

DDBJ・遺伝研スパコン入門

国立遺伝学研究所／総合研究大学院大学
大量遺伝情報研究室 教授

中村 保一

Aug 24, 2017
石川県立大学

遺伝研は静岡県三島市にあります



1996 *Synechocystis* sp. PCC 6803 (光合成細菌世界初)

2000 *Arabidopsis thaliana* (高等植物世界初)

2000 *Mesorhizobium loti* (窒素固定細菌世界初)

2001 *Anabaena (Nostoc)* sp. PCC 7120

2002 *Bradyrhizobium japonicum*

2002 *Thermosynechococcus elongatus* BP-1

2003 *Gloeobacter violaceus* PCC 7421

2007 *Microcystis aeruginosa* NIES-843

2008 *Lotus japonicus*

2012 *Bradyrhizobium* sp. S23321

2012 *Solanum lycopersicum* (トマト)

2012 *Eucalyptus globulus* (ユーカリ)

2012 *Hevea brasiliensis* draft (ゴムノキ)

2014 *Klebsormidium flaccidum* (藻類)

2014 *Weissella oryzae* SG25T, *Lactobacillus oryzae* SG293T,
Lactobacillus hokkaidonensis LOOC260T

2016 *Citrus sinensis* Unshu (ウンシュウミカン)

2017 *Marchantia polymorpha*

Cyanobacteria
Rhizobia
Plants
LAB



弊ラボで進行中のゲノムプロジェクト



ゼニゴケ *Marchantia polymorpha*

ゲノムサイズ 220 Mb

4400 スキャフールド / N50: 1.3 Mb

ゴムノキ *Hevea brasiliensis*

ゲノムサイズ 1.4 Gb

5万2千 スキャフールド / N50: 120 kb



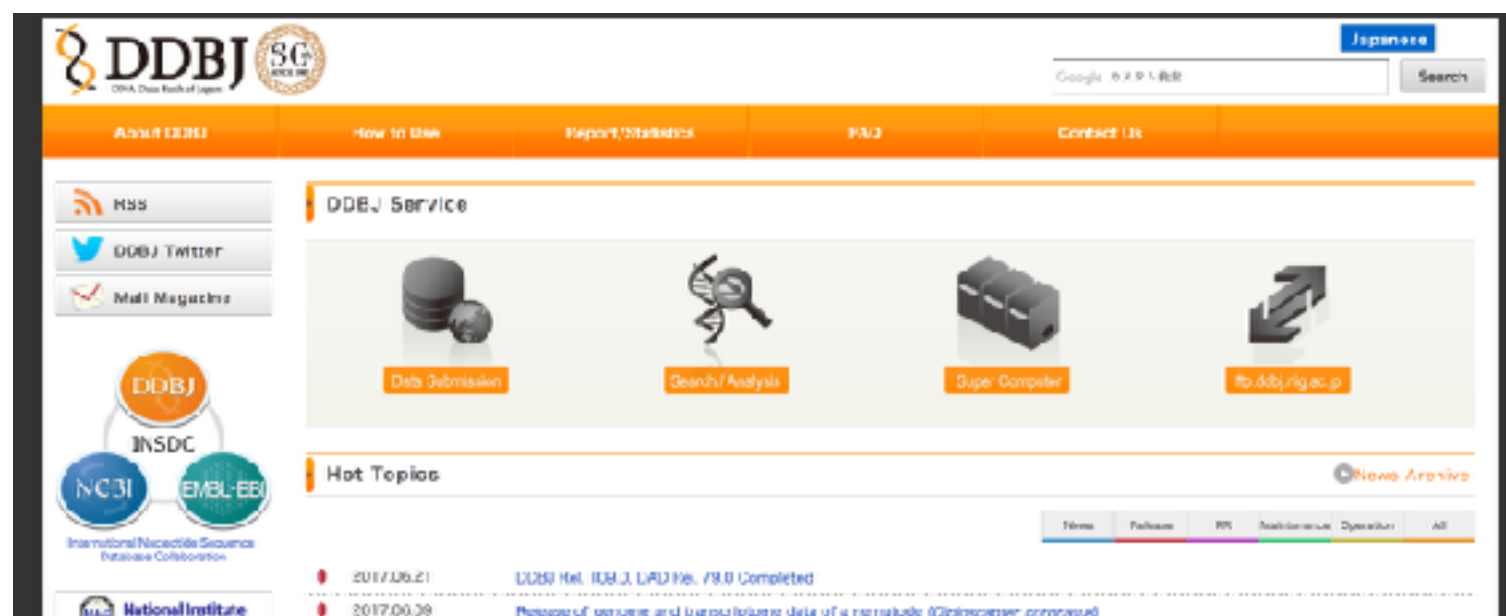
ウンシュウミカン *Citrus unshiu*

ゲノムサイズ 360 Mb

2万1千 スキャフールド / N50: 385 kb

DDBJ

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



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



 **DDBJ**  DNA Data Bank of Japan

[Japanese](#)

[About DDBJ](#) [How to Use](#) [Report/Statistics](#) [FAQ](#)

 [RSS](#)

 [DDBJ Twitter](#)

 [Mail Magazine](#)


International Nucleotide Sequence Database Collaboration

 **National Institute of Genetics**
Research Organization of Information and Systems

 大学共同利用機関法人
情報・システム研究機構
Research Organization of Information and Systems

 **DBCLS**
Database Center for Life Science

 **NBDC**
National Bioscience Database Center

DDBJ Service

 [Data Submission](#)

 [Search / Analysis](#)

 [Super Computer](#)

 <ftp.ddbj.nig.ac.jp>

Hot Topics

[News Archive](#)

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

INSDC

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



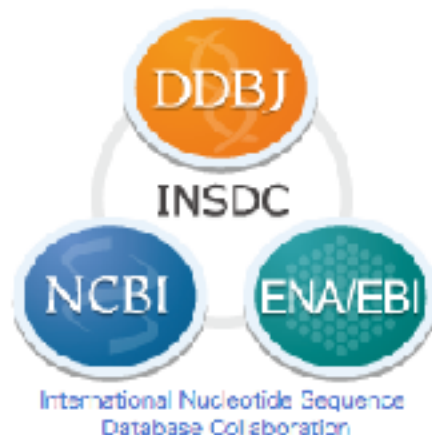
The screenshot shows the INSDC website. The header features the INSDC logo and the text "International Nucleotide Sequence Database Collaboration". Below the header is a navigation menu with links: ABOUT INSDC, POLICY, ADVISORS, and DOCUMENTS. The main content area is titled "International Nucleotide Sequence Database Collaboration" and contains a bullet point describing the INSDC initiative. To the left of the main content are logos for NCBI, ENA, and DDBJ. Below the text is a table summarizing the data types and their corresponding databases.

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

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

[DDBJ の紹介](#)[利用の手引き](#)[レポート・統計](#)[FAQ](#)

第31回
DDBJing 講習会 in 東京
6月12日 × JST 東京本部別館

 [RSSを購読する](#) [DDBJ Twitter](#) [Mail Magazine](#)

DDBJ Service



登録
Data Submission



検索・解析
Search / Analysis



スパコン
Super Computer



アーカイブ
ftp.ddbj.nig.ac.jp

Hot Topics

[ニュースアーカイブ](#)[お知らせ](#)[データ公開](#)[広報](#)[メンテナンス](#)[運用情報](#)[一覧](#)

- 2015.06.10 アノクビダイコン (*Raphanus sativus* cv. Aokubi) ゲノムと転写産物配列データの公開
- 2015.06.05 国立遺伝学研究所 特任研究員 (DDBJ アノテータ) 募集
- 2015.06.03 UniProt 2015_06 公開
- 2015.04.20 「第31回 DDBJing 講習会 in 東京」開催のお知らせ

 大学共同利用機関法人 情報・システム研究機構
国立遺伝学研究所 大学共同利用機関法人
情報・システム研究機構
Research Organization of Information and Systems

DBCLS
Database Center
for Life Science

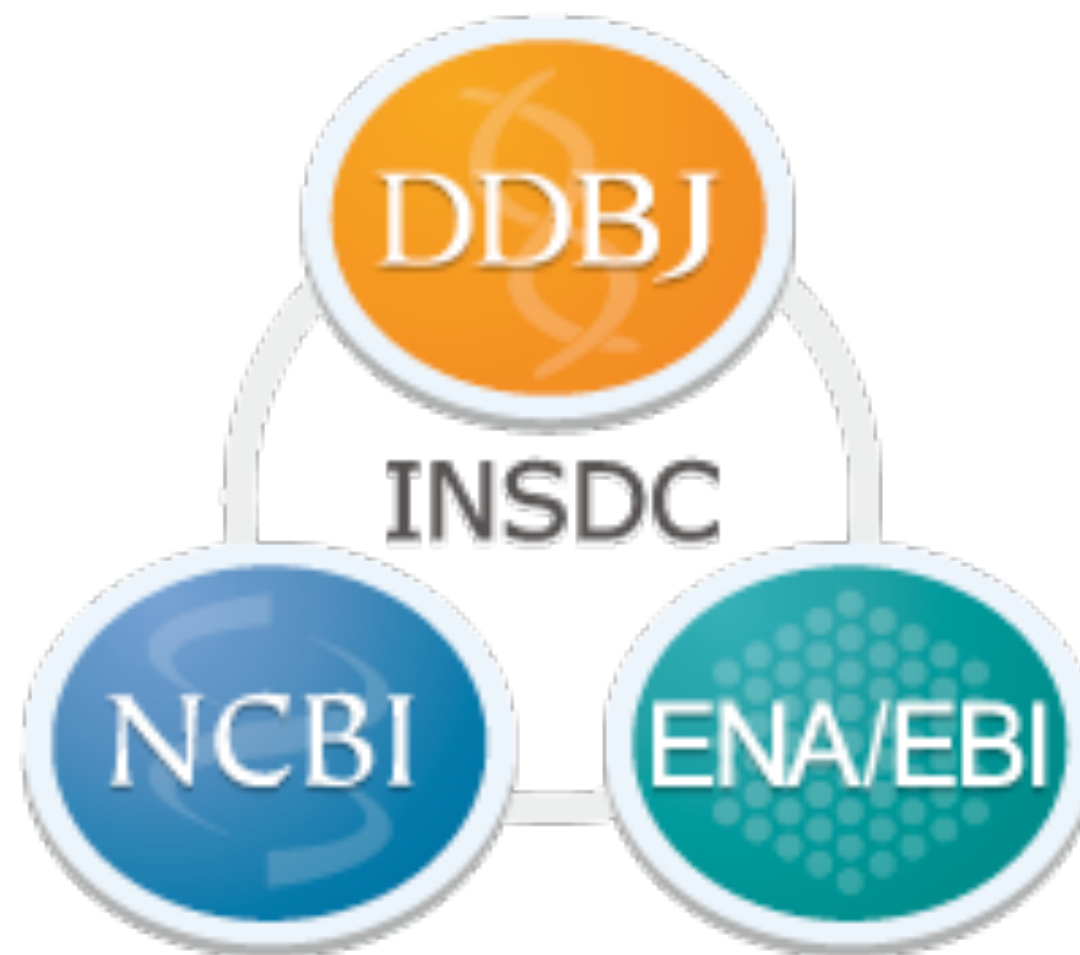
塩基配列データベースとはこのような事業

- 全世界で解読された塩基配列情報を
 - 査定して受入れ
 - データベースに蓄積し
 - 公開して共有する



国際塩基配列データベース (INSDC) の一員

- 米国: GenBank (NCBI)
- 欧州: ENA (EBI)
- 日本: DDBJ



DDBJ (from Release note 92) 44

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

ENA (from Release note 115) 27

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

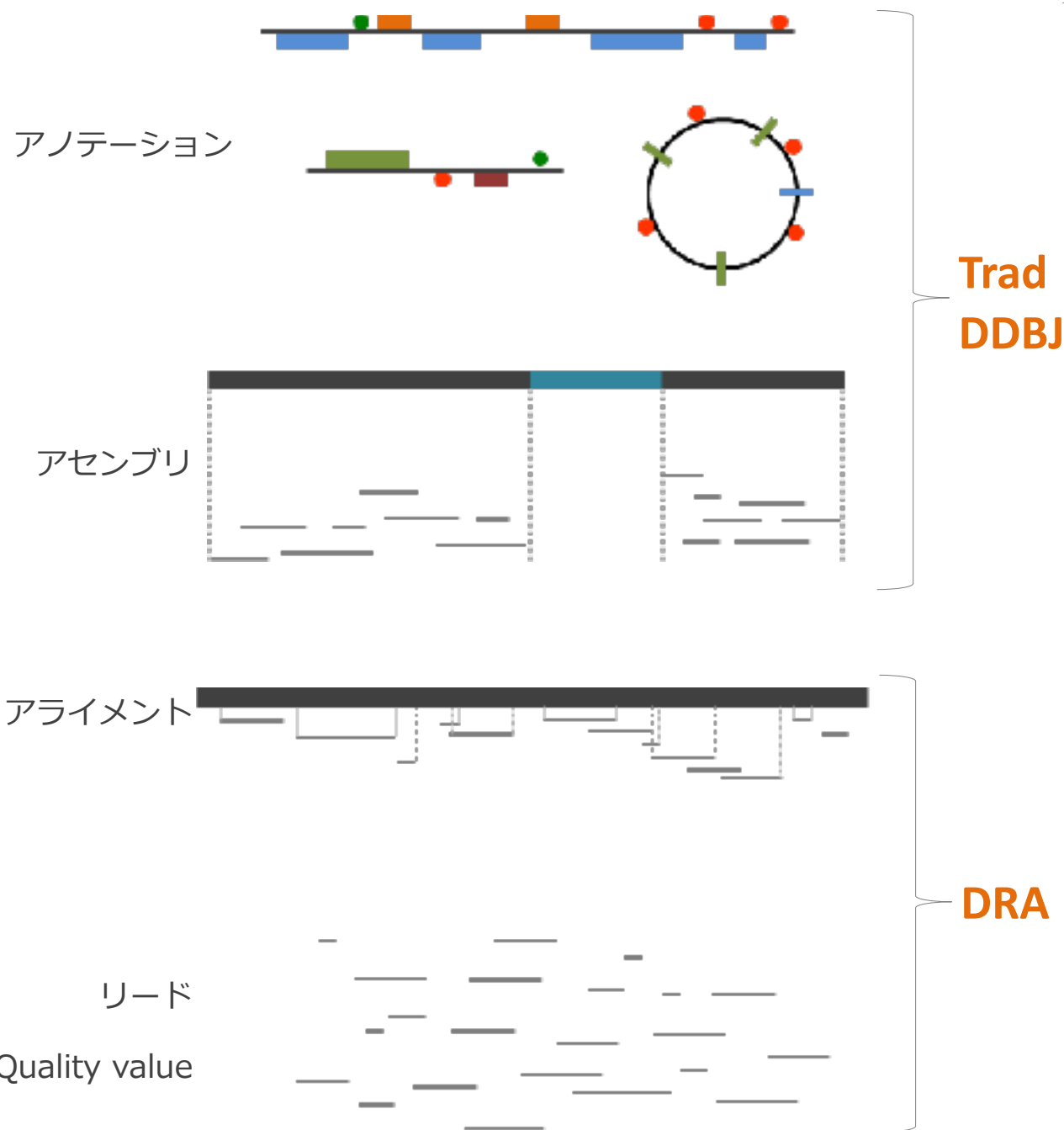
GenBank (from Release note 195) 68

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



DDBJが運営しているデータベース

DDBJ センター



アクセス制限

JGA

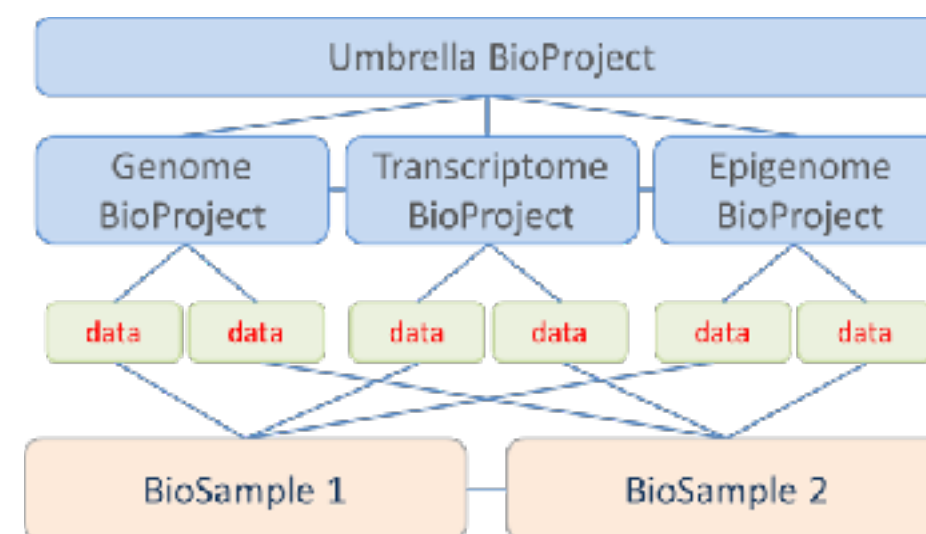
個人の遺伝型と表現型



ヒトデータ審査委員会



BioProject BioSample



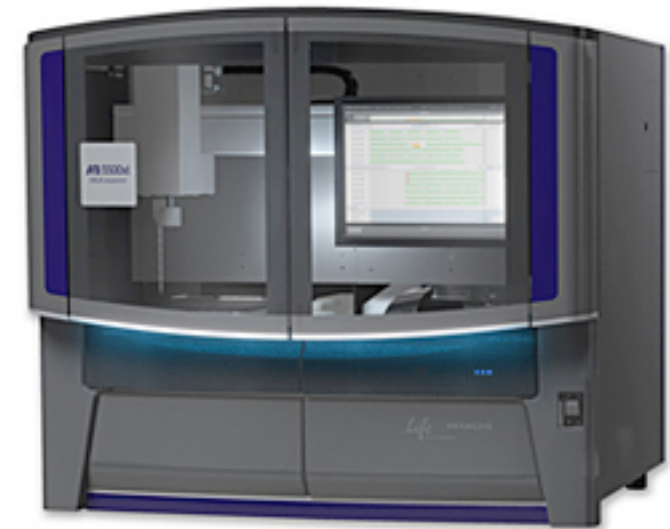
INSDC: オープンアクセス

DRA: DDBJ Sequence Read Archive

JGA: Japanese Genotype-phenotype Archive

生物学の 情報爆発

代表的 NGS 機材 (2010年頃のスライド)



(左) Roche (454): GS FLX+ System

(中) illumina: Genome Analyzer IIx System

(右) Life Technologies: 5500 xl SOLiD System

1000 Genomes

A Deep Catalog of Human Genetic Variation



[Home](#) [About](#) [Data](#) [Analysis](#) [Participants](#) [Contact](#) [Browser](#) [Wiki](#) [FTP search](#)

LATEST ANNOUNCEMENTS

THURSDAY NOVEMBER 06, 2014

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

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

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

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

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

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

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

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

Please send any questions about this data set to info@1000genomes.org

Recent project announcements

THURSDAY NOVEMBER 20, 2014

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

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

NAVIGATION

[Frequently Asked Questions](#)

LINKS



[All Project Announcements](#)



[Sample and Project Information](#)



[Media Archive](#)



[Download the 1000 Genomes Pilot Paper](#)



[Project Contacts](#)



[RSS Feed](#)

1001 Genomes

A Catalog of *Arabidopsis thaliana* Genetic Variation


[Home](#)
[Data Providers](#)
[Accessions](#)
[Tools](#)
[Software](#)
[Data Center](#)
[About](#)
[Help desk](#)

Welcome to the 1001 Genomes Project

Check

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

[Go »](#)

Browse

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

[Go »](#)

Download

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

[Go »](#)

Get Seeds

Seed sets of natural accessions are available for

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

Links

- [> GBrowse](#)
- [> WGS of 80 strains](#)
- [> Assemblies project](#)
- [> POLYMORPH](#)
- [> NCBI SRA Genomes Project](#)
- [> Map resource for 1001 Genomes](#)

News

February 23, 2014

Chi-2 and Seattle-0 assemblies generated by SLU are now available in the [Data Center](#).

August 1, 2013

SHOREmap v2.1 released.

August 6, 2012

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

November 29, 2011

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

November 15, 2011

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

August 28, 2011

Cap et al. Whole-genome

The 1001 Genomes Vision

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

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

今時（いまどき）のシーケンサー



(左) PacBio RSII System

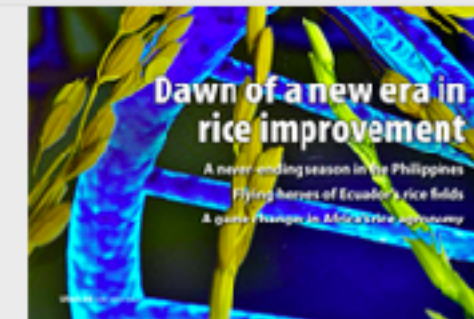
(中) illumina: HiSeq 4000 System

(右) Oxford NANOPORE MinION/SmidgION

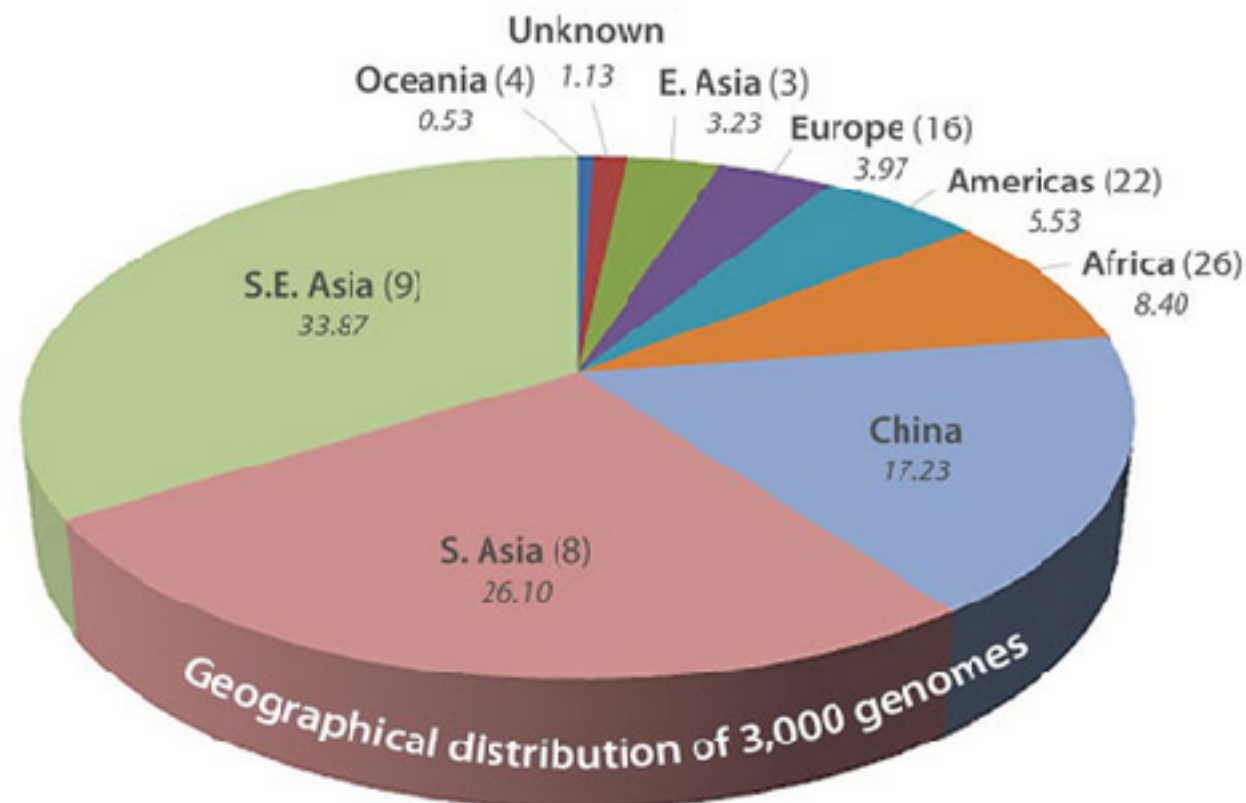
3k Rice Genomes Project

The 3,000 Rice Genomes Project

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



[View the full issue](#)



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



UK10K

Rare Genetic Variants in Health and Disease

What is UK10K?

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

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



Wellcome Library,
London

Project Design

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

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



The project received a £10.5 million funding award from the Wellcome Trust in March 2010 and sequencing started in late 2010. For more information, please use the links on the right hand side.

[HOME](#)

[GOALS](#)

[FUNDING](#)

[INVESTIGATORS](#)

[CONSORTIUM MEMBERSHIP](#)

[STUDY SAMPLES](#)

[ETHICS](#)

[DATA ACCESS](#)

[DATA & METHODS](#)

[DALLIANCE GENOME BROWSER](#)

[POSTERS](#)

[PUBLICATIONS](#)

[CONTACT US](#)

GGBN Global Genome Biodiversity Network

GGBN-GGI Awards Program

Mobilize your sample data!

Sample

DNA	148543
Tissues	251333
Cornea	10
Repetitive	1

Voucher

Cultures	21628
Vouchers	2
Specimens	154517
Unknown	6069
Collections	10

Taxa

Families	2019
Genera	11661
Species	30816

Total

606148

News

21 June 2017 - University of Helsinki Biodiversity Institute joins GGBN
 18 June 2017 - GGBN North American Regional Workshop Report now available
 08 June 2017 - GGBN-GGI Awards Program, 2017 Awardees
 22 May 2017 - GGBN Asia Pacific Summit and FAJGOW 2017
 02 May 2017 - GGBN Asia Regional Workshop (July 31-02) updates and RSVP deadline extended to May 31
 26 April 2017 - Job Opening at Natural History Museum, United Kingdom
 05 April 2017 - GGBN March 2017 newsletter now available
 01 April 2017 - Honorary Assistant and Learning Resources at Texas A&M University join GGBN
 27 March 2017 - Bishop Museum joins GGBN
 02 March 2017 - GGBN/ICP/WHO AGG Survey: Please participate by 29 March 2017
 01 March 2017 - Denver Botanic Gardens joins GGBN
 02 February 2017 - GGBN-GGI Awards Program 2017 Awardees and Honorary Assistant

Explore GGBN

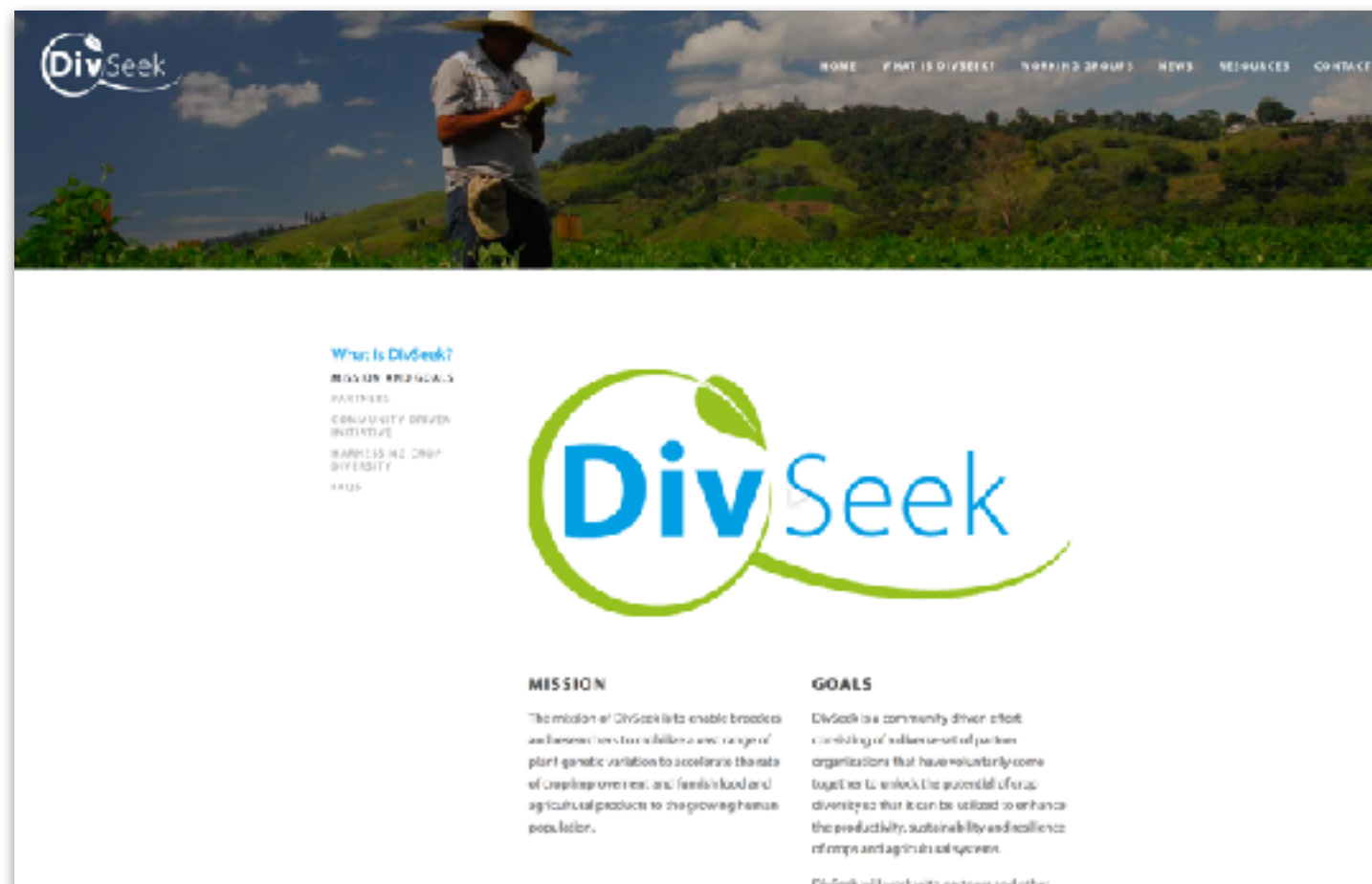
Repositories Search Documents About

GGBN

www.ggbn.org

GGBN provides a platform for biodiversity biobanks from across the world to:

- Collaborate to ensure consistent quality standards for DNA and tissue collections,
- Improve best practices for the preservation and use of such collections and
- Harmonize exchange and use of material in accordance with national and international legislation and conventions.



DivSeek

HOME WHAT IS DIVSEEK? JOINING GROUPS NEWS RESOURCES CONTACT

What is DivSeek?

MISSION

The mission of DivSeek is to enable breeders and researchers to mobilize a wide range of plant genetic variation to accelerate the rate of crop improvement and food security and agricultural products to the growing human population.

GOALS

DivSeek is a community driven effort consisting of a consortium of partner organizations that have voluntarily come together to unlock the potential of crop diversity so that it can be utilized to enhance the productivity, sustainability and resilience of crops and agricultural systems. DivSeek will work with breeders and other...

DivSeek

www.divseek.org

The mission of DivSeek is to enable breeders and researchers to mobilize a vast range of plant genetic variation to accelerate the rate of crop improvement and furnish food and agricultural products to the growing human population.

**GOLD**

GENOMES ONLINE DATABASE

[JGI HOME](#) [LOG IN](#)

genome gold

[Home](#) [Search](#) [Distribution Graphs](#) [Biogeographical Metadata](#) [Statistics](#) [GOLD Usage Policy](#) [Team](#) [Help](#) [News](#)

Studies	29,300
Biosamples	24,719
Sequencing Projects	143,738
Analysis Projects	120,223
Organisms	278,521

Welcome to the Genomes OnLine Database

GOLD Release v.6

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

1. Register



Register your project information and Metadata in the Genomes Online Database

[Register](#)

2. Annotate



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

[Annotate](#)

3. Publish



Standards in Genomic Sciences

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

[Publish](#)

Studies

Metagenomic [1,060](#)Non-Metagenomic [28,239](#)

Biosamples

[Classification](#)

Ecosystems

Host-associated [8,296](#)Engineered [3,325](#)Environmental [13,096](#)

Sequencing Projects

[Complete Projects](#) [11,592](#)[Permanent Drafts](#) [82,987](#)[Incomplete Projects](#) [47,073](#)[Targeted Projects](#) [1,233](#)

Analysis Projects

Genome Analysis [85,195](#)Metagenome Analysis [14,671](#)Metagenome - Cell Enrichment [830](#)Metagenome - Single Particle Sort [1,697](#)Metagenome - Assembled Genome (MAG) [5,051](#)Metatranscriptome Analysis [2,616](#)Combined Assembly [115](#)Single Cell - Screened (SAG) [2,200](#)Single Cell - Unscreened (SAG) [1,044](#)Transcriptome Analysis [117](#)

Special Projects

Type Strain Projects [5,845](#)Strains at Genbank [4,326](#)GEBA Projects [3,137](#)HMP Projects [2,913](#)

Projects with Genbank Data

Seq. Projects [80,587](#)Archaeal Projects [728](#)Bacterial Projects [68,566](#)Eukaryal Projects [3,325](#)Viral Projects [7,949](#)

JGI Projects

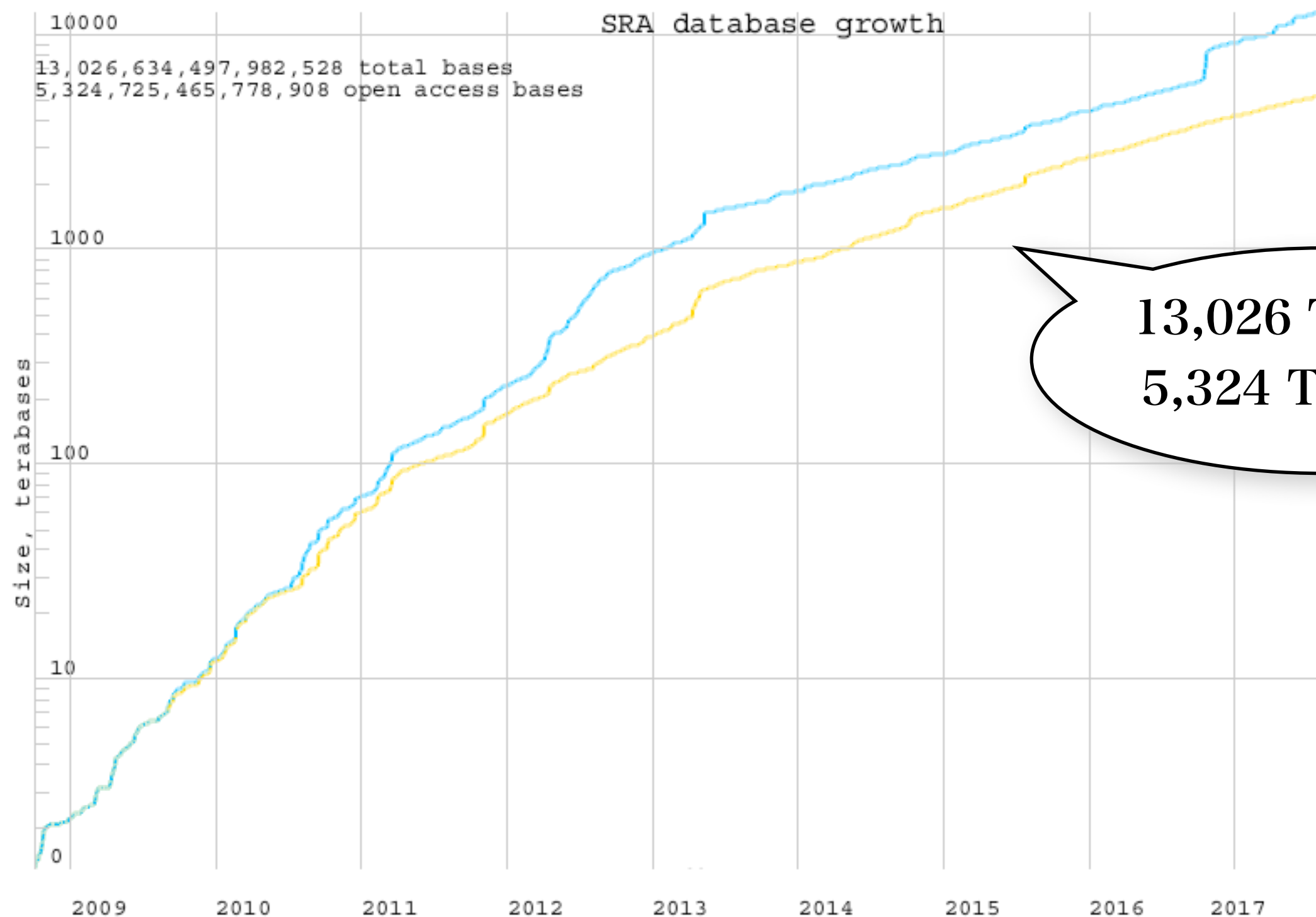
JGI Studies [1,190](#)JGI Biosamples [11,224](#)JGI Sequencing Projects [50,670](#)JGI Analysis Projects [29,071](#)

Organisms

Organisms [278,504](#)Archaea [2,410](#)Bacteria [249,671](#)Eukarya [18,138](#)Viruses [8,256](#)Bacterial Type Strains [10,694](#)

SRA growth (NGS row/bam data amount)

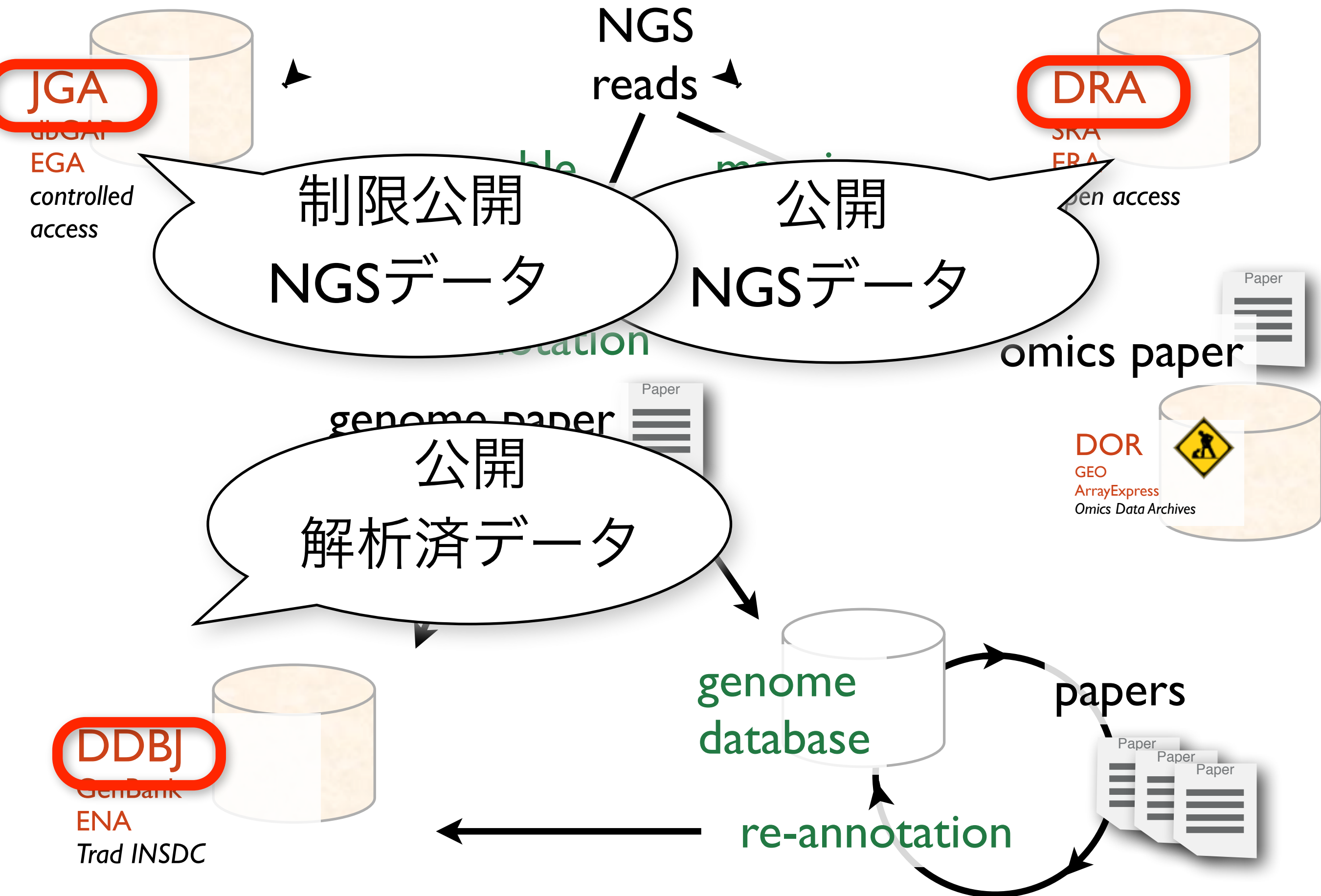
NCBI Sequence Read Archive



13,026 TB total
5,324 TB open

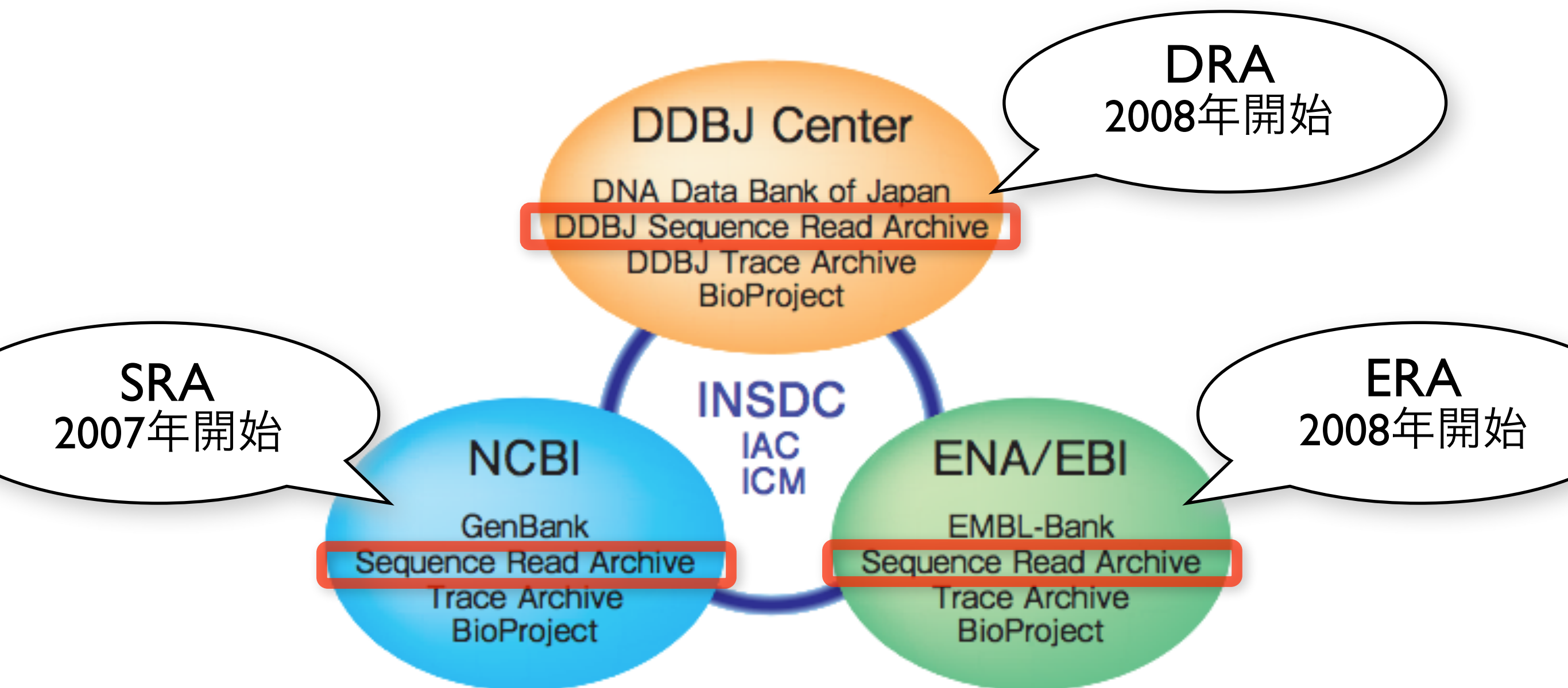
Total bases —
Open access bases —

The ecosystem of sequence data



DDBJ Sequence Read Archive (DRA)

新世代シーケンサから出力される配列や
アライメントデータを登録・公開



今時（いまどき）のシーケンサー【再掲】



(左) PacBio RSII System

(中) illumina: HiSeq 4000 System

(右) Oxford NANOPORE MinION/SmidgION

STAP細胞 遺伝子データに説明と矛盾する点

6月11日 18時09分

STAP細胞の染色体異常
[マウス生まれず 説明と矛盾]

era plugin for fast download

Show RUNs

Accession	Spots	Bases	Download
SRS559090	71.8M	14.5G	8.4G
SRX472647	33.4M	6.8G	3.9G
SR	8番目の染色体に異常あるマウスは 胎児の段階で死に 通常は生まれてこず		4.5G

NHK

理化学研究所の小保方晴子研究ユニットリーダーらが発表したSTAP細胞の論文を巡っては、新たな疑義が次々と指摘されていますが、STAP細胞のものだとしている遺伝子のデータにも、これまでの説明と矛盾する不自然な点のあることが専門家のグループの分析で新たに分かりました。

この分析を行ったのは、横浜にある理化学研究所統合生命医科学研究センターの遠藤高帆上級研究員らと東京大学の2つのグループで、それぞれ独自に行いました。

それによりますと、小保方リーダーらがインターネット上に公開しているSTAP細胞のものだとする遺伝子の情報を分析したところ、ほぼすべての細胞に8番目の染色体が通常の2本より1本多くなる「トリソミー」と呼ばれる異常のあることが分かったということです。

8番目の染色体がトリソミーを起こしたマウスは、母親のおなかの中にいる胎児の段階で死んでしまい、通常、生まれてこないということで、専門家グループは、小保方リーダーらが生後1週間ほどのマウスからSTAP細胞を作ったとするこれまでの主張と矛盾するとしています。

また8番染色体のトリソミーは、すでに研究で広く使われている万能細胞「ES細胞」を長い間培養すると起きることがある異常としても知られています。

<http://www3.nhk.or.jp/news/html/20140611/k10015134871000.html> より

DRAウェブサイト ⇒ [DRA DDBJ] で検索

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

登録関係情報

Sequence Read Archive

Login & Submit | Databases ▼ | English | Contact

Google™ カスタム検索



Home

Submission ▼

Search

Download ▼

Pipeline

About

解析パイプライン

DDBJ Sequence Read Archive (DRA) は, the 454 GS System®, Illumina Genome Analyzer®, Applied Biosystems SOLiD® System などの次世代シーケンサからの出力データを登録・解析・公開する国際的なデータベース (International Nucleotide Sequence Database Collaboration (INSDC) のメンバーであり、GenBank, EMBL, and DDBJ との国際協力ののもと、運営されています。従来のキャピタシオン・シーケンサからの出力データは EMBL/GenBank Archive にご登録ください。

データ検索

データ取得



検索

データをキーワード、生物名、シーケンサなどで検索する



登録

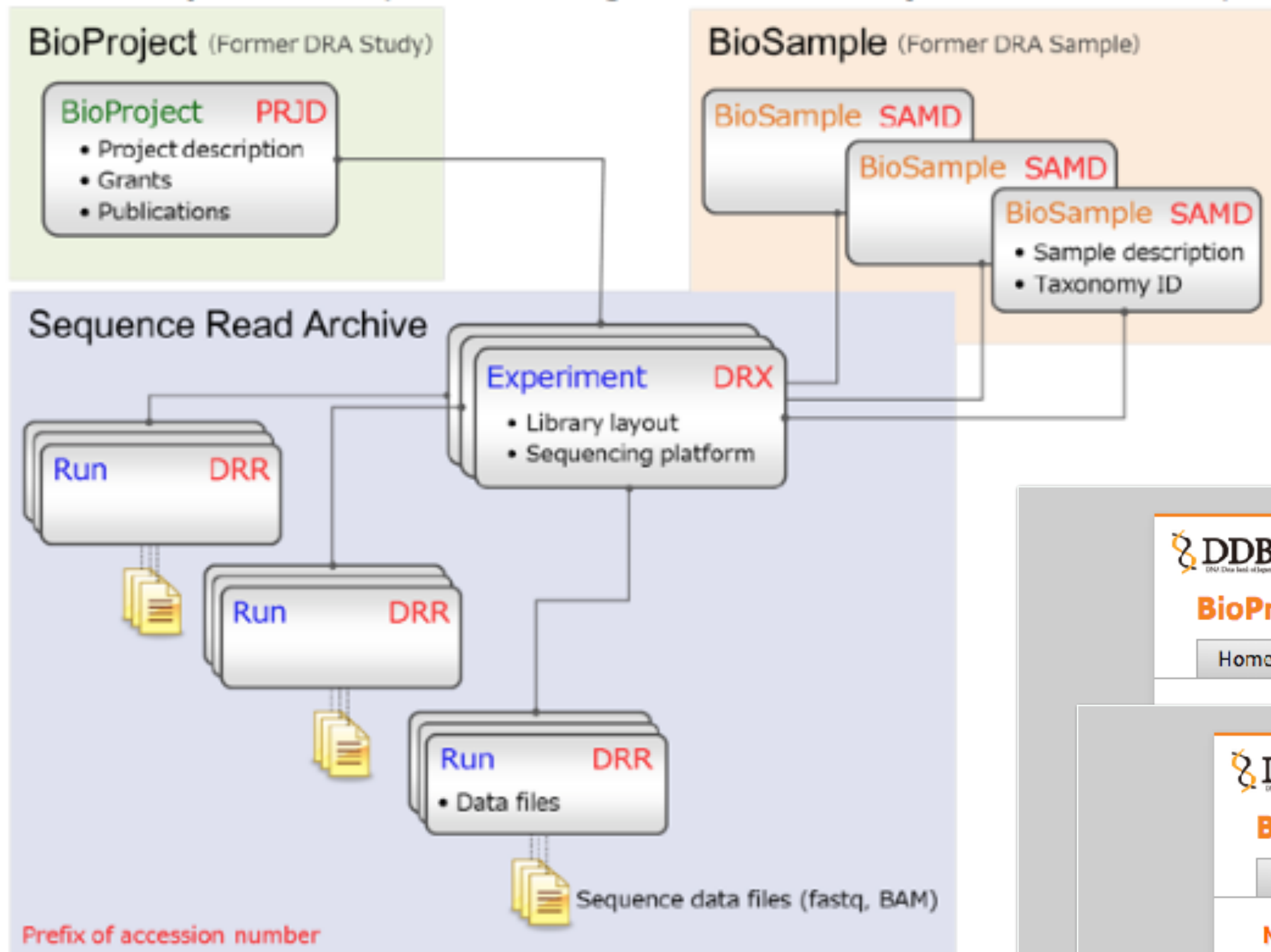
新型シーケンサからの生データやアライメントデータを登録する



動画マニュアル

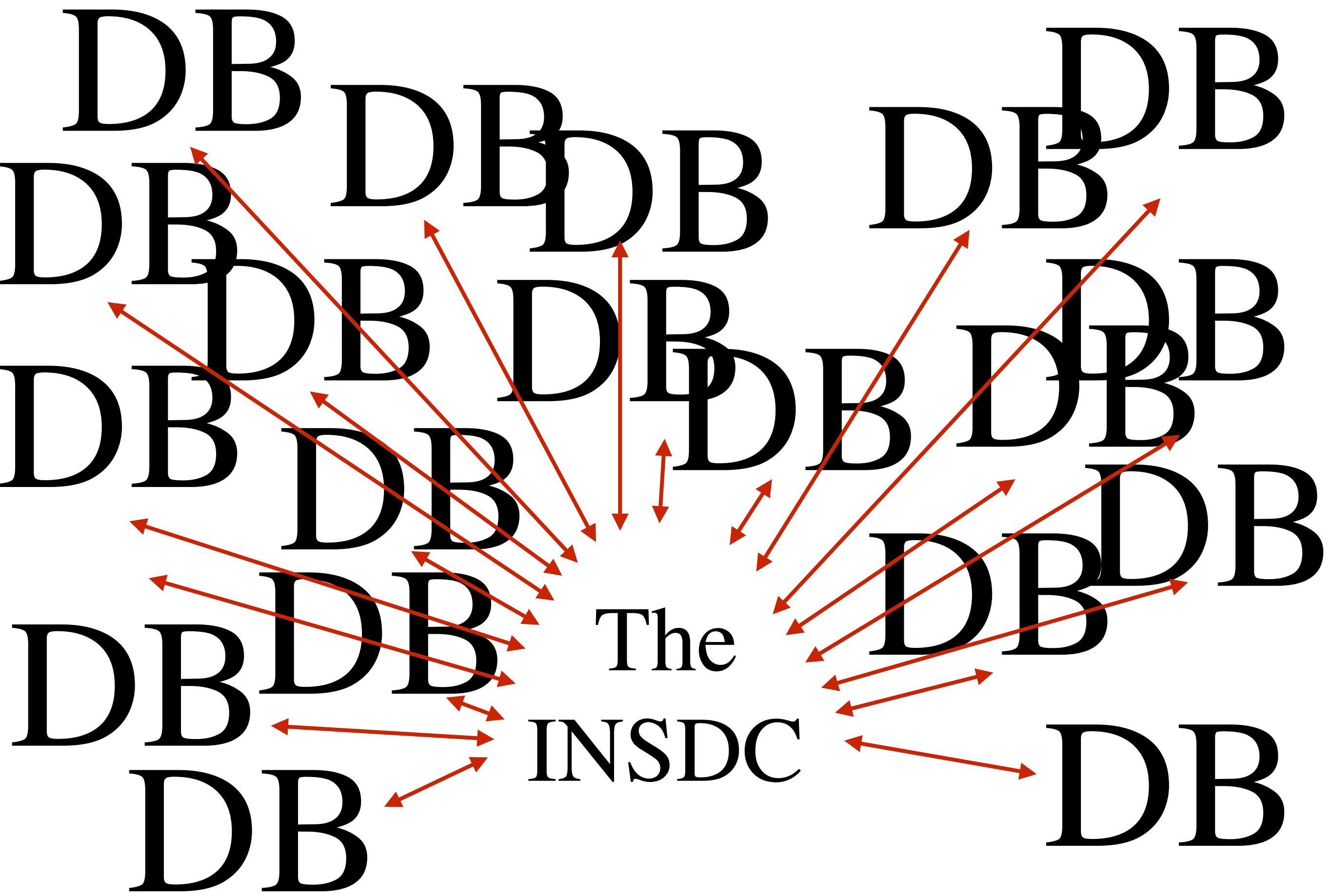
DRA の利用方法や登録方法を解説している動画を見る

DRA study and sample were migrated to BioProject and BioSample, respectively.



> 1,500

[illegible]



遺伝研スーパー
コンピュータ

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71 72 73 74 75 76 77 78 79 80 81 82 83 84 85 86 87 88 89 90 91 92 93 94 95 96 97 98 99 100 101 102 103 104 105 106 107 108 109 110 111 112 113 114 115 116 117 118 119 120 121 122 123 124 125 126 127 128 129 130 131 132 133 134 135 136 137 138 139 140 141 142 143 144 145 146 147 148 149 150 151 152 153 154 155 156 157 158 159 160 161 162 163 164 165 166 167 168 169 170 171 172 173 174 175 176 177 178 179 180 181 182 183 184 185 186 187 188 189 190 191 192 193 194 195 196 197 198 199 200 201 202 203 204 205 206 207 208 209 210 211 212 213 214 215 216 217 218 219 220 221 222 223 224 225 226 227 228 229 230 231 232 233 234 235 236 237 238 239 240 241 242 243 244 245 246 247 248 249 250 251 252 253 254 255 256 257 258 259 260 261 262 263 264 265 266 267 268 269 270 271 272 273 274 275 276 277 278 279 280 281 282 283 284 285 286 287 288 289 290 291 292 293 294 295 296 297 298 299 300 301 302 303 304 305 306 307 308 309 310 311 312 313 314 315 316 317 318 319 320 321 322 323 324 325 326 327 328 329 330 331 332 333 334 335 336 337 338 339 340 341 342 343 344 345 346 347 348 349 350 351 352 353 354 355 356 357 358 359 360 361 362 363 364 365 366 367 368 369 370 371 372 373 374 375 376 377 378 379 380 381 382 383 384 385 386 387 388 389 390 391 392 393 394 395 396 397 398 399 400 401 402 403 404 405 406 407 408 409 410 411 412 413 414 415 416 417 418 419 420 421 422 423 424 425 426 427 428 429 430 431 432 433 434 435 436 437 438 439 440 441 442 443 444 445 446 447 448 449 450 451 452 453 454 455 456 457 458 459 460 461 462 463 464 465 466 467 468 469 470 471 472 473 474 475 476 477 478 479 480 481 482 483 484 485 486 487 488 489 490 491 492 493 494 495 496 497 498 499 500 501 502 503 504 505 506 507 508 509 510 511 512 513 514 515 516 517 518 519 520 521 522 523 524 525 526 527 528 529 530 531 532 533 534 535 536 537 538 539 540 541 542 543 544 545 546 547 548 549 550 551 552 553 554 555 556 557 558 559 560 561 562 563 564 565 566 567 568 569 570 571 572 573 574 575 576 577 578 579 580 581 582 583 584 585 586 587 588 589 590 591 592 593 594 595 596 597 598 599 600 601 602 603 604 605 606 607 608 609 610 611 612 613 614 615 616 617 618 619 620 621 622 623 624 625 626 627 628 629 630 631 632 633 634 635 636 637 638 639 640 641 642 643 644 645 646 647 648 649 650 651 652 653 654 655 656 657 658 659 660 661 662 663 664 665 666 667 668 669 670 671 672 673 674 675 676 677 678 679 680 681 682 683 684 685 686 687 688 689 690 691 692 693 694 695 696 697 698 699 700 701 702 703 704 705 706 707 708 709 710 711 712 713 714 715 716 717 718 719 720 721 722 723 724 725 726 727 728 729 730 731 732 733 734 735 736 737 738 739 740 741 742 743 744 745 746 747 748 749 750 751 752 753 754 755 756 757 758 759 760 761 762 763 764 765 766 767 768 769 770 771 772 773 774 775 776 777 778 779 780 781 782 783 784 785 786 787 788 789 790 791 792 793 794 795 796 797 798 799 800 801 802 803 804 805 806 807 808 809 810 811 812 813 814 815 816 817 818 819 820 821 822 823 824 825 826 827 828 829 830 831 832 833 834 835 836 837 838 839 840 841 842 843 844 845 846 847 848 849 850 851 852 853 854 855 856 857 858 859 860 861 862 863 864 865 866 867 868 869 870 871 872 873 874 875 876 877 878 879 880 881 882 883 884 885 886 887 888 889 890 891 892 893 894 895 896 897 898 899 900 901 902 903 904 905 906 907 908 909 910 911 912 913 914 915 916 917 918 919 920 921 922 923 924 925 926 927 928 929 930 931 932 933 934 935 936 937 938 939 940 941 942 943 944 945 946 947 948 949 950 951 952 953 954 955 956 957 958 959 960 961 962 963 964 965 966 967 968 969 970 971 972 973 974 975 976 977 978 979 980 981 982 983 984 985 986 987 988 989 990 991 992 993 994 995 996 997 998 999 1000 1001 1002 1003 1004 1005 1006 1007 1008 1009 1010 1011 1012 1013 1014 1015 1016 1017 1018 1019 1020 1021 1022 1023 1024 1025 1026 1027 1028 1029 1030 1031 1032 1033 1034 1035 1036 1037 1038 1039 104

A circular graphic with a black outline, containing a DNA sequence in bold, black, uppercase letters. The sequence is arranged in seven lines, with the last line being partially cut off. The sequence is: TGTGTTGCTTC, TCAGAATTTTG, AGACCTGACAGAA, AACGGATGTTGCT, AGAACGCTTGGT, GGTATGAAGT. The background of the slide is a light blue gradient with faint, repeating text from the DNA sequence.

遺伝研スパコン

DDBJ の紹介

利用の手引き

レポート・統計

FAQ

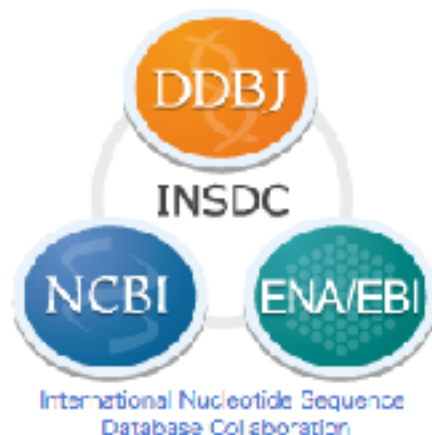
お問い合わせ

第31回
DDBJing 講習会 in 東京
6月12日 × JST 東京本部別館

RSSを購読する

DDBJ Twitter

Mail Magazine



International Nucleotide Sequence
Database Collaboration

大学共同利用機関法人 情報・システム研究機構
国立遺伝学研究所

大学共同利用機関法人
情報・システム研究機構
Research Organization of Information and Systems

DBCLS
Database Center
for Life Science

DDBJ Service



登録

Data Submission



検索・解析

Search / Analysis



スパコン

Super Computer



アーカイブ

ftp.ddbj.nig.ac.jp

Hot Topics

ニュースアーカイブ

お知らせ

データ公開

広報

メンテナンス

運用情報

一覧

- 2015.06.10 アノクビダイコン (*Raphanus sativus* cv. Aokubi) ゲノムと転写産物配列データの公開
- 2015.06.05 国立遺伝学研究所 特任研究員 (DDBJ アノテータ) 募集
- 2015.06.03 UniProt 2015_06 公開
- 2015.04.20 「第31回 DDBJing 講習会 in 東京」開催のお知らせ

遺伝研スーパーコンピュータサイト top

<http://sc.ddbj.nig.ac.jp/>



The screenshot shows the homepage of the NIG SuperComputer system. The browser's address bar displays sc.ddbj.nig.ac.jp. The website header includes the NIG logo and the text "大学共同利用法人 情報・システム研究機構 国立遺伝学研究所 スーパーコンピュータシステム SuperComputer Facilities of National Institute of Genetics". A search bar is located on the right. The main content area features a "重要なお知らせ" (Important Notice) section with a list of updates from 2014, including system maintenance and software installation. Below this is a section titled "国立遺伝学研究所 スーパーコンピュータシステム(NIG SUPERCOMPUTER)とは" (What is the NIG SuperComputer System?), which describes the system's capabilities for large-scale genomic data analysis. A large red arrow points to a row of five photographs showing the server racks of the supercomputer system. The left sidebar contains navigation links for language, home, login, system configuration, and usage.

Language/言語
日本語
English

ホーム

このサイトへのログイン
Login
(スパコンユーザでログイン可)

システム構成
ハードウェア構成
ソフトウェア構成
プログラミング環境
利用可能バイオツール
利用可能OSS
利用可能DB

システム使用方法
基本的利用方法
その他UGE利用方法
ファイル転送方法
システム利用TIPS

重要なお知らせ

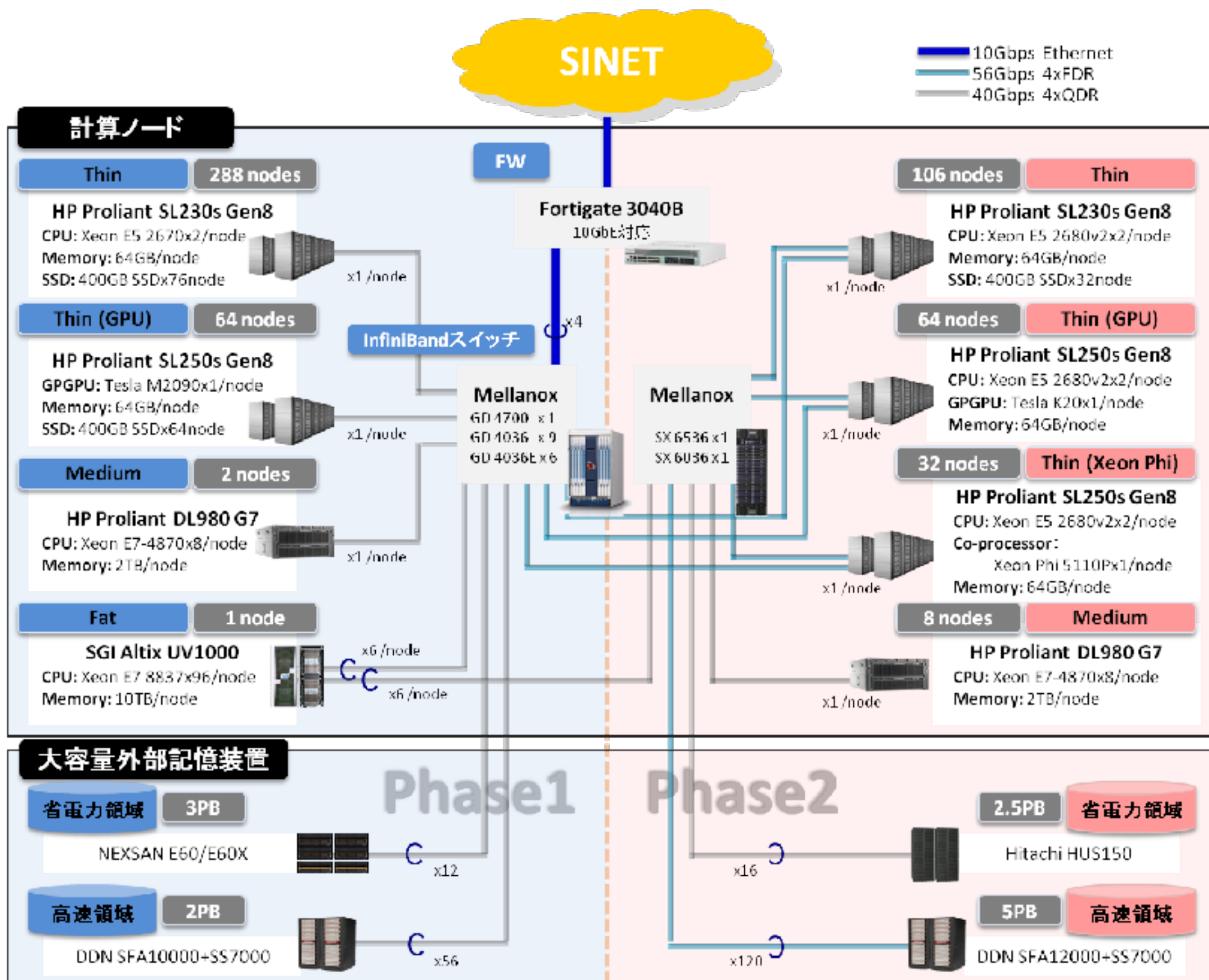
公開日 表題
2014年7月22日 【スパコンユーザ会】三島開催会場変更のお知らせ
2014年7月18日 ソフトウェアインストールのお知らせ
2014年7月15日 【fat障害：2】fat計算ノードサービス再開のお知らせ
2014年7月14日 【通信障害：2】Phase1システム通信不具合の復旧のお知らせ
2014年7月14日 【通信障害】Phase1システム通信不具合のお知らせ
2014年3月4日 2014年3月5日からのスパコンPhase2システムご利用方法について

国立遺伝学研究所 スーパーコンピュータシステム(NIG SUPERCOMPUTER)とは

大学共同利用機関法人 情報システム研究機構 国立遺伝学研究所は、2012年3月にスーパーコンピュータシステムを更新しました。新しいスーパーコンピュータシステムはゲノム解析を主な目的とした大規模計算機利用拠点として 最新鋭の大規模クラスター型計算機、大規模メモリ共有型型計算機、および大容量高速ディスク装置で構成されたスーパーコンピュータシステムサービスを提供しています。

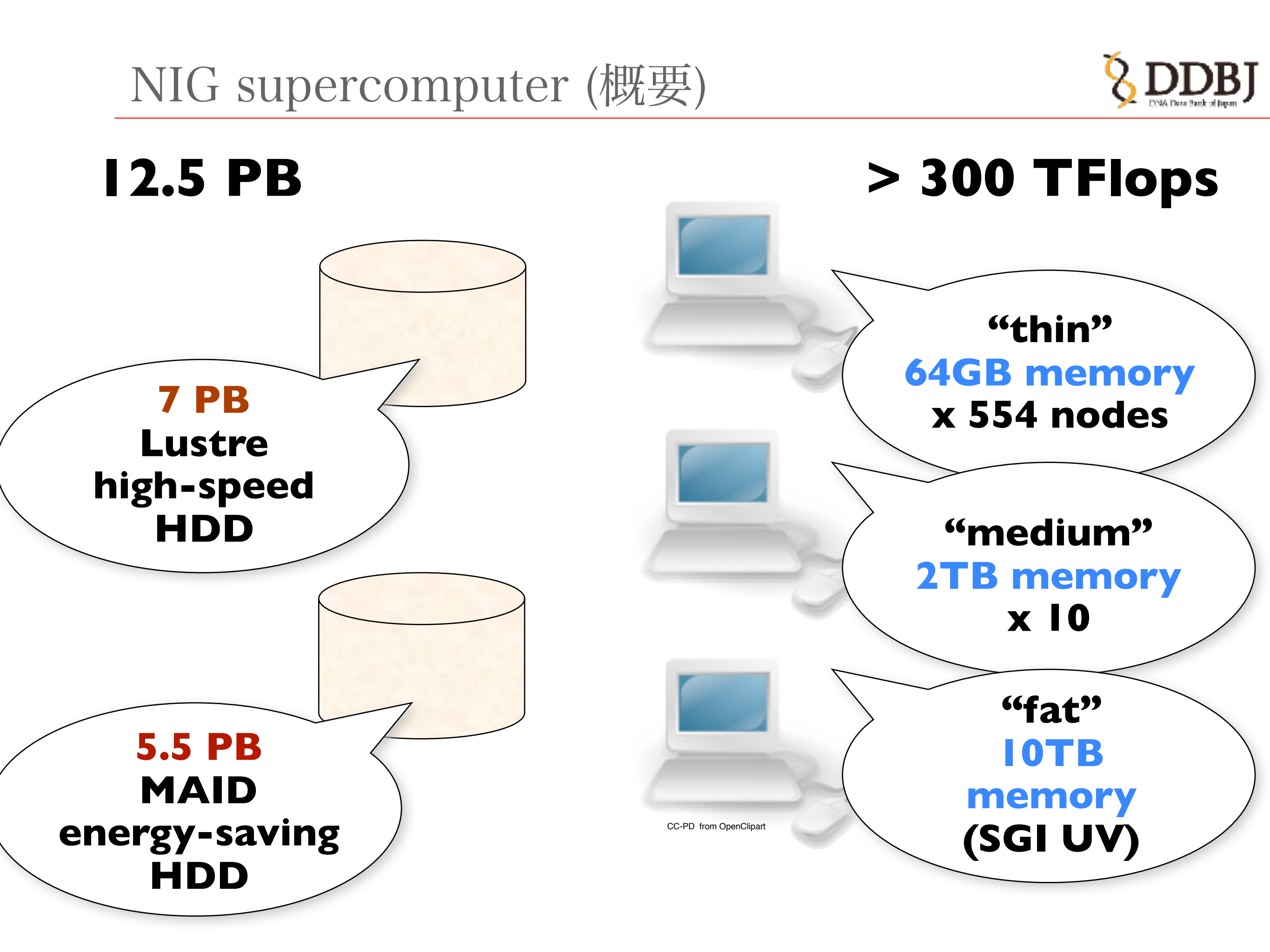
システムハードウェア構成
システムソフトウェア構成
システム稼働状況

遺伝研スーパーコンピュータ全容



12.5 PB

> 300 TFlops



The diagram illustrates the architecture of the NIG supercomputer. On the left, two large orange cylinders represent storage systems. The top cylinder is connected to a speech bubble containing text about Lustre HDD, and the bottom cylinder is connected to a speech bubble about MAID HDD. On the right, three computer icons represent different node types. Each icon is connected to a speech bubble describing its memory configuration: 'thin' nodes (64GB x 554), 'medium' nodes (2TB x 10), and 'fat' nodes (10TB SGI UV). The overall system capacity is noted as 12.5 PB and performance as > 300 TFlops.

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)

解析パイプラインも提供してます

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



Sequence Read Archive

Home

Submission ▼

Search

Download ▼

Pipeline

About

解析パイプライン

DDBJ Sequence Read Archive (DRA) は Roche 454 GS System®, Illumina Genome Analyzer®, Applied Biosystems SOLiD® System などの次世代シーケンサからの出力データのためのデータベースです。DRA は International Nucleotide Sequence Database Collaboration (INSDC) のメンバーであり、NCBI Sequence Read Archive (SRA) と EBI Sequence Read Archive (ERA) との国際協力のもと、運営されています。従来のキャピラリー式シーケンサからの出力データは DDBJ Trace Archive にご登録ください。



検索

データをキーワード、生物名、シーケンサなどで検索する



登録

新型シーケンサからの生データやアライメントデータを登録する



動画マニュアル

DRA の利用方法や登録方法を解説している動画をみる

DDBJ pipeline: ソフトウェア

よく用いられる
解析用ソフトウェアを
用意。クリックだけで
実行可能



ACCOUNT

login ID [yaskaz]

Logout

Change password

ANALYSIS

Data setup

DRA Start

FTP upload

HTTP upload

DRA Import

Preprocessing Start

step-1

Preprocessing

Mapping /
de novo Assembly

step-2

Workflow

Genome (SNP/Short
Indel)

RNA-seq (Tag count)

ChIP-seq

JOB STATUS

step1.

Preprocessing

step1.

Mapping

step1.

de novo Assembly

step2-All status

HELP

HELP

TUTORIAL

Contact Us.
DDBJ Read Annotation
Pipeline.
Development Team.

Select Query Files

Select Tools

Set QuerySet

Set Genom

Running Status

Selecting Tools for Basic Analysis of DDBJ ANNOTATION PIPELINE

BACK

NEXT

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		0.12.7	✓	✓	✓			✓	✓				✓	
☐	TopHat		1.0.11	✓		✓			✓					✓	
☐	Bowtie2		2.0.0	✓	✓	✓			✓	✓				✓	For reads longer than about 50 bp, Bowtie2 is generally faster, more sensitive, and uses less memory than Bowtie1.

de novo Assembly

Total limit = 22 Gbp

	Tool	Help	Version	Base space	Color space	Paired-end	MSS(WGS)	Comment
☐	SOAPdenovo		1.05			✓		
☐	ABYSS		1.3.2			✓		Maximum K-mer value is 64.
☐	Velvet		1.2.03			✓	✓	We severe recommend when performing Velvet, total length of those reads is up to 22G bp.Maximum K-mer value is 64.
☐	Euler		r2013-02-			✓		

良く使われるソフトウェアの特徴

マッピング

BLAT	高速シーケンサー登場以前からあるアライメントツール。発現データはイントロンを想定したギャップを考慮。
MAQ	高速シーケンサー登場初期にショートリードに対応。リード長が長くなるに従い開発はBWAに引き継がれる。
BWA	MAQより速く、Titaniumのリードもオプションで対応。
SOAP	メモリ消費量少なく、高速。精度はBWAより弱冠落ちる。
Bowtie/ Bowtie2	ギャップは考慮しないが処理は速い。BWA、SOAP2、Bowtieは Burrows-Wheeler変換というアルゴリズムでゲノムDNAにたいしてインデックスを作成、高速でマッピングする。Bowtie2は50bp以上に最適化。
TopHat	RNA-Seqのリードを内部でBowtieを利用してマッピング、スプライスジャンクションを特定する。

アセンブル

SOAPdenovo	ヒト、パンダ等大型ゲノムのアセンブリで使用された。比較的高速。
Abyss	初期に並列処理に対応したアセンブラ。
Velvet	高速シーケンサー登場初期に開発された。メモリ消費多め。
Trinity	RNA-Seq配列のアセンブラ。 上記3つともにde bruijn graphというアルゴリズムを使用

DDBJ pipeline: 比較対象

イネ、マウスなど
解析比較対象となる
配列を多数用意

DDBJ
DNA Data Bank of Japan

ACCOUNT
login ID [yaskaz]
Logout
Change password

ANALYSIS
Data setup
DRA Start
FTP upload

JOB STATUS
step1. Preprocessing
step1. Mapping
step1. de novo Assembly
step2-All status

HELP

Select Query Files → Select Tools → Set QuerySet → Running Status

Specifying Database of Reference

Major genome sets

Organisms: Arabidopsis thaliana
Genome sets: TAIR8, TAIR9, TAIR10

Major genome sets

Organisms: Oryza sativa japonica
Genome sets: IRGSP Releases Build 4.0, IRGSP Releases Build 5.0, IRGSP Releases Build 5.0 masked by RepeatMasker with tigr version5.0, tigr version6.0, tigr version6.1, tigr mitochondrion, tigr chloroplast

Major genome sets

Organisms: Homo sapiens
Genome sets: Homo sapiens Feb. 2009 (hg19), Mar.2006 (hg18), May.2004 (hg17), NCBI build 36.1_CRA, NCBI build 36.1_Celera, NCBI build 36.1_ref, NCBI build 36.2_CRA, NCBI build 36.2_Celera, NCBI build 36.2_ref, NCBI build 36.3_CRA, NCBI build 36.3_Celera, NCBI build 36.3_ref, NCBI build 36.3_HuRef, NCBI build 37.1_CRA, NCBI build 37.1_Celera, NCBI build 37.1_GRCh, NCBI build 37.1_HuRef

NGS's + SC's in Biology



“medium”
2TB memory
x 10



“fat”
10TB memory
(SGI UV)

