DDBJ
DNA Data Bank of Japan

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

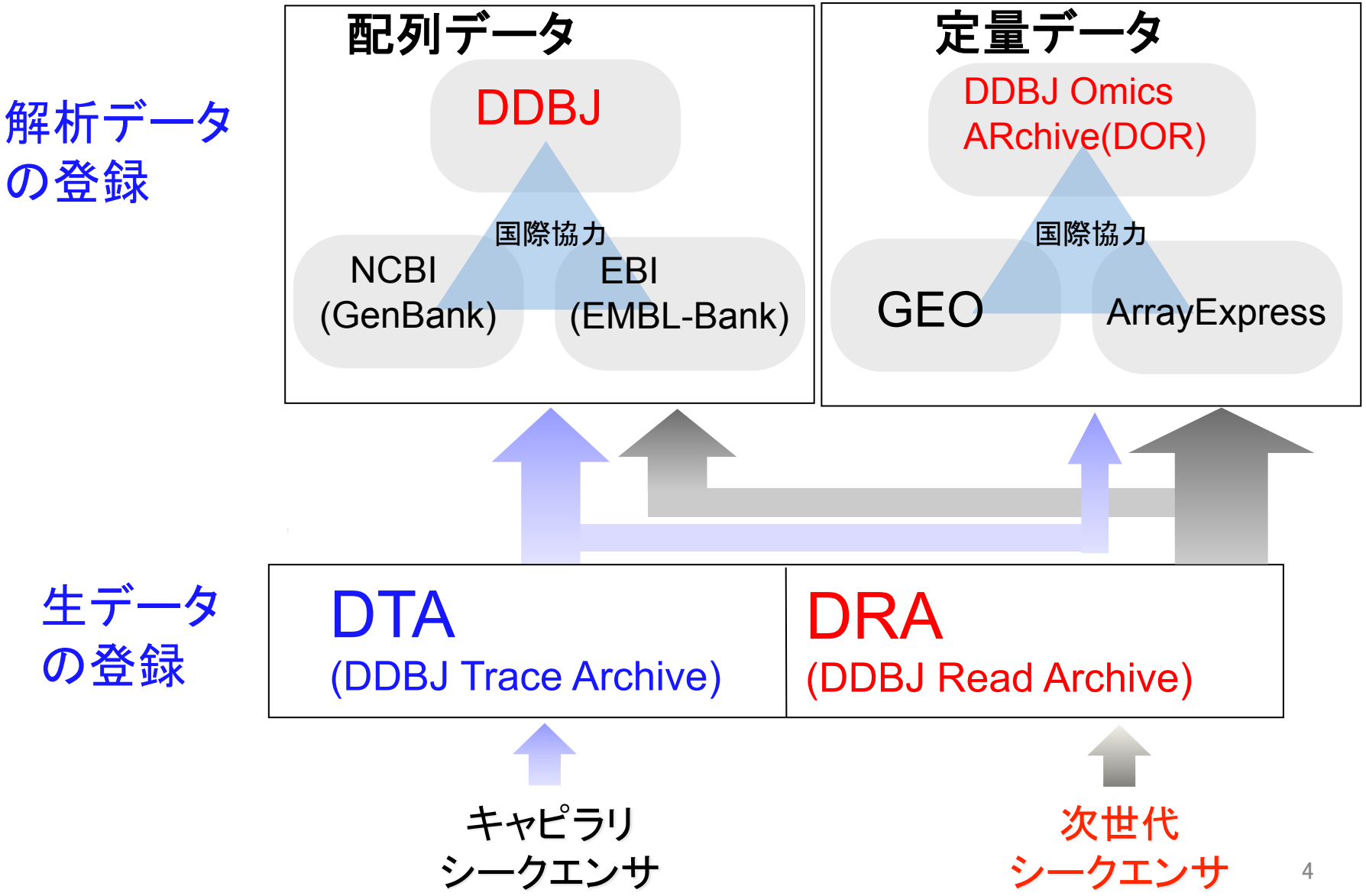
AJACS湘南3
2010年6月4日（金）14:20-15:20
日本大学生物資源科学部 2号館231教室

DDBJ
DNA Data Bank of Japan

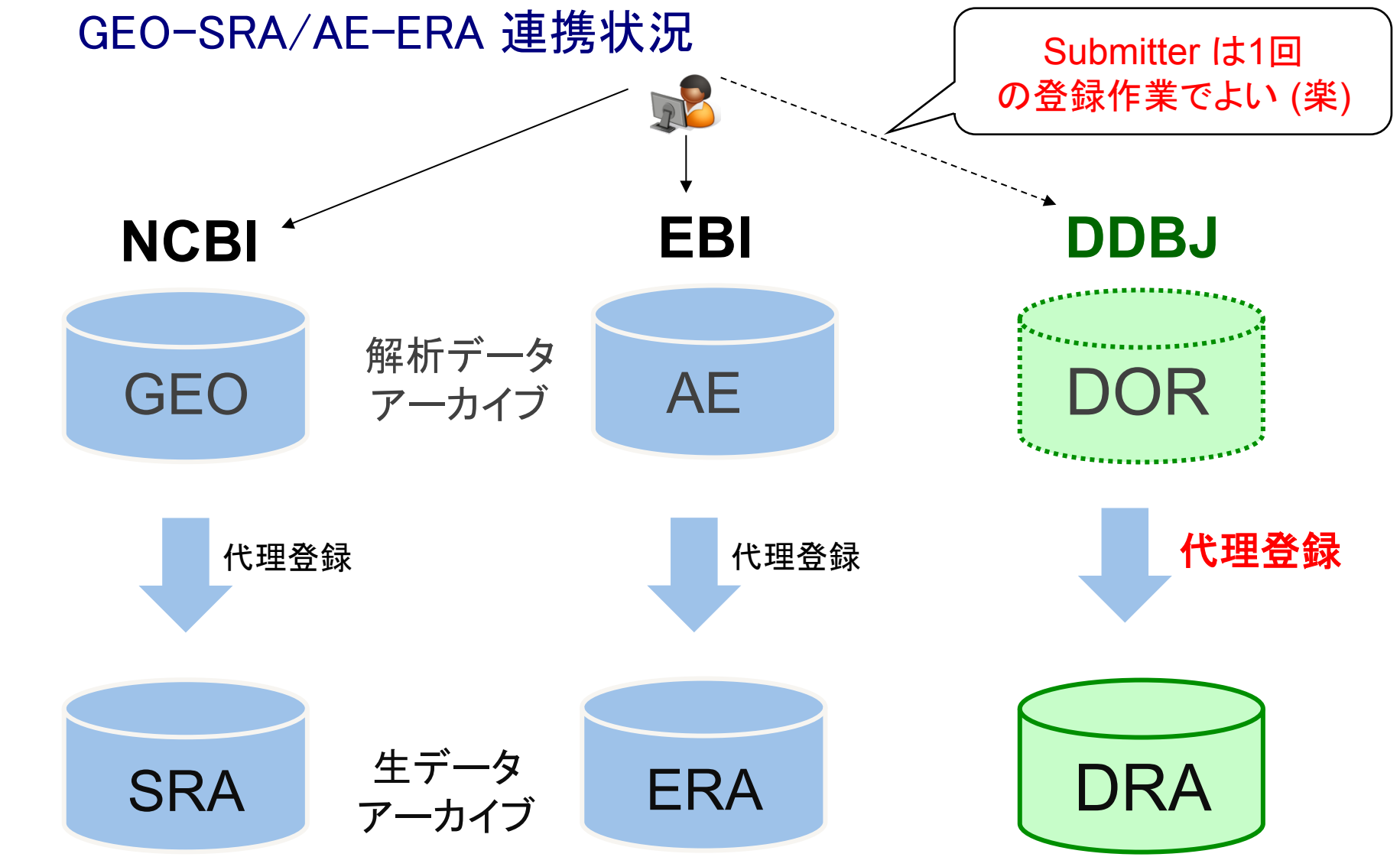
目次

③次世代シークエンサ・データの登録・解析

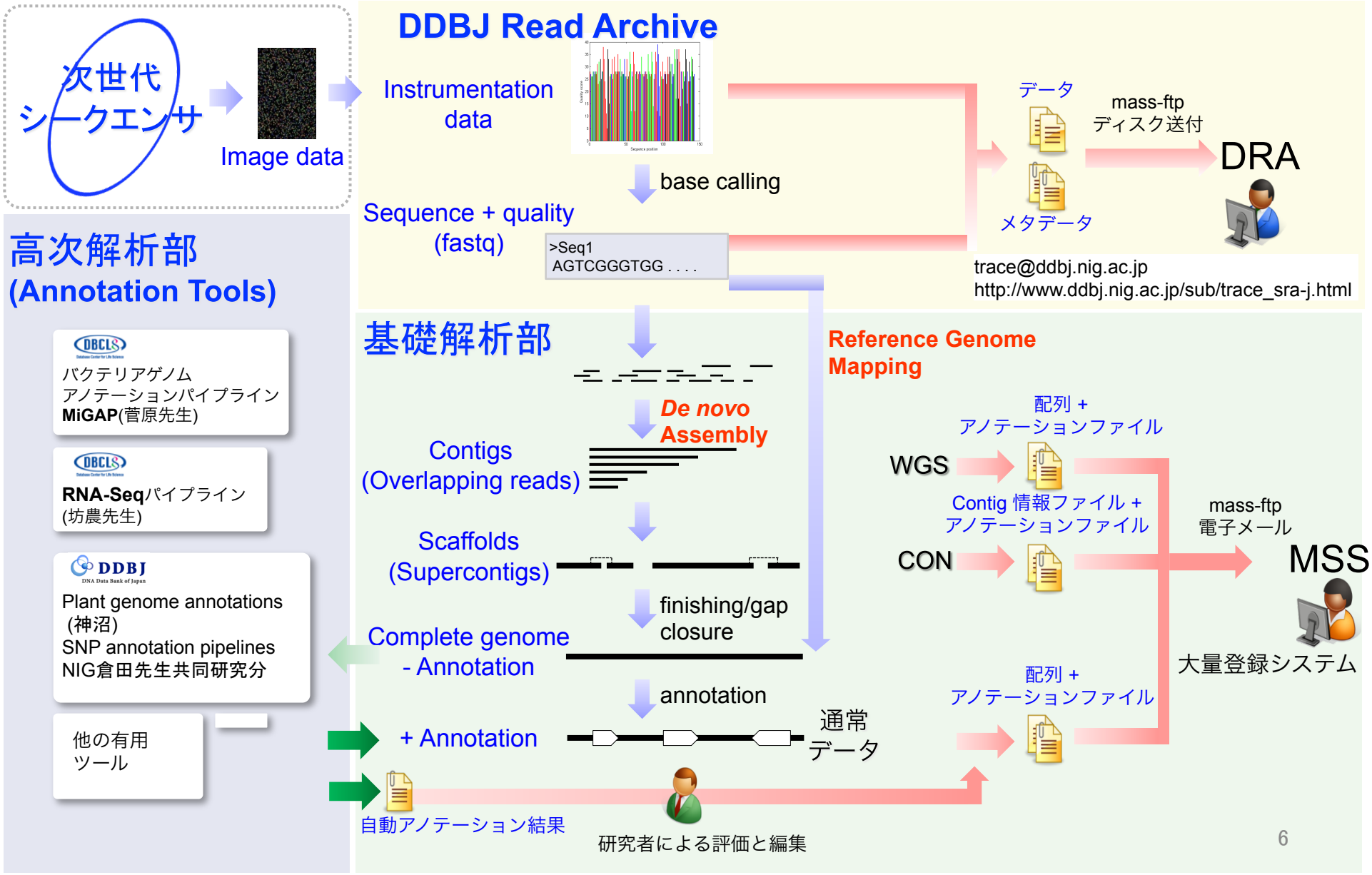
DDBJにおける塩基配列データの登録
→論文用アクセッション番号発行



DDBJ Omics Archiveから DRAへの代理登録



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
hhUhhhhhYhhhhhhhhehcaa`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

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

↑

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。

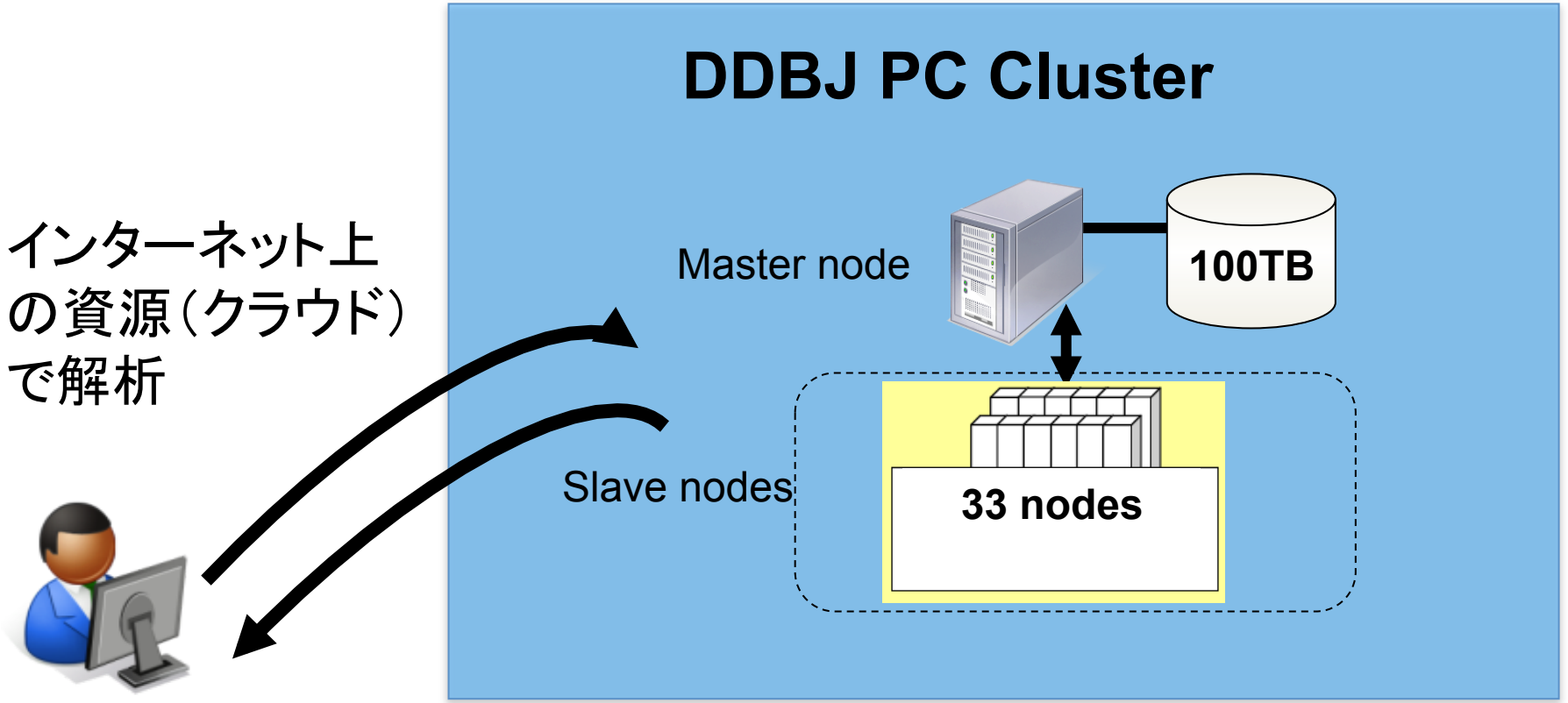
DDBJ
DNA Data Bank of Japan

パイプラインの特徴①

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

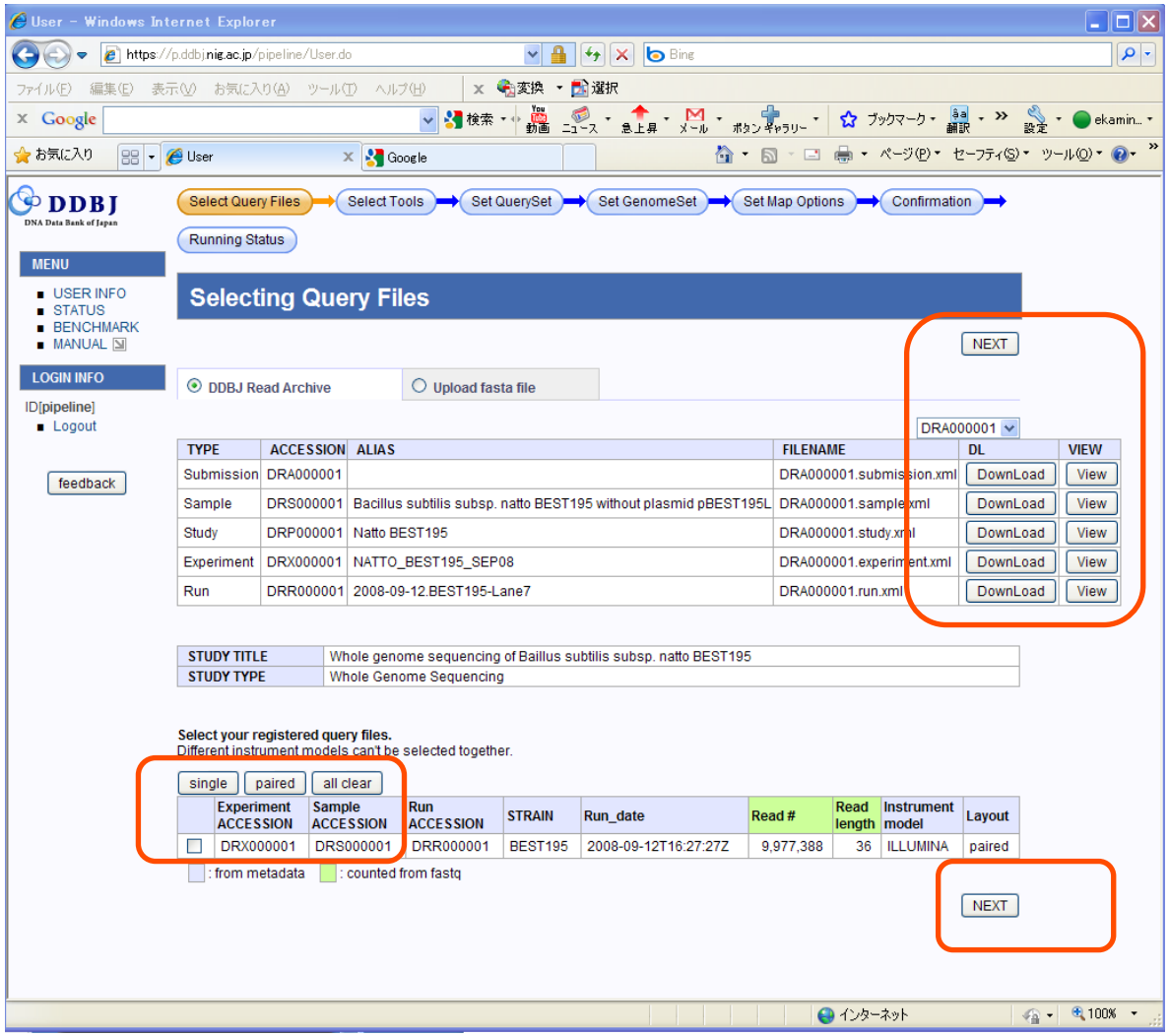


	CPU	Memory	Time	Maximum Job#
Computer nodes	2*3.16GHz	4GB	168H	42

DDBJ
DNA Data Bank of Japan

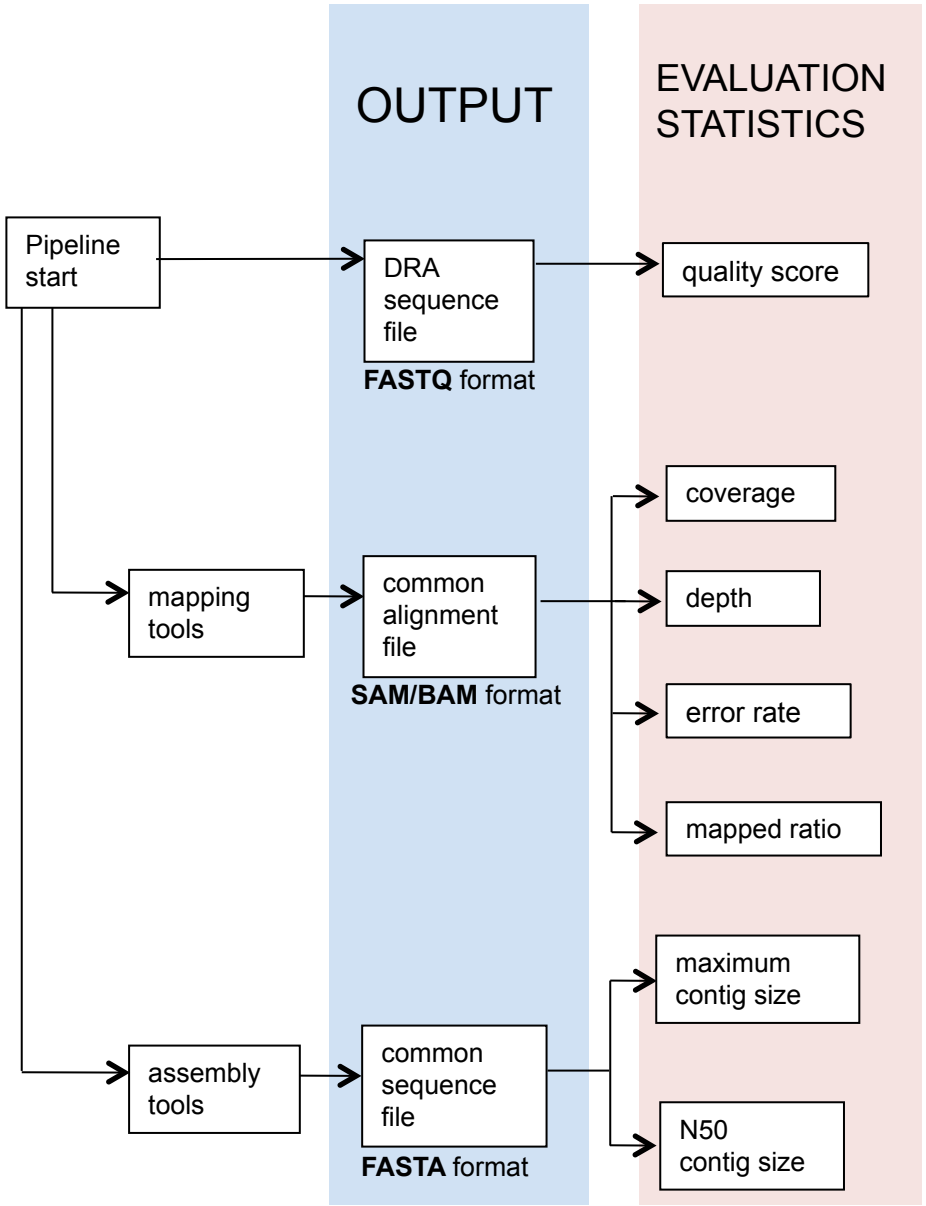
パイプラインの特徴②
実験研究者が自分1人で解析可能

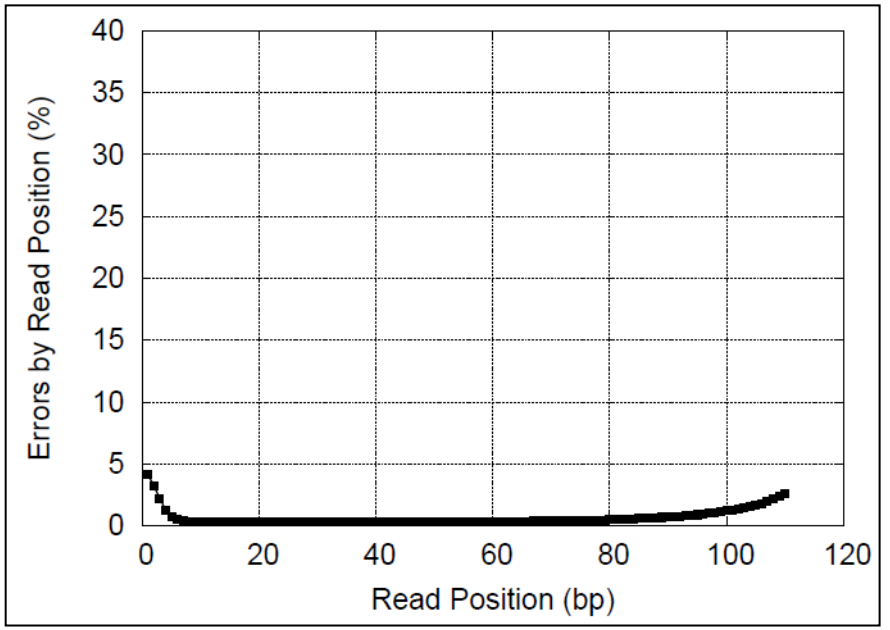
ボタン操作、テキスト入力のみで解析結果取得



プログラミング
しなくても解析可

パイプラインの特徴③
基本統計量(Quality Score/Coverage等)を自動計算



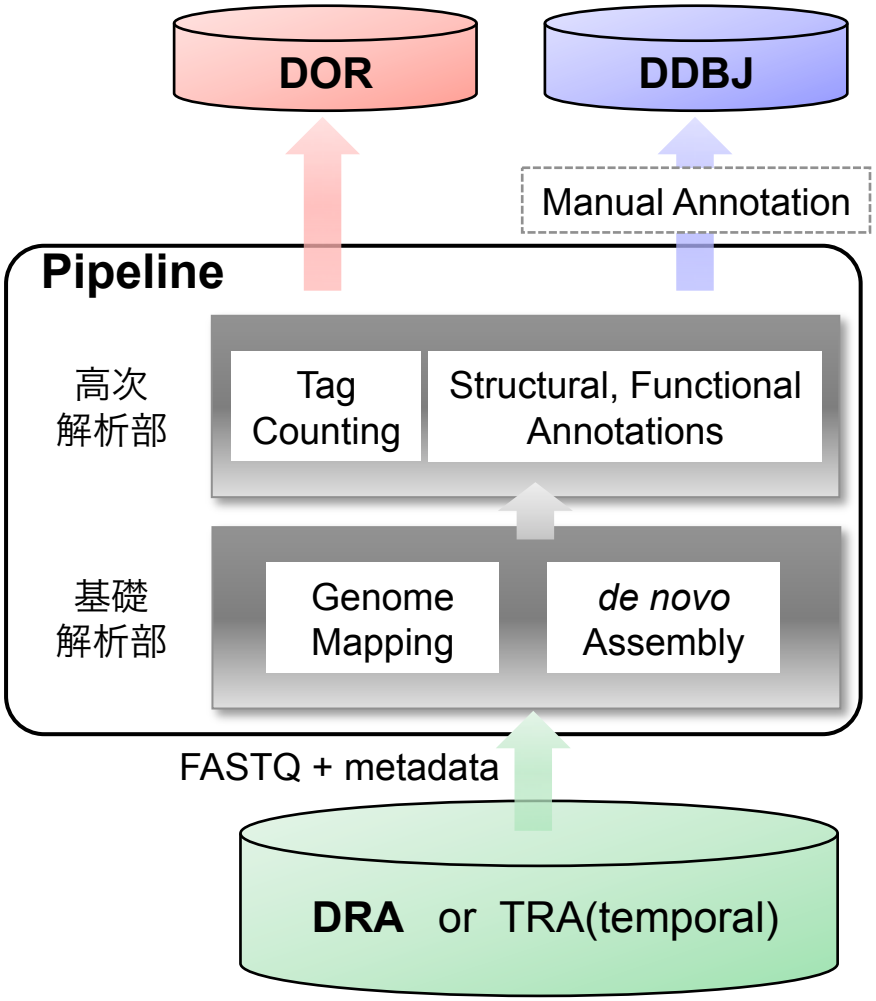


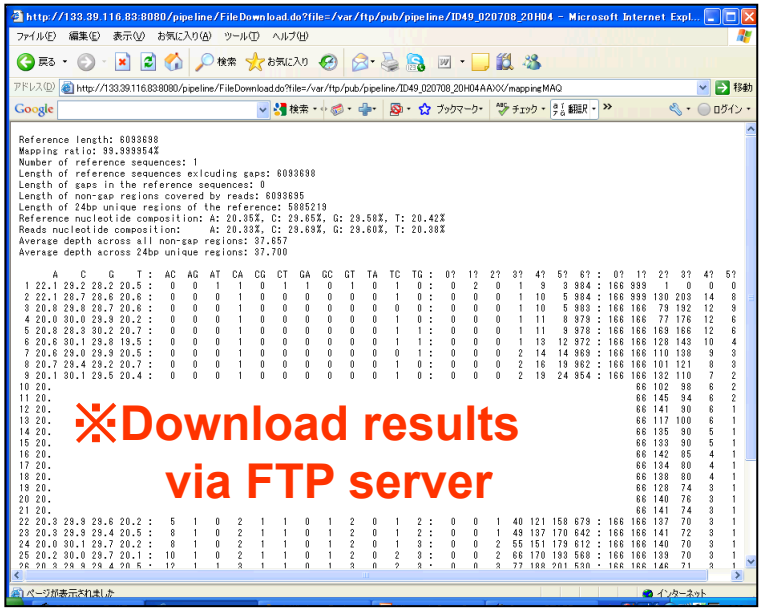
論文用の基本統計量、図を生成

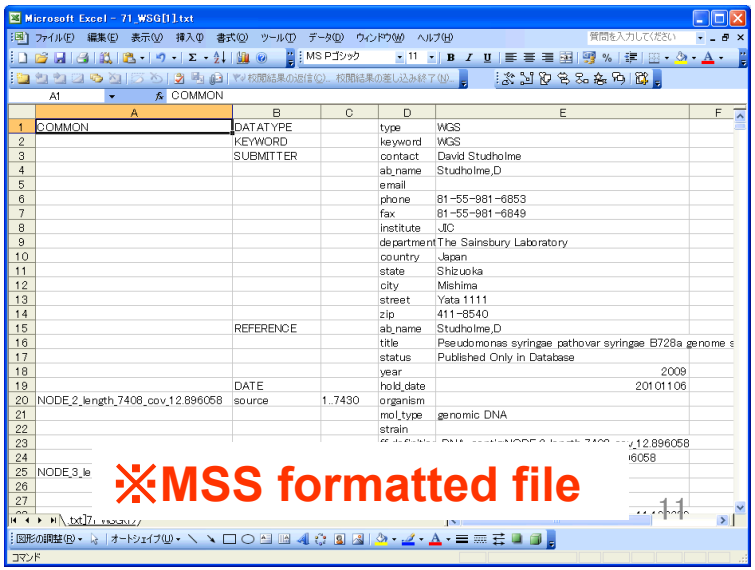
DDBJ
DNA Data Bank of Japan

パイプラインの特徴④

＜配列登録支援＞DDBJ登録ファイル自動生成

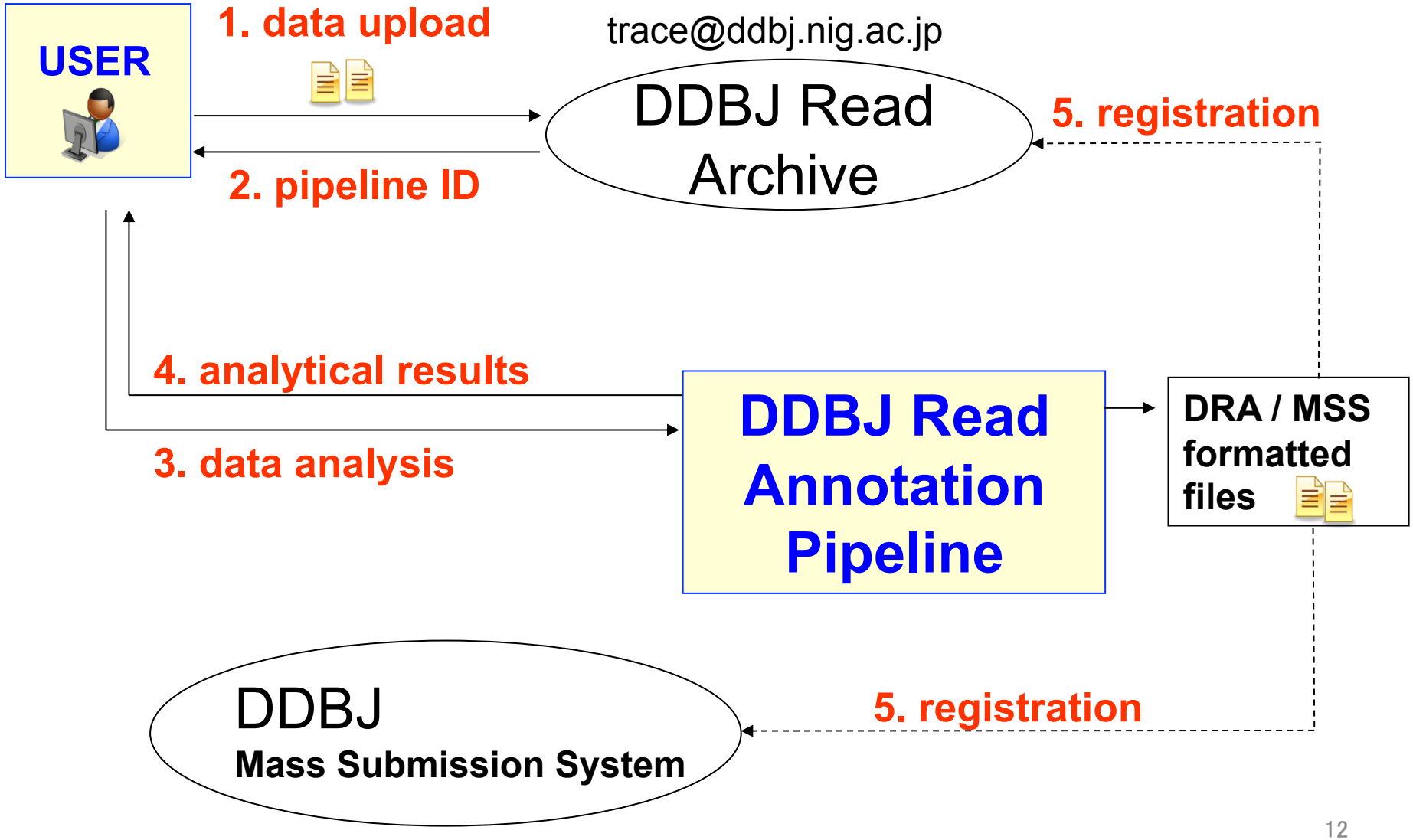






解析配列の登録手間を短縮

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

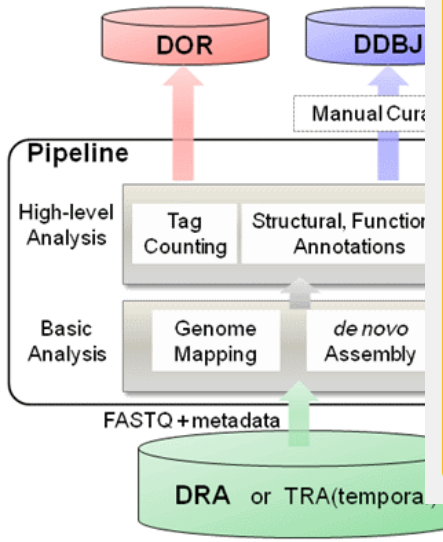


④DDBJクラウド型解析パイプラインのデモ

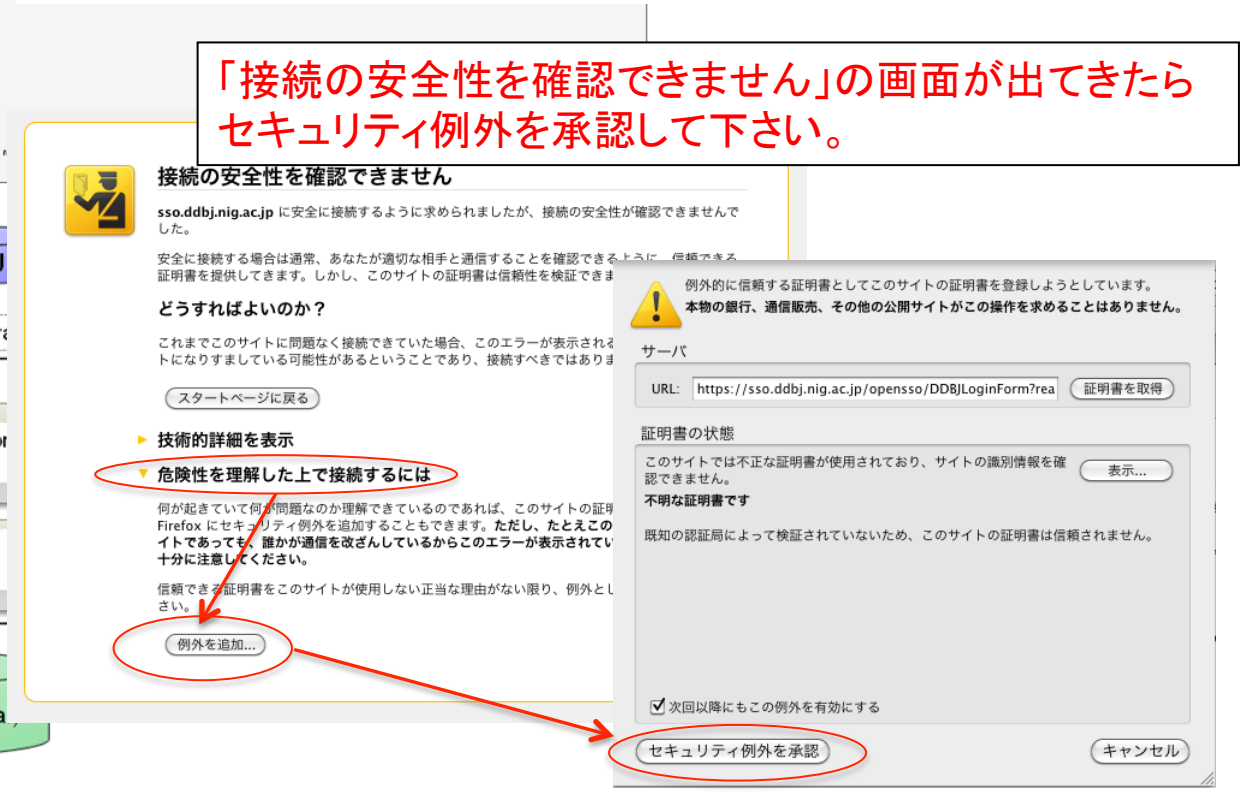
解析パイプライン
DDBJ Read Annotation Pipeline

解析パイプライン　https://p.ddbj.nig.ac.jp/

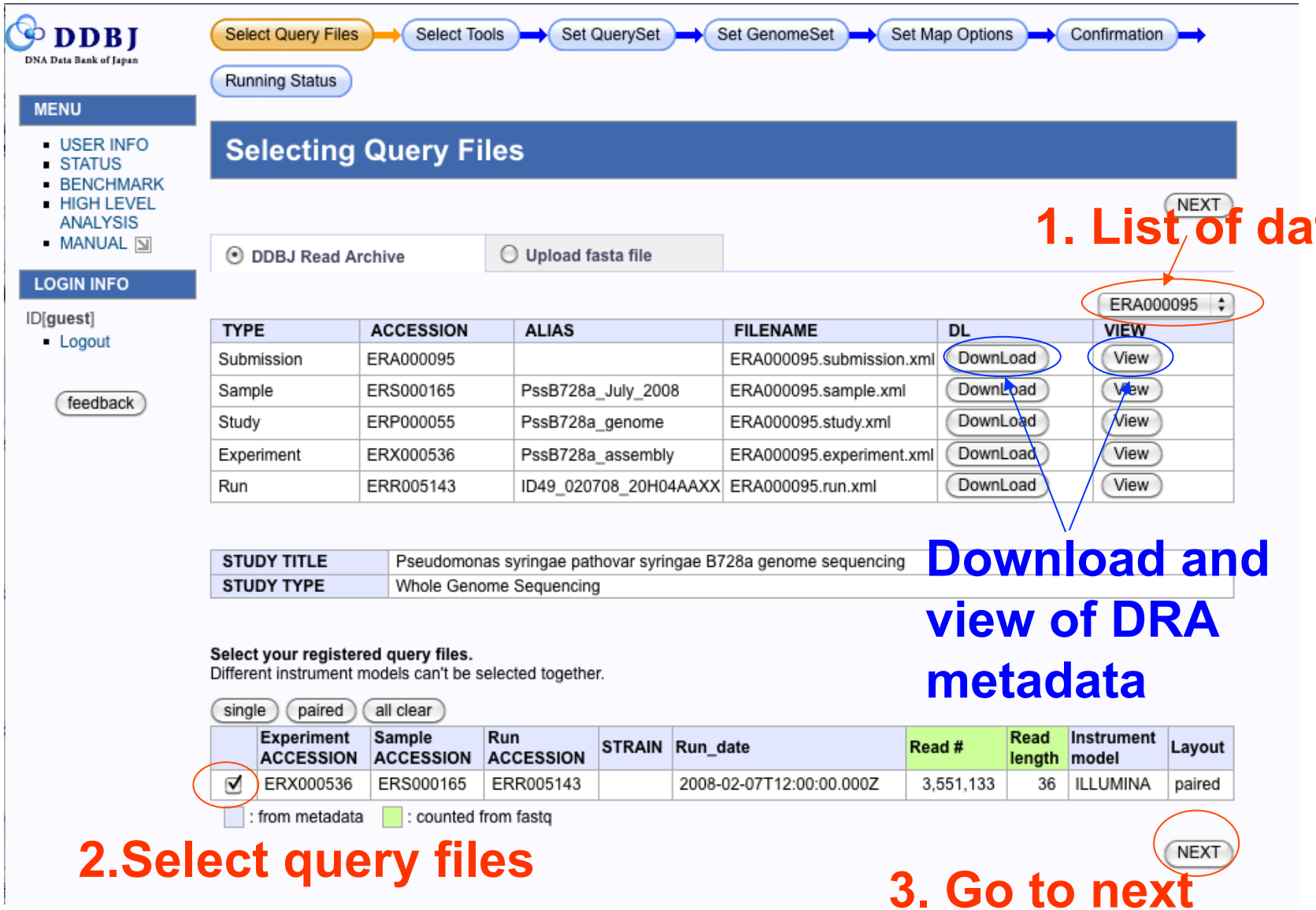
User ID: guest (パスワードなし)
でデモ画面を参照できます。

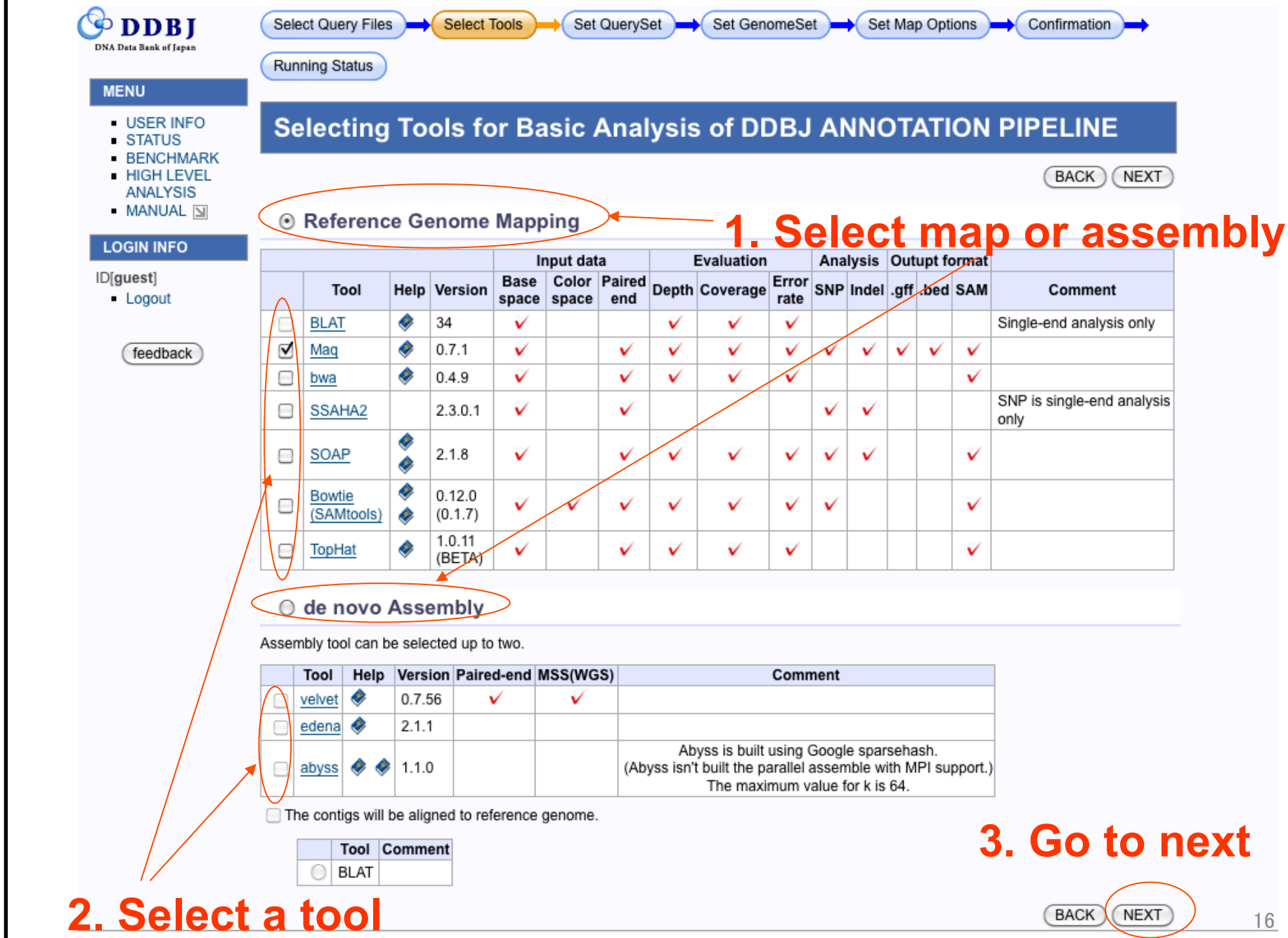


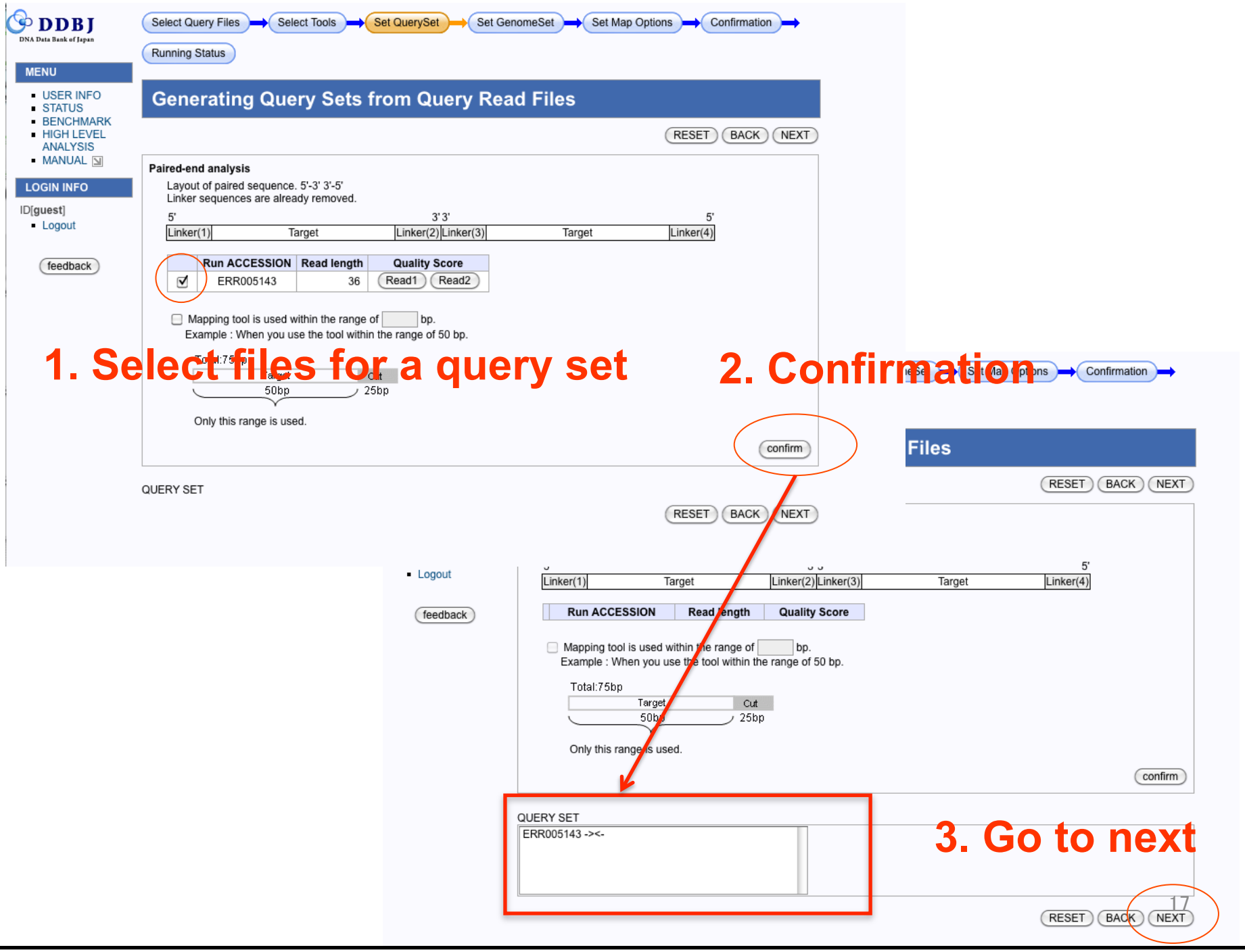
「接続の安全性を確認できません」の画面が出てきたら
セキュリティ例外を承認して下さい。

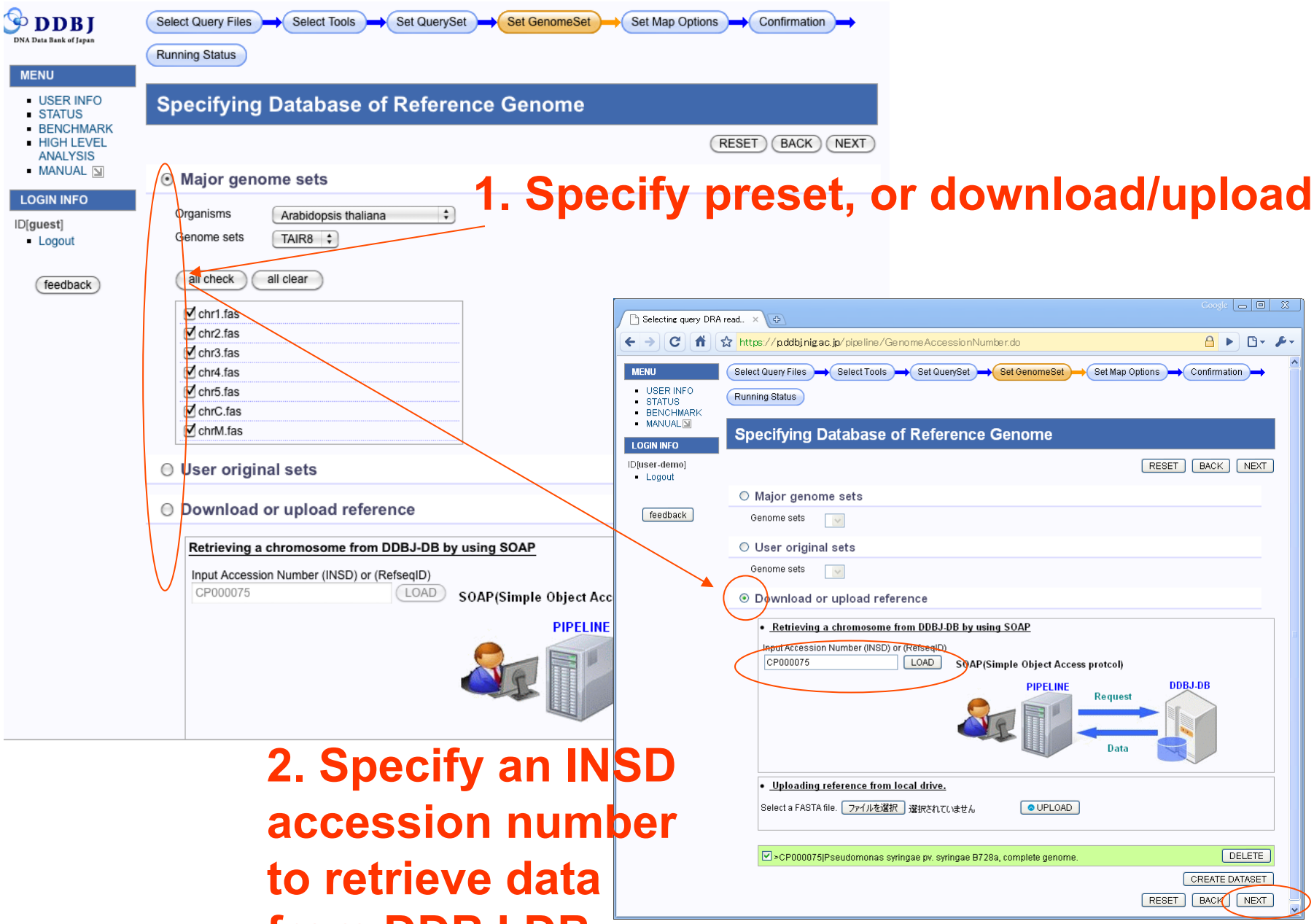


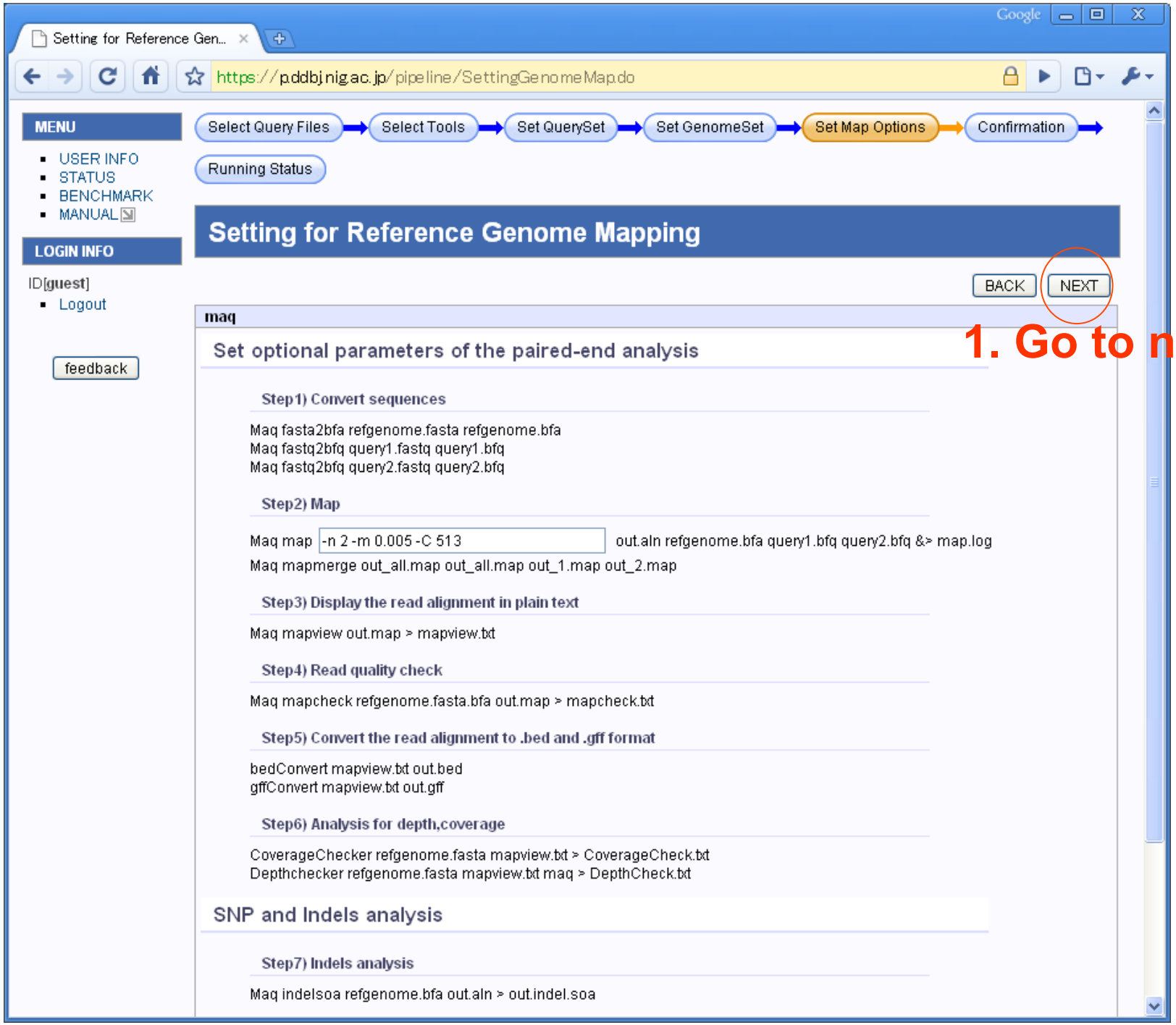
クエリ用FASTQ形式READファイルの選択

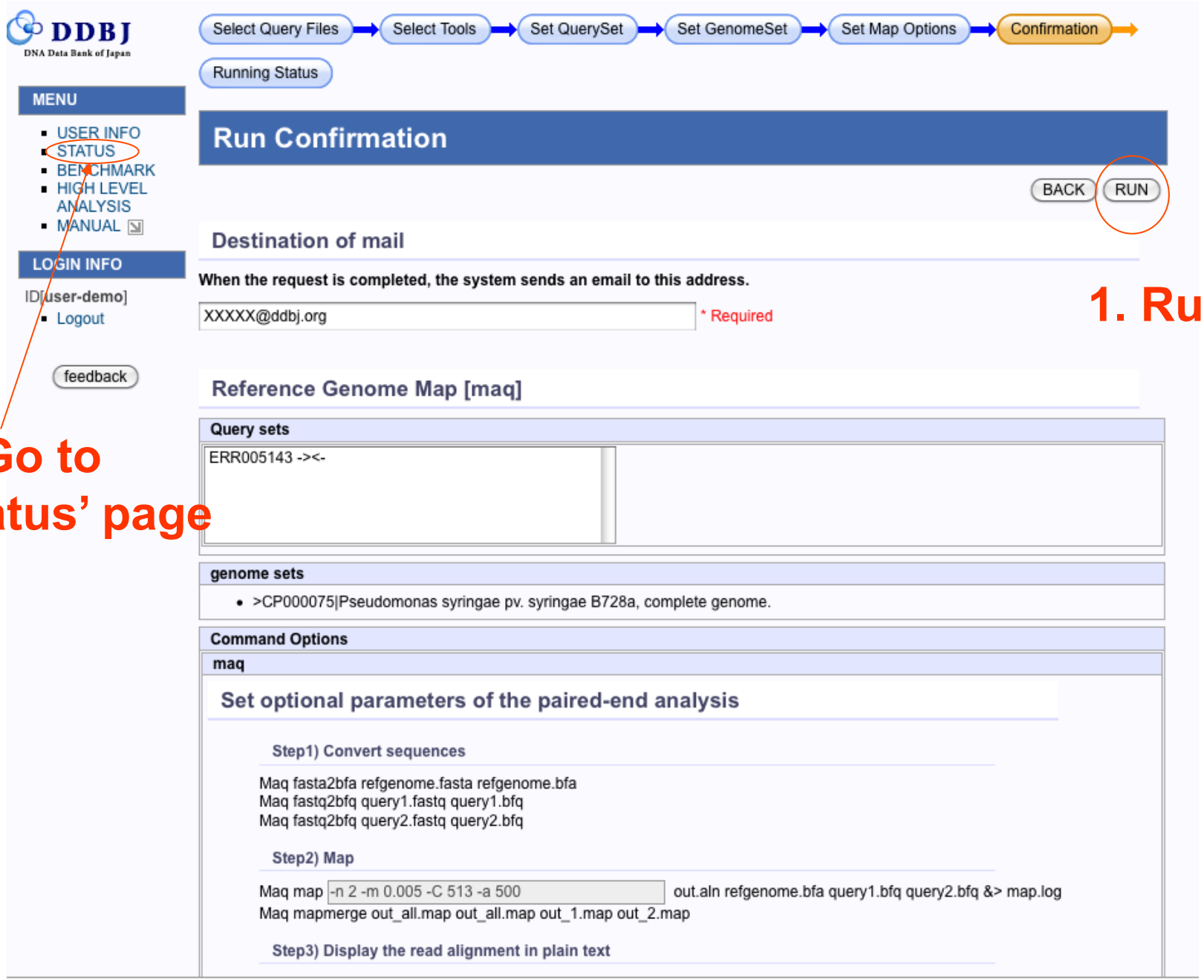



DDBJ
DNA Data Bank of Japan
Select Query Files
Select Tools
Set QuerySet
Set GenomeSet
Set Map Options
Confirmation
Running Status
MENU
USER INFO
STATUS
BENCHMARK
HIGH LEVEL ANALYSIS
MANUAL
LOGIN INFO
ID[guest]
Logout
feedback
Selecting Tools for Basic Analysis of DDBJ ANNOTATION PIPELINE
BACK
NEXT
Reference Genome Mapping
1. Select map or assembly
Input data
Evaluation
Analysis
Outupt format
Tool
Help
Version
Base space
Color space
Paired end
Depth
Coverage
Error rate
SNP
Indel
.gff
.bed
SAM
Comment
BLAT
34
Single-end analysis only
Maq
0.7.1
bwa
0.4.9
SSAHA2
2.3.0.1
SNP is single-end analysis only
SOAP
2.1.8
Bowtie (SAMtools)
0.12.0 (0.1.7)
TopHat
1.0.11 (BETA)
de novo Assembly
Assembly tool can be selected up to two.
Tool
Help
Version
Paired-end
MSS(WGS)
Comment
velvet
0.7.56
edena
2.1.1
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
BLAT
2. Select a tool
3. Go to next
BACK
NEXT
16


DDBJ
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
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.
5'
3' 3'
5'
Linker(1)
Target
Linker(2)
Linker(3)
Target
Linker(4)
Run ACCESSION
Read length
Quality Score
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.
Total:75bp
Target
Cut
50bp
25bp
Only this range is used.
confirm
QUERY SET
1. Select files for a query set
2. Confirmation
Files
3. Go to next
ERR005143 -><-
17

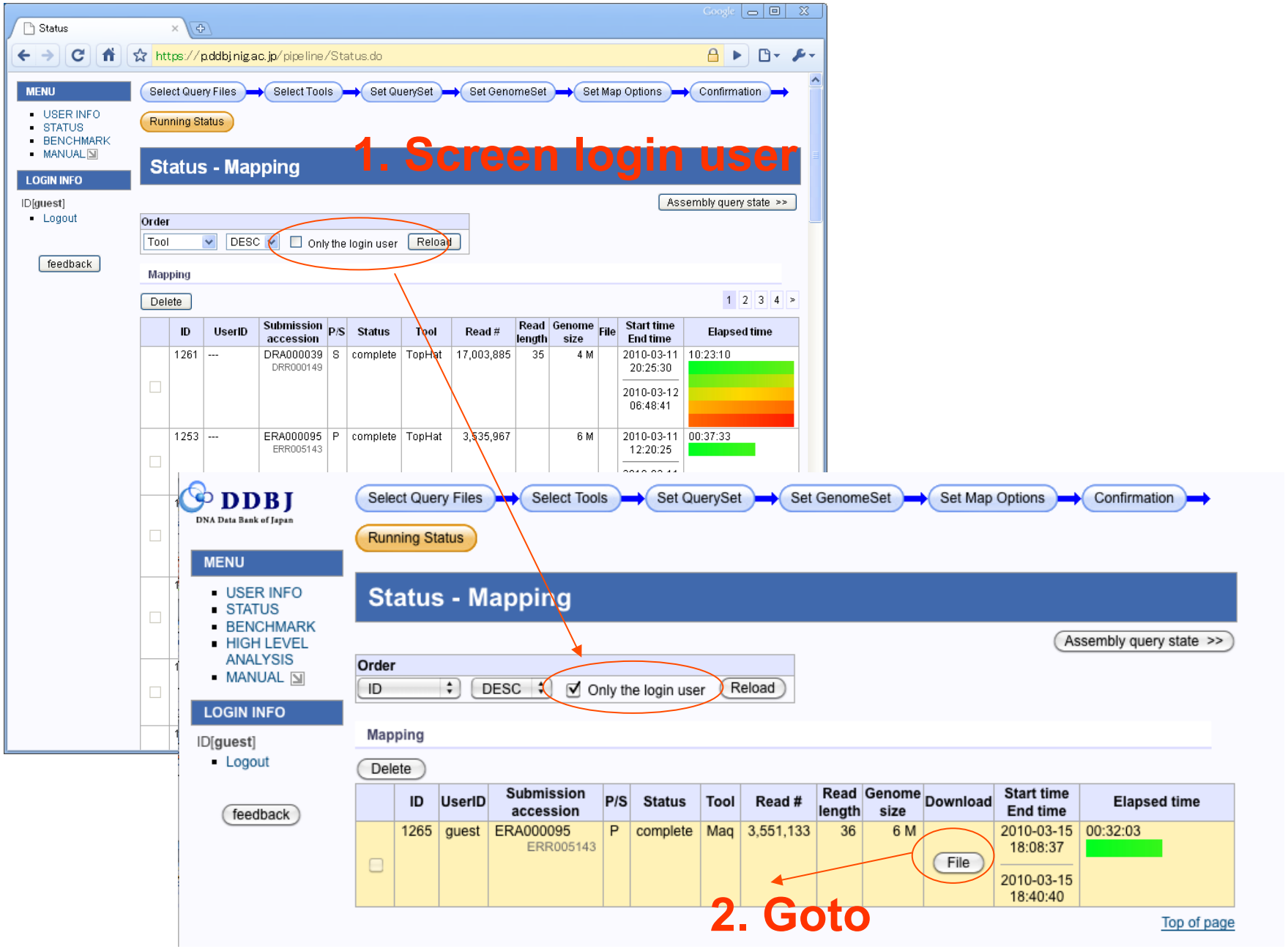

DDBJ
DNA Data Bank of Japan
Select Query Files
Select Tools
Set QuerySet
Set GenomeSet
Set Map Options
Confirmation
Running Status
MENU
USER INFO
STATUS
BENCHMARK
HIGH LEVEL ANALYSIS
MANUAL
LOGIN INFO
ID[guest]
Logout
feedback
Specifying Database of Reference Genome
RESET
BACK
NEXT
Major genome sets
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 Acc
PIPELINE
1. Specify preset, or download/upload
2. Specify an INSD accession number to retrieve data from DDBJ DB
3. Go to next
Selecting query DRA read...
https://ddbj.nig.ac.jp/pipeline/GenomeAccessionNumber.do
ID[user-demo]
SOAP(Simple Object Access protcol)
DDBJ-DB
Request
Data
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


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.txt
Step4) Read quality check
Maq mapcheck refgenome.fasta.bfa out.map > mapcheck.txt
Step5) Convert the read alignment to .bed and .gff format
bedConvert mapview.txt out.bed
gffConvert mapview.txt out.gff
Step6) Analysis for depth,coverage
CoverageChecker refgenome.fasta mapview.txt > CoverageCheck.txt
Depthchecker refgenome.fasta mapview.txt maq > DepthCheck.txt
SNP and Indels analysis
Step7) Indels analysis
Maq indelsoa refgenome.bfa out.aln > out.indel.soa
1. Go to next



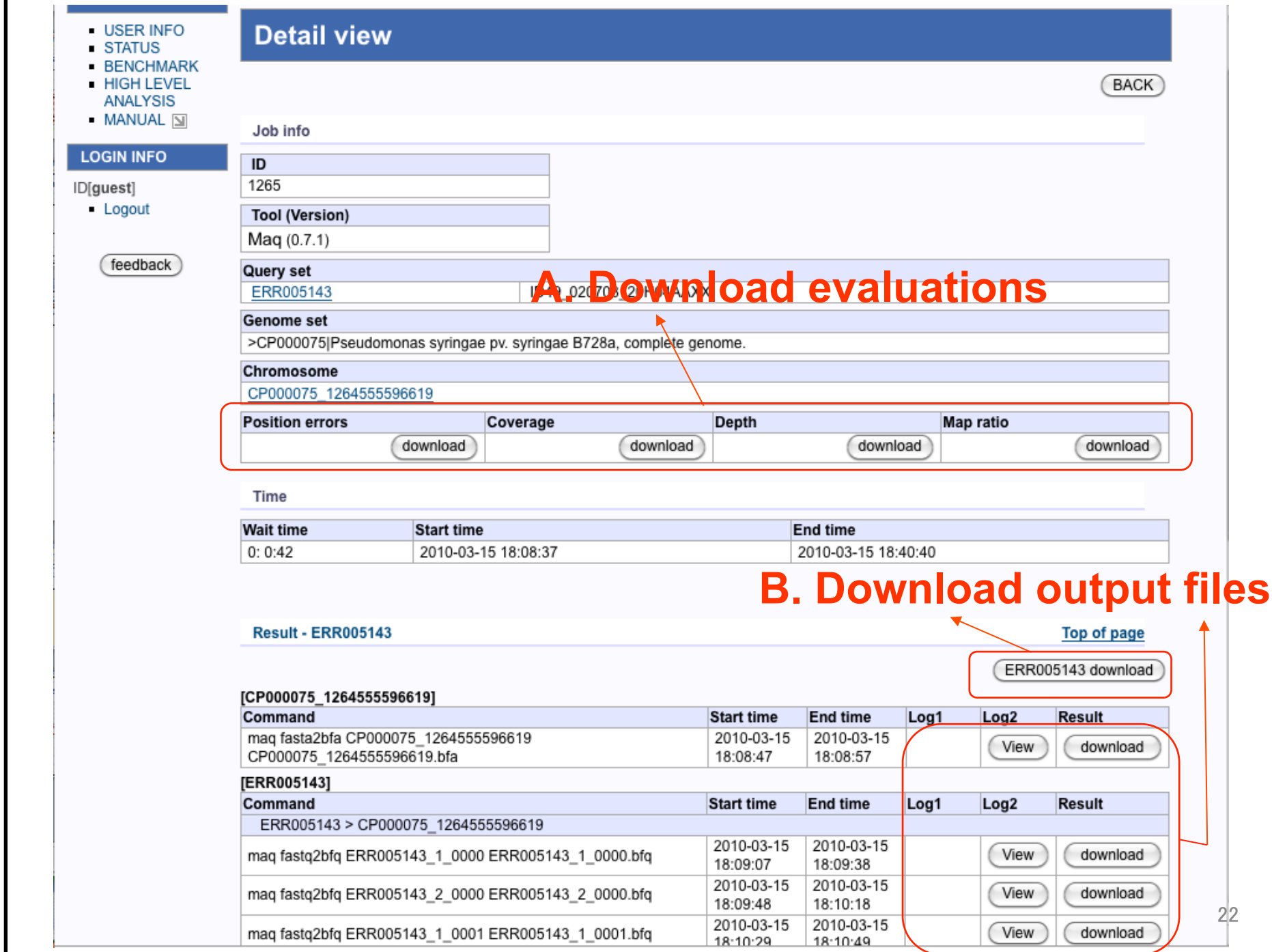
1. Run

2. Go to ‘status’ page


Status
https://ddbj.nig.ac.jp/pipeline/Status.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
Status - Mapping
1. Screen login user
Assembly query state >>
Order
Tool
DESC
Only the login user
Reload
Mapping
Delete
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
DDBJ
DNA Data Bank of Japan
MENU
USER INFO
STATUS
BENCHMARK
HIGH LEVEL ANALYSIS
MANUAL
LOGIN INFO
ID[guest]
Logout
feedback
Select Query Files
Select Tools
Set QuerySet
Set GenomeSet
Set Map Options
Confirmation
Running Status
Status - Mapping
Assembly query state >>
Order
ID
DESC
Only the login user
Reload
Mapping
Delete
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
Top of page
2. Goto
‘Download’ page


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
Coverage
Depth
Map ratio
download
A. Download evaluations
Time
Wait time
Start time
End time
0: 0:42
2010-03-15 18:08:37
2010-03-15 18:40:40
B. Download output files
Result - ERR005143
Top of page
ERR005143 download
[CP000075_1264555596619]
Command
Start time
End time
Log1
Log2
Result
maq fasta2bfa CP000075_1264555596619 CP000075_1264555596619.bfa
2010-03-15 18:08:47
2010-03-15 18:08:57
View
download
[ERR005143]
ERR005143 > CP000075_1264555596619
maq fastq2bfq ERR005143_1_0000 ERR005143_1_0000.bfq
2010-03-15 18:09:07
2010-03-15 18:09:38
maq fastq2bfq ERR005143_2_0000 ERR005143_2_0000.bfq
2010-03-15 18:09:48
2010-03-15 18:10:18
maq fastq2bfq ERR005143_1_0001 ERR005143_1_0001.bfq
2010-03-15 18:10:29
2010-03-15 18:10:49

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

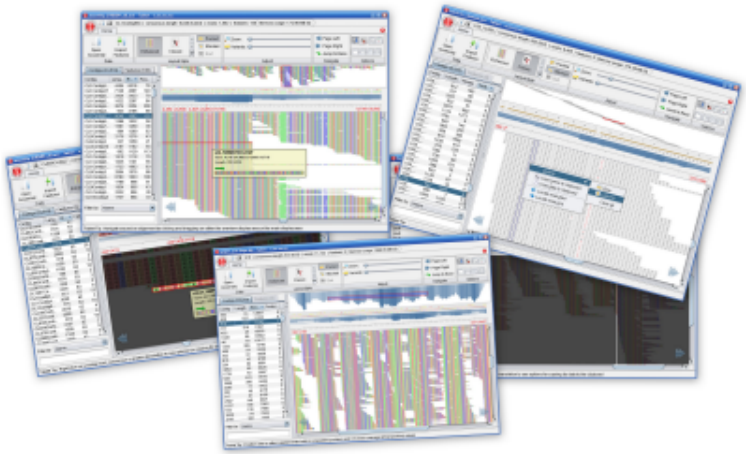
- Tablet Homepage
- Download Tablet
- Screenshots
- Tablet FAQ
- Sample Data
- Assembly Conversion
- Papers and Presentations
- Privacy Policy
- Tablet World Map
- Online Help

Our Software

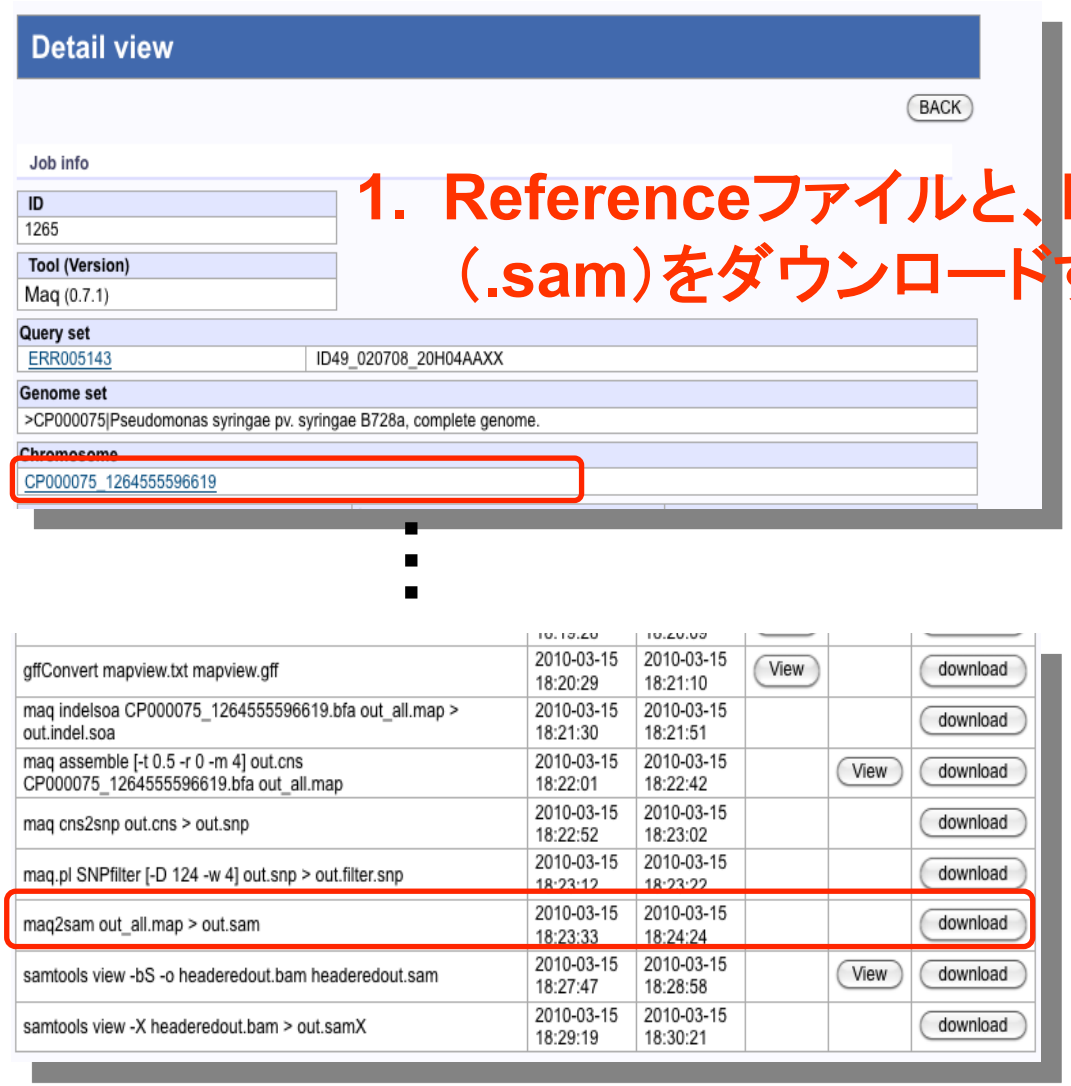
- CurlyWhirly
- Flapjack
- OPTIRas
- Strudel
- Tablet (new)
- TetraploidMap

Tablet - Next Generation Sequence Assembly Visualization

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

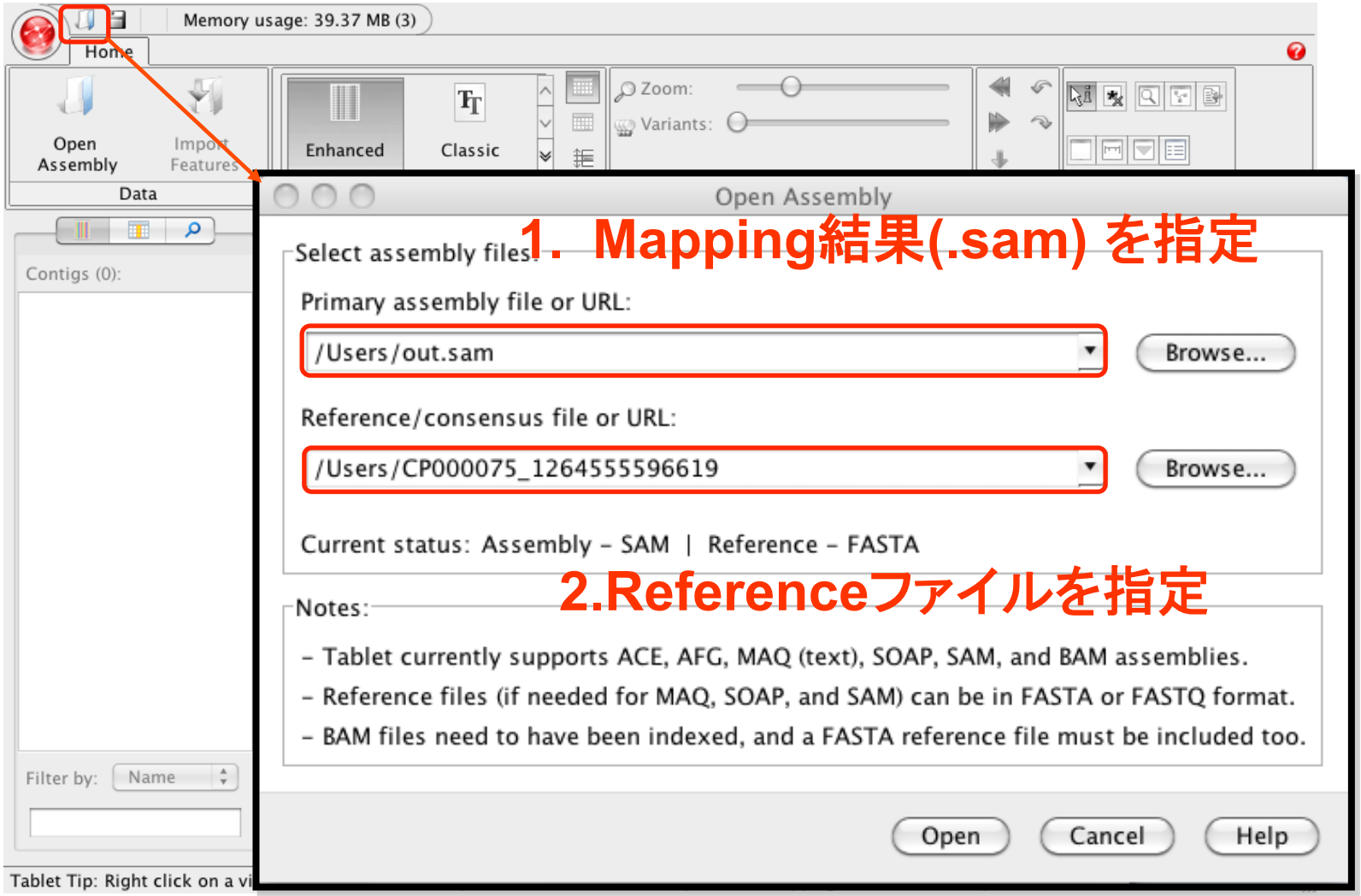


Mapping結果をダウンロード



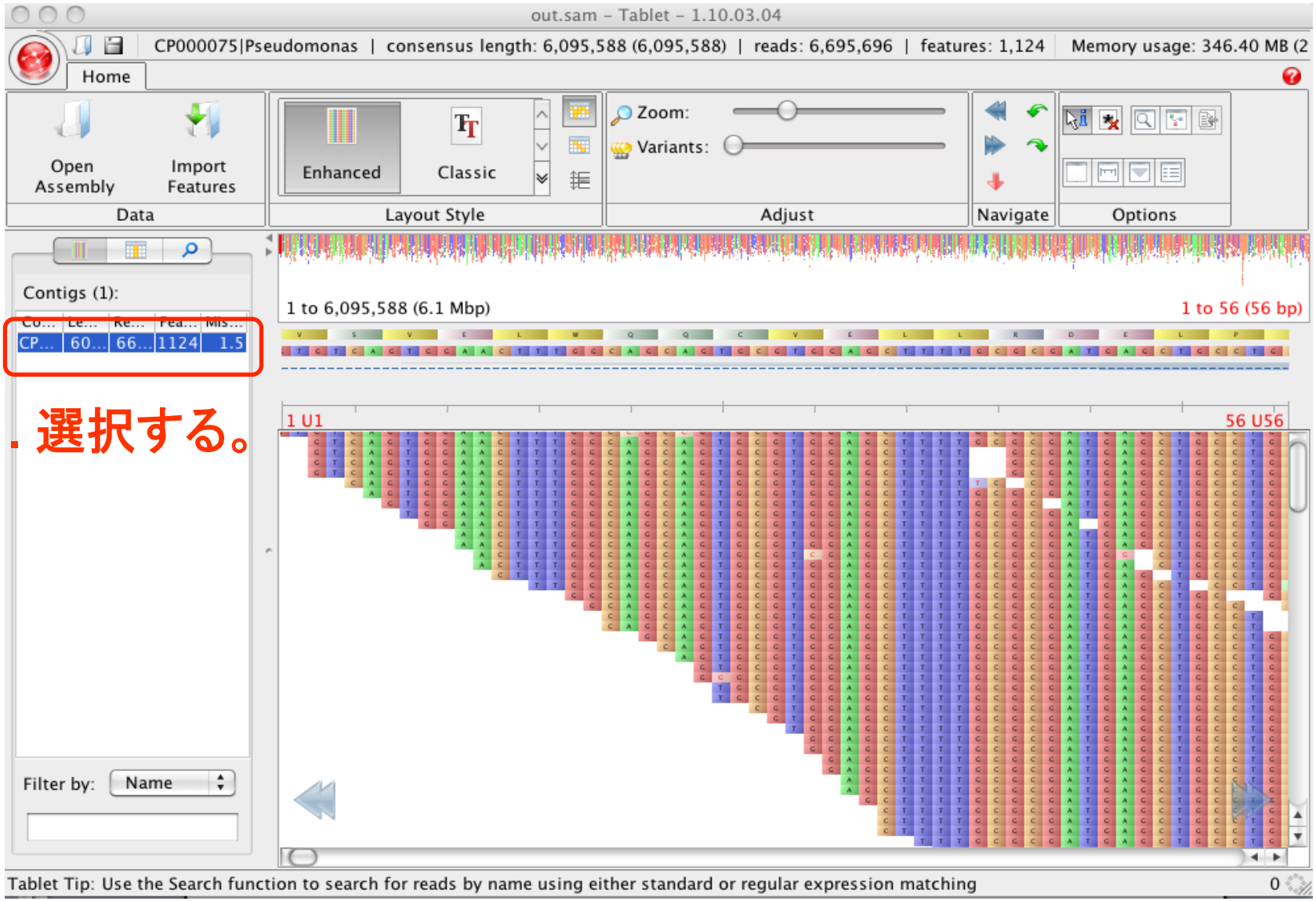
1. Referenceファイルと、Mapping結果ファイル（.sam）をダウンロードする。

結果ファイルとreference配列を指定



DDBJ
DNA Data Bank of Japan

Mapping結果をviewerで表示



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.