Science Advances
AAAS

advances.sciencemag.org/cgi/content/full/6/22/eaaz7835/DC1

Supplementary Materials for

Korean Genome Project: 1094 Korean personal genomes with clinical information

Sungwon Jeon, Youngjune Bhak, Yeonsong Choi, Yeonsu Jeon, Seunghoon Kim, Jaeyoung Jang, Jinho Jang, Asta Blazyte, Changjae Kim, Yeonkyung Kim, Jungae Shim, Nayeong Kim, Yeo Jin Kim, Seung Gu Park, Jungeun Kim, Yun Sung Cho, Yeshin Park, Hak-Min Kim, Byoung-Chul Kim, Neung-Hwa Park, Eun-Seok Shin, Byung Chul Kim, Dan Bolser, Andrea Manica, Jeremy S. Edwards, George Church*, Semin Lee*, Jong Bhak*

*Corresponding author. Email: gchurch@genetics.med.harvard.edu (G.C.); seminlee@unist.ac.kr (S.L.); jongbhak@genomics.org (J.B.)

Published 27 May 2020, *Sci. Adv.* **6**, eaaz7835 (2020)
DOI: 10.1126/sciadv.aaz7835

The PDF file includes:

Figs. S1 to S34
Tables S1 to S3
References

Other Supplementary Material for this manuscript includes the following:

(available at advances.sciencemag.org/cgi/content/full/6/22/eaaz7835/DC1)

Data S1 to S5

Supplementary materials and methods

Calling of CNVs

Copy number variations (CNVs) were identified via CNVnator (*56*) with default parameters and a 100-bp bin size from 1,094 samples. Thereafter, we excluded 23 samples in accordance with the following criteria: the total number of CNV exceeds one standard deviation (SD) from the average count of CNVs per sample (average CNV count: 525, SD of CNV count: 129). Reliable CNVnator calls were filtered in accordance with the following criteria:

1) e-values (e-val1, e-val2, e-val3, and e-val4) are less than 10^{-5}
2) q0 <0.5 (q0 is the fraction of reads mapped with zero quality)
3) Gap and centromere regions from UCSC hg38 data were filtered out.
4) For deletion calls, only those with <0.75 of normalized read depth*(1+q0) were used.
5) If the bases in the called region contained more than 90% of the "N," the calls were filtered out.
6) Segmental duplication regions from UCSC hg38 data were filtered out.

In total, 6,131 CNVs were identified in Korea1K (Fig S17). As expected, since copy number variants are quite variable, more than 50% of the CNVs were categorized as very rare (sample frequency < 0.001). After individual calling, the calls with >80% reciprocally overlapped regions from each individual were combined using the igraph package (*57*) in R. The start and end positions of the representative calls were assigned to the average of the locations from the combined calls. We annotated the gene symbol of the CNV calls with Ensembl database (*58*). We then checked the overlap between our call set and 1KGP phase 3 (*59*). Only CNVs that showed more than 80% of the overlap with CNVs of 1KGP were used for further analysis. We additionally used Control-FREEC (*60*) with a window size 100 bp and a breakpoint threshold 0.6 to validate the common CNVs which contained protein-coding genes in Korea1K. We filtered out the common CNVs which have lower than 0.85 of the recovery rate of the CNV calls between the two callers.

Calling of TE insertions

TE insertions were identified in Korea1K samples using Mobile Element Locator Tool (MELT; ver. 2.1.4) (*61*), a tool to detect TE insertions in ALU, LINE1, and SVA elements using discordant read pairs to define potential TE sites and split reads to identify breakpoints and target site duplications (*59*). We filtered out TE sites with <70% and >130% of average depth of 100bp flanking regions to control for variations at candidate TE sites. The allele frequency of TE insertions was calculated as the number of presented TE insertions normalized by the total number of alleles in the population (*62*). In total, 29,143 TE insertions were identified in Korea1K (Alu: 23,915, LINE1: 3,707, SVA: 1,521) from the WGS data (Table S2). More than 50% of the TE insertions identified in Korean1K were rare variants (allele frequency <1%, 16,225 TE insertions, Fig. S19); this pattern was similar to that of SNVs and indels. Allele frequencies of TE insertions were compared between Korea1K and 26 other populations from the 1KGP phase 3 data (*59*, *62*, *63*). We compared Korea1K TE insertions with 1KGP by the genomic position and TE types. Only TE insertions with frequencies of >5% and which overlapped with 1KGP call were used to PCA. Chi-square analysis was performed for each TE insertion and TE insertions with *Q*-values of <0.05 were determined to identify TE insertions differing significantly from those in the Korea1K dataset.

HLA typing

The HLA gene complex encodes MHC proteins responsible for antigen presentation. HLA typing was carried out using OptiType (ver. 1.3.1) (*64*), which predicts information regarding HLA class 1 alleles from WGS data. Reads were mapped to HLA reference sequences in the OptiType program using BWA (*43*) (ver. 0.7.15) and all unmapped reads were filtered out using

SAMtools (ver. 1.6) (*53*). Thereafter, we run OptiType's pipeline with default parameters. A*24:02, B*44:03, and C*01:02 were the most dominant types of HLA-A, B, and C, respectively (Fig. S23). To compare frequencies from multiple populations, we downloaded an HLA allele frequency database from The Allele Frequency Net Database (*65*).

Supplementary figures

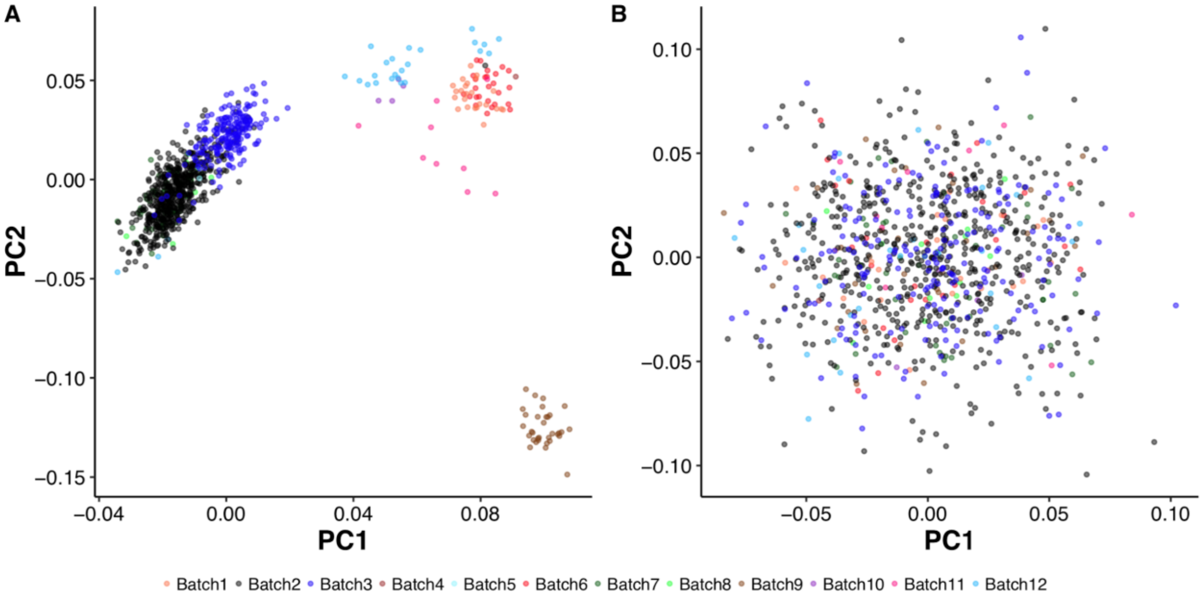


Fig. S1 Principal component analysis (PCA) plot using SNVs and Indels in Korea1K set. **(a)** before removing the batch effect **(b)** after removing the batch effect.

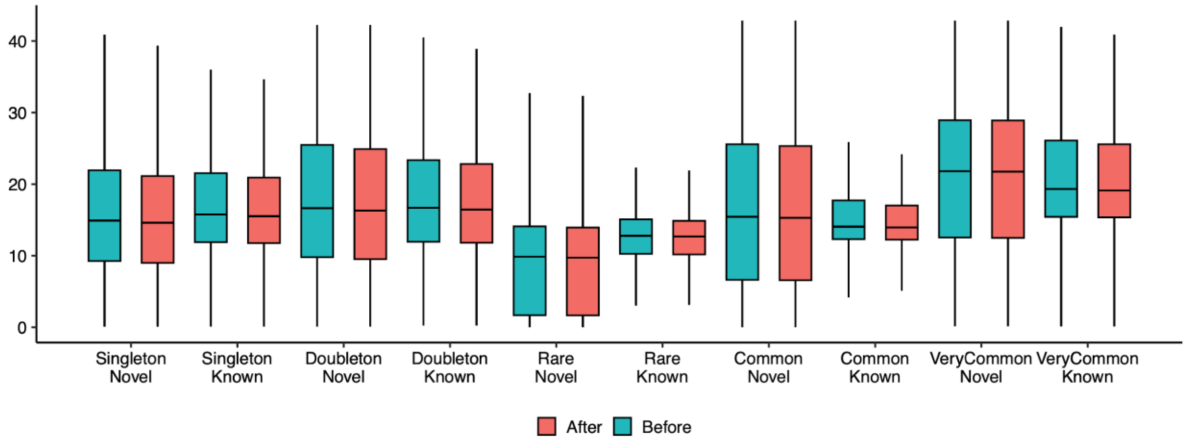


Fig. S2 Boxplot of variants quality normalized by depth based on allele frequency category and existence in dbSNP v.150 before and after batch effect filtering. (Singleton: allele count =1; Doubleton allele count =2; rare: allele count > 2 and allele frequency ≤ 0.01; common: allele frequency > 0.01 and allele frequency ≤ 0.05; very common: allele frequency > 0.05)

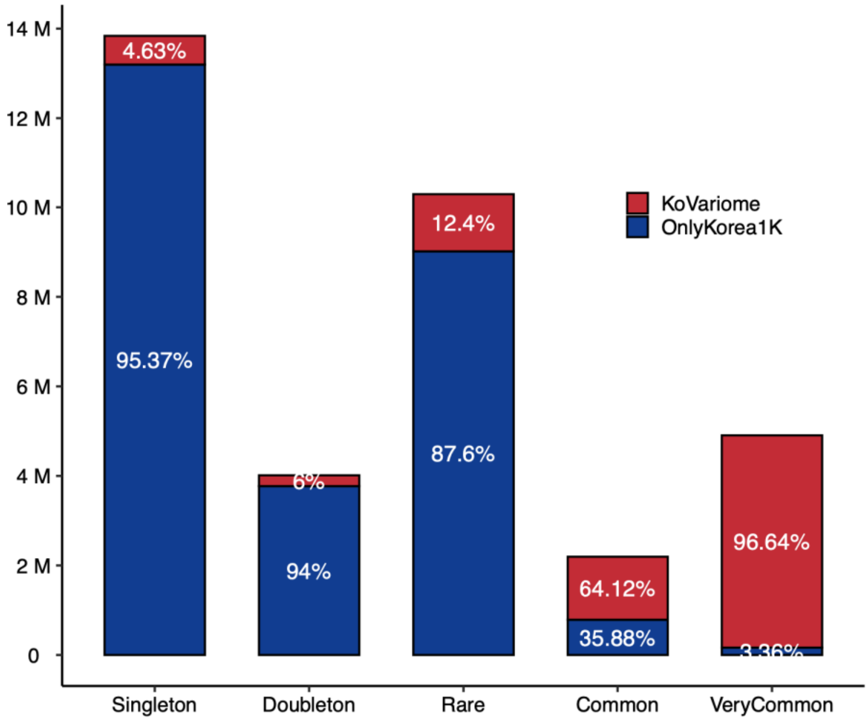


Fig. S3 Percentage of overlapped SNVs with KoVariome. Singleton: allele count =1; Doubleton allele count =2; rare: allele count > 2 and allele frequency ≤ 0.01; common: allele frequency > 0.01 and allele frequency ≤ 0.05; very common: allele frequency > 0.05

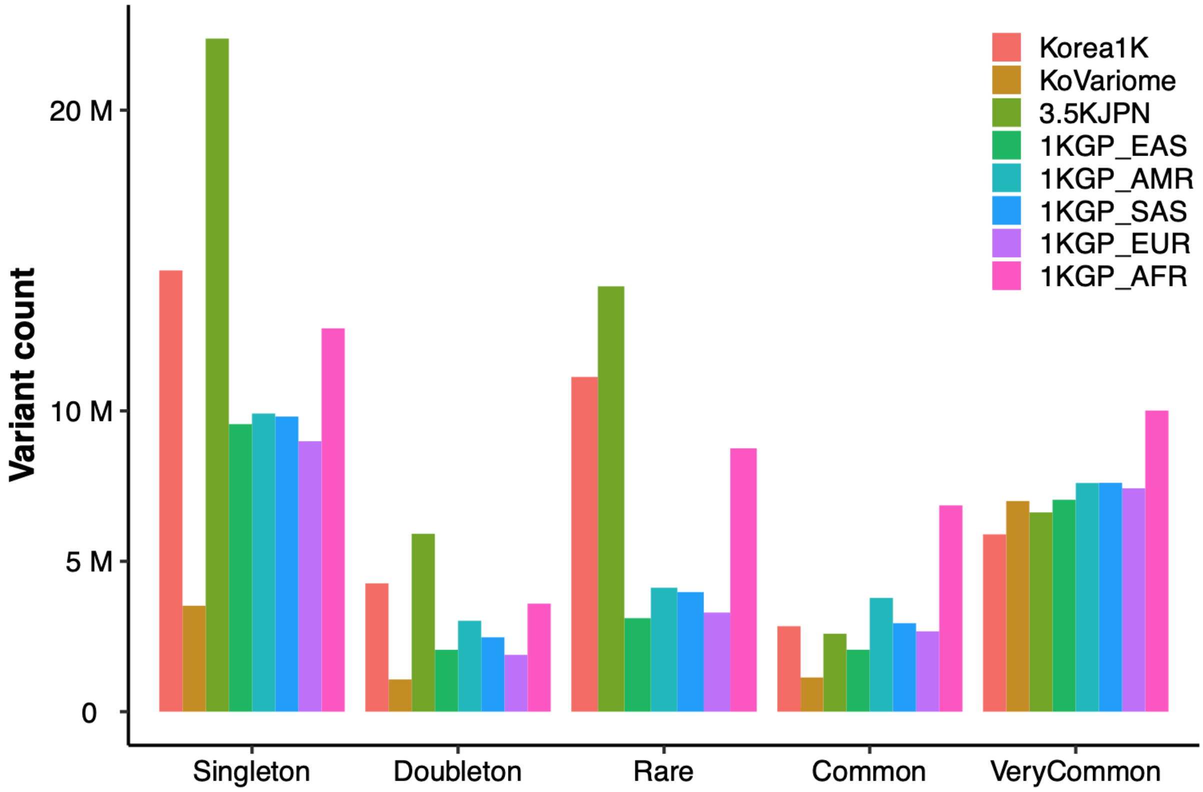


Fig. S4 Number of variants from variome databases based on allele frequencies. Singleton: allele count =1; Doubleton allele count =2; rare: allele count > 2 and allele frequency ≤ 0.01; common: allele frequency > 0.01 and allele frequency ≤ 0.05; very common: allele frequency > 0.05. The color indicates variome database. Note that there are no variants that have allele frequency ≤ 0.01 and allele count >2 for KoVariome because of the small size of samples.

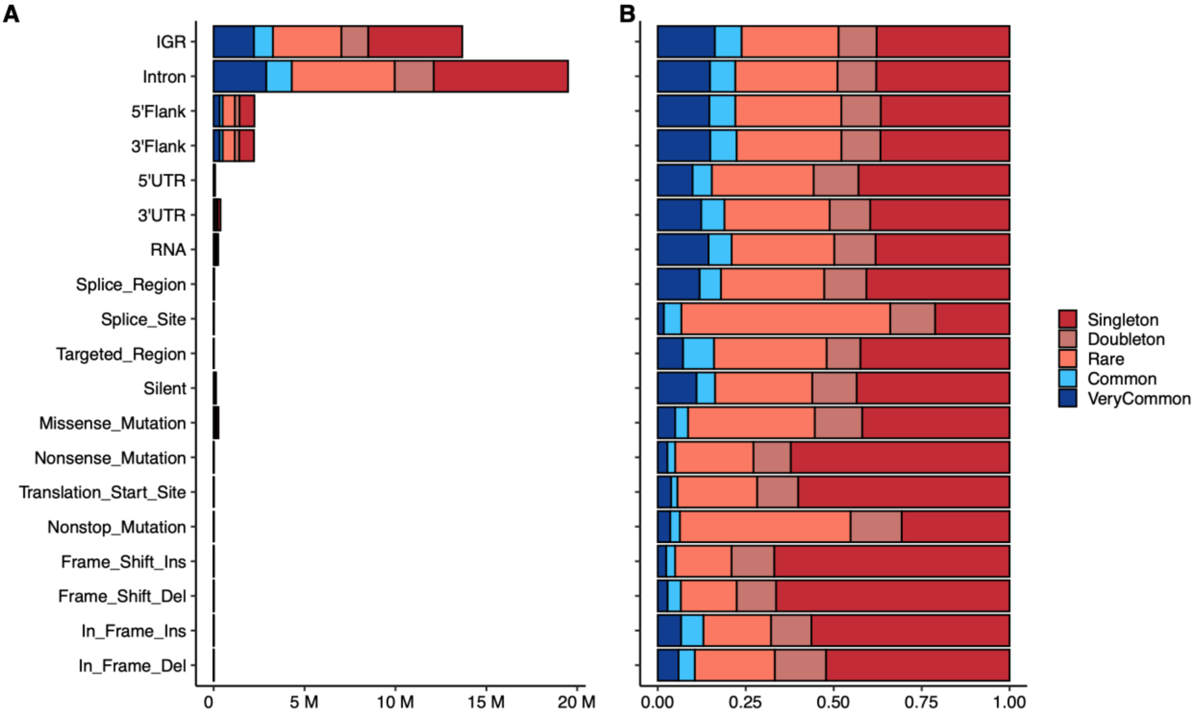


Fig. S5 Variants distribution based on variant location and allele frequency in Korea1K. **(A)** Variants counts, and **(B)** proportions of the number of variants based on allele frequency categories. IGR: inter-genic region except for 5’ and 3’ Flank variants; UTR: untranslated region. Singleton: allele count =1; Doubleton allele count =2; rare: allele count > 2 and allele frequency ≤ 0.01; common: allele frequency > 0.01 and allele frequency ≤ 0.05; very common: allele frequency > 0.05

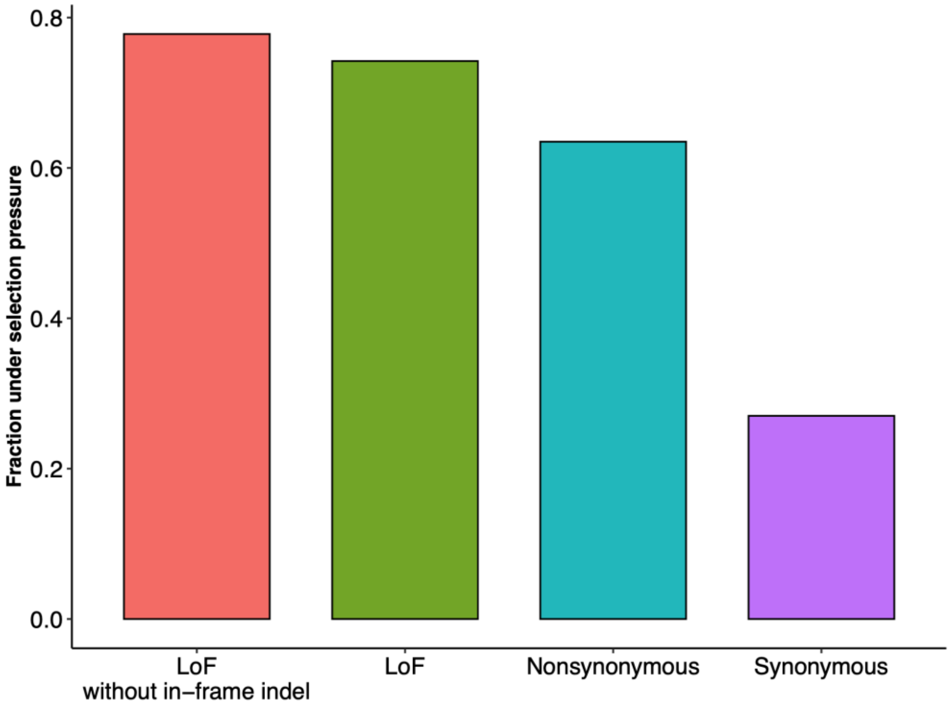


Fig. S6 Fraction under selection based on variants type. LoF indicates loss-of-function.

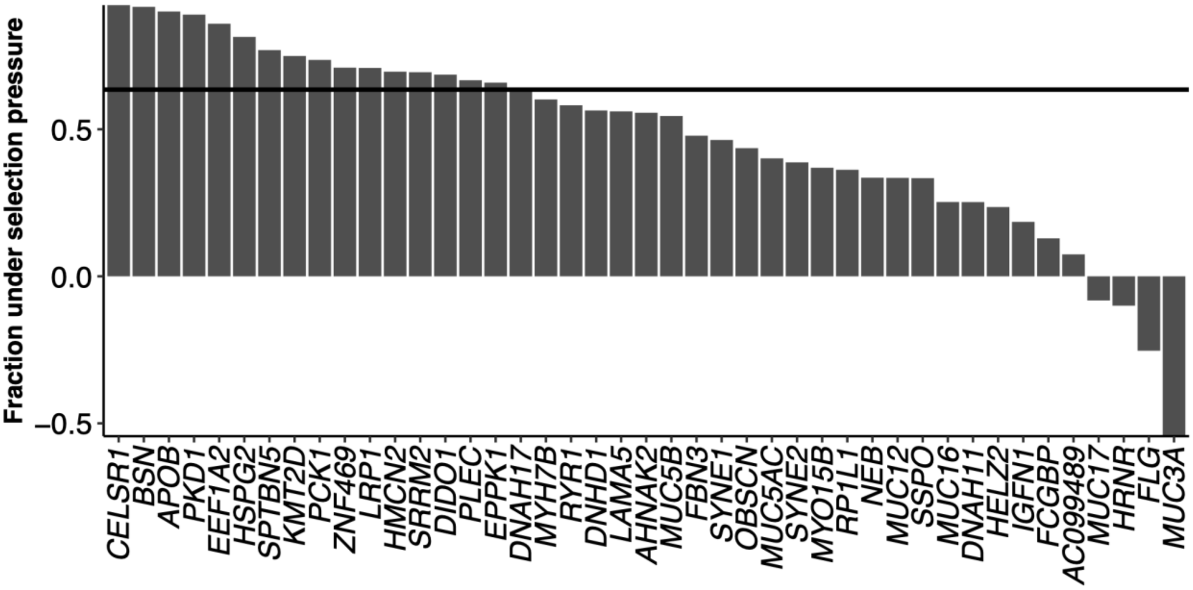


Fig. S7 Fraction under selection based on genes. The horizontal line indicates the fraction under the selection pressure of nonsynonymous variants.

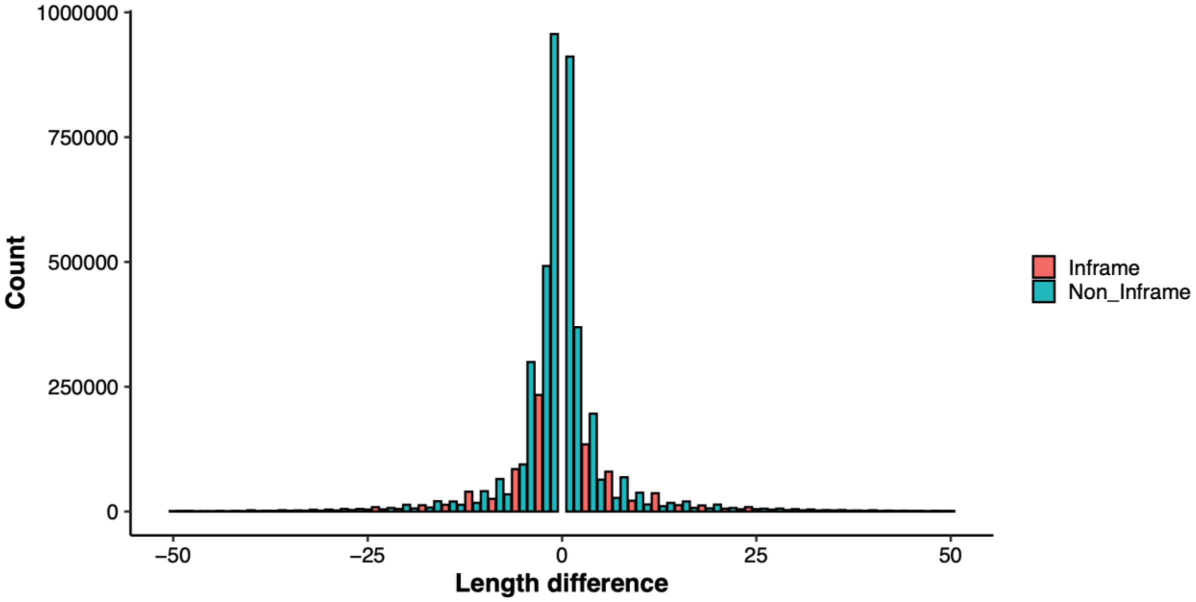


Fig. S8 Length distribution of Indels.

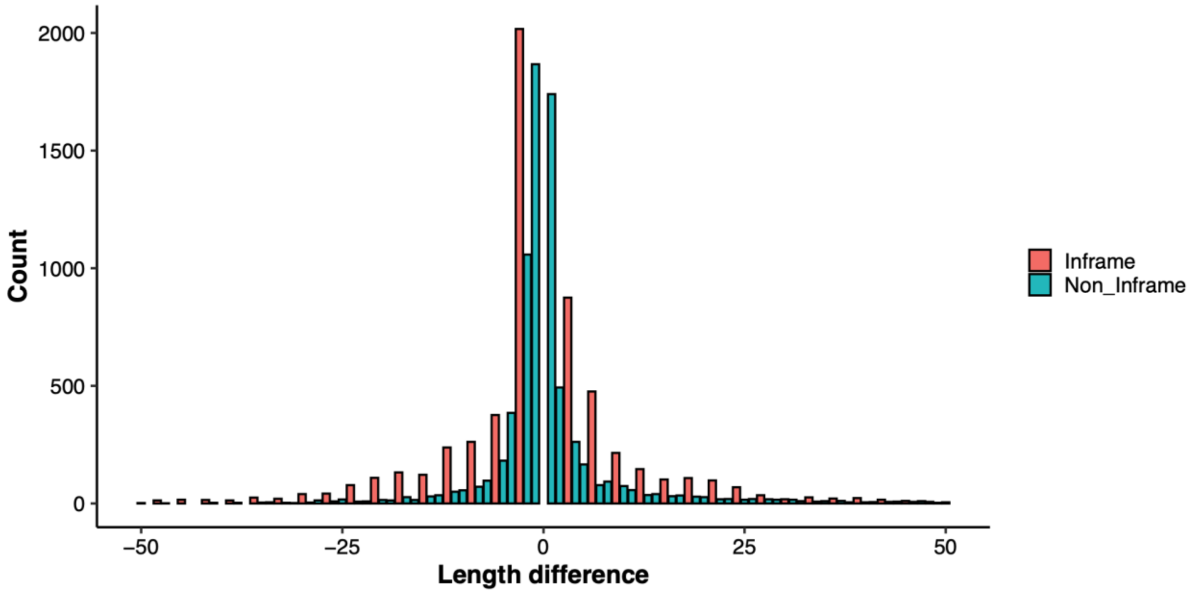


Fig. S9 Length distribution of Indels in the coding region.

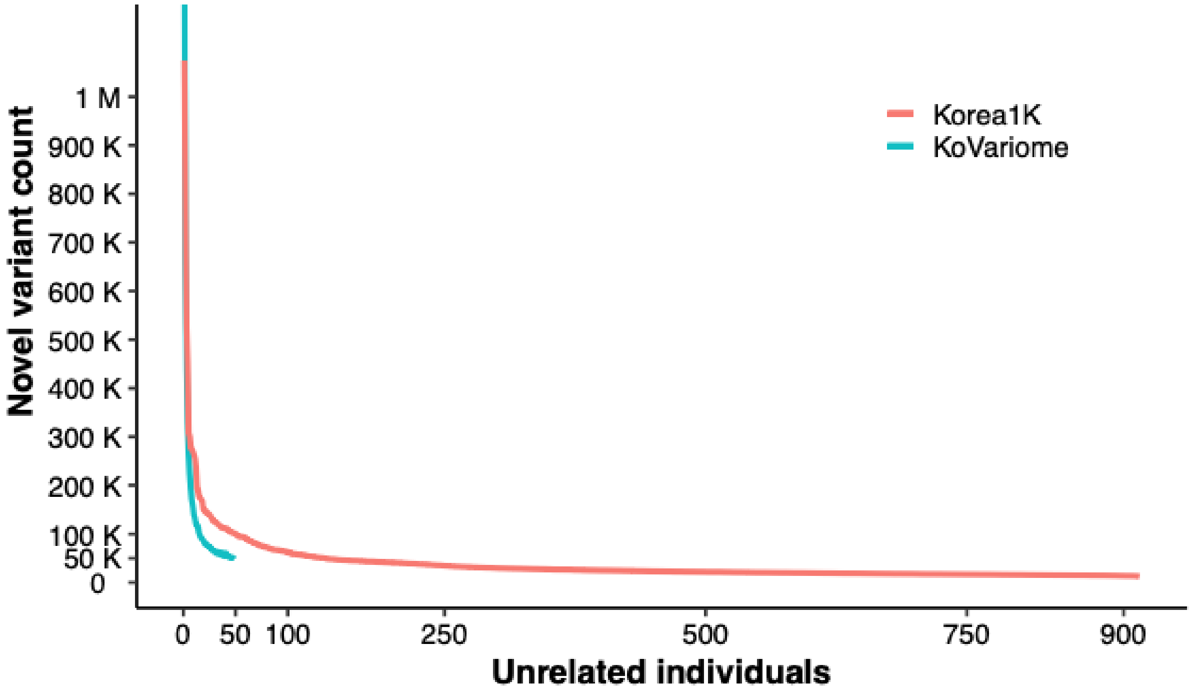


Fig. S10 Number of novel variants as a function of new unrelated individuals.

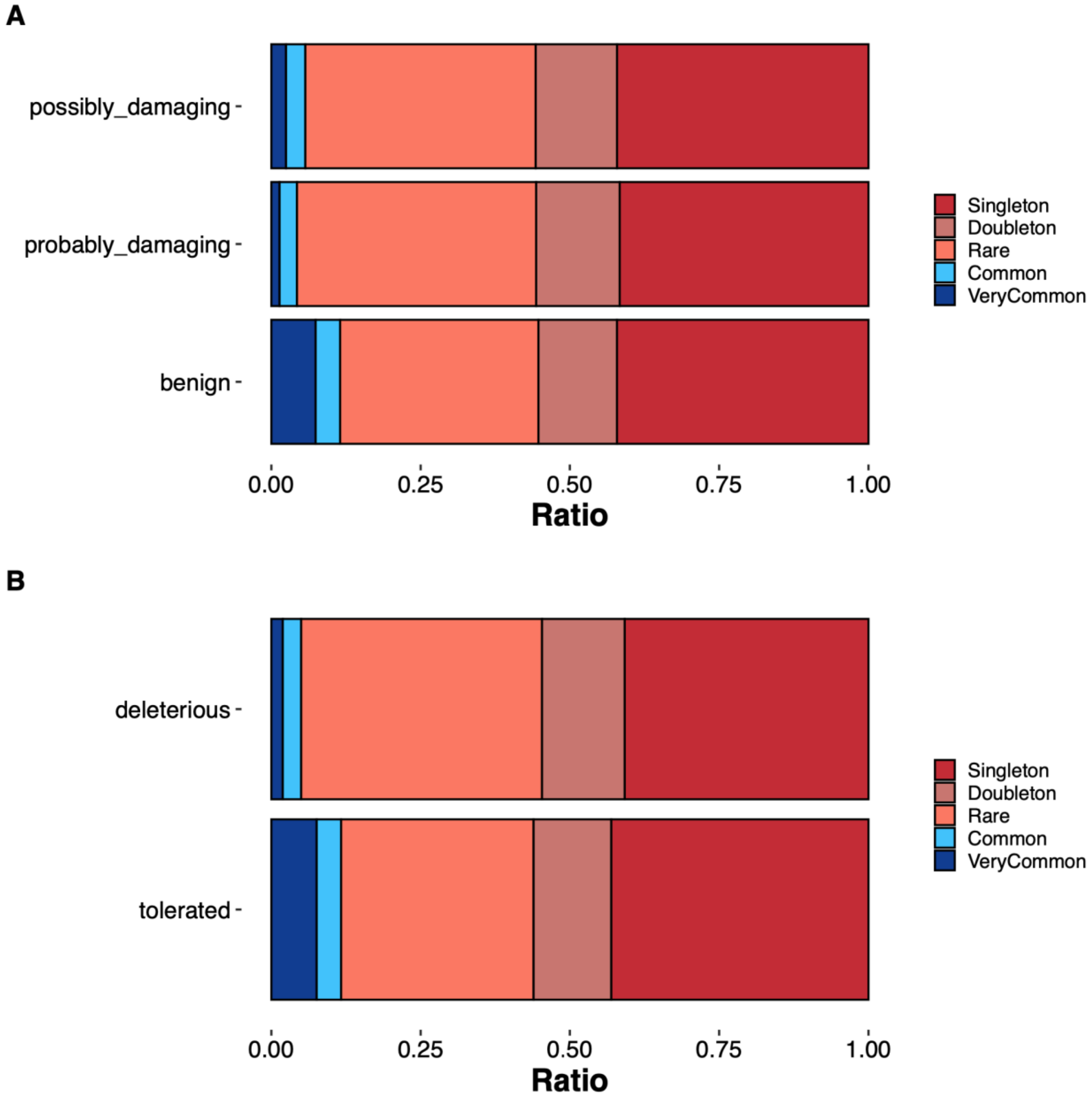


Fig. S11 Proportion of variants based on allele categories for **A)** PolyPhen and **B)** SIFT estimation. Singleton: allele count =1; Doubleton allele count =2; rare: allele count > 2 and allele frequency ≤ 0.01; common: allele frequency > 0.01 and allele frequency ≤ 0.05; very common: allele frequency > 0.05

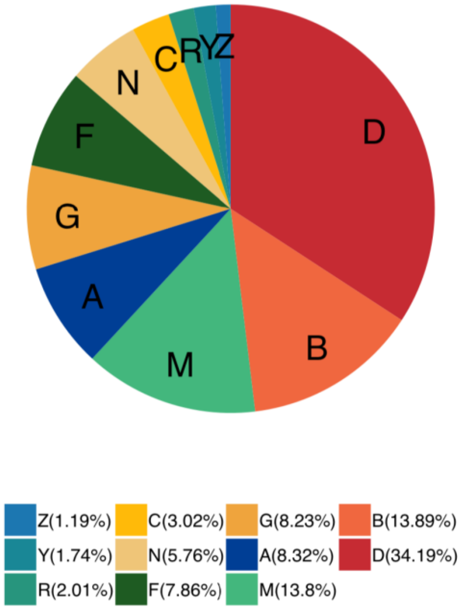


Fig. S12 Mitochondrial haplogroup distribution in Korea1K.

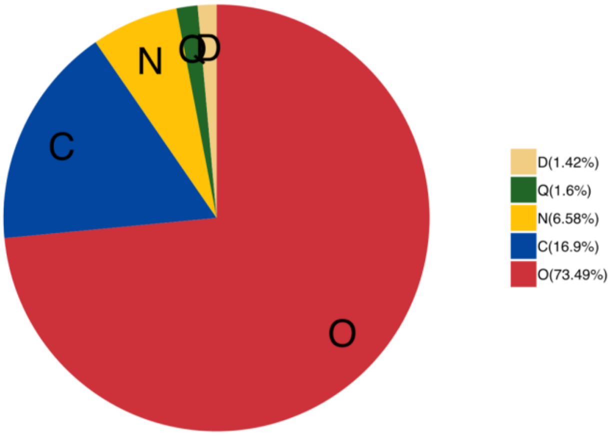


Fig. S13 Chromosome Y haplogroup distribution in Korea1K.

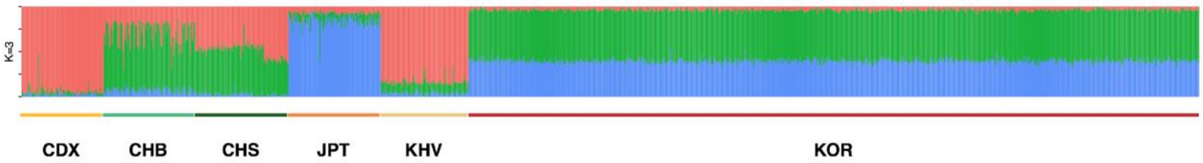


Fig. S14 ADMIXTURE plot for Korea1K and 1KGP East Asians. We used *K*=3 which showed the smallest cross-validation error. (CDX: Dai Chinese; CHB: Han Chinese; CHS: Southern Han Chinese; JPT: Japanese; KHV: Kinh Vietnamese)

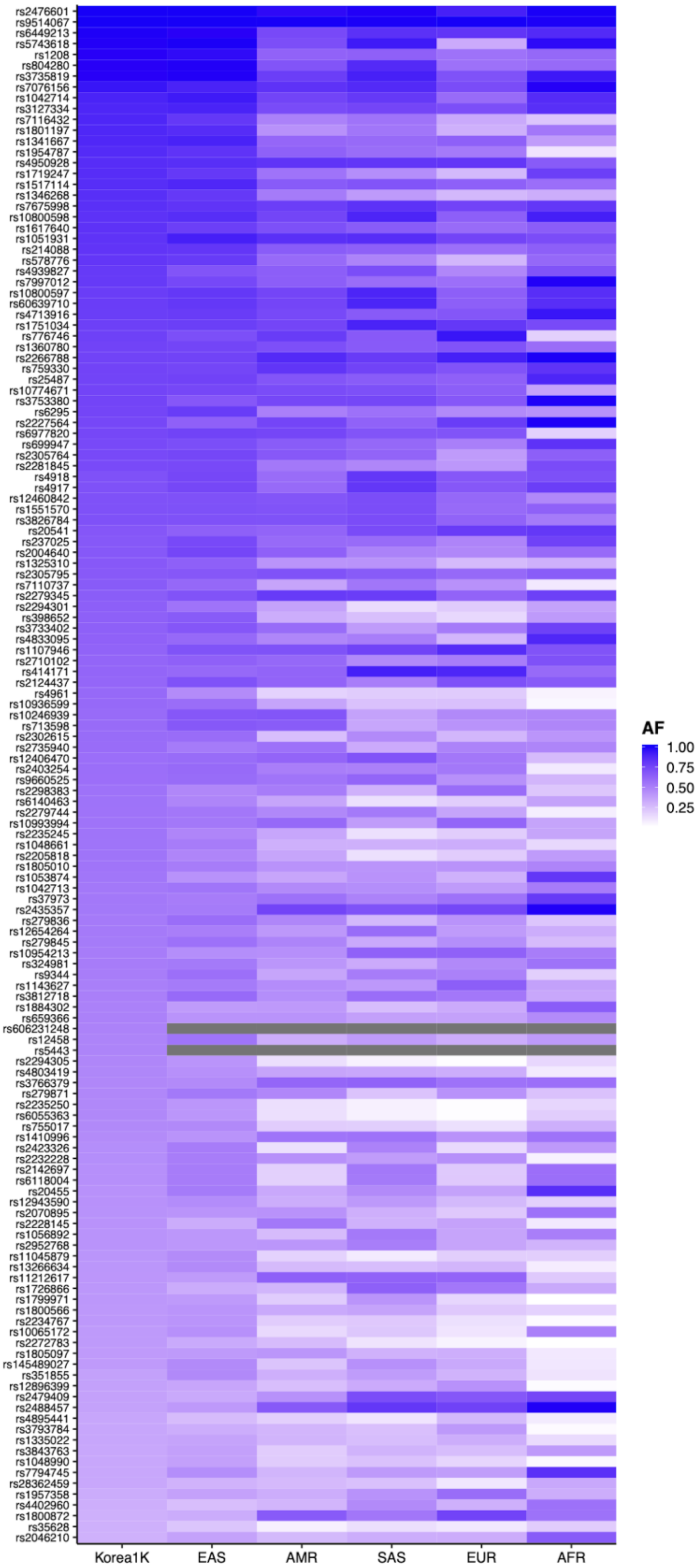
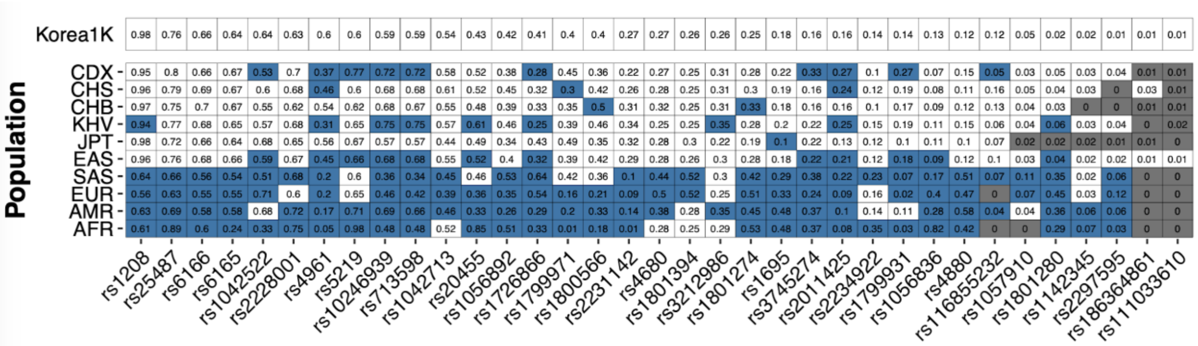


Fig. S15 ClinVar variants which have more than 10% of allele frequency in the Korea1K. The allele frequencies for the super population of 1KGP were also presented. (EAS: East Asian; SAS: South Asian; EUR: European; AMR: American; AFR: African)



	rs1208	rs25487	rs6166	rs6165	rs1042522	rs2228001	rs4961	rs5219	rs10246939	rs713598	rs1042713	rs20455
Korea1K	0.98	0.76	0.66	0.64	0.64	0.63	0.6	0.6	0.59	0.59	0.54	0.43
CDX	0.95	0.8	0.66	0.67	0.53	0.7	0.37	0.77	0.72	0.72	0.58	0.52
CHS	0.96	0.79	0.69	0.67	0.6	0.68	0.46	0.6	0.68	0.68	0.61	0.52
CHB	0.97	0.75	0.7	0.67	0.55	0.62	0.54	0.62	0.68	0.67	0.55	0.48
KHV	0.94	0.77	0.68	0.65	0.57	0.68	0.31	0.65	0.75	0.75	0.57	0.61
JPT	0.98	0.72	0.66	0.64	0.68	0.65	0.56	0.67	0.57	0.57	0.44	0.49
EAS	0.96	0.76	0.68	0.66	0.59	0.67	0.45	0.66	0.68	0.68	0.55	0.52
SAS	0.64	0.66	0.56	0.54	0.51	0.68	0.2	0.6	0.36	0.34	0.45	0.46
EUR	0.56	0.63	0.55	0.55	0.71	0.6	0.2	0.65	0.46	0.42	0.39	0.36
AMR	0.63	0.69	0.58	0.58	0.68	0.72	0.17	0.71	0.69	0.66	0.46	0.33
AFR	0.61	0.89	0.6	0.24	0.33	0.75	0.05	0.98	0.48	0.48	0.52	0.85

	rs1056892	rs1726866	rs1799971	rs1800566	rs2231142	rs4680	rs1801394	rs3212986	rs1801274	rs1695	rs3745274	rs2011425
Korea1K	0.42	0.41	0.4	0.4	0.27	0.27	0.26	0.26	0.25	0.18	0.16	0.16
CDX	0.38	0.28	0.45	0.36	0.22	0.27	0.25	0.31	0.28	0.22	0.33	0.27
CHS	0.45	0.32	0.3	0.42	0.26	0.28	0.25	0.31	0.3	0.19	0.16	0.24
CHB	0.39	0.33	0.35	0.5	0.31	0.32	0.25	0.31	0.33	0.18	0.16	0.16
KHV	0.46	0.25	0.39	0.46	0.34	0.25	0.25	0.35	0.28	0.2	0.22	0.25
JPT	0.34	0.43	0.49	0.35	0.32	0.28	0.3	0.22	0.19	0.1	0.22	0.13
EAS	0.4	0.32	0.39	0.42	0.29	0.28	0.26	0.3	0.28	0.18	0.22	0.21
SAS	0.53	0.64	0.42	0.36	0.1	0.44	0.52	0.3	0.42	0.29	0.38	0.22
EUR	0.35	0.54	0.16	0.21	0.09	0.5	0.52	0.25	0.51	0.33	0.24	0.09
AMR	0.26	0.29	0.2	0.33	0.14	0.38	0.28	0.35	0.45	0.48	0.37	0.1
AFR	0.51	0.33	0.01	0.18	0.01	0.28	0.25	0.29	0.53	0.48	0.37	0.08

	rs2234922	rs1799931	rs1056836	rs4880	rs116855232	rs1057910	rs1801280	rs1142345	rs2297595	rs186364861	rs111033610
Korea1K	0.14	0.14	0.13	0.12	0.12	0.05	0.02	0.02	0.01	0.01	0.01
CDX	0.1	0.27	0.07	0.15	0.05	0.03	0.05	0.03	0.04	0.01	0.01
CHS	0.12	0.19	0.08	0.11	0.16	0.05	0.04	0.03	0	0.03	0.01
CHB	0.1	0.17	0.09	0.12	0.13	0.04	0.03	0	0	0.01	0.01
KHV	0.15	0.19	0.11	0.15	0.06	0.04	0.06	0.03	0.04	0	0.02
JPT	0.12	0.1	0.11	0.1	0.07	0.02	0.02	0.02	0.01	0.01	0
EAS	0.12	0.18	0.09	0.12	0.1	0.03	0.04	0.02	0.02	0.01	0.01
SAS	0.23	0.07	0.17	0.51	0.07	0.11	0.35	0.02	0.06	0	0
EUR	0.16	0.02	0.4	0.47	0	0.07	0.45	0.03	0.12	0	0
AMR	0.14	0.11	0.28	0.58	0.04	0.04	0.36	0.06	0.06	0	0
AFR	0.35	0.03	0.82	0.42	0	0	0.29	0.07	0.03	0	0

Fig. S16 Drug response variants found in Korea1K. Blue indicates significantly different allele frequencies between the Korea1K dataset and the population from the Chi-square test. White indicates not significant. Grey indicates a Chi-square test could not be performed because of low allele count. Abbreviation on Y-axis is the same population code as 1KGP (CDX: Dai Chinese; CHB: Han Chinese; CHS: Southern Han Chinese; JPT: Japanese; KHV: Kinh Vietnamese; EAS: East Asian; SAS: South Asian; EUR: European; AMR: American; AFR: African).

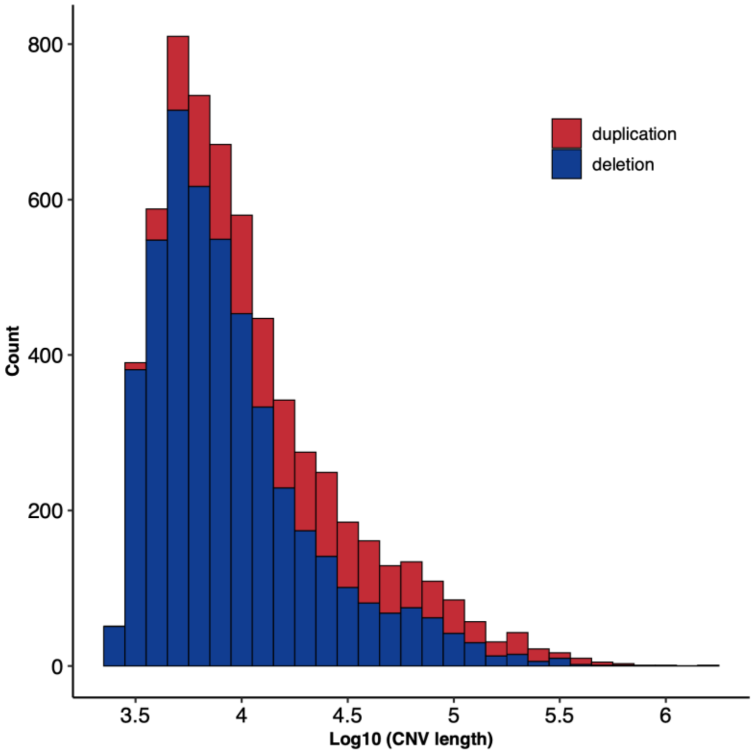


Fig. S17 Length distribution of copy number variations.

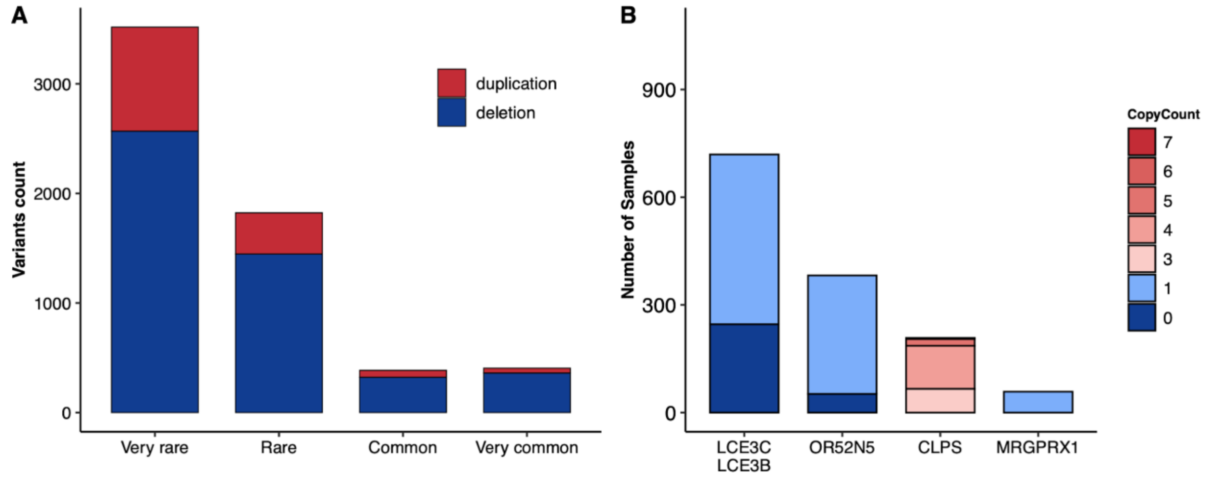


Fig. S18 Copy number variations in Korea1K. **(A)** The number of copy number variations (CNVs) based on categories of sample frequency. Very rare: sample frequency ≤ 0.001; rare: sample frequency > 0.001 and sample frequency ≤ 0.01; common: sample frequency > 0.01 and sample frequency ≤ 0.05; very common: sample frequency > 0.05. The colors indicate the types of CNVs. **(B)** Common CNVs overlapped with 1KGP set and protein-coding genes. Colors indicates the copy number.

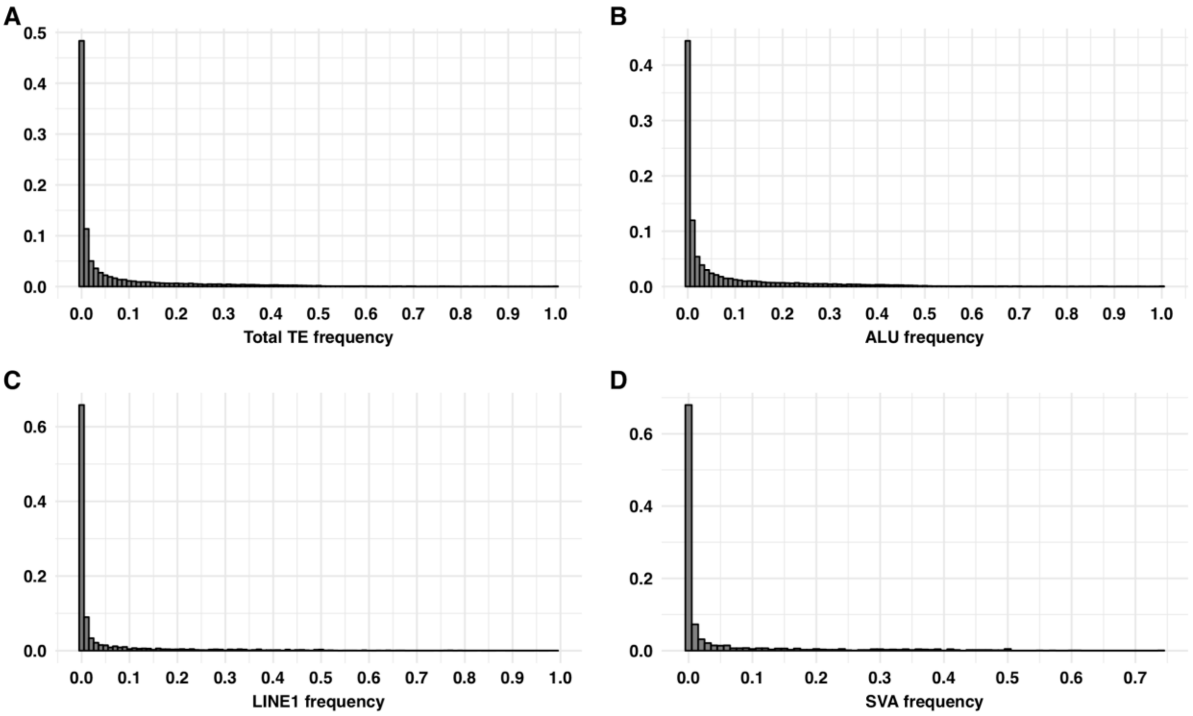


Fig. S19 Transposable element (TE) insertion frequency distribution in Korea1K. **(A)** All TE type. **(B)** ALU. **(C)** LINE1. **(D)** SVA.

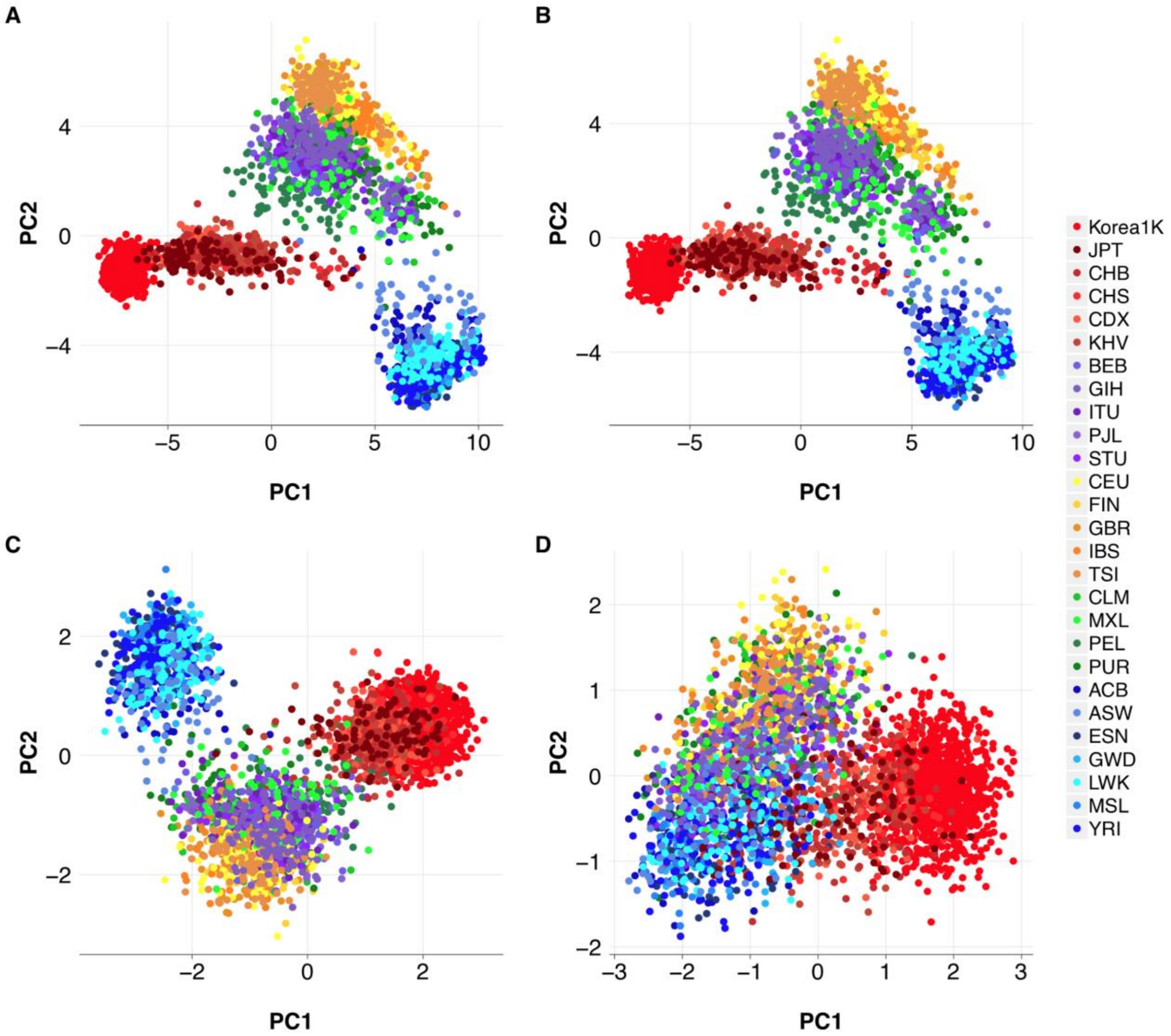


Fig. S20 PCA plot using Transposable element (TE) insertion. **(A)** All TE types. **(B)** ALU. **(C)** LINE1. **(D)** SVA.

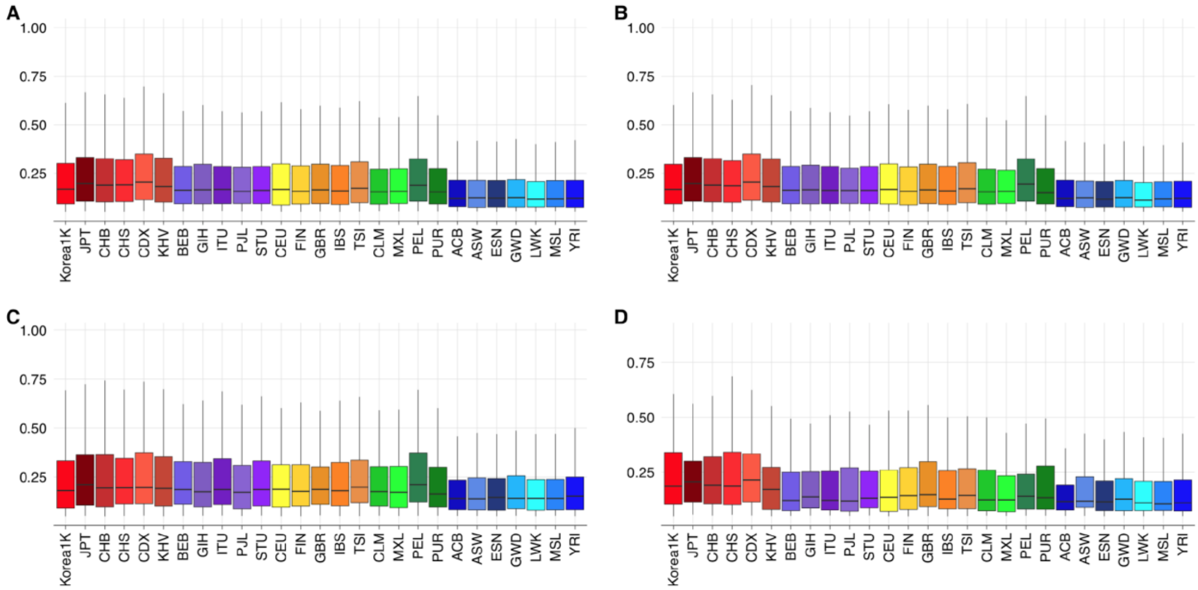


Fig. S21 Transposable element (TE) insertion frequency distribution of Korea1K and 1KGP populations. **(A)** All TE types. **(B)** ALU. **(C)** LINE1. **(D)** SVA.

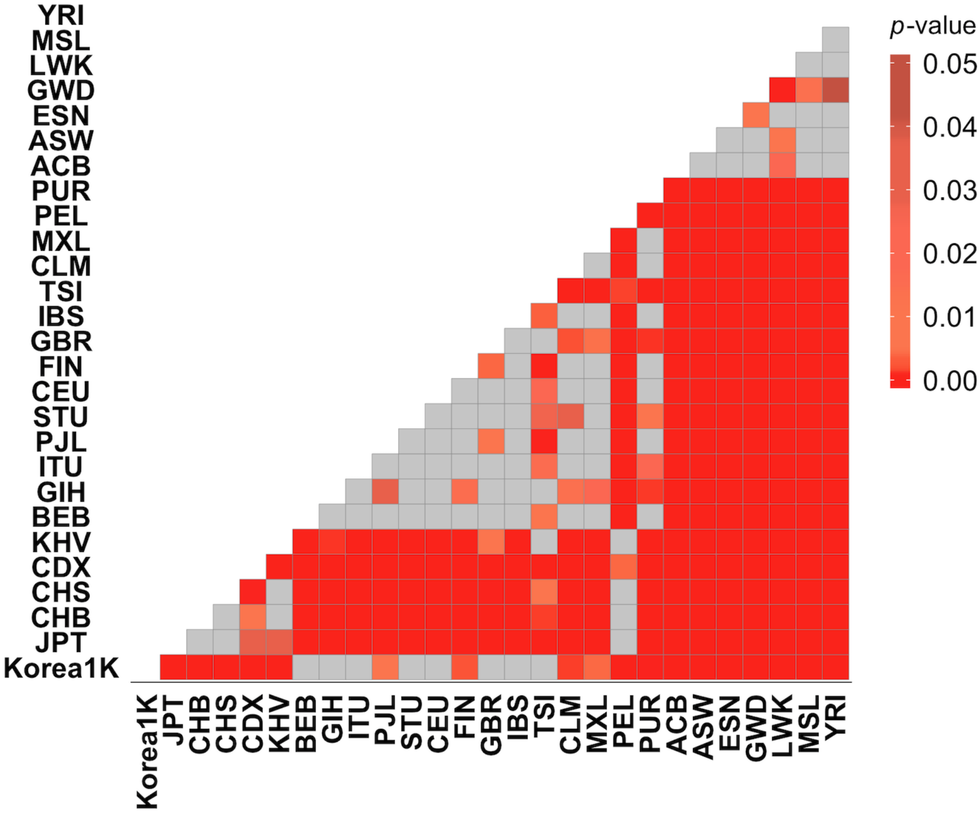


Fig. S22 Significance of TE insertion allele frequency difference. Colors represent *P*-value. The red box indicates a significant difference in TE insertion allele frequency distribution calculated by the Wilcoxon rank-sum test. The box which does not show a significant difference (*P*-value >0.05) colored into gray.

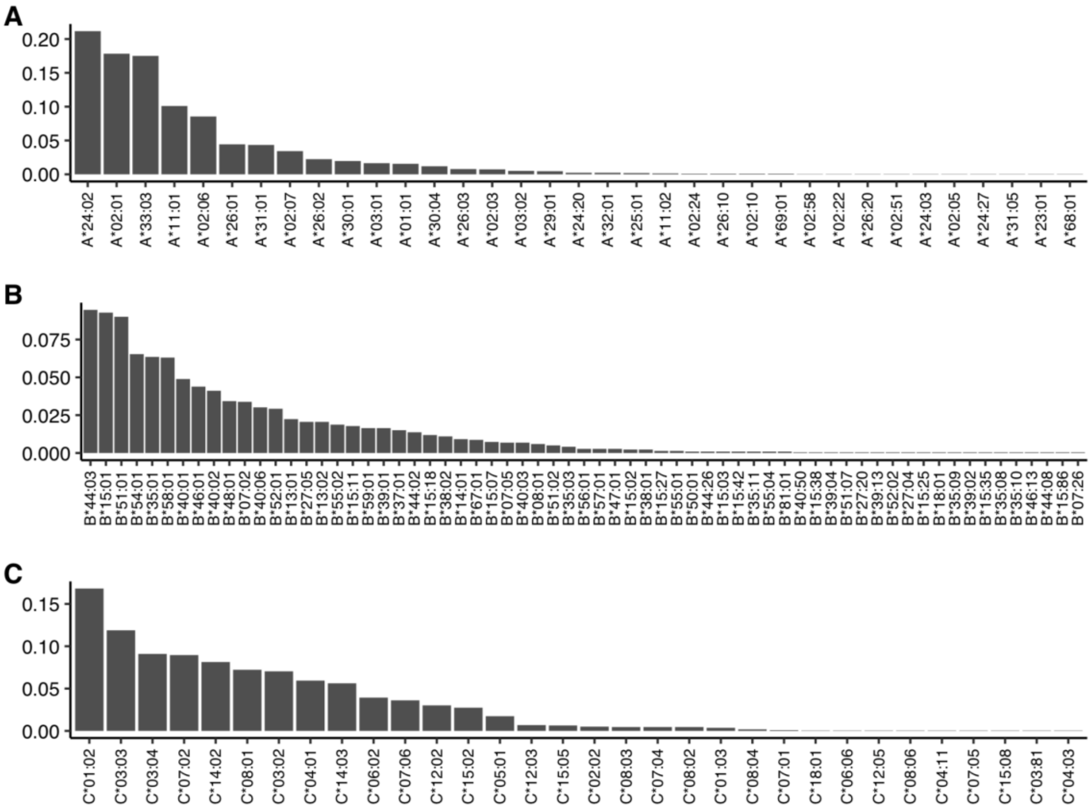


Fig. S23 HLA allele distribution in Korea1K. (A) HLA-A. (B) HLA-B. (C) HLA-C. The X-axis indicates the HLA allele type. Y-axis indicates the proportion of each type in Korea1K.

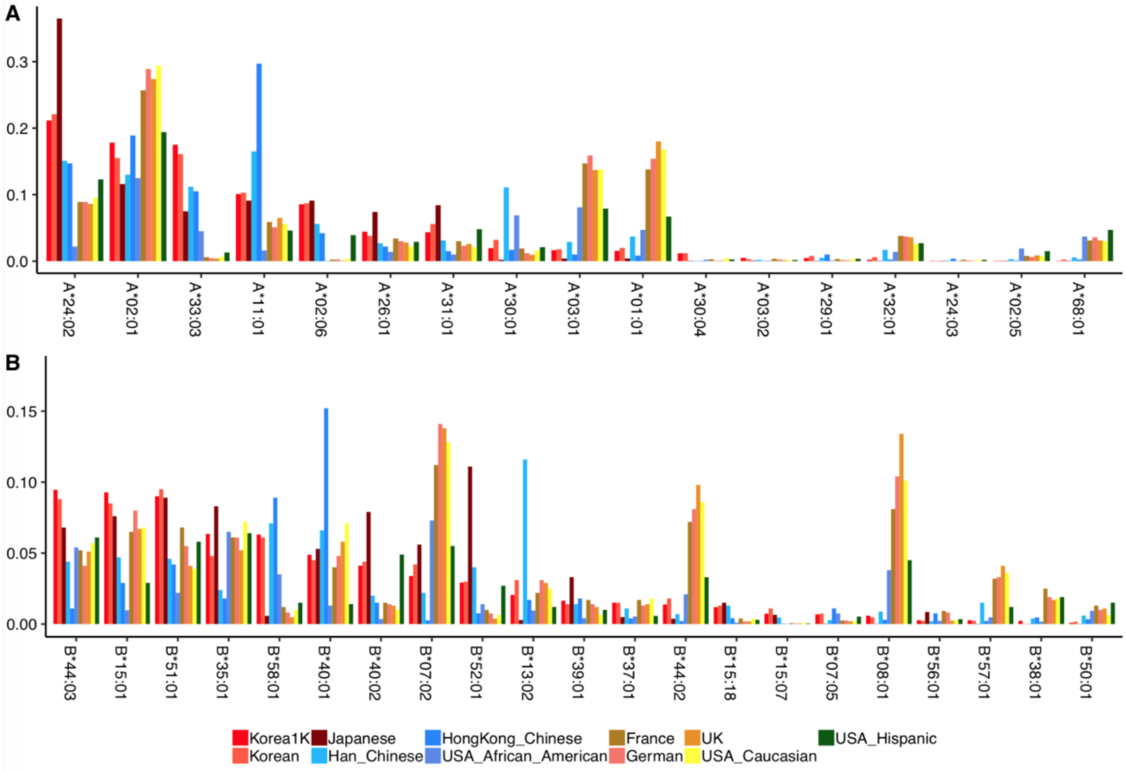


Fig. S24 Comparison of HLA type frequency to the public database. **(A)** Allele frequency of HLA-A loci. **(B)** Allele frequency of HLA-B loci.

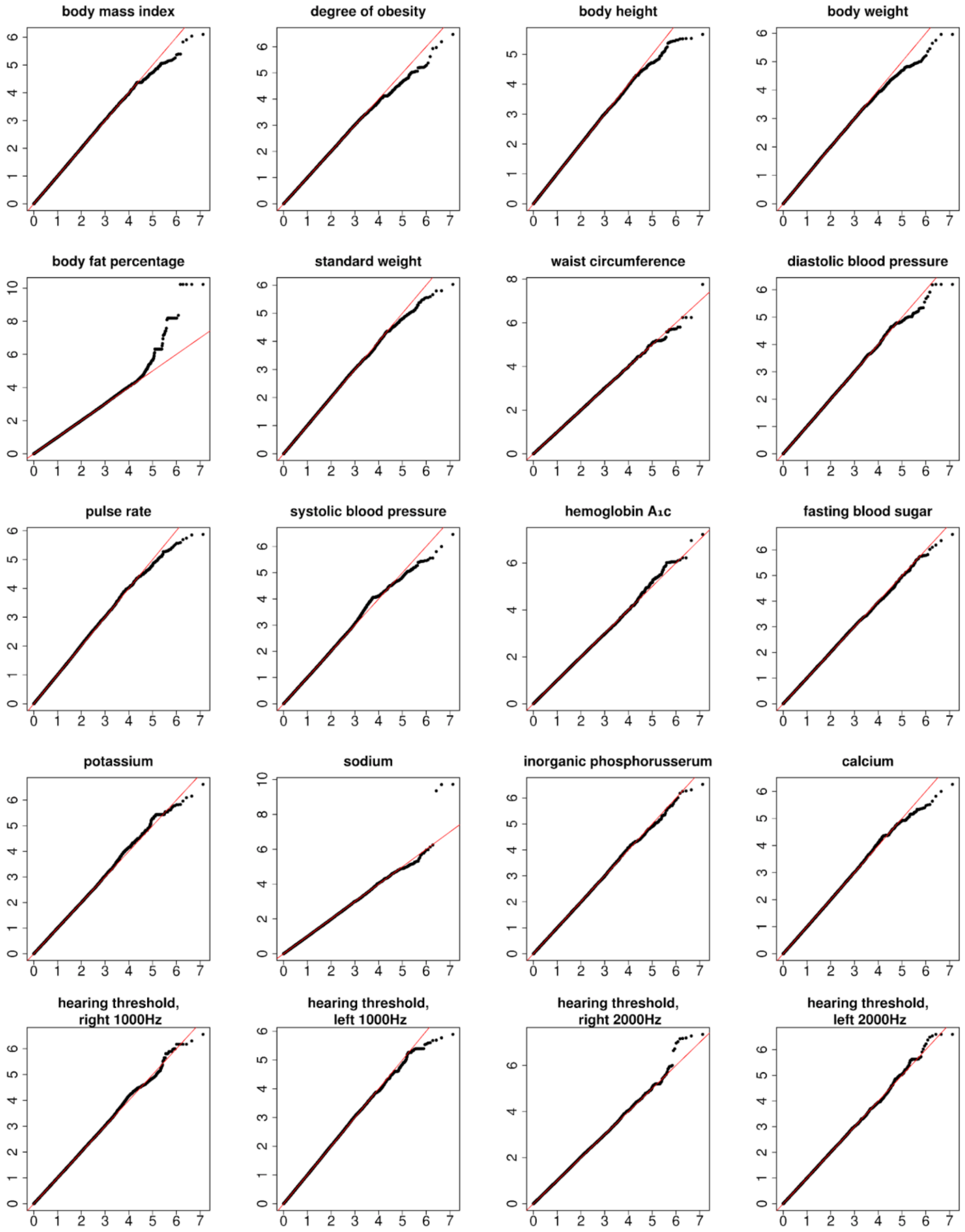


Fig. S25 QQplots for the GWA tests of the 20 traits. X-axis indicates expected -log$_{10}$ *P*-value. Y-axis indicates observed -log$_{10}$ *P*-value.

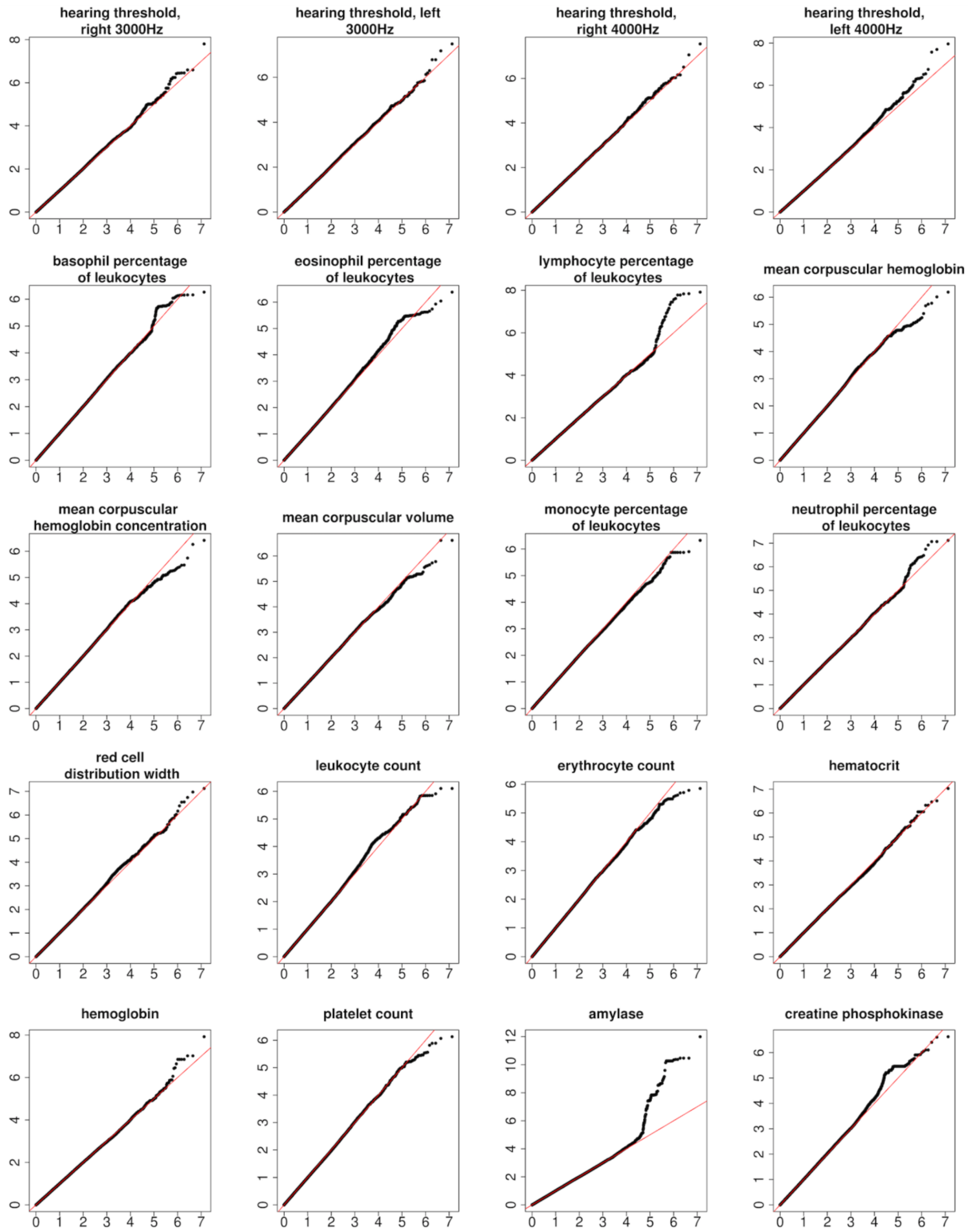


Fig. S26 QQplots for the GWA tests of the 20 traits. X-axis indicates expected -$\log_{10}$ *P*-value. Y-axis indicates observed -$\log_{10}$ *P*-value.

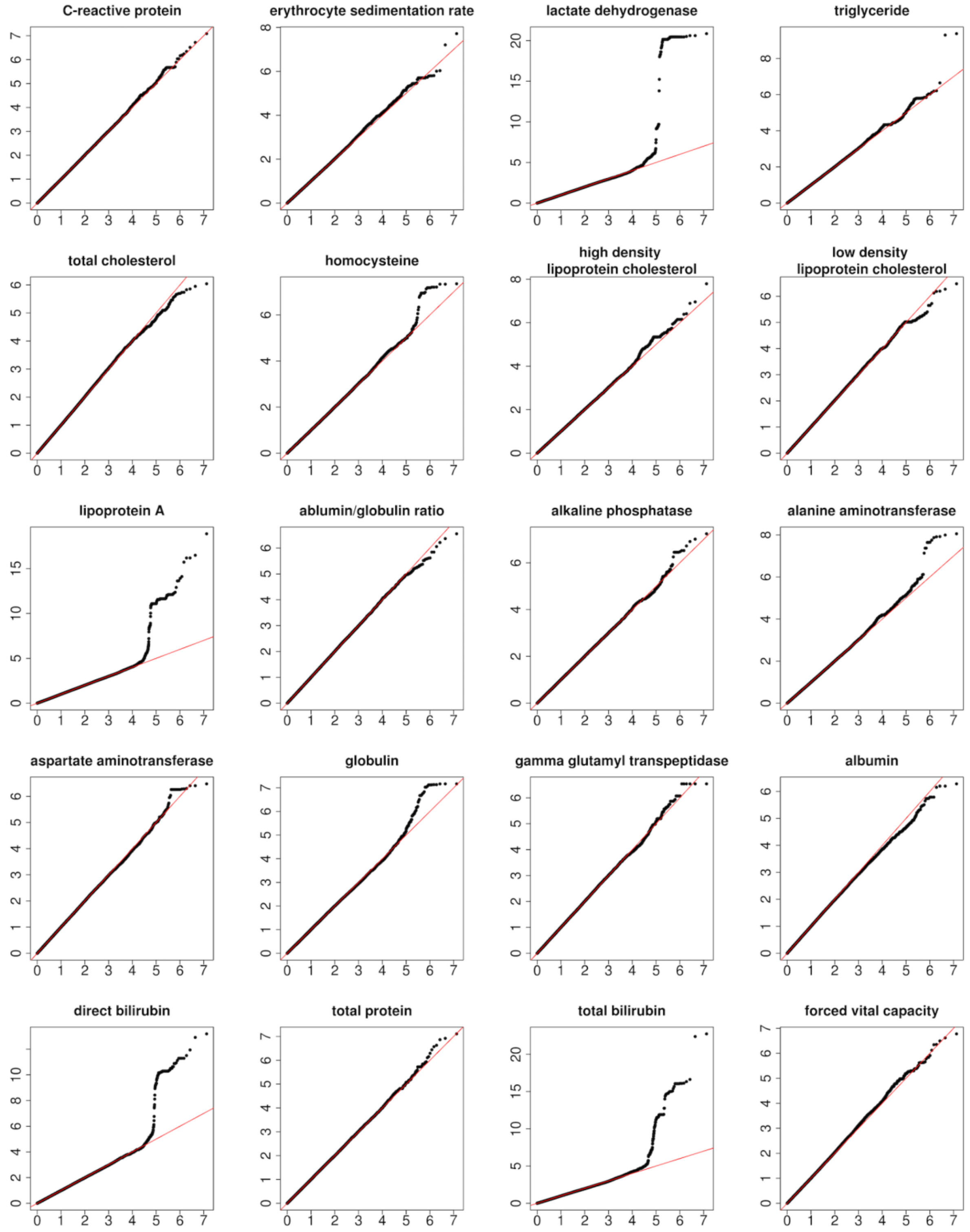


Fig. S27 QQplots for the GWA tests of the 20 traits. X-axis indicates expected -$\log_{10}$ P-value. Y-axis indicates observed -$\log_{10}$ *P*-value.

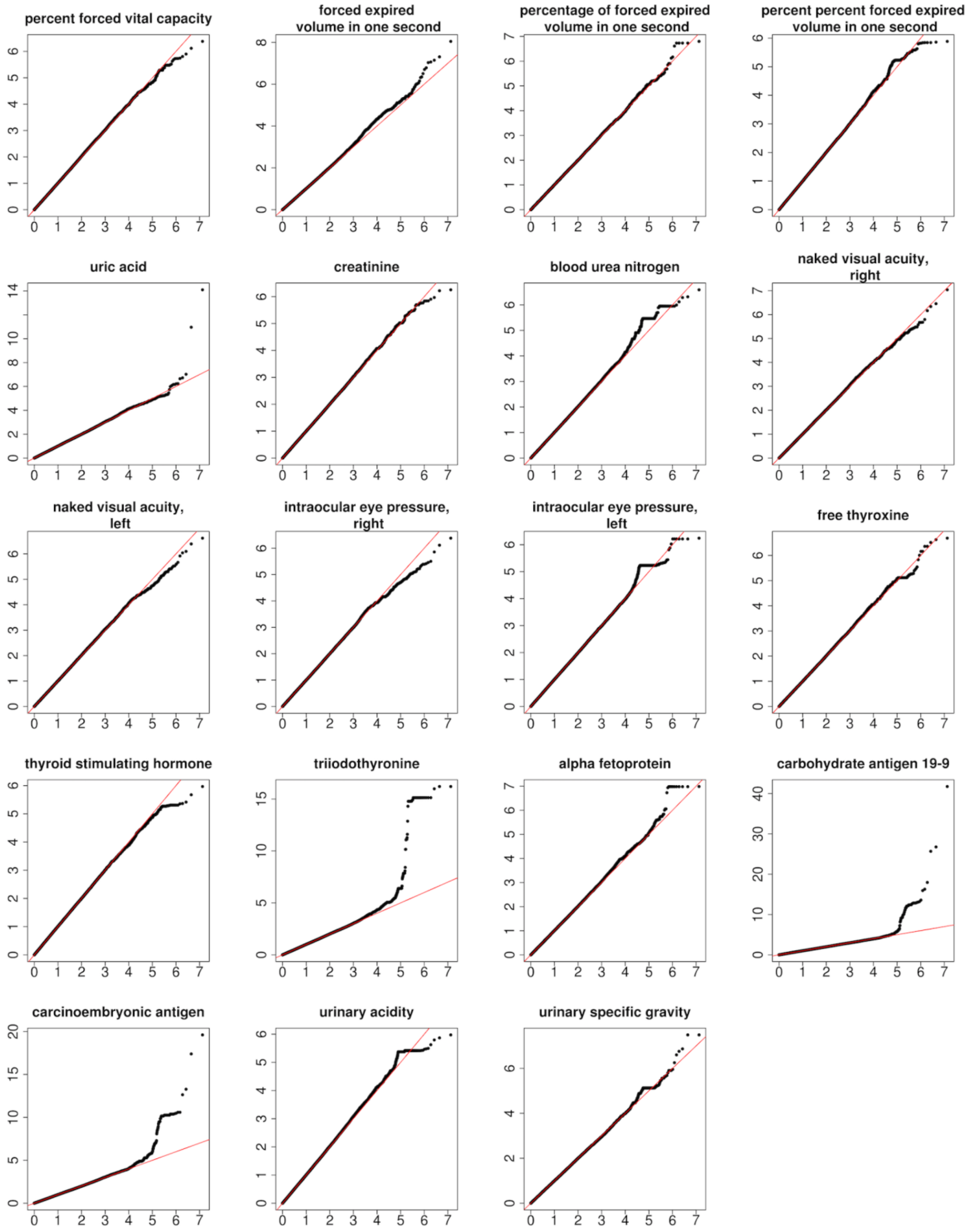


Fig. S28 QQplots for the GWA tests of the 19 traits. X-axis indicates expected -$\log_{10}$ P-value. Y-axis indicates observed -$\log_{10}$ *P*-value.

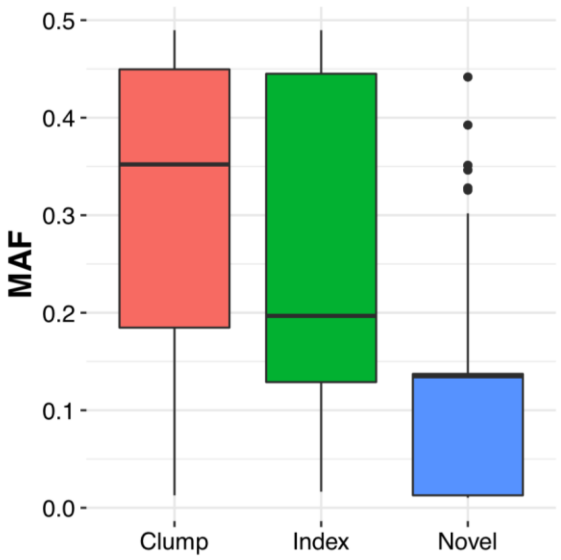


Fig. S29 Minor allele frequency (MAF) of the most significant variant on the loci from GWA analysis. 'Clump' means that clump variants in the loci were reported. 'Index' means that index variants in the loci were reported. 'Novel' means that no variants in the loci were reported.

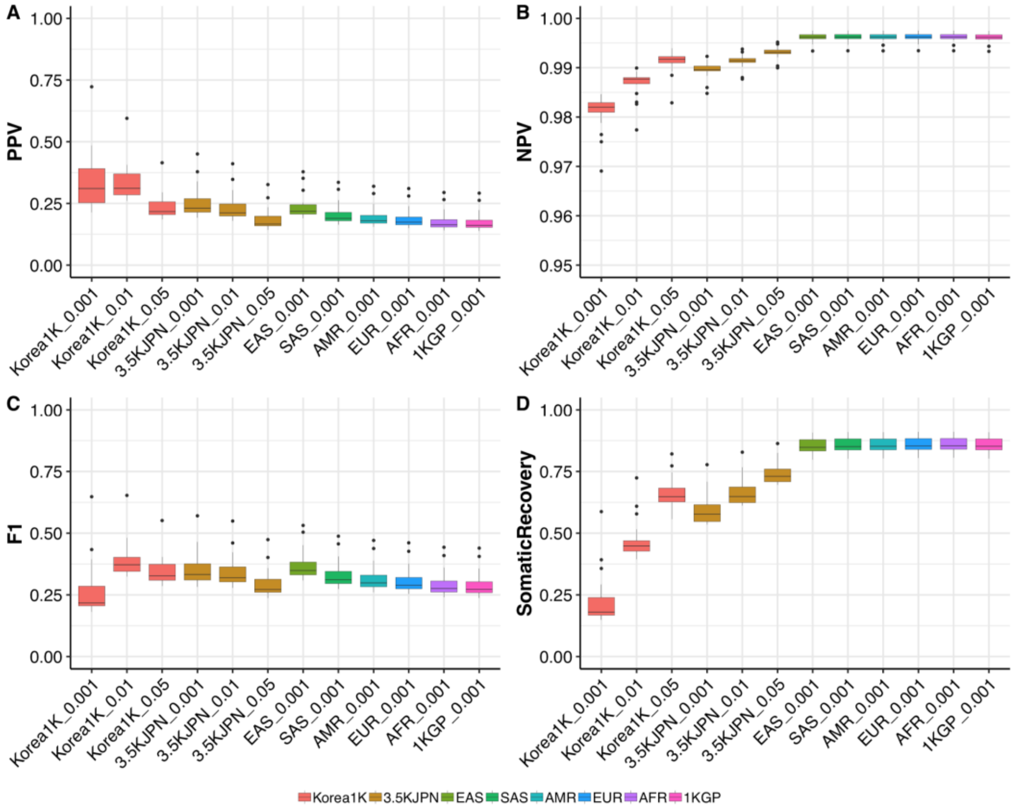


Fig. S30 Performance of the variant classification using different panels of normals**.** Numbers on the X-axis indicate allele frequency cut-off for selecting variants from the panel. PPV and NPV mean positive and negative predictive values, respectively.

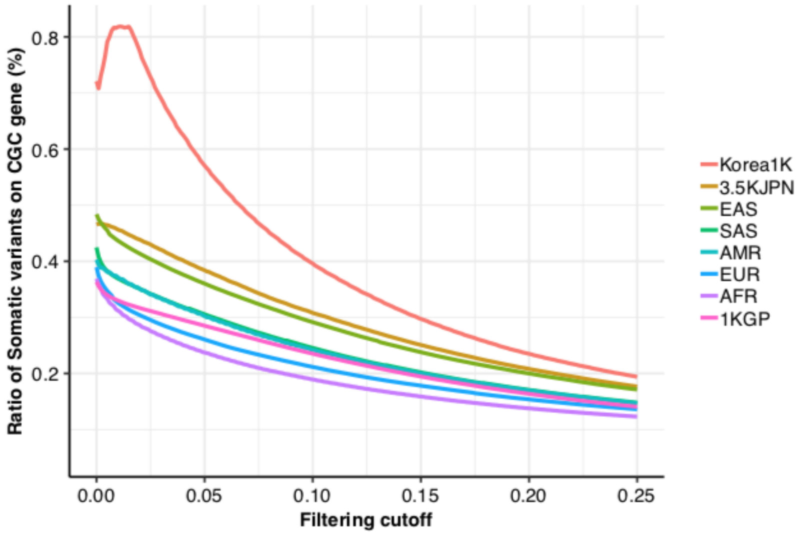


Fig. S31 Ratio of true somatic variants in CGC genes based on predicted somatic variants using a panel of normal.

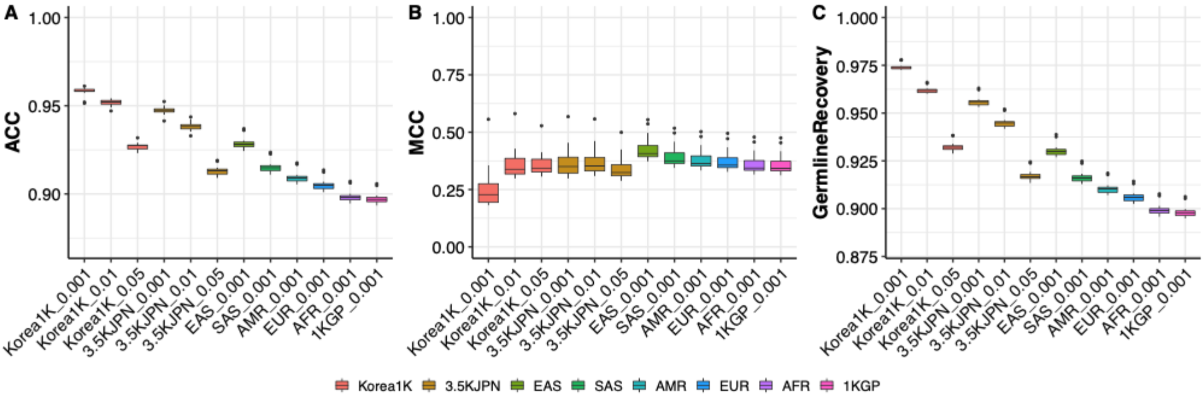


Fig. S32 Performance of the variant classification using different panels of normal when the only lift-over possible region was applied. **(A)** Accuracy of classification. **(B)** Matthews correlation coefficient (MCC) values. **(C)** Germline recovery rate. Numbers on the X-axis indicate allele frequency cut-off for selecting variants from each panel.

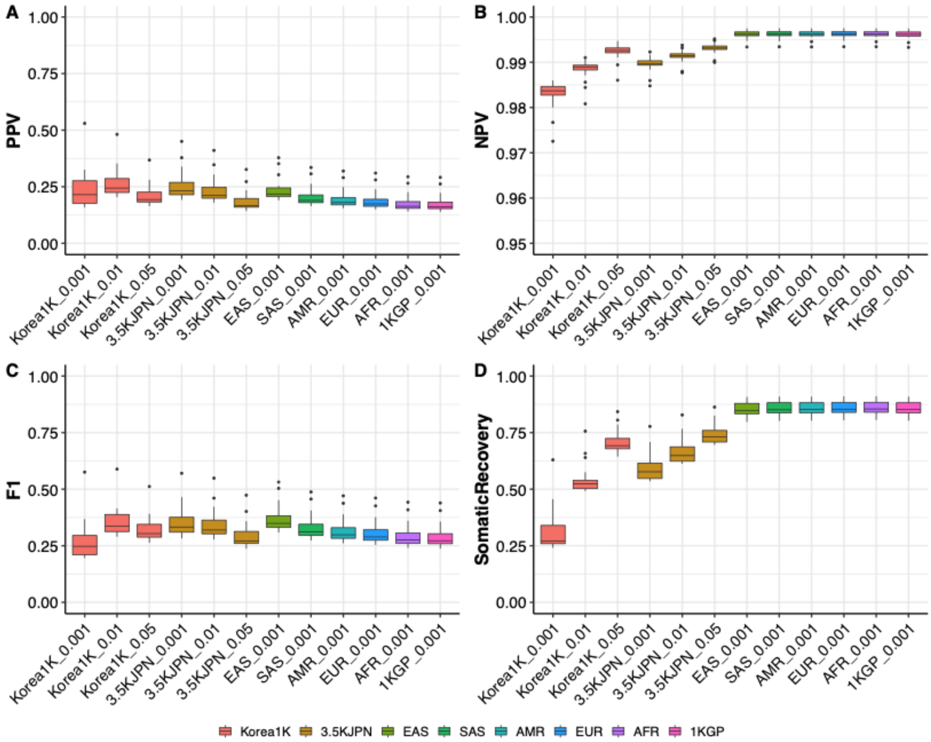


Fig. S33 Performance of the variant classification using different panels of normal when the only lift-over possible region was applied**.** Numbers on the X-axis indicate allele frequency cut-off for selecting variants from the panel. PPV and NPV mean positive and negative predictive values, respectively.

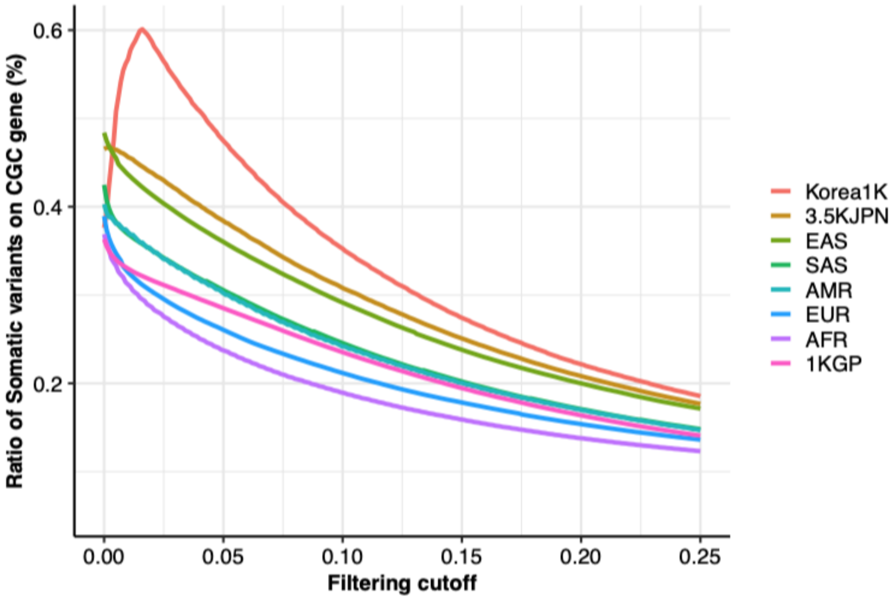


Fig. S34 Ratio of true somatic variants in CGC genes based on predicted somatic variants using a panel of normal when the only lift-over possible region was applied.

Supplementary tables

Table S1 Variant count before and after removing batch effect. Singleton: allele count =1; Doubleton allele count =2; rare: allele count > 2 and allele frequency ≤ 0.01; common: allele frequency > 0.01 and allele frequency ≤ 0.05; very common: allele frequency > 0.05. The allele frequency category in this table was based on 1,094 individuals.

Variant type	**Allele frequency category**	**Reported**	**Before removing batch effect**	**After removing batch effect**	**Remaining percentage**
SNV	Very Common	Novel	237,739	68,290	28.72%
	Very Common	dbSNP	6,122,812	4,589,385	74.96%
	Common	Novel	536,623	410,651	76.53%
	Common	dbSNP	2,104,239	1,612,718	76.64%
	Rare	Novel	7,777,745	6,988,611	89.85%
	Rare	dbSNP	5,380,191	4,212,315	78.29%
	Doubleton	Novel	3,034,057	2,804,907	92.45%
	Doubleton	dbSNP	1,549,851	1,365,313	88.09%
	Singleton	Novel	8,787,498	8,729,585	99.34%
	Singleton	dbSNP	3,659,762	3,434,077	93.83%
	Total SNV		39,190,517	34,215,852	87.31%
Indel	Very Common	Novel	673,458	236,402	35.10%
	Very Common	dbSNP	1,478,967	850,680	57.52%
	Common	Novel	942,429	408,410	43.34%
	Common	dbSNP	543,954	280,808	51.62%
	Rare	Novel	1,407,307	900,776	64.01%
	Rare	dbSNP	563,066	383,256	68.07%
	Doubleton	Novel	440,245	362,171	82.27%
	Doubleton	dbSNP	107,841	88,174	81.76%
	Singleton	Novel	1,191,058	1,118,475	93.91%
	Singleton	dbSNP	207,223	180,358	87.04%
	Total Indel		7,555,548	4,809,510	63.66%
Total variants			46,746,065	39,025,362	83.48%

Table S2 Number of Transposable element (TE) insertions before and after filtering.

TE type	**Number of TE loci before filtering**	**Number of TE loci after filtering**
ALU	23,924	23,915
LINE1	3,708	3,707
SVA	1,522	1,521
Total	29,154	29,143

Table S3 Average base quality by position in the reads of sequencing data.

Base Position	**Average Quality**
1	31.00
2	31.35
3	35.12
4	35.72
5	35.98
6	39.37
7	39.49
8	39.56
9	39.62
10-14	39.64
15-19	39.60
20-24	39.52
25-29	39.42
30-34	39.34
35-39	39.26
40-44	39.19
45-49	39.08
50-54	38.99
55-59	38.91
60-64	38.82
65-69	38.72
70-74	38.63
75-79	38.33
80-84	38.57
85-89	38.47
90-94	38.36
95-99	38.23
100-104	38.08
105-109	37.91
110-114	37.71
115-119	37.48
120-124	37.22
125-129	36.93
130-134	36.60
135-139	36.23
140-144	35.84
145-149	35.42
150	35.57

REFERENCES AND NOTES

1. V. Siska, E. R. Jones, S. Jeon, Y. Bhak, H. M. Kim, Y. S. Cho, H. Kim, K. Lee, E. Veselovskaya, T. Balueva, M. Gallego-Llorente, M. Hofreiter, D. G. Bradley, A. Eriksson, R. Pinhasi, J. Bhak, A. Manica, Genome-wide data from two early Neolithic East Asian individuals dating to 7700 years ago. *Sci. Adv.* **3**, e1601877 (2017).

2. HUGO Pan-Asian SNP Consortium, M. A. Abdulla, I. Ahmed, A. Assawamakin, J. Bhak, S. K. Brahmachari, G. C. Calacal, A. Chaurasia, C. H. Chen, J. Chen, Y. T. Chen, J. Chu, E. M. Cutiongco-de la Paz, M. C. De Ungria, F. C. Delfin, J. Edo, S. Fuchareon, H. Ghang, T. Gojobori, J. Han, S. F. Ho, B. P. Hoh, W. Huang, H. Inoko, P. Jha, T. A. Jinam, L. Jin, J. Jung, D. Kangwanpong, J. Kampuansai, G. C. Kennedy, P. Khurana, H. L. Kim, K. Kim, S. Kim, W. Y. Kim, K. Kimm, R. Kimura, T. Koike, S. Kulawonganunchai, V. Kumar, P. S. Lai, J. Y. Lee, S. Lee, E. T. Liu, P. P. Majumder, K. K. Mandapati, S. Marzuki, W. Mitchell, M. Mukerji, K. Naritomi, C. Ngamphiw, N. Niikawa, N. Nishida, B. Oh, S. Oh, J. Ohashi, A. Oka, R. Ong, C. D. Padilla, P. Palittapongarnpim, H. B. Perdigon, M. E. Phipps, E. Png, Y. Sakaki, J. M. Salvador, Y. Sandraling, V. Scaria, M. Seielstad, M. R. Sidek, A. Sinha, M. Srikummool, H. Sudoyo, S. Sugano, H. Suryadi, Y. Suzuki, K. A. Tabbada, A. Tan, K. Tokunaga, S. Tongsima, L. P. Villamor, E. Wang, Y. Wang, H. Wang, J. Y. Wu, H. Xiao, S. Xu, J. O. Yang, Y. Y. Shugart, H. S. Yoo, W. Yuan, G. Zhao, Zilfalil BA; Indian Genome Variation Consortium, Mapping human genetic diversity in Asia. *Science* **326**, 1541–1545 (2009).

3. R. O. K. M. F. Affairs, *Total Number of Overseas Koreans* (2017).

4. Databank, *Population Total* (2018).

5. J.-S. Seo, A. Rhie, J. Kim, S. Lee, M.-H. Sohn, C.-U. Kim, A. Hastie, H. Cao, J.-Y. Yun, J. Kim, J. Kuk, G. H. Park, J. Kim, H. Ryu, J. Kim, M. Roh, J. Baek, M. W. Hunkapiller, J. Korlach, J.-Y. Shin, C. Kim, *De novo* assembly and phasing of a Korean human genome. *Nature* **538**, 243–247 (2016).

6. S. Lee, J. Seo, J. Park, J.-Y. Nam, A. Choi, J. S. Ignatius, R. D. Bjornson, J.-H. Chae, I.-J. Jang, S. Lee, W.-Y. Park, D. Baek, M. Choi, Korean variant archive (KOVA): A reference database of genetic variations in the Korean population. *Sci. Rep.* **7**, 4287 (2017).

7. S.-M. Ahn, T.-H. Kim, S. Lee, D. Kim, H. Ghang, D.-S. Kim, B.-C. Kim, S.-Y. Kim, W.-Y. Kim, C. Kim, D. Park, Y. S. Lee, S. Kim, R. Reja, S. Jho, C. G. Kim, J.-Y. Cha, K.-H. Kim, B. Lee, J. Bhak, S.-J. Kim, The first Korean genome sequence and analysis: Full genome sequencing for a socio-ethnic group. *Genome Res.* **19**, 1622–1629 (2009).

8. Y. S. Cho, H. Kim, H.-M. Kim, S. Jho, J. H. Jun, Y. J. Lee, K. S. Chae, C. G. Kim, S. Kim, A. Eriksson, J. S. Edwards, S. Lee, B. C. Kim, A. Manica, T.-K. Oh, G. M. Church, J. Bhak, An ethnically relevant consensus Korean reference genome is a step towards personal reference genomes. *Nat. Commun.* **7**, 13637 (2016).

9. J. Kim, J. A. Weber, S. Jho, J. Jang, J. H. Jun, Y. S. Cho, H.-M. Kim, H. Kim, Y. Kim, O. S. Chung, C. G. Kim, H. J. Lee, B. C. Kim, K. Han, I. S. Koh, K. S. Chae, S. Lee, J. S. Edwards, J. Bhak, KoVariome: Korean national standard reference variome database of whole genomes with comprehensive SNV, indel, CNV, and SV analyses. *Sci. Rep.* **8**, 5677 (2018).

10. D. Hong, S. S. Park, Y. S. Ju, S. Kim, J. Y. Shin, S. Kim, S. B. Yu, W. C. Lee, S. Lee, H. Park, J. I. Kim, J. S. Seo, TIARA: A database for accurate analysis of multiple personal genomes based on cross-technology. *Nucleic Acids Res.* **39**, D883–D888 (2011).

11. 1000 Genomes Project Consortium, A. Auton, L. D. Brooks, R. M. Durbin, E. P. Garrison, H. M. Kang, J. O. Korbel, J. L. Marchini, S. McCarthy, G. A. McVean, G. R. Abecasis, A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).

12. UK10K Consortium, K. Walter, J. L. Min, J. Huang, L. Crooks, Y. Memari, S. McCarthy, J. R. Perry, C. Xu, M. Futema, D. Lawson, V. Iotchkova, S. Schiffels, A. E. Hendricks, P. Danecek, R. Li, J. Floyd, L. V. Wain, I. Barroso, S. E. Humphries, M. E. Hurles, E. Zeggini, J. C. Barrett, V. Plagnol, J. B. Richards, C. M. Greenwood, N. J. Timpson, R. Durbin, N. Soranzo, The UK10K project identifies rare variants in health and disease. *Nature* **526**, 82–90 (2015).

13. Genome of the Netherlands Consortium, Whole-genome sequence variation, population structure and demographic history of the Dutch population. *Nat. Genet.* **46**, 818–825 (2014).

14. R. M. Sherman, J. Forman, V. Antonescu, D. Puiu, M. Daya, N. Rafaels, M. P. Boorgula, S. Chavan, C. Vergara, V. E. Ortega, A. M. Levin, C. Eng, M. Yazdanbakhsh, J. G. Wilson, J. Marrugo, L. A. Lange, L. K. Williams, H. Watson, L. B. Ware, C. O. Olopade, O. Olopade, R. R. Oliveira, C. Ober, D. L. Nicolae, D. A. Meyers, A. Mayorga, J. Knight-Madden, T. Hartert, N. N. Hansel, M. G. Foreman, J. G. Ford, M. U. Faruque, G. M. Dunston, L. Caraballo, E. G. Burchard, E. R. Bleecker, M. I. Araujo, E. F. Herrera-Paz, M. Campbell, C. Foster, M. A. Taub, T. H. Beaty, I. Ruczinski, R. A. Mathias, K. C. Barnes, S. L. Salzberg, Assembly of a pan-genome from deep sequencing of 910 humans of African descent. *Nat. Genet.* **51**, 30–35 (2019).

15. D. F. Gudbjartsson, H. Helgason, S. A. Gudjonsson, F. Zink, A. Oddson, A. Gylfason, S. Besenbacher, G. Magnusson, B. V. Halldorsson, E. Hjartarson, G. T. Sigurdsson, S. N. Stacey, M. L. Frigge, H. Holm, J. Saemundsdottir, H. T. Helgadottir, H. Johannsdottir, G. Sigfusson, G. Thorgeirsson, J. T. Sverrisson, S. Gretarsdottir, G. B. Walters, T. Rafnar, B. Thjodleifsson, E. S. Bjornsson, S. Olafsson, H. Thorarinsdottir, T. Steingrimsdottir, T. S. Gudmundsdottir, A. Theodors, J. G. Jonasson, A. Sigurdsson, G. Bjornsdottir, J. J. Jonsson, O. Thorarensen, P. Ludvigsson, H. Gudbjartsson, G. I. Eyjolfsson, O. Sigurdardottir, I. Olafsson, D. O. Arnar, O. T. Magnusson, A. Kong, G. Masson, U. Thorsteinsdottir, A. Helgason, P. Sulem, K. Stefansson, Large-scale whole-genome sequencing of the Icelandic population. *Nat. Genet.* **47**, 435–444 (2015).

16. L. Maretty, J. M. Jensen, B. Petersen, J. A. Sibbesen, S. Liu, P. Villesen, L. Skov, K. Belling, C. Theil Have, J. M. G. Izarzugaza, M. Grosjean, J. Bork-Jensen, J. Grove, T. D. Als, S. Huang, Y. Chang, R. Xu, W. Ye, J. Rao, X. Guo, J. Sun, H. Cao, C. Ye, J. van Beusekom, T. Espeseth, E. Flindt, R. M. Friborg, A. E. Halager, S. le Hellard, C. M. Hultman, F. Lescai, S. Li, O. Lund, P. Løngren, T. Mailund, M. L. Matey-Hernandez, O. Mors, C. N. S. Pedersen, T. Sicheritz-Pontén, P. Sullivan, A. Syed, D. Westergaard, R. Yadav, N. Li, X. Xu, T. Hansen, A. Krogh, L. Bolund, T. I. A. Sørensen, O. Pedersen, R. Gupta, S. Rasmussen, S. Besenbacher, A. D. Børglum, J. Wang, H. Eiberg, K. Kristiansen, S. Brunak, M. H. Schierup, Sequencing and de novo assembly of 150 genomes from Denmark as a population reference. *Nature* **548**, 87–91 (2017).

17. M. Nagasaki, J. Yasuda, F. Katsuoka, N. Nariai, K. Kojima, Y. Kawai, Y. Yamaguchi-Kabata, J. Yokozawa, I. Danjoh, S. Saito, Y. Sato, T. Mimori, K. Tsuda, R. Saito, X. Pan, S. Nishikawa, S. Ito, Y. Kuroki, O. Tanabe, N. Fuse, S. Kuriyama, H. Kiyomoto, A. Hozawa, N. Minegishi, J. Douglas Engel, K. Kinoshita, S. Kure, N. Yaegashi; ToMMo Japanese Reference Panel Project, M. Yamamoto, Rare variant discovery by deep whole-genome sequencing of 1,070 Japanese individuals. *Nat. Commun.* **6**, 8018 (2015).

18. Y. Okada, Y. Momozawa, S. Sakaue, M. Kanai, K. Ishigaki, M. Akiyama, T. Kishikawa, Y. Arai, T. Sasaki, K. Kosaki, M. Suematsu, K. Matsuda, K. Yamamoto, M. Kubo, N. Hirose, Y. Kamatani, Deep whole-genome sequencing reveals recent selection signatures linked to evolution and disease risk of Japanese. *Nat. Commun.* **9**, 1631 (2018).

19. K. Yoon, S. Lee, T. S. Han, S. Y. Moon, S. M. Yun, S. H. Kong, S. Jho, J. Choe, J. Yu, H. J. Lee, J. H. Park, H. M. Kim, S. Y. Lee, J. Park, W. H. Kim, J. Bhak, H. K. Yang, S. J. Kim, Comprehensive genome- and transcriptome-wide analyses of mutations associated with microsatellite instability in Korean gastric cancers. *Genome Res.* **23**, 1109–1117 (2013).

20. S. T. Sherry, M. H. Ward, M. Kholodov, J. Baker, L. Phan, E. M. Smigielski, K. Sirotkin, dbSNP: The NCBI database of genetic variation. *Nucleic Acids Res.* **29**, 308–311 (2001).

21. S. Moon, J. M. Akey, A flexible method for estimating the fraction of fitness influencing mutations from large sequencing data sets. *Genome Res.* **26**, 834–843 (2016).

22. T. G. Clark, T. Andrew, G. M. Cooper, E. H. Margulies, J. C. Mullikin, D. J. Balding, Functional constraint and small insertions and deletions in the ENCODE regions of the human genome. *Genome Biol.* **8**, R180 (2007).

23. A. Telenti, L. C. T. Pierce, W. H. Biggs, J. di Iulio, E. H. M. Wong, M. M. Fabani, E. F. Kirkness, A. Moustafa, N. Shah, C. Xie, S. C. Brewerton, N. Bulsara, C. Garner, G. Metzker, E. Sandoval, B. A. Perkins, F. J. Och, Y. Turpaz, J. C. Venter, Deep sequencing of 10,000 human genomes. *Proc. Natl. Acad. Sci. U.S.A.* **113**, 11901–11906 (2016).

24. I. Adzhubei, D. M. Jordan, S. R. Sunyaev, Predicting functional effect of human missense mutations using PolyPhen-2. *Curr. Protoc. Hum. Genet.* **76**, 7.20.1–7.20.41 (2013).

25. N. L. Sim, P. Kumar, J. Hu, S. Henikoff, G. Schneider, P. C. Ng, SIFT web server: Predicting effects of amino acid substitutions on proteins. *Nucleic Acids Res.* **40**, W452–W457 (2012).

26. H. J. Jin, K. D. Kwak, M. F. Hammer, Y. Nakahori, T. Shinka, J. W. Lee, F. Jin, X. Jia, C. Tyler-Smith, W. Kim, Y-chromosomal DNA haplogroups and their implications for the dual origins of the Koreans. *Hum. Genet.* **114**, 27–35 (2003).

27. H. J. Jin, C. Tyler-Smith, W. Kim, The peopling of Korea revealed by analyses of mitochondrial DNA and Y-chromosomal markers. *PLOS One* **4**, e4210 (2009).

28. M. Tanaka, V. M. Cabrera, A. M. González, J. M. Larruga, T. Takeyasu, N. Fuku, L. J. Guo, R. Hirose, Y. Fujita, M. Kurata, K. Shinoda, K. Umetsu, Y. Yamada, Y. Oshida, Y. Sato, N. Hattori, Y. Mizuno, Y. Arai, N. Hirose, S. Ohta, O. Ogawa, Y. Tanaka, R. Kawamori, M. Shamoto-Nagai, W. Maruyama, H. Shimokata, R. Suzuki, H. Shimodaira, Mitochondrial genome variation in eastern Asia and the peopling of Japan. *Genome Res.* **14**, 1832–1850 (2004).

29. Y. Wang, D. Lu, Y. J. Chung, S. Xu, Genetic structure, divergence and admixture of Han Chinese, Japanese and Korean populations. *Hereditas* **155**, 19 (2018).

30. D. Cusi, C. Barlassina, T. Azzani, G. Casari, L. Citterio, M. Devoto, N. Glorioso, C. Lanzani, P. Manunta, M. Righetti, R. Rivera, P. Stella, C. Troffa, L. Zagato, G. Bianchi, Polymorphisms of α-adducin and salt sensitivity in patients with essential hypertension. *Lancet* **349**, 1353–1357 (1997).

31. B. M. Psaty, N. L. Smith, S. R. Heckbert, H. L. Vos, R. N. Lemaitre, A. P. Reiner, D. S. Siscovick, J. Bis, T. Lumley, W. T. Longstreth Jr., F. R. Rosendaal, Diuretic therapy, the α-Adducin gene variant, and the risk of myocardial infarction or stroke in persons with treated hypertension. *JAMA* **287**, 1680–1689 (2002).

32. C. Bycroft, C. Freeman, D. Petkova, G. Band, L. T. Elliott, K. Sharp, A. Motyer, D. Vukcevic, O. Delaneau, J. O'Connell, A. Cortes, S. Welsh, A. Young, M. Effingham, G. McVean, S. Leslie, N. Allen, P. Donnelly, J. Marchini, The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).

33. J. MacArthur, E. Bowler, M. Cerezo, L. Gil, P. Hall, E. Hastings, H. Junkins, A. McMahon, A. Milano, J. Morales, Z. M. Pendlington, D. Welter, T. Burdett, L. Hindorff, P. Flicek, F. Cunningham, H. Parkinson, The new NHGRI-EBI catalog of published genome-wide association studies (GWAS catalog). *Nucleic Acids Res.* **45**, D896–D901 (2017).

34. T.-W. Kang, H.-J. Kim, H. Ju, J.-H. Kim, Y.-J. Jeon, H.-C. Lee, K.-K. Kim, J.-W. Kim, S. Lee, J. Y. Kim, S.-Y. Kim, Y. S. Kim, Genome-wide association of serum bilirubin levels in Korean population. *Hum. Mol. Genet.* **19**, 3672–3678 (2010).

35. Y. J. Kim, M. J. Go, C. Hu, C. B. Hong, Y. K. Kim, J. Y. Lee, J. Y. Hwang, J. H. Oh, D. J. Kim, N. H. Kim, S. Kim, E. J. Hong, J. H. Kim, H. Min, Y. Kim, R. Zhang, W. Jia, Y. Okada, A. Takahashi, M. Kubo, T. Tanaka, N. Kamatani, K. Matsuda; MAGIC consortium,

T. Park, B. Oh, K. Kimm, D. Kang, C. Shin, N. H. Cho, H. L. Kim, B. G. Han, J. Y. Lee, Y. S. Cho, Large-scale genome-wide association studies in East Asians identify new genetic loci influencing metabolic traits. *Nat. Genet.* **43**, 990–995 (2011).

36. S. McCarthy, S. Das, W. Kretzschmar, O. Delaneau, A. R. Wood, A. Teumer, H. M. Kang, C. Fuchsberger, P. Danecek, K. Sharp, Y. Luo, C. Sidore, A. Kwong, N. Timpson, S. Koskinen, S. Vrieze, L. J. Scott, H. Zhang, A. Mahajan, J. Veldink, U. Peters, C. Pato, C. van Duijn, C. E. Gillies, I. Gandin, M. Mezzavilla, A. Gilly, M. Cocca, M. Traglia, A. Angius, J. C. Barrett, D. Boomsma, K. Branham, G. Breen, C. M. Brummett, F. Busonero, H. Campbell, A. Chan, S. Chen, E. Chew, F. S. Collins, L. J. Corbin, G. D. Smith, G. Dedoussis, M. Dorr, A. E. Farmaki, L. Ferrucci, L. Forer, R. M. Fraser, S. Gabriel, S. Levy, L. Groop, T. Harrison, A. Hattersley, O. L. Holmen, K. Hveem, M. Kretzler, J. C. Lee, M. McGue, T. Meitinger, D. Melzer, J. L. Min, K. L. Mohlke, J. B. Vincent, M. Nauck, D. Nickerson, A. Palotie, M. Pato, N. Pirastu, M. McInnis, J. B. Richards, C. Sala, V. Salomaa, D. Schlessinger, S. Schoenherr, P. E. Slagboom, K. Small, T. Spector, D. Stambolian, M. Tuke, J. Tuomilehto, L. van den Berg, W. van Rheenen, U. Volker, C. Wijmenga, D. Toniolo, E. Zeggini, P. Gasparini, M. G. Sampson, J. F. Wilson, T. Frayling, P. I. de Bakker, M. A. Swertz, S. McCarroll, C. Kooperberg, A. Dekker, D. Altshuler, C. Willer, W. Iacono, S. Ripatti, N. Soranzo, K. Walter, A. Swaroop, F. Cucca, C. A. Anderson, R. M. Myers, M. Boehnke, M. I. McCarthy, R. Durbin; Haplotype Reference Consortium, A reference panel of 64,976 haplotypes for genotype imputation. *Nat. Genet.* **48**, 1279–1283 (2016).

37. S. Das, L. Forer, S. Schönherr, C. Sidore, A. E. Locke, A. Kwong, S. I. Vrieze, E. Y. Chew, S. Levy, M. McGue, D. Schlessinger, D. Stambolian, P.R. Loh, W. G. Iacono, A. Swaroop, L. J. Scott, F. Cucca, F. Kronenberg, M. Boehnke, G. R. Abecasis, C. Fuchsberger, Next-generation genotype imputation service and methods. *Nat. Genet.* **48**, 1284–1287 (2016).

38. Y. Dou, H. D. Gold, L. J. Luquette, P. J. Park, Detecting somatic mutations in normal cells. *Trends Genet.* **34**, 545–557 (2018).

39. K. Cibulskis, M. S. Lawrence, S. L. Carter, A. Sivachenko, D. Jaffe, C. Sougnez, S. Gabriel, M. Meyerson, E. S. Lander, G. Getz, Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat. Biotechnol.* **31**, 213–219 (2013).

40. T. S. Alioto, I. Buchhalter, S. Derdak, B. Hutter, M. D. Eldridge, E. Hovig, L. E. Heisler, T. A. Beck, J. T. Simpson, L. Tonon, A. S. Sertier, A. M. Patch, N. Jäger, P. Ginsbach, R. Drews, N. Paramasivam, R. Kabbe, S. Chotewutmontri, N. Diessl, C. Previti, S. Schmidt, B. Brors, L. Feuerbach, M. Heinold, S. Gröbner, A. Korshunov, P. S. Tarpey, A. P. Butler, J. Hinton, D. Jones, A. Menzies, K. Raine, R. Shepherd, L. Stebbings, J. W. Teague, P. Ribeca, F. C. Giner, S. Beltran, E. Raineri, M. Dabad, S. C. Heath, M. Gut, R. E. Denroche, N. J. Harding, T. N. Yamaguchi, A. Fujimoto, H. Nakagawa, V. Quesada, R. Valdés-Mas, S. Nakken, D. Vodák, L. Bower, A. G. Lynch, C. L. Anderson, N. Waddell, J. V. Pearson, S. M. Grimmond, M. Peto, P. Spellman, M. He, C. Kandoth, S. Lee, J. Zhang, L. Létourneau, S. Ma, S. Seth, D. Torrents, L. Xi, D. A. Wheeler, C. López-Otín, E. Campo, P. J. Campbell, P. C. Boutros, X. S. Puente, D. S. Gerhard, S. M. Pfister, J. D. McPherson, T. J. Hudson, M.

Schlesner, P. Lichter, R. Eils, D. T. W. Jones, I. G. Gut, A comprehensive assessment of somatic mutation detection in cancer using whole-genome sequencing. *Nat. Commun.* **6**, 10001 (2015).

41. S. Hiltemann, G. Jenster, J. Trapman, P. van der Spek, A. Stubbs, Discriminating somatic and germline mutations in tumor DNA samples without matching normals. *Genome Res.* **25**, 1382–1390 (2015).

42. M. Martin, Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet J.* **17**, 10–12 (2011).

43. H. Li, R. Durbin, Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).

44. R. Poplin, V. Ruano-Rubio, M. A. DePristo, T. J. Fennell, M. O. Carneiro, G. A. Van der Auwera, D. E. Kling, L. D. Gauthier, A. Levy-Moonshine, D. Roazen, K. Shakir, J. Thibault, S. Chandran, C. Whelan, M. Lek, S. Gabriel, M. J. Daly, B. Neale, D. G. MacArthur, E. Banks, Scaling accurate genetic variant discovery to tens of thousands of samples. *bioRxiv* 201178 (2017).

45. W. McLaren, L. Gil, S. E. Hunt, H. S. Riat, G. R. S. Ritchie, A. Thormann, P. Flicek, F. Cunningham, The Ensembl variant effect predictor. *Genome Biol.* **17**, 122 (2016).

46. A. L. Price, N. J. Patterson, R. M. Plenge, M. E. Weinblatt, N. A. Shadick, D. Reich, Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38**, 904–909 (2006).

47. S. Purcell, B. Neale, K. Todd-Brown, L. Thomas, M. A. R. Ferreira, D. Bender, J. Maller, P. Sklar, P. I. W. de Bakker, M. J. Daly, P. C. Sham, PLINK: A tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).

48. D. H. Alexander, J. Novembre, K. Lange, Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* **19**, 1655–1664 (2009).

49. M. van Oven, PhyloTree build 17: Growing the human mitochondrial DNA tree. *Forensic Sci. Int.-Gen. Supp. Ser.* **5**, e392–e394 (2015).

50. H. Weissensteiner, D. Pacher, A. Kloss-Brandstätter, L. Forer, G. Specht, H. J. Bandelt, F. Kronenberg, A. Salas, S. Schönherr, HaploGrep 2: Mitochondrial haplogroup classification in the era of high-throughput sequencing. *Nucleic Acids Res.* **44**, W58–W63 (2016).

51. L. Jostins, Y. Xu, S. McCarthy, Q. Ayub, R. Durbin, J. Barrett, C. Tyler-Smith, YFitter: Maximum likelihood assignment of Y chromosome haplogroups from low-coverage sequence data. *arXiv Preprint arXiv* 1407.7988 (2014).

52. H. Zhao, Z. Sun, J. Wang, H. Huang, J. P. Kocher, L. Wang, CrossMap: A versatile tool for coordinate conversion between genome assemblies. *Bioinformatics* **30**, 1006–1007 (2014).

53. H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin; 1000 Genome Project Data Processing Subgroup, The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).

54. J. G. Tate, S. Bamford, H. C. Jubb, Z. Sondka, D. M. Beare, N. Bindal, H. Boutselakis, C. G. Cole, C. Creatore, E. Dawson, P. Fish, B. Harsha, C. Hathaway, S. C. Jupe, C. Y. Kok, K. Noble, L. Ponting, C. C. Ramshaw, C. E. Rye, H. E. Speedy, R. Stefancsik, S. L. Thompson, S. Wang, S. Ward, P. J. Campbell, S. A. Forbes, COSMIC: The catalogue of somatic mutations in cancer. *Nucleic Acids Res.* **47**, D941–D947 (2019).

55. W. J. Kent, C. W. Sugnet, T. S. Furey, K. M. Roskin, T. H. Pringle, A. M. Zahler, D. Haussler, The human genome browser at UCSC. *Genome Res.* **12**, 996–1006 (2002).

56. A. Abyzov, A. E. Urban, M. Snyder, M. Gerstein, CNVnator: An approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. *Genome Res.* **21**, 974–984 (2011).

57. M. G. Csardi, Package. *igraph* **3**, 214–217 (2013).

58. D. R. Zerbino, P. Achuthan, W. Akanni, M. R. Amode, D. Barrell, J. Bhai, K. Billis, C. Cummins, A. Gall, C. G. Girón, L. Gil, L. Gordon, L. Haggerty, E. Haskell, T. Hourlier, O. G. Izuogu, S. H. Janacek, T. Juettemann, J. K. To, M. R. Laird, I. Lavidas, Z. Liu, J. E. Loveland, T. Maurel, W. McLaren, B. Moore, J. Mudge, D. N. Murphy, V. Newman, M. Nuhn, D. Ogeh, C. K. Ong, A. Parker, M. Patricio, H. S. Riat, H. Schuilenburg, D. Sheppard, H. Sparrow, K. Taylor, A. Thormann, A. Vullo, B. Walts, A. Zadissa, A. Frankish, S. E. Hunt, M. Kostadima, N. Langridge, F. J. Martin, M. Muffato, E. Perry, M. Ruffier, D. M. Staines, S. J. Trevanion, B. L. Aken, F. Cunningham, A. Yates, P. Flicek, Ensembl 2018. *Nucleic Acids Res.* **46**, D754–D761 (2018).

59. P. H. Sudmant, T. Rausch, E. J. Gardner, R. E. Handsaker, A. Abyzov, J. Huddleston, Y. Zhang, K. Ye, G. Jun, M. H. Fritz, M. K. Konkel, A. Malhotra, A. M. Stütz, X. Shi, F. P. Casale, J. Chen, F. Hormozdiari, G. Dayama, K. Chen, M. Malig, M. J. P. Chaisson, K. Walter, S. Meiers, S. Kashin, E. Garrison, A. Auton, H. Y. K. Lam, X. J. Mu, C. Alkan, D. Antaki, T. Bae, E. Cerveira, P. Chines, Z. Chong, L. Clarke, E. Dal, L. Ding, S. Emery, X. Fan, M. Gujral, F. Kahveci, J. M. Kidd, Y. Kong, E. W. Lameijer, S. McCarthy, P. Flicek, R. A. Gibbs, G. Marth, C. E. Mason, A. Menelaou, D. M. Muzny, B. J. Nelson, A. Noor, N. F. Parrish, M. Pendleton, A. Quitadamo, B. Raeder, E. E. Schadt, M. Romanovitch, A. Schlattl, R. Sebra, A. A. Shabalin, A. Untergasser, J. A. Walker, M. Wang, F. Yu, C. Zhang, J. Zhang, X. Zheng-Bradley, W. Zhou, T. Zichner, J. Sebat, M. A. Batzer, S. A. McCarroll; 1000 Genomes Project Consortium, R. E. Mills, M. B. Gerstein, A. Bashir, O. Stegle, S. E. Devine, C. Lee, E. E. Eichler, J. O. Korbel, An integrated map of structural variation in 2,504 human genomes. *Nature* **526**, 75–81 (2015).

60. V. Boeva, T. Popova, K. Bleakley, P. Chiche, J. Cappo, G. Schleiermacher, I. Janoueix-Lerosey, O. Delattre, E. Barillot, Control-FREEC: A tool for assessing copy number and allelic content using next-generation sequencing data. *Bioinformatics* **28**, 423–425 (2012).

61. E. J. Gardner, V. K. Lam, D. N. Harris, N. T. Chuang, E. C. Scott, W. S. Pittard, R. E. Mills; 1000 Genomes Project Consortium, S. E. Devine, The mobile element locator tool (MELT): Population-scale mobile element discovery and biology. *Genome Res.* **27**, 1916–1929 (2017).

62. L. Rishishwar, C. E. Tellez Villa, I. K. Jordan, Transposable element polymorphisms recapitulate human evolution. *Mob. DNA* **6**, 21 (2015).

63. 1000 Genomes Project Consortium, G. R. Abecasis, D. Altshuler, A. Auton, L. D. Brooks, R. M. Durbin, R. A. Gibbs, M. E. Hurles, G. A. McVean, A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061–1073 (2010).

64. A. Szolek, B. Schubert, C. Mohr, M. Sturm, M. Feldhahn, O. Kohlbacher, OptiType: Precision HLA typing from next-generation sequencing data. *Bioinformatics* **30**, 3310–3316 (2014).

65. F. F. González-Galarza, L. Y. C. Takeshita, E. J. M. Santos, F. Kempson, M. H. T. Maia, A. L. S. da Silva, A. L. Teles e Silva, G. S. Ghattaoraya, A. Alfirevic, A. R. Jones, D. Middleton, Allele frequency net 2015 update: New features for HLA epitopes, KIR and disease and HLA adverse drug reaction associations. *Nucleic Acids Res.* **43**, D784–D788 (2015).