DDBJ Annual Report 2015

---------- 日本語目次 ----------

---------- Contents (English) ----------

1. 2015 ACTIVITY OVERVIEW

1-1. 2015 Main Events

	Event	Date
1	*Nucleic Acids Research* papers released. (Kodama, Nakamura *et al.*)	2015/1/28
2	Meeting at Korean Bioinformation Center (KOBIC) and Korean Intellectual Property Office (KIPO) (Aono and Lee)	2015/1/19-22
3	DDBJ Rel.100.0, DAD (DDBJ amino acid database) Rel. 70.0 (Team Const, Team SE)	2015/3/25
4	Japan Society for Bioscience, Biotechnology, and Agrochemistry, annual meeting 2015 at Okayama (Kohira, Aono, Okido, Team Joho)	2015/3/26-29
5	Open House in National Institute of Genetics (Team Joho)	2015/4/4
6	28th International Collaborators Meeting (ICM) of International Nucleotide Sequence Database Collaboration (INSDC) in USA (Mashima, Kodama, Lee, Tsutsui, and Nakamura)	2015/5/18-21
7	31st DDBJing lecture in Tokyo (Team Joho, Kodama, Lee, and Nakamura)	2015/6/12
8	DDBJ Rel. 101.0, DAD Rel. 71.0 (Team Const, Team SE)	2015/6/24
9	Collaborative meeting with DBCLS (Nakamura, Team Const)	2015/5/19-21
10	NGS Field 4th Meeting in Tsukuba (Fukuda, Kodama, Mashima, Team Joho)	2015/7/2-3
11	1st All-in-One NBDC/DBCLS/PDBj/DDBJ Joint Training Course in Osaka (Mashima, Nakamura, Ogasawara)	2015/7/18
12	32nd DDBJing lecture in Okinawa (on-demand) (Nakamura)	2015/7/29
13	NIG-OIST joint Symposium Evolutionary Bioinformatics (Nakamura)	2015/8/10-12
14	DDBJ Rel. 102.0, DAD Rel. 72.0 (Team Const, Team SE)	2015/9/29
15	BioJapan 2015 World Business Forum (Aono)	2015/10/15
16	*Nucleic Acids Research* papers released under Advance Access. (Mashima, Takagi *et al.*)	2015/11/17
17	33rd DDBJing lecture in Tokyo (Fukuda, Kodama, Lee, Kaminuma, Team Joho)	2015/11/11
18	The 38th Annual Meeting of the Molecular Biology Society of Japan (Team Const, Team Joho)	2015/12/1-4
19	RIKEN training seminar (Kodama)	2015/12/16
20	DDBJ Rel.103.0, DAD Rel. 73.0 (Team Const, Team SE)	2016/1/7
21	The 2015 DNA Database Advisory Committee Meeting	2016/2/29

1-2. Background

The DNA Data Bank of Japan (DDBJ, http://www.ddbj.nig.ac.jp) maintains a primary nucleotide sequence

database and provides analytical resources for biological information to researchers. This database content is exchanged with the US National Center for Biotechnology Information (NCBI) and the European Bioinformatics Institute (EBI) within the framework of the International Nucleotide Sequence Database Collaboration (INSDC).

1-3. Overview of 2015 DDBJ activities

In 2015, resources provided by the DDBJ include traditional nucleotide sequence data released in the form of 34,768,958 entries or 23,899,016,670 nucleotides (as of December 2015), and raw reads of new-generation sequencers in the sequence read archive (SRA). The 2015 DDBJ activities are summarized as the following three highlights.

First, we have been improving the online submission tool for the Japanese Genotype-phenotype Archive (JGA) to transfer an increasing volume of personal genetic and phenotypic data. We have also released the offline file encryption tool to allow secure data submission of a large file by sending a hard disk drive.

Second, DDBJ strengthened physically collaborative communication with National Bioscience Database Center (NBDC), Database Center for Life Science (DBCLS), and Protein Data Bank of Japan (PDBj), to promote international competitiveness. The NBDC Human Database represents a close collaboration with DDBJ JGA. Several collaborative studies between DBCLS and DDBJ developed Semantic Web databases for DDBJ entries, and pilot studies constructed a virtualization platform of Docker containers on the NIG supercomputer. Furthermore, the DDBJ center started collaborative educational seminars with three institutes: PDBj, NBDC, and DBCLS. This physical collaborative communication is expected to promote the production of helpful service tools for the biology community.

Third, the 28th International Collaborators Meeting (ICM) was held in NCBI on 19-21 May. The ICM is held annually for the purpose of international agreement of INSDC metadata standards covering three institutions: DDBJ, ENA/EBI, and NCBI. In 2015, NCBI was the host institution for the ICM. The meeting attendees included six from NCBI, four from EBI, and five from DDBJ. These attendees discussed practical matters for maintaining and updating the following nucleotide sequence data archives: DDBJ, ENA, GenBank, Sequence Read Archive (SRA), Trace Archive, BioProject, and BioSample.

2. OVERVIEW OF SUBMITTED DATA IN 2015

2-1. 28th ICM report

International Nucleotide Sequence Database Collaboration (INSDC), consisting of DDBJ, ENA/EBI, and NCBI, hold an international collaborators meeting every year to discuss practical matters for the maintenance of and improvements to the following nucleotide sequence data archives: DDBJ, ENA, GenBank, Sequence Read Archive (SRA), Trace Archive, BioProject, and BioSample.

Place: NCBI (Bethesda, USA)
Dates: 19-21 May 2015
Attendees: Nakamura, Y., Mashima, J., Kodama, Y., Lee, K.B., Tsutsui, H.
See the following site: http://www.ddbj.nig.ac.jp/insdc/icm2015-e.html

2-2. Released data from DDBJ in 2015

DDBJ traditional + WGS + TSA

Submission section: Kosuge, Okido,　Lee,　Tsutsui, Aono,　Ejima,

Update section: Sakai, Sugita, Mimura, Aono, Ejima

DRA: Kodama, Fukuda

BioProject: Kodama, Fukuda

BioSample: Kodama, Fukuda

JGA: Kodama

If not specified, statistics cover the period from January to December 2015.

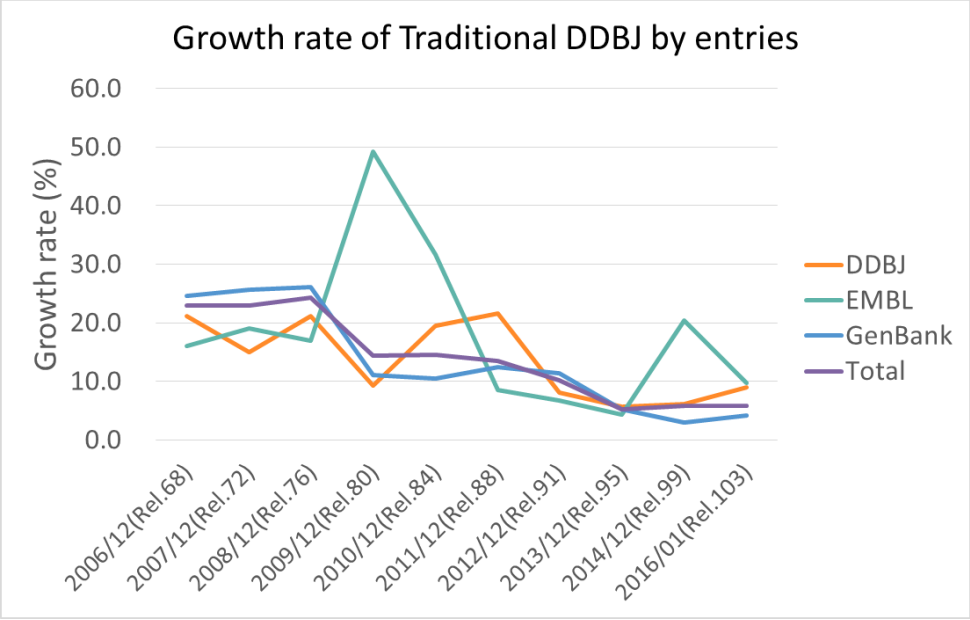
- Submitted to DDBJ; Indicating activities of DDBJ submission section
- Distributed/Redistributed from DDBJ; Indicating activities of update section
- Increment of published data; To consider degree of DDBJ contribution and total INSDC data storage

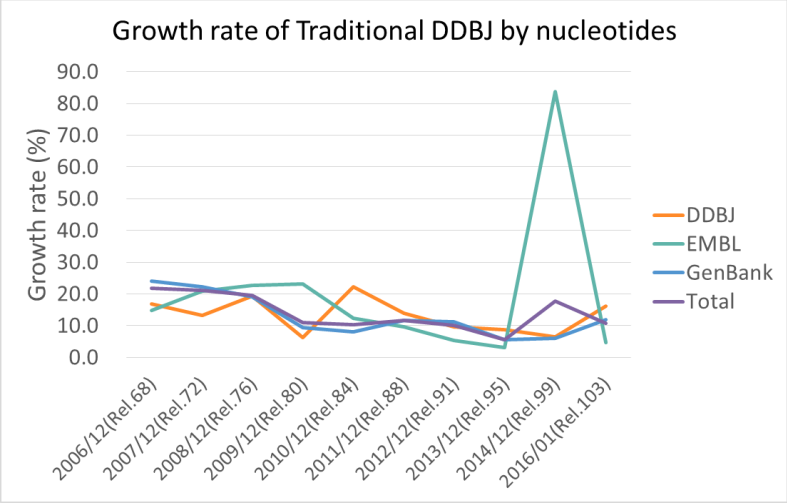
(1) Traditional DDBJ (DDBJ/EMBL-Bank/GenBank)

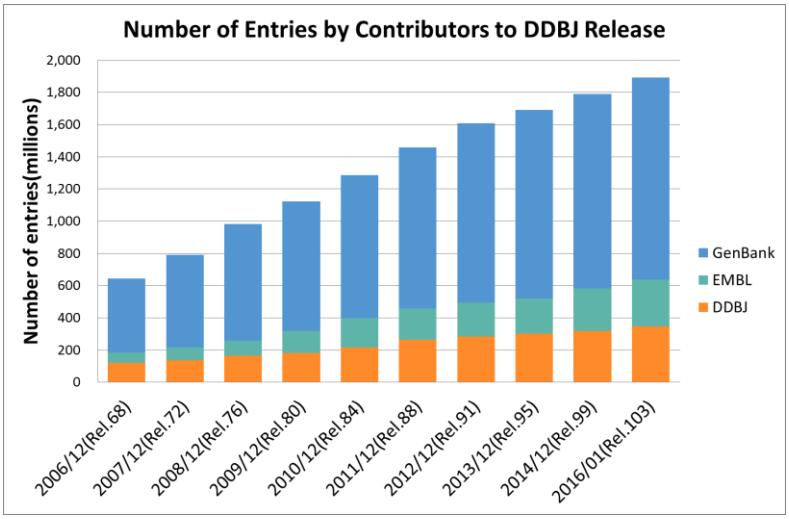
Submitted to DDBJ: 1,156,001 entries (including unpublished data)

Distributed/Redistributed from DDBJ: 2,605,495 entries

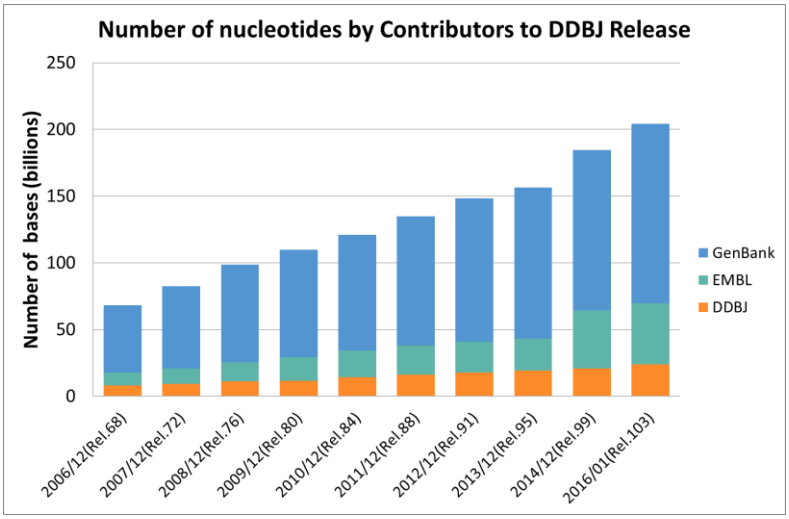
Increment of published data: DDBJ 1,282,165 entries







Number of entries: http://www.ddbj.nig.ac.jp/breakdown_stats/prop_ent-e.html



Number of nucleotides: http://www.ddbj.nig.ac.jp/breakdown_stats/prop_bp-e.html

(2) TSA data

Submitted to DDBJ: 4,218,864 entries (including unpublished data)

Distributed/Redistributed from DDBJ: 3,794,082 entries

Increment of published data: DDBJ 2,127,597 entries

(3) WGS data (draft genome assemblies)

Submitted to DDBJ: 2,834,598 entries (including unpublished data)

Distributed/Redistributed from DDBJ: 16,429,640 entries

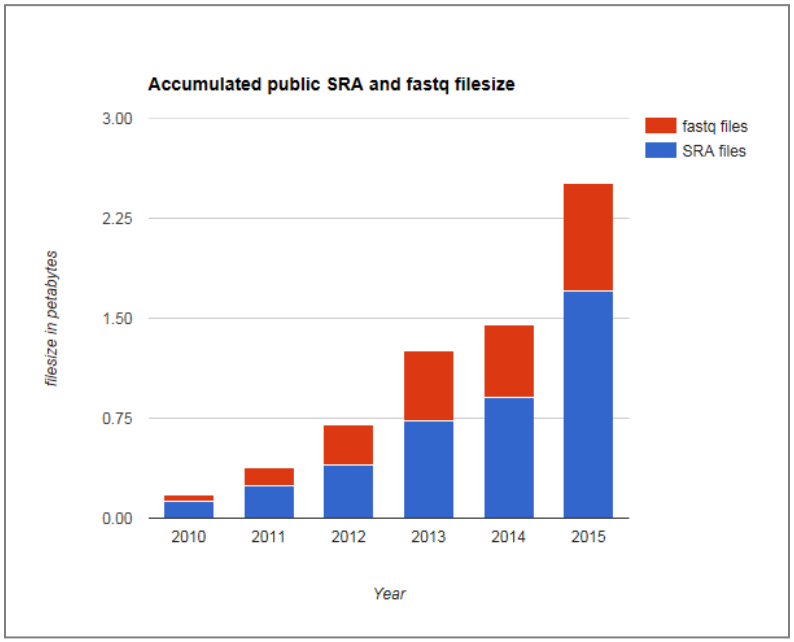
Increment of published data: DDBJ 4,055,937 entries, INSDC 192,593,956 entries

DDBJ: http://www.ddbj.nig.ac.jp/breakdown_stats/sub-wgs-e.html

INSDC: ftp.ddbj.nig.ac.jp/ddbj_database/wgs/WGS_ORGANISM_LIST.html

(4) DRA (DDBJ Sequence Read Archive)

As of 25 January 2016, DRA stored 2.51 petabytes (PB) of sequencing data in the form of SRA (1.7 PB) and fastq (0.81 PB) files. The DDBJ center generates fastq files from our own and exchanged SRA files; however, this process is delayed. Total size increased 1.06 PB compared with last year.



Statistics: DRA total filesize (25 January 2016)

(5) BioProject

In 2015, DDBJ issued accession numbers to 1,044 projects.

(6) BioSample

In 2015, DDBJ issued accession numbers to 20,182 samples.

(7) JGA (Japanese Genotype-phenotype Archive)

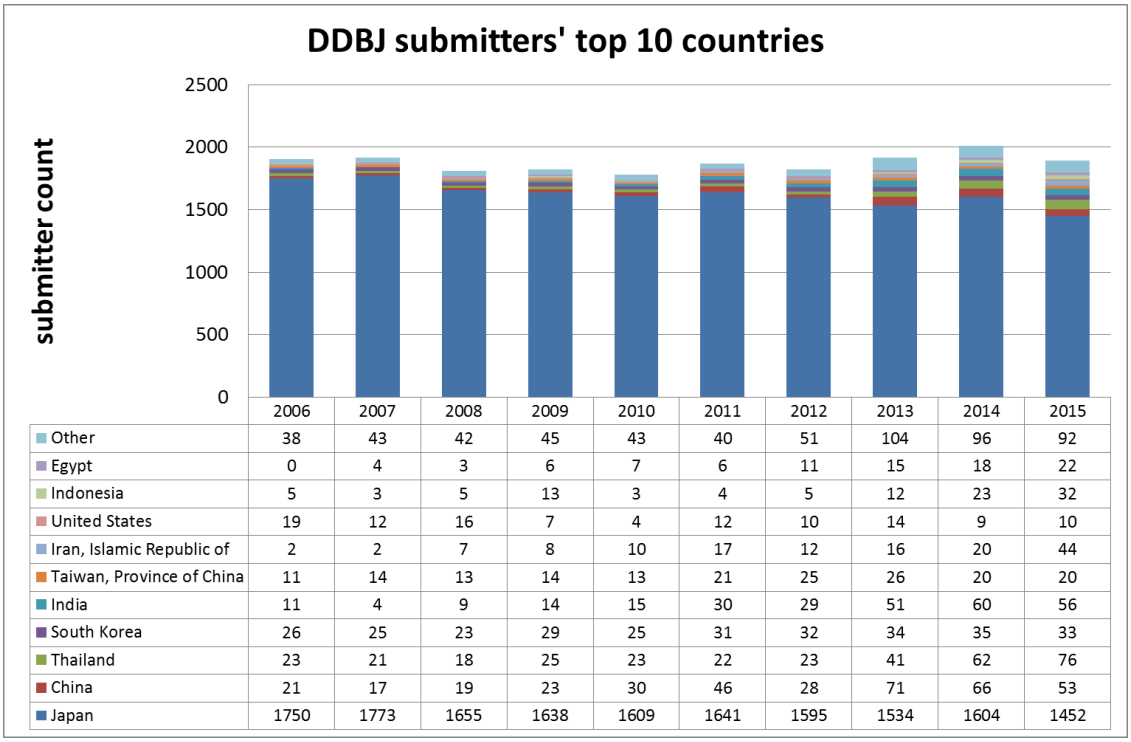
Submission to JGA is measured in terms of the number of samples and submitted file size (TB). Files were submitted in various formats (fastq, bam, microarray CEL, etc.). As of 25 January 2016, JGA had archived 41 studies (6,223 samples and 16.3 TB) and distributed 25 studies, including the cancer exome sequencing, HLA typing, and brain medical image data.

Released JGA studies (25 January 2016): https://ddbj.nig.ac.jp/jga/viewer/view/studies

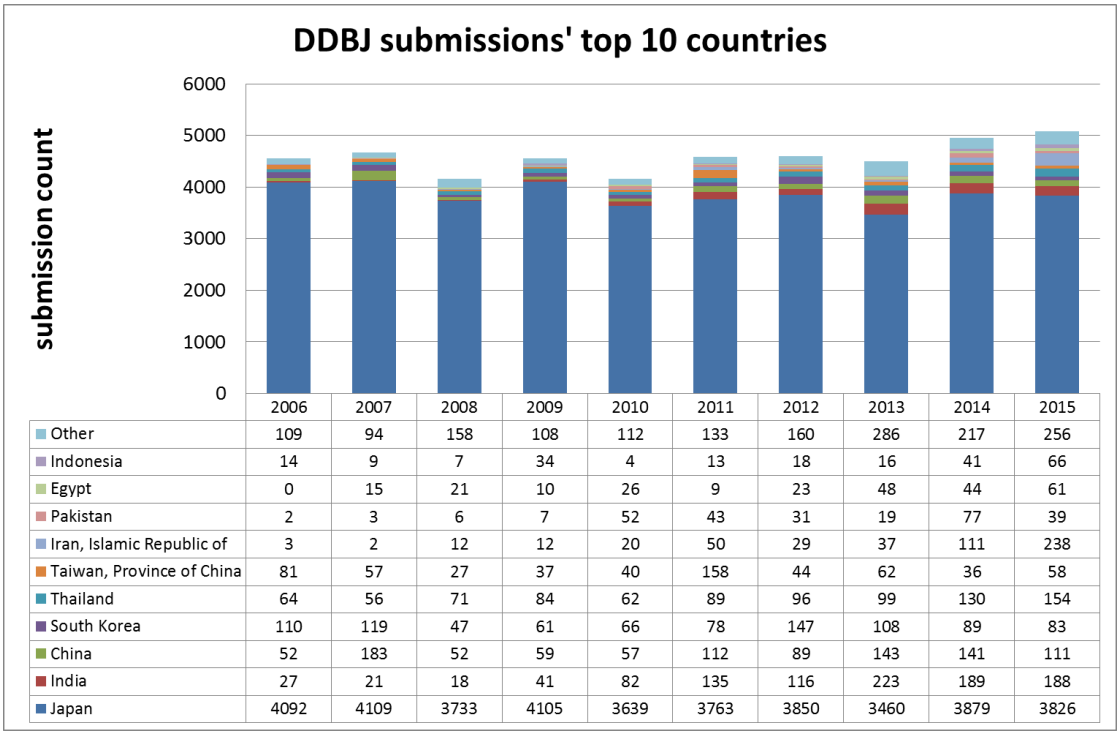
In 2015, two Japanese researchers received approval from NBDC to download the individual-level JGA dataset.

2-3. Statistical analyses of sequence data

(1) Classification of submitters by country



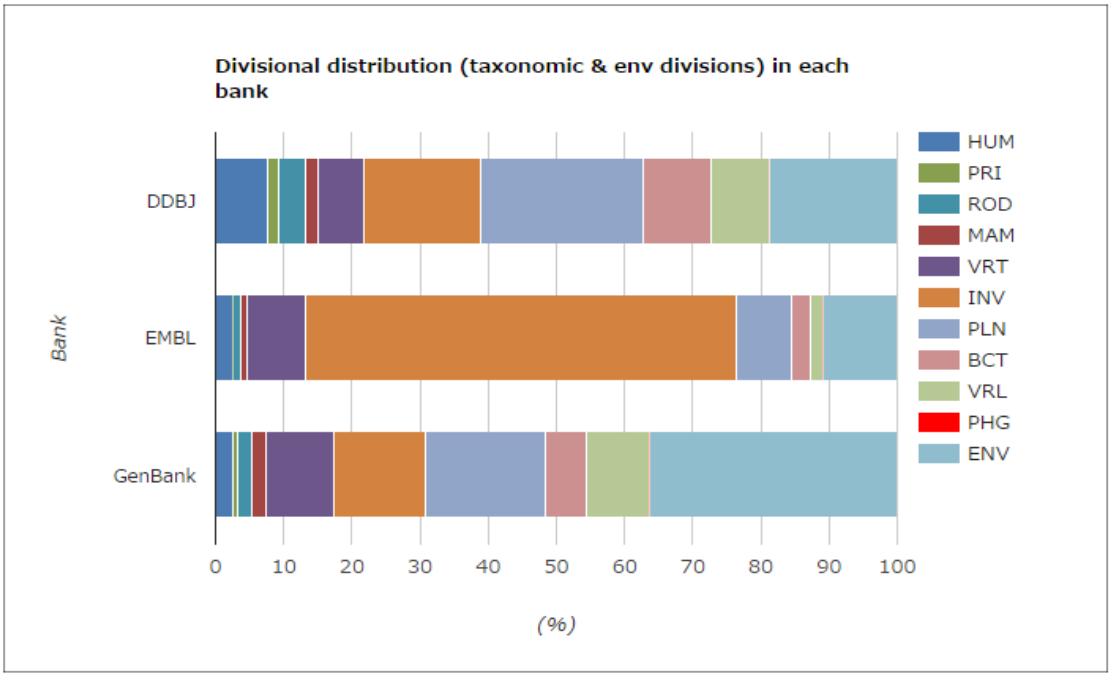
	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015
Other	38	43	42	45	43	40	51	104	96	92
Egypt	0	4	3	6	7	6	11	15	18	22
Indonesia	5	3	5	13	3	4	5	12	23	32
United States	19	12	16	7	4	12	10	14	9	10
Iran, Islamic Republic of	2	2	7	8	10	17	12	16	20	44
Taiwan, Province of China	11	14	13	14	13	21	25	26	20	20
India	11	4	9	14	15	30	29	51	60	56
South Korea	26	25	23	29	25	31	32	34	35	33
Thailand	23	21	18	25	23	22	23	41	62	76
China	21	17	19	23	30	46	28	71	66	53
Japan	1750	1773	1655	1638	1609	1641	1595	1534	1604	1452

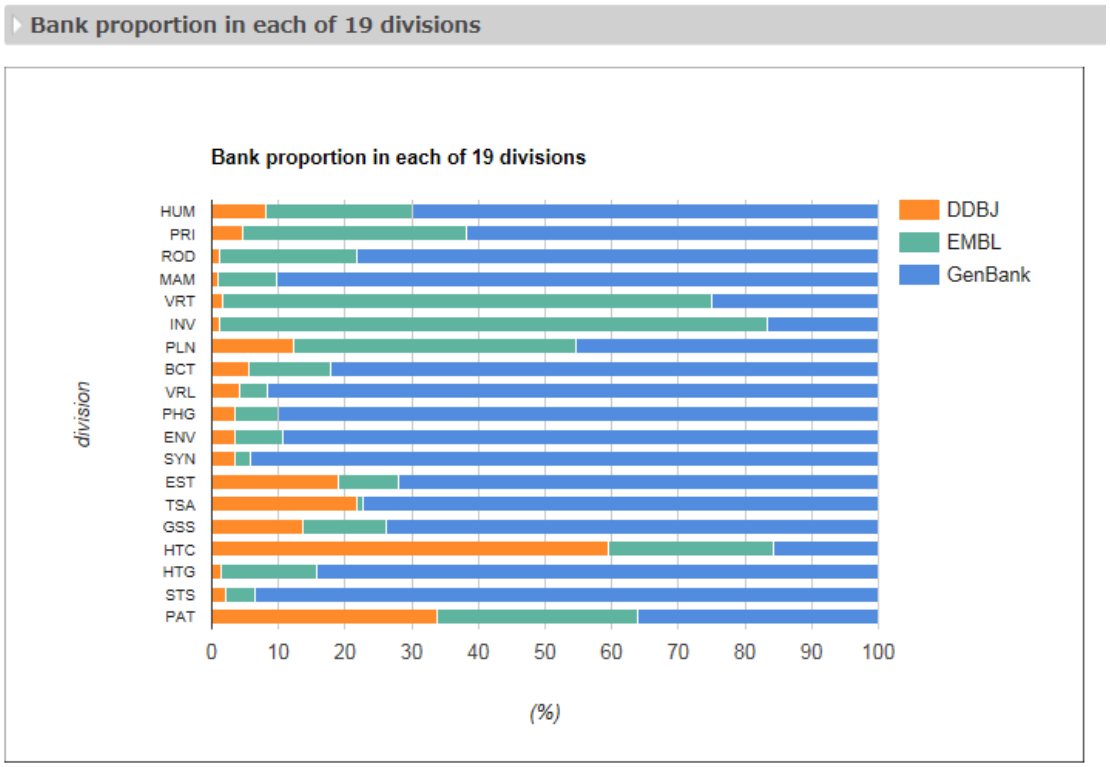


	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015
Other	109	94	158	108	112	133	160	286	217	256
Indonesia	14	9	7	34	4	13	18	16	41	66
Egypt	0	15	21	10	26	9	23	48	44	61
Pakistan	2	3	6	7	52	43	31	19	77	39
Iran, Islamic Republic of	3	2	12	12	20	50	29	37	111	238
Taiwan, Province of China	81	57	27	37	40	158	44	62	36	58
Thailand	64	56	71	84	62	89	96	99	130	154
South Korea	110	119	47	61	66	78	147	108	89	83
China	52	183	52	59	57	112	89	143	141	111
India	27	21	18	41	82	135	116	223	189	188
Japan	4092	4109	3733	4105	3639	3763	3850	3460	3879	3826

(2) Classification of entries by divisions and banks

Divisional distribution in each bank (taxonomic & env divisions)





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

3. SYSTEM DEVELOPMENT FOR SUBMISSION

3-1. Improvement of accession number assignment system for TSA submissions

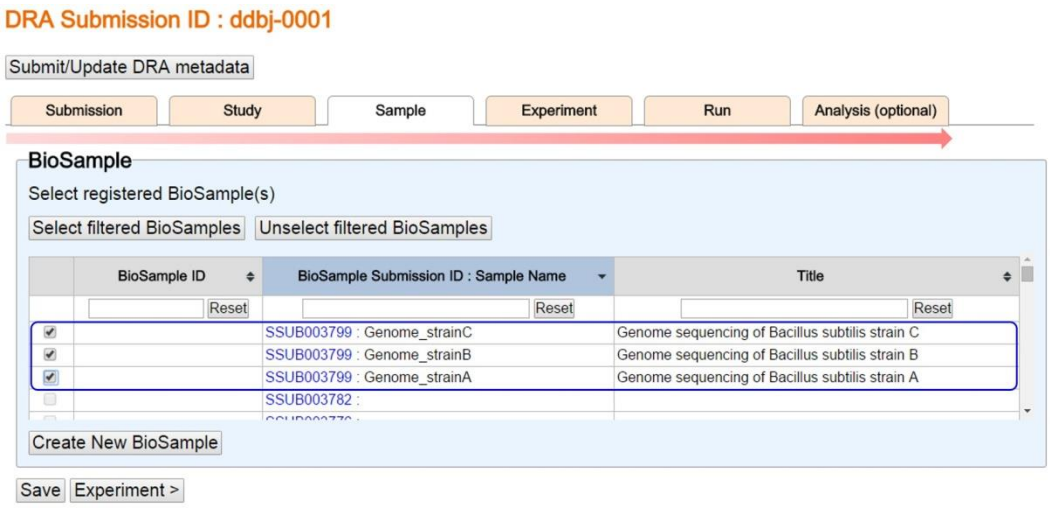
[Mashima, Team SE]

In 2015, data volume of Transcriptome Shotgun Assembly (TSA) submissions to DDBJ dramatically increased with individual submissions of 100,000 sequences. Thus, we decided to improve the DDBJ accession number assignment system to accept such bulk TSA submissions. Since October 2015, DDBJ has assigned accession numbers with four-letter prefixes for TSA data submitted to DDBJ the same as that for WGS data. In November 2015, DDBJ released TSA data with a four-letter prefix IAAA (IAAA01000001 - IAAA01132843) for the first time. See also prefix letter list for large-scale data and the anonymous FTP site of TSA data: ftp://ftp.ddbj.nig.ac.jp/ddbj_database/tsa/

3-2. New functions for DRA/BioProject/BioSample submission systems

[Kodama, Fukuda, Sakon, Fujimoto, Watanabe]

In April 2015, we released the upgraded BioProject/BioSample/DRA submission system. This system enables users to submit a DRA submission referencing submitted but as yet un-accessioned BioProjects and BioSample objects; thus, users do not need to wait for BioProject and BioSample accession numbers before submitting sequencing data to DRA. This upgrade was funded by the Genome Science Project of MEXT.



New BioProject/BioSample/DRA submission interface. The new BioProject, BioSample, and DRA submission interface enables submission of DRA metadata objects (Submission, Experiment, Run, and Analysis) referencing submitted but as yet un-accessioned BioProject and BioSample objects. Users can submit new BioProject, BioSample, and DRA metadata at the same time.

3-3. JGA system

[Kodama, Sato, Sakon, Fujimoto, Watanabe]

We have released a stand-alone JGA data file encryption tool. This tool allows submitter to submit large-scale data securely by sending a hard disk drive. We have established a remote JGA data backup workflow. We will back up the entire JGA file and system to Tohoku Medical Megabank Organization (ToMMo) in 2016.

4. SEQUENCE ANALYTICAL SERVICES

DDBJ provides sequence analytical services to the research community, which are implemented in the NIG supercomputer. These are classified according to interface into two types: command line interface and web-based graphical user interface (GUI).

4-1. Command line interface-based tools and databases

[Kawagoe and SE team]

By connecting to the NIG supercomputer, analytical services are available from a command line interface. The user connects to the NIG supercomputer by the SSH command and can immediately use various analytical tools (shown in “(1) Analytical tools”) with various databases (shown in “(2) Databases”). The system environment utilizes the Univa Grid Engine (UGE) job management system so that multiple users can efficiently share it.

(1) Analytical tools

☐ ALLPATHS-LG	☐ EDENA	☐ Oases	☐ SOAPdenovo	☐ Trinity
☐ Velvet	☐ spiral			
☐ Bowtie	☐ Bowtie2	☐ BWA	☐ SOAP	☐ SOAP3
☐ SOAP3-dp	☐ SSAHA2	☐ TopHat	☐ TopHat2	
☐ Cufflinks	☐ RSEM	☐ SRA-Tools	☐ gmap	
☐ Cutadapt				
☐ MACS				
☐ augustus	☐ bamtools			
☐ BLAT	☐ mothur	☐ NCBI-BLAST+	☐ NCBI-CXX-toolkit	☐ NCBI-toolkit
☐ FASTA	☐ InterProScan	☐ MAFFT		
☐ ClustalW	☐ ClustalW2	☐ PHYLIP	☐ RAxML	☐ Mrbayes5d
☐ HMMER				
☐ Pindel				
☐ GATK				
☐ BEDTools	☐ IGVViewer	☐ Picard	☐ SAMtools	☐ circos
☐ express				
☐ ACML-GFortran	☐ ACML-ifort	☐ ACML-PCI	☐ GMP	☐ GROMACS-GPU
☐ GSL	☐ Libtool	☐ libxslt	☐ MPC	☐ MPFR
☐ OpenMM	☐ OpenMPI			
☐ CUDA	☐ GCC	☐ Java	☐ Perl	☐ Python
☐ R	☐ Ruby	☐ pdftk		

(2) Databases

☐ DDBJ-unified-all	☐ DDBJ-unified-new	☐ GenBank	☐ GenPept
☐ GenPept-UPD	☐ NCBI-nt	☐ NCBI-nr	☐ NCBI-dbEST
☐ NCBI-dbGSS	☐ NCBI-HTGS	☐ NCBI-patnt	☐ NCBI-STS
☐ NCBI-Taxonomy	☐ NCBI-WGS	☐ RefSeq-Genomic	☐ RefSeq-Protein
☐ RefSeq-RNA	☐ RefSeq-UPD-AA	☐ RefSeq-UPD-NA	☐ EMBL
☐ EMBL-UPD	☐ PDB	☐ Pfam	☐ Swiss-Prot
☐ TrEMBL	☐ UniProt		

4-2. Web GUI-based tools

(1) WebBlast, VecScreen, TXSearch, and ClustalW.
[Ogasawara, Kosuge, Watanabe, Okubo]
DDBJ restored a web-based BLAST tool with its original graphical user interface. The original package was NCBI BLAST 2.2.25 and the prepared blast databases span all of DDBJ's databases, including patents and the latest daily sequences. A selection box for INSDC division is also available.
VecScreen, which is useful for determining vector sequence contamination from nucleotide sequences, has been released. Its main program is NCBI vecscreen, and we have equipped this as an API that works on the NIG supercomputer system.
TXSearch is a DDBJ-developed taxonomy database retrieval system that was unified by DDBJ, GenBank, and ENA. This system is expected to be helpful as a reference for taxonomic names when researchers submit nucleotide sequences to the DDBJ. TXSearch utilizes the NCBI taxonomy database, and thus taxonomic accessions.
ClustalW, based on the original ClustalW 2.1 package, is also available as one of the DDBJ services. The DDBJ's ClustalW system contains original matrices developed at the National Institute of Genetics for genetics research.

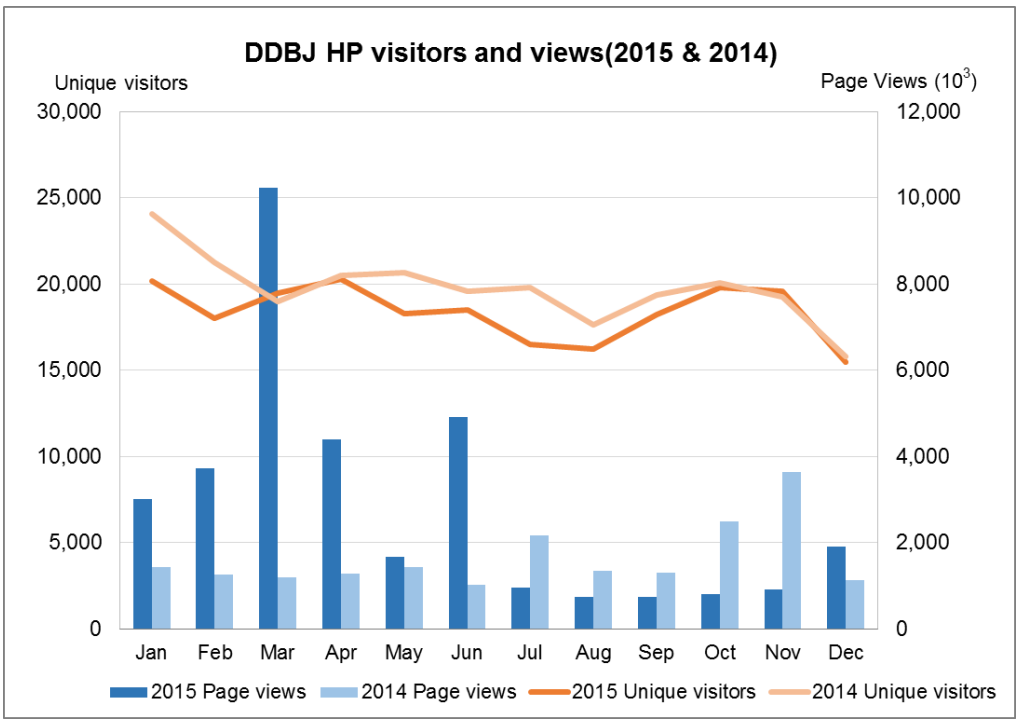
(2) MiGAP: Microbial Annotation Server
[Sugawara, Kurokawa from TITECH]
MiGAP (Microbial Genome Annotation Pipeline) provides novices and veterans alike with a mechanical annotation to microbial contigs and genomes. MiGAP identifies ORFs and RNA regions and infers the functions of ORFs by referring to popular public databases. MiGAP UserID links NIG supercomputer accounts.

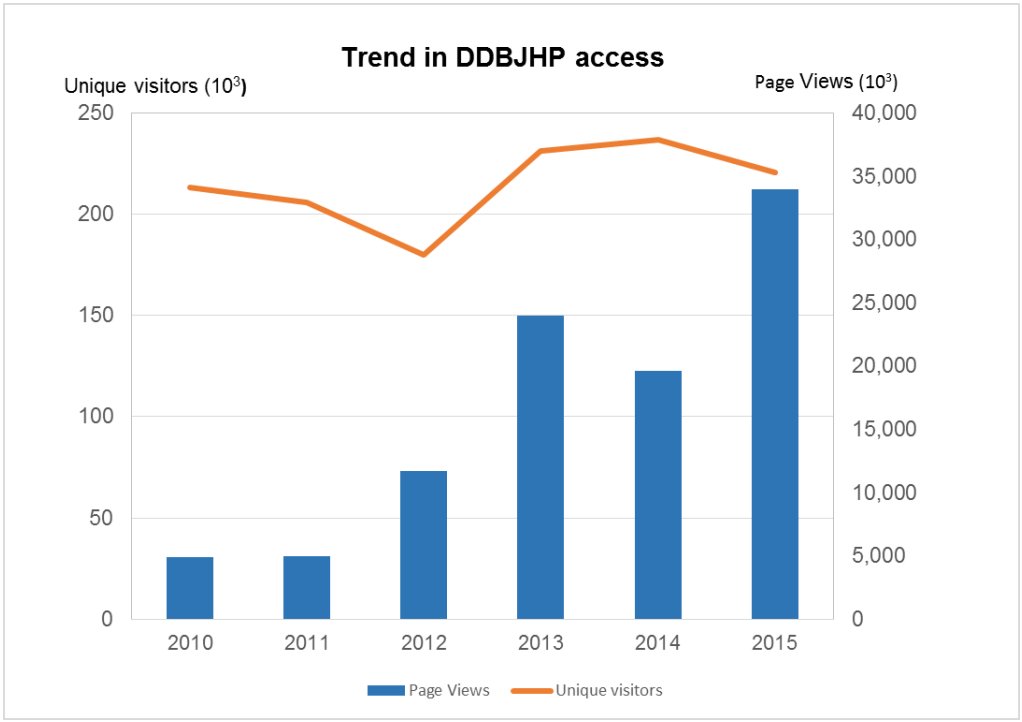
(3) DDBJ Pipeline: NGS analysis server
[Nagasaki, Mochizuki, Tanizawa, Saka, Kaminuma, Nakamura]
The DDBJ Read Annotation Pipeline (DDBJ Pipeline) is an NGS analytical system based on the NIG supercomputer. The DDBJ Pipeline began operation in 2009 and, since that time, 977 worldwide users have been registered (an increase of 40% over the last year). In 2015, a total of 5,637 jobs (a decrease of 20%) for mapping/denovo/pre-processing analysis were performed using this system. In addition, the total number of individual jobs on galaxy workflows was 6,481 (a decrease of 38%).

5. USAGE STATISTICS OF DDBJ WEB SERVICES

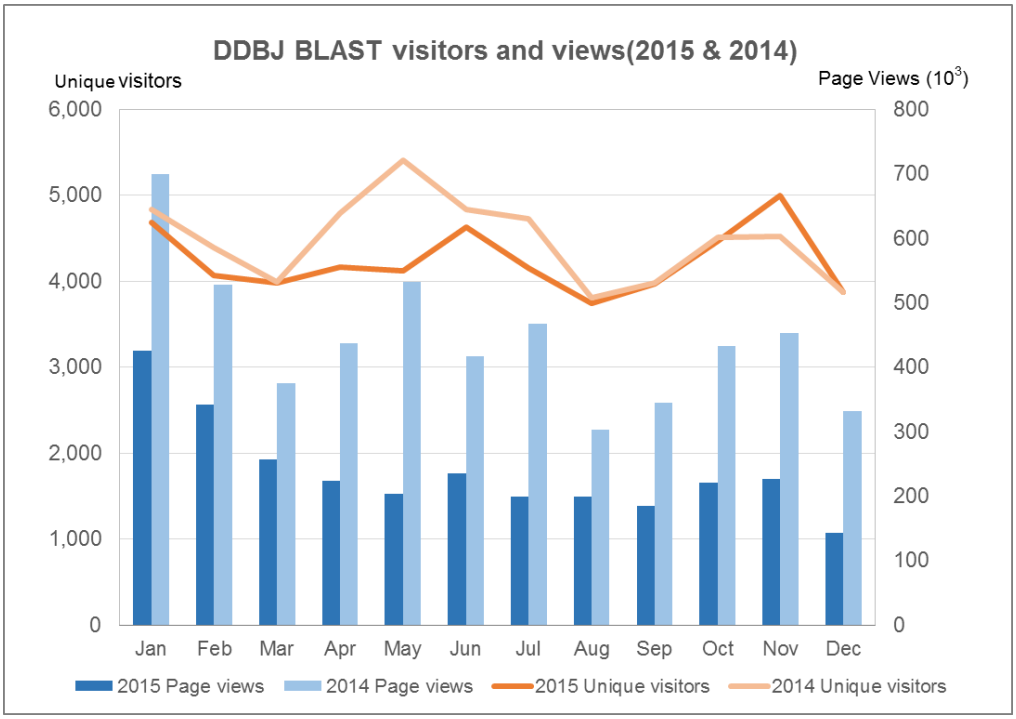
5-1. DDBJ HP, BLAST, CLUSTALW, ARSA (keyword search)

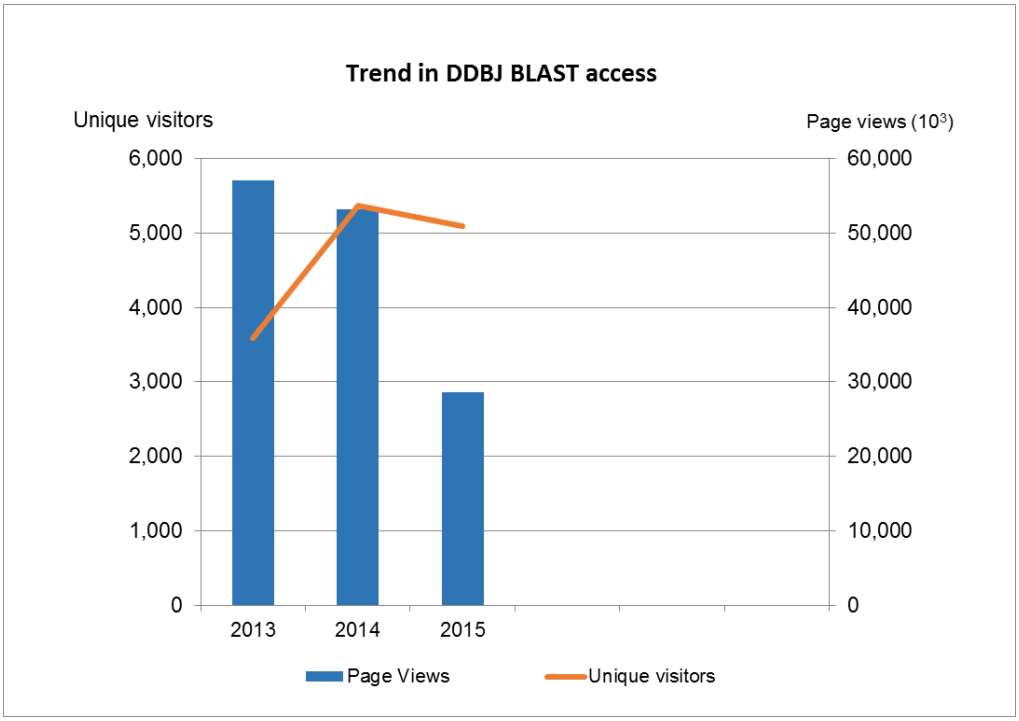
(1) DDBJ Home Page



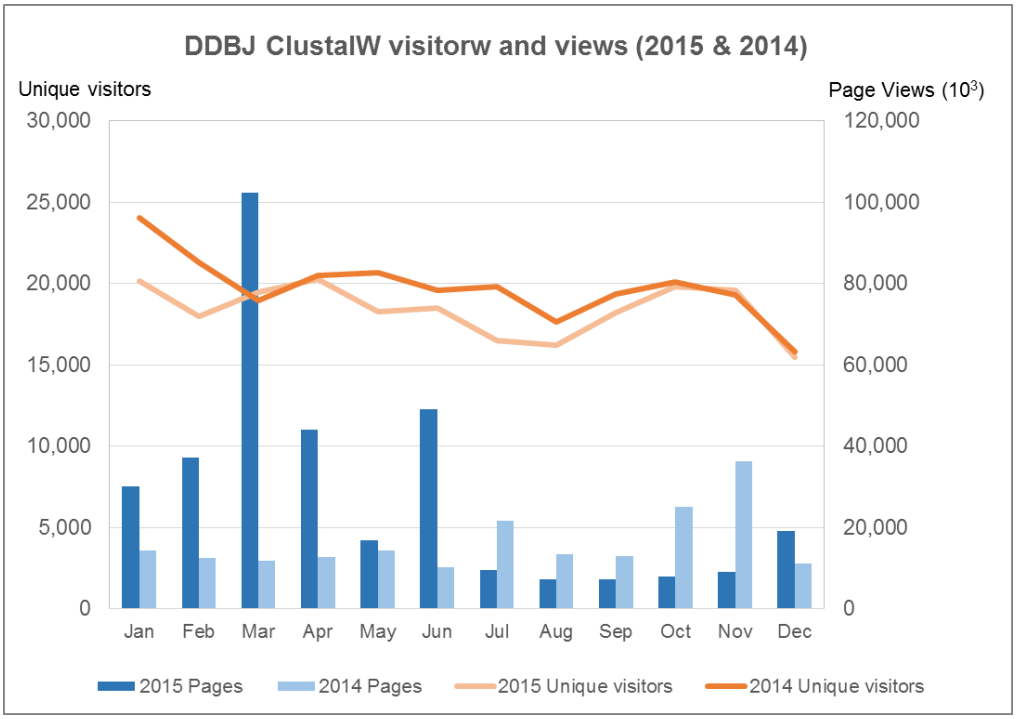


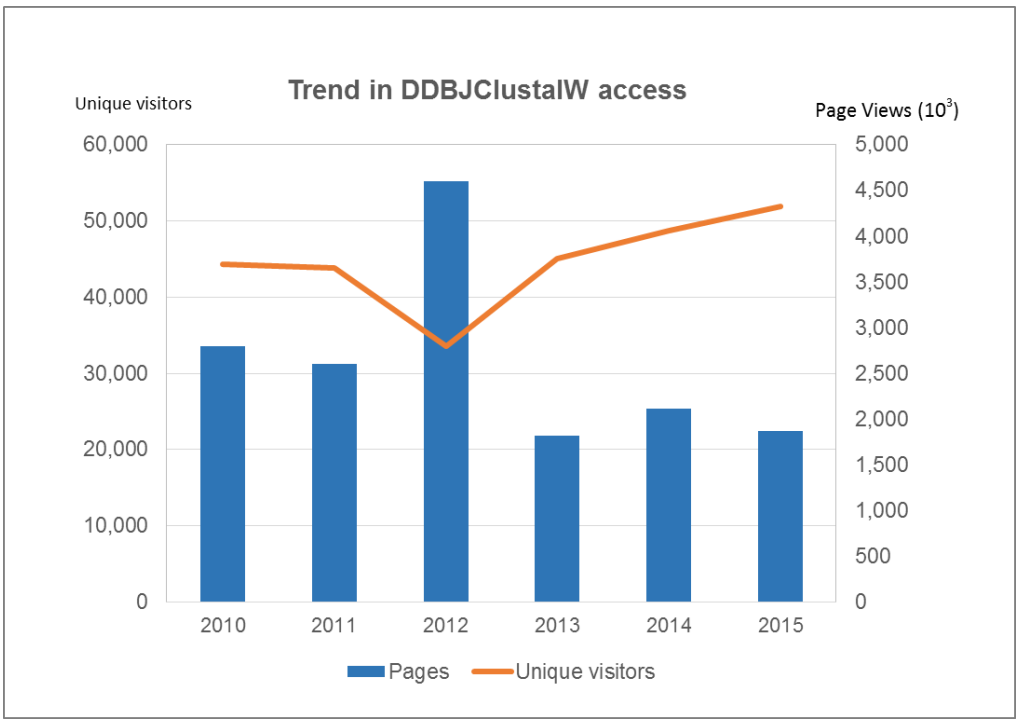
(2) BLAST


DDBJ BLAST visitors and views(2015 & 2014)
Unique visitors
Page Views (10^3)
6,000
5,000
4,000
3,000
2,000
1,000
0
800
700
600
500
400
300
200
100
0
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
2015 Page views
2014 Page views
2015 Unique visitors
2014 Unique visitors

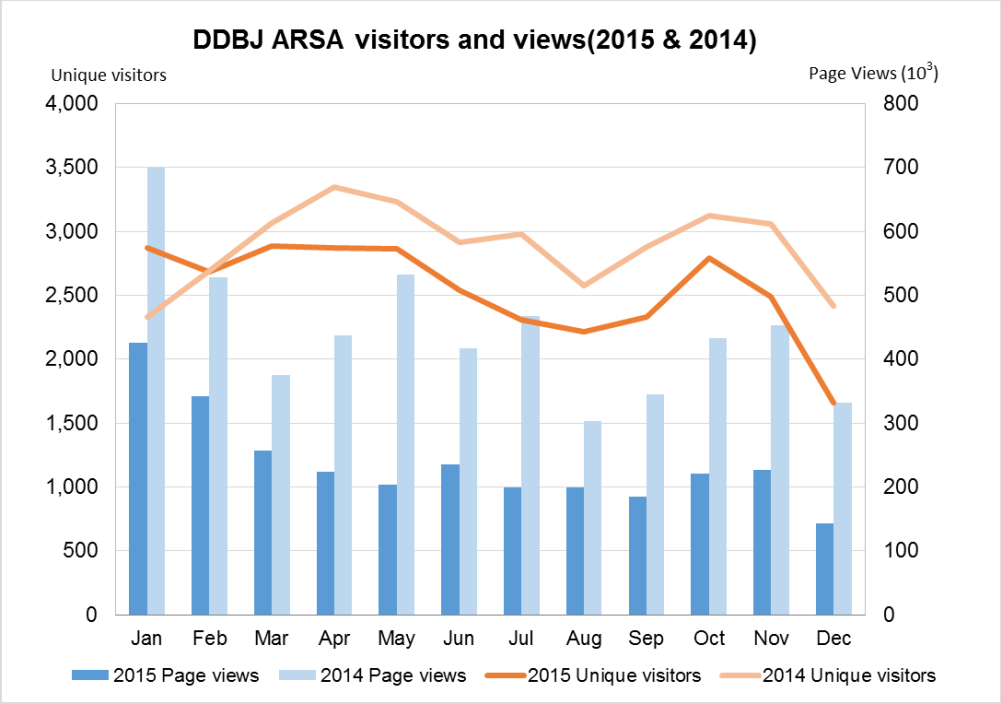

Trend in DDBJ BLAST access
Unique visitors
Page views (10^3)
6,000
5,000
4,000
3,000
2,000
1,000
0
60,000
50,000
40,000
30,000
20,000
10,000
0
2013
2014
2015
Page Views
Unique visitors

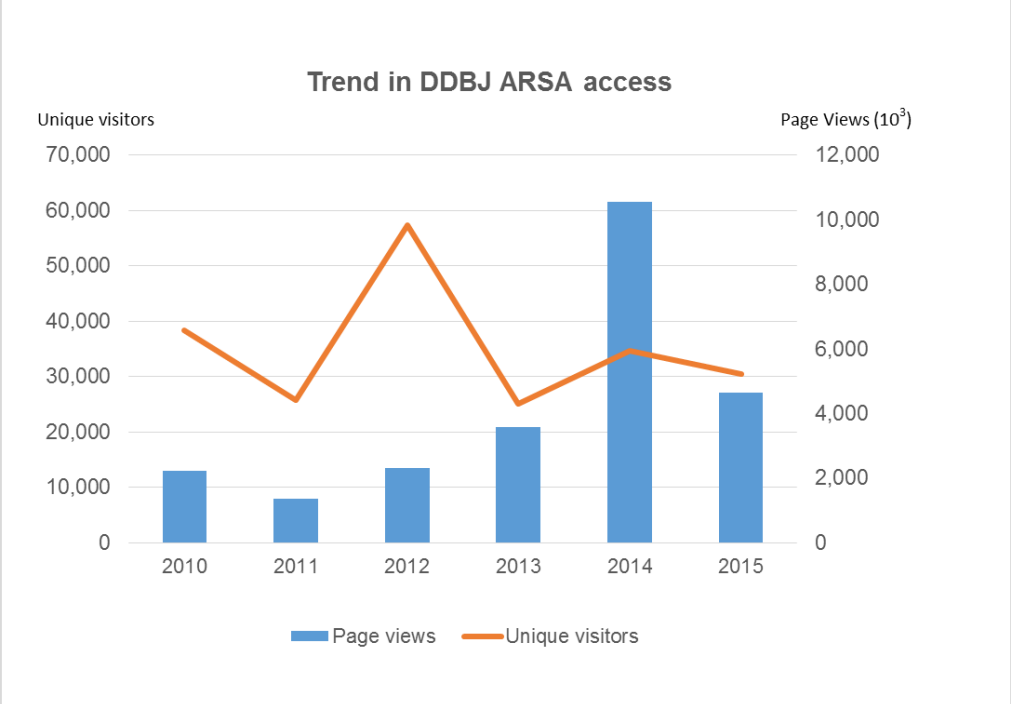
(3) ClustalW


DDBJ ClustalW visitorw and views (2015 & 2014)
Unique visitors
Page Views (10^3)
30,000
25,000
20,000
15,000
10,000
5,000
0
120,000
100,000
80,000
60,000
40,000
20,000
0
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
2015 Pages
2014 Pages
2015 Unique visitors
2014 Unique visitors


Trend in DDBJClustalW access
Unique visitors
Page Views (10^3)
60,000
50,000
40,000
30,000
20,000
10,000
0
5,000
4,500
4,000
3,500
3,000
2,500
2,000
1,500
1,000
500
0
2010
2011
2012
2013
2014
2015
Pages
Unique visitors

(4) ARSA (keyword search system)

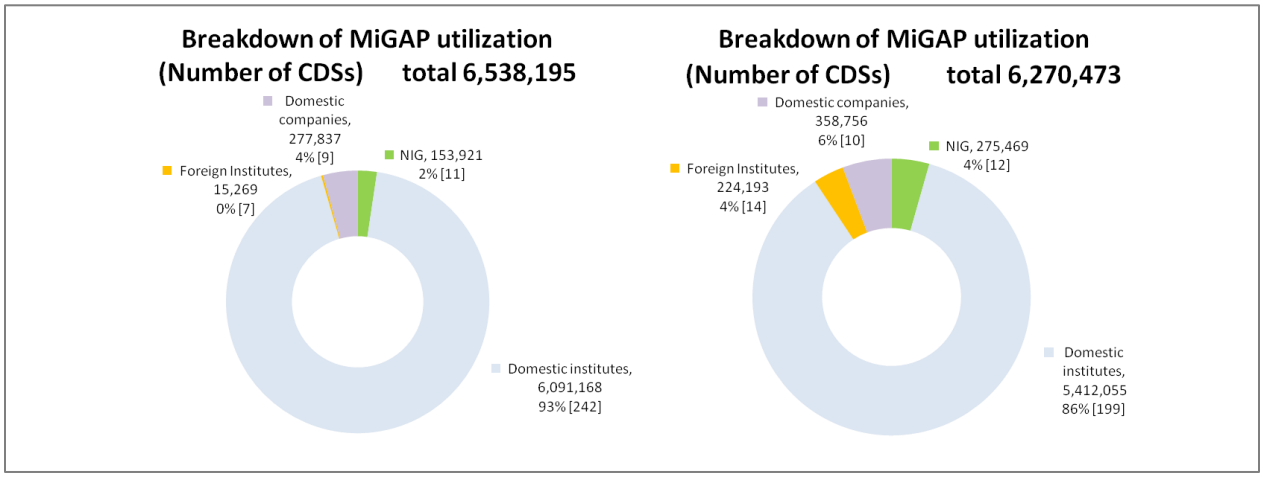




5-2. Sequence annotation analysis (MiGAP, DDBJ pipeline)

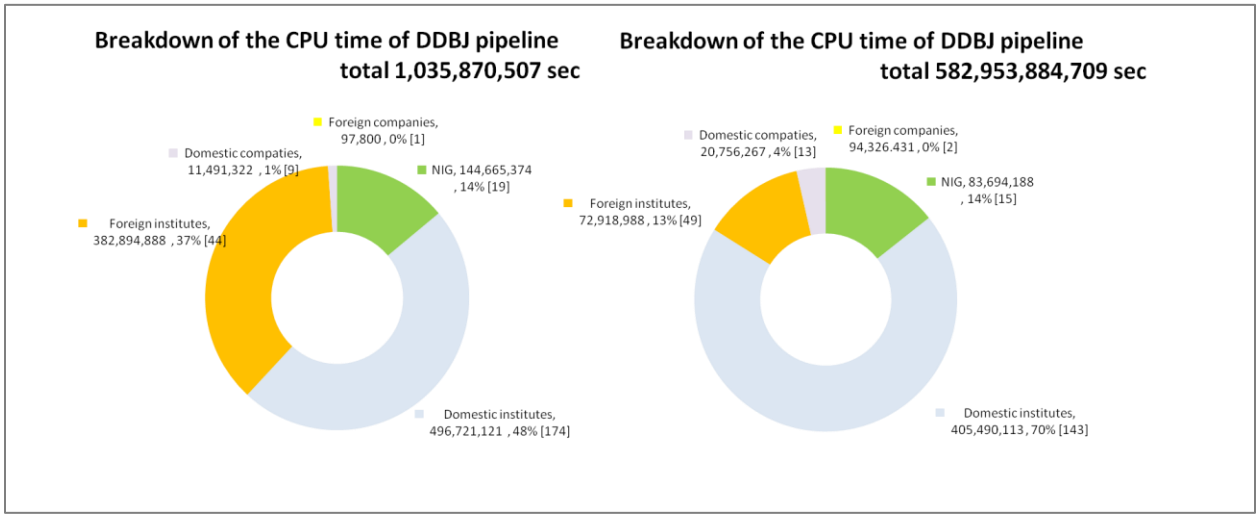
(1) Breakdown of MiGAP utilization (Number of CDSs)

2015　　2014



(2) Breakdown of CPUtime of DDBJ pipeline

2015　　2014



6. NIG SUPERCOMPUTER SYSTEM

(Through this chapter, “Phase I (abbrev. Ph 1 or Ph I) ” is an initial installed part introduced in March 2012, and “Phase II (abbrev. Ph 2 or Ph II)” is an enhanced part installed in March 2014)

6-1. Overview of the NIG Supercomputer System

The purposes of the NIG supercomputer are (1) construction of the DDBJ database and archiving

nucleotide sequence data; (2) analysis of the sequence data, e.g., computational annotation of genome sequences, indexing, and searching the databases; and (3) provision of computational resources to researchers (both inside and outside of NIG).

In order to cope with the ongoing genomic data deluge, in March 2012, our previous computer system was completely replaced by a commodity cluster-based system that boasts 122.5 TFlops of CPU capacity and 5 PB of storage space. During this upgrade, it was considered crucial to replace and refactor substantial portions of the DDBJ software systems as well. As a result of the replacement process, which took more than 2 years to perform, we have achieved significant (about 15-fold relative to the 2007 NIG supercomputer system) improvements in system performance. In order to meet the increasing electricity demand of NIG, the central substation of NIG had to be enhanced (from 2 MW to 2.4 MW or more) prior to the full deployment of the NIG supercomputer system. Accordingly, the installation process had to be divided into two phases, the initial installation on 1 March 2012 and the system enhancement on 1 March 2014, which roughly doubled all aspects of the newly installed system excluding the fat node. Calculation nodes within each subsystem are interconnected with full-bisection fat tree topology, and the subsystems (the phase I system and the phase II system) have 40 Gbps of connection. (Figure 6-1)

Since the system enhancement, the fully deployed NIG supercomputer system consists of (1) a distributed memory cluster system (64 GB memory/node); (2) shared memory (non-uniform memory access, NUMA) computer systems (10 TB memory fat node and 2 TB memory medium nodes for the use of memory-intensive calculation including *de novo* assembly of NGS data); (3) a 2 PB high-speed storage system for general calculation (Lustre file system); (4) a 3 PB power saving storage system (massive arrays of inactive disks, MAID) for the DDBJ sequence data archive; and (5) accelerators/coprocessors (64 nodes of NVIDIA Tesla M2090, 64 nodes of NVIDIA Tesla K20, and 32 nodes of Intel Xeon Phi 5110P).

The theoretical peak performance of the CPUs in the initial installation system (Phase I) is 128.4 TFlops (maximum performance Rmax is 82.90 TFlops, only using the thin nodes, Rpeak 117 TFlops), which ranks our system as 280th in the Supercomputer Top 500 list (as of June 2012, http://www.top500.org/).

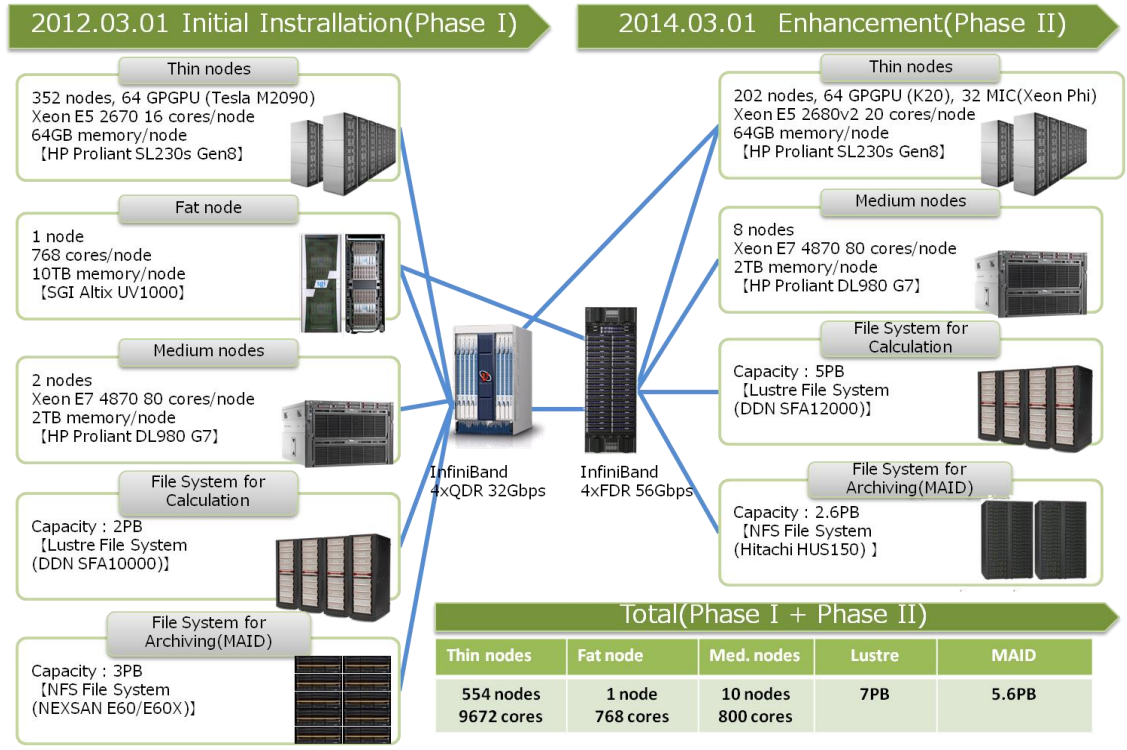


Figure 6-1. Overview of the NIG supercomputer system

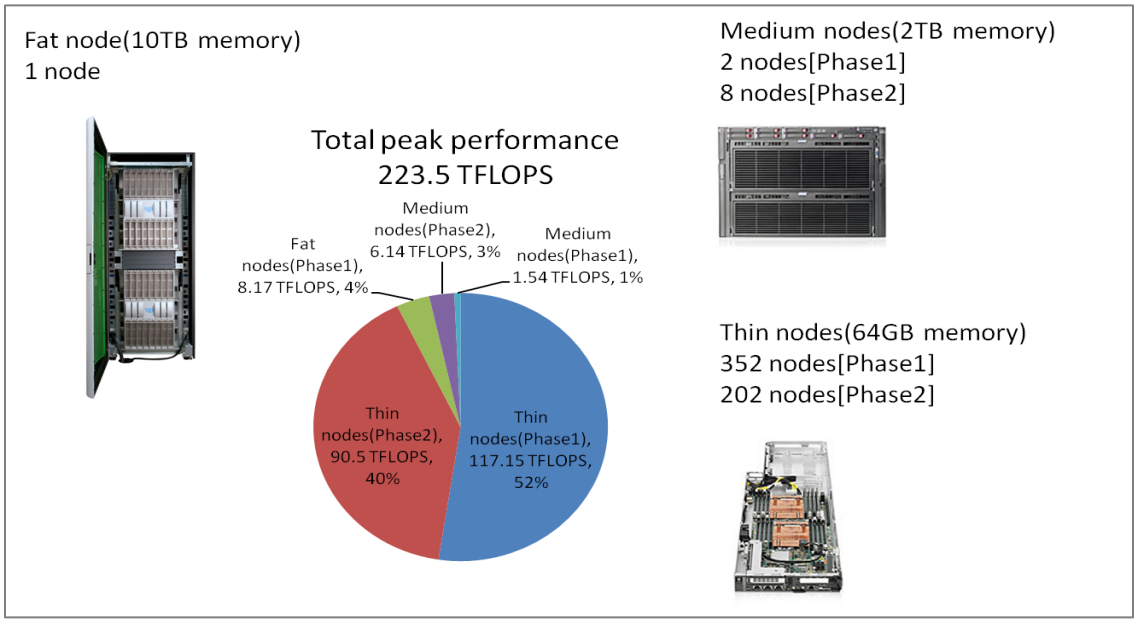


Figure 6-2. Proportions of the peak performances of calculation nodes.

Electric power consumption

Table 6-1. Power consumption of the NIG supercomputer and the capacity of the power receiving facilities of NIG.

	Rated power consumption	Actual power consumption	Contract demand of NIG
Previous system (Phases I + II)	517.9 kW Excluding AC*1	About 450 kW	2,119 kW
New system (Phase I)	537.76 kVA	About 300 kW	2,119 kW
New system (Phases I + II)	1004.63 kVA	About 600 kW	2,700 kW*2

*1 Air conditioner

*2 Electric power receiving facilities of NIG is enhanced in 2014.

2015 2014

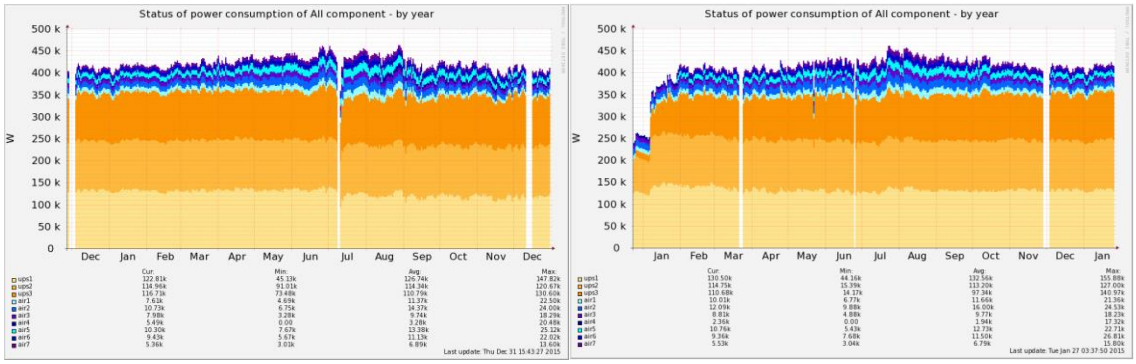


Figure 6-3. Power consumption of the NIG supercomputer (2015 and 2014).
Blues indicate air conditioner systems and oranges indicate computer systems.

2015 2014

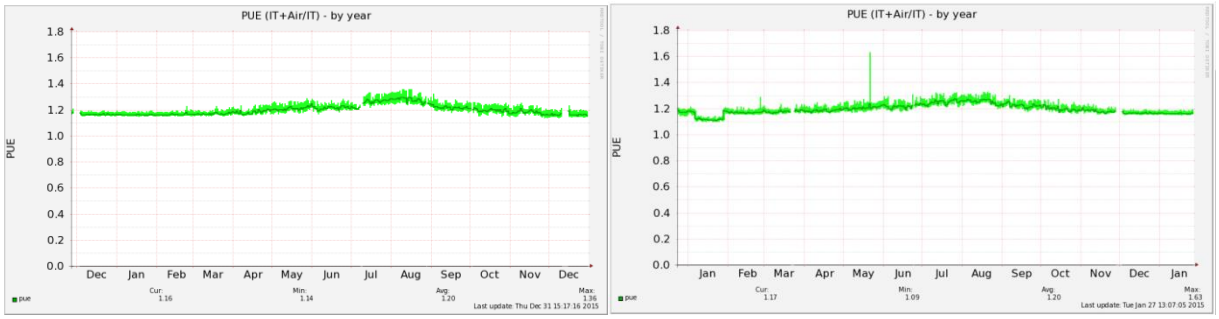


Figure 6-4. Power usage effectiveness (PUE) (2015 and 2014).
PUE is defined as (power consumption of AC + Computer)/Computer.

6-2. User Account Policy

With the replacement of the NIG supercomputer system, we have created a draft of a new user account policy to define the types of user accounts and the processes for creating/terminating these accounts. This policy states that NIG provides the resources of the supercomputer system to the users for research purposes. Users may include not only students and professional researchers at academic institutes, but also researchers at private companies. The usage purposes and year-end reports of all users will be openly displayed on the home page of the NIG supercomputer.

The following have been defined as categories of users.

Users are categorized into the following 5 groups

(1) Login users (normal): Normal login users are given access to the gateway node of the NIG supercomputer systems via SSH. The normal login users may use interactive nodes and submit their jobs to a batch queuing system. The users may use up to 1 TB of storage for calculation. Foreign users are restricted from using the supercomputer owing to the Export Administration Regulations.

(2) Login user (large scale): Login users who need storage spaces exceeding the limit of the normal login users must fill out an application to become a login user of large-scale analysis that includes a written statement of reasons.

(3) Large-scale web application users (MiGAP and DDBJ pipeline)

Web user (MiGAP and DDBJ pipeline): These users may use web applications running on the NIG supercomputer system but may not login to the supercomputer. This type of account is provided for the purpose of relaxing the requirement for creating a user account for the benefit of users whose usage is limited to web applications.

(4) DDBJ accounts: for DDBJ staff including the annotators and the DB construction team.

(5) System Engineer's accounts: System engineers who have root privileges on the system.

6-3. Workload Analysis

(1) Number of user accounts

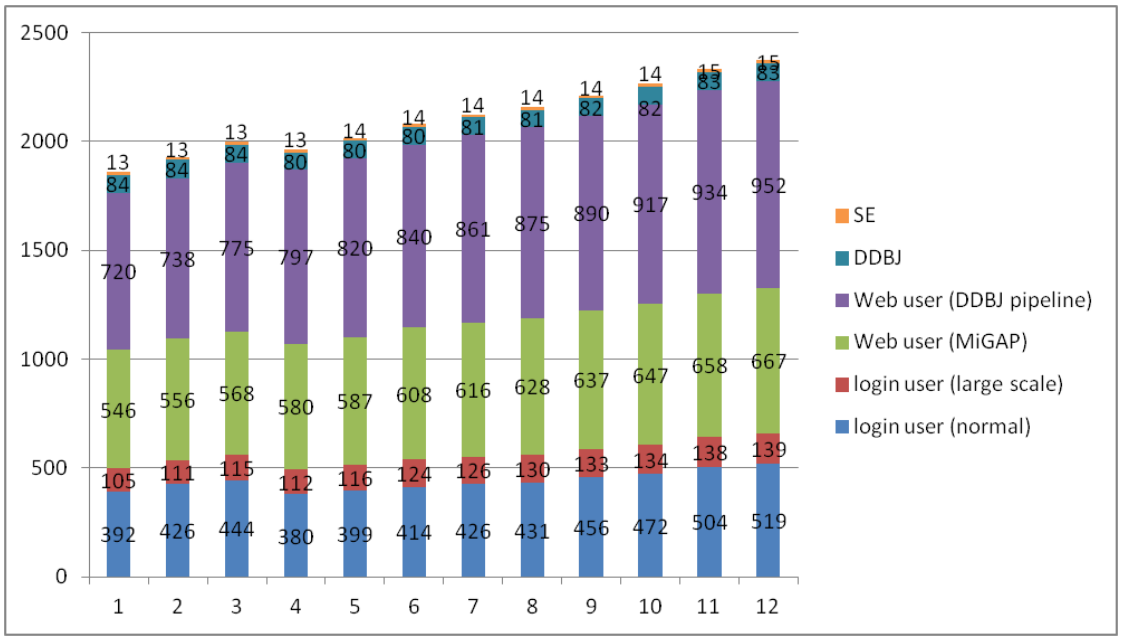


Figure 6-5. Number of user accounts (2015).

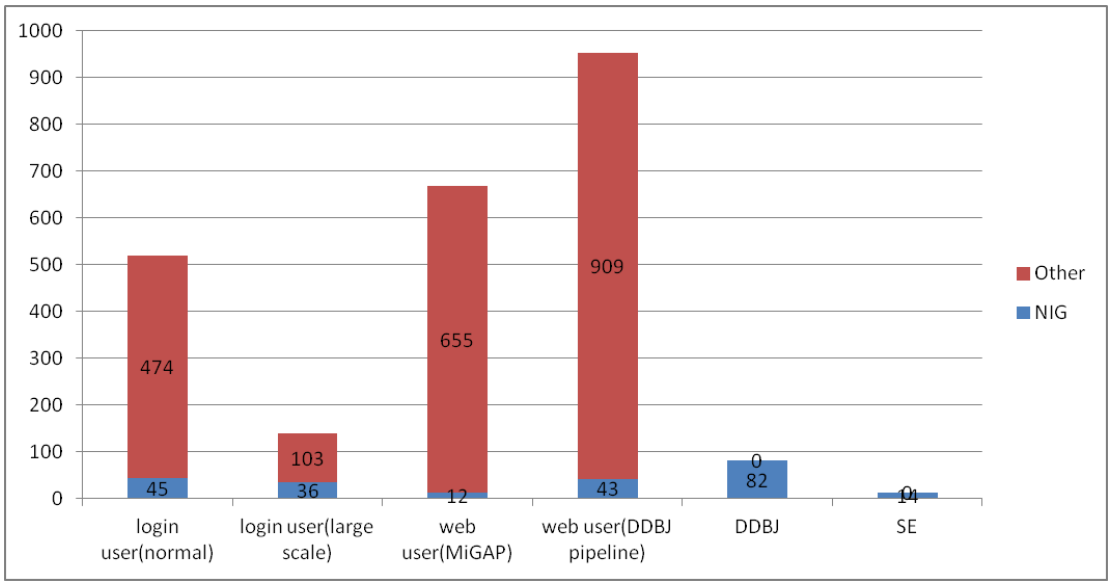


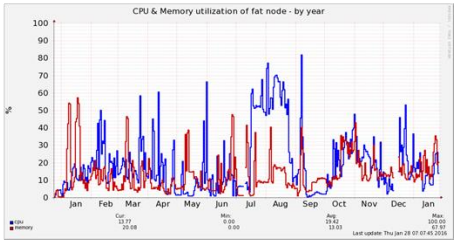
Figure 6-6. Ratio of user affiliation (NIG and others) for each user type (December 2015)

(2) CPU, memory and storage utilization rates (2015)

Fat node: 1 node

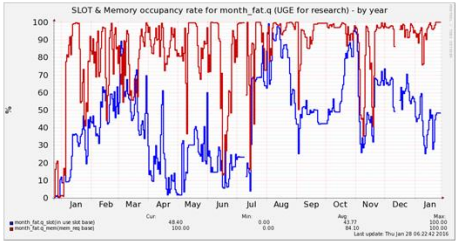
SGI Altix UV1000
Intel Xeon E7-8837,768 cores/node
10 TB memory/node

CPU & Memory Utilization



CPU : 19.4%
Memory : 13.0%

UGE slot occupancy rate

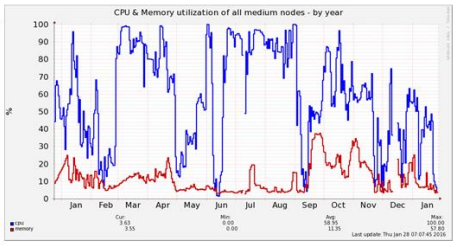


CPU : 43.8%
Memory : 84.1%

Medium node (Ph1): 2 nodes

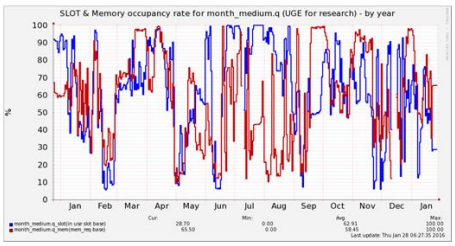
HP ProLiant DL980 G7
Intel Xeon E7-4870,80 cores/node
2 TB memory/node

CPU & Memory Utilization



CPU : 59.0%
Memory : 11.4%

UGE slot occupancy rate

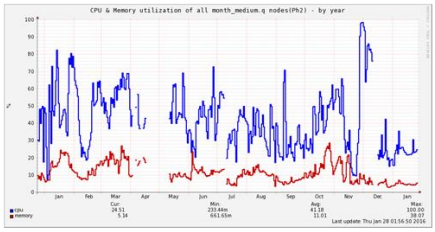


CPU : 62.9%
Memory : 58.5%

Medium node (Ph1): 8 nodes

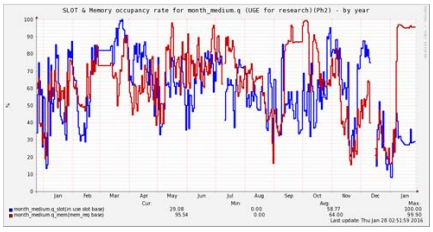
HP ProLiant DL980 G7
Intel Xeon E7-4870,80 cores/node
2 TB memory/node

CPU & Memory Utilization



CPU : 41.2%
Memory : 11.0%

UGE slot occupancy rate

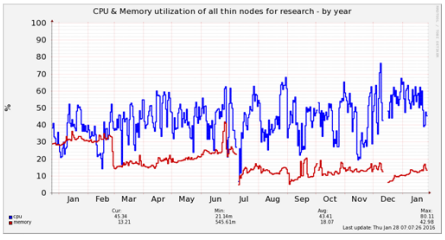


CPU : 58.8%
Memory : 64.0%

Thin node (Ph1, for researches; login users and web users): 168 nodes

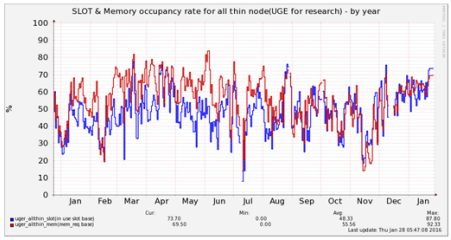
HP ProLiant SL230s,SL250s Gen8
Intel SandyBridge EP(Xeon E5-2670)
16 cores/node,64 GB memory/node

CPU & Memory Utilization



CPU : 43.4%
Memory : 18.1%

UGE slot occupancy rate

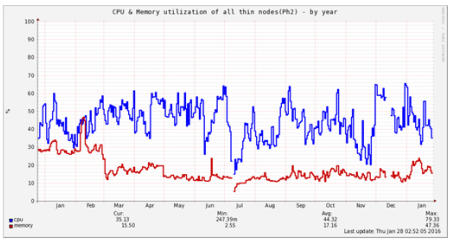


CPU : 48.3%
Memory : 55.6%

Thin node (Ph2, for researches; login users and web users): 193 nodes

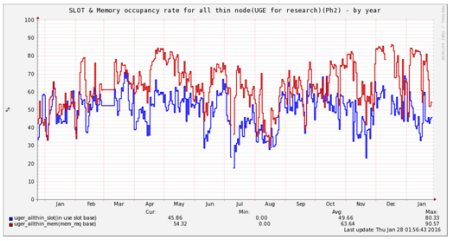
HP ProLiant SL230s,SL250s Gen8
Intel IVY Bridge EP(Xeon E5-2680v2)
20 cores/node,64 GB memory/node

CPU & Memory Utilization



CPU : 44.3%
Memory : 17.2%

UGE slot occupancy rate

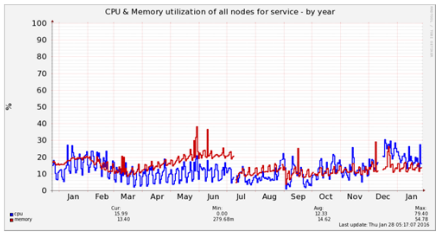


CPU : 49.7%
Memory : 63.6%

Thin node (Ph1, for services): 33 nodes

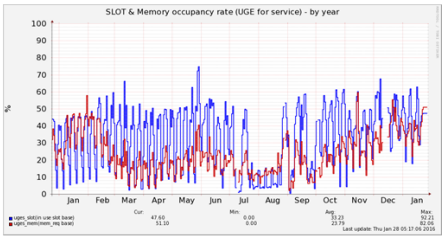
HP ProLiant SL230s,SL250s Gen8
Intel SandyBridge EP(Xeon E5-2670)
16 cores/node,64 GB memory/node

CPU & Memory Utilization



CPU : 12.3%
Memory : 14.6%

UGE slot occupancy rate



CPU : 33.2%
Memory : 23.8%

Figure 6-7: CPU and memory utilization rates (left) and Univa Grid Engine (UGE) slot and memory occupancy rate (right) (2015).

Blue lines are CPU utilization rates and red lines are memory utilization rates.
The UGE occupancy rate indicates the amount of user demand for the resources.
Thin nodes omitted from the figure are used for Web servers, software development, and testing purposes, as well as various administrative servers.

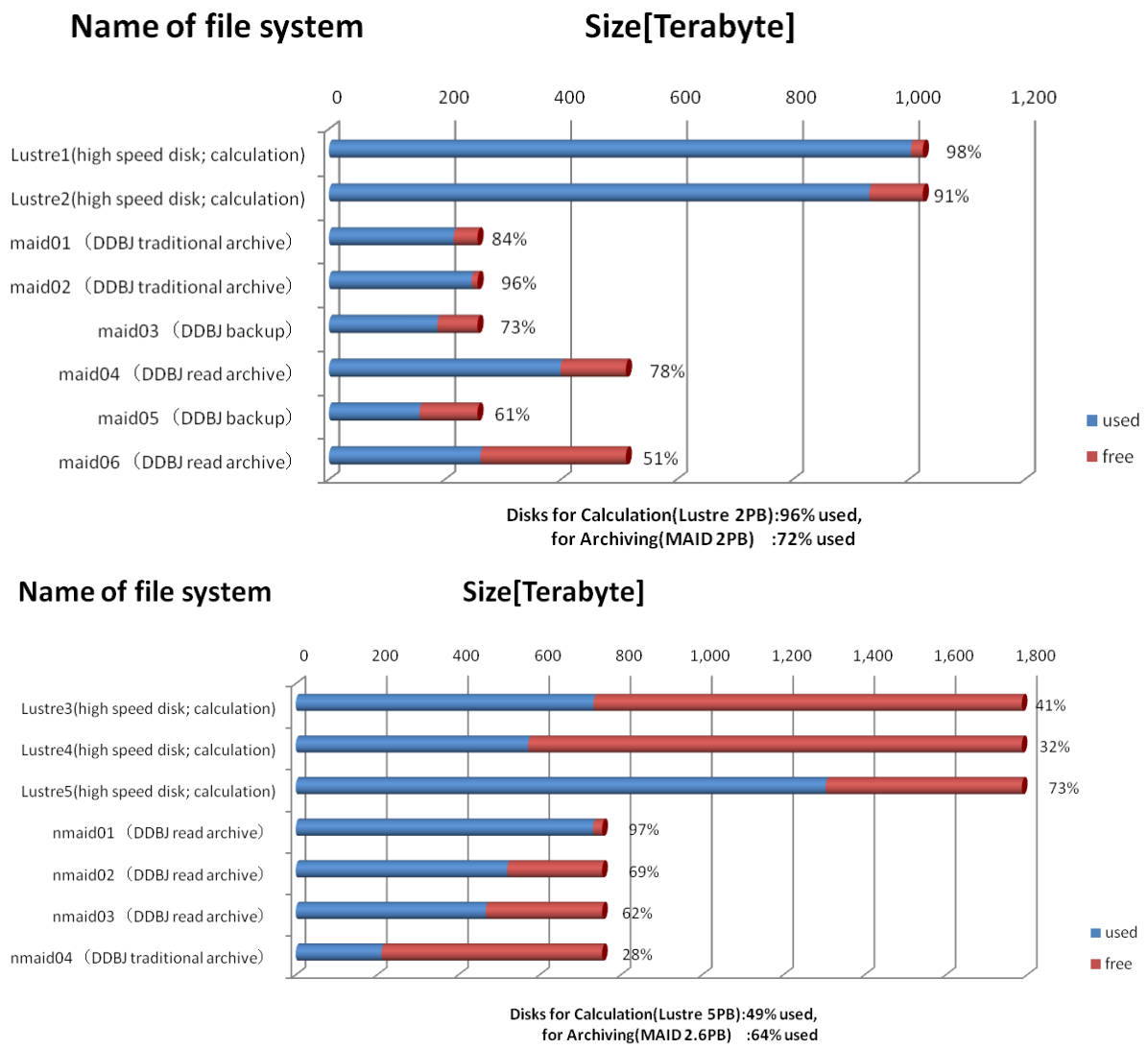
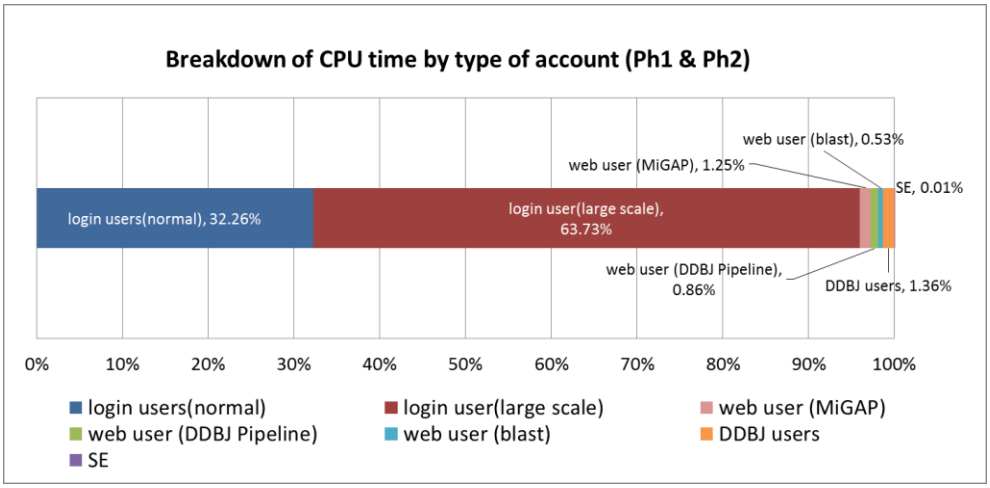
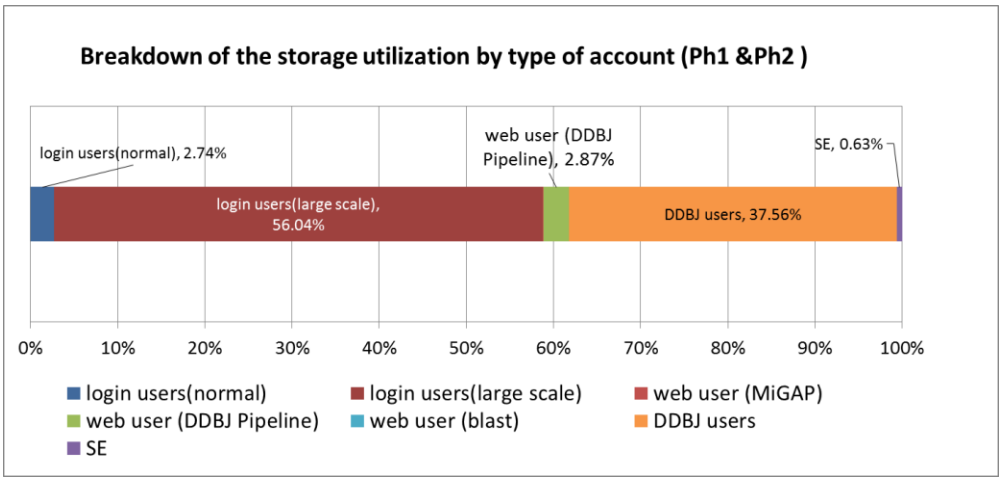
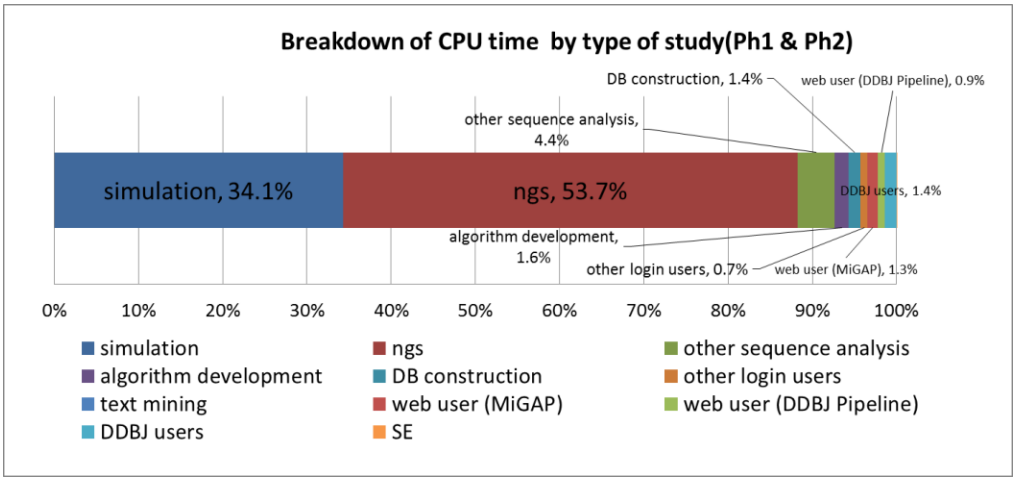


Figure 6-8: Storage utilization rate (January 2016).


Breakdown of CPU time by type of account (Ph1 & Ph2)
web user (blast), 0.53%
web user (MiGAP), 1.25%
SE, 0.01%
login users(normal), 32.26%
login user(large scale), 63.73%
web user (DDBJ Pipeline), 0.86%
DDBJ users, 1.36%
0% 10% 20% 30% 40% 50% 60% 70% 80% 90% 100%
login users(normal)
login user(large scale)
web user (MiGAP)
web user (DDBJ Pipeline)
web user (blast)
DDBJ users
SE


Breakdown of the storage utilization by type of account (Ph1 &Ph2)
login users(normal), 2.74%
web user (DDBJ Pipeline), 2.87%
SE, 0.63%
login users(large scale), 56.04%
DDBJ users, 37.56%
0% 10% 20% 30% 40% 50% 60% 70% 80% 90% 100%
login users(normal)
login users(large scale)
web user (MiGAP)
web user (DDBJ Pipeline)
web user (blast)
DDBJ users
SE


Breakdown of CPU time by type of study(Ph1 & Ph2)
DB construction, 1.4%
web user (DDBJ Pipeline), 0.9%
other sequence analysis, 4.4%
simulation, 34.1%
ngs, 53.7%
DDBJ users, 1.4%
algorithm development, 1.6%
other login users, 0.7%
web user (MiGAP), 1.3%
0% 10% 20% 30% 40% 50% 60% 70% 80% 90% 100%
simulation
ngs
other sequence analysis
algorithm development
DB construction
other login users
text mining
web user (MiGAP)
web user (DDBJ Pipeline)
DDBJ users
SE

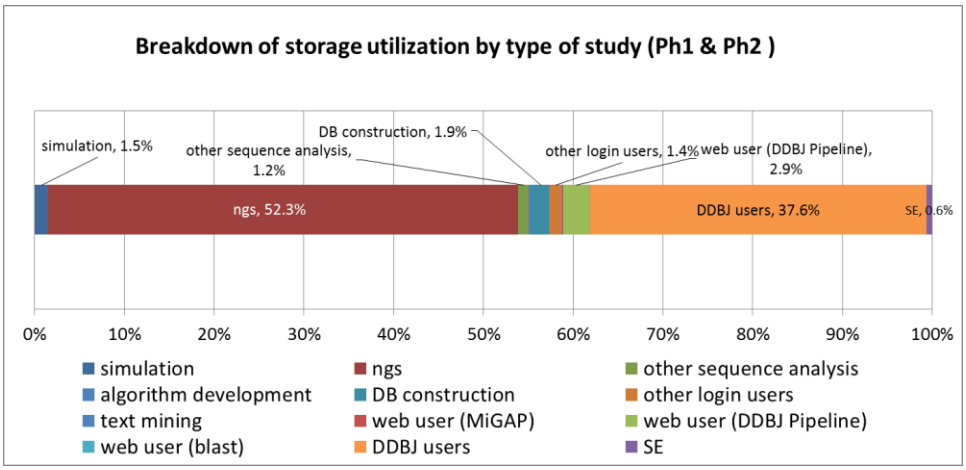


Figure 6-9 Breakdowns of CPU time and storage utilization (2015)

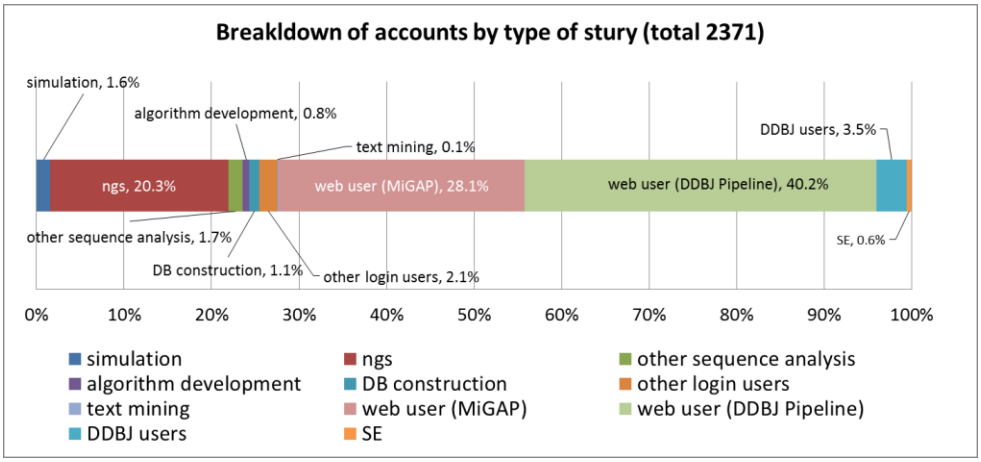


Figure 6-10: Breakdown of number of user accounts by study type (2015).

7. PUBLIC RELATIONS

7-1. Academic Presentation

Journal papers

(1) DNA data bank of Japan (DDBJ) progress report.
Mashima J, Kodama Y, Kosuge T, Fujisawa T, Katayama T, Nagasaki H, Okuda Y, Kaminuma E, Ogasawara O, Okubo K, Nakamura Y, Takagi T
Nucleic acids research 44(D1): D51-7, 2016. doi: 10.1093/nar/gkv1105. 2016 Jan

(2) The International Nucleotide Sequence Database Collaboration.
Cochrane G, Karsch-Mizrachi I, Takagi T, International Nucleotide Sequence Database Collaboration.
Nucleic acids research 44(D1): D48-50, 2016. doi: 10.1093/nar/gkv1323. 2016 Jan

(3) The DDBJ Japanese Genotype-phenotype Archive for genetic and phenotypic human data.
Kodama Y, Mashima J, Kosuge T, Katayama T, Fujisawa T, Kaminuma E, Ogasawara O, Okubo K, Takagi T, and Nakamura Y
Nucleic acids research 43(D1): D18-22, 2015. doi: 10.1093/nar/gku1120. 2015 Jan

Book, Article, etc. (in Japanese)

(1) **Introduction of bioinformatics**
Nakamura Y
Japanese Society of Bioinformatics (A part of chap.1)
Keio University Press Aug 2015.

(2) **How to use the NIG supercomputer and command-line based NGS data analysis**
Kaminuma E, Mochizuki T, Monden Y, Kobayashi T, Ohyanagi H, Kentaro Y,
56th J.Soc of Breeding Meeting (Symposium/Workshop Report),
Breeding Research Vol.17 No.2, July 2015.

Conference/Workshop/Seminar (in Japanese)

(1) **The RIKEN course "CAGE analysis in the post-FANTOM5 era"**. Dec 16, 2015
Submission of next generation sequencing data
Kodama Y

(2) **Togo-day Symposium 2015, Oct 5, 2015.**
DDBJ,
Nakamura Y

(3) **SIG Technical Reports, The 3rd IOT workshop**, Sep.27, 2015.
A Scientific Paper Reproducible Environment with Overlay Cloud Architecture,
Yokoyama S, Masatani Y, Ogasawara O, Ohta T, Yoshioka N, Aida K.

(4) **13th JCK Bioinformatics Symposium**, Aug 11, 2015.
Towards better genome annotation
Nakamura Y

(5) **Tokyo Nodai Genomic Center 2015 seminar, July 27, 2015.**
NGS sequence analysis and Introduction of public databases
Nakamura Y

(6) **NGS Genbano-Kai: 4th meeting** July 1, 2015.
DDBJ Pipeline Q&A and trend of analytical tool usage,
Saka K, Tanizawa Y, Nagasaki H, Mochizuki T, Sakamoto N, Kaminuma E,Nakamura Y

(7) **NGS Genbano-Kai: 4th meeting** July 1, 2015.
Enhancement of Data Analysis Infrastructure with the cooperation of the NIG supercomputer and inter-cloud system,
Ogasawara O, Ohta T, Shigetoshi Yokoyama, Yoshinobu Masatani,
Nobukazu Yoshioka, Kento Aida.

(8) **SSH Hikawa High-School Scientific Seminar,** June 18, 2015.
Introduction to Bioinformatics
Nakamura Y

(9) **The Japanese Association for Cancer and Hypoxia Research,** June 5, 2015.
DDBJ and Large-scale Sequence Analysis using NIG Super Computer
Nakamura Y.

(10) **IEICE Service Computing Workshop: Applications of Docker Containers and Cloud Infrastructures to NGS Analyses**, June 5, 2015.
Constructions,Operations and Managements of Web services and Cloud Systems
Osamu Ogasawara and Tazuro Ohta

(11) **HPCS2015**, May 19, 2015.
Construction of A Bioinformatics Pipeline Reproducibility Environment with Overlay Cloud Architecture,
Shigetoshi Yokoyama, Yoshinobu Masatani, Osamu Ogasawara, Tazro Ohta,Nobukazu Yoshioka, Kento Aida.

(12) **Science Information Infrastructure Open Forum 2015**, Feb 3, 2015.
On the Computational Infrastructure for Biological Studies with a focus on Nucleotide Databases
Osamu Ogasawara, Tazuro Ohta
Science Information Center (the host organization: National Institute of Information)

7-2. Training courses

[Kohira, Suzuki, Yokoyama, Fukuda, Kodama, Lee, Mashima, Kaminuma, Ogasawara, Nakamura]

33rd DDBJing Training Course (in Japanese)

Date: 11 November 2015
Venue: JST Science Plaza, Tokyo
Number of Trainees: 18
Theme: New Utilization Method for a Supercomputer in NGS Data Analysis and Data Submission to DDBJ
Lectures:

- **Introduction of DDBJ, Mass Sequence Data Analysis Using DDBJ Services and the NIG Supercomputer**
  Yasukazu Nakamura (Professor, DDBJ/NIG)
- **Introduction of RAD-Seq for SNP Analysis**
  Kenta Shirasawa (Research Staff, Laboratory of Plant Genomics and Genetics, Kazusa DNA Research Institute)
- **Introduction of NGS Genotyping for SNP Analysis**
  Hiromi Kanegae (Post Doc, Laboratory of Biometry and Bioinformatics, Graduate School of Agriculture and Life Sciences, University of Tokyo)
- **Introduction of DDBJ Sequence Read Archive (DRA), DDBJ BioProject, and DDBJ BioSample**
  Asami Fukuda (Senior Annotator, DDBJ)
- **Introduction of "Japanese Genotype-phenotype Archive"**
  Yuichi Kodama (Senior Annotator, DDBJ)
- **Introduction of and Training on DDBJ Read Annotation Pipeline**
  Takako Mochizuki (Genome Informatics Laboratory, NIG)

- **Introduction of Mass Submission System for a Large-scale Assembled Sequence Data Submission**
  LEE Kyungbum (Senior Annotator, DDBJ)

32nd DDBJing Training Course (in Japanese)

*This training course was held at the request of Naha Plant Protection Station.
Date: 29 July 2015
Venue: Naha Plant Protection Station, Okinawa
Trainees: only the staff of Naha Plant Protection Station
Lectures:

- **Use of Sequencing Data for Organism Identification**
  Hiroshi Mori (Assistant Professor, Graduate School of Bioscience and Biotechnology Biological Information, Tokyo Institute of Technology)
- **Primer Design**
  Asao Fujiyama (Professor, Comparative Genomics Laboratory, NIG)
- **Database Search (Keyword Search, Homology Search, Multiple Alignment)**
  Yasukazu Nakamura (Professor, DDBJ/NIG)

1st All-in-One NBDC/DBCLS/PDBj/DDBJ Joint Training Course (in Japanese)

Date: 18 July 2015
Venue: Osaka University Nakanoshima-Center, Osaka
Number of Trainees: about 22
Theme: How to Use Life-Science Databases and Tools of DDBJ, DBCLS, NBDC, and PDBj
Lectures:

- **Introduction of DDBJ Databases and Search/Analysis Tools**
  Yasukazu Nakamura (Professor, DDBJ/NIG)
- **Introduction of Data Submission to DDBJ**
  Jun Mashima (Chief Annotator, DDBJ)
- **Introduction of NIG Supercomputer**
  Osamu Ogasawara (Assistant Professor, DDBJ/NIG)

31st DDBJing Training Course (in Japanese)

Date: 12 June 2015
Venue: JST Science Plaza, Tokyo
Number of Trainees: 20
Theme: New Utilization Method for a Supercomputer in NGS Data Analysis and Data Submission to DDBJ
Lectures:

- **Introduction of DDBJ, Mass Sequence Data Analysis Using DDBJ Services and the NIG Supercomputer**
  Yasukazu Nakamura (Professor, DDBJ/NIG)
- **Construction of a Reproducible Bioinformatics Analysis Environment using BioDevOps**
  Itoshi Nikaido (Unit Leader, Bioinformatics Research Unit, RIKEN Advanced Center for Computing and Communication)
- **Reproducibility as a Service: Virtual Appliance for NGS Data Analysis**
  Tazro Ohta (Project Researcher, Database Center for Life Science (DBCLS))
- **Introduction of DDBJ Sequence Read Archive (DRA), DDBJ BioProject, and DDBJ BioSample**
  Yuichi Kodama (Genome Informatics Laboratory, NIG)

- **Introduction of Mass Submission System for a Large-scale Assembled Sequence Data Submission**
  LEE Kyungbum (Senior Annotator, DDBJ)
- **Introduction of Japanese Genotype-phenotype Archive**
  Yuichi Kodama (Senior Annotator, DDBJ)

7-3. Visitor tour

[Kohira, Yokoyama, Kawagoe, Ishikawa, Miyazaki, Kaminuma, Ogasawara, Arita, Nakamura]
DDBJ accepts visit requests from the outside. The following people came to visit DDBJ in 2015.

1	2016-01-27	Nihon University Mishima High School, Shizuoka (students)	27
2	2015-11-12	Mishima Yamada Junior High School, Shizuoka	5
3	2015-10-07	JSPS*** members	2
4	2015-08-20	Nirayama High School, Shizuoka and Shizuoka-Futaba Junior High School	12
5	2015-08-07	Haibara High School, Shizuoka	35
6	2015-07-28	Koryo High School in Kitamori-city, Yamanashi Prefecture	26
7	2015-07-13	NIGINTERN program students	40
8	2015-07-10	SOKENDAI- CAIS**	5
9	2015-06-10	Science teachers of high schools in the eastern part of Shizuoka	40
10	2015-06-04	MEXT*, Research Promotion Bureau, Life Science Division staff members, NBDC	6
11	2015-05-08	Asahi Kasei Pharma Corp. (laboratory researchers)	3
12	2015-01-28	Nihon University Mishima High School, Shizuoka	26

* Ministry of Education, Culture, Sports, Science and Technology
** The Center for Academic Information Services, SOKENDAI
*** Japan Society for the Promotion of Science

7-4. News releases on the Web, Mail Magazine, and inquiries from users

7-5. Public relations at the academic meetings

[Mashima, Kodama, Lee, Fukuda, Okido, Aono, Suzuki, Kohira]
[The 38th Annual Meeting of the Molecular Biology Society of Japan]
Period: 1-4 December 2015
Venue: Kobe Port Island
Activities: Demonstration of DDBJ services, distribution of brochures, personal assistance for submissions to DDBJ, at the exhibition booth

[NGS Field 4th Meeting]
Period: 2&3 July 2015
Tsukuba International Congress Center
Activities: Demonstration of DDBJ services, distribution of brochures, personal assistance for submissions to DDBJ, at the exhibition booth

[The 2015 Annual Conference of the Japan Society for Bioscience, Biotechnology and Agrochemistry] **in** Okayama
Period: 27-29 March 2015
Venue: Okayama University Tsushima Campus
Activities: Demonstration of DDBJ services, distribution of brochures, personal assistance for DDBJ service use at the exhibition booth
(Kohira, Aono, Okido, Team Joho)

8. COOPERATIVE RELATIONS

8-1. Cooperation with JPO/KIPO (KOBIC): Patent sequence

[Aono, Ogasawara, Okubo, Nakamura, Takagi]

DDBJ cooperates with Japan Patent Office (JPO) and Korean Intellectual Property Office (KIPO)/Korean Bioinformation Center (KOBIC) on the distribution of patent application sequences to INSD. DDBJ also makes available patent office data for United States Patent Office (USPTO) and European Patent Office (EPO) submitted to GenBank and ENA.

8-1-1 JPO

JPO data submission started in 1993. A private line for Internet connection was constructed between JPO and DDBJ that enables JPO submission data and INSDC data to be safely shared. The network servers were updated in 2015.

8-1-1-1: Data submission

DDBJ received JPO data in the form of monthly data (12 times) and additional data (3 times) in 2015.

(Number of distribution entries in 2015)

Nucleotide sequence data: 739,698 entries

Amino acid sequence data: 233,273 entries

8-1-1-2: ST.26

DDBJ has cooperated with the intellectual property office (IPO) task force team for JPO, EPO, and USPTO on the construction of the new sequence listing guideline ST.26 since 2012.

8-1-2 KIPO

KIPO data submission began in March 2008. DDBJ visited KOBIC, Korea Institute of Patent Information (KIPI), and KIPO to discuss the submission problem in Daejeon, Korea 3-6 December 2013 (Aono, Lee, and Nakamura) and 19-22 January 2015 (Aono and Lee).

8-1-2-1: Data submission

DDBJ has received KIPO data twice in 2015.

(Number of distribution entries in 2015)

Nucleotide sequence data: 155,806 entries

Amino acid sequence data: 95,881 entries

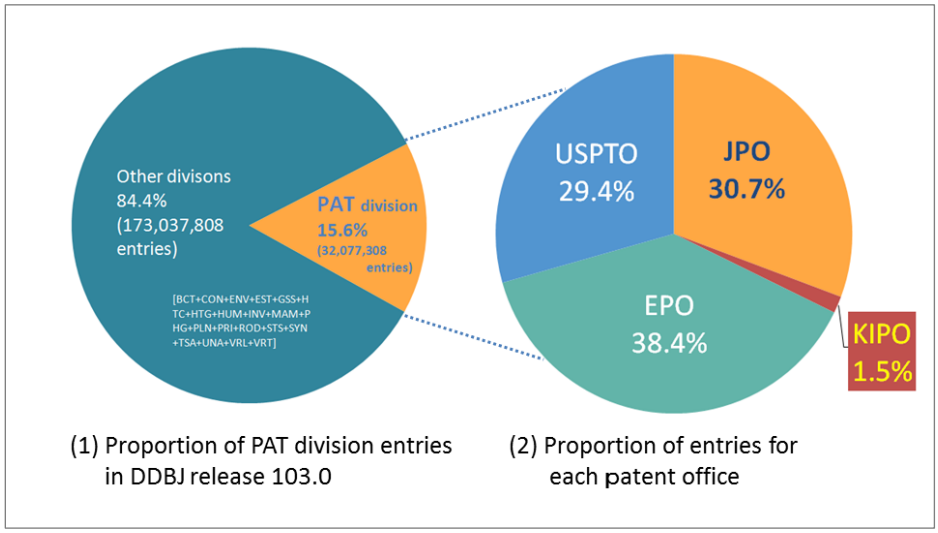
8-1-2-2: Changes to KIPO submission in 2015

*1: KIPO submission has started to include RNA data.

*2: The new KOBIC submission format was constructed.

*3: KIPO Flat File (FF) has added new information lines for CC (comment information), FH (fixed character string), and FT (feature information).

Figure 8-1: Statistics for patent data in DDBJ release 103.0



8-2. Collaboration with DBCLS for data integration of DDBJ resources

[Fujisawa, Kaminuma, Nakamura, DBCLS: Katayama]

To improve reusability of the sequence annotation data, we have developed the Resource Description Framework (RDF) version of DDBJ records in collaboration with DBCLS. To extend genomic context information, we developed the assembly records RDF using NCBI resources. Assembly records RDF has information about the structure of assembled genomes as a collection of WGS, CON entries, or completely sequenced chromosomes. In 2016, we will release the resulting RDF datasets.

8-3. Collaboration with RIKEN for describing RDF for metadata submission in DRA, BioProject, and BioSample

DDBJ and RIKEN have started a collaborative study about how to carry out DRA large submission and its application to Semantic Web technologies. We are working on the development of an RDF version of the the DRA (DDBJ Sequence Read Archive) submission metadata, including BioSample and BioProject.

8-4. Collaboration with NBDC for sharing data from human subject research

[Kodama, Mashima, NBDC: Minowa, Kawashima, Mitsuhashi, Miyazaki, Takagi (DDBJ and NBDC)]

The JGA service is provided in collaboration with the National Bioscience Database Center (NBDC) of the Japan Science and Technology Agency. Data storage, management, and distribution by JGA are governed by the NBDC policies and procedures for sharing human subject data.

The Data Access Committee (DAC) at NBDC reviews applications to submit data to JGA. Descriptions of these policies and guidelines can be found on the NBDC human database website at http://humandbs.biosciencedbc.jp. The English version of the NBDC website was prepared in March 2015 for overseas researchers. The DAC also reviews requests from researchers to use JGA datasets for research. The DAC ensures that the stated research purposes are compatible with participant consent and that the Principal Investigator and institution will abide by the NBDC guidelines and the specific terms and conditions imposed on a given dataset.

In 2015, two JGA annotators (Kodama, Mashima) joined the Data Access Committee and Data Sharing Subcommittee meetings held at NBDC.

8-5. Collaboration with ToMMo for maintaining backup copies of each other's genomic data

[Ogasawara, Watanabe, Okuda, ToMMo: Nagasaki]

DDBJ and Tohoku Medical Megabank Organization (ToMMo) have large amounts of genomic data which they have accumulated independently. To limit the damage caused when a disaster strikes, it is desirable to store backup copies of these data at remote locations. DDBJ and ToMMo have started a collaborative study about how to store each other's backup data safely and securely.

9. NIG AND ROIS RELATIONS

9-1. Cooperation with NIG IT support team

[Okuda, Ogasawara, Arita, Takagi] [Nagura, Nagira, Arita from NIG IT Unit team]

DDBJ shares information with NIG IT Unit team. Members of the NIG IT Unit team participate in a DDBJ meeting once a month and report their activities to DDBJ members. All system engineers of DDBJ and the NIG IT Unit team work in a common room, which contributes to their interacting with each other.

9-2. Cooperation with Intellectual Property Unit in NIG

[Aono, Nakamura, Takagi]

Aono cooperated with the Intellectual Property Unit to make a brochure about patent office data submission in DDBJ to distribute at BioJapan 2015 World Business Forum, and attended to introduce it at the NIG booth on 15 October 2015.

9-3. Cooperation with the NII SINET team and ISM supercomputer team

[Ogasawara, Okuda, Takagi] [Ohta from DBCLS, Aida from NII SINET team, Nakano from ISM supercomputer team, Okada from INPR]

To address intrinsically multi-disciplinary issues involving NGS data analysis, network and security engineering, and supercomputer management of multiple sites, DDBJ started transdisciplinary research with Database Center for Life Science (DBCLS), the SINET team of National Institute of Information (NII), the supercomputer center of the Institute of Statistical Mathematics (ISM), and National Institute of Polar Research (INPR) in 2014.

We are developing a framework that allows users to run programs on a light-weight virtual environment system such as Docker on several computing infrastructures, including the NIG and ISM supercomputer systems, AWS, and cluster servers in general. This framework aims to enhance research reproducibility. Since a Docker container can hold all the computational environments required to execute an analysis pipeline, the containers allow users to share the pipelines and execute them on various kinds of computer environments quite easily.

10. BUDGET PLAN OF FISCAL YEAR 2016

10-1. DDBJ budget in FY 2015

Item	Detailed description	Price (thousands of yen)
Rental expenses	NIG supercomputer	687,475
Subcontract expenses	DDBJ operation, Software development	187,884
Personnel expenses	Annotators	88,390
Hardware / repair	Relocation of servers for power saving, Repair of air-conditioners, etc.	3,307
Office supplies / consumable	PC, OA related	1,171
Travel expenses	Collaborative meeting, Publicity	3,176
Others	Postage, License fees, Exhibition fees, etc.	3,941
Total amount		975,344

10-2. DDBJ main efforts in FY 2016

DDBJ plans to focus on the following targets in FY 2016.

1. **Computing Platform for Personal Genomics**
   Investigating secure platforms on the NIG supercomputer for analyzing personal genomic information.
2. **Analysis-to-submission High-throughput Pipeline**
   Connecting data annotation pipelines and DDBJ submission tools to reduce submitter workload and incidentally save annotator costs.
3. **Next Supercomputer Design**
   Designing hardware systems of the next NIG supercomputer while considering not only supporting large-volume storage but also a high-performance environment for big data analysis.
4. **Closer Collaborative Relationship with NBDC/DBCLS/PDBj**
   Promoting collaborative research and educational seminars with NBDC, DBCLS, and PDBj toward establishing a unified service portal in the future.

10-3. Procurement plan of the next NIG supercomputer system (March 2017)

NIG plans to deploy a new supercomputer system in 2017 that replaces the current system. The next supercomputer system will include more than 25 PB of storage for DDBJ data archives and nearly the same size of calculation storage. NIG also plans to allow supercomputer users to analyze human personal genome data on the next supercomputer system.

Appendix:

Contributors of DDBJ databases from latest Release note (Rel.103, as of Nov. 27, 2015)

Jun Mashima, Hideo Aono, Yuji Ashizawa, Yukino Dobashi, Mayumi Ejima, Masahiro Fujimoto, Asami Fukuda, Tomohiro Hirai, Naofumi Ishikawa, Chiharu Kawagoe, Yuichi Kodama, Junko Kohira, Takehide Kosuge, Kyungbum Lee, Mika Maki, Hisako Mashima, Fujitaka Matsumori,
Kimiko Mimura, Hiroshi Miyazaki, Naoko Murakata, Satoshi Muraoka, Toshihisa Okido, Yoshihiro Okuda, Katsunaga Sakai, Yukie Sakon, Makoto Sato, Aimi Shiida, Rie Sugita, Kimiko Suzuki, Haru Tsutsui, Koji Watanabe, Tomohiko Yasuda, Emi Yokoyama,
Masanori Arita, Eli Kaminuma, Osamu Ogasawara, Kosaku Okubo, Toshihisa Takagi, and Yasukazu Nakamura