Published online 16 October 2022 *Nucleic Acids Research, 2023, Vol. 51, Database issue* **D1109–D1116**
https://doi.org/10.1093/nar/gkac905

PGG.SV: a whole-genome-sequencing-based structural variant resource and data analysis platform

Yimin Wang[1,†], Yunchao Ling[1,†], Jiao Gong[2,3,†], Xiaohan Zhao[2,3], Hanwen Zhou[1], Bo Xie[1], Haiyi Lou[2], Xinhao Zhuang[1], Li Jin[2,3], The Han100K Initiative[§], Shaohua Fan[2,*], Guoqing Zhang[1,*] and Shuhua Xu[2,3,4,5,6,*]

[1]Key Laboratory of Computational Biology, National Genomics Data Center & Bio-Med Big Data Center, Shanghai Institute of Nutrition and Health, University of Chinese Academy of Sciences, Chinese Academy of Sciences, Shanghai 200031, China, [2]State Key Laboratory of Genetic Engineering, Center for Evolutionary Biology, Collaborative Innovation Center of Genetics and Development, School of Life Sciences, Fudan University, Shanghai 200438, China, [3]Human Phenome Institute, Zhangjiang Fudan International Innovation Center, and Ministry of Education Key Laboratory of Contemporary Anthropology, Fudan University, Shanghai 201203, China, [4]Department of Liver Surgery and Transplantation Liver Cancer Institute, Zhongshan Hospital, Fudan University, Shanghai 200032, China, [5]School of Life Science and Technology, ShanghaiTech University, Shanghai 201210, China and [6]Center for Excellence in Animal Evolution and Genetics, Chinese Academy of Sciences, Kunming 650223, China

Received August 14, 2022; Revised September 21, 2022; Editorial Decision September 30, 2022; Accepted October 04, 2022

ABSTRACT

Structural variations (SVs) play important roles in human evolution and diseases, but there is a lack of data resources concerning representative samples, especially for East Asians. Taking advantage of both next-generation sequencing and third-generation sequencing data at the whole-genome level, we developed the database *PGG*.SV to provide a practical platform for both regionally and globally representative structural variants. In its current version, *PGG*.SV archives 584 277 SVs obtained from whole-genome sequencing data of 6048 samples, including 1030 long-read sequencing genomes representing 177 global populations. *PGG*.SV provides (i) high-quality SVs with fine-scale and precise genomic locations in both GRCh37 and GRCh38, covering underrepresented SVs in existing sequencing and microarray data; (ii) hierarchical estimation of SV prevalence in geographical populations; (iii) informative annotations of SV-related genes, potential functions and clinical effects; (iv) an analysis platform to facilitate SV-based case-control association studies and (v) various visualization tools for understanding the SV structures in the human genome. Taken together, *PGG*.SV provides a user-friendly online interface, easy-to-use analysis tools and a detailed presentation of results. *PGG*.SV is freely accessible via https://www.biosino.org/pggsv.

INTRODUCTION

Structural variations have attracted remarkable attention in both evolutionary and medical studies over the last two decades. Many important genetic diseases, including cancer (1,2), autism (3–6), Alzheimer's disease (7), Parkinson's disease (8), amyotrophic lateral sclerosis (9), heart failure (10), neurodevelopmental disorders (11) and autoimmune diseases (12) were found to be associated with SVs. SVs have also played an increasingly important role in aiding clinical diagnoses, such as in the diagnosis of retinopathy (13), identifying protective variants against malaria (14) and predicting deleterious mutations (15–17). Several large-scale projects have addressed the issues of population stratification, local population characteristics, and adaptation using SVs (18,19), including a study on the enrichment of SVs associated with cardiometabolic disease in Finns (20). However, due to the complex structure and unclear phenotypic effects of SVs, current studies generally pay attention to screening enriched SVs from patient samples, making the representative sample control very important. Furthermore, the prediction of the functions of SVs is expected to effectively narrow the scope of pathogenic mutations, and

[*]To whom correspondence should be addressed. Tel: +86 21 31246617; Fax: +86 21 31246617; Email: xushua@fudan.edu.cn
Correspondence may also be addressed to Guoqing Zhang. Email: gqzhang@picb.ac.cn
Correspondence may also be addressed to Shaohua Fan. Email: shaohua_fan@fudan.edu.cn
[†]The authors wish it to be known that, in their opinion, the first three authors should be regarded as Joint First Authors.
[§]Full list of participants (collaborators) of the Han100K Initiative can be found online via https://pog.fudan.edu.cn/han100k.

this has also become another hot topic in the field of genomics.

Several projects focusing on the human genome have released SV datasets, including the Simons Genome Diversity Project (SGDP) (21), the Human Genetic Diversity Project (HGDP) (22) and the 1000 Genomes Project (1kGP) (23). Several SV-oriented databases have also been released, including dbVar (24), the Database of Genomic Variants (DGV) (25), and the Genome Aggregation Database (gnomAD) (26). These datasets and databases provide a good grounding for large-scale SV studies, but some issues remained to be solved, such as the lack of East Asian samples (47 in SGDP, 220 in HGDP and 503 in 1kGP); some databases, such as gnomAD, not using the coordinates based on the GRCh38 reference genome, and limitations in data generation and SV detection. Currently, microarray and next-generation sequencing (NGS) data are the basis for large-scale SV database construction. However, recently long-read sequencing data have been demonstrated to be advantageous in the detection of SVs, but no database has as yet archived SVs based on long-read sequencing, which has led to a considerable number of under-represented SVs in the available databases (27–31). In addition, SV results released from various datasets have difficulty being used in an integrated context due to the inconsistent workflows of SV calling, making it difficult to maximize the utility of high-quality local datasets in studies targeting global populations.

Here, we developed an SV database, *PGG*.SV, and we constructed a reference dataset consisting of 6048 representative samples, of which 1030 were generated by long-read sequencing (Supplementary Table S1). We published a total of 684 047 SV entries available in both GRCh37 and GRCh38. Removing duplicates in different reference genomes, we provided a total of 584 277 independent SV entries (Supplementary Table S2). *PGG*.SV has integrated accessible public data and newly-generated data based on the same workflow. The data complement the East Asian population samples and allow for greater comparability between global ethnic groups (Supplementary Table S3). *PGG*.SV has many valuable applications, including the prediction of SV-associated genes, annotation of potential functions of SVs, and analysis of SVs in natural populations as the control in disease studies. *PGG*.SV also provides visualization tools for human genomic SVs as well as user-friendly SV comparison, filtering and download functions to help users apply SV data to medical and genetic research.

DATA PROCESSING

Variant calling

Detecting SVs from the raw reads is the first step in data processing (Figure 1). For NGS data, the joint use of multiple methods is effective in improving the quality of SV calling (32,33). We used BWA's mem algorithm to map fastq format data to the human reference genomes GRCh37 and GRCh38 (34). The MarkDuplicates function of GATK was used to process the mapped data (35). We then used a variety of SV calling software, including Breakdancer (36), Breakseq2 (37), CNVnator (38), Pindel (39), Lumpy (40) and Manta (41), covering the three main NGS-based SV calling algorithms: read-pair, read-depth, and split-read (42). Main SV types include Deletion (DEL), Duplication (DUP), Insertion (INS), Inversion (INV), multiallelic copy number variation (mCNV) and Translocation (TRA) with their respective length distributions. We filtered the results according to the type and length of SV that each software specializes in, including using Manta's INS results to take advantage of its strength in detecting INS as well as utilizing the property that the read-depth algorithm performs well for detecting large CNVs, thus retaining the results for CNVnator greater than 10 kb. SVs longer than 10 Mb were not considered. For other SV types and those of various lengths, we required them to be validated by at least two programs. Next, we integrated the above results with MetaSV (43) and an in-house pipeline, PGGSVpipeline to obtain individual-level SV results in vcf format. For long-read sequencing data, we used NGMLR to map fastq data to the reference genome and used Sniffles for SV calling (44). In all, 993 samples based on long-read sequencing had corresponding NGS data, and the results were integrated and carefully compared to improve the quality of the data (45). For conflicting results, we gave priority to the SVs obtained from long-read sequencing data.

Data integration

Unlike the variant positions recorded by SNPs, SVs in each individual are stored as regions with start and endpoints. This leads to the possibility that SVs may be assigned to different locations in the genome for samples due to differences in genetic composition and sequencing quality, even though they were inherited from the same SV event. To solve this problem, an integration algorithm was developed and applied to the data for *PGG*.SV (Supplementary Figure S1). We used vcf files under the same reference genome as input, extracted SVs of all individuals into an SV list, and merged clusters of SVs that were close to each other into a single SV event. At each iteration, we merged the closest SVs and repeated the process until no SV could be merged (the threshold was 50% reciprocal overlap). For the integration of INSs, we require the positions of INS to be located within 50 bp from each other and the length difference of inserted sequences to be less than 30%. Then for the merged SVs, we calculated the starting and ending positions based on the average of all samples carrying the current SV; this can avoid the influence of extreme values of the samples on the combined results. The final output SV list is a set of independent SV data sets that can represent all of the input sample individuals.

Quality control

Since our samples contain multiple independently acquired populations and multiple public data sets (Supplementary Table S1), a quality control process was applied to the integrated SV data set to control for batch effects. We first performed a genotyping process to determine the reliability of the SV results. Multiple graph genome-based SV genotyping methods have been shown to work well (46–48), and thus we used the graph-based SV genotyper Paragraph (49) to perform genotyping for DEL/DUP/CNV less than 200

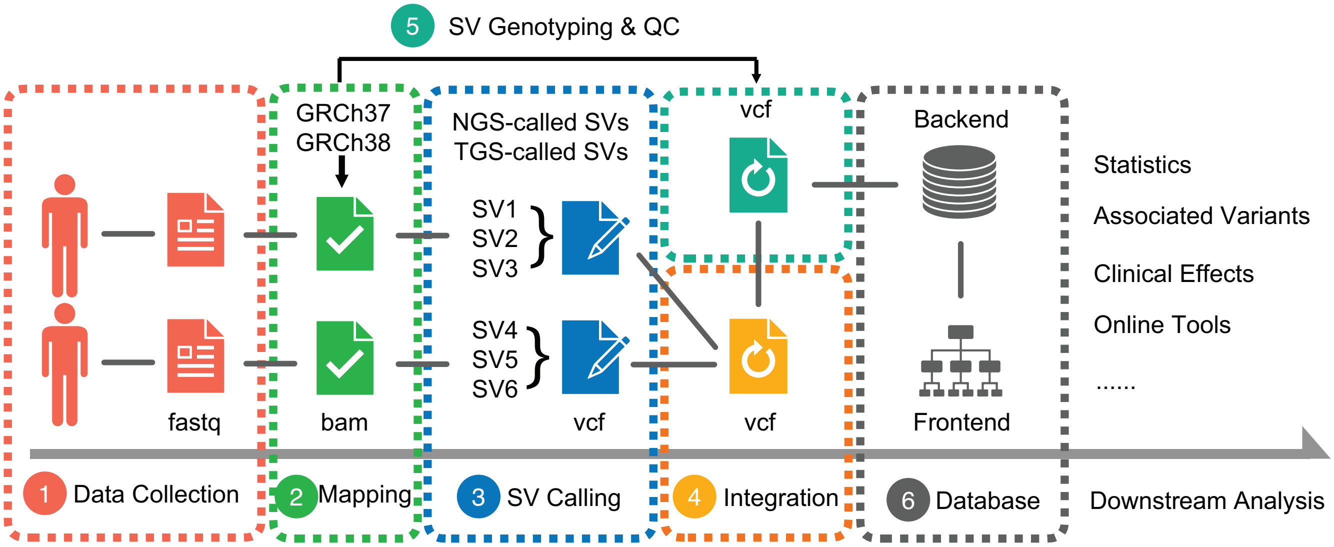


Figure 1. The analytical framework for data collection, processing, integration and quality control. Numbering indicates the order of steps.

bp as well as all INSs. For DEL/DUP/CNV above 200 bp, CNVnator was used to obtain better genotyping results (38). For INV, we used SVtyper for genotyping (50). All of the genotyping results were used only to correct for the presence and copy numbers of SVs (Figure 1).

Additional procedures for quality control included filtering SVs in the Gap region of the genome (cumulative gap length over 50%) and filtering SVs with a missing data rate >40% in the samples. In the data with the reference genome of GRCh37, we found that DEL/DUP/CNV of lengths up to 300 bp constituted a major part of the low-quality data. We added filtering for this part of the SV, including the requirement that the mappability of the SV region was ≥60% and that the SV was not located in a simple repeat region (>50% cumulative length). Meanwhile, we required SVs to have the same copy number in paired sequencing of seven identical samples in our data. These data–after the above measures–were further examined using principal component analysis, and no significant batch effects were found (Supplementary Figure S2).

Summary

We performed SV calling, integration, and quality control processes on 2095 and 4515 samples (Supplementary Table S3) using GRCh37 and GRCh38 as reference genomes and obtained 202 721 and 481 326 SVs, respectively (Supplementary Table S2). To make evaluations and facilitate applications of SV query in a more flexible way, SV calling were performed in 562 samples with sequence reads mapped to both GRCh37 and GRCh38. After positional transformation using CrossMap (51), nearly half of the SVs in GRCh37 can be found in GRCh38, indicating a higher frequency of common variants, while the other unmatched SVs were more likely rare variants at a lower frequency (Supplementary Figure S3). For GRCh38, as the sample size increased, a greater number of low-frequency variants were identified. Notably, SVs detected only in long-read sequencing data accounted for 33.3% of the GRCh38-specific SVs, reflecting the advantage of long-read sequencing data in SV detection.

DATABASE CONTENT

Overview

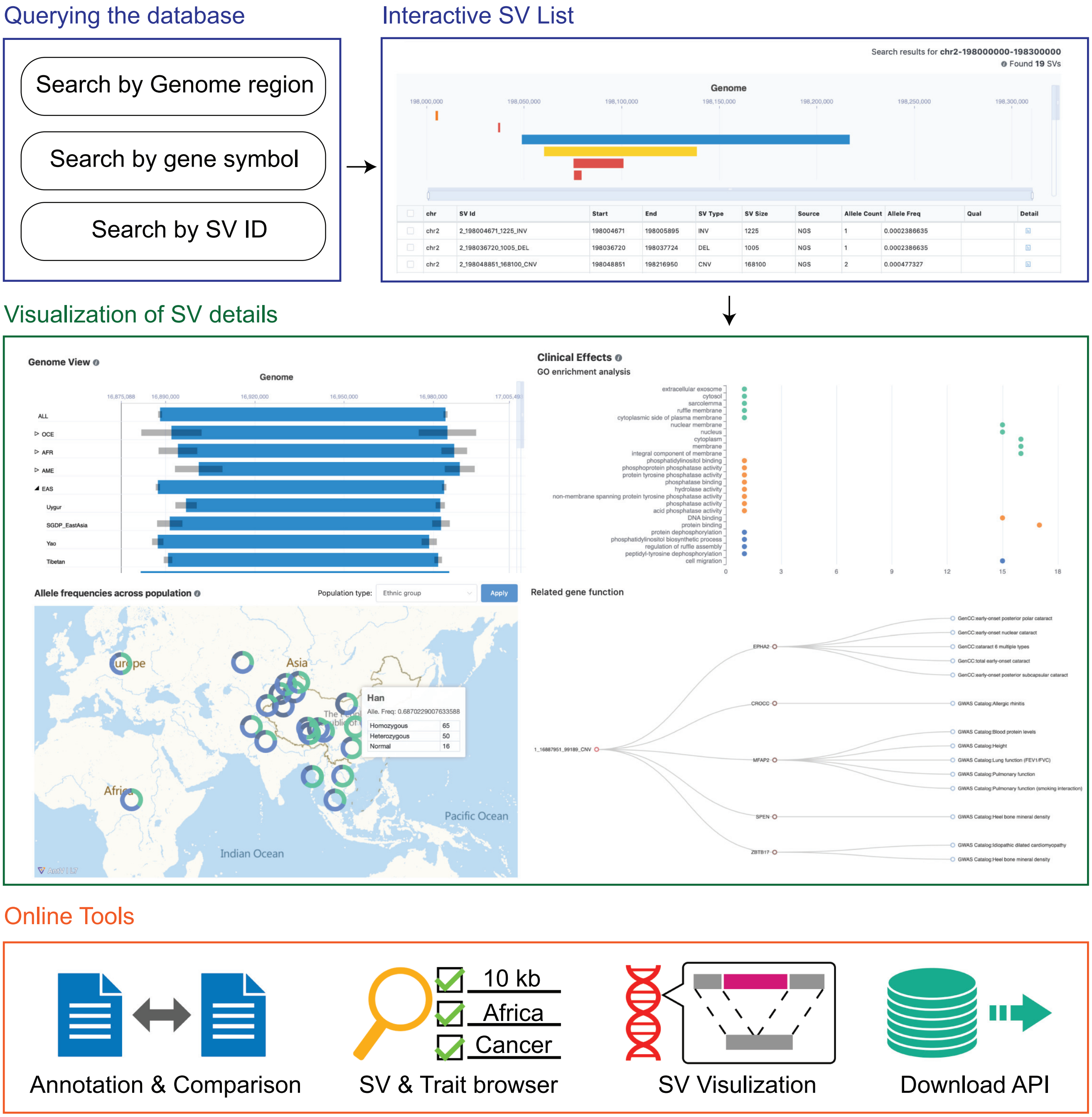
PGG.SV provides three query strategies: (i) query by genomic location, including GRCh37 and GRCh38; (ii) query by gene name and (iii) query by SV ID. We provide filtering to facilitate finding SV results that match expectations in terms of SV length, type, function, frequency and genetic distribution as well as the downloadable and visual presentation of results (Figure 2).

Database construction

The *PGG*.SV web server was built with a backend for frontend (BFF) architecture to enhance and improve the user experience. The backend was developed using Spring boot (https://spring.io/) to view the applications, make checks, and provides standard configurations, while the frontend was built using Vue.js (https://vuejs.org/), which is a lightweight JavaScript framework for building responsive user interfaces. The structure variation data and the associated population and geographic information were stored in MongoDB (https://www.mongodb.com/), which is a general-purpose, document-based distributed database. All of the modules were packaged into docker to ensure flexibility in website deployment.

Genome view

To date, no dataset has considered the genomic location differences of SVs in different populations. The locations of SVs can directly affect gene coding and regulatory regions (52–54) and may also affect phenotypes by altering the spatial structure of the genome (55). *PGG*.SV collects the distributions of starting and endpoint positions for all SVs and


Querying the database
Search by Genome region
Search by gene symbol
Search by SV ID
Interactive SV List
Search results for chr2-198000000-198300000
Found 19 SVs
Genome
198,000,000
198,050,000
198,100,000
198,150,000
198,200,000
198,250,000
198,300,000
Visualization of SV details
Genome View
Genome
16,875,088
16,890,000
16,920,000
16,950,000
16,980,000
17,005,493
ALL
OCE
AFR
AME
EAS
Uygur
SGDP_EastAsia
Yao
Tibetan
Allele frequencies across population
Population type:
Ethnic group
Apply
Europe
Asia
Africa
Indian Ocean
Pacific Ocean
The People's Republic of China
AntV | L7
Clinical Effects
GO enrichment analysis
extracellular exosome
cytosol
sarcolemma
ruffle membrane
cytoplasmic side of plasma membrane
nuclear membrane
nucleus
cytoplasm
membrane
integral component of membrane
phosphatidylinositol binding
phosphoprotein phosphatase activity
protein tyrosine phosphatase activity
phosphatase binding
hydrolase activity
non-membrane spanning protein tyrosine phosphatase activity
phosphatase activity
acid phosphatase activity
DNA binding
protein binding
protein dephosphorylation
phosphatidylinositol biosynthetic process
regulation of ruffle assembly
peptidyl-tyrosine dephosphorylation
cell migration
0
3
6
9
12
15
18
Related gene function
1_16887951_99189_CNV
EPHA2
CROCC
MFAP2
SPEN
ZBTB17
GenCC:early-onset posterior polar cataract
GenCC:early-onset nuclear cataract
GenCC:cataract 6 multiple types
GenCC:total early-onset cataract
GenCC:early-onset posterior subcapsular cataract
GWAS Catalog:Allergic rhinitis
GWAS Catalog:Blood protein levels
GWAS Catalog:Height
GWAS Catalog:Lung function (FEV1/FVC)
GWAS Catalog:Pulmonary function
GWAS Catalog:Pulmonary function (smoking interaction)
GWAS Catalog:Heel bone mineral density
GWAS Catalog:Idiopathic dilated cardiomyopathy
GWAS Catalog:Heel bone mineral density
Online Tools
10 kb
Africa
Cancer
Annotation & Comparison
SV & Trait browser
SV Visulization
Download API

	chr	SV Id	Start	End	SV Type	SV Size	Source	Allele Count	Allele Freq	Qual	Detail
□	chr2	2_198004671_1225_INV	198004671	198005895	INV	1225	NGS	1	0.0002386635		
□	chr2	2_198036720_1005_DEL	198036720	198037724	DEL	1005	NGS	1	0.0002386635		
□	chr2	2_198048851_168100_CNV	198048851	198216950	CNV	168100	NGS	2	0.000477327		

Han

Alle. Freq: 0.6870229007633588

Homozygous	65
Heterozygous	50
Normal	16

Figure 2. Site map of PGG.SV. The database consists of three main functional modules: (i) multiple ways of querying the database and displaying the results (blue box); (ii) visualization of SV details (green box) and (iii) an online analysis tool (orange box). The arrows represent the display order during querying SV, and the online analysis tool can be used freely without login.

calculates means and 95% confidence intervals for ethnic populations, intercontinental populations (Supplementary Table S3), and all of the samples.

Allele frequency

The enrichment of SVs in specific populations reflects environmental adaptations in evolution (56–59). Comparison of population SV frequencies helps us to understand natural selection and the way these variants received their functions. We used the copy number in SV calls and genotyping results as the basis for frequency statistics of copy number variants (including DEL and DUP). For other SV types, the statistics were performed by the presence of variants as for SNVs. Frequency statistics are also available in multiple units and can be toggled for display.

SV-associated variants

Current functional studies such as GWAS consider the association between SNVs and phenotypes, while the addition of SVs can effectively expand the association signal between genomic variants and phenotypes (52,60). We performed SV-SNV/INDEL linkage analysis with PLINK (61) to obtain variants that were highly correlated with the current SV within 1Mb. The existing annotations of other variant types allow us to indirectly understand the function of the target SV. Some functional SVs may be identified by association with other surrounding SVs/SNVs/INDELs, a situation that can help to further explain the functional mechanisms relating these variants to the phenotype. In other cases, SVs may play an important role in the modification of phenotypes, but are not taken into account by the GWAS analysis, resulting in some of the missing heritability (29,52,62). Users can go directly to the corresponding SNV page in our sister database *PGG*.SNV (https://www.pggsnv.org/) (63) to obtain richer information.

Clinical effects

SVs can affect a variety of phenotypes through gene expression and regulation, where the main approaches include direct alteration of gene coding regions and dosage effects triggered by changes in the copy number of regulatory elements (62,64–67). In other words, an SV is expected to be more likely to produce a potential clinical effect if it is located in the coding or regulatory element region of a gene. With the promoters, enhancers, and super-enhancers from GeneHancer (68) and SEdb (69), we combed through genes directly covered by the target SV as well as genes affected by regulatory elements covered by the SV to suggest potential clinical functions of the focal SV. We performed GO enrichment analyses of genes that may be affected by this SV (70,71), and we also annotated the phenotypic and clinical effects of these genes in the GWAS Catalog (72), ClinGen (73) and GenCC (74).

SV sequences

PGG.SV provides downloads of SV sequences for downstream analysis. INS sequences were generated by Manta and Sniffles during SV calling, and other SV types were obtained from the corresponding regions on the GRCh37/GRCh38 reference genome by SAMtools (75).

ONLINE TOOLS

Annotations and comparative analysis

PGG.SV provides online SV annotation and comparison tools. Users can employ the Annotation function to search for SV presence and related genes in *PGG*.SV and databases including dbVar, DGV, and gnomAD. For case–control analysis aimed at requirements such as disease research, the Comparative analysis function can provide control data for any sample size and genetic component (Supplementary Figure S4). The results of the annotations and comparisons will be generated in a short time for the user to download.

SV and trait browser

A table browser interface is a quick way to search, filter, sort and export SVs. Users can intuitively select populations, variant types, variant lengths, allele frequencies, functional annotations, and other characteristics of SVs; alternatively, they may search by specific phenotypes or diseases, thus helping researchers to quickly locate SVs of interest (Supplementary Figure S5).

SV structure visualization

PGG.SV provides users with a convenient tool for visualizing the SV structure. With Miropeats (76), users can generate a visual structure presentation of two DNA sequences with different SV states online. The tool achieves a clean, intuitive presentation for all SV types, including complex SVs (Supplementary Figure S6). Other complex SV events are non-tandem replication or translocation, for which we use MUMmer (77) to help users visually identify distant SV events that occur across the human genome. Both visualization tools can be accessed online.

FUTURE DIRECTIONS

In the current version, we provide not only detailed SV results based on representative samples, including population characteristics and SV function predictions, but also easy-to-use analysis tools and a user-friendly online interface. In our future plans, we will expand population sampling and move forward to a more comprehensive collection of diverse and highly accurate, complete, haplotype-phased genome sequence assemblies. Moreover, our current workflow does not cover well some of the SV types, for example, Translocations are also an important part of SVs to be considered in the future updates of the PGG.SV. In fact, we are making efforts to complete the calling of Translocations from the data in hands. We will release these results and update the database frequently.

DATA AVAILABILITY

PGGSVpipeline is an open-source workflow for NGS-based SV calling available on the group website (https://pog.fudan.edu.cn/#/software) and the GitHub repository (https://github.com/Shuhua-Group/PGGSVpipeline). The use of the data by this work is approved by the Ministry of Science and Technology of the People's Republic of China (No. 2022BAT2236).

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors are grateful for the helpful discussions of many group members during the early stage of the design and development of *PGG*.SV.

FUNDING

This study was supported by the Basic Science Center Program (32288101); the Strategic Priority Research Program [XDB38030100] of the Chinese Academy of Sciences (CAS); National Natural Science Foundation of China

(NSFC) [32030020, 31961130380]; UK Royal Society-Newton Advanced Fellowship [NAF\R1\191094]; Shanghai Municipal Science and Technology Major Project [2017SHZDZX01]. Funding for open access charge: NSFC [31961130380]; UK Royal Society-Newton Advanced Fellowship [NAF\R1\191094].
Conflict of interest statement. None declared.

REFERENCES

1. Hadi,K., Yao,X., Behr,J.M., Deshpande,A., Xanthopoulakis,C., Tian,H., Kudman,S., Rosiene,J., Darmofal,M., DeRose,J. *et al.* (2020) Distinct classes of complex structural variation uncovered across thousands of cancer genome graphs. *Cell*, **183**, 197–210.
2. 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.
3. Mitra,I., Huang,B., Mousavi,N., Ma,N., Lamkin,M., Yanicky,R., Shleizer-Burko,S., Lohmueller,K.E. and Gymrek,M. (2021) Patterns of de novo tandem repeat mutations and their role in autism. *Nature*, **589**, 246–250.
4. Trost,B., Engchuan,W., Nguyen,C.M., Thiruvahindrapuram,B., Dolzhenko,E., Backstrom,I., Mirceta,M., Mojarad,B.A., Yin,Y., Dov,A. *et al.* (2020) Genome-wide detection of tandem DNA repeats that are expanded in autism. *Nature*, **586**, 80–86.
5. Collins,R.L., Brand,H., Redin,C.E., Hanscom,C., Antolik,C., Stone,M.R., Glessner,J.T., Mason,T., Pregno,G., Dorrani,N. *et al.* (2017) Defining the diverse spectrum of inversions, complex structural variation, and chromothripsis in the morbid human genome. *Genome Biol.*, **18**, 36.
6. Leppa,V.M., Kravitz,S.N., Martin,C.L., Andrieux,J., Le Caignec,C., Martin-Coignard,D., DyBuncio,C., Sanders,S.J., Lowe,J.K., Cantor,R.M. *et al.* (2016) Rare inherited and de novo CNVs reveal complex contributions to ASD risk in multiplex families. *Am. J. Hum. Genet.*, **99**, 540–554.
7. Vialle,R.A., de Paiva Lopes,K., Bennett,D.A., Crary,J.F. and Raj,T. (2022) Integrating whole-genome sequencing with multi-omic data reveals the impact of structural variants on gene regulation in the human brain. *Nat. Neurosci.*, **25**, 504–514.
8. Huttenlocher,J., Stefansson,H., Steinberg,S., Helgadottir,H.T., Sveinbjörnsdóttir,S., Riess,O., Bauer,P. and Stefansson,K. (2015) Heterozygote carriers for CNVs in PARK2 are at increased risk of Parkinson's disease. *Hum. Mol. Genet.*, **24**, 5637–5643.
9. Course,M.M., Gudsnuk,K., Smukowski,S.N., Winston,K., Desai,N., Ross,J.P., Sulovari,A., Bourassa,C.V., Spiegelman,D., Couthouis,J. *et al.* (2020) Evolution of a human-specific tandem repeat associated with ALS. *Am. J. Hum. Genet.*, **107**, 445–460.
10. Haas,J., Mester,S., Lai,A., Frese,K.S., Sedaghat-Hamedani,F., Kayvanpour,E., Rausch,T., Nietsch,R., Boeckel,J.N., Carstensen,A. *et al.* (2018) Genomic structural variations lead to dysregulation of important coding and non-coding RNA species in dilated cardiomyopathy. *EMBO Mol. Med.*, **10**, 107–120.
11. Porubsky,D., Sanders,A.D., Höps,W., Hsieh,P., Sulovari,A., Li,R., Mercuri,L., Sorensen,M., Murali,S.C., Gordon,D. *et al.* (2020) Recurrent inversion toggling and great ape genome evolution. *Nat. Genet.*, **52**, 849–858.
12. Li,Y.R., Glessner,J.T., Coe,B.P., Li,J., Mohebnasab,M., Chang,X., Connolly,J., Kao,C., Wei,Z., Bradfield,J. *et al.* (2020) Rare copy number variants in over 100,000 european ancestry subjects reveal multiple disease associations. *Nat. Commun.*, **11**, 255.
13. Zampaglione,E., Kinde,B., Place,E.M., Navarro-Gomez,D., Maher,M., Jamshidi,F., Nassiri,S., Mazzone,J.A., Finn,C., Schlegel,D. *et al.* (2020) Copy-number variation contributes 9% of pathogenicity in the inherited retinal degenerations. *Genet. Med.*, **22**, 1079–1087.
14. Leffler,E.M., Band,G., Busby,G.B.J., Kivinen,K., Le,Q.S., Clarke,G.M., Bojang,K.A., Conway,D.J., Jallow,M., Sisay-Joof,F. *et al.* (2017) Resistance to malaria through structural variation of red blood cell invasion receptors. *Science*, **356**, eaam6393.
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.
16. Han,L., Zhao,X., Benton,M.L., Perumal,T., Collins,R.L., Hoffman,G.E., Johnson,J.S., Sloofman,L., Wang,H.Z., Stone,M.R. *et al.* (2020) Functional annotation of rare structural variation in the human brain. *Nat. Commun.*, **11**, 2990.
17. Middelkamp,S., Vlaar,J.M., Giltay,J., Korzelius,J., Besselink,N., Boymans,S., Janssen,R., de la Fonteijne,L., van Binsbergen,E., van Roosmalen,M.J. *et al.* (2019) Prioritization of genes driving congenital phenotypes of patients with de novo genomic structural variants. *Genome Med.*, **11**, 79.
18. Hsieh,P., Vollger,M.R., Dang,V., Porubsky,D., Baker,C., Cantsilieris,S., Hoekzema,K., Lewis,A.P., Munson,K.M., Sorensen,M. *et al.* (2019) Adaptive archaic introgression of copy number variants and the discovery of previously unknown human genes. *Science*, **366**, eaax2083.
19. Sudmant,P.H., Mallick,S., Nelson,B.J., Hormozdiari,F., Krumm,N., Huddleston,J., Coe,B.P., Baker,C., Nordenfelt,S., Bamshad,M. *et al.* (2015) Global diversity, population stratification, and selection of human copy-number variation. *Science*, **349**, aab3761.
20. Chen,L., Abel,H.J., Das,I., Larson,D.E., Ganel,L., Kanchi,K.L., Regier,A.A., Young,E.P., Kang,C.J., Scott,A.J. *et al.* (2021) Association of structural variation with cardiometabolic traits in finns. *Am. J. Hum. Genet.*, **108**, 583–596.
21. Mallick,S., Li,H., Lipson,M., Mathieson,I., Gymrek,M., Racimo,F., Zhao,M., Chennagiri,N., Nordenfelt,S., Tandon,A. *et al.* (2016) The simons genome diversity project: 300 genomes from 142 diverse populations. *Nature*, **538**, 201–206.
22. 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.
23. Sudmant,P.H., Rausch,T., Gardner,E.J., Handsaker,R.E., Abyzov,A., Huddleston,J., Zhang,Y., Ye,K., Jun,G., Fritz,M.H. *et al.* (2015) An integrated map of structural variation in 2,504 human genomes. *Nature*, **526**, 75–81.
24. Lappalainen,I., Lopez,J., Skipper,L., Hefferon,T., Spalding,J.D., Garner,J., Chen,C., Maguire,M., Corbett,M., Zhou,G. *et al.* (2013) DbVar and DGVa: public archives for genomic structural variation. *Nucleic Acids Res.*, **41**, D936–D941.
25. MacDonald,J.R., Ziman,R., Yuen,R.K., Feuk,L. and Scherer,S.W. (2014) The database of genomic variants: a curated collection of structural variation in the human genome. *Nucleic Acids Res.*, **42**, D986–D992.
26. 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.
27. Audano,P.A., Sulovari,A., Graves-Lindsay,T.A., Cantsilieris,S., Sorensen,M., Welch,A.E., Dougherty,M.L., Nelson,B.J., Shah,A., Dutcher,S.K. *et al.* (2019) Characterizing the major structural variant alleles of the human genome. *Cell*, **176**, 663–675.
28. 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.
29. Logsdon,G.A., Vollger,M.R. and Eichler,E.E. (2020) Long-read human genome sequencing and its applications. *Nat. Rev. Genet.*, **21**, 597–614.
30. Zhao,X., Collins,R.L., Lee,W.P., Weber,A.M., Jun,Y., Zhu,Q., Weisburd,B., Huang,Y., Audano,P.A., Wang,H. *et al.* (2021) Expectations and blind spots for structural variation detection from long-read assemblies and short-read genome sequencing technologies. *Am. J. Hum. Genet.*, **108**, 919–928.
31. Huddleston,J., Chaisson,M.J.P., Steinberg,K.M., Warren,W., Hoekzema,K., Gordon,D., Graves-Lindsay,T.A., Munson,K.M., Kronenberg,Z.N., Vives,L. *et al.* (2017) Discovery and genotyping of structural variation from long-read haploid genome sequence data. *Genome Res.*, **27**, 677–685.
32. Ho,S.S., Urban,A.E. and Mills,R.E. (2020) Structural variation in the sequencing era. *Nat. Rev. Genet.*, **21**, 171–189.
33. 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.

34. Li,H. and Durbin,R. (2010) Fast and accurate long-read alignment with burrows-wheeler transform. *Bioinformatics*, **26**, 589–595.
35. McKenna,A., Hanna,M., Banks,E., Sivachenko,A., Cibulskis,K., Kernytsky,A., Garimella,K., Altshuler,D., Gabriel,S., Daly,M. *et al.* (2010) The genome analysis toolkit: a mapreduce framework for analyzing next-generation DNA sequencing data. *Genome Res.*, **20**, 1297–1303.
36. 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.
37. Abyzov,A., Li,S., Kim,D.R., Mohiyuddin,M., Stütz,A.M., Parrish,N.F., Mu,X.J., Clark,W., Chen,K., Hurles,M. *et al.* (2015) Analysis of deletion breakpoints from 1,092 humans reveals details of mutation mechanisms. *Nat. Commun.*, **6**, 7256.
38. 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.
39. 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.
40. 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.
41. 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.
42. Alkan,C., Coe,B.P. and Eichler,E.E. (2011) Genome structural variation discovery and genotyping. *Nat. Rev. Genet.*, **12**, 363–376.
43. Mohiyuddin,M., Mu,J.C., Li,J., Bani Asadi,N., Gerstein,M.B., Abyzov,A., Wong,W.H. and Lam,H.Y. (2015) MetaSV: an accurate and integrative structural-variant caller for next generation sequencing. *Bioinformatics*, **31**, 2741–2744.
44. 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.
45. 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.
46. Sirén,J., Monlong,J., Chang,X., Novak,A.M., Eizenga,J.M., Markello,C., Sibbesen,J.A., Hickey,G., Chang,P.C., Carroll,A. *et al.* (2021) Pangenomics enables genotyping of known structural variants in 5202 diverse genomes. *Science*, **374**, abg8871.
47. Yan,S.M., Sherman,R.M., Taylor,D.J., Nair,D.R., Bortvin,A.N., Schatz,M.C. and McCoy,R.C. (2021) Local adaptation and archaic introgression shape global diversity at human structural variant loci. *Elife*, **10**, e67615.
48. Kolmogorov,M., Yuan,J., Lin,Y. and Pevzner,P.A. (2019) Assembly of long, error-prone reads using repeat graphs. *Nat. Biotechnol.*, **37**, 540–546.
49. Chen,S., Krusche,P., Dolzhenko,E., Sherman,R.M., Petrovski,R., Schlesinger,F., Kirsche,M., Bentley,D.R., Schatz,M.C., Sedlazeck,F.J. *et al.* (2019) Paragraph: a graph-based structural variant genotyper for short-read sequence data. *Genome Biol.*, **20**, 291.
50. 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.
51. Zhao,H., Sun,Z., Wang,J., Huang,H., Kocher,J.P. and Wang,L. (2014) CrossMap: a versatile tool for coordinate conversion between genome assemblies. *Bioinformatics*, **30**, 1006–1007.
52. Chiang,C., Scott,A.J., Davis,J.R., Tsang,E.K., Li,X., Kim,Y., Hadzic,T., Damani,F.N., Ganel,L., Montgomery,S.B. *et al.* (2017) The impact of structural variation on human gene expression. *Nat. Genet.*, **49**, 692–699.
53. Fotsing,S.F., Margoliash,J., Wang,C., Saini,S., Yanicky,R., Shleizer-Burko,S., Goren,A. and Gymrek,M. (2019) The impact of short tandem repeat variation on gene expression. *Nat. Genet.*, **51**, 1652–1659.
54. Jakubosky,D., D'Antonio,M., Bonder,M.J., Smail,C., Donovan,M.K.R., Young Greenwald,W.W., Matsui,H., D'Antonio-Chronowska,A., Stegle,O., Smith,E.N. *et al.* (2020) Properties of structural variants and short tandem repeats associated with gene expression and complex traits. *Nat. Commun.*, **11**, 2927.
55. Zhang,D., Huang,P., Sharma,M., Keller,C.A., Giardine,B., Zhang,H., Gilgenast,T.G., Phillips-Cremins,J.E., Hardison,R.C. and Blobel,G.A. (2020) Alteration of genome folding via contact domain boundary insertion. *Nat. Genet.*, **52**, 1076–1087.
56. Lan,X. and Pritchard,J.K. (2016) Coregulation of tandem duplicate genes slows evolution of subfunctionalization in mammals. *Science*, **352**, 1009–1013.
57. Mérot,C., Oomen,R.A., Tigano,A. and Wellenreuther,M. (2020) A roadmap for understanding the evolutionary significance of structural genomic variation. *Trends Ecol. Evol.*, **35**, 561–572.
58. Hollox,E.J., Zuccherato,L.W. and Tucci,S. (2022) Genome structural variation in human evolution. *Trends Genet.*, **38**, 45–58.
59. Hsieh,P., Dang,V., Vollger,M.R., Mao,Y., Huang,T.H., Dishuck,P.C., Baker,C., Cantsilieris,S., Lewis,A.P., Munson,K.M. *et al.* (2021) Evidence for opposing selective forces operating on human-specific duplicated TCAF genes in neanderthals and humans. *Nat. Commun.*, **12**, 5118.
60. Auwerx,C., Lepamets,M., Sadler,M.C., Patxot,M., Stojanov,M., Baud,D., Mägi,R., Porcu,E., Reymond,A. and Kutalik,Z. (2022) The individual and global impact of copy-number variants on complex human traits. *Am. J. Hum. Genet.*, **109**, 647–668.
61. Chang,C.C., Chow,C.C., Tellier,L.C., Vattikuti,S., Purcell,S.M. and Lee,J.J. (2015) Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience*, **4**, 7.
62. Gymrek,M., Willems,T., Guilmatre,A., Zeng,H., Markus,B., Georgiev,S., Daly,M.J., Price,A.L., Pritchard,J.K., Sharp,A.J. *et al.* (2016) Abundant contribution of short tandem repeats to gene expression variation in humans. *Nat. Genet.*, **48**, 22–29.
63. Zhang,C., Gao,Y., Ning,Z., Lu,Y., Zhang,X., Liu,J., Xie,B., Xue,Z., Wang,X., Yuan,K. *et al.* (2019) PGG.SNV: understanding the evolutionary and medical implications of human single nucleotide variations in diverse populations. *Genome Biol.*, **20**, 215.
64. 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.
65. Mukamel,R.E., Handsaker,R.E., Sherman,M.A., Barton,A.R., Zheng,Y., McCarroll,S.A. and Loh,P.R. (2021) Protein-coding repeat polymorphisms strongly shape diverse human phenotypes. *Science*, **373**, 1499–1505.
66. 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.
67. Rice,A.M. and McLysaght,A. (2017) Dosage sensitivity is a major determinant of human copy number variant pathogenicity. *Nat. Commun.*, **8**, 14366.
68. Fishilevich,S., Nudel,R., Rappaport,N., Hadar,R., Plaschkes,I., Iny Stein,T., Rosen,N., Kohn,A., Twik,M., Safran,M. *et al.* (2017) GeneHancer: genome-wide integration of enhancers and target genes in genecards. *Database (Oxford)*, **2017**, bax028.
69. Jiang,Y., Qian,F., Bai,X., Liu,Y., Wang,Q., Ai,B., Han,X., Shi,S., Zhang,J., Li,X. *et al.* (2019) SEdb: a comprehensive human super-enhancer database. *Nucleic Acids Res.*, **47**, D235–D243.
70. Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P., Dolinski,K., Dwight,S.S., Eppig,J.T. *et al.* (2000) Gene ontology: tool for the unification of biology. The gene ontology consortium. *Nat. Genet.*, **25**, 25–29.
71. (2021) The gene ontology resource: enriching a GOld mine. *Nucleic Acids Res.*, **49**, D325–D334.
72. Buniello,A., MacArthur,J.A.L., Cerezo,M., Harris,L.W., Hayhurst,J., Malangone,C., McMahon,A., Morales,J., Mountjoy,E., Sollis,E. *et al.* (2019) The NHGRI-EBI GWAS catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res.*, **47**, D1005–D1012.
73. Rehm,H.L., Berg,J.S., Brooks,L.D., Bustamante,C.D., Evans,J.P., Landrum,M.J., Ledbetter,D.H., Maglott,D.R., Martin,C.L., Nussbaum,R.L. *et al.* (2015) ClinGen–the clinical genome resource. *N. Engl. J. Med.*, **372**, 2235–2242.

74. DiStefano,M.T., Goehringer,S., Babb,L., Alkuraya,F.S., Amberger,J., Amin,M., Austin-Tse,C., Balzotti,M., Berg,J.S., Birney,E. *et al.* (2022) The gene curation coalition: a global effort to harmonize gene-disease evidence resources. *Genet. Med.*, **24**, 1732–1742.
75. Li,H., Handsaker,B., Wysoker,A., Fennell,T., Ruan,J., Homer,N., Marth,G., Abecasis,G. and Durbin,R. (2009) The sequence alignment/map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.
76. Parsons,J.D. (1995) Miropeats: graphical DNA sequence comparisons. *Comput. Appl. Biosci.*, **11**, 615–619.
77. Marçais,G., Delcher,A.L., Phillippy,A.M., Coston,R., Salzberg,S.L. and Zimin,A. (2018) MUMmer4: a fast and versatile genome alignment system. *PLoS Comput. Biol.*, **14**, e1005944.