**Workshop title:** Database Resources of the BIG Data Center

**Lead organizers:** Yiming Bao, Wemin Zhu, Zhang Zhang, Wenming Zhao, Jingfa Xiao

**Time:** 1.5 *hrs*

**Objective:**

The BIG Data Center advances life & health sciences by providing open access to a variety of resources, with the aim to translate big data into big discoveries and support activities in both academia and industry. With the vast amounts of omics data generated at ever-greater scales and rates, the BIG Data Center is continually expanding, updating and enriching its core database resources through big-data integration and value-added curation. We will present an overview of core database resources and ongoing projects. We will also discuss the challenges we face and share our collective views on big data integration and translation in the future. This workshop will present a combination of short talks, and open guided discussion.

**Relevance:**

This workshop is targeted to all conference attendees from all scientific and personal backgrounds and career stages.

**Proposed format (times are approximate) with confirmed speakers:**

1) **Overview of BIG Data Center resources|** Short talk by Yiming Bao **|** 10 minutes

*Summary*: The BIG Data Center at Beijing Institute of Genomics (BIG) of the Chinese Academy of Sciences provides freely open access to a suite of database resources in support of worldwide research activities in both academia and industry. With the vast amounts of omics data generated at ever-greater scales and rates, the BIG Data Center is continually expanding, updating and enriching its core database resources through big-data integration and value-added curation, including BioCode, BioProject, BioSample, Genome Sequence Archive, Genome Warehouse, Genome Variation Map, Gene Expression Nebulas, Methylation Bank, and Science Wikis. In addition, three featured web services are provided, viz., BIG Search, BIG SSO and Gsub. All of these resources are publicly accessible through the home page of the BIG Data Center at http://bigd.big.ac.cn.

2) ***De novo* assembly of a Chinese genome |** Short talk by Zhenglin Du **|** 8 minutes

*Summary:* Advances in genome sequencing and assembly technology provide an opportunity to investigate the human genetic diversity across different population groups. Given the huge genetic diversity between the population of North and South China, here we report a *de novo* assembly of a Northern Han individual (NH1.0), using the single-molecule real-time sequencing platform (PacBio), 10x Genomics linked reads, Illumina pair-end reads and Bionano Saphyr optical mapping system. The genome of NH1.0 was assembled with a contig N50 size of 3.6Mbp and a scaffold N50 size of 46.63 Mbp, which covered 15 chromosome arms with coverage more than 85% of euchromatic regions. It is indicated that the hybrid approach of combining PacBio and 10X Genomics technologies can highly improve the integrity of genome assembly. 2,218,371 SNPs and 18,613 structural variations (insertion and deletion) were detected in the NH1.0 genome compared with the GRCh38 genome, and 55.9% SNPs and 10.1% SVs shared with other two southern Chinese genomes. This work presents the most contiguous human genome assembly for the Chinese population, with extensive investigation of population-specific genetic variation for precision medicine.

3) **Genome variation map: a repository of genome variations for global animals and plants |** Short talk by Shuhui Song | 8 minutes

*Summary:* 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 animals and plants, 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 8,669 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.

4) **GSA: Genome Sequence Archive |** Short talk by Yanqing Wang | 8 minutes

*Summary*: 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 GSA (Genome Sequence Archive; http://bigd.big.ac.cn/gsa or http://gsa.big.ac.cn), a data repository specialized for archiving raw sequence data. In compliance with data standards and structures of 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 sequencing data deluge, but also takes the significant responsibility for global big data archive. As of December 2017, GSA archives a total of 22,501 Experiments and 24,591 Runs and houses more than 430 Terabytes of sequencing data in size, submitted from 275 submitters of 91 Organizations. All released data in GSA are publicly available through the FTP site at ftp://download.big.ac.cn/gsa.

5) **MethBank: a database of DNA methylomes across a variety of species** | Short talk by Rujiao Li | 8 minutes

*Summary*: MethBank (http://bigd.big.ac.cn/methbank) is a database that integrates high-quality DNA methylomes across a variety of species, provides an interactive browser for visualization of methylation data, as well as equips friendly web interfaces for data presentation and search.

MethBank features large-scale integration of high-quality methylomes, involving 34 consensus reference methylomes derived from a large number of human samples, 336 single-base resolution methylomes from different developmental stages and/or tissues of five plants, and 18 single-base resolution methylomes from gametes and early embryos at multiple stages of two animals. Additionally, it enables systematic identification of methylation sites closely associated with age, sites with constant methylation levels across different ages, differentially methylated promoters, age-specific differentially methylated cytosines/regions, and methylated CpG islands. Moreover, MethBank provides tools to estimate human methylation age online and to identify differentially methylated promoters, respectively. As an important resource in the BIG Data Center, MethBank will be frequently upgraded and improved. We will integrate more high-quality methylomes from a wider range of species and develop new functionalities that allow users to submit analyzed data to MethBank. Taken together, MethBank is of great help for deciphering DNA methylation regulatory mechanisms for epigenetic studies.

6) **LncRNAWiki: community curation of human long non-coding RNAs |** Short talk by Lina Ma | 8 minutes

*Summary*: LncRNAWiki (http://lncrna.big.ac.cn) is a wiki-based, publicly editable and open-content platform for community curation of human long non-coding RNAs (lncRNAs). Since its inception, LncRNAWiki has achieved more than 3.3 million views and over 1.3 million edits (http://lncrna.big.ac.cn/index.php/Special:Statistics). In the past two years, LncRNAWiki has been significantly improved and enriched. LncRNAWiki keeps frequent updates by community annotation of human lncRNAs and integration of newly identified lncRNAs with experimental evidence. In contrast to the previous version that had 86 curated lncRNAs, the updated version has a total of 959 lncRNAs that were community-curated (a detailed list is available at http://lncrna.big.ac.cn/index.php/LncRNAWiki:Featured). Among these 959 curated lncRNAs, 322 have been experimentally proved to be associated with cancer and other diseases. We identified disease-lncRNA associations and catalogued lncRNAs based on disease types (http://lncrna.big.ac.cn/index.php/LncRNAWiki:Disease), providing users with easy access to a comprehensive collection of disease and lncRNA associations. To provide users with the most comprehensive list of human lncRNAs, we also integrate the predicted lncRNAs from other databases, and identify novel lncRNAs based on RNA-seq data. Up to now, LncRNAWiki has integrated 252,252 lncRNAs from the existing databases and identified 36,528 novel lncRNAs, including a total of 141,596 lncRNA genes.

7) **Genome Warehouse: a public repository housing genome-scale data |** Short talk by Meili Chen | 8 minutes

*Summary*: The Genome Warehouse (GWH; http://bigd.big.ac.cn/gwh) is a public repository housing genome-scale data for a wide range of species and delivering a series of web services for genome data submission, storage, release and sharing. For each species, GWH contains detailed genome-related information including species metadata, genome assembly, sequence data and the corresponding annotations. Particularly, to archive high-quality genome sequences and genome annotation information, GWH adopts a uniform standardized procedure for quality control. Since the availability of online submission service in July 2017, GWH has accommodated 12 direct genome submissions, viz., four animals (one has been released), four plants (two have been released), one fungi and three bacteria, showing the great promise to have more and more genome data submissions in the wake of high-throughput sequencing capability and large-scale sequencing-based projects. Besides, GWH is also enriched by integrating 138 newly released genomes (61 animals and 77 plants) from NCBI.

8) **Big data integration of rice in IC4R** | Short talk by Lili Hao | 8 minutes
*Summary*: Rice is the most important staple food for a large part of the world's human population and also a key model organism for plant research. Information Commons for Rice (IC4R; http://ic4r.org), a rice knowledgebase featuring adoption of an extensible and sustainable architecture, integrates multiple omics data through community-contributed modules. Each module is developed and maintained by different committed groups, deals with data collection, processing and visualization, and delivers data on-demand via web services. In the current version, IC4R incorporates a variety of rice data through multiple committed modules, including genome-wide expression profiles derived entirely from RNA-Seq data, resequencing-based genomic variations obtained from re-sequencing data of thousands of rice varieties, plant homologous genes covering multiple diverse plant species, post-translational modifications, rice-related literatures, and gene annotations contributed by the rice research community. Unlike extant related databases, IC4R is designed for scalability and sustainability and thus also features collaborative integration of rice data and low costs for database update and maintenance. Future directions of IC4R include incorporation of other omics data and association of multiple omics data with agronomically important traits, dedicating to build IC4R into a valuable knowledgebase for both basic and translational researches in rice.

9) **Open guided discussion** | 20 minutes

**About the organizers:**
**Yiming Bao** is the Director of BIG Data Center, Beijing Institute of Genomics (BIG), Chinese Academy of Sciences (CAS), and a Professor in "100Talent" Program of CAS. Before joining BIG in 2017, Dr. Bao was a Staff Scientist at the National Center for Biotechnology Information (NCBI)/NLM/NIH, USA, where he managed the NCBI Influenza Virus Resource and the virus classification tool PASC, among other duties on the NCBI viral genome project. Dr. Bao authored more than 50 research articles, and received the 2006 NIH Merit Award. His research interests are on bioinformatics and viral genomes.

**Wemin Zhu** is a senior bioinformatitian, currently appointed as distinguished-professor at Beijing Proteome Research Center and previously worked at EMBL-EBI for several years. His main research interests include the collection, manangement, integration, annotation, analysis and mining of biological big data, and information system and infrastructure development for database and standardization. He serves as a scientific advisory board member for BIG Data Center, Beijing Institute of Genomics, Chinese Academy of Sciences.

**Zhang Zhang** is a Professor of Beijing Institute of Genomics (BIG), Chinese Academy of Sciences (CAS) and serves as Executive Director of BIG Data Center. Dr. Zhang was elected in the CAS 100-Talent Program in 2011. His research focuses on big data integration and mining and computational precision health genomics. He is an Executive Committee Member of International Society for Biocuration and acts as Associate Editor-in-Chief for Genomics Proteomics & Bioinformatics and Asian Associate Editor for Briefings in Bioinformatics.

**Wenming Zhao** is the associate director of BIG data center and the lead of the GSA working group, he got the "CAS Key Technology Talent Program" in 2015. He mainly focused on the methodology research for NGS data analysis and the bioinformatics database construction. Wenming's Team has constructed the first Genome Sequence Archive database in China, and got the approval by lots of journals. He also focused on the High performance computer, and constructed a strong computing environment for BIG.

**Jingfa Xiao** is a Professor at the Beijing Institute of Genomics (BIG), Chinese Academy of Sciences (CAS). Prof. Xiao leads the group of precision medicine omics databank in BIG Data Center. Prof. Xiao is the board member of genomics committee of genetics society of china and computational systems biology society of ORSC.

**Speakers:**
**Yiming Bao,** Director of BIG Data Center, BIG, CAS, Beijing, China
**Linzheng Du,** Team leader of CRG, BIG Data Center, BIG, CAS, Beijing, China
**Shuhui Song,** Team leader of GVM & EHR, BIG Data Center, BIG, CAS, Beijing, China
**Yanqing Wang,** Team leader of GSA, BIG Data Center, BIG, CAS, Beijing, China
**Rujiao Li,** Team leader of MethBank, BIG Data Center, BIG, CAS, Beijing, China
**Lina Ma,** Team leader of ScienceWikis, BIG Data Center, BIG, CAS, Beijing, China
**Meili Chen,** Team leader of GWH, BIG Data Center, BIG, CAS, Beijing, China
**Lili Hao,** Team leader of IC4R & GEN, BIG Data Center, BIG, CAS, Beijing, China