統合データベース講習会：AJACS千里

2015年6月16日

次世代シーケンサー（NGS）と関連するデータベース・ツール

情報・システム研究機構（ROIS） ライフサイエンス統合データベースセンター（DBCLS）
科学技術振興機構（JST） バイオサイエンスデータベースセンター（NBDC）

河野　信

本日紹介する内容（次世代シーケンスDB）

- 次世代シーケンスデータベースの概要
- 【実習】DRAsearch
  - 次世代シーケンスデータベース検索
- 【実習】DDBJ解析パイプライン
  - ウェブ上で次世代シーケンスデータを解析する
- 【応用】p-Galaxy/MiGAP
  - 配列データのアノテーション

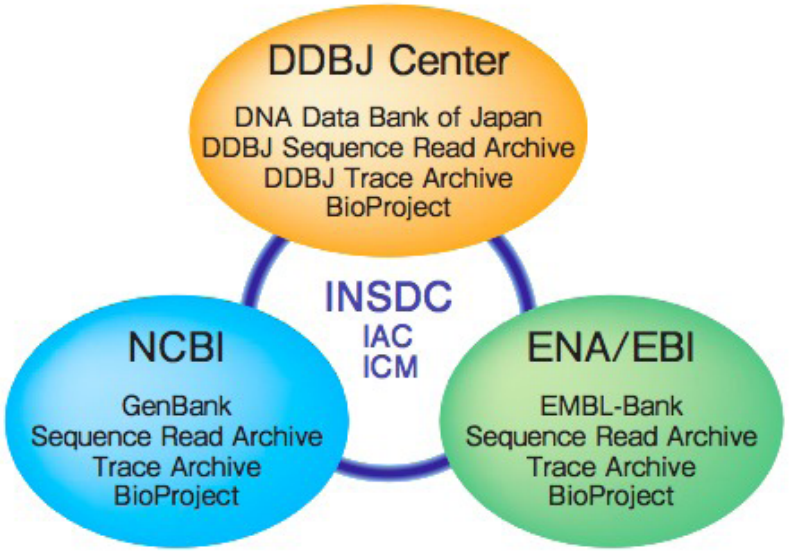
次世代シーケンサデータベースの概要

次世代シーケンサー（NGS）

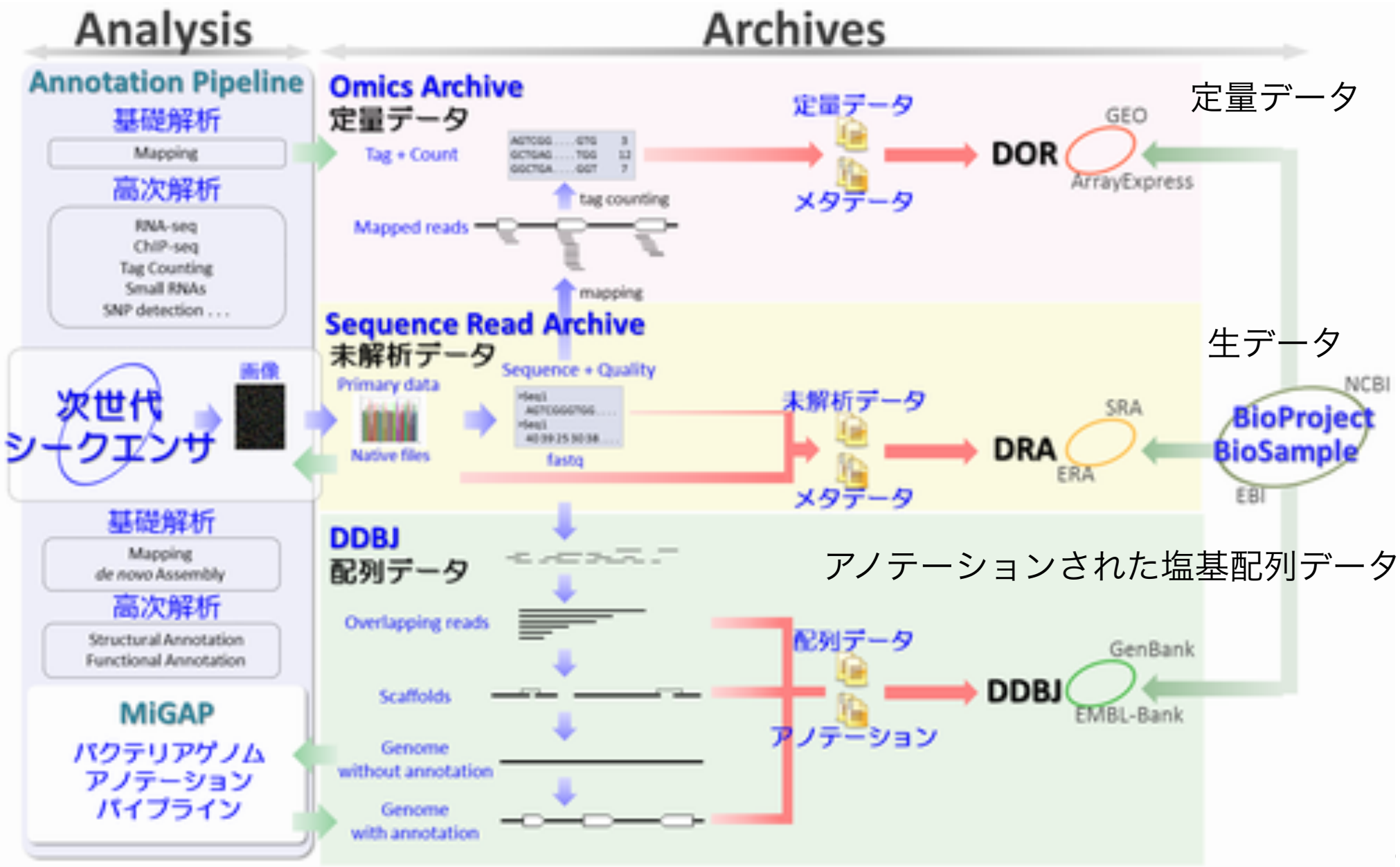
- 第一世代：キャピラリーシーケンサー
- 第二世代
  - 短い配列（～150 bp）を超並列（～12G）に読む
  - Illumina HiSeq/NextSeq/MiSeq, 454, SOLiD
- 第三世代
  - PacBio
    - 超ロングリード
  - NanoPore
    - USBメモリサイズ

Sequence Read Archive

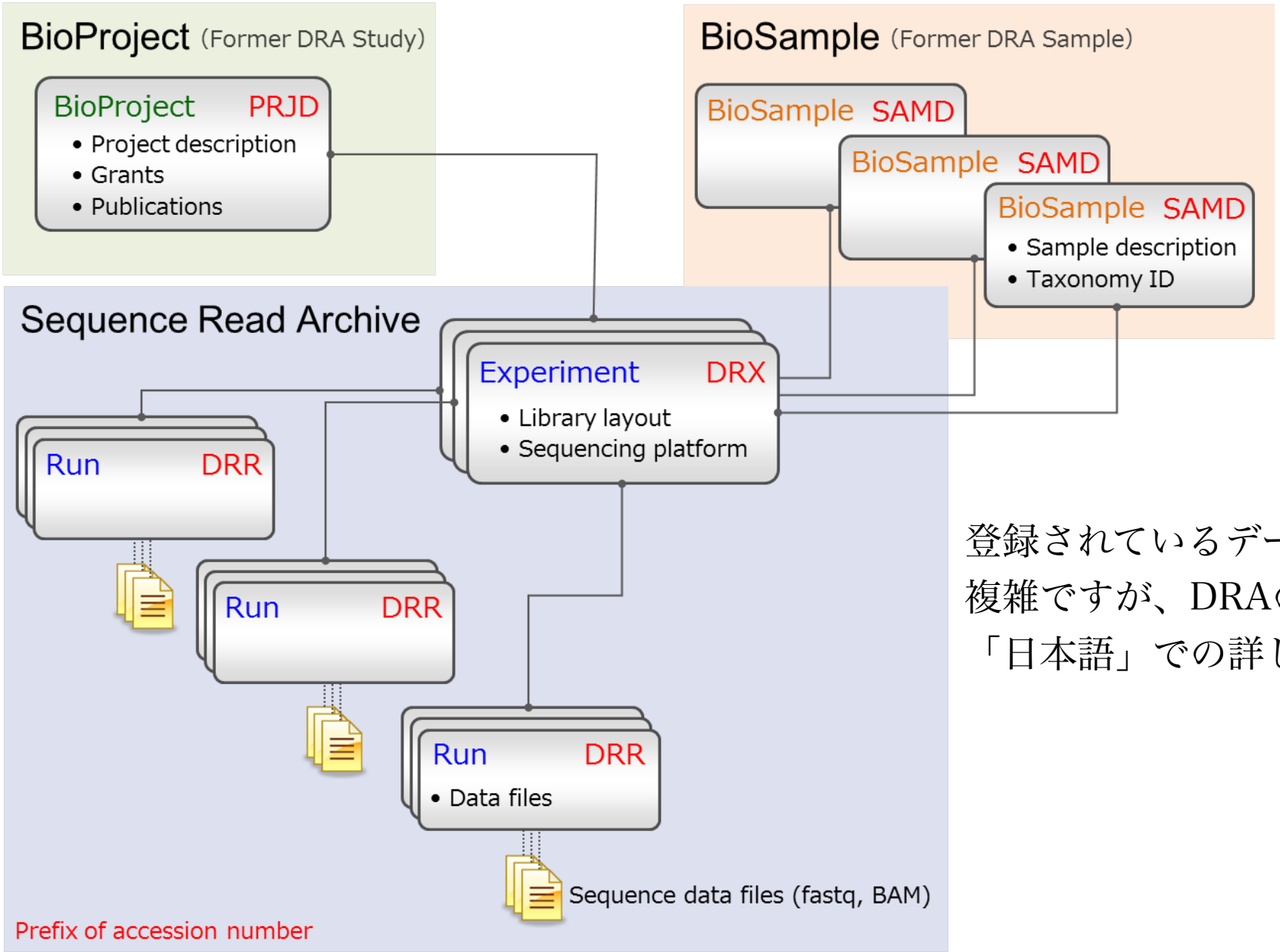
- SRA (NCBI)/ENA (EBI)/DRA (DDBJ)
  - 次世代シーケンスデータはどれかに登録する
    - 生データとマッピングデータ、メタデータのみ登録
    - 解析データは従来通りGenBank/ENA/DDBJに登録する
    - 定量データはGEO/ArrayExpress/(DOR)に登録する
  - 3データベースでデータを毎日交換しており、データの内容は同じ



データの種類とデータベース



次世代シーケンスのデータ構造



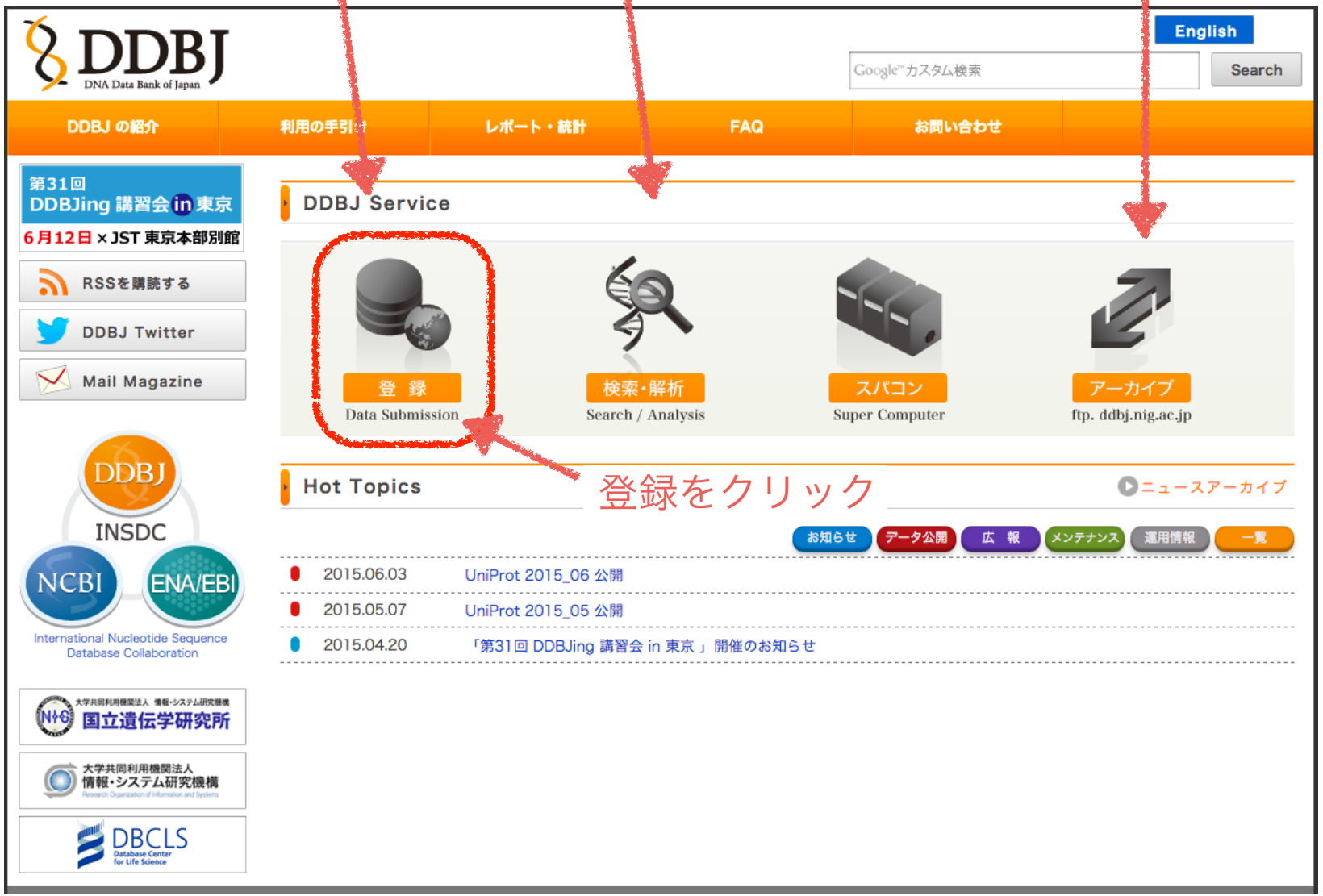
登録されているデータ構造は少々複雑ですが、DRAのページでは「日本語」での詳しい説明がある

実習

DRAsearchを使って
次世代シーケンスデータを検索する

DDBJトップページ　http://www.ddbj.nig.ac.jp/

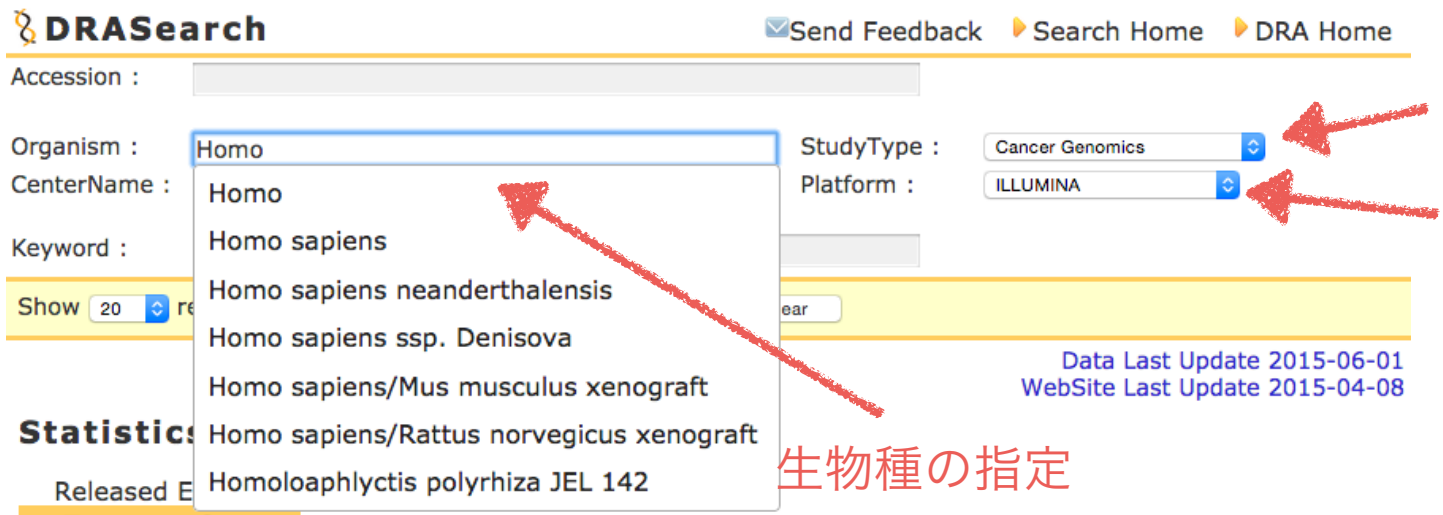
DDBJ/DRA登録　データ検索・解析　　データダウンロード



DDBJデータ登録画面

DRAウェブサイト http://trace.ddbj.nig.ac.jp/dra/

データセットの検索：カテゴリを指定



StudyTypeを指定

Platformを指定

生物種の指定
（数文字入力すると候補が表示される）

Type	Count
Submission	411042
Study	54015
Experiment	1087606
Sample	1021379
Run	1333157

Organism

#	Organism Name	Study
1	Homo sapiens	4723
2	Mus musculus	3390
3	Populus trichocarpa	979
4	Drosophila melanogaster	806
5	soil metagenome	600
6	Arabidopsis thaliana	546
7	Caenorhabditis elegans	466
8	Neurospora crassa	433
9	Saccharomyces cerevisiae	432
10	marine metagenome	428

Study Type

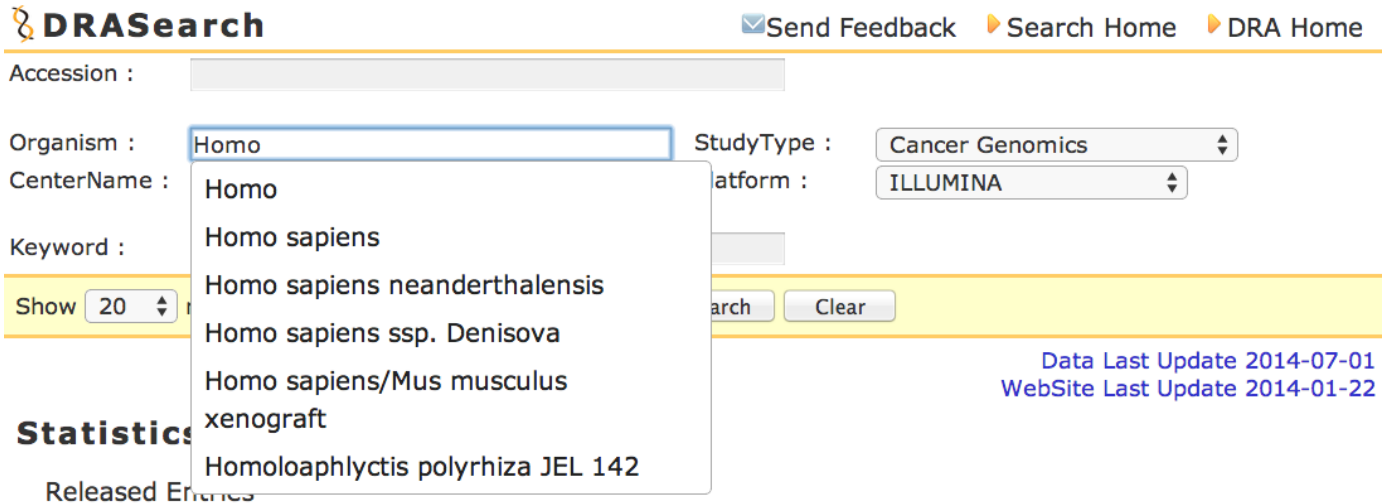
#	Study Type	Study
1	Whole Genome Sequencing	25829
2	Other	13030
3	Transcriptome Analysis	6691
4	Metagenomics	4693
5		2224
6	Population Genomics	698
7	Epigenetics	583
8	Exome Sequencing	153
9	Cancer Genomics	73
10	Pooled Clone Sequencing	32

Center Name

#	Center Name	Study
1	BioProject	24721
2	GEO	8000
3	UMIGS	2556
4	JGI	2368
5		2224
6	SC	1155
7	JCVI	1144
8	BI	962
9	WUGSC	692
10	WUSTL	535

メタデータによる分類

参考：約1年前の登録数



Type	Count
Submission	258755
Study	38619
Experiment	714047
Sample	688775
Run	831192

Organism

#	Organism Name	Study
1	Homo sapiens	2858
2	Mus musculus	2046
3	Drosophila melanogaster	649
4	unidentified	620
5	Caenorhabditis elegans	407
6	Arabidopsis thaliana	360
7	soil metagenome	342
8	marine metagenome	287
9	Saccharomyces cerevisiae	281
10	Solanum lycopersicum	187

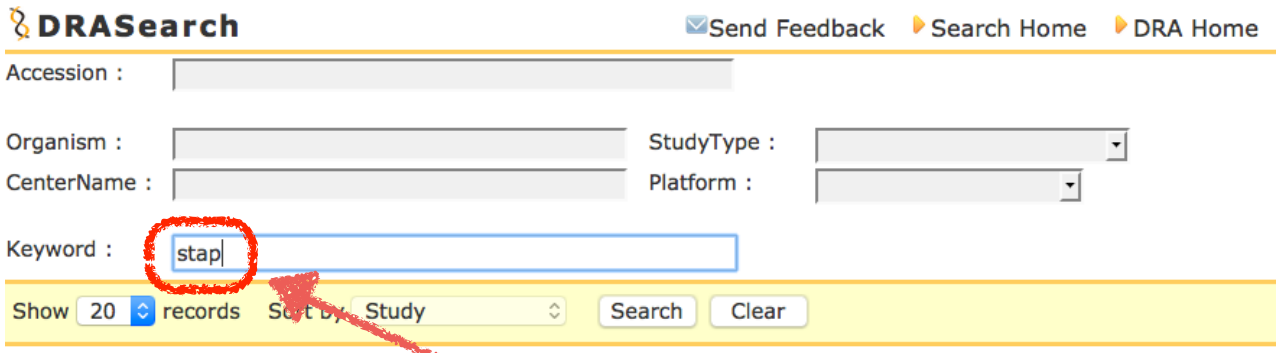
Study Type

#	Study Type	Study
1	Whole Genome Sequencing	21417
2	Other	6387
3	Transcriptome Analysis	4541
4	Metagenomics	2656
5	Epigenetics	1528
6		1440
7	Population Genomics	465
8	Exome Sequencing	89
9	Cancer Genomics	50
10	Pooled Clone Sequencing	31

Center Name [All List]

#	Center Name	Study
1	BioProject	16234
2	GEO	5291
3	JGI	2665
4	UMIGS	2575
5		1440
6	JCVI	1161
7	BI	962
8	SC	790
9	WUSTL	535
10	WUGSC	506

データセットの検索：キーワード検索



何か適当なキーワードで検索

Statistics

Released Entries

Type	Count
Submission	396380
Study	54456
Experiment	1060330
Sample	1029843
Run	1311201

Organism

#	Organism Name	Study
1	Homo sapiens	4717
2	Mus musculus	3403
3	Populus trichocarpa	978
4	Drosophila melanogaster	810
5	soil metagenome	596
6	Arabidopsis thaliana	550
7	Caenorhabditis elegans	473
8	Neurospora crassa	434
9	Saccharomyces cerevisiae	432
10	marine metagenome	426

Study Type

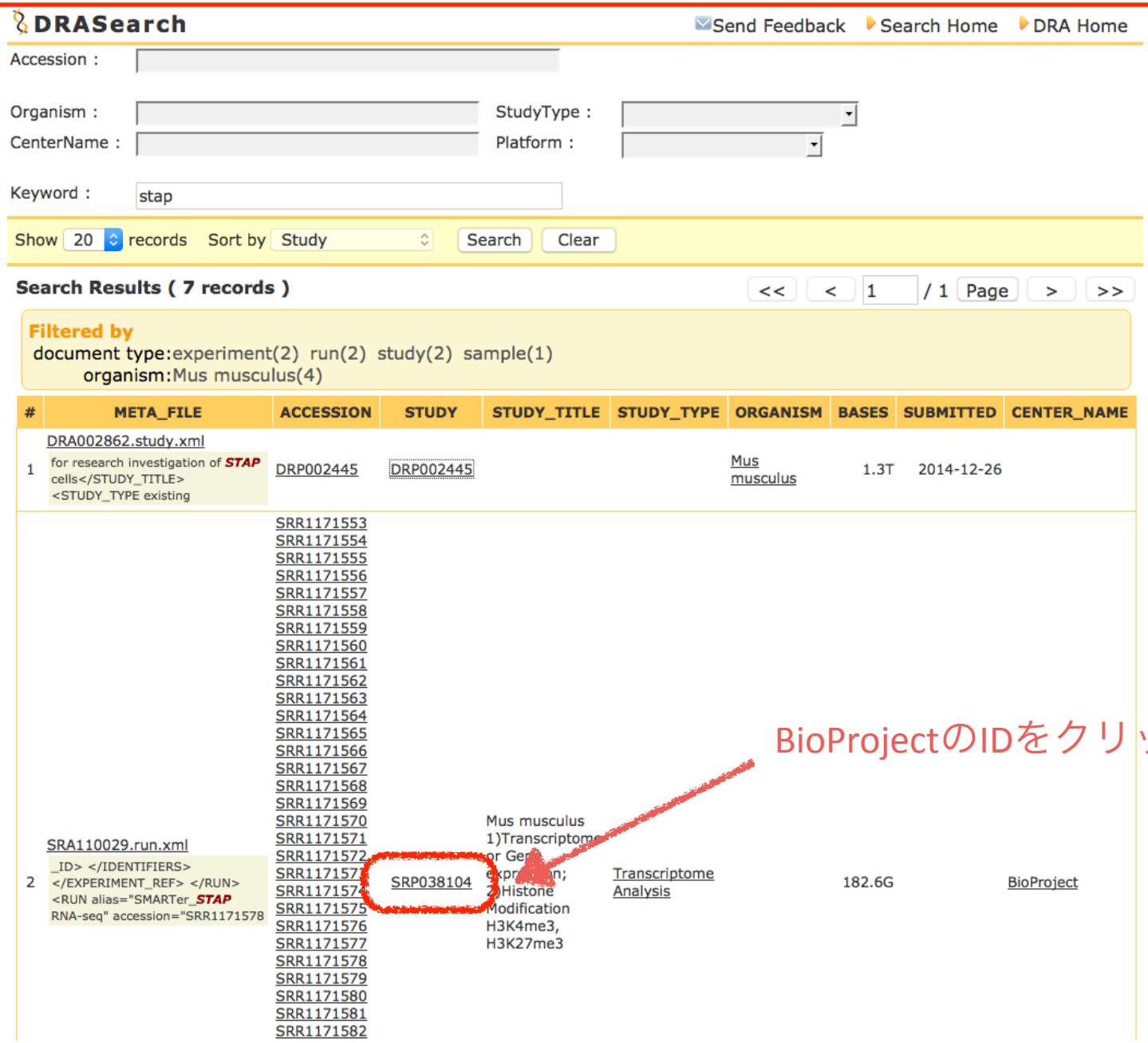
#	Study Type	Study
1	Whole Genome Sequencing	25878
2	Other	13214
3	Transcriptome Analysis	6776
4	Metagenomics	4811
5		2228
6	Population Genomics	701
7	Epigenetics	581
8	Exome Sequencing	154
9	Cancer Genomics	72
10	Pooled Clone Sequencing	32

Center Name

#	Center Name	Study
1	BioProject	25015
2	GEO	8141
3	UMIGS	2553
4	JGI	2363
5		2230
6	JCVI	1139
7	SC	1123
8	BI	962
9	WUGSC	691
10	WUSTL	535

メタデータによる分類

DRA検索結果



BioProjectエントリ

DRASearch　Send Feedback　Search Home　DRA Home

SRP038104

Study Detail	
Title	Mus musculus 1)Transcriptome or Gene expression; 2)Histone Modification H3K4me3, H3K27me3
Study Type	Transcriptome Analysis
Abstract	Global Expression Profile and Epigenetic profile of STAP and other types of ES cells.
Description	
Center Name	BioProject

Navigation			
Submission	SRA110029	FTP	
Experiment	SRX472627	FASTQ	SRA
	SRX472628	FASTQ	SRA
	SRX472629	FASTQ	SRA
	SRX472630	FASTQ	SRA
	SRX472631	FASTQ	SRA
	SRX472632	FASTQ	SRA
	SRX472633	FASTQ	SRA
	SRX472634	FASTQ	SRA
	SRX472635	FASTQ	SRA
	SRX472636	FASTQ	SRA
	SRX472637	FASTQ	SRA
	SRX472638	FASTQ	SRA
	SRX472639	FASTQ	SRA

Sample	SRS559080
	SRS559081
	SRS559082
	SRS559083

ExperimentのIDをクリック

Experimentエントリ

DRASearch　Send Feedback　Search Home　DRA Home

SRX472627 FASTQ SRA

Experiment Detail	
Title	CD45 positive Cells ; RNASeq_Rep1
Design Description	TruSeq RNA Sample Prep Kit v2
Organism	

Library Description	
Name	
Strategy	RNA-Seq
Source	TRANSCRIPTOMIC
Selection	cDNA
Layout	PAIRED
Orientation	
Nominal Length	
Nominal Sdev	
Construction Protocol	

Platform	
Platform	ILLUMINA
Instrument Model	Illumina HiSeq 1500

Processing	
Base Calls	
Sequence Space	
Base Caller	

Navigation		
Submission	SRA110029	FTP
Study	SRP038104	
Sample	SRS559080	
Run	SRR1171556	FASTQ SRA

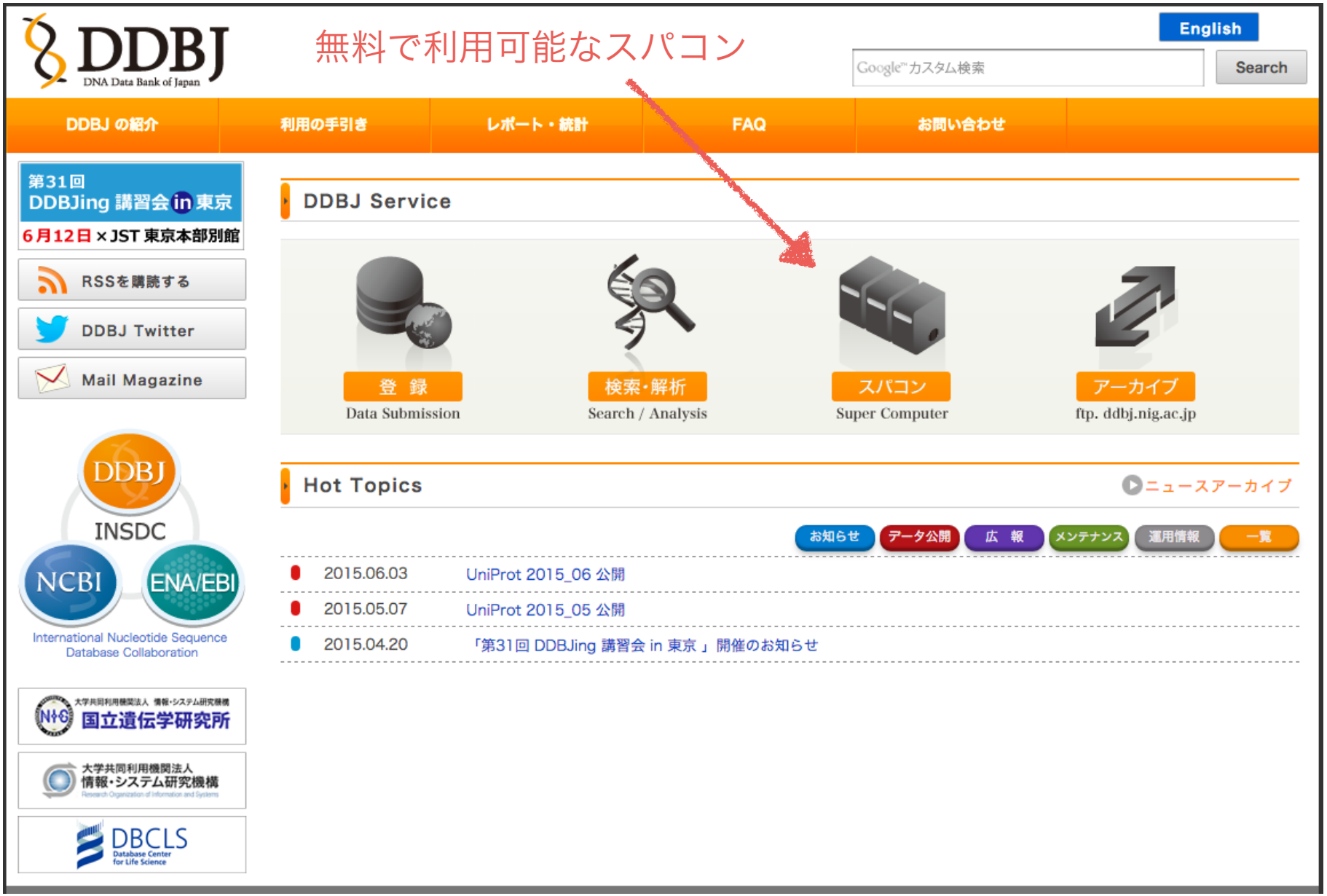
サンプル情報

データをダウンロード

公共リポジトリに登録されている
データをダンロードできるのはわかったけど、
どう使うの？

DDBJトップページ

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



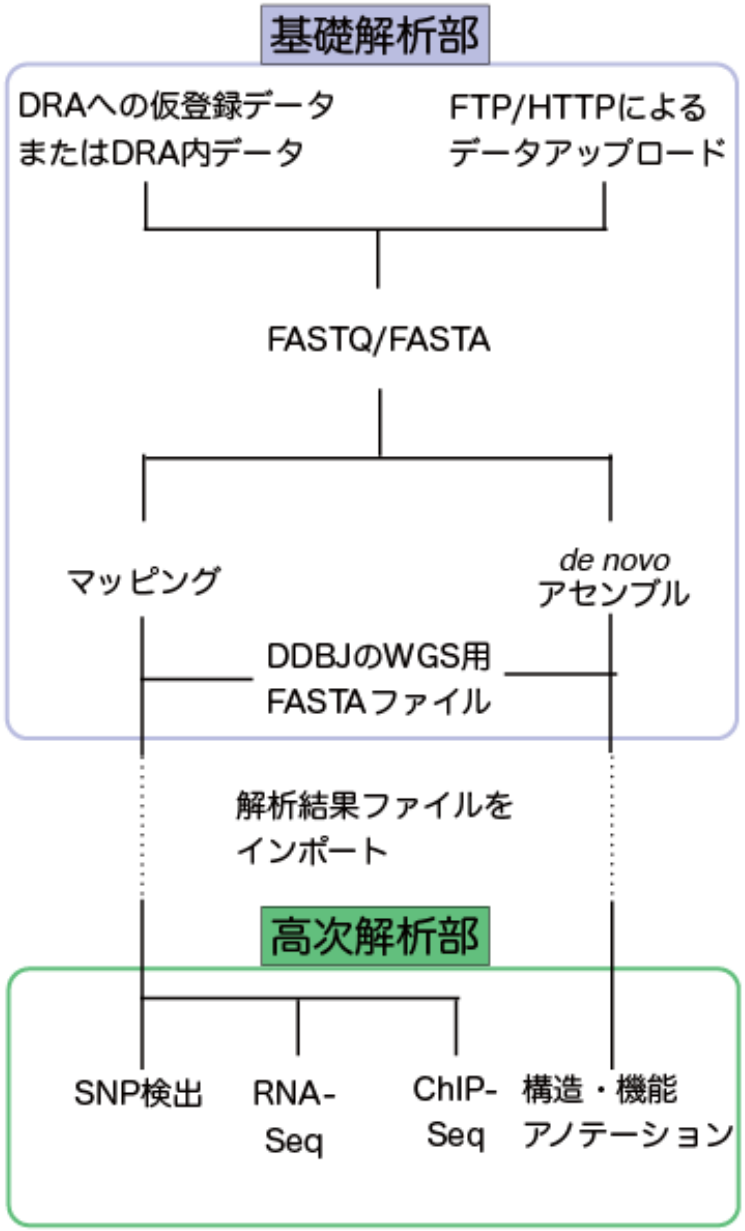
遺伝研スーパーコンピュータ

- 計算ノード
  - 64GB memory x 554 nodes
  - 2TB memory x 10 nodes
  - 10TB memory x 1 node
- Storage
  - 7 PB 高速HDD
  - 5.5 PB 大容量省電力HDD
- 利用申請することで無料で利用できます！
  - コマンドラインが基本
  - （停電で結構頻繁に落ちる）

スパコン利用申請はこちら

いや、
そうぢゃない

DDBJ解析パイプライン



DDBJ Read Annotation Pipeline

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

Galaxy / P-GALAXY

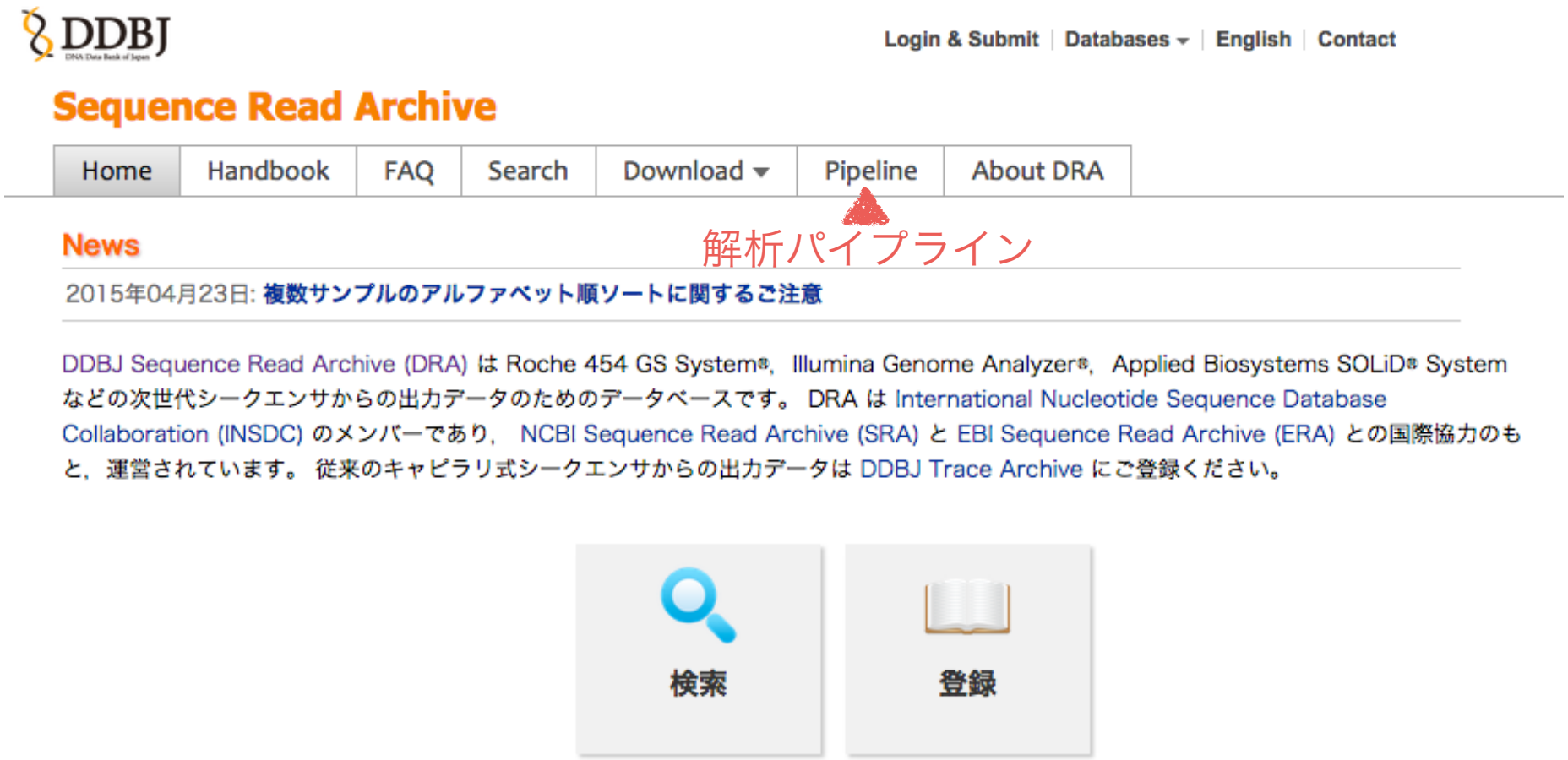
https://p-galaxy.ddbj.nig.ac.jp/

DDBJのスパコンをウェブ経由で利用

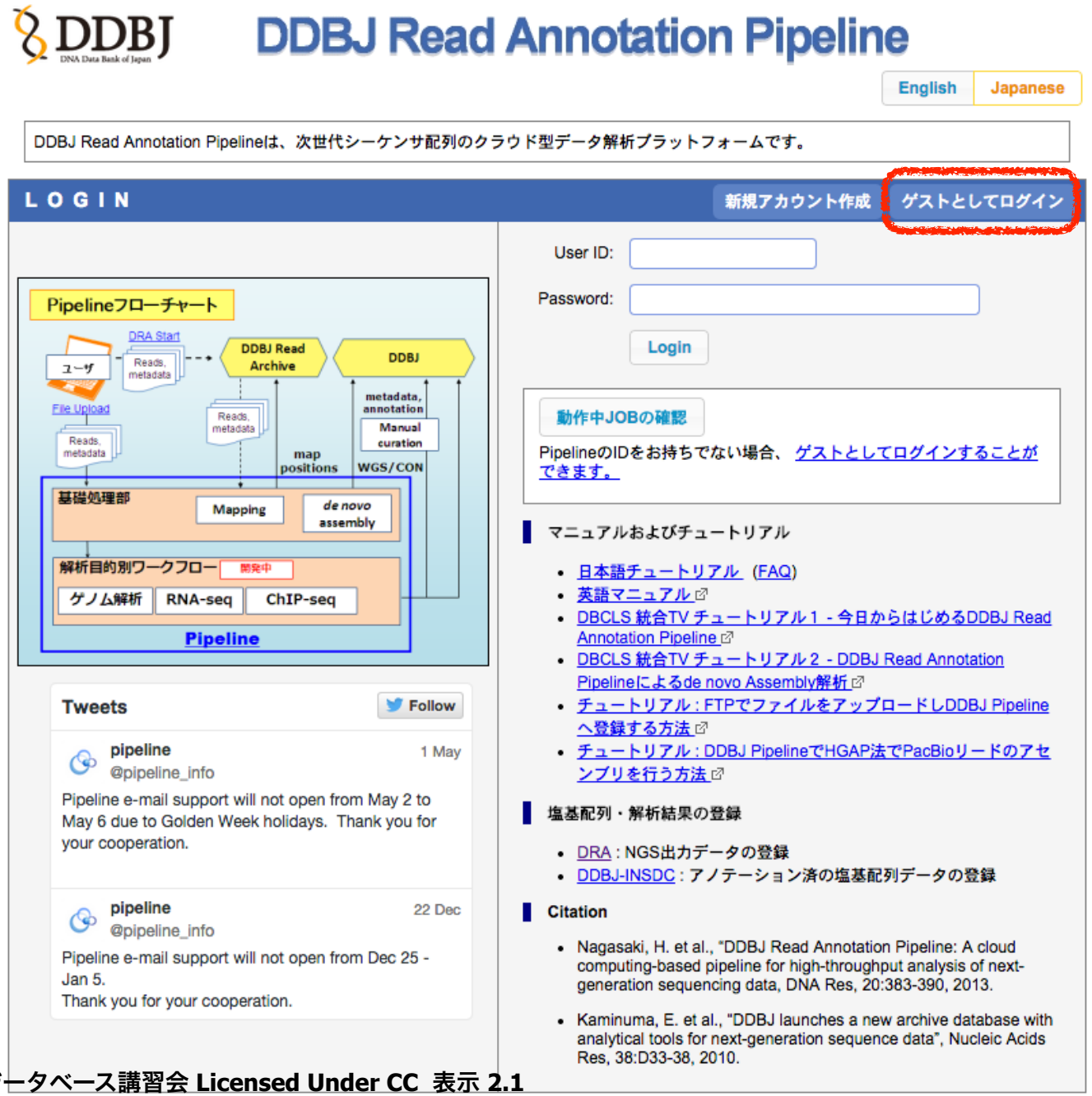
実習

DDBJ Read Annotation Pipeline
を使って次世代シーケンスのデータを解析する

DRAウェブサイト http://trace.ddbj.nig.ac.jp/dra/



DDBJ Read Annotation Pipeline

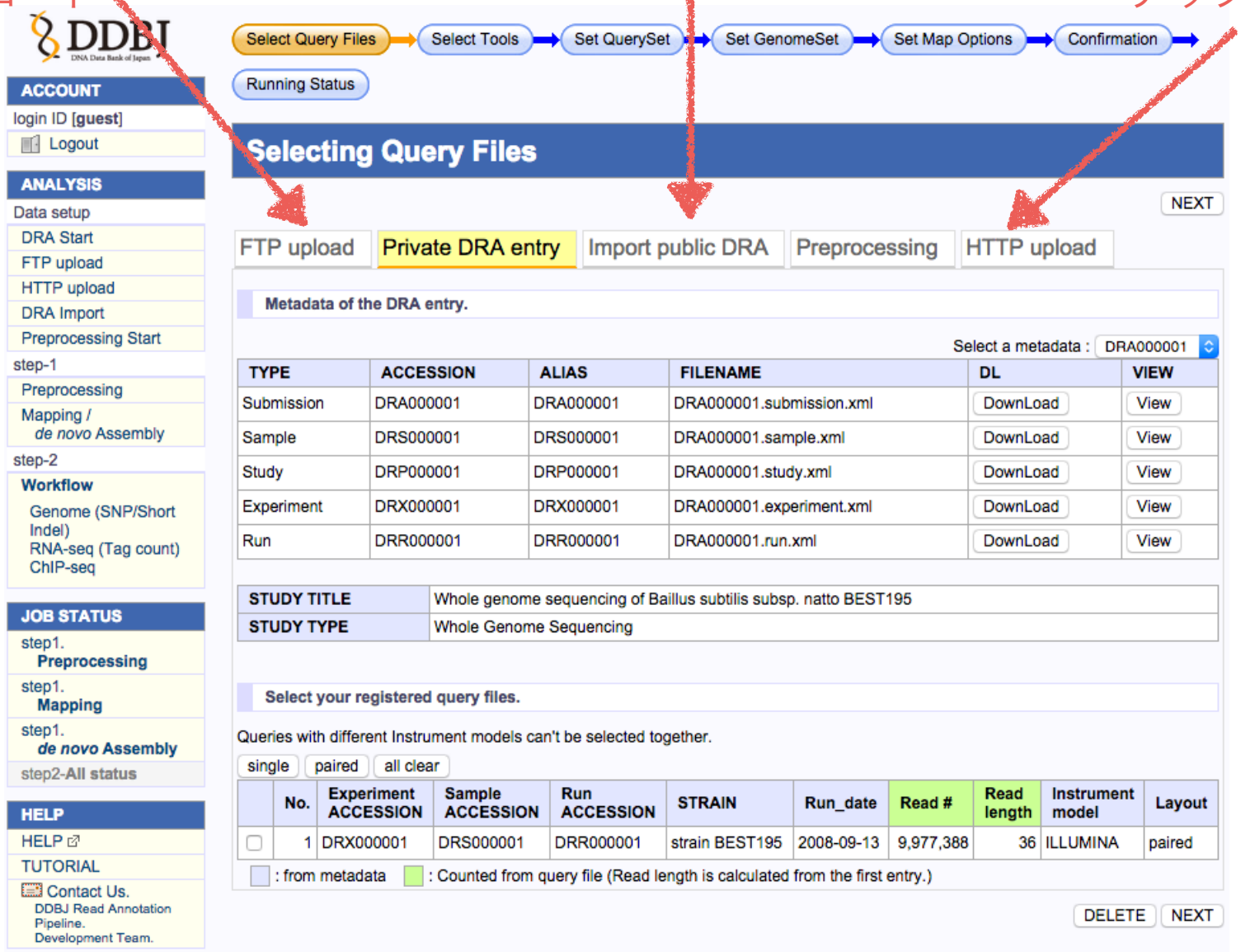


データセットの選択

FTPで手持ちのデータを
アップロード

DRAからデータをインポート

HTTPで手持ちのデータを
アップロード

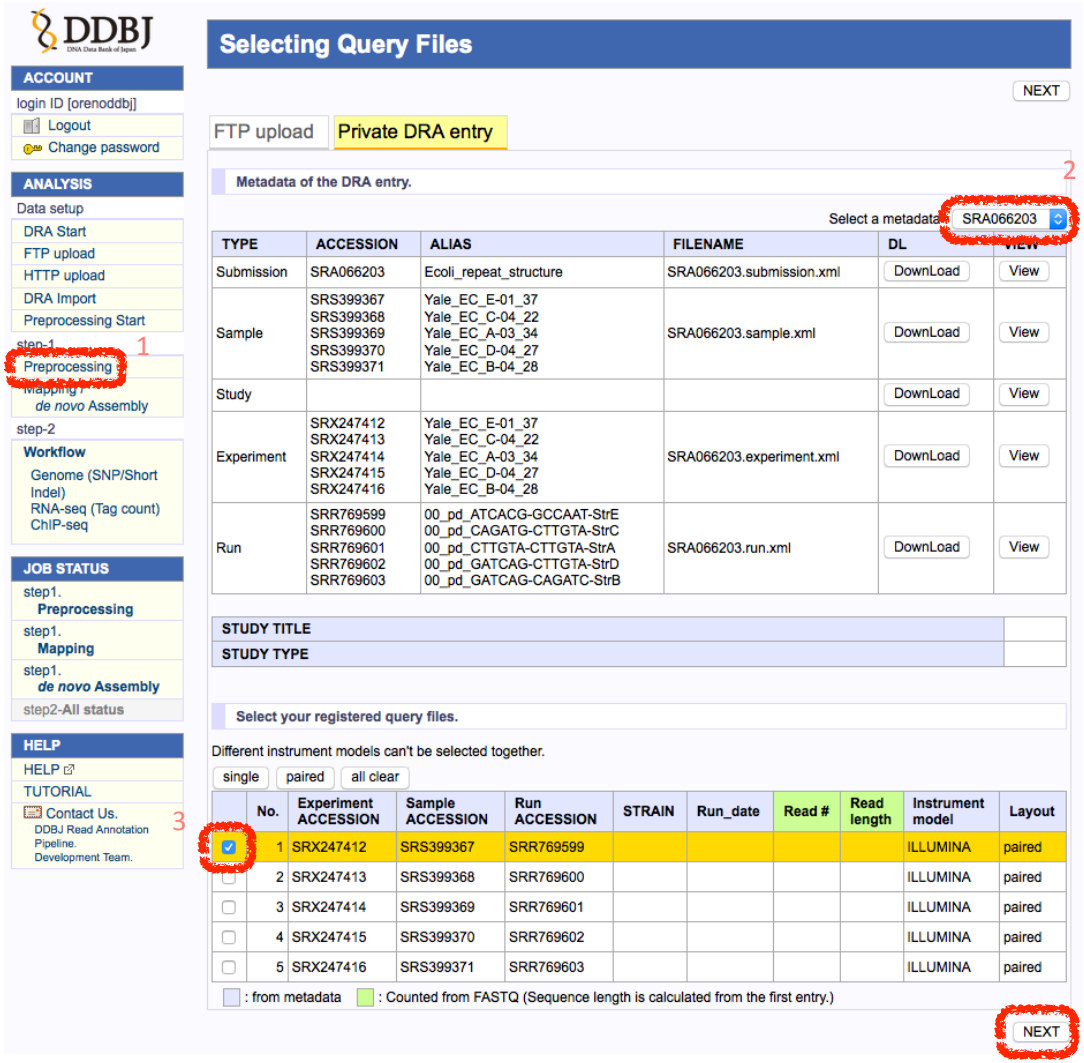


SRA/DRAデータのインポート

操作1

配列のクオリティをチェックする

インポートされたデータ



SRA066203 を選択

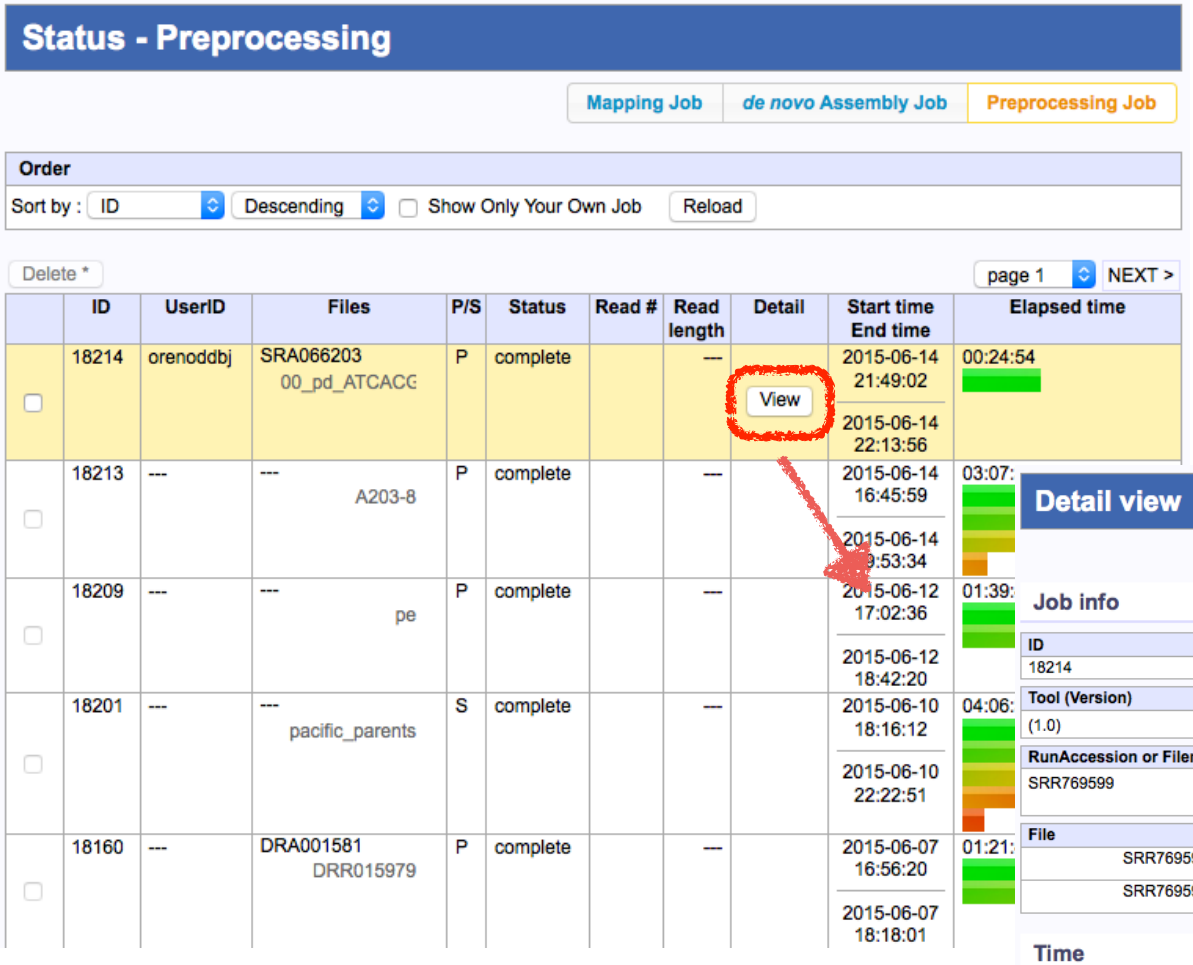
画面右の「Preprocessing」から、
配列のクオリティーなどをチェックできる

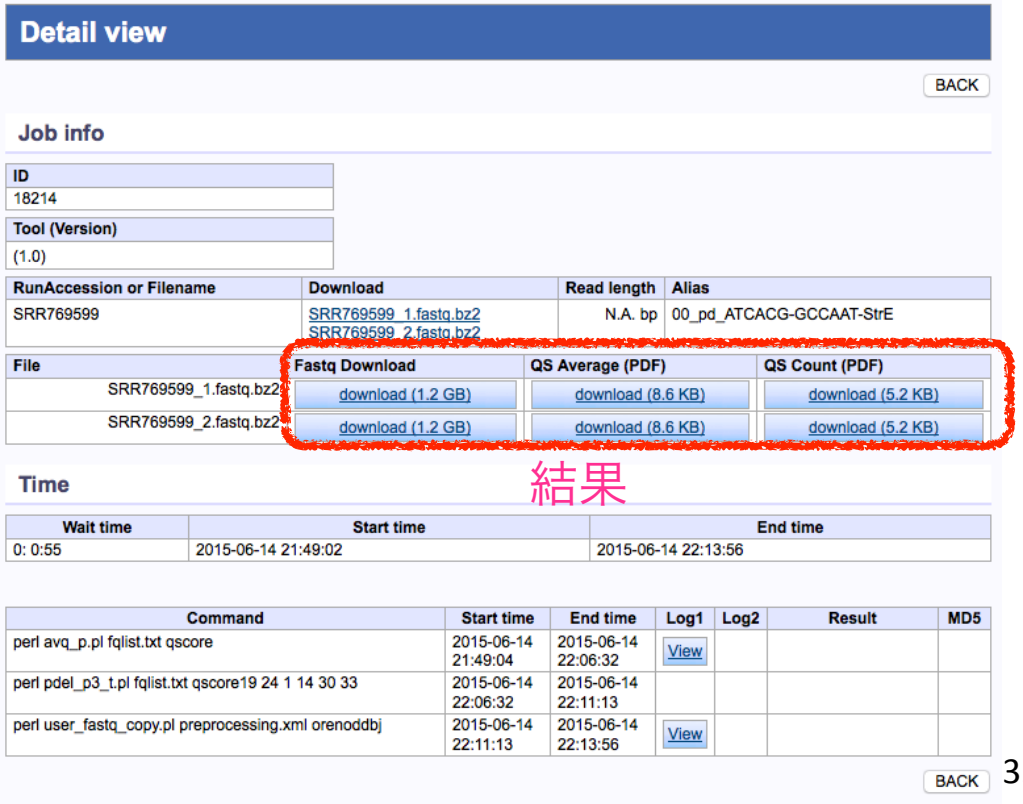
Preprocessing -> 計算対象ランデータを選択 -> NEXT

Preprocessing（クオリティチェック）

※ログインしないと実行できません（Runボタンがでてきません）

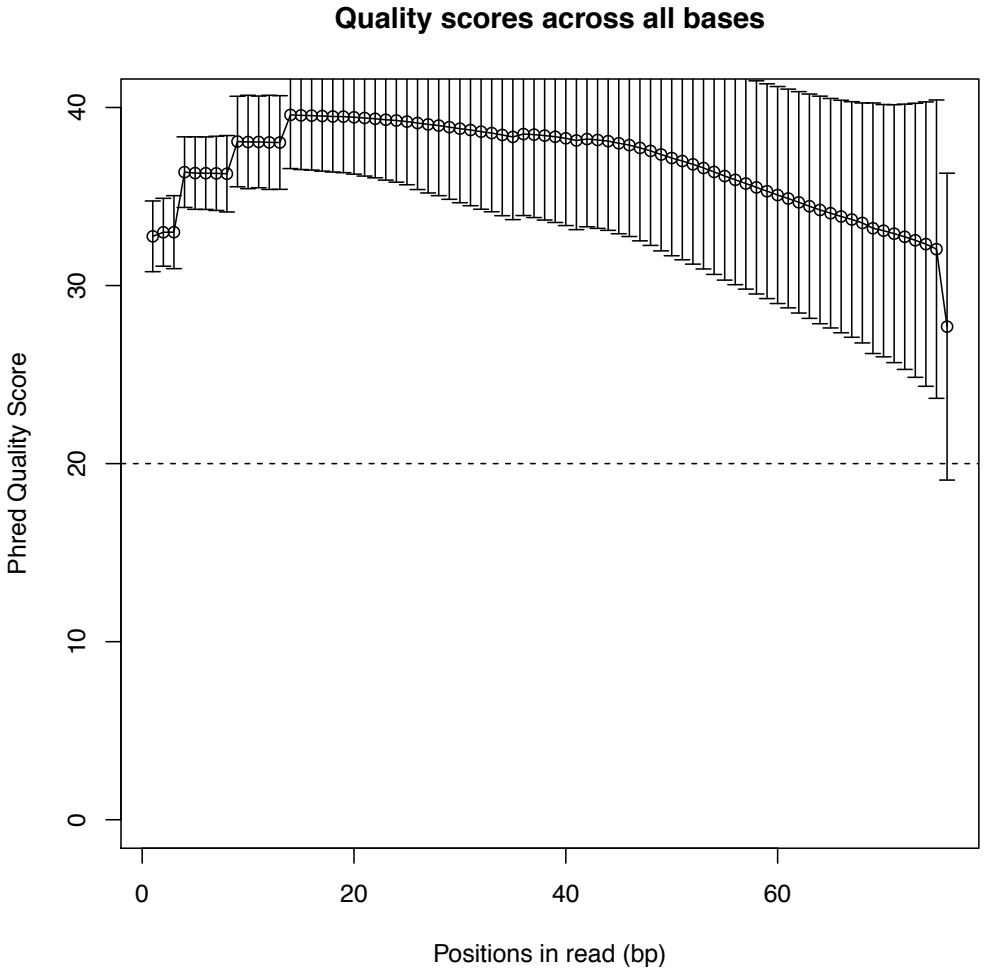
進捗確認画面-Preprocessing

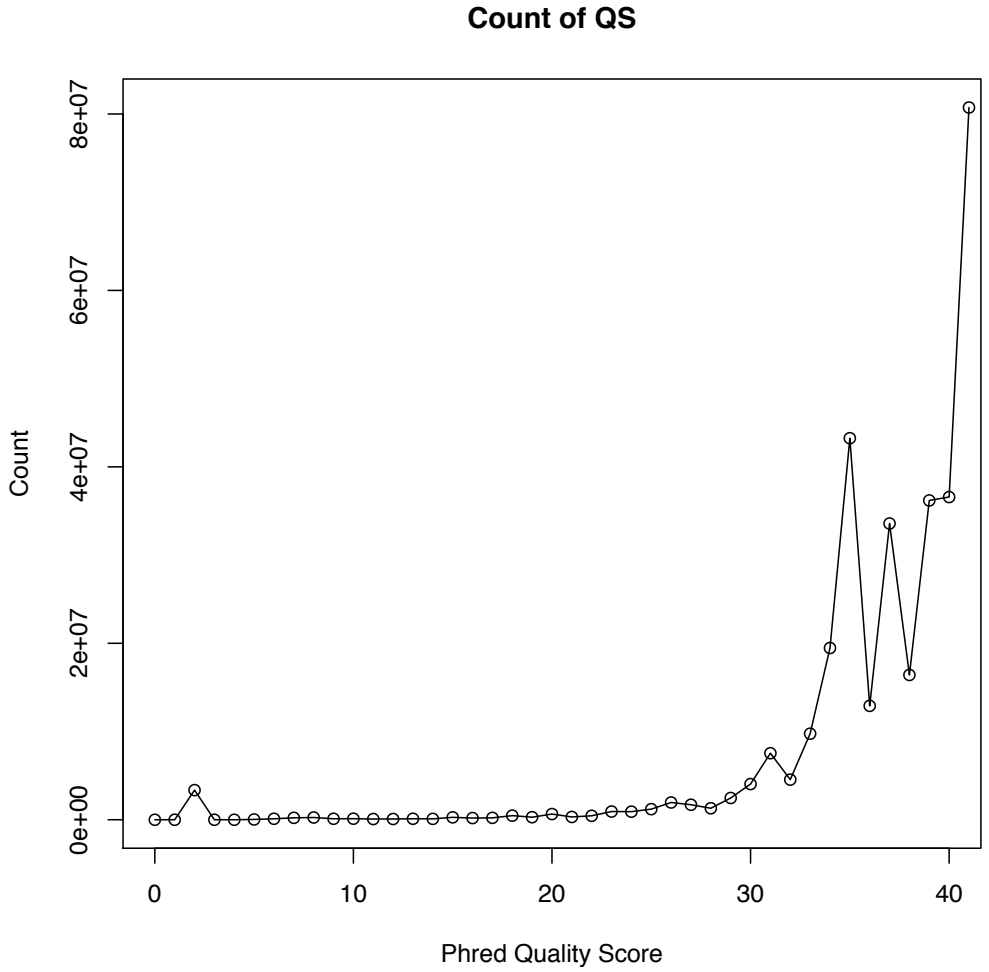




背景無色（白）が他のユーザ
背景黄色が自分のジョブ

Preprocessing 結果





塩基部位ごとのQuality score分布

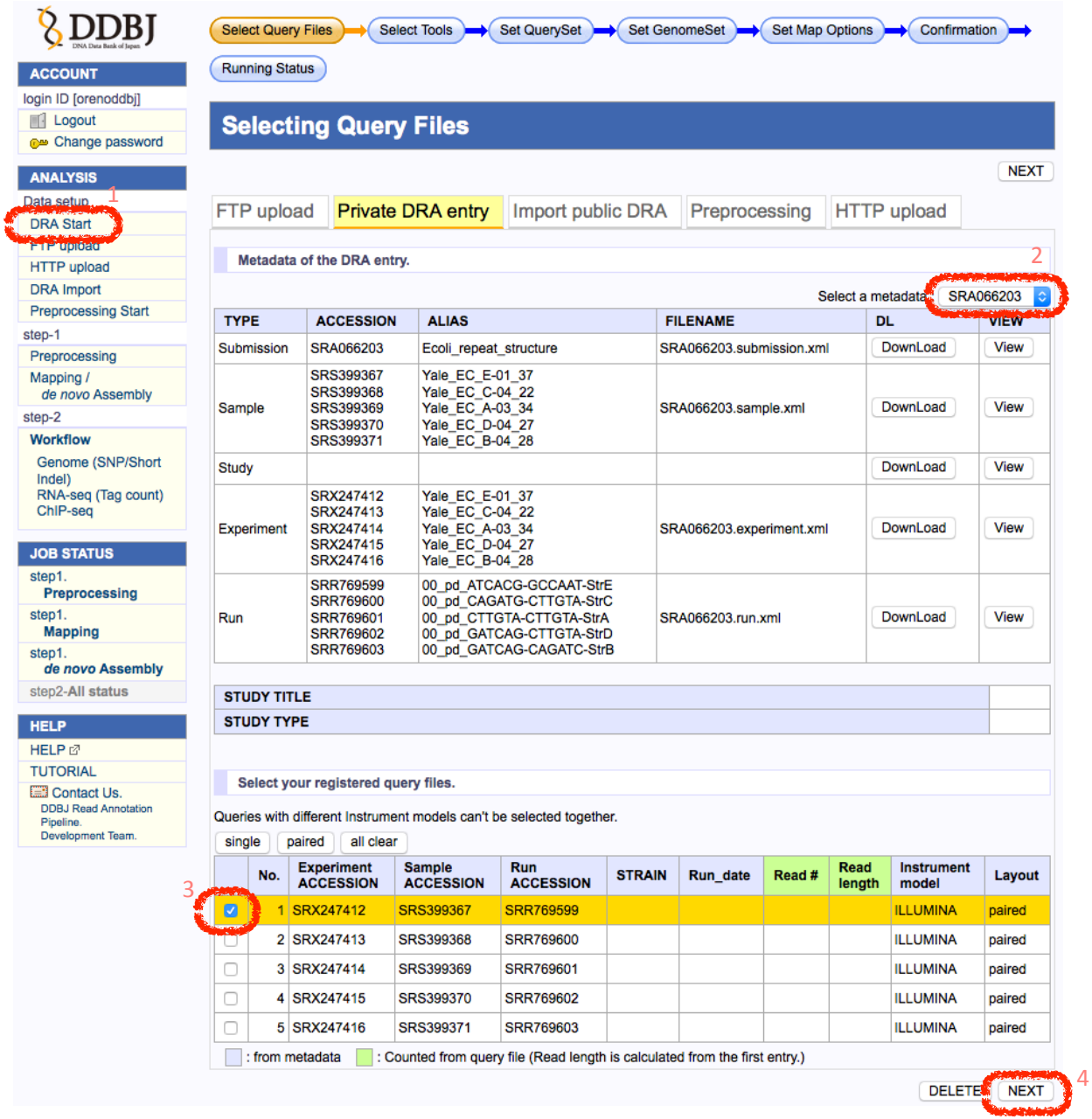
スコアの分布

どの塩基まで解析に使用するかの目安になる

操作2

配列データを既知のゲノム配列にマッピングする

パイプラインの実行〜マッピング



DRA Startで初期画面に戻る

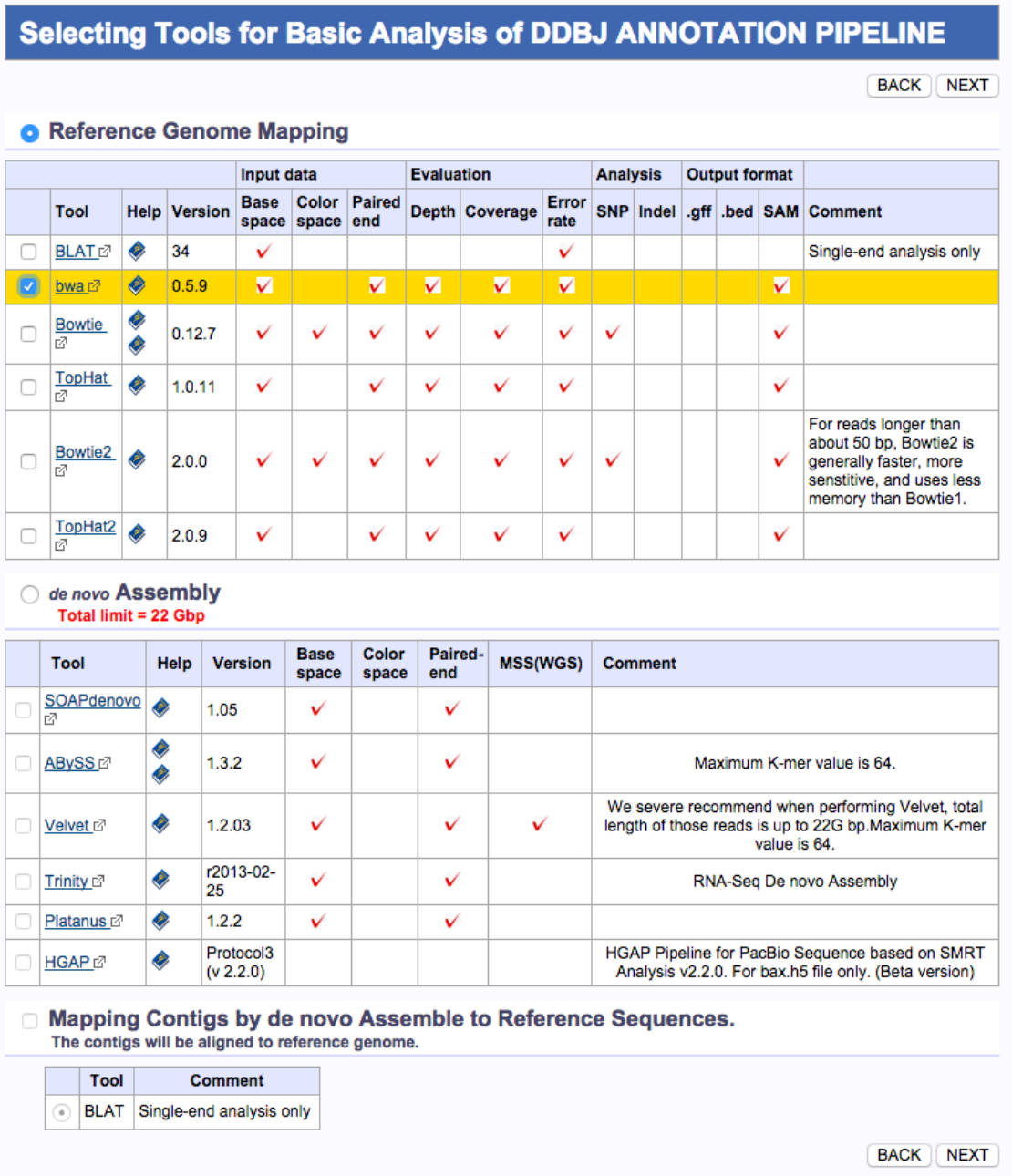
計算対象ランデータを選択 -> NEXT

SRA066203 を選択

解析ツールの選択～マッピング

マッピング系ツール

アセンブリ系ツール



Selecting Tools for Basic Analysis of DDBJ ANNOTATION PIPELINE

BACK NEXT

Reference Genome Mapping

	Tool	Help	Version	Input data: Base space	Input data: Color space	Input data: Paired end	Evaluation: Depth	Evaluation: Coverage	Evaluation: Error rate	Analysis: SNP	Analysis: Indel	Output format: .gff	Output format: .bed	Output format: SAM	Comment
☐	BLAT		34	✓					✓						Single-end analysis only
☑	bwa		0.5.9	✓		✓	✓	✓	✓					✓	
☐	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.
☐	TopHat2		2.0.9	✓		✓	✓	✓	✓					✓	

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		r2013-02-25	✓		✓		RNA-Seq De novo Assembly
☐	Platanus		1.2.2	✓		✓		
☐	HGAP		Protocol3 (v 2.2.0)					HGAP Pipeline for PacBio Sequence based on SMRT Analysis v2.2.0. For bax.h5 file only. (Beta version)

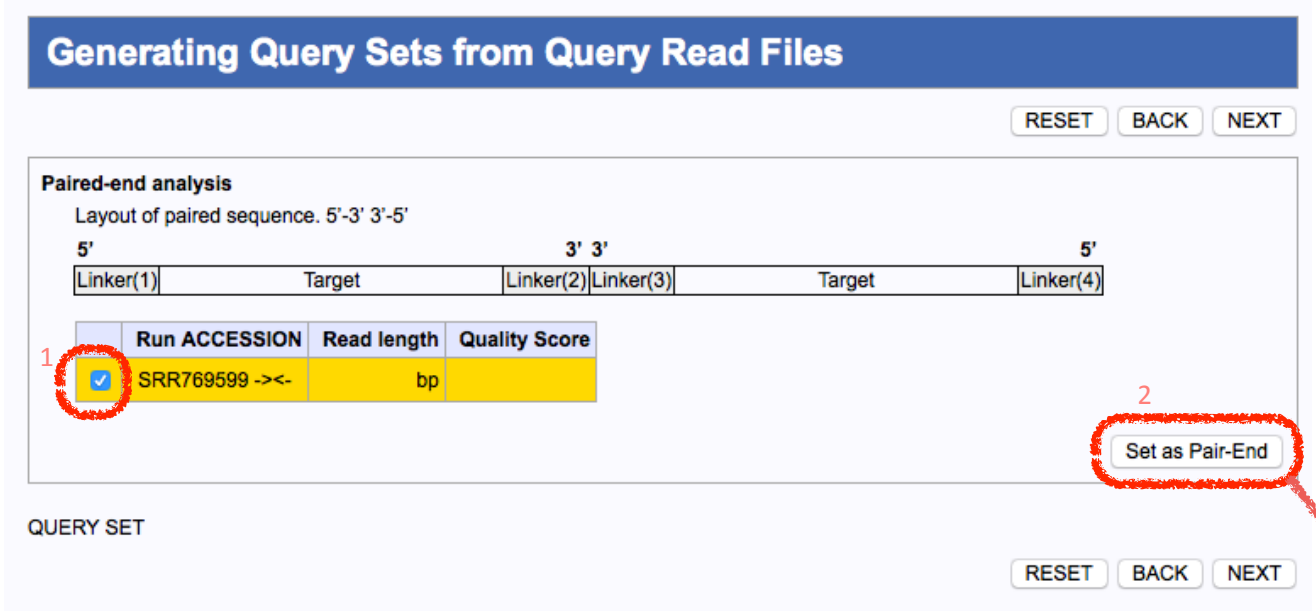
Mapping Contigs by de novo Assemble to Reference Sequences.

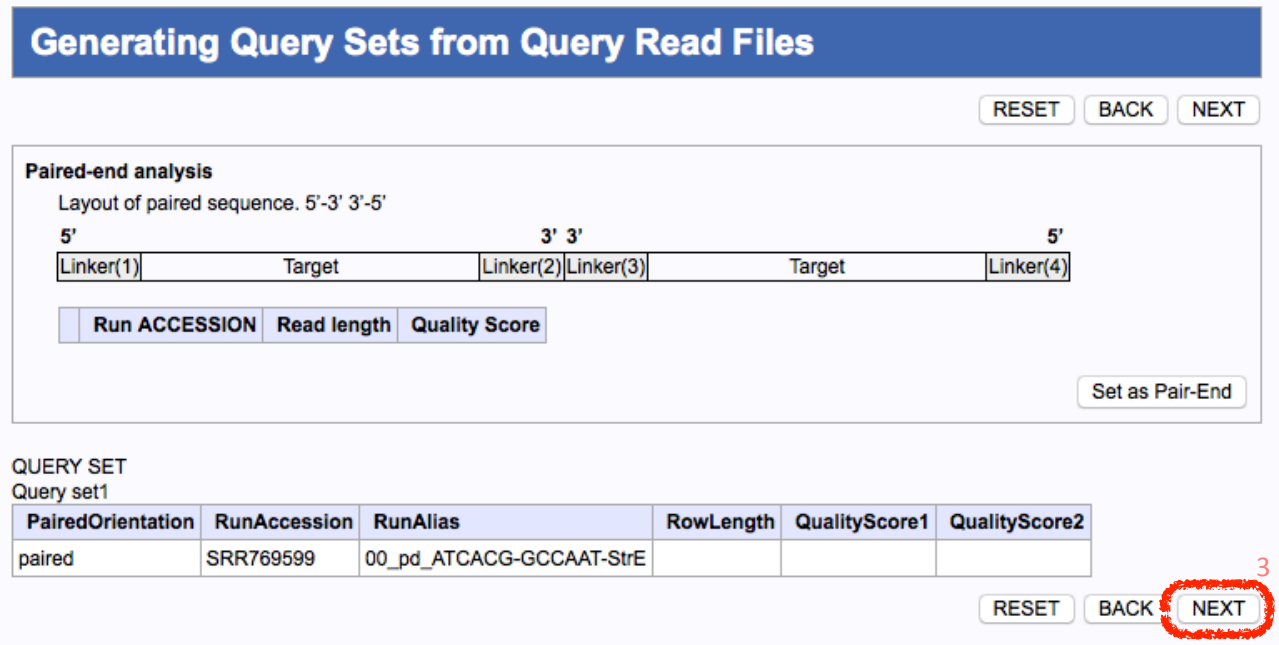
The contigs will be aligned to reference genome.

	Tool	Comment
◉	BLAT	Single-end analysis only

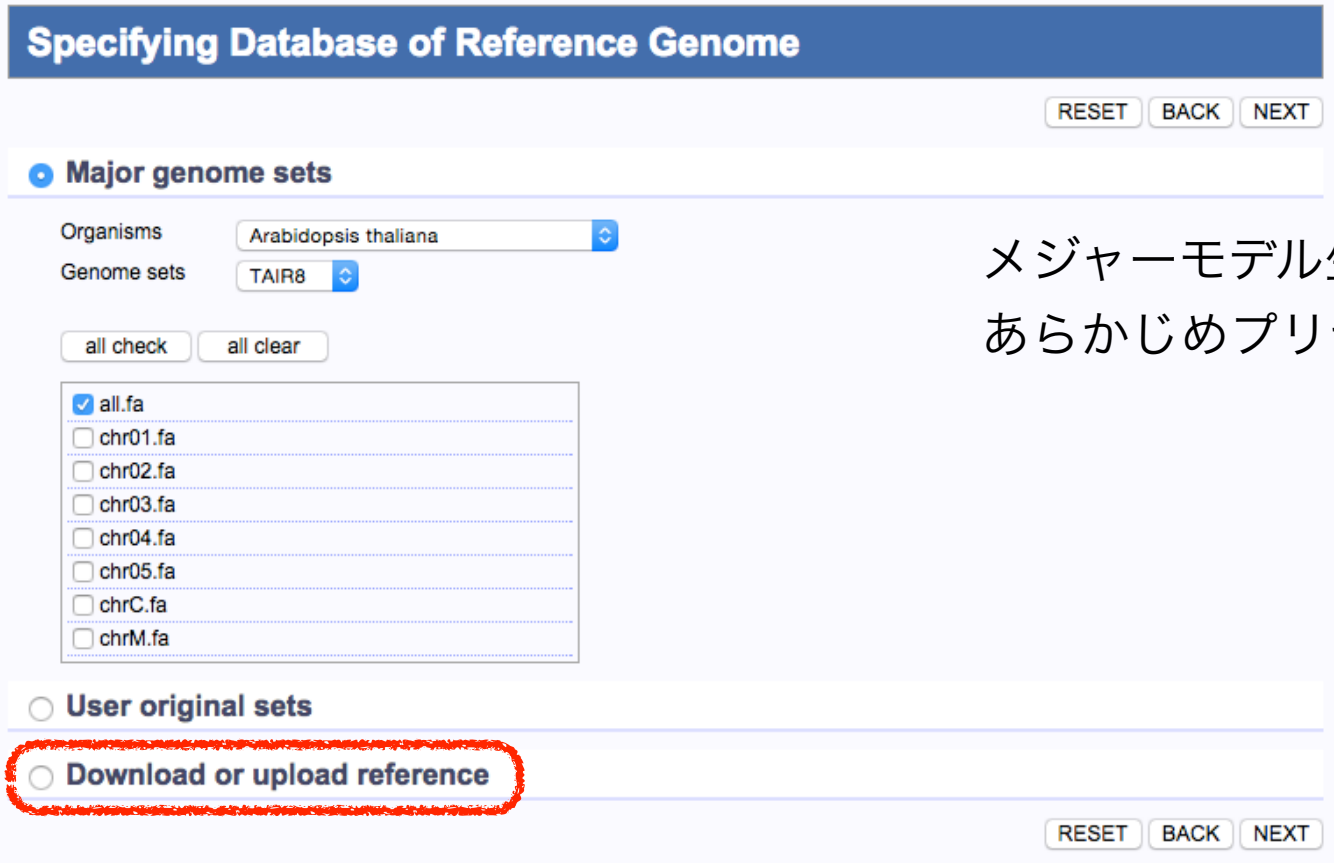
BACK NEXT

解析するランの指定





リファレンスゲノムの指定



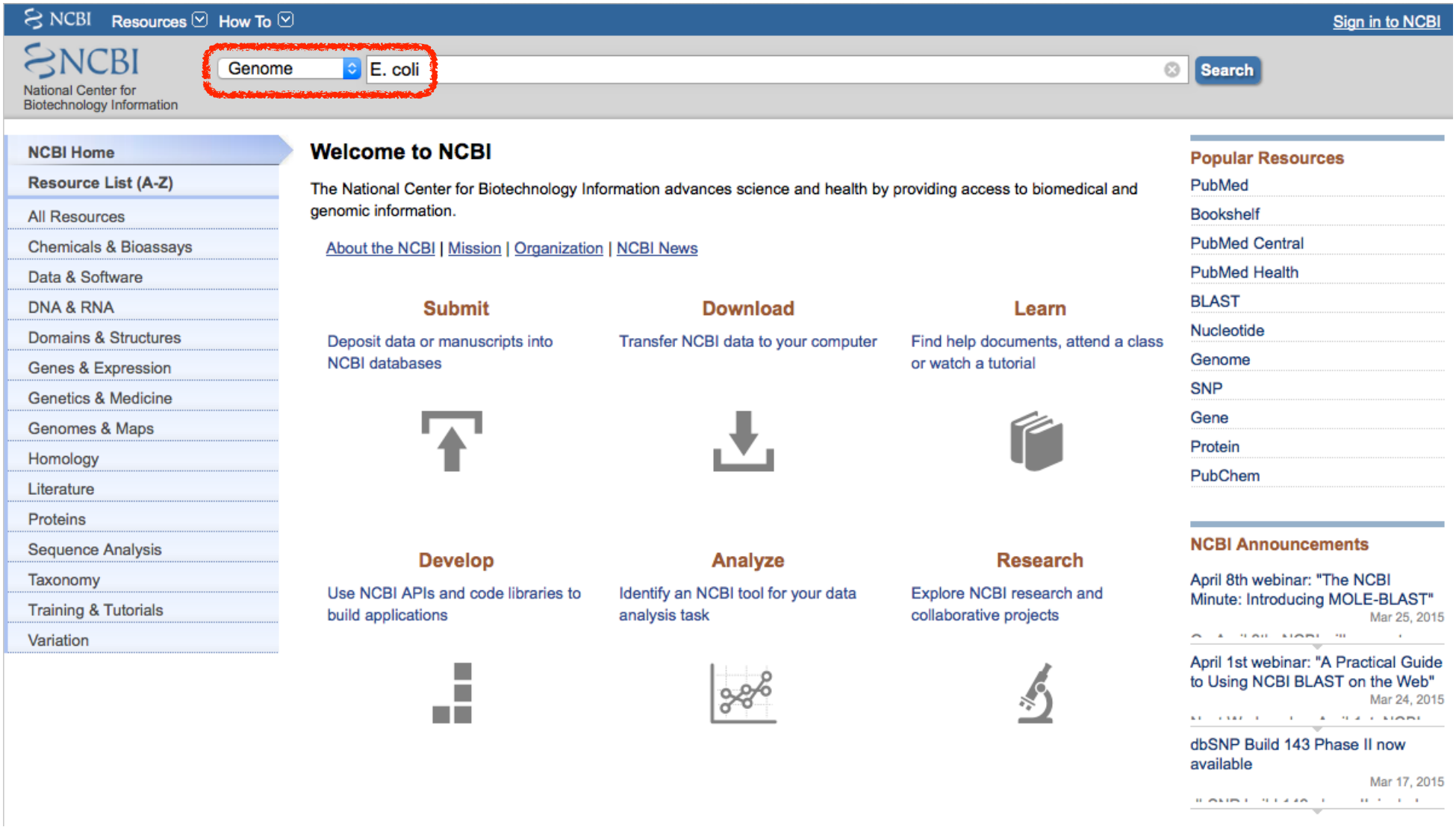
メジャーモデル生物のリファレンスゲノムは
あらかじめプリセットされている

ヒト
マウス
線虫
イネ
シロイヌナズナ
酵母
…

その他のリファレンスゲノムは
・自分でゲノム配列をアップロードする
・DDBJからインポートする

NCBIページ - 大腸菌のゲノムIDを調べる

ゲノムデータベースを選択して大腸菌を検索



http://www.ncbi.nlm.nih.gov/

NCBIゲノムページ

NCBI Resources How To

Genome

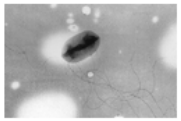
Genome Search

Create alert Limits Advanced

Escherichia coli
Reference genome: Escherichia coli str. K-12 substr. MG1655
Download sequences in FASTA format for **genome**, **protein**
Download genome annotation in **GFF**, **GenBank** or **tabular** format
BLAST against Escherichia coli **genome**, **protein**
All 3071 genomes for species:
Browse the **list**
Download sequence and annotation from **RefSeq** or **GenBank**

Display Settings: Overview

Organism Overview ; Genome Assembly and Annotation report [**3071**] ; Genome Grou



Escherichia coli

A well-studied enteric bacterium

Lineage: Bacteria[6344]; Proteobacteria[2198]; Gammaproteobacteria[953]; Escherichia coli[1]

Escherichia coli. This organism is typically present in the lower intestine of humans, wh only one minor constituent of the complete intestinal microflora. E.coli is easily grown in manipulation making it one of the most More...

Sequence data: genome assemblies: 3071; sequence reads: 135 (See Genome Assembly genome groups: 31 (See Genome Groups report)

Publications

1. Bacterial resistance to leucyl-tRNA synthetase inhibitor GSK2251052 develo infections. O'Dwyer K, et al. Antimicrob Agents Chemother 2015 Jan
2. First report of a clinical, multidrug-resistant Enterobacteriaceae isolate coha carbapenemase gene blaKPC-2 on the same transposon, Tn1721. Li G, et a
3. Large-scale genomic sequencing of extraintestinal pathogenic Escherichia coli strains. Salipante SJ, et al. Genome Res 2015 Jan

More...

Representative (genome information for reference and representative genomes)

Dendrogram (based on genomic BLAST)

1-182-04_S3_C2 (represents 3 genomes)
KTE102 (represents 3 genomes)

Representative (genome information for reference and representative genomes)

Reference genomes: [see all organisms]

- ***Escherichia coli str. K-12 substr. MG1655***
  Submitter: Univ. Wisconsin
  Morphology: Gram:Negative, Shape:Bacilli, Motility:Yes
  Environment: OxygenReq:Facultative, OptimumTemperature:37, TemperatureRange:Mesophilic, Habitat:HostAssociated

Type	Name	RefSeq	INSDC	Size (Mb)	GC%	Protein	rRNA	tRNA	Other RNA	Gene	Pseudogene
Chr	-	NC_000913.3	U00096.3	4.64	50.8	4,140	22	89	67	4,498	184

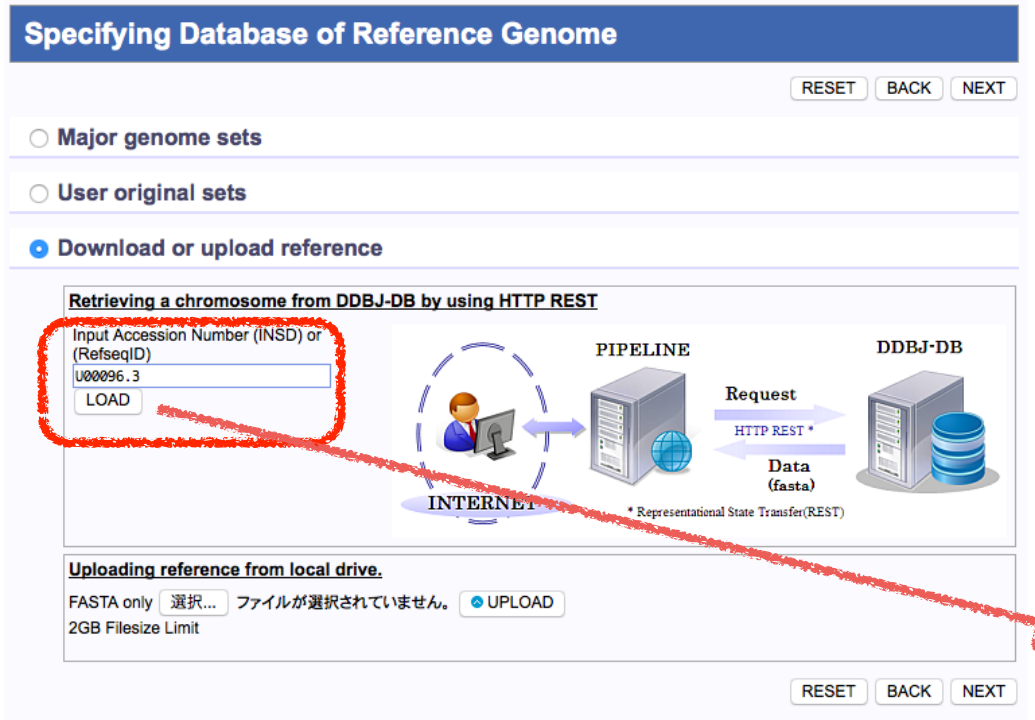
- ***Escherichia coli O157:H7 str. Sakai***
  Submitter: GIRC
  Human Pathogen
  Morphology: Gram:Negative, Shape:Bacilli, Motility:Yes
  Environment: OxygenReq:Facultative, OptimumTemperature:37, TemperatureRange:Mesophilic, Habitat:HostAssociated
  Phenotype: Disease:Hemorrhagic colitis

Type	Name	RefSeq	INSDC	Size (Mb)	GC%	Protein	rRNA	tRNA	Other RNA	Gene	Pseudogene
Chr	-	NC_002695.1	BA000007.2	5.5	50.5	5,204	22	105	14	5,360	14
Plsm	pOSAK1	NC_002127.1	AB011548.2	0.003306	43.4	3	-	-	-	3	-
Plsm	pO157	NC_002128.1	AB011549.2	0.092721	47.6	85	-	-	-	85	-

- ***Escherichia coli IAI39***
  Submitter: Genoscope
  Human Pathogen
  Morphology: Gram:Negative, Shape:Bacilli, Motility:No
  Environment: OxygenReq:Facultative, TemperatureRange:Mesophilic, Habitat:Multiple
  Phenotype: Disease:Urinary tract infection

Type	Name	RefSeq	INSDC	Size (Mb)	GC%	Protein	rRNA	tRNA	Other RNA	Gene	Pseudogene
Chr	-	NC_011750.1	CU928164.2	5.13	50.6	4,725	22	88	97	5,092	160

リファレンスゲノムデータのインポート

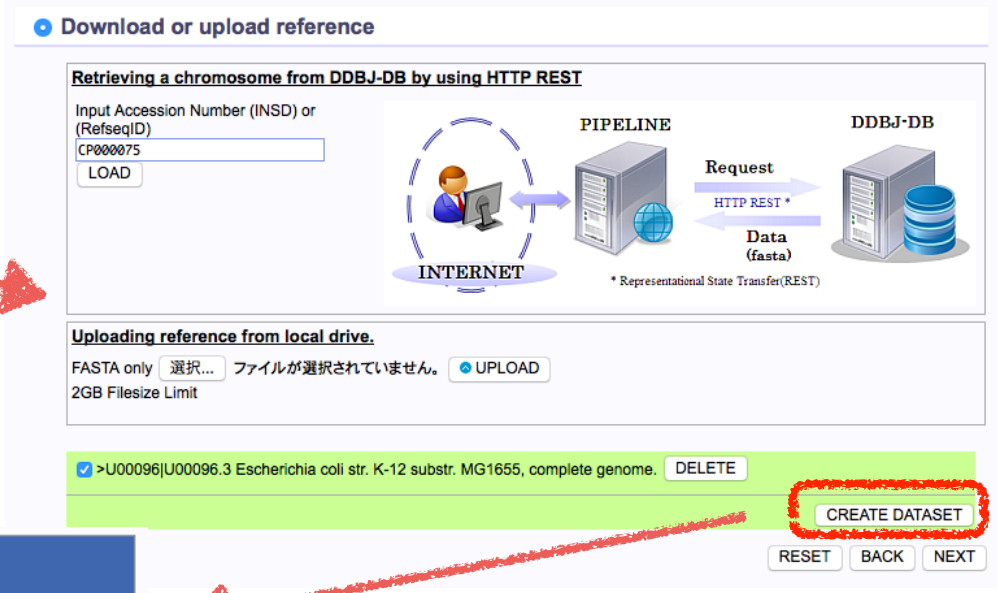


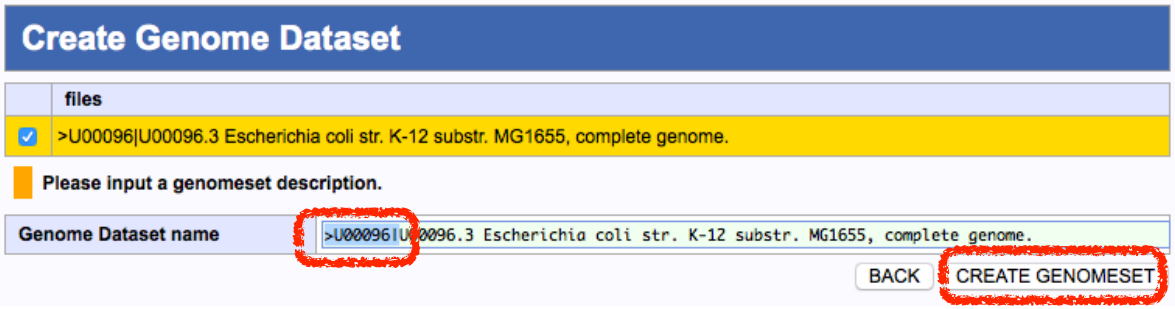
Escherichia coli のINSD IDを

→ 入力、LOAD

→ CREATE DATASET

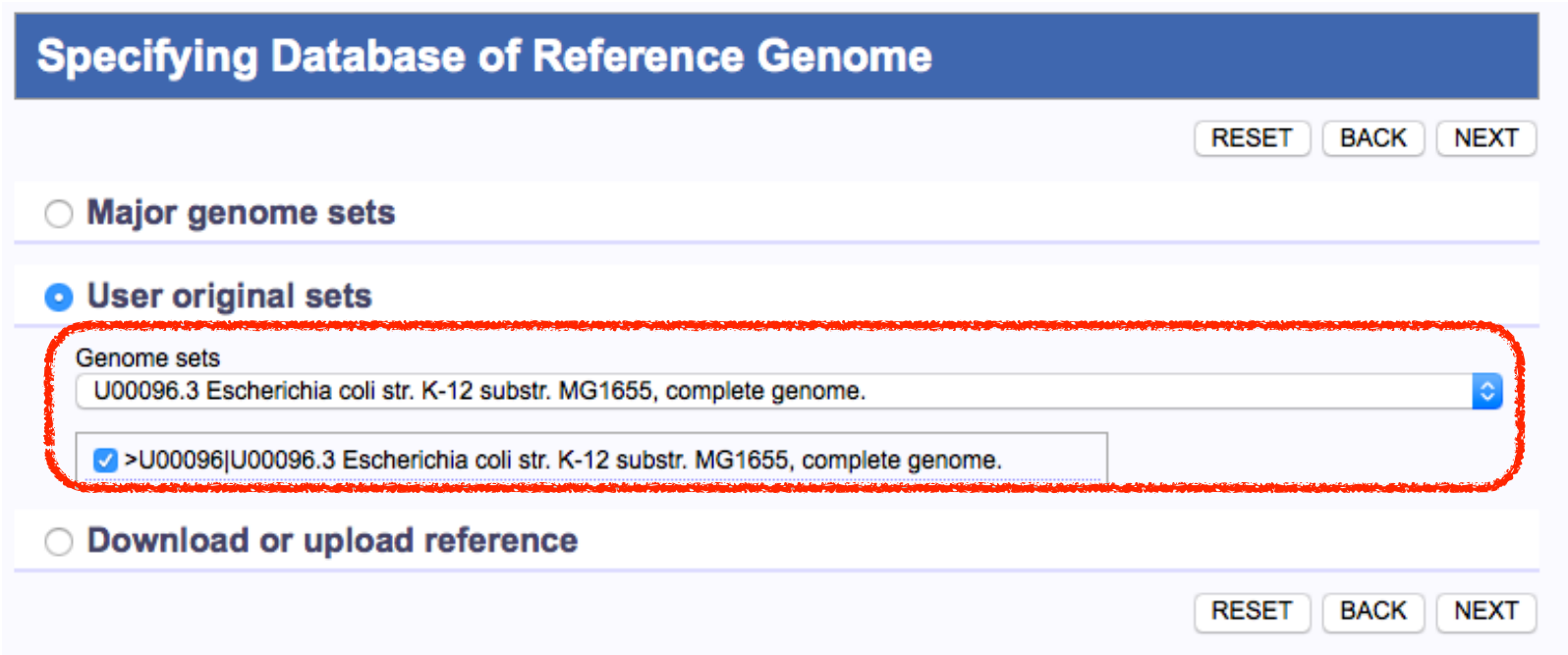
→ CREATE GENOMESET
でリファレンスゲノムを登録する





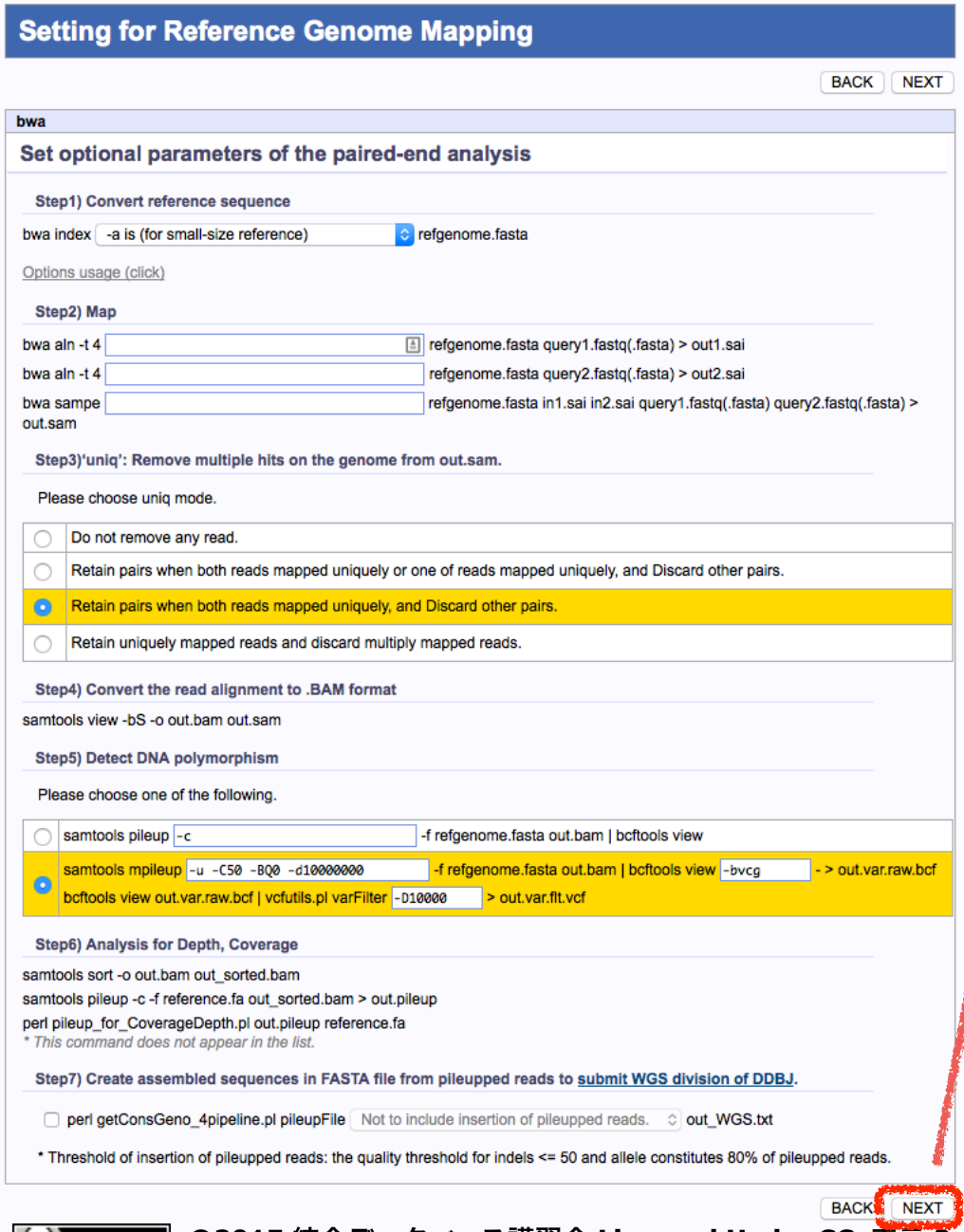
Genome Datase name の
縦棒｜より左側を削除
（エラーになる）

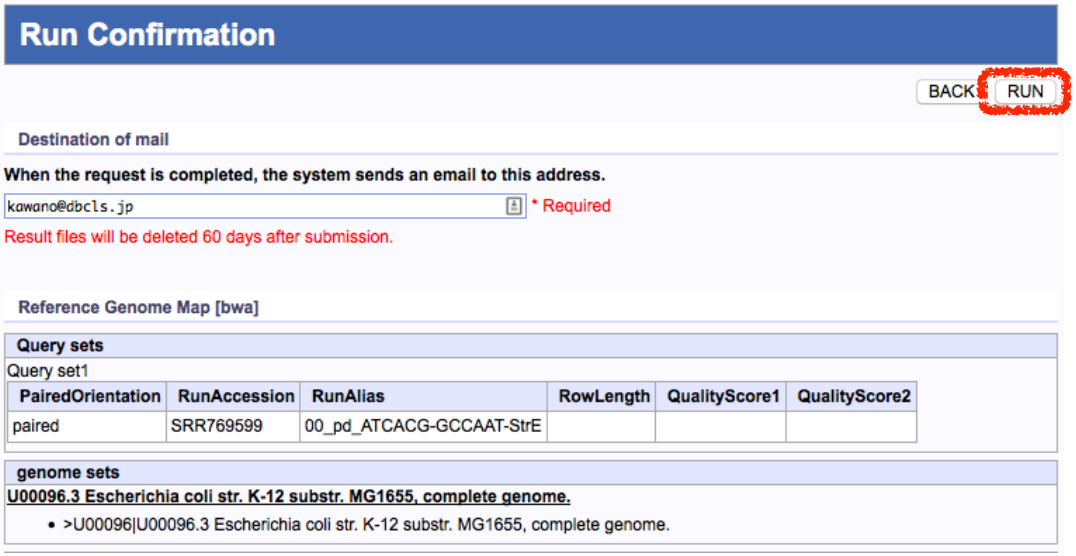
リファレンスゲノムの選択



先ほど登録したゲノムを選択

パラメータ設定、確認、実行



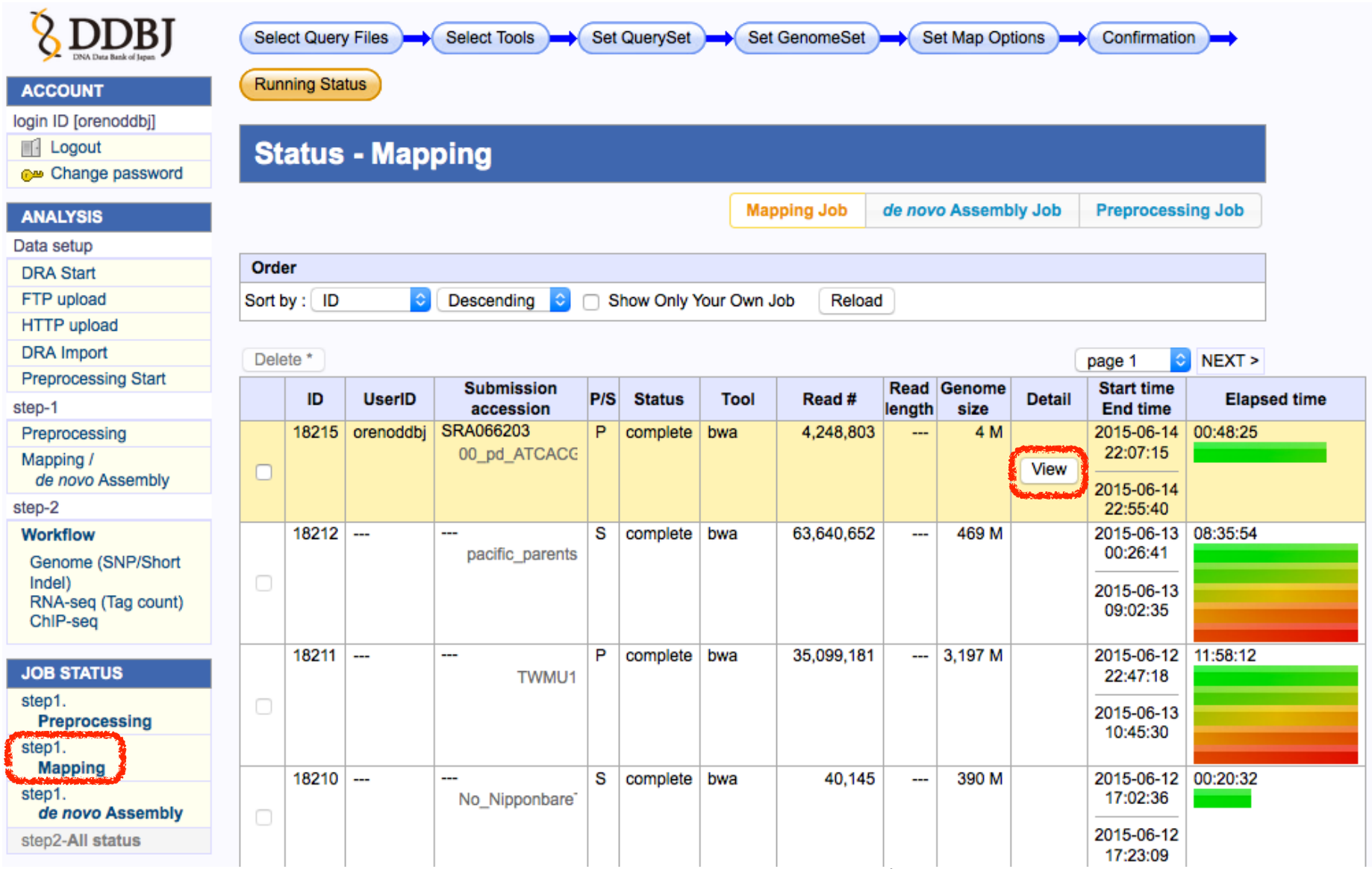


計算に時間がかかるのでメールアドレスの登録は必須

ログインしないと実行できません

（Runボタンが出てきません）

進捗確認画面-Mapping



背景無色（白）が他のユーザ
背景黄色が自分のジョブ

マッピング結果

Detail view

BACK

Job info

ID
18215

Tool (Version)
bwa (0.5.9)

RunAccession or Filename	Download	Read length	Alias
SRR769599	SRR769599_1.fastq.bz2 SRR769599_2.fastq.bz2	N.A. bp	00_pd_ATCACG-GCCAAT-StrE

Genome set

Chromosome
U00096.3_150614220106489

Download modified queries

- SRR769599_1.fastq.gz (Original size 1.2 GB)
- SRR769599_2.fastq.gz (Original size 1.2 GB)

結果ファイル

Download merged pileup file

Uniq.sam files have been merged if you specified 'uniq' option.

- merged.var.flt.vcf.gz (Original size 9.6 MB)
- merged.sam.gz (Original size 1.3 GB)

Download wgs file

- out_WGS.fasta.gz (Original size 1.5 MB)

Position errors	Map ratio	Depth, Coverage
PDF download	total query # : 4,248,803 mapped query # : 2,741,734 map ratio : 64.530 %	coverage : 4105488 / 4641652 * 100 = 88.449 depth : 406246179 / 4105488 = 98.952

統計量

Time

Wait time	Start time	End time
0: 0:56	2015-06-14 22:07:15	2015-06-14 22:55:40

U00096.3_150614220106489

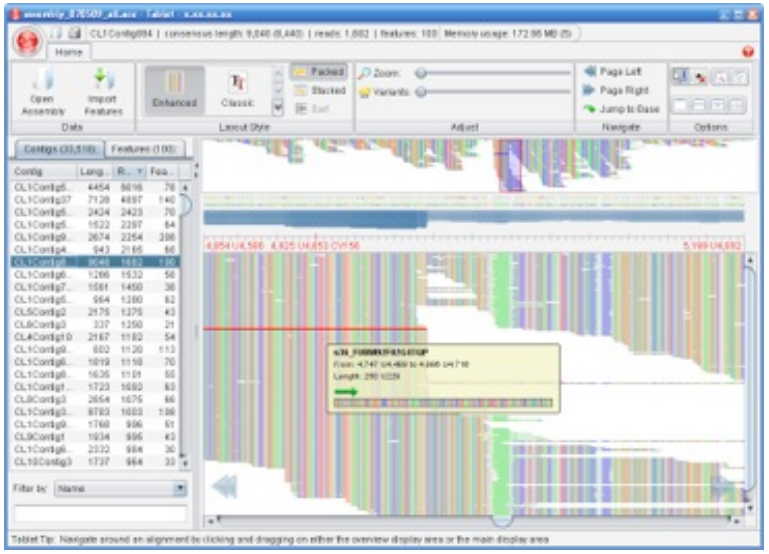
Command	Start time	End time	Log1	Log2	Result	MD5
Create BWA Index File bwa index [-a is] U00096.3_150614220106489	2015-06-14 22:07:15	2015-06-14 22:07:25		View		
BWA : Alignment bwa aln U00096.3_150614220106489 SRR769599_1.fastq > 1.sai	2015-06-14 22:07:25	2015-06-14 22:08:16		View		
BWA : Alignment bwa aln U00096.3_150614220106489 SRR769599_2.fastq > 2.sai	2015-06-14 22:08:17	2015-06-14 22:09:18		View		
BWA : SAMPE bwa sampe U00096.3_150614220106489 1.sai 2.sai SRR769599_1.fastq SRR769599_2.fastq > out.sam	2015-06-14 22:09:18	2015-06-14 22:12:04		View	Download(627.6 MB)	MD5
Extract Unmapped Reads python extractUnmappedFASTQ.py SRR769599_1.fastq SRR769599_2.fastq out.sam	2015-06-14 22:15:56	2015-06-14 22:17:18			Download(213.7 MB)	MD5
Convert SAM to BAM samtools view -bS -o out.bam out.sam	2015-06-14 22:19:52	2015-06-14 22:22:05		View	Download(649.7 MB)	MD5
Sort BAM File samtools sort out.bam out2	2015-06-14 22:22:36	2015-06-14 22:25:12		View	Download(485.6 MB)	MD5
Create BAM Index File samtools index out2.bam	2015-06-14 22:25:52	2015-06-14 22:26:13			Download(5.9 KB)	MD5
Uniquify SAM (Remove Multiple Hits) perl sam2uniq.pl out.sam UBE > uniqout.sam	2015-06-14 22:26:23	2015-06-14 22:27:38			Download(340.3 MB)	MD5
Convert SAM to BAM [For Unique SAM] samtools view -bS -o uniqout.bam uniqout.sam	2015-06-14 22:29:41	2015-06-14 22:30:53		View	Download(694.1 MB)	MD5
Sort BAM File [For Unique SAM] samtools sort uniqout.bam out2	2015-06-14 22:33:16	2015-06-14 22:34:41		View	Download(239.0 MB)	MD5
Create BAM Index File [For Unique SAM] samtools index out2.bam	2015-06-14 22:35:02	2015-06-14 22:35:12			Download(5.2 KB)	MD5
Mpileup and Create BCF File [For Unique SAM] samtools mpileup -u -C50 -BQ0 -d10000000 -f U00096.3_150614220106489 out2.bam | bcftools view -bvcg - > uniq.var.bcf	2015-06-14 22:35:24	2015-06-14 22:37:26		View		
Filter BCF and Convert to VCF File [For Unique SAM] bcftools view uniq.var.bcf | perl vcfutils.pl varFilter -D10000 > out-unique.var.flt.vcf	2015-06-14 22:37:27	2015-06-14 22:37:37			Download(1.5 MB)	MD5
Mpileup and Create BCF File samtools mpileup -u -C50 -BQ0 -d10000000 -f U00096.3_150614220106489 out2.bam | bcftools view -bvcg - > non-uniq.var.bcf	2015-06-14 22:37:48	2015-06-14 22:40:48		View		
Filter BCF and Convert to VCF File bcftools view non-uniq.var.bcf | perl vcfutils.pl varFilter -D10000 > out.var.flt.vcf	2015-06-14 22:40:48	2015-06-14 22:40:59			Download(1.6 MB)	MD5
PileUp from Sorted BAM File For DepthCoverage samtools pileup -c -f U00096.3_150614220106489 out2.bam > out.pileup	2015-06-14 22:41:10	2015-06-14 22:46:16				
Convert BAM to SAMX For ErrorRate samtools view -hX out.bam > out.samX	2015-06-14 22:46:17	2015-06-14 22:46:47				
Sort BAM File For MapRatio samtools sort -n out.bam out_sorted_by_name	2015-06-14 22:46:48	2015-06-14 22:50:04		View		
Convert BAM to SAMX For MapRatio samtools view -hX out_sorted_by_name.bam > out_sorted_by_name.samX	2015-06-14 22:50:04	2015-06-14 22:50:35				

BACK

中間ファイル

マッピング結果のビューア

- IGV
  - https://www.broadinstitute.org/igv/
  - windows, Mac OSX, Linux
- Tablet
  - http://ics.hutton.ac.uk/tablet/
  - windows, Mac OSX, Linux
- samtools tview
  - http://www.htslib.org/
  - Mac OSX, Linux
- …



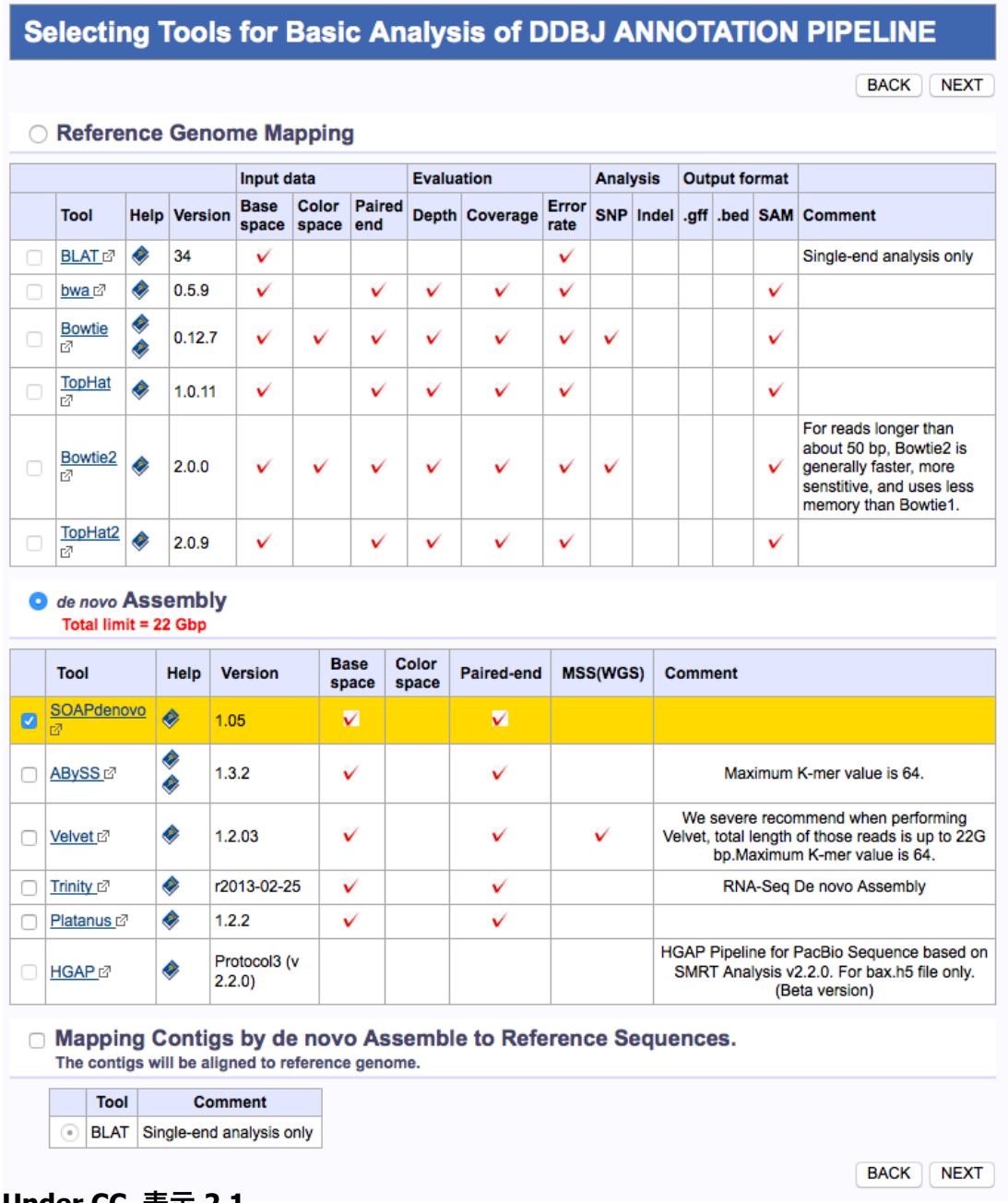
操作3

配列データから新規にゲノムをアセンブリする
(de novo assembly)

解析ツールの選択～de novo アセンブリ

マッピング系ツール

アセンブリ系ツール



Selecting Tools for Basic Analysis of DDBJ ANNOTATION PIPELINE

BACK NEXT

Reference Genome Mapping

	Tool	Help	Version	Input data: Base space	Input data: Color space	Input data: Paired end	Evaluation: Depth	Evaluation: Coverage	Evaluation: Error rate	Analysis: SNP	Analysis: Indel	Output format: .gff	Output format: .bed	Output format: SAM	Comment
☐	BLAT		34	✓					✓						Single-end analysis only
☐	bwa		0.5.9	✓		✓	✓	✓	✓					✓	
☐	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.
☐	TopHat2		2.0.9	✓		✓	✓	✓	✓					✓	

***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		r2013-02-25	✓		✓		RNA-Seq De novo Assembly
☐	Platanus		1.2.2	✓		✓		
☐	HGAP		Protocol3 (v 2.2.0)					HGAP Pipeline for PacBio Sequence based on SMRT Analysis v2.2.0. For bax.h5 file only. (Beta version)

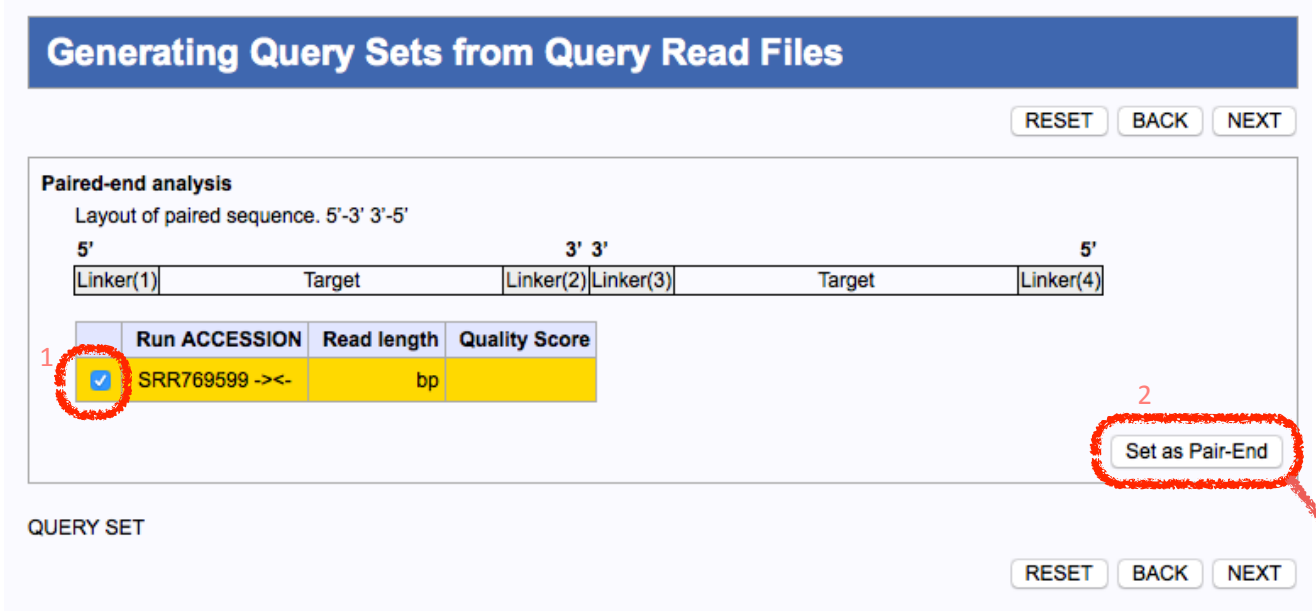
Mapping Contigs by de novo Assemble to Reference Sequences.

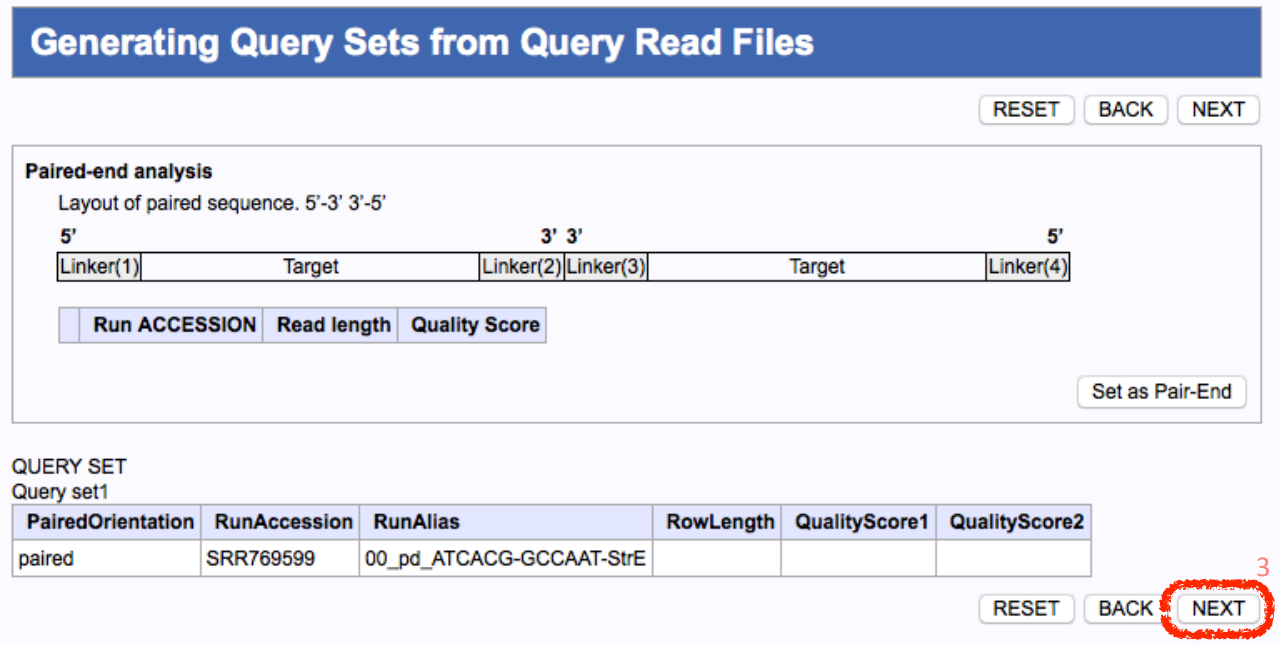
The contigs will be aligned to reference genome.

	Tool	Comment
○	BLAT	Single-end analysis only

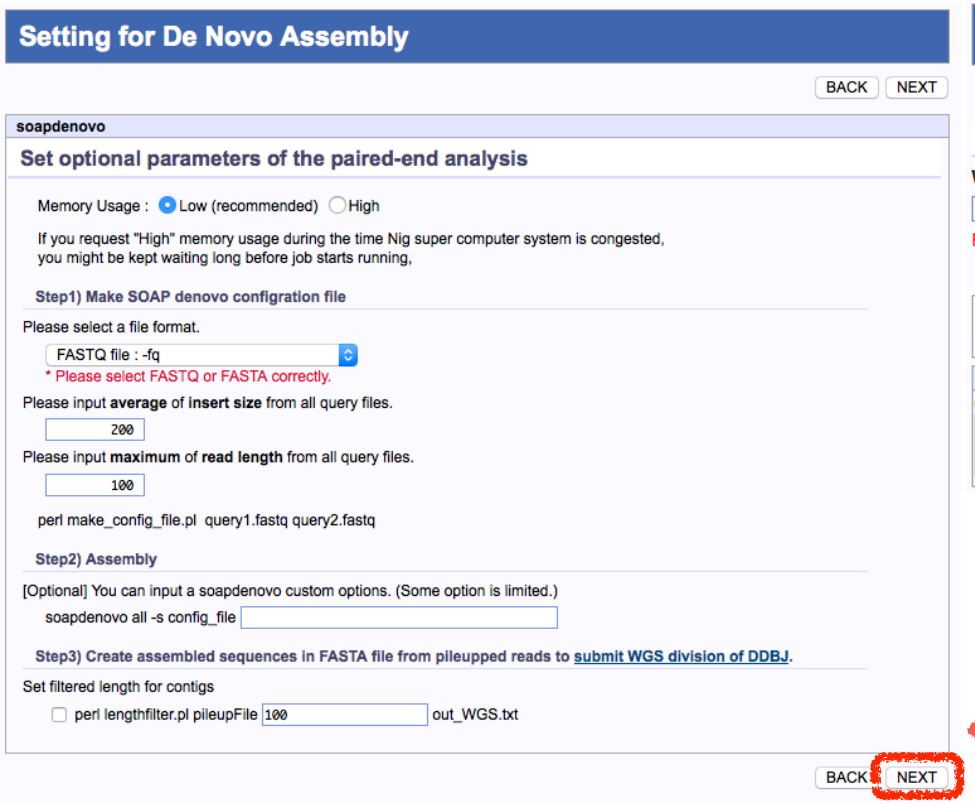
BACK NEXT

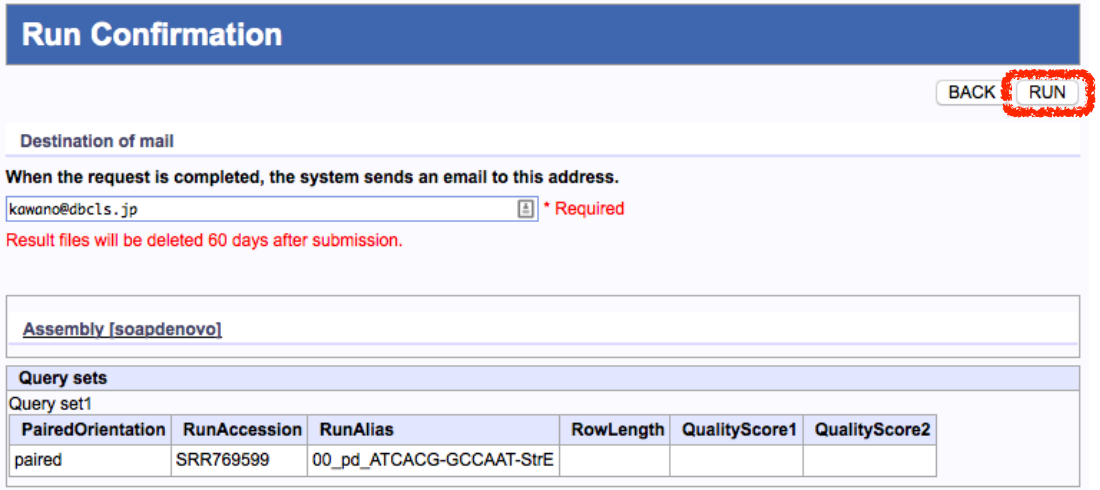
解析するランの指定





パラメータ設定、確認、実行



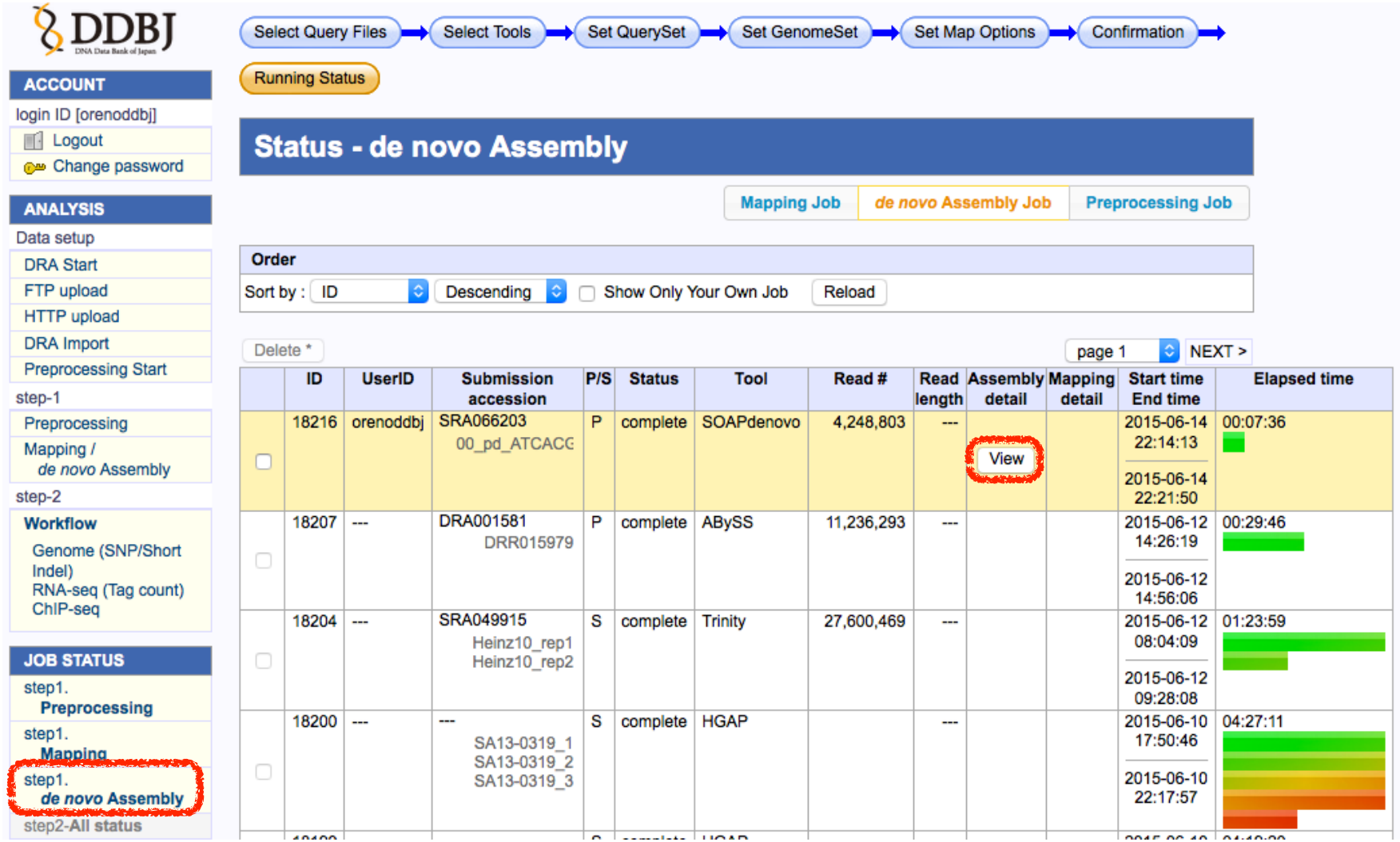


計算に時間がかかるのでメールアドレスの登録は必須

ログインしないと実行できません

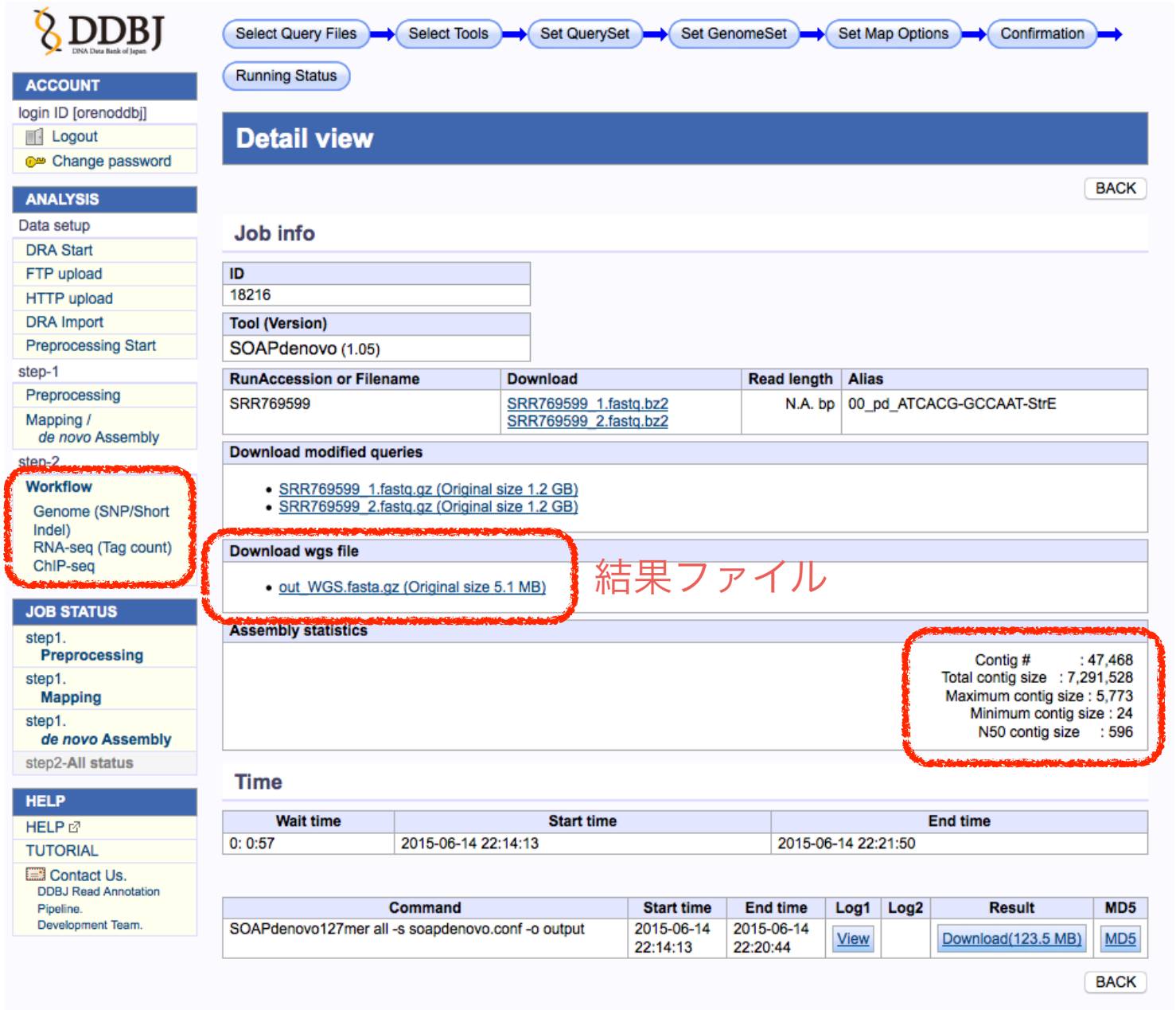
（Runボタンが出てきません）

進捗確認画面-Assembly



背景無色（白）が他のユーザ
背景黄色が自分のジョブ

アセンブリ結果



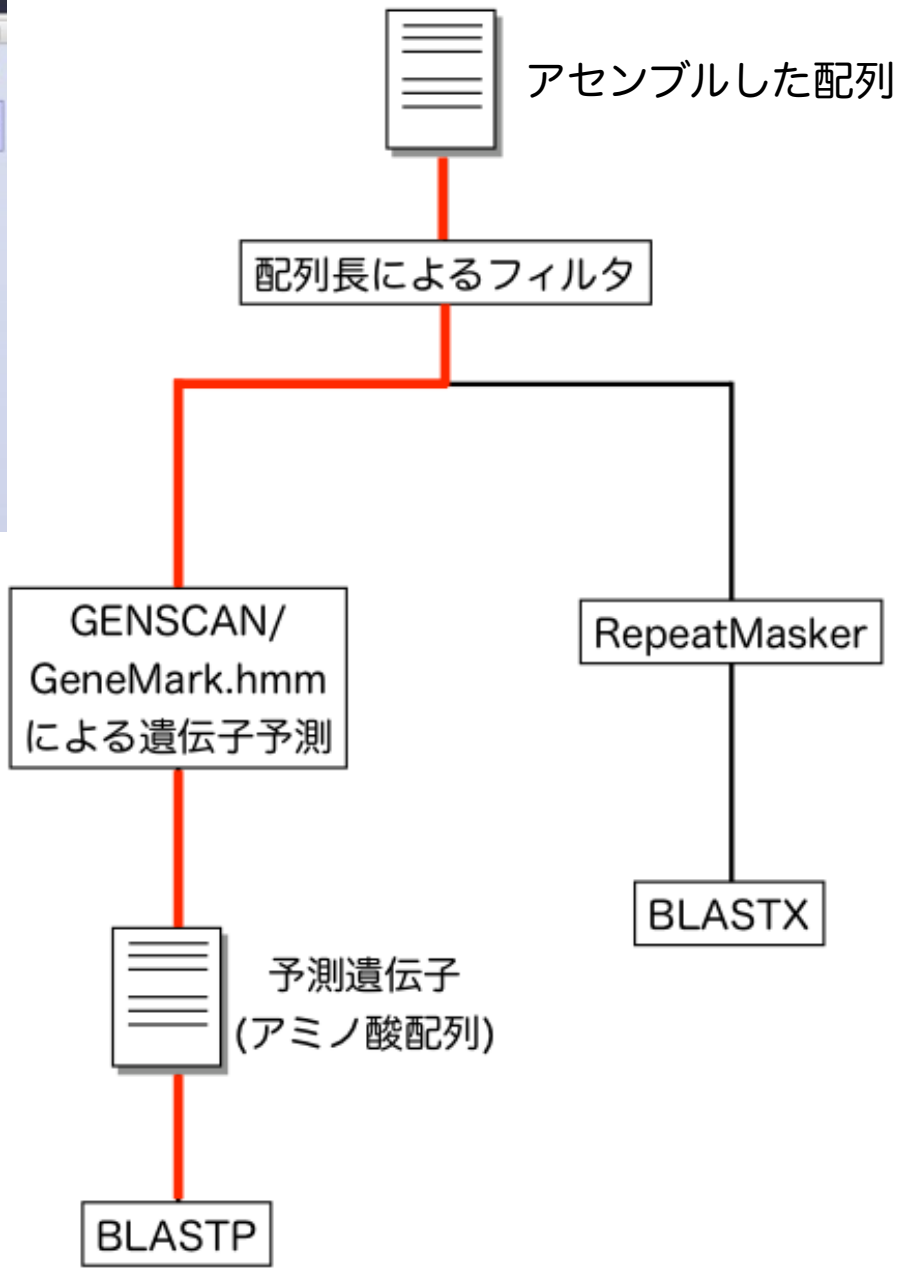
応用1

p-Galaxyを使った配列アノテーション

p-Galaxyを使った配列アノテーション

https://p-galaxy.ddbj.nig.ac.jp/

（開発中のサービスのため、一部サービスが使えなかったり、一時的にアクセスできなくなったりする可能性があります）

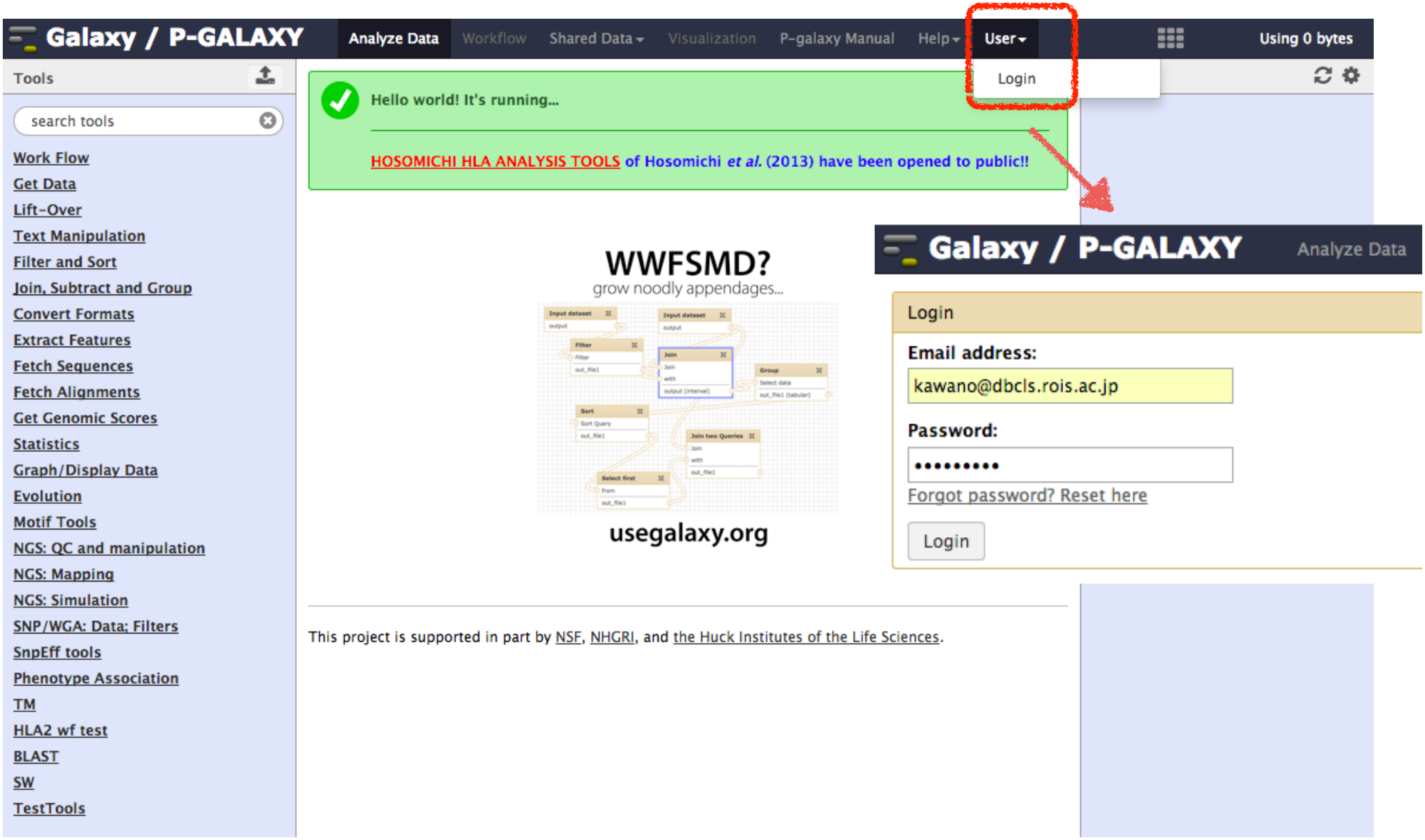


Galaxyとは？

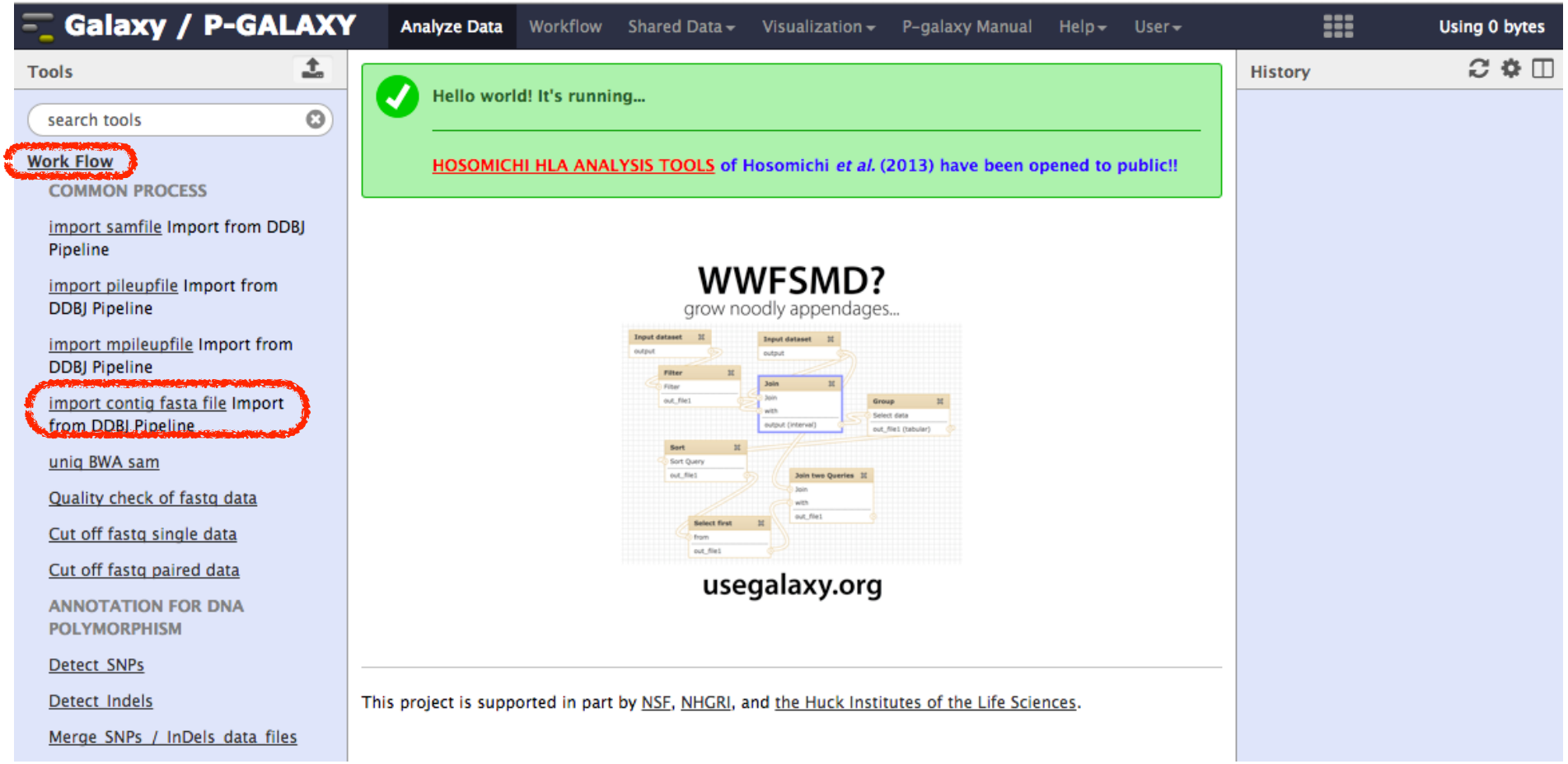
- ゲノムなど生物学データを対象とした、データ解析ワークフローの共有・公開のためのプラットフォーム
- オリジナルはペンシルバニア州立大学、エモリー大学を中心としたGalaxy teamが開発
- 独自のツールを追加して公開可能
  - DBCLS Galaxy
    - テキスト系ツール
  - p-Galaxy
    - 次世代解析ツール

http://usegalaxy.org/

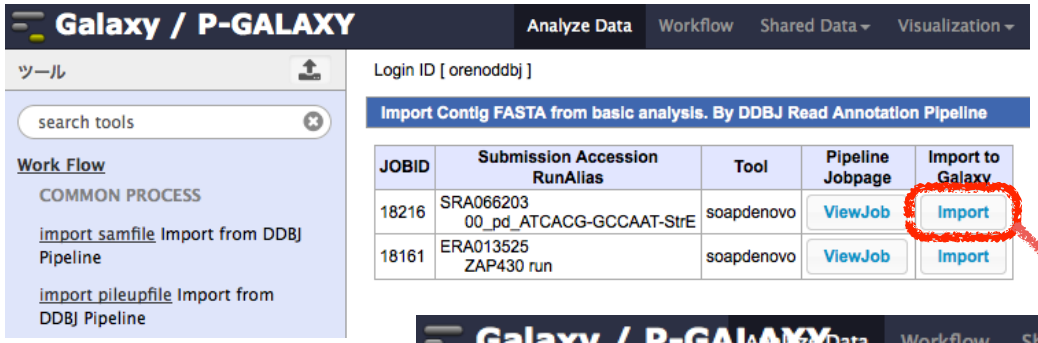
p-Galaxyにログイン

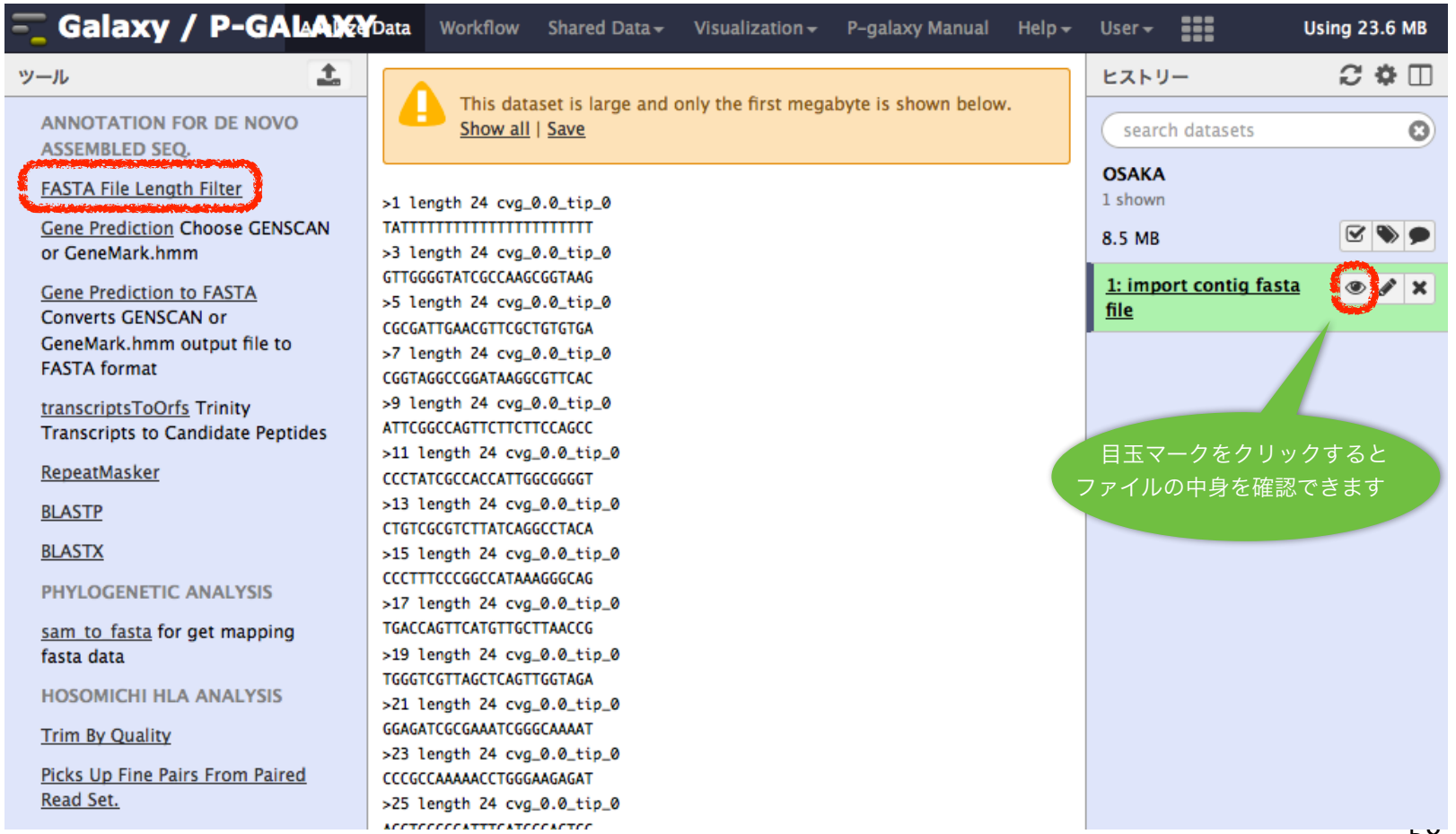


pipeline計算結果のデータインポート

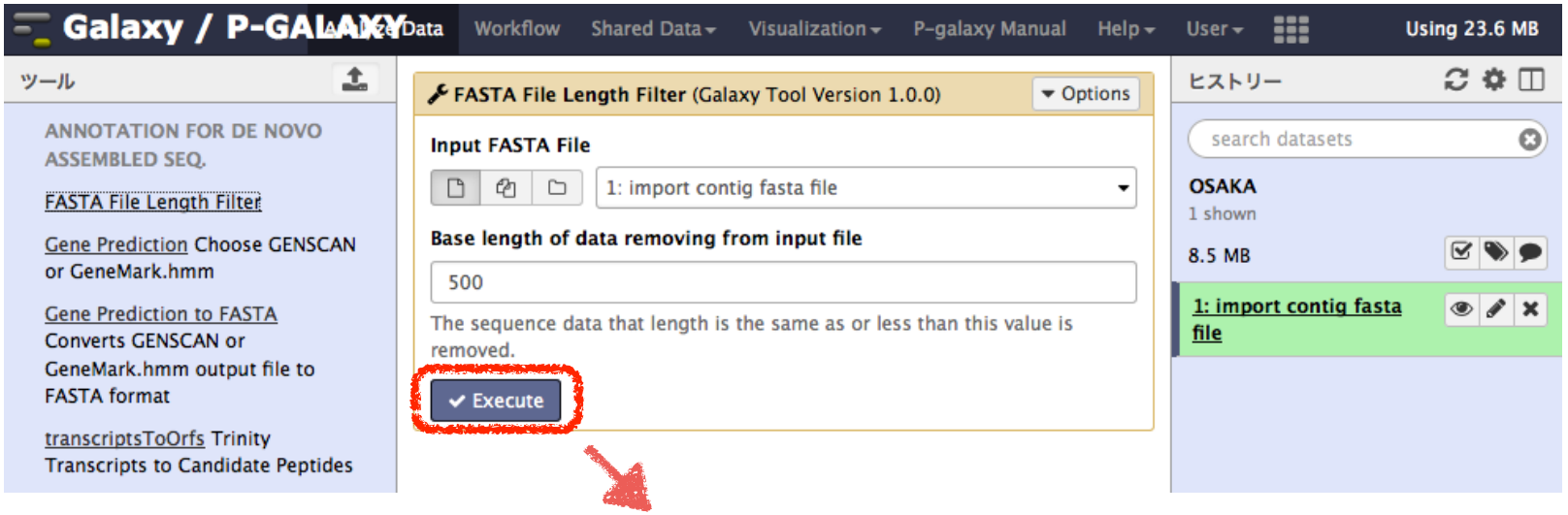


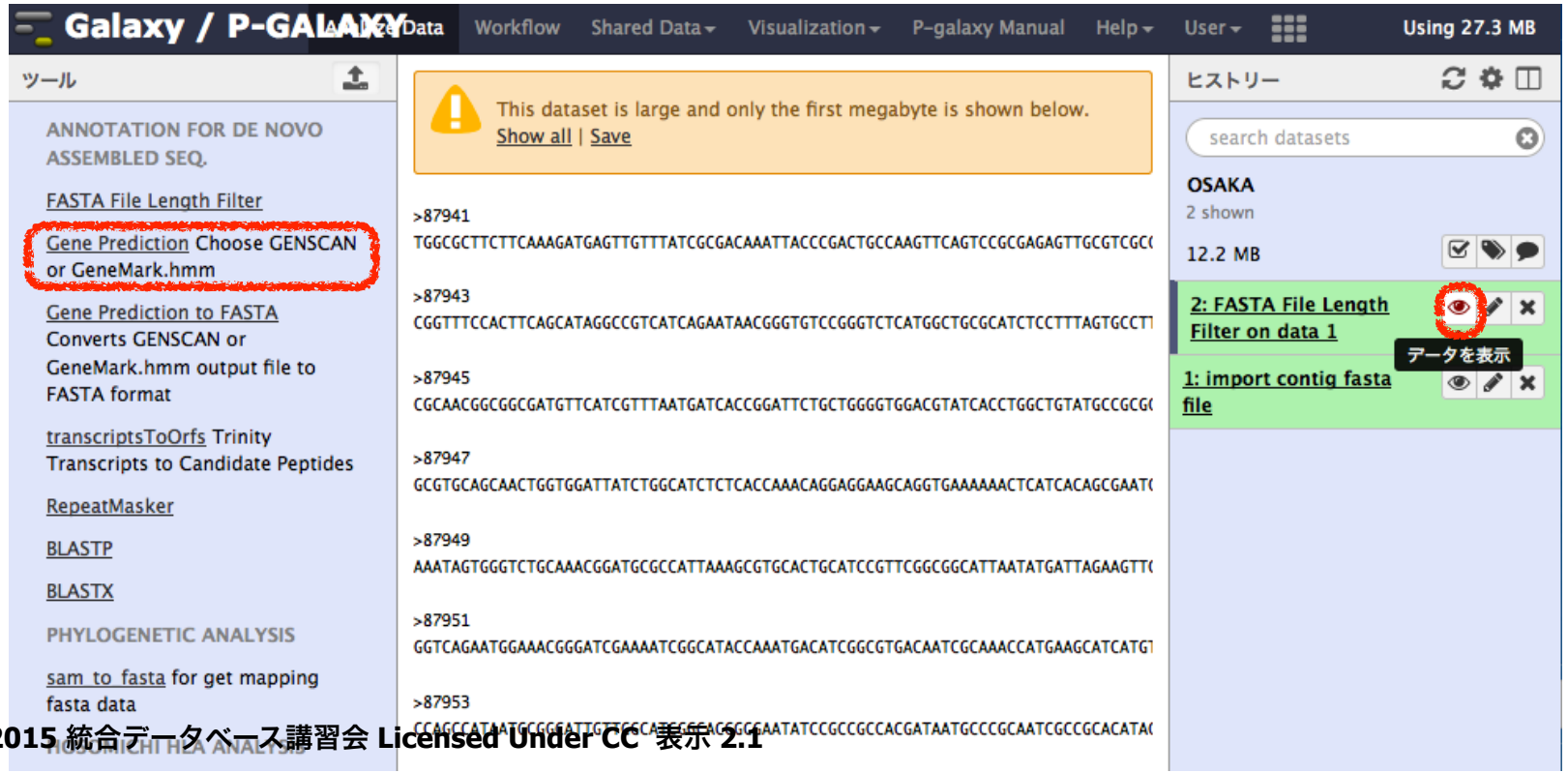
データのインポート





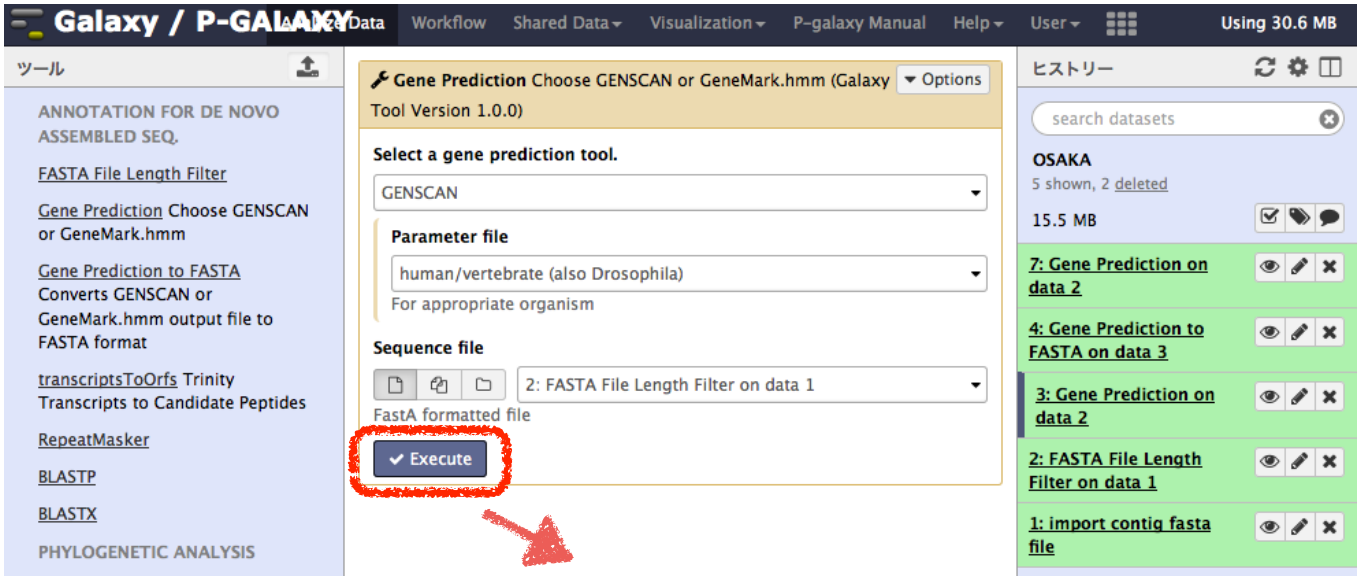
配列長でのフィルタリング

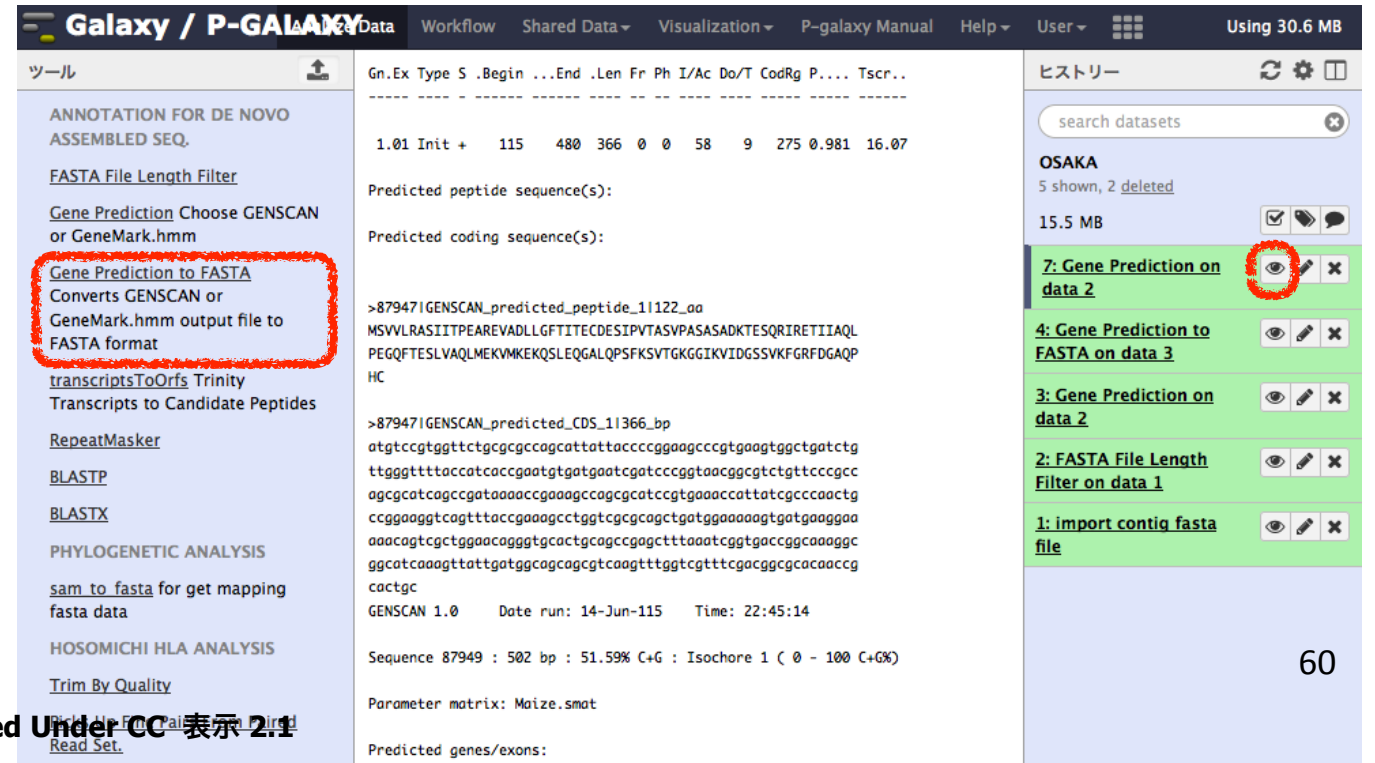




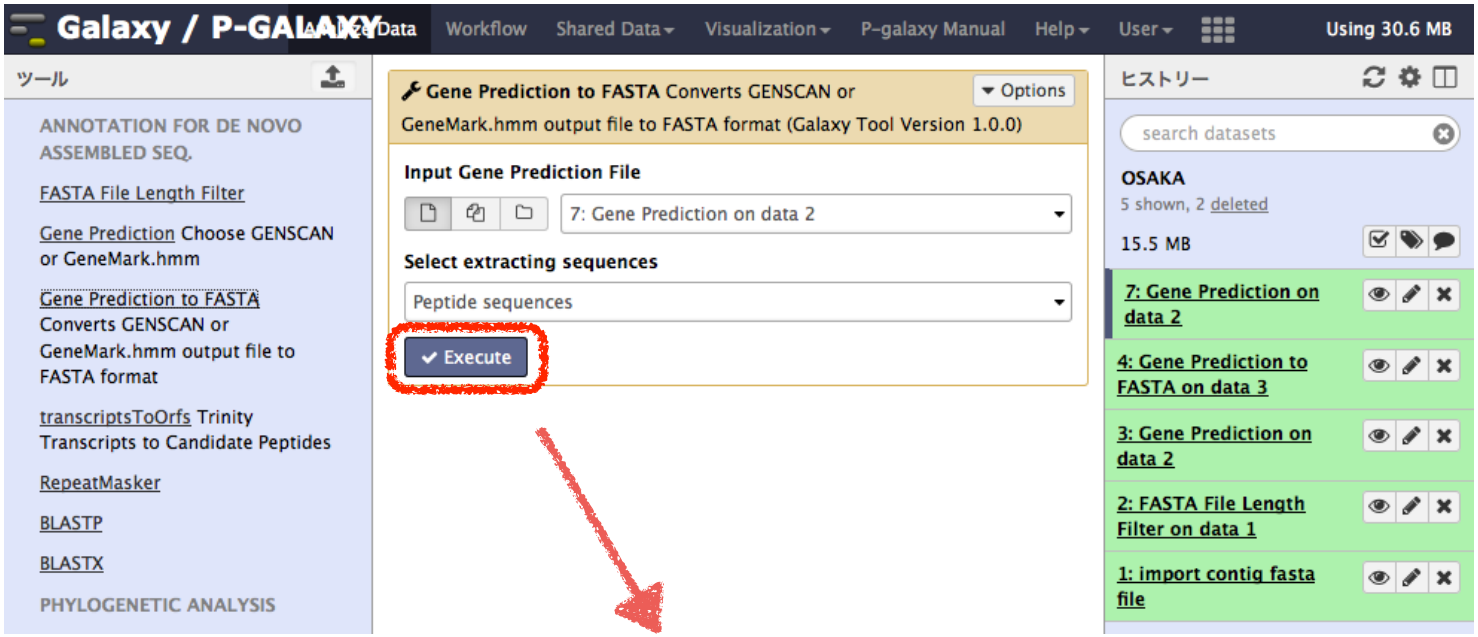
遺伝子予測

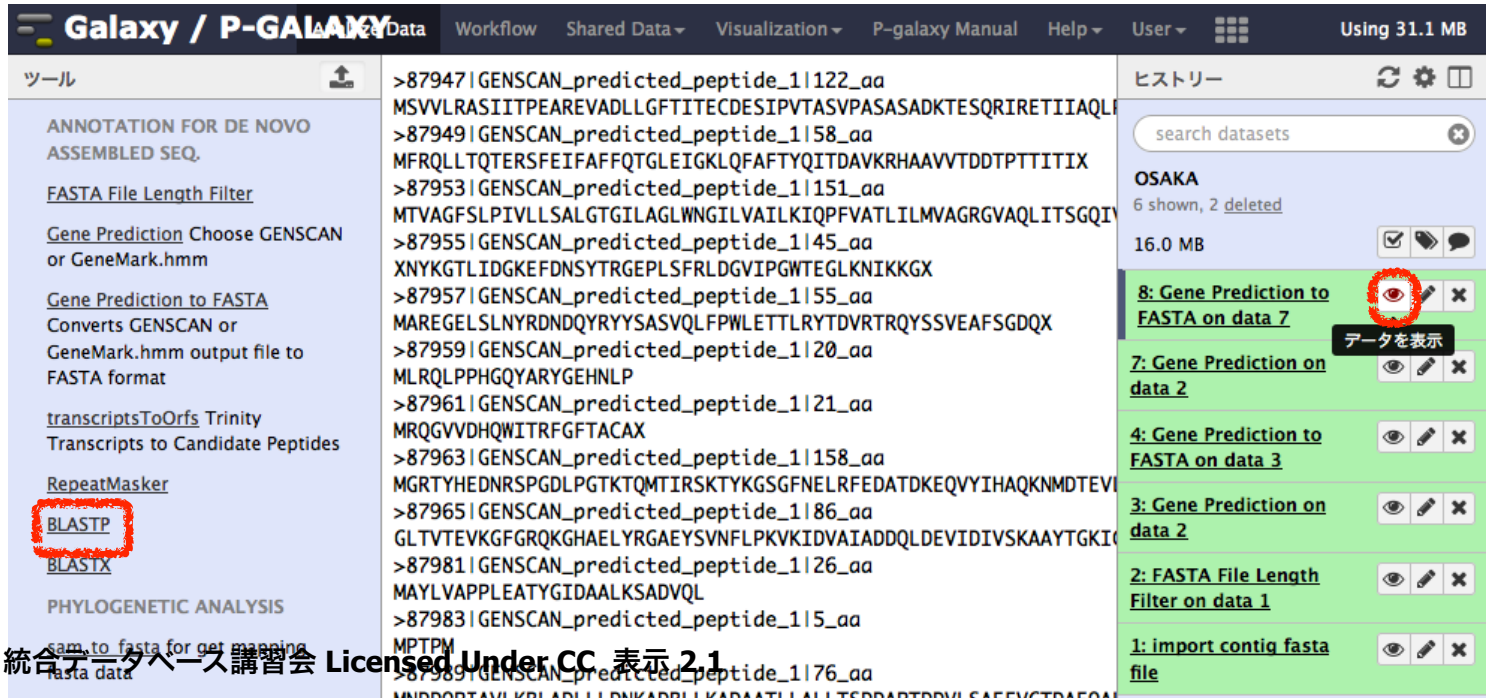
注：本来であれば、原核生物用の遺伝子モデルを使うべきですが、なぜかうまくいかないので（おそらく設定ミス）
今回は真核生物の遺伝子モデルを使って遺伝子を予測しています



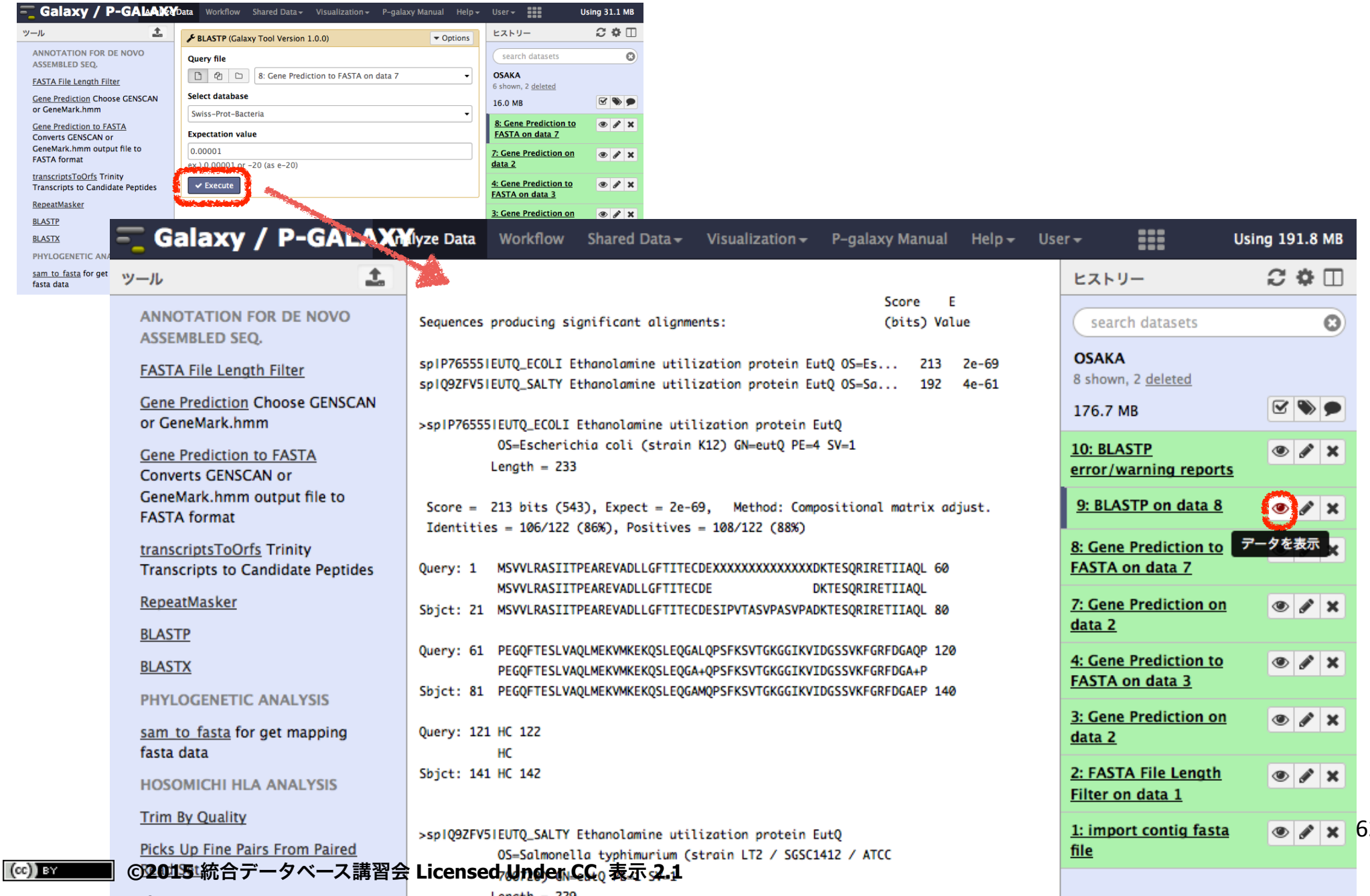


予測配列の取り出し





BLASTによる配列の機能アノテーション



DDBJ Pipeline 参考資料

- 詳しくはDDBJing講習会で

  http://www.ddbj.nig.ac.jp/ddbjing/dl.html

応用2

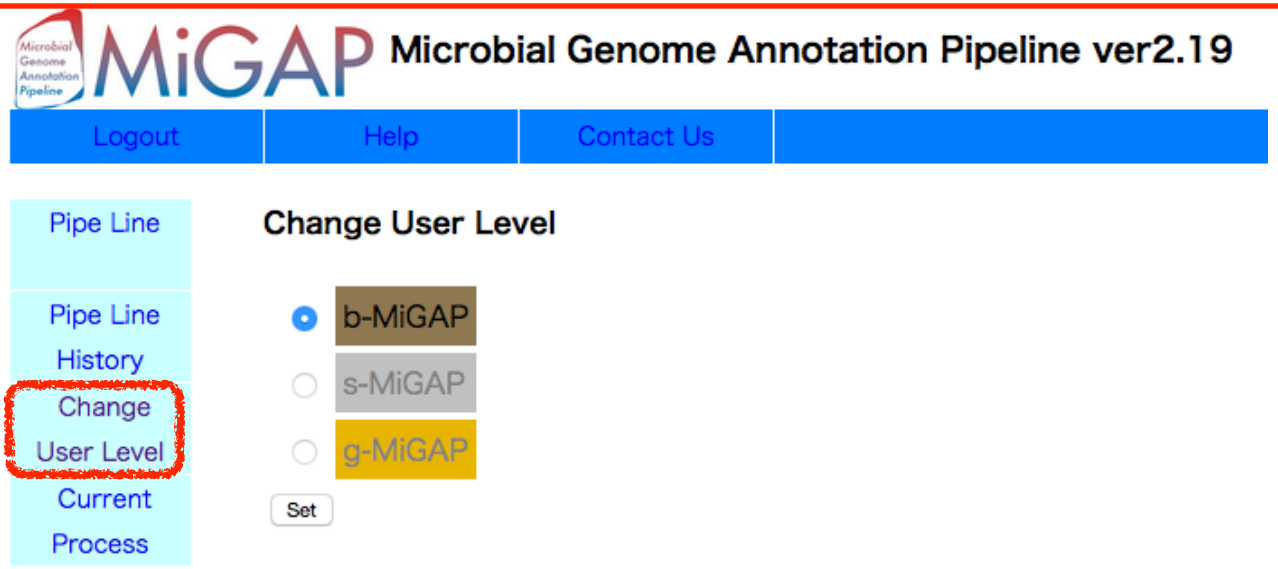
MiGAPを使った配列アノテーション

MiGAP

- 微生物ゲノムのアノテーションワークフロー
- MiGAPで付加するアノテーション情報
  - ORF (CDS)の同定
    - de novo 予測 ⇒ MetaGeneAnnotator, Glimmer, Augustus
  - rRNAの予測
    - de novo 予測 ⇒ RNAmmer
    - 既知rRNA配列との配列類似性
  - tRNAの予測
    - de novo 予測 ⇒ tRNAScan-SE
  - ORFの機能予測
    - BLAST

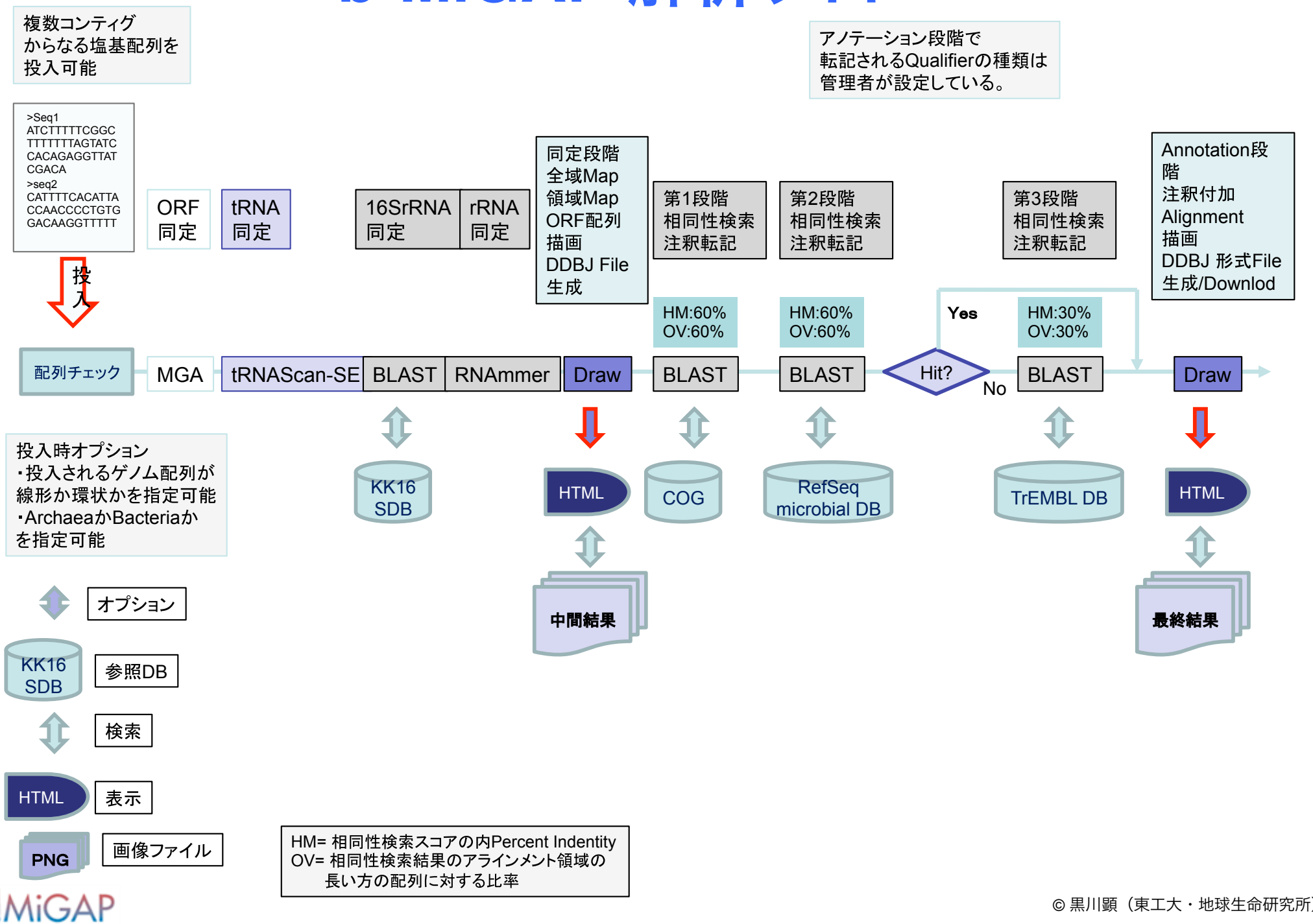
MiGAPウェブサイト　http://www.migap.org/

ユーザレベルの設定

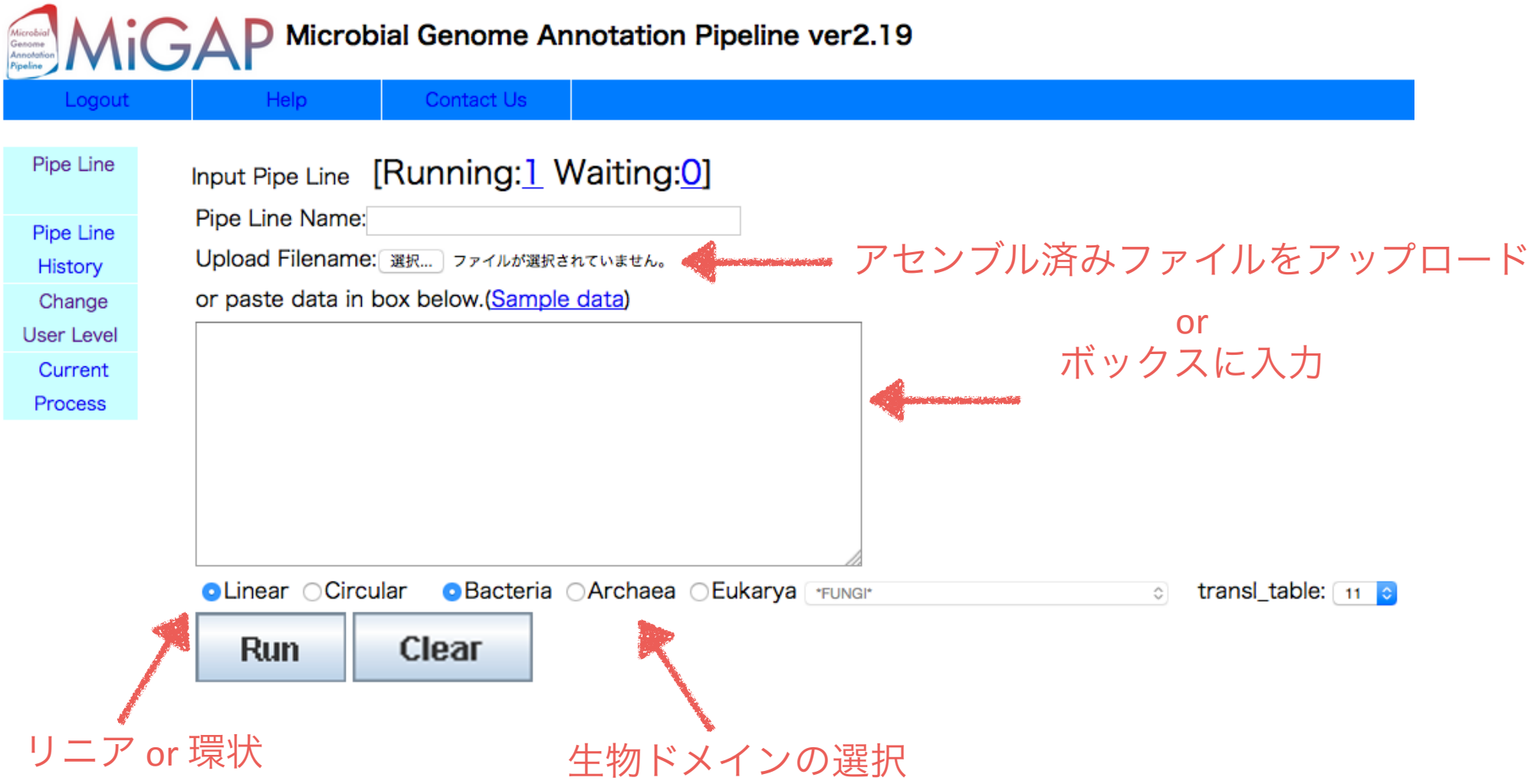


- b-MiGAP
  - ツール・パラメータ固定（ブロンズ）
- s-MiGAP
  - パラメータ設定が可能（シルバー）
- g-MiGAP
  - DBやツール、パラメータ設定が可能（ゴールド）

b-MiGAP 解析フロー



MiGAPデータ入力



MiGAP結果

MiGAP Microbial Genome Annotation Pipeline ver2.19

Logout | Help | Contact Us

LDAP_orenomigap (b-MiGAP) 2015/06/14 17:50:47 [View Menu] [Hide Menu]

Pipe Line
Pipe Line History
Change User Level
Current Process

History

LDAP_orenomigap | Hidden Data List

<< < [1/1] > >>

2015/06/14 17:58:12
2015/06/14 17:56:45
fuga
hogggge

Contig

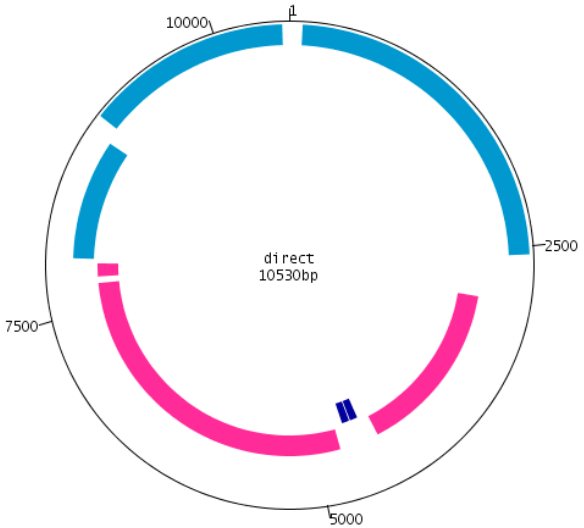
<< < [1/1] > >>

contig

Hide

Basic Information

Filename: direct
Contig: 1
Total Length: 10530
rRNA: 3
tRNA: 2
CDS: 3
RBS: 3
Run: 2015/06/14 17:58:12



ORF & RNA Extract

Start: 2015/06/14 17:58:12
End: 2015/06/14 17:58:23
Software: MetaGeneAnnotator 1.0
tRNAscan-SE 1.23
NCBI BLAST 2.2.18
RNAmmer 1.2

Annotation

Start: 2015/06/14 17:58:23
End: ---
Software: NCBI BLAST 2.2.18
DB: COG[20030417]:3
RefSeq[20140911]:0
TrEMBL release2014_04[20140417]:0

Download

Log File pipeline.log
N.A.: result-na.fasta.tar.gz
A.A.: result-aa.fasta.tar.gz

MiGAP参考資料

- 統合TV

✓微生物ゲノムアノテーションツールMiGAP

- 開発者である黒川先生による講習会の動画
  - http://youtu.be/ujxI6LJIbUE
  - http://togotv.dbcls.jp/20131024.html

✓MiGAPの使い方～導入と基本操作

- 動画による解説
  - http://youtu.be/oXAEZgoc5Eo
  - http://togotv.dbcls.jp/20100624.html

参考サイト

- 統合TV 

  - http://togotv.dbcls.jp/index.rb?category=NGS
- NGS Surfer's wiki 

  - https://cell-innovation.nig.ac.jp/wiki/tiki-index.php
- Q and A 

  - ライフサイエンスQA (日本語)、BioStar、SEQanswers
    - http://qa.lifesciencedb.jp/
    - https://www.biostars.org/
    - http://seqanswers.com/
- (Rで)塩基配列解析
  - http://www.iu.a.u-tokyo.ac.jp/~kadota/r_seq.html