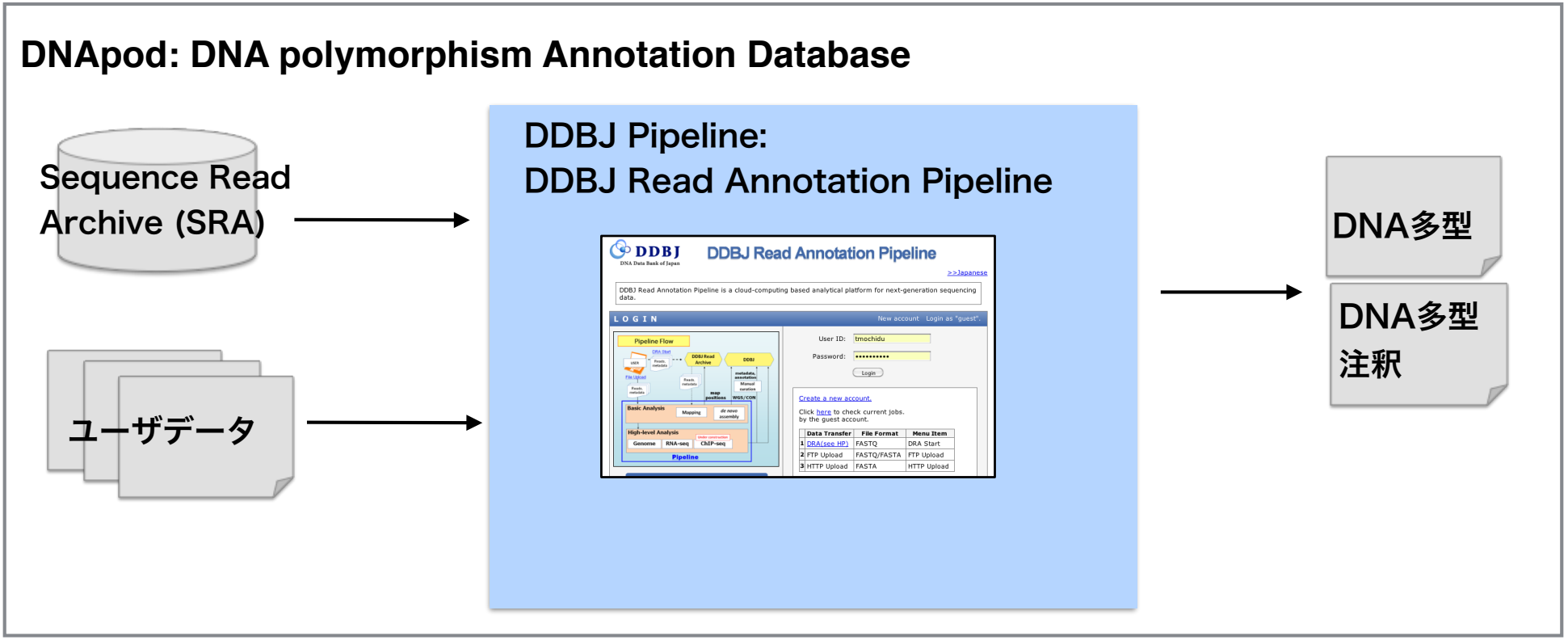
DDBJ Pipelineを用いたDNA多型注釈解析の実習



国立遺伝学研究所 大量遺伝情報研究室
望月孝子

NGS: Next Generation Sequencer 概要

NGS: Next Generation Sequencer 代表的な機種

illumina　　短鎖並列型

Hiseq 4000

リード長 150bp の場合

1300–1500Gb

PacBio　　長鎖並列型

Sequel System

平均リード長

約10kb

※1ランあたりフローセル2枚を使用した場合

NGSの活用

ゲノム (リシーケンス、*de novo* アセンブリ)、発現量解析、メタゲノム解析、
ChIP-Seq（転写因子解析）、SNP / InDel 解析、…

アプリケーション別に必要なリードスペック

application / 実験種	total bases / 総塩基数	read length / リード長	read number (M) / リード数
ヒトゲノムリシーケンス	90-150Gb	2x100	900-1500
ターゲットリシーケンス	<1Gb	2x100	10
exome sequence	5~7Gb	2x100	70
RNA-Seq	5Gb	2x100	50
TSS-Seq	1Gb	1x50	20
small RNA	0.35Gb	1x35	>10
微生物ゲノム	>150Mb	2x100	>1.5
真核生物ゲノム	>4Gb	2x100	>40
Bisulfite-Seq	90-150Gb	2x100	900-1500
ChIP-Seq	>6Gb	1x100	60

注: 対象のゲノムサイズなどで数字が変わることがあります．また，既に情報が古くなっている可能性もあります

細胞工学別冊 次世代シーケンサー目的別アドバンストメソッド p21より引用

機器に合った利用を

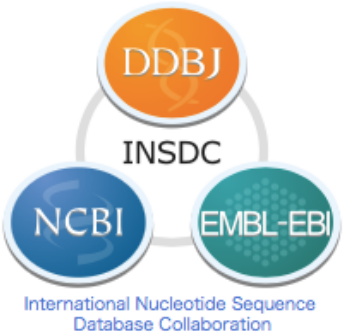
AJACS 53 仲里さんの資料より。

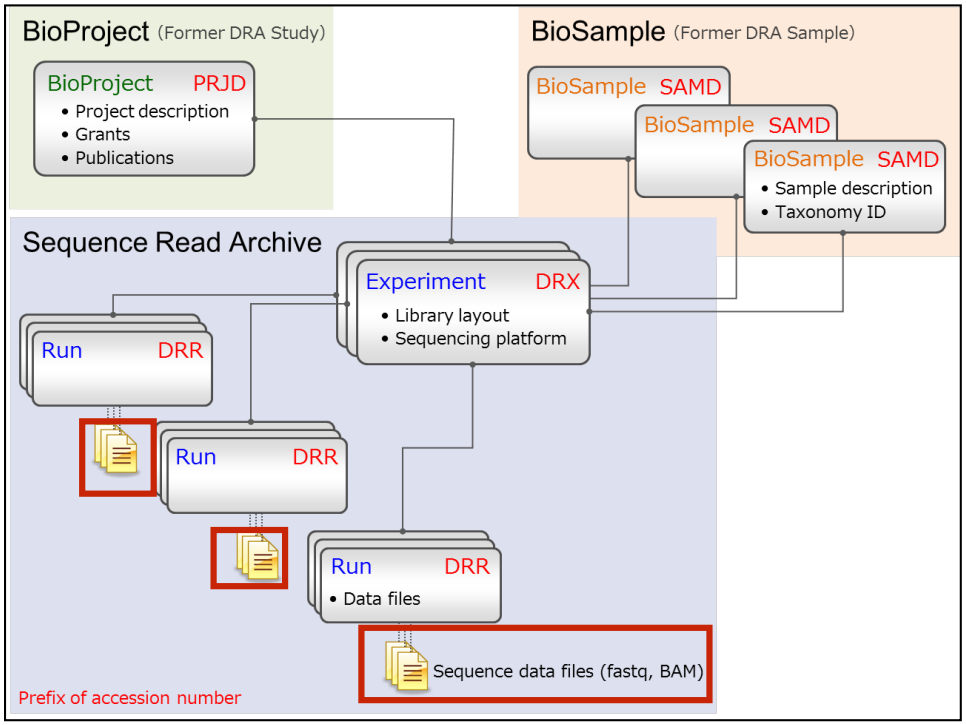
https://github.com/AJACS-training/AJACS53/tree/master/nakazato

NGS 公共データベース

SRA

Sequence Read Archive　　データ構成 メタデータ＋配列情報





JGA

Japanese Genotype-phenotype Archive

Japanese Genotype-phenotype Archive (JGA) は個人レベルの遺伝学的なデータと匿名化された表現型情報を保存し，提供しています。 データが収集された個人との間の同意に基づく協定により，JGA のデータ利用は特定の研究目的に制限されています。 JGA は厳格なプロトコールに従い，情報を管理，格納，提供しています。 登録処理が終わった全てのデータは暗号化されます。なお，JGA に登録されるデータおよびデータの利用についての審査は独立行政法人科学技術振興機構 (JST)/バイオサイエンスデータベースセンター (NBDC) が実施しています。 JGA は科学技術振興機構 National Bioscience Database Center (NBDC) と共同で運営されています。

DDBJ HP より引用 (http://trace.ddbj.nig.ac.jp/jga/index.html)

FASTQ format

```
@SEQ_ID
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
+
!''*((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65
```

4行で1セット

1行目: @ から始まり、配列の ID を記載する。

2行目: 配列

3行目: + を記載する。

4行目: クォリティスコア

クォリティスコアをASCII コードで表示

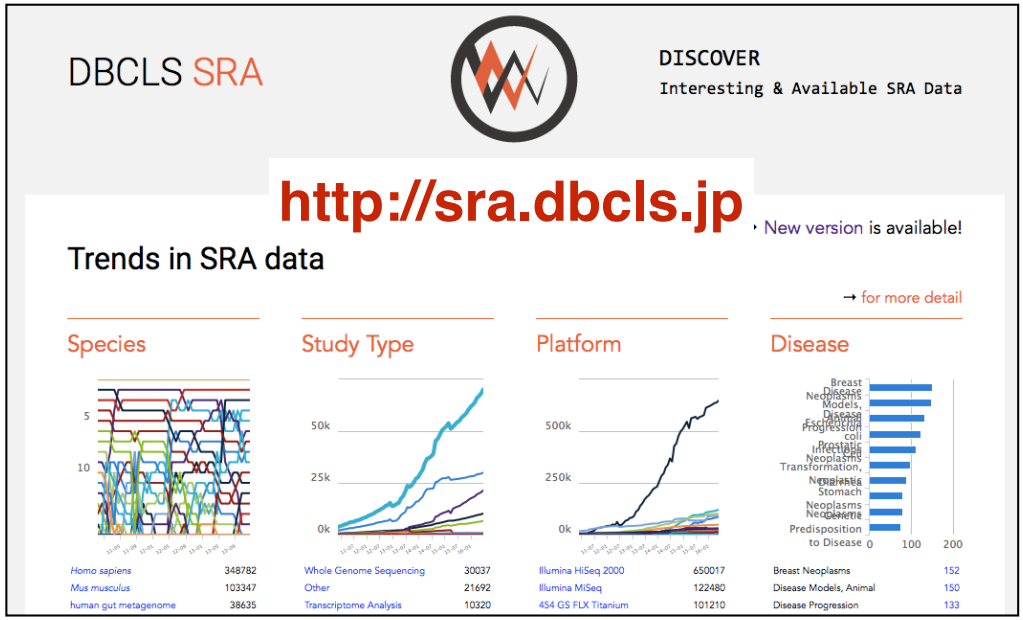
```
  SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS.....................................................
  ..........................XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX.....................
  ............................IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII.....................
  ............................JJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ.....................
  LLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLL.....................................................
  !"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}~
  |                         |    |        |                              |                     |
 33                        59   64       73                            104                   126
  0........................26...31.......40
                           -5....0........9.............................40
                                0........9.............................40
                                   3.....9.............................40
  0.2......................26...31........41

  S - Sanger        Phred+33,  raw reads typically (0, 40)
  X - Solexa        Solexa+64, raw reads typically (-5, 40)
  I - Illumina 1.3+ Phred+64,  raw reads typically (0, 40)
  J - Illumina 1.5+ Phred+64,  raw reads typically (3, 40)
      with 0=unused, 1=unused, 2=Read Segment Quality Control Indicator (bold)
      (Note: See discussion above).
  L - Illumina 1.8+ Phred+33,  raw reads typically (0, 41)
```

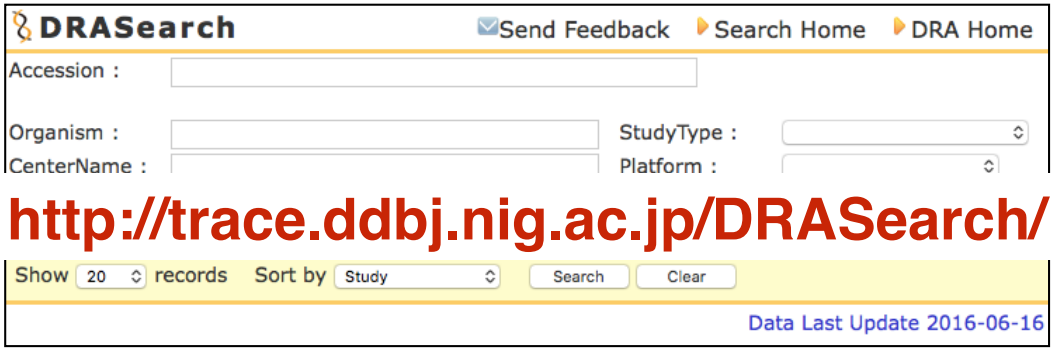
←SRA

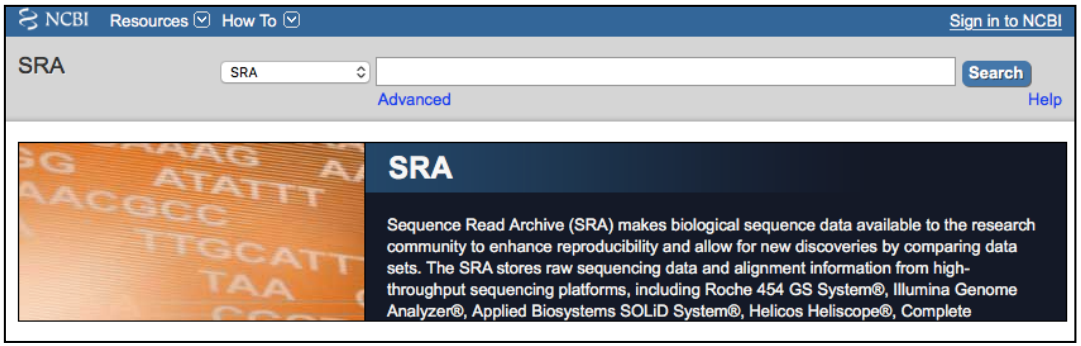
wikipedia FASTQ format より

.sra 形式でデータが提供されている場合は、SRA tool kit にてFASTQ format への変換が必要。

NGS 公共データの検索



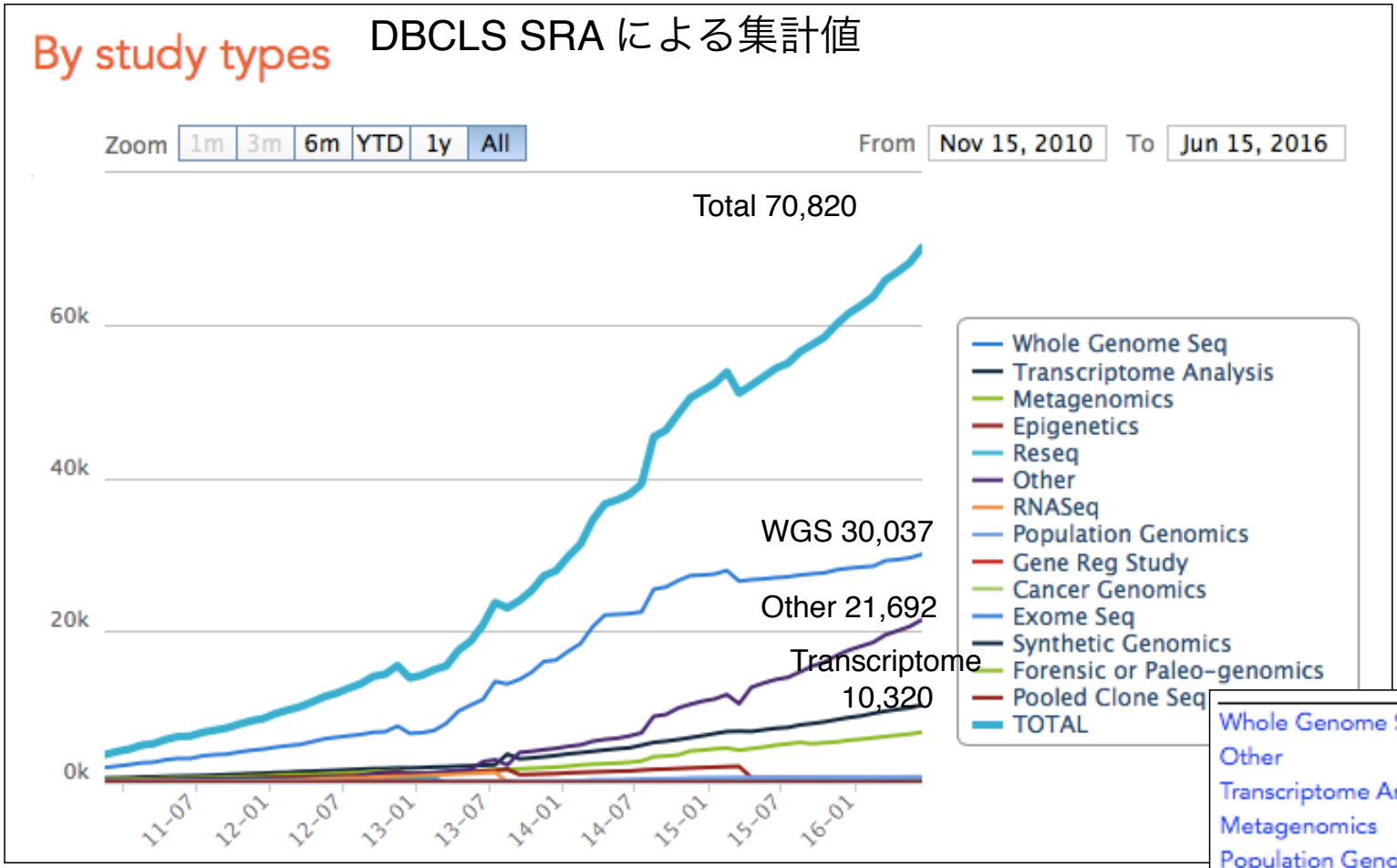




http://www.ncbi.nlm.nih.gov/sra/

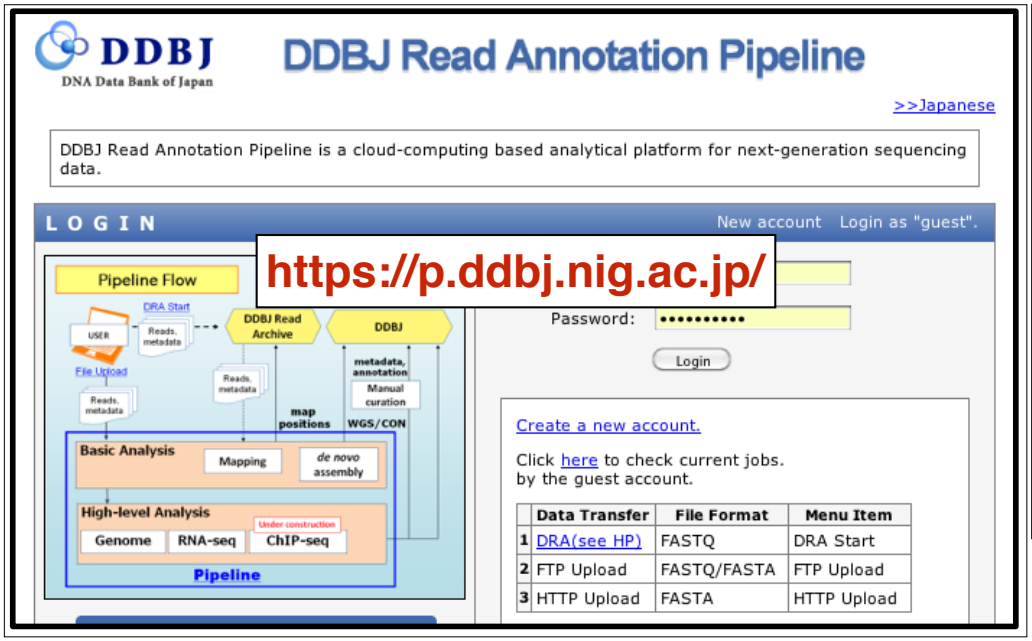
SRA データ登録状況

DBCLS SRA による集計値

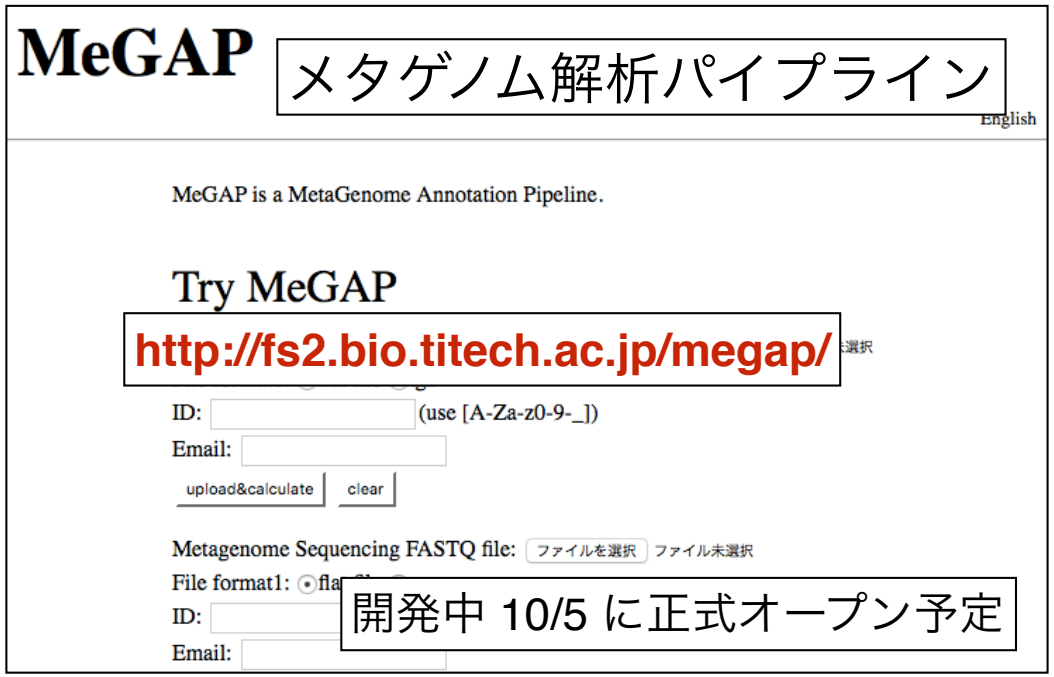


Whole Genome Sequencing	30037
Other	21692
Transcriptome Analysis	10320
Metagenomics	6698
Population Genomics	778
Epigenetics	684
Exome Sequencing	196
RNASeq	160
Cancer Genomics	136
Resequencing	58
Pooled Clone Sequencing	35
Gene Regulation Study	15
Synthetic Genomics	9
Whole Genome Sequencing	1
Transcriptome Sequencing	1
Total	70820

NGS 解析パイプライン 紹介



日本でのGalaxy コミュニティ。
研究者が作成した解析ワークフローを
Pitagora-Galaxy から配布している。



Galaxyとは ….
Galaxy はゲノムなどの生物学データを
対象としたウェブベースの解析環境。
ジョンズ・ホプキンス大学を中心とした
Galaxy team が開発
(本家 Galaxy サイトhttp://galaxyproject.org/)

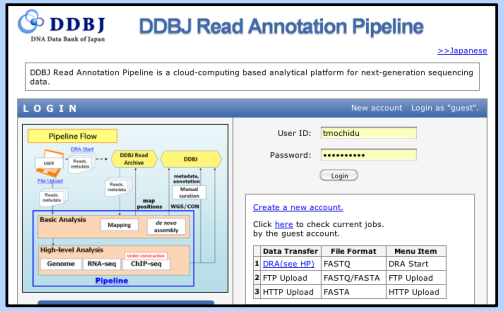
DDBJ pipeline: DDBJ Read Annotation Pipeline

DDBJ Sequence Read Archive (DRA)

ユーザデータ

DDBJ Pipeline: DDBJ Read Annotation Pipeline

基礎処理部



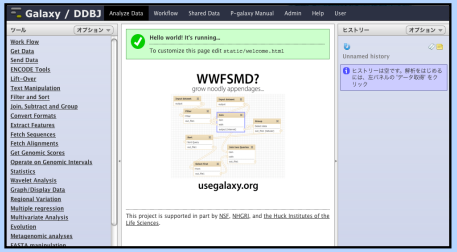
Quality Value (QV) フィルタ

マッピング

de novo アセンブリ

高次処理部 (Pitagora-Galaxy)

解析目的別ワークフロー



DNA多型注釈 (DNApod)

Contig, Scaffold注釈 (開発中)

Pitagora-Galaxy

Pitagora-Galaxy
Galaxy Japan Community + VM + Cloud

http://www.pitagora-galaxy.org

Pitagora-Galaxy にて公開しているワークフロー

現在のワークフロー

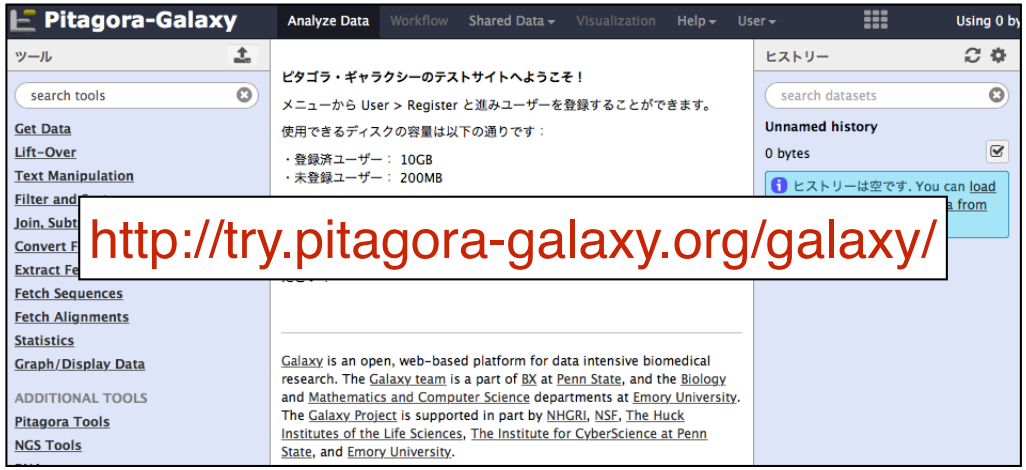
- 各ワークフローのベンチマーク結果はこちら

ワークフロー名	概要
RNA-seq 01	FastQC – TopHat2 – Cufflinks – Cuffmerge – Cuffdiff
RNA-seq 02	FastQC – FastqMcf – Sailfish
RNA-seq 03	bam2readcount – edgeR – vehp
ChIP-seq 02	Bowtie2 – MACS – 遺伝子の TSS 前後 1,000 bp と重なっているピークを取得
ChIP-seq 03	Bowtie2 – 4種の peak calling: MACS(1.3) / MACS(1.4) / MACS2 / SICER – 同じ BED 形式で出力
ChIP-seq 04	BWA – 4種の peak calling (上の ChIP-seq 03 と同じ)
BS-seq 01	Bisulfighter によるメチル化変化の検出
Variant Calling 02	GATK 3.3_0 を利用したバリアントの検出 (GenomyAnalysisTK, Picard)

DNApod ワークフローは現在、
https://sites.google.com/a/g.nig.ac.jp/dnapod-help/howto/workflow
から公開中。
Pitagora-Galaxy 次バージョンよりDNApod ワークフローを統合予定。

利用方法

(1)テスト用galaxyを利用



http://try.pitagora-galaxy.org/galaxy/

(2) 自分のマシン上で Galaxy の仮想マシンをセットアップ

(3) クラウド計算環境を使ってすぐに Galaxy を起動

DNApod: DNA polymorphism annotation database

ワークフローとデータベースを公開

DDBJ Sequence Read Archive (DRA)

WGS: whole-genome sequencing

ユーザデータ

Reference

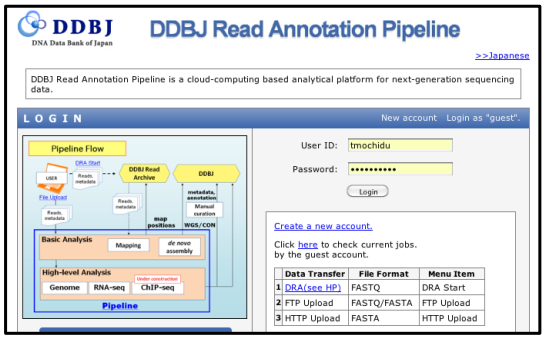
GACCGAGCTACGCCTCCTGTGGA

Reads (BWA)

GAGCTACGCCACCTG
GAGCTACGCCACCTG
GAGCTACGCCACCTG
AGCTACGCCACCTGT
GCTACGCCACCTGTG
GCTACGCCACCTGTG

SNP (samtools mpileup)

DDBJ Read Annotation Pipeline
基礎処理部



高次処理部 (Pitagora-Galaxy)
DNApod workflow

Reference

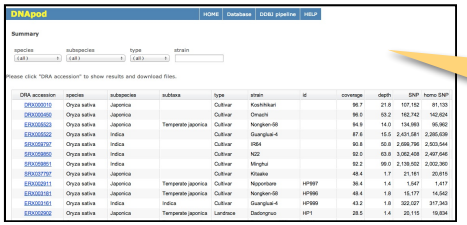
gene

exon

intron

バクテリア～動物、植物を網羅して行く予定

DNApod



現在、イネ679系統、トウモロコシ404系統、ソルガム66系統

DNApod ワークフロー 実習

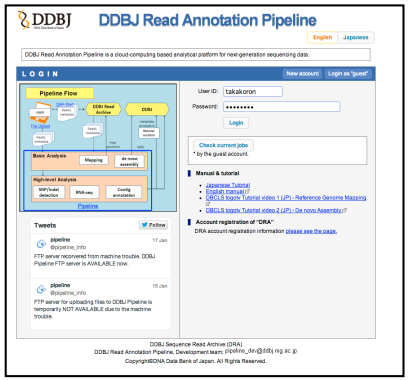
DNApod ワークフロー 実習の流れ

DRA

DRA000307

O. sativa subsp. *japonica* cv. Omachi

基礎処理部



マッピング

コンセンサス配列の決定

実習

インポート

高次処理部 (Pitagora-Galaxy)

DNA多型検出＆注釈

ホモSNPs検出

ホモInDels検出

ゲノム上のDNA多型の分布の描画

既知遺伝子による注釈付け

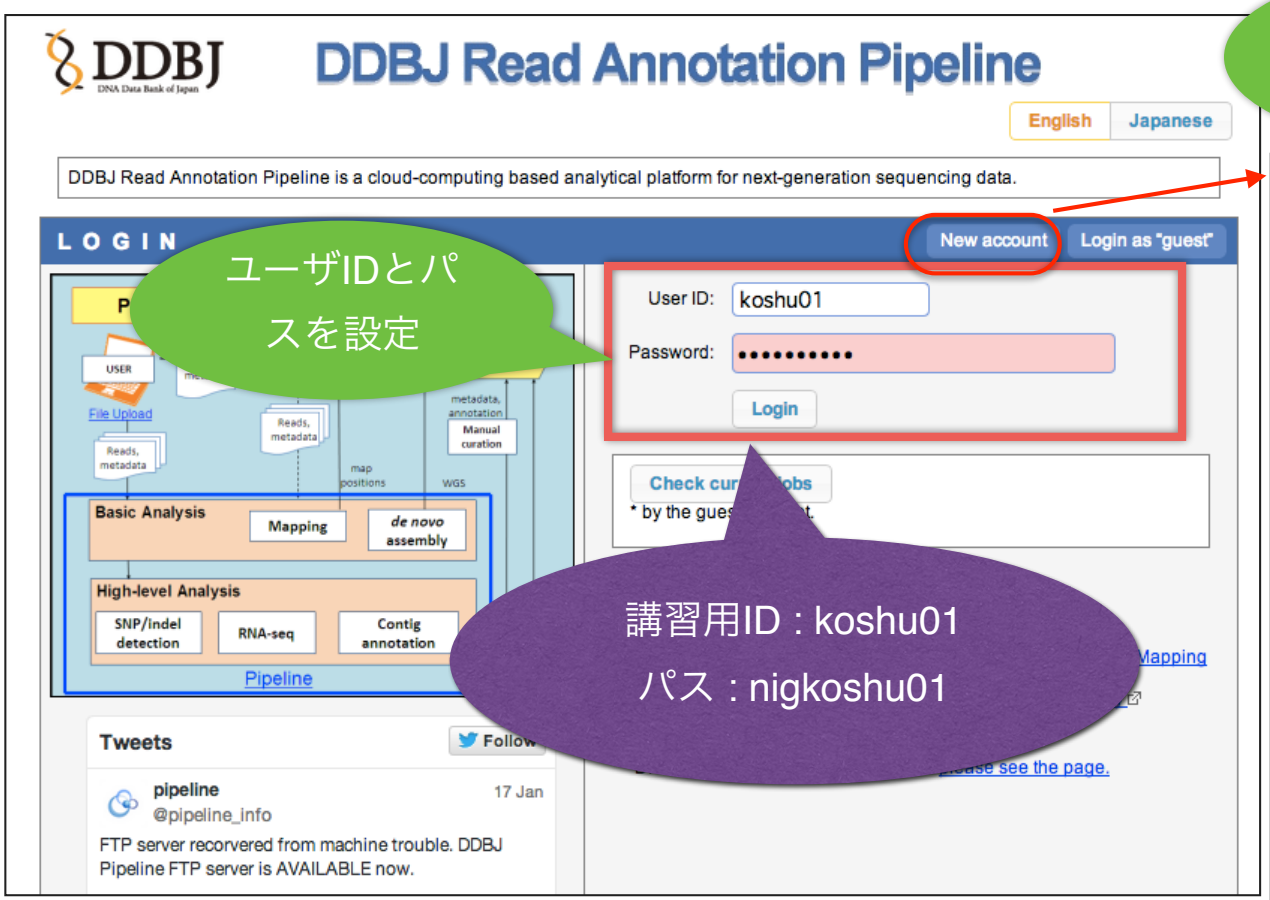
デモ

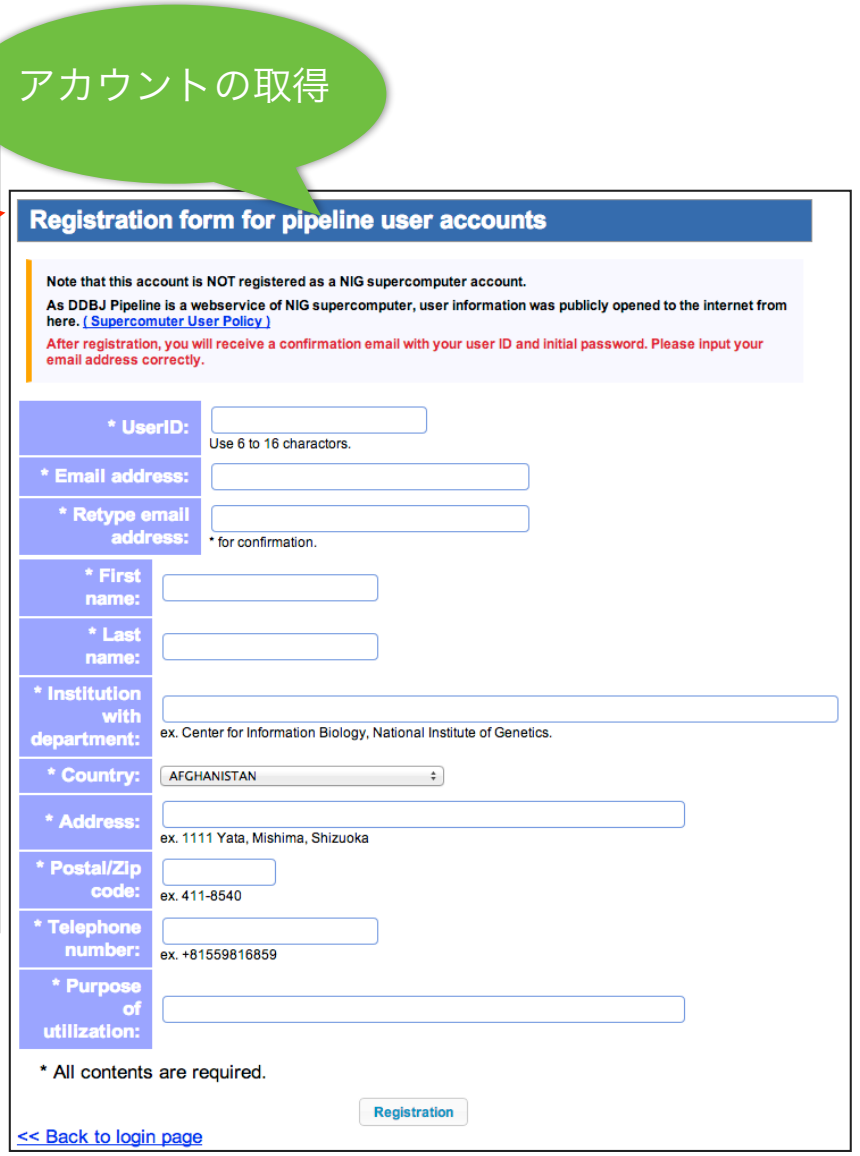
アクセス


DDBJ pipeline
+孝子さん
共有
ウェブ
ニュース
画像
キーワード検索
約 136,000 件 （0.14 秒）
DDBJ Read Annotation Pipeline
https://p.ddbj.nig.ac.jp/
DDBJ Read Annotation Pipelineは、次世代シーケンサ配列のクラウド型データ解析プラットフォームです。
DDBJ Pipeline | ynlab@nig
charles.genes.nig.ac.jp/about/research1/
DDBJ Pipeline. ◎次世代シークエンサ(NGS)の大量データ配列解析. 大量配列データの効率的処理を目的としたNGS配列解析手法の研究。遺伝研の計算機資源を利用したNGS自動配列解析システム「DDBJ Read Annotation Pipeline」を構築中。
[PDF] NGS クラウド型解析ツール DDBJ Pipeline 実習 - 国立遺伝学研...
www.ddbj.nig.ac.jp/ddbjing/dl/23-2-4.pdf
DDBJing 講習会(23) & PDBj 講習会 in 長浜. 国立遺伝学研究所. 生命情報・DDBJ研究セ


DDBJ
DNA Data Bank of Japan
English
Search
DDBJ の紹介
利用の手引き
レポート・統計
FAQ
お問い合わせ
HOME > 検索・解析
最終更新日：2014.12.18.
検索・解析
データベース検索
getentry
アクセッション番号などによるエントリの検索
ARSA
高速なキーワード検索
TXSearch
生物分類データベース検索
BLAST
相同性検索
VecScreen
ベクター配列データベースを対象にした相同性検索サービス
DRA Search
DRA に登録されたデータを，キーワード，生物名，シークエンサなどで検索
系統解析
ClustalW
WABI (We
WABI は，DDBJの検索サー
DDBJ では現在、次のサービスの Web
WABI BLAST
getentry WebAPI
ARSA WebAPI
DDBJのHPからのリンク
次世代 Sequence 解析
DDBJ Read Annotation Pipeline
次世代シークセンサー出力データの解析システム
（ご利用には申請が必要です）
ゲノム解析
MiGAP
微生物ゲノム配列のアノテーションツール
（ご利用には申請が必要です）
MiGAP-OLD (2012年２月以前の解析結果利用)
微生物ゲノム配列のアノテーションツール
（ご利用時のDBCLS のOpenID が必要です）
GTPS
共通プロトコルに基づくバクテリアゲノムの再アノテーション
GTOP
ゲノム配列からタンパク質の構造へ
DBCLS* の検索ツール
AOE
公共遺伝子発現データベースの目次
CRISPRdirect
CRISPR/Cas9システムのガイドRNAを設計することができるツール
S SRA
世代シーケンスデータの統計情報と検索機能を提供するサービス
、疾患について、関連する疾患、薬剤、臓器、生命現象などの特徴をキーワードでリスト表示するツール
GGGenome
高速ゲノム配列検索ツール
GGRNA
高速な遺伝子検索エンジン
RefEx
ヒトおよびマウス、ラットにおける遺伝子発現データのリファレンスデータセット
*DBCLS：ライフサイエンス統合データベースセンター（Database Center for Life Science）
タンパク質データベース及び構造解析
PMD
変異タンパク質データベース

ログイン


DDBJ
DNA Data Bank of Japan
DDBJ Read Annotation Pipeline
English
Japanese
DDBJ Read Annotation Pipeline is a cloud-computing based analytical platform for next-generation sequencing data.
LOGIN
New account
Login as "guest"
User ID:
koshu01
Password:
Login
ユーザIDとパスを設定
講習用ID：koshu01
パス：nigkoshu01
Basic Analysis
Mapping
de novo assembly
High-level Analysis
SNP/indel detection
RNA-seq
Contig annotation
Pipeline
Tweets
Follow
pipeline
@pipeline_info
17 Jan
FTP server recorvered from machine trouble. DDBJ Pipeline FTP server is AVAILABLE now.


アカウントの取得
Registration form for pipeline user accounts
Note that this account is NOT registered as a NIG supercomputer account.
As DDBJ Pipeline is a webservice of NIG supercomputer, user information was publicly opened to the internet from here. (Supercomuter User Policy)
After registration, you will receive a confirmation email with your user ID and initial password. Please input your email address correctly.
* UserID:
Use 6 to 16 charactors.
* Email address:
* Retype email address:
* for confirmation.
* First name:
* Last name:
* Institution with department:
ex. Center for Information Biology, National Institute of Genetics.
* Country:
AFGHANISTAN
* Address:
ex. 1111 Yata, Mishima, Shizuoka
* Postal/Zip code:
ex. 411-8540
* Telephone number:
ex. +81559816859
* Purpose of utilization:
* All contents are required.
Registration
<< Back to login page

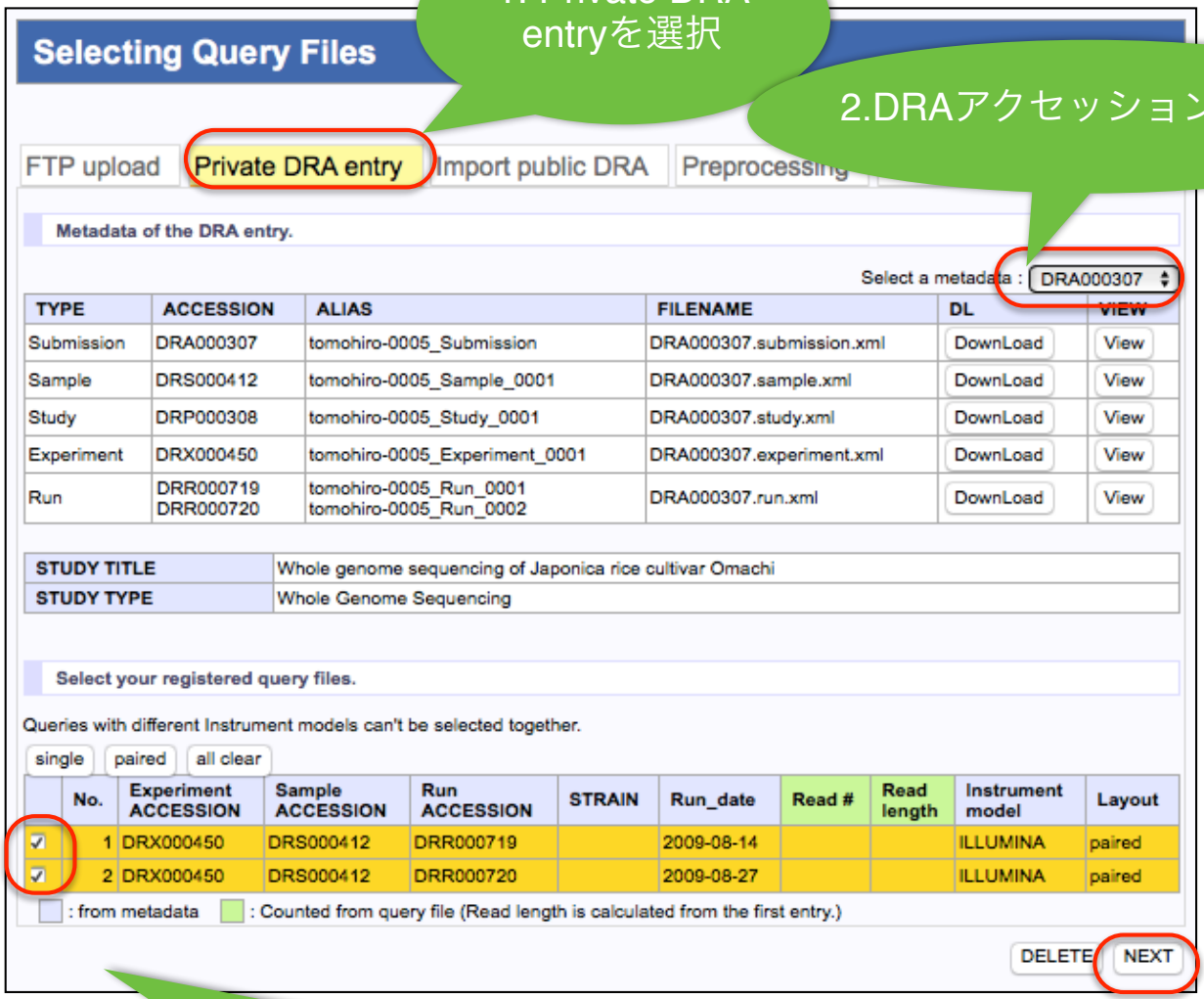
クエリの選択

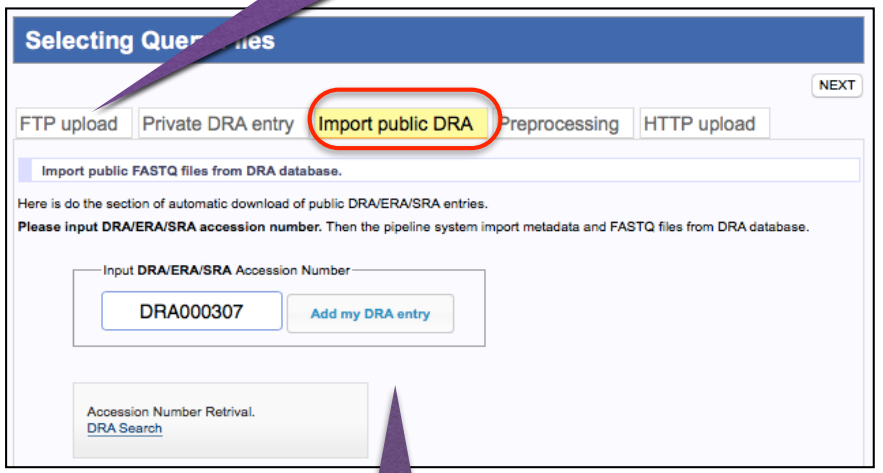
O. sativa subsp. *japonica* cv. Omach
DRA000307

1. Private DRA entryを選択

2.DRAアクセッションを選択

ユーザオリジナルデータを使用する場合は、FTP upload



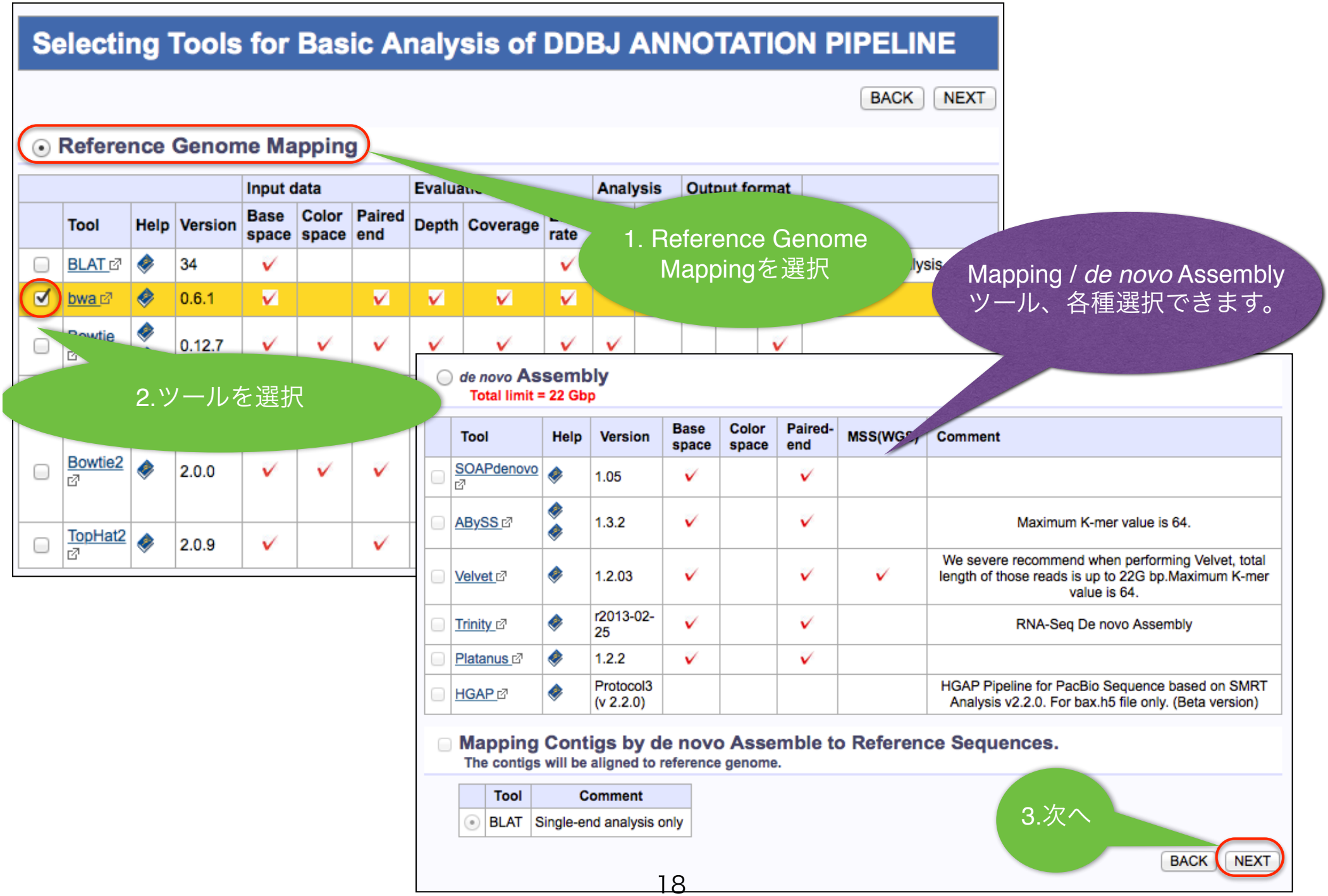


DRAデータを用いて解析するには、まず、「import public DRA」でデータをインポートしなければならない。
（今回の講習データはインポート済み）

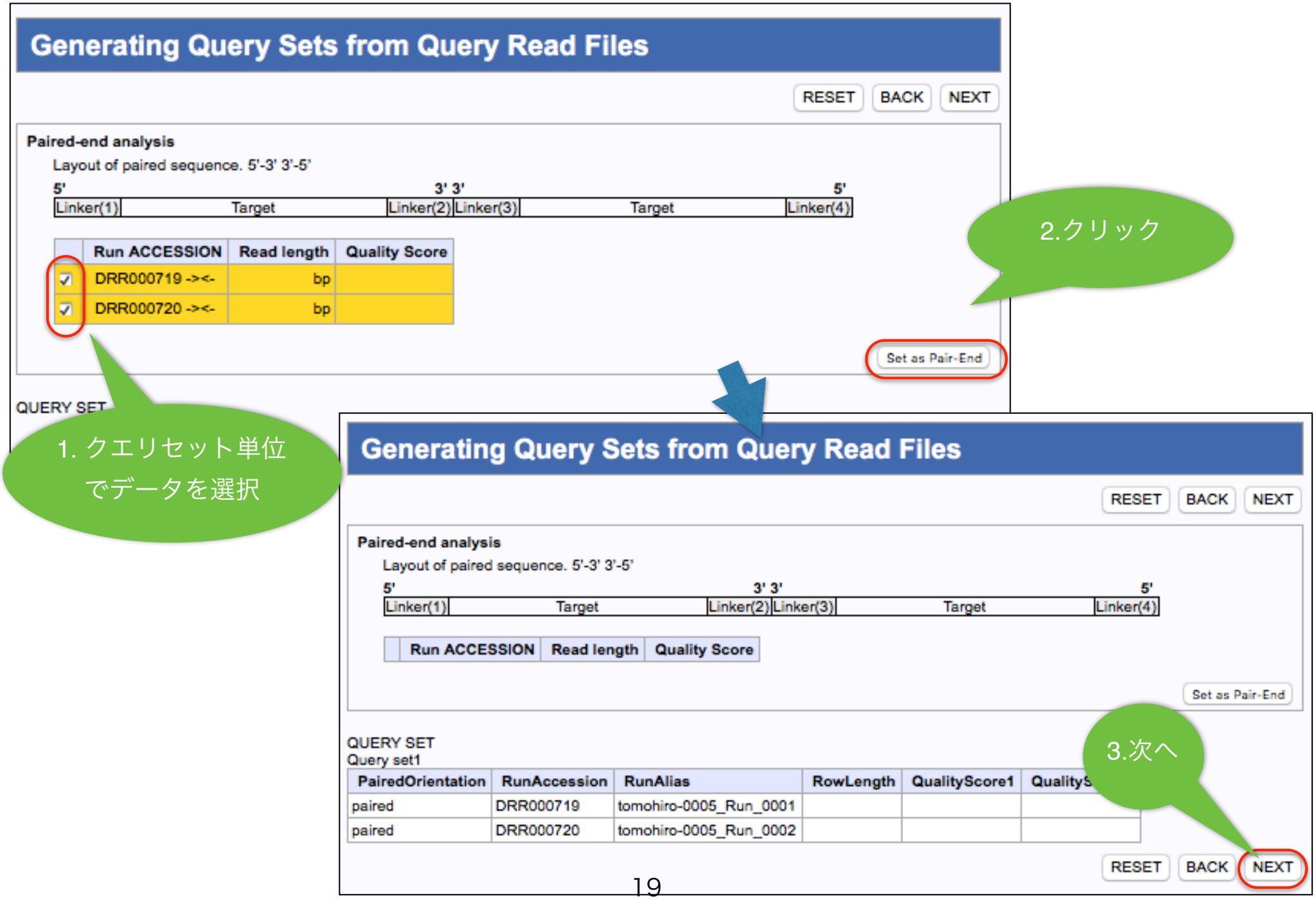
3.解析に使用するデータを選択

4.次へ

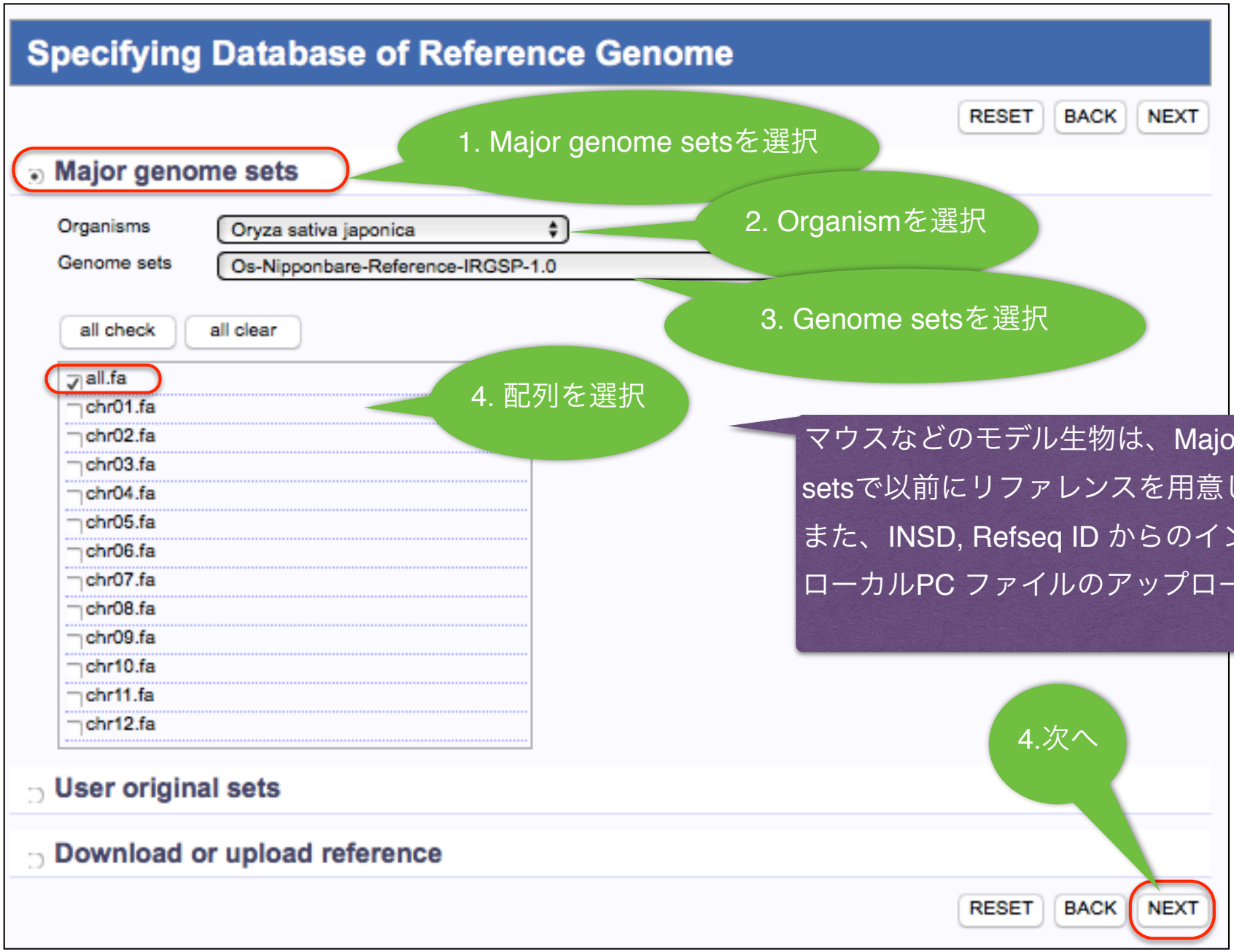
マッピングツールの選択



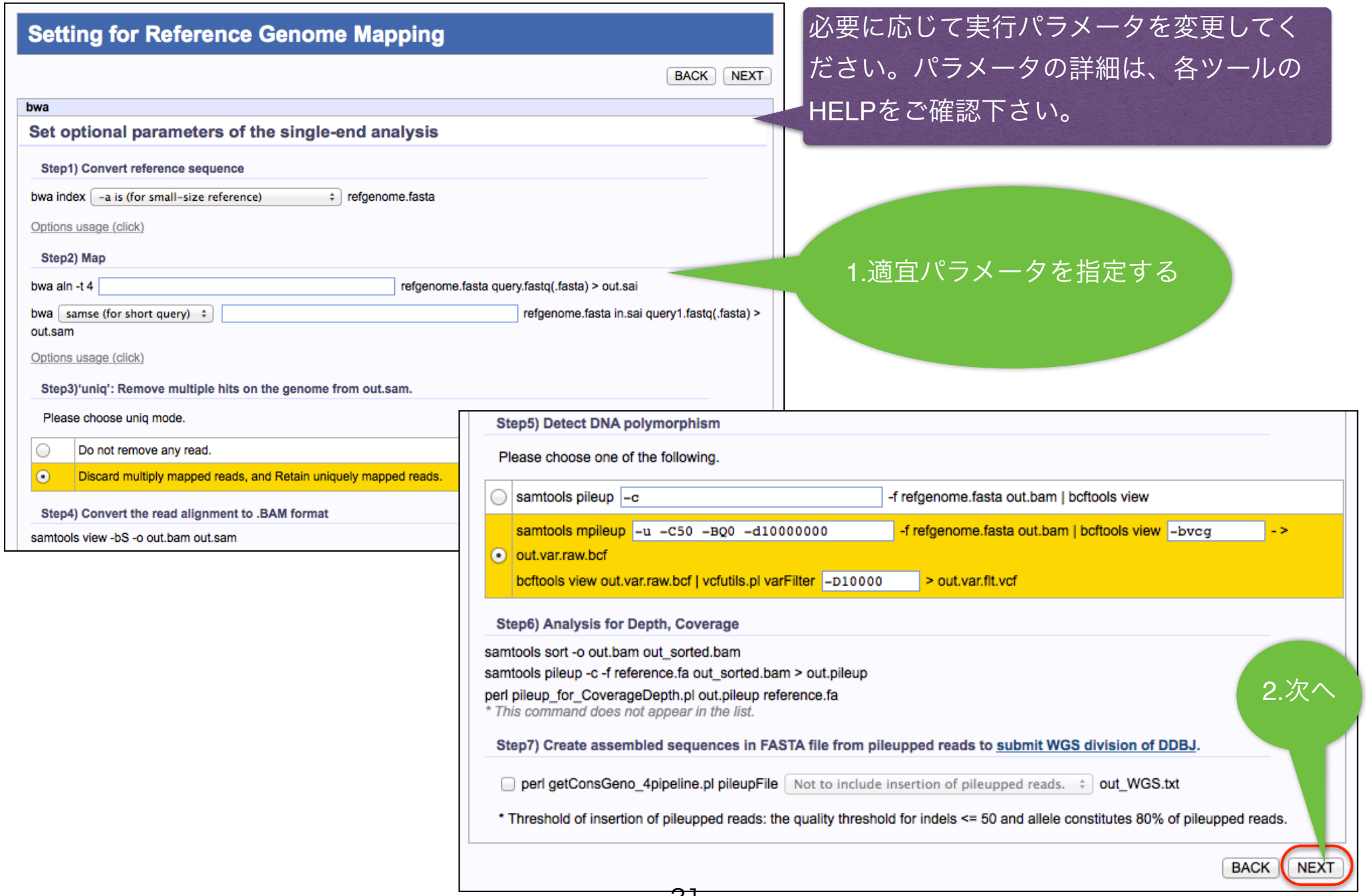
クエリセットの作成


Generating Query Sets from Query Read Files
RESET
BACK
NEXT
Paired-end analysis
Layout of paired sequence. 5'-3' 3'-5'
5'
3' 3'
5'
Linker(1)
Target
Linker(2)
Linker(3)
Target
Linker(4)
Run ACCESSION
Read length
Quality Score
DRR000719 -><-
bp
DRR000720 -><-
bp
Set as Pair-End
QUERY SET
2.クリック
1. クエリセット単位でデータを選択
Generating Query Sets from Query Read Files
RESET
BACK
NEXT
Paired-end analysis
Layout of paired sequence. 5'-3' 3'-5'
5'
3' 3'
5'
Linker(1)
Target
Linker(2)
Linker(3)
Target
Linker(4)
Run ACCESSION
Read length
Quality Score
Set as Pair-End
QUERY SET
Query set1
PairedOrientation
RunAccession
RunAlias
RowLength
QualityScore1
paired
DRR000719
tomohiro-0005_Run_0001
paired
DRR000720
tomohiro-0005_Run_0002
3.次へ
RESET
BACK
NEXT

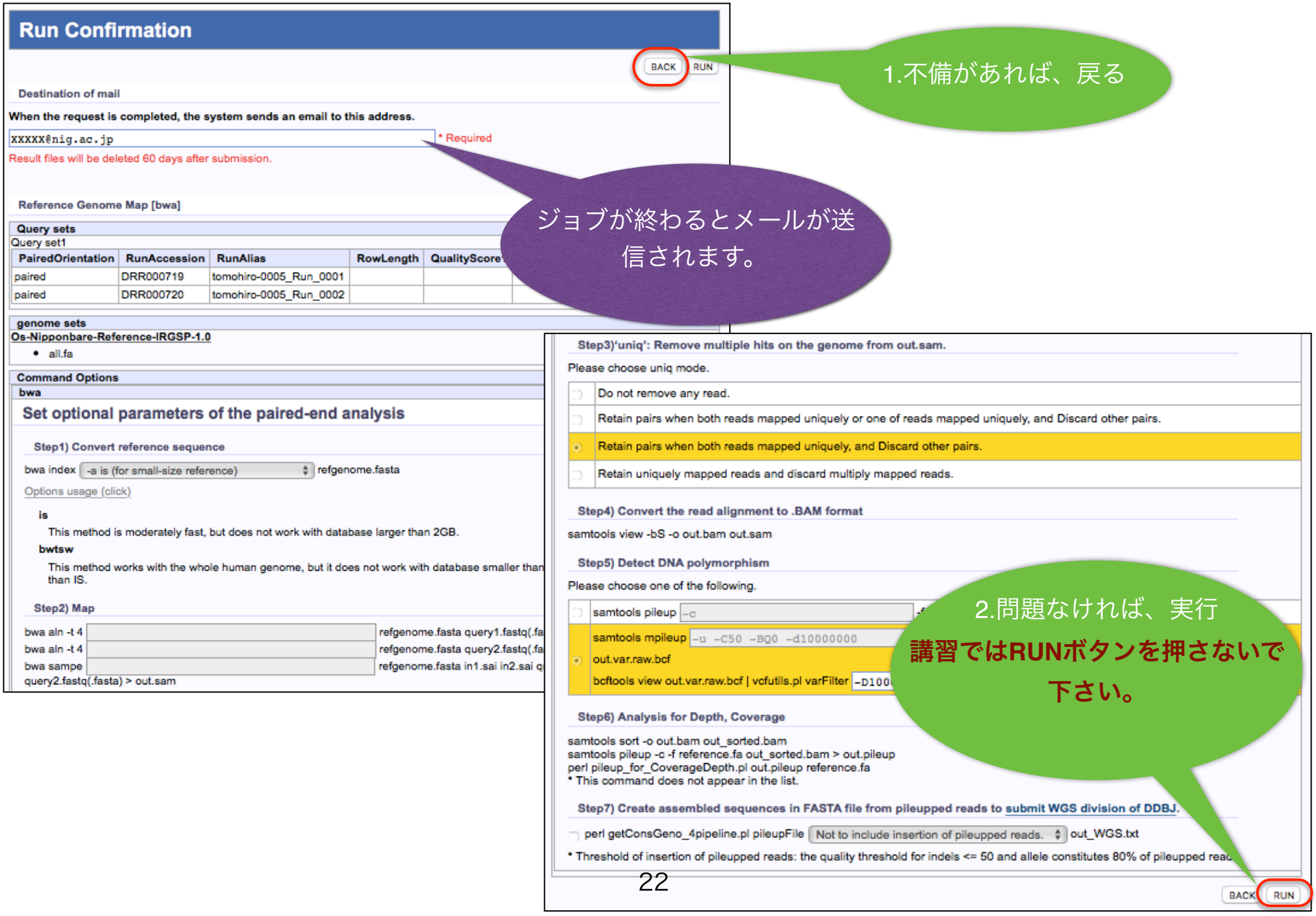
リファレンスの指定


Specifying Database of Reference Genome
RESET
BACK
NEXT
1. Major genome setsを選択
Major genome sets
Organisms
Oryza sativa japonica
2. Organismを選択
Genome sets
Os-Nipponbare-Reference-IRGSP-1.0
3. Genome setsを選択
all check
all clear
all.fa
chr01.fa
chr02.fa
chr03.fa
chr04.fa
chr05.fa
chr06.fa
chr07.fa
chr08.fa
chr09.fa
chr10.fa
chr11.fa
chr12.fa
4. 配列を選択
マウスなどのモデル生物は、Major genome setsで以前にリファレンスを用意しています。また、INSD, Refseq ID からのインポート、ローカルPC ファイルのアップロードもできま
User original sets
Download or upload reference
4.次へ
RESET
BACK
NEXT

実行パラメータの設定


Setting for Reference Genome Mapping
BACK
NEXT
bwa
Set optional parameters of the single-end analysis
Step1) Convert reference sequence
bwa index
-a is (for small-size reference)
refgenome.fasta
Options usage (click)
Step2) Map
bwa aln -t 4
refgenome.fasta query.fastq(.fasta) > out.sai
bwa
samse (for short query)
refgenome.fasta in.sai query1.fastq(.fasta) > out.sam
Options usage (click)
Step3)'uniq': Remove multiple hits on the genome from out.sam.
Please choose uniq mode.
Do not remove any read.
Discard multiply mapped reads, and Retain uniquely mapped reads.
Step4) Convert the read alignment to .BAM format
samtools view -bS -o out.bam out.sam
必要に応じて実行パラメータを変更してください。パラメータの詳細は、各ツールのHELPをご確認下さい。
1.適宜パラメータを指定する
Step5) Detect DNA polymorphism
Please choose one of the following.
samtools pileup -c
-f refgenome.fasta out.bam | bcftools view
samtools mpileup -u -C50 -BQ0 -d10000000
-f refgenome.fasta out.bam | bcftools view -bvcg - > out.var.raw.bcf
bcftools view out.var.raw.bcf | vcfutils.pl varFilter -D10000 > out.var.flt.vcf
Step6) Analysis for Depth, Coverage
samtools sort -o out.bam out_sorted.bam
samtools pileup -c -f reference.fa out_sorted.bam > out.pileup
perl pileup_for_CoverageDepth.pl out.pileup reference.fa
* This command does not appear in the list.
Step7) Create assembled sequences in FASTA file from pileupped reads to submit WGS division of DDBJ.
perl getConsGeno_4pipeline.pl pileupFile
Not to include insertion of pileupped reads.
out_WGS.txt
* Threshold of insertion of pileupped reads: the quality threshold for indels <= 50 and allele constitutes 80% of pileupped reads.
BACK
NEXT
2.次へ

実行条件の確認



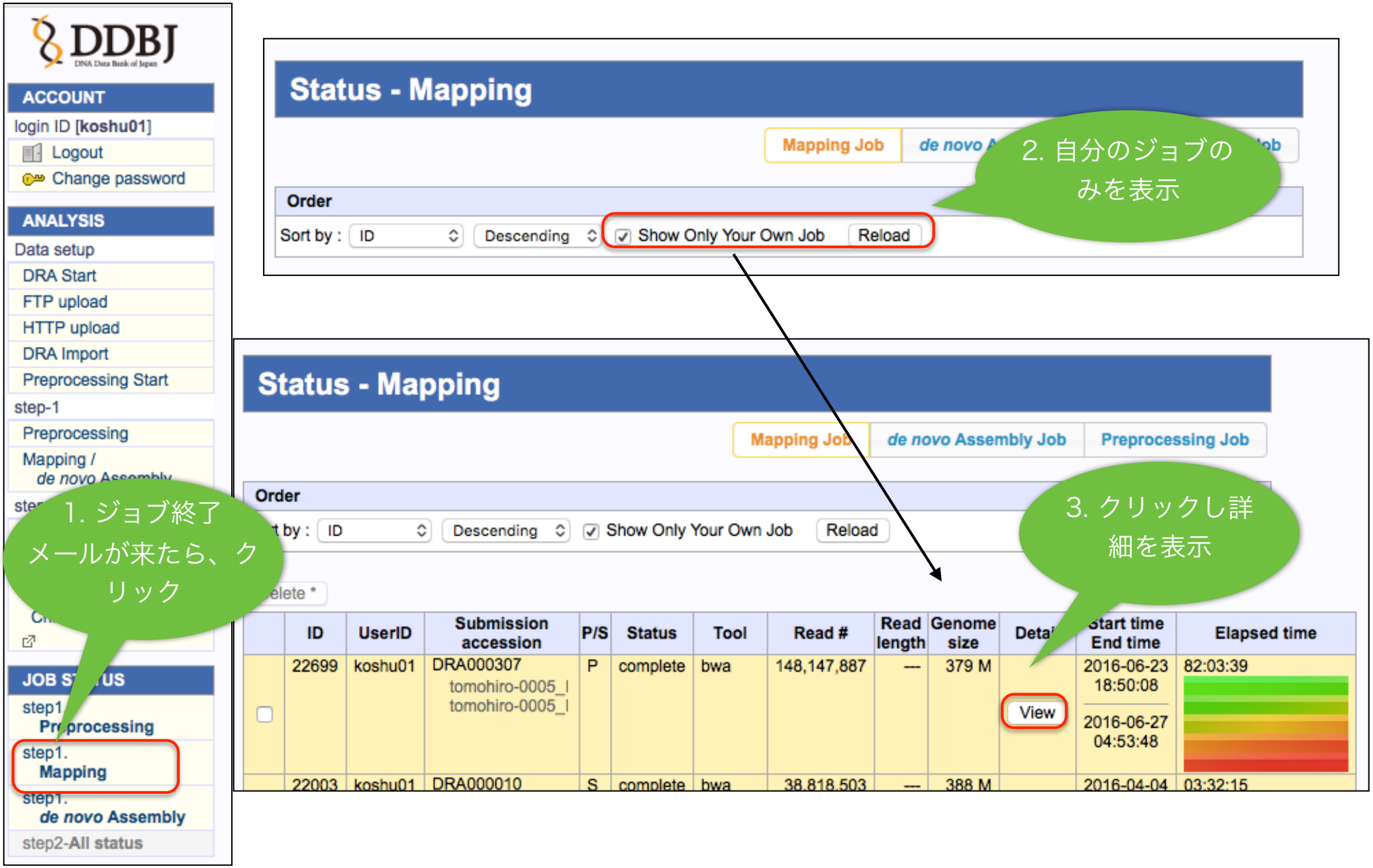
1.不備があれば、戻る

ジョブが終わるとメールが送信されます。

2.問題なければ、実行

講習ではRUNボタンを押さないで下さい。

実行結果の確認


DDBJ
ACCOUNT
login ID [koshu01]
Logout
Change password
ANALYSIS
Data setup
DRA Start
FTP upload
HTTP upload
DRA Import
Preprocessing Start
step-1
Preprocessing
Mapping /
de novo Assembly
JOB STATUS
step1.
Preprocessing
step1.
Mapping
step1.
de novo Assembly
step2-All status
Status - Mapping
Mapping Job
de novo Assembly Job
Preprocessing Job
Order
Sort by : ID
Descending
Show Only Your Own Job
Reload
ID
UserID
Submission accession
P/S
Status
Tool
Read #
Read length
Genome size
Start time
End time
Elapsed time
22699
koshu01
DRA000307
tomohiro-0005_I
tomohiro-0005_I
P
complete
bwa
148,147,887
---
379 M
View
2016-06-23 18:50:08
2016-06-27 04:53:48
82:03:39
22003
koshu01
DRA000010
S
complete
bwa
38,818,503
---
388 M
2016-04-04
03:32:15
1. ジョブ終了メールが来たら、クリック
2. 自分のジョブのみを表示
3. クリックし詳細を表示

実行結果の確認　- 詳細 -

Detail view

BACK

Job info

ID
22699

Tool (Version)
bwa (0.6.1)

RunAccession or Filename	Download	Read length	Alias
DRR000719	DRR000719.fastq.bz2 DRR000719_1.fastq.bz2 DRR000719_2.fastq.bz2	N.A. bp	tomohiro-0005_Run_0001
DRR000720	DRR000720.fastq.bz2 DRR000720_1.fastq.bz2 DRR000720_2.fastq.bz2	N.A. bp	tomohiro-0005_Run_0002

Genome set
Os-Nipponbare-Reference-IRGSP-1.0

Chromosome
all.fa

Download modified queries

- mergedout_1.fastq.gz (Original size 39.7 GB)
- mergedout_2.fastq.gz (Original size 39.7 GB)

Download merged pileup file

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

- merged.var.flt.vcf.gz (Original size 16.2 MB)
- merged.sam.gz (Original size 55.5 GB)

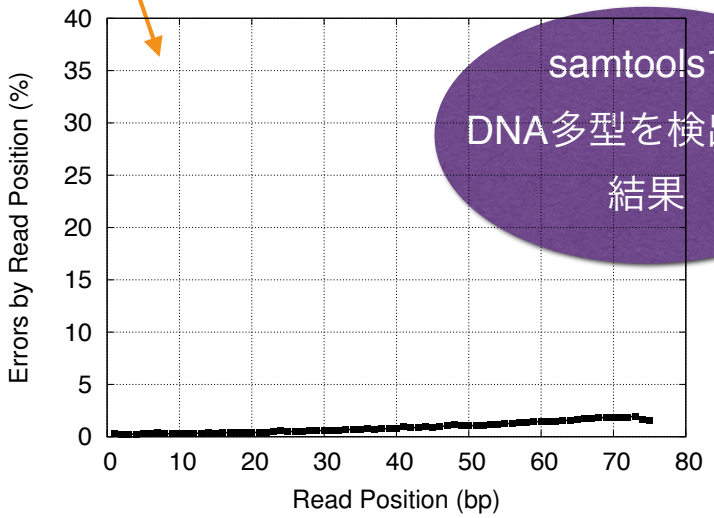
Download wgs file

- out_WGS.fasta.gz (Original size 3.7 MB)

Position errors	Depth, Coverage
PDF download	coverage : 367720365 / 373128865 * 100 = 98.551 depth : 20764213494 / 367720365 = 56.467

統計量

ユニーク化



samtoolsでDNA多型を検出した結果

bwaにてマッピング

Time

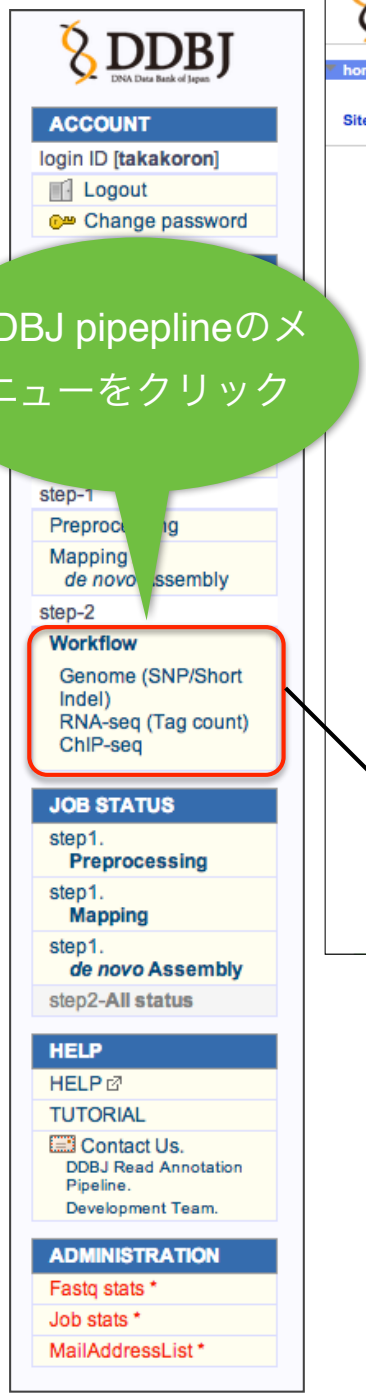
Wait time	Start time
0: 42:48	2016-06-23 18:50:08

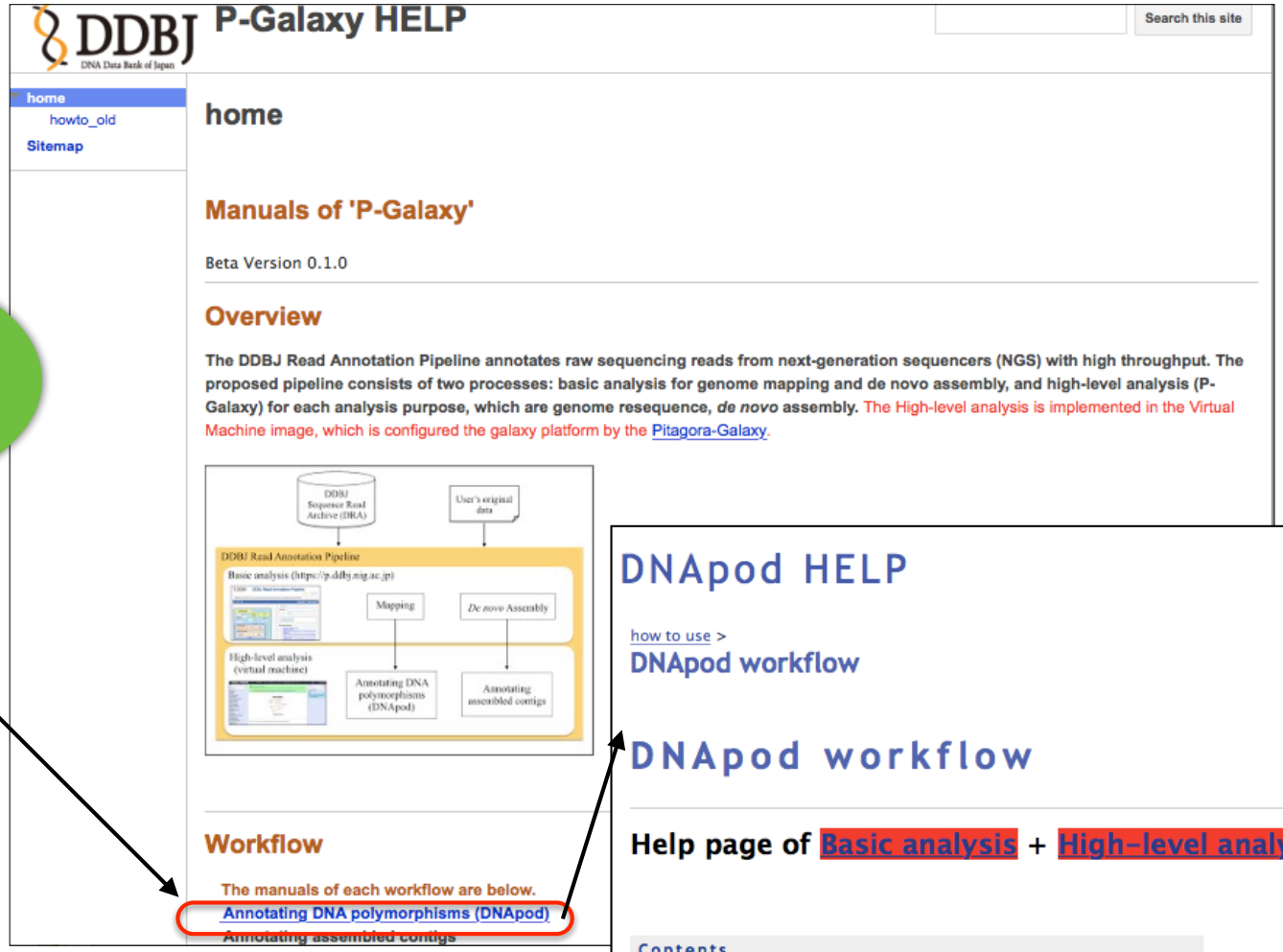
all.fa

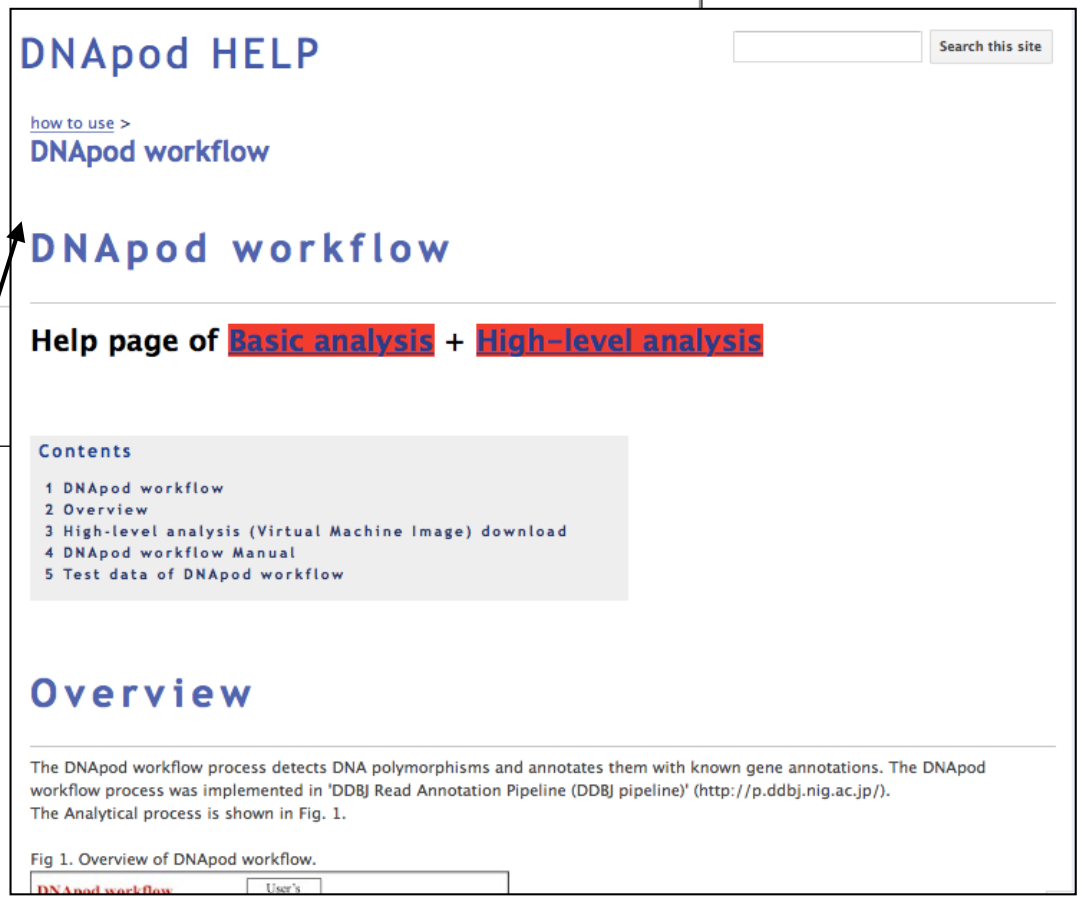
Command	Start time	End time	Log1	Log2	Result	MD5
Create BWA Index File bwa index [-a is] all.fa	2016-06-23 18:50:08	2016-06-23 18:55:40		View		
BWA : Alignment bwa aln all.fa mergedout_1.fastq > 1.sai	2016-06-23 18:55:41	2016-06-23 21:21:57		View		
BWA : Alignment bwa aln all.fa mergedout_2.fastq > 2.sai	2016-06-23 21:21:59	2016-06-23 22:56:55		View		
BWA : SAMPE bwa sampe all.fa 1.sai 2.sai mergedout_1.fastq mergedout_2.fastq > out.sam	2016-06-23 22:56:56	2016-06-24 02:03:14		View	Download(26.7 GB)	MD5
Extract Unmapped Reads python extractUnmappedFASTQ.py mergedout_1.fastq mergedout_2.fastq out.sam	2016-06-26 01:13:19	2016-06-26 02:06:16			Download(585.6 MB)	MD5
Convert SAM to BAM samtools view -bS -o out.bam out.sam	2016-06-26 02:09:48	2016-06-26 03:30:03		View	Download(27.8 GB)	MD5
Sort BAM File samtools sort out.bam out2	2016-06-26 03:52:32	2016-06-26 05:37:56		View	Download(20.4 GB)	MD5
Create BAM Index File samtools index out2.bam	2016-06-26 05:58:10	2016-06-26 05:57:31			Download(560.5 KB)	MD5
Uniquify SAM (Remove Multiple Hits) perl sam2uniq.pl out.sam UBE > uniqout.sam	2016-06-26 05:57:42	2016-06-26 06:31:05			Download(17.0 GB)	MD5
Convert SAM to BAM [For Unique SAM] samtools view -bS -o uniqout.bam uniqout.sam	2016-06-26 08:07:31	2016-06-26 09:00:33		View	Download(34.6 GB)	MD5
Sort BAM File [For Unique SAM] samtools sort uniqout.bam out2	2016-06-26 10:46:49	2016-06-26 11:50:55		View	Download(13.0 GB)	MD5
Create BAM Index File [For Unique SAM] samtools index out2.bam	2016-06-26 11:59:49	2016-06-26 12:03:06			Download(497.1 KB)	MD5
Mpileup and Create BCF File [For Unique SAM] samtools mpileup -u -C50 -BQ0 -d10000000 -f all.fa out2.bam | bcftools view -bvcg - > uniq.var.bcf	2016-06-26 12:03:17	2016-06-26 13:44:33		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	2016-06-26 13:44:34	2016-06-26 13:44:44			Download(3.1 MB)	MD5
Mpileup and Create BCF File samtools mpileup -u -C50 -BQ0 -d10000000 -f all.fa out2.bam | bcftools view -bvcg - > non-uniq.var.bcf	2016-06-26 13:44:55	2016-06-26 18:59:12		View		
Filter BCF and Convert to VCF File bcftools view non-uniq.var.bcf | perl vcfutils.pl varFilter -D10000 > out.var.flt.vcf	2016-06-26 18:59:13	2016-06-26 18:59:23			Download(4.5 MB)	MD5
PileUp from Sorted BAM File For DepthCoverage samtools pileup -c -f all.fa out2.bam > out.pileup	2016-06-26 18:59:34	2016-06-26 23:59:50				
Convert BAM to SAMX For ErrorRate samtools view -hX out.bam > out.samX	2016-06-26 23:59:50	2016-06-27 00:13:51				
Sort BAM File For MapRatio samtools sort -n out.bam out_sorted_by_name	2016-06-27 00:13:51	2016-06-27 02:18:59		View		
Convert BAM to SAMX For MapRatio samtools view -hX out_sorted_by_name.bam > out_sorted_by_name.samX	2016-06-27 02:19:00	2016-06-27 02:33:10				

アクセス

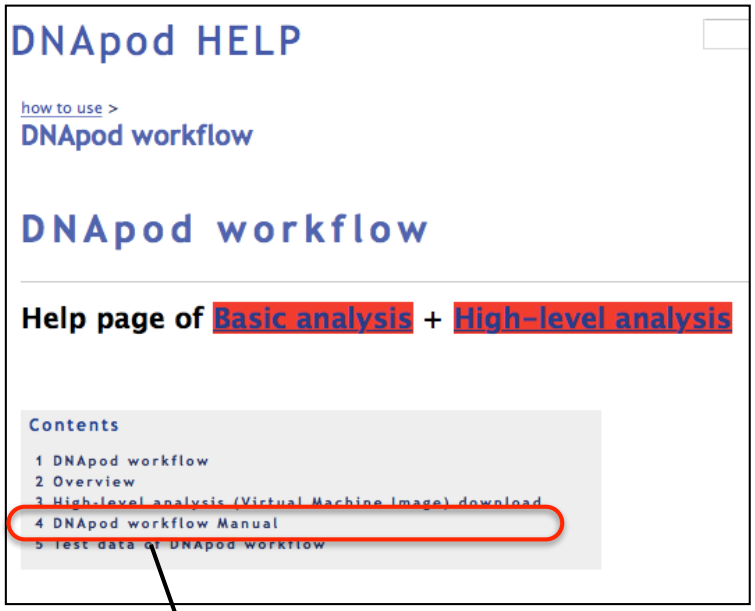
DDBJ pipeplineのメニューをクリック

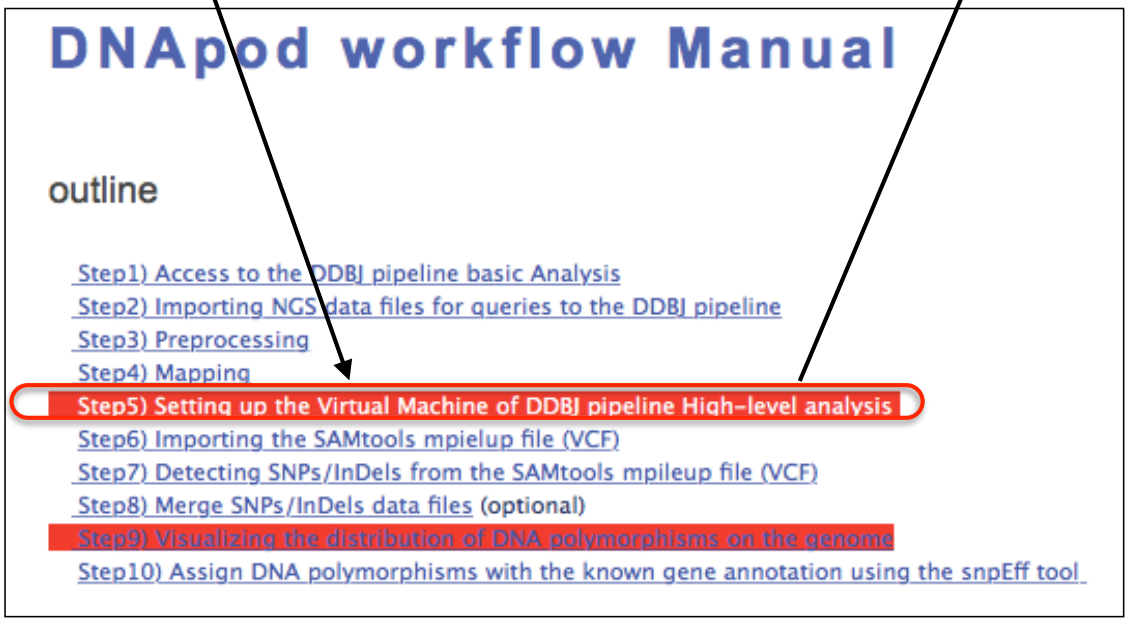






P-Galaxy 環境セットアップ (1)

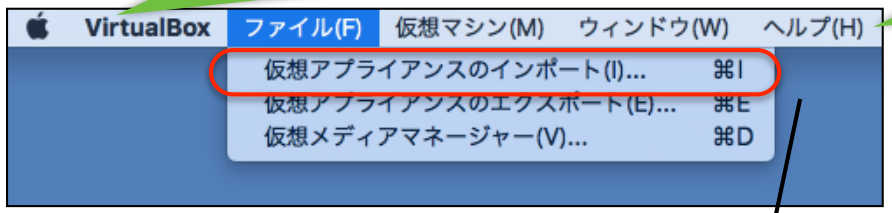




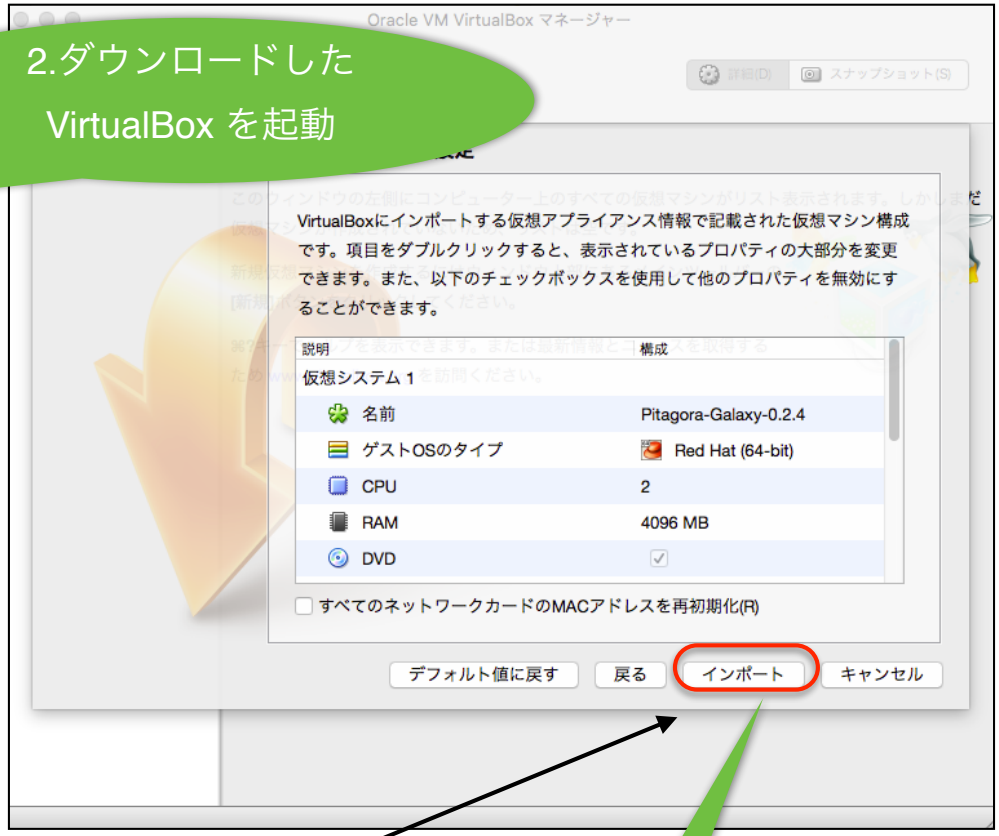
環境に合わせてダウンロード。

P-Galaxy 環境セットアップ (3)

1.ダウンロードした
VirtualBox を起動



2.ダウンロードした
VirtualBox を起動



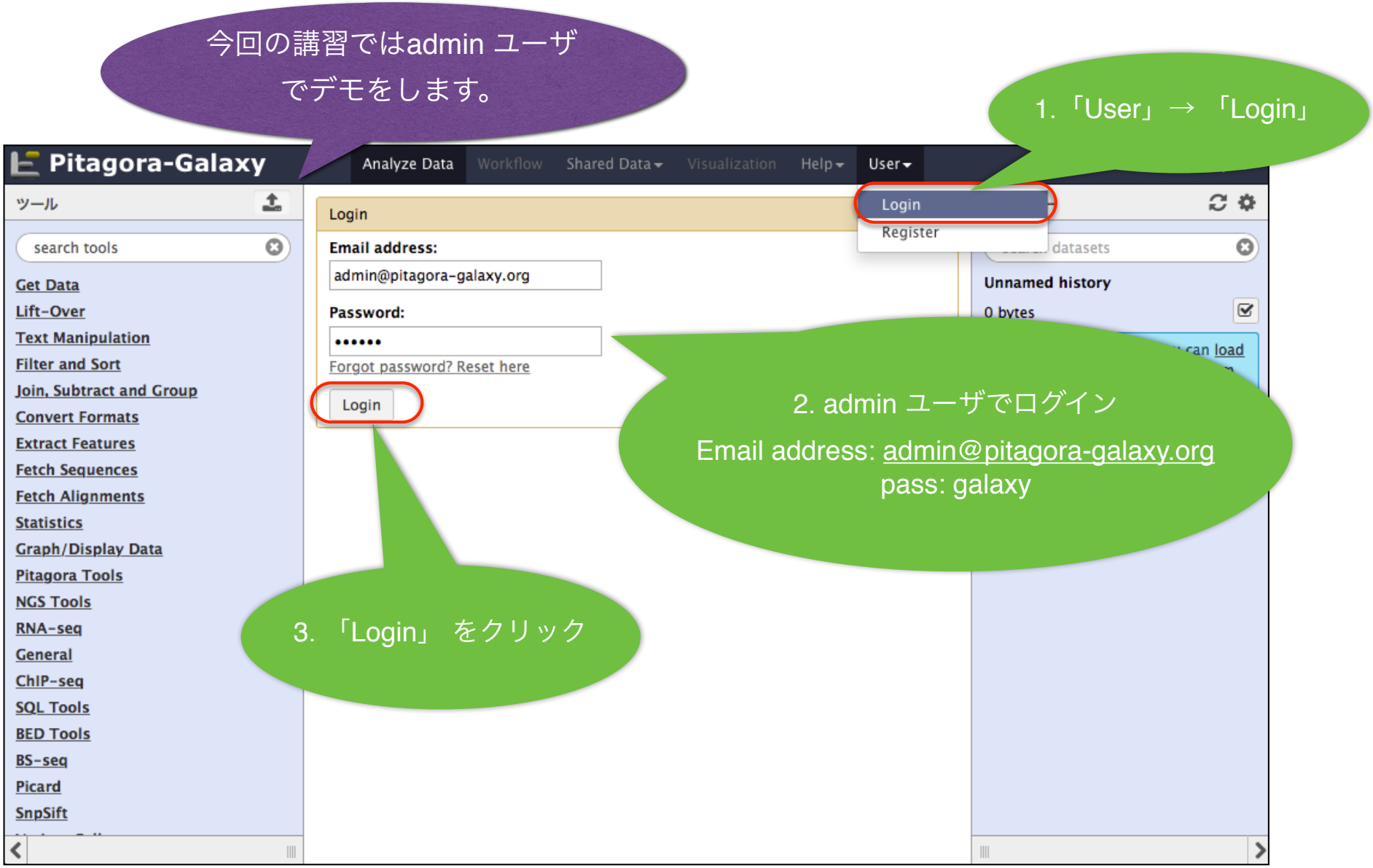
3.ダウンロードした
Virtual Machine Image を起動

4. 「続ける」 をクリック

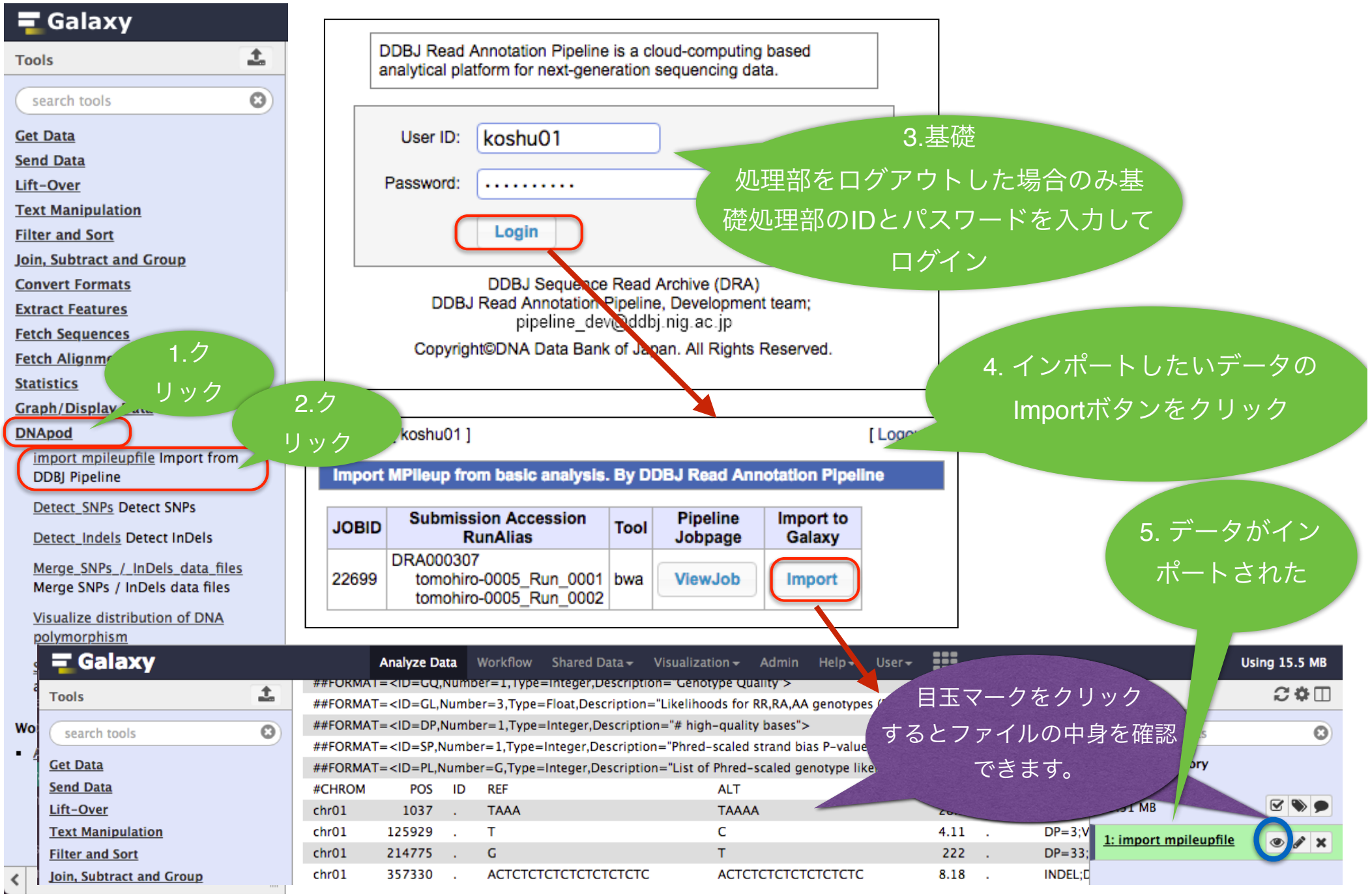
5. 「インポート」 を
クリック

P-Galaxy 環境セットアップ (4)

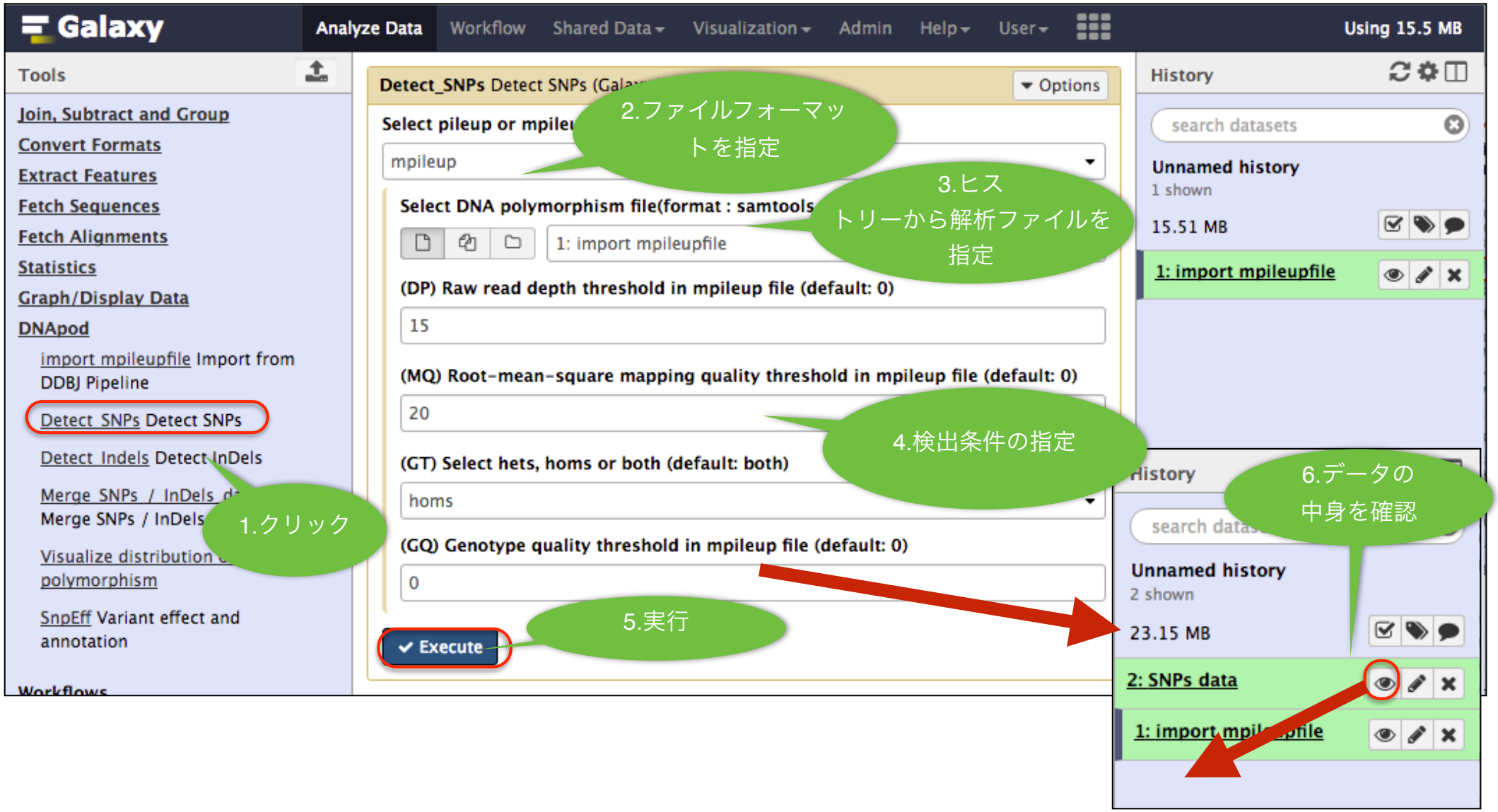
ログイン


今回の講習ではadmin ユーザ
でデモをします。
1.「User」→「Login」
Pitagora-Galaxy
Analyze Data
Workflow
Shared Data
Visualization
Help
User
Login
Register
ツール
search tools
Get Data
Lift-Over
Text Manipulation
Filter and Sort
Join, Subtract and Group
Convert Formats
Extract Features
Fetch Sequences
Fetch Alignments
Statistics
Graph/Display Data
Pitagora Tools
NGS Tools
RNA-seq
General
ChIP-seq
SQL Tools
BED Tools
BS-seq
Picard
SnpSift
Login
Email address:
admin@pitagora-galaxy.org
Password:
Forgot password? Reset here
Login
Unnamed history
0 bytes
2. admin ユーザでログイン
Email address: admin@pitagora-galaxy.org
pass: galaxy
3.「Login」 をクリック

基礎処理部のsamtools mpileupのデータインポート



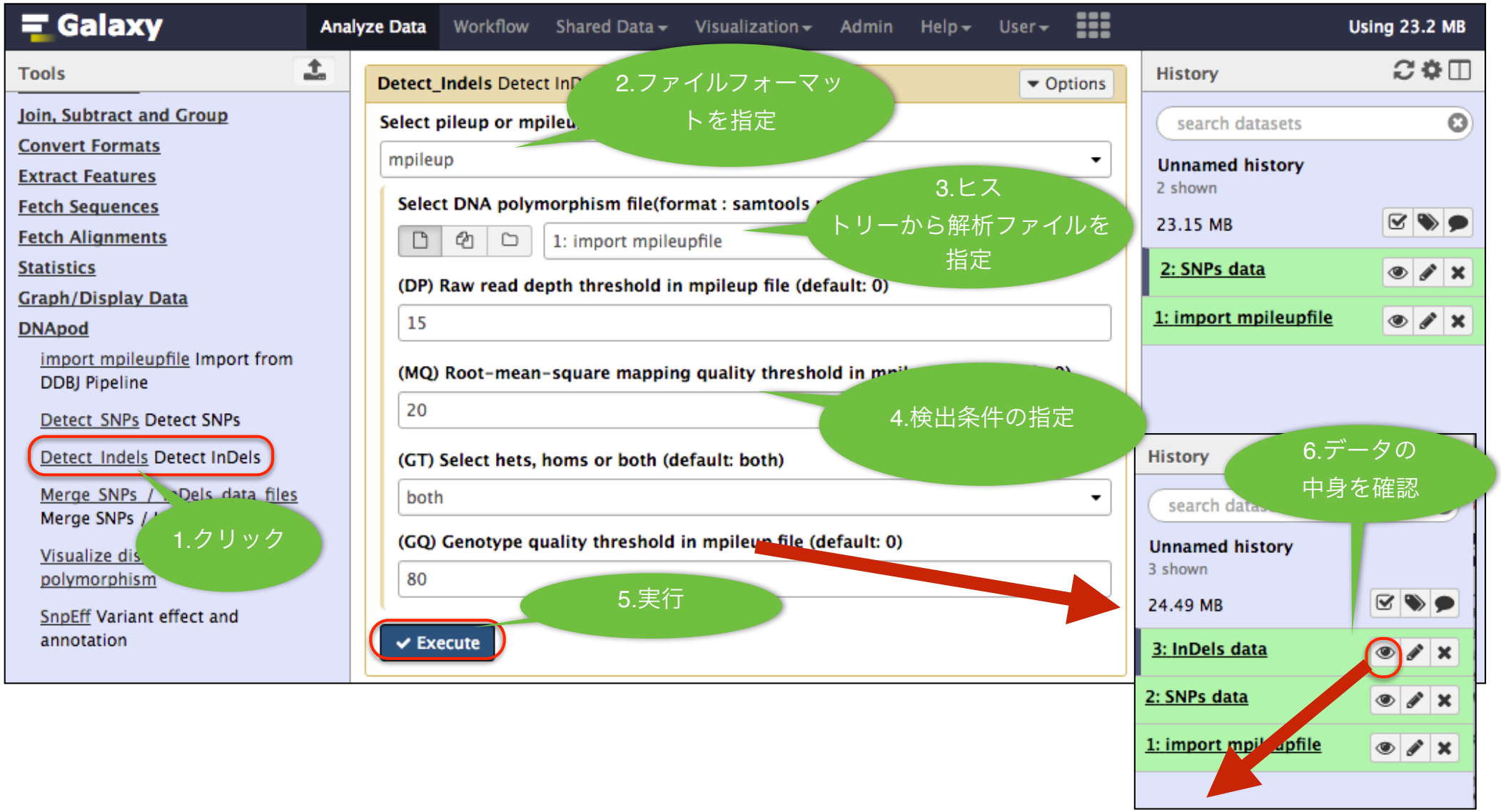
ホモSNPsの検出



#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	/home/w3pipeline/re
chr01	214775	.	G	T	222	.	DP=33;VDB=0.0539;AF1=1;AC1=2;DP4=0,0,20,13;MQ=45;FQ=-126	GT:PL:GQ	1/1:255,99,0:99
chr01	401692	.	A	T	210	.	DP=42;VDB=0.0535;AF1=1;AC1=2;DP4=0,1,13,28;MQ=49;FQ=-131;PV4=1,1,0.25,0.18	GT:PL:GQ	1/1:243,104,0:99
chr01	618781	.	C	A	219	.	DP=16;VDB=0.0475;AF1=1;AC1=2;DP4=0,0,5,11;MQ=45;FQ=-75	GT:PL:GQ	1/1:252,48,0:93

指定した閾値以上かつ、GT 1/1 でホモSNPのデータのみが出力されている。

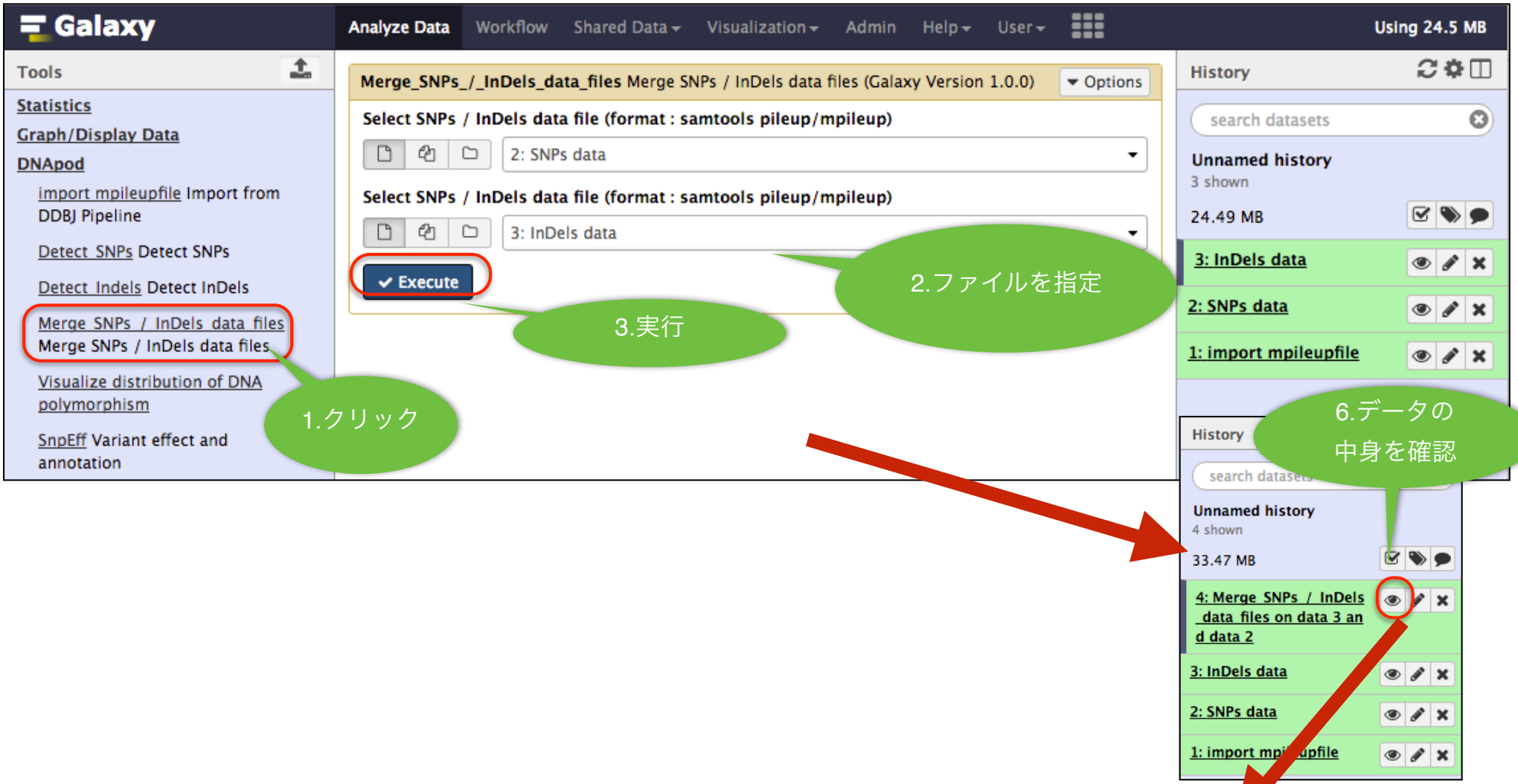
ホモInDelsの検出



#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	/home/w3pipeline/refdata/
chr01	431976	.	gttttttt	gttttt	80.5	.	INDEL;DP=46;VDB=0.0535;AF1=1;AC1=2;DP4=0,0,26,20;MQ=50;FQ=-173	GT:PL:GQ	1/1:121,138,0:99
chr01	626128	.	CTCTTCTTCTT	CTCTTCTT	214	.	INDEL;DP=41;VDB=0.0535;AF1=1;AC1=2;DP4=0,0,11,30;MQ=50;FQ=-158	GT:PL:GQ	1/1:255,123,0:99
chr01	639993	.	CTTTTTTTTTT	CTTTTTTTTTT,CTTTTTT	14.6	.	INDEL;DP=61;VDB=0.0535;AF1=1;AC1=2;DP4=0,0,35,26;MQ=50;FQ=-183	GT:PL:GQ	1/1:55,148,0,89,59,52:84

指定した閾値以上かつ、GT 1/1 でホモSNPのデータのみが出力されている。

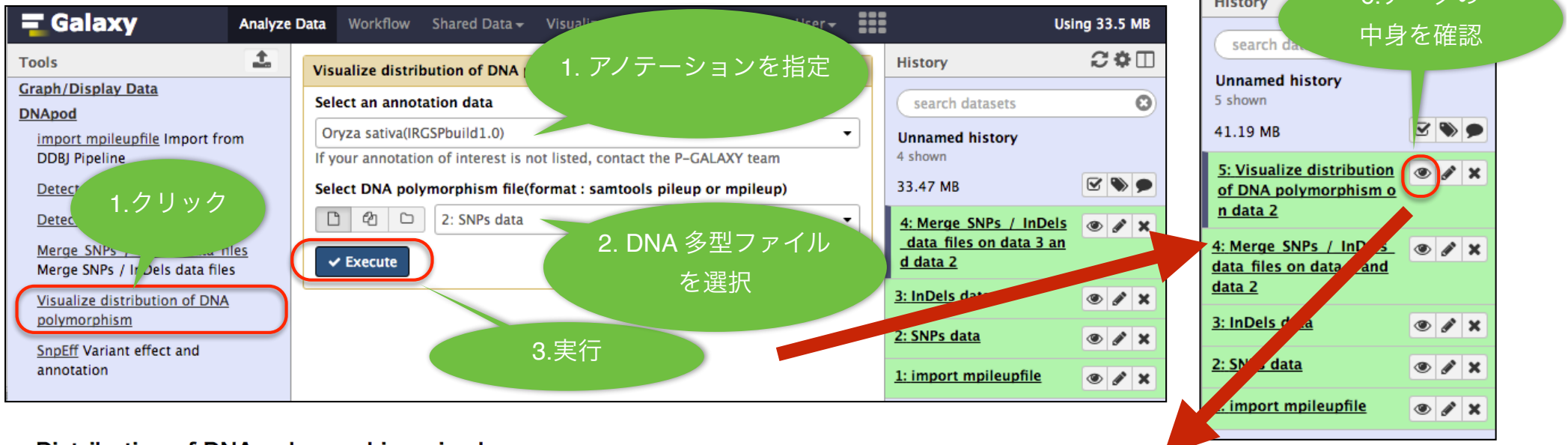
ホモSNPs, InDels ファイルのマージ

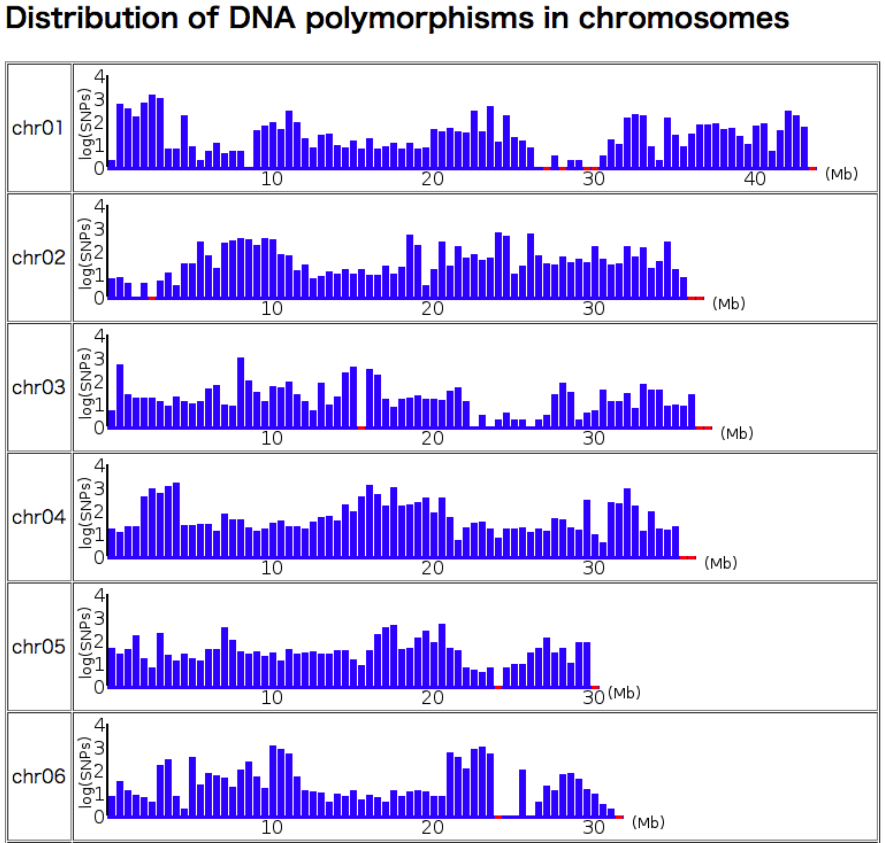


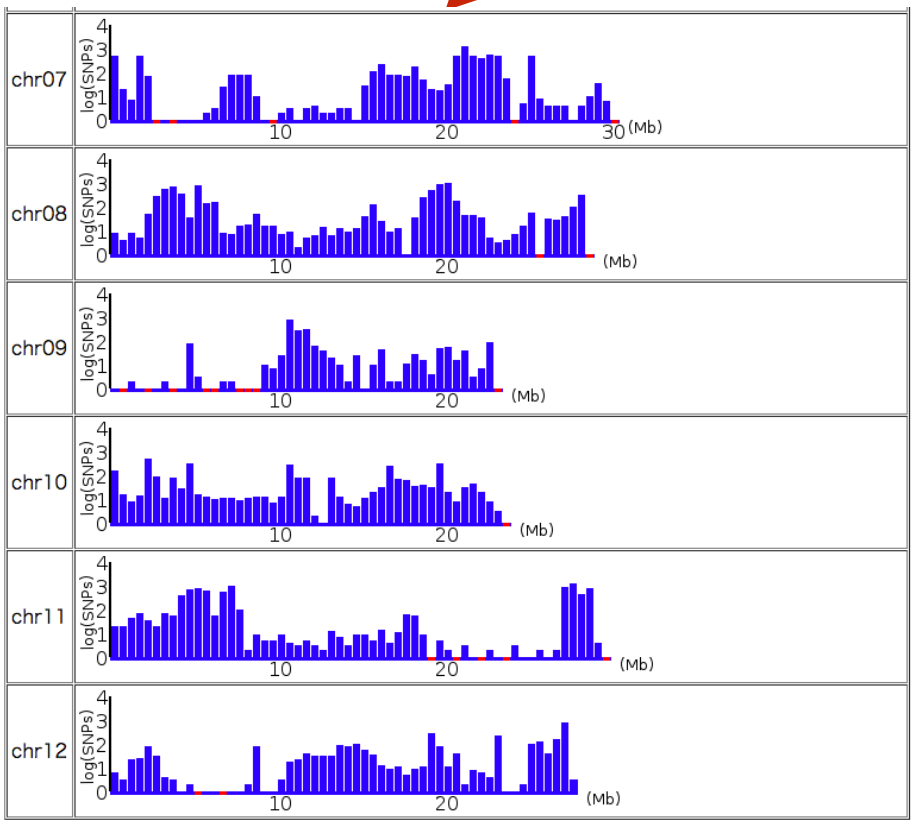
#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	/home/w3pipeline/refdata
chr01	214775	.	G	T	222	.	DP=33;VDB=0.0539;AF1=1;AC1=2;DP4=0,0,20,13;MQ=45;FQ=-126	GT:PL:GQ	1/1:255,99,0:99
chr01	401692	.	A	T	210	.	DP=42;VDB=0.0535;AF1=1;AC1=2;DP4=0,1,13,28;MQ=49;FQ=-131;PV4=1,1,0.25,0.18	GT:PL:GQ	1/1:243,104,0:99
chr01	431976	.	gttttttt	gttttttt	80.5	.	INDEL;DP=46;VDB=0.0535;AF1=1;AC1=2;DP4=0,0,26,20;MQ=50;FQ=-173	GT:PL:GQ	1/1:121,138,0:99
chr01	618781	.	C	A	219	.	DP=16;VDB=0.0475;AF1=1;AC1=2;DP4=0,0,5,11;MQ=45;FQ=-75	GT:PL:GQ	1/1:252,48,0:93
chr01	626128	.	CTCTTCTTCTT	CTCTTCTT	214	.	INDEL;DP=41;VDB=0.0535;AF1=1;AC1=2;DP4=0,0,11,30;MQ=50;FQ=-158	GT:PL:GQ	1/1:255,123,0:99
chr01	639993	.	CTTTTTTTTTTT	CTTTTTTTTTT,CTTTTTTTTT	14.6	.	INDEL;DP=61;VDB=0.0535;AF1=1;AC1=2;DP4=0,0,35,26;MQ=50;FQ=-183	GT:PL:GQ	1/1:55,148,0,89,59,52:84

SNPs と InDels ファイルがマージされている。

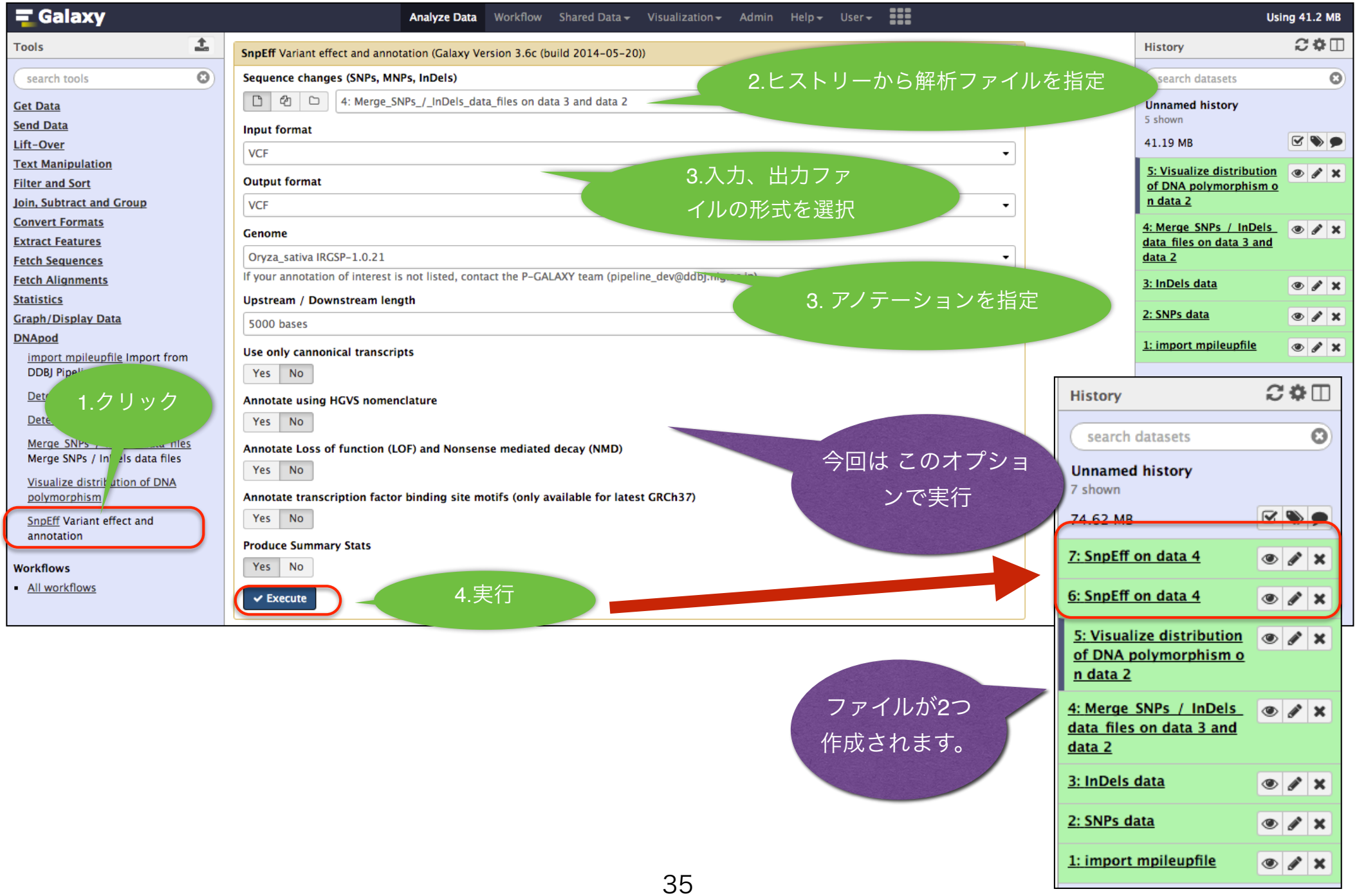
ゲノム上のDNA多型の分布を描画







SNPsアノテーション　SnpEff



SNPsアノテーション SnpEff 出力ファイル (1) アノテーション情報

vcfファイルのINFOフィールド内にEFF=でアノテーションが付与される。

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO
chr01	842425	.	T	C	138	.	DP=17;VDB=0.0059;AF1=1;AC1=2;DP4=0,0,17,0;MQ=44;FQ
chr01	842629	.	A	G	217	.	DP=33;VDB=0.0264;AF1=1;AC1=2;DP4=0,0,5,28;MQ=47;FQ
chr01	843085	.	T	A	124	.	DP=16;VDB=0.0004;AF1=1;AC1=2;DP4=0,0,16,0;MQ=37;FQ
chr01	843466	.	G	A	222	.	DP=45;VDB=0.0410;AF1=1;AC1=2;DP4=0,0,40,5;MQ=45;FQ
chr01	843604	.	C	A	222	.	DP=59;VDB=0.0534;AF1=1;AC1=2;DP4=0,0,50,9;MQ=46;FQ

…

;EFF=EXON(MODIFIER||||OS01G0115533|ncRNA|NON_CODING|OS01T0115533-00|2|1),SYNONYMOUS_CODING(LOW|SILENT|tcA/tcG|S111|251|OS01G0115566|protein_coding|CODING|OS01T0115566-00|1|1)
6;EFF=EXON(MODIFIER||||OS01G0115533|ncRNA|NON_CODING|OS01T0115533-00|2|1),SYNONYMOUS_CODING(LOW|SILENT|ctT/ctC|L43|251|OS01G0115566|protein_coding|CODING|OS01T0115566-00|1|1)
;EFF=DOWNSTREAM(MODIFIER||342|||OS01G0115533|ncRNA|NON_CODING|OS01T0115533-00||1),INTERGENIC(MODIFIER|||||||||1)
2;EFF=DOWNSTREAM(MODIFIER||723|||OS01G0115533|ncRNA|NON_CODING|OS01T0115533-00||1),INTERGENIC(MODIFIER|||||||||1)
5;EFF=NON_SYNONYMOUS_CODING(MODERATE|MISSENSE|tCt/tAt|S21Y|620|OS01G0115600|protein_coding|CODING|OS01T0115600-01|1|1|WARNING_TRANSCRIPT_INCOMPLETE)

FORMAT	
GT:PL:GQ	1/1:171,51,0:99
GT:PL:GQ	1/1:250,99,0:99
GT:PL:GQ	1/1:157,48,0:93
GT:PL:GQ	1/1:255,135,0:99
GT:PL:GQ	1/1:255,178,0:99

詳細はSnpEffのサイトを参照

http://snpeff.sourceforge.net/SnpEff_manual.html#input

SNPsアノテーション SnpEff 出力ファイル (2) 統計情報

解析統計情報を表示

Change rate details

Chromosome	Length	Changes	Change rate
1	43,270,923	9,469	4,569
2	35,937,250	8,238	4,362
3	36,413,819	4,860	7,492
4	35,502,694	13,819	2,569
5	29,958,434	4,940	6,064
6	31,248,787	8,792	3,554
7	29,697,621	7,486	3,967
8	28,443,022	8,091	3,515
9	23,012,720	2,389	9,632
10	23,207,287	3,339	6,950
11	29,021,106	9,280	3,127
12	27,531,856	3,231	8,521
Total	373,245,519	83,934	4,446

Number changes by type

Type	Total	Homo	Hetero
SNP	73,147	73,147	0
MNP	0	0	0
INS	4,087	3,961	126
DEL	6,700	6,429	271
MIXED	0	0	0
Interval	0	0	0
Total	83,934	83,537	397

Number of effects by impact

Type (alphabetical order)	Count	Percent
HIGH	307	0.093%
LOW	3,030	0.916%
MODERATE	3,431	1.038%
MODIFIER	323,849	97.953%

Number of effects by functional class

Type (alphabetical order)	Count	Percent
MISSENSE	3,472	57.512%
NONSENSE	75	1.242%
SILENT	2,490	41.246%

Missense / Silent ratio: 1.3944

Number of effects by type and region

Type (alphabetical order)	Count	Percent
CODON_CHANGE_PLUS_CODON_DELETION	14	0.004%
CODON_CHANGE_PLUS_CODON_INSERTION	3	0.001%
CODON_DELETION	17	0.005%
CODON_INSERTION	8	0.002%
DOWNSTREAM	119,103	36.024%
EXON	2,657	0.804%
FRAME_SHIFT	117	0.035%
INTERGENIC	64,193	19.416%
INTRON	12,539	3.793%
NON_SYNONYMOUS_CODING	3,389	1.025%
SPLICE_SITE_ACCEPTOR	12	0.004%
SPLICE_SITE_DONOR	14	0.004%
SPLICE_SITE_REGION	307	0.093%
START_GAINED	233	0.07%
START_LOST	9	0.003%
STOP_GAINED	76	0.023%
STOP_LOST	79	0.024%
SYNONYMOUS_CODING	2,479	0.75%
SYNONYMOUS_STOP	11	0.003%
UPSTREAM	120,553	36.463%
UTR_3_PRIME	3,155	0.954%
UTR_5_PRIME	1,649	0.499%

Region

Type (alphabetical order)	Count	Percent
DOWNSTREAM	119,103	36.024%
EXON	8,859	2.68%
INTERGENIC	64,193	19.416%
INTRON	12,539	3.793%
SPLICE_SITE_ACCEPTOR	12	0.004%
SPLICE_SITE_DONOR	14	0.004%
SPLICE_SITE_REGION	307	0.093%
UPSTREAM	120,553	36.463%
UTR_3_PRIME	3,155	0.954%
UTR_5_PRIME	1,882	0.569%

Base changes (SNPs)

	A	C	G	T
A	0	2,841	12,573	3,225
C	2,922	0	2,114	12,945
G	12,916	1,956	0	2,944
T	3,267	12,673	2,771	0

Ts/Tv (transitions / transversions)

Note: Only SNPs are used for this statistic.
Note: This Ts/Tv ratio is a 'raw' ratio. Some people prefer to use a ratio of rates, not observed events. In that case, you need to multiply by 2.0 (since there are twice as many possible transitions than transversions, E[Ts/Tv] ratio is twice the ratio of events).

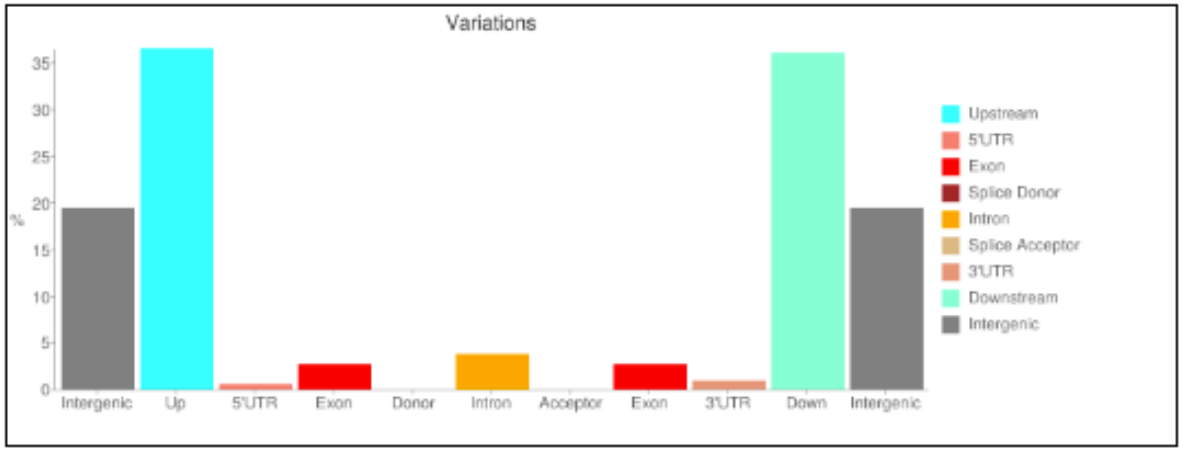
Transitions	51,107
Transversions	22,040
Ts/Tv ratio	2.3188

All variants:

Sample ,/home/w3pipeline/refdata/result/koshu01/22699/22506/36398/uniqsam/out2.bam,Total
Transitions ,102214,102214
Transversions ,44080,44080
Ts/Tv ,2.319,2.319

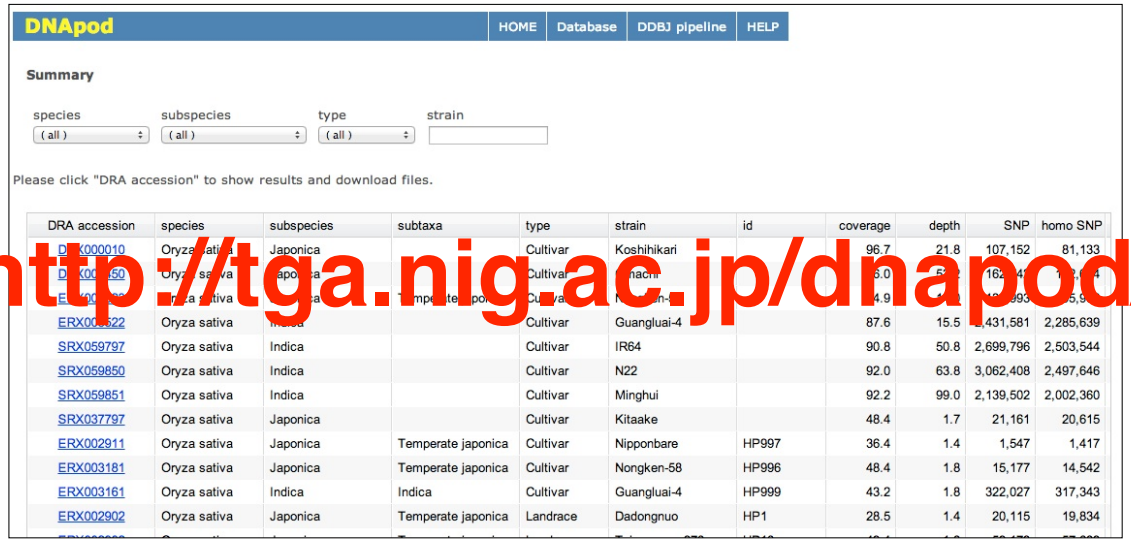
Only known variants (i.e. the ones having a non-empty ID field):

No results available (empty input?)



ご清聴ありがとうございました

DNApod データベース

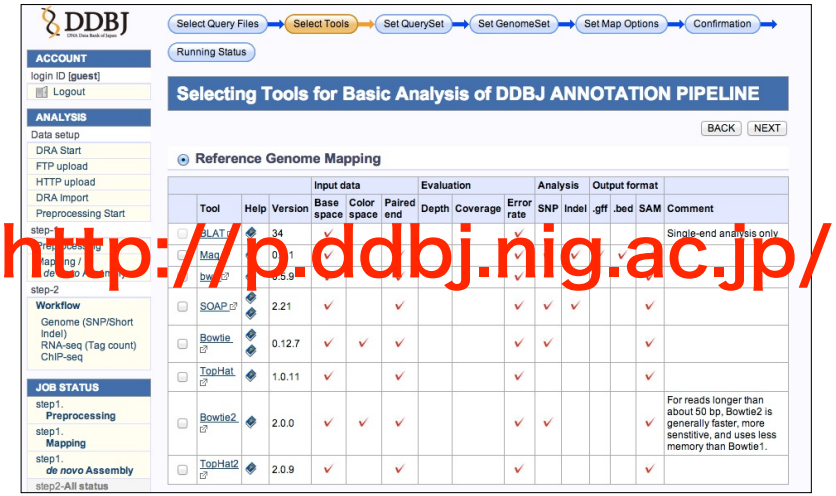


DRA accession	species	subspecies	subtaxa	type	strain	id	coverage	depth	SNP	homo SNP
DRX000010	Oryza sativa	Japonica		Cultivar	Koshihikari		96.7	21.8	107,152	81,133
ERX000322	Oryza sativa			Cultivar	Guangluai-4		87.6	15.5	1,431,581	2,285,639
SRX059797	Oryza sativa	Indica		Cultivar	IR64		90.8	50.8	2,699,796	2,503,544
SRX059850	Oryza sativa	Indica		Cultivar	N22		92.0	63.8	3,062,408	2,497,646
SRX059851	Oryza sativa	Indica		Cultivar	Minghui		92.2	99.0	2,139,502	2,002,360
SRX037797	Oryza sativa	Japonica		Cultivar	Kitaake		48.4	1.7	21,161	20,615
ERX002911	Oryza sativa	Japonica	Temperate japonica	Cultivar	Nipponbare	HP997	36.4	1.4	1,547	1,417
ERX003181	Oryza sativa	Japonica	Temperate japonica	Cultivar	Nongken-58	HP996	48.4	1.8	15,177	14,542
ERX003161	Oryza sativa	Indica	Indica	Cultivar	Guangluai-4	HP999	43.2	1.8	322,027	317,343
ERX002902	Oryza sativa	Japonica	Temperate japonica	Landrace	Dadongnuo	HP1	28.5	1.4	20,115	19,834

http://tga.nig.ac.jp/dnapod/

DDBJ Read Annotation Pipeline

基礎処理部



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

高次処理部　DNApod ワークフロー

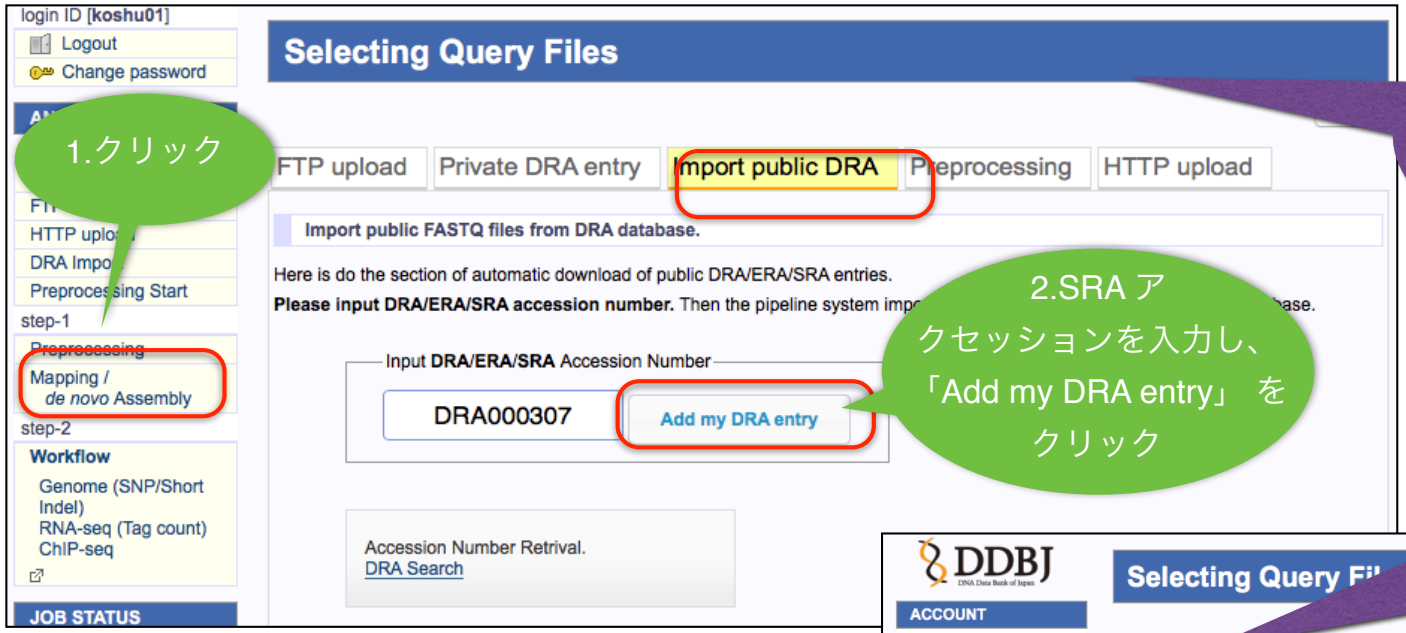
https://sites.google.com/a/g.nig.ac.jp/p-galaxy-help/home

追加資料

DDBJ Pipeline 基礎処理部

Quality Value (QV) フィルタ

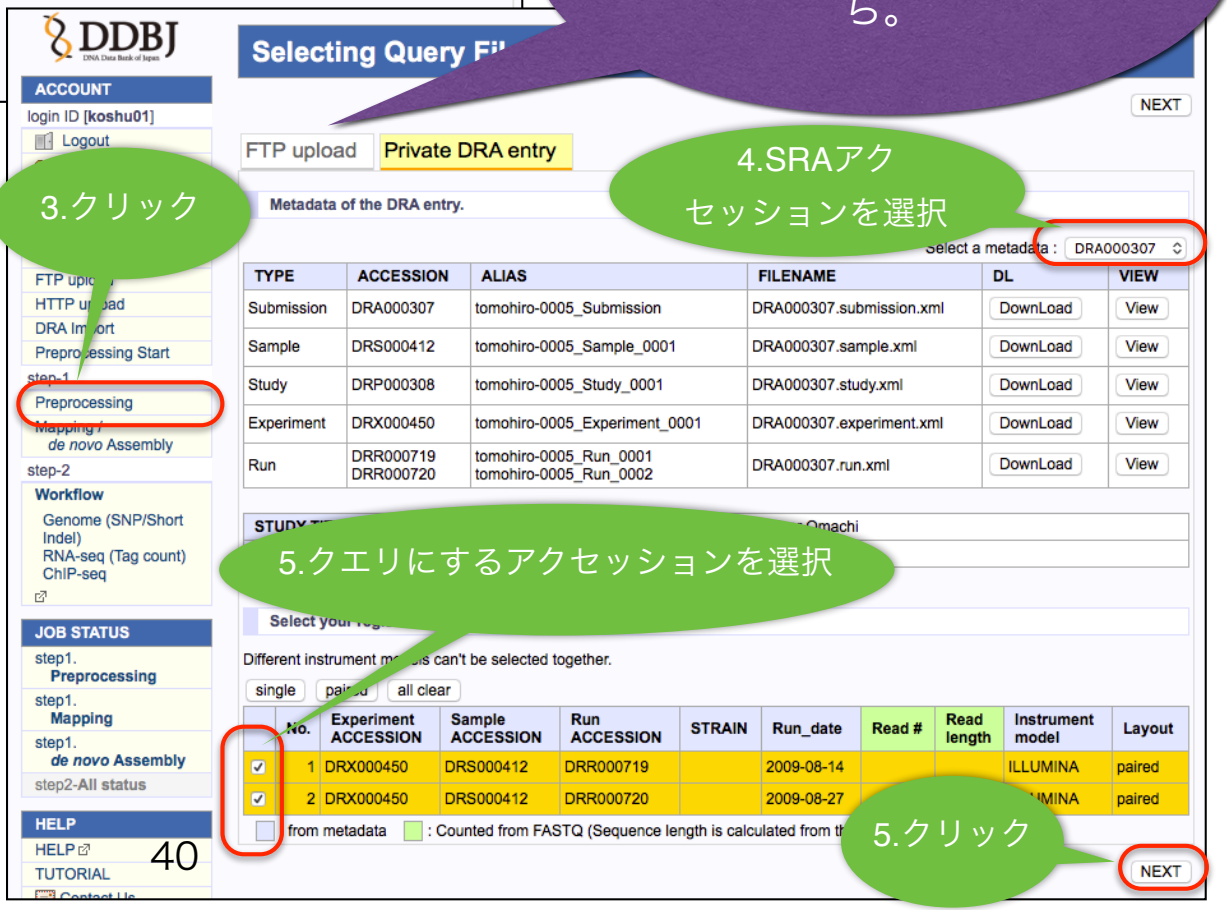
クエリの選択



SRA データを使用する場合は、データインポート処理が必要

2.SRA アクセッションを入力し、「Add my DRA entry」 をクリック

ローカルPC からFASTQ ファイルをアップロードする場合は、こちら。



パラメータの指定

Set Parameters for Preprocessing

BACK NEXT

Your selected queries

Run ACCESSION	Read length	Quality Score	Read Layout
DRR000719 -><-	bp		paired
DRR000720 -><-	bp		paired

Steps of preprocessing workflow

Step1: Set the encoding type of the quality values for sequence.

- (•) Phred+33 ( ) Phred+64

If you don't know it, please see '2.2 Encoding' of this site.

Step2: BASE TRIMMING with low quality from 5'end and 3'end of each read.

Bases with low quality (QV <= THRESHOLD) are trimmed from 5'end and 3'end of each read. The first and last bases of the trimmed read indicate high quality (QV > THRESHOLD).
If read length after base trimming is too short (length <= 24 bp), the read is removed. Thus the minimum read length will be 25bp.

- QV THRESHOLD : 19

Step3: READ REMOVING to discard trimmed reads including low quality bases with high percentage.

Trimmed reads with high percentage (>= Low quality bases# / Total bases#) of the low quality bases (QV <= THRESHOLD) are discarded.

- QV THRESHOLD : 14
- Percentage THRESHOLD : 30

Step 4: In the case of paired-end read, the pair is discarded when one read of the pair is removed at 'Step2' or 'Step3'.

BACK NEXT

パラメータの指定

Set Parameters for Preprocessing

BACK NEXT

Your selected queries

Run ACCESSION	Read length	Quality Score	Read Layout
DRR000719 -><-	bp		paired
DRR000720 -><-	bp		paired

Steps of preprocessing workflow

Step1: Set the encoding type of the quality values for sequence.

- (•) Phred+33 ( ) Phred+64

If you don't know it, please see '2.2 Encoding' of this site.

Step2: BASE TRIMMING with low quality from 5'end and 3'end of each read.

Bases with low quality (QV <= THRESHOLD) are trimmed from 5'end and 3'end of each read. The first and last bases of the trimmed read indicate high quality (QV > THRESHOLD).
If read length after base trimming is too short (length <= 24 bp), the read is removed. Thus the minimum read length will be 25bp.

- QV THRESHOLD : 19

Step3: READ REMOVING to discard trimmed reads including low quality bases with high percentage.

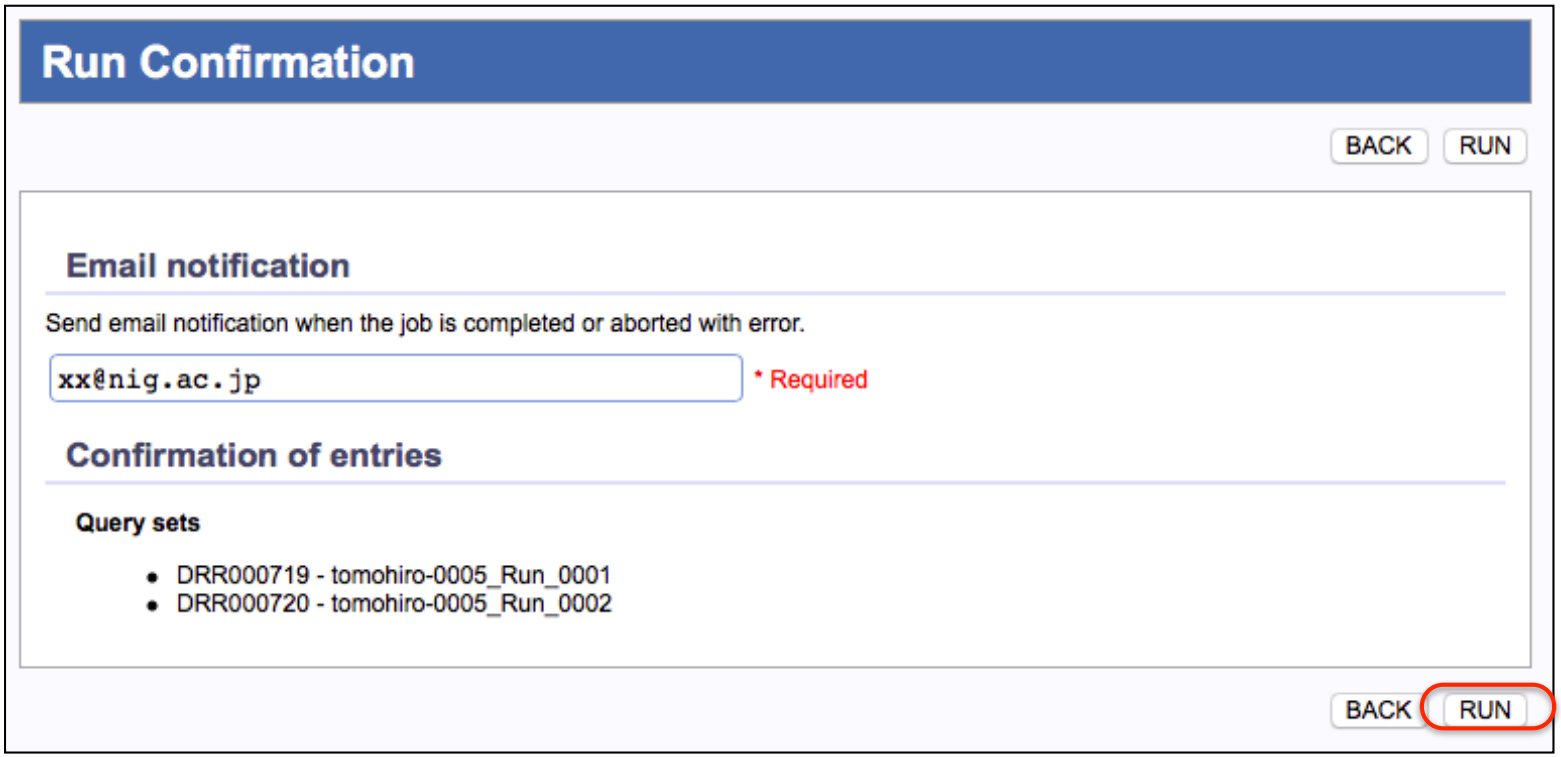
Trimmed reads with high percentage (>= Low quality bases# / Total bases#) of the low quality bases (QV <= THRESHOLD) are discarded.

- QV THRESHOLD : 14
- Percentage THRESHOLD : 30

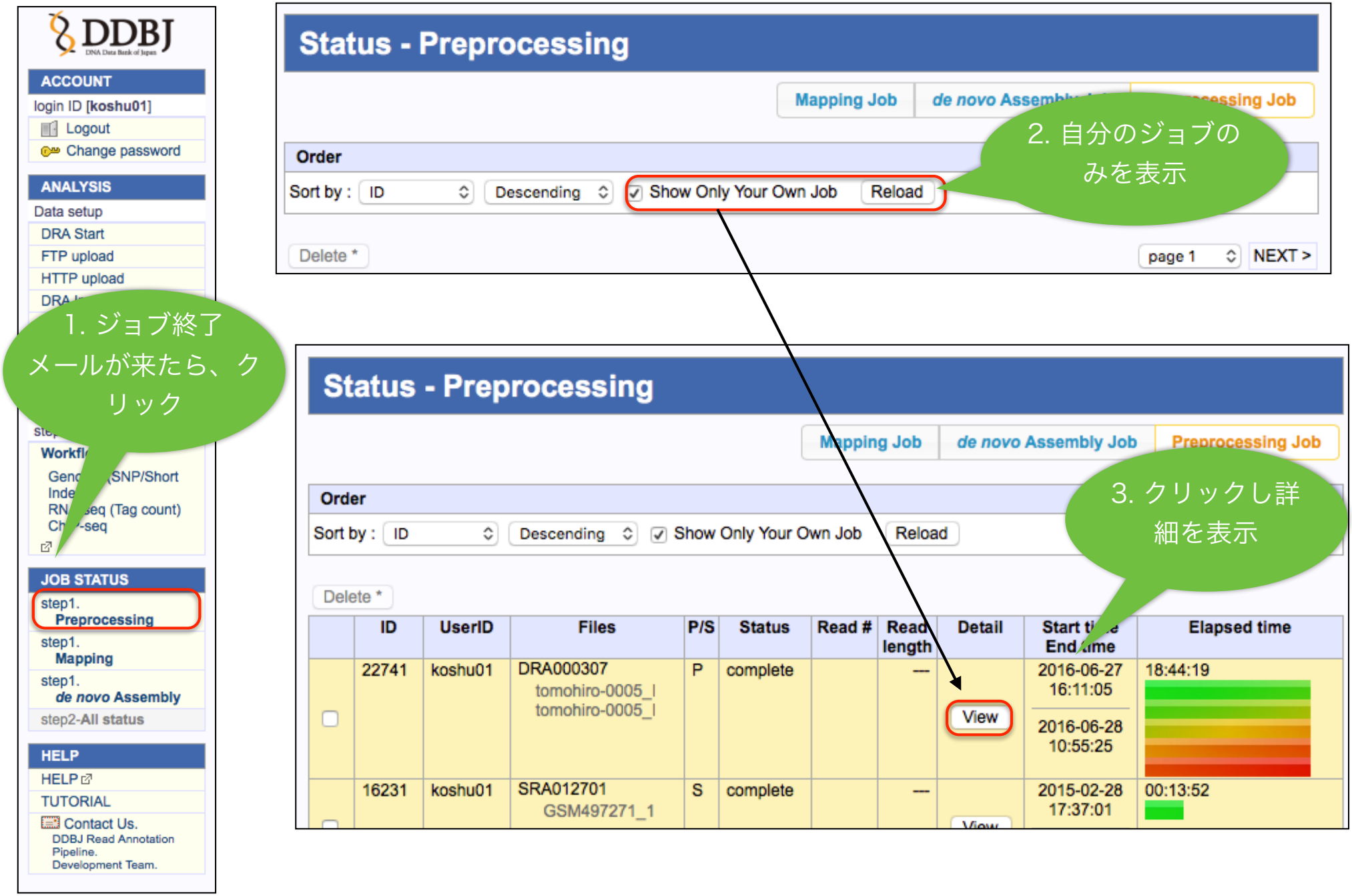
Step 4: In the case of paired-end read, the pair is discarded when one read of the pair is removed at 'Step2' or 'Step3'.

BACK NEXT

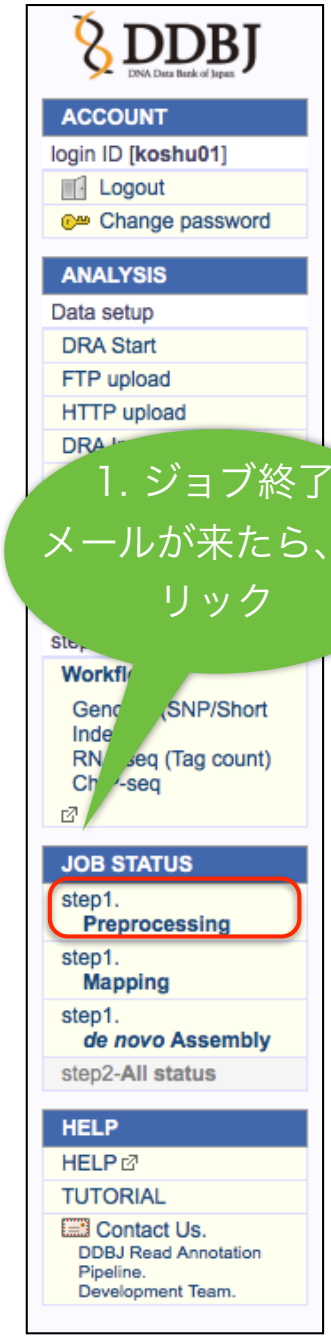
実行


Run Confirmation
BACK
RUN
Email notification
Send email notification when the job is completed or aborted with error.
xx@nig.ac.jp
* Required
Confirmation of entries
Query sets
DRR000719 - tomohiro-0005_Run_0001
DRR000720 - tomohiro-0005_Run_0002
BACK
RUN

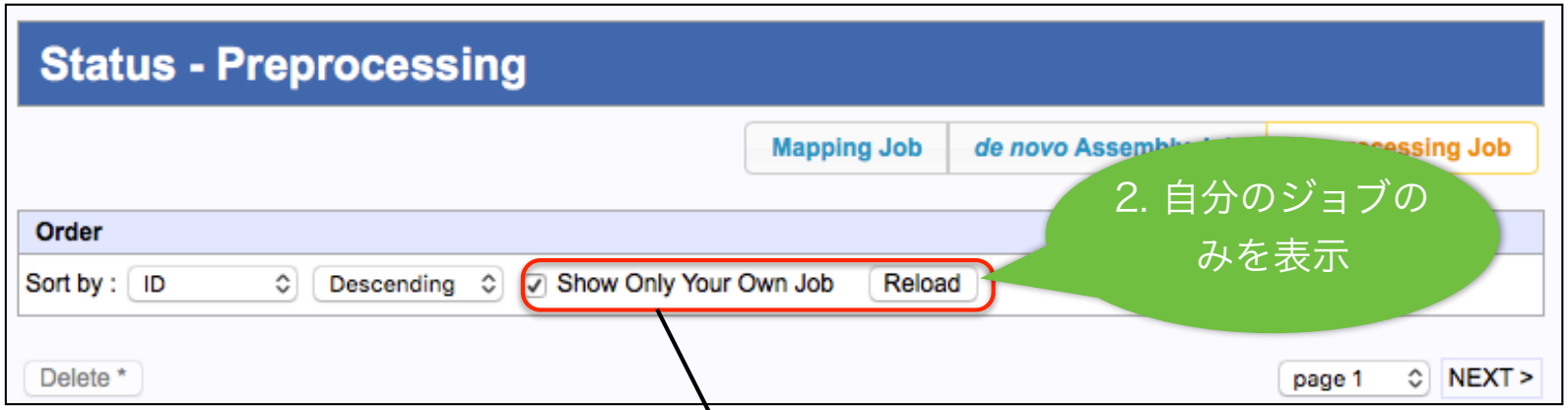
実行結果の確認


DDBJ
ACCOUNT
login ID [koshu01]
Logout
Change password
ANALYSIS
Data setup
DRA Start
FTP upload
HTTP upload
JOB STATUS
step1. Preprocessing
step1. Mapping
step1. de novo Assembly
step2-All status
HELP
HELP
TUTORIAL
Contact Us.
DDBJ Read Annotation Pipeline. Development Team.
1. ジョブ終了メールが来たら、クリック
Status - Preprocessing
Mapping Job
de novo Assembly Job
Preprocessing Job
Order
Sort by : ID
Descending
Show Only Your Own Job
Reload
2. 自分のジョブのみを表示
Delete *
page 1
NEXT >
Status - Preprocessing
Mapping Job
de novo Assembly Job
Preprocessing Job
Order
Sort by : ID
Descending
Show Only Your Own Job
Reload
Delete *
ID
UserID
Files
P/S
Status
Read #
Read length
Detail
Start time
End time
Elapsed time
22741
koshu01
DRA000307
tomohiro-0005_l
tomohiro-0005_l
P
complete
—
View
2016-06-27 16:11:05
2016-06-28 10:55:25
18:44:19
16231
koshu01
SRA012701
GSM497271_1
S
complete
—
2015-02-28 17:37:01
00:13:52
3. クリックし詳細を表示

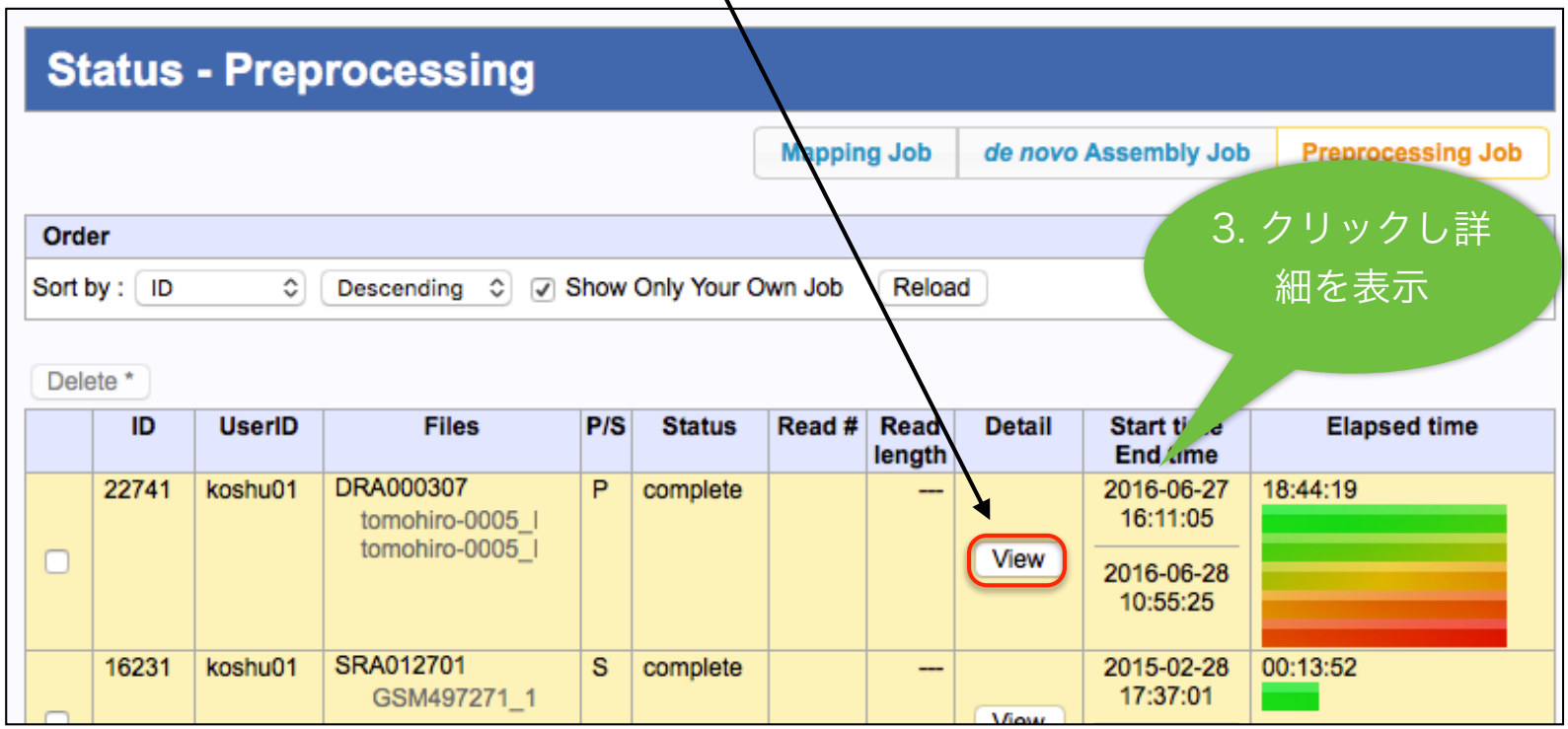
実行結果の確認



1. ジョブ終了メールが来たら、クリック



2. 自分のジョブのみを表示



3. クリックし詳細を表示

実行結果の確認 - 詳細 -

Detail view

BACK

Job info

ID
22741

Tool (Version)
(1.0)

RunAccession or Filename	Download	Read length	Alias
DRR000719	DRR000719.fastq.bz2 DRR000719_1.fastq.bz2 DRR000719_2.fastq.bz2	N.A. bp	tomohiro-0005_Run_0001
DRR000720	DRR000720.fastq.bz2 DRR000720_1.fastq.bz2 DRR000720_2.fastq.bz2	N.A. bp	tomohiro-0005_Run_0002

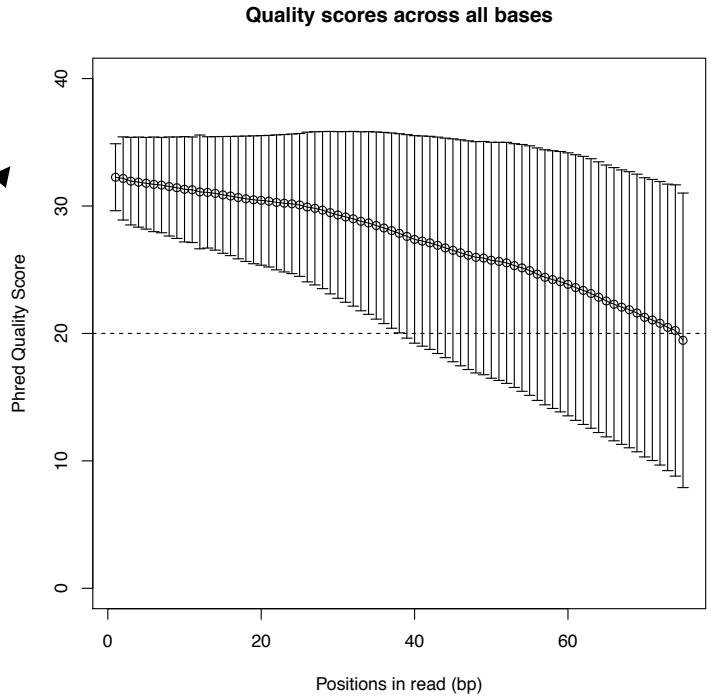
File	Fastq Download	QS Average (PDF)	QS Count (PDF)
DRR000719_1.fastq.bz2	download (4.2 GB)	download (8.7 KB)	download (5.2 KB)
DRR000719_2.fastq.bz2	download (4.2 GB)	download (8.6 KB)	download (5.2 KB)
DRR000720_1.fastq.bz2	download (32.6 GB)	download (8.6 KB)	download (5.2 KB)
DRR000720_2.fastq.bz2	download (32.4 GB)	download (8.7 KB)	download (5.1 KB)

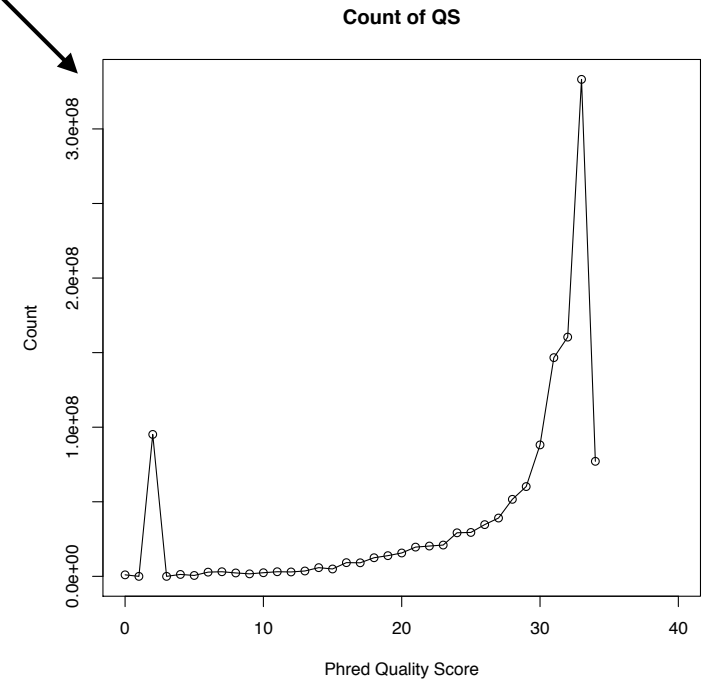
Time

Wait time	Start time	End time
0: 29:49	2016-06-27 16:11:05	2016-06-28 10:55:25

Command	Start time	End time	Log1	Log2	Result	MD5
perl avq_p.pl fqlist.txt qscore	2016-06-27 16:11:06	2016-06-28 07:06:04	View			
perl pdel_p3_t.pl fqlist.txt qscore19 24 1 14 30 33	2016-06-28 07:06:05	2016-06-28 09:46:31				
perl user_fastq_copy.pl preprocessing.xml koshu01	2016-06-28 09:46:33	2016-06-28 10:55:23	View			

BACK





Mapping の後工程は、SRA データを使用したときと同じです。