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

自己紹介

中村保一　検索

@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

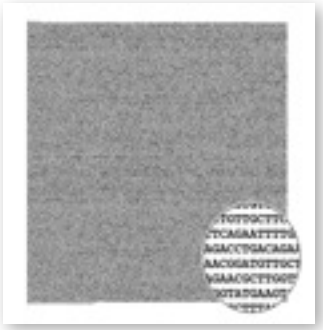
http://genome.kazusa.or.jp/cyanobase/

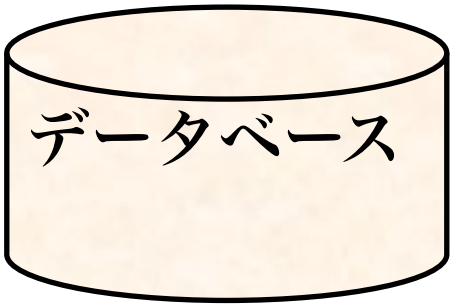
光合成細菌のゲノム解析＋データベース。Social Bookmark による遺伝子注釈系

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

塩基配列データバンクとはこのような事業

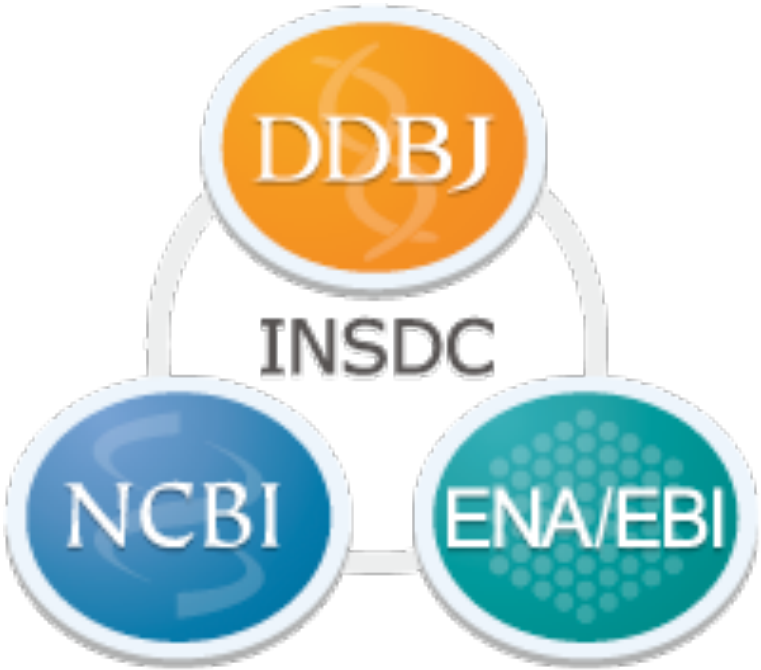
- 全世界で解読された塩基配列情報を
  - 査定して受入れ
  - データベースに蓄積し
  - 公開して共有する





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

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

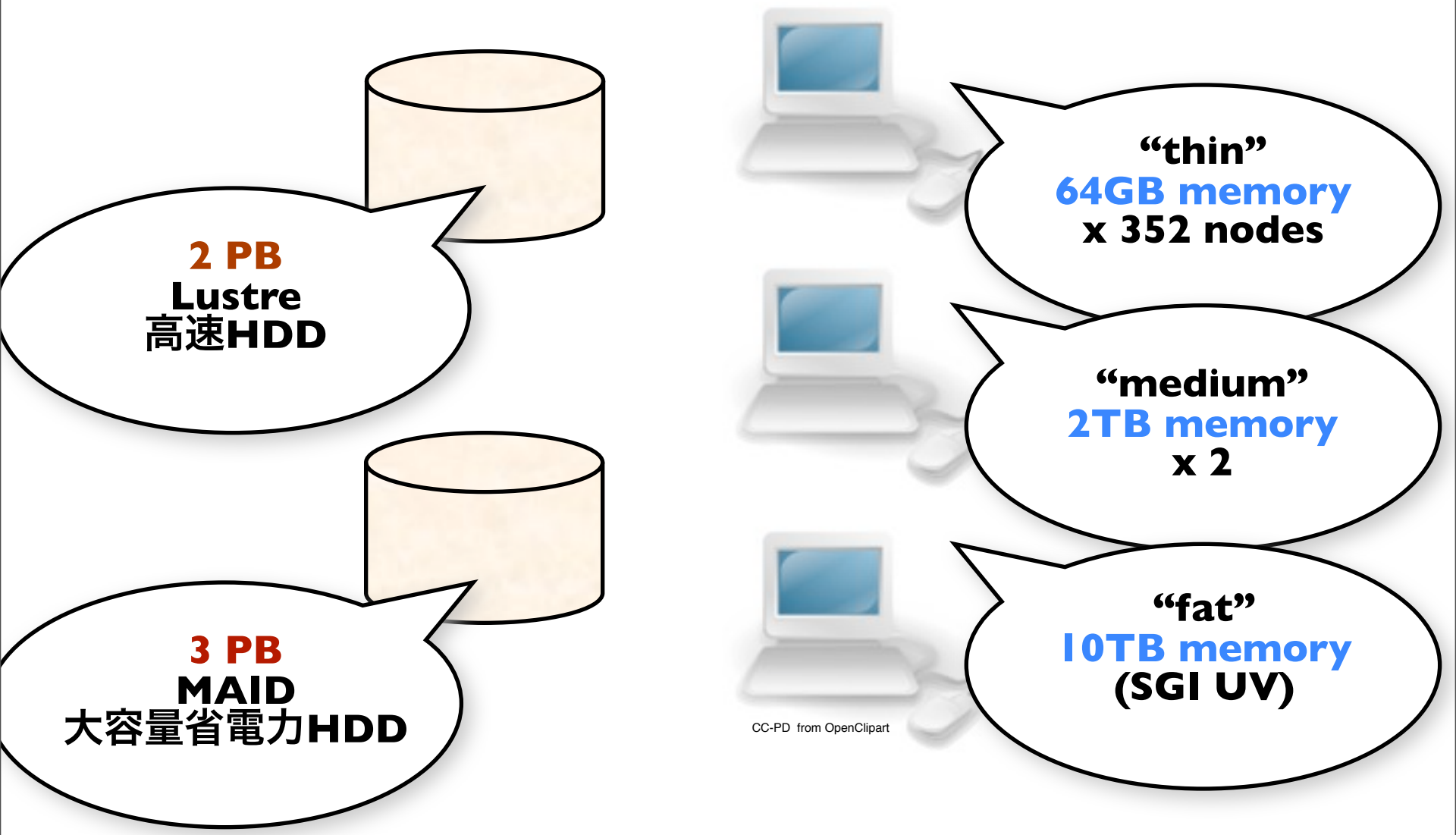
GenBank (from Release note 195)
Mark Cavanaugh, Ilene Mizrachi, Yiming Bao, Michael Baxter, Lori Black, Larissa Brown, Vincent Calhoun, Larry Chlumsky, Karen Clark, Jianli Dai, Michel Eschenbrenner, Irene Fang, Michael Fetchko, Linda Frisse, Andrea Gocke, Anjanette Johnston, Mark Landree, Jason Lowry,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

遺伝研スパコン


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
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
DDBJ
INSDC
NCBI
ENA/EBI
International Nucleotide Sequence Database Collaboration
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
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

遺伝研スーパーコンピュータ

DDBJ・新スパコン概要 (2012.3)



Unix Operating System

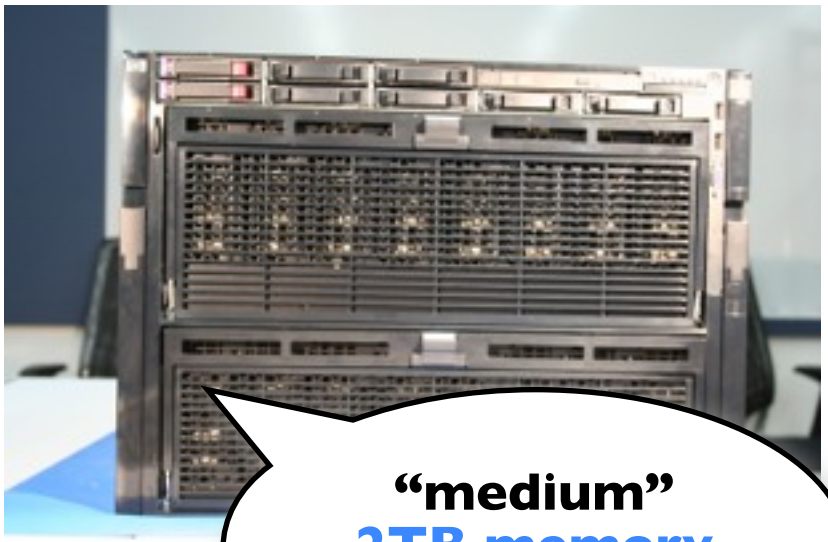
Bell 研の Ken Thompson, Dennis Ritchieらが “Space Travel” で遊ぶために部屋のスミに置いてあった借り物の PDP-7 で「えいやっ」と作ってみた「小さくて」「軽い」オペレーティングシステム (1968年)

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

DDBJ
DNA Data Bank of Japan

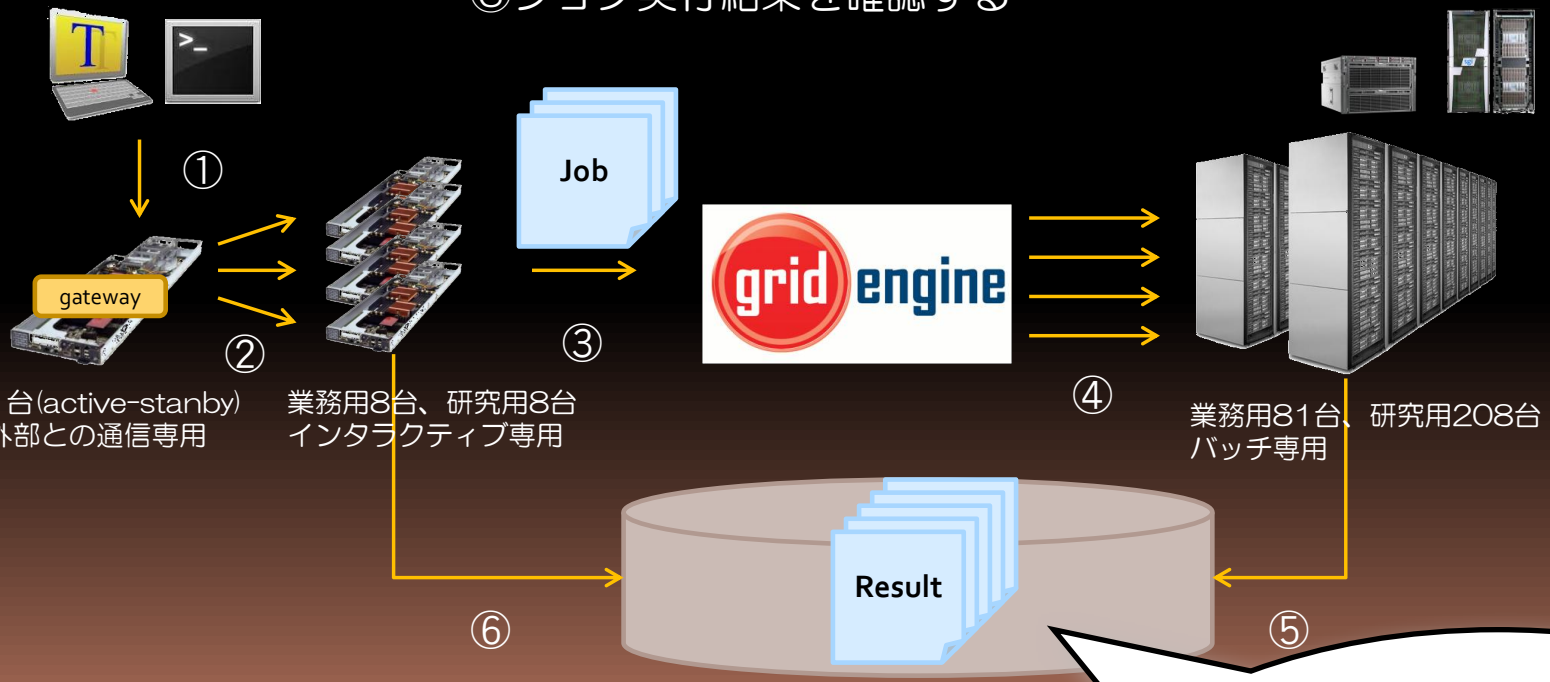


"medium"
2TB memory
x 2

"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



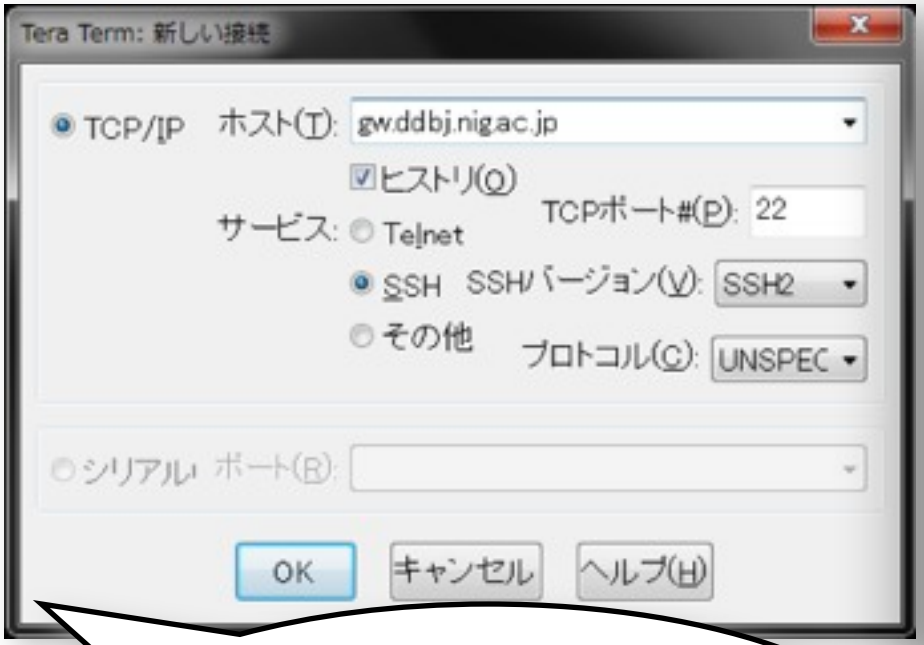
バッチジョブ投入⇒
キューの順番で処理

ログインして

みましょう

ログイン (1)

- Windowsから
  - Tera Term 起動



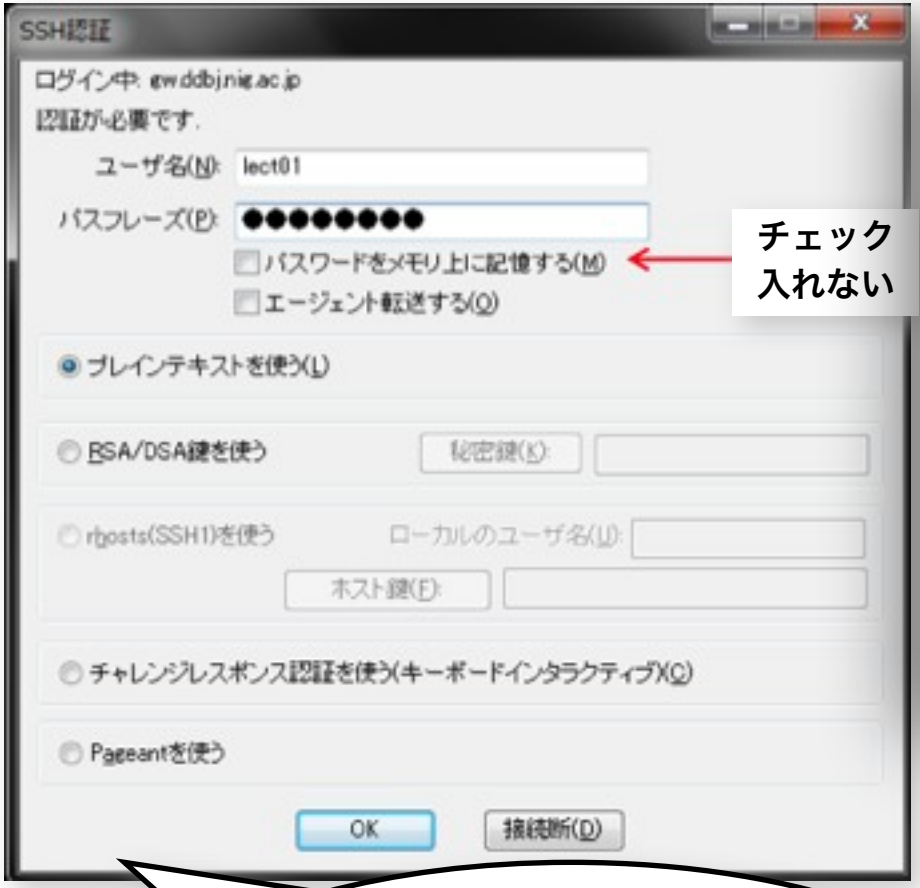
ホスト：**gw.ddbj.nig.ac.jp**
サービス：**SSH**
⇒ **OK**

- 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
- Macから



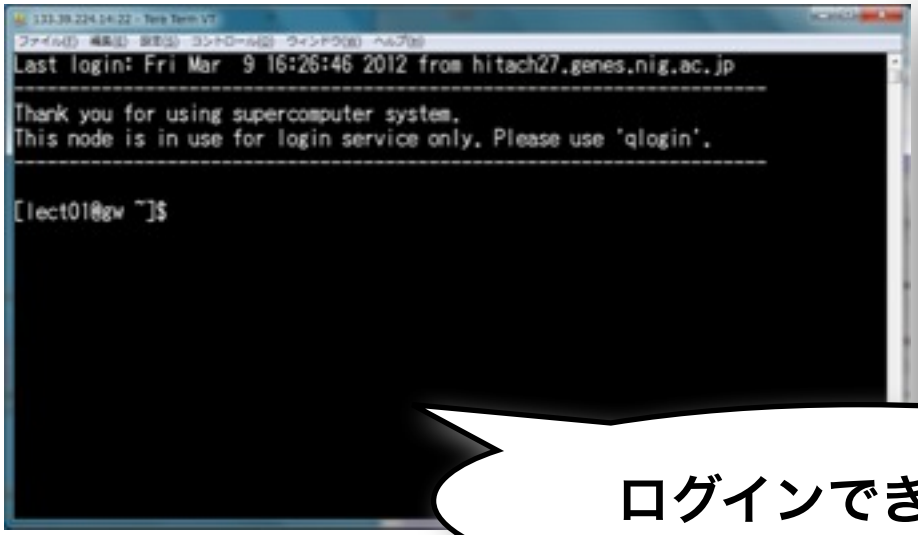
チェック
入れない

ユーザ名とパスワードを入力
⇒ **OK**

パスワードを入力
⇒ **Enter**

ログイン (3)

- Windowsから
- Macから



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

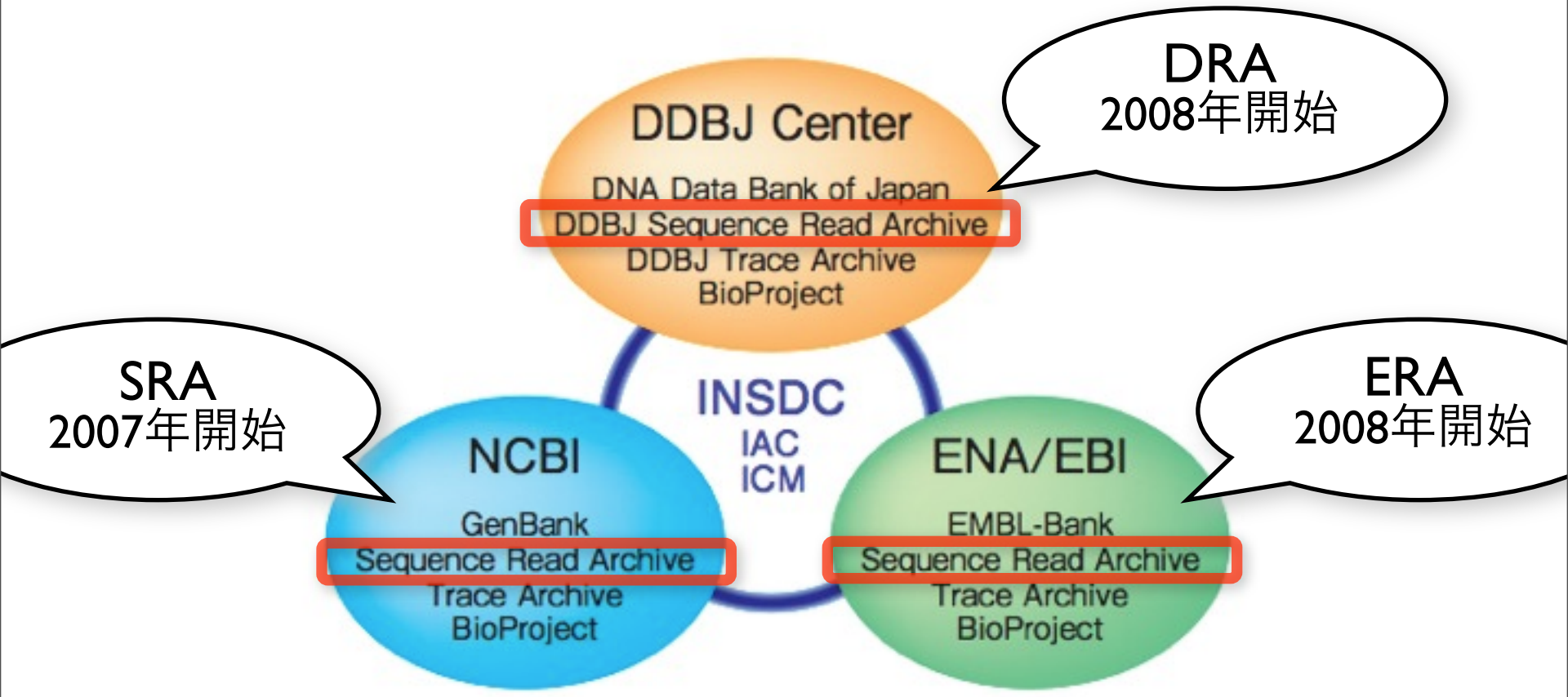
ロックされた場合には
sc-info@nig.ac.jp まで

DRA

DDBJ Sequence Read Archive

DDBJ Sequence Read Archive (DRA)

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

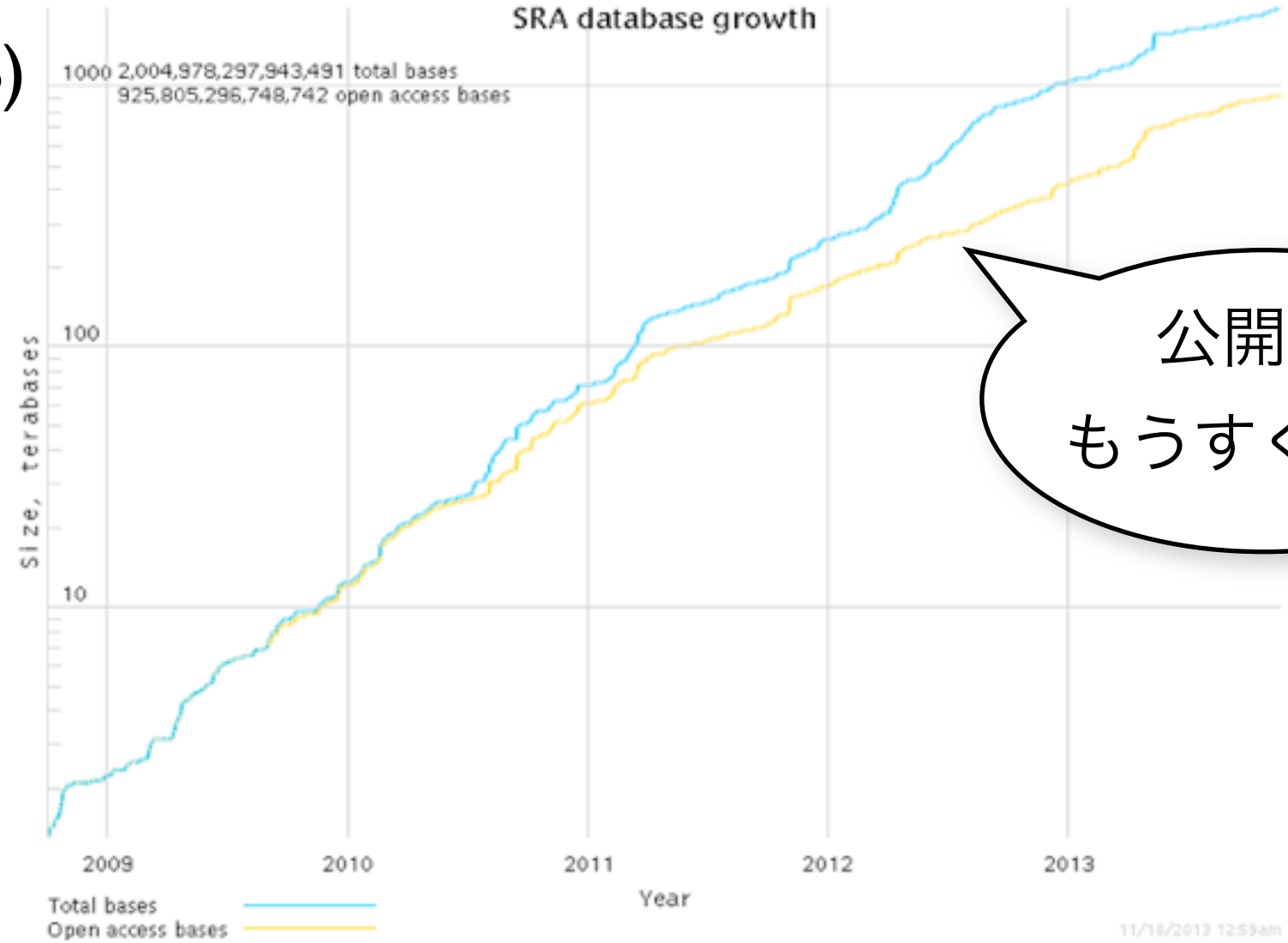


International Nucleotide Sequence Databank Collaboration

SRA growth (NCBI)

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

(TB)



公開分
もうすぐ1 PB

DRAウェブサイト ⇒ [DRA] で検索

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

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

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:C0A4UACXX:4:1101:1649:2314 length=88
TATCCTGGTCGGCTTGCAGGACGCCATCGAGGCAGAACTCANNNNNTTGCACGTGACACCT
GGCCCGCGCCTGCGTGTGCATCATGCG
+ERR171441.2 HWI-962:55:C0A4UACXX:4:1101:1649:2314 length=88
HFHFHIIJHIHJGIJJEFGHBDHGGIHIIEHJHGH;CDHHI#####,,;AAEDDDBBDDDD
DDDDDDBBBB@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）

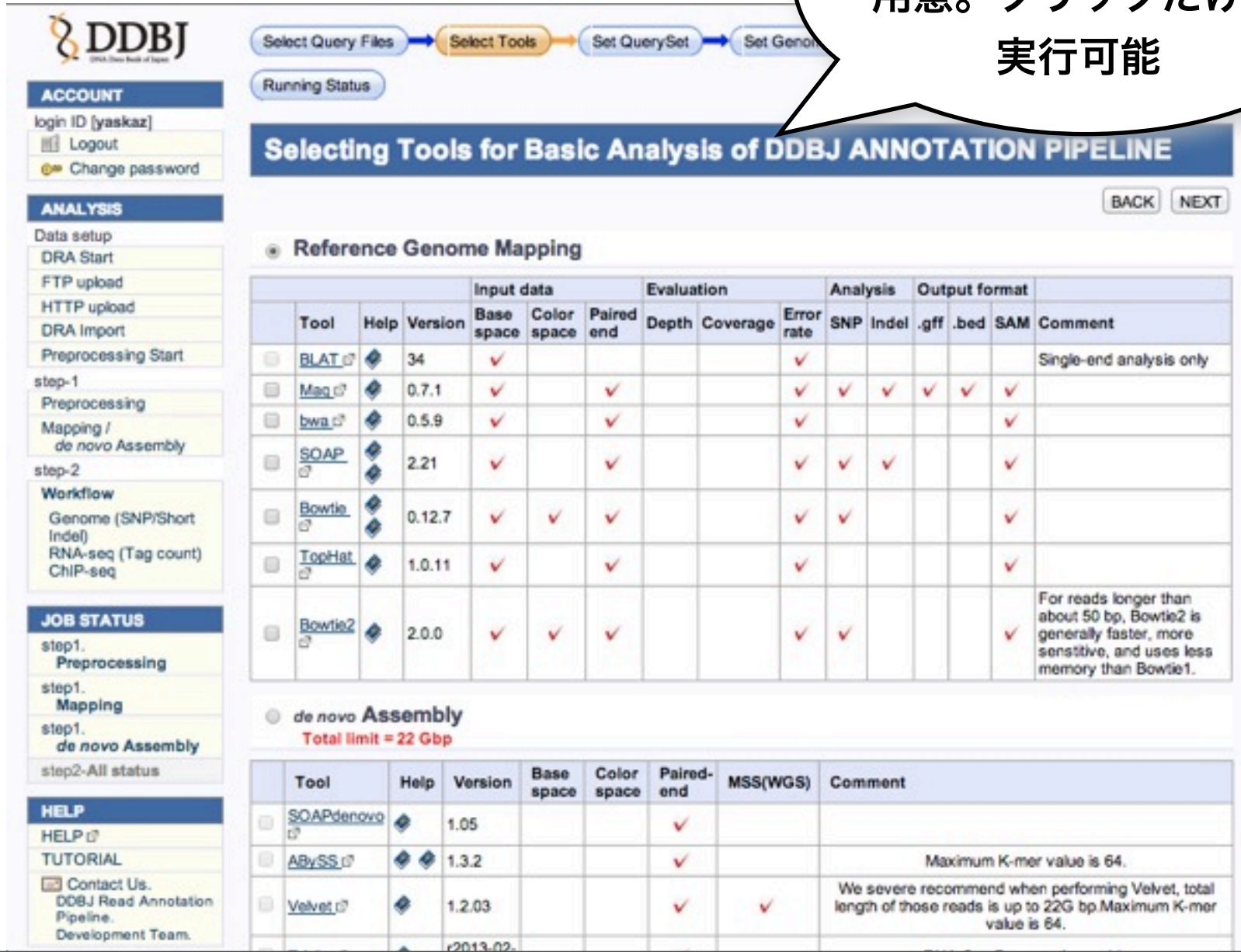
参照：http://ja.wikipedia.org/wiki/Fastq

解析パイプラインも提供してます

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

DRA 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



DDBJ

Select Query Files → Select Tools → Set QuerySet → Set Genom

Running Status

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.

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 senstitive, and uses less memory than Bowtie1.

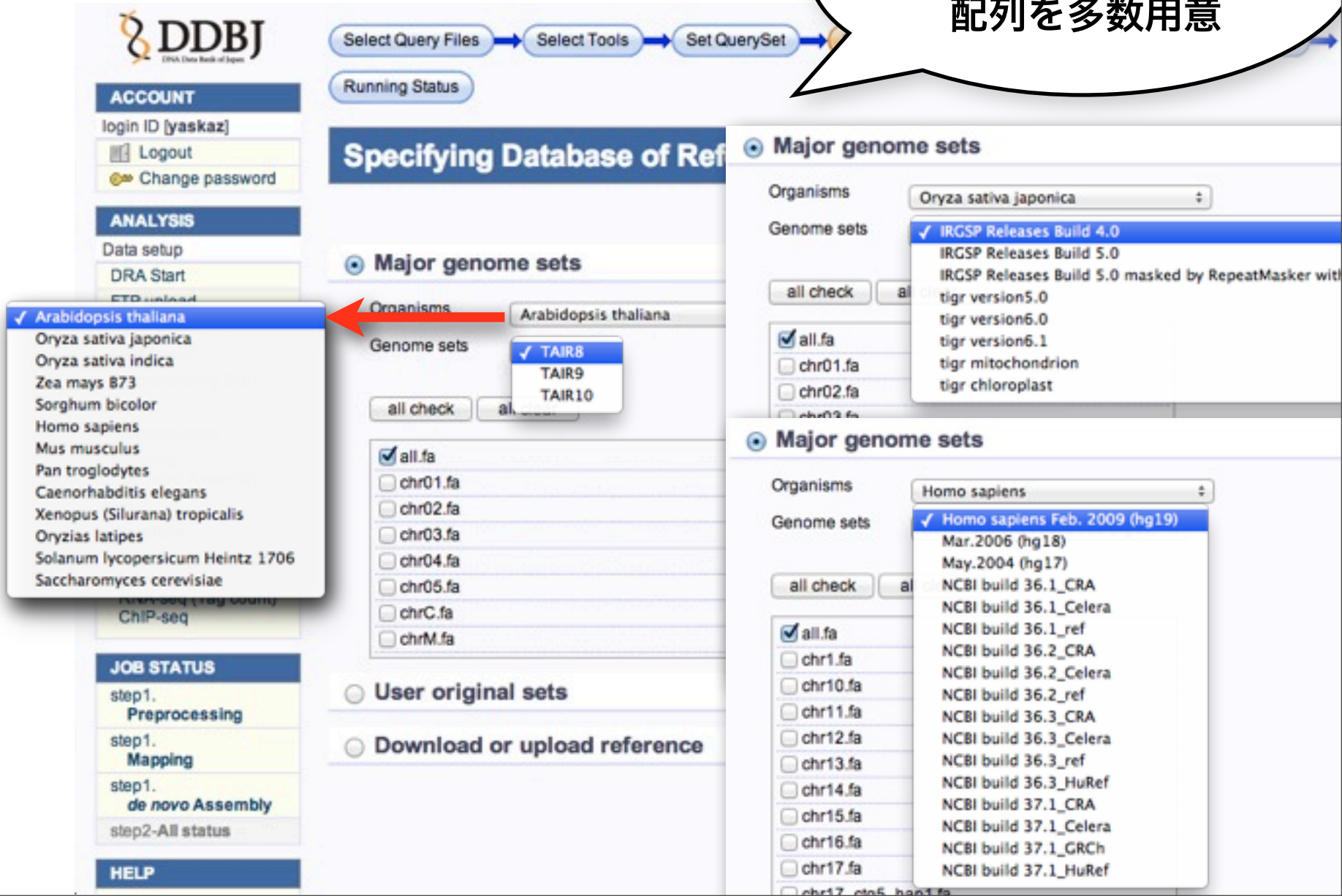
de novo Assembly

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.
			r2013-02-					

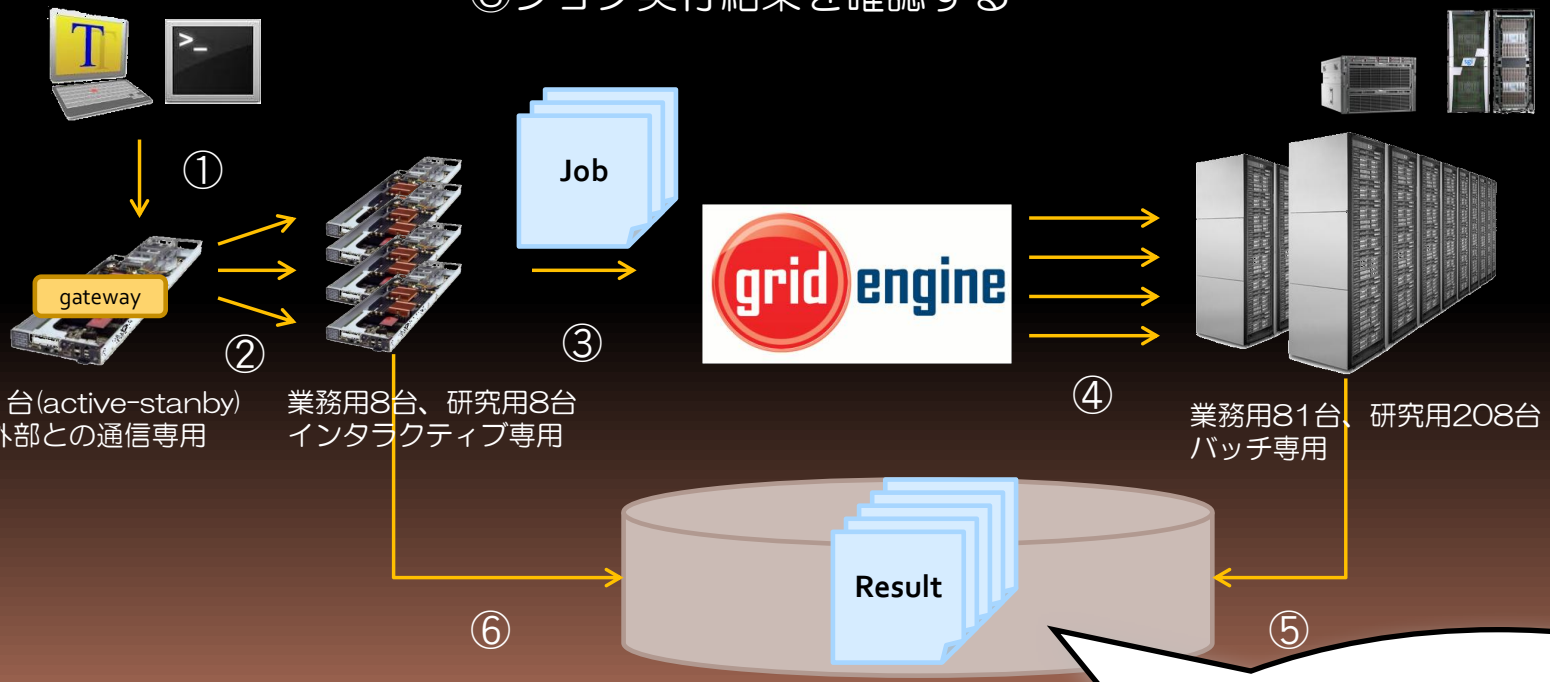
DRA pipeline: 比較対象

イネ、マウスなど
解析比較対象となる
配列を多数用意



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



今回はログインして
コマンドを打つ演習です

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)