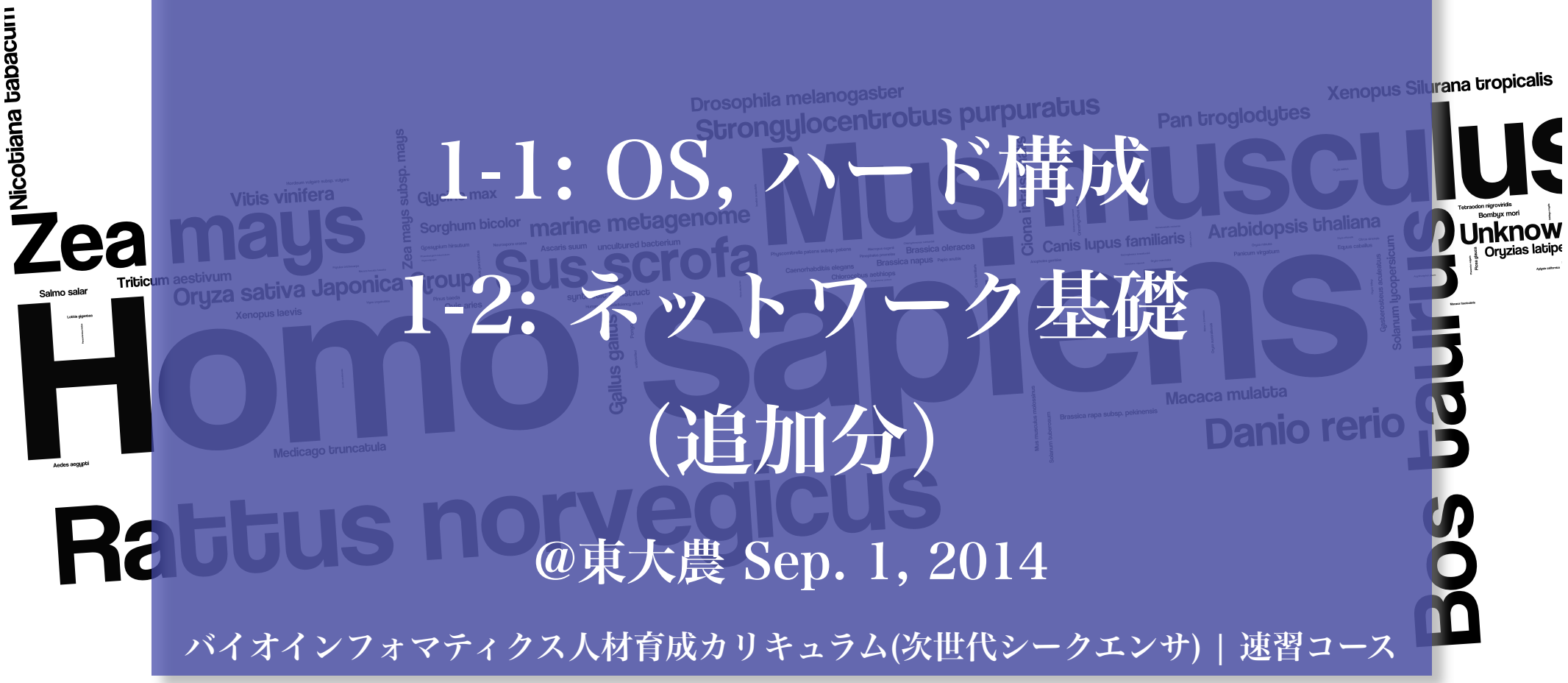


1-1: OS, ハード構成
1-2: ネットワーク基礎
（追加分）

@東大農 Sep. 1, 2014

バイオインフォマティクス人材育成カリキュラム(次世代シークエンサ) | 速習コース

中村 保一
NAKAMURA Yasukazu, Professor

国立遺伝学研究所 大量遺伝情報研究室

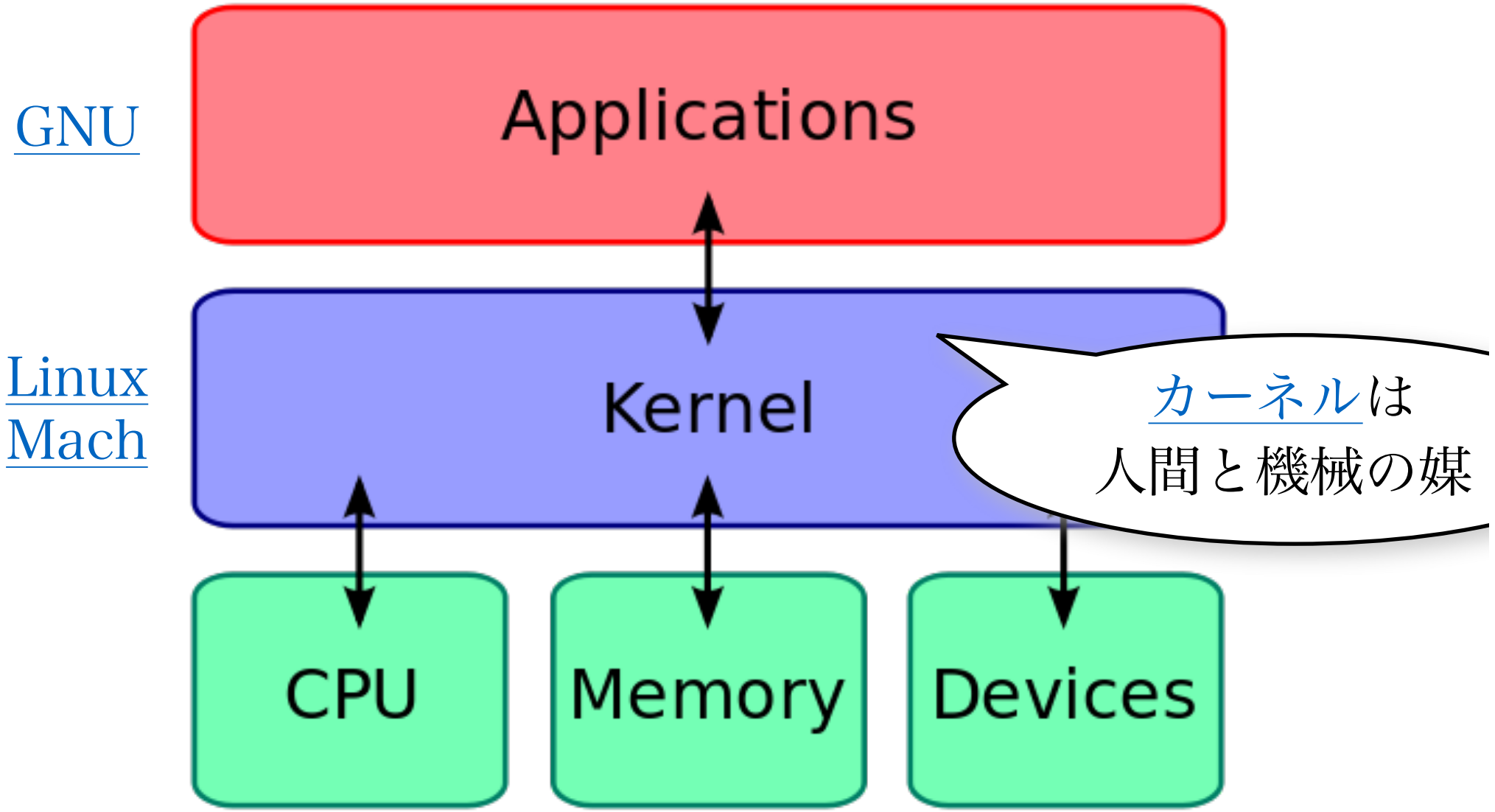
http://charles.genes.nig.ac.jp　yn@nig.ac.jp

Linux とは

- Linuxは、狭義ではLinuxカーネルを意味し、広義ではそれをカーネルとして用いたオペレーティングシステム (OS) を意味する。Linuxカーネルを用いたオペレーティングシステムは、Unix に類似した振る舞いをするので、Unix系のオペレーティングシステムの一種に分類される。

http://ja.wikipedia.org/wiki/Linux

UNIX の構成



http://commons.wikimedia.org/wiki/File:Kernel_Layout.svg

Linux カーネル

リーナス・ベネディクト・トーバルズ（Linus Benedict Torvalds; 1969-）

フィンランド、ヘルシンキ出身のプログラマ。Linuxカーネルを開発し、1991年に一般に公開した

- 1991年、ヘルシンキ大学の学生であったリーナス・トーバルズはオペレーティングシステムに好奇心を抱き Intel 80386 CPU を搭載した PC に UNIX 互換の独自のカーネルを開発した。最終的にこれが現在の Linux カーネルへと成長した。

http://ja.wikipedia.org/wiki/Linux

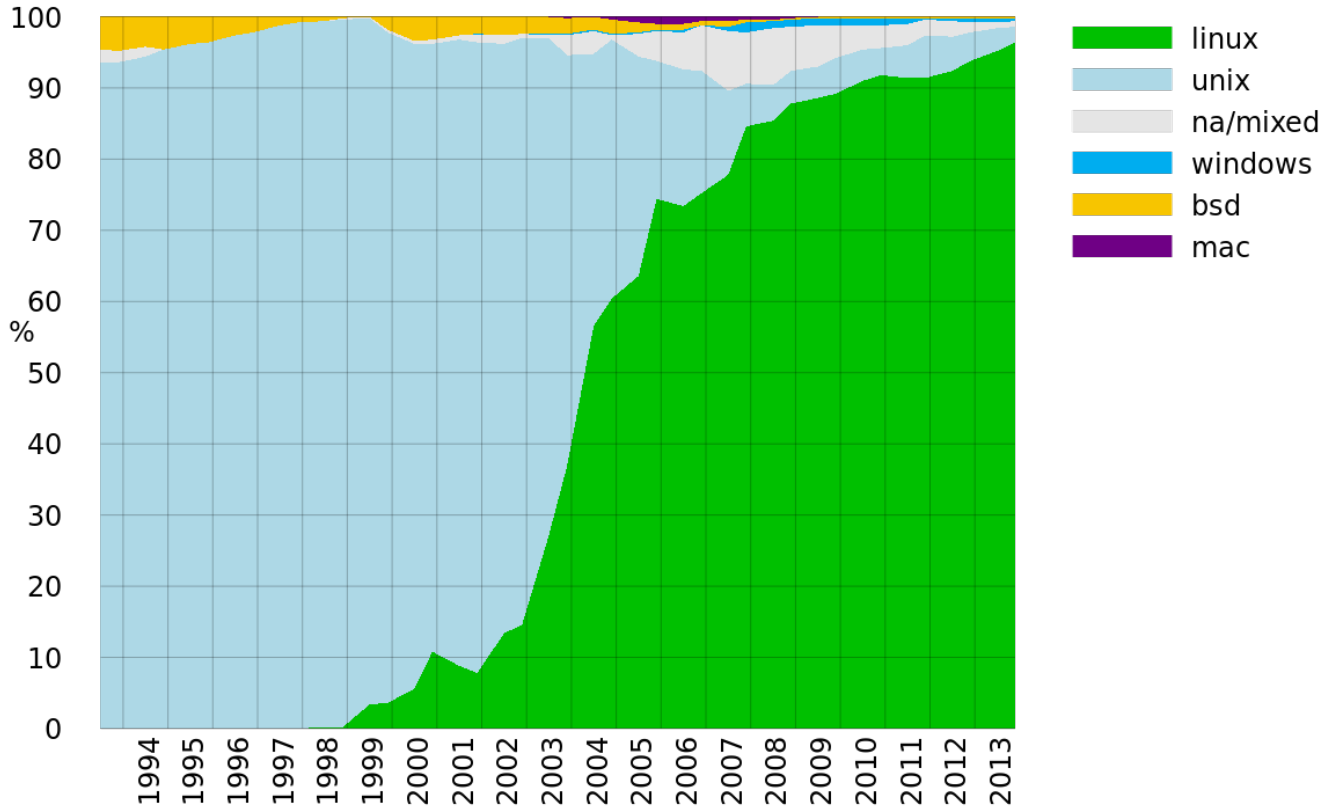
見えてるのは実は Linux でなくて GNU

DDBJ
DNA Data Bank of Japan

- GNU's Not Unix
- GNU は典型的には Linux と呼ばれるカーネルとともに使われます。この組み合わせが GNU/Linux オペレーティング・システムです。GNU/Linux は何百万人もの人に使われています。間違って“Linux”と呼ばれていますけれども。

http://www.gnu.org/home.ja.html

GNU/Linux



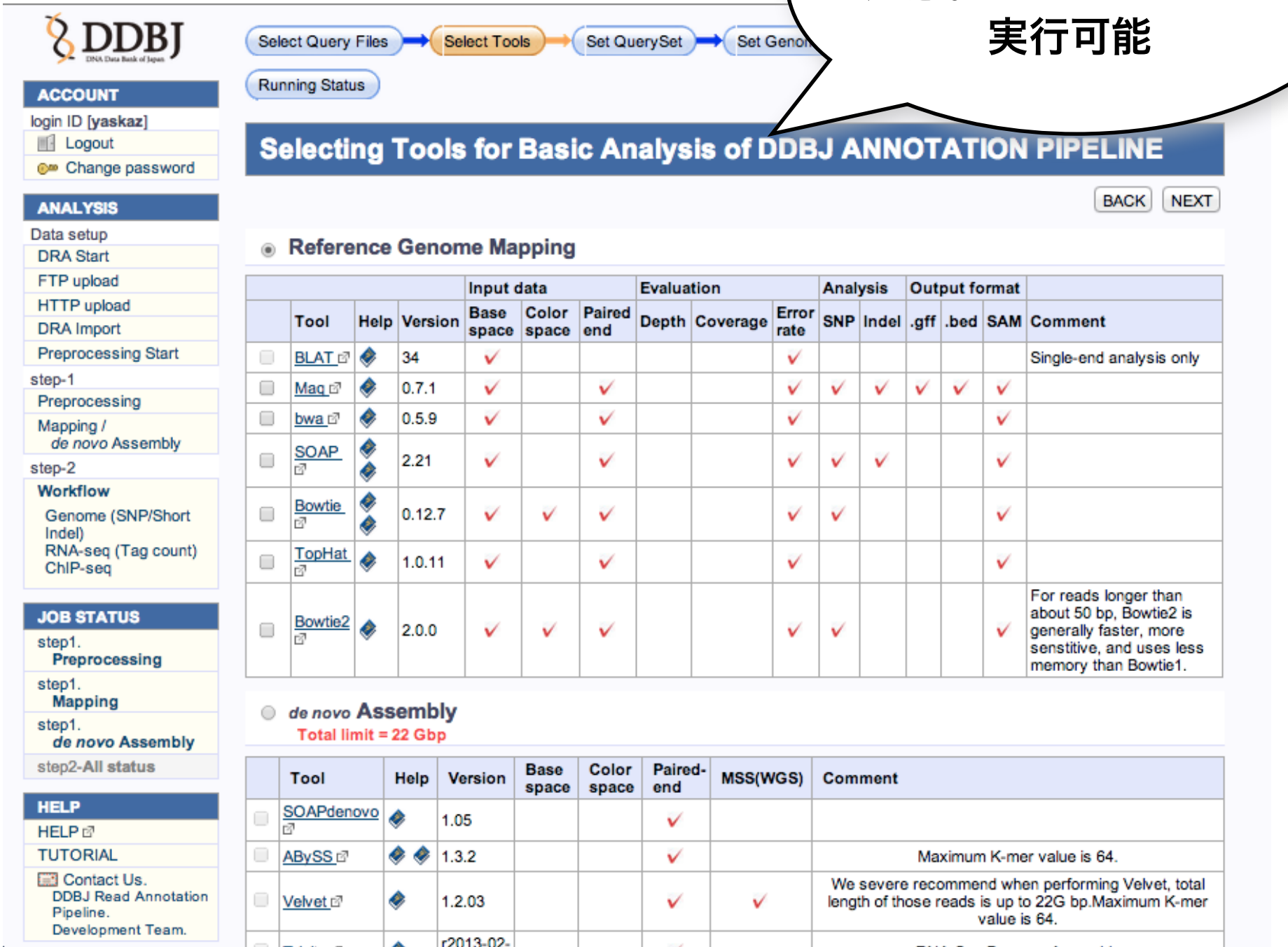
Operating systems used on top 500 supercomputers

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
DNA Data Bank of Japan

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.

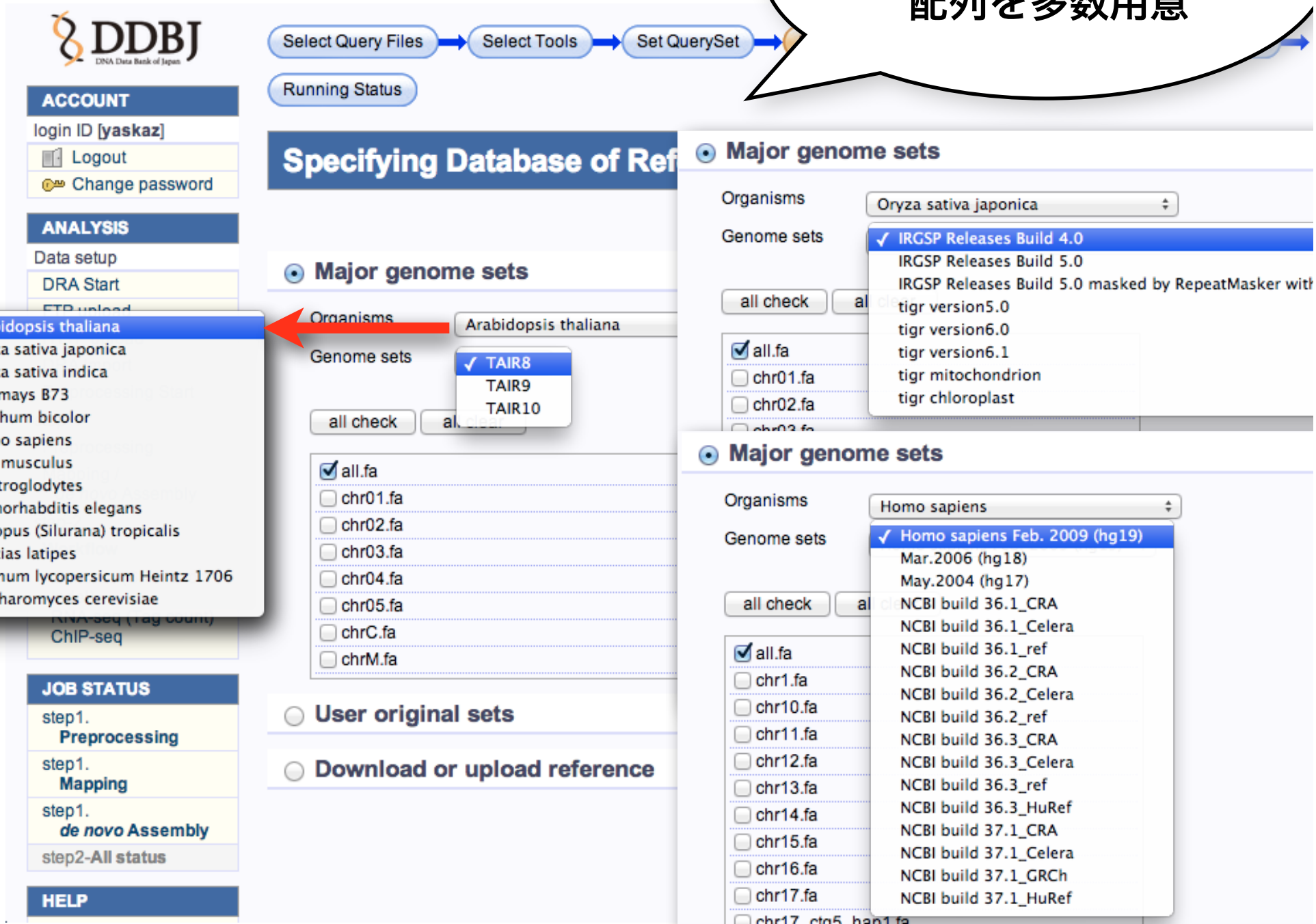
de novo Assembly

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-					

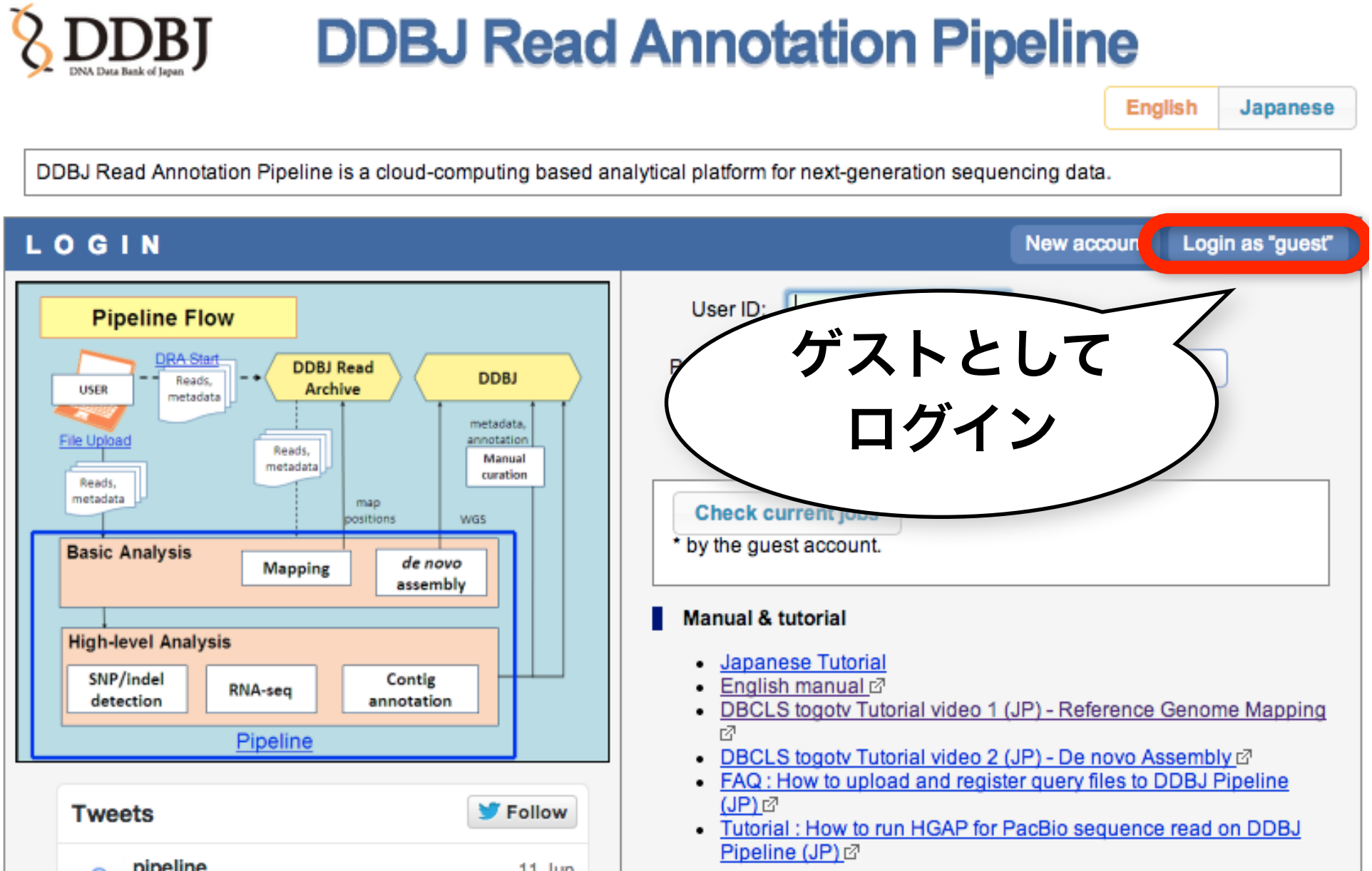
DRA pipeline: 比較対象

イネ、マウスなど
解析比較対象となる
配列を多数用意

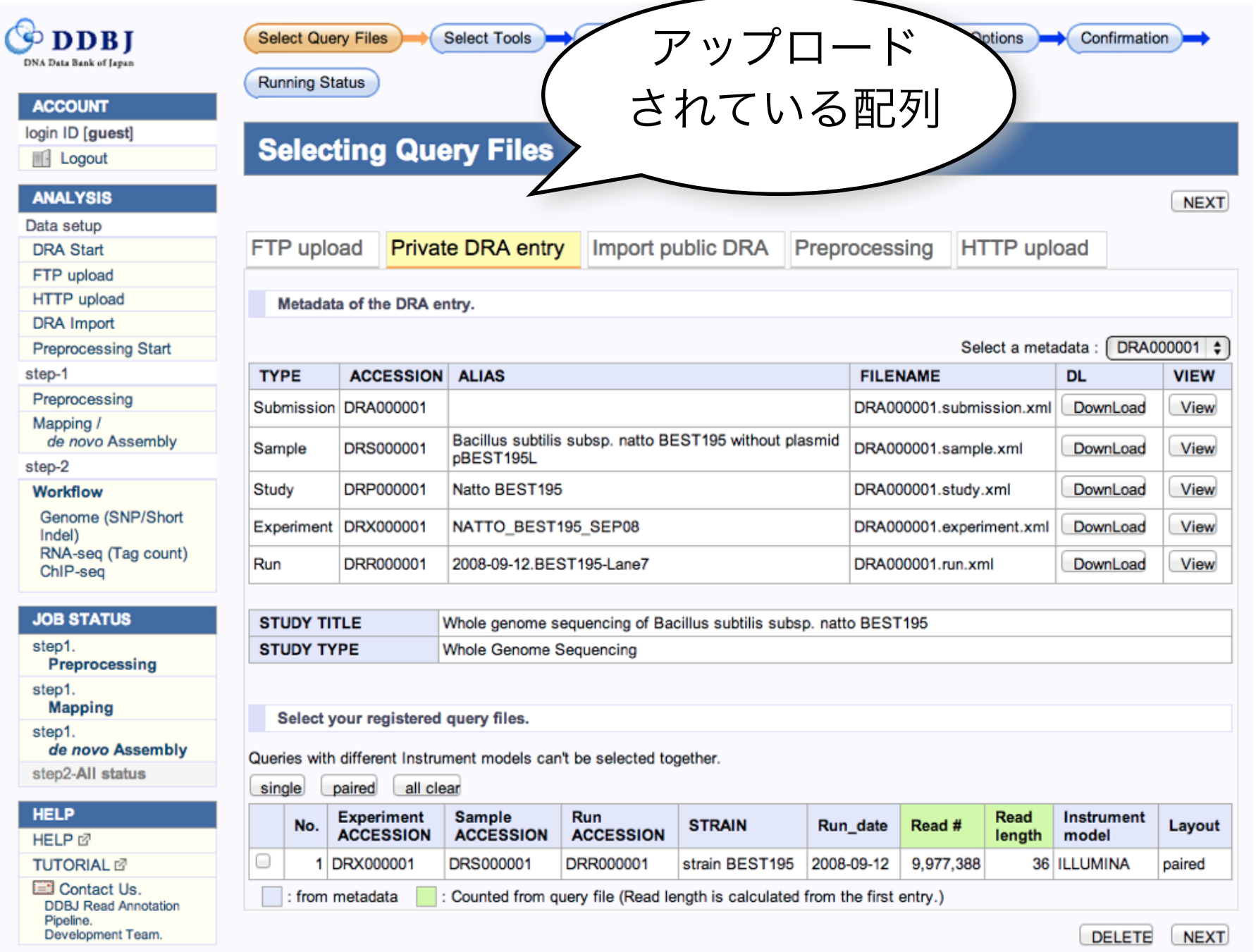


DDBJ パイプラインの解析の流れ

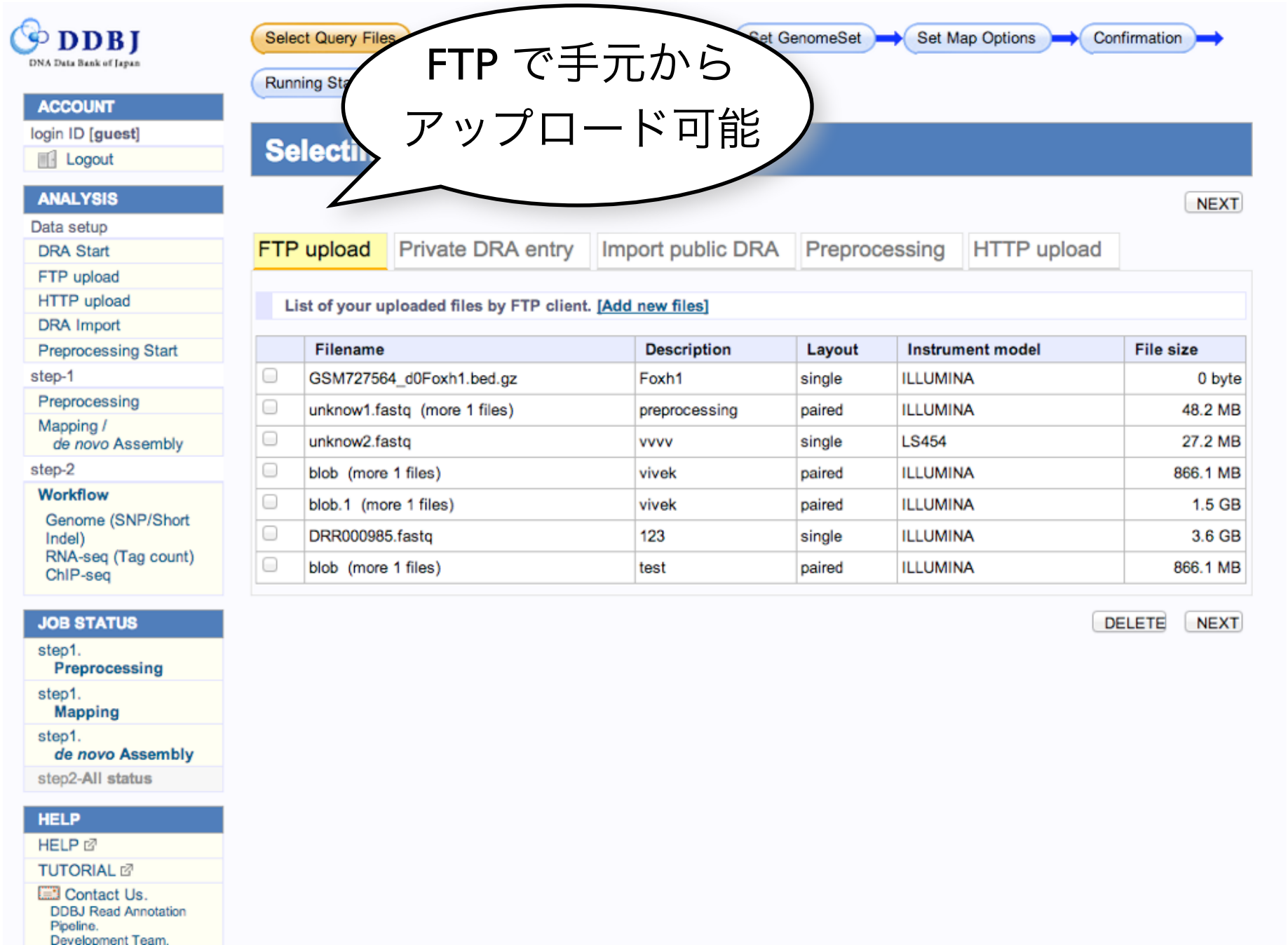
http://p.ddbj.nig.ac.jp



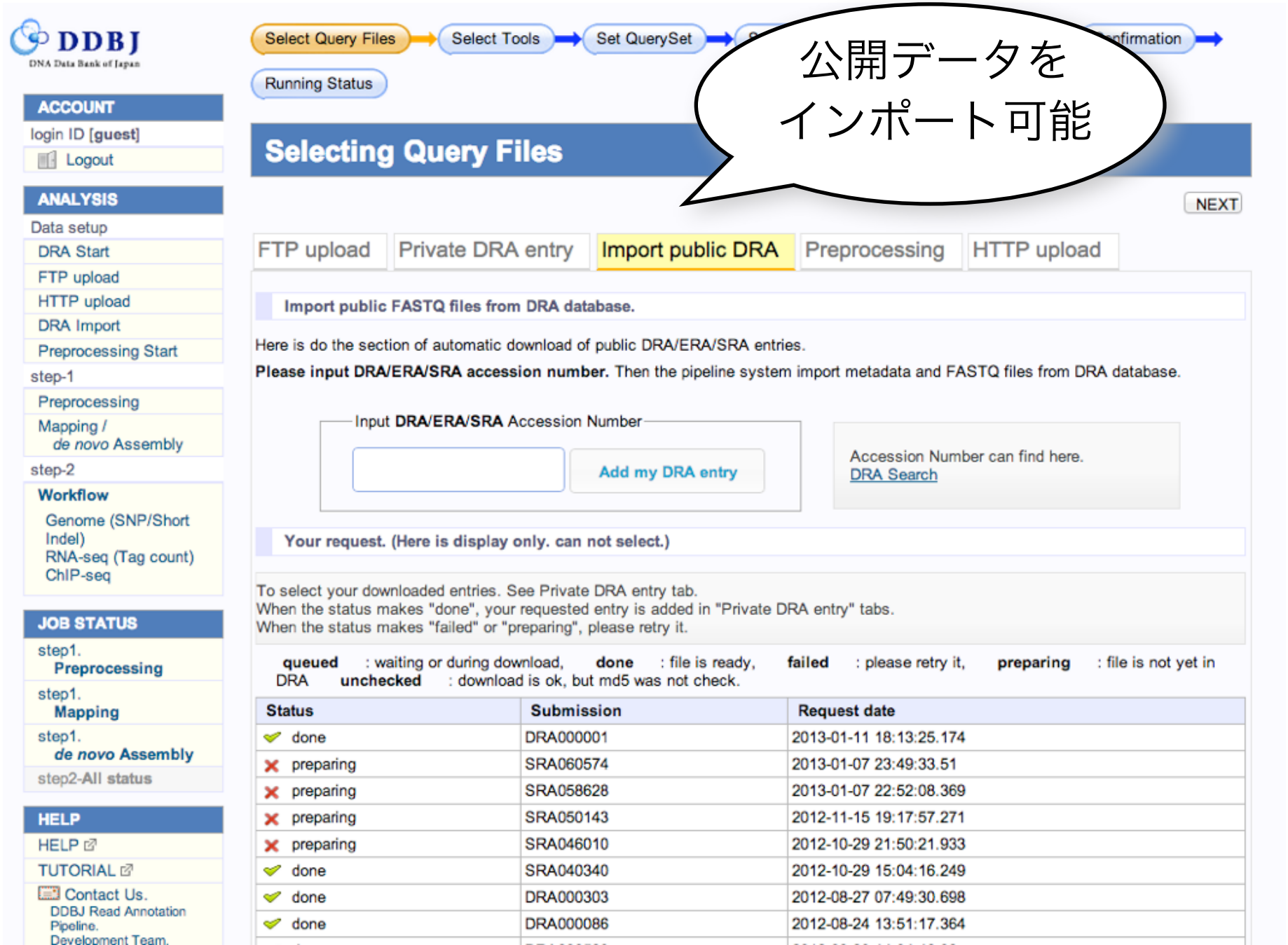
処理に使うNGSの配列ファイルの用意



処理に使うNGSの配列ファイルの用意



処理に使うNGSの配列ファイルの用意



今回はupload済のエントリから

ACCOUNT
login ID [guest]
Logout

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 le… about…

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.
☐	Trinity		r2012-06-08			✓		RNA-Seq De novo Assembly

Mapping Contigs by de novo Assemble to Reference Sequences.

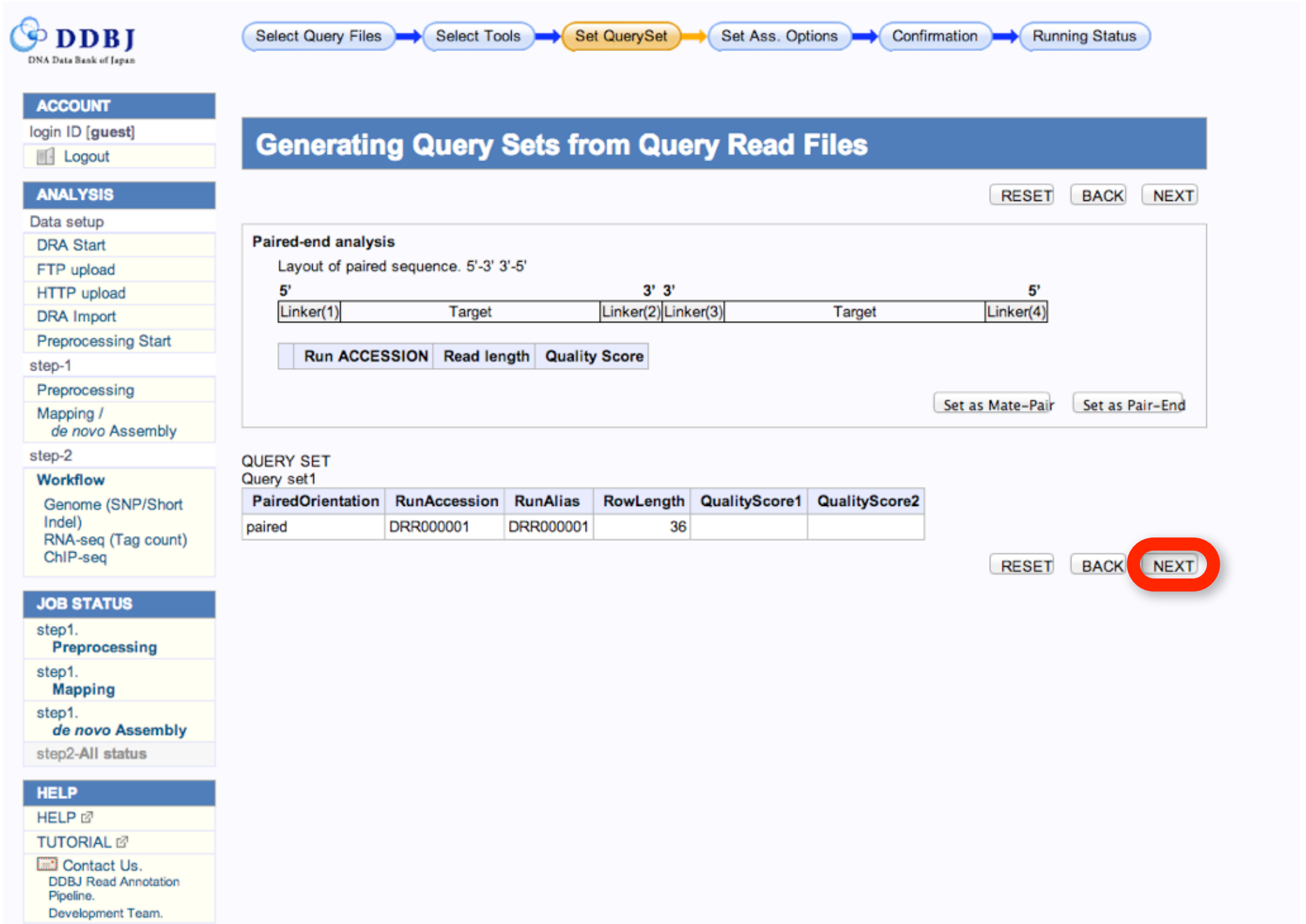
The contigs will be aligned to reference genome.

	Tool	Comment
◉	BLAT	Single-end analysis only

BACK NEXT

velvet で
アセンブル
しましょう

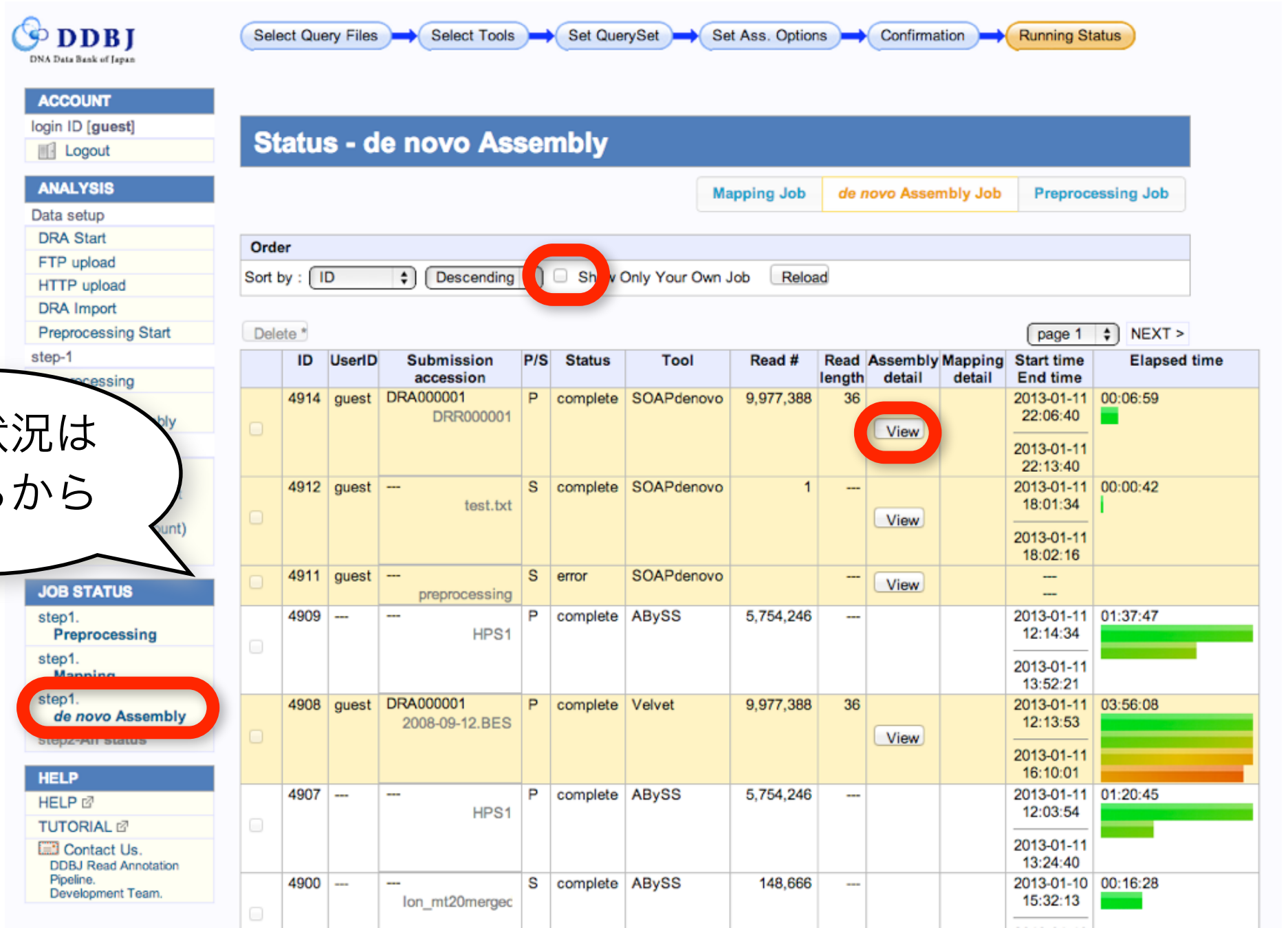
配列のセットの形式を選んで次へ



オプションのパラメータを選べます

終了したらメイルが来ます

「RUN を押した」と思ってください



Detail view

BACK

Job info

ID
4914

Tool (Version)
SOAPdenovo (1.05)

RunAccession or Filename	Download	Read length	Alias
DRR000001	DRR000001.fastq.gz DRR000001_1.fastq.gz DRR000001_2.fastq.bz2	36 bp	DRR000001

Download modified queries

- DRR000001_1.fastq.gz (Original size 1.7 GB)
- DRR000001_2.fastq.gz (Original size 1.7 GB)

Download wgs file

- out_WGS.fasta.gz (Original size 3.9 MB)

Assembly statistics

Contig # : 5,300
Total contig size : 4,138,179
Maximum contig size : 49,938
Minimum contig size : 24
N50 contig size : 13,255

アセンブル結果の基本情報

Time

Wait time	Start time	End time
0: 1:22	2013-01-11 22:06:40	2013-01-11 22:13:40

Command	Start time	End time	Log1	Log2	Result	MD5
SOAPdenovo127mer all -s soapdenovo.conf -o output	2013-01-11 22:06:40	2013-01-11 22:12:28	View	View	Download(174.7 MB)	MD5

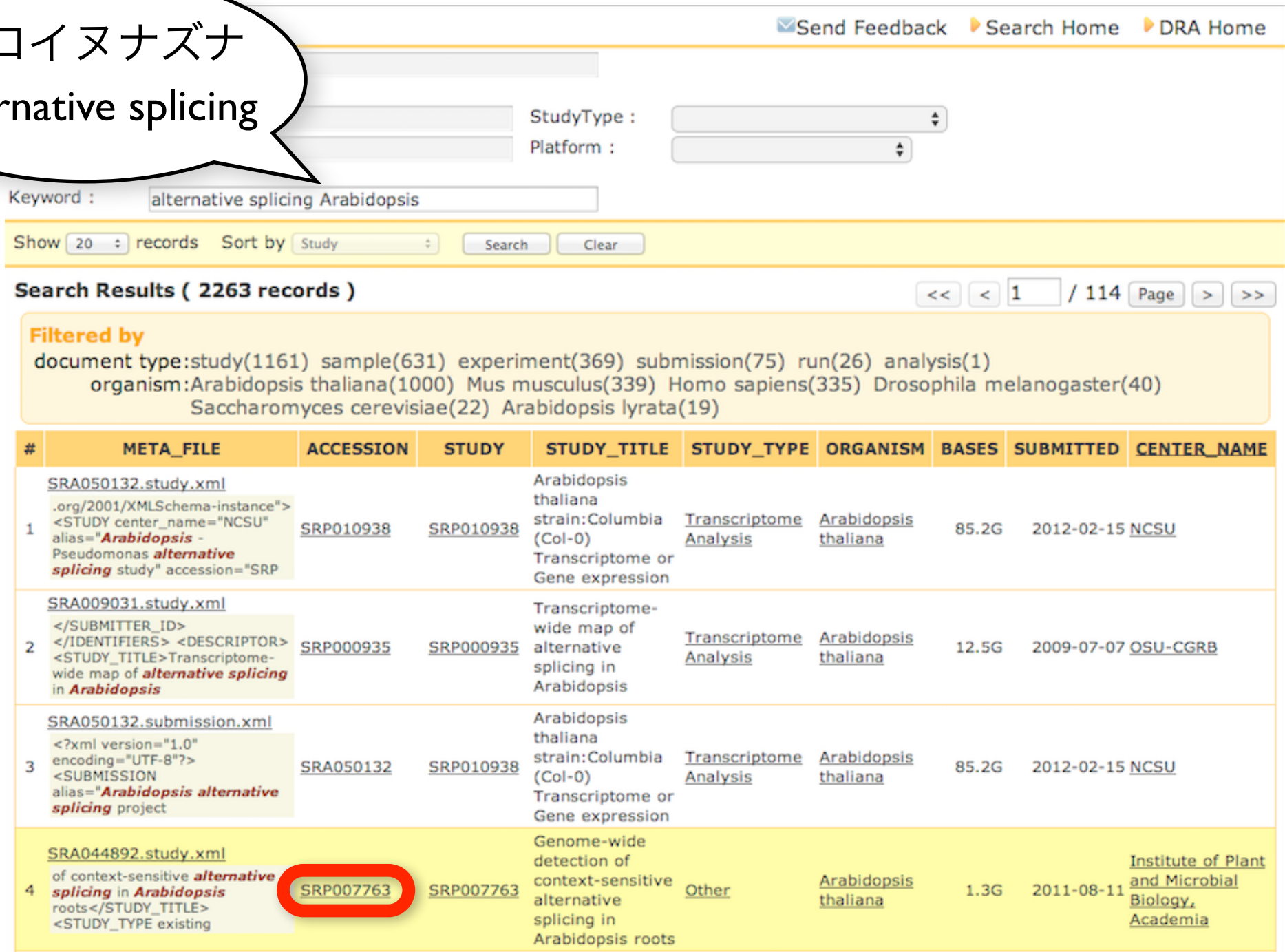
BACK

結果ファイル

Mappingの例 (DRAsearch+pipeline)

シロイヌナズナ

alternative splicing



Send Feedback | Search Home | DRA Home

StudyType :

Platform :

Keyword : alternative splicing Arabidopsis

Show 20 records Sort by Study Search Clear

Search Results (2263 records)

<< < 1 / 114 Page > >>

Filtered by

document type:study(1161) sample(631) experiment(369) submission(75) run(26) analysis(1)

organism:Arabidopsis thaliana(1000) Mus musculus(339) Homo sapiens(335) Drosophila melanogaster(40) Saccharomyces cerevisiae(22) Arabidopsis lyrata(19)

#	META_FILE	ACCESSION	STUDY	STUDY_TITLE	STUDY_TYPE	ORGANISM	BASES	SUBMITTED	CENTER_NAME
1	SRA050132.study.xml .org/2001/XMLSchema-instance"> <STUDY center_name="NCSU" alias="*Arabidopsis* - Pseudomonas *alternative splicing* study" accession="SRP	SRP010938	SRP010938	Arabidopsis thaliana strain:Columbia (Col-0) Transcriptome or Gene expression	Transcriptome Analysis	Arabidopsis thaliana	85.2G	2012-02-15	NCSU
2	SRA009031.study.xml </SUBMITTER_ID> </IDENTIFIERS> <DESCRIPTOR> <STUDY_TITLE>Transcriptome-wide map of *alternative splicing* in *Arabidopsis*	SRP000935	SRP000935	Transcriptome-wide map of alternative splicing in Arabidopsis	Transcriptome Analysis	Arabidopsis thaliana	12.5G	2009-07-07	OSU-CGRB
3	SRA050132.submission.xml <?xml version="1.0" encoding="UTF-8"?> <SUBMISSION alias="*Arabidopsis alternative splicing* project	SRA050132	SRP010938	Arabidopsis thaliana strain:Columbia (Col-0) Transcriptome or Gene expression	Transcriptome Analysis	Arabidopsis thaliana	85.2G	2012-02-15	NCSU
4	SRA044892.study.xml of context-sensitive *alternative splicing* in *Arabidopsis* roots</STUDY_TITLE> <STUDY_TYPE existing	SRP007763	SRP007763	Genome-wide detection of context-sensitive alternative splicing in Arabidopsis roots	Other	Arabidopsis thaliana	1.3G	2011-08-11	Institute of Plant and Microbial Biology, Academia

データのIDはこちら

DRASearch

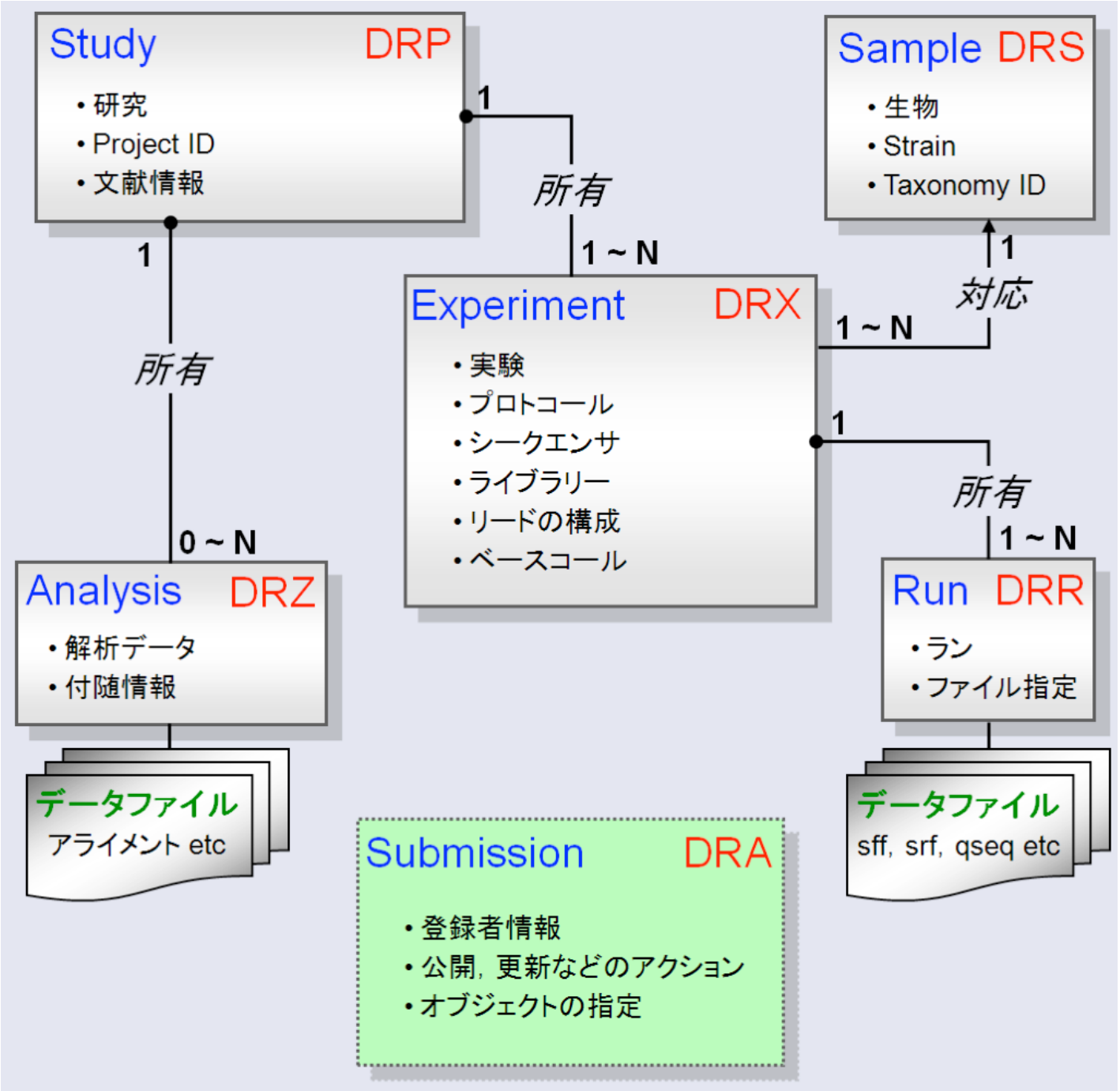
Send Feedback | Search Home | DRA Home

SRP007763

Study Detail	
Title	Genome-wide detection of context-sensitive alternative splicing in Arabidopsis roots
Study Type	Other
Abstract	To analyze context-sensitive changes in pre-mRNA splicing pattern and gene expression, we mapped the transcriptome of iron-deficient and iron-sufficient Arabidopsis roots using the RNA-seq technology. RNA-seq data were analyzed with a newly developed software package, RACKJ (Read Analysis & Comparis .. [more]
Description	
Center Name	Institute of Plant and Microbial Biology, Academia

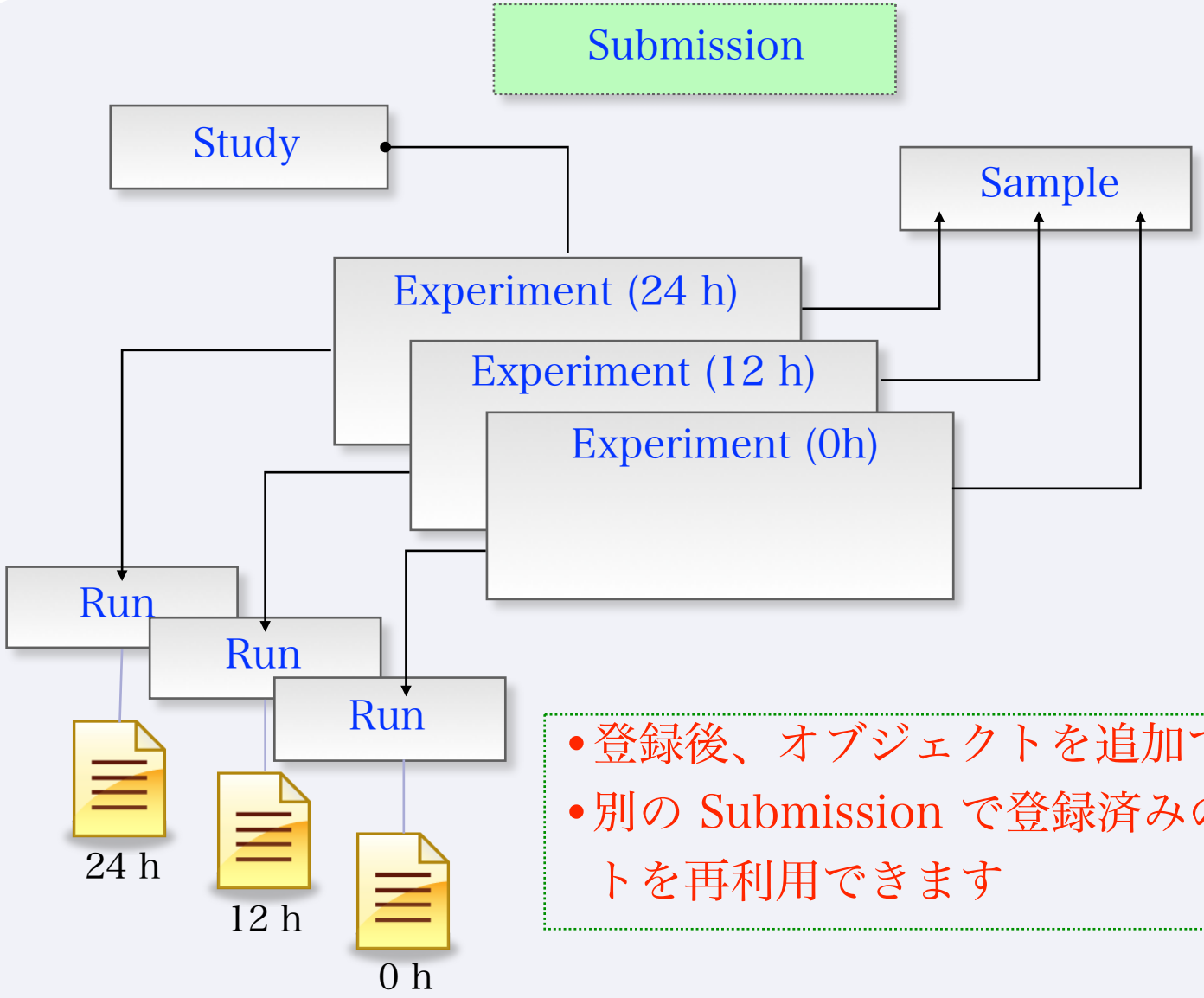
Navigation			
Submissio	SRA044892	FTP	
Experiment	SRX092046	FASTQ	SRA
Sample	SRS256250		

SRA の Metadata



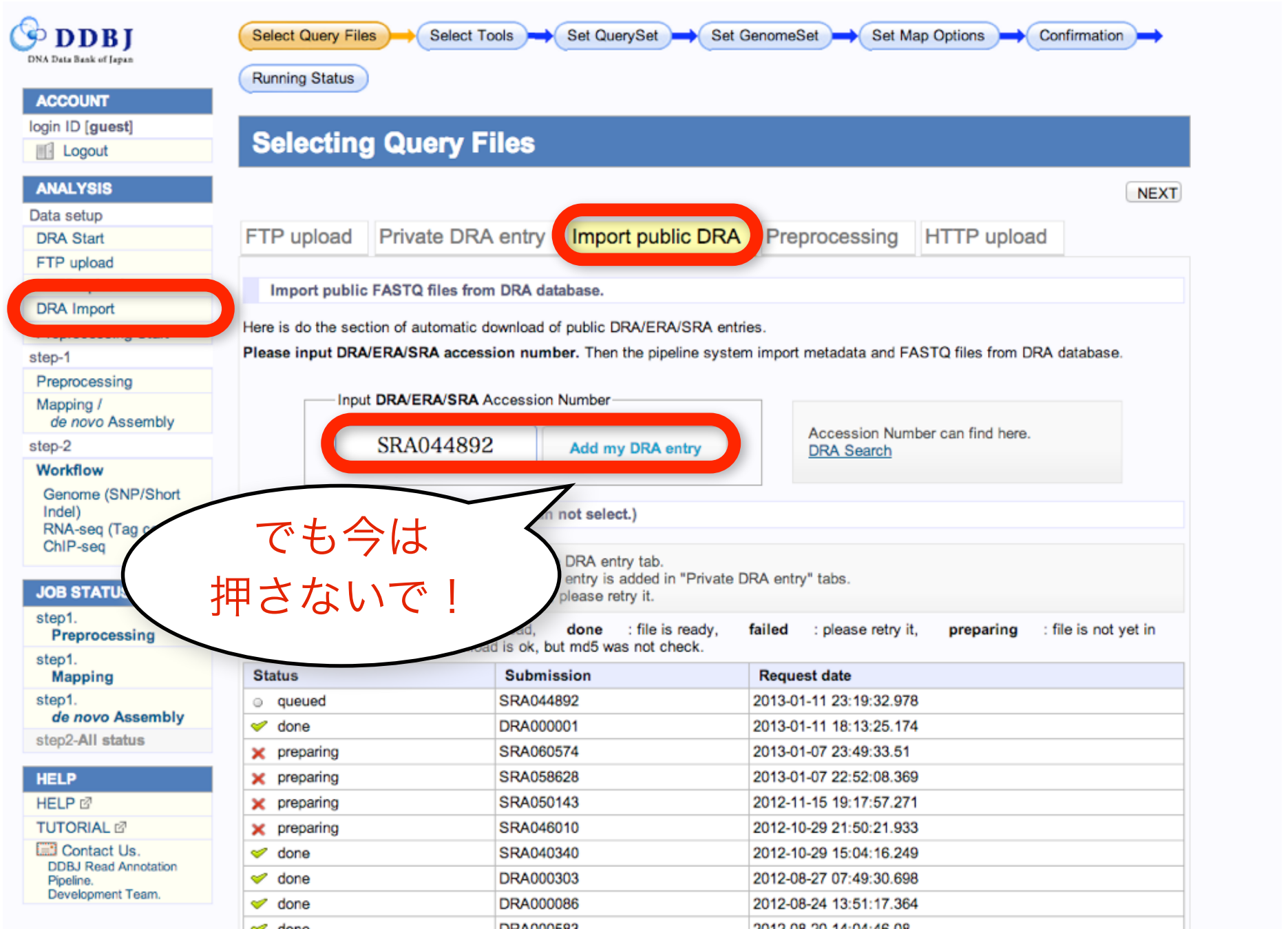
メタデータ構成の例

例) 培養細胞: 薬剤処理 0, 12, 24 h 後の転写プロファイル解析

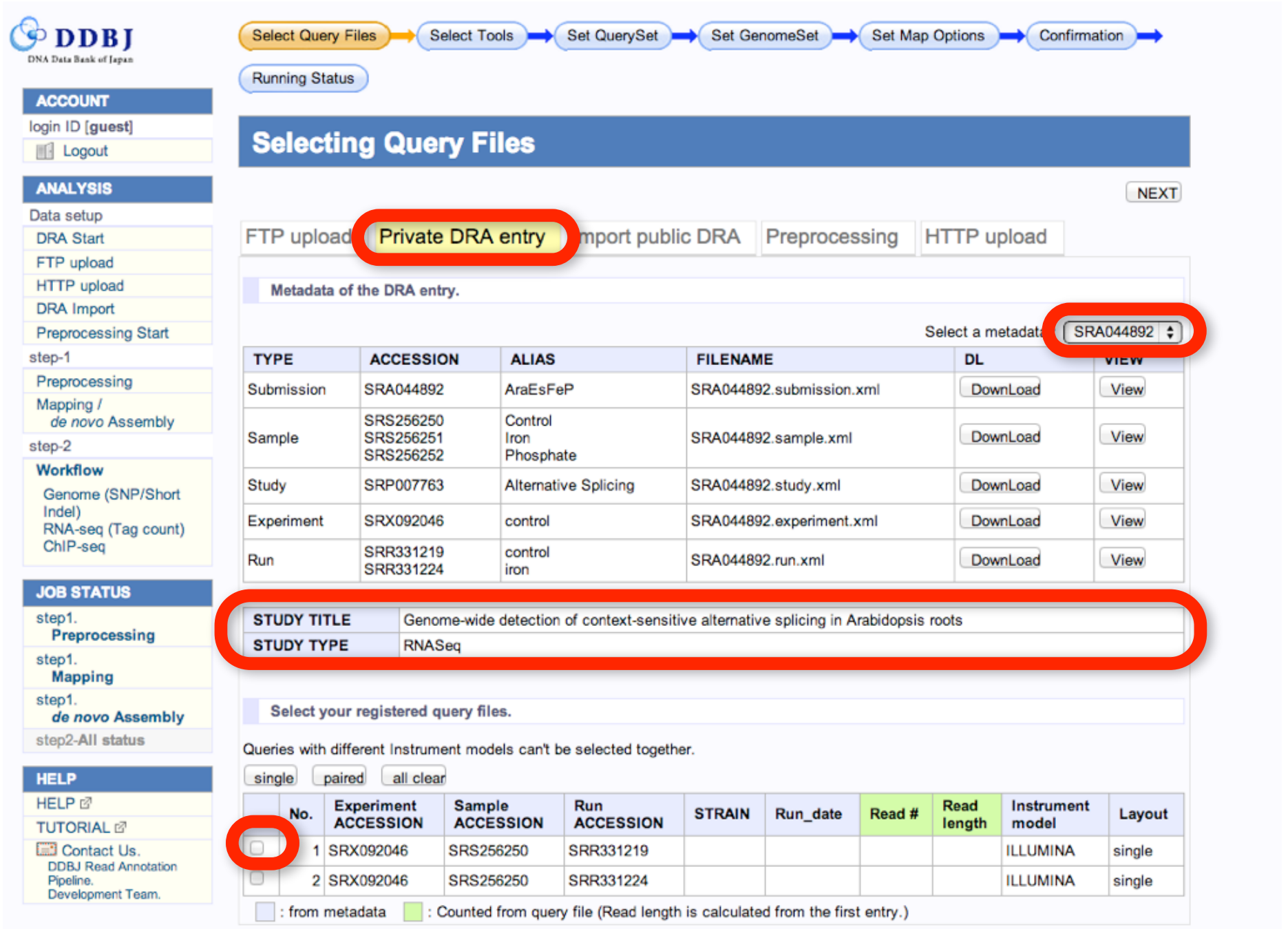


- 登録後、オブジェクトを追加できます
- 別の Submission で登録済みのオブジェクトを再利用できます

p.ddbj.nig.ac.jp を開き、さっきのIDを入力



あらかじめ、ロードしておきました



Selecting Query Files

NEXT

FTP upload | Private DRA entry | Import public DRA | Preprocessing | HTTP upload

Metadata of the DRA entry.

Select a metadata : SRA044892

TYPE	ACCESSION	ALIAS	FILENAME	DL	VIEW
Submission	SRA044892	AraEsFeP	SRA044892.submission.xml	DownLoad	View
Sample	SRS256250 SRS256251 SRS256252	Control Iron Phosphate	SRA044892.sample.xml	DownLoad	View
Study	SRP007763	Alternative Splicing	SRA044892.study.xml	DownLoad	View
Experiment	SRX092046	control	SRA044892.experiment.xml	DownLoad	View
Run	SRR331219 SRR331224	control iron	SRA044892.run.xml	DownLoad	View

STUDY TITLE	Genome-wide detection of context-sensitive alternative splicing in Arabidopsis roots
STUDY TYPE	RNASeq

Select your registered query files.

Queries with different Instrument models can't be selected together.

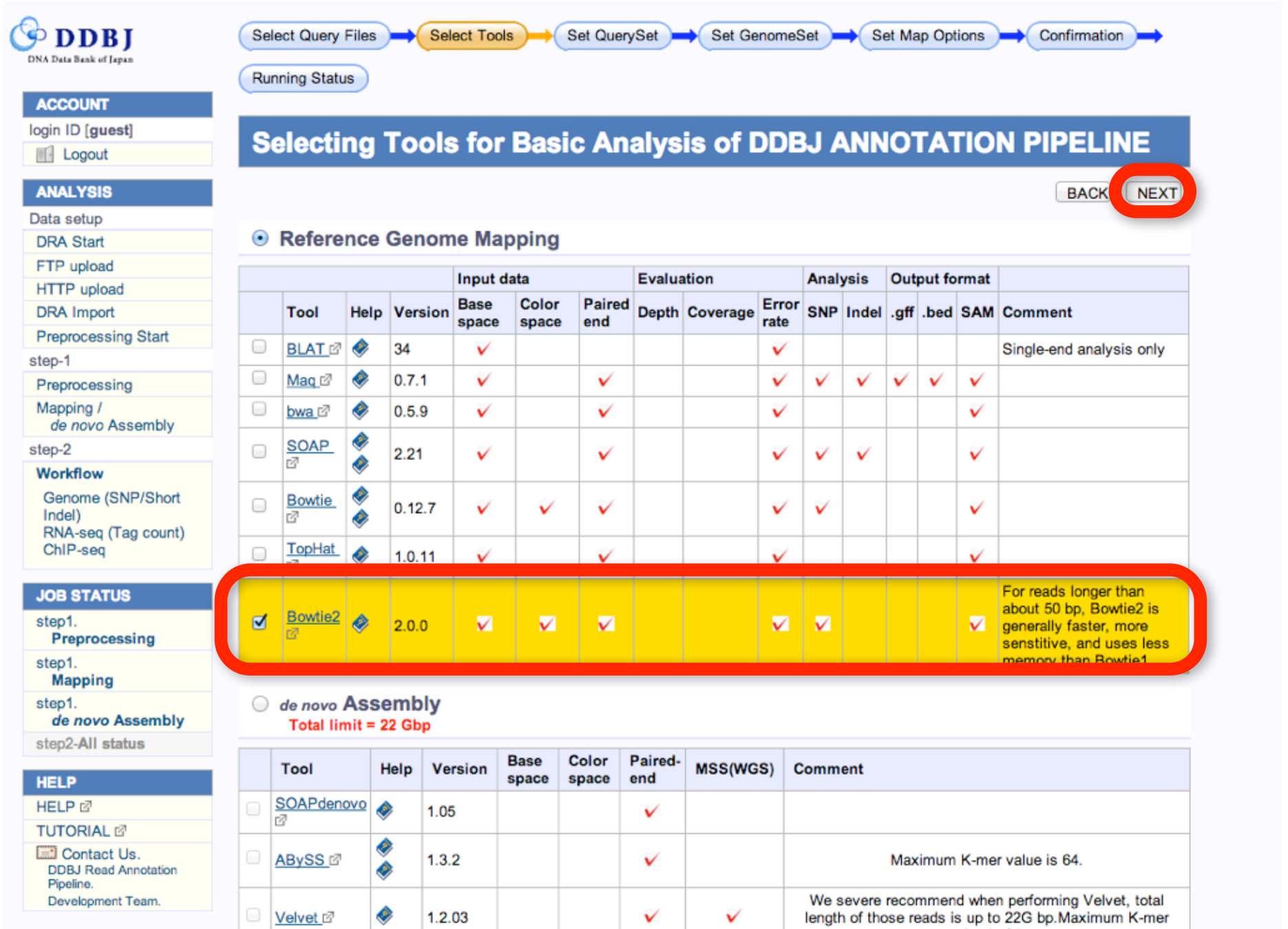
single | paired | all clear

	No.	Experiment ACCESSION	Sample ACCESSION	Run ACCESSION	STRAIN	Run_date	Read #	Read length	Instrument model	Layout
☑	1	SRX092046	SRS256250	SRR331219					ILLUMINA	single
☐	2	SRX092046	SRS256250	SRR331224					ILLUMINA	single

: from metadata : Counted from query file (Read length is calculated from the first entry.)

DELETE NEXT

Bowtie2 を選んで NEXT



DDBJ
DNA Data Bank of Japan

Select Query Files → Select Tools → Set QuerySet → Set GenomeSet → Set Map Options → Confirmation → Running Status

ACCOUNT
login ID [guest]
Logout

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

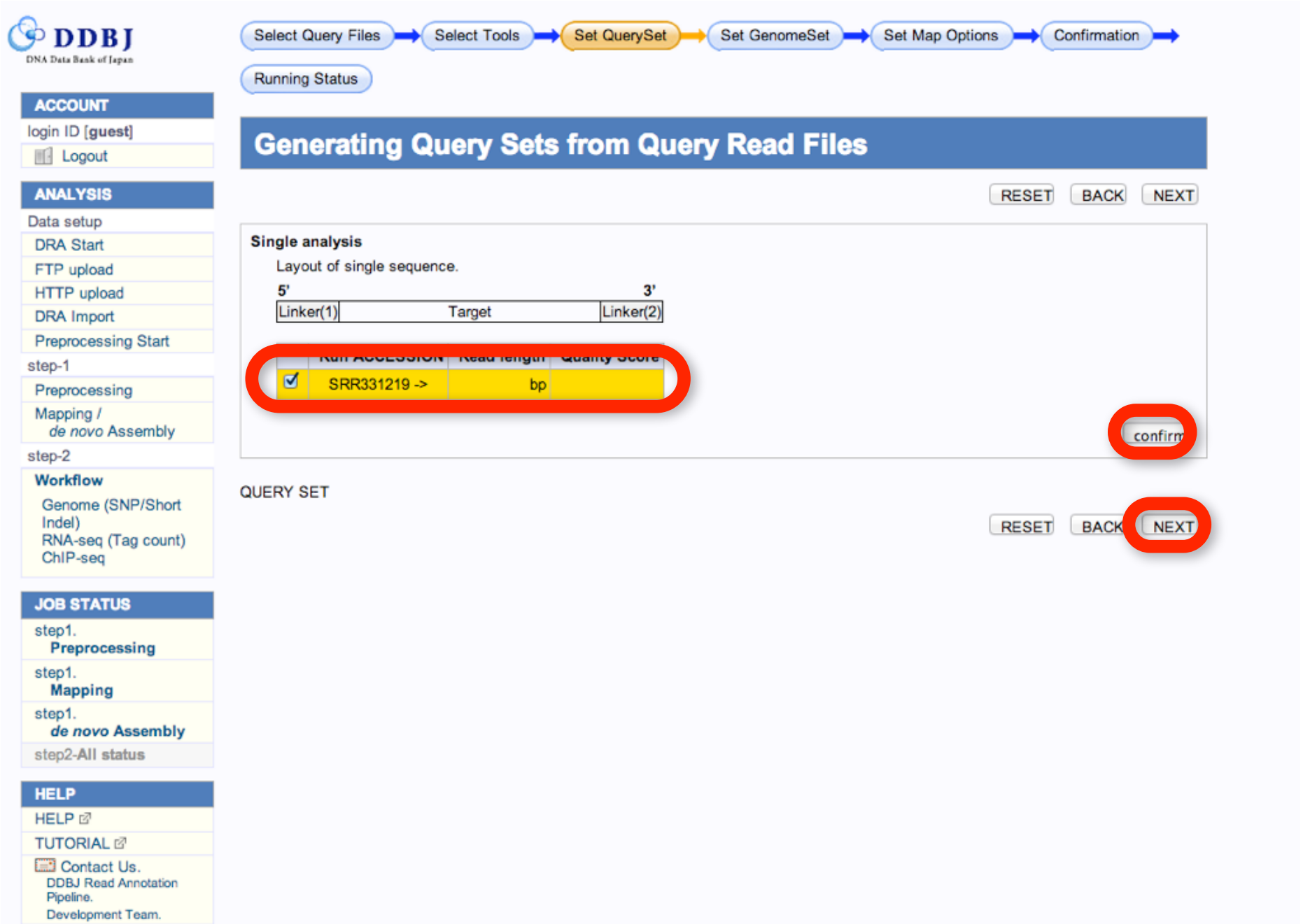
				Input data			Evaluation			Analysis		Output format			
	Tool	Help	Version	Base space	Color space	Paired end	Depth	Coverage	Error rate	SNP	Indel	.gff	.bed	SAM	Comment
☐	BLAT		34	✓					✓						Single-end analysis only
☐	Maq		0.7.1	✓		✓			✓	✓	✓	✓	✓	✓	
☐	bwa		0.5.9	✓		✓			✓					✓	
☐	SOAP		2.21	✓		✓			✓	✓	✓			✓	
☐	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

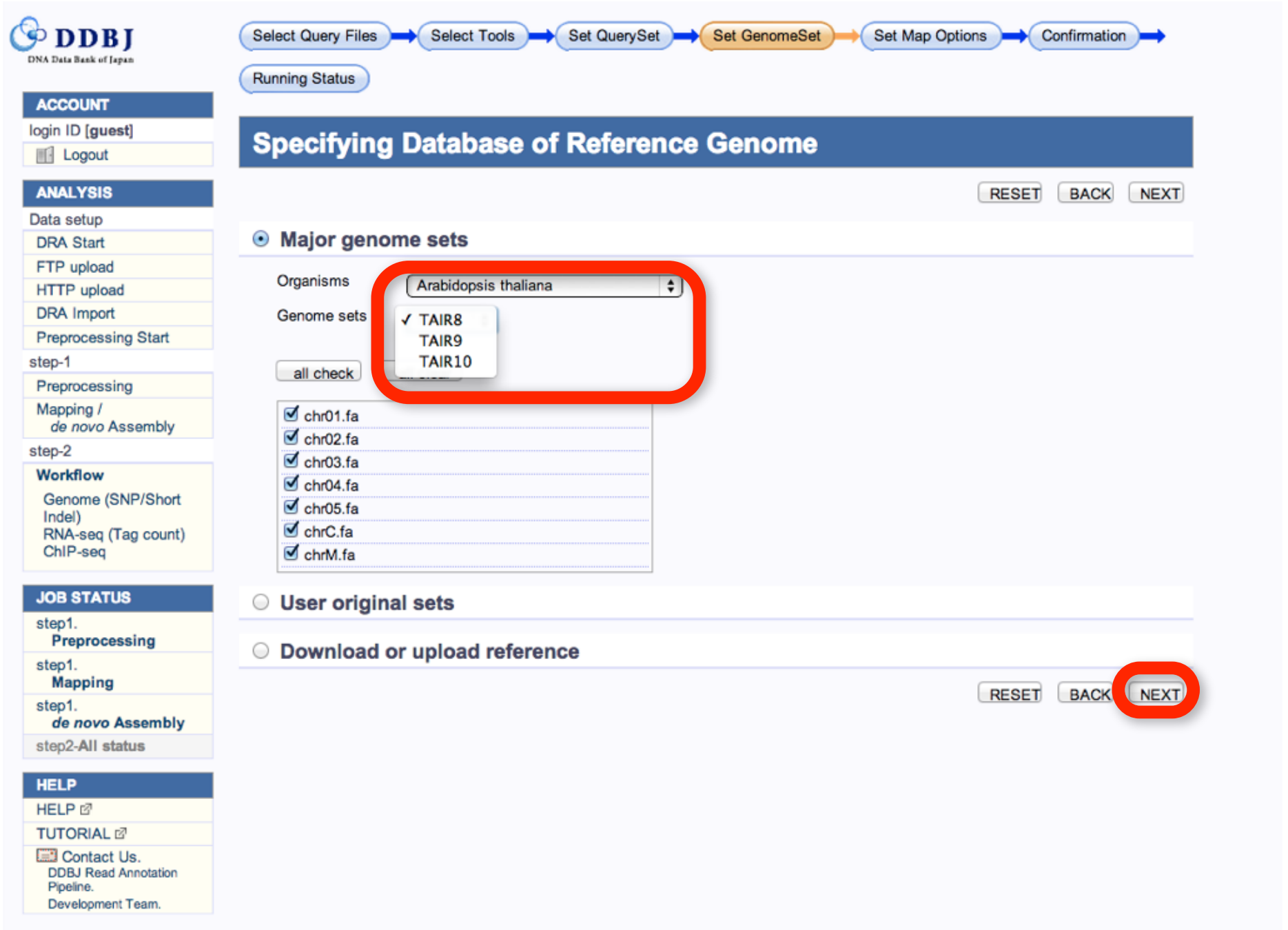
Total limit = 22 Gbp

	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

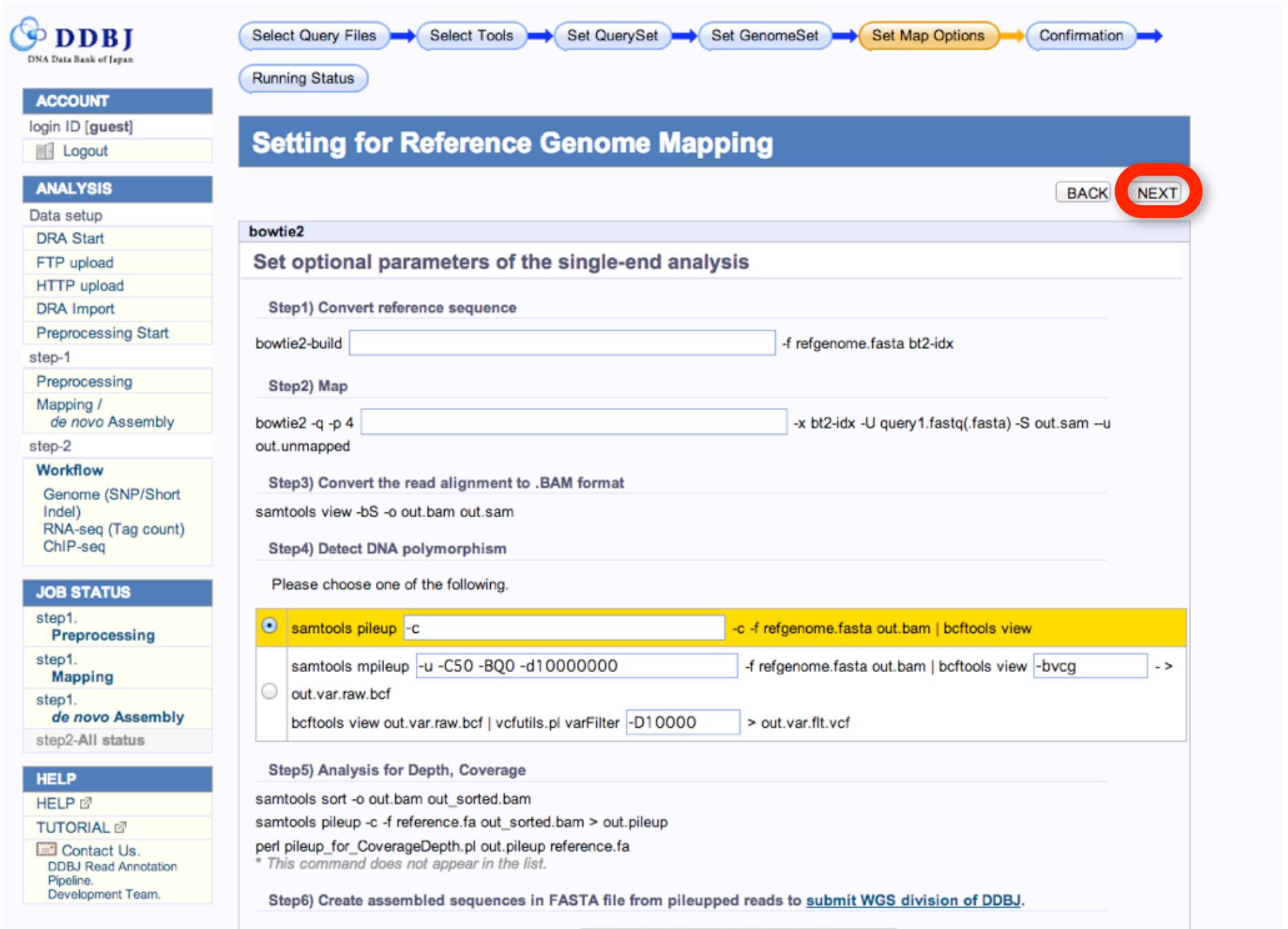
配列を選んで confirm, NEXT



TAIR10（最新）を選んでNEXT

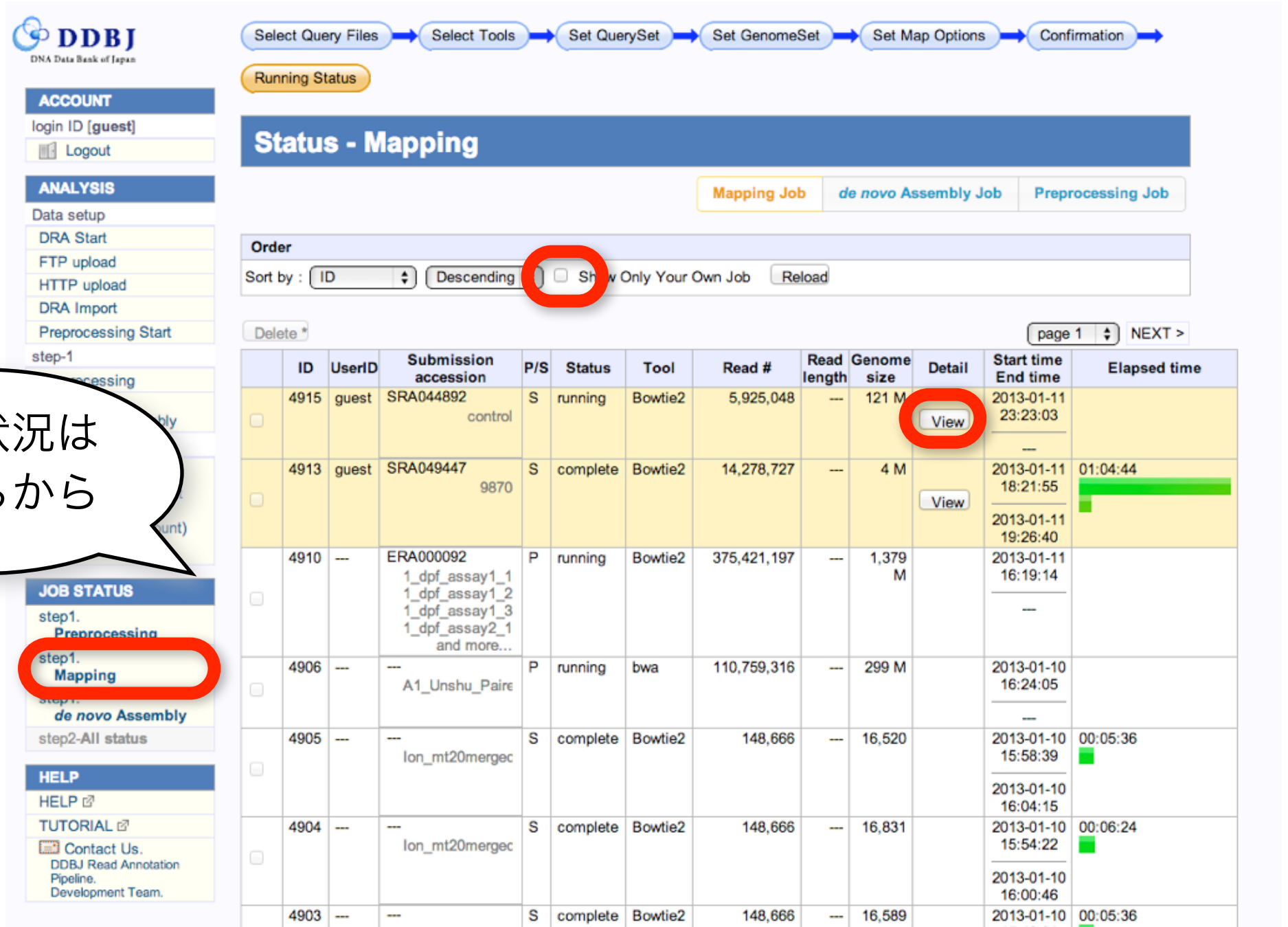


option 変更なければそのままNEXT



終了したらメイルが来ます

「RUN を押した」と思ってください



処理状況は
こちらから

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

実行結果

Development Team.

Tool (Version)
Bowtie2 (2.0.0)

RunAccession or Filename	Download	Read length	Alias
SRR331219	SRR331219.fastq.bz2	N.A. bp	control

Genome set
TAIR10

Chromosome
all.fa

Download modified queries

- SRR331219.fastq.gz (Original size 1.5 GB)

Download merged pileup file

- merged.pileup.gz (Original size 1.6 GB)
- merged.sam.gz (Original size 1.4 GB)

Download wgs file

- out_WGS.fasta.gz (Original size 67.7 MB)

Position errors	Map ratio	Depth, Coverage
PDF download	total query # : 5,925,048 mapped query # : 5,037,456 map ratio : 85.020 %	coverage : 29886366 / 119482012 * 100 = 25.013 depth : 384468141 / 29886366 = 12.864

Time

Wait time	Start time	End time
0: 1:12	2013-01-11 23:23:03	2013-01-12 00:19:10

all.fa

Command	Start time	End time	Log1	Log2	Result	MD5
bowtie2-build -f all.fa refgenome	2013-01-11 23:23:03	2013-01-11 23:25:26	View		Download(194.4 MB)	MD5
bowtie2 -p 4 -q -x refgenome -U SRR331219.fastq -S out.map --un out.unmapped	2013-01-11 23:26:07	2013-01-11 23:38:02		View	Download(463.0 MB)	MD5
samtools view -bS -o out.bam out.map	2013-01-11 23:40:56	2013-01-11 23:42:28		View	Download(488.9 MB)	MD5
samtools sort out.bam out2	2013-01-11 23:42:59	2013-01-11 23:44:52		View	Download(357.5 MB)	MD5