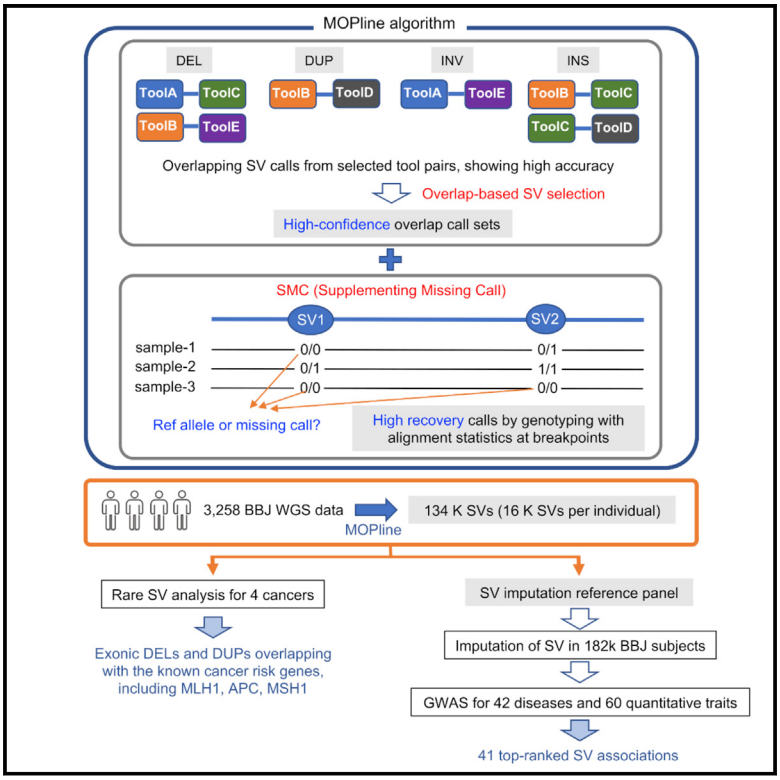
Cell Genomics

Technology

Detection of trait-associated structural variations using short-read sequencing

Graphical abstract



Highlights

- Development of MOPline to efficiently detect SVs from short-read WGS data
- MOPline detected 16,122 SVs per individual from 3,258 BBJ WGS datasets
- The BBJ SV panels were constructed to impute SVs in 181,622 Japanese individuals
- GWASs using the imputed SVs identified 41 top-ranked SVs associated with many traits

Authors

Shunichi Kosugi, Yoichiro Kamatani, Katsutoshi Harada, ..., Takayuki Morisaki, The BioBank Japan Project, Chikashi Terao

Correspondence

chikashi.terao@riken.jp

In brief

Kosugi et al. have developed MOPline, a structural variation (SV) detection tool. MOPline is flexible and scalable to accurately and sensitively detect SVs from short-read whole-genome sequencing (WGS) data by combining reliable SV call selection and missing call recovery algorithms. SVs detected by MOPline in 3,258 WGS datasets were subsequently imputed using 181,622 individual DNA microarray datasets. GWASs for 42 diseases and 60 quantitative traits with the imputed SVs identified top-ranked SV-trait associations for numerous complex traits.

Kosugi et al., 2023, Cell Genomics *3*, 100328
June 14, 2023
https://doi.org/10.1016/j.xgen.2023.100328

CellPress

Cell Genomics

CellPress
OPEN ACCESS

Technology

Detection of trait-associated structural variations using short-read sequencing

Shunichi Kosugi,[1,2] Yoichiro Kamatani,[3] Katsutoshi Harada,[1] Kohei Tomizuka,[1] Yukihide Momozawa,[4] Takayuki Morisaki,[5] The BioBank Japan Project, and Chikashi Terao[1,2,6,7,*]

[1]Laboratory for Statistical and Translational Genetics, RIKEN Center for Integrative Medical Sciences, 1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama, Kanagawa 230-0045, Japan

[2]Clinical Research Center, Shizuoka General Hospital, Shizuoka, Japan

[3]Department of Computational Biology and Medical Sciences, Graduate School of Frontier Sciences, The University of Tokyo, 5-1-5, Kashiwanoha, Kashiwa-shi, Chiba 277-8562, Japan

[4]Laboratory for Genotyping Development, RIKEN Center for Integrative Medical Sciences, 1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama City, Kanagawa 230-0045, Japan

[5]Division of Molecular Pathology, Institute of Medical Science, The University of Tokyo, 4-6-1, Shirokane-dai, Minato-ku, Tokyo 108-8639, Japan

[6]The Department of Applied Genetics, The School of Pharmaceutical Sciences, University of Shizuoka, Shizuoka, Japan

[7]Lead contact

*Correspondence: chikashi.terao@riken.jp

https://doi.org/10.1016/j.xgen.2023.100328

SUMMARY

Genomic structural variation (SV) affects genetic and phenotypic characteristics in diverse organisms, but the lack of reliable methods to detect SV has hindered genetic analysis. We developed a computational algorithm (MOPline) that includes missing call recovery combined with high-confidence SV call selection and genotyping using short-read whole-genome sequencing (WGS) data. Using 3,672 high-coverage WGS datasets, MOPline stably detected ~16,000 SVs per individual, which is over ~1.7–3.3-fold higher than previous large-scale projects while exhibiting a comparable level of statistical quality metrics. We imputed SVs from 181,622 Japanese individuals for 42 diseases and 60 quantitative traits. A genome-wide association study with the imputed SVs revealed 41 top-ranked or nearly top-ranked genome-wide significant SVs, including 8 exonic SVs with 5 novel associations and enriched mobile element insertions. This study demonstrates that short-read WGS data can be used to identify rare and common SVs associated with a variety of traits.

INTRODUCTION

Structural variation (SV) is generally defined as a mutation at least 50 bp larger than a short insertion or deletion (indel) and consists of deletions (DELs), insertions (INSs), duplications (DUPs), inversions (INVs), and translocations (TRA). SVs are the primary determinants of genomic variation at the species and individual level.[1–3] Because of the large size of SVs compared with SNVs and indels, SVs have much greater potential to alter gene function and gene regulation and modify coding regions, *cis*-regulatory regions, or stretches of topologically related domain sequences.[4–7] Thus, SVs have been implicated in many human diseases, including neurodevelopmental disorders and cancer, as well as in differences in gene expression between individuals.[8–11]

Although effective detection of SVs remains challenging because of their large size and variety,[1,12] sequencing-based and array-based-methods have been developed. Sequencing-based methods are more sensitive to detect SVs and have higher resolution to determine breakpoints (BPs) compared with array-based methods. Recent advances in single-molecule sequencing and linked-read sequencing technologies that generate long reads (LRs) have enabled more efficient detection of SVs than sequencing of short reads (SRs) because LRs can span more SVs and repeat genomic regions than SRs.[13–15] Comprehensive SV detection has been performed on several pilot human samples using multiple sequencing platforms, including SR or LR whole-genome sequencing (WGS), as represented by the Genome in a Bottle (GIAB) consortium and the Human Genome Structural Variation Consortium (HGSVC).[16–21] SV detection with LR WGS data generates approximately about three times as many SVs per individual (~24,000) as detection with short-lead WGS data (~7,500). However, the high cost of LR sequencing and the high demands on the quality and quantity of input DNA make SR-based SV detection an effective method, especially for multiple samples. Large-scale SV identification with 1,000–10,000 SR WGS datasets has been reported on a population scale in the 1000 Genomes (1KG) Project,[2] the gnomAD project,[22] the NHGRI Centers,[23] the Human Genome Diversity Project (HGDP),[16] and the HGSVC.[24]

The computational algorithm for SV detection based on SRs employs a basic method with multiple alignment signals, such as read pair (RP), split read (SP), read depth (RD), and assembly (AS).[1,13,25,26] These indirect alignment signals depend on SV detection algorithms, SR preparation, and read alignment methods and often lead to misassignment to SVs and false negative calls because of ambiguous and imprecise alignment of SRs.[27] Thus, current computational methods are hampered from stable and accurate detection of SVs, and no single algorithm can accurately and sensitively detect all types of SVs.[28,29] Many projects use multiple algorithms to call SVs and then merge the output to increase accuracy and/or recall.[2,14–16,22,23,30–37] In our previous systematic evaluation of overlapping SV calls, some specific pairs of algorithms, but not combinations of methods used in the algorithm (i.e., RP, SP, RD, or AS), showed higher accuracy with specific SV types and size ranges compared with other pairs.[29] Therefore, careful selection of overlap calls is necessary to improve SV detection accuracy.

We developed MOPline, a computational algorithm that iteratively merges optimized overlapping calls from multiple algorithms in each SV category to increase precision and genotype reference alleles of all samples at SV sites to increase recall. We detected ~16,000 SVs per individual from 414 high-coverage 1KG WGS data and 3,258 BioBank Japan (BBJ) WGS data (Figure 1A). A method called supplementing missing calls (SMC), which restores missing variants, increased high reliability variants by 42%. The SR-based SVs detected included many LR-based SVs in non-repeat regions despite depleting many LR-based SVs detected in repeat regions. BBJ SVs include many rare coding SVs that disrupt known and potential novel disease risk genes and common SVs in high linkage disequilibrium with published disease genome-wide association study (GWAS) variants. SV genotypes of BBJ 181,622 individuals were imputed using a reference panel containing the BBJ SVs (Figure 1B). GWASs for binary and quantitative traits using these imputed genotypes showed that many genome-wide significant loci contained SVs, some of which were likely causal variants.

DESIGN

Overlap calls between multiple SV detection algorithms show overall high accuracy, but the degree of precision and recall of overlap calls depends on the combination of algorithms, as in our previous study.[29] Using WGS data from NA12878, the precision and recall of overlap calls were systematically determined with various combinations of SV detection algorithms (STAR Methods). The precision and recall of overlap calls varied significantly between pairs from four or six randomly selected algorithms for each type of SV (Figures S1 and S2). These results indicate that selecting the appropriate pairs and number of algorithms is necessary to achieve a good balance between precision and recall for SV. We expected that repeated merging of overlapping calls obtained only from "good" pairs of algorithms that exhibited high precision would yield a highly reliable set of SV calls. We named this merging method MOP (merging overlap calls from selected pairs of algorithms).

Using the MOP merging method and a recovery method for variants missed through MOP-based selection, we developed a computational algorithm (MOPline) for efficient detection and genotyping of SVs with high reliability. MOPline is flexible enough to incorporate user-specified algorithms, but here we use a version of MOPline (MOPline-7t) that uses seven existing algorithms. An overview of MOPline-7t is given in Figure 1A (see Table S1 for the algorithms used in the other MOPline derivatives). MOPline uses single and multiple WGS datasets from human and non-human species, and the process consists of five steps: (1) selecting and merging overlap calls from pre-selected pairs of algorithms, (2) joint calling of multiple SV call sets, (3) genotyping, (4) SMC, and (5) adding annotation and filtering. The first step selects high-confidence SV calls for each sample, and the fourth step recovers missing calls from joint-called SV data. The SMC step is coupled with SV genotyping, which is based on multinominal logistic regression with alignment statistics of RD and split reads. SMC increased the median number of SV calls per sample by about 1.4–2.3 times, although the total number of SV sites did not change. See STAR Methods for a more detailed explanation of the MOPline algorithm.

RESULTS

The MOPline algorithm efficiently detects and genotypes reliable SVs from single or multiple SR WGS datasets

To evaluate the performance of MOPline, we determined the precision and recall of the SV calls detected with MOPline using multiple real WGS datasets and reference SV datasets (STAR Methods; see Table S2 for reference SV sets) and compared them with those determined for 32 existing SR-based SV detection algorithms, which showed high precision and recall in a previous study.[29] MOPline achieved superior performances in almost all categories except INV (see Figure 2 for NA12878, Figure S3 for all datasets, and Figures S4–S7 for DELs and DUPs in three size ranges). Note that the lower recall value for NA19240 than for NA12878 is due to the difference in the number of reference SVs for each (Table S2). The simple merge method (Simplemerge-7t; simple merging of SV call sets from 7 tools), in contrast, showed a significantly lower level of precision despite a high level of recall (Figure 2). MOPline with other combinations of 4–14 algorithms also showed similarly high levels of precision and recall in almost all cases (Figures S3–S7). These results suggest that six to nine algorithms are required for high-quality detection of all types and sizes of SVs with MOPline. In the evaluation of SV genotyping, the accuracy of the MOPline genotyping for DELs, INSs, and INVs were shown to be among the best among the existing SR-based SV genotyping algorithms (Figure S8). In particular, the superiority of MOPline over INS genotyping seems high because graph-based genotyping algorithms, such as BayesTyper, GraphTyper2, and Paragraph, require INS sequences for INS genotyping. In addition, the SMC function increased true positive (TP) calls by 42%, as explained in the next section.

The ensemble SV detection pipelines GATK-SV and sv-pipeline (svtools), which combine the results of five SV detection algorithms and three SV detection algorithms, respectively, have

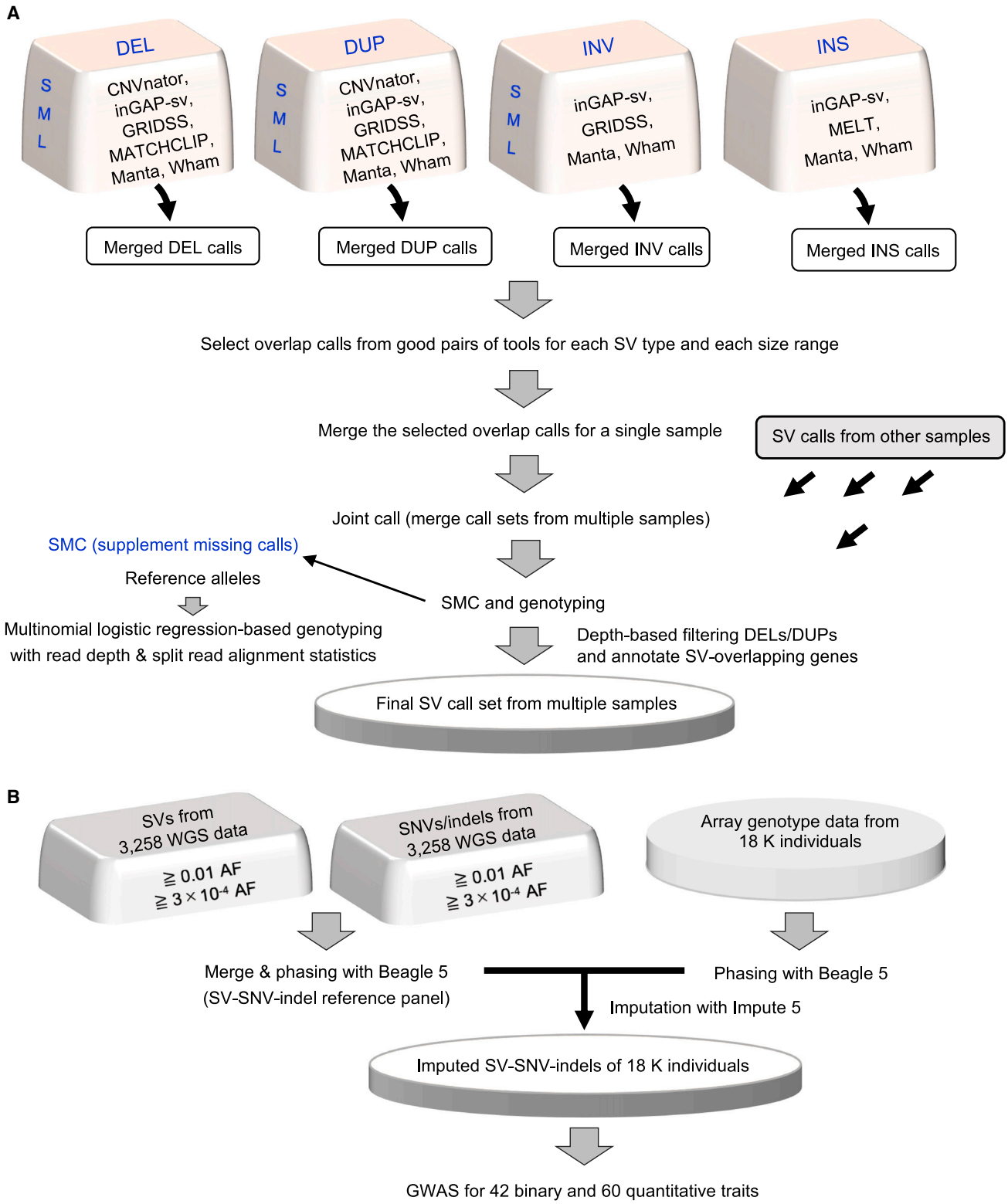


Figure 1. Overview of the framework for SV detection and SV imputation

(A) Overview of MOPline-7t. MOPline-7t is a derivative of MOPline and uses seven SV detection algorithms, six for DEL and DUP and four for INS and INV, as indicated. The algorithm used can be selected flexibly according to the research purpose and environment. The SV call sets for each SV type are simply merged without redundancy (merged SV calls). Then, from the merged SV calls, overlap calls from pre-selected algorithm pairs are selected for each SV type and size

(legend continued on next page)

been used in large-scale SV detection projects.[22,24,36] To fairly compare the SV calling performance of these pipelines and MOPline, we detected SVs with these tools using 100 high-coverage WGS datasets from the 1KG project (see STAR Methods for details). MOPline outperformed the other two pipelines in number of calls per sample and TP calls for NA12878 while maintaining almost 95% precision (Figures S9A, S9J, and S9K). The sv-pipeline had many DELs, DUPs, and INSs that were not shared with the other pipelines (Figures S9F–S9H), likely because of the high number of false-positive DEL/DUP calls and low-frequency INSs (Figures S9A, S9I, and S9K). For DELs/INSs specific to each pipeline in NA12878, many were located in the simple tandem repeat (STR) region, and there were significantly more SVs specific to MOPline compared with the SVs of the other pipelines (Figure S9I). The run time of the MOPline SV calling step, which runs seven external SV detection algorithms, was 29 h/sample using 2 CPU cores, while the run time of GATK-SV was 20 h/sample, consuming a large amount of memory (146 GB) and CPU (80 cores) (Figure S9L). The run time and memory usage for the MOPline merging and post-merging steps were significantly less than those of the sv-pipeline (Figures S9M and S9N).

SV detection from high-coverage 1KG WGS data with MOPline

We then evaluated MOPline's performance with 414 multiethnic high-coverage (30×) WGS datasets from 1KG. The total number of SVs detected was 98,393, and the median number of SVs per individual (14,575) was approximately 1.6–3.3 times higher than previous results using WGS data from diverse populations[22–24] (Table 1). Comparison between MOPline 1KG-SVs, gnomAD-SVs, and HGSVC-SVs (data corresponding to the same 1KG samples used in this study) indicated that SVs unique to this study were abundant in the STR region (Figure S10). Because of the higher mutation rate of STR,[38] the average AF of SVs in STRs was significantly higher than in non-repeat regions (Figures S10D and S10E), and the higher number of SVs per individual in this study may be in part due to the enhanced detection of SVs in STRs.

Consistent with the previous observations,[2,22] the total number of SVs detected in Yoruba in Ibadan, Nigeria (YRI) was the highest of any SV type (Figure S11A), reflecting the high number of SVs specific to YRI (Figure S11B). Allele frequencies at each SV site showed moderate variation among populations, with the highest agreement between Han Chinese in Beijing, China (CHB) and Japanese in Tokyo, Japan (JPT) (mean, 0.99) and lower agreement between YRI and other populations (0.82–0.83) (Figures S11C and S11D). Principal-component analysis (PCA) using the detected SVs clustered the four populations into three structures overlapping CHB-JPT (Figure S11E), consistent with the results using tag SNPs (Figure S11F). These results indicate that MOPline faithfully captures the differences in genetic architecture composed of SVs in multiple populations. In addition, compared with the result of single sample calls for NA12878, NA12878 included in Utah residents with ancestry from northern and western Europe (CEU) had 42% more TP calls and less loss of precision (Figures S11G and S11H), indicating high efficacy of the MOPline SMC function in SV call recovery.

We were interested in high-frequency (HF) SVs observed in any population that should be integrated into future human references. We found 3,068 SVs with site frequencies greater than or equal to 0.9 in any population (Table S3). Over 70% of INSs and DUPs overlapped with repeat regions including STRs and segmental DUPs. These SVs were validated with NA12878 PacBio CCS LR alignment data and the IGV viewer and showed a high validation rate of 93%; for HF DELs, 252 were homozygous DELs in ≧90% of 414 individuals. Many of the HF SVs found in GRCh37 were also observed in GRCh38. Approximately 64% of these 252 homozygous DELs matched the Alu elements annotated in the GRCh37 reference, with a size of approximately 300 bp. Reducing the proportion of homozygotes in 414 individuals decreased the Alu content but not the content of other types of retroelements (Figure S12). These results suggest that the human reference sequence (i.e., GRCh37) contains low-frequency INSs, including Alu INSs, which have recently been transposed in some individuals derived from the reference, resulting in detection of HF DELs in many individual genomes.

The SR-based MOPline results (SR-SVs) were compared with LR-based SV calling data (LR-SVs) from the recently reported LR-based, haplotype-resolved HGSVC SV set.[17] Approximately 35% of the 1KG SR-SVs overlapped with repeat regions, including STRs and segmental DUPs (Figure S13A). In contrast, more than 70% of LR-SVs overlapped with repeat regions (Figure S13B). About 40% of the total number of SR-SVs and LR-SVs were shared with each other (Figures S13C and S13E), and the shared proportion of SR-SVs increased to 75% when AF was limited to ≧ 0.05 (Figure S13D). For SVs specific to SR-SVs, approximately 60% were located in non-repeat regions (Figure S13C). In contrast, only 6%–8% of LR-SV-specific SVs were located in non-repeat regions (Figure S13E). These results indicate that most of LR-SVs located in non-repeat regions are common to SR-SVs. In addition, SR-DELs tend to contain more SVs that are ≧ 10 kb in length than LR-DELs (Figure S13G), which is consistent with the results detected with the NA12878 LR data (Figures S13H–S13L), although some of the large DELs we confirmed may contain false positives because of ambiguous alignments in segmental DUPs.

range. The calls selected for each SV category are merged into one SV call set for each sample, and SR alignment statistics on coverage and soft-clipped ends for each SV site are added to the vcf file. SV call sets from multiple samples are merged into a single vcf file (joint call). SVs from each sample are genotyped based on multinominal logistic regression with the read alignment statistics. Reference alleles for missing calls are regenotyped by the genotyping-coupled SMC function. Finally, DELs/DUPs with inconsistent characteristics related to read coverage or other read alignment signals are filtered, and SVs are annotated for their overlapping genes.

(B) Overview of SV imputation for GWASs performed in this study. To create reference panels for imputation, the BBJ-SV dataset detected from 3,258 WGS datasets using MOPline-7t was integrated with SNVs/indels detected from the same WGS data. In the integrated data, SV genotypes, except INVs, were converted to pseudo-SNP genotypes at the first BP of SV. Two different reference panels with ≧ 0.01 AF and ≧ 0.0003 AF were generated after phasing with Beagle 5. Array-based SNP genotype data of 181,622 BBJ samples were imputed with the reference panels using Impute 5, and imputed genotypes with an INFO score of 0.3 or higher were used for GWAS.

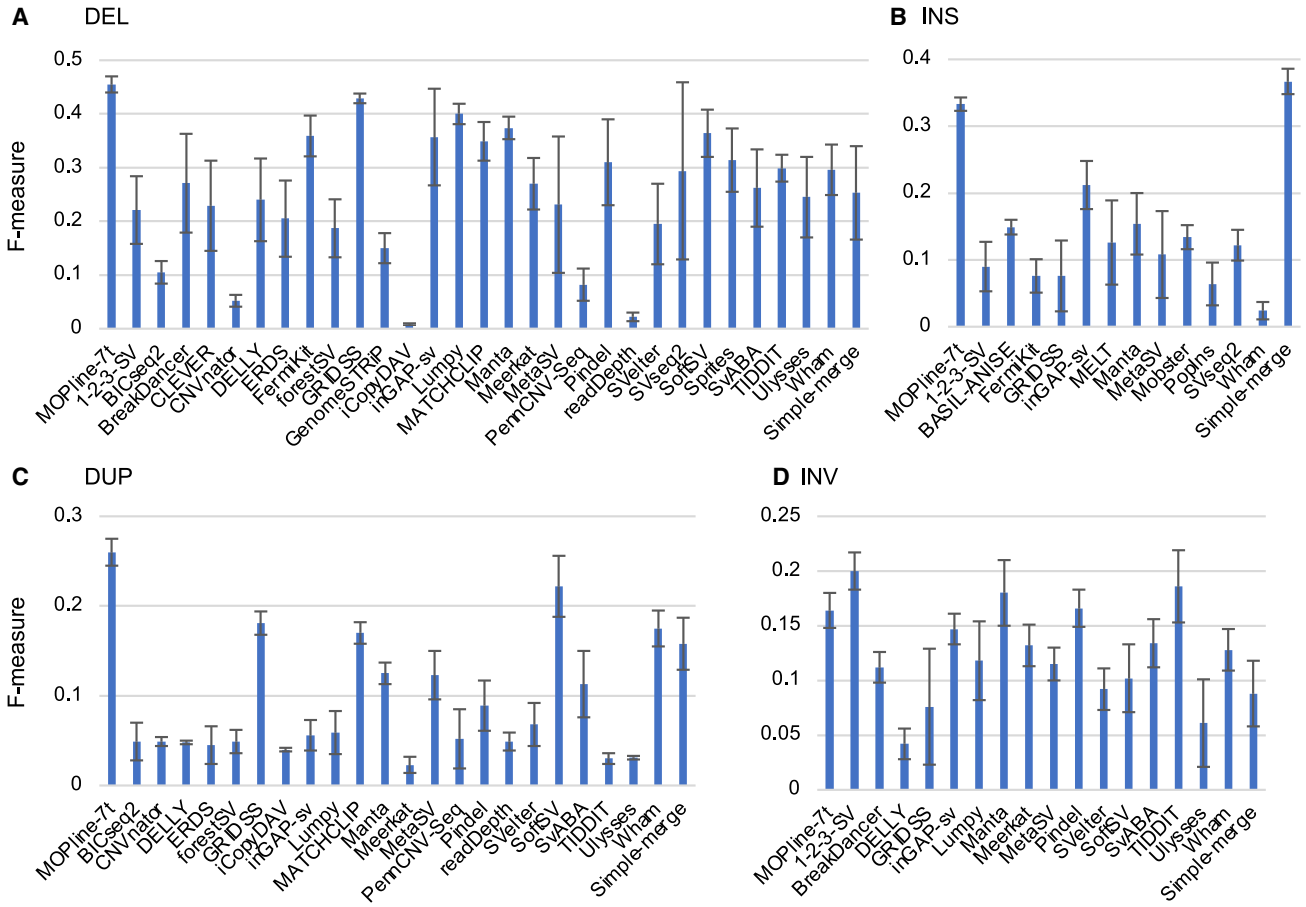


Figure 2. MOPline achieves higher accuracy in SV detection than existing SV detection algorithms

Using five NA12878 WGS datasets (data1–data5), we called DELs (A), INSs (B), DUPs (C), and INVs (D) with the indicated SV detection algorithms, including MOPline-7t. F-measures were determined for each SV type. The mean values of the F-measures are indicated by blue bars. Error bars represent standard error.

Robust and reliable detection of SVs from 3,258 BBJ WGS datasets with MOPline

SVs were called with MOPline-7t using 3,258 high-coverage WGS datasets from BBJ project participants. We detected a total number of 133,841 SVs, including 55,284 DELs, 15,222 DUPs, 61,750 INSs, and 1,585 INVs (Table 1; hereafter, these SVs are called BBJ-SVs), of which 51,087 (38%) were singleton SVs. The median number of SVs per individual was 16,122, which is approximately 1.7- to 3.7-fold higher than previous results with SR WGS data from various populations. Statistical evaluation using autosomal common BBJ-SVs (AF ≧ 0.01, excluding SVs in repeat regions) showed statistics similar to the gnomAD-SV study[22] (Figures 3A–3D). The median peak *r2* of linkage disequilibrium (LD; 0.81) was comparable with 0.85, a value reported in the gnomAD-SV study. The common SVs that matched the gnomAD-SVs showed a median peak LD *r2* of 0.93 (Figure 3B). 91% of the SVs in the BBJ were in Hardy- Weinberg equilibrium (HWE) (Figure 3C), which is slightly higher than the reported value (86%) in the gnomAD-SV study. The overall AF concordance for the same East Asian (EAS) population between the two studies was 0.94 (Pearson correlation coefficient) (Figure 3D). About 80% of the common EAS gnomAD-SVs matched the common BBJ-SVs, whereas only 54% of the common BBJ-SVs were shared with the EAS gnomAD-SVs. Validation in the 1KG-SVs using NA12878 PacBio CCS LR alignment data showed 93%–98% precision for any SV types except for INVs (Figure 3E). In the analysis of HWE and LD, the genotyping quality of DUPs was lower than the other types of SVs, mainly because of the inefficient ability to determine the copy number allele state of DUPs. However, the presence of detected DUPs was confirmed with high precision by validation with LR data. These results indicate that MOPline provides high-quality detection and genotyping of SVs.

The distribution of AF of SVs detected in the 3,258 samples was similar to the distribution of the SNVs/indels detected in the same datasets (Figures 3F and 3G). DELs, DUPs, and INVs tended to increase in size with lower AF (Figure S14), consistent with previous studies.[22,23] The numbers of SVs detected per sample deviated considerably less between samples than with a single algorithm despite the WGS data having been generated with different sequencing libraries and platforms (Figures S15 and S16), indicating robustly stable SV detection by MOPline. MOPline detected DELs derived from Alu and LINE1 mobile elements, corresponding to an HF peak around 300 bp and a low-frequency peak of ~5–6 kb, respectively (Figure 3H). The proportion of SVs,

Table 1. SVs detected using SR WGS data in current and previous studies

Study	Sample size	Coverage[a]	Total/per individual[b]	Total SVs	DELs	DUPs	INVs	INSs (MEIs)	Other types[c]
This study	414 (1KG)	37.6×	total	98,393	41,406	12,606	1,590	42,791 (21,344)	0
			per individual	14,575	5,244	1,714	133	7,484 (1,741)	0
	3,258 (BBJ)	27.9×	total	133,841	55,284	15,222	1,585	61,750 (20,919)	0
			per individual	16,122	5,511	2,160	108	8,343 (1,544)	0
gnomAD[d]	12,653	32×	total	335,470	172,637	47,463	788	109,278 (77,582)	5,304
			per individual	7,439	3,725	1,051	14	2,612	37
CCDG[e]	17,795	>20×	total	241,031	N/A	N/A	N/A	N/A	N/A
			per individual	4,442	(35%)	(11%)	N/A	(27%)	N/A
HGSVC/ 1KG c[f]	3,202	34×	total	173,355	90,259	28,242	920	49,693 (34,828)	4,241
			per individual	9,452	4,075	1,184	68	3,530 (1,194)	595
	414[g]	37.6×	total	71,110	33,929	10,714	471	24,047 (14,356)	1,949
			per individual	9,019	3,931	1,089	65	3,449 (1,169)	485

[a]Mean coverage per sample.
[b]Median of SVs per individual.
[c]TRA, CNV, or complex SV.
[d]Collins et al.[22]
[e]Abel et al.[23]
[f]Byrska-Bishop et al.[24]
[g]GRCh37-based data for 414 samples matching the 1KG data used in this study.

except for INSs, overlapping exons of protein-coding genes was higher than SNVs/indels (Figure 3I). The number of SVs per individual overlapping exon of protein-coding genes was 30%–40% of the sum of loss-of-function (LoF) SNVs and indels (Figure 3J), comparable with the gnomAD-SV study.[22] We examined the pLI scores[39] of the SV-overlapping genes. DEL- and DUP-overlapping genes with pLI ≧ 0.9 per individual were rare and comparable with ≧ 0.9 pLI genes with LOFTEE LoF SNVs (Figure 3K). SVs overlapping with ≧ 0.99 pLI genes were rich in rare variants with AF < 0.001 (odds ratio [OR] = 2.1, $p = 4 \times 10^{-9}$, Fisher's exact test) (Figure 3L). In addition, four and two knockouts of protein-coding genes caused by compound heterozygous exonic SVs and SVs-SNVs, respectively, were detected (Table S4). We further examined the constraint of the BBJ DELs and DUPs on annotated coding and noncoding regions. In DELs and DUPs, the phastCons super-conserved regions, the VISTA experimentally determined enhancer regions, and the coding and the proximal 5′-flanking regions (≦ 1 kb upstream of the first exon) of the high-pLI genes were highly constrained (Figure 3M; Figure S17), in agreement with previous studies.[22,23] For DELs/DUPs, protein-coding genes and noncoding genes were constrained at a similar level, suggesting that noncoding genes also play an important role in evolution.

BBJ-SVs include rare and common SVs for disease risk

The BBJ WGS data used were obtained from patients with any of seven diseases, including four cancers, and with different Illumina sequencing platforms between some sample sets (STAR Methods). These were individuals with coronary artery disease (CAD; n = 1,964), drug eruption (n = 189), colorectal cancer (n = 196), breast cancer (n = 237), prostate cancer (n = 215), gastric cancer (n = 257), and dementia (n = 200). Despite possible bias because of differences in sequencing platforms and small sample sizes, we performed a gene-based burden test and searched the ClinVar database for known disease risk genes with overlapping exons with SV. As controls, 2,353 non-cancer samples (361 female samples for breast cancer [BrCa] and 1,992 male samples for prostate cancer [PrCa]) were used, and 3,058 controls for dementia. We found that 17 known cancer risk genes deposited in ClinVar, including *MLH1*, *MSH2*, *APC*, *ATM*, and *BRCA1*, overlapped SVs with coding exons in the corresponding cancer patient samples (Table 2). These exonic SVs included all six pathogenic SVs found in the cancer gene panel in our previous study[40] and five additional pathogenic SVs not found in that study. Two dementia samples contained a 16-kb coding DEL at chr5:88114001 that overlapped the exons of the psychiatric disease gene *MEF2C*. All of these coding risk SVs were rare SVs with AF ≦ 0.002 and heterozygous except for the 173-kb DUP of the *RNF43* gene. Rare exonic SVs overlapping the known disease risk genes were enriched in case samples (21:3; OR, 7.7; $p = 1.9 \times 10^{-4}$, Fisher's exact test). Some genes not deposited in ClinVar did not show statistical significance for many SVs because of limited sample size and low frequency of associated SVs but, in certain disease groups, had enriched coding SVs (Table S5). Of the 16 SVs observed in multiple samples, 10 were also found in gnomAD-SVs, most of which were rare SVs with less than 0.001 AF, and 2 SVs were specific to EAS populations (Table S5). Overall, all the SVs found were rare SVs that were enriched in the corresponding disease case samples, although they did not reach the Bonferroni-corrected threshold because of small sample sizes. We also found the common SVs, including 12 DELs and 8 INSs that were in a strong LD with published GWAS signals, to be associated with these diseases (Table S6). The 6.8-kb DEL in strong LD with the PrCa GWAS SNP was a coding SV overlapping the *LILR3* exon, suggesting that this SV is a candidate for a causal variant. The associated INSs were enriched for Alu INSs (7/8, 87.5%), compared with Alu

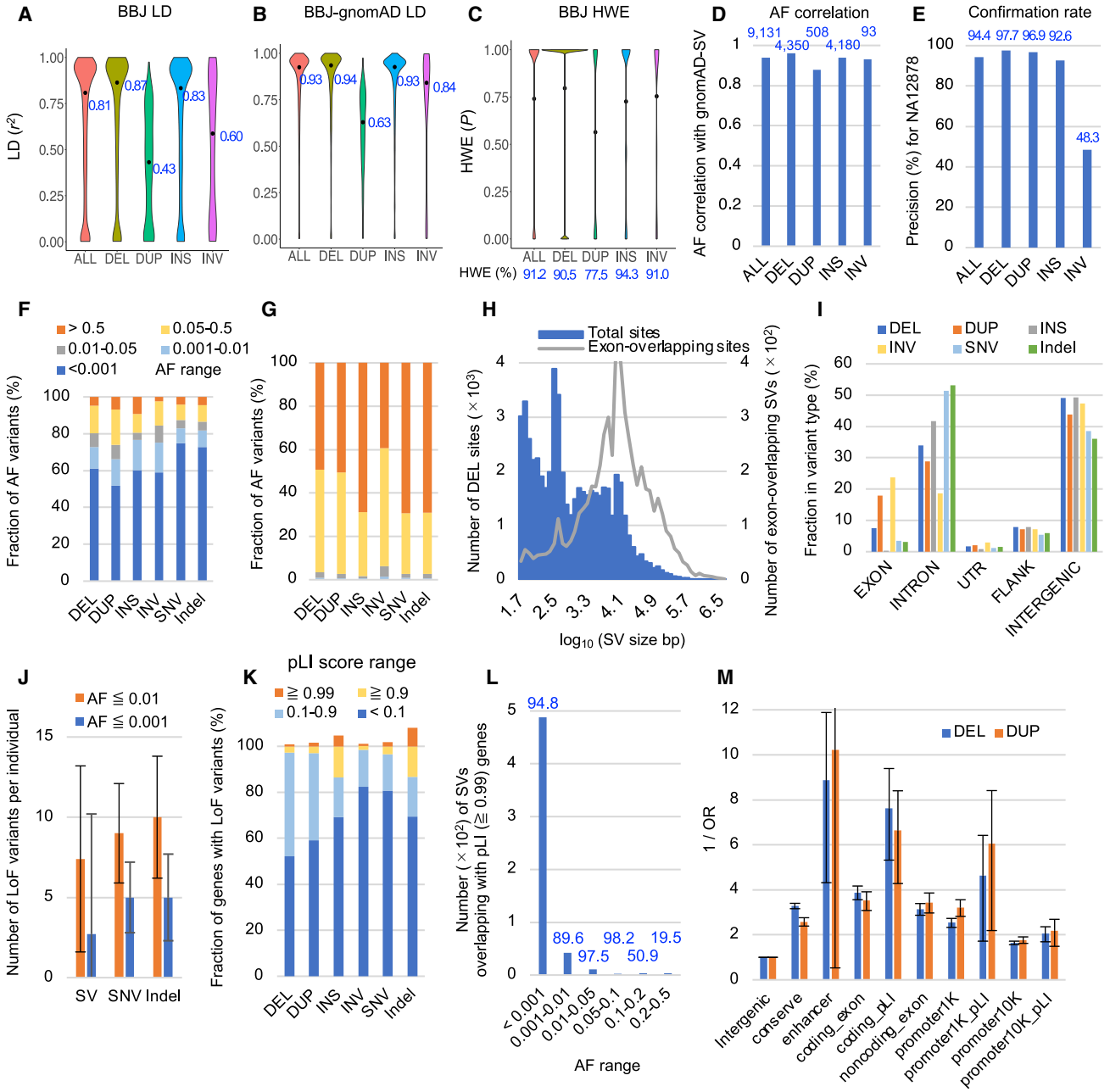


Figure 3. Quality and characteristics of SVs detected by MOPline from 3,258 BBJ WGS datasets

(A) Linkage disequilibrium (LD) between the BBJ-SVs and their neighboring SNVs or indels. The top LD correlation coefficients (*r2*) of each autosomal common (AF ≧ 0.01) SV site are plotted for the indicated SV type, with median values indicated by dots and blue text. A total of 14,310 SVs were tested, excluding SVs overlapping with repeat regions (see STAR Methods for details).

(B) LD between BBJ-gnomAD SVs and their neighboring SNVs or indels. We selected 7,440 BBJ-SVs overlapping with the common (AF ≧ 0.01) gnomAD-SVs found in both European and African populations, excluding those overlapping with repeat regions. The plot is the same as in (A).

(C) Hardy-Weinberg equilibrium (HWE) of the BBJ-SVs. p values obtained by HWE test for all autosomal BBJ-SVs are plotted for the indicated SV types. The mean p values are indicated by dots. HWE (percent) represents the percentage of SVs for which the HWE p value is greater than the Bonferroni-corrected p value.

(D) Correlation of AF of SVs between BBJ-SVs and gnomAD-SVs. We selected common SVs (AF ≧ 0.01, excluding SVs that overlap repeat regions) that matched between BBJ-SVs and the gnomAD-SVs. Matched SVs were selected based on a BP distance of ≦ 200 bp for INSs and ≧ 50% reciprocal overlap for the other types. Pearson correlation coefficients between AF of BBJ SVs and EAS-AF of gnomAD-SVs were calculated for all SVs and for each SV type. SV counts for each SV type are indicated on each bar.

(E) Confirmation rate of NA12878 SVs detected with MOPline. The SVs (N = 11,851) of NA12878 contained in the 1KG CEU SVs were evaluated using two reference datasets of NA12878 (the DGV-based and LR-based SV sets). The false-positive calls evaluated were further confirmed with alignment bam files

(legend continued on next page)

INSs (34.3%) in INSs with AF ≧ 0.01 (OR = 13.4, 95% confidence interval [CI] = 1.7–109, $p = 3.1 \times 10^{-3}$, Fisher's exact test; 34.3% of INSs were estimated to be Alu; see STAR Methods for details). Taken together, these results indicate that BBJ-SVs can identify disease risk SVs, including coding rare SVs and noncoding common SVs tagged with GWAS-significant SNVs.

SV imputation using BBJ-SVs

Because SVs and SNVs are associated with complex traits and show genetic linkages, we were motivated to construct an imputation reference panel for SVs, combining BBJ-SVs and SNVs/indels (Figure 1B). Using this reference panel, we imputed array genotype data for 181,622 BBJ samples (see details in STAR Methods), which included the genotype data used in our recent study.[41] SVs had a lower imputation quality than SNVs in the lower AF range, but almost 90% were imputed across all AF ranges (Figure 4). To examine how faithfully SVs are imputed, we used the 200 test samples randomly selected from 3,258 WGS samples for which the reference panel was created and imputed the array genotype data of the selected test samples using the reference panel of the remaining 3,058 samples to calculate precision and recall. We achieved overall precision (86%) and recall (72%) of SV imputation (Figures S18A–S18E). We then validated the false positives of imputed SVs from a selected single sample using the IGV viewer with the bam alignment file of the sample because MOPline could insufficiently genotype for some samples to generate false-negative calls. This validation confirmed the alignment signals associated with SVs in about 70% of the false positives, leading to overall precision (95%) and recall (80%) of SV imputation (Figure S18F).

GWAS with imputed SVs identifies a number of associated genes

GWAS was performed using the imputed data containing SVs from BBJ 181,622 samples related to 42 diseases and 60 quantitative traits.[42–45] We were interested in whether imputed SVs could be identified as causal variants of previously identified or novel GWAS loci. We identified 360 and 1,796 genome-wide significant SV associations for 26 diseases and 55 quantitative traits, respectively (Table S7). For the binary disease traits, 11 loci contained significant SVs with p values comparable with the top SNPs (Figures 5A and S19; Table S8). The DEL of 253 bp for hyperlipidemia (DPL) was more significant than the top SNP in p value and effect size (Figure 5B). The DEL of 30 kb for BrCa was the previously reported common risk DEL for BrCa (Figure 5C), which deletes the N-terminal region of the APOBEC3B protein and the C-terminal region of the APOBEC3A protein.[46] The Alu INS for type 2 diabetes (T2D) was also associated with body mass index (BMI) (Figure 5D) in the known locus[47] and reported to increase expression by 1.7-fold in an ectopic luciferase reporter assay.[48] Another Alu INS for PrCa (Figure 5E) was also a previously reported Alu INS associated with PrCa,[49] which showed a 2.7-fold increase in expression in a reporter assay.[48] Because the genetic component of PrCa, which is highly heritable, would be useful in predicting future mortality (C.T., unpublished data), we conducted survival analysis focusing on the Alu INS of PrCa. We found that homozygous carriers of this PrCa-associated INS who had not developed cancer at enrollment had a higher mortality from PrCa than non-carriers (hazard ratio = 2.43, 95% CI = 1.27–4.65, p = 0.0074, in the Cox proportional hazards model) (Figure 5M).

For quantitative traits, we found a total of 21 top and 29 nearly top associations with 22 traits (Figures 5 and S20; Table S8). Many of the top-ranked GWAS SVs colocalized with array SNPs that are in high LD with the SVs and have strong association signals similar to those of the SNPs (Table S9). The mean effect size (0.72, N = 6) of the top GWAS SVs in the rare/low-frequency range (AF < 0.01) were larger than the effect size (0.34, N = 227) for the top SNVs/indels in the same AF range. Notably, six DELs and one DUP were in exons, probably causal LoF SVs (Figures 5I, 5J, and 5L; Table S8). The associations with height, hemoglobin A1c (HbA1c), mean corpuscular hemoglobin concentration (MCHC),

generated with NA12878 CCS PacBio LRs and Illumina SRs using IGV viewer. SV calls with alignment evidence were judged as TPs. The percentage of TP calls for all SVs and each SV type is shown with bars and blue letters.

(F) Percentage of total variants stratified by AF for each type of variant. The percentage of variants in the indicated AF ranges is shown in each color bar. Variants (%) with AF ranges of >0.5, 0.05–0.5, 0.01–0.05, 0.001–0.01, and <0.001 are shown as orange, yellow, gray, light blue, and blue bars, respectively.

(G) Percentage of variants per individual stratified by AF. The color bars corresponding to the indicated AF ranges are the same as in (F).

(H) Size ranges for DELs. The number of DELs in the size ranges indicated on the x axis is indicated by blue bars on the left y axis. The number of DELs overlapping with exons of 20,268 protein-coding genes is indicated by gray lines on the right y axis. The ranges of SV size (base pairs) are indicated using a $\log_{10}$ scale on the x axis.

(I) Percentage of variants located in the gene regions. The percentage of variants located in a given region of a protein-coding gene is indicated by color bars according to the type of variant: DEL, DUP, INS, INV, SNV, and indel are indicated by blue, orange, gray, yellow, light blue, and green bars, respectively. "FLANK" indicates the gene-flanking regions located within 5 kb of the terminal exon.

(J) Mean number of LoF variants per individual. The LoF SVs include only variants overlapping exons. The LoF SNVs or indels are disruptive variants annotated with snpEff, followed by LOFTEE, annotated as "high confidence" or "high impact." Error bars represent standard errors.

(K) pLI-stratified percentage of protein-coding genes with LoF variants per individual. The percentages of genes in the indicated range of pLI scores among all pLI-annotated genes are shown for each type of variant. For SNVs and indels, LoF variants annotated with LOFTEE were used. Genes (percent) with pLI of <0.1, 0.1–0.9, ≧0.9, and ≧0.99 are indicated by blue, light blue, yellow, and orange bars, respectively.

(L) Distribution of SVs overlapping with exons of pLI genes with pLI ≧ 0.99 across AF ranges. The number of SVs overlapping exons of ≧ 0.99 pLI genes in the BBJ-SV data is indicated by blue bars for the indicated AF range. The average percentage of heterozygous SVs is indicated with blue letters on each bar.

(M) DELs and DUPs overlapping with specific genomic regions are constrained. Genomic regions analyzed include phastCons evolutionarily conserved regions (conserve), VISTA enhancers (enhancer), exons of protein-coding genes (coding_exon) and noncoding genes (noncoding_exon), and promoter regions 1 kb (promoter1K) and 10 kb (promoter10K) upstream of the transcriptional start sites of all genes. coding_pLI, promoter1K_pLI, and promoter10K_pLI are restricted to genes with pLI ≧ 0.9. Regions excluding the above regions and intron regions were designated intergenic regions. The constrained degree of SVs overlapping each region against those located in the intergenic regions was expressed as the inverse of the odds ratio (OR). Inverse ORs of DEL and DUP are shown as blue and orange bars, respectively, with CIs.

Table 2. Known disease risk genes with rare exonic SVs in the BBJ WGS data

Disease	Gene	Source[a]	Region[b]	SV type	SV size (kb)	Chr:pos	Case	Control	OR	p Value
Colorectal cancer (CoCa) (n = 196)	MLH1	ClinVar	exon	DEL	109	3:36940070	2	0	60.3	5.9E−3
					1.2	3:37048066				
	APC	ClinVar	exon-A	DEL	2760	5:109469013	2	0	60.3	5.9E−3
					825	5:111446233				
	MSH2	ClinVar	exon	DEL	31.0	2:47605652	2	0	60.3	5.9E−3
					11.2	2:47636876				
	NTHL1	ClinVar	exon-A	DUP	80.0	16:2068893	1	0	36.1	7.7E−2
	EPCAM	ClinVar	exon	DEL	31.0	2:47605652	1	1	12.0	0.15
	SASH1	ClinVar-c	exon	DEL	84.3	6:148683204	1	0	36.1	7.7E−2
	SDHD	ClinVar-c	exon	DEL	2.1	11:111963575	1	0	36.1	7.7E−2
Breast cancer (BrCa) (n = 237)	APC	ClinVar	exon-A	DUP	977	5:111695439	1	0	4.6[c] (29.8)	0.4[c] (9.2E−2)
	MLH1	ClinVar	intron	DEL	4.6	3:37076354	1	0	4.6	0.4
	PPM1D	ClinVar	exon-A	DUP	179	17:58589824	1	0	4.6	0.4
	MSMB	ClinVar-c	exon-A	DEL	135	10:51470001	1	0	4.6	0.4
Prostate cancer (PrCa) (n = 215)	RNASEL	ClinVar	exon-A	DEL	170	1:182442803	4	10	3.7	4.0E−2
	ATM	ClinVar-c	exon	INV	7.3	11:108137195	1	1	9.3	0.19
	RECQL	ClinVar-c	exon	DEL	3.0	12:21623979	1	0	27.8	9.7E−2
	BRCA1	ClinVar-c	exon-A	DUP	411	17:41202158	1	2	4.6	0.26
Gastric cancer (GaCa) (n = 257)	RAF1	ClinVar	exon	DEL	12.1	3:12657398	1	0	27.5	9.9E−2
	IDH1	ClinVar-c	exon-A	DUP	169	2:209041733	1	0	27.5	9.9E−2
	SDHD	ClinVar-c	exon	DUP	145	11:111968279	1	0	27.5	9.9E−2
	RNF43	ClinVar-c	exon-A	DUP	173	17:56409879	1	0	27.5	9.9E−2
	ATM	ClinVar-c	exon	DEL	4.4	11:108094086	1	0	27.5	9.9E−2
Dementia (n = 200)	MEF2C	ClinVar	exon	DEL	16.0	5:88114001	2	0	76.7	3.8E−3

[a]ClinVar, risk genes with "pathogenic" germline mutations for the corresponding disease in ClinVar; ClinVar-c, risk genes with "pathigenic" germline mutations for all cancer diseases in ClinVar in exons, introns, or UTRs.
[b]Exon-A, all exons of the corresponding genes overlap the SV.
[c]Values in parentheses indicate ORs and p values determined with control samples, including male samples.

alkaline phosphatase (ALP), and albumin/globulin ratio (A/G) found in four exonic SVs of the *MUC22*, *GYPA/GYPB*, *FUT2*, and *RP11-219A15.2* genes were novel. A 4-kb low-frequency DEL was associated with 4 hemoglobin-related traits (Red blood cell count [RBC], MCV, MCH, and MCHC) and overlaps the exons of the *HBA1* gene encoding hemoglobin subunit alpha (Figure 5I), which reflects loss of *HBA1* gene function. A 119-kb low-frequency DUP overlapping *GYPA/GYPB* genes was associated with HbA1c and MCHC (Table S8). *GYPA/GYPB* encode glycophorin A/B, the major and minor glycosylated membrane proteins of the erythrocyte, respectively, suggesting association of an extra copy of *GYPA* or *GYPB* with the HbA1c and MCHC traits. Another 1.6-kb DEL associated with LDL-C and triglyceride (TG) overlaps an exon of the *APOC1* gene, which encodes a member of the apolipoprotein family known to have an association with these traits. The association strength of this DEL is considerably lower than those of the overlapping SNPs at this locus, but this low-frequency DEL is not in LD with other associated common SNPs/indels (Figure 5L). The exonic SVs of the *GYPA/GYPB*, *RP11-219A15.2*, and *HBA1* genes overlapped with segmental DUPs, raising suspicion of a false positive, but manual visual inspection with 20 case and 20 control samples showed clear differences in read alignment between cases and controls. In addition, the 4-kb DEL overlapping the *HBA1* gene has been reported by previous studies,[50] and the 119-kb DUP overlapping the *GYPA/GYPB* genes was also observed in gnomAD-SVs, supporting the validity of our findings. Of the top-ranked 30 SVs for the binary and quantitative traits that matched with gnomAD-SVs, 7 SVs (23%) were rare or low-frequency SVs with AFs 5-fold lower in the European population than in the Japanese population (Table S8), indicating a high population-specific association of SVs.

To infer the causality of the GWAS variants on the traits, we used the transcription factor (TF)-binding footprints responsible for gene regulation in noncoding regions.[51] Of the 41 top-ranked GWAS SVs, 13 DELs (59% of the 22 top-ranked DELs) overlapped with the TF-binding footprints, most of which were footprints from tissues relevant to the corresponding traits (Table S8). To determine the empirical p value based on random expectation, we estimated how many randomly selected non-repeat genomic regions corresponding to top-ranked GWAS DELs can overlap with TF footprints. After 1,000 iterations of this simulation, an average of 5.8 of the 21 DELs tested (SD = 1.59) overlapped with the TF footprint, indicating that the top GWAS DELs found in this study were enriched in the TF footprint (11/21,

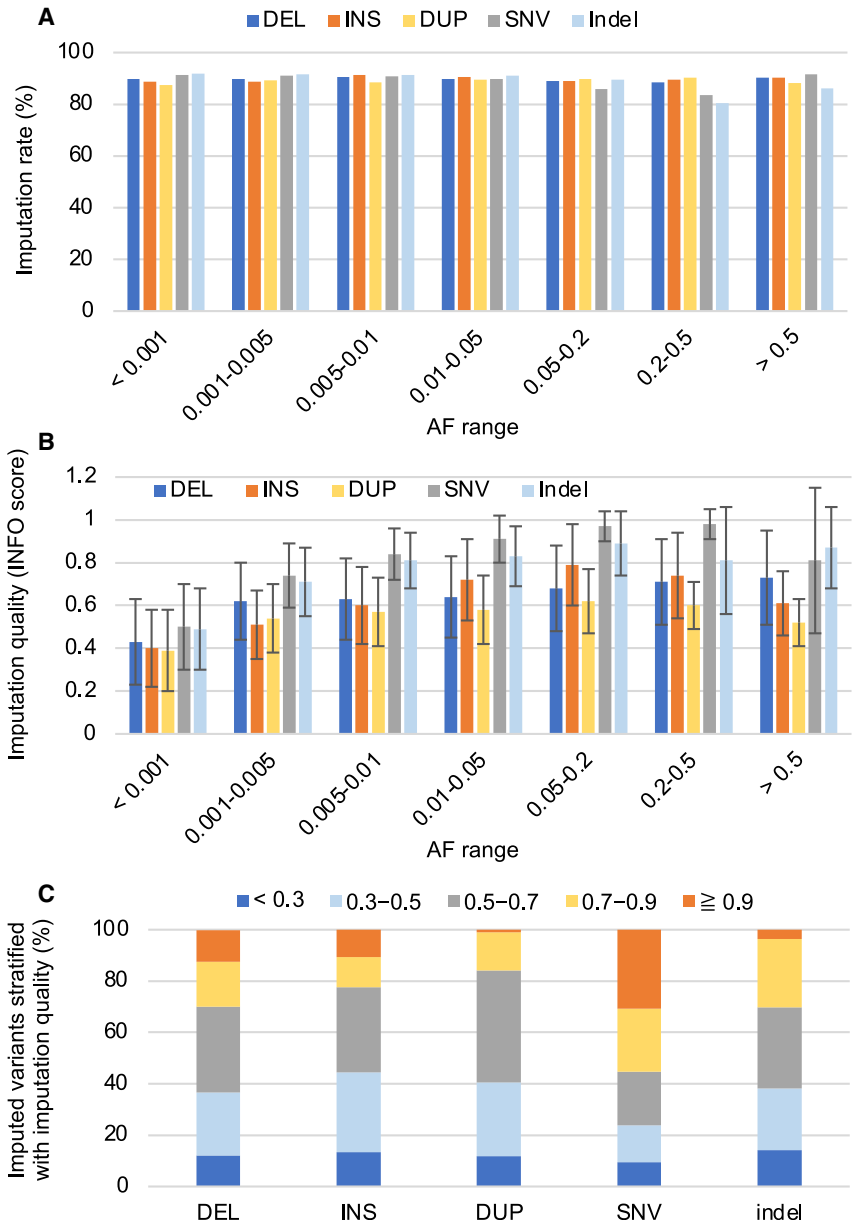


Figure 4. Imputation efficiency and quality of array genotype data from 181,622 samples

(A) Imputation efficiency of DELs, INSs, DUPs, SNVs, and indels in the AF range. The bars indicate the percentage of variants imputed with the array data of 181,622 samples using the 13 M SNP-indel-SV reference panel with ≧ AF 0.0003 (22,806,626 of SNVs, 901,910 of indels, 31,380 of DELs, 8,852 of DUPs, 36,518 of INSs) for a given variant type and AF range. DELs, INSs, DUPs, SNVs, and indels are shown in blue, orange, yellow, gray, and pale blue, respectively.

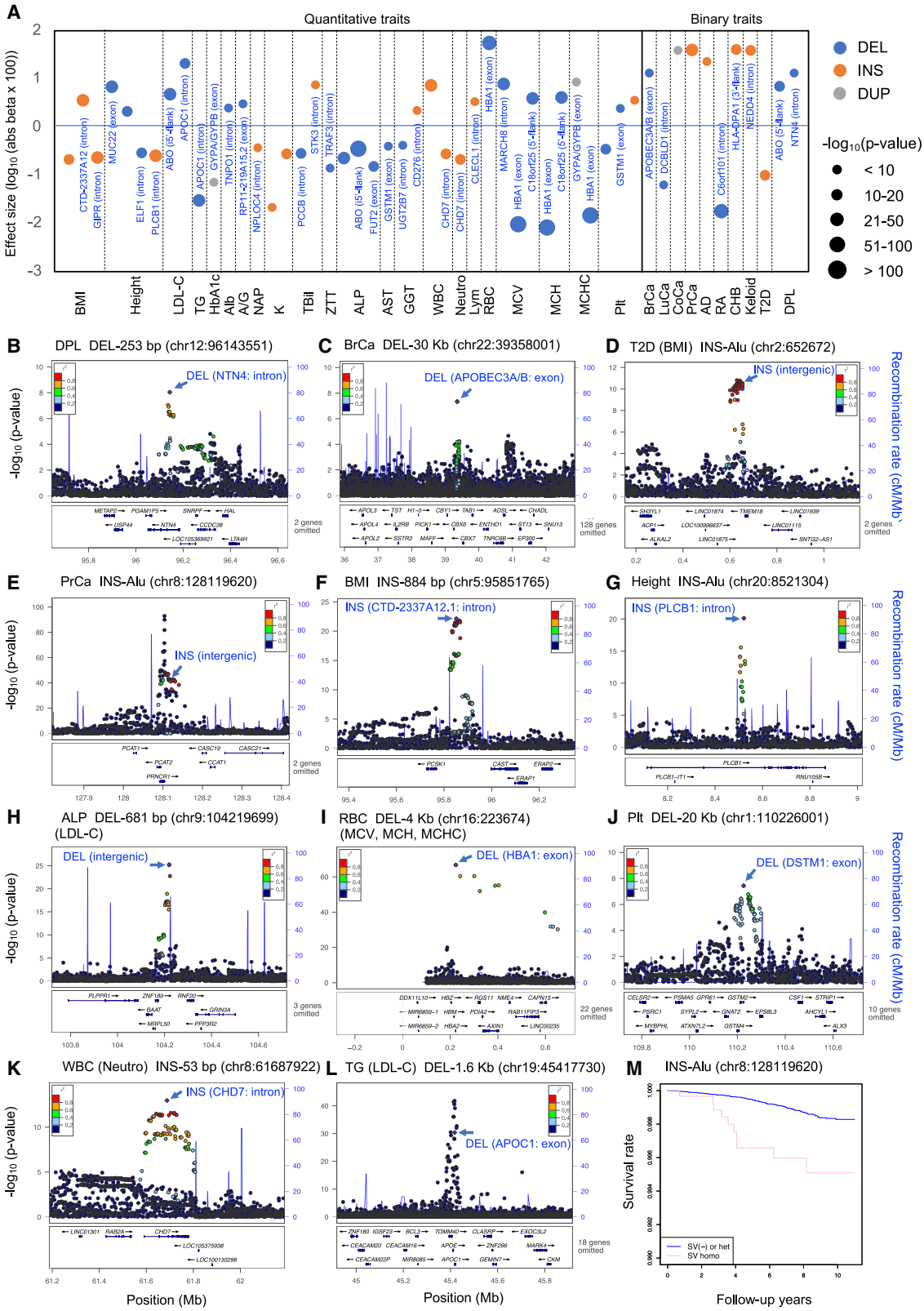
(B) Imputation quality for each variant type in the AF range. The bars indicate the average INFO score for a given variant type and a given AF range. Error bars represent standard errors.

(C) Imputed variants fractionated by imputation quality. The percentage of variants with a given quality among all imputed variants of the corresponding type is indicated by colored bars (blue, <0.3; light blue, 0.3–0.5; gray, 0.5–0.7; yellow, 0.7–0.9; orange, ≧ 0.9).

$p = 5.4 \times 10^{-4}$). Notably, for the top-ranked GWAS INSs, mobile element INSs, including Alu, L1, and SVA INSs, were significantly enriched (14/18; OR = 5.2, CI = 1.7–15.8; $p = 1.5 \times 10^{-3}$, Fisher's exact test; $p = 2.3 \times 10^{-4}$, hypergeometric distribution test; 40.2% of INSs were estimated to be mobile element INSs [MEIs]; see STAR Methods for details). This was also observed for the INSs associated with published GWAS signals in the previous section. Collectively, our results show that understudied SVs likely play a significant role in diseases and traits.

DISCUSSION

This study demonstrates efficient detection of high-quality SVs using SR WGS data from thousands of samples. The fully genotyped SV set generated with MOPline enabled us to detect trait-associated SVs, including population-specific ones, by burden tests and GWAS with imputed SVs. The quality of the detected SVs is comparable with that of gnomAD-SVs, which provides high-quality SVs for reference,[22] whereas the median number of SVs per individual reaches about 16,000, more than twice the number of gnomAD-SVs. The high quality of SVs is attributed to the strategy of selecting high-quality SVs that are overlap calls from selected pairs of SV detection algorithms. The strategy of selecting overlap calls from multiple SV call sets has been often used, but because overlap calls from many algorithm pairs do not have high accuracy,[29] it is important to select specific algorithm pairs that have high accuracy for overlap calls. The high per-individual SV recovery rate can be attributed to the SMC function employed for joint-called SVs from multiple samples. In the evaluation using the SV call set of NA12878, SMC was able to increase the SV TP calls by 42% of the initially selected overlap calls while achieving approximately 95% accuracy. SMC recovers missing calls, which has been initially defined as a reference allele for samples at an SV site, using the original SV calls from the algorithms used and re-genotyping with SR alignment signals. Because the overlap-based SV selection in the first step generates only high-quality SVs to remove low-quality ones, SVs have often been lost in a fraction of samples because SV calls often vary among WGS data and among SV detection algorithms used. Overlap-based selection of high-quality SVs followed by recovering missing calls with the SMC function enables accurate and sensitive detection of SVs. In addition, MOPline can buffer unstable calls from a single algorithm to call a constant number of SVs per sample (Figures S15 and S16).



(legend on next page)

The LR WGS data yielded more than 24,000 SV calls per individual, which is 1.5 times the number of SVs detected in this study. Many of the SVs unique to the LR data are located in highly polymorphic short tandem repeats/STRs, where the copy number of the repeat unit increases or decreases among individuals. Because LRs can resolve the structure of repeat units, the LR-based SV detection strategy can detect more copy number alleles in STRs than the SR-based one. However, even with LRs, it is difficult to precisely assign the position and size of SVs within the repeat regions, so many fragmented DELs and INSs observed in the STR region may seemingly increase the number of SV calls in the current LR-based SV call data. Despite the disadvantage of SRs to resolve repeat regions, it is likely that SR-based SV detection can more effectively detect large (≧ 10 kb) DELs than LR-based ones; comparisons for the other SV types could not be made because of limited data. This difference is attributed to the coverage-, split read-, and RP-based methods for SRs. LR-based SV detection is likely to have difficulty detecting such indirect signals of large SVs that fail to be aligned within an LR, especially for low-coverage data or reads that are not error corrected.

On average, approximately 2.7 SVs (AF ≦ 0.001) and 7.4 SVs (AF ≦ 0.01) per individual in the BBJ-SVs overlapped with exons of protein-coding genes. These numbers are equivalent to 27% and 39% of the LoF SNV/indels annotated with LOFTEE, respectively (Figure 3J), although all exon-overlapping DUPs and INVs on gene function may not be necessarily LoF variants, and we did not consider the effects of the SVs located in the intron regions on splicing. SVs could have a higher potential for affecting the function of noncoding genes and the transcriptional or splicing function in noncoding regulatory regions because of their large size than SNVs and indels. In fact, SVs of noncoding genes and promoter regions are more constrained than SVs of intergenic regions, as observed in this study. Furthermore, for haploinsufficiency genes (i.e., pLI ≧ 0.9), SVs of coding and promoter regions are more constrained. Specifically, DELs have a greater impact on disrupting regulatory elements, as observed in GWAS DELs, and have a higher frequency of overlapping TF footprints than GWAS SNVs/indels. Given that DELs and DUPs in regulatory regions are constrained by negative selection, such SVs are less frequent and may be less likely to participate as causal variants in polygenic common diseases. However, our GWAS identifies potentially causal SVs for several traits, which include common exonic SVs, such as a GSTM1 LoF DEL for the platelet count (Plt) and AST traits, FUT2 LoF DEL for the ALP trait, and an APOBEC3 LoF DEL for BrCa. Furthermore, a significant enrichment of MEIs in the GWAS INSs suggests that INSs, particularly MEIs, alter expression of the surrounding or distant genes through regulatory elements within INS sequences, as shown in a recent study,[48] or through methylation induced by the mobile elements. These observations suggest that many rare and common SVs in noncoding and coding regions should be involved in the causality of common diseases and traits. Thus, incorporating SVs into association studies using traditional SNPs and indels would increase the potential for identifying the cause of a disease or trait. In addition, our observation that at least 23% of the BBJ GWAS SVs are much less frequent in the European population illustrates the importance of performing association analyses of SV in many populations.

This study demonstrates a framework for detecting high-quality SVs and for identifying SVs associated with diseases and traits using multiple datasets, including imputed SVs. MOPline can detect SVs from SR sequencing data for single and tens of thousands of samples and can also handle non-human samples. This paper also shows pseudo-DELs and INVs that are prone to misidentification because of interspersed DUP signals in SR data and presents thousands of HF SVs common to different populations as useful insights for further SV call improvement. The methodology and the SV datasets created in this study provide a valuable resource for SV analyses in diverse research areas.

Limitations of the study

MOPline has several limitations; it cannot detect TRAs, genotyping DUPs is incomplete, and there are false positive calls. The first problem is the limited number of TRA detection algorithms and the lack of reference TRA information to evaluate the calling accuracy. The second is due to the difficulty of resolving the number of copies of DUPs for each haplotype in SR data. For the last problem, MOPline is still imperfect in its accuracy. In addition, most short-read-based SV detection algorithms, including MOPline, cannot accurately determine the length of INSs at many sites and often only detect their BPs. Thus, some of MOPline's SV calls would contain INSs shorter than 50 bp. Many of the false calls in MOPline are due to incorrect coverage caused by misaligned reads or split reads caused by short indels. GWASs with imputed SVs would produce some false signals because of incomplete SV imputation. Therefore, identified SVs, such as trait-associated SVs, need to be verified by alignment views or PCR to confirm that the identified SVs are indeed correct.

Figure 5. GWASs for binary and quantitative traits identified a number of genome-wide significant SVs

(A) Top-ranked GWAS SVs for quantitative and binary traits. Effect sizes for the top-ranked or nearly top-ranked SVs identified for each trait are plotted. For clarity, SV effect sizes are $\log_{10}$-transformed absolute values of the beta coefficient while keeping the direction of beta. DELs, INSs, and DUPs are indicated by blue, orange, and gray circles, with different sizes reflecting different p value ranges, as shown in the right panel. Trait abbreviations are shown in black letters at the bottom, and SV-overlapping genes and their gene regions are indicated in blue letters.

(B–L) Regional association plots for DPL (hyperlipidemia; B), BrCa (breast cancer; C), T2D (type 2 diabetes; D), PrCa (prostate cancer; E), BMI (body mass index; F), height (G), ALP (alkaline phosphatase; H), RBC (red blood cell count; I), Plt (platelet count; J), WBC (white blood cell count; K), and TG (triglyceride; L). SVs highlighted by purple diamonds and arrows indicate the top-ranked p values within the 1-Mb region, and the gene regions overlapping the SVs are indicated in blue letters. The ALP-associated DELs and WBC-associated INSs are also top-ranked variants associated with the low-density-lipoprotein cholesterol (LDL-C) and neutrophil count (Neutro) traits, respectively. The 4-kb DEL at chr16:223674 (I) is the low-frequency top variant common to all 4 traits (RBC, MCV, MCH, and MCHC).

(M) Survival analysis of PrCa-associated INSs. Survival rates for homozygous carriers and others of the PrCa-associated INS (chr8:128119620) are plotted over 12 years. The mortality data were obtained from 140,000 subjects in the BBJ follow-up study.

CONSORTIA

The BioBank Japan Project: Yuji Yamanashi, Yoichi Furukawa, Takayuki Morisaki, Yoshinori Murakami, Yoichiro Kamatani, Kaori Muto, Akiko Nagai, Wataru Obara, Ken Yamaji, Kazuhisa Takahashi, Satoshi Asai, Yasuo Takahashi, Takao Suzuki, Nobuaki Sinozaki, Hiroki Yamaguchi, Shiro Minami, Shigeo Murayama, Kozo Yoshimori, Satoshi Nagayama, Daisuke Obata, Masahiko Higashiyama, Akihide Masumoto, and Yukihiro Koretsune

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- METHOD DETAILS
  - WGS datasets
  - Reference SV dataset for real data
  - SV calling with simulated and real datasets
  - SV calling using GATK-SV and sv-pipeline
  - Analysis of SVs overlapping genomic elements
  - SNV and short indel calling
  - Compound heterozygous variants
  - Evaluation of SV detection algorithms
  - MOPline algorithm
  - INSs and DELs of mobile elements
  - PCA
  - LD test
  - HWE test
  - SV-associated disease risk genes
  - SVs associated with published GWAS signals
  - Imputation of SVs
  - GWAS
  - Survival analyses of prostate cancer
  - GWAS DELs overlapping TF footprints

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.xgen.2023.100328.

ACKNOWLEDGMENTS

We thank Dr. Takeshi Usui (Shizuoka General Hospital) for providing the environment for data analysis. We thank the members of our laboratory (Laboratory for Statistical and Translational Genetics, RIKEN Center for Integrative Medical Sciences) for their valuable comments and useful assistance. This study was supported by AMED under grants JP21kk0305013, JP21tm0424220, and JP21ck0106642; Japan Society for the Promotion of Science KAKENHI grants JP20H00462 and JP17K07264; the BioBank Japan project, which was supported by the Ministry of Education, Culture, Sports, Sciences, and Technology of the Japanese Government; and AMED Japan under grants 17km0305002 and 18km0605001.

AUTHOR CONTRIBUTIONS

S.K., Y.K., and C.T. conceived and designed the project. T.M., Y.M., and BBJ project members provided genome DNA samples, array genotype data, and WGS data from BBJ. K.T. created short variant calls from the BBJ WGS data. K.H. conducted installation and execution of the GATK-SV pipeline and created a singularity definition file for external tools and MOPline. C.T. conducted the survival analysis for PrCa. S.K. and C.T. conducted GWASs. S.K. performed the other analyses, including SV calling and imputation. S.K. created the MOPline code. S.K. and C.T. wrote the manuscript. All authors reviewed the paper and approved the final version.

DECLARATION OF INTERESTS

The authors declare no competing interests.

INCLUSION AND DIVERSITY

We support inclusive, diverse, and equitable conduct of research.

Received: May 19, 2022
Revised: February 17, 2023
Accepted: April 25, 2023
Published: May 18, 2023

REFERENCES

1. Alkan, C., Coe, B.P., and Eichler, E.E. (2011). Genome structural variation discovery and genotyping. Nat. Rev. Genet. *12*, 363–376. https://doi.org/10.1038/nrg2958.
2. Sudmant, P.H., Rausch, T., Gardner, E.J., Handsaker, R.E., Abyzov, A., Huddleston, J., Zhang, Y., Ye, K., Jun, G., Fritz, M.H.Y., et al. (2015). An integrated map of structural variation in 2,504 human genomes. Nature *526*, 75–81. https://doi.org/10.1038/nature15394.
3. Pang, A.W., MacDonald, J.R., Pinto, D., Wei, J., Rafiq, M.A., Conrad, D.F., Park, H., Hurles, M.E., Lee, C., Venter, J.C., et al. (2010). Towards a comprehensive structural variation map of an individual human genome. Genome Biol. *11*, R52. https://doi.org/10.1186/gb-2010-11-5-r52.
4. D'Haene, E., and Vergult, S. (2021). Interpreting the impact of noncoding structural variation in neurodevelopmental disorders. Genet. Med. *23*, 34–46. https://doi.org/10.1038/s41436-020-00974-1.
5. Chiang, C., Scott, A.J., Davis, J.R., Tsang, E.K., Li, X., Kim, Y., Hadzic, T., Damani, F.N., Ganel, L., et al.; GTEx Consortium (2017). The impact of structural variation on human gene expression. Nat. Genet. *49*, 692–699. https://doi.org/10.1038/ng.3834.
6. Halvorsen, M., Huh, R., Oskolkov, N., Wen, J., Netotea, S., Giusti-Rodriguez, P., Karlsson, R., Bryois, J., Nystedt, B., Ameur, A., et al. (2020). Increased burden of ultra-rare structural variants localizing to boundaries of topologically associated domains in schizophrenia. Nat. Commun. *11*, 1842. https://doi.org/10.1038/s41467-020-15707-w.
7. Scott, A.J., Chiang, C., and Hall, I.M. (2021). Structural variants are a major source of gene expression differences in humans and often affect multiple nearby genes. Genome Res. *31*, 2249–2257. https://doi.org/10.1101/gr.275488.121.
8. Stankiewicz, P., and Lupski, J.R. (2010). Structural variation in the human genome and its role in disease. Annu. Rev. Med. *61*, 437–455. https://doi.org/10.1146/annurev-med-100708-204735.
9. Weischenfeldt, J., Symmons, O., Spitz, F., and Korbel, J.O. (2013). Phenotypic impact of genomic structural variation: insights from and for human disease. Nat. Rev. Genet. *14*, 125–138. https://doi.org/10.1038/nrg3373.
10. Quigley, D.A., Dang, H.X., Zhao, S.G., Lloyd, P., Aggarwal, R., Alumkal, J.J., Foye, A., Kothari, V., Perry, M.D., Bailey, A.M., et al. (2018). Genomic

hallmarks and structural variation in metastatic prostate cancer. Cell *174*, 758–769.e9. https://doi.org/10.1016/j.cell.2018.06.039.

11. Dixon, J.R., Xu, J., Dileep, V., Zhan, Y., Song, F., Le, V.T., Yardımcı, G.G., Chakraborty, A., Bann, D.V., Wang, Y., et al. (2018). Integrative detection and analysis of structural variation in cancer genomes. Nat. Genet. *50*, 1388–1398. https://doi.org/10.1038/s41588-018-0195-8.
12. Li, W., and Olivier, M. (2013). Current analysis platforms and methods for detecting copy number variation. Physiol. Genomics *45*, 1–16. https://doi.org/10.1152/physiolgenomics.00082.2012.
13. Mahmoud, M., Gobet, N., Cruz-Dávalos, D.I., Mounier, N., Dessimoz, C., and Sedlazeck, F.J. (2019). Structural variant calling: the long and the short of it. Genome Biol. *20*, 246. https://doi.org/10.1186/s13059-019-1828-7.
14. Ho, S.S., Urban, A.E., and Mills, R.E. (2020). Structural variation in the sequencing era. Nat. Rev. Genet. *21*, 171–189. https://doi.org/10.1038/s41576-019-0180-9.
15. Liu, Z., Roberts, R., Mercer, T.R., Xu, J., Sedlazeck, F.J., and Tong, W. (2022). Towards accurate and reliable resolution of structural variants for clinical diagnosis. Genome Biol. *23*, 68. https://doi.org/10.1186/s13059-022-02636-8.
16. Almarri, M.A., Bergström, A., Prado-Martinez, J., Yang, F., Fu, B., Dunham, A.S., Chen, Y., Hurles, M.E., Tyler-Smith, C., and Xue, Y. (2020). Population structure, stratification, and introgression of human structural variation. Cell *182*, 189–199.e15. https://doi.org/10.1016/j.cell.2020.05.024.
17. Ebert, P., Audano, P.A., Zhu, Q., Rodriguez-Martin, B., Porubsky, D., Bonder, M.J., Sulovari, A., Ebler, J., Zhou, W., Serra Mari, R., et al. (2021). Haplotype-resolved diverse human genomes and integrated analysis of structural variation. Science *372*, eabf7117. https://doi.org/10.1126/science.abf7117.
18. Porubsky, D., Ebert, P., Audano, P.A., Vollger, M.R., Harvey, W.T., Marijon, P., Ebler, J., Munson, K.M., Sorensen, M., Sulovari, A., et al. (2021). Fully phased human genome assembly without parental data using single-cell strand sequencing and long reads. Nat. Biotechnol. *39*, 302–308. https://doi.org/10.1038/s41587-020-0719-5.
19. Garg, S., Fungtammasan, A., Carroll, A., Chou, M., Schmitt, A., Zhou, X., Mac, S., Peluso, P., Hatas, E., Ghurye, J., et al. (2021). Chromosome-scale, haplotype-resolved assembly of human genomes. Nat. Biotechnol. *39*, 309–312. https://doi.org/10.1038/s41587-020-0711-0.
20. Beyter, D., Ingimundardottir, H., Oddsson, A., Eggertsson, H.P., Bjornsson, E., Jonsson, H., Atlason, B.A., Kristmundsdottir, S., Mehringer, S., Hardarson, M.T., et al. (2021). Long-read sequencing of 3,622 Icelanders provides insight into the role of structural variants in human diseases and other traits. Nat. Genet. *53*, 779–786. https://doi.org/10.1038/s41588-021-00865-4.
21. Zook, J.M., Hansen, N.F., Olson, N.D., Chapman, L., Mullikin, J.C., Xiao, C., Sherry, S., Koren, S., Phillippy, A.M., Boutros, P.C., et al. (2020). A robust benchmark for detection of germline large deletions and insertions. Nat. Biotechnol. *38*, 1347–1355. https://doi.org/10.1038/s41587-020-0538-8.
22. Collins, R.L., Brand, H., Karczewski, K.J., Zhao, X., Alföldi, J., Francioli, L.C., Khera, A.V., Lowther, C., Gauthier, L.D., Wang, H., et al. (2020). A structural variation reference for medical and population genetics. Nature *581*, 444–451. https://doi.org/10.1038/s41586-020-2287-8.
23. Abel, H.J., Larson, D.E., Regier, A.A., Chiang, C., Das, I., Kanchi, K.L., Layer, R.M., Neale, B.M., Salerno, W.J., Reeves, C., et al. (2020). Mapping and characterization of structural variation in 17,795 human genomes. Nature *583*, 83–89. https://doi.org/10.1038/s41586-020-2371-0.
24. Byrska-Bishop, M., Evani, U.S., Zhao, X., Basile, A.O., Abel, H.J., Regier, A.A., Corvelo, A., Clarke, W.E., Musunuri, R., Nagulapalli, K., et al. (2022). High-coverage whole-genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios. Cell *185*, 3426–3440.e19. https://doi.org/10.1016/j.cell.2022.08.004.
25. Lin, K., Smit, S., Bonnema, G., Sanchez-Perez, G., and de Ridder, D. (2015). Making the difference: integrating structural variation detection tools. Brief. Bioinform. *16*, 852–864. https://doi.org/10.1093/bib/bbu047.
26. Pirooznia, M., Goes, F.S., and Zandi, P.P. (2015). Whole-genome CNV analysis: advances in computational approaches. Front. Genet. *6*, 138. https://doi.org/10.3389/fgene.2015.00138.
27. Khayat, M.M., Sahraeian, S.M.E., Zarate, S., Carroll, A., Hong, H., Pan, B., Shi, L., Gibbs, R.A., Mohiyuddin, M., Zheng, Y., and Sedlazeck, F.J. (2021). Hidden biases in germline structural variant detection. Genome Biol. *22*, 347. https://doi.org/10.1186/s13059-021-02558-x.
28. Pabinger, S., Dander, A., Fischer, M., Snajder, R., Sperk, M., Efremova, M., Krabichler, B., Speicher, M.R., Zschocke, J., and Trajanoski, Z. (2014). A survey of tools for variant analysis of next-generation genome sequencing data. Brief. Bioinform. *15*, 256–278. https://doi.org/10.1093/bib/bbs086.
29. Kosugi, S., Momozawa, Y., Liu, X., Terao, C., Kubo, M., and Kamatani, Y. (2019). Comprehensive evaluation of structural variation detection algorithms for whole genome sequencing. Genome Biol. *20*, 117. https://doi.org/10.1186/s13059-019-1720-5.
30. Genome of the Netherlands Consortium (2014). Whole-genome sequence variation, population structure and demographic history of the Dutch population. Nat. Genet. *46*, 818–825. https://doi.org/10.1038/ng.3021.
31. Mills, R.E., Walter, K., Stewart, C., Handsaker, R.E., Chen, K., Alkan, C., Abyzov, A., Yoon, S.C., Ye, K., Cheetham, R.K., et al. (2011). Mapping copy number variation by population-scale genome sequencing. Nature *470*, 59–65. https://doi.org/10.1038/nature09708.
32. Nagasaki, M., Yasuda, J., Katsuoka, F., Nariai, N., Kojima, K., Kawai, Y., Yamaguchi-Kabata, Y., Yokozawa, J., Danjoh, I., Saito, S., et al. (2015). Rare variant discovery by deep whole-genome sequencing of 1,070 Japanese individuals. Nat. Commun. *6*, 8018. https://doi.org/10.1038/ncomms9018.
33. Brandler, W.M., Antaki, D., Gujral, M., Noor, A., Rosanio, G., Chapman, T.R., Barrera, D.J., Lin, G.N., Malhotra, D., Watts, A.C., et al. (2016). Frequency and complexity of de novo structural mutation in autism. Am. J. Hum. Genet. *98*, 667–679. https://doi.org/10.1016/j.ajhg.2016.02.018.
34. Gokcumen, O., Tischler, V., Tica, J., Zhu, Q., Iskow, R.C., Lee, E., Fritz, M.H.Y., Langdon, A., Stütz, A.M., Pavlidis, P., et al. (2013). Primate genome architecture influences structural variation mechanisms and functional consequences. Proc. Natl. Acad. Sci. USA *110*, 15764–15769. https://doi.org/10.1073/pnas.1305904110.
35. Kloosterman, W.P., Francioli, L.C., Hormozdiari, F., Marschall, T., Hehir-Kwa, J.Y., Abdellaoui, A., Lameijer, E.W., Moed, M.H., Koval, V., Renkens, I., et al. (2015). Characteristics of de novo structural changes in the human genome. Genome Res. *25*, 792–801. https://doi.org/10.1101/gr.185041.114.
36. Werling, D.M., Brand, H., An, J.Y., Stone, M.R., Zhu, L., Glessner, J.T., Collins, R.L., Dong, S., Layer, R.M., Markenscoff-Papadimitriou, E., et al. (2018). An analytical framework for whole-genome sequence association studies and its implications for autism spectrum disorder. Nat. Genet. *50*, 727–736. https://doi.org/10.1038/s41588-018-0107-y.
37. Jakubosky, D., Smith, E.N., D'Antonio, M., Jan Bonder, M., Young Greenwald, W.W., D'Antonio-Chronowska, A., Matsui, H., i2QTL Consortium; Stegle, O., Montgomery, S.B., et al. (2020). Discovery and quality analysis of a comprehensive set of structural variants and short tandem repeats. Nat. Commun. *11*, 2928. https://doi.org/10.1038/s41467-020-16481-5.
38. Verbiest, M., Maksimov, M., Jin, Y., Anisimova, M., Gymrek, M., and Bilgin Sonay, T. (2023). Mutation and selection processes regulating short tandem repeats give rise to genetic and phenotypic diversity across species. J. Evol. Biol. *36*, 321–336. https://doi.org/10.1111/jeb.14106.
39. Lek, M., Karczewski, K.J., Minikel, E.V., Samocha, K.E., Banks, E., Fennell, T., O'Donnell-Luria, A.H., Ware, J.S., Hill, A.J., Cummings, B.B.,

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

40. Liu, X., Takata, S., Ashikawa, K., Aoi, T., Kosugi, S., Terao, C., Parrish, N.F., Matsuda, K., Nakagawa, H., Kamatani, Y., et al. (2020). Prevalence and spectrum of pathogenic germline variants in Japanese patients with early-onset colorectal, breast, and prostate cancer. JCO Precis. Oncol. *4*, 183–191. https://doi.org/10.1200/PO.19.00224.
41. Terao, C., Suzuki, A., Momozawa, Y., Akiyama, M., Ishigaki, K., Yamamoto, K., Matsuda, K., Murakami, Y., McCarroll, S.A., Kubo, M., et al. (2020). Chromosomal alterations among age-related haematopoietic clones in Japan. Nature *584*, 130–135. https://doi.org/10.1038/s41586-020-2426-2.
42. Akiyama, M., Okada, Y., Kanai, M., Takahashi, A., Momozawa, Y., Ikeda, M., Iwata, N., Ikegawa, S., Hirata, M., Matsuda, K., et al. (2017). Genome-wide association study identifies 112 new loci for body mass index in the Japanese population. Nat. Genet. *49*, 1458–1467. https://doi.org/10.1038/ng.3951.
43. Ishigaki, K., Akiyama, M., Kanai, M., Takahashi, A., Kawakami, E., Sugishita, H., Sakaue, S., Matoba, N., Low, S.K., Okada, Y., et al. (2020). Large-scale genome-wide association study in a Japanese population identifies novel susceptibility loci across different diseases. Nat. Genet. *52*, 669–679. https://doi.org/10.1038/s41588-020-0640-3.
44. Kanai, M., Akiyama, M., Takahashi, A., Matoba, N., Momozawa, Y., Ikeda, M., Iwata, N., Ikegawa, S., Hirata, M., Matsuda, K., et al. (2018). Genetic analysis of quantitative traits in the Japanese population links cell types to complex human diseases. Nat. Genet. *50*, 390–400. https://doi.org/10.1038/s41588-018-0047-6.
45. Akiyama, M., Ishigaki, K., Sakaue, S., Momozawa, Y., Horikoshi, M., Hirata, M., Matsuda, K., Ikegawa, S., Takahashi, A., Kanai, M., et al. (2019). Characterizing rare and low-frequency height-associated variants in the Japanese population. Nat. Commun. *10*, 4393. https://doi.org/10.1038/s41467-019-12276-5.
46. Nik-Zainal, S., Wedge, D.C., Alexandrov, L.B., Petljak, M., Butler, A.P., Bolli, N., Davies, H.R., Knappskog, S., Martin, S., Papaemmanuil, E., et al. (2014). Association of a germline copy number polymorphism of APOBEC3A and APOBEC3B with burden of putative APOBEC-dependent mutations in breast cancer. Nat. Genet. *46*, 487–491. https://doi.org/10.1038/ng.2955.
47. Speliotes, E.K., Willer, C.J., Berndt, S.I., Monda, K.L., Thorleifsson, G., Jackson, A.U., Lango Allen, H., Lindgren, C.M., Luan, J., Mägi, R., et al. (2010). Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. Nat. Genet. *42*, 937–948. https://doi.org/10.1038/ng.686.
48. Payer, L.M., Steranka, J.P., Kryatova, M.S., Grillo, G., Lupien, M., Rocha, P.P., and Burns, K.H. (2021). Alu insertion variants alter gene transcript levels. Genome Res. *31*, 2236–2248. https://doi.org/10.1101/gr.261305.120.
49. Payer, L.M., Steranka, J.P., Yang, W.R., Kryatova, M., Medabalimi, S., Ardeljan, D., Liu, C., Boeke, J.D., Avramopoulos, D., and Burns, K.H. (2017). Structural variants caused by Alu insertions are associated with risks for many human diseases. Proc. Natl. Acad. Sci. USA *114*, E3984–E3992. https://doi.org/10.1073/pnas.1704117114.
50. Galanello, R., and Cao, A. (2011). Gene test review. Alpha-thalassemia. Genet. Med. *13*, 83–88. https://doi.org/10.1097/GIM.0b013e3181fcb468.
51. Vierstra, J., Lazar, J., Sandstrom, R., Halow, J., Lee, K., Bates, D., Diegel, M., Dunn, D., Neri, F., Haugen, E., et al. (2020). Global reference mapping of human transcription factor footprints. Nature *583*, 729–736. https://doi.org/10.1038/s41586-020-2528-x.
52. Okada, Y., Momozawa, Y., Sakaue, S., Kanai, M., Ishigaki, K., Akiyama, M., Kishikawa, T., Arai, Y., Sasaki, T., Kosaki, K., et al. (2018). Deep whole-genome sequencing reveals recent selection signatures linked to evolution and disease risk of Japanese. Nat. Commun. *9*, 1631. https://doi.org/10.1038/s41467-018-03274-0.
53. Koyama, S., Ito, K., Terao, C., Akiyama, M., Horikoshi, M., Momozawa, Y., Matsunaga, H., Ieki, H., Ozaki, K., Onouchi, Y., et al. (2020). Population-specific and trans-ancestry genome-wide analyses identify distinct and shared genetic risk loci for coronary artery disease. Nat. Genet. *52*, 1169–1177. https://doi.org/10.1038/s41588-020-0705-3.
54. Felsenstein, J., and Churchill, G.A. (1996). A Hidden Markov Model approach to variation among sites in rate of evolution. Mol. Biol. Evol. *13*, 93–104. https://doi.org/10.1093/oxfordjournals.molbev.a025575.
55. Holtgrewe, M., Kuchenbecker, L., and Reinert, K. (2015). Methods for the detection and assembly of novel sequence in high-throughput sequencing data. Bioinformatics *31*, 1904–1912. https://doi.org/10.1093/bioinformatics/btv051.
56. Xi, R., Lee, S., Xia, Y., Kim, T.M., and Park, P.J. (2016). Copy number analysis of whole-genome data using BIC-seq2 and its application to detection of cancer susceptibility variants. Nucleic Acids Res. *44*, 6274–6286. https://doi.org/10.1093/nar/gkw491.
57. Chen, K., Wallis, J.W., McLellan, M.D., Larson, D.E., Kalicki, J.M., Pohl, C.S., McGrath, S.D., Wendl, M.C., Zhang, Q., Locke, D.P., et al. (2009). BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. Nat. Methods *6*, 677–681. https://doi.org/10.1038/nmeth.1363.
58. Marschall, T., Costa, I.G., Canzar, S., Bauer, M., Klau, G.W., Schliep, A., and Schönhuth, A. (2012). CLEVER: clique-enumerating variant finder. Bioinformatics *28*, 2875–2882. https://doi.org/10.1093/bioinformatics/bts566.
59. Abyzov, A., Urban, A.E., Snyder, M., and Gerstein, M. (2011). CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. Genome Res. *21*, 974–984. https://doi.org/10.1101/gr.114876.110.
60. Rausch, T., Zichner, T., Schlattl, A., Stütz, A.M., Benes, V., and Korbel, J.O. (2012). DELLY: structural variant discovery by integrated paired-end and split-read analysis. Bioinformatics *28*, i333–i339. https://doi.org/10.1093/bioinformatics/bts378.
61. Zhu, M., Need, A.C., Han, Y., Ge, D., Maia, J.M., Zhu, Q., Heinzen, E.L., Cirulli, E.T., Pelak, K., He, M., et al. (2012). Using ERDS to infer copy-number variants in high-coverage genomes. Am. J. Hum. Genet. *91*, 408–421. https://doi.org/10.1016/j.ajhg.2012.07.004.
62. Li, H. (2015). FermiKit: assembly-based variant calling for Illumina resequencing data. Bioinformatics *31*, 3694–3696. https://doi.org/10.1093/bioinformatics/btv440.
63. Michaelson, J.J., and Sebat, J. (2012). forestSV: structural variant discovery through statistical learning. Nat. Methods *9*, 819–821. https://doi.org/10.1038/nmeth.2085.
64. Handsaker, R.E., Korn, J.M., Nemesh, J., and McCarroll, S.A. (2011). Discovery and genotyping of genome structural polymorphism by sequencing on a population scale. Nat. Genet. *43*, 269–276. https://doi.org/10.1038/ng.768.
65. Cameron, D.L., Schröder, J., Penington, J.S., Do, H., Molania, R., Dobrovic, A., Speed, T.P., and Papenfuss, A.T. (2017). GRIDSS: sensitive and specific genomic rearrangement detection using positional de Bruijn graph assembly. Genome Res. *27*, 2050–2060. https://doi.org/10.1101/gr.222109.117.
66. Dharanipragada, P., Vogeti, S., and Parekh, N. (2018). iCopyDAV: integrated platform for copy number variations-Detection, annotation and visualization. PLoS One *13*, e0195334. https://doi.org/10.1371/journal.pone.0195334.
67. Qi, J., and Zhao, F. (2011). inGAP-sv: a novel scheme to identify and visualize structural variation from paired end mapping data. Nucleic Acids Res. *39*, W567–W575. https://doi.org/10.1093/nar/gkr506.
68. Layer, R.M., Chiang, C., Quinlan, A.R., and Hall, I.M. (2014). LUMPY: a probabilistic framework for structural variant discovery. Genome Biol. *15*, R84. https://doi.org/10.1186/gb-2014-15-6-r84.

69. Chen, X., Schulz-Trieglaff, O., Shaw, R., Barnes, B., Schlesinger, F., Källberg, M., Cox, A.J., Kruglyak, S., and Saunders, C.T. (2016). Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. Bioinformatics *32*, 1220–1222. https://doi.org/10.1093/bioinformatics/btv710.

70. Wu, Y., Tian, L., Pirastu, M., Stambolian, D., and Li, H. (2013). MATCHCLIP: locate precise breakpoints for copy number variation using CIGAR string by matching soft clipped reads. Front. Genet. *4*, 157. https://doi.org/10.3389/fgene.2013.00157.

71. Yang, L., Luquette, L.J., Gehlenborg, N., Xi, R., Haseley, P.S., Hsieh, C.H., Zhang, C., Ren, X., Protopopov, A., Chin, L., et al. (2013). Diverse mechanisms of somatic structural variations in human cancer genomes. Cell *153*, 919–929. https://doi.org/10.1016/j.cell.2013.04.010.

72. Gardner, E.J., Lam, V.K., Harris, D.N., Chuang, N.T., Scott, E.C., Pittard, W.S., Mills, R.E., and 1000 Genomes Project Consortium; and Devine, S.E. (2017). The mobile element locator tool (MELT): population-scale mobile element discovery and biology. Genome Res. https://doi.org/10.1101/gr.218032.116.

73. Mohiyuddin, M., Mu, J.C., Li, J., Bani Asadi, N., Gerstein, M.B., Abyzov, A., Wong, W.H., and Lam, H.Y.K. (2015). MetaSV: an accurate and integrative structural-variant caller for next generation sequencing. Bioinformatics *31*, 2741–2744. https://doi.org/10.1093/bioinformatics/btv204.

74. Thung, D.T., de Ligt, J., Vissers, L.E.M., Steehouwer, M., Kroon, M., de Vries, P., Slagboom, E.P., Ye, K., Veltman, J.A., and Hehir-Kwa, J.Y. (2014). Mobster: accurate detection of mobile element insertions in next generation sequencing data. Genome Biol. *15*, 488. https://doi.org/10.1186/s13059-014-0488-x.

75. de Araújo Lima, L., and Wang, K. (2017). PennCNV in whole-genome sequencing data. BMC Bioinf. *18*, 383. https://doi.org/10.1186/s12859-017-1802-x.

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

77. Kehr, B., Melsted, P., and Halldórsson, B.V. (2016). PopIns: population-scale detection of novel sequence insertions. Bioinformatics *32*, 961–967. https://doi.org/10.1093/bioinformatics/btv273.

78. Miller, C.A., Hampton, O., Coarfa, C., and Milosavljevic, A. (2011). ReadDepth: a parallel R package for detecting copy number alterations from short sequencing reads. PLoS One *6*, e16327. https://doi.org/10.1371/journal.pone.0016327.

79. Bartenhagen, C., and Dugas, M. (2016). Robust and exact structural variation detection with paired-end and soft-clipped alignments: SoftSV compared with eight algorithms. Brief. Bioinform. *17*, 51–62. https://doi.org/10.1093/bib/bbv028.

80. Zhang, Z., Wang, J., Luo, J., Ding, X., Zhong, J., Wang, J., Wu, F.X., and Pan, Y. (2016). Sprites: detection of deletions from sequencing data by re-aligning split reads. Bioinformatics *32*, 1788–1796. https://doi.org/10.1093/bioinformatics/btw053.

81. Wala, J.A., Bandopadhayay, P., Greenwald, N.F., O'Rourke, R., Sharpe, T., Stewart, C., Schumacher, S., Li, Y., Weischenfeldt, J., Yao, X., et al. (2018). SvABA: genome-wide detection of structural variants and indels by local assembly. Genome Res. *28*, 581–591. https://doi.org/10.1101/gr.221028.117.

82. Zhao, X., Emery, S.B., Myers, B., Kidd, J.M., and Mills, R.E. (2016). Resolving complex structural genomic rearrangements using a randomized approach. Genome Biol. *17*, 126. https://doi.org/10.1186/s13059-016-0993-1.

83. Zhang, J., Wang, J., and Wu, Y. (2012). An improved approach for accurate and efficient calling of structural variations with low-coverage sequence data. BMC Bioinf. *13*. https://doi.org/10.1186/1471-2105-13-S6-S6.

84. Gillet-Markowska, A., Richard, H., Fischer, G., and Lafontaine, I. (2015). Ulysses: accurate detection of low-frequency structural variations in large insert-size sequencing libraries. Bioinformatics *31*, 801–808. https://doi.org/10.1093/bioinformatics/btu730.

85. Kronenberg, Z.N., Osborne, E.J., Cone, K.R., Kennedy, B.J., Domyan, E.T., Shapiro, M.D., Elde, N.C., and Yandell, M. (2015). Wham: identifying structural variants of biological consequence. PLoS Comput. Biol. *11*, e1004572. https://doi.org/10.1371/journal.pcbi.1004572.

86. Jiang, T., Liu, Y., Jiang, Y., Li, J., Gao, Y., Cui, Z., Liu, Y., Liu, B., and Wang, Y. (2020). Long-read-based human genomic structural variation detection with cuteSV. Genome Biol. *21*, 189. https://doi.org/10.1186/s13059-020-02107-y.

87. Tham, C.Y., Tirado-Magallanes, R., Goh, Y., Fullwood, M.J., Koh, B.T.H., Wang, W., Ng, C.H., Chng, W.J., Thiery, A., Tenen, D.G., and Benoukraf, T. (2020). NanoVar: accurate characterization of patients' genomic structural variants using low-depth nanopore sequencing. Genome Biol. *21*, 56. https://doi.org/10.1186/s13059-020-01968-7.

88. Sedlazeck, F.J., Rescheneder, P., Smolka, M., Fang, H., Nattestad, M., von Haeseler, A., and Schatz, M.C. (2018). Accurate detection of complex structural variations using single-molecule sequencing. Nat. Methods *15*, 461–468. https://doi.org/10.1038/s41592-018-0001-7.

89. Heller, D., and Vingron, M. (2019). SVIM: structural variant identification using mapped long reads. Bioinformatics *35*, 2907–2915. https://doi.org/10.1093/bioinformatics/btz041.

90. Sibbesen, J.A., Maretty, L., and Danish Pan-Genome Consortium; and Krogh, A. (2018). Accurate genotyping across variant classes and lengths using variant graphs. Nat. Genet. *50*, 1054–1059. https://doi.org/10.1038/s41588-018-0145-5.

91. Eggertsson, H.P., Jonsson, H., Kristmundsdottir, S., Hjartarson, E., Kehr, B., Masson, G., Zink, F., Hjorleifsson, K.E., Jonasdottir, A., Jonasdottir, A., et al. (2017). Graphtyper enables population-scale genotyping using pangenome graphs. Nat. Genet. *49*, 1654–1660. https://doi.org/10.1038/ng.3964.

92. Chen, S., Krusche, P., Dolzhenko, E., Sherman, R.M., Petrovski, R., Schlesinger, F., Kirsche, M., Bentley, D.R., Schatz, M.C., Sedlazeck, F.J., and Eberle, M.A. (2019). Paragraph: a graph-based structural variant genotyper for short-read sequence data. Genome Biol. *20*, 291. https://doi.org/10.1186/s13059-019-1909-7.

93. Antaki, D., Brandler, W.M., and Sebat, J. (2018). SV2: accurate structural variation genotyping and de novo mutation detection from whole genomes. Bioinformatics *34*, 1774–1777. https://doi.org/10.1093/bioinformatics/btx813.

94. Larson, D.E., Abel, H.J., Chiang, C., Badve, A., Das, I., Eldred, J.M., Layer, R.M., and Hall, I.M. (2019). svtools: population-scale analysis of structural variation. Bioinformatics *35*, 4782–4787. https://doi.org/10.1093/bioinformatics/btz492.

95. Chiang, C., Layer, R.M., Faust, G.G., Lindberg, M.R., Rose, D.B., Garrison, E.P., Marth, G.T., Quinlan, A.R., and Hall, I.M. (2015). SpeedSeq: ultra-fast personal genome analysis and interpretation. Nat. Methods *12*, 966–968. https://doi.org/10.1038/nmeth.3505.

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

97. Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics *34*, 3094–3100. https://doi.org/10.1093/bioinformatics/bty191.

98. Danecek, P., Auton, A., Abecasis, G., Albers, C.A., Banks, E., DePristo, M.A., Handsaker, R.E., Lunter, G., Marth, G.T., Sherry, S.T., et al. (2011). The variant call format and VCFtools. Bioinformatics *27*, 2156–2158. https://doi.org/10.1093/bioinformatics/btr330.

99. DePristo, M.A., Banks, E., Poplin, R., Garimella, K.V., Maguire, J.R., Hartl, C., Philippakis, A.A., del Angel, G., Rivas, M.A., Hanna, M., et al. (2011). A framework for variation discovery and genotyping using next-generation

DNA sequencing data. Nat. Genet. *43*, 491–498. https://doi.org/10.1038/ng.806.

100. Cingolani, P., Platts, A., Wang, L.L., Coon, M., Nguyen, T., Wang, L., Land, S.J., Lu, X., and Ruden, D.M. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly (Austin) *6*, 80–92. https://doi.org/10.4161/fly.19695.

101. Karczewski, K.J., Francioli, L.C., Tiao, G., Cummings, B.B., Alföldi, J., Wang, Q., Collins, R.L., Laricchia, K.M., Ganna, A., Birnbaum, D.P., et al. (2020). The mutational constraint spectrum quantified from variation in 141,456 humans. Nature *581*, 434–443. https://doi.org/10.1038/s41586-020-2308-7.

102. Zheng, X., Levine, D., Shen, J., Gogarten, S.M., Laurie, C., and Weir, B.S. (2012). A high-performance computing toolset for relatedness and principal component analysis of SNP data. Bioinformatics *28*, 3326–3328. https://doi.org/10.1093/bioinformatics/bts606.

103. Browning, B.L., Tian, X., Zhou, Y., and Browning, S.R. (2021). Fast two-stage phasing of large-scale sequence data. Am. J. Hum. Genet. *108*, 1880–1890. https://doi.org/10.1016/j.ajhg.2021.08.005.

104. Howie, B.N., Donnelly, P., and Marchini, J. (2009). A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS Genet. *5*, e1000529. https://doi.org/10.1371/journal.pgen.1000529.

105. Zhou, W., Nielsen, J.B., Fritsche, L.G., Dey, R., Gabrielsen, M.E., Wolford, B.N., LeFaive, J., VandeHaar, P., Gagliano, S.A., Gifford, A., et al. (2018). Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. Nat. Genet. *50*, 1335–1341. https://doi.org/10.1038/s41588-018-0184-y.

106. Mbatchou, J., Barnard, L., Backman, J., Marcketta, A., Kosmicki, J.A., Ziyatdinov, A., Benner, C., O'Dushlaine, C., Barber, M., Boutkov, B., et al. (2021). Computationally efficient whole-genome regression for quantitative and binary traits. Nat. Genet. *53*, 1097–1103. https://doi.org/10.1038/s41588-021-00870-7.

107. Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y.C., Laslo, P., Cheng, J.X., Murre, C., Singh, H., and Glass, C.K. (2010). Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Mol. Cell *38*, 576–589. https://doi.org/10.1016/j.molcel.2010.05.004.

108. Noé, L., and Kucherov, G. (2005). YASS: enhancing the sensitivity of DNA similarity search. Nucleic Acids Res. *33*, W540–W543. https://doi.org/10.1093/nar/gki478.

109. Pendleton, M., Sebra, R., Pang, A.W.C., Ummat, A., Franzen, O., Rausch, T., Stütz, A.M., Stedman, W., Anantharaman, T., Hastie, A., et al. (2015). Assembly and diploid architecture of an individual human genome via single-molecule technologies. Nat. Methods *12*, 780–786. https://doi.org/10.1038/nmeth.3454.

110. Chaisson, M.J.P., Sanders, A.D., Zhao, X., Malhotra, A., Porubsky, D., Rausch, T., Gardner, E.J., Rodriguez, O.L., Guo, L., Collins, R.L., et al. (2019). Multi-platform discovery of haplotype-resolved structural variation in human genomes. Nat. Commun. *10*, 1784. https://doi.org/10.1038/s41467-018-08148-z.

111. Cameron, D.L., Baber, J., Shale, C., Valle-Inclan, J.E., Besselink, N., van Hoeck, A., Janssen, R., Cuppen, E., Priestley, P., and Papenfuss, A.T. (2021). GRIDSS2: comprehensive characterisation of somatic structural variation using single breakend variants and structural variant phasing. Genome Biol. *22*, 202. https://doi.org/10.1186/s13059-021-02423-x.

112. Klambauer, G., Schwarzbauer, K., Mayr, A., Clevert, D.A., Mitterecker, A., Bodenhofer, U., and Hochreiter, S. (2012). cn.MOPS: mixture of Poissons for discovering copy number variations in next-generation sequencing data with a low false discovery rate. Nucleic Acids Res. *40*, e69. https://doi.org/10.1093/nar/gks003.

113. English, A.C., Menon, V.K., Gibbs, R.A., Metcalf, G.A., and Sedlazeck, F.J. (2022). Truvari: refined structural variant comparison preserves allelic diversity. Genome Biol. *23*, 271. https://doi.org/10.1186/s13059-022-02840-6.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCES	SOURCE	IDENTIFIER
Deposited data		
SV vcf file from the 414 1KG WGS data	This paper	http://jenger.riken.jp/en/data (1KG-SV)
SV vcf file from the 3,258 BBJ WGS data	This paper	http://jenger.riken.jp/en/data (BBJ-SV)
GWAS summary statistics of SV for the binary traits	This paper	http://jenger.riken.jp/en/result (Case-control GWAS ID: 107-133)
GWAS summary statistics of SV for the quantitative traits	This paper	http://jenger.riken.jp/en/result (QTL GWAS ID: 140-199)
NA12878 (data1) WGS data	Illumina platinum genomes	https://www.ebi.ac.uk/ena/browser/view/PRJEB3246 (ERR174336–ERR174340)
NA12878 (data2) WGS data	Illumina	https://www.ebi.ac.uk/ena/browser/view/ERX069505 (ERR091571–ERR091573)
NA12878 (data3) WGS data	Genome in a Bottle Consortium	https://www.ebi.ac.uk/ena/browser/view/PRJNA200694 (SRR2052337–SRR2052352)
NA12878 (data4) WGS data	Broad Institute	https://www.ebi.ac.uk/ena/browser/view/SRR1910373
NA12878 (data5) WGS data	1000 Genomes project	ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp/1000G_2504_high_coverage/data (ERR3239334)
NA12878 (PacBio CCS) WGS data	Pacific Biosciences	https://www.ebi.ac.uk/ena/browser/view/PRJNA540705 (SRR9001768–SRR9001773)
NA19240 (data1) WGS data	Genome Institute at Washington University School of Medicine	https://www.ebi.ac.uk/ena/browser/view/SRR3189761
NA19240 (data2) WGS data	Genome Institute at Washington University School of Medicine	https://www.ebi.ac.uk/ena/browser/view/SRR7782669
Simulated WGS data (Sim-A)	Kosugi et al.[29]	https://drive.google.com/file/d/1xtan87fLd966RPuL360w8HU_saUO-bYy/view?usp=sharing
3,258 BBJ WGS data	BBJ Project: Okada et al.,[52] Koyama et al.[53]	BBJ Project samples: coronary artery disease (n = 1,964), drug eruption (n = 189), colorectal cancer (n = 196), breast cancer (n = 237), prostate cancer (n = 215), gastric cancer (n = 257), dementia (n = 200)
414 1KG WGS data	1000 Genomes project	ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp/1000G_2504_high_coverage/data
NA12878 reference SV set	Kosugi et al.[29]	https://github.com/stat-lab/EvalSVcallers/tree/master/Ref_SV
NA19240 reference SV set	This paper	https://github.com/stat-lab/EvalSVcallers/tree/master/Ref_SV
Sim-A reference SV set	Kosugi et al.[29]	https://github.com/stat-lab/EvalSVcallers/tree/master/Ref_SV
gnomAD-SV vcf file	Collins et al.[22]	https://gnomad.broadinstitute.org/downloads#v2-structural-variants
HGSVC long read-based SV vcf file	Ebert et al.[17]	http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/HGSVC2/release/v2.0/integrated_callset/variants_freeze4_sv_insdel_alt.vcf.gz
Simple/short tandem repeat (STR) data	UCSC Genome Browser	https://hgdownload.soe.ucsc.edu/goldenPath/hg19/database/simpleRepeat.txt.gz
Segmental duplication data	UCSC Genome Browser	https://hgdownload.soe.ucsc.edu/goldenPath/hg19/database/genomicSuperDups.txt.gz
Reference gap data	UCSC Genome Browser	https://hgdownload.soe.ucsc.edu/goldenPath/hg19/database/gad.txt.gz
Gene annotation gff3 file	Ensembl	ftp://ftp.ensembl.org/pub/grch37/release-87/gff3/homo_sapiens/Homo_sapiens.GRCh37.87.gff3

(*Continued on next page*)

Continued

REAGENT or RESOURCES	SOURCE	IDENTIFIER
pLI data	Lek et al.[39]	https://static-content.springer.com/esm/art%3A10.1038%2Fnature19057/MediaObjects/41586_2016_BFnature19057_MOESM241_ESM.zip
phastCons evolutionarily conserved data	Felsenstein and Churchill[54]	https://bds.mpi-cbg.de/hillerlab/120MammalAlignment/Human120way/data/conservation/phastConsElements_hg38_multiz120Mammals.bed.gz
Vista enhancer data	VISTA enhancer browser	https://enhancer.lbl.gov
TF footprints data	Vierstra et al.[51]	https://www.vierstra.org/resources/dgf/consensus_footprints_and_collapsed_motifs_hg38.bed,/consensus_index_matrix_full_hg38.txt
ClinVar data	NCBI	https://www.ncbi.nlm.nih.gov/clinvar
GWAS catalog	NHGRI-EBI GWAS Catalog	https://www.ebi.ac.uk/gwas/download/gwas_catalog_v1.0.2-associations.tsv
dbSNP	NCBI	ftp://ftp.ncbi.nih.gov/snp/.redesign/pre_build152/organisms/human_9606_b151_GRCh37p13/BED
Software and algorithms		
SV detection/genotyping algorithms		
MOPline	This paper	https://github.com/stat-lab/MOPline (https://doi.org/10.5281/zenodo.7820277)
1-2-3-SV	Unpublished	https://github.com/Vityay/1-2-3-SV
BASIL-ANISE	Holtgrewe et al.[55]	https://github.com/seqan/anise_basil
BICseq2	Xi et al.[56]	https://github.com/ding-lab/BICSEQ2
BreakDancer	Chen et al.[57]	https://github.com/genome/breakdancer
CLEVER	Marschall et al.[58]	http://clever-sv.googlecode.com (currently unavailable)
CNVnator	Abyzov et al.[59]	https://github.com/abyzovlab/CNVnator
DELLY	Rausch et al.[60]	https://github.com/dellytools/delly
ERDS	Zhu et al.[61]	https://github.com/igm-team/ERDS
FermiKit	Li[62]	https://github.com/lh3/fermikit
forestSV	Michaelson and Sebat[63]	https://sebatlab.org/data-software/
GATK-SV	Collins et al.[22]	https://github.com/broadinstitute/gatk-sv
GenomeSTRiP	Handsaker et al.[64]	https://software.broadinstitute.org/software/genomestrip
GRIDSS	Cameron et al.[65]	https://github.com/PapenfussLab/gridss
iCopyDAV	Dharanipragada et al.[66]	https://github.com/vogetihrsh/icopydav
inGAP-sv	Qi and Zhao[67]	http://ingap.sourceforge.net
Lumpy	Layer et al.[68]	https://github.com/arq5x/lumpy-sv
Manta	Chen et al.[69]	https://github.com/Illumina/manta
MATCHCLIP	Wu et al.[70]	https://github.com/yhwu/matchclips
Meerkat	Yang et al.[71]	http://compbio.med.harvard.edu/Meerkat
MELT	Gardner et al.[72]	https://melt.igs.umaryland.edu
MetaSV	Mohiyuddin et al.[73]	https://github.com/bioinform/metasv
Mobster	Thung et al.[74]	https://jyhehir.github.io/mobster
PennCNV-Seq	de Araujo Lima and Wang 2017[75]	https://github.com/WGLab/PennCNV-Seq
Pindel	Ye et al.[76]	https://github.com/genome/pindel
PopIns	Kehr et al.[77]	https://github.com/bkehr/popins
readDepth	Miller et al.[78]	https://github.com/chrisamiller/readDepth
SoftSV	Bartenhagen and Dugas[79]	https://sourceforge.net/projects/softsv/
Sprites	Zhang et al.[80]	https://github.com/zhangzhen/sprites

(*Continued on next page*)

Continued

REAGENT or RESOURCES	SOURCE	IDENTIFIER
SvABA	Wala et al.[81]	https://github.com/walaj/svaba
SVelter	Zhao et al.[82]	https://github.com/mills-lab/svelter
SVSeq2	Zhang et al.[83]	https://sourceforge.net/projects/svseq2/files/SVseq2_2
Ulysses	Gillet-Markowska et al.[84]	https://github.com/gillet/ulysses
Wham	Kronenberg et al.[85]	https://github.com/zeeev/wham
cuteSV	Jiang et al.[86]	https://github.com/tjiangHIT/cuteSV
NanoVar	Tham et al.[87]	https://github.com/benoukraflab/NanoVar
Sniffles	Sedlazeck et al.[88]	https://github.com/fritzsedlazeck/Sniffles
SVIM	Heller and Vingron[89]	https://github.com/eldariont/svim
BayesTyper	Sibbesen et al.[90]	https://github.com/bioinformatics-centre/BayesTyper
GrapghTyper2	Eggertsson et al.[91]	https://github.com/DecodeGenetics/graphtyper
Paragraph	Chen et al.[92]	https://github.com/Illumina/paragraph
SV2	Antaki et al.[93]	https://github.com/dantaki/SV2
sv-pipeline	Larson et al.[94]	https://github.com/hall-lab/sv-pipeline
svtools	Larson et al.[94]	https://github.com/hall-lab/svtools
SVtyper	Chiang et al.[95]	https://github.com/hall-lab/svtyper
Other software		
BWA	Li and Durbin[96]	http://bio-bwa.sourceforge.net/
Cromwell	Broad Institute	https://github.com/broadinstitute/cromwell
Minimap2	Li[97]	https://github.com/lh3/minimap2
NGM-LR	Sedlazeck et al.[88]	https://github.com/philres/ngmlr
vcftools	Danecek et al.[98]	https://vcftools.github.io/index.html
GATK HaplotypeCaller	DePristo et al.[99]	https://software.broadinstitute.org/gatk/
liftOver	UCSC Genome Browser	http://genome.ucsc.edu/cgi-bin/hgLiftOver
SnpEff	Cingolani et al.[100]	http://pcingola.github.io/SnpEff/
LOFTEE	Karczewski et al.[101]	https://registry.opendata.aws/hail-vep-pipeline/
nnet	CRAN	https://cran.r-project.org/web/packages/nnet/index.html
SNPRelate	Zheng et al.[102]	https://github.com/zhengxwen/SNPRelate
Beagle 5	Browning et al.[103]	http://faculty.washington.edu/browning/beagle/beagle.html
IMPUTE 5	Howie et al.[104]	https://innovation.ox.ac.uk/licence-details/impute-5/
SAIGE	Zhou et al.[105]	https://github.com/weizhouUMICH/SAIGE
REGENIE	Mbatchou et al.[106]	https://rgcgithub.github.io/regenie/
Homer2	Heinz et al.[107]	http://homer.ucsd.edu/homer/download.html
yass	Noe and Kucherov[108]	https://bioinfo.lifl.fr/yass/

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Chikashi Terao (chikashi.terao@riken.jp).

Materials availability

This study did not generate new unique reagents.

Data and code availability

SV vcf files generated using MOPline with the 414 1KG and the 3,258 BBJ WGS data, as well as GWAS summary statistics of SV for the binary and quantitative traits, are available at the JENGER website (http://jenger.riken.jp/en/).

The MOPline code and related data used in MOPline are available at https://github.com/stat-lab/MOPline (https://doi.org/10.5281/zenodo.7820277). The detailed protocol of MOPline can be found in Data S1 and at https://github.com/stat-lab/MOPline. A sample dataset for testing the execution of MOPline is available at http://jenger.riken.jp/en/data. The BBJ WGS data is available at https://humandbs.biosciencedbc.jp/hum0014-v19 and https://gr-sharingdbs.biosciencedbc.jp/agd0008-v1 through registration and review process in accordance with the database's policies.

METHOD DETAILS

WGS datasets

WGS datasets used in this study are summarized in the key resources table. The real short-read WGS datasets of NA12878 and NA19240 were used for the evaluation of SV detection algorithms. The NA12878 datasets, including Illumina HiSeq and NovaSeq, were downloaded from the European Nucleotide Archive (ENA, https://www.ebi.ac.uk/ena/browser/home), DDBJ (http://www.ddbj.nig.ac.jp) or the 1000 Genomes Project (1KGP) (ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp/1000G_2504_high_coverage/data). The NA12878 datasets included five (data1 to data5) datasets derived from different sources or libraries. The NA12878 data1 to data4 were the same as those used in our previous study,[29] and NA12878 data5 was obtained from 1KGP, which had been generated with the Illumina NovaSeq platform. Two independent read datasets (data1 and data2) of NA19240 and the PacBio CCS long read set of NA12878 were obtained from the European Nucleotide Archive (ENA, https://www.ebi.ac.uk/ena/browser/home), and were aligned to the hs37d5 reference or the GRCh38 reference (GRCh38_full_analysis_set_plus_decoy_hla.fa from 1KGP) using NGMLR v0.2.7 (https://github.com/philres/ngmlr) and Minimap2 v2.24 (https://github.com/lh3/minimap2).

The simulated WGS dataset Sim-A was also used to evaluate SV detection algorithms including MOPline, as used in our previous study[29] (available at https://github.com/stat-lab/EvalSVcallers). The Sim-A diploid genome contained a total of 8,310 SVs (3,526 DELs, 1,656 DUPs, 2,819 INSs, and 309 INVs) ranging in size from 50 bp to 1 Mb, and the Sim-A WGS dataset consisted of 125 bp of paired-end reads with 30× coverage.

BBJ high coverage WGS datasets were obtained from 3,258 individuals with any of seven different diseases, that were enrolled in the BioBank Japan Project, as described in previous study.[52,53] For the dementia samples and 1,764 CAD samples, WGS was conducted on the Illumina HiSeq X Five platform using the Illumina TruSeq DNA PCR-Free Library Preparation Kit to generate 2 × 150-bp paired-end reads with approximately 43× coverage for dementia and 23× coverage for CAD. For the other samples, WGS was conducted on the Illumina HiSeq2500 platform using the Illumina TruSeq Nano DNA Library Preparation Kit to generate 2 × 160-bp paired-end reads (2 × 125-bp paired-end reads for the gastric cancer samples) with 29∼35× coverage.

1KG WGS data were downloaded from the 1KGP ftp site. 1KG WGS data were composed of 99 CEU (Utah residents with Northern and Western European ancestry), 103 CHB (Han Chinese in Beijing, China), 104 JPT (Japanese in Tokyo, Japan), and 108 YRI (Yoruba in Ibadan, Nigeria) WGS data, which were generated on the Illumina NovaSeq platform. Because the provided data were CRAM files aligned to GRCh38, paired-end reads were extracted from the CRAM files and were aligned back to the hs37d5 reference.

Reference SV dataset for real data

The reference SV dataset corresponding to NA12878 was generated as previously described.[29] Briefly, the NA12878 reference SV set was generated mainly by combining the DGV variant data (the 2016-05-15 version for GRCh37 and the 2020-02-25 versions for GRCh38) obtained from the Database of Genomic Variants (http://dgv.tcag.ca/dgv/app/home) with the PacBio SV data identified from the NA12878 assembly generated with long reads[109] for GRCh37 or NA12878 SV set extracted from the long read-based, haplotype-resolved HGSVC SV call set[17] for GRCh38 (Table S2). The HGSVC assembly-based SV call set (variants_freeze4_sv_insdel_alt.vcf.gz) was obtained at http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/HGSVC2/release/v2.0/integrated_callset/, which contained 68,180 INSs and 43,150 DELs from a diverse population of 35 individuals. For the build37-based SV set, the coordinates of this SV set were converted to build37 using liftOver with the hg38ToHg19.over.chain file (downloaded at UCSC: https://genome.ucsc.edu). Long read assembly-based SVs of NA12878 were extracted from the build37-based and the build38-based vcf files, and calls with undefined genotypes were removed. Merging between different datasets was conducted based on a BP distance of ≦ 125 bp for INS and ≧ 70% reciprocal overlap for the other types, and only one of the overlaps (i.e., long read-based variants) was incorporated. Another NA12878 reference based on long reads was used for the evaluation of SVs called with MOPline and several other tools. This reference SV data was created by calling the NA12878 PacBio CCS long read data using Sniffles.[88] The reference SVs of NA19240 was derived from the study of the Human Genome Structural Variation Consortium, where SVs had been detected with multiple short and long read sets[110] (Table S2). A minimal 30 bp of NA19240 variants were extracted from nstd152.GRCh37.variant_call.vcf.gz, which was obtained at the NCBI dbVar site (ftp://ftp-trace.ncbi.nlm.nih.gov//pub/dbVar/data/Homo_sapiens/by_study/vcf). The NA19240 DGV variants (the 2016-05-15 version for GRCh37), including 2,045 DELs, 4,147 DUPs, 990 INSs, and 78 INVs, were integrated with the HGSV variants to generate a nonredundant NA19240 reference SV dataset, as in the NA12878 reference SVs. The reference SVs for NA19240 (38,562) contained 1.5-fold more SVs than the reference SVs for NA12878 (25,736), probably due to the remaining redundant SVs in the HGSV data (Table S2). Therefore, the recall value for SV detection using the NA19240 data would be lower than that of NA12878. A gnomAD-SV vcf file was downloaded from https://gnomad.broadinstitute.org/downloads#v2-structural-variants.

SV calling with simulated and real datasets

The simulated and real WGS datasets were aligned with the hs37d5 reference using BWA-MEM v0.7.5a (http://bio-bwa.sourceforge.net/) to generate bam files. For Meerkat and Mobster, bam files were modified by adding XA tags and removing hard-clipped reads to mimic bam files generated with bwa aln. Using these bam files, we performed SV calls with various short-read-based SV detection algorithms, including MOPline, as described in the previous study.[29] SV call sets were converted to MOPline-compatible vcf files with algorithm-specific conversion scripts, which were included in the available resource packages in this study or in the previous study.[29] Because many short read-based SV detection algorithms, including MOPline, cannot determine the length of INSs at many sites, the length of the INS was recorded as 0 or 1 at sites where only BPs were called. Overlap calls from two SV detection algorithms were selected with the criteria; a BP distance of $\leqq$ 200 bp for INS and $\geqq$ 60% reciprocal overlap for the other types. SV calling of MOPline were performed with different combinations of algorithms; MOPline derivatives with different algorithm combinations were named MOPline-4t, -6t, 6t(G), -7t, -9t, -11t, and -14t (Table S1). SV detection algorithms used in MOPline-7t included CNVnator v0.3.2,[59] GRIDSS v2.10.2,[111] inGAP-sv v3.1.1,[67] Manta v0.29.6,[69] MATCHCLIP v2,[70] MELT v2.0.1,[72] and Wham v1.8.[85] For the 3,258 high-coverage BBJ WGS dataset and the 414 high-coverage 1KG WGS dataset, SVs were called for each sample with MOPline-7t (v1.7). After the addition of read alignment statistics, regarding read coverage and soft-clipped read ends, to each SV site, all the SV call sets from the BBJ samples or the 1KG samples were merged (joint called) and genotyped based on multinominal logistic regression with the short read alignment signals supporting SVs. Finally, SVs were filtered and annotated for SVs overlapping with genes.

SV calling using GATK-SV and sv-pipeline

A total of 100 WGS data from 1KG CEU (n = 99) in addition to 1KG NA18525 were used for SV calling with GATK-SV (https://github.com/broadinstitute/gatk-sv), sv-pipeline (https://github.com/hall-lab/sv-pipeline), and MOPline-7t. These data were included in the 1KG WGS dataset used in this study, but GATK-SV used GRCh38-based data due to its limitations. GATK-SV is an ensemble tool of Manta, MELT,[72] Wham,[85] cn.MOPS,[112] and GATK gCNV, and all of these tools, except MELT, were executed using the provided Docker images. GATK-SV is intended to run on the Google Cloud. We continued to run GATK-SV of the 100 sample set in the cohort mode with a Cromwell server on the Google Cloud over one month while dealing with occasional errors, but the run was not completed. We then executed GATK-SV locally in the single sample mode for each sample. The docker image of GATK-SV was created with build_docker.py (https://github.com/broadinstitute/gatk-sv) and converted to a singularity image. A singularity image of MELT was also created for inclusion in the pipeline. WDL files containing GATKSVPipelineSingleSample.wdl were modified for local execution and the reference panel data were copied to the local server. The json files (GATKPipelineSingleSample.tmp.json and cromwell_config.json) for running with Cromwell and singularity in the single sample mode were created according to the instructions (https://github.com/broadinstitute/gatk-sv#quickstart). A Cromwell server was used to submit jobs for each sample as follows: cromshell submit GATKSVPipelineSingleSample.wdl JOBS/GATKPipelineSingleSample.${sampleID}.json cromwell_config.json dep.zip. The final SV calling set (*.annotated.final_cleanup.vcf.gz) of the 100-sample set and another set (GATK-SK (pass)) containing 'PASS' in the PASS field of the vcf files were joint-called using the truvari collapse command of Truvari,[113] with the options, -p 0 -S 50000000 -O 0.5 -P 0.5. BND type variants indicating the break-ends of undefined variants were removed from the final vcf file of GATK-SV.

sv-pipeline is an ensemble tool of Lumpy,[68] Manta,[69] and CNVnator,[59] based on the svtools algorithm.[94] sv-pipeline is also implemented in a wdl-based workflow consisting of the steps of pre-merge (SV discovery with Lumy and Manta), merge, and post-merge (per-sample genotyping with SVTyper and copy number estimation with CNVnator). Since the provided WDL scripts could not achieve full execution in our computing environment, we converted them to standard scripts so that they could be executed sequentially. The output files of the 100-sample set generated with Manta (v1.6.0) and Lumpy (v0.2.13) in the pre-merge step were filtered, sorted, and merged using svtools (v0.5.1) and associated scripts to produce a single merged vcf file. In the post-merge step, the merged vcf file was split into three vcf files, corresponding to DEL, INS, and other types (DUP and INV), and BND variants were removed. For DEL and other types, genotypes were determined for each sample using SVTyper (v0.7.1) and sample bam files. This step was the most time-consuming step, taking more than one hour per sample and per type. INS genotypes were obtained from Manta output files based on the sample information specified with the SNAME tag in the merged INS vcf, and the INS genotypes for each sample were joint-called across the coordinates of the merged INS vcf file. Copy number annotations were added to the genotyped vcf files, except INS vcfs, using CNVnator (v0.4.1) and svtools. The vcf files generated after genotype-pasting and sv-pruning were filtered for SVs less than 50 bp, and redundant overlapping SVs with a BP distance of $\leqq$ 150 bp for INSs and with $\geqq$ 50% reciprocal overlap for the other types were merged. Finally, the vcf files for each SV type were merged to obtain a single vcf file.

To examine overlap calls between the pipelines, the coordinates of GATK-SV calls were converted to the build37-based ones using liftOver. The overlap calls between the two or three pipelines were determined based on a BP distance of $\leqq$ 150 bp for INS and $\geqq$ 50% reciprocal size overlap for the other types. SV calling accuracy of NA12878, which was contained in the 100-sample set, was determined by manual visual inspection using the alignment data of NA12878 PacBio CCS long read and Illumina short reads after evaluation with the NA12878 reference SVs. The size of NA12878 INSs specific to each pipeline was determined using INSs called using Sniffles with the NA12878 PacBio CCS long reads, followed by manual measurements using the PacBio CCS long read alignment data and the IGV viewer for the remaining INSs with undetermined length.

Analysis of SVs overlapping genomic elements

To examine the overlap with repeat regions, we used the simple/short tandem repeat data (simpleRepeat.txt.gz) and the segmental duplication data (genomicSuperDups.txt.gz), which were obtained from the UCSC Genome Browser site (https://hgdownload.soe.ucsc.edu/goldenPath/hg19/database/). Additional repeats of retroelements, including long interspersed nuclear element (LINE), short interspersed nuclear element (SINE), and long terminal repeat (LTR) were extracted from a hg19-based RepeatMasker file, which was obtained from the UCSC Genome Browser site. We counted as overlaps the DELs or DUPs that have at least 30% of their size overlap with a repeat or the INSs within the repeat. A phastCons evolutionarily conserved region file (phastConsElements_hg38_multiz120Mammals.bed.gz) was downloaded from https://bds.mpi-cbg.de/hillerlab/120MammalAlignment/Human120way/data/conservation/, and converted to hg19 coordinates with liftOver. The Vista enhancer sequences were downloaded from https://enhancer.lbl.gov. The information for protein coding genes and non-coding genes was obtained from GRCh37-based gene annotation file (Homo_sapiens.GRCh37.87.gff3), which was downloaded from the Ensembl site (ftp://ftp.ensembl.org/pub/grch37/release-87/gff3/homo_sapiens). Overlaps between SVs and these non-coding elements were counted if they occupied $\geqq$ 50% of the size of either the SV or the genomic element. The constrained degree of SVs overlapping the functional elements against those located in the intergenic regions was determined using ORs with confidence intervals or mean AFs.

SNV and short indel calling

SNVs and short indels were called with the bam files of the 3,258 high coverage BBJ WGS dataset using HaplotypeCaller of GATK ver. 3.8 according to the GATK best practice with a minor modification (https://software.broadinstitute.org/gatk/). Variant calls from all the individuals were joint-called using GATK to merge them into a single vcf file. From joint-called variants, variants with DP < 5 and GQ < 20 or with DP > 60 and GQ < 95 were filtered out. Base quality score calibration of the filtered joint-called variant data was conducted again using GATK. The resulting variant vcf file was annotated using SnpEff v4.3m (http://pcingola.github.io/SnpEff/). LoF variants were annotated based on the SnpEff-annotation 'Effect', describing any of the terms 'stop_gained', 'stop_lost', 'disruptive_inframe_insertion', 'disruptive_inframe_deletion', 'frameshift_variant', and 'splice'. A VEP plugin, LOFTEE (https://registry.opendata.aws/hail-vep-pipeline/), was used for further LoF annotation, in which the annotated variants with LoF=HC or IMPACT=HIGH were selected as LoF variants. Variant sites with multiple alleles, containing both SNVs and indels, were excluded from the downstream analyses. For the 1KG SNVs, genotyped short variant vcf files were downloaded from ftp://ftp.1000genomes.ebi.ac.uk//vol1/ftp/release/20130502. SNVs corresponding to the 1KG-SV 414 sample set or 1KG 100 sample set were extracted from the vcf files.

Compound heterozygous variants

Compound heterozygous SVs and SNVs/indels were searched using LoF SVs and LOFTEE LoF SNVs/indels for protein coding genes. In samples with heterozygous LoF SVs and/or LoF SNVs/indels pairs in a gene, we checked whether they were in different alleles using an integrated vcf file of SVs and SNVs/indels that was phased with Beagle 5,[103] as described below. Finally, these variants were verified by manual visual inspection using the IGV viewer.

Evaluation of SV detection algorithms

The evaluation of SV calling accuracy was performed as described.[29] Briefly, DELs and DUPs were divided into three fractions each according to their size, and precision and recall were calculated for each SV type and size range (S, M, and L). Translocations were not evaluated because there are few known translocations in the databases. For INVs, the SV reference datasets does not contain enough INVs, which may make it difficult to make a reliable evaluation. Precision was calculated by dividing the number of true positive (TP) calls by the total number of calls, and recall was calculated by dividing the number of TP calls with the total number of reference SVs of the corresponding types or size-range. The TP calls were judged when the called DELs, DUPs, and INVs had $\geqq$ 50% reciprocal overlaps with the reference SVs, or when the BPs of the called INSs were placed within 200 bp of the reference INS BPs. Because INSs and DUPs are sometimes complementary and could be confusedly invoked by several different algorithms. Thus, when the called INSs had no matched INS references, we also searched them against the BPs of the reference DUPs. Similarly, called DUPs were also searched against the reference INSs when the called DUPs had no matched DUP references. In this case, a second BP of the reference INS was given at the upstream and the downstream of the INS length from the INS BP so that the reference INSs could be regarded as a provisional DUP. The precision and recall values for many algorithms varied depending on threshold values of the RSS (a minimum number of Reads Supporting an SV) or related scores.[29] To determine the best precision/recall points for each algorithm and each SV category, we selected an RSS threshold at which the numbers of calls for an SV type approximates but does not exceed 90% of the expected SV number in an individual (DEL: 3500, DUP: 550, INS: 3000, and INV: 100, estimated from the previous studies). The simple merging method (Simple-merge-7t) is a non-redundant simple merge of the SV call set from the seven algorithms used in MOPline-7t, filtered by RSS $\geqq$ 3. SV genotyping was evaluated using the Sim-A data, as in our previous study,[29] because real SV genotypes suitable for evaluation were not available.

To determine true positive (TP) calls from NA12878 SVs detected with MOPline, GATK-SV, and sv-pipeline, false positive (FP) calls evaluated using the NA12878 reference SV set were further evaluated using the second NA12878 reference set created using the NA12878 PacBio CCS long reads and Sniffles. The calls evaluated as FP were further validated with visual inspection with the IGV viewer using alignment data of the NA12878 PacBio CCS long reads and the NA12878 short reads. The visual validation was

determined by the presence or absence of evidence of alignments, including secondary alignments, indicating the presence of SV. Large DELs and DUPs with multiple long read split ends at the BPs or clear change in short and long lead depth were judged TP. For INS, only when the BPs of the test INS and the long-read alignment-derived INS were placed within 200 bp and the size of the alignment INS was ≧10 bp, it was determined to be a TP. The visual inspection for NA12878 MOPline calls revealed that a few percent of the total INS TP calls appeared to be smaller than 50 bp. For the other SV types, test SVs were determined as TP when more than half of the lengths of either the test SV or the long-lead alignment-derived SV overlapped and the ratio of the lengths of the SVs was between 0.5 and 2.0. If the long read alignment region corresponding to the test DEL was highly polymorphic with clustered SNVs and short indels, and the short-read alignment region corresponding to the test DEL was low coverage, the test DEL was determined as a TP although such a DEL could be an INV or a complex SV. For INVs and complex SVs, TP was determined if the corresponding long read alignment region was highly polymorphic with clustered SNVs and short indels or if split read alignments of long reads are observed around both BPs of the test SVs. For complex SVs, TP was determined by the presence of DELs and INSs 0.5 to 2.0 times the test SV size around the test SV position in the long read alignments. Short DUPs are detected as INSs in long read data, and in the short tandem repeat regions the sizes of DUP in short read data and INS in long read data are often considerably different (Figure S22). In view of this fact, TP for DUP was also determined if a long-read-derived INS longer than the test DUP was placed within 200 bp of either of the BPs of the test DUP, or if the length of the test DUP was less than 100 bp and the length of a closely adjacent long-lead-derived INS was ≥40% of the test DUP.

MOPline algorithm

Selection of overlap calls by MOP method

The main algorithm of MOPline is based on the MOP method, which merges overlapping SV calls from selected pairs of SV detection algorithms for each SV type and size range. We have currently prepared seven preset pipelines using four to fourteen algorithms (MOPline-4t, -6t, 6t(G), -7t, -9t, -11t, and -14t) although MOPline can use any numbers and any types of existing algorithms. The SV calling results from each algorithm were converted to a MOPline-specific vcf file, where RSS was appended to each call with a 'READS' key. For algorithms not reporting RSS, provisional RSS values, that were converted from some SV-supporting scores specific to the algorithm, were added to each call, as described previously.[29] To ensure a high level of SV detection accuracy, SV calls from each algorithm were filtered by RSS thresholds prior to the selection of overlap calls. RSS thresholds were selected a few points lower than the optimal RSS that would yield the highest sum of precision and recall values for each SV category and SV detection algorithm. If a given level of the RSS threshold achieved sufficiently high precision for a particular SV type and a particular algorithm in the evaluation analysis (e.g., > 95% precision for INS calling by inGAP-sv), MOPline accepted the corresponding SV calls without selecting that overlap calls. To simplify this step, MOPline did not merge the selected overlap calls, but rather used a merge-selection procedure that first merged all SV calls from multiple algorithms for each SV type and then selected the overlap calls for the pair of specified algorithms. In the merging step, overlapping calls (≧ 50% reciprocal overlap for DELs, DUPs, and INVs, and BP ±200 bp for INSs) were merged into one call without redundancy, and information (algorithm name, RSS, etc.) of the merged calls was added. In the selection step, overlap calls of specified algorithm pairs were selected from the merged SV calls for each SV category, satisfying the RSS criteria for each algorithm. Finally, all the SV calls for every SV category were merged into a single vcf file. The choice of algorithms and RSS parameters used in MOPline was optimized primarily based on the results of evaluation with the real data.

Adding read depth and split read information

MOPline adds read depth information DPR (ratio of read depths between SV internal and external regions) to each SV site. For this purpose, the average read depth in the 50-bp window for each chromosome in the bam file was calculated and was recorded in separate "cov" files, where the mean read depth was divided into two cases for all aligned reads and > 0 mapping quality reads only. Using the recorded read depth values, we calculated the DPR value by dividing the average internal depth of SV by the average flanking depth. The 5′-flanking region of the DEL/DUP corresponded to the region from 1 Kb (10% of SV size for > 10 Kb SV) upstream of a corrected first BP (100 bp [400 bp for CNVnator calls] upstream of the first BP) to the corrected first BP. The 3′-flanking region of a DEL/DUP corresponded to the region from a corrected second BP (the first BP + SV size + 100 bp) to 1 Kb (10% of SV size for > 10 Kb SV) downstream of the corrected second BP. For DELs, DPRs was calculated with > 0 mapping quality reads. In addition, MOPline adds the inconsistent DPR rate (DPS), which is the ratio of 50-bp regions with inconsistent DPRs in the internal DEL/DUP region. When a 50-bp window of an internal region has a DEL DPR of > 0.8 or a DUP DPR of < 1.1, the region is judged to have an inconsistent DPR. If there are five 50-bp regions with inconsistent DPR in a 1 Kb DEL, DPS of 0.25 is given. MOPline also adds the split read information SRR (ratio of break-ends of soft-clipped reads at SV BPs to the mean read depth of flanking regions). Split reads were divided into 5′-clipped and 3′-clipped ones and recorded in 50 bp windows per chromosome as well as read depth. These statistics, including DPR and SRR, are useful for true/false identification (Figure S21), and are used for both SV genotyping and SV filtering, as described later.

MOPline also integrates genotypes called by several specific algorithms in a way that depends on the accuracy of the genotyping. Based on the evaluation of SV genotyping in the previous study,[29] the genotyping scores of several genotyping algorithms were set as follows: CNVnator DEL: 96, CNVnator DUP: 93, DELLY DEL: DELLY INV: 80, Lumpy DEL: 92, Lumpy INV: 80, Manta DEL: 94, Manta INV: 80, Manta INS: 98, MELT INS: 70. Genotyping information was added to an SV call if all of the following criteria were met: the SV call is derived from algorithm(s) with the genotyping score, the genotypes from multiple algorithms are identical, and

the sum of the genotyping score(s) exceeds 84. The percentage of SVs genotyped at this step was about 58% and some genotypes of these SVs were changed based on the following multinomial logistic regression-based genotyping results.

Joint calling

MOPline performs joint calling of multi-sample SV call sets and generate a single vcf file. Joint calling combines genetically identical (similar) SV calls from multiple samples into a single site. ≧ 1 Kb DELs with DPR ≧ 0.9 or DPS ≧ 0.5, ≧ 100 Kb DELs with DPR ≧ 0.8 or DPS ≧ 0.35, and ≧ 1 Kb DUPs with DPR ≦ 1.1 were prefiltered. Overlapping SV sites from multiple samples were clustered primarily by the following criteria: ≧ 50% reciprocal overlap for DELs, DUPs, and INVs and BP ±200 bp for INSs. Clustering was done by stepwise merging of closer sites for each SV type. For each clustered site, the sites from multiple samples were integrated into a single site. Proximal DUPs and INSs with a BP distance of ≦ 10 bp were integrated into a single site since DUPs are a type of INSs and the calling of DUPs or INSs depends on SV calling algorithm. The median BP and size and the mean DPR were added to the POS and INFO fields of the output vcf file, and the genotype, BP, SV size, DPR, DPS, and SRR for each sample were added to the FORMAT field. In addition, if more than 50% of the median SV size at an SV site overlapped with the gap region of the reference, the SV site was removed. INVs greater than 200 Kb were removed because of the high likelihood of false positives. DELs, DUPs, and INVs were limited to the size of ≧ 50 bp whereas INSs could contain a fraction of < 50 bp INSs because many short read-based INS detection algorithms detect only INS BPs and fail to determine INS sizes precisely.

Genotyping

SV genotyping was performed by multinomial logistic regression with DPR and SRR statistics. The multinomial logistic regression was conducted using the multinom function in the nnet library in R (https://cran.r-project.org/web/packages/nnet/index.html). To prepare training datasets, we first collected the frequency data across DPR and SRR site-level bins (with 0.1 widths) for each type of SVs using MOPline joint call data for 1,494 BBJ samples. DELs and DUPs were further divided into size ranges (≦ 150 bp, 150-1000 bp, and > 1000 bp for DELs, and ≦ 100 bp, 100-2000 bp, and > 2000 bp for DUPs) and each was further divided into the two classes, overlapping repeat and non-repeat regions. The repeat region includes the STR and segmental duplication regions. The DPR-based or SRR-based frequency exhibited a binomial distribution in typical cases, each peak corresponding to heterozygous (Het) or homozygous (Hom) SVs. We selected SVs, whose DPR and/or SRR values exhibited a binomial distribution at the sites, and used the DPRs and SRRs of Het and Hom in each SV class as training data for multinomial logistic regression (only SRRs were used for INSs and INVs). The control training DPR/SRR datasets as reference allele data were obtained from regions where SVs were not present. The DEL and INS training datasets contained at least 10,000 DPR/SRR data for each genotype, whereas some classes of DUP and INV training data, including > 2 Kb Hom DUP in non-repeats and Hom INV in repeats, contained less than 100 DPR/SRR data. Training of the multinomial logistic regression was conducted using the selected DPRs/SRRs and the corresponding genotypes for each subclass (i.e., genotype as the objective valuable, and SRR for INS or SRR and DPR for the other type as the explanation valuables). The mean SRR and DPR values calculated at a site were also used as the explanation valuable. SVs that were more likely to be Ref alleles, such as ≧ 200 bp DELs with DPR ≧ 1.0, ≧ 200 bp DUPs with DPR ≦ 1.0, and INSs with SRR ≦ 0.1, were excluded for genotyping. Finally, genotypes of SVs were predicted with models created using the trained dataset corresponding to the subclasses. Genotype quality (GQ) was calculated using the probability value (Pr) of the predicted genotype with the following formula.

$$GQ = GL1 - GL2$$

where GL1 is the likelihood of the first predicted genotype and GL2 is the likelihood of the second predicted genotype.

$$GL1 = -10 \times \log_{10}(1 - Pr\,1)$$

$$GL2 = -10 \times \log_{10}(1 - Pr\,2)$$

where Pr1 is the probability of the first predicted genotype and Pr2 is the probability of the second predicted genotype.

SMC

Supplementing Missing Calls (SMC) genotypes plausible missing calls (ref alleles) among samples at high confidence SV sites in a joint-called vcf file. SMC is implemented at two levels: the first level finds matched SVs from the single SV detection algorithms used in MOPline, the second predicts SV genotypes based on DPR and SRR statistics. In the first level, when a sample has a reference allele at an SV site in the joint call data and the corresponding SV call is contained in the original SV calls of the sample (for MOPline-7t, CNVnator [DEL, DUP], GRIDSS [DEL, DUP, and INV], inGAP-sv [DEL and INS], Manta [DEL, DUP, INS, and INV], MATCHCLIP [DEL and DUP], MELT [INS], and Wham [DEL, DUP, and INV]) and when each SV must meet the RSS thresholds optimized for each algorithms and each SV category, which are specified in a SVtool_param.txt file, the reference allele was converted to a non-reference SV allele. At the second level, all genotypes at many selected SV sites were predicted using the multinomial logistic regression-based method with DPR and SRR values of the corresponding sites as described. SVs eligible for the second level of SMC were restricted to ≧ 50 bp DELs/DUPs with at least three non-Ref genotyped samples at a site, INSs with AF ≧ 0.01 and with at least five non-Ref genotyped samples at a site, and ≧ 1Kb INVs with AF ≧ 0.8 and with at least five non-Ref genotyped samples at a site. Genotypes assigned in the first level were determined to have a Ref allele if the probability of the Ref allele was ≧ 0.9. Genotypes with a Ref allele were judged as Het allele if the probability of Het allele was ≧ 0.99; otherwise, genotypes were judged as Ref allele. Finally,

the SMC-based genotypes were corrected with the DPR and SRR statistics calculated for each site (third-level SMC). For example, INSs with SRR ≧ 0.3 genotyped as Ref allele were converted to a Het allele if the SRR was greater than the minimum SRR value of the non-SMC-derived Het allele at that site or the mean SRR value minus the standard deviation (SD) of the SRR of the Het allele at that site. INSs genotyped as non-Ref alleles were converted to Ref alleles if the SRR was smaller than the minimum SRR value of the non-SMC Het allele at that site and smaller than the mean SRR value minus 3 SDs of the SRR of the Het allele at that site. DELs of ≧ 1 Kb with ≧ DPR 0.9 or DPS ≧ 0.5 and ≧ 1 Kb DUPs with DPR ≦ 1.1 were converted to Ref alleles. The SMC levels of SV genotypes were labeled in the FORMAT/SAMPLE fields of the vcf files with MC tags ranging from 0 to 3. SVs with unequal proportions of split-reads direction (5′- and -3′-soft-clipped read ends) in BP were considered false-positive calls; INSs with more than 8-fold difference and DELs/DUPs with more than 4-fold difference were considered ref alleles. SV sites with a frequency of ≧ 90% of samples with the unequal proportions of split-read direction were deleted.

Filtering

The SV filtering of MOPline can optionally be applied after any steps to a vcf file with the DPR or DR tags in the INFO or FORMAT fields. DELs and DUPs can be filtered by parameters based on read coverage while INS can be filtered by parameters based on split read signals (Figure S21). The joint call and SMC steps involve primary filtering of DELs/DUPs based on coverage and INSs based on split reads.

(a) Coverage-based DEL/DUP filtering: DELs and DUPs with inconsistent DPR and/or DPS are filtered with the MOPline DPR-based filtering feature. The filtering was applied for ≧ 10 Kb DELs when satisfying any of the following criteria: DPR > 0.75; DPR > 0.85 and DPS > 0.2. The filtering of DELs was not applied when ≧ 70% of the DEL length overlapped the segmental duplications. The filtering of DUPs was applied when satisfying any of the following criteria: DPR < 1.25 and >1 Kb in size; DPR < 1.35, DPS > 0.1, and > 200 bp in size. However, the DPR-based filtering was less effective for small DELs/DUPs and could miss SVs with small changes in read depth, such as translocations or a gain or loss of a single copy from multiple copy segments.

(b) Gap-based filtering: The SV detection algorithms based on read depth was found to miscall DELs that overlapped the gap regions in the reference likely because reads cannot be aligned to the reference gaps. When more than half of the size of the DEL overlapped the gap region, the corresponding DEL was filtered out (this is also conducted in the joint calling step). DUPs in close proximity to the gap regions were more likely to be miscalled. This type of DUP miscalling could occur due to misalignment of sequencing reads to be assigned to a gap region when the gap-flanking sequence is homologous to the unassigned sequence in the gap region, possibly with repetitive nature. The MOPline filtered out gap-flanking DUPs when the flanking gap size is ≧ 20 Kb or ≧ 0.5-fold of the DUP size. The 5′-flanking region corresponds to the region from 1.2 × DUP size upstream of the first BP to 0.2 × DUP size upstream of the first BP. The 3′-flanking region corresponds to the region from 0.2 × DUP size downstream of the second BP to 1.2 × DUP size downstream of the second BP. The gap-based DUP filtering was restricted to ≧ 50 Kb DUPs by default.

(c) DUPs overlapping segmental duplications: DUPs overlapping the segmental duplications were found to be ambiguous and inaccurate. We filtered out ≧ 5 Kb DUPs completely overlapping the segmental duplications with ≧ 3 copies of the unit in the genome (Figure S21).

(d) Overlapping INS-DUP calls: Since DUPs are a type of INSs, INS calls may overlap DUP calls. If the terminal segment of an INS overlaps the adjacent reference sequence at a BP, some SV detection algorithms consider the terminal segment of the INS to be DUP. When a DUP is in a short tandem repeat region, the INS size observed in the long-read data is often larger or shorter than the size of the DUP called with short-read data (Figure S22). In the latter case, the size of the DUP is incorrectly called because the DUP alignment signals often skip one or several copies of the short tandem repeat unit in the short-read data (Figure S22B). Calling of DUP or INS depends on the short read-based SV detection algorithms. In MOPline, integrating multiple SV detection call sets, redundant calls of a DUP and an INS at the same or nearly identical positions are often observed. MOPline deletes redundant INSs of ≦ 100 bp or redundant DUPs of >100 bp.

(e) Overlapping DEL-DUP signals causing pseudo-DEL/DUP calls: Short read-based SV detection algorithms often call both a DEL and a DUP with a similar size at nearly identical sites. We found that this overlapping DEL-DUP call could be caused largely by an interspersed duplication of a distantly located segment on the same chromosome. The interspersed duplication (i.e., insertion) of a local segment generates pseudo-read pair signals to detect both a DEL and a DUP with a size corresponding to the interspersed distance of the DUP (Figure S23). In this case, there is often no clear change in the read coverage of the pseudo-DEL and DUP. We observed at least 76 of this complex type of pseudo-DEL-DUP calls in the NA12878 SV call set detected with MOPline-7t (Table S10). Such miscalled DELs/DUPs were filtered out at the first SV-merging step. However, MOPline was not able to effectively remove such DELs or DUPs when DELs and DUPs of similar size were not called at similar sites in a given sample. Such miscalls can be effectively eliminated in the validation step using the alignment viewer.

(f) Overlapping INV-DUP signals causing pseudo-INV calls: An interspersed DUP often generates false DEL signals, as described above. An inverted interspersed DUP often generates wrong INV signals (Figure S24). We observed at least 22 of this complex type of pseudo-INV calls in the NA12878 SV call set detected with MOPline-7t (Table S10). However, MOPline was unable to effectively remove such pseudo-INV calls. Such miscalls can be effectively eliminated in the validation step using the alignment viewer.

Annotation of SV-overlapping genes

Information on gene regions that overlap with SV sites can be added in the INFO field of the vcf file. MOPline used the gene information for 41,911 genes, excluding pseudogenes and including 20,268 protein-coding genes, from Homo_sapiens.GRCh37.87.gff3 (or Homo_sapiens.GRCh38.104.gff3.gz for GRCh38), which is available at the Ensembl site. For SVs overlapping with exons/CDSs, introns, untranslated regions (UTRs), 5′- or 3′-flanking regions of the genes, the gene name, gene ID, and the overlapped gene region were added to the INFO field with the SVANN tag. These annotations were also added to the FORMAT AN subfield for each sample because the same integration site often has different SV sizes and BPs in different samples. When an SV overlapped with multiple regions of a gene, the region at the higher level of hierarchy (CDS/exon > UTR > intron > flanking) was indicated. Two ranges of flanking length were specified by default (5 Kb and 50 Kb).

INSs and DELs of mobile elements

MELT was used to detect mobile (retrovirus) element insertions (MEIs) for Alu, L1, SVA, and HERVK from the 3,258 high coverage BBJ WGS dataset, resulting in a total of 20,919 MEIs. We precisely measured the content of MEIs in the BBJ INSs using the long read-based HGSVC INS call set because the BBJ INS call set may contain false negative or false positive MEIs that MELT failed to call. The HGSVC INSs matching the BBJ-INSs (maximum distance 50 bp between BPs) were selected and the INS (ALT) sequences were extracted from the HGSVC SV vcf file. These INS sequences were aligned to the Alu, L1, and SVA reference sequences in a pair-wise manner using the yass alignment tool (https://bioinfo.lifl.fr/yass/). The alignments with mismatch rate of $\leqq$ 15% and continuous alignment length of $\geqq$ 50 bp were determined to be MEI sequences. Of the tested 10,274 INSs with AF $\geqq$ 0.01, 4,127 (40.2%) and 3,520 (34.3%) were found to be MEI and Alu sequences, respectively.

Mobile element deletions were detected using repeat element annotation information in rmsk.txt, which was obtained from the UCSC Genome Browser site (https://genome.ucsc.edu). The total length of the annotated regions was approximately 308 Mb for Alu, 513 Mb for L1, and 4 Mb for SVA. In the DELs detected with MOPline from the BBJ WGS data, DELs located on the Alu, L1, and SVA regions in the reference were annotated as Alu DELs, L1 DELs, and SVA DELs, respectively. Only when $\geqq$ 70% of the DEL size overlapped with these mobile element regions, the DELs were considered MEI DELs.

PCA

Principal component analysis (PCA) was performed using SNPRelate (https://github.com/zhengxwen/SNPRelate). The 1KG-SVs with AF $\geqq$ 0.05 were converted to pseudo-SNPs by converting the reference base at the first BP of the SV to another base. For PCA using SNPs, the SNPs used in the Illumina SNP array were selected from the 1KG SNPs with AF $\geqq$ 0.05. These SNPs and the sample information from the vcf files were converted to ped and map files to create gds files. The SNPs were pruned using 'ld.threshold = 0.2' in SNPRelate and PCA was performed.

LD test

Linkage disequilibrium (LD) between the BBJ-SVs and their neighboring SNVs or indels was determined as done in the gnomAD-SV study.[22] SVs were restricted to autosomal SVs with AF $\geqq$ 0.01 that did not overlap repeat regions, including the STR and the segmental duplication regions, and the SVs to be tested were a total of 14,310 SVs. The overlap between SVs and repeats was determined when $\geqq$ 30% of SV size overlapped with a repeat region. LDs between SV and nearby SNVs/indels with a distance of < 1 Mb were measured using vcftools (https://vcftools.github.io/index.html) with options –ld-window-bp 1000000 –min-r2 0.0001, and the top LDs were selected for each SV. The average of the selected top LDs was calculated for each type of SV.

HWE test

The Hardy-Weinberg equilibrium (HWE) of autosomal SV genotypes in the BBJ-SV dataset was tested for each SV type as in the gnomAD-SV study. Chi-square tests were performed to determine how likely the observed frequencies matched the expected HWE values and were calculated as follows: $N * AF^2$ for the number of homozygous variants and $N * AF * (1 - AF) * 2$ for the number of heterozygous variants, where N is the number of samples in the population. The percentage of SVs for which the HWE p-value exceeded the Bonferroni corrected p-value (0.05 divided by the number of SVs tested) was determined for each SV type.

SV-associated disease risk genes

Gene-based burden tests were performed for colorectal cancer (n = 196), breast cancer (n = 237), prostate cancer (n = 215), gastric cancer (n = 257), dementia (n = 200), and CAD (n = 1,964) using BBJ-SVs. SV data from 2,353 non-cancer disease samples (CAD, drug eruption, and dementia) were used as controls for colorectal and gastric cancers. For breast and prostate cancer, only the female (n = 361) and male (n = 1,992) samples in this control set were used as controls, respectively. For dementia and CAD, all samples other than the corresponding samples (n = 3,058 for dementia and n = 1,294 for CAD) were used as controls, respectively. We counted the number of case or control samples in which the SV overlapped with an exon of a gene. Known risk genes for each disease were obtained from ClinVar (https://www.ncbi.nlm.nih.gov/clinvar). For all the overlapped genes, ORs and p-values were determined by the number of cases and controls in which the SVs overlapped with a gene. p-values were determined by two-tailed Fisher's

exact test or chi-square test, depending on whether there were five or fewer cases with the mutated gene. All identified SVs with a significant association were confirmed using the IGV viewer with regional BAM files of up to 5 randomly selected samples for each carrier and non-carrier.

SVs associated with published GWAS signals

The GWAS data file (gwas_catalog_v1.0.2-associations.tsv) was downloaded from the GWAS catalog site (https://www.ebi.ac.uk/gwas/download). This file was converted to a file compatible with the GRCh37 reference using liftOver and with the dbSNP bed files (GRCh37 build 152, downloaded at NCBI: ftp://ftp.ncbi.nih.gov/snp/.redesign/pre_build152/organisms/human_9606_b151_GRCh37p13/BED). From the converted GWAS catalog file, GWAS data matching disease traits in the BBJ WGS data except drug eruption were extracted. For dementia, the data with the ''Alzheimer'' trait were also extracted. For CAD, the data with the ''coronary heart/artery disease'' or ''myocardial infarction'' trait were extracted. The GWAS lead SNPs within 1 Mb of a disease were considered as variants at the same locus and were merged into one site with the lowest p-value. For each disease, we selected the GWAS lead SNPs and its nearly located SVs (upstream 500 Kb to downstream 500 Kb of the lead SNP). We also selected SNVs from the BBJ WGS data that matched the GWAS lead SNPs. LD r^2 values between the pairs of the selected BBJ SNVs and SVs were determined using vcftools with the options '–geno-r2-positions <SNV.vcf> –ld-window-bp 1000000 –min-r2 0.1'. For SNV/SV pairs with LD $r^2 \geqq 0.1$, ORs and p-values were also determined for each SNVs and SVs. The same control sample set used in the disease burden test was used for each disease to calculate ORs and p-values. All disease-associated SVs in LD ($r2 \geqq 0.8$) with GWAS variants were confirmed using the IGV viewer with regional BAM files of up to 5 randomly selected samples for each carrier and non-carrier.

Imputation of SVs

To test whether SVs are associated with common diseases and traits, genotype data from the array were imputed with SV-containing imputation panels. To generate a reference imputation panel in vcf format, SVs (DELs, DUPs, and INSs) in BBJ-SVs were integrated with SNVs and indels detected from the same WGS data, in which the first BP of SVs was considered a pseudo-SNP. Variants with a genotyping rate of <95% in 3,258 samples and multiallelic variants were excluded. Several imputation panels containing variants with different AF thresholds ($\geqq 0.01$, $\geqq 0.001$, $\geqq 0.0003$, and no AF threshold) were generated for each autosome. The imputation panels for each chromosome were phased using Beagle 5.[103] (http://faculty.washington.edu/browning/beagle/beagle.html) with the HapMap genetic map files for GRCh37 before imputation. To evaluate SV phasing accuracy, we used long read-based haplotype-resolved HGSVC genotype data[17] as truth phased data. Since the HGSVC data shared five 1KG samples used in this study (NA18534, NA18939, NA12878, NA19238, and NA19239), we phased the combined SNV and SV data for the 1KG detected in this study. We calculated the flip rate for pairs of SVs and their flanking SNVs on both sides to determine the SV phasing accuracy for the shared five samples, as described in the recent HGSVC study.[24] Briefly, heterozygous variants matched between the HGSVC data and the 1KG data in this study were selected based on a minimal reciprocal overlap ratio of 0.8 for DEL and a maximal BP distance of 20 bp for INS. Since there is no DUP in the HGSVC data, the 1KG DUP was considered as INS. The closest matched SNVs within 20 Kb upstream and downstream of each of the matched SVs were selected. If the phase (on the same or different chromosomes) of a pair of heterozygous SNVs and SVs in this study matched that of a pair of HGSVC truth data, it was counted as true. A total of 4,500 to 6,500 SVs were tested in each sample, resulting in an average true positive rate of 98.3% for DEL and 96.5% for INS.

Quality-controlled array genotype data for 181,622 BBJ samples from participants each with any of 44 diseases was imputed with the phased reference panel using IMPUTE 5 v1.1.4[104] (https://innovation.ox.ac.uk/licence-details/impute-5/). Imputation with IMPUTE 5 was performed with the default options, chunk size 5 Mb (chunk size 3 Mb for chromosome 6 only) and genetic map files for GRCh37 provided by SHAPEIT4 (https://odelaneau.github.io/shapeit4/). Imputed files for chromosome 1 to chromosome 22 were combined into a single vcf file, compressed with bgzip, and indexed with tabix. In GWAS, imputed variants with INFO scores < 0.3 were excluded. To evaluate the accuracy of imputation, array genotype data from 200 randomly selected individuals from the 3,258 individuals used in the reference panel were used as a test data set. The test dataset was imputed with the reference panel of the remaining 3,058 individuals using IMPUTE 5. Because 200 samples of the test data were included in the original reference panel generated with the WGS data, a sample's imputed calls were determined to be a true positive if it matched that of the original reference panel (i.e., reference allele or non-reference allele). Precision was calculated as the percentage of true positives in the total number of imputed calls for each of the 200 test samples, and recall as the percentage of true positives in the total number of non-reference alleles for each of the 200 test samples. Overall precision or recall was the average of precision or recall per individual for each variant type and each AF range. To further validate the false positive imputation calls, we selected one sample from the test samples and checked the alignment signals supporting the SV calls using the IGV viewer with the bam file of the selected sample. SVs supported by only ambiguously aligned reads with mapping quality 0 were excluded from the analysis although it is possible that SVs supported by misaligned reads were evaluated in a difficult-to-align region containing segmental duplications.

GWAS

In GWAS for binary and quantitative traits, we used two imputation datasets (with INFO score $\geqq 0.3$) generated with the $\geqq 0.01$ AF and $\geqq 0.0003$ AF imputation panels, because imputation using the $\geqq 0.0003$ AF imputation panel resulted in low imputation efficiency in the AF range $\geqq 0.01$ (Figure S25). GWAS for 42 diseases was performed with the imputed SV-SNV-indel data for the

BBJ 181,622 samples using SAIGE v0.35.8,[105] which implements a generalized mixed model association method that controls unbalanced case-control ratios, as in a previous study.[43] The GWAS in this study differs from the previous study in the imputation panel and control sample; additional 32,793 controls from a different cohort were included in the study by Ishigaki et al.[43] The selection of control samples for several diseases was done as in the study by Ishigaki et al. In step 1 of SAIGE for fitting the null logistic/linear mixed model, binary trait was used as the trait type, and age, sex, and top 5 principal components were used as covariates for all diseases. This step was performed using plink bed files, that was converted from the imputed vcf file and pruned using plink with the options –indep-pairwise 500 50 0.2 –maf 0.01. The second step was done for each chromosome with the default options. For 60 quantitative traits, we used REGENIE v2.2.4,[106] which implements a machine-learning-based whole-genome regression model with high computational efficiency. The quantitative data for each trait were adjusted for age, sex, top 10 principal components, and disease status for the 47 target diseases of BBJ in a linear regression model, and the resulting residues were normalized, as described.[42,44,45] In step 1 of REGENIE fitting model, bsize 1000 was specified, and the pruned bed files were used with bsize 400 in step 2. From the GWAS results with the $\geqq$ 0.0003 AF imputation dataset, only genome-wide significant variants with AF < 0.01 were selected. Top-ranked SVs were selected from genome-wide significant variants (p-value $\leqq 5 \times 10^{-8}$) as SVs exhibiting the lowest p-value or a similar p-value ($\geqq$ 90% or $\geqq$ 84% of the -log10 p-value of the top variant for quantitative and binary traits, respectively) to the top variant at each GWAS loci. In the GWAS summary statistics, the direction of the effect size was inverted if the reference allele corresponded to an alternative allele (second allele). In addition, genome-wide significant SVs overlapping exons were searched to find SVs hidden in adjacent strongly associated SNPs. We confirmed whether identified top-ranked GWAS SVs were correctly genotyped or not, using the IGV viewer with regional BAM files from the original WGS data of up to 5 randomly selected samples for each carrier and non-carrier. For non-sporadic SVs with supporting SNPs at certain loci, alignment signals supporting SVs for all the SVs were observed in the selected positive sample data, but not in the negative sample data. However, for many sporadic SVs found in CAD, PrCa, and CoCa, alignment signals were inconsistent or ambiguous in several positive or negative samples, and these unreliable SVs were excluded from the analysis. To further assess the accuracy of the top-ranked associated SVs, we searched for array SNPs with strong association signals and in high LD with the SVs. For 45 top-ranked SVs with AF > 0.01, SNPs derived from array-genotype data within 200 Kb upstream or downstream of the SVs were selected. LDs between SV and SNP pairs were measured using vcftools, and the p-values for the corresponding association test were obtained from the GWAS summary statistics. Manhattan plots were drawn with the GWAS summary statistics using the R qqman library, with highlighted genome-wide significant SVs. Regional Manhattan plots were drawn using locuszoom (http://locuszoom.org) with focusing on a genome-wide significant SV specified with the –refsnp and with –flank 500kb options.

Survival analyses of prostate cancer

Among SVs significantly associated with complex traits, we focused on the SV associated with prostate cancer (PrCa) because of its high heritability (~58%) and many PrCa associations observed in previous GWAS. Furthermore, we found that polygenic risk score would predict development of PrCa (unpublished work). Given the possibility that SV may have a strong effect on coding genes, we hypothesized that SV would show a strong association with mortality in PrCa in the follow-up study. In the BBJ follow-up study, approximately 140,000 BBJ subjects were followed up for approximately 12 years to monitor mortality and its causes coded by ICD10. We took a very similar approach (including data usage) to our previous paper.[41] We restricted this analysis to male subjects without malignancy at enrollment and analyzed the association between PrCa mortality and SV using the Cox proportional hazards model with age, disease status at enrollment and smoking as covariates.

GWAS DELs overlapping TF footprints

To infer the causality of genome-wide significant variants, we searched for the GWAS variants that overlap with the transcription factor (TF) footprints. TF footprint data was obtained from the study by Vierstra et al.,[51] which identified 4.6 M consensus TF footprints from 243 biosamples by genomic DNaseI footprinting experiments. Compared to the data from Chip-Seq and ATAC-Seq, the higher resolution of the footprints (typically 7-30 bp per footprint) allows for more clearly defining TF-binding sites affected by variants. First, GWAS SNVs/indels in LD with GWAS SVs were defined using plink with the option, –r2 –ld-window 1000 –ld-window-r2 0.2 –ld-snp-list. For each of the defined GWAS loci, the overlap of the GWAS variants with TF footprints was searched using the data (consensus_footprints_and_collapsed_motifs_hg38.bed and consensus_index_matrix_full_hg38.txt), obtained from https://www.vierstra.org/resources/dgf. The coordinates in the file (consensus_footprints_and_collapsed_motifs_hg38.bed) were converted to those of GRCh37 using liftOver. The histological types derived from the footprints were selected based on the score ($\geqq$ 1.0) in the consensus_index_matrix_full_hg38.txt file. Overlap of DELs and footprints was restricted to only when $\geqq$ 50% of the footprint size overlapped with the DEL. SNVs and short indels that overlapped with footprints were further evaluated to determine whether the TF binding sites in the footprints were functionally disrupted by the mutation. This evaluation was performed using Homer (http://homer.ucsd.edu/homer/). For SNVs, we used the fasta file of the 21 bp genomic sequence centered on the SNV and the file of known vertebrate motifs provided by Homer to execute the homer2 find command. The SNV was considered a variant affecting the TF footprint if the Homer evaluation showed that its SNV functionally altered TF-binding motifs (loss or gain) on the TF footprint. We focused on only the inhibitory effect of the mutation in the TF-binding motif on the TF footprint because it is difficult to assess the effect of a newly created TF-binding motif by the mutation on the already formed TF footprint.

To determine the empirical p-value for TF footprint overlap based on random expectations using a permutation test, we used 21 non-redundant GWAS top-ranked DELs (the 30 Kb DEL associated with BrCa was excluded due to lack of TF footprint-derived histological types related to breast cancer). Simulated DELs corresponding to the GWAS DELs were created at randomly selected positions from the same chromosome with the same length, and regions corresponding to simple repeats, segmental duplications, and gaps were excluded because these regions are difficult to detect and genotype SVs using short reads. DEL overlap events were counted when the DEL overlapped with at least one TF footprint region and the overlapping footprint-derived histological type was related to the corresponding trait. The histological types related to traits were 'Respiratory' for LuCa, 'Connective' for RA, 'Hematopoietic/Musculoskeletal/Connective' for Height, 'Hematopoietic/Hepatic' for ALP, AST, Plt, GGT, TBil, and ZTT, and 'Hematopoietic' for the other traits. This test was repeated 1,000 times, and overlap counts for the 21 DELs were determined for each repeated test. The resulting counts followed a normal distribution with a mean of 5.8 and a standard deviation of 1.59. The p-value for the TF footprints enrichment of the GWAS top DELs (11/21) was approximated under the assumption of standard normal distribution.