10,239 whole genomes with multi-omic and clinical health information as the Korean population multi-omic reference dataset

Kyungwhan An[1,2,†], Sungwon Jeon[1,3,†], Yoonsung Kwon[1,2,†], Changhan Yoon[1,2], Yookyung Choi[1,2], Yeonsu Jeon[4], Hyoung-Jin Choi[1,2], Jihun Bhak[1,2], Dong-Hyun Shin[1,2], Hyomin Lee[1,2], Yeo Jin Kim[1], Eun-Seok Shin[5], Hyojung Ryu[1,3], Asta Blazyte[1,6], Sangsoo Park[1,3], Juok Cho[1,3], Dan Bolser[7], Jongbum Jeon[4], Soobok Joe[4], Jin Ok Yang[4], Jong-Hwan Kim[4], Byoung-Chul Kim[8], Jungeun Kim[8], Neung-Hwa Park[9,10], Siwoo Lee[11], Hee-jeong Jin[12], Younghwa Baek[12], Dooyoung Jung[2], Yunsung Cho[13,14], Yong Min Ahn[15], Hyung-Tae Jung[16], Byungjoo Ham[17,18,19], Yoon Young Choi[20], Seung Ah Choi[21,22], Seung-Ki Kim[21,22], Ji Hoon Phi[21,22], Heon Yung Gee[23], Borah Kim[24], Woojae Myung[25], Chong Kun Cheon[26], Dong Uk Kim[27], Gang-Nam Jin[28], Hojune Lee[28], Kyun Shik Chae[29], Chang Geun Kim[29], BongHee Lee[30], Jaesuk Lee[30], Semin Lee[1,2], George Church[31,32,33,*], Haeyoung Jeong[4,*], Jong Bhak[1,2,3,*]

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

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

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

[6]School of Medicine, Gachon University, Yeonsu-gu, Incheon 21565, Republic of Korea

[7]Geromics Ltd., 222 Mill Road, Cambridge, CB1 3NF, United Kingdom

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

[23]Department of Pharmacology, Graduate School of Medical Science, Yonsei University College of Medicine, Brain Korea 21 Project, Seoul, 03722, Republic of Korea

[24]Department of Psychiatry, CHA Bundang Medical Center, CHA University School of Medicine, Seongnam 13520, Republic of Korea

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

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

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

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

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

[30]nSAGE, Incheon 21999, Republic of Korea

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

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

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

†These authors contributed equally to this work.

*Correspondence to: gchurch@genetics.med.harvard.edu (George C.), hyjeong@kribb.re.kr (Haeyoung J.), jongbhak@genomics.org (Jong B.)

Abstract

We present Korea10K, the largest genomic dataset of the Korean population, comprising 10,239 high-coverage whole genomes (mean depth 30×) with matched multi-omic profiles and phenotype data. Korea10K achieves complete and near-complete discovery of very rare and ultra-rare alleles, respectively, at 9,000 Korean genomes. This dataset provides the high-quality population-specific imputation panel, enabling accurate inference of low-frequency variants. Admixture analyses confirm the overall genetic homogeneity of the Korean population, despite its diverse Y-chromosomal, mitochondrial, and HLA repertoires. This pattern reflects a long and continuous lineage history characterized by persistent internal admixture and genomic homogenization over thousands of years on the Korean peninsula. We also identified 16.8 million genomic variants that directly modify CG sites by creating or abolishing CG dinucleotides, providing the population-scale evidence of coordinated genomic-epigenomic regulatory mechanism in Koreans.

Keywords

Korean Genome Project, Korea10K, whole-genome sequencing, population genomics, variant discovery, population reference imputation panel, multi-omics, CGV

Introduction

Population-scale whole-genome sequencing (WGS) has enabled high-resolution investigations into global genetic diversity, population structure, and disease susceptibility. Landmark initiatives such as the 1,000 Genomes Project (1KGP) (Genomes Project et al., 2015) and the Genome Aggregation Database (gnomAD) (Karczewski et al., 2020) have collectively established a foundation for cataloging genetic variants across global populations. More recent large-scale biobank projects, including the UK Biobank (Consortium, 2025) and FinnGen (Kurki et al., 2023), 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 (Feng et al., 2022), Biobank Japan (Nagai et al., 2017), and Singapore's PRECISE-SG100K project (Wu et al., 2019), 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) (Jeon et al., 2020), Korean Variant Archive 2 (KOVA2; 2022) with 5,305 genomes and exomes which partly includes KGP data (Lee et al., 2022), followed by Korea4K (2024) with 4,157 high-depth whole genomes (Jeon et al., 2024), revealing Korean-specific genetic variants and genetic-phenotypic associations, establishing critical population references for East Asian genomics.

Despite these advances, no single resource to date has conclusively captured the genetic spectrum of Koreans. Prior studies were limited by cohort size, sequencing coverage and depth, or lack of integrated omics and clinical data. The absence of a unified, deeply sequenced, and phenotypically annotated Korean genomic reference has hindered both population genetics and translational research.

Here, we present Korea10K, the most comprehensive genomic and multi-omic resource of the Korean population to date, encompassing 10,239 high-depth whole genomes coupled with matched transcriptomic, epigenomic, and phenotypic datasets. This resource marks a new phase of Korean population genomics.

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. We identified 61,223,704 autosomal variants, comprising 57,184,188 single-nucleotide variants (SNVs) and 4,039,516 short insertions and deletions (indels) from 9,000 unrelated Korean samples. This reduction in sample number reflects stringent quality control, kinship filtering, and removal of ancestral outliers, to ensure unbiased variant discovery and accurate population-scale

allele frequency estimation Among the identified variants, 9,880,066 (16.1%) were novel and absent from dbSNP (v157), which already includes the Korea4K (from 3,617 unrelated Korean genomes) variome dataset. Low-frequency variants dominated these previously unseen variations, with 33.6% classified as singletons (AC = 1), 14.9% as doubletons (AC = 2), and 3.4% as ultra-rare (AC > 2, AF ≤ 0.0005), revealing previously uncharacterized variation within the Korean population (Figure 1A). Across all frequency categories, most of the novel variants were classified as SNVs, except for the very rare ones, where 34.9% were small indels (22.1% deletions and 12.8% insertions; Supplementary Fig. S1).

Cumulative variant discovery curves showed a continuous increase in novel variant counts with expanding sample size (Figure 1B), including singletons and doubletons. Saturation analysis indicated that ultra-rare variants (AC > 2, AF ≤ 0.0005) reached a mean discovery plateau at 8,955 samples (95% CI: 8,951–8,958), while very rare (AF > 0.0005, AF ≤ 0.001) and rare variants (AF > 0.001, AF ≤ 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, 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) (Supplementary Fig. S2; Supplementary Table S1). These results underscore the importance of large-scale sequencing to capture low-frequency Korean variation and justify the sample size of Korea10K for near-complete variant discovery for the ethnic Korean population.

The Korea10K population reference panel demonstrated markedly improved imputation performance compared with earlier Korean panels (Korea1K and Korea4K), particularly for low-frequency variants (MAF < 1%; Figure 1C). Relative to Korea1K, aggregated imputation R^2 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 (Supplementary Table S2). The enhancement was most striking for ultra-rare variants, including singletons and doubletons (Figure 1D). Furthermore, the Korea10K panel imputed 606,935 variants, representing a 26% increase over Korea1K (Figure 1C, upper left), underscoring the value of progressive variant discovery for genotype imputation.

Korea10K provides a population reference of 61.2 million fully phased, functionally annotated autosomal variants, including 9.9 million novel alleles. Combined with a high-resolution Korean imputation panel, it enables accurate genotype imputation, rare-variant analyses, and population-specific functional interpretation.

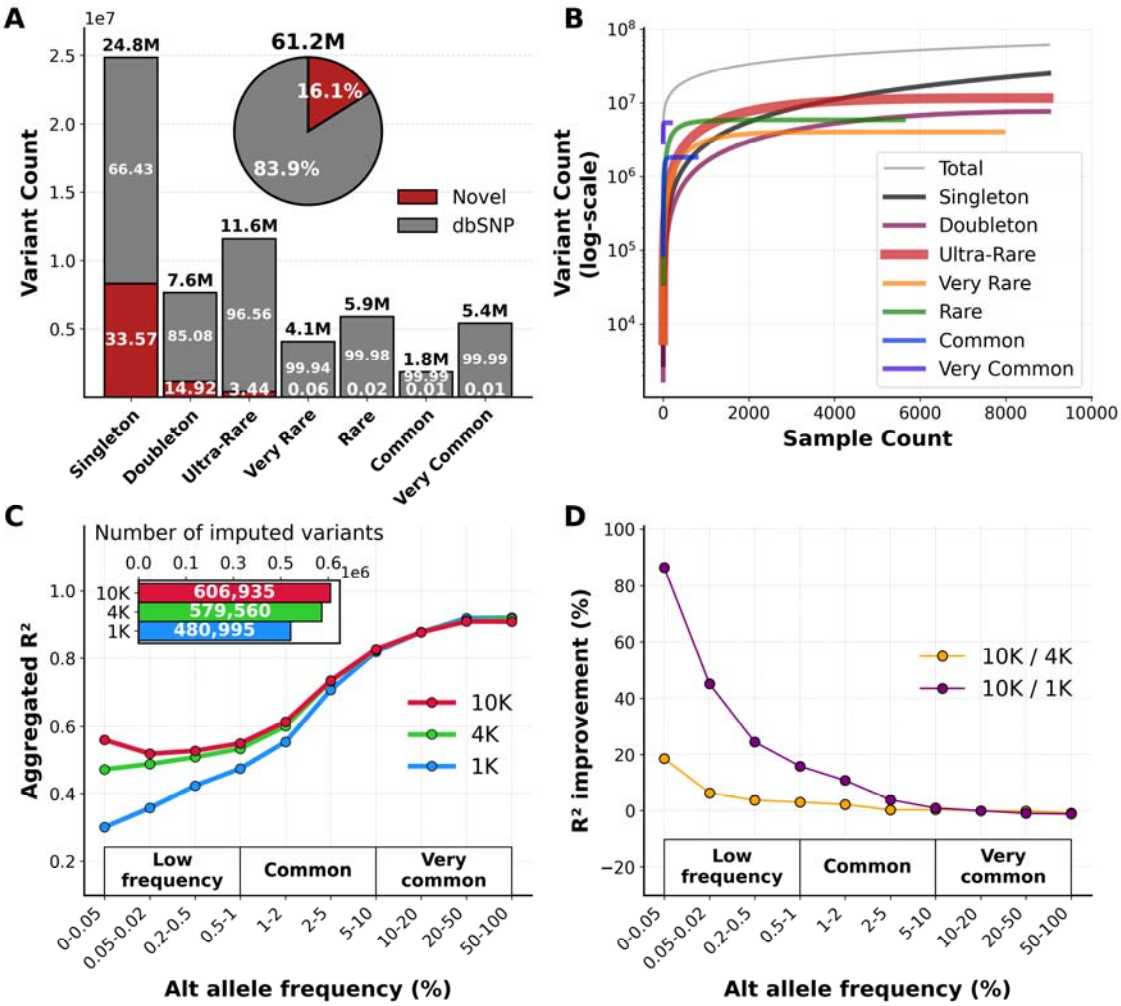


Figure 1. Variome landscape of the Korea10K whole genomes. (A) Stacked bar charts highlighting the distribution of discovered variant counts (y-axis) by allele frequency categories (x-axis), partitioned into novel (red) and dbSNP-reported (grey) variants. The pie chart summarizes overall variant statistics. **(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, AF $\leq$ 0.0005), Very Rare (AF > 0.0005, AF $\leq$ 0.001), Rare (AF > 0.001, AF $\leq$ 0.01), Common (AF > 0.01, AF $\leq$ 0.05), and Very Common (AF > 0.05). Abbreviations: Alt allele = Alternative allele, AC = allele count, AF = allele frequency.

Population Structure and Genetic Homogeneity of Koreans

Principal component analysis (PCA) of autosomal genotypes revealed Koreans as a cluster within the East Asian genetic continuum, clearly separated from Han Chinese (CHB, CHS) and Japanese (JPT) populations from the 1000 Genomes Project (Figure 2A, B). Ethnically-relevant reference genomes, KOREF1 (Cho et al., 2016) and its successor KOREF2, co-localized within the Korean cluster, validating them as well-suited genomic representatives. Admixture analysis revealed a single dominant ancestry component when modeling global autosomal variation with seven ancestral populations (K = 7), which best captures global autosomal variation (Figure 2C, top; Supplementary Fig. S3). At higher cluster numbers (K = 9), Koreans exhibited a dual-ancestry pattern shared with northern Han Chinese and Japanese populations. Even among 485 individuals selected based on genetic distance to maximize within-population diversity, the same two ancestral components persisted (Figure 2C, bottom), 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-Y (N = 4,053, 75.7%), comprising both O1-Y (N = 2,105, 39.8%) and O2-Y (N = 1,948, 36.8%), followed by C-Y (N = 847, 16.0%). This distribution is broadly consistent with patterns observed across northern East Asia (Figure 2D). In contrast, population-specific signatures include markedly higher frequencies of haplogroup D-Y in the Japanese (35.7%) and a stronger representation of O2 in northern Han Chinese (50.05%). Outside East Asia, Europeans are dominated by haplogroup R-Y (58%), whereas Africans show overwhelming prevalence of haplogroup E-Y (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 lack of phylogenetic structure reflects shallow sequence divergence among lineages due to recent demographic expansions into the Korean peninsula and the persistence of shared ancestral polymorphisms (Supplementary Fig. S4A).

The mitochondria of Koreans exhibited markedly greater diversity compared to the Y chromosome. A total of 10,239 mitochondrial assemblies and alignments were generated, representing one of the largest population-level mitogenomic resources to date. Phylogenetic reconstruction with the complete circular assemblies revealed a deeply branching topology, indicating highly diverse maternal lineages in Korea (Supplementary Fig. S4B). Despite extensive overall diversity, mitochondrial lineages in Koreans were primarily partitioned into two major macrohaplogroups, M-mt and N-mt, encompassing nearly all dominant subclades, namely D-mt (N = 2,988, 33.2%), M-mt (N = 1,315, 14.6%), and B-mt (N = 1,285, 14.3%) in 9,000 unrelated Koreans (Figure 2E). The absence of Western Eurasian (H-mt and X-mt) and African haplogroups (L-mt) supports prolonged maternal lineage isolation due to Korea's geographic remoteness (Supplementary Fig. S4B). The Principal Component Analysis (PCA) of the mitochondrial variants further captured fine-scale maternal sub-structure concordant with haplogroup-defined clusters (Supplementary Fig. S5). These findings support multiple prehistoric admixture events involving female migrants from neighboring East Asian populations into the Korean peninsula.

At the immunogenetic level, HLA class I (HLA-A) allele frequencies illustrate characteristic features of the Korean population that reflect long-term immunological purifying pressures. Among 9,000 unrelated Koreans (18,000 haplotypes), four alleles, including A24:02 (N = 4,059, 22.6%), A02:01 (N = 3,003, 16.7%), A33:03 (N = 2,879, 16.0%), and A11:01 (N = 1,871, 10.4%), collectively accounted for nearly two-thirds of all haplotypes (Figure 2F), indicating a narrow and coherent allelic spectrum. In comparison, Japanese individuals show a stronger enrichment of A24:02 (37.0%), whereas Han

Chinese exhibit a predominance of A11:01 (25.3%); both populations display reduced representation of A33:03. By contrast, Europeans and Africans exhibit markedly broader HLA polymorphism, with A02:01 comprising 30.7% of European haplotypes but accompanied by numerous low-frequency variants, and no single allele exceeding 13% in African populations.

Collectively, Koreans represent a genetically homogeneous population shaped by long-term geographic isolation and demographic continuity, with additional genetic contributions from multiple prehistoric admixture events involving both male and female migrants from neighboring East Asian populations. The ancestral and immunogenetic resources generated by Korea10K offer a valuable foundation for population genetics, forensic analyses, and immunological research.

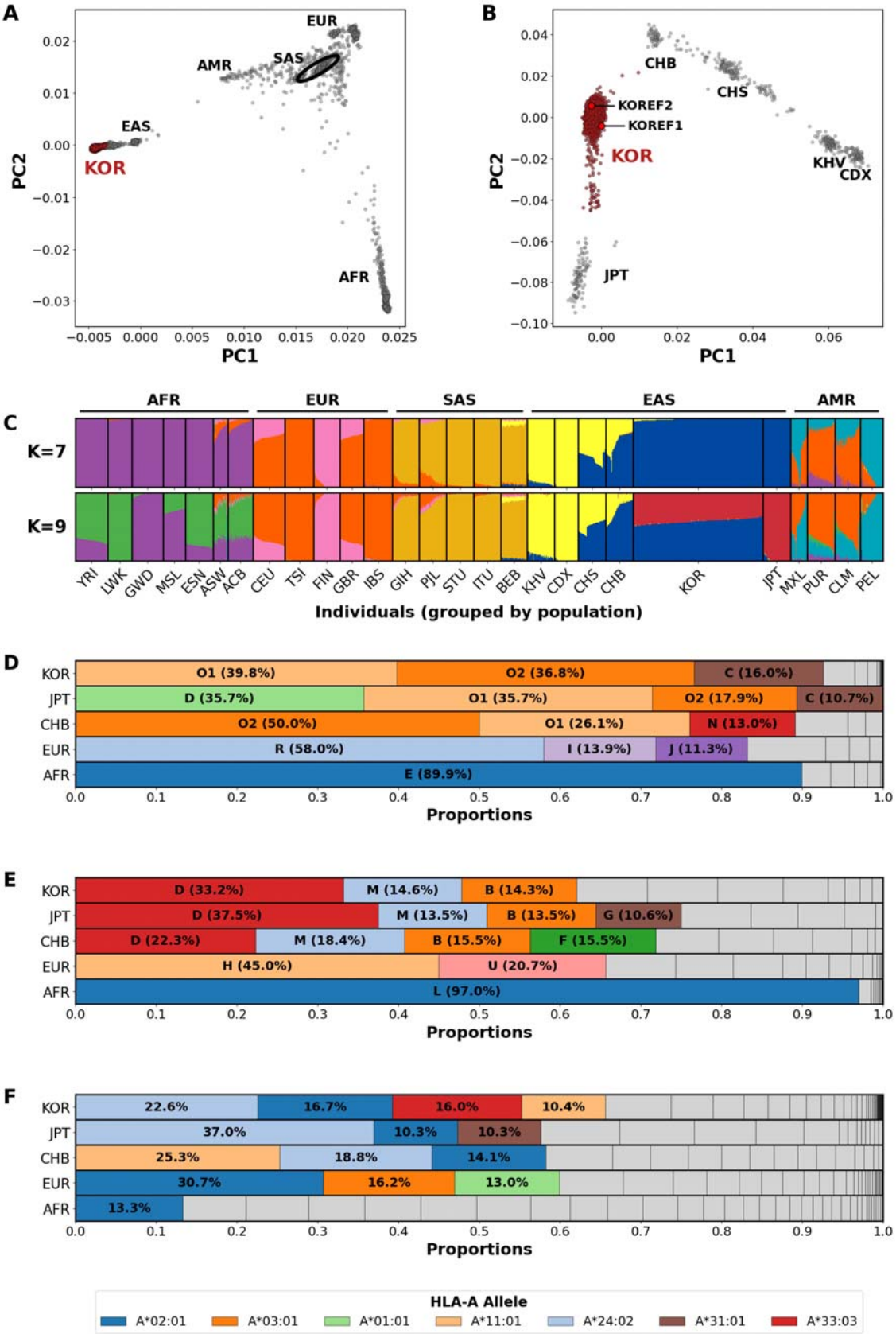

A
EUR
SAS
AMR
EAS
KOR
AFR
PC1
PC2
B
CHB
CHS
KOREF2
KOREF1
KOR
KHV
CDX
JPT
PC1
PC2
C
AFR
EUR
SAS
EAS
AMR
K=7
K=9
YRI LWK GWD MSL ESN ASW ACB CEU TSI FIN GBR IBS GIH PJL STU ITU BEB KHV CDX CHS CHB KOR JPT MXL PUR CLM PEL
Individuals (grouped by population)
D
KOR O1 (39.8%) O2 (36.8%) C (16.0%)
JPT D (35.7%) O1 (35.7%) O2 (17.9%) C (10.7%)
CHB O2 (50.0%) O1 (26.1%) N (13.0%)
EUR R (58.0%) I (13.9%) J (11.3%)
AFR E (89.9%)
Proportions
E
KOR D (33.2%) M (14.6%) B (14.3%)
JPT D (37.5%) M (13.5%) B (13.5%) G (10.6%)
CHB D (22.3%) M (18.4%) B (15.5%) F (15.5%)
EUR H (45.0%) U (20.7%)
AFR L (97.0%)
Proportions
F
KOR 22.6% 16.7% 16.0% 10.4%
JPT 37.0% 10.3% 10.3%
CHB 25.3% 18.8% 14.1%
EUR 30.7% 16.2% 13.0%
AFR 13.3%
Proportions
HLA-A Allele
A*02:01 A*03:01 A*01:01 A*11:01 A*24:02 A*31:01 A*33:03

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.

Population-Specific CG Context-associated Variants (CGV) in Koreans

DNA mutations can directly alter CG dinucleotide contexts, which in turn can modulate the RNA expression of specific genes. Although direct methylation data were not analyzed here, sequence-driven changes in CG context-associated variants (known as CGVs, CG-SNP, or meSNP), a form of geno-epigenetic polymorphism, were found to influence dinucleotide context. These CGVs suggest population-level differences in epigenetic regulation and gene expression, particularly in pathways related to psychiatric and metabolic disease risk in the Korean population.

We performed a whole-genome analysis of 9,000 unrelated Korean individuals and identified extensive variation affecting the CG dinucleotide context. This provides a population-scale view of CG sequence diversity. Among 61,223,704 high-quality biallelic variants, 16,388,405 (26.8%) directly altered CG sites, comprising 7,154,078 (11.7%) CG-creating and 9,364,566 (15.3%) CG-eliminating variants (Figure 3A, B). The proportion of CG-altering variants was consistent across allele frequency categories, indicating that CG context changes occur broadly across the frequency spectrum (Supplementary Fig. S6).

At the population level, 416,796 variants were both CG context–related and common in Koreans (allele frequency $\geq$ 0.8; Figure 3C), including 212,621 CG-creating and 206,845 CG-eliminating variants. A smaller subset (108,824 variants; 0.17%) simultaneously created and eliminated CG motifs through adjacent substitutions or small indels, leading to positional motif shifts without altering total CG abundance, an underappreciated source of bias in methylation-based analyses.

Regional aggregation of CGVs in 10-kb windows revealed that 97.6% of the genome differed by <5% in CG density relative to GRCh38 (Supplementary Fig. S7), whereas 408 regions (0.15%) exhibited >10% deviation (266 increases, 142 decreases; Figure 3D). The top five CG-altering “hotspots” overlapped functionally important loci, including major histocompatibility complex (MHC) and leptin receptor (LEPR) regions, both key regulators of immune and metabolic homeostasis (Abella et al., 2017). Notably, LEPR exhibited a 23.1% relative CG gain, ranking among the top loci (Supplementary Table S5), suggesting potential methylation-mediated modulation of expression.

Comparison with high-coverage 1KGP European genomes (N = 522) showed that 72.8% of the CGVs in the hotspot regions (with >10% deviation against GRCh38) were significantly more frequent in Koreans (2,078/2,853 variants; CG-creating: 1,320/1,803; CG-eliminating: 758/1,050; Figure 3E, F).

Genes within CG-gain regions were enriched for functions related to neuropsychiatric disorders, synaptic signaling, and metabolic regulation (Supplementary Fig. S8, 9).

Together, these findings establish a catalog of Korean-specific CG context-associated variants (CGVs) that reshape the CG landscape at the population level. This resource supports variant-aware epigenetic modeling, enhances population-specific methylation probe design, and provides a foundation for interpreting genetic-epigenetic interactions underlying disease risk and regulatory evolution in East Asians.

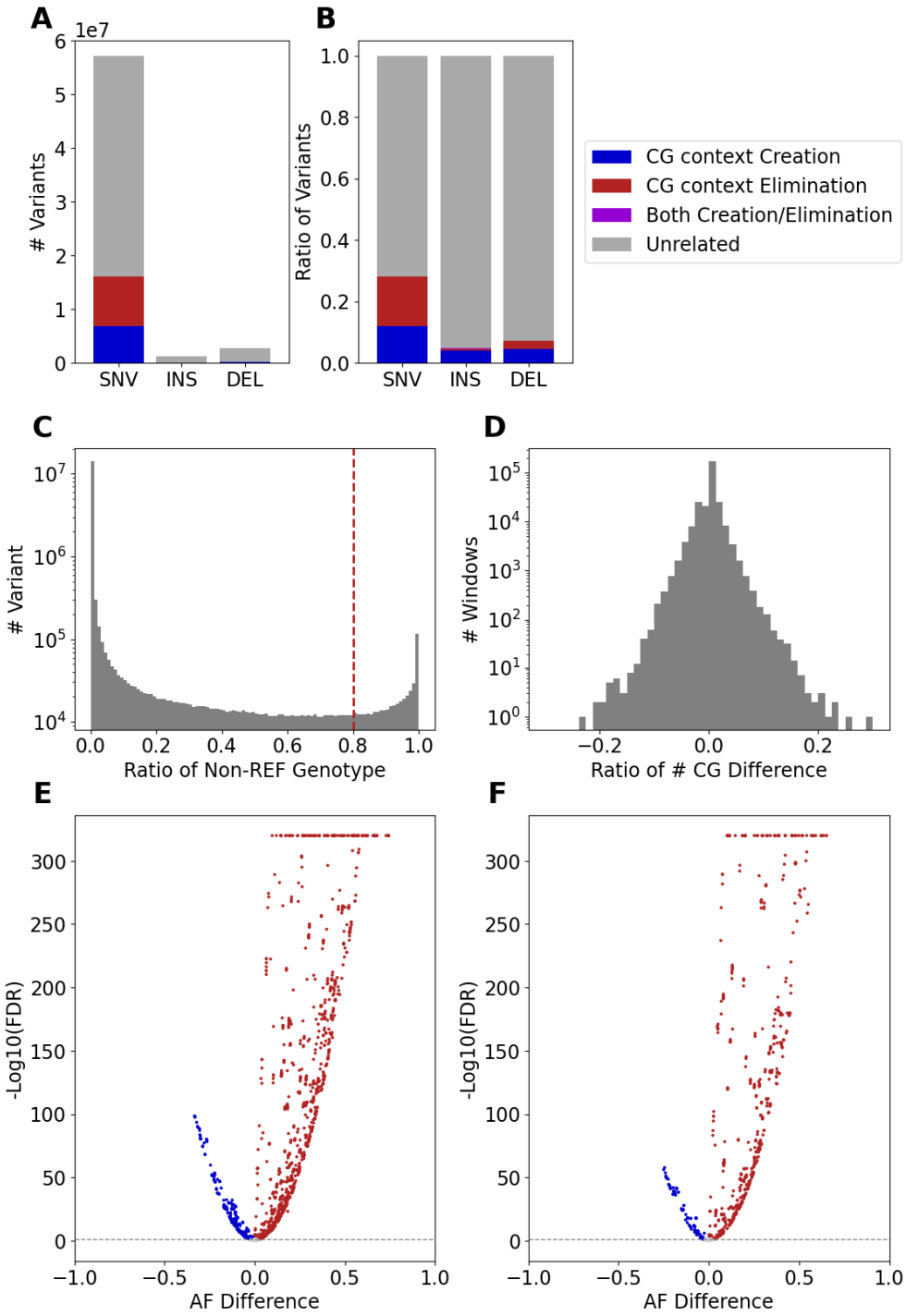


Figure 3. Population-specific CG context-associated variants (CGVs) in Koreans. (A) Total number (y-axis) of single nucleotide variants (SNVs), insertions (INS), and deletions (DEL) (x-axis) affecting CG dinucleotide context. **(B)** Proportion (y-axis) of different variant types (x-axis) categorized by CG context impact: creation (blue), elimination (red), both creation/elimination (purple), and unrelated (gray). **(C)** Distribution of variant counts (y-axis) over the ratio of non-reference genotypes (x-axis) across all individuals (dashed line = Korean-popular threshold, 0.8). **(D)** Relationship between the number (y-axis) and ratio (x-axis) of CG site differences per genomic window. **(E-F)** Volcano plots showing allele frequency (AF) differences between Koreans and Europeans (x-axis) for CG-creating and CG-eliminating variants, respectively, highlighting population-specific enrichment (y-axis, as negative of $\log_{10}$FDR).

Korea10K as a Genomic, Multi-omic, and Phenomic Data Resource

The Korea10K project represents the most comprehensive phase of the Korean Genome Project (KGP), expanding upon earlier phases including KPGP (N = 118), Korea1K (N = 1,094), and Korea4K (N = 4,157). Korea10K established a nationwide, population-scale genomic and multi-omic resource comprising 10,239 high-depth whole-genome sequences (WGS) from Korean individuals (Figure 4A). Participants were recruited through multiple regional centers and institutional partners across sequential collection phases, achieving broad demographic representation and phenotypic diversity.

Among all participants, 44.9% (N = 4,596) were healthy volunteers without apparent disease onset, whereas 55.1% (N = 5,643) had clinically verified diagnoses (Figure 4B). Disease sub-cohorts encompass cardiovascular, neuropsychiatric, metabolic, oncologic, and congenital conditions. Cardiovascular diseases represented the largest group (56.6%), including acute myocardial infarction (AMI) and angina, followed by mental and mood disorders such as depression and sleep disorders (29.4%), with additional subgroups to support disease-specific analyses (Figure 4C).

Korea10K marks a major expansion of the Korean Genome Project into the multi-omic era. Population-scale molecular profiles include 868 whole-blood transcriptomes, 1,211 DNA methylomes, and 102 chromatin accessibility datasets, and six blood proteomes, providing a systems-level view of regulatory effects of genetic variation and mechanistic insights into phenotypes. Complementary phenotypic resources include the Health Check-up (HC) dataset (6,583 participants, 336 phenotypes; Supplementary Table S3) and the Lifestyle Questionnaire (LQ) dataset (4,923 participants, 370 survey measures; Supplementary Table S4), encompassing diet, physical activity, smoking, alcohol use, sleep, and psychosocial factors (Table 1).

The core of Korea10K is a set of omics reference data. KOREF (KOrean REFerence) represents one of the most precise longitudinal omics profiling resources available in Korea. KOREF1 (middle-aged donor, PGP ID of hu3D760A) and KOREF-S2 (younger donor) provide high-precision datasets, including 39 WGS, 36 epigenomes (6 Hi-C, 19 DNA methylomes, 11 ATAC-seq), 73 transcriptomes, 6 proteomes (SomaScan assays), 6 microbiomes (four 16S rRNA-seq, two metagenomic WGS), and 12 phenomes (10 clinical records and two lifestyle information datasets), collected over a decade and a half-decade, respectively. Also, KOREF1 is the first Korean multi-omic reference accompanied by the first fully phased diploid Korean genome, achieving high-accuracy and telomere-to-telomere completeness (unpublished). KOREF2 constitutes the next-generation reference with a rich set of third-generation sequencing data (PacBio HiFi and ultra-long Oxford Nanopore data) toward near-complete assembly, improving mappability of the multi-omic data for highly accurate reference-based analyses.

Collectively, Korea10K provides an extensive resource capturing population-scale genomic and multi-omic variation representative of the Korean population, complemented by reference-quality longitudinal datasets that enable molecular resolution and temporal precision.

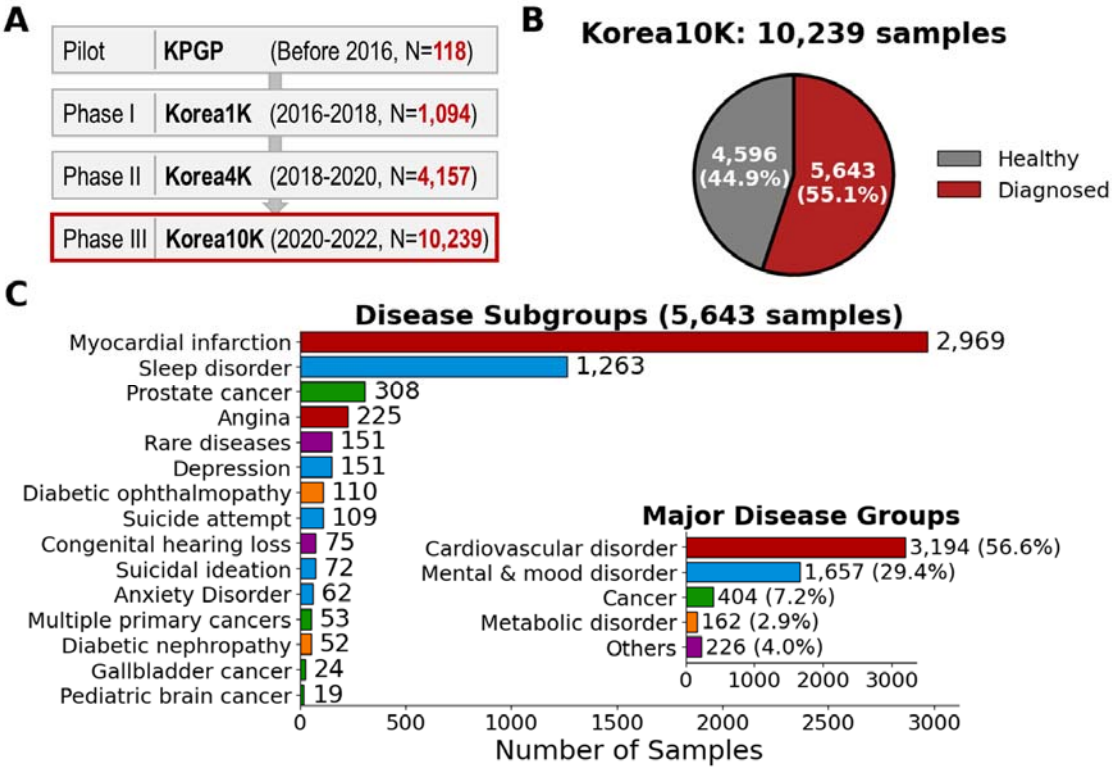


Figure 4. Overview of the Korea10K cohort. (A) Sequential recruitment phases of the Korea10K project, expanding from a pilot study (KPGP) to large-scale national inclusion through Korea1K, Korea4K, and Korea10K, comprising 10,239 high-depth whole genomes collected in total. **(B)** Distribution of healthy and clinically diagnosed 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.

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

Table 1. Overview of Multi-omic and Phenomic Data Resource in Korea10K. PBMC= Peripheral Blood Mononuclear Cell.

Discussion

There are several limitations to this study. From a design perspective, Korea10K is currently a cross-sectional resource except two KOREF samples; the absence of longitudinal outcome data precludes tracking of biological aging, disease progression, and causal inference over time. Although we classified participants in Korea10K as either healthy or clinically diagnosed, the putative healthy group cannot be assumed to be physiologically normal or free of undetected conditions, and diagnostic misclassification remains possible within the clinically diagnosed cohort despite clinician verification. More precise, data-driven stratification of participants into biologically informed subgroups will be critical for improving the resolution and interpretability of downstream association analyses (Ahlqvist et al., 2018). Clinical and lifestyle questionnaire data were collected only for a subset of individuals directly enrolled through participating institutions. Samples acquired via Material Transfer Agreements (MTAs) from external Korean biobanks lacked detailed phenotypic information. This incomplete metadata reduced the effective sample size for genome-wide association studies (GWAS), limiting power for large-scale genotype–phenotype mapping beyond the Korea4K dataset. Budgetary constraints restricted the generation of additional multi-omic datasets, limiting systems-level interpretation of genetic effects. For instance, population-matched whole blood RNA-seq would enable transcriptome-wide association analyses (TWAS) and expression QTL (eQTL) mapping tailored to the Korean population, improving fine-mapping and gene prioritization (Gusev et al., 2016). DNA methylation profiling and ATAC-seq would allow investigation of methylation QTLs (meQTLs) and chromatin-accessibility QTLs (caQTLs) facilitating genome-epigenome integration to reveal the core regulatory mechanisms (Banovich et al., 2014; Gate et al., 2018).

Technical limitations also remain. Reliance on short-read sequencing and an incomplete linear reference genome (GRCh38) left substantial genomic regions, including centromeres, sub-telomeres, segmental duplications, and the Y chromosome, poorly resolved. Consequently, structural variants, complex rearrangements, and tandem-repeat expansions were incompletely captured. Reference-based epigenomic analyses similarly remain constrained by reference bias. CG context-associated variants (CGVs) delineate regions where the Korean genomic CG landscape diverges from the GRCh38 human reference genome, limiting accurate characterization of population-specific methylation states and chromatin architecture. The prevalence of CGVs underscores the need for sequence-context aware modelling. Adoption of long-read sequencing (Kim et al., 2025), ethnically-specific (or even

personal) assemblies such as KOREF1 (Cho et al., 2016), and graph-based pangenomes (Liao et al., 2023) will also be essential for achieving comprehensive variant representation and their accurate functional annotation.

Future expansions incorporating longitudinal phenotyping, long-read sequencing, and multi-omic profiling will enable more precise capture of structural variation, population-specific epigenetic landscapes, and elucidation of molecular mechanisms underlying aging and disease. We expect that cross-platform multi-omics signatures will comprehensively elucidate the gene regulatory mechanisms underlying Korean-specific disease biomarkers through integrated pathway analysis. The development of AI-based models trained on the expansive multi-omics and phenotypic datasets will contribute to advancing both general and Korean-tailored precision medicine.

Despite limitations, Korea10K extends the scale and depth of Korean population genomics. Compared with Korea4K with 4,157 individuals (3,617 unrelated, 45 million variants), the cohort expanded to 10,239 genomes (9,000 unrelated) identified 61 million, including nearly 10 million previously unseen alleles, approaching the saturation of ultra-rare alleles. This near-complete discovery of low-frequency and private variants improved imputation performance and fine-mapping resolution, thereby enhancing the statistical power of future GWAS and rare-variant association studies in Koreans.

Population structure analyses confirmed again that Koreans form a distinct cluster within East Asia. Autosomal PCA positions them along the East Asian continuum, adjacent to Japanese and northern Han Chinese populations, reflecting shared ancestry while preserving a unique genetic profile. Rapid saturation of common autosomal variants evidences long-term isolation and demographic continuity. Y-chromosomal haplogroups point to recent male-mediated expansions consistent with Neolithic agricultural dispersals (Kim et al., 2011). Big mitochondrial diversity preserves signatures of ancient female-mediated admixture and Bronze Age demographic expansions (Kim et al., 2020). Immunogenetic profiles also reveal a broadly shared dominance of HLA-A alleles across northern East Asia, with a selective enrichment of A*33:03 in Koreans. Population-specific alleles suggest local selection pressure from endemic pathogens, with implications for vaccine response (Crocchiolo et al., 2022), autoimmune disease susceptibility (Kang et al., 2011), and drug contraindications (Li et al., 2017). Collectively, Koreans are genetically homogeneous, yet historically interconnected population shaped by layered demographic processes across northern East Asia. The exceptional homogeneity of Koreans minimizes population stratification bias and enhances the power, precision, and reproducibility of genomic and multi-omic association studies (Hellwege et al., 2017).

Beyond genomics, Korea10K incorporates a multi-omic dimension. The inclusion of detailed clinical and lifestyle metadata also extends its translational interpretability. The project captures population-level breadth while achieving individual-level depth through reference- and near-reference-quality multi-omic datasets called KOREFs. Korea10K thus bridges population genetics and precision medicine, enabling exploration of genetic and environmental architectures underlying disease risk and aging.

Korea10K has already facilitated discovery of genomic and multi-omic biomarkers associated with diverse phenotypes, including cardiovascular disorders (acute myocardial infarction) (Bhak et al., 2021; Jeon, Jeon, An, et al., 2023; Jeon, Jeon, Choi, et al., 2023; Jeon et al., 2025), neuropsychiatric traits (depression, anxiety, and suicidal behavior) (Bhak et al., 2019; Kwon et al., 2025), infectious diseases (COVID-19) (Ryu et al., 2025), and biological aging (An et al., 2025), providing mechanistic insights into disease etiology and progression.

By leveraging genetic homogeneity, high-resolution multi-omic data, and deep phenotyping, Korea10K provides a robust platform for biomarker discovery, disease mechanism elucidation, and the development of precision medicine strategies in Korea.

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 (Thermo Fisher Scientific) and Bioanalyzer 2100 (Agilent) for library preparation.

Whole-Genome Sequencing

Whole-genome sequencing (WGS) was performed at a mean depth of 30× using MGI DNBSEQ-T7 (N = 6,311; 61.6%), Illumina NovaSeq 6000 (N = 2,832; 27.7%), and HiSeq X Ten (N = 1,096; 10.7%) (Supplementary Fig. S10; Supplementary 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 v0.23.2 (Chen et al., 2018) 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 v0.7.17 (Li & Durbin, 2009) with read group information. Alignments were coordinate-sorted and duplicates marked with Picard v2.27.5 (*Picard Toolkit*, 2019). Base quality score recalibration was performed using GATK BaseRecalibrator and ApplyBQSR v4.6.1.0 (Poplin et al., 2018).

Variant Calling and Joint Genotyping

Per-sample variant calling was performed using GATK HaplotypeCaller v4.6.1.0 in GVCF mode, adjusting ploidy for sex chromosomes. GVCFs from 10,239 samples were aggregated using GenomicsDBImport and jointly genotyped with GATK v4.6.1.0. Post-calling, variant quality score recalibration (VQSR) was applied separately to SNPs and indels using dbSNP 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 from PLINK v2.0.0-a.6.13LM (Chang et al., 2015), 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-value < 1×10^{-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) v115 (McLaren et al., 2016).

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) (Hofmeister et al., 2023), and imputation was performed with Minimac3 v2.0.1 (Das et al., 2016). 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) (Taliun et al., 2021). 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 their imputation performance against same dataset as in the previous study (Jeon et al., 2020).

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 (Genomes Project et al., 2015). 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 PLINK2 (--extract-intersect), ensuring positional and allelic concordance prior to merging. The two VCFs were then merged using

PLINK v1.9.0-b.7.11 (--bmerge) (Purcell et al., 2007), 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□□, or (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 v1.3.0 (Alexander et al., 2009) 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 PLINK1 (--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 Korea10K samples 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 (Supplementary Fig. S2).

Y-chromosome phylogenetic reconstruction and haplogroup classification

For 5,975 male Korean participants, Y-chromosome variants were lifted to GRCh37 using GATK LiftoverVcf v4.6.1.0 and analyzed with yhaplo v2.1.14.dev7 + gf8b300b (Poznik, 2016) 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 v1.54 (Xu et al., 2025) 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-TREE3 v3.0.1 (Nguyen et al., 2015).

Mitochondrial DNA assembly and quality control

Filtered mitochondrial reads (chrM) were extracted from whole-genome BAM files using samtools v1.18 and converted to paired-end FASTQ format. Mitochondrial genomes were then assembled using NOVOPlasty v4.3 (Dierckxsens et al., 2017), 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 v1.21 (consensus) (Bonfield et al., 2021).

Annotation of 10,166 circularized assemblies to check their completeness was performed using MitoZ v3.5 (Meng et al., 2019) 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 Haplogrep3 v3.2.1 (Schonherr et al., 2023) with the PhyloTree Build 17.2 database (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 (Durbin et al., 2023). Multiple sequence alignment was conducted using MAFFT v7.525 (Katoh & Standley, 2013) with the --auto parameter.

Phylogenetic tree reconstruction was performed using IQ-TREE3 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).

Divergence time estimation was subsequently conducted in MEGA12 (Kumar et al., 2024) using the RelTime framework. For this analysis, we imported the IQ-TREE Newick file and applied four previously reported calibration constraints corresponding to haplogroups A, B, C, and D, each defined by established time intervals (Bonatto & Salzano, 1997). RelTime analysis was implemented under a maximum-likelihood statistical model.

To characterize rate variation across sites, we employed a General Time Reversible (GTR) nucleotide substitution model with gamma-distributed rate heterogeneity and a proportion of invariant sites (G+I). Rate variability was approximated using four discrete gamma categories. Sites with excessive missing data were filtered using partial deletion, retaining positions with at least 95% coverage. All RelTime computations were performed with 3 threads, enabling efficient processing of the large mitochondrial dataset.

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 v1.13 (Danecek et al., 2021) 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 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 (https://github.com/humanlongevity/HLA). 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.

Comparing Allele Frequencies of CGVs Across Populations

To evaluate population-specific differences in CGVs, allele frequencies were compared between Koreans and individuals of European ancestry from the 1000 Genomes Project (1KGP). Variant allele counts were obtained using PLINK2 without prior biallelic filtering. Multiallelic variants were decomposed into biallelic representations and left-normalized using bcftools v1.21 for coordinate consistency with the Korean dataset.

For genomic windows in Koreans exhibiting ≥ 10% CG context alteration, we selected variants whose directional effects matched the overall trend in that region (i.e., variants increasing CG content in gain regions or reducing CG content in loss regions). Differences in allele counts between Korean and European populations were assessed using chi-squared tests. Resulting P-values were adjusted for multiple comparisons using the Benjamini–Hochberg false discovery rate (FDR) method.

Genomic Annotation and Functional Enrichment of Hotspot CGVs

Genomic intervals exhibiting substantial CG context alteration were annotated using Annotatr [ref] with the “hg38_cpgs”, “hg38_basicgenes”, and “hg38_enhancers_fantom” annotations to map proximal genes and regulatory features. Enrichment analyses for Gene Ontology (GO) terms and pathway associations were performed using ShinyGO (Ge et al., 2020) and GeDiPNet (Kundu et al., 2023). Significance of enrichment was assessed using FDR correction, with terms considered significant at FDR < 0.05.

Abbreviations

ABHet = Heterozygotic Allele Balance

AC = Allele Count

AF = Allele Frequency

AFR = African

AI = Artificial Intelligence

Alt = Alternative allele

AMR = Ad Mixed American

BEB = Bengali in Bangladesh

CDX = Chinese Dai in Xishuangbanna, China

CGV = CG context-associated Variant

CHB = Han Chinese in Beijing, China

CHS = Han Chinese South, China

CLM = Colombian in Medellin, Colombia

COVID-19 = Coronavirus Disease 2019

EAS = East Asian

EBI = The European Bioinformatics Institute

EUR = European

ESN = Esan in Nigeria

FDR = False Discovery Rate

FIN = Finnish in Finland

GATK = Genome Analysis Tool Kit

GBR = British from England and Scotland

GeDiPNet = Genes, Diseases and Pathway Networks

GIH = Gujarati Indians in Houston, Texas, USA

gnomAD = Genome Aggregation Database

GO = Gene Ontology

GVCF = Genomic Variant Call Format

GWAS = Genome-Wide Association Study

GWD = Gambian in Western Division - Mandinka

HC = Health Check-up

HLA = Histocompatibility Leukocyte Antigen

HWE = Hardy-Weinburg Equilibrium

IBS = Iberian populations in Spain

ISOGG = International Society of Genetic Genealogy

ITU = Indian Telugu in the UK

JPT = Japanese in Tokyo, Japan

KGP = Korean Genome Project

KHV = Kinh in Ho Chi Minh City, Vietnam

KOR = Korean in Korea10K

KOREF = Korean Reference

KPGP = Korean Personal Genome Project

KEGG = Kyoto Encyclopedia of Gene and Genome

LD = Linkage Disequilibrium

LEPR = Leptin Receptor

LQ = Lifestyle Questionnaire

LWK = Luhya in Webuye, Kenya

MAF = Minor Allele Frequency

MFP = ModelFinder

MHC = Major Histocompatibility Complex

MSL = Mende in Sierra Leone

MTA = Material Transfer Agreement

MXL = Mexican Ancestry in Los Angeles CA USA

PBMC = Peripheral Blood Mononuclear Cells

PCA = Principal Component Analysis

PEL = Peruvian in Lima, Peru

PUR = Punjabi in Lahore, Pakistan\

QC = Quality Control

QTL = Quantitative Trait Loci

R^2 = Squared Correlation Coefficients

rCRS = revised Cambridge Reference Sequence

RSRS = Reconstructed Sapiens Reference Sequence

SAS = South Asian

SD = Standard Deviation

SLC15A5 = Solute Carrier Family 15 Member 5

SNV = Single Nucleotide Variant

TOPMed = Trans-Omics for Precision Medicine

TSI = Toscani in Italia

TWAS = Transcriptome-Wide Association Study

VEP = Variants Effect Predictor

VQSR = Variant Quality Score Recalibration

WGS = Whole-Genome Sequencing

YRI = Yoruba in Ibadan, Nigeria

1KGP = The 1000 Genome Project

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, Seolbin An, 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 are grateful to Kiyuk Chang, Sang Yeub Lee, Weon Kim, Min Gyu Kang, Ae-Young Her, and Moo Hyun Kim for their contributions to the collection of cardiovascular disease samples. We thank Soo-Jung Shin, Byung-Ryel Yoo, Jeong-Sik Yang for the deep involvement initiating this project. 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 NCSRD, KRISS. The biospecimens and clinical information for this study were provided by Ulsan University Hospital Biobank (60SA2016001-002, 60SA2016001-003, 60SA2016001-005, 60SA2016001-010-5, 60SA2017002-001, 60SA2017002-004, and 60SA2017002-006-4), Ulsan Medical Center and the Biobanks of Gyeongsang National University Hospital, Chungbuk National University Hospital (18-27, 20-04), Kyung Hee University Hospital (2018-4, 2019-4, 2019-6), and the members of the National Biobank of Korea, which is supported by the Ministry of Health, Welfare and Family Affairs.

Funding Information

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].

Author Contributions

K.A., S.Jeon, and Y.K. contributed equally to this work and jointly led the study, including conceptualization, design, data analysis, and manuscript writing. C.Y., Y.Choi., H-J.C., J.H.B., D.S., H.L., Y.J.K., and Semin L. contributed to primary data analyses and interpretation. Members of the Korean Genomics Center (KOGIC; S.Jeon, Y.K., K.A., C.Y., Y.Choi, H-J.C., Jihun B., D.S., Hyomin L., S.P., J.C., H.R., and Semin L.) and Genome Research Foundation (GRF; B-C.K. and J.K.) contributed to sample collection, sequencing, data processing, quality control, and additional analyses. The Korea Bioinformation Center (KOBIC; Y.Jeon., J-H.K., S.Joe, J.O.Y., J.J.) contributed to project coordination, data management, and computational support. Clinical collaborators from affiliated hospitals and institutions (E-S.S., A.B., D.B., N-H.P., Siwoo L., H-J.J., Y.B., D.J., Y.Cho., Y.M.A., H-T.J., B.H., Y.Y.C., S.A.C., S-K.K., J.H.P., H.Y.G., B.K., W.M., C.K.C., D.U.K., G-N.J., Hojune L., K.S.C., C.G.K., B.L., and J.L.) facilitated sample recruitment, clinical characterization, and interpretation of phenotypic data. S.J., Y.K., K.A., and J.B. wrote the manuscript with contributions from all authors. G.C., H.J., and J.B. supervised the study and provided overall project coordination.

Data Availability

Raw data of Korea10K will be made available through the Korea BioData Station (K-BDS), provided by the Korea Bioinformation Center (KOBIC), following publication. Prior to public release, access to the Korea10K data can be granted upon request under appropriate data use agreements. Access to individual-level data requires approval by the relevant Institutional Review Board and Data Access Committee in accordance with established international guidelines for genomic data sharing. The 1000 Genomes Project (1KGP) Phase 3 high-coverage (30×) dataset used in this paper was retrieved from the Amazon Web Services (AWS) server (https://registry.opendata.aws/1000-genomes). The 1000 Genomes Project (1KGP) Phase 1 legacy data can be found on their FTP server (https://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/).

Code Availability

Analysis scripts and computational workflows used in Korea10K data generation, and analysis are available through the project GitHub repository (https://github.com/korean-genomics-center/Korea10K). This includes variant calling pipelines, quality control procedures, population genetic analysis scripts, and imputation panel construction workflows.

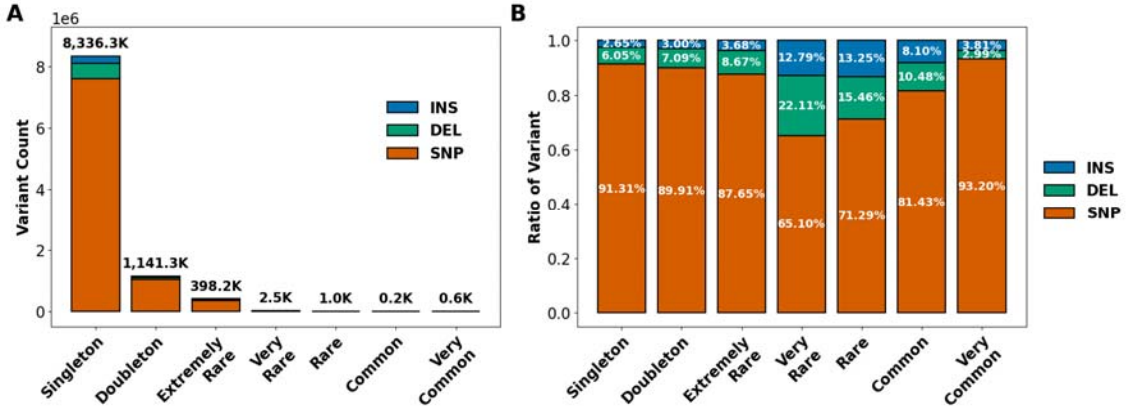
Competing Interests

S. Jeon is CEO and H.R. is an employee of AgingLab Inc. Y. Cho is an employee of CG Invites Co., LTD and Invites Genomics Co., LTD. B.L. is CEO and J.L. is an employee of nSAGE. G.C. is co-founder of Nebula Genomics and Glottatech.com. The remaining authors declare no competing interests.

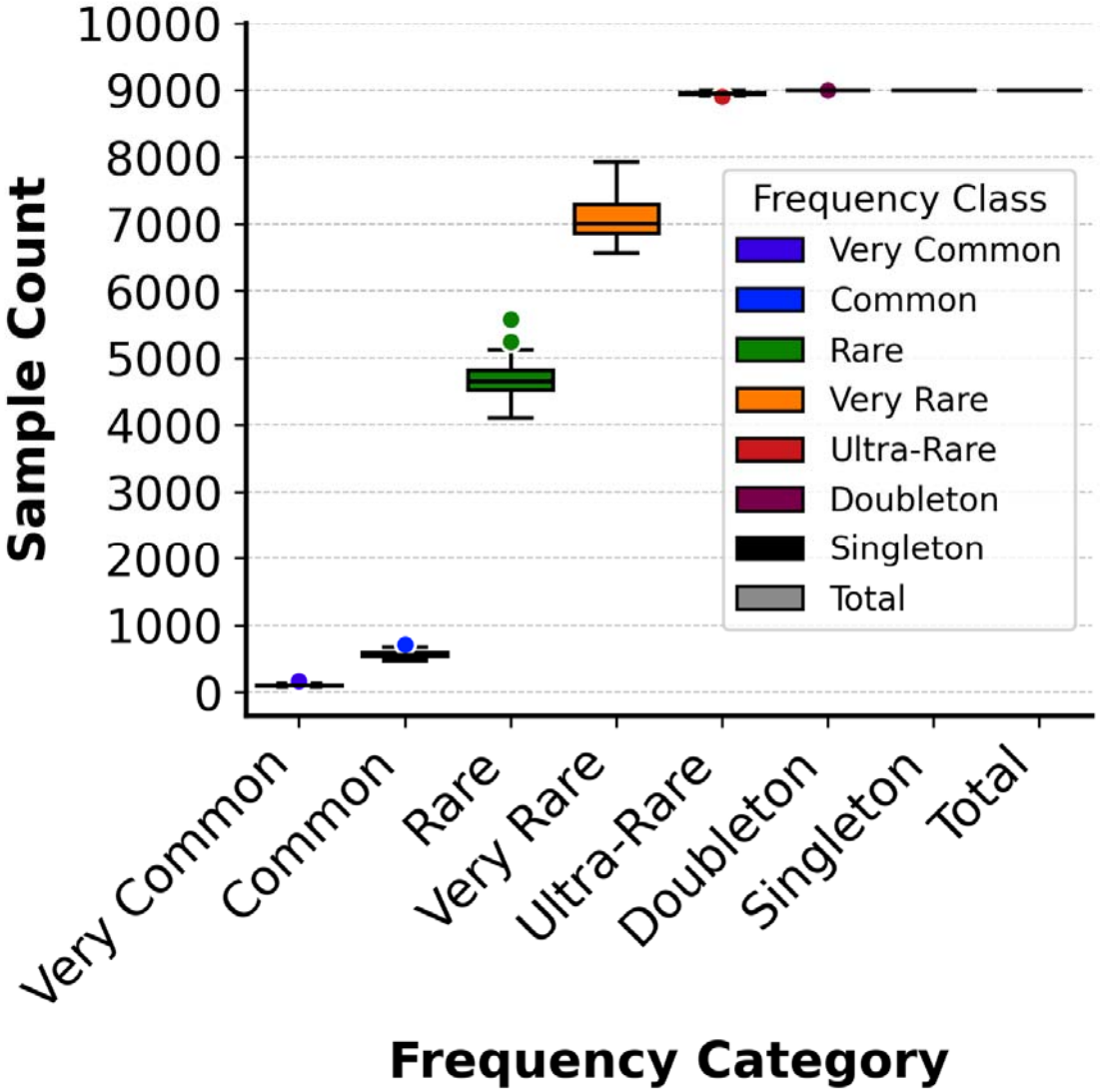
Ethics, Consent, and Permissions

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 all 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. Furthermore, our study adheres to the ethical guidelines and regulations stipulated by the IRB. As a result, we can make the data of 10,239 individuals available under controlled access while respecting the privacy and consent of the participants.

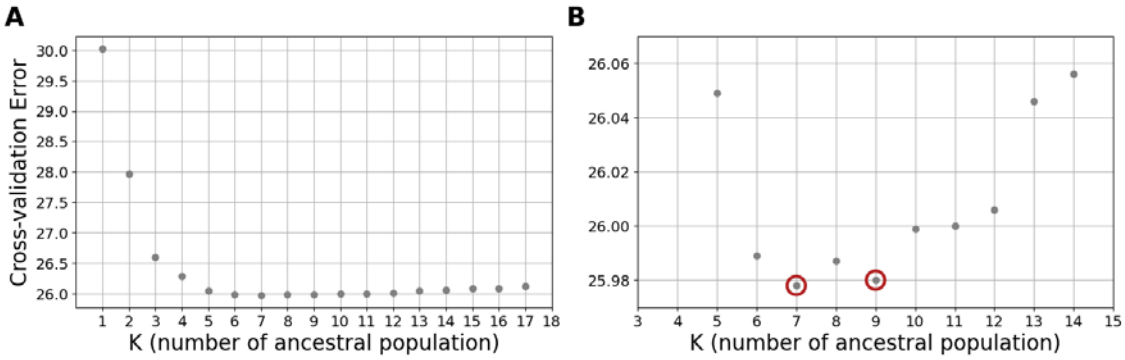
Supplementary Data



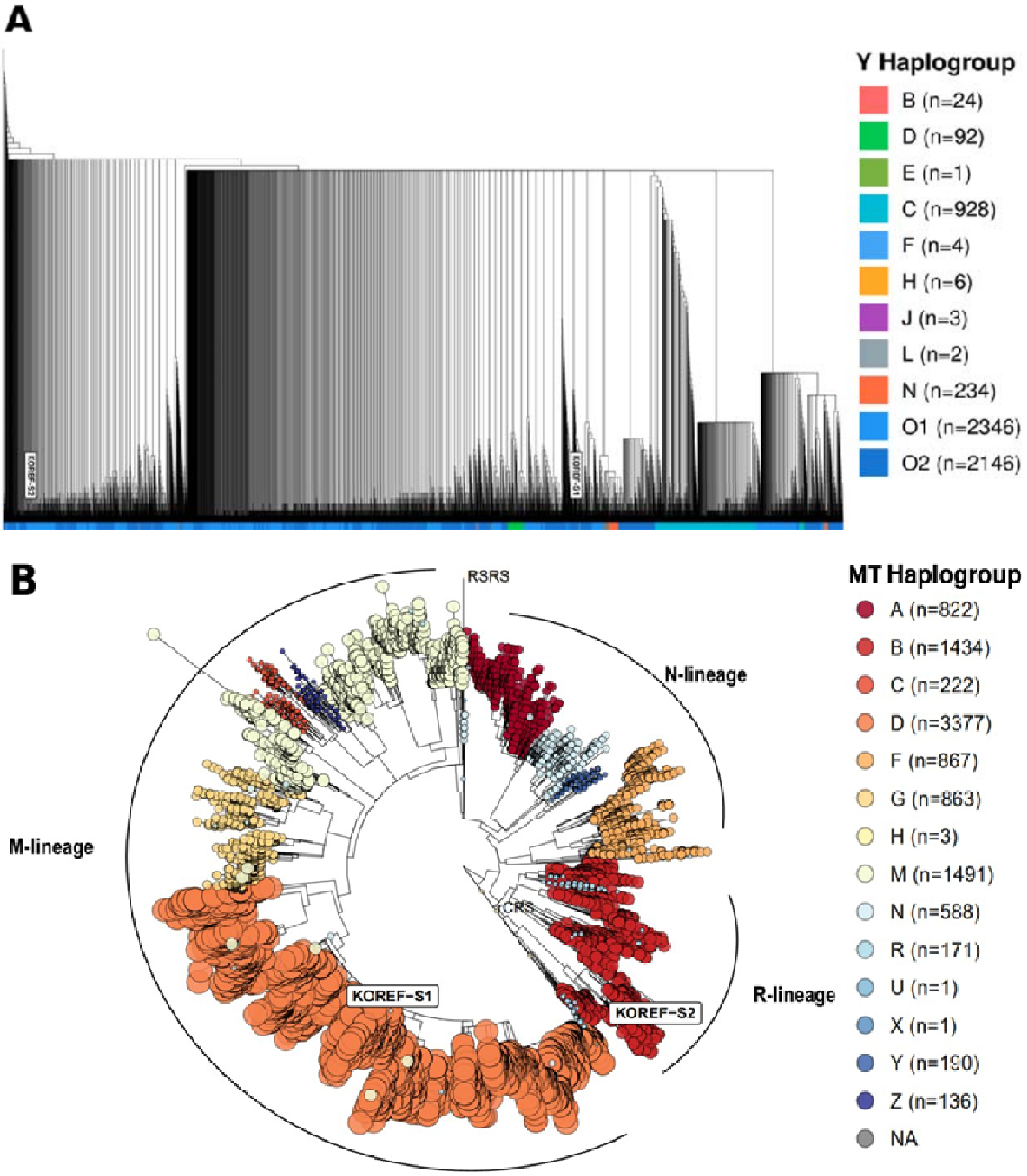
Supplementary Fig. S1. Characteristics of novel variants across allele frequency categories. Bar plots showing **(A)** count and **(B)** proportion of novel variants (y-axis) across allele frequency bins (x-axis) stratified by variant type: SNP (orange), deletion (DEL; green), and insertion (INS; blue). Variant counts are shown as stacked bars with absolute values annotated above, while proportions are displayed as relative contributions within each frequency bin.



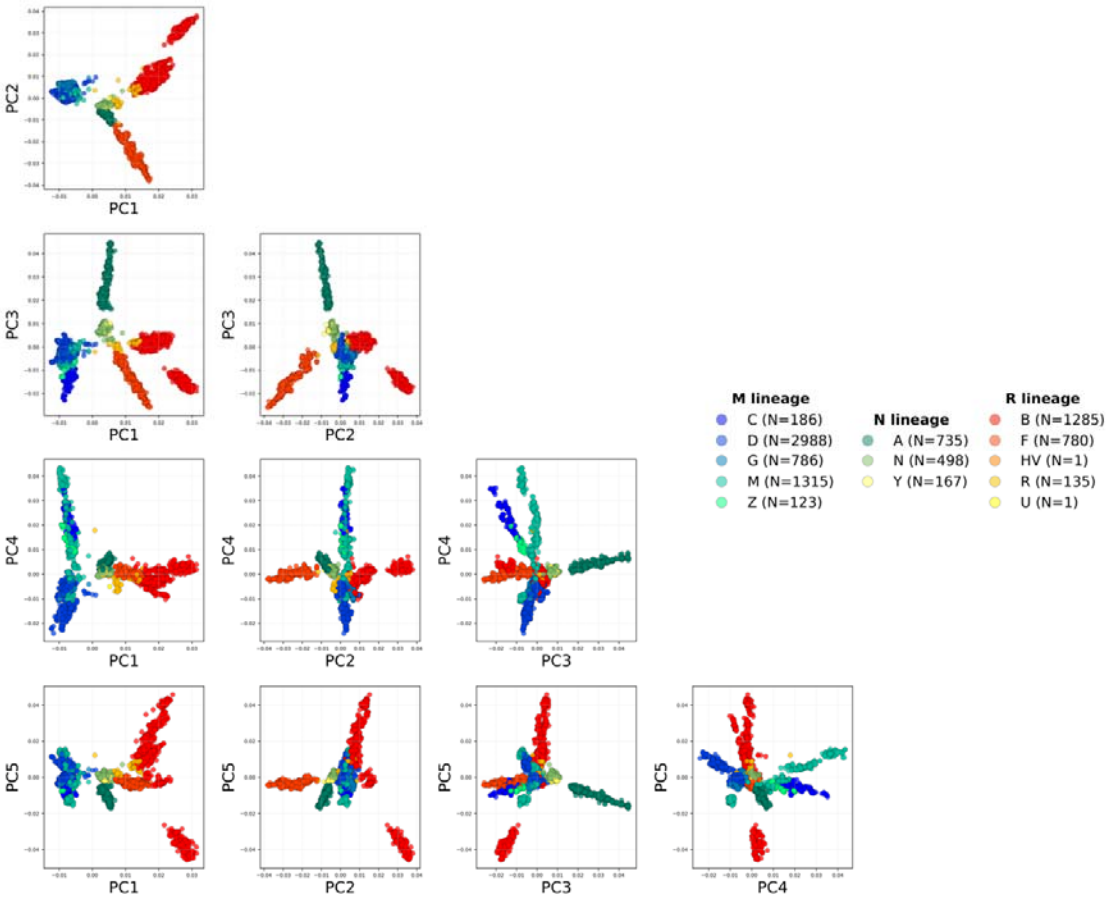
Supplementary Fig. S2. Saturation analysis across allele frequency categories in 9,000 unrelated Koreans. Boxplots depict the distribution of sample counts for variant discovery (y-axis) stratified by minor allele frequency (MAF) category (x-axis). Each box represents the interquartile range (IQR, 25th–75th percentile) with the median indicated by a horizontal line; whiskers extend to 1.5× IQR. Individual variants beyond the whiskers are shown as colored dots corresponding to frequency class. Sample counts reaching 9,000 samples, as exemplified by singleton and doubleton, indicates failure of the saturation in the cohort.



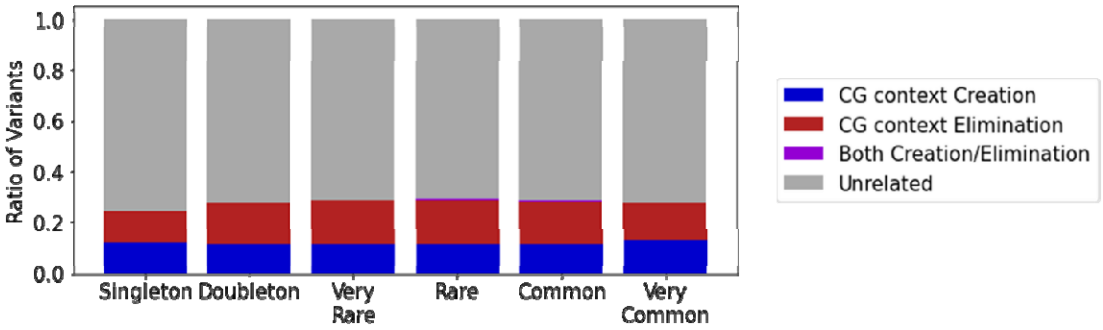
Supplementary Fig. S3. Selection of Optimal ADMIXTURE K-value(s) with minimal cross-validation errors. (A) Scatter plot showing the cross-validation (CV) error (y-axis) and K (number of hypothetical ancestral populations; x-axis). The CV error decreases sharply, reaching a plateau around $K=7$, suggesting this value captures the major structure in the dataset. **(B)** Magnified view of the cross-validation error between K = 3 and K = 15. The plot highlights two distinct local minima at K = 7 and K = 9 (indicated by red open circles).



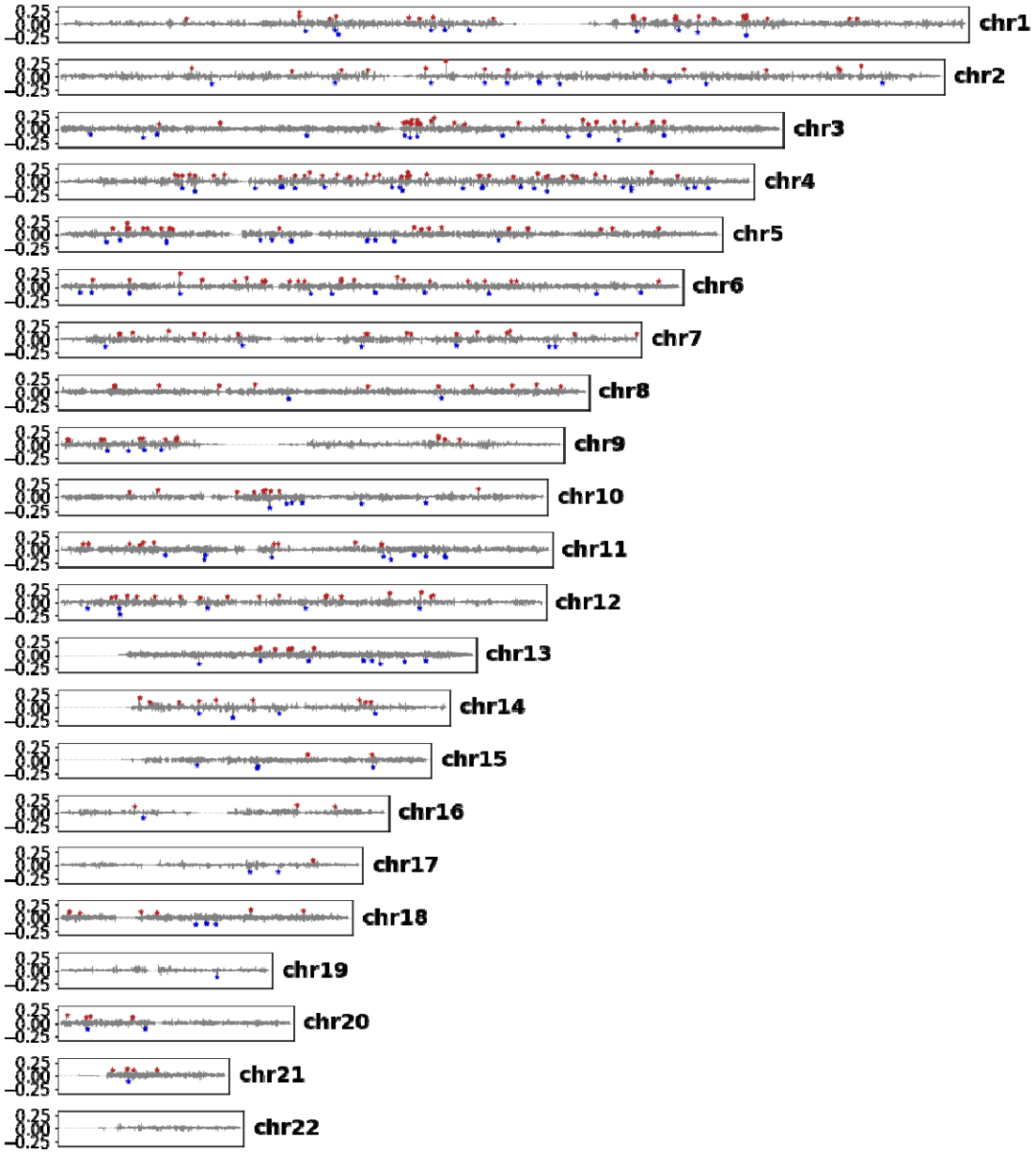
Supplementary Fig. S4. Phylogenetic reconstruction of Korean uniparental lineages in Korea10K. (A) A Y-chromosomal lineage tree based on p-distance showing 5,871 Y haplogroup-classified Korean males. **(B)** A mitochondrial lineage tree based on maximum likelihood portraying 10,166 MT haplogroup-assigned Korean individuals. Reference sequences, including Korean (KOREF1 or -S1 and KOREF2 or -S2; S = Single Individual), European (rCRS), and hypothetical human ancestor (RSRS) are annotated within the trees.



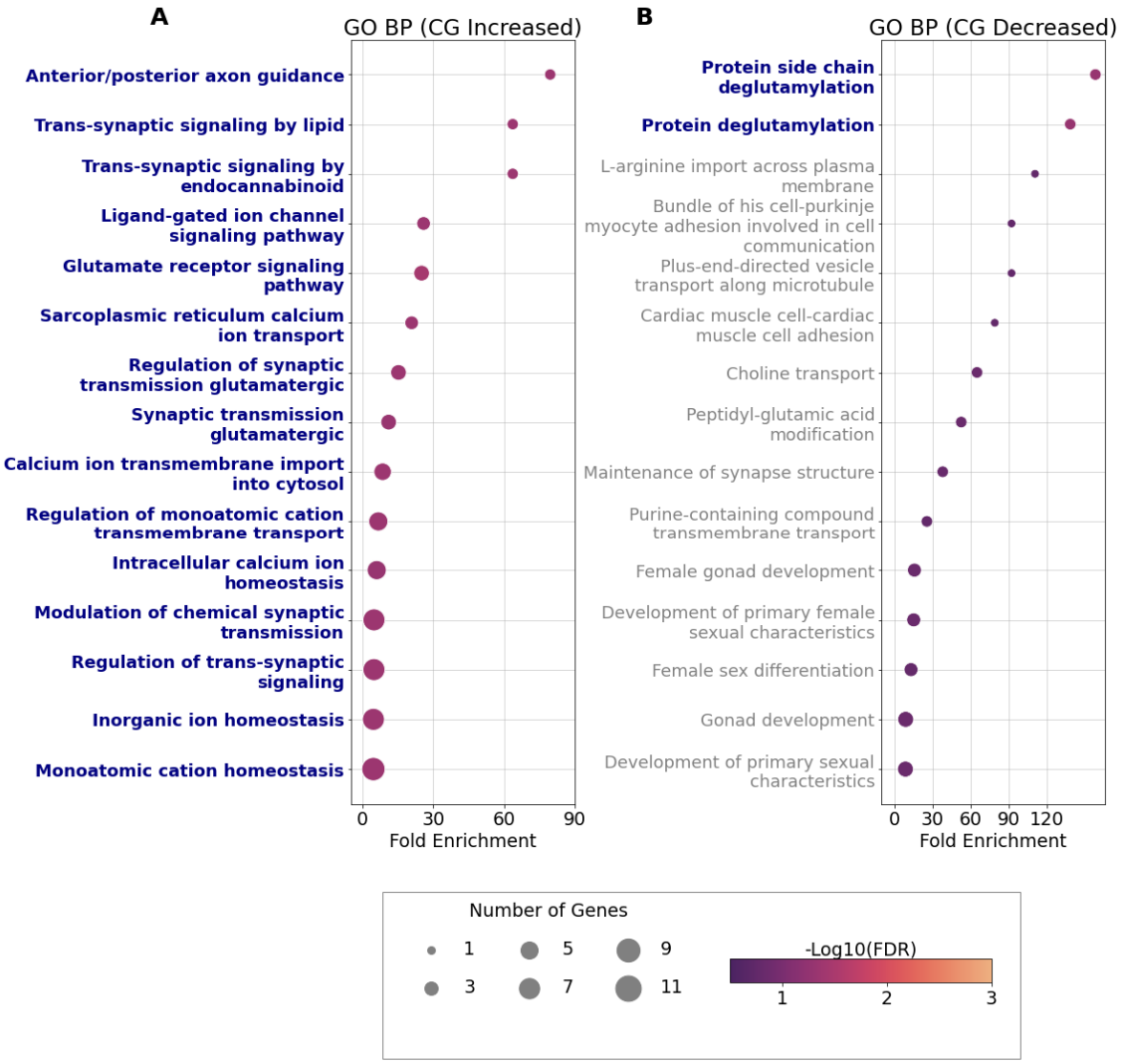
Supplementary Fig. S5. Combinatorial PC-biplots of mitochondrial variants in Korea10K. Each biplot of PCs depicts the distribution of mitochondrial variants along principal component axes (from PC1 to PC5 in a combinatorial manner), highlighting sample clustering according to haplogroup assignments. Colors indicate ancestral lineages (macrohaplogroups) corresponding to each haplogroup.



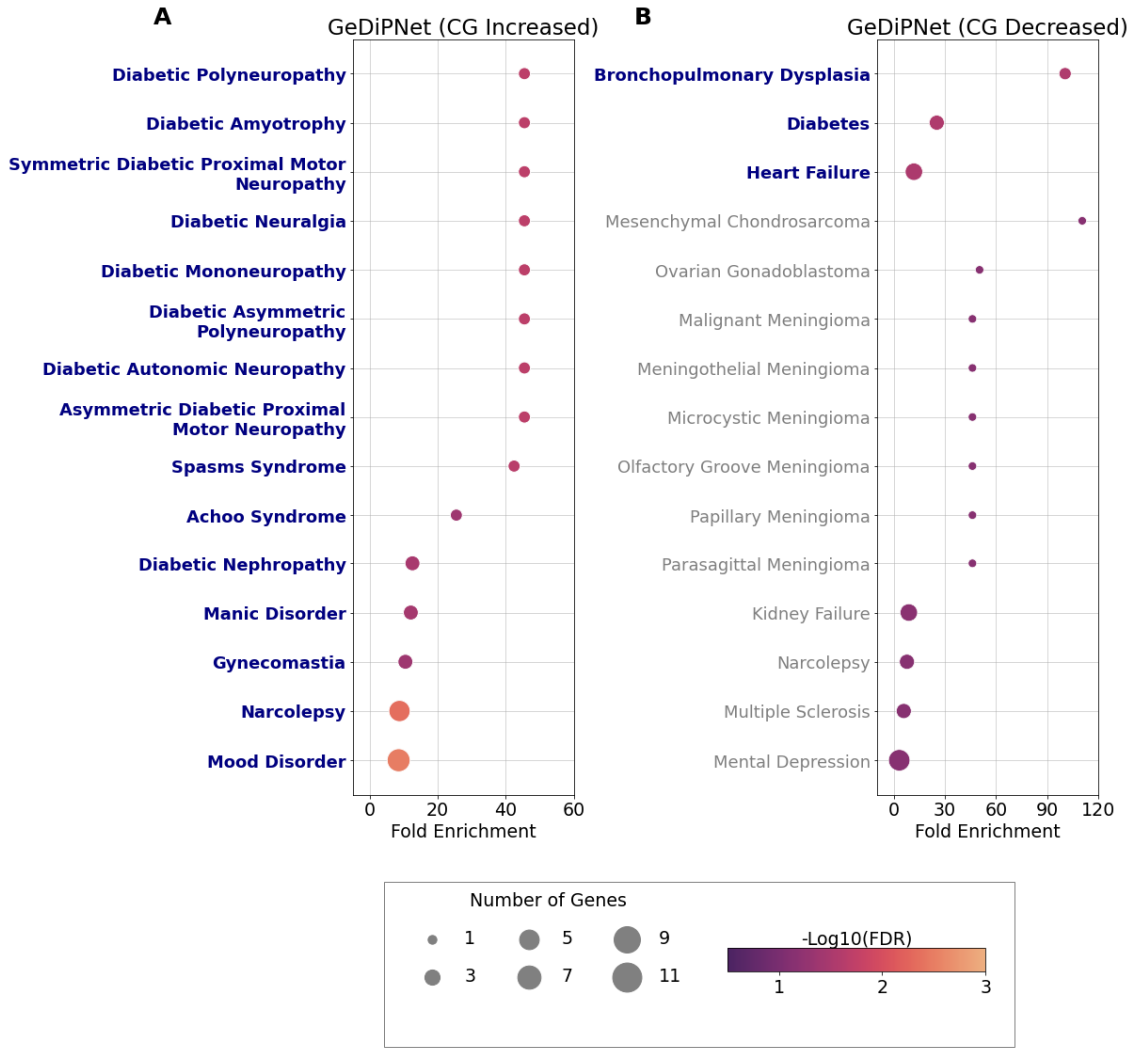
Supplementary Fig. S6. Bar plot depicting the characteristics of CGVs across allele frequency categories. Proportion of CGVs (y-axis) across allele frequency bins (x-axis) categorized by CG context impact: creation (blue), elimination (red), both creation/elimination (purple), and unrelated (gray).



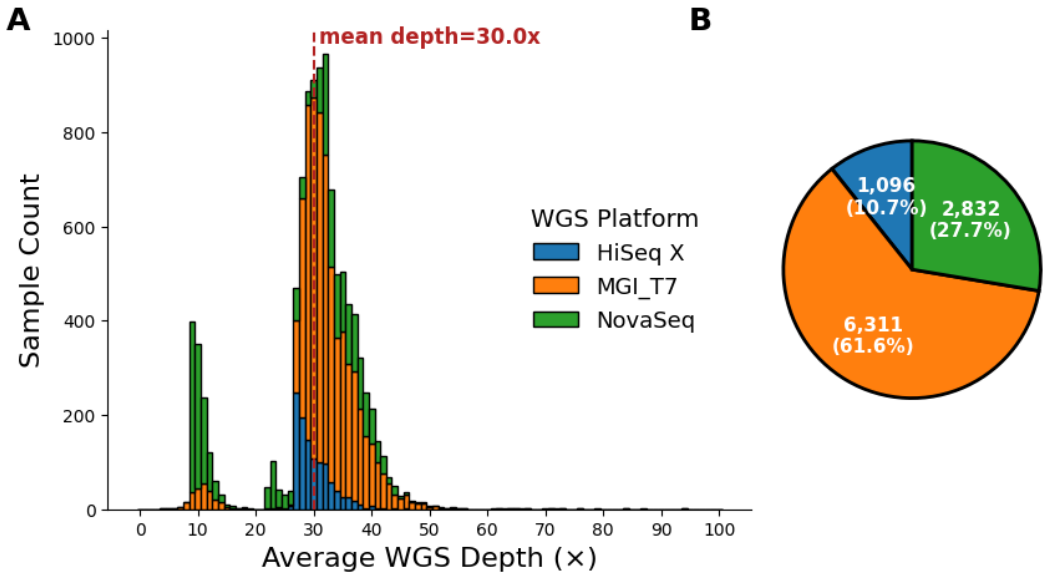
Supplementary Fig. S7. Ratio difference of CG context by chromosome region. Visualization of the proportion of CG context changes (y-axis) across 10,000 bp windows for each chromosome in GRCh38 (x-axis). The y-axis represents the relative difference (ratio) of CG context compared to GRCh38. Regions showing more than a 10% change are marked with stars.



Supplementary Fig. S8. Gene set enrichment analysis of regions with more than 10% relative change in CG context (GO Biological Process). (A) Regions with a relative increase over 10%; (B) Regions with a relative decrease over 10%; Only the top 15 enriched terms are shown. The x-axis represents fold enrichment, while the colors indicate significance (negative of $\log_{10}$FDR, ligher = more significant). The y-axis lists GO terms in GO Biological Process. The names of terms with a significant false discovery rate (FDR) (FDR < 0.05) are shown in blue.



Supplementary Fig. S9. Gene set enrichment analysis of regions with more than 10% relative change in CG context (GeDiPNet). (A) Regions with a relative increase over 10%; (B) Regions with a relative decrease over 10%; Only the top 15 enriched terms are shown. The names of terms with a significant false discovery rate (FDR) (FDR < 0.05) are shown in blue. The x-axis represents fold enrichment, while the colors indicate significance (negative of $\log_{10}$FDR, ligher = more significant). The y-axis lists GO terms in GO Biological Process. The names of terms with a significant false discovery rate (FDR) (FDR < 0.05) are shown in blue.



Supplementary Fig. S10. Distributions of sequencing depths and platforms across 10,239 whole genomes. (A) Stacked bar plot showing the distribution of average sequencing depth per sample (x-axis, as ×) across 10,239 Korean genomes (y-axis). Samples are grouped by rounded average depth (0–100×) on the x-axis, with colors indicating WGS platform: HiSeq X (blue), MGI T7 (orange), and NovaSeq (green). The vertical red dashed line and annotation indicates the mean depth (30.0×). The y-axis denotes the number of samples. **(B)** Pie chart showing the proportion of samples sequenced on each platform, with absolute counts and percentages indicated on the wedges.

Supplementary Table 1. Summary statistics of saturation analysis for full variant discovery stratified by allele frequency. The table shows the number of genomes required to reach saturation in variant detection across different frequency categories. For each category, the minimum (Min), median (Median), maximum (Max), and mean (Mean) number of genomes at saturation are reported, along with the 95% confidence interval (95% CI [lower, upper]) of the mean. Red highlights statistics reaching the maximum number (9,000 samples) without full variant discovery, which indicates failure to saturate.

Supplementary Table 2. Summary statistics of imputation performance stratified by allele frequency. The table shows percentage improvement in imputation quality (R^2) was calculated for each allele frequency (AF) bin (%) by comparing imputation panels of Korea10K to Korea1K (10K/1K) and Korea10K to Korea4K (10K/4K).

Supplementary Table 3. Overview of Health Check-up (HC) measured in Korea10K. The table summarizes the variables collected from participants across multiple domains including anthropometry, physiological and functional measures, hematology, and urinalysis. Normal reference ranges reflect commonly accepted clinical standards and are stratified by sex when appropriate.

Supplementary Table 4. Overview of Lifestyle Questionnaire in Korea10K. The table lists self-reported lifestyle variables, captured through structured questionnaires, that have been answered by survey respondents in Korea10K cohort.

Supplementary Table S5. Top five hotspot CGVs in Korea10K. The table lists the genomic regions exhibiting the most pronounced changes in CG dinucleotide context relative to the GRCh38 reference genome. Columns indicate the chromosome, start and end coordinates of the 100 kb window, percentage change in CG content (% CG Difference), absolute number of CG changes (# CG Difference), total number of CG dinucleotides in the reference (# CG from GRCh38), and gene annotation or relevant genomic feature (Note). Positive values indicate an gain in CG sites, whereas negative values indicate a loss.

Supplementary Table S6. Summary statistics of Whole-Genome Sequencing (WGS) data in Korea10K. The table summarizes key sequencing metrics for Korea10K genomes. Columns include the sample identifier (SampleID), total sequencing output (Production_Amount, Gbp), calculated mean coverage depth (Average_Depth, ×), integer rounded coverage depth (Rounded_Depth, ×), and sequencing platform used (WGS Platform).

References

Abella, V., Scotece, M., Conde, J., Pino, J., Gonzalez-Gay, M. A., Gomez-Reino, J. J., Mera, A., Lago, F., Gomez, R., & Gualillo, O. (2017). Leptin in the interplay of inflammation, metabolism and immune system disorders. *Nat Rev Rheumatol*, *13*(2), 100-109. https://doi.org/10.1038/nrrheum.2016.209

Ahlqvist, E., Storm, P., Karajamaki, A., Martinell, M., Dorkhan, M., Carlsson, A., Vikman, P., Prasad, R. B., Aly, D. M., Almgren, P., Wessman, Y., Shaat, N., Spegel, P., Mulder, H., Lindholm, E., Melander, O., Hansson, O., Malmqvist, U., Lernmark, A., . . . Groop, L. (2018). Novel subgroups of adult-onset diabetes and their association with outcomes: a data-driven cluster analysis of six variables. *Lancet Diabetes Endocrinol*, *6*(5), 361-369. https://doi.org/10.1016/S2213-8587(18)30051-2

Alexander, D. H., Novembre, J., & Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. *Genome Res*, *19*(9), 1655-1664. https://doi.org/10.1101/gr.094052.109

An, K., Kwon, Y., Bhak, J., Ryu, H., Jeon, S., Nam, D., & Bhak, J. (2025). Dynamic and reversible transcriptomic age shifts induced by COVID-19 in Korean whole blood. *Aging (Albany NY)*, *17*(6), 1484-1510. https://doi.org/10.18632/aging.206270

Banovich, N. E., Lan, X., McVicker, G., van de Geijn, B., Degner, J. F., Blischak, J. D., Roux, J., Pritchard, J. K., & Gilad, Y. (2014). Methylation QTLs are associated with coordinated changes in transcription factor binding, histone modifications, and gene expression levels. *PLoS Genet*, *10*(9), e1004663. https://doi.org/10.1371/journal.pgen.1004663

Bhak, Y., Jeon, Y., Jeon, S., Yoon, C., Kim, M., Blazyte, A., Kim, Y., Kang, Y., Kim, C., Lee, S. Y., Bae, J. W., Kim, W., Kim, Y. J., Shim, J., Kim, N., Chun, S., Kim, B. C., Kim, B. C., Lee, S., . . . Shin, E. S. (2021). Polygenic risk score validation using Korean genomes of 265 early-onset acute myocardial infarction patients and 636 healthy controls. *PLoS One*, *16*(2), e0246538. https://doi.org/10.1371/journal.pone.0246538

Bhak, Y., Jeong, H. O., Cho, Y. S., Jeon, S., Cho, J., Gim, J. A., Jeon, Y., Blazyte, A., Park, S. G., Kim, H. M., Shin, E. S., Paik, J. W., Lee, H. W., Kang, W., Kim, A., Kim, Y., Kim, B. C., Ham, B. J., Bhak, J., & Lee, S. (2019). Depression and suicide risk prediction models using blood-derived multi-omics data. *Transl Psychiatry*, *9*(1), 262. https://doi.org/10.1038/s41398-019-0595-2

Bonatto, S. L., & Salzano, F. M. (1997). Diversity and age of the four major mtDNA haplogroups, and their implications for the peopling of the New World. *Am J Hum Genet*, *61*(6), 1413-1423. https://doi.org/10.1086/301629

Bonfield, J. K., Marshall, J., Danecek, P., Li, H., Ohan, V., Whitwham, A., Keane, T., & Davies, R. M. (2021). HTSlib: C library for reading/writing high-throughput sequencing data. *Gigascience*, *10*(2). https://doi.org/10.1093/gigascience/giab007

Chang, C. C., Chow, C. C., Tellier, L. C., Vattikuti, S., Purcell, S. M., & Lee, J. J. (2015). Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience*, *4*, 7. https://doi.org/10.1186/s13742-015-0047-8

Chen, S., Zhou, Y., Chen, Y., & Gu, J. (2018). fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*, *34*(17), i884-i890. https://doi.org/10.1093/bioinformatics/bty560

Cho, Y. S., Kim, H., Kim, H. M., Jho, S., Jun, J., Lee, Y. J., Chae, K. S., Kim, C. G., Kim, S., Eriksson, A., Edwards, J. S., Lee, S., Kim, B. C., Manica, A., Oh, T. K., Church, G. M., & Bhak, J. (2016). An ethnically relevant consensus Korean reference genome is a step towards personal reference genomes. *Nat Commun*, *7*, 13637. https://doi.org/10.1038/ncomms13637

Consortium, U. K. B. W.-G. S. (2025). Whole-genome sequencing of 490,640 UK Biobank participants. *Nature*, *645*(8081), 692-701. https://doi.org/10.1038/s41586-025-09272-9

Crocchiolo, R., Gallina, A. M., Pani, A., Campisi, D., Cento, V., Sacchi, N., Miotti, V., Gagliardi, O. M., D'Amico, F., Vismara, C., Cornacchini, G., Lando, G., Cuppari, I., Scaglione, F., & Rossini, S. (2022). Polymorphism of the HLA system and weak antibody response to BNT162b2 mRNA vaccine. *HLA*, *99*(3), 183-191. https://doi.org/10.1111/tan.14546

Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., Whitwham, A., Keane, T., McCarthy, S. A., Davies, R. M., & Li, H. (2021). Twelve years of SAMtools and BCFtools. *Gigascience*, *10*(2). https://doi.org/10.1093/gigascience/giab008

Das, S., Forer, L., Schonherr, S., Sidore, C., Locke, A. E., Kwong, A., Vrieze, S. I., Chew, E. Y., Levy, S., McGue, M., Schlessinger, D., Stambolian, D., Loh, P. R., Iacono, W. G., Swaroop, A., Scott, L. J., Cucca, F., Kronenberg, F., Boehnke, M., . . . Fuchsberger, C. (2016). Next-generation genotype imputation service and methods. *Nat Genet*, *48*(10), 1284-1287. https://doi.org/10.1038/ng.3656

Dierckxsens, N., Mardulyn, P., & Smits, G. (2017). NOVOPlasty: de novo assembly of organelle genomes from whole genome data. *Nucleic Acids Res*, *45*(4), e18. https://doi.org/10.1093/nar/gkw955

Durbin, R., De Sanctis, B., & Blumer, M. (2023). Rotate: A command-line program to rotate circular DNA sequences to start at a given position or string. *Wellcome Open Res*, *8*, 401. https://doi.org/10.12688/wellcomeopenres.19568.1

Feng, Y. A., Chen, C. Y., Chen, T. T., Kuo, P. H., Hsu, Y. H., Yang, H. I., Chen, W. J., Su, M. W., Chu, H. W., Shen, C. Y., Ge, T., Huang, H., & Lin, Y. F. (2022). Taiwan Biobank: A rich biomedical research database of the Taiwanese population. *Cell Genom*, *2*(11), 100197. https://doi.org/10.1016/j.xgen.2022.100197

Gate, R. E., Cheng, C. S., Aiden, A. P., Siba, A., Tabaka, M., Lituiev, D., Machol, I., Gordon, M. G., Subramaniam, M., Shamim, M., Hougen, K. L., Wortman, I., Huang, S. C., Durand, N. C., Feng, T., De Jager, P. L., Chang, H. Y., Aiden, E. L., Benoist, C., . . . Regev, A. (2018). Genetic determinants of co-accessible chromatin regions in activated T cells across humans. *Nat Genet*, *50*(8), 1140-1150. https://doi.org/10.1038/s41588-018-0156-2

Ge, S. X., Jung, D., & Yao, R. (2020). ShinyGO: a graphical gene-set enrichment tool for animals and plants. *Bioinformatics*, *36*(8), 2628-2629. https://doi.org/10.1093/bioinformatics/btz931

Genomes Project, C., Auton, A., Brooks, L. D., Durbin, R. M., Garrison, E. P., Kang, H. M., Korbel, J. O., Marchini, J. L., McCarthy, S., McVean, G. A., & Abecasis, G. R. (2015). A global reference for human genetic variation. *Nature*, *526*(7571), 68-74. https://doi.org/10.1038/nature15393

Gusev, A., Ko, A., Shi, H., Bhatia, G., Chung, W., Penninx, B. W., Jansen, R., de Geus, E. J., Boomsma, D. I., Wright, F. A., Sullivan, P. F., Nikkola, E., Alvarez, M., Civelek, M., Lusis, A. J., Lehtimaki, T., Raitoharju, E., Kahonen, M., Seppala, I., . . . Pasaniuc, B. (2016). Integrative approaches for large-scale transcriptome-wide association studies. *Nat Genet*, *48*(3), 245-252. https://doi.org/10.1038/ng.3506

Hellwege, J. N., Keaton, J. M., Giri, A., Gao, X., Velez Edwards, D. R., & Edwards, T. L. (2017). Population Stratification in Genetic Association Studies. *Curr Protoc Hum Genet*, *95*, 1 22 21-21 22 23. https://doi.org/10.1002/cphg.48

Hofmeister, R. J., Ribeiro, D. M., Rubinacci, S., & Delaneau, O. (2023). Accurate rare variant phasing of whole-genome and whole-exome sequencing data in the UK Biobank. *Nat Genet*, *55*(7), 1243-1249. https://doi.org/10.1038/s41588-023-01415-w

Jeon, S., Bhak, Y., Choi, Y., Jeon, Y., Kim, S., Jang, J., Jang, J., Blazyte, A., Kim, C., Kim, Y., Shim, J., Kim, N., Kim, Y. J., Park, S. G., Kim, J., Cho, Y. S., Park, Y., Kim, H. M., Kim, B. C., . . . Bhak, J. (2020). Korean Genome Project: 1094 Korean personal genomes with clinical information. *Sci Adv*, *6*(22), eaaz7835. https://doi.org/10.1126/sciadv.aaz7835

Jeon, S., Choi, H., Jeon, Y., Choi, W. H., Choi, H., An, K., Ryu, H., Bhak, J., Lee, H., Kwon, Y., Ha, S., Kim, Y. J., Blazyte, A., Kim, C., Kim, Y., Kang, Y., Woo, Y. J., Lee, C., Seo, J., . . . Bhak, J. (2024). Korea4K: whole genome sequences of 4,157 Koreans with 107 phenotypes derived from extensive health check-ups. *Gigascience*, *13*. https://doi.org/10.1093/gigascience/giae014

Jeon, Y., Jeon, S., An, K., Kim, Y. J., Kim, B. C., Ryu, H., Choi, W. H., Choi, H., Kim, W., Lee, S. Y., Bae, J. W., Hwang, J. Y., Kang, M. G., An, S., Kim, Y., Kang, Y., Kim, B. C., Bhak, J., & Shin, E. S. (2023). Identification and validation of six acute myocardial infarction-associated variants, including a novel prognostic marker for cardiac mortality. *Front Cardiovasc Med*, *10*, 1226971. https://doi.org/10.3389/fcvm.2023.1226971

Jeon, Y., Jeon, S., Choi, W. H., An, K., Choi, H., Kim, B. C., Kim, W., Lee, S. Y., Bae, J. W., Hwang, J. Y., Kang, M. G., Kim, Y., Kang, Y., Kim, Y. J., Kim, B. C., Bhak, J., & Shin, E. S. (2023). Genome-wide analyses of early-onset acute myocardial infarction identify 29 novel loci by whole genome sequencing. *Hum Genet*, *142*(2), 231-243. https://doi.org/10.1007/s00439-022-02495-0

Jeon, Y., Kwon, Y., Kim, Y. J., Jeon, S., Ryu, H., An, K., Kim, B. C., Kim, W., Lee, S. Y., Bae, J. W., Hwang, J. Y., Kang, M. G., Kang, Y., Bhak, J., & Shin, E. S. (2025). Epigenetic Modulation of Vascular Smooth Muscle Cell Phenotype Switching in Early-Onset Acute Myocardial Infarction. *Arterioscler Thromb Vasc Biol*, *45*(6), e217-e230. https://doi.org/10.1161/ATVBAHA.125.322503

Kang, E. H., Kim, J. Y., Takeuchi, F., Kim, J. W., Shin, K., Lee, E. Y., Lee, Y. J., Lee, E. B., Park, M. H., & Song, Y. W. (2011). Associations between the HLA-A polymorphism and the clinical manifestations of Behcet's disease. *Arthritis Res Ther*, *13*(2), R49. https://doi.org/10.1186/ar3292

Karczewski, K. J., Francioli, L. C., Tiao, G., Cummings, B. B., Alfoldi, J., Wang, Q., Collins, R. L., Laricchia, K. M., Ganna, A., Birnbaum, D. P., Gauthier, L. D., Brand, H., Solomonson, M., Watts, N. A., Rhodes, D., Singer-Berk, M., England, E. M., Seaby, E. G., Kosmicki, J. A., . . . MacArthur, D. G. (2020). The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature*, *581*(7809), 434-443. https://doi.org/10.1038/s41586-020-2308-7

Katoh, K., & Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol*, *30*(4), 772-780. https://doi.org/10.1093/molbev/mst010

Kim, J., Jeon, S., Choi, J. P., Blazyte, A., Jeon, Y., Kim, J. I., Ohashi, J., Tokunaga, K., Sugano, S., Fucharoen, S., Al-Mulla, F., & Bhak, J. (2020). The Origin and Composition of Korean Ethnicity Analyzed by Ancient and Present-Day Genome Sequences. *Genome Biol Evol*, *12*(5), 553-565. https://doi.org/10.1093/gbe/evaa062

Kim, J., Park, J. L., Yang, J. O., Kim, S., Joe, S., Park, G., Hwang, T., Cho, M. J., Lee, S., Lee, J. E., Park, J. H., Yeo, M. K., & Kim, S. Y. (2025). Highly accurate Korean draft genomes reveal structural variation highlighting human telomere evolution. *Nucleic Acids Res*, *53*(1). https://doi.org/10.1093/nar/gkae1294

Kim, S. H., Kim, K. C., Shin, D. J., Jin, H. J., Kwak, K. D., Han, M. S., Song, J. M., Kim, W., & Kim, W. (2011). High frequencies of Y-chromosome haplogroup O2b-SRY465 lineages in Korea: a genetic perspective on the peopling of Korea. *Investig Genet*, *2*(1), 10. https://doi.org/10.1186/2041-2223-2-10

Kumar, S., Stecher, G., Suleski, M., Sanderford, M., Sharma, S., & Tamura, K. (2024). MEGA12: Molecular Evolutionary Genetic Analysis Version 12 for Adaptive and Green Computing. *Mol Biol Evol*, *41*(12). https://doi.org/10.1093/molbev/msae263

Kundu, I., Sharma, M., Barai, R. S., Pokar, K., & Idicula-Thomas, S. (2023). GeDiPNet: Online resource of curated gene-disease associations for polypharmacological targets discovery. *Genes Dis*, *10*(3), 647-649. https://doi.org/10.1016/j.gendis.2022.05.034

Kurki, M. I., Karjalainen, J., Palta, P., Sipila, T. P., Kristiansson, K., Donner, K. M., Reeve, M. P., Laivuori, H., Aavikko, M., Kaunisto, M. A., Loukola, A., Lahtela, E., Mattsson, H., Laiho, P., Della Briotta Parolo, P., Lehisto, A. A., Kanai, M., Mars, N., Ramo, J., . . . Palotie, A. (2023). FinnGen provides genetic insights from a well-phenotyped isolated population. *Nature*, *613*(7944), 508-518. https://doi.org/10.1038/s41586-022-05473-8

Kwon, Y., Blazyte, A., Jeon, Y., Kim, Y. J., An, K., Jeon, S., Ryu, H., Shin, D. H., Ahn, J., Um, H., Kang, Y., Bak, H., Kim, B. C., Lee, S., Jung, H. T., Shin, E. S., & Bhak, J. (2025). Identification of 17 novel epigenetic biomarkers associated with anxiety disorders using differential methylation analysis followed by machine learning-based validation. *Clin Epigenetics*, *17*(1), 24. https://doi.org/10.1186/s13148-025-01819-x

Lee, J., Lee, J., Jeon, S., Lee, J., Jang, I., Yang, J. O., Park, S., Lee, B., Choi, J., Choi, B. O., Gee, H. Y., Oh, J., Jang, I. J., Lee, S., Baek, D., Koh, Y., Yoon, S. S., Kim, Y. J., Chae, J. H., . . . Choi, M. (2022). A database of 5305 healthy Korean individuals reveals genetic and clinical

implications for an East Asian population. *Exp Mol Med*, *54*(11), 1862-1871. https://doi.org/10.1038/s12276-022-00871-4

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, *25*(14), 1754-1760. https://doi.org/10.1093/bioinformatics/btp324

Li, X., Zhao, Z., & Sun, S. S. (2017). Association of human leukocyte antigen variants and allopurinol-induced Stevens-Johnson syndrome and toxic epidermal necrolysis: A meta-analysis. *Am J Health Syst Pharm*, *74*(9), e183-e192. https://doi.org/10.2146/ajhp160243

Liao, W. W., Asri, M., Ebler, J., Doerr, D., Haukness, M., Hickey, G., Lu, S., Lucas, J. K., Monlong, J., Abel, H. J., Buonaiuto, S., Chang, X. H., Cheng, H., Chu, J., Colonna, V., Eizenga, J. M., Feng, X., Fischer, C., Fulton, R. S., . . . Paten, B. (2023). A draft human pangenome reference. *Nature*, *617*(7960), 312-324. https://doi.org/10.1038/s41586-023-05896-x

McLaren, W., Gil, L., Hunt, S. E., Riat, H. S., Ritchie, G. R., Thormann, A., Flicek, P., & Cunningham, F. (2016). The Ensembl Variant Effect Predictor. *Genome Biol*, *17*(1), 122. https://doi.org/10.1186/s13059-016-0974-4

Meng, G., Li, Y., Yang, C., & Liu, S. (2019). MitoZ: a toolkit for animal mitochondrial genome assembly, annotation and visualization. *Nucleic Acids Res*, *47*(11), e63. https://doi.org/10.1093/nar/gkz173

Nagai, A., Hirata, M., Kamatani, Y., Muto, K., Matsuda, K., Kiyohara, Y., Ninomiya, T., Tamakoshi, A., Yamagata, Z., Mushiroda, T., Murakami, Y., Yuji, K., Furukawa, Y., Zembutsu, H., Tanaka, T., Ohnishi, Y., Nakamura, Y., BioBank Japan Cooperative Hospital, G., & Kubo, M. (2017). Overview of the BioBank Japan Project: Study design and profile. *J Epidemiol*, *27*(3S), S2-S8. https://doi.org/10.1016/j.je.2016.12.005

Nguyen, L. T., Schmidt, H. A., von Haeseler, A., & Minh, B. Q. (2015). IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol Biol Evol*, *32*(1), 268-274. https://doi.org/10.1093/molbev/msu300

Picard Toolkit. (2019). Broad Institute. https://broadinstitute.github.io/picard/

Poplin, R., Ruano-Rubio, V., DePristo, M. A., Fennell, T. J., Carneiro, M. O., Van der Auwera, G. A., Kling, D. E., Gauthier, L. D., Levy-Moonshine, A., Roazen, D., Shakir, K., Thibault, J., Chandran, S., Whelan, C., Lek, M., Gabriel, S., Daly, M. J., Neale, B., MacArthur, D. G., & Banks, E. (2018). Scaling accurate genetic variant discovery to tens of thousands of samples. *bioRxiv*, 201178. https://doi.org/10.1101/201178

Poznik, G. D. (2016). Identifying Y-chromosome haplogroups in arbitrarily large samples of sequenced or genotyped men. *bioRxiv*, 088716. https://doi.org/10.1101/088716

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., Maller, J., Sklar, P., de Bakker, P. I., Daly, M. J., & Sham, P. C. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet*, *81*(3), 559-575. https://doi.org/10.1086/519795

Ryu, H., An, K., Kwon, Y., Jeon, Y., Jeon, S., Choi, H., Kim, Y. J., Kim, S., Sul, O. J., Lee, S., Chun, A. Y., Shin, E. S., Ra, S. W., & Bhak, J. (2025). Identification of reactive CpGs and RNA expression in early COVID-19 through cis-eQTM analysis reflecting disease severity and recovery. *Commun Biol*, *8*(1), 1174. https://doi.org/10.1038/s42003-025-08609-4

Schonherr, S., Weissensteiner, H., Kronenberg, F., & Forer, L. (2023). Haplogrep 3 - an interactive haplogroup classification and analysis platform. *Nucleic Acids Res*, *51*(W1), W263-W268. https://doi.org/10.1093/nar/gkad284

Taliun, D., Harris, D. N., Kessler, M. D., Carlson, J., Szpiech, Z. A., Torres, R., Taliun, S. A. G., Corvelo, A., Gogarten, S. M., Kang, H. M., Pitsillides, A. N., LeFaive, J., Lee, S. B., Tian, X., Browning, B. L., Das, S., Emde, A. K., Clarke, W. E., Loesch, D. P., . . . Abecasis, G. R. (2021). Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *Nature*, *590*(7845), 290-299. https://doi.org/10.1038/s41586-021-03205-y

Wu, D., Dou, J., Chai, X., Bellis, C., Wilm, A., Shih, C. C., Soon, W. W. J., Bertin, N., Lin, C. B., Khor, C. C., DeGiorgio, M., Cheng, S., Bao, L., Karnani, N., Hwang, W. Y. K., Davila, S., Tan, P., Shabbir, A., Moh, A., . . . Wang, C. (2019). Large-Scale Whole-Genome Sequencing of Three Diverse Asian Populations in Singapore. *Cell*, *179*(3), 736-749 e715. https://doi.org/10.1016/j.cell.2019.09.019

Xu, L., He, W., Tai, S., Huang, X., Qin, M., Liao, X., Jing, Y., Yang, J., Fang, X., Shi, J., & Jin, N. (2025). VCF2Dis: an ultra-fast and efficient tool to calculate pairwise genetic distance and construct population phylogeny from VCF files. *Gigascience*, *14*. https://doi.org/10.1093/gigascience/giaf032