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

Genome Informatics Laboratory &
DDBJ, National Institute of Genetics, JAPAN
Yaz Nakamura
中村保一

NGS's in 2010

Roche (454): GS FLX+ System

illumina: Genome Analyzer IIx System

Life Technologies: 5500 xl SOLiD System

www.1000genomes.org

1000 Genomes

A Deep Catalog of Human Genetic Variation

LATEST ANNOUNCEMENTS

THURSDAY NOVEMBER 06, 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 site 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.

The VCF files in the main release directory are also now available here 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

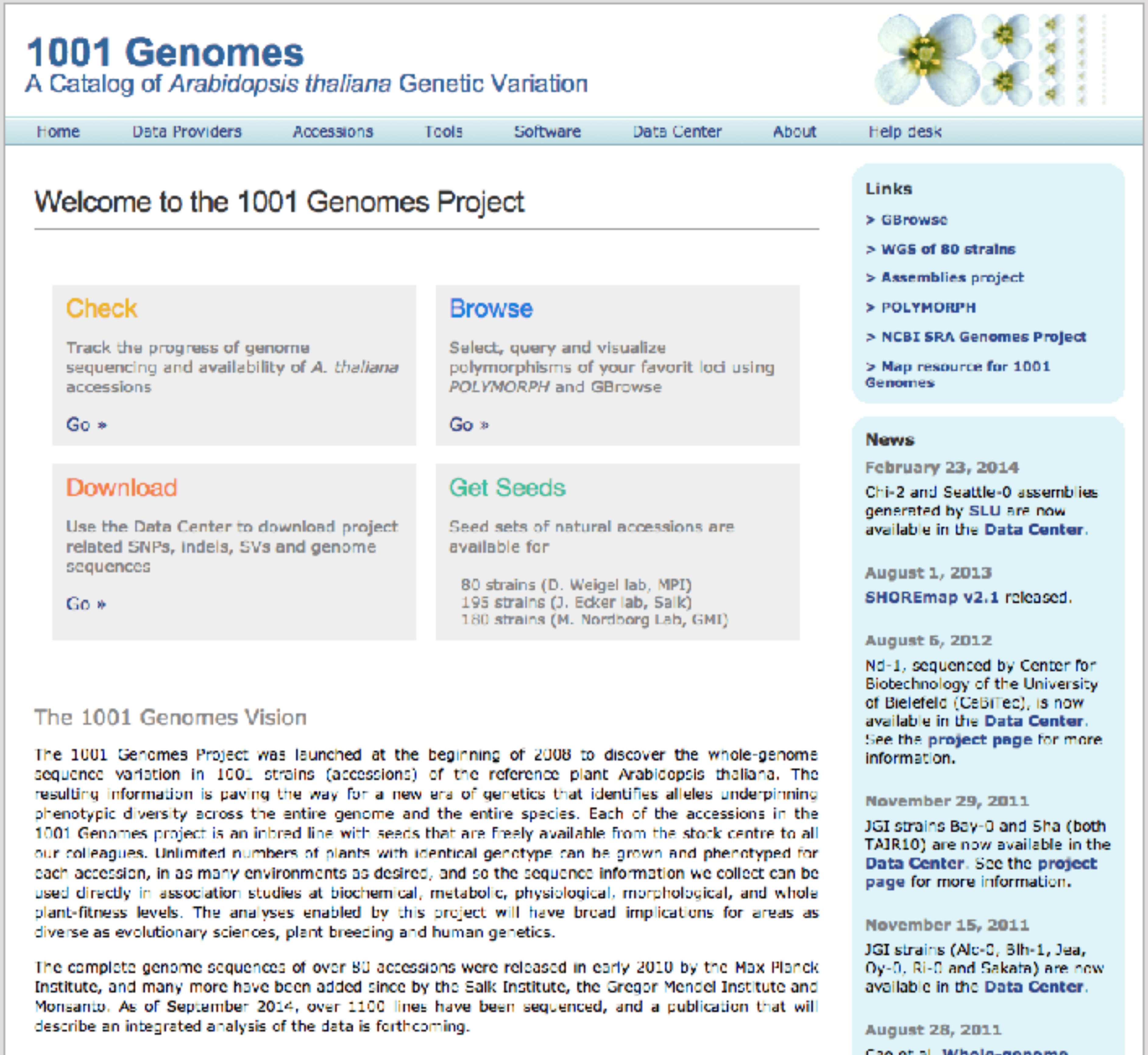
Recent project announcements

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 wconsolidating its web infrastructure into a single data centre.

1001genomes.org



Welcome to the 1001 Genomes Project

Check

Track the progress of genome sequencing and availability of *A. thaliana* accessions

Go »

Browse

Select, query and visualize polymorphisms of your favorit loci using *POLYMORPH* and GBrowse

Go »

Download

Use the Data Center to download project related SNPs, indels, SVs and genome sequences

Go »

Get Seeds

Seed sets of natural accessions are available for

80 strains (D. Weigel lab, MPI)
195 strains (J. Ecker lab, Salk)
180 strains (M. Nordborg Lab, GMI)

The 1001 Genomes Vision

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 underpinning phenotypic diversity across the entire genome and the entire species. Each of the accessions in the 1001 Genomes project is an inbred 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 studies at biochemical, metabolic, physiological, morphological, and whole plant-fitness levels. 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 Gregor 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.

Links

- > GBrowse
- > WGS of 80 strains
- > Assemblies project
- > POLYMORPH
- > NCBI SRA Genomes Project
- > Map resource for 1001 Genomes

News

February 23, 2014

Chi-2 and Seattle-0 assemblies generated by **SLU** are now available in the **Data Center**.

August 1, 2013

SHOREmap v2.1 released.

August 6, 2012

Nd-1, sequenced by Center for Biotechnology of the University of Bielefeld (CeBiTec), is now available in the **Data Center**. See the **project page** for more information.

November 29, 2011

JGI strains Bay-0 and Sha (both TAIR10) are now available in the **Data Center**. See the **project page** for more information.

November 15, 2011

JGI strains (Alc-0, Blh-1, Jea, Oy-0, Ri-0 and Sakata) are now available in the **Data Center**.

August 28, 2011

Cao et al. **Whole-genome**

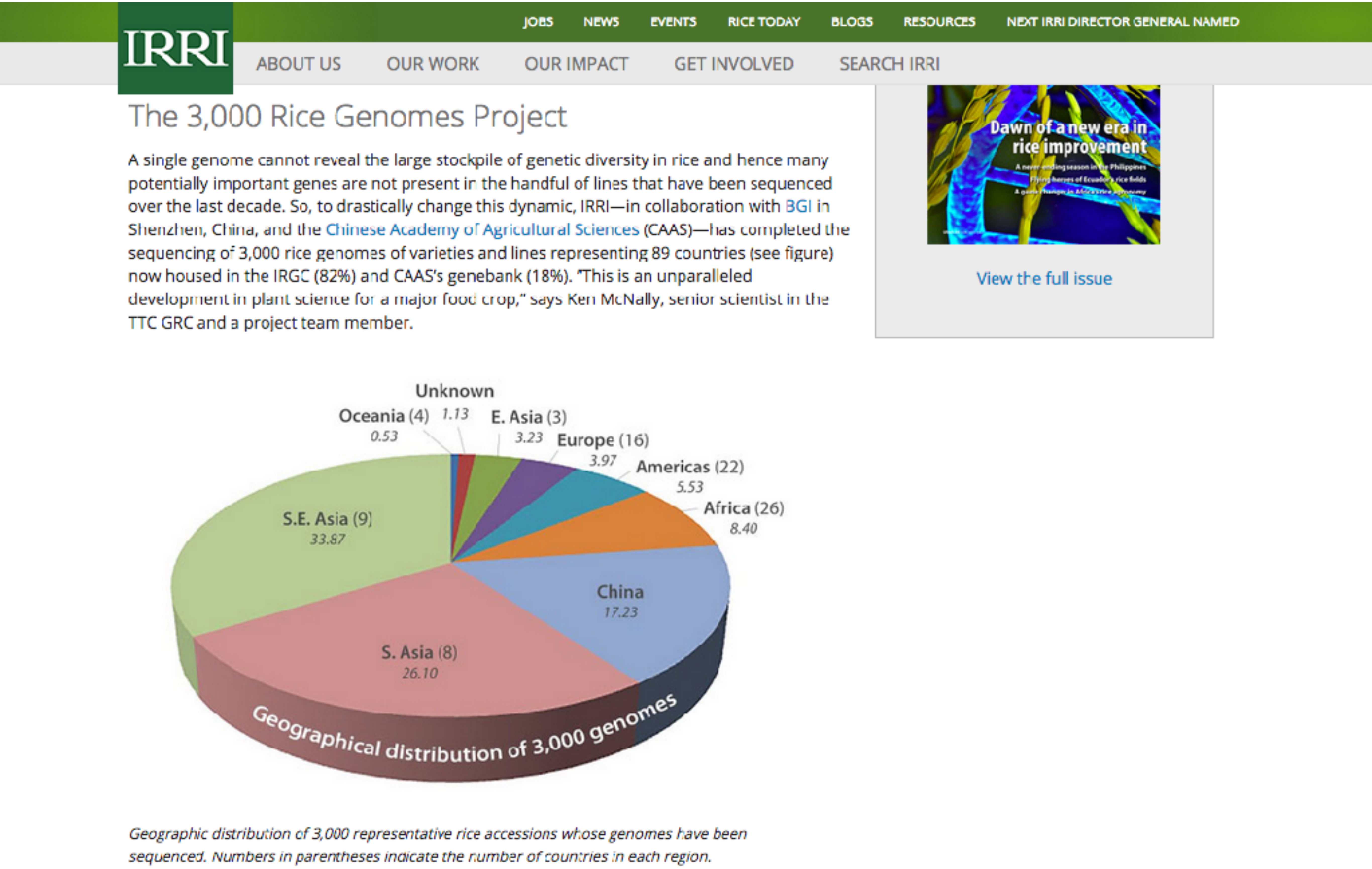
NGS's in 2017

PacBio RSII System

illumina: HiSeq 2500 System

Oxford NANOPORE MinION/SmidgION

3k Rice Genomes Project



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.

www.uk10k.org

UK10K

Rare Genetic Variants in Health and Disease

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.

Wellcome Library, London

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.

Last edited: 10 March 2014

© 2010 Wellcome Trust Sanger Institute

This site is hosted by the **Wellcome Trust Sanger Institute**.

Cookies policy | **Terms & Conditions**

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*

Cyanobacteria
Rhizobia
Plants
Lactic acid bacteria

Kazusa DNA Res. Inst.

NATIONAL INSTITUTE OF GENETICS · JAPAN

Genome projects undergo

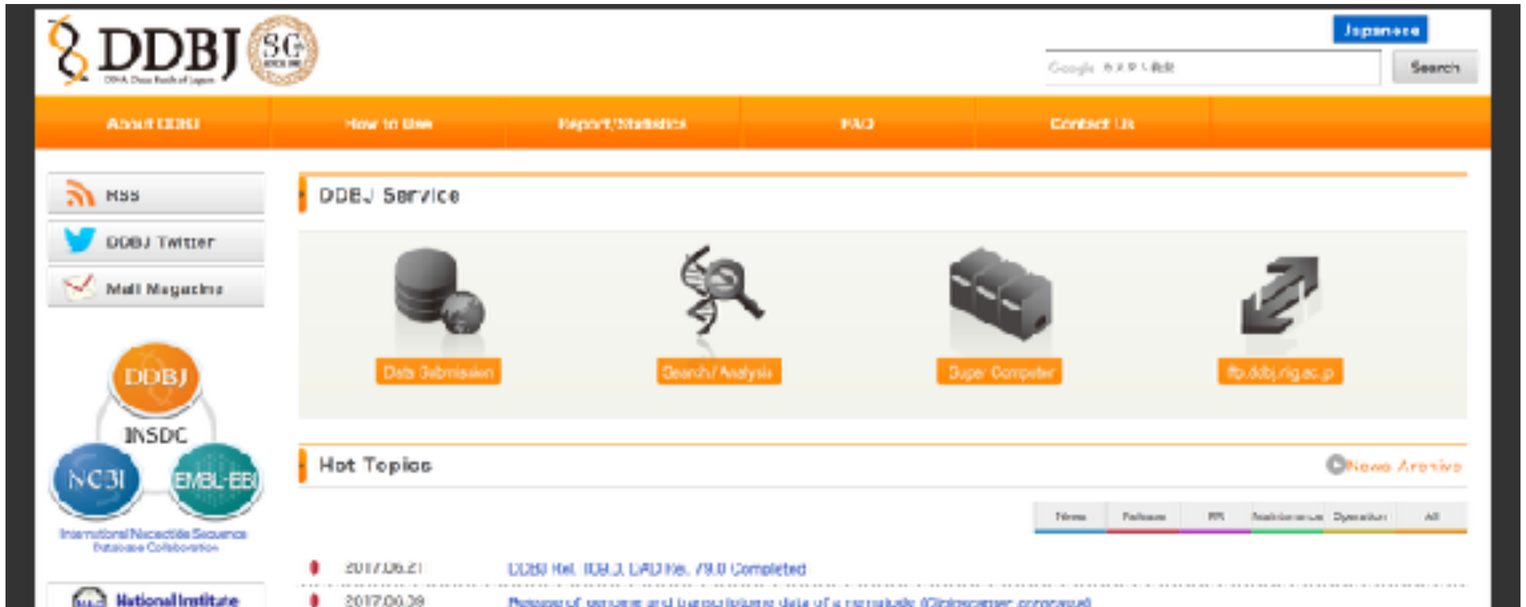
a liverwort, *Marchantia polymorpha*
220 Mb genome
4.4 k scaffolds / N50: 1.3 Mb

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

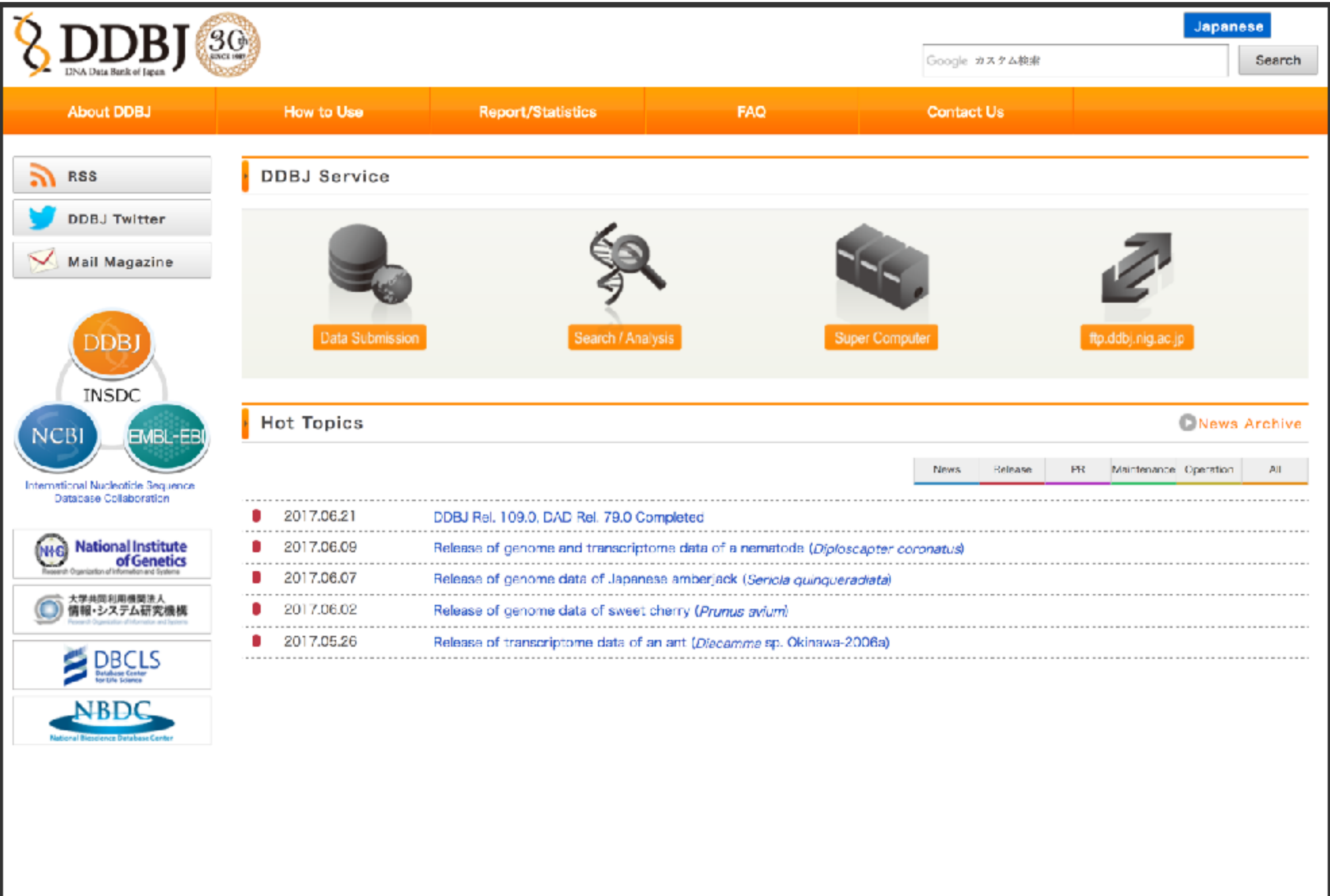
a citrus tree, *Citrus unshiu*
360 Mb genome
21.1 k scaffolds / N50: 385 kb

DDBJ

Mashima J, Kodama Y, Kosuge T, Fujisawa T, Katayama T, Nagasaki H, Okuda Y, Kaminuma E, Ogasawara O, Okubo K, Nakamura Y, Takagi T. (2016) **DNA data bank of Japan (DDBJ) progress report.** *Nucleic Acids Res.*, **44(D1):** D51-57.



DDBJ (http://www.ddbj.nig.ac.jp/)



DDBJ Service

Hot Topics

2017.06.21	DDBJ Rel. 109.0, DAD Rel. 79.0 Completed
2017.06.09	Release of genome and transcriptome data of a nematode (*Diploscapter coronatus*)
2017.06.07	Release of genome data of Japanese amberjack (*Seriola quinqueradiata*)
2017.06.02	Release of genome data of sweet cherry (*Prunus avium*)
2017.05.26	Release of transcriptome data of an ant (*Diacamma* sp. Okinawa-2006a)

INSDC

Cochrane G, Karsch-Mizrachi I, Takagi T, International Nucleotide Sequence Database Collaboration. (2016) **The International Nucleotide Sequence Database Collaboration.** *Nucleic Acids Res.*, **44(D1):** D48-50.

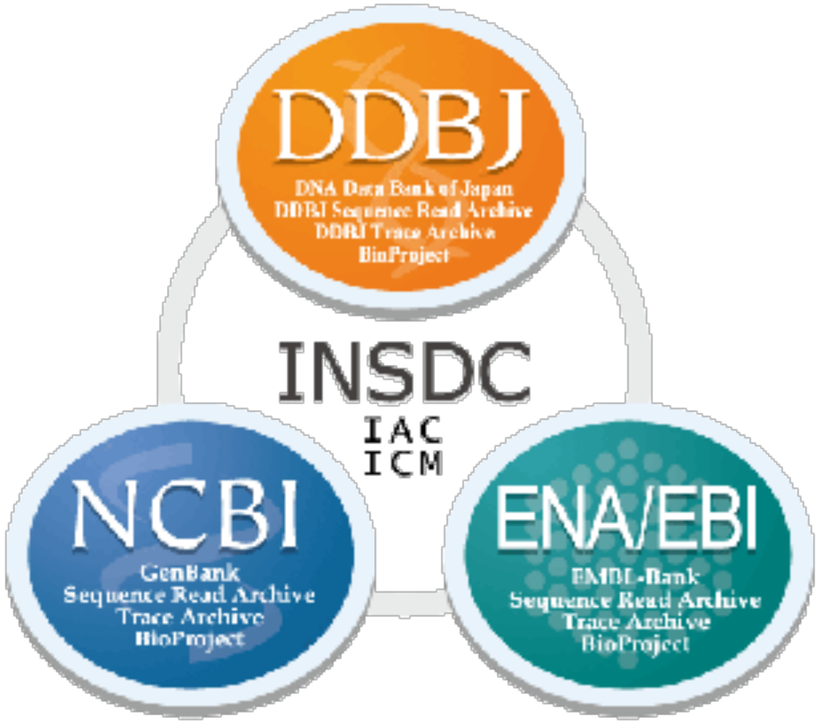
International Nucleotide Sequence Database Collaboration

- The International Nucleotide Sequence Database Collaboration (INSDC) is a long-standing foundational initiative that operates between DDBJ, EMBL-EBI and NCBI. INSDC covers the spectrum of data raw reads, though alignments and assemblies to functional annotation, enriched with contextual information relating to samples and experimental configurations.

Data type	DDBJ	EMBL-EBI	NCBI
Next generation reads	Sequence Read Archive	European Nucleotide Archive (ENA)	Sequence Read Archive
Capillary reads	Trace Archive		Trace Archive
Annotated sequences	DDBJ		GenBank
Samples	BioSample		BioSample
Studies	BioProject		BioProject

DDBJ is a partner of INSDC

International Nucleotide Sequence Databank Collaboration



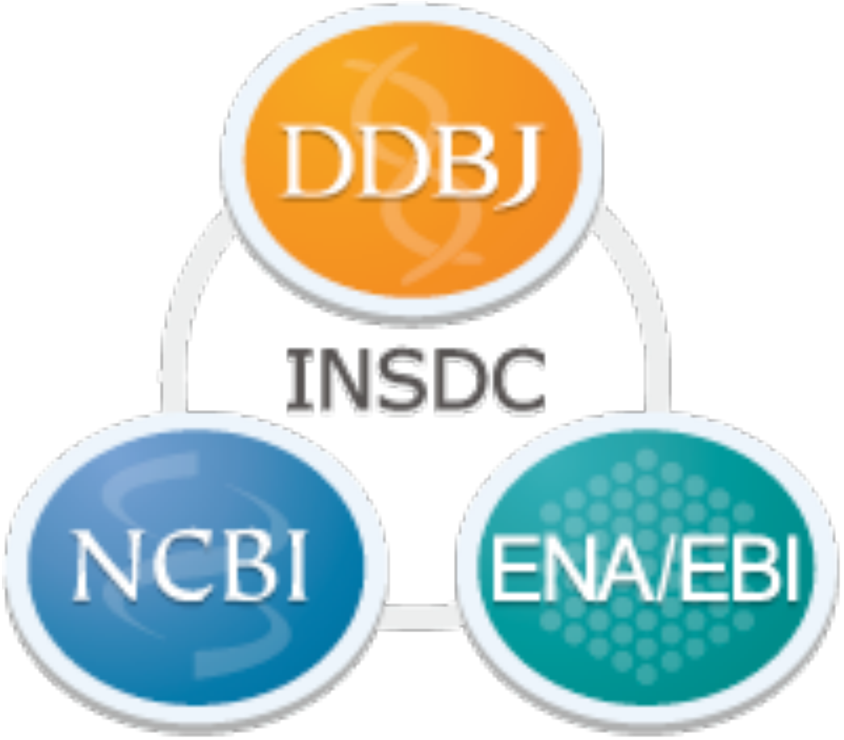
IAC: International Advisory Committee
ICM: International Collaborative Meeting

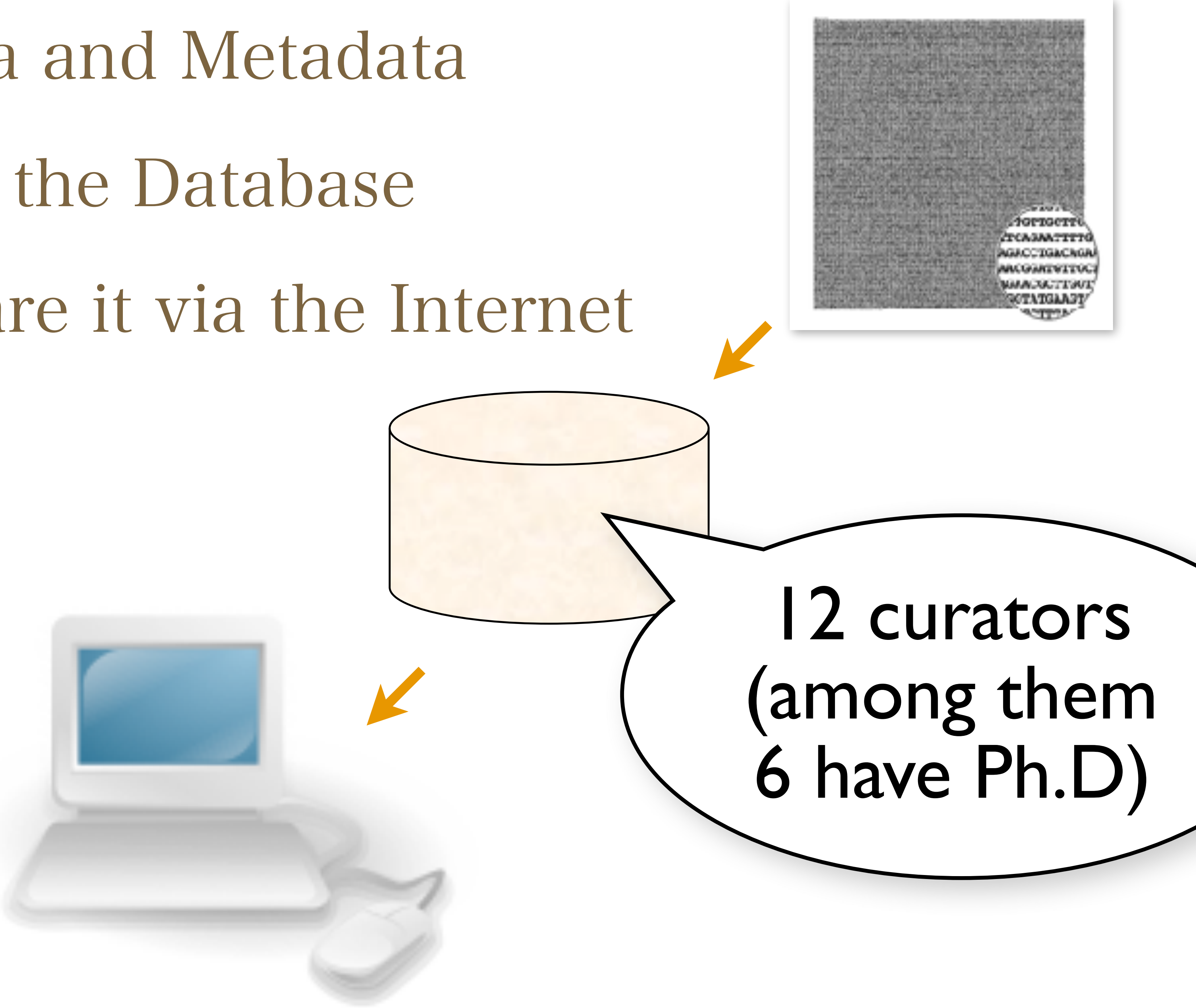
Data type	DDBJ Center	EMBL-EBI	NCBI
Next generation reads	Sequence Read Archive	European Nucleotide Archive (ENA)	Sequence Read Archive
Capillary reads	Trace Archive		Trace Archive
Annotated sequence	DDBJ		GenBank
Samples	BioSample		BioSample
Studies	BioProject		BioProject

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





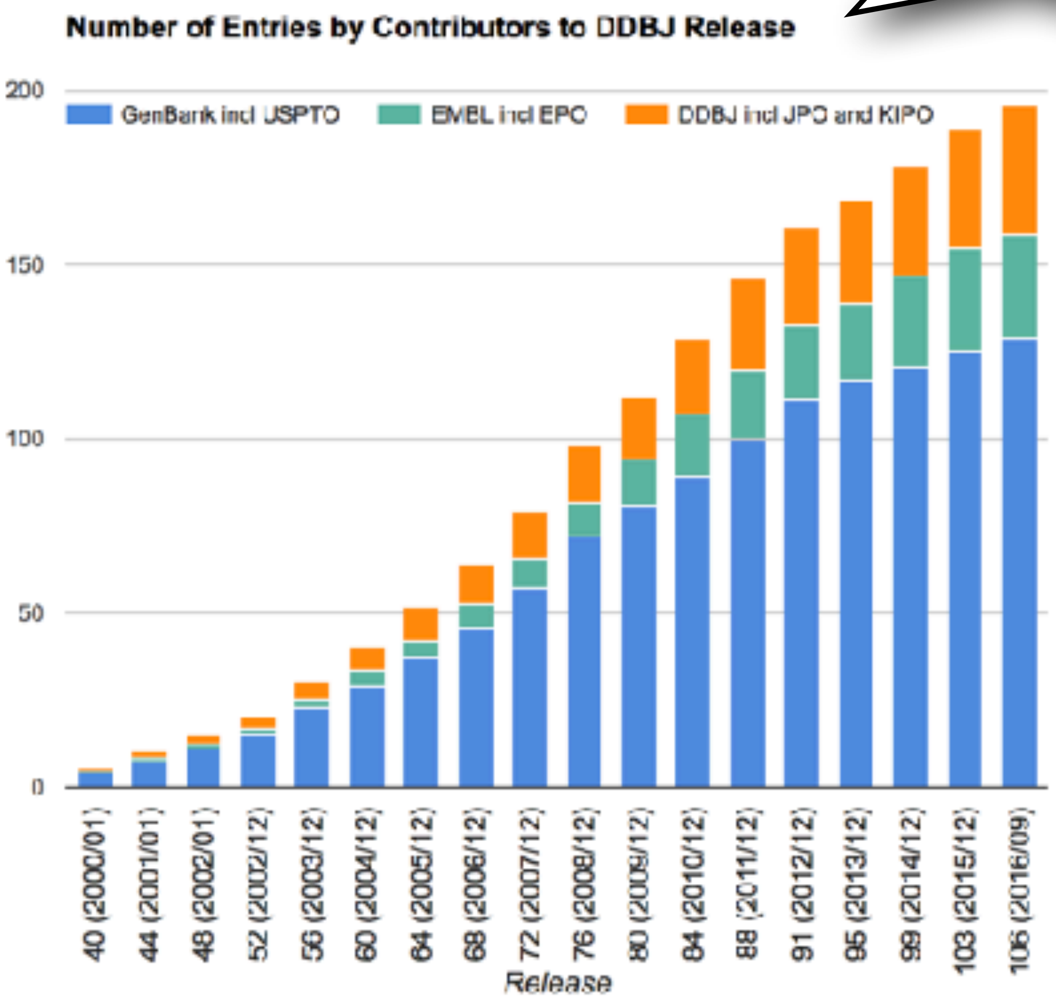
12 curators
(among them
6 have Ph.D)

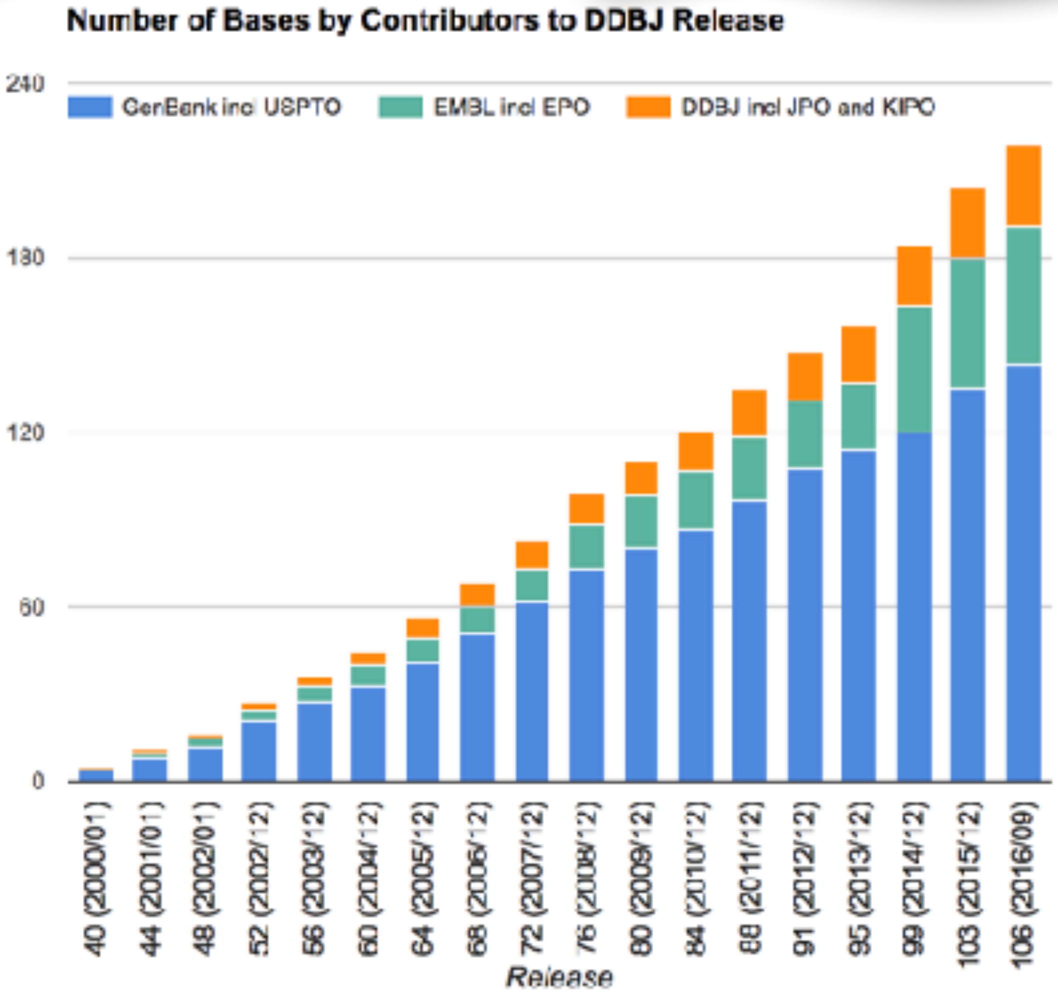
The amount of data continues to grow.

Traditional DDBJ / INSDC

entries:
196 million

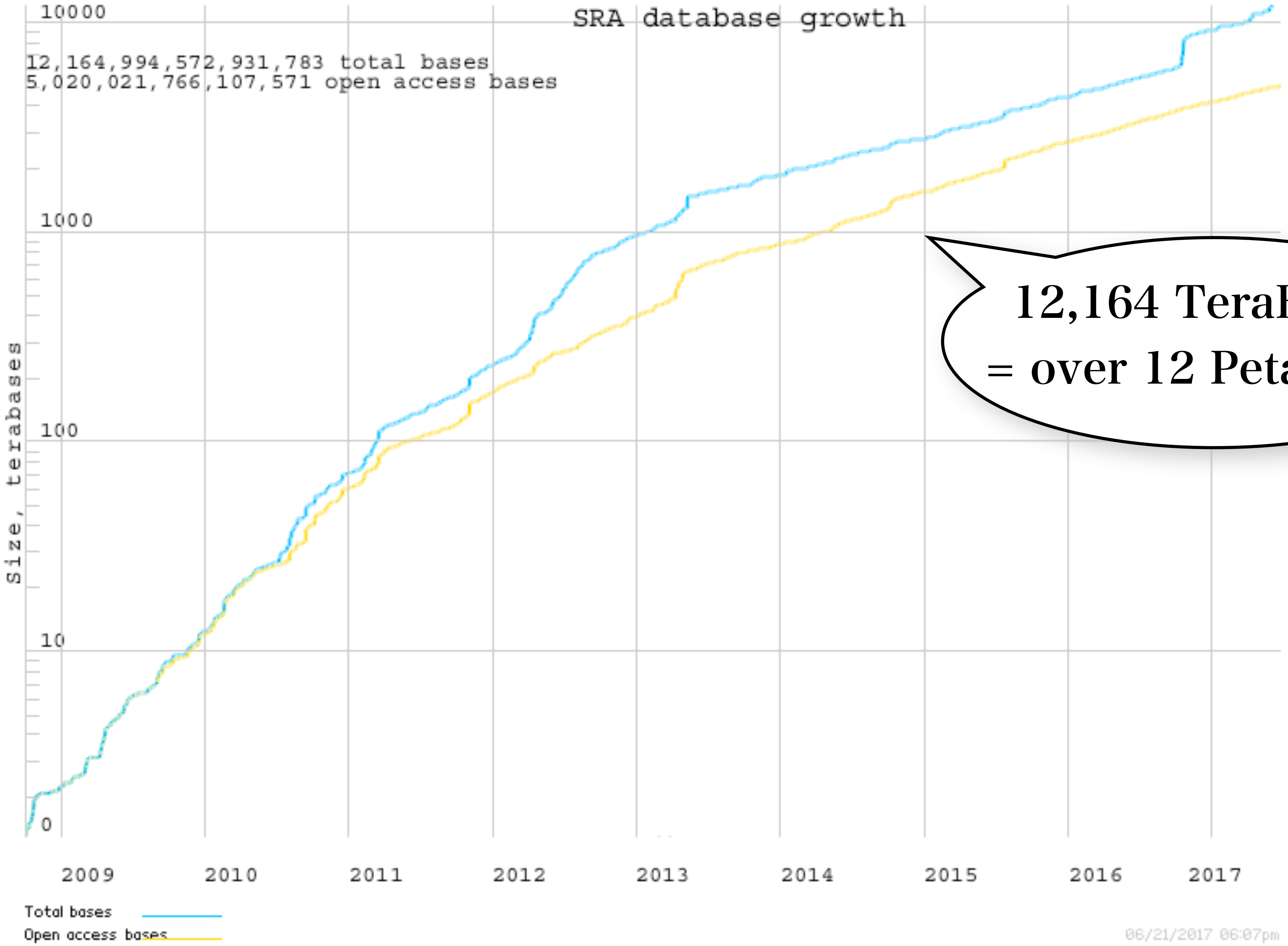
bases:
220 billion


Number of Entries by Contributors to DDBJ Release


Number of Bases by Contributors to DDBJ Release

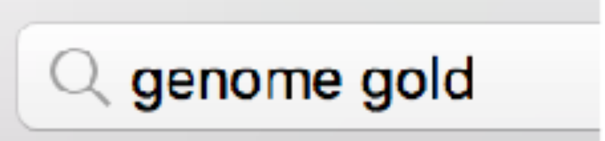
SRA growth (NGS row/bam data amount)

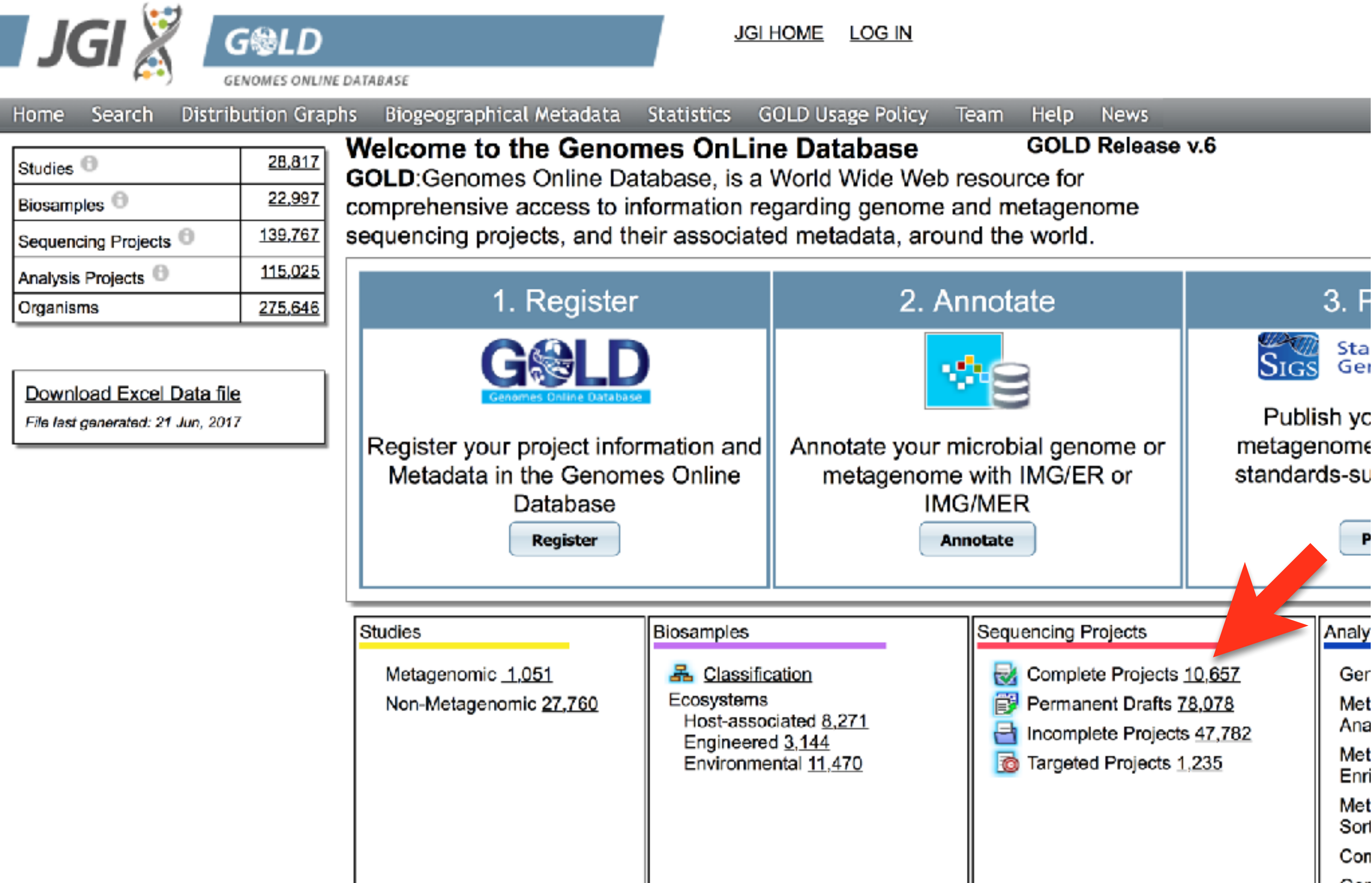
https://trace.ncbi.nlm.nih.gov/Traces/sra



12,164 TeraBases
= over 12 PetaBases

gold.jgi.doe.gov


genome gold



Studies	28,817
Biosamples	22,997
Sequencing Projects	139,767
Analysis Projects	115,025
Organisms	275,646

Download Excel Data file

File last generated: 21 Jun, 2017

Welcome to the Genomes OnLine Database

GOLD Release v.6

GOLD:Genomes Online Database, is a World Wide Web resource for comprehensive access to information regarding genome and metagenome sequencing projects, and their associated metadata, around the world.

1. Register

Register your project information and Metadata in the Genomes Online Database

2. Annotate

Annotate your microbial genome or metagenome with IMG/ER or IMG/MER

3. P

Publish yc metagenome standards-su

Studies

Metagenomic 1,051
Non-Metagenomic 27,760

Biosamples

Classification
Ecosystems
  Host-associated 8,271
  Engineered 3,144
  Environmental 11,470

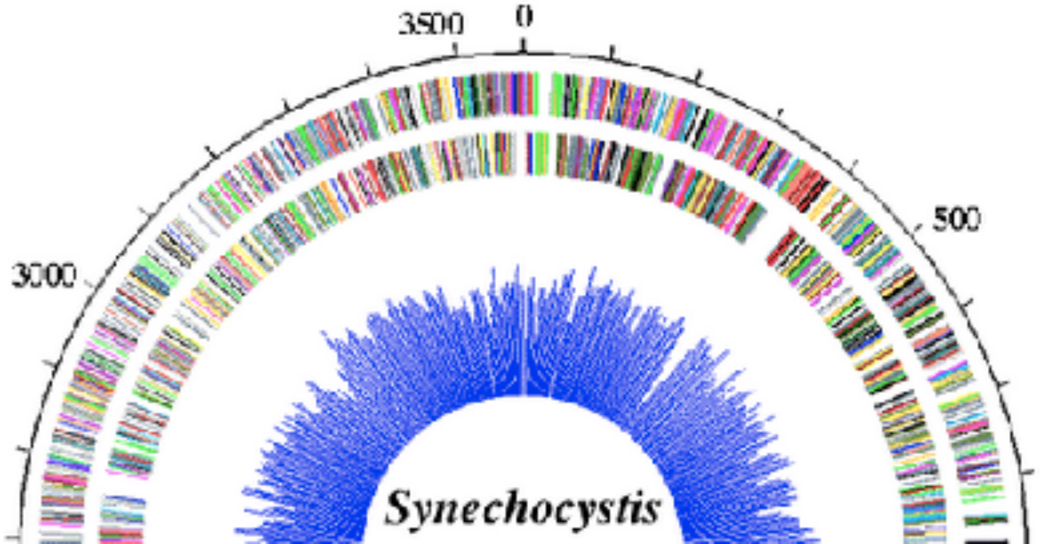
Sequencing Projects

Complete Projects 10,657
Permanent Drafts 78,078
Incomplete Projects 47,782
Targeted Projects 1,235

Analy

CyanoBase

Nakamura, Y., Kaneko, T., Hirosawa, M., Miyajima, N. and Tabata, S. (1998) CyanoBase, a www database containing the complete nucleotide sequence of the genome of *Synechocystis* sp. strain PCC6803. *Nucleic Acids Res*. **26,** 63-67.



Number of species in CyanoBase

1996: 1

2007: 5

2010: 39

2016: 376

The budget continues to shrink.

It costs.

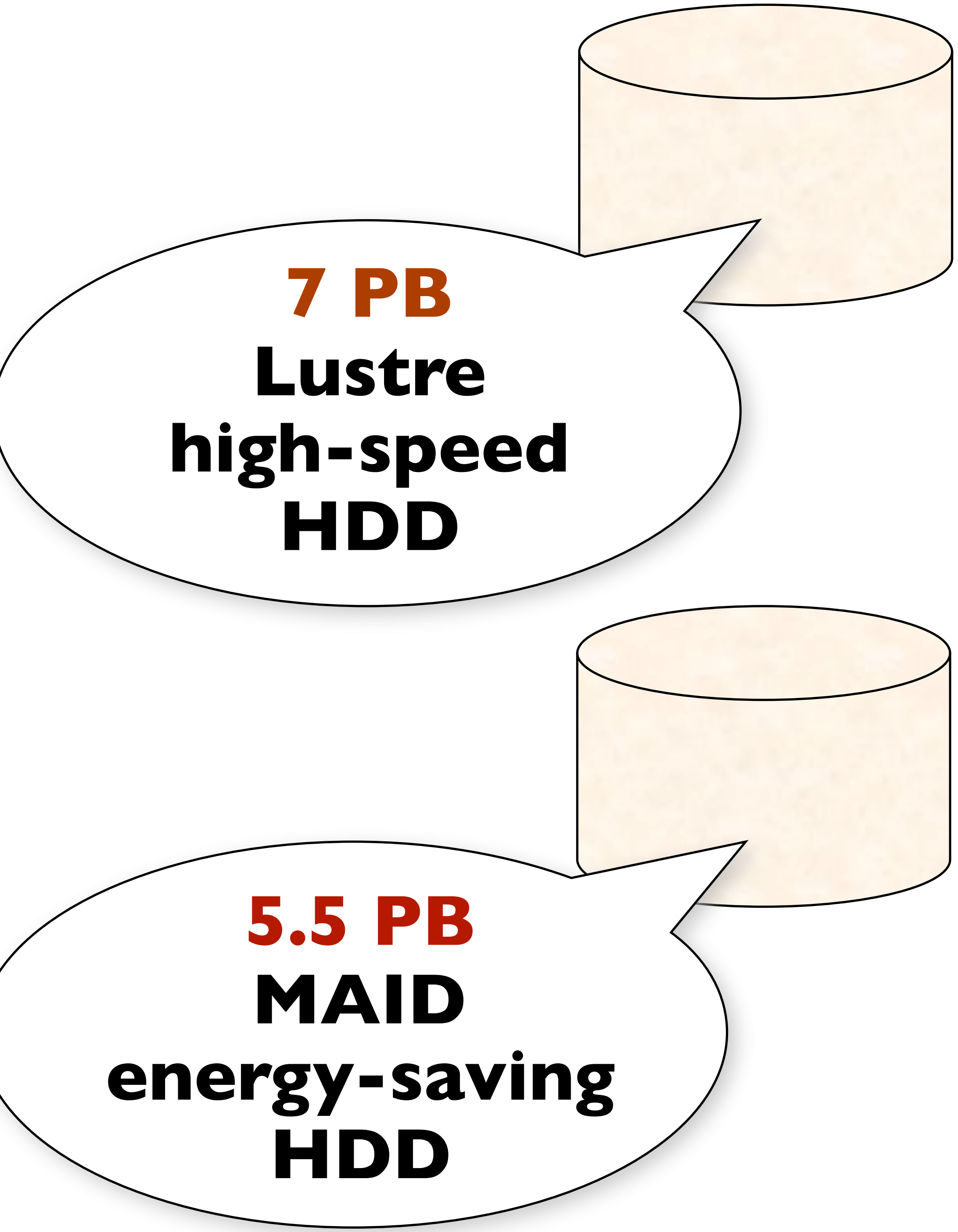
DDBJ (from Release note 92) 44
Jun Mashima, Hideo Aono, Yuji Ashizawa, Yukino Dobashi, Mayumi Ejima, Masahiro Fujimoto, Asami Fukuda, Tomohiro Hirai, Fumie Hirata, Naofumi Ishikawa, Toshikazu Katsumata, Chiharu Kawagoe, Shingo Kawahara, Yuichi Kodama, Junko Kohira, Takehide Kosuge, Kyungbum Lee, Mika Maki, Kimiko Mimura, Takeshi Moriyama, Yoshihisa Munakata, Naoko Murakata, Keiichi Nagai, Toshihisa Okido, Yoshihiro Okuda, Katsunaga Sakai, Makoto Sato, Yoshihiro Serizawa, Aimi Shiida, Yukie Shinyama, Rie Sugita, Kimiko Suzuki, Daisuke Takagi, Daisuke Takai, Haru Tsutsui, Koji Watanabe, Tomohiko Yasuda, Shigeru Yatsuzuka, Emi Yokoyama, Eli Kaminuma, Osamu Ogasawara, Kosaku Okubo, Toshihisa Takagi, and Yasukazu Nakamura

ENA (from Release note 115) 27
Blaise Alako, Clara Amid, Lawrence Bower, Ana Cerdeno-Taraga, Iain Cleland, Richard Gibson, Neil Goodgame, Petra ten Hoopen, Mikyung Jang, Simon Kay, Rasko Leinonen, Xin Liu, Arnaud Oisel, Rodrigo Lopez, Hamish McWilliam, Nima Pakseresht, Sheila Plaister, Rajesh Radhakrishnan, Kethy Reddy, Stephane Riviere, Marc Rossello, Nicole Silvester, Dmitriy Smirnov, Ana Luisa Toribio, Daniel Vaughan, Vadim Zalunin and Guy Cochrane

GenBank (from Release note 195) 68
Mark Cavanaugh, Ilene Mizrachi, Yiming Bao, Michael Baxter, Lori Black, Larissa Brown, Vincent Calhoun, Larry Chlumsky, Karen Clark, Jianli Dai, Michel Eschenbrenner, Irene Fang, Michael Fetchko, Linda Frisse, Andrea Gocke, Anjanette Johnston, Mark Landree, Jason Lowry, Suzanne Mate, Richard McVeigh, DeAnne Olsen Cravaritis, Leigh Riley, Susan Schafer, Beverly Underwood, Melissa Wright, Linda Yankie, Serge Bazhin, Evgueni Belyi, Colleen Bollin, Mark Cavanaugh, Yoon Choi, Ilya Dondoshansky, J. Bradley Holmes, WonHee Jang, Jonathan Kans, Leonid Khotomliansky, Michael Kimelman, Michael Kornbluh, Jim Ostell, Denis Sinyakov, Karl Sirotkin, Vladimir Soussov, Elena Starchenko, Hanzhen Sun, Tatiana Tatusova, Lukas Wagner, Eugene Yaschenko, Sergey Zhdanov, Slava Khotomliansky, Igor Lozitskiy, Craig Oakley, Eugene Semenov, Ben Slade, Constantin Vasilyev, Peter Cooper, Hanguan Liu, Wayne Matten, Scott McGinnis, Rana Morris, Steve Pechous, Monica Romiti, Eric Sayers, Tao Tao, Majda Valjavec-Gratian and David Lipman

NIG supercomputer (2014.3 ver.)

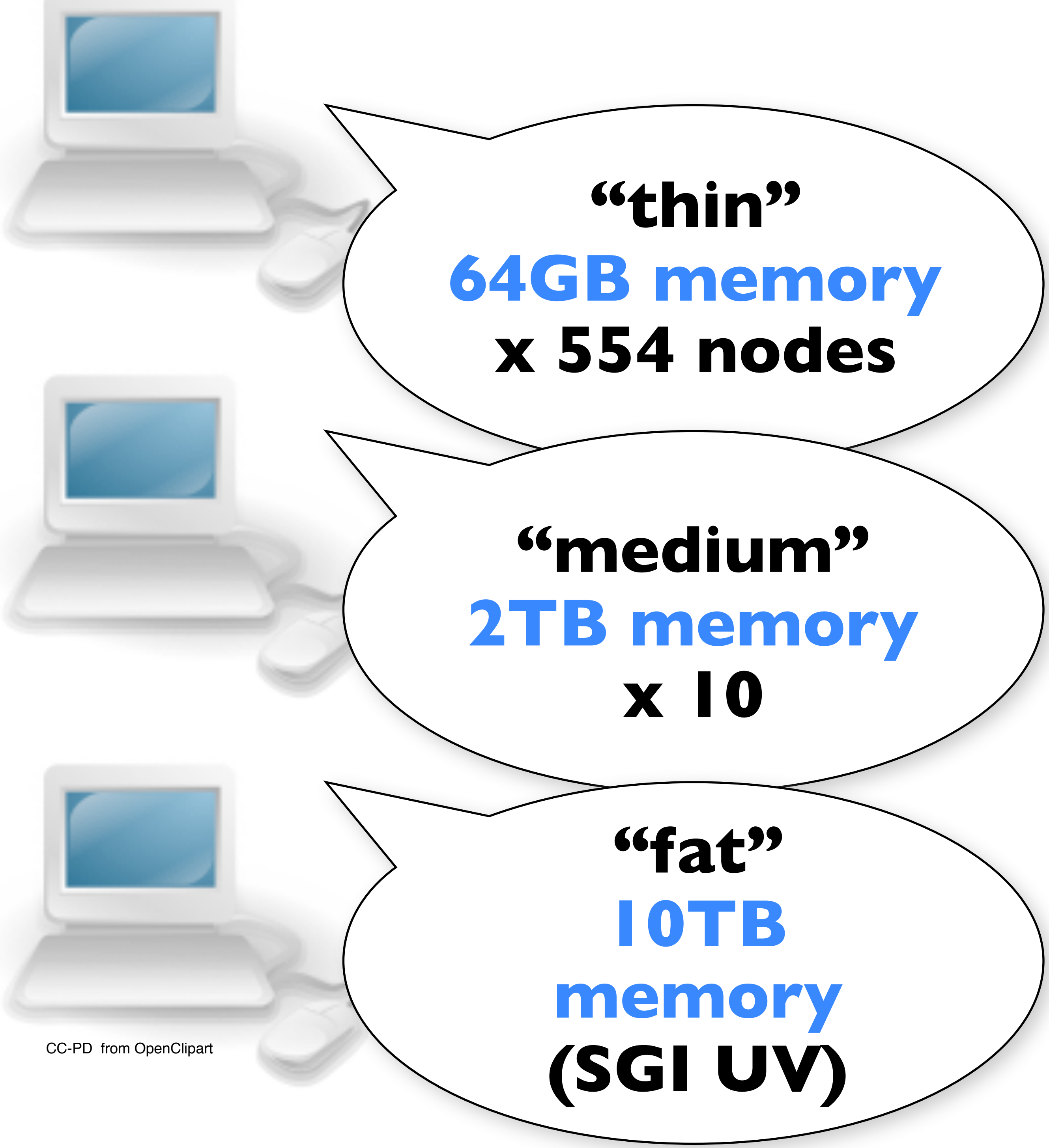
12.5 PB



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)

CC-PD from OpenClipart

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

News

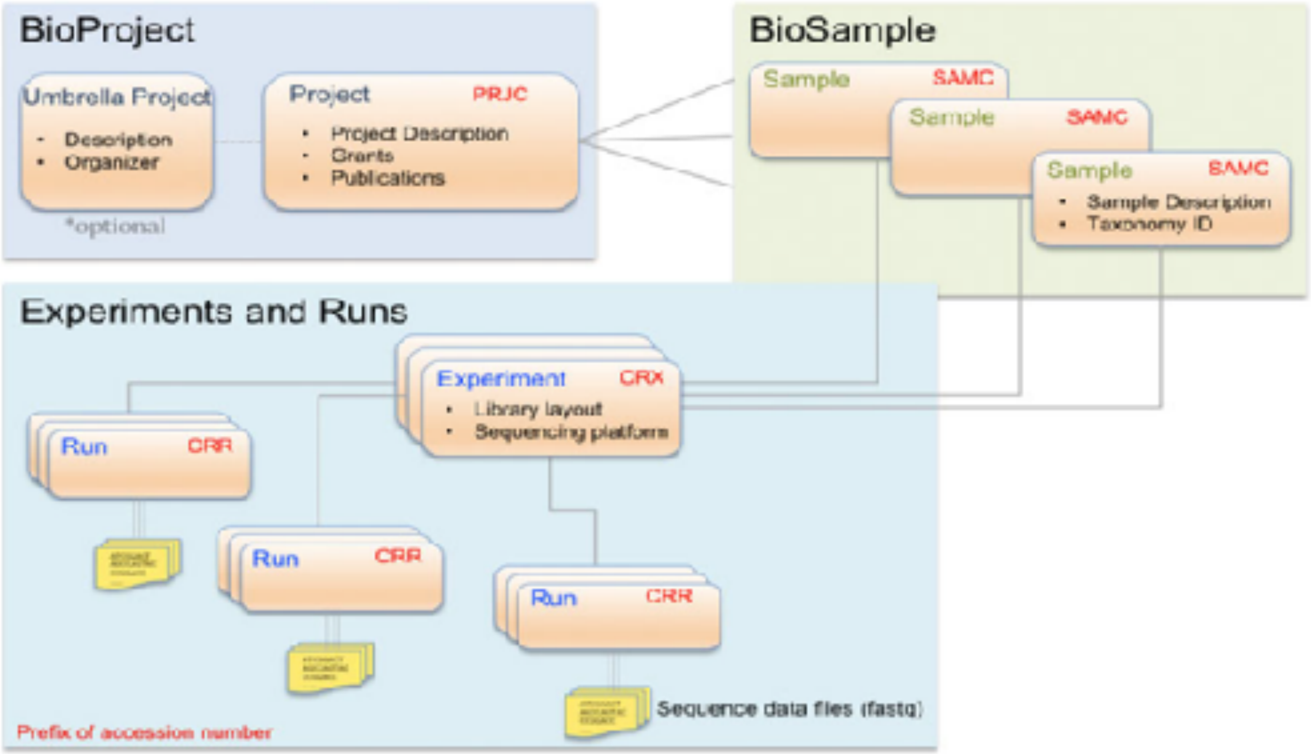
- Data deposited to GSA has been reported by a paper published in Nature Communications. (2017-06-20)
- Data deposited to GSA has been reported by a paper published in Genome Biology. (2017-06-16)
- Data deposited to GSA has been reported by a paper published in Cell Research. (2017-05-02)
- Data deposited 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 Genome Research. (2016-10-21)
- 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. (2016-08-20)
- Data deposited to GSA has been reported by a paper published in Stem Cell Reports. (2016-07-14)
- Data deposited to GSA has been reported by a paper published in Journal of Cell Science. (2016-05-17)
- Data deposited to GSA has been reported by a paper published in PNAS. (2015-11-10)

Latest Released Projects

Accession	Description
PRJCA000445 (2017-05-11)	Expression profiling of SCC15 cells treated by PAM
PRJCA000313 (2017-05-08)	Shuhui498 rice genome sequencing and assembly

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 revelant 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.

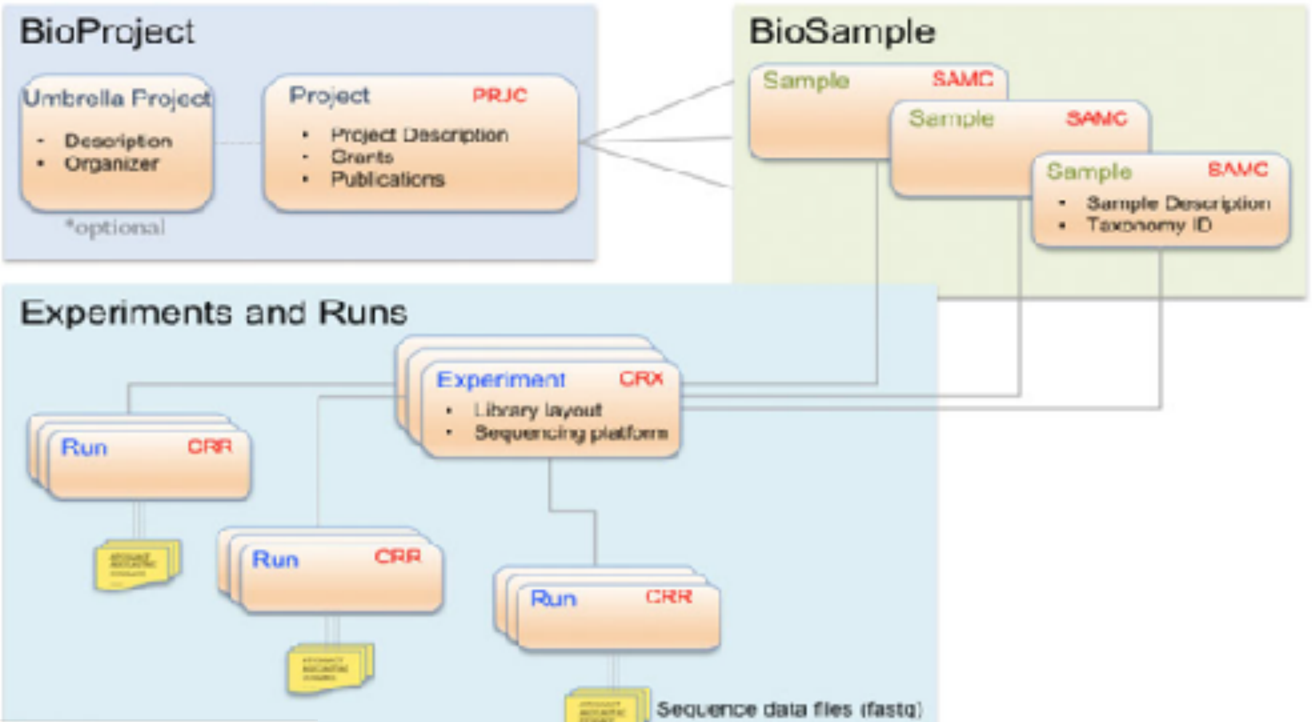
Run describes technical batch related files that belong to a specific Experiment. Each Run may have multiple files.

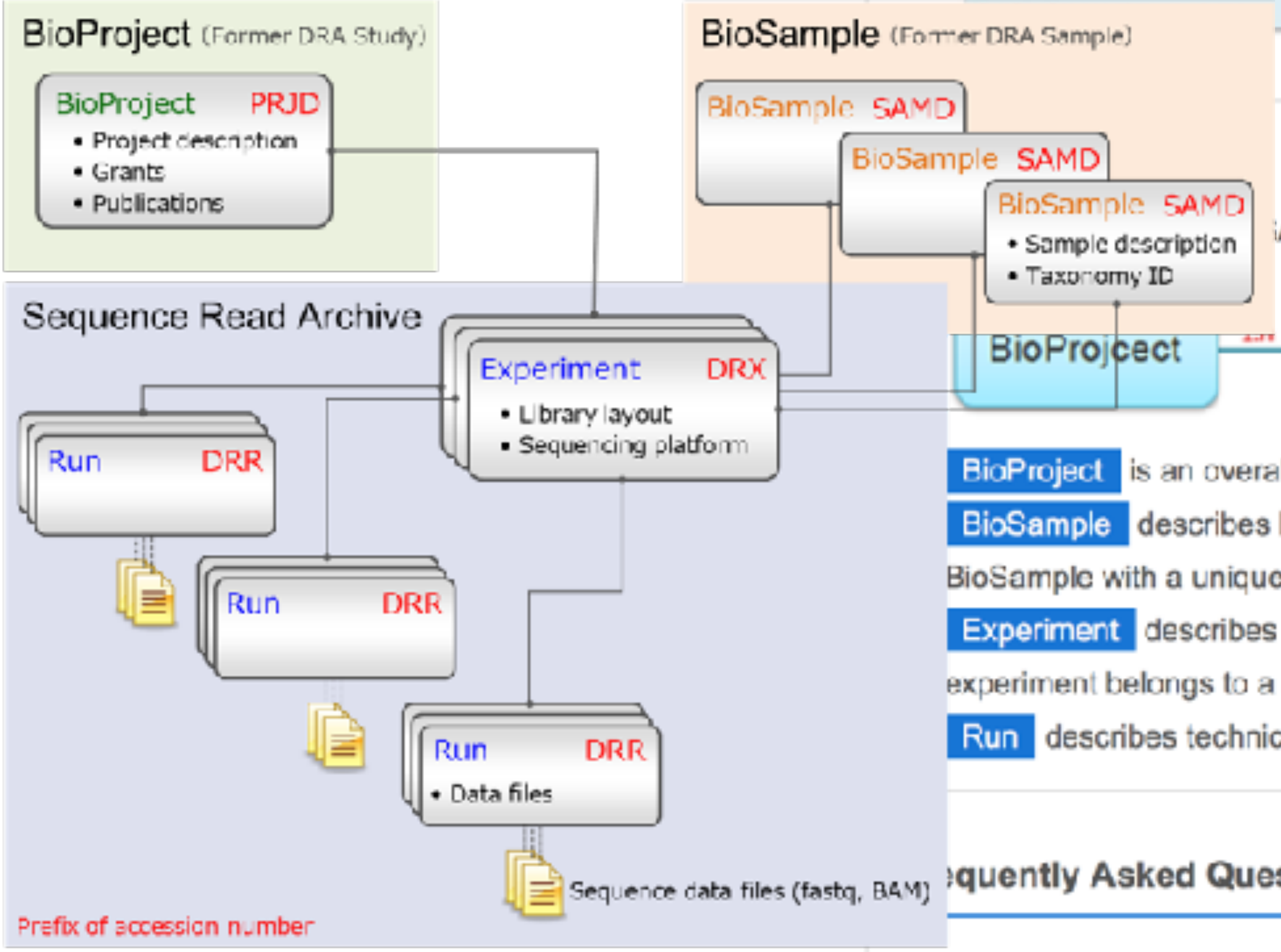
Frequently Asked Questions

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

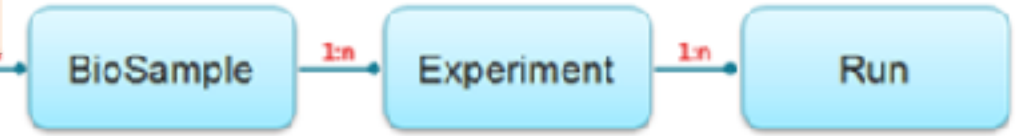
GSA Data Model

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 revelant projects supported by a collaborative grant or mega grant, e.g., 1000 Human Genomes Project, Dog 10K Genomes Project.





A 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.

quently Asked Questions

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

Accession	Description
PRJCA000445 (2017-05-11)	Expression profiling of SCC15 cells treated by PAM
PRJCA000313 (2017-05-08)	Shuhui498 rice genome sequencing and assembly
PRJCA000307 (2017-05-01)	MTBC
PRJCA000315 (2017-04-30)	RNA m5C sequencing in human HeLa cells and mouse tissues
PRJCA000416 (2017-04-11)	A Comparative Transcriptomic Analysis of Uveal Melanoma...

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 PRJCAxxxxxx, PRJCAyyyyyy that are publicly accessible at ***http://bigd.big.ac.cn/gsa***.

Please cite the following required publications.

- **GSA: Genome Sequence Archive.** *Genomics, Proteomics & Bioinformatics* 2017, 15(1): 14-18. doi:10.1016/j.gpb.2017.01.001.
- **The BIG Data Center: from deposition to integration to translation.** *Nucleic Acids Res* 2017, 45(D1): D18-D24. **[PMID=27899658]**

? Help & Support

If you have any question or would like to give us any suggestion/comment or report a bug, please feel free to contact us.

- Email: gsa@big.ac.cn
- QQ group: 548170081

We highly appreciate your comments and suggestions for further improving its functionalities and providing better services.

Partner Databases

- DDBJ-DRA
- EBI-ENA
- NCBI-SRA

BIGD
About BIGD
Mission
Advisory Board
Organizational Structure
History
Annual Report

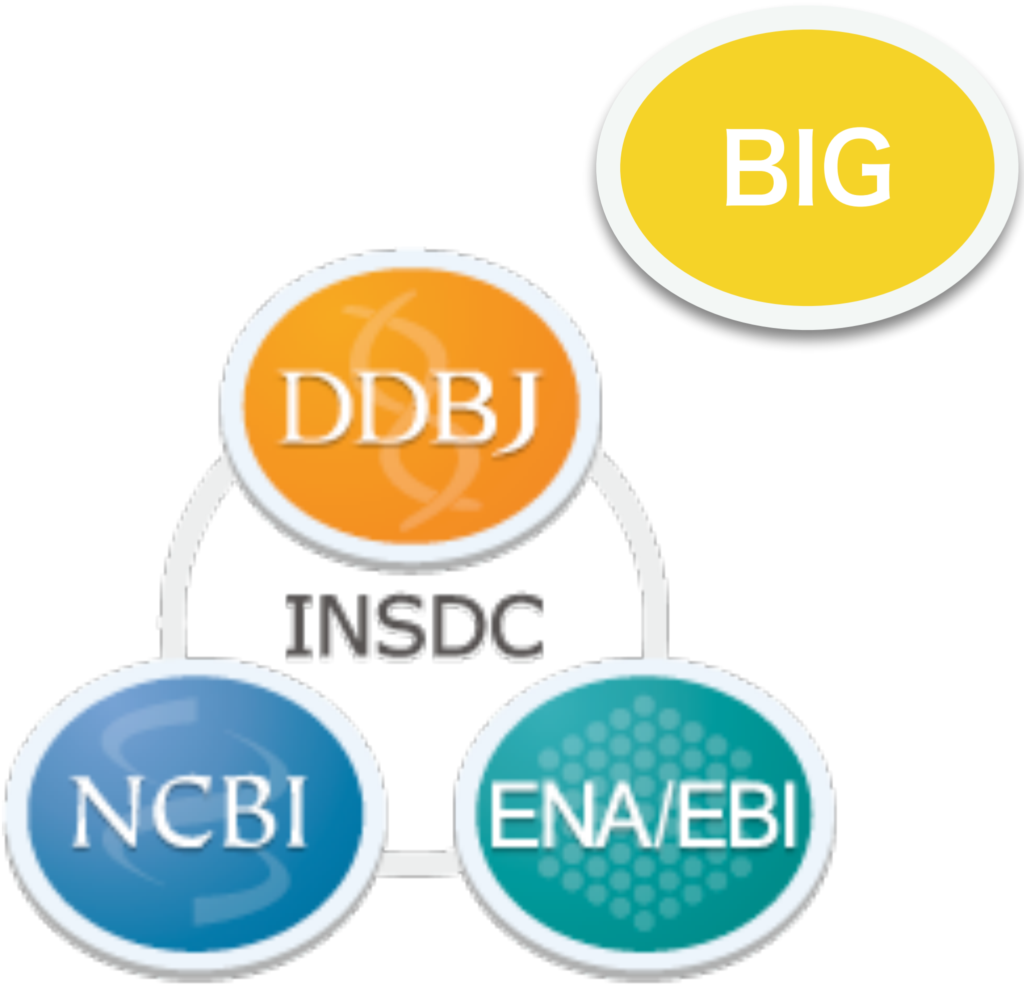
Resources
Databases
Tools

Research
Topics & Projects
Publications

Featured
Database Commons
GSA
IC4R
MethBank
VCGDB

Conferences & Training
Conferences
Training
Calendar

About Us
People
Collaborations
Funding
Contact Us



BIG

DDBJ

INSDC

NCBI

ENA/EBI

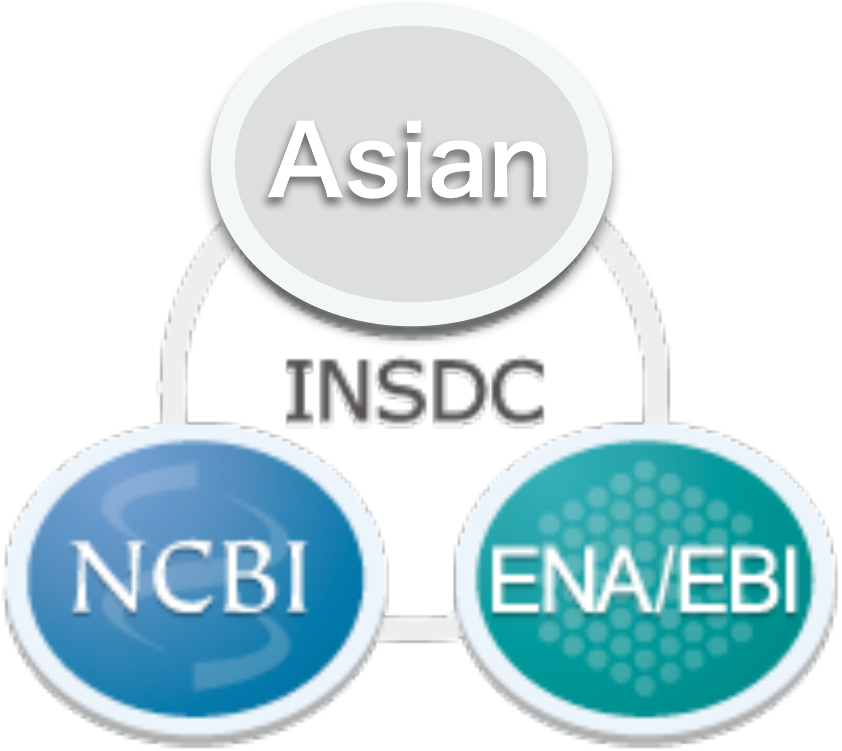
Asian

DDBJ

INSDC

NCBI

ENA/EBI


Asian
INSDC
NCBI
ENA/EBI

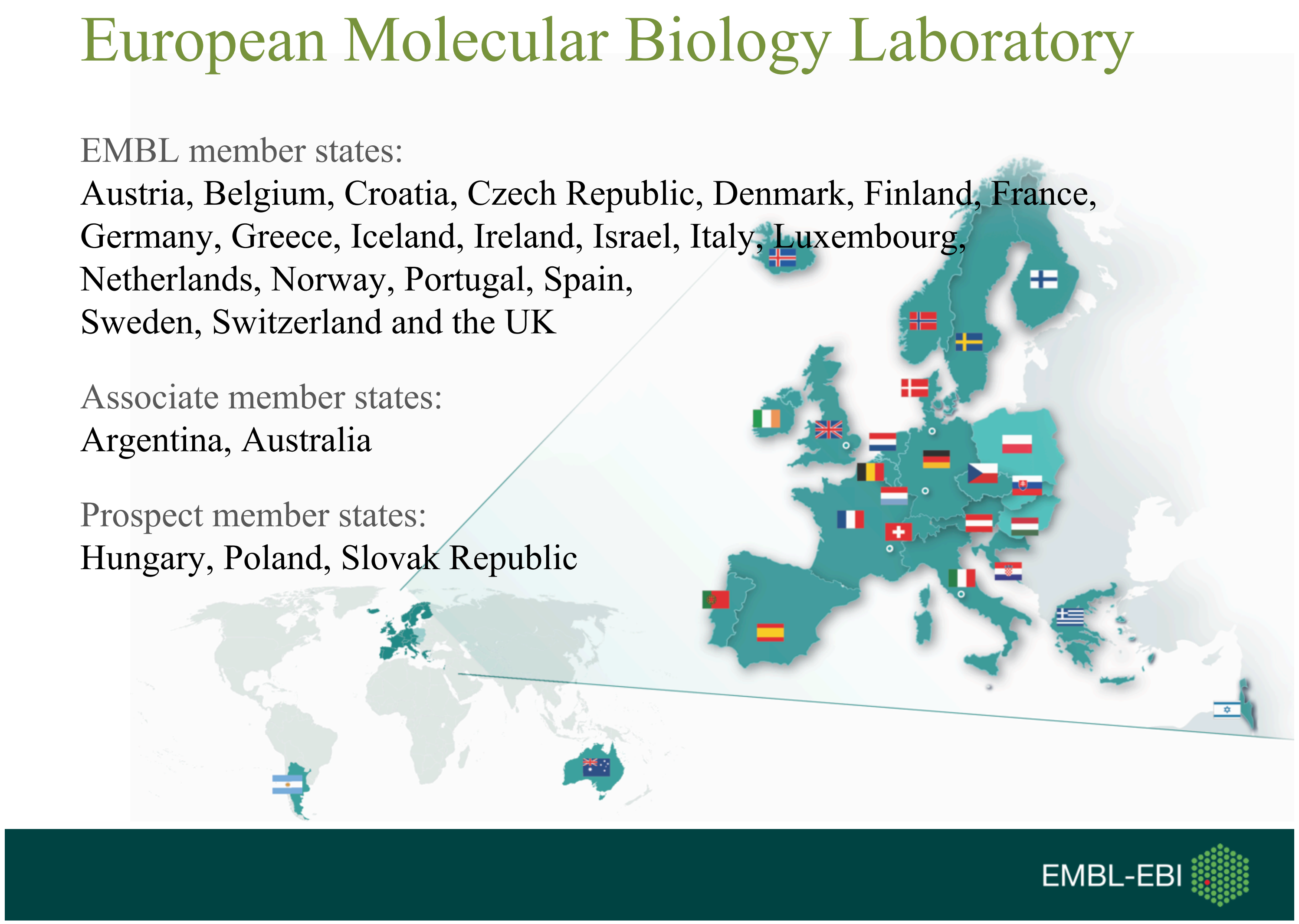
Hints:

TOWARDS COORDINATED INTERNATIONAL SUPPORT OF CORE DATA RESOURCES FOR THE LIFE SCIENCES

W. Anderson[1], R. Apweiler[2], A. Bateman[2], G.A. Bauer[1], H. Berman[3], J.A. Blake[9], N. Blomberg[4], S.K. Burley[5], G. Cochrane[2], V. Di Francesco[6], T. Donohue[21], C. Durinx[10], A. Game[23], E.D. Green[6], T. Gojobori[14], P. Goodhand[15], A. Hamosh[16], H. Hermjakob[2], M. Kanehisa[22], R. Kiley[17], J. McEntyre[2], R. McKibbin[18], S. Miyano[19], B. Pauly[1], N. Perrimon[12], M.A. Ragan[13], G. Richards[1], Y-Y. Teo[20], M. Westerfield[11], E. Westhof[7], and P.F. Lasko[8]

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.***

Lessons from EMBL / Elixir



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

EMBL-EBI