공학박사 학위논문

A Study on Genetic Implications of Korean Individuals through the Establishment of Genome Dataset

유전체 데이터세트 구축을 통한 한국인의 유전적 함의에 관한 연구

2023년 2월

서울대학교 대학원

협동과정 바이오엔지니어링 전공

이 정 은

Ph.D. Dissertation

A Study on Genetic Implications of Korean Individuals through the Establishment of Genome Dataset

유전체 데이터세트 구축을 통한 한국인의 유전적 함의에 관한 연구

February 2023

Interdisciplinary Program in Bioengineering
The Graduate School
Seoul National University

Jeongeun Lee

A Study on Genetic Implications of Korean Individuals through the Establishment of Genome Dataset

지도교수 최 진 욱, 최 무 림

이 논문을 공학박사 학위논문으로 제출함
2022년 12월

서울대학교 대학원
협동과정 바이오엔지니어링 전공
이 정 은

이정은의 공학박사 학위논문을 인준함
2023년 1월

위 원 장 김 종 일 (인)

부위원장 최 진 욱 (인)

부위원장 최 무 림 (인)

위 원 최 지 엽 (인)

위 원 조 성 엽 (인)

위 원 최 정 민 (인)

A Study on Genetic Implications of Korean Individuals through the Establishment of Genome Dataset

Advisor: Jinwook Choi, Murim Choi

Submitting a Ph.D. Dissertation

December 2022

Interdisciplinary Program in Bioengineering
The Graduate School
Seoul National University

Jeongeun Lee

Confirming the Ph.D. Dissertation written by
Jeongeun Lee
January 2023

Chair	Jong-Il Kim (Seal)
Vice Chair	Jinwook Choi (Seal)
Vice Chair	Murim Choi (Seal)
Examiner	Ji-Yeob Choi (Seal)
Examiner	Sung-Yup Cho (Seal)
Examiner	Jungmin Choi (Seal)

Abstract

A Study on Genetic Implications of Korean Individuals through the Establishment of Genome Dataset

Jeongeun Lee
Interdisciplinary Program in Bioengineering
The Graduate School
Seoul National University

Understanding genetic architectures of healthy individuals is fundamental in the study of physiology of human development and disease, as well as clinical diagnosis of genetic disease. Accordingly, in line with substantial advances in disease genetics, the importance of the genome database of general population has also emerged. However, studies to date have largely focused on individuals of European descent. This limits further discoveries of novel functional genetic variants in other ethnic groups. As a result, efforts to establish independent population-specific genome databases for each East Asian country have gradually grown, but the current state of Korean genome database construction is not reaching the database construction speed of neighboring East Asian countries.

In this study, in order to resolve the paucity of Korean

population genome resources and contribute to the establishment of an East Asian population genome database, a Korean genome database (KOVA2) consisting of 1,896 whole genome sequences and 3,409 whole exome sequences of healthy Koreans was established. This is the largest Korean-specific genome database ever, surpassing the 1,909 Korean genome data included in gnomAD. The constructed genome database detected mutations through the newly developed pipeline which takes the raw sequence data as an input, and only high quality variants from the healthy Koreans were included in the database. In total, 40,414,379 SNVs and 2,888,275 insertions/deletions were obtained, and 144,388 structural variants called from whole genome data were cataloged. A sample from KOVA2 was sequenced with another sequencing platform to evaluate the integrity of the calling pipeline, and it showed high concordance rate between sequencing platforms. Also, all known genetic characteristics reported from previously published genome databases were identified from KOVA 2.

The KOVA2 database was analyzed to additionally characterize the Korean-specific genetic features including the runs of homozygosity (ROH), the positively selected regions, allele age. In the process, we found loci that are strongly selected in Koreans compared to other East Asian populations, such as the ADH1A/1B and UHRF1BP1 loci. Analysis of allele age revealed a correlation between variant functionality and allele age. There was no significant difference in ROH regions of Koreans with other East

Asians. Estimation of the effective population size by time showed similar results that match to the population statistics record of Koreans.

Called variants from KOVA2, including the estimated allele age and scores reflecting degree of positive selection were made available for search and download from public websites. The results of this study will serve as valuable resource that can provide a new insight for various genetic studies targeting East Asian populations.

* This thesis is based on a published article; Lee *et al.*, *Exp. Mol. Med.* 54:1862-1871 (2022) [1].

Keywords : genome database, whole-exome sequencing, whole-genome sequencing, East Asian, Korean, positive selection, allele age
Student Number : 2014-30270

Table of Contents

List of Figures and Tables

Chapter 1

Chapter 2

Chapter 3

List of Abbreviations

AFR	African
CDX	Chinese individuals Dai from Xishuangbanna
CHB	Chinese individuals from Beijing,
CHS	Han Chinese individuals from South China
EAS	East Asian
EUR	European
gnomAD	Genome aggregation database
GWAS	Genome-Wide Association Study
HWE	Hardy-Weinberg equilibrium
INDEL	Insertion or deletion variants
JPT	Japanese
KG	1000 Genomes Project Phase 3
KHV	Kinh individuals from Ho Chi Minh City
KOVA 2	Korean Variant Archive 2
LD	Linkage Disequilibrium
LoF	Loss of function variant
MAF	Minor Allele Frequency
PCA	Principal component analysis
pLI	The probability of being loss-of-function intolerant
ROH	Runs of homozygosity
SAS	South Asian
SNV	Single nucleotide variant

SV	Structural variants
VCF	Variant call format
WDL	Workflow Description Language
WES	Whole exome sequencing
WGS	Whole genome sequencing

Chapter 1. Introduction

The reference population database composed of the genomes of healthy people is fundamental resource in the study of physiology of human development and disease, study of human evolution, and clinical diagnosis of genetic disease. The recommendation of The American College of Medical Genetics and Genomics (ACMG) for mutation interpretation is also highlighting the importance of its usage in clinical practice as a tool to separate the putative pathogenic variants from benign variants which are variants observed in many healthy patients [2-3].

The control database has been developed in a way that covers more individuals and various sequencing type. As rare variants are found more in the larger genetic studies, the cohort size of the control database also has been increased to determine whether the rarity of the variant is not due to the lack of a sample group. Whole exome sequencing (WES) has been done a lot due to the ease of interpretation, but whole genome sequencing (WGS) is getting attention, as the need for study of non-coding region has become more prominent.

However, the problem with the current control database is that diversity of the population is limited. Especially, despite making up about 5% of the global population, only 8.2% of participants in genome-wide association studies (GWAS) [4] are from East Asia. Comparatively few control databases are available for East Asian

populations. The Genome Aggregation Database (gnomAD) [5], the most widely used control genome database due to the largest cohort size, has been increased its sample size every year until 2020 since the Exome Aggregation Consortium (ExAC) [6], which consists of WES data of 60,706 people, was released in 2016. Currently, 191,830 WES samples and 76,156 WGS samples are included. Of these, only 9977 WES samples from East Asia were included, and among them, only 1909 Korean samples were analyzed (Figure 1.1). Notably, small databases of genetic information from Korean individuals (with sizes of approximately one thousand individuals) have been released [7–9]. However, as the cohort sizes of human genetics studies increase, it becomes necessary to construct larger Korean control databases.

At least 40,000 years ago, people from the Korean peninsula are known to have traveled there; this migration likely took place via two routes, from northeast and southeast Asia [10–11]. Throughout history, there has been extensive but constant mixing with the nearby populations of the Chinese and Japanese [12], but numerous studies have revealed that ethnic Korean people are genetically distinct from Chinese and Japanese people. The ethnic Korean population is the world's 15th largest ethnic group with a population of about 83 million. Modern national health care systems, particularly in South Korea, can offer a chance to investigate the genetics of numerous diseases in this population if the necessary genetic infrastructure is put in place.

The goal of this thesis is to establish a genomic database of healthy Koreans and to identify Korean-specific genetic and clinical characteristics from it (Figure 1.2). In Chapter 2, I introduced the Korean Variant Archive 2 (KOVA 2), the largest genome database from the ethnic Korean population to date. which is composed of 1,896 whole genome sequences and 3,409 whole exome sequences from healthy individuals of Korean ethnicity. In Chapter 3, the patterns of runs of homozygosity, positive selected intervals, and allele age of variants were identified to investigate Korean-specific genetic features using the constructed KOVA2. In this process, 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. Also, mutations known to be pathogenic among KOVA2 variants are listed.

As the data is deposited and enabled to download from public genome browser ((https://www.kobic.re.kr/kova/), I hope that this study serve as a valuable resource for genetic studies of Korean and East Asian populations.

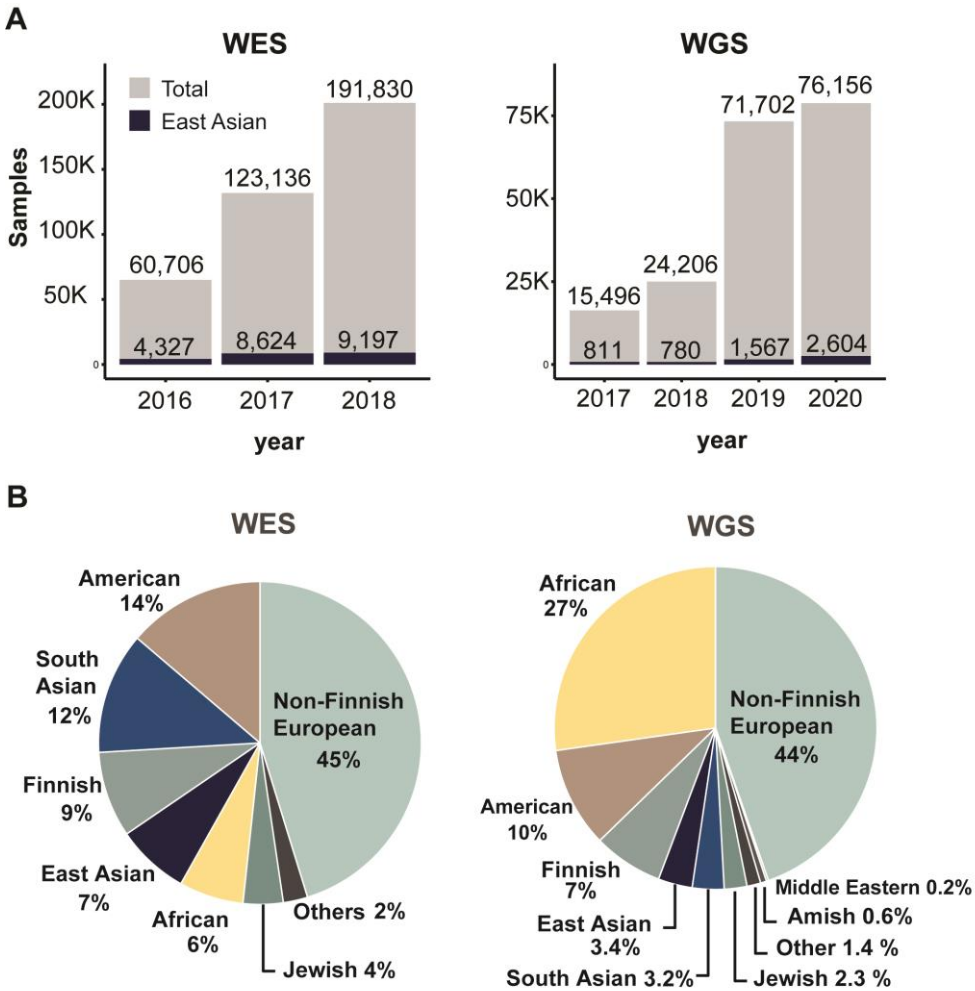

Figure 1.1. Small proportion of East Asians in the gnomAD database. (a) East Asian sample size increase compared with overall sample size increase. The gray bar shows the total number of samples included, and the blue bar shows the number of East Asian individuals included. (b) population composition of gnomAD database.

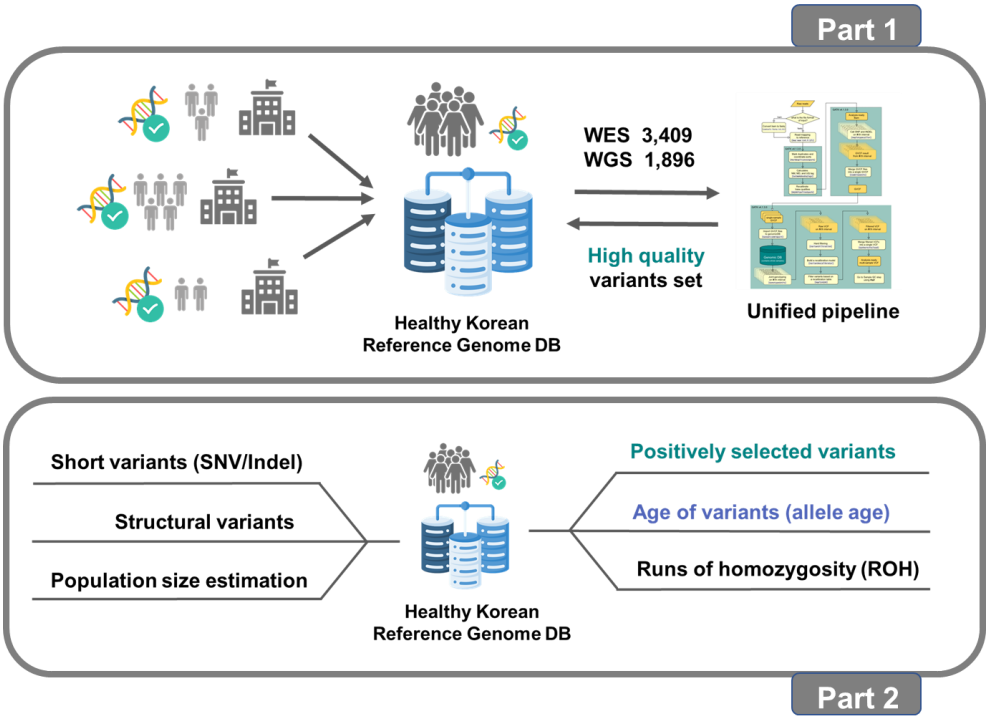

Figure 1.2. schematics of the thesis

Chapter 2. Korean Genome Database

2.1. Background

As the size of genetic studies increases, the importance of the control genome database has increased. As introduced in Chapter 1 above, the size of the gnomAD database covering the global population has been increasing, but recently, the genome database of a single ethnicity is expanding. Korea has also built WES and WGS databases in line with this trend, but the sample size is still small. (Table 2.1)

Most of the existing genomic databases were collected through a consortium consisting of several researchers [5-6]. The consortium has the advantage of generating vast data in a comparatively short time. However, if the protocol to detect variants is not unified from the time of DNA extraction, genetic characteristics may be represented in an unintentional way.

Many of the samples used in this study were previously sequenced by various researchers for other disease research, so the type of whole exome sequencing library was different or the sequencing coverage depth was different for each data center. Also, the pipeline used for variant calling was not unified. As a result, there was a problem that the read coverage of a specific region was low and under-represented, and a variant that was originally found

commonly was represented as a rare variant in the database, or the variant calling quality of a specific region could be low. Therefore, it was necessary to build a pipeline that can trust the quality of variants.

In addition, when recruiting donors on a large scale, the off-target sample in the study may be included in the cohort due to mistakes in the experimental process or administrative errors. A patient sample and a healthy control sample may be swapped due to sample labeling mistake, and one sample may be duplicated in the database due to a file naming mistake during the transferring data to other center. When family samples are registered to sample collection, the family samples should be removed except one sample [13]. Since samples with family relationships have a high genetic correlation with each other, genetic bias may occur. Especially, rare genetic mutations that are passed down from a particular family could be represented as common. Therefore, in order to construct a population database that reflects various genetic characteristics, efforts should be made to collect only one sample of healthy individuals without kinship.

Finally, it is necessary to operate a computing resource to handle big data. In order to detect variants of all samples from raw sequence data using a unified pipeline, capability of storage, memory, and CPU must be considered. Even if the number of WES samples is only 1000, at least 5 TB of disk space for the FASTA file containing the sequence information is required. Also, it

requires at least 2.5 times of the storage space during the workflow. The problem is that in order to process 1000 whole genome sequencing data, at least 10 times of the storage space is required for WES data. Accordingly, the computing time for the process is also considerable. Therefore, it is necessary to shorten the time required for variant calling in one sample as much as possible.

In this study, tools utilizing Apache spark which is a general-purpose distributed data processing were largely included in the pipeline, and this pipeline was implemented using Workflow Description Language (WDL) [14] to handle cloud computing resources. To ensure that only high quality variants are included in the Korean genome database, stringent quality control for sample and called-variant were applied.

Table 2.1 Population-specific genome database of East Asian

Name of database	Major population	Sequencing type (coverage depth)	Number of Healthy donors	Publication date
HuaBiao [15]	Han Chinese	WES (>x100)	5000	2021.11.
Taiwan Biobank [16]	Han Chinese	WGS (>30x)	1,445	2021.02.
NyuWa [17]	Han Chinese	WGS (~26.2x)	2,902	2021.11.
Westlake BioBank for Chinese [18]	Han Chinese	WGS (~13x)	1,151	2022.05
1KJPN [19]	Japanese	WGS (30x)	1,070	2015.08
3.5KJPN [20]	Japanese	WGS (30x)	3,552	2016.06
8.3KJPN [21]	Japanese	WGS (30x)	8,380	2021.11.
KOVA [7]	Korean	WES (>50x)	1,055	2017.06
Korea1K [9]	Korean	WGS (30x)	1,094	2020.05

2.2. Materials and Methods

2.2.1. Cohorts and sample preparation

The whole-exome sequencing (WES) and whole-genome sequencing (WGS) data for Korean individuals were collected from independent research groups in Korea (Table 2.2). All sequencing data were obtained from normal tissues or blood samples following standard protocols [7, 22]. This research 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 de-identified samples and in accordance with relevant guidelines and regulations. Sample collection was done in collaboration with Jean Lee.

Table 2.2. Sample collection. Numbers denote number of samples after sample filtering. Note “KOVA I” denotes the data was also used in the first version of KOVA (Lee et al., Sci Reports 2017).

Group leader	Center	WES	WGS	Total	Note
Woong-Yang Park	Samsung Genome Institute	1,181	-	1,181	KOVA I
Jong Hwa Bhak	Ulsan National Institute of Science and Technology	-	903	903	-
Murim Choi	Seoul National University	587	23	610	-
Jong-Hee Chae	Seoul National University Children’s Hospital	545	-	545	KOVA I
National Biobank of Korea	Korea Biobank Project	-	347	347	-
Young-Joon Kim	Yonsei University	-	324	324	-
The National Center for Medical Information and Knowledge	Clinical & Omics Data Archive (CODA)	-	299	299	-
Youngil Koh	Seoul National University Hospital	284	-	284	-
Daehyun Baek	Seoul National University	222	-	222	KOVA I
Sanghyuk Lee	Ewha Womans University	194	-	194	KOVA I
In-Jin Jang	Seoul National University	118	-	118	KOVA I
ETC		224	-	224	-
Heon Yung Gee	Yonsei University	45	-	45	-
Byung-Ok Choi	Samsung Medical Center	9	-	9	-
Total		**3,409**	**1,896**	**5,305**	**-**

2.2.2. Variant calling

To map raw reads to the GRCh38+decoy reference sequence BWA mem v0.7.17 [23] was used with default options. After marking duplicates and sorting by coordinate with MarkDuplicatesSpark, the mapping quality was recalibrated by BQSRPipelineSpark, implemented in GATK version 4.1.3.0 [22]. Qualimap v2.2.1 [24] 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 guideline [22]. Briefly, SNVs and indels were recalibrated by GATK's VQSR model to select 99.7% and 99.0% of true sites, respectively, from the training set. The detailed workflow is described in Figure 2.1. In the case of a tool that does not support spark, the workflow performed parallel processing by giving the finely divided target region interval as an input parameter. By creating a pipeline with WDL, merging the processes scattered over the divided intervals was stably performed. Furthermore, the pipeline was submitted to the Cromwell server which manage the cloud computing resource, so that computing time per sample could be shorten. By sharing the pipeline to Sungwon Jeon from Ulsan

National Institute of Science and Technology (UNIST), raw reads of 903 WGS data from Jong Wha Bhak were processed in UNIST computing cluster to get GVCF files. Also, all WGS files were jointly genotyped on the UNIST computing cluster. Further analyses adopted a modified version of gnomAD QC steps [5] and were mostly performed with Hail [25], which is an open-source Python library for genome data analysis. Hail is also a spark-based library, so it greatly helped speed up post processing. After merging the WES and WGS data using Hail, we excluded multi-allelic variants or were in low complexity regions [26]. Genotype calls that had a genotype quality (GQ) < 20, read depth (DP) < 10, allelic balance (AB) < 0.2 were also excluded from counting of alleles (Figure 2.2).

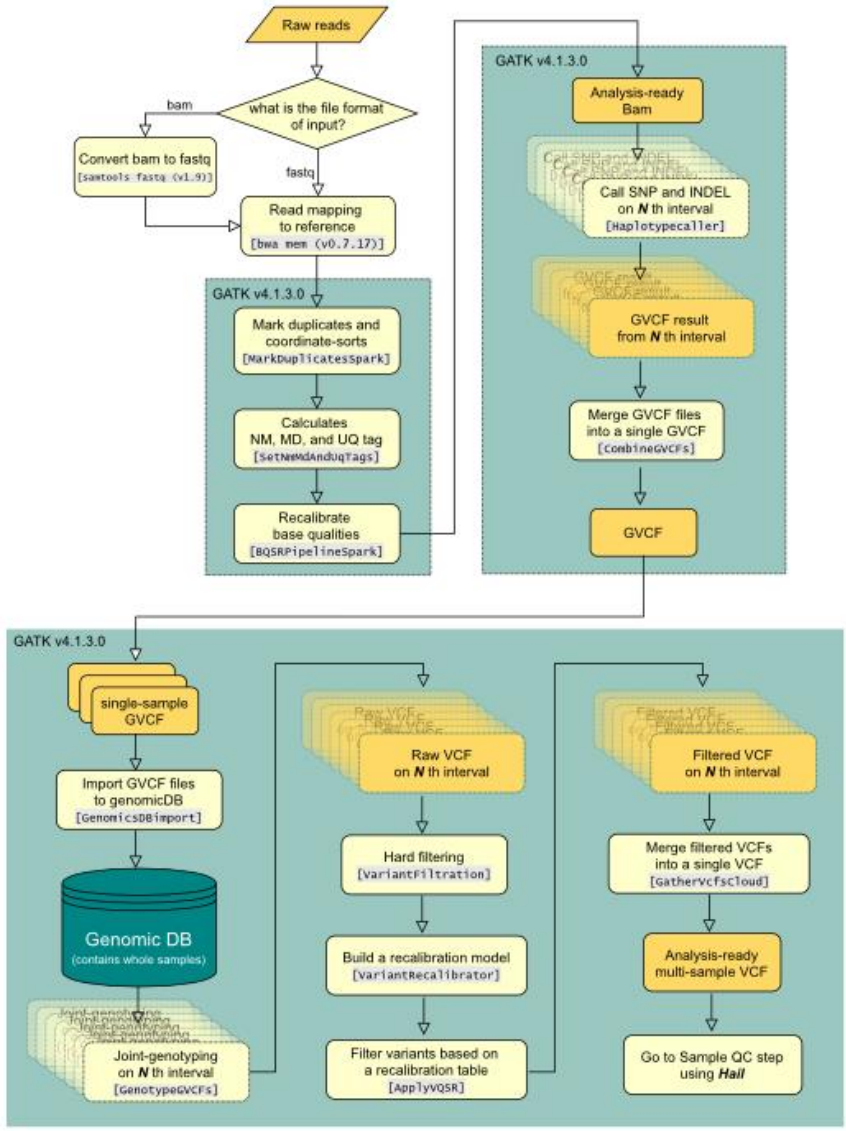

Figure 2.1. Variant calling pipeline. The name of the tool or function for each step is described in the bracket in a gray background. If the same procedure is performed in a reiterative manner, the step is depicted in overlapped boxes. The subsequent steps after BWA mapping are all proceeded using GATK v4.1.3.0.

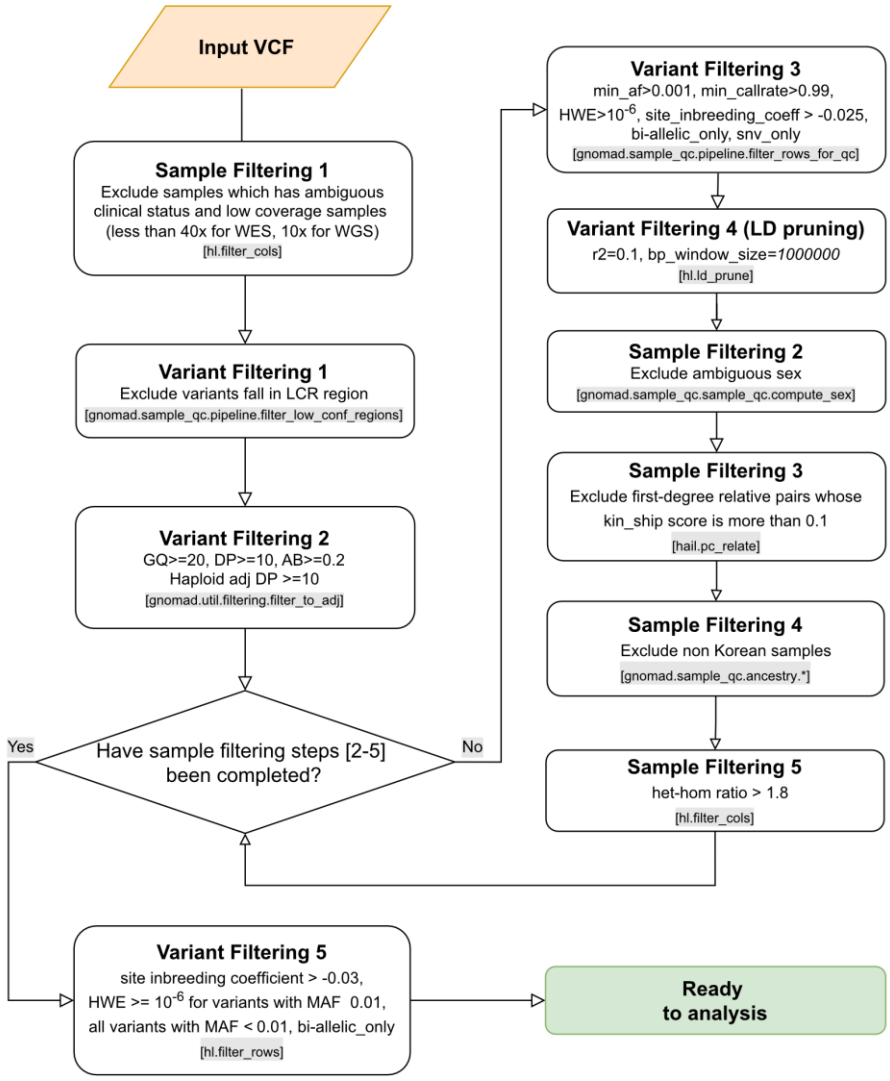

Figure 2.2. Quality control process. Input VCF is the output of the pipeline in Fig S1. Each block contains the Hail function in gray background.

2.2.3. 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 R2 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 (Figure 2.3, Table 2.3). 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.

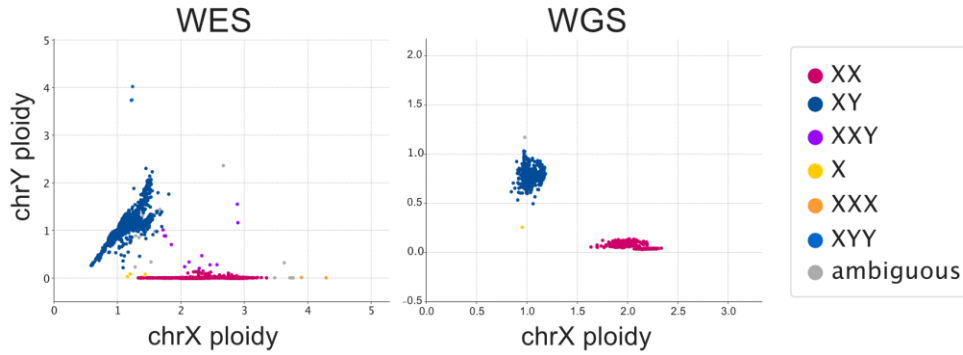

Figure 2.3. Sex inference. The X-axis and Y-axis represent the ploidy of chromosome X and Y, respectively, normalized by the coverage of chromosome 21.

Table 2.3. Number of samples based on inferred sex information

Inferred sex	WES	WGS	Total
Female	1627	943	2570
Male	1782	953	2735
Total	3409	1896	5305

2.2.4. 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 $\geq$ 0.001, call-rate > 0.99, HWE P > 1.0 x 10-8, inbreeding coefficient score > -0.025, and R2 for LD pruning < 0.1. After calculating kinship using the 'pc_relate' feature [27] in Hail, we selected the maximal independent set of samples with kinship < 0.1 using the 'maximal_independent_set [28] from Hail. For related sample pairs, we chose the one with a higher coverage depth.

2.2.5. Population structure analysis

All biallelic autosomal SNVs from our dataset and the 1000 Genomes Project Phase 3 (KG) [29] were merged and filtered; variants were retained if they had a MAF > 0.001, call-rate > 0.99, HWE P > 1.0 x 10-8, and inbreeding coefficient score > -0.025. We then pruned the variants to those with an LD R2 < 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.

2.2.6. Sample quality control

The overall process is summarized in Figure 2.2. First, we excluded samples with ambiguous clinical status or having a mean coverage depth < 40X and < 10X for WES and WGS, respectively. To remove unintentionally enrolled cancer samples, samples with likely pathogenic oncogene variants were removed. Among the cosmic mutations [30] identified in the oncogene defined by oncoKB [31], the variants used for removal satisfy the following conditions:

1) MUTATION_SIGNIFICANCE_TIER: 1 or 2
2) COSMIC_SAMPLE_MUTATED >= 100
3) KOVA_AC < 300

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 (Figure 2.4). Ti/Tv and Het/Hom scores were computed using the 'compute_sample_qc_metric' function implemented in Hail. 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.

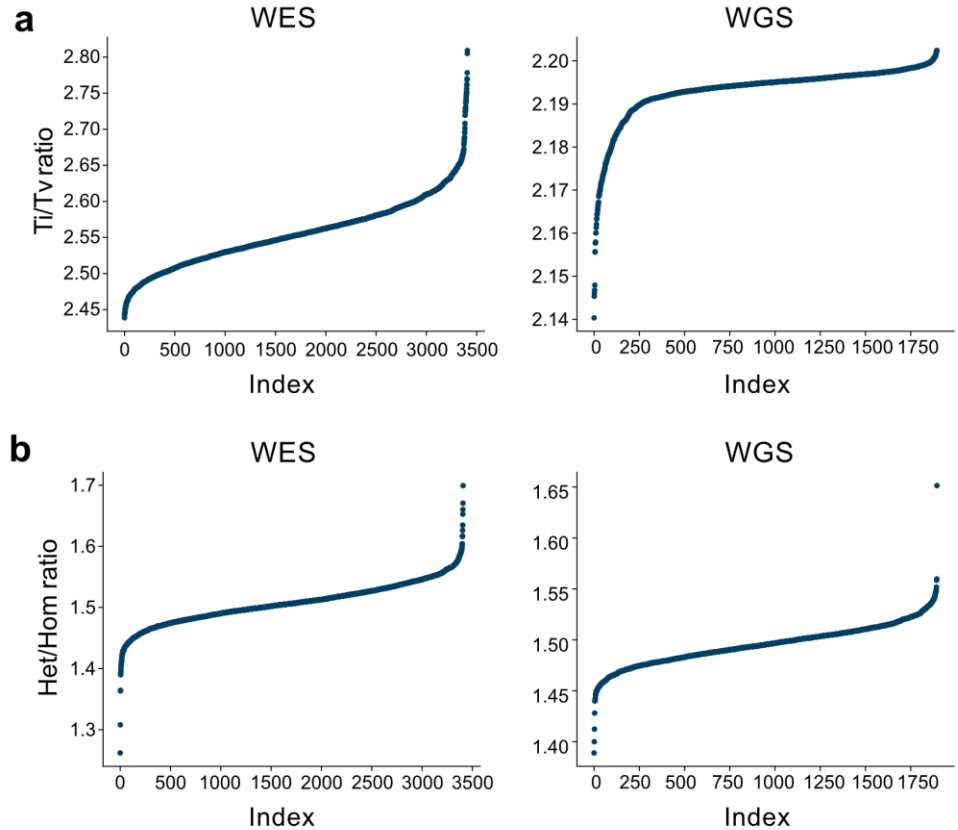

Figure 2.4. Quality control of KOVA 2 samples. (a) A transition/transversion (Ti/Tv) ratio value distribution for WES (left) and WGS (right) samples. (b) A heterozygous/homozygous ratio value distribution on WES (left) and on WGS (right) samples.

2.2.7. Variant quality control

The overall process is summarized in Figure 2.2. 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. This QC procedure differs slightly from the standard gnomAD filtering procedure [5] (Table 2.4). Functional annotation was performed by the Variants Effect Predictor (VEP) version 101 [32]. For each variant, we selected the most severe functional consequences using the gnomAD package of Hail.

Table 2.4. Comparison of variant QC steps between gnomAD and KOVA 2.

Filtering steps	gnomAD v2.1	KOVA 2
Variant Quality Score Recalibration (VQSR)	allele−specific VQSR using random forests classifier	VQSR using GATK
Low complexity region (LCR)	Flagged as lcr	Filtered
Segmental duplication	Flagged as segdup	−
Low quality genotype	GQ < 20, DP < 10, AB ≤ 0.2 for heterozygotes	GQ < 20, DP < 10, AB ≤ 0.2 for heterozygotes
Hardy−Weinberg equilibrium test	−	HWE < 10−6 for variants with MAF ≥ 0.01
Inbreeding coefficient (F)	F < −0.3	F < −0.3

2.2.8. Phasing

After carrying out sample-level and variant-level quality control, WGS data were phased with SHAPEIT4 version 4.2.2 [33]. 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 a missingness < 10% and MAF > 0.001. We used the genetic maps for reference version hg38 that are provided by SHAPEIT4 [34]. We also phased our data with Beagle 5.2 (beagle.21Apr 21.304.jar) [35], for which we used the hg38 genetic map available at the Beagle website [36] and the reference panel created by the KG Project.

2.2.9. Imputation of array data

Imputation of variants based on KOVA 2 was performed as previously described [9] and evaluation of imputation performance was done in collaboration with Jeongha Lee. 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 Impute2 [37]. Panel imputation accuracy was compared using the aggregated squared Pearson correlation coefficient (R2) determined between the imputed genotype dosages and the true genotypes from genome data.

2.2.10. Calling of structural variants (SVs)

Manta v1.6 [38] 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 v2.1 [39]. Slightly different SV representations across VCF files were merged using svimmer [40]. An SV was defined as known if it overlapped with any entry in the gnomAD SV v2.1 dataset.

2.3. Result

Korean-targeted WES and WGS data collected from various projects were used to build a Korean control database. Raw reads from 4,258 WES and 2,396 WGS data produced with different libraries were processed and filtered using an updated version of the pipeline than previous one [5,7,22]. The Methods section describes the exclusion criteria for samples and variants. After filtering out 1,349 samples (20.3% of initial samples), variants from the remaining 5,305 samples (3,409 WES and 1,896 WGS) were used in subsequent analyses (Table2.5, Table2.6). The remained samples consist of healthy volunteers (31.4%), healthy parents of rare disease patients (28.1%), or cancer patients' normal tissues (40.2%).

To assess the integrity of our calling pipeline, three distinct evaluations were performed. (1) A set of variants from a sample analyzed using 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 (Tables 2.7 and 2.8). (2) Additionally, our pipeline called 97.2% of Sanger-validated variants, and the missing calls were entirely caused by low-coverage regions (Table 2.9). (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 (Table 2.10). 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 individuals [41].

Two aspects were examined to ensure that KOVA2 does not contain confounding factors introduced by cancer normal tissue samples. First, whether more GWAS [4] signals associated with cancer were found in cancer normal tissues was determined (Table 2.11), and secondly, whether more cosmic [30] oncogene variants were found in cancer normal samples was determined (Table 2.12). Among 1673 GWAS cancer variants, 1108 mutations that did not overlap with Panel of Normal (PON) [42] publicly provided by GATK were first listed for analysis using GWAS. Among these 1108 GWAS cancer variants, 18 variants from WES and 160 variants from WGS overlapped with KOVA AF < 5% variants. Of these, three mutations from WGS were found more frequently in the cancer normal sample group (Fisher's exact test, $p < 0.05$). However, these three mutations are also commonly found in gnomAD (Table 2.11). In the same way, out of the 70,472 COSMIC oncogene variants that were obtained by choosing variants on oncogene defined by oncoKB from cosmic cancer mutation cencus v92, 70,151 mutations that did not overlap with PON were first listed for analysis using COSMIC. 753 WES and 613 WGS variants out of the 70,151 COSMIC oncogene variants overlapped with KOVA AF 5% variants. Of these, the cancer normal sample group

had higher rates of two WES mutations and seven WGS mutations ($p < 0.05$) (Table 2.12). These variants have the “Other” flag on the mutation significance tier which are assigned by COSMIC, indicating that they are likely to have a low impact. Similar results were obtained when comparing the number of cancer-related variants per sample (Figure 2.5, Figure 2.6)

Table 2.5. Sample quality control process of WES data.

Step	Condition	Samples before filtering	Samples with condition
1	Ambiguous clinical status	4,235	306
2	Low coverage depth (meanCoverage < 40)	3,929	77
3	Ambiguous sex	3,852	92
4	Duplicated (kin > 0.35)	3,760	164
5	Related (0.1< kin <=0.35)	3,596	22
6	Ambiguous ethnicity	3,574	91
7	Het/hom ratio outlier (ratio > 1.8)	3,483	22
8	In both WES and WGS	3,461	52
	Total	3,409	849

Table 2.6. Sample quality control process of WGS data

Step	Condition	Samples before filtering	Samples with condition
1	Low coverage depth (meanCoverage < 10)	2,396	165
2	Ambiguous sex	2,231	10
3	Duplicated (kin > 0.35)	2,221	144
4	Related (0.1< kin <=0.35)	2,077	149
5	Ambiguous ethnicity	1,928	32
	Total	1,896	500

Table 2.7. The number of concordant or discordant variant pairs between PacBio and KOVA pipeline from a single sample. 0/0, homozygous reference; 0/1, heterozygous variant; 1/1, homozygous variant;

	Genotype	KOVA2		
		0/0	0/1	1/1
PacBio	0/0	-	3,820	610
	0/1	11,411	1,974,392	675
	1/1	742	3,613	1,548,512

Table 2.8. The number of concordant or discordant variant pairs between NovaSeq and KOVA pipeline from a single sample. 0/0, homozygous reference; 0/1, heterozygous variant; 1/1, homozygous variant;

	Genotype	KOVA2		
		0/0	0/1	1/1
NovaSeq	0/0	–	–	–
	0/1	5,562	2,064,129	1,215
	1/1	56	1,028	1,495,016

Table 2.9. Comparison of Sanger-validate calls and KOVA 2 calls. 0/0, homozygous reference; 0/1, heterozygous variant

No.	KOVA2 Sample ID	Chr:position(hg38)	Sanger result	KOVA2 call			Concordant?
				Call	Ref. coverage	Nonref. coverage	
1	KVE0617	chr4:15059272	0/0	0/0	65	0	Yes
2	KVE0632	chr8:60743012	0/0	0/0	17	0	Yes
3	KVE0633	chr8:60743012	0/0	0/0	16	0	Yes
4	KVE0853	chr4:15059272	0/0	0/0	84	0	Yes
5	KVE0909	chr10:72551287	0/0	0/0	39	0	Yes
6	KVE2741	chr6:75087652	0/0	0/0	39	0	Yes
7	KVE2758	Chr9:137162182	0/0	0/0	28	0	Yes
8	KVE2759	Chr9:137162182	0/0	0/0	38	0	Yes
9	KVE2778	chr16:48361909	0/0	0/0	152	0	Yes
10	KVE2779	chr16:48361909	0/0	0/0	225	0	Yes
11	KVE2782	chr22:27751027	0/0	0/0	46	0	Yes
12	KVE2783	chr22:27751027	0/0	0/0	49	0	Yes
13	KVE2785	chr22:23787200	0/0	0/0	128	0	Yes

14	KVE2786	chr22:23787200	0/0	0/0	143	0	Yes
15	KVE2787	chr6:157184324	0/0	0/0	190	0	Yes
16	KVE2788	chr6:157184324	0/0	0/0	177	0	Yes
17	KVE2791	chr16:56354885	0/0	0/0	236	1	Yes
18	KVE2792	chr16:56354885	0/0	0/0	160	0	Yes
19	KVE2797	chr17:63964667	0/0	0/0	13	0	Yes
20	KVE2798	chr17:63964667	0/0	0/0	9	0	Yes
21	KVE2799	chrX:53382505	0/0	0/0	128	0	Yes
22	KVE2800	chrX:53382505	0/0	0/0	63	0	Yes
23	KVE2807	chr12:45837577	0/0	0/0	242	0	Yes
24	KVE2808	chr12:45837577	0/0	0/0	223	2	Yes
25	KVE2809	chr16:67539861	0/0	0/0	37	1	Yes
26	KVE2810	chr16:67539861	0/0	0/0	41	0	Yes
27	KVE2811	chr12:32733792	0/0	0/0	41	0	Yes
28	KVE2812	chr12:32733792	0/0	0/0	44	0	Yes
29	KVE2816	chrX:71564608	0/0	0/0	40	0	Yes
30	KVE2822	chr14:101980380	0/0	0/0	25	0	Yes

31	KVE2823	chr14:101980380	0/0	0/0	16	0	Yes
32	KVE2840	chr12:49033931	0/0	0/0	43	0	Yes
33	KVE2841	chr12:49033931	0/0	0/0	43	0	Yes
34	KVE3585	chr2:86252036	0/0	0/0	77	0	Yes
35	KVE3586	chr3:155084298	0/0	0/0	35	0	Yes
36	KVE3640	chr13:110176904	0/0	0/0	86	0	Yes
37	KVE3641	chr13:110176904	0/0	0/0	100	0	Yes
38	KVE3645	chr9:2081979	0/0	0/0	34	0	Yes
39	KVE3646	chr9:2081979	0/0	0/0	37	0	Yes
40	KVE3780	chr19:50323104	0/0	0/0	18	0	Yes
41	KVE3805	chr18:33740159	0/0	0/0	36	0	Yes
42	KVE3812	chrX:115165459	0/0	0/0	8	0	Yes
43	KVE3825	chr6:33451838	0/0	0/0	25	0	Yes
44	KVE3826	chr6:33451838	0/0	0/0	16	0	Yes
45	KVE3835	chrX:53234559	0/0	0/0	94	0	Yes
46	KVE3836	chrX:53234559	0/0	0/0	50	1	Yes
47	KVE3837	chr1:181651440	0/0	0/0	105	0	Yes

48	KVE3838	chr1:181651440	0/0	0/0	68	0	Yes
49	KVE4140	chr22:50675152	0/0	0/0	13	0	Yes
50	KVE4141	chr22:50675152	0/0	0/0	6	0	Yes
51	KVE4142	chr1:27552049	0/0	0/0	11	0	Yes
52	KVE4143	chr1:27552049	0/0	0/0	20	0	Yes
53	KVE4144	chr11:118758879	0/0	0/0	34	0	Yes
54	KVE4145	chr11:118758879	0/0	0/0	38	0	Yes
55	KVE0634	chr3:33018452	0/1	0/1	31	23	Yes
56	KVE0635	chr3:33068263	0/1	0/1	44	39	Yes
57	KVE0749	chr3:33018452	0/1	0/1	39	29	Yes
58	KVE0750	chr3:33018452	0/1	0/1	31	27	Yes
59	KVE0908	chr10:72551287	0/1	0/1	84	50	Yes
60	KVE2742	chr6:75087652	0/1	0/1	21	16	Yes
61	KVE2789	chr1:180274413	0/1	0/1	71	61	Yes
62	KVE2790	chr1:180274571	0/1	0/1	37	36	Yes
63	KVE2815(F)	chrX:71564608	0/1	0/1	53	50	Yes
64	KVE2836	chr10:133364730	0/1	0/1	51	67	Yes

65	KVE2837	chr10:133373332	0/1	0/1	5	12	Yes
66	KVE3587	chr3:155084298	0/1	0/1	28	27	Yes
67	KVE3638	chr19:55137102	0/1	0/1	36	30	Yes
68	KVE3639	chr19:55134092	0/1	0/1	29	25	Yes
69	KVE3806	chr18:33740159	0/1	0/1	47	53	Yes
70	KVE3779	chr19:50323104	0/1	No call	1	8	No
71	KVE3811(F)	chrX:115165459	0/1	No call	1	4	No

Table 2.10. The number of variants covered by WES, WGS only or by both methods.

MAF	Number of variants			Concordance
	Both	WES only	WGS only	
0.05−0.1	7,041	55	911	87.9%
0.1−0.2	8,200	60	1,090	87.7%
0.2−0.3	5,335	34	662	88.5%
0.3−0.4	4,147	30	557	87.6%
0.4−0.5	3,323	29	389	88.8%
0.5−0.6	2,949	23	329	89.3%
0.6−0.7	2,425	20	266	89.5%
0.7−0.8	2,063	8	233	89.5%
0.8−0.9	2,054	12	216	90.0%
0.9−1.0	2,657	24	271	90.0%
Total	40,194	295	4,924	88.5%

Table 2.11. GWAS cancer signals enriched in samples from cancer normal tissue

Locus	Base change	CN		HP		HV		gnomAD		CN vs HP	CN vs HV	HV vs HP
		AC	AF	AC	AF	AC	AF	AC	AF	p-value	p-value	p-value
chr11:69513996	C>T	7	0.011	1	0.022	9	0.003	20500	0.135	0.427	0.013	0.141
chr11:69516215	G>A	7	0.011	1	0.022	8	0.003	18094	0.119	0.432	0.007	0.125
chr11:69516650	C>T	6	0.010	1	0.022	7	0.002	7953	0.052	0.393	0.014	0.113

Table 2.12. COSMIC oncogene variants enriched in samples from cancer normal tissue

Locus	Base change	Gene	COSMIC		CN		HP		HV		gnomAD		CN vs HP	CN vs HV	HV vs HP
			AC	AF	AC	AF	AC	AF	AC	AF	AC	AF	p-value	p-value	p-value
chr1:11212411*	C>G	MTOR	5	8E-05	146	0.039	81	0.029	7	0.030	1706	0.011	0.029	0.600	0.840
chr1:39327259*	A>G	MACF1	3	8E-05	21	0.006	5	0.002	0	0.000	24	0.000	0.016	0.632	1.000
chr7:106905140	G>A	PIK3CG	3	6E-05	2	0.003	0	0.000	0	0.000	9	0.000	1.000	0.028	NaN
chr7:116699097	G>A	MET	1	2E-05	3	0.005	0	0.000	1	0.000	5	0.000	1.000	0.018	1.000
chr9:134731639	G>A	COL5A1	4	1E-04	4	0.006	0	0.000	0	0.000	3	0.000	1.000	0.001	NaN
chr19:10500009	C>T	KEAP1	1	3E-05	9	0.015	0	0.000	16	0.006	35	0.000	1.000	0.030	1.000
chr19:18168762	G>A	PIK3R2	1	3E-05	3	0.010	0	0.000	2	0.001	11	0.000	1.000	0.007	1.000
chr19:33301725	C>A	CEBPA	1	3E-05	5	0.018	1	0.250	5	0.003	23314	0.158	0.084	0.005	0.013
chr20:18316456	C>T	ZNF133	1	3E-05	2	0.003	0	0.000	0	0.000	3	0.000	1.000	0.031	NaN

Variants with * calculated from WES, while others from WGS

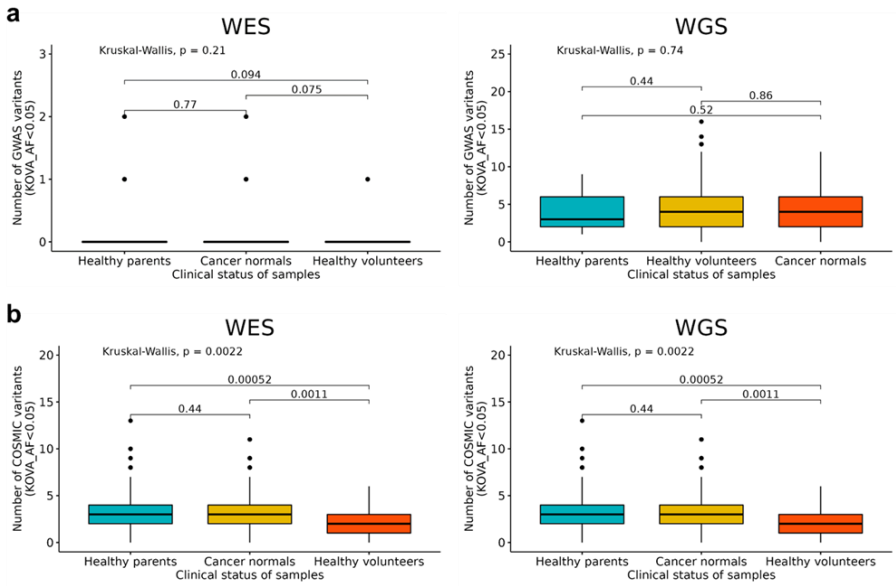

Figure 2.5. Enrichment test of cancer-related variants with KOVA AF < 0.05 on cancer normal samples. (a) comparison of allele frequency for GWAS cancer variants. (b) comparison of allele frequency for COSMIC oncogene variants.

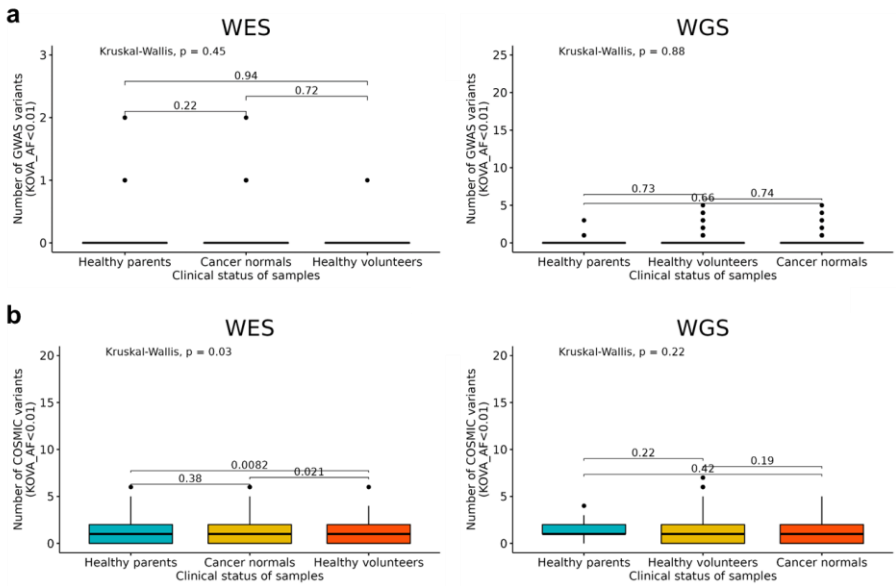

Figure 2.6. Enrichment test of cancer-related variants with KOVA AF < 0.01 on cancer normal samples. (a) comparison of allele frequency for GWAS cancer variants (b) comparison of allele frequency for COSMIC oncogene variants.

PCA analysis of WES and WGS data separately revealed no batch effect in either data, but quite a few outlier samples were discovered when only WGS data was examined (Figure2.7). However, neither batch effects nor outlier samples were discovered when PCA analysis was carried out once again on the data produced by integrating WES and WGS data (Figure 2.8). PCA found KOVA 2 samples in a cluster distinct from samples from Japanese, northern Chinese, southern Chinese, and Southeast Asian individuals (Figure2.9, Figure2.10), reflecting different genetic characteristics among each race in East Asian countries. Also, this results indicate that the QC works were done properly.

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 (Figure2.11, Figure2.12).

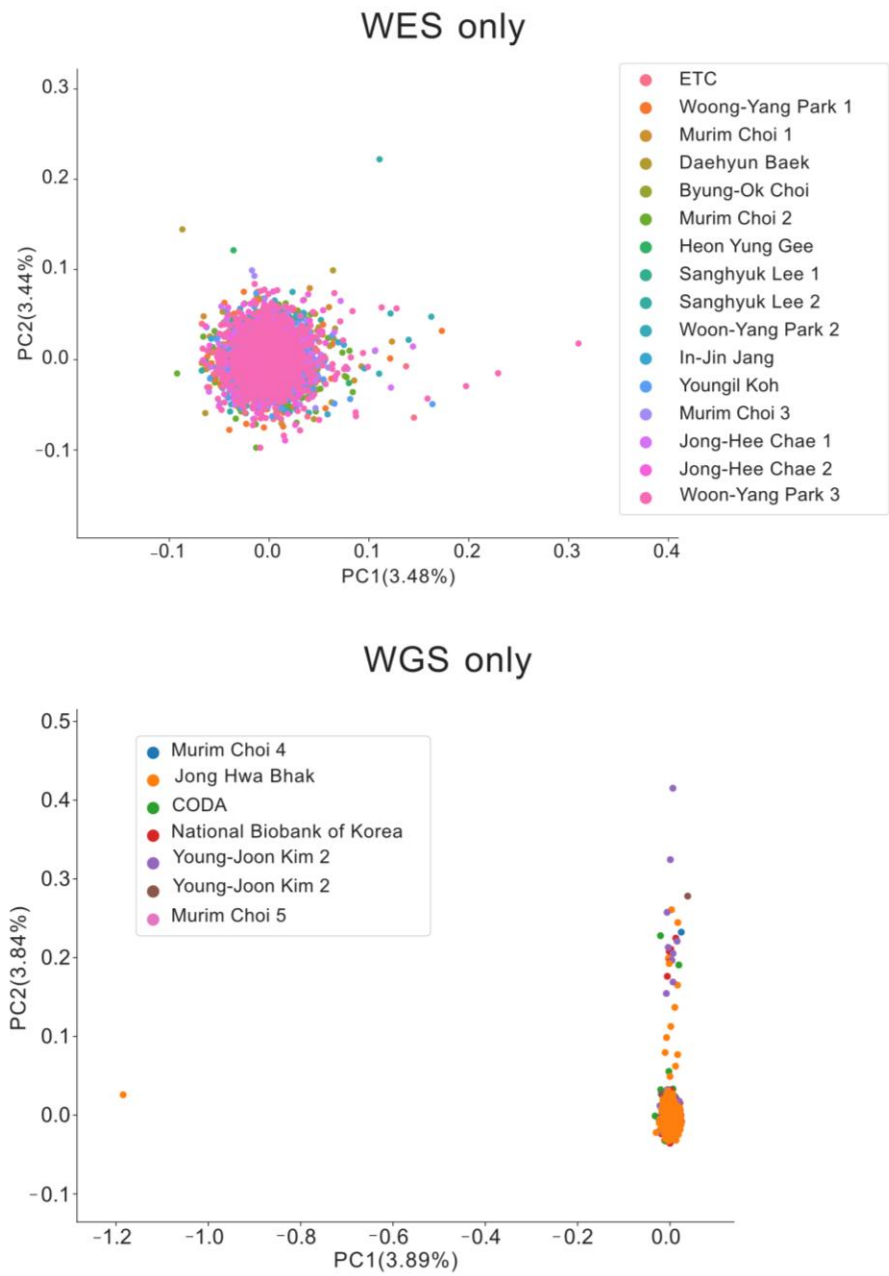

Figure 2.7. PCA of KOVA 2 WES(top), WGS samples (bottom). Legend shows batch information of samples.

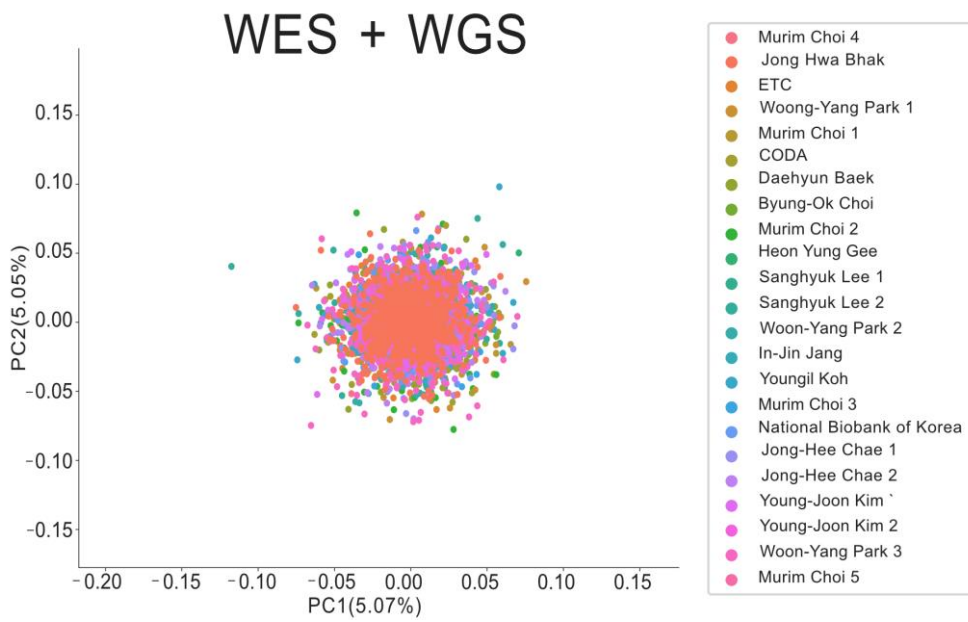

Figure 2.8. PCA of KOVA 2 WES-WGS combined data. Legend shows batch information of samples.

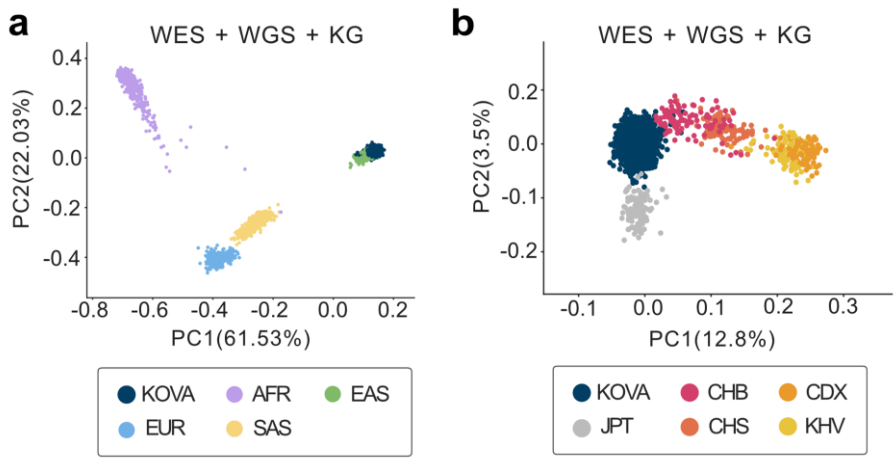

Figure 2.9. PCA of KOVA 2 with KG individuals. (a) PCA of KOVA 2 and all population included in KG project. AFR African EAS East Asian EUR European SAS South Asian (b) PCA of KOVA 2 and the neighboring East Asian populations. CHB Han Chinese individuals from Beijing, CDX Chinese individuals Dai from Xishuangbanna JPT Japanese individuals from Tokyo, CHS Han Chinese individuals from South China KHV Kinh individuals from Ho Chi Minh City.

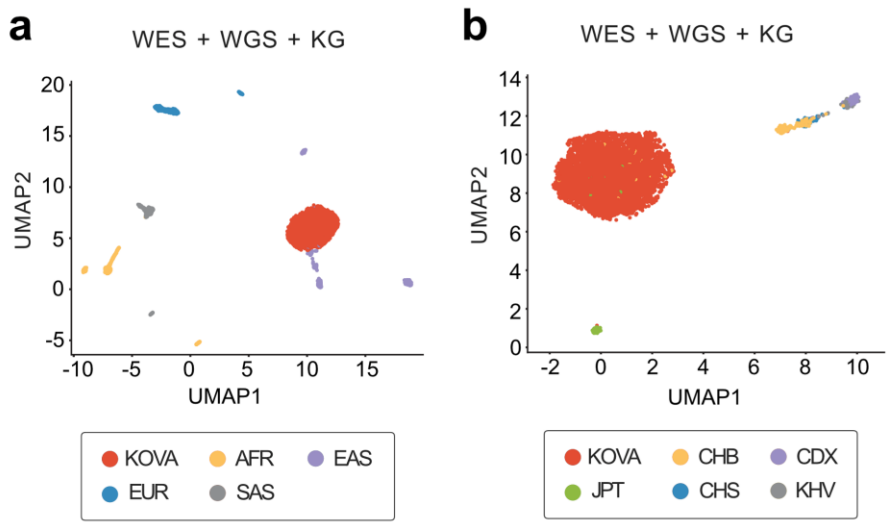

Figure 2.10. UMAP of KOVA 2 with KG individuals. (a) UMAP of KOVA 2 and all population included in KG project. (b) PCA of KOVA 2 and the neighboring East Asian populations. The labels are the same as those in Figure 2.7.

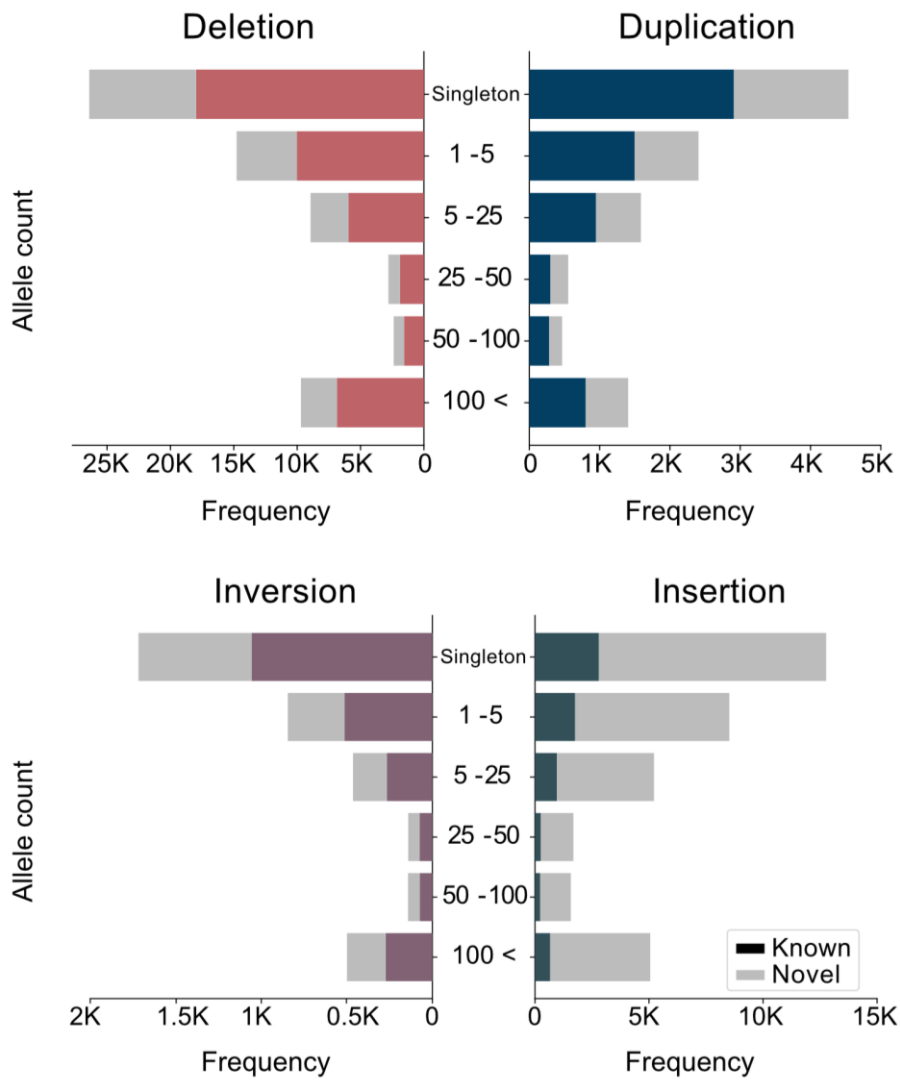

Figure 2.11. Number of SV variants by allele counts, divided by SV type, and known (dark-colored) and novel (gray-colored) variants according to gnomAD SV database.

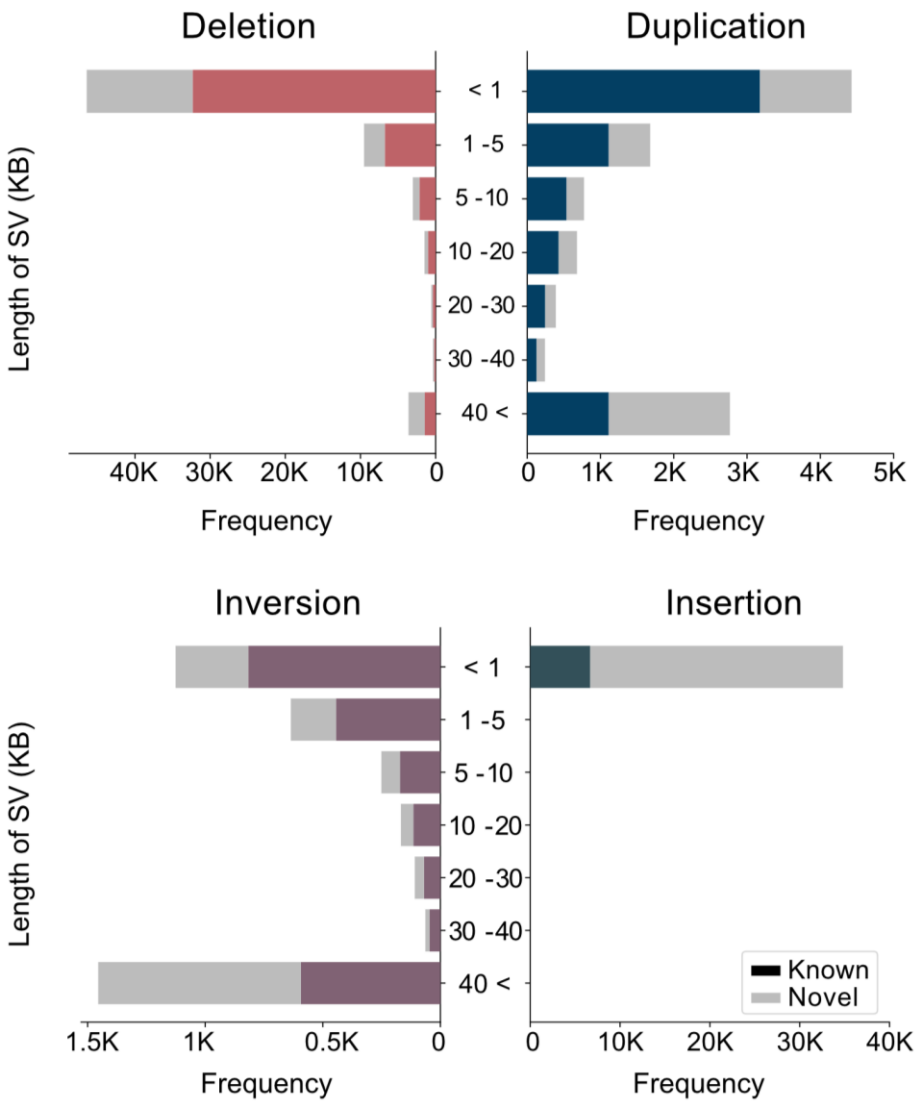

Figure 2.12. Number of SV variants by the length of SV in kilobases, divided by SV type, and known (dark-colored) and novel (gray-colored) variants according to gnomAD SV database.

When the minor allele frequency (MAF) distribution was examined, it was discovered that there was a high enrichment of rare variants (<1%), including a greater percentage of novel variants that weren't present in the control gnomAD v3.11 [5] database (Figure 2.13, Figure 2.14, Figure 2.15, Table 2.13, Table 2.14). In contrast to prevalent variations (>5% frequency in gnomAD v3.1), which were rapidly saturated at <500 samples, adding data from more Korean individuals was insufficient to saturate newly identified variants, whether in coding or noncoding regions (Figure 2.16). Interestingly, while KOVA 2 common variants (>5% frequency) were saturated before adding 100 samples, gnomAD common non-coding variants continued to show a slightly increasing trend even after analyzing 1,800 WGS samples. This finding suggests that in order to completely cover this set of variants, a higher sample size is required (Figure 2.17). As anticipated, variant function indicators such as the nonsilent/silent (NS/S) ratio, CADD [40], ReMM [41], FunSeq2 [42], and LINSIGHT [43] all exhibited increasing functionality as MAF decreased (Figure 2.18, Figure 2.19). Finally, the distribution of variants in the proximal intron regions indicated strong selection against any base change as variants approach exon-intron borders (Figure 2.20).

Imputing variants Based on KOVA 2 had the best coverage of common variants in Koreans as well as rare variants, leading to superior performance to other reference panels (Figure 2.21).

These outcomes show the high quality of the KOVA 2 variant set.

This KOVA2 variant set was released in public website (https://www.kobic.re.kr/kova), so users can search the interested genes or loci in a table format or in a genome browser format. Also, individuals can download the variant set from the browser with a minimum registration process (Figure 2.22).

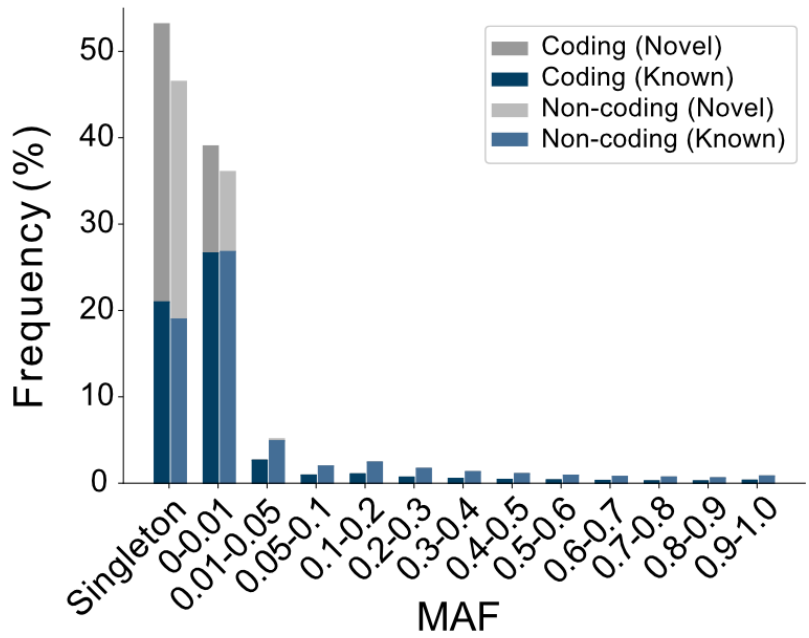

Figure 2.13. Variant frequency by MAF according to variant status: coding and noncoding, known and novel.

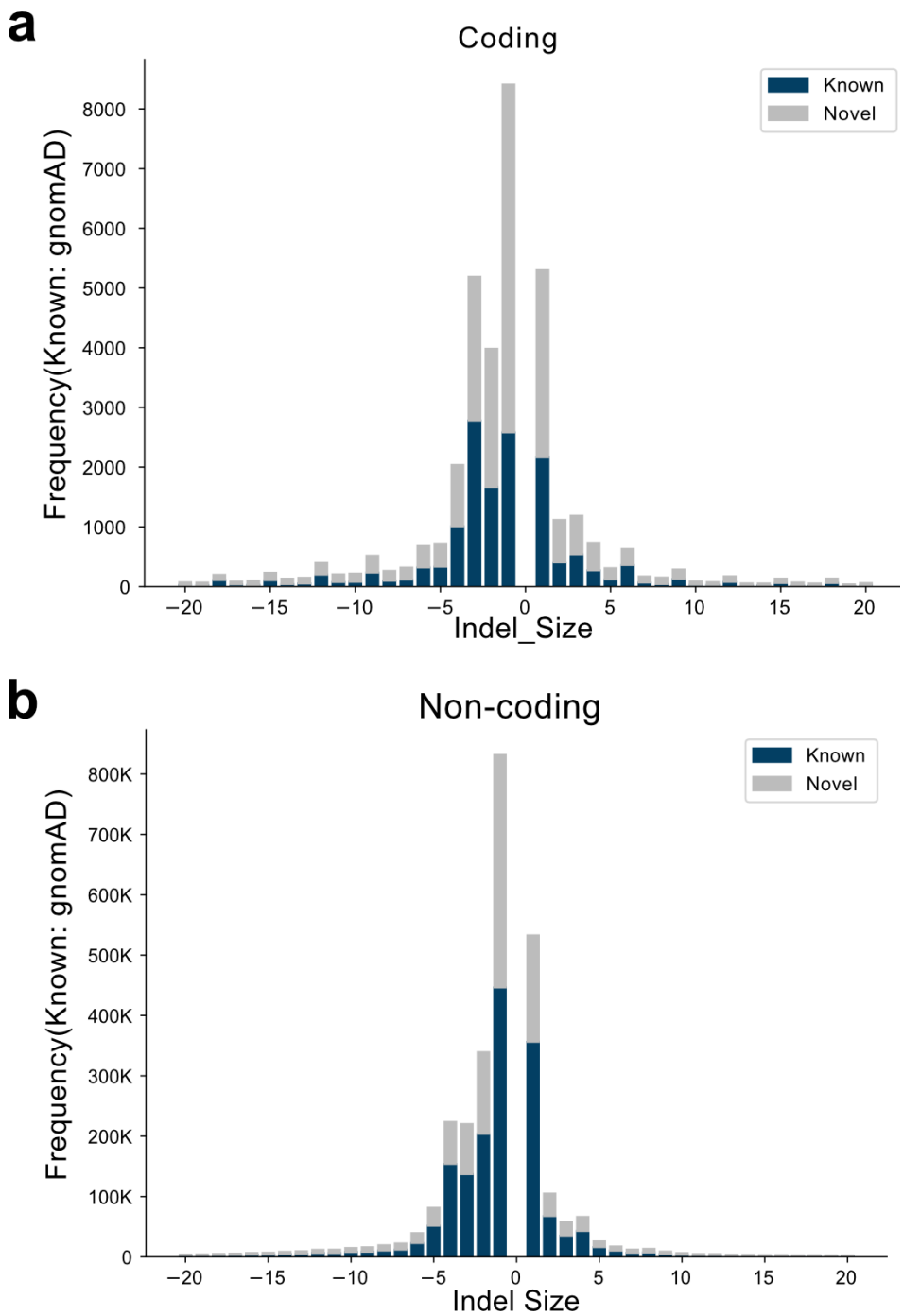

Figure 2.14. The distributions of indel sizes in (a) coding and (b) non-coding regions. The frequency of known variants is in dark blue.

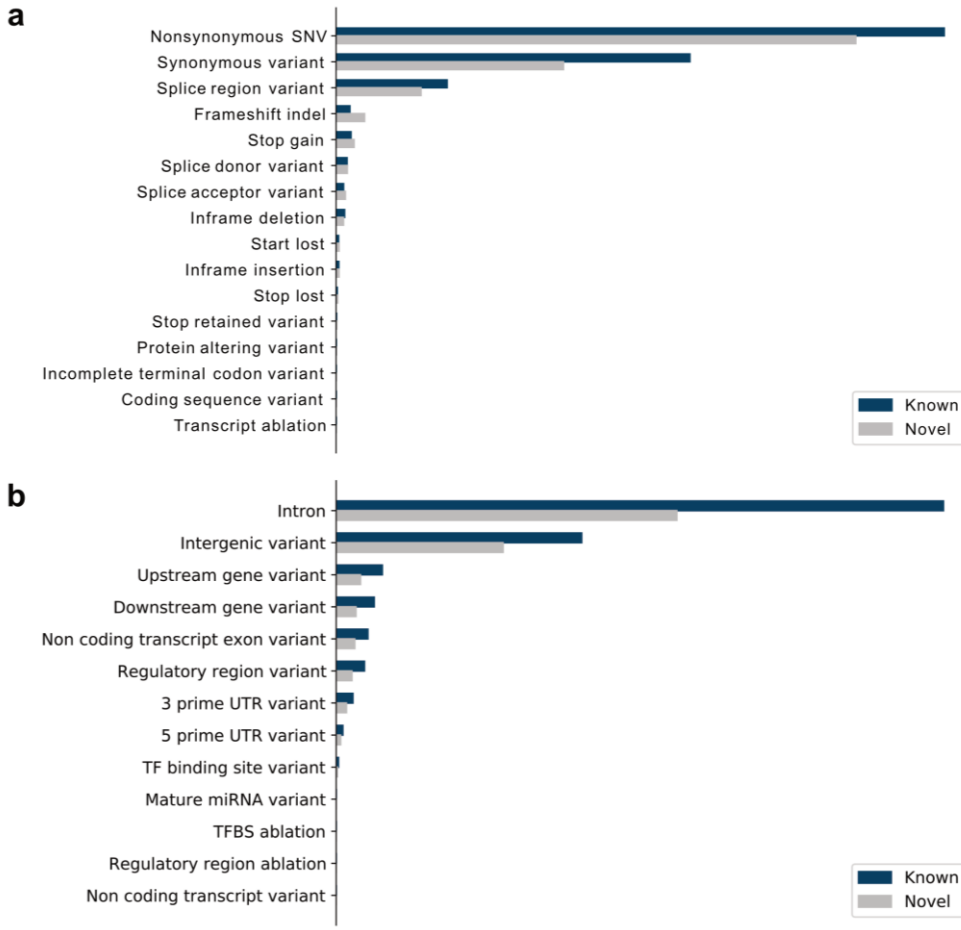

Figure 2.15. The number of variants by annotated function on (a) coding and (b) non-coding regions. The frequency of known variants is in dark blue.

Table 2.13. Coding variant counts by functional class

Variant function	Number of Singletons	Number of non-singletons	Total counts	Number of known variants	Number of novel variants
Transcript ablation	1	3	4	3	1
Coding sequence variant	14	16	30	22	8
Incomplete terminal codon variant	15	19	34	13	21
Protein altering variant	130	31	161	43	118
Stop retained variant	237	215	452	203	249
Stop lost	616	567	1,183	502	681
Inframe insertion	1,421	1,073	2,494	1,206	1,288
Start lost	1,389	1,017	2,406	1,106	1,300
Inframe deletion	4,043	3,021	7,064	3,780	3,284
Splice acceptor variant	4,227	3,212	7,439	3,317	4,122
Splice donor variant	5,243	4,618	9,861	4,887	4,974
Stop gain	9,246	5,399	14,645	6,647	7,998
Frameshift indel	12,239	6,583	18,822	6,157	12,665
Splice region variant	43,437	43,699	87,136	49,383	37,753
Synonymous variant	127,844	130,603	258,447	157,345	101,102
Nonsynonymous SNV	274,693	226,818	501,511	270,422	231,089
Total	**484,795**	**426,894**	**911,689**	**505,036**	**406,653**

Table 2.14. Noncoding variant counts by functional class

Variant function	Number of Singletons	Number of non-singletons	Total counts	Number of known variants	Number of novel variants
Non coding transcript variant	-	1	1	1	-
Regulatory region ablation	1	-	1	1	-
TFBS ablation	112	121	233	143	90
Mature miRNA variant	564	563	1,127	620	507
TF binding site variant	41,715	48,332	90,047	58,693	31,354
5 prime UTR variant	141,718	145,808	287,526	172,423	115,103
3 prime UTR variant	341,553	361,410	702,963	436,082	266,881
Regulatory region variant	529,226	619,532	1,148,758	737,657	411,101
Non coding transcript exon variant	615,209	696,435	1,311,644	826,591	485,053
Downstream gene variant	690,767	820,003	1,510,770	993,375	517,395
Upstream gene variant	843,801	993,065	1,836,866	1,203,204	633,662
Intergenic variant	4,929,501	5,832,094	10,761,595	6,406,273	4,355,322
Intron	11,580,871	13,158,563	24,739,434	15,849,148	8,890,286
Total	**19,715,038**	**22,675,927**	**42,390,965**	**26,684,211**	**15,706,754**

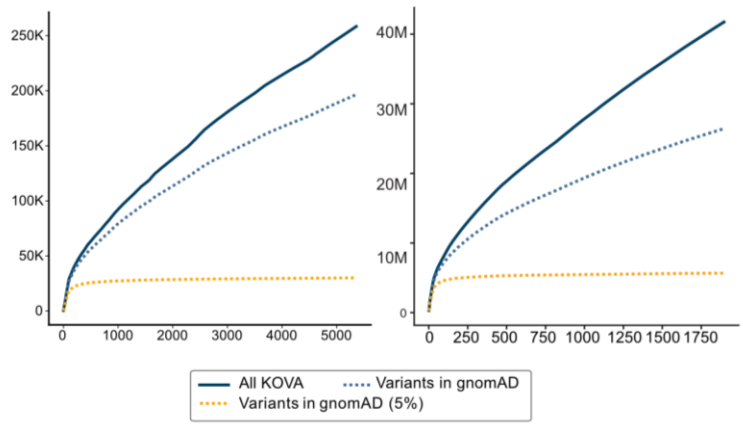

Figure 2.16. The number of variants identified according to increment the KOVA 2 samples. The number of variants identified is divided by the coding (left) and noncoding (right) status

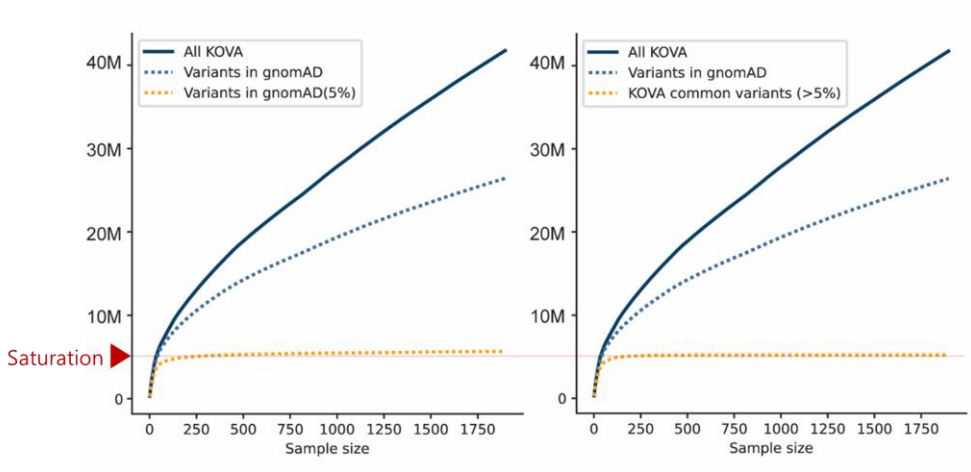

Figure 2.17. Saturation on the proportion of common non-coding variants with KOVA 2 AF > 5%.

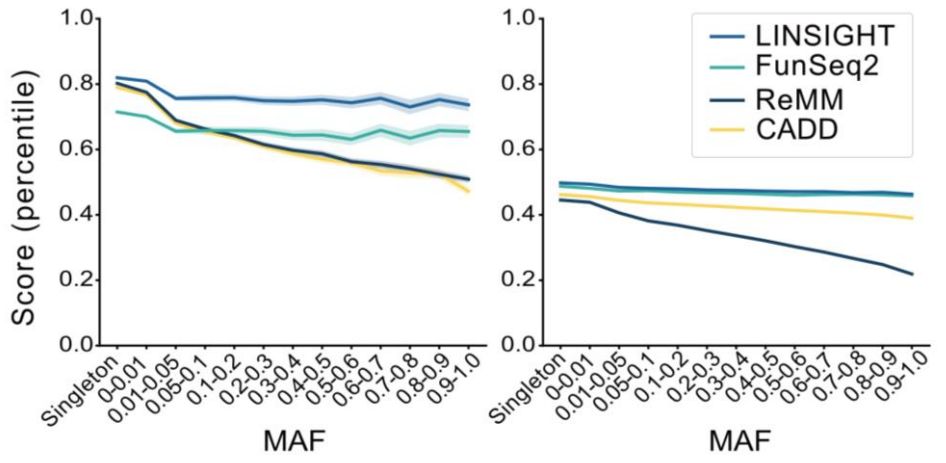

Figure 2.18. The patterns of variant functionality predicted by different software according to MAF are divided by the coding (left) and noncoding (right) status. As each program produces scores with different scales, and each scoring system was converted to percentiles.

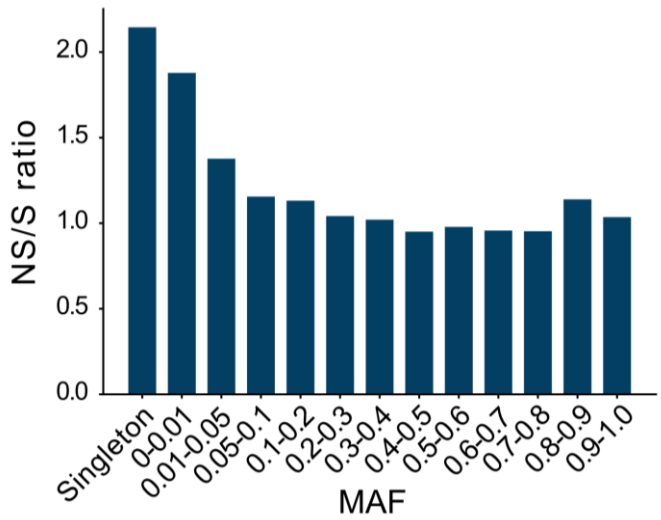

Figure 2.19. The nonsilent/silent (NS/S) ratio of coding variants by MAF

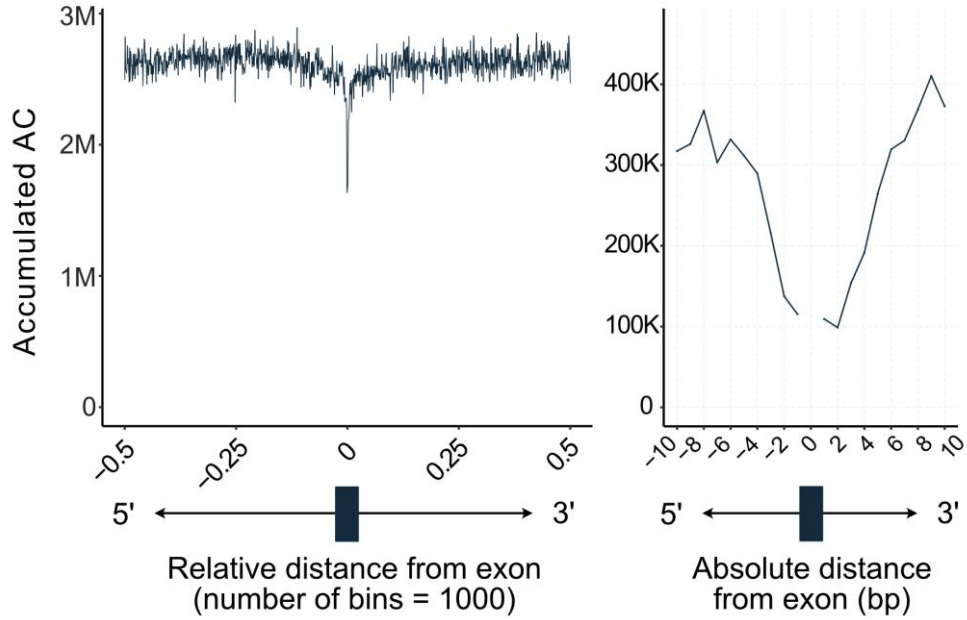

Figure 2.20. Intron variant burden according to the relative (left) or absolute (right) distance from exons.

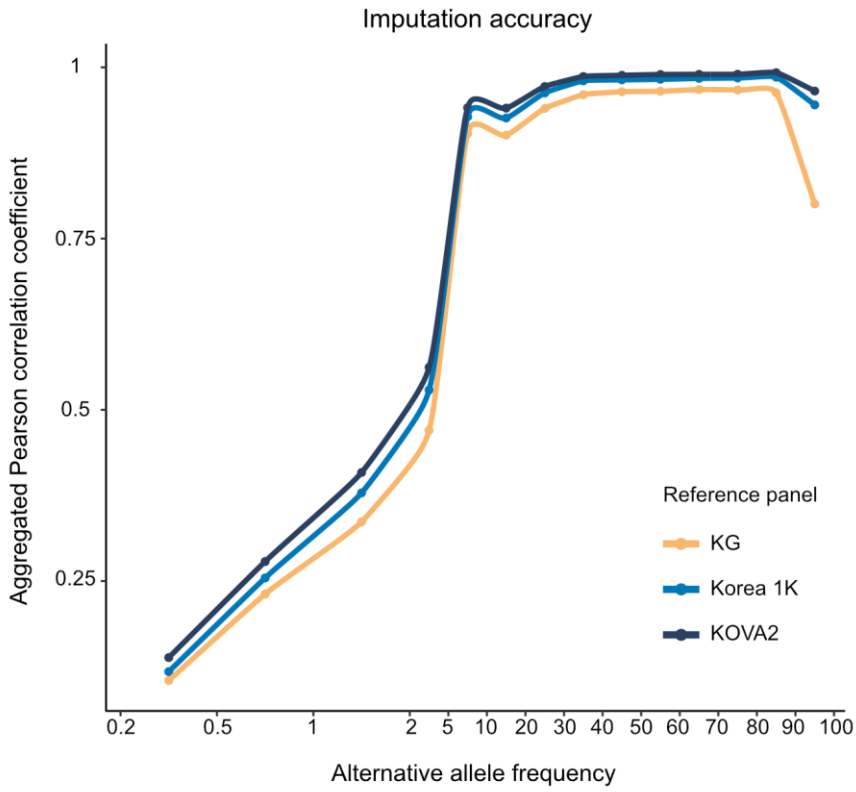

Figure 2.21. Imputation performance of KOVA 2 reference panel. The aggregated Pearson correlation coefficient (R2) between known genotypes from WGS data and imputed genotypes by the percentage of stratified alternative allele frequency. Imputation was performed in collaboration with Jeongha Lee.

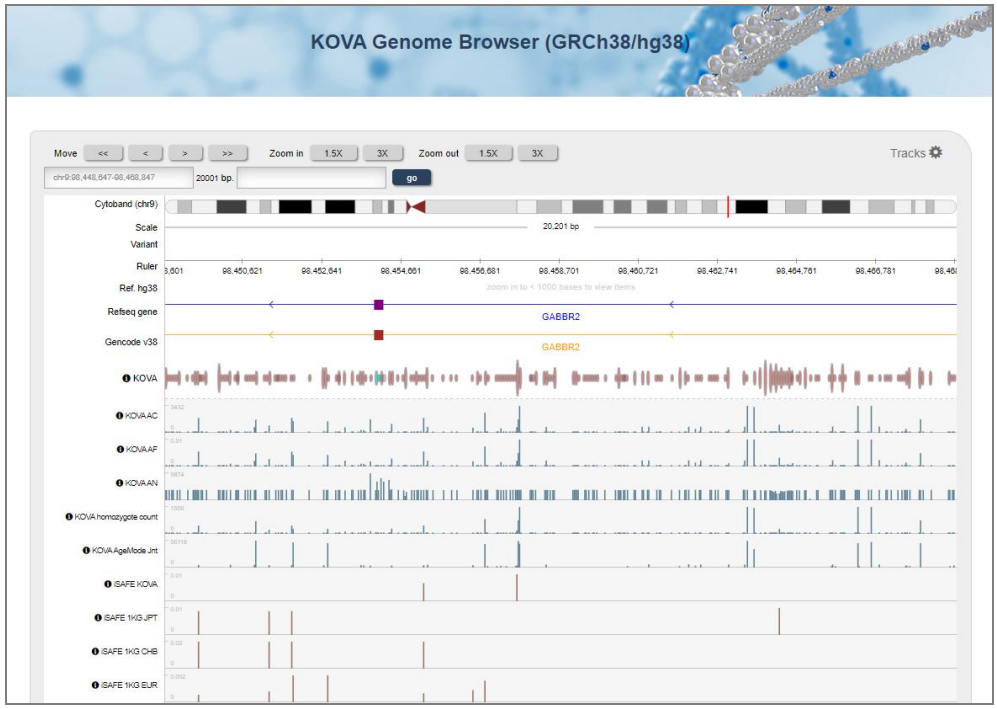

Figure 2.22. Genome browser for KOVA 2. Screenshot of result page of *GABBR2* gene search from https://www.kobic.re.kr/kova

Chapter 3. Genetic and Clinical Implications of Korean Genome

3.1. Background

As introduced in Chapter 1, Koreans have genetic similarities with the Japanese and Chinese, the neighboring races, but there are distinct genetic characteristics. These characteristics may also differ in the composition of haplotypes. Haplotype refers to a set of alleles that are likely to be inherited together when genetic information is passed on from single parent to offspring (Figure 3.1a, b). When a specific variant has an evolutionary advantage, a haplotype covering that variant prevalently exists in the population by positive selection, leading to selective sweep [47, 48]. Therefore, there are differences in the composition of haplotypes by race (Figure 3.1c).

Using the type and frequency of haplotypes from the population genome database, it is possible to infer which regions have strong positive selection pressure. Akbari *et al.* [47] devised a haplotype allele frequency score using the frequency information of variants present in the haplotype, and then tried to find evolutionary favored mutations based on this.

On the other hand, recombination and mutation events

accumulates on a haplotype over generations. If the number of mutations present in a particular haplotype is large or the length of the haplotype is short, we could assume that the haplotype has been passed down over several generations. Albers *et al.* [49] developed an allele age estimation model with gamma distribution using mutation count and haplotype length as parameters.

Understanding favored mutations during the selective sweep process is helpful in figuring out where selection came from or how disease develops. It can be particularly useful in identifying the causes of racial disparities in particular environments and temperaments [50]. In the Hemoglobin-B gene (*HBB*), for instance, the sickle cell mutation was the target of selection for malaria resistance, which explains why diseases affecting red blood cells are prevalent in regions with high rates of malaria [51]. Another compelling example is that human populations living at high altitudes adapted to low oxygen levels most likely through genetic adaptation mediated by genes in the Hypoxia Inducible Factor (HIF) pathway [52]. Likewise, identifying favored mutations during selective sweep can elucidate candidate genetic loci or regions for adaptive traits. Furthermore, because favored mutations tend to have functional impact on the phenotype, positively selected variants can shed light on disease physiology [53].

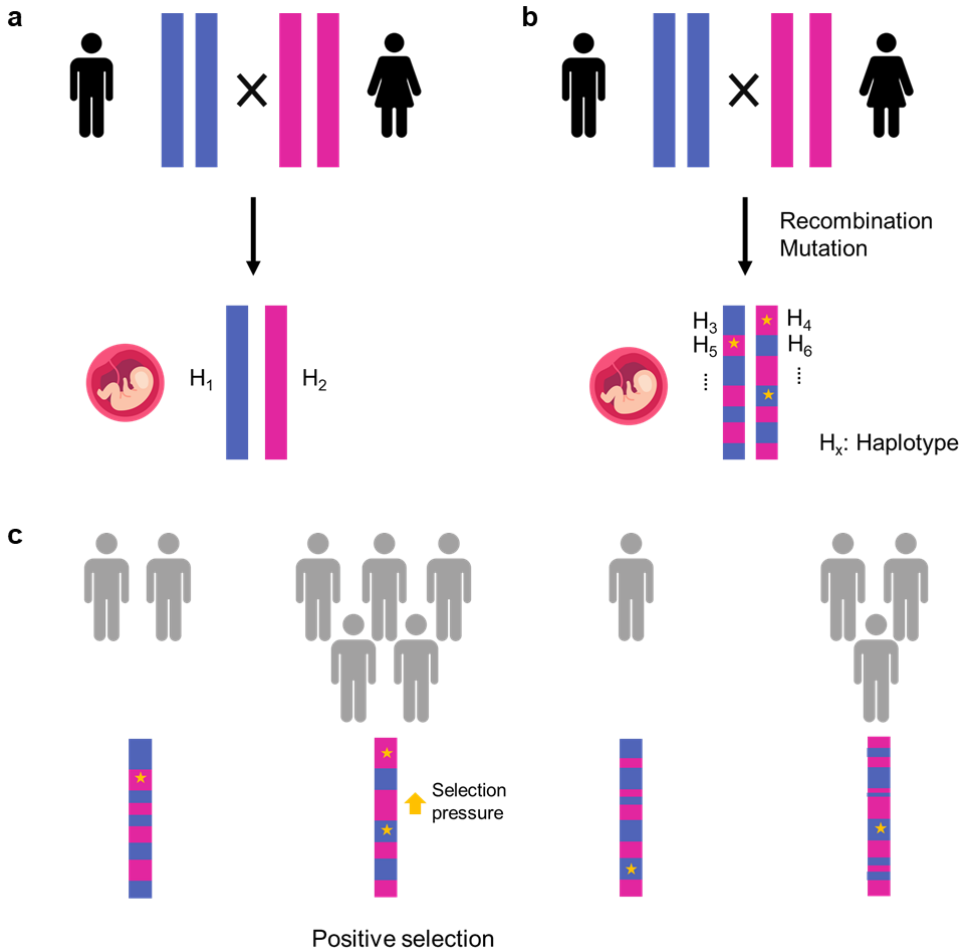

Figure 3.1. Introduction to haplotype and positive selection. Composition of haplotype a) without events of recombination and mutation, b) with the events. H_x denotes type of haplotype. c. population increase of positively selected haplotype .

3.2. Materials and Methods

3.2.1. Cohorts and sample preparation

In Chapter 3, analysis was conducted on 1,896 WGS data that had been phased using SHAPEIT described in Chapter 2. KG data [29] was used when phased data from other races was needed to identify Korean-specific genetic traits. GEM-J WGA panel [19] was used for Japanese allele frequency, and gnomAD data [5] was used for allele frequency of other races.

3.2.2. Runs of homozygosity (ROH)

PLINK v1.90b6.12 [54,55] 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 FROH scores, representing inbreeding levels, using the 'Froh_inbreeding' function of detectRuns package version 0.9.6 [56].

3.2.3. Regions of positive selection

Selected variants in positive selection sweeps were captured from phased KOVA 2 and KG data using iSAFE v1.0.7 [47] 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. This work was done in collaboration with Jean Lee.

3.2.4. Effective population size estimation

To estimate the historical effective population size, we used IBDNe software [57] according to the recommended protocol. Briefly, after detecting IBD segments with hap-IBD.jar [58], we refined them through removal of any breaks and short gaps from the segments using merge-ibd-segments.17Jan20.102.jar [59]. Finally, we used ibdne.23Apr20.ae9.jar [57] with default options to estimate the effective population size from the refined IBD segments.

3.2.5. Allele ages

Genealogical Estimation of Variant Age (GEVA) version v1beta [49] with default 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. From the output, AgeMode were used for further analysis. To compare allele ages as estimated by our data with those estimated from the KG data, we downloaded the Atlas of Variant Age from the developer's website [60]. Chimpanzee variants called from 25 individuals were downloaded from the Great Ape Genome Project [61].

3.3. Result

Homozygous pathogenic mutations in ROH are uncommon in an outbred community, such as the Korean population, compared to populations with a larger burden of consanguinity [62]. Instead, these regions can be used to denote a population-wide positive selection that was imposed through a selective sweep [63]. In terms of the ROH profile, the population of KOVA 2 does not deviate significantly from East Asian populations generally (Figure 3.2). The iSAFE algorithm [47] was used to further characterize intervals that reflect positive selection in KOVA 2. It identified a total of 16,272 loci that were selected in at least one population (iSAFE > 0.2) and revealed a number of distinctive loci specific to each population (172 for KOVA 2, 149 for the Japanese population, 77 for the Chinese population, and 364 for the European population) (Figure 3.3). A well-known locus in LCT showed a significant selection signal in the European population (Figure 3.4), reflecting reliability of the reported signals.

On the basis of the iSAFE score, 172 Korean-specific iSAFE signal loci were ranked and reviewed. Signals that overlap with GWAS signals or are located in genes that have been extensively studied in the past were scrutinized more closely. In the case of an unknown gene, we attempted to interpret its meaning by examining the gene's expression level in each tissue using the Genotype-Tissue Expression (GTEx) project. Interestingly, when compared

to the Japanese, Chinese, and European populations, two loci – ADH1A/1B and UHRF1BP1 – were among the most strongly selected loci in the Korean population (Figure 3.4, Figure 3.5). ADH1A and ADH1B encode alcohol dehydrogenases 1A and 1B, respectively, and are known to comprise recently selected loci in East Asian individuals [64,65]. Here, the Korean population in KOVA 2 showed the strongest signal among the East Asian populations examined. Among the haplotypes that encompass these selected loci, "haplotype #1", previously reported as the East Asian haplotype [64], showed the highest frequency (Figure 3.5). The difference in haplotype frequency between Korean and Japanese was not statistically significant (Fisher's exact test, $p = 0.38$), but it was statistically significant between Korean and Chinese (Fisher's exact test, $p < 0.05$). Additionally, this signal is reflected in the minor allele frequency of rs1229984 as well, which was the lowest among the populations studied, so to yield the most His48 in the populations (Figure 3.5, Figure 3.6). For rs1229984 and rs3811801, the difference in allele frequency between Korean and Japanese was not statistically significant, but it was statistically significant between Korean and Chinese for rs3811801 (Fisher's exact test, $p < 0.05$) (Figure 3.6, Figure 3.7). The prominent Korean–specific signal we found in UHRF1BP1 has not been reported elsewhere, and the function of the gene remains largely unknown.

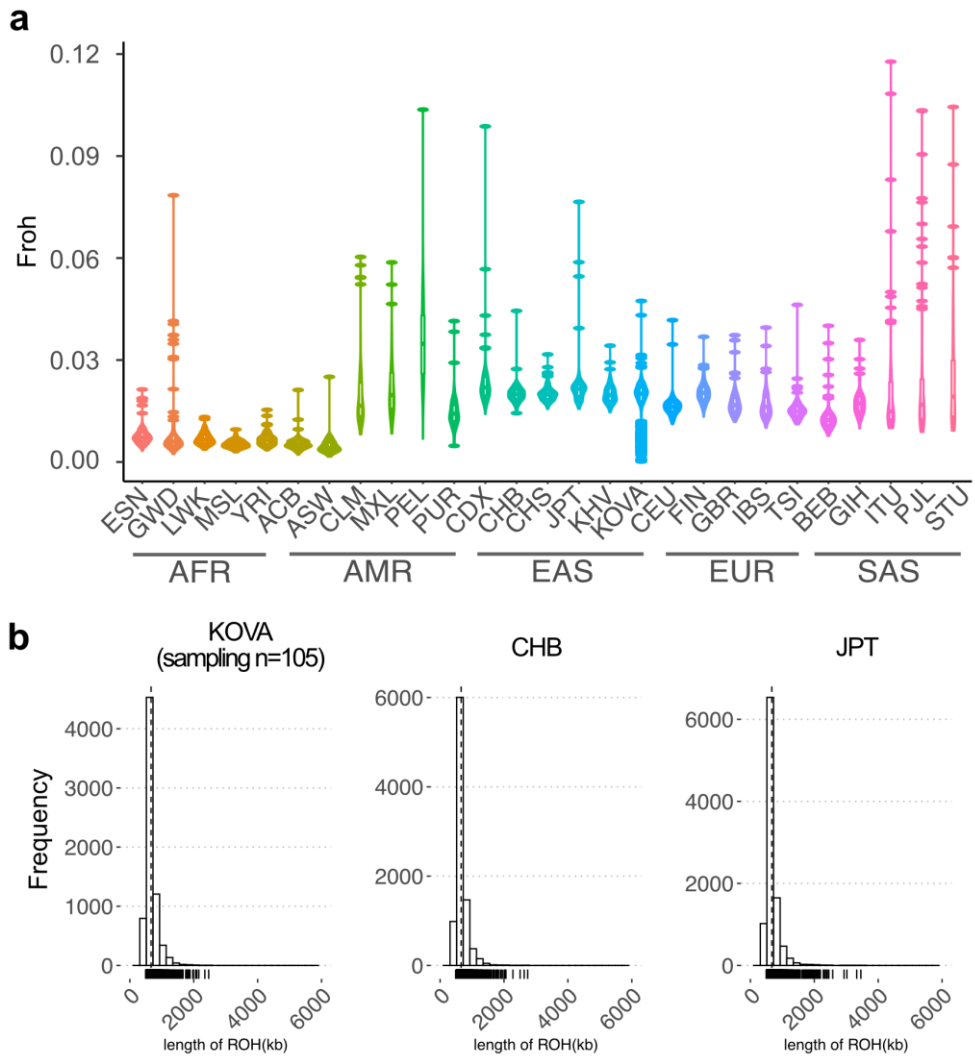

Figure 3.2. ROH profile of KOVA 2 samples. (a) Fraction of ROH per individual (Froh) by population, from KG. (b) Distribution of ROH interval length in KOVA 2, Han Chinese in Beijing (CHB), and Japanese (JPT).

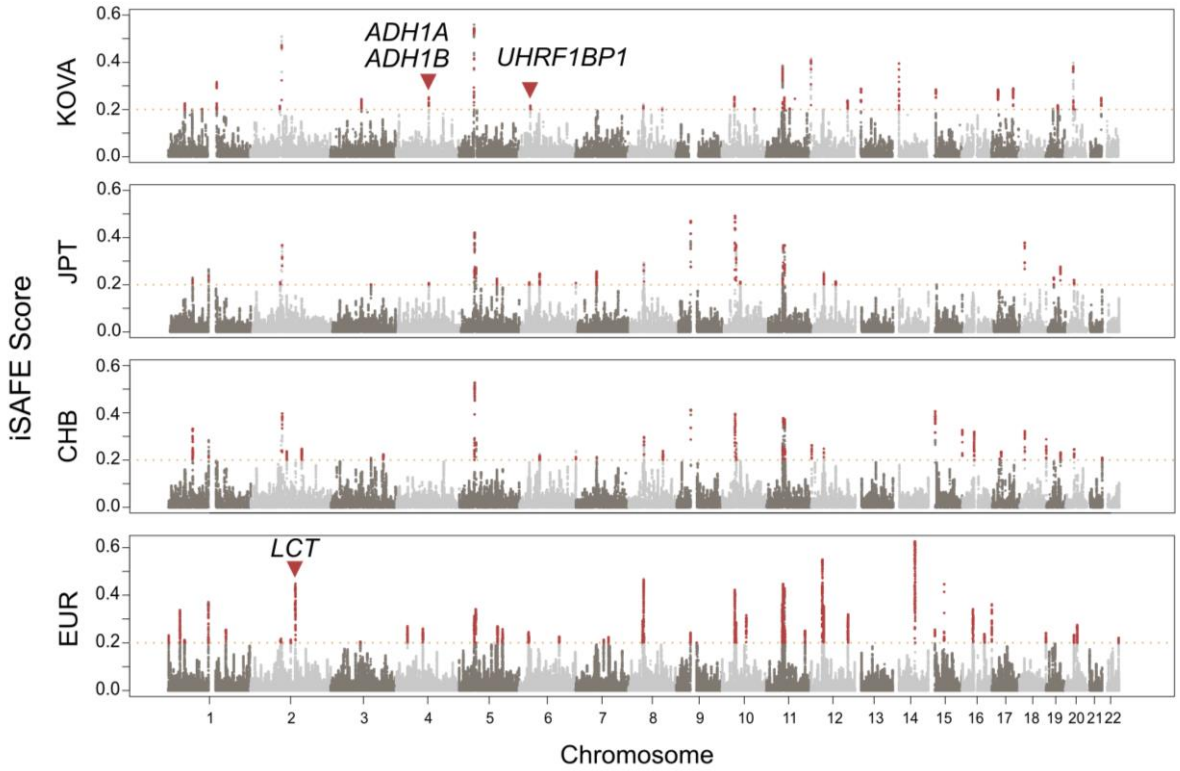

Figure 3.3. Signature of positive selection as indicated by the iSAFE score. Genome-wide iSAFE values were obtained using KOVA 2, Japanese, Chinese, and European cohorts. Gene loci indicated with triangles are separately displayed in Figure 3.3. This figure was created in collaboration with Jean Lee.

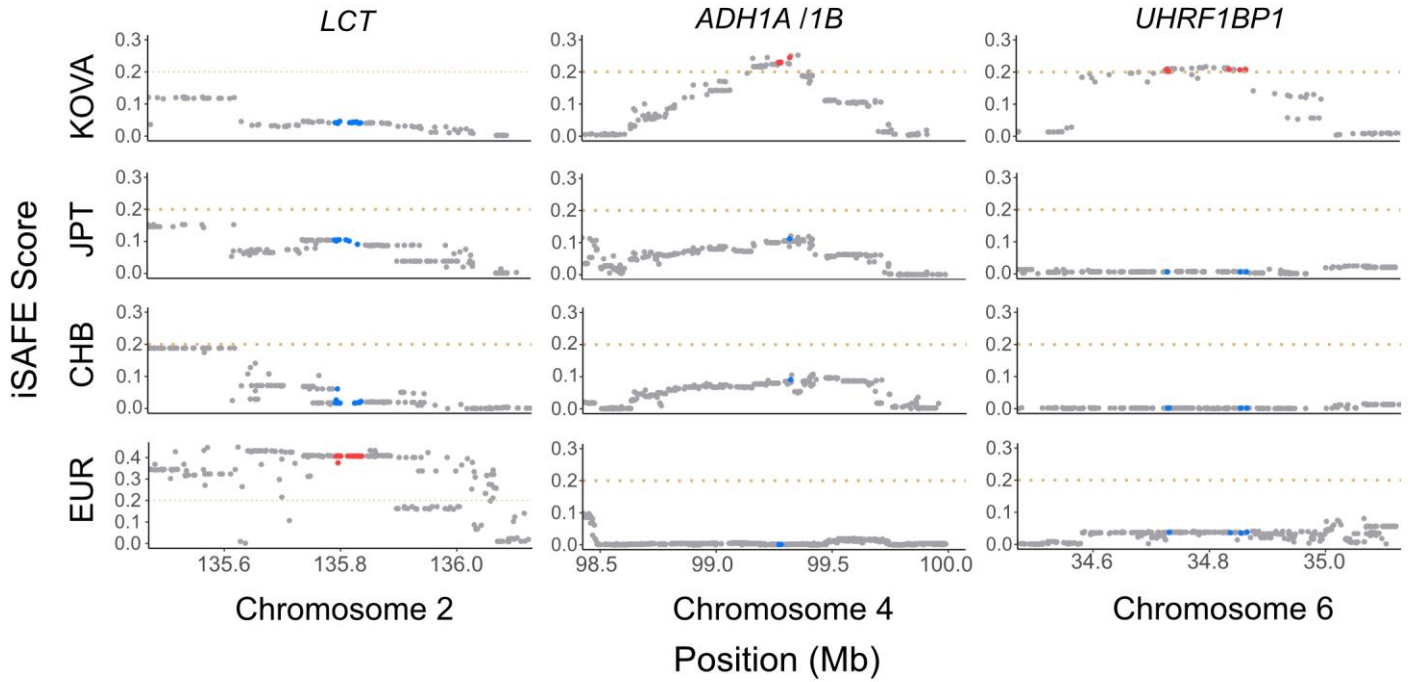

Figure 3.4. Regional plots of the iSAFE values from the same set of ethnic cohorts, as marked in Figure 3.3. This figure was created in collaboration with Jean Lee.

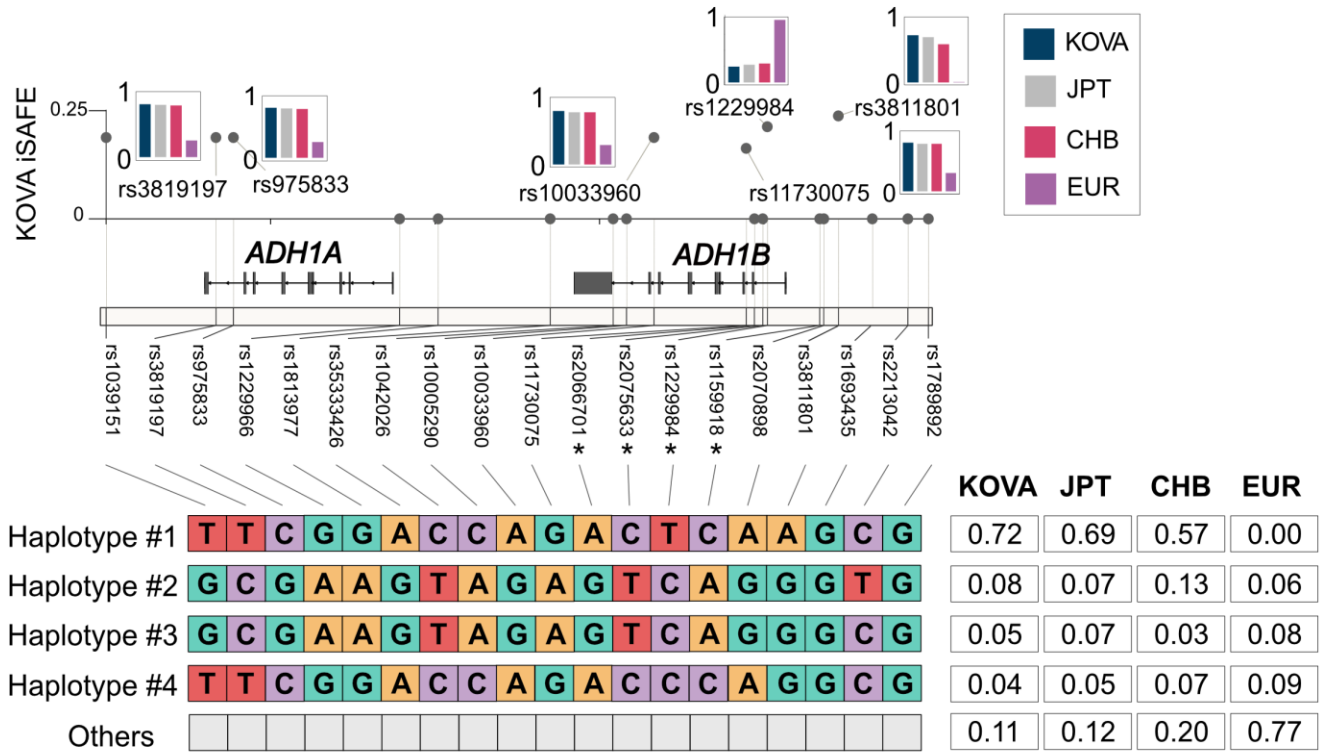

Figure 3.5. The KOVA iSAFE scores and haplotype frequency of ADH1A/1B locus. The KOVA iSAFE scores of selected tag SNPs in the ADH1A/1B locus (top) and their haplotype frequencies by population (bottom). Bar plots denote the MAF of designated SNPs in each population, and SNPs with asterisks denote major markers for haplotype identification used in Han et al., 2007 [56].

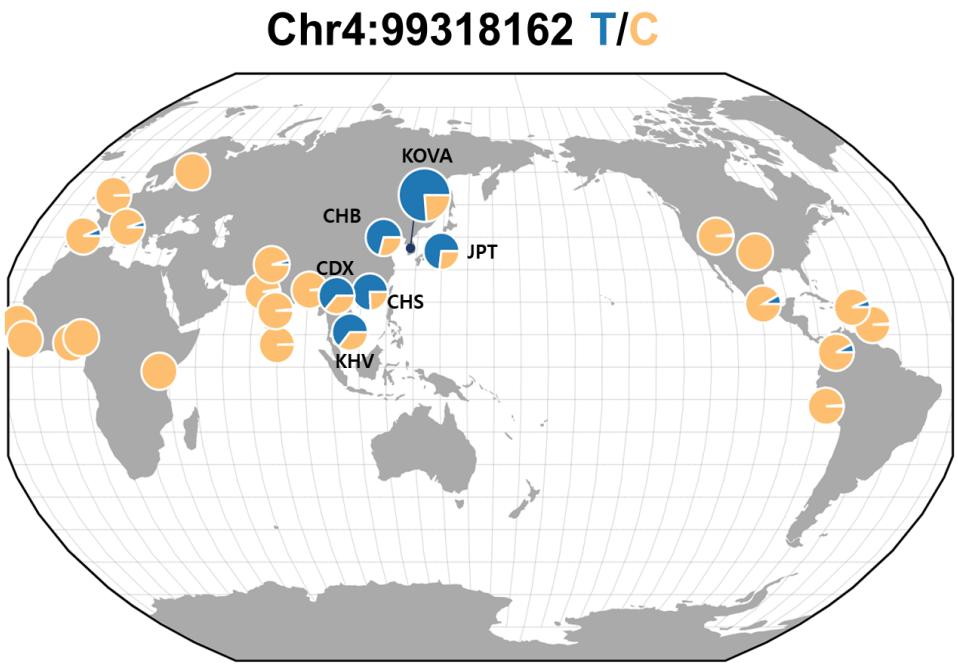

Figure 3.6. Allele frequency map for rs1229984, the tag SNP from the haplotype covering ADH1B. Blue and orange represent the frequency proportions of the effect (T) and non-effect (C) alleles, respectively. Obtained from the Geography of Genetic Variants Browser Beta v0.4

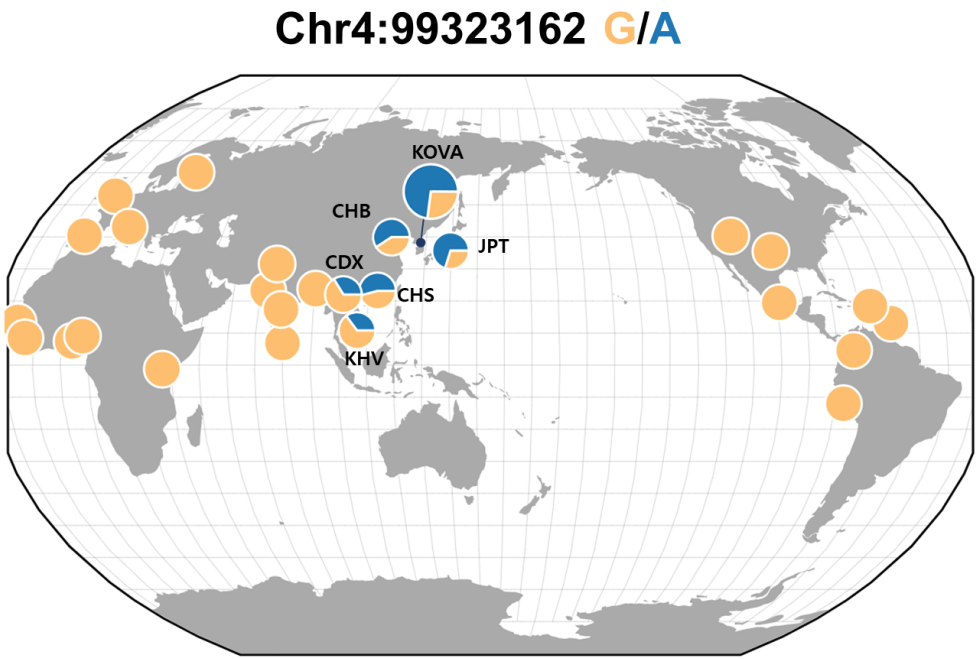

Figure 3.7. Allele frequency map for rs3811801, the tag SNP from the haplotype covering ADH1B. Orange and blue represent the frequency proportions of the effect (A) and non-effect (G) alleles, respectively.

Following that, I sought to determine whether KOVA 2 could be used to estimate the dates of origin for variant, also known as allele ages, and to understand the implications of this information with regard to the frequency and function of variants. Notably, estimating allele ages may lead to the discovery of recently emerged population-specific variants. To carry out this analysis, WGS-originated KOVA 2 variants were phased using a previously reported method [9,29,66]. This enabled us to estimate the population size, which ranged between 10 and 20 million people. This is a value comparable to the current Korean population size of approximately 50 million, especially given the recent population explosion (e.g., the Korean population was approximately 20 million in 1950 and 13 million in 1925 [67]). (Figure 3.8). Next, allele ages were estimated using variants with a frequency greater than 1%. The obtained allele ages correlated strongly with the MAF, as expected. In variants with a high MAF, the allele age was greater, and vice versa. Interestingly, older variants exhibited higher overlap with variants from chimpanzees, implying that some of these variants may have a primate-level origin (Figure 3.9a and Figure 3.10a). When variants were separated by function, it was discovered that older allele ages and higher overlap with chimpanzees corresponded to less functionality, as indicated by annotation (Figure 3.9b, Table 3.1, Figure. 3.10b and Figure 3.11). Remarkably, high confidence LoF and missense variants with high CADD score were the youngest and showed minimal overlap with

chimpanzees. Furthermore, all functional classes of rare variants with a MAF less than 5% were young and did not overlap with chimpanzees (Figure 3.9c and Figure 3.10c). When variants were classified by pLI score, this trend was not clearly replicated (Figure 3.12). Overall, these findings suggest that the majority of rare variants are of relatively recent origin and thus tend to be population specific. The allele age of unique variants found specifically in ethnic outliers which samples are inferred as not Korean was compared to the allele age of variants shared by ethnic outliers and the Korean cohort (Figure 3.13). The difference in allele ages between the unique and common variants was minimal, most likely because allele frequency and function of a variant are stronger predictor of allele age than ethnic distribution.

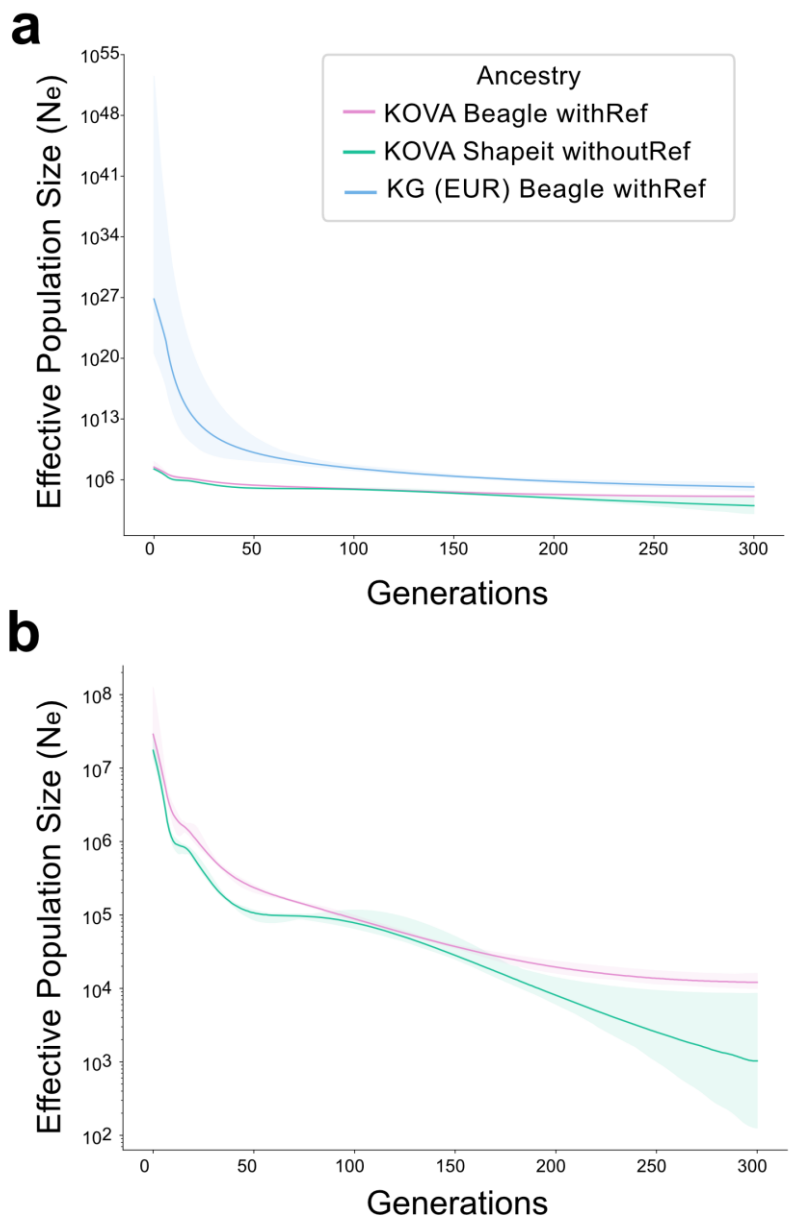

Figure 3.8. Estimated effective population size based on KOVA. (a) Effective Korean population size by generations before the present. (b) Estimated population size of Korean population calculated based on KOVA data, subset of (a).

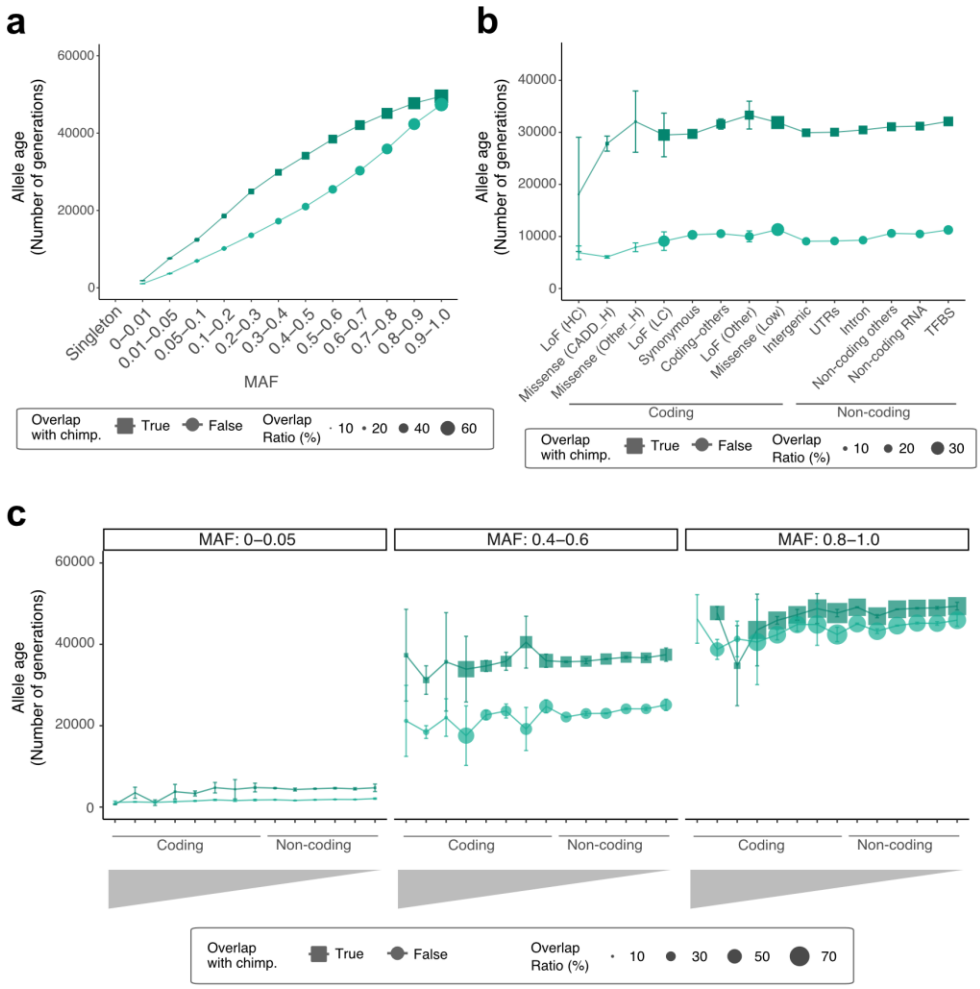

Figure 3.9. Allele ages of KOVA 2 variants based on KOVA. (a) Allele ages by MAF, divided by whether the allele cooccurs in chimpanzees (squares) or not (circles). (b) Allele age by predicted function, divided by whether the allele cooccurs in chimpanzees (squares) or not (circles). TFBS denotes the 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.

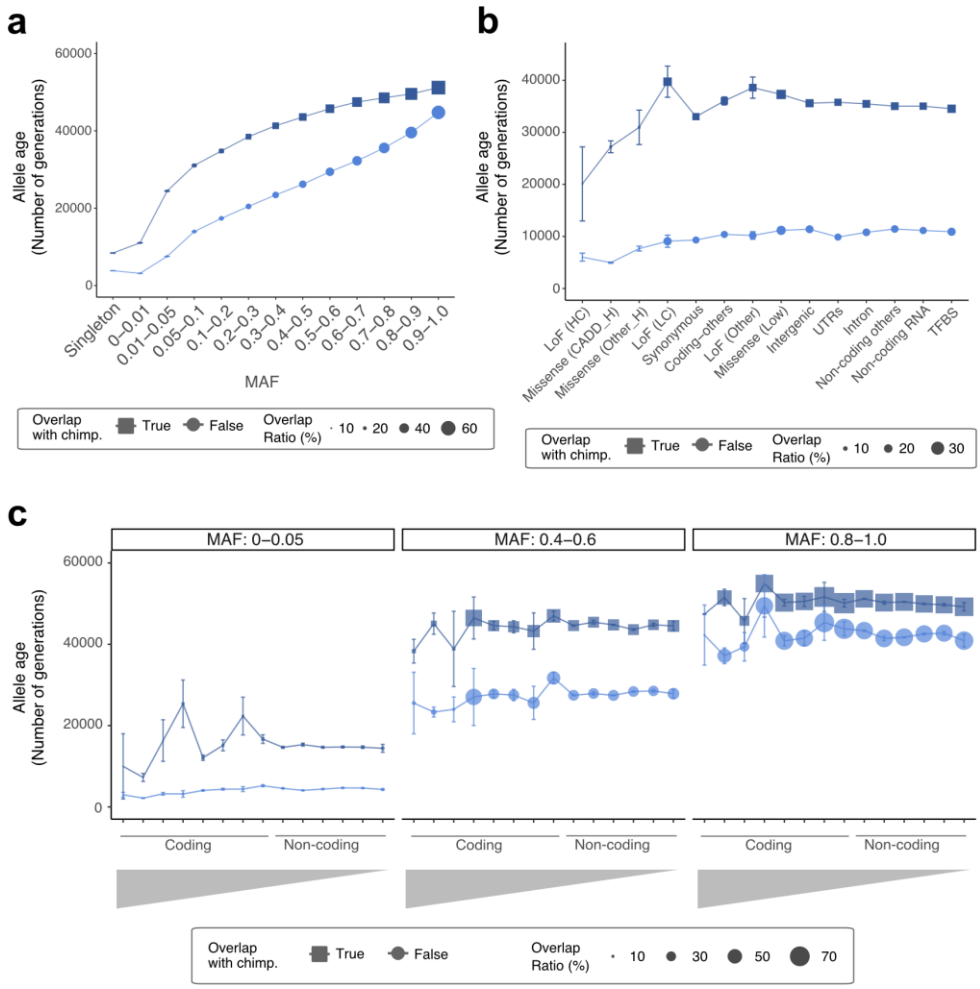

Figure 3.10. Allele ages of KOVA 2 variants based on KG data. (a) Allele ages by MAF, divided by the co-occurrence from chimpanzees (squares) or not (circles). (b) Allele ages by predicted function, divided by the co-occurrence from chimpanzees (squares) or not (circles). (c) Allele ages by MAF and predicted function. Three MAF intervals are displayed. The X-axis bins depicted in grey triangles in (c) are the same as that of (b)

Table 3.1. Allele age by variant class

	All variants				Variants co-occurred from chimpanzee				Ratio (The co-occurrence from Chimp/All)			
Variant class	**kova allele age**	**kova allele cnt**	**kg allele age**	**kg allele cnt**	**kova allele age**	**kova allele cnt**	**kg allele age**	**kg allele cnt**	**kova allele age**	**kova allele cnt**	**kg allele age**	**kg allele cnt**
LoF (HC)	7,190	431	6,343	1,123	18,075	12	20,090	25	2.51	0.03	3.17	0.02
Missense_OH*	8,962	1,164	8,759	3,484	32,063	50	30,963	163	3.58	0.04	3.53	0.05
Missense_CH**	7,704	12,567	6,075	36,661	27,832	945	27,227	1,815	3.61	0.08	4.48	0.05
intergenic	13,374	1,898,606	16,082	2,608,107	29,926	391,635	35,596	505,764	2.24	0.21	2.21	0.19
intron	13,742	4,353,161	15,409	6,593,118	30,459	915,522	35,470	1,233,526	2.22	0.21	2.30	0.19
ncRNA	14,846	190,890	15,313	326,043	31,192	40,277	35,030	56,884	2.10	0.21	2.29	0.17
UTRs	13,565	121,912	14,302	225,036	30,031	25,877	35,776	38,390	2.21	0.21	2.50	0.17
nc-others	15,093	747,565	15,805	1,234,003	31,091	163,582	35,018	229,091	2.06	0.22	2.22	0.19
coding-others	15,137	8,618	14,787	16,819	31,490	1,901	36,065	2,883	2.08	0.22	2.44	0.17
synonymous	15,062	18,811	13,362	48,377	29,723	4,600	32,997	8,260	1.97	0.24	2.47	0.17
LoF (LC)	14,914	400	15,325	840	29,496	114	39,728	171	1.98	0.29	2.59	0.20
Missense_Lo***	17,979	10,403	17,105	24,103	31,886	3,367	37,284	5,475	1.77	0.32	2.18	0.23

*: **Missense_OH: missense variants which have high pathogenic-scores other than CADD**

: **Missense_CH: missense variants which have high CADD score

***: **Missense_Lo: missense variants which have low pathogenic-scores including CADD**

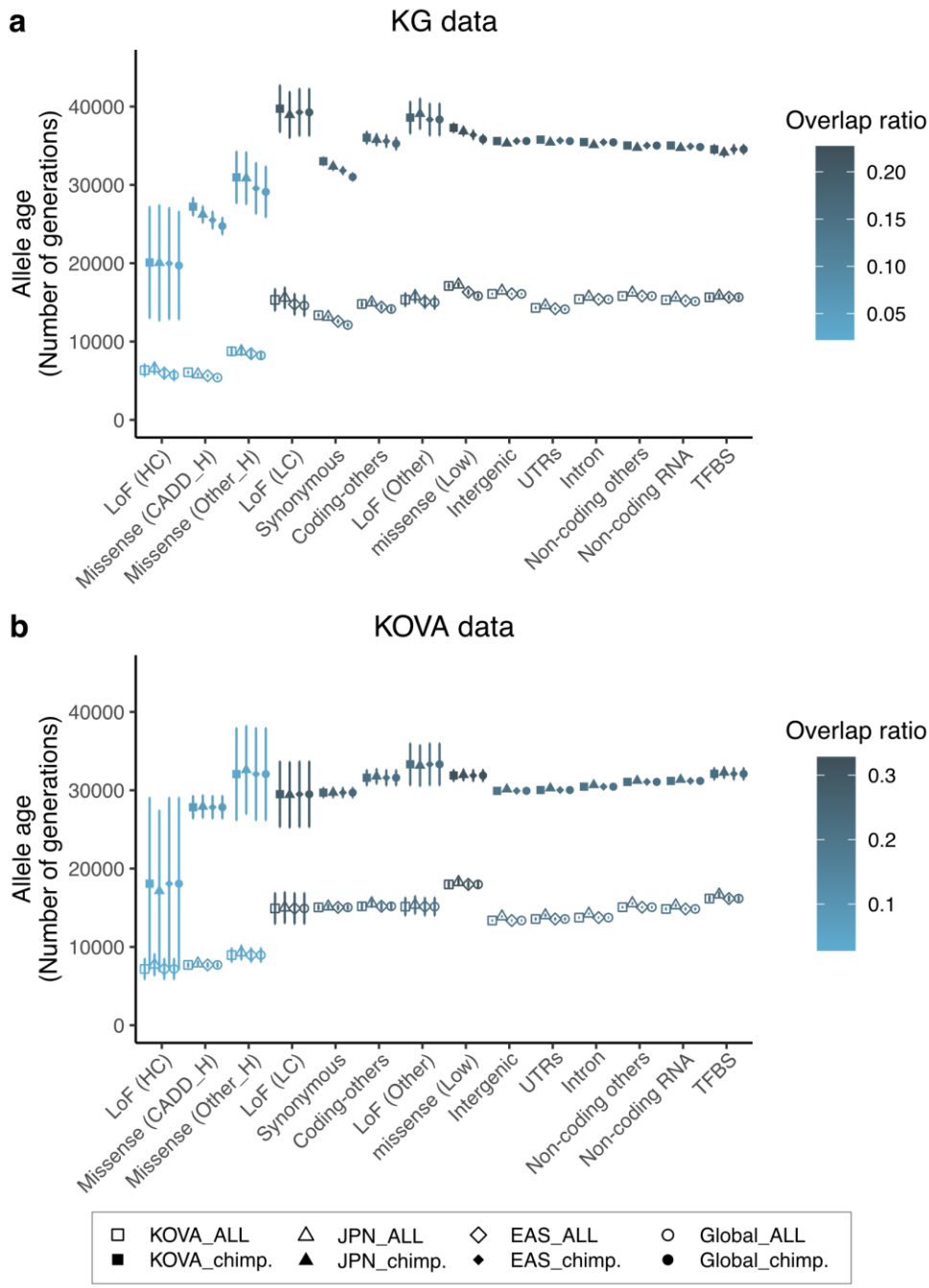

Figure 3.11. Allele ages of KOVA 2 variants by population. (a) Allele ages based on KG data by predicted function, divided by the co-occurrence from chimpanzees (filled) or not (blank). (b) Allele ages based on KOVA data by predicted function, divided by the co-occurrence from chimpanzees (filled) or not (blank).

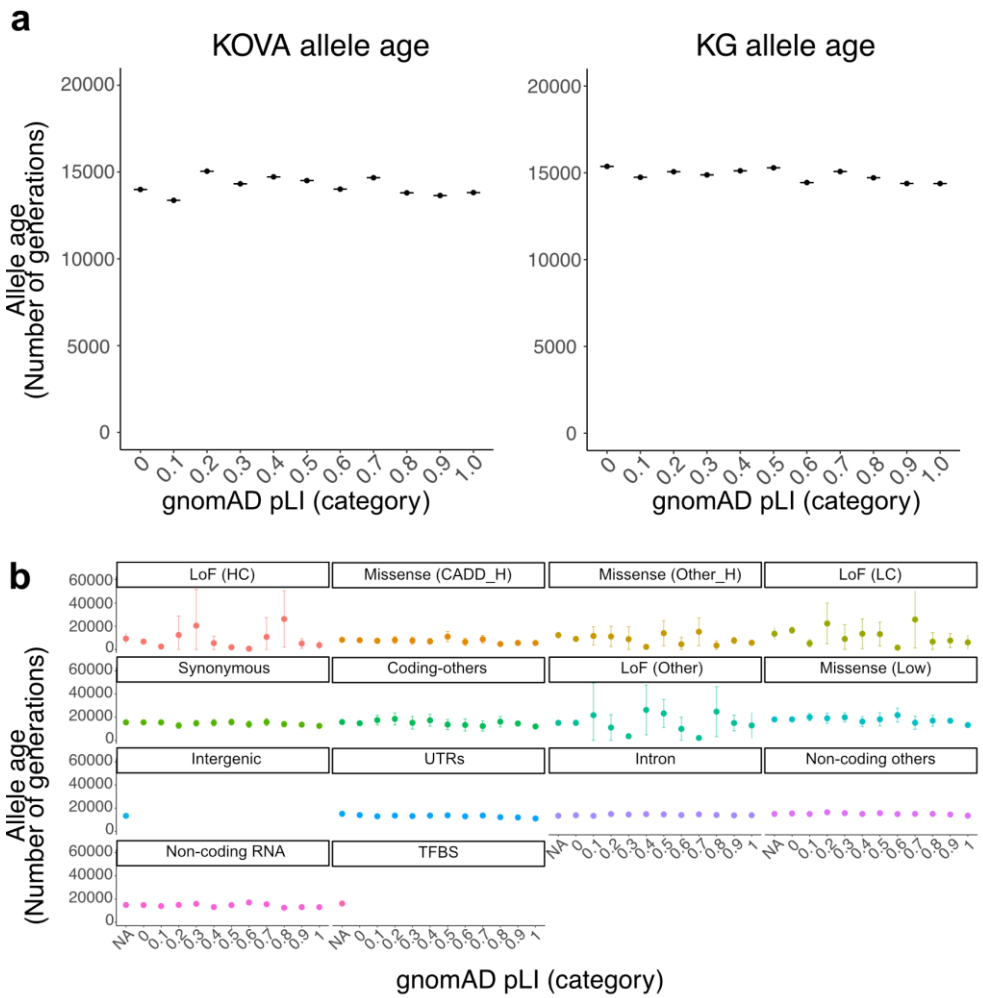

Figure 3.12. Allele age of KOVA 2 variants by pLI score. (a) Allele ages based on KOVA 2 data (left) and on KG data (right) by decile of gnomAD pLI score. NA represents variants without pLI scores. (b) Allele ages based on KOVA 2 data by pLI per predicted variant function.

.

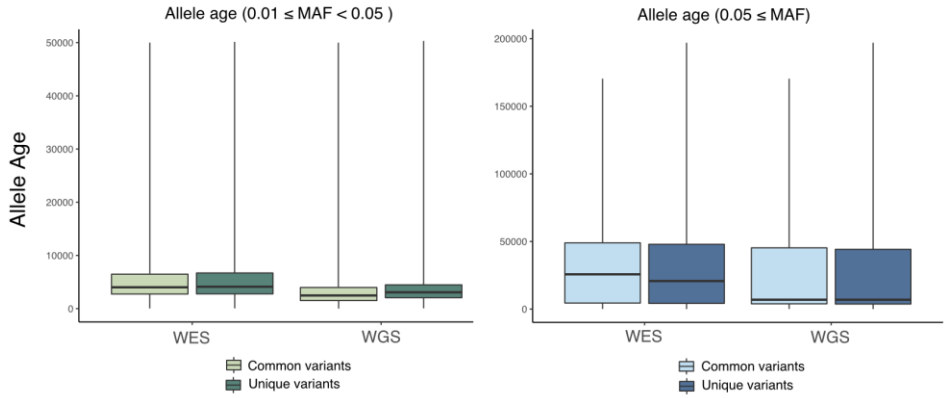

Figure 3.13. Comparison of allele age for variants found specifically to the ethnic outliers and variants common between the ethnic outliers and Korean cohort. This figure was created in collaboration with Jean Lee.

To determine whether the KOVA 2 set contains variants that have previously been annotated as pathogenic, KOVA-specific rare variants (MAF < 0.001) in high pLI genes were selected 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 with a dominant inheritance pattern (Table 3.2). This finding implies that these variants may not be as pathogenic as previously thought. Alternatively, because KOVA 2 is made up of three main types of individuals, i.e., healthy volunteers, normal genomes of cancer patients, and healthy parents of rare disease patients, one could argue that the variants predispose carriers to develop cancer or their children to manifest rare diseases.

Table 3.2. ClinVar pathogenic variants in KOVA 2. 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	Clin Var **	Dominant or Recessive ***	ClinVar condition
LoF (HC)	chr3:128 481942	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
	chr3:412 36467	CAG>C	1	12,148	0.8×10^{-4}	V	*CTNNB1*	1.00	P	D	Mental retardation, autosomal dominant 19; Inborn genetic diseases
	chr6:790 26060	A>C	1	12,150	0.8×10^{-4}	P	*PHIP*	1.00	P	D	Developmental delay, intellectual disability, obesity, and dysmorphic features
	chr7:128 846444	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
	chr9:954 58142	G>T	1	12,120	0.8×10^{-4}	V	*PTCH1*	1.00	P	D	Gorlin syndrome
	chr12:86 8379	C>T	1	12,134	0.8×10^{-4}	C	*WNK1*	1.00	P	D/R	Hereditary sensory and autonomic neuropathy type IIA
	chrX:400 64351	G>A	1	12,224	0.8×10^{-4}	C	*BCOR*	1.00	P	D	Oculofaciocardiodental syndrome

Missense	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	9,310	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
	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
	chr7:55528486	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
	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
	chr11:11908974 7	G>A	1	12,144	0.8×10^{-4}	V	*HMBS*	0.95	P	D	Acute intermittent porphyria
	chr11:11909278 5	G>A	1	12,138	0.8×10^{-4}	V	*HMBS*	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

Chapter 4. Discussion

In this study, I established the largest Korean control genome database to date, along with information on its genetic characteristics and uses. KOVA 2 displayed the major features of population genome databases, and considerable genetic information was additionally analyzed from the dataset. The KOVA 2 variant set has been uploaded and will be shared to the community for use as a control set in East Asian genetic studies.

The identified variants in the KOVA 2 dataset showed typical patterns of purifying selection and frequency–functionality relationships. The sample size was insufficient to encompass all rare variants in the population, as is the case with larger population genome database. However, it exhibited the best coverage of common variants in Koreans and thus performed better when imputing variants. Although KOVA 2 can serve as a control set to screen nonpathogenic variants for rare Mendelian diseases, a list of ClinVar pathogenic variants that are present in KOVA 2 at low frequency was identified. It should be further clarified whether these variants are nonpathogenic in the Korean population or whether their carriers were able to avoid developing the associated diseases because of their genomic background. A combined analysis of the positive selection signatures and allele age estimation may lead to the discovery of genetic loci that have recently emerged and been selected in a population.

Not surprisingly, the top signals in this Korean population overlapped with those in neighboring East Asian populations. This result implies a recent divergence, ongoing admixture, and similar environmental constraints that were placed on these populations during recent evolution. Nonetheless, our findings identified loci worthy of further investigation. For example, detailed dissection of a previously reported East Asian-selected alcohol dehydrogenase gene locus discovered that it was the most strongly selected locus among the three East Asian populations studied here. The major haplotype in East Asian populations ("Haplotype #1" in Figure 3.5) was the most abundant in the Korean population, and the *ADH1B* His48 allele indicated by low frequency of rs1229984 was the most common in the Korean population among East Asian populations (Figure 3.5). This His48 allele is known to increase aldehyde production when compared to the wild-type counterpart. This is due to increased ethanol oxidation, which causes adverse reactions such as flushing and nausea [68]. In the long term, this allele is also protective against alcohol dependency [69]. The functional consequence of the second locus of interest, *UHRF1BP1*, is unknown due to a lack of research. Nonetheless, it is remarkable that associations between this variant and systematic lupus erythematosus have been repeatedly reported in East Asian populations [70-72]. This gene is most strongly expressed in the testes (Figure 4.1), implying that it can confer selection by influencing the reproductive process in males. A new algorithm

based on large-scale population data may discover novel loci that were missed in this study, in addition to these two loci.

Furthermore, to provide a guideline for future decision making on national-wide genomic studies or large-scale disease cohorts, the number of ClinVar pathogenic variants from WES and WGS were compared. Although not many pathogenic variants were included in KOVA 2, the result showed the complete concordance (100%) between these two sequencing platform.

Currently, several biobank projects have collected demographic information as well as a wide range of clinical laboratory test data, but KOVA 2 contains only genotype information along with sequencing-related information. As a result, future efforts to create a database containing such information are required.

Finally, the KOVA 2 data was uploaded to a genome browser and enabled users to download the variant set with a simple registration process. The creation of a Korean-specific variant set and comparative analysis will bolster a wide range of genetic and genomic studies involving East Asian populations. It will also serve as a precursor for much larger genome datasets that will be available soon, particularly if they can be combined with data from North Korean individuals.

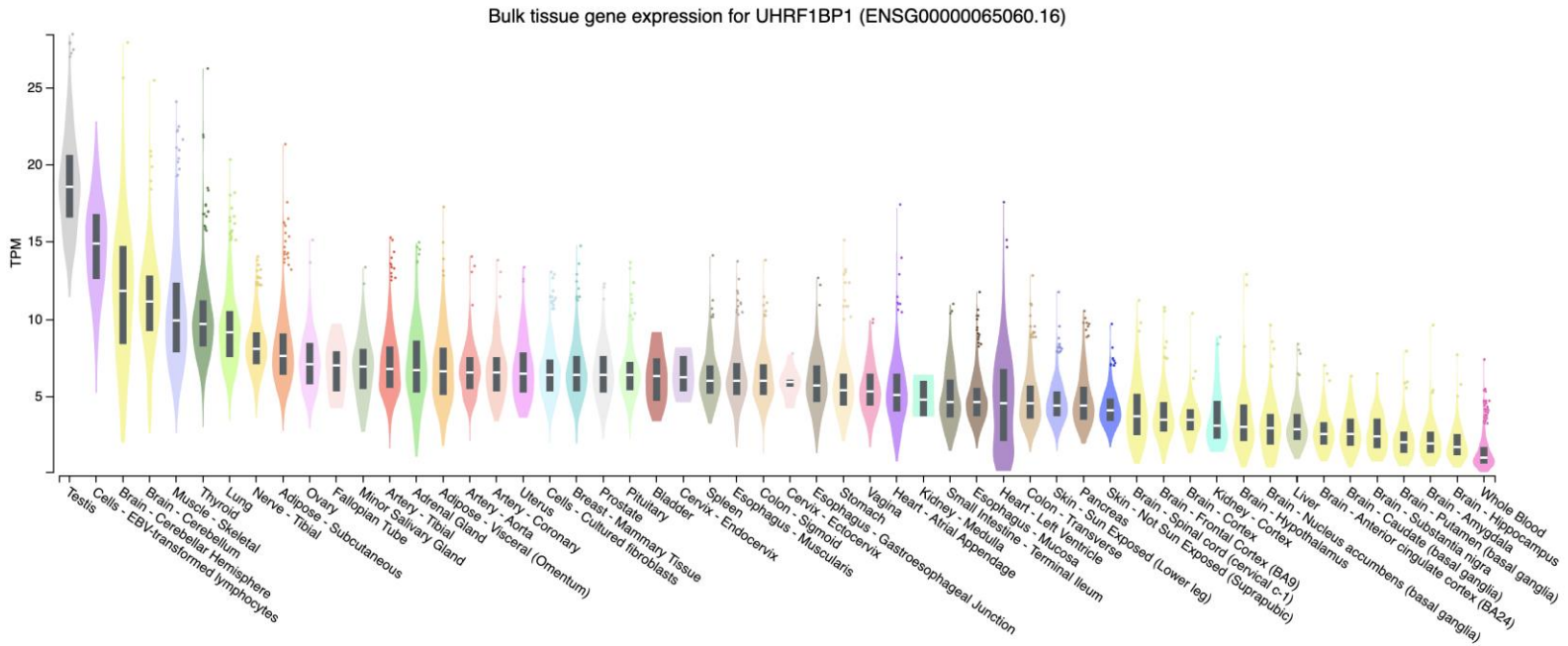

Figure 4.1. Tissue expression profile of UHRF1BP1. Displaying the highest expression in testes (on the far left; https://gtexportal.org/home/gene/UHRF1BP1). This figure was created in collaboration with Jean Lee.

Reference

1. Lee, J. et al. A database of 5305 healthy Korean individuals reveals genetic and clinical implications for an East Asian population. *Exp. Mol. Med.* **54**, 1862–1871 (2022).

2. Richards, S. et al. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet. Med.* **17**, 405–424 (2015).

3. Li, MM, et al. Standards and guidelines for the interpretation and reporting of sequence variants in cancer: a joint consensus recommendation of the Association for Molecular Pathology, American Society of Clinical Oncology, and College of American Pathologists. *J. Mol. Diagn.* **19**, 4–23 (2017).

4. Sirugo, G., Williams, S. M. & Tishkoff, S. A. The Missing Diversity in Human Genetic Studies. *Cell* **177**, 26–31 (2019).

5. Karczewski, K. J. et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).

6. Lek, M. et al. Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285-291 (2016).

7. Lee, S. et al. Korean Variant Archive (KOVA): a reference database of genetic variations in the Korean population. *Sci. Rep.* **7**, 4287 (2017).

8. Kwak, S. H. et al. Findings of a 1303 Korean whole-exome sequencing study. *Exp. Mol. Med.* **49**, e356–e356 (2017).

9. Jeon, S. et al. Korean Genome Project: 1094 Korean personal genomes with clinical information. *Sci. Adv.* **6**, eaaz7835 (2020).

10. 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).

11. 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).

12. 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).

13. Manichaikul, A. et al. Robust relationship inference in genome-wide association studies. *Bioinformatics* **26**, 2867-2873 (2010).

14. Gentry, J. et al. OpenWDL. https://openwdl.org (2022).

15. Hao, M. et al. The HuaBiao project: whole-exome sequencing of 5000 Han Chinese individuals. *J. Genet. Genomics* **48**, 1032-1035 (2021).

16. Wei, CY. et al. Genetic profiles of 103,106 individuals in the Taiwan Biobank provide insights into the health and history of Han Chinese. *npj Genom. Med.* **6**, 10 (2021).

17. Zhang, P. et al. NyuWa Genome resource: A deep whole-genome sequencing-based variation profile and reference panel for the Chinese population. *Cell Rep.* **37**, 110017 (2021).

18. Cong, PK. et al. Genomic analyses of 10,376 individuals in the Westlake BioBank for Chinese (WBBC) pilot project. *Nat. Commun.* **13**, 2939 (2022).

19. Nagasaki, M. et al. Rare variant discovery by deep whole-genome sequencing of 1,070 Japanese individuals. *Nat. Commun.* **6**, 8018 (2015).

20. Tadaka, S. et al. 3.5KJPNv2: an allele frequency panel of 3552 Japanese individuals including the X chromosome. *Hum Genome. Var.* **6**, 28 (2019).

21. Tadaka, Shu. et al. jMorp updates in 2020: large enhancement of multi-omics data resources on the general Japanese population. *Nucleic Acids Res.* **49**, D536-D544 (2021).

22. Auwera, G. A. V. der & O'Connor, B. D. *Genomics in the cloud: using Docker, GATK, and WDL in Terra*. (O'Reilly Media, 2020).

23. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv https://doi.org/10.48550/arXiv.1303.3997 (2013).

24. Okonechnikov, K., Conesa, A. & García-Alcalde, F. Qualimap 2: advanced multi-sample quality control for high-throughput sequencing data. *Bioinformatics* **32**, 292–294 (2016).

25. Team, H. Hail 0.2.77-684f32d73643. https://github.com/hail-is/hail/commit/684f32d73643 (2021).

26. Li, H. Toward better understanding of artifacts in variant calling from high-coverage samples. *Bioinformatics* 30, 2843–2851 (2014).

27. 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).

28. Team, H. "maximal independent set" method. https://hail.is/docs/0.2/methods/misc.html#hail.methods.maximal_independent_set (2021).

29. Auton, A. et al. A global reference for human genetic variation. *Nature* 526, 68–74 (2015).

30. Tate, John G. et al. COSMIC: the catalogue of somatic mutations in cancer. *Nucleic Acids Res.* 47, D941-D947 (2019).

31. Chakravarty, D. et al. OncoKB: A Precision Oncology Knowledge Base. *JCO Precis. Oncol.* 2017, PO.17.00011 (2017).

32. McLaren, W. et al. The Ensembl Variant Effect Predictor. *Genome Biol.* **17**, 122 (2016).

33. 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).

34. 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).

35. 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).

36. 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).

37. 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).

38. Chen, X. et al. Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinformatics* 32, 1220–1222 (2016).

39. Collins, R. L. et al. A structural variation reference for medical and population genetics. *Nature* 581, 444–451 (2020).

40. Eggertsson, H. P. et al. GraphTyper2 enables population-scale genotyping of structural variation using pangenome graphs. *Nat. Commun.* 10, 5402 (2019).

41. Halldorsson, B. V. et al. The sequences of 150,119 genomes in the UK Biobank. *Nature* 607, 732-740 (2022).

42. Public GATK Panel of Normals for hg38. gs://gatk-best-practices/somatic-hg38/1000g_pon.hg38.vcf.gz (2022).

.43. 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).

44. 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).

45. Fu, Y. et al. FunSeq2: a framework for prioritizing noncoding regulatory variants in cancer. *Genome Biol.* **15**, 480 (2014).

46. Huang, Y.-F., Gulko, B. & Siepel, A. Fast, scalable prediction of deleterious noncoding variants from functional and population genomic data. *Nat. Genet.* **49**, 618–624 (2017).

47. Akbari, A. et al. Identifying the Favored Mutation in a Positive Selective Sweep. *Nat. Methods* **15**, 279–282 (2018).

48. Pemberton, T. J. et al. Genomic Patterns of Homozygosity in Worldwide Human Populations. *Am. J. Hum. Genet.* **91**, 275–292 (2012).

49. Albers, P. K. & McVean, G. Dating genomic variants and shared ancestry in population-scale sequencing data. *PLoS Biol.* **18**, e3000586 (2020).

50. Vitti, J. J., Grossman, S. R., & Sabeti, P. C. Detecting natural selection in genomic data. *Annu. Rev. Genet.* **47**, 97–120 (2013).

51. Allison AC. Protection afforded by sickle–cell trait against subtertian malarial infection. *BMJ* **1**, 290–294 (1954).

52. Eichstaedt, CA. et. al. Genetic and phenotypic differentiation of an Andean intermediate altitude population. *Physiol. Rep.* **3**, e12376 (2015).

53. Hancock, AM. et. al. Adaptations to climate in candidate genes for common metabolic disorders. *PLoS Genet.* **4**, e32 (2008).

54. Purcell, S. & Chang, C. PLINK 1.9. www.cog–genomics.org/plink/1.9/ (2019).

55. Chang, C. C. et al. Second–generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**, 1–16 (2015).

56. Biscarini, F., Cozzi, P., Gaspa, G. & Marras, G. detectRUNS: Detect runs of homozygosity and runs of heterozygosity in diploid genomes. CRAN (The Comprehensive R Archive Network) (2019).

57. 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).

58. 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).

59. Browning, B. L. & Browning, S. R. Improving the Accuracy and Efficiency of Identity-by-Descent Detection in Population Data. *Genetics* **194**, 459–471 (2013).

60. Albers, P. K. & McVean, G. Human Genome Dating. https://human.genome.dating/download/index (2020).

61. Prado-Martinez, J. et al. Great ape genetic diversity and population history. *Nature* **499**, 471–475 (2013).

62. 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).

63. Pemberton, T. J. et al. Genomic Patterns of Homozygosity in Worldwide Human Populations. *Am. J. Hum. Genet.* **91**, 275–292 (2012).

64. Han, Y. et al. Evidence of Positive Selection on a Class I ADH Locus. *Am. J. Hum. Genet.* **80**, 441–456 (2007).

65. 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).

66. Wall, J. D. et al. The GenomeAsia 100K Project enables genetic discoveries across Asia. *Nature* **576**, 106–111 (2019).

67. Korea, S. Korean statistical information service. https://kosis.kr/eng/ (2022).

68. Edenberg, H. J. The genetics of alcohol metabolism: role of alcohol dehydrogenase and aldehyde dehydrogenase variants. *Alcohol Res. Health* **30**, 5–13 (2007).

69. Li, D., Zhao, H. & Gelernter, J. Strong association of the alohol dehydrogenase 1B gene (ADH1B) with alcohol dependence and alcohol-induced medical diseases. *Biol. Psychiatry* **70**, 504–12 (2011).

70. Wu, J. et al. The Rare Variant rs35356162 in UHRF1BP1 Increases Bladder Cancer Risk in Han Chinese Population. *Front. Oncol.* **10**, 134 (2020).

71. 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).

72. 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).

국문 초록

건강한 개인의 유전체에 대한 이해는 인간 발생 및 질병 생리학 연구, 유전 질환에 대한 임상적 진단의 근간이 된다. 따라서 질병유전학의 상당한 발전에 발맞추어 일반 인구에 대한 유전체 데이터베이스의 중요성 또한 대두되고 있다. 그러나 현재까지의 연구는 주로 유럽계 개인에 초점을 맞추어 진행되어왔기에, 다른 인종 그룹에서 새로운 기능적 유전 변이의 추가 발견이 제한적으로 이루어졌다. 이에 따라 동아시아 국가별로 독자적인 유전체 데이터베이스를 구축하고자 하는 노력이 점차 커지고 있으나, 한국인의 유전체 데이터베이스는 인접 동아시아 국가의 데이터베이스 구축 속도에 못 미치고 있는 실정이다.

본 연구에서는 한국인 인구 유전체 자원의 부족을 해소하고 동아시아 인구 유전체 데이터베이스 구축에 기여하기 위해, 건강한 한국인의 1,896개의 전장 유전체 염기서열 정보와 3,409개의 전장 엑솜 염기서열 정보로 구성된 한국인 유전체 데이터베이스 (KOVA2)를 구축하였다. 이는 gnomAD에 포함된 한국인 1,909명에 대한 유전체 데이터를 넘어 역대 최대 규모의 한국인 특화 유전체 데이터베이스이다. 구축된 유전체 데이터베이스는 초기 데이터부터 통일된 파이프라인을 통해 변이를 검출하였으며, 건강한 한국인만의 높은 정확도의 유전 변이만이 데이터베이스에 포함되도록 하였다. 이를 통해 40,414,379개의 단일 염기 변이와 2,888,275 삽입/삭제 변이 정보를 얻었으며, 전장 유전체 데이터를 이용하여 144,388개의 구조 변이에 대한 정보를 정리하였다. KOVA 2 데이터베이스를 이용해 분석된 한명의 샘플을 다른 시퀀싱 플랫폼으로도

시퀀싱하여 변이 검출의 정확도를 평가한 결과 시퀀싱 플랫폼간에 높은 일치율을 보였다. 또한 이전에 발표된 유전체 데이터베이스를 통해 알려진 유전적 특징 모두 보임으로써 KOVA2 변이의 신뢰도를 검증하였다.

구축된 KOVA2 데이터베이스는 동형 접합성의 연속성, 진화적으로 양성적 선택이 이루어진 영역, 변이의 나이, 그리고 인구수의 변화를 추정하는 데 추가적으로 활용하여 한국인 특이적 유전적 특징을 분석하였다. 그 과정에서 ADH1A/1B 및 UHRF1BP1 유전자좌와 같이 다른 동아시아 인구에 비해 한국인에게서 진화적으로 강하게 선택되는 유전자좌를 발견했다. 대립형질의 나이를 분석한 결과는 유전변이의 기능과 진화적 나이 사이에 존재하는 상관관계를 밝혔다. 동형 접합성의 연속성을 파악한 결과는 한국인 특이적인 차이를 보이지 않았으며, 한국인의 인구수 통계 기록과 유사한 시간별 인구수를 추정할 수 있었다.

변이별로 추정된 변이의 나이와 양성적 선택의 크기를 포함한 한국인의 유전 변이 정보는 공개 웹사이트에서 검색 및 다운로드할 수 있도록 하였다. 본 연구 결과는 동아시아 인구를 대상으로 하는 유전학 연구에 새로운 귀감을 줄 수 있는 귀중한 자료가 될 것이다.

* 본 학위 논문은 출판된 논문 (Lee *et al.*, *Exp. Mol. Med.* 54:1862-1871 (2022)) 을 바탕으로 작성되었음 [1].

키워드: 유전체 데이터베이스, 엑솜 시퀀싱, 전장 유전체, 동아시아인, 한국인, 양성적 선택, 대립형질 나이

학 번: 2014-30270