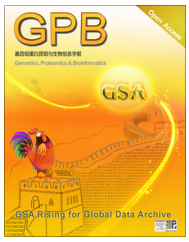
Genomics Proteomics Bioinformatics 15 (2017) 14–18

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