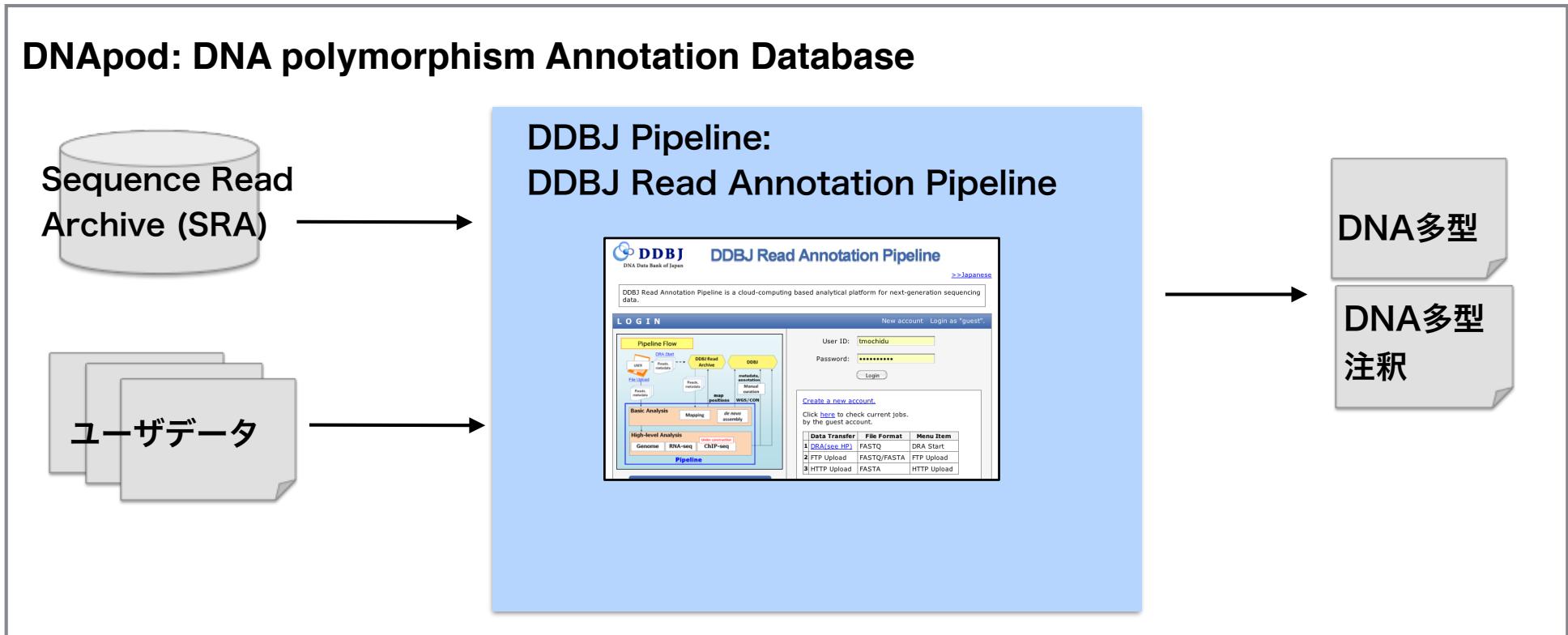


# DDBJ Pipelineを用いたDNA多型注釈解析の実習



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

PacBio

長鎖並列型



Sequel System

リード長 150bp の場合

1300–1500Gb

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。

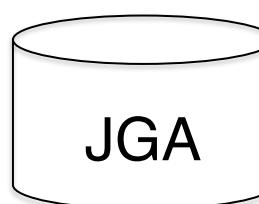
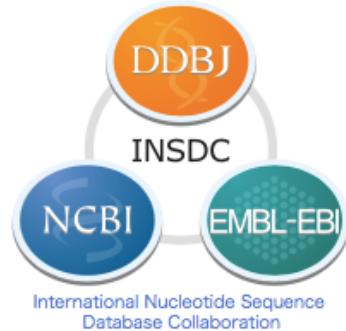
<https://github.com/AJACS-training/AJACS53/tree/master/nakazato>

# NGS 公共データベース

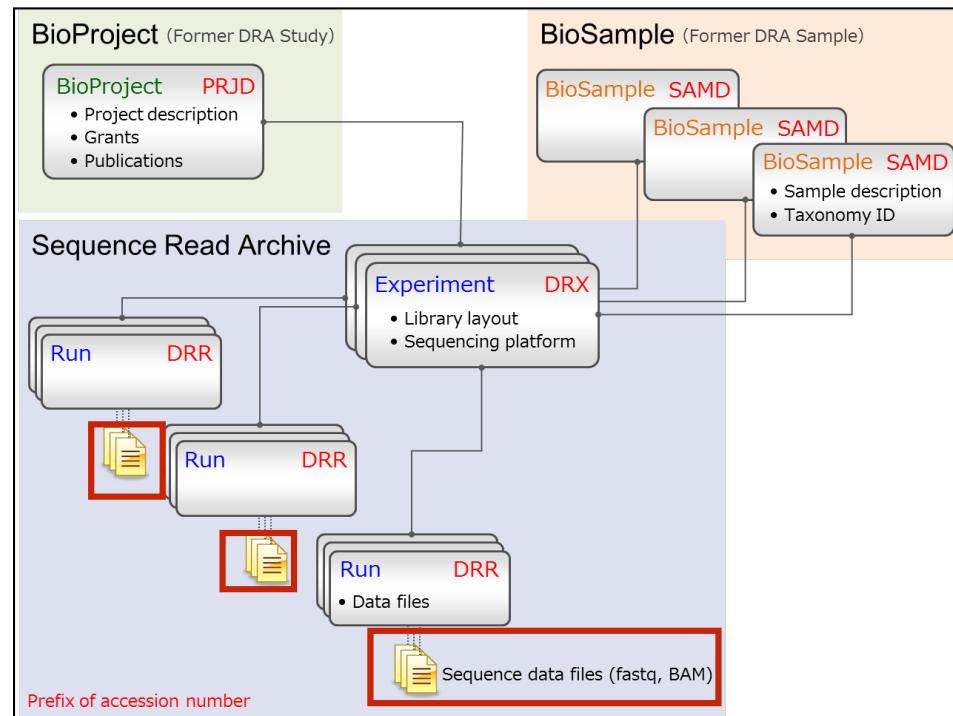


Sequence Read Archive

データ構成 メタデータ + 配列情報



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。

# FASTQ format

```
@SEQ_ID  
GATTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTGTTCAACTCACAGTTT  
+  
! ' ' *((((*++))%%++)(%%%).1***-+*'')**55CCF>>>>CCCCCCCC65
```

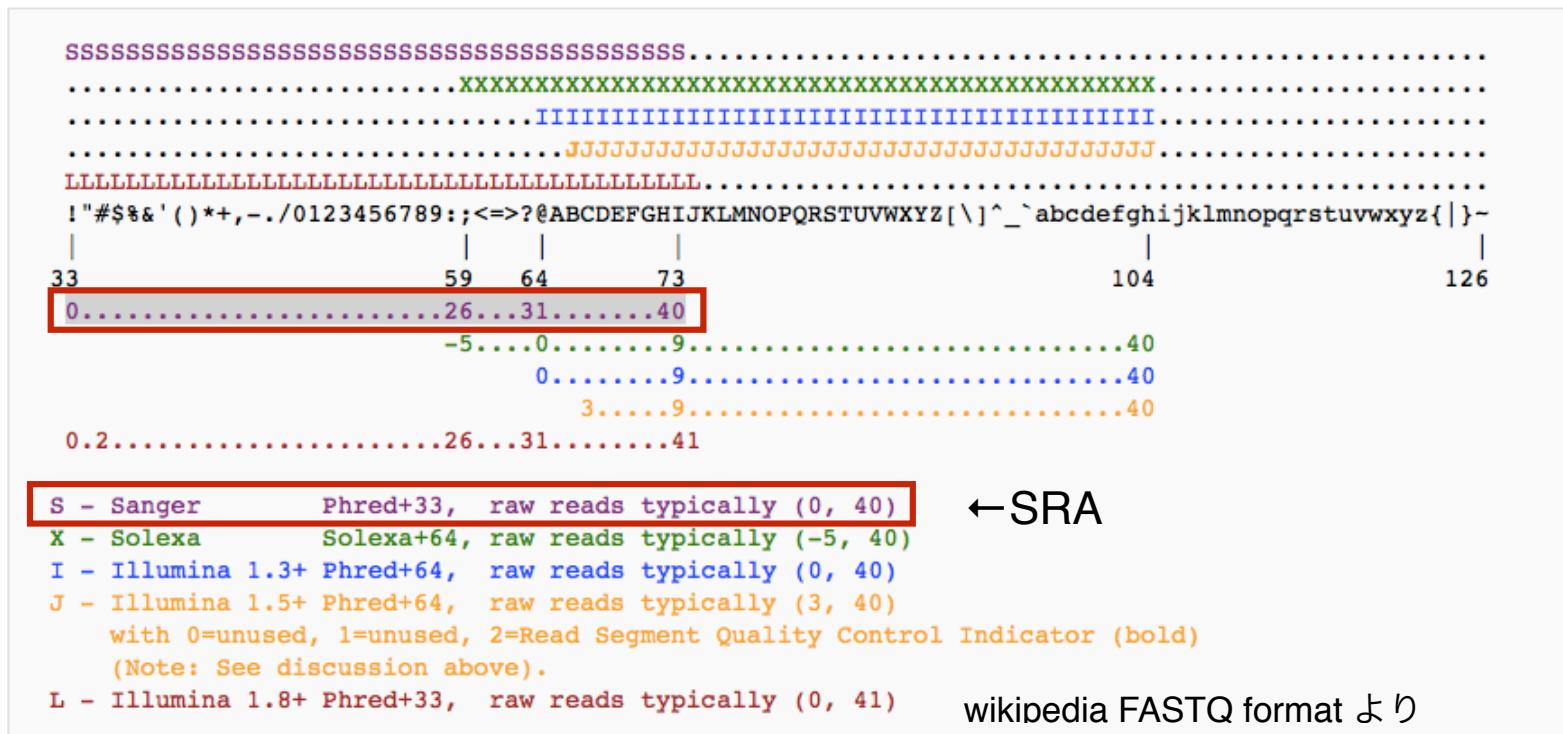
1行目: @ から始まり、配列の ID を記載する。

2行目: 配列

3行目: + を記載する。

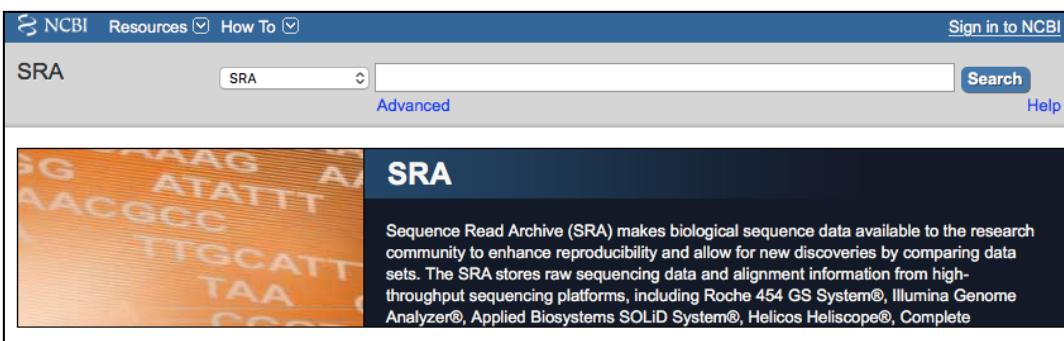
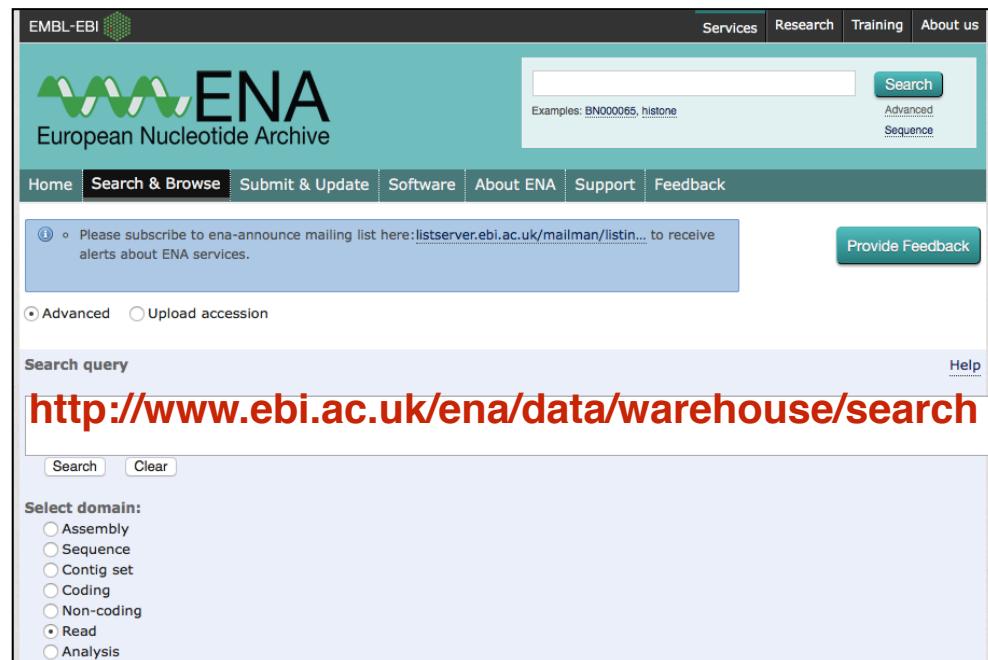
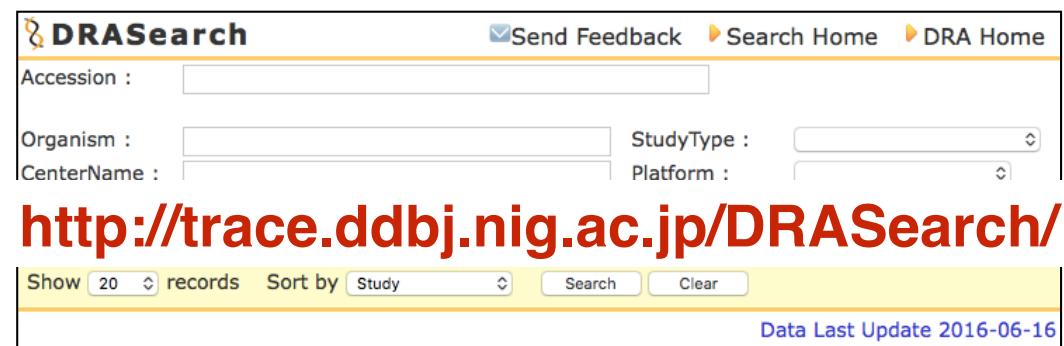
4行目: クオリティスコア

クオリティスコアをASCIIコードで表示



.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

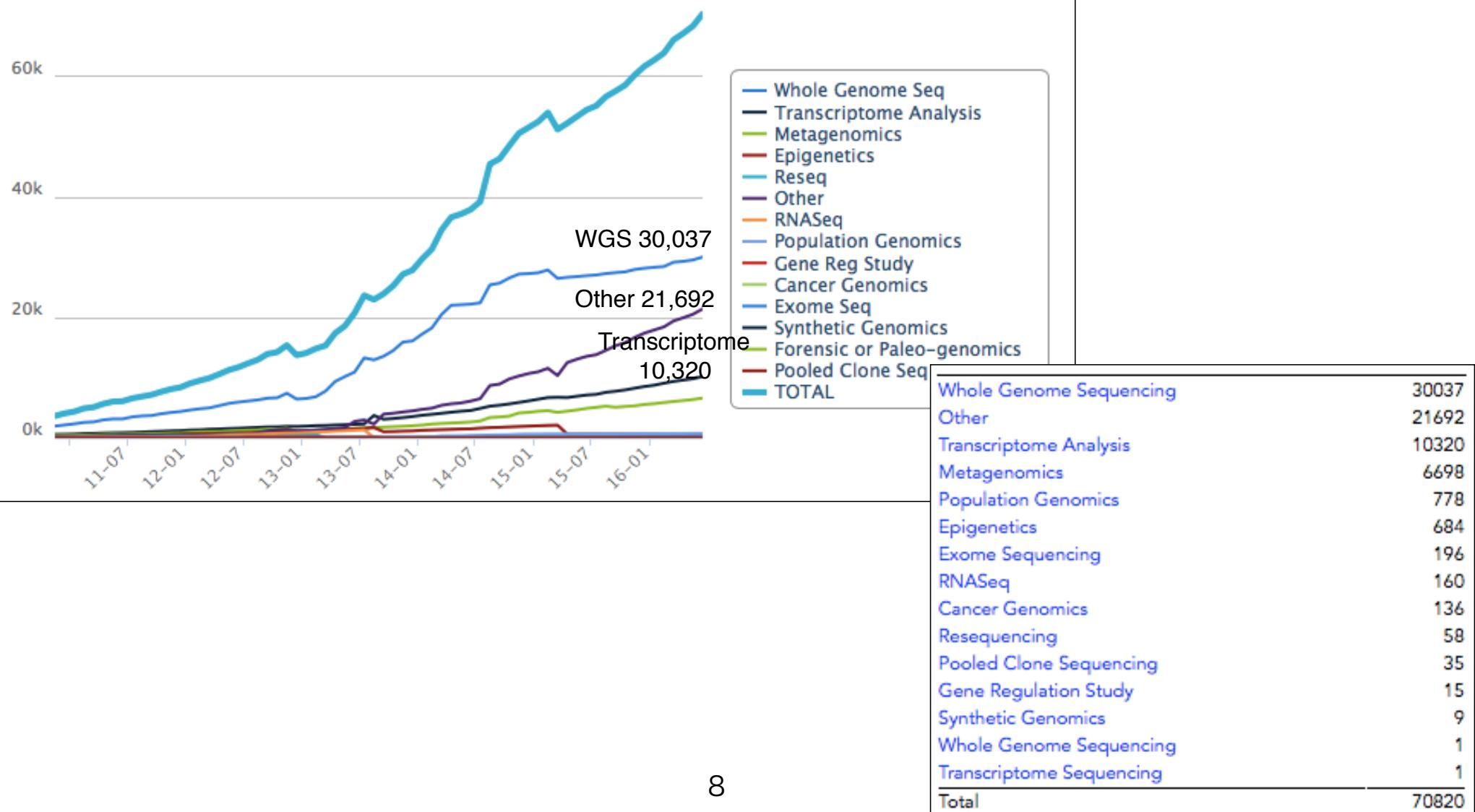
By study types

DBCLS SRA による集計値

Zoom

From  To

Total 70,820



# NGS 解析パイプライン 紹介

**DDBJ Read Annotation Pipeline**

DDBJ Read Annotation Pipeline is a cloud-computing based analytical platform for next-generation sequencing data.

**LOG IN**

New account Login as "guest".  
<https://p.ddbj.nig.ac.jp/>

Password:

Login

Create a new account.  
Click [here](#) to check current jobs.  
by the guest account.

Basic Analysis  
Mapping de novo assembly

High-level Analysis  
Genome RNA-seq ChIP-seq  
Pipeline

File Transfer Data Format Menu Item

1 DRA(see HP)	FASTQ	DRA Start
2 FTP Upload	FASTQ/FASTA	FTP Upload
3 HTTP Upload	FASTA	HTTP Upload

**Pitagora-Galaxy**  
Galaxy Japan Community + VM + Cloud

このサイトでは動作検証したツールとワークフローの一覧を記載しており、これらは全て Galaxy プロジェクトの Tool Shed からダウンロードすることができます。まだ Galaxy を開いてください。

**ピタゴラ装置でデータ解析**

方法 1. テストサイトを使う

このサイトでは動作検証したツールとワークフローの一覧を記載しており、これらは全て Galaxy プロジェクトの Tool Shed からダウンロードすることができます。まだ Galaxy を開いてください。

<http://www.pitagora-galaxy.org> ナイトへ

方法 2. 仮想環境をダウンロードする

このサイトでは Oracle VirtualBox を使ってあなたのパソコンで今すぐ無料で Galaxy を使う方法を紹介しています。

ツールとワークフロー ダウンロード

**MeGAP** メタゲノム解析パイプライン

English

MeGAP is a MetaGenome Annotation Pipeline.

Try MeGAP

<http://fs2.bio.titech.ac.jp/megap/>

ID:  (use [A-Za-z0-9-\_])

Email:

upload&calculate clear

Metagenome Sequencing FASTQ file:  ファイルを選択 ファイル未選択

File format:  fastq

ID:   
Email:

開発中 10/5 に正式オープン予定

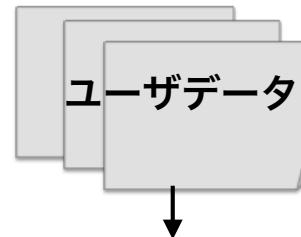
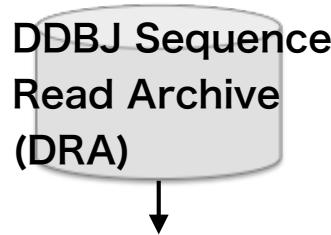
日本での Galaxy コミュニティ。  
研究者が作成した解析ワークフローを Pitagora-Galaxy から配布している。

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 Pipeline: DDBJ Read Annotation Pipeline

基礎処理部

The screenshot shows the DDBJ Read Annotation Pipeline web interface. On the left, there is a 'LOG IN' form with fields for 'User ID' (tmochidu) and 'Password'. On the right, there is a 'Create a new account' section and a 'Data Transfer' menu with options like 'DRA(see IUP)' and 'FTP Upload'. In the center, there is a 'Pipeline Flow' diagram illustrating the data processing workflow from 'DRA Start' through 'Basic Analysis' and 'High-level Analysis' stages to 'WGS/CON'.

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

The screenshot shows the Galaxy / DDBJ interface. It displays a workflow step titled 'WWFSMD?' which is part of a larger pipeline. The interface includes various tool icons and a history panel.

DNA多型注釈  
(DNApod)

Contig, Scaffold注釈  
(開発中)

# Pitagora-Galaxy



<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 – vep
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 を利用したバリエントの検出 (GenomeAnalysisTK, 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

ツール  
ツール  
search tools  
Get Data  
Lift-Over  
Text Manipulation  
Filter and...  
Join, Subt...  
Convert F...  
Extract F...  
Fetch Sequences  
Fetch Alignments  
Statistics  
Graph/Display Data  
ADDITIONAL TOOLS  
Pitagora Tools  
NGS Tools

ピタゴラ・ギャラクシーのテストサイトへようこそ！  
メニューから User > Register と進みユーザーを登録することができます。  
使用できるディスクの容量は以下の通りです：  
・登録済ユーザー： 10GB  
・未登録ユーザー： 200MB

ヒストリー  
search datasets  
Unnamed history  
0 bytes  
ヒストリーは空です。 You can load data from

Galaxy is an open, web-based platform for data intensive biomedical research. The Galaxy team is a part of BX at Penn State, and the Biology and Mathematics and Computer Science departments at Emory University. The Galaxy Project is supported in part by NHGRI, NSF, The Huck Institutes of the Life Sciences, The Institute for CyberScience at Penn State, and Emory University.

<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

#### 仮想環境の入手

- 最新版 Version 0.2.4 : [Windows](#) / [Mac](#)

#### 準備

- VirtualBox をインストールしてください。  
この仮想環境は VirtualBox のバージョン 4.3 を用いています。

#### 起動手順

- 上記のリンクからOVAイメージをダウンロードします。
- VirtualBox マネージャーで、ファイル > 仮想アプライアンスのインポート、を開きます。
- ダウンロードしたOVAイメージを選択し、設定を変更せずにインポートします。
- VirtualBox マネージャーで、インポートされたVMを右クリックし「起動」します。
- VMの起動後、ブラウザで <http://192.168.56.10:8080/> にアクセスします。

ova

### (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

GACCGAGCTACGCCCTCCTGTGGA

Reads  
(BWA)

GAGCTACGCCACCTG  
GAGCTACGCCACCTG  
GAGCTACGCCACCTG  
AGCTACGCCACCTGT  
GCTACGCCACCTGTG  
GCTACGCCACCTGTG

SNP  
(samtools mpileup)

Reference

gene

exon

intron

バクテリア～動  
物、植物を網羅して  
行く予定

DNApod

12

DDBJ Read Annotation Pipeline  
基礎処理部

DDBJ Read Annotation Pipeline

DDBJ Read Annotation Pipeline is a cloud-computing based analytical platform for next-generation sequencing data.

**LOGIN**

Pipeline Flow

User ID: tmochidu  
Password:   
Login

Create a new account.  
Click here to check current jobs.  
by the guest account.

Data Transfer File Format Menu Item  
1 DRA(see HP) FASTQ DRA Start  
2 FTP Upload FASTQ/FASTA FTP Upload  
3 HTTP Upload FASTA HTTP Upload

高次処理部 (Pitagora-Galaxy)  
DNApod workflow

Galaxy / DDBJ Analyze Data Workflow Shared Data P-galaxy Manual Admin Help User

Hello world! It's running...  
To customize this page edit atactic/welcome.html

WWFSMD? grow needy appendages...

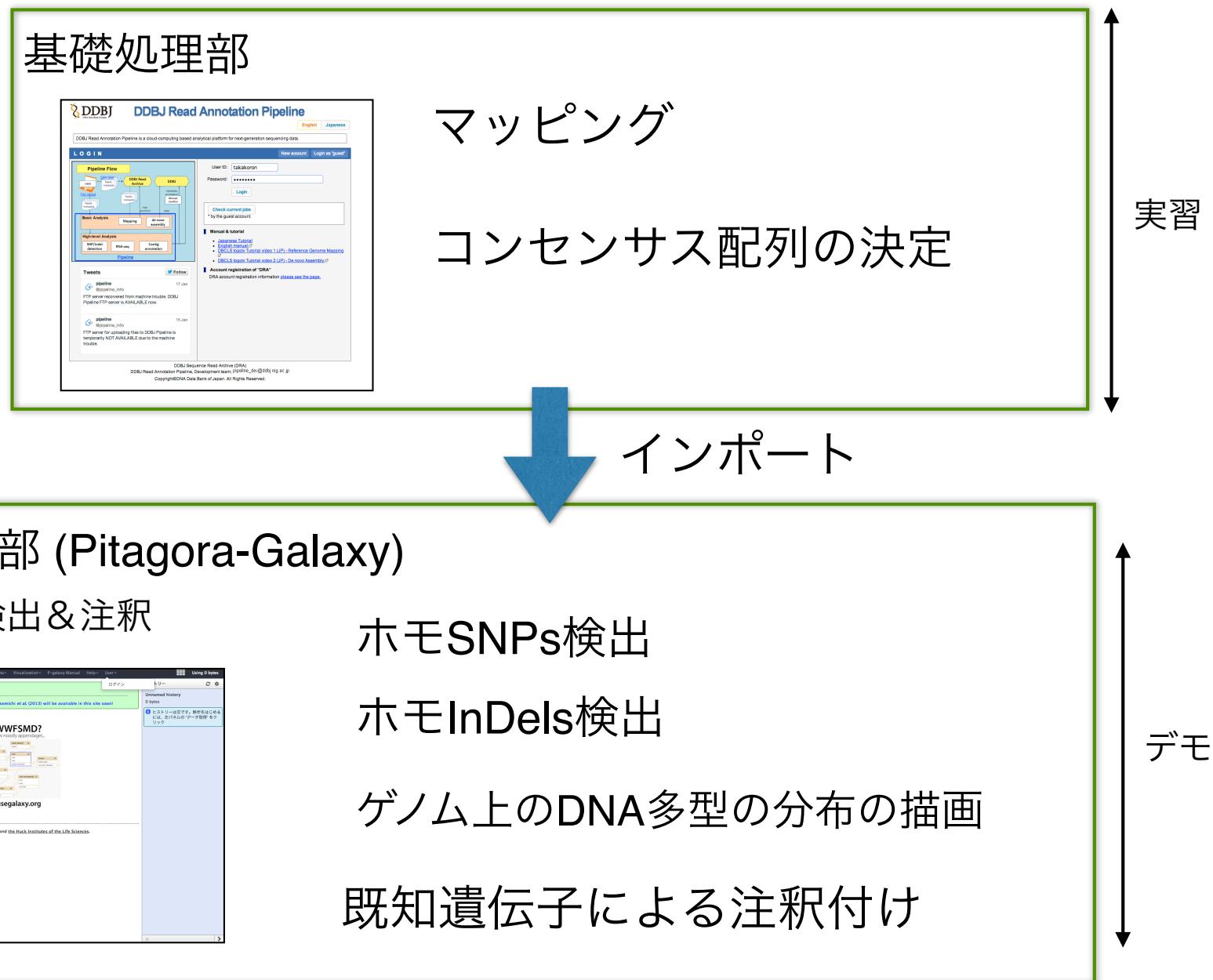
This project is supported in part by NED, NHGRI, and the Huck Institutes of the Life Sciences.

現在、イネ679系統、  
トウモロコシ404系統、  
ソルガム66系統

DNApod									
HOME Database DDBJ pipeline HELP									
Summary									
DRA accession									
species	subspecies	address	site	strain	M	coverage	depth	DNA - home STR	
EZ000001	Oryza sativa	Japanes	Cultivar	Koshihikari	96.7	213	107,162	91,130	
EZ000005	Oryza sativa	Japanes	Cultivar	Yamada	96.0	53.2	182,442	142,604	
EZ000006	Oryza sativa	Japanes	Cultivar	Yamada	94.3	53.2	182,442	142,604	
EZ000022	Oryza sativa	India	Cultivar	Gongjue-4	87.6	15.2	2,624,581	2,365,036	
EZ000070	Oryza sativa	India	Cultivar	IR8	90.8	55.0	2,624,780	2,365,544	
EZ000095	Oryza sativa	India	Cultivar	N22	92.0	63.0	3,042,401	2,497,946	
EZ000100	Oryza sativa	India	Cultivar	IR64	93.0	63.0	3,042,401	2,497,946	
EZ000270	Oryza sativa	Japanes	Cultivar	Kitaake	48.4	1.7	21,161	20,615	
EZ000271	Oryza sativa	Japanes	Cultivar	Nipponbare	36.4	1.4	1,547	1,417	
EZ000272	Oryza sativa	Japanes	Cultivar	Yamada	94.1	53.2	182,442	142,604	
EZ000273	Oryza sativa	India	Cultivar	Gongjue-4	HP900	41.2	1.0	322,027	317,361
EZ000292	Oryza sativa	Temperate japonica	Cultivar	Delonghus	HP1	28.5	1.4	25,115	18,024

# DNApod ワークフロー 実習

# DNApod ワークフロー 実習の流れ



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.** ◎次世代シーケンサ(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](http://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

Google™ カスタム検索 Search English

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、シーケンサなどで検索

**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
- RefSeq 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リスト表示するツール
- GGGenome 高速ゲノム配列検索ツール
- GRNA 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マネージメントツール
- \*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)

**次世代 Sequence 解析**

**DDBJ Read Annotation Pipeline**

次世代シーケ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）

**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

DDBJのHPからのリンク

WABI (Web Application for Bioinformatics Integration)

WABI は、DDBJの検索サービスを統合するための Web アプリケーションです。DDBJ では現在、次のサービスの Web API を実装しています。

- WABI BLAST
- getentry WebAPI
- ARSA WebAPI

**アカウントの取得**

DDBJ Read Annotation Pipeline is a cloud-computing based analytical platform for next-generation sequencing data.

**L O G I N**

User ID: koshu01  
Password: nigkoshu01

New account    Login as "guest"

User ID: koshu01  
Password: nigkoshu01

Login

Check current jobs  
\* by the guest account.

講習用ID : koshu01  
パス : nigkoshu01

Tweets

pipeline @pipeline\_info 17 Jan  
FTP server recovered from machine trouble. DDBJ Pipeline FTP server is AVAILABLE now.

**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. ([Supercomputer 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 characters.

\* 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](#)

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

DRA000307

**Selecting Query Files**

FTP upload **Private DRA entry** Import public DRA Preprocessing

Metadata of the DRA entry.

TYPE	ACCESSION	ALIAS	FILENAME	DL	VIEW
Submission	DRA000307	tomohiro-0005_Submission	DRA000307.submission.xml	DownLoad	View
Sample	DRS000412	tomohiro-0005_Sample_0001	DRA000307.sample.xml	DownLoad	View
Study	DRP000308	tomohiro-0005_Study_0001	DRA000307.study.xml	DownLoad	View
Experiment	DRX000450	tomohiro-0005_Experiment_0001	DRA000307.experiment.xml	DownLoad	View
Run	DRR000719 DRR000720	tomohiro-0005_Run_0001 tomohiro-0005_Run_0002	DRA000307.run.xml	DownLoad	View

STUDY TITLE: Whole genome sequencing of Japonica rice cultivar Omachi  
STUDY TYPE: Whole Genome Sequencing

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
<input checked="" type="checkbox"/>	1 DRX000450	DRS000412	DRR000719		2009-08-14			ILLUMINA	paired
<input checked="" type="checkbox"/>	2 DRX000450	DRS000412	DRR000720		2009-08-27			ILLUMINA	paired

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

DELETE **NEXT**

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

**Selecting Query Files**

FTP upload **Import public DRA** Preprocessing HTTP upload

Import public FASTQ files from DRA database.

Here is the section of automatic download of public DRA/ERA/SRA entries.  
Please input DRA/ERA/SRA accession number. Then the pipeline system import metadata and FASTQ files from DRA database.

Input DRA/ERA/SRA Accession Number  
**DRA000307** Add my DRA entry

Accession Number Retrieval.  
DRA Search

3.解析に使用するデータを選択

4.次へ

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)

## 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	Rate
<input type="checkbox"/>	BLAT		34	✓					✓
<input checked="" type="checkbox"/>	bwa		0.6.1	✓		✓	✓	✓	✓
<input type="checkbox"/>	Bowtie		0.12.7	✓	✓	✓	✓	✓	✓

1. Reference Genome Mappingを選択

2. ツールを選択

	Tool	Help	Version	Base space	Color space	Paired-end
<input type="checkbox"/>	Bowtie2		2.0.0	✓	✓	✓
<input type="checkbox"/>	TopHat2		2.0.9	✓		✓

Mapping / de novo Assemblyツール、各種選択できます。

Tool	Help	Version	Base space	Color space	Paired-end	MSS(WGS)	Comment
<input type="checkbox"/>	SOAPdenovo		1.05	✓		✓	
<input type="checkbox"/>	ABYSS		1.3.2	✓		✓	Maximum K-mer value is 64.
<input type="checkbox"/>	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.
<input type="checkbox"/>	Trinity		2013-02-25	✓		✓	RNA-Seq De novo Assembly
<input type="checkbox"/>	Platanus		1.2.2	✓		✓	
<input type="checkbox"/>	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
<input checked="" type="radio"/>	BLAT	Single-end analysis only

3. 次へ

BACK  NEXT

**Generating Query Sets from Query Read Files**

Paired-end analysis  
Layout of paired sequence. 5'-3' 3'-5'

5'	Linker(1)	Target	Linker(2)	Linker(3)	Target	Linker(4)	5'
----	-----------	--------	-----------	-----------	--------	-----------	----

**QUERY SET**

1. クエリセット単位でデータを選択

	Run ACCESSION	Read length	Quality Score
<input checked="" type="checkbox"/>	DRR000719 -><-	bp	
<input checked="" type="checkbox"/>	DRR000720 -><-	bp	

**Set as Pair-End**

2. クリック

**Generating Query Sets from Query Read Files**

Paired-end analysis  
Layout of paired sequence. 5'-3' 3'-5'

5'	Linker(1)	Target	Linker(2)	Linker(3)	Target	Linker(4)	5'
----	-----------	--------	-----------	-----------	--------	-----------	----

**QUERY SET**  
Query set1

PairedOrientation	RunAccession	RunAlias	RowLength	QualityScore1	QualityS
paired	DRR000719	tomohiro-0005_Run_0001			
paired	DRR000720	tomohiro-0005_Run_0002			

**RESET** **BACK** **NEXT**

3. 次へ

**Specifying Database of Reference Genome**

1. Major genome setsを選択

2. Organismを選択

3. Genome setsを選択

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す

4. 次へ

**Major genome sets**

Organisms: Oryza sativa japonica

Genome sets: Os-Nipponbare-Reference-IRGSP-1.0

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

User original sets

Download or upload reference

RESET BACK NEXT

**Setting for Reference Genome Mapping**

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**

**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 -bvbg -> out.var.raw.bcf  
bcftools view out.var.raw.bcf | vcftools pl varFilter -d10000 > out.var.fltr.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**

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

2. 次へ

**Run Confirmation**

Destination of mail  
When the request is completed, the system sends an email to this address.  
XXXXXX@nig.ac.jp \* Required  
Result files will be deleted 60 days after submission.

Reference Genome Map [bwa]

Query sets  
Query set1

PairedOrientation	RunAccession	RunAlias	RowLength	QualityScore
paired	DRR000719	tomohiro-0005_Run_0001		
paired	DRR000720	tomohiro-0005_Run_0002		

genome sets  
Os-Nipponbare-Reference-IRGSP-1.0  
• all.fa

Command Options  
bwa

**Set optional parameters of the paired-end analysis**

Step1) Convert reference sequence  
bwa index -a is (for small-size reference) refgenome.fasta  
Options usage (click)

is  
This method is moderately fast, but does not work with database larger than 2GB.

bwtsw  
This method works with the whole human genome, but it does not work with database smaller than IS.

Step2) Map  
bwa aln -t 4 refgenome.fasta query1.fastq(.fa)  
bwa aln -t 4 refgenome.fasta query2.fastq(.fa)  
bwa sampe refgenome.fasta in1.sai in2.sai query2.fastq(.fasta) > out.sam

Step3) 'uniq': Remove multiple hits on the genome from out.sam.  
Please choose uniq mode.  
 Do not remove any read.  
 Retain pairs when both reads mapped uniquely or one of reads mapped uniquely, and Discard other pairs.  
 Retain pairs when both reads mapped uniquely, and Discard other pairs.  
 Retain uniquely mapped reads and discard multiply mapped reads.

Step4) Convert the read alignment to .BAM format  
samtools view -bS -o out.bam out.sam

Step5) Detect DNA polymorphism  
Please choose one of the following.  
 samtools pileup -c  
 samtools mpileup -u -C50 -BQ0 -d10000000 out.var.raw.bcf  
 bcftools view out.var.raw.bcf | vcftools.pl varFilter -D100

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 read

BACK RUN

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。

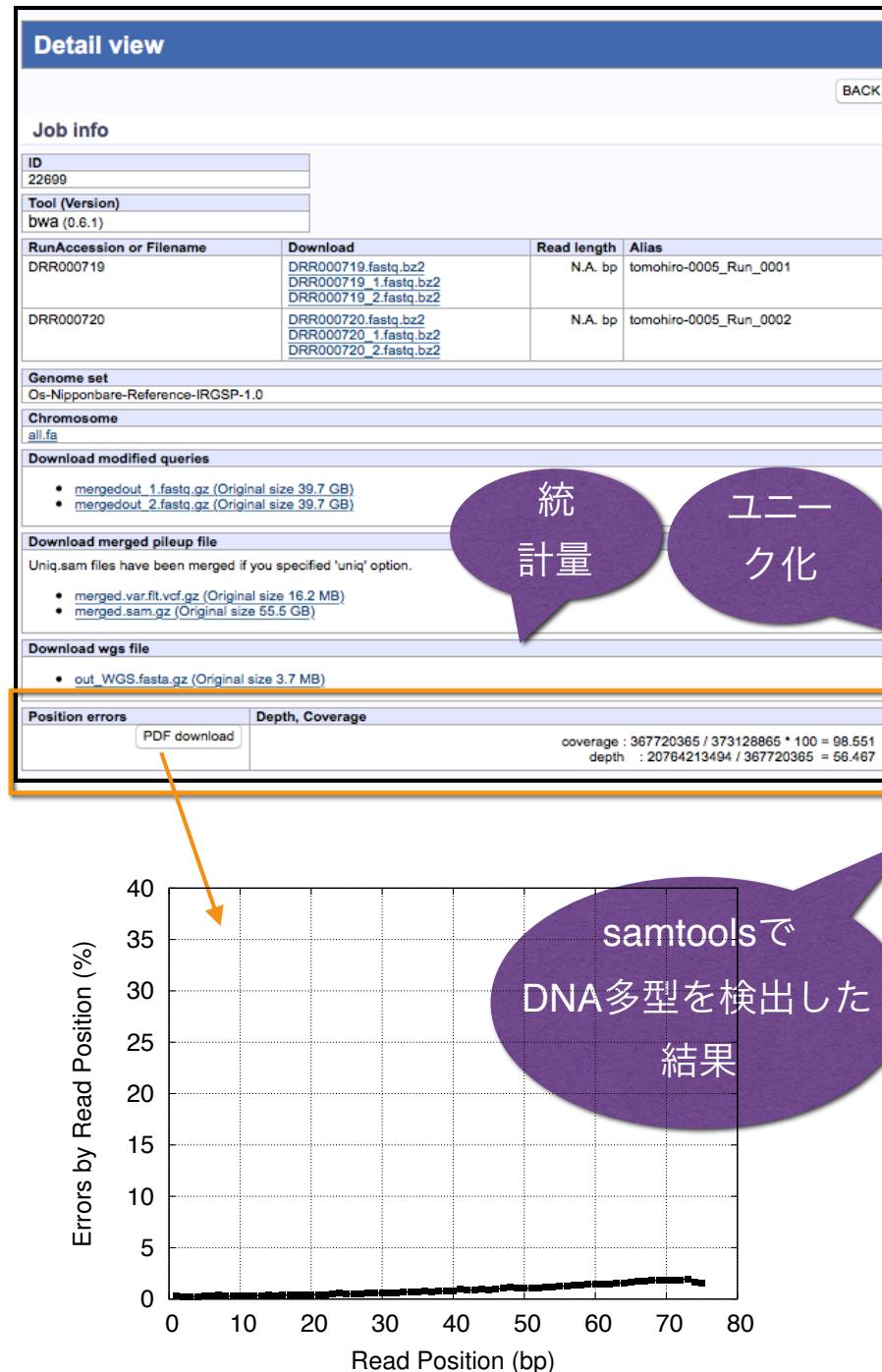
2. 問題なければ、実行  
講習ではRUNボタンを押さないで下さい。

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

	ID	UserID	Submission accession	P/S	Status	Tool	Read #	Read length	Genome size	Detail	Start time	End time	Elapsed time
<input type="checkbox"/>	22699	koshu01	DRA000307 tomohiro-0005_I tomohiro-0005_I	P	complete	bwa	148,147,887	—	379 M	<a href="#">View</a>	2016-06-23 18:50:08	82:03:39	
<input type="checkbox"/>	22003	koshu01	DRA000010	S	complete	bwa	38,818,503	—	388 M	<a href="#">View</a>	2016-06-27 04:53:48	03:32:15	



**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			<a href="#">View</a>	
BWA : Alignment bwa aln all.fa mergedout_1.fastq > 1.sai	2016-06-23 18:55:41	2016-06-23 21:21:57			<a href="#">View</a>	
BWA : Alignment bwa aln all.fa mergedout_2.fastq > 2.sai	2016-06-23 21:21:59	2016-06-23 22:56:55			<a href="#">View</a>	
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			<a href="#">View</a>	<a href="#">Download(26.7 GB)</a>
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			<a href="#">View</a>	<a href="#">Download(585.6 MB)</a>
Convert SAM to BAM samtools view -bS -o out.bam out.sam	2016-06-26 02:09:48	2016-06-26 03:30:03			<a href="#">View</a>	<a href="#">Download(27.8 GB)</a>
Sort BAM File samtools sort out.bam out2	2016-06-26 03:52:32	2016-06-26 05:37:56			<a href="#">View</a>	<a href="#">Download(20.4 GB)</a>
Create BAM Index File samtools index out2.bam	2016-06-26 05:52:40	2016-06-26 05:57:24			<a href="#">View</a>	<a href="#">Download(560.5 KB)</a>
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			<a href="#">View</a>	<a href="#">Download(17.0 GB)</a>
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			<a href="#">View</a>	<a href="#">Download(34.6 GB)</a>
Sort BAM File [ For Unique SAM ] samtools sort uniqout.bam out2	2016-06-26 10:46:49	2016-06-26 11:50:55			<a href="#">View</a>	<a href="#">Download(13.0 GB)</a>
Create BAM Index File [ For Unique SAM ] samtools index out2.bam	2016-06-26 11:59:49	2016-06-26 12:03:06			<a href="#">View</a>	<a href="#">Download(497.1 KB)</a>
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			<a href="#">View</a>	
Filter BCF and Convert to VCF File [ For Unique SAM ] bcftools view uniq.var.bcf   perl vcftools.pl varFilter -D10000 > out-unique.var.vcf	2016-06-26 13:44:34	2016-06-26 13:44:44			<a href="#">View</a>	<a href="#">Download(3.1 MB)</a>
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			<a href="#">View</a>	
Filter BCF and Convert to VCF File bcftools view non-uniq.var.bcf   perl vcftools.pl varFilter -D10000 > out.var.vcf	2016-06-26 18:59:13	2016-06-26 18:59:23			<a href="#">View</a>	<a href="#">Download(4.5 MB)</a>
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			<a href="#">View</a>	
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				

**bwaにてマッピング**

DDBJ pipelineのメニューをクリック

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.

ADMINISTRATION  
Fastq stats \*  
Job stats \*  
MailAddressList \*

## P-Galaxy HELP

home

### Manuals of 'P-Galaxy'

Beta Version 0.1.0

#### Overview

The DDBJ Read Annotation Pipeline annotates raw sequencing reads from next-generation sequencers (NGS) with high throughput. The proposed pipeline consists of two processes: basic analysis for genome mapping and de novo assembly, and high-level analysis (P-Galaxy) for each analysis purpose, which are genome resequence, *de novo* assembly. The High-level analysis is implemented in the Virtual Machine image, which is configured the galaxy platform by the [Pitagora-Galaxy](#).

The manuals of each workflow are below.  
[Annotating DNA polymorphisms \(DNApod\)](#)

## DNApod HELP

how to use >  
**DNApod workflow**

### DNApod workflow

Help page of [Basic analysis](#) + [High-level analysis](#)

#### Contents

- 1 DNApod workflow
- 2 Overview
- 3 High-level analysis (Virtual Machine Image) download
- 4 DNApod workflow Manual
- 5 Test data of DNApod workflow

### Overview

The DNApod workflow process detects DNA polymorphisms and annotates them with known gene annotations. The DNApod workflow process was implemented in 'DDBJ Read Annotation Pipeline (DDBJ pipeline)' (<http://p.ddbj.nig.ac.jp/>). The Analytical process is shown in Fig. 1.

Fig 1. Overview of DNApod workflow.

DNApod HELP

how to use >  
DNApod workflow

## DNApod workflow

Help page of [Basic analysis](#) + [High-level analysis](#)

Contents

- 1 DNApod workflow
- 2 Overview
- 3 High-level analysis (Virtual Machine Image) download
- 4 DNApod workflow Manual**
- 5 Test data of DNApod workflow

home > [howto\\_old](#) >

## Setting up a Virtual Machine

### Downloading the Virtual Machine Image

- Latest Version 1.0.0 : [Windows](#) / [Mac](#)

### Preparation

- Installing the [VirtualBox](#) .

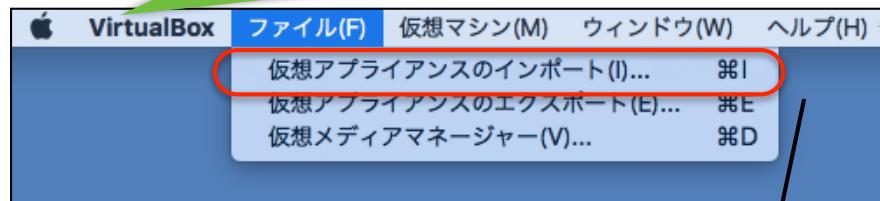
## DNApod workflow Manual

### outline

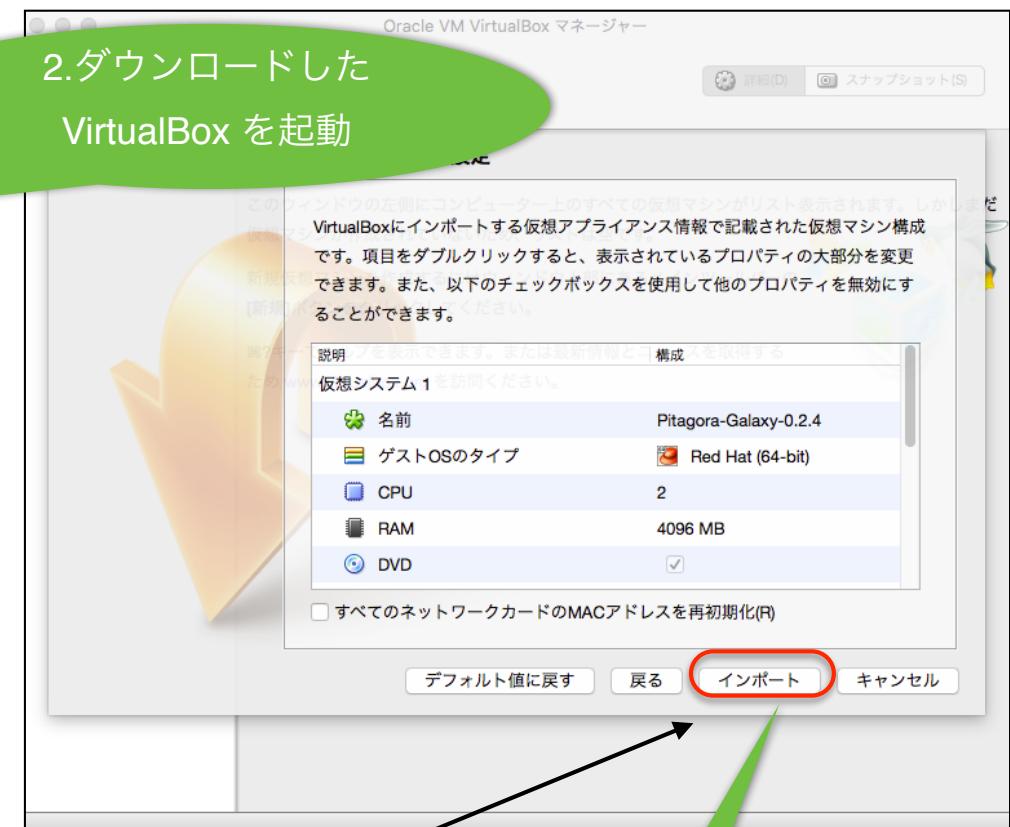
- [Step1\) Access to the DDBJ pipeline basic Analysis](#)
- [Step2\) Importing NGS data files for queries to the DDBJ pipeline](#)
- [Step3\) Preprocessing](#)
- [Step4\) Mapping](#)
- Step5) Setting up the Virtual Machine of DDBJ pipeline High-level analysis**
- [Step6\) Importing the SAMtools mpileup file \(VCF\)](#)
- [Step7\) Detecting SNPs/InDels from the SAMtools mpileup file \(VCF\)](#)
- [Step8\) Merge SNPs/InDels data files \(optional\)](#)
- [Step9\) Visualizing the distribution of DNA polymorphisms on the genome](#)
- [Step10\) Assign DNA polymorphisms with the known gene annotation using the snpEff tool](#)

環境に合わせてダウ  
ンロード。

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

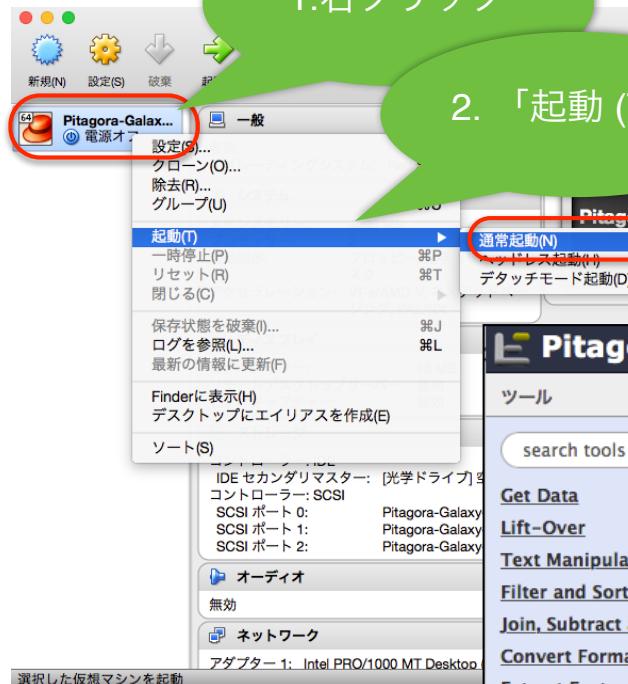


3.ダウンロードした  
Virtual Machine Image を起動



4.「続ける」をクリック

5.「インポート」を  
クリック



Pitagora-Galaxy DNApod ワークフローが起動！

Hello world! It's running...

To customize this page edit static/welcome.html

ピタゴラ・ギャラクシーへようこそ！

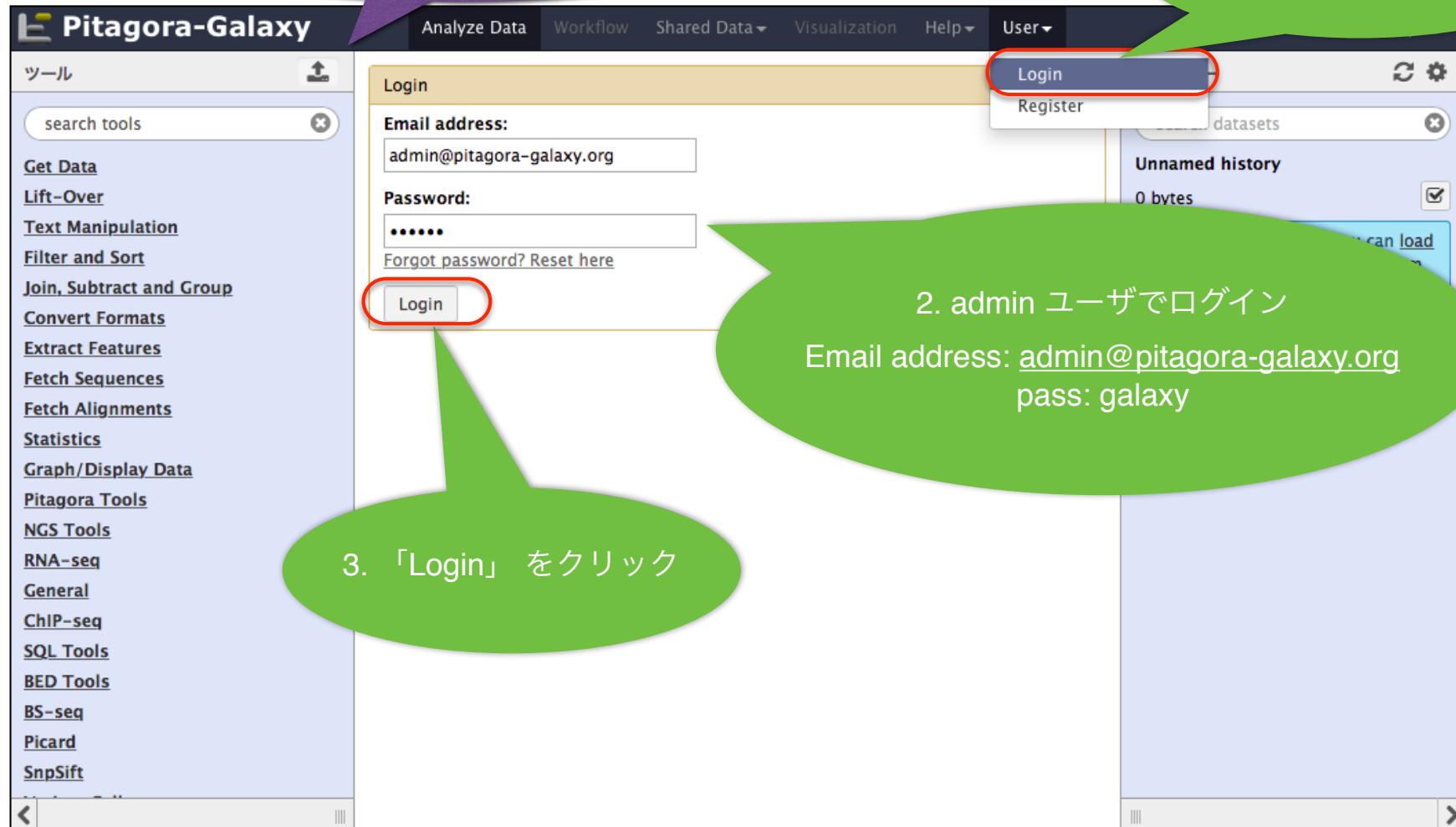
上のメニューから User > Login と進み、次のどちらかのユーザーでログインしてください。

- ・管理用ユーザー： admin@pitagora-galaxy.org (password: galaxy)
- ・通常のユーザー： test@pitagora-galaxy.org (password: galaxy)

Galaxy is an open, web-based platform for data intensive biomedical research. The Galaxy team is a part of BX at Penn State, and the Biology and Mathematics and Computer Science departments at Emory University. The Galaxy Project is supported in part by NHGRI, NSF, The Huck Institutes of the Life Sciences, The Institute for CyberScience at Penn State, and Emory University.

4. 「http://192.168.56.10:8080」にアクセス

今回の講習ではadmin ユーザ  
でデモをします。



Galaxy Tools search tools Get Data Send Data Lift-Over Text Manipulation Filter and Sort Join, Subtract and Group Convert Formats Extract Features Fetch Sequences Fetch Alignments Statistics Graph/Display Data DNApod import mpileupfile Import from DDBJ Pipeline Detect SNPs Detect InDels Merge SNPs / InDels data files Visualize distribution of DNA polymorphism

1. クリック

2. クリック

DDBJ Read Annotation Pipeline is a cloud-computing based analytical platform for next-generation sequencing data.

User ID: koshu01  
Password: .....  
Login

DDBJ Sequence Read Archive (DRA)  
DDBJ Read Annotation Pipeline, Development team;  
pipeline\_dev@ddbj.nig.ac.jp  
Copyright©DNA Data Bank of Japan. All Rights Reserved.

[Logout]

Import MPileup from basic analysis. By DDBJ Read Annotation Pipeline

JOBID	Submission Accession RunAlias	Tool	Pipeline Jobpage	Import to Galaxy
22699	DRA000307 tomohiro-0005_Run_0001 tomohiro-0005_Run_0002	bwa	<a href="#">ViewJob</a>	<a href="#">Import</a>

Analyze Data Workflow Shared Data Visualization Admin Help User Using 15.5 MB

##FORMAT!<ID=GQ,Number=1,Type=Integer>Description="Genotype Quality">  
##FORMAT=<ID=GL,Number=3,Type=Float>Description="Likelihoods for RR,RA,AA genotypes">  
##FORMAT=<ID=DP,Number=1,Type=Integer>Description="# high-quality bases">  
##FORMAT=<ID=SP,Number=1,Type=Integer>Description="Phred-scaled strand bias P-value">  
##FORMAT=<ID=PL,Number=G,Type=Integer>Description="List of Phred-scaled genotype likelihoods">

#CHROM	POS	ID	REF	ALT	...	...	...
chr01	1037	.	TAAA	TAAAA	...	...	...
chr01	125929	.	T	C	4.11	.	DP=3;V
chr01	214775	.	G	T	222	.	DP=33;I
chr01	357330	.	ACTCTCTCTCTCTCTCTC	ACTCTCTCTCTCTCTC	8.18	.	INDEL;D

1: import mpileupfile

3. 基礎  
処理部をログアウトした場合のみ基礎処理部のIDとパスワードを入力してログイン

4. インポートしたいデータの Importボタンをクリック

5. データがインポートされた

目玉マークをクリックするとファイルの中身を確認できます。

## ホモSNPsの検出

Galaxy Analyze Data Workflow Shared Data Visualization Admin Help User Using 15.5 MB

Tools

- [Join, Subtract and Group](#)
- [Convert Formats](#)
- [Extract Features](#)
- [Fetch Sequences](#)
- [Fetch Alignments](#)
- [Statistics](#)
- [Graph/Display Data](#)
- [DNApod](#)
- [import mpileupfile Import from DDBJ Pipeline](#)
- [Detect SNPs Detect SNPs](#) (selected)
- [Detect Indels Detect InDels](#)
- [Merge SNPs / InDels](#)
- [Merge SNPs / InDels](#)
- [Visualize distribution of polymorphism](#)
- [SnpEff Variant effect and annotation](#)

Detect\_SNP Detect SNPs (Galaxy)

Select pileup or mpileup file: mpileup

Select DNA polymorphism file(format : samtools): 1: import mpileupfile

(DP) Raw read depth threshold in mpileup file (default: 0): 15

(MQ) Root-mean-square mapping quality threshold in mpileup file (default: 0): 20

(GT) Select hets, homs or both (default: both): homs

(GQ) Genotype quality threshold in mpileup file (default: 0): 0

Execute

History search datasets Unnamed history 1 shown 15.51 MB 1: import mpileupfile

History search datasets Unnamed history 2 shown 23.15 MB 2: SNPs data 1: import mpileupfile

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	
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

Galaxy Analyze Data Workflow Shared Data Visualization Admin Help User Using 23.2 MB

Tools  
 Join, Subtract and Group  
 Convert Formats  
 Extract Features  
 Fetch Sequences  
 Fetch Alignments  
 Statistics  
 Graph/Display Data  
 DNApod  
 Import mpileupfile Import from DDBJ Pipeline  
 Detect SNPs Detect SNPs  
 Detect Indels Detect Indels **1. クリック**  
 Merge SNPs / Indels data files  
 Merge SNPs / Indels data files  
 Visualize discovered polymorphism  
 SnpEff Variant effect and annotation

Detect Indels Detect Indels **2. ファイルフォーマットを指定**  
 Select pileup or mpileup  
 mpileup  
 Select DNA polymorphism file(format : samtools)  
 1: import mpileupfile **3. ヒストリーから解析ファイルを指定**  
 (DP) Raw read depth threshold in mpileup file (default: 0)  
 15  
 (MQ) Root-mean-square mapping quality threshold in mpileup file (default: 0)  
 20  
 (GT) Select hets, homs or both (default: both)  
 both  
 (GQ) Genotype quality threshold in mpileup file (default: 0)  
 80  
**4. 検出条件の指定**  
 Execute **5. 実行**

History  
 search datasets  
 Unnamed history 2 shown  
 23.15 MB  
 2: SNPs data  
 1: import mpileupfile  
 History  
 search datasets  
 Unnamed history 3 shown  
 24.49 MB  
 3: Indels data **6. データの中身を確認**  
 2: SNPs data  
 1: import mpileupfile

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	FILE
chr01	431976	.	gttttttt	gtttttt	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

**Galaxy** Using 24.5 MB

Analyze Data Workflow Shared Data Visualization Admin Help User

Tools Statistics Graph/Display Data DNApod import mpileupfile Import from DDBJ Pipeline Detect SNPs Detect SNPs Detect Indels Detect Indels Merge SNPs / InDels data files Merge SNPs / InDels data files Visualize distribution of DNA polymorphism SnpEff Variant effect and annotation

Merge\_SNPs\_/\_InDels\_data\_files Merge SNPs / InDels data files (Galaxy Version 1.0.0) Options

Select SNPs / InDels data file (format : samtools pileup/mpileup)  
2: SNPs data

Select SNPs / InDels data file (format : samtools pileup/mpileup)  
3: InDels data

✓ Execute

History search datasets Unnamed history 3 shown 24.49 MB 3: InDels data 2: SNPs data 1: import mpileupfile

1. クリック 2. ファイルを指定 3. 実行

History search datasets Unnamed history 4 shown 33.47 MB 4: Merge SNPs / InDels data files on data 3 and data 2 3: InDels data 2: SNPs data 1: import mpileupfile

6. データの中身を確認

#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	gtttttt	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	.	CTTTTTTTTTT	CTTTTTTTTTT,CTTTTTTTT	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。

33

Galaxy Analyze Data Workflow Shared Data Visualizations User Tools

Tools

Graph/Display Data

DNApod

import mpileupfile Import from DDBJ Pipeline

Detect SNPs / InDels

Merge SNPs / InDels

Merge SNPs / InDels data files

**Visualize distribution of DNA polymorphism**

SnpEff Variant effect and annotation

**Visualize distribution of DNA polymorphism**

Select an annotation data  
Oryza sativa(IRGSPbuild1.0)

Select DNA polymorphism file(format : samtools pileup or mpileup)  
2: SNPs data

✓ Execute

Using 33.5 MB

History

search datasets

Unnamed history

4 shown

33.47 MB

4: Merge SNPs / InDels data files on data 3 and data 2

3: InDels data

2: SNPs data

1: import mpileupfile

6. データの中身を確認

History

search datasets

Unnamed history

5 shown

41.19 MB

5: Visualize distribution of DNA polymorphism on data 2

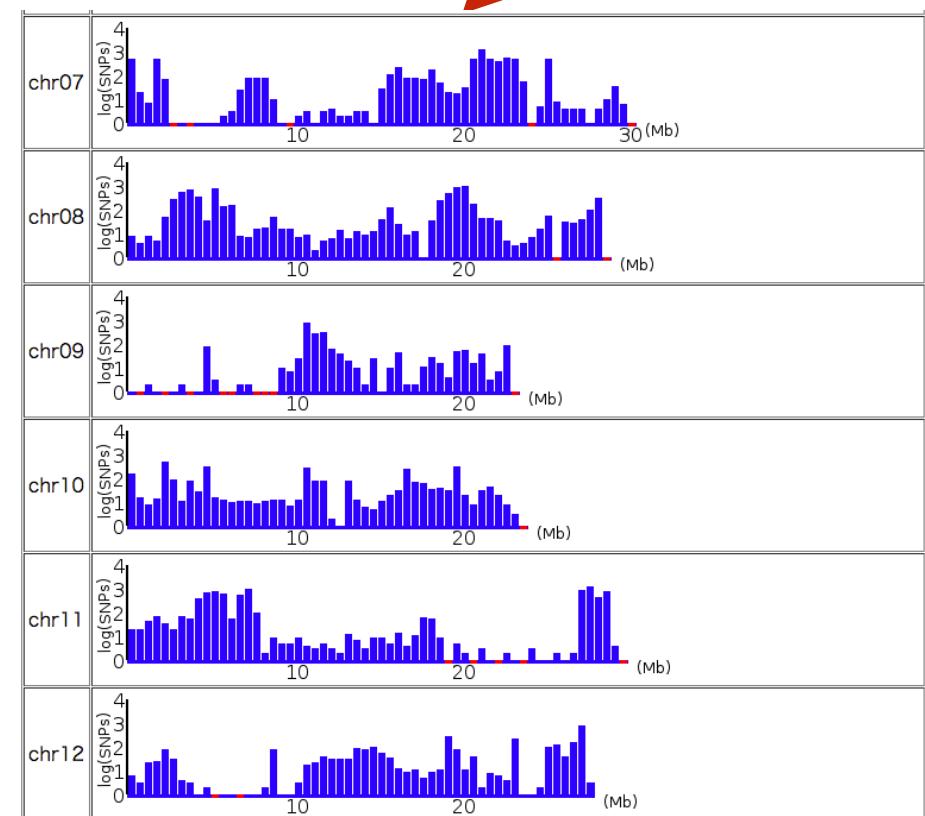
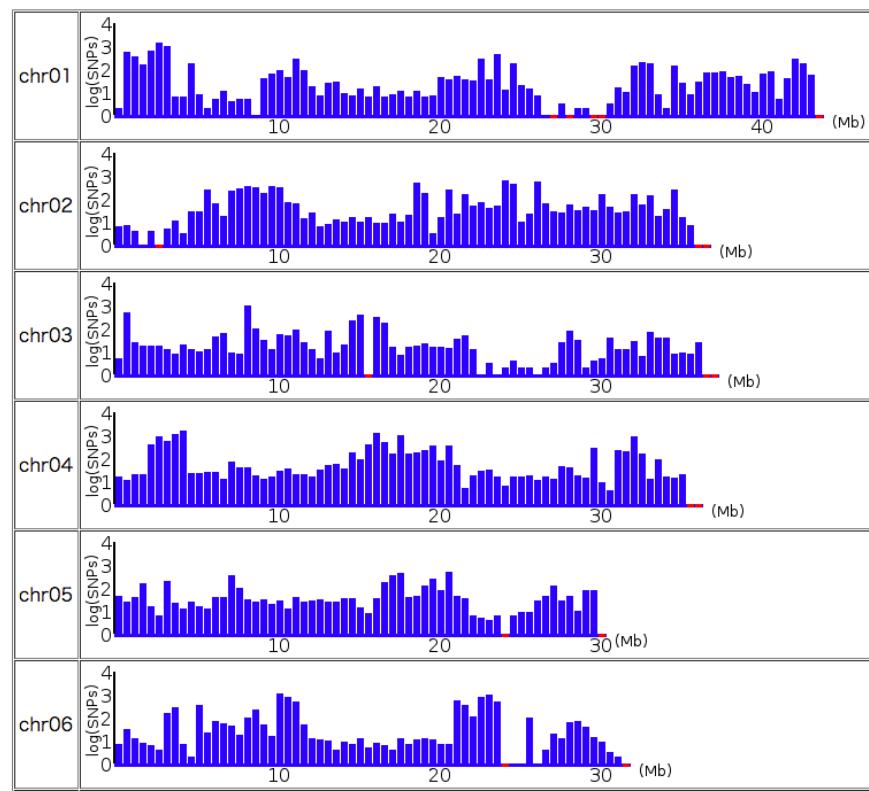
4: Merge SNPs / InDels data files on data 3 and data 2

3: InDels data

2: SNPs data

1: import mpileupfile

Distribution of DNA polymorphisms in chromosomes



Galaxy Using 41.2 MB

Tools search tools

**Get Data**

**Send Data**

**Lift-Over**

**Text Manipulation**

**Filter and Sort**

**Join, Subtract and Group**

**Convert Formats**

**Extract Features**

**Fetch Sequences**

**Fetch Alignments**

**Statistics**

**Graph/Display Data**

**DNApod**

import mpileupfile Import from DDBJ Pipeline

Merge SNPs / InDels data files Merge SNPs / InDels data files Visualize distribution of DNA polymorphism **SnpEff Variant effect and annotation** 1.クリック

Workflows All workflows

Analyze Data Workflow Shared Data Visualization Admin Help User

SnpEff Variant effect and annotation (Galaxy Version 3.6c (build 2014-05-20))

Sequence changes (SNPs, MNPs, InDels) 4: Merge\_SNPs/\_InDels\_data\_files on data 3 and data 2

Input format VCF

Output format VCF

Genome Oryza\_sativa IRGSP-1.0.21

If your annotation of interest is not listed, contact the P-GALAXY team (pipeline\_dev@ddbj.nig.ac.jp)

Upstream / Downstream length 5000 bases

Use only canonical transcripts Yes No

Annotate using HGVS nomenclature Yes No

Annotate Loss of function (LOF) and Nonsense mediated decay (NMD) Yes No

Annotate transcription factor binding site motifs (only available for latest GRCh37) Yes No

Produce Summary Stats Yes No

Execute 4.実行

History search datasets Unnamed history 5 shown 41.19 MB 5: Visualize distribution of DNA polymorphism on data 2 4: Merge SNPs / InDels data files on data 3 and data 2 3: InDels data 2: SNPs data 1: import mpileupfile

History search datasets Unnamed history 7 shown 74.62 MB 7: SnpEff on data 4 6: SnpEff on data 4 5: Visualize distribution of DNA polymorphism on data 2 4: Merge SNPs / InDels data files on data 3 and data 2 3: InDels data 2: SNPs data 1: import mpileupfile

2.ヒストリーから解析ファイルを指定

3.入力、出力ファイルの形式を選択

3.アノテーションを指定

今日はこのオプションで実行

ファイルが2つ作成されます。

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|S11|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:QQ	1/1:171,51,0:99
GT:PL:QQ	1/1:250,99,0:99
GT:PL:QQ	1/1:157,48,0:93
GT:PL:QQ	1/1:255,135,0:99
GT:PL:QQ	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](http://snpeff.sourceforge.net/SnpEff_manual.html#input)

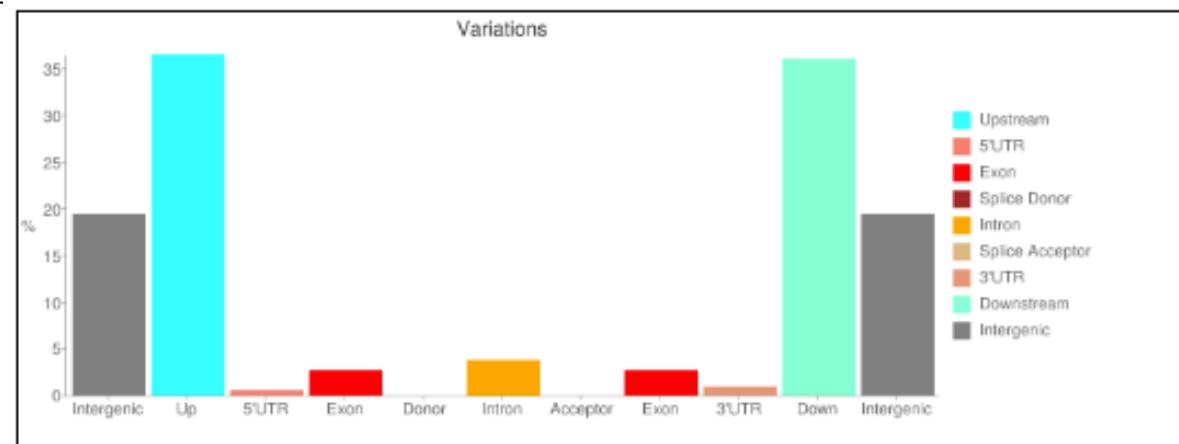
## 解析統計情報を表示

## 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



# ご清聴ありがとうございました

## DNApod データベース

**DNApod**

HOME Database DDBJ pipeline HELP

Summary

species      subspecies      type      strain

(all)      (all)      (all)     

Please click "DRA accession" to show results and download files.

DRA accession	species	subspecies	subtaxa	type	strain	id	coverage	depth	SNP	homo SNP
D_K000010	Oryza sativa	Japonica		Cultivar	Koshihikari		96.7	21.8	107,152	81,133
D_K000090	Oryza sativa	Japonica		Cultivar	Nipponbare		5.0	5.0	16,441	11,144
ERX000322	Oryza sativa	Indica		Cultivar	Guangluai-4		4.9	4.9	1,993	5,993
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	2,497,846
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		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

**DDBJ**

ACCOUNT  
login ID [guest]  
Logout

ANALYSIS  
Data setup  
DRA Start  
FTP upload  
HTTP upload  
DRA Import  
Preprocessing Start

step-1  
fastQC  
mapping /  
de novo assembly

step-2  
Workflow  
Genome (SNP/Short Indel)  
RNA-seq (Tag count)  
ChIP-seq

JOB STATUS  
step 1.  
Preprocessing  
step 1.  
Mapping  
step 1.  
de novo Assembly  
step 2-All status

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

**Selecting Tools for Basic Analysis of DDBJ ANNOTATION PIPELINE**

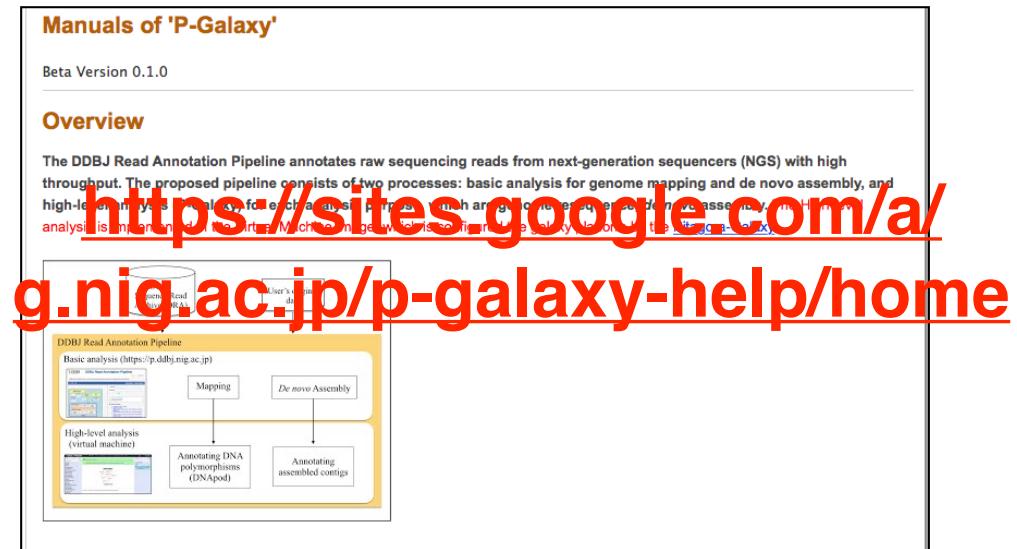
**Reference Genome Mapping**

Tool	Help	Version	Input data	Evaluation	Analysis	Output format						
BLAT	3.4		Base space	Paired space end	Depth Coverage	Error rate	SNP	Indel	.gff	.bed	SAM	Comment
Map	0.1		Base space	Paired space end	Depth Coverage	Error rate	SNP	Indel	.gff	.bed	SAM	Comment
bwa	0.5.9		Base space	Paired space end	Depth Coverage	Error rate	SNP	Indel	.gff	.bed	SAM	Comment
SOAP	2.21		✓	✓	✓	✓	✓	✓	✓	✓	✓	Single-end analysis only
Bowtie	0.12.7		✓	✓	✓	✓	✓	✓	✓	✓	✓	
TopHat	1.0.11		✓	✓	✓	✓	✓	✓	✓	✓	✓	
Bowtie2	2.0.0		✓	✓	✓	✓	✓	✓	✓	✓	✓	
TopHat2	2.0.9		✓	✓	✓	✓	✓	✓	✓	✓	✓	

For reads longer than about 50 bp, Bowtie2 is generally faster, more sensitive, and uses less memory than Bowtie1.

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

### 高次処理部 DNApod ワークフロー



# 追加資料

**DDBJ Pipeline 基礎処理部**

Quality Value (QV) フィルタ

## クエリの選択

**1. クリック**

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

Import public FASTQ files from DRA database.  
Please input DRA/ERA/SRA accession number. Then the pipeline system imports the data.

Input DRA/ERA/SRA Accession Number  
DRA000307 Add my DRA entry

**3. クリック**

**4. SRAアクセッションを選択**

Select a metadata : DRA000307

TYPE	ACCESSION	ALIAS	FILENAME	DL	VIEW
Submission	DRA000307	tomohiro-0005_Submission	DRA000307.submission.xml	DownLoad	View
Sample	DRS000412	tomohiro-0005_Sample_0001	DRA000307.sample.xml	DownLoad	View
Study	DRP000308	tomohiro-0005_Study_0001	DRA000307.study.xml	DownLoad	View
Experiment	DRX000450	tomohiro-0005_Experiment_0001	DRA000307.experiment.xml	DownLoad	View
Run	DRR000719	tomohiro-0005_Run_0001	DRA000307.run.xml	DownLoad	View
	DRR000720	tomohiro-0005_Run_0002			

**5. クエリにするアクションを選択**

Select your query

Single paired all clear

No.	Experiment ACCESSION	Sample ACCESSION	Run ACCESSION	STRAIN	Run_date	Read #	Read length	Instrument model	Layout
1	DRX000450	DRS000412	DRR000719		2009-08-14			ILLUMINA	paired
2	DRX000450	DRS000412	DRR000720		2009-08-27			ILLUMINA	paired

from metadata : Counted from FASTQ (Sequence length is calculated from the first 1000 bases)

**5. クリック**

NEXT

## 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 :

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

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

QV THRESHOLD :   
 Percentage THRESHOLD :

**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**

## 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 :

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

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

QV THRESHOLD :   
 Percentage THRESHOLD :

**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**

**Run Confirmation**

**BACK** **RUN**

**Email notification**

Send email notification when the job is completed or aborted with error.

\* Required

**Confirmation of entries**

**Query sets**

- DRR000719 - tomohiro-0005\_Run\_0001
- DRR000720 - tomohiro-0005\_Run\_0002

**BACK** **RUN**

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

ID	User ID	Files	P/S	Status	Read #	Read length	Detail	Start time	Elapsed time
22741	koshu01	DRA000307 tomohiro-0005_I tomohiro-0005_I	P	complete		--	<a href="#">View</a>	2016-06-27 16:11:05	18:44:19
16231	koshu01	SRA012701 GSM497271_1	S	complete		--	<a href="#">View</a>	2015-02-28 17:37:01	00:13:52

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

ID	User ID	Files	P/S	Status	Read #	Read length	Detail	Start time	End time	Elapsed time
22741	koshu01	DRA000307 tomohiro-0005_I tomohiro-0005_I	P	complete		--	<a href="#">View</a>	2016-06-27 16:11:05		18:44:19
16231	koshu01	SRA012701 GSM497271_1	S	complete		--	<a href="#">View</a>	2015-02-28 17:37:01		00:13:52

**Detail view**

**Job info**

ID	22741
Tool (Version)	(1.0)

RunAccession or Filename	Download	Read length	Alias
DRR000719	<a href="#">DRR000719.fastq.bz2</a> <a href="#">DRR000719_1.fastq.bz2</a> <a href="#">DRR000719_2.fastq.bz2</a>	N.A. bp	tomohiro-0005_Run_0001
DRR000720	<a href="#">DRR000720.fastq.bz2</a> <a href="#">DRR000720_1.fastq.bz2</a> <a href="#">DRR000720_2.fastq.bz2</a>	N.A. bp	tomohiro-0005_Run_0002

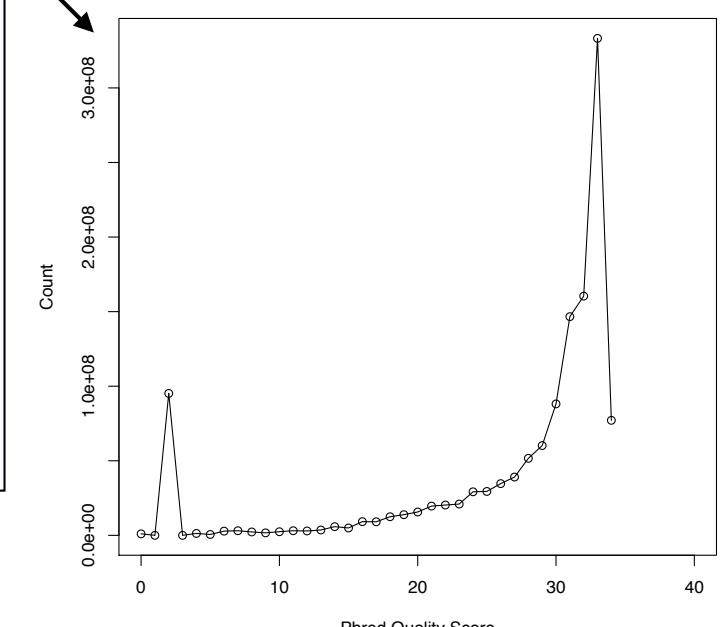
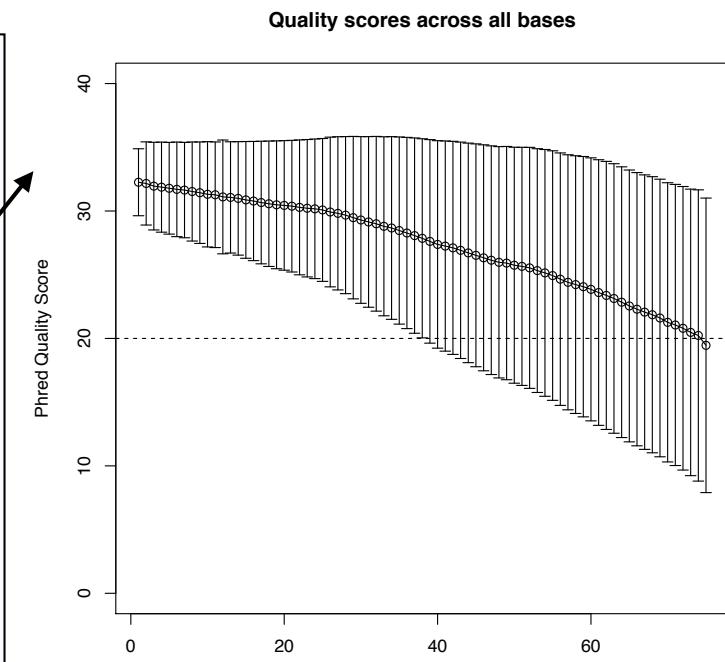
File	Fastq Download	QS Average (PDF)	QS Count (PDF)
DRR000719_1.fastq.bz2	<a href="#">download (4.2 GB)</a>	<a href="#">download (8.7 KB)</a>	<a href="#">download (5.2 KB)</a>
DRR000719_2.fastq.bz2	<a href="#">download (4.2 GB)</a>	<a href="#">download (8.6 KB)</a>	<a href="#">download (5.2 KB)</a>
DRR000720_1.fastq.bz2	<a href="#">download (32.6 GB)</a>	<a href="#">download (8.6 KB)</a>	<a href="#">download (5.2 KB)</a>
DRR000720_2.fastq.bz2	<a href="#">download (32.4 GB)</a>	<a href="#">download (8.7 KB)</a>	<a href="#">download (5.1 KB)</a>

**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	<a href="#">View</a>			
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	<a href="#">View</a>			

[BACK](#)



**1. クリック**

**2. Preprocessing をクリック**

**2. データを選択**

QV フィルタのJob ID + Run accession ID  
でデータを識別する。

**2. 次へ**

Filename	Layout	File size
22741_DR000720_1_e.fastq.bz2 (more 1 files)	paired	16.1 GB
22741_DR000719_1_e.fastq.bz2 (more 1 files)	paired	2.0 GB
22716_mapped.unmapped.fastq.bz2 (more 1 files)	paired	486.5 MB
22003_mapped.unmapped.fastq.bz2	single	51.4 MB
20654_mapped.unmapped.fastq.bz2	single	355.8 MB
19716_mapped.unmapped.fastq.bz2	single	55.9 MB
17855_E...unmapped.fastq.bz2	single	55.9 MB
17854_ERR018562.unmapped.fastq.bz2	single	14 byte
17851_ERR018562.unmapped.fastq.bz2	single	14 byte
17811_ERR018562.unmapped.fastq.bz2	single	55.9 MB
16721_mergedout.unmapped.fastq.bz2	single	49.8 MB
16231_SRR042533_e.fastq.bz2	single	239.7 MB
5906_SRR042533_e.fastq.bz2	single	239.7 MB

Mapping の後工程は、SRA データを使用したときと同じです。

47