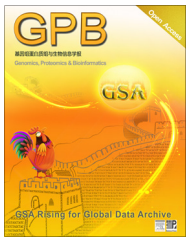
Genomics Proteomics Bioinformatics 15 (2017) 14–18

HOSTED BY

ELSEVIER

Genomics Proteomics Bioinformatics

www.elsevier.com/locate/gpb
www.sciencedirect.com



DATABASE

GSA: Genome Sequence Archive*

Yanqing Wang [1,#,a], Fuhai Song [2,3,#,b], Junwei Zhu [1,#,c], Sisi Zhang [1,#,d], Yadong Yang [2,3,#,e], Tingting Chen [1,f], Bixia Tang [1,3,g], Lili Dong [1,h], Nan Ding [2,i], Qian Zhang [2,j], Zhouxian Bai [2,3,k], Xunong Dong [2,3,l], Huanxin Chen [1,m], Mingyuan Sun [1,n], Shuang Zhai [1,o], Yubin Sun [1,p], Lei Yu [1,q], Li Lan [1,r], Jingfa Xiao [1,2,3,4,s], Xiangdong Fang [2,3,4,*,t], Hongxing Lei [2,3,5,*,u], Zhang Zhang [1,2,3,4,*,v], Wenming Zhao [1,3,4,*,w]

[1] *BIG Data Center, Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing 100101, China*
[2] *CAS Key Laboratory of Genome Sciences and Information, Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing 100101, China*
[3] *University of Chinese Academy of Sciences, Beijing 100049, China*
[4] *Collaborative Innovation Center of Genetics and Development, Fudan University, Shanghai 200438, China*
[5] *Center of Alzheimer's Disease, Beijing Institute for Brain Disorders, Beijing 100053, China*

* Corresponding authors.
E-mail: fangxd@big.ac.cn (Fang X), leihx@big.ac.cn (Lei H), zhangzhang@big.ac.cn (Zhang Z), zhaowm@big.ac.cn (Zhao W).
# Equal contribution.
[a] ORCID: 0000-0002-7985-7941.
[b] ORCID: 0000-0003-0848-8349.
[c] ORCID: 0000-0003-4689-3513.
[d] ORCID: 0000-0002-3852-4796.
[e] ORCID: 0000-0003-2936-1574.
[f] ORCID: 0000-0003-1296-3093.
[g] ORCID: 0000-0002-9357-4411.
[h] ORCID: 0000-0003-0953-6306.
[i] ORCID: 0000-0002-1045-1695.
[j] ORCID: 0000-0003-4580-171X.
[k] ORCID: 0000-0001-7071-666X.
[l] ORCID: 0000-0002-0956-502X.
[m] ORCID: 0000-0003-1293-4495.
[n] ORCID: 0000-0003-0688-3978.
[o] ORCID: 0000-0002-2084-7132.
[p] ORCID: 0000-0003-3810-7156.
[q] ORCID: 0000-0002-8057-0913.
[r] ORCID: 0000-0002-4761-2245.
[s] ORCID: 0000-0002-2835-4340.
[t] ORCID: 0000-0002-6628-8620.
[u] ORCID: 0000-0003-0496-0386.
[v] ORCID: 0000-0001-6603-5060.
[w] ORCID: 0000-0002-4396-8287.

*The Chinese version of this article is available at http://gpb.big.ac.cn.

Peer review under responsibility of Beijing Institute of Genomics, Chinese Academy of Sciences and Genetics Society of China.

http://dx.doi.org/10.1016/j.gpb.2017.01.001

Received 5 January 2017; accepted 7 January 2017
Available online 2 February 2017

Handled by Fangqing Zhao

KEYWORDS

Genome Sequence Archive;
GSA;
Big data;
Raw sequence data;
INSDC

Abstract With the rapid development of sequencing technologies towards higher throughput and lower cost, sequence data are generated at an unprecedentedly explosive rate. To provide an efficient and easy-to-use platform for managing huge sequence data, here we present **Genome Sequence Archive** (**GSA**; http://bigd.big.ac.cn/gsa or http://gsa.big.ac.cn), a data repository for archiving **raw sequence data**. In compliance with data standards and structures of the International Nucleotide Sequence Database Collaboration (**INSDC**), GSA adopts four data objects (BioProject, BioSample, Experiment, and Run) for data organization, accepts raw sequence reads produced by a variety of sequencing platforms, stores both sequence reads and metadata submitted from all over the world, and makes all these data publicly available to worldwide scientific communities. In the era of **big data**, GSA is not only an important complement to existing INSDC members by alleviating the increasing burdens of handling sequence data deluge, but also takes the significant responsibility for global big data archive and provides free unrestricted access to all publicly available data in support of research activities throughout the world.

Introduction

Next-generation sequencing (NGS) technologies have been extensively and routinely applied to a wide range of important issues in life and health sciences, leading to an unprecedented explosion in sequence data. Considering the increasingly higher throughput and lower costs attributable to rapid advancements of NGS technologies, large-scale sequencing projects for population genomics and precision medicine are ongoing or in the planning stages around the world, *e.g.*, the US Precision Medicine Initiative (PMI) [1], UK10 K Project [2], Icelandic Population Genome Project [3], and Dog 10 K Project [4]. As a corollary, such deluge of sequencing data poses great challenges in big data deposition, integration, and translation [5,6]. Accordingly, it is fundamentally crucial to store and manage sequencing data in support of integrative in-depth analyses and large-scale data mining.

The International Nucleotide Sequence Database Collaboration (INSDC) [7] operating between the DNA Data Bank of Japan (DDBJ) [8], the European Molecular Biology Laboratory-European Bioinformatics Institute (EMBL-EBI) [9], and the National Center for Biotechnology Information (NCBI) [10], provides valuable services for archiving a broad spectrum of sequence data. However, with the exponentially accumulating volume of sequence data, submitting big data to INSDC database resources becomes increasingly daunting and time-consuming, simply because network bandwidth is a formidable bottleneck for big data transfer across countries/regions. This situation is particularly severer in China; to our experience, for instance, submission of ~1 terabyte (TB) data to the NCBI Sequence Read Archive (SRA) takes ~2 weeks based on the 150-Mbps upload bandwidth over a shared international network in Beijing Institute of Genomics (BIG), Chinese Academy of Sciences (CAS). China, with the increasing funding support in biomedical research, has been a powerhouse in generating enormous amounts of sequencing data. Given the huge population and rich biodiversities in China, it is undoubted that data generated from sequencing projects for the Chinese population (*e.g.*, CAS PMI at http://news.xinhuanet.com/english/2016–01/09/c_134993997.htm) and domestically featured species would be growing strikingly at extraordinarily exponential rates, which accordingly brings an insurmountable challenge and burden to current practice of data submission and sharing.

To address this issue, here we present Genome Sequence Archive (GSA; http://bigd.big.ac.cn/gsa or http://gsa.big.ac.cn), a data repository for archiving raw sequence data. As a core database resource of BIG Data Center [11] (http://bigd.big.ac.cn), GSA is built based on INSDC data standards and structures and provides data archival services for scientific communities not only in China but also throughout the world. GSA accepts raw sequence reads produced by a variety of sequencing platforms, stores both sequence reads and metadata, and provides free and unrestricted access to all publicly available data for worldwide scientific communities.

Implementation

GSA is implemented with Java Server Pages (JSP; a Java programming framework for constructing dynamic web pages), Spring (an application framework and inversion of control container; http://www.springsource.org), Struts (a Model-View-Controller framework for creating Java web applications; http://struts.apache.org), and MyBatis (a persistence framework for the database connection and operation; http://www.mybatis.org). GSA adopts MySQL (http://www.mysql.org) as relational database management system to store metadata information. All codes are developed using Eclipse (http://www.eclipse.org), an integrated development environment (IDE) that features rapid development of Java-based web applications. To provide stable web services, GSA is hosted on a CentOS-7 operating system with four servers, namely, Apache serving static content, Tomcat serving

dynamic content, a MySQL server for database management, as well as a FTP server for file upload and download.

Database content and usage

Data structure and organization

Designed for compatibility, GSA follows INSDC data standards and structures. All data are organized into four objects, *i.e.*, BioProject, BioSample, Experiment, and Run (Figure 1). "BioProject", bearing an accession number prefixed with "PRJC" (where C, hereinafter, stands for China), provides an overall description for an individual research initiative, including basic description, organism, data type, submitter, funding information, and publication(s) if available. "BioSample", possessing an accession number prefixed with SAMC, contains descriptive information about biological materials used in the experiments, including sample types and attributes. "Experiment", having an accession number prefixed with CRX, provides a detailed description of treatments for a specific BioSample, including experiment intention, library method, and sequencing type. "Run", adopting an accession number prefixed with CRR, includes a list of sequence data file(s) related to a specific experiment. It is noted that "Experiment" and "Run" constitute China Read Archive (CRA). Based on these standardized data objects, GSA not only facilitates data submission and deposition, but also enables data sharing and exchange.

In addition, GSA features umbrella projects and provides an organizational structure for a large collaborative project consisting of multiple sub-projects that are funded by a same grant and have very close collaborations. GSA is well supported by CAS that functions as the national scientific think tank and academic governing body. Currently, two umbrella projects from CAS Strategic Priority Research Programs and one CAS Key Research Program make it officially mandatory to submit sequencing data to GSA.

Data archive and statistics

GSA accepts data submissions from all over the world, covers the spectrum of sequence reads generated by a variety of sequencing platforms, and accommodates several commonly-used file formats, like FASTQ, BAM, and VCF. GSA performs validations for all submitted data items to ensure data integrity and increase data reusability. Similar to INSDC members, GSA allows users to set data as either public or controlled, indicating that the data is publicly accessible or placed under controlled access over a given period of time, respectively. Regarding data security, all submitted data have copies stored in physically separate disks. Since its inception in August 2015, GSA presents a dramatic increase on data submissions in terms of the numbers of BioProjects and BioSamples, Experiments, and Runs, as well as file size (Figure 2). As of December 2016, GSA houses a total of 198 BioProjects, 8674 BioSamples, 9263 Experiments and 10,745 Runs for more than 80 species, submitted by more than 160 data providers from a total of 39 institutions, and archives more than 200 TBs of sequence data.

Data submission and retrieval

To create a submission, users need to register and log into the GSA system. Basically, to submit data to GSA, there are five straightforward steps involving BioProject, BioSample, Experiment, Run, and Sequence Files (Figure 3). In order to maximally simplify the submission procedure, GSA is equipped with a user-friendly input wizard for metadata collection. To ease sequence file uploading, GSA provides a FTP server supporting two Internet Protocols (IPv4 and IPv6). In addition, GSA provides user-friendly web interfaces for data query and browsing. Users can search the data of interest by specifying a given BioProject, BioSample, Experiment, or Run ID. Moreover, GSA allows users to conduct advanced search by inputting species name, sequencing type, sequencing platform,

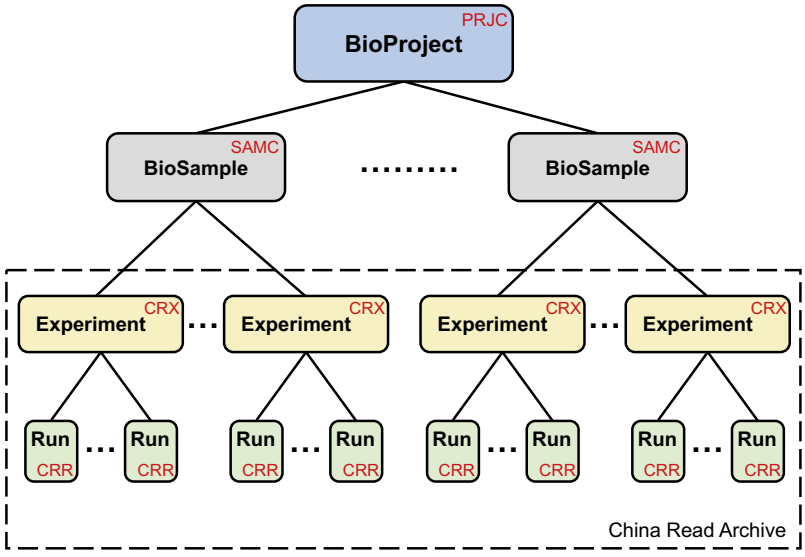


Figure 1 Data model in GSA

Prefixes of accession numbers for data objects, including BioProject, BioSample, Experiment, and Run, are indicated in red. Data objects Experiment and Run constitute China Read Archive.

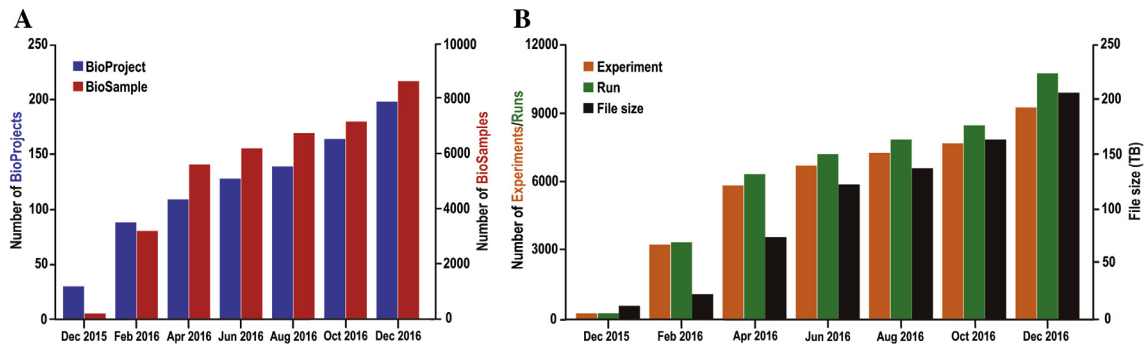


Figure 2 Data statistics of GSA

A. Numbers of BioProjects and BioSamples in GSA. **B.** Numbers of Experiments and Runs, as well as file size in GSA. All statistics are based on data submissions ranging from December 2015 to December 2016.

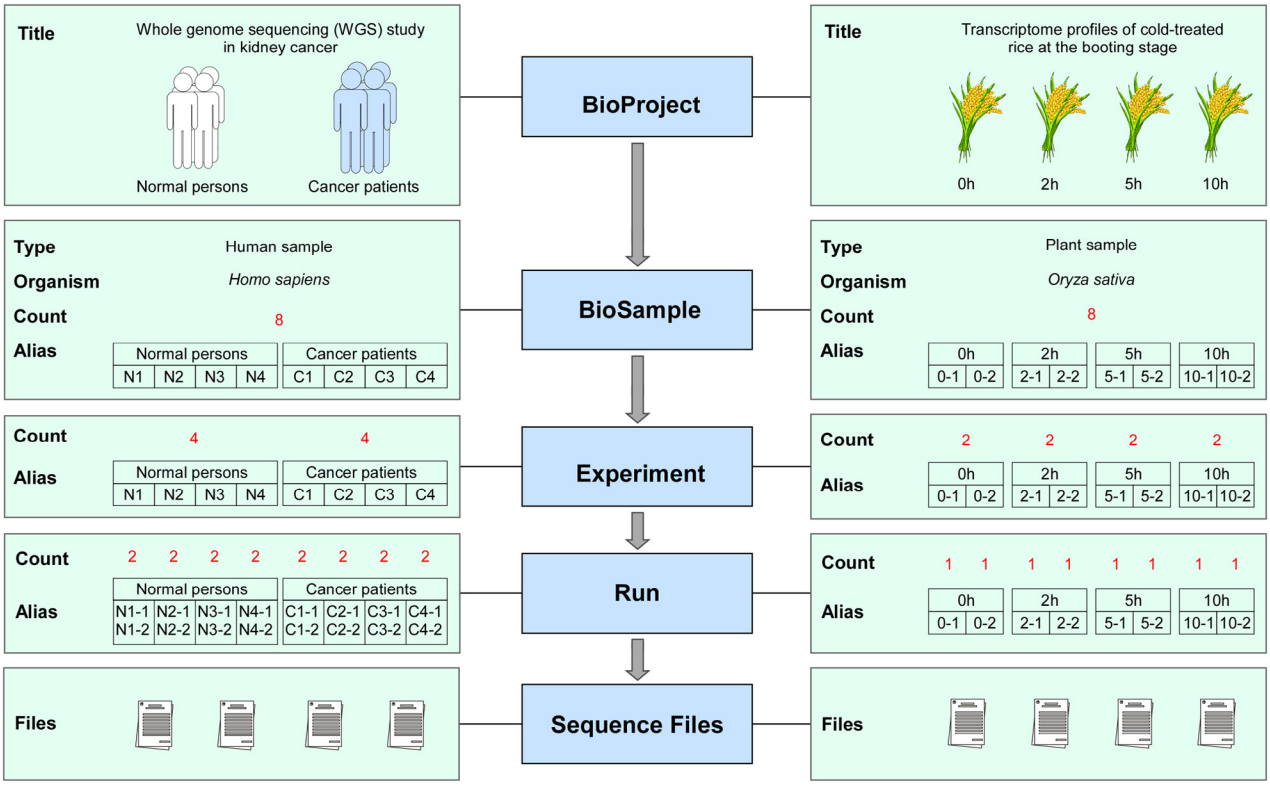


Figure 3 Graphic illustration of data submissions to GSA

Two representative studies are provided here as examples to depict the data objects involved in data submission.

disease/phenotype/trait, tissue/cell line, *etc*. GSA also allows users to browse all publicly available BioProjects, BioSamples, and Experiments.

Perspectives and concluding remarks

"*With great power comes with great responsibility*". Nowadays, China is the second largest economy, playing an increasingly important and influential role in the global economy. Equally, in academia, it is time for us to implement the practice of archiving sequence data for worldwide scientific communities, especially considering the larger quantities of sequence data generated in China. Equivalent to INSDC members, GSA is committed to archiving raw sequence data. GSA's ultimate goal, which is also the expectation from funding agencies, is to provide free archival services for raw sequence data, establish and promote a centralized archival practice in China, play an important role in global sequence data archive, and support research activities in both academia and industry throughout the world. In addition, there are also strong domestic incentives and agreements from academia, industry, and government (over 1000 supporters from more than 380 organizations; http://bigd.big.ac.cn/gdsd) to deposit data into GSA and make GSA a centralized archival resource in China.

To sum up, GSA is a data repository for archiving raw sequence data. Designed for compatibility, GSA adopts INSDC data standards and structures, archives both sequence

reads and metadata submitted from all over the world, and makes all these data publicly available to worldwide scientific communities. In the era of big data, GSA is not only an important complement to existing INSDC members by alleviating the increasing burdens of handling sequence data deluge, but also takes the significant responsibility for global big data archive and provides free unrestricted access to all publicly available data in support of research activities throughout the world. In future, we will not only upgrade infrastructure of GSA to achieve big data storage, exchange and sharing, but also will develop new functionalities to archive population-based PMI data and a variety of metagenome data.

Authors' contributions

WZ, ZZ, HL, and XF conceived of the idea and supervised the project. WZ, YW, and BT designed the system architecture. YW, JZ, FS, YY and ZB wrote the source code. QZ, ND, TC and XD tested the system. TC, LD and SSZ conducted data quality control and provided feedback service. HC, MS, YS, SZ, LL and LY constructed and maintained the network and hardware infrastructure. WZ, SSZ, and ZZ drafted the manuscript. ZZ, WZ and JX revised the manuscript. All authors read and approved the final manuscript.

Competing interests

The authors have declared no competing interests.

Acknowledgments

We sincerely thank Prof. Jingchu Luo and Prof. Weimin Zhu for their valuable suggestions and constructive discussions on this work. This work was supported by grants from the Strategic Priority Research Program of the Chinese Academy of Sciences (Grant Nos. XDB13040500 and XDA08020102); the National High-tech R&D Program (863 Program; Grant Nos. 2014AA021503 and 2015AA020108); the National Key Research Program of China (Grant Nos. 2016YFC0901603, 2016YFB0201702, 2016YFC0901903, and 2016YFC0901701); the International Partnership Program of the Chinese Academy of Sciences (Grant No. 153F11KYSB20160008); the Key Program of the Chinese Academy of Sciences (Grant No. KJZD-EW-L14); the Key Technology Talent Program of the Chinese Academy of Sciences (awarded to WZ); and the 100 Talent Program of the Chinese Academy of Sciences (awarded to ZZ).

References

[1] Collins FS, Varmus H. A new initiative on precision medicine. N Engl J Med 2015;372:793–5.
[2] Taylor PN, Porcu E, Chew S, Campbell PJ, Traglia M, Brown SJ, et al. Whole-genome sequence-based analysis of thyroid function. Nat Commun 2015;6:5681.
[3] Gudbjartsson DF, Helgason H, Gudjonsson SA, Zink F, Oddson A, Gylfason A, et al. Large-scale whole-genome sequencing of the Icelandic population. Nat Genet 2015;47:435–44.
[4] Bai B, Zhao WM, Tang BX, Wang YQ, Wang L, Zhang Z, et al. DoGSD: the dog and wolf genome SNP database. Nucleic Acids Res 2015;43:D777–83.
[5] Xue Y, Lameijer EW, Ye K, Zhang K, Chang S, Wang X, et al. Precision medicine: what challenges are we facing? Genomics Proteomics Bioinformatics 2016;14:253–61.
[6] Zhang Z, Bajic VB, Yu J, Cheung KH, Townsend JP. Data integration in bioinformatics: current efforts and challenges. In: Mahdavi MA, editor. Bioinformatics—trends and methodologies. Rijeka: InTech; 2011. p. 41–56.
[7] Cochrane G, Karsch-Mizrachi I, Takagi T. International Nucleotide Sequence Database Collaboration. The International Nucleotide Sequence Database Collaboration. Nucleic Acids Res 2016;44:D48–50.
[8] Mashima J, Kodama Y, Kosuge T, Fujisawa T, Katayama T, Nagasaki H, et al. DNA data bank of Japan (DDBJ) progress report. Nucleic Acids Res 2016;44:D51–7.
[9] Cook CE, Bergman MT, Finn RD, Cochrane G, Birney E, Apweiler R. The European Bioinformatics Institute in 2016: data growth and integration. Nucleic Acids Res 2016;44:D20–6.
[10] NCBI Resource Coordinators. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res 2016;44:D7–19.
[11] BIG Data Center Members. The BIG Data Center: from deposition to integration to translation. Nucleic Acids Res 2017;45:D18–24.