

DDBJ Read Annotation Pipeline の紹介と実習

(RNA-Seq配列の*de novo*アセンブリを中心に)

国立遺伝学研究所
大量遺伝情報研究室
長崎 英樹

長崎は遺伝研 大量遺伝情報研究室の所属です。

国立遺伝学研究所



生命情報研究センター



欧州EBIと米国NCBIと密接に協力しながら
DDBJ/EMBL/GenBank国際塩基配列データ
ベースを構築しています。

私たちは

塩基配列登録を支援するシステムづくり

登録データの活用するシステムづくり

高速シークエンス配列の情報解析

を行なっています。

高速シークンサー配列の登場で短期間、低成本で大量の塩基配列データを出力されるようになった。



illumina社 HiSeq2000



Life Technologies社 ion torrent



Pacific Bioscience社 PacBio RS II

▶ Illumina Sequencing Technology - YouTube

https://www.youtube.com/watch?v=womKfikWlxM

YouTube JP

Cluster Generation
Bridge amplification

Illumina sequencing

作者: Ilya Flyamer
再生回数 147,045 回

3:04

1st, 2nd, and 3rd Generation Genome Sequencing Technologies

作者: ImGenTechWPI
再生回数 27,044 回

14:40

Intro to Sequencing by Synthesis: Industry-leading Data Quality

作者: Illumina Inc
再生回数 5,393 回

4:23

Polymerase Chain Reaction (PCR) | MIT 7.01SC Fundamentals of

作者: MIT OpenCourseWare
再生回数 70,043 回

8:35

How to sequence the human genome - Mark J. Kiel

作者: TED-Ed
再生回数 90,421 回

5:05

Agarose Gel Electrophoresis, DNA Sequencing, PCR, Excerpt 2 | MIT

作者: MIT OpenCourseWare
再生回数 30,725 回

42:27

Next-Generation Sequencing Technologies - Elaine Mardis (2012)

作者: GenomeTV

Illumina Sequencing Technology

Illumina Inc

チャンネル登録 1,174

48,743

1,51 / 5:03

||

+ 追加 < 共有 ... その他

200 6

This screenshot shows a YouTube search results page for "Illumina Sequencing Technology". The main video thumbnail on the left illustrates the process of cluster generation and bridge amplification. To the right, there are several recommended videos from the Illumina channel, including topics like 1st, 2nd, and 3rd generation sequencing technologies, sequencing by synthesis, PCR, agarose gel electrophoresis, and next-generation sequencing technologies. The video player at the bottom indicates the current video is at 1:51 of 5:03.

高速シークンサー配列の登場で短期間、低成本で大量の塩基配列データを出力されるようになった。



illumina社 HiSeq2000



Life Technologies社 ion torrent



Pacific Bioscience社 PacBio RS II

その結果...

データ保管場所の確保

計算機不足

解析のための人員不足

といった問題がでてきた。

DDBJの高速シーケンサー配列の諸問題への対応

データ保管場所
の確保

DDBJ Sequence
Read Archive (DRA)

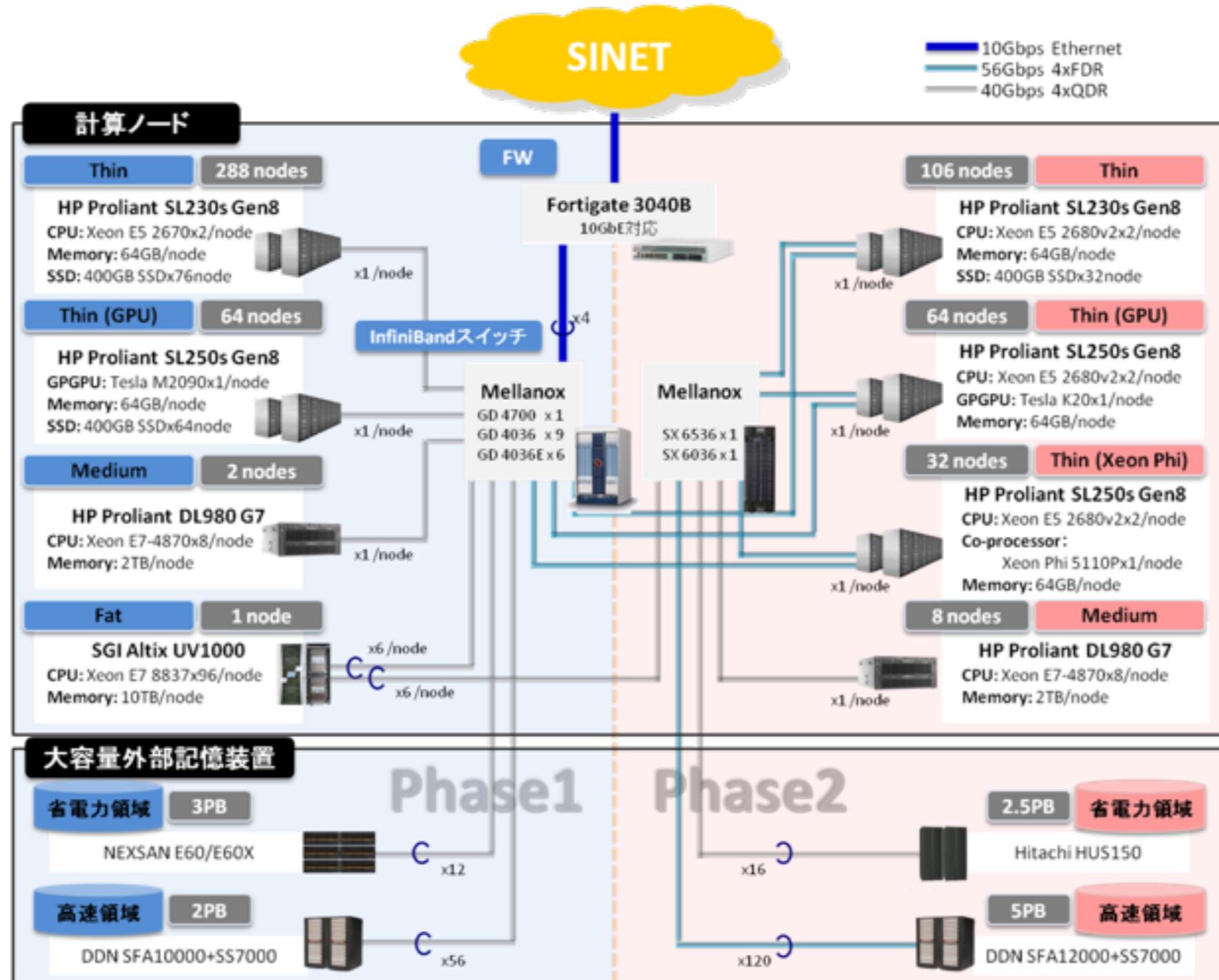
計算機不足

NIG(遺伝研)スパコンシステム

解析のための
人員不足



遺伝研スーパーコンピュータシステム(NIGスパコンシステム)



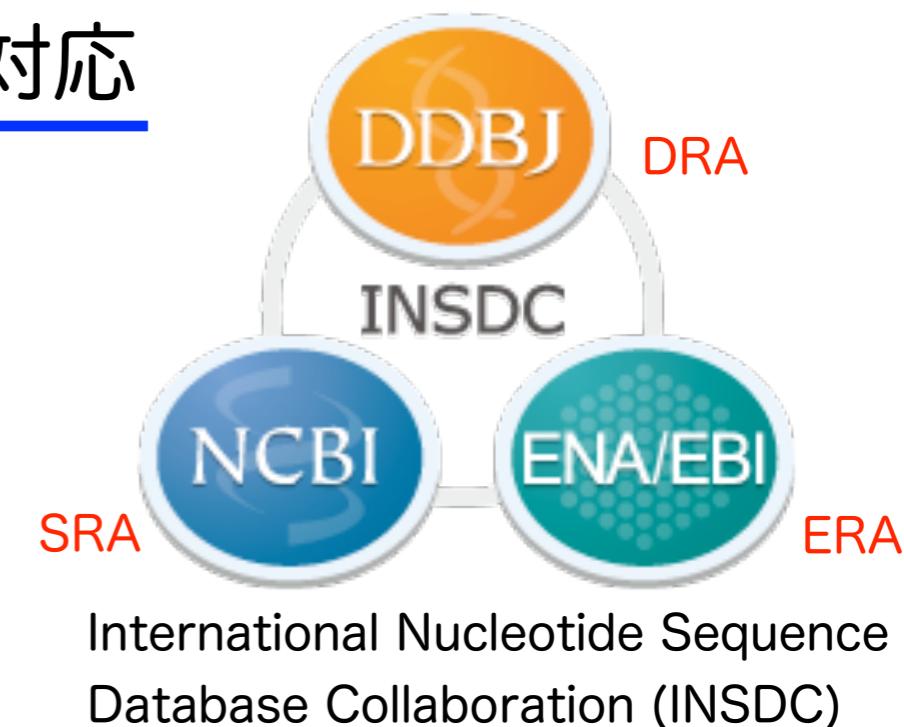
ゲノム解析を主な目的とした大規模計算機利用拠点として 最新鋭の大規模クラスタ型計算機、大規模メモリ共有型型計算機、および大容量高速ディスク装置で構成されたスーパーコンピューティングシステムサービスを提供しています。

<http://sc.ddbj.nig.ac.jp/index.php>

DDBJの高速シークエンサー配列の諸問題への対応

データ保管場所
の確保

DDBJ Sequence
Read Archive (DRA)



計算機不足

NIG(遺伝研)スパコンシステム

- アカウント登録で無償利用
- ✗ コマンドラインによる操作
- ✗ データ規模や使用メモリ量等で計算機ノードを選択などコツがいる。

解析のための
人員不足

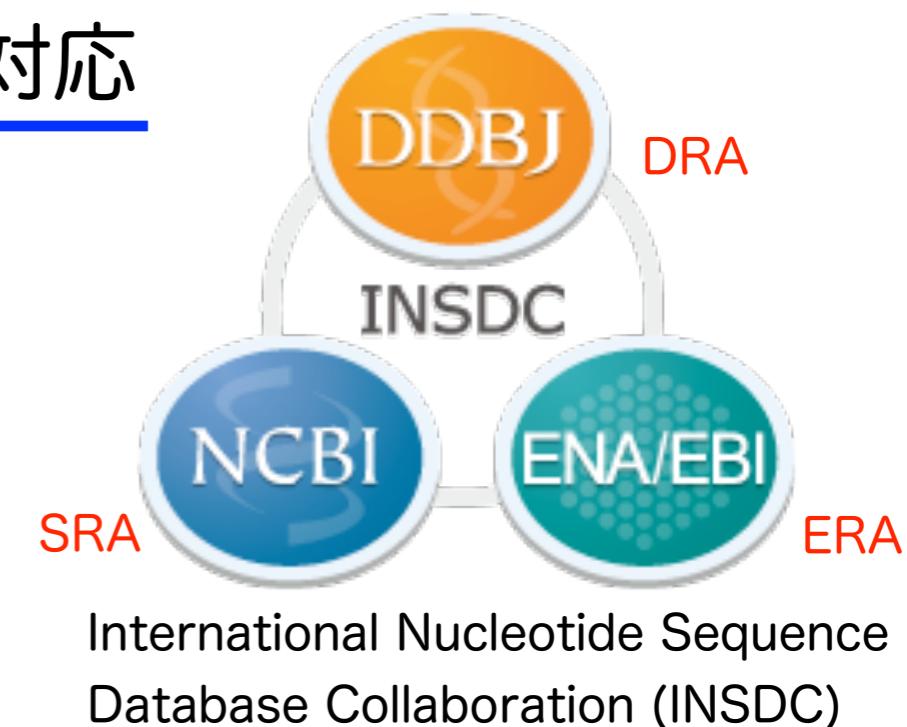
DDBJの高速シークエンサー配列の諸問題への対応

データ保管場所
の確保

DDBJ Sequence
Read Archive (DRA)

計算機不足

解析のための
人員不足

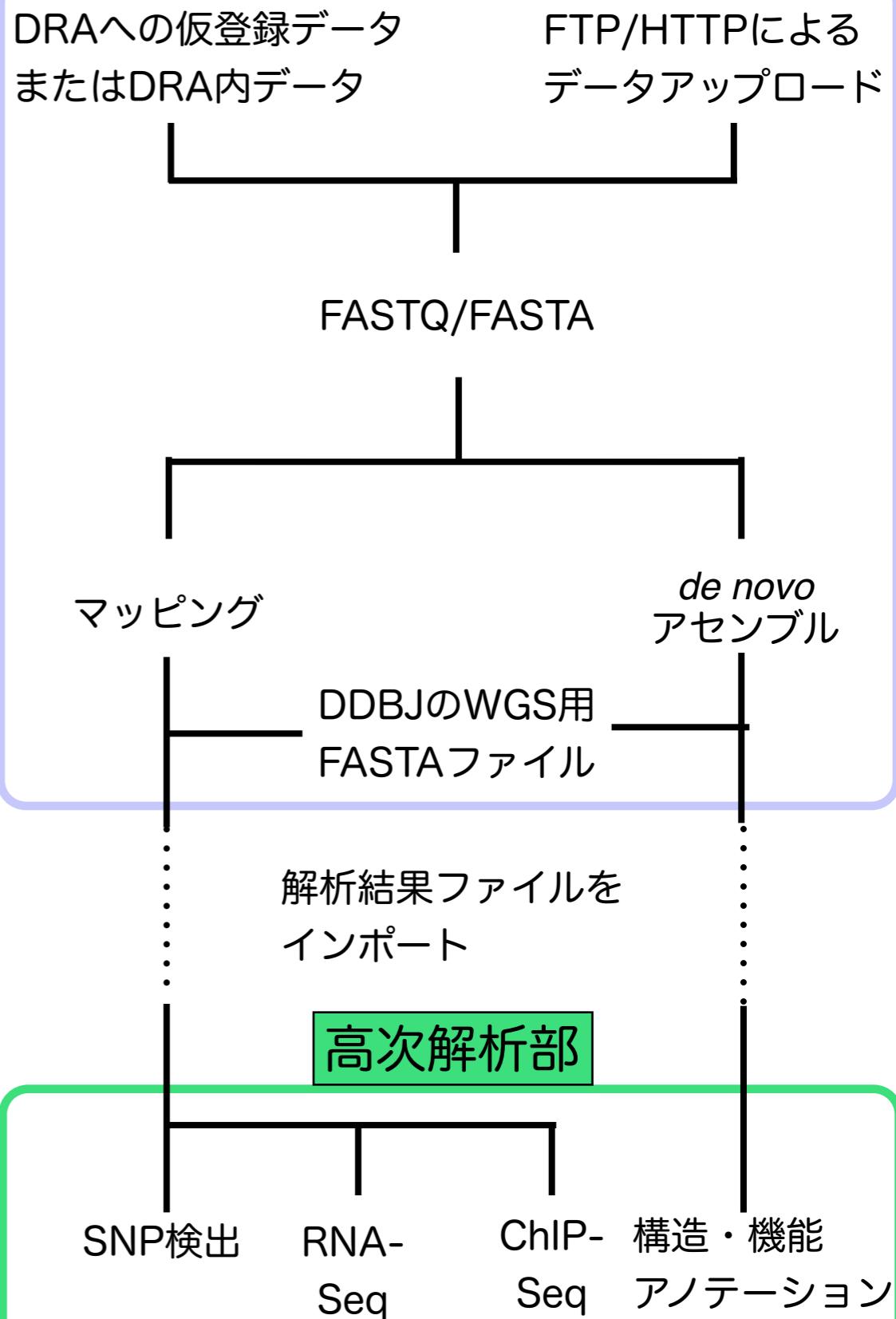


DDBJ Read Annotation Pipeline
(DDBJ パイプライン)

DDBJ パイプラインの特徴

基礎解析部

- ・遺伝研の計算機で分散処理を実行、高速シークンスデータを解析するクラウド型パイプライン
- ・オンラインで無償で利用可。
- ・**基礎解析部** (マッピング、*de novo* アセンブル)と
高次解析部 (構造・機能のアノテーション)で構成



DDBJ Read Annotation Pipeline

<https://sso.ddbj.nig.ac.jp/opensso/UI/Login?realm=ddbj&goto=https://p.ddbj.nig.ac.jp/>

DDBJ DNA Data Bank of Japan

DDBJ Read Annotation Pipeline

>>Japanese

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

LOGIN

New account - Login as "guest".

User ID:
Password:

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

Manual & tutorial

- [Japanese manual](#)
- [English manual](#)
- [FTP client Manual. \(JP\)](#)
- [DBCLS togotv Tutorial video 1 \(JP\) - Reference Genome Mapping](#)
- [DBCLS togotv Tutorial video 2 \(JP\) - De novo Assembly](#)

System usage of DDBJ Read Annotation Pipeline

- [Display statistics of monthly pipeline's web access.](#)
- [Display figures of pipeline's disc usage.](#)
- [Display statistics of monthly pipeline's job.](#)

Account registration of "DRA"

DRA account registration information [please see the page.](#)

pipeline_info

We changed the accessible letter number of login passwords of DDBJ Pipeline from 9 to 16.
17 days ago · reply · retweet · favorite

!!!!Notice!!!! 'de novo Assembly' is out of service. ('Reference Genome Mapping' is available.)
19 days ago · reply · retweet · favorite

DDBJ pipeline is now available : system maintenance was finished.
27 days ago · reply · retweet · favorite

!!!!Notice!!!! DDBJ pipeline services will not be available due to system

[twitter](#) [Join the conversation](#)

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.

- 13種類のマッピング・アセンブルソフト対応
マッピング

BLAT	高速シーケンサー登場以前からあるアライメントツール。 発現データはイントロンを想定したギャップを考慮。
MAQ	高速シーケンサー登場初期にショートリードに対応。 リード長が長くなるに従い開発はBWAに引き継がれる。
BWA	MAQより速く、Titaniumのリードもオプションで対応。
SOAP	メモリ消費量少なく、より高速、精度はBWAより弱冠落ちる。
Bowtie/ Bowtie2	ギャップは考慮しないが処理は速い。BWA、SOAP2、BowtieはBurrows-Wheeler変換というアルゴリズムでゲノムDNAにたいしてインデックスを作成、高速でマッピングする。Bowtie2は50bp以上に最適化。
TopHat	RNA-Seqのリードを内部でBowtieを利用してマッピング、スプライスジャンクションを特定する。

アセンブル

SOAPdenovo	ヒト、パンダ等大型ゲノムのアセンブリで使用された。比較的高速。
Abyss	初期に並列処理に対応したアセンブリ。
Velvet	高速シーケンサー登場初期に開発された。メモリ消費多め。
Trinity	RNA-Seq配列のアセンブリ。上記3つともにde bruijn graphというアルゴリズムを使用。

The screenshot shows a web-based tool selection interface for DDBJ's annotation pipeline. It includes sections for 'Reference Genome Mapping' and 'de novo Assembly'. The 'Reference Genome Mapping' section lists tools like BLAT, MAQ, SOAP, Bowtie, and TopHat, each with a brief description and a table of features. The 'de novo Assembly' section lists tools like SOAPdenovo, ABySS, Velvet, and Trinity, also with descriptions and feature tables. A note indicates a total limit of 22 Gbp for assembly.

東工大のPlatanusとPacBioデータ用のアセンブリ、HGAPも追加されました!

- 13種類のマッピング・アセンブルソフト対応

Selecting Tools for Basic Analysis of DDBJ ANNOTATION PIPELINE

Reference Genome Mapping

Tool	Help	Version	Input data			Evaluation			Analysis	Output format			
			Base space	Color space	Paired end	Depth	Coverage	Error rate		SNP	Indel	.gff	.bed
BLAT	34		✓						✓				Single-end analysis only
Maq	0.7.1		✓		✓				✓	✓	✓	✓	
bwa	0.5.9		✓		✓				✓				
SOAP	2.21		✓		✓				✓	✓	✓		
Bowtie	0.12.7		✓	✓	✓				✓	✓			
TopHat	1.0.11		✓		✓				✓				

de novo Assembly
Total limit = 22 Gbp

Tool	Help	Version	Base space	Color space	Paired-end	MSS(WGS)	Comment
SOAPdenovo	1.05				✓		
ABySS	1.3.2				✓		Maximum K-mer value is 64.
Velvet	1.2.03				✓	✓	We severe recommend when performing Velvet, total length of those reads is up to 22G bp. Maximum K-mer value is 64.
Trinity	r2012-06-08				✓		RNA-Seq De novo Assembly

Mapping Contigs by de novo Assemble to Reference Sequences.
The contigs will be aligned to reference genome.

Tool	Comment
BLAT	Single-end analysis only

BACK NEXT

- 公開配列データの活用が容易

公開データと比較、レファレンスとしての活用

- 13種類のマッピング・アセンブルソフト対応

- 公開配列データの活用が容易

公開データと比較、レファレンスとしての活用

- ジョブステータスで実行状態を確認可能

NIGスパコンで実行
マッピング

Intel Xeon 2.60GHz 16 core, 64GB RAM * 352 nodes

アセンブル

Intel Xeon 2.40GHz 80 cores, 2TB RAM * 2 nodes

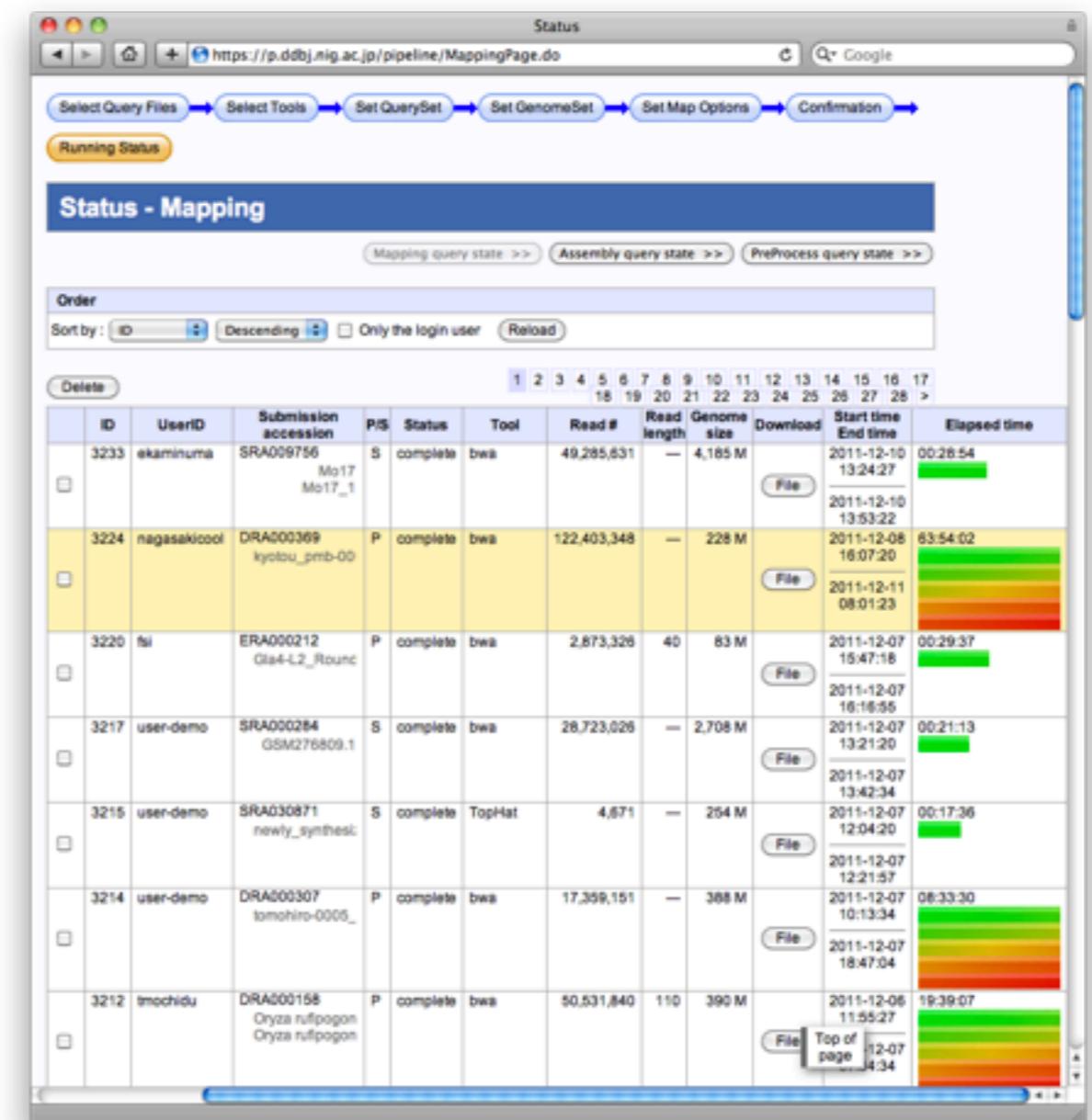
Intel Xeon 2.66GHz 768 cores, 10TB RAM

ストレージ

2PB storage

解析終了をメールで通知

- SAMtools/FASTAによる共通フォーマットでの出力



基礎解析部

DRAへの仮登録データ
またはDRA内データ

FTP/HTTPによる
データアップロード

FASTQ/FASTA

マッピング

de novo
アセンブル

DDBJのWGS用
FASTAファイル

解析結果ファイルを
インポート

高次解析部

SNP検出

RNA-
SeqChIP-
Seq

構造・機能
アノテーション



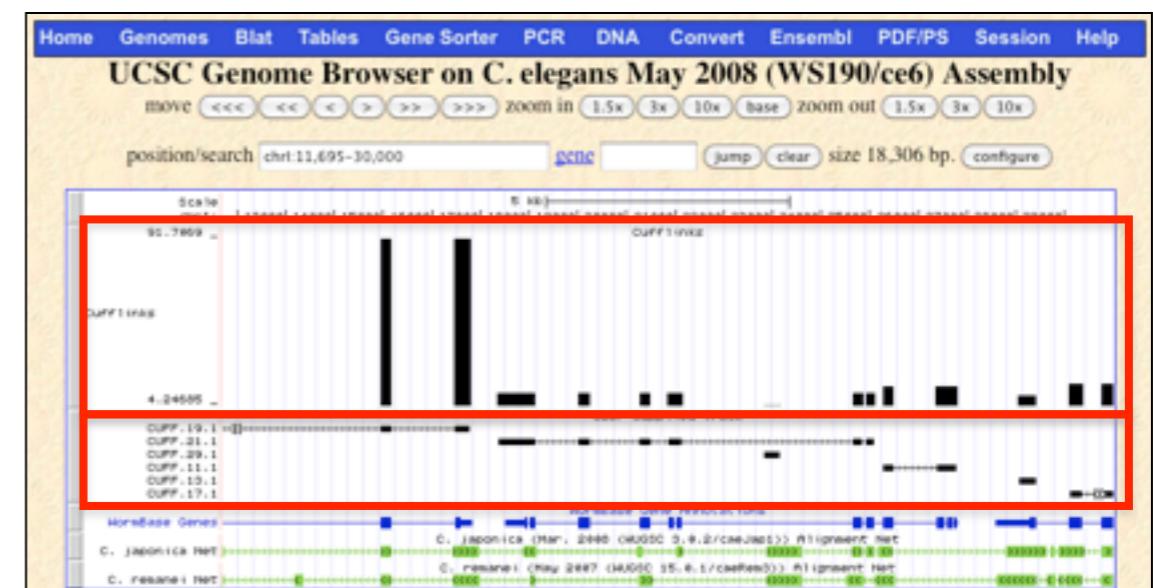
- Galaxyで多様な構造・機能のアノテーションに対応
- 基礎解析部のデータファイルを活用
(SAMや(m)pileup、FASTAファイルを参照)

NGSデータのマッピング結果の解析

SNPのゲノム上の分布の表示

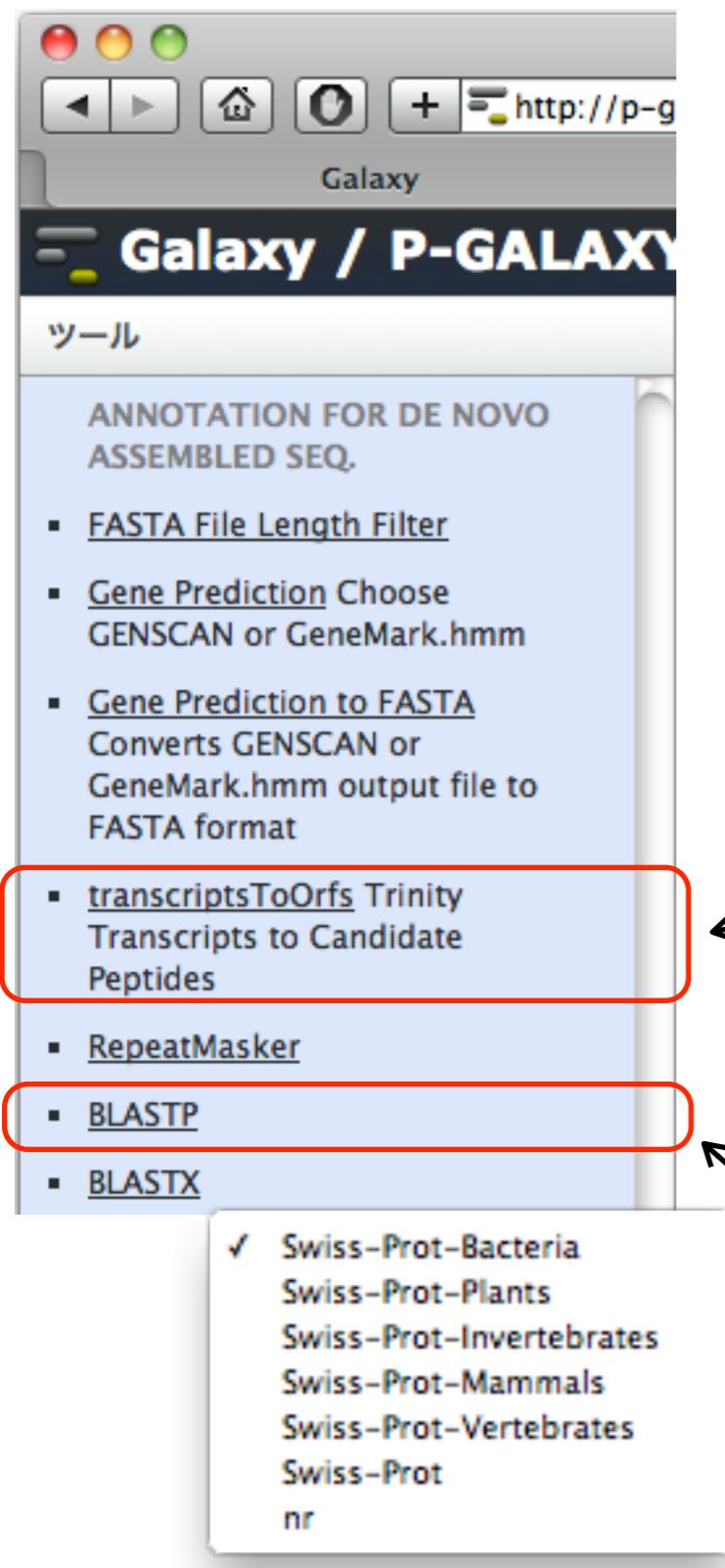


RNA-SeqのCufflinks実行(発現量の正規化)
gtf->wigフォーマット変換
UCSC genome browser siteでの可視化



(<http://genome.ucsc.edu/cgi-bin/hgGateway>)

ChIP-Seq
MACSによるDNA結合タンパク質の結合部位候補の同定

RNA-Seqの*de novo* アセンブル結果の解析

Trinityによるアセンブル



FASTAファイル



配列長フィルター



アミノ酸変換

長いORFかつ
HMMERによるモチーフ検索UniProtKB/Swiss-Prot、
nrに対するBLASTP

BLASTP 2.2.25 [Feb-01-2011]

Reference: Altschul, Stephen F., Thomas L. Madden, Alejandro A. Schaffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman (1997), "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs", Nucleic Acids Res. 25:3389-3402.

Reference for compositional score matrix adjustment: Altschul, Stephen F., John C. Wootton, E. Michael Gertz, Richa Agarwala, Aleksandr Morgulis, Alejandro A. Schaffer, and Yi-Kuo Yu (2005) "Protein database searches using compositionally adjusted substitution matrices", FEBS J. 272:5101-5109.

Query= gene_2|Genemark.hmm|51_aa|+|161|316 >9641
(51 letters)

Database: uniprot_sprot.fasta
529,056 sequences; 107,423,367 total letters

Searching.....done

Sequences producing significant alignments: Score E (bits) Value

sp|C0H402|YKZP_BACSU Uncharacterized protein ykzP OS=Bacillus su... 107 2e-23
>sp|C0H402|YKZP_BACSU Uncharacterized protein ykzP OS=Bacillus subtilis GN=ykzP PE=4
ST=1 Length = 51

Score = 107 bits (267), Expect = 2e-23, Method: Compositional matrix adjust.
Identities = 51/51 (100%), Positives = 51/51 (100%)

Query: 1 MGRRAEVNEAIKKNNPTTESMIDPNNSYKTQYHDDPNFPGANRNSHQQQGGef 51
MGRRAEVNEAIKKNNPTTESMIDPNNSYKTQYHDDPNFPGANRNSHQQQGGef
Sbjct: 1 MGRRAEVNEAIKKNNPTTESMIDPNNSYKTQYHDDPNFPGANRNSHQQQGGef 51

BLASTP 2.2.25 [Feb-01-2011]

Reference: Altschul, Stephen F., Thomas L. Madden, Alejandro A. Schaffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman (1997), "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs", Nucleic Acids Res. 25:3389-3402.

Reference for compositional score matrix adjustment: Altschul, Stephen F., John C. Wootton, E. Michael Gertz, Richa Agarwala, Aleksandr Morgulis, Alejandro A. Schaffer, and Yi-Kuo Yu (2005) "Protein database searches using compositionally adjusted substitution matrices", FEBS J. 272:5101-5109.

Query= gene_3|Genemark.hmm|49_aa|+|346|495 >9641
(49 letters)

Database: uniprot_sprot.fasta
529,056 sequences; 107,423,367 total letters

Searching.....done

Sequences producing significant alignments: Score E (bits) Value

sp|031659|YKZE_BACSU Uncharacterized protein ykzE OS=Bacillus su... 75 8e-14
>sp|031659|YKZE_BACSU Uncharacterized protein ykzE OS=Bacillus subtilis GN=ykzE PE=4
ST=1 Length = 58

Score = 75.5 bits (184), Expect = 8e-14, Method: Compositional matrix adjust.
Identities = 36/36 (100%), Positives = 36/36 (100%)

Query: 1 HQNRRKXPINRKTVLEEEFSSSELGDYNAGKIIETLEVT 36
HQNRRKXPINRKTVLEEEFSSSELGDYNAGKIIETLEVT
Sbjct: 10 HQNRRKXPINRKTVLEEEFSSSELGDYNAGKIIETLEVT 45

DDBJパイプラインで実行するTrinityについて

Inchworm:

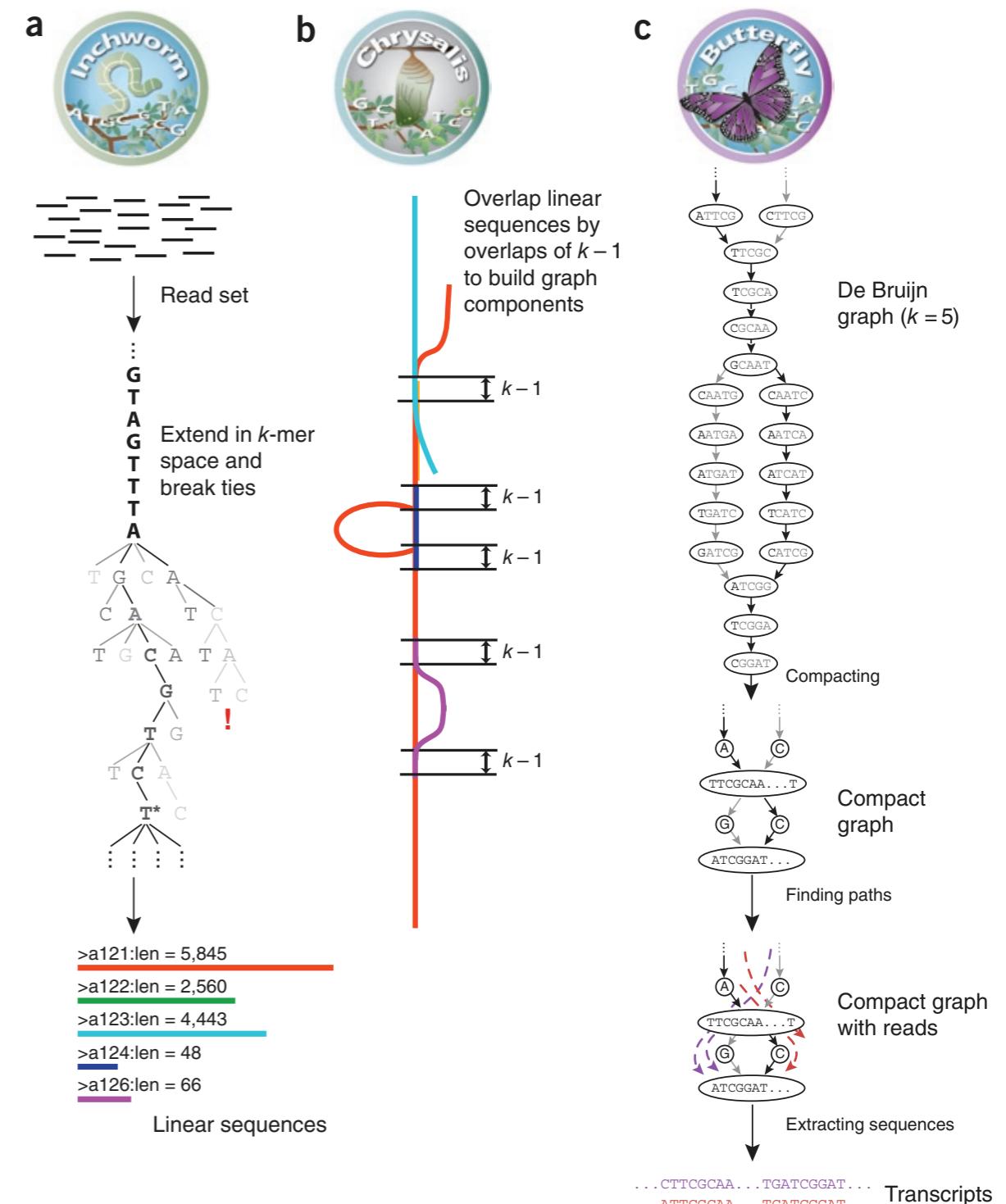
k -mer($k=25$)でざっくりアセンブルしてコンティグをつくる。

Chrysalis:

スプライスバリエントやパラログ由来のコンティグを含めてクラスター化
コンティグの共通部分を基にどういう経路をとってつながっていくか? >グラフを作成

Butterfly:

グラフを精査していくってスプライスバリエントやパラログも再構成する。



Trinityについては

Nat Biotechnol. 2011 May 15;29(7):644-52.

グラフアルゴリズムについては

http://d.hatena.ne.jp/hoxo_m/20100930/p1

等ご参考ください。

Figure 1 Overview of Trinity. (a) Inchworm assembles the read data set (short black lines, top) by greedily searching for paths in a k -mer graph (middle), resulting in a collection of linear contigs (color lines, bottom), with each k -mer present only once in the contigs. (b) Chrysalis pools contigs (colored lines) if they share at least one $k-1$ -mer and if reads span the junction between contigs, and then it builds individual de Bruijn graphs from each pool. (c) Butterfly takes each de Bruijn graph from Chrysalis (top), and trims spurious edges and compacts linear paths (middle). It then reconciles the graph with reads (dashed colored arrows, bottom) and pairs (not shown), and outputs one linear sequence for each splice form and/or paralogous transcript represented in the graph (bottom, colored sequences).

今回はミドリフグのRNA-Seqデータを使用します



Tetraodon nigroviridis

最大で全長17 cm。

観賞魚としてポピュラーであり、2-3 cm程度の幼魚
が多く熱帯魚店等で売られている。

SRR579565 (エントリー: SRA059267)
76bpの150,435,952リード
ペアエンド

謝辞

大量遺伝情報研究室の方々
富士ソフト株式会社 森崎さん

DDBJの方々



本研究は、文部科学省科学研究費新学術領域研究『生命科学系3分野支援活動』
「ゲノム支援」および科学研究費基盤(C)の支援を受けております。

大量研ではDDBJパイプラインをカンキツ類、野生イネ、ミニトマト、ゼニゴケ等
の変異解析、パラゴムの木のアセンブルに使用しております。

DDBJ Read Annotation Pipeline: a cloud computing-based pipeline for high-throughput analysis
of next-generation sequencing data.
DNA Res. 2013 Aug;20(4):383-90.

実習内容

DDBJ パイプラインを用いた denovo RNAseq アセンブリ

DRA (DDBJ Sequence Read Archive)からの配列データのインポート

DDBJパイプライン基礎部での Preprocessing ジョブ実行

DDBJパイプライン基礎部での Trinity ジョブ実行

DDBJパイプライン高次解析部(Galaxy)でのジョブ実行

参考資料

DDBJパイプライン(基礎部)へのアカウント作成

DDBJパイプライン(基礎部)のFTPによるデータ転送

DRAからの配列データインポート

今回使用する高速シーケンサー配列の確認

DRAで検索すると早い

DRA: <http://trace.ddbj.nig.ac.jp/dra>

今回は実習用サンプルとしてミドリフグの高速シーケンサーで出力された RNAseq 配列を用いる。

DRAのwebサイトから「検索」をクリック



DRASearchのwebサイトが表示

「Organism:」に「*Tetraodon nigroviridis*」と入力し、「Search」をクリック。

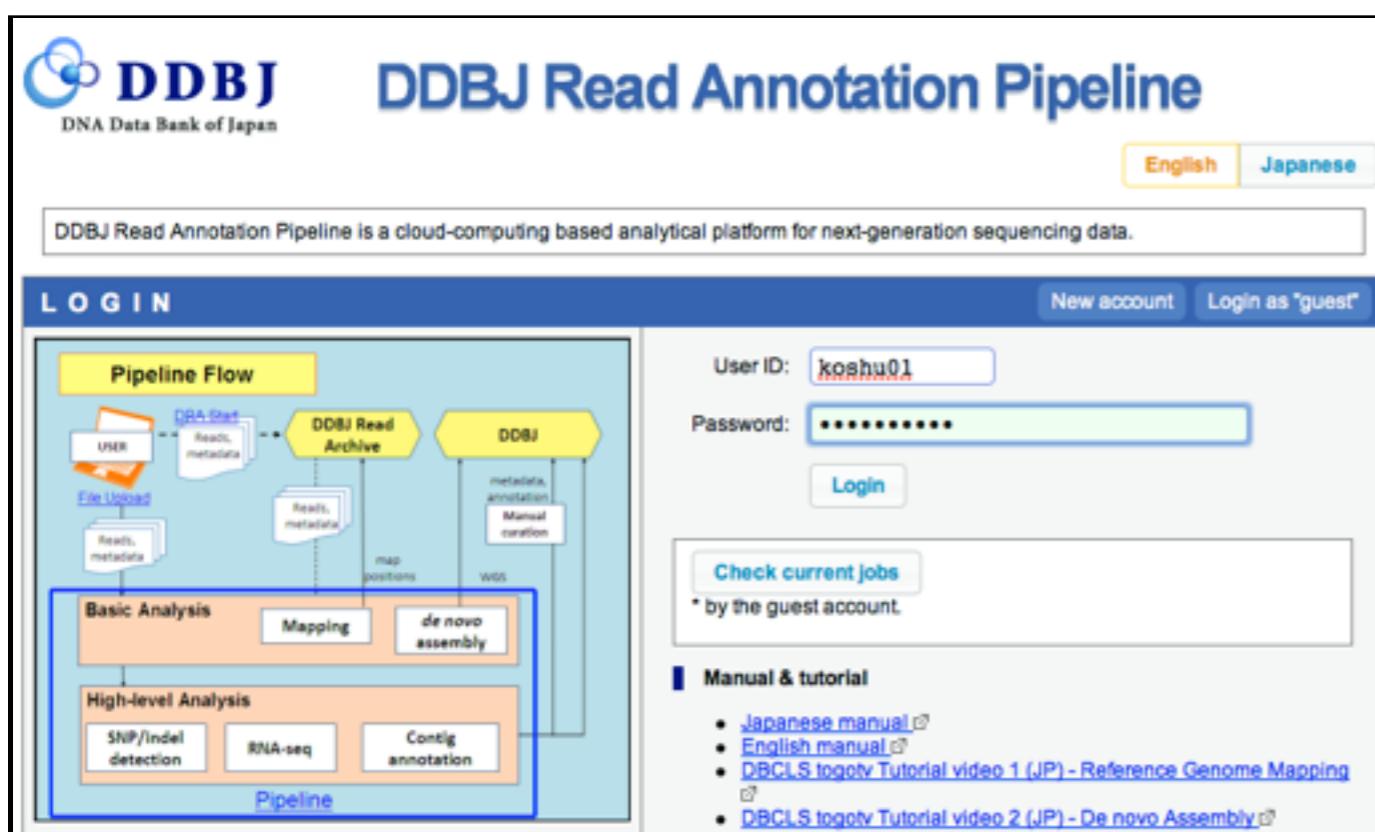
今回はアクセション番号「SRA059267」のデータをサンプルに用いる。Pipelineからインポートするのに必要なので、アクセションをメモしておく。

#	STUDY	SUBMISSION	STUDY_TITLE	STUDY_TYPE	ORGANISM	BASES	SUBMITTED	CENTER_NAME
1	SRP00117	SRA012701	GSE19824: Genome-wide evolutionary analysis of eukaryotic DNA methylation (RNA-Seq)	Transcriptome Analysis	<i>Tetraodon nigroviridis</i>	5.1G	2010-07-21	GEO
2	SRP002418	SRA012701	GSE19824: Genome-wide evolutionary analysis of eukaryotic DNA methylation (ChIP-Seq)	Epigenetics	<i>Tetraodon nigroviridis</i>	1.3G	2010-07-21	GEO

The screenshot shows the DRASearch results page. It has a search interface at the top with fields for Accession, Organism (set to 'Tetraodon nigroviridis'), StudyType, Platform, and Keyword. Below the search bar, it says 'Show 20 records Sort by Study' and has 'Search' and 'Clear' buttons. The results table is titled 'Search Results (4 studies)'. The first row is highlighted with a red circle around the study ID 'SRP00117'. The table columns are: #, STUDY, SUBMISSION, STUDY_TITLE, STUDY_TYPE, ORGANISM, BASES, SUBMITTED, and CENTER_NAME. The results show two entries for 'SRP00117' and one for 'SRP002418'. The 'Organism' column lists various species, and the 'CENTER_NAME' column shows 'GEO' for all entries.

DDBJパイプラインにログイン

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



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

DDBJ, pipeline で検索すると早い

デモ用アカウントは
講習内でお伝えします

DRAから配列データをインポート

DDBJパイプラインログインする。

「Import public DRA」をクリック

The screenshot shows the 'Selecting Query Files' step of the DDBJ Pipeline. The left sidebar shows account and analysis options. The main area has tabs for 'FTP upload', 'Private DRA entry' (highlighted with a red circle), 'Import public DRA' (highlighted with a red circle), 'Preprocessing', and 'HTTP upload'. Below the tabs is a table of metadata for a DRA entry, including columns for Type, Accession, Alias, Filename, DL, and View. A 'NEXT' button is at the top right.

TYPE	ACCESSION	ALIAS	FILENAME	DL	VIEW
Submission	DRA000001		DRA000001.submission.xml	<input type="button" value="DownLoad"/>	<input type="button" value="View"/>
Sample	DRS000001	Bacillus subtilis subsp. natto BEST195 without plasmid pBEST195L	DRA000001.sample.xml	<input type="button" value="DownLoad"/>	<input type="button" value="View"/>
Study	DRP000001	Natto BEST195	DRA000001.study.xml	<input type="button" value="DownLoad"/>	<input type="button" value="View"/>
Experiment	DRX000001	NATTO_BEST195_SEP08	DRA000001.experiment.xml	<input type="button" value="DownLoad"/>	<input type="button" value="View"/>
Run	DRR000001	2008-09-12 BEST195-Lane7	DRA000001.run.xml	<input type="button" value="DownLoad"/>	<input type="button" value="View"/>

「Input DRA/ERA/SRA Accession Number」に

「SRA059267」と入力

「Add my DRA entry」をクリック

The screenshot shows the 'Selecting Query Files' step again, with the 'Import public DRA' tab selected. It includes instructions for importing public FASTQ files from the DRA database and a note about automatic download. Below is a form for inputting an accession number, with a red circle around the input field containing 'SRA059267'. A 'NEXT' button is at the top right.

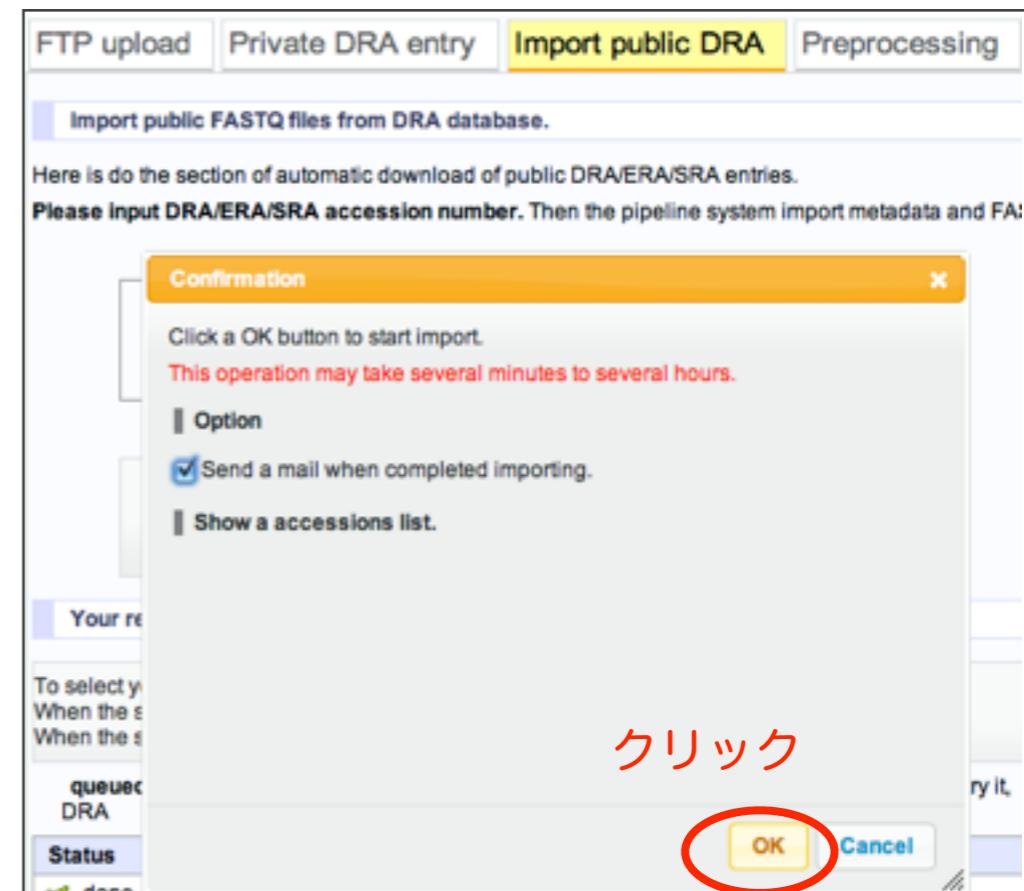
Input DRA/ERA/SRA Accession Number
SRA059267 クリック

DRAから配列データをインポート

「Confirmation」のダイアログが現れる。

「Send a mail when completed importing」のチェックを確認。チェックしておくとimport終了時にメールが届く。

「OK」をクリック。



importの進行状況は、「Import public DRA」タブ内で確認できます。

webブラウザをリロードして下方の入手リストを確認。

実行中のDRAのアクセッションが「queued」から「done」になったら完了。

ブラウザリロードで確認

A screenshot of the "Import public DRA" tab. At the top, the tab is highlighted with a red circle. Below it, there's a section for entering a DRA accession number with a "Add my DRA entry" button. A note says "Accession Number can find here. DRA Search". Under "Your request. (Here is display only, can not select.)", there's a note about selecting entries and their status. A table at the bottom shows a single entry: Status: queued, Submission: DRA000307, Request date: 2012-06-22 13:40:50.392. The "queued" row is circled in red.

Preprocessing

リードのクオリティ値によるフィルタリング

Preprocessing 実行するクエリファイルを選択

Trinity 実行の前に、インポートしたデータの前処理として、QV によるフィルタリングを行う

- ①左のメニューから「Preprocessing」を選択し、②「Private DRA entry」タブをクリックする。

ドロップダウンリストから、

- ③先ほどインポートしたアクセション「SRA059267」を選択する。

(FTPアップロードしたファイルを選択することも可能)

DDBJ DNA Data Bank of Japan

ACCOUNT

login ID [koshu01]
Logout
Change password

ANALYSIS

Data setup

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

NEXT

Selecting Query Files

FTP upload **Private DRA entry** (circled in red) ②

Metadata of the DRA entry.

TYPE	ACCESSION	ALIAS	FILENAME	DL
Submission	DRA000307	tomohiro-0005_Submission	DRA000307.submission.xml	Download view (circled in red) ③ SRA012701
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 a metadata ✓ DRA000307
Select your registered query files.
Queries with different Instrument models can't be selected together.
single paired all clear

No.	Experiment ACCESSION	Sample ACCESSION	Run ACCESSION	STRAIN	Run_date	Read #	Read length	Instrument model	Layout
1	SRX191149	SRS366858	SRR579545	strain C57BL/6				ILLUMINA	paired
2	SRX191150	SRS366859	SRR579546	strain C57BL/6				ILLUMINA	paired
3	SRX191151	SRS366860	SRR579547	strain C57BL/6				ILLUMINA	paired
4	SRX191152	SRS366861	SRR579548	strain C57BL/6				ILLUMINA	paired
5	SRX191153	SRS366862	SRR579549	strain C57BL/6				ILLUMINA	paired
6	SRX191154	SRS366863	SRR579550	strain C57BL/6				ILLUMINA	paired
7	SRX191155	SRS366864	SRR579551					ILLUMINA	paired
8	SRX191156	SRS366865	SRR579552					ILLUMINA	paired
9	SRX191157	SRS366866	SRR579553					ILLUMINA	paired
16	SRX191164	SRS366873	SRR579560					ILLUMINA	paired
17	SRX191165	SRS366874	SRR579561					ILLUMINA	paired
18	SRX191166	SRS366875	SRR579562					ILLUMINA	paired
19	SRX191167	SRS366876	SRR579563					ILLUMINA	paired
20	SRX191168	SRS366877	SRR579564					ILLUMINA	paired
21	SRX191169	SRS366878	SRR579565					ILLUMINA	paired

ウィンドウ下部にメタデータおよびファイル一覧が表示されるので、この中から、Tetraodon_nigroviridis_RNA-Seq に該当する Experimental ACCESSION SRX191169 のものをチェック。

最下部の「NEXT」を押し、次画面に進む。

NEXT

Preprocessing 実行条件の指定

Trinity 実行の前に、インポートしたデータの前処理として、QV によるフィルタリングを行う

Your selected queries

Run ACCESSION	Read length	Quality Score	Read Layout
SRR042533 ->	bp		single

Steps of preprocessing workflow

Step1: Set the type of the quality value.

- クオリティ値の選択 DRA からインポートされた
◦ Phred+33 Phred+64 → データはすべて Phred+33 形式になっています。

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.

リードの両端から QV <=19 となる塩基をトリム。

- QV THRESHOLD : → トリム後の長さが 25 bp 未満となった場合は、リード全体を削除。
(ペアの場合は、ペアとなるもう一方も同時に除かれる)

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

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

- QV THRESHOLD : トリム後のリードの中に、QV <= 14 のリードが 30 %
→ 以上含まれていた場合、リード全体を削除。
◦ Percentage THRESHOLD : (ペアの場合は、ペアとなるもう一方も同時に除かれる)

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'.

最下部の「NEXT」を押し、次画面に進む。

BACK

NEXT

Preprocessing 実行および実行状況の確認

Trinity 実行の前に、インポートしたデータの前処理として、QV によるフィルタリングを行う

メールを入力して「Run」ボタンを押す。

The screenshot shows a configuration interface for a preprocessing job. It includes fields for 'Email notification' (with a note that it's required), 'Confirmation of entries', and a 'Query sets' section containing 'SRR042533 - GSM497271_1'. At the bottom right are 'BACK' and 'RUN' buttons.

ステータス画面でジョブの実行状況の確認。

Preprocessing でフィルタリングをした
クエリファイルを利用してdenovo Assembly
/ mapping を行う場合、ジョブIDが必要になる
ので、覚えておくこと。

「View」ボタンで詳細を確認。

The screenshot shows a 'Status - Preprocessing' page with a table of jobs. The columns are: ID, UserID, Files, P/S, Status, Read #, Read length, Detail, Start time, End time, and Elapsed time. The 'Preprocessing Job' tab is selected. Two rows are highlighted with red circles: the first row for job ID 5509 (Status: running) and the second row for job ID 5455 (Status: complete). The 'View' button in the Detail column for job 5509 is also circled.

ID	User ID	Files	P/S	Status	Read #	Read length	Detail	Start time	End time	Elapsed time
5509	oshu01	SRA012701 GSM497271_1	S	running	—	—	View	2013-04-30 17:42:30	—	—
5455	--	-- FY23KIH080_pl	P	complete	—	—		2013-04-25 11:55:45	2013-04-25 13:15:56	01:20:10
5452	--	-- FY22KIH033_pl	P	complete	—	—		2013-04-25 11:16:44	2013-04-25 12:44:28	01:27:43

Preprocessing 結果の確認

Trinity 実行の前に、インポートしたデータの前処理として、QV によるフィルタリングを行う

Detail view

Job info

ID 5509	Tool (Version) (1.0)	RunAccession or Filename SRR042533	Download SRR042533.fastq.bz2	Read length N.A. bp	Alias GSM497271_1
File SRR042533.fastq.bz2	Fastq Download download (1.3 GB)	QS Average (PDF) download (6.6 KB)	QS Count (PDF) download (5.1 KB)		

Time

Wait time	Start time	End time
0:0:59	2013-04-30 17:42:30	2013-04-30 17:56:24

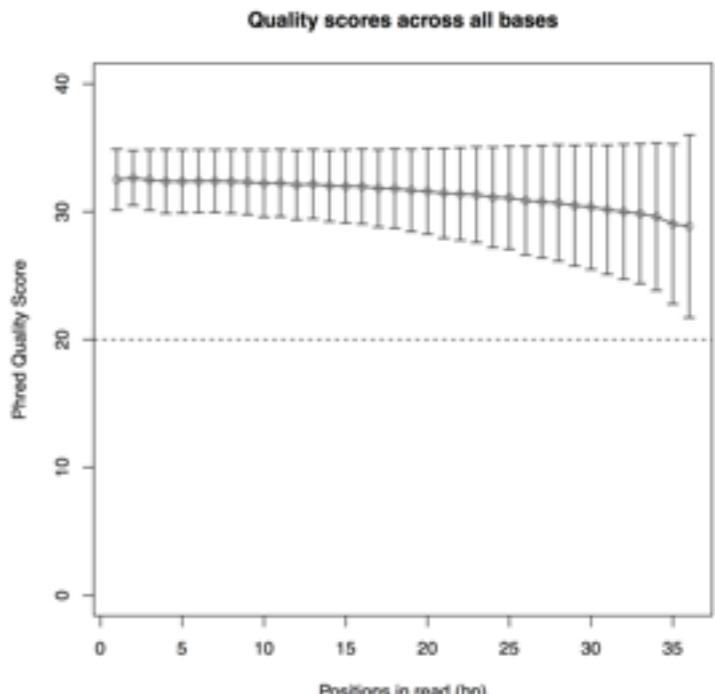
Command	Start time	End time	Log1	Log2	Result	MD5
perl avq_p.pl fqlist.txt qscore	2013-04-30 17:42:30	2013-04-30 17:50:28	View			
perl pdel_p3_t.pl fqlist.txt qscore19 24 0 14 30 33	2013-04-30 17:50:28	2013-04-30 17:53:51				
perl user_fastq_copy.pl preprocessing.xml koshu01	2013-04-30 17:53:51	2013-04-30 17:56:24	View			

ログの確認

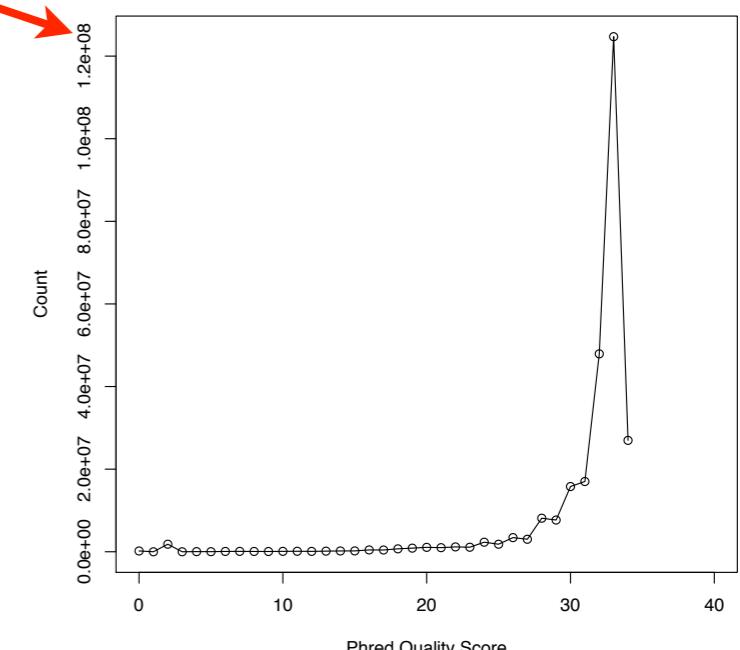
BACK

「BACK」ボタンでジョブ履歴画面に戻る

リード位置ごとの平均クオリティ値



クオリティ値ごとの塩基数



denovo Assembly Trinity の実行

Trinityの実行 クエリファイルの選択

クエリとなるFASTQ/FASTA配列を選択する方法としてDDBJパイプラインでは、下記の4通りの方法がある。

- FTPクライアントソフトでアップロードした配列を使用
「FTP upload」
- webブラウザでアップロードした配列を利用
「HTTP upload」
- DRAからインポートした配列を使用する
「Private DRA entry」
- Preprocessing で処理した配列を使用
「Preprocessing」

選択

Filename	Layout	File size
4927_DRR000719_1.unmapped.fastq_4741.bz2 (more 1 files)	paired	65.8 MB
4932_DRR000719_1.unmapped.fastq_4746.bz2 (more 1 files)	paired	65.8 MB
4971_DRR000719_1.unmapped.fastq_4785.bz2 (more 1 files)	paired	65.8 MB
5509_SRR042533_e.fastq.bz2	single	239.7 MB

DELETE NEXT

今回は Preprocessing で処理したクエリを使用する。

画面左のメニューから、「Preprocessing Start」を選択。

Preprocessing で処理されたファイルは、

「(PreprocessingのジョブID) _ もとのファイル名_e.fastq.bz2」という形式のファイル名になっているので、先ほど確認しておいたジョブIDで始まるものを選択。

最下部の「NEXT」をクリック。

次へ

Trinityの実行 ツールの選択

「denovo Assembly」 → 「Trinity」 の順に選択

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 checked="" type="checkbox"/>	Trinity 	 	r2012-06-08			✓		RNA-Seq De novo Assembly

最下部の「NEXT」をクリック。

Trinityの実行 クエリのレイアウト選択

実行するAccessionの横のチェックボックスをクリック

右側の「confirm」ボタンをクリック。（ペアエンドのクエリの場合「Set as PairEnd」ボタン）

Single analysis

Layout of single sequence.

5' 3'

Linker(1)	Target	Linker(2)
-----------	--------	-----------

Run ACCESSION **Read length** **Quality Score**

	Run ACCESSION	Read length	Quality Score
<input checked="" type="checkbox"/>	5509_SRR042533_e.fastq.bz2	bp	

confirm

画面下に確定したレイアウトが表示されるので、最下部の「NEXT」をクリック。

QUERY SET
Query set1

PairedOrientation	RunAccession	RunAlias	RowLength	QualityScore1	QualityScore2
single	1819	SRR042533 by Preprocessing			

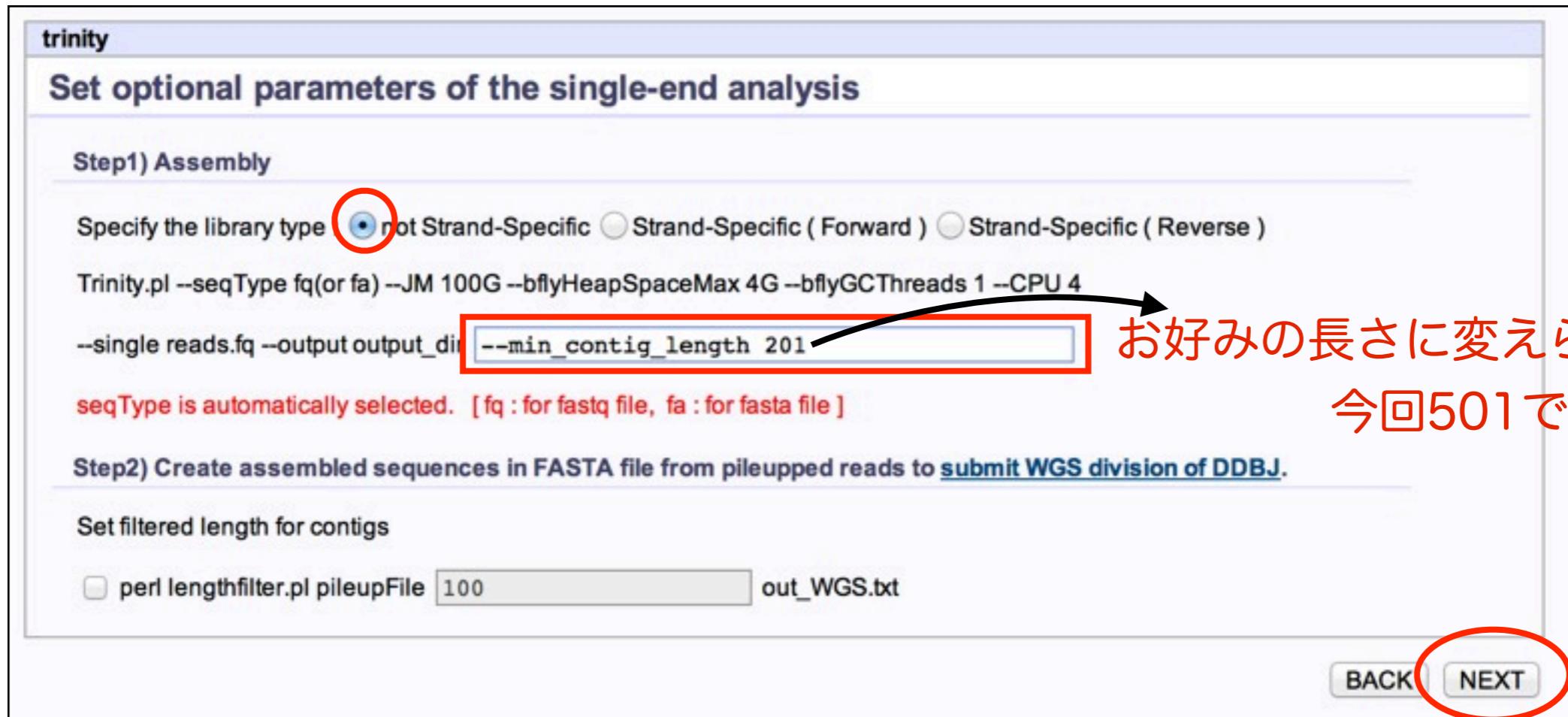
RESET BACK **NEXT**

今回はクエリファイルを1つしか選択していないので、あまり意味はないが、複数のファイルを選択していた場合、それらをすべて結合して実行するか、あるいは、別々に連続して実行するかをこの画面で選択する。

Trinityの実行 実行オプションの指定

library type および 実行時のオプションを指定。

今回は501の条件で実行する、



参考) Pipelineで使用している Trinity 実行コマンド

クエリファイルの種類	メモリ、CPU 関係の指定(固定)
FASTA or FASTQ (自動で指定される)	
Trinity.pl <u>--seqType fq</u> <u>--JM 100G</u> <u>--bflyHeapSpaceMax 4G</u> <u>--bflyGCThreads 1</u> <u>--CPU 4</u>	
<u>--single <クエリファイル名></u> <u>--output <出力ディレクトリ名></u> <u>--min_contig_length 201</u>	
入力ファイル・出力ファイルの指定 (自動で指定される)	ユーザーの指定するオプション

Trinityの実行 実行オプションの確認

メールアドレスを入力して、「RUN」ボタンを押す

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 [trinity]

Query sets
Query set1

PairedOrientation	RunAccession	RunAlias	RowLength	QualityScore1	QualityScore2
single	1819	SRR042533 by Preprocessing			

Assembly commands
trinity
Set optional parameters of the single-end analysis

Step1) Assembly
Specify the library type : not Strand-Specific Strand-Specific (Forward) Strand-Specific (Reverse)
Trinity.pl --seqType fq(or fa) --JM 100G --bflyHeapSpaceMax 4G --bflyGCThreads 1 --CPU 4
--single reads.fq --output output_dir
seqType is automatically selected. [fq : for fastq file, fa : for fasta file]

Step2) 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

[BACK](#) [RUN](#)

Trinityの実行 実行状況の確認

Status → denovo Assembly から、実行したジョブの確認をする

	ID	User ID	Submission accession	P/S	Status	Tool	Read #	Read length	Assembly detail	Mapping detail	Start time	End time	Elapsed time
<input type="checkbox"/>	5516	--	-- Whole transcript	S	running	Trinity	46,765,342	--			2013-04-30 18:26:32	--	
<input type="checkbox"/>	5515	koshu01	SRA012701 GSM497271_1	S	complete	Trinity	7,468,448	--	View		2013-04-30 18:15:21	02:40:24	
<input type="checkbox"/>	5514	koshu01	-- SRR042533 by	S	complete	Trinity	7,420,316	--	View		2013-04-30 18:13:59	02:42:20	
<input type="checkbox"/>	5508	--	-- Drosophila RNA	S	complete	Trinity	18,524,700	--			2013-04-30 17:23:42	03:43:55	
<input type="checkbox"/>	5507	--	SRA009364 42CRDAAXX	P	complete	Trinity	9,262,350	--			2013-04-30 15:45:56	05:21:38	

「View」ボタンをクリックして、詳細確認。

Trinityの実行 実行状況の確認

Status → denovo Assembly から、実行したジョブの確認をする

Job info

ID 5514	Tool (Version) Trinity (r2012-06-08)		
RunAccession or Filename 1819	Download 5509_SRR042533_e.fastq.bz2	Read length N.A. bp	Alias SRR042533 by Preprocessing

Download modified queries

- [5509_SRR042533_e.fastq.gz \(Original size 1.3 GB\)](#)

Download wgs file

- [out_WGS.fasta.gz \(Original size 1.0 MB\)](#)

Assembly statistics

結果ファイルの統計値

Contig # : 2,466
Total contig size : 1,018,683
Maximum contig size : 4,351
Minimum contig size : 202
N50 contig size : 450

Time

Wait time	Start time	End time
0:0:47	2013-04-30 18:13:59	2013-04-30 20:56:20

結果ファイルのダウンロード

Command	Start time	End time	Log1	Log2	Result	MD5
Trinity.pl --seqType fq --JM 100G --bflyHeapSpaceMax 4G --bflyGCThreads 1 --CPU 4 --single 5509_SRR042533_e.fastq --output output_dir -- min_contig_length 201	2013-04-30 18:13:59	2013-04-30 20:55:45	View		Download(353.7 KB)	MD5

[BACK](#)

「BACK」ボタンで、一覧画面に戻る

これで基礎部は終了です。

DDBJパイプライン高次解析部による RNA-Seqアセンブル結果の解析

高次解析部起動

パイプライン基礎部の左のメニュー欄から「step-2/Workflow」をクリック。

The screenshot shows a table of pipeline steps. The 'Workflow' section, which includes options like 'Genome (SNP/Short Indel)', 'RNA-seq (Tag count)', and 'ChIP-seq', is circled in red.

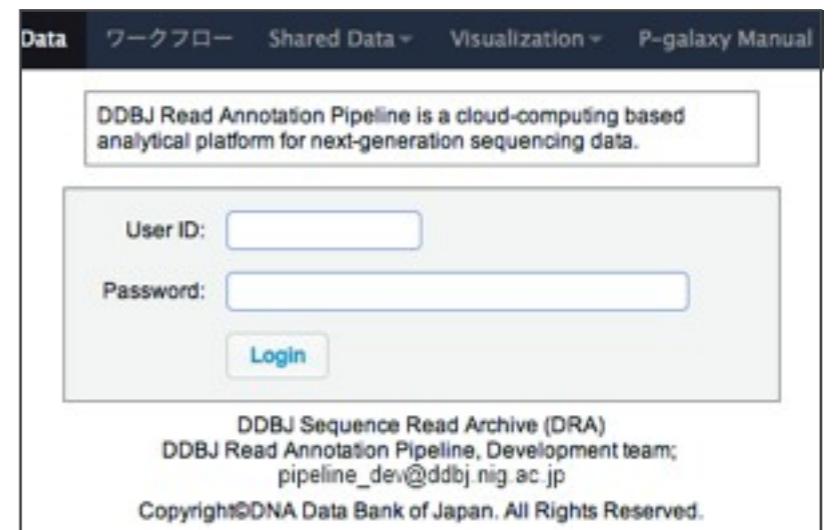
ID	User ID	Status
4927	koshu01	待機
4407	koshu01	実行中
4291	koshu01	待機
4289	koshu01	待機

高次解析部(GALAXY)が起動



Tips:

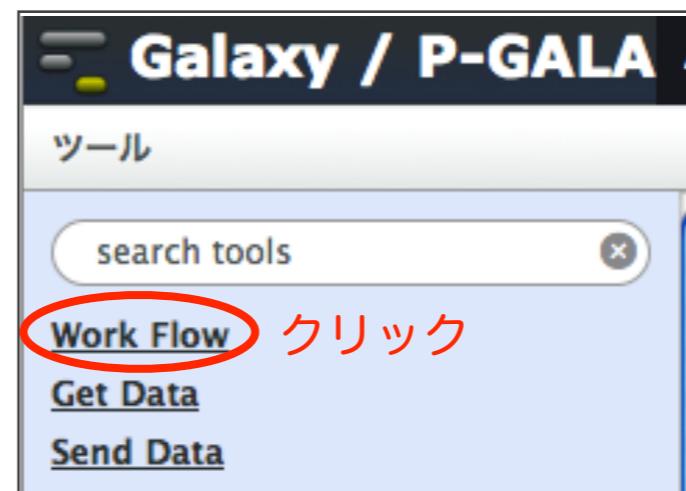
<http://p-galaxy.ddbj.nig.ac.jp>でURL直打ちして、「ツール」メニューの「Work Flow」をクリック。
基礎解析と同じパイプライン登録時のメールアドレスとパスワードを入力しても起動可能。



RNA-Seqのアセンブル結果をインポート

TrinityによるRNA-Seqのアセンブル結果を
GALAXYにインポートする。

左側「ツール」メニューの「Work Flow」をクリック



左側「ツール」メニューの「COMMON PROCESS」
の下「import contig form DDBJ Pipeline」を
クリック

実行したジョブのsamfileのリストのうち、今回は
「SRR042533 by Preprocessing」の「import」を
クリック

左側は「COMMON PROCESS」メニューで「import contig fasta file Import from DDBJ Pipeline」が赤い丸で囲まれて「クリック」と指示されています。右側は「Login ID [koshu01]」でログインしている状態で、表題「Import Contig FASTA from basic analysis. By DDBJ Read Annotation Pipeline」の下にジョブ一覧が表示されています。

JOBID	Submission Accession RunAlias	Tool	Pipeline Jobpage	Import to Galaxy
5519	SRA009364 42CRDAAXX	trinity	ViewJob	Import
5515	SRA012701 GSM497271_1	trinity	ViewJob	Import
5514	SRR042533 by Preprocessing	trinity	ViewJob	Import

「SRR042533」を確認と「クリック」の指示が右側に記載されています。

中央にツール実行開始の表示が現れ...



左側のヒストリーに読み込み中のファイルが
表示される(緑色になったら終了)
ヒストリーの目のアイコンをクリックすると
中央にプレビューされる。



アミノ酸変換

さらにその下の「transcriptsToOrfs (N.A.)
Trinity Transcripts to Candidate Peptides」
をクリック

CPU: 16くらい推奨
「Execute」をクリック

結果としては

- 1) アミノ酸配列
 - 2) pfamのドメインとのマッチング
 - 3) その他ORF候補
- が返ってくる。

クリック 結果1

transcriptsToOrfs (N.A.) Trinity
Transcripts to Candidate Peptides

transcriptsToOrfs (version 0.0.1)

Transcripts sequences in fastA format:
2: FASTA File Length...r on data 1

Minimum peptide length (in amino acids):
100

Strand specificity type:
NOT strand specific, examine both strands

Pfam database:
Pfam-A.hmm

CPU:
16 16くらい推奨

Number of CPUs to use by hmmscan

Execute クリック

>m.565 g.565 ORF g.565 m.565 type:internal len:207 (-)...
DLEMQIEGLKEELIFLKKNHEEELLAMRAQMSGQVHVEVAAPAPAEDLTKVMADIREHYES
ITAKNQKELETWFNSKSEALNKEMMTQTVTQLTSRSEVTEVKRSQALQIELESLLGMKA
SLEGTIQLDTQNRYSMMLAGYQQQVTSLEQQLVQLRADLVRQGQDYQMLLDIKTRLEIA
EYRRILLEGEAAASSSTSSTSSTKTRRL
>m.566 g.566 ORF g.566 m.566 type:complete len:216 (+)...
MAQSVPVVMFKLVLVGDGGTGKTTFVKRHLTGEFEKKYVATLGVEVHPLFFNTNRGNVKF
NVWDTAGQEKFGLRDGYYIQAQCAIIMFDVTSRVTYKNVPNWHRLVRCENIPIVLCG
NKVDIKDRKVKAKSIVFHRKKNLQYYDISAKSNYNFEKPFLWLARKLIGDPNLEFVEMPA
LAPPEVTMDPALAVQYEKELHVASQTA LPDDEDL*
>m.568 g.568 ORF g.568 m.568 type:internal len:227 (-)...
GDRFKEDRKAKRLPEKSIDMIILLTDGDPNSGESRIPVIQENVKAAIGGQMSLFLSGFGN
DVKYPLFDVMSRENNGLARRIYEGSDAALQLQGFYDEVSSPLLVDLRYPDNAVDSLTT
NQFSQLFNGSEIVVAGRLKDNDIDNFPVEVFGQGLNDFSEQQFSVLDWSGMYPDDYIF
GDFTERLWAYLTIQQLLDKSKTGDAEEKANASAEALDMSLRYSFVTP
>m.571 g.571 ORF g.571 m.571 type:5prime_partial len:394...
ASGGEGTHSSCGSWFNAGAKDFPSVPYSYLDNFNDYKCKTSSGEIESYHDVHQVRDCRLVS
LLDLALEKDYVRGKVADYMNRVLDMGVAGFRVDACKHMWPGDLSAVYGRNNLNTKWFPE
GSRPFIFQEVIDLGGEAISYTYYVHLGRVTEFKYGAKGTVFRKWNEKLMYTKNWGEGW
GFMPNGNAVFIDNHNDNQRGHGAGGAAIVTFWDSRLHKMAVAYMLAHPYGVTRVMSSFRW
NRHIVNGKDQNDWMGPSPHDGSTKSVPINPDETCDGFWCEHRWRQIKNMVIFRNVVNG
QPHSNWWDDNNSNQVAFGRGNRGFIIFNNDDWDLDVTNTGLPAGTYCDVISQKEAGRCT
GKQIHVGSDGRAHFRISNRDEDPFVAIHVESKL*
>m.573 g.573 ORF g.573 m.573 type:5prime_partial len:224...
WEPSWPWQVSLQEYTFHFCGGSLINENWWVTAACNCVRTSHRVILGEHDRSSNNENIQV
MQVGQVFKHPNYSYTINNDITLIKASPAQLNIRVSPVCVAETSDVFPGMKCVTSGWG
LTRYNAPDTPPRLQVALPLTNEECRKHWGSKITDLMVCAGASGASSCMGDSGGPLVCE
KAGAWTLVGVISWGSFCVSSPGVYARVTLRAWMDQIIAAN*

結果2

#	# target name	accession	query name	accession	--- full sequence -----			best 1 domain -----			domain number estimation -----								
					E-value	score	bias	E-value	score	bias	exp	reg	clu	ov	env	dom	rep	inc	description of target
	Actin	PF00022.14	m.1	-	2.8e-162	539.5	0.0	3.2e-162	539.3	0.0	1.0	1	0	0	1	1	1	1	Actin
	Apolipoprotein domain	PF01442.13	m.3	-	1.1e-38	132.6	10.6	1.1e-38	132.6	7.3	1.8	2	0	0	2	2	2	2	Apolipoprotein A1/A4/E

結果3

comp1002_c0_seq10	621	ID=m.565;Name=ORF_g.565_m.565_type:internal_len:207_(-)_ (g.565,_m.565);	0	-	0	621	1	621	0
comp1006_c0_seq137	685	ID=m.566;Name=ORF_g.566_m.566_type:complete_len:216_(+)_ (g.566,_m.566);	0	+	37	685	1	648	0
comp1010_c0_seq12	683	ID=m.568;Name=ORF_g.568_m.568_type:internal_len:227_(-)_ (g.568,_m.568);	0	-	2	683	1	681	0

RNA-Seq由来のアミノ酸配列をBLASTPにかける

左側「ツール」メニューの「Work Flow」の下、「ANNOTATION FOR DE NOVO ASSEMBLED SEQ.」の下、「BLASTP」をクリック

BLASTP

クリック

BLASTP (version 1.0.0)

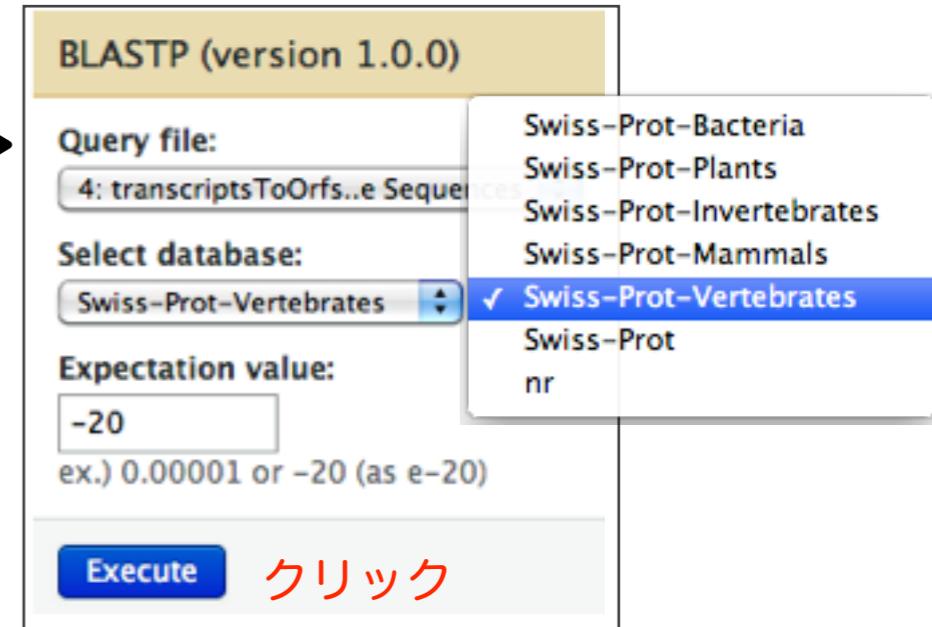
Query file:
4: transcriptsToOrfs...e Sequel

Select database:
Swiss-Prot-Vertebrates

Expectation value:
-20
ex.) 0.00001 or -20 (as e-20)

Execute

クリック



「Select database:」は今回「Swiss-Prot-Vertebrates」を選択
「Expectation Value:」は今回 -20と入力

「Execute」をクリック

「BLASTP error/warning reports」はBLASTのエラー表示など

7: BLASTP error/warning reports

6: BLASTP on data 4
94,939 lines

クリック

クリック

BLASTP 2.2.26 [Sep-21-2011]
Reference: Altschul, Stephen F., Tho
Jinghui Zhang, Zheng Zhang, Webb Mil
"Gapped BLAST and PSI-BLAST: a new g



「BLASTP on data...」をクリックするとフロッピーのアイコンが出てくるのでそのアイコンをクリックするとBLASTP結果のダウンロードが始まる。

ワークフローの保存も可能

(GALAXYがメールアドレスを訊いてきたりするのでパイプラインのユーザーアカウント取得後)

参考: <https://main.g2.bx.psu.edu/u/aun1/p/galaxy101> の”4. Converting histories into workflows”など

The screenshot shows the Galaxy P-GALAXY web interface. The top navigation bar includes links for Analyze Data, ワークフロー (Workflow), Shared Data, Visualization, P-galaxy Manual, Admin, Help, User, and a status indicator showing "Using 44.0 GB".

The main content area displays a workflow conversion interface. It lists tools used to create datasets from a history:

- vcfutils.pl varFilter**
- vcfutils.pl vcf2fq**
- Filter pileup on coverage and SNPs**
- 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
- BLASTP**
- RepeatMasker**
- BLASTX**
- transcriptsToOrfs (N.A.) Trinity**
Transcripts to Candidate Peptides
- PHYLOGENETIC ANALYSIS**
- sam_to_fasta for get mapping fasta data**
- HOSOMICHI HLA ANALYSIS**
TOOLS ARE UNDER CONSTRUCTION!
- Picks Up Fine Pairs From Paired Read Set.**

Below this, there's a section for "Workflow name" with a dropdown menu set to "Workflow constructed from history 'Drosophila paired 1'". Buttons for "Create Workflow", "Check all", and "Uncheck all" are available.

The "Tool" section lists the tools included in the workflow, each with a checkbox to "Include" them:

- import contig fasta file**: This tool cannot be used in workflows.
- FASTA File Length Filter**: Include "FASTA File Length Filter" in workflow
- transcriptsToOrfs**: Include "transcriptsToOrfs" in workflow
- BLASTP**: Include "BLASTP" in workflow

The "History items created" section shows the steps generated by the tools:

- 1: import contig fasta file
- 2: FASTA File Length Filter on data 1
- 3: transcriptsToOrfs on data 2: Pfam matches to Candidate Peptide Sequences
- 4: transcriptsToOrfs on data 2: Candidate Peptide Sequences
- 5: transcriptsToOrfs on data 2 Candidate Peptide Coordinates
- 15: BLASTP on data 4

The right side of the interface features a "ヒストリー" (History) panel with a tree view of history lists and actions:

- HISTORY LISTS**
 - Saved Histories
 - Histories Shared with Me
- CURRENT HISTORY**
 - Create New
 - Clone
 - Copy Datasets
 - Share or Publish
 - Extract Workflow** (highlighted)
 - Dataset Security
 - Show Deleted Datasets
 - Show Hidden Datasets
 - Purge Deleted Datasets
 - Show Structure
 - Export to File
 - 削除する (Delete)
 - Delete Permanently
- OTHER ACTIONS**
 - Import from File

参考資料

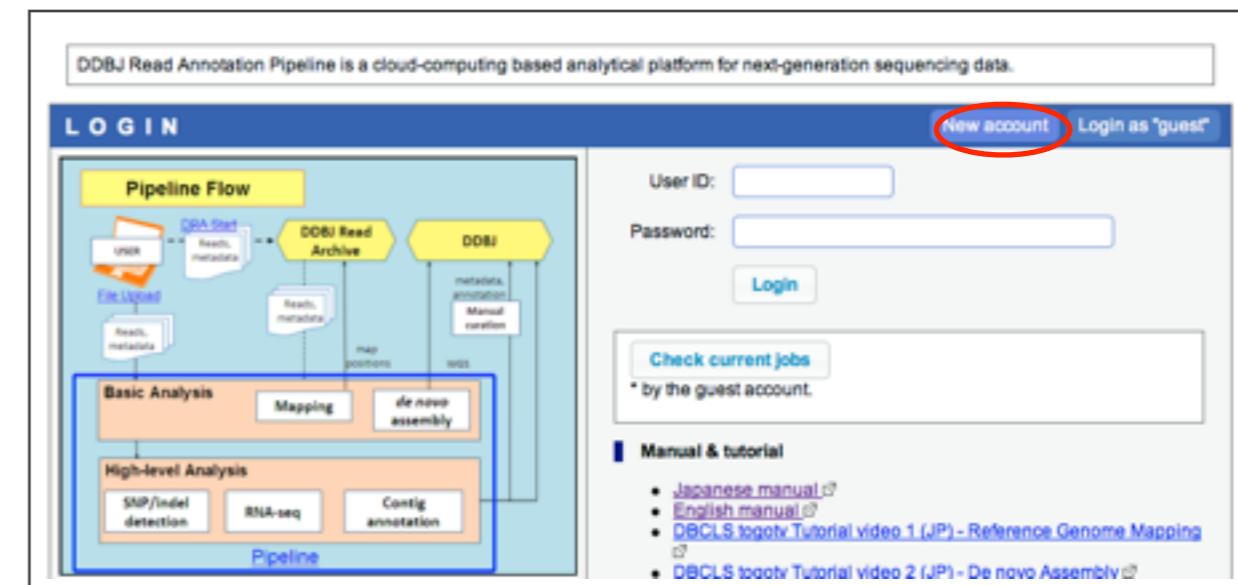
DDBJパイプライン(基礎部)へのアカウント作成

DDBJパイプライン(基礎部)に新規登録

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

DDBJパイプライン(<http://p.ddbj.nig.ac.jp/>)に入る。

「New account」をクリック



UserIDを決めて必要情報を入力

The screenshot shows the 'Registration form for pipeline user accounts'. The form consists of several input fields: 'UserID' (purple background), 'Email address', 'Retype email address' (purple background), 'First name', 'Last name', and 'Institution with department'. A note at the top states: '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](#))'. Below this, a message in red text says: 'After registration, you will receive a confirmation email with your user ID and initial password. Please input your email address correctly.' At the bottom right, there is a 'Registration' button.

「Registration」をクリック

パスワードがeメールで届くのでそのパスワードでログイン

Registration

Query file指定方法

FTP Upload画面へ遷移

ACCOUNT
login ID [guest]
[Logout](#)

ANALYSIS
Data setup
DRA Start
FTP upload (Red Circled)
HTTP upload

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

Running Status

Selecting Query Files

NEXT

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

Metadata of the DRA entry.

入力ファイルの指定方法4種類

1.メニューのFTP Uploadをクリック

step-2

Workflow

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

Genome (Large Indel)

Job Confirmation

- step1. PreProcess status
- step1. Mapping status
- step1. Assembly status
- step2-All status

Help

- HELP
- MANUAL
- BENCHMARK
- Contact Us.
- DDBJ Read Annotation Pipeline.
- Development Team.

Select a metadata : DRA000001

	FILENAME	DL	VIEW
Submission	DRA000001	DRA000001.submission.xml	DownLoad View
Sample	DRS000001	Bacillus subtilis subsp. natto BEST195 without plasmid pBEST195L	DownLoad View
Study	DRP000001	Natto BEST195	DownLoad View
Experiment	DRX000001	NATTO_BEST195_SEP08	DownLoad View
Run	DRR000001	2008-09-12.BEST195-Lane7	DownLoad View

STUDY TITLE Whole genome sequencing of Baillus subtilis subsp. natto BEST195

STUDY TYPE Whole Genome Sequencing

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 type="checkbox"/>	1 DRX000001	DRS000001	DRR000001	strain BEST195	2008-09-12	9,977,388	36	ILLUMINA	paired

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

[DELETE](#) [NEXT](#)

Query file指定方法

FTP clientによるUpload

Registration of fastq/fasta files

1. Upload FASTA/FASTQ files 2. Select FASTA/FASTQ files 3. Registration

Please upload query files.

To use your fasta or fastq files as pipeline query, you need to upload files to our server via FTP or HTTP.
FTP uploading works faster than HTTP uploading. Therefore we recommend using FTP rather than HTTP

By FTP (Recommended)

FTP Configuration.

Server : Port	pdata.nig.ac.jp:21
Security	SSL Explicit encryption
User ID/password	Your Pipeline login ID/password If you can't login via FTP, retry after changing password .

[FTP setting manual](#) ↗

Recommended FTP client softwares.

Windows	FFFTP ↗ WinSCP ↗
Mac OS X	Cyberduck ↗
LinuxOS	FileZilla ↗

For security our FTP server utilizes FTP over SSL protocol (FTPS).
Other FTP client softwares can be used if they support FTPS.

**FTP clientをローカルPCにインストールし、
DDBJのサーバーへFTP転送をする。
※転送方法は次ページに記述**

Query file指定方法

FTP client (Cyberduck)のインストール

FTP client Cyberduckの場合

1.http://cyberduck.ch/へアクセス

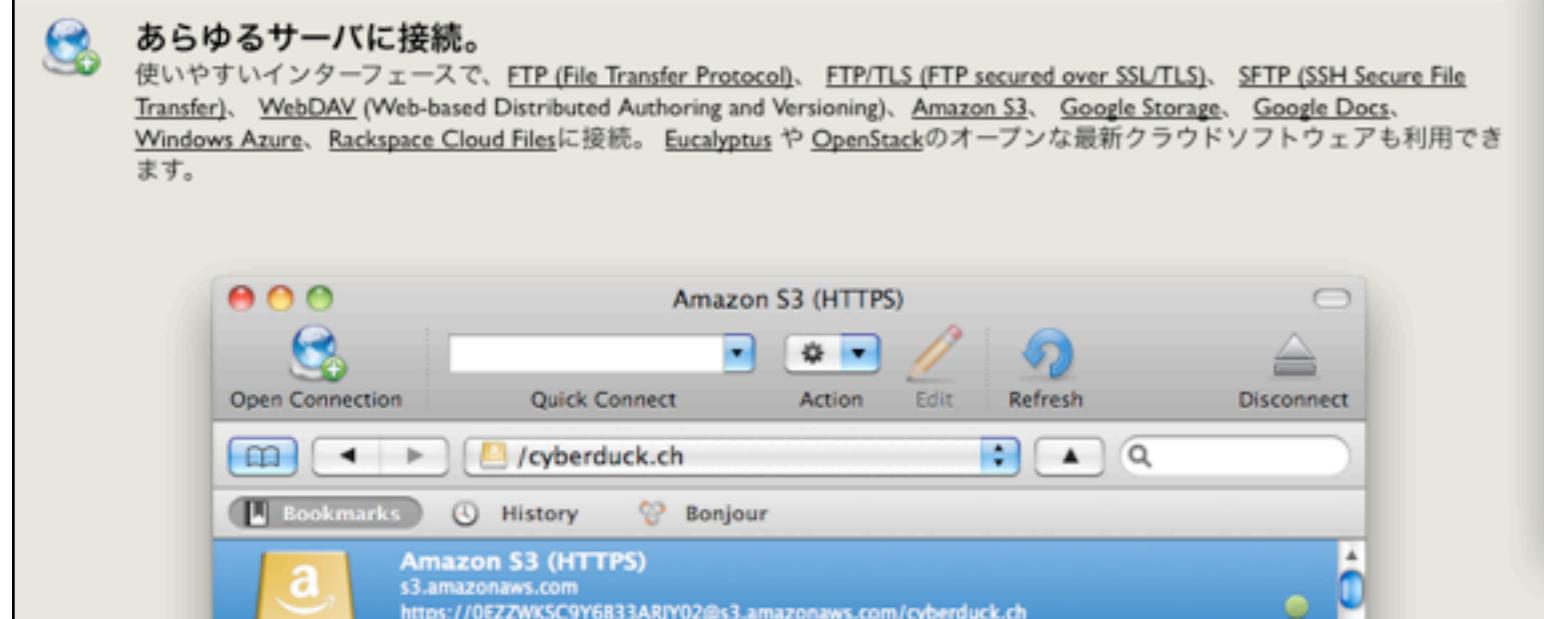


The screenshot shows the Cyberduck website. A yellow rubber duck icon is on the left. The main heading is "Cyberduck" with the subtitle "オープンソース の FTP、SFTP、WebDAV、Cloud Files、Google Docs、Amazon S3用ブラウザ、MacとWindowsに対応。". Below the heading, there's a "Cyberduckについて" menu bar with links to "ニュース", "更新履歴", "開発", "ヘルプ", and "寄付". A sidebar on the right contains download links for different versions:

- ダウンロード バージョン3.8.1 2010年12月6日 Cyberduck-3.8.1.zip ユニバーサルバイナリ、Mac OS X 10.5以降が必要
- バージョン4.0パブリックベータ 2010年12月13日 Cyberduck-Installer-4.0b8.exe Windows XP、Windows VistaまたはWindows 7が必要

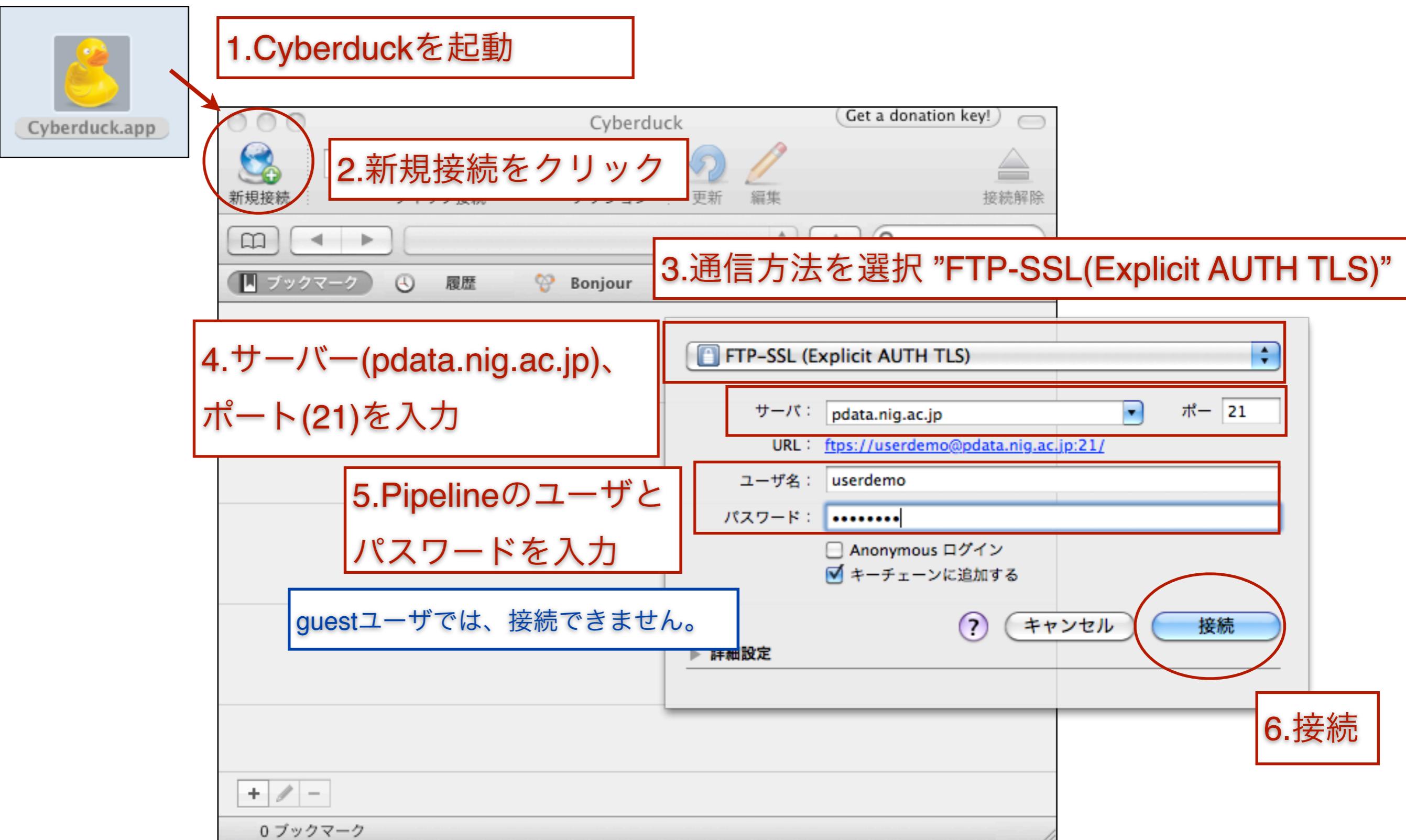
A red box highlights the "ダウンロード" link. A red box also highlights the "寄付" button at the bottom right.

2.ダウンロード

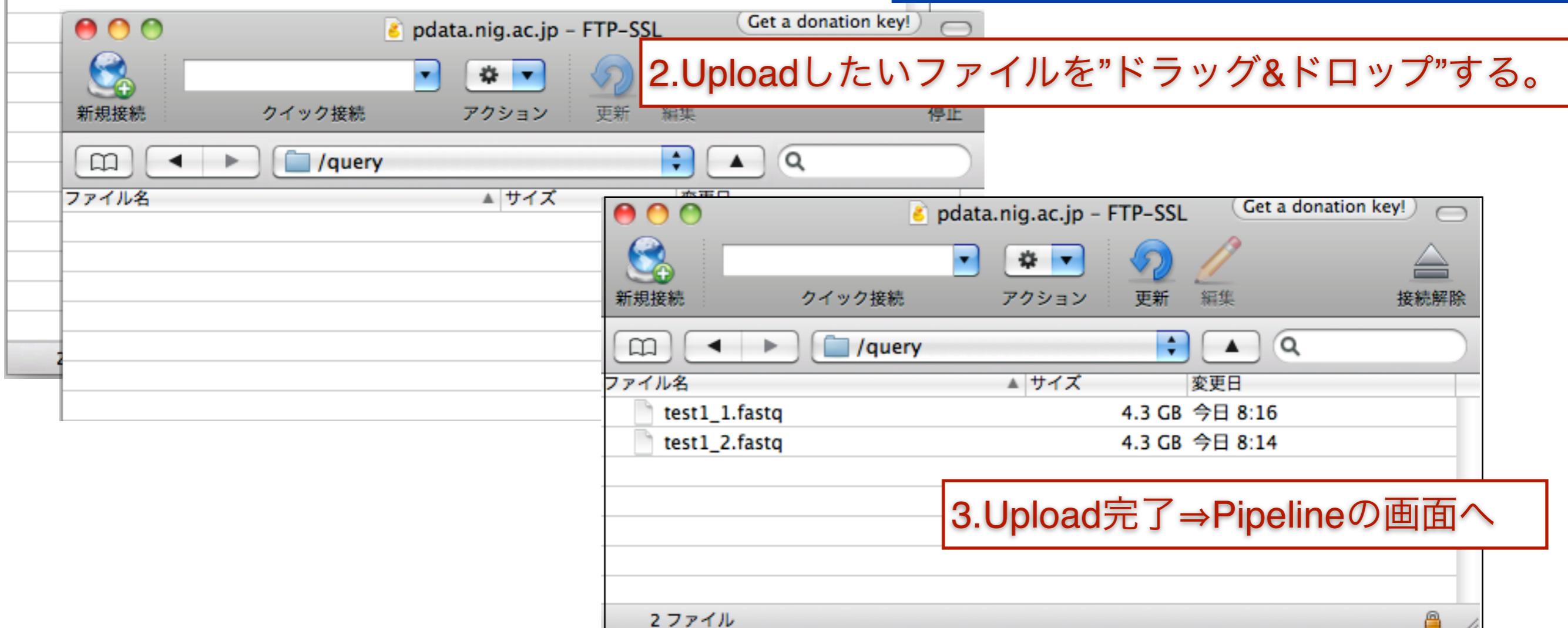
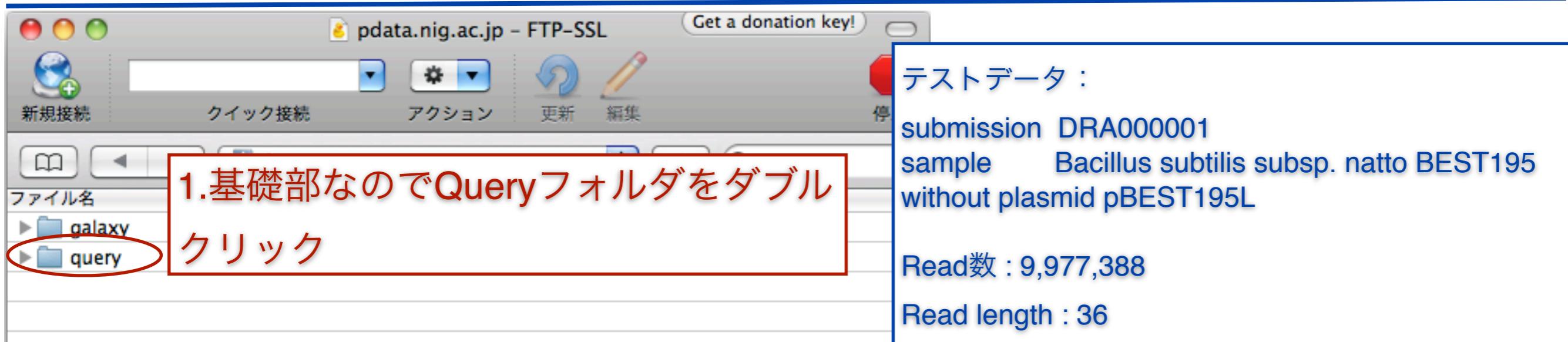


The screenshot shows the Cyberduck application window. The title bar says "Amazon S3 (HTTPS)". The address bar shows "https://0EZZWKSC9Y6B33ARJY02@s3.amazonaws.com/cyberduck.ch". The interface includes "Open Connection", "Quick Connect", "Action", "Edit", "Refresh", and "Disconnect" buttons. A "Bookmarks" tab is visible. The status bar at the bottom shows "Amazon S3 (HTTPS)" and the URL "https://0EZZWKSC9Y6B33ARJY02@s3.amazonaws.com/cyberduck.ch".

Query file指定方法 通信先サーバ情報を設定



Query file指定方法 Upload



Query file指定方法

Uploadしたファイルの注釈づけ 1

1.Pipelineの画面に戻る

UploadしたファイルがSingle-endの場合

2.Select a FASTA/FASTQ fileを選択

Registration of fastq/fasta files

Upload FASTA/FASTQ files [Select a FASTA/FASTQ file](#) Input a specification

Please specify read layout and correspondent files.

Read layout : [Single-end](#)

3.Single-endを選択

filename	type	size	timestamp
<input type="radio"/> Not select			
<input type="radio"/> test1_1.fastq	fastq	392.3 MB	2011-01-04 05:23:00
<input type="radio"/> test1_2.fastq	fastq	392.3 MB	2011-01-04 05:24:05
<input checked="" type="radio"/> test2.fastq	fastq	392.3 MB	2011-01-04 05:33:20

4.readを選択

Go to the next page after you select a file.

5.次へ

[Next STEP](#)

UploadしたファイルがPaired-endの場合

2.Select a FASTA/FASTQ fileを選択

Registration of fastq/fasta files

Upload FASTA/FASTQ files [Select a FASTA/FASTQ file](#) Input a specification

Please specify read layout and correspondent files.

Read layout : [Paired-end](#)

3.Paired-endを選択

filename	type	size	timestamp
<input type="radio"/> Not select			
<input checked="" type="radio"/> test1_1.fastq	fastq	392.3 MB	2011-01-04 05:23:00
<input type="radio"/> test1_2.fastq	fastq	392.3 MB	2011-01-04 05:24:05

4.read1 fileを選択

If you want a paired-end query, please select a "Reverse" query file.

Select a reverse FASTQ file:

filename	type	size	timestamp
<input type="radio"/> Not select			
<input type="radio"/> test1_1.fastq	fastq	392.3 MB	2011-01-04 05:23:00
<input checked="" type="radio"/> test1_2.fastq	fastq	392.3 MB	2011-01-04 05:24:05

5.read1 fileと対になるread2 fileを選択

Go to the next page after you select a file.

6.次へ

[Next STEP](#)

Query file指定方法 Uploadしたファイルの注釈づけ 2

Registration of fastq/fasta files

Upload FASTA/FASTQ files > Select a FASTA/FASTQ file > Input a specification

Please specify instrument model.

SelectedFile 1	test1_1.fastq
SelectedFile 2	test1_2.fastq
Read layout	Paired-end
Instrument model	ILLUMINA
(Required) Study title	test data

NOTICE: After confirming your entries, push the SUBMIT button to register uploaded files.

1. シーケンサの機種を選択

2. Study titleを入力

3. 登録

4. Assembly/Mappingの実行画面へ

Registration complete.
Press "Mapping / Assembly" button, to goto job input pages.

Assembly / Mapping

Query file指定方法

Uploadしたファイルの確認

Uploadしたファイルを使用して解析が可能になって
いる。

Selecting Query Files

1. Upload FASTA/FASTQ(FTP client)を選択

FTP upload Private DRA entry Import public DRA HTTP upload

NEXT

List of your uploaded files by FTP client. [\[Add new files\]](#)

	Filename	Description	Layout	Instrument model	File size
<input type="checkbox"/>	GSM727564_d0Foxh1.bed.gz	Foxh1	single	ILLUMINA	124.4 KB
<input checked="" type="checkbox"/>	test1_1.fastq (more 1 files)	test data	paired	ILLUMINA	3.4 GB

DELETE **NEXT**

2. 解析に使用したいファイルを選択

3. 次へ