

平成25年度情報解析講習会

DDBJ スパコン概要

～まずはログイン

国立遺伝学研究所
大量遺伝情報研究室
中村保一



@yaskaz

a.k.a. catlover, ikasumipapa,
猫教授

✿ 使い倒し系バイオインフォマティスト



✿ 植物とか微生物のゲノム解析+DB屋



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

シロイヌナズナの 1/4
(27 Mb, 6200 genes) の解析



[http://genome.kazusa.or.jp/
cyanobase/](http://genome.kazusa.or.jp/cyanobase/)

光合成細菌のゲノム解析+データ
ベース。Social Bookmark による
遺伝子注釈系



DNA Data Bank of Japan

English

Google™カスタム検索

Search

DDBJ の紹介

利用の手引き

レポート・統計

Q and A

お問い合わせ

 Web Magazine

 RSSを購読する

 DDBJ Twitter

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

一覧

- 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)

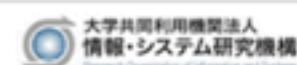
Maintenance

一覧

Information

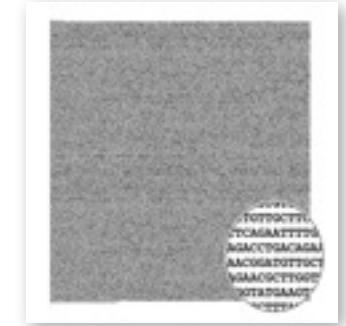


大学共同利用機関法人
情報・システム研究機構
国立遺伝学研究所

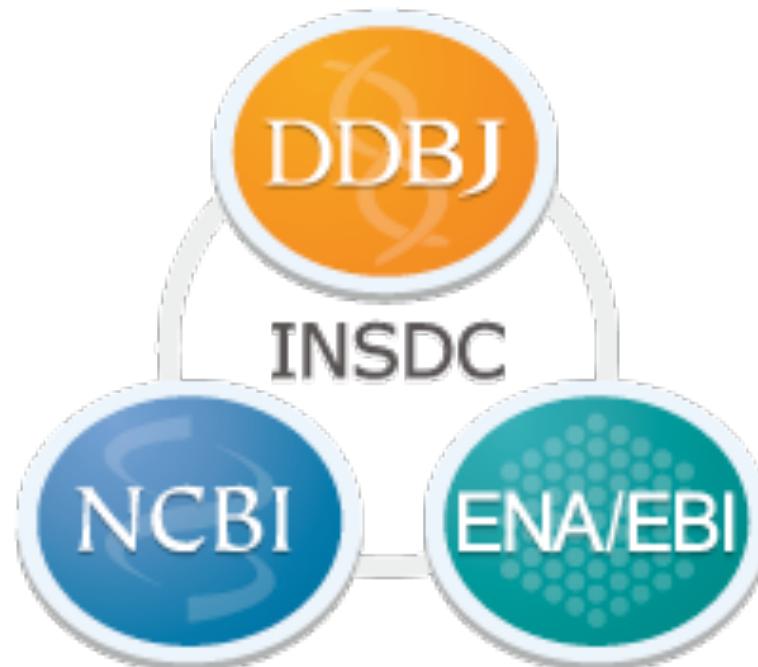


大学共同利用機関法人
情報・システム研究機構
Research Organization of Information and Systems

- 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



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



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, Yoshihisa 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, Yoshihisa 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 Siroткин, 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

DNA Data Bank of Japan

English

Google™カスタム検索

Search

DDBJ の紹介 利用の手引き レポート・統計 Q and A お問い合わせ

Web Magazine

RSSを購読する

DDBJ Twitter

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

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)

Maintenance

Information

一覧

一覧

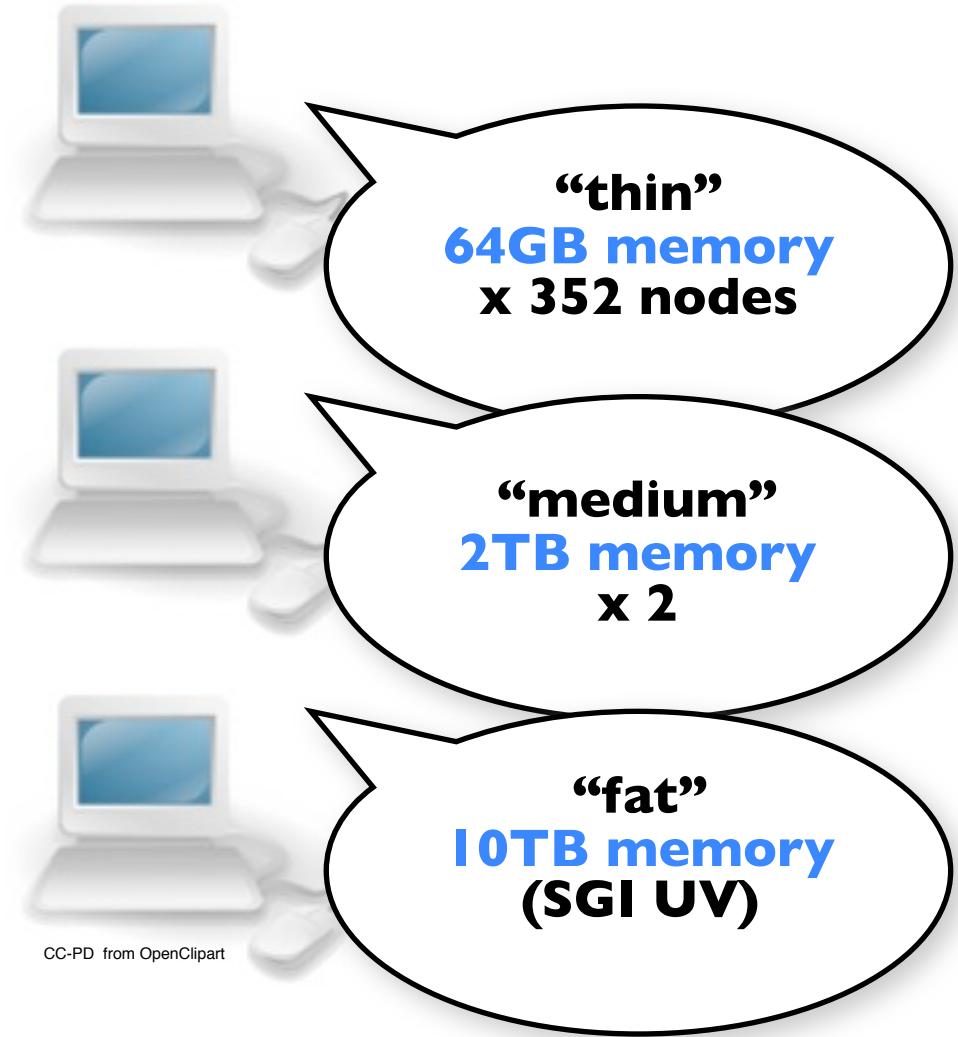
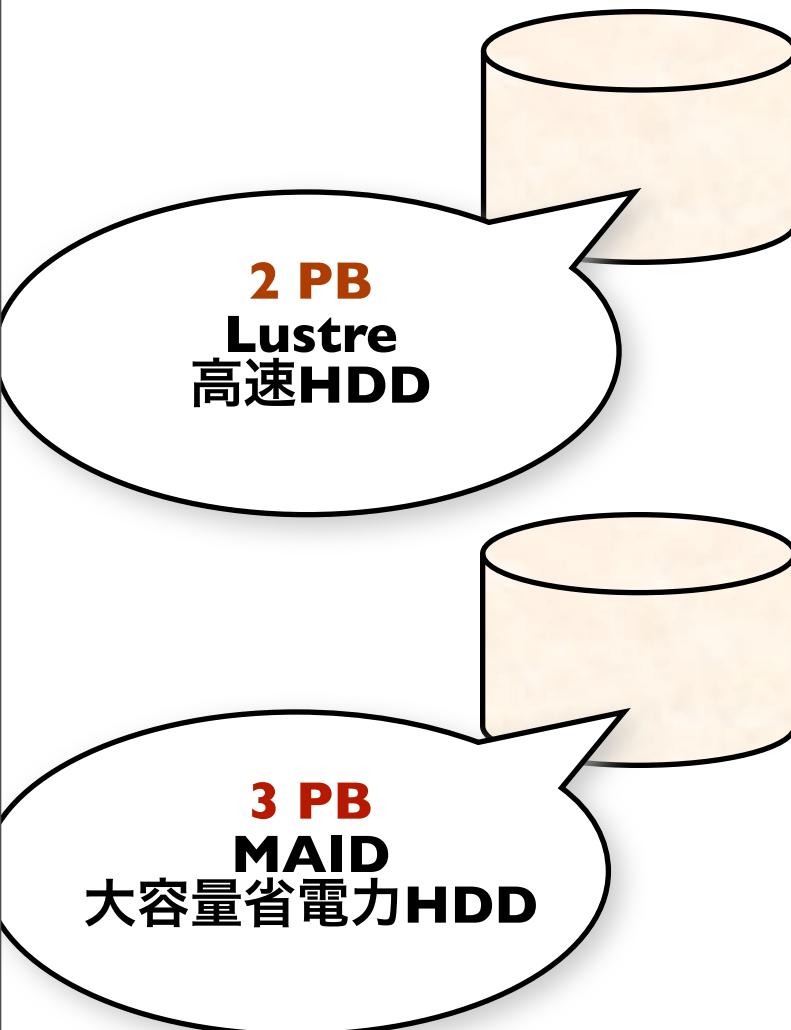
NIG 国立遺伝学研究所

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

遺伝研スーパー

コンピュータ

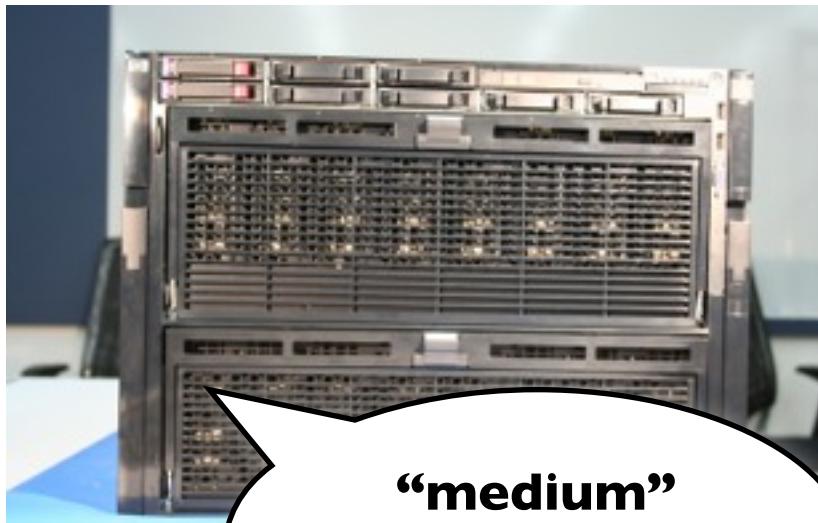


Bell 研の Ken Thompson,
Dennis Ritchieらが “Space
Travel” で遊ぶために部屋のス
ミに置いてあった借り物の
PDP-7 で「えいやつ」と作って
みた「小さくて」「軽い」オペ
レーティングシステム (1968年)

Unix の特徴

- 初期のマシンはメモリ空間が狭かったために個々のプログラムのサイズが限られていた。
- このことが、小さなコマンドをパイプでつなげて、出力を次のコマンドに入れていく「ツールボックス」アプローチを促進した
- 現在、大型の計算機の多くは Unix か、少なくとも Unix のフロントエンドを持っている
- Unix-like で PC で使う OS に Linux がある
- MacOSX は Mac UI の皮をかぶった Unix

NGS's + SC's in Biology



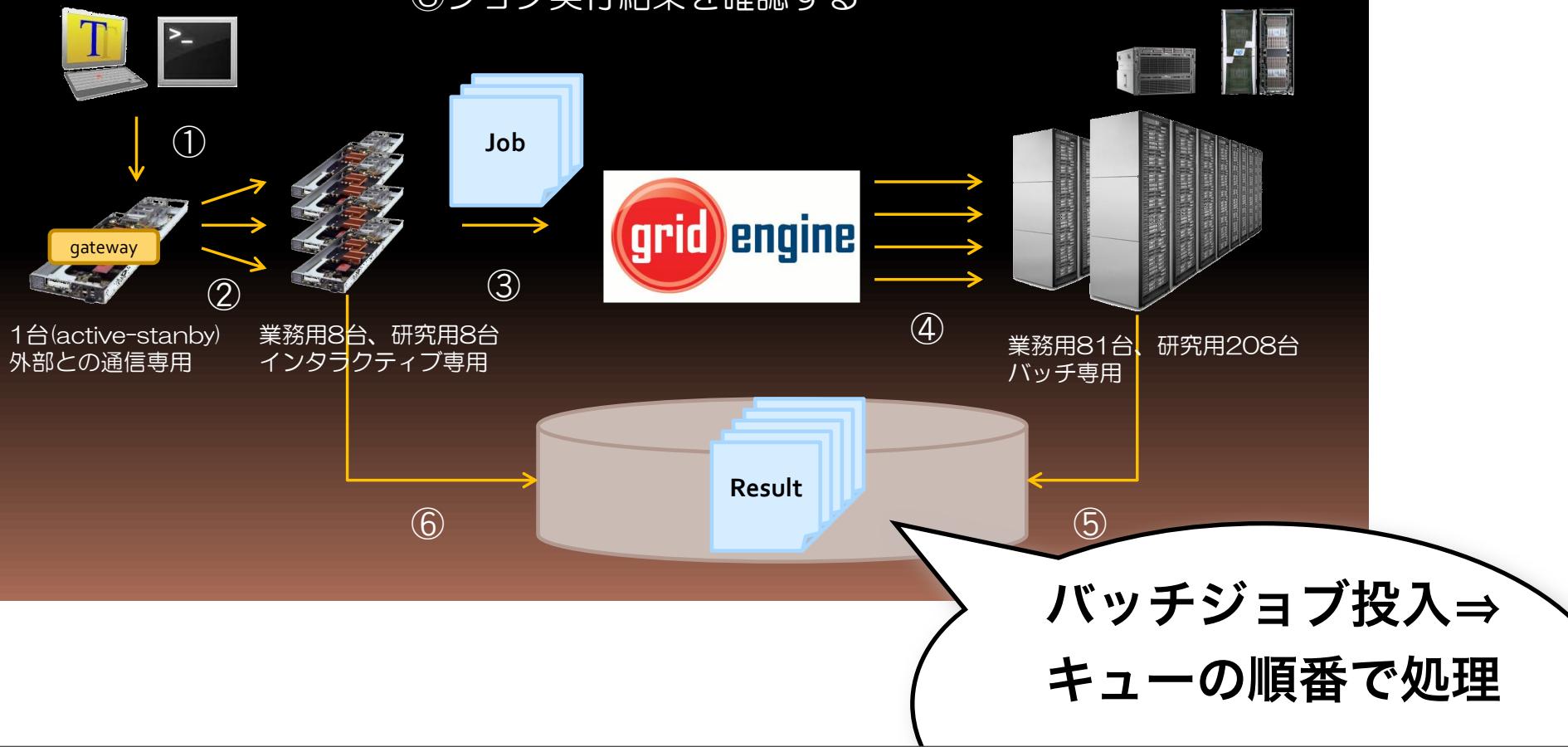
**“medium”
2TB memory
x 2**



**“fat”
10TB memory
(SGI UV)**

スパコン使用方法(イメージ)

- ①ゲートウェイノード(gw.ddbj.nig.ac.jp)にログインする
- ②qloginを実行しインタラクティブノードにログインする
- ③qloginしたホストからジョブをUGEに投入する
- ④UGEは負荷の低いノードでジョブを実行する
- ⑤ジョブ実行結果をlustreのホームディレクトリに出力する
- ⑥ジョブ実行結果を確認する

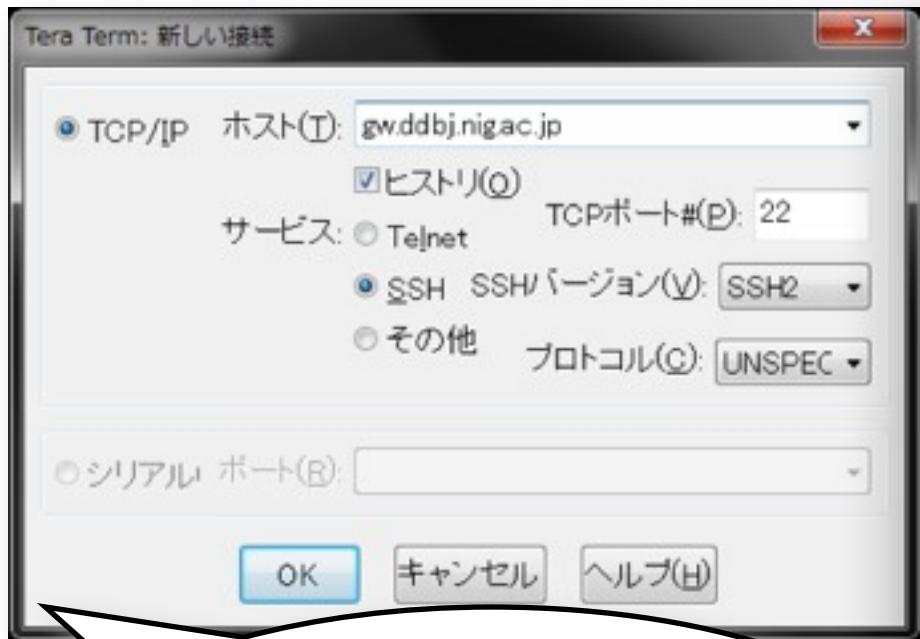


ログインして

みましよう

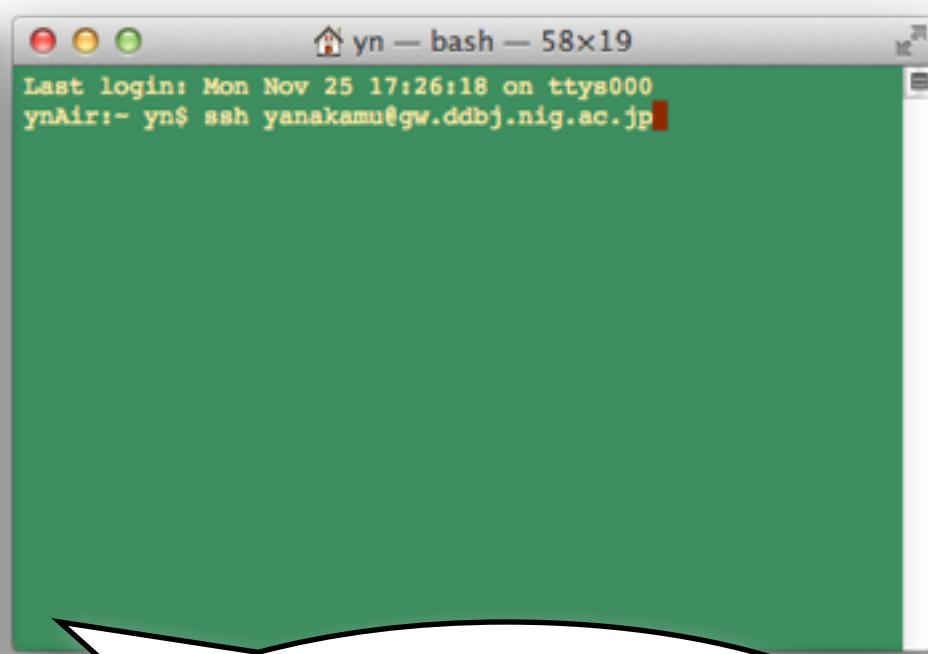
ログイン (1)

- Windowsから
 - Tera Term 起動



ホスト : **gw.ddbj.nig.ac.jp**
 サービス : **SSH**
 ⇒ **OK**

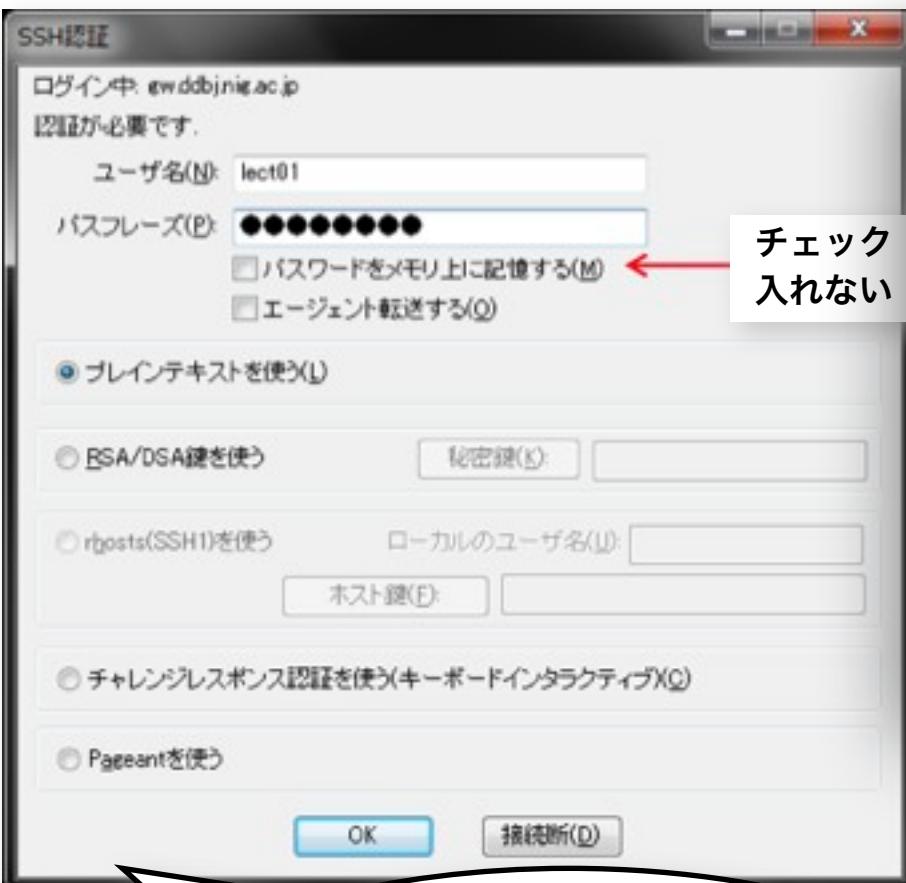
- Macから
 - ターミナル 起動



ssh ユーザ名@gw.ddbj.nig.ac.jp
 ⇒ **Enter**

ログイン (2)

- Windowsから



ユーザ名とパスワードを入力
⇒ OK

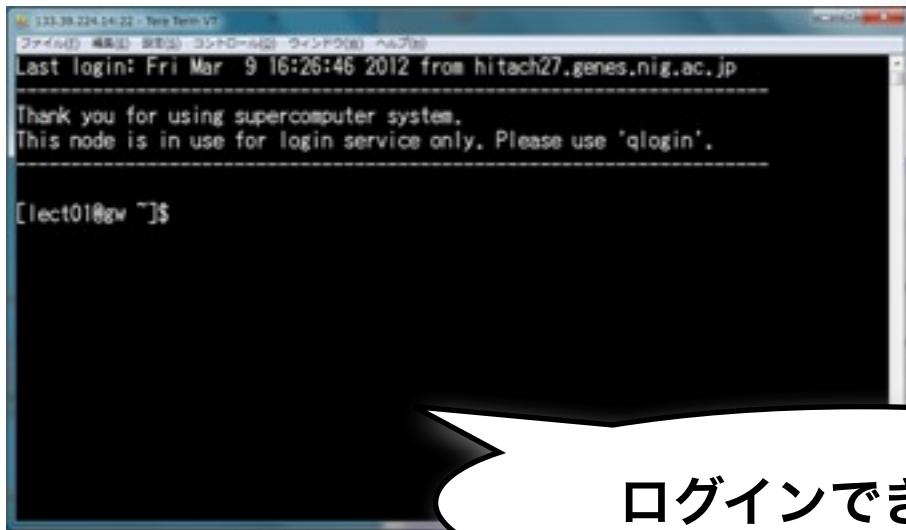
- Macから



パスワードを入力
⇒ Enter

ログイン (3)

• Windowsから



```
133.39.234.146.22 - Test Term V7
[ファイル] [編集] [表示] [コントロール] [ライントラッカー] [ヘルプ]
Last login: Fri Mar  9 16:26:46 2012 from hitach27.genes.nig.ac.jp
Thank you for using supercomputer system.
This node is in use for login service only. Please use 'qlogin'.
[lect01@gw ~]$
```

• Macから



```
yn — yanakamu@t351:~ — bash — 72x30
>Last login: Mon Nov 25 17:26:18 on ttys000
ynAir:- yn$ ssh yanakamu@gw.ddbj.nig.ac.jp
yanakamu@gw.ddbj.nig.ac.jp's password:
Last login: Mon Nov 25 18:17:43 2013 from 133.39.20.15
Thank you for using supercomputer system.
This node is in use for login service only. Please use 'qlogin'.
[yanakamu@gw ~]$
```

ログインできました！

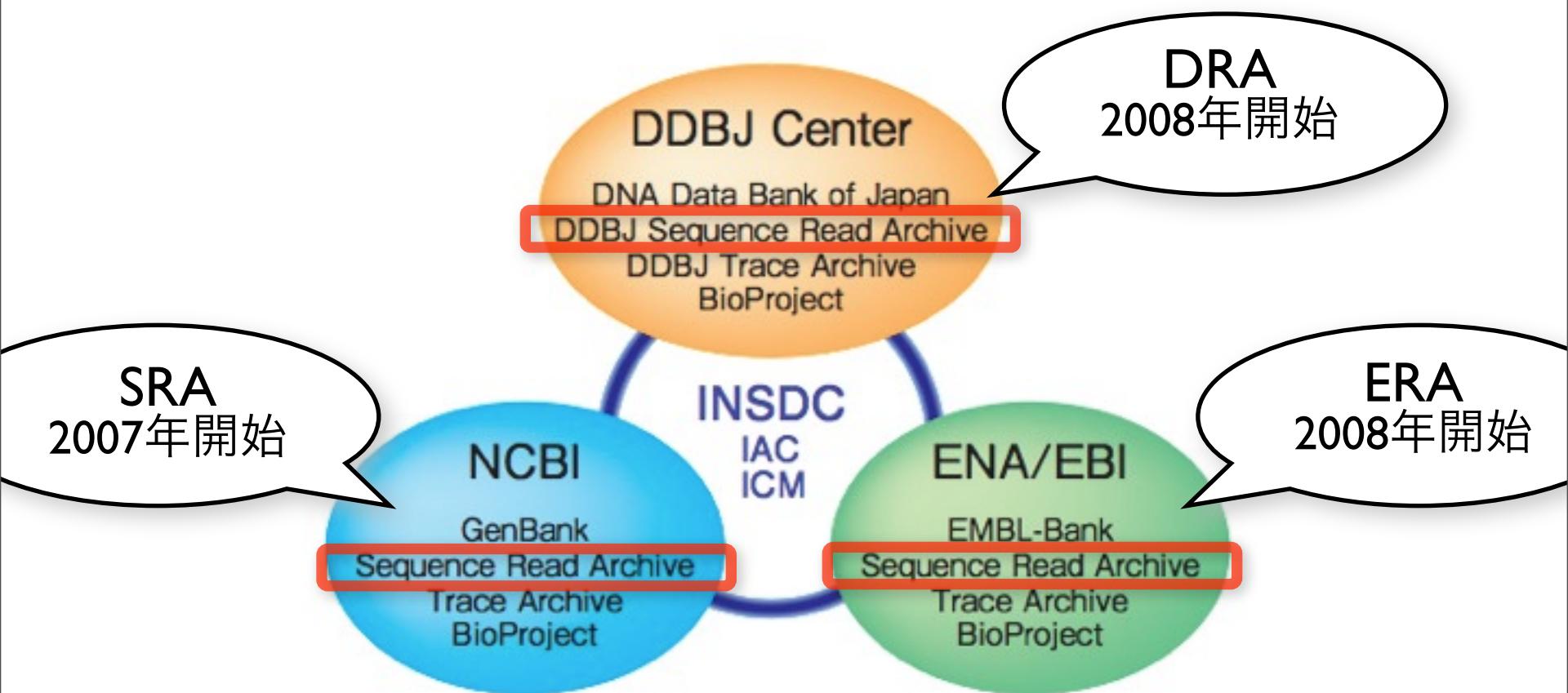
パスワードを4回間違えると
アカウントロックされます

ロックされた場合には
sc-info@nig.ac.jp まで

DRA

DDBJ Sequence Read Archive

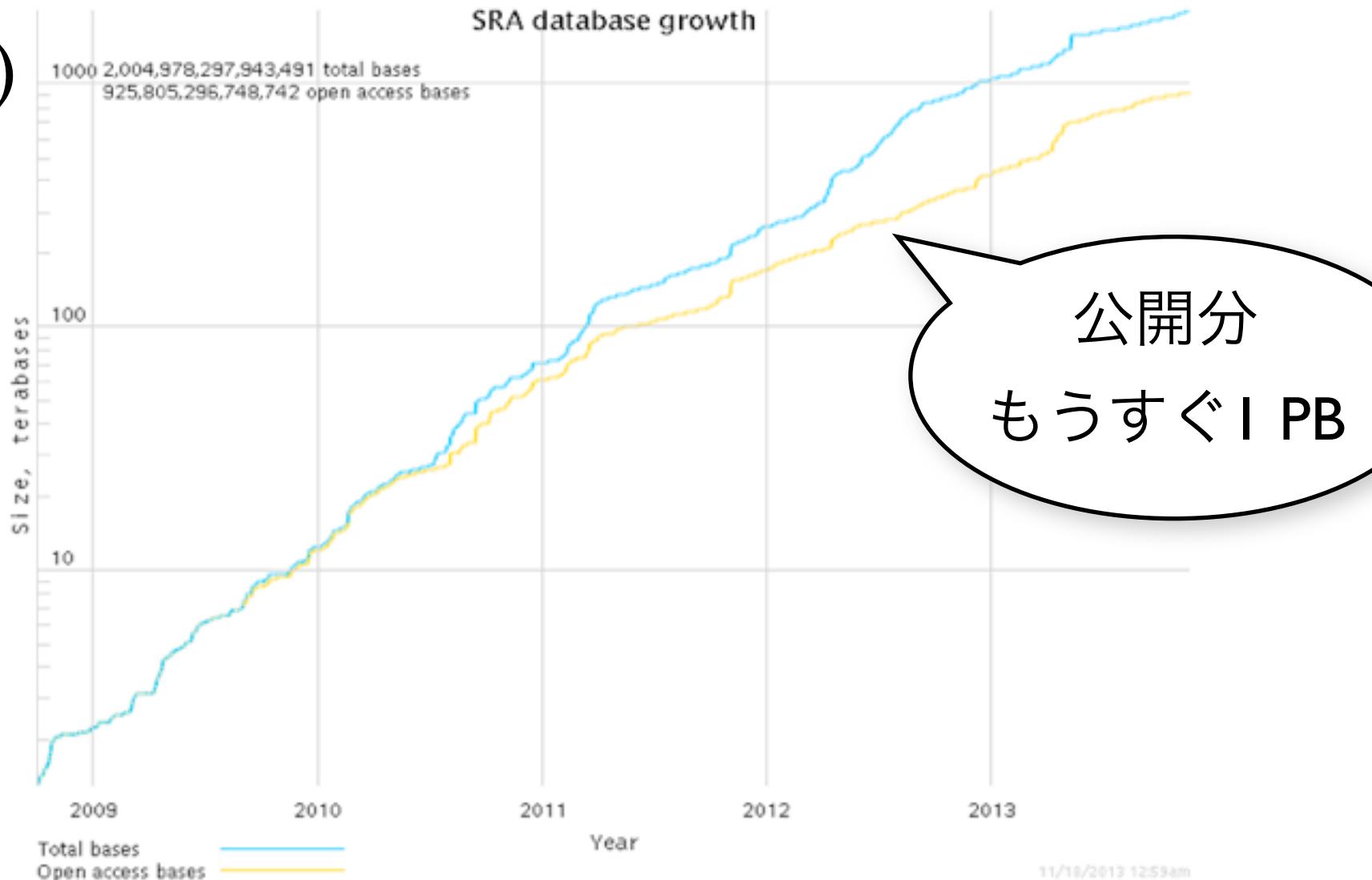
新世代シークエンサから出力される配列や
アライメントデータを登録・公開



SRA growth (NCBI)

<http://trace.ncbi.nlm.nih.gov/Traces/sra>

(TB)



DRAウェブサイト ⇒ [DRA] で検索

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

登録関係情報

解析パイプライン

データ検索

データ取得

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を登録するデータベースです。DRA は, International Nucleotide Sequence Database Collaboration (INSDC) のメンバーであり, DRA は, Sequence Read Archive (ERA) との国際協力のもと, 運営されています。従来のキャビラリーセquencing データを登録する場合は, DDBJ Sequence Read 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見る

公開データの DRA Search での検索

公開データは EBI SRA / NCBI SRA と共有されています

The screenshot shows the DRA Search interface. At the top, there is a search bar with fields for Accession (DRA000001), Organism (Abiotrophia defectiva ATCC 49176), StudyType (Epigenetics), and Platform (ILLUMINA). Below the search bar are buttons for 'Show 20 records', 'Sort by Study', 'Search', and 'Clear'. To the right of the search bar, a black oval highlights the search bar area with the text '生物名 etc での絞り込み' (Filtering by organism name etc.). A large blue arrow points down to the search results table. To the right of the arrow, another black oval highlights the results table with the text '検索結果リスト' (Search results list). The results table is titled 'Search Results (358 studies)' and shows a list of studies with columns for #, STUDY, SUBMISSION, STUDY_TITLE, STUDY_TYPE, ORGANISM, and CENTER_NAME. The first study in the list is highlighted with a yellow background and the text 'DRP000003'. To the right of the results table, a navigation panel is visible with sections for 'Submission (DRA000001)', 'Experiment (DRP000003)', 'Sample (DRP000003)', and 'Run (DRP000003)'. A black oval highlights this panel with the text 'ダウンロード' (Download). At the bottom right, a black oval highlights the 'Project ID' and 'Center Name' fields with the text '詳細 (メタデータ記述)' (Details (Metadata description)).

Type	Count
Submission	23770
Study	3423
Experiment	29624
Sample	111241
Run	71620

#	Organism Name	Study	#
1	Homo sapiens	293	1
2	metagenome sequence	169	2
3	Mus musculus	163	3
4	Drosophila melanogaster	121	4
5	marine metagenome	79	5
6	Caenorhabditis elegans	39	6
7	Arabidopsis thaliana	38	7
8	synthetic construct	37	8
9	Saccharomyces cerevisiae	35	9
10	Panicum virgatum	21	10

#	STUDY	SUBMISSION	STUDY_TITLE	STUDY_TYPE	ORGANISM	CENTER_NAME
1	DRP000003	DRA000003	Comprehensive identification and characterization of the nucleosome structure	Transcriptome Analysis	Homo sapiens	UT-MGS
2	DRP000004	DRA000004	Comprehensive identification and characterization of the transcripts, their expression level	Transcriptome Analysis	Homo sapiens	UT-MGS
3	DRP000005	DRA000005	Comprehensive characterization of the expression level	Transcriptome Analysis	Homo sapiens	UT-MGS
4	DRP000006	DRA000006	Comprehensive characterization of the expression level	Transcriptome Analysis	Homo sapiens	UT-MGS
5	DRP000007	DRA000007	Comprehensive characterization of the expression level	Transcriptome Analysis	Homo sapiens	UT-MGS
6	DRP000008	DRA000008	Comprehensive characterization of the expression level	Transcriptome Analysis	Homo sapiens	UT-MGS

DRP000003

Study Detail	
Title	Comprehensive identification and characterization of the nucleosome structure
Abstract	Comprehensive identification and characterization of the nucleosome structure in mammalian genes were attempted. We used Nucleosome-Seq method, in which next generation sequencing technology and micrococcal nuclease digestion assay were combined.
Description	Although recent studies have revealed that the majority of human genes are subjected to regulation of alternative promoters (APs), the biological relevance of this phenomenon remains unclear. In order to understand biological significance of the presence of diverg ... [more]
Project ID	94559
Center Name	UT-MGS (University of Tokyo, Medical Genome Sciences)

Navigation

- Submission (DRA000001)
- Experiment (DRP000003)
- Sample (DRP000003)
- Run (DRP000003)

詳細 (メタデータ記述)

FASTQ フォーマット

- テキストベースの形式で、DNAなどの塩基配列とそのクオリティスコアを1つのファイルに一緒に保存する

```
@ERR171441.2 HWI-962:55:COA4UACXX:4:1101:1649:2314 length=88
TATCCTGGTCGGCTTGCAGGACGCCATCGAGGCAGAACTCANNNNNTGCACGTGACACCT
GGCCCGCGCCTGCGTGTGCATCATGCG
+ERR171441.2 HWI-962:55:COA4UACXX:4:1101:1649:2314 length=88
HFHFHIIJHIHJGIJJEFGHBDHGGIHHIEHJHGH;CDHHI#####,,;AAEDDDBBDDDD
DDDDDDDBBBB@BDD<BBB>@CCCCD@>
```

1本の配列は4行で記述される。1行目は文字「@」で始まり、その後ろに配列のIDと、オプションとして説明を記述する。2行目は塩基配列を記述する。3行目には文字「+」を記載する。またその後ろに配列のIDを記載することもある。4行目には2行目に記述した配列のクオリティ値を記述する。このクオリティ値は2行目の配列と同じ文字数でなければならない（例はERR171441の先頭4行）

参照：<http://ja.wikipedia.org/wiki/Fastq>

解析パイプラインも提供しています

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

 [DDBJ](#) | DNA Data Bank of Japan

[Login & Submit](#) | [Databases](#) | [English](#) | [Contact](#)

[Google™ カスタム検索](#) 

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。従来のキャビラリ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見る

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

DRA pipeline: ソフトウェア

DDBJ (DNA Data Bank of Japan)

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 Genome

Running Status

Selecting Tools for Basic Analysis of DDBJ ANNOTATION PIPELINE

Reference Genome Mapping

Tool	Help	Version	Input data			Evaluation		Error rate	SNP	Indel	.gff	.bed	SAM	Comment
			Base space	Color space	Paired end	Depth	Coverage							
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 severely recommend when performing Velvet, total length of those reads is up to 22G bp. Maximum K-mer value is 64.
		r2013-02-					

DRA 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

Specifying Database of Reference

ACCOUNT

login ID [yaskaz]

Logout Change password

ANALYSIS

Data setup DRA Start FTP upload

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

ChIP-seq

JOB STATUS

step1. Preprocessing

step1. Mapping

step1. de novo Assembly

step2-All status

HELP

Select Query Files → Select Tools → Set QuerySet →

Running Status

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 mitochondrial, tigr chloroplast

all check all

all.fa chr01.fa chr02.fa chr03.fa chr04.fa chr05.fa chrC.fa chrM.fa

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

all check all

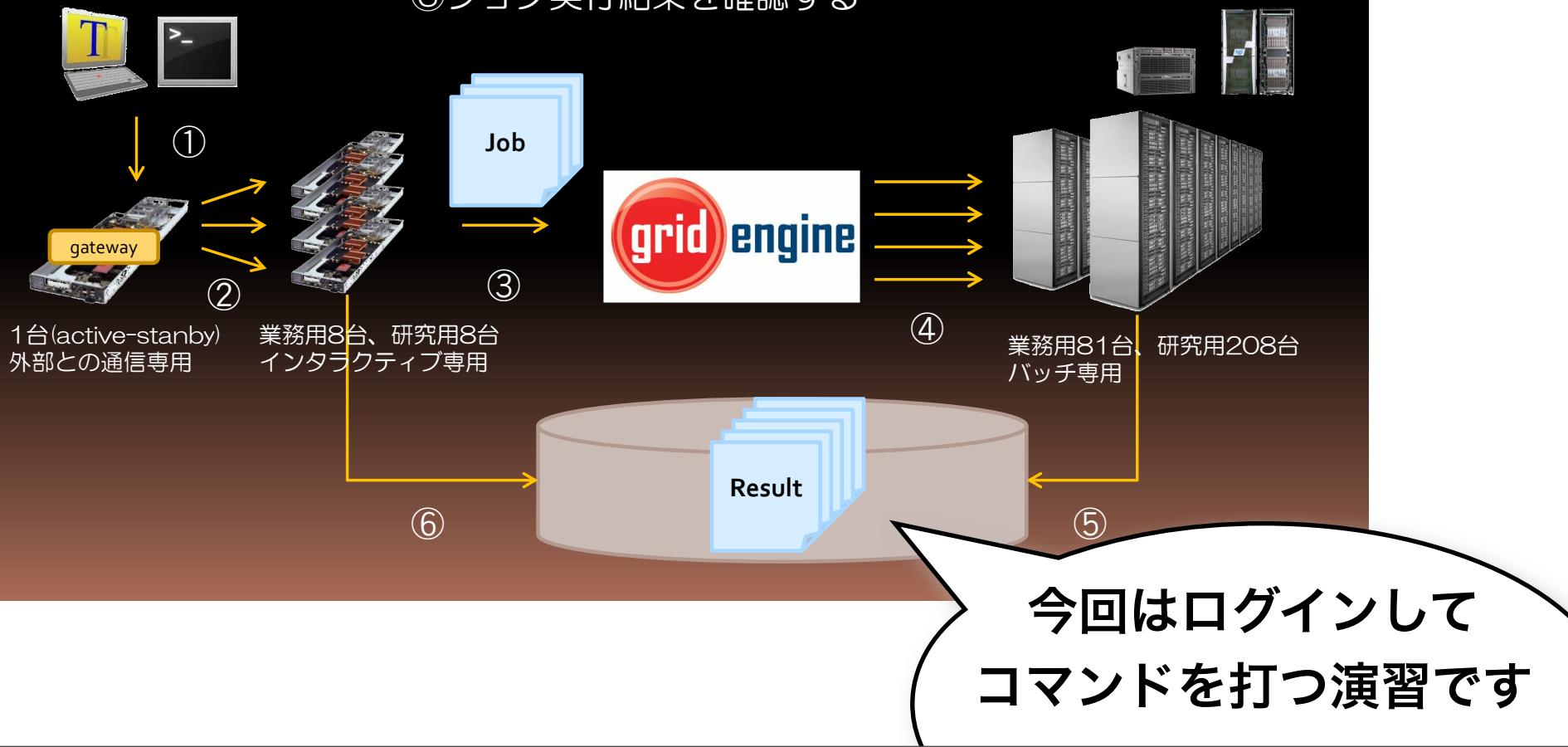
all.fa chr01.fa chr02.fa chr03.fa chr04.fa chr05.fa chrC.fa chrM.fa

User original sets

Download or upload reference

スパコン使用方法(イメージ)

- ①ゲートウェイノード(gw.ddbj.nig.ac.jp)にログインする
- ②qloginを実行しインタラクティブノードにログインする
- ③qloginしたホストからジョブをUGEに投入する
- ④UGEは負荷の低いノードでジョブを実行する
- ⑤ジョブ実行結果をlustreのホームディレクトリに出力する
- ⑥ジョブ実行結果を確認する



あわせてご参照ください

- 全般的な操作方法：とくにデータ転送方法などについてこちらをご参照ください（過去の講習会資料）
 - http://sc.ddbj.nig.ac.jp/images/stories/meetingdoc/20120510/ja/ja_Basic_usage-1.pdf
- Unix / Linux の生い立ち・思想
 - UNIXという考え方—その設計思想と哲学 (www.amazon.co.jp/dp/4274064069/)
 - それがぼくには楽しかったから (www.amazon.co.jp/dp/4274064069/)
 - ハッカーズ (www.amazon.co.jp/dp/487593100X)