Big genomics data, big scientific impact: New challenges for further development of life science

November 28 2012

BGI, the world's largest genomics organization, today announced its latest advances in the analysis, management and dissemination of "Big Genomics Data" at their 3rd bioinformatics software and data release conference. Released at the conference include new bioinformatics analysis pipelines and software, including SOAPhecate v2.5 and SOAPgaea, as well as an updated version of EasyGenomics, one cloud-based bioinformatics solution. Additionally, BGI's big-data journal *GigaScience* also provided an update on its integrated GigaDB database and reported on plans for its data analysis platform based on the Galaxy workflow system.

Genomics and next-generation sequencing (NGS) technologies have revolutionized life sciences research. With the cost of DNA sequencing steadily plummeting in price, the amount of data generated with NGS technologies continues to grow at an unprecedented pace. This has led in recent years to an over 400,000% increase in daily sequencing data generation. In the age of "Big Genomics Data", the handling, storing and sharing of these tremendous volumes of data has become a significant research bottleneck.

For dealing with big data efficiently, BGI has integrated the Apache™ Hadoop™ MapReduce framework into algorithms for NGS analyses. Based on this framework, they have also developed two new algorithms: SOAP-Hecate and SOAP-Gaea. These algorithms are two of the key

components of the flexible green cloud computing infrastructure at BGI for de novo Assembly and NGS Analysis. They have successfully been applied into analyzing the sequencing data in clinical and biological research with a fast turnaround time, high efficiency and low cost.

In the conference, Yan Li, Director of Bioinformatics Products from the BGI, announced a new version of distributed genome assembler – SOAP-Hecate v2.5. This software not only outputs linearized sequences, but is also a flexible and easy-to-use platform to study the complexity and characteristics of a genome. The scalability of SOAP-Hecate enables researchers to control the assembly time by choosing different size of the cluster. SOAP-Hecate v2.5 is able to finish the assembly of a human whole genome within two days.

Yan Li also presented the application of SOAP-Gaea in genetics research. The new version of SOAP-Gaea is integrated fully into a computational pipeline for variation detection. At present, the application of SOAP-Gaea in Cancer genomics studies has reduced the analysis time from two weeks to two days. EasyGenomics- Next Generation Bioinformatics on the Cloud

"Our goal is to make NGS analysis easier and faster, and get answers everyone can trust." said, Dr. Xing Xu, Director of Cloud Computing Products at BGI. Dr. Xu gave an overview of EasyGenomics and its advantages at the conference. He says "EasyGenomics is a robust cloud-based bioinformatics solution for NGS data analysis, which is engineered with BGI's state-of-the-art technologies and platforms. Through this excellent platform, researchers have the ability to access these informatics resources at their fingertips".

"EasyGenomics can liberate scientists from the burden of handling and processing the huge amount of NGS data." says Dr. Xu. As the sequencing cost drops sharply, a vast amounts of NGS data has been and

will continue to be generated, BGI for example generating terabytes of data per day. This flood of data becomes one of the biggest obstacles for researchers to further explore the mysteries of life science, and so Dr. Xu added, "EasyGenomics emerges as the times require."

According to Xu's introduction, EasyGenomics offers a SaaS (Software as a Service) based NGS bioinformatics analysis solution based upon BGI's cloud infrastructure and cutting-edge Aspera fasp file transport technology. Users only need a standard web browser to access the services of EasyGenomics from anywhere in the world. With EasyGenomics, users are freed from tedious resource maintenance and management as well as counter-intuitive command-line tools. Xu said,

"Performing bioinformatics analyses becomes as easy as doing online shopping with just a few clicks." EasyGenomics offers high-speed data exchange solutions that are 10-100X faster than conventional ftp://ftp. This robust platform contains popular workflows for whole genome and exome sequencing analysis as well as RNA-seq, de novo assembly, and small RNA data analysis tools. It also provides extensive quality control statistics and various informative reports, such as reports on sequencing quality, mapping, coverage, and enrichment statistics, allowing users to access data quality, evaluate analysis performance, and identify potential issues.

GigaDB and Galaxy - Revolutionizing Data Dissemination, Organization and Analysis

GigaDB is the database of the journal *GigaScience*, which is a repository to host publicly available, large-scale data sets, each with a unique Digital Object Identifier (DOI) for their citation and tracking. *GigaScience* was launched by BGI and its publishing partner BioMed Central. It is an online, open-access and open-data journal that publishes

studies involving large-scale data from the life and biomedical sciences. This big-data journal provides a new model in scientific publishing: combining article and data publication.

In the conference, Peter Li from *GigaScience* is presenting what can be expected from the new version of GigaDB to be released in early 2013. This newly updated database will provide a user-friendly interface for querying and downloading data sets. It currently contains over 39 data sets, many previously unpublished from the BGI, including genomic, mass spectrometry, transcriptomic, epigenomic and metagenomic data.

In addition to GigaDB, Peter says they and the CUHK-BGI Innovation Institute of Trans-omics (CBIIT) have been developing a data analysis platform based on the Galaxy workflow system through which we will be making the tools and data processing pipelines reported in *GigaScience* available for the research community. As a pilot project, they have integrated the next generation sequencing data analysis tools from the BGI SOAP package suite of tools into their Galaxy platform to provide an opportunity for the community to use its functionality from within automated pipelines.

Peter said, "We expect to make *GigaScience* workflows available from myExperiment, an online repository of scientific workflows, in the near future. GigaDB will eventually be integrated with our Galaxy-based data analysis platform so the results in our papers and the data hosted by us can be analyzed and viewed in a more reproducible and reusable manner."

Provided by BGI Shenzhen

genomics-scientific-impact-life.html