

次世代シーケンサの活用法

神沼英里¹, 望月孝子¹, 児玉悠一¹, 猿橋智¹, 菅原秀明¹,
大久保公策^{1,2}, 高木利久^{1,2}, 中村保一¹

1 国立遺伝学研究所 生命情報・DDBJ研究センター

2 ライフサイエンス統合データベースセンター

AJACS湘南3

2010年6月4日(金) 14:20-15:20

日本大学生物資源科学部 2号館231教室

目次

次世代シーケンサの活用法

①次世代シーケンサの特徴・原理

- 新旧ゲノム解析技術の比較
- 次世代シーケンサーの特徴
- 次世代シーケンサーの原理

②次世代シーケンサの活用事例

- 解析事例

③次世代シーケンサ・データの登録・解析

- DDBJのアーカイブ・データベース
- DDBJのクラウド型解析パイプライン

④DDBJクラウド型解析パイプラインのデモ

③次世代シークエンサ・データの 登録・解析

DDBJにおける塩基配列データの登録 →論文用アクセッション番号発行

解析データ
の登録

配列データ

DDBJ

国際協力

NCBI
(GenBank)

EBI
(EMBL-Bank)

定量データ

**DDBJ Omics
ARchive(DOR)**

国際協力

GEO

ArrayExpress

生データ
の登録

DTA

(DDBJ Trace Archive)

DRA

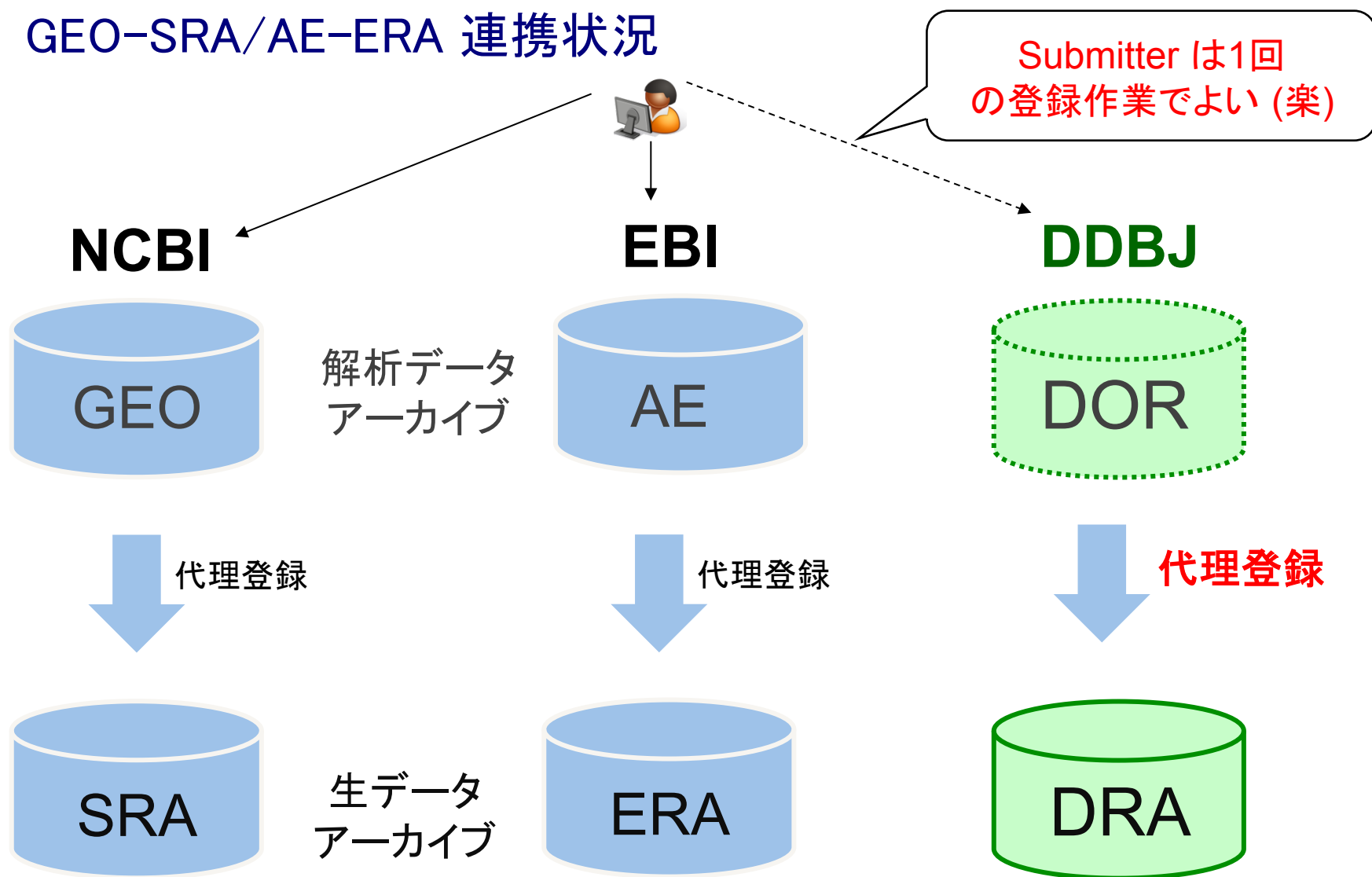
(DDBJ Read Archive)

キャピタリ
シーケンサ

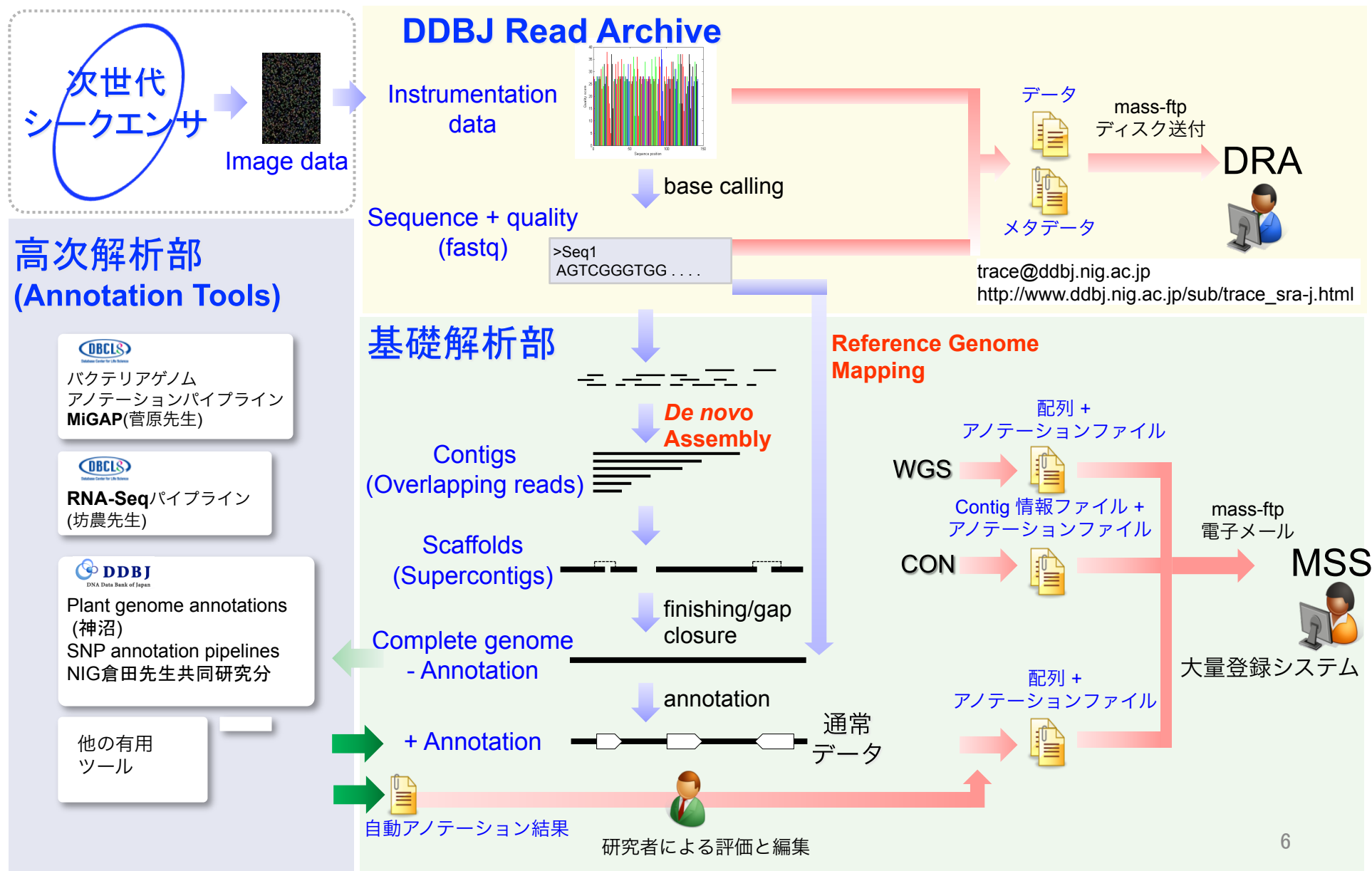
次世代
シーケンサ

DDBJ Omics Archiveから DRAへの代理登録

GEO-SRA/AE-ERA 連携状況



DDBJ Read Annotation Pipeline



FASTQ ファイルと形式とBase Call精度

FASTQ format

@配列ID
Sequence
+配列ID
Sequenceの精度



```
@ID49_20708_20H04AAXX_R1:7:300:39:401
GTCTCGACCAGCCTCGACAACCTC ...
+ID49_20708_20H04AAXX_R1:7:300:39:401
hhUhhhhhYhhhhhhehcaa`BSKhh\XH ...
```

Q score (Phred)	ASCII dec	ASCII Glyph
0	64	@
:	:	
40	104	h

Q-score (evaluation measure of base calls)

Phred quality scores are logarithmically linked to error probabilities

Quality of Phred Score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90 %
20	1 in 100	99 %
30	1 in 1000	99.9 %
40	1 in 10000	99.99 %
50	1 in 100000	99.999 %

Ref. http://en.wikipedia.org/wiki/FASTQ_format

$$Q_{\text{phred}} = -10 \log_{10} p$$

$$Q_{\text{solexa}} = -10 \log_{10} \frac{p}{1-p}$$



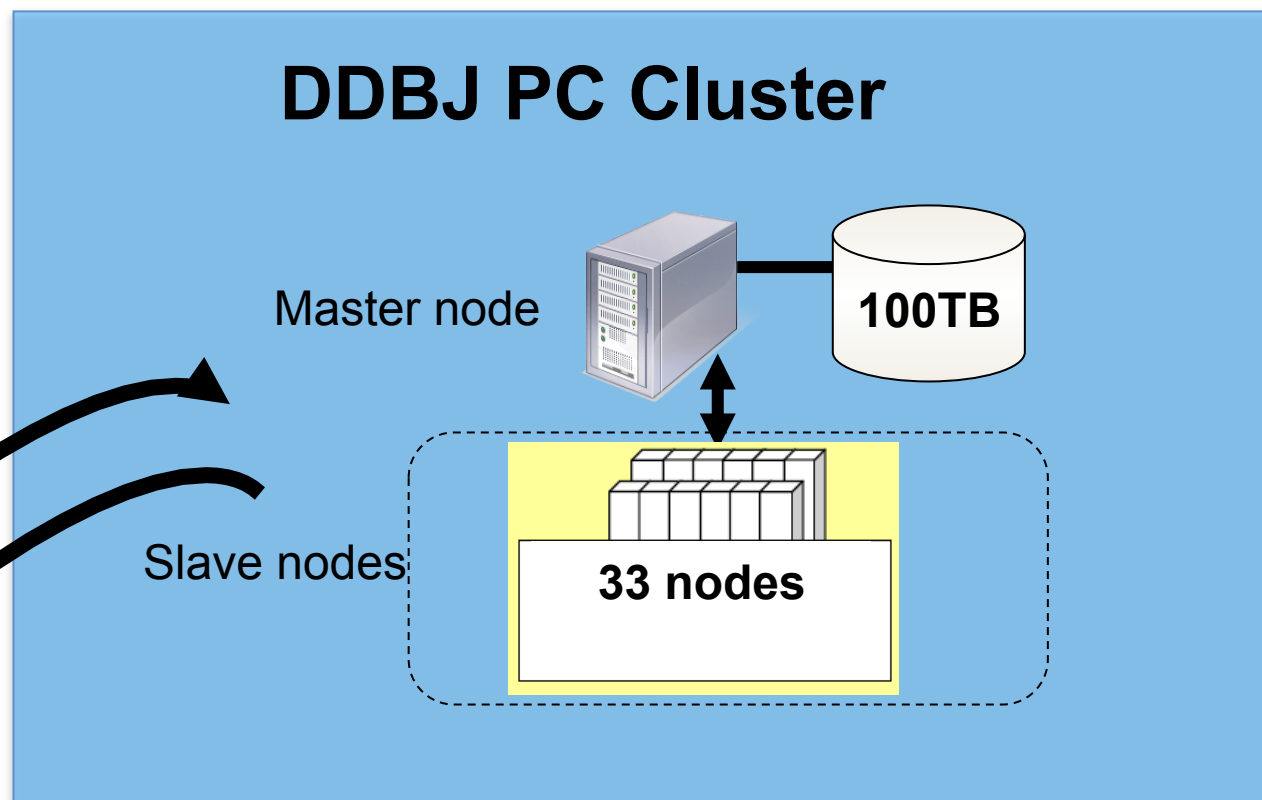
Solexa現バージョン
(illumina 1.3 format)

通常＝Q20以上しか使わない
ラボによりQ15以上使う所もあり。

クラウド型の解析システム

国立遺伝学研究所のスーパーコンピュータで高速化

インターネット上
の資源(クラウド)
で解析



	CPU	Memory	Time	Maximum Job#
Computer nodes	2*3.16GHz	4GB	168H	42

The screenshot shows the DDBJ pipeline web interface. The main content area is titled 'Selecting Query Files'. It includes a table of query files with columns: TYPE, ACCESSION, ALIAS, FILENAME, DL, and VIEW. The table lists several files, including Submission, Sample, Study, Experiment, and Run. A 'NEXT' button is visible next to the table. Below the table, there is a section for 'Select your registered query files.' with a table of query files and a 'NEXT' button.

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

STUDY TITLE: Whole genome sequencing of Baillus subtilis subsp. natto BEST195
STUDY TYPE: Whole Genome Sequencing

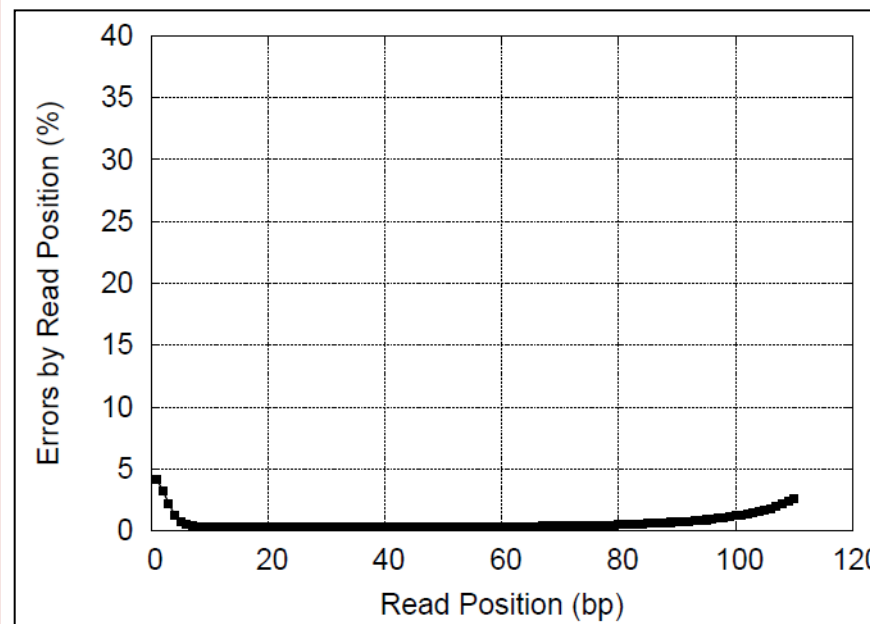
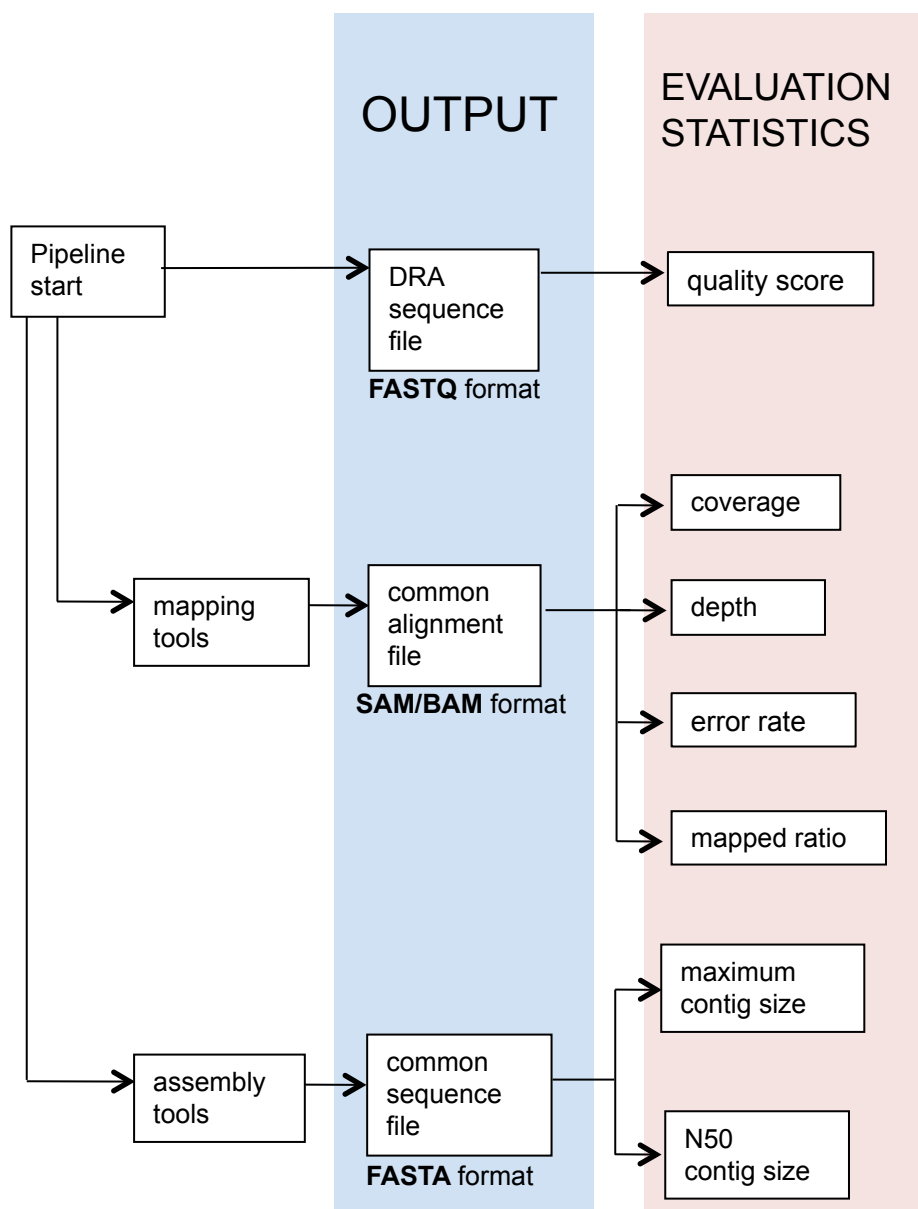
Select your registered query files.
Different instrument models can't be selected together.

single paired all clear

Experiment ACCESSION	Sample ACCESSION	Run ACCESSION	STRAIN	Run_date	Read #	Read length	Instrument model	Layout
<input type="checkbox"/> DRX000001	<input type="checkbox"/> DRS000001	<input type="checkbox"/> DRR000001	BEST195	2008-09-12T16:27:27Z	9,977,388	36	ILLUMINA	paired

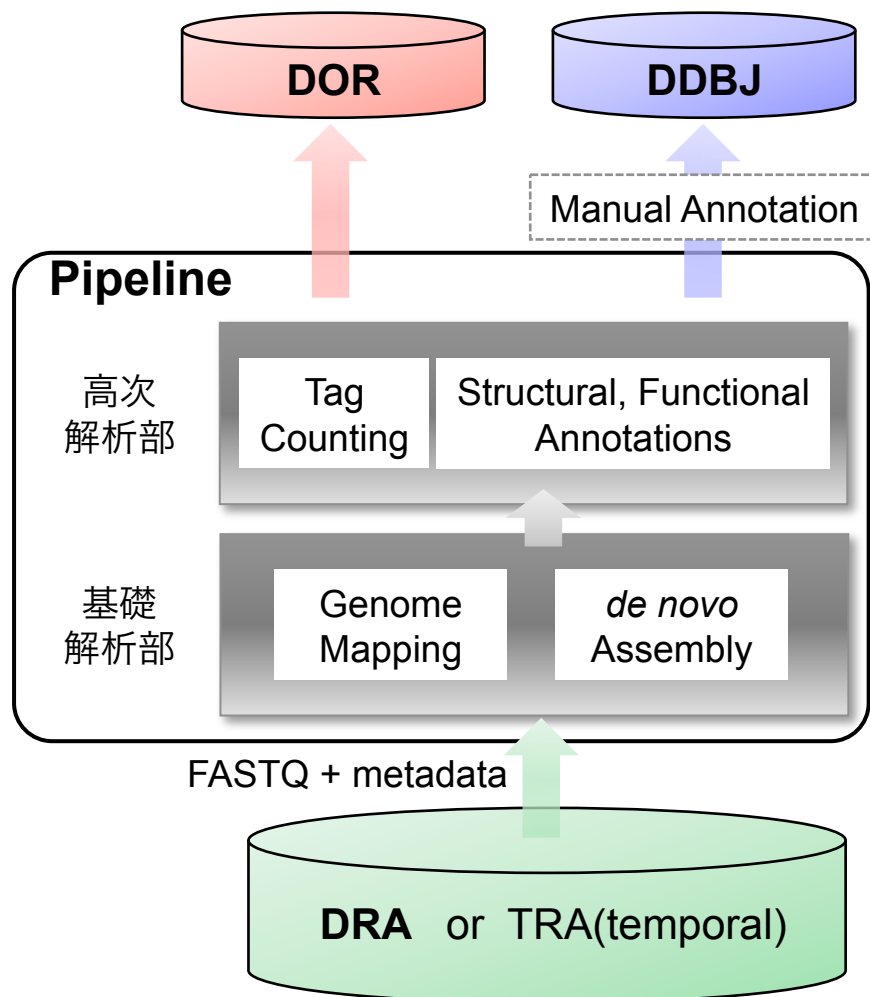
: from metadata : counted from fastq

プログラミング
しなくても解析可

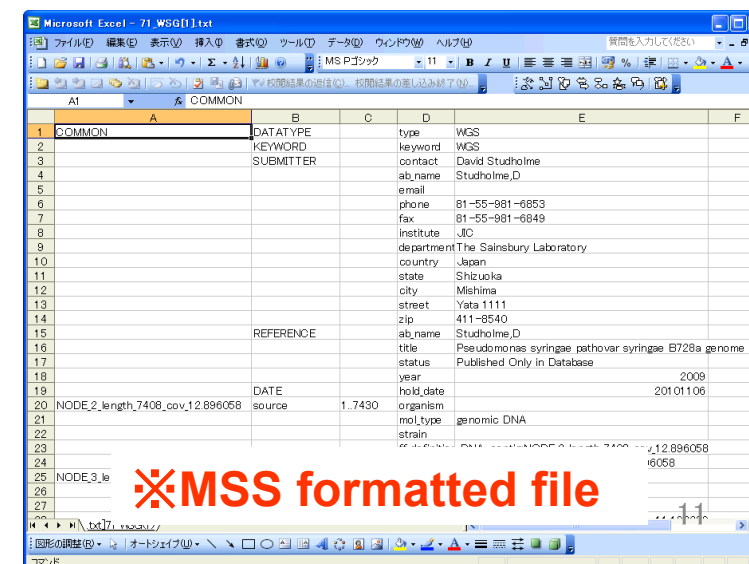
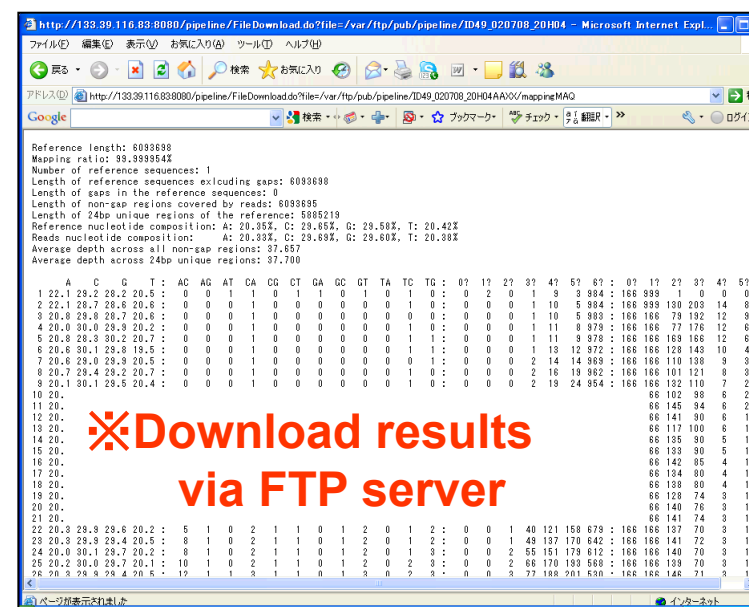


論文用の基本統計量、
図を生成

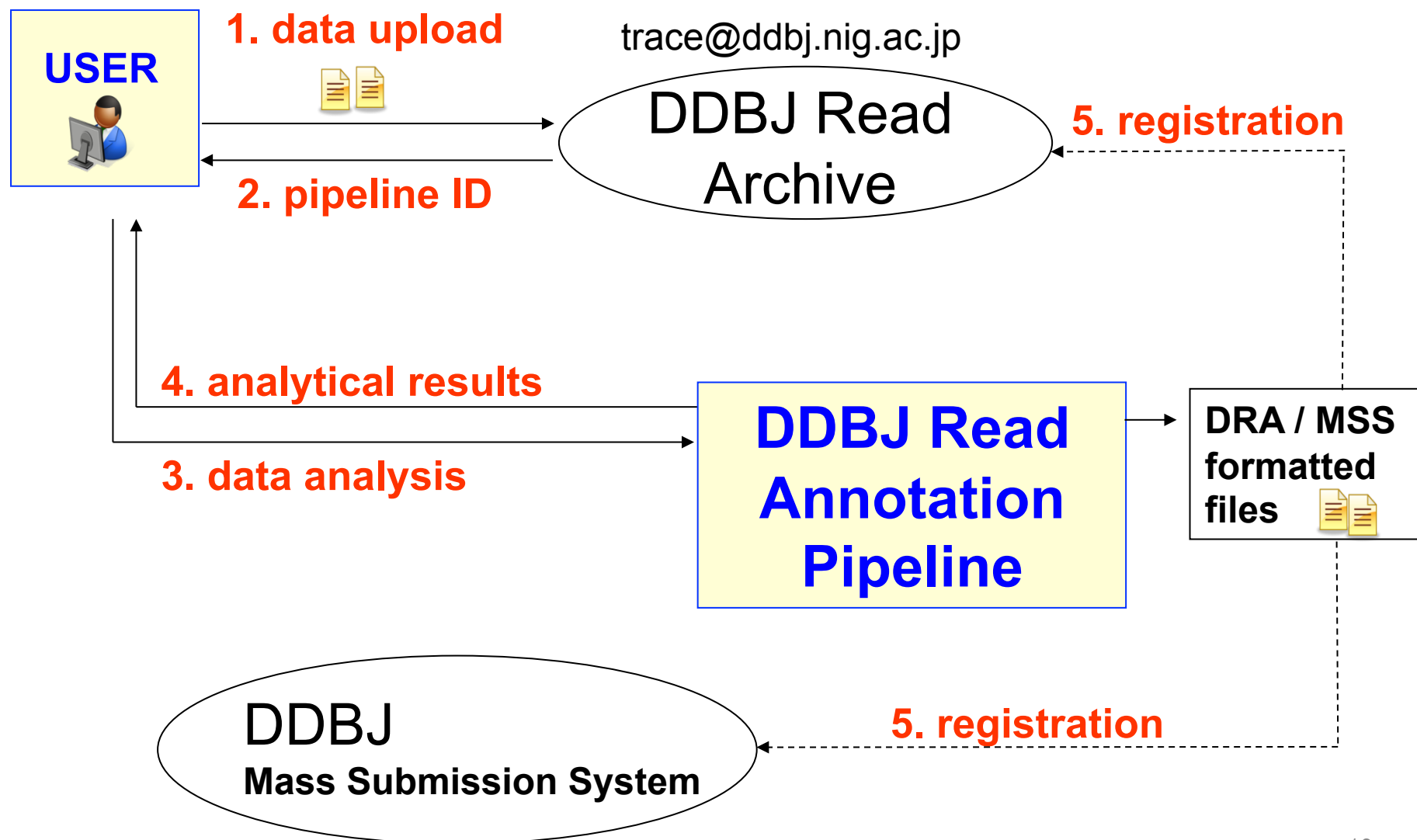
< 配列登録支援 > DDBJ登録ファイル自動生成



解析配列の登録手間を短縮



解析パイプラインの利用には、まず
trace@ddbj.nig.ac.jpへ連絡



④DDBJクラウド型解析パイプライン のデモ

解析パイプライン

DDBJ Read Annotation Pipeline

解析パイプライン <https://p.ddbj.nig.ac.jp/>

User ID: guest (パスワードなし)
でデモ画面を参照できます。

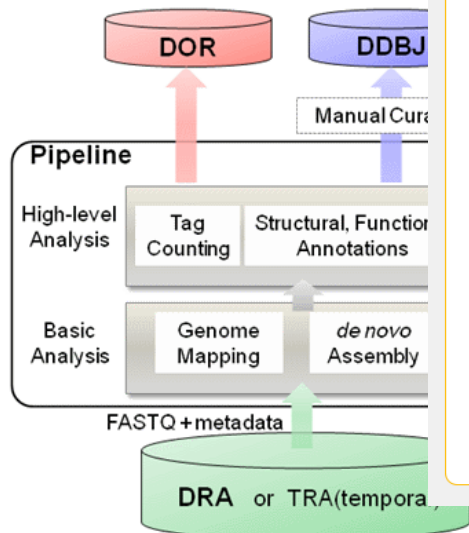
LOGIN

User ID:

Password:

Login

If you log in to the system as an anonymous user...



「接続の安全性を確認できません」の画面が出てきたら
セキュリティ例外を承認して下さい。

接続の安全性を確認できません

sso.ddbj.nig.ac.jp に安全に接続するように求められましたが、接続の安全性が確認できませんでした。

安全に接続する場合は通常、あなたが適切な相手と通信することを確認できるように信頼できる証明書を提供してきます。しかし、このサイトの証明書は信頼性を検証できません。

どうすればよいのか？

これまでこのサイトに問題なく接続できていた場合、このエラーが表示される可能性があります。接続すべきではありません。

スタートページに戻る

技術の詳細を表示

危険性を理解した上で接続するには

何が起きていて何が問題なのか理解できているのであれば、このサイトの証明書をFirefoxにセキュリティ例外を追加することもできます。ただし、たとえこのサイトであっても、誰かが通信を改ざんしているからこのエラーが表示されていることに注意してください。

信頼できる証明書をこのサイトが使用しない正当な理由がない限り、例外として承認してください。

例外を追加...

URL: 証明書を取得

証明書の状態

このサイトでは不正な証明書が使用されており、サイトの識別情報を確認できません。

不明な証明書です

既知の認証局によって検証されていないため、このサイトの証明書は信頼されません。

☒ 次回以降にもこの例外を有効にする

セキュリティ例外を承認

キャンセル

クエリ用FASTQ形式READファイルの選択

DNA Data Bank of Japan

[Select Query Files](#)
[Select Tools](#)
[Set QuerySet](#)
[Set GenomeSet](#)
[Set Map Options](#)
[Confirmation](#)

[Running Status](#)

MENU

- USER INFO
- STATUS
- BENCHMARK
- HIGH LEVEL ANALYSIS
- MANUAL

LOGIN INFO

ID[guest]

- Logout

[feedback](#)

☒ **DDBJ Read Archive**
☐ Upload fasta file

Selecting Query Files

[NEXT](#)

1. List of data

TYPE	ACCESSION	ALIAS	FILENAME	DL	VIEW
Submission	ERA000095		ERA000095.submission.xml	Download	View
Sample	ERS000165	PssB728a_July_2008	ERA000095.sample.xml	Download	View
Study	ERP000055	PssB728a_genome	ERA000095.study.xml	Download	View
Experiment	ERX000536	PssB728a_assembly	ERA000095.experiment.xml	Download	View
Run	ERR005143	ID49_020708_20H04AAXX	ERA000095.run.xml	Download	View

Download and view of DRA metadata

STUDY TITLE	Pseudomonas syringae pathovar syringae B728a genome sequencing
STUDY TYPE	Whole Genome Sequencing

Select your registered query files.
Different instrument models can't be selected together.

☐ single
 ☐ paired
 ☐ all clear

Experiment ACCESSION	Sample ACCESSION	Run ACCESSION	STRAIN	Run_date	Read #	Read length	Instrument model	Layout
<input checked="" type="checkbox"/> ERX000536	ERS000165	ERR005143		2008-02-07T12:00:00.000Z	3,551,133	36	ILLUMINA	paired

☐ : from metadata
 ☒ : counted from fastq

2. Select query files

3. Go to next

[NEXT](#)

MENU

- USER INFO
- STATUS
- BENCHMARK
- HIGH LEVEL ANALYSIS
- MANUAL

LOGIN INFO

ID[guest]
Logout

feedback

Selecting Tools for Basic Analysis of DDBJ ANNOTATION PIPELINE

BACK NEXT

Reference Genome Mapping

1. Select map or assembly

	Tool	Help	Version	Base space	Color space	Paired end	Depth	Coverage	Error rate	Analysis	Output format	Comment
<input type="checkbox"/>	BLAT		34	✓			✓	✓	✓			Single-end analysis only
<input checked="" type="checkbox"/>	Maq		0.7.1	✓		✓	✓	✓	✓	✓	✓	
<input type="checkbox"/>	bwa		0.4.9	✓		✓	✓	✓	✓			
<input type="checkbox"/>	SSAHA2		2.3.0.1	✓		✓				✓		SNP is single-end analysis only
<input type="checkbox"/>	SOAP		2.1.8	✓		✓	✓	✓	✓	✓		
<input type="checkbox"/>	Bowtie (SAMtools)		0.12.0 (0.1.7)	✓	✓	✓	✓	✓	✓			
<input type="checkbox"/>	TopHat		1.0.11 (BETA)	✓		✓	✓	✓	✓			

de novo Assembly

Assembly tool can be selected up to two.

	Tool	Help	Version	Paired-end	MSS(WGS)	Comment
<input type="checkbox"/>	velvet		0.7.56	✓	✓	
<input type="checkbox"/>	edena		2.1.1			
<input type="checkbox"/>	abyss		1.1.0			Abyss is built using Google sparsehash. (Abyss isn't built the parallel assemble with MPI support.) The maximum value for k is 64.

☐ The contigs will be aligned to reference genome.

Tool	Comment
<input type="radio"/> BLAT	

3. Go to next

BACK NEXT

2. Select a tool

MENU

- USER INFO
- STATUS
- BENCHMARK
- HIGH LEVEL ANALYSIS
- MANUAL

LOGIN INFO

ID[guest]
Logout

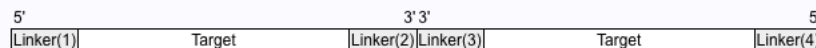
feedback

Generating Query Sets from Query Read Files

RESET BACK NEXT

Paired-end analysis

Layout of paired sequence. 5'-3' 3'-5'
Linker sequences are already removed.



Run	ACCESSION	Read length	Quality Score
<input checked="" type="checkbox"/>	ERR005143	36	Read1 Read2

☐ Mapping tool is used within the range of bp.
Example : When you use the tool within the range of 50 bp.



Only this range is used.

QUERY SET

1. Select files for a query set

2. Confirmation

confirm

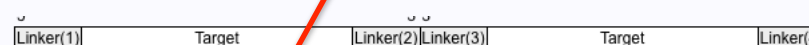
Files

RESET BACK NEXT

RESET BACK NEXT

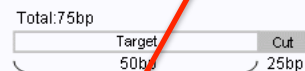
Logout

feedback



Run	ACCESSION	Read length	Quality Score
-----	-----------	-------------	---------------

☐ Mapping tool is used within the range of bp.
Example : When you use the tool within the range of 50 bp.



Only this range is used.

confirm

QUERY SET

ERR005143 -><-

3. Go to next

RESET BACK NEXT

17

MENU

- USER INFO
- STATUS
- BENCHMARK
- HIGH LEVEL ANALYSIS
- MANUAL

LOGIN INFO

ID[guest]
Logout

feedback

Specifying Database of Reference Genome

RESET BACK NEXT

Major genome sets

Organisms: Arabidopsis thaliana

Genome sets: TAIR8

all check all clear

- ☒ chr1.fas
- ☒ chr2.fas
- ☒ chr3.fas
- ☒ chr4.fas
- ☒ chr5.fas
- ☒ chrC.fas
- ☒ chrM.fas

User original sets

Download or upload reference

Retrieving a chromosome from DDBJ-DB by using SOAP

Input Accession Number (INSD) or (RefseqID)
CP000075

LOAD

SOAP(Simple Object Access

PIPELINE



2. Specify an INSD accession number to retrieve data from DDBJ DB

1. Specify preset, or download/upload

Selecting query DRA read. x

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

MENU

- USER INFO
- STATUS
- BENCHMARK
- MANUAL

LOGIN INFO

ID[user-demo]
Logout

feedback

Specifying Database of Reference Genome

RESET BACK NEXT

☐ Major genome sets

Genome sets: [dropdown]

☐ User original sets

Genome sets: [dropdown]

☒ **Download or upload reference**

Retrieving a chromosome from DDBJ-DB by using SOAP

Input Accession Number (INSD) or (RefseqID)
CP000075

LOAD

SOAP(Simple Object Access protocol)

PIPELINE

Request

Data

DDBJ-DB

Uploading reference from local drive.

Select a FASTA file. ファイルを選択 選択されていません

UPLOAD

☒ >CP000075|Pseudomonas syringae pv. syringae B728a, complete genome.

DELETE

CREATE DATASET

RESET BACK **NEXT**

3. Go to next

Setting for Reference Gen...

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

MENU

- USER INFO
- STATUS
- BENCHMARK
- MANUAL

LOGIN INFO

ID[guest]

- Logout

feedback

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

Running Status

Setting for Reference Genome Mapping

BACK NEXT

maq

Set optional parameters of the paired-end analysis

Step1) Convert sequences

```
Maq fasta2bfa refgenome.fasta refgenome.bfa
Maq fastq2bfq query1.fastq query1.bfq
Maq fastq2bfq query2.fastq query2.bfq
```

Step2) Map

```
Maq map -n 2 -m 0.005 -C 513 out.aln refgenome.bfa query1.bfq query2.bfq &> map.log
Maq mapmerge out_all.map out_all.map out_1.map out_2.map
```

Step3) Display the read alignment in plain text

```
Maq mapview out.map > mapview.bt
```

Step4) Read quality check

```
Maq mapcheck refgenome.fasta.bfa out.map > mapcheck.bt
```

Step5) Convert the read alignment to .bed and .gff format

```
bedConvert mapview.bt out.bed
gffConvert mapview.bt out.gff
```

Step6) Analysis for depth,coverage

```
CoverageChecker refgenome.fasta mapview.bt > CoverageCheck.bt
Depthchecker refgenome.fasta mapview.bt maq > DepthCheck.bt
```

SNP and Indels analysis

Step7) Indels analysis

```
Maq indelsoa refgenome.bfa out.aln > out.indel.soa
```

1. Go to next

MENU

- USER INFO
- STATUS**
- BENCHMARK
- HIGH LEVEL ANALYSIS
- MANUAL

LOGIN INFO

ID[user-demo]

- Logout

feedback

Run Confirmation

BACK

RUN

Destination of mail

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

XXXXX@ddbj.org

* Required

Reference Genome Map [maq]

Query sets

ERR005143 -><-

genome sets

- >CP000075|Pseudomonas syringae pv. syringae B728a, complete genome.

Command Options

maq

Set optional parameters of the paired-end analysis

Step1) Convert sequences

Maq fasta2bfa refgenome.fasta refgenome.bfa
Maq fastq2bfq query1.fastq query1.bfq
Maq fastq2bfq query2.fastq query2.bfq

Step2) Map

Maq map -n 2 -m 0.005 -C 513 -a 500 out.aln refgenome.bfa query1.bfq query2.bfq &> map.log
Maq mapmerge out_all.map out_all.map out_1.map out_2.map

Step3) Display the read alignment in plain text

1. Run

**2. Go to
'status' page**

1. Screen login user

2. Goto 'Download' page

The screenshot shows the 'Status - Mapping' page of the pddb.nig.ac.jp/pipeline/Status.do application. The page features a navigation menu on the left with links for USER INFO, STATUS, BENCHMARK, and MANUAL. A 'Running Status' button is located at the top. The main content area displays a table of mapping results. A red circle highlights the 'Only the login user' checkbox in the 'Order' section, and another red circle highlights the 'File' button in the 'Download' column of the table. An arrow points from the first circle to the second, indicating the next step in the process.

ID	UserID	Submission accession	P/S	Status	Tool	Read #	Read length	Genome size	File	Start time	End time	Elapsed time
1261	---	DRA000039 DRR000149	S	complete	TopHat	17,003,885	35	4 M		2010-03-11 20:25:30	2010-03-12 06:48:41	10:23:10
1253	---	ERA000095 ERR005143	P	complete	TopHat	3,535,967		6 M		2010-03-11 12:20:25		00:37:33

ID	UserID	Submission accession	P/S	Status	Tool	Read #	Read length	Genome size	Download	Start time	End time	Elapsed time
1265	guest	ERA000095 ERR005143	P	complete	Maq	3,551,133	36	6 M	File	2010-03-15 18:08:37	2010-03-15 18:40:40	00:32:03

- USER INFO
- STATUS
- BENCHMARK
- HIGH LEVEL ANALYSIS
- MANUAL 

LOGIN INFO

ID[guest]

- Logout

[feedback](#)

Detail view

[BACK](#)

Job info

ID

1265

Tool (Version)

Maq (0.7.1)

Query set

[ERR005143](#)

Genome set

>CP000075|Pseudomonas syringae pv. syringae B728a, complete genome.

Chromosome

[CP000075_1264555596619](#)

Position errors

[download](#)

Coverage

[download](#)

Depth

[download](#)

Map ratio

[download](#)

Time

Wait time

0: 0:42

Start time

2010-03-15 18:08:37

End time

2010-03-15 18:40:40

Result - ERR005143

[Top of page](#)

[ERR005143 download](#)

[CP000075_1264555596619]

Command

maq fasta2bfa CP000075_1264555596619
CP000075_1264555596619.bfa

Start time

2010-03-15
18:08:47

End time

2010-03-15
18:08:57

Log1

Log2

[View](#)

Result

[download](#)

[ERR005143]

Command

ERR005143 > CP000075_1264555596619

Start time

2010-03-15
18:09:07

End time

2010-03-15
18:09:38

Log1

Log2

[View](#)

Result

[download](#)

maq fastq2bfq ERR005143_1_0000 ERR005143_1_0000.bfq

2010-03-15
18:09:48

2010-03-15
18:10:18

[View](#)

[download](#)

maq fastq2bfq ERR005143_2_0000 ERR005143_2_0000.bfq

2010-03-15
18:10:29

2010-03-15
18:10:49

[View](#)

[download](#)

maq fastq2bfq ERR005143_1_0001 ERR005143_1_0001.bfq

2010-03-15
18:10:29

2010-03-15
18:10:49

[View](#)

[download](#)

Viewerを起動(Tablet)



<http://bioinf.scri.ac.uk/tablet/>



1. ローカルディレクトリにTabletをインストール Plant Bioinformatics Group

Tablet

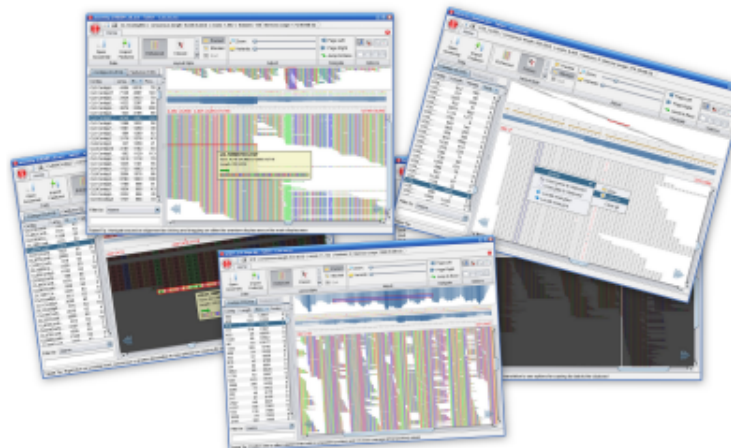
- Tablet Homepage
- **Download Tablet**
- Screenshots
- Tablet FAQ
- Sample Data
- Assembly Conversion
- Papers and Presentations
- Privacy Policy
- Tablet World Map
- Online Help

Our Software

- CurlyWhirly
- Flapjack
- OPTIRas
- Strudel
- Tablet (new)
- TetraploidMap

Tablet - Next Generation Sequence Assembly Visualization

Tablet is a lightweight, high-performance graphical viewer for next generation sequence assemblies and alignments.



Mapping結果をダウンロード

Detail view

BACK

Job info

ID

1265

Tool (Version)

Maq (0.7.1)

Query set

[ERR005143](#)

ID49_020708_20H04AAXX

Genome set

>CP000075|Pseudomonas syringae pv. syringae B728a, complete genome.

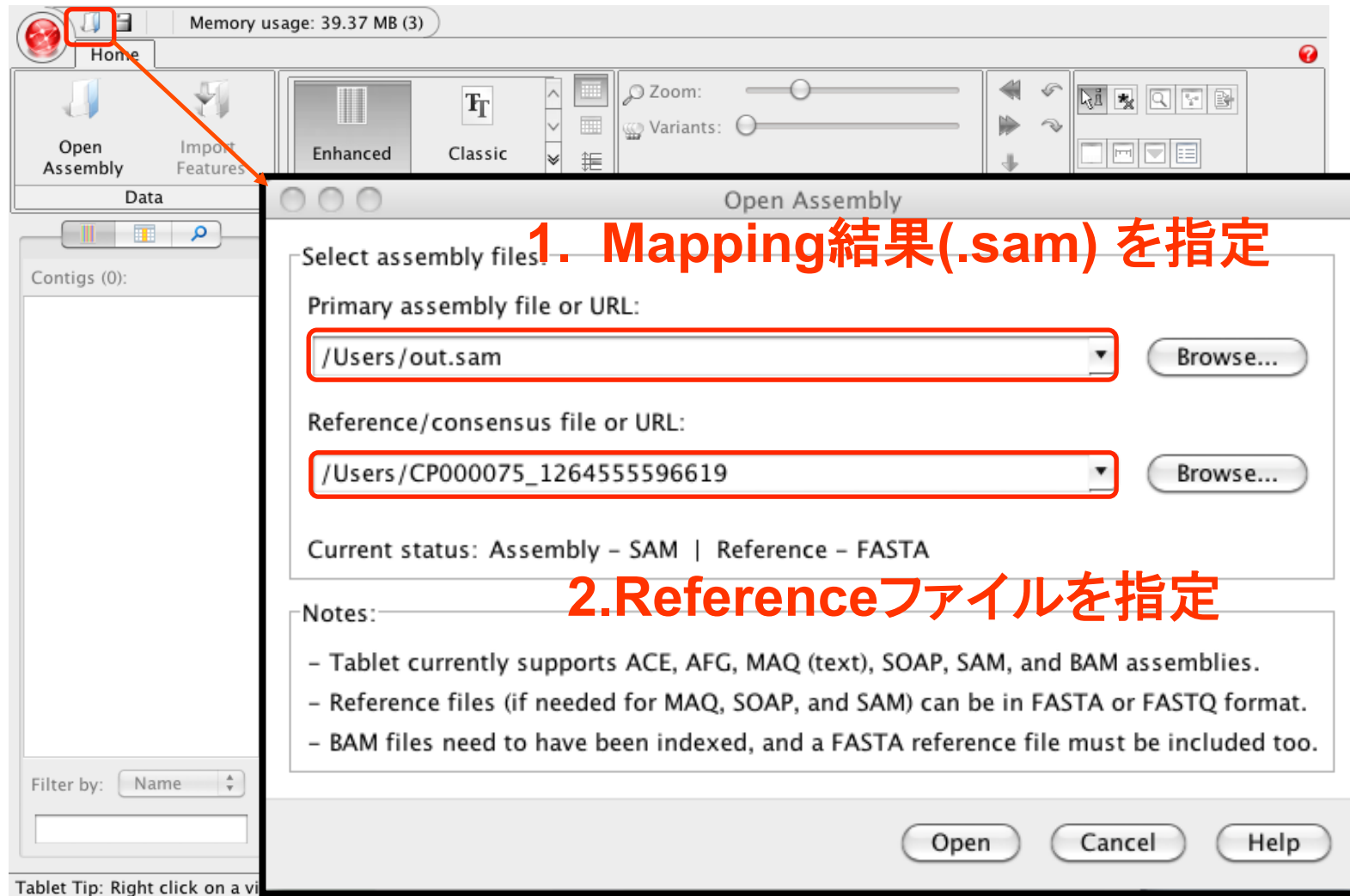
Chromosome

[CP000075_1264555596619](#)

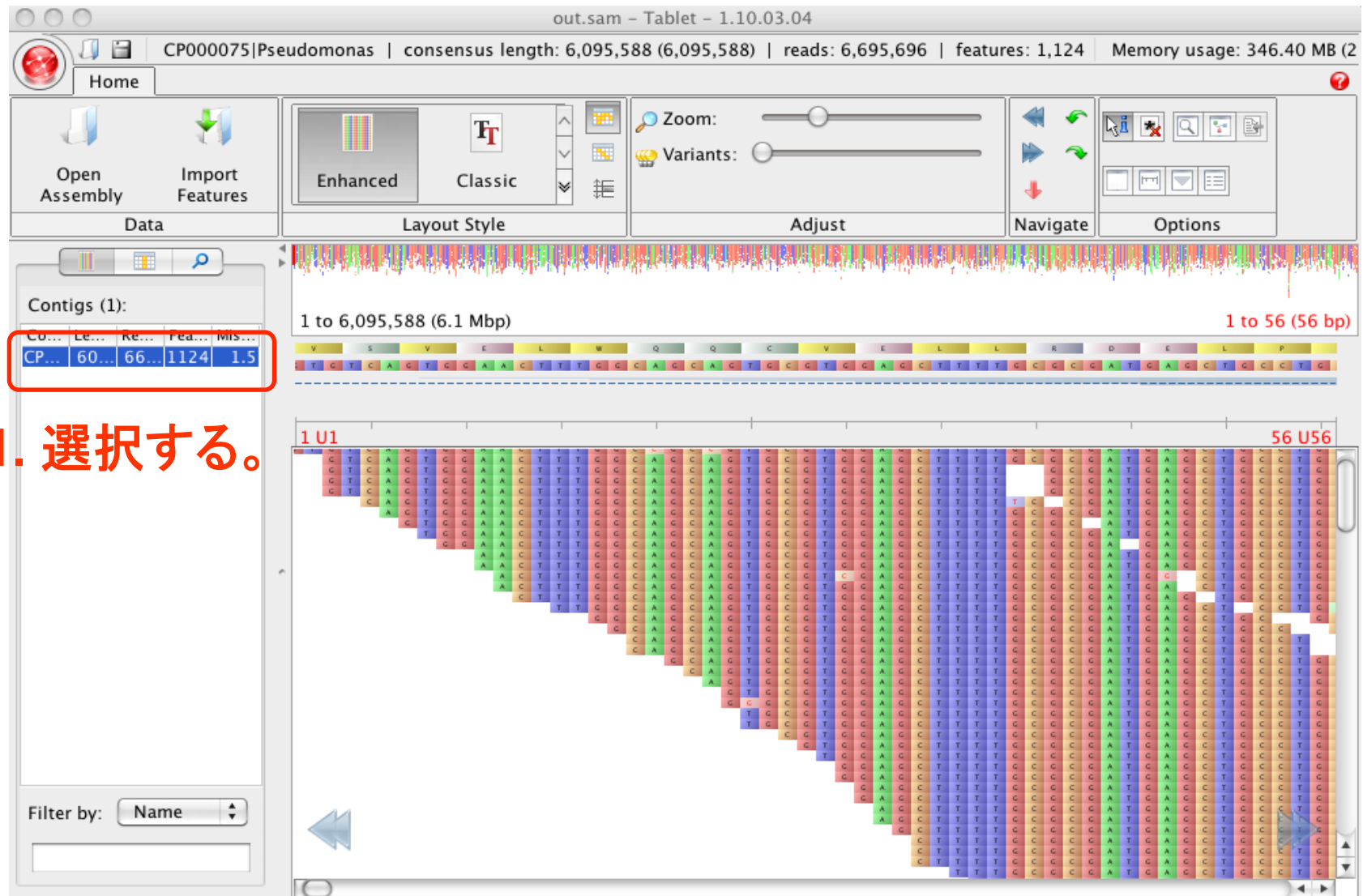
1. Referenceファイルと、Mapping結果ファイル(.sam)をダウンロードする。

gffConvert mapview.txt mapview.gff	2010-03-15 18:20:29	2010-03-15 18:21:10	View	download
maq indelsoa CP000075_1264555596619.bfa out_all.map > out.indel.so	2010-03-15 18:21:30	2010-03-15 18:21:51		download
maq assemble [-t 0.5 -r 0 -m 4] out.cns CP000075_1264555596619.bfa out_all.map	2010-03-15 18:22:01	2010-03-15 18:22:42	View	download
maq cns2snp out.cns > out.snp	2010-03-15 18:22:52	2010-03-15 18:23:02		download
maq.pl SNPfilter [-D 124 -w 4] out.snp > out.filter.snp	2010-03-15 18:23:12	2010-03-15 18:23:22		download
maq2sam out_all.map > out.sam	2010-03-15 18:23:33	2010-03-15 18:24:24		download
samtools view -bS -o headeredout.bam headeredout.sam	2010-03-15 18:27:47	2010-03-15 18:28:58	View	download
samtools view -X headeredout.bam > out.samX	2010-03-15 18:29:19	2010-03-15 18:30:21		download

結果ファイルとreference配列を指定



Mapping結果をviewerで表示



1. 選択する。

Acknowledgements

- Jun Mashima
- Toshihisa Okido
- Asami Nozaki

- Hitoshi Kunii
- Daisuke Ikumi
- Takeshi Konno
- Nobuhiro Hoshi
- Shouta Morizaki

- Yoshio Tateno
- DDBJ Annotators, members

DRA is part of the National project of integrating life science databases, and is supported by the Japan Science and Technology Agency Institute for Bioinformatics Research and Development.