

DDBJ Read Annotation Pipeline の紹介と実習

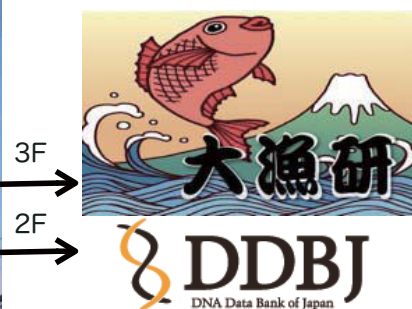
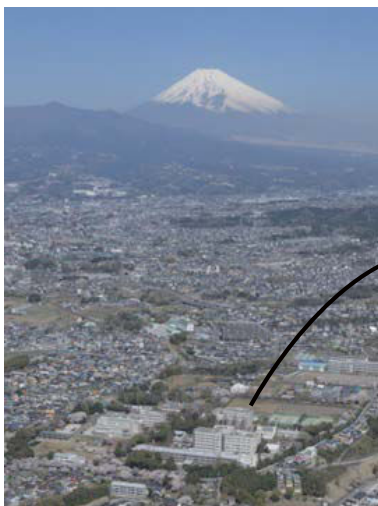
国立遺伝学研究所
大量遺伝情報研究室

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

高速シーケンス配列の情報解析

を行なっています。

高速シーケンサー配列の登場で短期間、低コストで大量の塩基配列データが出力されるようになった。



illumina社 HiSeq2000



Roche社 GS FLX



Life Technologies社 SOLiD

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

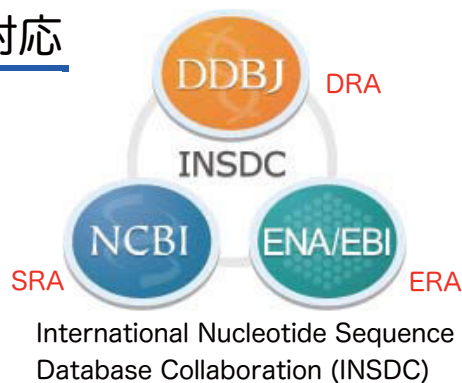
といった問題がでてきた。

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)

>中村、児玉の発表をご参考
ください。



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。

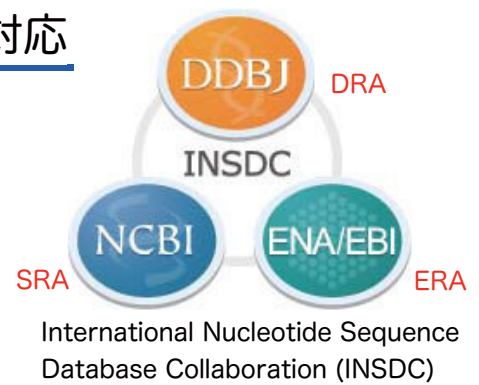
解析のための
人員不足

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)

>中村、児玉の発表をご参考
ください。



計算機不足

解析のための
人員不足

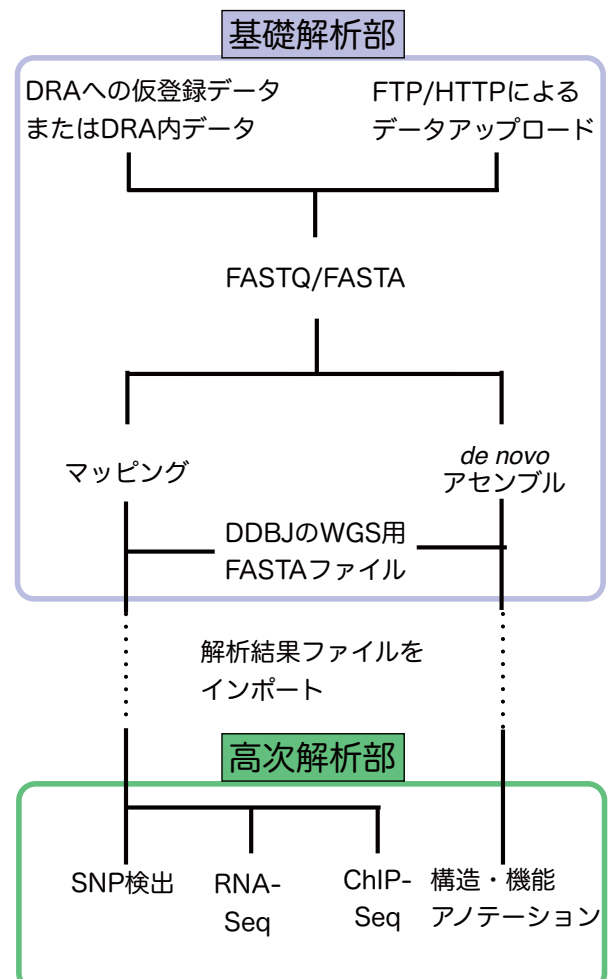
DDBJ Read Annotation Pipeline
(DDBJ パイプライン)

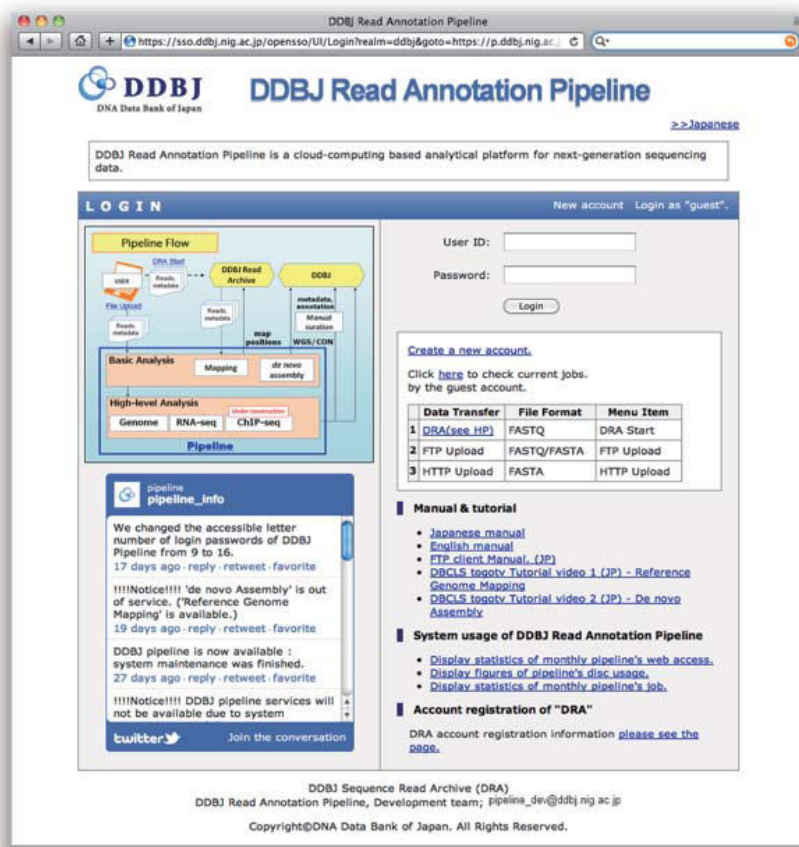
DDBJ パイプラインの特徴

・ 遺伝研の計算機で分散処理を実行、高速シーケン
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



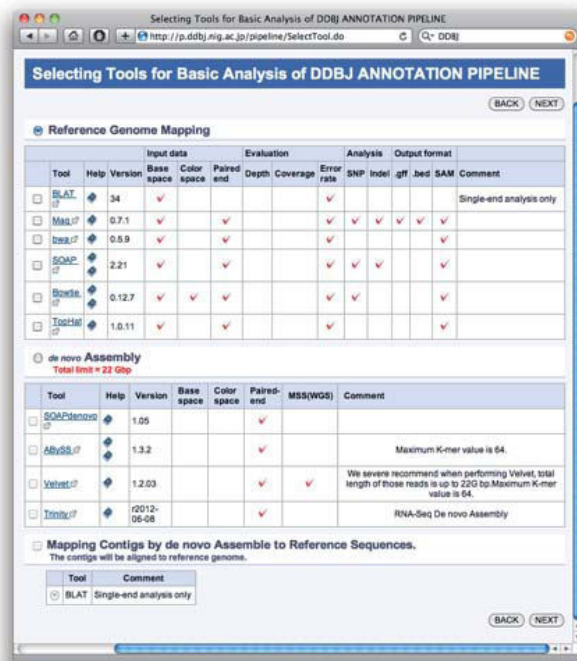


・11種類のマッピング・アセンブルソフト対応 マッピング

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

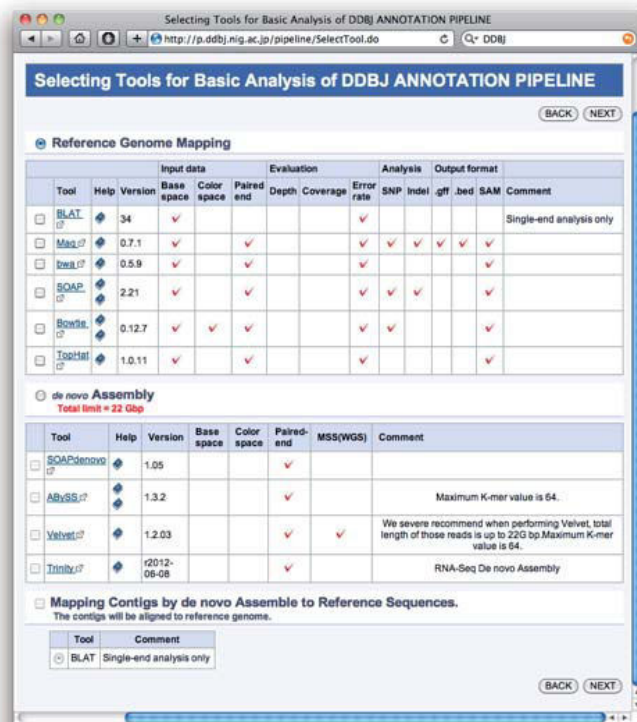
SOAPdenovo	ヒト、パンダ等大型ゲノムのアセンブリで使用された。比較的高速。
Abyss	初期に並列処理に対応したアセンブラ。
Velvet	高速シーケンサー登場初期に開発された。メモリ消費多め。
Trinity	RNA-Seq配列のアセンブラ。上記3つともにde bruijn graphというアルゴリズムを使用。



- ・ 11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



- ・ 11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

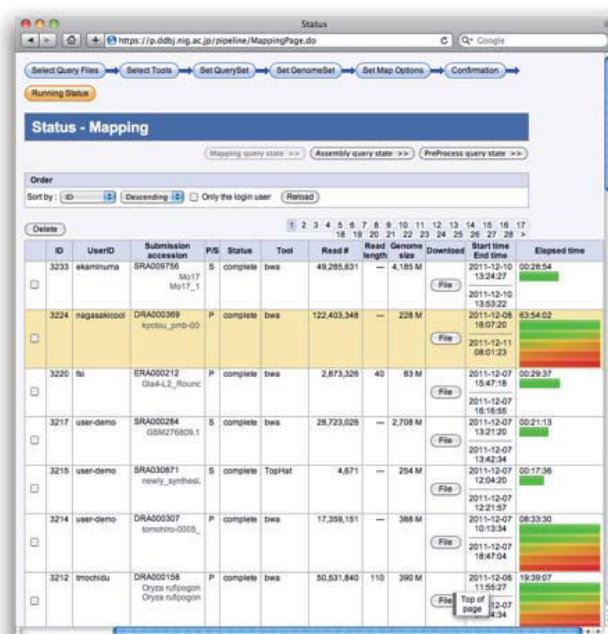
Intel Xeon 2.66GHz 768 cores, 10TB RAM

ストレージ

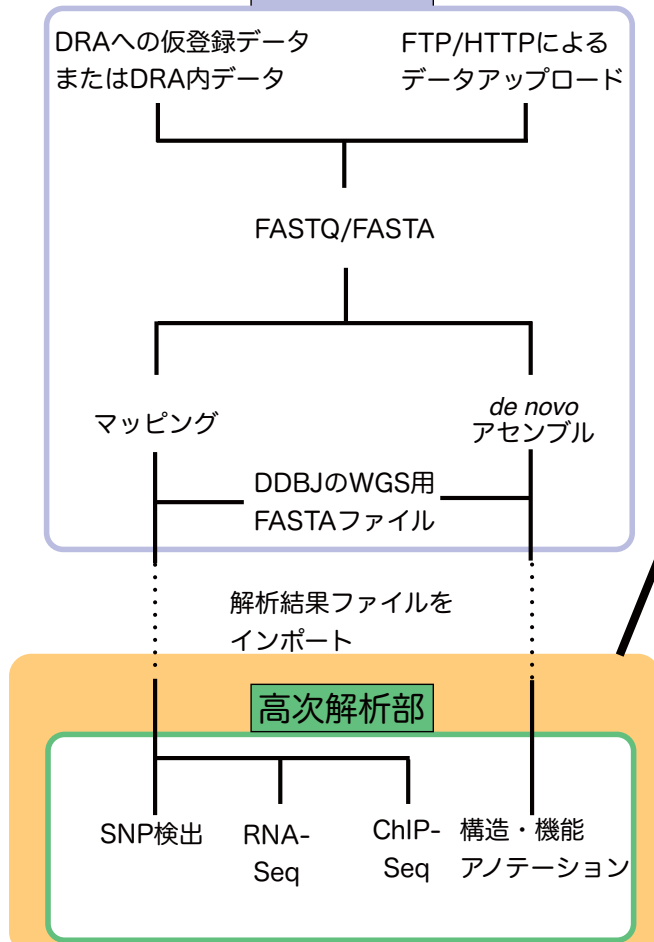
2PB storage

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



- ・ 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)

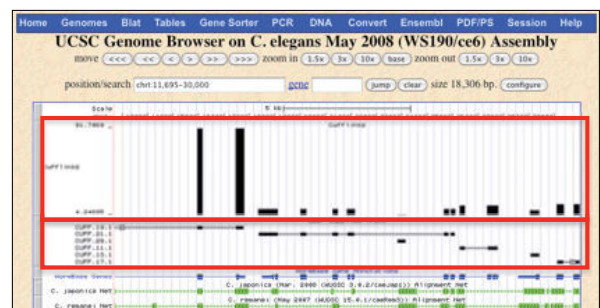
DDBJパイプライン 高次解析部

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



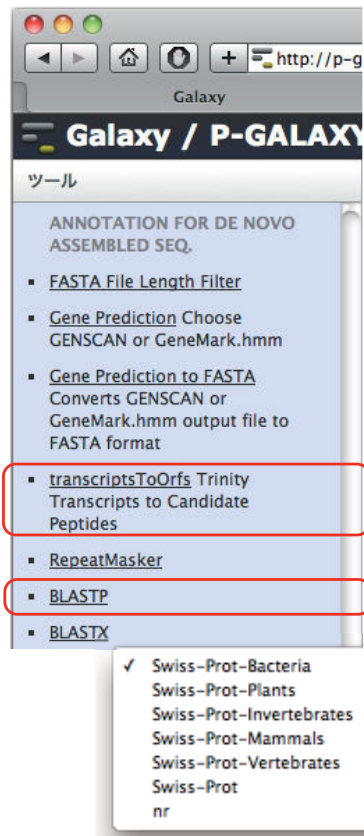
RNA-SeqのCufflinks実行(発現量の正規化)
gtf->wigフォーマット変換
UCSC genome browser siteでの可視化



(<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

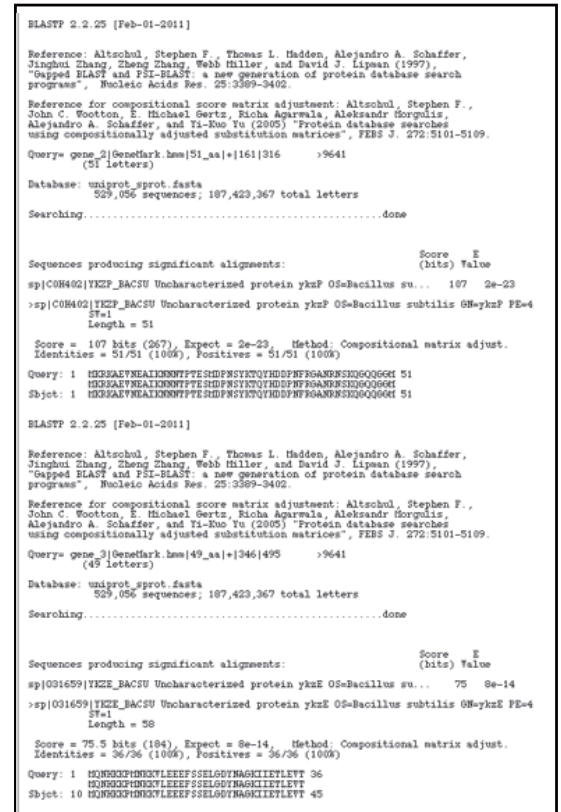
FASTAファイル

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



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:

スプライスバリエーション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った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ったスプライスバリエーションやパラログも再構成する。

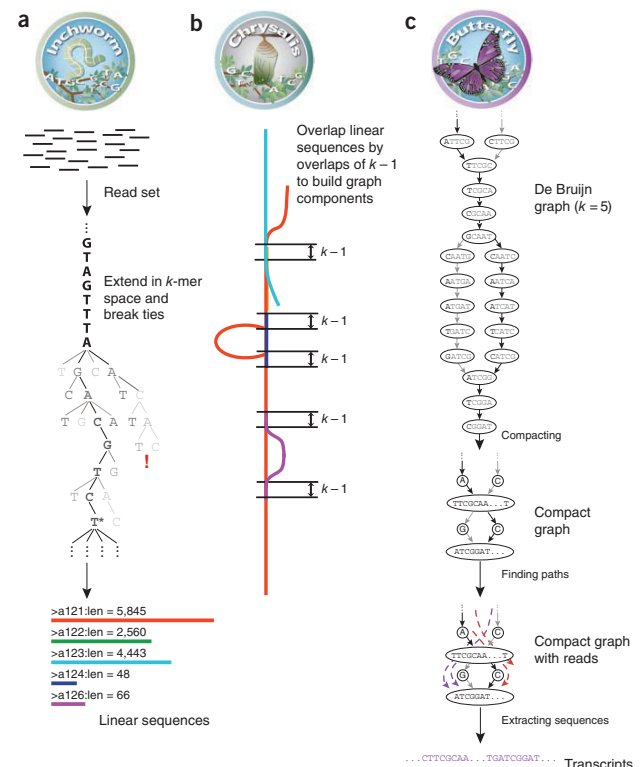


Figure 1 Overview of Trinity. (a) Inchworm assembles the read data set (short black lines, top) by greedily searching for paths in a k -mer graph (middle), resulting in a collection of linear contigs (color lines, bottom), with each k -mer present only once in the contigs. (b) Chrysalis pools contigs (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 outputs one linear sequence for each splice form and/or paralogous transcript represented in the graph (bottom, colored sequences).

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。

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
が多くの熱帯魚店等で売られている。

SRR042533 (エントリー: SRA012701)
36bpの7,468,448リード
シングルエンド

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

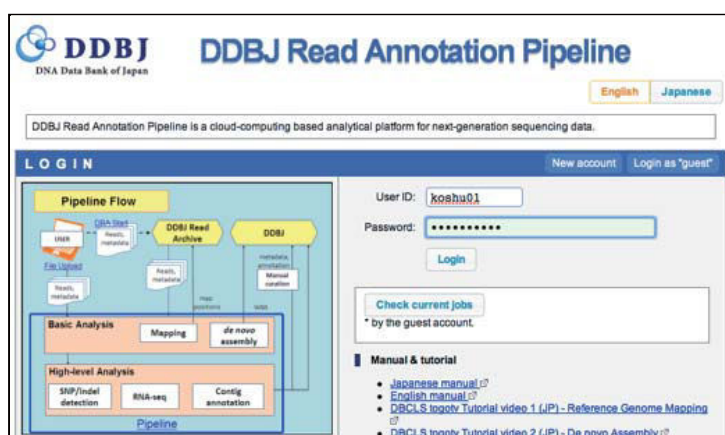
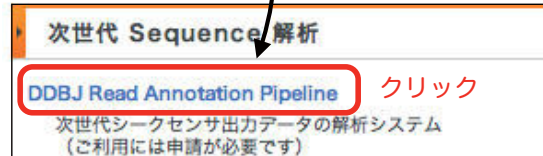
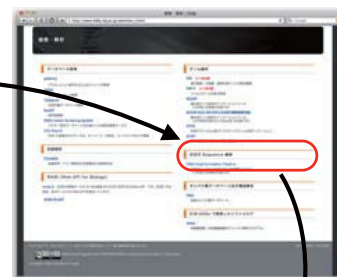
「Organism:」に「Tetraodon nigroviridis」と入力し、「Search」をクリック。

今回はアクセッション番号「SRA012701」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、アクセッションをメモしておく。



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

TYPE	ACCESSION	ALIAS	FILENAME	DL	VIEW
Submission	DRA000001		DRA000001.submission.xml	Download	View
Sample	DRS000001	Bacillus subtilis subsp. natto BEST195 without plasmid pBEST195L	DRA000001.sample.xml	Download	View
Study	DRP000001	Natto BEST195	DRA000001.study.xml	Download	View
Experiment	DRX000001	NATTO_BEST195_SEP08	DRA000001.experiment.xml	Download	View
Run	DRR000001	2008-09-12 BEST195-Lane7	DRA000001.run.xml	Download	View

「Input DRA/ERA/SRA Accession Number」に「SRA012701」と入力

「Add my DRA entry」をクリック

Input DRA/ERA/SRA Accession Number: SRA012701

Add my DRA entry

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もimport終了時にメールが届く。

「OK」をクリック。

Confirmation

Click a OK button to start import.

This operation may take several minutes to several hours.

Option

☒ Send a mail when completed importing.

☐ Show a accessions list.

OK Cancel

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のアクセスションが「queued」から「done」になったら完了。

Status	Submission	Request date
queued	DRA000307	2012-08-22 13:40:50.392

ブラウザリロードで確認

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ッション「SRAA012701」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)

①

②

③

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 ACCESSION20122 のものをチェック。

最下部の「NEXT」を押し、次画面に進む。

NEXT

No.	Experiment ACCESSION	Sample ACCESSION	Run ACCESSION	STRAIN	Run_date	Read #	Read length	Instrument model	Layout
1	SRX020112	SRS070561	SRR042523	strain n/a				ILLUMINA	single
2	SRX020113	SRS070562	SRR042524	strain n/a				ILLUMINA	single
3	SRX020114	SRS070563	SRR042525	strain n/a				ILLUMINA	single
4	SRX020115	SRS070564	SRR042526	strain Okayama 7				ILLUMINA	single
5	SRX020116	SRS070565	SRR042527	strain ATCC #64925				ILLUMINA	single
6	SRX020117	SRS070566	SRR042528	strain n/a				ILLUMINA	single
7	SRX020118	SRS070567	SRR042529	strain japonica cultivar Nipponbare				ILLUMINA	single
8	SRX020119	SRS070568	SRR042530	strain NRRL 1555				ILLUMINA	single
9	SRX020120	SRS070569	SRR042531	strain ATCC #11538				ILLUMINA	single
10	SRX020121	SRS070570	SRR042532	strain n/a				ILLUMINA	single
11	SRX020122	SRS070571	SRR042533	strain n/a				ILLUMINA	single
12	SRX020123	SRS070572	SRR042534	strain #UAMH 1704				ILLUMINA	single
13	SRX020124	SRS070573	SRR042535	strain UTEX #LB 1885				ILLUMINA	single
14	SRX020125	SRS070574	SRR042536	strain ATCC #50258				ILLUMINA	single
15	SRX020126	SRS070575	SRR042538	strain n/a				ILLUMINA	single

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データはすべて Phred+33 形式になっています。

☒ Phred+33 ☐ Phred+64 →

If you don't know it, please see ['2.2 Encoding' of this site](#).

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

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

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。

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。

Email notification

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

* Required

Confirmation of entries

Query sets

- SRR042533 - GSM497271_1

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。

Status - Preprocessing										
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	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

処理済みの FASTQ ファイルのダウンロード

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

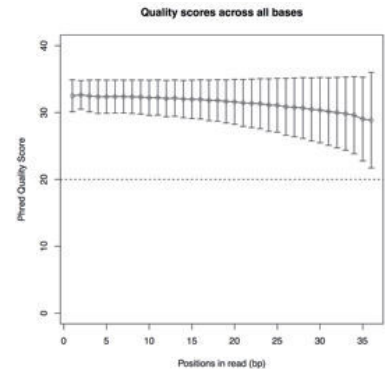
Wait time	Start time	End time
0: 0:59	2013-04-30 17:42:30	2013-04-30 17:56:24

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			

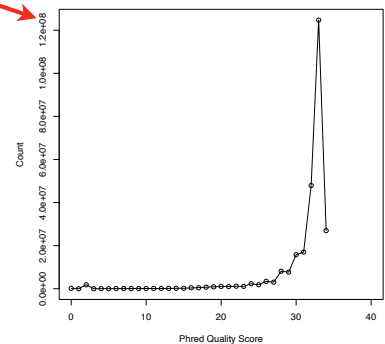
ログの確認

「BACK」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
Count of QS



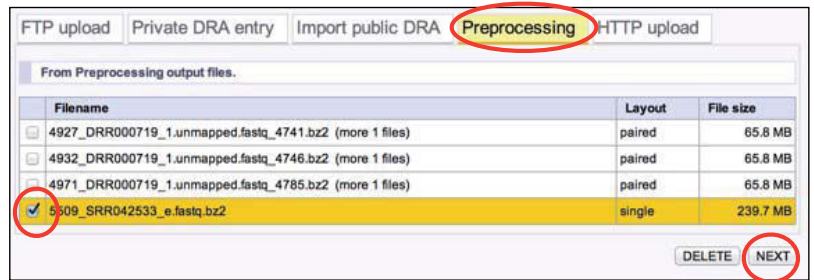
denovo Assembly
Trinity の実行

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



次へ

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。

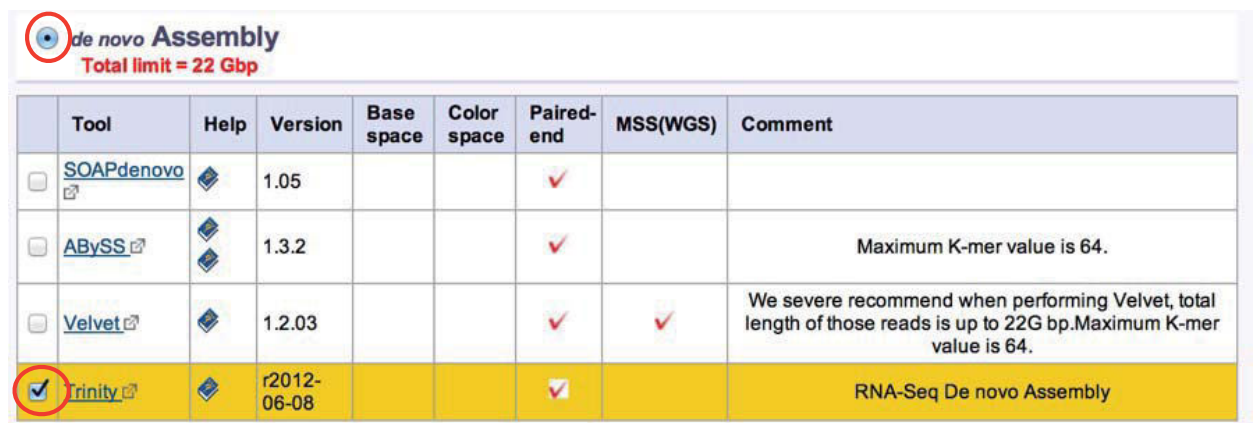
Preprocessing で処理されたファイルは、

「(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。

Trinityの実行 ツールの選択

「denovo Assembly」 → 「Trinity」の順に選択



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
☒ 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。

今回はデフォルトの条件で実行する、

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 --min_contig_length 201

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 100 out_WGS.txt

BACK NEXT

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
FASTA or FASTQ (自動で指定される)

メモリ、CPU 関係の指定(固定)

Trinity.pl --seqType fq --JM 100G --bflyHeapSpaceMax 4G --bflyGCThreads 1 --CPU 4

--single <クエリファイル名> --output <出力ディレクトリ名> --min_contig_length 201

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 --bityHeapSpaceMax 4G --bityGCThreads 1 --CPU 4

--single reads.fq --output output_dir --min_contig_length 201

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	UserID	Submission accession	P/S	Status	Tool	Read #	Read length	Assembly detail	Mapping detail	Start time End time	Elapsed time
☐	5516	---	Whole transcrip	S	running	Trinity	46,765,342	---			2013-04-30 18:26:32	
☐	5515	koshu01	SRA012701 GSM497271_1	S	complete	Trinity	7,468,448	---	View		2013-04-30 18:15:21	02:40:24
☐	5514	koshu01	SRR042533 by	S	complete	Trinity	7,420,316	---	View		2013-04-30 18:13:59	02:42:20
☐	5508	---	Drosophila RNA	S	complete	Trinity	18,524,700	---			2013-04-30 17:23:42	03:43:55
☐	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	Download	Read length	Alias
1819	5509_SRR042533_e.fastq.bz2	N.A. bp	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

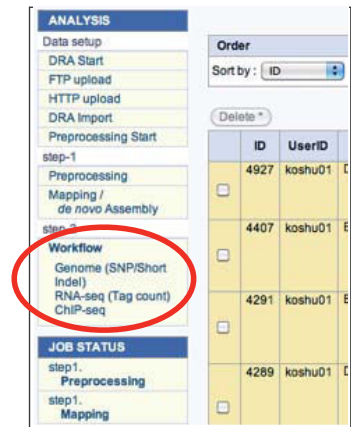
「BACK」ボタンで、一覧画面に戻る

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カラムから「step-2/Workflow」をクリック。

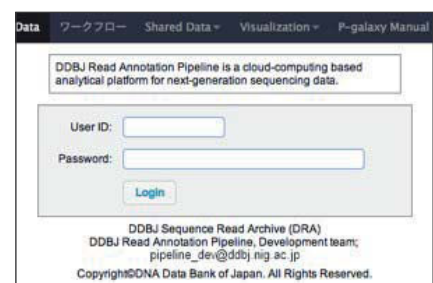
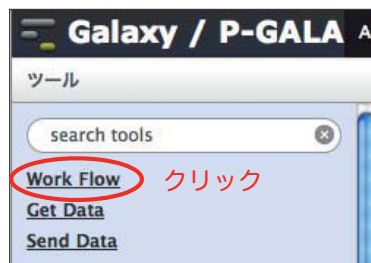


高次解析部(GALAXY)が起動



Tips:

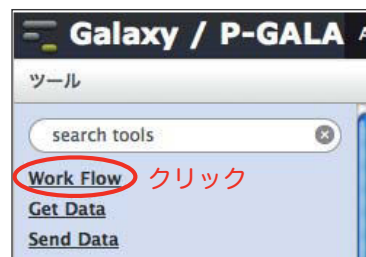
http://p-galaxy.ddbj.nig.ac.jpでURL直打ちして、「ツール」メニューの「Work Flow」をクリック。
基礎解析と同じパイプライン登録時のメールアドレスとパスワードを入力しても起動可能。



RNA-Seqのアセンブル結果をインポート

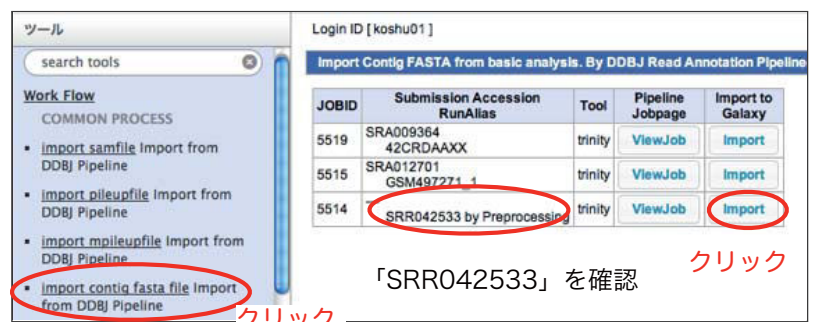
TrinityによるRNA-Seqのアセンブル結果をGALAXYにインポートする。

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 from 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

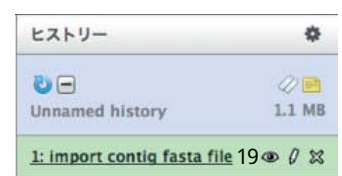


「SRR042533」を確認

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。

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

Number of CPUs to use by hmmscan

Execute

結果1

```

>m.565 g.565 ORF g.565 m.565 type:internal len:207 (-)...
DLEWQIEGLKEELIFLKKHHEELAMRAQMSGGVHVEVPAEDLTKVMADITREHYE
ITAKNQKLETWFSNKSSEALNKKMMHTVTTLQTSRSEVTEVKRSLOALQIELESLLGMKA
SLEGTLDQTPNRNYSMLAQYQQOQVTSLEQQLVQLRADLVRGQGDYQMLDIKTRLELEIA
EYRRLLEGEAAASSSTSSSTKTRRL
>m.566 g.566 ORF g.566 m.566 type:complete len:216 (+)...
MAQSVVPMFKLVLDGGGTGKTTFVKRHLTGFEKKYVATLGVEVHPLFPNTNRGNVVF
NVWDTAGQEGFGLRDGGYIQAQCAIIMFDVTSRVTYKNPNHRDLVRVCENIPVLGC
NKVDIKDRKVKAKSIVFHRKKNLQYYDISAKSNYNFEKFLWLARKLIGDPNLEFVEMPA
LAPPEVTMDPALAVQYKEKELHVASQTLAPDDEDDL*
>m.568 g.568 ORF g.568 m.568 type:internal len:227 (-)...
GDRFKEDRKAKRLPEKSIDMILLTDGDPNSGESRIPVIQENVKAAIGGQMSLFSLGFGN
DVKYPLDVMSENNGARRIYEGSDAALQLGFDYVSSPLLDDVLDLRYPDNAVDSLTT
NQFSQLFNGSEIVVAGRLKNDINDNFVEVFGQGLNDFSEQQQFSVLDSWGMYPDDDYIF
GDFTERLWAYLTIQQLLDKSKTGAEEKANASAEALDMSLRYSFVTP
>m.571 g.571 ORF g.571 m.571 type:Sprime_partial len:394...
ASGGEHTSSCGSWFNAGKDFPSVPYSYLDNFYCKTSSGEIESYHDVHQVRDCLRLVS
LLDLALEKDYVRGVADYMNRLVDMGVAGVRACKHWPGLDSAVYGRLLNLTWKWFFE
GSRPFIQEVIDLGGESISYTVYVHLGRVTEFKYGAKLGTVERKWNNEKLMYTKNWGEW
GFMPPNGNAVVFIDNHQNRHGAGGAALVTFWDSRLHMAVAYMLAHPYGVTRVNSFRW
NRHIVNGKDQNDWMGPPSPDGSKTSVPINPDETCDGQGWVCEHRWRQIKNMVIFRNVNG
QPHSNWDDNNSQVAFGRNGRGIIFNNDWDLDVTLMTGLPAGTYCDVISGQKEAGRCT
GRQIHVSGSDGAIFHFIENDEDFPFAIVEKSL*
>m.573 g.573 ORF g.573 m.573 type:Sprime_partial len:224...
WEPSPWQVSLQEYTGPHFCGSLNENWVVAACHNVTSHRVLGEHDSRNSNENIOV
MQVGVKPHNYNYSYITNDITLIKLASPAQLNIRVSPVCAETSVDVFGGMKCVTSGWG
LTRYNAPDTPRLQVALPLTLNTECRKHGWSKITDLMVCAGASGSSCMGDSGGLVCE
KAGAWTLVIGVSWSGFCSVSSGPPYARVMTLRAMMDQITIAAN*

```

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	dm	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

「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

「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。

BLASTP

クリック

BLASTP (version 1.0.0)

Query file: 4: transcriptsToOrfs...e Sequ

Select database: Swiss-Prot-Vertebrates

Expectation value: -20

Execute

Swiss-Prot-Bacteria

Swiss-Prot-Plants

Swiss-Prot-Invertebrates

Swiss-Prot-Mammals

Swiss-Prot-Vertebrates

Swiss-Prot

nr

7: BLASTP error/warning reports

6: BLASTP on data 4

94,939 lines

フォーマット: txt, データベース: 2

情報: /usr/local/bin/blastall -a 10 -p blastp -i /home/w3galaxy/galaxy-refs/database/files/002/dataset_28_70.dat -o /home/w3galaxy/galaxy-refs/database/files/002/dataset_28_72.dat -d /home/w3galaxy/galaxy-refs/data/blastxdb/unirot_swisspro_t_vertebrates.fasta -e

BLASTP 2.2.26 [Sep-21-2011]

Reference: Altschul, Stephen F., The

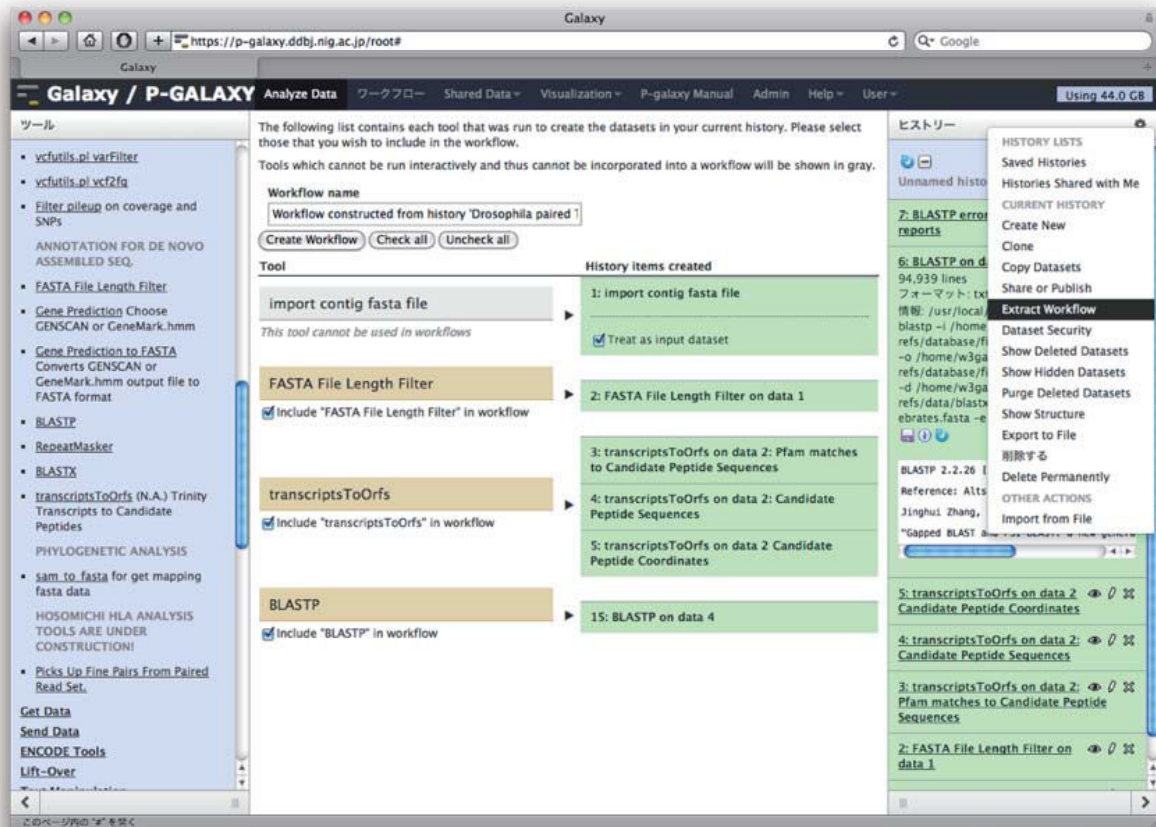
Jinghui Zhang, Zheng Zhang, Webb Mi

"Gapped BLAST and PSI-BLAST: a new c

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



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

「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



Query file指定方法 FTP Upload画面へ遷移

1.メニューのFTP Uploadをクリック

入力ファイルの指定方法4種類

	FILENAME	DL	VIEW	
Submission	DRA000001	DRA000001.submission.xml	Download View	
Sample	DRS000001	Bacillus subtilis subsp. natto BEST195 without plasmid pBEST195L	DRA000001.sample.xml	Download View
Study	DRP000001	Natto BEST195	DRA000001.study.xml	Download View
Experiment	DRX000001	NATTO_BEST195_SEPO8	DRA000001.experiment.xml	Download View
Run	DRR000001	2008-09-12.BEST195-Lane7	DRA000001.run.xml	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
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/fastq files

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

Please upload the file to be used.

To use your fasta or fastq files, needs to upload to our server using ftp system.
For security this ftp server is using FTP over SSL protocol. Therefore, please use FTP client that supports the file transfer protocol FTPS.

FTP settings.

Server : Port	p.ddbj.nig.ac.jp:21
Security	SSL
User ID/password	Your Pipeline login ID/password

FTP client software.

Windows	WinSCP, FileZilla
MacOSX	FileZilla, Cyberduck etc...

The upload directory is not open to the other users.
FTP transfer is secured by ssl.
Go to the next page after you upload a file.

Next STEP

1.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

2.データ転送後、次へ

Query file指定方法

FTP client (Cyberduck)のインストール

FTP client Cyberduckの場合

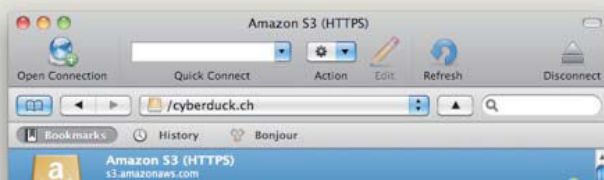


Cyberduck
オープンソースのFTP、SFTP、WebDAV、Cloud Files、
Google Docs、Amazon S3用ブラウザ、MacとWindowsに対応。

[Cyberduckについて](#) | [ニュース](#) | [更新履歴](#) | [開発](#) | [ヘルプ](#) | [寄付](#)

Macをリモートコントロール 遠くのMacも遠隔操作 ソフトのダウンロードは無料！ [www.TeamViewer.com](#)
[FTP Site - Free Trial](#) 1,515,353+ Business Users Trust Our Award-Winning Solution. Free Trial! [Unlimited ShareFile.com](#)
[MacOS X Snow Leopard 登場](#) さらに進化 より高速に もっと簡単に 世界で最も先進的なOS。今すぐ購入 [store.apple](#).
Ads by Google

あらゆるサーバに接続。
使いやすいインターフェースで、FTP (File Transfer Protocol)、FTP/TLS (FTP secured over SSL/TLS)、SFTP (SSH Secure File Transfer)、WebDAV (Web-based Distributed Authoring and Versioning)、Amazon S3、Google Storage、Google Docs、Windows Azure、Rackspace Cloud Filesに接続。Eucalyptus や OpenStackのオープンな最新クラウドソフトウェアも利用できます。



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
Downloads hosted by [Carbotix CDN Open Source Content Delivery](#)

寄付
VISA

1.http://cyberduck.ch/へアクセス

2.ダウンロード

Query file指定方法 通信先サーバ情報を設定

1. Cyberduckを起動

2. 新規接続をクリック

3. 通信方法を選択 "FTP-SSL(Explicit AUTH TLS)"

4. サーバ(133.39.116.60)、ポート(21)を入力

5. Pipelineのユーザとパスワードを入力

guestユーザでは、接続できません。

6. 接続

Query file指定方法 Upload

1. 基礎部なのでQueryフォルダをダブルクリック

テストデータ :

submission DRA000001

sample Bacillus subtilis subsp. natto BEST195 without plasmid pBEST195L

Read数 : 9,977,388

Read length : 36

2. Uploadしたいファイルを"ドラッグ&ドロップ"する。

3. Upload完了⇒Pipelineの画面へ

ファイル名	サイズ	変更日
test1_1.fastq	4.3 GB	今日 8:16
test1_2.fastq	4.3 GB	今日 8:14

2 ファイル

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/fastq files

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

Please specify read layout and correspondent files.

Read layout: **Single-end**

Select a FASTA/FASTQ file:

filename	type	size	timestamp
☐ Not select			
☐ test1_1.fastq	fastq	392.3 MB	2011-01-04 05:23:00
☐ test1_2.fastq	fastq	392.3 MB	2011-01-04 05:24:05
☒ test2.fastq	fastq	392.3 MB	2011-01-04 05:33:20

Go to the next page after you select a file.

Next STEP

3. Single-endを選択

4. readを選択

5. 次へ

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/fastq files

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

Please specify read layout and correspondent files.

Read layout: **Paired-end**

Select a FASTA/FASTQ file:

filename	type	size	timestamp
☐ Not select			
☒ test1_1.fastq	fastq	392.3 MB	2011-01-04 05:23:00
☐ test1_2.fastq	fastq	392.3 MB	2011-01-04 05:24:05

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

Select a reverse FASTQ file:

filename	type	size	timestamp
☐ Not select			
☐ test1_1.fastq	fastq	392.3 MB	2011-01-04 05:23:00
☒ test1_2.fastq	fastq	392.3 MB	2011-01-04 05:24:05

Go to the next page after you select a file.

Next STEP

3. Paired-endを選択

4. read1 fileを選択

5. read1 fileと対になるread2 fileを選択

6. 次へ

Query file指定方法

Uploadしたファイルの注釈づけ 2

Registration of fastq/fastq 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.

SUBMIT

1. シークエンサの機種を選択

2. Study titleを入力

3. 登録

Registration complete.

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

Assembly / Mapping

4. Assembly/Mappingの実行画面へ

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
☐	GSM727564_d0Foxh1.bed.gz	Foxh1	single	ILLUMINA	124.4 KB
☒	test1_1.fastq (more 1 files)	test data	paired	ILLUMINA	3.4 GB

2.解析に使用したいファイルを選択

DELETE

NEXT

3.次へ