

統合データベース講習会：AJACS千里

2015年6月16日

次世代シーケンサー（NGS）と 関連するデータベース・ツール

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

河野 信

本日紹介する内容（次世代シーケンスDB）

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

次世代シーケンサデータベースの概要

次世代シーケンサー (NGS)

◆ 第一世代：キャピラリーシーケンサー

◆ 第二世代

- 短い配列 (~ 150 bp) を超並列 ($\sim 12G$) に読む
- Illumina HiSeq/NextSeq/MiSeq, 454, SOLiD



Togo picture gallery by DBCLS is Licensed under a Creative Commons 表示 2.1 日本 (c)

◆ 第三世代

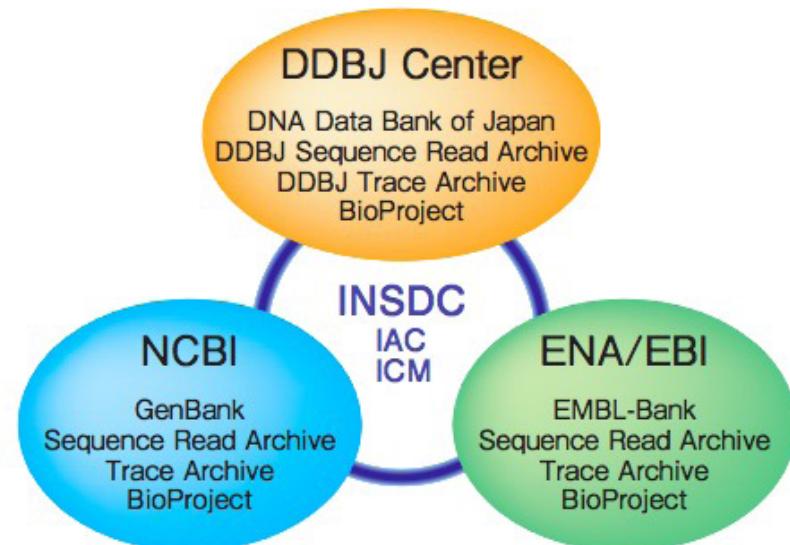
- PacBio
 - 超ロングリード
- NanoPore
 - USBメモリサイズ



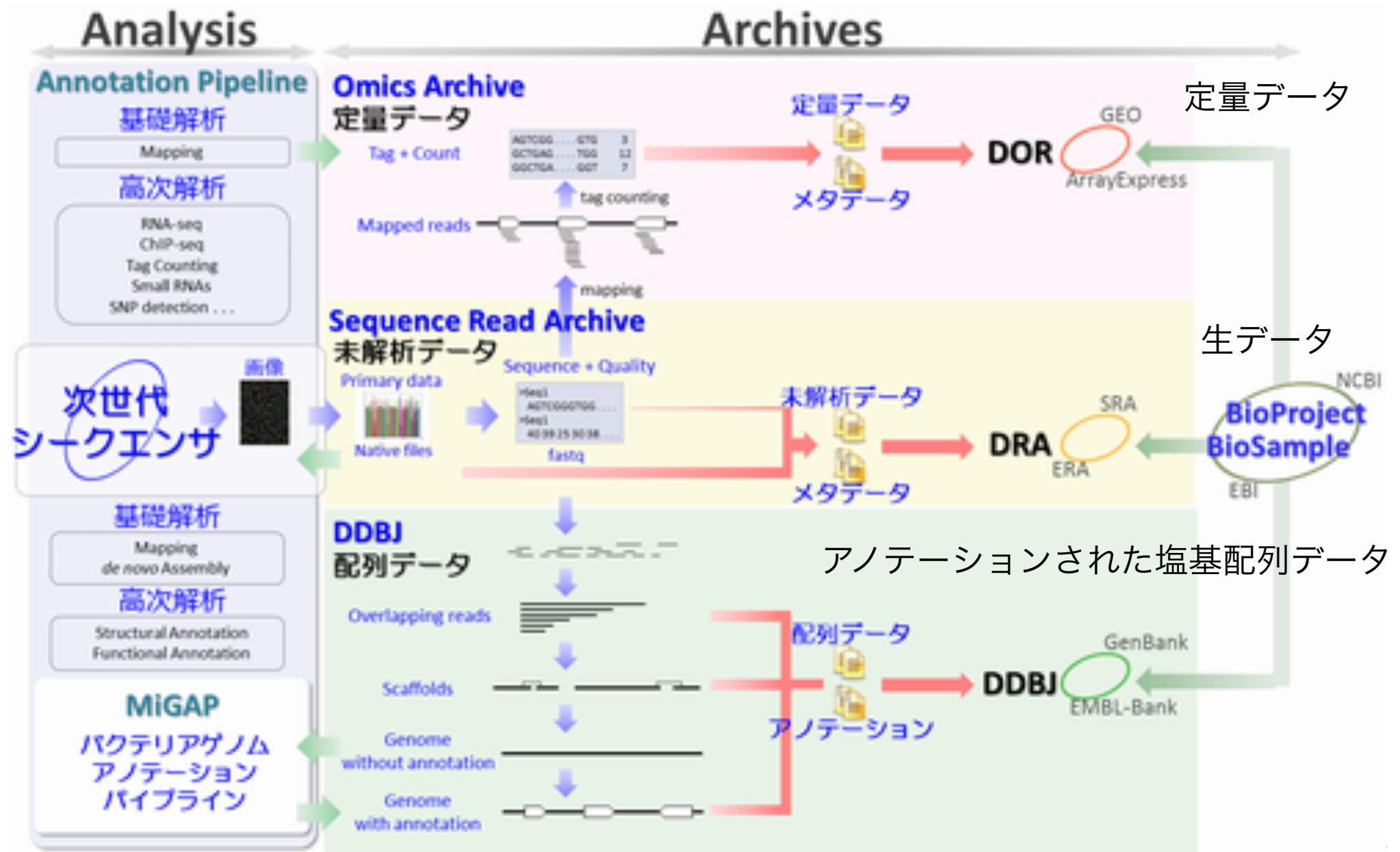
Sequence Read Archive

◆ SRA (NCBI)/ENA (EBI)/DRA (DDBJ)

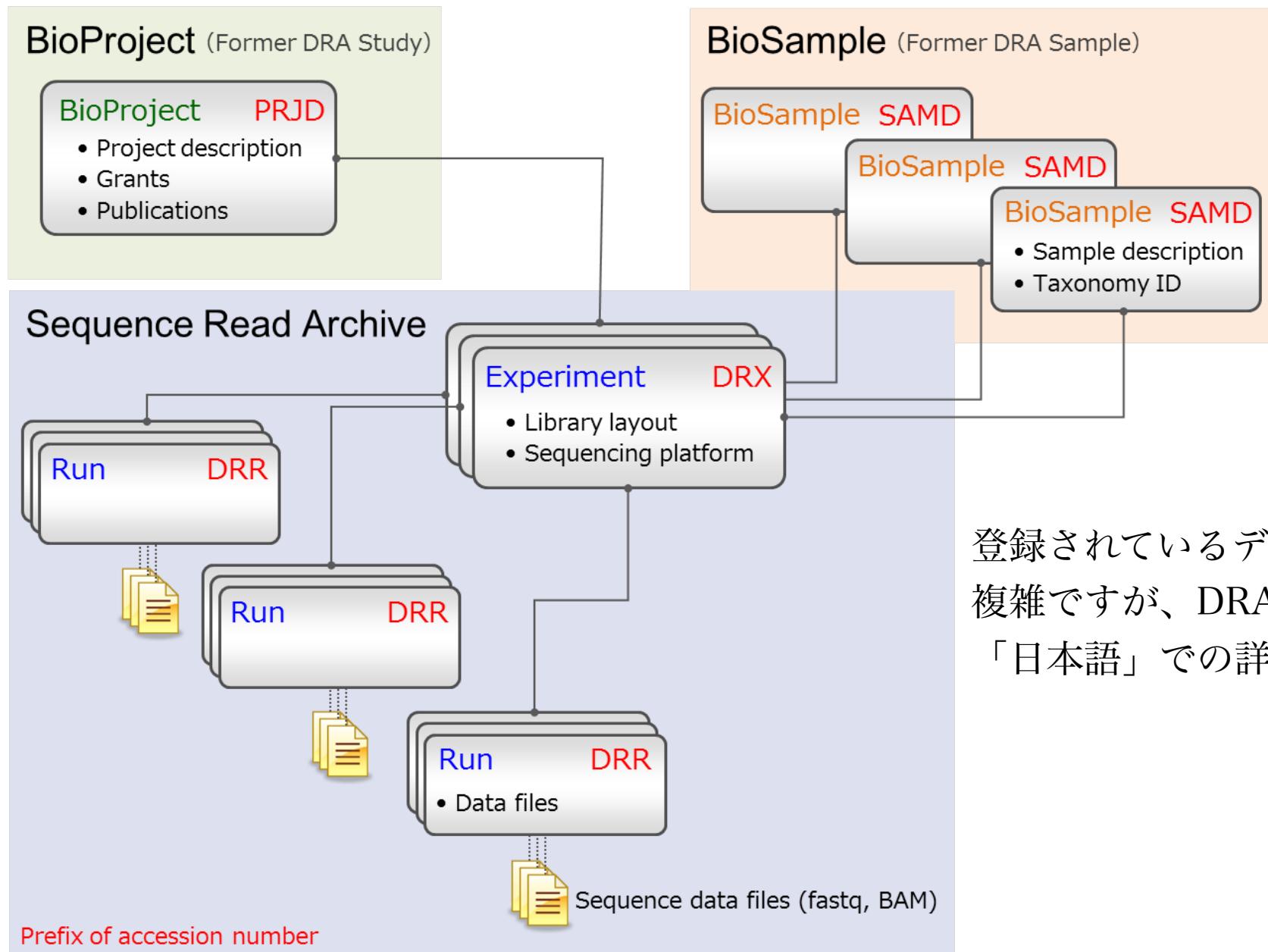
- 次世代シーケンスデータはどれかに登録する
 - 生データとマッピングデータ、メタデータのみ登録
 - 解析データは従来通りGenBank/ENA/DDBJに登録する
 - 定量データはGEO/ArrayExpress/(DOR)に登録する
- 3データベースでデータを毎日交換しており、データの内容は同じ



データの種類とデータベース



次世代シーケンスのデータ構造



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

実習

DRAsearchを使って
次世代シーケンスデータを検索する

DDBJ トップページ

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

DDBJ/DRA登録 データ検索・解析

データダウンロード

The screenshot shows the DDBJ homepage with several red annotations:

- A red circle highlights the "Data Submission" icon under the "DDBJ Service" section.
- Three red arrows point from the top right towards the "Search / Analysis", "Super Computer", and "Archive" icons.
- A large red arrow points from the "Data Submission" icon down to the "Hot Topics" section.

DDBJ Service

- 登録 (Data Submission)
- 検索・解析 (Search / Analysis)
- スパコン (Super Computer)
- アーカイブ (ftp. ddbj.nig.ac.jp)

Hot Topics

- お知らせ (Alert) | データ公開 (Data Release) | 広報 (Press Release) | メンテナンス (Maintenance) | 運用情報 (Operational Information) | 一覧 (List)
- 2015.06.03 UniProt 2015_06 公開
- 2015.05.07 UniProt 2015_05 公開
- 2015.04.20 「第31回 DDBJing 講習会 in 東京」開催のお知らせ

DDBJデータ登録画面

English

Google™カスタム検索 Search

塩基配列の登録 プロジェクトの登録 塩基配列登録の前に Flat File の説明 お問い合わせ

HOME > データ登録 最終更新日 : 2015.3.19.

データ登録

データ登録方法

- アノテーションをつけた塩基配列の登録
 - DDBJ Nucleotide Sequence Submission System
Web経由の塩基配列登録システム
 - Mass Submission System (MSS)
登録予定データが、件数が多い、多数の Feature を持つ、配列が長大、などの場合や、web経由の登録システムが対応していないデータ（例：WGS）の登録
 - DDBJ Sequence Read Archive (DRA)**
次世代シーケンサーから出力されたデータの登録
- DDBJ Trace Archive (DTA)
Sanger法をベースとしたシーケンサーから得られるクロマトグラム（traces）の登録
- DDBJ BioProject Database
研究プロジェクトの登録。データ（WGS, complete genome, transcriptome project, DRA, DTAなど）をプロジェクト単位でグループ化するためのBioProject IDを発行。
- DDBJ BioSample Database
実験データを得るために使用された生物学的な試料（サンプル）についての情報を集中して管理するデータベース。BioSample IDを発行。
- Japanese Genotype-phenotype Archive (JGA)
個人に由来する遺伝学的なデータと匿名化された表現型情報の登録。NBDCヒトデータ共有ガイドラインに則って運用されています。

登録データ種別

- 塩基配列の登録・更新・修正
- 登録前にお読みください
 - 塩基配列の登録について
登録に必要なデータや登録情報の具体的な記述方法など
 - 修正・更新時にお読みください
塩基配列登録データの修正・更新
- その他のデータの登録・更新・修正
 - DRA/DTA/BioProject/BioSample ご利用の際にお読みください
 - D-way 登録アカウントマニュアル
データの登録に必要なD-wayアカウント取得の方法など
 - DRA マニュアル
 - BioProject マニュアル
 - BioSample マニュアル
 - DTA マニュアル

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

 DDBJ
DNA Data Bank of Japan

Sequence Read Archive

Login & Submit | Databases ▾ | English | Contact

Home | Handbook | FAQ | **Search** | Download ▾ | Pipeline | About DRA

データ取得

News 登録関連情報 データ検索 — 解析パイプライン
2015年04月23日: 優歎サンノルのアルノアベット順ソートに関するご注意

DDBJ Sequence Read Archive (DRA) は Roche 454 GS System®, Illumina Genome Analyzer®, Applied Biosystems SOLiD® System などの次世代シークエンサからの出力データのためのデータベースです。DRA は International Nucleotide Sequence Database Collaboration (INSDC) のメンバーであり、NCBI Sequence Read Archive (SRA) と EBI Sequence Read Archive (ERA) との国際協力のもと、運営されています。従来のキャピラリ式シークエンサからの出力データは DDBJ Trace Archive にご登録ください。

 検索

 登録

データセットの検索：カテゴリを指定

DRA Search

Accession :

Organism : StudyType : StudyTypeを指定

CenterName : Platform : Platformを指定

Keyword :

Show 20 results

Statistics

Released Entries

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

生物種の指定
(数文字入力すると候補が表示される)

Organism

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

Study Type

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

Center Name

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

メタデータによる分類

12

参考：約1年前の登録数

DRA Search

Send Feedback | Search Home | DRA Home

Accession :

Organism : Homo

CenterName : Homo

Keyword : Homo sapiens

Show 20

StudyType : Cancer Genomics

Platform : ILLUMINA

Search Clear

Data Last Update 2014-07-01
WebSite Last Update 2014-01-22

Statistics

Released Entries

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

Organism

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

Study Type

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

Center Name [All List]

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

データセットの検索：キーワード検索

DRA Search

Accession :

Organism : StudyType :

CenterName : Platform :

Keyword : stap

Show 20 records Sort by Study Search Clear

Data Last Update 2015-06-14
WebSite Last Update 2015-04-08

Statistics

Released Entries

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

何か適当なキーワードで検索

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

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

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

メタデータによる分類

DRA検索結果

DRA Search

Accession :

Organism : StudyType :

CenterName : Platform :

Keyword :

Show 20 records Sort by Study

Search Results (7 records) / 1 Page

Filtered by
document type:experiment(2) run(2) study(2) sample(1)
organism:Mus musculus(4)

#	META_FILE	ACCESSION	STUDY	STUDY_TITLE	STUDY_TYPE	ORGANISM	BASES	SUBMITTED	CENTER_NAME
1	DRA002862.study.xml for research investigation of STAP cells</STUDY_TITLE> <STUDY_TYPE existing	DRP002445	DRP002445			Mus musculus	1.3T	2014-12-26	
2	SRA110029.run.xml _ID> </IDENTIFIERS> </EXPERIMENT_REF> </RUN> <RUN alias="SMARTer_ STAP RNA-seq" accession="SRR1171578	SRR1171553 SRR1171554 SRR1171555 SRR1171556 SRR1171557 SRR1171558 SRR1171559 SRR1171560 SRR1171561 SRR1171562 SRR1171563 SRR1171564 SRR1171565 SRR1171566 SRR1171567 SRR1171568 SRR1171569 SRR1171570 SRR1171571 SRR1171572 SRR1171573 SRR1171574 SRR1171575 SRR1171576 SRR1171577 SRR1171578 SRR1171579 SRR1171580 SRR1171581 SRR1171582	SRP038104	Mus musculus 1)Transcriptome or Gen expression; 2)Histone Modification H3K4me3, H3K27me3	Transcriptome Analysis	182.6G	BioProject		

BioProjectのIDをクリック

BioProjectエントリ

DRA Search

Send Feedback

 Search Home

 DRA Home

SRP038104

Study Detail

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

Navigation

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

ExperimentのIDをクリック

 Sample	SRS559080
	SRS559081
	SRS559082
	SRS559083

Experimentエントリ

DRA Search

Send Feedback

 Search Home

 DRA Home

SRX472627  

Experiment Detail

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

Library Description

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

Platform

Platform	ILLUMINA
Instrument Model	Illumina HiSeq 1500

Processing

Base Calls

Sequence Space	
Base Caller	

Send Feedback

 Search Home

 DRA Home

Navigation

 Submission	SRA110029	
 Study	SRP038104	
 Sample	SRS559080	
 Run	SRR1171556	 

サンプル情報

データをダウンロード

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

DDBJ トップページ

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

The screenshot shows the DDBJ homepage with a red arrow pointing from the text "無料で利用可能なスパコン" (Free to use Super Computer) to the "Super Computer" icon in the "DDBJ Service" section.

DDBJ Service

- 登録 Data Submission
- 検索・解析 Search / Analysis
- スパコン Super Computer
- アーカイブ ftp. ddbj.nig.ac.jp

Hot Topics

- 2015.06.03 UniProt 2015_06 公開
- 2015.05.07 UniProt 2015_05 公開
- 2015.04.20 「第31回 DDBJing 講習会 in 東京」開催のお知らせ

Logos of Collaborating Institutions:

- INSDC
- NCBI
- ENA/EBI
- NIG 国立遺伝学研究所
- RISE 大学共同利用機関法人 情報・システム研究機構
- DBCLS Database Center for Life Science

遺伝研スーパーコンピュータ

◆ 計算ノード

- 64GB memory x 554 nodes
- 2TB memory x 10 nodes
- 10TB memory x 1 node

◆ Storage

- 7 PB 高速HDD
- 5.5 PB 大容量省電力HDD

◆ 利用申請することで無料で利用できます！

- コマンドラインが基本
- (停電で結構頻繁に落ちる)

スパコン利用申請はこちら

 大学共同利用法人 情報・システム研究機構 国立遺伝学研究所
スーパーコンピュータシステム
SuperComputer Facilities of National Institute of Genetics

現在地: Home 2015年06月06日 サイトポリシー サイトマップ 検索... 検索

Language/言語
: ● : 

[ホーム](#)

このサイトへのログイン
 [Login](#)
(スパコンユーザでログイン可)

システム構成
[ハードウェア構成](#)
[ソフトウェア構成](#)
[プログラミング環境](#)
[利用可能バイオツール](#)
[利用可能OSS](#)
[利用可能DB](#)

システム使用方法
[基本的利用方法](#)
[その他UGE利用方法](#)
[ファイル転送方法](#)
[システム利用TIPS](#)
[稼働スケジュール](#)

各種申請
[システムの利用条件](#)
[各種申請窓口について](#)
[新規ユーザ登録申請](#)
[大規模利用申請](#)
[MiGAP利用申請](#)
[DDBJ Pipeline利用申請](#)
[OSSインストール申請](#)
[アカウント継続・停止申請](#)

重要なお知らせ

一覧へ

公開日 表題

2015年5月27日 Javaアプリケーションの標準利用環境のバージョン更新について

2015年5月12日 【Lustre3障害】Lustre3障害のお知らせ

2015年5月12日 【定期メンテナンス】7月8日～7月11日 定期メンテナンスに伴うサービス停止のお知らせ

2014年9月10日 【スパコンユーザ会】会議報告

2014年3月4日 2014年3月5日からのスパコンPhase2システムご利用方法について

国立遺伝学研究所 スーパーコンピュータシステム(NIG SUPERCOMPUTER)とは

大学共同利用機関法人 情報システム研究機構 国立遺伝学研究所は、2012年3月にスーパーコンピュータシステムを更新しました。新しいスーパーコンピュータシステムはゲノム解析を主な目的とした大規模計算機利用拠点として最新鋭の大規模クラスタ型計算機、大規模メモリ共有型計算機、および大容量高速ディスク装置で構成されたスーパーコンピューティングシステムサービスを提供しています。



[システムハードウェア構成](#)
[システムソフトウェア構成](#)
[システム稼働状況](#)

本サイトは国立遺伝学研究所スーパーコンピュータシステムが提供する計算機リソース、各種アプリケーション、それらの利用方法についての各種情報を提供します。DDBJセンターとして提供する各種サービスについては[DDBJセンターのホームページ](#)からご参照ください。

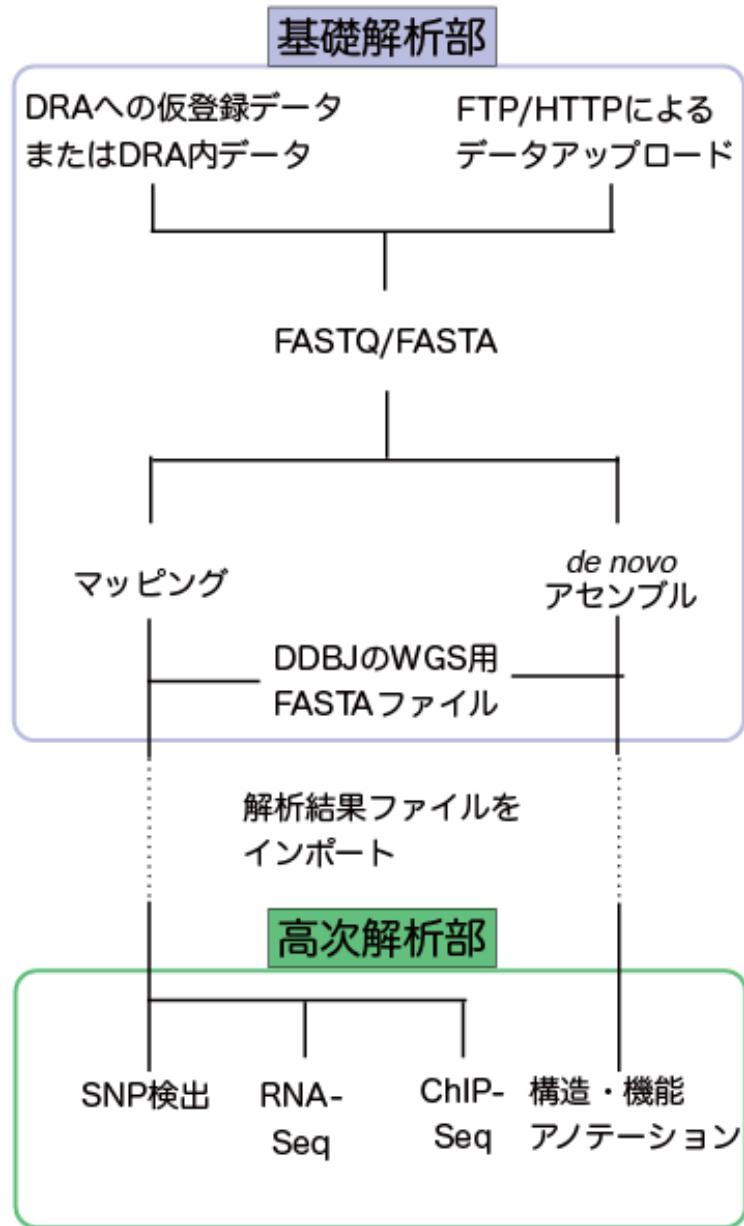
ディスク利用状況

Lustreのsize,file項目は、現在のディスク使用量（全ユーザ合計）／ディスク容量で表現しています。
quotaのsize項目は、申請ディスク使用量（全ユーザ合計）／ディスク容量で表現しています。
sizeの単位は、Tbyteです。

ディスク	size	quota	状況
lustre1	960.76 / 1,024.90	93%	lustre1 file : 195,218,045 / 731,840,512 26%
quota	1,981.85 / 1,024.90	193%	
lustre2	928.15 / 1,024.90	90%	lustre2 file : 643,447,895 / 731,840,512 87%
quota	1,089.65 / 1,024.90	106%	
lustre3	518.25 / 1,787.70	28%	lustre3 file : 78,260,318 / 731,840,512 10%
quota	1,401.64 / 1,787.70	78%	

いや、
そうぢやない

DDBJ解析パイプライン



DDBJ Read Annotation Pipeline

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

Galaxy / P-GALAXY

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

DDBJのスパコンをウェブ経由で利用

実習

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

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

 DDBJ
DNA Data Bank of Japan

Login & Submit | Databases ▾ | English | Contact

Sequence Read Archive

Home | Handbook | FAQ | Search | Download ▾ | Pipeline | About DRA

News  **解析パイプライン**

2015年04月23日: 複数サンプルのアルファベット順ソートに関するご注意

DDBJ Sequence Read Archive (DRA) は Roche 454 GS System®, Illumina Genome Analyzer®, Applied Biosystems SOLiD® System などの次世代シークエンサからの出力データのためのデータベースです。DRA は International Nucleotide Sequence Database Collaboration (INSDC) のメンバーであり、NCBI Sequence Read Archive (SRA) と EBI Sequence Read Archive (ERA) との国際協力のもと、運営されています。従来のキャピラリ式シークエンサからの出力データは DDBJ Trace Archive にご登録ください。


検索


登録

DDBJ Read Annotation Pipeline

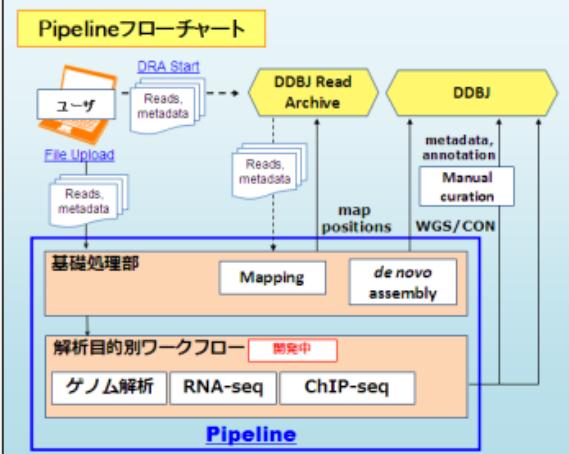


DDBJ Read Annotation Pipeline

English Japanese

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。

LOG IN



新規アカウント作成

ゲストとしてログイン

User ID:

Password:

Login

動作中JOBの確認

PipelineのIDをお持ちでない場合、[ゲストとしてログインすることができます](#)。

マニュアルおよびチュートリアル

- [日本語チュートリアル \(FAQ\)](#)
- [英語マニュアル](#)
- [DBCLS 統合TV チュートリアル1 - 今日からはじめるDDBJ Read Annotation Pipeline](#)
- [DBCLS 統合TV チュートリアル2 - DDBJ Read Annotation Pipelineによるde novo Assembly解析](#)
- [チュートリアル : FTPでファイルをアップロードしDDBJ Pipelineへ登録する方法](#)
- [チュートリアル : DDBJ PipelineでHGAP法でPacBioリードのアンブリを行ふ方法](#)

塩基配列・解析結果の登録

- [DRA : NGS出力データの登録](#)
- [DDBJ-INSDC : アノテーション済の塩基配列データの登録](#)

Citation

- Nagasaki, H. et al., "DDBJ Read Annotation Pipeline: A cloud computing-based pipeline for high-throughput analysis of next-generation sequencing data", DNA Res, 20:383-390, 2013.
- Kaminuma, E. et al., "DDBJ launches a new archive database with analytical tools for next-generation sequence data", Nucleic Acids Res, 38:D33-38, 2010.

Tweets

Follow

pipeline
@pipeline_info

1 May

Pipeline e-mail support will not open from May 2 to May 6 due to Golden Week holidays. Thank you for your cooperation.

pipeline
@pipeline_info

22 Dec

Pipeline e-mail support will not open from Dec 25 - Jan 5. Thank you for your cooperation.



データセットの選択

FTPで手持ちのデータを
アップロード



ACCOUNT
login ID [guest]

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

step-1
Preprocessing
Mapping /
de novo Assembly

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

JOB STATUS
step1.
Preprocessing
step1.
Mapping
step1.
de novo Assembly
step2-All status

HELP
HELP
TUTORIAL
 Contact Us.
DDBJ Read Annotation
Pipeline.
Development Team.

DRAからデータをインポート

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

Running Status

Selecting Query Files

FTP upload Private DRA entry Import public DRA Preprocessing HTTP upload

NEXT

Metadata of the DRA entry.

Select a metadata : DRA000001

TYPE	ACCESSION	ALIAS	FILENAME	DL	VIEW
Submission	DRA000001	DRA000001	DRA000001.submission.xml	<input type="button" value="DownLoad"/>	<input type="button" value="View"/>
Sample	DRS000001	DRS000001	DRA000001.sample.xml	<input type="button" value="DownLoad"/>	<input type="button" value="View"/>
Study	DRP000001	DRP000001	DRA000001.study.xml	<input type="button" value="DownLoad"/>	<input type="button" value="View"/>
Experiment	DRX000001	DRX000001	DRA000001.experiment.xml	<input type="button" value="DownLoad"/>	<input type="button" value="View"/>
Run	DRR000001	DRR000001	DRA000001.run.xml	<input type="button" value="DownLoad"/>	<input type="button" value="View"/>

STUDY TITLE Whole genome sequencing of *Baillus subtilis* subsp. *natto* BEST195
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 type="checkbox"/>	1 DRX000001	DRS000001	DRR000001	strain BEST195	2008-09-13	9,977,388	36	ILLUMINA	paired

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

DELETE NEXT

HTTPで手持ちのデータを
アップロード

SRA/DRAデータのインポート

The screenshot shows the DDBJ Pipeline interface for importing SRA/DRA data. The main title is "Selecting Query Files". The "Import public DRA" tab is selected. In the center, there is a form to input a DRA accession number, with "SRA066203" entered. Below it, a button labeled "Add my DRA entry" is highlighted with a red circle. A red arrow points from this button to a "Confirmation" dialog box on the right. The dialog box contains instructions: "Click a OK button to start import. This operation may take several minutes to several hours." It also includes options: "Send a mail when completed importing." and "Show a accessions list." At the bottom of the dialog, there are "OK" and "Cancel" buttons. On the left side of the interface, there is a sidebar with various menu items like "ACCOUNT", "ANALYSIS", "JOB STATUS", and "HELP". The "JOB STATUS" section shows a list of pending tasks: "step1. Preprocessing", "step1. Mapping", and "step1. de novo Assembly". The "HELP" section includes links to "HELP", "TUTORIAL", "Contact Us.", and "Development Team.". At the bottom, there is a "DELETE" button and a "NEXT" button.

※実行しないでください！

Confirmation

Click a OK button to start import.
This operation may take several minutes to several hours.

Reload the page, to update your importing status.

Option

Send a mail when completed importing.

Show a accessions list.

OK Cancel

DELETE NEXT

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

Running Status

ACCOUNT

login ID [orenoddbj]
Logout Change password

ANALYSIS

Data setup
DRA Start
FTP upload
HTTP upload
DRA Import
Preprocessing Start
step-1
Preprocessing
Mapping /
de novo Assembly
step-2
Workflow
Genome (SNP/Short
Indel)
RNA-seq (Tag count)
ChIP-seq

JOB STATUS

step1.
Preprocessing
step1.
Mapping
step1.
de novo Assembly
step2-All status

HELP

HELP
TUTORIAL
Contact Us.
DDBJ Read Annotation
Pipeline.
Development Team.

FTP upload Private DRA entry Import public DRA Preprocessing HTTP upload

Import public FASTQ files from DRA database.

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

SRA066203 Add my DRA entry

Accession Number can find here.
[DRA Search](#)

Your request. (Here is display only. can not select.)

To select your downloaded entries. See Private DRA entry tab.
When the status makes "done", your requested entry is added in
When the status makes "failed" or "preparing", please retry it.

queued : waiting or during download, done : file is
DRA uncheckd : download is ok, but md5 was not ch

Status	Submission
queued	SRA066203
done	ERA013525
done	SRA009211
preparing	SRA026538

操作1

配列のクオリティをチェックする

インポートされたデータ

DDBJ DNA Data Bank of Japan

ACCOUNT
Login ID [orenoddbj]
Logout
Change password

ANALYSIS
Data setup
DRA Start
FTP upload
HTTP upload
DRA Import
Preprocessing Start
step-1
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.
3 DDBJ Read Annotation Pipeline Development Team.

Selecting Query Files

FTP upload Private DRA entry 2

Metadata of the DRA entry.

Select a metadata: SRA066203

TYPE	ACCESSION	ALIAS	FILENAME	DL	VIEW
Submission	SRA066203	Ecoli_repeat_structure	SRA066203.submission.xml	DownLoad	View
Sample	SRS399367 SRS399368 SRS399369 SRS399370 SRS399371	Yale_EC_E-01_37 Yale_EC_C-04_22 Yale_EC_A-03_34 Yale_EC_D-04_27 Yale_EC_B-04_28	SRA066203.sample.xml	DownLoad	View
Study				DownLoad	View
Experiment	SRX247412 SRX247413 SRX247414 SRX247415 SRX247416	Yale_EC_E-01_37 Yale_EC_C-04_22 Yale_EC_A-03_34 Yale_EC_D-04_27 Yale_EC_B-04_28	SRA066203.experiment.xml	DownLoad	View
Run	SRR769599 SRR769600 SRR769601 SRR769602 SRR769603	00_pd_ATCACG-GCCAAT-StrE 00_pd_CAGATG-CTTGTA-StrC 00_pd_CTTGTA-CTTGTA-StrA 00_pd_GATCAG-CTTGTA-StrD 00_pd_GATCAG-CAGATC-StrB	SRA066203.run.xml	DownLoad	View

STUDY TITLE

STUDY TYPE

Select your registered query files.

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	SRX247412	SRS399367	SRR769599					ILLUMINA	paired
<input type="checkbox"/> 2	SRX247413	SRS399368	SRR769600					ILLUMINA	paired
<input type="checkbox"/> 3	SRX247414	SRS399369	SRR769601					ILLUMINA	paired
<input type="checkbox"/> 4	SRX247415	SRS399370	SRR769602					ILLUMINA	paired
<input type="checkbox"/> 5	SRX247416	SRS399371	SRR769603					ILLUMINA	paired

: from metadata : Counted from FASTQ (Sequence length is calculated from the first entry.)

NEXT 4

SRA066203 を選択

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

Preprocessing -> 計算対象ランデータを選択 -> NEXT

Preprocessing (クオリティチェック)

DDBJ DNA Data Bank of Japan

ACCOUNT
login ID [orenoddb]

ANALYSIS
Data setup
DRA Start
FTP upload
HTTP upload
DRA Import
Preprocessing Start
step-1
Preprocessing
Mapping /
de novo Assembly
step-2
Workflow
Genome (SNP/Short Indel)
RNA-seq (Tag count)
ChIP-seq

JOB STATUS
step1.
Preprocessing 3
step1.
Mapping
step1.
de novo Assembly
step2-All status

HELP
HELP
TUTORIAL

DDBJ Read Annotation Pipeline.
Development Team.

Set Parameters for Preprocessing

BACK NEXT

Your selected queries

Run ACCESSION	Read length	Quality Score	Read Layout
SRR769599 ->-	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 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.
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 (\geq 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'.

Run Confirmation

BACK RUN

Email notification
Send email notification when the job is completed or aborted with error.
kawano@dbcls.jp * Required

Confirmation of entries
Query sets
• SRR769599 - 00_pd_ATCACG-GCCAT-StrE

確認画面

1 BACK NEXT 2 RUN

パラメータ設定

*ログインしないと実行できません (Runボタンがでできません)

進捗確認画面-Preprocessing

Status - Preprocessing

Mapping Job de novo Assembly Job Preprocessing Job

Order
Sort by : ID Descending Show Only Your Own Job Reload

	ID	User ID	Files	P/S	Status	Read #	Read length	Detail	Start time	End time	Elapsed time
<input type="checkbox"/>	18214	orenoddbj	SRA066203 00_pd_ATCACG	P	complete		--	View	2015-06-14 21:49:02	2015-06-14 22:13:56	00:24:54
<input type="checkbox"/>	18213	---	A203-8	P	complete		--		2015-06-14 16:45:59		03:07:
<input type="checkbox"/>	18209	---	pe	P	complete		--		2015-06-12 17:02:36		01:39:
<input type="checkbox"/>	18201	---	pacific_parents	S	complete		--		2015-06-10 18:16:12		04:06:
<input type="checkbox"/>	18160	---	DRA001581 DRR015979	P	complete		--		2015-06-07 16:56:20		01:21:

Delete * page 1 NEXT >

[Detail view](#) BACK

Job info

- ID: 18214
- Tool (Version): (1.0)
- RunAccession or Filename: SRR769599
- Download: SRR769599_1.fastq.bz2, SRR769599_2.fastq.bz2
- Read length: N.A. bp
- Alias: 00_pd_ATCACG-GCCAAT-StrE

File

File	Fastq Download	QS Average (PDF)	QS Count (PDF)
SRR769599_1.fastq.bz2	download (1.2 GB)	download (8.6 KB)	download (5.2 KB)
SRR769599_2.fastq.bz2	download (1.2 GB)	download (8.6 KB)	download (5.2 KB)

Time 結果

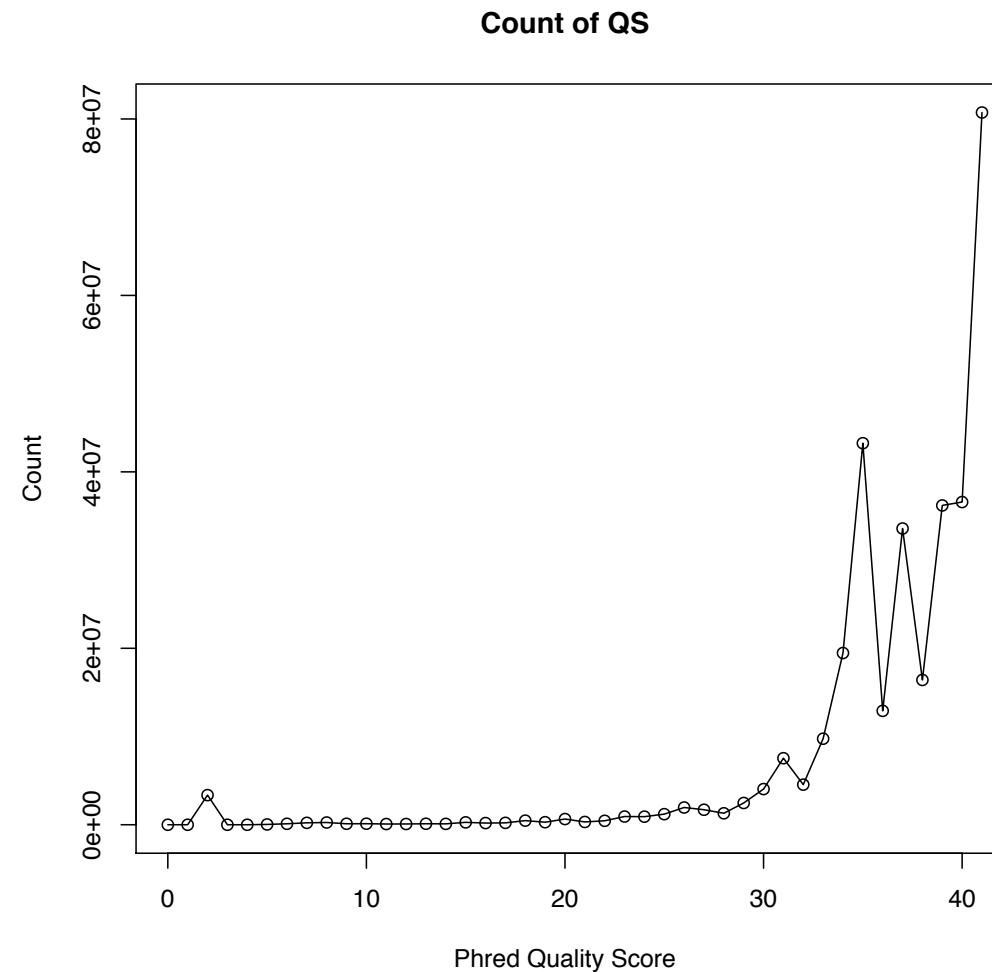
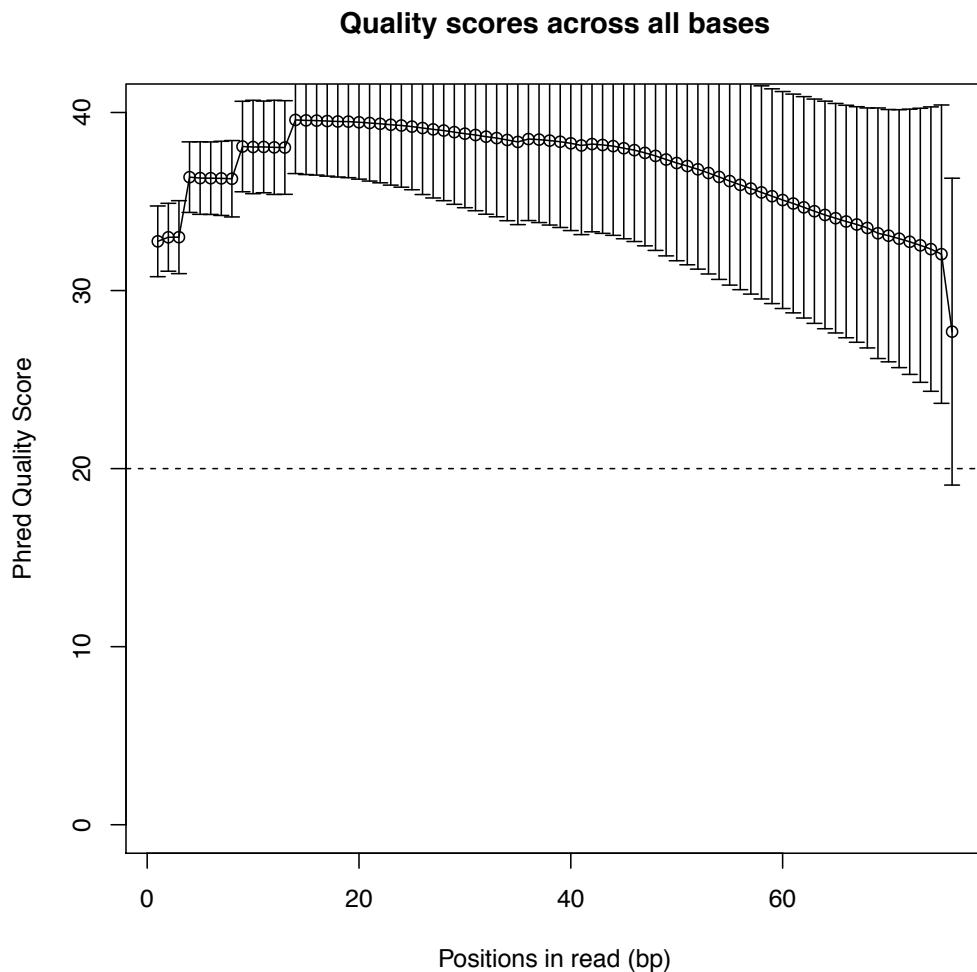
Wait time	Start time	End time
0: 0:55	2015-06-14 21:49:02	2015-06-14 22:13:56

Command

Command	Start time	End time	Log1	Log2	Result	MD5
perl avq_p.pl fqlist.txt qscore	2015-06-14 21:49:04	2015-06-14 22:06:32	View			
perl pdel_p3_t.pl fqlist.txt qscore19 24 1 14 30 33	2015-06-14 22:06:32	2015-06-14 22:11:13				
perl user_fastq_copy.pl preprocessing.xml orenoddbj	2015-06-14 22:11:13	2015-06-14 22:13:56	View			

BACK

Preprocessing 結果



塩基部位ごとのQuality score分布

スコアの分布

どの塩基まで解析に使用するかの目安になる

操作2

配列データを既知のゲノム配列にマッピングする

パイプラインの実行～マッピング

ACCOUNT
login ID [orenoddb]
Logout
Change password

ANALYSIS
Data setup 1
DRA Start (circled)
FTP upload
HTTP upload
DRA Import
Preprocessing Start
step-1
Preprocessing
Mapping /
de novo Assembly
step-2
Workflow
Genome (SNP/Short
Indel)
RNA-seq (Tag count)
ChIP-seq

JOB STATUS
step1.
Preprocessing
step1.
Mapping
step1.
de novo Assembly
step2-All status

HELP
HELP
TUTORIAL
Contact Us.
DDBJ Read Annotation
Pipeline.
Development Team.

Selecting Query Files

FTP upload Private DRA entry Import public DRA Preprocessing HTTP upload

Metadata of the DRA entry. 2

TYPE	ACCESSION	ALIAS	FILENAME	DL	VIEW
Submission	SRA066203	Ecoli_repeat_structure	SRA066203.submission.xml	DownLoad	View
Sample	SRX399367	Yale_EC_E-01_37			
	SRX399368	Yale_EC_C-04_22			
	SRX399369	Yale_EC_A-03_34			
	SRX399370	Yale_EC_D-04_27			
	SRX399371	Yale_EC_B-04_28			
Study	SRX247412	Yale_EC_E-01_37			
	SRX247413	Yale_EC_C-04_22			
	SRX247414	Yale_EC_A-03_34	SRA066203.experiment.xml		
	SRX247415	Yale_EC_D-04_27			
	SRX247416	Yale_EC_B-04_28			
Experiment	SRX247412	Yale_EC_E-01_37			
	SRX247413	Yale_EC_C-04_22			
	SRX247414	Yale_EC_A-03_34	SRA066203.experiment.xml		
	SRX247415	Yale_EC_D-04_27			
	SRX247416	Yale_EC_B-04_28			
Run	SRX247412	Yale_EC_E-01_37			
	SRX247413	Yale_EC_C-04_22			
	SRX247414	Yale_EC_A-03_34	SRA066203.experiment.xml		
	SRX247415	Yale_EC_D-04_27			
	SRX247416	Yale_EC_B-04_28			

STUDY TITLE
STUDY TYPE

Select your registered query files.

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

No.	Experiment ACCESSION	Sample ACCESSION	Run ACCESSION	STRAIN	Run_date	Read #	Read length	Instrument model	Layout
1	SRX247412	SRX399367	SRR769599					ILLUMINA	paired
2	SRX247413	SRX399368	SRR769600					ILLUMINA	paired
3	SRX247414	SRX399369	SRR769601					ILLUMINA	paired
4	SRX247415	SRX399370	SRR769602					ILLUMINA	paired
5	SRX247416	SRX399371	SRR769603					ILLUMINA	paired

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

DELETE NEXT

DRA Startで初期画面に戻る

計算対象ランデータを選択 -> NEXT

SRA066203 を選択

解析ツールの選択～マッピング

マッピング系ツール



Selecting Tools for Basic Analysis of DDBJ ANNOTATION PIPELINE

Reference Genome Mapping

Tool	Help	Version	Input data			Evaluation			Analysis		Output format			Comment
			Base space	Color space	Paired end	Depth	Coverage	Error rate	SNP	Indel	.gff	.bed	SAM	
<input type="checkbox"/> BLAT		34	✓											Single-end analysis only
<input checked="" type="checkbox"/> bwa		0.5.9	✓		✓	✓	✓	✓	✓				✓	
<input type="checkbox"/> Bowtie		0.12.7	✓	✓	✓	✓	✓	✓	✓					✓
<input type="checkbox"/> TopHat		1.0.11	✓		✓	✓	✓	✓						✓
<input type="checkbox"/> Bowtie2		2.0.0	✓	✓	✓	✓	✓	✓	✓	✓				For reads longer than about 50 bp, Bowtie2 is generally faster, more sensitive, and uses less memory than Bowtie1.
<input type="checkbox"/> TopHat2		2.0.9	✓		✓	✓	✓	✓	✓				✓	

de novo Assembly
Total limit = 22 Gbp

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		r2013-02-25	✓		✓		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

BACK **NEXT**

解析するランの指定

Generating Query Sets from Query Read Files

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
1	<input checked="" type="checkbox"/> SRR769599 -><-	bp	

Set as Pair-End

QUERY SET

RESET BACK NEXT

Generating Query Sets from Query Read Files

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	QualityScore2
paired	SRR769599	00_pd_ATCACG-GCCAAT-StrE			

RESET BACK NEXT

リファレンスゲノムの指定

Specifying Database of Reference Genome

RESET BACK NEXT

Major genome sets

Organisms: Arabidopsis thaliana
Genome sets: TAIR8

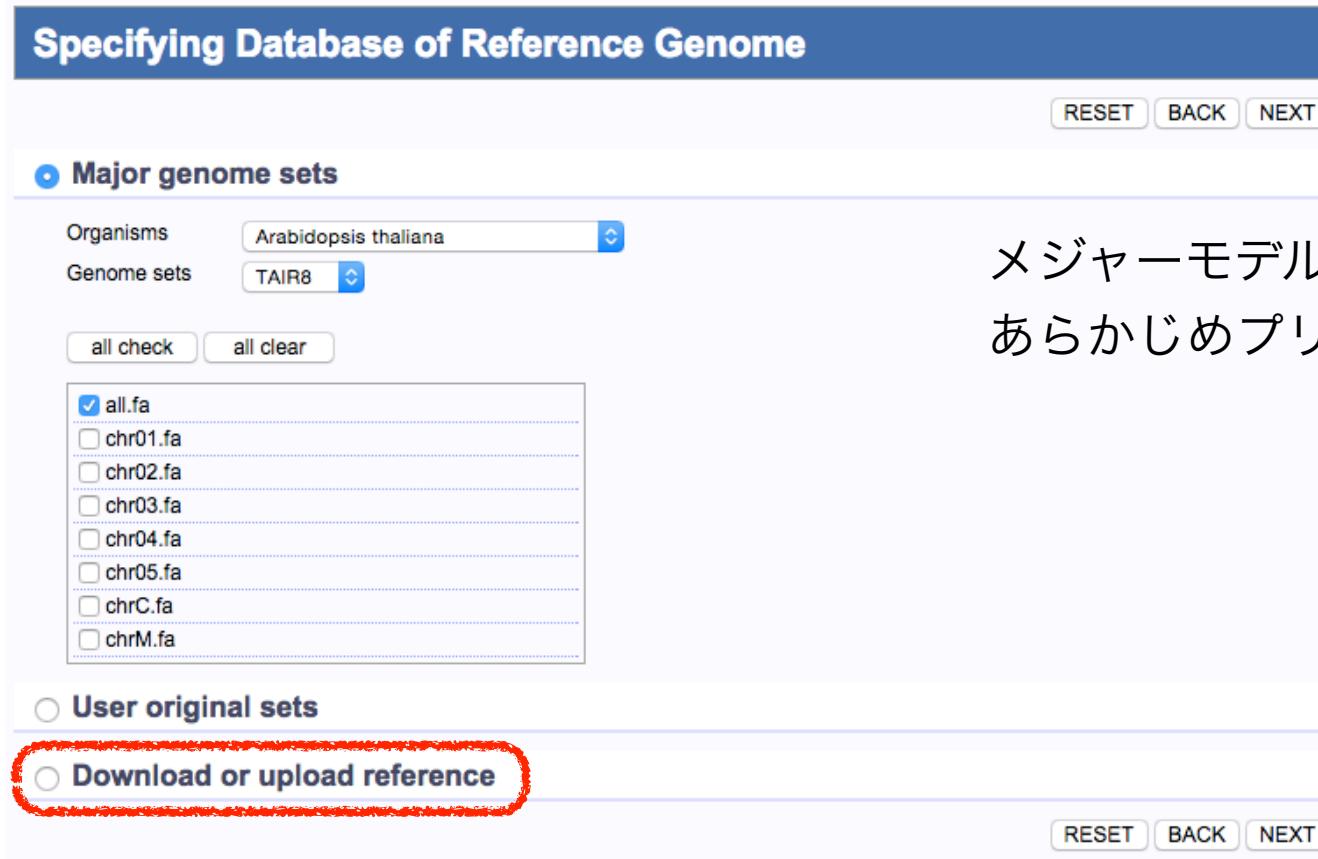
all check all clear

all.fa
 chr01.fa
 chr02.fa
 chr03.fa
 chr04.fa
 chr05.fa
 chrC.fa
 chrM.fa

User original sets

Download or upload reference

RESET BACK NEXT



メジャー モデル生物のリファレンスゲノムはあらかじめプリセットされている

ヒト
マウス
線虫
イネ
シロイヌナズナ
酵母
...

その他のリファレンスゲノムは

- 自分でゲノム配列をアップロードする
- DDBJからインポートする

NCBIページ - 大腸菌のゲノムIDを調べる

ゲノムデータベースを選択して大腸菌を検索

NCBI Resources How To Sign in to NCBI

NCBI National Center for Biotechnology Information

Genome E. coli Search

NCBI Home Resource List (A-Z) All Resources Chemicals & Bioassays Data & Software DNA & RNA Domains & Structures Genes & Expression Genetics & Medicine Genomes & Maps Homology Literature Proteins Sequence Analysis Taxonomy Training & Tutorials Variation

Welcome to NCBI

The National Center for Biotechnology Information advances science and health by providing access to biomedical and genomic information.

About the NCBI | Mission | Organization | NCBI News

Submit Deposit data or manuscripts into NCBI databases

Download Transfer NCBI data to your computer

Learn Find help documents, attend a class or watch a tutorial

Develop Use NCBI APIs and code libraries to build applications

Analyze Identify an NCBI tool for your data analysis task

Research Explore NCBI research and collaborative projects

Popular Resources PubMed Bookshelf PubMed Central PubMed Health BLAST Nucleotide Genome SNP Gene Protein PubChem

NCBI Announcements

April 8th webinar: "The NCBI Minute: Introducing MOLE-BLAST" Mar 25, 2015

April 1st webinar: "A Practical Guide to Using NCBI BLAST on the Web" Mar 24, 2015

dbSNP Build 143 Phase II now available Mar 17, 2015

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

NCBIゲノムページ

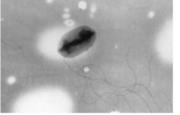
NCBI Resources How To

Genome Genome Search Create alert Limits Advanced

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

Display Settings: Overview

[Organism Overview](#) ; [Genome Assembly and Annotation report \[3071\]](#) ; [Genome Groups](#)

 **Escherichia coli**
A well-studied enteric bacterium

Lineage: Bacteria[6344]; Proteobacteria[2198]; Gammaproteobacteria[953]; Escherichia coli[1]
Escherichia coli. This organism is typically present in the lower intestine of humans, where it is a major constituent of the complete intestinal microflora. E. coli is easily grown in culture and is often used as a model organism for bacterial genetics and molecular biology. It is also a common cause of foodborne illness. [More...](#)

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

Publications

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

[More...](#)

Representative (genome information for reference and representative genomes)

Dendrogram (based on genomic BLAST)

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

Representative (genome information for reference and representative genomes)

Reference genomes: [\[see all organisms\]](#)

- Escherichia coli str. K-12 substr. MG1655
 - Submitter: Univ. Wisconsin
 - Morphology: Gram:Negative, Shape:Bacilli, Motility:Yes
 - Environment: OxygenReq:Facultative, OptimumTemperature:37, TemperatureRange:Mesophilic, Habitat:HostAssociated
- Escherichia coli O157:H7 str. Sakai
 - Submitter: GIRC
 - Human Pathogen
 - Morphology: Gram:Negative, Shape:Bacilli, Motility:Yes
 - Environment: OxygenReq:Facultative, OptimumTemperature:37, TemperatureRange:Mesophilic, Habitat:HostAssociated
 - Phenotype: Disease:Hemorrhagic colitis
- Escherichia coli IAI39
 - Submitter: Genoscope
 - Human Pathogen
 - Morphology: Gram:Negative, Shape:Bacilli, Motility:No
 - Environment: OxygenReq:Facultative, TemperatureRange:Mesophilic, Habitat:Multiple
 - Phenotype: Disease:Urinary tract infection

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

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

リファレンスゲノムデータのインポート

Specifying Database of Reference Genome

RESET BACK NEXT

Major genome sets
 User original sets
 Download or upload reference

Retrieving a chromosome from DDBJ-DB by using HTTP REST

Input Accession Number (INSD) or (RefseqID)
U00096.3
LOAD

PIPELINE INTERNET DDBJ-DB
Request HTTP REST * Data (fasta)
* Representational State Transfer(REST)

Uploading reference from local drive.
FASTA only 選択... ファイルが選択されていません。 UPLOAD
2GB Filesize Limit

RESET BACK NEXT

Escherichia coli のINSD IDを
→ 入力、 LOAD
→ CREATE DATASET
→ CREATE GENOMESET
でリファレンスゲノムを登録する

Download or upload reference

Retrieving a chromosome from DDBJ-DB by using HTTP REST

Input Accession Number (INSD) or (RefseqID)
CP000075
LOAD

PIPELINE INTERNET DDBJ-DB
Request HTTP REST * Data (fasta)
* Representational State Transfer(REST)

Uploading reference from local drive.
FASTA only 選択... ファイルが選択されていません。 UPLOAD
2GB Filesize Limit

CREATE DATASET

RESET BACK NEXT

Create Genome Dataset

files
✓ >U00096|U00096.3 Escherichia coli str. K-12 substr. MG1655, complete genome.

Please input a genomeset description.

Genome Dataset name >U00096|U00096.3 Escherichia coli str. K-12 substr. MG1655, complete genome.
BACK CREATE GENOMESET

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

リファレンスゲノムの選択

Specifying Database of Reference Genome

Major genome sets

User original sets

Genome sets

U00096.3 Escherichia coli str. K-12 substr. MG1655, complete genome.

>U00096|U00096.3 Escherichia coli str. K-12 substr. MG1655, complete genome.

Download or upload reference

RESET BACK NEXT

先ほど登録したゲノムを選択

パラメータ設定、確認、実行

Setting for Reference Genome Mapping

BACK NEXT

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)

Step2) Map

bwa aln -t 4 ↗ refgenome.fasta query1.fastq(.fasta) > out1.sai

bwa aln -t 4 ↗ refgenome.fasta query2.fastq(.fasta) > out2.sai

bwa sampe ↗ refgenome.fasta in1.sai in2.sai query1.fastq(.fasta) 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 ↗ refgenome.fasta out.bam | bcftools view
- samtools mpileup -u -C50 -BQ0 -d10000000 ↗ refgenome.fasta out.bam | bcftools view -bvcg - > 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

Run Confirmation

BACK RUN

Destination of mail

When the request is completed, the system sends an email to this address.

kawano@dbcls.jp  * Required

Result files will be deleted 60 days after submission.

Reference Genome Map [bwa]

Query sets

Query set1

PairedOrientation	RunAccession	RunAlias	RowLength	QualityScore1	QualityScore2
paired	SRR769599	00_pd_ATCACG-GCCAAT-StrE			

genome sets

U00096.3 Escherichia coli str. K-12 substr. MG1655, complete genome.

• >U00096|U00096.3 Escherichia coli str. K-12 substr. MG1655, complete genome.



計算に時間がかかるのでメールアドレスの登録は必須
ログインしないと実行できません
(Runボタンが出てきません)

進捗確認画面-Mapping

The screenshot shows the DDBJ Status - Mapping page. On the left, there's a sidebar with sections for ACCOUNT (login ID [orenoddb]), ANALYSIS (Data setup, DRA Start, FTP upload, HTTP upload, DRA Import, Preprocessing Start), and JOB STATUS (step1: Preprocessing, step1: Mapping, step1: de novo Assembly, step2: All status). The main area is titled 'Status - Mapping' and has tabs for Mapping Job, de novo Assembly Job, and Preprocessing Job. It shows a table of jobs with columns: Order, ID, UserID, Submission accession, P/S, Status, Tool, Read #, Read length, Genome size, Detail, Start time End time, and Elapsed time. The 'View' button in the Detail column for the first job is circled in red.

ID	UserID	Submission accession	P/S	Status	Tool	Read #	Read length	Genome size	Detail	Start time	End time	Elapsed time
18215	orenoddbj	SRA066203 00_pd_ATCAGG	P	complete	bwa	4,248,803	---	4 M	View	2015-06-14 22:07:15	2015-06-14 22:55:40	00:48:25
18212	---	---	S	complete	bwa	63,640,652	---	469 M		2015-06-13 00:26:41	2015-06-13 09:02:35	08:35:54
18211	---	TWMU1	P	complete	bwa	35,099,181	---	3,197 M		2015-06-12 22:47:18	2015-06-13 10:45:30	11:58:12
18210	---	No_Nipponbare	S	complete	bwa	40,145	---	390 M		2015-06-12 17:02:36	2015-06-12 17:23:09	00:20:32

背景無色（白）が他のユーザ
背景黄色が自分のジョブ

マッピング結果

Detail view

BACK

Job info

ID 18215	Tool (Version) bwa (0.5.9)		
RunAccession or Filename SRR769599	Download SRR769599_1.fastq.gz (<small>Original size 1.2 GB</small>) SRR769599_2.fastq.gz (<small>Original size 1.2 GB</small>)	Read length N.A. bp	Alias 00_pd_ATCACG-GCCAAT-StrE

Genome set

Chromosome
[U00096.3_150614220106489](#)

Download modified queries

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

Download merged pileup file

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

- [merged.var.vcf.gz](#) (Original size 9.6 MB)
- [merged.sam.gz](#) (Original size 1.3 GB)

Download wgs file

- [out_WGS.fasta.gz](#) (Original size 1.5 MB)

Position errors

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

Time

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

結果ファイル

統計量

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

BACK

中間ファイル

マッピング結果のビューア

◆ IGV

○ <https://www.broadinstitute.org/igv/>

○ windows, Mac OSX, Linux

○ Tablet

○ <http://ics.hutton.ac.uk/tablet/>

○ windows, Mac OSX, Linux

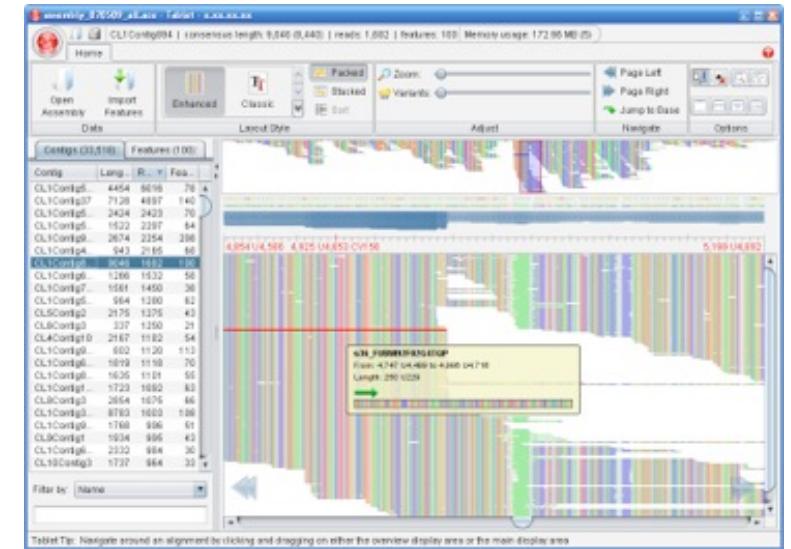
◆ samtools tview

○ <http://www.htslib.org/>

○ Mac OSX, Linux



...



操作3

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

解析ツールの選択～de novo アセンブリ

マッピング系ツール



Selecting Tools for Basic Analysis of DDBJ ANNOTATION PIPELINE

BACK NEXT

Reference Genome Mapping

	Tool	Help	Version	Base space	Color space	Paired end	Depth	Coverage	Error rate	SNP	Indel	.gff	.bed	SAM	Comment
<input type="checkbox"/>	BLAT		34	✓					✓					Single-end analysis only	
<input type="checkbox"/>	bwa		0.5.9	✓			✓	✓	✓	✓				✓	
<input type="checkbox"/>	Bowtie		0.12.7	✓	✓	✓	✓	✓	✓	✓	✓			✓	
<input type="checkbox"/>	TopHat		1.0.11	✓			✓	✓	✓	✓				✓	
<input type="checkbox"/>	Bowtie2		2.0.0	✓	✓	✓	✓	✓	✓	✓	✓			For reads longer than about 50 bp, Bowtie2 is generally faster, more sensitive, and uses less memory than Bowtie1.	
<input type="checkbox"/>	TopHat2		2.0.9	✓			✓	✓	✓	✓				✓	

de novo Assembly

Total limit = 22 Gbp

	Tool	Help	Version	Base space	Color space	Paired-end	MSS(WGS)	Comment
<input checked="" 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		r2013-02-25	✓		✓		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

BACK NEXT

解析するランの指定

Generating Query Sets from Query Read Files

Paired-end analysis

Layout of paired sequence. 5'-3' 3'-5'

Run	ACCESSION	Read length	Quality Score
1	<input checked="" type="checkbox"/> SRR769599 -><-	bp	

2

QUERY SET

RESET **BACK** **NEXT**

Generating Query Sets from Query Read Files

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 ACCESION Read length Quality Score

[Set as Pair-End](#)

QUERY SET
Query set1

PairedOrientation	RunAccession	RunAlias	RowLength	QualityScore1	QualityScore2
paired	SRR769599	00_pd_ATCACG-GCCAAT-StrE			

[RESET](#) [BACK](#) [NEXT](#) 3

パラメータ設定、確認、実行

Setting for De Novo Assembly

soapdenovo

Set optional parameters of the paired-end analysis

Memory Usage : Low (recommended) High
If you request "High" memory usage during the time Nig super computer system is congested, you might be kept waiting long before job starts running.

Step1) Make SOAP denovo configuration file

Please select a file format.
FASTQ file : -fq

* Please select FASTQ or FASTA correctly.

Please input average of insert size from all query files.

Please input maximum of read length from all query files.

perl make_config_file.pl query1.fastq query2.fastq

Step2) Assembly

[Optional] You can input a soapdenovo custom options. (Some option is limited.)
soapdenovo all -s config_file

Step3) Create assembled sequences in FASTA file from pileupped reads to submit WGS division of DDBJ.

Set filtered length for contigs
 perl lengthfilter.pl pileupFile out_WGS.txt

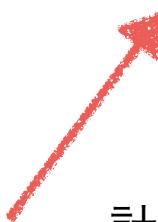
Run Confirmation

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

Assembly [soapdenovo]

Query sets

Query set1	PairedOrientation	RunAccession	RunAlias	RowLength	QualityScore1	QualityScore2
	paired	SRR769599	00_pd_ATCACG-GCCAAT-StrE			



計算に時間がかかるのでメールアドレスの登録は必須
ログインしないと実行できません
(Runボタンが出てきません)

進捗確認画面-Assembly

ACCOUNT
login ID [orenoddb]
Logout Change password

ANALYSIS
Data setup
DRA Start
FTP upload
HTTP upload
DRA Import
Preprocessing Start
step-1
Preprocessing
Mapping /
de novo Assembly
step-2
Workflow
Genome (SNP/Short Indel)
RNA-seq (Tag count)
ChIP-seq

JOB STATUS
step1.
Preprocessing
step1.
Mapping
step1.
de novo Assembly
step2-All status

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

Running Status

Status - de novo Assembly

Mapping Job de novo Assembly Job Preprocessing Job

Order											
Sort by : ID Descending Show Only Your Own Job Reload											
Delete * page 1 NEXT >											
ID	UserID	Submission accession	P/S	Status	Tool	Read #	Read length	Assembly detail	Mapping detail	Start time	Elapsed time
18216	orenoddb	SRA066203 00_pd_ATCACG	P	complete	SOAPdenovo	4,248,803	---	View		2015-06-14 22:14:13	00:07:36
18207	---	DRA001581 DRR015979	P	complete	ABySS	11,236,293	---			2015-06-12 14:26:19	00:29:46
18204	---	SRA049915 Heinz10_rep1 Heinz10_rep2	S	complete	Trinity	27,600,469	---			2015-06-12 08:04:09	01:23:59
18200	---	SA13-0319_1 SA13-0319_2 SA13-0319_3	S	complete	HGAP		---			2015-06-10 17:50:46	04:27:11

背景無色（白）が他のユーザ
背景黄色が自分のジョブ

アセンブリ結果

The screenshot shows the DDBJ Analysis tool interface for a sequencing project. The left sidebar contains navigation links for ACCOUNT, ANALYSIS, and JOB STATUS. The main area displays the 'Detail view' of the assembly process.

Job info:

ID	18216
Tool (Version)	SOAPdenovo (1.05)

RunAccession or Filename | **Download** | **Read length** | **Alias**

SRR769599	SRR769599_1.fastq.gz SRR769599_2.fastq.gz	N.A. bp	00_pd_ATCACG-GCCAAT-StrE
-----------	--	---------	--------------------------

Download modified queries

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

Download wgs file

- [out_WGS.fasta.gz](#) (Original size 5.1 MB)

Assembly statistics:

Contig # : 47,468
Total contig size : 7,291,528
Maximum contig size : 5,773
Minimum contig size : 24
N50 contig size : 596

Time:

Wait time	Start time	End time
0: 0:57	2015-06-14 22:14:13	2015-06-14 22:21:50

Command: SOAPdenovo127mer all -s soapdenovo.conf -o output

Start time	End time	Log1	Log2	Result	MD5
2015-06-14 22:14:13	2015-06-14 22:20:44	View		Download(123.5 MB)	MD5

Red boxes highlight the 'Workflow' section, the 'Download wgs file' section, and the assembly statistics table. Red text annotations '結果ファイル' (Results file) and '統計量' (Statistics) point to the highlighted areas.

応用1

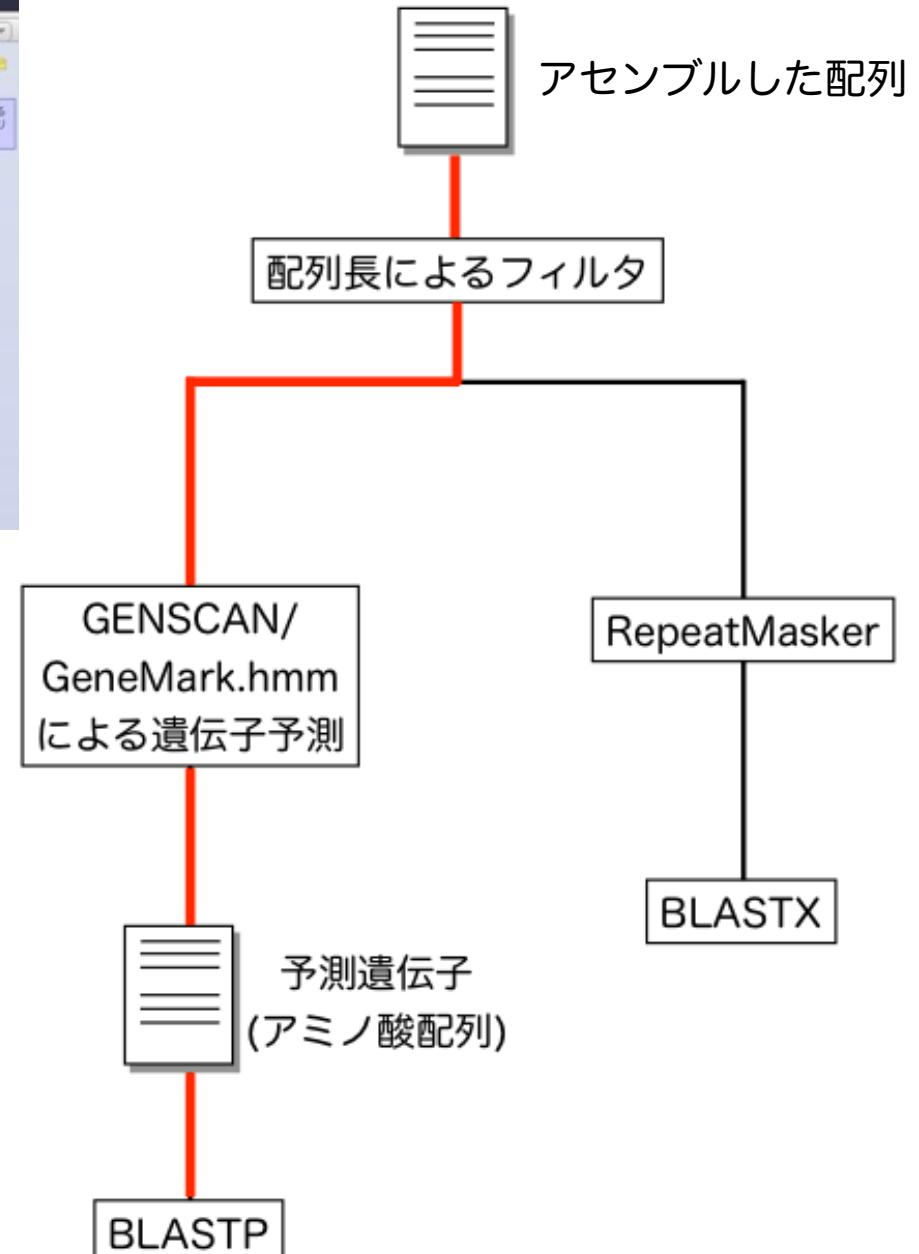
p-Galaxyを使った配列アノテーション

p-Galaxyを使った配列アノテーション

The screenshot shows the p-Galaxy web interface. At the top, there's a navigation bar with 'Galaxy / DDBJ' and links for 'Analyze Data', 'Workflow', 'Shared Data', 'P-galaxy Manual', 'Help', and 'User'. On the left, a sidebar lists various tools and analyses, including 'Work Flow', 'Get Data', 'Send Data', 'ENCODE Tools', 'Lift-Over', 'Text Manipulation', 'Filter and Sort', 'Join, Subtract and Group', 'Convert Formats', 'Extract Features', 'Fetch Sequences', 'Fetch Alignments', 'Get Genomic Scores', 'Operate on Genomic Intervals', 'Statistics', 'Wavelet Analysis', 'Graph/Display Data', 'Regional Variation', 'Multiple regression', 'Multivariate Analysis', 'Evolution', 'Metagenomic analyses', 'FASTA manipulation', 'NGS: QC and manipulation', 'NGS: Mapping', 'NGS: Indel Analysis', and 'NGS: Expression Analysis'. A central panel displays a workflow titled 'WWFSMD?' with the sub-instruction 'grow noodly appendages...'. Below the workflow is the URL 'usegalaxy.org'. A green banner at the top says 'Hello world! It's running...' and 'To customize this page edit static/welcome.html'. A history panel on the right shows an 'Unnamed history' entry.

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

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



Galaxyとは？

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

- DBCLS Galaxy

- テキスト系ツール

- p-Galaxy

- 次世代解析ツール



55

p-Galaxyにログイン

The screenshot shows the Galaxy / P-GALAXY web interface. At the top, there is a navigation bar with links for Analyze Data, Workflow, Shared Data, Visualization, P-galaxy Manual, Help, and User (with a dropdown menu). A red box highlights the "User" dropdown, and a red arrow points from it to a separate "Login" page.

The main content area displays a green banner with a checkmark icon and the text "Hello world! It's running..." and "HOSOMICHI HLA ANALYSIS TOOLS of Hosomichi et al. (2013) have been opened to public!!". Below the banner, there is a complex workflow diagram titled "WWFSMD? grow noodly appendages..." with various nodes like Input dataset, Filter, Join, Sort, Group, and Join two Queries. The "usegalaxy.org" logo is visible below the diagram.

The "Login" page on the right has fields for "Email address" (containing "kawano@dbcls.rois.ac.jp") and "Password" (containing masked text). It also includes links for "Forgot password? Reset here" and a "Login" button.

At the bottom of the main page, there is a note: "This project is supported in part by NSF, NHGRI, and the Huck Institutes of the Life Sciences."

pipeline計算結果のデータインポート

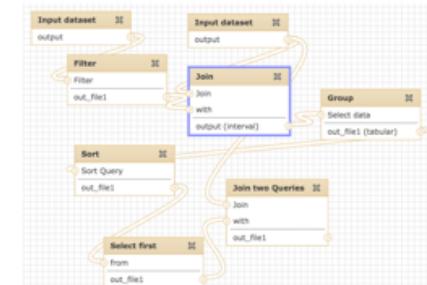
Galaxy / P-GALAXY Analyze Data Workflow Shared Data Visualization P-galaxy Manual Help User Using 0 bytes

Tools search tools Work Flow COMMON PROCESS

[import samfile](#) Import from DDBJ Pipeline
[import pileupfile](#) Import from DDBJ Pipeline
[import mpileupfile](#) Import from DDBJ Pipeline
[import contig fasta file](#) Import from DDBJ Pipeline (highlighted)
[uniq BWA sam](#)
[Quality check of fastq data](#)
[Cut off fastq single data](#)
[Cut off fastq paired data](#)
ANNOTATION FOR DNA POLYMORPHISM
[Detect SNPs](#)
[Detect Indels](#)
[Merge SNPs / InDels data files](#)

Hello world! It's running...
HOSOMICHI HLA ANALYSIS TOOLS of Hosomichi et al. (2013) have been opened to public!!

WWFSMD?
grow noodly appendages...



usegalaxy.org

This project is supported in part by [NSF](#), [NHGRI](#), and [the Huck Institutes of the Life Sciences](#).

History

データのインポート

Galaxy / P-GALAXY

Analyze Data Workflow Shared Data Visualization

ツール

search tools

Work Flow

COMMON PROCESS

[import samfile](#) Import from DDBJ Pipeline

[import pileupfile](#) Import from DDBJ Pipeline

Login ID [orenoddbj]

Import Contig FASTA from basic analysis. By DDBJ Read Annotation Pipeline

JOBID	Submission Accession RunAlias	Tool	Pipeline Jobpage	Import to Galaxy
18216	SRA066203 00_pd_ATCACG-GCCAAT-StrE	soapdenovo	ViewJob	Import
18161	ERA013525 ZAP430 run	soapdenovo	ViewJob	Import

Galaxy / P-GALAXY

Analyze Data Workflow Shared Data Visualization P-galaxy Manual Help User Using 23.6 MB

ツール

ANNOTATION FOR DE NOVO ASSEMBLED SEQ.

[FASTA File Length Filter](#)

Gene Prediction Choose GENSCAN or GeneMark.hmm

Gene Prediction to FASTA Converts GENSCAN or GeneMark.hmm output file to FASTA format

[transcriptsToOrfs](#) Trinity Transcripts to Candidate Peptides

[RepeatMasker](#)

[BLASTP](#)

[BLASTX](#)

PHYLOGENETIC ANALYSIS

[sam_to_fasta](#) for get mapping fasta data

HOSOMICHI HLA ANALYSIS

[Trim By Quality](#)

[Picks Up Fine Pairs From Paired Read Set.](#)

This dataset is large and only the first megabyte is shown below.
[Show all](#) | [Save](#)

```
>1 length 24 cvg_0.0_tip_0
TATTTTTTTTTTTTTTTTTTTTT
>3 length 24 cvg_0.0_tip_0
GTTGGGGTATGCCAACCGGTAAAG
>5 length 24 cvg_0.0_tip_0
CGCGATTGAACGTTCGCTGTGTA
>7 length 24 cvg_0.0_tip_0
CGGTAGGCCGATAAGGCCTTCAC
>9 length 24 cvg_0.0_tip_0
ATTCCGCCAGTTCTTCTTCAGCC
>11 length 24 cvg_0.0_tip_0
CCCTATGCCACCATTGGCGGGGT
>13 length 24 cvg_0.0_tip_0
CTGTCGCGTCTTATCAGGCCCTACA
>15 length 24 cvg_0.0_tip_0
CCCTTTCCCGGCCATAAAGGGCAG
>17 length 24 cvg_0.0_tip_0
TGACCAAGTTCATGTTGCTTAACCG
>19 length 24 cvg_0.0_tip_0
TGGGTGTTAGCTCAGTTGGTAGA
>21 length 24 cvg_0.0_tip_0
GGAGATCGCGAAATCGGGCAAAT
>23 length 24 cvg_0.0_tip_0
CCCCGCCAAAAACCTGGGAAGAGAT
>25 length 24 cvg_0.0_tip_0
ACCTGGGGGATTTTATGGGCTCTC
```

ヒストリー

search datasets

OSAKA

1 shown

8.5 MB

1: import contig fasta file

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

配列長でのフィルタリング

Galaxy / P-GALAXY

ツール

ANNOTATION FOR DE NOVO ASSEMBLED SEQ.

FASTA File Length Filter

Gene Prediction Choose GENSCAN or GeneMark.hmm

Gene Prediction to FASTA Converts GENSCAN or GeneMark.hmm output file to FASTA format

transcriptsToOrfs Trinity Transcripts to Candidate Peptides

FASTA File Length Filter (Galaxy Tool Version 1.0.0)

Input FASTA File
1: import contig fasta file

Base length of data removing from input file
500

The sequence data that length is the same as or less than this value is removed.

Execute

ヒストリー

search datasets

OSAKA 1 shown 8.5 MB

1: import contig fasta file

Galaxy / P-GALAXY

ツール

ANNOTATION FOR DE NOVO ASSEMBLED SEQ.

FASTA File Length Filter

Gene Prediction Choose GENSCAN or GeneMark.hmm

Gene Prediction to FASTA Converts GENSCAN or GeneMark.hmm output file to FASTA format

transcriptsToOrfs Trinity Transcripts to Candidate Peptides

RepeatMasker

BLASTP

BLASTX

PHYLOGENETIC ANALYSIS

sam to fasta for get mapping fasta data

This dataset is large and only the first megabyte is shown below.
Show all | Save

>87941 TGGCGCTTCTCAAAGATGAGTTGTTATCGCGACAAATTACCGACTGCCAGTTAGTCAGTCCCGAGAGTTGCGTCGC

>87943 CGGTTTCACTTCAGCATAGGCCGTATCAGATAACGGGTGTCGGGTCATGGCTGCGATCTCCTTAGTGCCT

>87945 CGCAACGGCGCGATGTTCATCGTTAATGATCACCGATTCTGCTGGGTGGACGTACCTGGCTATGCCGCG

>87947 GCGTGAGCACTGGTGGATTATCTGGCATCTCTCACAAACAGGAGGAAGCAGGTAAAAAAACTCATCACAGCGAT

>87949 AAATAGTGGTCTGAAACGGATGCGCATTAAAGCGTGCAGTCATCCGTCGGCGCATTAAATATGATTAGTT

>87951 GGTCAAGATGGAAACGGGATCGAAAATCGGCATACCAATGACATCGCGTACAATCGCAAACCATGAAGCATCG

>87953 CCAGCATAATCGGGATTGTCGCAATGACCGGAAATATCCGCGCCACGATAATGCCCAATGCCGACATA

ヒストリー

search datasets

OSAKA 2 shown 12.2 MB

2: FASTA File Length Filter on data 1

1: import contig fasta file

遺伝子予測

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

The screenshot shows the Galaxy / P-GALAXY interface. In the center, there is a 'Gene Prediction' tool configuration window. The 'Tool Version 1.0.0' dropdown is set to 'GENSCAN'. The 'Parameter file' dropdown is set to 'human/vertebrate (also Drosophila)'. The 'Sequence file' dropdown shows '2: FASTA File Length Filter on data 1'. A large red arrow points to the 'Execute' button at the bottom left of the configuration window.

The screenshot shows the Galaxy / P-GALAXY interface after the execution of the gene prediction. The results are displayed in a table and text sections. The 'Predicted peptide sequence(s)' section contains the sequence: >879471GENSCAN_predicted_peptide_11122_aa MSVVLRASITTPEARADLLGFTITECDESIPVTASVPASASADKTESQRIRETIIAQI PEGQFTESLVAQLMEKVMKEKQSLEQGALQPSFKSVTGGIKVIDGSSVKGRFDGAQP HC. The 'Predicted coding sequence(s)' section contains the sequence: >879471GENSCAN_predicted_CDS_1|366_bp atgtccgttgttctcgccgcacgttattaccggaaagccgtgaatgtggctgtatcg tttgggttttaccatcaccgaatgtatcgatccggtaacggcgctgttcccgcc agcgcacatcgccataaaaccgaaaggccgcacgttccgtgaaaccattatcgcccaactc cccggaaaggccatcgatccggaaaggccgcacgttccgtgaaaccattatcgcccaactc aacatcgccatcgccaaaggccgcacgttccgtgaaaccattatcgcccaactc ggcacatcgccatcgccaaaggccgcacgttccgtgaaaccattatcgcccaactc cactgc GENSCAN 1.0 Date run: 14-Jun-115 Time: 22:45:14 Sequence 87949 : 502 bp : 51.59% C+G : Isochore 1 (0 - 100 C+G%) Parameter matrix: Maize.smat Predicted genes/exons:

予測配列の取り出し

Galaxy / P-GALAXY

ツール

- ANNOTATION FOR DE NOVO ASSEMBLED SEQ.
- FASTA File Length Filter
- Gene Prediction Choose GENSCAN or GeneMark.hmm
- Gene Prediction to FASTA Converts GENSCAN or GeneMark.hmm output file to FASTA format
- transcriptsToOrfs Trinity Transcripts to Candidate Peptides
- RepeatMasker
- BLASTP
- BLASTX
- PHYLOGENETIC ANALYSIS

Gene Prediction to FASTA Converts GENSCAN or GeneMark.hmm output file to FASTA format (Galaxy Tool Version 1.0.0)

Input Gene Prediction File
7: Gene Prediction on data 2

Select extracting sequences
Peptide sequences

Execute

ヒストリー

- search datasets
- OSAKA
- 5 shown, 2 deleted
- 15.5 MB
- 7: Gene Prediction on data 2
- 4: Gene Prediction to FASTA on data 3
- 3: Gene Prediction on data 2
- 2: FASTA File Length Filter on data 1
- 1: import contig fasta file

Galaxy / P-GALAXY

ツール

- ANNOTATION FOR DE NOVO ASSEMBLED SEQ.
- FASTA File Length Filter
- Gene Prediction Choose GENSCAN or GeneMark.hmm
- Gene Prediction to FASTA Converts GENSCAN or GeneMark.hmm output file to FASTA format
- transcriptsToOrfs Trinity Transcripts to Candidate Peptides
- RepeatMasker
- BLASTP
- BLASTX
- PHYLOGENETIC ANALYSIS

>87947|GENSCAN_predicted_peptide_1|122_aa
MSVVRASITTEPAREVADLLGFTITECDESIPVTASVPASASADKTESQRIRETIIAQLE
>87949|GENSCAN_predicted_peptide_1|58_aa
MFRQLLTQTERSFEIFIAFFQTGLEIGKLQFAFTYQITDAVKRHAAVTDDTPPTTIX
>87953|GENSCAN_predicted_peptide_1|151_aa
MTVAGFSLPIVLLSALGTGILLAGLWNGILVAILKIQPFVATLILMVAGRGAQLITSQIV
>87955|GENSCAN_predicted_peptide_1|45_aa
XNYKGTLLIDGKEFDNSYTRGEPLSFRLDGVIPGWTEGLKNIKKGX
>87957|GENSCAN_predicted_peptide_1|55_aa
MAREGELSLNYRDNDQYRYYASVQLFPWLETTLRYTDVRTRQYSSVEAFSGDQX
>87959|GENSCAN_predicted_peptide_1|20_aa
MLRQLPPHGQYARYGEHNL
>87961|GENSCAN_predicted_peptide_1|21_aa
MRQGVVDHQWITRGFTACAX
>87963|GENSCAN_predicted_peptide_1|158_aa
MGRTYHEDNRSPGDLPGTKTQMITSKTYKGSGFNLRFEDATDKEQVYIHAQKNMDTEV
>87965|GENSCAN_predicted_peptide_1|86_aa
GLTVTEVKGFGRQKGHAEYLRYGAESVNLPKVKIDVAIADDQLDEVIDIVSKAAYTGKIC
>87981|GENSCAN_predicted_peptide_1|26_aa
MAYLVAPPLEYGIDAALKSADVQL
>87983|GENSCAN_predicted_peptide_1|5_aa
MPTPM
>87989|GENSCAN_predicted_peptide_1|76_aa
MNDDPDTAVLKPKAPLILDPNKAQBLIKADATHLALLTSPPARTPRDVLSAEEVCTDAEAM

ヒストリー

- search datasets
- OSAKA
- 6 shown, 2 deleted
- 16.0 MB
- 8: Gene Prediction to FASTA on data 7
- 7: Gene Prediction on data 2
- 4: Gene Prediction to FASTA on data 3
- 3: Gene Prediction on data 2
- 2: FASTA File Length Filter on data 1
- 1: import contig fasta file

BLASTによる配列の機能アノテーション

Galaxy / P-GALAXY Data Workflow Shared Data Visualization P-galaxy Manual Help User Using 31.1 MB

ツール
ANNOTATION FOR DE NOVO ASSEMBLED SEQ.
FASTA File Length Filter
Gene Prediction Choose GENSCAN or GeneMark.hmm
Gene Prediction to FASTA Converts GENSCAN or GeneMark.hmm output file to FASTA format
transcriptsToOrfs Trinity Transcripts to Candidate Peptides
RepeatMasker
BLASTP
BLASTX
PHYLOGENETIC ANALYSIS sam_to_fasta for get mapping fasta data

BLASTP (Galaxy Tool Version 1.0.0)
Query file 8: Gene Prediction to FASTA on data 7
Select database Swiss-Prot-Bacteria
Expectation value 0.00001 ex 1.0.00001 or -20 (as e-20)
Execute

ヒストリー search datasets
OSAKA 6 shown, 2 deleted 16.0 MB
8: Gene Prediction to FASTA on data 7
7: Gene Prediction on data 2
4: Gene Prediction to FASTA on data 3
3: Gene Prediction on

Galaxy / P-GALAXY Analyze Data Workflow Shared Data Visualization P-galaxy Manual Help User Using 191.8 MB

ツール
ANNOTATION FOR DE NOVO ASSEMBLED SEQ.
FASTA File Length Filter
Gene Prediction Choose GENSCAN or GeneMark.hmm
Gene Prediction to FASTA Converts GENSCAN or GeneMark.hmm output file to FASTA format
transcriptsToOrfs Trinity Transcripts to Candidate Peptides
RepeatMasker
BLASTP
BLASTX
PHYLOGENETIC ANALYSIS sam_to_fasta for get mapping fasta data
HOSOMICHI HLA ANALYSIS Trim By Quality
Picks Up Fine Pairs From Paired

ヒストリー search datasets
OSAKA 8 shown, 2 deleted 176.7 MB
10: BLASTP error/warning reports
9: BLASTP on data 8
8: Gene Prediction to FASTA on data 7 データを表示
7: Gene Prediction on data 2
4: Gene Prediction to FASTA on data 3
3: Gene Prediction on data 2
2: FASTA File Length Filter on data 1
1: import contig fasta file

Sequences producing significant alignments:

	Score	E
spIP765551EUTQ_ECOLI Ethanolamine utilization protein EutQ OS=Es...	213	2e-69
spIQ9ZFV51EUTQ_SALTY Ethanolamine utilization protein EutQ OS=Sa...	192	4e-61

>spIP765551EUTQ_ECOLI Ethanolamine utilization protein EutQ
OS=Escherichia coli (strain K12) GN=eutQ PE=4 SV=1
Length = 233

Score = 213 bits (543), Expect = 2e-69, Method: Compositional matrix adjust.
Identities = 106/122 (86%), Positives = 108/122 (88%)

Query: 1 MSVVLRASIITPEAREVADELLGFTITECDEXXXXXXXXXXXXXDKTESQRIRETIIAQ 60
MSVVLRASIITPEAREVADELLGFTITECDE DKTESQRIRETIIAQ
Sbjct: 21 MSVVLRASIITPEAREVADELLGFTITECDESIPVTASVPASVPADKTESQRIRETIIAQ 80

Query: 61 PEGQFTESLVAQLMEKVMKEKQSLEQGALQPSFKSVTGKGGIKVIDGSSVKFGRFDGAQP 120
PEGQFTESLVAQLMEKVMKEKQSLEQGA+QPSFKSVTGKGGIKVIDGSSVKFGRFDGA+P
Sbjct: 81 PEGQFTESLVAQLMEKVMKEKQSLEQGAMQPSFKSVTGKGGIKVIDGSSVKFGRFDGAEP 140

Query: 121 HC 122
HC
Sbjct: 141 HC 142

>spIQ9ZFV51EUTQ_SALTY Ethanolamine utilization protein EutQ
OS=Salmonella typhimurium (strain LT2 / SGSC1412 / ATCC 13048) PE=1 SV=1
Length = 233

(cc) BY ©2015 統合データベース講習会 Licensed Under CC 表示 2.1 62

DDBJ Pipeline 参考資料

◆ 詳しくはDDBJing講習会で

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

The screenshot shows the DDBJ website's main navigation bar at the top, featuring the DDBJ logo, a search bar, and links for "DDBJ の紹介", "利用の手引き", "レポート・統計", "FAQ", and "お問い合わせ". Below the navigation is a breadcrumb trail: HOME > 利用の手引き > DDBJing 講習会 > DDBJing講習会 アーカイブ. To the right, it says "最終更新日 : 2015.1.8.". A large black header box contains the text "DDBJing講習会 アーカイブ". Below this, a message encourages users to check past DDBJing workshops on YouTube and Slideshare. It also notes that the content is updated periodically and advises contacting the workshop if materials are used or reused. A section titled "「第30回 DDBJing 講習会 in 東京」 2014年12月18日" provides details about the 30th workshop, including its theme ("NGS大規模配列解析とDDBJへの登録") and location ("独立行政法人 科学技術振興機構 東京本部 アクセス"). At the bottom, a table lists seven presentations from the workshop, each with a speaker's name, title, and links to YouTube and Slideshare.

タイトル	講師	資料
DDBJ と NIG SuperComputer の紹介、大量配列情報解析	中村 保一（国立遺伝学研究所 教授）	YouTube Slideshare
BioProject, BioSample, DDBJ Sequence Read Archive の紹介	児玉 悠一（DDBJ 構築チーム アノテータ）	YouTube Slideshare
Mass Submission System の紹介	李 慶範（DDBJ 構築チーム アノテータ）	YouTube Slideshare
DDBJ Read Annotation Pipeline の紹介と実習	長崎 英樹（国立遺伝学研究所 大量遺伝研究室 研究員）	YouTube Slideshare
Japanese Genotype-phenotype Archive の紹介	児玉 悠一（DDBJ 構築チーム アノテータ）	YouTube Slideshare
メタゲノム解析と微生物統合データベース	森 宙史（東京工業大学大学院生命理工学研究科 助教）	YouTube Slideshare

※以降はご覧になりたい回をクリックすると内容が表示されます。



©2015 統合データベース講習会 Licensed Under CC 表示 2.1

応用2

MiGAPを使った配列アノテーション

MiGAP

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

MiGAPウェブサイト

<http://www.migap.org/>

The screenshot shows the MiGAP website homepage. The top navigation bar includes links for フォーラム (Forum), ホーム (Home), ヘルプ (Help), 管理情報 (Management Information), and コンタクト (Contact). The main content area features several maintenance notices:

- 遺伝研スパコンの定期メンテナンスに伴うMiGAPサービス停止のお知らせ** (Notice of MiGAP service suspension due to regular maintenance of the Genetic Research Supercomputer)
 - Date: 2015年5月13日(水曜日) 17:21 | Author: 齊藤 仁浩 | [Edit](#) | [Delete](#) | [Print](#)
 - Message: 以下の日程で国立立遺伝学研究所内スパコンの定期メンテナンスが実施される予定です。
 - Maintenance Period: 2015年7月8日(水)9:00 ~ 7月11日(土)24:00
 - Note: メンテナンスに伴い、MiGAPのサービスが停止となりますのでご協力ををお願い致します。
 - MIGAPへの投入受付: 2015年7月6日 (月) 17:00
 - Note: なお、実行待ちジョブが多数ある場合は、ジョブ投入受付停止時刻を早める場合があります。
 - Remaining Job强制終了時刻: 2015年7月7日 (火)12:00
 - Note: サービス再開予定は以下の通りです。
 - Job Submission Reception Resumption预定時刻: 2015年7月13日 (月) 10:00
- MiGAPサービスを再開いたします** (MiGAP service has been restored)
 - Date: 2015年4月28日(火曜日) 18:01 | Author: 齊藤 仁浩 | [Edit](#) | [Delete](#) | [Print](#)
 - Message: メンテナンスの為に停止しておりましたMiGAPのサービスを再開いたします。
- トラブル再発による発生緊急メンテナンスのおしらせ** (Notice of emergency maintenance due to recurring trouble)
 - Date: 2015年4月28日(火曜日) 14:58 | Author: 齊藤 仁浩 | [Edit](#) | [Delete](#) | [Print](#)
 - Message: 昨日に発生したトラブルの再発により、緊急メンテナンスを行っております。
ご利用の皆様にはご不便、ご迷惑をお掛けいたします。

Left sidebar menu items include:

- トップメニュー (Top Menu):
 - MiGAPについて (About MiGAP)
 - ヘルプ (Help)
 - お知らせ (Announcements)
 - バイオラインについて (About BioLine)
 - MiGAPサーバの運用主体 (MiGAP server operator)
 - MiGAP引用・関連文献リスト (Citations and related literature)
 - 謝辞 (Acknowledgments)
- バイオラインにログイン (Log in to BioLine)
- 旧バイオラインにログイン (Log in to old BioLine) (highlighted with a red box)
- 利用登録する (Register)
- アカウントを取得する (Get account)
- パスワードを変更する (Change password)
- パスワード再発行を依頼する (Request password reset)

Right sidebar menu items include:

- 最新ニュース (Latest News):
 - 遺伝研スパコンの定期メンテナンスに伴うMiGAPサービス停止のお知らせ (Notice of MiGAP service suspension due to regular maintenance of the Genetic Research Supercomputer)
 - MiGAPサービスを再開いたします (MiGAP service has been restored)
 - トラブル再発による発生緊急メンテナンスのおしらせ (Notice of emergency maintenance due to recurring trouble)
 - MiGAPサービスを再開しました (MiGAP service has been opened)
 - 緊急メンテナンスによるサービスの停止のお知らせ (Notice of service suspension due to emergency maintenance)
 - MiGAPサービス再開しました (MiGAP service has been opened)
 - 緊急メンテナンスによるサービスの停止のお知らせ (Notice of service suspension due to emergency maintenance)
 - MiGAPサービス再開しました (MiGAP service has been opened)
 - MiGAP投入受付停止のお知らせ (Notice of job submission reception stop)
 - MiGAPサービス再開しました (MiGAP service has been opened)
- 閲覧ランキング (Viewing ranking)



ユーザレベルの設定

The screenshot shows the MiGAP interface with a sidebar on the left containing links: Pipe Line, Pipe Line History, Change (highlighted with a red box), User Level (highlighted with a red box), Current, and Process. To the right, a 'Change User Level' section is displayed with three radio button options: b-MiGAP (selected), s-MiGAP, and g-MiGAP. A 'Set' button is located below the radio buttons.

◆ b-MiGAP

- ツール・パラメータ固定（ブロンズ）

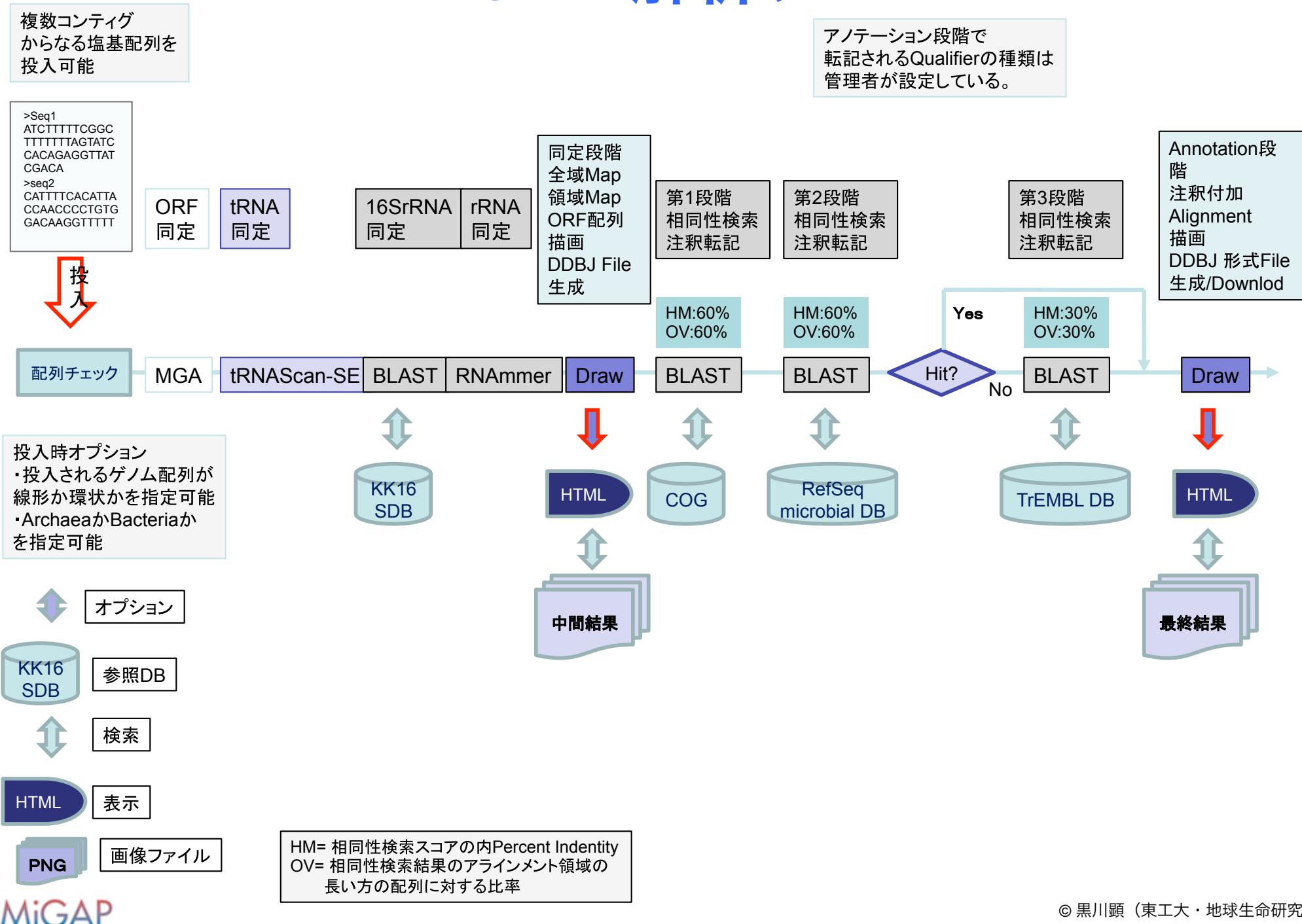
◆ s-MiGAP

- パラメータ設定が可能（シルバー）

◆ g-MiGAP

- DBやツール、パラメータ設定が可能（ゴールド）

b-MiGAP 解析フロー



MiGAPデータ入力

MiGAP Microbial Genome Annotation Pipeline ver2.19

Logout Help Contact Us

Pipe Line

Input Pipe Line [Running:1 Waiting:0]

Pipe Line Name:

Upload Filename: ファイルが選択されていません。 ← アセンブル済みファイルをアップロード
or paste data in box below. ([Sample data](#)) ← ボックスに入力

or
User Level
Current Process

Linear Circular Bacteria Archaea Eukarya *FUNGI* transl_table: 11

Run Clear

MiGAP結果

MiGAP Microbial Genome Annotation Pipeline ver2.19

Logout Help Contact Us

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

Pipe Line History

Pipe Line History Data List Contig

Change User Level

Current Process

History

LDAP_orenomigap | Hidden [1/1] contig

<< < > >>

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

Basic Information

Hide

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

ORF & RNA Extract

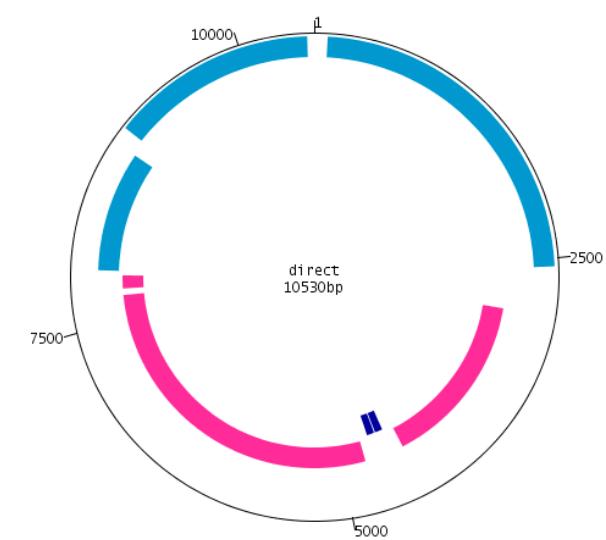
Start: 2015/06/14 17:58:12
End: 2015/06/14 17:58:23
Software: MetaGeneAnnotator 1.0
tRNAscan-SE 1.23
NCBI BLAST 2.2.18
RNAmmer 1.2

Annotation

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

Download

Log File [pipeline.log](#)
N.A.: [result-na.fasta.tar.gz](#)
A.A.: [result-aa.fasta.tar.gz](#)



MiGAP参考資料

- 統合TV

- ✓ 微生物ゲノムアノテーションツールMiGAP

- ▶ 開発者である黒川先生による講習会の動画
 - <http://youtu.be/ujxl6LJlbUE>
 - <http://togotv.dbcls.jp/20131024.html>

- ✓ MiGAPの使い方～導入と基本操作

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

参考サイト

◆ 統合TV



- <http://togotv.dbcls.jp/index.rb?category=NGS>

◆ NGS Surfer's wiki



- <https://cell-innovation.nig.ac.jp/wiki/tiki-index.php>

◆ Q and A



- ライフサイエンスQA (日本語)、BioStar、SEQanswers

- <http://qa.lifesciencedb.jp/>
- <https://www.biostars.org/>
- <http://seqanswers.com/>

◦ (Rで)塩基配列解析

- http://www.iu.a.u-tokyo.ac.jp/~kadota/r_seq.html