

ARTICLE OPEN



A database of 5305 healthy Korean individuals reveals genetic and clinical implications for an East Asian population

Jeongeun Lee^{1,17}, Jean Lee^{2,17}, Sungwon Jeon³, Jeongha Lee², Insu Jang⁴, Jin Ok Yang^{4,5}, Soojin Park⁶, Byungwook Lee⁴, Jinwook Choi^{1,7}, Byung-Ok Choi⁸, Heon Yung Gee⁹, Jaeseong Oh¹⁰, In-Jin Jang¹⁰, Sanghyuk Lee¹¹, Daehyun Baek¹², Youngil Koh¹³, Sung-Soo Yoon¹³, Young-Joon Kim¹⁴, Jong-Hee Chae^{6,15}, Woong-Yang Park¹⁶, Jong Hwa Bhak³✉ and Murim Choi^{10,2}✉

© The Author(s) 2022

Despite substantial advances in disease genetics, studies to date have largely focused on individuals of European descent. This limits further discoveries of novel functional genetic variants in other ethnic groups. To alleviate the paucity of East Asian population genome resources, we established the Korean Variant Archive 2 (KOVA 2), which is composed of 1896 whole-genome sequences and 3409 whole-exome sequences from healthy individuals of Korean ethnicity. This is the largest genome database from the ethnic Korean population to date, surpassing the 1909 Korean individuals deposited in gnomAD. The variants in KOVA 2 displayed all the known genetic features of those from previous genome databases, and we compiled data from Korean-specific runs of homozygosity, positively selected intervals, and structural variants. In doing so, we found loci, such as the loci of *ADH1A/1B* and *UHRF1BP1*, that are strongly selected in the Korean population relative to other East Asian populations. Our analysis of allele ages revealed a correlation between variant functionality and evolutionary age. The data can be browsed and downloaded from a public website (<https://www.kobic.re.kr/kova/>). We anticipate that KOVA 2 will serve as a valuable resource for genetic studies involving East Asian populations.

Experimental & Molecular Medicine (2022) 54:1862–1871; <https://doi.org/10.1038/s12276-022-00871-4>

INTRODUCTION

Korean individuals are known to have migrated to the Korean peninsula at least 40,000 years ago; this migration probably occurred through two routes, from northeast and southeast Asia^{1,2}. Complex but constant admixture with the neighboring Chinese and Japanese populations ensued throughout history³, yet many studies have suggested that ethnic Korean individuals are genetically distinct from Chinese and Japanese individuals. At ~83 million people, the ethnic Korean population is the 15th largest ethnic group in the world. Especially in South Korea, modern nationwide healthcare systems can provide an opportunity to study the genetics of various diseases in this population if the appropriate genetic infrastructure is provided.

As the field of human genetics advances, more attention is being paid to non-European populations as a new venue for obtaining novel insights into the genetics and physiology of

human development, physiology, and disease. Although East Asian individuals make up nearly a fifth of the world population, they comprise only 8.2% of participants in genome-wide association studies (GWAS)⁴. Likewise, control databases—compilations of apparently healthy genomes—for East Asian populations are scarce compared to those for Europeans. For example, among the 141,156 participants in gnomAD version 2, only 9977 East Asian individuals were listed, of which only 1909 were Korean individuals. Notably, small databases of genetic information from Korean individuals (with sizes of ~1000 individuals) have been released^{5–7}, including one established by our group (Korean Variant Archive [KOVA]⁵). However, as the cohort sizes of human genetics studies increase, it becomes necessary to construct larger Korean control databases.

Here, we introduce KOVA 2, a Korean control database that includes 5305 individuals. Using the variant set from KOVA 2, we

¹Interdisciplinary Program in Bioengineering, Graduate School, Seoul National University, Seoul 03080, Republic of Korea. ²Department of Biomedical Sciences, Seoul National University College of Medicine, Seoul 03080, Republic of Korea. ³Department of Biomedical Engineering, College of Information and Biotechnology, Ulsan National Institute of Science and Technology (UNIST), Ulsan 44919, Republic of Korea. ⁴Korea Bioinformatics Center (KOBIC), Korea Research Institute of Bioscience and Biotechnology (KRIBB), Daejeon 34141, Republic of Korea. ⁵Department of Bio and Brain Engineering, Korea Advanced Institute of Science and Technology (KAIST), Daejeon 34141, Republic of Korea. ⁶Department of Pediatrics, Seoul National University College of Medicine, Seoul 03080, Republic of Korea. ⁷Department of Biomedical Engineering, Seoul National University College of Medicine, Seoul 03080, Republic of Korea. ⁸Department of Neurology, Samsung Medical Center, Sungkyunkwan University School of Medicine, Seoul 06351, Republic of Korea. ⁹Department of Pharmacology, Brain Korea 21 PLUS Project for Medical Sciences, Yonsei University College of Medicine, Seoul 03722, Republic of Korea. ¹⁰Department of Clinical Pharmacology and Therapeutics, Seoul National University College of Medicine and Hospital, Seoul 03080, Republic of Korea. ¹¹Department of Bio-Information Science, Ewha Womans University, Seoul 03760, Republic of Korea. ¹²School of Biological Sciences, Seoul National University, Seoul 08826, Republic of Korea. ¹³Department of Internal Medicine, Seoul National University Hospital, Seoul 03080, Republic of Korea. ¹⁴Department of Biochemistry, College of Life Science and Biotechnology, Yonsei University, Seoul 03722, Republic of Korea. ¹⁵Department of Genomic Medicine, Seoul National University Hospital, Seoul 03080, Republic of Korea. ¹⁶Samsung Genome Institute, Samsung Medical Center, Seoul 06351, Republic of Korea. ¹⁷These authors contributed equally: Jeongeun Lee, Jean Lee. ✉email: jongbhak@gmail.com; murimchoi@snu.ac.kr

Received: 30 May 2022 Revised: 21 July 2022 Accepted: 8 August 2022

Published online: 2 November 2022

determined Korean-specific runs of homozygosity (ROH) regions, intervals of positive selection, structural variants, and allelic ages. As a public resource, KOVA 2 will serve as an essential tool for genetic studies of Korean and East Asian populations.

MATERIALS AND METHODS

Cohorts and sample preparation

We collected the whole-exome sequencing (WES) and whole-genome sequencing (WGS) data for Korean individuals from independent research groups in Korea (Supplementary Table 1). All sequencing data were obtained from normal tissues or blood samples following standard protocols⁵. This project was performed with the approval of the Institutional Review Board of each group (Seoul National University and others), in which all donors provided written informed consent if available. All experiments were performed on deidentified samples and in accordance with relevant guidelines and regulations.

Variant calling

We used BWA mem v0.7.17⁸ with default options to map raw reads to the GRCh38+decoy reference sequence. After marking duplicates and sorting by coordinating with MarkDuplicatesSpark, the mapping quality was recalibrated by BQSRPipelineSpark, implemented in GATK version 4.1.3.0⁹. Qualimap v2.2.1¹⁰ was used to generate quality control metrics for the mapped sequence data. Single nucleotide variants (SNVs) and small insertions and deletions (indels) were then called for each sample using GATK HaplotypeCaller with the option “-ERC GVCF”. To jointly genotype samples, we created a genomicsDB using GenomicsDBImport in GATK and followed the GATK best practice guideline⁹. Briefly, SNVs and indels were recalibrated by GATK’s VQS model to select 99.7 and 99.0% of true sites, respectively, from the training set. The detailed workflow is described in Supplementary Fig. 1. Further analyses adopted a modified version of gnomAD QC steps¹¹ and were mostly performed with Hail¹², which is an open-source Python library for genome data analysis. After merging the WES and WGS data using Hail, we excluded multiallelic variants and variants that had a genotype quality (GQ) <20, read depth (DP) <10, allelic balance (AB) <0.2, or were in low complexity regions¹³ (Supplementary Fig. 2).

Sex inference

We inferred the sex of each sample by calculating sex chromosome ploidy, which is defined as the coverage of sex chromosomes divided by the coverage of chromosome 21. To assign X and Y ploidy cutoffs, we calculated F-stat scores based on the linkage disequilibrium (LD)-pruned biallelic SNVs (MAF >0.05, call-rate >0.99, inbreeding coefficient score ≥ -0.03 and R^2 for LD pruning <0.1) using the “annotate_sex” function of the gnomAD Hail library with the parameters “male_threshold = 0.8, female_threshold = 0.5”. An XX karyotype was defined if X chromosome ploidy ranged between [1.7, 3.4] and [1.55, 2.45] for WES and WGS, respectively. An XY karyotype was assigned when Y chromosome ploidy ranged between [0.2, 2.3] and [0.45, 1.11] and X chromosome ploidy was below 1.65 and 1.50 for WES and WGS, respectively (Supplementary Fig. 3). In subsequent analyses, only samples assigned to the XX or XY karyotype were used. In total, 92 samples were excluded because they were determined to be of ambiguous sex.

Relatedness inference

To remove close relatives, we estimated kinship and the probability of identity-by-descent (IBD) being zero for every pair of samples based on the LD-pruned variants with a MAF ≥ 0.001 , call-rate >0.99, HWE $P > 1.0 \times 10^{-8}$, inbreeding coefficient score > -0.025 , and R^2 for LD pruning <0.1. After calculating kinship using the “pc_relate” feature¹⁴ in Hail, we selected the maximal independent set of samples with kinship <0.1 using the “maximal_independent_set”¹⁵ from Hail. For related sample pairs, we chose the one with a higher coverage depth.

Population structure analysis

All biallelic autosomal SNVs from our dataset and the 1000 Genomes Project Phase 3 (KG)¹⁶ were merged and filtered; variants were retained if they had a MAF >0.001, call-rate >0.99, HWE $P > 1.0 \times 10^{-8}$, and inbreeding coefficient score > -0.025 . We then pruned the variants to those with an LD $R^2 < 0.1$. To perform a principal component analysis (PCA) on the

Hardy–Weinberg-normalized variants, we used the “hwe_normalized_pca” function of Hail with $k = 30$. Each sample was assigned to an ancestry, determined as the ancestry with maximum probability emitted from a random forest model trained on the KG PCA result. We removed non-Korean or Korean-outlier samples iteratively until the Chinese, Japanese, Korean, and Vietnamese populations all became distinguishable based on PCs 1 and 2.

Sample QC

The overall process is summarized in Supplementary Fig. 2. First, we excluded samples with ambiguous clinical status or having a mean coverage depth of <40X and <10X for WES and WGS, respectively. Samples with ambiguous or abnormal sex were then excluded, as were duplicated samples and closely related samples. We further removed samples with ambiguous ethnicity, followed by samples with a Het/Hom ratio >1.8 (Supplementary Figs. 2, 4, 5). Finally, after combining the WES and WGS data, we reperformed the relatedness inference procedure to remove WES samples that overlapped or were related to WGS samples.

Variant quality control

The overall process is summarized in Supplementary Fig. 2. Variants were considered to have violated Hardy–Weinberg equilibrium (HWE) on allelic frequency ($P < 1.0 \times 10^{-6}$) when the allele frequency was >0.01 or the inbreeding coefficient score was < -0.03 , and those variants were removed. Functional annotation was performed by the Variants Effect Predictor (VEP) version 101¹⁷. For each variant, we selected the most severe functional consequences using the gnomAD package of Hail. Ti/Tv and Het/Hom scores were computed using the “compute_sample_qc_metric” function implemented in Hail (Supplementary Fig. 4).

Phasing

After carrying out sample-level and variant-level quality control, WGS data were phased with SHAPEIT4 version 4.2.2¹⁸. After input to SHAPEIT4, we converted the VCF file to a PLINK file format with the option “-geno 0.1 -maf 0.001” to keep SNVs with missingness <10% and MAF >0.001. We used the genetic maps for reference version hg38 that are provided by SHAPEIT4¹⁹. We also phased our data with Beagle 5.2 (beagle.21-Apr21.304.jar)²⁰, for which we used the hg38 genetic map available at the Beagle website²¹ and the reference panel created by the 1000 Genome Project.

Runs of homozygosity (ROH)

PLINK v1.90b6.12^{22,23} was used to call ROH regions from SHAPEIT-phased data with the options “-maf 0.05 -hwe 0.00005 -homozyg -homozyg-snp 50 -homozyg-kb 500 -homozyg-density 10 -homozyg-gap 10 -homozyg-window-snp 50 -homozyg-window-missing 5 -homozyg-window-het 1 -homozyg-window-threshold 0.05”. To ensure the fair comparison of ROH intervals from KOVA 2 with other populations in the KG, the regions were called from randomly selected sets of 105 samples from KOVA 2. After merging the ROH results from KOVA 2 and KG data, we calculated F_{ROH} scores, representing inbreeding levels, using the “Froh_inbreeding” function of detectRuns package version 0.9.6²⁴.

Regions of positive selection

Selected variants in positive selection sweeps were captured from phased KOVA 2 and KG data using iSAFE v1.0.7²⁵ software. iSAFE uses a statistic generated from population genetics signals to precisely identify the preferred variant in a large region (~5 Mbp). A variant is favored if its iSAFE score is larger than 0.1 ($P < 1.0 \times 10^{-4}$), and a high iSAFE score signifies that the variant is strongly positively selected. We used iSAFE with default options (-MaxRegionSize 6000000 -window 300 -MaxRank 15 -MaxFreq 0.95 -IgnoreGaps) plus the performance-improving parameter “-vcf-cont” with random outgroup (nontarget) samples comprising 10% of the data.

Effective population size estimation

To estimate the historical effective population size, we used IBDNe software²⁶ according to the recommended protocol. Briefly, after detecting IBD segments with hap-IBD.jar²⁷, we refined them through the removal of any breaks and short gaps from the segments using merge-ibd-segments.17Jan20.102.jar²⁸. Finally, we used ibdne.23Apr20.ae9.jar²⁶ with default options to estimate the effective population size from the refined IBD segments.

Allele ages

Genealogical Estimation of Variant Age (GEVA) version v1beta²⁹ with parameters “--Ne 10000 --mut 1e-8 --maxConcordant 500 --maxDiscordant 500” was used to estimate the ages of variants from autosomal haplotype data phased by SHAPEIT4. Allele ages were computed by the joint clock model, which combines the mutation and recombination clock models. To compare allele ages as estimated by our data with those estimated from the 1000 Genomes data, we downloaded the Atlas of Variant Age from the developer’s website³⁰. Chimpanzee variants called from 25 individuals were downloaded from the Great Ape Genome Project³¹.

Imputation of array data

Imputation of variants based on KOVA 2 was performed as previously described⁷. Variants present on the Infinium Global Screening Array (GSA-24v3-0_A1) were extracted from WGS data of 197 COVID-19 patients and imputed using Impute2³². Panel imputation accuracy was compared using the aggregated squared Pearson correlation coefficient (R^2) determined between the imputed genotype dosages and the true genotypes from genome data.

Calling of structural variants (SVs)

Manta v1.6³³ was used to call structural variants for individual WGS samples. The convertInversion.py script provided with Manta was applied to represent inversion events in the manner of gnomAD SV v2.1³⁴. Slightly different SV representations across VCF files were merged using swimmer³⁵. An SV was defined as known if it overlapped with any entry in the gnomAD SV v2.1 dataset.

RESULTS

Characteristics of the genetic variants in the KOVA 2 dataset

To construct a Korean control database, we collected WES and WGS data generated from multiple projects that targeted Korean individuals (Supplementary Table 1). Samples originated from the normal tissues of cancer patients (40.2%), healthy parents of rare disease patients (28.4%), or healthy volunteers (31.4%) (Supplementary Table 1). Raw reads from 6654 sequencing libraries (4258 WES and 2396 WGS) were collected, processed, and filtered according to criteria from our previous experience⁵ and other studies. Exclusion criteria for samples and variants are described in the Methods, Supplementary Tables 2, 3 and Supplementary Fig. 2. After filtering out 1349 samples (20.3% of initial samples), variants from the remaining 5305 samples (3409 WES and 1896 WGS) were used in subsequent analyses. A total of 40,414,379 SNVs (874,026 coding and 39,540,353 noncoding) and 2,888,275 indels (37,663 coding and 2,850,612 noncoding) were called. From WGS data only, 144,388 CNVs (65,017 deletions, 10,956 duplications, and 68,415 others) were called (Supplementary Fig. 6).

Evaluation of the minor allele frequency (MAF) distribution revealed high enrichment of rare variants (<1%) that included a larger proportion of novel variants not found in control gnomAD v3.1¹¹ database (Fig. 1a, Supplementary Table 4, and Supplementary Fig. 7). As seen in other population datasets, adding data from more Korean individuals was not sufficient to saturate newly discovered variants, whether in coding or noncoding regions, whereas common coding variants (>5% frequency in gnomAD v3.1) were quickly saturated at <500 samples (Fig. 1b). Interestingly, common noncoding variants still displayed an increasing trend by analyzing 1800 WGS samples. This finding indicates that larger sample size is needed to fully cover this group of variants (Fig. 1b). As expected, variant function indicators, such as the nonsilent/silent (NS/S) ratio, CADD³⁶, ReMM³⁷, FunSeq2³⁸, and LINSIGHT³⁹, all showed increased functionality as MAF decreased (Fig. 1c, d). PCA located KOVA 2 samples in a cluster that was separate from the samples from Japanese, northern Chinese, southern Chinese, and Southeast Asian individuals (Fig. 1e and Supplementary Fig. 5). Finally, we found that the distribution of variants in the proximal intron regions indicates strong selection

against any base change as variants approach exon–intron boundaries (Fig. 1f). These results demonstrate the high quality of the KOVA 2 variant set.

Pathogenic variants

To determine if the KOVA 2 set contains variants that have previously been annotated as pathogenic, we selected KOVA-specific rare variants (MAF <0.001) in high pLI genes and compared them against ClinVar. A total of 25 variants (seven loss-of-function (LoF) and 18 missense variants) that were identified in the KOVA 2 participants were labeled as “likely pathogenic” or “pathogenic” in relation to diseases that follow a dominant inheritance pattern (Table 1). This observation suggests that these variants may not be pathogenic, as they were previously thought to be. Alternately, since KOVA 2 is composed of three main types of individuals, i.e., healthy volunteers, normal genomes of cancer patients, and healthy parents of rare disease patients, one may argue that the variants may predispose carriers to develop cancer or their children to manifest rare diseases.

Regions of homozygosity and positive selection

Compared to populations with a higher burden of consanguinity, homozygous pathogenic variants in ROH are rarely found in an outbred population, such as the Korean population⁴⁰. Rather, such regions can be used to signify a positive selection that was imposed on the population in the form of a selective sweep⁴¹. In terms of the ROH profile, the population in KOVA 2 does not deviate much from East Asian populations in general (Supplementary Fig. 8). To further characterize intervals that represent positive selection in KOVA 2, we applied the iSAFE algorithm²⁵. It yielded a total of 16,272 loci that were selected in at least one population (iSAFE >0.2) and identified a number of unique loci unique to each population (172 for KOVA 2, 149 for the Japanese population, 77 for the Chinese population, and 364 for the European population) (Fig. 2a). Although the functional and expression analyses of these loci did not yield notable features, a well-known locus in *LCT* showed a strong selection signal in the European population (Fig. 2b). Interestingly, we provide evidence that two loci—*ADH1A/1B* and *UHRF1BP1*—are among the most strongly selected loci in the Korean population when compared to the Japanese, Chinese, and European populations (Fig. 2b). *ADH1A* and *ADH1B* encode alcohol dehydrogenases 1A and 1B and are known to comprise a recently selected locus in East Asian individuals^{42,43}. Here, we show that the Korean population displayed the strongest signal among the East Asian populations that we evaluated. It has the highest frequency of “haplotype #1”, which represents the East Asian haplotype identified by a previous study⁴² (Fig. 2c). This signal is also reflected in the minor allele frequency of rs1229984, which was the lowest among the studied populations (Fig. 2c). The prominent Korean-specific signal we observed in *UHRF1BP1* has not been reported elsewhere, and the function of the gene remains largely unknown.

Allele ages

Next, we sought to determine whether KOVA 2 can be used to estimate the dates of origin for variants or allele ages and the implications of such information with regard to the function and frequency of variants. Notably, the estimation of allele ages may lead to the discovery of population-specific variants that have emerged recently. To carry out this analysis, we first phased our WGS-originated variants using a previously reported method^{7,16,44}. This allowed us to estimate the population size, which came to 10–20 million. This is a value that is similar to that of the current Korean population size of ~50 million—especially given the recent population explosion (e.g., the Korean population was ~20 million in 1950 and 13 million in 1925⁴⁵) (Supplementary Fig. 9). Next,

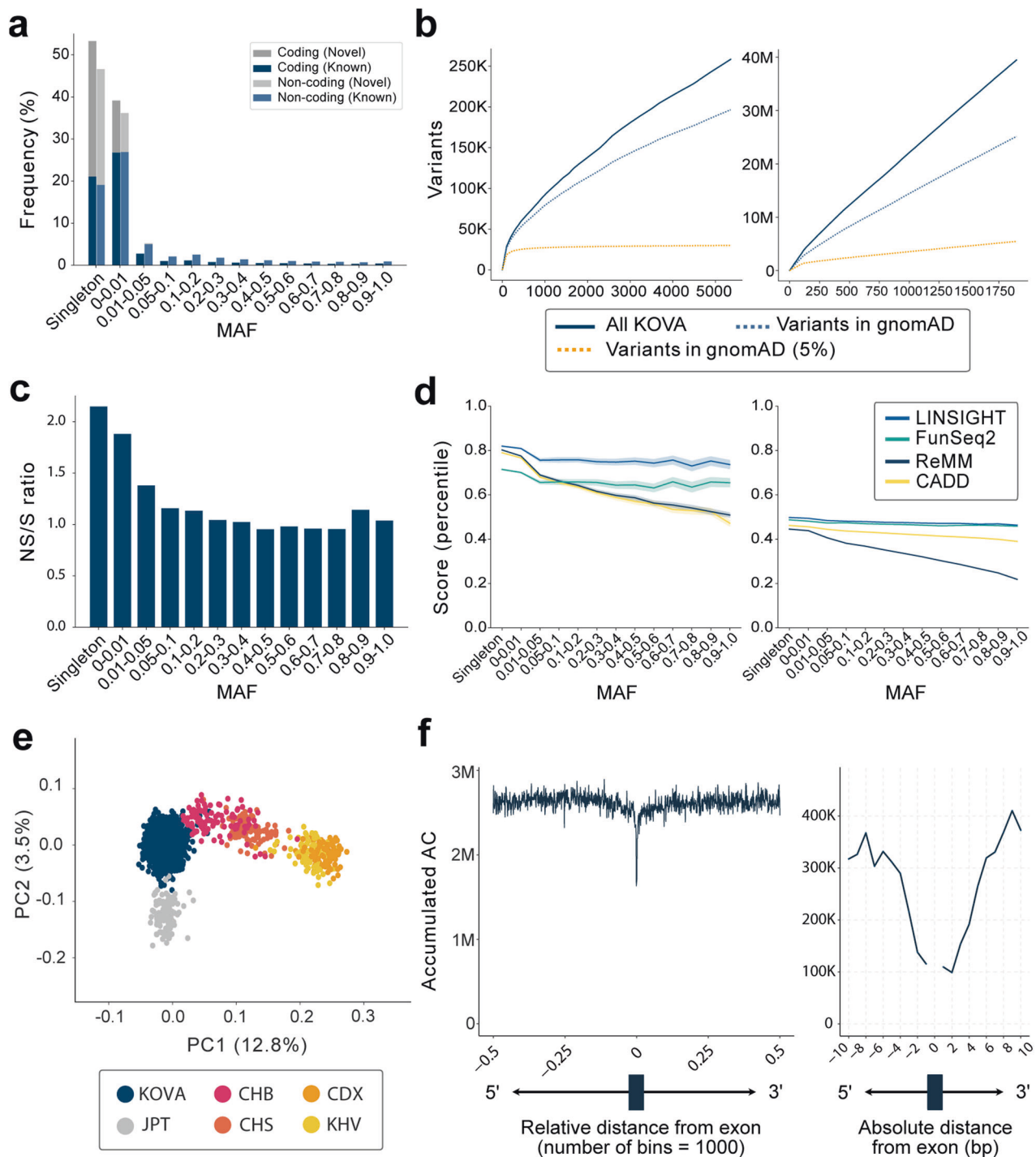


Fig. 1 Profile of variants in the KOVA 2 dataset. **a** Variant frequency by MAF according to variant status: coding and noncoding, known and novel. **b** The number of variants identified as the number of included KOVA 2 individuals was increased and is divided by the coding (left) and noncoding (right) status. **c** The nonsilent/silent (NS/S) ratio of coding variants by MAF. **d** The patterns of variant functionality predicted by different software according to MAF are divided by the coding (left) and noncoding (right) status. As each program produces scores with different scales, and each scoring system was converted to percentiles. **e** The PCA of KOVA 2 and the neighboring East Asian populations. CHB Han Chinese individuals from Beijing, CDX Chinese individuals Dai from Xishuangbanna, JPT Japanese individuals from Tokyo, CHS Han Chinese individuals from South China, and KHV Kinh individuals from Ho Chi Minh City. **f** Intron variant burden according to the relative (left) or absolute (right) distance from exons. The X-axis bins in **c** and **d** are the same as those in **a**.

variants with frequency >1% were used to estimate allele ages. As expected, the obtained allele ages showed a strong correlation with the MAF. The allele age was greater in variants with a high MAF or vice versa. Interestingly, variants of greater age showed

higher overlap with variants from chimpanzees, suggesting that some of these variants may have a primate-level origin (Fig. 3a and Supplementary Fig. 10). Separating variants by function revealed that older allele ages and higher overlap with

Table 1. Missense and high confidence (HC) loss-of-function (LoF) variants identified in KOVA 2 that are pathogenic or likely pathogenic in ClinVar but not found in gnomAD.

Variant class	Locus (hg38)	Base change	AC	AN	AF	Carrier type*	Gene symbol	pLI	ClinVar**	Dominant or Recessive***	ClinVar condition
LoF (HC)	chr3:128481942	CG > C	1	12,234	0.8 × 10 ⁻⁴	C	GATA2	0.98	P	D	Lymphedema, primary, with myelodysplasia; GATA2 deficiency with susceptibility to MDS/AML
	chr3:41236467	CAG > C	1	12,148	0.8 × 10 ⁻⁴	V	CTNNB1	1.00	P	D	Mental retardation, autosomal dominant 19; Inborn genetic diseases
	chr6:79026060	A > C	1	12,150	0.8 × 10 ⁻⁴	P	PHIP	1.00	P	D	Developmental delay, intellectual disability, obesity, and dysmorphic features
	chr7:128846444	C > T	1	12,136	0.8 × 10 ⁻⁴	C	FLNC	1.00	P	D	Myofibrillar myopathy, filamin C-related; Myopathy, distal, 4; Cardiomyopathy, familial hypertrophic, 26; Dilated cardiomyopathy, dominant
	chr9:95458142	G > T	1	12,120	0.8 × 10 ⁻⁴	V	PTCH1	1.00	P	D	Gorlin syndrome
	chr12:868379	C > T	1	12,134	0.8 × 10 ⁻⁴	C	WNK1	1.00	P	D/R	Hereditary sensory and autonomic neuropathy type IIA
	chrX:40064351	G > A	1	12,224	0.8 × 10 ⁻⁴	C	BCOR	1.00	P	D	Oculofaciocardiodental syndrome
	chr1:42927147	C > T	1	12,152	0.8 × 10 ⁻⁴	C	SLC2A1	0.99	LP	D/R	Not provided
	chr2:108753474	A > G	1	9310	1.1 × 10 ⁻⁴	C	RANBP2	1.00	P	D	Encephalopathy, acute, infection-induced, 3, susceptibility to
	chr3:123296110	G > A	1	12,122	0.8 × 10 ⁻⁴	P	ADCY5	0.99	LP	D/R	Inborn genetic diseases
Missense	chr3:128483925	C > T	1	12,238	0.8 × 10 ⁻⁴	P	GATA2	0.98	P	D	Lymphedema, primary, with myelodysplasia; GATA2 deficiency with susceptibility to MDS/AML
	chr5:128395182	C > T	1	12,160	0.8 × 10 ⁻⁴	P	FBN2	1.00	C	D	Congenital contractural arachnodactyly
	chr5:138570987	T > C	3	10,530	2.8 × 10 ⁻⁴	P, V	HSPA9	0.97	P	D/R	Even-plus syndrome
	chr6:157206668	C > T	1	12,144	0.8 × 10 ⁻⁴	V	ARID1B	1.00	LP	D	Coffin-Siris syndrome 1
	chr6:3154909	C > T	1	12,116	0.8 × 10 ⁻⁴	V	TUBB2A	0.94	P/LP	D	Cortical dysplasia, complex, with other brain malformations 5
	chr7:150952508	G > A	1	12,128	0.8 × 10 ⁻⁴	C	KCNH2	0.99	LP	D	Arrhythmia; Long QT syndrome 2; Congenital long QT syndrome
	chr7:5528486	G > C	1	12,126	0.8 × 10 ⁻⁴	V	ACTB	0.99	LP	D	Not provided
	chr9:130872896	C > T	1	12,244	0.8 × 10 ⁻⁴	C	ABL1	1.00	P/LP	D	Chronic myelogenous leukemia, BCR-ABL1-positive; Lymphoblastic leukemia, acute, with lymphomatous features; Leukemia, Philadelphia chromosome-positive, resistant to imatinib

Table 1. continued

Variant class	Locus (hg38)	Base change	AC	AN	AF	Carrier type*	Gene symbol	pLI	ClinVar**	Dominant or Recessive***	ClinVar condition
	chr9:132328351	A > G	1	12,152	0.8×10^{-4}	P	SETX	0.96	P	D/R	Spinocerebellar ataxia, autosomal recessive, with axonal neuropathy 2
	chr11:119089747	G > A	1	12,144	0.8×10^{-4}	V	HMB5	0.95	P	D	Acute intermittent porphyria
	chr11:119092785	G > A	1	12,138	0.8×10^{-4}	V	HMB5	0.95	LP	D	Not provided
	chr12:47978736	G > A	8	10,014	8.0×10^{-4}	C, P, V	COL2A1	1.00	LP	D	Spondyloepiphyseal dysplasia, Namaqualand type
	chr15:48470646	C > T	1	12,140	0.8×10^{-4}	P	FBN1	1.00	LP	D	Not provided
	chr16:9840706	G > A	1	12,148	0.8×10^{-4}	V	GRIN2A	1.00	P/LP	D	Epilepsy, focal, with speech disorder and with or without mental retardation;
	chr18:44951948	G > A	1	12,250	0.8×10^{-4}	C	SETBP1	1.00	P	D	Chronic myelogenous leukemia, BCR-ABL1 positive; Schinzel-Giedion syndrome

*C: normal sample of a cancer patient; P: parent of a rare disease patient; V: healthy volunteer.

**P: pathogenic, LP: likely pathogenic.

***D: dominant, R: recessive, D/R: observed in both patterns.

chimpanzees corresponded to less functionality, as indicated by annotation (Fig. 3b, Supplementary Table 5, and Supplementary Figs. 10, 11). Remarkably, high confidence LoF and high CADD score missense variants were the youngest and showed minimal overlap with chimpanzees. In addition, all functional classes of rare variants with a MAF <5% were young and did not overlap with chimpanzees (Fig. 3c and Supplementary Fig. 10). This trend was not clearly replicated when variants were categorized by pLI score (Supplementary Fig. 12). Overall, these findings suggest that most rare variants are of relatively recent origin and therefore tend to be population-specific.

DISCUSSION

Here, we report the establishment of the largest Korean control genome database to date, along with its genetic features and applications. KOVA 2 displayed the major features of ethnic genome databases, and we added considerable genetic information to the dataset. The variant set in KOVA 2 has been uploaded and will be shared among the community to be used as a control set in East Asian genetic studies.

First, to validate the integrity of our calling pipeline, we performed three distinct comparison evaluations. (1) A variant set from a sample that was analyzed by different sequencing techniques was compared. The variants showed a 99.4% concordance rate between HiSeq and PacBio calls and a 99.8% concordance rate between HiSeq and NovaSeq calls (Supplementary Tables 6, 7). (2) Additionally, our pipeline called 97.2% of Sanger-validated variants, and the missing calls were entirely caused by low-coverage regions (Supplementary Table 8). (3) Finally, a comparison of common variant calls across the WES and WGS platforms was performed. Among the 45,413 common coding variants (>5% frequency), 40,489 were detected by WES, and 45,118 were detected by WGS, showing 88.5% concordance (Supplementary Table 9). The missed calls were primarily from WES. This concordance is similar to one that was calculated from a recent study using 150,119 UK Biobank individuals⁴⁶.

Within the KOVA 2 dataset, the identified variants displayed typical patterns of purifying selection and frequency-functionality relationships. The sample size does not confer enough power to cover all rare variants in the population, as is the case with other larger variant sets. However, it exhibits the best coverage of common variants in Koreans and hence better performance when imputing variants (Supplementary Fig. 13). Although KOVA 2 can serve as a control set for filtering nonpathogenic variants for rare Mendelian diseases, we identified a list of ClinVar pathogenic variants that are present at low frequency. Whether these variants are nonpathogenic in the Korean population or the genomic background enabled these carriers to evade developing the associated diseases should be further elucidated.

A combined analysis of the positive selection signatures and allele age estimation may lead to the discovery of genetic loci that have recently arisen and been selected in a population. Not surprisingly, the top signals in this Korean population overlapped with those from neighboring populations in East Asia. This finding demonstrates a recent diversion, the continuous admixture, and the similar environmental constraints that were exerted on these populations during recent evolution. Nevertheless, our results identified loci that merit further investigation. For example, we further dissected a known East Asian-selected alcohol dehydrogenase gene locus and found that it was the most strongly selected locus among the three East Asian populations analyzed here. The major haplotype in East Asian populations ("Haplotype #1" in Fig. 2c) was the most abundant in the Korean population, and among East Asian populations, the Korean population had the highest frequency of the

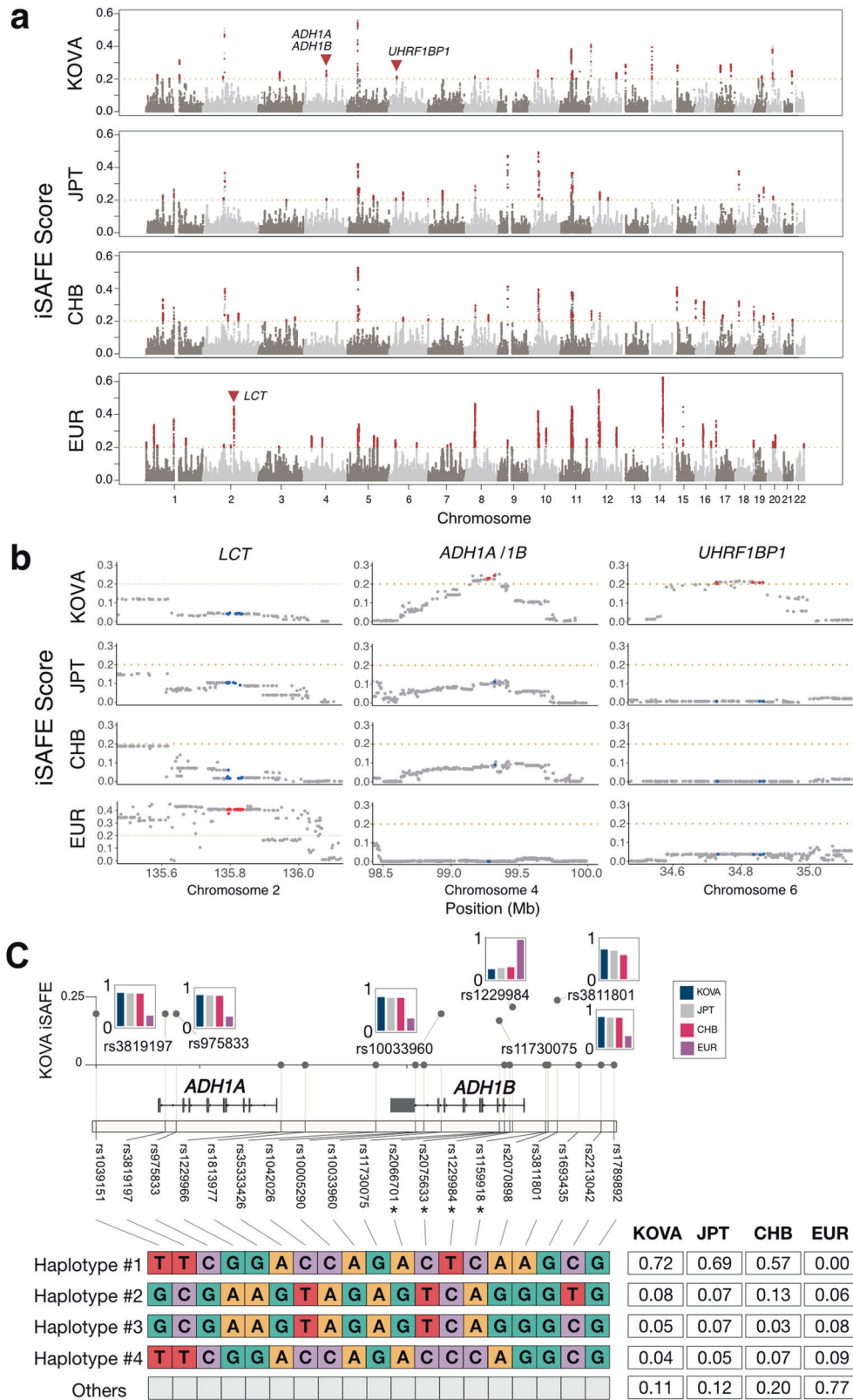


Fig. 2 Signature of positive selection as indicated by the iSAFE score. a Genome-wide iSAFE values were obtained using KOVA 2, Japanese, Chinese, and European cohorts. Gene loci indicated with triangles are separately displayed in **b**. **b** Regional plots of the iSAFE values from the same set of ethnic cohorts, as marked in **a**. **c** The KOVA iSAFE scores of selected tag SNPs in the *ADH1A/1B* locus (top) and their haplotype frequencies by population (bottom). Bar plots denote the MAF of designated SNPs in each population, and SNPs with asterisks denote major markers for haplotype identification used in Han et al., 2007⁴².

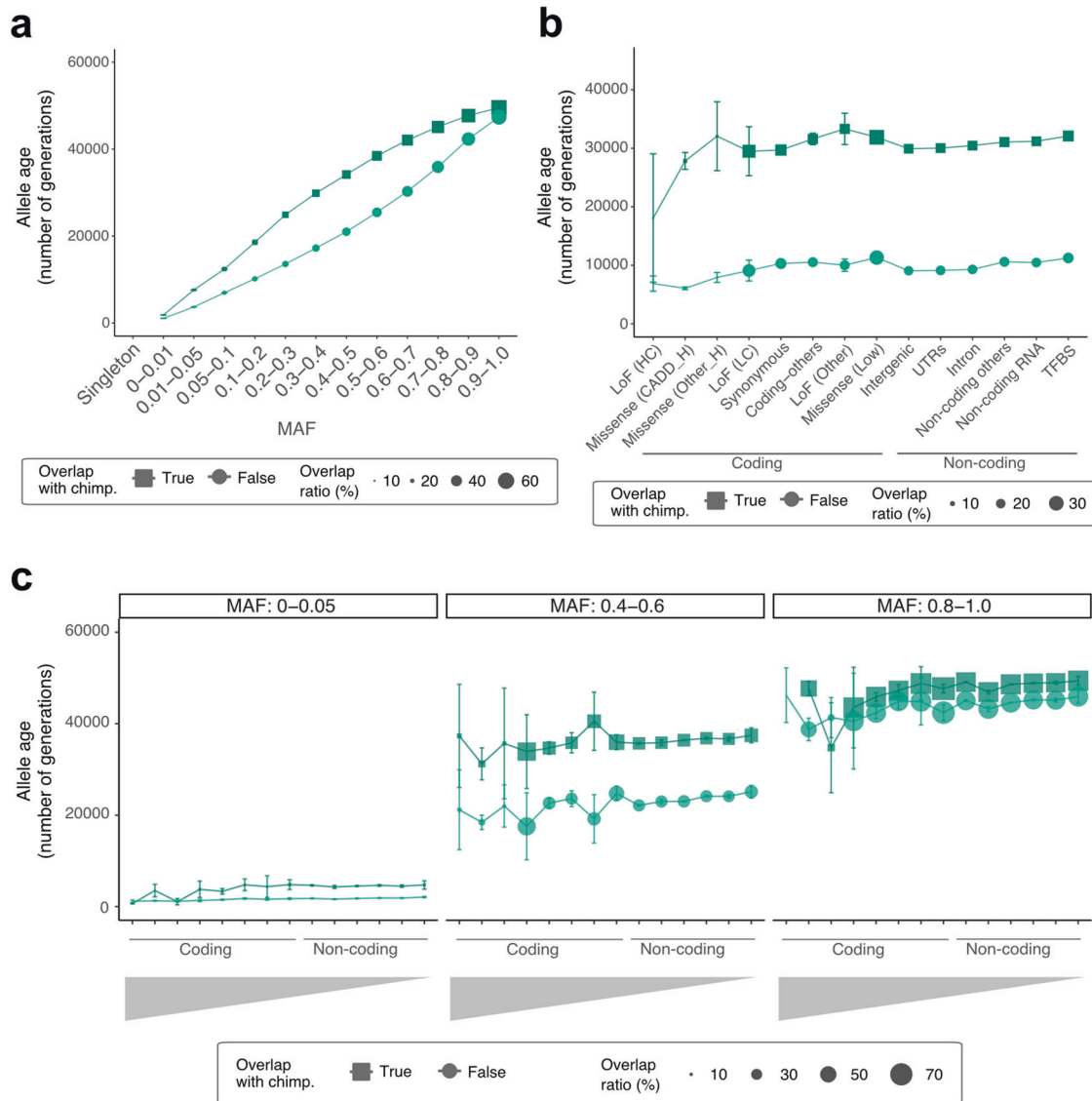


Fig. 3 Allele ages of KOVA 2 variants. **a** Allele ages by MAF, divided by whether the allele cooccurs in chimpanzees (squares) or not (circles). **b** Allele age by predicted function, divided by whether the allele cooccurs in chimpanzees (squares) or not (circles). TFBS denotes the transcription factor-binding site. **c** Allele age by MAF and predicted function. Three MAF intervals are displayed. The X-axis bins in **c** are the same as those in **b**.

functional variant *ADH1B* p.Arg48His (Fig. 2c). This variant is known to cause increased aldehyde production relative to its wild-type counterpart. This is due to the increased oxidation of ethanol and subsequently triggers adverse reactions, such as flushing and nausea⁴⁷. In the long term, the variant is also protective against alcohol dependency⁴⁸. The functional consequence of the second locus of interest, *UHRF1BP1*, remains uncertain, as it has received little study. Nevertheless, it is remarkable that associations between the variants in this gene and systematic lupus erythematosus have been repeatedly reported in East Asian populations^{49–51}. This gene is most strongly expressed in the testes (Supplementary Fig. 14), making it possible to infer that it can confer selection by affecting the reproductive process in males. Looking beyond these two loci, a new algorithm based on large-scale population data may discover novel loci that were missed in our study.

Finally, we deposited the data in a genome browser and enabled the downloading of the variant set by users with a minimal registration process. The establishment of a Korean-

specific variant set and comparative analysis will bolster various types of genetic and genomic studies involving East Asian populations. In addition, it will serve as a precursor for much larger genome datasets that will be available soon, especially if they can be merged with data from North Korean individuals.

DATA AVAILABILITY

A list of the annotated SNVs, indels, and CNVs with frequency information can be viewed and downloaded from the KOVA 2 website (<https://www.kobic.re.kr/kova/>).

REFERENCES

- Jin, H.-J. et al. Y-chromosomal DNA haplogroups and their implications for the dual origins of the Koreans. *Hum. Genet.* **114**, 27–35 (2003).
- Kim, W., Shin, D. J., Harihara, S. & Kim, Y. J. Y chromosomal DNA variation in East Asian populations and its potential for inferring the peopling of Korea. *J. Hum. Genet.* **45**, 76–83 (2000).
- Wang, Y., Lu, D., Chung, Y.-J. & Xu, S. Genetic structure, divergence and admixture of Han Chinese, Japanese and Korean populations. *Hereditas* **155**, 19 (2018).

4. Sirugo, G., Williams, S. M. & Tishkoff, S. A. The missing diversity in human genetic studies. *Cell* **177**, 26–31 (2019).
5. Lee, S. et al. Korean variant archive (KOVA): a reference database of genetic variations in the Korean population. *Sci. Rep.* **7**, 4287 (2017).
6. Kwak, S. H. et al. Findings of a 1303 Korean whole-exome sequencing study. *Exp. Mol. Med.* **49**, e356–e356 (2017).
7. Jeon, S. et al. Korean Genome Project: 1094 Korean personal genomes with clinical information. *Sci. Adv.* **6**, eaaz7835 (2020).
8. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. Preprint at arXiv <https://doi.org/10.48550/arXiv.1303.3997> (2013).
9. Auwera, G. A. V. der & O'Connor, B. D. *Genomics in the Cloud: Using Docker, GATK, and WDL in Terra*. (O'Reilly Media, 2020).
10. Okonechnikov, K., Conesa, A. & Garcia-Alcalde, F. Qualimap 2: advanced multi-sample quality control for high-throughput sequencing data. *Bioinformatics* **32**, 292–294 (2016).
11. Karczewski, K. J. et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).
12. Hail Team. Hail 0.2.77-684f32d73643. <https://github.com/hail-is/hail/releases/tag/0.2.77> (2021).
13. Li, H. Toward better understanding of artifacts in variant calling from high-coverage samples. *Bioinformatics* **30**, 2843–2851 (2014).
14. Conomos, M. P., Reiner, A. P., Weir, B. S. & Thornton, T. A. Model-free estimation of recent genetic relatedness. *Am. J. Hum. Genet.* **98**, 127–148 (2016).
15. Team, H. “maximal independent set” method. https://hail.is/docs/0.2/methods/misc.html#hail.methods.maximal_independent_set (2021).
16. Auton, A. et al. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
17. McLaren, W. et al. The Ensembl variant effect predictor. *Genome Biol.* **17**, 122 (2016).
18. Delaneau, O., Zagury, J.-F., Robinson, M. R., Marchini, J. L. & Dermitzakis, E. T. Accurate, scalable and integrative haplotype estimation. *Nat. Commun.* **10**, 5436 (2019).
19. Delaneau, O., Zagury, J.-F., Robinson, M. R., Marchini, J. L. & Dermitzakis, E. T. Genetic map for reference version hg38 by SHAPEIT4. https://github.com/odelaneau/shapeit4/blob/master/maps/genetic_maps.b38.tar.gz (2018).
20. Browning, B. L., Tian, X., Zhou, Y. & Browning, S. R. Fast two-stage phasing of large-scale sequence data. *Am. J. Hum. Genet.* **108**, 1880–1890 (2021).
21. Browning, B. L., Tian, X., Zhou, Y. & Browning, S. R. Genetic map for reference version hg38 by Beagle 5.2. http://bochet.gcc.biostat.washington.edu/beagle/genetic_maps/plink.GRCh38.map.zip (2018).
22. Purcell, S. & Chang, C. PLINK 1.9. www.cog-genomics.org/plink/1.9/ (2019).
23. Chang, C. C. et al. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**, 1–16 (2015).
24. Biscarini, F., Cozzi, P., Gaspa, G. & Marras, G. detectRUNS: Detect runs of homozygosity and runs of heterozygosity in diploid genomes. (2018).
25. Akbari, A. et al. Identifying the favored mutation in a positive selective sweep. *Nat. Methods* **15**, 279–282 (2018).
26. Browning, S. R. & Browning, B. L. Accurate non-parametric estimation of recent effective population size from segments of identity by descent. *Am. J. Hum. Genet.* **97**, 404–418 (2015).
27. Zhou, Y., Browning, S. R. & Browning, B. L. A fast and simple method for detecting identity-by-descent segments in large-scale data. *Am. J. Hum. Genet.* **106**, 426–437 (2020).
28. Browning, B. L. & Browning, S. R. Improving the accuracy and efficiency of identity-by-descent detection in population data. *Genetics* **194**, 459–471 (2013).
29. Albers, P. K. & McVean, G. Dating genomic variants and shared ancestry in population-scale sequencing data. *PLoS Biol.* **18**, e3000586 (2020).
30. Albers, P. K. & McVean, G. Human genome dating. <https://human.genome.dating/download/index> (2020).
31. Prado-Martinez, J. et al. Great ape genetic diversity and population history. *Nature* **499**, 471–475 (2013).
32. Howie, B. N., Donnelly, P. & Marchini, J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet.* **5**, e1000529 (2009).
33. Chen, X. et al. Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinformatics* **32**, 1220–1222 (2016).
34. Collins, R. L. et al. A structural variation reference for medical and population genetics. *Nature* **581**, 444–451 (2020).
35. Eggertsson, H. P. et al. GraphTyper2 enables population-scale genotyping of structural variation using pangenome graphs. *Nat. Commun.* **10**, 5402 (2019).
36. Rentzsch, P., Witten, D., Cooper, G. M., Shendure, J. & Kircher, M. CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Res.* **47**, D886–D894 (2018).
37. Smedley, D. et al. A whole-genome analysis framework for effective identification of pathogenic regulatory variants in Mendelian disease. *Am. J. Hum. Genet.* **99**, 595–606 (2016).
38. Fu, Y. et al. FunSeq2: a framework for prioritizing noncoding regulatory variants in cancer. *Genome Biol.* **15**, 480 (2014).
39. Huang, Y.-F., Gulko, B. & Siepel, A. Fast, scalable prediction of deleterious non-coding variants from functional and population genomic data. *Nat. Genet.* **49**, 618–624 (2017).
40. Lee, Y. et al. Genomic profiling of 553 uncharacterized neurodevelopment patients reveals a high proportion of recessive pathogenic variant carriers in an outbred population. *Sci. Rep.* **10**, 1413 (2020).
41. Pemberton, T. J. et al. Genomic patterns of homozygosity in worldwide human populations. *Am. J. Hum. Genet.* **91**, 275–292 (2012).
42. Han, Y. et al. Evidence of positive selection on a class I ADH locus. *Am. J. Hum. Genet.* **80**, 441–456 (2007).
43. Okada, Y. et al. Deep whole-genome sequencing reveals recent selection signatures linked to evolution and disease risk of Japanese. *Nat. Commun.* **9**, 1631 (2018).
44. Wall, J. D. et al. The GenomeAsia 100K Project enables genetic discoveries across Asia. *Nature* **576**, 106–111 (2019).
45. Korea, S. Korean statistical information service. <https://kosis.kr/eng/> (2022).
46. Halldorsson, B. V. et al. The sequences of 150,119 genomes in the UK Biobank. *Nature* **607**, 732–740 (2022).
47. Edenberg, H. J. The genetics of alcohol metabolism: role of alcohol dehydrogenase and aldehyde dehydrogenase variants. *Alcohol Res. Health* **30**, 5–13 (2007).
48. Li, D., Zhao, H. & Gelernter, J. Strong association of the alcohol dehydrogenase 1B gene (ADH1B) with alcohol dependence and alcohol-induced medical diseases. *Biol. Psychiatry* **70**, 504–512 (2011).
49. Wu, J. et al. The rare variant rs35356162 in UHRF1BP1 increases bladder cancer risk in Han Chinese population. *Front. Oncol.* **10**, 134 (2020).
50. Morris, D. L. et al. Genome-wide association meta-analysis in Chinese and European individuals identifies ten new loci associated with systemic lupus erythematosus. *Nat. Genet.* **48**, 940–946 (2016).
51. Yin, X. et al. Meta-analysis of 208370 East Asians identifies 113 susceptibility loci for systemic lupus erythematosus. *Ann. Rheum. Dis.* **80**, 632–640 (2021).

ACKNOWLEDGEMENTS

We thank the Korea Bioinformation Center (KOBIC) K-BDS and Clinical and Omics Data Archive for sharing Korean genome data and Drs. Semin Lee, Jungmin Choi, Joon Yong An, and Choong Won Chung for critical comments. A part of this study was supported by the research programs through the National Research Foundation funded by the Ministry of Science and Technology (2014M3C9A2064686, 2018M3C9A5064708, and 2020M3E5D7086836).

AUTHOR CONTRIBUTIONS

Jeongeun L. and Jean L. compiled the genome data, processed and analyzed the data, and wrote the manuscript. S.J., Jeongha L., and S.P. analyzed the data. I.J., J.O.Y., and B.L. constructed the data-sharing website. J.C., B.-O.C., H.Y.G., J.O., I.-J.J., S.L., D.B., Y.K., S.-S.Y., Y.-J.K., J.-H.C., and W.-Y.P. provided the genome data. J.H.B. provided the genome data and analyzed the data. M.C. designed the research studies, analyzed the data, and wrote the manuscript.

COMPETING INTERESTS

The authors declare no competing interests.

ADDITIONAL INFORMATION

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s12276-022-00871-4>.

Correspondence and requests for materials should be addressed to Jong Hwa Bhak or Murim Choi.

Reprints and permission information is available at <http://www.nature.com/reprints>

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



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

© The Author(s) 2022