] · Ae-Young Her[19] · Suk Chon[20] · Jeong-Taek Woo[20] · Sang Youl Rhee[20] · Siwoo Lee[21] · Hyo-Jeong Ban[22] · Hee-Jeong Jin[22] · Younghwa Baek[22] · Yong Min Ahn[23,24,25] · Sang Jin Rhee[23,24] · Min Ji Kim[26] · Sang Yeol Lee[27] · Chan-Mo Yang[27] · Se-Hoon Shim[28] · Seong-Jin Cho[29] · Shin Gyeom Kim[30] · Hyung-Tae Jung[31] · Byung-Joo Ham[32,33,34] · Yoon Young Choi[35] · Jae-Ho Cheong[36] · Seung-Ki Kim[37,38] · Ji Hoon Phi[37,38] · Seung Ah Choi[37,38] · Heon Yung Gee[39] · Sun Young Joo[39] · Jinsei Jung[40] · Wonsuk Shin[41,42] · Sang-Hyuk Lee[41,43] · Borah Kim[43,44] · Woojae Myung[24,45] · Chong Kun Cheon[46] · Dong Uk Kim[47] · Seok-Soo Byun[48,49] · Gangnam Jin[50] · Hojun Lee[5,50] · Kyun Shik Chae[51] · Chang Geun Kim[51] · Bonghee Lee[7,52] · Jaesuk Lee[52] · Kwangwoo Kim[53,54] · Semin Lee[1,2] · Neung-Hwa Park[55,56] · Haeyoung Jeong[5] · George M. Church[57,58,59] · Jong Bhak[1,2,4]

✉ Neung-Hwa Park
nhparkmd@gmail.com

✉ Haeyoung Jeong
hyjeong@kribb.re.kr

✉ George M. Church
gchurch@genetics.med.harvard.edu

✉ Jong Bhak
jongbhak@genomics.org; jongbhak@unist.ac.kr

Kyungwhan An
micolak0115@postech.ac.kr

Sungwon Jeon
jsw0061@gmail.com

Yoonsung Kwon
ponyo1203@unist.ac.kr

Yookyung Choi
chldbrud1230@gmail.com

Changhan Yoon
changhan1110@gmail.com

Yeonsu Jeon
brain0106@gmail.com

Jihun Bhak
jihun147278@gmail.com

Dong-Hyun Shin
sdh9t1@unist.ac.kr

Hyoung-Jin Choi
hjchoi349@gmail.com

Hyomin Lee
hyomin@unist.ac.kr

Yeo Jin Kim
yjin.kim09@gmail.com

Eun-Seok Shin
sesim1989@gmail.com

Hyojung Ryu
hjryu3094@gmail.com

Asta Blazyte
blazyte.asta@gmail.com

Sangsoo Park
s4ngsoo@unist.ac.kr

Juok Cho
juokcho@gmail.com

Dan Bolser
dan.geromics@gmail.com

Soobok Joe
soobok@kribb.re.kr

Jin Ok Yang
joy@kribb.re.kr

Jongbum Jeon
jongbumj@kribb.re.kr

Jong-Hwan Kim
kkjjhhk@kribb.re.kr

Jungeun Kim
jungeunkim079@gmail.com

Dooyoung Jung
dooyoung@unist.ac.kr

Yun Sung Cho
yunsung.cho@inviteseco.com

Kiyuk Chang
kiyuk@catholic.ac.kr

Eun Ho Choo
cmcchu@catholic.ac.kr

Eunmin Kim
kimeunmin@gmail.com

Sang Yeub Lee
louisahj@gmail.com

Weon Kim
mylovekw@hanmail.net

Min Gyu Kang
med2floyd@naver.com

Ae-Young Her
hermartha1@gmail.com

Suk Chon
imdrjs@khu.ac.kr

Jeong-Taek Woo
jtwoo@khu.ac.kr

Sang Youl Rhee
bard95@hanmail.net

Siwoo Lee
bard95@hanmail.net

Hyo-Jeong Ban
hjban@kiom.re.kr

Hee-Jeong Jin
hjjin@kiom.re.kr

Younghwa Baek
aori79@kiom.re.kr

Yong Min Ahn
aym@snu.ac.kr

Sang Jin Rhee
hellojr1123@hanmail.net

Min Ji Kim
demonic89@naver.com

Sang Yeol Lee
psysangyeol@hanmail.net

Chan-Mo Yang
ychanmo@wku.ac.kr

Se-Hoon Shim
shshim2k@daum.net

Seong-Jin Cho
sjcho@gilhospital.com

Shin Gyeom Kim
redmensch@schmc.ac.kr

Hyung-Tae Jung
lukie004@naver.com

Byung-Joo Ham
byungjoo.ham@gmail.com

Yoon Young Choi
laki98@naver.com

Jae-Ho Cheong
jhcheong@yuhs.ac

Seung-Ki Kim
nsthomas@snu.ac.kr

Ji Hoon Phi
phi.jihoon@gmail.com

Seung Ah Choi
aiippo7@gmail.com

Heon Yung Gee
hygee@yuhs.ac

Sun Young Joo
kin2844@yuhs.ac

Jinsei Jung
jsjung@yuhs.ac

Wonsuk Shin
wonsug89@chamc.co.kr

Sang-Hyuk Lee
drshlee@cha.ac.kr

Borah Kim
borahkim27@gmail.com

Woojae Myung
wmyung@snu.ac.kr

Chong Kun Cheon
chongkun@pusan.ac.kr

Dong Uk Kim
amlm374@gmail.com

Seok-Soo Byun
seoksoo.byeon@gmail.com

Gangnam Jin
morrison@korea.kr

Hojun Lee
ehojune@gmail.com

Kyun Shik Chae
cks@kriss.re.kr

Chang Geun Kim
cgkim@kriss.re.kr

Bonghee Lee
bhlee@gachon.ac.kr

Jaesuk Lee
jaesuklee88@gmail.com

Kwangwoo Kim
kkim@khu.ac.kr

Semin Lee
seminlee@unist.ac.kr

[1] Korean Genomics Center (KOGIC), Ulsan National Institute of Science and Technology (UNIST), Ulsan 44919, Republic of Korea

[2] Department of Biomedical Engineering, College of Information-Bio Convergence Engineering, Ulsan National Institute of Science and Technology (UNIST), Ulsan 44919, Republic of Korea

[3] Graduate School of Artificial Intelligence, Pohang University of Science and Technology (POSTECH), Pohang 37673, Republic of Korea

[4] Aging Lab Inc., Ulsan 44919, Republic of Korea

[5] Institute of Bioscience and Biotechnology, Korea Bioinformation Center (KOBIC), Korea Research, Daejeon 34141, Republic of Korea

[6] Department of Cardiology, Ulsan University Hospital, University of Ulsan College of Medicine, Ulsan 44033, Republic of Korea

[7] School of Medicine, Gachon University, Yeonsu-Gu, Incheon 21565, Republic of Korea

[8] Geromics Ltd., 222 Mill Road, Cambridge CB1 3NF, UK

[9] Department of Bioinformatics, KRIBB School of Bioscience, University of Science and Technology (UST), Daejeon 34141, Republic of Korea

[10] Personal Genomics Institute, Genome Research Foundation, Cheongju 28190, Republic of Korea

[11] Graduate School of Health Science and Technology, Ulsan National Institute of Science and Technology (UNIST), Ulsan 44919, Republic of Korea

[12] CG Invites Co., LTD, Seoul 07802, Republic of Korea

[13] Invites Genomics Co., LTD, Jeju 63243, Republic of Korea

[14] Cardiovascular Research Institute for Intractable Disease, College of Medicine, The Catholic University of Korea, Seoul 06591, Republic of Korea

[15] Division of Cardiology, College of Medicine, Seoul St. Mary's Hospital, The Catholic University of Korea, Seoul, Republic of Korea

[16] Division of Cardiology, Chung-Ang University Gwangmyeong Hospital and Department of Internal Medicine, Chung-Ang University College of Medicine, Seoul, Republic of Korea

[17] Division of Cardiology, Department of Internal Medicine, Kyung Hee University Hospital, Seoul, Republic of Korea

[18] Department of Internal Medicine, Gyeongsang National University Hospital, Gyeongsang National University School of Medicine, Jinju, Korea

[19] Division of Cardiology, Department of Internal Medicine, Kangwon National University School of Medicine, Chuncheon, Republic of Korea

[20] Department of Endocrinology and Metabolism, Kyung Hee University School of Medicine, Seoul 02447, Korea

[21] College of Korean Medicine, Gachon University, Seongnam 13120, Republic of Korea

[22] KM Data Division, Korea Institute of Oriental Medicine, Daejeon 90703, Republic of Korea

[23] Department of Neuropsychiatry, Seoul National University Hospital, Seoul 03080, Republic of Korea

[24] Department of Psychiatry, Seoul National University College of Medicine, Seoul 03080, Republic of Korea

[25] Institute of Human Behavioral Medicine, Seoul National University College of Medicine, Seoul 03080, Republic of Korea

[26] Department of Psychiatry, Columbia University, New York, NY, USA

[27] Department of Psychiatry, Wonkwang University Hospital, Iksan, Republic of Korea

[28] Department of Psychiatry, Soon Chun Hyang University Cheonan Hospital, Soon Chun Hyang University, Cheonan, Republic of Korea

[29] Department of Psychiatry, Gachon University Gil Medical Center, Incheon, Republic of Korea

[30] Department of Neuropsychiatry, Soon Chun Hyang University Bucheon Hospital, Bucheon, Republic of Korea

[31] Department of Psychiatry, Ulsan Medical Center, Ulsan 44686, Republic of Korea

[32] Department of Biomedical Sciences, Korea University College of Medicine, Seoul, Republic of Korea

[33] Department of Psychiatry, Korea University Anam Hospital, Korea University College of Medicine, Seoul, Republic of Korea

[34] Brain Convergence Research Center, Korea University Anam Hospital, Seoul, Republic of Korea

[35] Department of Surgery, Soonchunhyang Bucheon Hospital, Soonchunhyang University College of Medicine, Bucheon 14584, Republic of Korea

[36] Department of Biomedical Systems Informatics, Yonsei University College of Medicine, 250 Seongsanno Seodaemun-Gu, Seoul 120-752, Korea

[37] Division of Pediatric Neurosurgery, Seoul National University Children's Hospital, Seoul, Korea

[38] Neuroscience Research Institute, Seoul National University College of Medicine, Seoul, Korea

[39] Department of Pharmacology, Brain Korea 21 PLUS Project for Medical Sciences, Yonsei University College of Medicine, Seoul 03722, Republic of Korea

[40] Department of Otorhinolaryngology, Yonsei University College of Medicine, Seoul 03722, Republic of Korea

[41] Department of Clinical Pharmacology and Therapeutics, CHA University Bundang Medical Center, Seongnam-Si, Gyeonggi-Do 13496, Republic of Korea

[42] Department of Physiology, School of Medicine, CHA University, Pocheon-Si, Gyeonggi-Do 11160, Republic of Korea

[43] Department of Psychiatry, CHA University Bundang Medical Center, Seongnam-Si, Gyeonggi-Do 13496, Republic of Korea

[44] CHA Hollywood Presbyterian Medical Center, Los Angeles, CA 90027, USA

[45] Department of Neuropsychiatry, Seoul National University Bundang Hospital, Seongnam, South Korea

[46] Division of Medical Genetics, Department of Pediatrics, Pusan National University School of Medicine, Pusan National University Children's Hospital, Yangsan 50612, Republic of Korea

[47] Department of Internal Medicine, College of Medicine and Medical Research Institute, Pusan National University, Pusan National University Hospital, Busan 49241, Republic of Korea

[48] Department of Urology, Seoul National University Bundang Hospital, Seongnam 13620, Republic of Korea

[49] Department of Medical Device Development, Seoul National University College of Medicine, Seoul 03080, Republic of Korea

[50] Forensic DNA Division, National Forensic Service, Seoul 08036, Republic of Korea

[51] National Standard Reference Center, Korea Research Institute of Standards and Science (KRISS), Daejeon 34113, Republic of Korea

[52] nSAGE, Incheon 21999, Republic of Korea

[53] Department of Biology, Kyung Hee University, Seoul, Republic of Korea

[54] Department of Biomedical and Pharmaceutical Sciences, Kyung Hee University, Seoul, Republic of Korea

[55] Department of Internal Medicine, University of Ulsan College of Medicine, Ulsan University Hospital, Ulsan 44033, Republic of Korea

[56] Biomedical Research Center, University of Ulsan College of Medicine, Ulsan University Hospital, Ulsan 44033, Republic of Korea

[57] Wyss Institute for Biologically Inspired Engineering, Harvard University, Boston, MA 02115, USA

[58] Department of Genetics, Harvard Medical School, Boston, MA 02115, USA

[59] Harvard-MIT Program in Health Sciences and Technology, Cambridge, MA 02139, USA