Journal of Genetics and Genomics 48 (2021) 1032–1035

Contents lists available at ScienceDirect

Journal of Genetics and Genomics

Journal homepage: www.journals.elsevier.com/journal-of-genetics-and-genomics/

Letter to the editor

The HuaBiao project: whole-exome sequencing of 5000 Han Chinese individuals

Next-generation sequencing technologies have significantly accelerated the identification of disease-causing mutations and facilitated the emergence of personalized medicine (Genomes Project Consortium et al., 2015; Goodwin et al., 2016; Sirugo et al., 2019). In comparison with whole-genome sequencing, whole-exome sequencing (WES), which covers the coding regions of the genome, offers a cost-efficacy balance. WES provides deeper sequencing depth (>100×) and allows the more accurate detection of rare variants that are tailored for clinical applications (Lek et al., 2016).

To identify disease-causing mutations, benign variants that exist in a significantly higher number in the genome are often filtered out, and the remaining variants are narrowed down to a few plausible pathogenic variants. The benign variants are genetically tolerated and have chances of reaching a certain frequency in populations due to genetic drift. The established variants with certain allele frequencies (e.g., >0.1%) in any population are usually considered as likely benign (Lek et al., 2016). On the contrary, disease-causing variants lead to reduced fitness in carriers and thus are kept rare by natural selection (Kiezun et al., 2012). Therefore, the most basic yet effective filter is based on the allele frequency in a large reference database (Povysil et al., 2019). Furthermore, genetic diversity in exonic regions in different populations has been a research interest in human genetics, including population structure and history, complex trait dissection, and pharmacogenetics applications (Kiezun et al., 2012; Genomes Project Consortium et al., 2015; Lek et al., 2016). Thus, the construction of a large WES database for specific populations is essential and valuable (Kiezun et al., 2012).

Several large public WES databases have been established for medical genetics, such as the Exome Aggregation Consortium (ExAC) and the Genome Aggregation Database (gnomAD) (Lek et al., 2016; Karczewski et al., 2020). However, the number of Han Chinese samples in these databases is limited (Genomes Project Consortium et al., 2015; Bycroft et al., 2018). The Han Chinese population is the largest ethnic group in the world (Liu et al., 2018), exhibiting genetic differentiation among the populations from the north to the south of China (Xu et al., 2009). Therefore, these databases are not tailored for medical genetics studies on Han Chinese populations. Accordingly, it is urgently needed for a high-quality WES database of the Han Chinese subpopulations.

Here, we present an exon database named "HuaBiao", which contains deep sequencing (>100×) of 5000 unrelated healthy samples. These samples were collected mainly from three representative Han Chinese populations at Zhengzhou in north China, Taizhou in east China, and Nanning in south China (Wen et al., 2004; Xu et al., 2009). High-quality reads (mean Phred score > 30) and high sequencing depth of samples (mean depth > 100×) were processed by the standardized bioinformatics pipeline, yielding high-confidence variants. In summary, the "HuaBiao" database contains allele frequencies of 2.07 million variations (1924K single-nucleotide polymorphisms [SNPs] and 145K insertions/deletions [INDELs]), 46.4% of which are novel compared with gnomAD (Collins et al., 2020). Researchers and clinical geneticists can quickly retrieve allele frequency information from the website (https://www.biosino.org/wepd).

We first assessed the sequencing quality of the raw reads. After quality control and preprocessing, the mean Phred score of each base in all samples was above 30 (Fig. S1A and S1B). We also checked the percentage of reads mapping to the reference genome for each sample and found high mapping rates (>95%) in the majority of samples (>98%; Fig. S1C). Moreover, the rate of good mappings (MAPQ >20) was calculated (Fig. S1D). The MAPQ20 rates of most samples were above 88%. Therefore, the summaries of the read quality, mapping rate, and MAPQ20 rates show that the high-quality reads in our database are reliable for downstream analyses.

Capture ratio and sequencing depth were calculated after read alignment. The capture ratio refers to the percentage of reads mapped to target regions over all the mapped reads. The median capture ratio was about 80%, with the lower and upper quartiles at 76% and 82%, respectively (Fig. S2A). The distribution of capture ratio was thus indicative of the excellent efficiency of the AIExome V1–CNV (iGeneTech, Beijing) exome capture kit. Meanwhile, the density peak of sequencing coverage (depth) was between 100× and 110 × (Fig. S2B). The median value of mean coverage was 107×, whereas the lower and upper quartiles were 90× and 122×, respectively. Similarly, the sequencing depth of each autosome was also evaluated. The mean coverages (depth) of different chromosomes were consistent, and almost all were above 100× (Fig. S2C). Furthermore, the coverage of each locus was shown, and the density peak of locus coverage was about 62× (Fig. S2D). The high coverage of samples, chromosomes, and loci enabled high-confidence variant calling.

With stringent quality control procedures and variant calling, we obtained a total of 2,069,303 variants, including 1,924,056 SNPs and 145,247 INDELs (Fig. S3A). The variants are mainly distributed in exonic, intronic, and noncoding RNA (ncRNA) regions. In partic-

https://doi.org/10.1016/j.jgg.2021.07.013

ular, there were 1,015,183 (49.1%) exonic variants, 409,892 (19.8%) intronic variants, and 319,000 (15.4%) ncRNA variants. Furthermore, we classified the variants in exonic regions based on their functional annotations (Fig. S3B). The exonic SNPs consisted of 611,764 non-synonymous, 342,893 synonymous, and 15,714 stop-gain/stop-loss variants. The exonic INDELs comprised 13,501 frameshift, 18,139 nonframeshift, and 732 stop-gain/stop-loss variants. The allele frequencies of 92.8% of the variants were below 1%. The allele frequencies of 85.4% of the variants were below 0.1% (Fig. S3C). In particular, singletons accounted for 51.5% of the total variants in our data set (Fig. S3D). The transition-transversion ratio (Ti/Tv) in our overall data sets and exonic region were 2.09 and 2.45, respectively.

To further validate the robustness and precision of our dataset, we used the Chinese Quartet reference samples (http://chinese-quartet.org/), whose genomes have been well characterized. Compared with the reference genotypes, the genotypes of these reference samples that were called in our study reached a high level of precision (96.8%) and recall (96.1%). Meanwhile, the concordance rate of the technical replicates in our dataset also reached 98.9%. Besides, the variants were further validated using Illumina Infinium Global Screening Array. The concordance rate reached 99.8% between genotypes of WES and genotyping array. These evaluations indicated the robustness of genotype calling (Chen et al., 2009; HUGO Pan-Asian SNP Consortium et al., 2009).

Approximately 46.4% of the variants in our database have not been reported in gnomAD, which indicates that our database provides a valuable resource for rare variant analysis on Han Chinese populations. Comparison of Minor allele frequency (MAF)s of shared variants between our database and the East Asian samples from gnomAD (Fig. S4) revealed high consistent in all variants

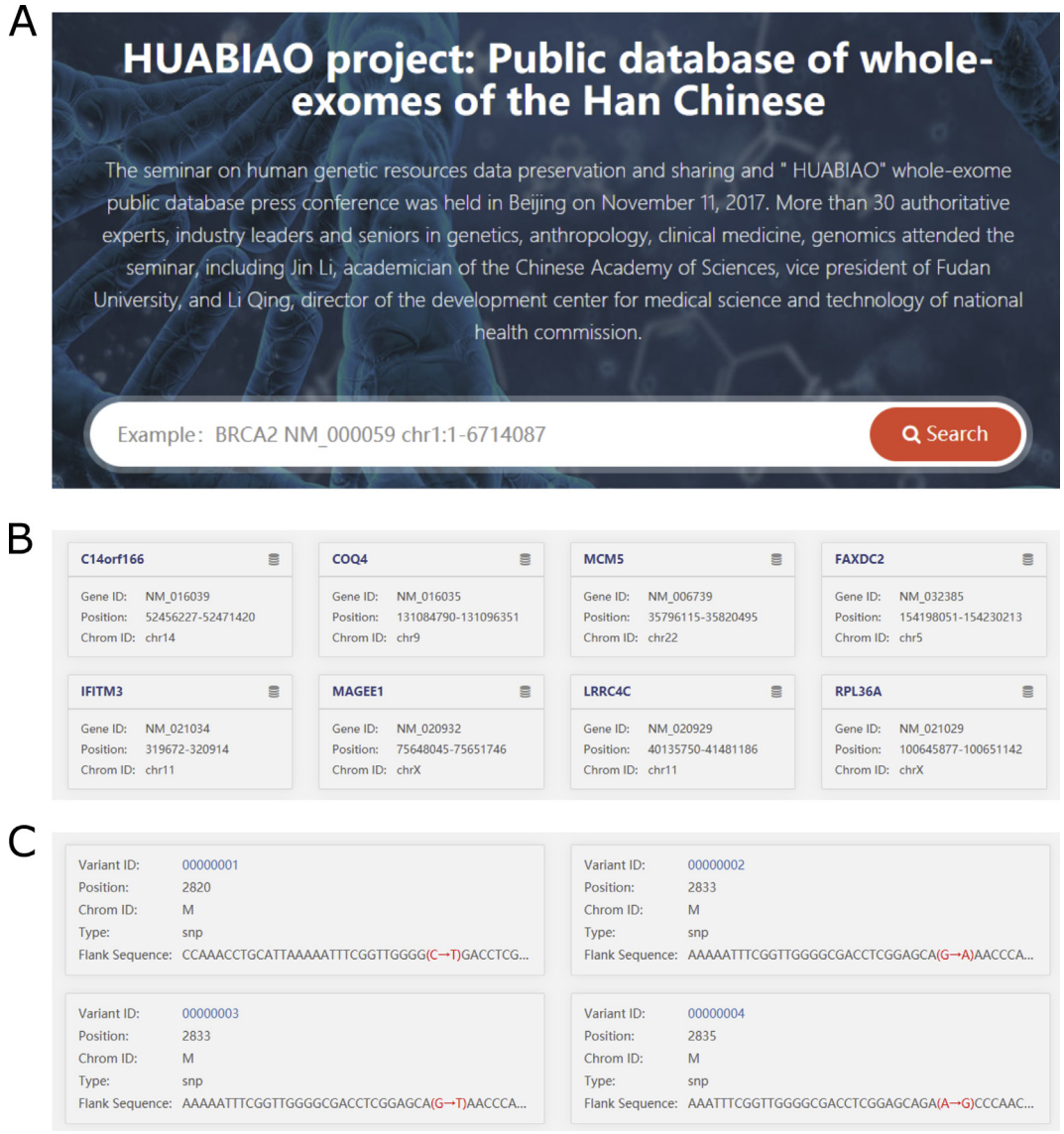

Fig. 1. A concise description of the HuaBiao database. **A**: The home page of the HuaBiao database. **B**: Examples of Gene View in the HuaBiao database. **C**: Examples of Variation View in the HuaBiao database.

($R^2 = 0.99$) and exonic variants ($R^2 = 0.99$). In addition, we found that some outliers of the MAF comparison resulted from inaccurate MAFs in gnomAD due to insufficient allele counts (e.g., $N < 10$).

Principal component analysis (PCA) of autosomal variants revealed distinct separations of the three major groups (Zhengzhou, Taizhou, and Nanning) in PC1, which was in accordance with previous studies on genetic stratification in the Han Chinese population (HUGO Pan-Asian SNP Consortium et al., 2009; Xu et al., 2009) (Fig. S4C).

The frequency information of genetic variants in "HuaBiao" database is available online (https://www.biosino.org/wepd). Researchers can retrieve allele frequency information in the database by variant queries. The "HuaBiao" database provides a general search function for three input types on the home page: gene symbol, gene ID, or variant position (hg19). There are tips for the three input types in the search box (Fig. 1A). Furthermore, the "HuaBiao" database includes Gene View and Variation View. By clicking on the gene symbol labeled blue, the page will transit to the gene detail page, which shows the summary information and variations located in this gene (Fig. 1B). The variant details page mainly contains general information, as well as allele and gene information (Fig. 1C). The sequencing data of "HuaBiao" project have been deposited in the National Omics Data Encyclopedia, which is an integrated, compatible, comparable, and scalable multi-omics resource platform (https://www.biosino.org/node/). The whole frequency information can be downloaded with registration and request. To ensure the security of genetic data, the sequence data are not publicly available. Collaborating researchers can contact the corresponding authors to access and analyze these sequencing data.

Previous studies have shown that the performance of genotype calling as a function of average sequencing depth reaches saturation at about 50× (Gao et al., 2020). Therefore, the high coverage of exome regions (~100×) of our database should be enough to yield high-confidence variants and be used in medical genetics applications.

The Han Chinese population is substructured, with three major observed clusters corresponding roughly to the northern Han, central Han, and southern Han (Wen et al., 2004; Chen et al., 2009; HUGO Pan-Asian SNP Consortium et al., 2009; Xu et al., 2009; Liu et al., 2018; Cao et al., 2020; Gao et al., 2020). Furthermore, the haplotype diversity of East Asia is strongly correlated with latitude, with diversity decreasing from the south to the north (Gao et al., 2020). To better characterize the genetic diversities of the Han Chinese population, we collected 5000 unrelated healthy samples, mainly from Zhengzhou in north China, Taizhou in east China, and Nanning in south China. Analysis of our data set further confirmed the three clusters of the Han Chinese population and characterized their genetic gradient in the exon region by allele frequency variations. The differentiation of allele frequency in the three populations was examined through Fst (Table S1). Several genes with significance across latitudes, such as *MTHFR* and *CR1*, were also reported in previous studies (Xu et al., 2009; Liu et al., 2018).

Due to the large diversity of Chinese populations, the sample in our database may not be representative of all Chinese subpopulations. In the future, more samples from different Chinese subpopulations, particularly those other than the Han Chinese populations, should be investigated.

Compared with public WES databases (e.g., ExAC and gnomAD), the "HuaBiao" database is the first that includes a large-scale Han Chinese population with high-confidence variants. It is worth noting that 46.4% of our variants are novel and may be specific to the Chinese population. These novel rare variants in our database could provide valuable resources for the diagnosis of Mendelian diseases in China.

Conflict of interest

The authors declare that they have no competing interests.

Acknowledgments

We thank the technical support of iGeneTech in this study. This work was supported by Shanghai Municipal Science and Technology Major Project (2017SHZDZX01), CAMS Innovation Fund for Medical Sciences (2019-I2M-5-066), and the National Basic Research Program of China (2015FY111700). This work was also supported by the Postdoctoral Science Foundation of China (2018M640333 and 2019M651354).

Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jgg.2021.07.013.

References

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.

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.

Chen, J., Zheng, H., Bei, J.X., Sun, L., Jia, W.H., Li, T., Zhang, F., Seielstad, M., Zeng, Y.X., Zhang, X., et al., 2009. Genetic structure of the Han Chinese population revealed by genome-wide SNP variation. Am. J. Hum. Genet. 85, 775−785.

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., 2020. A structural variation reference for medical and population genetics. Nature 581, 444−451.

Gao, Y., Zhang, C., Yuan, L., Ling, Y., 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.

Genomes Project Consortium, Auton, A., Brooks, L.D., Durbin, R.M., Garrison, E.P., Kang, H.M., Korbel, J.O., Marchini, J.L., McCarthy, S., McVean, G.A., et al., 2015. A global reference for human genetic variation. Nature 526, 68−74.

Goodwin, S., McPherson, J.D., McCombie, W.R., 2016. Coming of age: ten years of next-generation sequencing technologies. Nat. Rev. Genet. 17, 333−351.

HUGO Pan-Asian SNP Consortium, Abdulla, M.A., Ahmed, I., Assawamakin, A., Bhak, J., Brahmachari, S.K., Calacal, G.C., Chaurasia, A., Chen, C.H., Chen, J., et al., 2009. Mapping human genetic diversity in Asia. Science 326, 1541−1545.

Karczewski, K.J., Francioli, L.C., Tiao, G., Cummings, B.B., Alfoldi, 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.

Kiezun, A., Garimella, K., Do, R., Stitziel, N.O., Neale, B.M., McLaren, P.J., Gupta, N., Sklar, P., Sullivan, P.F., Moran, J.L., et al., 2012. Exome sequencing and the genetic basis of complex traits. Nat. Genet. 44, 623−630.

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.

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. e314.
Povysil, G., Petrovski, S., Hostyk, J., Aggarwal, V., Allen, A.S., Goldstein, D.B., 2019. Rare-variant collapsing analyses for complex traits: guidelines and applications. Nat. Rev. Genet. 20, 747–759.
Sirugo, G., Williams, S.M., Tishkoff, S.A., 2019. The missing diversity in human genetic studies. Cell 177, 26–31.
Wen, B., Li, H., Lu, D., Song, X., Zhang, F., He, Y., Li, F., Gao, Y., Mao, X., Zhang, L., et al., 2004. Genetic evidence supports demic diffusion of Han culture. Nature 431, 302–305.
Xu, S., Yin, X., Li, S., Jin, W., Lou, H., Yang, L., Gong, X., Wang, H., Shen, Y., Pan, X., et al., 2009. Genomic dissection of population substructure of Han Chinese and its implication in association studies. Am. J. Hum. Genet. 85, 762–774.

Meng Hao[1], Weilin Pu[1]
State Key Laboratory of Genetic Engineering, Collaborative Innovation Center for Genetics and Development, Human Phenome Institute, Fudan University, Shanghai 200433, China

Yi Li[1]
State Key Laboratory of Genetic Engineering, Collaborative Innovation Center for Genetics and Development, Human Phenome Institute, Institute for Six-sector Economy, Fudan University, Shanghai 200433, China

Shaoqing Wen[1]
Ministry of Education Key Laboratory of Contemporary Anthropology, Department of Anthropology and Human Genetics, School of Life Sciences, Institute of Archaeological Science, Fudan University, Shanghai 200433, China

Chang Sun
Ministry of Education Key Laboratory of Contemporary Anthropology, Department of Anthropology and Human Genetics, School of Life Sciences, Fudan University, Shanghai 200433, China

Yanyun Ma
Ministry of Education Key Laboratory of Contemporary Anthropology, Department of Anthropology and Human Genetics, School of Life Sciences, Institute for Six-sector Economy, Fudan University, Shanghai 200433, China

Hongxiang Zheng
Ministry of Education Key Laboratory of Contemporary Anthropology, Department of Anthropology and Human Genetics, School of Life Sciences, Fudan University, Shanghai 200433, China

Xingdong Chen
Ministry of Education Key Laboratory of Contemporary Anthropology, Department of Anthropology and Human Genetics, School of Life Sciences, Fudan University, Shanghai 200433, China

Taizhou Institute of Health and Sciences, Fudan University, Taizhou, Jiangsu 225300, China

Jingze Tan
Ministry of Education Key Laboratory of Contemporary Anthropology, Department of Anthropology and Human Genetics, School of Life Sciences, Fudan University, Shanghai 200433, China

Guoqing Zhang
Key Laboratory of Computational Biology, Bio-Med Big Data Center, Shanghai Institute of Nutrition and Health, University of Chinese Academy of Sciences, Chinese Academy of Sciences, Shanghai 200031, China

Menghan Zhang
Ministry of Education Key Laboratory of Contemporary Anthropology, Department of Anthropology and Human Genetics, School of Life Sciences, Fudan University, Shanghai 200433, China

Shuhua Xu
Ministry of Education Key Laboratory of Contemporary Anthropology, Department of Anthropology and Human Genetics, School of Life Sciences, Human Phenome Institute, Fudan University, Shanghai 200433, China

Yi Wang[*], Hui Li[**]
Ministry of Education Key Laboratory of Contemporary Anthropology, Department of Anthropology and Human Genetics, School of Life Sciences, Fudan University, Shanghai 200433, China

Jiucun Wang[***], Li Jin[****]
State Key Laboratory of Genetic Engineering, Collaborative Innovation Center for Genetics and Development, Human Phenome Institute, Fudan University, Shanghai 200433, China

Taizhou Institute of Health and Sciences, Fudan University, Taizhou, Jiangsu 225300, China

[*] Corresponding author.

[**] Corresponding author.

[***] Corresponding author.

[****] Corresponding author.
E-mail addresses: godspeed.wang@gmail.com (Y. Wang), lhca@fudan.edu.cn (H. Li), jcwang@fudan.edu.cn (J. Wang), lijin@fudan.edu.cn (L. Jin).

22 April 2021
Available online 18 August 2021

[1] These authors contributed equally to this article.