KoVariome: Korean National Standard Reference Variome database of whole genomes with comprehensive SNV, indel, CNV, and SV analyses

Jungeun Kim[1†], Jessica A. Weber[2†], Sungwoong Jho[1†], Jinho Jang[3,4], JeHoon Jun[1, 5], Yun Sung Cho[5], Hak-Min Kim[3,4], Hyunho Kim[5], Yumi Kim[5], OkSung Chung[1,5], Chang Geun Kim[6], HyeJin Lee[1], Byung Chul Kim[7], Kyudong Han[8], InSong Koh[9], Kyun Shik Chae[6], Semin Lee[3,4], Jeremy S. Edwards[10,*], and Jong Bhak[1,3,4,5,*]

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

[2] Department of Biology, University of New Mexico, Albuquerque, NM 87131, USA.

[3] Department of Biomedical Engineering, School of Life Sciences, Ulsan National Institute of Science and Technology (UNIST), Ulsan 44919, Republic of Korea

[4] The Genomics Institute, Ulsan National Institute of Science and Technology (UNIST), Ulsan 44919, Republic of Korea

[5] Geromics, Ulsan 44919, Republic of Korea.

[6] National Standard Reference Center, Korea Research Institute of Standards and Science, Daejeon 34113, Republic of Korea.

[7] Clinomics, Ulsan 44919, Republic of Korea.

[8] Department of Nanabiomedical Science & BK21 PLUS NBM Global Research Center for Regenerative Medicine, Dankook University, Cheonan 31116, Republic of Korea.

[9] Department of Physiology, College of Medicine, Hanyang University, Seoul 04763, Republic of Korea.

[10] Chemistry and Chemical Biology, UNM Comprehensive Cancer Center, University of New Mexico, Albuquerque, NM 87131, USA.

†Contributed equally

*Corresponding authors

E-mail: jongbhak@genomics.org (JB), jsedwards@salud.unm.edu (JE).

Abstract

High-coverage whole-genome sequencing data of a single ethnicity can provide a useful catalogue of population-specific genetic variations. Herein, we report a comprehensive analysis of the Korean population, and present the Korean National Standard Reference Variome (KoVariome). As a part of the Korean Personal Genome Project (KPGP), we constructed the KoVariome database using 5.5 terabases of whole genome sequence data from 50 healthy Korean individuals with an average coverage depth of 31×. In total, KoVariome includes 12.7M single-nucleotide variants (SNVs), 1.7M short insertions and deletions (indels), 4K structural variations (SVs), and 3.6K copy number variations (CNVs). Among them, 2.4M (19%) SNVs and 0.4M (24%) indels were identified as novel. We also discovered selective enrichment of 3.8M SNVs and 0.5M indels in Korean individuals, which were used to filter out 1,271 coding-SNVs not originally removed from the 1,000 Genomes Project data when prioritizing disease-causing variants. CNV analyses revealed gene losses related to bone mineral densities and duplicated genes involved in brain development and fat reduction. Finally, KoVariome health records were used to identify novel disease-causing variants in the Korean population, demonstrating the value of high-quality ethnic variation databases for the accurate interpretation of individual genomes and the precise characterization of genetic variations.

Introduction

The human reference genome[1] was a milestone of scientific achievement and provides the foundation for biomedical research and personalized healthcare[2]. The completion of the human genome marked the beginning of our concerted efforts to understand and catalogue genetic variation across human populations. The International HapMap project resolved human haplotypes into more than one million common single nucleotide polymorphisms (SNPs) in an effort to catalogue genetic variations associated with diseases[3]. Subsequently, other large-scale genomic studies have identified 360M copy number variations (CNVs)[4] and 6.4M small insertions and deletions (indels)[5]. These efforts laid the groundwork for approximately 1,800 genome-wide association (GWA) studies that investigated the genetic basis of complex diseases such as diabetes, cancer, and heart disease[6]. While these GWA studies have identified a wide range of disease-associated alleles that can be used as diagnostic tools[7], the majority of the findings are associated with low disease risks and have led to a renewed focus on the detection of rare variants that are more predictive of disease[8].

To identify pathogenic rare variants in GWA studies, disease cohorts are compared to population-scale variomes generated from healthy controls to remove common and low frequency variants in diverse human ethnic groups[9,10]. As a result, numerous population genomic studies have been performed to characterize ethnicity-relevant variations. One of the largest of such efforts, the 1,000 Genomes Project (1000GP), reported a total of 88M genetic variants, including SNPs, indels, and structural variations (SVs) from 2,504 healthy individuals[11], and resolved population stratification by sampling 26 populations across five continents; Africa (AFR), East Asia (EAS), Europe (EUR), South Asia (SAS), and Americas (AMR). More recently, the Exome Aggregation Consortium (ExAC) released ten million human genetic variants from 60,706 individuals with a resolution of one exonic variant for every eight base-pairs[12]. Analysis of high coverage sequencing data (more than 30x) from 10,000 individuals showed that each newly

analyzed genome added roughly 0.7MB of new sequences to the human reference genome and contributed an average of 8,579 new SNVs to the existing human variation data set [13]. Large-scale variome studies, such as those previous discussed, have significantly increased our understanding of variation in the human population, however, the population composition is still broadly biased towards Europeans (54.97% in ExAC[12] and 78.55% in Telenti *et al.*[13]). Consequently, many groups have initiated small variome studies of more targeted populations, i.e. the Malays[14], Dutch (GoNL)[15], Danish[16], Japanese (1KGPN)[17], Finland, and United Kingdom[18]. The large number of population-specific variations discovered in these studies highlights the importance of single population variomes in creating comprehensive databases of population heterogeneity and stratification.

SVs are also an important type of genomic variation in the human population that contribute significantly to genomic diversity[19]. SVs include large insertions (INSs), deletions (DELs), inversions (INVs), translocations, and CNVs[20]. Unlike SNVs and small indels, however, the identification of SVs remains challenging largely because of genome complexities and the limitations of short-read sequencing technologies[21]. Current efforts to resolve SVs reported several population-scale SVs[16,19] and CNVs[17,22] from whole genome sequencing (WGS) data, and these analyses characterized population-specific traits such as amylase gene duplication in high-starch diet populations[17,23] and revealed associations for specific diseases such as hemophilia A[24], hunter syndrome[25], autism[26], schizophrenia[27], and Crohn's disease[28]. Nevertheless, SVs identified in healthy individuals also contain a substantial number of individual- and population-specific SVs with no disease association. Taken together, these results have demonstrated the importance of constructing population-specific SV and CNV profiles for the precise characterization of disease association and identifying diagnostic markers for personalized medicine.

The Korean population is regarded as a relatively homogeneous ethnic group in East Asia[29], from which a relatively small set of samples can produce a high-coverage population

variome. Since the first Korean whole genome sequences were reported in 2009[30], further variome studies in the Korean population have been conducted in the last decade using low-cost next generation sequencing (NGS) technologies[31-36]. Two exonic variomes of more than 1,000 Koreans were reported, though sampling was focused on disease cohorts containing patients with type II diabetes mellitus, hemophilia, cancer, and other rare diseases [35,36]. Consequently, these studies are not suitable for parsing benign, demographic variants from disease variants. As the Korean center of the Personal Genome Project (PGP)[17], the Korean Personal Genome Project (KPGP or PGP-Korea) was initiated in 2006 by the Korean Bioinformation Center (KOBIC) to resolve ethnicity-relevant variation in Korea by providing a comprehensive genomic, phenomic, and enviromic dataset accessible to researchers across the world. In 2009, KPGP published the first Korean genome with NGS data[30] and the number of complete genomes has increased to 60 genomes as of 2016. This population was used to construct the first Korean Reference genome standard (KOREF)[37], which was registered as a standard reference data for ethnic Korean genome sequence by evaluating its traceability, uncertainty, and consistency in the beginning of 2017.

To characterize the genomic variations across the Korean population, we selected and analyzed WGS data from 50 unrelated, healthy Korean individuals in KPGP cohorts with associated clinical diagnoses and family histories related to major diseases. In this report, we describe the general features of KoVariome and characterize all four types of genomic variations, which include 12.7M SNVs, 1.7M indels, 4K SVs, and 3.6K CNVs. This comprehensive database of genomic variations and corresponding metadata will be a valuable resource to the genomic community as researchers search for the genetic basis of disease.

Results and Discussion

Construction of the Korean standard Variome: KoVariome

Since 2010, the Korean variome data center, as a part of the KPGP, has been recruiting volunteers to generate WGS and whole exome sequencing (WES) data. The current KoVariome (version 20160815) has been constructed based on WGS data from 50 unrelated Korean individuals who responded to questionnaires detailing body characteristics, habits, allergies, family histories, and physical conditions related to 19 disease classes (Supplementary Table S1). A total of 5.5 TB of high-quality paired-end WGS data were generated, containing an average of 31× coverage per individual (Table 1 and Supplementary Table S2). WGS data from each individual covered 95% of the human reference genome (hg19) on average. From these data, we identified approximately 3.8M SNVs (ranged 3.7-3.9M) and 0.5M indels (0.4-0.7M) per Korean individual (Table 1 and Supplementary Fig. S1A). The hetero-to-homozygosity ratio of the autosomal SNVs was 1.49, which is consistent with previously reported data[38]. The length distributions of the indel loci were symmetric, with the majority of indel sizes shorter than six bases (94.8% for insertions, 97.8% for deletions) (Supplementary Fig. S1B). We identified approximately 20,097 (0.53%) SNVs and 258 (0.05%) indels in the coding regions including 10,394 (0.22%) non-synonymous changes per individual (Table 1).

Novel KoVariome SNVs were counted by adding individual samples one by one (Fig. 1A), and the number of novel SNVs logarithmically decreased and became depleted after the 9th donor. In total, we observed 59K novel SNVs, including 1.2K (2.03%) coding-SNVs, per individual. To assess the relatedness of the KoVariome individuals, we compared the pairwise genetic distance of KoVariome with those of family data (Fig. 1B). WGS data from thirty families were downloaded from the KPGP database, which included two monozygotic twins, 14 parent-children pairs, seven siblings, five grandparents-grandchildren, six uncles-nephews, and three

cousins. We analyzed familial SNVs using the same method as in KoVariome and also compared genetic distances between the two groups (see Methods). The genetic distance among KoVariome individuals was higher (pi=8.8e-4) than those found in the familial data, such as monozygotic twins (4.8e-4), siblings (6.7e-4), parent-child (6.8e-4), uncle-nephew (7.7e-4) and grandparents-grandchild (7.8e-4), and cousins (8.2e-4); verifying that no genetic bias was present in the sample collection stage. In accordance with previous reports, the multidimensional scaling (MDS) of the variants among Korean, Chinese, and Japanese individuals showed a clear separation of the three populations (Fig. S2) despite the geographical and historical associations between these groups[35,37]. These analyses reinforce the need for distinct KOREF and KoVariome reference resources to parse disease variants from demographic variants in this population.

Accuracy test of SNVs and indels in KoVariome

We evaluated the accuracy of KoVariome SNV and indel predictions by comparing genotype results from the Axiom™ Genome-ASI 1 Array with WGS data from 35 individuals. A total of 503,694 SNV positions were compared, from which we obtained an average of 0.9993 precision (ranged: 0.9984-0.9996) and 0.9980 recall (ranged: 0.9817-0.9994) (Supplementary Table S3). In addition, there was a 99.65% (ranged: 98.62-99.87%) concordance of the SNVs called by the WGS and Axiom array calls. Compared to similar variome studies, this genotype accuracy was slightly lower than the high-depth trio data in the Danish population study (99.8%)[16] but higher than that of the Dutch population SNVs (99.4-99.5%) analyzed with intermediate depths[39]. The accuracy of the SNV calls was analyzed across the genome, and a total of 499,889 (99.24%) SNVs showed a genotype concordance higher than 0.99, while 0.4% of SNVs showed the genotype accuracy less than 0.95 (Supplementary Table S4). Similar levels of genotype concordances were observed in the repetitive regions of the genome (99.56% of SNVs with the

genotype correspondence > 0.95, Supplementary Table S5), suggesting that SNV calling accuracy is not reduced in repetitive regions of the genome.

We also compared the accuracy of indel variant calls with the 1,981 indel markers on the Axiom™ Genome-ASI 1 Array. A genotype comparison showed an average accuracy of 98.49% for indels, which was slightly lower than those observed in SNVs (Supplementary Table S3), and comparable to the false positive (FP) rate for indels that was reported in the Danish data[16]. In terms of genomic loci, 1,343 (91.11%) indels showed perfect genotype concordance with array data and 1,446 (98.10%) indels had an accuracy higher than 90% (Supplementary Fig. S3).

Genome-wide features of KoVariome

By merging the variants of 50 unrelated Korean individuals, we identified 12.7M SNVs and 1.7M small indels shorter than 100bp (Table 1); approximately 1.5 times the number of SNVs previously reported from preliminary KPGP data (0.8M)[33]. Both types of variants were primarily distributed in the non-coding regions (about 98%), including intergenic and intron regions (Supplementary Table S6). Approximately 10.3M (81.10%) SNVs and 1.3M (76.47%) indels were present in dbSNP (ver. 146); while 2.4M SNVs and 0.4M indels were novel (Table 1). A total of 9M (70.42%) SNVs and 0.8M (48.68%) indels were found in the 1000GP variome (Supplementary Table S6); and based on allele frequencies, 4.6M (51.03%) and 4.4M (48.82%) of these SNVs were classified into the categories '1000GP common' and '1000GP low frequency', respectively (Fig. 2A). Most notably, 13,584 (0.15%) KoVariome SNVs were rarely observed in the 1000GP continental groups with a MAF < 0.1%. A similar distribution was observed with the indels, where 64.2% and 35.8% of the KoVariome indels were classified into the '1000GP common' (0.5M) and '1000GP low frequency' (0.3M) classes, respectively. Only ten indels were classified into the '1000GP rare' category. Almost all of the variants in the '1000GP common' category were also frequently observed in KoVariome, representing 4.5M

(98.33%) SNVs and 0.5M (93.37%) indels in this class (Fig. 2A and Supplementary Table S6). Surprisingly, however, roughly half of the variants in '1000GP low frequency' were classified as 'frequent in KoVariome'.

Next, we compared the allele frequencies in five the continental 1000GP groups to KoVariome. In total, we observed 3.4M (77.19%) SNVs and 0.2M (74.21%) indels that were statistically enriched in at least one of the continental groups or the Korean population (Fig. 2B), suggesting a population stratification. To further explore the population stratification, we identified the variants uniquely enriched in each continental group, and the enriched variants that were in common between the continental groups. In total, nearly three million (2.7M) SNVs and 156K indels were frequently found in the Korean population. Among them, 2.5M (95.20%) SNVs and 143K (94.47%) indels showed Korean specific enrichments, while the other enriched variants were shared by other continents (Figure 2B). Among the five continental groups, as expected, EAS shared the largest number of enriched variants (89.5K SNPs and 5.3K indels) with the Korean population [40].

Interpretation of the KoVariome-specific variants

Characterizing ethnicity specific variants is necessary to understand the demographic differences between populations and can be used to filter out low frequency clustered variants in a specific group. In KoVariome, there were 3.8M SNVs and 0.9M indels not observed in the 1000GP variome (Supplementary Table S6). Among them, 1.1M (29.16%) SNVs and 0.4M (40.88%) indels were classified as 'frequent in KoVariome' (Fig. 2A). Of the 15,279 non-synonymous SNVs and 480 frame-shift indels specific to KoVariome, 11,746 (76.88%) and 397 (82.71%) were rare in KoVariome ($n < 3$), respectively; whereas 3,533 SNVs were frequently observed (occurring at least three times) in KoVariome but not observed in the 1000GP variome.

To identify the clinical relevance of these variants, we compared the genomic loci of these SNVs against the ClinVar database and identified six pathogenic or likely pathogenic loci with associated disease information (Table 2). Two pathogenic SNVs (rs386834119 and rs1136743) were autosomal recessive (AR) diseases, and therefore, no phenotypes were expected since all of the KoVariome SNVs were heterozygotes in the KPGP. We observed a high allele frequency (three males and two females) of the cancer-associated SNV (rs200564819) in *RAD51*, which is known to increase the risk of developing ovarian and breast cancers[41]. While the inheritance type is not available in the Online Mendelian Inheritance in Man (OMIM), we speculate that it is autosomal dominant (AD) with incomplete penetrance since four out of 14 male *RARD51*-deficient carriers (heterozygous) were diagnosed with colorectal cancers. However, none of the donors with this SNV have been diagnosed cancer and have no familial cancer history. We also observed a splicing-donor ('GU') candidate five nucleotides downstream of this SNV, although further confirmation is required. This site may induce the null effect of rs200564819 by the creation of new splice-sites according to the guidelines from the American College of Medical Genetics and Genomics (ACMC)[42]. In addition, we observed two pathogenic missense SNVs (rs121912678 and rs20016664) associated with fibrodysplasia ossificans progressive (FOP) and Van der Woude syndrome (VWS), respectively (Table 2). A chr2:g158630626C>G SNV was rarely observed in the ExAc database (MAF=0.0002) and another variant (C>T) at this position revealed a pathogenic effect for FOP disease by changing R206H in the *activin receptor type I* (*ACVR1*) gene[43]. While the pathogenicity of R206P in *ACVR1* due to a C>G mutation is not yet known, we suggest that it is likely benign because of the high MAF (0.14) of this allele without any FOP phenotypes, skeletal malformation, or progressive extraskeletal ossification recorded in the KPGP survey. Furthermore, the 400th amino acid of the *interferon regulatory factor 6* (*IRF6*) gene is known to be a hot spot of VWS, orofacial clefting disorders. Two pathogenic residues, R400W[44] and R400Q[45], were reported for VWS; however, the pathogenicity of R400P arisen by chr1:209961970C>G is not yet confirmed. A total of 14 heterozygous SNVs had no phenotype

for VWS symptom, despite the AD inheritance pattern of this disease; and consequently, the R400P substitution also seems to be benign. Taken together, the KoVariome-specific frequent variants demonstrate the importance of using population-scale health data to identify pathogenic loci in specific diseases, and for the accurate identification of benign variants that are not annotated because of population stratification.

Functional impact of rare variants

We investigated the proportion of the SNVs in four SNV classes (1000GP Common, 1000GP Low Frequency, 1000GP Rare, KoVariome Specific; Supplementary Fig. S4). Our analyses showed that a higher portion of the coding SNVs were enriched in the '1000GP rare' class, while the SNVs in the non-coding regions were similarly distributed in all other variant classes. The portion of non-synonymous SNVs in the '1000GP rare' class was more than twice what was observed in the other classes. It is possible that these patterns are associated with purifying selection to rapidly remove deleterious alleles in the population[46], though it was not possible to identify this pattern in frame-shift indels because of the small number of variants (981) in this class. To analyze the tendencies of purifying selection in KoVariome, we defined rare variant ratios (RVRs) as the number of SNVs in the 'rare in KoVariome' class divided by the number of SNVs in the 'frequent in KoVariome' class. We then compared RVRs across genomic regions (Fig. 2C). In both SNVs and indels, RVRs in the intergenic region were lowest (0.66), while similar levels of RVRs were observed in other non-coding regions (0.66-0.87). Under the assumption that mutations occur randomly throughout the genome, lower rates of RVR in non-coding regions suggest neutral selection with no or weak selection pressures in the population. Conversely, the highest RVR of frame-shift indels (1.45) suggests there was some purifying selection against these variants in the Korean population. Furthermore, about twice as many RVRs were observed in the non-synonymous (1.16) and splice-site (1.33) SNVs compared to

intergenic regions. Although SNVs in the coding region can be deleterious to protein function, selection pressure on the non-synonymous and splice-site SNVs seem to be slightly lower than that of the frame-shift indels.

Interpretation of disease-causing variants among Korean individuals

Rare SNVs in an individual genome are more likely to be pathogenic than common variants. Because genetic variants are known to be geographically clustered, characterizing population stratification is a critical first step to identifying disease-causing variants[47]. With this concept, we examined rare SNVs in each individual after filtering out common SNVs that were classified as '1000GP common', '1000GP low frequency', or 'frequent' in KoVariome. From an average of 3.8M SNVs per individual, 3.4M (88.70%) and 0.4M (9.39%) SNVs were filtered out using the 1000GP variome or KoVariome, respectively (Fig. 3A and Table 3). Overall, KoVariome allowed 1,231 (12.25%, median value) non-synonymous SNVs and 40 (24.01%) splice-site SNVs to be filtered out as common variants in the Korean population, which significantly improves the ability to pin-point disease causative variants.

After filtering, Korean donors had a median of 47,957 (1.26%) rare SNVs, most of which (98.33%) were located in non-coding regions. Among these rare SNVs, we observed an average of 219 (67.17%) non-synonymous SNVs and seven (0.87 %) splice-site SNVs per individual (Fig. 3B and Table 2). On average, 166 (73.45%) of these SNVs were present in dbSNP (ver. 146), but not in the 1000GP variome (Fig. 3C). Of the 12,445 non-synonymous rare SNVs distributed in 50 Korean individuals, we identified 7,645 (61.43%) pathogenic or probably pathogenic SNVs predicted by at least one computational algorithm (see methods section, Table S7). In total, 38 (0.5%) pathogenic rare SNVs in KoVariome were homozygotes and the remaining (99.5%) were heterozygotes. In addition, 29 (58%) of the donors had no homozygous pathogenic rare SNVs. To obtain clinical information concerning these pathogenic rare-SNVs, we searched the genomic loci

for these SNVs against the ClinVar database. A total of 127 of the rare SNVs were found in ClinVar, 53 of which showed clear clinical significance. Eight (6.39%) and thirteen (10.24%) were listed as benign and likely benign in ClinVar, respectively, and not fatal for a specific disease. Conversely, 29 (22.83%) and three (2.36%) were pathogenic and likely pathogenic, respectively (Table 4). These rare SNVs contribute to disease according to their inheritance patterns, and a manual investigation of the inheritance type using the OMIM database identified seven AD and 17 AR SNVs for specific loci; although we failed to identify the inheritance types for eight SNV loci (Table 4). All 17 of the AR SNVs were heterozygous in KoVariome, so it was not possible to assign phenotypes to these loci. Within the donor group with pathogenic rare AD SNVs, we searched for phenotypes or familial histories associated with target diseases in the questionnaire. We identified a familial history for type II diabetes mellitus associated with rs121918673 allele KPGP participants; however, one donor with the rs121918673 allele was nondiabetic and reported no family history of this disease Additionally, one donor was heterozygous for the rs121912749 allele, which has been associated with spherocytosis, and this donor reported associated symptoms but no anemia (Supplementary Table S1 and S7). However, it is clinically known that spherocytosis has heterogenetic symptoms ranging from asymptomatic to hemolytic anemia. These examples highlight the disease-relevant genetic information this resource can provide to patients, and emphasize the utility of KoVariome to the Korean population at large as WGS becomes a more routine component of healthcare.

Structural variations in KoVariome

SVs are common across the human genome, though identifying and defining the impact of SVs is more difficult than SNVs or indels (<100bp). We predicted on average 6,534 individual SVs, including 450 INVs, 354 intra-chromosomal translocations (ITXs), 478 INSs, and 5,252 DELs using BreakDancer (BD) and Pindel programs (Supplementary Table S8). To identify SVs with

clear break points, we removed 15-32% spurious SVs per individual (see Methods; Supplementary Fig. S5 and Table S8). After filtering, we obtained 40,179 non-redundant SVs; including 4,896 INVs, 2,131 ITXs, 12,171 INSs, and 20,981 DELs. Within the Korean donor group, individuals contained 3,294 SVs (median), 82.36% of which were DELs (Fig. 4A). The median length of individual SVs was 2.3Kb for INVs, 5.8Kb for ITXs, 1.3Kb for INSs, and 342bp for DELs (Fig. 4B). A high proportion of SVs were specific to an individual genome (Fig. 4C), consistent with findings from the 1KJPN[17]. The portion of individual-specific SVs was greatest for INSs (92.51%), followed by INVs (88.87%), ITXs (68.93%), and DELs (47.82%) (Table S8). A substantial proportion of SVs (98.5% INSs and 61% DELs) were novel and were not previously deposited in the Database of Genomic Variants (DGV). Overall, the non-redundant combined SVs ranged in size up to 10M and all classes were enriched in the 1-2Kb size range (Fig. 4D, Supplementary Fig. S6).

Finally, we analyzed the SVs to determine whether they were enriched for repetitive elements. Within the SVs, we cataloged repeat types and searched for Korean-specific enrichments compared to those present in other populations. Among the SVs, we found that 13% contained short interspersed elements (SINEs), 20% contained long interspersed elements (LINEs), 3.4% contained DNA transposons, and 8.6% contained long terminal repeats (LTRs). The majority of SINEs were observed in DELs of 200-300bp, which is consistent with *de novo* assembled SVs[16] and the predicted SVs[15]. These results suggest that SVs are enriched for SINEs in the 1-4Kb INVs, and LINEs in the 4-40Kb INVs (Supplementary Fig. S6A). Additionally, simple repeats were predominantly observed in INSs (Fig. 4D) and 3-5Kb ITXs (Supplementary Fig. S6B).

Copy number variations in KoVariome

The high coverage WGS data used to construct KoVariome provides sufficient data to

characterize CNVs in a single genome. The FREEC program[48] predicted an average of 199 deletions and 336 duplications per genome (Supplementary Table S9). After filtering out spurious CNVs (Supplementary Fig. S7), 161.74 (81.46%) deletions and 296.72 (88.29%) duplications remained from the original calls. In total, we predicted 2,038 non-redundant deletions and 1,564 non-redundant duplications, and the unified CNVs were approximately 5Kb-100Kb in length (Fig. 5A). When compared to the DGV, we identified 3.6K known CNVs, including 1,169 (57.36%) deletions and 846 (54.09%) duplications. Repeat composition analyses of CNV regions revealed that deletions smaller than 5K and duplications smaller than 10K contained a 20-fold more simple repeats compared to their overall frequencies in the human genome. In addition, SINEs were 2-fold more frequent in the > 600Kb deletions. These associations differ from the repeat distributions in SVs. By examining the genes in the unified CNVs, 869 (46.47%) deletions and 1,105 (70.65%) duplications were found to contain at least one gene. In addition, only two deletions and three duplications were conserved in the 50 Korean individuals (Table 5). Interestingly, a long 2M genomic block on chromosome 10, containing seven genes, was found to be duplicated an average of 4.22 times in the KPGP donors. Included among these genes is *G protein regulated inducer of neurite outgrowth 2* (*GPRIN2*), which is associated with brain development and neurite outgrowth [49]. Previous reports identified this duplication in Asian, European, and Yoruba populations (three-six copies), while no duplications were reported in the chimpanzee, orangutan, or gorilla [22]. We also identified 444 CNVs conserved in 1000GP (Supplementary Table S10), which are probably shared East Asian CNVs and are not specific to Koreans. Five deletions and nine duplications were found to be enriched in the Korean population using the following criteria; i) odds ratio > 10 comparing with CNV ratio in any continents, ii) p-values < 0.01, and iii) more than five individuals in KoVariome. Phenotypic features were examined by searching genes against the OMIM database, resulting in the identification of three deletions and three duplications containing genes associated with known phenotypes (Fig. 5B). A high copy number deletion of *UDP glucuronosyltransferase family 2 member B17* (*UGT2B17*),

which is associated with bone mineral density and osteoporosis[50], was observed by comparing our Korean individuals with EUR, AFR, and AMR populations. This finding is consistent with previous studies which reported that 66.7% of Korean males have a deletion of this gene, compared to only 9.3% of Swedish males[51]. We also observed frequent deletions of *acyl-CoA thioesterase 1* (*ACOT1*), which functions to maintain the cellular levels of acyl-CoA and free fatty acids[52]. We identified the duplication of *hydroxycarboxylic acid receptor 2* (*HCAR2*) in 12% of the Koreans, which is associated with lipid-lowering effects[53]. We excluded the gene duplications of *NBPF15* and *HERC2* because they were located at the CNV break points. These CNVs will be useful for detecting Korean-specific genetic associations with specific phenotypes in future studies, which is especially important since CNVs are analyzed less often than SNVs even though they likely contain important disease-relevant variations.

Conclusions

To discover disease-causing genetic variants, researchers rely on comprehensive, population-specific databases containing the benign genetic variation present within specific ethnic groups. The KoVariome database was created to fill this need for the Korean population, and includes 5.5 TB of WGS data from 50 healthy, unrelated Korean individuals with corresponding health metadata. Using this database, we characterized all four variation types and identified 12.7M SNVs, 1.7M indels, 4K SVs, and 3.6K CNVs, many of which were novel or selectively enriched in the Korean population. Despite their close geographic proximity, the Korean population was shown to be genetically distinct from the Chinese and Japanese populations, highlighting the need for a Korean-specific variome to accurately identify rare disease variants in this population. Accordingly, KoVariome was used to predict candidate loci, inheritance patterns, and genetic risk for several diseases, including cancer, fibrodysplasia ossificans progressive, Van der Woude syndrome, type II diabetes mellitus, and spherocytosis. As this database grows and the accuracy

of predicting disease associations increases, genetic tests will increasingly become increasingly routine components of precision healthcare. KoVariome will be an invaluable resource for biomedical researchers and health practitioners, and will directly benefit patients by ensuring they are presented with the most accurate genetic predictions of disease risks.

Methods

Sample collection and data distribution

Since 2010, the Korean variome data center (KOVAC) recruited volunteers for the Korea Personal Genome Project (KPGP). Consent was acquired from all participants in accordance with the Korean Life Ethics bill. In addition to providing a blood sample for WGS, each individual responded to a questionnaire regarding body characteristics, habits, response to 16 allergies, family histories, and physical condition related to 19 disease classes (Table S1). Genomic DNA was extracted using a QIAamp DNA Blood Mini Kit (Qiagen, CA, USA) and 69 WGS libraries were constructed using TruSeq DNA sample preparation kits (Illumina, CA, USA). Sequencing was performed using Illumina HiSeq sequencers following the manufacturer's instruction. WGS data from 50 healthy unrelated Korean individuals were analyzed to create the KoVariome database, which was released through the national FTP portal server of the KOBIC (ftp://ftp.kobic.re.kr/pub/KPGP/) and distributed through GRF (http://pgi.re.kr) and Variome.net. All data analyzed in this study were deposited in NCBI SRA (PRJNA284338) and accessions for each sample were listed in Supplementary Table S2.

Analysis of SNVs and indels

The WGS data were processed according to a protocol that was evaluated by the technical committee of the Korean Research Institute of Standards and Science (KRISS). Genomic

resources were downloaded from UCSC Genome bioinformatics (http://hgdownload.cse.ucsc.edu/goldenPath/hg19/bigZips/), including the reference human genome (GRCH37/hg19), reference genes, and repeat annotations. Raw DNA reads were cleaned by Sickle (https://github.com/najoshi/sickle) with a quality score > 20 and read length > 50 bp. Cleaned paired-end reads were mapped to the human reference genome using BWA [54]and indels were realigned and recalibrated after removing the PCR duplicates. Finally, we identified SNVs and indels for each individual using the GATK UnifiedGenotyper (ver. GATK-Lite-2.3-9)[55]. To improve the quality of identified SNVs, we applied SNV meeting criteria of: i) read depth (DP) is 20× or higher, ii) mapping rate is 90% or higher. Low-quality indels were removed from future analyses using the following criteria: i) quality score <27 and DP <6, ii) heterozygous indels with mapped allelic valance less than 0.3.

Functional effect of the variants

To analyze the functional effects of variations, we implemented SnpEff-3.3[56]. The deleterious effects of the non-synonymous SNVs were obtained by searching dbNSFP (ver. 2.9.1), a portal database providing deleterious non-synonymous SNVs[57]. We then predicted the effects of each variant on protein function using SIFT, Polyphen2, PROVEAN, MetaSVM, and MetaLE, and further annotated variants using the Interpro_domain and COSMIC (Catalogue of Somatic Mutations in Cancer, ver. 71) databases. Previously reported SNVs and indels were identified using the dbSNP database (ver. 146). All variants shorter than 50 bp were then stored in this database[58]. The databases ClinVar (ver. 20161101)[59] and OMIM (generated 2016-11-22)[60] were searched to identify known pathogenic variants.

Genetic distance calculation

The genetic distance (pi) between two samples was calculated using the following formula:

$$pi = D / N,$$

where D is the nucleotide difference between two samples and N is the number of compared positions. The sum of the nucleotide difference was calculated between two samples for each genomic position, which ranged from 0-1. A homozygous genotype composed of a reference allele was adopted as the genotype for uncalled sites.

Multidimensional Scaling (MDS) analysis

Genotype data for 84 Chinese and 86 Japanese individuals were obtained from Phase 3 of the HapMap project[3]. A total of 1,387,956 SNV loci were merged with KoVariome. The PLINK program was used to remove the genomic loci with MAF < 0.05, call rates < 0.05, and SNPs in linkage disequilibrium blocks[61]. In total, 117,521 SNPs remained after filtering and were used in the MDS analysis. Five dimensional components were calculated in R with the distance matrix method “canberra” and MDS plots were generated using the MASS package [62].

Accuracy of the SNVs

To measure the accuracy of SNV predictions, 35 individuals were genotyped with the Axiom™ Genome-Wide East Asian (ASI) 1 Array (Affymetrix, Inc.). The accuracy and recalls were analyzed using a contingency table constructed with the presence and absence of the alternative alleles analyzed from our pipeline and the genotyping results from the Axiom™ Genome-ASI 1 Array. The precision of calls was calculated by analyzing the concordance and denoted as true positive predictions (TP) from all predicted SNVs. The recalls were defined as TPs divided by the number of genotypes represented on the Axiom™ Genome-ASI 1 Array. The genotype accuracies were measured by analyzing the concordance of the genotypes between the GATK prediction and the results from the Axiom™ Genome-ASI 1 Array. The accuracy of the indel predictions were

calculated by comparing genotypes between GATK predictions and the Axiom™ Genome-ASI 1 Array.

Structural variants

We applied two programs, BD[63] and pindel[64], to predict genome-wide SVs based on the discordant mate-pair and split-read information, respectively. From the bam files for each individual, insertions and deletions of a length between 100 and 1 Kb were predicted by pindel (ver. 0.2.4t) and those longer than 1Kb were predicted by BD (ver. 1.4.5)[65]. We next constructed unassembled genomic blocks ('N') from the hg19 reference genome and examined the SVs that overlapped with these unassembled genomic regions. From this analysis, we discovered a high portion of spurious SVs in these regions (Supplementary Fig. S5), with the majority of them >100M in size. The following criteria were used to filter out spurious SVs; i) reciprocally > 10% overlaps between SVs and un-assembled genomic blocks, ii) 'N's more than 50% coverage of SVs, and iii) more than 2 un-assembled genomic blocks in the predicted SVs. After filtering, we clustered SVs that reciprocally overlapped > 70% in any individual. Unified SVs were defined by the average start and end positions in each SV cluster. The novelty of each SV was defined by comparing unified SVs with those in the DGV[66], with 70% reciprocal overlaps.

Copy number variations

CNVs were predicted with FREEC (ver. 10.6) using window size =100, step size =50, and breakpoint =0.6 [48]. The spurious CNVs were enriched in >1M in length (Figure S7), which were filtered using the same criteria described in the SV methods above. Unified CNVs were constructed by merging individual's CNVs that reciprocally overlapped by >=70%. The start and end positions of the unified CNVs were defined as average position of the original calls. Known CNVs were defined by comparing with CNVs in the DGV database[66].

Additional information

Data resource access: http://variome.net, http://kpgp.kr, http://koreangenome.org

SNP data have been deposited in dbSNP under batch_id 1062763.

Competing Interests: The authors declare that they have no competing interests.

References

1 International Human Genome Sequencing, C. Finishing the euchromatic sequence of the human genome. *Nature* **431**, 931-945, doi:10.1038/nature03001 (2004).

2 Collins, F. S. & McKusick, V. A. Implications of the Human Genome Project for medical science. *Jama* **285**, 540-544 (2001).

3 International HapMap, C. The International HapMap Project. *Nature* **426**, 789-796, doi:10.1038/nature02168 (2003).

4 Redon, R. *et al.* Global variation in copy number in the human genome. *Nature* **444**, 444-454, doi:10.1038/nature05329 (2006).

5 Mills, R. E. *et al.* An initial map of insertion and deletion (INDEL) variation in the human genome. *Genome research* **16**, 1182-1190, doi:10.1101/gr.4565806 (2006).

6 Welter, D. *et al.* The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic acids research* **42**, D1001-1006, doi:10.1093/nar/gkt1229 (2014).

7 Reich, D. E. & Lander, E. S. On the allelic spectrum of human disease. *Trends in genetics : TIG* **17**, 502-510 (2001).

8 Kraft, P. & Hunter, D. J. Genetic risk prediction--are we there yet? *The New England journal of medicine* **360**, 1701-1703, doi:10.1056/NEJMp0810107 (2009).

9 Bamshad, M. J. *et al.* Exome sequencing as a tool for Mendelian disease gene discovery. *Nature reviews. Genetics* **12**, 745-755, doi:10.1038/nrg3031 (2011).

10 MacArthur, D. G. *et al.* Guidelines for investigating causality of sequence variants in human disease. *Nature* **508**, 469-476, doi:10.1038/nature13127 (2014).

11 Genomes Project, C. *et al.* A global reference for human genetic variation. *Nature* **526**, 68-74, doi:10.1038/nature15393 (2015).

12 Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285-291, doi:10.1038/nature19057 (2016).

13 Telenti, A. *et al.* Deep sequencing of 10,000 human genomes. *Proceedings of the National Academy of Sciences of the United States of America* **113**, 11901-11906, doi:10.1073/pnas.1613365113 (2016).

14 Wong, L. P. *et al.* Deep whole-genome sequencing of 100 southeast Asian Malays.

American journal of human genetics **92**, 52-66, doi:10.1016/j.ajhg.2012.12.005 (2013).

15 Genome of the Netherlands, C. Whole-genome sequence variation, population structure and demographic history of the Dutch population. *Nature genetics* **46**, 818-825, doi:10.1038/ng.3021 (2014).

16 Besenbacher, S. *et al.* Novel variation and de novo mutation rates in population-wide de novo assembled Danish trios. *Nature communications* **6**, 5969, doi:10.1038/ncomms6969 (2015).

17 Nagasaki, M. *et al.* Rare variant discovery by deep whole-genome sequencing of 1,070 Japanese individuals. *Nature communications* **6**, 8018, doi:10.1038/ncomms9018 (2015).

18 Chheda, H. *et al.* Whole-genome view of the consequences of a population bottleneck using 2926 genome sequences from Finland and United Kingdom. *European journal of human genetics : EJHG*, doi:10.1038/ejhg.2016.205 (2017).

19 Sudmant, P. H. *et al.* An integrated map of structural variation in 2,504 human genomes. *Nature* **526**, 75-81, doi:10.1038/nature15394 (2015).

20 Feuk, L., Carson, A. R. & Scherer, S. W. Structural variation in the human genome. *Nature reviews. Genetics* **7**, 85-97, doi:10.1038/nrg1767 (2006).

21 Huddleston, J. *et al.* Discovery and genotyping of structural variation from long-read haploid genome sequence data. *Genome research* **27**, 677-685, doi:10.1101/gr.214007.116 (2017).

22 Sudmant, P. H. *et al.* Diversity of human copy number variation and multicopy genes. *Science* **330**, 641-646, doi:10.1126/science.1197005 (2010).

23 Perry, G. H. *et al.* Diet and the evolution of human amylase gene copy number variation. *Nature genetics* **39**, 1256-1260, doi:10.1038/ng2123 (2007).

24 Lakich, D., Kazazian, H. H., Jr., Antonarakis, S. E. & Gitschier, J. Inversions disrupting the factor VIII gene are a common cause of severe haemophilia A. *Nature genetics* **5**, 236-241, doi:10.1038/ng1193-236 (1993).

25 Bondeson, M. L. *et al.* Inversion of the IDS gene resulting from recombination with IDS-related sequences is a common cause of the Hunter syndrome. *Human molecular genetics* **4**, 615-621 (1995).

26 Pinto, D. *et al.* Functional impact of global rare copy number variation in autism spectrum disorders. *Nature* **466**, 368-372, doi:10.1038/nature09146 (2010).

27 Stefansson, H. *et al.* Large recurrent microdeletions associated with schizophrenia. *Nature* **455**, 232-236, doi:10.1038/nature07229 (2008).

28 McCarroll, S. A. *et al.* Deletion polymorphism upstream of IRGM associated with altered IRGM expression and Crohn's disease. *Nature genetics* **40**, 1107-1112, doi:10.1038/ng.215 (2008).

29 Consortium, H. P.-A. S. *et al.* Mapping human genetic diversity in Asia. *Science* **326**,

1541-1545, doi:10.1126/science.1177074 (2009).

30 Ahn, S. M. *et al.* The first Korean genome sequence and analysis: full genome sequencing for a socio-ethnic group. *Genome research* **19**, 1622-1629, doi:10.1101/gr.092197.109 (2009).

31 Kim, J. I. *et al.* A highly annotated whole-genome sequence of a Korean individual. *Nature* **460**, 1011-1015, doi:10.1038/nature08211 (2009).

32 Ju, Y. S. *et al.* Extensive genomic and transcriptional diversity identified through massively parallel DNA and RNA sequencing of eighteen Korean individuals. *Nature genetics* **43**, 745-752, doi:10.1038/ng.872 (2011).

33 Zhang, W. *et al.* Whole genome sequencing of 35 individuals provides insights into the genetic architecture of Korean population. *BMC bioinformatics* **15 Suppl 11**, S6, doi:10.1186/1471-2105-15-S11-S6 (2014).

34 Hong, D. *et al.* TIARA genome database: update 2013. *Database : the journal of biological databases and curation* **2013**, bat003, doi:10.1093/database/bat003 (2013).

35 Lee, S. *et al.* Korean Variant Archive (KOVA): a reference database of genetic variations in the Korean population. *Scientific reports* **7**, 4287, doi:10.1038/s41598-017-04642-4 (2017).

36 Kwak, S. H. *et al.* Findings of a 1303 Korean whole-exome sequencing study. *Experimental & molecular medicine* **49**, e356, doi:10.1038/emm.2017.142 (2017).

37 Cho, Y. S. *et al.* An ethnically relevant consensus Korean reference genome is a step towards personal reference genomes. *Nature communications* **7**, 13637, doi:10.1038/ncomms13637 (2016).

38 Wang, J., Raskin, L., Samuels, D. C., Shyr, Y. & Guo, Y. Genome measures used for quality control are dependent on gene function and ancestry. *Bioinformatics* **31**, 318-323, doi:10.1093/bioinformatics/btu668 (2015).

39 Boomsma, D. I. *et al.* The Genome of the Netherlands: design, and project goals. *European journal of human genetics : EJHG* **22**, 221-227, doi:10.1038/ejhg.2013.118 (2014).

40 Scally, A. & Durbin, R. Revising the human mutation rate: implications for understanding human evolution. *Nature reviews. Genetics* **13**, 745-753, doi:10.1038/nrg3295 (2012).

41 Loveday, C. *et al.* Germline mutations in RAD51D confer susceptibility to ovarian cancer. *Nature genetics* **43**, 879-882, doi:10.1038/ng.893 (2011).

42 Richards, C. S. *et al.* ACMG recommendations for standards for interpretation and reporting of sequence variations: Revisions 2007. *Genetics in medicine : official journal of the American College of Medical Genetics* **10**, 294-300, doi:10.1097/GIM.0b013e31816b5cae (2008).

43 Shore, E. M. *et al.* A recurrent mutation in the BMP type I receptor ACVR1 causes

inherited and sporadic fibrodysplasia ossificans progressiva. *Nature genetics* **38**, 525-527, doi:10.1038/ng1783 (2006).

44 Wang, X. *et al.* Novel mutations in the IRF6 gene for Van der Woude syndrome. *Human genetics* **113**, 382-386, doi:10.1007/s00439-003-0989-2 (2003).

45 Malik, S. *et al.* Epidemiology of Van der Woude syndrome from mutational analyses in affected patients from Pakistan. *Clinical genetics* **78**, 247-256, doi:10.1111/j.1399-0004.2010.01375.x (2010).

46 Charlesworth, B., Morgan, M. T. & Charlesworth, D. The effect of deleterious mutations on neutral molecular variation. *Genetics* **134**, 1289-1303 (1993).

47 Alshatwi, A. A., Hasan, T. N., Syed, N. A., Shafi, G. & Grace, B. L. Identification of functional SNPs in BARD1 gene and in silico analysis of damaging SNPs: based on data procured from dbSNP database. *PloS one* **7**, e43939, doi:10.1371/journal.pone.0043939 (2012).

48 Boeva, V. *et al.* Control-FREEC: a tool for assessing copy number and allelic content using next-generation sequencing data. *Bioinformatics* **28**, 423-425, doi:10.1093/bioinformatics/btr670 (2012).

49 Chen, L. T., Gilman, A. G. & Kozasa, T. A candidate target for G protein action in brain. *The Journal of biological chemistry* **274**, 26931-26938 (1999).

50 Yang, T. L. *et al.* Genome-wide copy-number-variation study identified a susceptibility gene, UGT2B17, for osteoporosis. *American journal of human genetics* **83**, 663-674, doi:10.1016/j.ajhg.2008.10.006 (2008).

51 Jakobsson, J. *et al.* Large differences in testosterone excretion in Korean and Swedish men are strongly associated with a UDP-glucuronosyl transferase 2B17 polymorphism. *The Journal of clinical endocrinology and metabolism* **91**, 687-693, doi:10.1210/jc.2005-1643 (2006).

52 Hunt, M. C., Rautanen, A., Westin, M. A., Svensson, L. T. & Alexson, S. E. Analysis of the mouse and human acyl-CoA thioesterase (ACOT) gene clusters shows that convergent, functional evolution results in a reduced number of human peroxisomal ACOTs. *FASEB journal : official publication of the Federation of American Societies for Experimental Biology* **20**, 1855-1864, doi:10.1096/fj.06-6042com (2006).

53 Tunaru, S. *et al.* PUMA-G and HM74 are receptors for nicotinic acid and mediate its anti-lipolytic effect. *Nature medicine* **9**, 352-355, doi:10.1038/nm824 (2003).

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

55 McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome research* **20**, 1297-1303, doi:10.1101/gr.107524.110 (2010).

56 Cingolani, P. *et al.* A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly* **6**, 80-92, doi:10.4161/fly.19695 (2012).

57 Dong, C. *et al.* Comparison and integration of deleteriousness prediction methods for nonsynonymous SNVs in whole exome sequencing studies. *Human molecular genetics* **24**, 2125-2137, doi:10.1093/hmg/ddu733 (2015).

58 Sherry, S. T. *et al.* dbSNP: the NCBI database of genetic variation. *Nucleic acids research* **29**, 308-311 (2001).

59 Landrum, M. J. *et al.* ClinVar: public archive of interpretations of clinically relevant variants. *Nucleic acids research* **44**, D862-868, doi:10.1093/nar/gkv1222 (2016).

60 Baxevanis, A. D. Searching Online Mendelian Inheritance in Man (OMIM) for information for genetic loci involved in human disease. *Current protocols in bioinformatics* **Chapter 1**, Unit 1 2, doi:10.1002/0471250953.bi0102s00 (2002).

61 Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *American journal of human genetics* **81**, 559-575, doi:10.1086/519795 (2007).

62 Ihaka, R. & Gentleman, R. R: a language for data analysis and graphics. *Comput Graph Stat* **5**, 299-134 (1996).

63 Chen, K. *et al.* BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. *Nature methods* **6**, 677-681, doi:10.1038/nmeth.1363 (2009).

64 Ye, K., Schulz, M. H., Long, Q., Apweiler, R. & Ning, Z. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics* **25**, 2865-2871, doi:10.1093/bioinformatics/btp394 (2009).

65 Mimori, T. *et al.* iSVP: an integrated structural variant calling pipeline from high-throughput sequencing data. *BMC systems biology* **7 Suppl 6**, S8, doi:10.1186/1752-0509-7-S6-S8 (2013).

66 MacDonald, J. R., Ziman, R., Yuen, R. K., Feuk, L. & Scherer, S. W. The Database of Genomic Variants: a curated collection of structural variation in the human genome. *Nucleic acids research* **42**, D986-992, doi:10.1093/nar/gkt958 (2014).

Acknowledgements

The authors thank many people not listed as authors who provided analyses, data, feedback, samples, and encouragement. Especially, thanks for Sunghoon Lee, Taehyung Kim, Sanghoon Song, Sangsoo Kim, and George Church. This work was supported by the 2014 Research Fund (1.140113.01 and 1.150014.01) and the 2016 Research Fund (1.160052.01) of Ulsan National Institute of Science & Technology (UNIST) and supported by the Ministry of Trade, Industry & Energy (MOTIE, Korea) under Industrial Technology Innovation Programs 'National Center for standard Reference Data', No.10075262. This work was also supported by the Research Fund (14-BR-SS-03) of Civil-Military Technology Cooperation Program and the Next-Generation Information Computing Development Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT & Future Planning (NRF-2016M3C4A7952635).

Author Contributions

J.B. and B.K. planned and designed this study. J.K., S.J., J.J., H.-M.K., H.K., and O.C. processed the sequencing data. Y.K, J.B., and J.J. contributed to recruitment of individual for whole genome sequencing. J.K., S.L., J.B., Y.C., and J.W., contributed to the interpretation of the data. J.W., C.K., H.L., B.K., K.H., I.K., J.E., J.B., K.C., and J.E. wrote and reviewed draft. All authors commented on the manuscript and approved the final version to be submitted.

Supplementary information

1. SupplementaryTables
2. SupplementaryFigures

Comments

By submitting a comment you agree to abide by our Terms and Community Guidelines. If you find something abusive or that does not comply with our terms or guidelines please flag it as inappropriate.

Figure legends

Figure 1. Status of KPGP variomes analyzed using 50 unrelated Korean individuals

A. Accumulation of novel SNV alleles. The number of novel SNV alleles were defined as newly identified nucleotides compared with previously constructed SNVs in KoVariome. B. Genetic distance according to the familial relationships. Abbreviations: Monozygotic Twin (MT), Parent and child (PC), Brothers (Br), Grandparents vs. grand children (GPC), Uncle vs. Nephew (UN), and Cousins (Co)

Figure 2. Genetic features of KoVariome

A. Two dimensional classification of KoVariome. SNVs and indels observed in 1000GP data were classified based on the minor allele frequencies (MAF); '1000GP Common': MAF >=5% in all five continents, '1000GP Low frequency': MAF >= 0.1% in any continent, and '1000GP Rare'; MAF < 0.1% in all five continents. The five continental populations included African (AFR), European (EUR), Native American (AMR), South Asian (SAS), and East Asian (EAS). The second group was classified by the number of variants in KoVariome; 'Frequent in KoVariome' (>= 3) and 'Rare in KoVariome' (< 3). B. The Venn diagrams represent the number of variants enriched in specific continents for both SNVs (left) and indels (right). The enrichment was analyzed by Fisher's exact test based on odds ratio > 3 and p-value < 0.01. The total numbers of enriched variants in the Korean (KOR) population are denoted in the white space of the Venn diagram. The numbers next to the continental population abbreviations represent the total number of enriched variants in that 1000GP continental group. The numbers within each ellipse denote the number of variants enriched both in KOR and a specific continent (left) and the number of variants enriched exclusively in the represented continent (right). C. Rare variant ratios (RVRs) observed in each genomic region. RVRs were calculated by dividing the number rare variants by the number of frequent variants in KoVariome.

Figure 3. Individual variants describing functional effects

A. Classification of individual variants based on frequency in 1000GP and KoVariome. Gray represents the portion of individual variants classified in the '1000GP common' and '1000GP Low frequency'. Blue represents the portion of the individual variants classified in the 'Frequent in KoVariome'. Red represents rare variants in both 1000GP and KoVariome 'Rare in Both'. B. Individual variants in the 'Rare in Both' were classified by gene coordinates. To more clearly represent the portion functionally important rare variants, 98% of the rare variants in the non-coding regions were not represented. C. Number of pathogenic variants for each individual. Red and blue bars represent the number of pathogenic variants previously reported in dbSNP and novel, respectively.

Figure 4. Properties of structural variants discovered in KoVariome

A. The boxplot represents the number of variants per Korean individual by variant type (n=50). The lower and upper hinges of the boxes correspond to the 25^{th} and 75^{th} percentiles and the whiskers represent the 1.5x inter-quartile range (IQR) extending from the hinges. Abbreviations of the variants: inversions (INV), intra-chromosomal translocation (ITX), insertions (INS), and deletions (DEL). B. Length of the variants present in the individual genome. See variant types and boxplot definition in A. C. Frequency of variants in KoVariome. D. The upper graph represents the number of SVs identified at specific length ranges. The KoVariome specific variants were defined by comparing SVs in the Database of Genomic Variants (DGV) with 70% reciprocal overlap. The lower graph represents the portion of repeats distributed in the variants. Repeat classes were defined by the repeat annotations provided in the UCSC Genome bioinformatics. Simple repeats contained both microsatellites and low complexity (e.g., AT-rich). Abbreviations of repeats: short interspersed element (SINE), long interspersed element (LINE), and long terminal repeat (LTR).

Figure 5. Properties of copy number variations in KoVariome

A. The number CNVs in the Korean population and the portion of the repeats in a specific length range. The conserved CNVs were defined by searching the Database of Genomic Variants (DGV) with 70% reciprocal overlaps. See the abbreviations of repeats in Fig. 4B. Korean enriched CNVs were identified by searching the CNVs reported in the 1000GP. No. represents the number of CNVs predicted in KoVariome. The heatmap represents the odds ratio of the CNVs compared to the CNV ratio in a specific 1000GP continental group. Associated genes were identified by searching the OMIM database. Abbreviations of continent group: European (EUR), African (AFR), Native American (AMR), South Asian (SAS), and East Asian (EAS).

Tables

Table 1. Statistics of KoVariome

Sample information for KoVariome	
No. of samples (Male/Female)	50 (31/19)
Total NGS yield	5.5 tera bases
Average sequenced depth	31x
Average mapped read rates	95%
SNVs	
Total No. of SNVs	12,735,004
No. of known variants in 1000GP [a]	8,967,464
No. of known variants in dbSNP [b]	10,286,599
Average No. of SNV per sample	3,813,311
Average No. of Coding SNVs [c]	20,097
Average No. of non-synonymous SNVs [c]	10,394
Average No. of SNVs with high effects [c]	287
Indels	
Total No. of indels	1,743,117
No. of known variants in 1000GP [a]	848, 471
No. of known variants in dbSNP [b]	1,307,000
Average No. of indel per sample	503,553
Average No. of Coding indels [c]	258
Average No. of LOF indels [c]	157

Variants deposited in [a] 1000GP and [b] the dbSNP (ver. 146). [c] predicted with SNPEff

Table 2. ClinVar annotation of the KoVariome frequent SNVs

Chr.	Position	Ref.	Alt.	rs No [a]	Gene	Codon Changes	Disease	Inheritance Type [b]	No. [c]	MAF [d]
17	33,445,518	A	C	rs200564819 [*]	*RAD51D*	Splice-site	Familial breast-ovarian cancer 4	n.a	5	0.05
1	161,599,571	T	C	rs2290834	*FCGR3B*	I106V	Neutrophil-specific antigens na1/na2	UNKNOWN	3	0.15
8	100,844,596	G	T	rs386834119	*VPS13B*	Splice-site	Cohen syndrome	AR	13	0.26
2	158,630,626	C	G	rs121912678	*ACVR1*	R206P	Fibrodysplasia ossificans progressive	AD	14	0.14
1	209,961,970	C	G	rs200166664	*IRF6*	R400P	Van der Woude syndrome	AD	14	0.14
11	18,290,859	C	T	rs1136743	*SAA1*	A70V	Systemic amyloidosis	AR	22	0.66

AR: autosomal recessive; AD: autosomal dominant; Chr.: chromosome; Ref. reference allele; Alt. alternative allele

[a] KoVariome frequent SNVs with the Reference SNP cluster IDs (rs number) in ClinVar. We were only included pathogenic or likely pathogenic ([*]) SNVs.

[b] Inheritance types were searched against OMIM database with rs numbers and phenotypes represented in ClinVar database. 'n.a.' represents there are no data in the OMIM database. 'UNKNOWN' represents inheritance type for corresponding phenotype was not reported in OMIM database.

[c] No. of alternative allele in Korean population, [d] minor allele frequencies (MAF) in KoVariome.

Table 3. Statistics of individual SNVs

Statistics of individual variants	**No. of SNVs**	**(%)**
1000GP common and 1000GP low frequency SNPs	**3.4 M**	**(88.70)**
Frequent SNVs in KoVariome	**0.4M**	**(9.39)**
1000GP rare and KoVariome rare SNVs	**47,957**	**(1.26)**
Statistics of individual rare SNVs		
Protein-Coding	**326**	**(40.72)**
Synonymous SNVs	**107**	**(13.37)**
Non-synonymous SNVs	**219**	**(27.36)**
Splice-site SNVs	**7**	**(0.87)**
RNA-Coding	**80**	**(9.93)**
Other statistics		
Median No. of pathogenic rare SNVs [a]	**137**	**(65.06)**

[a] Pathogenicity of the rare SNVs were predicted by at least one program among SIFT, Polyphen2, PROVEAN, MetaSVM, and MetaLE.

Table 4. Known pathogenic rare variants associated with disease

Individual ID	rs No.	Genotype	Codon change	Inheritance type [a]	gene	ClinVarTraits
KPGP-00001	rs563607795	A/G	L385P	n.a.	*SLC19A3*	Thiamine metabolism dysfunction syndrome
KPGP-00001	rs199769221*	G/C	R116P	AD	*PRSS1*	Hereditary pancreatitis
KPGP-00032	rs387907164	T/C	C32R	AR	*KIAA1530*	UV-sensitive syndrome 3
KPGP-00033	rs119490107	C/A	D234Y	UNKNOWN	*RAD54B*	Carcinoma of colon
KPGP-00039	rs199476197	A/C	H331P	AR	*CYP4V2*	Bietti crystalline corneoretinal dystrophy
KPGP-00088	rs28940280	G/A	D279N	AR	*CLN5*	Ceroid lipofuscinosis neuronal 5
KPGP-00122	rs587782989	C/T	R464H	AD	*CCDC88C*	Spinocerebellar ataxia 40
KPGP-00124	rs142808899	C/T	G303R	AR	*DHCR7*	Smith-Lemli-Opitz syndrome
KPGP-00127	rs111033744	A/G	Y100C	AR	*GALT*	Galoctosemia
KPGP-00127	rs137852972	T/C	N88S	AD	*BSCL2*	Silver spastic paraplegia syndrome
KPGP-00129	rs137853022	C/T	R696Q	AR	*IKBKAP*	Familial dysautonomia
KPGP-00129	rs386833823*	G/A	S238F	AR	*SLC7A7*	Lysinuric protein intolerance
KPGP-00131	rs200088377	G/A	P191L	n.a.	*IL17RD*	Delayed puberty
KPGP-00136	rs121908099	G/A	R405Q	AR	*CYP27A1*	Cholestanol storage disease
KPGP-00136	rs750218942	C/G	Splice-site	AR	*XPA*	Xeroderma pigmentosum
KPGP-00136	rs727502791	G/A	R158*	AD	*MFAP5*	Aortic aneurysm (familial thoracic 9)
KPGP-00136	rs545215807	G/A	G109S	AR	*ACADVL*	VLCAD deficiency
KPGP-00139	rs387907033	G/C	G401A	AR	*SYT14*	Spinocerebellar ataxia
KPGP-00139	rs748486078	G/A	S95L	UNKNOWN	*IL17F*	Candidiasis
KPGP-00144	rs119480073	C/T	R801	AR	*LPIN1*	Myoglobinuria
KPGP-00144	rs104895438	G/A	A612T	AD	*NOD2*	Sarcoidosis

KPGP-00205	rs121913050	G/A	R153H	UNKNOWN	*ERCC4*	XFE progeroid syndrome
KPGP-00220	rs121918673	G/C	S439R	AD	*HNF1B*	Diabetes mellitus type 2
KPGP-00266	rs104894085	G/A	Q258*	AR	*STAR*	Cholesterol monooxygenase deficiency
KPGP-00227	rs121909569	A/G	S148P	AD, AR	*SERPINC1*	Antithrombin III deficiency
KPGP-00228	rs121434426	G/A	Q356*	UNKNOWN	*FANCG*	Fanconi anemia
KPGP-00232	rs121909385	T/C	L623P	AR	*SLC12A3*	Familial hypokalemia hypomagnesemia
KPGP-00233	rs672601312	G/T	E127*	AR	*ISG15*	Immunodeficiency 38 with basal ganglia calcification
KPGP-00233	rs749462358	C/T	E924K	n.a.	*ASPM*	Not provided
KPGP-00245	rs137854500	C/T	D1289N	AR	*ABCA1*	Tangier disease
KPGP-00254	rs201968272	G/A	R237Q	AR	*DDX11*	Warsaw breakage syndrome
KPGP-00325	rs121912749	C/T	G130R	AD	*SLC4A1*	Spherocytosis type 4

Abbreviations: Chr. chromosome; Ref. reference allele; Alt. alternative allele; AD: autosomal dominant; AR: autosomal recessive

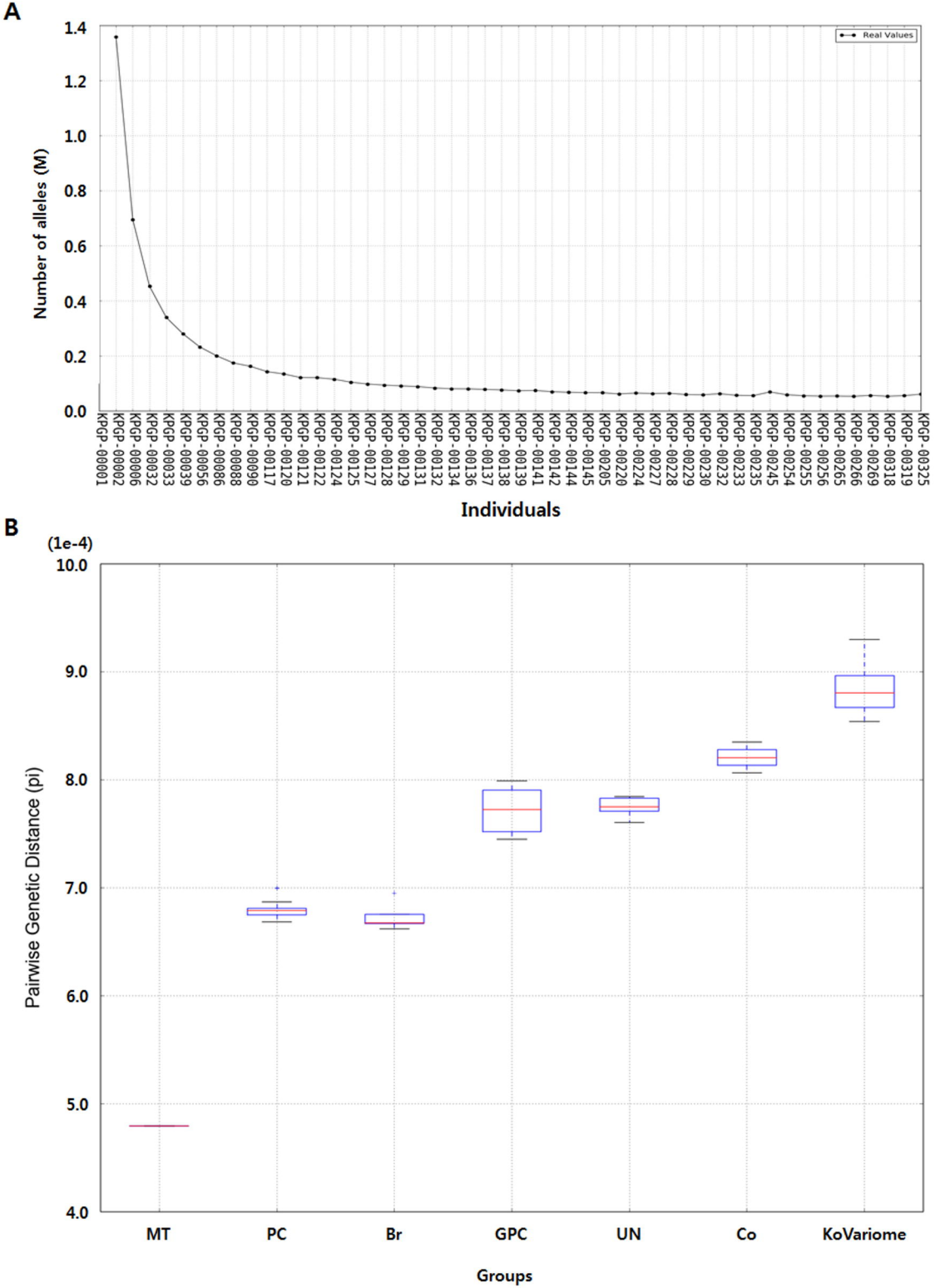
[*] The clinical significance of SNV locus was defined as likely pathogenic in the ClinVar database.

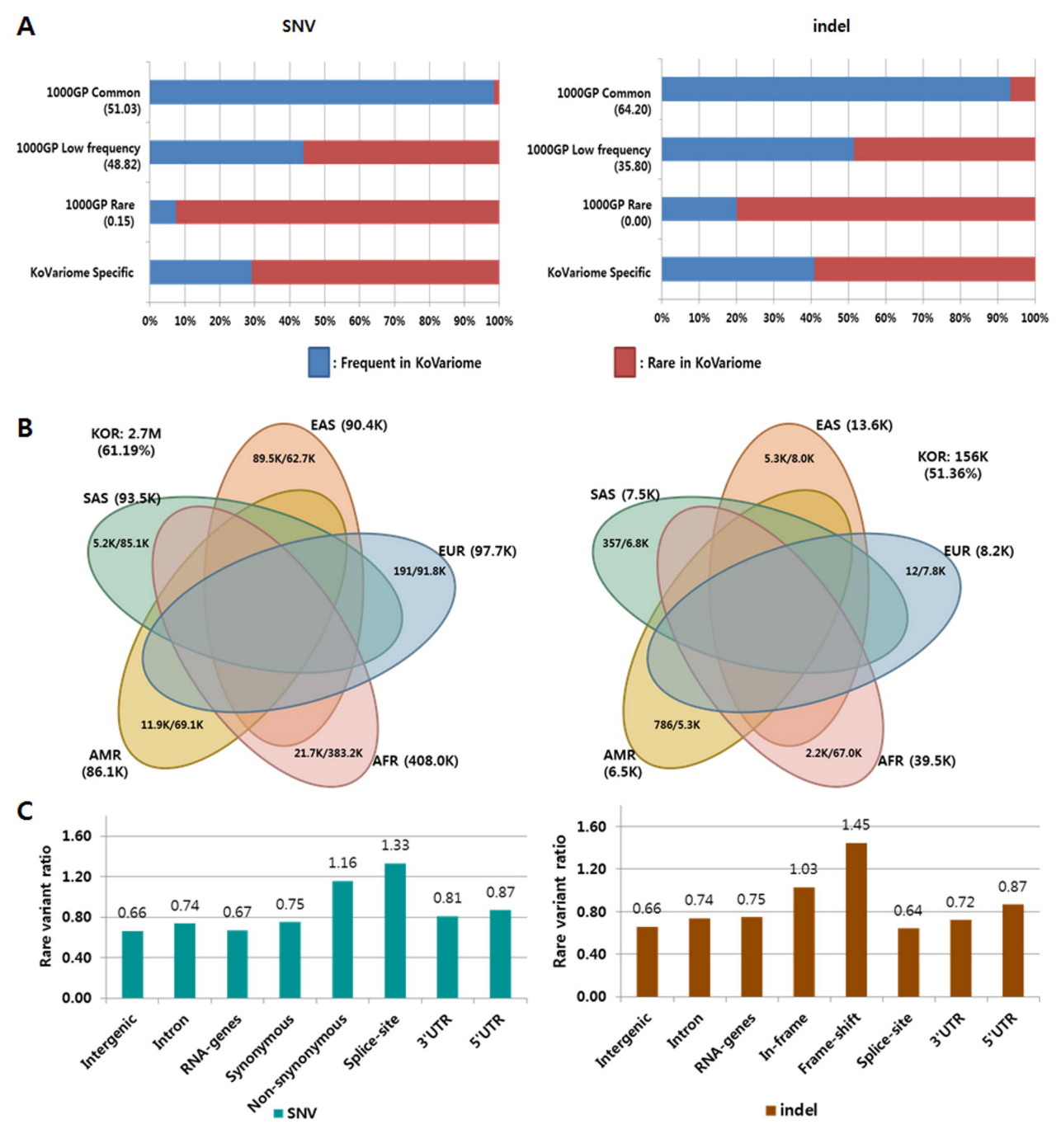
[a] Inheritance type were searched against OMIM database with rs numbers and phenotypes in the ClinVar database. 'n.a.' represents there are no data in the OMIM database. 'UNKNOWN' represents inheritance type for corresponding phenotype was not reported in OMIM database.

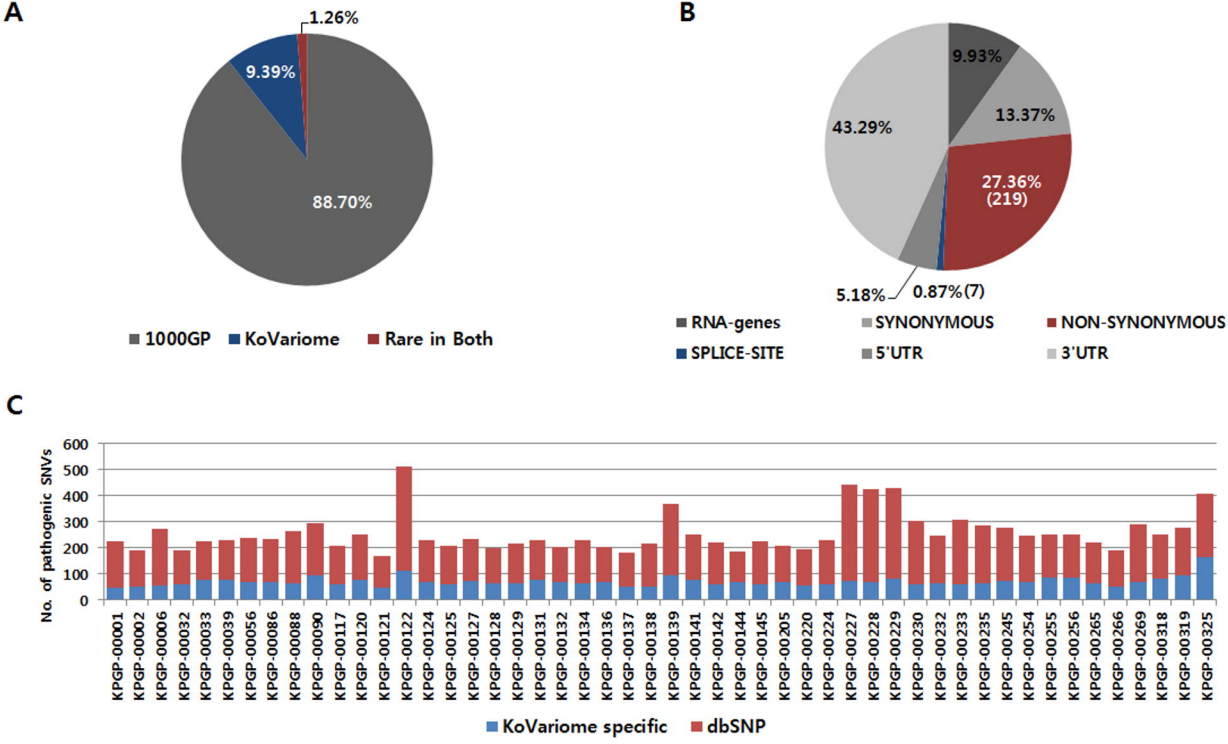
Table 5. Copy number variations conserved in 50 Korean individuals

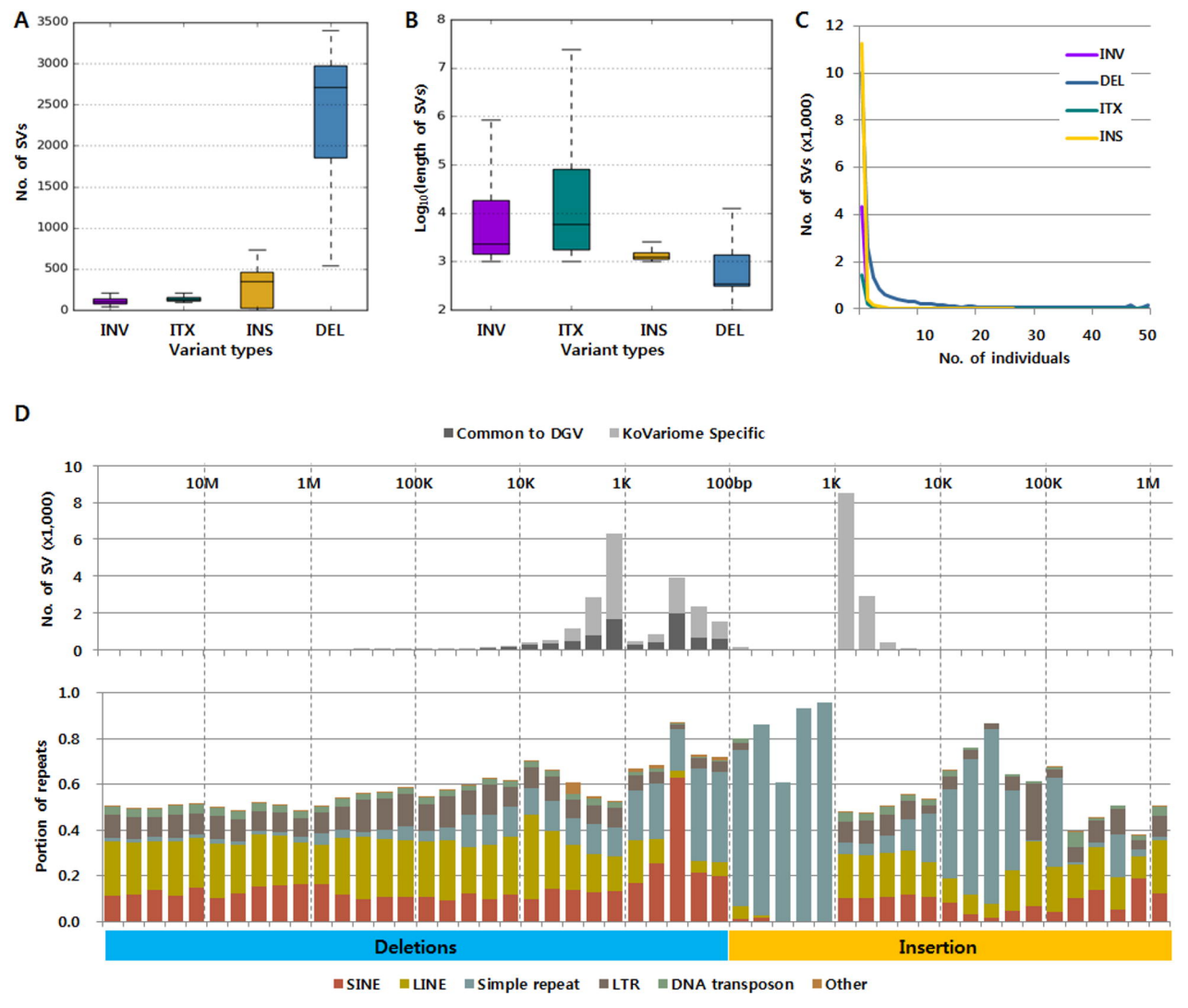
Chr.	Start	End	CNV Types	Average copy number	Genes [a]
chr2	132,964,050	133,121,849	Dup.	4.02	*MIR663B, FAM201B, ZNF806, ANKRD30BL*
chr10	46,222,900	46,946,499	Del.	1.0	*PTPN20, FAM35BP, AGAP4, FRMPD2B, FAM21C, BMS1P5*
chr10	46,946,200	47,150,299	Dup.	4.22	*NPY4R, GPRIN2, CH17-360D5.1, LINC00842, LOC102724593, HNRNPA1P33, SYT15*
chr10	47,147,400	47,384,499	Del.	1.0	*ANXA8, FAM35DP, LINC00842, FAM25C, AGAP9, FAM25G, BMS1P6*
chr15	21,885,000	21,944,149	Dup.	6.4	*LOC646214*

[a] Genes in the identified CNV region. Chr. Chromosome; Dup. duplication; Del. deletions.

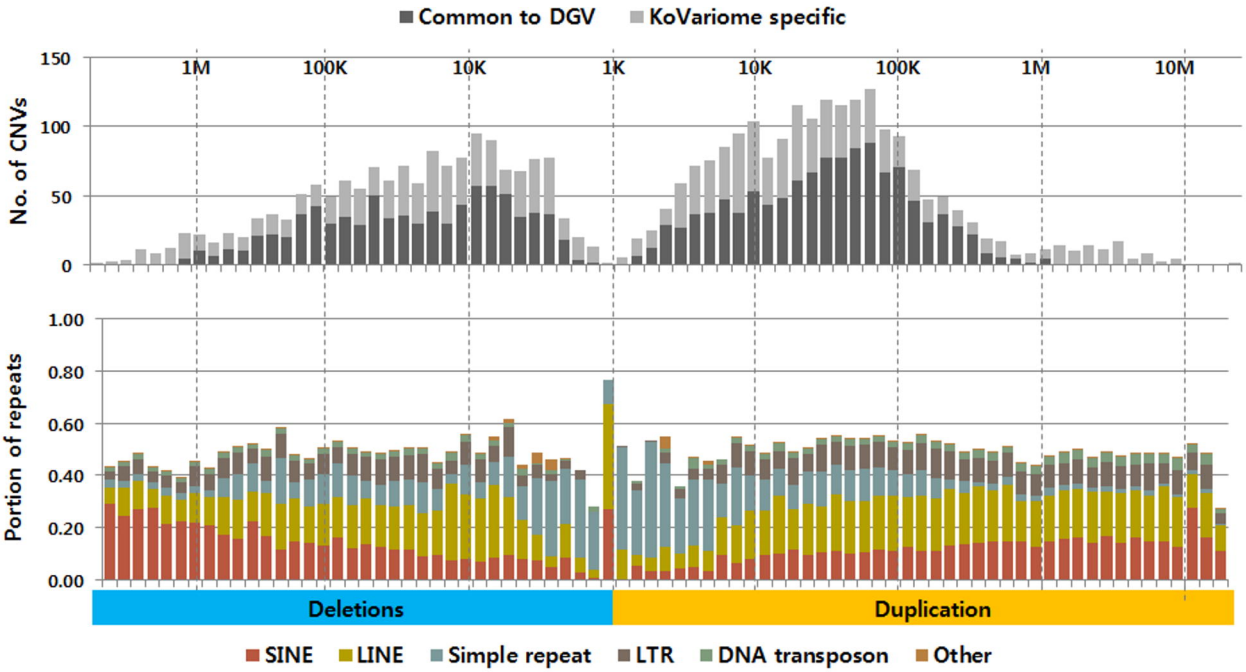

A
Real Values
Number of alleles (M)
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
KPGP-00001
KPGP-00002
KPGP-00006
KPGP-00032
KPGP-00033
KPGP-00039
KPGP-00056
KPGP-00086
KPGP-00088
KPGP-00090
KPGP-00117
KPGP-00120
KPGP-00121
KPGP-00122
KPGP-00124
KPGP-00125
KPGP-00127
KPGP-00128
KPGP-00129
KPGP-00131
KPGP-00132
KPGP-00134
KPGP-00136
KPGP-00137
KPGP-00138
KPGP-00139
KPGP-00141
KPGP-00142
KPGP-00144
KPGP-00145
KPGP-00205
KPGP-00220
KPGP-00224
KPGP-00227
KPGP-00228
KPGP-00229
KPGP-00230
KPGP-00232
KPGP-00233
KPGP-00235
KPGP-00245
KPGP-00254
KPGP-00255
KPGP-00256
KPGP-00265
KPGP-00266
KPGP-00269
KPGP-00318
KPGP-00319
KPGP-00325
Individuals
B
(1e-4)
10.0
9.0
8.0
7.0
6.0
5.0
4.0
Pairwise Genetic Distance (pi)
MT
PC
Br
GPC
UN
Co
KoVariome
Groups


A
SNV
1000GP Common (51.03)
1000GP Low frequency (48.82)
1000GP Rare (0.15)
KoVariome Specific
0% 10% 20% 30% 40% 50% 60% 70% 80% 90% 100%
: Frequent in KoVariome
indel
1000GP Common (64.20)
1000GP Low frequency (35.80)
1000GP Rare (0.00)
KoVariome Specific
0% 10% 20% 30% 40% 50% 60% 70% 80% 90% 100%
: Rare in KoVariome
B
KOR: 2.7M (61.19%)
EAS (90.4K)
89.5K/62.7K
SAS (93.5K)
5.2K/85.1K
EUR (97.7K)
191/91.8K
11.9K/69.1K
AMR (86.1K)
21.7K/383.2K
AFR (408.0K)
EAS (13.6K)
5.3K/8.0K
KOR: 156K (51.36%)
SAS (7.5K)
357/6.8K
EUR (8.2K)
12/7.8K
786/5.3K
AMR (6.5K)
2.2K/67.0K
AFR (39.5K)
C
Rare variant ratio
1.60
1.20
0.80
0.40
0.00
0.66 0.74 0.67 0.75 1.16 1.33 0.81 0.87
Intergenic Intron RNA-genes Synonymous Non-snynonymous Splice-site 3'UTR 5'UTR
SNV
Rare variant ratio
1.60
1.20
0.80
0.40
0.00
0.66 0.74 0.75 1.03 1.45 0.64 0.72 0.87
Intergenic Intron RNA-genes In-frame Frame-shift Splice-site 3'UTR 5'UTR
indel


A
1.26%
9.39%
88.70%
1000GP
KoVariome
Rare in Both
B
9.93%
13.37%
27.36%
(219)
43.29%
5.18%
0.87%(7)
RNA-genes
SYNONYMOUS
NON-SYNONYMOUS
SPLICE-SITE
5'UTR
3'UTR
C
No. of pathogenic SNVs
600
500
400
300
200
100
0
KPGP-00001
KPGP-00002
KPGP-00006
KPGP-00032
KPGP-00033
KPGP-00039
KPGP-00056
KPGP-00086
KPGP-00088
KPGP-00090
KPGP-00117
KPGP-00120
KPGP-00121
KPGP-00122
KPGP-00124
KPGP-00125
KPGP-00127
KPGP-00128
KPGP-00129
KPGP-00131
KPGP-00132
KPGP-00134
KPGP-00136
KPGP-00137
KPGP-00138
KPGP-00139
KPGP-00141
KPGP-00142
KPGP-00144
KPGP-00145
KPGP-00205
KPGP-00220
KPGP-00224
KPGP-00227
KPGP-00228
KPGP-00229
KPGP-00230
KPGP-00232
KPGP-00233
KPGP-00235
KPGP-00245
KPGP-00254
KPGP-00255
KPGP-00256
KPGP-00265
KPGP-00266
KPGP-00269
KPGP-00318
KPGP-00319
KPGP-00325
KoVariome specific
dbSNP


A
No. of SVs
3500
3000
2500
2000
1500
1000
500
0
INV
ITX
INS
DEL
Variant types
B
Log10(length of SVs)
8
7
6
5
4
3
2
INV
ITX
INS
DEL
Variant types
C
No. of SVs (x1,000)
12
10
8
6
4
2
0
INV
DEL
ITX
INS
10
20
30
40
50
No. of individuals
D
Common to DGV
KoVariome Specific
No. of SV (x1,000)
10
8
6
4
2
0
10M
1M
100K
10K
1K
100bp
1K
10K
100K
1M
Portion of repeats
1.0
0.8
0.6
0.4
0.2
0.0
Deletions
Insertion
SINE
LINE
Simple repeat
LTR
DNA transposon
Other

A



B

	DGV ID	No.	Continent					Genes	Average copy No.
			EUR	AFR	AMR	SAS	EAS		
Del.	esv3600874	43	7.69	15.80	10.29	3.98	1.64	*UGT2B17*	0.26
Del.	esv3634909	42	4.53	11.56	5.55	6.28	1.48	*ACOT1*	0.12
Del.	dgv1065e214	6	-	69.27	-	-	-	*NKD2*	0.50
Dup.	esv3587458	24	204.92	312.31	492.92	304.15	94.00	*NBPF15*	4.13
Dup.	esv3635993	18	16.54	1.29	1.44	2.59	0.22	*HERC2*	3.94
Dup.	esv3630980	6	12.90	69.27	24.18	29.91	0.94	*HCAR3,HCAR2*	3.00

Odds ratio: <3 | 3-5 | 5-10 | >10