D890–D895 *Nucleic Acids Research, 2023, Vol. 51, Database issue* *Published online 25 July 2022*
https://doi.org/10.1093/nar/gkac638

CMDB: the comprehensive population genome variation database of China

Zhichao Li[1,2,†], Xiaosen Jiang[1,2,†], Mingyan Fang[2], Yong Bai[2], Siyang Liu[2], Shujia Huang[2,*] and Xin Jin[2,3,*]

[1]College of Life Sciences, University of Chinese Academy of Sciences, Beijing 100049, China, [2]BGI-Shenzhen, Shenzhen 518083, Guangdong, China and [3]School of Medicine, South China University of Technology, Guangzhou, Guangdong 510006, China

Received June 20, 2022; Editorial Decision July 05, 2022; Accepted July 22, 2022

ABSTRACT

A high-quality genome variation database derived from a large-scale population is one of the most important infrastructures for genomics, clinical and translational medicine research. Here, we developed the Chinese Millionome Database (CMDB), a database that contains 9.04 million single nucleotide variants (SNV) with allele frequency information derived from low-coverage (0.06×–0.1×) whole-genome sequencing (WGS) data of 141 431 unrelated healthy Chinese individuals. These individuals were recruited from 31 out of the 34 administrative divisions in China, covering Han and 36 other ethnic minorities. CMDB, housing the WGS data of a multi-ethnic Chinese population featuring wide geographical distribution, has become the most representative and comprehensive Chinese population genome database to date. Researchers can quickly search for variant, gene or genomic regions to obtain the variant information, including mutation basic information, allele frequency, genic annotation and overview of frequencies in global populations. Furthermore, the CMDB also provides information on the association of the variants with a range of phenotypes, including height, BMI, maternal age and twin pregnancy. Based on these data, researchers can conduct meta-analysis of related phenotypes. CMDB is freely available at https://db.cngb.org/cmdb/.

INTRODUCTION

As the cost of sequencing has declined rapidly during the last decades, numerous large-scale population genomics research were conducted to support the precision medicine initiative (1–21). To date, several well-known large-scale international and national population genome variation databases have been published, such as the 1000 Genomes Project (1KGP) (15), Genomics England (16), Genome of the Netherlands (17), ExAC (18) and gnomAD (19). In addition, there are some genomic databases, such as NIH All of us (22) (https://allofus.nih.gov/), that are under construction. These databases have provided a considerable number of genome variants to assist in the fundamental understanding of human genetic architectures. However, the majority of these lack generalizability and are biased towards the baseline for European populations. While significant knowledge about human genetics has been achieved thanks to studies involving more than hundreds of thousands of participants for population with European ancestry (23,24), genetic information of the Chinese population were under-represented. The knowledge of genetic variants and their allele frequencies of the world's largest population were generally obtained from studies consisting of less than tens of thousands individuals (25).

China is the most populous country and the second-largest economy in the world. Therefore, a large-scale Chinese genome database based on whole-genome sequencing data of hundreds of thousands of Chinese individuals with various ethnic groups will not only be of great benefit to the practice of both evolutionary and translational medicine research of the Chinese population, but also addressing the existing sample distribution imbalance of global population in available databases. In the past decade, several Chinese genome projects have been launched. *PGG*.Han, released in 2019, archives approximately 12K WGS data and 102 586 high-density genome-wide genotyped data (26). ChinaMAP released the analytics of deep WGS about 10.5k individuals recruited from 8 ethnic populations in China (25). The NyuWa genome resource that includes deep WGS of 2999 Chinese individuals and constructed reference panel has been open access since 2021 (27). However, *PGG*.Han and NyuWa are predominantly biased towards the Han Chinese population while ChinaMAP is mostly biased

[*]To whom correspondence should be addressed. Email: jinxin@genomics.cn
Correspondence may also be addressed to Shujia Huang. Email: hshujia@qq.com
[†]The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

toward the Central and North Chinese population. Consequently, they may not be capable of representing the ethnic diversity of China.

Here, we have established the Chinese Millionome Database (CMDB, https://db.cngb.org/cmdb/), the largest and the most representative Chinese genome variation database to date. The CMDB database contains 9.04 million single nucleotide variants (SNVs) and the allele frequency information from low-coverage (0.06×–0.1×) WGS data of 141 431 unrelated healthy Chinese individuals (28). These individuals were recruited from 31 out of the 34 administrative divisions in China, covering the Han and 36 ethnic minorities. The main functions of the database are described below. (i) the database provides a query interface to search for allele frequency of variants. (ii) GWAS summary statistics for common phenotypes that were cataloged based on Genome-wide Association Study (GWAS).

DATA COLLECTION AND PROCESS

The CMDB was designed based on the whole genome analysis results from 141 431 unrelated individuals. The participants were all healthy pregnant women and recruited via the non-invasive fetal trisomy test at BGI between 2012 and 2013 (28,29), covering 31 out of the 34 administrative divisions with Han and 36 other ethnic minorities in China. All of the participants had signed informed consent and then been assigned an anonymous sample code before donating blood sample for excluding any personal identifiable information. After applying the low-coverage WGS (0.06× to 0.1×) for each participant, we filtered and aligned the reads using SOAPnuke (30) and bwa (31), removed duplicates with samtools rmdup module (32), recalibrated the base quality score using GATK(v3.8) (33). BaseVar (https://github.com/ShujiaHuang/basevar), a self-developed fast maximum likelihood estimation method, was afterwards run to identify polymorphic sites and infer allele frequencies (28). In addition, We annotated the variants with VEP (34), employed STITCH(v1.2.7) (35) to impute genotype, and Angsd (36) to call the significant loci associated with a phenotype. More details in sample summary and analysis have been described in the previous study (28).

DATABASE IMPLEMENTATION

Based on the allele frequency and the summary results of GWAS, we adopted the following IT strategies to build the database. We used flask as the CMDB backend framework (https://flask.palletsprojects.com/) and MongoDB as the database engine (https://www.mongodb.com/). Flask is a popular, free and open-source micro framework for building web application while MongoDB is a popular and open-source document-oriented NoSQL database that is designed to store a large scale of data. The web frontend interface was developed using the VUE framework (https://vuejs.org/), a progressive framework for building user interfaces. Moreover, JQuery (https://jquery.com/, a cross-platform and feature-rich JavaScript library) and Bootstrap (https://getbootstrap.com, an open-source toolkit for developing Web projects using HTML, CSS and JS) were used as plugins. In addition, the website used D3 (https://d3js.org/, a JavaScript library for manipulating documents based on data) and Echart (https://echarts.apache.org/, data visualization open-source toolkit) to visualize genomics data.

DATABASE CONTENT AND USAGE

Overview

CMDB contains 9.04 million SNVs and associated variants with four phenotypes. In these datasets, 81.7% of the variants have been called in the 1KGP Han Chinese individuals (15), and 2.6% (233 966) are completely novel not present in the dbSNP databse (28). Statistics information on the distribution of variants on all chromosomes is displayed on the overview page of the database, and the user can select any chromosome to view the statistics information for that chromosome (Supplementary Figure S1). Furthermore, CMDB yields various statistics summaries, including variant density, allele frequency spectrum, base change, ts/tv ratio, variant quality and variant type.

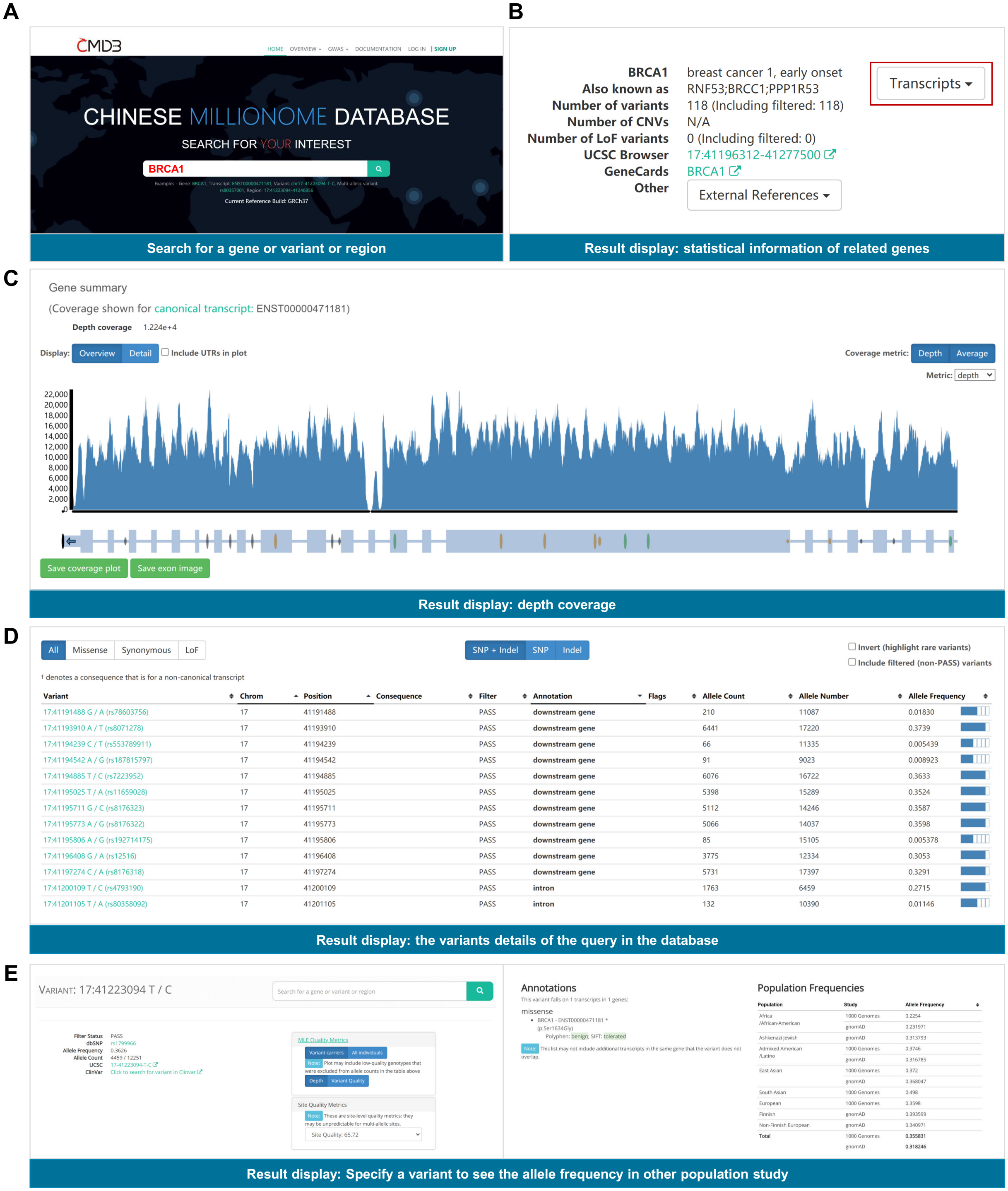
CMDB provides a web interface that enables quick search for variants, genes or genomic regions to query the variation information, including basic information of variants, allele frequency, genic annotation and frequency comparison to the global populations. Besides, the CMDB also provides information about the association of the variant with a range of phenotypes, including height, BMI, maternal age and twin pregnancy. Based on the GWAS data, researchers can perform large-scale meta-analyses of the phenotypes of interest. These functionalities were efficiently implemented by a self-developed tool calls CMDBtools (https://github.com/ShujiaHuang/cmdbtools).

Search for variants (Homepage)

CMDB allows users to learn about genomic variants in the large population study from five perspectives: Gene, Transcript, Variant, Multi-allelic variant and genomic Region that spanning not more than 100 kb. The relevant elements are available in any query result interface so that the user can further view the element information of interest. This means the corresponding transcripts and variants can be found when searching for genes, and vice versa. For example, researcher can search for *BRCA1,* one of the most common breast cancer genes, on the home page and get the gene summary including the coverage plot of the query transcripts or the genomic regions and corresponding variants data comprising of chromosome positions, annotation, allele frequency, etc. (Figure 1). Users can click on a transcript or variant to focus on the corresponding element at a more granular level.

GWAS interface

CMDB also provides interactive Manhattan plot and GWAS summary statistics based on Chinese population currently for four commonly measured traits: Height, BMI, Age and Twins-pregnancy and more traits in the future. When obtaining query phenotype results, user can select seven *P*-value thresholds to browse the distribution of significant variation associated with the phenotype. In addition, user can focus on any associated gene, variant and region by additional filtering fields (Figure 2).



CHINESE MILLIONOME DATABASE

SEARCH FOR YOUR INTEREST

BRCA1

Examples - Gene: BRCA1, Transcript: ENST00000471181, Variant: chr17-41223094-T-C, Multi-allelic variant: rs80357001, Region: 17:41223094-41246856

Current Reference Build: GRCh37

Search for a gene or variant or region

BRCA1	breast cancer 1, early onset
Also known as	RNF53;BRCC1;PPP1R53
Number of variants	118 (Including filtered: 118)
Number of CNVs	N/A
Number of LoF variants	0 (Including filtered: 0)
UCSC Browser	17:41196312-41277500
GeneCards	BRCA1
Other	External References

Result display: statistical information of related genes

Gene summary

(Coverage shown for canonical transcript: ENST00000471181)

Depth coverage 1.224e+4

□ Include UTRs in plot

Result display: depth coverage

□ Invert (highlight rare variants)

□ Include filtered (non-PASS) variants

† denotes a consequence that is for a non-canonical transcript

Variant	Chrom	Position	Consequence	Filter	Annotation	Flags	Allele Count	Allele Number	Allele Frequency
17:41191488 G / A (rs78603756)	17	41191488		PASS	downstream gene		210	11087	0.01830
17:41193910 A / T (rs8071278)	17	41193910		PASS	downstream gene		6441	17220	0.3739
17:41194239 C / T (rs553789911)	17	41194239		PASS	downstream gene		66	11335	0.005439
17:41194542 A / G (rs187815797)	17	41194542		PASS	downstream gene		91	9023	0.008923
17:41194885 T / C (rs7223952)	17	41194885		PASS	downstream gene		6076	16722	0.3633
17:41195025 T / A (rs11659028)	17	41195025		PASS	downstream gene		5398	15289	0.3524
17:41195711 G / C (rs8176323)	17	41195711		PASS	downstream gene		5112	14246	0.3587
17:41195773 A / G (rs8176322)	17	41195773		PASS	downstream gene		5066	14037	0.3598
17:41195806 A / G (rs192714175)	17	41195806		PASS	downstream gene		85	15105	0.005378
17:41196408 G / A (rs12516)	17	41196408		PASS	downstream gene		3775	12334	0.3053
17:41197274 C / A (rs8176318)	17	41197274		PASS	downstream gene		5731	17397	0.3291
17:41200109 T / C (rs4793190)	17	41200109		PASS	intron		1763	6459	0.2715
17:41201105 T / A (rs80358092)	17	41201105		PASS	intron		132	10390	0.01146

Result display: the variants details of the query in the database

VARIANT: 17:41223094 T / C

Search for a gene or variant or region

Filter Status	PASS
dbSNP	rs1799966
Allele Frequency	0.3626
Allele Count	4459 / 12251
UCSC	17-41223094-T-C
ClinVar	Click to search for variant in Clinvar

MLE Quality Metrics

Note: Plot may include low-quality genotypes that were excluded from allele counts in the table above

Site Quality Metrics

Note: These are site-level quality metrics: they may be unpredictable for multi-allelic sites.

Site Quality: 65.72

Annotations

This variant falls on 1 transcripts in 1 genes:

missense

- BRCA1 - ENST00000471181 * (p.Ser1634Gly)
  Polyphen: benign; SIFT: tolerated

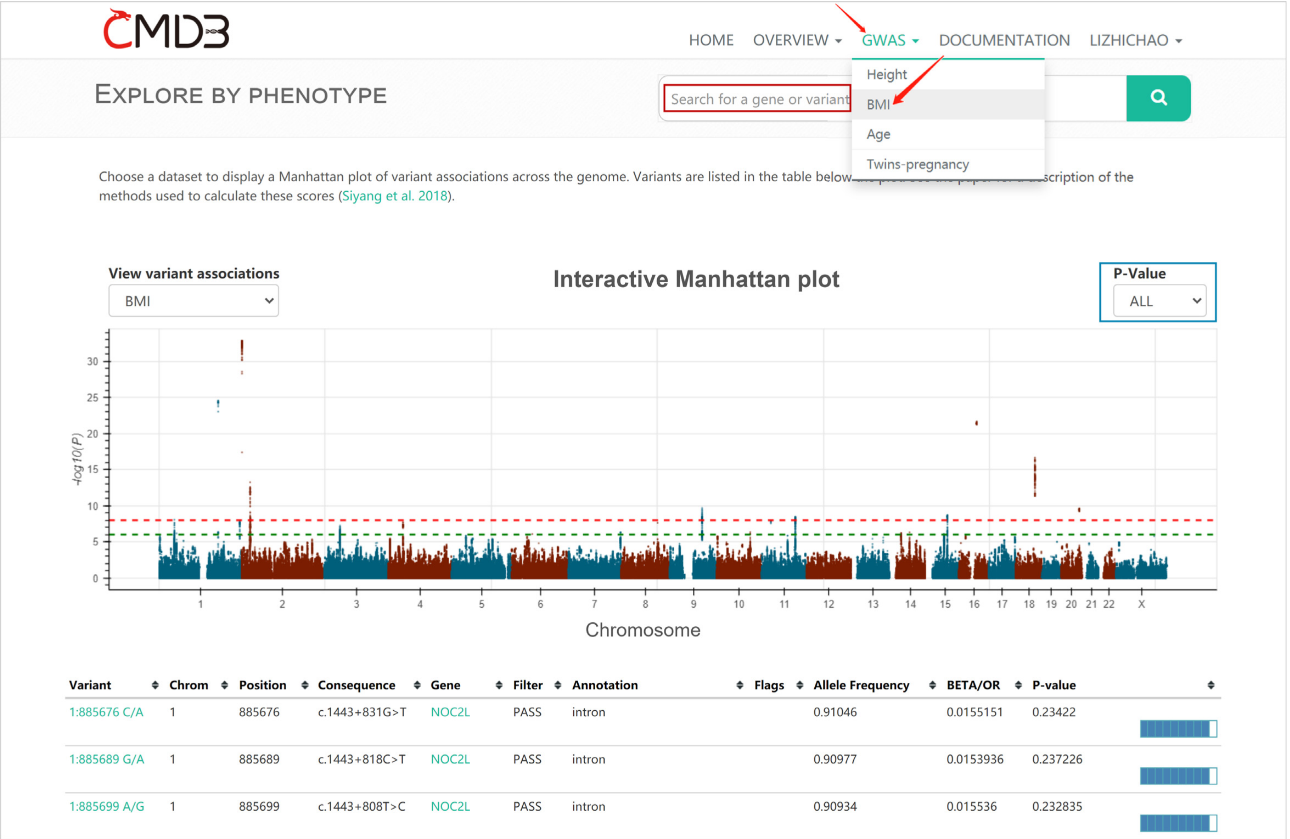
Note: This list may not include additional transcripts in the same gene that the variant does not overlap.

Population Frequencies

Population	Study	Allele Frequency
Africa /African-American	1000 Genomes	0.2254
	gnomAD	0.231971
Ashkenazi Jewish	gnomAD	0.313793
Admixed American /Latino	1000 Genomes	0.3746
	gnomAD	0.316785
East Asian	1000 Genomes	0.372
	gnomAD	0.368047
South Asian	1000 Genomes	0.498
European	1000 Genomes	0.3598
Finnish	gnomAD	0.393599
Non-Finnish European	gnomAD	0.340971
Total	1000 Genomes	**0.355831**
	gnomAD	**0.318246**

Result display: Specify a variant to see the allele frequency in other population study

Figure 1. Screenshots of *BRCA1* gene search and search result. (**A**) Search with *BRCA1* gene as an example, the corresponding transcript, variant, multi-allelic variant and region can all be used to get similar results page. The following is the query results. (**B**) The gene summary including the number of variants, UCSC browser link. (**C**) The coverage plot of a transcript. (**D**) The details of variants queried in the database. (**E**) Specify a variant to see the allele frequency in other population studies.



EXPLORE BY PHENOTYPE

Choose a dataset to display a Manhattan plot of variant associations across the genome. Variants are listed in the table below … scription of the methods used to calculate these scores (Siyang et al. 2018).

View variant associations: BMI

Interactive Manhattan plot

P-Value: ALL

Variant	Chrom	Position	Consequence	Gene	Filter	Annotation	Flags	Allele Frequency	BETA/OR	P-value
1:885676 C/A	1	885676	c.1443+831G>T	NOC2L	PASS	intron		0.91046	0.0155151	0.23422
1:885689 G/A	1	885689	c.1443+818C>T	NOC2L	PASS	intron		0.90977	0.0153936	0.237226
1:885699 A/G	1	885699	c.1443+808T>C	NOC2L	PASS	intron		0.90934	0.015536	0.232835

Figure 2. An example of GWAS query results. The associated variant with BMI under all *P*-value thresholds.

API DEVELOPMENT

The CMDB provides a Genomics API, which is a RESTful API (https://restfulapi.net/) that enables researcher to access CMDB variants data across remote sites via program scripts or command line. The API allows retrieval of CMDB variant information via variants' position or dbSNP Reference (rs) number. CMDBtools was developed based on the CMDB Genomics API. It is a Python toolkit that could backup queries for single variant and multiple variants, and annotate the local VCF (Variant Call Format) file with data including allele frequency, coverage information by population level, allele count by population level and filtration status in CMDB. CMDBtools can be easily install via the Python PyPI approach (https://pypi.org/search/?q=cmdbtools). The CMDBtools guidance is shown on the DOCUMENTATION page (http://cmdb.bgi.com/cmdb/apidoc-cmdbtools) and github repository simultaneously (https://github.com/ShujiaHuang/cmdbtools).

DISCUSSION AND FUTURE DIRECTIONS

Many studies have demonstrated that large-scale samples via low-coverage (<1×) sequencing performs good in calling common variants (37), calculating genome-wide polygenic scores (37), calling copy number variants (CNV) (38), even performs better than high coverage sequencing in many aspects when taking into account limited costs (28,37–42). Although sequencing at low depth, the development of the Chinese millionome database (CMDB) has filled the gap for a high-quality baseline genomic database of Chinese population. The CMDB database allows researchers to query variants, annotating their local VCF files and obtain several GWAS summary statistics results. Involving more than one hundred of thousands of participants with the broad geographic representation of multi-ethnic origin, CMDB is currently the largest and most representative Chinese population genome database. We believe that CMDB will greatly facilitate Chinese population genetic studies and medicine development.

In the future, we plan to improve the database in the following aspects. First, we will update the variants dataset including variant counts and allele frequencies by including additional samples and more comprehensive ethnics, and providing GWAS summary statistics with more phenotypes. Second, we will also add the population frequencies of SV (structure variation), including CNV, to the database since SV plays a very important role in genetic diseases or specific traits. Third, some general population genomics analysis tools including IMPUTE2 (43) and BEAGLE (44) used for imputation, plink (45) used for GWAS will be developed as online applications. Meanwhile, we will continuously improve the CMDBtools to speed up the response time towards a greater amount of user query.

DATA AVAILABILITY

The CMDBtools is available in the GitHub repository (https://github.com/ShujiaHuang/cmdbtools).

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We are especially grateful to Xiaofeng Wei, Kang Fang, Zhulidizi Yeerxin, Runming Lin, Ziyuan Ke from BGI for providing IT technical support to set up and upgrade CMDB database. This work was supported by China National GeneBank (CNGB).

FUNDING

National Natural Science Foundation of China [32000398]; National Natural Science Foundation of China [31900487]; Natural Science Foundation of Guangdong Province, China [2017A030306026]; Guangdong-Hong Kong Joint Laboratory on Immunological and Genetic Kidney Diseases [2019B121205005]. Funding for open access charge: Beijing Genomics Institute at Shenzhen.

Conflict of interest statement. None declared.

REFERENCES

1. Stark,Z., Dolman,L., Manolio,T.A., Ozenberger,B., Hill,S.L., Caulfied,M.J., Levy,Y., Glazer,D., Wilson,J., Lawler,M. *et al.* (2019) Integrating genomics into healthcare: a global responsibility. *Am. J. Hum. Genet.*, **104**, 13–20.
2. International HapMap Consortium (2003) The International HapMap Project. *Nature*, **426**, 789–796.
3. Goodwin,S., McPherson,J.D. and McCombie,W.R. (2016) Coming of age: ten years of next-generation sequencing technologies. *Nat. Rev. Genet.*, **17**, 333–351.
4. Metspalu,A., Koehler,F., Laschinski,G., Ganten,D. and Roots,I. (2004) The Estonian Genome Project in the context of European Genome Research. *Dtsch. Med. Wochenschr.*, **129**(Suppl. 1), S25–S28.
5. Lethimonnier,F. and Levy,Y. (2018) Genomic medicine france 2025. *Ann. Oncol.*, **29**, 783–784.
6. Tadaka,S., Katsuoka,F., Ueki,M., Kojima,K., Makino,S., Saito,S., Otsuki,A., Gocho,C., Sakurai-Yageta,M., Danjoh,I. *et al.* (2019) 3.5KJPNv2: an allele frequency panel of 3552 japanese individuals including the x chromosome. *Hum. Genome Var.*, **6**, 28.
7. Le,V.S., Tran,K.T., Bui,H.T.P., Le,H.T.T., Nguyen,C.D., Do,D.H., Ly,HaTT, Pham,L.T.D., Dao,L.T.M. and Nguyen,L.T. (2019) A vietnamese human genetic variation database. *Hum. Mutat.*, **40**, 1664–1675.
8. Kim,J., Weber,J.A., Jho,S., Jang,J., Jun,J., Cho,Y.S., Kim,H.-M., Kim,H., Kim,Y., Chung,O. *et al.* (2018) KoVariome: korean national standard reference variome database of whole genomes with comprehensive SNV, indel, CNV, and SV analyses OPEN. *Sci. Rep.*, **8**, 5677.
9. Gudbjartsson,D.F., Helgason,H., Gudjonsson,S.A., Zink,F., Oddson,A., Gylfason,A., Besenbacher,S., Magnusson,G., Halldorsson,B.V., Hjartarson,E. *et al.* (2015) Large-scale whole-genome sequencing of the icelandic population. *Nat. Genet.*, **47**, 435–444.
10. Walter,K., Min,J.L., Huang,J., Crooks,L., Memari,Y., McCarthy,S., Perry,J.R.B., Xu,C., Futema,M., Lawson,D. *et al.* (2015) The UK10K project identifies rare variants in health and disease. *Nature*, **526**, 82–89.
11. Turnbull,C., Scott,R.H., Thomas,E., Jones,L., Murugaesu,N., Pretty,F.B., Halai,D., Baple,E., Craig,C., Hamblin,A. *et al.* (2018) The 100 000 genomes project: bringing whole genome sequencing to the NHS. *BMJ*, **361**, k1687.
12. Hehir-Kwa,J.Y., Marschall,T., Kloosterman,W.P., Francioli,L.C., Baaijens,J.A., Dijkstra,L.J., Abdellaoui,A., Koval,V., Thung,D.T., Wardenaar,R. *et al.* (2016) A high-quality human reference panel reveals the complexity and distribution of genomic structural variants. *Nat. Commun.*, **7**, 12989.
13. Telenti,A., Pierce,L.C.T., Biggs,W.H., di Iulio,J., Wong,E.H.M., Fabani,M.M., Kirkness,E.F., Moustafa,A., Shah,N., Xie,C. *et al.* (2016) Deep sequencing of 10,000 human genomes. *Proc. Natl. Acad. Sci. U.S.A.*, **113**, 11901–11906.
14. 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**, 2–10.
15. Auton,A., Abecasis,G.R., Altshuler,D.M., Durbin,R.M., Bentley,D.R., Chakravarti,A., Clark,A.G., Donnelly,P., Eichler,E.E., Flicek,P. *et al.* (2015) A global reference for human genetic variation. *Nature*, **526**, 68–74.
16. Genomics England (2021) The national genomics research library v7.
17. Boomsma,D.I., Wijmenga,C., Slagboom,E.P., Swertz,M.A., Karssen,L.C., Abdellaoui,A., Ye,K., Guryev,V., Vermaat,M., van Dijk,F. *et al.* (2014) The genome of the netherlands: design, and project goals. *Eur. J. Hum. Genet.*, **22**, 221–227.
18. Karczewski,K.J., Weisburd,B., Thomas,B., Solomonson,M., Ruderfer,D.M., Kavanagh,D., Hamamsy,T., Lek,M., Samocha,K.E., Cummings,B.B. *et al.* (2017) The ExAC browser: displaying reference data information from over 60 000 exomes. *Nucleic Acids Res.*, **45**, D840–D845.
19. 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.
20. Mattingsdal,M., Ebenesersdóttir,S.S., Moore,K.H.S., Andreassen,O.A., Hansen,T.F., Werge,T., Kockum,I., Olsson,T., Alfredsson,L., Helgason,A. *et al.* (2021) The genetic structure of Norway. *Eur. J. Hum. Genet.*, **29**, 1710–1718.
21. Maretty,L., Jensen,J.M., Petersen,B., Sibbesen,J.A., Liu,S., Villesen,P., Skov,L., Belling,K., Theil Have,C., Izarzugaza,J.M.G. *et al.* (2017) Sequencing and de novo assembly of 150 genomes from denmark as a population reference. *Nature*, **548**, 87–91.
22. Denny,J.C., Rutter,J.L., Goldstein,D.B., Philippakis,A., Smoller,J.W., Jenkins,G. and Dishman,E. (2019) The 'All of us' research program. *N. Engl. J. Med.*, **381**, 668–676.
23. Taliun,D., Harris,D.N., Kessler,M.D., Carlson,J., Szpiech,Z.A., Torres,R., Taliun,S.A.G., Corvelo,A., Gogarten,S.M., Kang,H.M. *et al.* (2021) Sequencing of 53,831 diverse genomes from the NHLBI TOPMed program. *Nature*, **590**, 290–299.
24. Bycroft,C., Freeman,C., Petkova,D., Band,G., Elliott,L.T., Sharp,K., Motyer,A., Vukcevic,D., Delaneau,O., O'Connell,J. *et al.* (2018) The UK biobank resource with deep phenotyping and genomic data. *Nature*, **562**, 203–209.
25. Cao,Y., Li,L., Xu,M., Feng,Z., Sun,X., Lu,J., Xu,Y., Du,P., Wang,T., Hu,R. *et al.* (2020) The ChinaMAP analytics of deep whole genome sequences in 10,588 individuals. *Cell Res.*, **30**, 717–731.
26. Gao,Y., Zhang,C., Yuan,L., Ling,Y.C., Wang,X., Liu,C., Pan,Y., Zhang,X., Ma,X., Wang,Y. *et al.* (2020) PGG.Han: the han chinese genome database and analysis platform. *Nucleic Acids Res.*, **48**, D971–D976.
27. Zhang,P., Luo,H., Li,Y., Wang,Y., Wang,J., Zheng,Y., Niu,Y., Shi,Y., Zhou,H., Song,T. *et al.* (2021) NyuWa genome resource: a deep whole-genome sequencing-based variation profile and reference panel for the chinese population. *Cell Rep.*, **37**, 110017.
28. Liu,S., Huang,S., Chen,F., Zhao,L., Yuan,Y., Francis,S.S., Fang,L., Li,Z., Lin,L., Liu,R. *et al.* (2018) Genomic analyses from Non-invasive prenatal testing reveal genetic associations, patterns of viral infections, and chinese population history. *Cell*, **175**, 347–359.
29. Zhang,H., Gao,Y., Jiang,F., Fu,M., Yuan,Y., Guo,Y., Zhu,Z., Lin,M., Liu,Q., Tian,Z. *et al.* (2015) Non-invasive prenatal testing for trisomies 21, 18 and 13: clinical experience from 146 958 pregnancies. *Ultrasound Obstet. Gynecol.*, **45**, 530–538.
30. Chen,Y., Chen,Y., Shi,C., Huang,Z., Zhang,Y., Li,S., Li,Y., Ye,J., Yu,C., Li,Z. *et al.* (2018) SOAPnuke: a mapreduce acceleration-supported software for integrated quality control and preprocessing of high-throughput sequencing data. *Gigascience*, **7**, 1–6.
31. Li,H. and Durbin,R. (2009) Fast and accurate short read alignment with burrows-wheeler transform. *Bioinformatics*, **25**, 1754–1760.
32. 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.

33. DePristo,M.A., Banks,E., Poplin,R., Garimella,K.V, Maguire,J.R., Hartl,C., Philippakis,A.A., 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.
34. McLaren,W., Gil,L., Hunt,S.E., Riat,H.S., Ritchie,G.R.S., Thormann,A., Flicek,P. and Cunningham,F. (2016) The ensembl variant effect predictor. *Genome Biol.*, **17**, 122.
35. Davies,R.W., Flint,J., Myers,S. and Mott,R. (2016) Rapid genotype imputation from sequence without reference panels. *Nat. Genet.*, **48**, 965–969.
36. Korneliussen,T.S., Albrechtsen,A. and Nielsen,R. (2014) ANGSD: analysis of next generation sequencing data. *BMC Bioinf.*, **15**, 356.
37. Homburger,J.R., Neben,C.L., Mishne,G., Zhou,A.Y., Kathiresan,S. and Khera,A.V. (2019) Low coverage whole genome sequencing enables accurate assessment of common variants and calculation of genome-wide polygenic scores. *Genome Med*, **11**, 74.
38. Dong,Z., Xie,W., Chen,H., Xu,J., Wang,H., Li,Y., Wang,J., Chen,F., Choy,K.W. and Jiang,H. (2017) Copy-number variants detection by low-pass whole-genome sequencing. *Curr. Protoc. Hum. Genet.*, **2017**, 8.17.1–8.17.16.
39. Li,Y., Sidore,C., Kang,H.M., Boehnke,M. and Abecasis,G.R. (2011) Low-coverage sequencing: implications for design of complex trait association studies. *Genome Res.*, **21**, 940–951.
40. Pasaniuc,B., Rohland,N., McLaren,P.J., Garimella,K., Zaitlen,N., Li,H., Gupta,N., Neale,B.M., Daly,M.J., Sklar,P. *et al.* (2012) Extremely low-coverage sequencing and imputation increases power for genome-wide association studies. *Nat. Genet.*, **44**, 631–635.
41. Fumagalli,M. (2013) Assessing the effect of sequencing depth and sample size in population genetics inferences. *PLoS One*, **8**, 14–17.
42. Zhou,Bo, Ho,S.S., Zhang,X., Pattni,R., Haraksingh,R.R. and Urban,A.E. (2017) Whole-genome sequencing analysis of CNV using low-coverage and paired-end strategies is efficient and outperforms arraybased CNV analysis. *Physiol. Behav.*, **176**, 139–148.
43. 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.
44. Browning,B.L. and Browning,S.R. (2008) A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. *Am. J. Hum. Genet.*, **84**, 210–223.
45. Purcell,S., Neale,B., Todd-Brown,K., Thomas,L., Ferreira,M.A.R., Bender,D., Maller,J., Sklar,P., De Bakker,P.I.W., Daly,M.J. *et al.* (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.*, **81**, 559–575.