2015-11-26

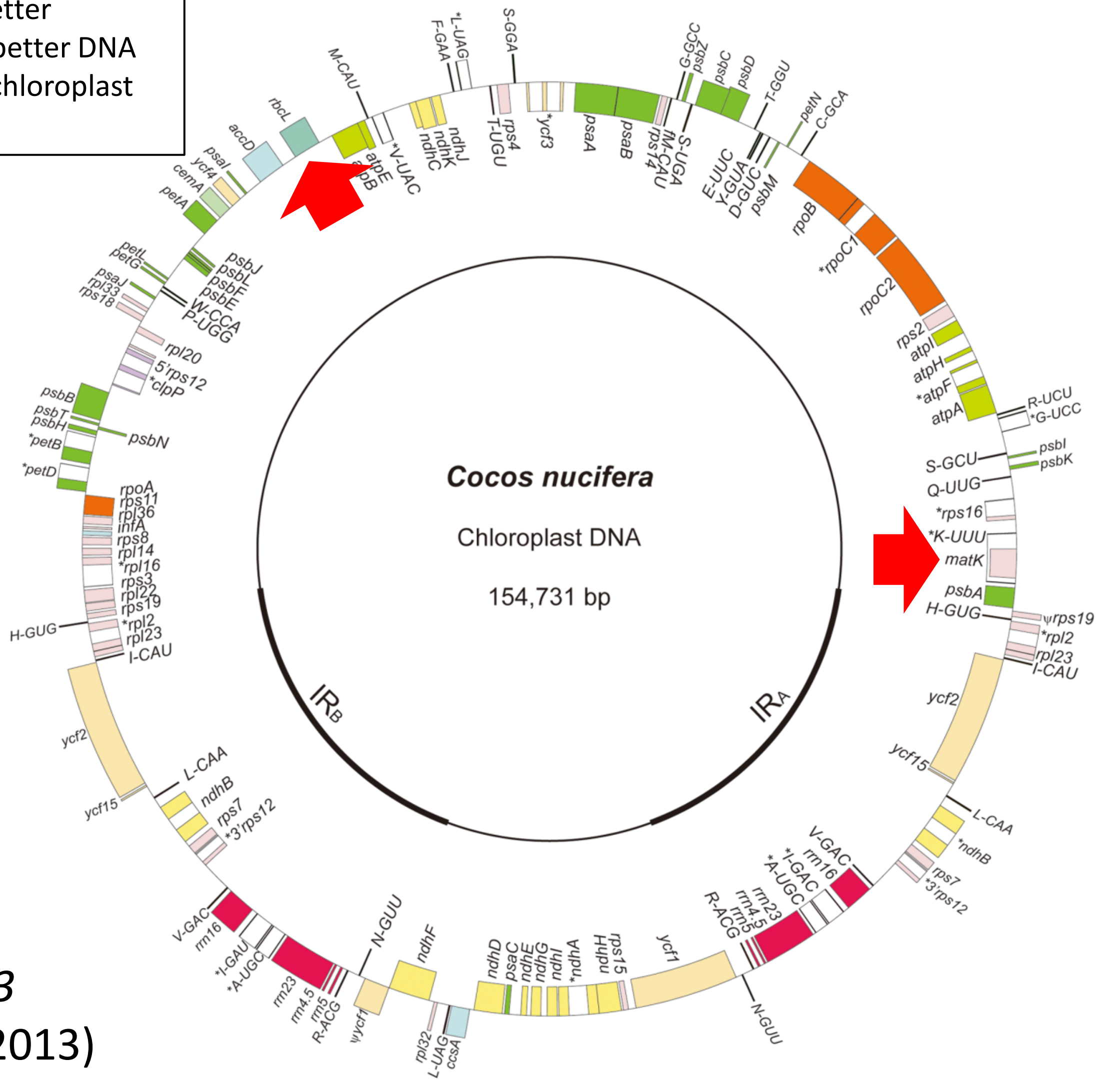
3. DNA Barcode

- Plant barcodes(matK, rbcL)
- BOLD/GBIF DB
- GBIF entries of ‘Palm Trees’

Plant DNA barcodes

In 2009, a collaboration of a large group of plant DNA barcode researchers proposed two chloroplast genes, rbcL and matK, taken together, as a barcode for plants.[6] Adding the nuclear internal transcribed spacer ITS2 region was proposed to provide better resolution between species.[21] As of 2015, the search for better DNA barcodes for plants continues, with the proposal that the chloroplast region *ycf1* may be suitable.

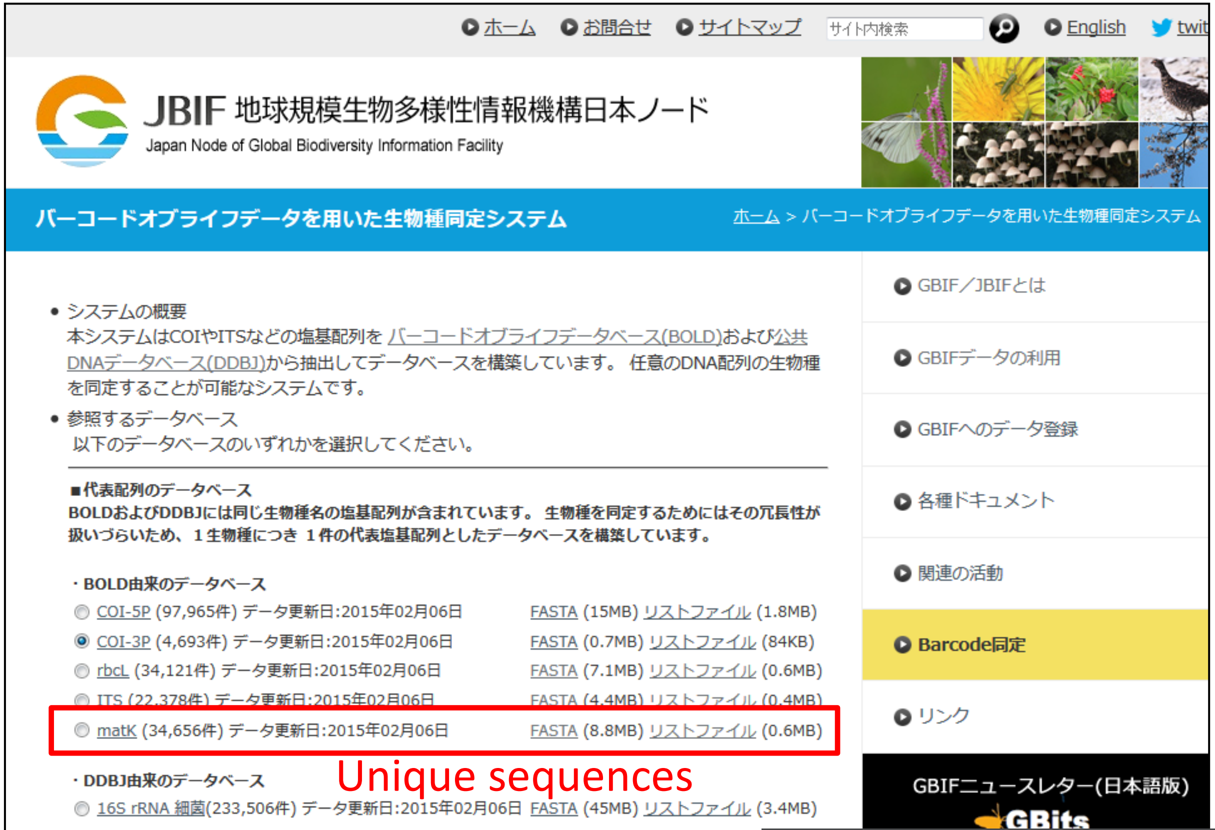
https://en.wikipedia.org/wiki/DNA_barcoding



PMID:*24023703*
(Huang et al., 2013)

GBIF (Global Biodiversity Information Facility)

http://www.gbif.jp/bol/



JBIF 地球規模生物多様性情報機構日本ノード
Japan Node of Global Biodiversity Information Facility

バーコードオブライフデータを用いた生物種同定システム

ホーム > バーコードオブライフデータを用いた生物種同定システム

- システムの概要
  本システムはCOIやITSなどの塩基配列を バーコードオブライフデータベース(BOLD)および公共DNAデータベース(DDBJ)から抽出してデータベースを構築しています。 任意のDNA配列の生物種を同定することが可能なシステムです。
- 参照するデータベース
  以下のデータベースのいずれかを選択してください。

■代表配列のデータベース
BOLDおよびDDBJには同じ生物種名の塩基配列が含まれています。 生物種を同定するためにはその冗長性が扱いづらいため、１生物種につき １件の代表塩基配列としたデータベースを構築しています。

・BOLD由来のデータベース

○ COI-5P (97,965件) データ更新日:2015年02月06日　FASTA (15MB) リストファイル (1.8MB)
● COI-3P (4,693件) データ更新日:2015年02月06日　FASTA (0.7MB) リストファイル (84KB)
○ rbcL (34,121件) データ更新日:2015年02月06日　FASTA (7.1MB) リストファイル (0.6MB)
○ ITS (22,378件) データ更新日:2015年02月06日　FASTA (4.4MB) リストファイル (0.4MB)
○ matK (34,656件) データ更新日:2015年02月06日　FASTA (8.8MB) リストファイル (0.6MB)

Unique sequences

・DDBJ由来のデータベース

○ 16S rRNA 細菌(233,506件) データ更新日:2015年02月06日 FASTA (45MB) リストファイル (3.4MB)

■全件のデータベース
データベースの塩基配列全件に対して比較することが可能です。

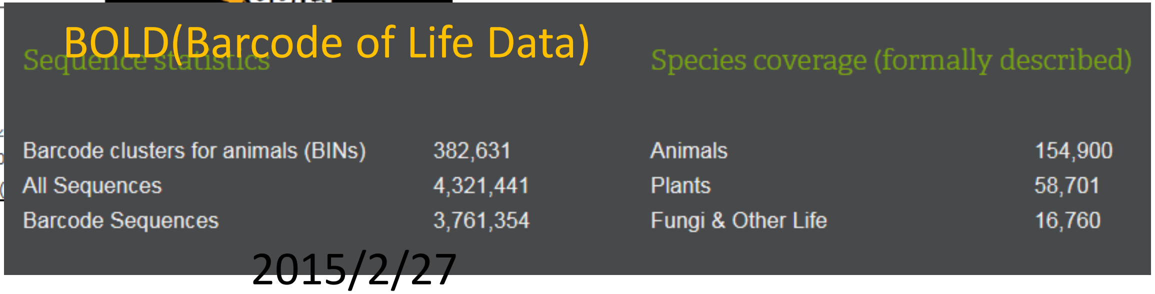
・BOLD由来のデータベース

○ COI-5P (2,807,009件) データ更新日:2015年02月06日　FASTA (380MB) リストファイル
○ COI-3P (20,898件) データ更新日:2015年02月06日　FASTA (3MB) リストファイル (0
○ rbcL (77,415件) データ更新日:2015年02月06日　FASTA (13MB) リストファイル (

GBIF　Japan Node
↓
JBIF

Dr.Yamazaki(NBRP) supports JBIF database.

BOLD and DDBJ sources



BOLD(Barcode of Life Data)

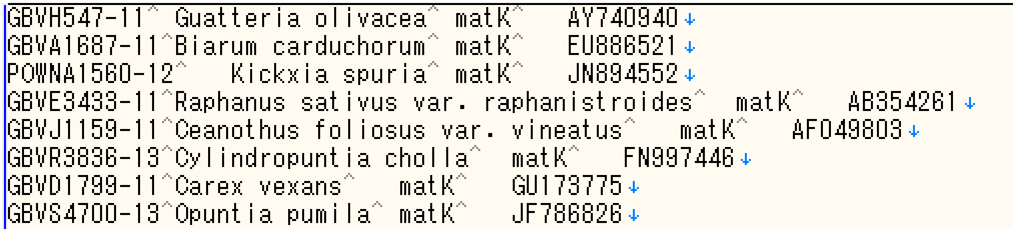
Sequence statistics		Species coverage (formally described)	
Barcode clusters for animals (BINs)	382,631	Animals	154,900
All Sequences	4,321,441	Plants	58,701
Barcode Sequences	3,761,354	Fungi & Other Life	16,760

2015/2/27

ASSIGNMENT[6]

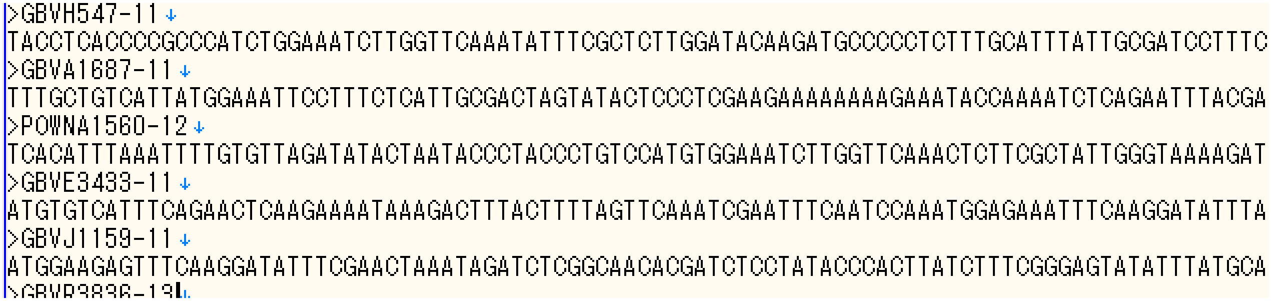
Download files: GBIF matK sequences

(1)matK_rpsv.list (TSV format: Accession ID, Species name, barcode name, GenBank ID)



```
GBVH547-11^ Guatteria olivacea^ matK^   AY740940↓
GBVA1687-11^Biarum carduchorum^ matK^   EU886521↓
POWNA1560-12^   Kickxia spuria^ matK^   JN894552↓
GBVE3433-11^Raphanus sativus var. raphanistroides^  matK^   AB354261↓
GBVJ1159-11^Ceanothus foliosus var. vineatus^   matK^   AF049803↓
GBVR3836-13^Cylindropuntia cholla^  matK^   FN997446↓
GBVD1799-11^Carex vexans^   matK^   GU173775↓
GBVS4700-13^Opuntia pumila^ matK^   JF786826↓
```

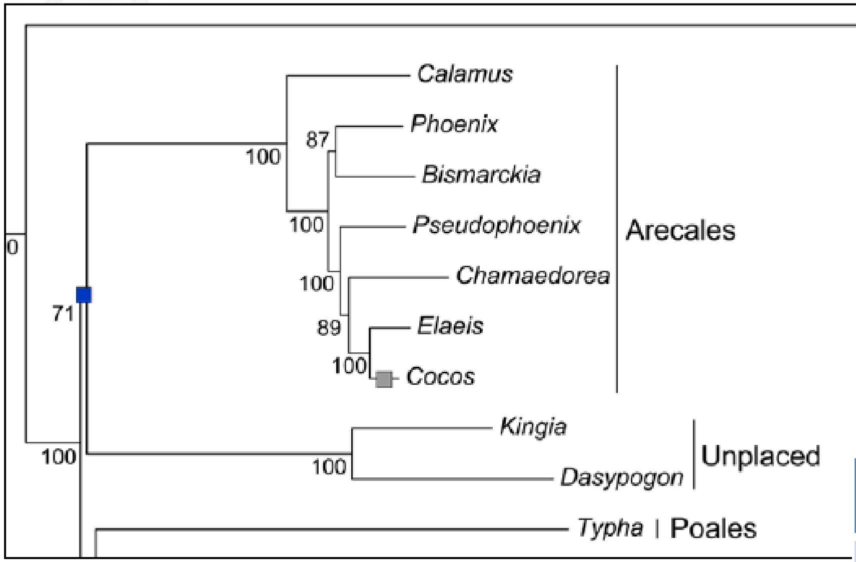
(2)matK_rpsv.fasta (fasta format: >Accession ID, Sequence)



```
>GBVH547-11↓
TACCTCACCCCGCCCATCTGGAAATCTTGGTTCAAATATTTCGCTCTTGGATACAAGATGCCCCCTCTTTGCATTTATTGCGATCCTTTC
>GBVA1687-11↓
TTTGCTGTCATTATGGAAATTCCTTTCTCATTGCGACTAGTATACTCCCTCGAAGAAAAAAAAGAAATACCAAAATCTCAGAATTTACGA
>POWNA1560-12↓
TCACATTTAAATTTTGTGTTAGATATACTAATACCCTACCCTGTCCATGTGGAAATCTTGGTTCAAACTCTTCGCTATTGGGTAAAAGAT
>GBVE3433-11↓
ATGTGTCATTTCAGAACTCAAGAAAATAAAGACTTTACTTTTAGTTCAAATCGAATTTCAATCCAAATGGAGAAATTTCAAGGATATTTA
>GBVJ1159-11↓
ATGGAAGAGTTTCAAGGATATTTCGAACTAAATAGATCTCGGCAACACGATCTCCTATACCCACTTATCTTTCGGGAGTATATTTATGCA
>GBVR3836-13↓
```

ASSIGNMENT[6]
Extract matK sequences of Cocos nucifera and other palm trees and by programming and perform phylogenetic analysis

Reference : Palm trees and matK entries



PMID:*24023703*
(YY Huang et al., 2013)

Genus	Ex. Species name	matK entries
Cocos	*Cocos nucifera*(Coconut)	1
Phoenix	*Phoenix dactylifera(Date Palm)*	6
Bismarckia	*Bismarckia nobilis*	1
Pseudophoenix	*Pseudophoenix lediniana*	6
Chamaedorea	*Chamaedorea elegans*	31
Elaeis	*Elaeis guineensis(*Oil Palm*)*	1
Calamus	Calamus sp.(Rattan)	42
Areca	*Areca catechu(*Betel nuts *)*	2
Metroxylon	*Metroxylon salomonense*	1

2015-11-26

4. NGS Read Alignment

- DDBJ Pipeline
- SAM/BAM format
- Visualization(Samtools tview)
- SRA100551(query)
- GU811709 (ref.)

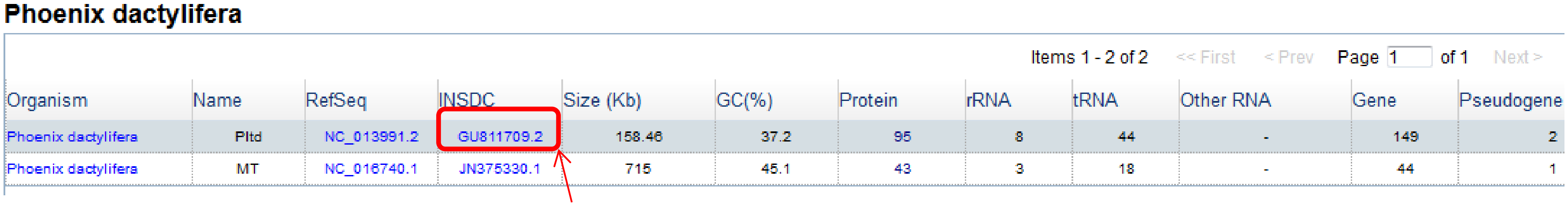
Date palm: datasets of chloroplast genome

Do not download data: the next page tool imports automatically

- Phoenix dactylifera (date palm) : taxid:42345

GU811709 (reference sequence)

http://www.ncbi.nlm.nih.gov/genome/organelles/2664?



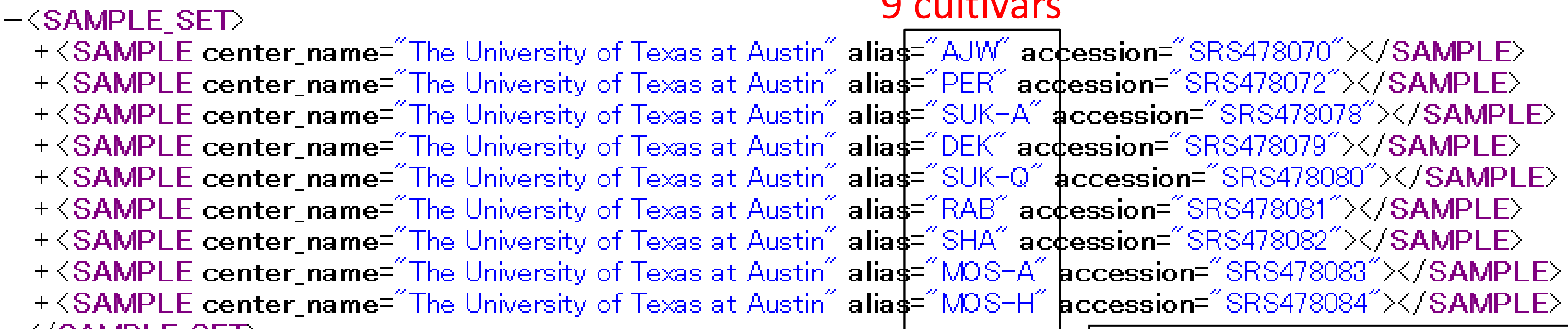
Phoenix dactylifera

Organism	Name	RefSeq	INSDC	Size (Kb)	GC(%)	Protein	rRNA	tRNA	Other RNA	Gene	Pseudogene
Phoenix dactylifera	Pltd	NC_013991.2	GU811709.2	158.46	37.2	95	8	44	-	149	2
Phoenix dactylifera	MT	NC_016740.1	JN375330.1	715	45.1	43	3	18	-	44	1

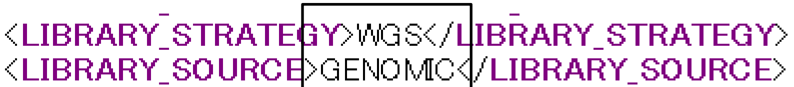
chloroplast genome

SRA100551 (query sequences)

9 cultivars



```
-<SAMPLE_SET>
  +<SAMPLE center_name="The University of Texas at Austin" alias="AJW" accession="SRS478070"></SAMPLE>
  +<SAMPLE center_name="The University of Texas at Austin" alias="PER" accession="SRS478072"></SAMPLE>
  +<SAMPLE center_name="The University of Texas at Austin" alias="SUK-A" accession="SRS478078"></SAMPLE>
  +<SAMPLE center_name="The University of Texas at Austin" alias="DEK" accession="SRS478079"></SAMPLE>
  +<SAMPLE center_name="The University of Texas at Austin" alias="SUK-Q" accession="SRS478080"></SAMPLE>
  +<SAMPLE center_name="The University of Texas at Austin" alias="RAB" accession="SRS478081"></SAMPLE>
  +<SAMPLE center_name="The University of Texas at Austin" alias="SHA" accession="SRS478082"></SAMPLE>
  +<SAMPLE center_name="The University of Texas at Austin" alias="MOS-A" accession="SRS478083"></SAMPLE>
  +<SAMPLE center_name="The University of Texas at Austin" alias="MOS-H" accession="SRS478084"></SAMPLE>
 </SAMPLE_SET>
```

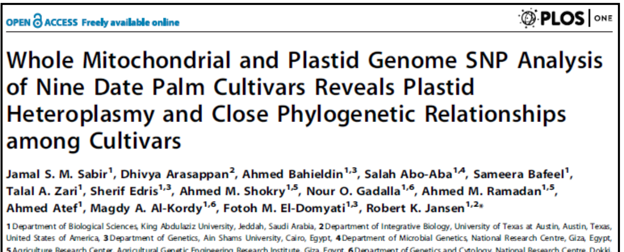


```
<LIBRARY_STRATEGY>WGS</LIBRARY_STRATEGY>
<LIBRARY_SOURCE>GENOMIC</LIBRARY_SOURCE>
```

whole genome sequencing

(Sabir et al., 2014)

PMID: 24718264



OPEN ACCESS Freely available online — PLOS | ONE

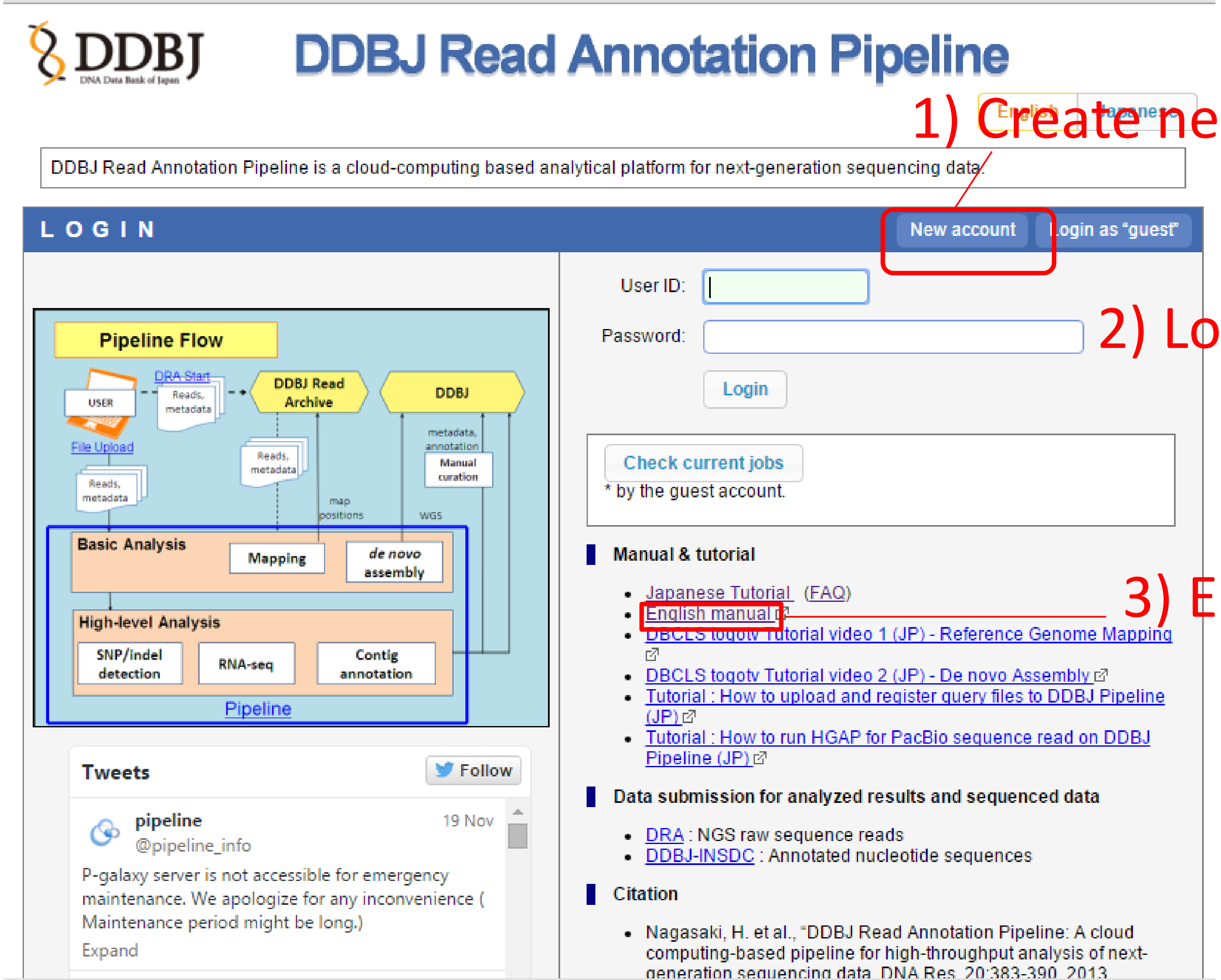
Whole Mitochondrial and Plastid Genome SNP Analysis of Nine Date Palm Cultivars Reveals Plastid Heteroplasmy and Close Phylogenetic Relationships among Cultivars

Jamal S. M. Sabir[1], Dhivya Arasappan[2], Ahmed Bahieldin[1,3], Salah Abo-Aba[1,4], Sameera Bafeel[1], Talal A. Zari[1], Sherif Edris[1,3], Ahmed M. Shokry[1,5], Nour O. Gadalla[1,6], Ahmed M. Ramadan[1,5], Ahmed Atef[1], Magdy A. Al-Kordy[1,6], Fotoh M. El-Domyati[1,3], Robert K. Jansen[1,2]*

1 Department of Biological Sciences, King Abdulaziz University, Jeddah, Saudi Arabia, 2 Department of Integrative Biology, University of Texas at Austin, Austin, Texas, United States of America, 3 Department of Genetics, Ain Shams University, Cairo, Egypt, 4 Department of Microbial Genetics, National Research Centre, Giza, Egypt, 5 Agriculture Research Center, Agricultural Genetic Engineering Research Institute, Giza, Egypt, 6 Department of Genetics and Cytology, National Research Centre, Dokki,

DDBJ pipeline : NGS read alignment

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



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 Tutorial (FAQ)
- English manual
- DBCLS togotv Tutorial video 1 (JP) - Reference Genome Mapping
- DBCLS togotv Tutorial video 2 (JP) - De novo Assembly
- Tutorial : How to upload and register query files to DDBJ Pipeline (JP)
- Tutorial : How to run HGAP for PacBio sequence read on DDBJ Pipeline (JP)

Data submission for analyzed results and sequenced data

- DRA : NGS raw sequence reads
- DDBJ-INSDC : Annotated nucleotide sequences

Citation

- Nagasaki, H. et al., "DDBJ Read Annotation Pipeline: A cloud computing-based pipeline for high-throughput analysis of next-generation sequencing data. DNA Res. 20:383-390, 2013

Tweets Follow

pipeline @pipeline_info 19 Nov

P-galaxy server is not accessible for emergency maintenance. We apologize for any inconvenience (Maintenance period might be long.)

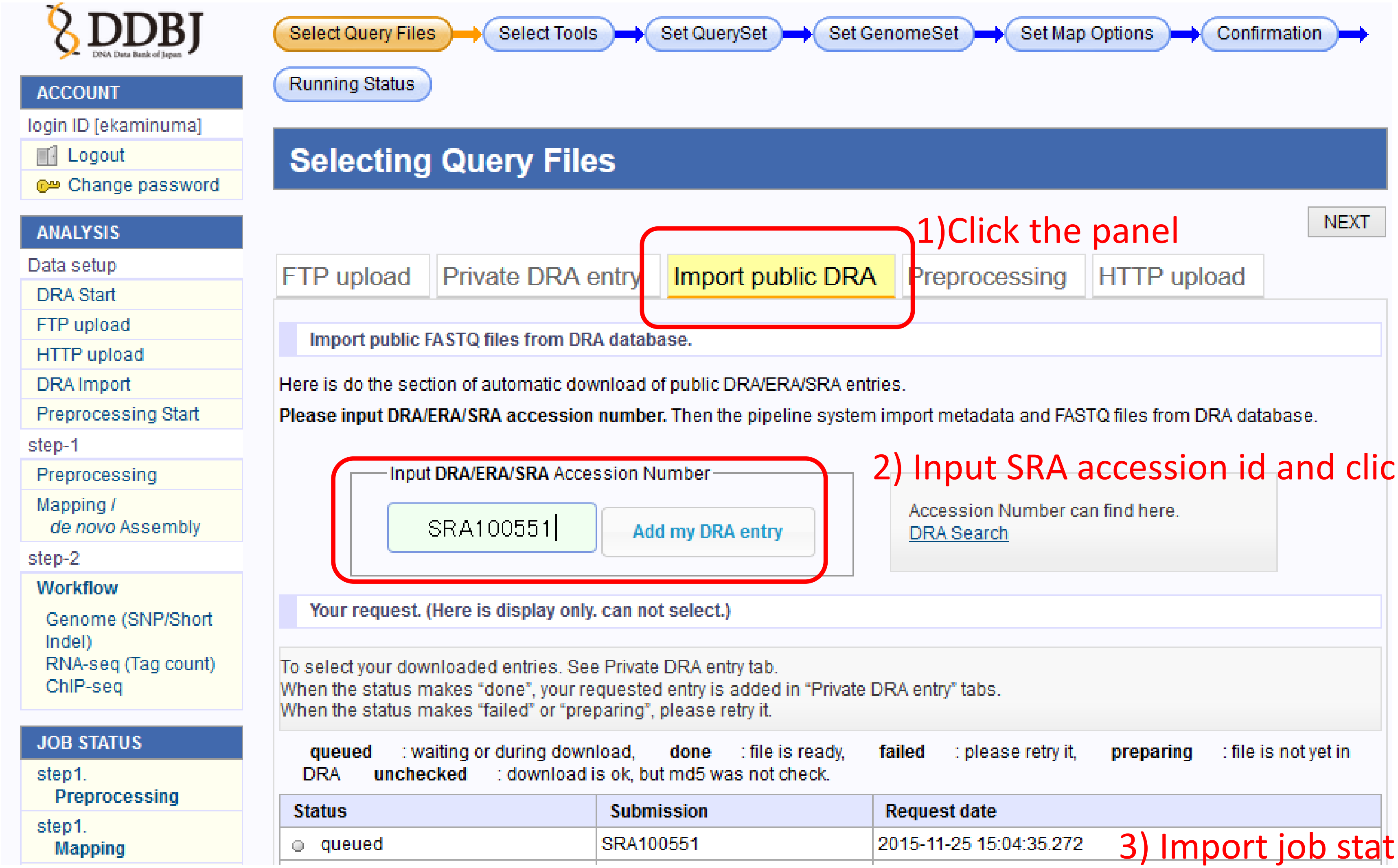
Expand

1) Create new account

2) Login

3) English manual

DDBJ pipeline 2: Import SRA data



Selecting Query Files

1)Click the panel

Import public FASTQ files from DRA database.

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.

Input **DRA/ERA/SRA** Accession Number

SRA100551

2) Input SRA accession id and click the butt

Accession Number can find here.
DRA Search

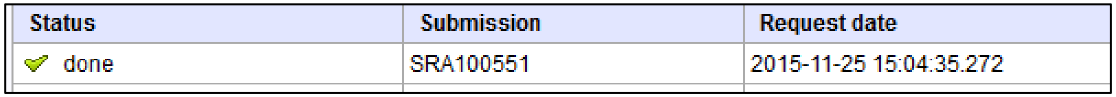
Your request. (Here is display only. can not select.)

To select your downloaded entries. See Private DRA entry tab.
When the status makes "done", your requested entry is added in "Private DRA entry" tabs.
When the status makes "failed" or "preparing", please retry it.

queued : waiting or during download, **done** : file is ready, **failed** : please retry it, **preparing** : file is not yet in DRA **unchecked** : download is ok, but md5 was not check.

Status	Submission	Request date
○ queued	SRA100551	2015-11-25 15:04:35.272

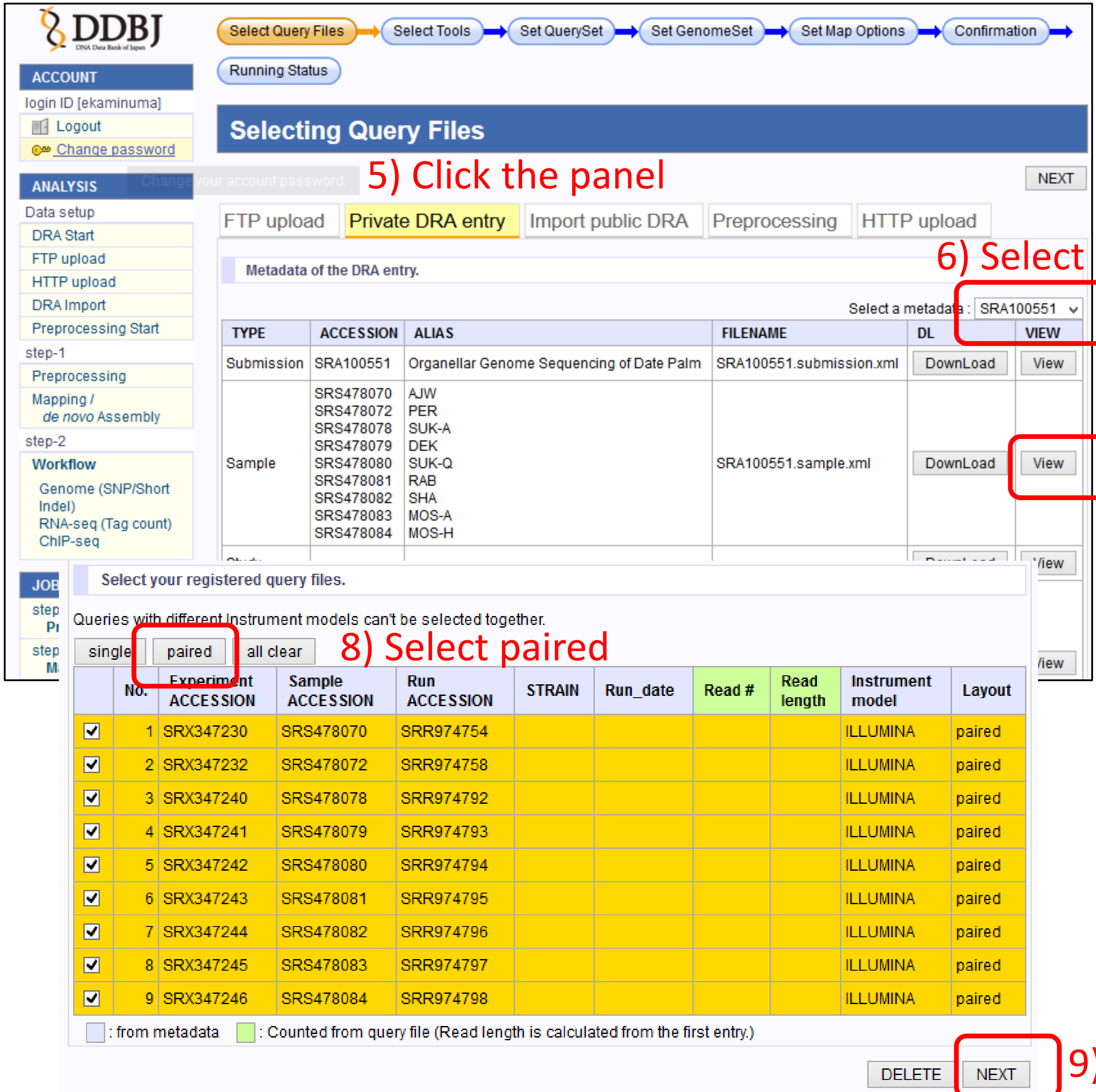
3) Import job status



Status	Submission	Request date
✓ done	SRA100551	2015-11-25 15:04:35.272

4) The pipeline will send
the e-mail notification after job completed.

DDBJ pipeline 3: Confirm SRA metadata



Selecting Query Files

5) Click the panel

6) Select "SRA100551" dataset

Metadata of the DRA entry.

Select a metadata : SRA100551

TYPE	ACCESSION	ALIAS	FILENAME	DL	VIEW
Submission	SRA100551	Organellar Genome Sequencing of Date Palm	SRA100551.submission.xml	DownLoad	View
Sample	SRS478070 SRS478072 SRS478078 SRS478079 SRS478080 SRS478081 SRS478082 SRS478083 SRS478084	AJW PER SUK-A DEK SUK-Q RAB SHA MOS-A MOS-H	SRA100551.sample.xml	DownLoad	View
Study				DownLoad	View
					View

Select your registered query files.

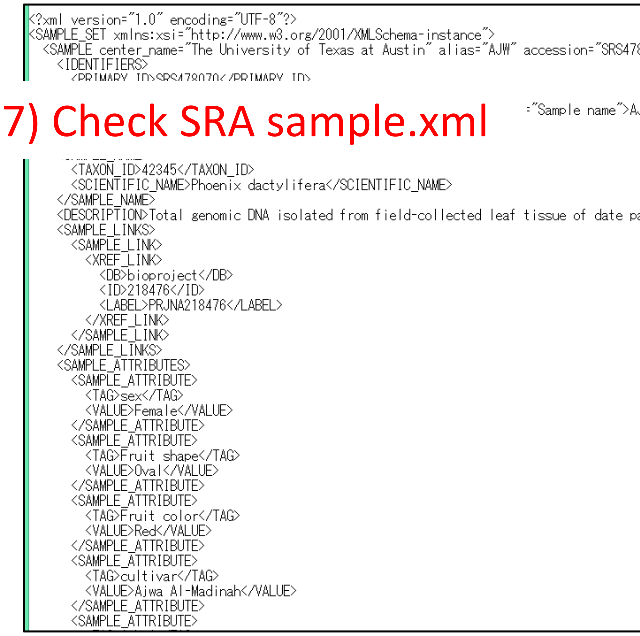
Queries with different Instrument models can't be selected together.

8) Select paired

	No.	Experiment ACCESSION	Sample ACCESSION	Run ACCESSION	STRAIN	Run_date	Read #	Read length	Instrument model	Layout
☑	1	SRX347230	SRS478070	SRR974754					ILLUMINA	paired
☑	2	SRX347232	SRS478072	SRR974758					ILLUMINA	paired
☑	3	SRX347240	SRS478078	SRR974792					ILLUMINA	paired
☑	4	SRX347241	SRS478079	SRR974793					ILLUMINA	paired
☑	5	SRX347242	SRS478080	SRR974794					ILLUMINA	paired
☑	6	SRX347243	SRS478081	SRR974795					ILLUMINA	paired
☑	7	SRX347244	SRS478082	SRR974796					ILLUMINA	paired
☑	8	SRX347245	SRS478083	SRR974797					ILLUMINA	paired
☑	9	SRX347246	SRS478084	SRR974798					ILLUMINA	paired

□ : from metadata 🟩 : Counted from query file (Read length is calculated from the first entry.)

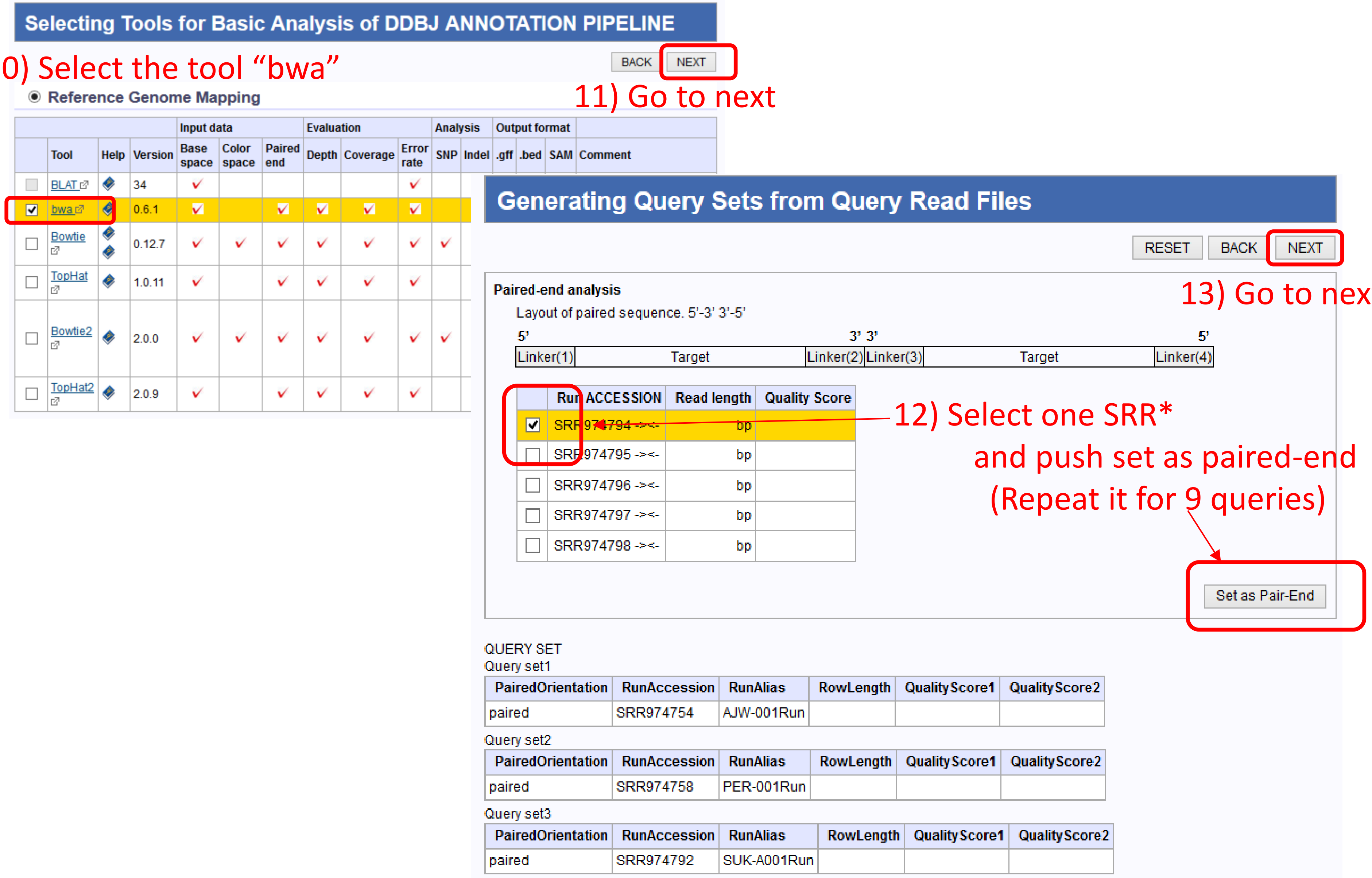
9) Go to next



7) Check SRA sample.xml

```xml
<?xml version="1.0" encoding="UTF-8"?>
<SAMPLE_SET xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance">
  <SAMPLE center_name="The University of Texas at Austin" alias="AJW" accession="SRS478
    <IDENTIFIERS>
      <PRIMARY_ID>SRS478070</PRIMARY_ID>
                                                                      ="Sample name">A
      <TAXON_ID>42345</TAXON_ID>
      <SCIENTIFIC_NAME>Phoenix dactylifera</SCIENTIFIC_NAME>
    </SAMPLE_NAME>
    <DESCRIPTION>Total genomic DNA isolated from field-collected leaf tissue of date pa
    <SAMPLE_LINKS>
      <SAMPLE_LINK>
        <XREF_LINK>
          <DB>bioproject</DB>
          <ID>218476</ID>
          <LABEL>PRJNA218476</LABEL>
        </XREF_LINK>
      </SAMPLE_LINK>
    </SAMPLE_LINKS>
    <SAMPLE_ATTRIBUTES>
      <SAMPLE_ATTRIBUTE>
        <TAG>sex</TAG>
        <VALUE>Female</VALUE>
      </SAMPLE_ATTRIBUTE>
      <SAMPLE_ATTRIBUTE>
        <TAG>Fruit shape</TAG>
        <VALUE>Oval</VALUE>
      </SAMPLE_ATTRIBUTE>
      <SAMPLE_ATTRIBUTE>
        <TAG>Fruit color</TAG>
        <VALUE>Red</VALUE>
      </SAMPLE_ATTRIBUTE>
      <SAMPLE_ATTRIBUTE>
        <TAG>cultivar</TAG>
        <VALUE>Ajwa Al-Madinah</VALUE>
      </SAMPLE_ATTRIBUTE>
      <SAMPLE_ATTRIBUTE>
```

DDBJ pipeline 4: Specify the alignment tool and generate 9 query sets



Selecting Tools for Basic Analysis of DDBJ ANNOTATION PIPELINE

10) Select the tool “bwa”

BACK NEXT

11) Go to 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	✓					✓						
☑	bwa		0.6.1	✓		✓	✓	✓	✓						
□	Bowtie		0.12.7	✓	✓	✓	✓	✓	✓	✓					
□	TopHat		1.0.11	✓		✓	✓	✓	✓						
□	Bowtie2		2.0.0	✓	✓	✓	✓	✓	✓	✓					
□	TopHat2		2.0.9	✓		✓	✓	✓	✓						

Generating Query Sets from Query Read Files

RESET BACK NEXT

13) Go to next

Paired-end analysis

Layout of paired sequence. 5'-3' 3'-5'

5'		3'	3'		5'
Linker(1)	Target	Linker(2)	Linker(3)	Target	Linker(4)

	Run ACCESSION	Read length	Quality Score
☑	SRR974794 -><-	bp	
□	SRR974795 -><-	bp	
□	SRR974796 -><-	bp	
□	SRR974797 -><-	bp	
□	SRR974798 -><-	bp	

12) Select one SRR*
and push set as paired-end
(Repeat it for 9 queries)

Set as Pair-End

QUERY SET

Query set1

PairedOrientation	RunAccession	RunAlias	RowLength	QualityScore1	QualityScore2
paired	SRR974754	AJW-001Run			

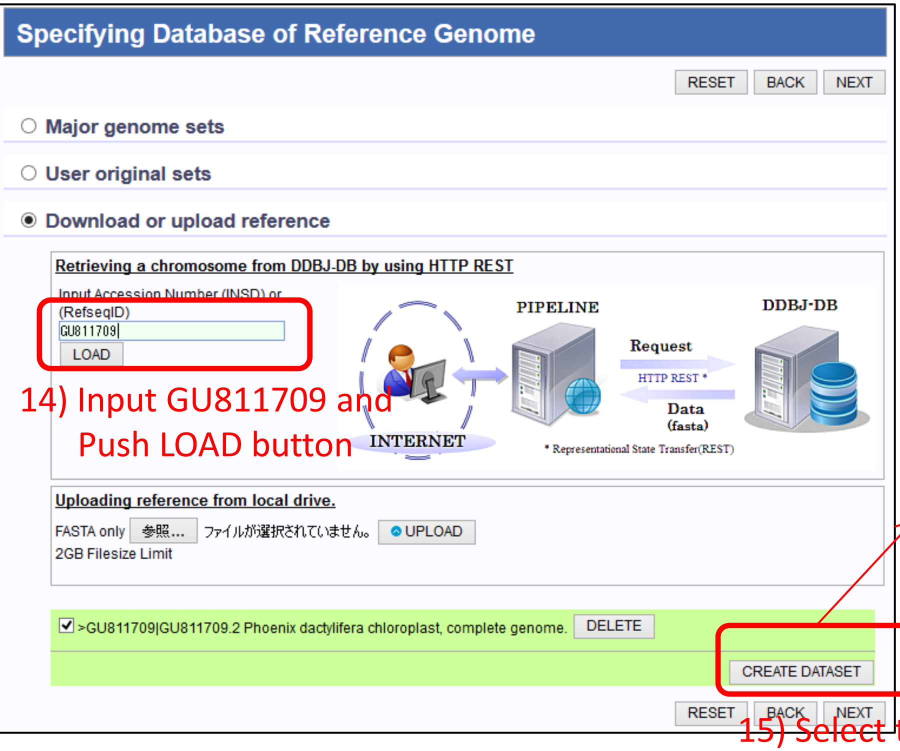
Query set2

PairedOrientation	RunAccession	RunAlias	RowLength	QualityScore1	QualityScore2
paired	SRR974758	PER-001Run			

Query set3

PairedOrientation	RunAccession	RunAlias	RowLength	QualityScore1	QualityScore2
paired	SRR974792	SUK-A001Run			

DDBJ pipeline 5: Specify the reference sequence for read alignment analysis



Specifying Database of Reference Genome

RESET BACK NEXT

○ Major genome sets

○ User original sets

● Download or upload reference

Retrieving a chromosome from DDBJ-DB by using HTTP REST

Input Accession Number (INSD) or (RefseqID)

GU811709

LOAD

14) Input GU811709 and Push LOAD button

Uploading reference from local drive.

FASTA only 参照... ファイルが選択されていません。 UPLOAD

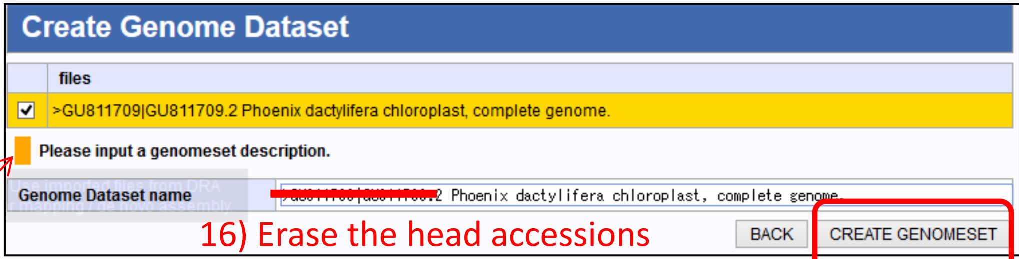
2GB Filesize Limit

☑ >GU811709|GU811709.2 Phoenix dactylifera chloroplast, complete genome. DELETE

CREATE DATASET

RESET BACK NEXT

15) Select the button



Create Genome Dataset

	files
☑	>GU811709|GU811709.2 Phoenix dactylifera chloroplast, complete genome.

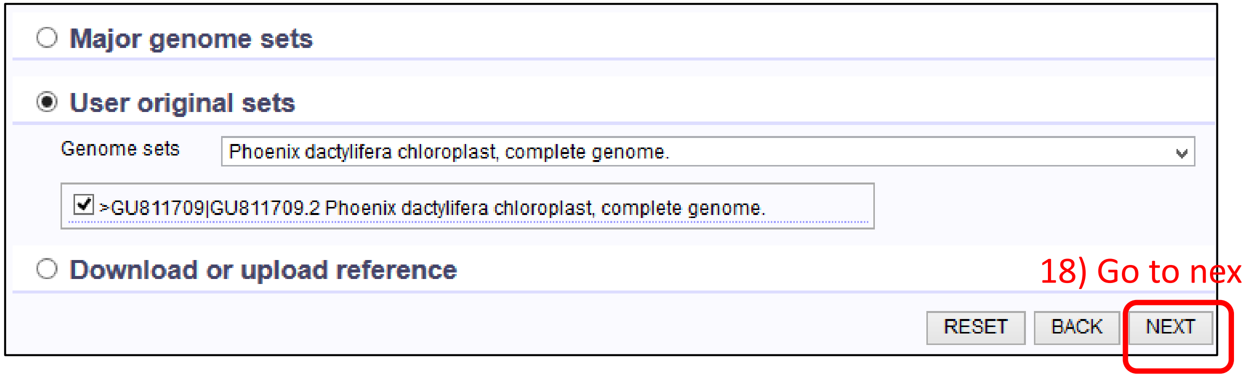
Please input a genomeset description.

Genome Dataset name: ~~>GU811709|GU811709.~~2 Phoenix dactylifera chloroplast, complete genome.

BACK CREATE GENOMESET

16) Erase the head accessions

17) Select the button



○ Major genome sets

● User original sets

Genome sets: Phoenix dactylifera chloroplast, complete genome.

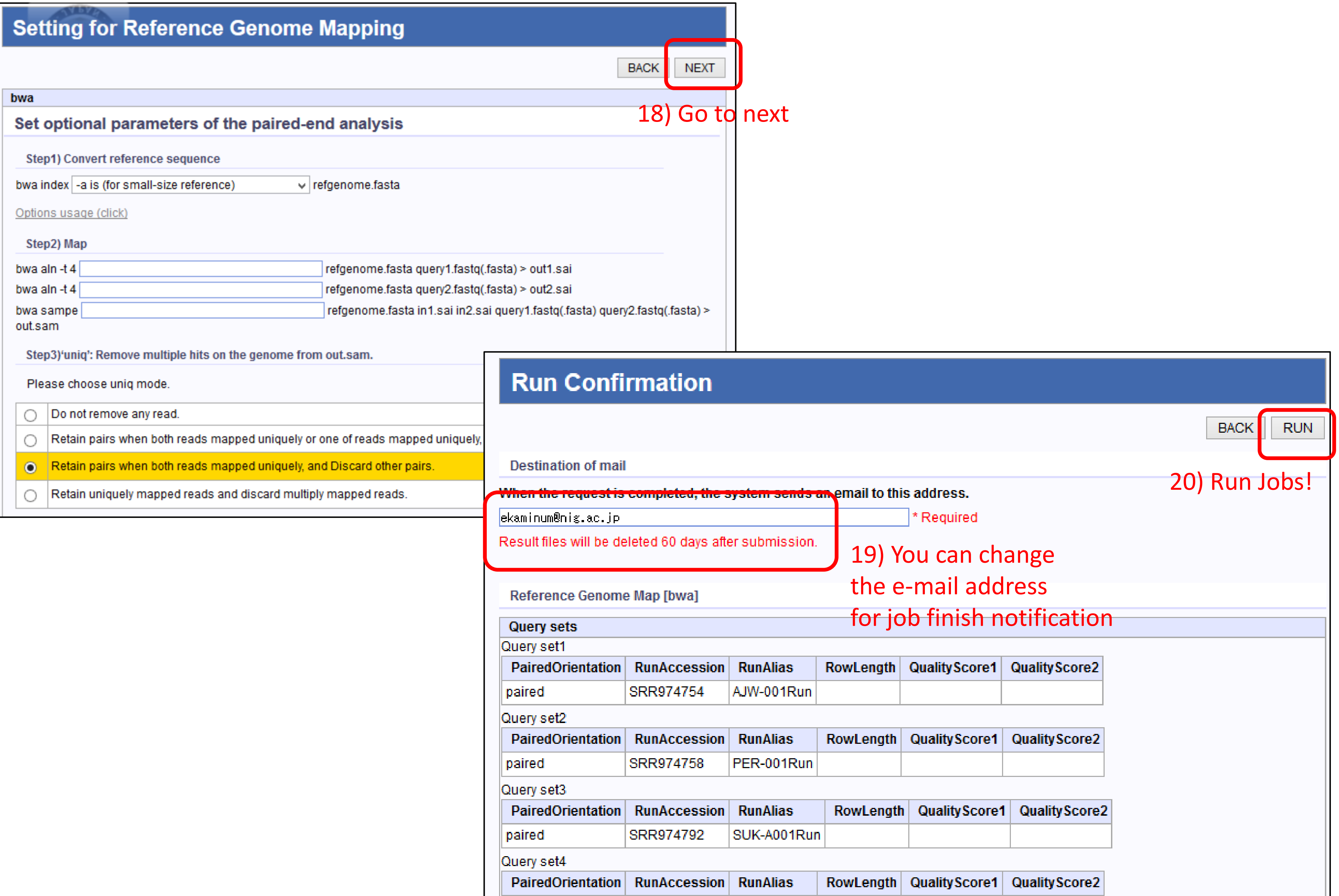
☑ >GU811709|GU811709.2 Phoenix dactylifera chloroplast, complete genome.

○ Download or upload reference

RESET BACK NEXT

18) Go to next

DDBJ pipeline 6: Set options and run all jobs



Setting for Reference Genome Mapping

18) Go to next

bwa

Set optional parameters of the paired-end analysis

Step1) Convert reference sequence

bwa index -a is (for small-size reference) refgenome.fasta

Options usage (click)

Step2) Map

bwa aln -t 4 refgenome.fasta query1.fastq(.fasta) > out1.sai

bwa aln -t 4 refgenome.fasta query2.fastq(.fasta) > out2.sai

bwa sampe refgenome.fasta in1.sai in2.sai query1.fastq(.fasta) query2.fastq(.fasta) > out.sam

Step3)'uniq': Remove multiple hits on the genome from out.sam.

Please choose uniq mode.

- ○ Do not remove any read.
- ○ Retain pairs when both reads mapped uniquely or one of reads mapped uniquely,
- ● Retain pairs when both reads mapped uniquely, and Discard other pairs.
- ○ Retain uniquely mapped reads and discard multiply mapped reads.

Run Confirmation

20) Run Jobs!

Destination of mail

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

ekaminum@nig.ac.jp * Required

Result files will be deleted 60 days after submission.

19) You can change
the e-mail address
for job finish notification

Reference Genome Map [bwa]

Query sets

Query set1

PairedOrientation	RunAccession	RunAlias	RowLength	QualityScore1	QualityScore2
paired	SRR974754	AJW-001Run			

Query set2

PairedOrientation	RunAccession	RunAlias	RowLength	QualityScore1	QualityScore2
paired	SRR974758	PER-001Run			

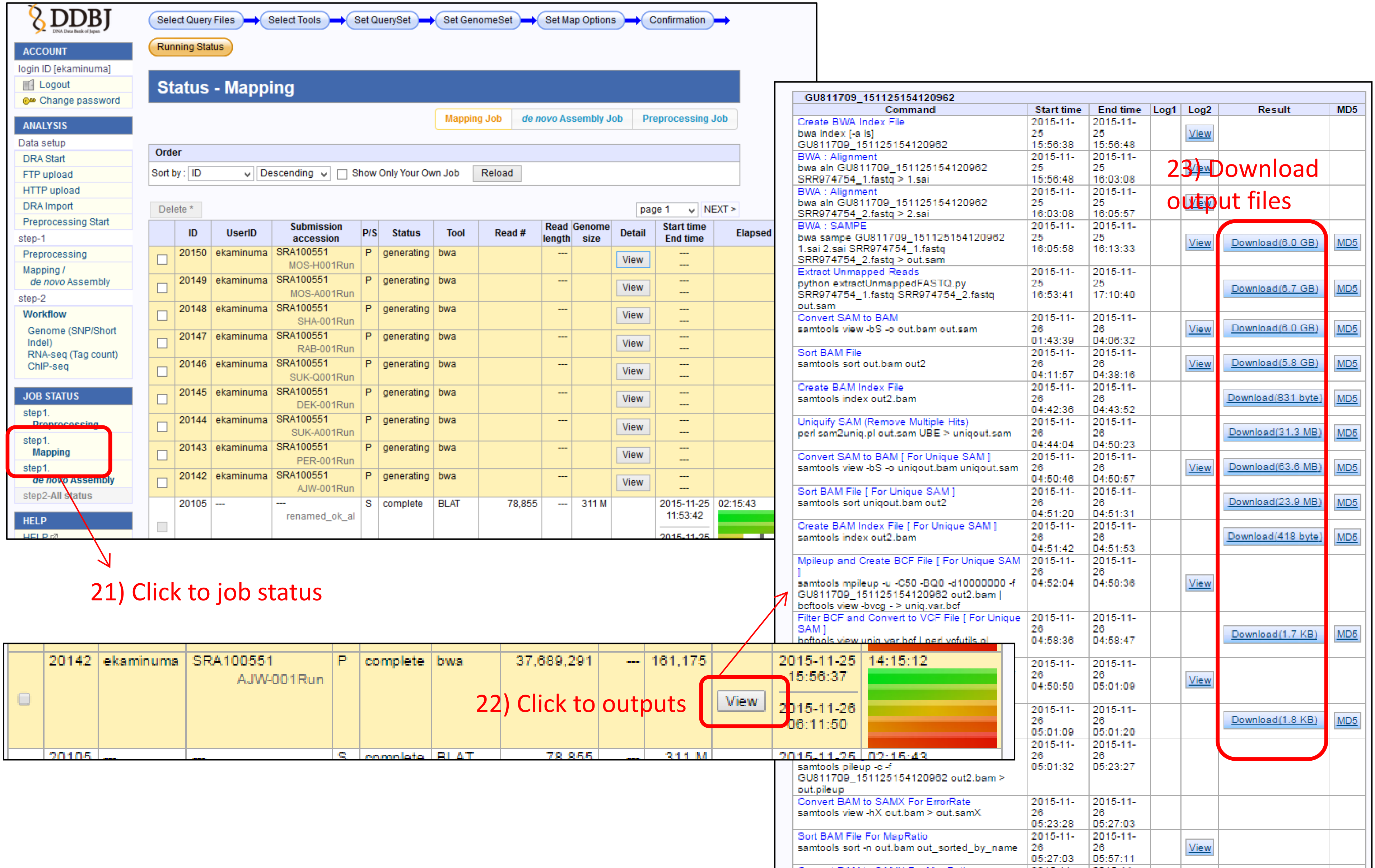
Query set3

PairedOrientation	RunAccession	RunAlias	RowLength	QualityScore1	QualityScore2
paired	SRR974792	SUK-A001Run			

Query set4

PairedOrientation	RunAccession	RunAlias	RowLength	QualityScore1	QualityScore2

DDBJ pipeline 7: Confirm job status and outputs



Status - Mapping

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	Submission accession	P/S	Status	Tool	Read #	Read length	Genome size	Detail	Start time End time	Elapsed
□	20150	ekaminuma	SRA100551 MOS-H001Run	P	generating	bwa		---		View	--- ---	
□	20149	ekaminuma	SRA100551 MOS-A001Run	P	generating	bwa		---		View	--- ---	
□	20148	ekaminuma	SRA100551 SHA-001Run	P	generating	bwa		---		View	--- ---	
□	20147	ekaminuma	SRA100551 RAB-001Run	P	generating	bwa		---		View	--- ---	
□	20146	ekaminuma	SRA100551 SUK-Q001Run	P	generating	bwa		---		View	--- ---	
□	20145	ekaminuma	SRA100551 DEK-001Run	P	generating	bwa		---		View	--- ---	
□	20144	ekaminuma	SRA100551 SUK-A001Run	P	generating	bwa		---		View	--- ---	
□	20143	ekaminuma	SRA100551 PER-001Run	P	generating	bwa		---		View	--- ---	
□	20142	ekaminuma	SRA100551 AJW-001Run	P	generating	bwa		---		View	--- ---	
□	20105	---	--- renamed_ok_al	S	complete	BLAT	78,855	---	311 M		2015-11-25 11:53:42 2015-11-25	02:15:43

21) Click to job status

□	20142	ekaminuma	SRA100551 AJW-001Run	P	complete	bwa	37,689,291	---	161,175	View	2015-11-25 15:56:37 2015-11-26 06:11:50	14:15:12
	20105	---	---	S	complete	BLAT	78,855	---	311 M		2015-11-25	02:15:43

22) Click to outputs

23) Download output files

GU811709_151125154120962

Command	Start time	End time	Log1	Log2	Result	MD5
Create BWA Index File bwa index [-a is] GU811709_151125154120962	2015-11-25 15:56:38	2015-11-25 15:56:48		View		
BWA : Alignment bwa aln GU811709_151125154120962 SRR974754_1.fastq > 1.sai	2015-11-25 15:56:48	2015-11-25 16:03:08		View		
BWA : Alignment bwa aln GU811709_151125154120962 SRR974754_2.fastq > 2.sai	2015-11-25 16:03:08	2015-11-25 16:05:57		View		
BWA : SAMPE bwa sampe GU811709_151125154120962 1.sai 2.sai SRR974754_1.fastq SRR974754_2.fastq > out.sam	2015-11-25 16:05:58	2015-11-25 16:13:33		View	Download(6.0 GB)	MD5
Extract Unmapped Reads python extractUnmappedFASTQ.py SRR974754_1.fastq SRR974754_2.fastq out.sam	2015-11-25 16:53:41	2015-11-25 17:10:40			Download(6.7 GB)	MD5
Convert SAM to BAM samtools view -bS -o out.bam out.sam	2015-11-26 01:43:39	2015-11-26 04:06:32		View	Download(6.0 GB)	MD5
Sort BAM File samtools sort out.bam out2	2015-11-26 04:11:57	2015-11-26 04:38:16		View	Download(5.8 GB)	MD5
Create BAM Index File samtools index out2.bam	2015-11-26 04:42:36	2015-11-26 04:43:52			Download(831 byte)	MD5
Uniquify SAM (Remove Multiple Hits) perl sam2uniq.pl out.sam UBE > uniqout.sam	2015-11-26 04:44:04	2015-11-26 04:50:23			Download(31.3 MB)	MD5
Convert SAM to BAM [For Unique SAM] samtools view -bS -o uniqout.bam uniqout.sam	2015-11-26 04:50:46	2015-11-26 04:50:57		View	Download(63.6 MB)	MD5
Sort BAM File [For Unique SAM] samtools sort uniqout.bam out2	2015-11-26 04:51:20	2015-11-26 04:51:31			Download(23.9 MB)	MD5
Create BAM Index File [For Unique SAM] samtools index out2.bam	2015-11-26 04:51:42	2015-11-26 04:51:53			Download(418 byte)	MD5
Mpileup and Create BCF File [For Unique SAM] samtools mpileup -u -C50 -BQ0 -d10000000 -f GU811709_151125154120962 out2.bam | bcftools view -bvcg - > uniq.var.bcf	2015-11-26 04:52:04	2015-11-26 04:58:36		View		
Filter BCF and Convert to VCF File [For Unique SAM] bcftools view uniq.var.bcf | perl vcfutils.pl	2015-11-26 04:58:36	2015-11-26 04:58:47			Download(1.7 KB)	MD5
	2015-11-26 04:58:58	2015-11-26 05:01:09		View		
	2015-11-26 05:01:09	2015-11-26 05:01:20			Download(1.8 KB)	MD5
samtools pileup -c -f GU811709_151125154120962 out2.bam > out.pileup	2015-11-26 05:01:32	2015-11-26 05:23:27				
Convert BAM to SAMX For ErrorRate samtools view -hX out.bam > out.samX	2015-11-26 05:23:28	2015-11-26 05:27:03				
Sort BAM File For MapRatio samtools sort -n out.bam out_sorted_by_name	2015-11-26 05:27:03	2015-11-26 05:57:11		View		

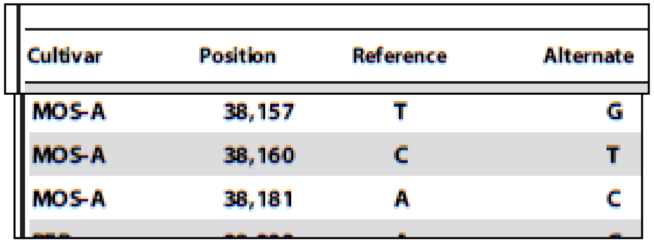
ASSIGNMENT[7]

Visualizing alignment reads using BAM files

ASSIGNMENT[7]
Confirming detected SNPs at 38,157-38,181 positions of "MOS-A cultivar" in the BAM file using "samtools" tview function and save figures of tview screenshot.

■ Samtools reference

1. http://www.htslib.org/doc/samtools.html
2. https://en.wikipedia.org/wiki/SAMtools



Cultivar	Position	Reference	Alternate
MOS-A	38,157	T	G
MOS-A	38,160	C	T
MOS-A	38,181	A	C

(Table 5. Sabir et al., 2014)

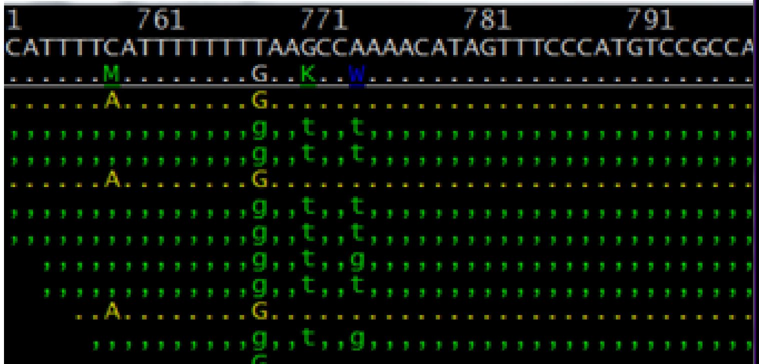
■ Commands in NIG supercomputer

```
> qlogin
>mkdir datepalm
>cp /home/kaminuma/tmp_data/SRA100551/* ~/datepalm/
>cd datepalm
>less SRR974797_uniqout.sam
>samtools tview SRR974797_out2.bam GU811709.fa
```

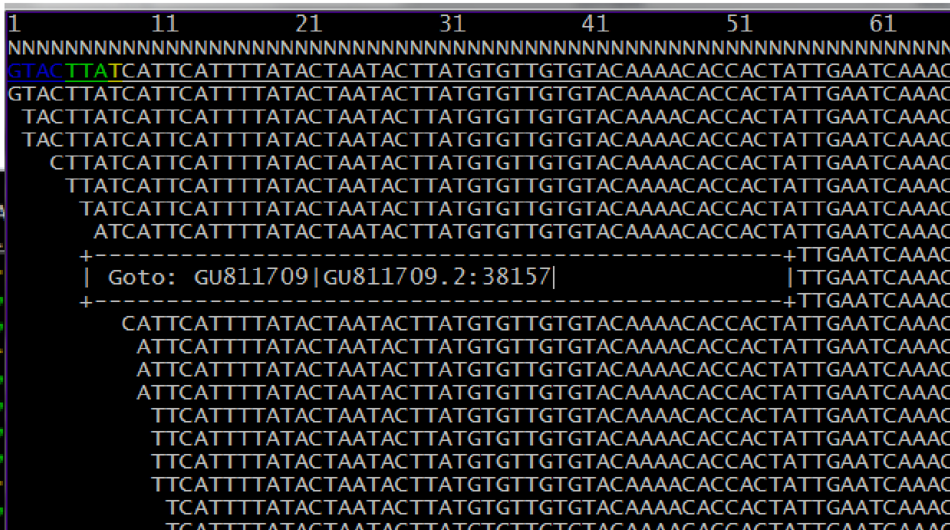
Example : tview screenshot

Reference

Query



```
1       761       771       781       791
CATTTTCATTTTTTTTAAGCCAAAACATAGTTTCCCATGTCCGCCA
......M........G..K..W........................
......A........G..............................
,,,,,,,,,,,,,,,g,,t,,t,,,,,,,,,,,,,,,,,,,,,,,,
,,,,,,,,,,,,,,,g,,t,,t,,,,,,,,,,,,,,,,,,,,,,,,
......A........G..............................
,,,,,,,,,,,,,,,g,,t,,t,,,,,,,,,,,,,,,,,,,,,,,,
,,,,,,,,,,,,,,,g,,t,,t,,,,,,,,,,,,,,,,,,,,,,,,
  ,,,,,,,,,,,,,g,,t,,g,,,,,,,,,,,,,,,,,,,,,,,,
  ,,,,,,,,,,,,,g,,t,,t,,,,,,,,,,,,,,,,,,,,,,,,
    ..A........G..............................
     ,,,,,,,,,,g,,t,,g,,,,,,,,,,,,,,,,,,,,,,,,
               G
```



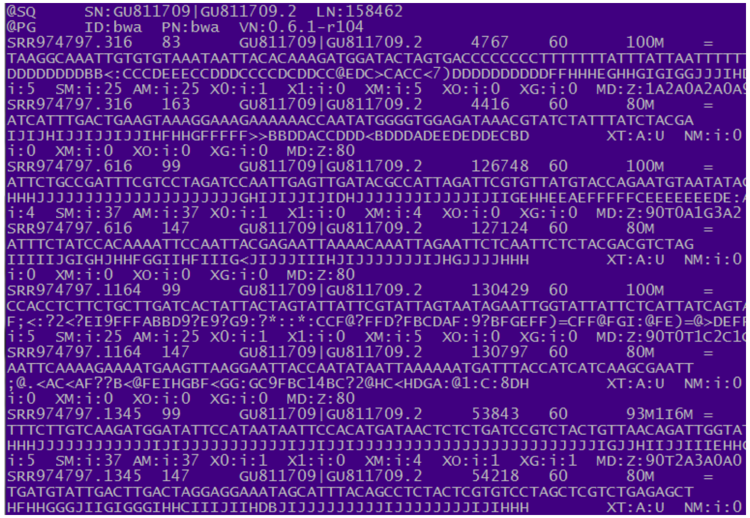
```
1         11        21        31        41        51        61
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
GTACTTATCATTCATTTTATACTAATACTTATGTGTTGTGTACAAAACACCACTATTGAATCAAAC
GTACTTATCATTCATTTTATACTAATACTTATGTGTTGTGTACAAAACACCACTATTGAATCAAAC
 TACTTATCATTCATTTTATACTAATACTTATGTGTTGTGTACAAAACACCACTATTGAATCAAAC
 TACTTATCATTCATTTTATACTAATACTTATGTGTTGTGTACAAAACACCACTATTGAATCAAAC
   CTTATCATTCATTTTATACTAATACTTATGTGTTGTGTACAAAACACCACTATTGAATCAAAC
    TTATCATTCATTTTATACTAATACTTATGTGTTGTGTACAAAACACCACTATTGAATCAAAC
     TATCATTCATTTTATACTAATACTTATGTGTTGTGTACAAAACACCACTATTGAATCAAAC
      ATCATTCATTTTATACTAATACTTATGTGTTGTGTACAAAACACCACTATTGAATCAAAC
     +------------------------------------------------+TTGAATCAAAC
     | Goto: GU811709|GU811709.2:38157                |TTGAATCAAAC
     +------------------------------------------------+TTGAATCAAAC
        CATTCATTTTATACTAATACTTATGTGTTGTGTACAAAACACCACTATTGAATCAAAC
         ATTCATTTTATACTAATACTTATGTGTTGTGTACAAAACACCACTATTGAATCAAAC
         ATTCATTTTATACTAATACTTATGTGTTGTGTACAAAACACCACTATTGAATCAAAC
         ATTCATTTTATACTAATACTTATGTGTTGTGTACAAAACACCACTATTGAATCAAAC
          TTCATTTTATACTAATACTTATGTGTTGTGTACAAAACACCACTATTGAATCAAAC
          TTCATTTTATACTAATACTTATGTGTTGTGTACAAAACACCACTATTGAATCAAAC
          TTCATTTTATACTAATACTTATGTGTTGTGTACAAAACACCACTATTGAATCAAAC
          TTCATTTTATACTAATACTTATGTGTTGTGTACAAAACACCACTATTGAATCAAAC
           TCATTTTATACTAATACTTATGTGTTGTGTACAAAACACCACTATTGAATCAAAC
           TCATTTTATACTAATACTTATGTGTTGTGTACAAAACACCACTATTGAATCAAAC
```

Reference : Alignment file format (SAM/BAM,pileup)

＜SAM/BAM format＞

■ Reference

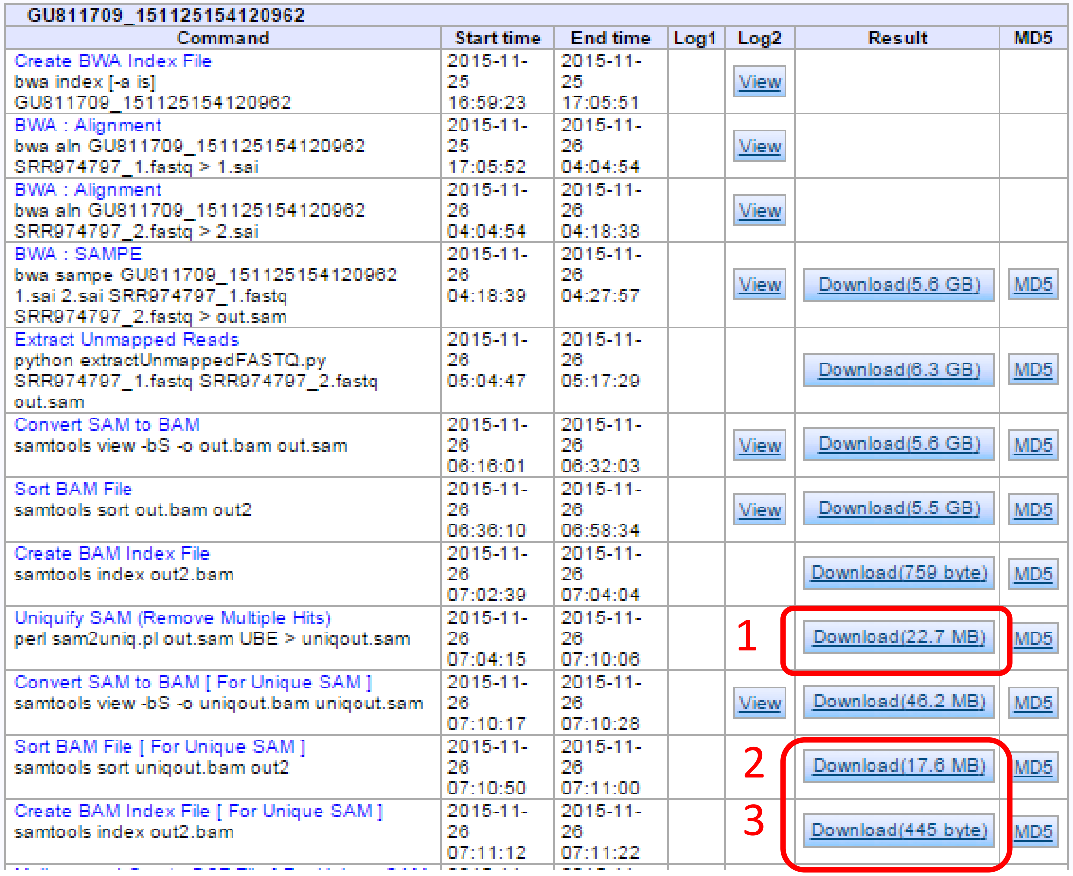
1. https://samtools.github.io/hts-specs/SAMv1.pdf
2. http://genome.sph.umich.edu/wiki/SAM



```
@SQ     SN:GU811709|GU811709.2  LN:158462
@PG     ID:bwa  PN:bwa  VN:0.6.1-r104
SRR974797.316   83      GU811709|GU811709.2     4767    60      100M    =
TAAGGCAAATTGTGTGTAAATAATTACACAAAGATGGATACTAGTGACCCCCCCCTTTTTTTATTTATTAATTTTT
DDDDDDDDBB<:CCCDEEECCDDDCCCCDCDDCC@EDC>CACC<7)DDDDDDDDDDFFHHHEGHHGIGIGGJJJIHD
i:5  SM:i:25 AM:i:25 X0:i:1  X1:i:0  XM:i:5  XO:i:0  XG:i:0  MD:Z:1A2A0A2A0A
SRR974797.316   163     GU811709|GU811709.2     4416    60      80M     =
ATCATTTGACTGAAGTAAAGGAAAGAAAAAACCAATATGGGGTGGAGATAAACGTATCTATTTATCTACGA
IJIJHIJJIJJIJJIHFHHGFFFFF>>BBDDACCDDD<BDDDADEEDEDDECBD          XT:A:U  NM:i:0
i:0  XM:i:0  XO:i:0  XG:i:0  MD:Z:80
SRR974797.616   99      GU811709|GU811709.2     126748  60      100M    =
ATTCTGCCGATTTCGTCCTAGATCCAATTGAGTTGATACGCCATTAGATTCGTGTTATGTACCAGAATGTAATATAC
HHHJJJJJJJJJJJJJJJJJJJJJJGHIJIJJIJIDHJJJJJJJIJJJJIJIIGEHHEEAEFFFFFCEEEEEEEDE:A
i:4  SM:i:37 AM:i:37 X0:i:1  X1:i:0  XM:i:4  XO:i:0  XG:i:0  MD:Z:90T0A1G3A2
SRR974797.616   147     GU811709|GU811709.2     127124  60      80M     =
ATTTCTATCCACAAAATTCCAATTACGAGAATTAAAACAAATTAGAATTCTCAATTCTCTACGACGTCTAG
IIIIIJGIGHJHHFGGIIHFIIIG<JIJJJIIIHJIJJJJJJJIJHGJJJJHHH          XT:A:U  NM:i:0
i:0  XM:i:0  XO:i:0  XG:i:0  MD:Z:80
SRR974797.1164  99      GU811709|GU811709.2     130429  60      100M    =
CCACCTCTTCTGCTTGATCACTATTACTAGTATTATTCGTATTAGTAATAGAATTGGTATTATTCTCATTATCAGTA
F;<:?2<?EI9FFFABBD9?E9?G9:?*::*:CCF@?FFD?FBCDAF:9?BFGEFF)=CFF@FGI:@FE)=@>DEFF
i:5  SM:i:25 AM:i:25 X0:i:1  X1:i:0  XM:i:5  XO:i:0  XG:i:0  MD:Z:90T0T1C2C1C
SRR974797.1164  147     GU811709|GU811709.2     130797  60      80M     =
AATTCAAAAGAAAATGAAGTTAAGGAATTACCAATATAATTAAAAAATGATTTACCATCATCAAGCGAATT
;@.<AC<AF??B<@FEIHGBF<GG:GC9FBC14BC?2@HC<HDGA:@1:C:8DH          XT:A:U  NM:i:0
i:0  XM:i:0  XO:i:0  XG:i:0  MD:Z:80
SRR974797.1345  99      GU811709|GU811709.2     53843   60      93M1I6M =
TTTCTTGTCAAGATGGATATTCCATAATAATTCCACATGATAACTCTCTGATCCGTCTACTGTTAACAGATTGGTAT
HHHJJJJJJJJJJJJJIJIJJJJJJJJJJJJJIJJIJJIJJJJJJJJJJJJJJJJJJJJJJJJJIGJJHIIJJIIIEHHG
i:5  SM:i:37 AM:i:37 X0:i:1  X1:i:0  XM:i:4  XO:i:1  XG:i:1  MD:Z:90T2A3A0A0
SRR974797.1345  147     GU811709|GU811709.2     54218   60      80M     =
TGATGTATTGACTTGACTAGGAGGAAATAGCATTTACAGCCTCTACTCGTGTCCTAGCTCGTCTGAGAGCT
HFHHGGGJIIGIGGGIHHCIIIJIIHDBJIJJJJJJJJJJIJJJJJJJJJIJIHHH          XT:A:U  NM:i:0
```

SAM format (by aligned read)
↓
[compressed]
↓
BAM format

＜DDBJ Pipeline download panel＞



GU811709_151125154120962

Command	Start time	End time	Log1	Log2	Result	MD5
Create BWA Index File bwa index [-a is] GU811709_151125154120962	2015-11-25 16:59:23	2015-11-25 17:05:51		View		
BWA : Alignment bwa aln GU811709_151125154120962 SRR974797_1.fastq > 1.sai	2015-11-25 17:05:52	2015-11-26 04:04:54		View		
BWA : Alignment bwa aln GU811709_151125154120962 SRR974797_2.fastq > 2.sai	2015-11-26 04:04:54	2015-11-26 04:18:38		View		
BWA : SAMPE bwa sampe GU811709_151125154120962 1.sai 2.sai SRR974797_1.fastq SRR974797_2.fastq > out.sam	2015-11-26 04:18:39	2015-11-26 04:27:57		View	Download(5.6 GB)	MD5
Extract Unmapped Reads python extractUnmappedFASTQ.py SRR974797_1.fastq SRR974797_2.fastq out.sam	2015-11-26 05:04:47	2015-11-26 05:17:29			Download(6.3 GB)	MD5
Convert SAM to BAM samtools view -bS -o out.bam out.sam	2015-11-26 06:16:01	2015-11-26 06:32:03		View	Download(5.6 GB)	MD5
Sort BAM File samtools sort out.bam out2	2015-11-26 06:36:10	2015-11-26 06:58:34		View	Download(5.5 GB)	MD5
Create BAM Index File samtools index out2.bam	2015-11-26 07:02:39	2015-11-26 07:04:04			Download(759 byte)	MD5
Uniquify SAM (Remove Multiple Hits) perl sam2uniq.pl out.sam UBE > uniqout.sam	2015-11-26 07:04:15	2015-11-26 07:10:06		1	Download(22.7 MB)	MD5
Convert SAM to BAM [For Unique SAM] samtools view -bS -o uniqout.bam uniqout.sam	2015-11-26 07:10:17	2015-11-26 07:10:28		View	Download(46.2 MB)	MD5
Sort BAM File [For Unique SAM] samtools sort uniqout.bam out2	2015-11-26 07:10:50	2015-11-26 07:11:00		2	Download(17.6 MB)	MD5
Create BAM Index File [For Unique SAM] samtools index out2.bam	2015-11-26 07:11:12	2015-11-26 07:11:22		3	Download(445 byte)	MD5

1. SRR974797_uniqout.sam
2. SRR974797_out2.bam
3. SRR974797_out2.bam.bai

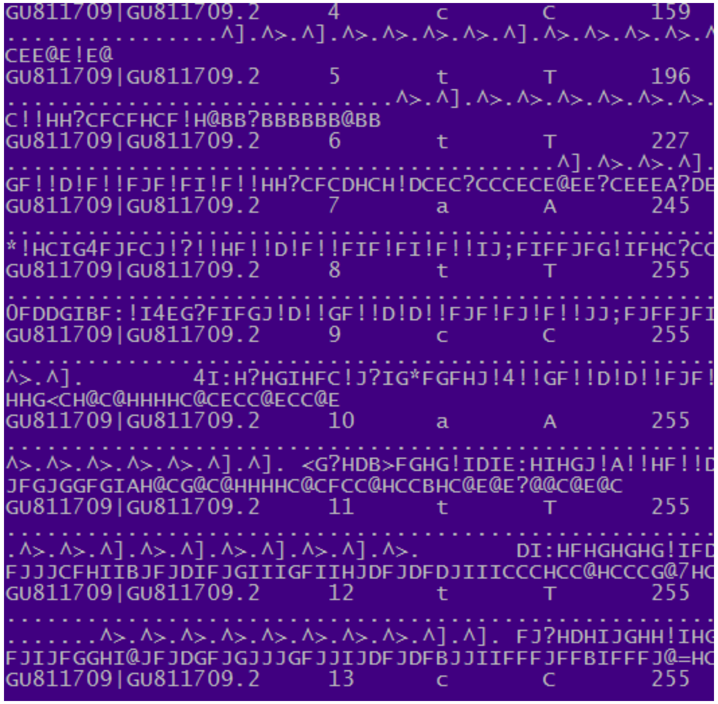
Extracting 9 cultivar sequences at psaA/psaB genes by programming

ASSIGNMENT[8]

mpileup format

<u>ls -l */SRA100551/pileup/*.pileup</u>

http://samtools.sourceforge.net/mpileup.shtml



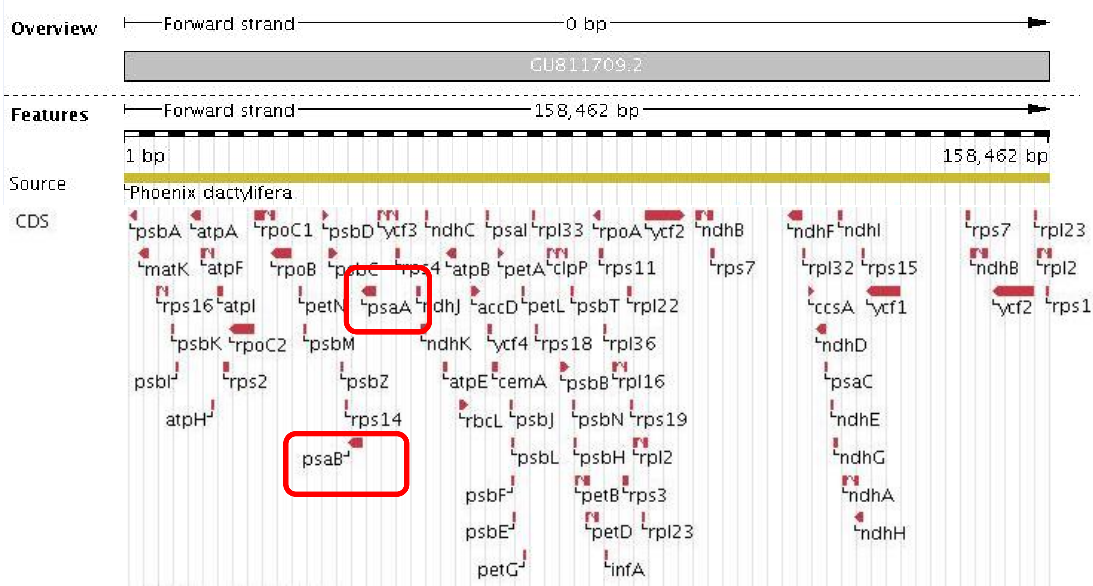
```
GU811709|GU811709.2     4       c       C       159
...................^].^>.^].^>.^>.^>.^>.^].^>.^>.^>.^>.^
CEE@E!E@
GU811709|GU811709.2     5       t       T       196
..................................^>.^].^>.^>.^>.^>.^>.^>.
C!!HH?CFCFHCF!H@BB?BBBBBB@BB
GU811709|GU811709.2     6       t       T       227
........................................^].^>.^>.^].
GF!!D!F!!FJF!FI!F!!HH?CFCDHCH!DCEC?CCCECE@EE?CEEEA?DE
GU811709|GU811709.2     7       a       A       245
.....................................................
*!HCIG4FJFCJ!?!!HF!!D!F!!FIF!FI!F!!IJ;FIFFJFG!IFHC?CC
GU811709|GU811709.2     8       t       T       255
.....................................................
0FDDGIBF:!I4EG?FIFGJ!D!!GF!!D!D!!FJF!FJ!F!!JJ;FJFFJFI
GU811709|GU811709.2     9       c       C       255
.....................................................
^>.^].          4I:H?HGIHFC!J?IG*FGFHJ!4!!GF!!D!D!!FJF!
HHG<CH@C@HHHHC@CECC@ECC@E
GU811709|GU811709.2     10      a       A       255
.....................................................
^>.^>.^>.^>.^>.^].^]. <G?HDB>FGHG!IDIE:HIHGJ!A!!HF!!D
JFGJGGFGIAH@CG@C@HHHHC@CFCC@HCCBHC@E@E?@@C@E@C
GU811709|GU811709.2     11      t       T       255
.....................................................
.^>.^>.^].^>.^].^>.^].^>.^].^>.         DI:HFHGHGHG!IFD
FJJJCFHIIBJFJDIFJGIIIGFIIHJDFJDFDJIIICCCHCC@HCCCG@7HC
GU811709|GU811709.2     12      t       T       255
.....................................................
.......^>.^>.^>.^>.^>.^>.^>.^>.^].^]. FJ?HDHIJGHH!IHG
FJIJFGGHI@JFJDGFJGJJJGFJJIJDFJDFBJJIIFFFJFFBIFFFJ@=HC
GU811709|GU811709.2     13      c       C       255
```

Genomic coordinate↑

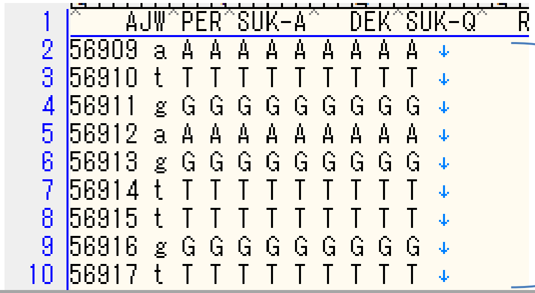
Reference base↑

<u>Query aligned base</u>↑

Ref. http://www.ebi.ac.uk/ena/data/view/GU811709



Example output : Tab-separated(TSV) file



```
^     AJW^PER^SUK-A^   DEK^SUK-Q^   R
56909 a A A A A A A A A A ↓
56910 t T T T T T T T T T ↓
56911 g G G G G G G G G G ↓
56912 a A A A A A A A A A ↓
56913 g G G G G G G G G G ↓
56914 t T T T T T T T T T ↓
56915 t T T T T T T T T T ↓
56916 g G G G G G G G G G ↓
56917 t T T T T T T T T T ↓
```

← 9 cultivar names

1st column: genomic pos.
2nd～10th column:
Aligned base by cultivars

ASSIGNMENT[8]

Extract 9 cultivar genomic sequences from <u>analyzed mpileup files</u> with psaA gene (genomic position: 40117..42369), and psaB gene (37887..40091) by programming.