Genome Variation Map: a data repository of genome variations in BIG Data Center

Shuhui Song1,†, Dongmei Tian1,†, Cuiping Li1,†, Bixia Tang1,2,3,†, Lili Dong1,†, Jingfa Xiao1,2,3,4, Yiming Bao1,2, Wenming Zhao1, Hang He5 and Zhang Zhang1,2,3,4,*

1BIG Data Center, Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing 100101, China, 2CAS Key Laboratory of Genome Sciences and Information, Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing 100101, China, 3University of Chinese Academy of Sciences, Beijing 100049, China, 4Collaborative Innovation Center of Genetics and Development, Fudan University, Shanghai 200438, China and 5School of Life Sciences, Peking University, Beijing 100871, China

Received August 14, 2017; Revised October 09, 2017; Editorial Decision October 10, 2017; Accepted October 11, 2017

ABSTRACT

The Genome Variation Map (GVM; http://bigd.big.ac.cn/gvm/) is a public data repository of genome variations. As a core resource in the BIG Data Center, Beijing Institute of Genomics, Chinese Academy of Sciences, GVM dedicates to collect, integrate and visualize genome variations for a wide range of species, accepts submissions of different types of genome variations from all over the world and provides free open access to all publicly available data in support of worldwide research activities. Unlike existing related databases, GVM features integration of a large number of genome variations for a broad diversity of species including human, cultivated plants and domesticated animals. Specifically, the current implementation of GVM not only houses a total of ~4.9 billion variants for 19 species including chicken, dog, goat, human, poplar, rice and tomato, but also incorporates 8669 individual genotypes and 13 262 manually curated high-quality genotype-to-phenotype associations for non-human species. In addition, GVM provides friendly intuitive web interfaces for data submission, browse, search and visualization. Collectively, GVM serves as an important resource for archiving genomic variation data, helpful for better understanding population genetic diversity and deciphering complex mechanisms associated with different phenotypes.

INTRODUCTION

With the rapid development of high-throughput sequencing technologies, biological sequence data have been generated exponentially over the past decade. The availability of high-quality reference genome sequences and the improvement of genome variation data analysis methodology enable large-scale identification of genome variations at unprecedented rates, making it possible to systematically conduct population evolution studies and decipher genotype-to-phenotype associations (1–3). Therefore, it is fundamentally vital to build a public data repository for managing different genome variations from a wide variety of species in aid of big data mining and integrative in-depth analyses.

Toward this end, valuable efforts have been made in the National Center for Biotechnology Information (NCBI) (4) and the European Bioinformatics Institute (EBI) (5). Specifically, dbSNP (6) and dbVar (7) are two major resources in NCBI for archiving worldwide genome variations, and the counterpart in EBI, European Variation Archive (EVA) (8), imports variation data primarily from these two resources. Unfortunately, it was just recently announced that dbSNP and dbVar will phase out support for non-human data and stop accepting non-human data submissions from 1 September 2017 (https://ncbiinsights.ncbi.nlm.nih.gov/2017/05/09/phasing-out-support-for-non-human-genome-organism-data-in-dbsnp-and-dbvar/), consequently presenting formidable challenges in deposition and integration of publicly available variation data at a global scale, especially for non-human variant data. Besides, existing related databases do not well manage phenotype information, particularly for non-human species; although they can be obtained from controlled-access repositories for human (such as dbGaP (9)), genotype-to-phenotype associations for non-human species are considerably absent in existing related databases.

Here we present GVM (Genome Variation Map; http://bigd.big.ac.cn/gvm/), a public data repository of genome variations, including single nucleotide polymorphisms...
(SNP) and small insertions and deletions (INDEL), with particular focuses on human as well as cultivated plants and domesticated animals. As a core resource of the BIG Data Center (10), part of Beijing Institute of Genomics, Chinese Academy of Sciences, GVM dedicates to collect, integrate and visualize genome variations for a wide variety of species, accepts submissions of different types of genome variations from all over the world and provides free open access to all publicly available data in support of worldwide research activities. Based on a large collection of raw sequence data from public repositories, GVM integrates a large number of genome variants for 19 species and provides friendly web interfaces for data submission, search, browse and visualization.

IMPLEMENTATION

GVM is built based on J2EE framework with MySQL (http://www.mysql.org; a free and popular relational database management system) as its database engine. Web user interfaces are developed by using JSP (JavaServer Pages; a technology facilitating rapid development of dynamic web pages based on the Java programming language) and AJAX (Asynchronous JavaScript and XML; a set of web development techniques to create asynchronous applications without interfering with the display and behaviour of the existing page). GBrowser (11) (http://gbrowser.sourceforge.net) is adopted for chromosome-based data visualization. All raw sequence data are obtained from Genome Sequence Archive (GSA) (12) in BIG Data Center (10) and Sequence Read Archive (SRA) (13) in NCBI as well as from our collaborators and partners. We filter low quality reads and bases and identify SNPs and INDELS using standard GATK pipeline (14). Variants Effect Predictor (VEP) (15) is used to predict all variants’ effects. All analyzed results are publicly available at the download page of GVM (http://bigd.big.ac.cn/gvm/download).

DATABASE CONTENT AND USAGE

The current version of GVM houses a total of ~4.9 billion variants for 19 species covering 8,884 individuals. The detailed statistics of variants, individuals, genotype-to-phenotype pairs and associated projects are displayed and maintained online at the home page of GVM and summarized in Table 1. GVM features comprehensive incorporation of SNPs and INDELs for not only human but also cultivated plants (e.g. maize, rice, tomato, sorghum and soybean) and domesticated animals (e.g. chicken, dog, goat and pig) and covers new featured species, viz. giant panda, killer whale, moso bamboo, rubber and wheat, that are absent in existing related databases (Table 1).

Based on a large collection of individual genotypes (e.g. 5152 for rice, 926 for poplar, 579 for tomato, 544 for soybean), GVM incorporates a large quantity of genome variants and accordingly includes more newly identified variants, which are of great significance for fully capturing genetic diversity and systematically deciphering population evolutionary history. Additionally, GVM accommodates high-quality genotype-to-phenotype associations (e.g. 7432 for rice, 3332 for maize, 449 for soybean, 326 for pig, 203 for dog) that are manually curated from a number of publications on genome-wide association studies. Particularly, GVM focuses on collection of genomic variants for Chinese population; it collects 16,348,637 Chinese genomic variants that are derived from the 1000 Genome Project (16,17) and integrates 180,911 genotype-to-phenotype pairs from ClinVar (18), GWAS-catalog (19) and OMIM (20), providing valuable resources for in-depth investigations on molecular mechanisms associated with different phenotypes.

To support information search and exploration, GVM provides friendly web interfaces to retrieve variant relevant information (Figure 1; http://bigd.big.ac.cn/gvm/search). Simply by specifying a variant identifier, users can obtain its related details including variant position, alleles, minor allele frequency, variant effect and hyperlinks to external databases (e.g. dbSNP). In addition, GVM allows users to obtain multiple variants by searching variant consequence type, minor allele frequency, genomic position or gene name and function, which are retrieved in a tabular format and can be displayed in GBrowser. Moreover, searched items can be further refined by multiple filters, greatly facilitating users to narrow down the items of interest in an efficient and intuitive manner. For any given variant, GVM provides detailed information such as variant basic information, population diversity, genotype-to-phenotype annotation and gene annotation.

To dynamically visualize variant genotype and allele frequency, GVM deploys an interactive and user-friendly genome browser (Figure 2) built based on GBrowser (11), enabling users to zoom and scroll to any region along the genome and to investigate variants of interest in a visualized manner. It includes a number of individual tracks for selection and provides an interactive visualization of variant genotype and allele frequency for user selected individuals. Moreover, whole-genome variants for all collected species are publicly available in VCF and FASTA formats at the download page of GVM (http://bigd.big.ac.cn/gvm/download).

In the era of big data, keeping a database comprehensive and up-to-date is increasingly challenging, accordingly demanding a large number of researchers getting involved in data submission. ‘Nothing great is ever accomplished in isolation—Yo-Yo Ma’. Similar to dbSNP, GVM accepts community submissions of different types of genome variations from all over the world (http://bigd.big.ac.cn/gvm/submission). For each submission, it is required to provide compulsory information including not only variants in VCF or HapMap format but also metadata regarding submitter details, project and sample information, and variants analysis method. Each submission will be reviewed and assigned an accession number prefixed with ‘GVM’, which is convenient for data citation in any publication and data exchange with related databases.

DISCUSSION AND FUTURE DIRECTIONS

GVM is a public data repository of genome variations. Different from existing related databases, GVM features comprehensive integration of different types of genome variations for a wide range of species. It accepts human and non-human variant data submissions from all over the world,
Figure 1. Screenshots of data search and representation. (A) Search items involving genome assembly, technology, position and type, consequence type, gene, minor allele frequency and genotype-to-phenotype. (B) Search results containing genome variations, links to external databases, individual genotype and genome browser. (C) Variant details including basic information, gene annotation, population diversity and genotype-to-phenotype annotation.
Figure 2. Screenshots of genome variation visualization. (A) Genome browser, taking *Sorghum bicolor* chromosome 1 as an example. (B) Individual tracks for selection. (C) Variant allele frequency information.
integrates high-quality genotype-to-phenotype associations curated from a number of scientific publications, and provides free open access to all publicly available data in aid of worldwide research activities. Moreover, it equips with friendly web interfaces for data submission, browse, search and visualization. Taken together, GVM bears significance in archiving human and non-human genome variations at a global scale, helpful for fully capturing population genetic diversity and better understanding complex mechanisms associated with different phenotypes.

Along with the ongoing projects for precision medicine and population genomics in China, future directions of GVM accordingly include integration of more variation data from human and a broader range of species and development of more interactive and intuitive web interfaces for big data submission, search and visualization. In addition, we will continue to make enhancements in genomic variant annotation and develop a web-based system that allows multiple curators to annotate, verify and publish genotype-to-phenotype associations. We will also develop standards for variation data representation, analysis and exchange and deploy a cloud-based variation data analysis pipeline that is used to link GVM with GSA (http://bigd.big.ac.cn/gsa; a database for archiving raw sequence reads), with the ultimate goal to achieve automatic variation data analysis after data submissions to GSA and then enable automatic integration of analysed results into GVM. Meanwhile, we will link GVM with other omics databases in BIG Data Center, such as GEN (Gene Expression Nebulas, a data portal of gene expression profiles), MethBank (a methylation databank) (21), and LncRNAWiki (a wiki-based knowledgebase of long non-coding RNAs) (22). We also call for collaborators to work together to build GVM into an integral repository covering more comprehensive genome variations across a broader range of species.

**REFERENCES**

1. Chen, W., Gao, Y., Xie, W., Gong, L., Lu, K., Wang, W., Li, Y., Liu, X., Zhang, H., Dong, H. *et al.* (2014) Genome-wide association analyses provide genetic and biochemical insights into natural variation in rice metabolism. *Nat. Genet.*, 46, 714–721.

**ACKNOWLEDGEMENTS**

We thank Dr Jun Yu for valuable discussions on this work and members of the BIG Data Center for reporting bugs and sending comments.

**FUNDING**

Strategic Priority Research Program of the Chinese Academy of Sciences [XDB13040500 to Z.Z., W.Z.; XDA08020102 to Z.Z.]; National Key Research & Development Program of China [2016YFE0206600 to Y.B.; 2017YFC0907502 to Z.Z.]; National Key Research Program of China [2016YFC0901603 to W.Z.; 2016YFB0201702, 2016YFC0901903 to J.X.]; National Programs for High Technology Research and Development [863 Program; 2015AA020108 to Z.Z.]; The Youth Innovation Promotion Association of Chinese Academy of Science [2017141 to S.S.]; National Natural Science Foundation of China [30900831 to S.S.]; International Partnership Program of the Chinese Academy of Sciences [153F11KYSB20160008]; Key Program of the Chinese Academy of Sciences [KJZD-EW-L14 to J.X.]; Key Technology Talent Program of the Chinese Academy of Sciences (to W.Z.); The 100 Talent Program of the Chinese Academy of Sciences (to Y.B. and Z.Z.). Funding for open access charge: Strategic Priority Research Program of the Chinese Academy of Sciences [KJZD-EW-L14 to J.X.]; National Key Research & Development Program of China [2016YFE0206600 to Y.B.; 2017YFC0907502 to Z.Z.]; National Key Research Program of China [2016YFC0901603 to W.Z.; 2016YFB0201702, 2016YFC0901903 to J.X.]; National Programs for High Technology Research and Development [863 Program; 2015AA020108 to Z.Z.]; The Youth Innovation Promotion Association of Chinese Academy of Science [2017141 to S.S.]; National Natural Science Foundation of China [30900831 to S.S.]; International Partnership Program of the Chinese Academy of Sciences [153F11KYSB20160008]; Key Program of the Chinese Academy of Sciences [KJZD-EW-L14 to J.X.]; Key Technology Talent Program of the Chinese Academy of Sciences (to W.Z.); The 100 Talent Program of the Chinese Academy of Sciences (to Y.B. and Z.Z.). Funding for open access charge: Strategic Priority Research Program of the Chinese Academy of Sciences.

**Conflict of interest statement.** None declared.

**Table 1.** GVM data content and statistics as of 1 August 2017

| Species                  | Number of variants | Project count | Individual count | G2P associations |
|--------------------------|--------------------|---------------|------------------|------------------|
| **Animals**              | SNP                | INDEL         |                  |                  |
| Human (Homo sapiens)     | 13 327 822         | 3 019 815     | 4                | 215              | 180 911          |
| Cattle (Bos taurus)      | 53 609 957         | 6 724 343     | 9                | 95               |                  |
| Chicken (Gallus gallus)  | 36 174 851         | 4 619 064     | 9                | 112              | 1 249            |
| Dog (Canis familiaris)   | 18 457 814         | —             | 3                | 78               | 203              |
| Duck (Anas platyrhynchos)| 8 213 041          | 1 484 245     | 3                | 3                |                  |
| **Giant panda** (Ailuropoda melanoleuca) | 11 820 056      | 2 544 981     | 1                | 34               |                  |
| **Groat** (Capa hircus)  | 48 505 769         | 6 389 438     | 8                | 233              |                  |
| **Killer whale** (Orcinus Orca) | 4 821 960       | 573 454       | 1                | 48               |                  |
| Pig (Sus scrofa)         | 64 709 967         | 12 087 428    | 8                | 247              | 326              |
| Sheep (Ovis aries)       | 60 025 622         | 10 395 242    | 10               | 125              | 271              |
| **Plants**               |                    |               |                  |                  |
| Maize (Zea mays)         | 1 501 581          | —             | 1                | 376              | 3 332            |
| Moso bamboo (Phyllostachys heterocycle) | 2 009 487   | —             | 1                | 1                |                  |
| Poplar (Populus trichocarpa) | 19 861 824     | 7 727 010     | 4                | 926              |                  |
| Rice (Oryza sativa)      | 18 161 579         | —             | 5                | 5 152            | 7 432            |
| **Rubber** (Hevea brasiliensis) | 9 584 819       | —             | 1                | 6                |                  |
| Sorghum (Sorghum bicolor) | 15 513 117       | —             | 3                | 48               |                  |
| Soybean (Glycine max)    | 19 921 434         | 3 050 299     | 8                | 544              | 449              |
| Tomato (Solanum lycopersicum) | 26 938 825     | 3 724 480     | 6                | 579              |                  |
| Wheat (Triticum aestivum) | 1 365 924         | 114 594       | 2                | 62               |                  |

Note: Species in bold are featured species in GVM, whereas they are absent in dbSNP and EVA; ‘––’ indicates that it is under construction; ‘G2P Associations’ means genotype-to-phenotype associations.
2. Huang, X., Zhao, Y., Wei, X., Li, C., Wang, A., Zhao, Q., Li, W., Guo, Y., Deng, L., Zhu, C. et al. (2011) Genome-wide association study of flowering time and grain yield traits in a worldwide collection of rice germplasm. *Nat. Genet.*, 43, 32–39.

3. Plassais, J., Lagoutte, L., Correard, S., Paradis, M., Guaguere, E., Hedan, B., Pommier, A., Bothrel, N., Cadiergues, M., Pilorge, P. et al. (2016) A point mutation in a lincRNA upstream of GDNF is associated to a canine insensitivity to pain: a spontaneous model for human sensory neuropathies. *PLoS Genet.*, 12, e1006482.

4. NCBI Resource Coordinators (2017) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, 45, D12–D17.

5. Cook, C.E., Bergman, M.T., Finn, R.D., Cochrane, G., Birney, E. and Apweiler, R. (2016) The European Bioinformatics Institute in 2016: Data growth and integration. *Nucleic Acids Res.*, 44, D20–D26.

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

7. Lappalainen, I., Lopez, J., Hefferon, T., Spalding, J.D., Garner, J., Chen, C., Maguire, M., Corbett, M., Zhou, G. et al. (2012) DVar and DGaV: public archives for genomic structural variation. *Nucleic Acids Res.*, 41, D936–D941.

8. Chen, Y., Cunningham, F., Rios, D., McLaren, W.M., Smith, J., Pritchard, B., Spudich, G.M., Brent, S., Kulesha, E., Marin-Garcia, P. et al. (2010) Ensembl variation resources. * BMC Genomics*, 11, 293.

9. Mailman, M.D., Feolo, M., Jin, Y., Kimura, M., Tryka, K., Bagoutdinov, R., Hao, L., Liang, A., Paschall, J., Phan, L. et al. (2007) The NCBI dbGaP database of genotypes and phenotypes. *Nat. Genet.*, 39, 1181–1186.

10. Members, B.D.C. (2016) The Big Data Center: from deposition to integration to translation. *Nucleic Acids Res.*, 45, D18–D24.

11. Donlin, M.J. (2009) Using the Generic Genome Browser (GBrowse). *Curr. Protoc. Bioinformatics*, doi:10.1002/0471250953.bi0909s17.

12. Wang, Y., Song, F., Zhu, J., Zhang, S., Yang, Y., Chen, T., Tang, B., Dong, L., Ding, N., Zhang, Q. et al. (2017) GSA: Genome Sequence Archive. *Genomics Proteomics Bioinformatics*, 15, 14–18.

13. Kodama, Y., Shumway, M. and Leinonen, R. (2012) The Sequence Read Archive: explosive growth of sequencing data. *Nucleic Acids Res.*, 40, D54–D56.

14. 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.

15. McLaren, W., Gil, L., Hunt, S.E., Riat, H.S., Ritchie, G.R., Thorffmann, A., Flicek, P. and Cunningham, F. (2016) The Ensembl Variant Effect Predictor. *Genome Biol.*, 17, 122.

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

17. Abecasis, G.R., Auton, A., Brooks, L.D., DePristo, M.A., Durbin, R.M., Handsaker, R.E., Kang, H.M., Marth, G.T. and McVean, G.A. (2012) An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491, 56–65.

18. Landrum, M.J., Lee, J.M., Benson, M., Brown, G., Chao, C., Chitipiralla, S., Gu, B., Hart, J., Hoffman, D., Hoover, J. et al. (2016) ClinVar: public archive of interpretations of clinically relevant variants. *Nucleic Acids Res.*, 44, D862–D868.

19. MacArthur, J., Bowler, E., Cerozo, M., Gil, L., Hall, P., Hastings, E., Junkins, H., McMahon, A., Milo, A., Morales, J. et al. (2016) The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Res.*, 45, D896–D901.

20. Amberger, J.S., Bocchini, C.A., Schiettecatte, F., Scott, A.F. and Hamosh, A. (2014) OMIM.org: Online Mendelian Inheritance in Man (OMIM(R)), an online catalog of human genes and genetic disorders. *Nucleic Acids Res.*, 43, D789–D798.

21. Zou, D., Sun, S., Li, R., Liu, J., Zhang, J. and Zhang, Z. (2014) MethBank: a database integrating next-generation sequencing single-base-resolution DNA methylation programming data. *Nucleic Acids Res.*, 43, D54–D58.

22. Ma, L., Li, A., Zou, D., Xu, X., Xia, L., Yu, J., Bajic, V.B. and Zhang, Z. (2014) LncRNAWiki: harnessing community knowledge in collaborative curation of human long non-coding RNAs. *Nucleic Acids Res.*, 43, D187–D192.