2017年度 「先進ゲノム支援」情報解析講習会

DDBJスパコンでの解析の実践I
（DDBJパイプライン）

国立遺伝学研究所 大量遺伝情報研究室
望月孝子

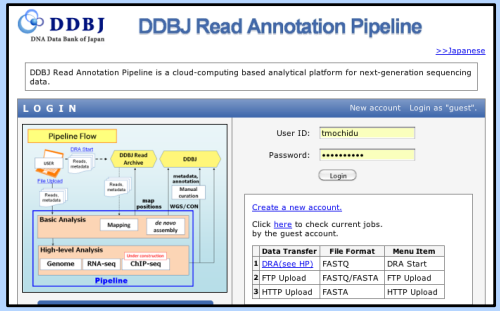
DDBJ Read Annotation Pipline 全体像

DDBJ Sequence Read Archive (DRA)

ユーザデータ

DDBJ Pipeline: DDBJ Read Annotation Pipeline

基礎処理部



マッピング

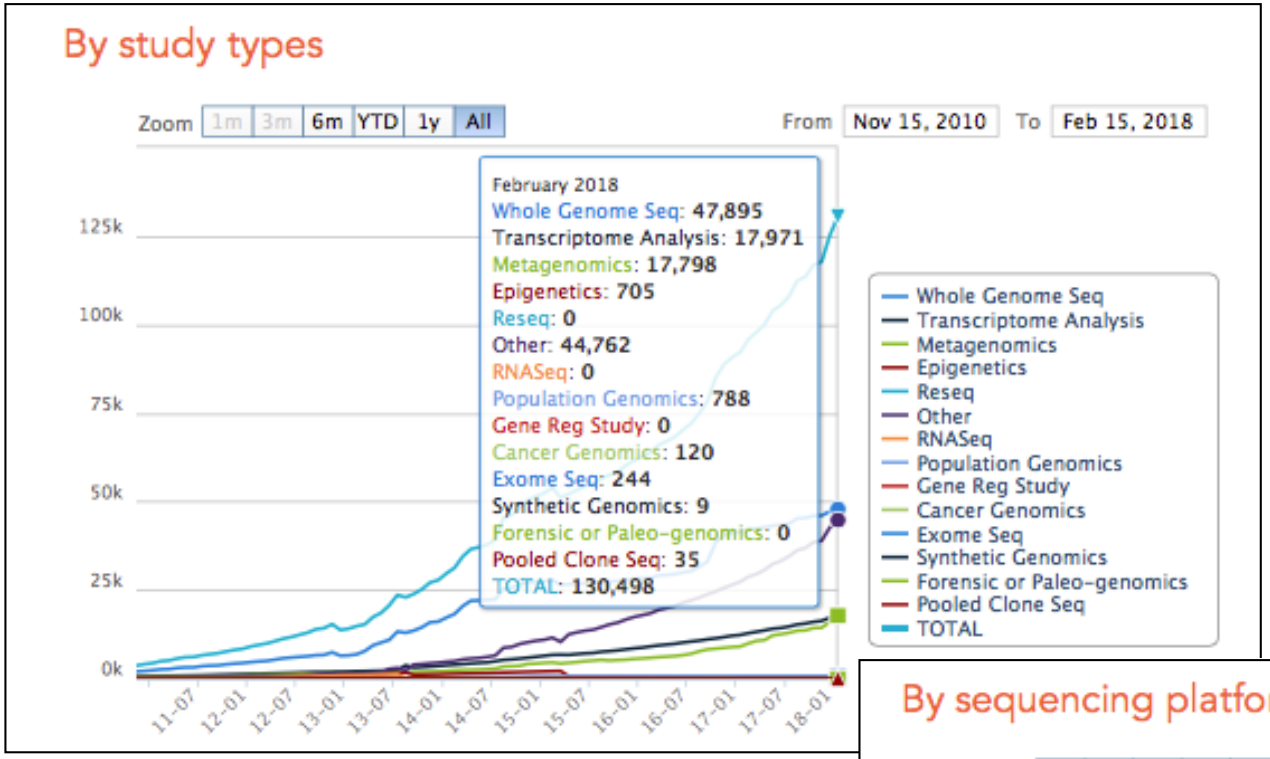
de novo アセンブリ

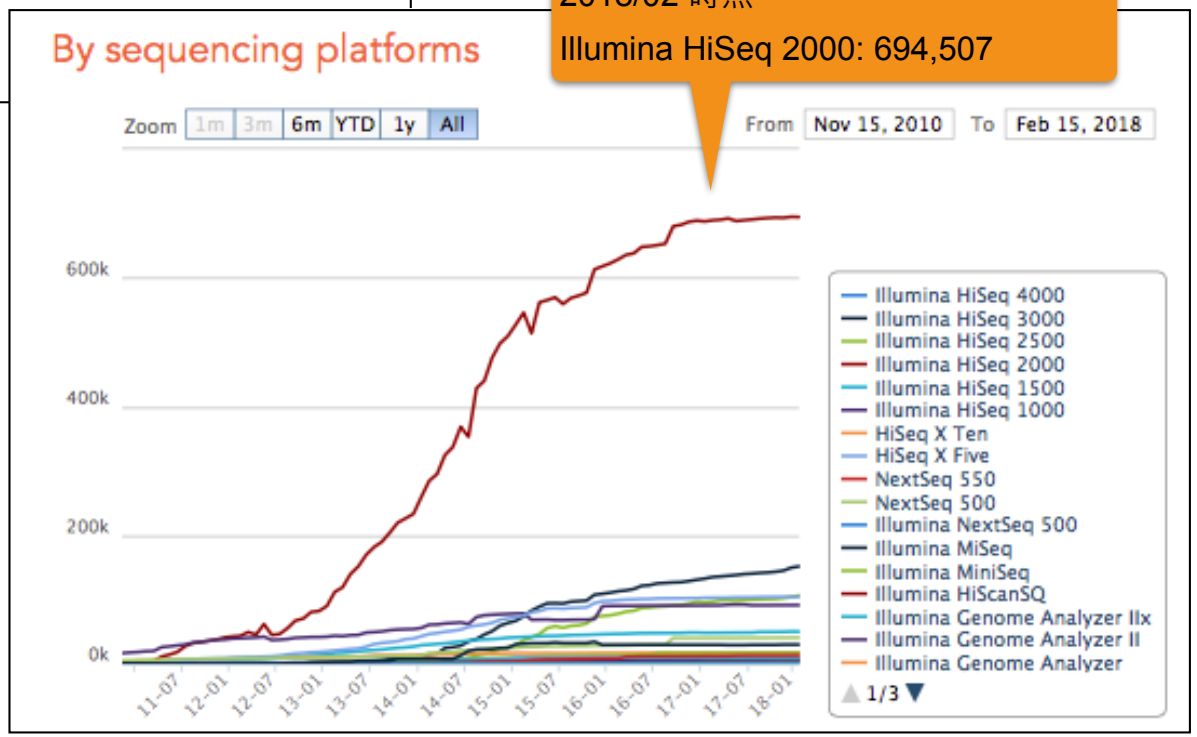
高次処理部

解析目的別ワークフロー

DNA多型
注釈
(DNApod)

SRA データ登録状況



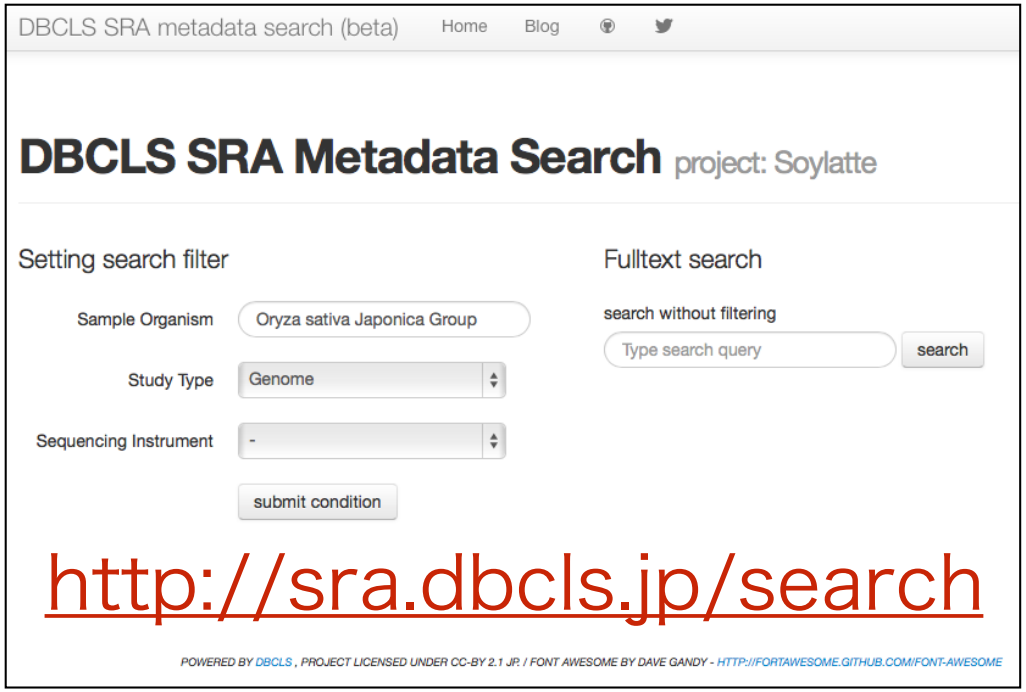


DBCLS SRA

DBCLS SRA

http://sra.dbcls.jp/trends.html

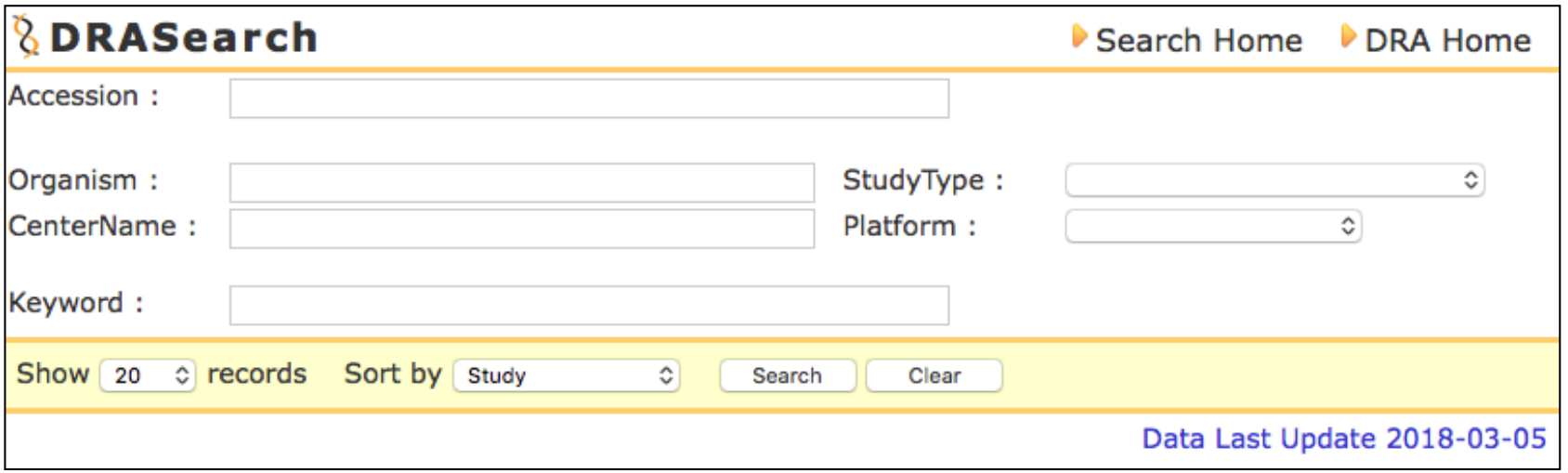
SRAデータサーチ



http://sra.dbcls.jp/search

https://www.ebi.ac.uk/ena/data/warehouse/search

http://ddbj.nig.ac.jp/DRASearch/



DDBJ pipeline 提供ツール一覧

Reference Genome Mapping

				Input data			Evaluation			Analysis		Output 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
☐	bwa		0.6.1	✓		✓	✓	✓	✓					✓	
☐	Bowtie		0.12.7	✓	✓	✓	✓	✓	✓	✓				✓	
☐	TopHat		1.0.11	✓		✓	✓	✓	✓					✓	
☐	Bowtie2		2.2.6	✓	✓	✓	✓	✓	✓	✓				✓	For reads longer than about 50 bp, Bowtie2 is generally faster, more senstitive, and uses less memory than Bowtie1.
☐	TopHat2		2.1.0	✓		✓	✓	✓	✓					✓	

de novo Assembly

	Tool	Help	Version	Base space	Color space	Paired-end	MSS(WGS)	Comment
☐	SOAPdenovo		2.04-r240	✓		✓		
☐	ABySS		1.3.2	✓		✓		Maximum K-mer value is 64.
☐	Velvet		1.2.10	✓		✓	✓	We severe recommend when performing Velvet, total length of those reads is up to 22G bp.Maximum K-mer value is 64.
☐	Trinity		2.1.1	✓		✓		RNA-Seq De novo Assembly
☐	Platanus		1.2.2	✓		✓		
☐	HGAP		Protocol3 (v 2.2.0)					HGAP Pipeline for PacBio Sequence based on SMRT Analysis v2.2.0. For bax.h5 file only. (Beta version)
☐	Canu		1.6					Nanopore only (Beta version)

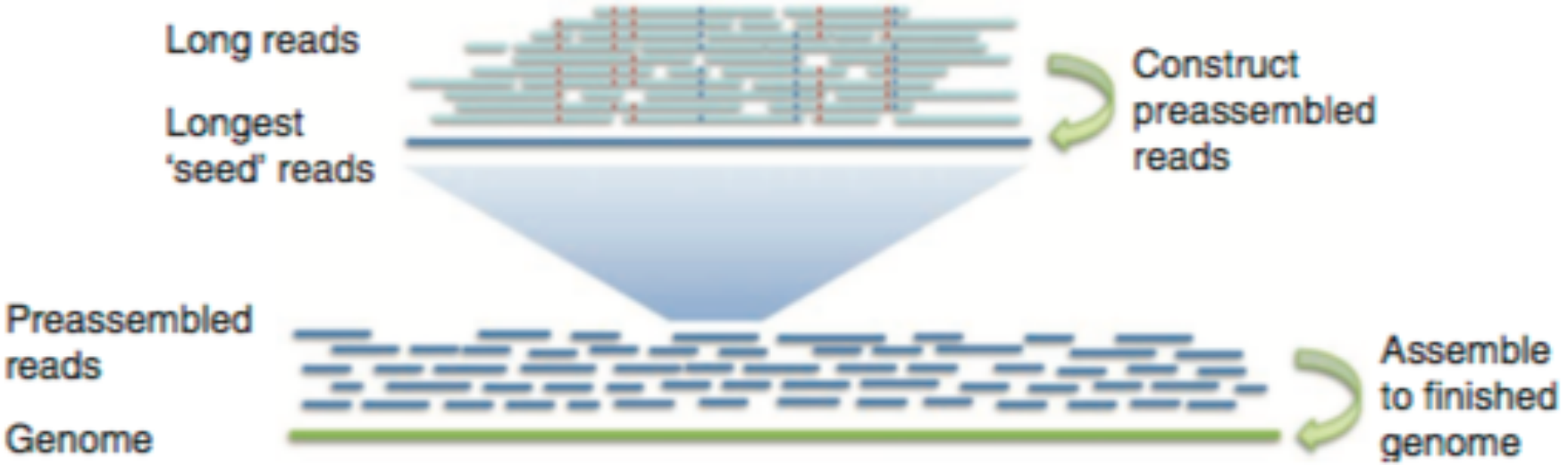
今回の講習はHGAP を使用します。

Mapping Contigs by de novo Assemble to Reference Sequences.

The contigs will be aligned to reference genome.

	Tool	Comment
◉	BLAT	Single-end analysis only

HGAP の処理の流れ



Chin C. S. *et al.*, Nature Methods, 10, 563–569, 2013

(1) Preassembly
　一定以上の長さのリード (シード) に、短いリードをマップし、エラー補正を行う。

(2) Assembly
　Preassemble reads をアセンブルする。

(3) Consensus Polishing
　コンティグの再補正を行う。

HGAP　version 2.2.0　　実行方法

入力ファイル

bax.h5 形式ファイル (DDBJ SRA では公開されていない)

1セルにつき 3 つの bax.h5 ファイル

パラメータ

GenomeSize: 推定ゲノムサイズ

Minimun Seed length: デフォルト 6000 及び Automatic Estimation

カバレージ が x25 以上あるのであれば、25x の リード で シードリード が補正されるように自動で 最小の シードリード 長を計算してくれる。もし、カバレージ が x25 よりも小さければ指定した値を最小の シードリード 長にする。

出力ファイル

アセンブル結果ファイル (fasta / fastq)

補正された ロングリードファイル

ログファイル

講習用データ

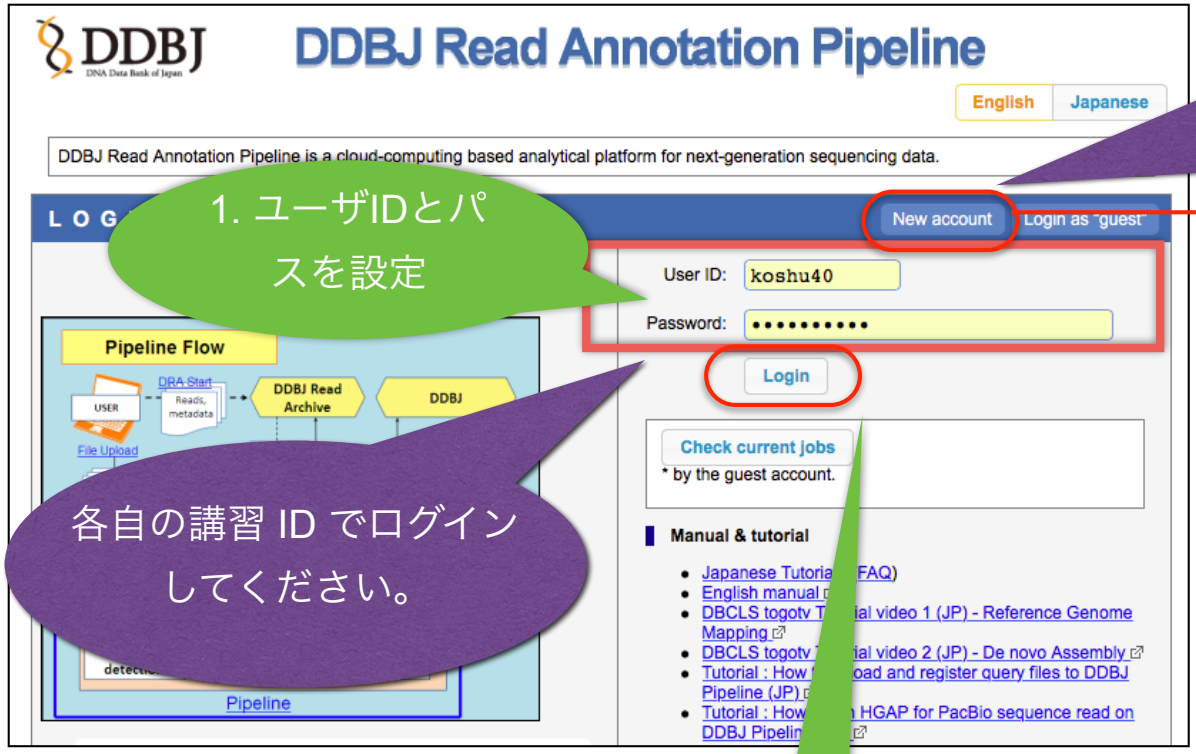
乳酸菌 *Lactobacillus hokkaidonensis LOOC260*^{T}

PacBio RS II で解読

アクセス

キーワード検索

ログイン

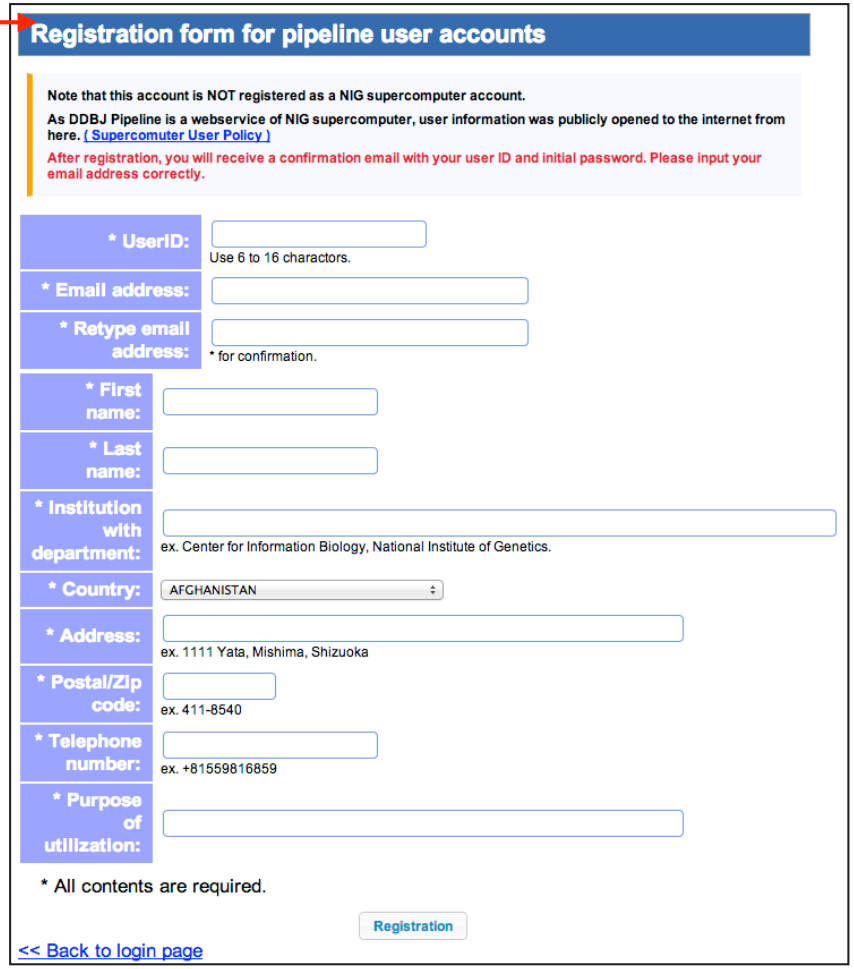


1. ユーザIDとパスを設定

各自の講習 ID でログインしてください。

2.Login をクリック

アカウントの作成はこちらから。

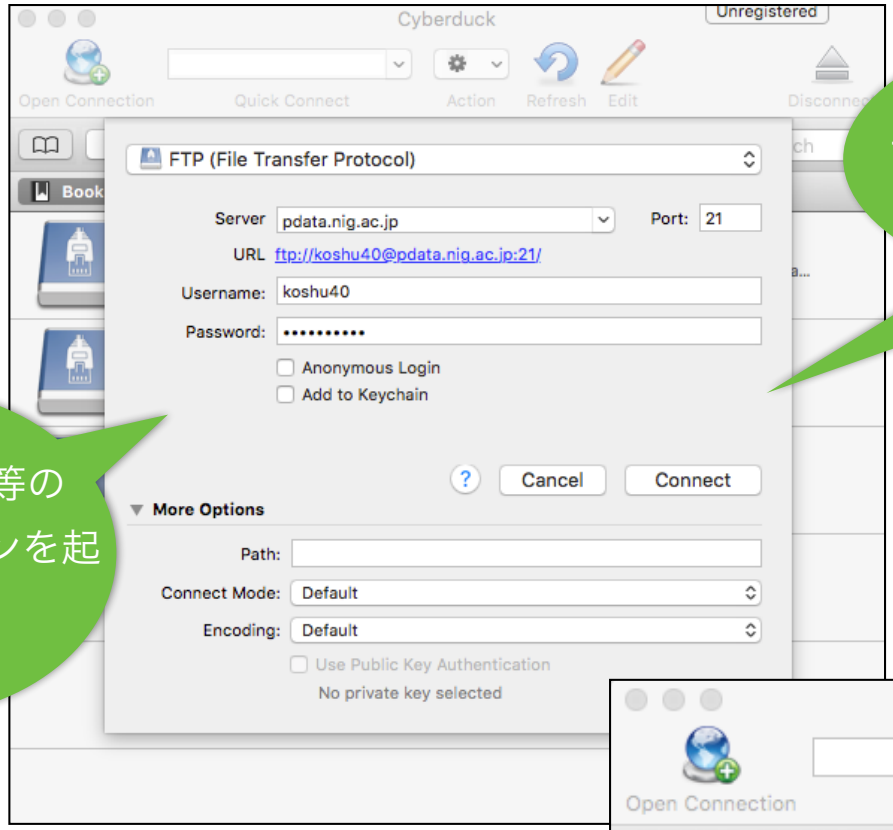


FTP ファイルアップロード

Step 1

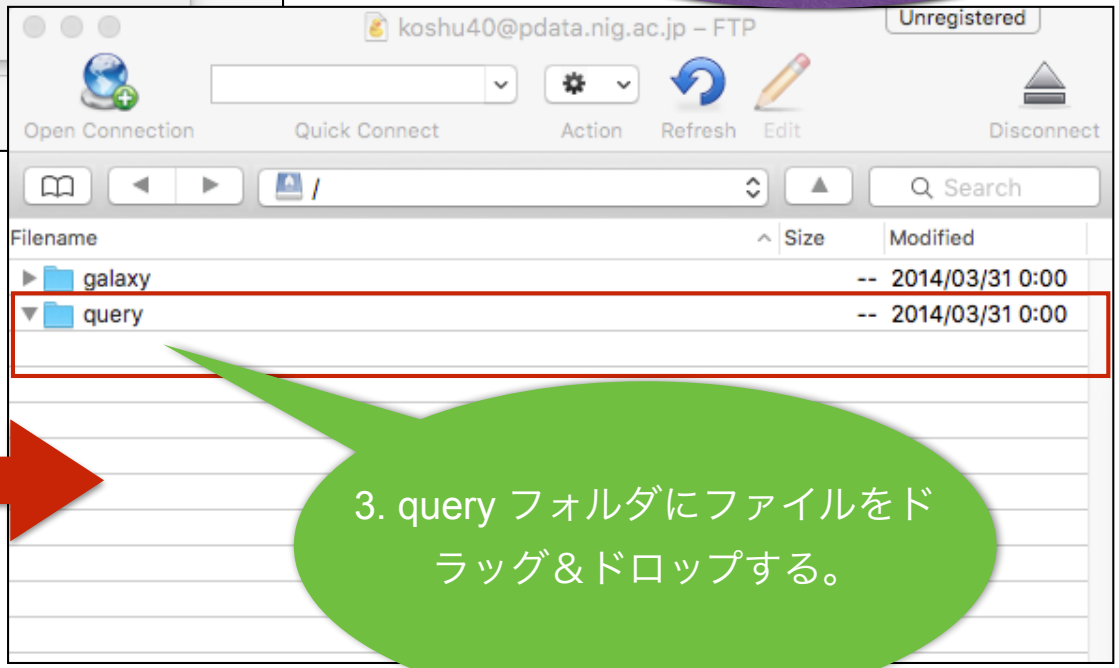
Cyberduck

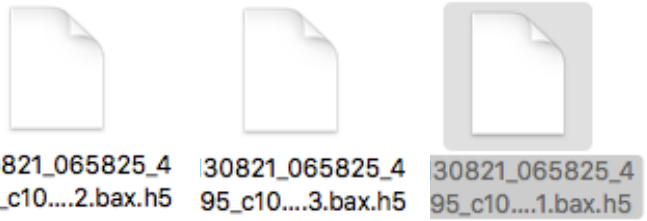
1. Cyberduck 等のアプリケーションを起動



2. FTP の情報を設定する。
サーバー名、ポート番号、ユーザID、パスワード

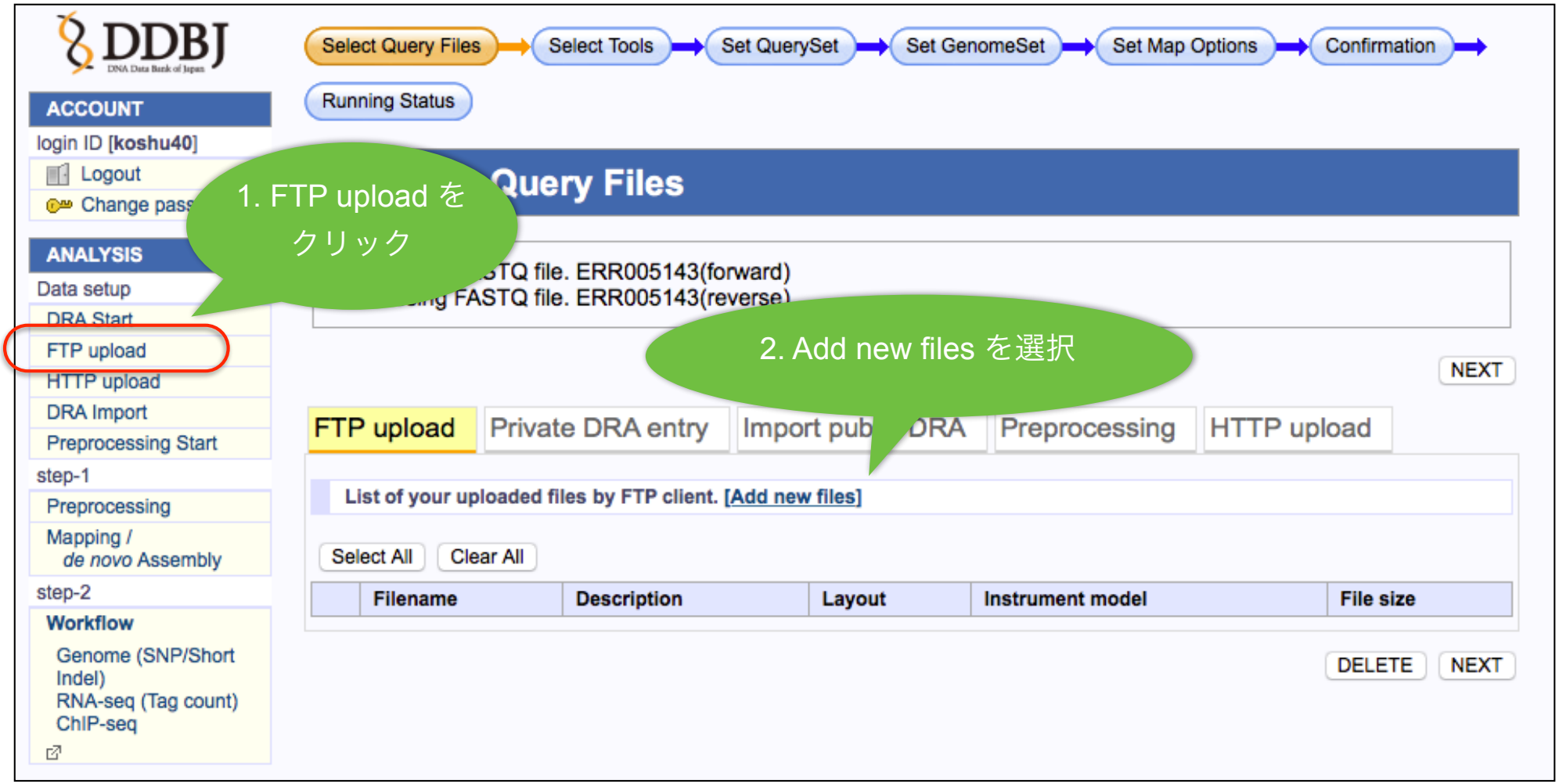
講習では 1 セル分のデータを使用。
このステップは準備済みです。





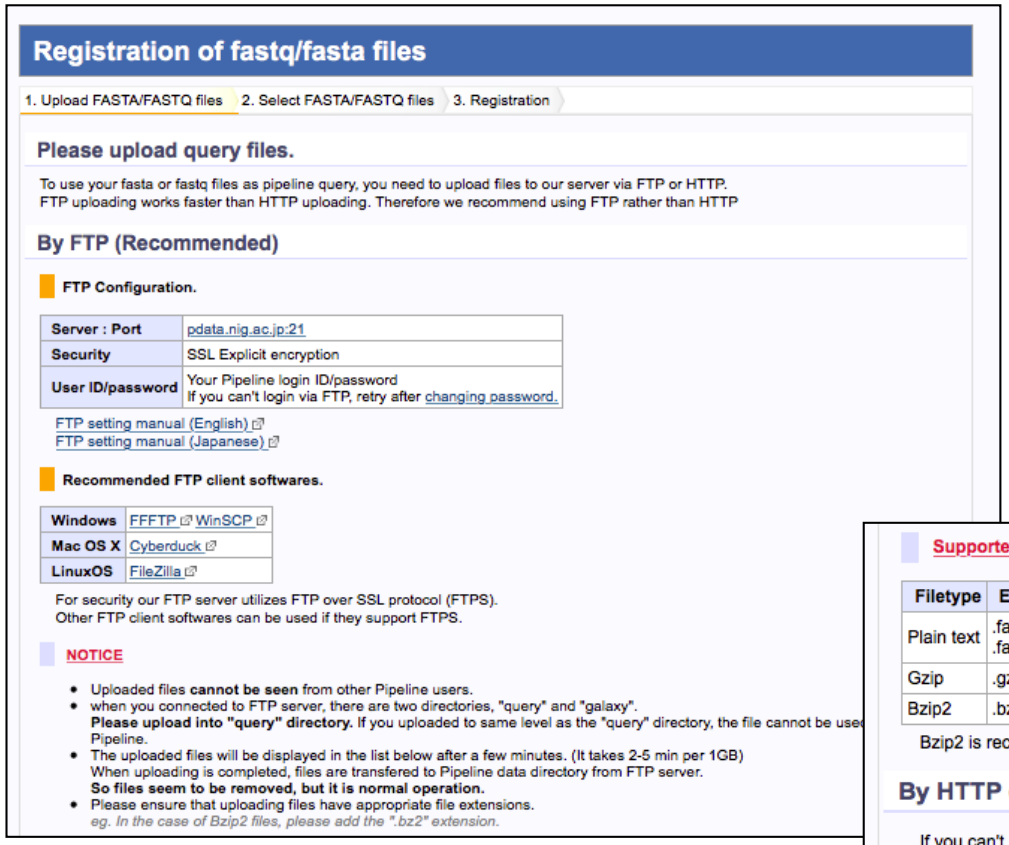
3. query フォルダにファイルをドラッグ＆ドロップする。

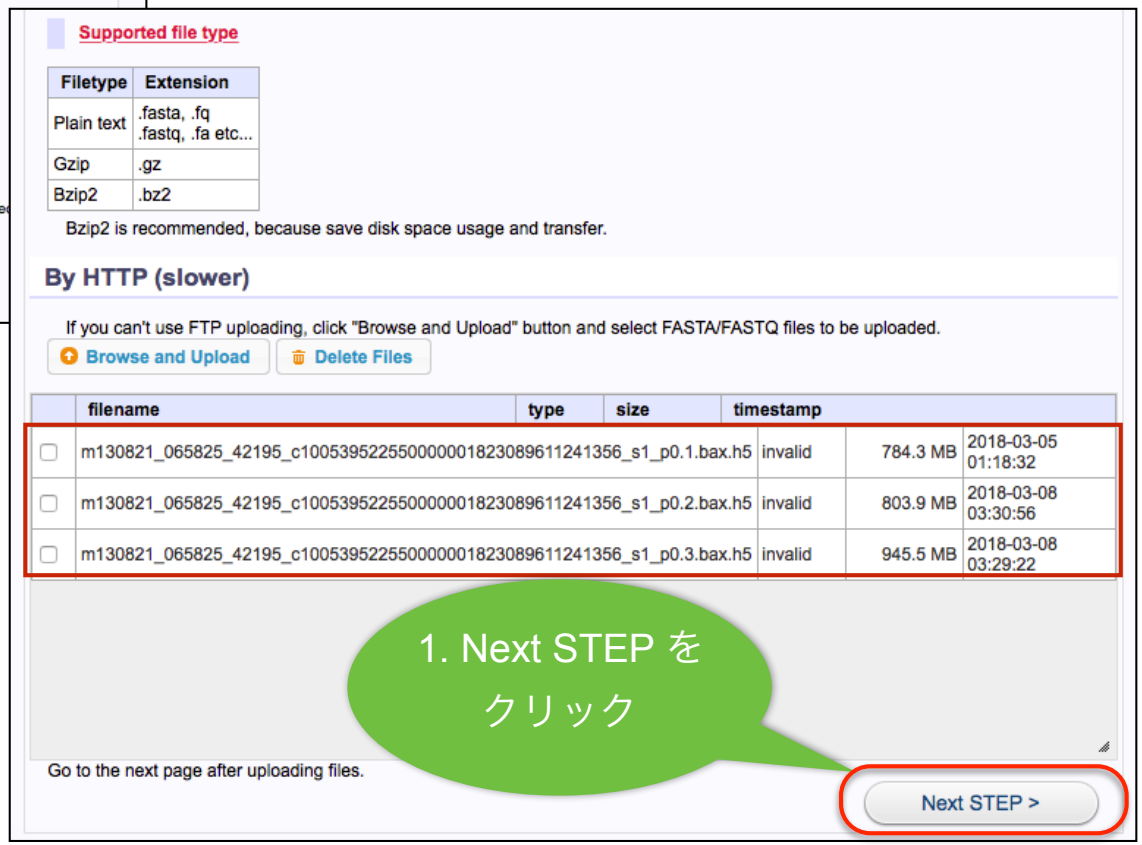
FTP ファイルアップロード　Step 2



FTP ファイルアップロード

Step 3



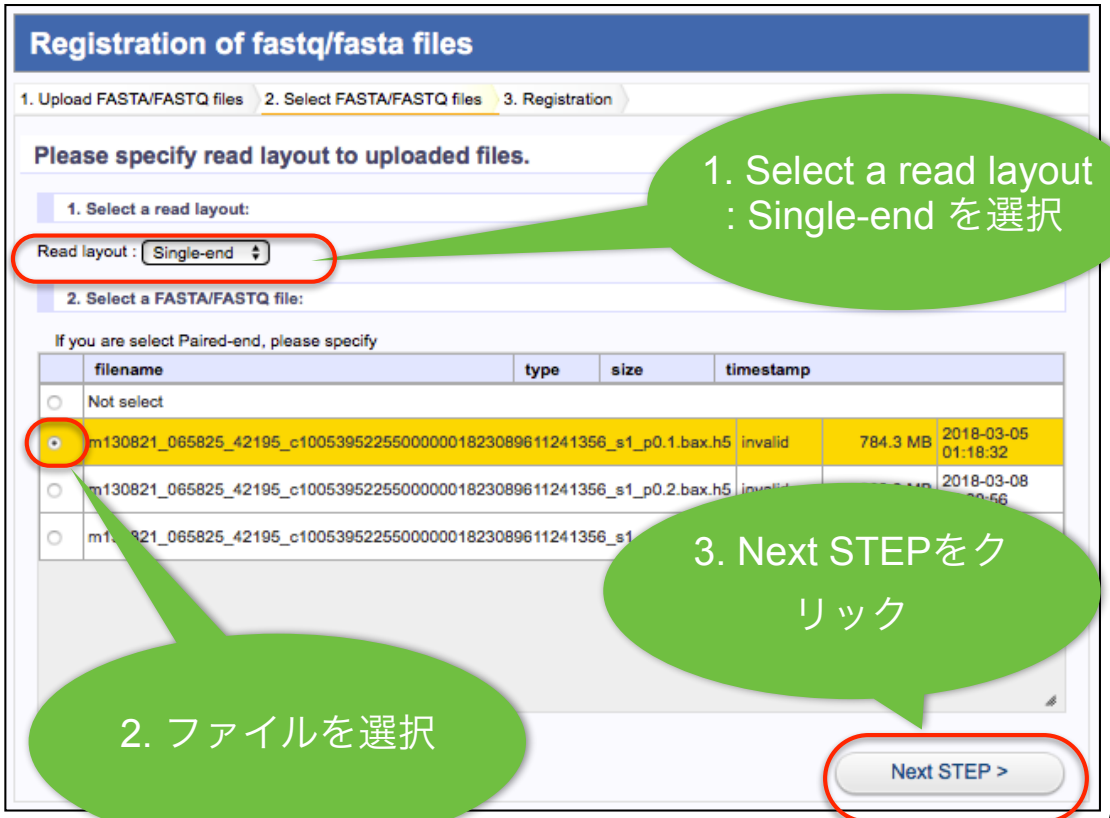


講習用のデータが3ファイルあるのを確認してください。

1. Next STEP をクリック

FTP ファイルアップロード

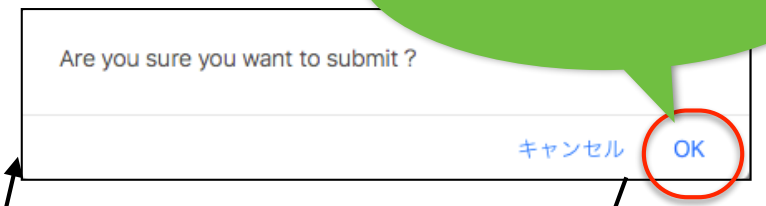
Step 4



1. Select a read layout : Single-end を選択

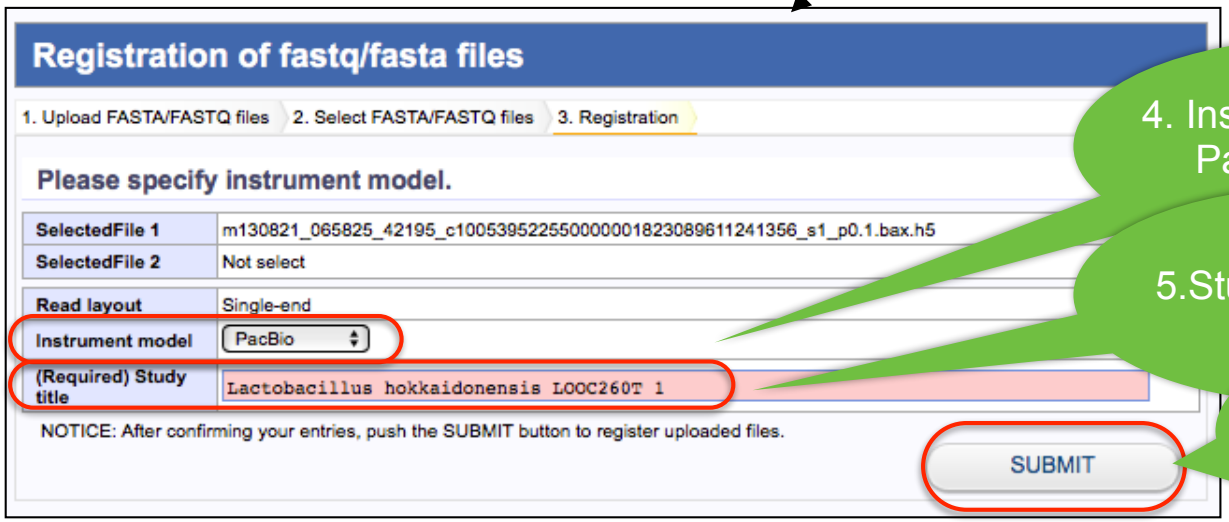
2. ファイルを選択

3. Next STEPをクリック



6. OK をクリック

7. Step 2 - 4 を3ファイル分行う。

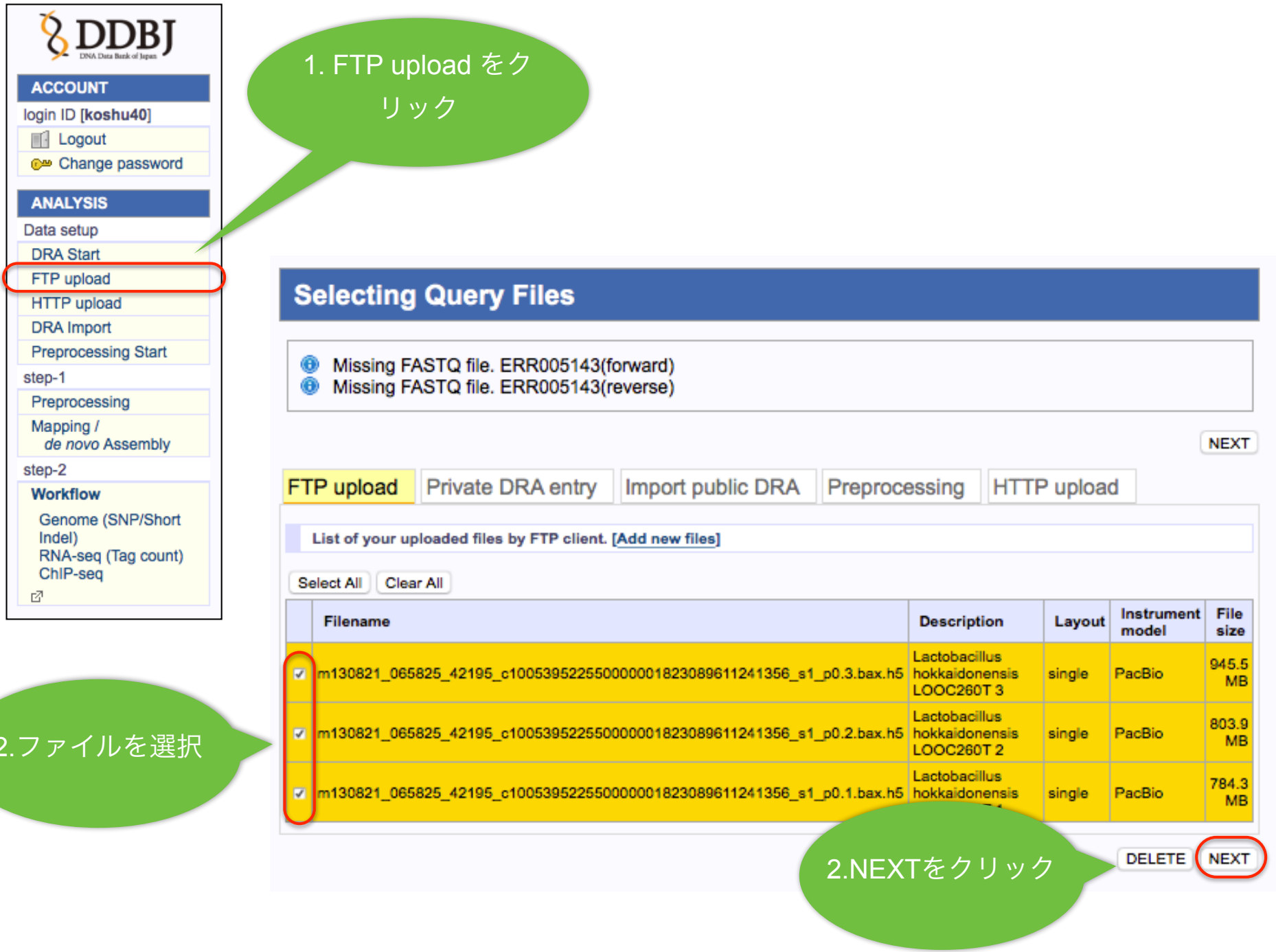


4. Instrument model: PacBio を選択

5.Study title を入力

6. SUBMIT をクリック

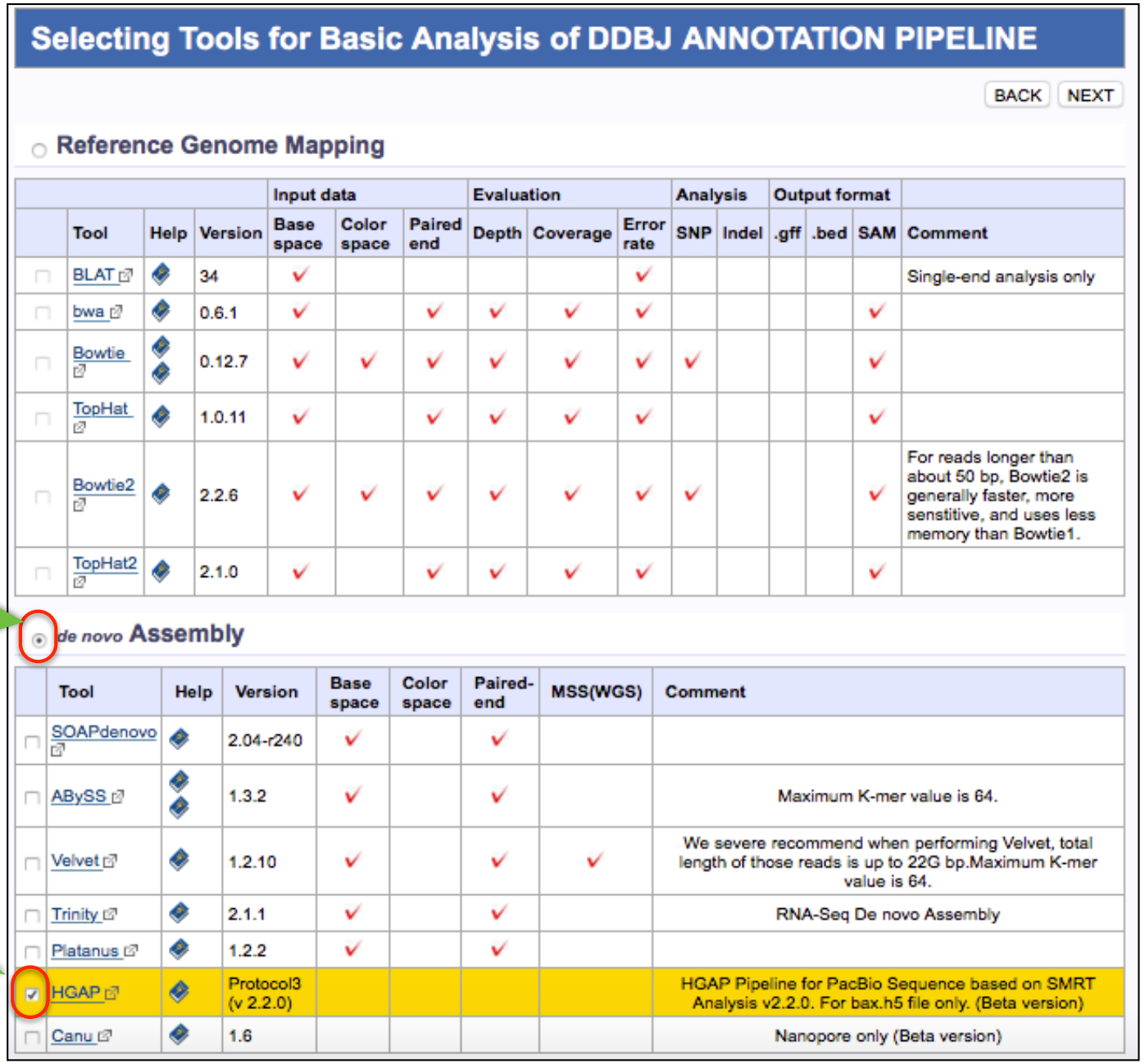
クエリの選択


DDBJ
DNA Data Bank of Japan
ACCOUNT
login ID [koshu40]
Logout
Change password
ANALYSIS
Data setup
DRA Start
FTP upload
HTTP upload
DRA Import
Preprocessing Start
step-1
Preprocessing
Mapping / de novo Assembly
step-2
Workflow
Genome (SNP/Short Indel)
RNA-seq (Tag count)
ChIP-seq
1. FTP upload をクリック
Selecting Query Files
Missing FASTQ file. ERR005143(forward)
Missing FASTQ file. ERR005143(reverse)
NEXT
FTP upload
Private DRA entry
Import public DRA
Preprocessing
HTTP upload
List of your uploaded files by FTP client. [Add new files]
Select All
Clear All
Filename
Description
Layout
Instrument model
File size
m130821_065825_42195_c100539522550000001823089611241356_s1_p0.3.bax.h5
Lactobacillus hokkaidonensis LOOC260T 3
single
PacBio
945.5 MB
m130821_065825_42195_c100539522550000001823089611241356_s1_p0.2.bax.h5
Lactobacillus hokkaidonensis LOOC260T 2
single
PacBio
803.9 MB
m130821_065825_42195_c100539522550000001823089611241356_s1_p0.1.bax.h5
Lactobacillus hokkaidonensis
single
PacBio
784.3 MB
2.ファイルを選択
DELETE
NEXT
2.NEXTをクリック

ツールの選択

1.*de novo* Assenbly を選択

2.HGAP を選択



Selecting Tools for Basic Analysis of DDBJ ANNOTATION PIPELINE

BACK NEXT

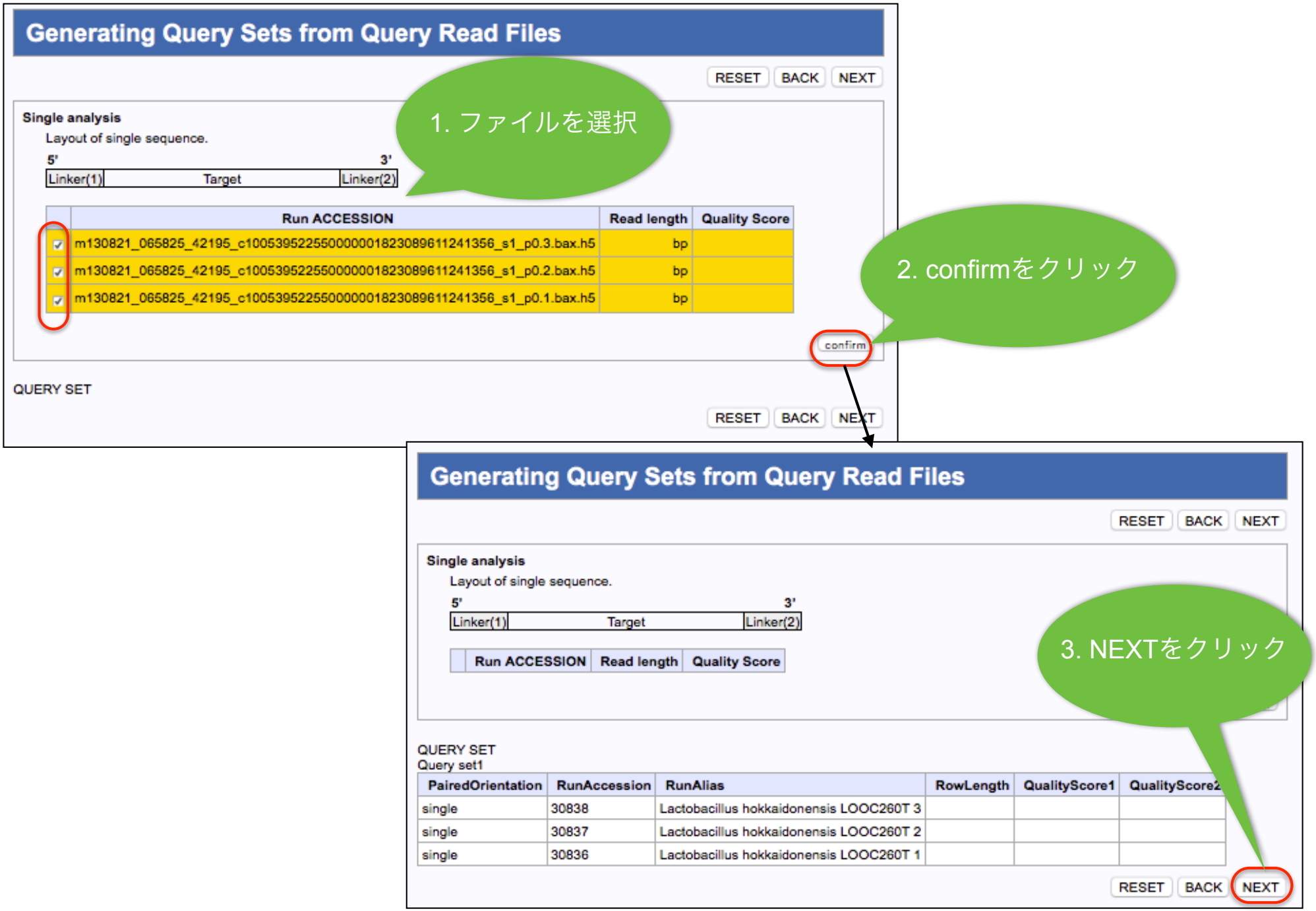
Reference Genome Mapping

				Input data			Evaluation			Analysis		Output 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
☐	bwa		0.6.1	✓		✓	✓	✓	✓					✓	
☐	Bowtie		0.12.7	✓	✓	✓	✓	✓	✓	✓				✓	
☐	TopHat		1.0.11	✓		✓	✓	✓	✓					✓	
☐	Bowtie2		2.2.6	✓	✓	✓	✓	✓	✓	✓				✓	For reads longer than about 50 bp, Bowtie2 is generally faster, more senstitive, and uses less memory than Bowtie1.
☐	TopHat2		2.1.0	✓		✓	✓	✓	✓					✓	

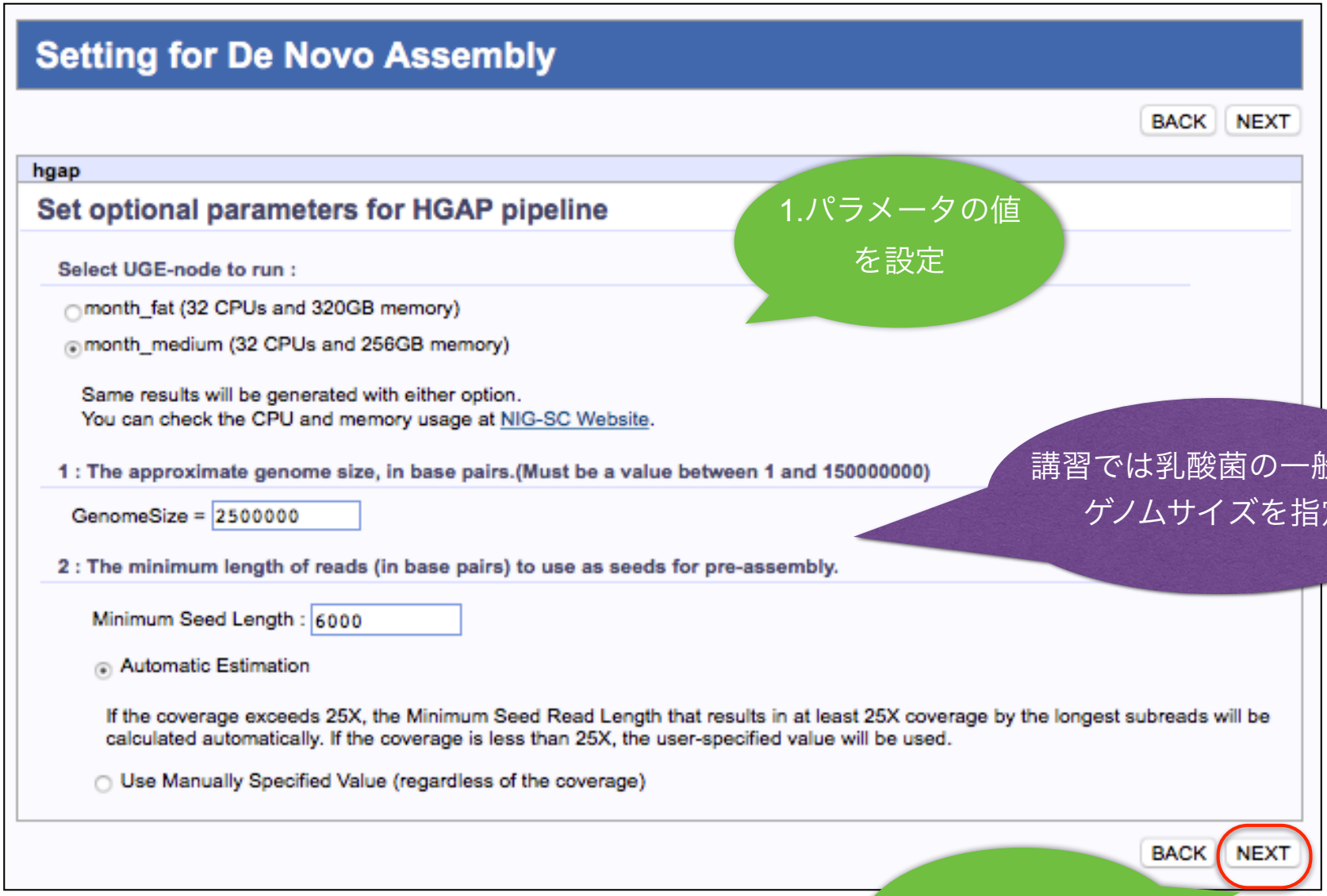
de novo Assembly

	Tool	Help	Version	Base space	Color space	Paired-end	MSS(WGS)	Comment
☐	SOAPdenovo		2.04-r240	✓		✓		
☐	ABySS		1.3.2	✓		✓		Maximum K-mer value is 64.
☐	Velvet		1.2.10	✓		✓	✓	We severe recommend when performing Velvet, total length of those reads is up to 22G bp.Maximum K-mer value is 64.
☐	Trinity		2.1.1	✓		✓		RNA-Seq De novo Assembly
☐	Platanus		1.2.2	✓		✓		
☑	HGAP		Protocol3 (v 2.2.0)					HGAP Pipeline for PacBio Sequence based on SMRT Analysis v2.2.0. For bax.h5 file only. (Beta version)
☐	Canu		1.6					Nanopore only (Beta version)

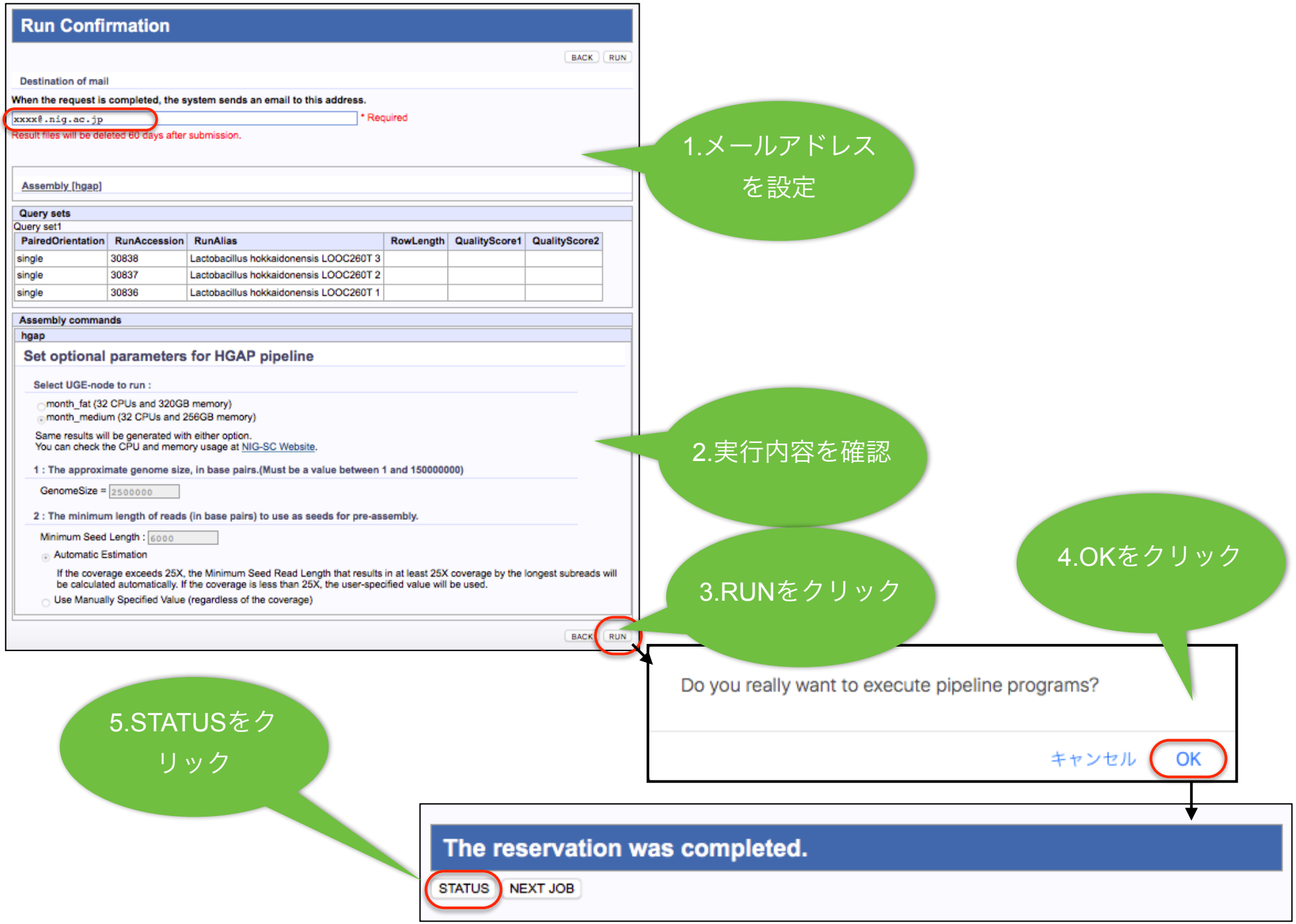
クエリセットの作成


Generating Query Sets from Query Read Files
RESET
BACK
NEXT
Single analysis
Layout of single sequence.
5'
3'
Linker(1)
Target
Linker(2)
Run ACCESSION
Read length
Quality Score
m130821_065825_42195_c100539522550000001823089611241356_s1_p0.3.bax.h5
bp
m130821_065825_42195_c100539522550000001823089611241356_s1_p0.2.bax.h5
bp
m130821_065825_42195_c100539522550000001823089611241356_s1_p0.1.bax.h5
bp
1. ファイルを選択
2. confirmをクリック
confirm
QUERY SET
RESET
BACK
NEXT
Generating Query Sets from Query Read Files
RESET
BACK
NEXT
Single analysis
Layout of single sequence.
5'
3'
Linker(1)
Target
Linker(2)
Run ACCESSION
Read length
Quality Score
3. NEXTをクリック
QUERY SET
Query set1
PairedOrientation
RunAccession
RunAlias
RowLength
QualityScore1
QualityScore2
single
30838
Lactobacillus hokkaidonensis LOOC260T 3
single
30837
Lactobacillus hokkaidonensis LOOC260T 2
single
30836
Lactobacillus hokkaidonensis LOOC260T 1
RESET
BACK
NEXT

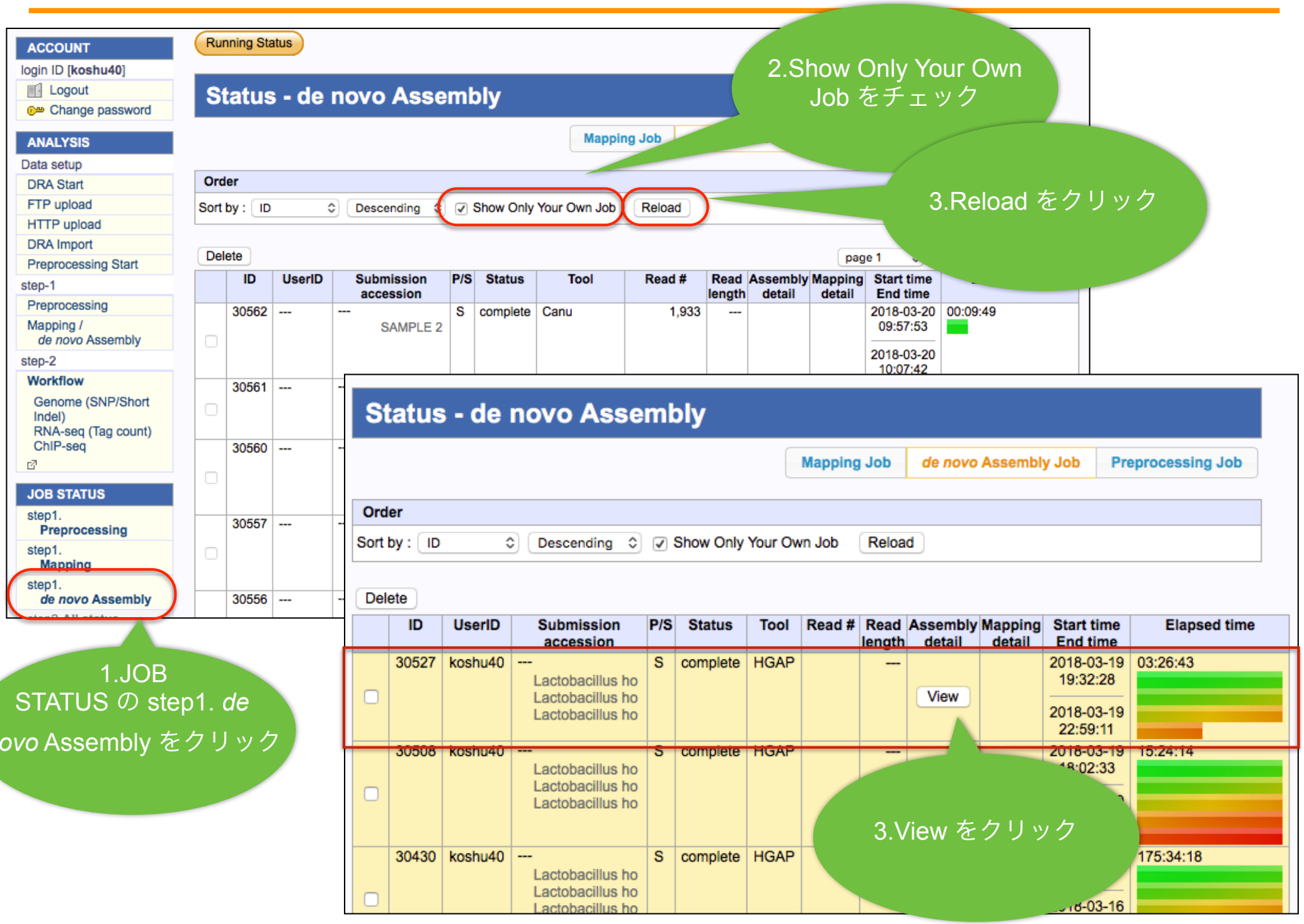
パラメータの設定


Setting for De Novo Assembly
BACK
NEXT
hgap
Set optional parameters for HGAP pipeline
1.パラメータの値を設定
Select UGE-node to run :
month_fat (32 CPUs and 320GB memory)
month_medium (32 CPUs and 256GB memory)
Same results will be generated with either option.
You can check the CPU and memory usage at NIG-SC Website.
1 : The approximate genome size, in base pairs.(Must be a value between 1 and 150000000)
GenomeSize = 2500000
講習では乳酸菌の一般的なゲノムサイズを指定
2 : The minimum length of reads (in base pairs) to use as seeds for pre-assembly.
Minimum Seed Length : 6000
Automatic Estimation
If the coverage exceeds 25X, the Minimum Seed Read Length that results in at least 25X coverage by the longest subreads will be calculated automatically. If the coverage is less than 25X, the user-specified value will be used.
Use Manually Specified Value (regardless of the coverage)
BACK
NEXT
2.NEXTの値を設定

実行の確認


Run Confirmation
BACK
RUN
Destination of mail
When the request is completed, the system sends an email to this address.
xxxx@.nig.ac.jp
* Required
Result files will be deleted 60 days after submission.
Assembly [hgap]
Query sets
Query set1
PairedOrientation
RunAccession
RunAlias
RowLength
QualityScore1
QualityScore2
single 30838 Lactobacillus hokkaidonensis LOOC260T 3
single 30837 Lactobacillus hokkaidonensis LOOC260T 2
single 30836 Lactobacillus hokkaidonensis LOOC260T 1
Assembly commands
hgap
Set optional parameters for HGAP pipeline
Select UGE-node to run :
month_fat (32 CPUs and 320GB memory)
month_medium (32 CPUs and 256GB memory)
Same results will be generated with either option.
You can check the CPU and memory usage at NIG-SC Website.
1 : The approximate genome size, in base pairs.(Must be a value between 1 and 150000000)
GenomeSize = 2500000
2 : The minimum length of reads (in base pairs) to use as seeds for pre-assembly.
Minimum Seed Length : 6000
Automatic Estimation
If the coverage exceeds 25X, the Minimum Seed Read Length that results in at least 25X coverage by the longest subreads will be calculated automatically. If the coverage is less than 25X, the user-specified value will be used.
Use Manually Specified Value (regardless of the coverage)
BACK
RUN
1.メールアドレスを設定
2.実行内容を確認
3.RUNをクリック
4.OKをクリック
Do you really want to execute pipeline programs?
キャンセル
OK
5.STATUSをクリック
The reservation was completed.
STATUS
NEXT JOB

実行の確認

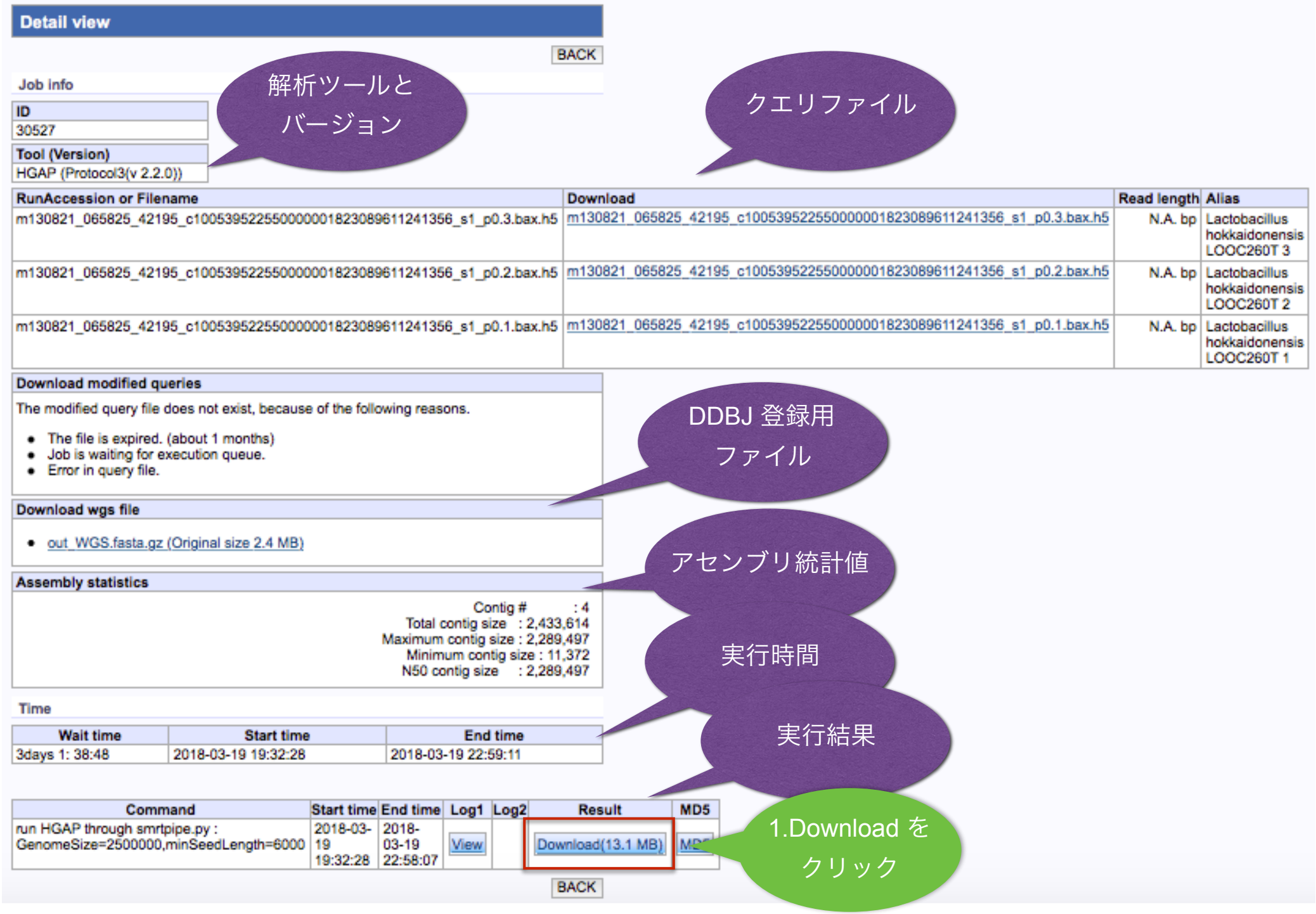

ACCOUNT
login ID [koshu40]
Logout
Change password
ANALYSIS
Data setup
DRA Start
FTP upload
HTTP upload
DRA Import
Preprocessing Start
step-1
Preprocessing
Mapping / de novo Assembly
step-2
Workflow
Genome (SNP/Short Indel)
RNA-seq (Tag count)
ChIP-seq
JOB STATUS
step1. Preprocessing
step1. Mapping
step1. de novo Assembly
Running Status
Status - de novo Assembly
Mapping Job
Order
Sort by : ID
Descending
Show Only Your Own Job
Reload
Delete
page 1
ID
UserID
Submission accession
P/S
Status
Tool
Read #
Read length
Assembly detail
Mapping detail
Start time
End time
30562

SAMPLE 2
S
complete
Canu
1,933
2018-03-20 09:57:53
2018-03-20 10:07:42
00:09:49
30561
30560
30557
30556
2.Show Only Your Own Job をチェック
3.Reload をクリック
Status - de novo Assembly
Mapping Job
de novo Assembly Job
Preprocessing Job
Order
Sort by : ID
Descending
Show Only Your Own Job
Reload
Delete
ID
UserID
Submission accession
P/S
Status
Tool
Read #
Read length
Assembly detail
Mapping detail
Start time
End time
Elapsed time
30527
koshu40

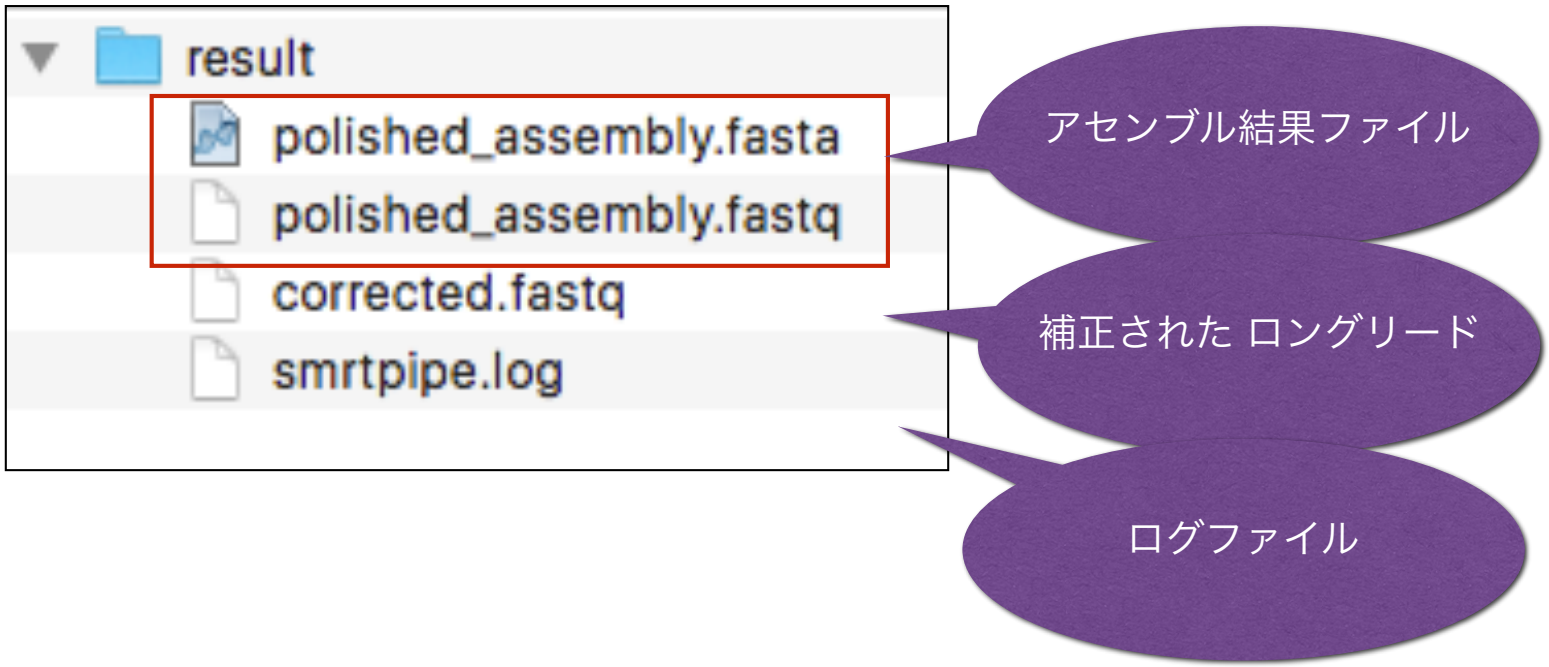
Lactobacillus ho
Lactobacillus ho
Lactobacillus ho
S
complete
HGAP

View
2018-03-19 19:32:28
2018-03-19 22:59:11
03:26:43
30508
15:24:14
30430
175:34:18
1.JOB STATUS の step1. de novo Assembly をクリック
3.View をクリック

結果画面


Detail view
BACK
Job info
ID
30527
Tool (Version)
HGAP (Protocol3(v 2.2.0))
解析ツールとバージョン
クエリファイル
RunAccession or Filename
Download
Read length
Alias
m130821_065825_42195_c100539522550000001823089611241356_s1_p0.3.bax.h5
m130821_065825_42195_c100539522550000001823089611241356_s1_p0.3.bax.h5
N.A. bp
Lactobacillus hokkaidonensis LOOC260T 3
m130821_065825_42195_c100539522550000001823089611241356_s1_p0.2.bax.h5
m130821_065825_42195_c100539522550000001823089611241356_s1_p0.2.bax.h5
N.A. bp
Lactobacillus hokkaidonensis LOOC260T 2
m130821_065825_42195_c100539522550000001823089611241356_s1_p0.1.bax.h5
m130821_065825_42195_c100539522550000001823089611241356_s1_p0.1.bax.h5
N.A. bp
Lactobacillus hokkaidonensis LOOC260T 1
Download modified queries
The modified query file does not exist, because of the following reasons.
The file is expired. (about 1 months)
Job is waiting for execution queue.
Error in query file.
DDBJ 登録用ファイル
Download wgs file
out_WGS.fasta.gz (Original size 2.4 MB)
アセンブリ統計値
Assembly statistics
Contig # : 4
Total contig size : 2,433,614
Maximum contig size : 2,289,497
Minimum contig size : 11,372
N50 contig size : 2,289,497
実行時間
Time
Wait time
Start time
End time
3days 1: 38:48
2018-03-19 19:32:28
2018-03-19 22:59:11
実行結果
Command
Start time
End time
Log1
Log2
Result
MD5
run HGAP through smrtpipe.py : GenomeSize=2500000,minSeedLength=6000
2018-03-19 19:32:28
2018-03-19 22:58:07
View
Download(13.1 MB)
1.Download をクリック
BACK

結果ファイル


result
polished_assembly.fasta
polished_assembly.fastq
corrected.fastq
smrtpipe.log
アセンブル結果ファイル
補正された ロングリード
ログファイル

ご清聴ありがとうございました