How to make bio-DB’s and services sustainable?

Genome Informatics Laboratory & DDBJ, National Institute of Genetics, JAPAN
Yaz Nakamura

中村保一
Roche (454): GS FLX+ System
illumina: Genome Analyzer IIx System
Life Technologies: 5500 xl SOLiD System
LATEST ANNOUNCEMENTS

THURSDAY NOVEMBER 27, 2014

**Phase3 variant calls for chrY are available, variant calls for chrX have been updated**

Our final release of the Phase 3 variant set is now available on the FTP site, including a newly added VCF file for chrY.

The chrY variant calls were made with a different process from that of the autosomes; a separate README is available in the release directory describing some details.

The chrX VCF file has been updated to include standard annotation including DP, continental super-population allele frequency.

The gsea file in the release directory is now wgs containing autosomes, chrX and Y.

Two algorithms were used to discover short tandem repeats (STRs) in the phase3 data. However the STRs did not make into the final integrated call set. They are now available separately [here](link).

The VCF files in the main release directory are also now available [here](link) in BCF format for faster processing time.

This release includes super population allele frequencies in the main release VCFs and **functional annotation** from the Ensembl Variant Effect Predictor along side many other datasets in the supporting directory. The complete list of data is covered in the Supporting Directory README. The issues which have been raised and resolved since our initial release are covered in the Known Issues README.

Please send any questions about this data set to [info@1000genomes.org](mailto:info@1000genomes.org)

---

**Recent project announcements**

THURSDAY NOVEMBER 20, 2014

**EMBL-EBI 1000 Genomes FTP site will be at reduced capacity between November 21th and December 8th**

The EMBL-EBI FTP site will be at reduced capacity between November 21st and December 8th due to EMBL-EBI consolidating its web infrastructure into a single data centre.
Welcome to the 1001 Genomes Project

The 1001 Genomes Project was launched at the beginning of 2008 to discover the whole-genome sequence variation in 1001 strains (accessions) of the reference plant Arabidopsis thaliana. The resulting information is paving the way for a new era of genetics that identifies alleles underlying phenotypic diversity across the entire genome and the entire species. Each of the accessions in the 1001 Genomes project is an introgressed line with seeds that are freely available from the stock centre to all our colleagues. Unlimited numbers of plants with identical genotype can be grown and phenotyped for each accession, in as many environments as desired, and so the sequence information we collect can be used directly in association and QTL studies. The analyses enabled by this project will have broad implications for areas as diverse as evolutionary sciences, plant breeding and human genetics.

The complete genome sequences of over 80 accessions were released in early 2010 by the Max Planck Institute, and many more have been added since by the Salk Institute, the Greger Mendel Institute and Monsanto. As of September 2014, over 1100 lines have been sequenced, and a publication that will describe an integrated analysis of the data is forthcoming.
NGS’s in 2017

PacBio RSII System

illumina: HiSeq 2500 System

Oxford NANOPORE MinION/SmidgION
The 3,000 Rice Genomes Project

A single genome cannot reveal the large stockpile of genetic diversity in rice and hence many potentially important genes are not present in the handful of lines that have been sequenced over the last decade. So, to drastically change this dynamic, IRRI—in collaboration with BGI in Shenzhen, China, and the Chinese Academy of Agricultural Sciences (CAAS)—has completed the sequencing of 3,000 rice genomes of varieties and lines representing 89 countries (see figure) now housed in the IRGC (82%) and CAAS’s genebank (18%). “This is an unparalleled development in plant science for a major food crop,” says Ken McNally, senior scientist in the TTC GRC and a project team member.

Geographic distribution of 3,000 representative rice accessions whose genomes have been sequenced. Numbers in parentheses indicate the number of countries in each region.
What is UK10K?

The UK10K project will enable researchers in the UK and beyond to better understand the link between low-frequency and rare genetic changes, and human disease caused by harmful changes to the proteins the body makes.

Although many hundreds of genes that are involved in causing disease have already been identified, it is believed that many more remain to be discovered. The UK10K project aims to help uncover them by studying the genetic code of 10,000 people in much finer detail than ever before.

Project Design

Not all genetic changes are harmful or lead to disease, so the project is taking a two-pronged approach to identify rare variants and their effects:

- by studying and comparing the DNA of 4,000 people whose physical characteristics are well documented, the project aims to identify those changes that have no discernible effect and those that may be linked to a particular disease;
- by studying the changes within protein-coding areas of DNA that tell the body how to make proteins of 6,000 people with extreme health problems and comparing them with the first group, it is hoped to find only those changes in DNA that are responsible for the particular health problems observed.

The project received a £10.5 million funding award from the Wellcome Trust in March 2010 and sequencing started in late 2010. For more information, please use the links on the right hand side.
1996 *Synechocystis* sp. PCC 6803
2000 *Arabidopsis thaliana*
2000 *Mesorhizobium loti*
2001 *Anabaena (Nostoc) sp. PCC 7120*
2002 *Bradyrhizobium japonicum*
2002 *Thermosynechococcus elongatus* BP-1
2003 *Gloeobacter violaceus* PCC 7421
2007 *Microcystis aeruginosa* NIES-843
2008 *Lotus japonicus*
2012 *Bradyrhizobium* sp. S23321
2012 *Solanum lycopersicum*
2012 *Eucalyptus globulus*
2012 *Hevea brasiliensis* (Para rubber tree) [draft]
2014 *Klebsormidium flaccidum* (an algae)
2014 *Weissella oryzae* SG25T
2014 *Lactobacillus oryzae* SG293T
2014 *Lactobacillus hokkaidonensis* LOOC260T
2015 *Nostoc* sp. NIES-3756
2016 *Hevea brasiliensis* ver. 2 [improved by using Pacbio]
2016 *Marchantia polymorpha* (a liverwort)
2017 *Citrus unshiu*
Genome projects undergo

a rubber tree, *Hevea brasiliensis*  
1.4 Gb genome  
52.7 k scaffolds / N50: 120 kb

a liverwort, *Marchantia polymorpha*  
220 Mb genome  
4.4 k scaffolds / N50: 1.3 Mb

a citrus tree, *Citrus unshiu*  
360 Mb genome  
21.1 k scaffolds / N50: 385 kb
DDBJ (http://www.ddbj.nig.ac.jp/)
INSDC

DDBJ is a partner of INSACD

International Nucleotide Sequence Databank Collaboration

IAC: International Advisory Committee
ICM: International Collaborative Meeting

<table>
<thead>
<tr>
<th>Data type</th>
<th>DDBJ Center</th>
<th>EMBL-EBI</th>
<th>NCBI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Next generation</td>
<td>Sequence Read Archive</td>
<td>European Nucleotide Archive (ENA)</td>
<td>Sequence Read Archive</td>
</tr>
<tr>
<td>reads</td>
<td>DDBJ</td>
<td></td>
<td>Trace Archive</td>
</tr>
<tr>
<td>Capillary reads</td>
<td>Trace Archive</td>
<td></td>
<td>GenBank</td>
</tr>
<tr>
<td>Annotated sequence</td>
<td>DDBJ</td>
<td></td>
<td>BioSample</td>
</tr>
<tr>
<td>Samples</td>
<td>BioSample</td>
<td></td>
<td>BioSample</td>
</tr>
<tr>
<td>Studies</td>
<td>BioProject</td>
<td></td>
<td>BioProject</td>
</tr>
</tbody>
</table>

from insdc.org
The business of DNA Databank

- Determined Nucleotide Sequence
- Checking Data and Metadata
- Putting it into the Database
- Open and Share it via the Internet

12 curators (among them 6 have Ph.D.)
The amount of data continues to grow.
Traditional DDBJ / INSDC

Number of Entries by Contributors to DDBJ Release

entries: 196 million

Number of Bases by Contributors to DDBJ Release

bases: 220 billion
SRA growth (NGS row/bam data amount)


12,164,994,572,931,783 total bases
5,020,021,766,107,571 open access bases

12,164 TeraBases
= over 12 PetaBases
Number of species in CyanoBase

1996: 1
2007: 5
2010: 39
2016: 376
The budget continues to shrink.
It costs.
DDBJ (from Release note 92) 44

ENA (from Release note 115) 27

GenBank (from Release note 195) 68
NIG supercomputer (2014.3 ver.)

12.5 PB

7 PB
Lustre
high-speed HDD

5.5 PB
MAID
energy-saving HDD

> 300 TFlops

“thin”
64GB memory
x 554 nodes

“medium”
2TB memory
x 10

“fat”
10TB memory
(SGI UV)
Trouble
Introduction

The Genome Sequence Archive (GSA) is a data repository for genome, transcriptome and other omics primitive sequencing data. It archives raw sequence data produced from a wide variety of sequencing platforms. In addition to raw sequencing data, GSA also accommodates secondary analyzed files in acceptable formats (like BAM, VCF).

GSA is one of database resources in BIG Data Center (BIGD), part of Beijing Institute of Genomics (BIG), Chinese Academy of Sciences (CAS), serving as a primary archive of genome sequencing data for worldwide institutions and laboratories.

Compatible with standards and structures adopted in extant archives in International Nucleotide Sequence Database Collaboration, GSA covers the spectrum of raw sequencing reads, accepts the submissions from all over the world, archives sequencing data and metadata and makes these data publicly available to worldwide scientific communities.

China Genomic Data Sharing Initiative

Data Submission

1. Register BioProject
2. Register BioSample
3. Register Experiment
4. Register Run
5. Upload Sequence File

Data deposited to GSA has been reported by a paper published in Nature Communications. (2017-06-25)
Data deposited to GSA has been reported by a paper published in Genome Biology. (2017-06-16)
Data deposited to GSA has been reported by a paper published in Cell Research. (2017-05-02)
Data deposited to GSA has been reported by a paper published in Mol Biol Evol. (2017-02-17)
Data deposited to GSA has been reported by a paper published in AJHG. (2016-09-03)
Data deposited to GSA has been reported by a paper published in Current Biology. (2016-06-20)
Data deposited to GSA has been reported by a paper published in Stem Cell Reports. (2016-07-16)
Data deposited to GSA has been reported by a paper published in Journal of Cell Science. (2015-05-17)
Data deposited to GSA has been reported by a paper published in PNAS. (2015-11-10)

Latest Released Projects

<table>
<thead>
<tr>
<th>Accession</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>PRJCA003445</td>
<td>Expression profiling of SCC15 cells treated by PAM</td>
</tr>
<tr>
<td>PRJCA00313</td>
<td>Shuhu498 rice genome sequencing and assembly</td>
</tr>
</tbody>
</table>
GSA Data Model

The data model adopted by GSA consists of Project, Sample, Experiment and Run. Unlike other data depositories, GSA features "Umbrella Project", which is used to, albeit optional when registering a BioProject, effectively manage multiple highly relevant projects supported by a collaborative grant or mega grant, e.g., 1000 Human Genomes Project, Dog 10K Genomes Project.

GSA Data Relationships

Data relationships in GSA are as follows.

- **BioProject** is an overall description of a single research initiative, typically involving multiple samples.
- **BioSample** describes biological source material; each physically unique specimen should be registered as a single BioSample with a unique set of attributes.
- **Experiment** describes detailed treatment for each BioSample. Each sample may have multiple experiments and each experiment belongs to a specific BioSample.
- **Run** describes technical batch related files that belong to a specific Experiment. Each Run may have multiple files.

Frequently Asked Questions

Answers to some of the most frequently asked questions submitted to the GSA are listed as follows.
GSA Data Model

The data model adopted by GSA consists of Project, Sample, Experiment and Run. Unlike other data depositories, GSA features "Umbrella Project", which is used to, albeit optional when registering a BioProject, effectively manage multiple highly relevant projects supported by a collaborative grant or mega grant, e.g., 1000 Human Genomes Project, Dog 10K Genomes Project.

BioProject (Former DRA Study)
- Project Description
- Grants
- Publications

BioSample (Former DRA Sample)
- Sample Description
- Taxonomy ID

Experiments and Runs
- Library layout
- Sequencing platform

BioProject
- BioSample
- Experiment
- Run

Sequence data files (fastq, BAM)

BioProject is an overall description of a single research initiative, typically involving multiple samples. BioSample describes biological source material; each physically unique specimen should be registered as a single BioSample with a unique set of attributes. Experiment describes detailed treatment for each BioSample. Each sample may have multiple experiments and each experiment belongs to a specific BioSample. Run describes technical batch related files that belong to a specific Experiment. Each Run may have multiple files.

Frequently Asked Questions

Answers to some of the most frequently asked questions submitted to the GSA are listed as follows.
How to Cite

When you have successfully submitted data to GSA, please consider to use the following words to describe data deposition in your manuscript.

The raw sequence data reported in this paper have been deposited in the Genome Sequence Archive (Genomics, Proteomics & Bioinformatics 2017) in BIG Data Center (Nucleic Acids Res 2017), Beijing Institute of Genomics (BIG), Chinese Academy of Sciences, under accession numbers PRJCAxxxxx, PRJCAyyyyy that are publicly accessible at http://bigd.big.ac.cn/gsa.

Please cite the following required publications.

- **The BIG Data Center: from deposition to integration to translation.** Nucleic Acids Res 2017, 45(D1): D18-D24. [PMID=27699656]
Hints:
TOWARDS COORDINATED INTERNATIONAL SUPPORT OF CORE DATA RESOURCES FOR THE LIFE SCIENCES

On November 18-19, 2016, the Human Frontier Science Program Organization (HFSPO) hosted a meeting of senior managers of key data resources and leaders of several major funding organizations to discuss the challenges associated with sustaining biological and biomedical (i.e., life sciences) data resources and associated infrastructure. A strong consensus emerged from the group that core data resources for the life sciences should be supported through a coordinated international effort(s) that better ensure long-term sustainability and that appropriately align funding with scientific impact. Ideally, funding for such data resources should allow for access at no charge, as is presently the usual (and preferred) mechanism.
European Molecular Biology Laboratory

EMBL member states:
Austria, Belgium, Croatia, Czech Republic, Denmark, Finland, France, Germany, Greece, Iceland, Ireland, Israel, Italy, Luxembourg, Netherlands, Norway, Portugal, Spain, Sweden, Switzerland and the UK

Associate member states:
Argentina, Australia

Prospect member states:
Hungary, Poland, Slovak Republic