GigaScience

Korea4K: whole genome sequences of 4,157 Koreans with 107 phenotypes derived from extensive health check-ups

--Manuscript Draft--

Manuscript Number:	GIGA-D-23-00109	
Full Title:	Korea4K: whole genome sequences of 4,157 Koreans with 107 phenotypes derived from extensive health check-ups	
Article Type:	Research	
Funding Information:	Ulsan National Institute of Science and Technology (1.200108.01)	Dr. Jong Hwa Bhak
	Ulsan National Institute of Science and Technology (1.200047.01)	Dr. Jong Hwa Bhak
	Small and Medium Business Administration (1425157253)	Dr. Jong Hwa Bhak
	Small and Medium Business Administration (1425157301)	Dr. Jong Hwa Bhak
	Small and Medium Business Administration (1425156792)	Dr. Jong Hwa Bhak
	Ministry of Trade, Industry and Energy (20016225)	Dr. Jong Hwa Bhak

Abstract:

Background

Phenome-wide association studies (PheWASs) have been conducted on Asian populations, including Koreans, but many were based on chip or exome genotyping data. Such studies have limitations regarding whole-genome-wide association analysis, making it crucial to have genome-to-phenome association information with the largest possible whole-genome and matched phenome data to conduct further population-genome studies and develop healthcare services based on population genomics.

Results

Here, we present 4,157 whole-genome sequences (Korea4K) coupled with 107 health check-up parameters as the largest genomic resource of the Korean Genome Project. It encompasses most of the variants with allele frequency > 0.001 in Koreans, indicating that it sufficiently covered most of the common and rare genetic variants with commonly measured phenotypes for Koreans. Korea4K provides 45,537,252 variants, and half of them were not present in Korea1K (1,094 samples). We also identified 1,356 new geno-phenotype associations which were not found by the Korea1K dataset. Phenomics analyses further revealed 24 significant genetic correlations, 1,131 pleiotropic variants, and 127 causal relationships based on Mendelian randomization among 37 traits. In addition, the Korea4K imputation reference panel, the largest Korean variants reference to date, showed a superior imputation performance to Korea1K across all allele frequency categories.

Conclusions

Collectively, Korea4K provides not only the largest Korean genome data but also corresponding health check-up parameters and novel genome-phenome associations. The large-scale pathological whole-genome-wide omics data will become a powerful set for genome-phenome level association studies to discover causal markers for the prediction and diagnosis of health conditions in future studies.

Corresponding Author:	Jong Hwa Bhak, Ph.D. Ulsan National Institute of Science and Technology Ulsan, Ulsan KOREA, REPUBLIC OF
Corresponding Author Secondary Information:	
Corresponding Author's Institution:	Ulsan National Institute of Science and Technology

Corresponding Author's Secondary Institution:	
First Author:	Sungwon Jeon, Ph.D.
First Author Secondary Information:	
Order of Authors:	Sungwon Jeon, Ph.D.
	Hansol Choi, Ph.D.
	Yeonsu Jeon, Ph.D.
	Whan-Hyuk Choi, Ph.D.
	Hyunjoo Choi, Master of Science
	Kyungwhan An, Bachelor of Science
	Hyojung Ryu, Ph.D.
	Jihun Bhak, Bachelor of Science
	Hyeonjae Lee, Bachelor of Science
	Yoonsung Kwon, Bachelor of Science
	Sukyeon Ha, Bachelor of Science
	Yeo Jin Kim, Ph.D.
	Asta Blazyte, Master of Science
	Changjae Kim, Ph.D.
	Yeonkyung Kim, Master of Science
	Younghui Kang, Bachelor of Science
	Yeong Ju Woo, Bachelor of Science
	Chanyoung Lee, Bachelor of Science
	Jeongwoo Seo, Bachelor of Science
	Changhan Yoon, Master of Science
	Dan Bolser, Ph.D.
	Orsolya Biro, Ph.D.
	Eun-Seok Shin, M.D., Ph.D.
	Byung Chul Kim, Ph.D.
	Seon-Young Kim, Ph.D.
	Ji-Hwan Park, Ph.D.
	Jongbum Jeon, Ph.D.
	Dooyoung Jung, Ph.D.
	Semin Lee, Ph.D.
	Jong Hwa Bhak, Ph.D.
Order of Authors Secondary Information:	
Additional Information:	
Question	**Response**
Are you submitting this manuscript to a special series or article collection?	No
Experimental design and statistics	Yes

Full details of the experimental design and statistical methods used should be given in the Methods section, as detailed in our Minimum Standards Reporting Checklist. Information essential to interpreting the data presented should be made available in the figure legends. Have you included all the information requested in your manuscript?	
Resources A description of all resources used, including antibodies, cell lines, animals and software tools, with enough information to allow them to be uniquely identified, should be included in the Methods section. Authors are strongly encouraged to cite Research Resource Identifiers (RRIDs) for antibodies, model organisms and tools, where possible. Have you included the information requested as detailed in our Minimum Standards Reporting Checklist?	Yes
Availability of data and materials All datasets and code on which the conclusions of the paper rely must be either included in your submission or deposited in publicly available repositories (where available and ethically appropriate), referencing such data using a unique identifier in the references and in the "Availability of Data and Materials" section of your manuscript. Have you have met the above requirement as detailed in our Minimum Standards Reporting Checklist?	Yes

Korea4K: whole genome sequences of 4,157 Koreans with 107 phenotypes derived from extensive health check-ups

Authors

Sungwon Jeon[1,2,†], Hansol Choi[1,3,†], Yeonsu Jeon[1,2], Whan-Hyuk Choi[1,3], Hyunjoo Choi[1,3], Kyungwhan An[1,3], Hyojung Ryu[1,2], Jihun Bhak[1,3], Hyeonjae Lee[1,3], Yoonsung Kwon[1,3], Sukyeon Ha[1,4], Yeo Jin Kim[2], Asta Blazyte[1,3], Changjae Kim[2], Yeonkyung Kim[2], Younghui Kang[1,2], Yeong Ju Woo[2], Chanyoung Lee[1,3], Jeongwoo Seo[1,3], Changhan Yoon[1,3], Dan Bolser[5], Orsolya Biro[6], Eun-Seok Shin[7], Byung Chul Kim[2], Seon-Young Kim[8], Ji-Hwan Park[8], Jongbum Jeon[8], Dooyoung Jung[3], Semin Lee[1,3,*], and Jong Bhak[1,2,3,9,*]

[1]Korean Genomics Center (KOGIC), Ulsan National Institute of Science and Technology (UNIST), Ulsan 44919, Republic of Korea

[2]Clinomics Inc., Ulsan 44919, Republic of Korea

[3]Department of Biomedical Engineering, College of Information-Bio Convergence Engineering, Ulsan National Institute of Science and Technology (UNIST), Ulsan 44919, Republic of Korea

[4]Department of Computer Science & Engineering (CSE), College of Information-Bio Convergence Engineering, Ulsan National Institute of Science and Technology (UNIST), Ulsan 44919, Republic of Korea

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

[6]Clinomics Europe Ltd., Budapest 1094, Hungary

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

[8]Korea Bioinformation Center, Korea Research Institute of Bioscience and Biotechnology, Daejeon 34141, Republic of Korea.

[9]Personal Genomics Institute (PGI), Genome Research Foundation (GRF), Osong, 28160, Republic of Korea

† These authors contributed equally to this work.

* Correspondence to: seminlee@unist.ac.kr (S.L.), and jongbhak@genomics.org (Jong B.)

Abstract

Background

Phenome-wide association studies (PheWASs) have been conducted on Asian populations, including Koreans, but many were based on chip or exome genotyping data. Such studies have limitations regarding whole-genome-wide association analysis, making it crucial to have genome-to-phenome association information with the largest possible whole-genome and matched phenome data to conduct further population-genome studies and develop healthcare services based on population genomics.

Results

Here, we present 4,157 whole-genome sequences (Korea4K) coupled with 107 health check-up parameters as the largest genomic resource of the Korean Genome Project. It encompasses most of the variants with allele frequency > 0.001 in Koreans, indicating that it sufficiently covered most of the common and rare genetic variants with commonly measured phenotypes for Koreans. Korea4K provides 45,537,252 variants, and half of them were not present in Korea1K (1,094 samples). We also identified 1,356 new geno-phenotype associations which were not found by the Korea1K dataset. Phenomics analyses further revealed 24 significant genetic correlations, 1,131 pleiotropic variants, and 127 causal relationships based on Mendelian randomization among 37 traits. In addition, the Korea4K imputation reference panel, the largest Korean variants reference to date, showed a superior imputation performance to Korea1K across all allele frequency categories.

Conclusions

Collectively, Korea4K provides not only the largest Korean genome data but also corresponding health check-up parameters and novel genome-phenome associations. The large-scale pathological whole-genome-wide omics data will become a powerful set for genome-phenome level association studies to discover causal markers for the prediction and diagnosis of health conditions in future studies.

Keywords

Korean Genome Project, Genome, Phenome, Population genomics, Variome

Background

South Korea has perhaps one of the most extensive and convenient annual health check-up services. Every year, almost all Koreans aged over 40 receive a standardized health check-up, and the amount of individual clinical data is very extensive [1]. In 2020, we published 1,094 whole genomes with clinical information (Korea1K) by providing all the participants with an extensive and free standard health check-up showing the value of whole-genome data accompanied by clinical information mapping the genome diversity with practical applications [2]. Here, we present the second phase of the Korean Genome Project (KGP) with 4,157 sets of whole-genome data, Korea4K. It is accompanied by 107 types of clinical traits that have been donated by 2,685 healthy participants who acquired the health check-up reports from the hospitals of their choice in the past years. We manually annotated thousands of donated health reports that are matched with the whole-genome information. Therefore, apart from the increased number of samples, the main difference between Korea1K and Korea4K is that Korea4K's clinical information is from very heterogeneous but fairly standard Korean health check-up centers, while Korea1K was from one very well-controlled university hospital health check-up center. This was also a testbed to assess how difficult it would be to merge data from the heterogeneous health check-up record system in a nation for a large-scale genome to phenome association analysis.

Previously, there were a few phenome-wide association studies (PheWASs) on Asian populations, but they were limited to chip or exome-based genotyping data. A Japanese PheWAS identified the genetic links among clinical traits, complex diseases, and cell-type specific patterns [3]. Another PheWAS using 10,000 Korean cohorts' health check-up data from multiple lab sources showed network relationships between genes and phenotypes [4]. However, none of these studies covered the entirety of genomic variation, and they have limitations on genome-wide data analyses [5, 6].

A scientific contribution of this version of KGP is that we provide extensive genome-to-phenome association information with the largest genomic and clinical data from Korea to date to estimate how many samples and clinical parameters cover the whole genomic and common phenotypic diversity of Koreans. Korea4K contains 4,157 Korean genomes from East Asian ancestry, and 2,685 of them are accompanied by 107 types of clinical information such as height, waist circumference, weight, albumin/globulin ratio, basophil, direct bilirubin, low-density lipoprotein, high-density lipoprotein, mean corpuscular volume, and total cholesterol. The rest does not contain such kind of data because the biobank does not have phenotype information, or we were not able to collect it from the participants. Korea4K extends the efforts to completely map the totality of Korean genomic diversity, which can be a useful scope reference for disease risk prediction, diagnosis, and treatments in the future for personalized medicine.

As the second phase of the KGP, Korea4K not only extends the previously reported Korea1K [2] but also includes new multi-phenotypic association analyses, that is, analyses on markers that are associated with multiple phenotypes (pleiotropy), the genetic correlation between traits, and estimated causality relationship among traits through Mendelian randomization (MR) and 3D structure models for Korean specific missense variants. Combining these two omics data, we provide the community with the most extensive geno-phenotype association of healthy Korean participants. We have also applied the genomic variation data to the genotype imputation of low-frequency variants in the Korean population.

Data Description

The goal of our project was to create a genome dataset for Korea4K, which included newly sequenced genomic data from 2,848 participants as well as 1,309 whole-genome sequencing (WGS) datasets from Korea1K and public data archives. Additionally, we established a phenome dataset for Korea4K by gathering or computing 107 clinical parameters and genome data from 2,685 samples. We collected a total of 3,383 clinical datasets, including multiple time points per sample, from regular health checkups conducted by various hospitals and clinics across Korea between 2016 and 2019. The genome and phenome datasets were produced and curated by the protocol in Material and Methods.

Analyses

The largest Korean whole-genome variants data: Korea4K variome

A total of 64,301,272 single nucleotide variants (SNVs) and 8,776,608 Indels were called against the human genome reference (hg38) from the 4,157 Korean whole genomes, including 3,071 healthy controls (Supplementary Table S1 and S2). It contains 3,063 newly added whole genomes sequenced by Illumina next-generation sequencing (NGS) platforms (HiSeq X10 and Novaseq 6000), in addition to the previous Korea1K dataset which was mostly generated by Illumina HiSeq X10. Using the variant data, we selected 3,617 samples with no kinship after initial sample filtering (see Methods). To exclude erroneous variants from sequencing batch effects from the heterogeneous Illumina NGS platforms and library preparation, we applied an allele balance bias measurement and finally acquired 12,713,580 erroneously called variant candidates (Supplementary Fig. S1). After additional variant filtering (see Methods), we identified 45,537,252 variants including 42,124,137 SNVs, 36,029 double nucleotide variants (DNVs), 26,135 triple nucleotide variants (TNVs), 3,261,682 indels, and 89,269 other types of small variants from the 3,617 unrelated samples. We named this filtered Korean dataset the Korea4K variome (Fig. 1). A total of 23,689,147 variants were not present in the previous Korea1K variome. This unexpectedly large difference is likely derived from different batch effect filtering, and variant calling and filtering procedures, as well as new variants from the larger sample size. Consistent with the Korea1K study [2], most variants were located in intronic or intergenic regions and rarely in splicing sites or coding regions (Supplementary Fig. S2), which is a sign of negative selection pressure in the population. Half of the total variants (21,941,879; 48.2%) were singleton or doubleton in the 3,617 unrelated samples, indicating that the Korean population's genetic diversity is very low as the population diversity could be covered by fewer than 4,000 unrelated samples (Fig. 1A, Supplementary Table S3). Almost all the common (allele frequency of > 0.01 and allele frequency of ≤ 0.05) and very common (allele frequency of > 0.05) variants were found to be already reported in the dbSNP database (99.70% and 99.97%, respectively), while more than half of the singleton and doubleton variants were newly discovered in this study (59.9% and 44.57%, respectively), indicating the new variant pool is well-exhausted in the Korean population by the 3,617 samples resulting in a large portion of individual specific novel variants in the Korean variome (Fig. 1A, Supplementary Table S3). Only 3,092 and 3,569 unrelated individuals were needed to discover all the rare (allele frequency of > 0.001 and allele frequency of ≤ 0.01) and very rare (allele count of > 2 and allele frequency of ≤ 0.001) variants in the Korea4K variome, respectively (Fig. 1B) indicating that the Korea4K variome includes almost all the rare and very rare variants of Korean people of East Asian ancestry. It is notable that in our previous Korea1K data, the accumulated variant number curves did not reach a plateau [2]. Regarding common variants, only 481 and 161 unrelated individuals were necessary for common and very common variants, respectively, to cover the diversity which is close to the Korea1K statistics (440 and 132 samples). Essentially, the Korea4K variome statistics indicate the saturation of population diversity detection among Koreans. However, as expected, in the case of singleton and doubleton variants, the Korea4K variant discovery curve did not reach a plateau. This is due to each individual's novel random variants and we will never reach a point of no novel variant discovery even with increased sample numbers.

As a practical application, we constructed a Korea4K imputation reference panel from the 3,614 unrelated whole-genomes that showed a consistently better imputation performance than the Korea1K. The Korea4K panel was able to impute 198,805 more genotypes than the Korea1K panel (7,551,095 loci compared to 7,352,290) with the same dataset. Moreover, as expected, the Korea4K panel had better accuracy across all allele frequency categories than the Korea1K panel (Fig. 1C). The difference in aggregated R^2 became larger for variants with allele frequency (AF) in Korea4K < 0.05 than for those in Korea1K, indicating higher accuracy in rare variants (Fig. 1C). In particular, the Korea4K imputation panel improved the imputation accuracy by 6% for the rare variants group compared to Korea1K on average.

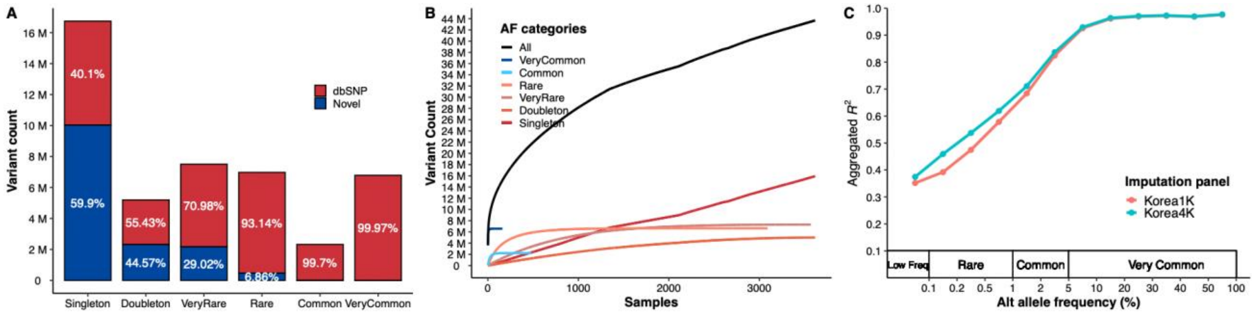
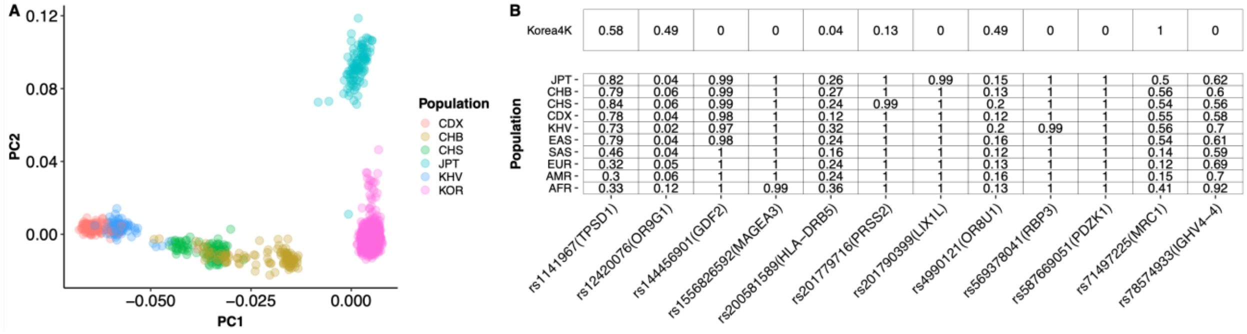


Figure 1 Korean variome profile and imputation evaluation using Korea4K (A) The number of variants in the Korea4K variome is categorized by allele frequencies (AFs) among unrelated Korea4K genomes. Singleton, allele count = 1; doubleton, allele count = 2; very rare, allele count

of > 2 and allele frequency of ≤ 0.001; rare, allele frequency of > 0.001 and allele frequency of ≤ 0.01; common, allele frequency of > 0.01 and allele frequency of ≤ 0.05; very common, allele frequency of > 0.05. **(B)** The number of discovered variants as a function of unrelated genomes. **(C)** Imputation performance evaluation using the Korea4K and Korea1K panels. The X-axis indicates alternative (Alt) allele frequency in the Korea4K variome. The Y-axis represents the aggregated R^2 values of variants. We used variants that were overlapped by imputed results across two panels.

As in Korea1K, the Korean population is genetically distinct from the Chinese and Japanese populations, confirmed by principal component analysis (PCA) with few outliers (Fig. 2A). We also found 62 missense variants out of 282,607 in Korea4K that had AFs significantly different from ten populations in the 1000 genome project (1KGP) from European Bioinformatics Institute (EBI), Cambridge, UK (Chi-squared test $P < 5 \times 10^{-5}$ against each of the ten populations, see Methods; Supplementary Table S4). The genes containing such Korean-specific missense variants included *LILRB3*, *HLA-DRB5*, *IGLV5-48*, and *IGHV4-4* that are known to be associated with adaptive immunity, and *OR9G1* and *OR8U1* for olfactory receptors. Additionally, we found that twelve Korean-specific missense variants were in protein functional domains (Fig. 2B). Four of them were predicted to facilitate increased structural stability calculated in the protein 3D models built by AlphaFlod2 [7], while the other eight variants were predicted to cause decreased stability (Supplementary Table S5).



	rs1141967(TPSD1)	rs12420076(OR9G1)	rs144456901(GDF2)	rs155826592(MAGEA3)	rs200581589(HLA–DRB5)	rs201779716(PRSS2)	rs201790399(LIX1L)	rs4990121(OR8U1)	rs569378041(RBP3)	rs587669051(PDZK1)	rs71497225(MRC1)	rs78574933(IGHV4–4)
Korea4K	0.58	0.49	0	0	0.04	0.13	0	0.49	0	0	1	0
JPT	0.82	0.04	0.99	1	0.26	1	0.99	0.15	1	1	0.5	0.62
CHB	0.79	0.06	0.99	1	0.27	1	1	0.13	1	1	0.56	0.6
CHS	0.84	0.06	0.99	1	0.24	0.99	1	0.2	1	1	0.54	0.56
CDX	0.78	0.04	0.98	1	0.12	1	1	0.12	1	1	0.55	0.58
KHV	0.73	0.02	0.97	1	0.32	1	1	0.2	0.99	1	0.56	0.7
EAS	0.79	0.04	0.98	1	0.24	1	1	0.16	1	1	0.54	0.61
SAS	0.46	0.04	1	1	0.16	1	1	0.12	1	1	0.14	0.59
EUR	0.32	0.05	1	1	0.24	1	1	0.13	1	1	0.12	0.69
AMR	0.3	0.06	1	1	0.24	1	1	0.16	1	1	0.15	0.7
AFR	0.33	0.12	1	0.99	0.36	1	1	0.13	1	1	0.41	0.92

Figure 2 Comparison of Korea4K and 1KGP (A) The results from principal component analysis of Korea4K and the 1KGP set of East Asian samples. **(B)** Allele frequency information of Korea4K and the populations in the 1KGP for the twelve Korean-specific missense variants located in protein functional domains. KOR: Korea4K; CDX: Dai Chinese; CHB: Han Chinese; CHS: Southern Han Chinese; JPT: Japanese; KHV: Kinh Vietnamese; EAS: East Asians; SAS: South Asians; EUR: European; AMR: American; AFR: African.

Whole-genome-wide association study (WGWAS)

Whole-genome-wide association studies (WGWASs) revealed that 2,324 variants from 157 unique loci had significant associations with 34 clinical traits from 37 WGWAS target traits ($P < 5 \times 10^{-8}$; Fig. 3A-F, Supplementary Table S6). We used 90 clinical traits from the 107 phenotypes after filtering 27 traits with a high missing rate and biased distribution for WGWASs (see Methods). Of the 90 traits, 54 were not confident in Quantile-Quantile plots and were excluded from further Mendelian randomization and pleiotropy analyses (see Methods). Among the 2,324 WGWAS significant variants, only 85 variants (31 loci) were reported in the GWAS catalog database [8]. The trait with the largest number of significantly associated loci was carbohydrate antigen 19-9

(CA19-9), a cancer antigen, with sixteen loci. Uric acid had the second highest number of significant loci with fourteen loci.

Korea4K showed much stronger statistical power than the previous Korea1K study, identifying 1,356 new WGWAS variants (107 loci) from 28 common traits between Korea4K and Korea1K. Also, Korea4K had much lower (i.e., more significant) *P*-values than Korea1K for all the commonly found association variants between the two datasets (Supplementary Fig. S3). Among the 107 loci containing the 1,356 new WGWAS variants, 798 Korea4K significant WGWAS variants from 73 loci had not been significant in Korea1K (Supplementary Table S6). Furthermore, twelve traits (albumin/globulin ratio, basophil, C-reactive protein, direct bilirubin, height, low-density lipoprotein, mean corpuscular volume, right hearing at 2000hz, thyroid stimulating hormone, total cholesterol, waist, weight) had 425 WGWAS variants that were significant uniquely in Korea4K, meaning no significant WGWAS variants from the twelve traits in Korea1K (Supplementary Table S6). For example, a missense variant, rs6431625 ($P = 1.41 \times 10^{-23}$), in *UGT1A3* was found to be associated with direct bilirubin in Korea4K. It was previously reported to be associated with circulating bilirubin levels [9]. Another Korea4K-specific missense variant is rs7412 ($P = 2.86 \times 10^{-14}$) in *APOE* which is associated with low-density lipoprotein (LDL) levels. Its association with cholesterol levels has been previously well-established [10]. Finding novel WGWAS variants in Korea4K was due to the increased sample size and subsequently increased variant number compared to Korea1K.

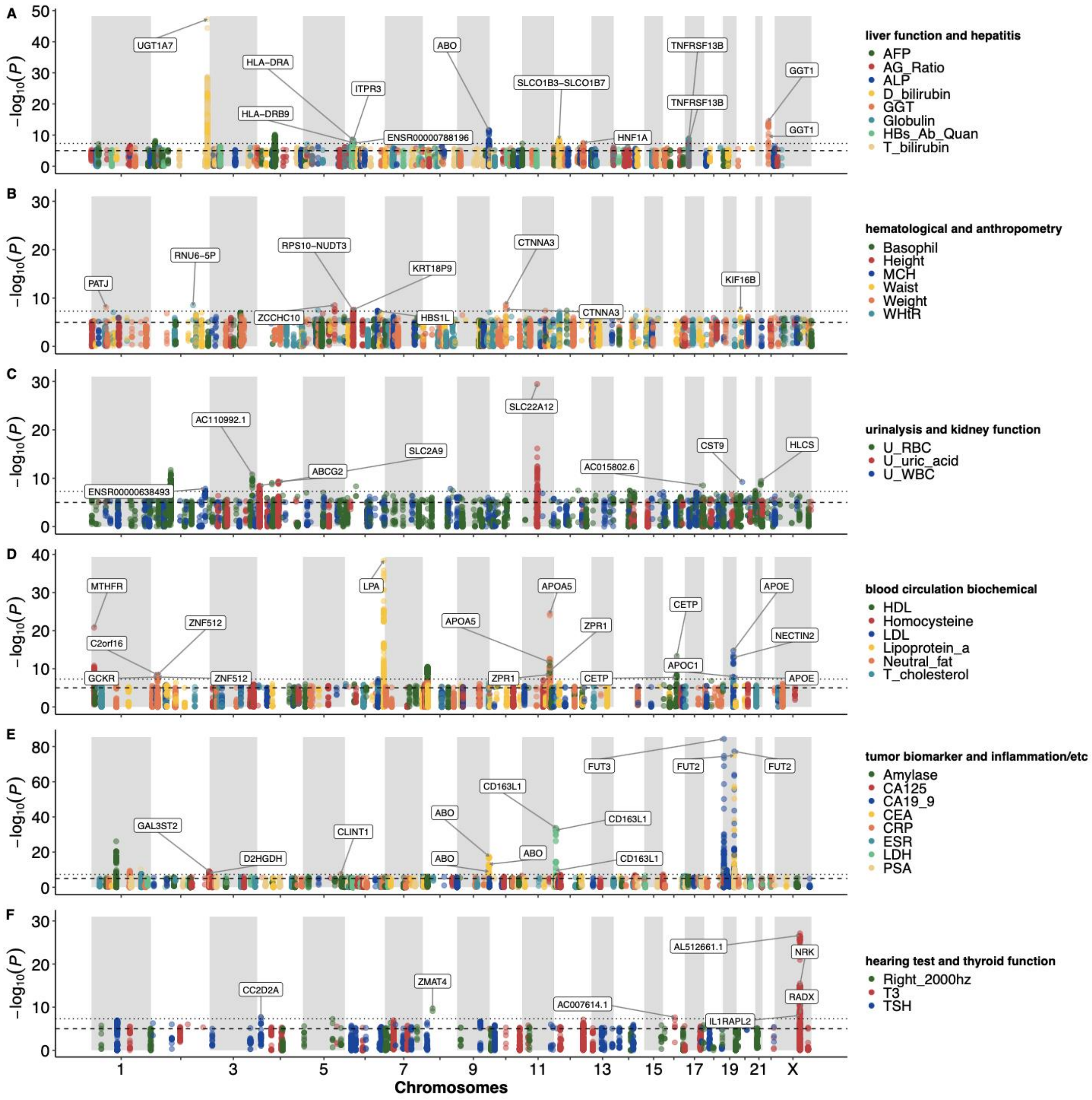


Figure 3 Whole-genome-wide association studies in Korea4K. (A-F) Whole-genome-wide association studies from 34 traits. Loci are presented only when index variants of the loci had significant *P*-value ($P < 5 \times 10^{-8}$) from the WGWAS. The dashed line indicates the suggestive threshold ($P < 10^{-5}$). The dotted line indicates the significant threshold ($P < 5 \times 10^{-8}$).

Genetic correlation (GC) and phenotypic correlation (PC)

We found 27 traits with significant heritability among 89 quantitative traits (Fig. 4A; the lower boundary of genetic heritability > 0 with 95% confidence interval; Supplementary Table S7). A total of 24 pairs of traits showed a significant genetic correlation ($FDR_{GC} < 0.05$), measured as rG value, among 351 trait pairs between the 27 traits that showed significant heritability (Fig. 4, Supplementary Table S8). We found consistent results of Weight-Waist and body mass index (BMI)-Waist pairs, showing a significant genetic correlation in the UK Biobank data (http://www.nealelab.is/uk-biobank) with the same trend as our result (rG = 0.9, $P = 10^{-308}$ in UK Biobank; rG = 0.9, $P = 10^{-308}$ in UK Biobank, respectively). We identified 2,274 trait-trait relationships that had significant phenotypic correlation (FDR < 0.05, its 95% CI does not include 0) from trait-trait associations between 3,916 pairs of 89 quantitative traits (Fig. 4B, Supplementary Table S9). Most genetic and phenotypic correlations showed the same direction of correlation. The only two exceptions were waist/weight ratio (WWtR) – Urine white blood cell (U_WBC) and Waist-Creatine which showed opposite directions. This trend of Waist-Creatine has also been reported in a correlation database using UK-biobank data [11].

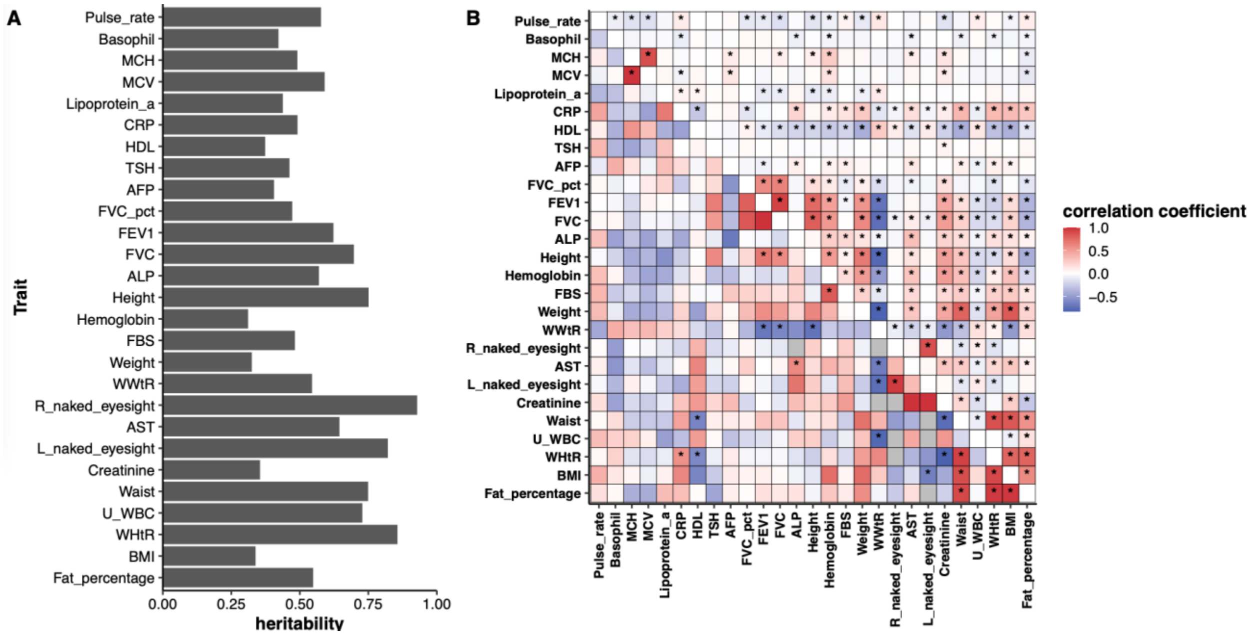


Figure 4 Genetic correlation and Phenotypic correlation in Korea4K. (A) Genetic heritability of 27 traits that showed at least a marginal significance. **(B)** Genetic correlation and phenotypic correlation between the 27 traits. The upper triangle indicates phenotypic correlation coefficient (Pearson's) and lower triangle indicates genetic correlation coefficient (rG).

Pleiotropy and Mendelian randomization (MR)

Out of the 37 WGWAS target traits, we detected 1,131 variants from 21 traits having suggestive associations ($P_{GWAS} < 10^{-5}$) with at least two traits, implying pleiotropic variants (Fig. 5, red edges; Supplementary Table S10). We devised the Variant-Sharing Index (VSI) to measure the degree of intersection between two phenotypes (Table 1; See methods). If the VSI equals zero, two traits share no suggestively associated variants, while 100 indicates the traits share all of them. The trait pairs with shared suggestive variants (SSVs) and the corresponding VSIs are listed in Table1. Notably, we had only one variant, rs77913154 (chr5:18853857), that was shared among three traits: Globulin, AG_Ratio, and ESR (Supplementary Table S10). Interestingly, we found fifteen variants residing on *SOD2P1-AC095032.2-AC095032.1* locus forming pleiotropy between the serum amylase level and the level of CA125, a known ovarian cancer marker (VSI=2.3). Fourteen variants of the fifteen variants conform to the alteration of *AMY2B* level based on cis-eQTL results from GTEx Portal (ver.8), four of which were associated with expression in the pancreatic tissue. Already, there have been reports that patients with ovarian cancer manifest hyperamylasemia [12-14]. In terms of causality evaluation, a total of 127 trait pairs among 1,332 pairs of the 37 WGWAS traits were estimated to have significant causal relationships (FDR < 0.05, Fig. 5, Supplementary Table S11) from at least two of three different Mendelian randomization (MR) analysis methods (IVW: 166 pairs; MRPRESSO: 139; MR-Egger: 23). We found 59 unidirectional relationships and 68 bidirectional causal relationships (Supplementary Table S11).

Table 1 Pleiotropic traits and Variant-Sharing Index (VSI)

Trait1	Trait2	Suggestive variants in trait1	Suggestive variants in trait2	Shared variants	Total variants	VSI
D_bilirubin	T_bilirubin	638	632	569	701	81.2
Globulin	AG_Ratio	294	230	147	377	39
HDL	Neutral_fat	348	398	191	555	34.4
CEA	CA19_9	221	264	74	411	18
T_cholesterol	LDL	74	238	38	274	13.9
WHtR	Waist	177	100	31	246	12.6
ALP	CEA	153	221	35	339	10.3
T3	GGT	542	125	23	644	3.6
CA125	Amylase	202	466	15	653	2.3
Weight	Waist	123	100	5	218	2.3
Height	Weight	173	123	2	294	0.7
ESR	AG_Ratio	163	230	1	392	0.3
Globulin	ESR	294	163	1	456	0.2
U_RBC	Globulin	627	294	1	920	0.1

Summary results of the four phenomics analyses

We summarized the four phenomics analyses (Genetic correlation, Phenotypic correlation, Mendelian randomization, and pleiotropy) by visualizing them in network plots (Fig. 5). In general, the discovered trait-trait pairs of genetic correlation, Mendelian randomization, and pleiotropy analysis results were not often overlapping. Nevertheless, the network visualization suggests distinguishable association patterns of the two secondary body measures, WHtR and WWtR with other phenotypes. WHtR had associations with the C-reactive protein (CRP), creatine, and HDL. On the other hand, WWtR was associated with aspartate aminotransferase (AST), forced expiratory volume (FEV1), forced vital capacity (FVC), and urine white blood cell (U_WBC). Despite their similarity, they may reflect different biological mechanisms.

Genetic correlation and pleiotropy are found exclusive of each other, despite both measures having shared genetic components of two different traits. GC is mainly observed from body measures such as waist, weight, height, and Left-naked eyesight. Pleiotropy was more on the relationship between metabolites in blood such as LDL, bilirubin, or CEA. The only overlap is WHtR-Waist, where one is derived from the other. MR analysis suggests a causal relationship between phenotypic correlations. For example, the Fat-percentage influences Waist and WHtR, which is followed by the influence on HDL and CRP. The result is concordant with previous reports that body fat percentage and CRP are correlated [15, 16]. ALP and CEA showed potential causality, as well as the shared variants between them (pleiotropy near *ABO* gene). Many previous studies reported them together as targets for diagnosing cancer and monitoring metastasis [17-19]. Nevertheless, their molecular-level relationship has not been suggested. Our phenomics results also suggested that Waist/Height ratio is a linked trait in the association of paired traits, and the association is probably derived from indirect causation.

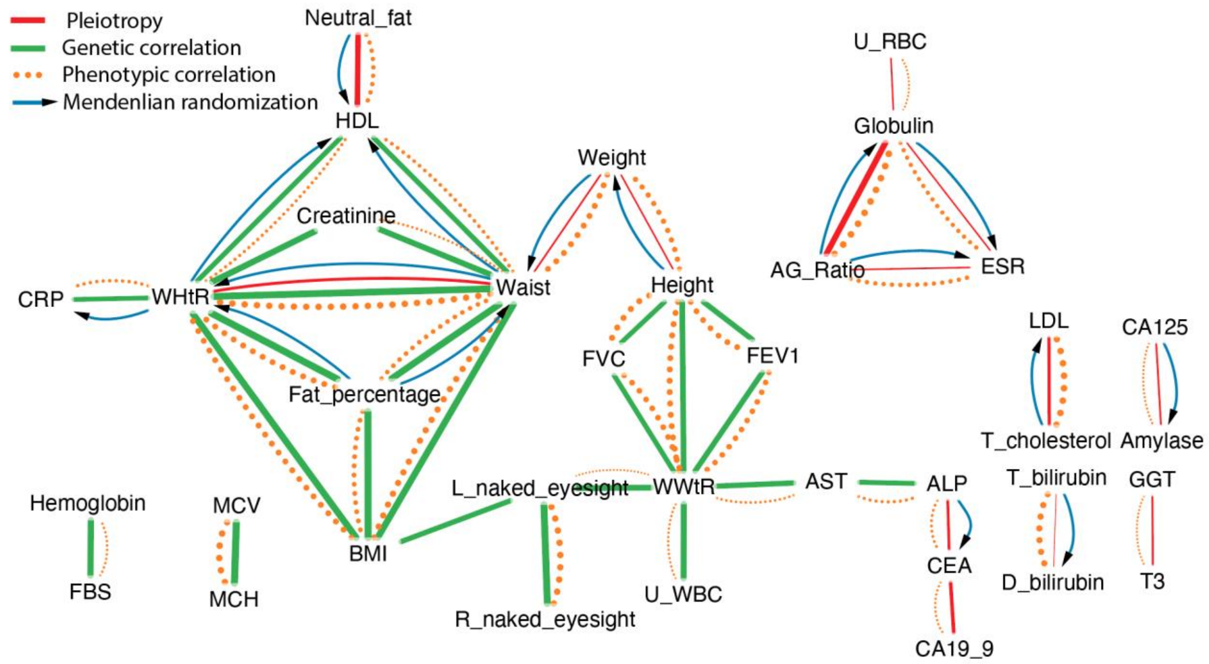


Figure 5 Graph visualization of genetic correlation, phenotypic correlation, pleiotropy, and Mendelian randomization. Green line indicates significant genetic correlation (GC), and the edge thickness indicates the absolute value of the correlation coefficient. Red line indicates trait pairs that have pleiotropic variants. Dotted orange lines indicate phenotypic correlation (PC), and the edge thickness indicates the absolute value of Pearson's correlation coefficient. Blue arrow line indicates a causal relationship from Mendelian randomization (MR). MR and PC were visualized only when at least one of GC or Pleiotropy relationships was observed between the traits.

Discussion

Batch effect exacerbated by sequencing platform and library preparation bias is a critical problem in very large population genome association studies, especially with clinical data from heterogeneous health check-up centers. In the future, more and more diverse whole-genome data with extensive clinical data will be publicly available, and it is inevitable that they will be merged for more precise whole genome-to-phenome association research. Korea4K is not an exception in that regard, and in one homogeneous population WGWAS, it was necessary to consider and factor in a great deal of sequencing and clinical data batch effects and errors. We attempted to minimize the errors by using allele balance with optimal filtering criteria and time-consuming manual checks on health reports that were donated by the participants. The largest challenge of Korea4K project was cleaning up heterogeneous clinical data from different health check-up centers. Another major issue was that the health check-up data heterogeneity caused reduced numbers of participants' common traits with which to compare. Some of the health data were from past years' health check-ups from heterogeneous hospitals throughout Korea. This heterogeneity in location and time was not an intentional experimental design but was in order to reduce the cost of performing expensive one-center health check-ups for the Korea4K participants. Therefore, WGWAS along with standardized and unified national and public health check-up data will greatly benefit future whole-genome-wide association studies.

Although 4,157 seems like a large number, we found the sample size in this study was still not large enough to detect weak association signals. The Korea4K variome with matched phenotype information has allowed us to estimate genomic correlation across various phenotypes using GREML [20]. GREML has been reported to have higher accuracy compared to methods, such as linkage disequilibrium score regression (LDSC), using only summary statistics from GWAS [21]. For example, the minimum heritability score was 0.34 (Degree of obesity) among the traits detected as statistically significant. The statistical power of our maximum 2,685 subjects and FDR < 0.05 is estimated to be 0.72 for detecting traits with heritability of 0.3 or higher (Calculated from GCTA-GREML Power Calculator) [22]. This will increase to 0.97 with 4,000 subjects.

WGWAS, whole-genome-wide association, not chip-based GWAS, performs better in geno-phenotype association studies, and we suggest WGS for future studies for its genetic data completeness. Our pleiotropy analysis based on the WGWAS made it possible to reveal the portions of genetic association across multiple traits. For example, we could identify the variants in the well-known pleiotropic relationships such as ALP-CEA by *ABO* locus (35 variants), Neutral_Fat-HDL by *LPL* locus (181 variants) and Total cholesterol-LDL by *TOMM40,* and *APOE* locus (4 and 2 variants, respectively). These loci and their corresponding trait pairs were previously reported from chip-based GWAS summary results [23, 24]. However, we found more pleiotropic variants thanks to whole-genome-wide, unbiased coverage of WGS. Notably, the four methods that we adopted produced discrete trait-trait relationships, which means that multiple phenomics methods should be applied to investigate specific relationships and mechanisms among clinical traits or diseases. In other words, phenomics analyses were limited and not powerful enough to discover novel and indirect associations with current datasets.

One of the purposes of Korea4K was to build a reference dataset to discover unknown whole-genome to phenome associations that can be detected from samples of healthy people. This, however, is contradictory and it limited us in discovering clear pathogenic associations because most of the participants examined in WGWAS were healthy without any severe aberrant phenotypes or diseases that could bring us clues for interesting omics analyses.

There are three important limitations of our study. The first is we failed to acquire long DNA sequencing reads from the healthy participants for building a structural variation reference set for the Korean population. The second is the lack of epigenomic data from the 4,157 samples. This was mostly due to high costs for generation and computing long-read based assemblies and sequencing methylated DNA sites. The third one, which is perhaps the most relevant for the purpose of performing association studies for healthcare is that we failed to acquire more rare and severe disease data from patients, accompanied by precise clinical and multiomics data. We have excluded a small number of rare disease cases, as those required a large amount of sequencing data from genome, transcriptome, and methylome to perform precise functional analyses. Large-scale pathological whole-genome-wide omics data will become a powerful set for genome-phenome level association studies to detect causal markers for the prediction and diagnosis of health conditions in future studies.

Potential Implications

The Korea4K dataset can be a valuable variome reference, as it contains matched phenome data for personalized medicine, large-scale population genome studies, and the understanding of anthropologic history in Korea. This large-scale Korean genome-phenome dataset can help identify genetic basis for diseases and phenotypes, enabling personalized treatment plans for individuals. Analyzing the genome-phenome association dataset can also be used to develop new drugs that target specific genetic variations in the Korean population. The Korea4K dataset can also be valuable for other populations, particularly East Asians, as it can be used to identify population-specific genome-phenome patterns by comparing the population's genome-phenome data to the Korea4K dataset. Furthermore, the Korea4K reference panel can be utilized for genotype imputation of DNA chip genotyping data for the Korean population and other East Asians.

Materials and Methods

Sample collection and whole-genome sequencing

We collected 2,848 blood samples or already processed DNA samples from Korean individuals. A total of 1,094 whole-genome sequencing (WGS) datasets originating from our previous study (Korea1K) and 215 WGS data from publicly available Clinical & Omics Data Archive (CODA) were added to the aforementioned dataset [2]. The genomic DNA was extracted using the DNeasy Blood & Tissue kit (Qiagen) from whole blood samples. We constructed the whole-genome sequencing library from the DNA by using the TruSeq Nano DNA Sample Prep kit (Illumina) kit. Whole-genome sequences of the 2,848 samples were generated by the Illumina Nova-seq 6000 platform.

Joint genotype calling

Adapter contamination was trimmed using Cutadapt (RRID:SCR_011841, ver. 1.9.1) [25] with a forward adapter (′GATCGGAAGAGCACACGTCTGAACTCCAGTCAC′) and reverse adapter (′GATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT′) and with a minimum read length of 50 bp after trimming. We mapped the whole-genome sequencing reads from 4,157 samples to the human reference genome (hg38) using BWA-mem (RRID:SCR_010910, ver. 0.7.17) with the '-M' option and alt-aware mode [26]. The mapped reads were sorted by genomic coordination using Picard (RRID:SCR_006525, ver. 2.20.3). We marked the PCR-duplicates and recalibrated the base quality of the mapped reads using the MarkDuplicates and BaseRecalibrator module in Picard (RRID:SCR_006525, ver. 2.20.3), respectively. A total of 3,156 samples had a mapping depth of ≥ 20 × (Supplementary Fig. S4). Individual genotypes were called in GVCF format by HaplotypeCaller in GATK (RRID:SCR_001876, ver. 4.1.3) with '--genotyping-mode DISCOVERY -stand-call-conf 30 -ERC GVCF' options [27]. We merged the individual genotypes to a single GVCF for each chromosome using CombineGVCFs in GATK (RRID:SCR_001876, ver. 4.1.3) [27]. We jointly genotyped the merged GVCF with the genotypeGVCF module in

GATK (RRID:SCR_001876, ver.4.1.3) [27]. Variant quality of the joint genotypes was recalibrated using the VQSR module in GATK (RRID:SCR_001876, ver. 4.1.3) [27].

Sample and variant filtering

After joint genotyping, we filtered out a total of 540 participants on the criteria that are listed below using SelectVariants in GATK (RRID:SCR_001876, ver. 4.1.3) with '--remove-unused-alternates' option to remove unused variants [27].

1. showing high genotype missing rate (>10%): nine samples
2. having too high or low heterozygous variants ratio compared to homozygous variants per sample (3 s.d.): four samples
3. having relatedness to other samples: 428 samples
4. having non-Korean genetic background from PCA analysis with 1KGP set: seven samples
5. reported to have a rare disease: 40 samples
6. 52 samples who became not applicable for this study

Finally, the Korea4K variome data included 3,617 participants' genomes. To detect variants which were probably called because of a sequencing batch effect, we measured average allele balance of the alleles. Then, we excluded 12,713,580 variants that had average allele balance of the loci out of the range of ± 1 × standard deviation (SD) from a genome-wide average of allele balance to remove the sequencing batch effect (Supplementary Fig. S1). We also excluded the variants which had a genotyping rate of < 0.9 for downstream variant analysis. The variants in the final variome set were annotated using Variant Effect Predictor (VEP) with Ensemble database (RRID:SCR_007931, ver. 101) [28].

Principal Component Analysis (PCA) with the EBI's 1KGP genome data

The interpopulation genomic structure was evaluated by projecting the first two PCs determined via PCA of SNVs from both Korea4K and East Asian populations from 1KGP. We merged variants from the Korea4K and 1KGP sets and then filtered out variants with the following criteria: (i) biallelic SNVs with a MAF < 1%; (ii) biallelic SNVs with an HWE $P < 10^{-6}$; (iii) biallelic SNVs with a missing genotype rate of > 0.01. Extracted variants were LD pruned using " --indep 200 4 0.1" option in PLINK (RRID:SCR_001757, ver. 1.90b3n) [29], yielding 330,350 sites. PCA was carried out using PLINK (RRID:SCR_001757, ver. 1.90b3n) [29].

Korean-specific missense variants

We collected allele frequency data from ten populations (African (AFR), American (AMR), European (EUR), South Asian (SAS), East Asian (EAS), Japanese in Tokyo (JPT), Kinh Vietnamese (KHV), Han Chinese in Beijing (CHB), Han Chinese Southern (CHS), and Chinese Dai in Xishuangbanna (CDX)) from EBI's 1KGP database [30]. For each Korea4K variant, we compared its allele frequency to the allele frequency of all of the ten populations using the Chi-squared test. We selected variants that were specific to the Korean when the *P*-value of the Chi-squared test to the ten populations was less than 5×10^{-5}.

Protein structure modeling and thermodynamic stability measurement

We constructed the mutant-type (MT) protein sequences of the Korean-specific missense variants by substituting the reference protein sequences found in the Ensembl database (RRID:SCR_002344, ver. 101) [31]. We modeled the structures of the wild-type (WT) and mutant-type protein models using AlphaFold2 (ver. 2.0) with the ‘--max_template_data 2022-05-09 --db_preset reduced_dbs’ option with default databases downloaded by AlphaFold2 [7]. We used the InterPro (RRID:SCR_006695) database [32] to determine whether a missense variant was located in the domain region within the protein sequence. We extracted the domain region from the WT and MT protein 3D models and excluded domains that had less than 50 amino acids. Afterwards, we calculated ΔG_{WT} and ΔG_{MT} using the ‘Stability’ command of foldX (RRID:SCR_008522) [33] to measure the protein thermodynamic stability. Finally, we measured the change in protein thermodynamic stability between the two models by calculating the difference between the WT and MT domain models ($\Delta\Delta G = \Delta G_{MT} - \Delta G_{WT}$).

Imputation

We constructed an imputation reference panel of Korea4K and Korea1K sets which includes 3,614, and 873 Korean individuals, respectively. A total of 26,210,741 and 15,649,303 autosomal biallelic variants with a missing genotype call rate of < 0.1 and minor allele count > 1 (not a singleton) were extracted for the Korea4K and Korea1K panel, respectively. The extracted variomes were phased into haplotype using SHAPEIT2 (ver. v2.r904) [34]. We used the same test dataset as in the previous study [2]. The phased test data was imputed using the imputation reference panel by Minimac3 (RRID:SCR_009292, ver. 2.0.1) [35]. We estimated imputation accuracies using squared Pearson’s correlation coefficients (R^2) between the true genotypes and imputed genotype dosages.

Clinical information

We collected or calculated 107 clinical parameters (93 quantitative and 14 qualitative traits; Supplementary Table S12) along with genome data from 2,685 samples among the Korea4K samples. A total of 3,383 clinical datasets (including multiple time points per sample) from regular health checkups carried out by various hospitals and clinics throughout Korea were collected from 2,685 participants between 2016 and 2019. When a single participant had multiple clinical datasets, the most recent one was chosen for the following analysis. Four quantitative clinical traits and 12 qualitative traits were excluded from the further analysis, since the traits were missing from more than 90% of participants due to health check-up reports heterogeneity, or the traits that were qualitative and biased to one category (more than 1:4). Standard Weight was also removed from the analysis, because the trait was not an inherently correct representation of the sample's clinical data but rather a recommended value. Three traits (Hepatitis B virus antibody, antigen, and hepatitis C antibody) contained both quantitative and qualitative values. Therefore, both of the values were utilized for analysis, i.e, Hbs_Ab_Quan and Hbs_Ab_Binary. Phenotypic correlations were calculated by Pearson's method.

Whole genome-wide association study (WGWAS)

SNVs and indels with a MAF <1%, HWE $P < 10^{-6}$, and a missing genotype rate of > 0.01 were excluded from the analysis using PLINK (ver. 1.90b3n) [29]. A total of 90 WGWAS (88 quantitative and 2 qualitative traits) were performed with a total of 3,617 individuals and 7,782,381 variants. Each WGWAS had a different number of individuals that included those who had the target clinical traits. The WGWAS was performed using linear and logistic regression under an additive genetic model with PLINK (ver. 2.00 alpha) [36] for quantitative and qualitative traits, respectively. Sex, age, age^2 (age squared), body mass index (BMI), and the top ten principal components of SNV genotypes were included in the model as covariates. BMI was excluded from covariates in the WGWAS for BMI itself and degree of obesity. We rejected 53 traits from further analysis based on QQ-plot analysis (Supplementary Fig. S5-S20). We used 5×10^{-8} for a whole-genome-wide significance threshold. The 7,782,381 variants were clumped into 466,938 loci based on linkage disequilibrium (LD) information using PLINK (ver. 1.90b3n) with '--clump-p1 1, --clump-p2 1, --clump-r2 0.1, --clump-kb 250, and --clump-index-first' options [29].

Measuring heritability and genetic correlation

We calculated genetic relatedness among individuals from SNPs by genetic relationship matrix (GRM) in genome-wide complex trait analysis (GCTA) (ver. 1.93.2) with ' --autosome --maf 0.01 --make-grm' options [20]. We estimated the genetic heritability of 87 quantitative traits using GCTA (ver. 1.93.2) with '--reml --grm' options [20]. We estimated the genetic correlations (GC) using the bivariate genome-based restricted maximum likelihood (GREML) algorithm [37] in the GCTA (ver. 1.93.2) with '--reml-bivar --grm --reml-bivar-lrt-rg' options [20]. Two of the 253 trait pairs were excluded since the log-likelihood did not converge.

Calculation of Variant Sharing Index (VSI)

The variant sharing index (VSI) is a Jaccard score to measure how many pleiotropic components exist out of all significant variants from i-th and j-th traits, which is defined as

$$VSI(i,j)= |S_i \cap S_j| / |S_i \cup S_j|$$

where S_i and S_j denote sets of significant variants for the i-th and j-th traits, respectively. The VSI increases as two traits have more pleiotropic variants among their significant variants.

Pleiotropic variants with tissue-specific expression regulatory function

We annotated the gene symbol of the pleiotropic variant by using Ensemble database (ver. 101) [31]. In case of intergenic variants, we annotated the genes which were located the nearest in both directions of the variant. The single tissue eQTL data (ver. 8) from the GTEx portal were used to investigate the eQTL of pleiotropic variants in Korea4K.

Investigation of potential causal relationships between traits based on Mendelian randomization (MR)

We used the Mendelian randomization method to investigate potential causal relationships among 1,332 combinations of an exposure trait and an outcome trait among 37 clinical traits. MR is computed from the linear regression analysis between the effects of SNPs on an exposure trait and their effects on an outcome trait. We chose the SNPs with suggestive WGWAS results (*P*-value < 10^{-5}) with exposure traits as the instrument variables. In case multiple SNPs existed in the LD block, the one with the smallest *P*-value was chosen. We rejected 40 SNPs, which were detected as outliers of linear regression from MR-PRESSO software (1.0) [38] with 'NbDistribution=10000 and SignifThreshold=0.05' options, from further analysis. MR coefficients were computed using the chosen SNPs by three different methods: the Inverse-variance weighted (IVW) and MR-Egger method of TwoSampleMR package (v.0.5.6) [39] and MR-PRESSO software (1.0) [38]. Finally, we selected 36 significant causal relationships that overlapped at least two of three methods (IVW, MR_Egger, and MRPRESSO). All analyses were performed with default options.

Acknowledgments

We appreciate all participants and Ulsan citizens who supported this project. We also thank Ju Yeon Park, Sangryoul Han, Jungae Shim, Nayoung Kim, Seung Gu Park, Byoung-Chul Kim, Jungeun Kim, Neung-hwa Park, Suan Cho, and Yeshin Park for supporting this project. This work was supported by Biodatafarm computing infrastructure funded by the Ulsan metropolitan city government. We thank the Korea Institute of Science and Technology Information (KISTI) for providing us with the Korea Research Environment Open NETwork (KREONET). We thank our collaborators in NCSRD, KRISS, and C.-G. Kim. The Genotype-Tissue Expression (GTEx) Project was supported by the Common Fund of the Office of the Director of the National Institutes of Health, and by NCI, NHGRI, NHLBI, NIDA, NIMH, and NINDS. The data used for the analyses described in this manuscript were obtained from the GTEx Portal on 11/25/2021. This work was supported by the U-K BRAND Research Fund (1.200108.01) of UNIST (Ulsan National Institute of Science & Technology). This work was supported by the Research Project Funded by Ulsan City Research Fund (1.200047.01) of UNIST. This work was supported by the Promotion of Innovative Businesses for Regulation-Free Special Zones funded by the Ministry of SMEs and Startups (MSS, Korea) (1425157301 and 1425156792). This work was also supported by the Establishment of Demonstration Infrastructure for Regulation-Free Special Zones funded by the Ministry of SMEs and Startups (MSS, Korea) (1425157253). This research was also supported by the Technology Innovation Program (20016225, Development and Dissemination on National Standard Reference Data) funded by the Ministry of Trade, Industry & Energy (MOTIE, Korea). The biospecimens for this study were provided by Ulsan Medical Center and the Biobanks of Gyeongsang National University Hospital, Chungbuk National University Hospital (18-27, 20-04), and Kyung Hee University Hospital (2018-4, 2019-4, 2019-6), the members of the National Biobank of Korea, which is supported by the Ministry of Health, Welfare and Family Affairs. 215 whole-genome-seq data used in this study were provided by the Clinical & Omics Data Archive (CODA; http://coda.nih.go.kr), CODA accession number S000680. We thank Jaesu Bhak and Maryana Bhak for editing the manuscript.

Author contributions

S.J., Hansol C., Y. J., W.C. and Jong B. wrote the manuscript. S.J., Hansol C., Y. J., Hyunjoo C., K.A., H. R., Jihun. B., H. L., Yoonsung. K., S. H., C.L., and J. S. conducted the data analysis. C. K., Yeonkyung K., Younghui K., and Y. J. W. performed wet-lab experiments. S. J., Yeo Jin K., B. C. K., S.L., and Jong B. designed the study. S.J., Hansol C., Y.J. W.C., A.B., C.Y., D. B., O. B., E. S., S. K., J. P., J. J., D. J., S. L., and Jong B. revised the manuscript. S. L. and Jong B. jointly supervised the study.

Ethics, consents and permissions

Sample collection and sequencing were approved by the Institutional Review Board (IRB) of the Ulsan National Institute of Science and Technology (UNISTIRB-15-19-A and UNISTIRB-16-13-C). Informed consent was obtained from all individuals for their participation in the Korean Ulsan genome project.

Competing interests

S.J., Y. J., H. R., Y.J.K., C.K, Yeonkyung K., Younghui K., Y. J. W., and B. C. K. are employees and Jong B. is the CEO of Clinomics Inc. The authors declare no other competing interests.

Data and materials availability

Allele frequency information of variants is publicly available under http://koreangenome.org. Raw sequencing data, individual genotype information, and clinical trait data will be as easily and freely available as possible upon request and after approval from the Korean Genomics Center's review board in UNIST. Information about the Korean Genome Project and other data sharing can be found at http://koreangenome.org.

Additional Files

Supplementary Fig. S1. Variants batch effect of DNA sequences.

Supplementary Fig. S2. Variants distribution based on variant location and allele frequency category in Korea4K.

Supplementary Fig. S3. Power comparison of whole-genome-wide association study between Korea4K and Korea1K.

Supplementary Fig. S4. Mapping depth distribution of Korea4K genomes.

Supplementary Fig. S5. QQplots for the whole-genome-wide association tests of the traits on the anthropometry category.

Supplementary Fig. S6. QQplots for the whole-genome-wide association tests of the traits on blood circulation biochemical category.

Supplementary Fig. S7. QQplots for the whole-genome-wide association tests of the traits on blood circulation physics category.

Supplementary Fig. S8. QQplots for the whole-genome-wide association tests of the traits on diabetes category.

Supplementary Fig. S9. QQplots for the whole-genome-wide association tests of the traits on electrolyte category.

Supplementary Fig. S10. QQplots for the whole-genome-wide association tests of the traits on hearing test category.

Supplementary Fig. S11. QQplots for the whole-genome-wide association tests of the traits on hematological category.

Supplementary Fig. S12. QQplots for the whole-genome-wide association tests of the traits on hepatitis category.

Supplementary Fig. S13. QQplots for the whole-genome-wide association tests of the traits on inflammation and etc category.

Supplementary Fig. S14. QQplots for the whole-genome-wide association tests of the traits on kidney function category.

Supplementary Fig. S15. QQplots for the whole-genome-wide association tests of the traits on liver function category.

Supplementary Fig. S16. QQplots for the whole-genome-wide association tests of the traits on pulmonary function category.

Supplementary Fig. S17. QQplots for the whole-genome-wide association tests of the traits on thyroid function category.

Supplementary Fig. S18. QQplots for the whole-genome-wide association tests of the traits on tumor biomarker category.

Supplementary Fig. S19. QQplots for the whole-genome-wide association tests of the traits on urinalysis category.

Supplementary Fig. S20. QQplots for the whole-genome-wide association tests of the traits on vision category.

Supplementary Table S1. Sample count in Korea4K.

Supplementary Table S2. Variant count in Korea4K before sample and variant filtering.

Supplementary Table S3. Variant count based on variant categories and reported to dbSNP.

Supplementary Table S4. Allele frequency information of populations for 62 Korean-specific missense variants

Supplementary Table S5. Prediction of changes in protein thermodynamic stability according to missense variant

Supplementary Table S6. List of the GWAS variants which have association significance *P*< 5E-8

Supplementary Table S7. Genetic heritability measurement

Supplementary Table S8. Genetic correlation measurement

Supplementary Table S9. Phenotypic correlation estimation

Supplementary Table S10. Pleiotropic variants

Supplementary Table S11. Mendelian randomization results

Supplementary Table S12. Statistics of clinical information

References

1. Song SO, Jung CH, Song YD, Park CY, Kwon HS, Cha BS, et al. Background and data configuration process of a nationwide population-based study using the korean national health insurance system. Diabetes Metab J. 2014;38 5:395-403. doi:10.4093/dmj.2014.38.5.395.
2. Jeon S, Bhak Y, Choi Y, Jeon Y, Kim S, Jang J, et al. Korean Genome Project: 1094 Korean personal genomes with clinical information. Sci Adv. 2020;6 22 doi:ARTN eaaz7835
10.1126/sciadv.aaz7835.
3. Kanai M, Akiyama M, Takahashi A, Matoba N, Momozawa Y, Ikeda M, et al. Genetic analysis of quantitative traits in the Japanese population links cell types to complex human diseases. Nat Genet. 2018;50 3:390-+. doi:10.1038/s41588-018-0047-6.
4. Choe EK, Shivakumar M, Verma A, Verma SS, Choi SH, Kim JS, et al. Leveraging deep phenotyping from health check-up cohort with 10,000 Korean individuals for phenome-wide association study of 136 traits. Sci Rep-Uk. 2022;12 1 doi:ARTN 1930
10.1038/s41598-021-04580-2.

5. Van Hout CV, Tachmazidou I, Backman JD, Hoffman JD, Liu D, Pandey AK, et al. Exome sequencing and characterization of 49,960 individuals in the UK Biobank. Nature. 2020;586 7831:749-+. doi:10.1038/s41586-020-2853-0.

6. Jiang LD, Zheng ZL, Qi T, Kemper KE, Wray NR, Visscher PM, et al. A resource-efficient tool for mixed model association analysis of large-scale data. Nat Genet. 2019;51 12:1749-+. doi:10.1038/s41588-019-0530-8.

7. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. Nature. 2021;596 7873:583-+. doi:10.1038/s41586-021-03819-2.

8. Buniello A, MacArthur JAL, Cerezo M, Harris LW, Hayhurst J, Malangone C, et al. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. Nucleic Acids Res. 2019;47 D1:D1005-D12. doi:10.1093/nar/gky1120.

9. Seyed Khoei N, Jenab M, Murphy N, Banbury BL, Carreras-Torres R, Viallon V, et al. Circulating bilirubin levels and risk of colorectal cancer: serological and Mendelian randomization analyses. BMC Med. 2020;18 1:229. doi:10.1186/s12916-020-01703-w.

10. Chang MH, Yesupriya A, Ned RM, Mueller PW and Dowling NF. Genetic variants associated with fasting blood lipids in the U.S. population: Third National Health and Nutrition Examination Survey. BMC Med Genet. 2010;11:62. doi:10.1186/1471-2350-11-62.

11. Canela-Xandri O, Rawlik K and Tenesa A. An atlas of genetic associations in UK Biobank. Nat Genet. 2018;50 11:1593-9. doi:10.1038/s41588-018-0248-z.

12. Guo S, Lv HT, Yan L and Rong FN. Hyperamylasemia may indicate the presence of ovarian carcinoma A case report. Medicine. 2018;97 49 doi:ARTN e13520
10.1097/MD.0000000000013520.

13. Shintani D, Yoshida H, Imai Y and Fujiwara K. Acute pancreatitis induced by paclitaxel and carboplatin therapy in an ovarian cancer patient. Eur J Gynaecol Oncol. 2016;37 2:286-7.

14. Zakrzewska I and Pietryńczak M. The activity of alpha-amylase and its salivary isoenzymes in serum and urine of patients with neoplastic diseases of female reproductive organs. Roczniki Akademii Medycznej w Bialymstoku (1995). 1996;41 2:492-8.

15. Forouhi NG, Sattar N and McKeigue PM. Relation of C-reactive protein to body fat distribution and features of the metabolic syndrome in Europeans and South Asians. Int J Obes Relat Metab Disord. 2001;25 9:1327-31. doi:10.1038/sj.ijo.0801723.

16. Lim S, Jang HC, Lee HK, Kimm KC, Park C and Cho NH. The relationship between body fat and C-reactive protein in middle-aged Korean population. Atherosclerosis. 2006;184 1:171-7. doi:10.1016/j.atherosclerosis.2005.04.003.

17. Aabo K, Pedersen H and Kjaer M. Carcinoembryonic antigen (CEA) and alkaline phosphatase in progressive colorectal cancer with special reference to patient survival. Eur J Cancer Clin Oncol. 1986;22 2:211-7. doi:10.1016/0277-5379(86)90033-7.

18. Tartter PI, Slater G, Gelernt I and Aufses AH, Jr. Screening for liver metastases from colorectal cancer with carcinoembryonic antigen and alkaline phosphatase. Ann Surg. 1981;193 3:357-60. doi:10.1097/00000658-198103000-00019.

19. Walach N, Guterman A, Zaidman JL, Kaufman S and Scharf S. Leukocyte alkaline phosphatase and carcinoembryonic antigen in breast cancer patients: clinical correlation with the markers. J Surg Oncol. 1989;40 2:85-7. doi:10.1002/jso.2930400205.

20. Yang JA, Lee SH, Goddard ME and Visscher PM. GCTA: A Tool for Genome-wide Complex Trait Analysis. Am J Hum Genet. 2011;88 1:76-82. doi:10.1016/j.ajhg.2010.11.011.

21. Zhang YL, Cheng YS, Jiang W, Ye YX, Lu QS and Zhao HY. Comparison of methods for estimating genetic correlation between complex traits using GWAS summary statistics. Brief Bioinform. 2021;22 5 doi:ARTN bbaa442
10.1093/bib/bbaa442.

22. Visscher PM, Hemani G, Vinkhuyzen AA, Chen GB, Lee SH, Wray NR, et al. Statistical power to detect genetic (co)variance of complex traits using SNP data in unrelated samples. PLoS Genet. 2014;10 4:e1004269. doi:10.1371/journal.pgen.1004269.

23. Li J, Gui L, Wu C, He Y, Zhou L, Guo H, et al. Genome-wide association study on serum alkaline phosphatase levels in a Chinese population. BMC Genomics. 2013;14:684. doi:10.1186/1471-2164-14-684.

24. Middelberg RP, Ferreira MA, Henders AK, Heath AC, Madden PA, Montgomery GW, et al. Genetic variants in LPL, OASL and TOMM40/APOE-C1-C2-C4 genes are associated with multiple cardiovascular-related traits. BMC Med Genet. 2011;12:123. doi:10.1186/1471-2350-12-123.

25. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet journal. 2011;17 1:10-2.

26. Li H and Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25 14:1754-60. doi:10.1093/bioinformatics/btp324.

27. Poplin R, Ruano-Rubio V, DePristo MA, Fennell TJ, Carneiro MO, Van der Auwera GA, et al. Scaling accurate genetic variant discovery to tens of thousands of samples. BioRxiv. 2018:201178.

28. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GR, Thormann A, et al. The ensembl variant effect predictor. Genome biology. 2016;17 1:1-14.

29. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: A tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007;81 3:559-75. doi:10.1086/519795.

30. 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. doi:10.1038/nature09534.

31. Cunningham F, Allen JE, Allen J, Alvarez-Jarreta J, Amode MR, Armean IM, et al. Ensembl 2022. Nucleic Acids Res. 2022;50 D1:D988-D95. doi:10.1093/nar/gkab1049.

32. Blum M, Chang HY, Chuguransky S, Grego T, Kandasaamy S, Mitchell A, et al. The InterPro protein families and domains database: 20 years on. Nucleic Acids Res. 2021;49 D1:D344-D54. doi:10.1093/nar/gkaa977.

33. Delgado J, Radusky LG, Cianferoni D and Serrano L. FoldX 5.0: working with RNA, small molecules and a new graphical interface. Bioinformatics. 2019;35 20:4168-9. doi:10.1093/bioinformatics/btz184.

34. 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. doi:10.1038/ng.3643.

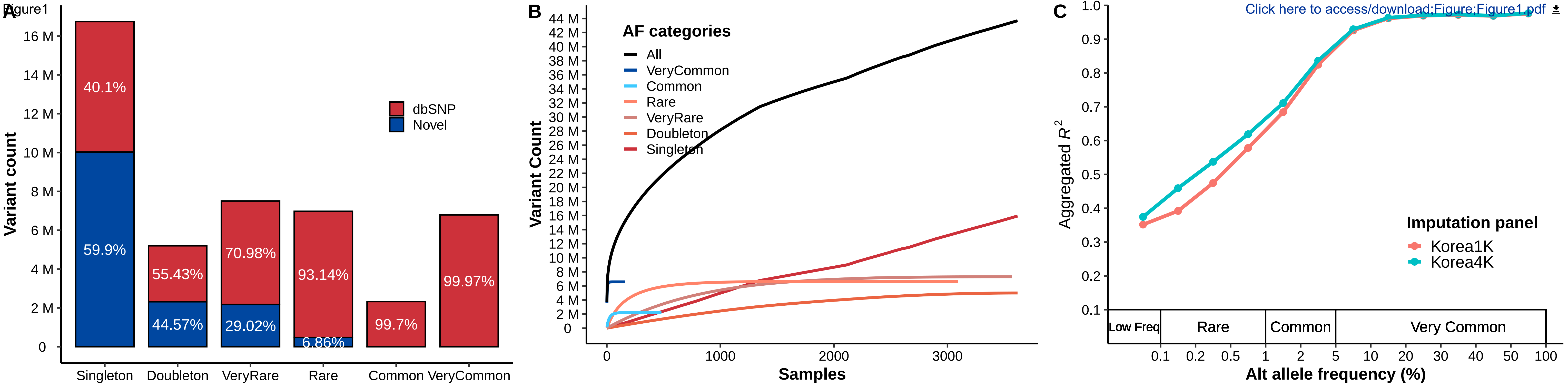
35. Das S, Forer L, Schonherr S, Sidore C, Locke AE, Kwong A, et al. Next-generation genotype imputation service and methods. Nat Genet. 2016;48 10:1284-7. doi:10.1038/ng.3656.

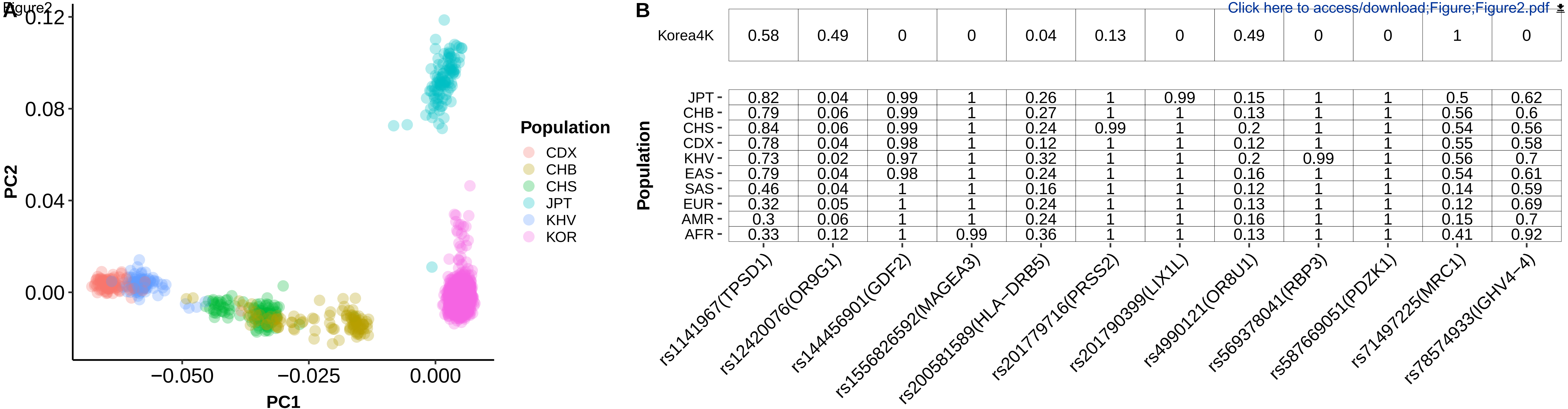
36. Chang CC, Chow CC, Tellier LCAM, Vattikuti S, Purcell SM and Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience. 2015;4 doi:ARTN 7
10.1186/s13742-015-0047-8.

37. Lee S, Yang J, Goddard M, Visscher P and Wray N. Estimation of pleiotropy between complex diseases using SNP-derived genomic relationships and restricted maximum likelihood. Bioinformatics. 2012;28 19:2540-2.

38. Verbanck M, Chen CY, Neale B and Do R. Detection of widespread horizontal pleiotropy in causal relationships inferred from Mendelian randomization between complex traits and diseases. Nat Genet. 2018;50 5:693-+. doi:10.1038/s41588-018-0099-7.

39. Hemani G, Zhengn J, Elsworth B, Wade KH, Haberland V, Baird D, et al. The MR-Base platform supports systematic causal inference across the human phenome. Elife. 2018;7 doi:ARTN e34408
10.7554/eLife.34408.


A
Variant count
16 M
14 M
12 M
10 M
8 M
6 M
4 M
2 M
0
40.1%
59.9%
55.43%
44.57%
70.98%
29.02%
93.14%
6.86%
99.7%
99.97%
dbSNP
Novel
Singleton
Doubleton
VeryRare
Rare
Common
VeryCommon
B
AF categories
All
VeryCommon
Common
Rare
VeryRare
Doubleton
Singleton
Variant Count
44 M
42 M
40 M
38 M
36 M
34 M
32 M
30 M
28 M
26 M
24 M
22 M
20 M
18 M
16 M
14 M
12 M
10 M
8 M
6 M
4 M
2 M
0
0
1000
2000
3000
Samples
C
Aggregated R^2
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
Imputation panel
Korea1K
Korea4K
Low Freq
Rare
Common
Very Common
0.1
0.2
0.5
1
2
5
10
20
30
40
50
100
Alt allele frequency (%)



	rs1141967(TPSD1)	rs12420076(OR9G1)	rs14456901(GDF2)	rs155826592(MAGEA3)	rs200581589(HLA−DRB5)	rs201779716(PRSS2)	rs201790399(LIX1L)	rs4990121(OR8U1)	rs569378041(RBP3)	rs587669051(PDZK1)	rs71497225(MRC1)	rs78574933(IGHV4−4)
Korea4K	0.58	0.49	0	0	0.04	0.13	0	0.49	0	0	1	0
JPT	0.82	0.04	0.99	1	0.26	1	0.99	0.15	1	1	0.5	0.62
CHB	0.79	0.06	0.99	1	0.27	1	1	0.13	1	1	0.56	0.6
CHS	0.84	0.06	0.99	1	0.24	0.99	1	0.2	1	1	0.54	0.56
CDX	0.78	0.04	0.98	1	0.12	1	1	0.12	1	1	0.55	0.58
KHV	0.73	0.02	0.97	1	0.32	1	1	0.2	0.99	1	0.56	0.7
EAS	0.79	0.04	0.98	1	0.24	1	1	0.16	1	1	0.54	0.61
SAS	0.46	0.04	1	1	0.16	1	1	0.12	1	1	0.14	0.59
EUR	0.32	0.05	1	1	0.24	1	1	0.13	1	1	0.12	0.69
AMR	0.3	0.06	1	1	0.24	1	1	0.16	1	1	0.15	0.7
AFR	0.33	0.12	1	0.99	0.36	1	1	0.13	1	1	0.41	0.92

Figure3

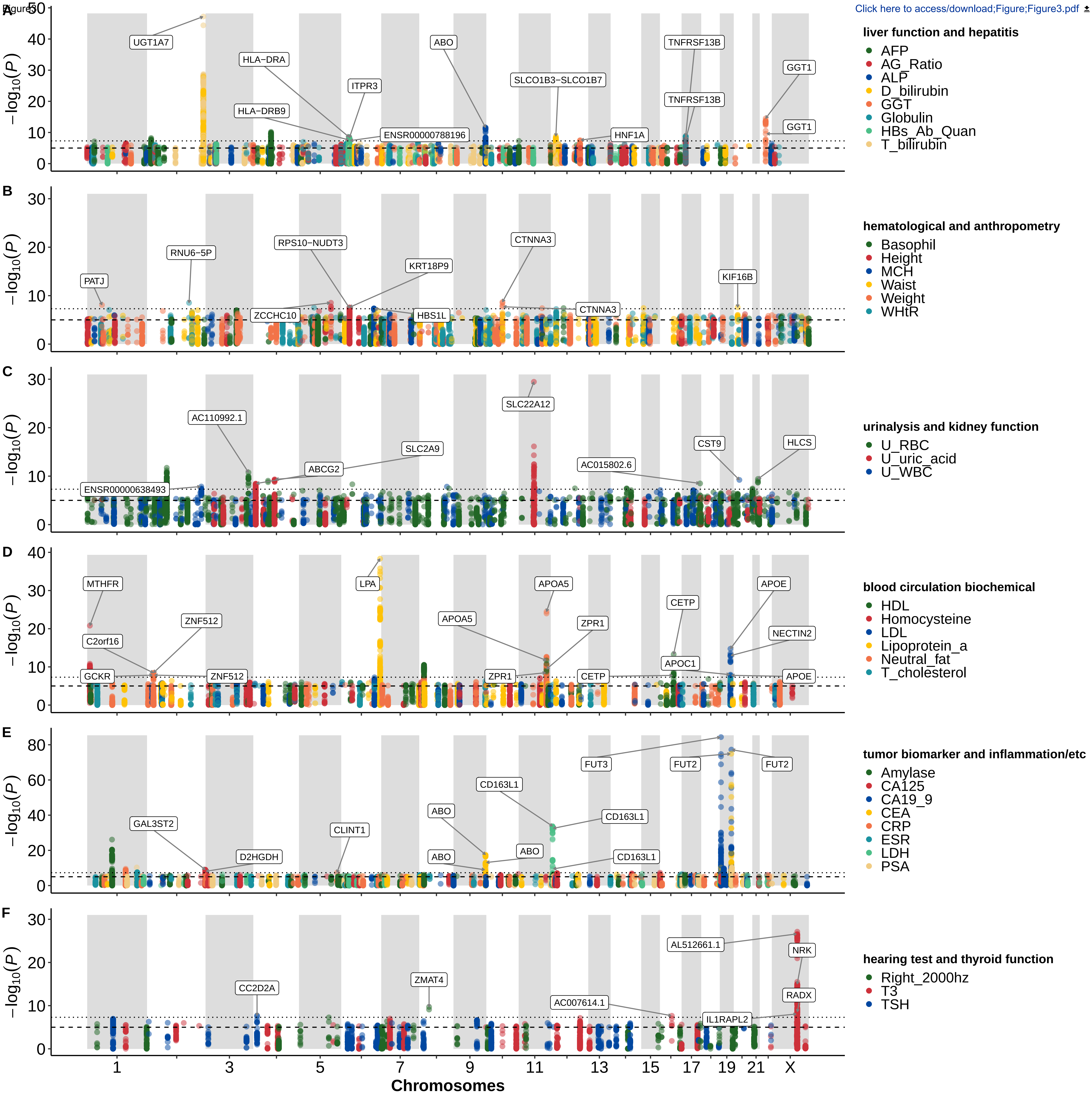

A
B
C
D
E
F
−log10(P)
0
10
20
30
40
50
60
80
Chromosomes
1
3
5
7
9
11
13
15
17
19
21
X
liver function and hepatitis
● AFP
● AG_Ratio
● ALP
● D_bilirubin
● GGT
● Globulin
● HBs_Ab_Quan
● T_bilirubin
UGT1A7
HLA−DRA
HLA−DRB9
ITPR3
ENSR00000788196
ABO
SLCO1B3−SLCO1B7
HNF1A
TNFRSF13B
TNFRSF13B
GGT1
GGT1
hematological and anthropometry
● Basophil
● Height
● MCH
● Waist
● Weight
● WHtR
PATJ
RNU6−5P
RPS10−NUDT3
KRT18P9
ZCCHC10
HBS1L
CTNNA3
CTNNA3
KIF16B
urinalysis and kidney function
● U_RBC
● U_uric_acid
● U_WBC
ENSR00000638493
AC110992.1
ABCG2
SLC2A9
SLC22A12
AC015802.6
CST9
HLCS
blood circulation biochemical
● HDL
● Homocysteine
● LDL
● Lipoprotein_a
● Neutral_fat
● T_cholesterol
MTHFR
C2orf16
GCKR
ZNF512
ZNF512
LPA
APOA5
APOA5
ZPR1
ZPR1
CETP
CETP
APOC1
APOE
NECTIN2
APOE
tumor biomarker and inflammation/etc
● Amylase
● CA125
● CA19_9
● CEA
● CRP
● ESR
● LDH
● PSA
GAL3ST2
D2HGDH
CLINT1
ABO
ABO
ABO
CD163L1
CD163L1
CD163L1
FUT3
FUT2
FUT2
hearing test and thyroid function
● Right_2000hz
● T3
● TSH
CC2D2A
ZMAT4
AC007614.1
AL512661.1
NRK
RADX
IL1RAPL2

Figure4

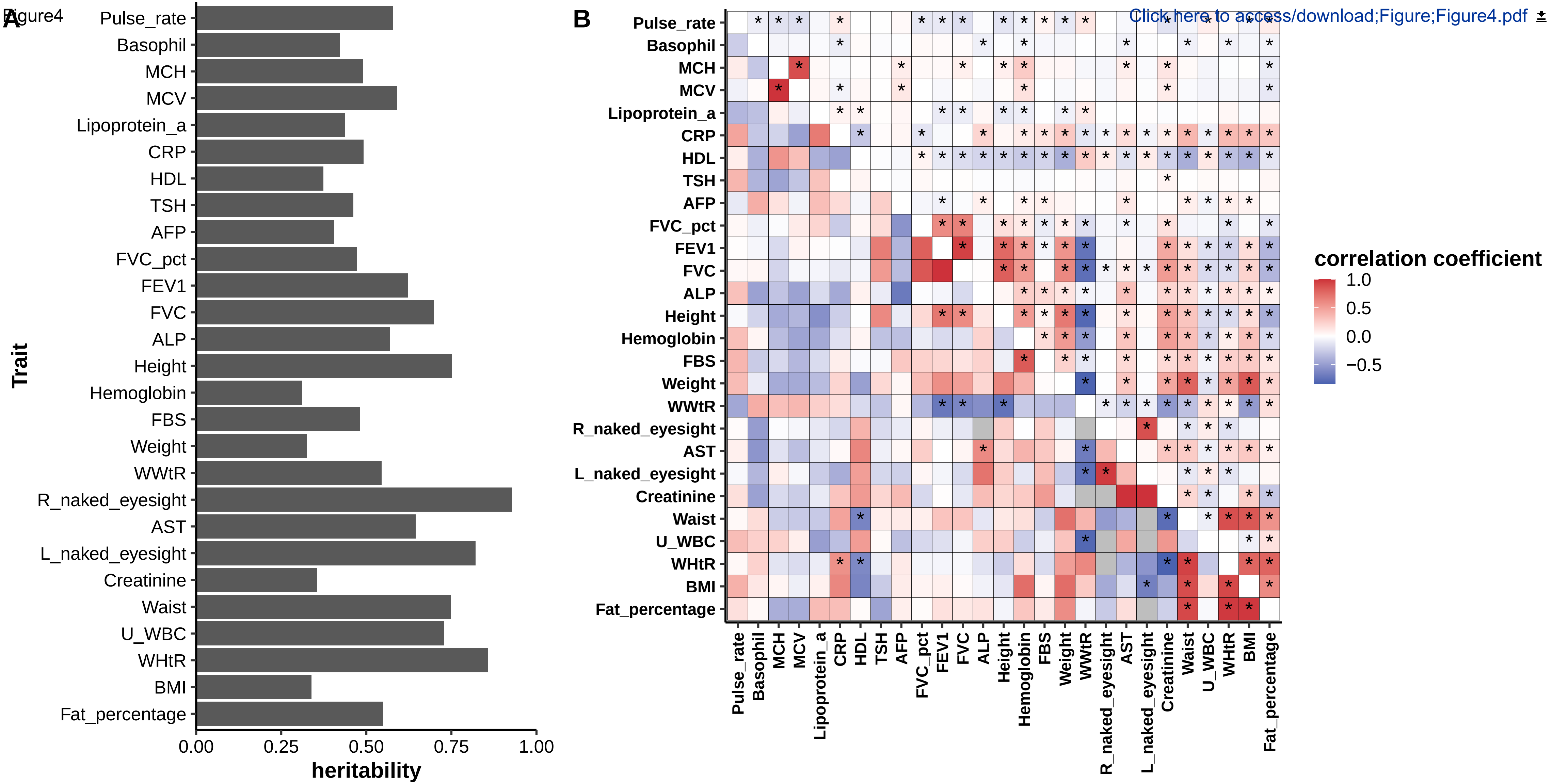
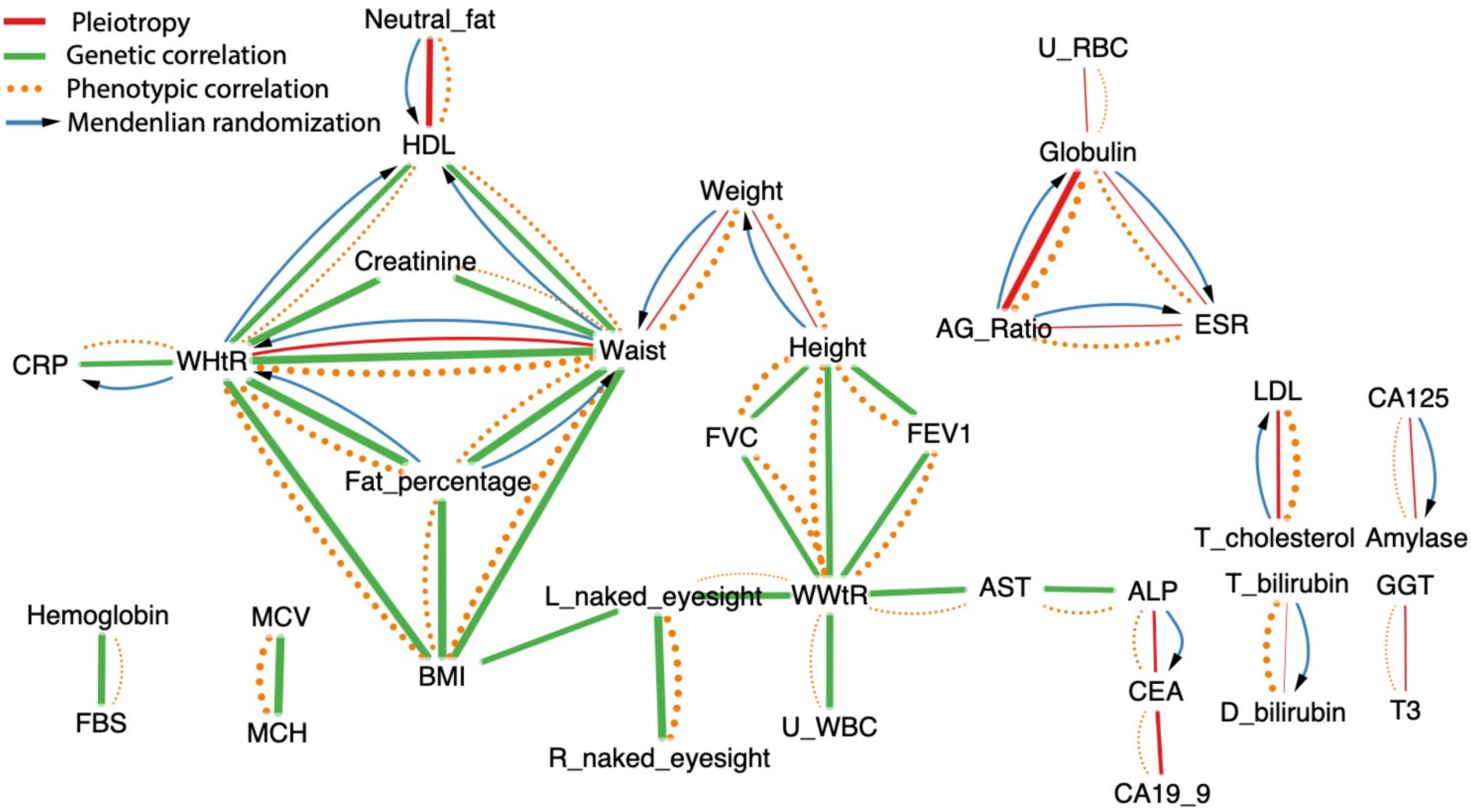

A
Pulse_rate
Basophil
MCH
MCV
Lipoprotein_a
CRP
HDL
TSH
AFP
FVC_pct
FEV1
FVC
ALP
Height
Hemoglobin
FBS
Weight
WWtR
R_naked_eyesight
AST
L_naked_eyesight
Creatinine
Waist
U_WBC
WHtR
BMI
Fat_percentage
Trait
0.00
0.25
0.50
0.75
1.00
heritability
B
Pulse_rate
Basophil
MCH
MCV
Lipoprotein_a
CRP
HDL
TSH
AFP
FVC_pct
FEV1
FVC
ALP
Height
Hemoglobin
FBS
Weight
WWtR
R_naked_eyesight
AST
L_naked_eyesight
Creatinine
Waist
U_WBC
WHtR
BMI
Fat_percentage
correlation coefficient
1.0
0.5
0.0
−0.5

Figure5


Pleiotropy
Genetic correlation
Phenotypic correlation
Mendenlian randomization
Neutral_fat
HDL
U_RBC
Globulin
Weight
Creatinine
AG_Ratio
ESR
CRP
WHtR
Waist
Height
LDL
CA125
FVC
FEV1
Fat_percentage
T_cholesterol
Amylase
AST
ALP
T_bilirubin
GGT
L_naked_eyesight
WWtR
Hemoglobin
MCV
BMI
CEA
D_bilirubin
T3
FBS
MCH
U_WBC
R_naked_eyesight
CA19_9

Supplementary Table S1

Supplementary Table S2

Supplementary Table S3

Supplementary Table S4

Supplementary Table S5

Supplementary Table S6

Supplementary Table S7

Supplementary Table S8

Supplementary Table S9

Supplementary Table S10

Supplementary Table S11

Supplementary Table S12

Supplementary Figures

Dear Editor,

We would like to submit our manuscript entitled “**Korea4K: whole genome sequences of 4,157 Koreans with 107 phenotypes derived from extensive health check-ups**” as a *Research Article* in *GigaScience*.

Since we reported Korea1K (1,094 Korean genomes with 79 clinical traits) in 2020 (Jeon et al., *Sci Adv*. 2020), we have pursued a more comprehensive study based on a larger cohort of Koreans (4,157 whole genomes with 107 clinical traits) as the second phase of the Korean Genome Project (KGP).

Here, we found that only around 4,000 whole genomes (Korea4K) were sufficient to cover the genomic diversity of the Korean population with East Asian ancestry by analyzing the statistics of common and rare SNP variants. We also present the Korea4K variome database as a part of the KGP, which could be a resource for a large-scale population genomics analysis of diverse ethnic groups in association with human evolution and diseases.

The major difference between Korea1K and Korear4K is not only in the sample size but also in the number of clinical traits derived from extensively curated reports covering the most common health check-up parameters. With the greater number of samples and clinical traits, we were able to identify 1,356 new associations between genotypes and phenotypes, which had not been detected in Korea1K. Furthermore, we performed genetic correlation, pleiotropy, and Mendelian randomization analyses to map the variome with the clinical traits from common health check-ups. We also confirmed that Korea4K, compared to Korea1K, could improve quality as a reference panel for genotype imputation.

As our study provides a possibly useful resource for exploring the relationship between the genome and the phenome, and the variome data will be publicly available as open as possible, we believe that this manuscript fits the scope of *GigaScience*.

All study participants provided informed consent, and the study design was approved by the appropriate ethics review board.

S.J., Y. J., H. R., Y.J.K., C.K, Yeonkyung K., Younghui K., Y. J. W., and B. C. K. are employees and Jong B. is the CEO of Clinomics Inc. The authors declare no other competing interests.

We confirm that all authors have approved the manuscript for submission and the content of the manuscript has not been published, or submitted for publication elsewhere.

We would like to suggest the following reviewers:

□ Tim Hubbard, Ph.D., Professor of Bioinformatics and Head of Department, Department of Medical & Molecular Genetics at King’s College London,
tim.hubbard@kcl.ac.uk
□ Masao Nagasaki, Ph.D., Professor at the Center for Genomic Medicine, Kyoto University,
nagasaki@genome.med.kyoto-u.ac.jp

Thank you for your consideration.

Sincerely,
Jong Bhak, Ph.D.

Korean Genomics Center
Ulsan National Institute of Science and Technology
Ulsan 44919, Republic of Korea
Email: jongbhak@genomics.org
Tel: +82 (0)10 4644 6754