

Nicotiana tabacum

Zea

Salmo salar
Larus argentatus
Acacia greggii

Rattus



DDBJパイプラインによる RNA-seq配列のde novoアセンブル 前座

中村 保一

NAKAMURA Yasukazu, Professor

国立遺伝学研究所 大量遺伝情報研究室

& DDBJセンター

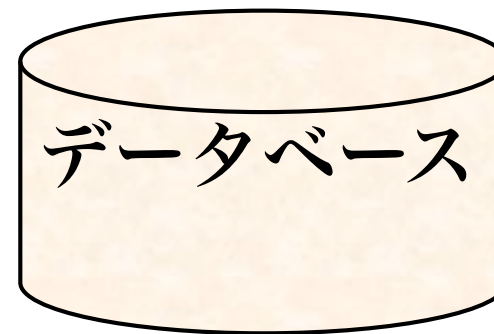
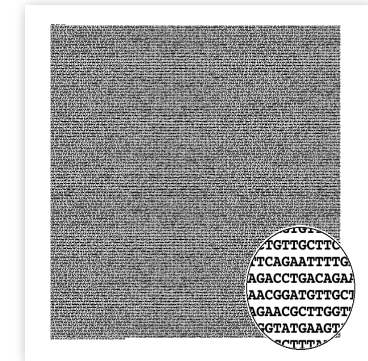
日本DNAデータバンク

DDBJ

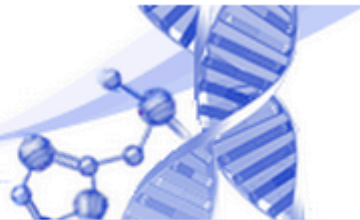
DNA Data Bank of Japan

塩基配列データベースとはこのような事業

- 📌 全世界で解読された塩基配列情報を
- 📌 査定して受入れ
- 📌 データベースに蓄積し
- 📌 公開して共有する



DDBJ (<http://www.ddbj.nig.ac.jp/>)



[ENGLISH](#)  

サイト内検索

- HOME
- 塩基配列の登録
- 利用の手引き
- 検索・解析
- FTP・WebAPI
- レポート・統計
- お問い合わせ

▶ DDBJの紹介

▶ Q&A集

▶ 塩基配列の登録

- ▶ [SAKURA](#)
- ▶ [DDBJ塩基配列登録システム](#)
- NEW
- ▶ [大量登録システム\(MSS\)](#)
- ▶ [データの修正・更新](#)
- ▶ [DDBJ Sequence Read Archive](#)
- ▶ [DDBJ Trace Archive](#)

▶ プロジェクトの登録

▶ [DDBJ BioProject Database](#)

▶ スーパーコンピュータ利用

▶ [スパコンの利用申込](#)

DDBJ : DNA Data Bank of Japan

DDBJ (日本DNAデータバンク) は欧州と米国の対応機関 (EBIおよび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る三大国際DNAデータバンクのひとつです



Photo by Hideki Nagasaki

Hot Topics [▶ 一覧へ](#)

- 2012.10.19 [DDBJリリース 90.1, DAD リリース 60.1 公開](#)
- 2012.10.15 [書籍「次世代シーケンサー：目的別アドバンスドメソッド」の紹介](#)
- 2012.10.12 [DDBJとDADのデータ不備についてのお詫び](#)

Maintenance [▶ 一覧へ](#)

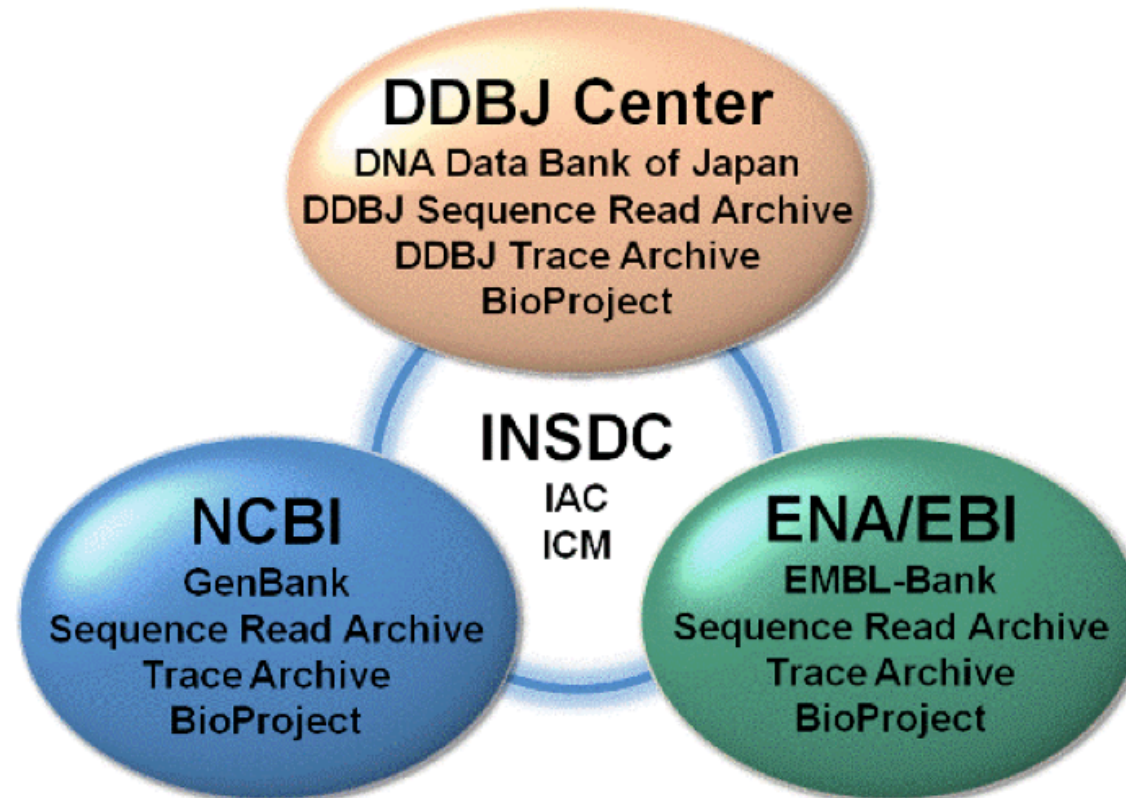
- 2012.09.11 (10/3) [国立遺伝学研究所ならびにDDBJネットワークの一時停止](#)
- 2012.08.28 [\(変更\) SAKURA から新登録システムへ切り替えのお知らせ](#)

国際塩基配列データベース (INSDC) の一員

📍 米国: GenBank (NCBI)

📍 欧州: ENA (EBI)

📍 日本: DDBJ



DDBJ登録ファイルの例

LOCUS AB091058 2109 bp DNA linear BCT 02-SEP-2003
DEFINITION *Gluconacetobacter xylinus* cmcase, ccp genes for endo-beta-1,4-glucanase, cellulose complementing protein, complete cds.
ACCESSION [AB091058](#)
VERSION AB091058.1
KEYWORDS .
SOURCE *Gluconacetobacter xylinus*
ORGANISM [Gluconacetobacter xylinus](#)
Bacteria; Proteobacteria; Alphaproteobacteria; Rhodospirillales; Acetobacteraceae; *Gluconacetobacter*.
REFERENCE 1 (bases 1 to 2109)
AUTHORS Kawano,S., Tajima,K., Uemori,Y., Yamashita,H., Erata,T., Munekata,M. and Takai,M.
TITLE Direct Submission
JOURNAL Submitted (28-AUG-2002) to the DDBJ/EMBL/GenBank databases. Contact:Kenji Tajima
Hokkaido University, Graduate School of Engineering; N13W8, Kita-ku, Sapporo, Hokkaido 060-8628, Japan
REFERENCE 2
AUTHORS Kawano,S., Tajima,K., Uemori,Y., Yamashita,H., Erata,T., Munekata,M. and Takai,M.
TITLE Cloning of Cellulose Synthesis Related Genes from *Acetobacter xylinum* ATCC23769 and ATCC53582: Comparison of Cellulose Synthetic Ability Between ATCC23769 and ATCC53582
JOURNAL Unpublished (2002)
COMMENT
FEATURES Location/Qualifiers
source 1..2109
/db_xref="taxon:28448"
/mol_type="genomic DNA"
/note="synonym:Acetobacter xylinum"
/organism="[Gluconacetobacter xylinus](#)"
/strain="ATCC 53582"
CDS 10..1038
/codon_start=1
/gene="cmcase"
/product="endo-beta-1,4-glucanase"
/protein_id="[BAC82540.1](#)"
/transl_table=11
/translation="MSVMAAMGGAQVLSSTGAFADTAPDAVAQQWAI FRAKYLRPSGR VVDTGNGGSESHSEGGYGLMFAASAGDLASFQSMWMMWARTNLQHTNDKLFWRFLKGH QPPVPDKNNATDGDLLIALALGRAGKRFQRPDIQDAMAIYGDVNLNMTMKAGPYVVL MPGAVGFTKKDSVILNLSYYVMPSSLQAFDLTADPRWRQVMEDGIRLVASGRFGQWRL PPDWLAVNRATGALSIA SGWPPRFSDAIRVPLYFYWAHMLAPNVLADFTRFWNNFGA NALPGWVDLTTGARSPYNAPPGYLVAECTGLDSAGELPTLDHAPDYSAALTLLVYI ARAEETIK"

CDS 1035..2096
/codon_start=1
/gene="ccp"
/product="cellulose complementing protein"
/protein_id="[BAC82541.1](#)"
/transl_table=11
/translation="MSASGSDEVAGGGQAGSPQDFQVRLRSFGVEGGQYSYRPFVDRS FVDTGVPEAVERHFDQAEHDTAVEEQVTPAPQIAVAPPPPPVPPPAIVTETAPPPP VVVSAPVTYEPAAAAPAEPPVQEAAPVQAPVPPAPVPPVIAEQAPPAAPDPASVPYAN VAAAPVPPDPAPVTPAPQARVTPGPNTRMVEPFPSRPQVVRTVQEGATPSRVPSRSMNAFP RTSASSISERPVDGRVADEWSPVPKARLSRPRPRPGDLSFFQGMRDTRDEKFFPV ASTRSVRSNVSRMTSMTKTDTNSSQASRPGSPVSPDGSPMTAEVFMFLGGRATELLS PRPSLREALLRRENEES"
BASE COUNT 343 a 661 c 661 g 444 t
ORIGIN
1 cgttccttta tgtcggatcat ggcggcgatg ggagggcgcg aggtgctttc atccaccggt
61 gcgttcgcag acaccgcccc cgatgcggtc gcgcagcaat gggccatctt ccgcgccaaag
121 tatcttcgct ccagcggacg tgtcgtggat acgggcaatg gtggcgaatc ccatagtgag
181 gggcagggct atggcatgct ctttcgccgcg tcggcggggg acctgctgct gttccagtcg
241 atgtggatgt gggcgcgcac caacctgcag cataccaatg acaagctggt ttcttgccgg
301 ttctcaagg ggcacagccg cccggtgccc gacaagaaca atgccacaga tggcgcacctg
361 ctgatcgcgc ttgcgcttgg ctgctgcggc aagcgtttcc agcgccccga ttacattcag
421 gacgccatgy ccatttatgg cgatgtgctg aacctgatga cgatgaagcg ggagccgtat
481 gtcgtcctca tgcccggctg tgtcggcttt accaagaagg acagcgtgat cctcaacctg
541 tctattacg tcatgcctc gctgctgcag gcgctcgacc ttacggccga cccgcgctgg
601 cgctcaggtg tggagacgg gattcgcctt gttccgcgcg gccgcttcgg gcagtggcgc
661 ctgccccccg actggctggc ggtgaatcgc gccaccgggt cgctgctgat cgcacggga
721 tggccgcgcg gcttttccca tgatgcatg cgggtgccc tttattttta ttgggcgat
781 atgctggcgc cgaacgtggt ggctgattc acccgattct ggaataattt cggggctaata
841 gccctgccag gatgggttga tctgacaaca gggcgcgctt cgccgtacaa cgccccgct
901 ggatattctg ctggtgccga atgcacgggg cttgattctg ccggggaact cccgacactg
961 gatcatgccc cggattatta ttccgcagcg ttgacgctgc tcggttacat cgcgcggcg
1021 gaggagacta taaagtgagt gcttcagggt ctgatgaggt ggctggggga gggcaggctg
1081 gaagtccgca ggattttcag cgggtcctgc gttcttttgg tgtcgaaggt gggcagatt
1141 cctaccggcc gtttggtagc cgttcctttg atgtgacagg cgtgccccga gctgttga
1201 ggcacttcga tcaggcggag catgacacgg cggttgagga caggtcact cccgcgccac
1261 aaatcgcggt cgaccgcca cgcaccgcca ccccgcccga tcggtcctga cccgcgcc
1321 aaaccgccc cccgcgcct gtcgtggtca gcgctccggt caggtatgaa cccccgctg
1381 ccgcccgtgc ggacagcct cccgttcagg aagccccggt gcagccggcg ccggttcccc
1441 ccgcccgtgc gcccccgat gcggagcagg ctctcccgcg ggcgcccga ccggcatccg
1501 tgccgtatgc gaactgcgcg gcagaccgcg ttccacctga tcccgcaccg gttacgctg
1561 cgccgcagc gcgctgacg gggccgaaca cccgtatggt ggagcccttt tcccgcgcg
1621 aggtccgcac ggtgcaggag ggggcaacc cgtcacgtgt accttcgctg tcaatgaacg
1681 ctttcccccg cacatcagca tcgtccataa gtgagcgtcc ggtgacagc ggtgttgccg
1741 atgaatggag tctgtttccg aaggcagccc tcagcccgcg ggagcgtccg cgtcccgcg
1801 atctgagctt ttctttcag gggatgccc acaccctga tgaaaagaag ttctttccg
1861 tggcgtccac gcgatcagtt cgttctaag ttccaggat gaccagcatg accaagacag
1921 acacgaattc ctctcaggt tctcgtccc gcagcccgt cgctcgcct gatgggtcgc
1981 ccacaatggc cgaagtgtc atgacgctg gtggtcgtgc gacggaactc ctacgcccc
2041 gtccttcgct gcgggagcgc ctggtgcgct gtcgtgaaa cgaagaaga tctaaggcc
2101 ctatattca

遺伝子・立体構造の論文には登録が不可欠



PLoS BIOLOGY
a peer-reviewed open-access journal published by the Public Library of Science

Login | Create Account | Feedback

Search articles... GO Advanced Search

Browse RSS

Home Browse Articles About For Readers For Authors and Reviewers Journals Hubs PLoS.org

Accession Numbers

All appropriate datasets, images, and information should be deposited in public resources. Please provide the relevant accession numbers (and version numbers, if appropriate). Accession numbers should be provided in parentheses after the entity on first use. Suggested databases include, but are not limited to:

- > [ArrayExpress](#)
- > [BioModels Database](#)
- > [Database of Interacting Proteins](#)
- > [DNA Data Bank of Japan \[DDBJ\]](#)
- > [DRYAD](#)
- > [EMBL Nucleotide Sequence Database](#)
- > [GenBank](#)
- > [Gene Expression Omnibus \[GEO\]](#)
- > [Protein Data Bank](#)
- > [UniProtKB/Swiss-Prot](#)
- > [ClinicalTrials.gov](#)

論文投稿時の注意：論文の著者は、論文で言及した塩基配列や立体構造などのデータについて、インターネットで参照可能な公共データベースの登録番号を掲載しなければならない

INSDCに多くの配列が登録された生物種

Images created by the Wordle.net web application are licensed under a Creative Commons Attribution 3.0 United States License.



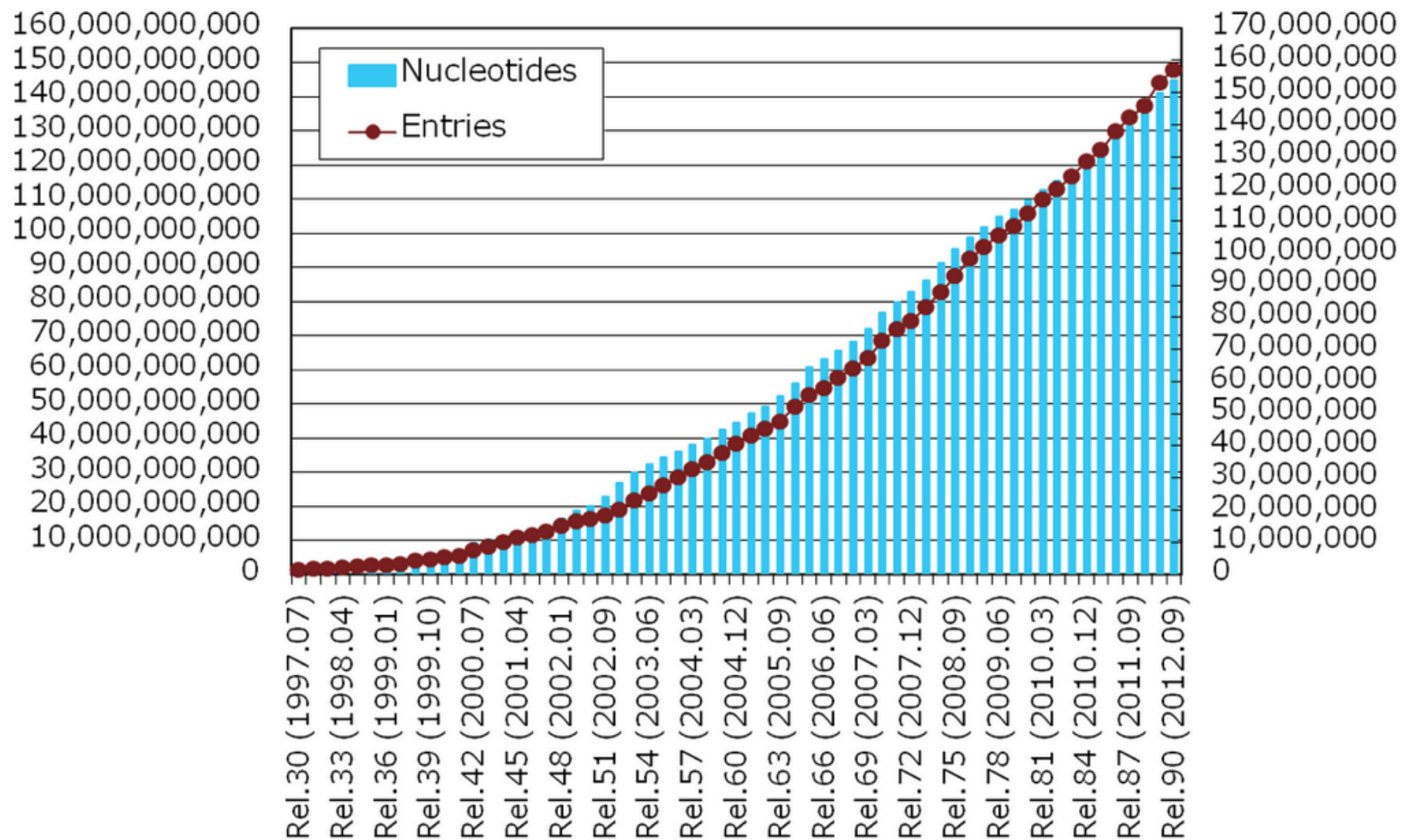
DDBJに登録されている生物種 Top 100の
ワードクラウド（数が多いほど大きい字
で表示）

INSDC塩基配列データの量

DDBJ/EMBL/GenBank database growth

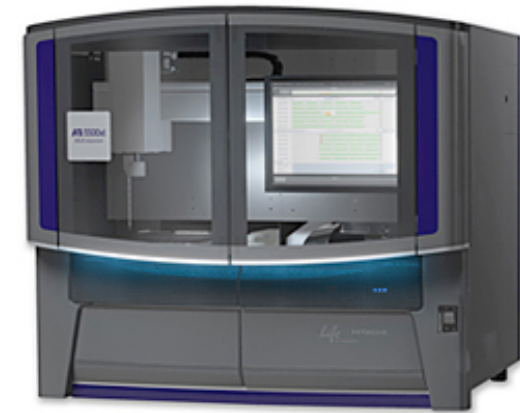
塩基数: 1,500億

登録数: 1.6億



生物学の 情報爆発

代表的な超並列DNA配列決定装置



(左) Roche (454): GS FLX+ System

(中) illumina: Genome Analyzer IIx System

(右) Life Technologies: 5500 xl SOLiD System

新型シーケンサーの特徴：高速・大量



http://www.illumina.co.jp/systems/hiseq_systems.ilmn
より引用

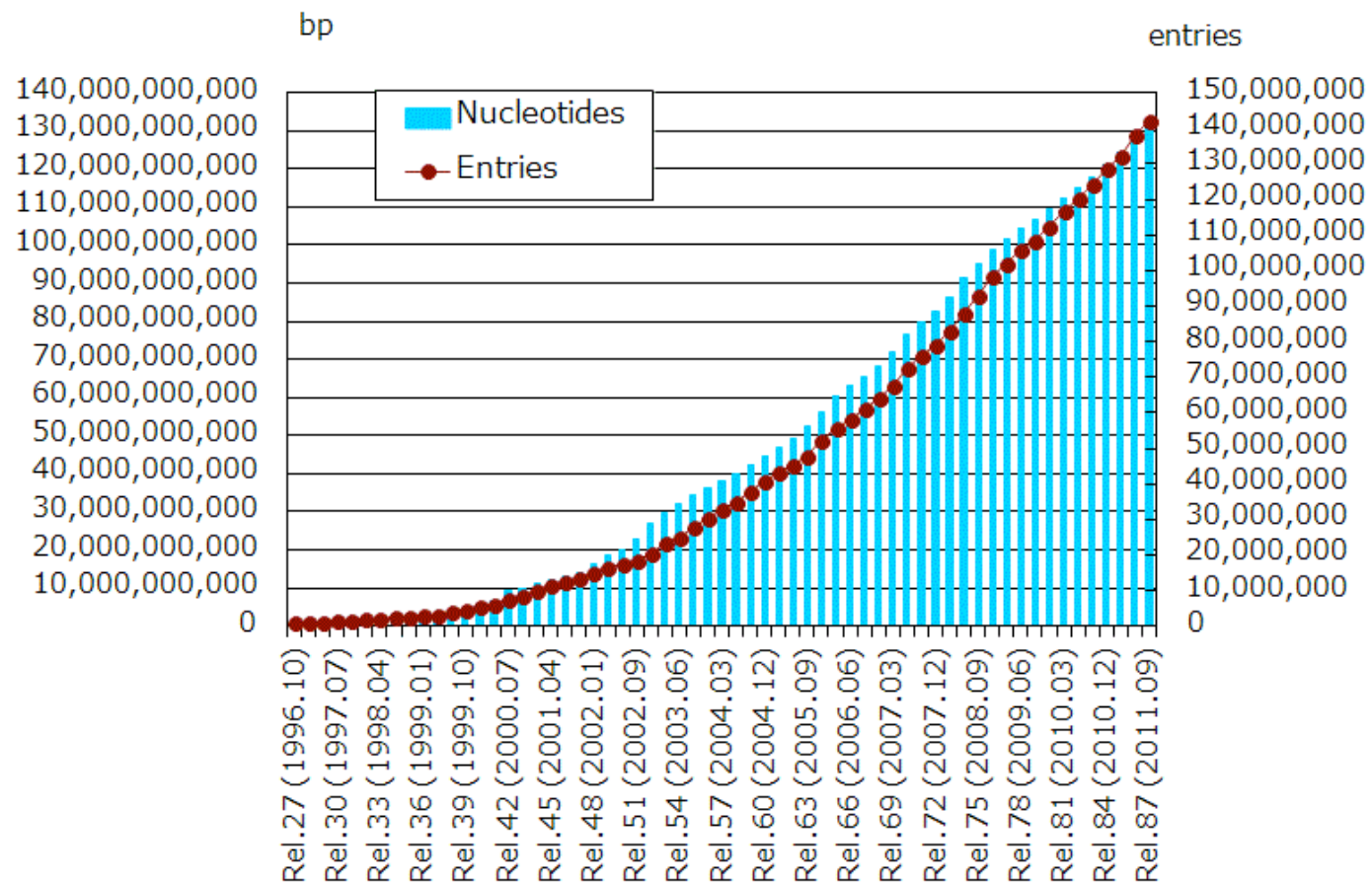
🎧 イルミナの最新型 HiSeq2000

- 🎧 一解析で6000億塩基 (600ギガベース)
- 🎧 ヒト一人のDNAがおよそ30億塩基対なので
- 🎧 一解析でざっくり200人分ゲノムが取得できる

塩基配列データの爆発

伝統的DNAデータベースの容量: **150GB**

DDBJ/EMBL/GenBank database growth

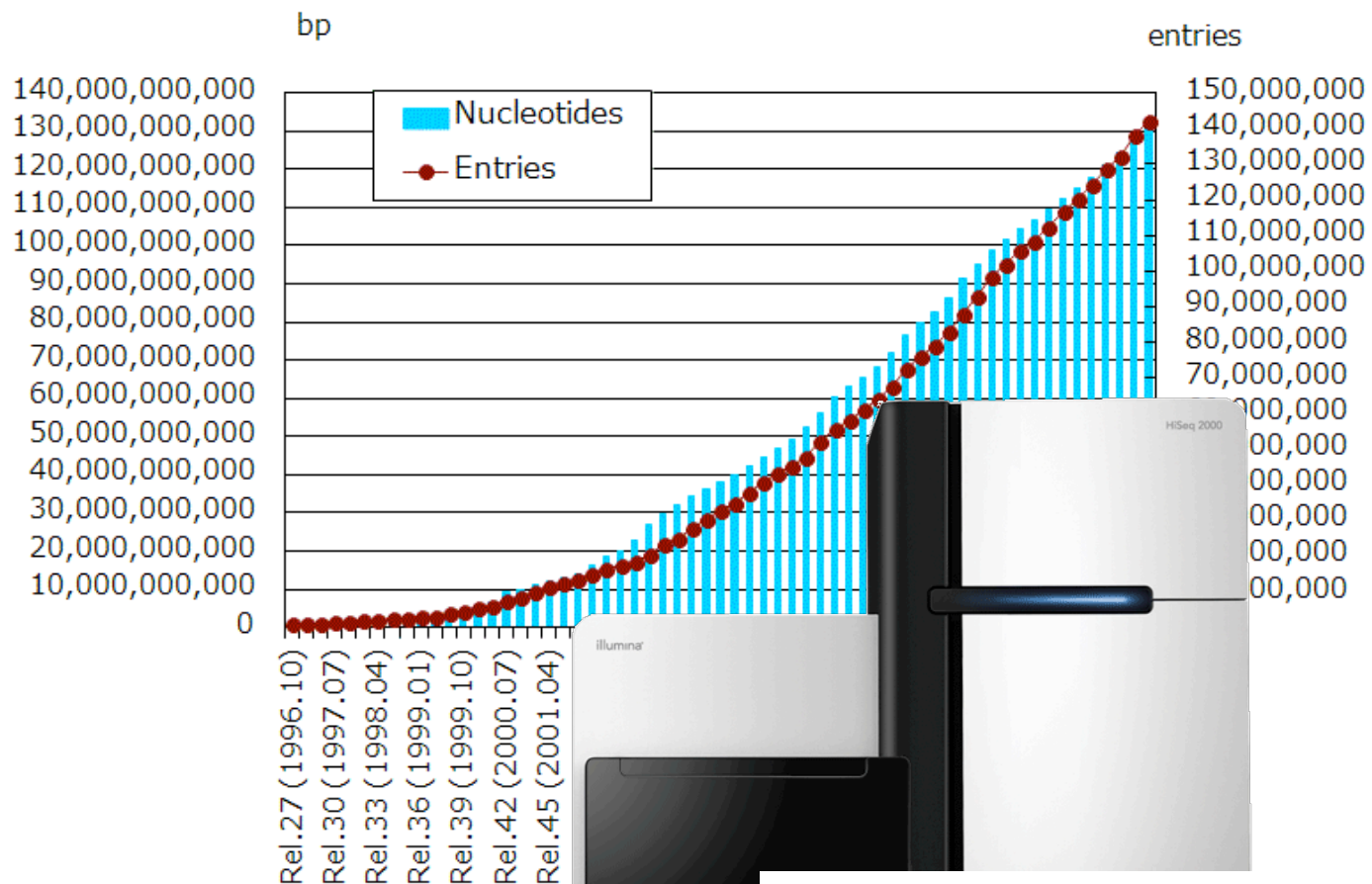


Note: CON division is not counted in statistics of DDBJ periodical

塩基配列データの爆発

伝統的DNAデータバンクの容量: **150GB**

DDBJ/EMBL/GenBank database growth



Note: CON divis

一解析で 600GB

遺伝研

スーパ

コンピュータ

遺伝研 Supercomputer (2012.3-)



2012.03.01

初期導入

- 165.1 TFlops
- 5 PB HDD
- Containing 10TB and 2TB shared memory system.

Rmax of LINPACK: 82.90 TFLOPS

Rank:170th in Top500 (Nov. 2011)

(Rank 11th in Japan)

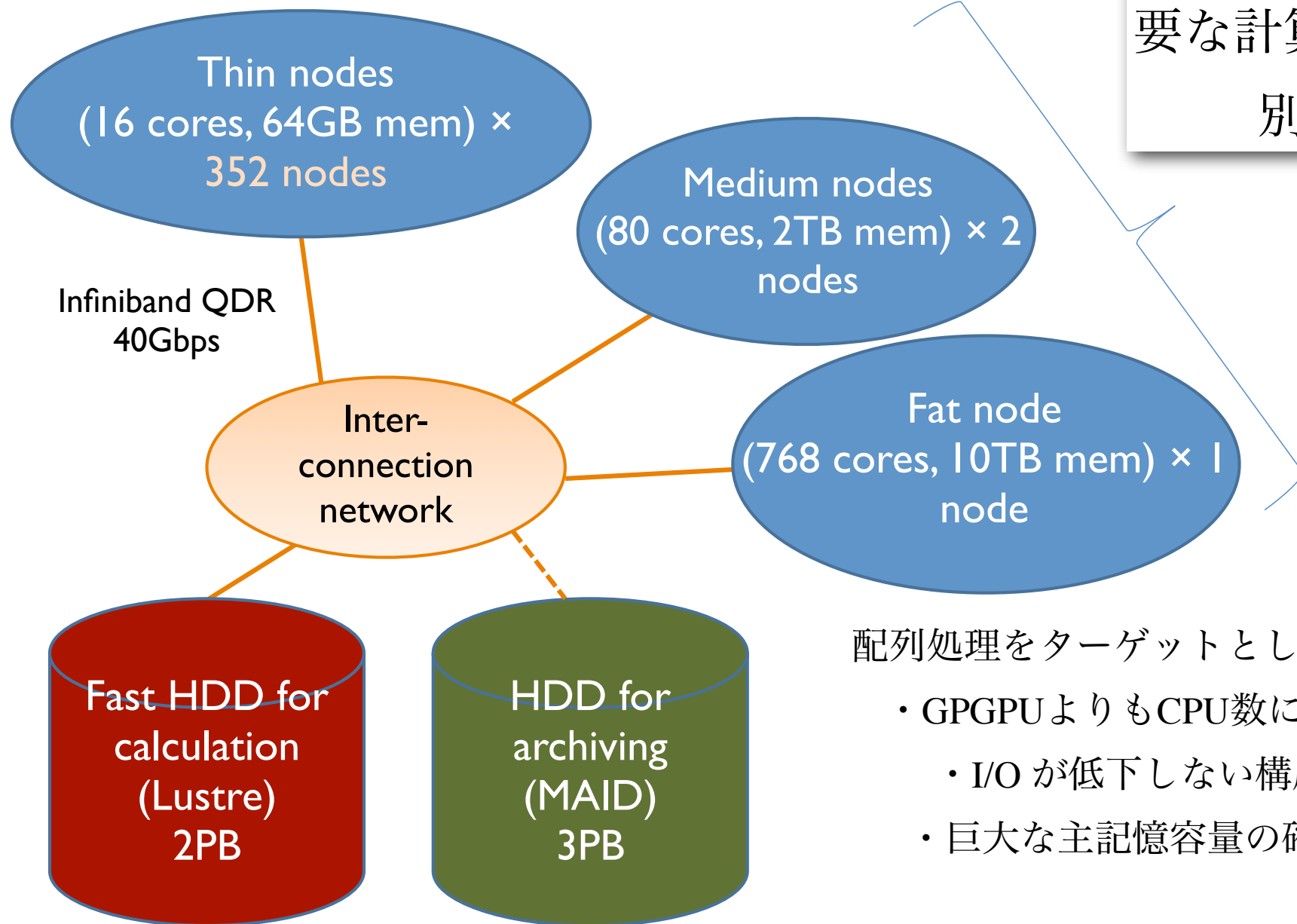
2014.03.01

追加導入

- about 400 TFlops (total)
- 12.5 PB HDD (total)

遺伝研スーパーコンピュータ：概要

大容量メモリが必要な計算専用の特別機器



配列処理をターゲットとした特徴

- GPGPUよりもCPU数に重点
 - I/O が低下しない構成
- 巨大な主記憶容量の確保

理研「京」
核融合研
東工大
東大
高エネ研
筑波大
東北大
京大
原子力機構
九州大

Rank	Site	System	Cores	Rmax (TFlop/s)	Rpeak (TFlop/s)	Power (kW)
2	RIKEN Advanced Institute for Computational Science (AICS) Japan	K computer, SPARC64 VIIIfx 2.0GHz, Tofu interconnect Fujitsu	705024	10510.0	11280.4	12659.9
12	International Fusion Energy Research Centre (IFERC), EU(F4E) - Japan Broader Approach collaboration Japan	Helios - Bulk B510, Xeon E5-2680 8C 2.700GHz, Infiniband QDR Bull SA	70560	1237.0	1524.1	2200
14	GSIC Center, Tokyo Institute of Technology Japan	TSUBAME 2.0 - HP ProLiant SL390s G7 Xeon 8C X5670, Nvidia GPU, Linux/Windows NEC/HP	73278	1192.0	2287.6	1398.6
18	Information Technology Center, The University of Tokyo Japan	Oakleaf-FX - PRIMEHPC FX10, SPARC64 IXfx 16C 1.848GHz, Tofu interconnect Fujitsu	76800	1043.0	1135.4	1176.8
36	High Energy Accelerator Research Organization /KEK Japan	BlueGene/Q, Power BQC 16C 1.60GHz, Custom IBM	49152	517.6	629.1	246.6
41	Center for Computational Sciences, University of Tsukuba Japan	HA-PACS - Xtream-X GreenBlade 8204, Xeon E5-2670 8C 2.600GHz, Infiniband QDR, NVIDIA 2090 Appro International	20800	421.6	778.1	407.3
70	Institute for Materials Research, Tohoku University (IMR) Japan	Hitachi SR16000 Model M1, POWER7 8C 3.836GHz, Custom Hitachi	10240	243.9	306.4	556.3
73	Kyoto University Japan	Camphor - Cray XE6, Opteron 16C 2.50GHz, Cray Gemini interconnect Cray Inc.	30080	239.4	300.8	
84	Japan Atomic Energy Agency (JAEA) Japan	BX900 Xeon X5570 2.93GHz, Infiniband QDR Fujitsu	17072	191.4	200.1	831
106	Research Institute for Information Technology, Kyushu University Japan	PRIMEHPC FX10, SPARC64 IXfx 16C 1.848GHz, Tofu interconnect Fujitsu	12288	166.7	181.7	
110	University of Tokyo/Institute for Solid State Physics Japan	SGI Altix ICE 8400EX Xeon X5570 4-core 2.93 GHz, Infiniband SGI	15360	161.8	180.0	719
126	Kyoto University Japan	Laurel - Xtreme-X, Xeon E5-2670 8C 2.600GHz, Infiniband FDR Appro International	9280	135.4	193.0	210.4
141	Financial Institution Japan	xSeries x3650M3, Xeon X56xx 2.66 GHz, GigE IBM	22272	125.5	237.0	690.4
145	Japan Agency for Marine -Earth Science and Technology Japan	Earth Simulator - SX-9/E/1280M160 NEC	1280	122.4	131.1	
146	Information Initiative Center, Hokkaido University Japan	Hitachi SR16000 Model M1, POWER7 8C 3.836GHz, Custom Hitachi	5632	121.6	168.9	354.6
160	JAXA Japan	Fujitsu FX1, Quadcore SPARC64 VII 2.52 GHz, Infiniband DDR Fujitsu	12032	110.6	121.3	1020
181	Information Technology Center, The University of Tokyo Japan	T2K Open Supercomputer (Total Combined Cluster) - Hitachi opteron QC 2.3 GHz Myrinet 10G Hitachi	15104	101.7	139.0	831
183	University of Tokyo/Human Genome Center, IMS Japan	HA8000-tc/HT225, Opteron 6276 16C 2.300GHz, Infiniband QDR Hitachi	16128	100.6	148.4	
190	Institute of Physical and Chemical Res. (RIKEN) Japan	RIKEN Integrated Cluster of Clusters, Xeon X5570 2.93GHz, Infiniband DDR Fujitsu	9048	97.9	106.0	
270	Numazu Complex, Fujitsu Limited Japan	PRIMEHPC FX10, SPARC64 IXfx 16C 1.848GHz, Tofu interconnect Fujitsu	6144	84.4	90.8	
280	National Institute of Genetics Japan	Cluster Platform SL230s/SL250s, Xeon E5-2670 8C 2.60GHz, Infiniband QDR Hewlett-Packard	5616	82.9	116.8	



2012.6

国内21位



世界280位

遺伝研スパコンを利用するには

- <http://www.ddbj.nig.ac.jp/system/supercom/supercom-apl.html>
- [DDBJ スーパーコンピュータ] で検索



The screenshot shows the DDBJ (DNA Data Bank of Japan) website. The header includes the DDBJ logo, navigation links (HOME, 塩基配列の登録, 利用の手引き, 検索・解析, FTP・WebAPI, レポート・統計, お問い合わせ), and a search bar. The main content area is titled 'スーパーコンピュータシステムの利用申込' (Supercomputer System Usage Application). It contains a paragraph explaining the login requirements for the supercomputer system and a section for '新規申込' (New Application) with a link for 'アカウント新規受付' (New Account Acceptance).

DDBJ
DNA Data Bank of Japan

ENGLISH

サイト内検索

HOME 塩基配列の登録 **利用の手引き** 検索・解析 FTP・WebAPI レポート・統計 お問い合わせ

HOME > [DDBJing目次](#) : [DDBJ利用の手引き](#) > [スーパーコンピュータシステムの利用申込](#)

▶ DDBJの紹介

▶ Q&A集

▶ 塩基配列の登録

- ▶ SAKURA
- ▶ 大量登録システム(MSS)
- ▶ データの修正・更新
- ▶ [DDBJ Sequence Read Archive](#)
- ▶ [DDBJ Trace Archive](#)

▶ プロジェクトの登録

- ▶ [DDBJ BioProject Database](#)

▶ 検索

スーパーコンピュータシステムの利用申込

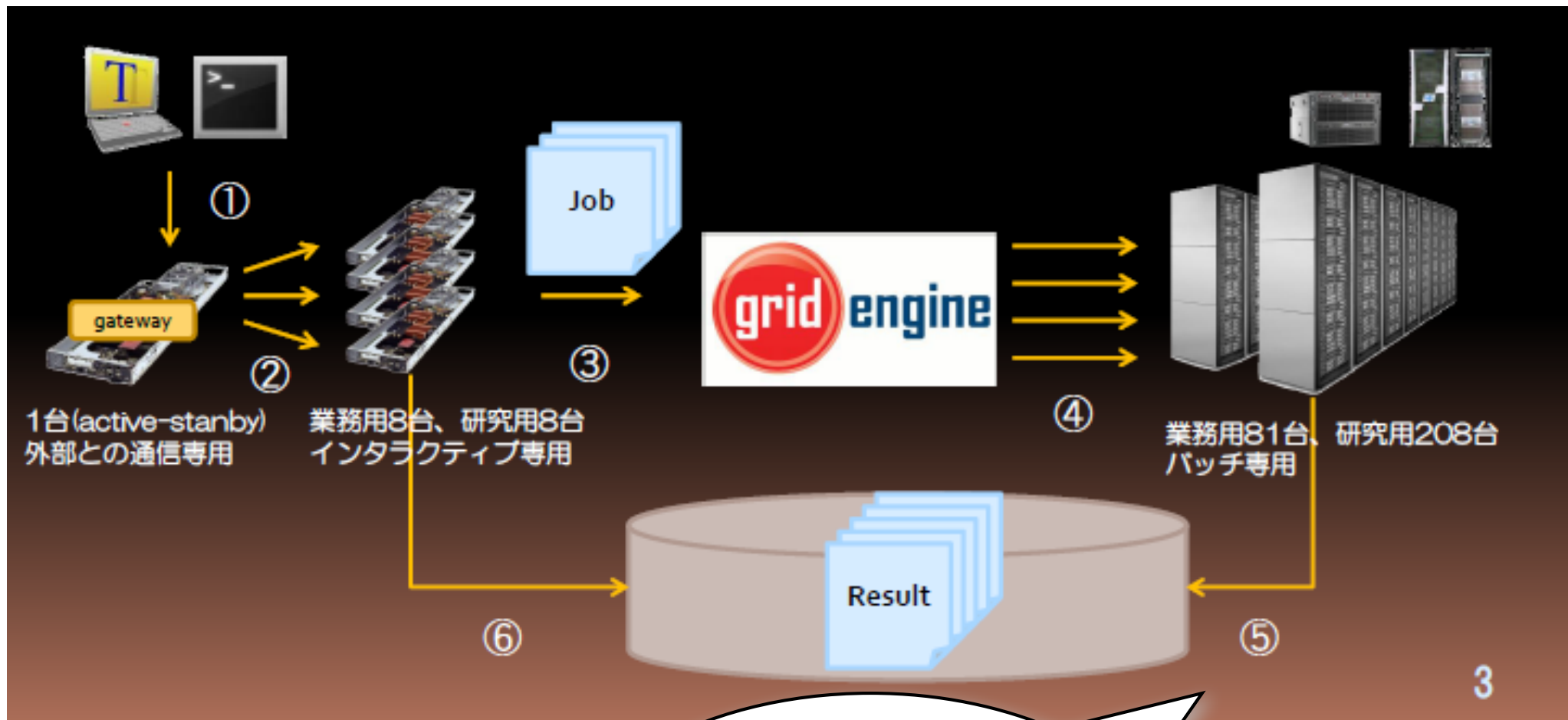
国立遺伝学研究所のスーパーコンピュータシステム(以下「スパコン」)にログインして利用する場合は、計算機のアカウントが必要です。[情報・システム研究機構国立遺伝学研究所 DDBJ塩基配列データベース等利用規程](#)をご覧ください。利用申請を行ってください。利用期間は一事業年度です。利用を継続する場合は、年度末に継続申請の手続きを行ってください。利用を中止する場合は、国立遺伝学研究所大型計算機利用中止申請を行ってください。

今後もこの方面のサービスの充実を図って参りますので、ご理解とご協力をお願いします。

▶ 新規申込

「[アカウント新規受付](#)」
スパコンにログインするアカウントを希望の方は、こちらから申込を行ってください。申請受理後、郵送にて登録証を発送いたします。

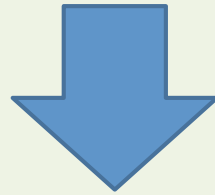
基本的な利用方法：バッチジョブの投入



バッチ（非対話）
ジョブ投入、
キューの順番で処理

バッチジョブ投入方法：qsub コマンド

```
# run a script on a login node.  
bash your_script.sh
```



```
# run a script on a calculation node.  
qsub -cwd -S /bin/bash your_script.sh
```

メモリ量を指定して実行する方法

```
# This job runs on 1 CPU core and 128GB  
memory.
```

```
qsub -cwd -l month -l medium  
-l s_vmem=128G,mem_req=128G  
-S /bin/bash your_script.sh
```

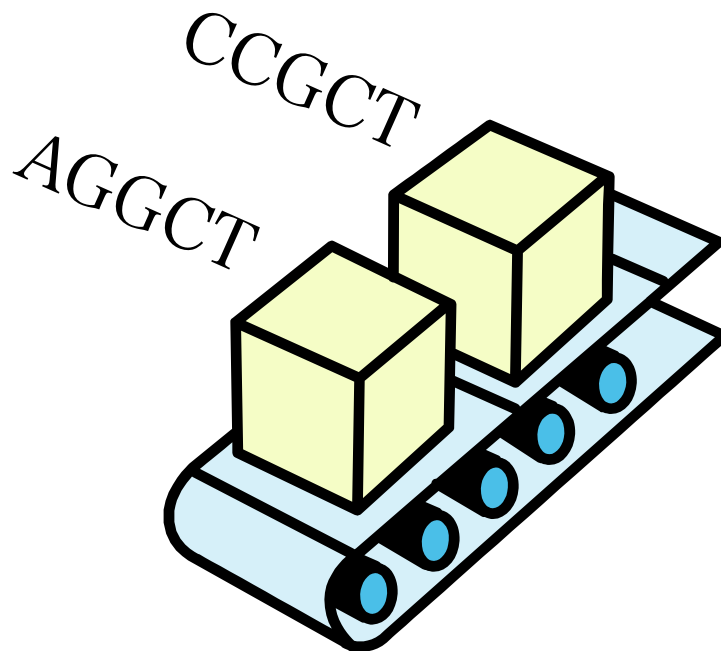
```
# This job runs on 10 CPU core (in the  
same node) and 1280GB memory.
```

```
qsub -cwd -l month -l medium  
-l s_vmem=128G,mem_req=128G  
-pe def_slot=10  
-S /bin/bash your_script.sh
```

一般の
生物学者には
難しい...

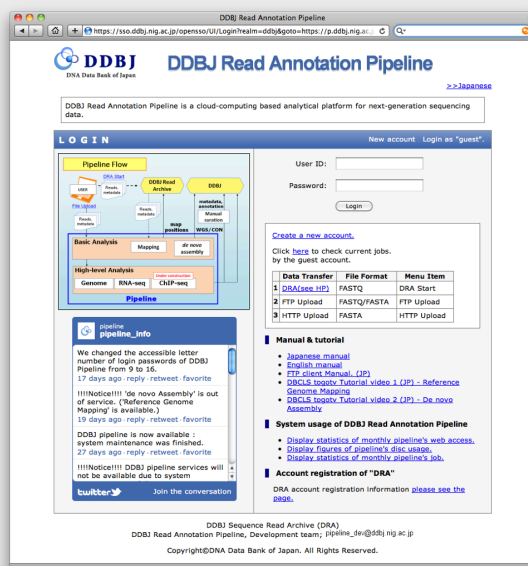
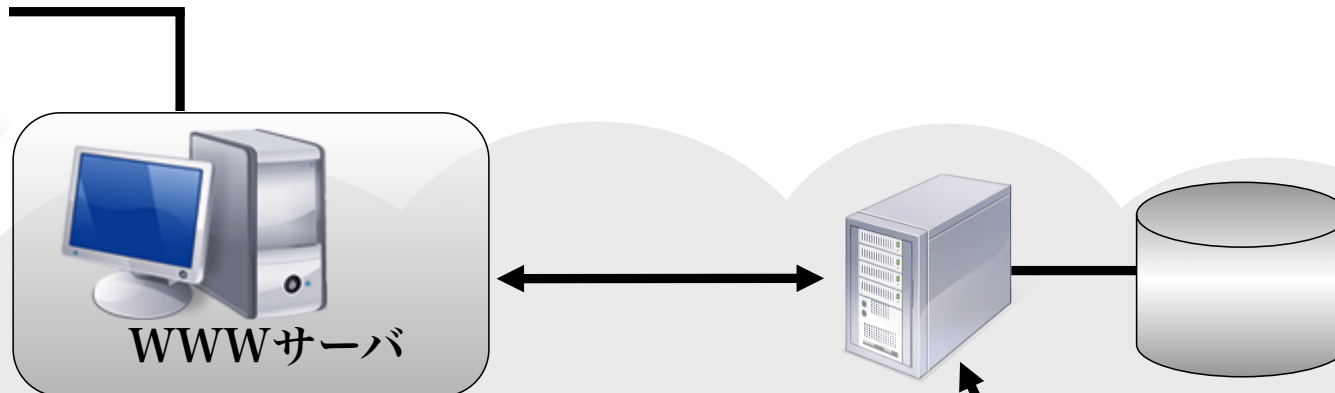
DDBJ Pipeline

新型シーケンサ配列のクラウド型解析ツール



DDBJ Pipeline: クラウド型解析ツール

研究者



DDBJ スーパーコンピュータ



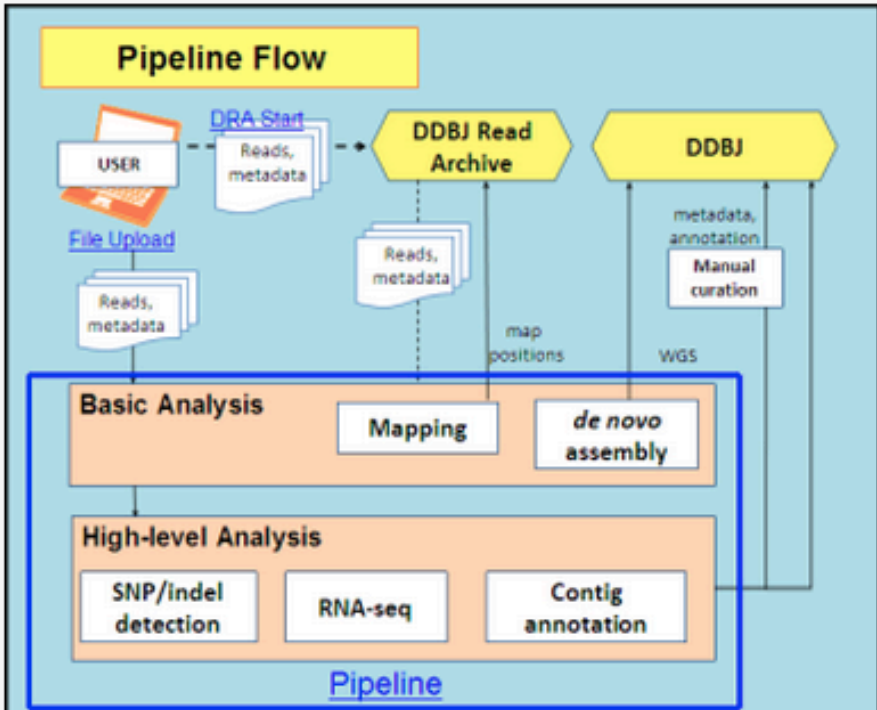
計算資源が足りない → DDBJスパコンを使おう
解析できない → web ブラウザから簡単に制御

DDBJ Read Annotation Pipeline

[English](#) [Japanese](#)

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

LOGIN [New account](#) [Login as "guest"](#)



User ID:

Password:

[Login](#)

[Check current jobs](#)
* by the guest account.

Manual & tutorial

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

Account registration of "DRA"

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

 pipeline_info

pipeline_info Reload bugs in 'HTTP upload' function were fixed. Please reload the web page of your uploaded data.

DDBJ pipeline: Software

よく用いられる
解析用ソフトウェアを
用意。クリックだけで
実行可能

Selecting Tools for Basic Analysis of DDBJ ANNOTATION

https://p.ddbj.nig.ac.jp/pipeline/SelectTool.do

Selecting Tools for Basic Analysis of DDBJ ANNOTATION

BACK NEXT

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 type="checkbox"/>	Maq		0.7.1	✓		✓	✓	✓	✓	✓	✓	✓	✓		
<input type="checkbox"/>	bwa		0.5.9	✓		✓	✓	✓	✓					✓	
<input type="checkbox"/>	SOAP		2.21	✓		✓			✓	✓				✓	
<input type="checkbox"/>	Bowtie (SAMtools)		0.12.7 (0.1.16)	✓	✓	✓	✓	✓	✓	✓				✓	
<input type="checkbox"/>	TopHat		1.0.11 (BETA)	✓		✓	✓	✓	✓					✓	

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		1.2.5	✓				ABYSS works slow in our pipeline-system.

DDBJ pipeline: references

イネ、マウスなど
解析比較対象となる
配列を多数用意



ACCOUNT

login ID [guest]

Logout

ANALYSIS

step-1

Mapping / Assembly

step-2

Genome
(SNP/Short Indel)

Genome
(Large Indel)

RNA-seq
(Tag count)

ChIP-seq

Job Confirmation

step-1 Status

step-2 Status

Help

MANUAL

BENCHMARK

feedback

Select Query Files

Running Status

Specifying Reference Genome

RESET BACK NEXT

Major genome

Organisms

Genome sets

all check

- chr01.fasta
- chr02.fasta
- chr03.fasta
- chr04.fasta
- chr05.fasta
- chr06.fasta
- chr07.fasta
- chr08.fasta
- chr09.fasta
- chr10.fasta
- chr11.fasta
- chr12.fasta

- Arabidopsis thaliana
- Oryza sativa japonica
- Oryza sativa indica
- Zea mays B73
- Sorghum bicolor
- Homo sapiens
- Mus musculus
- Pan troglodytes
- Caenorhabditis elegans
- Xenopus (Silurana) tropicalis
- Oryzias latipes
- Solanum lycopersicum Heintz 1706
- Saccharomyces cerevisiae

- IRGSP Releases Build 4.0
- IRGSP Releases Build 5.0
- IRGSP Releases Build 5.0 masked by RepeatMasker with MIPS repeat data
- tigr version5.0
- tigr version6.0
- tigr version6.1
- tigr mitochondrion
- tigr chloroplast

Organisms

Mus musculus

Genome sets

- Dec.2011 (mm10)
- Jul. 2007 (mm9)
- Mar.2006 (mm8)
- Aug.2005 (mm7)
- NCBI build 36
- NCBI build 37

all check

- chr1.fa
- chr10.fa
- chr11.fa

Organisms

Arabidopsis thaliana

Genome sets

- TAIR8
- TAIR9
- TAIR10

all check

- chr1.fas
- chr2.fas
- chr3.fas

User original sets

Download or upload reference

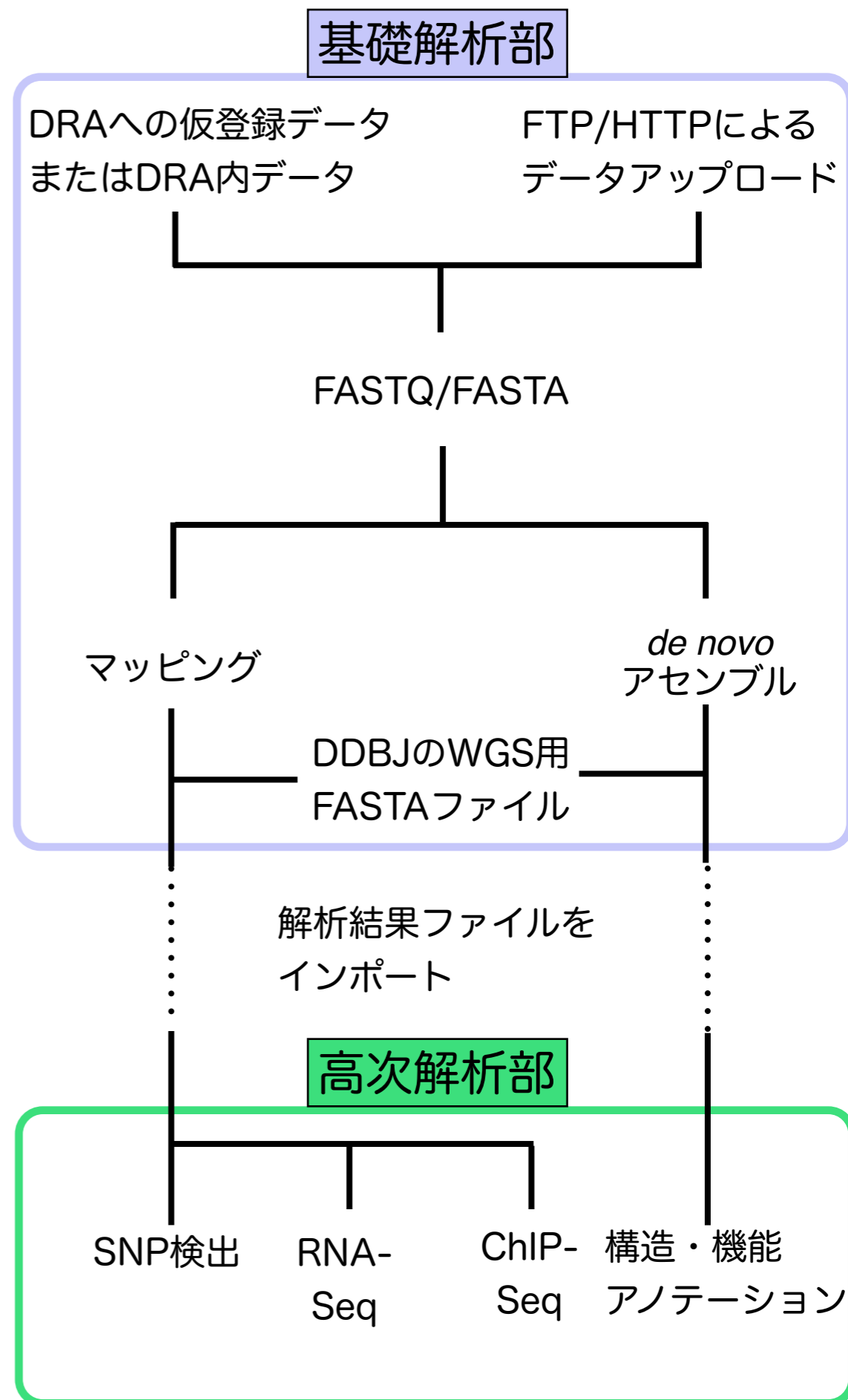
DDBJ pipeline: merits

- DDBJのPC cluster上で運用
- 大量データ転送問題が回避できる
- 最新の *de facto* standard ツールを用意
 - for *de novo* assemble 「アSEMBル」
 - for reference genome mapping 「マッピング」
- 処理はwww上のGUI簡単

「DDBJパイプラインによる RNA-seq配列のde novoアセンブル」

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



The screenshot shows the DDBJ Read Annotation Pipeline website. The page title is "DDBJ Read Annotation Pipeline" and the URL is "https://sso.ddbj.nig.ac.jp/opensso/UI/Login?realm=ddbj&goto=https://p.ddbj.nig.ac.jp". The DDBJ logo (DNA Data Bank of Japan) is in the top left. A description states: "DDBJ Read Annotation Pipeline is a cloud-computing based analytical platform for next-generation sequencing data." There is a link to ">>Japanese".

The main content area is divided into two columns. The left column features a "LOGIN" header with links for "New account" and "Login as 'guest'". Below this is a "Pipeline Flow" diagram showing the process from "USER" (File Upload) to "DDBJ Read Archive" and "DDBJ", including "Basic Analysis" (Mapping, de novo assembly) and "High-level Analysis" (Genome, RNA-seq, ChIP-seq). A "Pipeline" box highlights the analysis steps. Below the diagram is a "pipeline_info" section with a Twitter feed containing several announcements about password changes, service outages, and system maintenance.

The right column contains a login form with fields for "User ID:" and "Password:", and a "Login" button. Below the form is a link to "Create a new account." and a note: "Click [here](#) to check current jobs by the guest account." A table lists data transfer methods:

	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

Below the table are sections for "Manual & tutorial" (with links to Japanese and English manuals, FTP client manual, and DBCLS tutorial videos) and "System usage of DDBJ Read Annotation Pipeline" (with links to display statistics). The "Account registration of 'DRA'" section includes a link to registration information.

At the bottom of the page, it says: "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."

- 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

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	Help	34	✓						✓						Single-end analysis only
<input type="checkbox"/>	Maq	Help	0.7.1	✓		✓				✓	✓	✓	✓	✓		
<input type="checkbox"/>	bwa	Help	0.5.9	✓		✓				✓					✓	
<input type="checkbox"/>	SOAP	Help	2.21	✓		✓				✓	✓				✓	
<input type="checkbox"/>	Bowtie	Help	0.12.7	✓	✓	✓				✓	✓				✓	
<input type="checkbox"/>	TopHat	Help	1.0.11	✓		✓				✓					✓	

de novo Assembly
Total limit = 22 Gbp

	Tool	Help	Version	Base space	Color space	Paired-end	MSS(WGS)	Comment
<input type="checkbox"/>	SOAPdenovo	Help	1.05			✓		
<input type="checkbox"/>	ABySS	Help	1.3.2			✓		Maximum K-mer value is 64.
<input type="checkbox"/>	Velvet	Help	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	Help	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
<input checked="" type="radio"/>	BLAT	Single-end analysis only

- 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

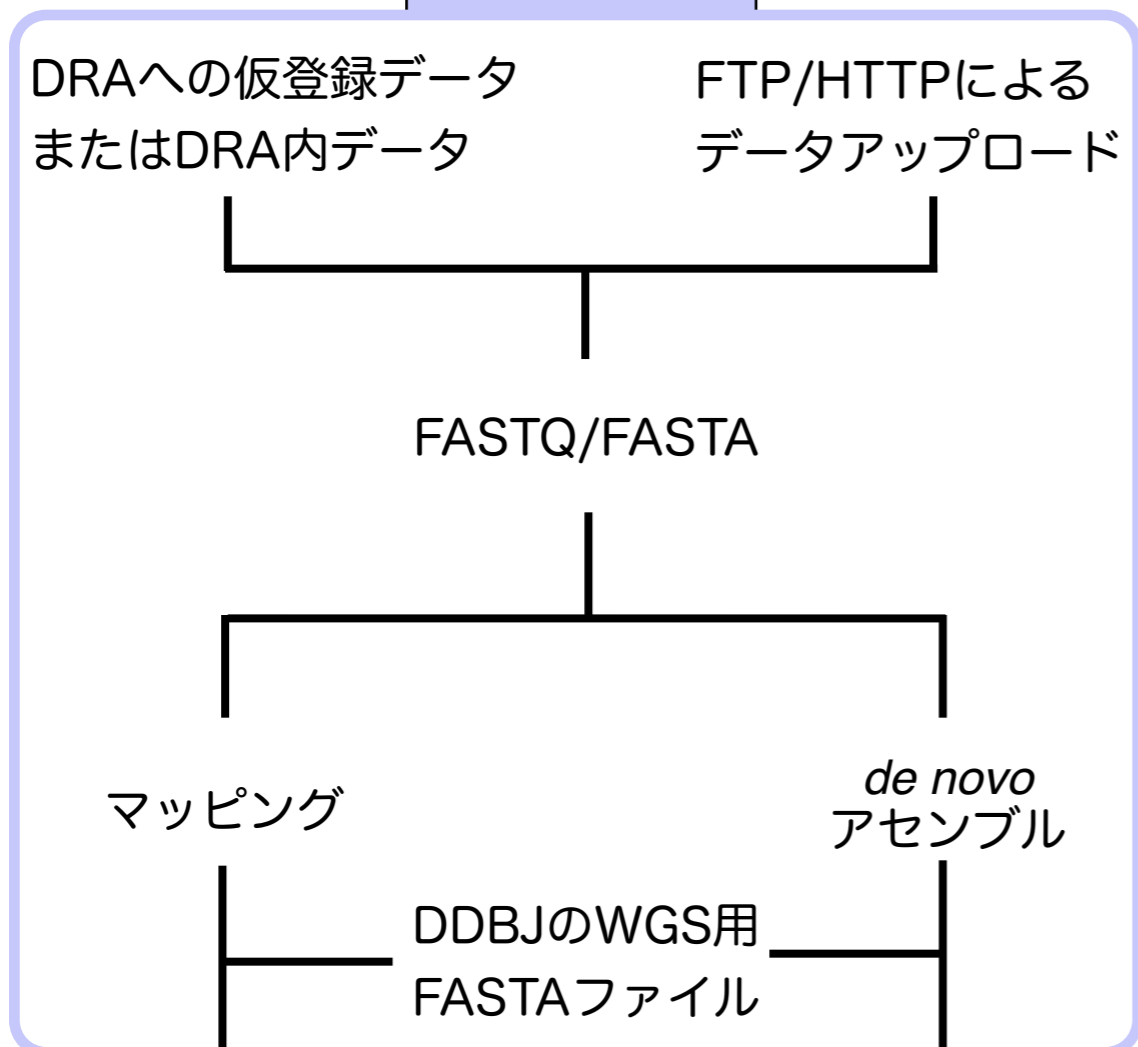
解析終了をメールで通知

- SAMtools/FASTAによる共通フォーマットでの出力

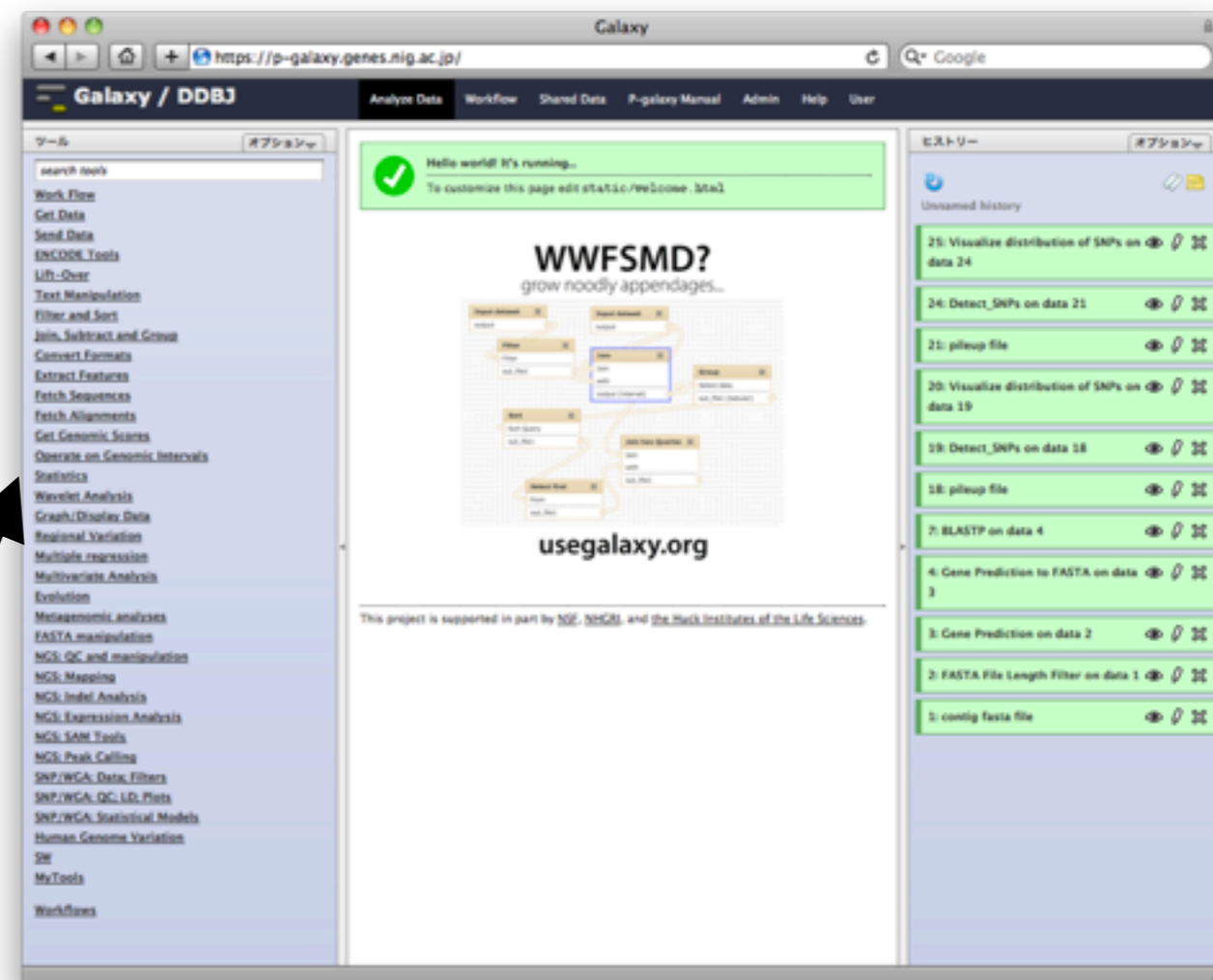
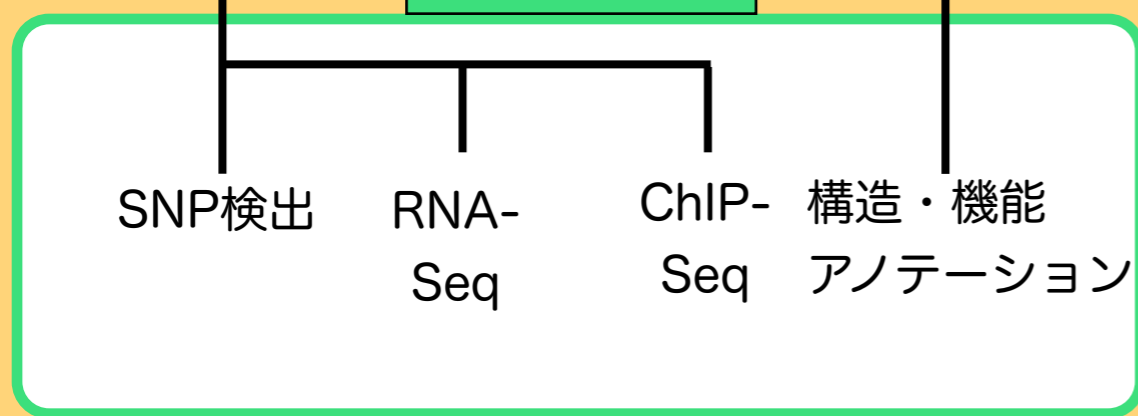
The screenshot shows the 'Status - Mapping' page of the DDBJ pipeline. It features a progress bar at the top with steps: Select Query Files, Select Tools, Set QuerySet, Set GenomeSet, Set Map Options, and Confirmation. Below the progress bar, there are buttons for 'Running Status', 'Mapping query state', 'Assembly query state', and 'PreProcess query state'. The main content is a table with columns: ID, UserID, Submission accession, P/S, Status, Tool, Read #, Read length, Genome size, Download, Start time, End time, and Elapsed time. The table lists several jobs, including those for 'Mo17', 'kyotou_pmb-00', 'Gla4-L2_Rounc', 'GSM276809.1', 'newly_synthesi', 'tomohiro-0005_', and 'Oryza rufipogon'. Each row includes a 'File' button and a color-coded progress bar for the elapsed time.

ID	UserID	Submission accession	P/S	Status	Tool	Read #	Read length	Genome size	Download	Start time	End time	Elapsed time
3233	ekaminuma	SRA009756 Mo17 Mo17_1	S	complete	bwa	49,285,831	—	4,185 M	File	2011-12-10 13:24:27	2011-12-10 13:53:22	00:28:54
3224	nagasakicool	DRA000369 kyotou_pmb-00	P	complete	bwa	122,403,348	—	228 M	File	2011-12-08 16:07:20	2011-12-11 08:01:23	83:54:02
3220	fu	ERA000212 Gla4-L2_Rounc	P	complete	bwa	2,873,326	40	83 M	File	2011-12-07 15:47:18	2011-12-07 16:16:55	00:29:37
3217	user-demo	SRA000284 GSM276809.1	S	complete	bwa	28,723,026	—	2,708 M	File	2011-12-07 13:21:20	2011-12-07 13:42:34	00:21:13
3215	user-demo	SRA030871 newly_synthesi	S	complete	TopHat	4,871	—	254 M	File	2011-12-07 12:04:20	2011-12-07 12:21:57	00:17:36
3214	user-demo	DRA000307 tomohiro-0005_	P	complete	bwa	17,359,151	—	388 M	File	2011-12-07 10:13:34	2011-12-07 18:47:04	08:33:30
3212	Imochidu	DRA000158 Oryza rufipogon Oryza rufipogon	P	complete	bwa	50,531,840	110	390 M	File	2011-12-06 11:55:27	2011-12-07 12:07:43	19:39:07

基礎解析部



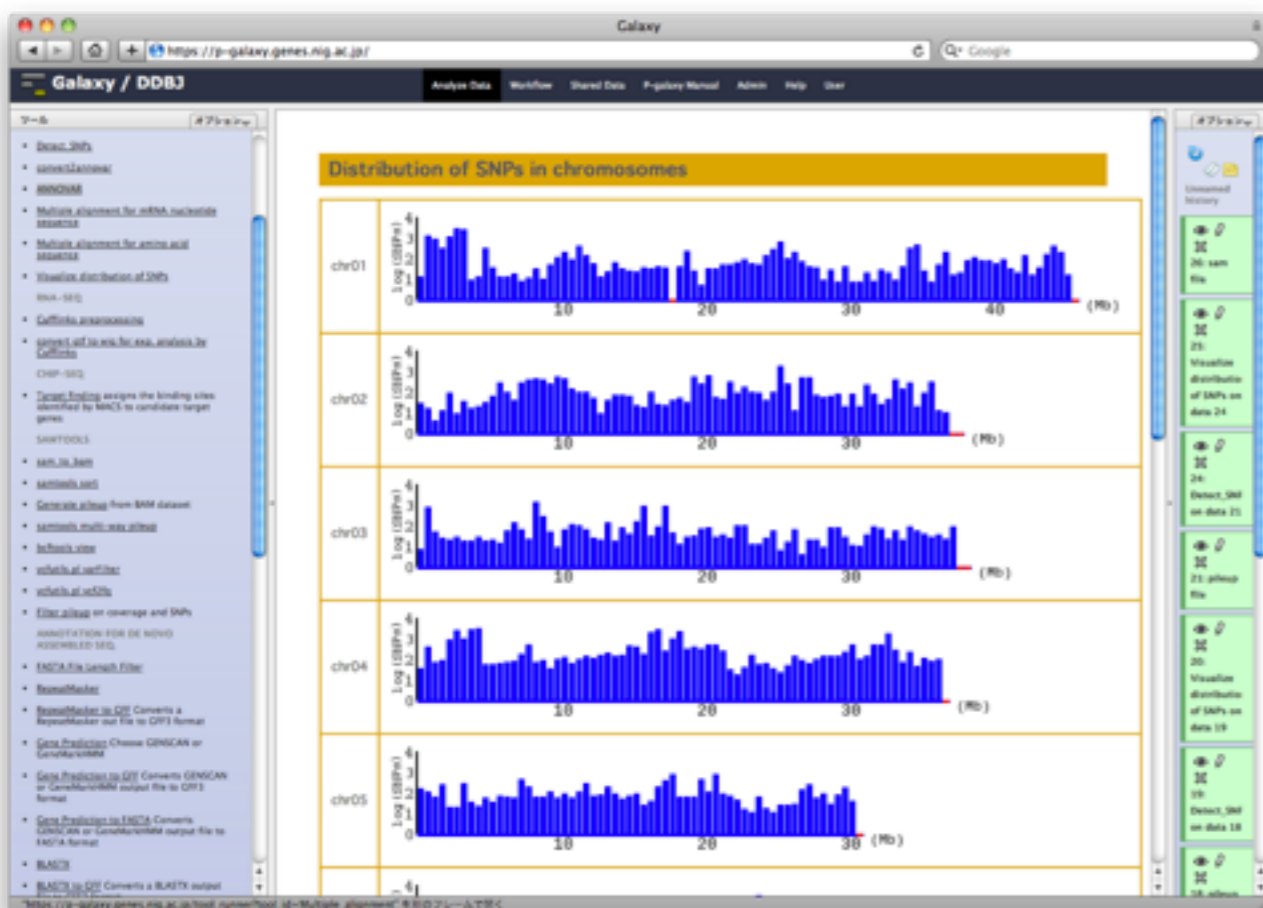
高次解析部



- 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)

NGSデータのマッピング結果の解析

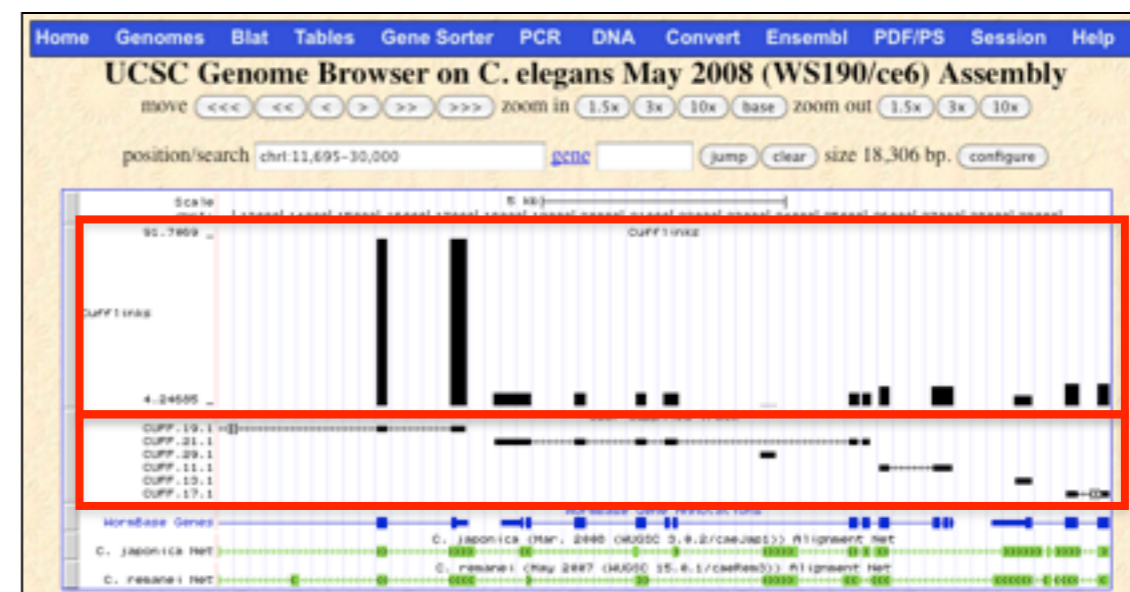
SNPのゲノム上の分布の表示



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

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

- ✓ Swiss-Prot-Bacteria
- Swiss-Prot-Plants
- Swiss-Prot-Invertebrates
- Swiss-Prot-Mammals
- Swiss-Prot-Vertebrates
- Swiss-Prot
- nr

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; 187,423,367 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
SF=1
Length = 51

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

Query: 1 MKRKAQVNEALIKNNWTFTESHDFNSYKIQYHDDPNFRGANRNSKQGGQGGI 51
MKRKAQVNEALIKNNWTFTESHDFNSYKIQYHDDPNFRGANRNSKQGGQGGI
Sbjct: 1 MKRKAQVNEALIKNNWTFTESHDFNSYKIQYHDDPNFRGANRNSKQGGQGGI 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; 187,423,367 total letters

Searching.....done

Sequences producing significant alignments:

Score E
(bits) Value
sp|O31659|YKZE_BACSU Uncharacterized protein ykzE OS=Bacillus su... 75 8e-14
>sp|O31659|YKZE_BACSU Uncharacterized protein ykzE OS=Bacillus subtilis GN=ykzE PE=4
SF=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 MQNKGKPHDKKTLLEEFSSSELGQYVAGKIIETLEVT 36
MQNKGKPHDKKTLLEEFSSSELGQYVAGKIIETLEVT
Sbjct: 10 MQNKGKPHDKKTLLEEFSSSELGQYVAGKIIETLEVT 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ったってスプライスバリエーション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。

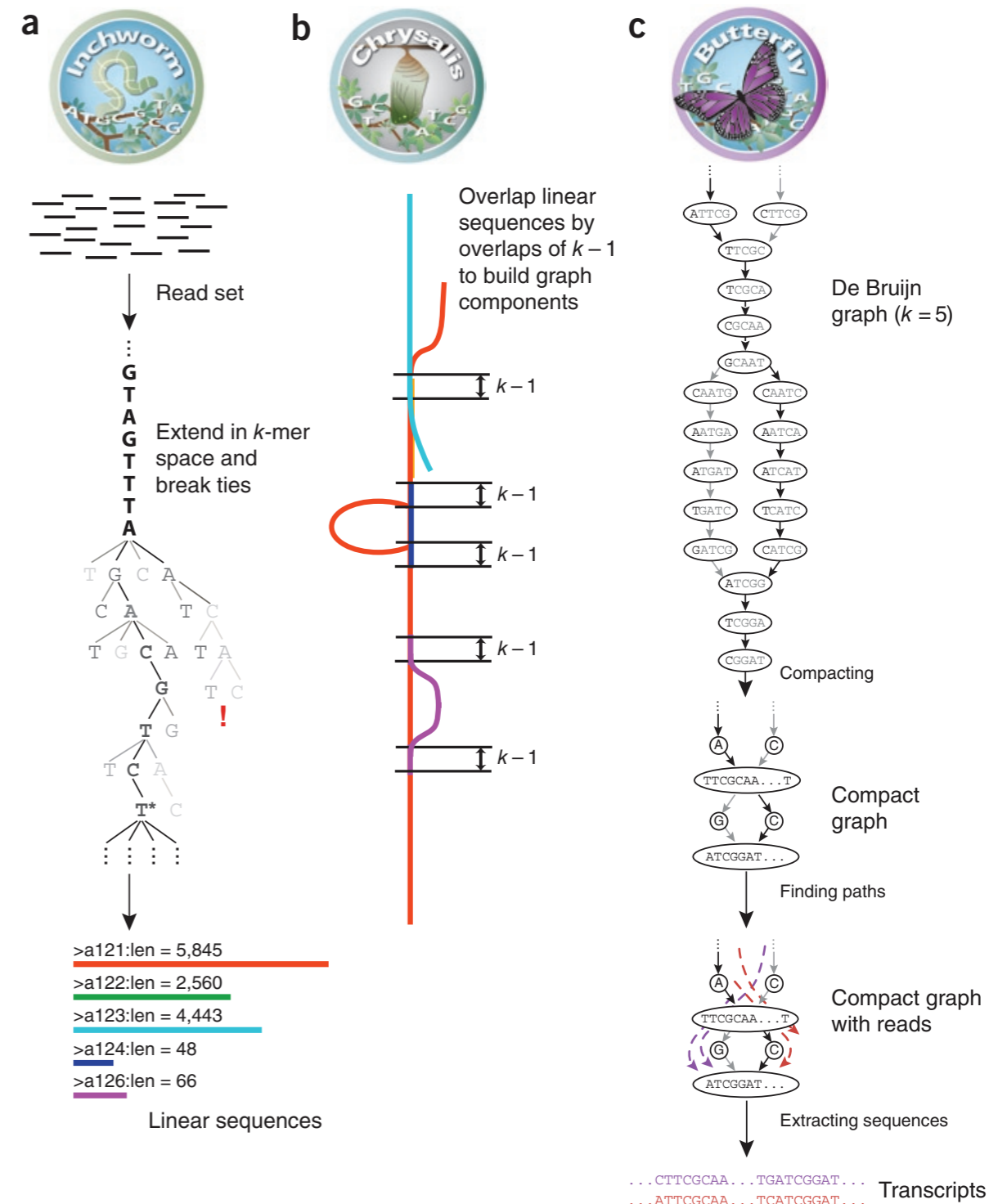


Figure 1 Overview of Trinity. (a) Inchworm assembles the read data set (short black lines, top) by greedily searching for paths in a k -mer graph (middle), resulting in a collection of linear contigs (color lines, bottom), with each k -mer present only once in the contigs. (b) Chrysalis pools contigs (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程度の幼魚が多くの特産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. in press.

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。

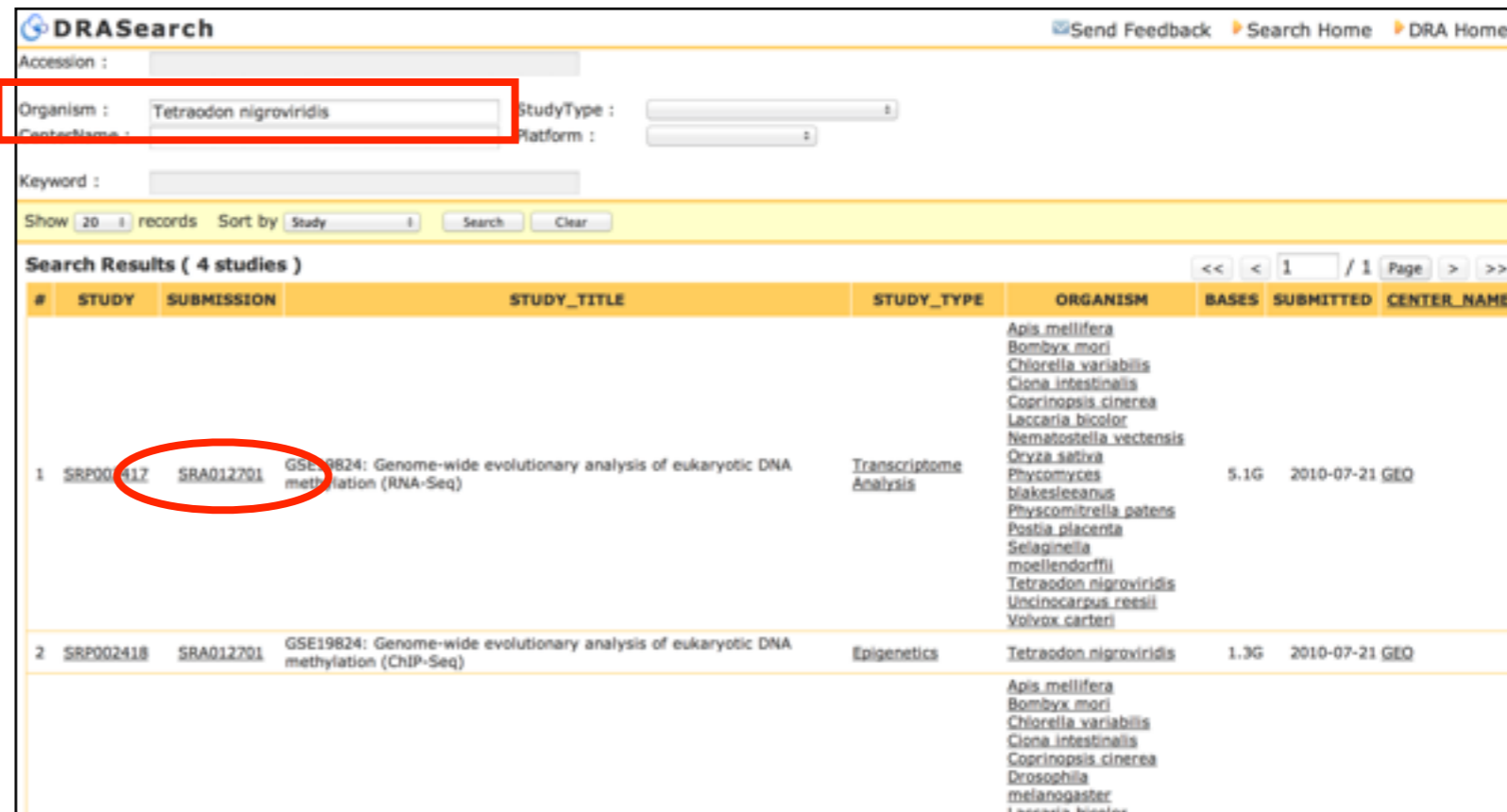
DRAのwebサイトから「検索」をクリック



DRA Search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

HOME 塩基配列の登録 利用の手引き 検索・解析 FTP・WebAPI レポート・統計 お問い合わせ

DDBJの紹介
Q&A集

塩基配列の登録
SAKURA
大量登録システム(MSS)
データの修正・更新
DDBJ Sequence Read Archive
DDBJ Trace Archive

プロジェクトの登録
DDBJ BioProject Database

スーパーコンピュータ利用
スローンの利用申込
スローンの利用方法
スローンマニュアル

検索
genentry
ARSA
TXSearch
BLAST

系統解析
ClustalW

NGSデータ解析
DDBJ Read Annotation Pipeline

ゲノム解析
MiGap
GIB (GIB-GIB-V-GTPS)
GTOP

タンパク質構造解析
PMQ

DDBJメールマガジン

DDBJ : DNA Data Bank of Japan

DDBJ (日本DNAデータバンク) は欧州と米国の対応機関 (EBIおよび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る三大国際DNAデータバンクのひとつです

Hot Topics

- 2012.08.08 ヨーロッパモノアラガイ (*Lymnaea stagnalis*) TSA データの公開
- 2012.07.31 富田勝教授の個人ゲノム配列が公開
- 2012.07.12 DDBJ Read Annotation Pipeline サービス再開

Maintenance

- 2012.07.31 (8/13-14)DDBJ 夏季休業
- 2012.07.02 DAD リリース59.0 にお

Information

- DDBJ Web Magazine No.73

塩基配列の登録・更新

- 塩基配列の登録
塩基配列の登録手順を御案内します。
- 登録データの修正・更新
登録データの修正をされる方は、最

日本DNAデータバンク (DDBJ)

DDBJ Center
DNA Data Bank of Japan
DDBJ Sequence Read Archive
DDBJ Trace Archive
BioProject

NCBI
GenBank
Sequence Read Archive

INSDC
JAC
ICM

ENA/EBI
EMBL
BioProject
Sequence Read Archive

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

DDBJ, pipeline で検索すると早い

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

DDBJ DNA Data Bank of Japan

DDBJ Read Annotation Pipeline

English Japanese

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

LOGIN

New account Login as "guest"

User ID:

Password:

Login

デモ用アカウントは講習内でお伝えします

Check current jobs

* by the guest account.

Manual & tutorial

- Japanese manual
- English manual
- DBCLS togotv Tutorial video 1 (JP) - Reference Genome Mapping
- DBCLS togotv Tutorial video 2 (JP) - De novo Assembly

Pipeline Flow

USER → DDBJ Read Archive → DDBJ

File Upload → Reads, metadata → DDBJ Read Archive → Reads, metadata → DDBJ → metadata, annotation → Manual curation

Basic Analysis: Mapping, de novo assembly

High-level Analysis: SNP/indel detection, RNA-seq, Contig annotation

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 DDBJ pipeline interface. At the top, a progress bar indicates the current step is 'Select Query Files'. The left sidebar contains navigation menus for ACCOUNT, ANALYSIS, and Workflow. The main content area is titled 'Selecting Query Files' and features a tabbed interface with options: FTP upload, Private DRA entry, Import public DRA (highlighted with a red circle and the word '選択' above it), Preprocessing, and HTTP upload. Below the tabs, there is a section for 'Metadata of the DRA entry.' with a dropdown menu set to 'DRA000001'. A table lists various metadata entries with columns for TYPE, ACCESSION, ALIAS, FILENAME, DL, and VIEW.

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

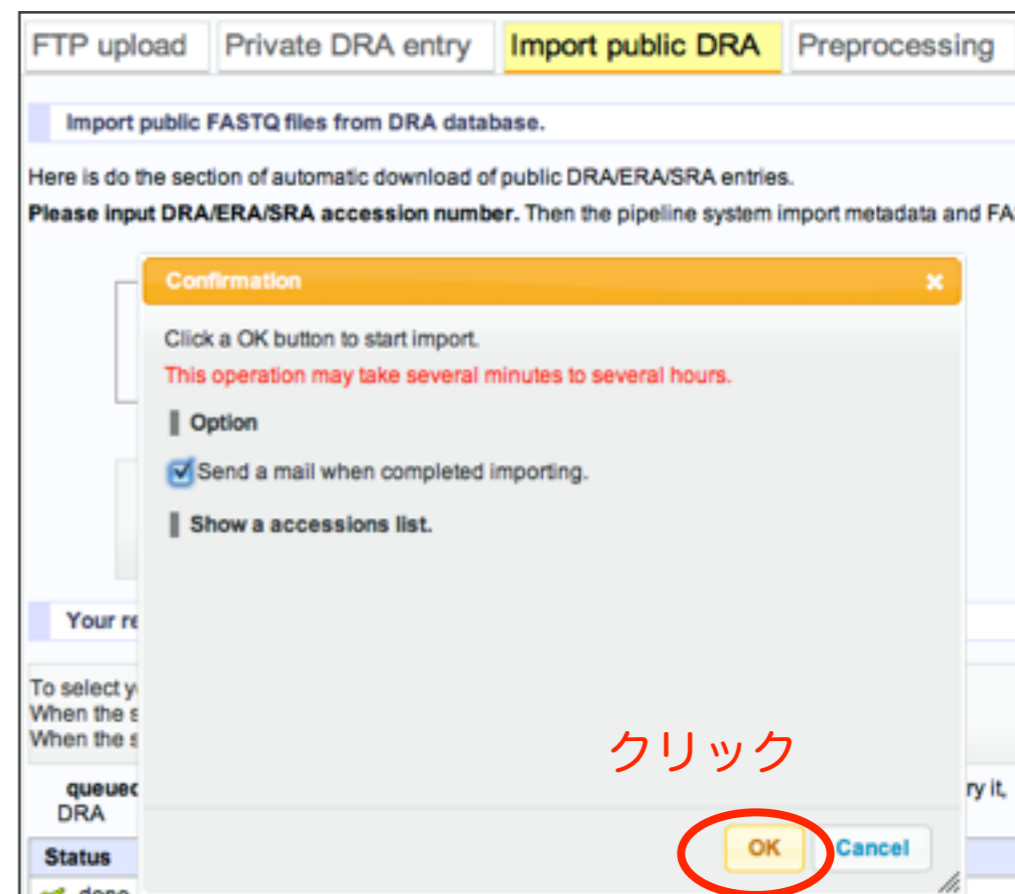
The screenshot shows the 'Import public FASTQ files from DRA database' section. It includes instructions: 'Here is do the section of automatic download of public DRA/ERA/SRA entries. Please input DRA/ERA/SRA accession number. Then the pipeline system import metadata and FASTQ files from DRA database.' Below this, there is an input field labeled 'Input DRA/ERA/SRA Accession Number' containing the text 'SRA012701', which is circled in red. To the right of the input field is a button labeled 'Add my DRA entry', which is also circled in red with the word 'クリック' written in red next to it. At the bottom, there is a link for 'Accession Number can find here. DRA Search'.

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。

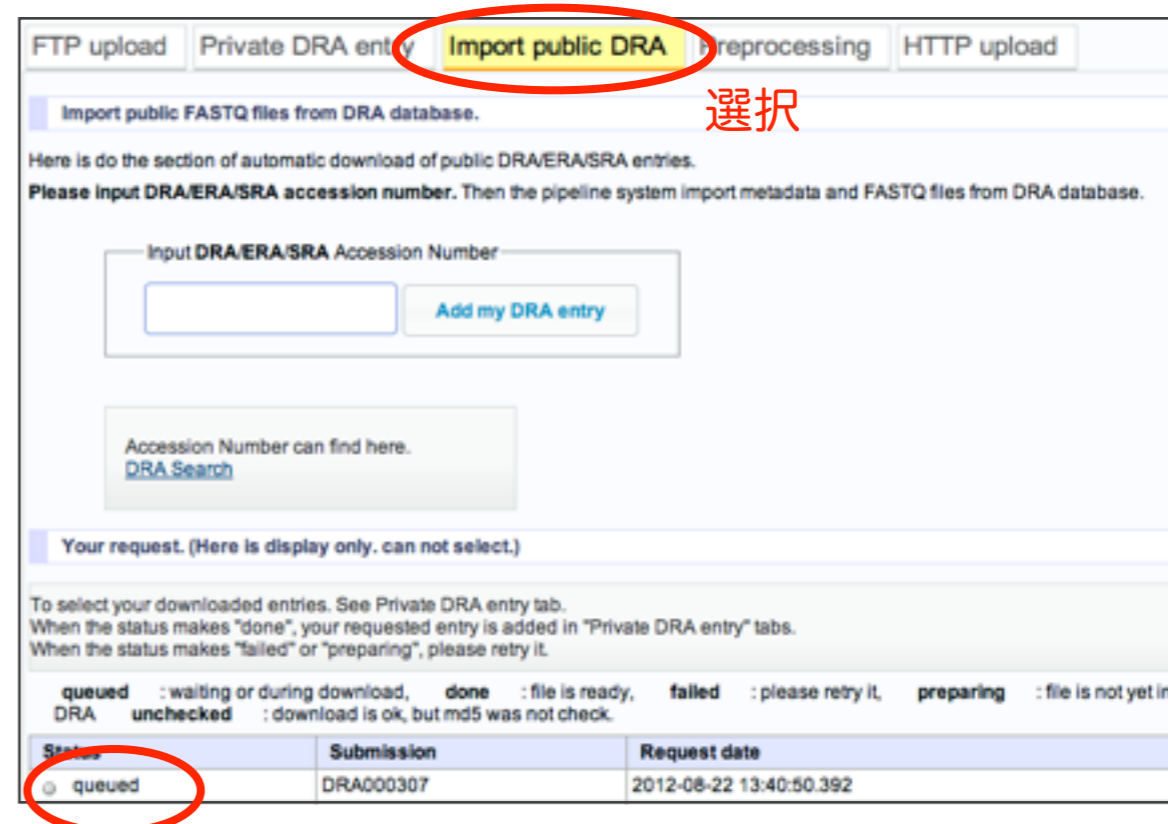
「OK」をクリック。



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

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)

TYPE	ACCESSION	ALIAS	FILENAME	DL
Submission	DRA000307	tomohiro-0005_Submission	DRA000307.submission.xml	Download
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

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。

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

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.

- Phred+33 Phred+64

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。

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 ($QV \leq THRESHOLD$) 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 ($length \leq 24 bp$), the read is removed. Thus the minimum read length will be 25bp.

リードの両端から $QV \leq 19$ 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 ($\geq Low\ quality\ bases\# / Total\ bases\#$) of the low quality bases ($QV \leq THRESHOLD$) are discarded.

トリム後のリードの中に、 $QV \leq 14$ のリードが 30 %

- QV THRESHOLD :

→ 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。

Email notification

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

* Required

Confirmation of entries

Query sets

- SRR042533 - GSM497271_1

BACK RUN

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
<input type="checkbox"/>	5509	koshu01	SRA012701 GSM497271_1	S	running		---	View	2013-04-30 17:42:30		
<input type="checkbox"/>	5455	---	--- FY23KIH080_pl	P	complete		---		2013-04-25 11:55:45		01:20:10
<input type="checkbox"/>	5452	---	--- FY22KIH033_pl	P	complete		---		2013-04-25 11:16:44		01:27:43
									2013-04-25 12:44:28		

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 BACK

Job info

ID: 5509

Tool (Version): (1.0)

RunAccession or Filename	Download	Read length	Alias
SRR042533	SRR042533.fastq.bz2	N.A. bp	GSM497271_1

File	Fastq Download	QS Average (PDF)	QS Count (PDF)
SRR042533.fastq.bz2	download (1.3 GB)	download (6.6 KB)	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

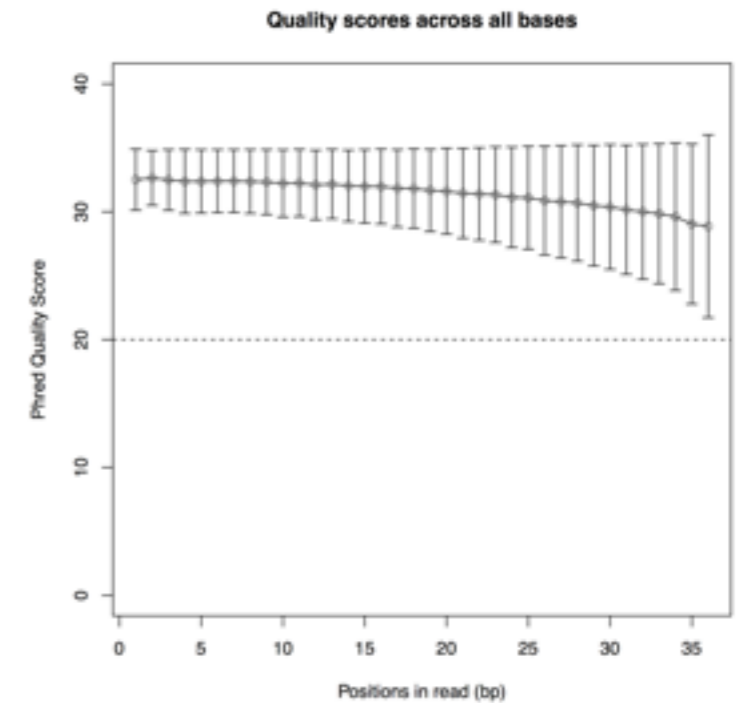
Command	Start time	End time	Log1	Log2	Result	MD5
perl avq_p.pl fqlist.txt qscore	2013-04-30 17:42:30	2013-04-30 17:50:28	View			
perl pdel_p3_t.pl fqlist.txt qscore 19 24 0 14 30 33	2013-04-30 17:50:28	2013-04-30 17:53:51				
perl user_fastq_copy.pl preprocessing.xml koshu01	2013-04-30 17:53:51	2013-04-30 17:56:24	View			

BACK

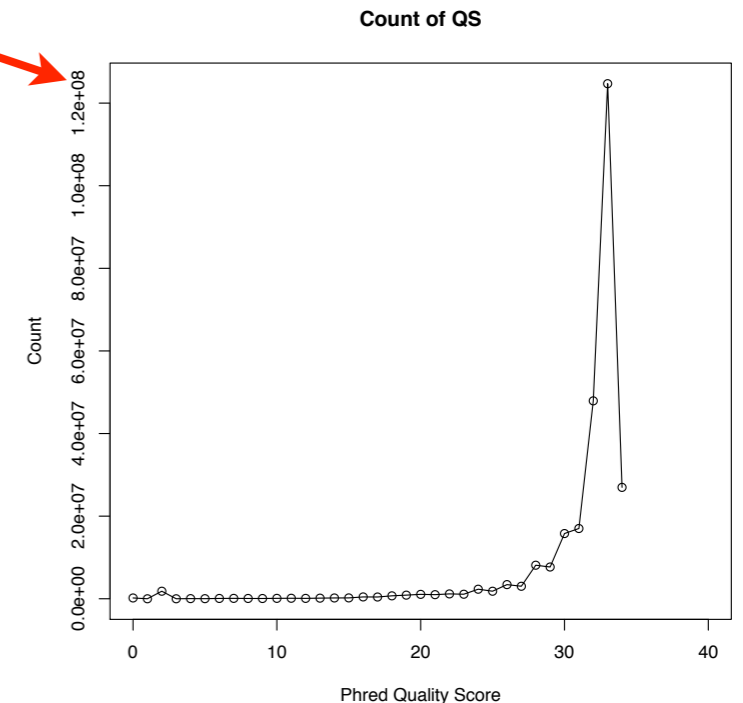
ログの確認

「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



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

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

From Preprocessing output files.

	Filename	Layout	File size
<input type="checkbox"/>	4927_DRR000719_1.unmapped.fastq_4741.bz2 (more 1 files)	paired	65.8 MB
<input type="checkbox"/>	4932_DRR000719_1.unmapped.fastq_4746.bz2 (more 1 files)	paired	65.8 MB
<input type="checkbox"/>	4971_DRR000719_1.unmapped.fastq_4785.bz2 (more 1 files)	paired	65.8 MB
<input checked="" type="checkbox"/>	5509_SRR042533_e.fastq.bz2	single	239.7 MB

DELETE **NEXT**

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

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
<input checked="" type="checkbox"/>	5509_SRR042533_e.fastq.bz2	bp	

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			

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

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 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 --bflyHeapSpaceMax 4G --bflyGCThreads 1 --CPU 4

--single reads.fq --output output_dir

seqType is automatically selected. [fq : for fastq file, fa : for fasta file]


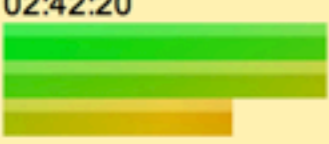
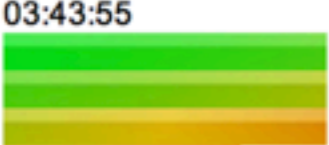

Step2) Create assembled sequences in FASTA file from pileupped reads to [submit WGS division of DDBJ](#).

Set filtered length for contigs

perl lengthfilter.pl pileupFile out_WGS.txt

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
<input type="checkbox"/>	5516	---	Whole transcrip	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 2013-04-30 20:55:45	02:40:24 
<input type="checkbox"/>	5514	koshu01	--- SRR042533 by	S	complete	Trinity	7,420,316	---	View		2013-04-30 18:13:59 2013-04-30 20:56:20	02:42:20 
<input type="checkbox"/>	5508	---	--- Drosophila RNA	S	complete	Trinity	18,524,700	---			2013-04-30 17:23:42 2013-04-30 21:07:38	03:43:55 
<input type="checkbox"/>	5507	---	SRA009364 42CRDAAXX	P	complete	Trinity	9,262,350	---			2013-04-30 15:45:56 2013-04-30 21:07:35	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

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

結果ファイルの統計値

Time

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

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

結果ファイルのダウンロード

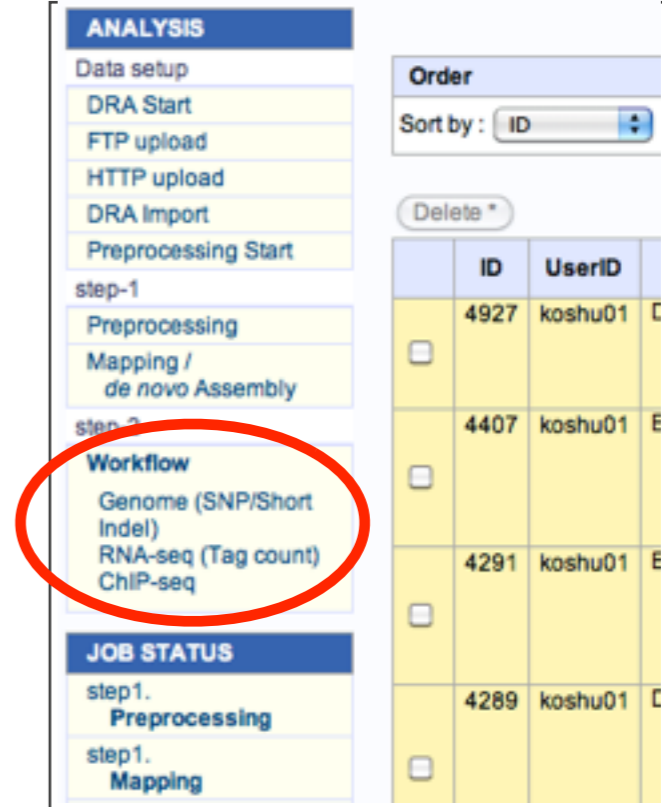
BACK

「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。



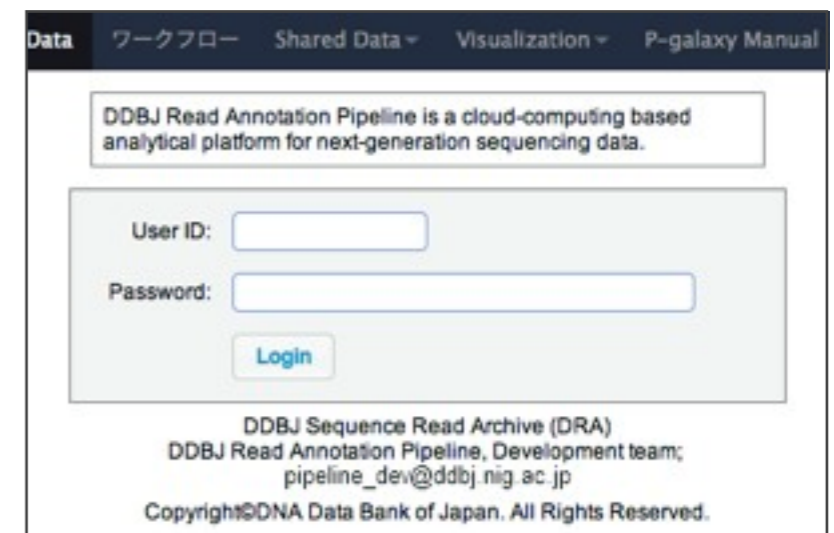
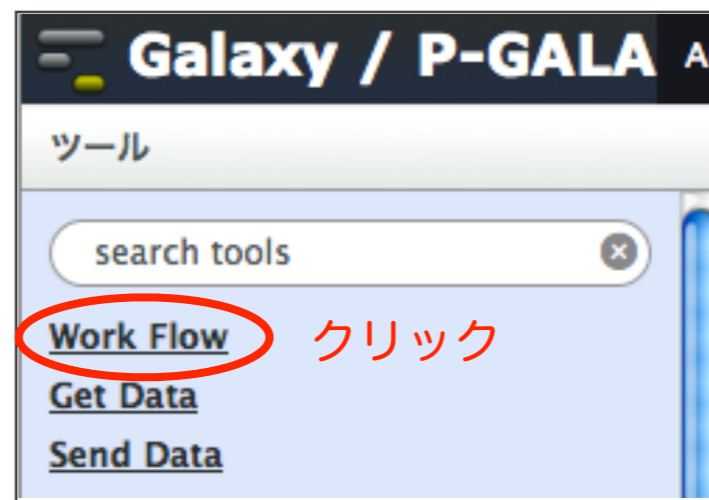
高次解析部(GALAXY)が起動



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。

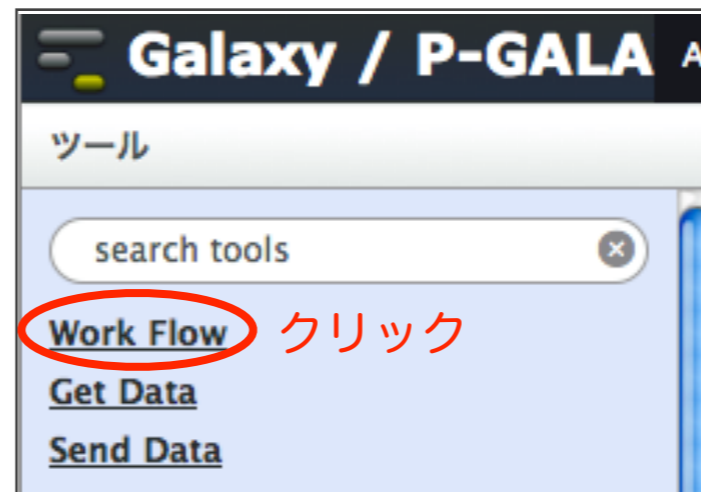
基礎解析と同じパイプライン登録時のメールアドレスとパスワードを入力しても起動可能。



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



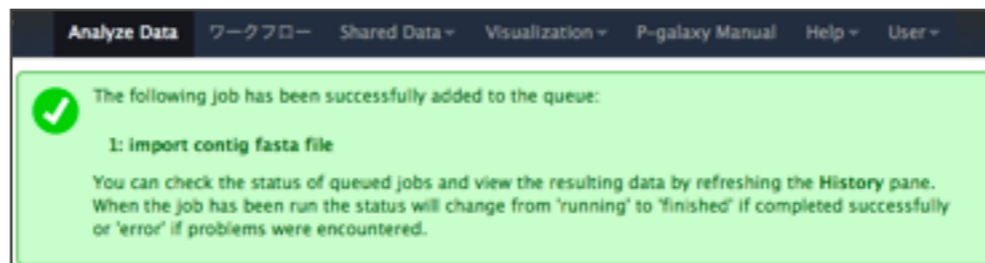
左側「ツール」メニューの「COMMON PROCESS」の下「import contig form DDBJ Pipeline」をクリック

The screenshot shows the Galaxy interface with the 'import contig fasta file' menu item circled in red. To the right, a table lists jobs with their submission accession numbers and 'import' buttons. The 'import' button for job 5514 is circled in red, with the text 'クリック' (click) written next to it. Below the table, the text '「SRR042533」を確認' (check 'SRR042533') is written.

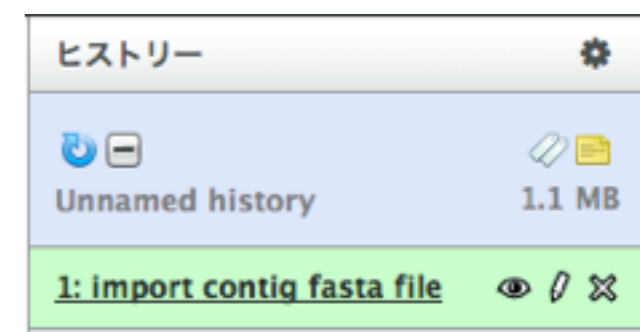
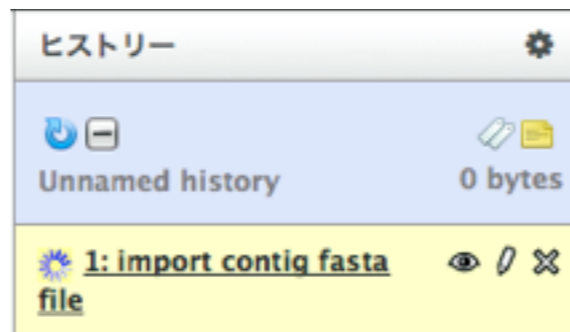
JOBID	Submission Accession RunAlias	Tool	Pipeline Jobpage	Import to Galaxy
5519	SRA009364 42CRDAAXX	trinity	ViewJob	Import
5515	SRA012701 GSM497271_1	trinity	ViewJob	Import
5514	SRR042533 by Preprocessing	trinity	ViewJob	Import

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

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。

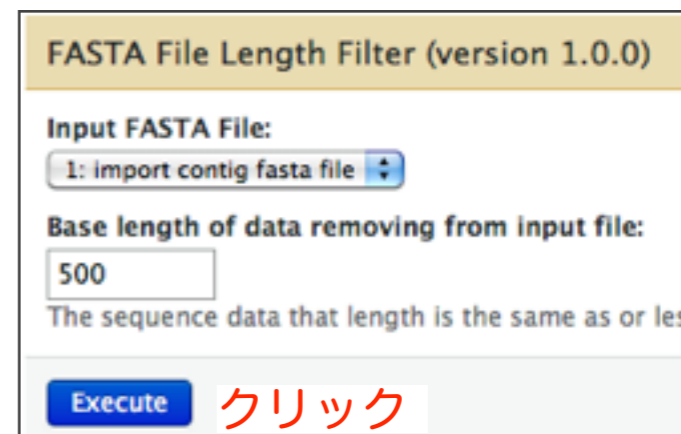


コンティグの長さ調節/アミノ酸変換

左側「ツール」メニューの「Work Flow」の下、さらに「ANNOTATION FOR DE NOVO ASSEMBLED SEQ.」の下に移動



クリック



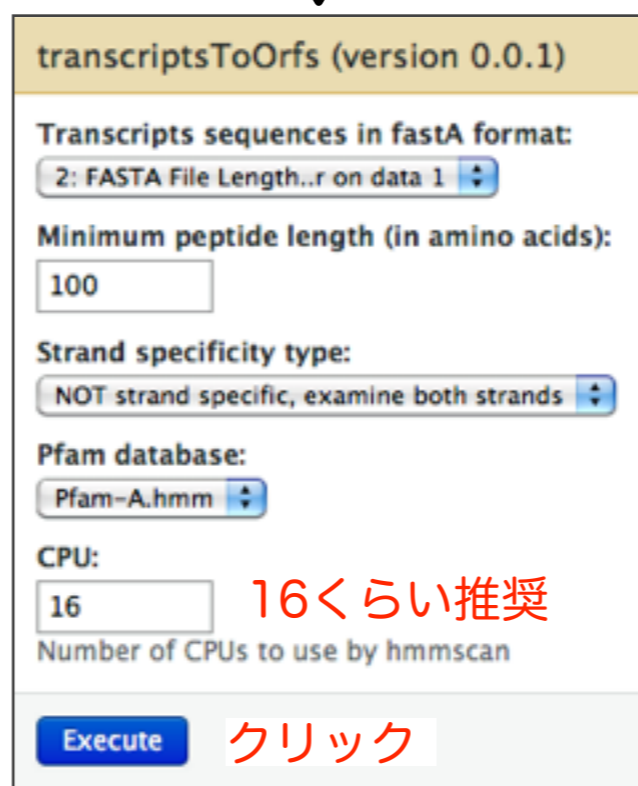
クリック

「FASTA File Length Filter」を今回「500」でもTrinityの--min_contig_lengthオプションで調整可
「Execute」をクリック

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



クリック



16くらい推奨

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。

結果1

```
>m.565 g.565 ORF g.565 m.565 type:internal len:207 (-)...
DLEMQIEGLKEELIFLKKNHEEELLAMRAQMSGQVHVEEAAPAEADLTKVMADIREHYES
ITAKNQKELETWFNSKSEALNKEMMTQVTLQTSRSEVTEVKRSLQALQIELESLLGMKA
SLEGTLDQTONRYSMMLAGYQQQVTSLEQQLVQLRADLVRQGDYQMLLDIKTRLELEIA
EYRRLLEGEAAASSSTSSTSTKTRRL
>m.566 g.566 ORF g.566 m.566 type:complete len:216 (+)...
MAQSVPVVMPKLVLVGDGGTGTTFVVRHLTGEFEKKYVATLGVEVHPLFFNTNRGNVVF
NVWDTAGQEKFGGLRDGYIIQAQCAIIMFDVTSRVTYKNVNPWHRDLVRVCENIPIVLCG
NKVDIKDRKVKAKSIVFHRKKNLQYYDISAKSNYNFEKPFLLWLRKLIIGDPNLEFVEMPA
LAPPEVTMDPALAVQYEKELHVASQTALPDEDDDL*
>m.568 g.568 ORF g.568 m.568 type:internal len:227 (-)...
GDRFKEDRKAKRLPEKSIDMIILLTDGPNSESRIPVIQENVKAAIGGQMSLFLGFGN
DVKYPFLDVMSRENGLARRIYEGSDAALQLQGFYDEVSSPLLLDVLRYPDNAVDSLTT
NQFSQLFNGSEIVVAGRLKDNIDNFPVEVFGQLNDFSEQGFVLDWVSGMYPDDDYIF
GDFTERLWAYLTIQQLLDKSKTGDAEEKANASAEALDMSLRYSFVTP
>m.571 g.571 ORF g.571 m.571 type:5prime_partial len:394...
ASGGEGTHSSCGSWFNAGAKDFPSPYSYLDNFYKCKTSSGEIESYHDVHQRDCRLVS
LLDLALEKDYVRGKVADYMNRLVDMGVAGFRVDACKHMWPGLSAVYGRLLNNTKWFPE
GSRPFIQEVIDLGGEAISYTVYVHLGRVTEFKYGAKLGTVFRKWNNEKLMYTKNWGEW
GFMPNGNAVVFIDNHDNQRGHGAGGAAIVTFWDSRLHKMAVAYMLAHPYGVTRVMSSFRW
NRHIVNGKDQNDWMGPPSHPDGSTKSVPINDETCDGWDVCEHRWRQIKNMVIFRNVVNG
QPHSNWWDNNSNQVAFGRGNRGIIFNDDWDLVTLNLTGLPAGTYCDVISGQKEAGRCT
GKQIHVGS DGRAHFRISNRDEDPFVAIHVESKL*
>m.573 g.573 ORF g.573 m.573 type:5prime_partial len:224...
WEPSPWPQVSLQEYTGFFHFCGSLINENWVVTAAHCNVRTSHRVLILGEHRRSSNNENIQV
MQVGQVFKHPNYSYTIINNDITLIKLASPAQLNIRVSPVCAETSDFPGGMKCVTSGWG
LTRYNAPDTPPRLQOVALPLLTNEECCRKHGWSKITDLMVCAGASGASSCMGDSGGPLVCE
KAGAWTLVGI VSWGSGFCSVSSPGVYARVTMLRAWMDQIIAAN*
```

結果2

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

結果3

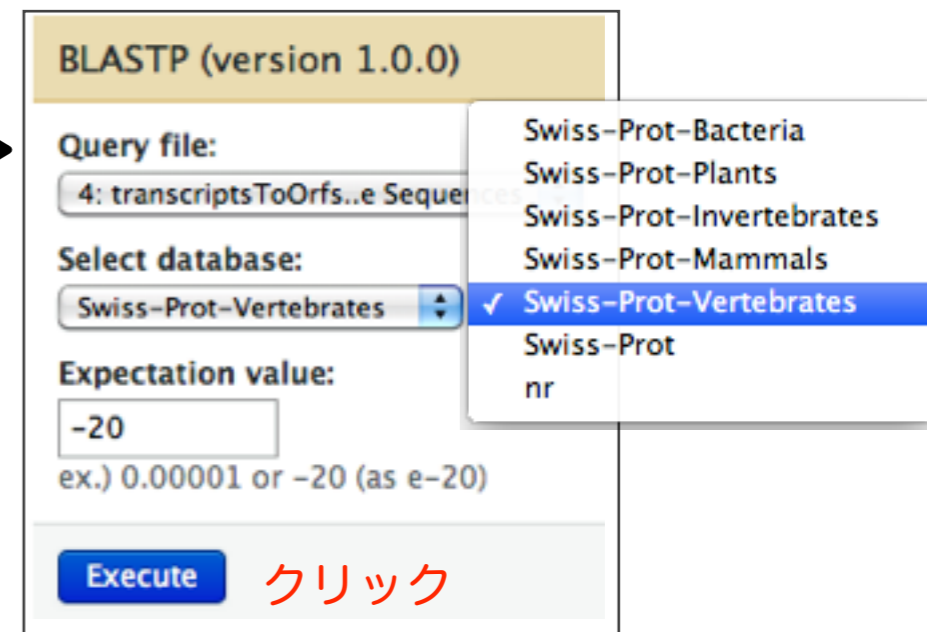
comp1002_c0_seq10	621	ID=m.565;Name=ORF_g.565_m.565_type:internal_len:207_-(g.565_m.565);	0	-	0	621	1	621	0
comp1006_c0_seq137	685	ID=m.566;Name=ORF_g.566_m.566_type:complete_len:216_+(g.566_m.566);	0	+	37	685	1	648	0
comp1010_c0_seq12	683	ID=m.568;Name=ORF_g.568_m.568_type:internal_len:227_-(g.568_m.568);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



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。

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 displays the Galaxy web interface for converting a history into a workflow. The main content area shows the following workflow configuration:

Workflow name: Workflow constructed from history "Drosophila paired 1"

Tools and History Items:

Tool	History items created
import contig fasta file <i>This tool cannot be used in workflows</i>	1: import contig fasta file <input checked="" type="checkbox"/> Treat as input dataset
FASTA File Length Filter <input checked="" type="checkbox"/> Include "FASTA File Length Filter" in workflow	2: FASTA File Length Filter on data 1
transcriptsToOrfs <input checked="" type="checkbox"/> Include "transcriptsToOrfs" in workflow	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
BLASTP <input checked="" type="checkbox"/> Include "BLASTP" in workflow	15: BLASTP on data 4

The right-hand panel shows the history list with a context menu open, highlighting the "Extract Workflow" option. The history list includes items like "7: BLASTP error reports", "6: BLASTP on data 4", "5: transcriptsToOrfs on data 2 Candidate Peptide Coordinates", "4: transcriptsToOrfs on data 2: Candidate Peptide Sequences", "3: transcriptsToOrfs on data 2: Pfam matches to Candidate Peptide Sequences", and "2: FASTA File Length Filter on data 1".

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

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:

* by the guest account.

Manual & tutorial

- [Japanese manual](#)
- [English manual](#)
- [DBCLS tootv Tutorial video 1 \(JP\) - Reference Genome Mapping](#)
- [DBCLS tootv Tutorial video 2 \(JP\) - De novo Assembly](#)

UserIDを決めて必要情報を入力

DDBJ
DNA Data Bank of Japan

Registration form for pipeline user accounts

Note that this account is NOT registered as a NIG supercomputer account.
As DDBJ Pipeline is a webservice of NIG supercomputer, user information was publicly opened to the internet from here. ([Supercomputer User Policy](#))

After registration, you will receive a confirmation email with your user ID and initial password. Please input your email address correctly.

* UserID:
Use 6 to 16 characters.

* Email address:

* Retype email address:
* for confirmation.

* First name:

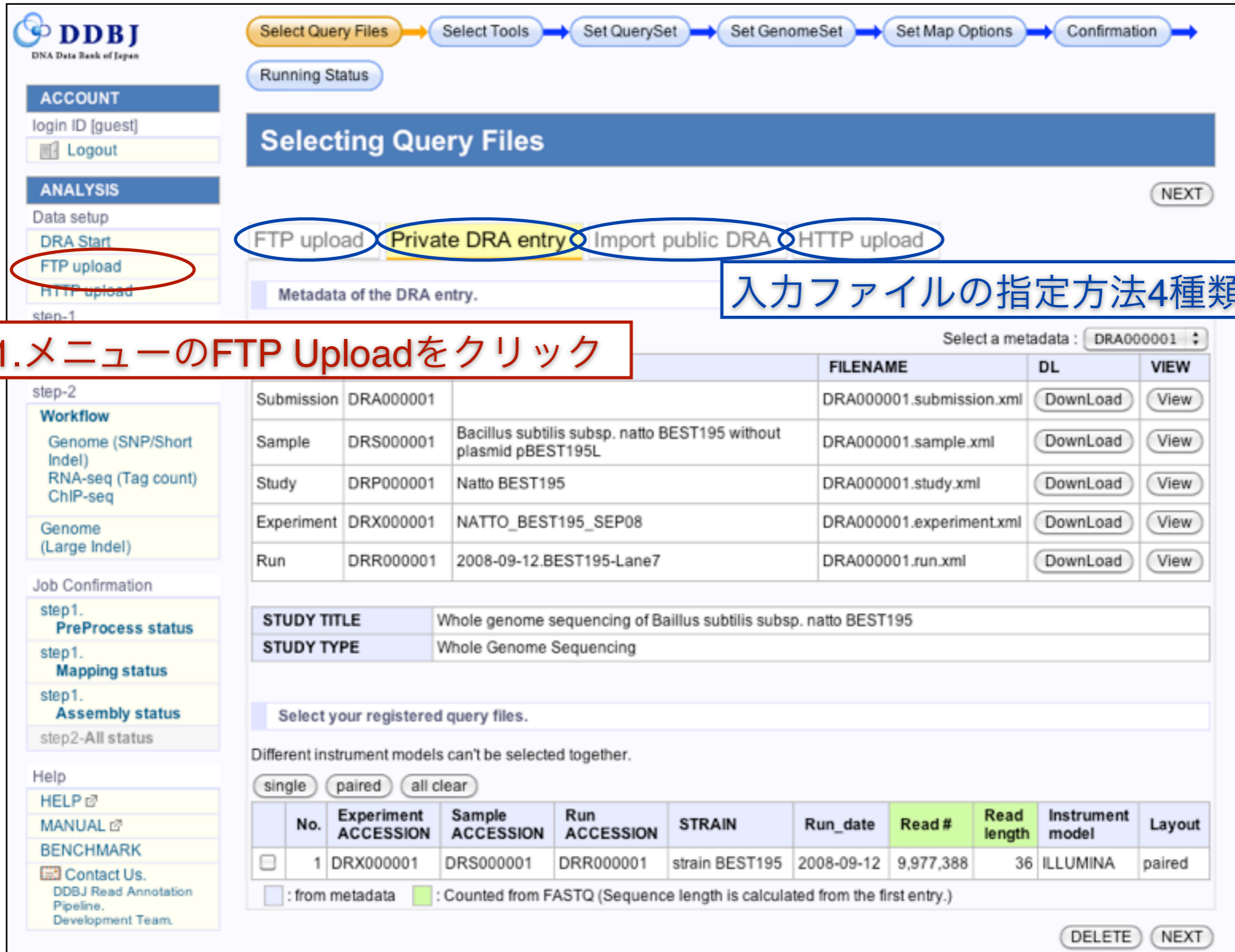
* Last name:

* Institution with department:
ex. Center for Information Biology, National Institute of Genetics.

「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



The screenshot shows the DDBJ FTP Upload interface. A navigation bar at the top includes steps: Select Query Files (highlighted), Select Tools, Set QuerySet, Set GenomeSet, Set Map Options, and Confirmation. A sidebar on the left contains menu items: ACCOUNT (login ID [guest], Logout), ANALYSIS (Data setup, DRA Start, FTP upload (circled in red), HTTP upload), and Job Confirmation (PreProcess status, Mapping status, Assembly status, All status). A central box titled 'Selecting Query Files' contains four options: FTP upload (circled in blue), Private DRA entry (circled in blue), Import public DRA, and HTTP upload (circled in blue). A blue box with white text points to these options, stating '入力ファイルの指定方法4種類' (4 types of input file specification methods). A red box with white text points to the 'FTP upload' menu item, stating '1.メニューのFTP Uploadをクリック' (1. Click FTP Upload in the menu). Below the options, there is a table of metadata for DRA000001, including Submission, Sample, Study, Experiment, and Run details. A legend at the bottom indicates that blue boxes represent files from metadata and green boxes represent files counted from FASTQ.

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	<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"/>

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.)

Query file指定方法

FTP clientによるUpload

Registration of fastq/fasta 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

FTP client Cyberduckの場合

1. <http://cyberduck.ch/>へアクセス



The screenshot shows the Cyberduck website homepage. At the top left is the Cyberduck logo (a yellow duck) and the name "Cyberduck". Below it, the text reads: "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。". A navigation bar contains links for "Cyberduckについて", "ニュース", "更新履歴", "開発", "ヘルプ", and "寄付". Below the navigation bar are several promotional banners. The main content area features a section titled "あらゆるサーバに接続。" (Connect to any server.) with a list of supported protocols and services. At the bottom of the main content area, there is a browser window titled "Amazon S3 (HTTPS)" showing a connection interface with buttons for "Open Connection", "Quick Connect", "Action", "Edit", "Refresh", and "Disconnect". The browser's address bar shows "/cyberduck.ch".

2. ダウンロード



This screenshot shows the download section of the Cyberduck website. It lists two download options:

- 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-4.0b.exe
Windows XP、Windows VistaまたはWindows 7が必要

Below the download information, it states "Downloads hosted by Cacheboy CDN: Open Source Content Delivery." and includes a "寄付" (Donate) button with logos for MasterCard, VISA, and others.

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

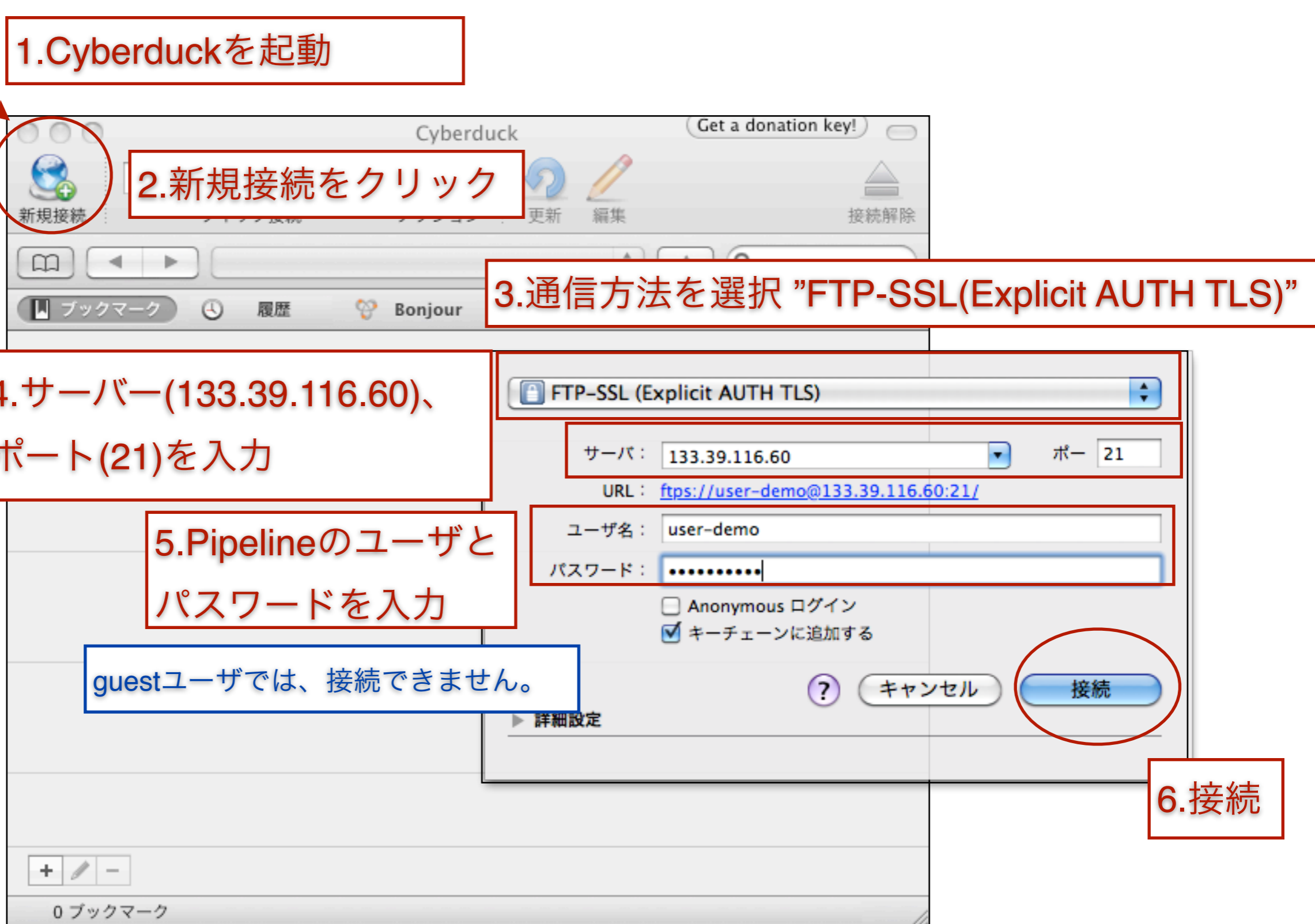
3. 通信方法を選択 "FTP-SSL(Explicit AUTH TLS)"

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

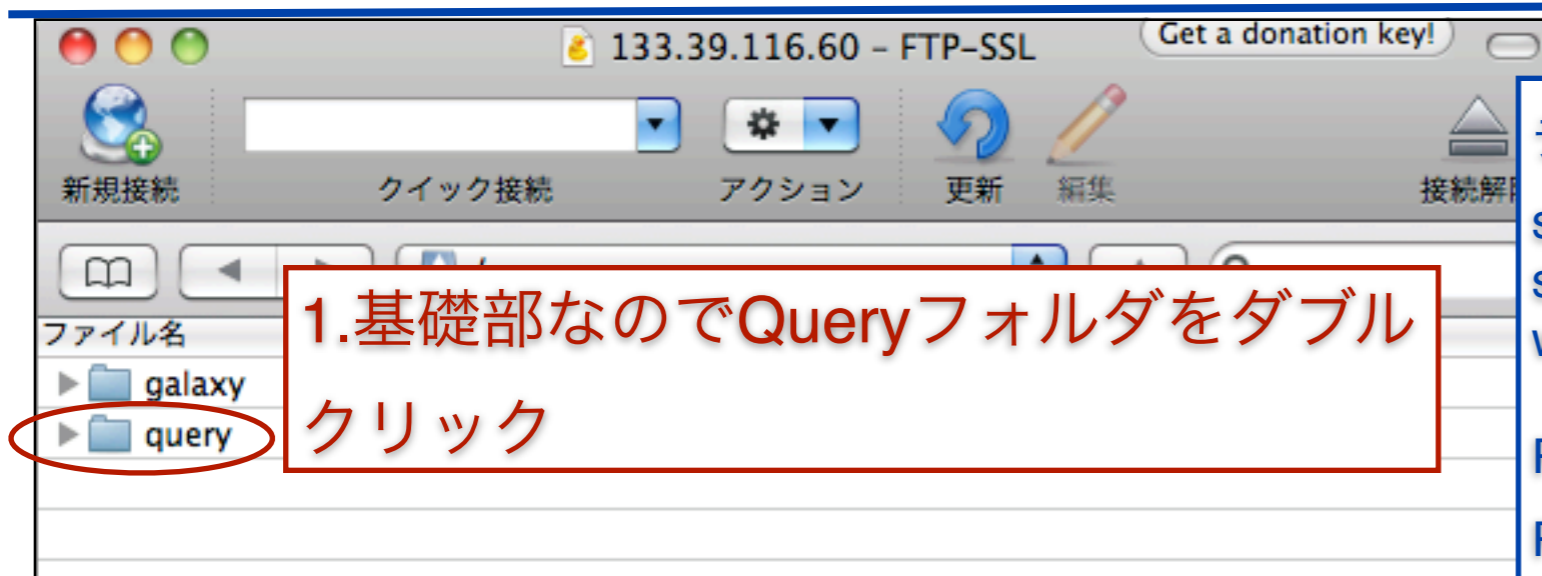
guestユーザでは、接続できません。

6. 接続



The screenshot shows the Cyberduck application window with several red boxes highlighting specific steps in the connection setup process. The interface includes a top menu bar with options like '更新' (Update) and '編集' (Edit), and a main area with a '新規接続' (New Connection) button. The connection details form is visible, showing the selected protocol as 'FTP-SSL (Explicit AUTH TLS)', the server address '133.39.116.60', port '21', and the URL 'ftps://user-demo@133.39.116.60:21/'. The user name is 'user-demo' and the password field is masked with dots. There are checkboxes for 'Anonymous ログイン' (Anonymous Login) and 'キーチェーンに追加する' (Add to Keychain). The '接続' (Connect) button is highlighted with a red circle. A blue box contains a warning message: 'guestユーザでは、接続できません。' (Cannot connect with guest user).

Query file指定方法 Upload

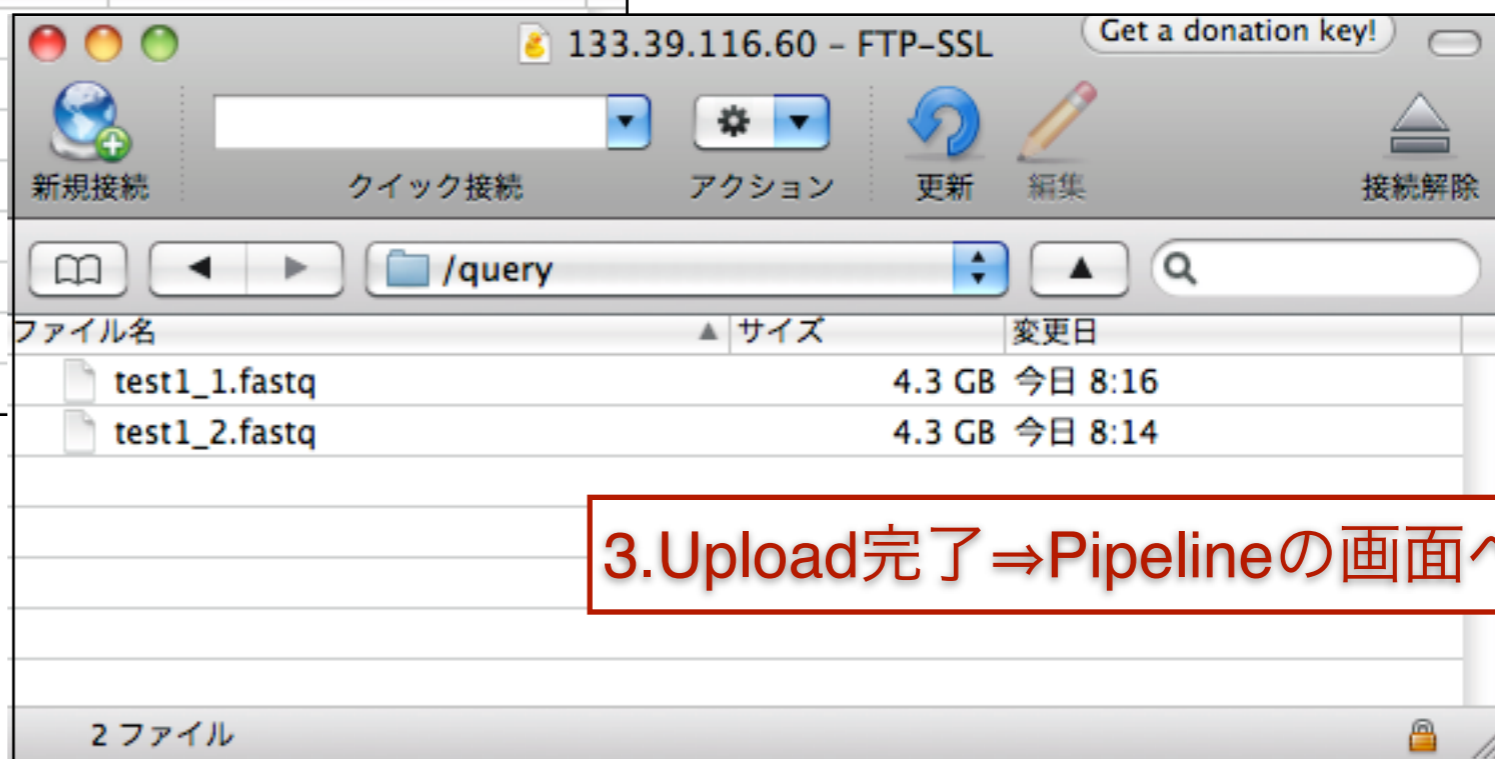
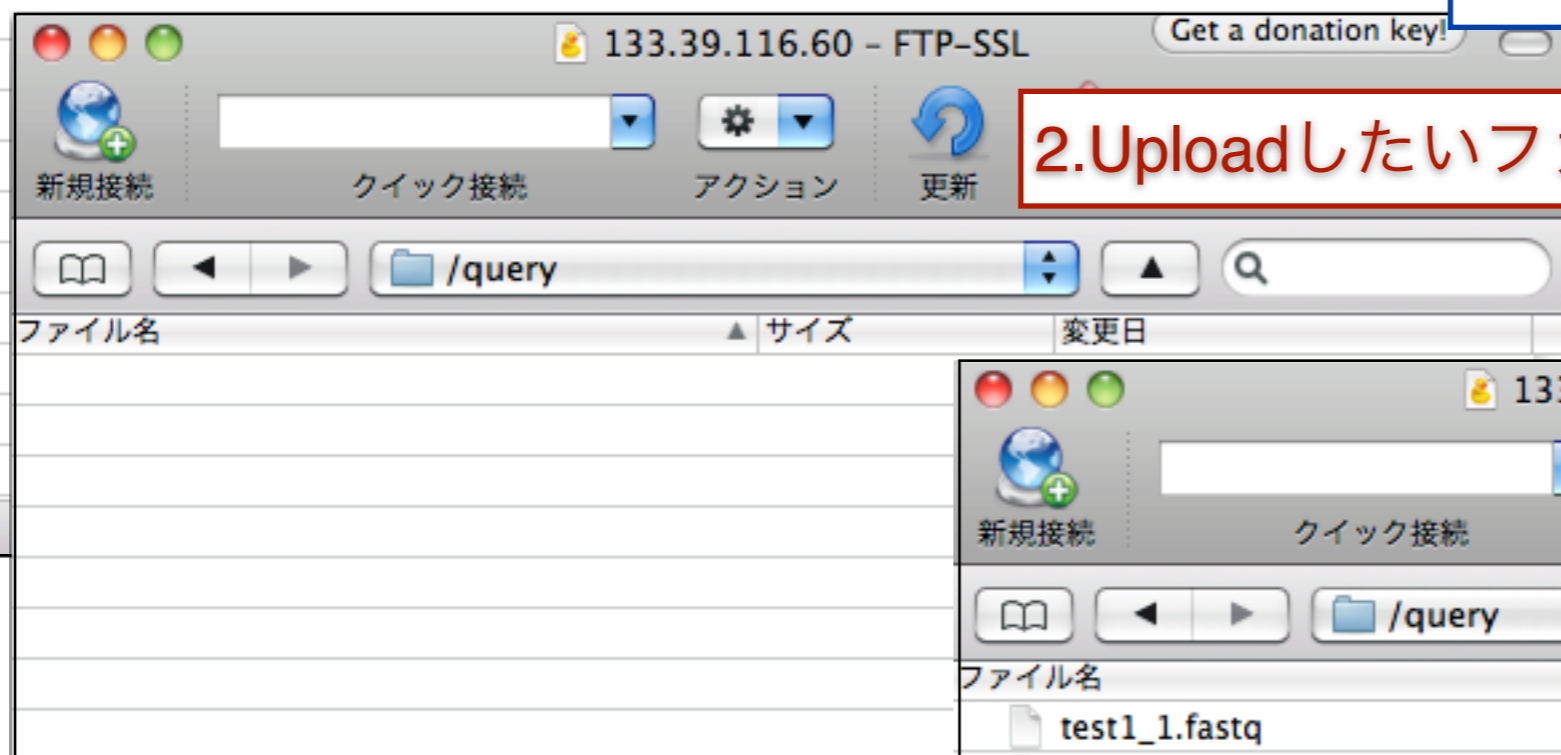


テストデータ :

submission DRA000001
sample Bacillus subtilis subsp. natto BEST195
without plasmid pBEST195L

Read数 : 9,977,388

Read length : 36



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

UploadしたファイルがPaired-endの場合

2. Select a FASTA/FASTQ fileを選択

3. Single-endを選択

3. Paired-endを選択

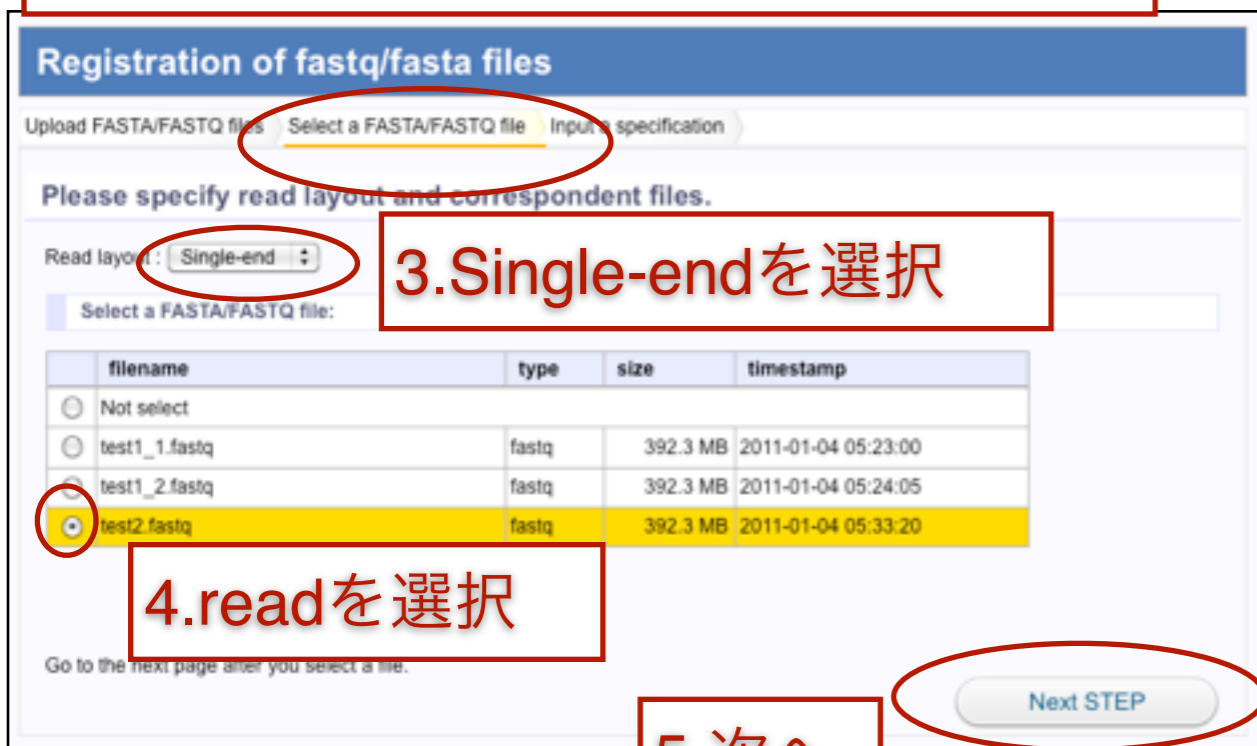
4. readを選択

4. read1 fileを選択

5. 次へ

5. read1 fileと対になるread2 fileを選択

6. 次へ



Registration of fastq/fastq files

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

Please specify read layout and correspondent files.

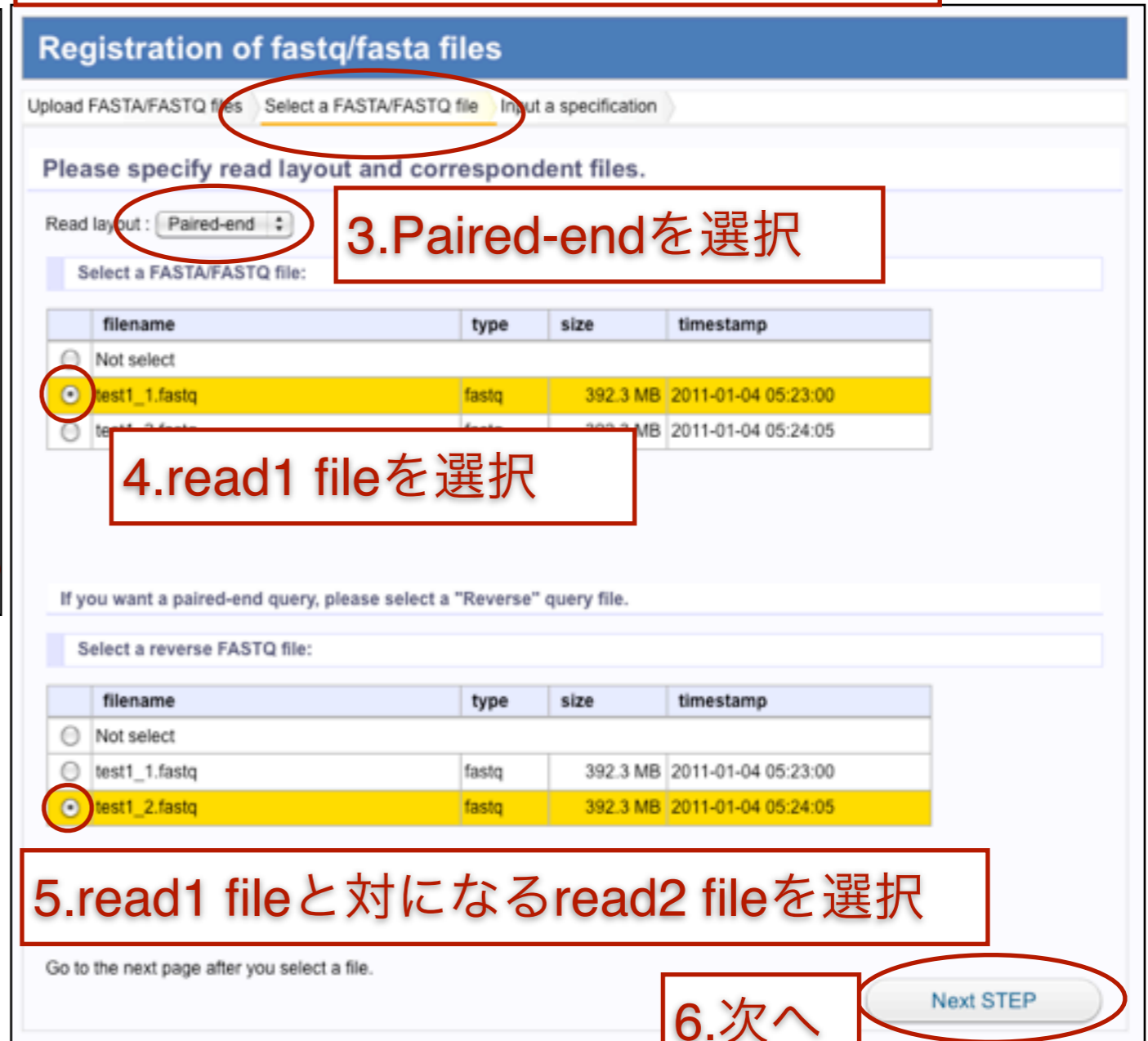
Read layout: **Single-end**

Select a FASTA/FASTQ file:

filename	type	size	timestamp
<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

Go to the next page after you select a file.

Next STEP



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
<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

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

Go to the next page after you select a file.

Next STEP

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.シーケンサの機種を選択

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

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