



저작자표시-비영리-변경금지 2.0 대한민국

이용자는 아래의 조건을 따르는 경우에 한하여 자유롭게

- 이 저작물을 복제, 배포, 전송, 전시, 공연 및 방송할 수 있습니다.

다음과 같은 조건을 따라야 합니다:



저작자표시. 귀하는 원저작자를 표시하여야 합니다.



비영리. 귀하는 이 저작물을 영리 목적으로 이용할 수 없습니다.



변경금지. 귀하는 이 저작물을 개작, 변형 또는 가공할 수 없습니다.

- 귀하는, 이 저작물의 재이용이나 배포의 경우, 이 저작물에 적용된 이용허락조건을 명확하게 나타내어야 합니다.
- 저작권자로부터 별도의 허가를 받으면 이러한 조건들은 적용되지 않습니다.

저작권법에 따른 이용자의 권리는 위의 내용에 의하여 영향을 받지 않습니다.

이것은 [이용허락규약\(Legal Code\)](#)을 이해하기 쉽게 요약한 것입니다.

[Disclaimer](#)

의학박사 학위논문

**1,779 whole-genome datasets unveil
population-specific genetic architecture
and pharmacogenomics profile
in Northeast Asian reference panel**

1,779 명 동북아시아인의
전장 유전체 데이터를 기반으로 한
참조 패널 생성과 유전학적 인구 특성 구조 및
약리 유전체학 프로파일의 연구

2020 년 8 월

서울대학교 대학원
의과학과 의과학 전공
김 창 욱


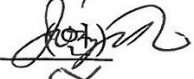



1,779명 동북아시아인의
전장 유전체 데이터를 기반으로 한
참조 패널 생성과 유전학적 인구 특성 구조 및
약리 유전체학 프로파일의 연구

지도교수 김 종 일

이 논문을 의학박사 학위논문으로 제출함
2020년 5월

서울대학교 대학원
의과학과 의과학 전공
김 창 욱

김창욱의 의학박사 학위논문을 인준함
2020년 7월

위원장	윤 홍 택	
부위원장	김 종 일	
위원	서 정 언	
위원	문 인 희	
위원	안 정 석	

**1,779 whole-genome datasets unveil
population-specific genetic architecture
and pharmacogenomics profile
in Northeast Asian reference panel**

by

Chang-Uk Kim

A thesis submitted to the Department of
Biomedical Sciences in partial fulfillment of the
requirement of the Degree of Doctor of Philosophy
College of Medicine

July 2020

Approved by Thesis Committee:

Professor Hong Duk Youn Chairman

Professor JONG-IL KIM Vice chairman

Professor SEO JONG-SUN

Professor Mook, Inhee

Professor Jung-Hyuck Ahn

ABSTRACT

1,779 whole-genome datasets unveil population-specific genetic architecture and pharmacogenomics profile in Northeast Asian reference panel

Chang-Uk Kim

Major in Biomedical Science

Department of Biomedical Science

Seoul National University Graduate School

Introduction: Whole-genome sequencing (WGS), an important technique in genome research, is becoming bigger the number of subjects thanks to both the increase of sequencing capacity and the decrease of sequencing cost. Large scale WGS for specific human populations with deep depth coverage is necessary to study population genomics. Moreover, the need for large-scale deep WGS datasets is emerging to precisely understand the pharmacogenomics profile for precision medicine in Northeast Asia in line with the global trend. However, most of the WGS studies are currently biased to Europe.

Methods: We constructed the Northeast Asian Reference Database (NARD) using whole-genome sequencing data of 1,779 individuals from Korea, Mongolia, Japan, China, and Hong Kong. The NARD provides the genetic diversity of Korean and Mongolian ancestries that were not present in the 1000 Genomes Project Phase 3 (1KGP3). We re-phased the genotypes merged from the NARD and the 1KGP3 to construct a more robust union set of haplotypes.

Mongol and Korean samples have never been released on the scale and the depth of the NARD level. It is expecting to shed light on novel and accurate insights to population genomics. To investigate the population structure, we performed PCA analysis, the fixation index (F_{ST}) analysis, phylogenetic tree construction, and ADMIXTURE analysis.

We also tried to reveal the pharmacogenetic characteristics of Northeast Asians. We looked at various types of variants specific to Northeast Asians, the single nucleotide polymorphisms (SNPs) related to drug responses including rs116855232 in *NUDT15*, the SVs including *BCL2L11* (*BIM*) intronic deletion, and the HLA haplotypes related to the responsiveness of immune checkpoint blockade (ICB) therapy.

Results: The re-phasing approach we used to enhance the panel merged of the NARD and the 1KGP3 established a robust imputation reference panel for Northeast Asians, which yields the greatest accuracy in the genotype imputation especially for rare and low-frequency variants of Northeast Asians compared to the existing panels.

Population genomics analyses demonstrated the significant differentiation among Koreans, Mongolians, Japanese, and mainland East Asians (Chinese and Southeast Asians), in contrast to previous studies that highlighted the close genetic relationships in Northeast Asian populations.

The NARD variants catalog covered 14.8 million novel SNPs, which is improving the disease-related variants discovery by reducing the potential pathogenic candidates with common frequency redefined from rare frequency. Pharmacogenomics profiling suggested that the inefficiency of tyrosine kinase and the inhibition of immune checkpoint prevailed in Northeast Asians.

The workbench of the imputation pipeline with the NARD panel is available at <https://nard.macrogen.com/>.

Conclusions: We constructed the most accurate genotype imputation panel for Northeast Asian with public availability. We also unveiled the detailed Northeast Asian population structure and pharmacogenomic observations. Our work will contribute to further studies into the era of precision medicine for not only Northeast Asian but also the global population.

*This work is published in Genome Medicine (1).

Keywords: Population genomics; Pharmacogenomics; Reference panel; Genotype imputation; Whole-genome sequencing; Northeast Asians; East Asians

Student number: 2015-31234

CONTENTS

Abstract	i
Contents	iv
List of Tables	v
List of Figures	vi
List of Abbreviations	ix
Introduction	1
Whole-genome sequencing for human genomics.....	2
Genotype imputation with a population-specific reference panel	6
Population genomics based on whole-genome sequencing.....	9
Pharmacogenomics and precision medicine.....	12
Material and Methods	14
Results	29
Discovery of genetic variants including SNPs, indels, and structural variations	30
Population genomics analyses to reveal genetic architecture.....	33
Imputation accuracy improved with the NARD reference panel	37
Pharmacogenomics for precision medicine in Northeast Asians	44
Discussion	84
References	89
Abstract in Korean	103

LIST OF TABLES

Table 1. The total number of variants in 1,779 Northeast Asians by MAF and functional category.....	48
Table 2. The basic statistics of structural variations.....	49
Table 3. Imputation performance according to types of reference panel	50

LIST OF FIGURES

Figure 1. The cross-validation error inferred by ADMIXTURE algorithm	28
Figure 2. Geographic map of the study area in the NARD.	51
Figure 3. Correlation between the sequencing depth and the number of variants.	52
Figure 4. Pearson correlation coefficient (R) was calculated excluding the two samples with an abnormal heterozygous/homozygous ratio.	53
Figure 5. Transition to transversion ratio of the populations in the NARD	54
Figure 6. Heterozygous to homozygous ratio of the global populations.....	55
Figure 7. The number of loss-of-function variants.....	56
Figure 8. Hardy-Weinberg Equilibrium test of variants in the NARD.....	57
Figure 9. The number of novel SNPs that were not identified elsewhere	58
Figure 10. Distribution of novel SNP per population.....	59
Figure 11. Distribution of novel SNPs based on RefSeq gene definition	60
Figure 12. The size distribution of ascertained SVs.....	61
Figure 13. The distribution of variant allele counts of structural variation.....	62
Figure 14. MAF differences of SNPs shared between the NARD and the gnomAD	63
Figure 15. The F_{ST} network among Asian populations of the NARD and the 1KGP3.....	64

Figure 16. PCA of global populations from the NARD and the 1KGP3	65
Figure 17. PCA of Northeast and Southeast Asians from the NARD and the 1KGP3.....	66
Figure 18. PCA of Northeast and Southeast Asians from the NARD and the 1KGP3.....	67
Figure 19. Population substructure of Northeast and Southeast Asians with five ancestral components inferred by ADMIXTURE algorithm.....	68
Figure 20. Population substructure of MNG with five ancestral components inferred by ADMIXTURE algorithm	69
Figure 21. The maximum likelihood trees generated by TreeMix	70
Figure 22. Imputation accuracy assessment using the five different reference panels in KOR individuals.....	71
Figure 23. Imputation accuracy assessment using the five different reference panels in CHN individuals.....	72
Figure 24. Imputation accuracy assessment using the five different reference panels in JPN individuals.....	73
Figure 25. Imputation accuracy assessment using the five different reference panels in FRA individuals.....	74
Figure 26. The number of imputed SNPs as a function of the estimated imputation accuracy and the types of imputation panel.....	75
Figure 27. Imputation performance evaluation of CHN individuals.....	76
Figure 28. Imputation performance evaluation of JPN individuals	77
Figure 29. Length distribution of shared IBD tracts between the two individuals in each population	78

Figure 30. The flow chart of the pipeline consisting of four major steps for the NARD imputation server.....	79
Figure 31. The number of uncommon (MAF < 5%) protein-altering variants (missense, nonsense, frameshift, and splicing variants) after filtration using the gnomAD with/without the NARD	80
Figure 32. Pharmacogenomic analysis.....	81
Figure 33. Distribution of HLA class I subtypes HLA-B*15:01 associated with ICB response in the NMDP database	82
Figure 34. Distribution of HLA class I subtypes HLA-B44 associated with ICB response in the NMDP database.....	83

LIST OF ABBREVIATIONS

NGS: next generation sequencing

WGS: whole-genome sequencing

WES: whole exome sequencing

GWAS: genome-wide association study

1KGP3: The 1000 Genomes Project Phase 3

HRC: Haplotype Reference Consortium

EUR: European

CHN: Chinese

JPN: Japanese

MNG: Mongolians

KOR: Koreans

HKG: Hong Kong

NARD: Northeast Asian Reference Database

SNP: single nucleotide polymorphism

indel: insertion and deletion

LPS: low-pass sequencing

IBD: identity-by-descent

F_{ST} : the fixation index

PharmGKB: The Pharmacogenomics Knowledgebase

IRB: institutional review board

DRAGEN: Dynamic Read Analysis for GENomics

VQSR: variant quality score recalibration

LCR: low complexity region

Kaviar: Known VARiants

gnomAD: The Genome Aggregation Database

ExAC: Exome Aggregation Consortium

dbSNP: The Single Nucleotide Polymorphism Database

loftee: Loss-Of-Function Transcript Effect Estimator

PolyPhen-2: Polymorphism Phenotyping v2

MAF: minor allele frequency

SV: structural variation

dbNSFP: The Database for Nonsynonymous SNPs Functional Predictions

VCF: variant call format

DGV: Database of Genomic Variants

HWE: Hardy-Weinberg equilibrium

R : Pearson correlation coefficient

HLA: human leukocyte antigen

NMDP: The National Marrow Donor Program

Het: heterozygous

Hom: homozygous

Ti: transition

Tv: transversion

PCA: principal component analysis

CHB: Han Chinese in Beijing, China

CHS: Southern Han Chinese

CDX: Chinese Dai in Xishuangbanna, China

KHV: Kinh in Ho Chi Minh City, Vietnam

GIH: Gujarati Indian in Houston, Texas

BEB: Bengali in Bangladesh

ITU: Indian Telugu in the UK

PJL: Punjabi in Lahore, Pakistan

STU: Sri Lankan Tamil in the UK

BUR: Buryats

KHA: Khalkha Mongols

OTH: other Mongolians including Barga, Daringanga, Kazakh, Khoton, Uuld,
Durvud, Khotogoid, and Zakhchin

FRA: French

CEU: Utah Residents (CEPH) with Northern and Western European Ancestry

TSI: Toscani in Italia

FIN: Finnish in Finland

GBR: British in England and Scotland

IBS: Iberian Population in Spain

JPT: Japanese in Tokyo, Japan

YRI: Yoruba in Ibadan, Nigeria

LWK: Luhya in Webuye, Kenya

GWD: Gambian in Western Divisions in the Gambia

MSL: Mende in Sierra Leone

ESN: Esan in Nigeria

ASW: Americans of African Ancestry in SW USA

ACB: African Caribbeans in Barbados

MXL: Mexican Ancestry from Los Angeles USA

PUR: Puerto Ricans from Puerto Rico

CLM: Colombians from Medellin, Colombia

PEL: Peruvians from Lima, Peru

AFR: African

AMR: Ad Mixed American

EAS: East Asian

SAS: South Asian

gnomAD-ALL: worldwide populations from the gnomAD

gnomAD-EAS: East Asians from the gnomAD

BIM: Bcl-2-like protein 11

ICB: immune checkpoint blockade

Introduction

Whole-genome sequencing for human genomics

Next generation sequencing (NGS) based on massively parallel sequencing has completely changed the paradigm in human genomics. NGS has enabled whole-genome sequencing (WGS) approach decoding entire regions of the genomes, which is compared to the other techniques that decode only the small parts of the targeted genomic regions (2). Whole exome sequencing (WES) technology targets most of the coding regions and microarray technology only targets the interesting regions discovered by previous studies. These coverages are enough for most applications. However, the leading research discovering novel markers or hidden structures need for targeting whole coverages not only previously defined regions.

WGS reads several times for each base pair on the entire genome to improve accuracy and to identify whole alleles for multiploidy, specifically diploid in the human genome. The deep coverage depth means the number of sequenced times on the specific locus of the genome is in the certain criteria (e.g. 10X ~ 20X is considered as the intermediate range and over 20x is considered as the deep range) (3). The size of genome data produced by WGS reaches hundreds of millions of base pairs. Handling the massive amount of the NGS data also requires the high-performance computational equipment

During the past decade, the reference panels with population-scale WGS have enabled extensive human genetic research (4-11). They have played an

imperative role in human genetic research, especially for clinical variant interpretation and the genotype imputation in complex genome-wide association study (GWAS) (4-6, 9, 10). The most used the imputation panels were constructed by the 1000 Genomes Project Phase 3 (1KGP3) and Haplotype Reference Consortium (HRC) studies, which are publicly available for researchers. As the genotype imputation is an essential step to increase the power of GWAS in a cost-efficient way, the confidence of imputed genotypes is most important for human genetic studies. To improve the quality of the genotype imputation, the large-scale population-specific reference panels with deep sequencing coverage are mandatory. Moreover, it is essential for a better understanding of the population structure and the demographic history (7-9, 11). Accordingly, several research groups have generated the large-scale WGS data to build their population-specific reference panels (6, 7, 9, 10, 12-15).

Huang et al. (6) created a reference panel with low depth WGS from 3,781 British individuals. They achieved large increases in imputations accuracy by the re-phasing approach for integrating with the 1KGP3 haplotypes. Bai et al. (7) created a reference panel with intermediate depth WGS from 175 Mongolian individuals representing six tribes. They identified alleles shared between Finns and Mongolians/Siberians. Genome of the Netherlands (9) created a reference panel with intermediate depth WGS from 250 Dutch parent-offspring families. They discovered 20.4 million SNPs and 1.2 million indels. Gudbjartsson et al. (10) created a reference panel with intermediate depth WGS from 2,636 Icelanders. They found 20 million SNPs and 1.5

million indels. Hou et al. (12) created a reference panel with deep depth WGS from 265 individuals of Amish and Mennonite ancestry. They discovered 12 million SNPs and indels with free imputation server availability. Mitt et al. (13) created a reference panel with deep depth WGS from 2,244 Estonian individuals. Nagasaki et al. (14) created a reference panel with deep depth WGS from 1,070 Japanese individuals. They found 21.2 million SNPs and 3.4 million indels. Sidore et al. (15) created a reference panel with low depth WGS from 2,120 Sardinians. They discovered 17.6 million variants and assessed the impact on GWAS studying circulating lipid levels and inflammatory.

Despite Northeast Asians account for 21.5% of the worldwide population (<http://www.worldometers.info/world-population/>), most of the genetic studies and reference panels are biased to European ancestries (EUR) (16). There are some population-scale studies for building reference panels of Han Chinese (CHN), Japanese (JPN), Mongolians (MNG), and Koreans (KOR), but several issues, including public unavailability (7, 14, 17, 18), inadequate sequencing coverage (18, 19), small sample size (7, 20), and restriction to exonic regions (21, 22), need to be resolved for the solid imputation reference panel.

Therefore, constructing a large-scale whole-genome reference panel covering the diverse population groups in Northeast Asia with deep sequencing coverage is still required to allow dense and accurate genotype imputation for the genetic research in these populations.

We present the Northeast Asian Reference Database (NARD), consisting of 1,779 individuals from KOR (n = 850), MNG (n = 384), JPN (n = 396), CHN (n = 91), and Hong Kong (HKG, n = 58) with deep (n = 834, $\geq 20X$) or intermediate (n = 947, $10X \sim 20X$) sequencing coverages. The goal of this study is to provide highly accurate imputation panel, to clarify the genetic architecture of Northeast Asians, to unveil the pharmacogenomics profile, and to establish a high-quality population-specific reference panel for the genetic studies and precision medicine in Northeast Asia.

Genotype imputation with a population-specific reference panel

The genotype imputation estimates unknown variants including single nucleotide polymorphisms (SNPs) and insertions and deletions (indels) using the surrounding known markers by the statistical method based on the characteristics that the groups of near SNPs are preserved when the haplotype including the SNPs is inherited (23). Through the genotype imputation, it is possible to obtain the extended variants from the input marker set of SNPs and indels produced with low coverage technology that can read hundreds of thousands of SNPs rapidly in a cost-effective way such as microarray and low-pass sequencing (LPS). The genotype imputation can be utilized in GWAS to effectively conduct large and complex studies by increasing the density of the variants. The genotype imputation is also applicable to solve the problem of the experimentally untyped variants.

As mentioned above, the genotype imputation estimates the untyped genotypes based on the nearby genotype information. Identity-by-descent (IBD) provides the basic principle for the genotype imputation and haplotype phasing. The IBD block is a part of the chromosome consist of the same genomic contents in different individuals because the origin of the part is inherited from the same ancestor. As variants on the IBD are identical, knowing only a small set of the determinant markers allows identifying the remaining genotypes on the specific IBD. The reference panel, the repository

of the IBD blocks, is essential for the genotype imputation. The bigger the size of the panel and the nearer the distance of the panel to target individuals make more chance to match appropriate IBDs enabling more accurate imputation.

The genotype imputation tends to be easy to estimate accurately on common variants but hard to estimate on rare or low-frequency variants, due to the lack of information for rare or low-frequency variants in the reference panel. The accuracy of the genotype imputation is influenced by various factors including the reasons related to the reference panel. The factors influencing the accuracy of the genotype imputations are:

- 1) the imputation algorithm,
- 2) the number of markers from the source, which is microarray or LPS,
- 3) the representativeness of the markers in the haplotypes,
- 4) the number of samples in the reference panel, and
- 5) the homology of ethnicity between the samples in the reference panel and the imputation target individuals.

The fourth and fifth factors mentioned above are related to the reference panels. There was no panel high level enough to be used for precision medicine in Northeast Asia to date. The NARD is a robust reference panel, providing the most accurate genotype imputation for Northeast Asians.

Several representative tools are implementing the imputation algorithm with different advantages. Minimac3 (24) utilizing the repeat haplotype patterns

makes reference panel into “m3vcf” file that is own simplified variant call format (25) to reduce the size to optimization increasing the loading speed and reducing the memory consumption. IMPUTE2 (26) is based on the re-phasing approach that statistically estimates the haplotype and imputes untyped genotypes. Beagle (27) uses the haplotype frequency model proposed by Li and Stephens, which reduces the computational burden. Beagle also provides the result of IBD segment detection.

Population genomics based on whole-genome sequencing

Population genomics utilizes the genomics to understand population genetics, providing more accurate and novel insights compared to the previous approaches considering the limited genetic clues. Technological advancements have made it possible to decode the whole genomes of lots of samples. This massive decoded dataset enables a precise and in-depth understanding of even microevolution.

Gel electrophoresis and restriction enzyme mapping were methods for previous population genomics (28). The information lacks due to the limitation of the number of individuals and the region of the targeted genome allowed by the previous approaches limited the accuracy and range of facts of the result of the research (29). Non-neutral phenomena in the genome are the features related to evolutionary effect or population specificity. WGS approach changes the paradigm from looking for specific interesting regions to scanning whole genomic regions to find the candidate features. Analyzing tools for population genomics are being developed continuously that include the principal component analysis (PCA), the fixation index (F_{ST}) (30), Treemix (31), and ADMIXTURE (32).

PCA is one of the multivariate analysis techniques used in the various domains to reduce multi variations into a small number of features without any annotations. PCA is unsupervised learning that can serve the plot from

high dimensional data into the low-level dimensional space recognizable to humans. Through calculating the data covariance matrix and then performing eigenvalue decomposition on the covariance matrix, PCA extracts the principal features from the original data. PCA reveals the population structure based on lots of individuals from different populations. Moreover, in the QC context, PCA can be used to define the erroneous sample based on the abnormal result.

F_{ST} is an implementation of Wright's F-statistics, which calculates the distance between populations through the genotype data such as SNPs (30). Most of the application uses the variance of the allele frequency between populations.

Treemix is a tool for statistical model inferring the patterns of population splits and mixtures in multiple populations (31). Treemix reveals the population structure by finding the common ancestors and migration events in the phylogenetic tree using lots of SNPs from even WGS. The analysis provides from the overview structures to detailed structures, in the population relationships.

ADMIXTURE reveals the stratification of populations and estimates the individual ancestry from the SNPs decoded with WGS (32). The result of ADMIXTURE is useful to study population genetics and genetic epidemiology. The tool analyses the population structure through cross-

validation or calculates the ancestry estimation by supervised learning with the datasets including well-defined ancestries.

Understanding of where we come from and how we related to specific ethnicities is not just for interesting, but also important in that it gives us accurate evolutionary insights and public health perspectives. The NARD enables population genomic studies at unprecedented high resolution for Northeast Asians with deep coverage and large scale WGS datasets.

Pharmacogenomics and precision medicine

Pharmacogenomics is a compound word of Pharmacology and Genomics, which studies the differences in drug responses under different genetic conditions. Understanding the genetic differences in individual responses to medications reveals the risk of side effects caused by applying the “one size fits all” approach. Many factors determine the differences, one of which is due to genetic innateness. The research into the genetic cause of drug reactions through genomics is important, and its application to clinical practice in this paradigm is important for precision medicine.

Pharmacogenomics database such as the Pharmacogenomics Knowledgebase (PharmGKB; <https://www.pharmgkb.org/>) (33), PharmVar (<https://www.pharmvar.org/>), SuperCYP Bioinformatics Tool (<http://bioinformatics.charite.de/supercyp/>), FINDbase (<http://www.findbase.org/>), Pharmacogenomics Biomarkers in Drug Labelling (<http://www.fda.gov/drugs/scienceresearch/researchareas/pharmacogenetics/cm083378.htm>), and Pharmacogenomics Research Network (<http://www.pgrn.org/>) can be used for clinical screening, by providing personalized drug responses. An individual who knows about his/her drug responses can be prescribed with the correct drug with adequate dosage. If a population knows the drug-related adverse effect caused by certain SNPs prevailed in the population, the guidelines can be made to consider testing the SNPs for public health perspective.

Precision medicine separates the population into subpopulations composite of individuals who have different features that must be treated with different medicine. New generation precision techniques today allow classifying the groups of individuals for the purpose. As one of the techniques allowing precision medicine, NGS decodes the genome of individuals with a feasible cost. The personal genome is different for each individual. The genome can determine the inherited disease to cancer for personal.

The NARD is an important dataset that has significant quantitative and qualitative meaning in Northeast Asia as the robust imputation reference panel and the population genomics database including diverse populations underrepresented in existing databases. Furthermore, the NARD will make a meaningful contribution to the research and application of precision medicine in Northeast Asia.

Material and Methods

1. Ethics statement

This study was approved by the institutional review board (IRB) of Seoul National University Hospital, in accordance with the Declaration of Helsinki (approved ID: C-1705-048-852). Written informed consents were obtained from all study subjects.

2. Whole-genome sequencing for 1,690 samples

For 1,690 individuals of KOR, JPN, MNG, and HKG, deep depth WGS was performed at Macrogen (Seoul, Korea). DNA libraries composite of the size of about 500 bp were sequenced using Hiseq X instrument (Illumina, San Diego, CA) as paired-end 100 base reads based on the manufacturer's instructions. We also included publicly available 91 CHN samples (34), which have been sequenced by Illumina Hiseq 2000 instrument (Illumina, San Diego, CA). This cohort includes YH cell line and samples from the HapMap and the 1KGP3 with deep sequencing depth (on average, 70X) (4, 34, 35).

3. Variant discovery of SNPs and indels

Read alignment to the human reference genome (hg19) without the alternate contigs, duplicate read removal, and joint calling of SNPs and indels were performed using Dynamic Read Analysis for GENomics (DRAGEN) platform

(version 01.003.024.02.00.01.23004; <http://edicogenome.com/dragen-bioit-platform/>) with the following parameters:

- 1) creating gVCF: “--enable-map-align-output true,” “--remove-duplicates true,” “--enable-bam-indexing true,” “--enable-variant-caller true,” and “--vc-emit-ref-confidence GVCF,” and
- 2) joint calling: “--enable-joint-genotyping true.”

For indels, we discarded variants that greater than or equal to 50 base pairs, which are classified as structural variants in general (7, 36). Variant quality score recalibration (VQSR) was applied to raw variants based on the GATK’s best practice (<https://software.broadinstitute.org/gatk/best-practices/>) with the parameters given below:

1) Annotations

SNP: DP, QD, MQ, MQRankSum, FS, SQR

Indel: DP, QD, MQ, MQRankSum, ReadPosRankSum, FS, SOR

2) Truth set

SNP: HapMap3.3 and 1KPG Omni2.5

Indel: Mills & 1KGP gold standard

3) Training set

SNP: HapMap3.3, 1KGP Omni2.5 and 1KGP phase 1 high confidence

Indel: Mills & 1KGP gold standard and 1KGP phase 1

4) Known set

SNP: dbsnp138

SNPs and indels below 99% of truth sensitivity levels from VQSR were initially filtered. Moreover, recalibrated variants were further filtered based on the following criteria:

- 1) located in the low complexity regions (LCRs) which were defined by the 1KGP3 (4),
- 2) genotype quality < 20, and
- 3) read depth < 5.

After these filtration processes, SNPs and indels were phased by SHAPEIT3 (version r884.1) which provides a fast population-scale phasing with low switch-error using the following model parameters: "--states 100," "--window 2," and "--effective-size 15000." This step also includes the imputation of missing genotypes.

4. Calculate the concordance of overlapped variants between the NARD and publicly available chip data for quality check

To validate variants in the NARD, we selected the 86 CHN samples in the NARD which have publicly available Illumina Omni 2.5M array data from the 1KGP3. 1,664,330 SNPs in total were overlapped between the NARD and the Omni chip, excluding mitochondrial DNA and pseudoautosomal regions, due to the worthlessness and problematic for these regions. The concordance is the cumulative sum of the matching alleles divided by the

total number of loci multiplied by two that is the maximum matching opportunity in a diploid.

$$\text{concordance} = \frac{\sum \text{matched alleles}}{\text{total number of loci} \times 2}$$

The variant call results for sex chromosomes in males are always represented as homozygous in non-error situations. For the uniformity, the sex chromosomes in males are considered as diploids for this calculation.

5. Annotation for SNPs and indels using ANNOVAR

All the SNPs and indels in this study were annotated by ANNOVAR based on RefSeq gene definition (37, 38). For novel variant classification, Known VARiants (Kaviar; <http://db.systemsbiology.net/kaviar/>), the Genome Aggregation Database (gnomAD; <http://gnomad.broadinstitute.org/>), the Exome Aggregation Consortium (ExAC), and the Single Nucleotide Polymorphism Database (dbSNP) build 150 were annotated (39-41). For loss-of-function variant annotation, we performed the Loss-Of-Function Transcript Effect Estimator (loftee; version 0.3-beta) (42) which is a plugin of Variant Effect Predictor (43) to remove low confidence annotations (44) with the following parameters: “--pick,” “--vcf,” “--cache,” “--offline,” and “--plugin LoF.” For the 1KGP3 dataset, we also removed the variants within LCR. Moreover, the functional impact of variants was measured by Polymorphism Phenotyping v2 (PolyPhen-2), which were embedded in the Database for Nonsynonymous SNPs Functional Predictions (dbNSFP) (45, 46). In the

pharmacogenomic analysis, we investigated SNPs that were annotated as drug response in Clinical significance on ClinVar (47). Then, we selected SNPs that were curated by the PharmGKB. SNPs that common but rare in non-Finnish Europeans of the gnomAD were considered as Northeast Asian specific variants. Researchers can download MAFs of variants in the NARD as `ANNOVAR` format (https://nard.macrogen.com/download/NARD_Annovar.zip).

6. Structural variation discovery using Delly

We used Delly (version 0.7.8) (48) to call structural variations (SVs) from 1,690 individuals in the NARD excluding CHN due to the composition of libraries that consists of multiple insertion sizes disturbing SV call (49). It was executed according to the germline SV calling procedure described in the manual (<https://github.com/dellytools/delly/>) with the default parameters.

To start Delly execution, we take the BAM files created by DRAGEN into BCF files by the calling method of Delly with the human genome reference using the below command:

```
$ delly call -x hg19.excl -o delly.bcf -g hg19.fa input.bam
```

The “hg19.excl” file defines the sex chromosomes and the centromeres of autosomes that will be excluded in the calling process. Then, creating the variant call format (VCF) files from the BCF files is performed with BCFtools using the below command:

```
$ bcftools view delly.bcf > delly.vcf
```

Multiple BCF files per each sample created by above works are merged using the below command:

```
$ delly merge -o sites.bcf *.bcf
```

Genotyping is performed with the merged “sites.bcf” file with the call method of Delly per sample BCF file using the below command:

```
$ delly call -g hg19.fa -v sites.bcf -o geno.bcf -x hg19.excl input.bam
```

All genotype bcf files are merged with the merge method of Delly using the below command:

```
$ bcftools merge -m id -O b -o merged.bcf *.geno.bcf
```

Delly provides germline filter method, which is performed using the below command:

```
$ delly filter -f germline -o result.bcf merged.bcf
```

Only variants longer than or equal to 50 base pairs length are considered as SV by the in-house script.

7. Annotation of structural variation for the novelty determination and the functional classification

We annotated novel SVs with the Database of Genomic Variants (DGV) and the 1KGP3 (4, 50), under the criteria that SVs from our study and other database were overlapped at least 20% reciprocally.

$$\begin{aligned} \text{Overlapped}(A, B) &= (\text{Min}(A_{end}, B_{end}) - \text{Max}(A_{start}, B_{end})) \\ &\geq \text{Min}(A_{length}, B_{length}) \times 0.2 \end{aligned}$$

We additionally used RefSeq (37) to annotate SVs with gene names and the functional regions including CDS, exonic, intronic, or intergenic. We classified Northeast Asian specific SVs that are common in the NARD but rare in non-Finnish Europeans of the 1KGP3 (4). We except the Finnish in Europe in this analysis because the ancestry of Finnish has migration flow from MNG, the population of Northeast Asia (7). Then, we kept only SVs that common in East Asians of the 1KGP3 to eliminate potential batch effects derived from SV detection software between two studies.

8. F_{ST} analysis to measure the genetic distance among populations

To measure the genetic distance between Northeast Asian populations and the other populations in the 1KGP3 (4), we calculated the pairwise Weir and Cockerham weighted F_{ST} (30) using VCFtools (version 0.1.16) (25, 34). The evaluation calculation for F_{ST} is

$$F_{ST} = \frac{\sigma_S^2}{\bar{p}(1 - \bar{p})}$$

, where \bar{p} is the average allele frequency of the whole population, and σ_S^2 is the variance in the frequency of the allele between different subpopulations. To visualize the F_{ST} network in the map, the opacity and thickness of the edges between the populations were determined by the intensity calculated as:

$$Intensity = \left(\frac{1 - F_{ST}}{F_{ST_{max}}} \right)^2$$

The intensity of the edges of the network graph on the global map indicates the affinities between nodes representing each population. The graph reveals the structure of both the grouping and the isolation of populations. The related populations will be grouped into the population group that linked with strong edges. The unrelated population groups will be isolated for each other, which is represented by the weak edges among the nodes of isolated groups.

9. Hardy-Weinberg Equilibrium calculation for SNPs to confirm the quality of variant discovery

To confirm the quality of SNPs, we calculated Hardy-Weinberg Equilibrium (HWE) of variants in the NARD using VCFtools (version 0.1.12b) with “--hardy” option (51). Following the HWE test, we plotted the distribution graph to inspect the shape of the numbers of the variants by HWE P-values. Especially the amount of the variants under the 10^{-5} of P-value. All the variants under the significant threshold of P-value are not only caused by error but also by the population-specific evolutionary effect including genetic drift, mate choice, assortative mating, natural selection, sexual selection, mutation, gene flow, meiotic drive, genetic hitchhiking, population bottleneck, founder effect and inbreeding (52). 99.2% of the variants passed HWE test.

10. Population structure analysis by PCA and ADMIXTURE

We converted VCF files of bi-allelic autosomal SNPs from both the NARD and the 1KGP3 into PLINK format using GotCloud (version 1.75.5) (4, 53, 54). Then, we merged two panels by PLINK (version 1.9) (53), and extracted SNPs with genotype rate are equals to 100% and MAF greater than or equals to 1%. Finally, we pruned SNPs with linkage disequilibrium ($R^2 > 0.1$) within 50 base pairs by sliding window using PLINK. With these processed data, we carried out PCA using Genome-wide Complex Trait Analysis (version 1.91.3beta) (55) for two combinations:

- 1) Northeast Asians of the NARD and worldwide populations of the 1KGP3 and
- 2) Northeast Asians of the NARD and East Asians of the 1KGP3, separately.

We also applied the unsupervised ADMIXTURE algorithm (version 1.3) (32) using cross-validation for the population structure analysis. The optimal number of clusters was determined by comparing K values with cross-validation error rates (**Fig. 1**). The number of ancestries was varied from $K=2$ to $K=10$. The results of ADMIXTURE analysis were visualized by Genesis (<http://www.bioinf.wits.ac.za/software/genesis/>).

11. Phylogenetic tree to reveal the population structure and the migration events among populations

The maximum likelihood trees were generated by TreeMix (version 1.13) (31) allowing to infer the migration events. The allele frequency stratification for population extracted from bed file by PLINK is performed using the below command:

```
$ plink --noweb --bfile input_bed_file --freq --out output
```

Following building the stratification, the stratification is compressed with gzip to “output.freq.strat.gz” file. Completed the compression, stratification file is converted to “output.treemix.freq.gz” file by “plink2treemix.py” program using the below command:

```
$ plink2treemix.py output.strat.freq.gz output.treemix.freq.gz
```

Finally, TreeMix is performed using the following command:

```
$ treemix output.treemix.freq.gz -o result -root AFR
```

Drawing the result to the pdf file is performed with R script “plotting_funcs.R” in the source of TreeMix.

12. Genotype imputation accuracy and reference panel building

For building the imputation panel, the singleton variants in the NARD were excluded, because they may be the novel mutation for an individual. To merge the NARD and the 1KGP3 panels, we used the same approach as the UK10K and IMPUTE2; NARD-specific variants were imputed into the 1KGP3 using Minimac3 (version 2.0.1) and vice versa. Then they were merged into a single reference panel. Besides, the merged panel was re-

phased by SHAPEIT3 using the model parameters mentioned above with “--early-stopping” and “--cluster-size 4000” parameters. We kept variants that are not located in LCR.

We separately processed 113 KOR (20, 56, 57), 79 CHN, 27 JPN (41, 58), and 24 French (FRA) (59) individuals that are not included in the reference panels for the genotype imputation accuracy evaluation. Then, we discarded 16 related individuals from a KOR cohort. Unrelated sample selection was achieved by kinship estimation using KING (60). Then, we extracted SNPs from sites on the Illumina Omni 2.5M array and monomorphic sites were excluded. As a result, 1,345,511, 1,320,123, 1,214,151, and 2,847,580 autosomal SNPs remained in the pseudo-GWAS panels of KOR, CHN, JPN, and FRA cohorts, respectively.

We performed the genotype imputation using Minimac3 with the five different types of reference panels. The genotype imputation using the HRC panel was performed at the Michigan Imputation Server (<https://imputationserver.sph.umich.edu/>). Before the imputation, the haplotypes of individuals in the four cohorts were estimated using Eagle2 (version 2.3.2). After imputation, we extracted 4,352,921, 5,427,462, 48,431,56, and 5,419,512 SNPs in the four cohorts, which were imputed by all reference panels, and none of them were present with missing genotype in the non-masked dataset. The squared Pearson correlation coefficients (R^2) were calculated between the imputed dosages and true genotypes, and those values

were aggregated into 11 MAF bins to measure the imputation accuracy. The 11 MAF bins are 0% ~ 0.2% for very rare, 0.2% ~ 0.5% for rare, 0.5% ~ 1%, 1% ~ 2%, and 2% ~ 5% for low-frequency, and 5% ~ 10%, 10% ~ 20%, 20% ~ 30%, 30% ~ 40%, 40% ~ 50%, and 50% ~ 100% for common.

13. HLA typing for profiling HLA distribution in Northeast Asia

We determined the human leukocyte antigen (HLA) class I haplotype of 1,779 individuals by xHLA (version 1.2) (61). To conduct a comparative analysis, we downloaded the HLA class I genotype of European populations from 1KGP [FTP](ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/working/20140725_hla_genotypes/) server (ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/working/20140725_hla_genotypes/), which were studied by another group (62). The worldwide frequencies of HLA-B*15:01 and HLA-B44 supertype from the National Marrow Donor Program (NMDP) (63) were downloaded from xHLA (<https://github.com/humanlongevity/HLA/>).

To type the HLA haplotype, we aligned the rawdata FastQ files onto the human genome reference file using BWA with the below parameters:

```
$ bwa mem hg38.fa 1.fq.gz 2.fq.gz > output.sam
```

Then, SAMtools is performed to make the sorted BAM file using the below command:

```
$ samtools view -u output.sam | samtools sort > output.bam
```

Completed BAM file creation, BAM files are indexed with the below command:

```
$ samtools index output.bam
```

Finally, we performed the xHLA using the below command:

```
$ docker run -w `pwd` humanlongevity/hla --sample_id sample  
--input_bam_path output.bam --output_path result
```

The result of the xHLA is generated in JSON format. The result files are parsed with the in-house script to make an integrated tabular file.

14. IBD analysis to evaluate the effect of the re-phasing approach

The shared IBD segments between two individuals were identified using RefinedIBD (version 12Jul18.a0b) with “length=2.0” parameter like below command line (39):

```
$ java -Xmx256g --jar refined-ibd.jar length=2.0 gt=input.vcf.gz
```

To evaluate the effect of the re-phasing approach on haplotype correction, we performed this analysis using the original and re-phased haplotypes of the NARD which were phased without and with the 1KGP3 panel, separately. The short gaps and breaks (> 0.6 cM) between IBD segments were discarded using merge-ibd-segments utility program. The command performing the program is below:

```
$ cat phased.ibd | java -jar merge-ibd-segments.jar phased.vcf.gz  
constrecomb.map 0.6 2 > result.ibd
```

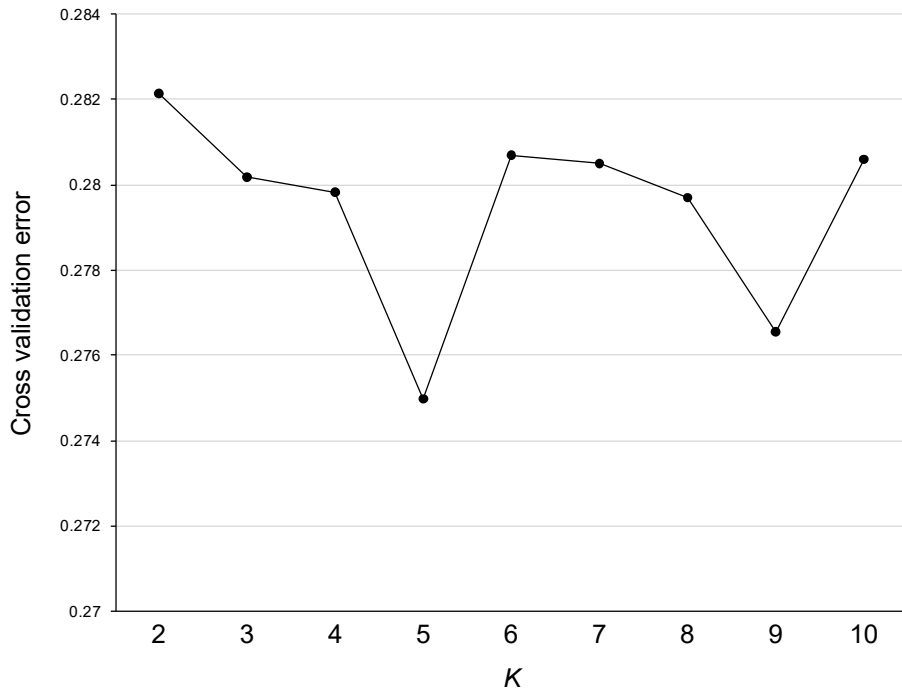


Figure 1 | The cross-validation error inferred by ADMIXTURE algorithm.

Results

1. Discovery of genetic variants including SNPs, indels, and structural variations

1.1 SNPs and indels

The NARD contains 1,779 Northeast Asians including KOR (n = 850), JPN (n = 396), MNG (n = 384), CHN (n = 91), and HKG (n = 58) with deep (20X \leq , n = 834) or intermediate (10X ~ 20X, n = 945) sequencing coverages (**Fig. 2**). The percentage of the KOR samples is 48%, compositing of a major proportion of the NARD dataset. JPN, MNG, CHN, and HKG are 22%, 22%, 5%, and 3% respectively. KOR and MNG, the uncommon populations in existing datasets to date, are 70% of the total. The proportion of the populations are affecting the numbers in every result.

Initially, WGS was performed on 1,779 Northeast Asians, but two MNG samples with low variant count and an abnormal ratio of heterozygous to homozygous genotypes (Het/Hom) were discarded in the downstream analysis (**Figs. 3-4**). Except for the two outliers, 1,779 samples are in the reasonable range of the criteria. We evaluated potential bias from inconsistent sequencing coverage of samples and found no significant correlation (Pearson correlation coefficient) between the sequencing depth and the number of variants: SNP ($R = 0.15$) and short indel ($R = -0.20$). Also, the transition to transversion (Ti/Tv) ratios was consistent across the samples

(2.1 on average; **Fig. 5**). The Ti/Tv of CHN is slightly smaller than of the other populations due to the difference in the sequencing platform, but no significant differences. The Het/Hom ratio is known to be ancestry dependent (64). The Het/Hom ratios (1.4 on average; **Fig. 6**) and the number of loss-of-function variants (35.4 on average; **Fig. 7**) in the NARD were similar to those in East Asians from the 1KGP3 (1.3 and 36.9 on average for each). The Het/Hom ratios and the number of loss-of-function variants are different in other population groups, but consistent with the related population in the 1KGP3. Also, 99.2% of the variants passed HWE test ($P > 1 \times 10^{-5}$; **Fig. 8**).

In the NARD, a total of 40.6 million SNPs and 3.8 million indels were discovered, and 77.1% were singletons or rare variants ($MAF < 0.5\%$; **Table 1**). On average, 3.3 million SNPs and 0.3 million indels were found for each individual. We identified 15.4 million novel SNPs (37.8% of the total) in the NARD (**Fig. 9**). Among them, 45.0% were specific to KOR, likely due to their large sample size in our dataset, and 12.6% were found across populations (**Fig. 10**). The other percentages of the novel SNPs for each population also reflected the numbers of populations. Most novel SNPs were singletons or rare variants and located in non-coding regions, supporting the purifying selection effect. (**Fig. 11**). We found the high integrity of our WGS variant call pipeline; the genotype concordance between WGS and Illumina Omni 2.5M array of 86 CHN samples in the NARD was 99.6%.

1.2 Structural variations

We discovered 93,764 SVs including 67,241 large deletions, 20,262 duplications, 4,664 inversions, and 1,597 large insertions, where 74.5% were novel and 77% were singleton or rare (**Table 2**). The reason that the major type of detected results is deletion is for the detection advantage for the deletion type of the algorithm exploited in Delly. The lengths of the SVs, particularly insertions, tend to be short also due to the limitation of the short read based approach (**Fig. 12**). NGS mapping algorithm aligns the short reads onto the reference based on the sequence similarity. Due to most of the insertions are positioned in repetitive regions, long sized insertions composed of similar motifs and cannot be accurately determined by the short read based approach.

More than half of the novel SVs (59.1% of the total) are shared among multiple populations (**Fig. 13**). Conversely, 40.9% of novel SVs are unique in a specific population. Since only looking at the ratio can cause misinterpretation, the absolute amount was shown in the figure for each section of the variant allele count. The proportion of the unique to a specific population is less than the result in the 1KGP3. This comparison is reasonable due to the genetic distances among our populations in the NARD is much closer than the distance among the super populations in the 1KGP3.

1.3 Redefinition of MAF

The frequencies of SNPs between the gnomAD and the NARD were compared. We redefined the frequency of 2.0 million SNPs that are rare in the gnomAD to low-frequency or common ($MAF \geq 5\%$). Moreover, 0.8 million rare SNPs in East Asian from the gnomAD were low-frequency or common variants in the NARD (**Fig. 14**). The redefinition of the allele frequency affects the interpretation of the SNPs in further studies, as rare SNPs can be considered as risk causing mutations. In this meaning, WGS for the underrepresented populations is important. In section 4.1, we show the effect of this redefinition in pharmacogenomics.

2. Population genomics analyses to reveal the genetic architecture

To examine the population structure of Northeast Asians, we conducted population genomics analyses for the NARD combining with the 1KGP3 data (4). F_{ST} (30) and PCA showed that the populations of the NARD were genetically close to East Asians of the 1KGP3, but distinctive from South Asians. Additionally, ADMIXTURE (32) and phylogenetic analyses supported the different ancestral components for each of KOR, MNG, JPN, and mainland East Asians.

2.1 F_{ST}

F_{ST} (30) network graph represents the two groups of populations (**Fig. 15**).

One group consists of MNG, KOR, JPN, HKG, Han Chinese in Beijing, China (CHB), Southern Han Chinese (CHS), Chinese Dai in Xishuangbanna, China (CDX), and Kinh in Ho Chi Minh City, Vietnam (KHV), and another group consists of Gujarati Indian in Houston, Texas (GIH), Bengali in Bangladesh (BEB), Indian Telugu in the UK (ITU), Punjabi in Lahore, Pakistan (PJI), and Sri Lankan Tamil in the UK (STU). The two super populations are related to Northeast Asia and South Asia, respectively. The super populations are strongly intra-connected, while inter-connection to each other is weak. The result represents a significant difference between Northeast Asian and South Asian.

2.2 PCA

We examined the ancestry composition of individuals in the NARD to illustrate how it covers the genetic diversity that was not present in existing reference panels. From the PCA result of global human populations with PC1 and PC2 explaining 1.0% and 0.5%, individuals from the NARD were closely related to East Asians from the 1KGP3 as expected (**Fig. 16**). MNG were separately clustered and positioned between East Asian and non-African populations as previously reported (7). When we applied PCA to only Northeast and Southeast Asians with PC1 and PC2 explaining 7.5% and 3.6%, a clear population differentiation pattern was observed among them (**Figs. 17-18**). MNG were most distinct from other populations based on PC1, and PC2 separated KOR, JPN, and mainland East Asians including CDX, CHB, CHS, HKG, and KHV. Interestingly, there were no overlapped samples between KOR and JPN except for a few outliers. This result implies that their ancestral compositions are distinctive enough to form separate clusters.

2.3 ADMIXTURE

Unsupervised ADMIXTURE analysis (32) supported the different ancestral components for each of KOR, MNG, JPN, and mainland East Asians (**Fig. 19**). In the case of MNG, there were Buryats (BUR, n = 299), Khalkha Mongols (KHA, n = 73), and other Mongolians including Barga, Daringanga, Kazakh, Khoton, Uuld, Durvud, Khotogoid, and Zakhchin (OTH, n = 12). They were

also genetically separated into BUR and KHA/OTH (**Fig. 20**). The results highlight that the NARD has the most diverse genetic compositions of Northeast Asian populations by adding the two ancestries, KOR and MNG, which have been underrepresented in current public datasets such as the 1KGP3 panel.

2.4 Phylogenetic Tree

Phylogenetic analysis using TreeMix (31) showed that MNG is the root of the populations in the NARD and supported the branch splitting KOR/JPN from mainland East Asians (**Fig. 21**). The in-depth analysis using the NARD revealed the fine-resolution population structure and emphasized the importance of the reference panel covering all populations across Northeast Asia. Moreover, the analysis of migration events ranged between one and five reveals the additional genetic relationships among populations.

3. Imputation accuracy improved with the NARD reference panel

Establishing an imputation reference panel for Northeast Asians was one of the main purposes of this study. We performed the imputation test on the simulated genotype data from the unrelated individuals of 97 KOR samples (20, 56, 57), 79 CHN samples, 27 JPN samples, and 24 FRA samples to evaluate the comparative accuracy of the NARD panel.

3.1 Imputation Accuracy

To illustrate the robustness of the NARD as an imputation reference panel, we built a pseudo-GWAS dataset using an independent cohort of 97 unrelated KOR individuals (20, 56, 57) and simulated the genotype imputation analysis. It was created from WGS data by masking the genotypes that were not included in the sites of Illumina Omni 2.5M array. Then, the genotype imputation was conducted by Minimac3 on pre-phased SNPs using five different types of reference panel.

HRC panel which is the largest in sample size showed poor imputation accuracy compared with other panels (**Figs. 22-24**), although the size of the reference panel is one of the major determinants of the imputation performance (9, 26). This might be due to the composition of the population skewed to European ancestry in HRC panel. Moreover, the NARD panel outperforms the 1KGP3 panel, even at very rare SNPs ($MAF < 0.2\%$). We

then merged the NARD and the 1KGP3 panels to further enhance the imputation performance. Consistent with HRC and UK10K (5, 6), we confirmed a large improvement in the imputation accuracy, particularly for very rare ($R^2 = 0.87$), rare [$\text{MAF} < 0.5\%$; $R^2 = 0.91$], and low-frequency ($0.5\% \leq \text{MAF} < 5\%$; $R^2 = 0.93$) SNPs, when the merged panel was re-phased by SHAPEIT3 (65).

The imputation was conducted by Minimac3 on the pre-phased SNPs using five types of reference panels:

- 1) NARD (n = 1,779),
- 2) 1KGP3 (n = 2,504),
- 3) HRC r1.1 (n = 32,470),
- 4) NARD + 1KGP3 (n = 4,200), and
- 5) NARD + 1KGP3 (re-phased, n = 4,200).

To measure the imputation accuracy, we calculated the R^2 between the true genotypes and the imputed dosages as a function of MAF in 850 KOR individuals from the NARD. The imputation performance of the NARD exceeded the 1KGP3 panel for every MAF bin (**Fig. 22**). Notably, the HRC panel, with the largest sample size including individuals from the 1KGP3, showed poor performance compared with other panels. Since the low imputation accuracy of the HRC panel is inconsistent with the original investigation, we performed the same analysis using 24 unrelated FRA individuals (59). In contrast to a KOR cohort, we confirmed that the HRC panel produced the most accurate genotype dosages for an FRA cohort, and

the NARD panel had poor suitability for Europeans (**Fig. 25**).

3.2 The improvement of imputation accuracy

We merged both the NARD and the 1KGP3 panels and performed re-phasing to enhance the imputation performance based on previous studies (5, 6). To merge the NARD and the 1KGP3 panels without missing genotypes, we used the same approach that was implemented in the UK10K (6, 23) and IMPUTE2. We reciprocally imputed two panels using Minimac3 to statistically infer the missing genotypes in the NARD or the 1KGP3 panels. Consistent with previous studies (7, 9, 12-14), combining two panels showed more accurate results of the imputation compared to the NARD or the 1KGP3 alone. Furthermore, we confirmed a large improvement of the imputation accuracy, particularly for very-rare ($MAF < 0.2\%$; $R^2 = 0.80$), rare ($0.2\% \leq MAF < 0.5\%$; $R^2 = 0.83$), and low-frequency ($0.5\% \leq MAF < 5\%$; $R^2 = 0.87$) variants, when the haplotypes in the merged panel were re-phased by SHAPEIT3 (65). In addition to measuring accuracy, we assessed the number of accurately imputed SNPs for each panel. For this analysis, we used the estimated R^2 values measured by Minimac3, as it is the standard for the quality control procedure in GWAS (66, 67). We found that the NARD + 1KGP3 (re-phased) panel produced the greatest number of high-confident SNPs ($R^2 \geq 0.9$) compared with other panels, especially 1KGP3 ($n = 7.5$ million versus 6.7 million), in concordance with the imputation accuracy (**Fig. 26**).

We also illustrated the potential of the NARD + 1KGP3 (re-phased) as a reference panel for diverse Northeast Asians by performing additional imputation tests using independent cohorts of unrelated CHN and JPN individuals ($n = 79$ and 27 , respectively) (41, 58). For measurement of the imputation accuracy, we used MAF bins defined by 10,639 CHN and 3,554 JPN individuals (17, 19). In agreement with the imputation result of a KOR cohort, the NARD + 1KGP3 (re-phased) panel provided the most accurate genotype imputation on very-rare ($R^2 = 0.71$ and 0.84 for CHN and JPN cohorts, respectively), rare ($R^2 = 0.71$ and 0.89 for CHN and JPN cohorts, respectively), and low-frequency ($R^2 = 0.81$ and 0.91 for CHN and JPN cohorts, respectively) variants (**Figs. 23-24**). The NARD + 1KGP3 (re-phased) panel also generated the largest number of accurately imputed genotypes compared with other panels, particularly the 1KGP3 ($n = 7.0$ million versus 6.8 million and 6.6 million versus 6.2 million for CHN and JPN cohorts, respectively; **Figs. 27-28**).

To investigate where the improvement of the NARD + 1KGP3 (re-phased) comes from, we divided the panel into the NARD (re-phased) and the 1KGP3 (re-phased) and assessed the imputation accuracy separately. The NARD (re-phased) panel had slightly lower imputation power than the NARD+1KGP3 (re-phased) panel, but greatly improved compared to the original NARD panel (**Table 3**). Meanwhile, the 1KGP3 (re-phased) panel showed no improvement in the imputation accuracy compared to the original 1KGP3 panel.

We examined the underlying reasons for improved the imputation performance caused by the re-phasing approach using identity-by-descent analysis. It is known that phasing or genotype errors cause the gaps within the real IBD tracts, hence the length of segments in phased genotype data tends to be shorter (68, 69). Based on this aspect, we expected that haplotype correction is occurred by re-phasing, and it would extend the length of shared IBD segments among individuals. Therefore, we measured the shared large IBD segments ($\geq 2\text{cM}$) between two individuals using the original (phased without the 1KGP3) and re-phased haplotypes of the NARD. As a result, we confirmed the significant increase in length and the number of shared IBD segments in re-phased haplotypes, which implies that the haplotype refinement in the NARD was achieved by the re-phasing process (Fig. 29).

3.4 Imputation server

The genotype imputations can be performed at the NARD imputation server for the academic purpose (<https://nard.macrogen.com/>). Also, researchers can download MAF data in the NARD as a VCF file (https://nard.macrogen.com/download/NARD_MAF.hg19.zip). The hg38 version of MAF data liftovered by CrossMap (version 0.3.6) (70) is also available (https://nard.macrogen.com/download/NARD_MAF.hg38.zip).

We developed a user-friendly web site to provide the imputation service using the NARD + 1KGP3 (re-phased) panel for researchers (**Fig. 30**). Our web site provides the imputation process for a wide range of genotype data format:

- 1) PLINK (ped or bed files paired with map or bim/fam files, respectively) (53),
- 2) 23andMe (Mountain View, CA) rawdata,
- 3) AncestryDNA (Lehi, UT) rawdata, and
- 4) VCF.

Results are processed through the imputation pipeline consisting of four major steps: pre-processing, phasing, imputation, and post-processing. The pre-processing step checks the format and content validity of uploaded files and converts them into VCF files for the next steps. Depending on the format of uploaded files, PLINK and 23andMe/AncestryDNA files will be converted into VCF files using GotCloud (54) and BCFtools (71), respectively, based on hg19 reference coordinate. When the input files have multiple chromosomes, the system will automatically separate them into multiple files for each chromosome. The subsequent analyses proceeded regardless of whether files have “chr” prefix in their contig names or not. The pre-processed data is phased using Eagle2 (72) or SHAPEIT2 (73), and Beagle5.0 with or without a reference panel, respectively. Then, the imputation is performed with Minimac4 (<https://github.com/statgen/Minimac4/>). In the post-processing step, the output is assessed and provided as gzip-compressed VCF and PLINK binary files. The server will provide the PLINK format with extra files containing predicted R^2 values per variant for the imputation quality check.

Once the imputation is finished, users will be notified by email and the result will be stored in the server for a week.

4. Pharmacogenomics for precision medicine in Northeast Asians

We evaluated the advantage of the NARD as a population-specific panel for clinical variant interpretation, as the exclusion of common variants is the first step to identify rare disease-causing genes (74). Next, we investigated Northeast Asian-specific pharmacogenomic features affected by different types of variants. We also cataloged Northeast Asian-specific SVs within genic regions. Finally, we examined HLA haplotypes to ascertain the prevalence of HLA related immunotherapy efficacy in Northeast Asia.

4.1 Filtration of rare variant for clinical purpose

Filtering common variants based on the population allele frequency is the first step to identify rare disease-causing genes (74). To examine the potential advantage of the NARD for clinical variant interpretation, the frequencies of SNPs between the gnomAD 2.1.1 release (42) and the NARD were compared. We redefined the frequency of 1.8 million genome-wide SNPs that are rare in worldwide populations from the gnomAD (gnomAD-ALL) to low-frequency or common ($MAF \geq 5\%$). Moreover, 0.5 million rare genome-wide SNPs in East Asians from the gnomAD (gnomAD-EAS) were low-frequency or common variants in the NARD (**Fig. 14**). We simulated rare disease variant discovery using 203 samples that were included in the three pseudo-GWAS panels for the imputation analysis. We applied variant filtering criteria ($MAF < 5\%$) from the guidelines of the American College of Medical Genetics for

the interpretation of sequence variants (25). Notably, the number of protein-altering variants (missense, nonsense, frameshift, and splicing variants) with rare frequency was significantly reduced when the exome catalog of the gnomAD-EAS and the NARD were jointly applied for variant filtration (**Fig. 31**). This result represents that the NARD could also contribute to the classification of pathogenic variant besides the genotype imputation for the Northeast Asians.

4.2 Asian-specific SNPs related to drug responses

We investigated Northeast Asian-specific pharmacogenomic features affected by different types of variants. We found 116 SNPs or indels associated with drug responses based on the ClinVar (47) annotation, and three of them were common in only Asian populations, especially in Northeast Asians: rs4148323 (in *UGT1A1*), rs4986893 (in *CYP2C19*), and rs116855232 (in *NUDT15*). Notably, rs116855232 is known to induce life-threatening leukopenia after thiopurine therapy (75). The frequencies of rs116855232 in CHN, KOR, and JPN were consistent with previous research (75); 14.3%, 11.8%, and 10.6%, respectively. Especially, it was most frequently found in MNG (15.2%; **Fig. 32a**).

4.3 Large deletion in intronic region of *BIM*

We also cataloged 83 Northeast Asian-specific SVs within genic regions. Our finding includes an intronic deletion (chr2:111,883,195 - 111,886,097) in *BCL2L11*, which encodes Bcl-2-like protein 11 (*BIM*) and mediates intrinsic resistance to tyrosine kinase inhibitor (49). It was predominantly identified in Northeast Asian populations, most and least frequently in KOR (8.5%) and MNG (3.2%), respectively (**Fig. 32b**).

4.4 HLA-B*15:01 influencing ICB response

The HLA was recently reported to be associated with the responsiveness to immune checkpoint blockade (ICB) (76), hence we examined HLA class I haplotypes of Northeast Asians. It was addressed that HLA-B*15:01 influences poor response to ICB therapy. We identified that HLA-B*15:01 was more frequently found in KOR and JPN compared to non-Finnish Europeans of the 1KGP3 (Chi-square $P = 0.04$ and $P = 0.009$ for each; **Fig. 32c**). This result was supported by the HLA catalog from the NMDP (63); KOR and JPN showed high incidences of HLA-B*15:01 followed by native Alaskans or Aleuts (**Fig. 33**).

4.5 HLA-B44 influencing ICB response

Unlike HLA-B*15:01, the cancer patients with HLA-B44 supertype (HLA-B*18:01, HLA-B*44:02, HLA-B*44:03, HLA-B*44:05, and HLA-B*50:0) are described to have the favorable response to ICB therapy (76). It was less

frequently found in Northeast Asians relative to non-Finnish Europeans (Chi-square $P < 0.0001$, **Fig. 32d**). According to results in the NMDP database (63), less than 5% of KOR, JPN, and CHN individuals have HLA-B44 supertype, while European Caucasians account for 22.2% of this supertype (**Fig. 34**). Both the results are consistent.

Table 1. The total number of variants in 1,779 Northeast Asians by MAF and functional category.

Type	Frequency ^a	Number of variants	Functional variation								
			Protein coding region				Non-coding region				
			Silent / Nonframeshift	Missense / Frameshift	Stoploss / Stopgain	Unknown	Intronic	Intergenic	Splicing	UTR	ncRNA
SNP	Singleton	17,811,366	86,804	146,480	3,722	2,690	6,842,300	9,370,754	2,110	247,422	1,109,084
	Rare	13,673,626	54,642	87,791	1,658	1,917	5,270,353	7,248,270	1,363	164,492	843,140
	Low	3,430,315	12,753	15,710	232	428	1,299,727	1,851,373	245	38,673	211,174
	Common	5,727,339	17,886	15,981	151	729	2,049,372	3,228,994	159	53,221	360,846
	Total	40,642,646	172,085	265,962	5,763	5,764	15,461,752	21,699,391	3,877	503,808	2,524,244
Indel	Singleton	1,402,707	3,191	5,068	157	129	558,772	717,182	517	27,748	89,943
	Rare	1,376,996	2,733	2,884	127	127	544,183	717,045	217	22,047	87,633
	Low	452,337	634	827	37	37	173,946	241,506	61	6,444	28,845
	Common	569,436	422	369	18	89	207,132	317,135	145	7,157	36,969
	Total	3,801,476	6,980	9,148	339	382	1,484,033	1,992,868	940	63,396	243,390

^a Rare: MAF<0.5%, low (low-frequency): 0.5%≤MAF<5%, and common: MAF≥5%.

Table 2. The basic statistics of structural variations.

	Total	Novelty		Type			
		Known	Novel	Insertion	Deletion	Duplication	Inversion
Total	93,764	23,886 (25.5%)	69,878 (74.5%)	1,597 (1.7%)	67,241 (71.7%)	20,262 (21.6%)	4,664 (5.0%)
Singleton	25,387 (27.1%)	3,715 (14.6%)	21,672 (85.4%)	170 (0.7%)	17,127 (67.5%)	6,459 (25.4%)	1,631 (6.4%)
Rare	46,807 (49.9%)	7,289 (15.6%)	39,518 (56.6%)	174 (0.4%)	33,329 (71.2%)	11,676 (24.9%)	1,628 (3.5%)
Low	7,857 (8.38%)	4,148 (52.8%)	3,709 (47.2%)	127 (1.6%)	6,139 (78.1%)	1,060 (13.5%)	531 (6.8%)
Common	13,713 (14.6%)	8,734 (63.7%)	4,979 (36.3%)	1,126 (8.2%)	10,646 (77.6%)	1,067 (7.8%)	874 (6.4%)

Table 3. Imputation performance according to types of reference panel.

Cohort	Panel	Imputation accuracy (aggregated R^2)											Number of imputed SNP		
		$0 < \text{MAF} < 0.002$	$0.002 \leq \text{MAF} < 0.005$	$0.005 \leq \text{MAF} < 0.01$	$0.01 \leq \text{MAF} < 0.02$	$0.02 \leq \text{MAF} < 0.05$	$0.05 \leq \text{MAF} < 0.1$	$0.1 \leq \text{MAF} < 0.2$	$0.2 \leq \text{MAF} < 0.3$	$0.3 \leq \text{MAF} < 0.4$	$0.4 \leq \text{MAF} < 0.5$	$0.5 \leq \text{MAF} \leq 1$	$0.7 \leq R^2 < 0.8$	$0.8 \leq R^2 < 0.9$	$0.9 \leq R^2$
KOR	1KGP3	0.634	0.651	0.673	0.715	0.833	0.942	0.977	0.986	0.988	0.989	0.987	475,346	602,461	6,656,207
	1KGP3 (re-phased)	0.635	0.655	0.679	0.721	0.837	0.943	0.978	0.987	0.988	0.989	0.987	525,997	650,791	6,552,317
	NARD	0.689	0.723	0.732	0.768	0.867	0.954	0.983	0.990	0.991	0.992	0.991	418,799	516,709	6,208,397
	NARD (re-phased)	0.795	0.828	0.834	0.855	0.916	0.971	0.988	0.992	0.993	0.993	0.993	615,889	829,310	7,361,976
	NARD+1KGP3 (re-phased)	0.800	0.830	0.836	0.856	0.916	0.970	0.988	0.992	0.993	0.993	0.993	618,023	881,042	7,511,378
CHN	1KGP3	0.711	0.694	0.677	0.715	0.833	0.938	0.974	0.983	0.986	0.987	0.982	461,459	603,108	6,836,755
	1KGP3 (re-phased)	0.695	0.685	0.684	0.734	0.849	0.944	0.975	0.983	0.985	0.986	0.983	497,468	628,264	6,644,335
	NARD	0.592	0.648	0.663	0.709	0.832	0.938	0.974	0.983	0.986	0.987	0.981	321,904	434,883	6,084,715
	NARD (re-phased)	0.696	0.724	0.732	0.775	0.873	0.953	0.980	0.986	0.988	0.989	0.987	485,193	590,093	6,842,105
	NARD+1KGP3 (re-phased)	0.750	0.747	0.760	0.794	0.881	0.955	0.981	0.987	0.989	0.990	0.986	537,328	670,956	6,970,990
JPN	1KGP3	0.663	0.735	0.765	0.804	0.885	0.951	0.978	0.985	0.986	0.987	0.986	247,958	400,617	6,191,701
	1KGP3 (re-phased)	0.671	0.744	0.774	0.811	0.889	0.953	0.979	0.985	0.987	0.988	0.987	261,419	418,217	6,089,905
	NARD	0.728	0.783	0.799	0.830	0.899	0.958	0.982	0.988	0.989	0.990	0.990	222,439	343,005	5,683,476
	NARD (re-phased)	0.836	0.883	0.891	0.906	0.941	0.974	0.988	0.991	0.992	0.993	0.992	250,310	407,568	6,511,709
	NARD+1KGP3 (re-phased)	0.843	0.887	0.893	0.908	0.943	0.974	0.988	0.991	0.992	0.992	0.992	245,796	414,965	6,578,176



Figure 2 | Geographic map of the study area in the NARD. The proportions of KOR, JPN, MNG, CHN, and HKG are 47.8%, 22.3%, 21.6%, 5.1%, and 3.3%, respectively.

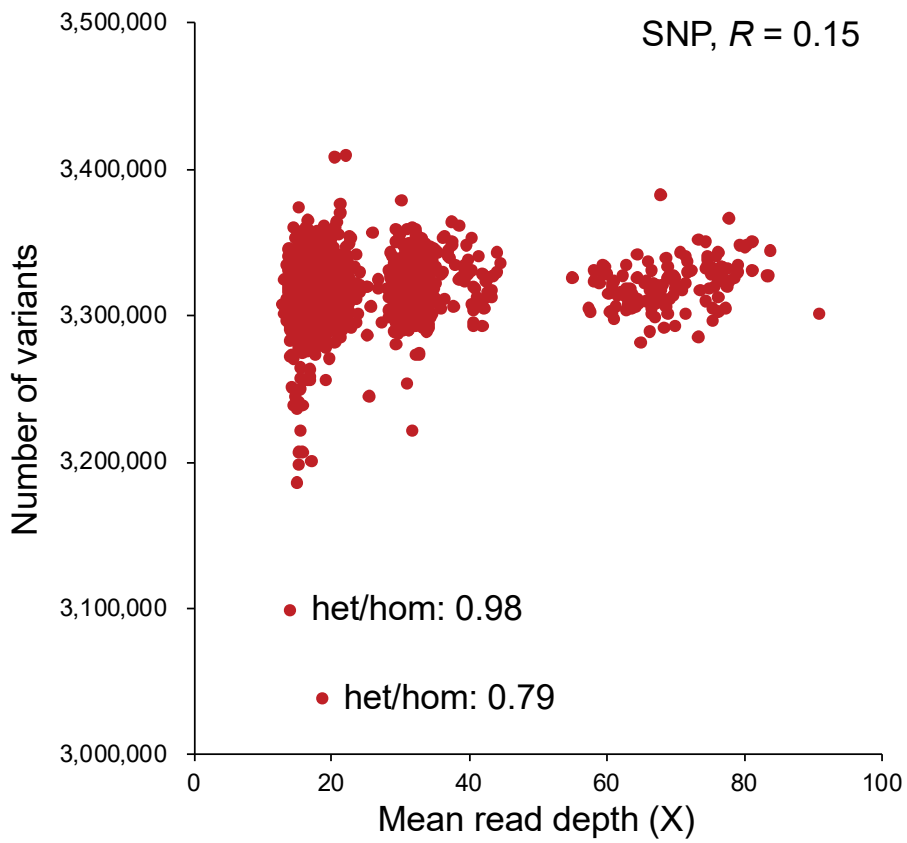


Figure 3 | Correlation between the sequencing depth and the number of variants.

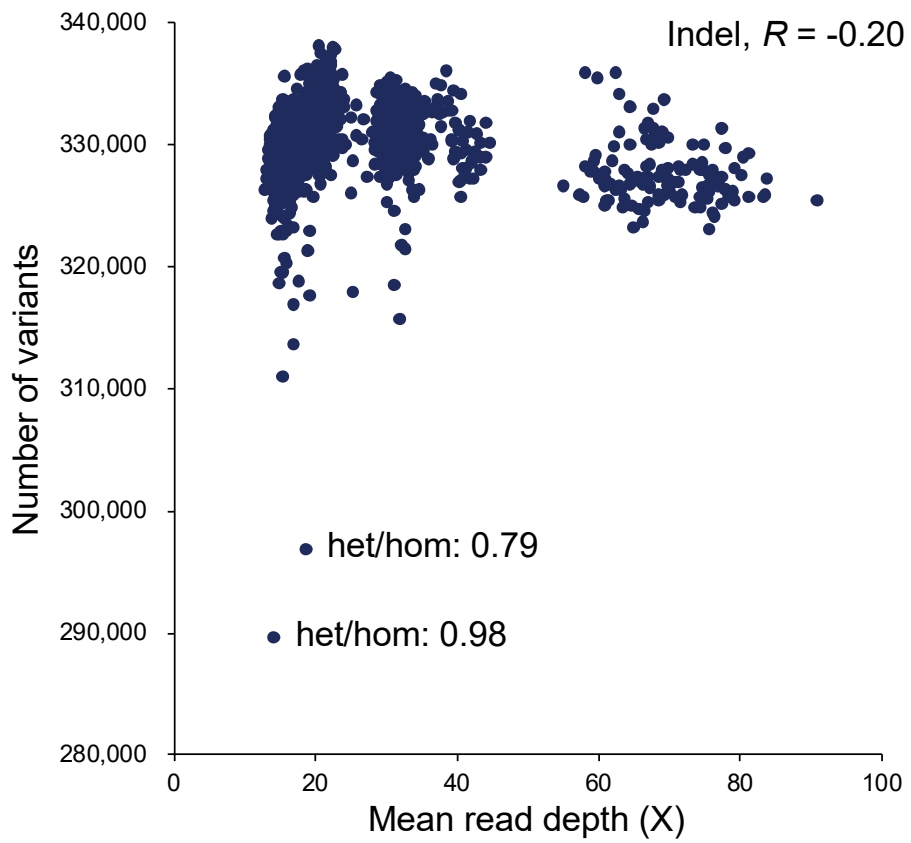


Figure 4 | Pearson correlation coefficient (R) was calculated excluding the two samples with abnormal heterozygous/homozygous ratios.

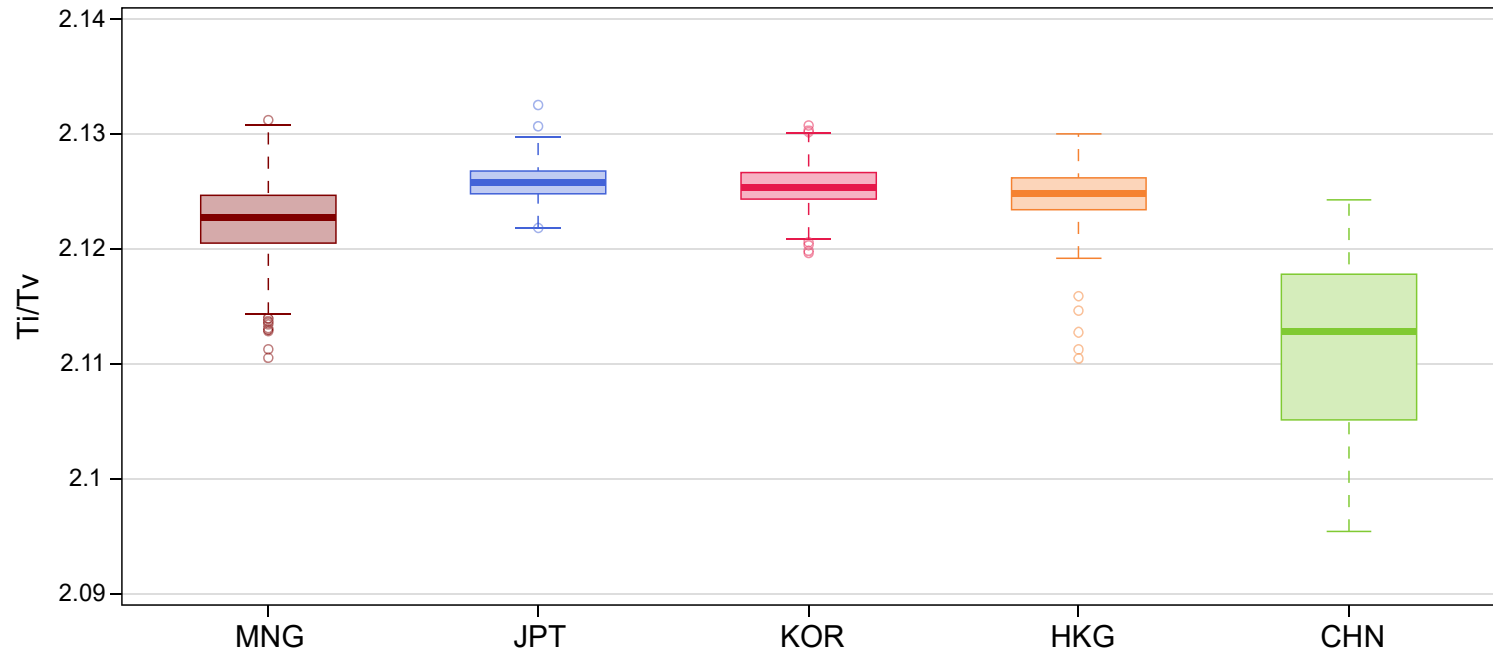


Figure 5 | The transition to transversion ratio of the populations in the NARD.

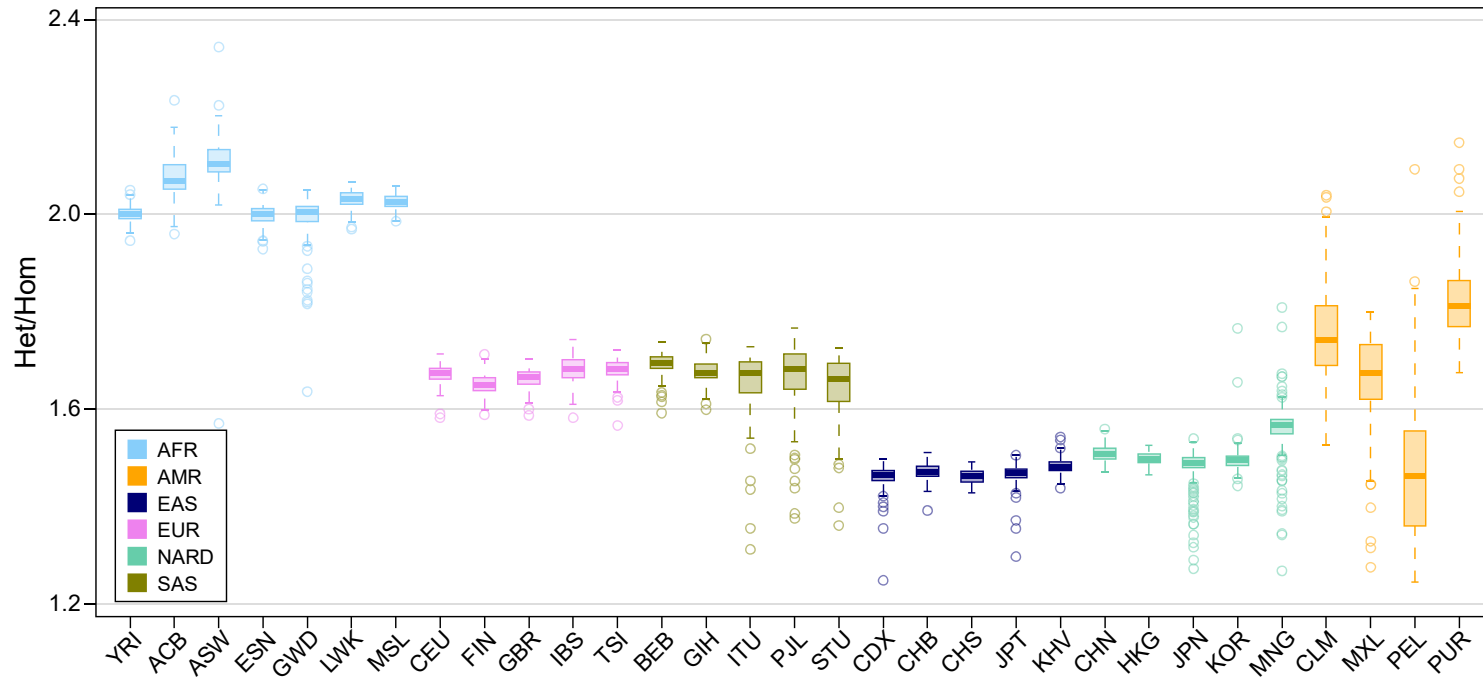


Figure 6 | Heterozygous to homozygous ratio of the global populations. The Het/Hom ratios for the populations of the NARD and the 1KGP3 show the trends specific for each population group. The trend of the NARD populations is similar to EAS of the 1KGP3.

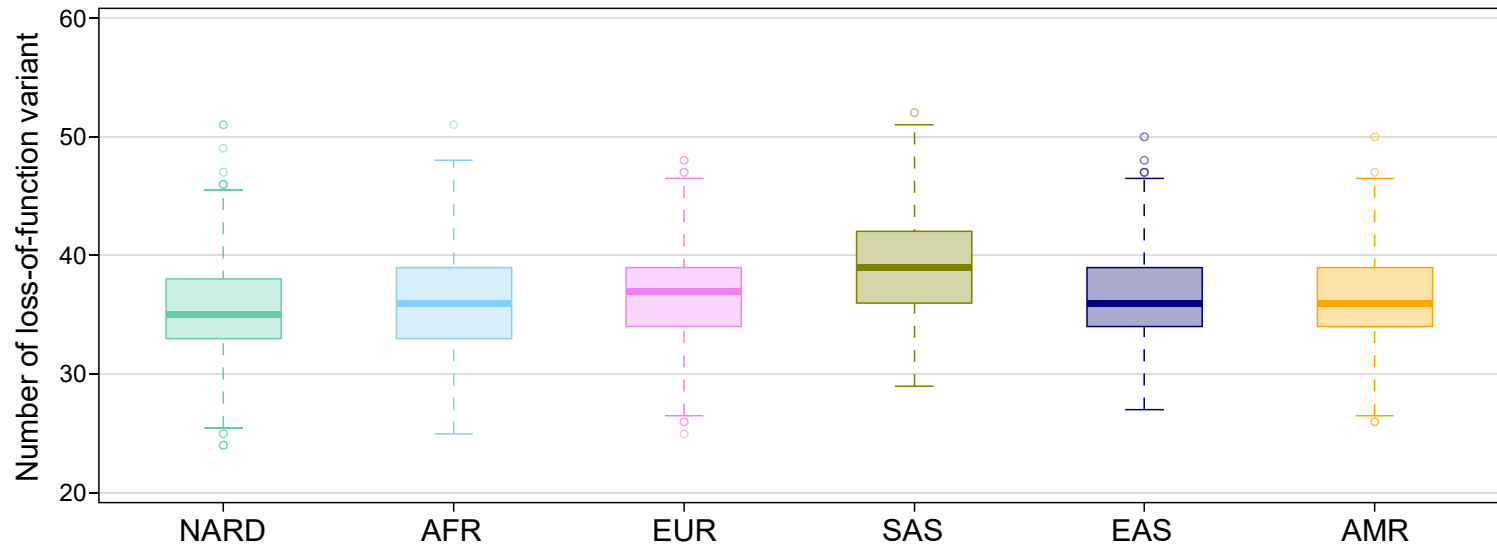


Figure 7 | The number of loss-of-function variants. The 1KGP3 dataset was divided into five super population codes.

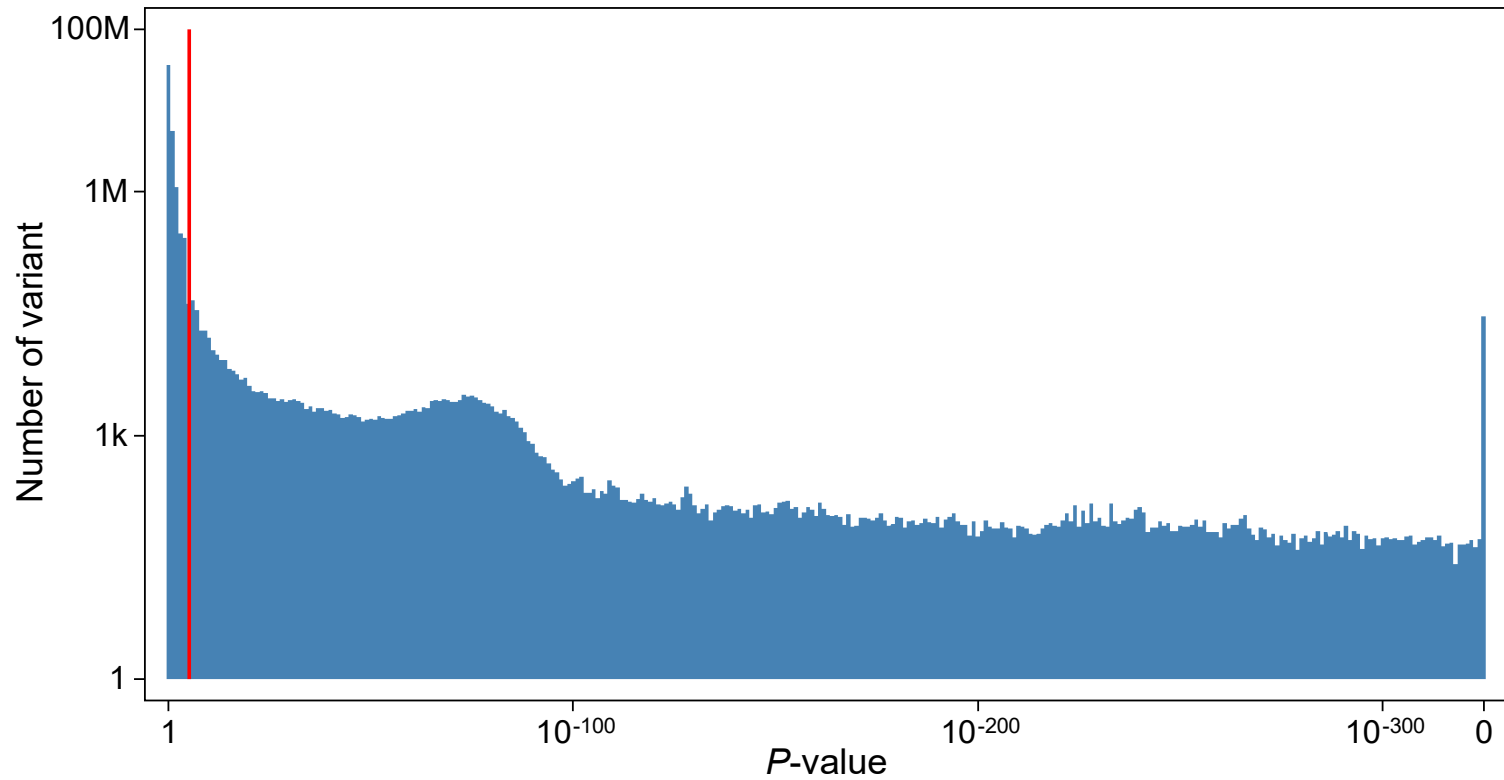


Figure 8 | Hardy-Weinberg Equilibrium test of variants in the NARD. The red line indicates the significance threshold (P-value = 10^{-5}).

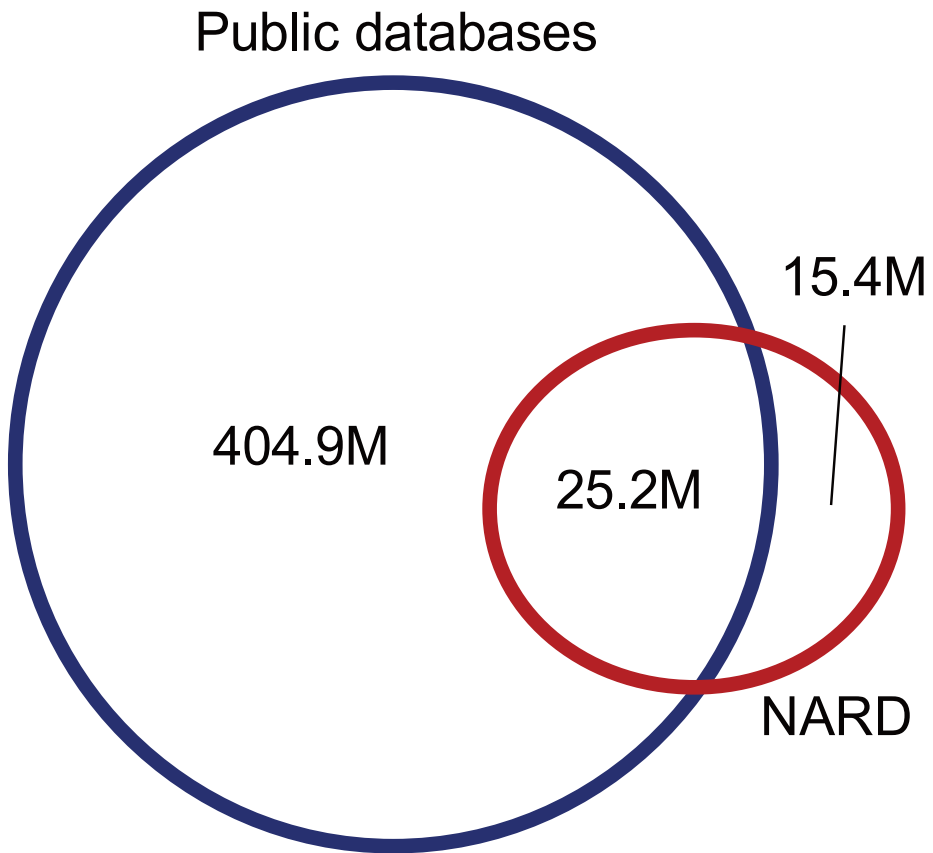


Figure 9 | The number of novel SNPs that were not identified elsewhere.

Public databases include Kaviar, gnomAD (2.1.1 release), and dbSNP150.

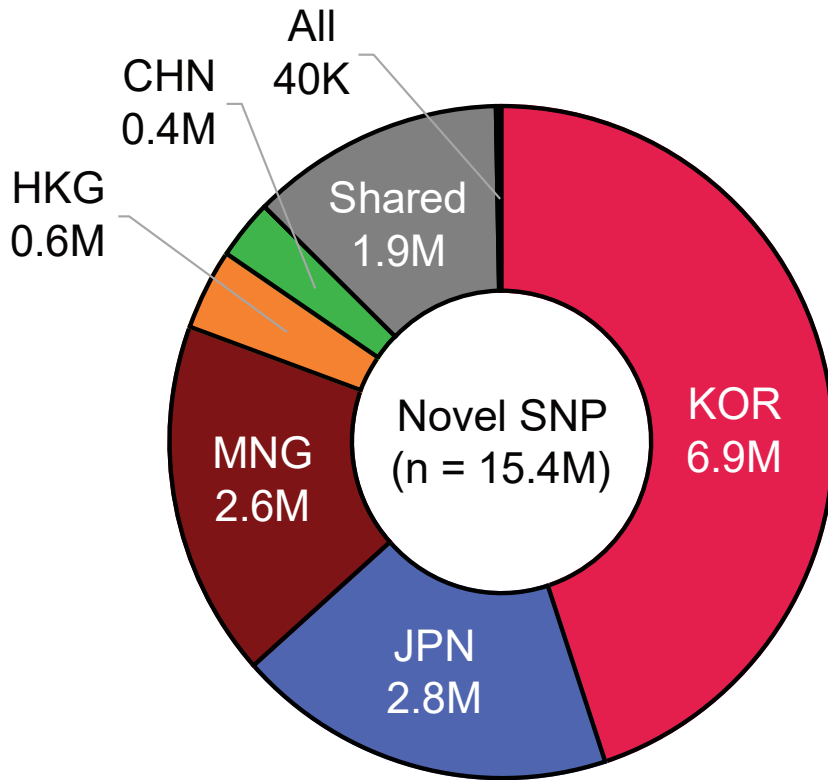


Figure 10 | Distribution of novel SNPs per population. Novel SNPs found in multiple and all populations were included in “Shared” and “All,” respectively.

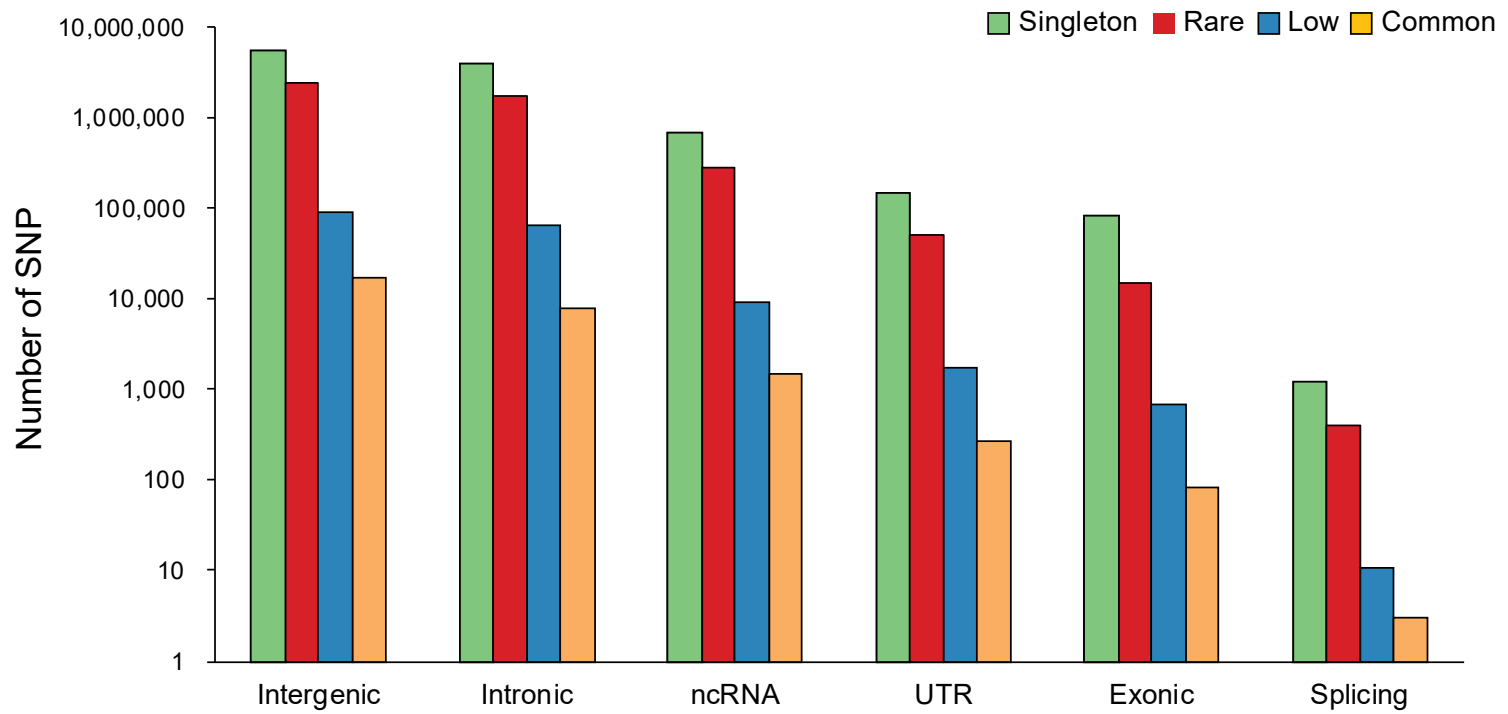


Figure 11 | Distribution of novel SNPs based on the RefSeq gene definition.

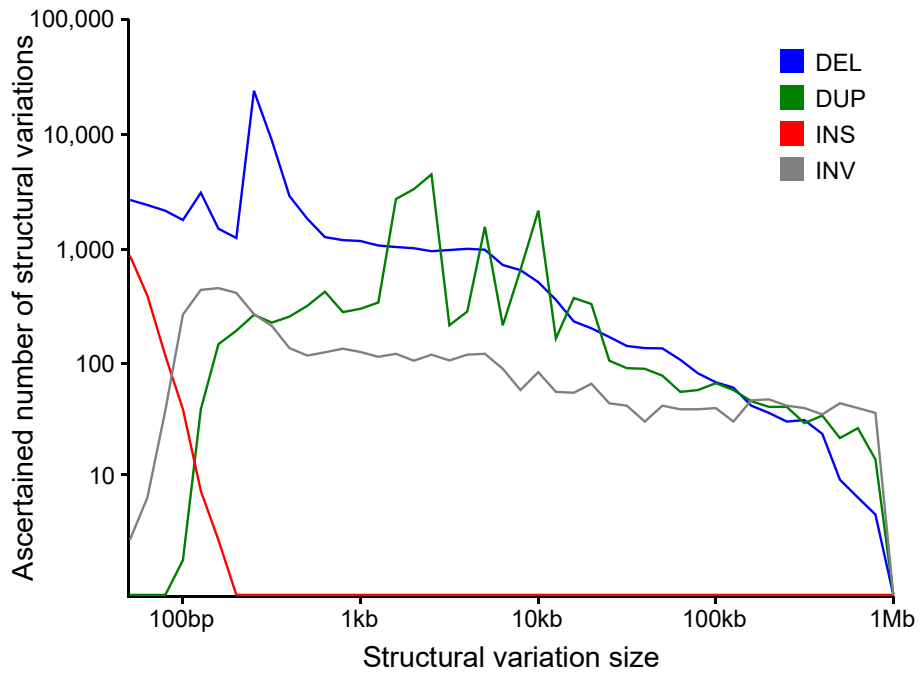


Figure 12 | The size distribution of ascertained SVs. DEL, DUP, INS, and INV denote deletions, duplications, insertions, and inversions, respectively.

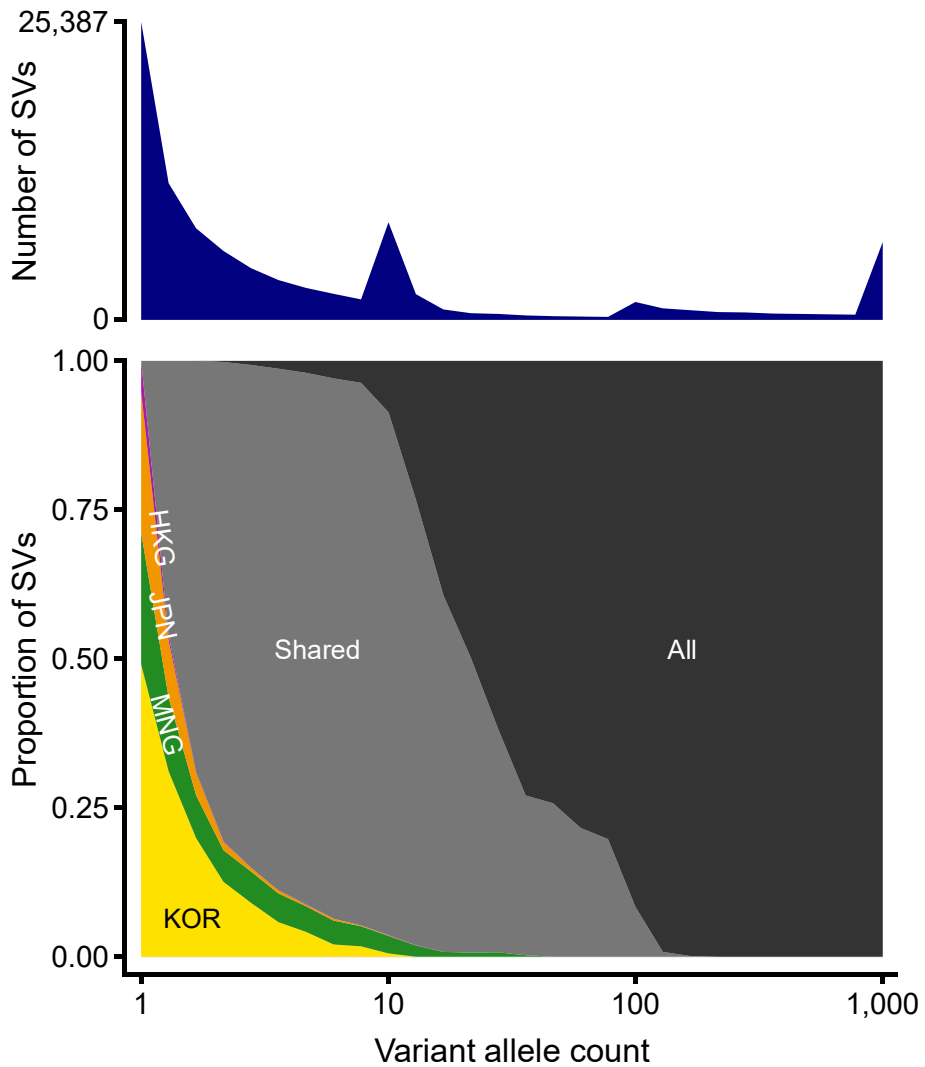


Figure 13 | The distribution of variant allele counts of structural variations. KOR, MNG, JPN, and HKG refer to the structural variations unique to each population. “Shared” is discovered in at least two populations and “All” is discovered in whole populations.

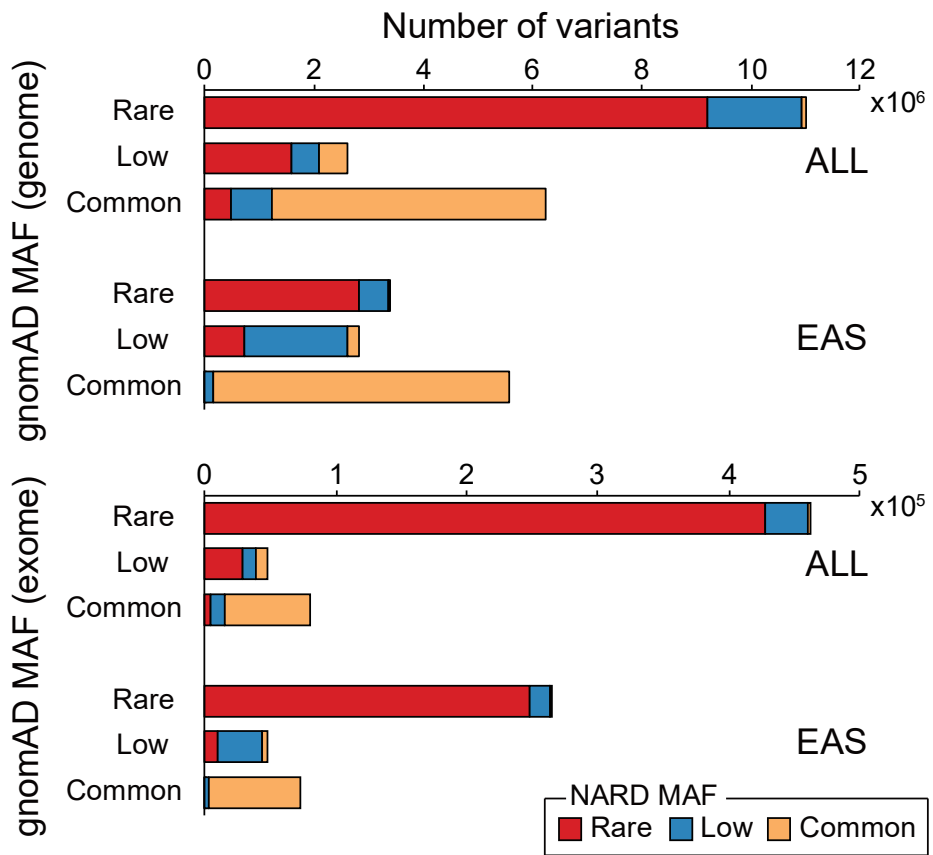


Figure 14 | MAF differences of SNPs shared between the NARD and the gnomAD in genome and exome regions. The x-axis denotes the MAF of SNPs in worldwide populations (ALL) or EAS from the gnomAD. Color represents the MAF of SNPs in 1,779 Northeast Asians from the NARD.

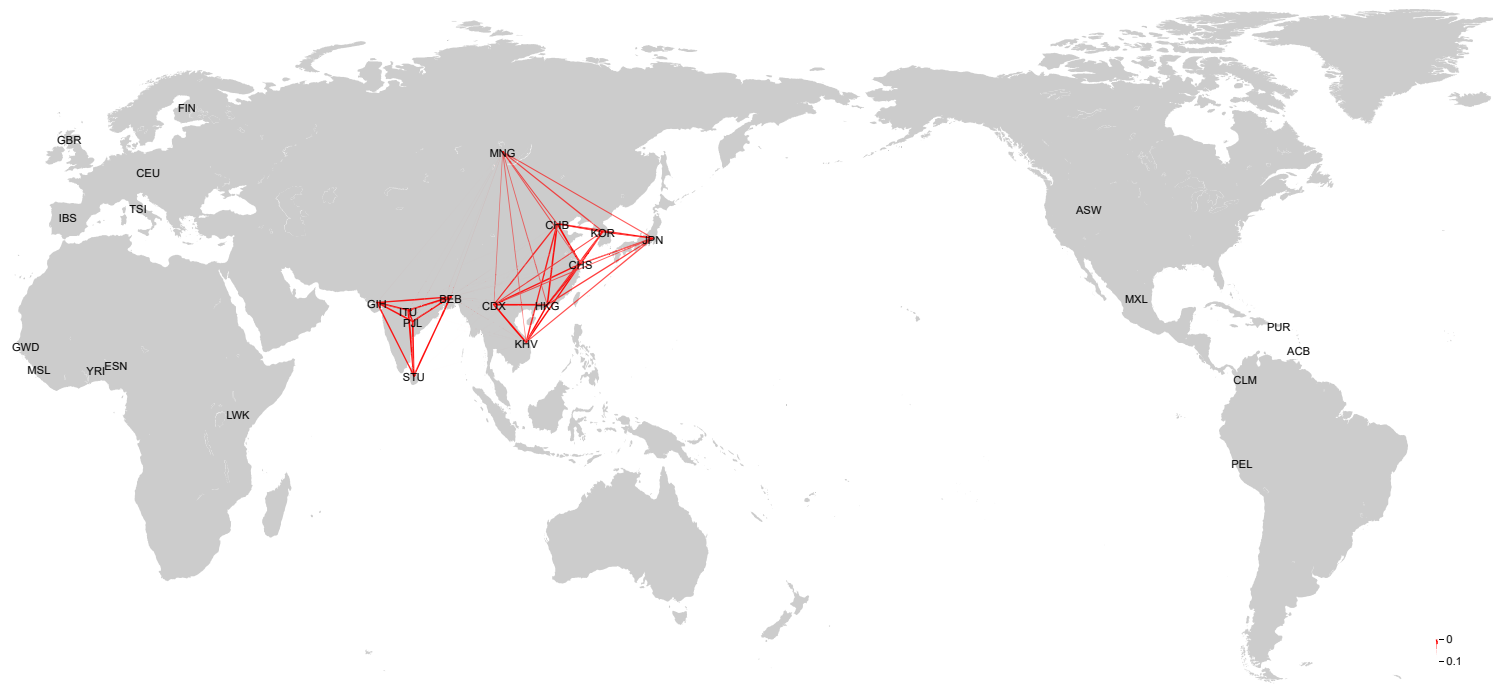


Figure 15 | The F_{ST} network among Asian populations of the NARD and the 1KGP3. The genetic structure of the Korean population. The thickness and the opacity of the red lines represent the genetic affinity between each population. JPT from the 1KGP3 was merged into JPN in this analysis.

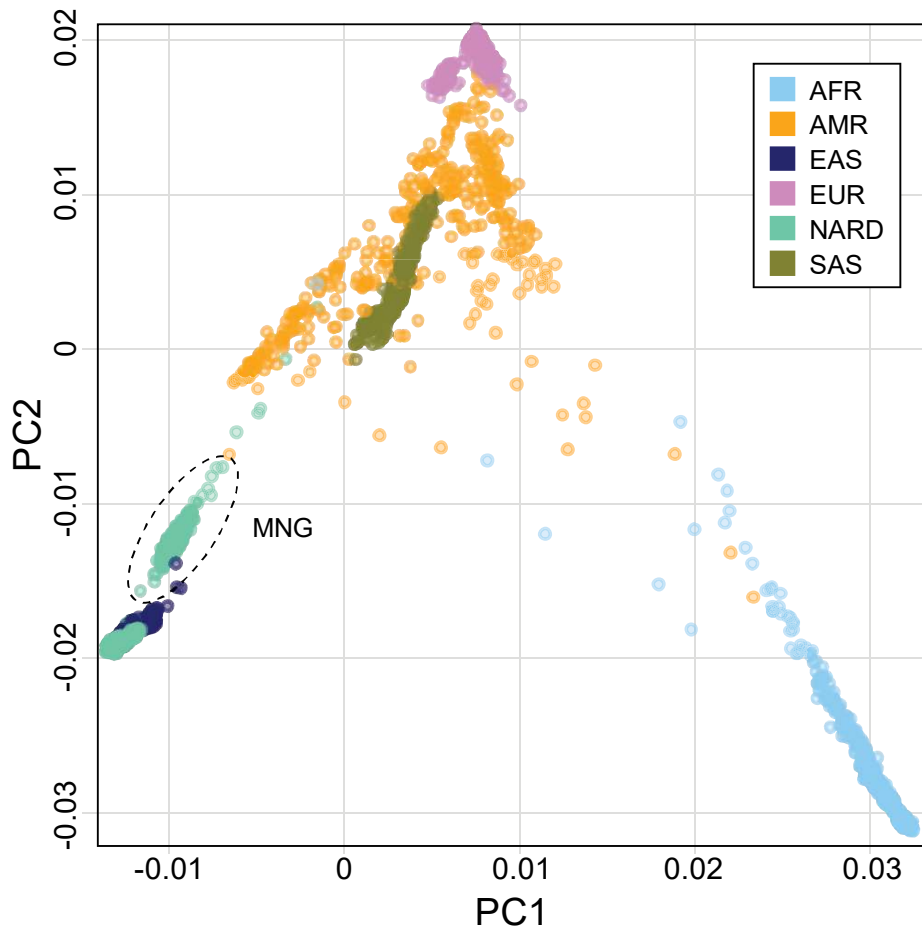


Figure 16 | PCA of global populations from the NARD and the 1KGP3.

AFR, AMR, EAS, EUR, and SAS denote Africans, Americans, East Asians, Europeans, and South Asians, respectively.

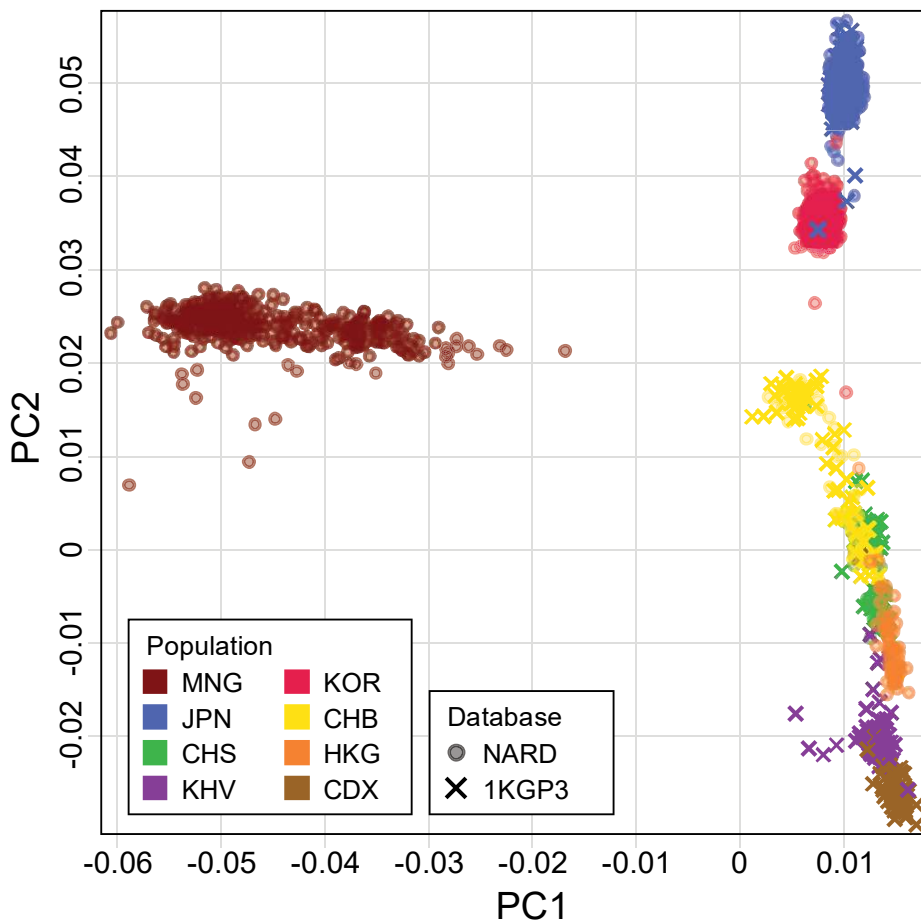


Figure 17 | PCA of Northeast and Southeast Asians from the NARD and the 1KGP3. Japanese in Tokyo from the 1KGP3 were merged into JPN and CHN of the NARD were categorized into CHB and CHS in this figure.

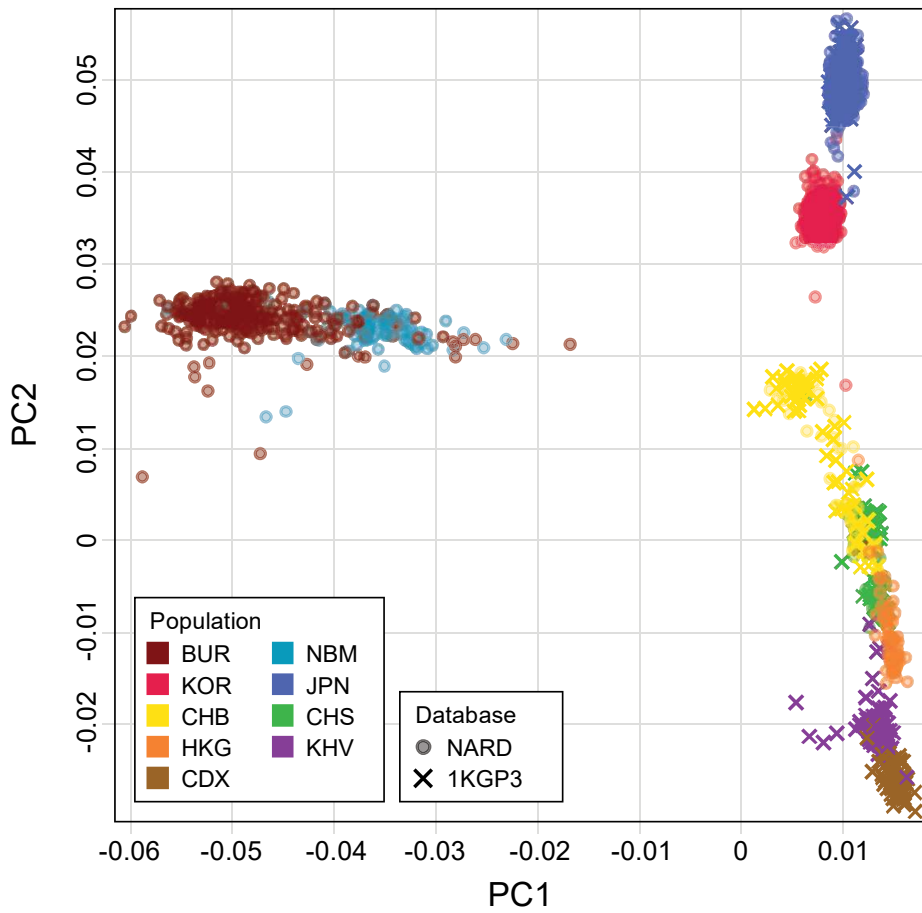


Figure 18 | PCA of Northeast and Southeast Asians from the NARD and the 1KGP3. MNG population is divided into BUR and NBM in this figure.

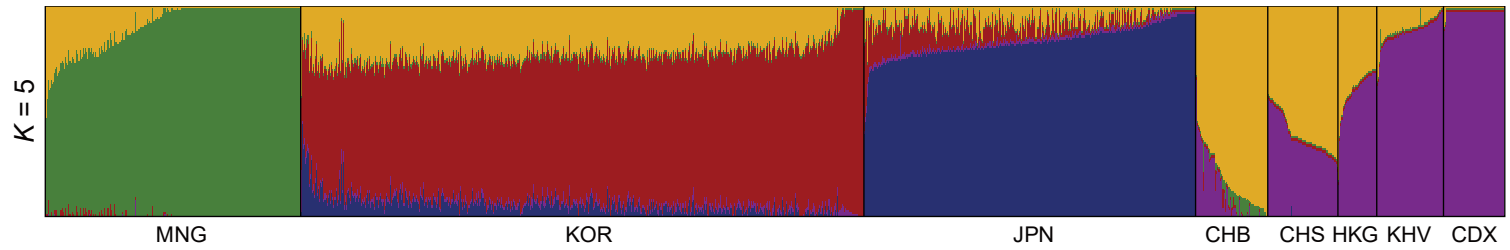


Figure 19 | Population substructure of Northeast and Southeast Asians with five ancestral components inferred by ADMIXTURE algorithm.



Figure 20 | Population substructure of MNG with five ancestral components inferred by ADMIXTURE algorithm.

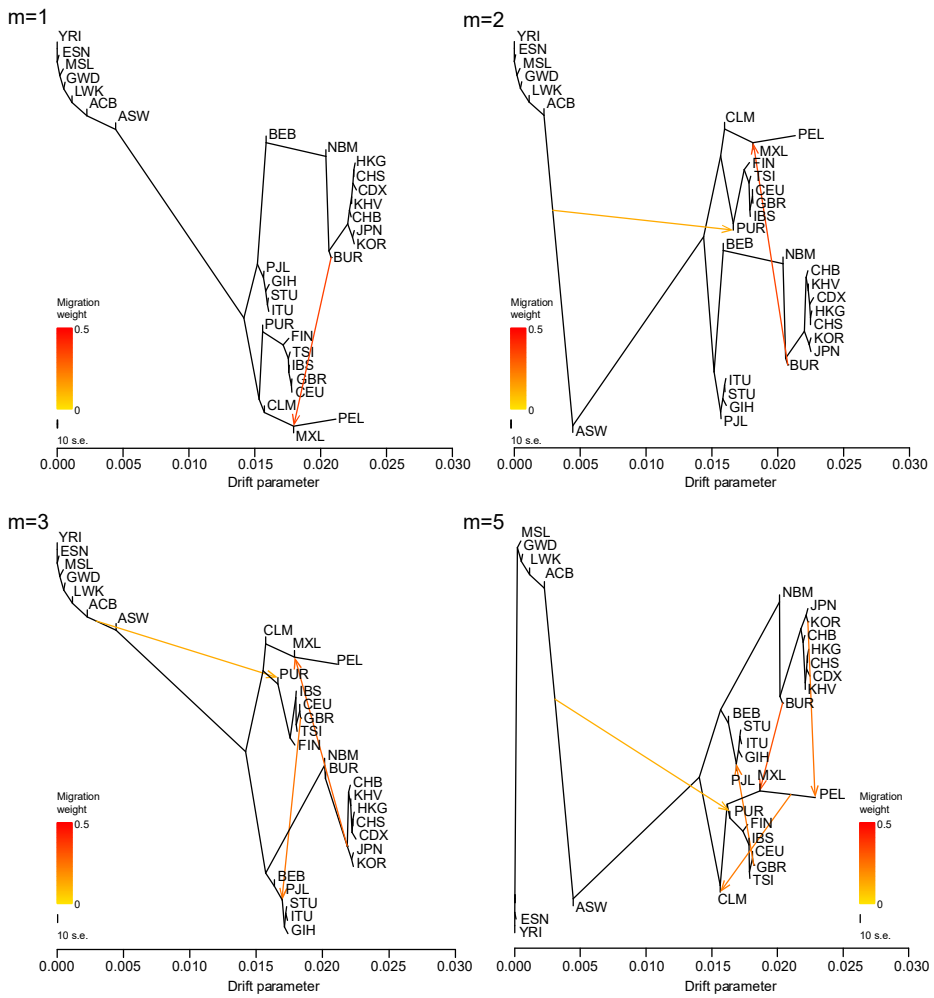


Figure 21 | The maximum likelihood trees generated by TreeMix. One to five migration edges were shown (except for m=4). JPT from the 1KGP3 was merged into JPN in this analysis.

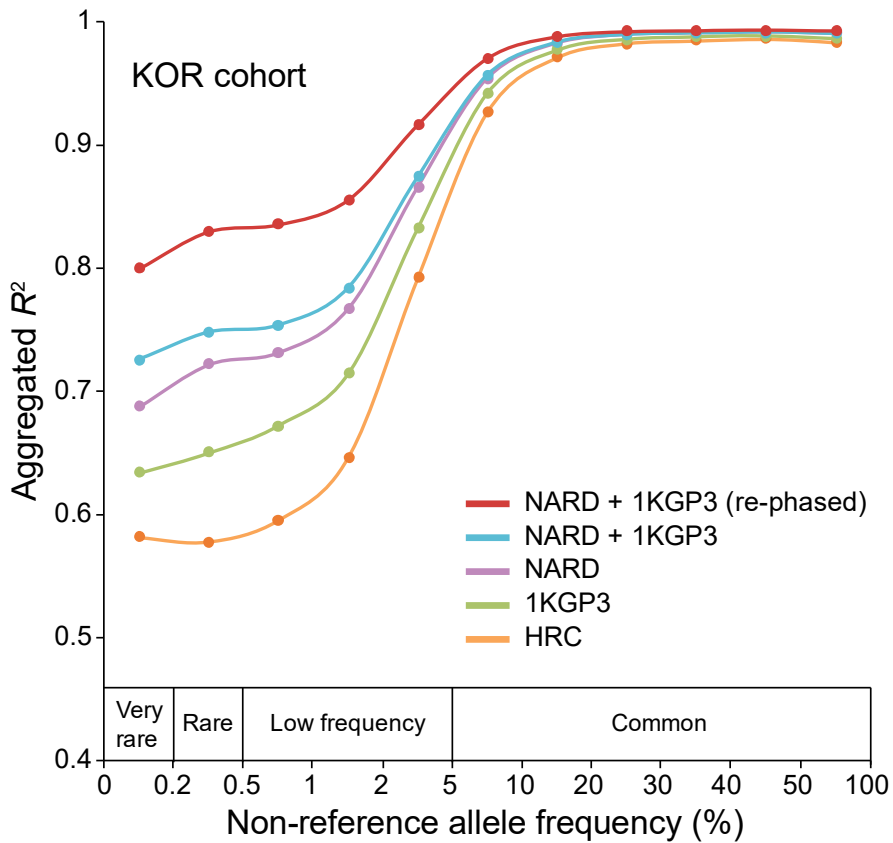


Figure 22 | Imputation accuracy assessment using the five different reference panels in KOR individuals. The pseudo-GWAS panel of 97 KOR was used for the imputation. The x-axis represents MAF of 850 KOR individuals from the NARD. The y-axis represents the aggregated R^2 values of SNPs, which were calculated by the true genotypes and the imputed dosages. Only SNPs that were imputed across all panels were used for the aggregation of R^2 values.

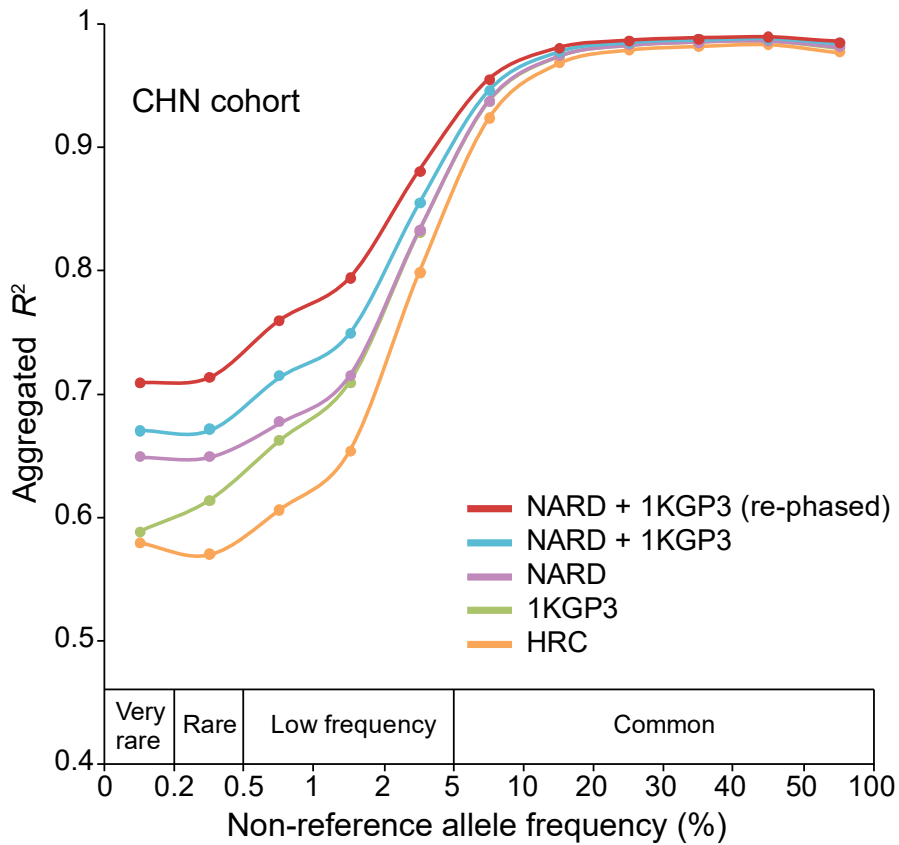


Figure 23 | Imputation accuracy assessment using the five different reference panels in CHN individuals. The pseudo-GWAS panel of 79 CHN individuals was used for the imputation. The x-axis represents MAF of 10,639 CHN individuals. The y-axis represents the aggregated R^2 values of SNPs, which were calculated by the true genotypes and the imputed dosages. Only SNPs that were imputed across all panels were used for the aggregation of R^2 values.

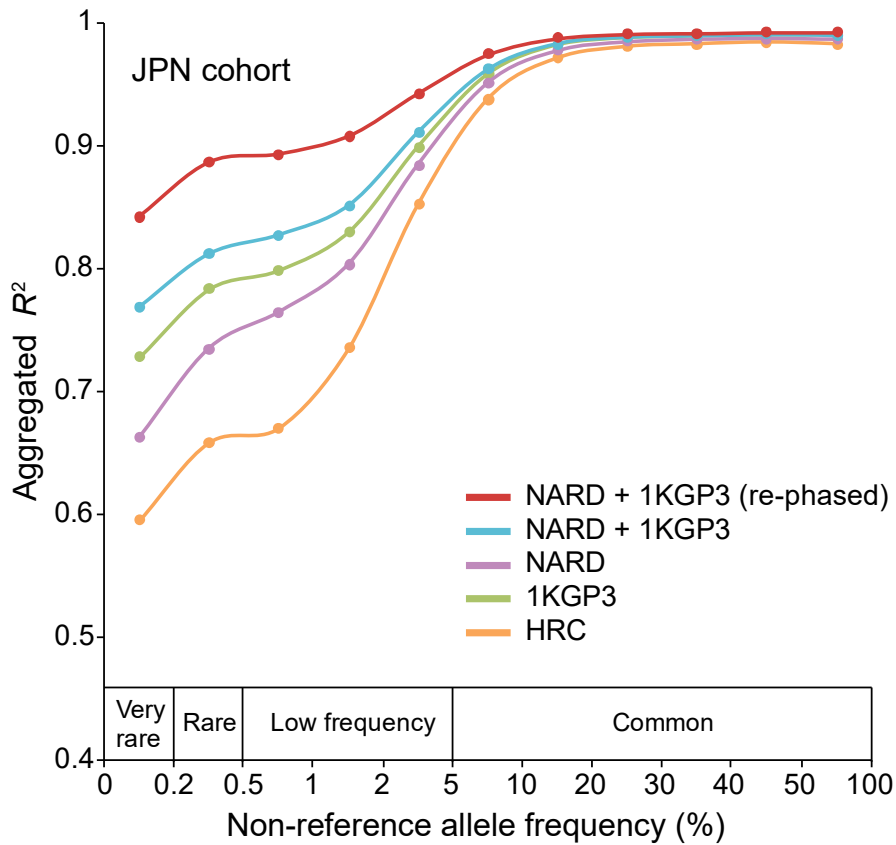


Figure 24 | Imputation accuracy assessment using the five different reference panels in JPN individuals. The pseudo-GWAS panel of 27 JPN individuals was used for the imputation. The x-axis represents MAF of 3,554 JPN individuals. The y-axis represents the aggregated R^2 values of SNPs, which were calculated by the true genotypes and the imputed dosages. Only SNPs that were imputed across all panels were used for the aggregation of R^2 values.

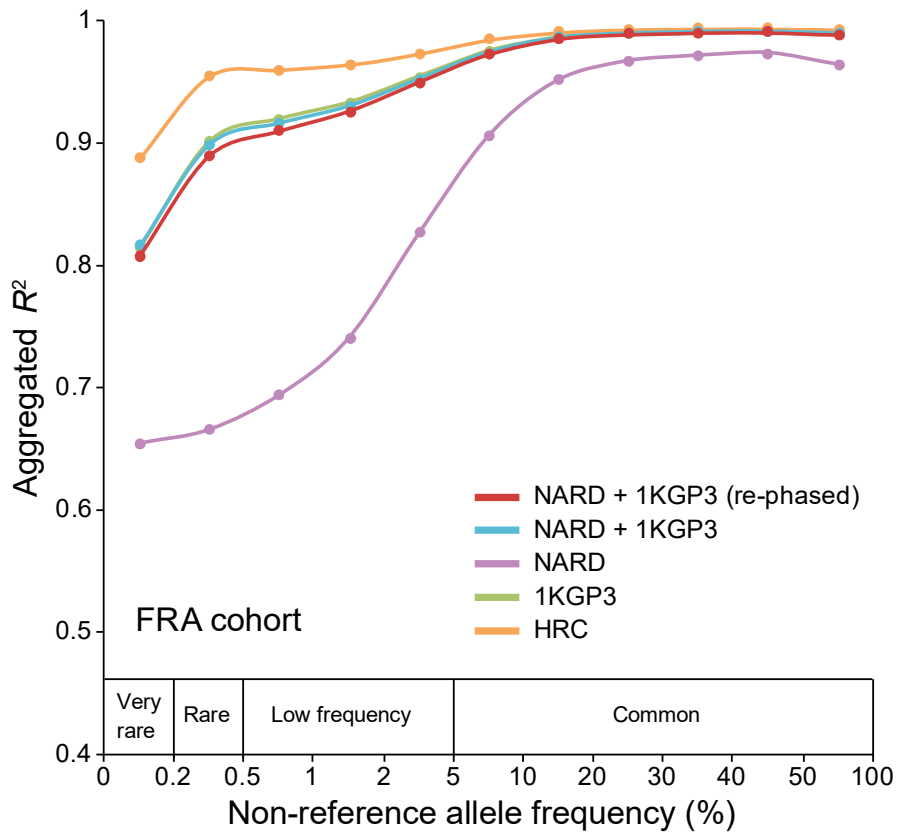


Figure 25 | Imputation accuracy assessment using the five different reference panels in FRA individuals. The pseudo-GWAS panel of 24 FRA was used for the imputation. The x-axis represents MAF of 7,718 non-Finnish European individuals from the gnomAD. The y-axis represents the aggregated R^2 values of SNPs, which were calculated by the true genotypes and the imputed dosages. Only SNPs that were imputed across all panels were used for the aggregation of R^2 values.

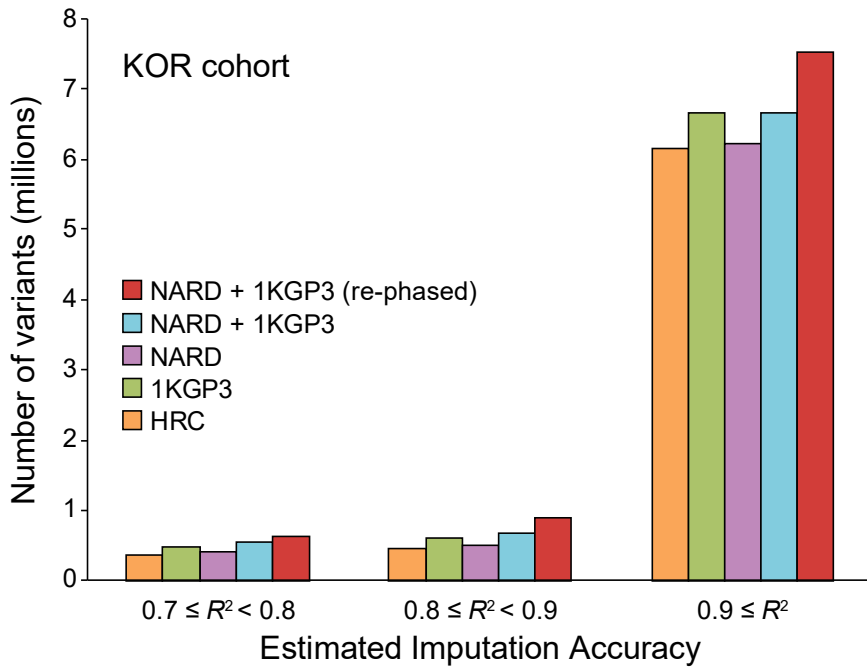


Figure 26 | The number of imputed SNPs as a function of the estimated imputation accuracy and the types of imputation panel. This result was generated based on the R^2 values that were estimated by Minimac3.

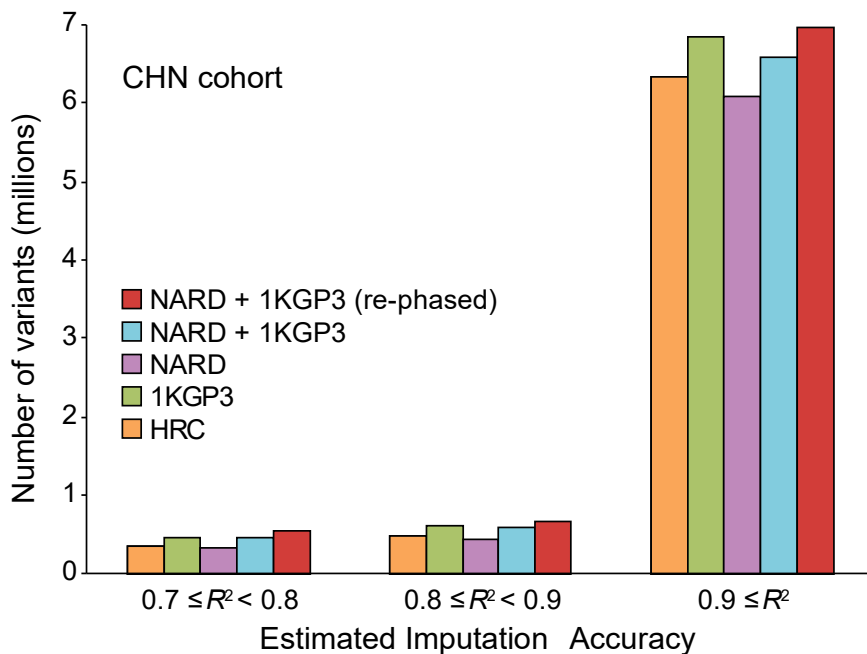


Figure 27 | Imputation performance evaluation of CHN individuals. The number of imputed SNPs as a function of the estimated imputation accuracy and the types of imputation panel. This result was generated based on the R^2 values that were estimated by Minimac3.

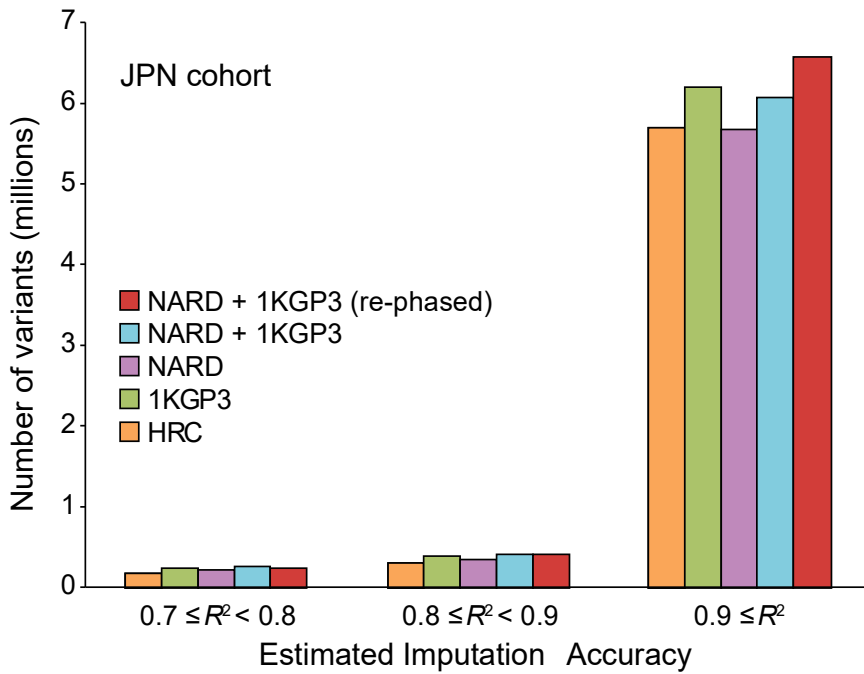


Figure 28 | Imputation performance evaluation of JPN individuals. The number of imputed SNPs as a function of the estimated imputation accuracy and the types of imputation panel. This result was generated based on the R^2 values that were estimated by Minimac3.

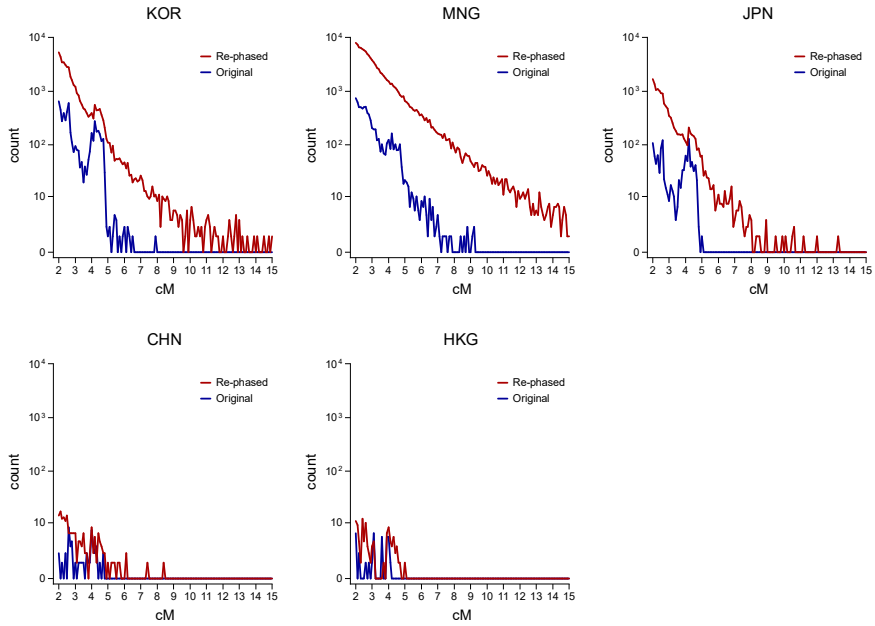


Figure 29 | Length distribution of shared IBD tracts between the two individuals in each population. Distribution of shared IBD tracts which were computed using the original and the re-phased haplotypes were displayed separately. The lengths of the shared IBDs are different by populations but consistently increased in the re-phased haplotypes across all.

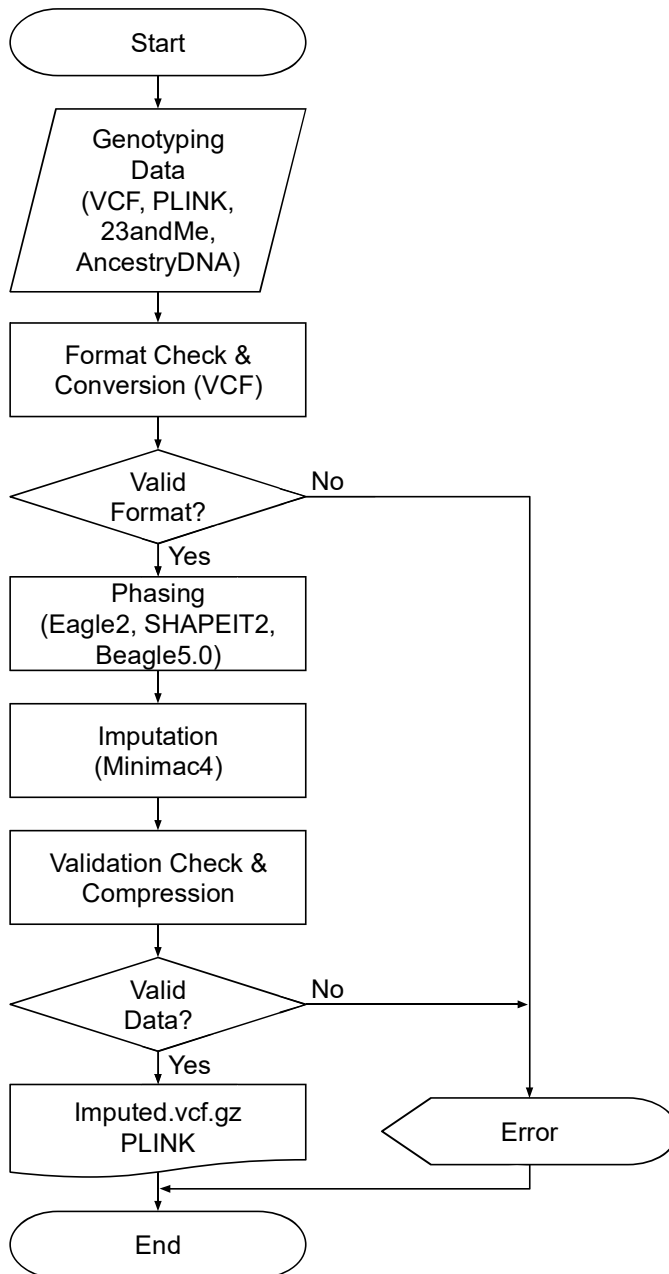


Figure 30 | The flow chart of the pipeline consisting of four major steps for the NARD imputation server.

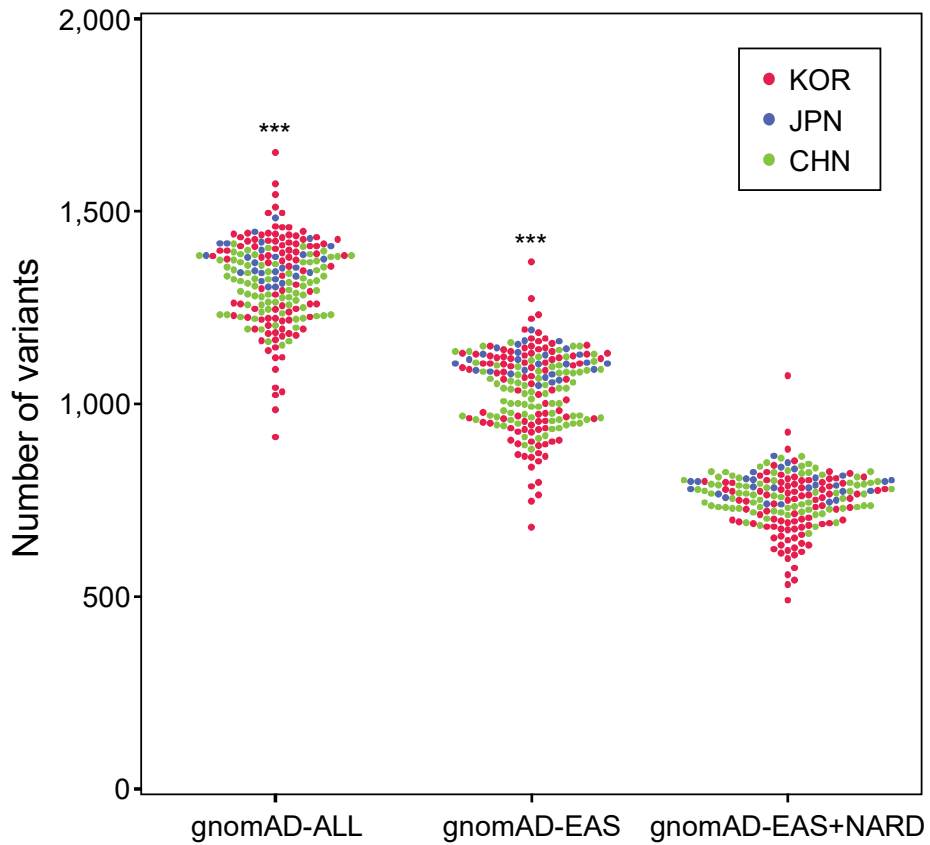


Figure 31 | The number of uncommon (MAF < 5%) protein-altering variants (missense, nonsense, frameshift, and splicing variants) after filtration using the gnomAD with/without the NARD. Variant catalog from the gnomAD (exome) was applied. ***P < 0.0001 by two-tailed Mann-Whitney U-test (compared with the gnomAD EAS + NARD).

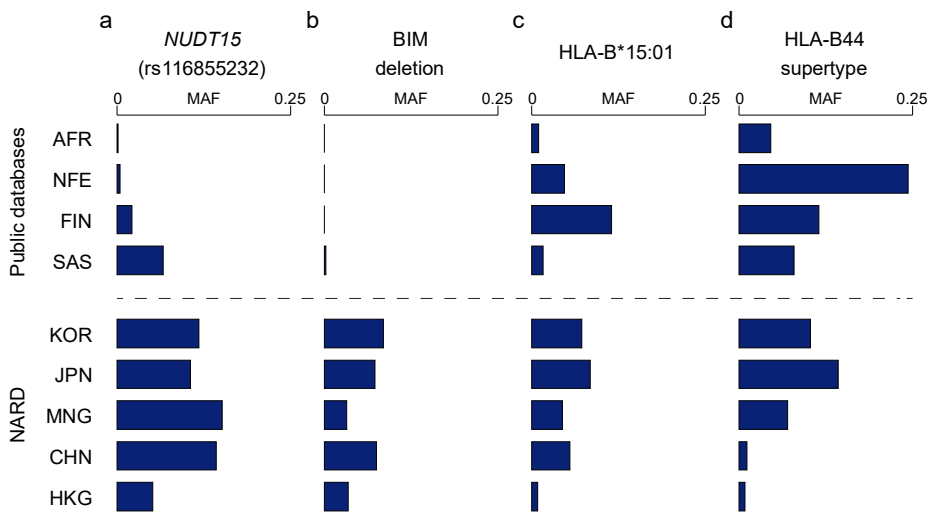


Figure 32 | Pharmacogenomic analysis. The frequency of variants and HLA class I subtypes associated with drug response. a, rs116855232 in *NUDT15*. b, Intronic deletion in *BIM* (chr2:111,883,195 - 111,886,097). c, HLA-B*15:01. d, HLA-B44 supertype. Public databases include the 1KGP3, the gnomAD, and the NMDP. MAFs of Africans, non-Finnish Europeans, Finns, and South Asians were from the gnomAD (for rs116855232) and the 1KGP3 (for *BIM* deletion). HLA frequencies of Africans, non-Finnish Europeans, and Finnish were from the 1KGP3 and South Asian Indians from the NMDP.

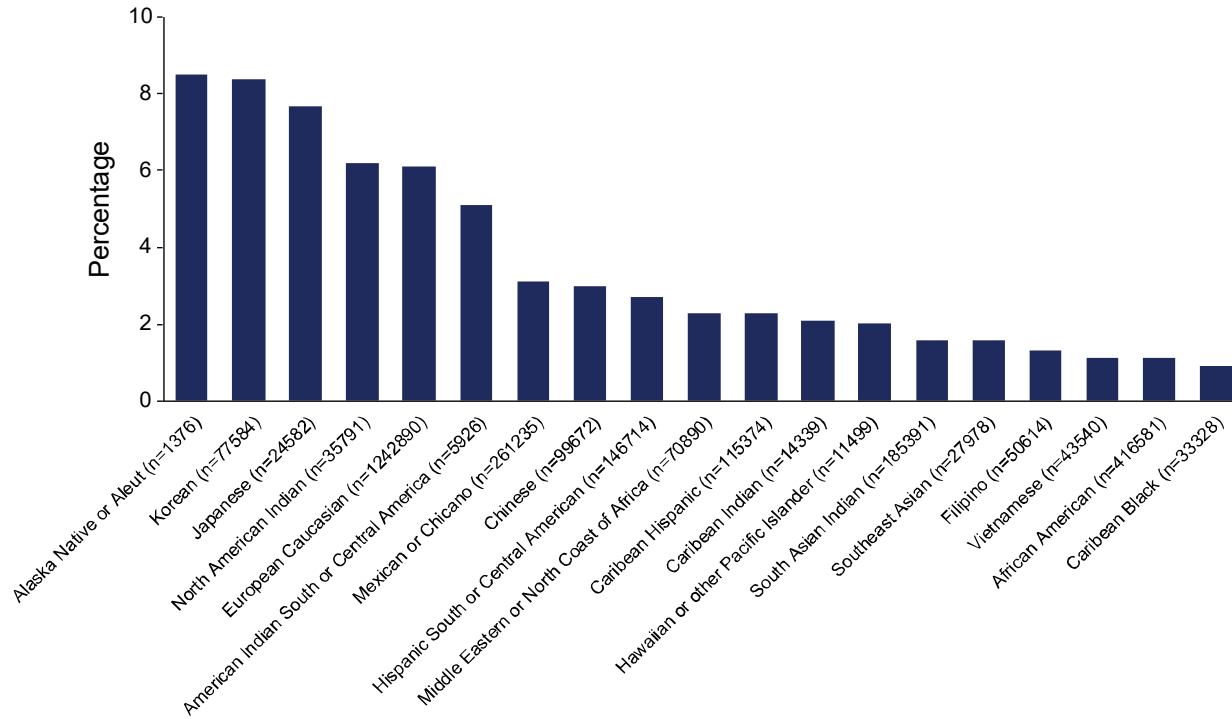


Figure 33 | Distribution of HLA class I subtypes HLA-B*15:01 associated with ICB response in the NMDP database.

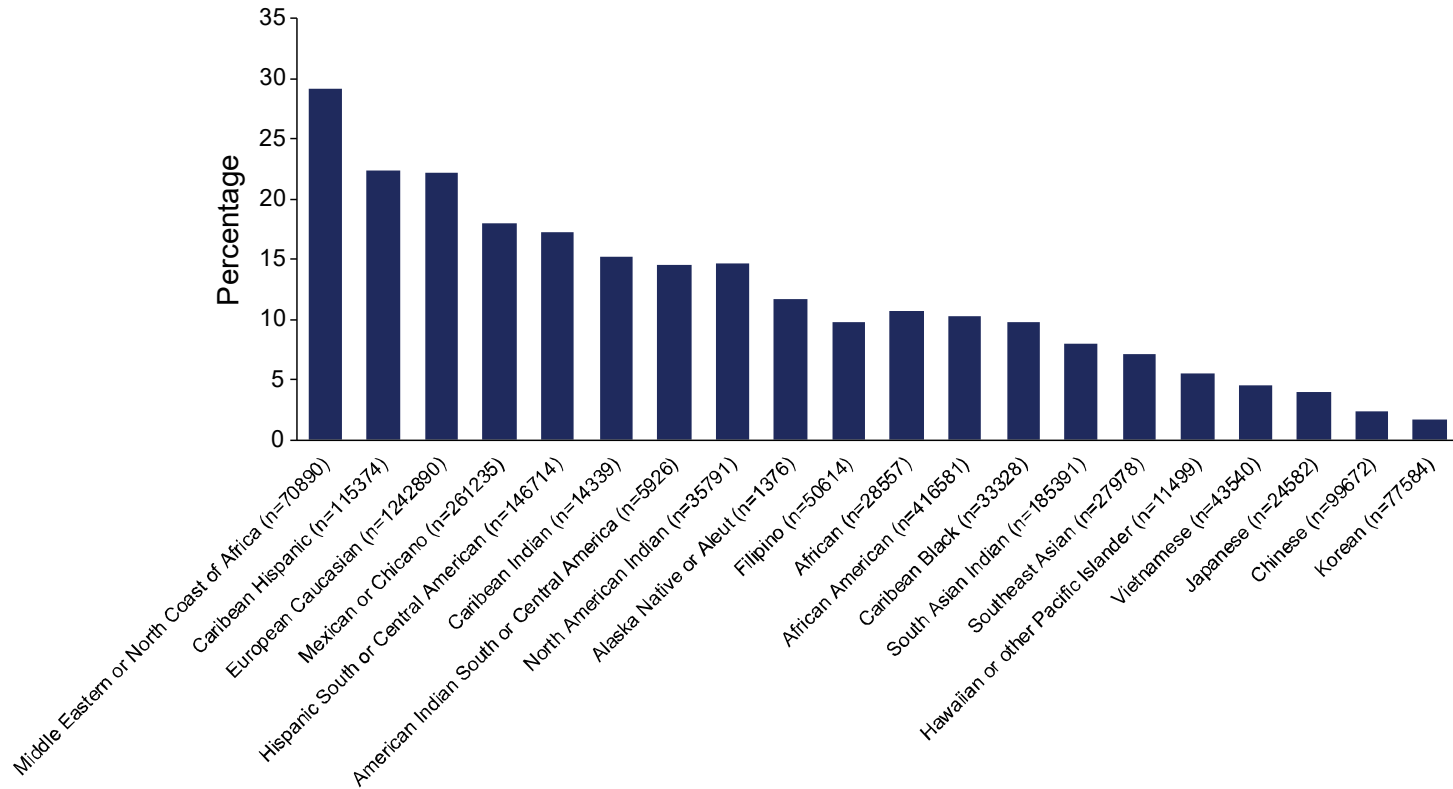


Figure 34 | Distribution of HLA class I subtypes HLA-B*44 associated with ICB response in the NMDP database.

Discussion

As the cost-reduction and technological advancements in WGS, several groups have focused on building the population-specific reference panels, specifically for underrepresented populations in the conventional panels such as the 1KGP3 (7, 9, 10, 12-15, 17). However, the Northeast Asian-specific reference panel with deep sequencing coverage and large samples has been barely created and most of them are not publicly available. In this study, we integrated the WGS variants of 1,779 Northeast Asian individuals to construct a reference panel, NARD, to resolve the uncertainty of the genotype imputation along with the pre-existing panels, and to facilitate more comprehensive genetic analysis of Northeast Asians.

The accuracy of the genotype imputation is one of the most concerns for the genomic reference panels. The imputation accuracy is known to be affected by several factors and one of the major determinants is the sample size of the reference panel (6, 73). Until now, most genotype imputations of Northeast Asians were relied on the panels with a large sample size (38, 39, 43, 63), although the ancestries are not matched between the study population and the reference panel. These panels showed lower imputation accuracy, compared to the well-matched population-specific panels even with smaller sample sizes (7, 9, 12-15, 77).

In terms of panel size, the HRC panel was constructed using the genotypes of more than 30,000 individuals, however mostly composed of European descent from various cohorts including the 1KGP3 study. The previous

studies demonstrated the poor imputation performance of this panel for CHN, admixed Africans, and Hispanic/Latino populations, even worse than the 1KGP3 panel (78, 79), and our analysis again supported these results. The investigation of the HRC is reasonably different to other investigations including our study because they only examined the imputation accuracy on European ancestries. Additionally, the TOPMed Freeze5 panel was constructed recently using 125,568 haplotypes from 62,784 individuals, mostly including the European descent followed by African descent (80). It is the largest publicly available reference panel to date, but the lack of EAS is still unsolved.

In the point of Korean genomic dataset, Korea1K project includes 1,094 WGS of individuals of which 1,007 genomes were newly generated (81). They also measured the 79 quantitative traits acquired from the blood and urine of the participants to assess GWAS. However, they failed to find any novel meaningful markers including SVs. While they generated lots of WGS data and did various kind of analyses including panel generation to impute, we cannot utilize their panel enabling further studies.

As several previous studies yield further increment of the imputation accuracy from their constructed panels by combining dataset of the 1KGP3 (7, 9, 12-14), we also confirmed the improvement of the imputation performance by combining the NARD and the 1KGP3 panels using a fast and simple approach as described in the UK10K and IMPUTE2. After merging the NARD and the

1KGP3, we enhanced the power of the merged panel by applying the re-phasing strategy. It is an advanced process that has not been applied in most of the previous studies (7, 9, 12-14), but the HRC study has shown a further improvement of the imputation accuracy with this approach. Based on this strategy, the NARD + 1KGP3 (re-phased) panel produced more accurately imputed genotypes, especially for uncommon variants ($MAF < 5\%$), than the NARD + 1KGP3 panel. We verified that the improvement might be due to haplotype correction in the NARD panel with the assistance of the haplotypes in the 1KGP3 panel.

Considering the importance of population-specific reference panel, we generated a large-scale WGS dataset of KOR and MNG that were not included in the existing databases such as the 1KGP3 panel. We confirmed that KOR and MNG were genetically differentiated from the other East Asian populations. Therefore, the major ancestries in Northeast Asia are finally covered as population-scale by the NARD. In addition to the two populations, JPN, CHN, and HKG were also included to increase the sample size effect for the imputation power and to build the NARD as a reference panel that can be applied to diverse Northeast Asian populations.

The NARD is the most diverse panel of Northeast Asian in terms of population genetic structure. ADMIXTURE analysis showed the unique components for each of MNG, KOR, JPN, and CHN and shared components of them. F_{ST} network presented that East Asian and South Asian are clearly

divided, and MNG grouped with East Asian populations. PCA analysis supported that MNG in East Asian forms the link with Europe, and more specifically in Treemix analysis, MNG is the root that branches the mainland East Asian and the group of KOR and JPN.

The pharmacogenomic analyses highlight the importance of ethnic-specific genetic screening that could enhance the efficiency of various therapies without severe side effects. We evaluated the advantage of the NARD as a population-specific panel for clinical variant interpretation. We also investigated Northeast Asian-specific pharmacogenomic features by different types of variants. We cataloged Northeast Asian-specific SVs and HLA haplotypes related to immunotherapies. Through these analyses, we examined the characteristics of Northeast Asians in terms of pharmacogenomics.

In summary, we generated a large-scale reference panel for Northeast Asians, which will be a highly valuable resource to resolve the current deficiency of Northeast Asian genome data. We believe that our efforts will remarkably contribute to precision medicine in Northeast Asia.

References

1. Yoo SK, Kim CU, Kim HL, Kim S, Shin JY, Kim N, et al. NARD: whole-genome reference panel of 1779 Northeast Asians improves imputation accuracy of rare and low-frequency variants. *Genome Med.* 2019;11(1):64.
2. Schwarze K, Buchanan J, Taylor JC, Wordsworth S. Are whole-exome and whole-genome sequencing approaches cost-effective? A systematic review of the literature. *Genet Med.* 2018;20(10):1122-30.
3. Telenti A, Pierce LC, Biggs WH, di Iulio J, Wong EH, Fabani MM, et al. Deep sequencing of 10,000 human genomes. *Proc Natl Acad Sci U S A.* 2016;113(42):11901-6.
4. Genomes Project C, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, et al. A global reference for human genetic variation. *Nature.* 2015;526(7571):68-74.
5. McCarthy S, Das S, Kretzschmar W, Delaneau O, Wood AR, Teumer A, et al. A reference panel of 64,976 haplotypes for genotype imputation. *Nat Genet.* 2016;48(10):1279-83.
6. Huang J, Howie B, McCarthy S, Memari Y, Walter K, Min JL, et al. Improved imputation of low-frequency and rare variants using the UK10K haplotype reference panel. *Nat Commun.* 2015;6:8111.
7. Bai H, Guo X, Narisu N, Lan T, Wu Q, Xing Y, et al. Whole-genome

sequencing of 175 Mongolians uncovers population-specific genetic architecture and gene flow throughout North and East Asia. *Nat Genet.* 2018;50(12):1696-704.

8. Chiang CWK, Marcus JH, Sidore C, Biddanda A, Al-Asadi H, Zoledziewska M, et al. Genomic history of the Sardinian population. *Nat Genet.* 2018;50(10):1426-34.

9. Genome of the Netherlands C. Whole-genome sequence variation, population structure and demographic history of the Dutch population. *Nat Genet.* 2014;46(8):818-25.

10. Gudbjartsson DF, Helgason H, Gudjonsson SA, Zink F, Oddson A, Gylfason A, et al. Large-scale whole-genome sequencing of the Icelandic population. *Nat Genet.* 2015;47(5):435-44.

11. Mallick S, Li H, Lipson M, Mathieson I, Gymrek M, Racimo F, et al. The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. *Nature.* 2016;538(7624):201-6.

12. Hou L, Kember RL, Roach JC, O'Connell JR, Craig DW, Bucan M, et al. A population-specific reference panel empowers genetic studies of Anabaptist populations. *Sci Rep.* 2017;7(1):6079.

13. Mitt M, Kals M, Parn K, Gabriel SB, Lander ES, Palotie A, et al.

Improved imputation accuracy of rare and low-frequency variants using population-specific high-coverage WGS-based imputation reference panel. *Eur J Hum Genet.* 2017;25(7):869-76.

14. Nagasaki M, Yasuda J, Katsuoka F, Nariai N, Kojima K, Kawai Y, et al. Rare variant discovery by deep whole-genome sequencing of 1,070 Japanese individuals. *Nat Commun.* 2015;6:8018.

15. Sidore C, Busonero F, Maschio A, Porcu E, Naitza S, Zoledziwska M, et al. Genome sequencing elucidates Sardinian genetic architecture and augments association analyses for lipid and blood inflammatory markers. *Nat Genet.* 2015;47(11):1272-81.

16. Sirugo G, Williams SM, Tishkoff SA. The Missing Diversity in Human Genetic Studies. *Cell.* 2019;177(4):1080.

17. Okada Y, Momozawa Y, Sakaue S, Kanai M, Ishigaki K, Akiyama M, et al. Deep whole-genome sequencing reveals recent selection signatures linked to evolution and disease risk of Japanese. *Nat Commun.* 2018;9(1):1631.

18. Liu S, Huang S, Chen F, Zhao L, Yuan Y, Francis SS, et al. Genomic Analyses from Non-invasive Prenatal Testing Reveal Genetic Associations, Patterns of Viral Infections, and Chinese Population History. *Cell.* 2018;175(2):347-59 e14.

19. Cai N, Bigdeli TB, Kretzschmar WW, Li Y, Liang J, Hu J, et al. 11,670 whole-genome sequences representative of the Han Chinese population from the CONVERGE project. *Sci Data*. 2017;4:170011.
20. Kim J, Weber JA, Jho S, Jang J, Jun J, Cho YS, et al. KoVariome: Korean National Standard Reference Variome database of whole genomes with comprehensive SNV, indel, CNV, and SV analyses. *Sci Rep*. 2018;8(1):5677.
21. Lee S, Seo J, Park J, Nam JY, Choi A, Ignatius JS, et al. Korean Variant Archive (KOVA): a reference database of genetic variations in the Korean population. *Sci Rep*. 2017;7(1):4287.
22. Kwak SH, Chae J, Choi S, Kim MJ, Choi M, Chae JH, et al. Findings of a 1303 Korean whole-exome sequencing study. *Exp Mol Med*. 2017;49(7):e356.
23. Howie BN, Donnelly P, Marchini J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet*. 2009;5(6):e1000529.
24. Das S, Forer L, Schonherr S, Sidore C, Locke AE, Kwong A, et al. Next-generation genotype imputation service and methods. *Nat Genet*. 2016;48(10):1284-7.

25. Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J, 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.* 2015;17(5):405-24.
26. Howie B, Marchini J, Stephens M. Genotype imputation with thousands of genomes. *G3 (Bethesda).* 2011;1(6):457-70.
27. Browning BL, Zhou Y, Browning SR. A One-Penny Imputed Genome from Next-Generation Reference Panels. *Am J Hum Genet.* 2018;103(3):338-48.
28. Charlesworth B. Molecular population genomics: a short history. *Genet Res (Camb).* 2010;92(5-6):397-411.
29. Schilling MP, Wolf PG, Duffy AM, Rai HS, Rowe CA, Richardson BA, et al. Genotyping-by-sequencing for *Populus* population genomics: an assessment of genome sampling patterns and filtering approaches. *PLoS One.* 2014;9(4):e95292.
30. Holsinger KE, Weir BS. Genetics in geographically structured populations: defining, estimating and interpreting F_{ST} . *Nat Rev Genet.* 2009;10(9):639-50.

31. Pickrell JK, Pritchard JK. Inference of population splits and mixtures from genome-wide allele frequency data. *PLoS Genet.* 2012;8(11):e1002967.
32. Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* 2009;19(9):1655-64.
33. Thorn CF, Klein TE, Altman RB. PharmGKB: the Pharmacogenomics Knowledge Base. *Methods Mol Biol.* 2013;1015:311-20.
34. Lan T, Lin H, Zhu W, Laurent T, Yang M, Liu X, et al. Deep whole-genome sequencing of 90 Han Chinese genomes. *Gigascience.* 2017;6(9):1-7.
35. International HapMap C. A haplotype map of the human genome. *Nature.* 2005;437(7063):1299-320.
36. Genomes Project C, Abecasis GR, Altshuler D, Auton A, Brooks LD, Durbin RM, et al. A map of human genome variation from population-scale sequencing. *Nature.* 2010;467(7319):1061-73.
37. Pruitt KD, Tatusova T, Maglott DR. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.* 2007;35(Database issue):D61-5.
38. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* 2010;38(16):e164.

39. Glusman G, Caballero J, Mauldin DE, Hood L, Roach JC. Kaviar: an accessible system for testing SNV novelty. *Bioinformatics*. 2011;27(22):3216-7.
40. Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, et al. Analysis of protein-coding genetic variation in 60,706 humans. *Nature*. 2016;536(7616):285-91.
41. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, et al. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res*. 2001;29(1):308-11.
42. Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alföldi J, Wang Q, et al. Variation across 141,456 human exomes and genomes reveals the spectrum of loss-of-function intolerance across human protein-coding genes. *bioRxiv*. 2019;2019:531210.
43. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GR, Thormann A, et al. The Ensembl Variant Effect Predictor. *Genome Biol*. 2016;17(1):122.
44. MacArthur DG, Balasubramanian S, Frankish A, Huang N, Morris J, Walter K, et al. A systematic survey of loss-of-function variants in human protein-coding genes. *Science*. 2012;335(6070):823-8.
45. Adzhubei I, Jordan DM, Sunyaev SR. Predicting functional effect of

human missense mutations using PolyPhen-2. *Curr Protoc Hum Genet.* 2013;Chapter 7:Unit7 20.

46. Liu X, Wu C, Li C, Boerwinkle E. dbNSFP v3.0: A One-Stop Database of Functional Predictions and Annotations for Human Nonsynonymous and Splice-Site SNVs. *Hum Mutat.* 2016;37(3):235-41.

47. Landrum MJ, Lee JM, Riley GR, Jang W, Rubinstein WS, Church DM, et al. ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res.* 2014;42(Database issue):D980-5.

48. Rausch T, Zichner T, Schlattl A, Stutz AM, Benes V, Korbel JO. DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics.* 2012;28(18):i333-i9.

49. Ng KP, Hillmer AM, Chuah CT, Juan WC, Ko TK, Teo AS, et al. A common BIM deletion polymorphism mediates intrinsic resistance and inferior responses to tyrosine kinase inhibitors in cancer. *Nat Med.* 2012;18(4):521-8.

50. MacDonald JR, Ziman R, Yuen RK, Feuk L, Scherer SW. The Database of Genomic Variants: a curated collection of structural variation in the human genome. *Nucleic Acids Res.* 2014;42(Database issue):D986-92.

51. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA,

et al. The variant call format and VCFtools. *Bioinformatics*. 2011;27(15):2156-8.

52. Wigginton JE, Cutler DJ, Abecasis GR. A note on exact tests of Hardy-Weinberg equilibrium. *Am J Hum Genet*. 2005;76(5):887-93.

53. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience*. 2015;4:7.

54. Jun G, Wing MK, Abecasis GR, Kang HM. An efficient and scalable analysis framework for variant extraction and refinement from population-scale DNA sequence data. *Genome Res*. 2015;25(6):918-25.

55. Yang J, Lee SH, Goddard ME, Visscher PM. GCTA: a tool for genome-wide complex trait analysis. *Am J Hum Genet*. 2011;88(1):76-82.

56. Ju YS, Kim JI, Kim S, Hong D, Park H, Shin JY, et al. Extensive genomic and transcriptional diversity identified through massively parallel DNA and RNA sequencing of eighteen Korean individuals. *Nat Genet*. 2011;43(8):745-52.

57. Seo JS, Rhie A, Kim J, Lee S, Sohn MH, Kim CU, et al. De novo assembly and phasing of a Korean human genome. *Nature*. 2016;538(7624):243-7.

58. Lu D, Lou H, Yuan K, Wang X, Wang Y, Zhang C, et al. Ancestral Origins and Genetic History of Tibetan Highlanders. *Am J Hum Genet.* 2016;99(3):580-94.
59. Cann HM, de Toma C, Cazes L, Legrand MF, Morel V, Piouffre L, et al. A human genome diversity cell line panel. *Science.* 2002;296(5566):261-2.
60. Manichaikul A, Mychaleckyj JC, Rich SS, Daly K, Sale M, Chen WM. Robust relationship inference in genome-wide association studies. *Bioinformatics.* 2010;26(22):2867-73.
61. Xie C, Yeo ZX, Wong M, Piper J, Long T, Kirkness EF, et al. Fast and accurate HLA typing from short-read next-generation sequence data with xHLA. *Proc Natl Acad Sci U S A.* 2017;114(30):8059-64.
62. Gourraud PA, Khankhanian P, Cereb N, Yang SY, Feolo M, Maiers M, et al. HLA diversity in the 1000 genomes dataset. *PLoS One.* 2014;9(7):e97282.
63. Bray RA, Hurley CK, Kamani NR, Woolfrey A, Muller C, Spellman S, et al. National marrow donor program HLA matching guidelines for unrelated adult donor hematopoietic cell transplants. *Biol Blood Marrow Transplant.* 2008;14(9 Suppl):45-53.
64. Wang J, Raskin L, Samuels DC, Shyr Y, Guo Y. Genome measures

used for quality control are dependent on gene function and ancestry. *Bioinformatics*. 2015;31(3):318-23.

65. O'Connell J, Sharp K, Shrine N, Wain L, Hall I, Tobin M, et al. Haplotype estimation for biobank-scale data sets. *Nat Genet*. 2016;48(7):817-20.

66. consortium C. Sparse whole-genome sequencing identifies two loci for major depressive disorder. *Nature*. 2015;523(7562):588-91.

67. Son HY, Hwangbo Y, Yoo SK, Im SW, Yang SD, Kwak SJ, et al. Genome-wide association and expression quantitative trait loci studies identify multiple susceptibility loci for thyroid cancer. *Nat Commun*. 2017;8:15966.

68. Bjelland DW, Lingala U, Patel PS, Jones M, Keller MC. A fast and accurate method for detection of IBD shared haplotypes in genome-wide SNP data. *Eur J Hum Genet*. 2017;25(5):617-24.

69. Browning BL, Browning SR. Improving the accuracy and efficiency of identity-by-descent detection in population data. *Genetics*. 2013;194(2):459-71.

70. Zhao H, Sun Z, Wang J, Huang H, Kocher JP, Wang L. CrossMap: a versatile tool for coordinate conversion between genome assemblies.

Bioinformatics. 2014;30(7):1006-7.

71. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*. 2009;25(16):2078-9.

72. Loh PR, Danecek P, Palamara PF, Fuchsberger C, Y AR, H KF, et al. Reference-based phasing using the Haplotype Reference Consortium panel. *Nat Genet*. 2016;48(11):1443-8.

73. Delaneau O, Zagury JF, Marchini J. Improved whole-chromosome phasing for disease and population genetic studies. *Nat Methods*. 2013;10(1):5-6.

74. Boycott KM, Vanstone MR, Bulman DE, MacKenzie AE. Rare-disease genetics in the era of next-generation sequencing: discovery to translation. *Nat Rev Genet*. 2013;14(10):681-91.

75. Yang SK, Hong M, Baek J, Choi H, Zhao W, Jung Y, et al. A common missense variant in NUDT15 confers susceptibility to thiopurine-induced leukopenia. *Nat Genet*. 2014;46(9):1017-20.

76. Chowell D, Morris LGT, Grigg CM, Weber JK, Samstein RM, Makarov V, et al. Patient HLA class I genotype influences cancer response to checkpoint blockade immunotherapy. *Science*. 2018;359(6375):582-7.

77. Yasuda J, Katsuoka F, Danjoh I, Kawai Y, Kojima K, Nagasaki M, et al. Regional genetic differences among Japanese populations and performance of genotype imputation using whole-genome reference panel of the Tohoku medical Megabank Project. *BMC Genomics*. 2018;2018;19:551.
78. Lin Y, Liu L, Yang S, Li Y, Lin D, Zhang X, et al. Genotype imputation for Han Chinese population using Haplotype Reference Consortium as reference. *Hum Genet*. 2018;137(6-7):431-6.
79. Kowalski MH, Qian H, Hou Z, Rosen JD, Tapia AL, Shan Y, et al. Use of >100,000 NHLBI Trans-Omics for Precision Medicine (TOPMed) Consortium whole genome sequences improves imputation quality and detection of rare variant associations in admixed African and Hispanic/Latino populations. *bioRxiv*. 2019;683201.
80. Taliun D, Harris DN, Kessler MD, Carlson J, Szpiech ZA, Torres R, et al. Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *bioRxiv*. 2019:563866.
81. Jeon S, Bhak Y, Choi Y, Jeon Y, Kim S, Jang J, et al. Korean Genome Project: 1094 Korean personal genomes with clinical information. *Science Advances*. 2020;6(22):eaaz7835.

국문 초록

1,779명 동북아시아인의 전장 유전체 데이터를 기반으로 한 참조 패널 생성과 유전학적 인구 특성 구조 및 약리 유전체학 프로파일의 연구

서울대학교 대학원 의과학과 의과학 전공

김창욱

서론: 전장 게놈 해독 (WGS)의 비용 감소와 생산량 증가로 인간 게놈 연구의 대상자 수가 점차 늘어나고 있다. 특정 인구에 대한 대규모의 WGS는 인간 대상 유전체학 연구에서 매우 중요하며, 더 나아가 정밀의학 실현을 위한 인구집단에 대한 집단 유전체학적 이해와 약리 유전체학적 프로파일의 정확한 구축을 위해 대규모 WGS 데이터의 필요성은 지속해서 대두되고 있다. 하지만 대부분의 그 연구의 대상이 유럽인 중심으로 편중 되어있는 것이 현실이다.

방법: 우리는 한국인과 몽골인, 일본인, 중국인, 홍콩인 1,779 명으로부터 생산한 전장 유전체 서열 분석 데이터를 활용하여 NARD (Northeast Asian Reference Database)를 구축하였다. NARD 는 1000 게놈 프로젝트 3 단계 (1KGP3)에 포함되지 않았던 한국계와 몽골계 인구의 새로운 유전적 다양성을 제공한다. 우리는 NARD 와 1KGP3 의 유전자형 데이터를 병합하고 re-phasing 방법으로 높은 성능의 통합 데이터 세트를 생성하였다.

앞서 한국과 몽골 인구에 대해 NARD 정도의 규모와 정밀성을 갖춘 데이터가 발표된 적은 없었다. NARD 는 앞으로 동북아시아의 유전체학 분야에 더 정확하고 새로운 통찰을 제공할 것이다. 우리는 이 데이터를 토대로 한 PCA, F_{ST} 분석과 계통수 분석을 통해 인구 유전체학적 연구를 진행하였다.

우리는 또한 동북아시아의 약리학적 특성을 밝히는 시도를 하였다. 약물 반응과 관련된 단일 염기 다형성 (SNP) 및 *BCL2L11* (*BIM*) 인트론 영역의 결손을 포함하는 구조적 변이, 면역 체크 포인트 차단 (ICB)에 대한 효험과 관련이 있는 HLA 영역을 포함한 동북아시아 특이적 변이를 조사하였다.

결과: re-phasing 방법으로 병합된 NARD 와 1KGP3 의 패널을 이용한 동아시아인 대상 imputation 은 기존 패널들의 성능과 비교하여 가장 높은 정확도를 보였으며 특히 희귀 변이와 저 빈도 변이에 대해 그 향상이 두드러졌다.

우리는 인구 구조 분석을 통해 기존에 알려진 것과 달리 한국인, 몽골인, 일본인과 중국인 및 동남아시아인 사이에 뚜렷한 차이가 존재한다는 것을 확인할 수 있었다.

NARD 에서 일정 이상 빈도로 존재하는 변이는 환자를 대상으로 한 검사나 연구에서 단백질을 변형시키는 변이의 허위 후보를 제거하는데 활용될 수 있다. NARD 에서는 총 1,480 만여 개의 기존에 보고되지 않았던 신규 변이가 발견되었다. 그리고 약리 유전체학적 분석에서 타이로신 키나아제와 면역 체크 포인트 억제제의 효과 감소가 다른 지역에 비해 동북아시아에서 더 빈번하게 나타남을 보였다. NARD 참조 패널은 <https://nard.macrogen.com/> 에서 임퓨테이션 파이프라인과 함께 제공된다.

결론: 우리는 동북아시아인 대상으로 가장 정확한 참조 패널을 구성하였다. 이 참조 패널은 연구목적으로 누구나 쉽게 사용할 수

있게 웹을 통해 제공된다. 또한 동북아시아의 인구 구조 및 약물 유전체학적으로 더욱 정밀한 통찰을 제공했다. 우리의 연구는 앞으로 동북아시아 정밀 의학 시대를 열기 위한 추가적인 연구의 초석이 될 것이다.

*본 내용은 Genome Medicine 에 출판이 완료된 내용임 (1)

주요어: 집단 유전체학, 약리 유전체학, 참조 패널, 유전형
임플리케이션, 전장 유전체 해독, 동북아시아, 동아시아

꿈을 이루어 가는 여정이 한없이 즐겁습니다.

그 과정에 학위를 받게 되는 것은 하나의 행운입니다.

길은 생각보다 길고, 굽이집니다.

순간들을 아내 박선재와 함께했고, 영원히 그럴 것입니다.

딸 연서와 예서, 하고 싶은 일들을 재미있게 해나가거라.

모든 순간에 저를 있게 한 것은 아버지 김영근, 어머니 노현숙,

그리고 동생 김창민입니다.

제자로 받아 주신 서정선 교수님께 존경과 감사를 표합니다.