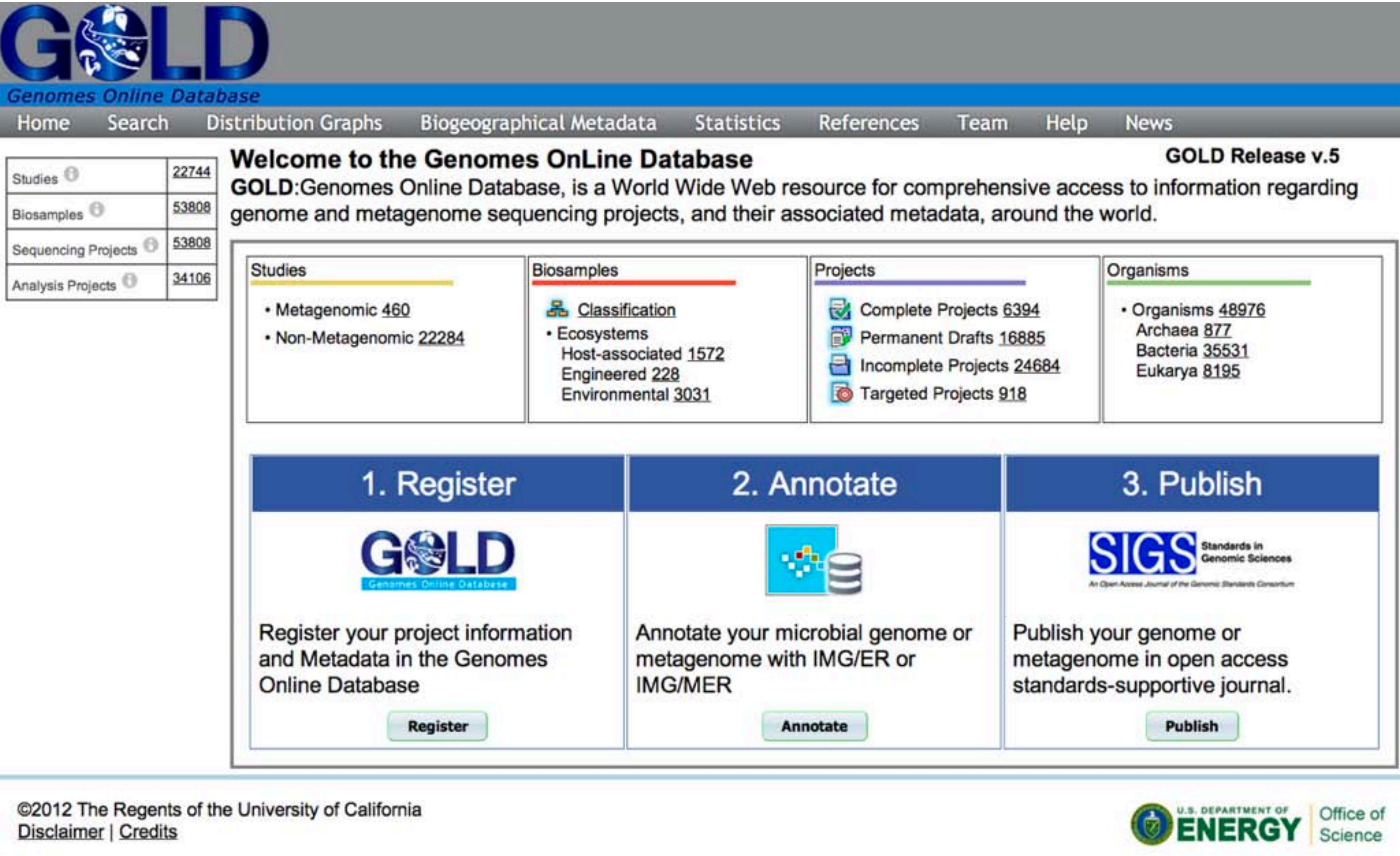
Introduction of DDBJ and microbe genome analysis pipelines

Genome Informatics Laboratory
National Institute of Genetics, JAPAN
Yaz Nakamura

Genome informatics helps your life

http://www.genomesonline.org/



Studies	22744
Biosamples	53808
Sequencing Projects	53808
Analysis Projects	34106

Welcome to the Genomes OnLine Database

GOLD Release v.5

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.

Studies
- Metagenomic 460
- Non-Metagenomic 22284

Biosamples

Classification
- Ecosystems
  Host-associated 1572
  Engineered 228
  Environmental 3031

Projects

Complete Projects 6394
Permanent Drafts 16885
Incomplete Projects 24684
Targeted Projects 918

Organisms
- Organisms 48976
  Archaea 877
  Bacteria 35531
  Eukarya 8195

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. Publish

Publish your genome or metagenome in open access standards-supportive journal.

What kind of skills are necessary?

[needed]

Operating a UNIX system.

Using databases effectively.

Making some easy scripts (to make your life easy).

[not needed]

High-level programming skill

(e.g. operating system construction)

Who am I?

I worked for Kazusa DNA Research Institute (1996-2008).
Analyses and DB construction for plant and plant-related bacteria genomes.
From 2009, I work for DDBJ, NIG.

nature

Arabidopsis thaliana genome sequence

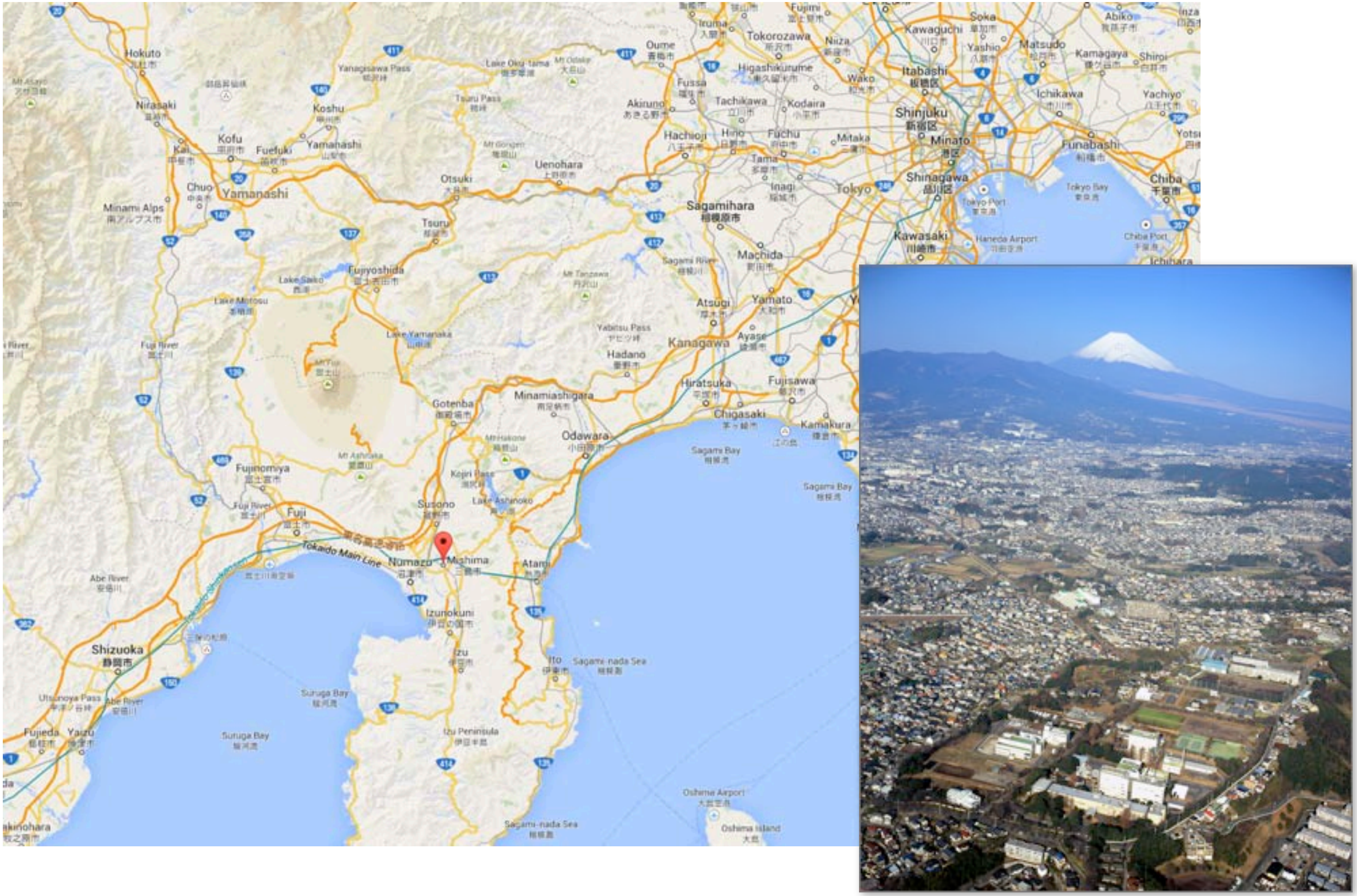
The Arabidopsis Genome Initiative (2000)
Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature*, 408, 796-815.

My team annotated 1/4 (27 Mb, 6200 genes) regions of *Arabidopsis* genome.

http://genome.microbedb.jp/cyanobase/
http://genome.microbedb.jp/rhizobase/

World central genome DB’s for Cyanobacteria and Rhizobia (plant-related bacteria)

Location of DDBJ/nig (Mishima city)



Genome Informatics Laboratory

Yaz Nakamura, Eli Kaminuma, Hideki Nagasaki, Takako Mochizuki, Takatomo Fujisawa, Naoko Iida, Yasuhiro Tanizawa, Kaori Fujieda, Naoko Murakata, Naoko Sakamoto

My Lab’s projects

CyanoBacteria, Rhizobia and Streptomyces DB’s and TogoAnnotation

with DBCLS and Onishi lab. at Tokyo Univ. as an activity of integrated DB project

Citrus species’ sequencing project

with National Institute of Fruit Tree Science and Fujiyama lab as a part of TRIC project

A liverwort (a moss) sequencing project

with Kohchi lab at Kyoto Univ. as a part of Genome Science project

A *Charophyceae* (an algae) annotation project

with Ohta lab at Titech

A Rubber tree sequencing project

with Bridgestone co ltd.

DNApod : DNA Polymorphism annOtation Database

as an activity of TRIC project

Supported by Grant-in-aids “Genome Science” by MEXT, Integrated Lifescience Database Project by JST and Transdisciplinary Research Integration Center (TRIC) project by ROIS

DDBJ

Kosuge T, Mashima J, Kodama Y, Fujisawa T, Kaminuma E, Ogasawara O, Okubo K, Takagi T, Nakamura Y. (2014) DDBJ progress report: a new submission system for leading to a correct annotation. *Nucleic Acids Res.*, **42(1):** D44-49. doi: 10.1093/nar/gkt1066.

INSDC

Nakamura Y, Cochrane G, Karsch-Mizrachi I on behalf of the International Nucleotide Sequence Database Collaboration. (2013) The International Nucleotide Sequence Database Collaboration. *Nucleic Acids Res.*, **41 (D1),** D21-D24. Epub 2012 Nov 24.

DDBJ is a member of INSDC

International Nucleotide Sequence Databank Collaboration

- USA: GenBank/NCBI
- EU: ENA/EBI
- Japan: DDBJ/NIG

IAC: International Advisory Comittee
ICM: International Collaborative Meeting

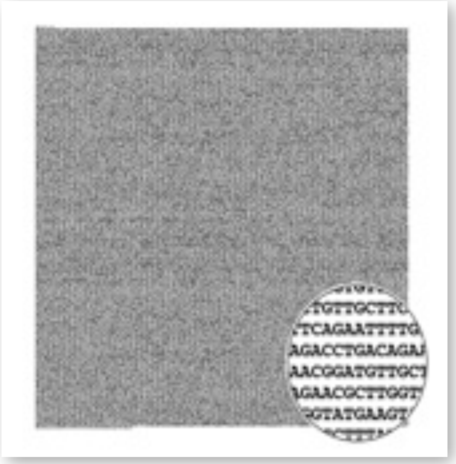
DDBJ (from Release note 92)
Jun Mashima, Hideo Aono, Yuji Ashizawa, Yukino Dobashi, Mayumi Ejima, Masahiro Fujimoto, Asami Fukuda, Tomohiro Hirai, Fumie Hirata, Naofumi Ishikawa, Toshikazu Katsumata, Chiharu Kawagoe, Shingo Kawahara, Yuichi Kodama, Junko Kohira, Takehide Kosuge, Kyungbum Lee, Mika Maki, 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

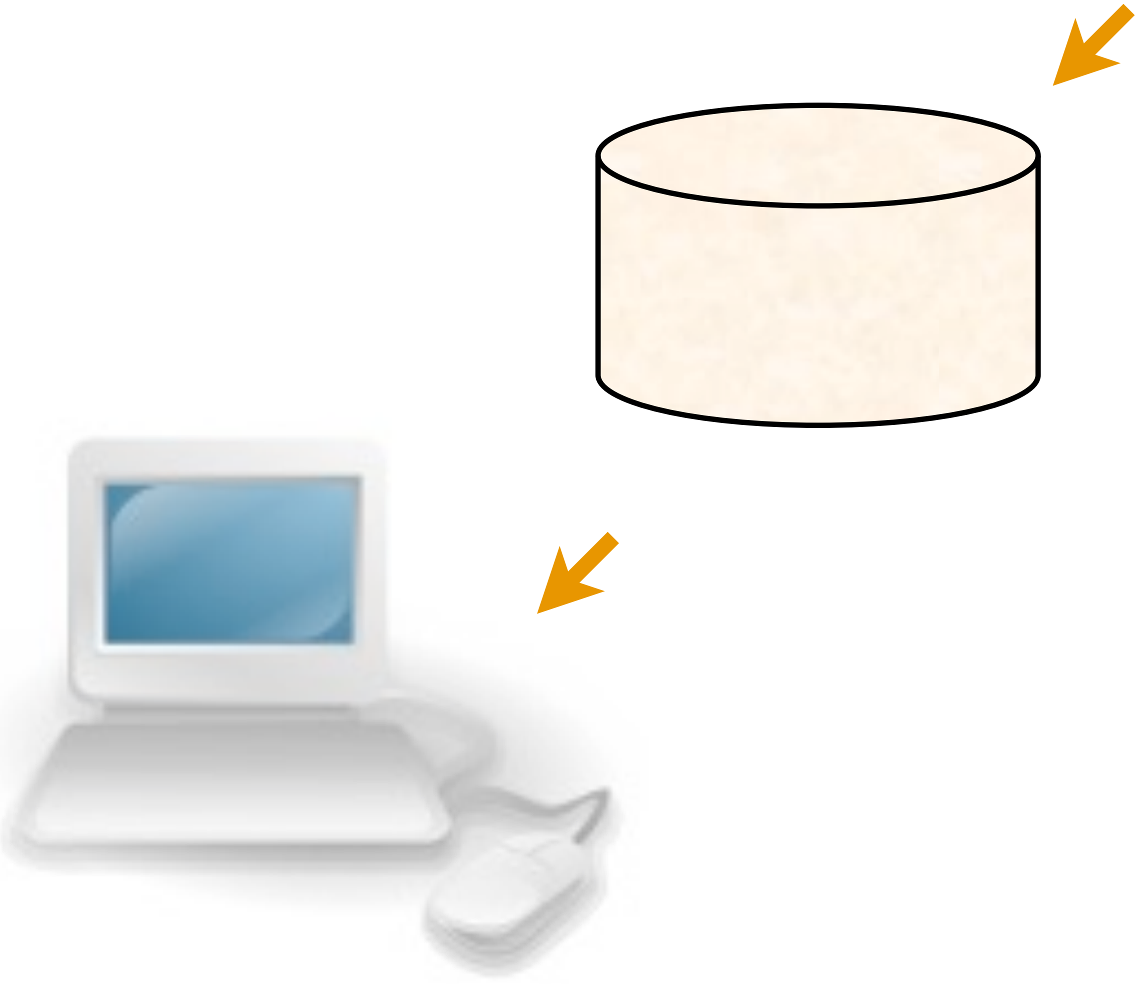
ENA (from Release note 115)
Blaise Alako, Clara Amid, Lawrence Bower, Ana Cerdeno-Taraga, Iain Cleland, Richard Gibson, Neil Goodgame, Petra ten Hoopen, Mikyung Jang, Simon Kay, Rasko Leinonen, Xin Liu, Arnaud Oisel, Rodrigo Lopez, Hamish McWilliam, Nima Pakseresht, Sheila Plaister, Rajesh Radhakrishnan, 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)
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

The business of DNA Databank

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

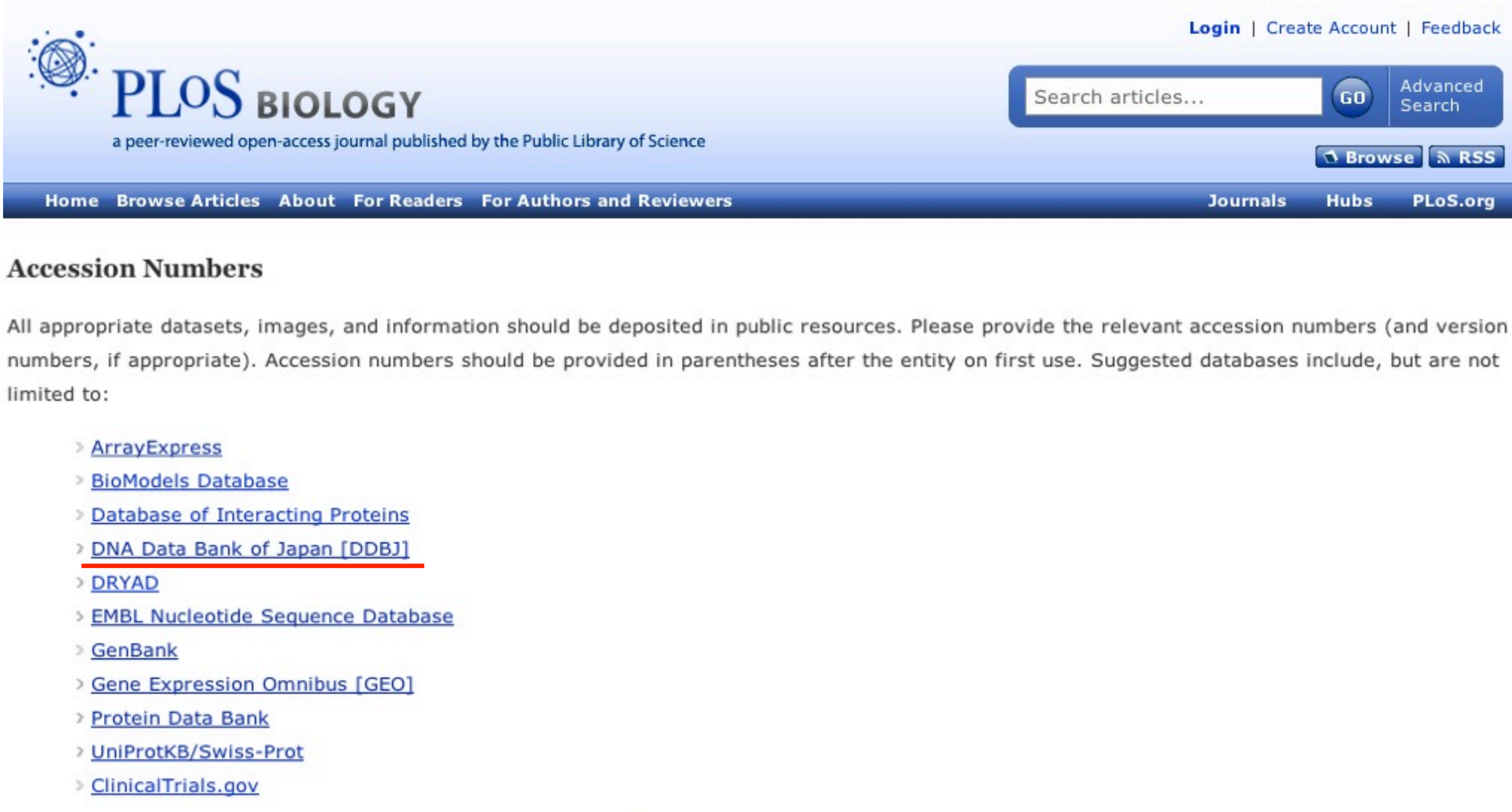




An example of DDBJ’s entry

```
LOCUS       HUMIL2HOM                397 bp    DNA     linear   HUM 27-APR-1993
DEFINITION  Human interleukin 2 (IL-2)-like DNA.
ACCESSION   M13784
VERSION     M13784.1
KEYWORDS    .
SOURCE      Homo sapiens (human)
  ORGANISM  Homo sapiens
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini;
            Catarrhini; Hominidae; Homo.
REFERENCE   1  (bases 1 to 397)
  AUTHORS   Mita,S., Maeda,S. and Shimada,K.
  TITLE     Characterization of human genomic DNA sequences homologous to the
            interleukin 2 cDNA
  JOURNAL   Biochem. Biophys. Res. Commun. 138 (2), 966-973 (1986)
   PUBMED   3017347
COMMENT     Original source text: Human placenta DNA, clone Lm HoIL2-3.
            Numerous stop codons are found in the interleukin 2-like IIa DNA.
FEATURES             Location/Qualifiers
     source          1..397
                     /organism="Homo sapiens"
                     /mol_type="genomic DNA"
                     /db_xref="taxon:9606"
BASE COUNT          117 a            84 c            48 g           148 t
ORIGIN      RsaI site.
        1 actgatttat ttttaataaa attacaagag attttaattt taaacccaaa agttctttta
       61 ttgcatctca ctgtgtttag ctttgtttac cctttgagaa ggcctgagat aataactttc
      121 ttcttcaact ctttcatcag ctcctgtaac cttttttcct taggttctta actgatgttg
      181 tggcctgctg ctaaaaacgc tttatcttaa agttctaaaa ggaaatgttt tcttctaaca
      241 taacattctg ggctcttgac tttatgaaat caaaaacttt cacttatgac caggatacac
      301 tcttcctctg tctaactaat tcaagcacta tcttcattca ttttgacttg cagattatcc
      361 aaacagactc cccataatga aaagcaatca cactgca
//
```

Accession No. is neccessary for publication



Accession Numbers

All appropriate datasets, images, and information should be deposited in public resources. Please provide the relevant accession numbers (and version numbers, if appropriate). Accession numbers should be provided in parentheses after the entity on first use. Suggested databases include, but are not limited to:

- ArrayExpress
- BioModels Database
- Database of Interacting Proteins
- DNA Data Bank of Japan [DDBJ]
- DRYAD
- EMBL Nucleotide Sequence Database
- GenBank
- Gene Expression Omnibus [GEO]
- Protein Data Bank
- UniProtKB/Swiss-Prot
- ClinicalTrials.gov

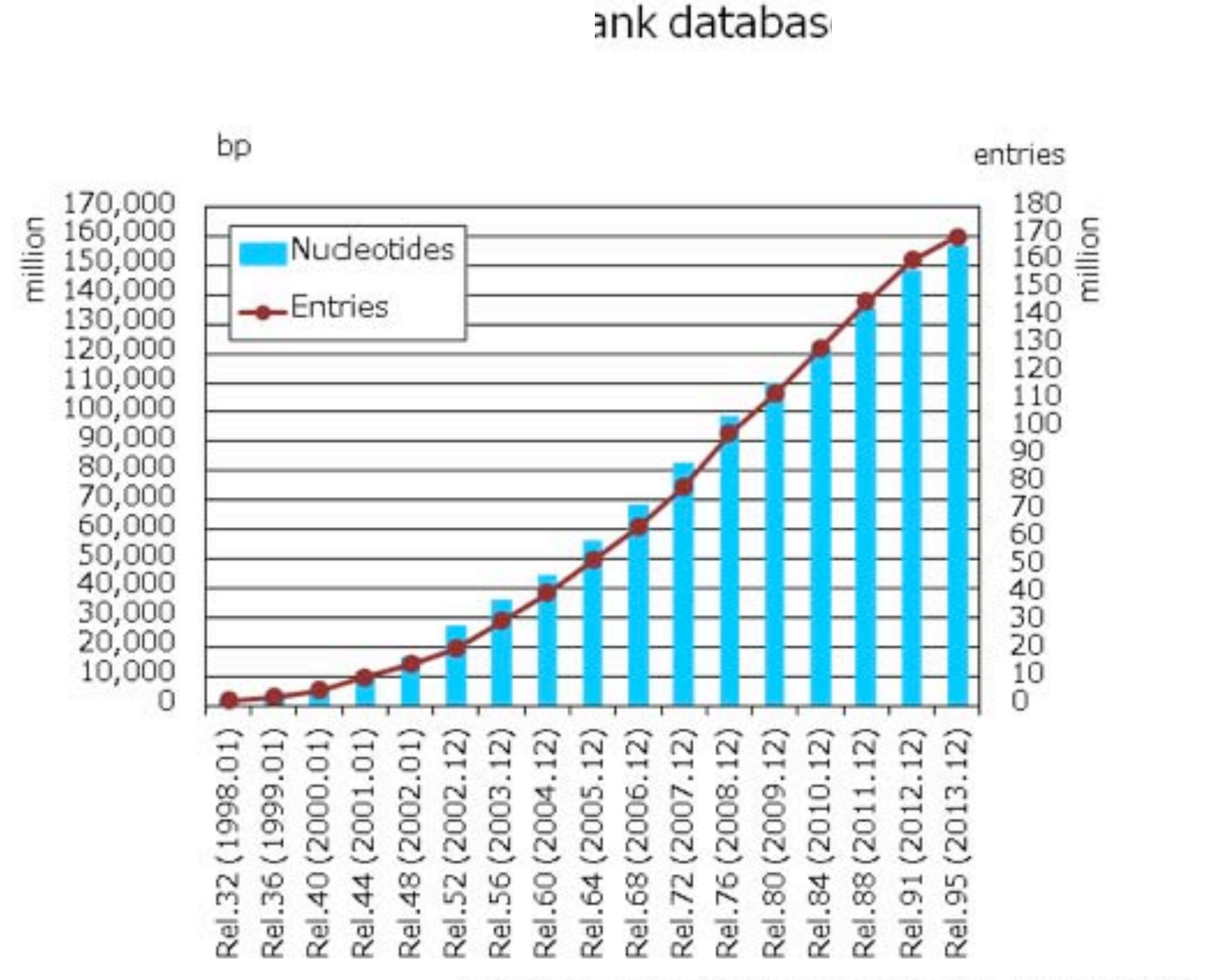
Current data amount of INSDC

160 billion nucleotides

170 million entries

1,600 億

1.7 億



Note: CON and TPA division are not counted in statistics of DDBJ

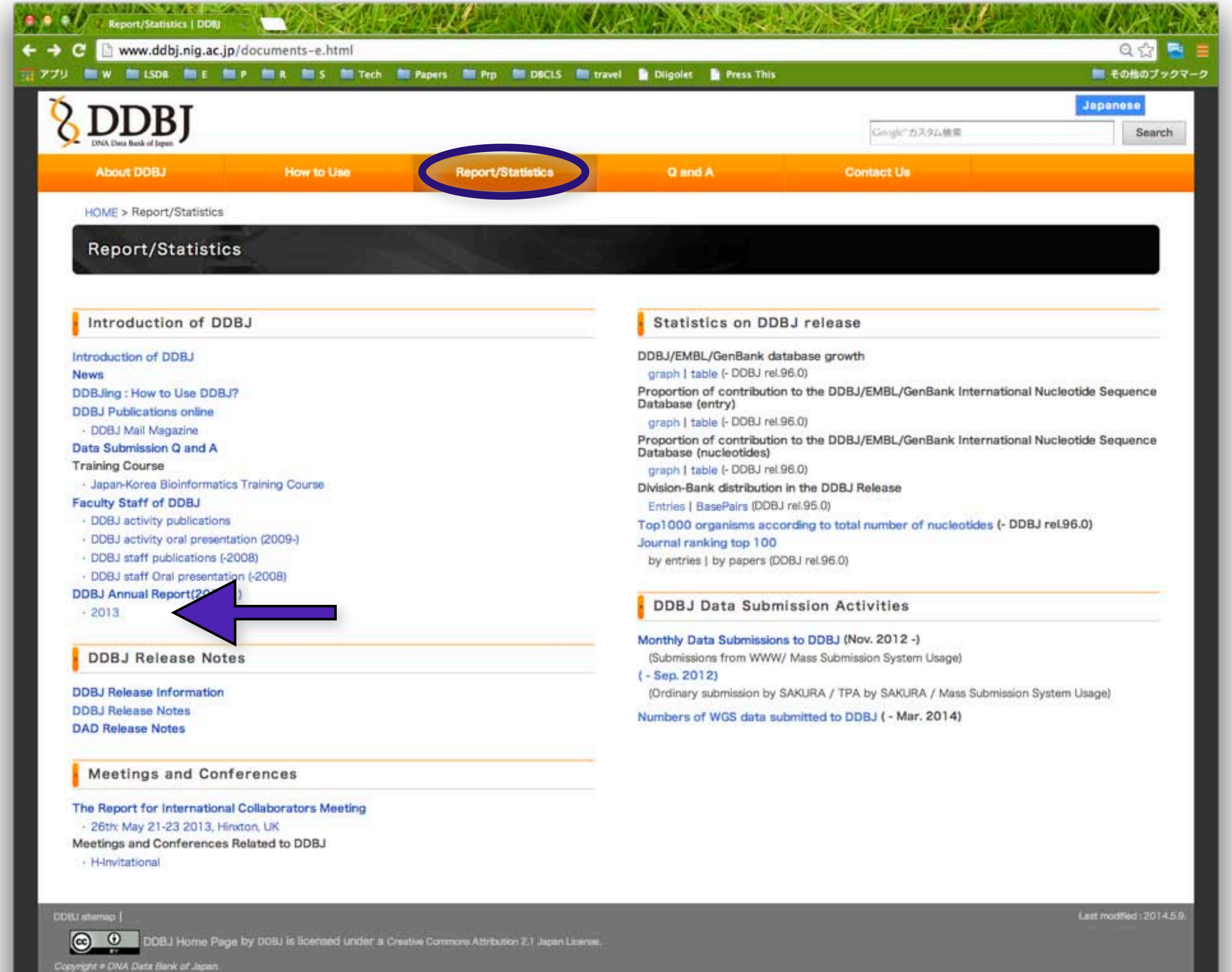
Species' word cloud in INSDC

The cloud gives greater prominence to words that appear more frequently in the source text.

DDBJ 2013-2014

DDBJ / NIG
Yaz Nakamura

ICM 2014 in Mishima



HOME > Report/Statistics

Report/Statistics

Introduction of DDBJ

Introduction of DDBJ
News
DDBJing : How to Use DDBJ?
DDBJ Publications online
- DDBJ Mail Magazine

Data Submission Q and A
Training Course
- Japan-Korea Bioinformatics Training Course

Faculty Staff of DDBJ
- DDBJ activity publications
- DDBJ activity oral presentation (2009-)
- DDBJ staff publications (-2008)
- DDBJ staff Oral presentation (-2008)

DDBJ Annual Report(20)
- 2013

DDBJ Release Notes

DDBJ Release Information
DDBJ Release Notes
DAD Release Notes

Meetings and Conferences

The Report for International Collaborators Meeting
- 26th: May 21-23 2013, Hinxton, UK

Meetings and Conferences Related to DDBJ
- H-Invitational

Statistics on DDBJ release

DDBJ/EMBL/GenBank database growth
graph | table (- DDBJ rel.96.0)
Proportion of contribution to the DDBJ/EMBL/GenBank International Nucleotide Sequence Database (entry)
graph | table (- DDBJ rel.96.0)
Proportion of contribution to the DDBJ/EMBL/GenBank International Nucleotide Sequence Database (nucleotides)
graph | table (- DDBJ rel.96.0)
Division-Bank distribution in the DDBJ Release
Entries | BasePairs (DDBJ rel.95.0)
Top1000 organisms according to total number of nucleotides (- DDBJ rel.96.0)
Journal ranking top 100
by entries | by papers (DDBJ rel.96.0)

DDBJ Data Submission Activities

Monthly Data Submissions to DDBJ (Nov. 2012 -)
(Submissions from WWW/ Mass Submission System Usage)
(- Sep. 2012)
(Ordinary submission by SAKURA / TPA by SAKURA / Mass Submission System Usage)
Numbers of WGS data submitted to DDBJ (- Mar. 2014)

DDBJ sitemap |

Last modified : 2014.5.9.

Overview

Faculty staff

Director: Toshihisa Takagi Prof., Eli Kaminuma Assist Prof., Yasukazu Nakamura Prof., Osamu Ogasawara Assist Prof., Kousaku Okubo Prof., Masanori Arita Prof. (6)

DB construction team

Leader: Jun Mashima Dr., Hideo Aono, Mayumi Ejima, Asami Fukuda, Noriko Furuya Dr., Yuichi Kodama Dr., Takehide Kosuge Dr., Kyungbum Lee Dr., Kimiko Mimura, Toshihisa Okido Dr., Katsunaga Sakai Dr., Rie Sugita, Haru Tsutsui (13)

Public Relation team

Leader: Junko Kohira, Kimiko Suzuki, Emi Yokoyama (3)

Engineering staff

DB management team / SC team

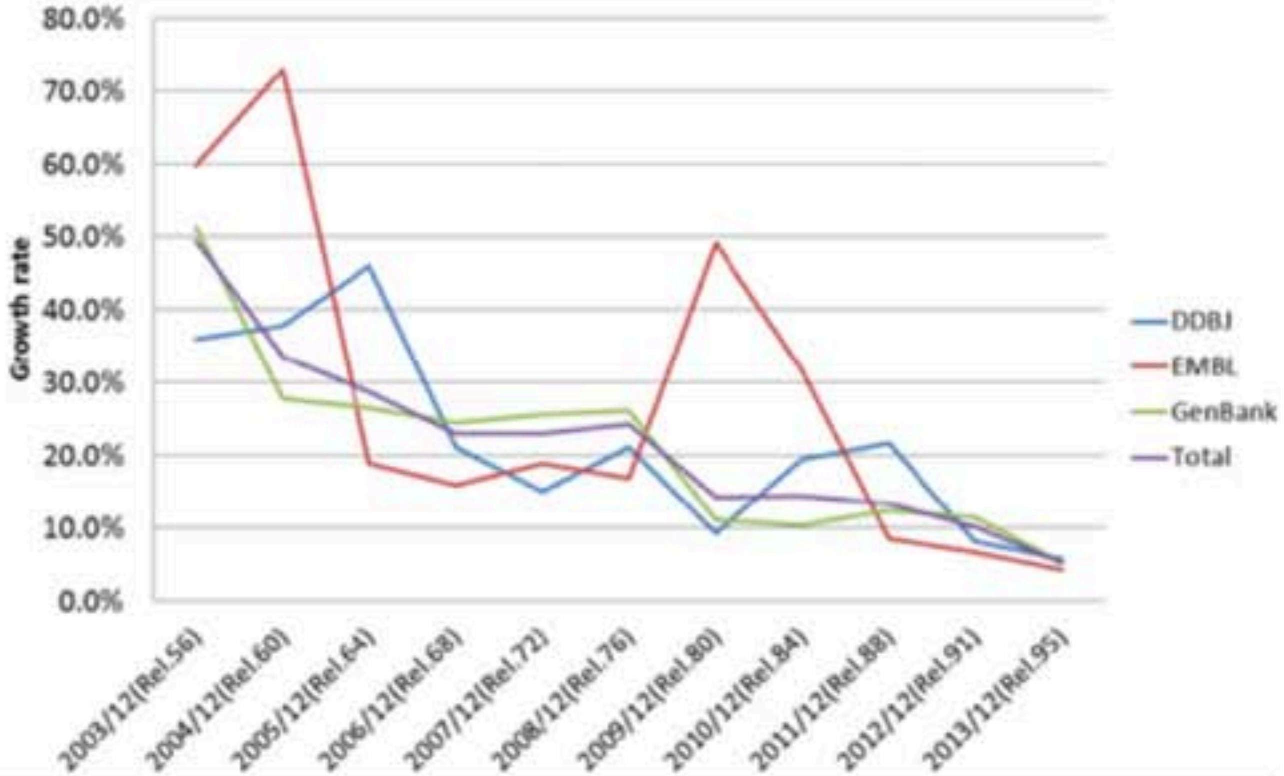
Leader: Koji Watanabe, Chiharu Kawagoe, Yuji Ashizawa, Yukino Dobashi, Masahiro Fujimoto, Tomohiro Hirai, Naofumi Ishikawa, Toshikazu Katsumata, Shingo Kawahara, Takeshi Moriyama, Yoshihiro Okuda, Makoto Sato, Yoshihiro Serizawa, Aimi Shiida, Yukie Shinyama, Daisuke Takagi, Daisuke Takai, Takeshi Tsurusawa, Tomohiko Yasuda, Shigeru Yatsuzuka (20)

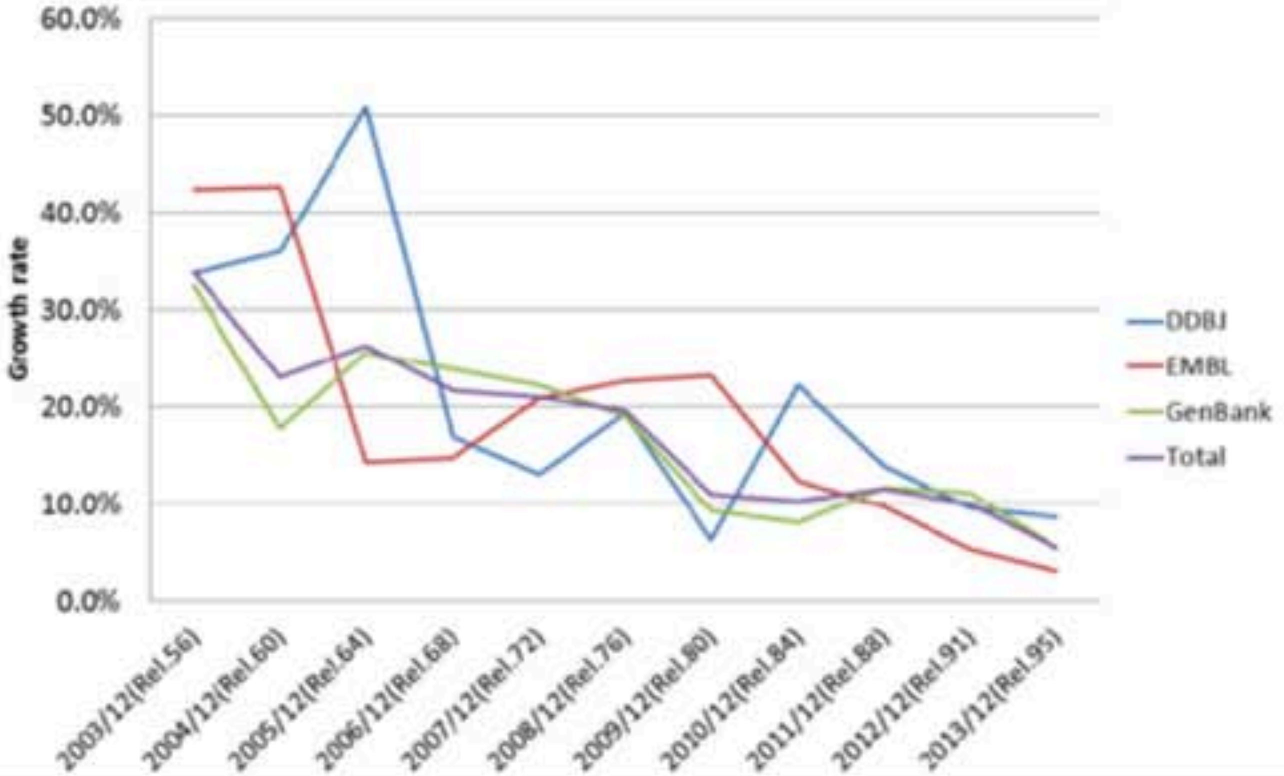
Secretary

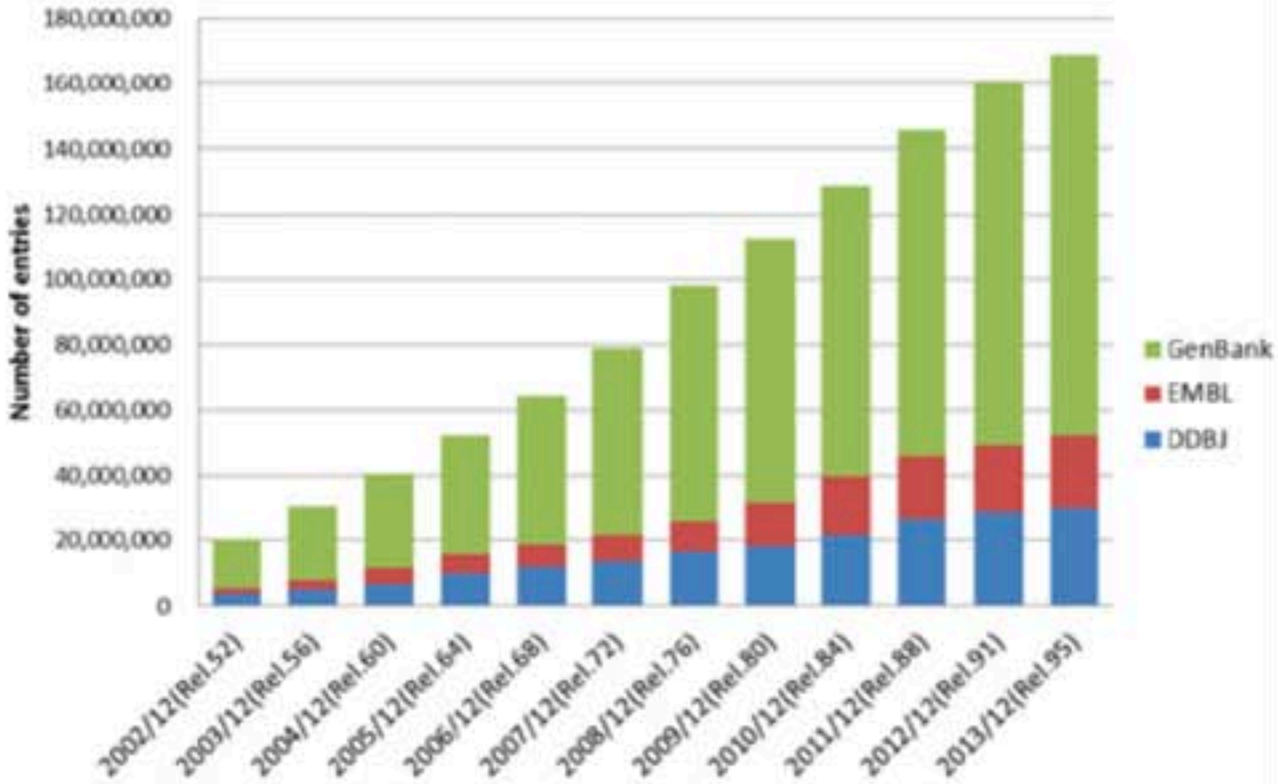
Mika Maki, Naoko Murakata (2)

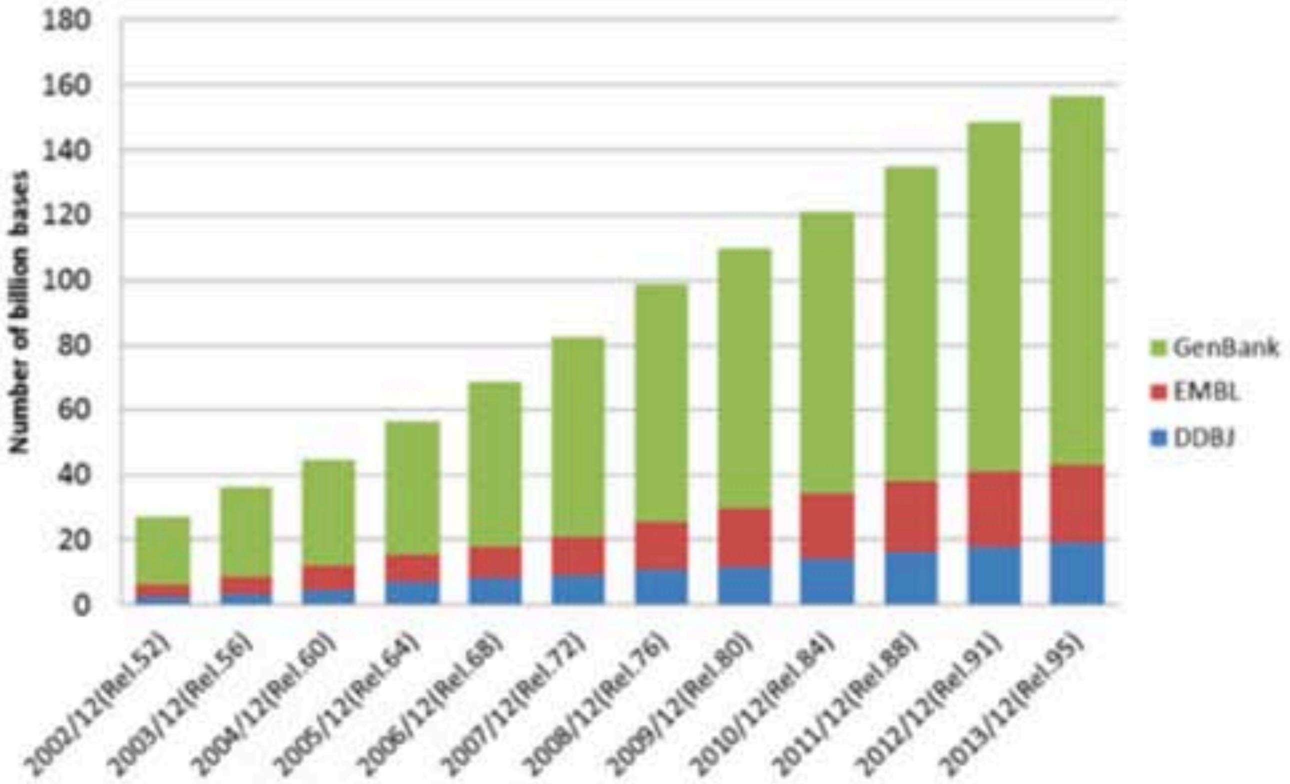
From Release Note for Rel. 96.0, Mar. 2014

Traditional DDBJ / INSDC


Traditional DDBJ growth rate by entries

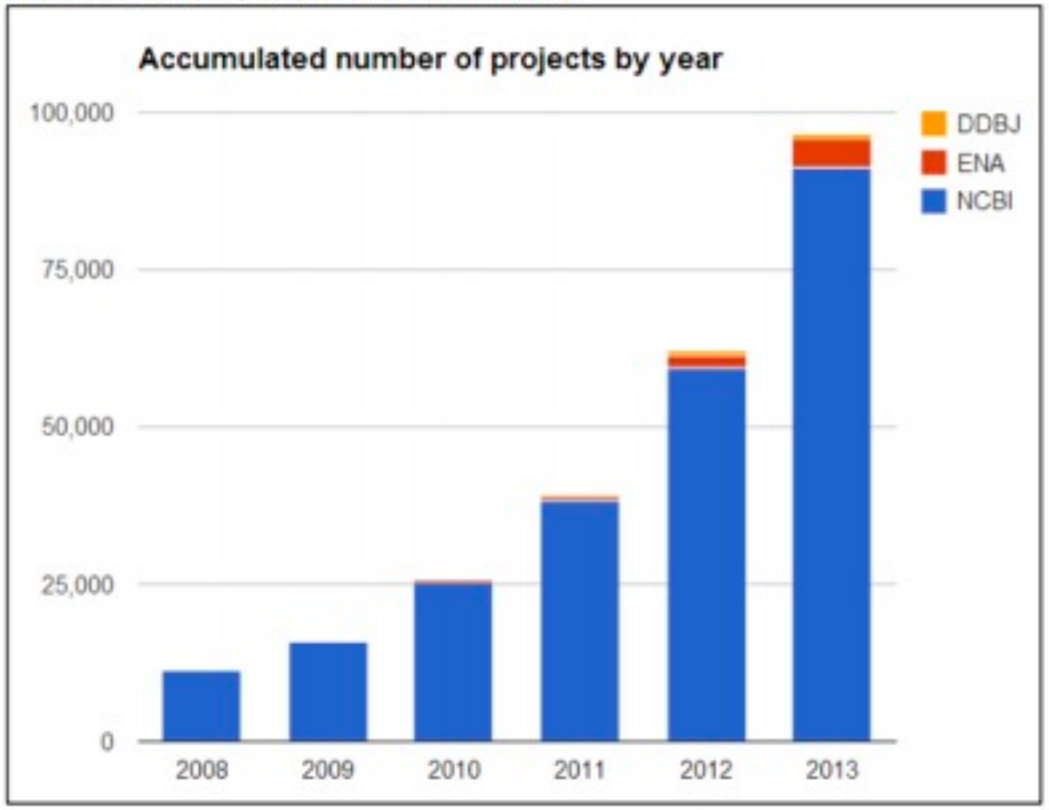

Traditional DDBJ growth rate by nucleotides


The Number of Entries by Contributors to DDBJ Release

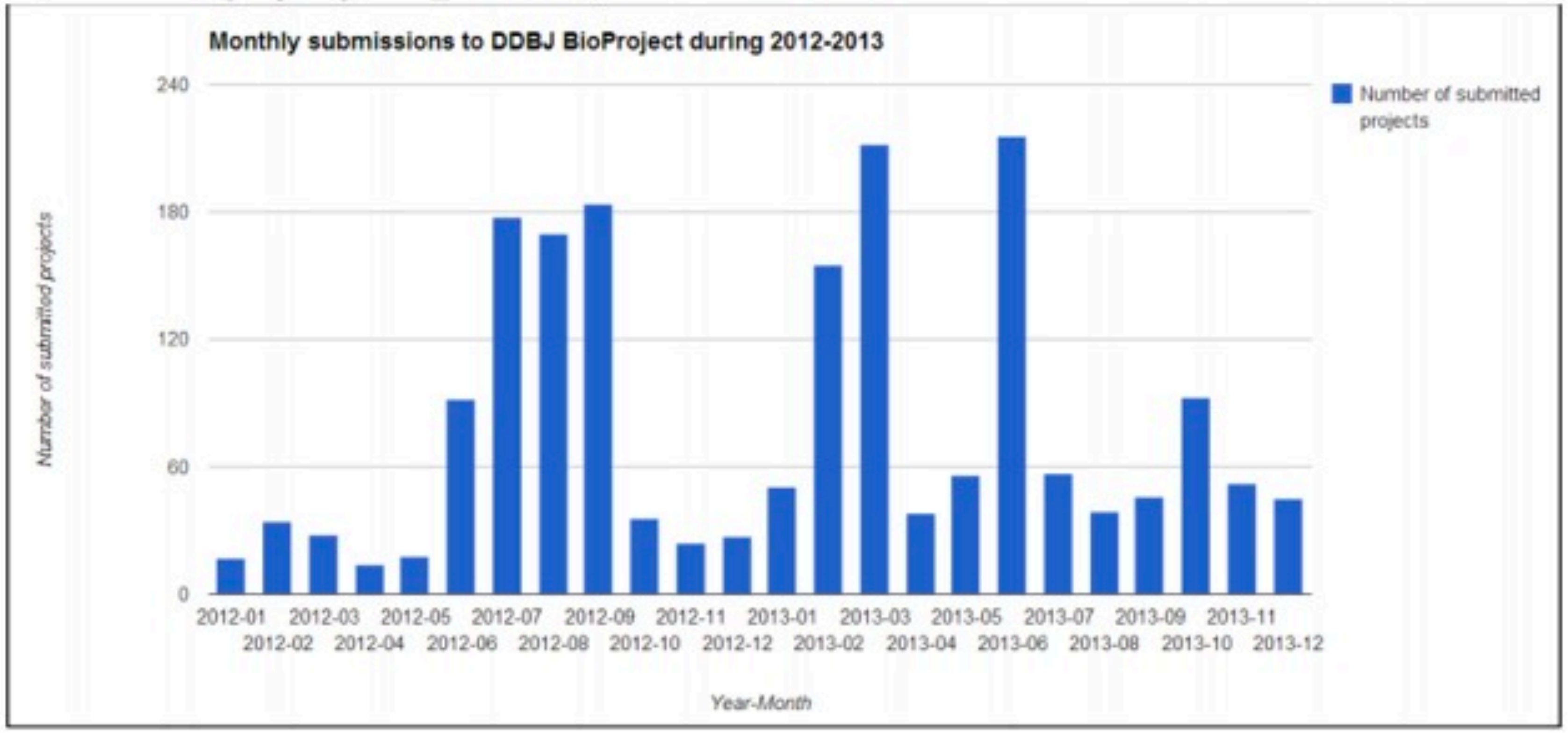

The Number of nucleotides by Contributors to DDBJ Release

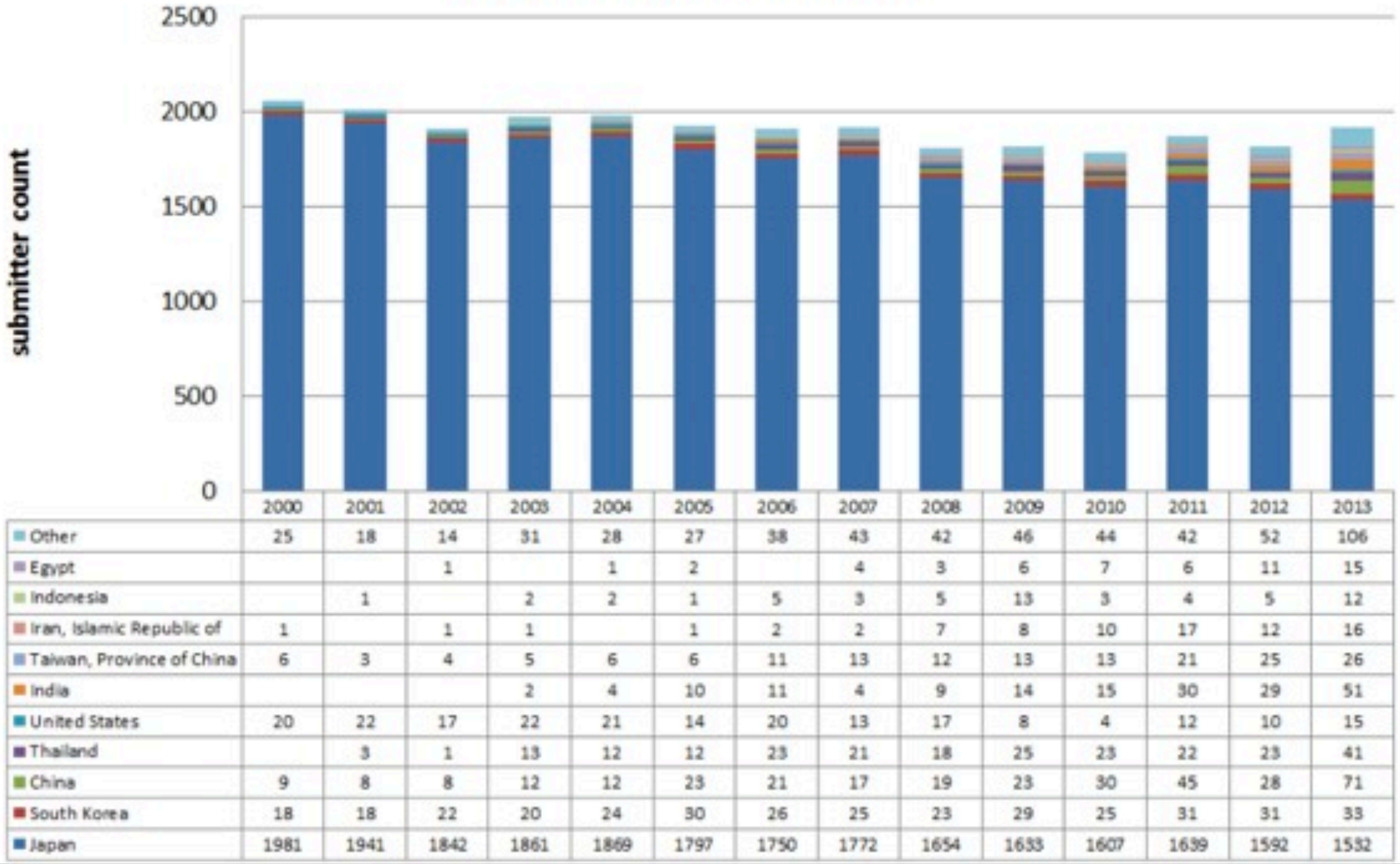
BioProject

Statistics page as of 3th, February, 2014:
http://trace.ddbj.nig.ac.jp/stat_e.html#project



Statistics page as of 3th, February, 2014:
http://trace.ddbj.nig.ac.jp/stat_e.html#bp





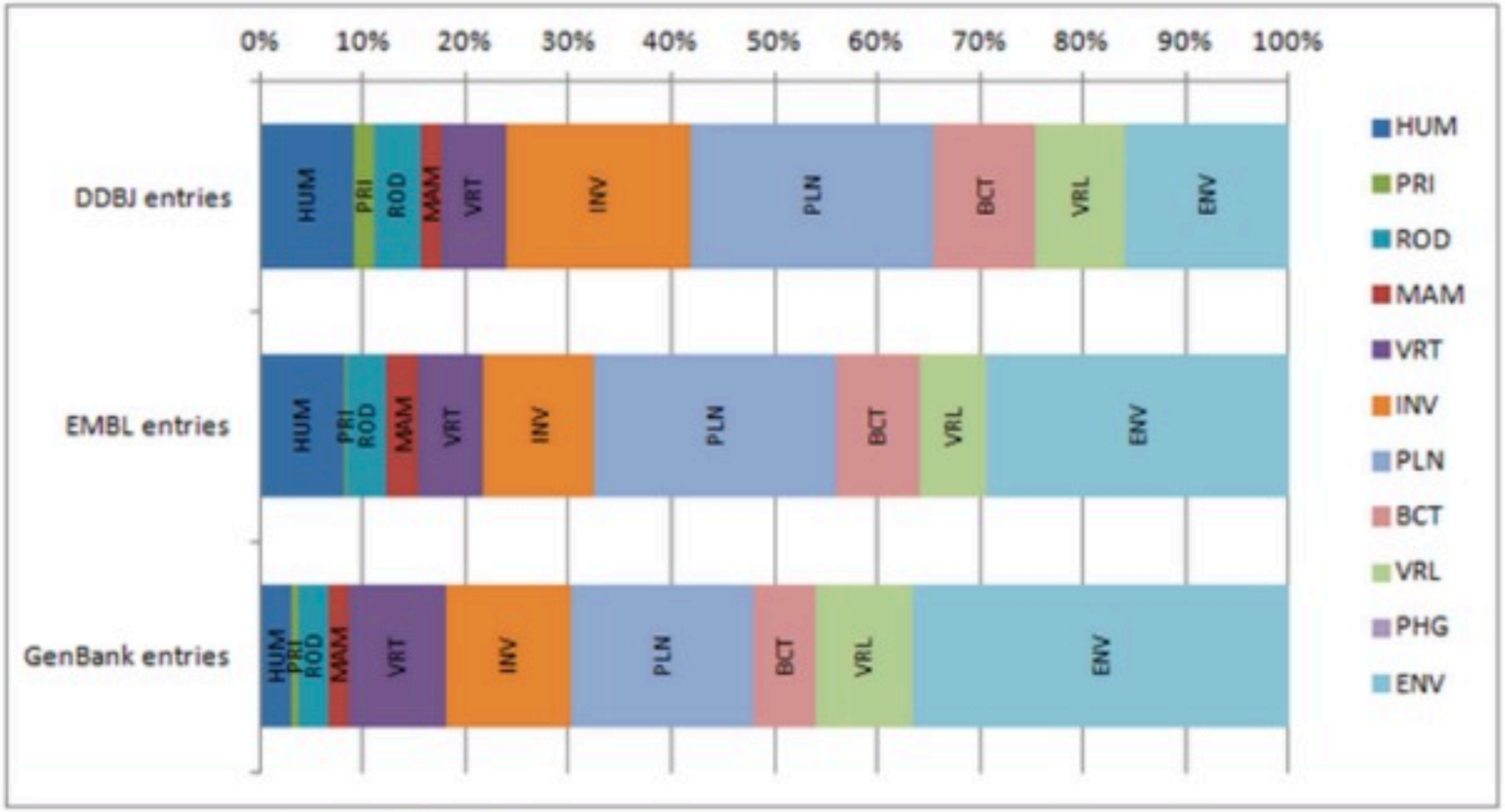
DDBJ submitter's country Top 10

	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013
Other	25	18	14	31	28	27	38	43	42	46	44	42	52	106
Egypt			1		1	2		4	3	6	7	6	11	15
Indonesia		1		2	2	1	5	3	5	13	3	4	5	12
Iran, Islamic Republic of	1		1	1		1	2	2	7	8	10	17	12	16
Taiwan, Province of China	6	3	4	5	6	6	11	13	12	13	13	21	25	26
India				2	4	10	11	4	9	14	15	30	29	51
United States	20	22	17	22	21	14	20	13	17	8	4	12	10	15
Thailand		3	1	13	12	12	23	21	18	25	23	22	23	41
China	9	8	8	12	12	23	21	17	19	23	30	45	28	71
South Korea	18	18	22	20	24	30	26	25	23	29	25	31	31	33
Japan	1981	1941	1842	1861	1869	1797	1750	1772	1654	1633	1607	1639	1592	1532

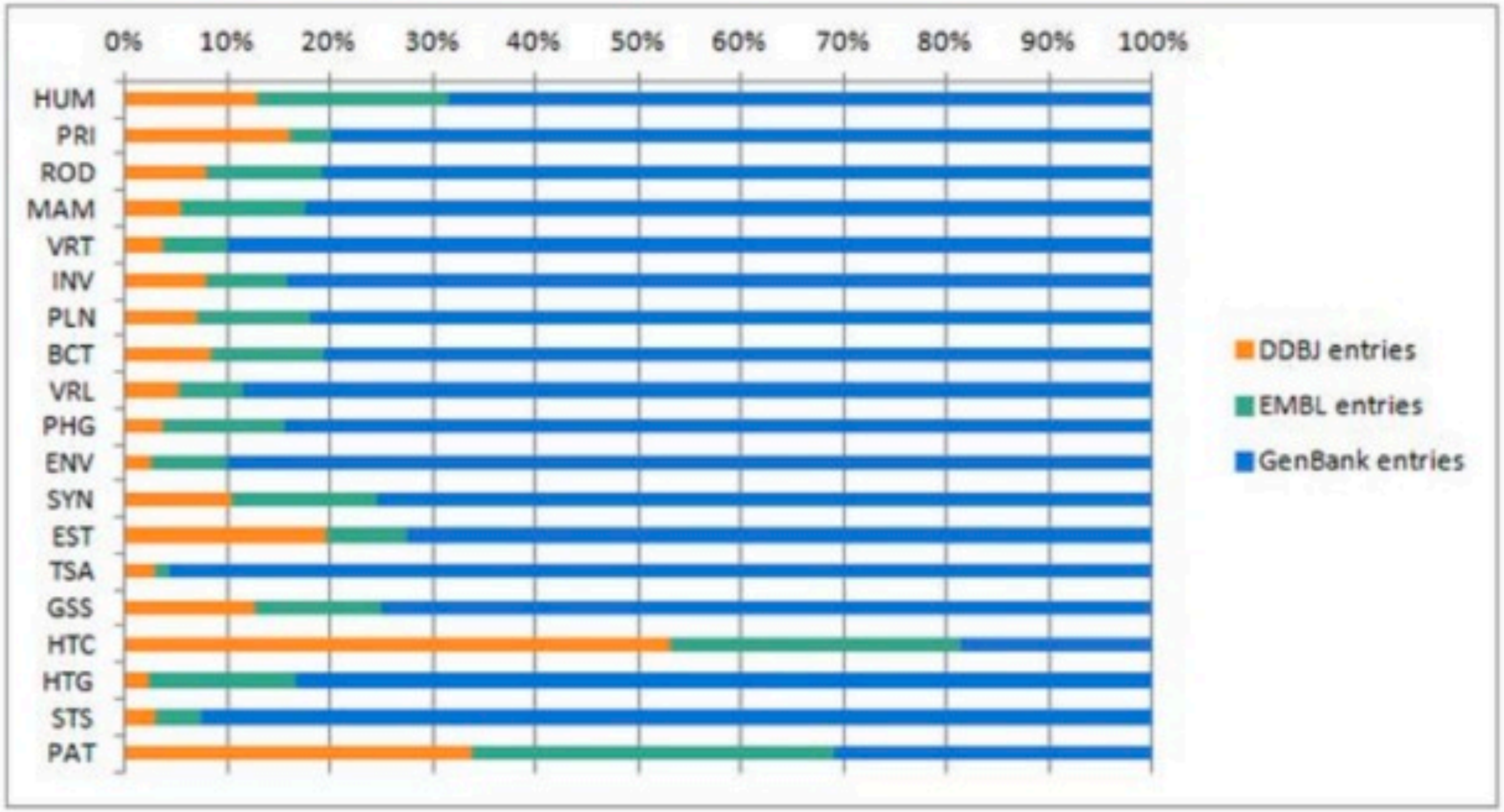
Top 10 country of submit to DDBJ

	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013
Other	39	26	25	127	68	52	82	86	144	135	120	141	165	287
Egypt	0	0	2	0	1	4	0	15	21	10	26	9	23	48
Pakistan	0	2	0	0	0	5	2	3	6	7	52	43	31	19
Iran, Islamic Republic of	40	0	1	43	0	1	3	2	12	12	20	50	29	37
United States	38	63	48	36	40	29	43	18	25	10	7	14	14	26
Taiwan, Province of China	24	61	8	24	8	12	81	56	22	36	40	158	44	62
India	0	0	0	3	9	34	27	21	18	41	82	135	116	223
Thailand	0	5	1	36	29	25	64	56	71	84	62	89	96	99
China	46	34	14	57	79	58	52	183	52	59	57	111	89	143
South Korea	69	66	61	55	95	101	110	119	47	61	66	78	146	106
Japan	4123	4110	4134	3884	4311	4084	4090	4109	3734	4103	3628	3755	3851	3450

Divisional distribution (taxonomic & env divisions) in each bank

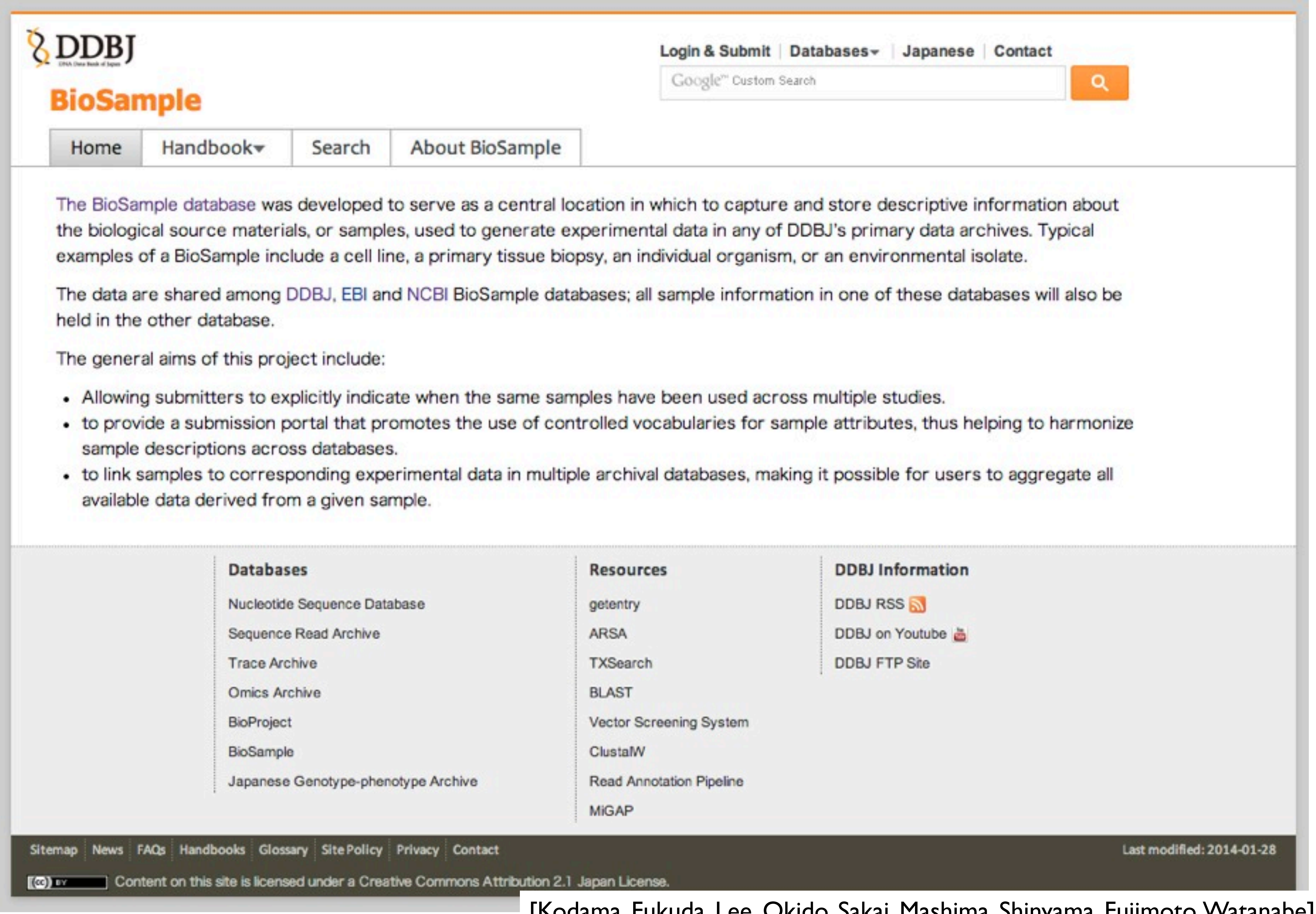


Bank proportion in each 19 divisions



http://www.ddbj.nig.ac.jp/breakdown_stats/div-bank_ent-e.html new

BioSample



The BioSample database was developed to serve as a central location in which to capture and store descriptive information about the biological source materials, or samples, used to generate experimental data in any of DDBJ's primary data archives. Typical examples of a BioSample include a cell line, a primary tissue biopsy, an individual organism, or an environmental isolate.

The data are shared among DDBJ, EBI and NCBI BioSample databases; all sample information in one of these databases will also be held in the other database.

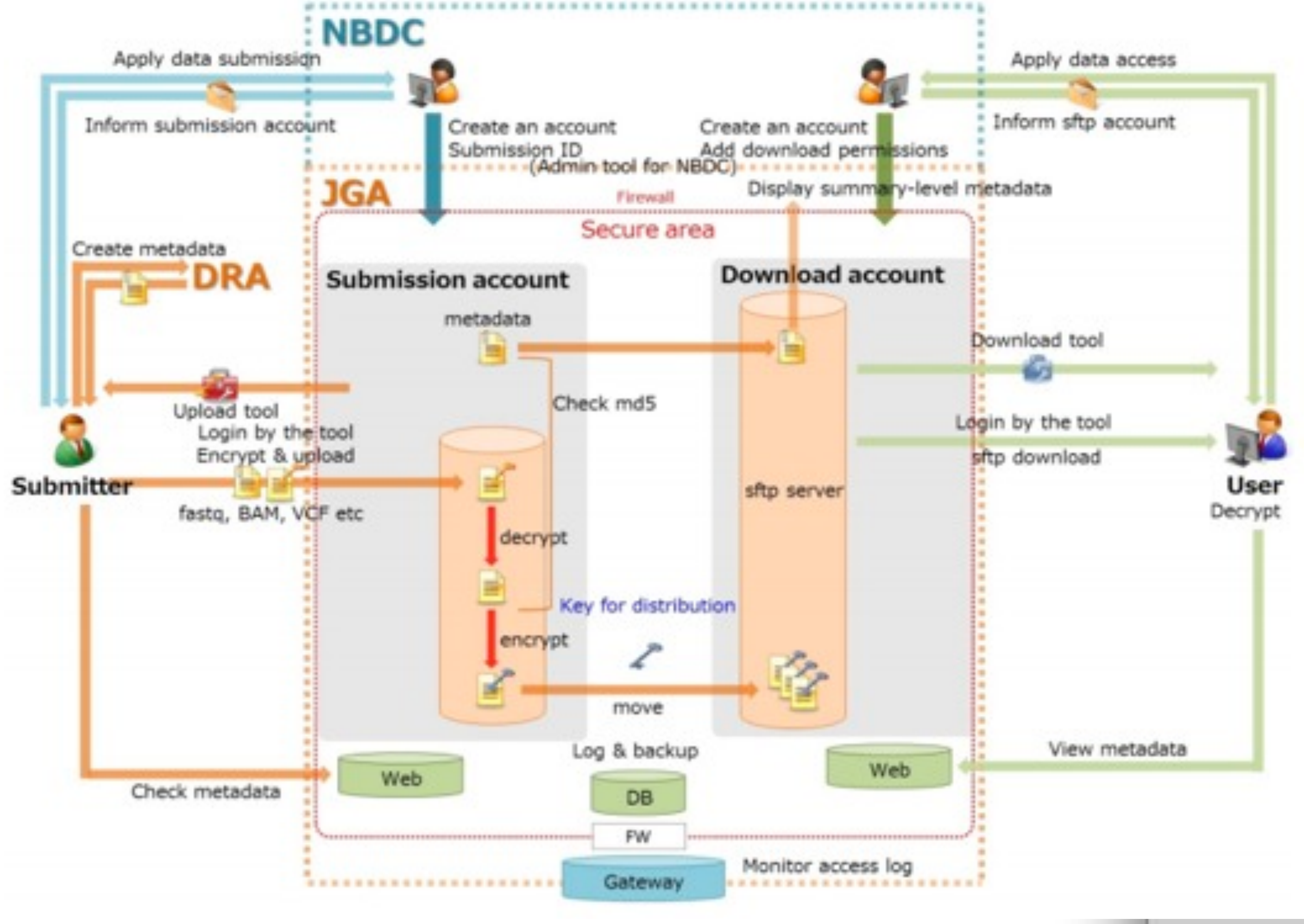
The general aims of this project include:

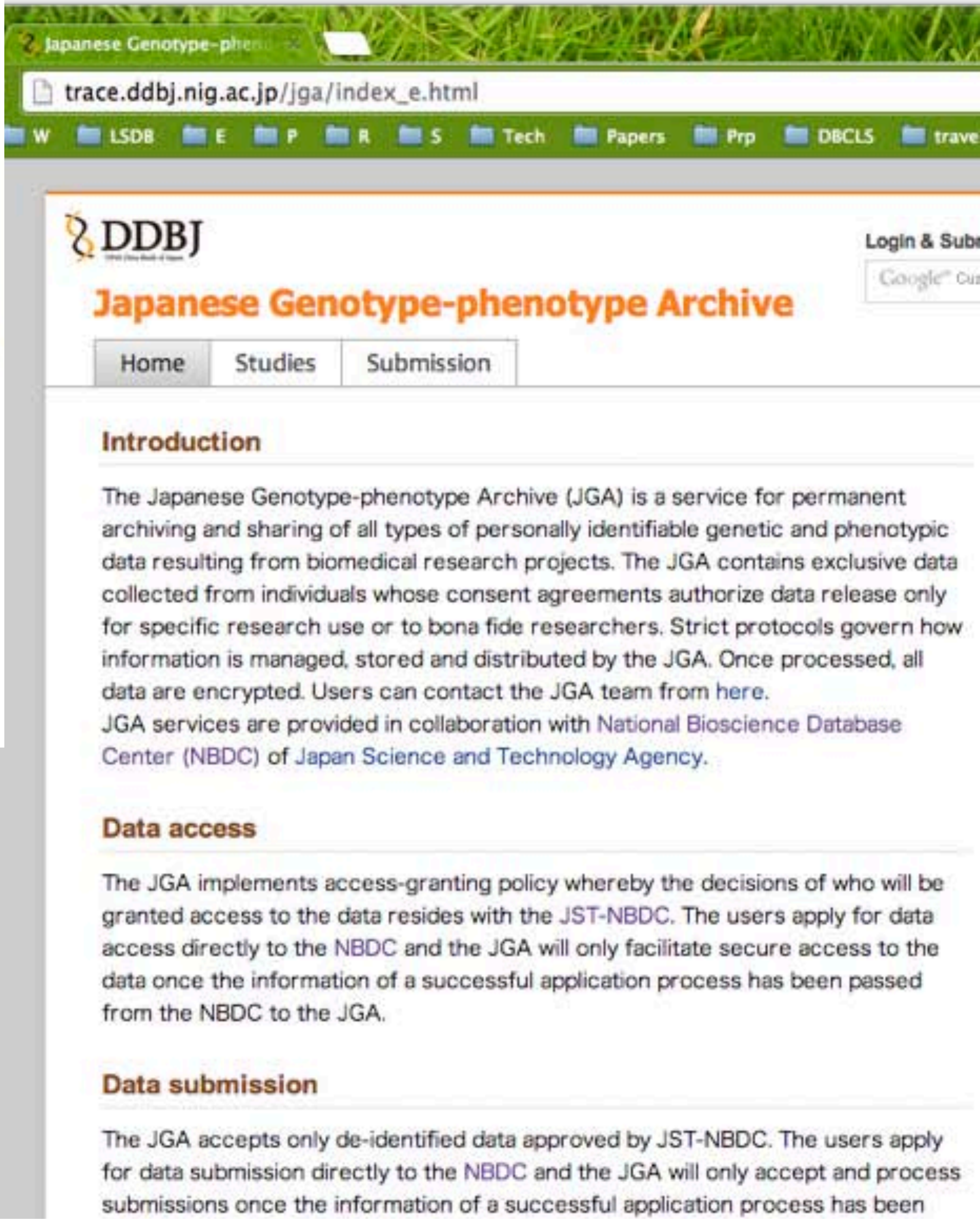
- Allowing submitters to explicitly indicate when the same samples have been used across multiple studies.
- to provide a submission portal that promotes the use of controlled vocabularies for sample attributes, thus helping to harmonize sample descriptions across databases.
- to link samples to corresponding experimental data in multiple archival databases, making it possible for users to aggregate all available data derived from a given sample.

Content on this site is licensed under a Creative Commons Attribution 2.1 Japan License.

[Kodama, Fukuda, Lee, Okido, Sakai, Mashima, Shinyama, Fujimoto, Watanabe]

JGA: The Japanese Genotype-phenotype Archive





DDBJ

Japanese Genotype-phenotype Archive

Introduction

The Japanese Genotype-phenotype Archive (JGA) is a service for permanent archiving and sharing of all types of personally identifiable genetic and phenotypic data resulting from biomedical research projects. The JGA contains exclusive data collected from individuals whose consent agreements authorize data release only for specific research use or to bona fide researchers. Strict protocols govern how information is managed, stored and distributed by the JGA. Once processed, all data are encrypted. Users can contact the JGA team from here.
JGA services are provided in collaboration with National Bioscience Database Center (NBDC) of Japan Science and Technology Agency.

Data access

The JGA implements access-granting policy whereby the decisions of who will be granted access to the data resides with the JST-NBDC. The users apply for data access directly to the NBDC and the JGA will only facilitate secure access to the data once the information of a successful application process has been passed from the NBDC to the JGA.

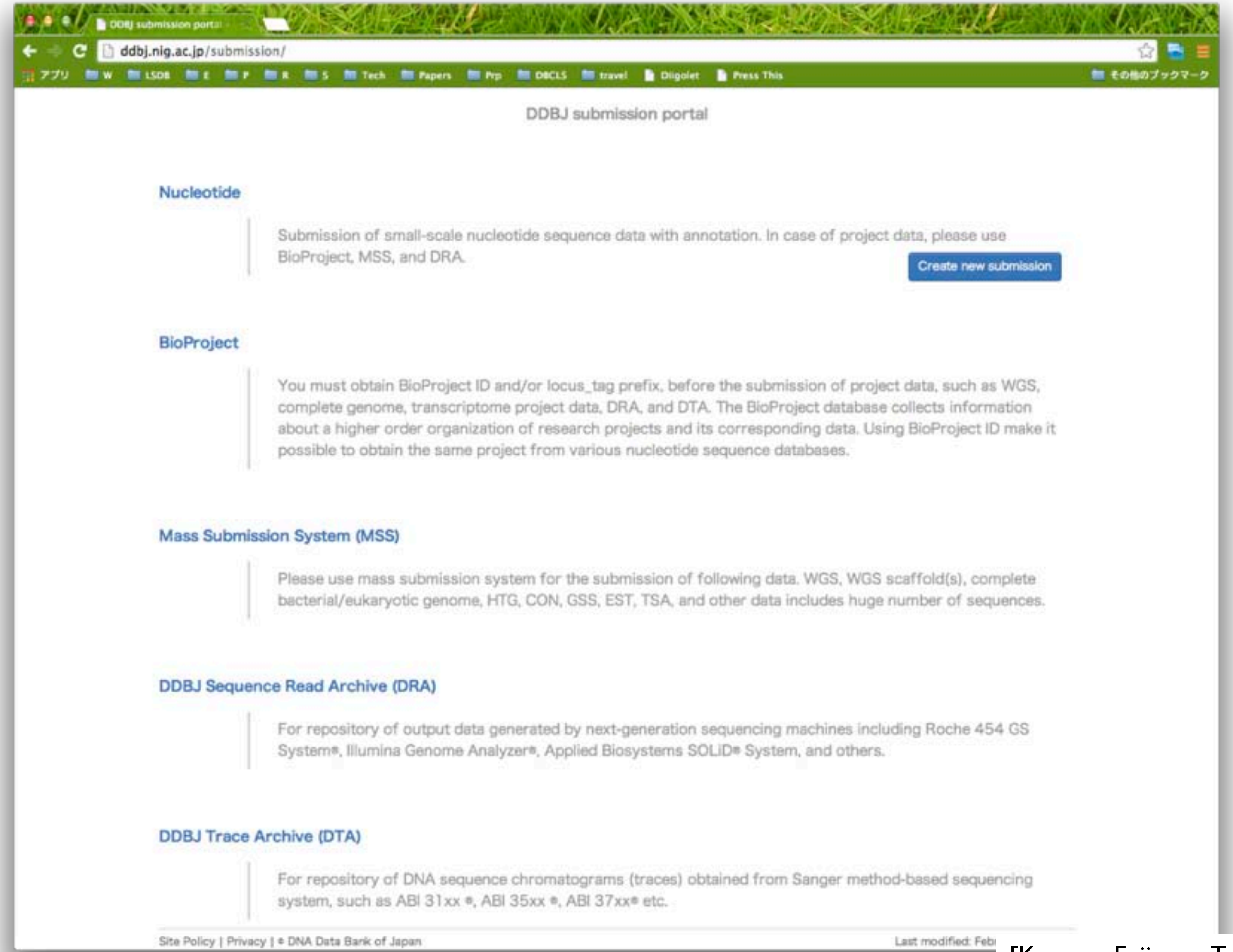
Data submission

The JGA accepts only de-identified data approved by JST-NBDC. The users apply for data submission directly to the NBDC and the JGA will only accept and process submissions once the information of a successful application process has been

* NBDC: National Bioscience Database Center (http://biosciencedbc.jp)

[Kodama, Mashima, Shinyama, Fujimoto, Watanabe, Tsurusawa, NBDC: Minowa, Kawashima, Mitsuhashi] refactored

Integrated submission portal



DDBJ submission portal

Nucleotide

Submission of small-scale nucleotide sequence data with annotation. In case of project data, please use BioProject, MSS, and DRA.

BioProject

You must obtain BioProject ID and/or locus_tag prefix, before the submission of project data, such as WGS, complete genome, transcriptome project data, DRA, and DTA. The BioProject database collects information about a higher order organization of research projects and its corresponding data. Using BioProject ID make it possible to obtain the same project from various nucleotide sequence databases.

Mass Submission System (MSS)

Please use mass submission system for the submission of following data. WGS, WGS scaffold(s), complete bacterial/eukaryotic genome, HTG, CON, GSS, EST, TSA, and other data includes huge number of sequences.

DDBJ Sequence Read Archive (DRA)

For repository of output data generated by next-generation sequencing machines including Roche 454 GS System®, Illumina Genome Analyzer®, Applied Biosystems SOLiD® System, and others.

DDBJ Trace Archive (DTA)

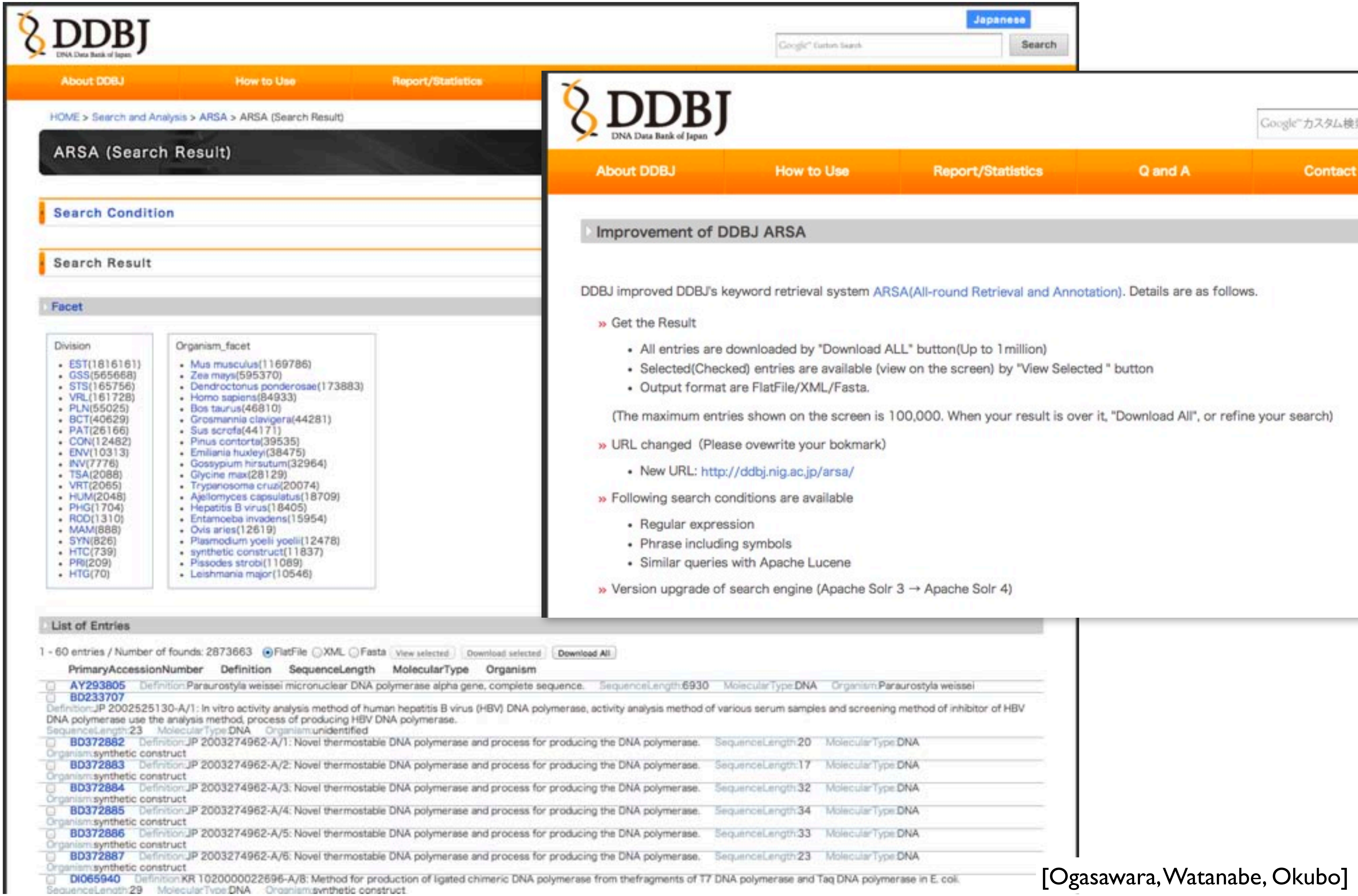
For repository of DNA sequence chromatograms (traces) obtained from Sanger method-based sequencing system, such as ABI 31xx ®, ABI 35xx ®, ABI 37xx® etc.

Site Policy | Privacy | © DNA Data Bank of Japan

Last modified: Feb

[Kosuge,, Fujisawa, Tsutsui, Ejima]

ARSA/getentry



HOME > Search and Analysis > ARSA > ARSA (Search Result)

ARSA (Search Result)

Search Condition

Search Result

Facet

Division
- EST(1816161)
- GSS(565668)
- STS(165756)
- VRL(161728)
- PLN(55025)
- BCT(40629)
- PAT(26166)
- CON(12482)
- ENV(10313)
- INV(7776)
- TSA(2088)
- VRT(2065)
- HUM(2048)
- PHG(1704)
- ROD(1310)
- MAM(888)
- SYN(826)
- HTC(739)
- PRI(209)
- HTG(70)

Organism_facet
- Mus musculus(1169786)
- Zea mays(595370)
- Dendroctonus ponderosae(173883)
- Homo sapiens(84933)
- Bos taurus(46810)
- Grosmannia clavigera(44281)
- Sus scrofa(44171)
- Pinus contorta(39535)
- Emiliania huxleyi(38475)
- Gossypium hirsutum(32964)
- Glycine max(28129)
- Trypanosoma cruzi(20074)
- Ajellomyces capsulatus(18709)
- Hepatitis B virus(18405)
- Entamoeba invadens(15954)
- Ovis aries(12619)
- Plasmodium yoelii yoelii(12478)
- synthetic construct(11837)
- Pissodes strobi(11089)
- Leishmania major(10546)

List of Entries

1 - 60 entries / Number of founds: 2873663 ● FlatFile ○ XML ○ Fasta

PrimaryAccessionNumber Definition SequenceLength MolecularType Organism

☐ AY293805 Definition:Paraurostyla weissei micronuclear DNA polymerase alpha gene, complete sequence. SequenceLength:6930 MolecularType:DNA Organism:Paraurostyla weissei

☐ BD233707
Definition:JP 2002525130-A/1: In vitro activity analysis method of human hepatitis B virus (HBV) DNA polymerase, activity analysis method of various serum samples and screening method of inhibitor of HBV DNA polymerase use the analysis method, process of producing HBV DNA polymerase.
SequenceLength:23 MolecularType:DNA Organism:unidentified

☐ BD372882 Definition:JP 2003274962-A/1: Novel thermostable DNA polymerase and process for producing the DNA polymerase. SequenceLength:20 MolecularType:DNA
Organism:synthetic construct

☐ BD372883 Definition:JP 2003274962-A/2: Novel thermostable DNA polymerase and process for producing the DNA polymerase. SequenceLength:17 MolecularType:DNA
Organism:synthetic construct

☐ BD372884 Definition:JP 2003274962-A/3: Novel thermostable DNA polymerase and process for producing the DNA polymerase. SequenceLength:32 MolecularType:DNA
Organism:synthetic construct

☐ BD372885 Definition:JP 2003274962-A/4: Novel thermostable DNA polymerase and process for producing the DNA polymerase. SequenceLength:34 MolecularType:DNA
Organism:synthetic construct

☐ BD372886 Definition:JP 2003274962-A/5: Novel thermostable DNA polymerase and process for producing the DNA polymerase. SequenceLength:33 MolecularType:DNA
Organism:synthetic construct

☐ BD372887 Definition:JP 2003274962-A/6: Novel thermostable DNA polymerase and process for producing the DNA polymerase. SequenceLength:23 MolecularType:DNA
Organism:synthetic construct

☐ DI065940 Definition:KR 1020000022696-A/8: Method for production of ligated chimeric DNA polymerase from thefragments of T7 DNA polymerase and Taq DNA polymerase in E. coli.
SequenceLength:29 MolecularType:DNA Organism:synthetic construct

Improvement of DDBJ ARSA

DDBJ improved DDBJ's keyword retrieval system ARSA(All-round Retrieval and Annotation). Details are as follows.

» Get the Result

- All entries are downloaded by "Download ALL" button(Up to 1million)
- Selected(Checked) entries are available (view on the screen) by "View Selected " button
- Output format are FlatFile/XML/Fasta.

(The maximum entries shown on the screen is 100,000. When your result is over it, "Download All", or refine your search)

» URL changed (Please ovewrite your bokmark)

- New URL: http://ddbj.nig.ac.jp/arsa/

» Following search conditions are available

- Regular expression
- Phrase including symbols
- Similar queries with Apache Lucene

» Version upgrade of search engine (Apache Solr 3 → Apache Solr 4)

[Ogasawara, Watanabe, Okubo]

Homepage renewal

DDBJ
DNA Data Bank of Japan

ENGLISH

サイト内検索

HOME | 塩基配列の登録 | 利用の手引き | 検索・解析 | FTP・WebAPI | レポート・統計 | お問い合わせ

- DDBJの紹介
- Q&A集

塩基配列の登録
- SAKURA
- 大量登録システム(MSS)
- データの修正・更新
- DDBJ Sequence Read Archive
- DDBJ Trace Archive

プロジェクトの登録
- DDBJ BioProject Database

検索
- getentry
- ARSA
- TXSearch
- BLAST

系統解析
- ClustalW

NGSデータ解析

DDBJ : DNA Data Bank of Japan

DDBJ（日本DNAデータバンク）は欧州と米国の対応機関（EBIおよびNCBI）と密接に協力しながら DDBJ/EMBL/GenBank 国際塩基配列データベースを構築している三大国際DNAデータバンクのひとつです

Hot Topics ▸ 一覧へ

- 2012.01.12 DDBJ リリース 88.0、DAD リリース 58.0 完成
- 2012.01.11 ブタ (Sus scrofa) full length enriched cDNA 配列 データの公開
- 2012.01.05 Nucleic Acids Research に DDBJ に関する論文発表

Maintenance ▸ 一覧へ

- 2011.12.21 （全サービス再開） (2012/1/12-18) DDBJ 公開サービスの一時停止予定

Information

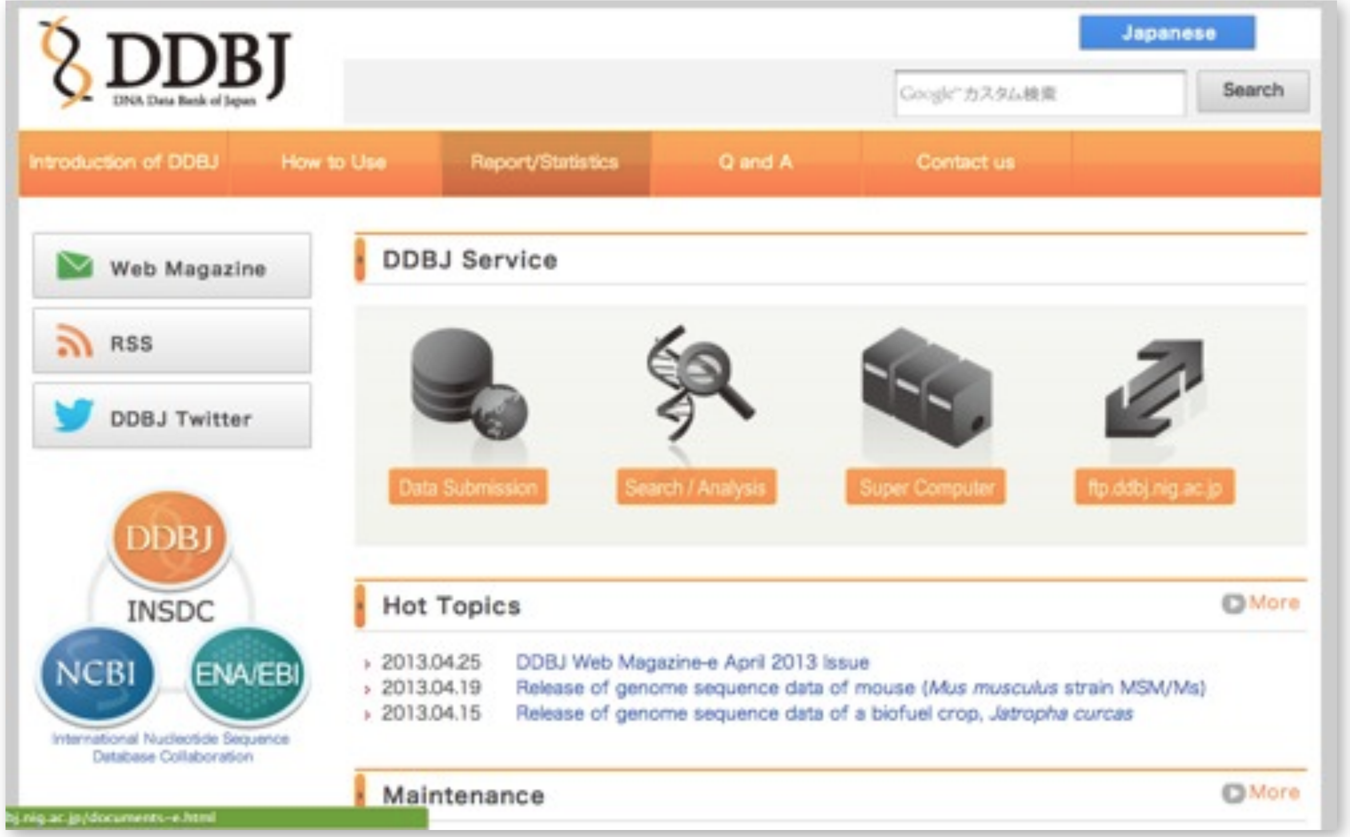
- 「第25回 DDBJing 講習会 in 三島」開催のお知らせ
- DDBJ メールマガジン No.67 配信

塩基配列の登録・更新

塩基配列の登録

FTP・Web API

FTP (ftp.ddbj.nig.ac.jp)



DDBJ
DNA Data Bank of Japan

Japanese

Google™カスタム検索 | Search

Introduction of DDBJ | How to Use | Report/Statistics | Q and A | Contact us

- Web Magazine
- RSS
- DDBJ Twitter

DDBJ
INSDC
NCBI
ENA/EBI
International Nucleotide Sequence Database Collaboration

DDBJ Service

Data Submission | Search / Analysis | Super Computer | ftp.ddbj.nig.ac.jp

Hot Topics More

- 2013.04.25 DDBJ Web Magazine-e April 2013 Issue
- 2013.04.19 Release of genome sequence data of mouse (*Mus musculus* strain MSM/Ms)
- 2013.04.15 Release of genome sequence data of a biofuel crop, *Jatropha curcas*

Maintenance More

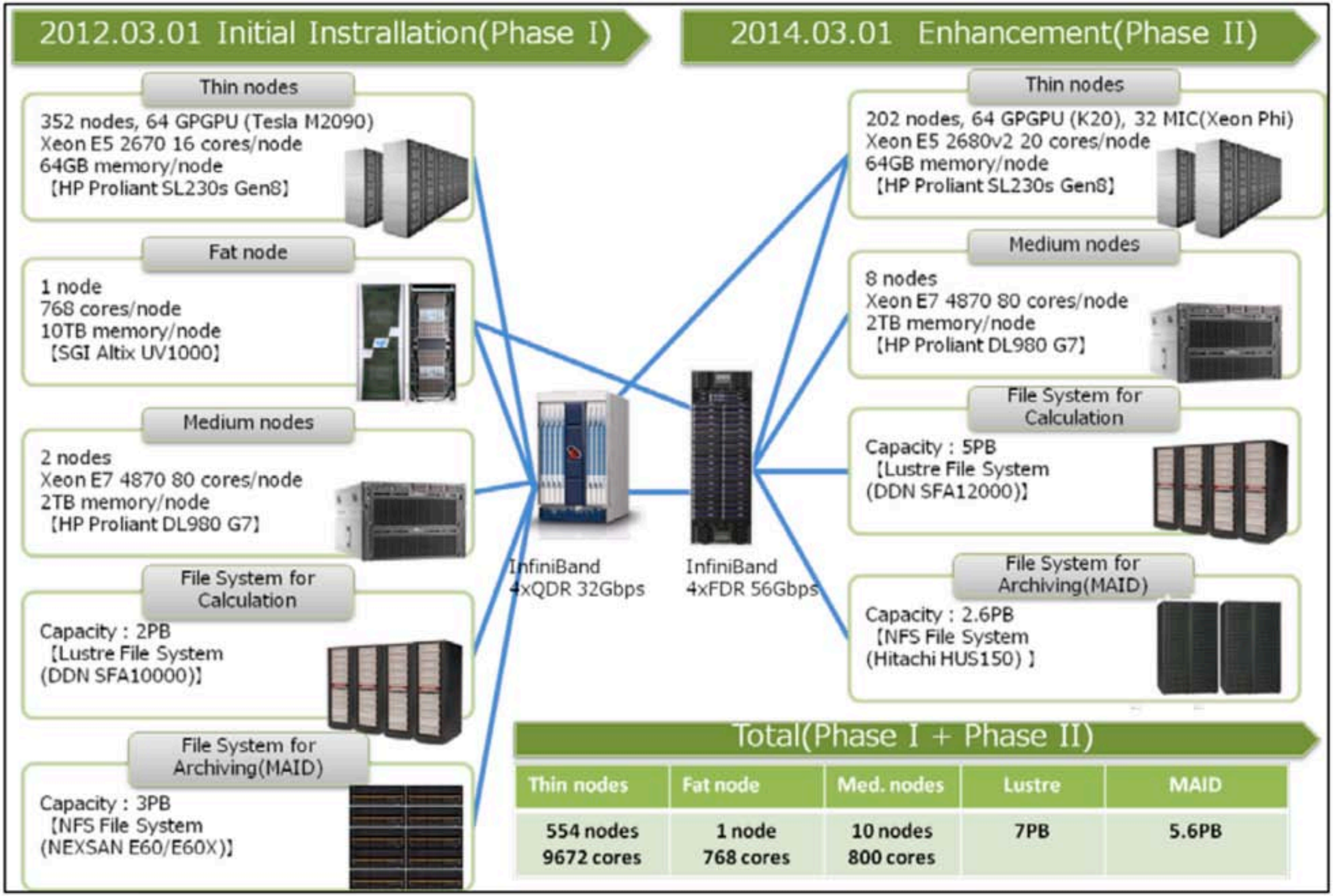
bj.nig.ac.jp/documents-e.html

WordPress

A free content management system (CMS)

enhancement

NIG Supercomputer System



2012.03.01 Initial Instrallation(Phase I)

Thin nodes

352 nodes, 64 GPGPU (Tesla M2090)
Xeon E5 2670 16 cores/node
64GB memory/node
【HP Proliant SL230s Gen8】

Fat node

1 node
768 cores/node
10TB memory/node
【SGI Altix UV1000】

Medium nodes

2 nodes
Xeon E7 4870 80 cores/node
2TB memory/node
【HP Proliant DL980 G7】

File System for Calculation

Capacity : 2PB
【Lustre File System
(DDN SFA10000)】

File System for Archiving(MAID)

Capacity : 3PB
【NFS File System
(NEXSAN E60/E60X)】

2014.03.01 Enhancement(Phase II)

Thin nodes

202 nodes, 64 GPGPU (K20), 32 MIC(Xeon Phi)
Xeon E5 2680v2 20 cores/node
64GB memory/node
【HP Proliant SL230s Gen8】

Medium nodes

8 nodes
Xeon E7 4870 80 cores/node
2TB memory/node
【HP Proliant DL980 G7】

File System for Calculation

Capacity : 5PB
【Lustre File System
(DDN SFA12000)】

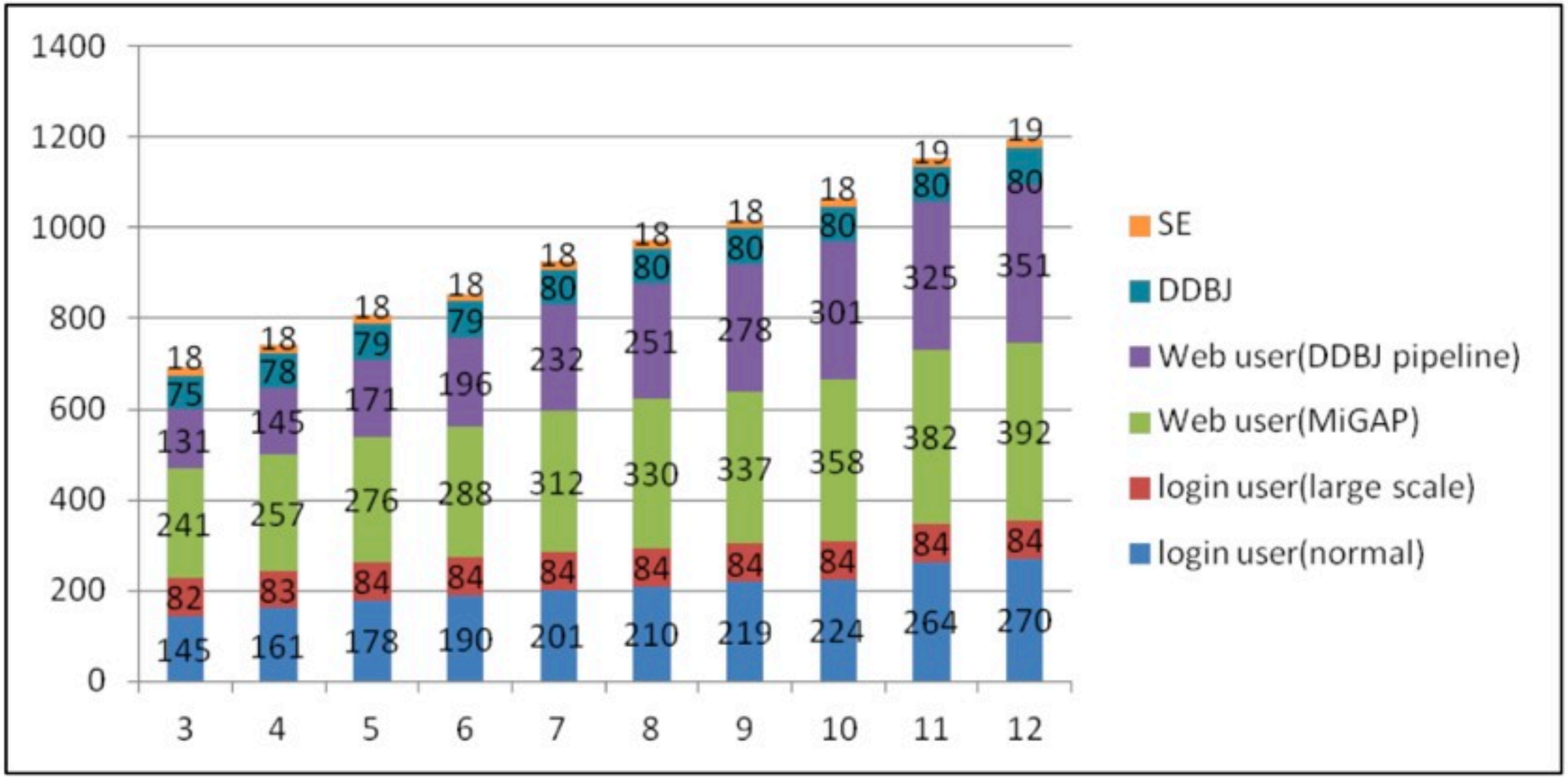
File System for Archiving(MAID)

Capacity : 2.6PB
【NFS File System
(Hitachi HUS150) 】

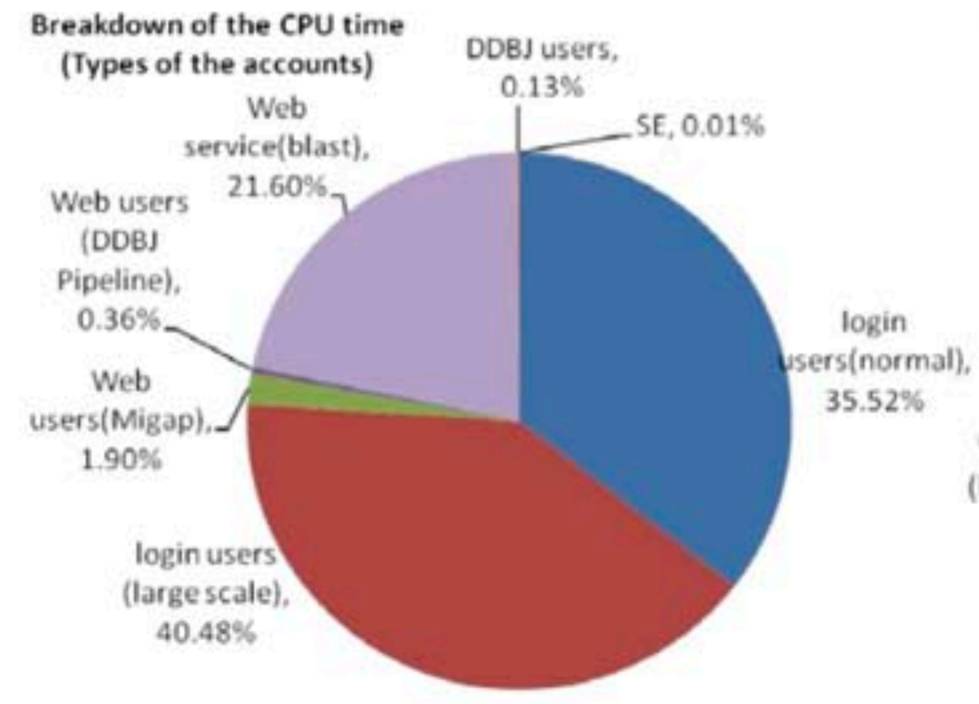
Total(Phase I + Phase II)

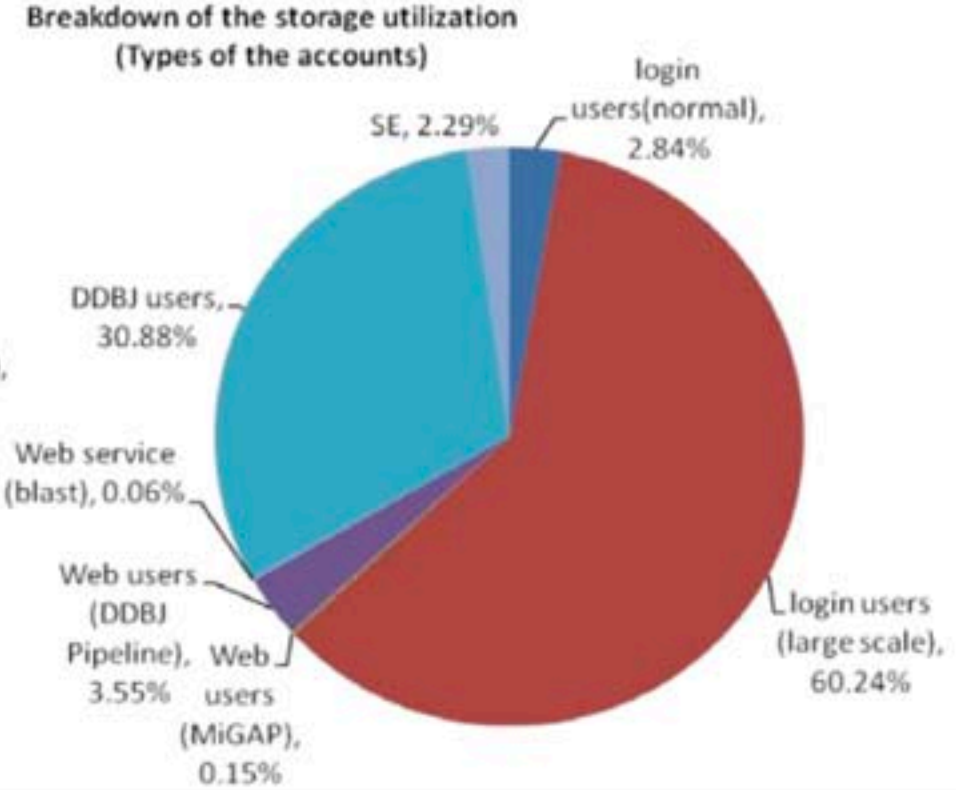
Thin nodes	Fat node	Med. nodes	Lustre	MAID
554 nodes 9672 cores	1 node 768 cores	10 nodes 800 cores	7PB	5.6PB

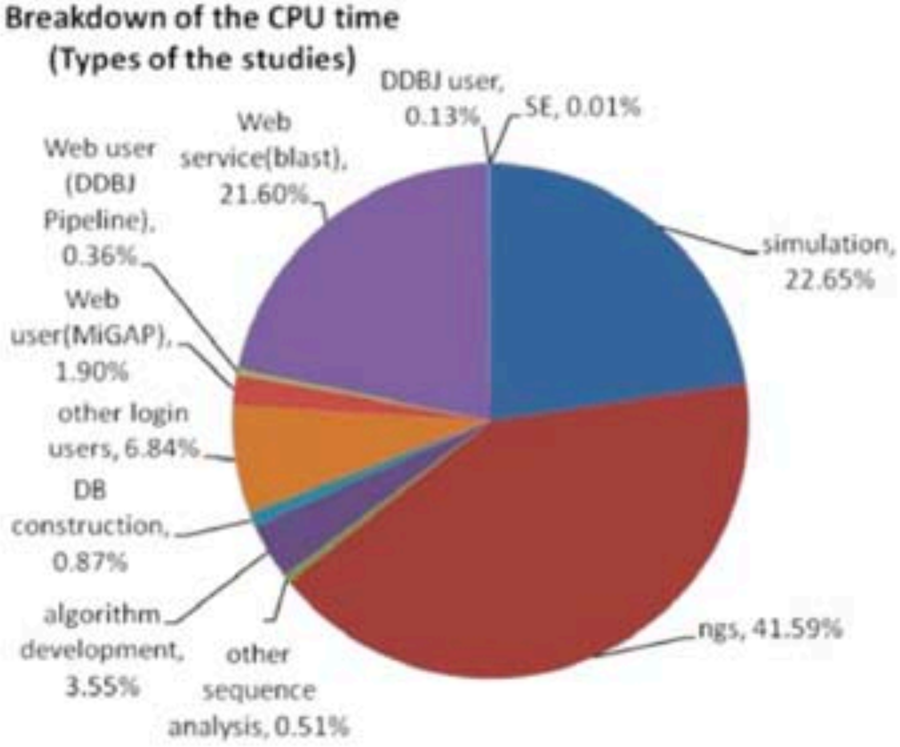
NIG Supercomputer System

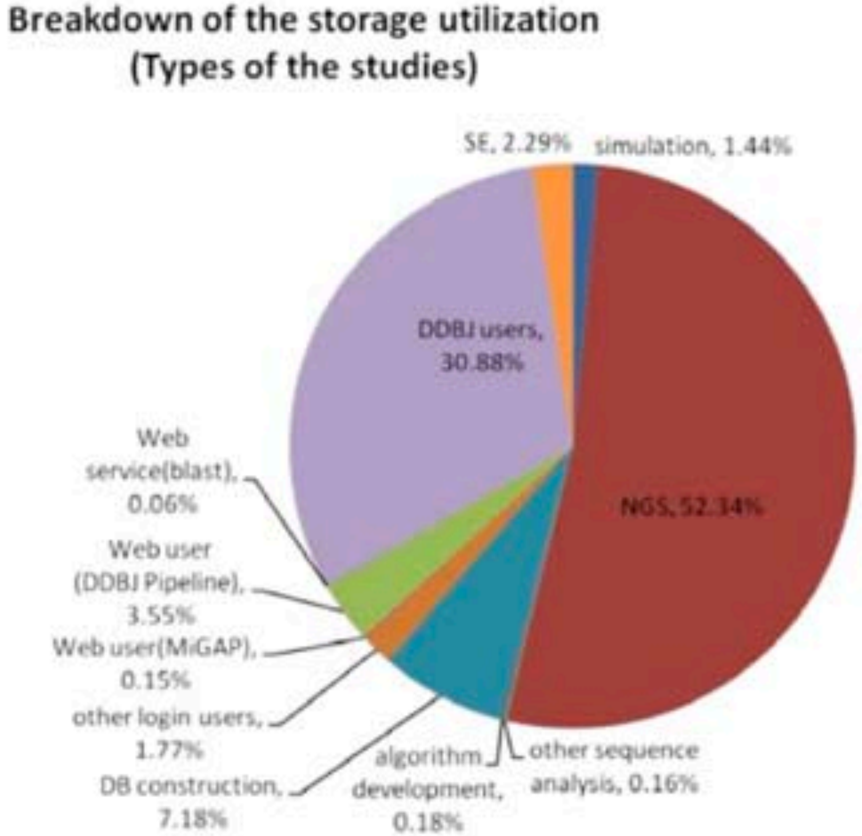


NIG Supercomputer System


Breakdown of the CPU time (Types of the accounts)

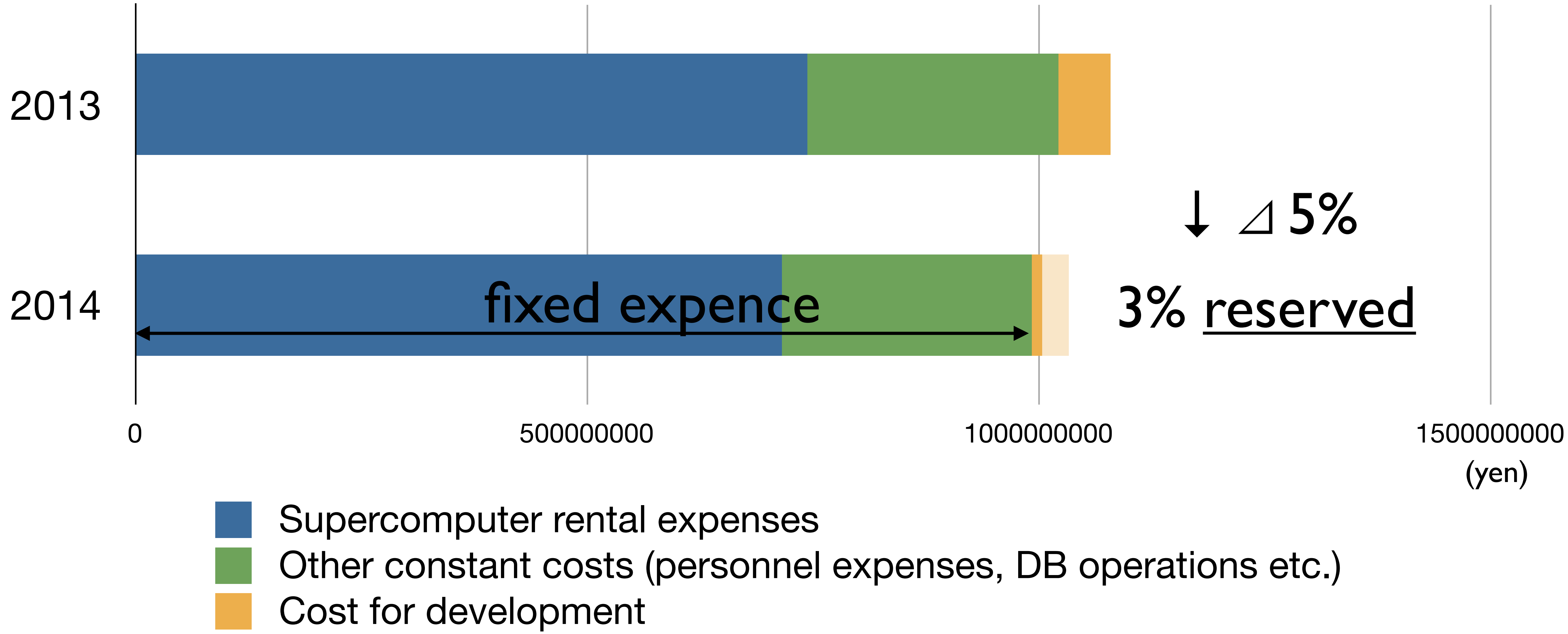

Breakdown of the storage utilization (Types of the accounts)


Breakdown of the CPU time (Types of the studies)


Breakdown of the storage utilization (Types of the studies)

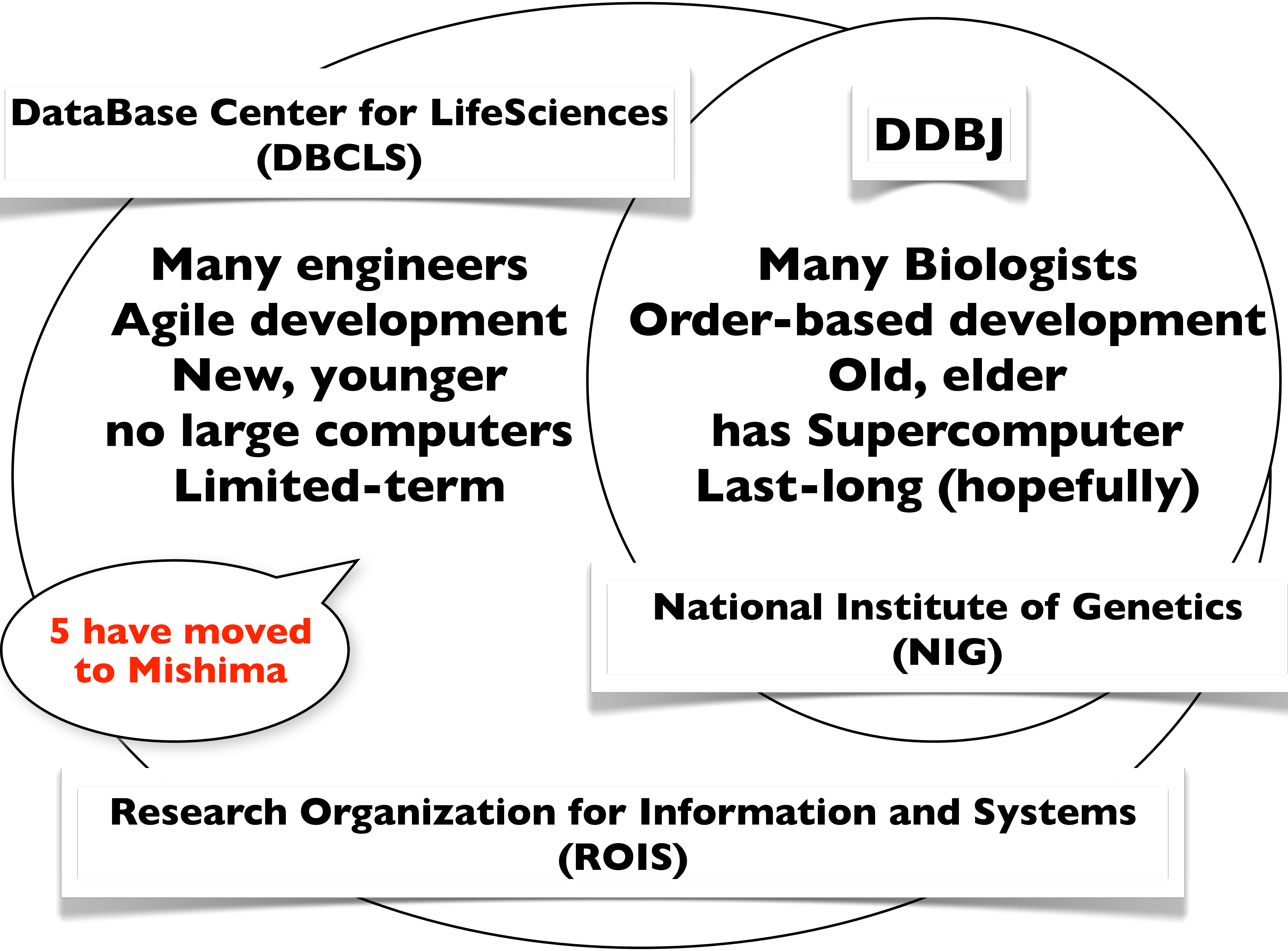
bad news

DDBJ Budget in FY 2014 (Apr. 2014 - Mar. 2015)



In a worst-case scenario, the cost for development in FY 2014 may be reduced to 1/5 (11,397 k yen) than that of FY2013 (57,785 k yen).

good news



DataBase Center for LifeSciences (DBCLS)

Many engineers
Agile development
New, younger
no large computers
Limited-term

DDBJ

Many Biologists
Order-based development
Old, elder
has Supercomputer
Last-long (hopefully)

5 have moved to Mishima

National Institute of Genetics (NIG)

Research Organization for Information and Systems (ROIS)

DDBJ efforts in FY 2014

- Based on INSDC cooperation, enhancement of the current services and development of new ones will be steadily performed.
- The JGA service for the controlled access data will be improved.
- Promoting collaboration of database research with NBDC and DBCLS* members and specific service developments based on needs from biological database communities
- Establishment of efficient operation after the secondary reinforcement of the NIG supercomputer.

* National Bioscience Database Center (http://biosciencedbc.jp) and DataBase Center for Life Science (http://dbcls.rois.ac.jp)

DB’s and Services for

NGS

example of the NGS’s

Roche (454): GS FLX+ System

illumina: Genome Analyzer IIx System

Life Technologies: 5500 xl SOLiD System

DDBJ Sequence Read Archive (DRA)

http://trace.ddbj.nig.ac.jp/dra/

Sequence Read Archive

News

2014-06-11 New: System temporarily unavailable less..

From 13:00 June 23th (Mon) until 12:00 June 26th (Thu), DRA, BioProject and BioSample submission and search services will be unavailable because of the NIG super computer's maintenance.

2014-05-13: New DRA submission system is released. less..

We have released the new DRA submission system. For major changes, please see the slides and new handbook.

(6th, June, 2014)
For submissions with status "new" which had been created before 12th, May. 2014, addition or deletion of metadata objects could cause errors. It is recommended that download metadata as a tab-delimited text file and upload it into a newly created submission.

DDBJ Sequence Read Archive Handbook

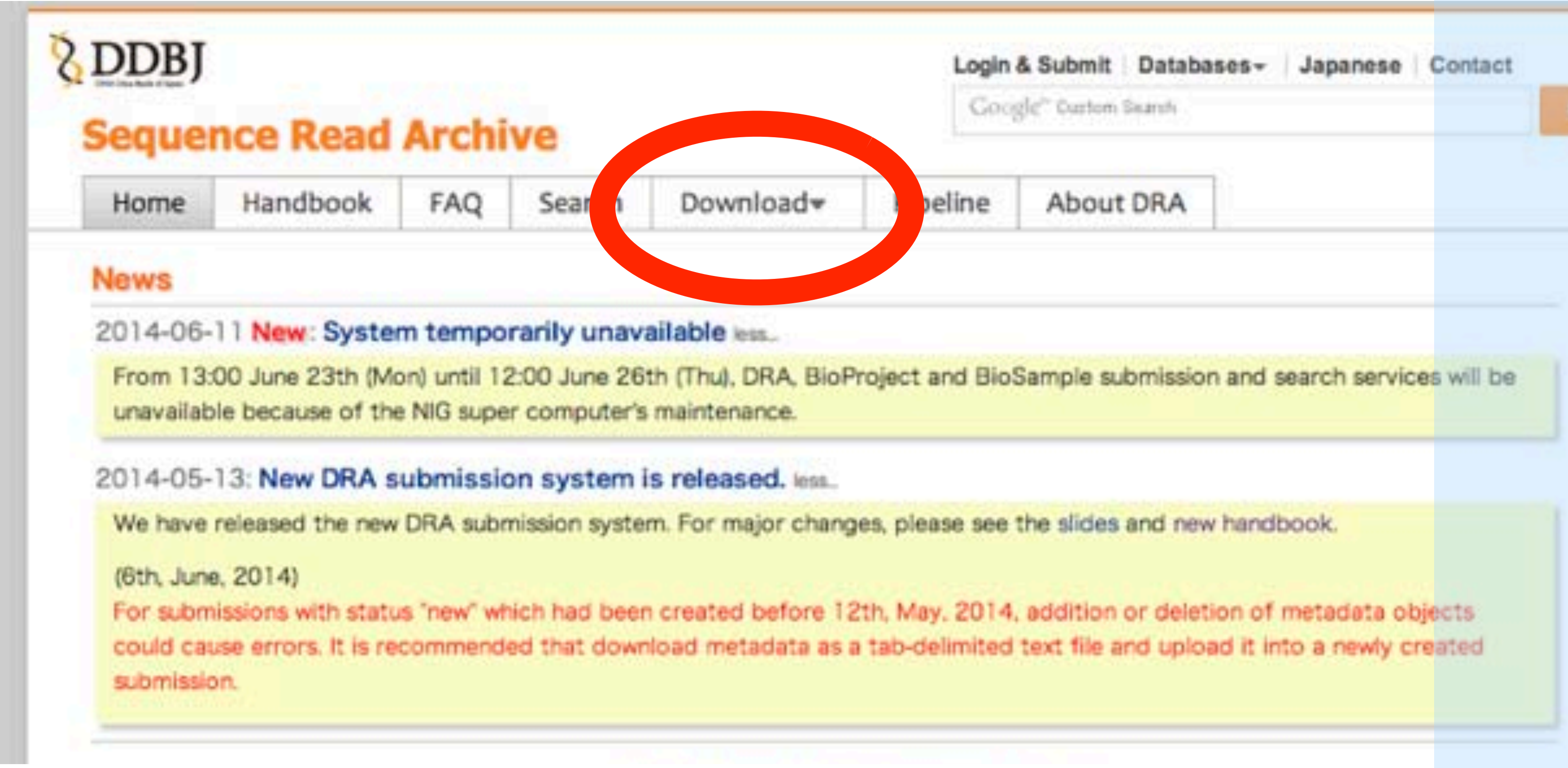
May 12th (Mon), 2014

New DRA submission system starts!

The DRA submission system will be reformed.

DDBJ Sequence Read Archive (DRA) is an archive database for output data generated by next-generation sequencing machines including Roche 454 GS System®, Illumina Genome Analyzer®, Applied Biosystems SOLiD® System, and others. DRA is a member of the International Nucleotide Sequence Database Collaboration (INSDC) and archiving the data in a close collaboration with NCBI Sequence Read Archive (SRA) and EBI Sequence Read Archive (ERA). Please submit the trace data from conventional capillary sequencers to DDBJ Trace Archive.

DRA: NGS data’s archive



Sequence Read Archive

News

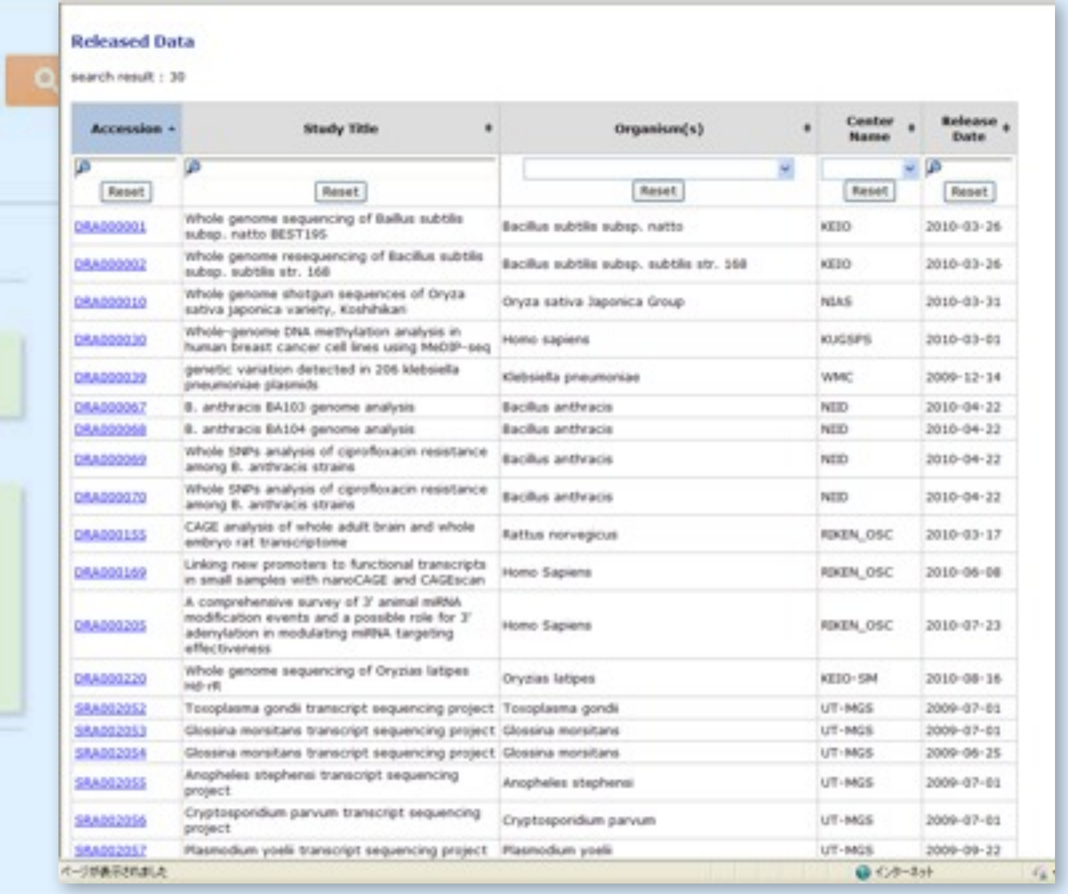
2014-06-11 **New: System temporarily unavailable** less..

From 13:00 June 23th (Mon) until 12:00 June 26th (Thu), DRA, BioProject and BioSample submission and search services will be unavailable because of the NIG super computer's maintenance.

2014-05-13: **New DRA submission system is released.** less..

We have released the new DRA submission system. For major changes, please see the slides and new handbook.

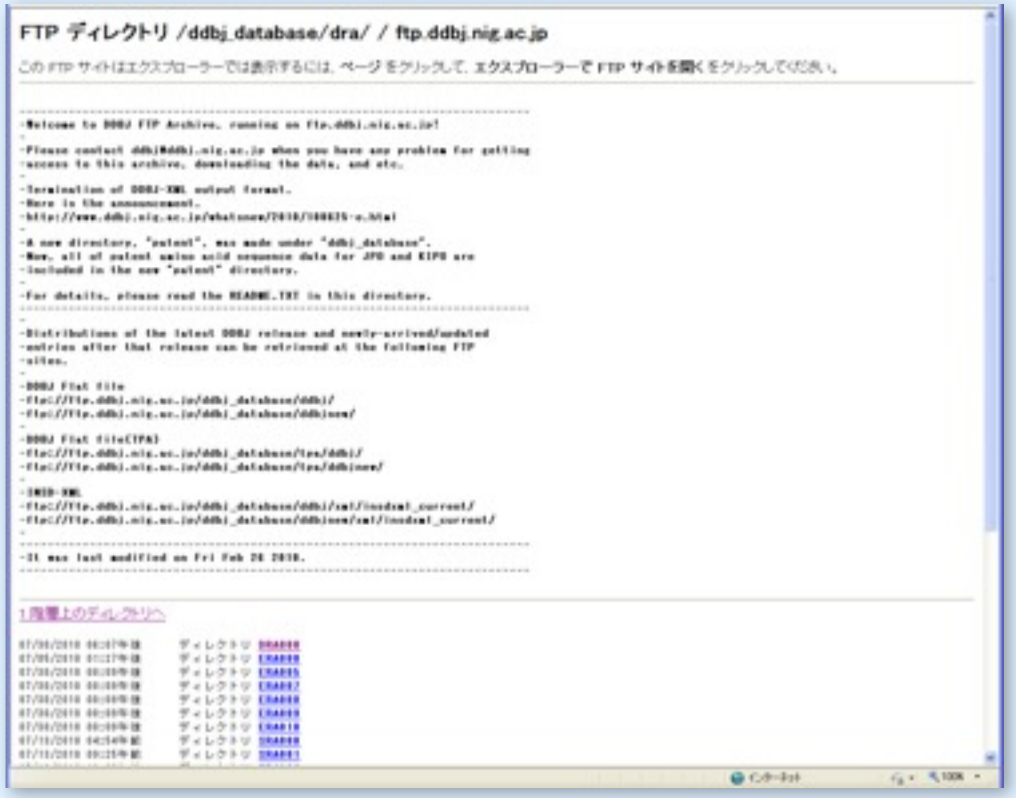
(6th, June, 2014)
For submissions with status "new" which had been created before 12th, May, 2014, addition or deletion of metadata objects could cause errors. It is recommended that download metadata as a tab-delimited text file and upload it into a newly created submission.



Released Data

search result : 30

Accession	Study Title	Organism(s)	Center Name	Release Date
DRA000001	Whole genome sequencing of Bailus subtilis subsp. natto BEST195	Bacillus subtilis subsp. natto	KEIO	2010-03-26
DRA000002	Whole genome resequencing of Bacillus subtilis subsp. subtilis str. 168	Bacillus subtilis subsp. subtilis str. 168	KEIO	2010-03-26
DRA000010	Whole genome shotgun sequences of Oryza sativa japonica variety, Koshihikari	Oryza sativa Japonica Group	NIAS	2010-03-31
DRA000030	Whole-genome DNA methylation analysis in human breast cancer cell lines using MeDIP-seq	Homo sapiens	KUGSPS	2010-03-01
DRA000039	genetic variation detected in 206 klebsiella pneumoniae plasmids	Klebsiella pneumoniae	WMC	2009-12-14
DRA000067	B. anthracis BA103 genome analysis	Bacillus anthracis	NIID	2010-04-22
DRA000068	B. anthracis BA104 genome analysis	Bacillus anthracis	NIID	2010-04-22
DRA000069	Whole SNPs analysis of ciprofloxacin resistance among B. anthracis strains	Bacillus anthracis	NIID	2010-04-22
DRA000070	Whole SNPs analysis of ciprofloxacin resistance among B. anthracis strains	Bacillus anthracis	NIID	2010-04-22
DRA000155	CAGE analysis of whole adult brain and whole embryo rat transcriptome	Rattus norvegicus	RIKEN_OSC	2010-03-17
DRA000169	Linking new promoters to functional transcripts in small samples with nanoCAGE and CAGEscan	Homo Sapiens	RIKEN_OSC	2010-06-08
DRA000205	A comprehensive survey of 3' animal miRNA modification events and a possible role for 3' adenylation in modulating miRNA targeting effectiveness	Homo Sapiens	RIKEN_OSC	2010-07-23
DRA000220	Whole genome sequencing of Oryzias latipes Hd-rR	Oryzias latipes	KEIO-SM	2010-08-16
SRA002052	Toxoplasma gondii transcript sequencing project	Toxoplasma gondii	UT-MGS	2009-07-01
SRA002053	Glossina morsitans transcript sequencing project	Glossina morsitans	UT-MGS	2009-07-01
SRA002054	Glossina morsitans transcript sequencing project	Glossina morsitans	UT-MGS	2009-06-25
SRA002055	Anopheles stephensi transcript sequencing project	Anopheles stephensi	UT-MGS	2009-07-01
SRA002056	Cryptosporidium parvum transcript sequencing project	Cryptosporidium parvum	UT-MGS	2009-07-01
SRA002057	Plasmodium yoelii transcript sequencing project	Plasmodium yoelii	UT-MGS	2009-09-22

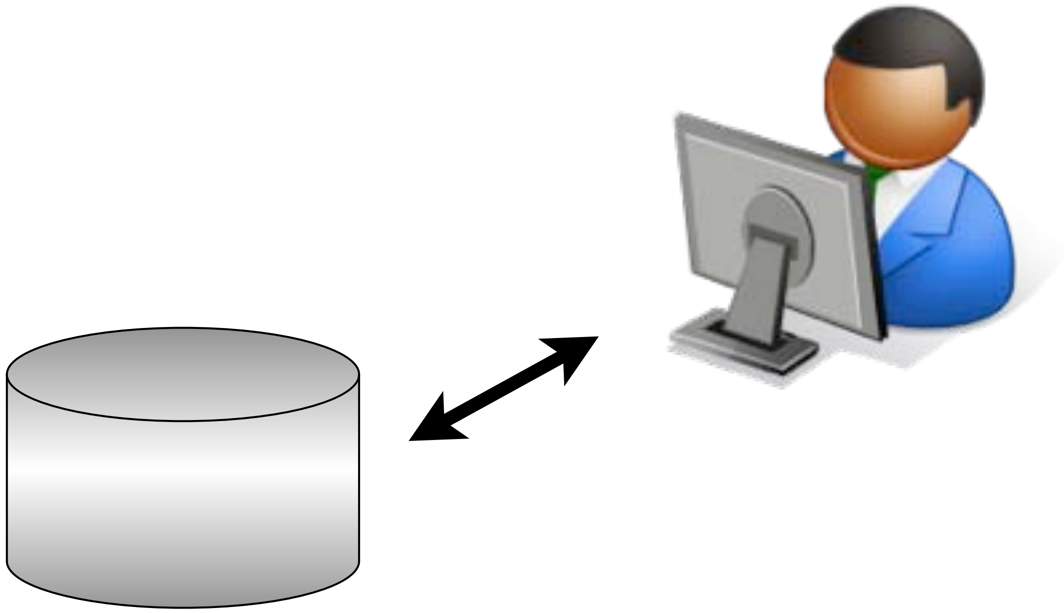


FTP ディレクトリ /ddbj_database/dra/ / ftp.ddbj.nig.ac.jp

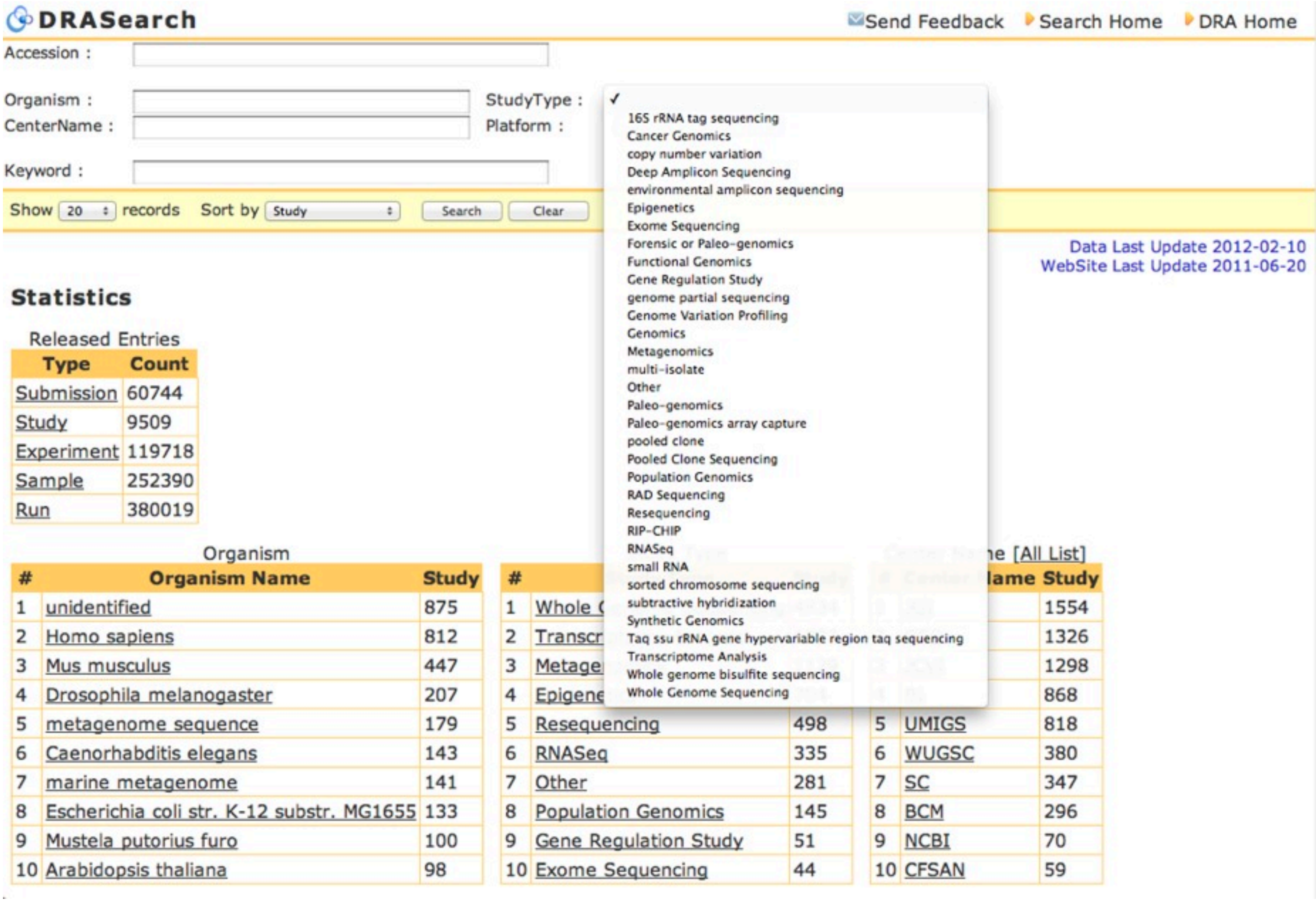
この FTP サイトはエクスプローラーでは表示するには、ページ をクリックして、エクスプローラーで FTP サイトを開く をクリックしてください。

```
-------------------------------------------------------------------------
-Welcome to DDBJ FTP Archive, running on ftp.ddbj.nig.ac.jp!
-
-Please contact ddbj@ddbj.nig.ac.jp when you have any problem for getting
-access to this archive, downloading the data, and etc.
-
-Termination of DDBJ-XML output format.
-Here is the announcement.
-http://www.ddbj.nig.ac.jp/whatsnew/2010/100625-e.html
-
-A new directory, "patent", was made under "ddbj_database".
-Now, all of patent amino acid sequence data for JPO and KIPO are
-included in the new "patent" directory.
-
-For details, please read the README.TXT in this directory.
-------------------------------------------------------------------------
-
-Distributions of the latest DDBJ release and newly-arrived/updated
-entries after that release can be retrieved at the following FTP
-sites.
-
-DDBJ Flat file
-ftp://ftp.ddbj.nig.ac.jp/ddbj_database/ddbj/
-ftp://ftp.ddbj.nig.ac.jp/ddbj_database/ddbjnew/
-
-DDBJ Flat file(TPA)
-ftp://ftp.ddbj.nig.ac.jp/ddbj_database/tpa/ddbj/
-ftp://ftp.ddbj.nig.ac.jp/ddbj_database/tpa/ddbjnew/
-
-INSD-XML
-ftp://ftp.ddbj.nig.ac.jp/ddbj_database/ddbj/xml/insdxml_current/
-ftp://ftp.ddbj.nig.ac.jp/ddbj_database/ddbjnew/xml/insdxml_current/
-
-------------------------------------------------------------------------
-It was last modified on Fri Feb 26 2010.
-------------------------------------------------------------------------
```

1階層上のディレクトリへ



DRAsearch system



Show 20 records Sort by Study

Data Last Update 2012-02-10
WebSite Last Update 2011-06-20

Statistics

Released Entries

Type	Count
Submission	60744
Study	9509
Experiment	119718
Sample	252390
Run	380019

Organism

#	Organism Name	Study
1	unidentified	875
2	Homo sapiens	812
3	Mus musculus	447
4	Drosophila melanogaster	207
5	metagenome sequence	179
6	Caenorhabditis elegans	143
7	marine metagenome	141
8	Escherichia coli str. K-12 substr. MG1655	133
9	Mustela putorius furo	100
10	Arabidopsis thaliana	98

#		
1	Whole	
2	Transcr	
3	Metage	
4	Epigene	
5	Resequencing	498
6	RNASeq	335
7	Other	281
8	Population Genomics	145
9	Gene Regulation Study	51
10	Exome Sequencing	44

ne [All List]

	ame	Study
		1554
		1326
		1298
		868
5	UMIGS	818
6	WUGSC	380
7	SC	347
8	BCM	296
9	NCBI	70
10	CFSAN	59

An user’s voice (tweets)

中村保一　博士（猫） @yaskaz 20時間
おーいぇー

愛ちゃん (本名) @dritoshi 21時間
公共のNGSデータを使う上での結論: DDBJ DRASearch から検索して、lftp script で DL が一番楽。SRA のサイトを使う必要はまったくない
中村保一　博士（猫）さんがリツイート

愛ちゃん (本名) @dritoshi 21時間
DDBJ DRA はやい!
中村保一　博士（猫）さんがリツイート

愛ちゃん (本名) @dritoshi 21時間
そしてわかりやすい!
中村保一　博士（猫）さんがリツイート

愛ちゃん (本名) @dritoshi 21時間
DRASearch めちゃべんり!
中村保一　博士（猫）さんがリツイート

Oh! Year!
(me)

Not nessessary to use SRA site.

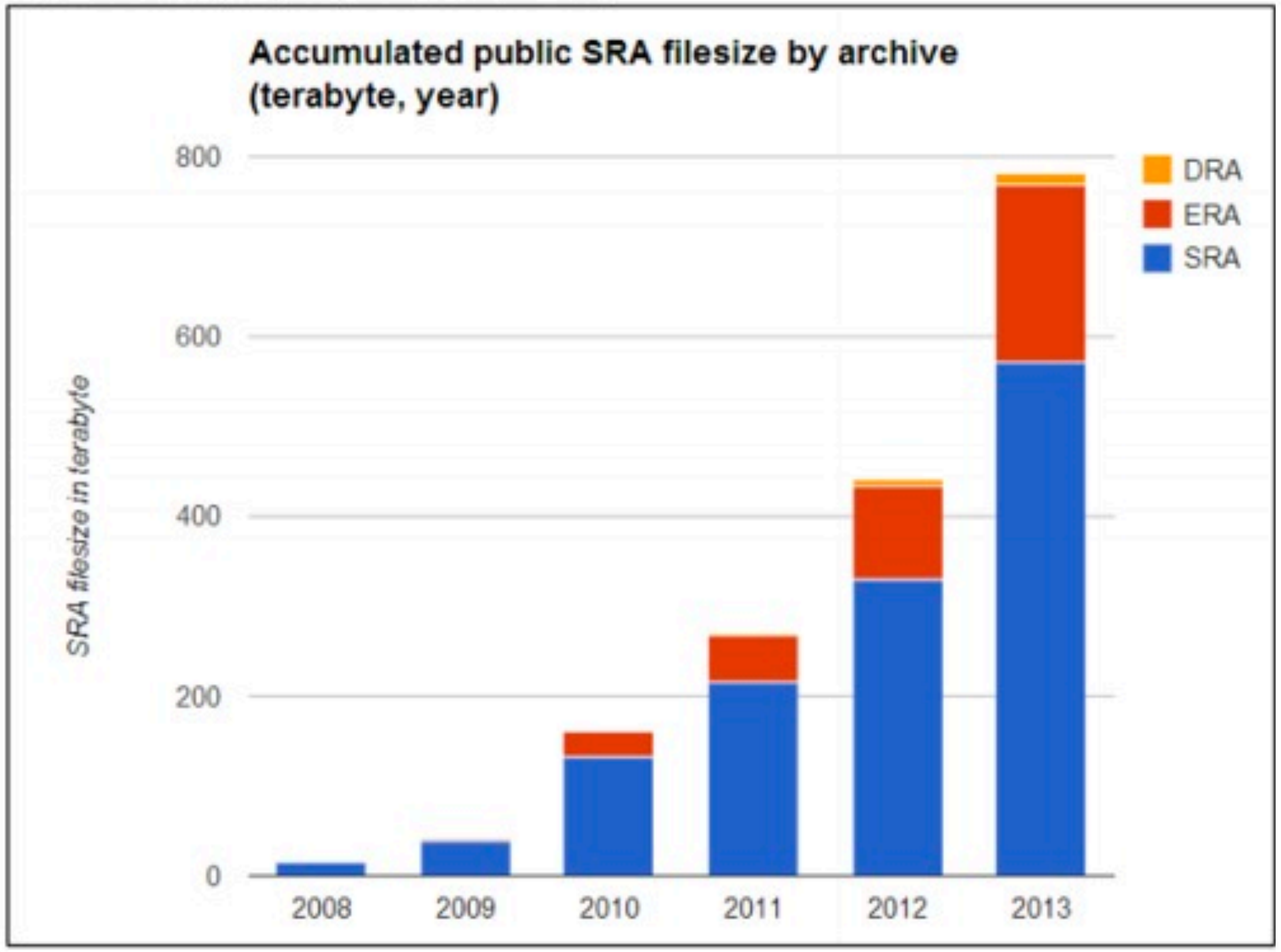
DDBJ DRA is fast to download!

Easy to understand!

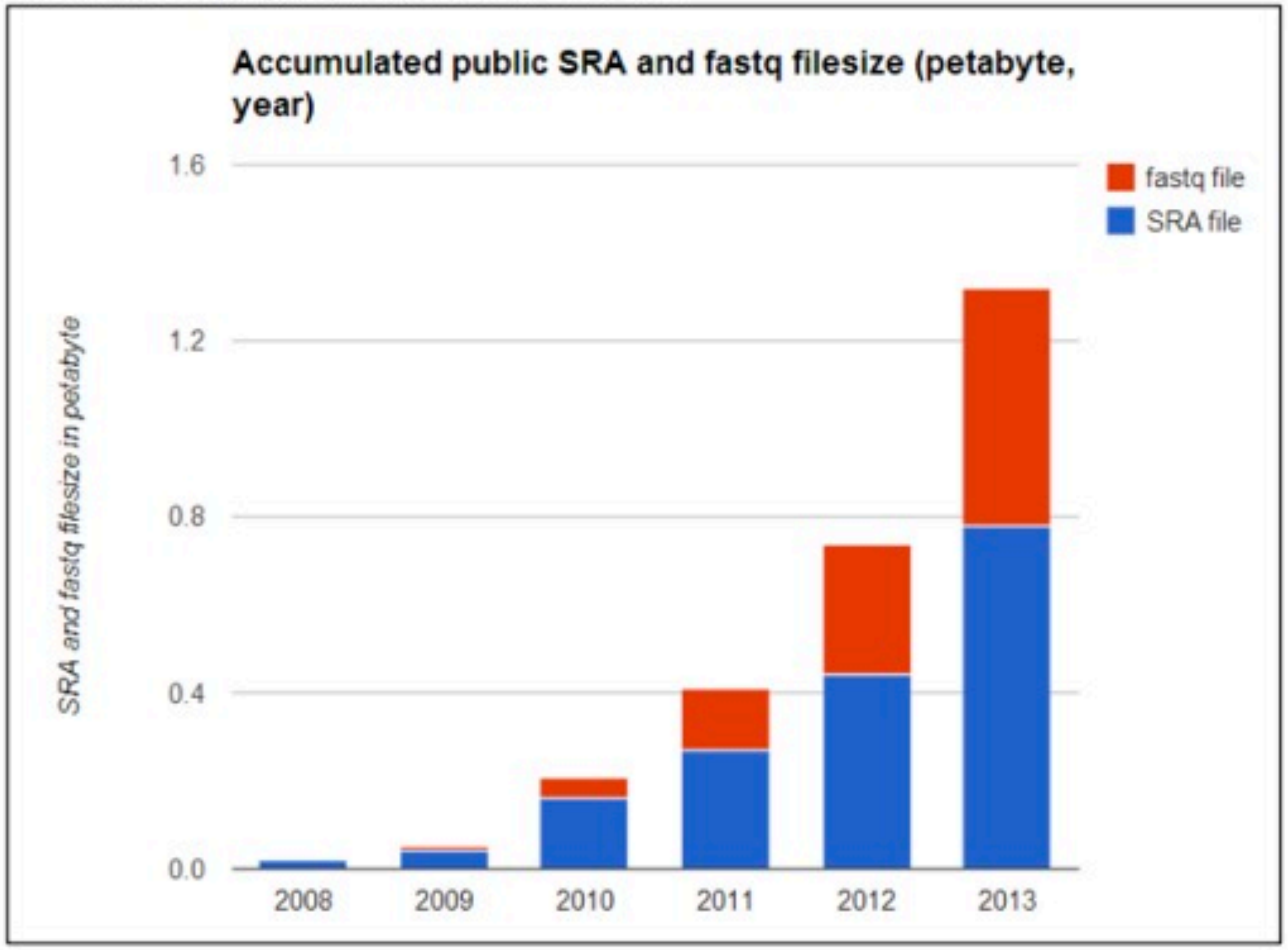
DRAsearch is extremely handy!

DRA's data amount: a problem in Japan

Statistics page as of 3th, February, 2014:
http://trace.ddbj.nig.ac.jp/stat_e.html#size

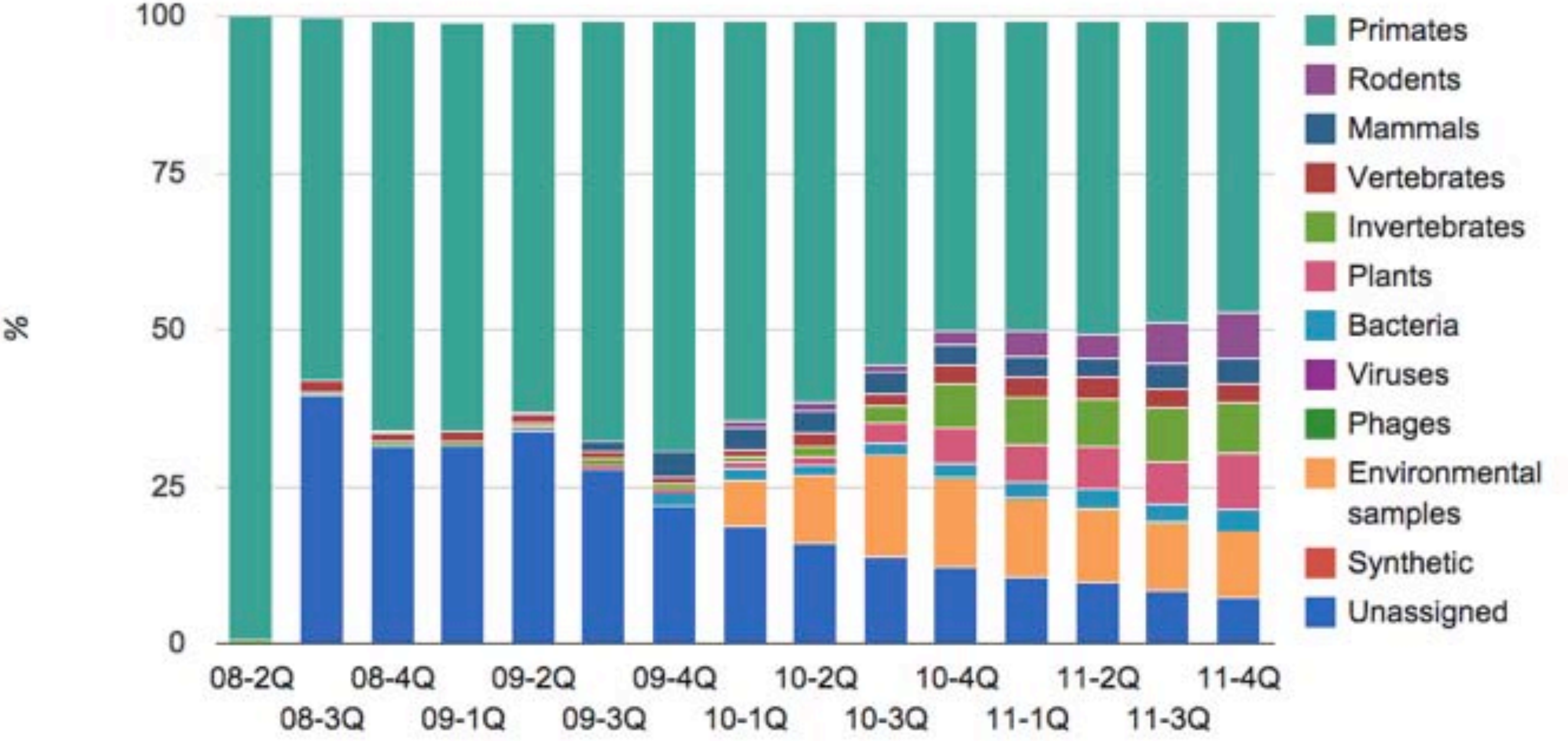


Statistics page as of 3th, February, 2014:
http://trace.ddbj.nig.ac.jp/stat_e.html#total_size



[DDBJ traditional + WGS
submission section: Kosuge, Okido, Lee, Tsutsui, Aono, Yokoyama, Ejima
update section: Sakai, Sugita, Mimura, Aono, Ejima
DRA: Kodama, Furuya
BioProject: Kodama, Furuya, Fukuda, Lee, Okido, Sakai, Yokoyama]

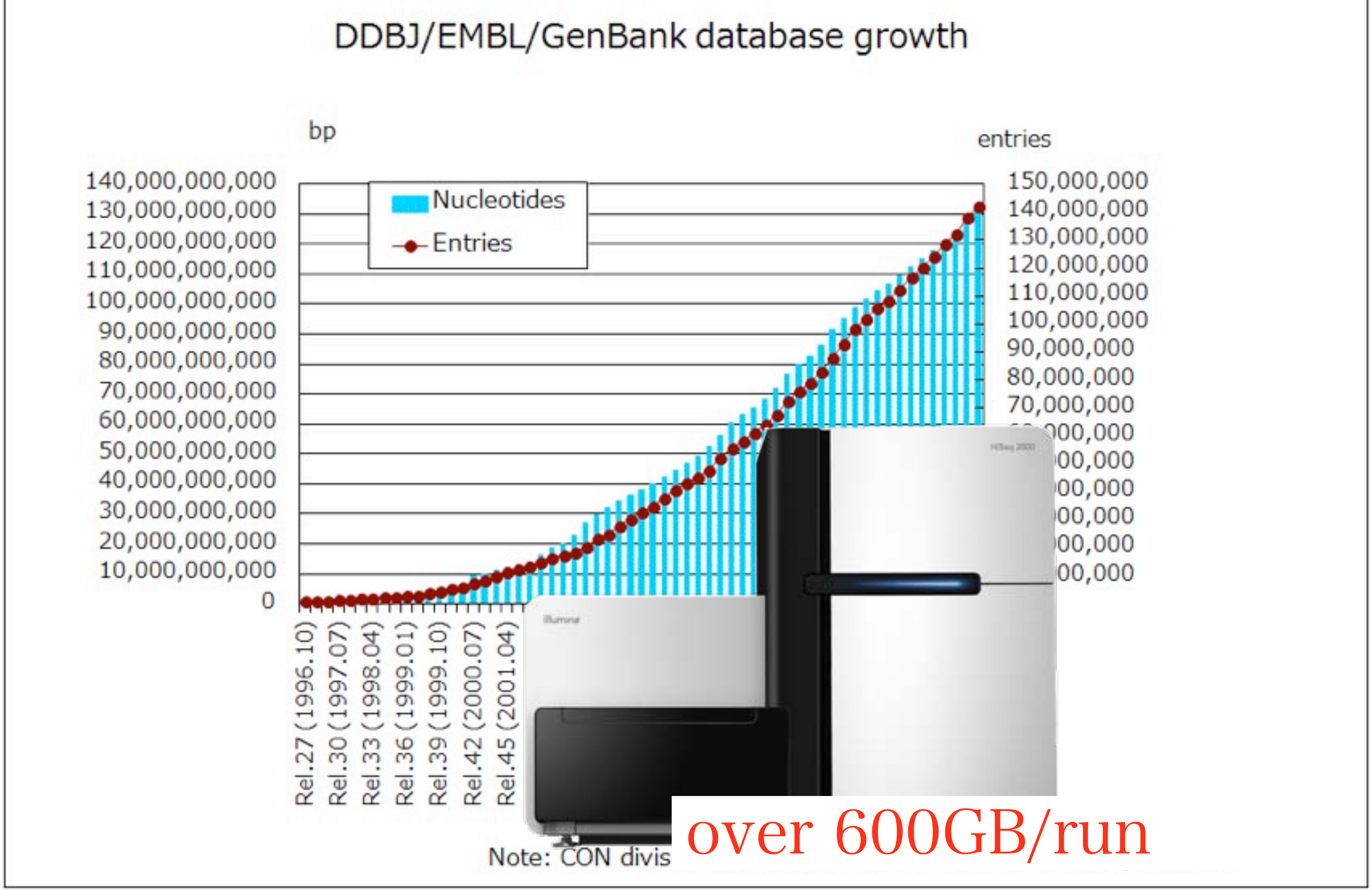
Taxonomic division in SRA (in TB)



http://trace.ddbj.nig.ac.jp/dra/stat_sra.shtml

Data explosion

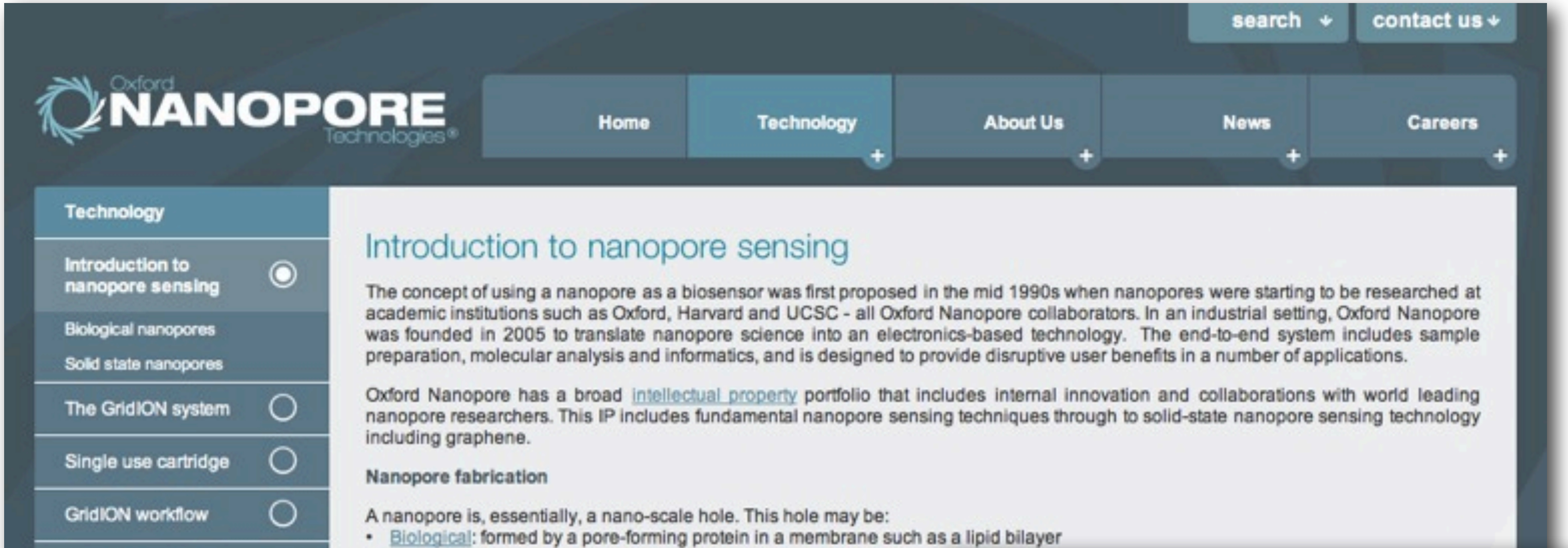
Trad DB: 150GB



over 600GB/run

“NANOPORE” sequencer will be emerged

http://www.nanoporetech.com/technology/introduction-to-nanopore-sensing/introduction-to-nanopore-sensing



Introduction to nanopore sensing

The concept of using a nanopore as a biosensor was first proposed in the mid 1990s when nanopores were starting to be researched at academic institutions such as Oxford, Harvard and UCSC - all Oxford Nanopore collaborators. In an industrial setting, Oxford Nanopore was founded in 2005 to translate nanopore science into an electronics-based technology. The end-to-end system includes sample preparation, molecular analysis and informatics, and is designed to provide disruptive user benefits in a number of applications.

Oxford Nanopore has a broad intellectual property portfolio that includes internal innovation and collaborations with world leading nanopore researchers. This IP includes fundamental nanopore sensing techniques through to solid-state nanopore sensing technology including graphene.

Nanopore fabrication

A nanopore is, essentially, a nano-scale hole. This hole may be:
- Biological: formed by a pore-forming protein in a membrane such as a lipid bilayer
- Solid-state: formed in synthetic materials such as silicon nitride or g
- Hybrid: formed by a pore-forming protein set in synthetic material

Nanopore sensing
A nanopore may be used to identify a target analyte as follows.

“MinION - $900 usb-powered DNA sequencer”

We are faced with “Big Data”

data

Strange things in the Refseq / nr DB

- protain (183) < protein
- imilar to (28) < similar to
- simila to (22) < similar to
- cromosome (4) < chromosome
- RNA olymerase < RNA polymerase
- dehydrogenas, ehydrogenase
- transposas, ransposase
- “2-Sep” for septin-2 < SEPT2

Copy & paste error!

>gi|91204169|emb|CAJ71822.1| strongly **imilar to** aspartate aminotransferase [Candidatus Kuenenia stuttgartiensis]
MIASRMSNIDSSGIRKVFDLAQKMKSPVNLSIGQPDFDVPGEIKEVAIKSINEGANKYTLTQGIPELRNV
...

>gi|31541577|gb|AAP56877.1| predicted methyl **transferas** [Mycoplasma gallisepticum R]
MSALYLVGLPIGNLSEINHRALEILNQLEIIYCENTDNFKKLLNLLNINFRDKKLISYHKFNETNRFIMI
...

similar to

transferase

“similar to similar to”

```
LOCUS       AL591981              347050 bp    DNA     linear   BCT 16-APR-2005
DEFINITION  Listeria monocytogenes strain EGD, complete genome, segment 9/12.
ACCESSION   AL591981 AL591824
VERSION     AL591981.1
KEYWORDS    .
SOURCE      Listeria monocytogenes
  ORGANISM  Listeria monocytogenes
            Bacteria; Firmicutes; Bacillales; Listeriaceae; Listeria.
REFERENCE   2  (bases 1 to 347050)
  AUTHORS   Glaser,P., Frangeul,L. and Rusniok,C.
  JOURNAL   Submitted (06-JUN-2001) to the EMBL/GenBank/DDBJ databases. Glaser
            P., Institut Pasteur, Genomique des Microorganismes Pathogenes, 25
            rue du Docteur Roux, 75724 Paris Cedex 15, FRANCE.

...
     CDS             complement(12915..14294)
                     /transl_table=11
                     /gene="lmo1703"
                     /note="similar to similar to RNA methyltransferases"
                     /db_xref="GOA:Q8Y6I1"
                     /db_xref="InterPro:IPR001566"
                     /db_xref="InterPro:IPR002792"
                     /db_xref="InterPro:IPR010280"
                     /db_xref="UniProtKB/Swiss-Prot:Q8Y6I1"
                     /protein_id="CAC99781.1"
                     /translation="MNQNPVEEGQKFPLTIRRMGINGEGIGYFKKAVVFVPGAITGEEV
                     VVEAVKVRDRFTEAKLNKIRKKSPNRVTAPCPVYEACGGCQLQHVAYSAQLELKRDIVI
                     QSIEKHTKIDPTKLKIRPTIGMEDPWRYRNKSQFQTRMVGSGQVETGLFGANSHQLVPI
                     EDCIVQQPVTIKVTNFVRDLLEKYGVPIYDEKAGSGIVRTIVVRTGVKTGETQLVFITN
                     SKKLPKKREMLAEIEAALPEVTSIMQNVNQAKSSLIFGDETFLLAGKESIEEKLMELEF
                     DLSARAFFQLNPFQTERLYQEVEKALVLTGSETLVDAYCGVGTIGQAFAGKVKEVRGMD
                     IIPESIEDAKRNAEKNGIENVYYEVGKAEDVLPKWVKEGFRPDAVIVDPPRSGCDQGLI
                     KSLLDVEAKQLVYVSCNPSTLARDLALLAKKYRIRYMQPVDMFPQTAHVETVVLLQLKD
                     K"
```

SEPT2 ⇒ 2-Sep case in Refseq

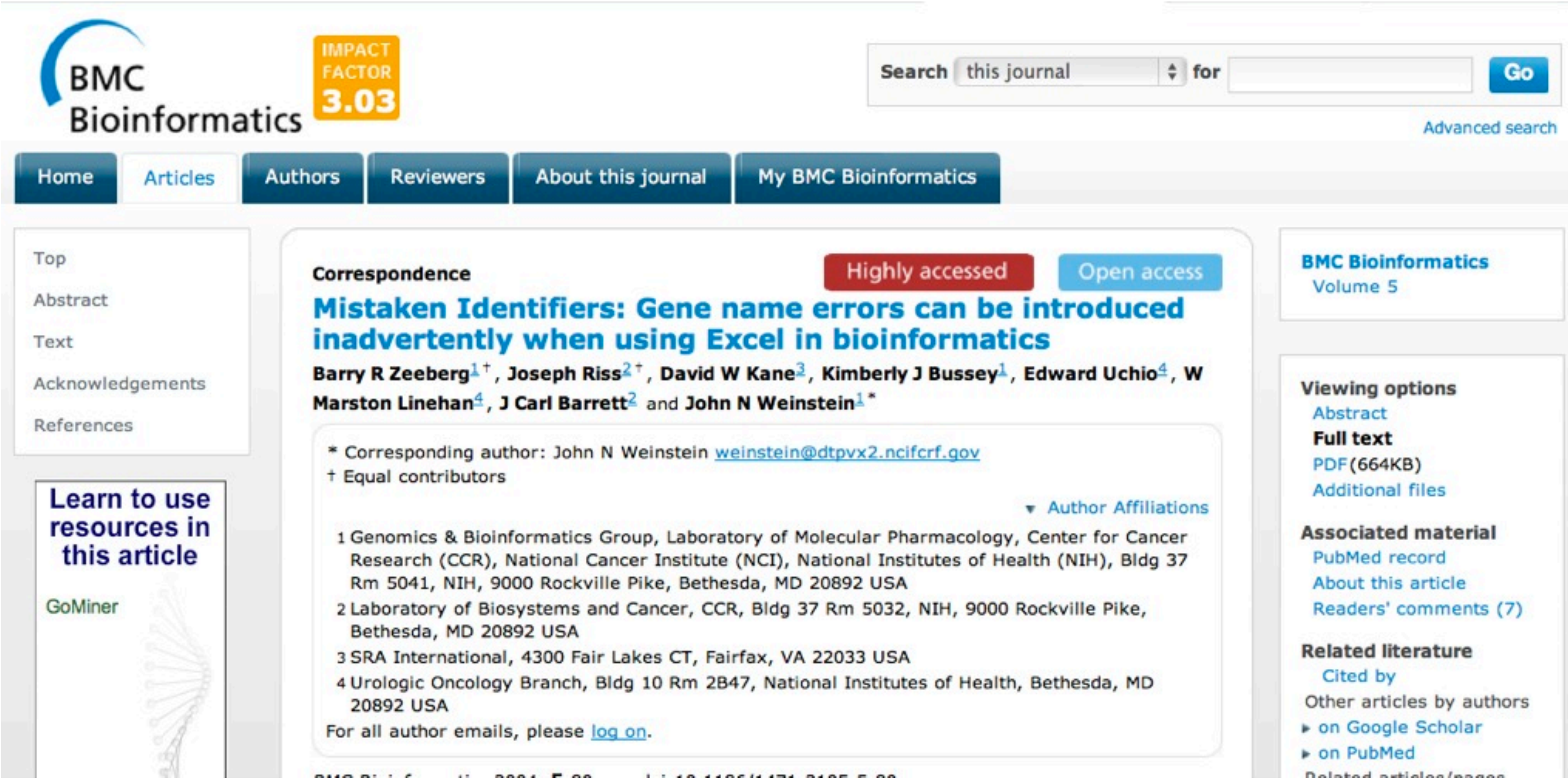
Introduced by MS-Excel's automatic correction

```
LOCUS       XM_392412               2125 bp    mRNA    linear   INV 12-APR-2011
DEFINITION  PREDICTED: Apis mellifera septin-2 (2-Sep), mRNA.
ACCESSION   XM_392412
VERSION     XM_392412.4  GI:328785636
KEYWORDS    .
SOURCE      Apis mellifera (honey bee)
  ORGANISM  Apis mellifera
            Eukaryota; Metazoa; Arthropoda; Hexapoda; Insecta; Pterygota;
            Neoptera; Endopterygota; Hymenoptera; Apocrita; Aculeata; Apoidea;
            Apidae; Apis.
COMMENT     MODEL REFSEQ:  This record is predicted by automated computational
            analysis. This record is derived from a genomic sequence
            (NW_003378075) annotated using gene prediction method: GNOMON,
            supported by EST evidence.
            Also see:
                Documentation of NCBI's Annotation Process

            On Apr 12, 2011 this sequence version replaced gi:110757583.
FEATURES             Location/Qualifiers
     source          1..2125
                     /organism="Apis mellifera"
                     /mol_type="mRNA"
                     /strain="DH4"
                     /db_xref="taxon:7460"
                     /linkage_group="LG6"
     gene            1..2125
                     /gene="2-Sep"
                     /note="Derived by automated computational analysis using
                     gene prediction method: GNOMON. Supporting evidence
                     includes similarity to: 436 ESTs, 11 Proteins"
                     /db_xref="BEEBASE:GB17411"
                     /db_xref="GeneID:408882"
     misc_feature    164..166
                     /gene="2-Sep"
                     /note="upstream in-frame stop codon"
     CDS             194..1444
                     /gene="2-Sep"
                     /codon_start=1
                     /product="septin-2"
                     /protein_id="XP_392412.2"
```

http://www.biomedcentral.com/1471-2105/5/80

Identifier “mutation” by Excel (2-Sep)



Correspondence

Mistaken Identifiers: Gene name errors can be introduced inadvertently when using Excel in bioinformatics

Barry R Zeeberg[1]†, **Joseph Riss**[2]†, **David W Kane**[3], **Kimberly J Bussey**[1], **Edward Uchio**[4], **W Marston Linehan**[4], **J Carl Barrett**[2] and **John N Weinstein**[1]*

* Corresponding author: John N Weinstein weinstein@dtpvx2.ncifcrf.gov
† Equal contributors

▾ Author Affiliations

1 Genomics & Bioinformatics Group, Laboratory of Molecular Pharmacology, Center for Cancer Research (CCR), National Cancer Institute (NCI), National Institutes of Health (NIH), Bldg 37 Rm 5041, NIH, 9000 Rockville Pike, Bethesda, MD 20892 USA
2 Laboratory of Biosystems and Cancer, CCR, Bldg 37 Rm 5032, NIH, 9000 Rockville Pike, Bethesda, MD 20892 USA
3 SRA International, 4300 Fair Lakes CT, Fairfax, VA 22033 USA
4 Urologic Oncology Branch, Bldg 10 Rm 2B47, National Institutes of Health, Bethesda, MD 20892 USA

For all author emails, please log on.

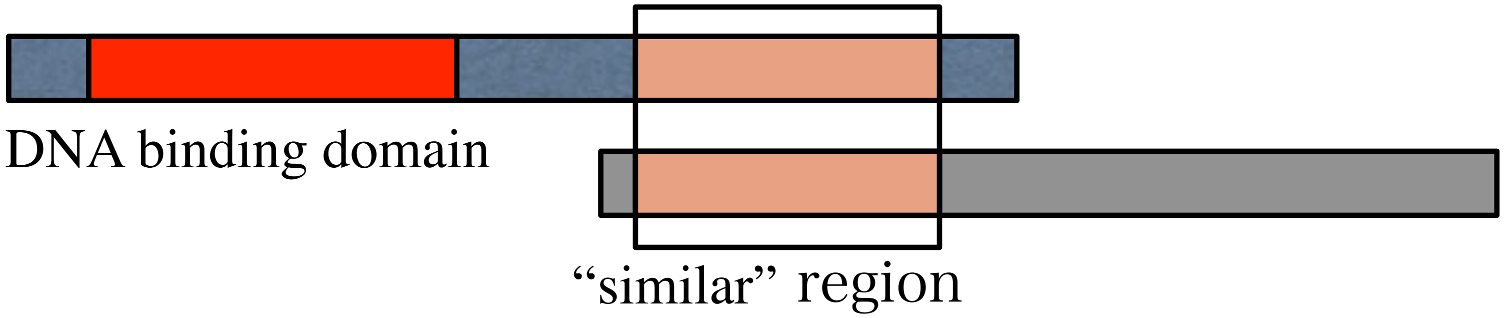
BMC Bioinformatics 2004, **5**:80 doi:10.1186/1471-2105-5-80

How to Avoid such stupid Errors?

- Good Reference Sequences
- Good Reference Annotations
- Fully Automated Annotation process
  - Ontology for gene and metadata
  - Rule-based gene description

- No copy and paste by hand
- No auto-correction by Excel

“contains similarity” ?



DNA binding protein

unknown protein

automated annotation:

“contains similarity to <u>DNA-binding protein</u>”

↑ wothout the domain!

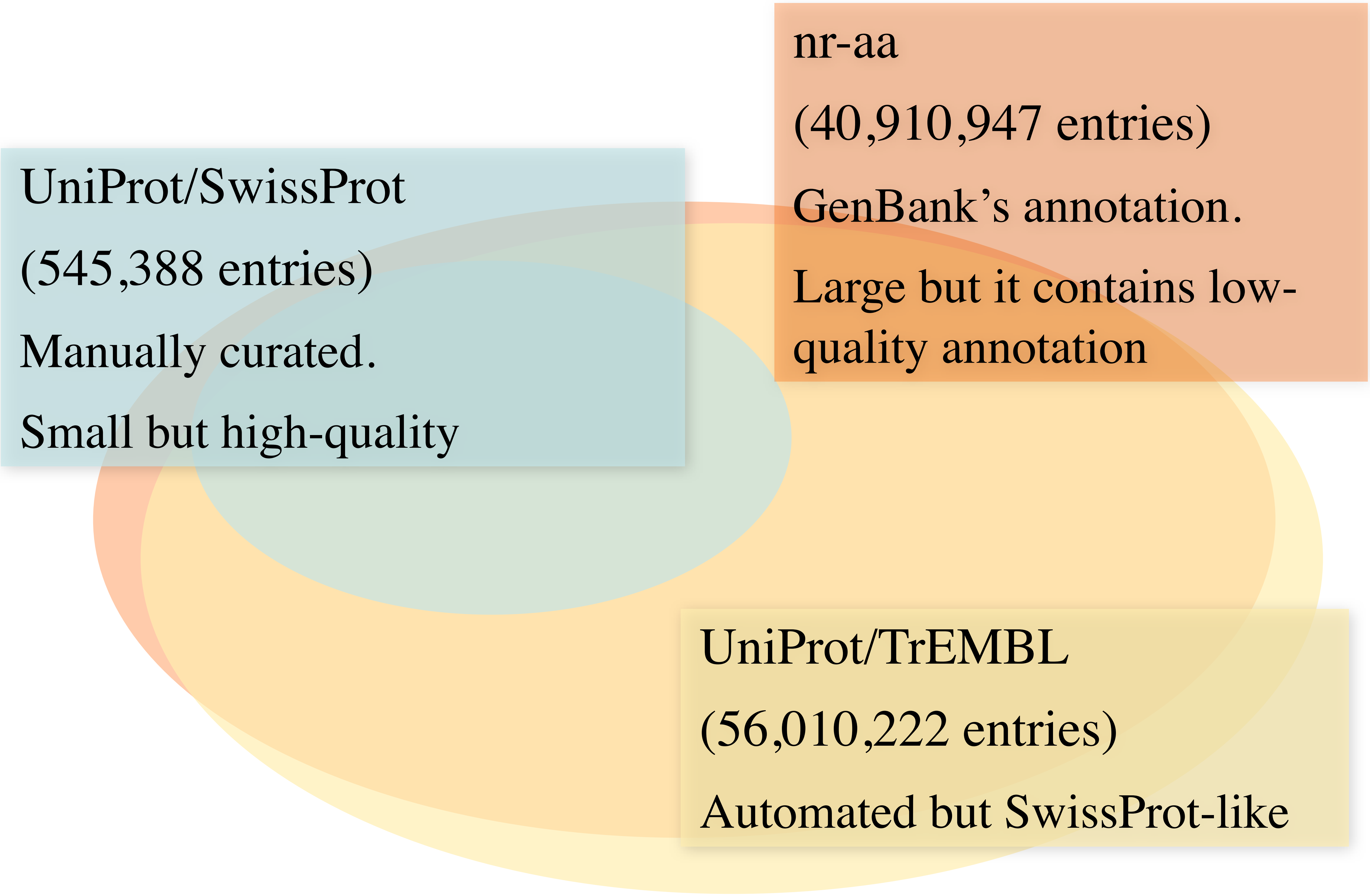
How to avoid the function inference problem?

- Use a good library
- Use functionally known patterns
- Use annotations with ‘evidence’

How to avoid the function inference problem?

- Use a good library
- Use functionally known patterns
- Use annotations with ‘evidence’

Use a small but better library



nr-aa

(40,910,947 entries)

GenBank’s annotation.

Large but it contains low-quality annotation

UniProt/SwissProt

(545,388 entries)

Manually curated.

Small but high-quality

UniProt/TrEMBL

(56,010,222 entries)

Automated but SwissProt-like

How to avoid the function inference problem?

- Use a good library
- Use functionally known patterns
- Use annotations with ‘evidence’

Motif databases

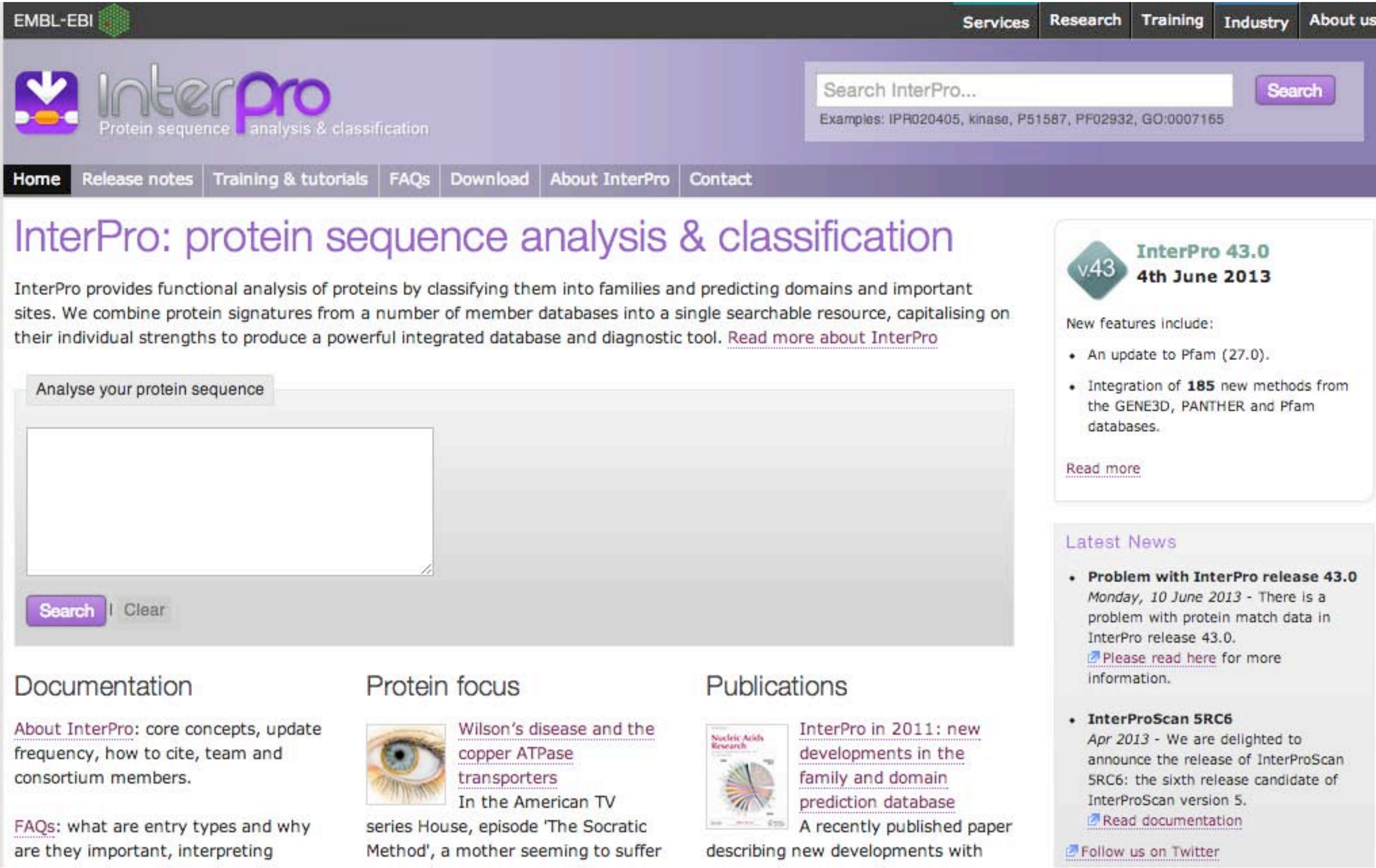
- PROSITE
  - http://www.expasy.ch/prosite/
- BLOCKS
  - http://www.blocks.fhcrc.org/
- PRINTS
  - http://bioinf.man.ac.uk/dbbrowser/PRINTS/PRINTS.html

Domain databases

- ProDom
  - http://prodes.toulouse.inra.fr/prodom/doc/prodom.html
- Pfam
  - http://www.sanger.ac.uk/Software/Pfam/
- SMART
  - http://smart.embl-heidelberg.de/
- TIGRFAMs
  - http://www.tigr.org/TIGRFAMs/

InterPro

- The integrated database for motifs and domains
- Pfam, PRINTS, PROSITE… and InterPro’s own anntation



Examples: IPR020405, kinase, P51587, PF02932, GO:0007165

InterPro: protein sequence analysis & classification

InterPro provides functional analysis of proteins by classifying them into families and predicting domains and important sites. We combine protein signatures from a number of member databases into a single searchable resource, capitalising on their individual strengths to produce a powerful integrated database and diagnostic tool. Read more about InterPro

Analyse your protein sequence

InterPro 43.0
4th June 2013

New features include:

- An update to Pfam (27.0).
- Integration of **185** new methods from the GENE3D, PANTHER and Pfam databases.

Read more

Documentation

About InterPro: core concepts, update frequency, how to cite, team and consortium members.

FAQs: what are entry types and why are they important, interpreting

Protein focus

Wilson's disease and the copper ATPase transporters
In the American TV series House, episode 'The Socratic Method', a mother seeming to suffer

Publications

InterPro in 2011: new developments in the family and domain prediction database
A recently published paper describing new developments with

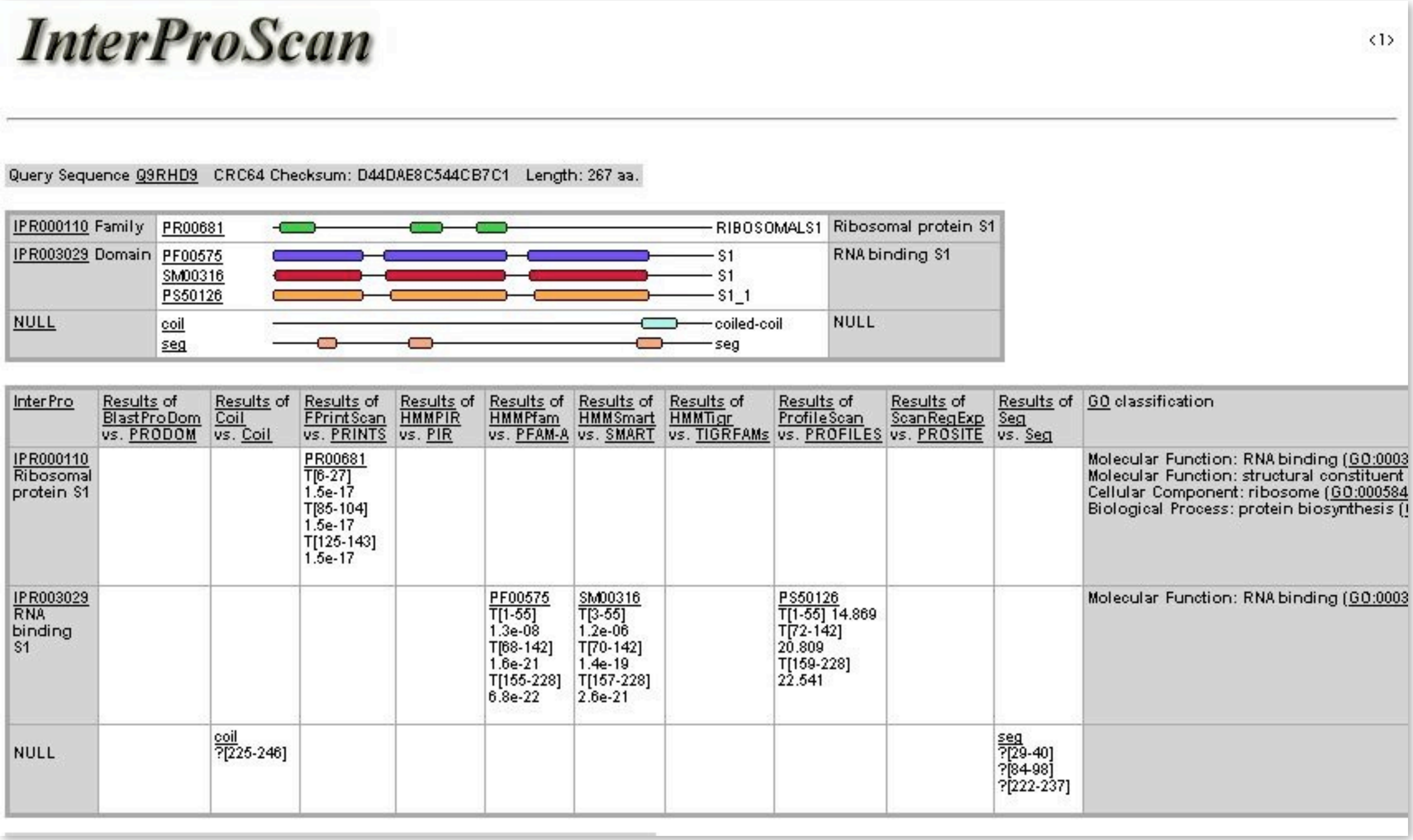
Latest News

- **Problem with InterPro release 43.0**
  Monday, 10 June 2013 - There is a problem with protein match data in InterPro release 43.0.
  Please read here for more information.
- **InterProScan 5RC6**
  Apr 2013 - We are delighted to announce the release of InterProScan 5RC6: the sixth release candidate of InterProScan version 5.
  Read documentation

Follow us on Twitter

http://www.ebi.ac.uk/interpro

InterProScan’s output



InterProScan

<1>

Query Sequence Q9RHD9 CRC64 Checksum: D44DAE8C544CB7C1 Length: 267 aa.

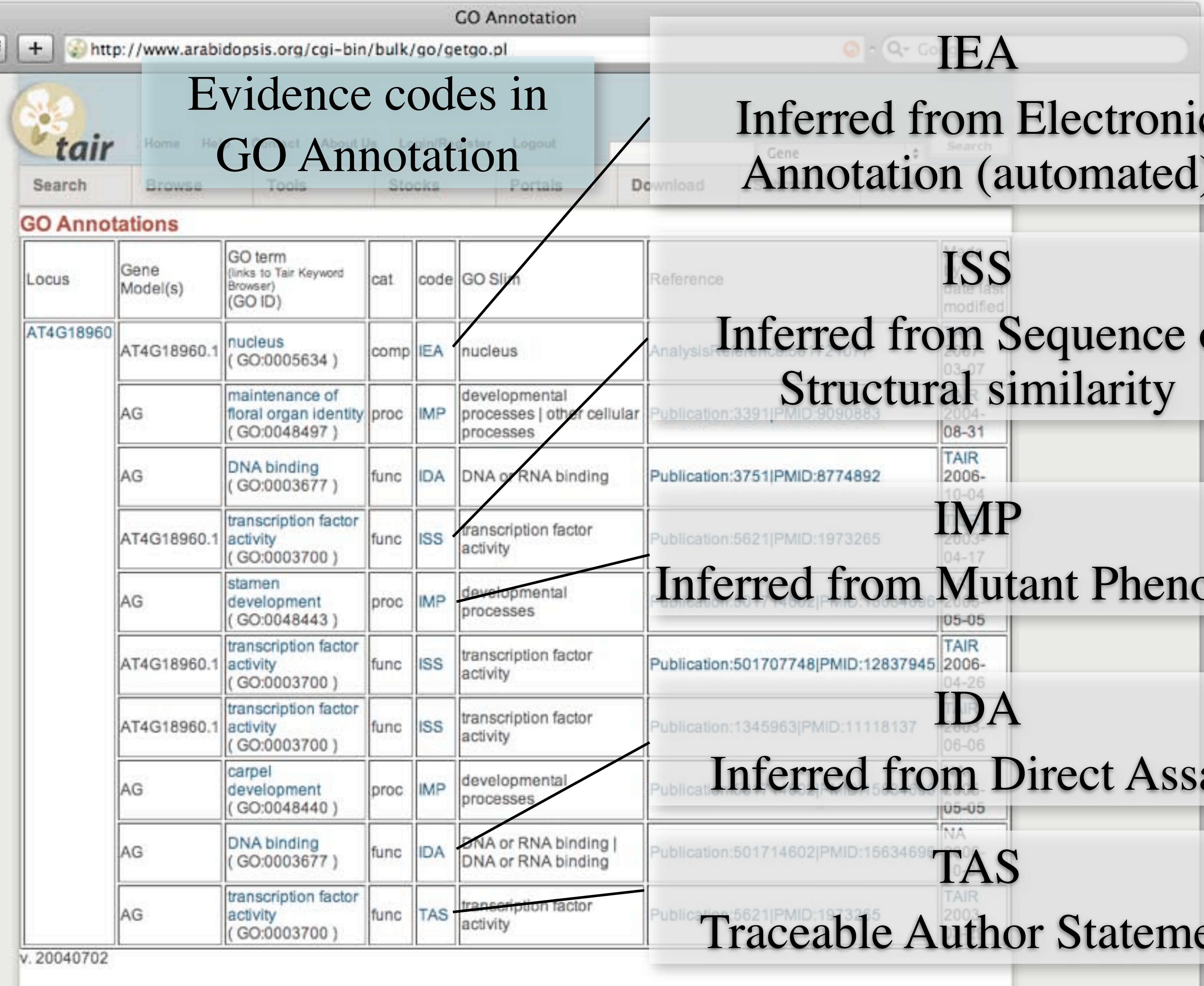
IPR000110 Family	PR00681	RIBOSOMALS1	Ribosomal protein S1
IPR003029 Domain	PF00575 SM00316 PS50126	S1 S1 S1_1	RNA binding S1
NULL	coil seg	coiled-coil seg	NULL

InterPro	Results of BlastProDom vs. PRODOM	Results of Coil vs. Coil	Results of FPrintScan vs. PRINTS	Results of HMMPIR vs. PIR	Results of HMMPfam vs. PFAM-A	Results of HMMSmart vs. SMART	Results of HMMTigr vs. TIGRFAMs	Results of ProfileScan vs. PROFILES	Results of ScanRegExp vs. PROSITE	Results of Seg vs. Seg	GO classification
IPR000110 Ribosomal protein S1			PR00681 T[6-27] 1.5e-17 T[85-104] 1.5e-17 T[125-143] 1.5e-17								Molecular Function: RNA binding (GO:0003 Molecular Function: structural constituent Cellular Component: ribosome (GO:000584 Biological Process: protein biosynthesis (
IPR003029 RNA binding S1					PF00575 T[1-55] 1.3e-08 T[68-142] 1.6e-21 T[155-228] 6.8e-22	SM00316 T[3-55] 1.2e-06 T[70-142] 1.4e-19 T[157-228] 2.6e-21		PS50126 T[1-55] 14.869 T[72-142] 20.809 T[159-228] 22.541			Molecular Function: RNA binding (GO:0003
NULL		coil ?[225-246]								seg ?[29-40] ?[84-98] ?[222-237]	

How to avoid the function inference problem?

- Use a good library
- Use functionally known patterns
- Use annotations with ‘evidence’

Evidence codes in GO Annotation



GO Annotations

Locus	Gene Model(s)	GO term (links to Tair Keyword Browser) (GO ID)	cat	code	GO Slim	Reference	Made [illegible] date last modified
AT4G18960	AT4G18960.1	nucleus (GO:0005634)	comp	IEA	nucleus	AnalysisReference:[illegible]	[illegible] 2007-03-07
	AG	maintenance of floral organ identity (GO:0048497)	proc	IMP	developmental processes | other cellular processes	Publication:3391|PMID:9090883	[illegible] 2004-08-31
	AG	DNA binding (GO:0003677)	func	IDA	DNA or RNA binding	Publication:3751|PMID:8774892	TAIR 2006-10-04
	AT4G18960.1	transcription factor activity (GO:0003700)	func	ISS	transcription factor activity	Publication:5621|PMID:1973265	[illegible] 2003-04-17
	AG	stamen development (GO:0048443)	proc	IMP	developmental processes	Publication:[illegible]	[illegible] 05-05
	AT4G18960.1	transcription factor activity (GO:0003700)	func	ISS	transcription factor activity	Publication:501707748|PMID:12837945	TAIR 2006-04-26
	AT4G18960.1	transcription factor activity (GO:0003700)	func	ISS	transcription factor activity	Publication:1345963|PMID:11118137	[illegible] 06-06
	AG	carpel development (GO:0048440)	proc	IMP	developmental processes	[illegible]	[illegible] 05-05
	AG	DNA binding (GO:0003677)	func	IDA	DNA or RNA binding | DNA or RNA binding	Publication:501714602|PMID:15634696	NA [illegible]
	AG	transcription factor activity (GO:0003700)	func	TAS	transcription factor activity	Publication:5621|PMID:1973265	TAIR 2003[illegible]

v. 20040702

IEA

Inferred from Electronic Annotation (automated)

ISS

Inferred from Sequence or Structural similarity

IMP

Inferred from Mutant Phenotype

IDA

Inferred from Direct Assay

TAS

Traceable Author Statement

cloud

+

crowd

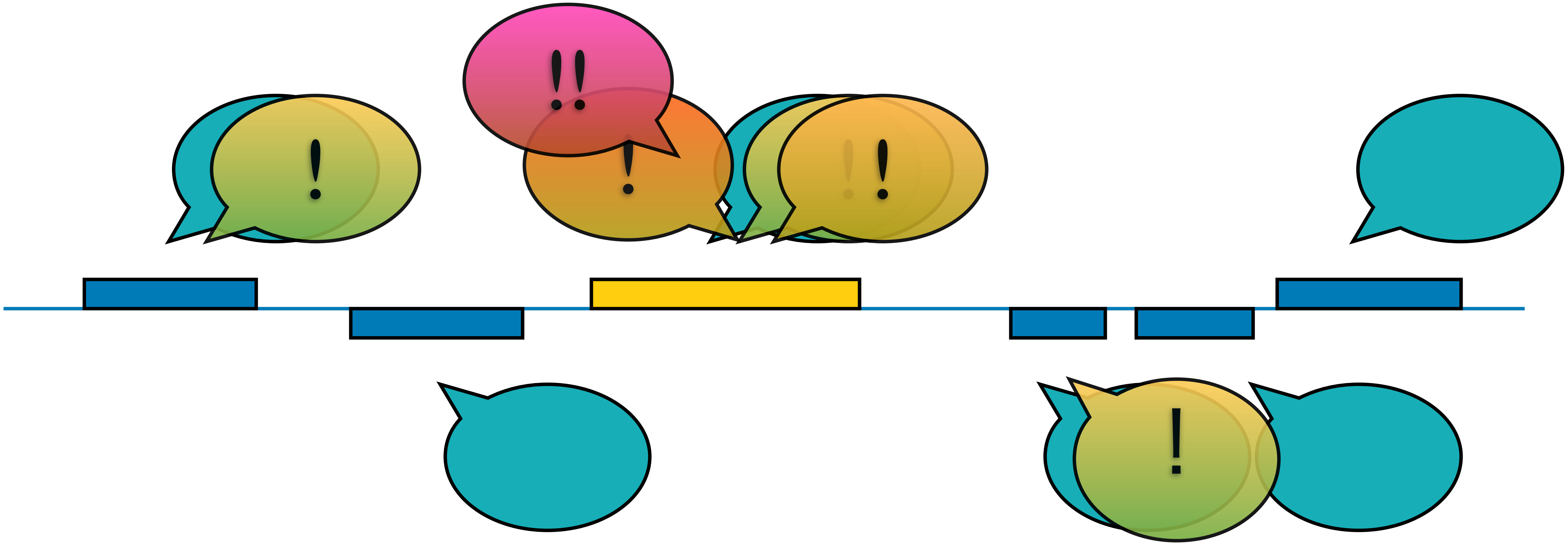
We need:

- good references
- good pipelines

We need:

- good references
- good pipelines

TogoAnnotation: a social genome annotation platform



http://togo.annotation.jp

http://togo.annotation.jp

About TogoAnnotation

統合アノテーションは、ソーシャルブックマークのしくみを利用して生物のデータに様々な注釈（アノテーション）をつけることが出来るサービスです。

News

統合アノテーション
TogoAnnotation

TogoAnnotation TogoAnnotation (KazusaAnnotation)を用いてコミュニティゲノムアノテーションを実施したBradyrhizobium sp. S23321の論文、配列が公開されました。 jstage.jst.go.jp/article/jsme2/... getentry.ddbj.nig.ac.jp/getentry?acces...
25 days ago · reply · retweet · favorite

bonohu TogoAnnotation祭――――(ﾟ∀ﾟ)――――!! (#AJACS live at ustre.am/1x4x/1)
38 days ago · reply · retweet · favorite

synobu 19種の光合成関連生物から約8000報の文献をキュレーターが読んで20万以上の文献中の遺伝子名記述を抽出しています. togo.annotation.jp
40 days ago · reply · retweet · favorite

synobu TogoAnnotation: ソーシャルブックマークでゲノム注釈を行なうプロジェクト. 遺伝子, 文献, ゲノムごとのサマリや各種JSON APIが付け加わりました. togo.annotation.jp
40 days ago · reply · retweet · favorite

yaskaz KazusaAnntation 改め TogoAnnotation review 中。さらにかっちょ良くなったお togo.annotation.jp

Join the conversation

Recent Annotations

Ann:TG:4742
http://streptomyces.nih.go.jp/gview/view_annon.cgi?molecule=TG&id=4...
1405 annotations annotation:157874

Ann:TG:2095
http://streptomyces.nih.go.jp/gview/view_annon.cgi?molecule=TG&id=2...
60 annotations annotation:158993

Ann:TG:4742
http://streptomyces.nih.go.jp/gview/view_annon.cgi?molecule=TG&id=4...

Recent Genes

AdpA, adpAg
http://streptomyces.nih.go.jp/gview/view_annon.cgi?molecule=TG&id=4742
1405 annotations
26 references

sgmA
http://streptomyces.nih.go.jp/gview/view_annon.cgi?molecule=TG&id=2095
60 annotations
14 references

adsA
http://streptomyces.nih.go.jp/gview/view_annon.cgi?molecule=TG&id=4151
78 annotations
17 references

more... Last updated: 2012-04-24 04:09

Recent References

Scherzinger, D. Ruch, S. Kloer, D. P. Wilde, A. Al-Babili, S.
Retinal is formed from apo-carotenoids in Nostoc sp. PCC7120:...
Biochem J. 2006 Sep 15;398(3):361-9.
72 annotations pmid:16759173

Tomono, A. Tsai, Y. Ohnishi, Y. Horinouchi, S.
Three chymotrypsin genes are members of the AdpA regulon in...
J Bacteriol. 2005 Sep;187(18):6341-53.
346 annotations pmid:16159767

Zhao, K. H. Zhang, J. Tu, J. M. Bohm, S. Ploscher, M. Eichacker, L. Bubenzer, C. Sc...
Lyase activities of CpcS- and CpcT-like proteins from Nostoc PC...
J Biol Chem. 2007 Nov 23;282(47):34093-103. Epub 2007 Sep 25.
9 annotations pmid:17895251

more... Last updated: 2012-04-24 08:36

Recent Genomes

http://togo.annotation.jp

CyanoBase [search] in [Gene symbols, Definition]

1. Shoumskaya, M. A. Paithoonrangsarid, K. Kanesaki, Y. Los, D. A. Zinchenko, V. V. Tanticharoen, M. Suzuki, I. Murata, N.
Identical Hik-Rre systems are involved in perception and transduction of salt signals and hyperosmotic signals but regulate the expression of individual genes to different extents in synechocystis.
J Biol Chem. 2005 Jun 3;280(22):21531-8. Epub 2005 Mar 31. PMID:15805106 | Abstract | MeSH | Related Articles | Sections | Gene Index |
Ab In Mm Re Di Fi Ta | CyanoGenes:1712 CyanoGenes:MedlineID gname:rre31 GI abstract introduction experimental procedures results discussion table1 table2 fig1 fig2 fig3 fig4

2. Ashby, M. K. Mullineaux, C. W.
Cyanobacterial ycf27 gene products regulate energy transfer from phycobilisomes to photosystems I and II.
FEMS Microbiol Lett. 1999 Dec 15;181(2):253-60. PMID:10585546 | Abstract | MeSH | Related Articles | Sections | Gene Index |
Ab Mm Re Di Fi Ta | CYORF:CYREF GI syn:slr0115 abstract materials and methods results discussion table1 table2 fig1 fig2 fig3

3. Ashby, M. K. Houmard, J. Mullineaux, C. W.
The ycf27 genes from cyanobacteria and eukaryotic algae: distribution and implications for chloroplast evolution.
FEMS Microbiol Lett. 2002 Aug 27;214(1):25-30. PMID:12204368 | Abstract | MeSH | Related Articles | Sections | Gene Index |
Ab In Mm Di Fi Ta | DBTCY GI tf_type:OmpR gname:rpaA tfac abstract introduction materials methods discussion table1 table3 table4 fig1

4. Paithoonrangsarid, K. Shoumskaya, M. A. Kanesaki, Y. Satoh, S. Tabata, S. Los, D. A. Zinchenko, V. V. Hayashi, H. Tanticharoen, M. Suzuki, I. Murata, N.
Five histidine kinases perceive osmotic stress and regulate distinct sets of genes in Synechocystis.
J Biol Chem. 2004 Dec 17;279(51):53078-86. Epub 2004 Oct 7. PMID:15471853 | Abstract | MeSH | Related Articles | Sections | Gene Index
Ab Re Di Fi Ta | CyanoGenes:1712 CyanoGenes:MedlineID DBTCY GI tf_type:OmpR tfac abs
results discussion table2 fig1 fig2 fig3

5. Hanke, G. T. Satomi, Y. Shinmura, K. Takao, T. Hase, T.
A screen for potential ferredoxin electron transfer partners uncovers new, redox dependent interactions.
Biochim Biophys Acta. 2011 Feb;1814(2):366-74. Epub 2010 Sep 22. PMID:20869472 | Abstract | MeSH | Related Articles | Sections | Gene Inde
Ab Re Di Co Fi Ta | GI gname:rpaA gname:rre31 abstract results discussion conclusions table3 fig2

6. Kappell, A. D. van Waasbergen, L. G.
The response regulator RpaB binds the high light regulatory 1 sequence upstream of the high-light-inducible hliB gene from the cyanobacterium Synechocystis PCC 6803.
Arch Microbiol. 2007 Apr;187(4):337-42. Epub 2007 Feb 10. PMID:17294172 | Abstract | MeSH | Related Articles | Sections | Gene Index
Mm Re Di Fi | GI materials and methods results discussion fig2 fig3

7. Tabei, Y. Okada, K. Tsuzuki, M.
Sll1330 controls the expression of glycolytic genes in Synechocystis sp. PCC 6803.
Biochem Biophys Res Commun. 2007 Apr 20;355(4):1045-50. Epub 2007 Feb 22. PMID:17331473 | Abstract | MeSH | Related Articles | Sections | Gene Index |
Re Di Fi Ta | GI gname:rpaA results discussion table1 fig2

8. Sugita, C. Ogata, K. Shikata, M. Jikuya, H. Takano, J. Furumichi, M. Kanehisa, M. Omata, T. Sugiura, M. Sugita, M.
Complete nucleotide sequence of the freshwater unicellular cyanobacterium Synechococcus elongatus PCC 6301 chromosome: gene content and organization.
Photosynth Res. 2007 Jul-Sep;93(1-3):55-67. Epub 2007 Jan 9. PMID:17211581 | Abstract | MeSH | Related Articles | Sections | Gene Index |
Re Di Ta | GI gname:rpaA gname:rre31 results discussion table3

9. Murata, N. Suzuki, I.
Exploitation of genomic sequences in a systematic analysis to access how cyanobacteria sense environmental stress.
J Exp Bot. 2006;57(2):235-47. Epub 2005 Nov 29. PMID:16317040 | Abstract | MeSH | Related Articles | Sections | Gene Index |
Fi | CYORF:CYREF CyanoGenes:1712 CyanoGenes:MedlineID GI gname:rre31 syn:slr0115 body fig4

10. Mary, I. Vaulot, D.
Two-component systems in Prochlorococcus MED4: genomic analysis and differential expression under stress.
FEMS Microbiol Lett. 2003 Sep 12;226(1):135-44. PMID:13129619 | Abstract | MeSH | Related Articles | Sections | Gene Index |
Di Ta | GI discussion table4

11. Takai, N. Nakajima, M. Oyama, T. Kito, R. Sugita, C. Sugita, M. Kondo, T. Iwasaki, H.
A KaiC-associating SasA-RpaA two-component regulatory system as a major circadian timing mediator in cyanobacteria.
Proc Natl Acad Sci U S A. 2006 Aug 8;103(32):12109-14. Epub 2006 Aug 1. PMID:16882723 | Abstract | MeSH | Related Articles | Sections | Gene Index |
Di | GI gname:rre31 gname:rpaA discussion

12. Mizuno, T. Kaneko, T. Tabata, S.
Compilation of all genes encoding bacterial two-component signal transducers in the genome of the cyanobacterium, Synechocystis sp. strain PCC 6803.
DNA Res. 1996 Dec 31;3(6):407-14. PMID:9097043 | Abstract | MeSH | Related Articles | Sections | Gene Index |
Ta | DBTCY GI tf_type:OmpR tfac table2

13. Lechno-Yossef, S. Fan, Q. Ehira, S. Sato, N. Wolk, C. P.
Mutations in four regulatory genes have interrelated effects on heterocyst maturation in Anabaena sp. strain PCC 7120.
J Bacteriol. 2006 Nov;188(21):7387-95. Epub 2006 Aug 25. PMID:16936023 | Abstract | MeSH | Related Articles | Sections | Gene Index |
In | GI gname:rre31 introduction

14. Hihara, Y. Kamei, A. Kanehisa, M. Kaplan, A. Ikeuchi, M.
DNA microarray analysis of cyanobacterial gene expression during acclimation to high light.

15. Midorikawa, T. Matsumoto, K. Narikawa, R. Ikeuchi, M.
An Rrf2-type transcriptional regulator is required for expression of psaAB genes in the cyanobacterium Synechocystis sp. PCC 6803.

16. Peter, E. Wallner, T. Wilde, A. Grimm, B.

Manually curated,
__complete publication__
list for each gene

For example, 429 papers for a gene *psbA2* at:
http://genome.kazusa.or.jp/cyanobase/Synechocystis/genes/slr1311

TogoAnnotation: applications

- Gene indexing: paper curation project (>5000)
  - Cyanobacteria, Rhizobia, Streptmyces
- Community annotation projects.
  - Comminity annotation for a *Rhizobium*

3 days training then 2 months remote

Microbes Environ. 2012 Mar 28. [Epub ahead of print]

Complete Genome Sequence of *Bradyrhizobium* sp. S23321: Insights into Symbiosis Evolution in Soil Oligotrophs.

Okubo T, Tsukui T, Maita H, Okamoto S, Oshima K, Fujisawa T, Saito A, Futamata H, Hattori R, Shimomura Y, Haruta S, Morimoto S, Wang Y, Sakai Y, Hattori M, Aizawa SI, Nagashima KV, Masuda S, Hattori T, Yamashita A, Bao Z, Hayatsu M, Kajiya-Kanegae H, Yoshinaga I, Sakamoto K, Toyota K, Nakao M, Kohara M, Anda M, Niwa R, Jung-Hwan P, Sameshima-Saito R, Tokuda SI, Yamamoto S, Yamamoto S, Yokoyama T, Akutsu T, Nakamura Y, Nakahira-Yanaka Y, Takada Hoshino Y, Hirakawa H, Mitsui H, Terasawa K, Itakura M, Sato S, Ikeda-Ohtsubo W, Sakakura N, Kaminuma E, Minamisawa K.

  - Currently we annotate an algae and a moss genomes.

http://togo.annotation.jp

About TogoAnnotation

統合アノテーションは、ソーシャルブックマークのしくみを利用して生物のデータに様々な注釈（アノテーション）をつけることが出来るサービスです。

News

統合アノテーション
TogoAnnotation

TogoAnnotation TogoAnnotation (KazusaAnnotation)を用いてコミュニティゲノムアノテーションを実施したBradyrhizobium sp. S23321の論文、配列が公開されました。 jstage.jst.go.jp/article/jsme2/... getentry.ddbj.nig.ac.jp/getentry?acces...
25 days ago · reply · retweet · favorite

bonohu TogoAnnotationキ────(∀)────!! (#AJACS live at ustre.am/1x4x/1)
38 days ago · reply · retweet · favorite

synobu 19種の光合成関連生物から約8000報の文献をキュレーターが読んで20万以上の文献中の遺伝子名記述を抽出しています. togo.annotation.jp
40 days ago · reply · retweet · favorite

synobu TogoAnnotation: ソーシャルブックマークでゲノム注釈を行なうプロジェクト. 遺伝子, 文献, ゲノムごとのサマリや各種JSON APIが付け加わりました. togo.annotation.jp
40 days ago · reply · retweet · favorite

yaskaz KazusaAnntation 改め TogoAnnotation review 中。さらにかっちょ良くなったお togo.annotation.jp

Join the conversation

Recent Annotations

Ann:TG:4742
http://streptomyces.nih.go.jp/gview/view_annon.cgi?molecule=TG&id=4...
1405 annotations annotation:157874

Ann:TG:2095
http://streptomyces.nih.go.jp/gview/view_annon.cgi?molecule=TG&id=2...
60 annotations annotation:158993

Ann:TG:4742
http://streptomyces.nih.go.jp/gview/view_annon.cgi?molecule=TG&id=4...

Recent Genes

AdpA, adpAg
http://streptomyces.nih.go.jp/gview/view_annon.cgi?molecule=TG&id=4742
1405 annotations
26 references

sgmA
http://streptomyces.nih.go.jp/gview/view_annon.cgi?molecule=TG&id=2095
60 annotations
14 references

adsA
http://streptomyces.nih.go.jp/gview/view_annon.cgi?molecule=TG&id=4151
78 annotations
17 references

more...
Last updated: 2012-04-24 04:09

Recent References

Scherzinger, D. Ruch, S. Kloer, D. P. Wilde, A. Al-Babili, S.
Retinal is formed from apo-carotenoids in Nostoc sp. PCC7120:...
Biochem J. 2006 Sep 15;398(3):361-9.
72 annotations pmid:16759173

Tomono, A. Tsai, Y. Ohnishi, Y. Horinouchi, S.
Three chymotrypsin genes are members of the AdpA regulon in...
J Bacteriol. 2005 Sep;187(18):6341-53.
346 annotations pmid:16159767

Zhao, K. H. Zhang, J. Tu, J. M. Bohm, S. Ploscher, M. Eichacker, L. Bubenzer, C. Sc...
Lyase activities of CpcS- and CpcT-like proteins from Nostoc PC...
J Biol Chem. 2007 Nov 23;282(47):34093-103. Epub 2007 Sep 25.
9 annotations pmid:17895251

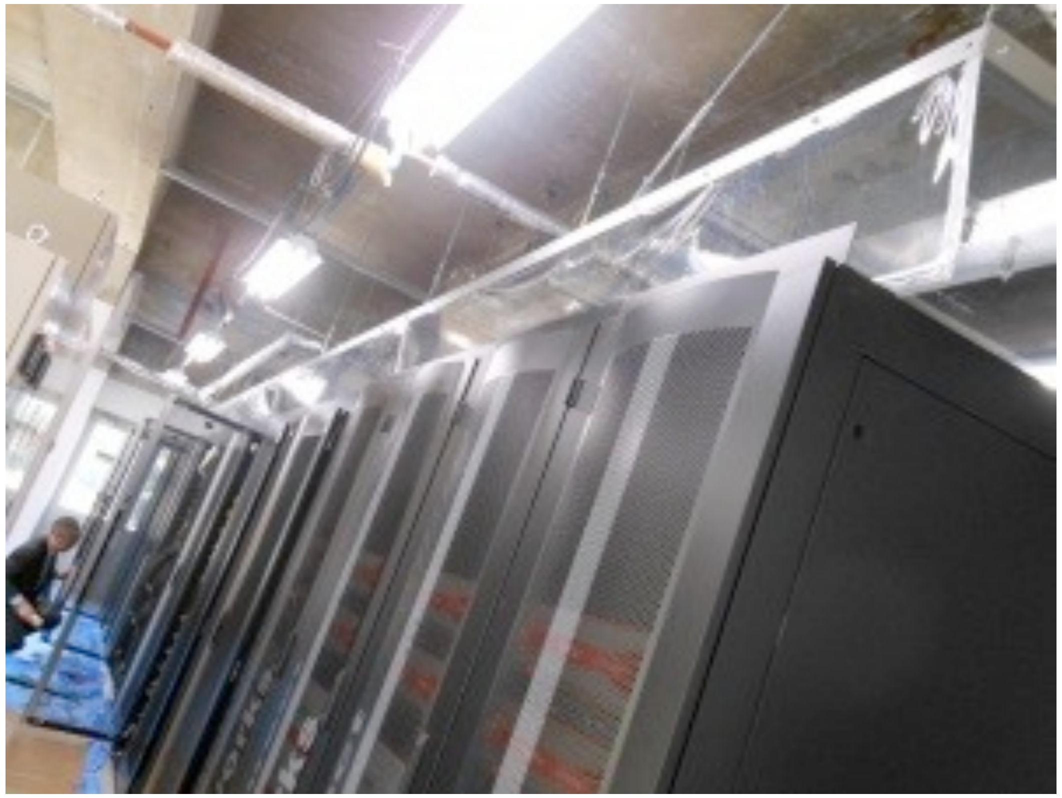
more...
Last updated: 2012-04-24 08:36

Recent Genomes

We need:

- good references
- good pipelines

NIG Supercomputer (2012.3-)



2012.03.01

Phase 1

- 165.1 TFlops
- 5 PB HDD
- Containing 10TB and 2TB shared memory system.

Rmax of LINPACK: 82.90 TFLOPS

Rank:170th in Top500 (Nov. 2011)

(Rank 11th in Japan)

2014.03.01

Phase 2

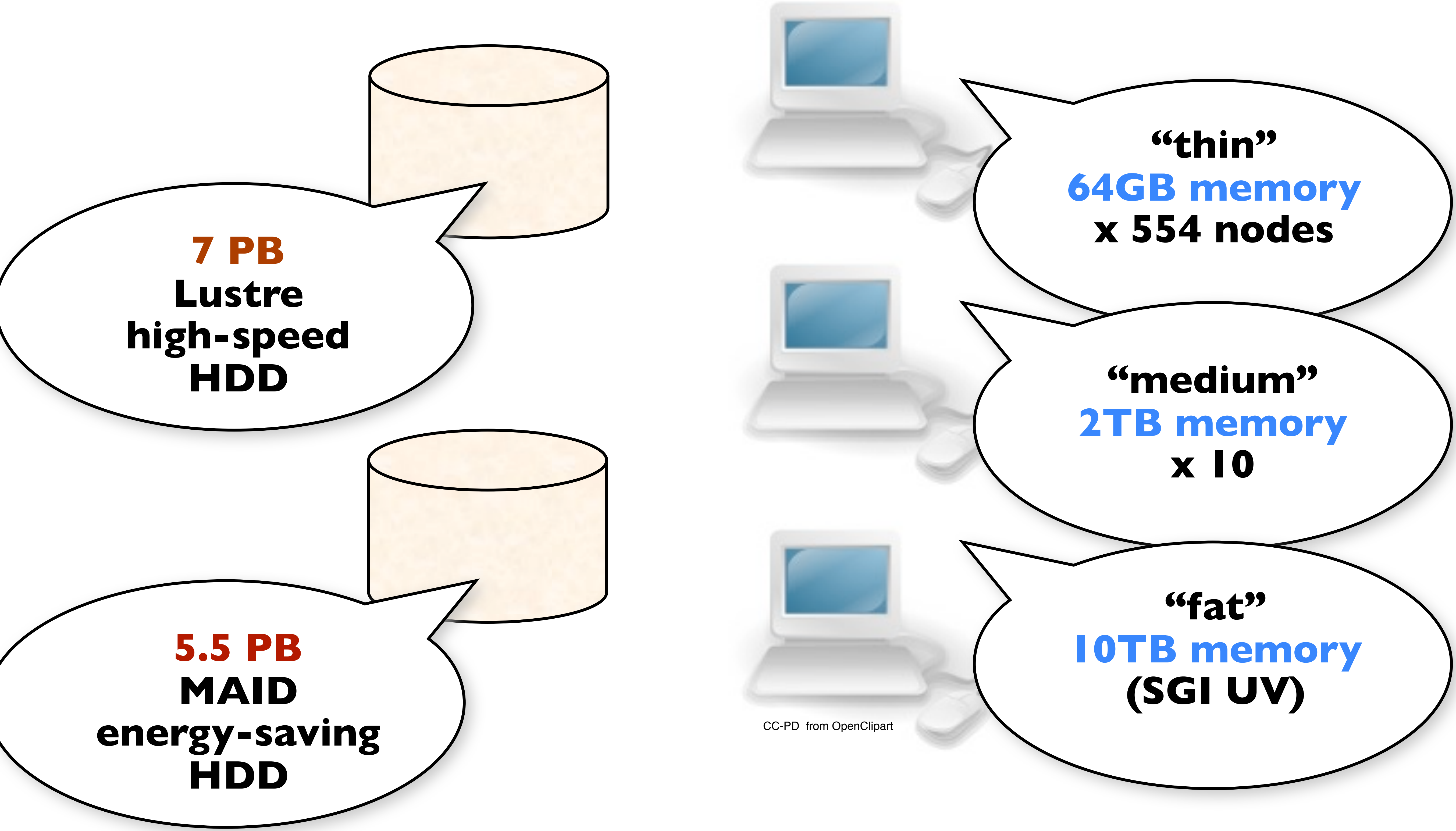
- about 400 TFlops (total)
- 12.5 PB HDD (total)

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

Hot Topics

- 2013.06.26 WABI (Web API for Biology) の再開
- 2013.06.11 DDBJ リリース 93.0, DAD リリース 63.0 完成
- 2013.05.15 「第27回 DDBJing 講習会 in 三島(2013.7.4.開催)」のご案内 (参加申込み受付中)

DDBJ supercomputer (2014.3 ver.)



7 PB
Lustre
high-speed
HDD

5.5 PB
MAID
energy-saving
HDD

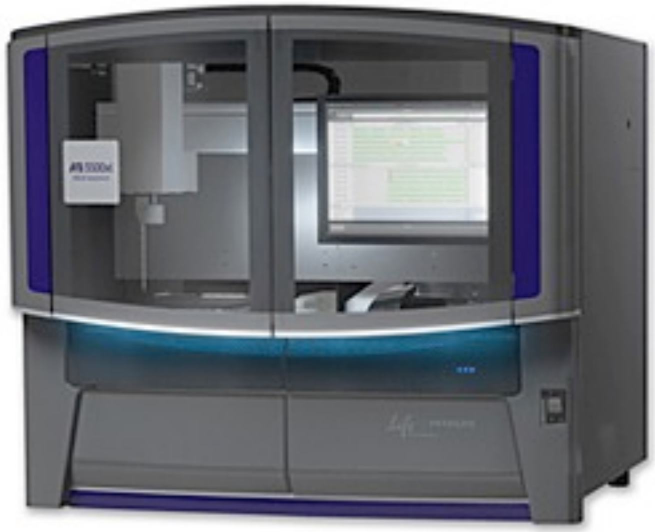
“thin”
64GB memory
x 554 nodes

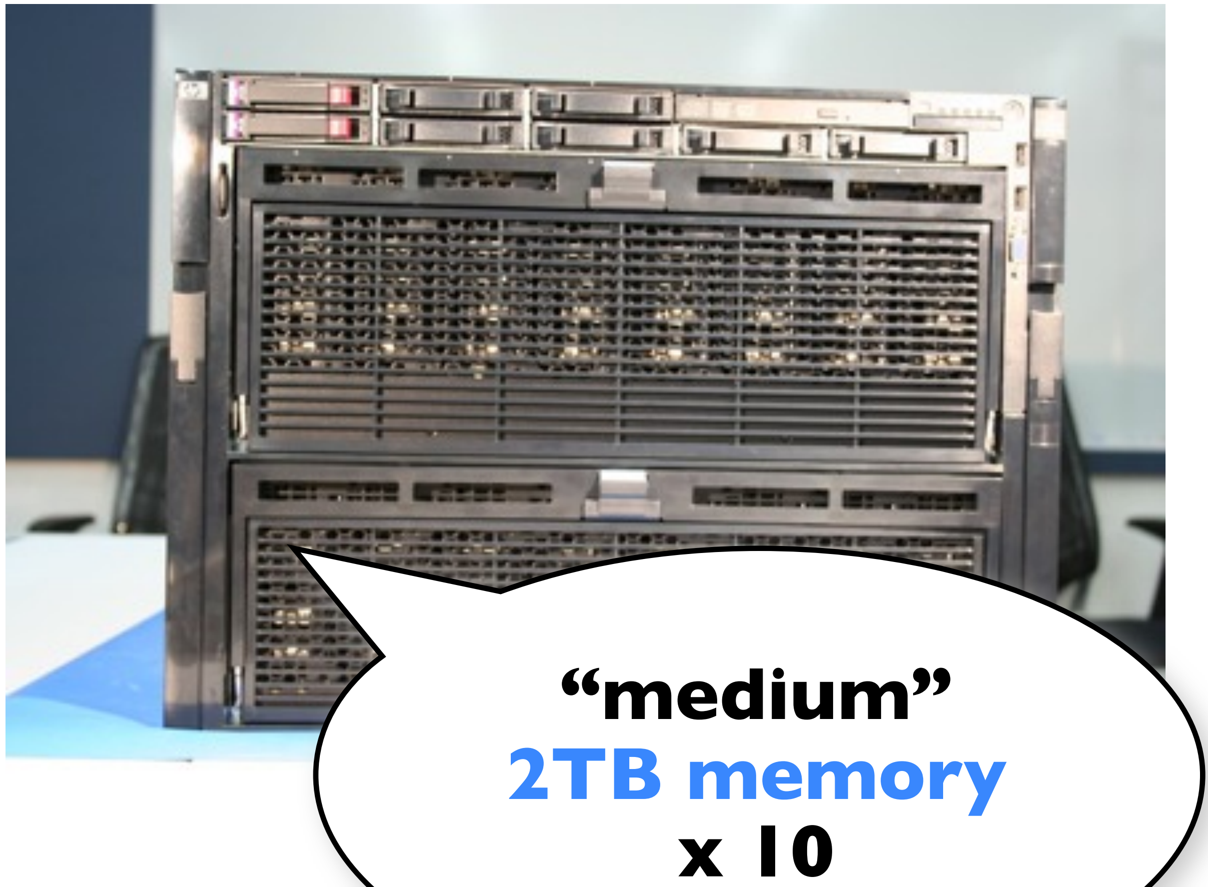
“medium”
2TB memory
x 10

“fat”
10TB memory
(SGI UV)

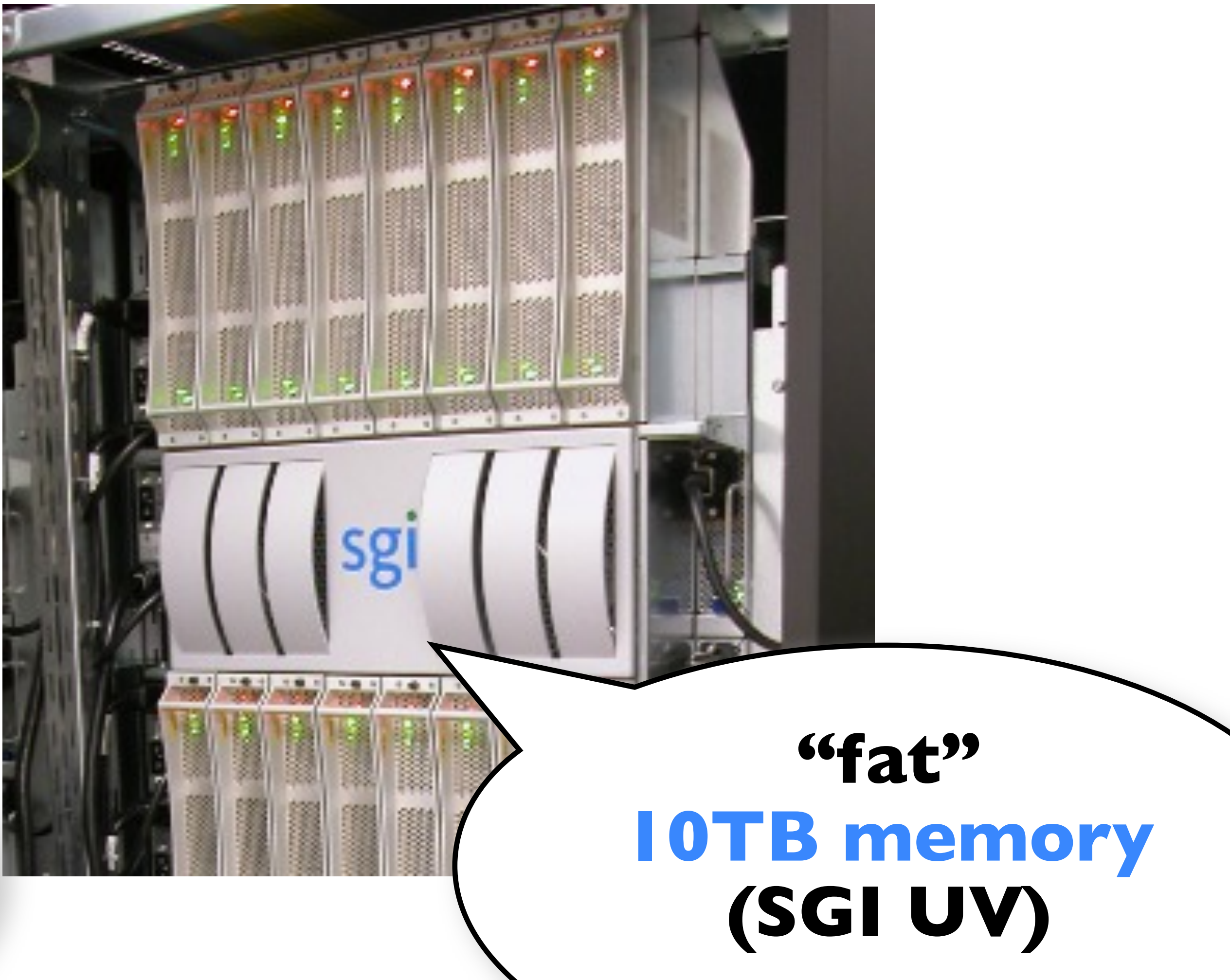
CC-PD from OpenClipart

NGS's + SC's in Biology



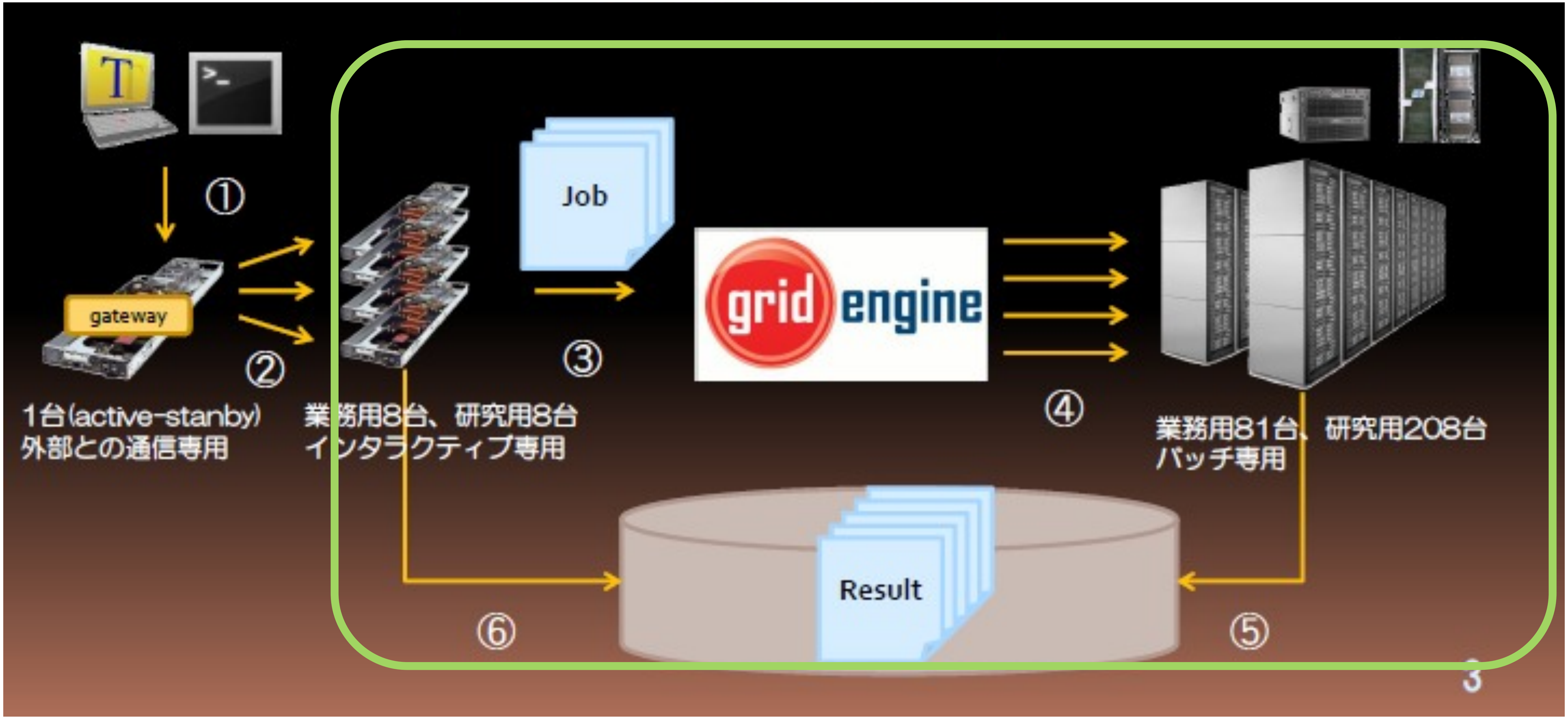


"medium"
2TB memory
x 10



"fat"
10TB memory
(SGI UV)

Running Batch Jobs



Running Jobs on Medium and Fat nodes

```
# run a script on a thin node.
qsub -cwd -S /bin/bash your_script.sh

# run a script on a medium node.
qsub -cwd -l month -l medium -S /bin/bash your_script.sh

# run a script on the fat node.
qsub -cwd -l month -l fat -S /bin/bash your_script.sh
```

Memory Request (for each CPU core)

```
# This job runs on 1 CPU core and 128GB
memory.
qsub -cwd -l month -l medium
 -l s_vmem=128G,mem_req=128G
 -S /bin/bash your_script.sh

# This job runs on 10 CPU core (in the same
node) and 1280GB memory.
qsub -cwd -l month -l medium
 -l s_vmem=128G,mem_req=128G
 -pe def_slot=10
 -S /bin/bash your_script.sh
```

It’s easy to use,

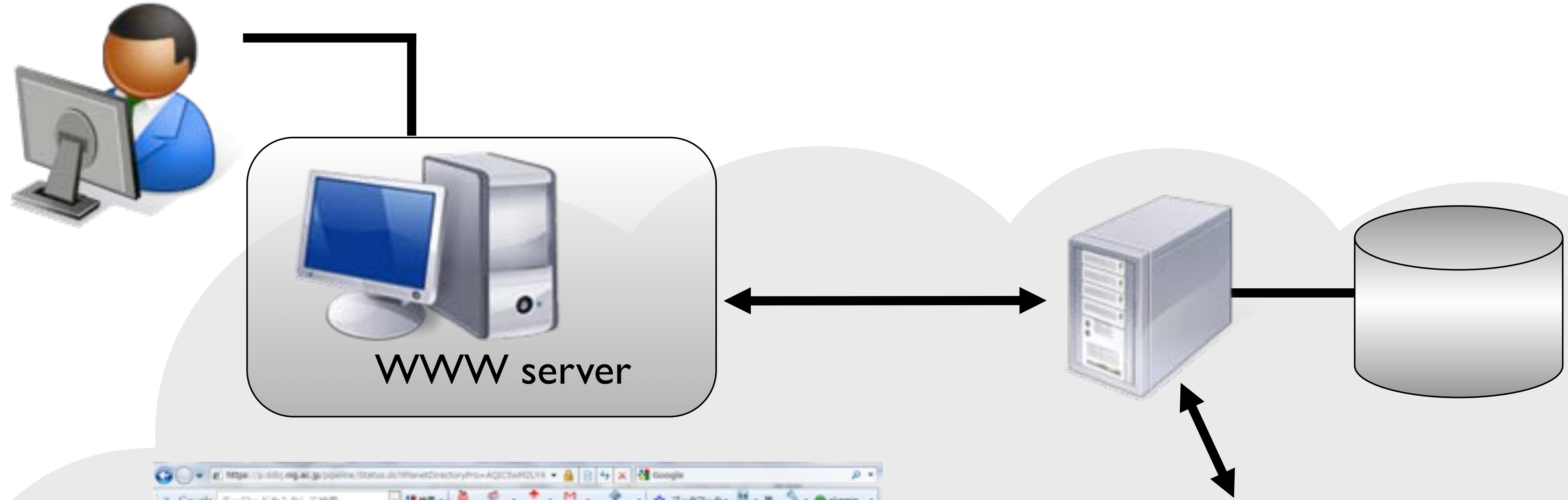
isn’t it?

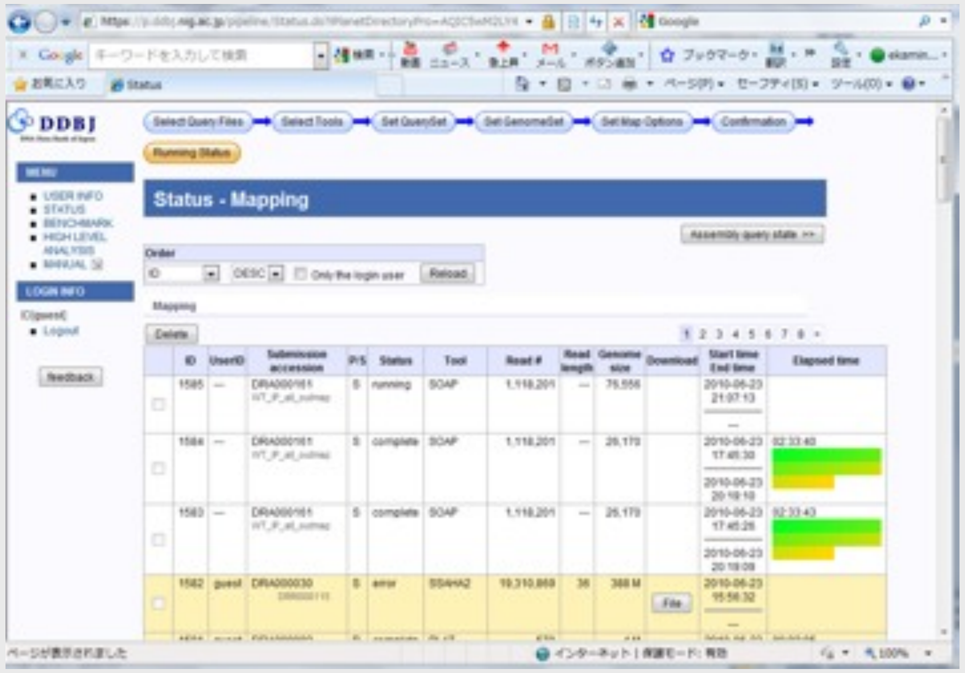
(簡単でしょ？)

(゜д゜) ...What?

(ハア？)

DDBJ Pipeline: Cloud-based analysis tool





Status - Mapping

Order: ID DESC ☐ Only the login user

Mapping

1 2 3 4 5 6 7 8 »

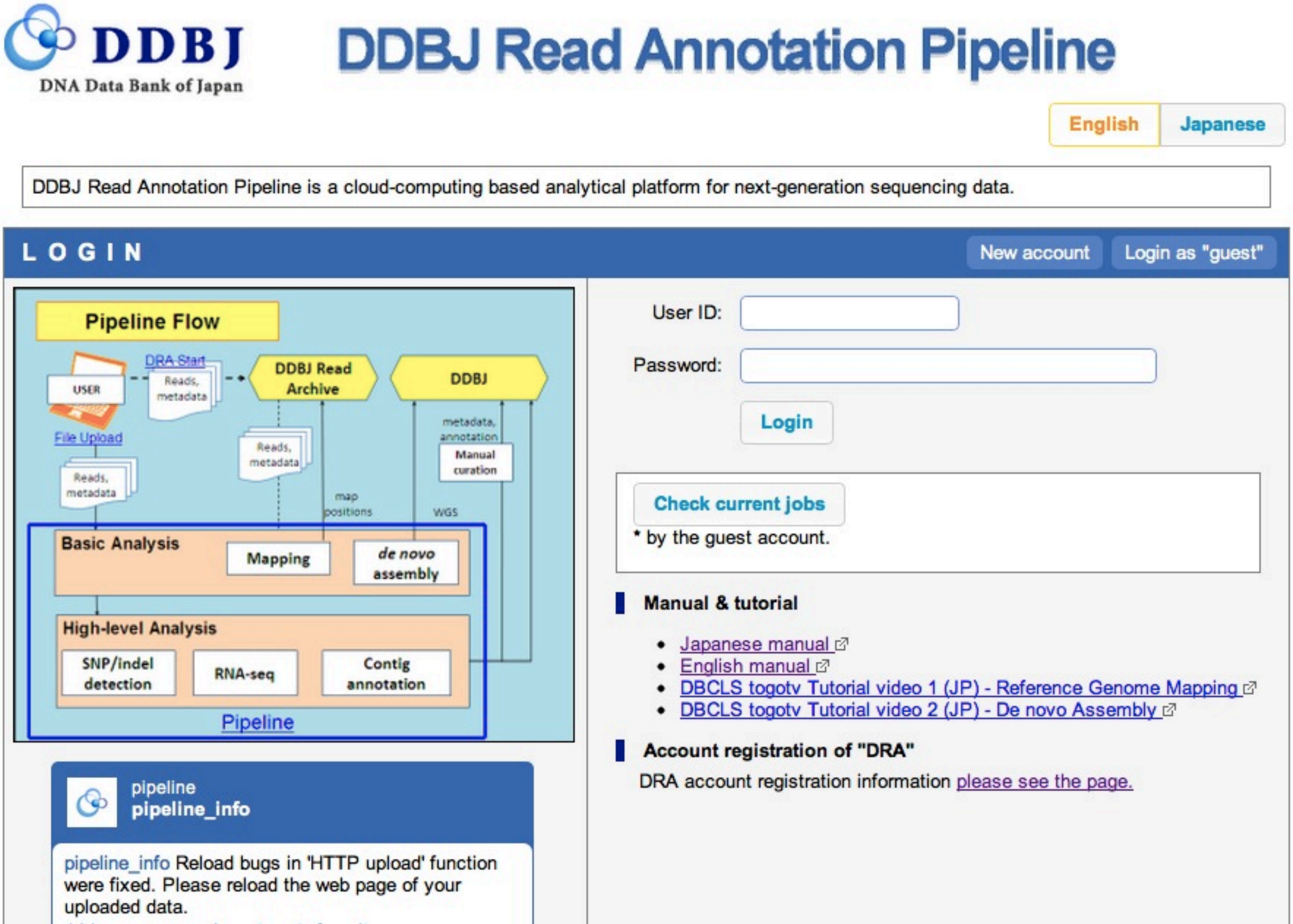
	ID	UserID	Submission accession	P/S	Status	Tool	Read #	Read length	Genome size	Download	Start time End time	Elapsed time
☐	1585	—	DRA000161	S	running	SOAP	1,118,201	—	76,556		2010-06-23 21:07:13 —	
☐	1584	—	DRA000161	S	complete	SOAP	1,118,201	—	26,170		2010-06-23 17:45:30 2010-06-23 20:19:10	02:33:40
☐	1583	—	DRA000161	S	complete	SOAP	1,118,201	—	26,170		2010-06-23 17:45:26 2010-06-23 20:19:09	02:33:43
☐	1582	guest	DRA000030	S	error	SSAHA2	10,310,869	36	388 M	File	2010-06-23 15:56:32 —	

DDBJ supercomputer

no resource? → use DDBJ's supercomputer

no skill? → web-based easy-operation

http://p.ddbj.nig.ac.jp



DDBJ Read Annotation Pipeline

DDBJ Read Annotation Pipeline is a cloud-computing based analytical platform for next-generation sequencing data.

User ID:

Password:

* by the guest account.

Manual & tutorial

- Japanese manual
- English manual
- DBCLS togotv Tutorial video 1 (JP) - Reference Genome Mapping
- DBCLS togotv Tutorial video 2 (JP) - De novo Assembly

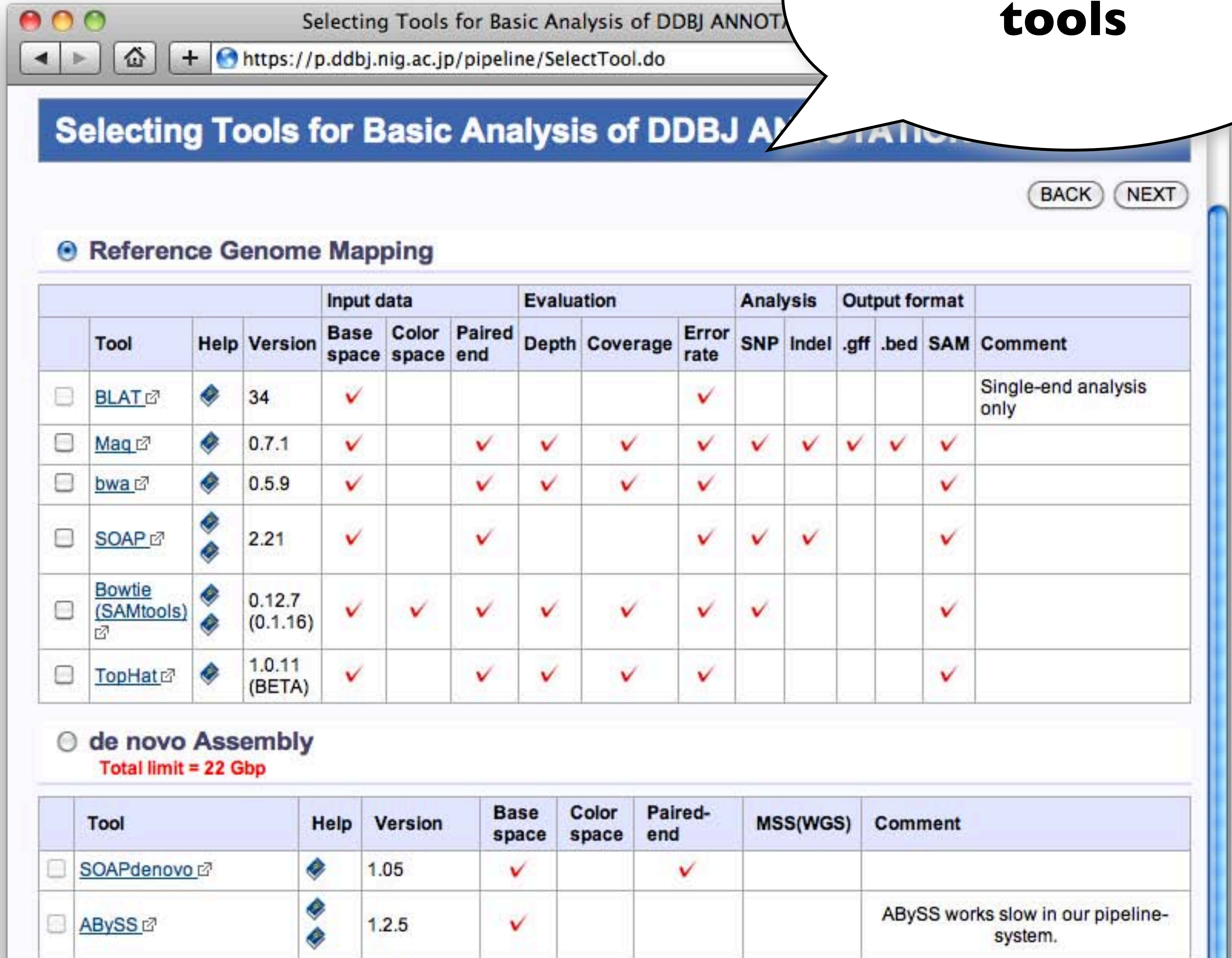
Account registration of "DRA"

DRA account registration information please see the page.

pipeline_info Reload bugs in 'HTTP upload' function were fixed. Please reload the web page of your uploaded data.

DDBJ pipeline: Software

***de facto* standard tools**



Selecting Tools for Basic Analysis of DDBJ A

● Reference Genome Mapping

				Input data			Evaluation			Analysis		Output format			
	Tool	Help	Version	Base space	Color space	Paired end	Depth	Coverage	Error rate	SNP	Indel	.gff	.bed	SAM	Comment
☐	BLAT	◆	34	✓					✓						Single-end analysis only
☐	Maq	◆	0.7.1	✓		✓	✓	✓	✓	✓	✓	✓	✓	✓	
☐	bwa	◆	0.5.9	✓		✓	✓	✓	✓					✓	
☐	SOAP	◆ ◆	2.21	✓		✓			✓	✓	✓			✓	
☐	Bowtie (SAMtools)	◆ ◆	0.12.7 (0.1.16)	✓	✓	✓	✓	✓	✓	✓				✓	
☐	TopHat	◆	1.0.11 (BETA)	✓		✓	✓	✓	✓					✓	

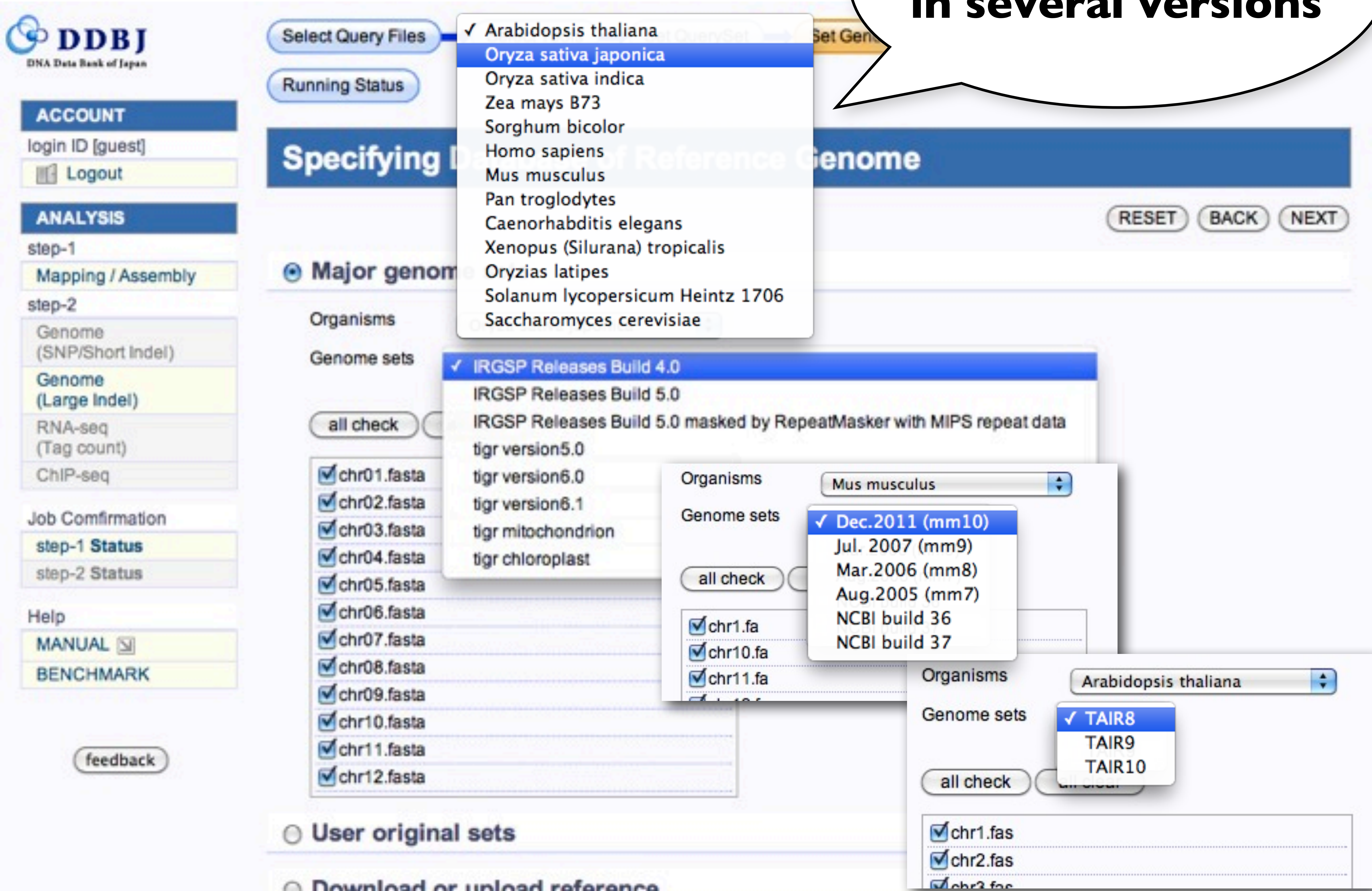
○ de novo Assembly

Total limit = 22 Gbp

	Tool	Help	Version	Base space	Color space	Paired-end	MSS(WGS)	Comment
☐	SOAPdenovo	◆	1.05	✓		✓		
☐	ABySS	◆ ◆	1.2.5	✓				ABySS works slow in our pipeline-system.

DDBJ pipeline: references

major genome sets in several versions



Specifying Data Set of Reference Genome

● Major genome

Organisms: Oryza sativa japonica
- ✓ Arabidopsis thaliana
- Oryza sativa japonica
- Oryza sativa indica
- Zea mays B73
- Sorghum bicolor
- Homo sapiens
- Mus musculus
- Pan troglodytes
- Caenorhabditis elegans
- Xenopus (Silurana) tropicalis
- Oryzias latipes
- Solanum lycopersicum Heintz 1706
- Saccharomyces cerevisiae

Genome sets:
- ✓ IRGSP Releases Build 4.0
- IRGSP Releases Build 5.0
- IRGSP Releases Build 5.0 masked by RepeatMasker with MIPS repeat data
- tigr version5.0
- tigr version6.0
- tigr version6.1
- tigr mitochondrion
- tigr chloroplast

☑ chr01.fasta
☑ chr02.fasta
☑ chr03.fasta
☑ chr04.fasta
☑ chr05.fasta
☑ chr06.fasta
☑ chr07.fasta
☑ chr08.fasta
☑ chr09.fasta
☑ chr10.fasta
☑ chr11.fasta
☑ chr12.fasta

○ User original sets

○ Download or upload reference

Organisms: Mus musculus

Genome sets:
- ✓ Dec.2011 (mm10)
- Jul. 2007 (mm9)
- Mar.2006 (mm8)
- Aug.2005 (mm7)
- NCBI build 36
- NCBI build 37

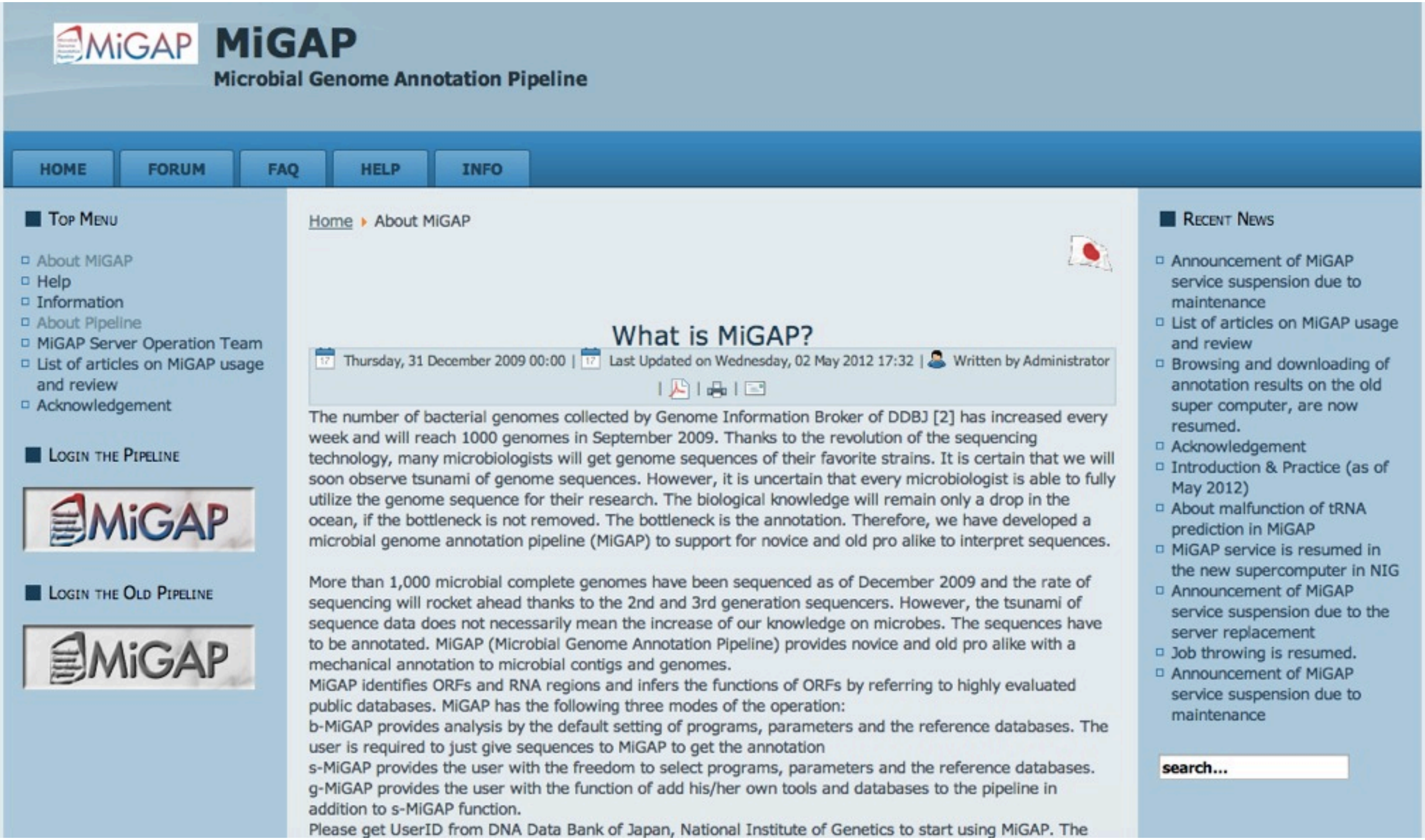
☑ chr1.fa
☑ chr10.fa
☑ chr11.fa

Organisms: Arabidopsis thaliana

Genome sets:
- ✓ TAIR8
- TAIR9
- TAIR10

☑ chr1.fas
☑ chr2.fas
☑ chr3.fas

http://www.migap.org



MiGAP

Microbial Genome Annotation Pipeline

TOP MENU

- About MiGAP
- Help
- Information
- About Pipeline
- MiGAP Server Operation Team
- List of articles on MiGAP usage and review
- Acknowledgement

LOGIN THE PIPELINE

LOGIN THE OLD PIPELINE

Home ▸ About MiGAP

What is MiGAP?

Thursday, 31 December 2009 00:00 | Last Updated on Wednesday, 02 May 2012 17:32 | Written by Administrator

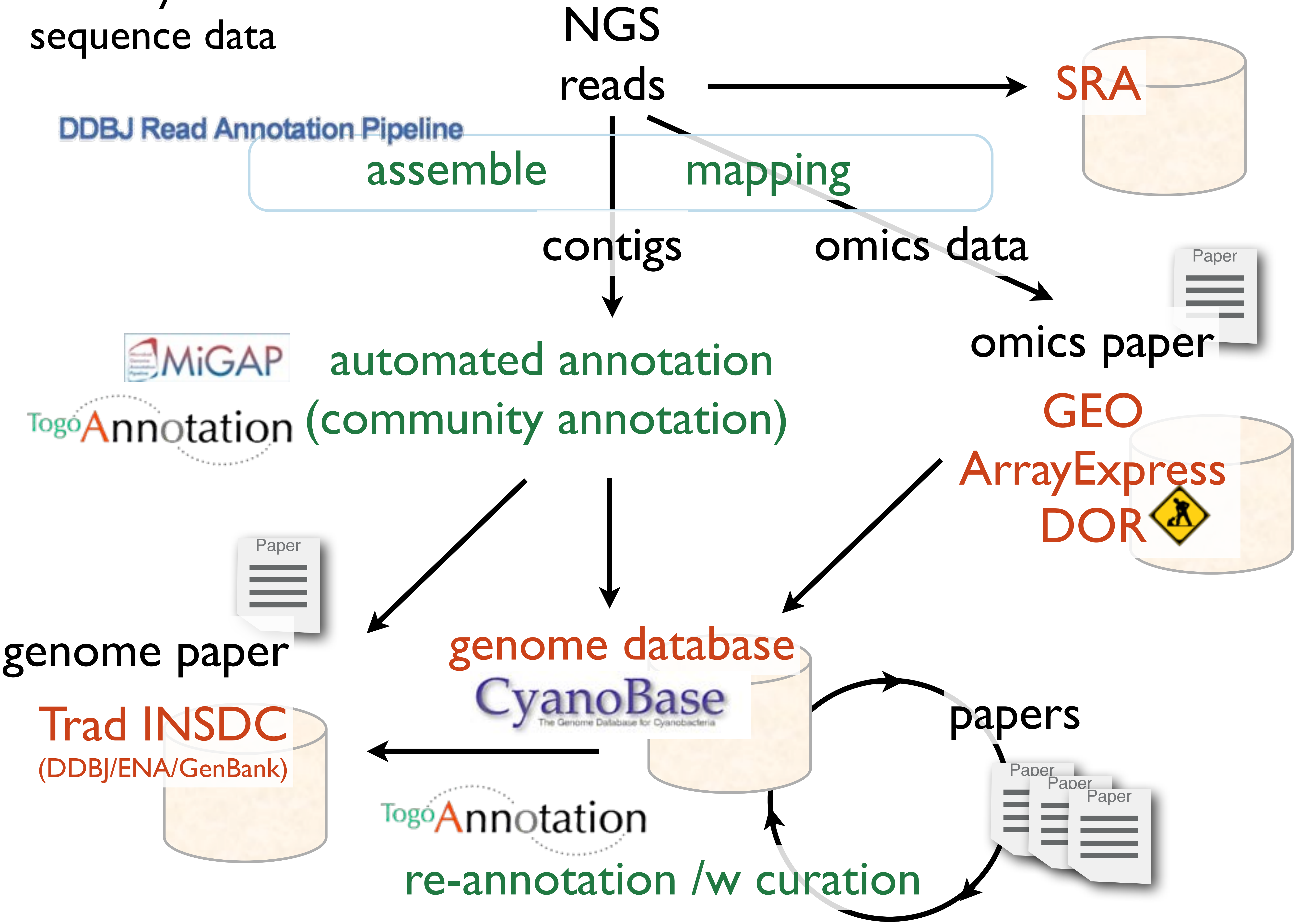
The number of bacterial genomes collected by Genome Information Broker of DDBJ [2] has increased every week and will reach 1000 genomes in September 2009. Thanks to the revolution of the sequencing technology, many microbiologists will get genome sequences of their favorite strains. It is certain that we will soon observe tsunami of genome sequences. However, it is uncertain that every microbiologist is able to fully utilize the genome sequence for their research. The biological knowledge will remain only a drop in the ocean, if the bottleneck is not removed. The bottleneck is the annotation. Therefore, we have developed a microbial genome annotation pipeline (MiGAP) to support for novice and old pro alike to interpret sequences.

More than 1,000 microbial complete genomes have been sequenced as of December 2009 and the rate of sequencing will rocket ahead thanks to the 2nd and 3rd generation sequencers. However, the tsunami of sequence data does not necessarily mean the increase of our knowledge on microbes. The sequences have to be annotated. MiGAP (Microbial Genome Annotation Pipeline) provides novice and old pro alike with a mechanical annotation to microbial contigs and genomes.
MiGAP identifies ORFs and RNA regions and infers the functions of ORFs by referring to highly evaluated public databases. MiGAP has the following three modes of the operation:
b-MiGAP provides analysis by the default setting of programs, parameters and the reference databases. The user is required to just give sequences to MiGAP to get the annotation
s-MiGAP provides the user with the freedom to select programs, parameters and the reference databases.
g-MiGAP provides the user with the function of add his/her own tools and databases to the pipeline in addition to s-MiGAP function.
Please get UserID from DNA Data Bank of Japan, National Institute of Genetics to start using MiGAP. The

RECENT NEWS

- Announcement of MiGAP service suspension due to maintenance
- List of articles on MiGAP usage and review
- Browsing and downloading of annotation results on the old super computer, are now resumed.
- Acknowledgement
- Introduction & Practice (as of May 2012)
- About malfunction of tRNA prediction in MiGAP
- MiGAP service is resumed in the new supercomputer in NIG
- Announcement of MiGAP service suspension due to the server replacement
- Job throwing is resumed.
- Announcement of MiGAP service suspension due to maintenance

The ecosystem of sequence data



NGS reads

SRA

DDBJ Read Annotation Pipeline

assemble

mapping

contigs

omics data

omics paper

MiGAP

TogoAnnotation

automated annotation (community annotation)

GEO

ArrayExpress

DOR

genome paper

genome database

CyanoBase

The Genome Database for Cyanobacteria

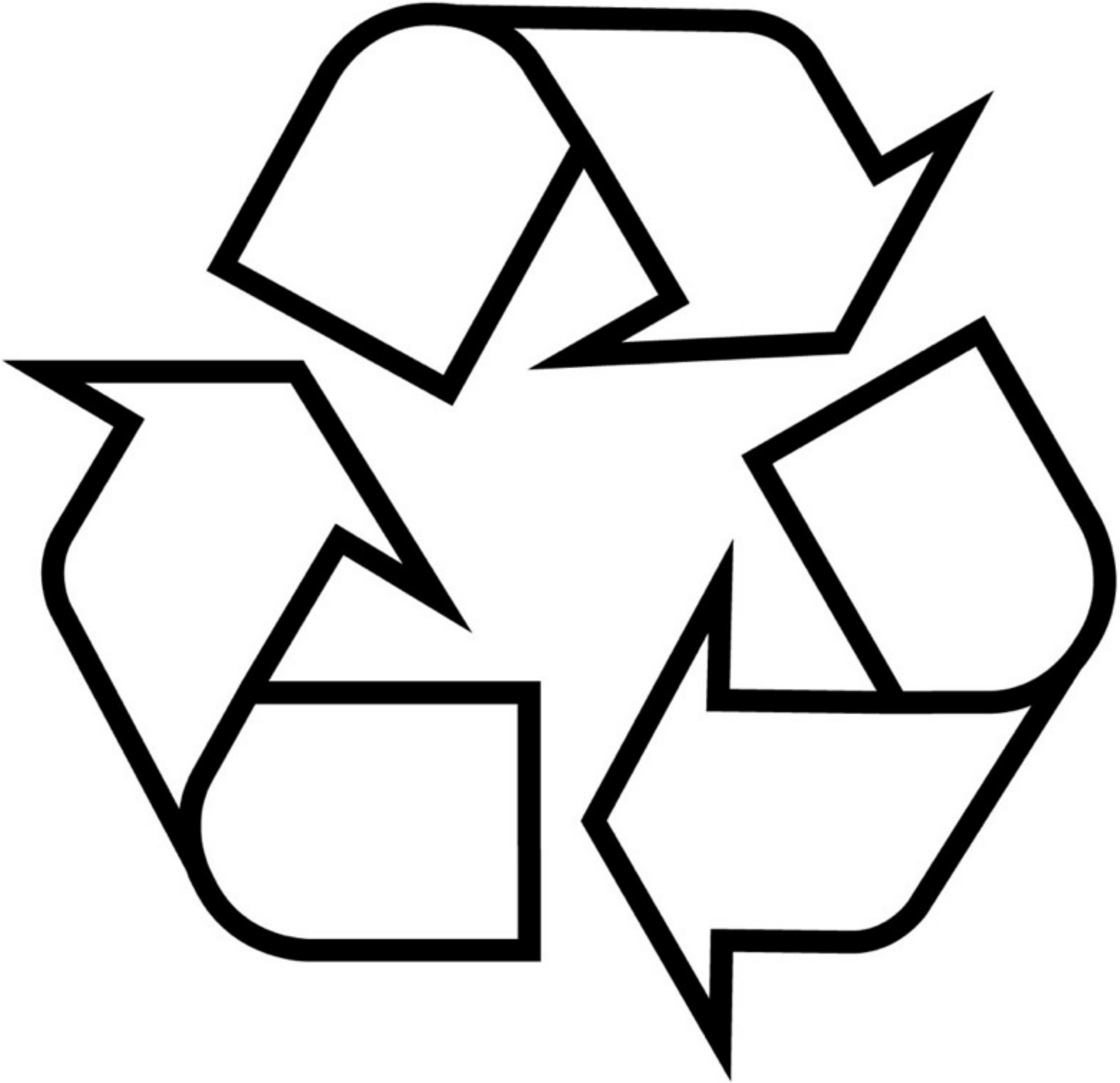
Trad INSDC

(DDBJ/ENA/GenBank)

papers

TogoAnnotation

re-annotation /w curation



Thanks
for your
attention!

Cc, the first-ever cloned cat.