Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Korea10K: 10,239 whole genomes with multiomic and clinical health information as the Korean multiomics reference dataset

Kyungwhan An
Korean Genomics Center (KOGIC), Ulsan National Institute of Science and Technology (UNIST)
Sungwon Jeon
Korean Genomics Center (KOGIC), Ulsan National Institute of Science and Technology (UNIST)
Yoonsung Kwon
Korean Genomics Center (KOGIC), Ulsan National Institute of Science and Technology (UNIST)
Yookyung Choi
Korean Genomics Center (KOGIC), Ulsan National Institute of Science and Technology (UNIST)
Changhan Yoon
Korean Genomics Center (KOGIC), Ulsan National Institute of Science and Technology (UNIST)
Yeonsu Jeon
Korea Bioinformation Center (KOBIC), Korea Research Institute of Bioscience and Biotechnology
Jihun Bhak
Korean Genomics Center (KOGIC), Ulsan National Institute of Science and Technology (UNIST)
Dong-Hyun Shin
Korean Genomics Center (KOGIC), Ulsan National Institute of Science and Technology (UNIST)
Hyoung-Jin Choi
Korean Genomics Center (KOGIC), Ulsan National Institute of Science and Technology (UNIST)
Hyomin Lee
Korean Genomics Center (KOGIC), Ulsan National Institute of Science and Technology (UNIST)
Yeo Jin Kim
Korean Genomics Center (KOGIC), Ulsan National Institute of Science and Technology (UNIST)
Eun-Seok Shin
Ulsan University Hospital, University of Ulsan College of Medicine
Hyojung Ryu
Korean Genomics Center (KOGIC), Ulsan National Institute of Science and Technology (UNIST)
Asta Blazyte
Korean Genomics Center (KOGIC), Ulsan National Institute of Science and Technology (UNIST)
Dan Bolser
Geromics Ltd
Sangsoo Park

Korean Genomics Center (KOGIC), Ulsan National Institute of Science and Technology (UNIST)

Juok Cho
Korean Genomics Center (KOGIC), Ulsan National Institute of Science and Technology (UNIST)

Soobok Joe
Korea Bioinformation Center (KOBIC), Korea Research Institute of Bioscience and Biotechnology

Jin Ok Yang
Korea Bioinformation Center (KOBIC), Korea Research Institute of Bioscience and Biotechnology

Jongbum Jeon
Korea Bioinformation Center (KOBIC), Korea Research Institute of Bioscience and Biotechnology

Jong-Hwan Kim
Korea Bioinformation Center (KOBIC), Korea Research Institute of Bioscience and Biotechnology

Jungeun Kim
Personal Genomics Institute, Genome Research Foundation

Dooyoung Jung
Ulsan National Institute of Science and Technology (UNIST)

Yun Sung Cho
CG Invites Co., LTD

Kiyuk Chang
The Catholic University of Korea

Eun Ho Choo
The Catholic University of Korea

Eunmin Kim
The Catholic University of Korea

Sang Yeub Lee
Chung-Ang University Gwangmyeong Hospital, Chung-Ang University College of Medicine

Weon Kim
Kyung Hee University Hospital

Min Gyu Kang
Gyeongsang National University Hospital, Gyeongsang National University School of Medicine

Ae-Young Her
Kangwon National University School of Medicine

Suk Chon
Kyung Hee University School of Medicine

Jeong-Taek Woo
Kyung Hee University School of Medicine

Sang Youl Rhee
Kyung Hee University School of Medicine

Siwoo Lee
Gachon University

Hee-Jeong Jin
Korea Institute of Oriental Medicine
Younghwa Baek
Korea Institute of Oriental Medicine
Hyo-Jeong Ban
Korea Institute of Oriental Medicine
Yong Min Ahn
Seoul National University Hospital
Sang Jin Rhee
Seoul National University Hospital
Min Ji Kim
Columbia University
Sang Yeol Lee
Wonkwang University Hospital
Chan-Mo Yang
Wonkwang University Hospital
Se-Hoon Shim
Soon Chun Hyang University Cheonan Hospital, Soon Chun Hyang University
Seong-Jin Cho
Gachon University Gil Medical Center
Shin Gyeom Kim
Soon Chun Hyang University Bucheon Hospital
Hyung-Tae Jung
Ulsan Medical Center
Byung-Joo Ham
Korea University College of Medicine
Yoon Young Choi
Soonchunhyang Bucheon Hospital, Soonchunhyang University College of Medicine
Jae-Ho Cheong
Yonsei University College of Medicine
Seung-Ki Kim
Seoul National University Children's Hospital
Ji Hoon Phi
Seoul National University Children's Hospital
Seung Ah Choi
Seoul National University Children's Hospital
Heon Yung Gee
Yonsei University College of Medicine
Sun Young Joo

Yonsei University College of Medicine

Jinsei Jung
Yonsei University College of Medicine

Wonsuk Shin
CHA University Bundang Medical Center

Sang-Hyuk Lee
CHA University Bundang Medical Center

Borah Kim
CHA University Bundang Medical Center

Woojae Myung
Seoul National University College of Medicine

Chong Kun Cheon
Pusan National University Children's Hospital

Dong Uk Kim
Pusan National University Hospital

Seok-Soo Byun
Seoul National University Bundang Hospital

Gangnam Jin
National Forensic Service

Hojun Lee
National Forensic Service

Kyun Shik Chae
Korea Research Institute of Standards and Science (KRISS)

Chang Geun Kim
Korea Research Institute of Standards and Science (KRISS)

Bonghee Lee
nSAGE

Jaesuk Lee
nSAGE

Kwangwoo Kim
Kyung Hee University

Semin Lee
Korean Genomics Center (KOGIC), Ulsan National Institute of Science and Technology (UNIST)

Neung-Hwa Park
Ulsan University Hospital, University of Ulsan College of Medicine

Haeyoung Jeong
Korea Bioinformation Center (KOBIC), Korea Research Institute of Bioscience and Biotechnology

George M. Church
Harvard University

Jong Bhak

jongbhak@genomics.org

Korean Genomics Center (KOGIC), Ulsan National Institute of Science and Technology (UNIST)

Research Article

Keywords: Korean Genome Project, Korea10K, whole-genome sequencing, population genomics, imputation panel, multiomics, CpG variation, epigenomics, epigenetic clocks

Posted Date: April 17th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9285712/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License

Additional Declarations: Competing interest reported. SJ is the CEO and HR is an employee of AgingLab Inc.; YCho is an employee of CG Invites Co., LTD and Invites Genomics Co., LTD; B.L. is the CEO and JL is an employee of nSAGE; and GC is a co-founder of Nebula Genomics and Glottatech.com. All other authors declare no competing interests.

Version of Record: A version of this preprint was published at Genome Biology on November 20th, 2026. See the published version at https://doi.org/10.1186/s13059-026-04321-6.

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 genomic and multiomic resource comprising 10,239 high-depth whole genomes (mean coverage 30×) with integrated molecular and phenotypic data. Analysis of 9,000 unrelated individuals enabled near-complete discovery of rare and ultra-rare variants and supported the construction of a high-resolution, population-specific imputation panel. Koreans exhibit pronounced autosomal genetic homogeneity despite substantial diversity in Y-chromosomal, mitochondrial, and HLA lineages, reflecting long-term demographic continuity. We further identified 16.4 million variants that alter CpG dinucleotide context, revealing widespread sequence-driven modulation of the genomic CG landscape. Notably, population-specific CG-eliminating variants disrupt CpG probe targets in widely used methylation arrays, introducing a systematic source of bias in epigenome-wide association studies and epigenetic clock estimation.

Conclusion

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.

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 Korea1K (2020) with 1,094 whole genomes under the Korean Genome Project (KGP) [8] and Korean Variant Archive 2 (KOVA2; 2022) with 5,305 genomes and exomes which partly includes KGP data [9], followed by Korea4K (2024) with 4,157 high-depth whole genomes [10], 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. This dataset enables systematic characterization of population-specific genetic variation and its functional consequences, providing a foundation for integrative genomic analyses in the Korean population.

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. 1A). 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; Fig. S1).

Cumulative variant discovery curves showed a continuous increase in novel variant counts with expanding sample size (Fig. 1B), especially singletons and doubletons. Saturation analysis indicated that ultra-rare variants reached a mean discovery plateau at 8,955 samples (95% CI: 8,951-8,958), while 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 mere 569 samples (95% CI: 557–581) while very common variants (AF > 0.05) saturated at astonishing 109 samples (95% CI: 106–111) (Fig. S2; Table S1). These results underscore the relative genetic homogeneity of Koreans and 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. 1C). 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 (Table S2). The

enhancement was most striking for ultra-rare variants, including singletons and doubletons (Fig. 1D). Furthermore, the Korea10K panel imputed 606,935 variants, representing a 26% increase over Korea1K (Fig. 1C, 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. 2A, B). Ethnically relevant reference genomes, KOREF1 [11] 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. 2C, top panel; 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. Notably, 485 Koreans used in analysis were selected based on genetic distance to maximize within-population diversity (Fig. 2C, bottom panel), reflecting long-standing demographic continuity across the Korean Peninsula.

Analysis of 5,290 unrelated Korean Y-chromosomal genomes revealed a strong dominance of haplogroup O (N = 4,053, 75.7%), 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. 2D), 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 [12], into the Korean peninsula and the persistence of shared ancestral polymorphisms (Fig. S4A).

Mitochondrial genomes in Koreans exhibited markedly greater diversity compared to the Y chromosome. Phylogenetic reconstruction of 10,239 mitochondrial assemblies revealed deeply branching maternal lineages, primarily partitioned into macrohaplogroups M and N (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. 2E). The absence of Western Eurasian (H and X) and African haplogroups (L) supports prolonged geographic isolation (Fig. S4B), whereas fine-scale structure captured by PCA of 1,247 mitochondrial variants indicates multiple maternal lineage contributions from East Asian populations (Fig. S5). These findings support multiple ancient migration and admixture events involving females from neighboring populations into the Korean peninsula [13].

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 haplotypes), 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%) – accounted for nearly two-thirds of all haplotypes (Fig. 2F), indicating a narrow and coherent allelic spectrum. In comparison, Japanese showed a stronger enrichment of A*24:02 (37.0%), whereas Han Chinese exhibit a predominance of A*11:01 (25.3%); both populations display reduced representation of A*33:03. By contrast, Europeans and Africans exhibit markedly broader HLA polymorphism. No single allele exceeded 13% in African populations.

Population-specific CG-eliminating variants systematically bias DNA methylation measurements and epigenetic clock estimates in Koreans

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) 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 (Fig. S6A, B). These variants were broadly distributed across the allele frequency spectrum (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 (Fig. S6D).

To assess their impact on epigenomic assays, we intersected CG-eliminating variants with CpG probe target sites across three widely used methylation arrays: Infinium HumanMethylation27 (27K), HumanMethylation450 (450K), and MethylationEPIC v2.0 (EPIC v2.0). Between 16.0% and 30.6% of probes overlapped at least one CG-eliminating variant, converting the target CpG into a non-CpG context and rendering the site constitutively unmethylated irrespective of biological regulation (27K: 16.0%; 450K: 25.1%; EPIC v2.0: 30.6%; Fig. 3A-C). This proportion increased monotonically with array density, reflecting a fundamental trade-off between genomic coverage and susceptibility to population-specific genetic disruption.

Importantly, this effect was not limited to rare variation. On the EPIC v2.0, 11,486 probes were disrupted by variants with allele frequency (AF) > 1%, and 339 probes were universally eliminated (AF = 1.00), indicating complete absence of the target CpG across all individuals and representing systematic measurement failure rather than true biological hypomethylation (Fig. 3C). These findings demonstrate that CG-eliminating variants generate false-negative methylation signals and introduce a systematic source of bias in epigenome-wide association studies, whereby genetic variation can masquerade as differential methylation while reducing the effective probe space.

Consistently, all six epigenetic clocks examined contained disrupted CpG markers (Fig. 3D); notably, GrimAge was most affected, with 47 probes disrupted by common variants (AF > 0.01), indicating that widely used methylation-based biomarkers are not robust to population-specific CpG loss. Together, these results establish CG-eliminating variants as a pervasive and previously underrecognized source of bias in array-based methylation profiling, with direct implications for EWAS interpretation and the cross-population portability of epigenetic biomarkers.

Genetic Association of Biological Aging in the Koreans

We performed a genome-wide association analysis of biological aging, quantified as Phenotypic Age Advance (PAA) [14], among 4,498 unrelated Koreans with available health check-up data (Supplementary Note). Association testing of 6,996,863 autosomal variants using inverse-normal transformed PAA (PAA-INT) identified four SNVs at two unique loci (*LRP1B* and *EIF2AK4*) surpassing the suggestive significance ($P < 10^{-6}$; Fig. 5E). Three SNVs (rs4954893, rs13016322, and rs13025185) were positioned on LRP1B locus, while one (rs75578790) mapped to EIF2AK4 (Fig. 5D).

LRP1B is implicated in cognitive performance for successful aging [15]. EIF2AK4 (also known as GCN2) is a central regulator of the integrated stress response [16]. The GWAS showed minimal genomic inflation ($\lambda_{GC} = 1.012$; Fig. 5F), indicating effective control of confounding despite the modest sample size.

Korea10K as a Genomic, Multiomic, and Phenomic Data Resource

Korea10K expands prior Korean Genome Project efforts (KPGP, Korea1K, and Korea4K) into a nationwide, population-scale resource comprising 10,239 high-depth whole genomes (Fig. 4A). Participants were recruited across multiple institutions, capturing broad demographic representation and phenotypic diversity.

The cohort includes 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. 4B, C). Cardiovascular disorders represent the largest subgroup (56.6%), followed by neuropsychiatric conditions, enabling disease-focused analyses at a population-scale.

Korea10K marks a major expansion of the KGP into the multiomic era. Multiomic profiles from blood include 868 transcriptomes, 1,211 DNA methylomes, and 102 chromatin accessibility datasets, and six proteomic measurements, providing a systems-level view of the molecular consequences of genetic variation. Complementary phenotypic resources include the Health Check-up (HC) dataset (6,583 participants, 336 phenotypes; Table S4) and the Lifestyle Questionnaire (LQ) dataset (4,923 participants, 370 survey measures; Table S5), 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 (Supplementary Note).

Table 1
Overview of Multiomic and Phenomic Data Resource in Korea10K. Abbreviation: PBMC = Peripheral Blood Mononuclear Cell.

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

Discussion

There are several limitations to this study. Korea10K is cross-sectional, except 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 for association analyses and limiting integrative modeling across molecular and clinical layers. Structurally complex regions, including centromeres, sub-telomeric regions, and the Y chromosome, were left unresolved. Reliance on short-read sequencing and a linear reference genome.

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 [17], ethnically-relevant assemblies such as the KOREF assembly [11], and graph-based pangenomes [18] will be essential to resolve complex variation and reduce reference bias. Transcriptomic and epigenomic profiling will further enable population-specific mapping of regulatory quantitative trait loci, including transcriptomic [19] and epigenomic QTLs [20, 21], 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 consistent with Neolithic agricultural dispersals [12], whereas extensive mitochondrial diversity preserves signatures of ancient female-mediated admixture and Bronze Age demographic expansions [13]. 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 [22], autoimmune disease susceptibility [23], and drug contraindications [24]. 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) [25–28], neuropsychiatric traits (depression, anxiety, and suicidal behavior) [29, 30], infectious diseases (COVID-19) [31], and biological aging [32], providing mechanistic insights into disease etiology and progression.

Importantly, our findings highlight a critical interaction between genetic variation and epigenomic measurement. Population-specific CG-eliminating variants systematically disrupt CpG sites targeted by methylation arrays, introducing a previously underrecognized source of bias in epigenome-wide association studies and epigenetic clock estimation. This reveals a fundamental limitation of array-based methylation profiling and underscores the importance of incorporating population-specific genetic context in epigenomic analyses.

Conclusion

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 imporve 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 identification of population-specific genetic effects on epigenomic measurements 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%) (Fig. S10; Table S6). 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 prepared using platform-specific kits: TruSeq Nano or TruSeq PCR-Free (Illumina) and MGIEasy FS DNA Library Prep (BGI), followed by circularization and rolling circle amplification for DNBSEQ-T7 libraries. Libraries were sequenced as 2 × 150bp (or 151bp for Illumina) paired-end reads.

Read Processing and Alignment

Raw sequencing reads were trimmed with fastp (RRID:SCR_016962, v0.23.2) [33] 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 (RID:SCR_010910, v0.7.17) [34] with read group information. Alignments were coordinate-sorted and duplicates marked with Picard (RRID:SCR_006525, v2.27.5) [35]. Base quality score recalibration was performed using GATK (RRID:SCR_001876, v4.6.1.0) BaseRecalibrator and ApplyBQSR [36].

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 & 1000G 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.6.13LM) [37], 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 “supposed” Koreans with 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) [38].

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 median 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) [39], and imputation was performed with Minimac3 (RRID:SCR_009292, v2.0.1) [40]. 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) [41]. 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 panels were tested their

imputation performance against 19 unrelated Korean patients with gastric cancers utilized in the previous study [8].

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 [42]. 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 putative Korean genomes. Variant harmonization was performed using reciprocal intersection of SNPs present in both datasets using PLINK (RRID:SCR_001757, v2.0.0-a.6.13LM) "--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" [43], 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 200kb 1 0.5, resulting in 850,281 variants for population structure analysis.

Merging the 9,013 Korean genomes with all 2,541 1KGP individuals yielded 850,871 LD-pruned variants shared across all populations. 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) [44] was run for from K = 1 to K = 17. To reduce computational burden and account for long-range LD, variants were further pruned using --indep-pairwise 500kb 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 (Fig. S2).

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 and analyzed with yhaplo (v2.1.14.dev7 + gf8b300b) [45] 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) [46] 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) [47].

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) [48], 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" [49].

Annotation of 10,166 circularized assemblies to check their completeness was performed using MitoZ (v3.5) [50] 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) [51] 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 [52]. Multiple sequence alignment was conducted using MAFFT (RRID:SCR_011811, v7.525) [53] 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.6.13LM) “--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) [54] 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 two-digit resolution.

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

Identification and Regional Aggregation of CG Context-Associated Variants (CGVs) for Population-specific profiling

Genomic variants altering local CG dinucleotide context were identified by comparing the flanking nucleotides of reference and alternate alleles in GRCh38 to generate a CG context index. Variants whose substitution altered this index were classified as CG context-associated.

To estimate the prevalence of CG context-associated variants (CGVs) across the Korean population, Population prevalence was assessed via allelic presence, scoring heterozygous or homozygous alternate genotypes as 1 and homozygous reference as 0. Variants present in ≥ 80% of individuals were defined as Korean-popular CGVs.

To quantify their impact on genomic CG composition, Korean-popular CGVs were mapped to 10-kb windows. Net gain or loss of CG sites per window was summed and normalized by the total CG count in GRCh38, producing a CG context change ratio. Windows with ≥ 10% net gain or loss were retained for downstream analysis.

Assessment of CpG probe disruption by CG-eliminating variants

To quantify the extent to which CG-eliminating variants produce artifactual hypomethylation signals in the Korean population, we intersected the genomic coordinates of CpG probe target sites with CG-eliminating variants across three widely used methylation array platforms: Infinium HumanMethylation27 BeadChip, Infinium HumanMethylation450 BeadChip, and Infinium MethylationEPIC v2.0. For the HumanMethylation27 and HumanMethylation450 platforms, probe coordinates were lifted over to the GRCh38 reference assembly to ensure consistency with variant annotations.

Probes were subjected to stringent filtering criteria prior to analysis. First, probes whose target dinucleotide no longer corresponded to a CpG context in the GRCh38 reference sequence were excluded, ensuring accurate representation of CpG-targeting probes. Second, probes mapping to unplaced or unlocalized scaffolds were removed. The remaining probes were used to assess the frequency and distribution of CG-disrupting variants and their potential to induce spurious hypomethylation signals in array-based measurements.

Harmonization of Clinical Biomarkers Across KGP and NHANES Cohorts

Health check-up (HC) records were obtained from participants in the Korean Genome Project (KGP) and from publicly available datasets in the U.S. National Health and Nutrition Examination Survey or NHANES (RRID:SCR_013201) III and IV. NHANES data were retrieved using the BioAge R toolkit (RRID:SCR_027788) [55] which provides harmonized clinical and mortality records released by the U.S. Centers for Disease Control and Prevention.

Raw clinical data from KGP participants (N = 6,583) were harmonized with NHANES III (N = 18,825) and NHANES IV (N = 40,839) using a unified naming and unit-conversion system for clinical biomarkers. To ensure comparability across cohorts, biomarkers exhibiting substantial missingness were removed. To guarantee statistical power for genome-wide association analysis, only biomarkers representing at least 5,000 KGP participants were retained, except lymphocyte and monocyte proportions (deliberately included for possible adjustments for cell-type composition). A total of 12 clinical biomarkers were included in the mortality modeling: blood creatinine, waist circumference, body-mass index (BMI), blood triglyceride, chronological age, blood glucose, systolic blood pressure, diastolic blood pressure, high-density lipoprotein, total cholesterol level, lymphocyte proportions, and monocyte proportions.

Finally, only samples with complete phenotypic and consistent mortality status data were included in downstream analyses (NHANES III: N = 11,562, NHANES IV: N = 16,173, and KGP: N = 4,498).

Mortality Risk Prediction and Selection of Mortality-Associated Biomarkers

Candidate mortality-associated biomarkers were identified from NHANES III using penalized Cox proportional hazards regression implemented in the lifelines Python package (RRID:SCR_024899, v0.30.0), applying elastic-net regularization ($\alpha = 0.5$). Each clinical variable was first examined by univariate Cox regression to assess the significance and direction of its hazard ratio. A multivariate penalized model was fitted to evaluate joint contributions of variables to mortality risk. Model performance was evaluated based on the concordance index (C-index). The resulting set of mortality-associated biomarkers, including creatinine, chronological age, glucose, and systolic blood pressure, was retained for downstream survival modeling.

Multivariate Modeling of Phenotypic Age (PhenoAge)

Phenotypic Age (PhenoAge) is defined as the chronological age at which a reference individual exhibits the same predicted 10-year mortality hazard as that derived from a participant's biomarker profile. Our model derives biological aging from clinical biomarkers only, not DNA methylation measures.

PhenoAge was calculated using a modified implementation of the Levine et al. (2018) algorithm [14]. The BioAge R toolkit (RRID:SCR_027788), specifically the "phenoage_nhanes" function, enabled the incorporation of the tailored biomarker panel determined from univariate mortality-risk modeling. A Gompertz proportional hazards model was fitted on NHANES III mortality data, linking the mortality-associated biomarkers to 10-year mortality risk.

The trained Gompertz model parameters were subsequently applied to the KGP dataset to generate PhenoAge estimates for Korean individuals. To align PhenoAge with chronological age and correct for regression bias, a linear calibration was performed.

$PhenoAge_{calibrated}$ = (PhenoAge – intercept) / slope,

where the slope and intercept were obtained from linear regression of PhenoAge versus chronological age. This transformation ensured a mean PhenoAge Advance (PAA = PhenoAge – chronological age) centered around zero at the population level.

Genome-wide Association Study on Phenotypic Age Advance (PAA)

From 9,000 unrelated Koreans, variants were retained after applying the following exclusion criteria: (i) minor allele frequency (MAF) < 1%, (ii) deviation from Hardy-Weinberg equilibrium (HWE) with $P < 10^{-6}$, and (iii) missing genotype rate > 1%, leaving 6,996,863 variants.

Phenotypic Age Advance (PAA) exhibited a strong right-skewness and heavily-tailed distributional properties, violating linear regression assumptions. We applied a rank-based inverse normal transformation (INT) prior to association testing, yielding PAA-INT.

GWAS was performed on PAA-INT using PLINK (RRID:SCR_001757, v2.0.0-a.6.13LM) under a linear additive model in 4,498 unrelated Koreans with matched genotype and clinical data. Covariates included chronological age, age^2 (age squared), and the top 10 principal components (PC1-PC10) derived from genome-wide SNVs to adjust for the effect of time (linear and non-linear) and residual population stratification, respectively.

Abbreviations

AC = Allele count; AF = Allele frequency; AI = Artificial Intelligence; AMI = Acute myocardial infarction; AMR = Admixed American; ATAC-seq = Assay for Transposase-Accessible Chromatin using sequencing; BGI = Beijing Genomics Institute; BWA = Burrows–Wheeler Aligner; CDX = Chinese Dai; CG = Cytosine–Guanine dinucleotide; CGV = CG context–associated variant; CHB = Han Chinese in Beijing; CHS = Southern Han Chinese; CI = Confidence interval; CoxPH = Cox proportional hazards; EAS = East Asian; EUR = European; FDR = False discovery rate; GC = Genomic control; gnomAD = Genome Aggregation Database; GWAS = Genome-wide association study; HC = Health Check-up; HLA = Human leukocyte antigen; HR = Hazard ratio; Hi-C = High-throughput chromosome conformation capture; INT = Inverse-normal transformation; JPT = Japanese in Tokyo; KGP = Korean Genome Project; KHV = Kinh in Ho Chi Minh City, Vietnam; KOREF = Korean Reference Genome; KOR = Korean in Korea10K; LQ = Lifestyle Questionnaire; MAF = Minor allele frequency; MHC = Major Histocompatibility Complex; mtDNA = Mitochondrial DNA; NHANES = National Health and Nutrition Examination Survey; PAA = Phenotypic Age Advance; PBMC = Peripheral blood mononuclear cell; PCA = Principal component analysis; PhenoAge =

Phenotypic age; QC = Quality control; SAS = South Asian; SNV = Single-nucleotide variant; WGS = Whole-genome sequencing; 1KGP = 1000 Genomes Project.

Declarations

Code Availability

The source codes used to generate data and calculate statistics are openly available on the GitHub page: https://github.com/korean-genomics-center/korea10k.

Data Availability

Genetic variations data from Korea10K that pose no privacy concerns are available in VCF format through the Korea BioData Station (K-BDS, https://kbds.re.kr/), provided by the Korea Bioinformation Center (KOBIC). Access to summary-level datasets, such as HLA typing and haplogrouping results can be provided upon request. Access to individual-level 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. Any dataset provided is subject to restriction or withdrawal in the future in response to policy updates or regulatory requirements.

The 1KGP Phase 3 high-depth (30×) dataset used in this paper was retrieved from the Amazon Web Services (AWS) server (https://registry.opendata.aws/1000-genomes).

The 1KGP Phase 1 legacy data can be found on their FTP server (https://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/).

Ethics, Consent, and Permissions

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 outline the intended use of their data for research purposes and underscore the voluntary nature of their participation. Our study adheres to the ethical guidelines and regulations stipulated by the IRB.

Competing interests

SJ is CEO and HR is an employee of AgingLab Inc. YCho is an employee of CG Invites Co., LTD and Invites Genomics Co., LTD. B.L. is CEO and JL is an employee of nSAGE. GC is co-founder of Nebula Genomics and Glottatech.com. All other authors declare no competing interests

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 Korea government (MSIT) (No. RS-2025-25449754).

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.

Acknowledgments

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 and Moo-Young Jung 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, 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, 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 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) for assistance with language editing, polishing, and improving the clarity of the manuscript.

References

1. Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alfoldi J, Wang Q, Collins RL, Laricchia KM, Ganna A, Birnbaum DP, 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, Reeve MP, Laivuori H, Aavikko M, Kaunisto MA, 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, Chen WJ, Su MW, Chu HW, Shen CY, 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, Ninomiya T, Tamakoshi A, Yamagata Z, Mushiroda T, 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, Soon WWJ, Bertin N, Lin CB, Khor CC, et al. Large-Scale Whole-Genome Sequencing of Three Diverse Asian Populations in Singapore. Cell. 2019;179:736–e749715.
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. Jeon S, Bhak Y, Choi Y, Jeon Y, Kim S, Jang J, Jang J, Blazyte A, Kim C, Kim Y, et al. Korean Genome Project: 1094 Korean personal genomes with clinical information. Sci Adv. 2020;6:eaaz7835.
9. Lee J, Lee J, Jeon S, Lee J, Jang I, Yang JO, Park S, Lee B, Choi J, Choi BO, 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.
10. Jeon S, Choi H, Jeon Y, Choi WH, Choi H, An K, Ryu H, Bhak J, Lee H, Kwon Y et al. Korea4K: whole genome sequences of 4,157 Koreans with 107 phenotypes derived from extensive health check-ups. Gigascience 2024, 13.
11. Cho YS, Kim H, Kim HM, Jho S, Jun J, Lee YJ, Chae KS, Kim CG, Kim S, Eriksson A, et al. An ethnically relevant consensus Korean reference genome is a step towards personal reference genomes. Nat Commun. 2016;7:13637.

12. Kim SH, Kim KC, Shin DJ, Jin HJ, Kwak KD, Han MS, Song JM, Kim W, Kim W. High frequencies of Y-chromosome haplogroup O2b-SRY465 lineages in Korea: a genetic perspective on the peopling of Korea. Investig Genet. 2011;2:10.
13. Kim J, Jeon S, Choi JP, Blazyte A, Jeon Y, Kim JI, Ohashi J, Tokunaga K, Sugano S, Fucharoen S, et al. The Origin and Composition of Korean Ethnicity Analyzed by Ancient and Present-Day Genome Sequences. Genome Biol Evol. 2020;12:553–65.
14. Levine ME, Lu AT, Quach A, Chen BH, Assimes TL, Bandinelli S, Hou L, Baccarelli AA, Stewart JD, Li Y, et al. An epigenetic biomarker of aging for lifespan and healthspan. Aging. 2018;10:573–91.
15. Poduslo SE, Huang R, Spiro A. 3rd: A genome screen of successful aging without cognitive decline identifies LRP1B by haplotype analysis. Am J Med Genet B Neuropsychiatr Genet. 2010;153B:114–9.
16. Pakos-Zebrucka K, Koryga I, Mnich K, Ljujic M, Samali A, Gorman AM. The integrated stress response. EMBO Rep. 2016;17:1374–95.
17. Kim J, Park JL, Yang JO, Kim S, Joe S, Park G, Hwang T, Cho MJ, Lee S, Lee JE et al. Highly accurate Korean draft genomes reveal structural variation highlighting human telomere evolution. Nucleic Acids Res 2025, 53.
18. Liao WW, Asri M, Ebler J, Doerr D, Haukness M, Hickey G, Lu S, Lucas JK, Monlong J, Abel HJ, et al. A draft human pangenome reference. Nature. 2023;617:312–24.
19. Gusev A, Ko A, Shi H, Bhatia G, Chung W, Penninx BW, Jansen R, de Geus EJ, Boomsma DI, Wright FA, et al. Integrative approaches for large-scale transcriptome-wide association studies. Nat Genet. 2016;48:245–52.
20. Gate RE, Cheng CS, Aiden AP, Siba A, Tabaka M, Lituiev D, Machol I, Gordon MG, Subramaniam M, Shamim M, et al. Genetic determinants of co-accessible chromatin regions in activated T cells across humans. Nat Genet. 2018;50:1140–50.
21. Banovich NE, Lan X, McVicker G, van de Geijn B, Degner JF, Blischak JD, Roux J, Pritchard JK, Gilad Y. Methylation QTLs are associated with coordinated changes in transcription factor binding, histone modifications, and gene expression levels. PLoS Genet. 2014;10:e1004663.
22. Crocchiolo R, Gallina AM, Pani A, Campisi D, Cento V, Sacchi N, Miotti V, Gagliardi OM, D'Amico F, Vismara C, et al. Polymorphism of the HLA system and weak antibody response to BNT162b2 mRNA vaccine. HLA. 2022;99:183–91.
23. Kang EH, Kim JY, Takeuchi F, Kim JW, Shin K, Lee EY, Lee YJ, Lee EB, Park MH, Song YW. Associations between the HLA-A polymorphism and the clinical manifestations of Behcet's disease. Arthritis Res Ther. 2011;13:R49.
24. 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.
25. Bhak Y, Jeon Y, Jeon S, Yoon C, Kim M, Blazyte A, Kim Y, Kang Y, Kim C, Lee SY, 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.

26. Jeon Y, Jeon S, Choi WH, An K, Choi H, Kim BC, Kim W, Lee SY, Bae JW, Hwang JY, 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.
27. Jeon Y, Jeon S, An K, Kim YJ, Kim BC, Ryu H, Choi WH, Choi H, Kim W, Lee SY, 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.
28. Jeon Y, Kwon Y, Kim YJ, Jeon S, Ryu H, An K, Kim BC, Kim W, Lee SY, Bae JW, 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.
29. Bhak Y, Jeong HO, Cho YS, Jeon S, Cho J, Gim JA, Jeon Y, Blazyte A, Park SG, Kim HM, et al. Depression and suicide risk prediction models using blood-derived multi-omics data. Transl Psychiatry. 2019;9:262.
30. Kwon Y, Blazyte A, Jeon Y, Kim YJ, An K, Jeon S, Ryu H, Shin DH, Ahn J, Um H, 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.
31. Ryu H, An K, Kwon Y, Jeon Y, Jeon S, Choi H, Kim YJ, Kim S, Sul OJ, Lee S, 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.
32. An K, Kwon Y, Bhak J, Ryu H, Jeon S, Nam D, Bhak J. Dynamic and reversible transcriptomic age shifts induced by COVID-19 in Korean whole blood. Aging. 2025;17:1484–510.
33. Chen S, Zhou Y, Chen Y, Gu J. fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics. 2018;34:i884–90.
34. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25:1754–60.
35. Picard, Toolkit. [https://broadinstitute.github.io/picard/]
36. 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.
37. 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.
38. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GR, Thormann A, Flicek P, Cunningham F. The Ensembl Variant Effect Predictor. Genome Biol. 2016;17:122.
39. 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.
40. Das S, Forer L, Schonherr S, Sidore C, Locke AE, Kwong A, Vrieze SI, Chew EY, Levy S, McGue M, et al. Next-generation genotype imputation service and methods. Nat Genet. 2016;48:1284–7.

41. Taliun D, Harris DN, Kessler MD, Carlson J, Szpiech ZA, Torres R, Taliun SAG, Corvelo A, Gogarten SM, Kang HM, et al. Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. Nature. 2021;590:290–9.
42. Genomes Project C, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, Marchini JL, McCarthy S, McVean GA, Abecasis GR. A global reference for human genetic variation. Nature. 2015;526:68–74.
43. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ, Sham PC. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007;81:559–75.
44. Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. Genome Res. 2009;19:1655–64.
45. Poznik GD. Identifying Y-chromosome haplogroups in arbitrarily large samples of sequenced or genotyped men. bioRxiv 2016:088716.
46. Xu L, He W, Tai S, Huang X, Qin M, Liao X, Jing Y, Yang J, Fang X, Shi J, Jin N. VCF2Dis: an ultra-fast and efficient tool to calculate pairwise genetic distance and construct population phylogeny from VCF files. Gigascience 2025, 14.
47. 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.
48. Dierckxsens N, Mardulyn P, Smits G. NOVOPlasty: de novo assembly of organelle genomes from whole genome data. Nucleic Acids Res. 2017;45:e18.
49. Bonfield JK, Marshall J, Danecek P, Li H, Ohan V, Whitwham A, Keane T, Davies RM. HTSlib: C library for reading/writing high-throughput sequencing data. Gigascience 2021, 10.
50. 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.
51. Schonherr S, Weissensteiner H, Kronenberg F, Forer L. Haplogrep 3 - an interactive haplogroup classification and analysis platform. Nucleic Acids Res. 2023;51:W263–8.
52. 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.
53. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol. 2013;30:772–80.
54. Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, Whitwham A, Keane T, McCarthy SA, Davies RM, Li H. Twelve years of SAMtools and BCFtools. *Gigascience* 2021, 10.
55. Kwon D, Belsky DW. A toolkit for quantification of biological age from blood chemistry and organ function test data: BioAge. Geroscience. 2021;43:2795–808.

Figures

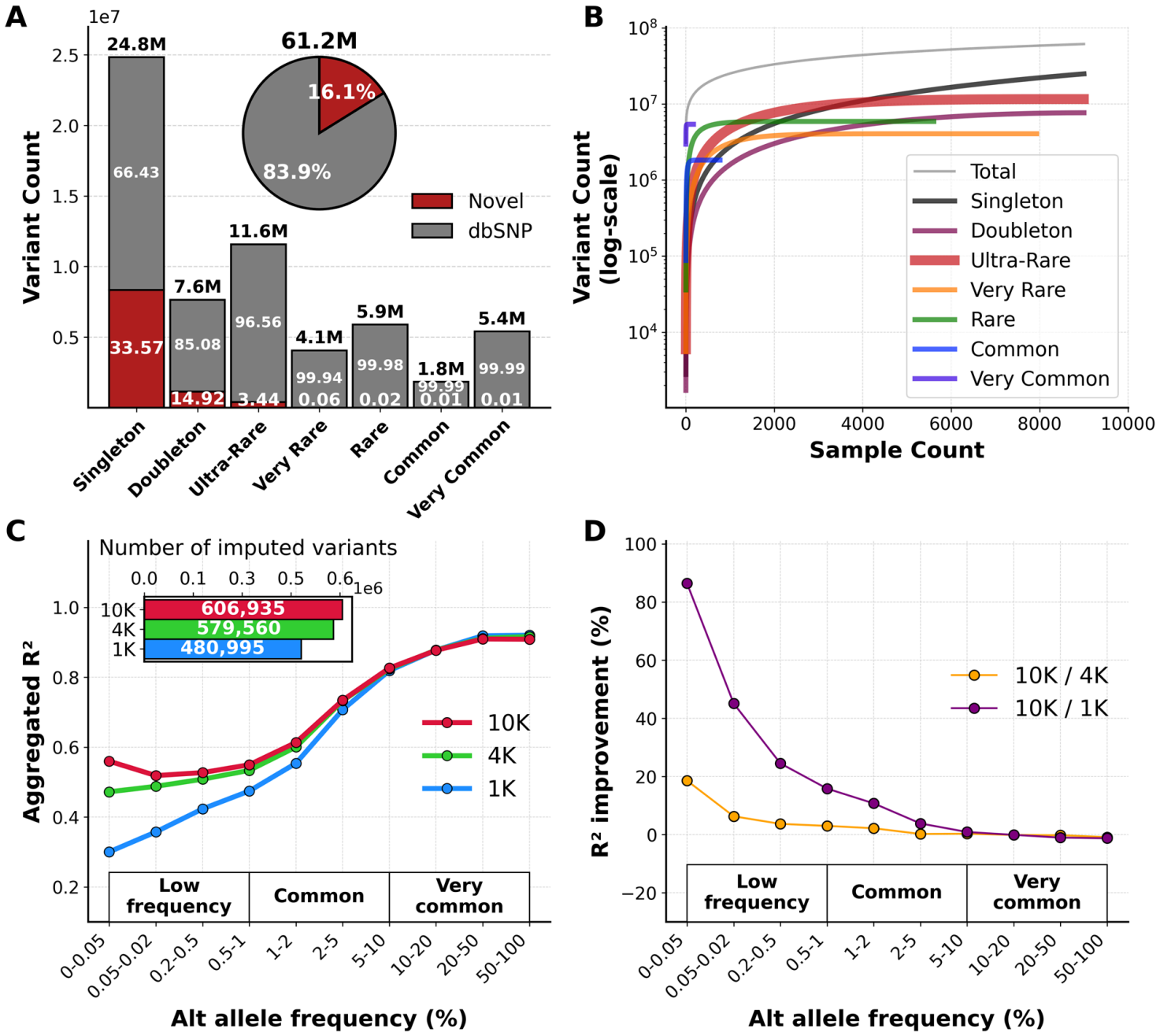


Figure 1

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 (red) and dbSNP-reported (grey) 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, in 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 (blue), Korea4K (green), and Korea10K (red) 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 percentage) comparing Korea10K to Korea1K (purple) and

Korea10K to Korea4K (yellow) 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.

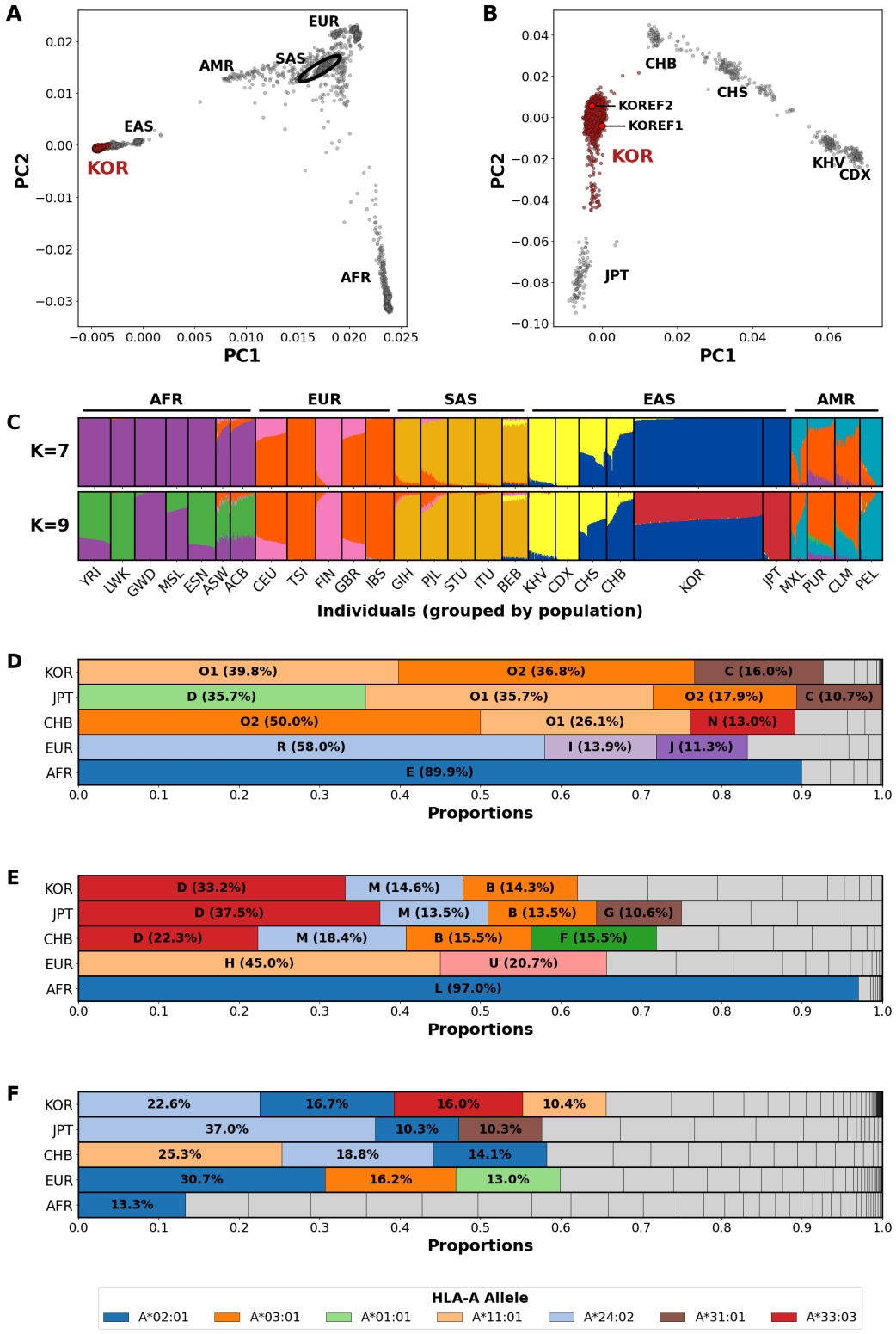


Figure 2

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). Koreans are displayed as closed red dots. 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 bright red dots. **(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 = Ad Mixed 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.

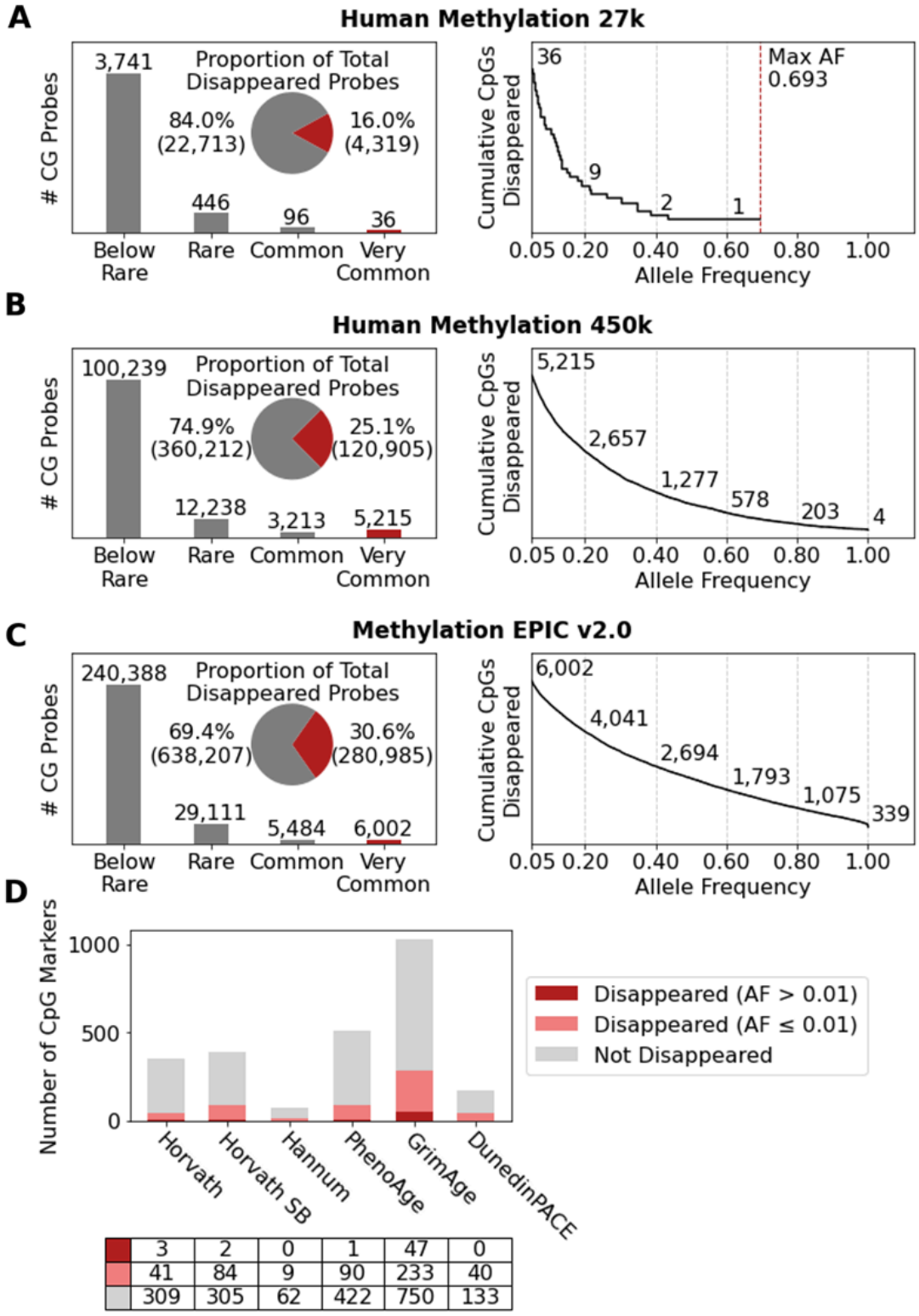


Figure 3

Impact of CG-eliminating variants on CpG probe across methylation array platforms. (A–C) Overlap between CG-eliminating variants identified in the Korean population and probes stratified by allele frequency (AF). Tested on three methylation array platforms: (A) Infinium HumanMethylation27 BeadChip (27K), (B) Infinium HumanMethylation450 BeadChip (450K), and (C) Infinium MethylationEPIC v2.0 Kit (EPIC v2.0). Bar charts showing the number of CpG probes disrupted by CG-eliminating variants

across four AF categories: Below Rare (AF ≤ 0.001), Rare (0.001 < AF ≤ 0.01), Common (0.01 < AF ≤ 0.05), and Very Common (AF > 0.05). Inset pie charts indicate the proportion of total probes on each array affected by CG-eliminating variants (red) relative to unaffected probes (gray). Cumulative number of CpG probes disrupted by Very Common variants (AF > 0.05) as a function of increasing AF threshold. **(D)** Stacked bar chart showing the number of epigenetic clock markers disrupted by CG-eliminating variants for five clocks (Horvath, Horvath Skin & Blood (SB), Hannum, PhenoAge, GrimAge, and DunedinPACE). Disrupted markers are stratified by AF > 0.01 (dark red) and AF ≤ 0.01 (light red), with unaffected markers shown in gray. The table below the chart provides exact counts of disrupted and retained markers for each clock.

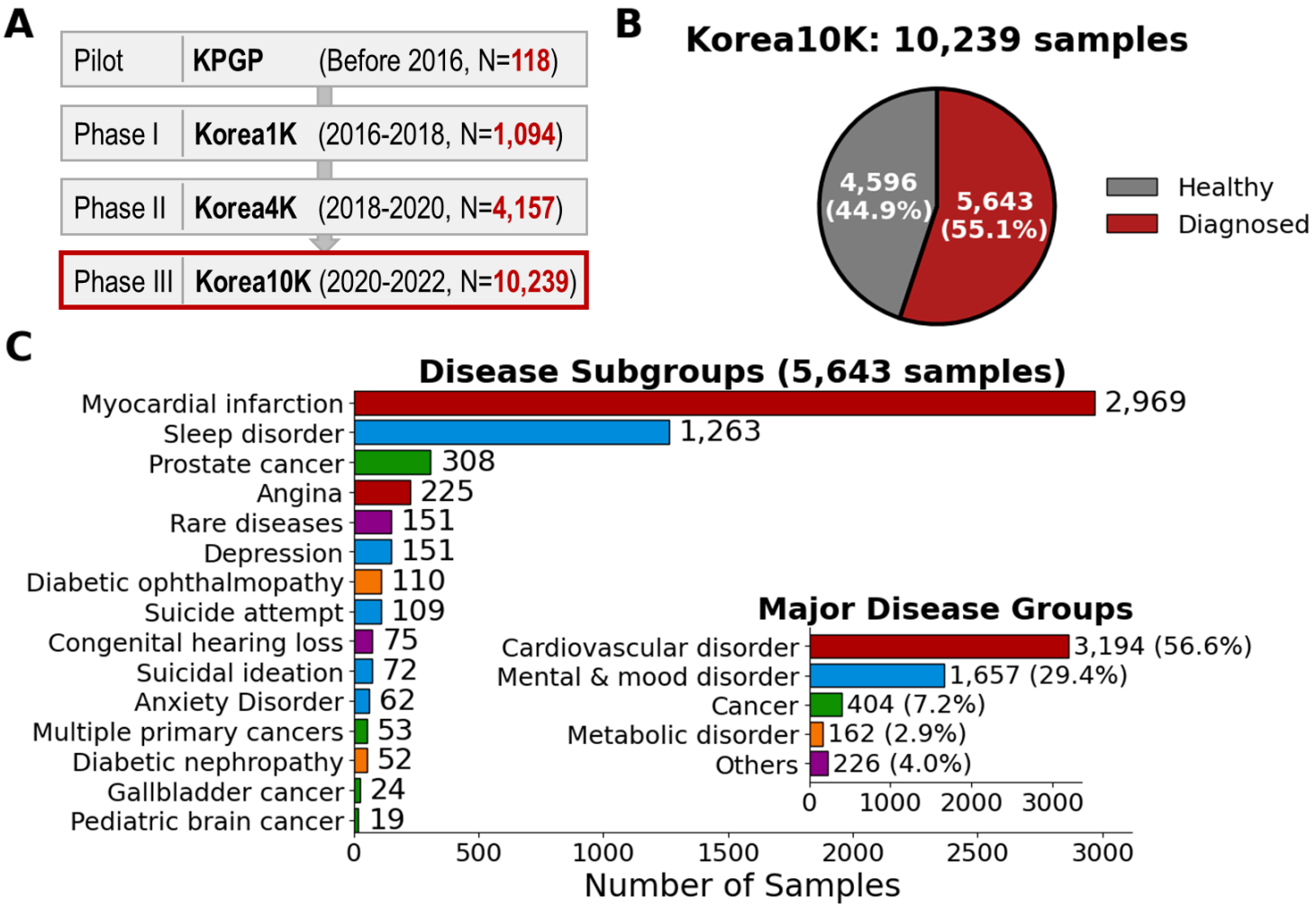


Figure 4

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 = 351),

metabolic disorders (N = 162), others (N = 226), and. Abbreviation: KPGP = Korean Personal Genome Project.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Korea10KSupplementaryGenomeBiologyResubmission.docx