Octopus_Docs Documentation

Release stable

Taemook

Apr 13, 2020

Contents

Octopus-toolkit

Note: `2018-09-05` : Octopus-toolkit supports Ubuntu18.04 version. Please feel free to contact us if you have any problems in use.

- Please cite the following paper :
  - Kim T, Seo HD, Hennighausen L, Lee D, Kang K. Octopus-toolkit: a workflow to automate mining of public epigenomic and transcriptomic next-generation sequencing data. Nucleic Acids Res. 2018 Feb 6. doi: 10.1093/nar/gky083. PubMed PMID: 29420797

<2017-04-06 17:02:12 by Prof. Keunsoo Kang, Taemook Kim in the Kangklab>

Octopus-toolkit is a stand-alone application for retrieving and processing large sets of next-generation sequencing (NGS) data with a single step. Octopus-toolkit is an automated set-up-and-analysis pipeline utilizing the Aspera, SRA Toolkit, bwtool, Samtools FastQC, Trimmomatic, HISAT2, STAR, and HOMER applications. All the applications will be installed on the user's computer when the program starts. Upon the installation, it can automatically retrieve original files (.SRA) of various data sets, including ChIP-seq, ATAC-seq, DNase-seq, MeDIP-seq, MNase-seq, and RNA-seq, from the gene expression omnibus data repository. The downloaded files can then be sequentially processed to generate BAM and BigWig files, which are used for advanced analyses and visualization. Currently, it can process NGS data from popular model genomes such as, human (Homo sapiens), mouse (Mus musculus), dog (Canis lupus familiaris), Fruit fly (Drosophila melanogaster), Zebrafish (Danio rerio), Arabidopsis (Arabidopsis thaliana), budding yeast (Saccharomyces cerevisiae), and Worm (c.elegans) genomes. With the processed files from Octopus-toolkit, the meta-analysis of various data sets, motif searches for DNA-binding proteins, and the identification of differentially expressed genes and/or protein-binding sites can be easily conducted with few commands by users. Octopus-toolkit can allow biologist and other researchers to run NGS analysis without understanding of computation behind the tools.

CHAPTER 1

Download

`Latest Version (2.2.0)`:(`Octopus-toolkit`), release 04/11/2020

- Version(`2.2.0`) is a major release with the following changes.
- Upgraded versions of some tools
- Changed the tool used to download raw data from NCBI
- Fixed the issue of Err006-1 what raw data was not downloaded.

`Latest Mac Version (2.2.0)`:(`Octopus-toolkit_mac`), release 04/11/2020

- Version(`2.2.0`) is a major release with the following changes.
- Upgraded versions of some tools
- Changed the tool used to download raw data from NCBI
- Fixed the issue of Err006-1 what raw data was not downloaded.

CHAPTER 2

Hardware/Software Requirement

Minimum Memory (RAM):

- `32Gb` memory for RNA-Seq.
- `8Gb` memory for Others (ChIP,ATAC,MNase,DNase,MeDIP)

Operating System:

- 32-64bit `Linux`, 64bit `MacOS`, 64bit `Window(Alpha Version).`

Operating System Version (tested):

- Linux : Ubuntu (`14.04`),(`16.04`, highly recommend), (`18.04`)
- Fedora (`22`),(`25`)
- Mint (`18`)
- CentOS (`7`)
- MacOS (`Sierra.10.12.6`)

CHAPTER 3

Program development

- Eclipse : Neon.1a Service Release(4.6.1)
- Language : Java Programming language (JDK1.8)
- Graphic User Interface(GUI) : Swing & Windowbuilder

CHAPTER 4

Previous version

`Latest Version (2.1.3)`, release 02/20/2019

- Version(`2.1.3`) is a minor release with the following changes.
- Optimized the analysis package for each operating system.
- Adjusted threshold of the read min length in Trimming process.

`Latest Mac Version (2.1.3)`, release 02/20/2019

- Version(`2.1.3`) is a minor release with the following changes.
- Optimized the analysis package for each operating system.
- Adjusted threshold of the read min length in Trimming process.

`Version (2.1.2)`, release 09/05/2018

- Version(`2.1.2`) is a minor release with the following changes.
- Added new operating system to support Ubuntu `18.04` version.
- Upgraded versions of R packages.
- Optimized the analysis package for each operating system.

`Mac Version (2.1.2)`, release 09/05/2018

- Version(`2.1.2`) is a minor release with the following changes.
- Upgraded versions of R packages.
- Optimized the analysis package for each operating system.

`Version (2.1.1)`, release 06/21/2018

- Version(`2.1.1`) is a minor release with the following changes.

- Changed source code (Download) of Octopus-toolkit to install Homer tool.

`Mac Version (2.1.1)`, release 06/21/2018

- Version(`2.1.1`) is a minor release with the following changes.
- Changed source code (Download) of Octopus-toolkit to install Homer tool.

`Version (2.1.0)`, release 02/13/2018

- Version(`2.1.0`) is a minor release with the following changes.
- Periodic inspection of source code
- Change the server storage and link

`Mac Version (2.1.0)`, release 02/13/2018

- Version(`2.1.0`) is a minor release with the following changes.
- Periodic inspection of source code
- Change the server storage and link

`Version (2.0.9)`, release 01/03/2018

- Version(`2.0.9`) is a minor release with the following changes.
- Optimized Paired-end classfication Method in Private Data.
- Changed source code (Private Table) of Octopus-toolkit to make maintenance easier.

`Mac Version (2.0.9)`, release 01/03/2018

- Version(`2.0.9`) is a minor release with the following changes.
- Optimized Paired-end classfication Method in Private Data.
- Changed source code (Private Table) of Octopus-toolkit to make maintenance easier.

`Version (2.0.8)`, release 12/22/2017

- Version(`2.0.8`) is a minor release with the following changes.
- Modified a list of the private table.
- Changed source code (Private Table) of Octopus-toolkit to make maintenance easier.

`Mac Version (2.0.8)`, release 12/22/2017

- Version(`2.0.8`) is a minor release with the following changes.
- Modified a list of the private table.
- Changed source code (Private Table) of Octopus-toolkit to make maintenance easier.

`Version (2.0.7)`, release 12/11/2017

- Version(`2.0.7`) is a minor release with the following changes.
- modified a parsing code because the format of the NCBI's GEO Accession display is changed.

`Mac Version (2.0.7)`, release 12/11/2017

- Version(`2.0.7`) is a minor release with the following changes.
- modified a parsing code because the format of the NCBI's GEO Accession display is changed.

`Version (2.0.6)`, release 11/28/2017

- Version(`2.0.6`) is a minor release with the following changes.
- Changed a method which to obtain modified url of the raw data in NCBI.
- Display Microarray in unsupported list (Err004-2) (NULL -> Microarray)

`Mac Version (2.0.6)`, release 11/28/2017

- Version(`2.0.5`) is a minor release with the following changes.
- Changed a method which to obtain modified url of the raw data in NCBI.
- Display Microarray in unsupported list (Err004-2) (NULL -> Microarray)

`Version (2.0.5)`, release 11/08/2017

- Version(`2.0.5`) is a minor release with the following changes.
- Periodic inspection of source code.

`Mac Version (2.0.5)`, release 11/08/2017

- Version(`2.0.5`) is a minor release with the following changes.
- Periodic inspection of source code.

`Version (2.0.4)`, release 11/07/2017

- Version(`2.0.4`) is a minor release with the following changes.
- Applied the modified url of raw data in GEO Dataset. (Issue : changed FTP path of SRA experiment data in NCBI)

`Mac Version (2.0.4)`, release 11/07/2017

- Version(`2.0.4`) is a minor release with the following changes.
- Applied the modified url of raw data in GEO Dataset. (Issue : changed FTP path of SRA experiment data in NCBI)

`Version (2.0.3)`, release 10/23/2017

- Version(`2.0.3`) is a minor release with the following changes.
- Updated `CentOS` version
- Optimized Mapping process.
- Changed source code of Octopus-toolkit to make maintenance easier.

`Mac Version (2.0.3)`, release 10/23/2017

- Version(`2.0.3`) is a minor release with the following changes.
- Optimized Mapping process.

- Changed source code of Octopus-toolkit to make maintenance easier.

`Version (2.0.2)` , release 09/09/2017

- Version(`2.0.2`) is a minor release with the following changes.
- Modified a Full parameter Option.
- Changed source code of Octopus-toolkit to make maintenance easier.

`Mac Version (2.0.2)`, release 09/09/2017

`Version (2.0.1)`, release 08/23/2017

- Version(`2.0.1`) is a minor release with the following changes.
- Change 3rd party tools to be installed without password.
- When analysis is completed/failed, notify the path of results in a terminal window.
- Changed source code of Octopus-toolkit to make maintenance easier.
- Changed User Interface(UI) of the installation progressbar.
- Added Ubuntu 14.04, Refer to *"How to install libraries"*, *"How to install R"*
- Added the `Tutorial` about how to use a custom adapter sequence generated by ownself and how to discover de novo and known motif using the output file of Octopus-toolkit.

`Beta Version (2.0.0)`, release 07/29/2017

4.1 Table of contents

4.1.1 0.Quick Start

To use the Octopus-toolkit right away, please follow these tutorials:

0-1. Installation Movie Clip

Tutorial for installation. (Youtube)

0-2. Ubuntu(16.04), Mint(18) (We highly recommend to use Ubuntu)

- Commands (`Quick_Start(Ubuntu,mint).txt`):

```
sudo apt-get update
sudo apt-get install openjdk-8-jdk
sudo apt-get install zlib1g-dev libpng12-dev libncurses5-dev build-essential r-
↪base
wget http://octopus-toolkit2.readthedocs.io/en/latest/_downloads/Octopus-toolkit.
↪zip -O Octopus-toolkit.zip
unzip Octopus-toolkit.zip
cd Octopus-toolkit/
java -jar Octopus-toolkit.jar
```

0-3. Fedora(25)

- Commands (`Quick_Start(Fedora).txt`):

```
sudo yum update
sudo yum install java-1.8.0-openjdk
sudo yum install zlib-devel.x86_64 libpng-devel.x86_64 libpng12-devel.x86_64␣
↪ncurses-devel.x86_64 gcc-c++ bzip2-devel xz-devel R
wget http://octopus-toolkit2.readthedocs.io/en/latest/_downloads/Octopus-toolkit.
↪zip -O Octopus-toolkit.zip
unzip Octopus-toolkit.zip
cd Octopus-toolkit/
java -jar Octopus-toolkit.jar
```

0-4. CentOS(7)

- Commands (`Quick_Start(CentOS).txt`):

```
sudo yum update
sudo yum install java-1.8.0-openjdk
sudo yum install zlib-devel.x86_64 libpng-devel.x86_64 libpng12-devel.x86_64␣
↪ncurses-devel.x86_64 gcc-c++ bzip2-devel xz-devel
sudo yum install epel-release
sudo yum install R
wget http://octopus-toolkit2.readthedocs.io/en/latest/_downloads/Octopus-toolkit.
↪zip -O Octopus-toolkit.zip
unzip Octopus-toolkit.zip
cd Octopus-toolkit/
java -jar Octopus-toolkit.jar
```

0-5. MacOS(Sierra 10.12.6)

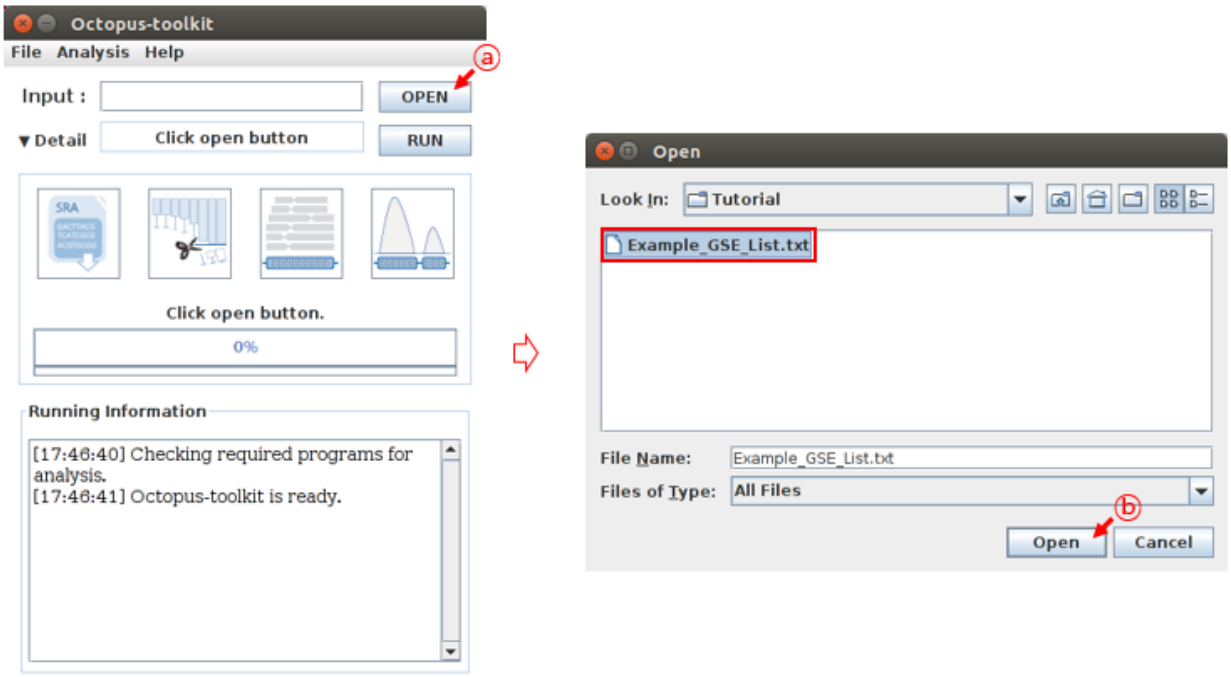
Note:

Please refer to the link below for MacOS Link : *1.Installation 1-6.MacOS(Sierra_10.12.6)*

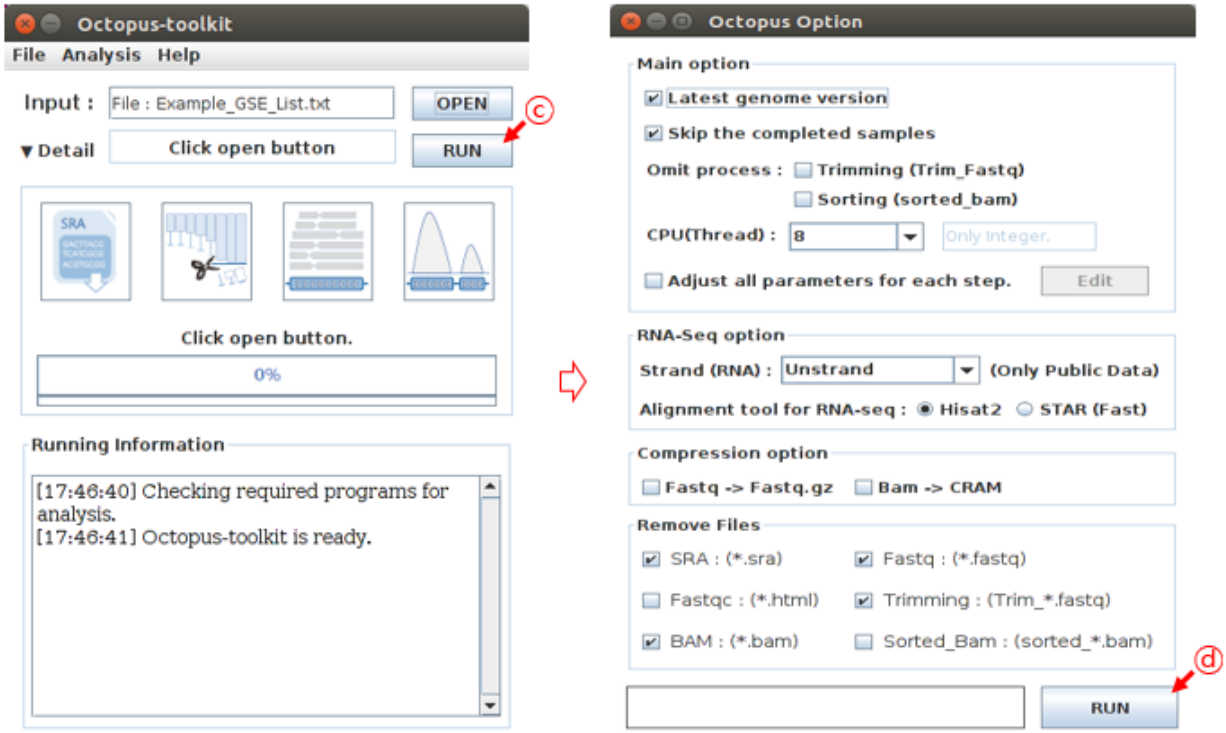
0-6. Quick Run (Public data)

Example GSE file (`Example_GSE_List.txt`)

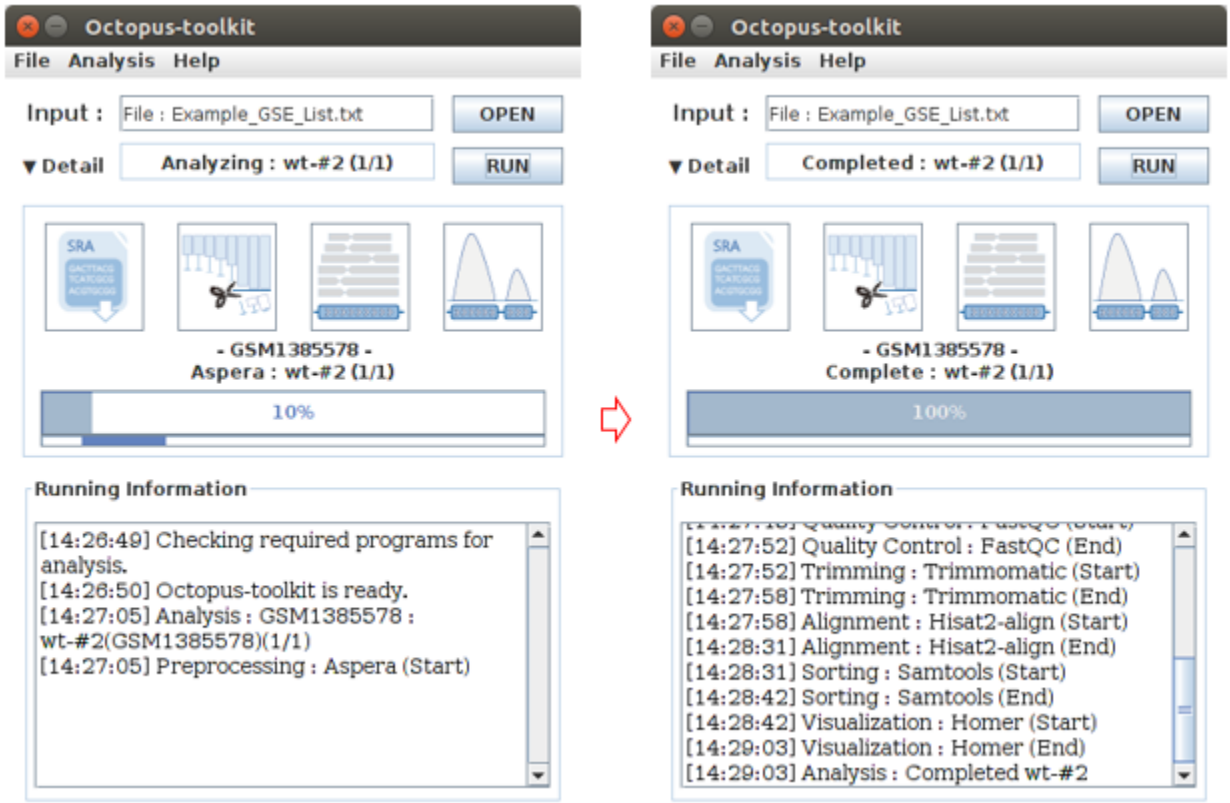
- `A` : Click the OPEN button.
- `B` : Select the `Example_GSE_List.txt` file.



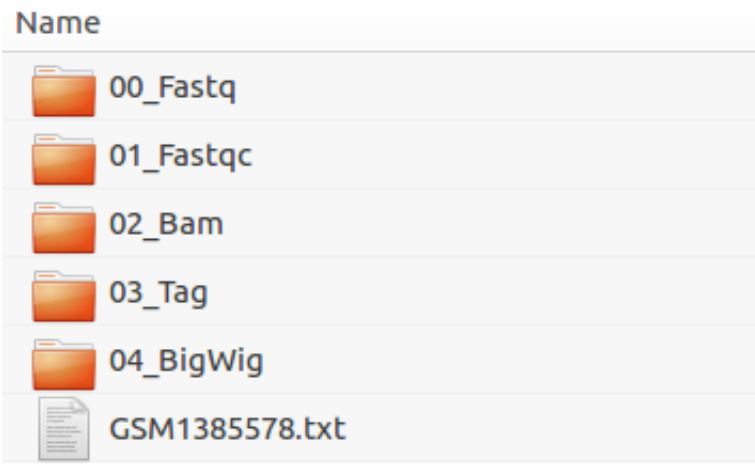
- C : Click the RUN button.
- D : Set parameters. Then, click the RUN button.



- Octopus-toolkit will sequentialy download and process the data specified in the list file. The analysis may take some time.



- Below shows output file of Octopus-toolkit



4.1.2 1.Installation

1-1.Download

Note: Applications in the `Requirement` section must be installed on your computer before running the Octopus-toolkit.

1. Octopus-toolkit_2.2.0 : (`Octopus-toolkit`) (64bit)
2. Octopus-toolkit_mac_2.2.0 : (`Octopus-toolkit_mac`)

1-2.Installation Movie Clip

- Tutorial for installation
- Ubuntu `16.04` version. (Youtube - Ubuntu 16.04)
- CentOS `6.9` version. (Youtube - CentOS 6.9) - Currently not supported.
- CentOS `7` version. (Youtube - CentOS 7)
- MacOS `Sierra 10.12.6` version (Youtube - MacOS Sierra 10.12.6)

1-3.Requirement

To run the Octopus-toolkit, `Java 8` (JDK, Java Development ToolKit) or higher, must be installed on your computer.

- Ubuntu, Mint (`Ubuntu 16.04 or Mint18`):

```
sudo apt-get update
sudo apt-get install openjdk-8-jdk
```

- Ubuntu (`14.04`):

```
sudo add-apt-repository ppa:openjdk-r/ppa
sudo apt-get update
sudo apt-get install openjdk-8-jdk
sudo update-alternatives --config java
sudo update-alternatives --config javac
```

- Fedora, CentOS(`Fedora 22~25 or CentOS 7`):

```
sudo yum update
sudo yum install java-1.8.0-openjdk
```

Octopus-toolkit utilizes several libraries for analysis. Each operating system such as ubuntu, mint and fedora differ in ways to install the applications. Please follow the installation guide below.

1-4.Ubuntu(14.04,16.04), Mint(18)

To run the Octopus-toolkit, you must install the following libraries: `zlib1g`, `libpng12`, `libncurses5`, `g++`, `liblzma`, `libbz2`

- `zlib1g-dev`

```
sudo apt-get install zlib1g-dev
```

- `libncurses5-dev`

```
sudo apt-get install libncurses5-dev
```

- `g++`

```
sudo apt-get install build-essential
```

- `liblzma-dev`

```
sudo apt-get install liblzma-dev
```

- `libbz2-dev`

```
sudo apt-get install libbz2-dev
```

OR

```
sudo apt-get install zlib1g-dev libpng12-dev libncurses5-dev build-essential liblzma-
↪dev libbz2-dev
```

In the Ubuntu version (`18.04`)

- `libpng-dev`

```
sudo apt-get install libpng-dev
```

Another Ubuntu version (`14.04``16.04`), Mint (`18`)

- `libpng12-dev`

```
sudo apt-get install libpng12-dev
```

1-5.Fedora(22~25), CentOS(7)

To run the Octopus-toolkit, you must install the following libraries: `zlib`, `libpng` , `libpng12`, `ncurses`, `gcc-c++`, `libbz2`, `liblzma`

- `zlib-devel`

```
sudo yum install zlib-devel.x86_64
```

- `libpng-devel`

```
sudo yum install libpng-devel.x86_64
```

- `libpng-devel12`

```
sudo yum install libpng12-devel.x86_64
```

- `ncurses-devel`

```
sudo yum install ncurses-devel.x86_64
```

- `gcc-c++`

```
sudo yum install gcc-c++
```

- `libbz2`

```
sudo yum install bzip2-devel
```

- `liblzma`

```
sudo yum install xz-devel
```

OR

```
sudo yum install zlib-devel.x86_64 libpng-devel.x86_64 libpng12-devel.x86_64 ncurses-
↪devel.x86_64 gcc-c++ bzip2-devel xz-devel
```

1-6.MacOS(Sierra_10.12.6)

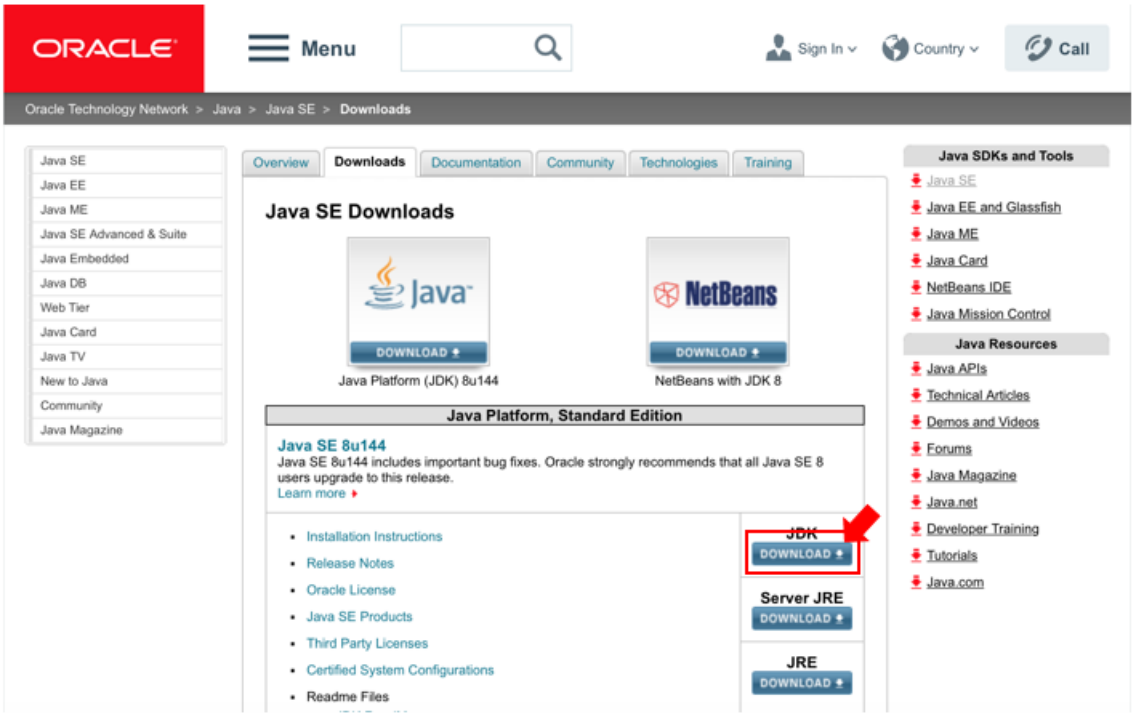
Note: Applications in the `Requirement` section must be installed on your computer before running the Octopus-toolkit (Mac version).

To run the Octopus-toolkit, `Java 8` (JDK, Java Development ToolKit) or higher, must be installed on your computer. (`Octopus-toolkit_mac_2.1.3`)

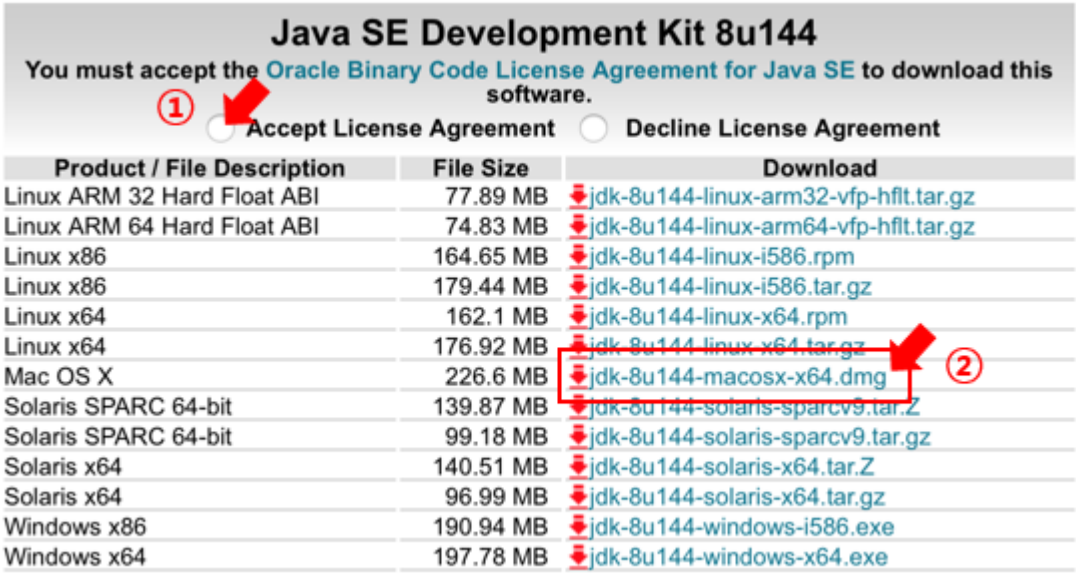
- `Java 8` (JDK, Java Development ToolKit) or higher:

```
Link : http://www.oracle.com/technetwork/java/javase/downloads/index.html
```

0. Click the JDK DOWNLOAD Button



1. Click the Accept License Agreement radio button.
2. Click the jdk-(version)-macosx-x64.dmg



3. Go to the Download folder. Execute the downloaded installation file.
4. Double click on icon to install.

Octopus-toolkit utilizes several libraries for analysis. Please follow the installation guide below.

- Xcode Update:

```
xcode-select --install
```

- Library (wget, liblzma,libpng):

```
/usr/bin/ruby -e "$(curl -fsSL https://raw.githubusercontent.com/Homebrew/install/
↪master/install)"
brew install wget
brew install xz
brew install libpng
```

1-7.R (3.1)

To draw heatmap and Line plot, `R (3.1)` or higher version of R must be installed on your computer.

- Ubuntu, Mint (`Ubuntu 16.04 or Mint18`):

```
sudo apt-get install r-base
```

- Ubuntu (`14.04`):

```
sudo apt-get update
sudo apt-get install r-base
sudo apt-key adv --keyserver keyserver.ubuntu.com --recv-keys E084DAB9
sudo add-apt-repository ppa:marutter/rdev
sudo apt-get update
sudo apt-get upgrade
sudo apt-get install r-base
```

- Fedora (`Fedora 22~25`):

```
sudo yum install R
```

- CentOS (`CentOS 7`):

```
sudo yum install epel-release
sudo yum install R
```

- MacOS (`Sierra`):

```
/usr/bin/ruby -e "$(curl -fsSL https://raw.githubusercontent.com/Homebrew/install/
↪master/install)"
brew install r
```

4.1.3 2.Run

Note: `Requirements` must be installed on a computer before running the Octopus-toolkit. (*Installation*)

2-1.How to run the Octopus-toolkit

Please follow the movie clip (Youtube)

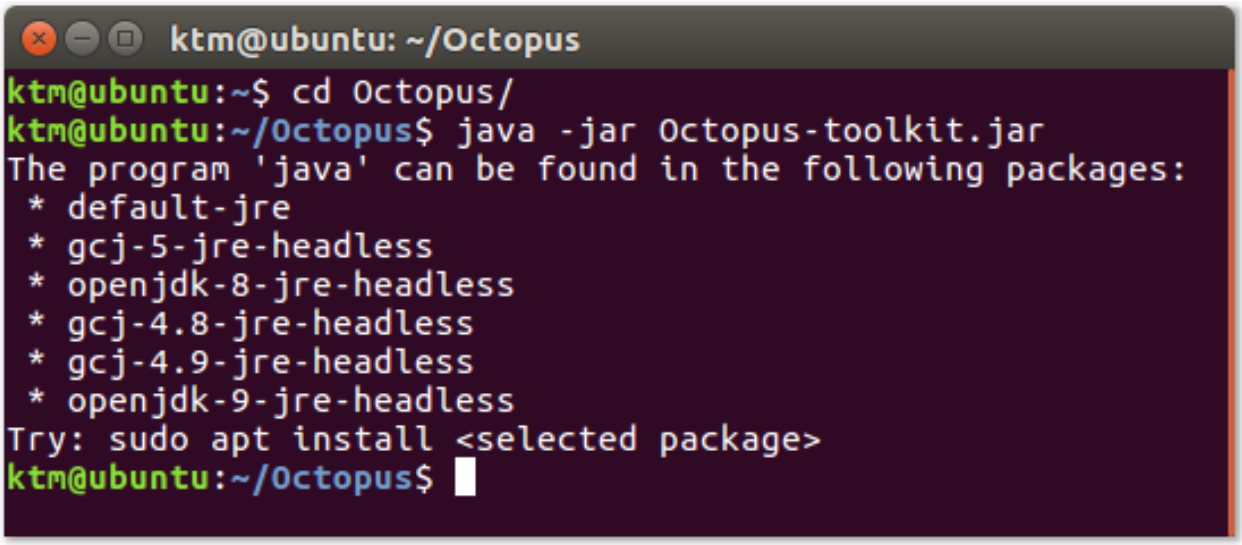
- Download (`Octopus-toolkit`) and uncompress it to the `folder` where you want it to be installed.
- Open the terminal application (cmd) and type the below command in

```
cd Octopus-toolkit/
java -jar Octopus-toolkit.jar
```

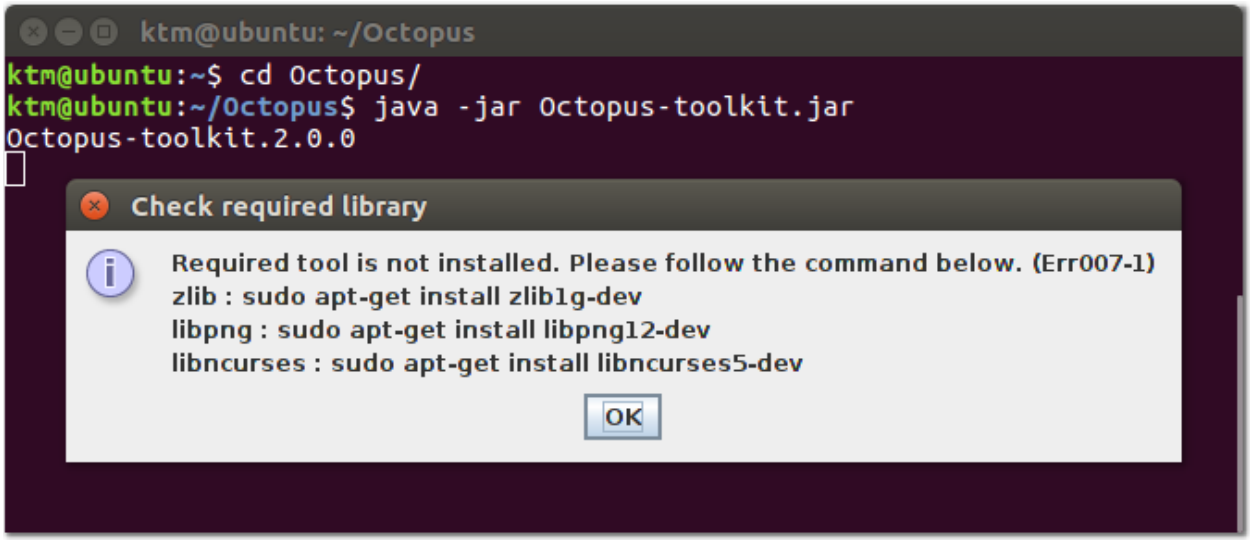
Or Command (Download ~ Run)

```
wget http://octopus-toolkit2.readthedocs.io/en/latest/_downloads/Octopus-toolkit.zip -
↪O Octopus-toolkit.zip
unzip Octopus-toolkit.zip
cd Octopus-toolkit/
java -jar Octopus-toolkit.jar
```

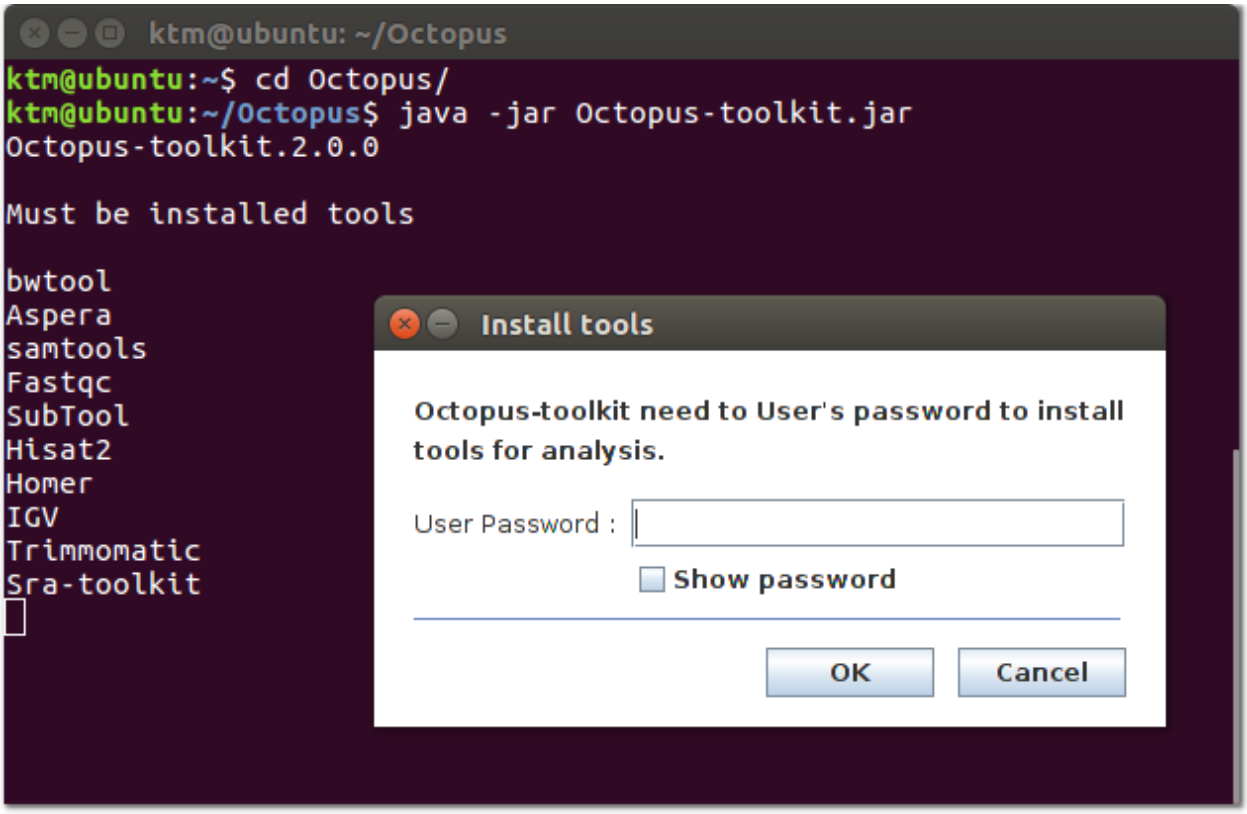
- No Java installation could be found. (Java 8 version) : (*Err007-1*)



- No prerequisite were found. (Libraries in system) : (*Err007-1*)



- Password for permission



2-2.Java Virtual Machine(VM) heap memory limited

Octopus-toolkit requires at least 8 Gb of the memory (32 Gb of memory for processing human/mouse RNA-seq) (Recommend : `32+` Gb RAM).

If you get errors related to running out of memory, please increase the heap memory of the Java Virtual Machine.

- If your memory is less than 16Gb:

```
java -jar Octopus-toolkit.jar -Xms2G -Xmx16G -XX:MaxPermSize=16G -XX:PermSize=2G
```

Argument	Description	Recommend
`Xms`	Initial Heap Size	`2Gb`
`Xmx`	Max Heap Size	Maximum of RAM
`XX:PermSize`	Initial Permanent Size	`2Gb`
`XX:MaxPermSize`	Max Permanent Size	Maximum of RAM

4.1.4 3.Octopus-toolkit output directory

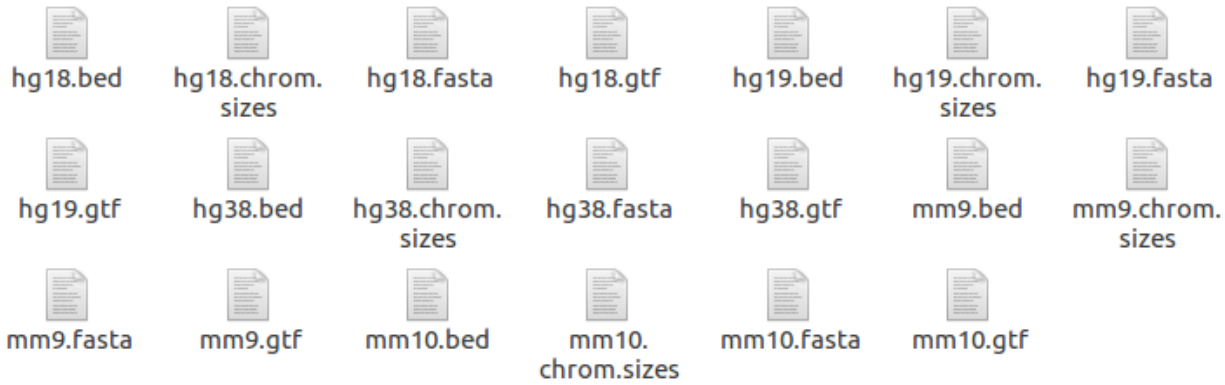
Octopus-toolkit creates five directories when you run the program.

- Octopus-toolkit main directory.

Main Folder	Sub Folder	Description
`Index`	Reference, Hisat2, STAR	Reference genome sequence and annotation files for analysis and alignment tools.
`Log`	Command, Run	Log file containing the commands used for analysis.
`Result`	GSE_Folder, P_Folder	The output files.
`Script`		Scripts used for analysis.
`Tools`	Analysis tools	Store the 3rd party tools used by Octopus-toolkit.

3-1.Index-Reference

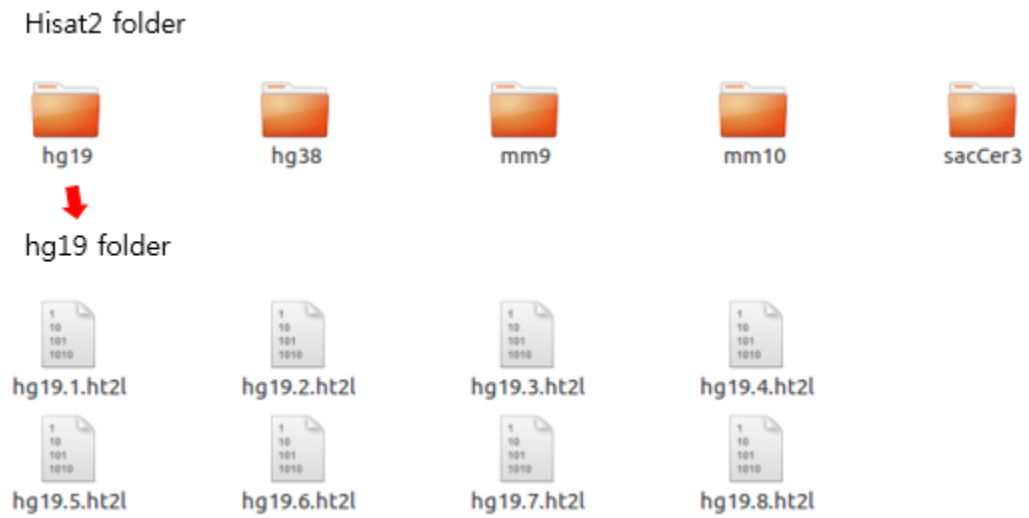
- `Reference folder`



The reference folder contains several reference files required for analysis.

Before starting each process, Octopus-toolkit checks the folder whether the reference files are prepared. If not, it automatically prepares the files.

3-2.Index-Hisat2

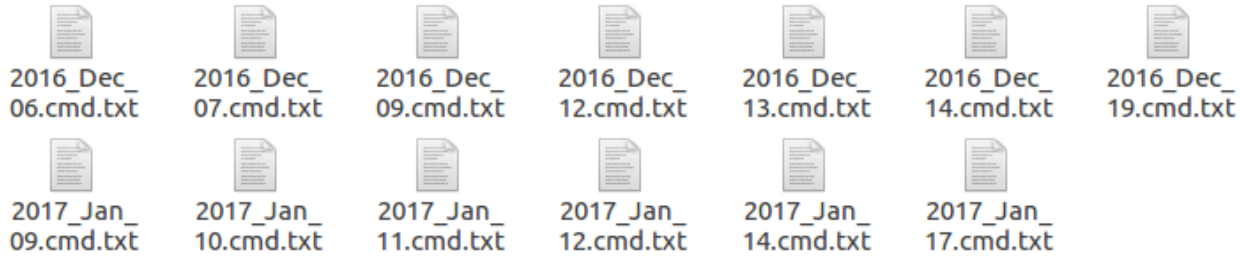


The reference genome sequence file should be indexed at least once before proceeding to the alignment step. The folder contains indexed genome sequence files used by the Hisat2 tool.

Octopus-toolkit inspects the index file of the genome before running the alignment process and runs the indexing step if it does not exist.

3-3.Log-Command

- `Command folder`

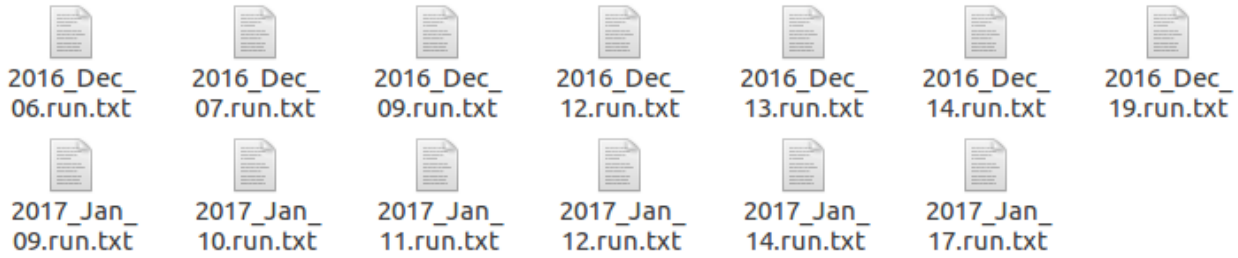


The Command directory contains log files containing the commands used during the analysis.

The file name adopts the date it is created. (`2016_Dec_06.cmd.txt`)

3-4.Log-Run

- `Run folder`

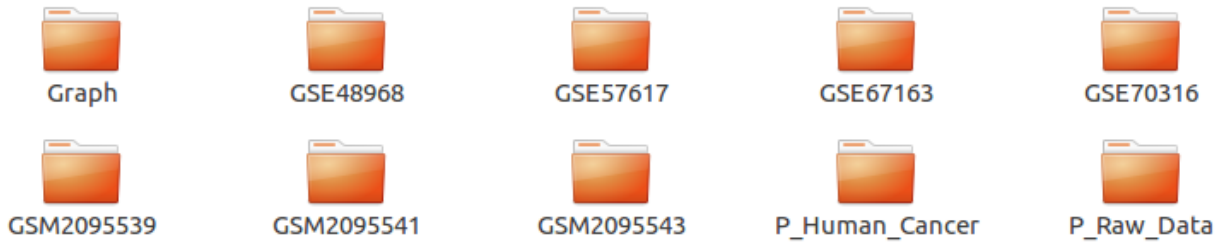


The Run directory contains log files containing running information recorded during analysis.

The file name adopts the date it is created. (`2016_Dec_06.run.txt`)

3-5.Result

- `Result folder`



The Result folder stores the output of Octopus-toolkit.

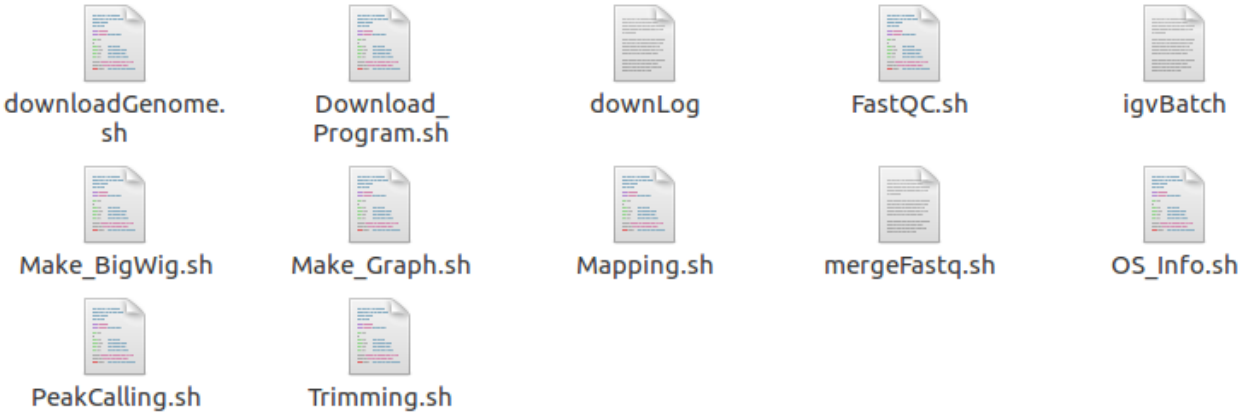
The folder name is based on the GEO accession number you entered. For the private data, the folder name begins with *P_*.

The Graph folder stores Heatmaps and Lineplots when you run the Graph function.

The detailed information regarding the output can be founded : *Output Link*

3-6.Script

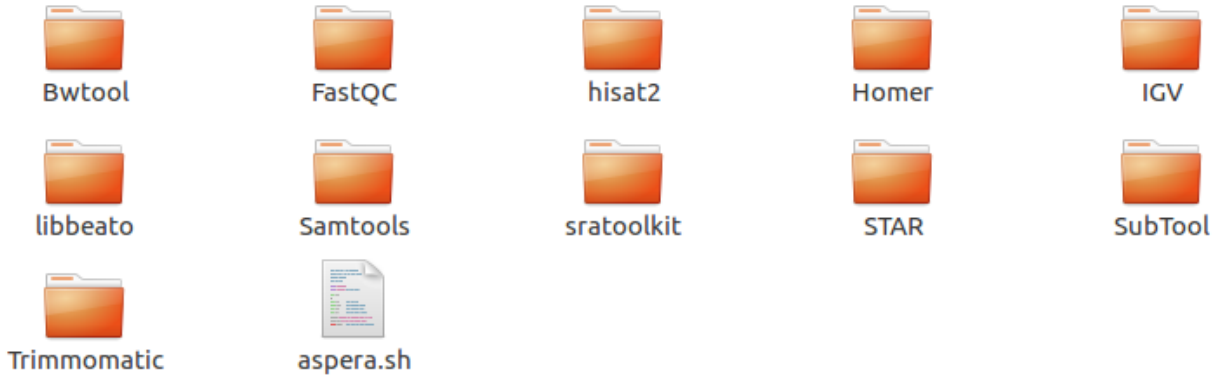
- `Script folder`



The Script folder stores the script files used by Octopus-toolkit.

3-7.Tools

- `Tools folder`

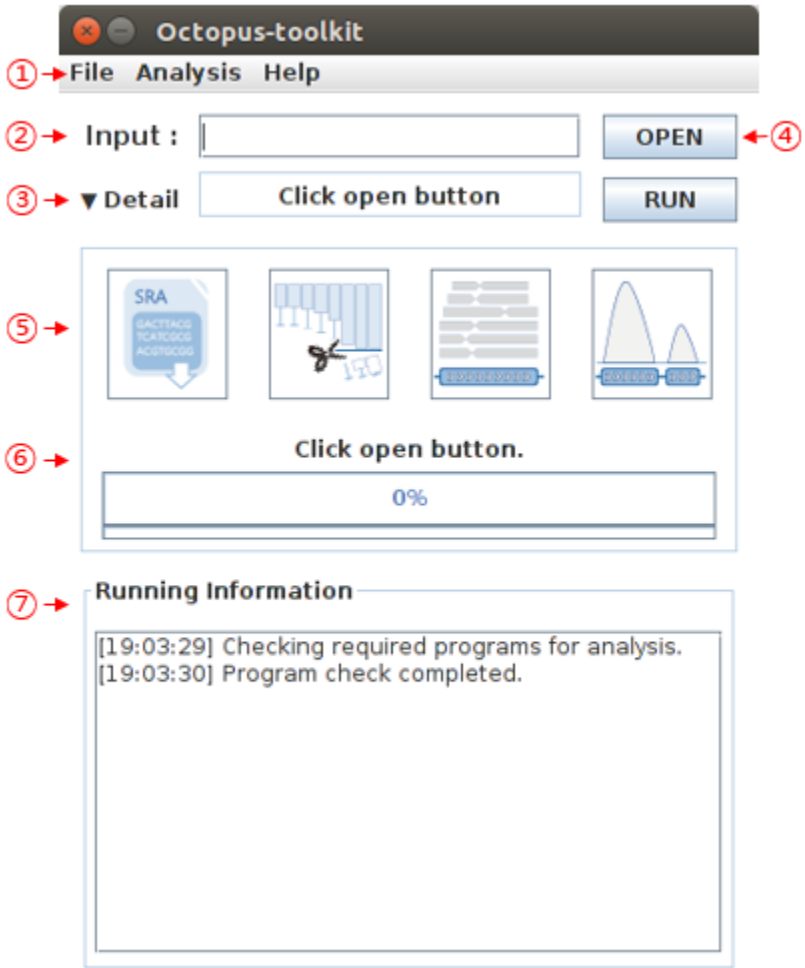


The Tools directory contains binary versions of 3rd party tools used in the Octopus-toolkit.

4.1.5 4.User Interface

4-1.Main UI

- The image below is the main UI of Octopus-toolkit



- The description of each part are as follows.

No	Name	Description
1	`Menu Bar`	Functions such as Private analysis (your data).
2	`Input`	Input GEO accession number (GSE or GSM) or a text file containing GEO accession numbers (one GSE or GSM per line).
3	`Status`	brief information regarding steps and errors.
4	`Open and Run`	Run the anlaysis.
5	`Full parameters`	Setting up the parameters for each tool.
6	`Progress bar`	Progress bar.
7	`Running information`	Status window

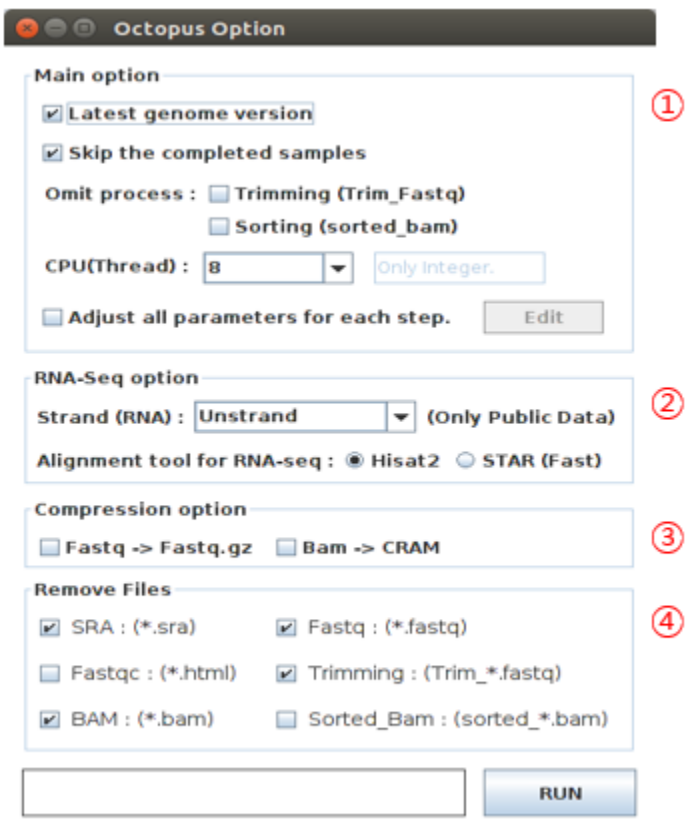
4-2.Menu Bar

- Below are the details.

Menu	Sub Menu	Description
File	`Exit`	Exit the Octopus-toolkit
Analysis	`Private Data`	Analyze your data
	`Peak Calling`	Find peaks (enriched regions) for ChIP-seq MNase-seq MeDIP-seq and ATAC-seq with HOMER.
	`Graph`	Draw the Heatmap and Line Plot from the output.
	`IGV`	Visualization using IGV (Integrative Genomics Viewer)
Help	`Manual(Tutorial)`	Go to the Octopus-toolkit manual page.
	`Error Code`	Go to the Octopus-toolkit *Error code* page.
	`Homepage`	Go to the Octopus-toolkit homepage

4-3.Octopus Option

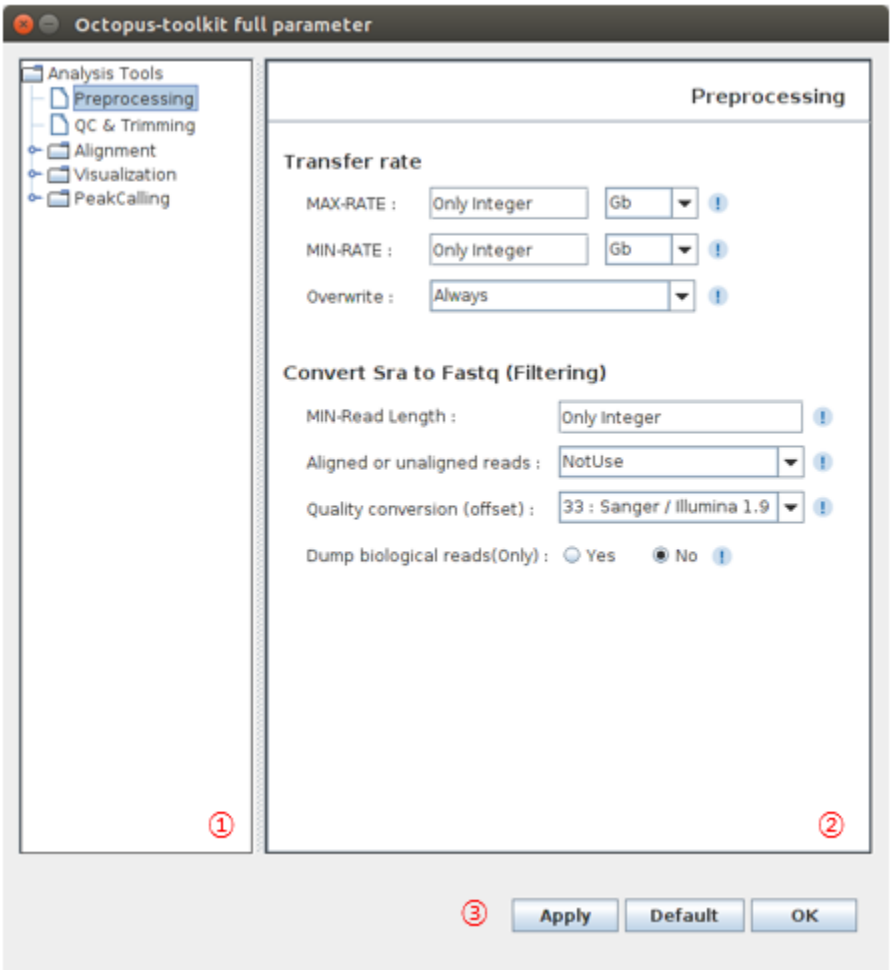
- Octopus-toolkit options



No	Name	Description
1	`Main option`	Main options.
2	`RNA-Seq option`	Options for RNA-seq data only.
3	`Compression option`	To save disk space.
4	`Remove Files`	Delete selected intermediate files after each process.

4-4.Full parameters

- The following image shows Full parameters window.

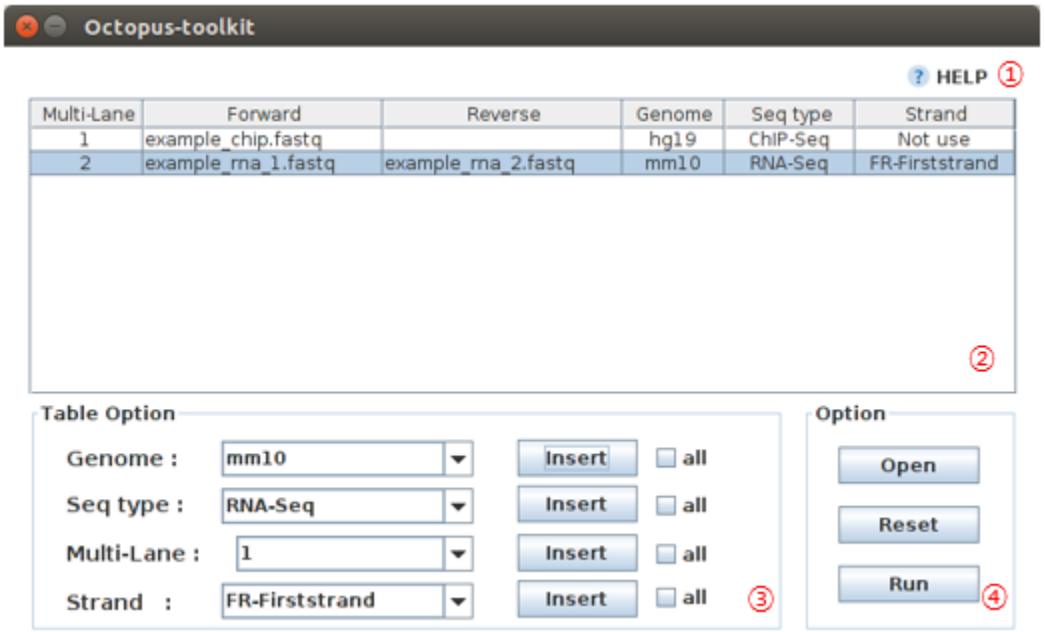


No	Name	Description
1	`Analysis tree`	Select one of steps
2	`Paremeter window`	Change parameters for the process you selected
3	`Apply`	You can apply or reset the option.

4-5.Private Table

You can analyze your own data (Fastq) using Octopus-toolkit (Analysis - Private Data). The private Table is a setup window for your data.

To analyze your own data, you must select appropriate information as follows.

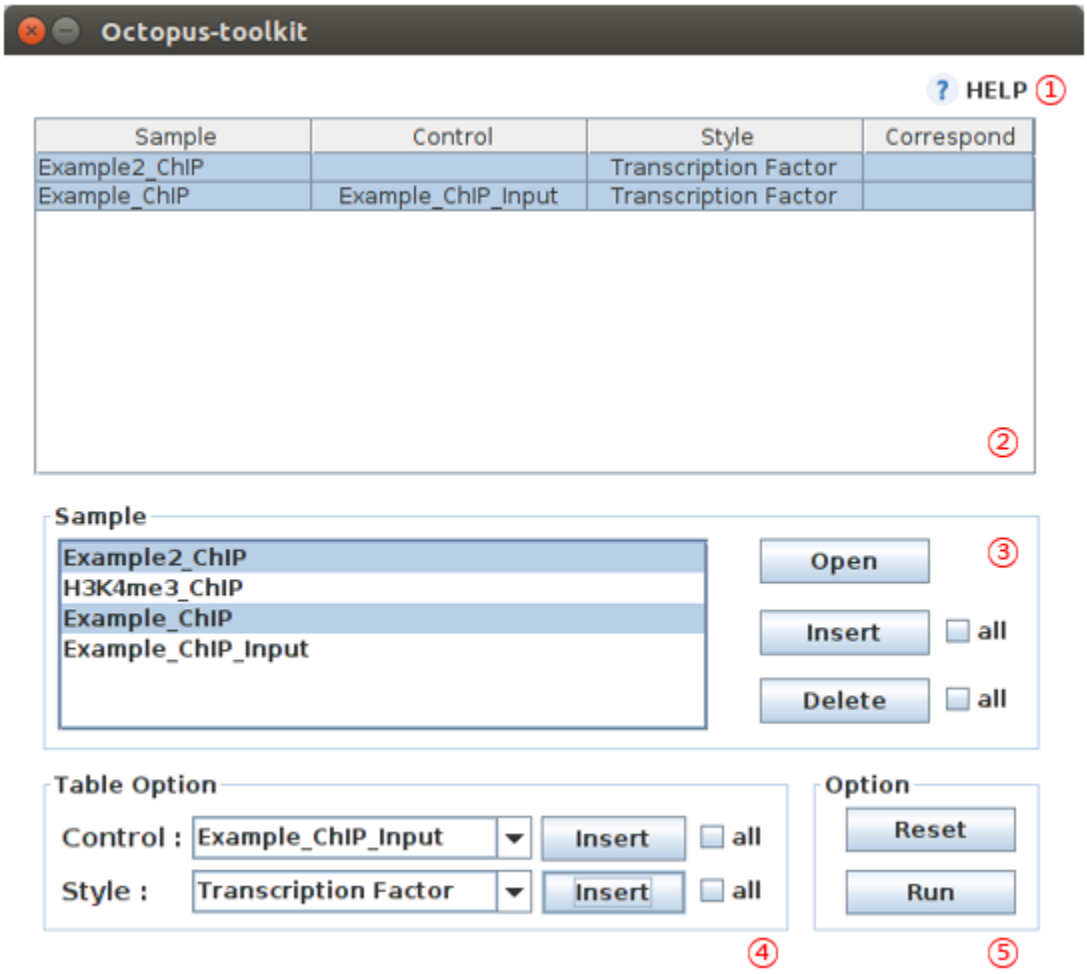


No	Name	Description
1	`Help`	Go to the tutorial page.
2	`Private table`	Files with related information.
3	`Setup`	The option window is used to set appropriate information needed for processing given files.
4	`Apply`	You can apply or reset the option.

4-6.Peak Calling Table

You can identify peaks using the Peak Calling function. You have to select appropriate options for each file from the setting window.

This function is not applicable for RNA-seq data.

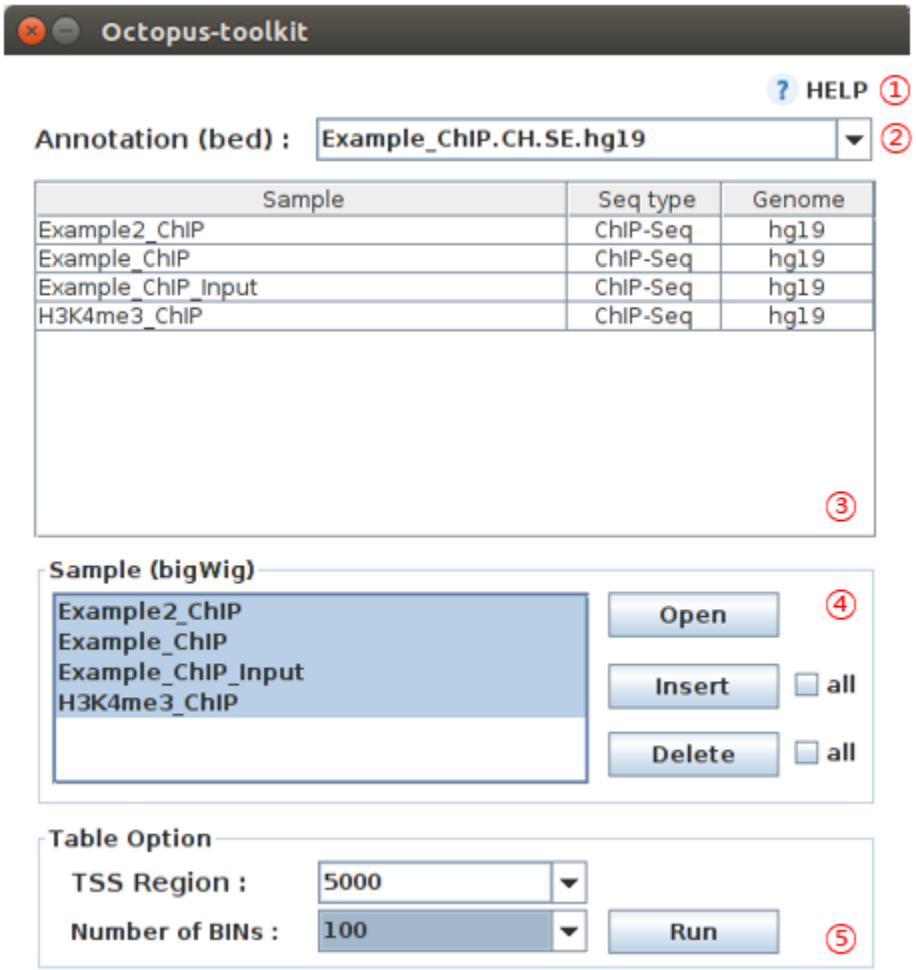


No	Name	Description
1	`Help`	Go to the tutorial.
2	`Set up table`	Parameters for peak calling
3	`Files`	Select files for analysis
4	`Setup`	Select appropriate options for given files
5	`Apply`	You can apply or reset the parameters

4-7.Graph Table

To draw heatmap and line plots with the identified regions.

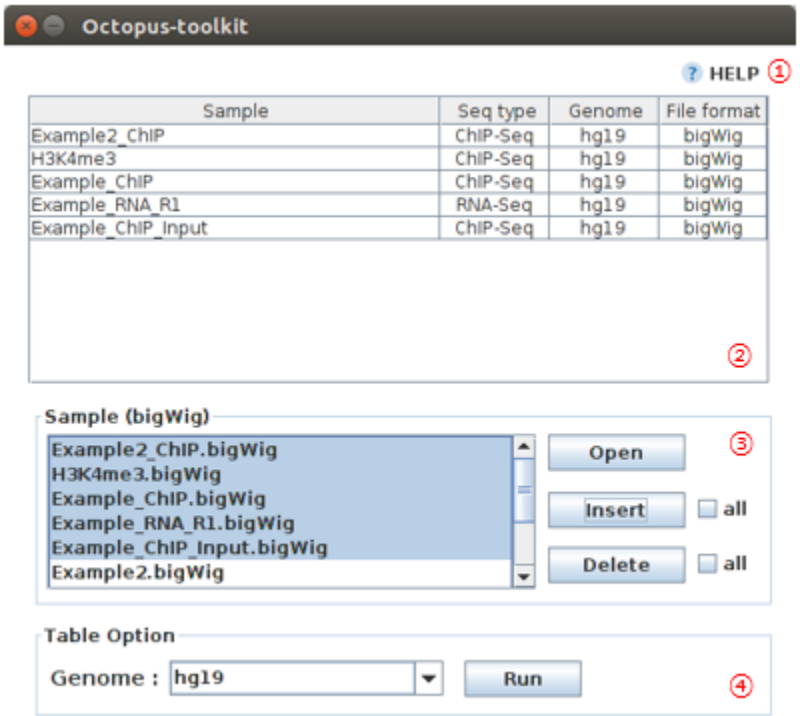
This function is not applicable for RNA-seq data.



No	Name	Description
1	`Help`	Go to the tutorial.
2	`Annotation`	Choose a peak file.
3	`Samples`	Status window
4	`Sample bigWig files`	Select bigWig files of samples you want to draw over the identified regions in the peak file.
5	`Option`	Define the range (bp) relate to the center of peaks.

4-8.IGV Table

You can visualize your data with bigWig files via IGV (IGV, Integrative Genomics Viewer).



No	Name	Description
1	`Help`	Go to the tutorial.
2	`Samples`	Status window.
3	`Sample bigWig files`	Select bigWig files for visualization.
4	`Genome`	Choose the reference genome.

4.1.6 5.User Guide

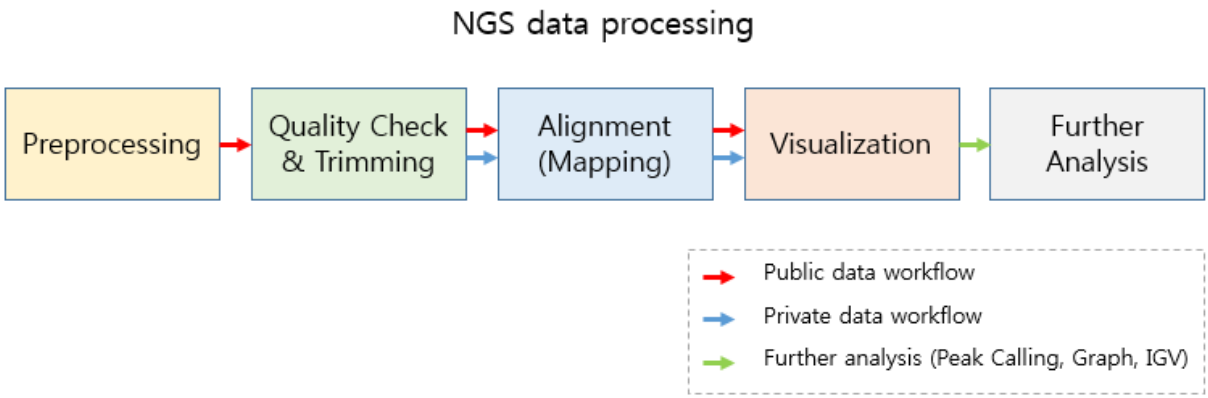
Octoput-toolkit can analyze a number of publicly available next-generation sequencing (NGS) data in with a single step. In addition, you can also analyze your own data (.fastq) using the same analysis pipeline that is provided by the Octopus-toolkit.

- `Supported NGS types` : RNA-seq, ChIP-seq, ATAC-seq, DNase-seq, MeDIP-seq, and MNase-seq
- `Public data` : NGS data released from gene expression omnibus (GEO).
- `Private data` : NGS data stored in your computer (.fastq or .fastq.gz)

Octopus-toolkit provides several additional functions for further analysis. * `Peak Calling` : Identification of read enriched regions (.bed) * `Drawing Graph` : Drawing line plot and heatmap on specified regions (.bed) * `Visualization` : Explore genome with bigWig files through IGV

Basically, Octopus-toolkit processes NGS data by the following steps.

- `NGS data processing`



5-1.3rd party tools used in Octopus-toolkit

Octopus-toolkit utilizes the following 3rd party tools during the process.

NGS Process	Function	3rd party tool	Sub-tools
`Preprocessing`	Download SRA files from NCBI	`Aspera`	ascp
	Convert SRA files to Fastq files	`SRAToolkit`	fastq-dump
`Quality check`	Quality check for raw data	`FastQC`	fastqc
`Trimming`	Trimming for adapter sequence and portions of low-quality reads	`Trimmomatic`	
`Alignment`	Indexing a reference genome	`Hisat2`, `STAR`	hisat2-build, STAR
	Mapping reads to the reference genome	`Hisat2`, `STAR`	hisat2-align, STAR
`Visualization`	Create bigWig files for visualization	`Homer`	makeTagDirectory,makeUCSCFile,analyzeRe
`Peak calling`	Detect enriched regions by mapped reads	`Homer`	findPeak, pos2bed,annotatePeaks
`Graph`	Calculate normalized values from bigWig files	`Bwtool`	matrix
	Draw the heatmap and line plot	`R`	pheatmap, ggplot2
`IGV`	Explore the genome with processed data (bigWig files)	`IGV`	

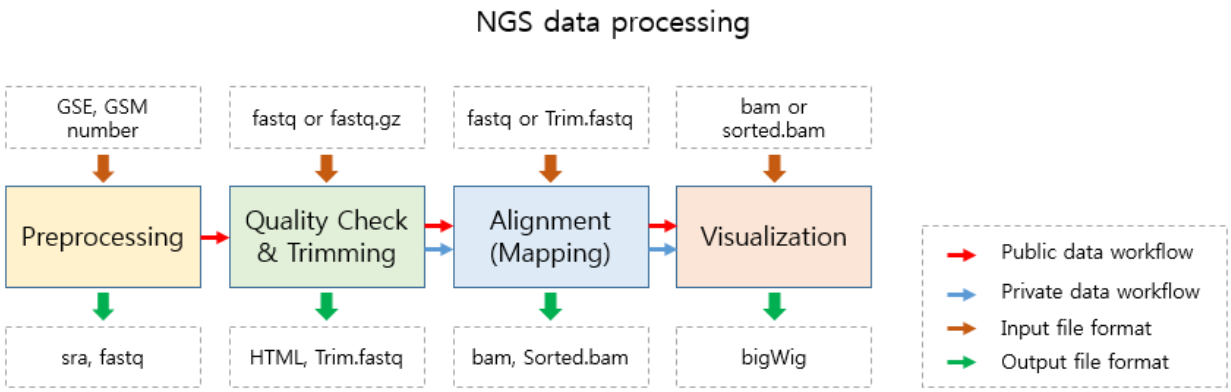
5-2.Public data

Quick Start

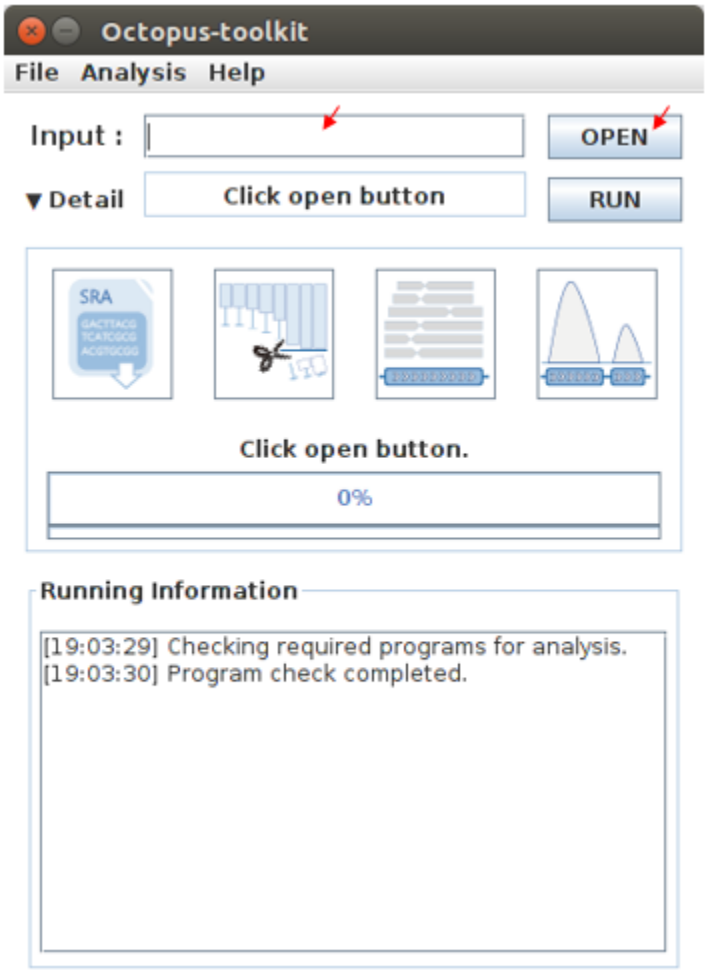
Note:

1. Enter a GEO (Gene Expression Omnibus) accession number or click the open button to load a list of GEO accession numbers.
2. Click the Run button.
3. Select appropriate options you want to use.
4. Click the Run button to begin the analysis.

Work flow



To analyze a single or a set of publicly available NGS data from GEO.



Please enter a single GEO Accession number or open a text file containing GEO Accession numbers.

- Input : `GEO Accession number`

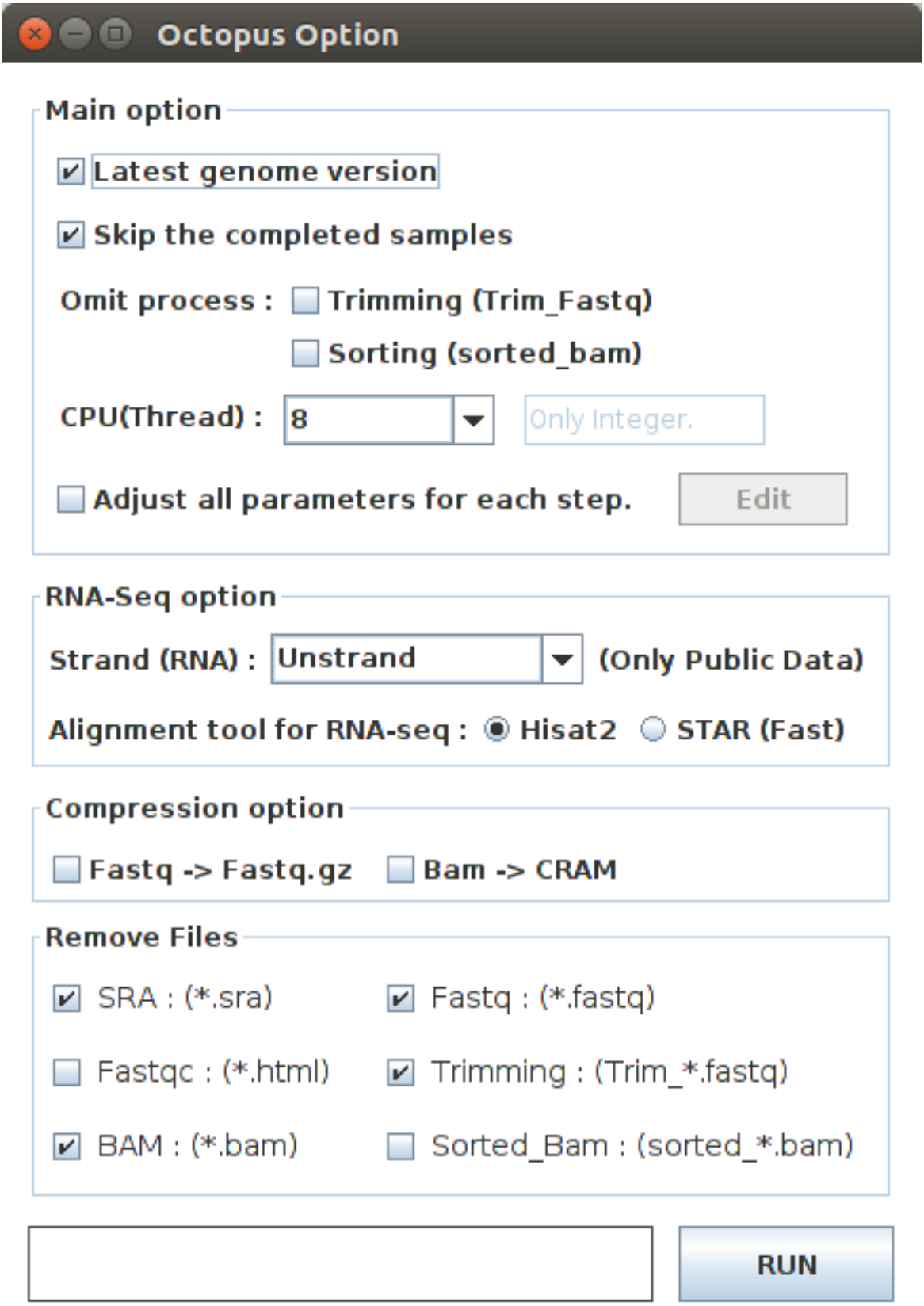
```
GSExxx : Each GSE (study) record is assigned to a single study which contains at
↪least a NGS data (GSM).
GSMxxx : Each GSM (sample) record is assigned to a single NGS data.
```

- Input : `GEO Accession number list` (`example.list`)

1. Enter a single GEO accession number or click the open button to load a list GEO accession numbers.
2. Click the Run button.

Octopus option

You can change a number of parameters provided by Octopus-toolkit or the integrated tools.


Octopus Option
Main option
Latest genome version
Skip the completed samples
Omit process :
Trimming (Trim_Fastq)
Sorting (sorted_bam)
CPU(Thread) :
8
Only Integer.
Adjust all parameters for each step.
Edit
RNA-Seq option
Strand (RNA) :
Unstrand
(Only Public Data)
Alignment tool for RNA-seq :
Hisat2
STAR (Fast)
Compression option
Fastq -> Fastq.gz
Bam -> CRAM
Remove Files
SRA : (*.sra)
Fastq : (*.fastq)
Fastqc : (*.html)
Trimming : (Trim_*.fastq)
BAM : (*.bam)
Sorted_Bam : (sorted_*.bam)
RUN

Option	Decription
`Latest genome version`	Use the latest genome rather than the genome used for the study.
`Skip the completed samples`	Skip the samples that have already been analyzed.
`Omit process`	Omit the selected processes such as trimming and sorting steps.
`CPU(Thread)`	Set the number of CPUs to use.
`Adjust all parameters for each step`	Change full parameters in each step.
`Edit`	Open the Full parameter option window.
`Strand (RNA)`	Set the library strand for RNA-Seq.
`Alignment tool for RNA-seq`	Set the alignment tool for RNA-seq.
`Fastq -> Fastq.gz`	Compress Fastq to Fastq.gz.
`Bam -> CRAM`	Compress Bam to CRAM.
`Remove Files`	Delete selected intermediate files once each process is completed to save space.

- `Latest genome version`

Octopus-toolkit can analyze the genomes of Homo sapiens, Mus musculus, Drosophila melanogaster, Saccharomyces cerevisiae, and Canis lupus familaris.

Table 1: Available analysis genome version

Organism	Genome version
Homo sapiens	hg38, hg19, hg18
Mus musculus	mm10, mm9
Drosophila melanogaster	dm6, dm3
Saccharomyces cerevisiae	sacCer3
Canis lupus familaris	canFam3
Arabidopsis thaliana	tair10
Danio rerio	danRer10
Caenorhabditis elegans	ce11

The latest genome version uses the latest version of the genome for analysis. If you don't select this option, Octopus-toolkit uses the genome defined by submitter.

- Latest genome (O) : hg38, mm10, dm6, sacCer3, canFam3, tair10, danRer10, ce11
- Latest genome (X) : hg19, mm9, dm3, sacCer3, canFam3, tair10, danRer10, ce11
- `Skip the completed samples`

While analyzing a number of GSE/GSM data, you can stop the analysis and resume it later.

Octoput-toolkit will skip the samples that have been analyzed completely.

If you have the samples that have been analyzed completely and you want to analyze it again, please do not check this option.

- `Omit process`

The omit process allows you to skip the trimming step and/or the sorting step. This shortens the anaysis time.

During the trimming process, all reads will be discarded if all of the reads have bad sequencing quality. Octopus-toolkit will analyze the original raw data (.fastq) in this case by skipping the trimming step.

During the sorting process, BAM file will be sorted by using Samtools. In general, many applications uses sorted BAM files. If you are not interested in analyzing the sorted BAM files, you may skip this process.

- `CPU(Thread)`

You can set the number of CPUs for analysis. (Default : Maximum number of cores depending on your computer)

- `Adjust all parameters for each step`

You can adjust many parameters for each stop. Check the box and click the Edit button. The parameter window will pop up.

Please refer to the link for details : *Full Parameter*

- `Edit`

When you click the Edit button, the parameter window will appear.

- `Strand (RNA)`

The strand option allows you to choose whether or not to take the stranded information into account. This is only available for stranded-specific RNA-seq.

Octopus-toolkit extracts information from the GEO website when analyzing the public data. However, stranded-specific information of RNA-seq is not well documented. Therefore, this may or may not be applicable depending on the data.

You can select either non-strand library or the strand-specific library such as FR-Firststrand, FR-Secondstrand using this option.

- `RNA-Seq alignment tools`

You can select an alignment tool to be used during the alignment process for RNA-seq: HISAT2 or STAR.

HISAT2 uses less memory (RAM) than STAR, but STAR is generally faster than HISAT2.

- `Fastq->Fastq.gz` or `Bam->CRAM`

You can compress intermediate files to save your disk space.

- `Remove Files`

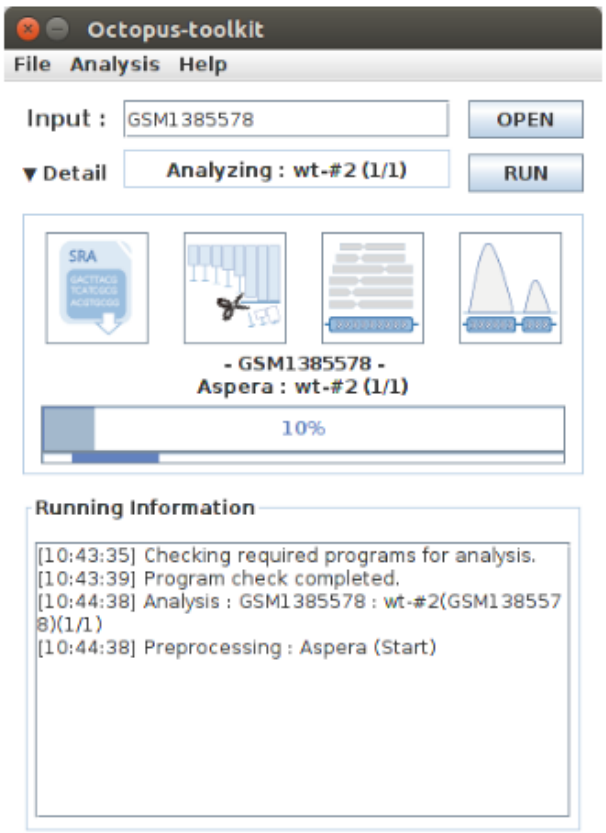
Each step creates intermediate files which may or may not be used. If you want to further analyze the processed data, you might want to keep those intermediate files. If not, you can remove intermediate files (up to few hundread gigabytes) by selecting the boxes in Reomve Files window.

Option	Extension	Description
`SRA`	sra	Compressed raw data downloaded from NCBI. (Sequence Read Archive)
`Fastq`	fastq or fq	Raw data converted during preprocessing. (A short read sequence file)
`Fastqc`	html or text	Output generated during Quality Check. (output of FastQC)
`Trimming`	fastq or fq	Output generated during Trimming. (Trimmed raw file(Fastq))
`BAM`	bam	Output generated during Alignment. (Mapped read to the genome)
`Sorted_Bam`	bam	Output generated during Sorting. (Sorted mapped read)

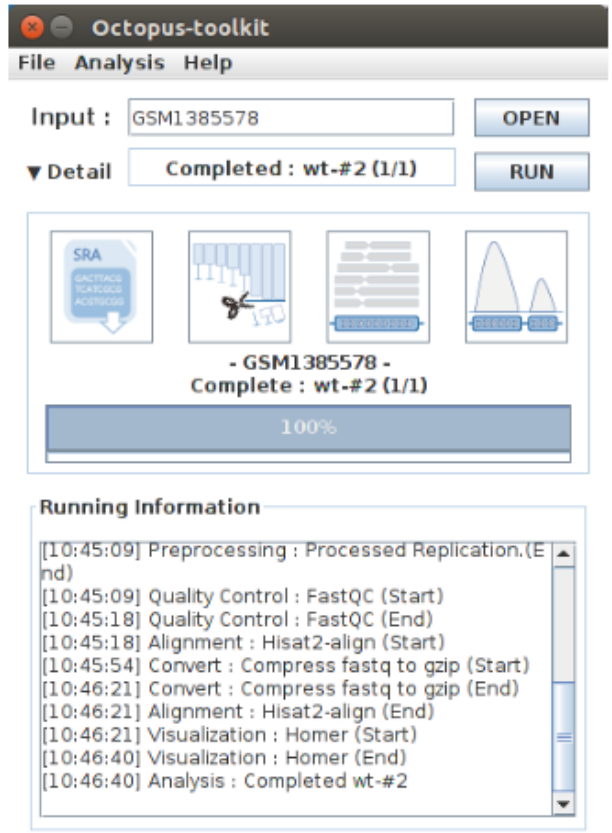
3. Set the paramters and options.
4. Click the Run button to begin the anlaysis.

Run

- Below shows progress bar and status window (GSM1385578).



< The start of analysis >



< The end of analysis >

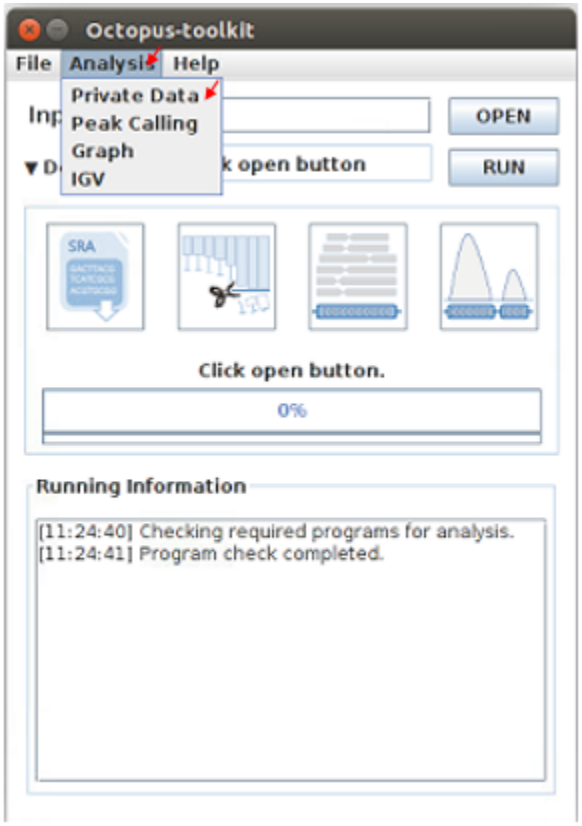
5-3.Private data

Quick Start

Note:

1. Select the analysis tab -> Select the Private Data function in the Menu bar.
2. Select raw files (.fastq) in your computer.
3. Add appropriate information for each sample in the private table.
4. Click the Run button in the private table.

Analyzing your data (private data)



Unlike the public data analysis, private data analysis does not require the converting step (.sra to .fastq).

Input files can be fastq (.fastq or .fq) files or compressed fastq (.fastq.gz or .fq.gz) files.

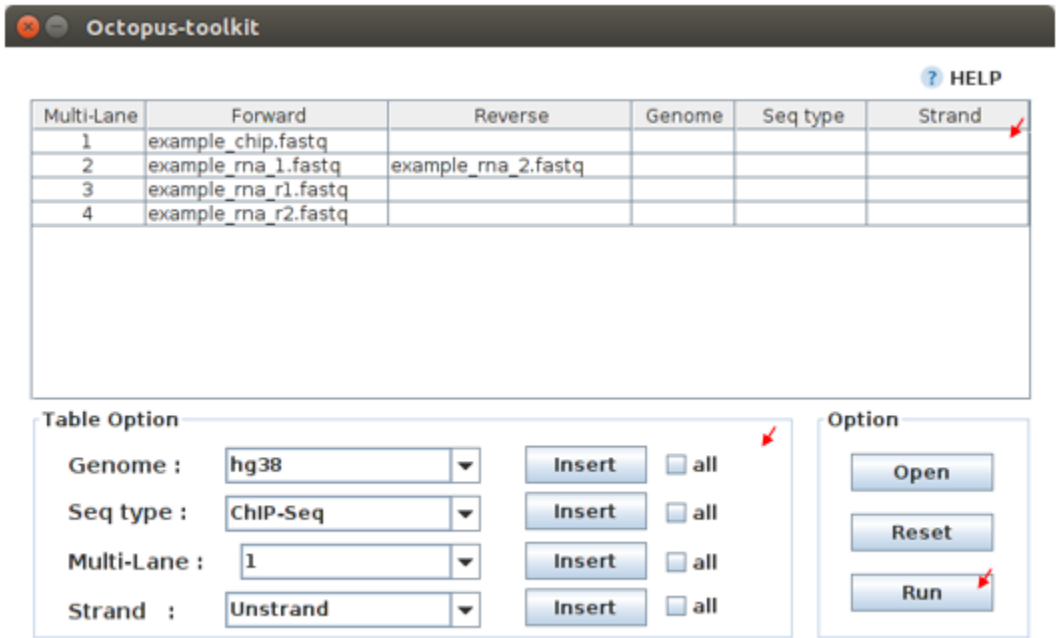
Files must follow the rules below.

Note:

- Raw data : Sample `.fastq` or Sample `.fq`
- compressed Raw data : Sample `.fastq.gz` or Sample `.fq.gz`
- Single-End data : Sample `.fastq` (or fq, fastq.gz, fq.gz)
- Paired-End data : Sample `_1.fastq`, Sample `_2.fastq`

Octopus-toolkit only loads files that match the above rules.

Private table



Octopus-toolkit requires appropriate sample information for each file. You need to specify the required information.

If any of the selected files does not appear in the list, please check the file name and format of your files.

You must specify the following information for each sample.

Option	Decription
`Genome`	Select the genome.
`Seq type`	Select the experimental type such as ChIP-seq.
`Multi-Lane`	Set the Multi-lane option.
`Strand`	Select the strand strategy if applicable.

- `Genome`

The following genomes are available in the Octopus-toolkit:

Species	Genome version
`Homo sapiens`	`hg38` (Dec.2013, GRCh38), `hg19` (Feb.2009,GRCh37), `hg18` (Mar.2006 NCBI36)
`Mus musculus`	`mm10` (Dec.2011 GRCm38), `mm9` (July.2007 NCBI37)
`Drosophila melanogaster`	`dm6` (Aug.2014 BDGP Release 6+ ISO1 MT), `dm3` (Apr.2006 BDGP R5)
`Saccharomyces cerevisiae`	`sacCer3` (Apr.2011 SacCer_Apr2011)
`Canis lupus familaris`	`canFam3` (Sep.2011 Broad CanFam3.1)
`Arabidopsis thaliana`	`tair10`
`Danio rerio`	`danRer10` (Sep.2014 GRCz10)
`Caenorhabditis elegans`	`ce11` (Feb.2013 WBcel235)

- `Seq type`

Octopus-toolkit supports the following experimenatal types: ChIP-Seq, RNA-Seq, MeDIP-Seq, ATAC-Seq, DNase-Seq and MNase-Seq.

- `Multi-Lane`

A single sample can be obtained from muliple lanes in a sequencing instrument. In this case, files from multe lanes can be merged by setting the same number in the Multi-Lane column.

Multi-lane files generally have the following filenames.

```
Sample.L001.fastq, Sample.L002.fastq, Sample.L003.fastq ... Sample.L008.fastq
```

To merge the above files, you must set the number of 'Multi-Lane columns' to the same number for each file.

- `Strand`

This option is to set the library strategy for RNA-seq.

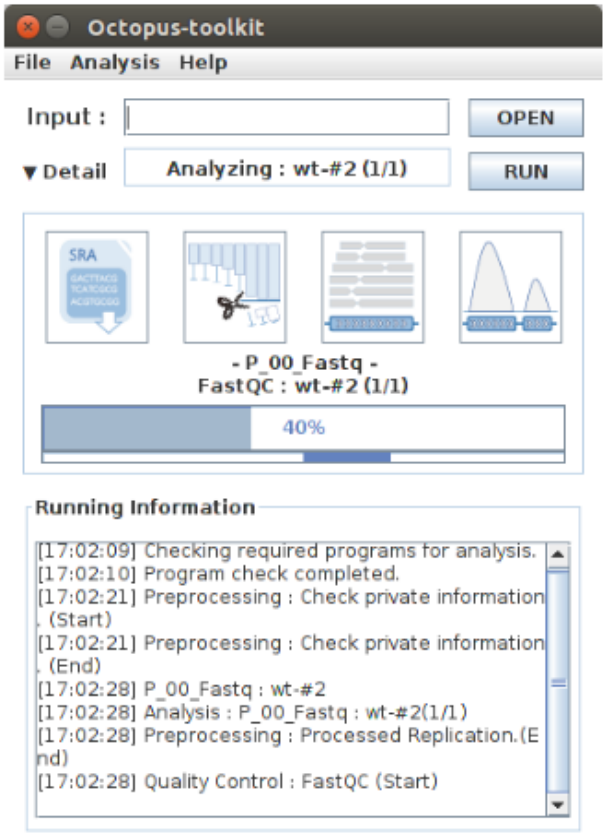
1. Unstranded library : `Unstrand` (Default)
2. Strand-specific library : `FR-Firststrand` or `FR-secondstrand`

Options

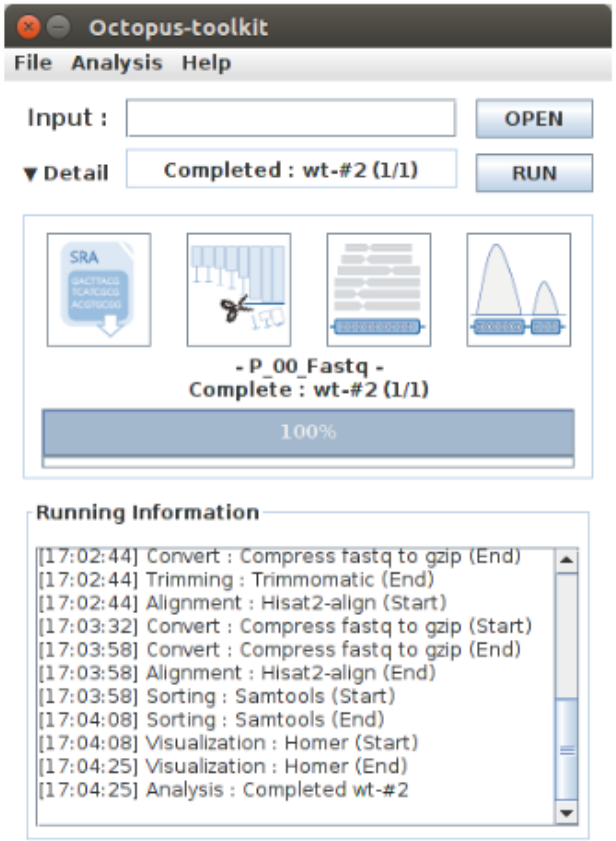
Options for private analysis is the same as public data analysis. Please refer to the public data analysis. (*Octopus option*)

Run

- Snapshots (Private data analysis)



< The start of analysis >



< The end of analysis >

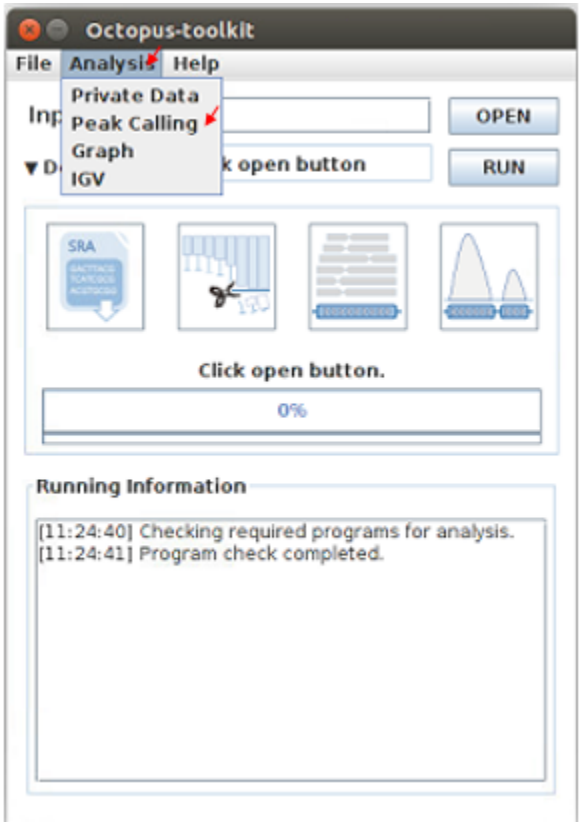
5-4.Peak Calling

Quick Start

Note:

1. Select the Analysis tab -> Click the Peak Calling function in the Menu bar.
2. Select the output folder (Result/GSExxxxx) in the Result directory generated by Octopus-toolkit.
3. Add information of each sample in the peak calling table.
4. Click the Run button.

Peak calling anlaysis

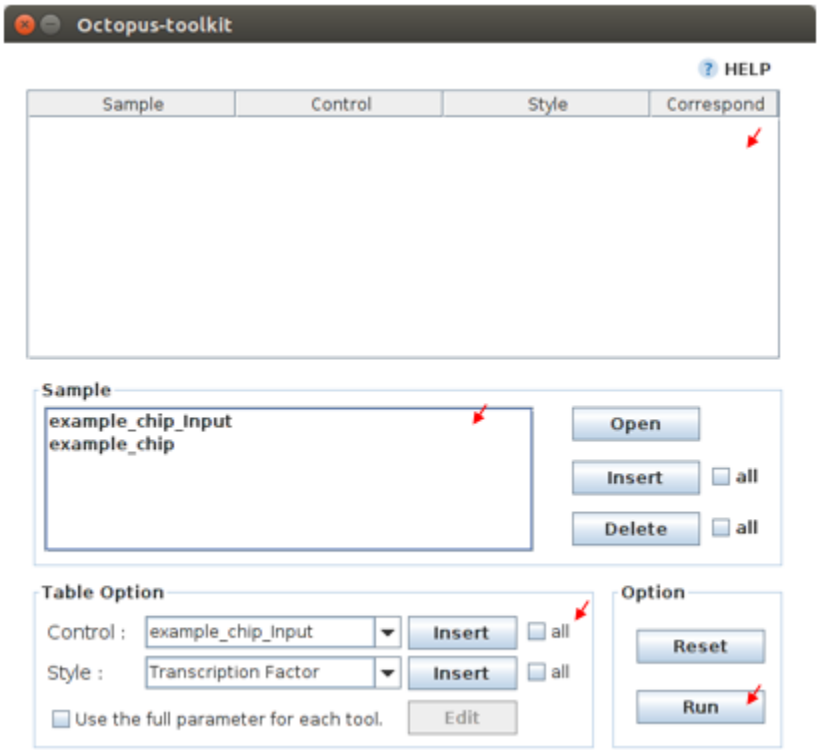


The purpose of the peak calling anlaysis is to identify regions enriched by mapped reads.

In order to perform the peak calling analysis, you must have the Octopus-toolkit output folders.

1. Select the Analysis tab -> Click the Peak Calling function in the Menu bar.
2. Select output directories generated by the Octopus-toolkit.

Peak calling table



To run the peak calling analysis, please select output folders (Result/GSExxxxx). Then, fill in the blanks using the Table Option functions.

- `Control`

If available, please select an appropriate control (IgG or input) per sample to filter out the background noise. (Recommended)

- `Style`

Based on experimental types, you can select a predefined paramter (by HOMER) for the Peak calling process.

option	Seq type	Description
`Transcription Factor`	ChIP-Seq, DNase-Seq	Peak finding for single contact or focal ChIP-Seq experiments or DNase-Seq.
`Histone`	ChIP-Seq	Peak finding for broad regions of enrichment found in ChIP-Seq experiments for various histone marks.
`DNase`	DNase-Seq	Adjusted parameters for DNase-Seq peak finding.
`mC`	MeDIP-Seq	DNA methylation analysis.

Please select a style option that meets your analysis needs.

3. Add appropriate information for each sample in the private table.
4. Click the Run button in the private table.

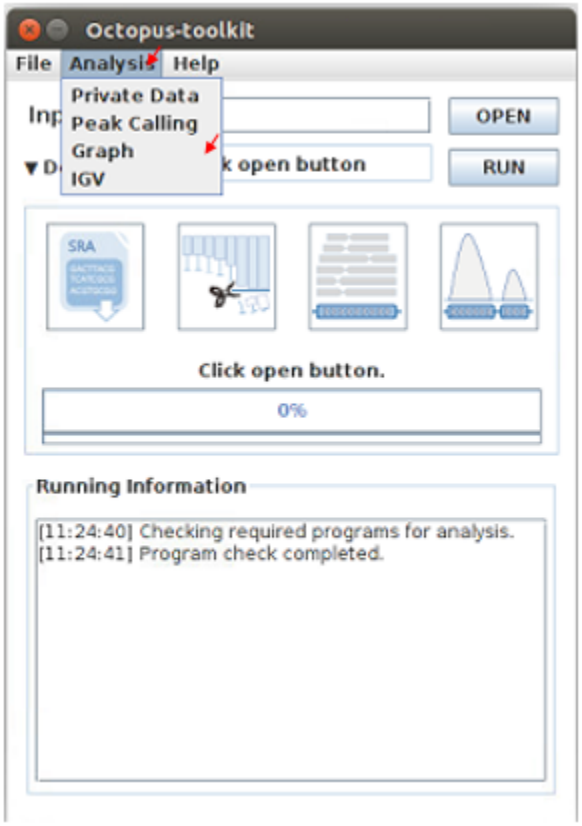
5-5.Graph

Quick Start

Note:

1. Select the Analysis tab -> Click the Graph function in the Menu bar.
2. Select output folders (Result/GSExxxxx). Multiple output folders can be selected.
3. Set the range of transcription start site (TSS) region and BIN size in the Graph table.
4. Click the Run button.

Start analyzing Graph



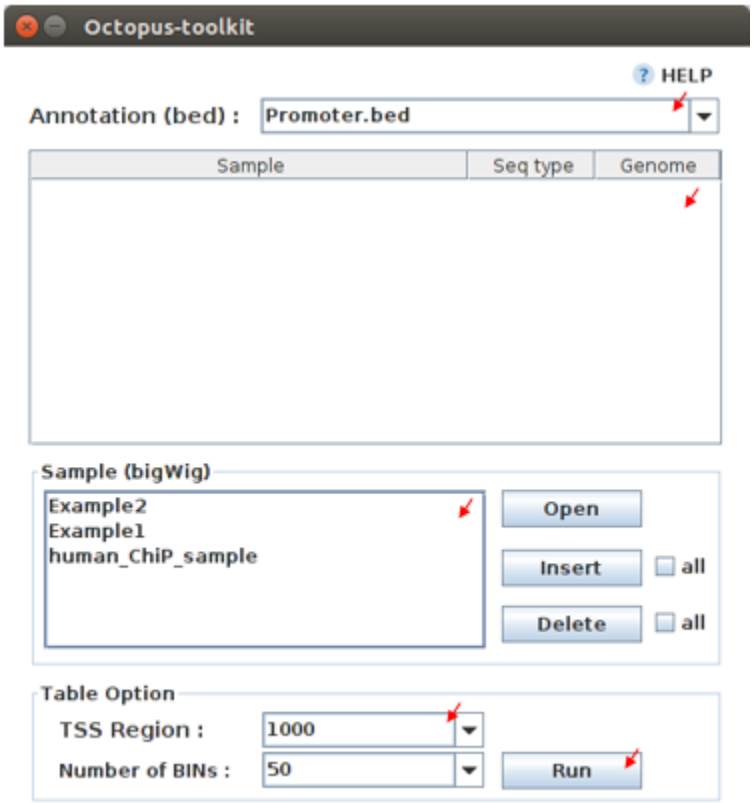
The Graph function is to draw average signal pattern on specificed regions which are defined by the user. Signals are extracted from bigWig (normalized to ten million mapped reads) files.

If you would like to draw plots on peaks, you need to complete the peak calling analysis for a sample of your interest.

- `Previous steps` : Public data or Private data analysis -> Peak Calling.

1. Select the Analysis tab -> Click the Graph function in the Menu bar.
2. Select output folders generated by either Public analysis or Private analysis.

Graph table



To draw graphs, Octopus-toolkit requires bigWig (signal) files, which are generated by either Public anlaysis or Private analysis.

- `bigWig` : Output of the Public data or Private data analysis.
- `bed` : Output of the Peak calling analysis.

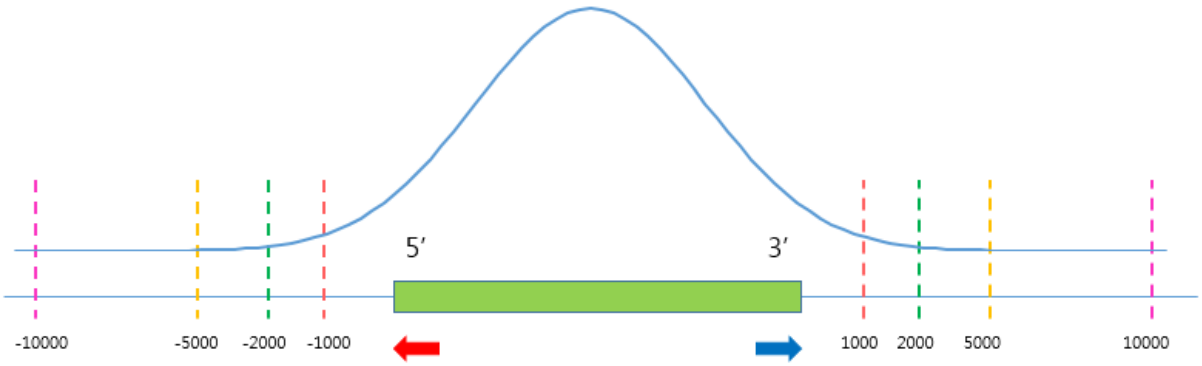
Finally, hit the Run button. The output (plots) will be stored in the Graph directory under the Result folder.

- `Annotation(bed)`

First, select loci (.bed) of your interest from the Annotation (bed) function. Second, select samples (.bigWig) of your interest from the Sample list.

- `TSS Region`

Third, set appropriate paramters from the Table option. The unit for this option is basepair (bp).



The default ranges of TSS-regions are 1000, 2000, 5000 and 10000 bp.

- `Number of BINs`

The region selected in the TSS region option is divided into n (number of BINs) BINs

The lager the bin size, the smoother the graph can be drawn.

3. set the TSS region and BIN size in the Graph table.
4. Click the Run button.

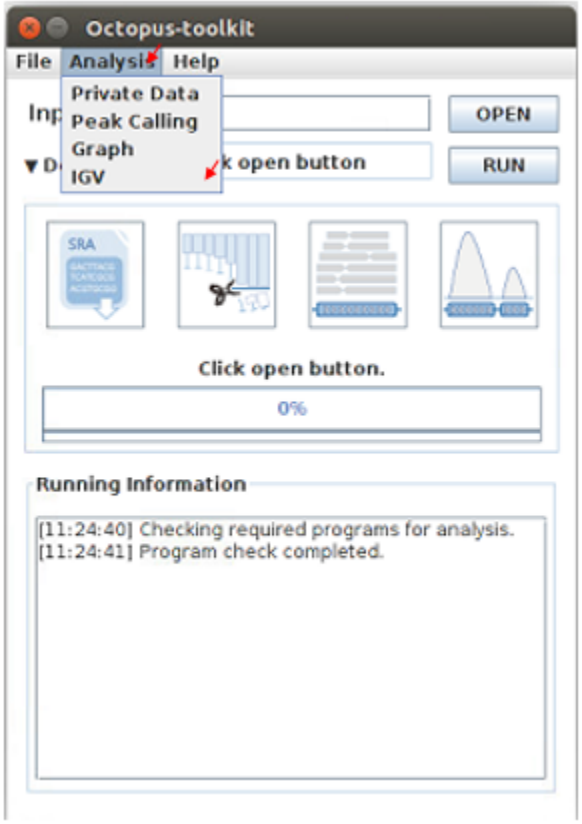
5-6.Visualization

Quick Start

Note:

1. Select the Analysis tab -> Click the IGV function in the Menu bar.
2. Select output folders (Result/GSExxxxx) of your interest.
3. In the sample window, select samples and then, click the Insert button.
4. Check see if all genomes are the same. Only data in the same genome can be loaded into the IGV.
5. Set the same genome in the Table option.
6. Click the Run button.

Start analyzing IGV

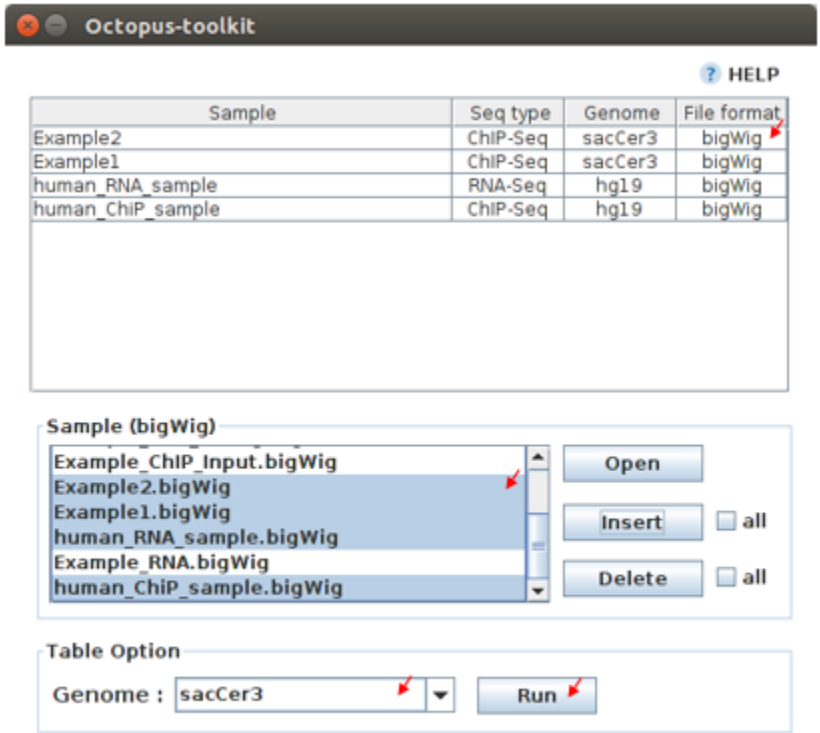


The IGV function is a process of visualizing analyzed data through IGV, a visualization tool.

IGV uses bigWig files.

1. Select the Analysis tab -> Click the IGV function in the Menu bar.
2. Select output folders (Result/GSExxxxx) of your interest.

IGV table



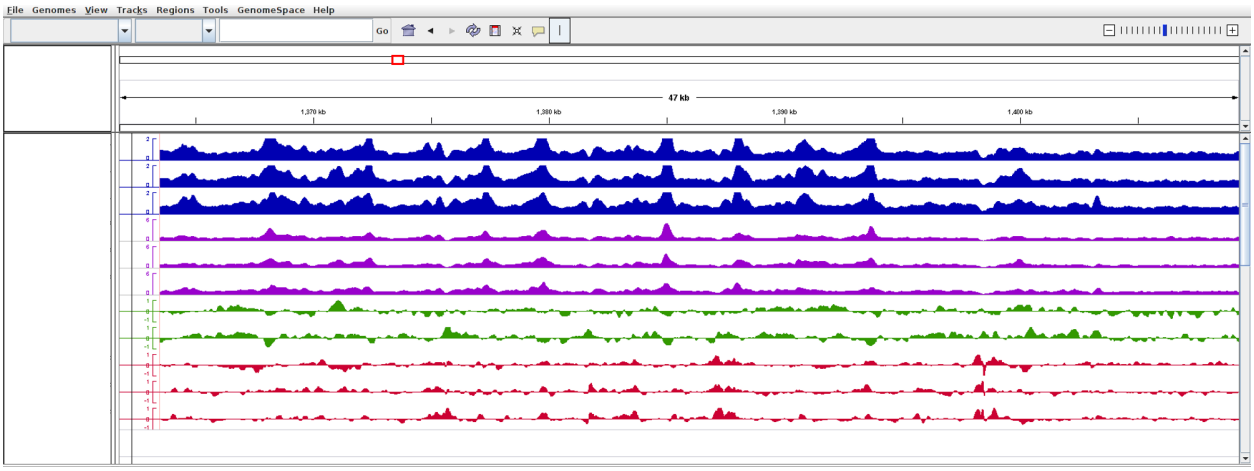
- `Genome`

Genome option shows the genome of the samples.

3. In the sample window, select samples and then, click the Insert button.
4. Check if all genomes are the same. Only the data in the same genome can be loaded into the IGV.
5. Set the same genome in the Table option.
6. Click the Run button.

Run

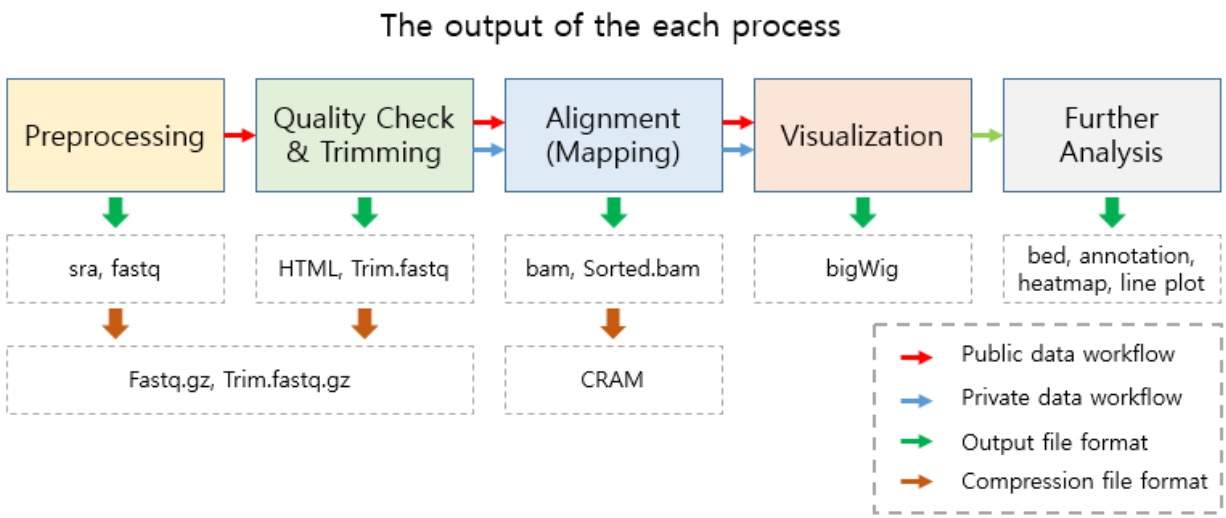
- Below shows run screen of IGV.



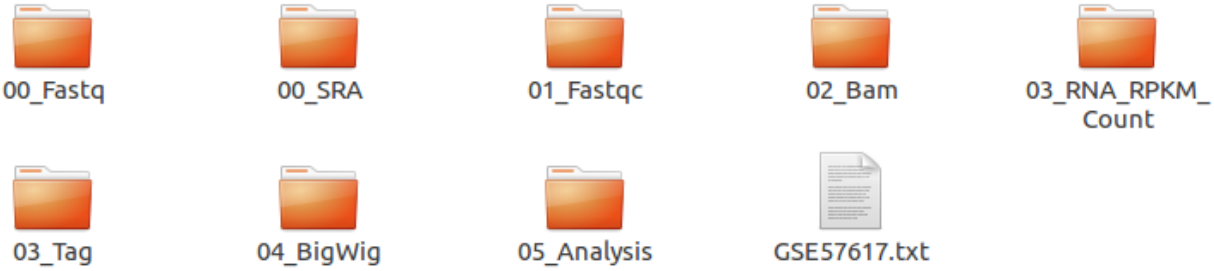
Unlike other tools integrated in Octopus-toolkit, the IGV tool runs separately from the Octopus-toolkit.

5-7.Output (important!)

The output files generated by each process are as follows:

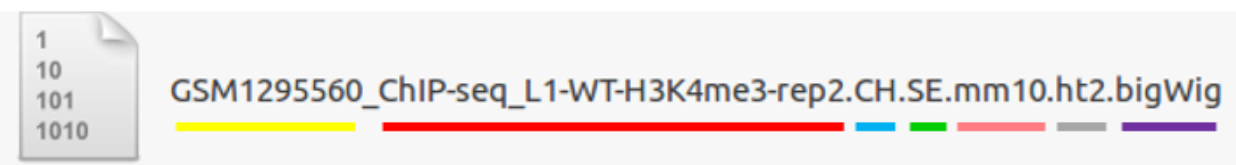


- In the Result folder



Folder name	Process	File format	Description
`00_Fastq`	Preprocessing,Trimming	`fastq`, `Trim.fastq`	Save the raw file and trimmed file.
`00_SRA`	Preprocessing	`sra`	Store the SRA file downloaded from NCBI
`01_Fastqc`	Quality check	`html`, `txt`	Save the result of the Quality check.
`02_Bam`	Alignment	`bam`, `sorted.bam`, `bai`	Save the Alignment and sorted files.
`03_RNA_RPKM_Count`	Normalization	`RPKM`, `Count`	Save the calculated RPKM and raw read count tables for the RNA-Seq data.
`03_Tag`	Downstream (motif) analyses by HOMER	`Tag folder`	Save the Tag folders created by the Homer tool.
`04_BigWig`	Visualization	`bigWig`	Save the bigWig files for visualization
`05_Analysis`	Peak Calling,Annotation	`bed`, `annotation`	Save the peak (.bed) and annotation files.
`GSE57617.txt`	Log file	`txt`	Sample.txt is a file that stores the analysis status and information.

5-8.File Naming



- **`Yellow`** [GSM Accession number] Only applicable for the public data.
- **`Red`** [ChIP-Seq_L1-WT-H3K4me3-rep2] Sample file name (title) defined on the GEO website.
- `Blue` : Experimental types described below.

Experimental types	Abbreviation	Experimental types	Abbreviation	Experimental types	Abbreviation
ChIP-Seq	`CH`	RNA-Seq	`RN`	MeDIP-Seq	`ME`
ATAC-Seq	`AT`	DNase-Seq	`DN`	MNase-Seq	`MN`

- **`Green`** [Sequencing strategy] SE : Single-End, PE : Paired-End
- **`Pink`** [Reference genome] Reference Genome
- **`Gray`** [Alignment tool] RNA-Seq alignment tools. (ht2 : Hisat2, str: STAR)
- **`Purple`** [File extension] Output Format

5-9.Full Parameters

You can adjust the parameters of 3rd party tools integrated into the Octopus-toolkit.

The 3rd party tools used in Octopus-toolkit : *3rd party tools*

Preprocessing

In the preprocessing step, Octopus-toolkit downloads selected NGS data from NCBI and converts the downloaded (.sra) files to FASTQ files. The 3rd party tools used in the preprocessing step are Aspera and SRAToolkit(fastq-dump)

- **Transfer rate**

  `MAX-RATE` : MAX transfer rate (Only Integers)

  `MIN-RATE` : MIN transfer rate (Only Integers)

  `Overwrite` : Overwrite-Method, Always(Default), Never, Older, Diff

- **Convert Sra to Fastq (Filtering)**

  `MIN-Read Length` : Filter by sequence length >= <Value> (Only Integers)

  `Aligned or unaligned reads` : Dump only aligned sequence or unaligned sequences, NotUse(Default), Both, Aligned, Unaligned

  `Quality conversion` (offset) : Offset to use for quality conversion, 33(Default), 64

  `Dump biological reads(Only)` : Dump only biological reads, No(Default)

QC & Trimming

QC & Trimming is the process of assessing the quality of the reads. If bad sequencing quality are detected, portions of low-quality reads are trimmed. The 3rd party tools used in QC & Trimming are FastQC and Trimmomatic.

- **Determining the quality of DNA Sequence**

  `K-Mer` : Specifies the length of Kmer to look for in the Kmer content module, Specified Kmer length must be between 2 and 10. Default length is 7 if not specified.

  `Allocated memory` : Set the momory available on your computer for Quality check. Provides a measure of currently available memory . (Octopus-toolkit option)

- **Trimmed DNA sequence data**

  `Illumina adapt Sequence` : Cut adapter and other illumina-specific sequences from the read.

  `Seed mismatches` : Specifies the maximum mismatch count which will still allow a full match to be performed

  `Palindrome clip threshold` : Specifies how accurate the match between the two 'adapter ligated' reads must be for PE palindrome read alignment.

  `Simple clip threshold` : Specifies how accurate the match between adapter or any sequence must be against a read.

  `Window size` : specifies the number of bases to be averaged.

  `Average quality` : Specifies the average quality required.

  `LEADING` : Specifies the minimum quality required to keep a base.

  `TRAILING` : Specifies the minimum quality required to keep a base.

  `HEADCROP` : The number of bases to keep, from the start of the read.

  `TAILCROP` : The number of bases to remove from the start of the read.

`Minimum length of reads to be kept` : Specifies the minimum length of reads to be kept.

Alignment-Hisat2

Alignment is the process of mapping reads to a reference genome. The 3rd party tool used in Alignment is Hisat2.

- **Input**

  `Skip N read` : Skip the first <int> reads/pairs in the input (none)

  `Stop after aligning N reads` : Stop after first <int> reads/pairs (no limit)

  `Trim N bases 5' end` : Trim <int> bases from 5'/left end of reads (0)

  `Trim N bases 3' end` : Trim <int> bases from 3'/right end of reads (0)

- **Scoring**

  `Ambiguous read penalty` : Penalty for non-A/C/G/Ts in read/ref

  `Mismatch penalty` : Max and min penalties for mismatch; lower qual = lower penalty <2,6>

  `Soft-Clipping penalty` : Max and min penalties for soft-clipping; lower qual = lower penalty <1,2>

  `Read gap penalty` : Read gap open, extend penalties (5,3)

  `Reference gap penalty` : Reference gap open, extend penalties (5,3)

- **Alignment**

  `Ignore all quality values` : Treat all quality values as 30 in Phred scale (no)

  `Do not align reverse of read` : Do not align forward (original) version of read (no)

  `Do not align forward of read` : Do not align reverse-complement version of read (no)

- **Spliced alignment**

  `Do not spliced alignment` : Disable spliced alignment

  `Canonical` : Penalty for a canonical splice site (0)

  `Non-canonical` : Penalty for a non-canonical splice site (12)

  `MIN-Length` : Minimum intron length (20)

  `MAX-Length` : Maximum intron length (500000)

Alignment-STAR

Alignment is the process of mapping reads to a reference genome. The 3rd party tool used in Alignment is STAR (RNA-Seq only).

- **Alignment**

  `AlignIntronMin` : Minimum intron size: genomic gap is considered intron if itsnlength>=alignIntronMin, otherwise it is considered as Deletion (21)

  `AlignIntronMax` : Maximum intron size, if 0, max intron size will be determined byn(2^winBinNbits)*winAnchorDistNbins (0)

`AlignMatesGapMax` : Maximum gap between two mates, if 0, max intron gap will be determined by (2^winBinNbits)*winAnchorDistNbins (0)

- **Output Filtering**

  `OutFilterMultimapNmax` : int: maximum number of loci the read is allowed to map to. Alignments (all of them) will be output only if the read maps to no more loci than this value. Otherwise no alignments will be output, and the read will be counted as "mapped to too many loci" in the Log.final.out. (10)

  `OutFilterMismatchNmax` : int: alignment will be output only if it has no more mismatches than this value. (10)

  `OutFilterMismatchNoverLmax` : float: alignment will be output only if its ratio of mismatches to *mapped* length is less than or equal to this value.(0.3)

Visualization-TagDirectory

Visualization-TagDirectory is the process of creating Tag directories. The 3rd party tool used in TagDirectory is Homer.

- **Create tag directory**

  `Fragment-Length` : (Set estimated fragment length - given: use read lengths), By default treats the sample as a single read ChIP-Seq experiment

  `Maximum tags per bp` : Maximum tags per bp, default: no maximum

  `Flip the strands of each read` : Flip strand of each read, i.e. might want to use with some RNA-seq

  `Length of the read to keep` : Filter reads with lengths outside this range

Visualization-MakeBigWig

MakeBigWig is the process of creating bigWig files from the Tag directories. The 3rd party tool used in MakeBigWig is Homer.

- **Make visualization data**

  `Size of the bigWig files` : Size of file, when gzipped, default: 1e10, i.e. no reduction

  `Fragment Length` : Approximate fragment length, default: auto

  `Resolution` : Resolution, in bp, of file, default: 1, `avg` report average coverage if resolution is larger than 1bp, default: max is reported

  `Tags per bp to count` : Minimum and maximum tags per bp to count, default: no limit

  `Plot negative values` : Plot negative values, i.e. for - strand transcription

- **Normalization**

  `Normalize the total number of reads` : Total number of tags to normalize experiment to, default: 1e7

  `Set the standard length` : Expected length of fragment to normalize to [0=off], default: 100

PeakCalling-ChIP-Seq/Histone

PeakCalling is the process of detecting enriched regions (peaks) by mapped reads. The 3rd party tool used in PeakCalling is Homer.

- **ChIP-Seq/Histone**

  `Peak size` : Peak size, default: 0

  `MIN-Distance` : Minimum distance between peaks, default: 0 (peak size x2)

  `Genome Size` : Set effective mappable genome size, default: 2e9

  `Fragment Length` : Approximate fragment length, default: auto

  `Input Fragment Length` : Approximate fragment length of input tags, default: auto

  `Tag` : Maximum tags per bp to count, 0 = no limit, default: auto

  `Input tag` : Maximum tags per bp to count in input, 0 = no limit, default: auto

  `Tag count to normalize` : Tag count to normalize to, default 10000000

  `Region Resolution` : Extends start/stop coordinates to cover full region considered "enriched" (YES), `Resolution` number of fractions peaks are divided in when extending 'regions', def: 4

PeakCalling-Peak Filter

- **Peak Filter**

  `Fold Enrichment(Input)` : Fold enrichment over input tag count, default: 4.0

  `Poisson p-value threshold(Input)` : Poisson p-value threshold relative to input tag count, default: 0.0001

  `Fold Enrichment(Local)` : Fold enrichment over local tag count, default: 4.0

  `Poisson p-value threshold(Local)` : Poisson p-value threshold relative to local tag count, default: 0.0001

  `Fold Enrichment(Unique Tag)` : Fold enrichment limit of expected unique tag positions, default: 2.0

  `Local Size(Local tag)` : Region to check for local tag enrichment, default: 10000

  `Input Size(Input tag)` : Size of region to search for control tags, default: 0

  `` `False Discovery Rate `` : False discovery rate, default = 0.001

  `Poisson p-value cutoff` : Set poisson p-value cutoff, default: 0.001

  `Set # of tags` : Set # of tags to define a peak, default: 25

  `Set # of normalized tags` : Set # of normalized tags to define a peak, by default uses 1e7 for norm

PeakCalling-Other analysis

- **MethylC-Seq/BS-Seq**

`Find Region` : Find unmethylated/methylated regions, default: -unmethyC

`Methyl Threshold` : Methylation threshold of regions, default: avg methylation/2

`Min cytosine per Methyl` : Minimum number of cytosines per methylation peak, default: 6

4.1.7 6.Tutorial

A case by case tutorial.

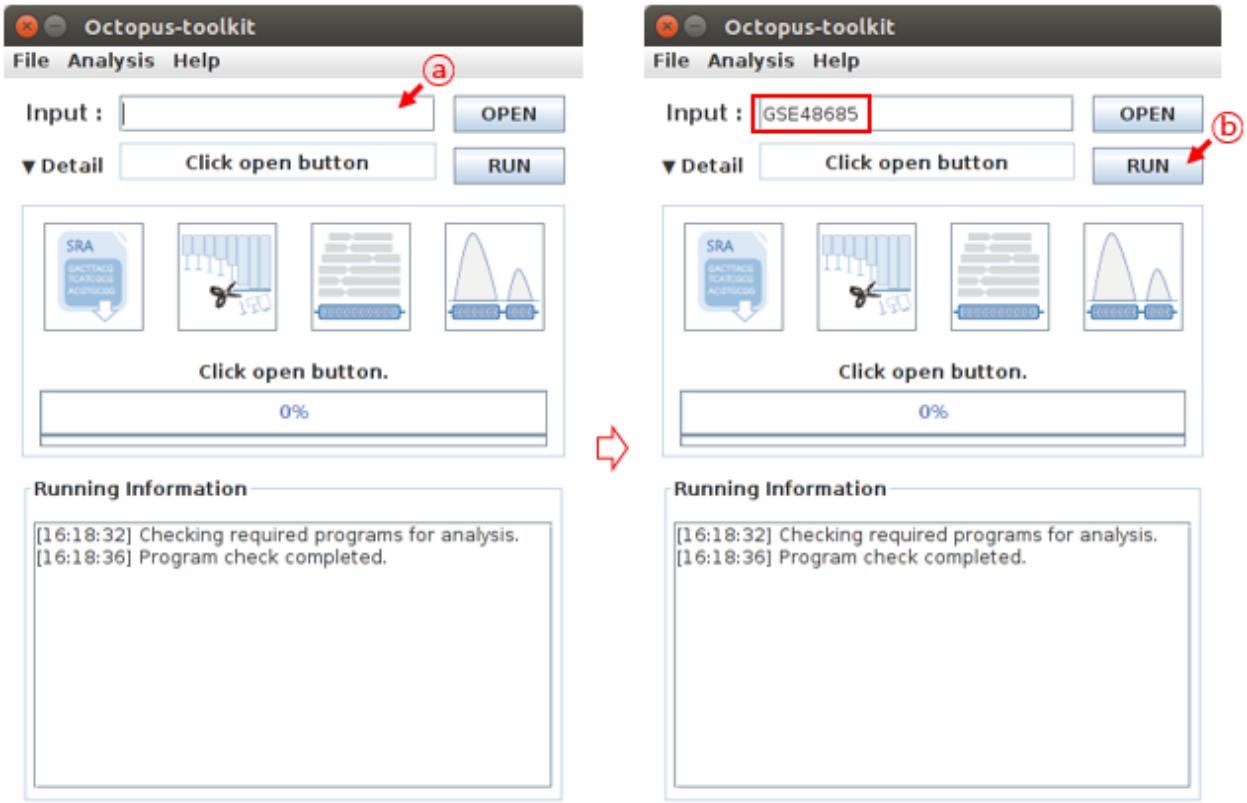
Tutorial ID	Description
6-1.Public data	How to analyze a public data with `GEO accession number`.
6-2.Public data	How to analyze a list of public data with a text file containing the `list` of GEO accession numbers.
6-3.Private data	How to setup Private table for `user's data`.
6-4.Private data	How to setup Private table when you have multiple files for a single sample: `Multi-lane`.
6-5.Peak Calling	How to identify peaks using `Peak Calling` with the output.
6-6.Graph	How to draw `Graph` with the output.
6-7.IGV	How to explore genome using `IGV` with the output.
6-8.Custom adapter sequence	how to use a `custom adapter sequence` generated by ownself.
6-9.Motif analysis	how to discover `de novo and known motif` using the output file of Octopus-toolkit.

6-1.Public data (Single GSE/GSM)

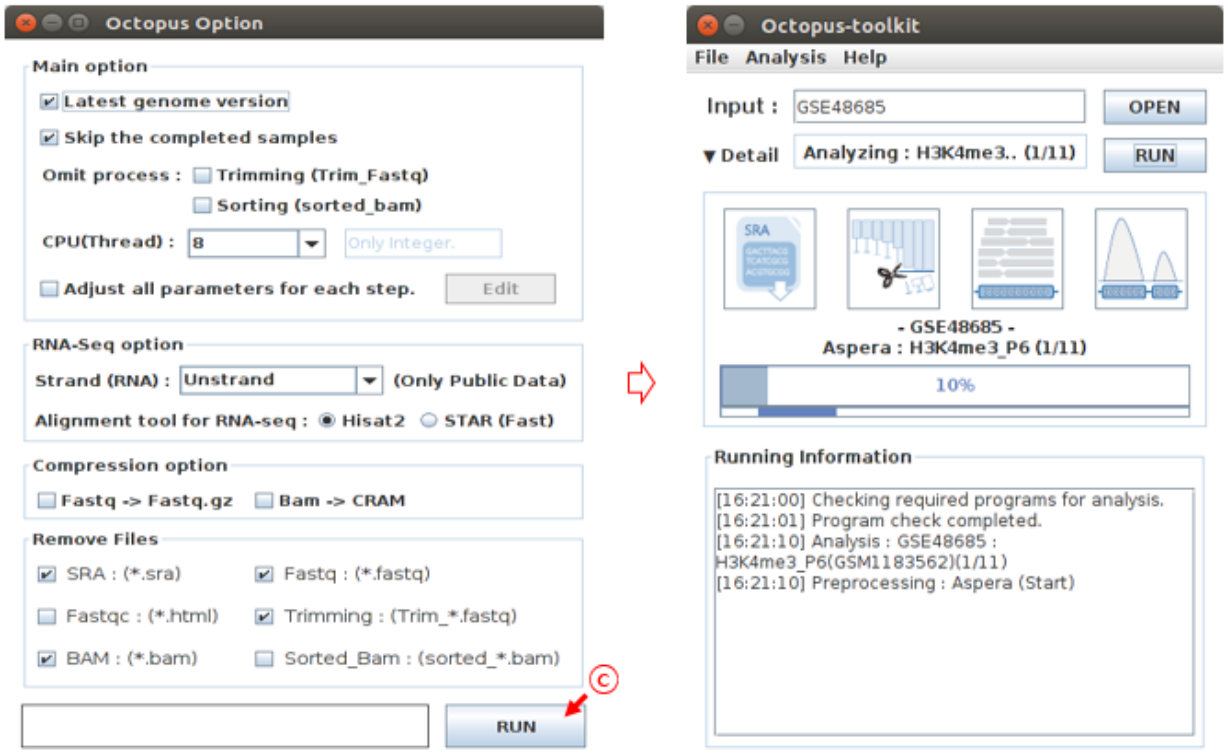
Note: `6-1.Public data (Single GSE/GSM)` describes how to process publicly available data by entering a single GEO accession number.

Analyzing published data is a simple process. Enter a GEO accession number in the input text area. Then click the Run button and Octopus-toolkit option window will appear. In the Option window, set the parameters for the analysis and click the RUN button to begin the analysis.

- `GEO accession number` : GSE48685 (ChIP-Seq:10, RNA-Seq:1)



- `A` : Enter GSE48685 in the input text area.
- `B` : Click the Run button



- `C` : Select the options to analyze and click the Run button. (Option : Defalut)

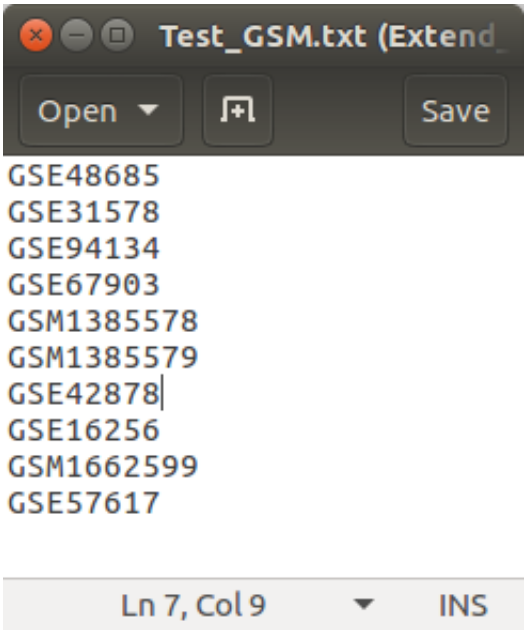
Finally, Octopus-toolkit will automatically download raw files in the GSE48685 ftp directory and subsequenty analyze the data. The output will be stored in a specified directory. No other action is required.

6-2.Public data (Multi GSE/GSM)

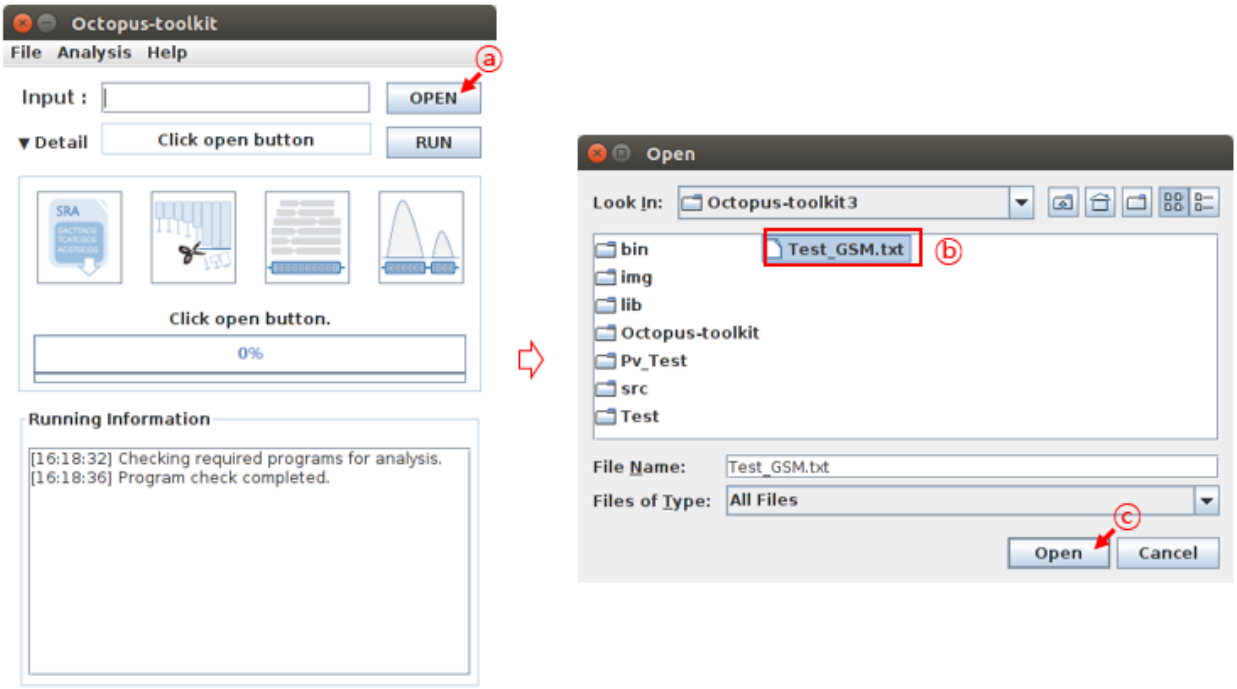
Note: `6-2.Public data (Multi GSE/GSM)` describes how to sequentially analyze a set of public data (a list of GSE accession numbers).

You may want to analyze samples (GSM) in a study (GSE) with several other studies (GSEs) altogether. In this case, you need to create a text file containing GSM ids for samples and GSE ids for studies.

An example is shown below. (`example.list`)

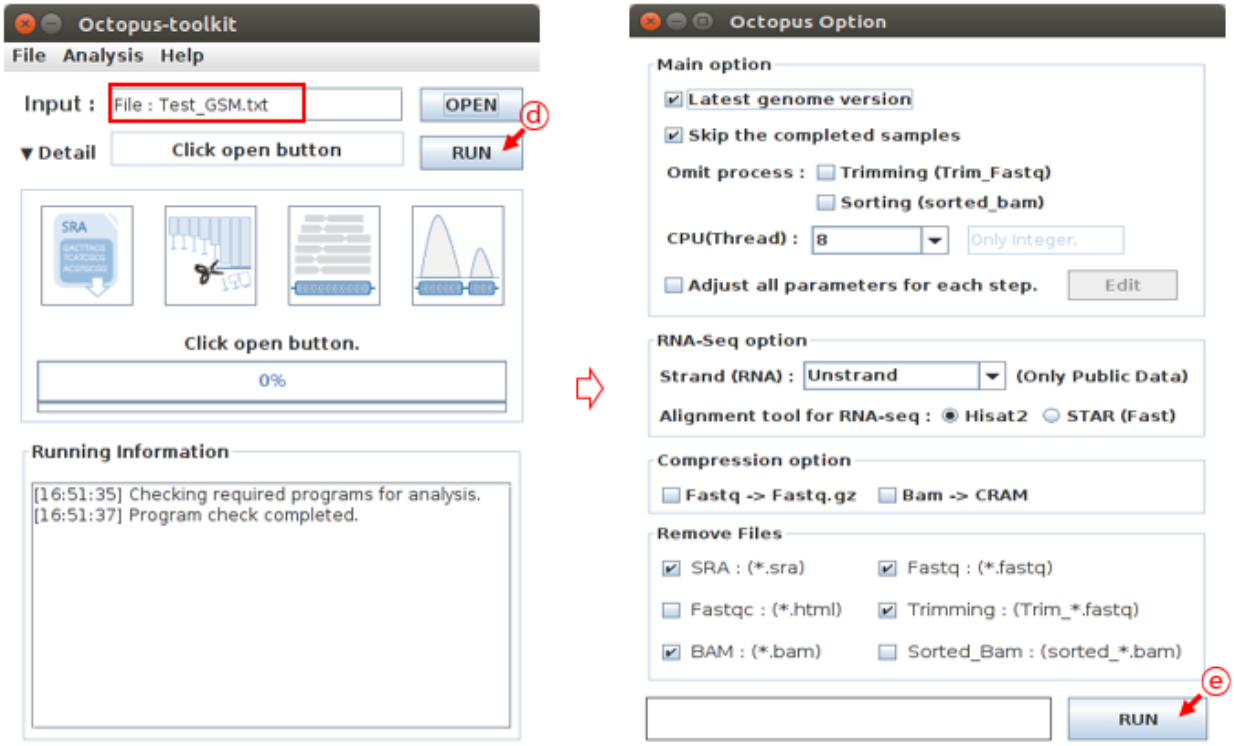


Then, click the OPEN button and select the list file you prepared.

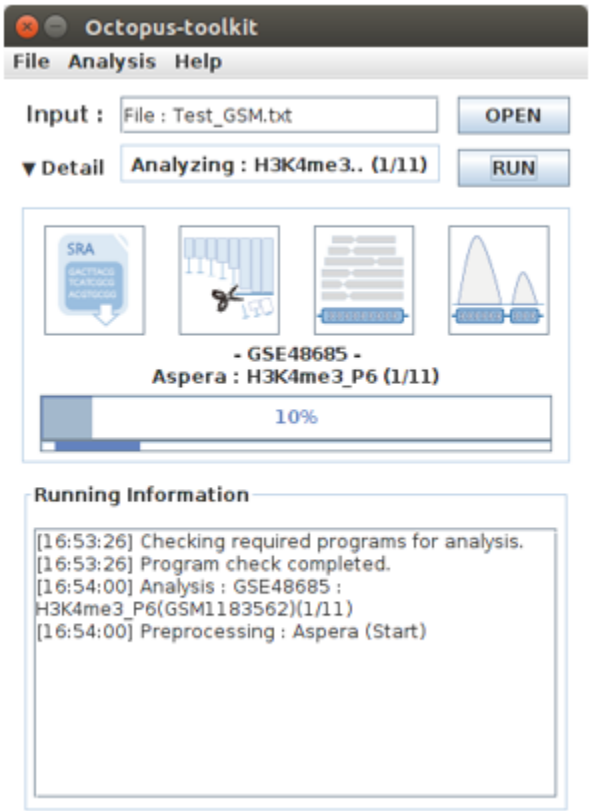


- A : Click the Open button
- B : Select the GEO accession number list file.
- C : Click the Open button

Then, click the RUN button. Octopus-toolkit option window will appear. In the Option window, set the parameters for the analysis and click the RUN button to begin the analysis.



- D : Click the RUN button
- E : Select the options to analyze and click the RUN button. (Option : Defalut)



Finally, Octopus-toolkit will automatically analyze the list of data. Sit back and relax until the results are out.

6-3.Private data (Basic)

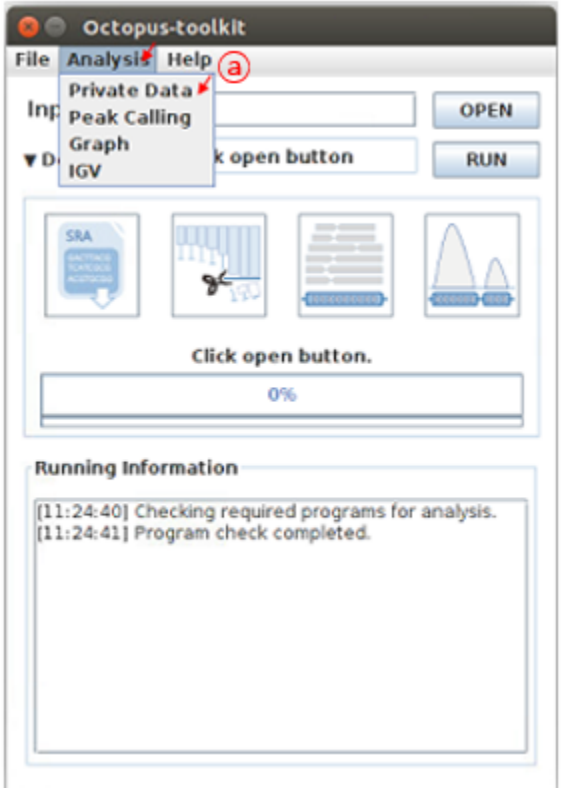
Note: `6-3.Private data (Basic)` describes how to analyze your own data using the same analysis pipeline for the public data.

Let's assume that you have the following data.

Table 2: Analysis situation.

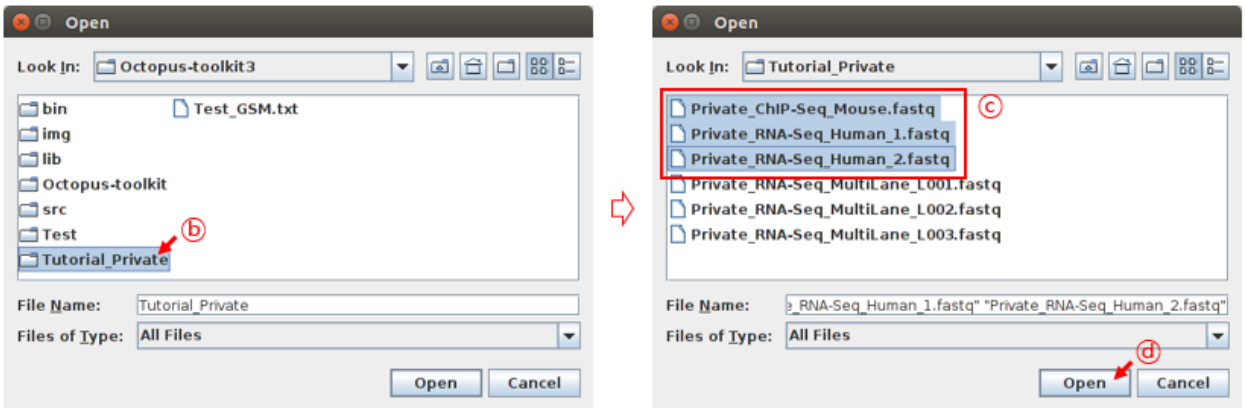
NO	File name	Genome	Seq Type	SE or PE	Strand
1	Private_ChIP-Seq_Mouse.fastq	mm10	ChIP-Seq	Single-End	Not use
2	Private_RNA-Seq_Human_1.fastq	hg38	RNA-Seq	Paired-End	FR-Firststrand
3	Private_RNA-Seq_Human_2.fastq	hg38	RNA-Seq	Paired-End	FR-Firststrand

First, open the Analysis tab and then, click Private data function.



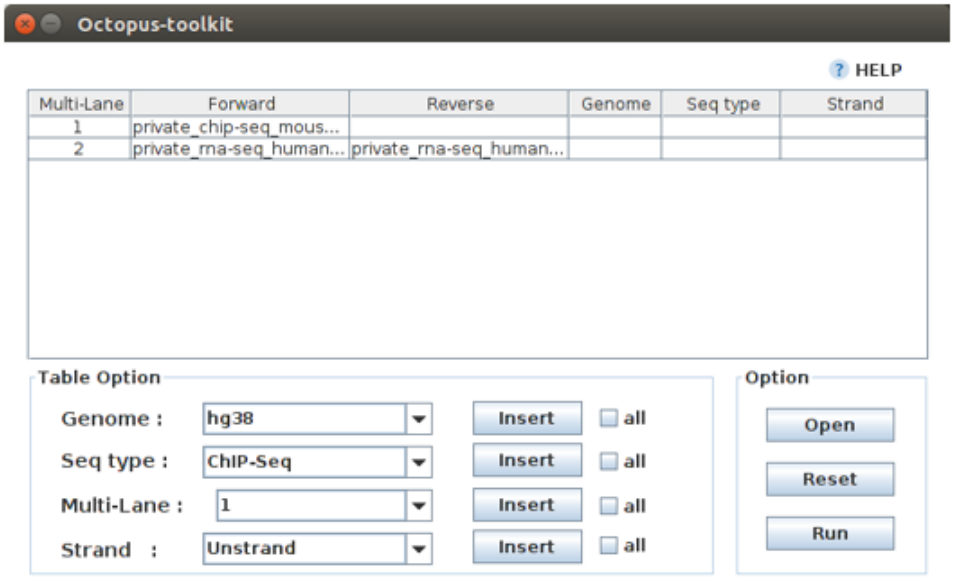
- A : Click the Private Data function in the Analysis menu bar.

Select your fastq files and click the Open button.

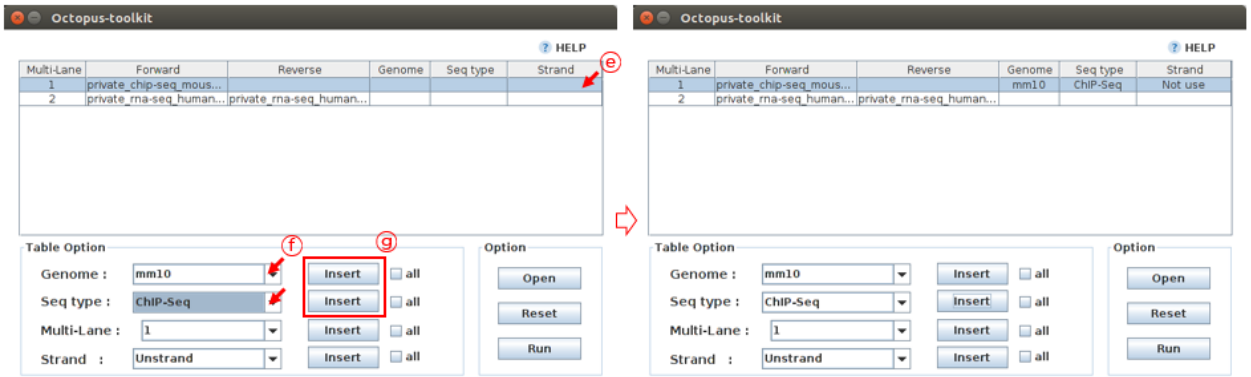


- B : select the folder
- C : Select the files
- D : Click the Open button

The following Private Table window will appear.



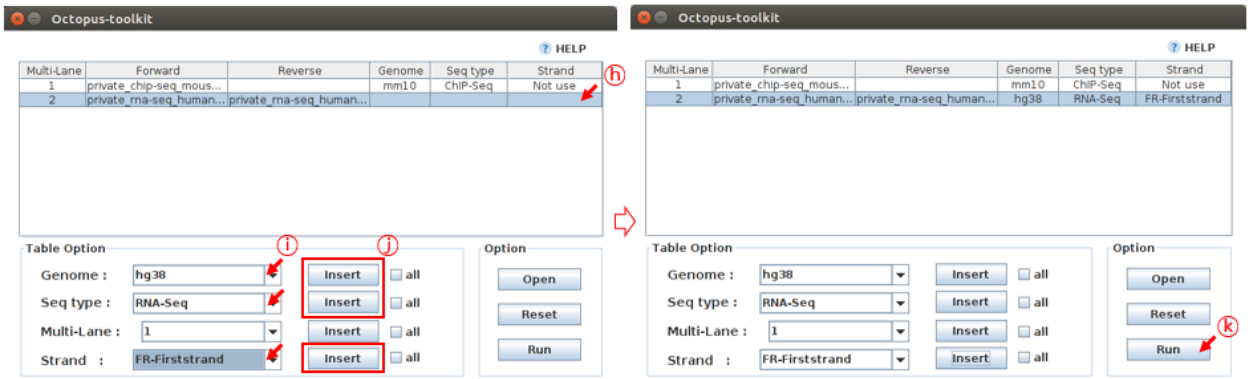
Case 1. Fill in the blank for the 1.Private_ChIP-Seq_Mouse fastq file. Reads in this ChIP-seq file (single-end) should be mapped to the mm10 genome.



- `E` : Select the Private_ChIP-Seq_Mouse.fastq sample.
- `F` : Select appropriate parameters regarding this sample. (Genome : `mm10`, Seq-Type : `ChIP-Seq`)
- `G` : Click the Insert button

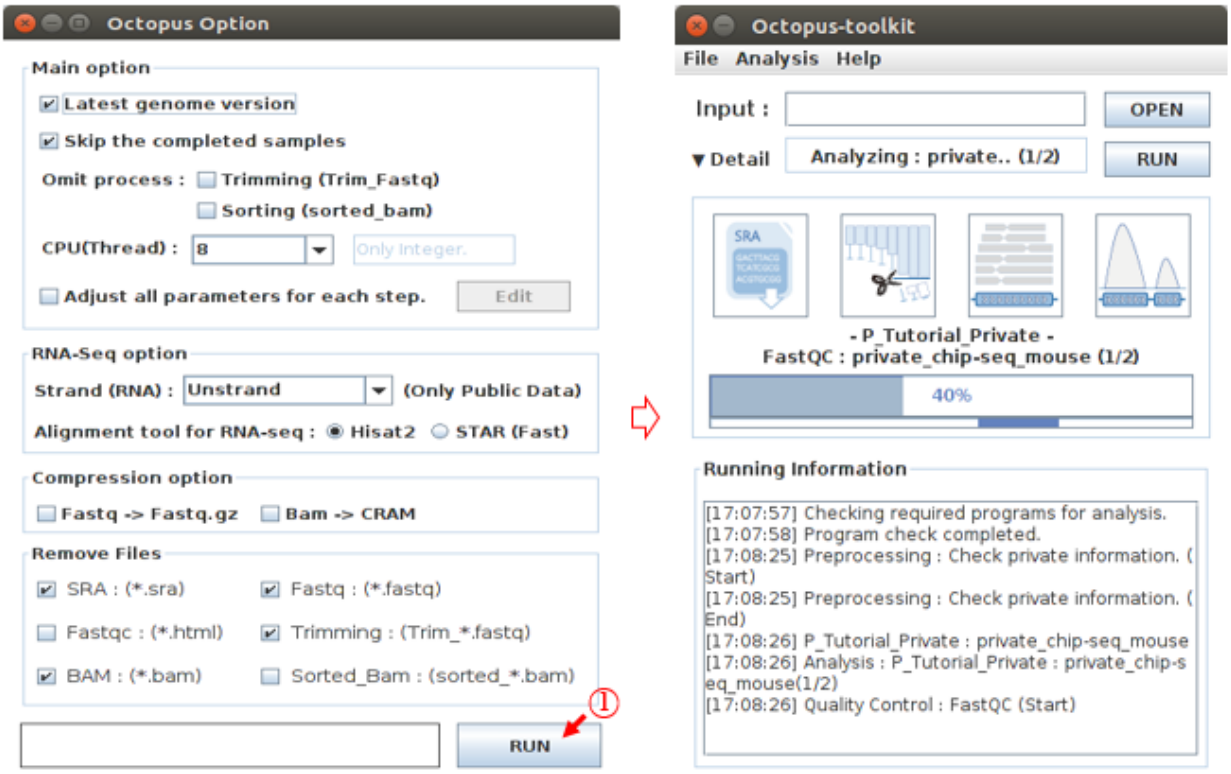
Case 2. Fill in the blank for the 2 and 3.Private_RNA-Seq_Human fastq files. Reads in this RNA-seq files (paired-end, FR-Firststrand) should be mapped to the hg38 genome.

Octopus-toolkit automatically recognizes Paired-End files. The name of the files must be the same and end with the suffix _1.fastq and _2.fastq



- H : Select the Private_RNA-Seq_Human.fastq sample.
- I : Select information about this sample. (Genome : hg38, Seq-Type : RNA-Seq, Strand : FR-Firststrand)
- J : Click the Insert button
- K : Click the Run button

The Octopus-toolkit option window will appear. In the Option window, set the parameters for the analysis and click the RUN button to begin the analysis.



- L : Click the Run button.

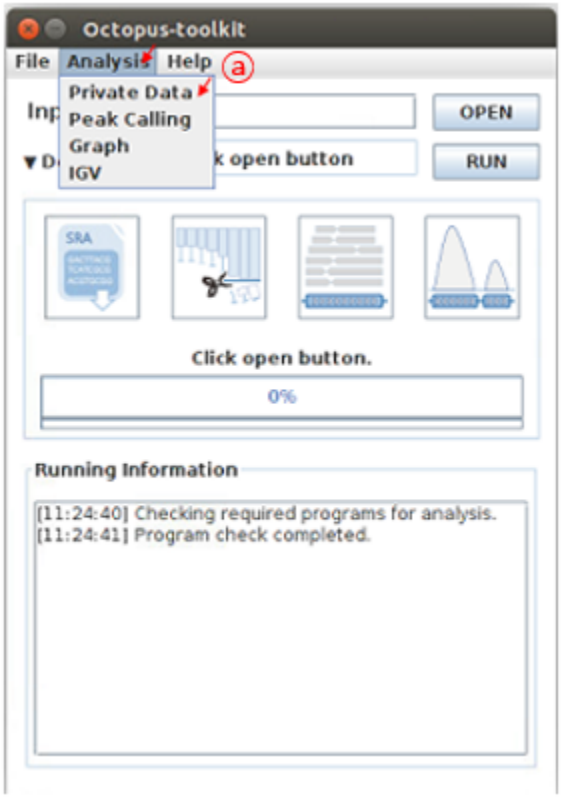
6-4.Private data (Multi-lane)

Note: `6-4.Private data (Multi-lane)` describes how to process your samples from multe lanes.

Table 3: Analysis situation.

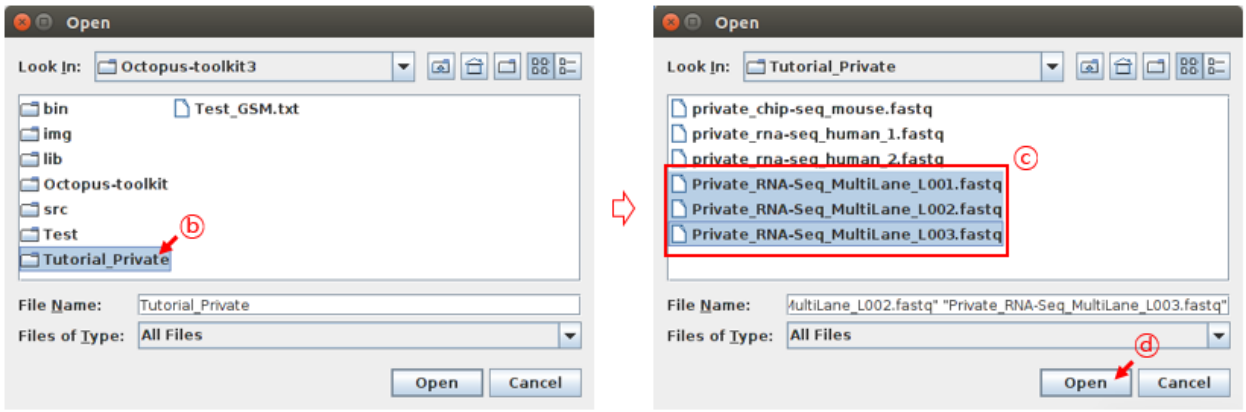
NO	File name	Genome	Seq Type	SE or PE	Strand
1	Private_ChIP-Seq_MultiLane_L001.fastq	hg38	ChIP-Seq	Single-End	Not use
2	Private_ChIP-Seq_MultiLane_L002.fastq	hg38	ChIP-Seq	Single-End	Not use
3	Private_ChIP-Seq_MultiLane_L003.fastq	hg38	ChIP-Seq	Single-End	Not use

First, open the Analysis tab and then, click Private data function.



- `A` : Click the Private Data in the Analysis menu bar.

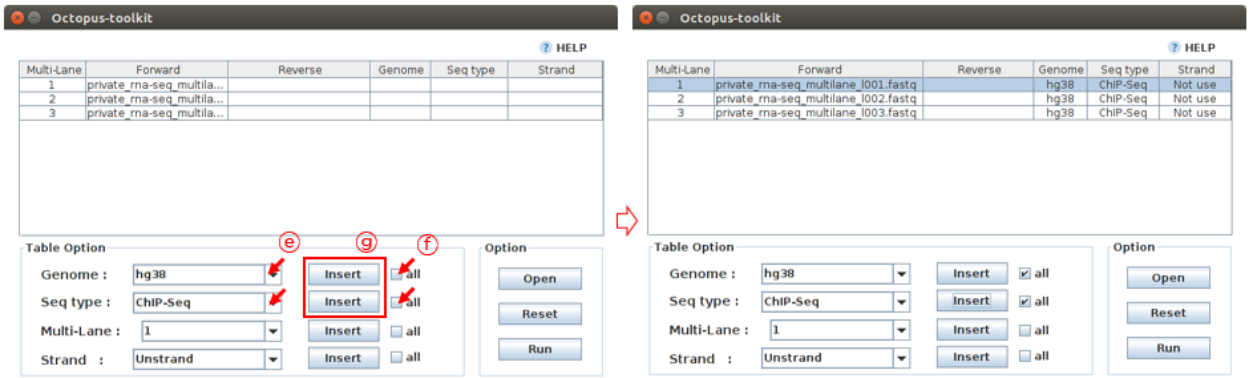
Select your fastq (multi-lane) files and click the Open button.



- B : select the folder
- C : Select the files
- D : Click the Open button

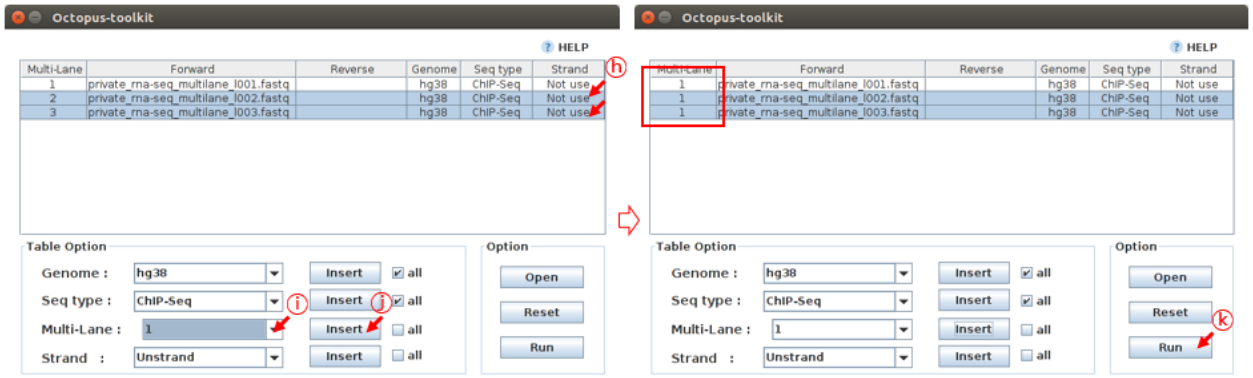
The following Private Table window will appear.

Case 1. let's fill in the blank for the Private_ChIP-Seq_MultiLane fastq file. Reads in these ChIP-seq files (single-end) should be mapped to the hg38 genome. Since all samples have the same information, you can use the all button to enter the same information at once.



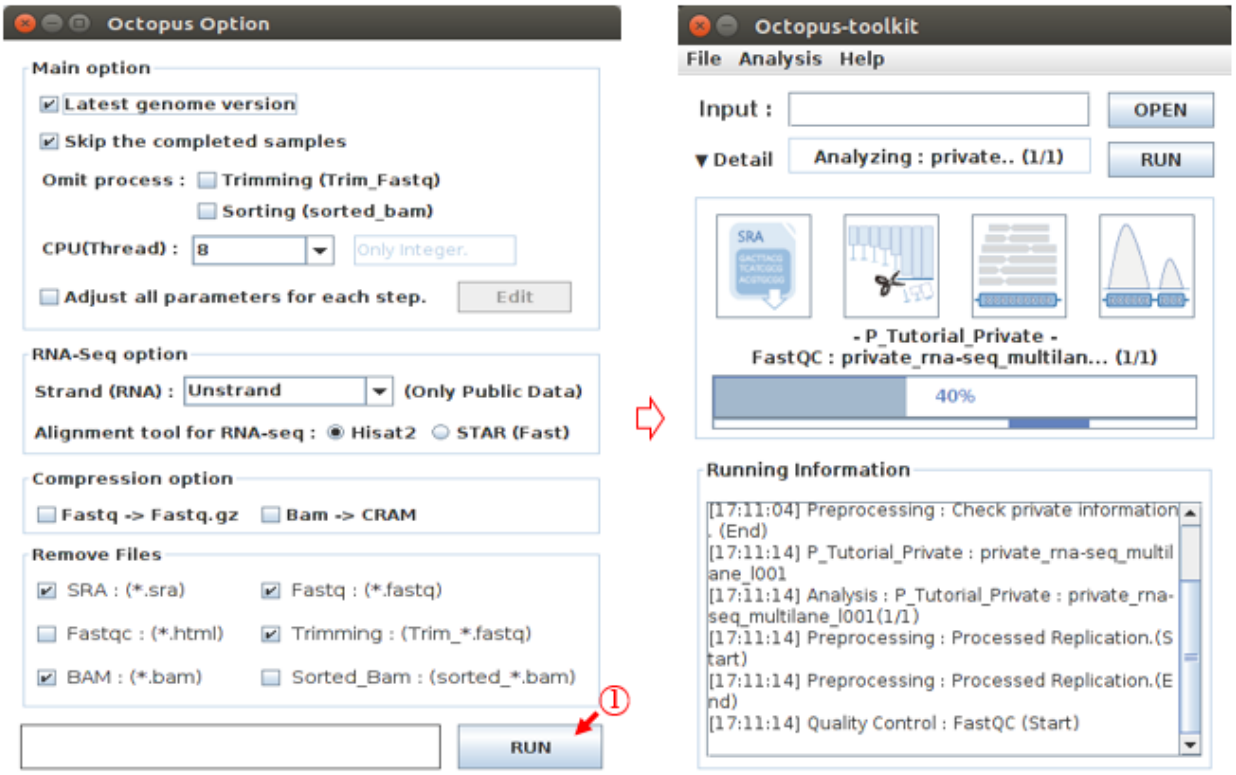
- E : Select information about this sample. (Genome : `hg38`, Seq-Type : `ChIP-Seq`)
- F : Click the all button
- G : Click the Insert button

Octopus-toolkit will merge the files with the same number in the Multi-Lane column prior to analysis. Please carefully assign the same number to multi-lane fastq files.



- H : Select the Private_RNA-Seq_MultiLane Files.
- I : Select the number 1 (Multi-Lane)
- J : Click the Insert button
- K : Click the Run button

The Octopus-toolkit option window will appear. In the Option window, set the parameters for the analysis and click the RUN button to begin the analysis.



- L : Click the Run button

6-5.Peak Calling

Note: `6-5.Peak Calling` describes how to identify peaks (enriched regions by mapped reads) with the Octopus-toolkit output.

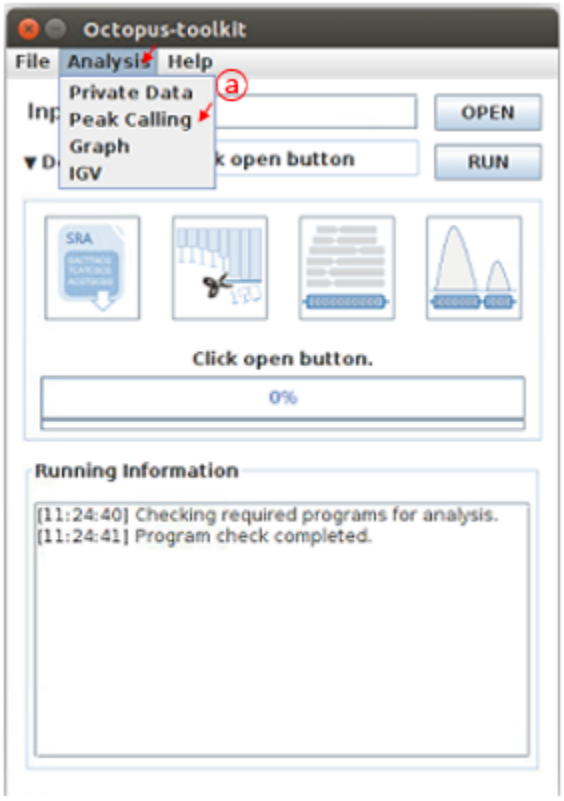
You can identify peaks from the output: 6-1 ~ 6-4.

Let's say you have the following ChIP-seq data.

Table 4: Analysis situation.

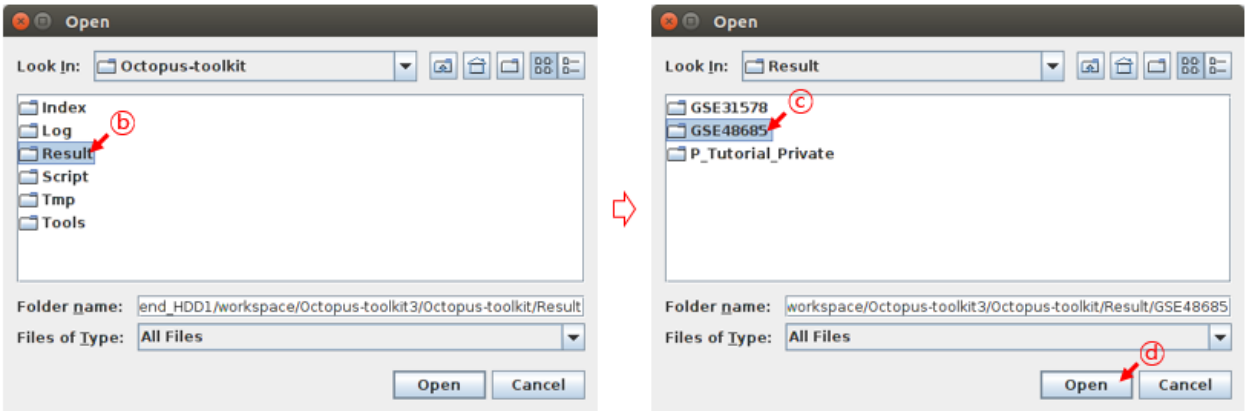
NO	Sample name	Input/Control/IgG	Style	Result Path
1	STAT5A_P6	Input_P6	Transcription Factor	Result/GSE48685

First, open the Analysis tab and then, click the Peak Calling function.



- `A` : Click the Peak Calling in the Analysis menu bar.

Octopus-toolkit output will be stored in the Result folder. You need to select an appropriate study (GSE directory) in the Result folder. For example, select the GSE48685 directory.

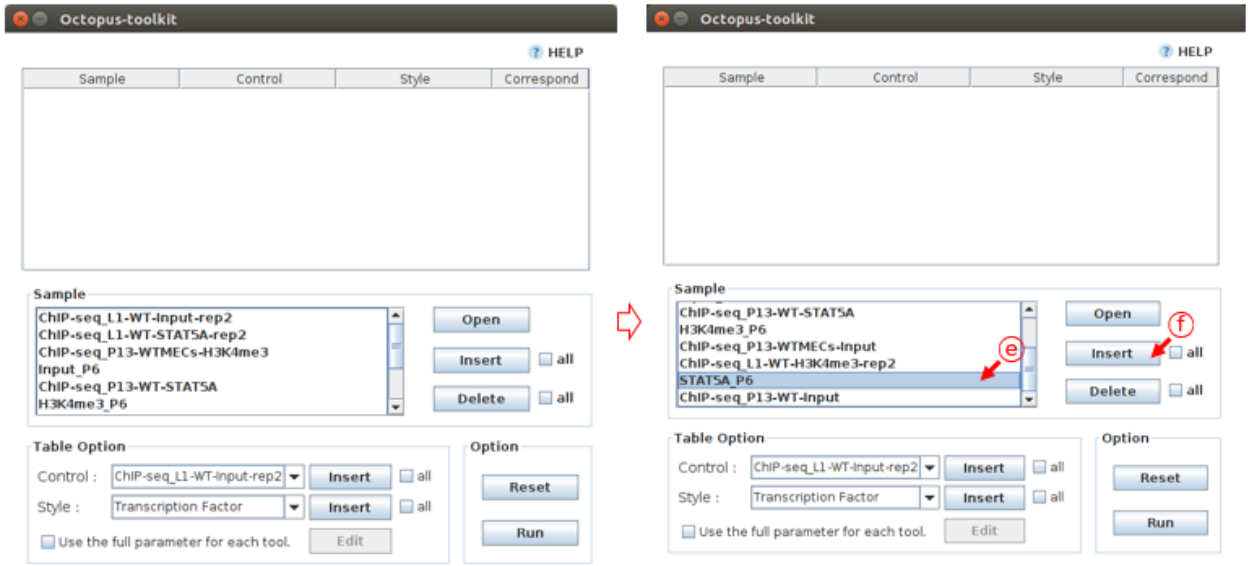


- B : Select the Result folder.
- C : Select the GSE48685 folder.
- D : Click the Open button.

Once you select an GSE folder (not double click), please click the Open button. Then, the Peak Calling Table will appear.

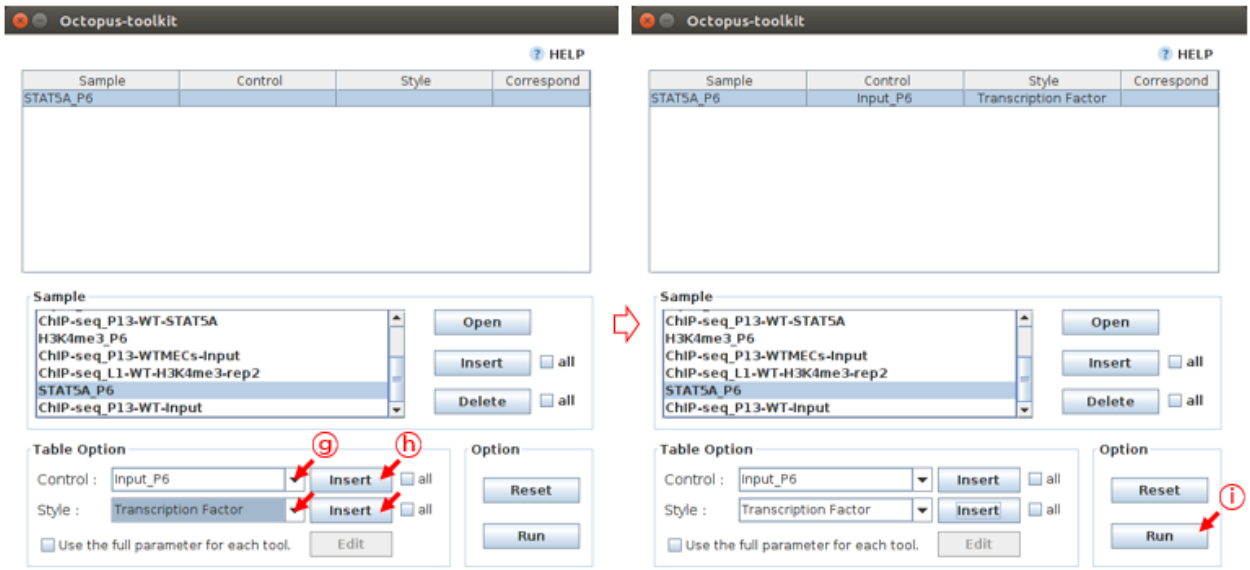
Samples of GSE48685, which were processed by Octopus-toolkit, will appear in the Sample area.

First, you need to add the processed samples to the Peak Calling table using the Insert function.



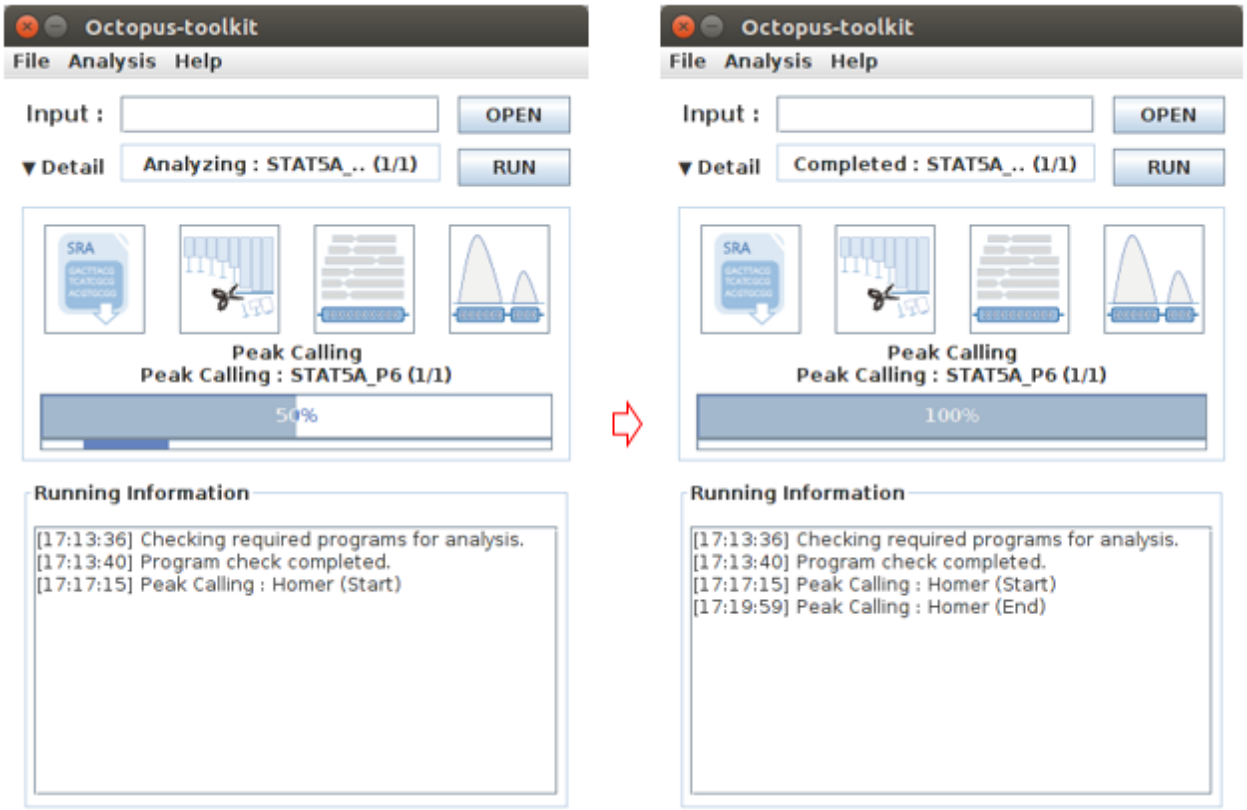
- E : Select the STAT5A_P6
- F : Click the Insert button

Then fill in the blanks for the selected samples using the Table option function. If there is a control (Control) sample to filter out background noise, you also need to add it to the Correspond column.

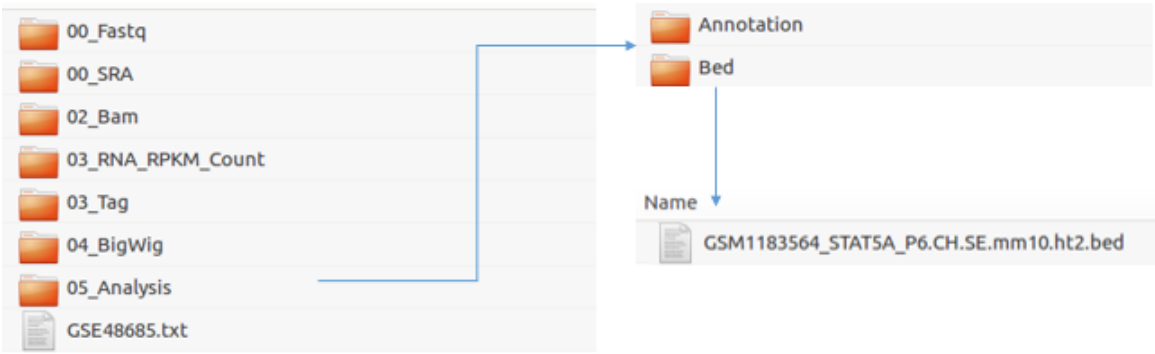


- `G` : Select the information about STAT5A_P6 (Control : `Input_P6`, Style : `Transcription Factor`)
- `H` : Click the Insert button
- `I` : Click the Run button

Peak Calling analysis will start according to the Table information.



Once completed, you can find the result files (.bed for peaks) in the 05_Analysis directory in the Result/GSE48685 directory.



- `Result Path` : Octopus-toolkit/Result/GSE48685

6-6.Graph

Note: `6-6.Graph` describes how to draw plots with the output: 6-1 ~ 6-5.

You can draw a heatmap and line plots with a few clicks.

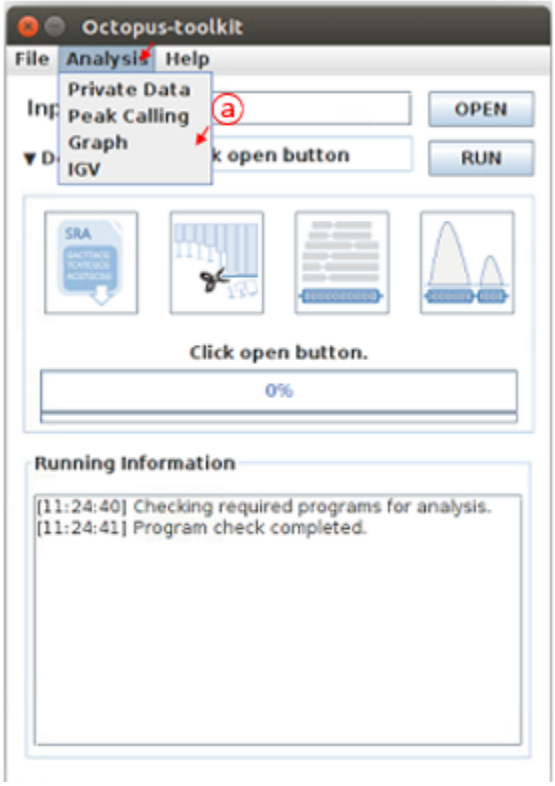
6-6.Graph tutorial describes how to draw plots for multiple outputs. Let say you have the following outputs processed by Octopus-toolkit.

Table 5: Analysis situation.

NO	Sample name	Peak(.bed)
1	STAT5A_P6	O
2	M_Bcl6_rep2_G50	X
3	MH_STAT5_rep2_G41	X

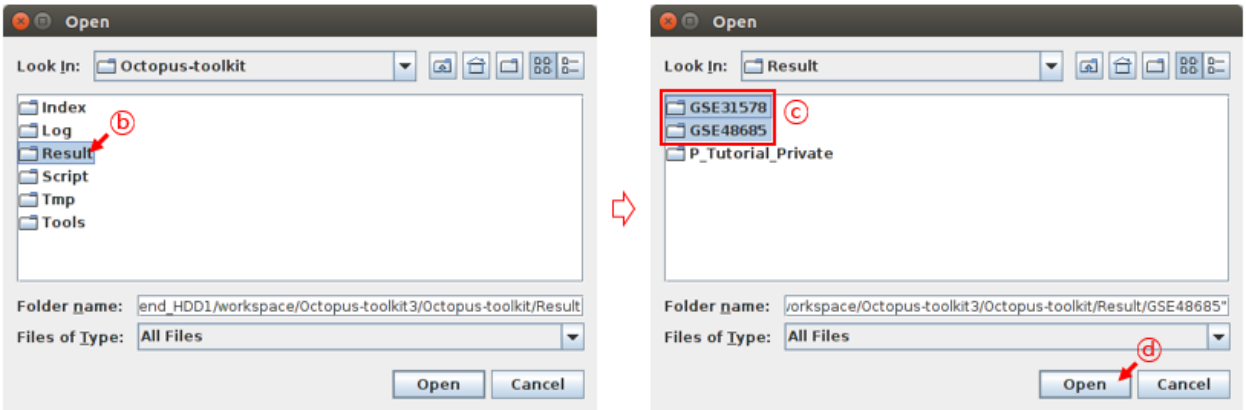
- Option : +- `1000 bp` based on TSS, Bin Size : `100`

First, open the Analysis tab and then, click the Graph function.



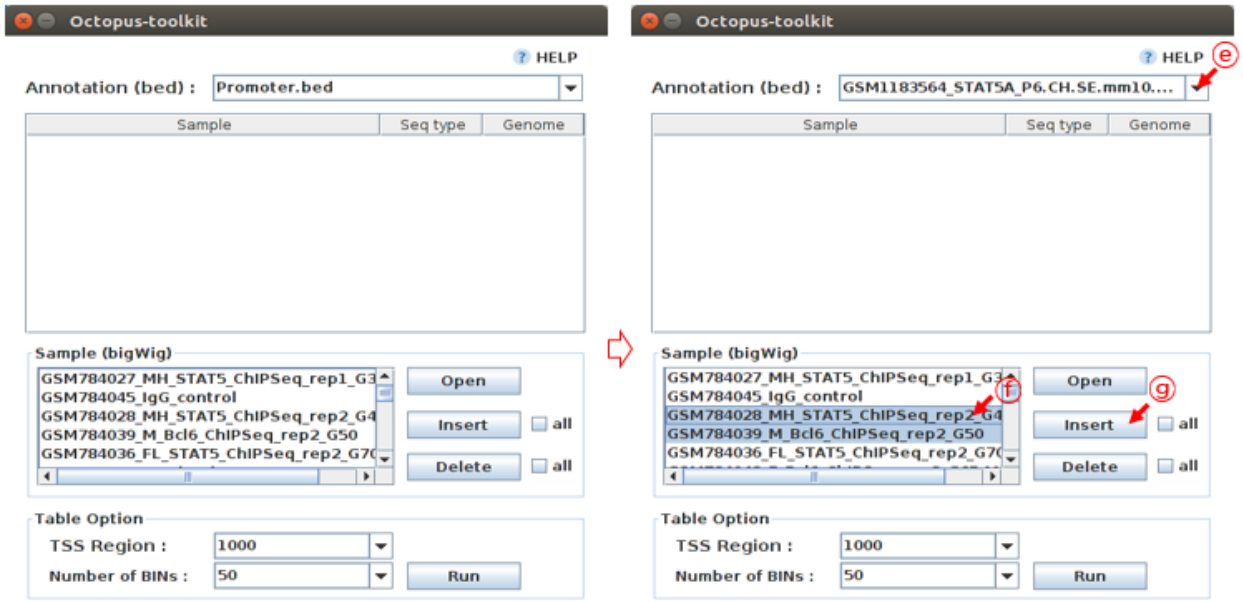
- A : Click the Graph in the Analysis menu bar.

Octopus-toolkit output will be stored in the Result folder. To draw heatmap and plot, you need to select appropriate studies (GSE directories) in the Result folder. For example, select the GSE48685 and GSE31578 directories.



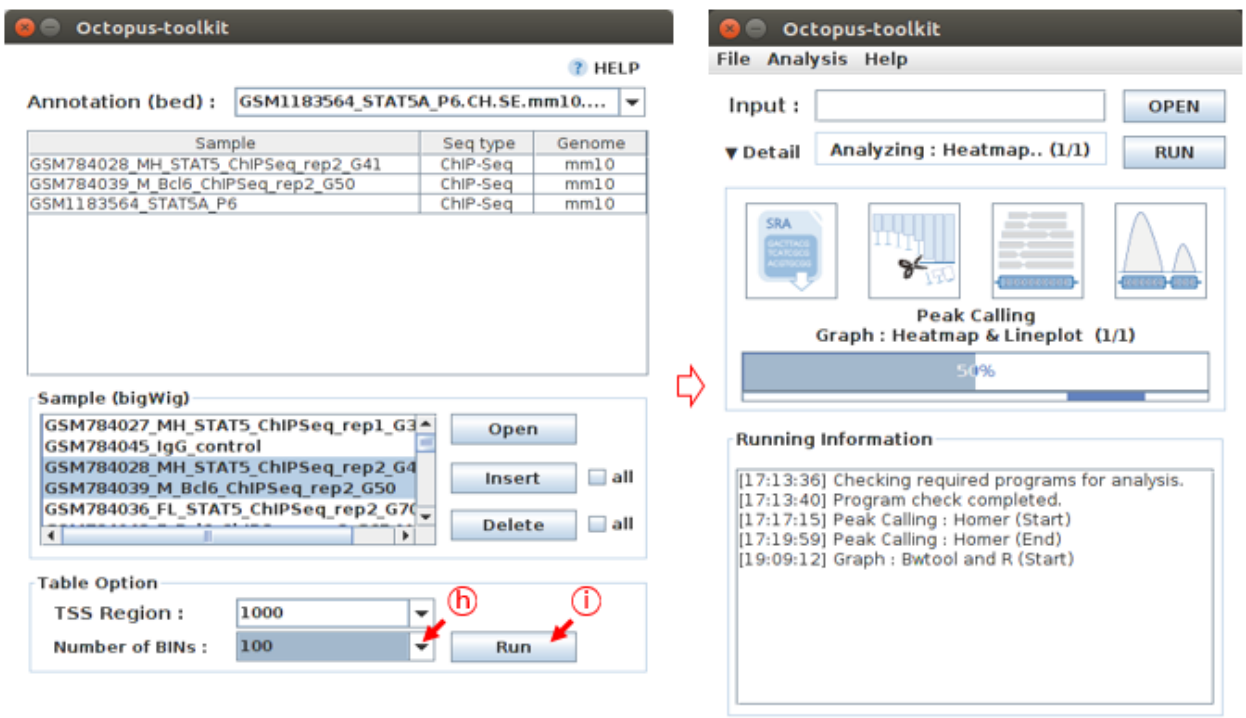
- B : Select the Result folder.
- C : Select the GSE48685 and GSE31578 folders.
- D : Click the Open button.

The heatmap and plot will be drawn based on an annotation file (reference). The default annotation file (.bed) contains promoter regions. You can replace it with peak file (.bed) generated by Octopus-toolkit if you perform the peak calling analysis.



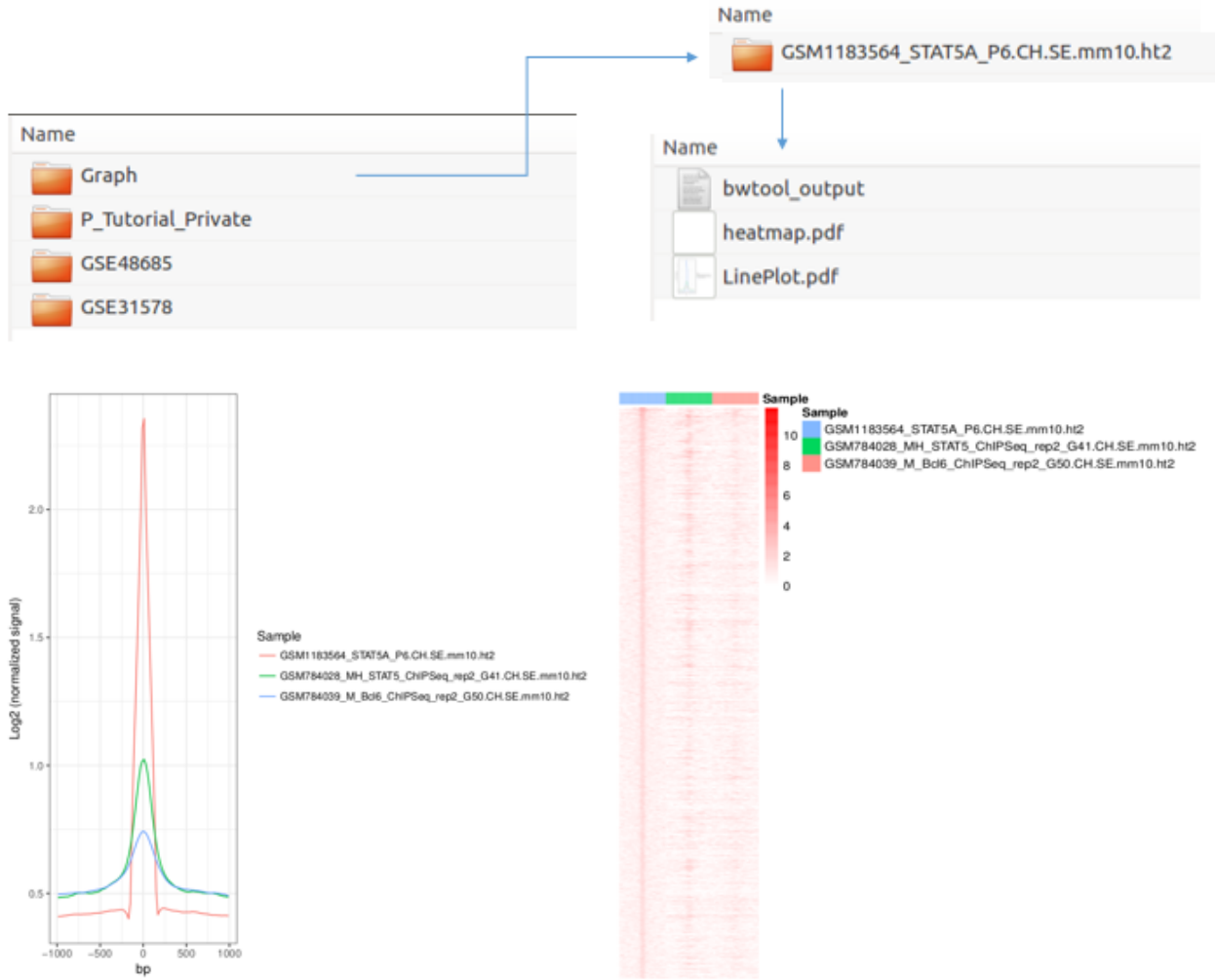
- E : Select STAT5A_P6_CH.SE.mm10 as the reference.
- F : Select samples of your interest from the list.
- G : Click the Insert button.

In the Table option, Adjust TSS region (bp) and Number of BINs (resolution) parameters. Click the Run button to perform the Graph analysis.



- H : Select the 1000 in TSS region and 100 in Number of BINs
- I : Click the Run button

Heatmap and plot will be stored in the Result/Graph folder.

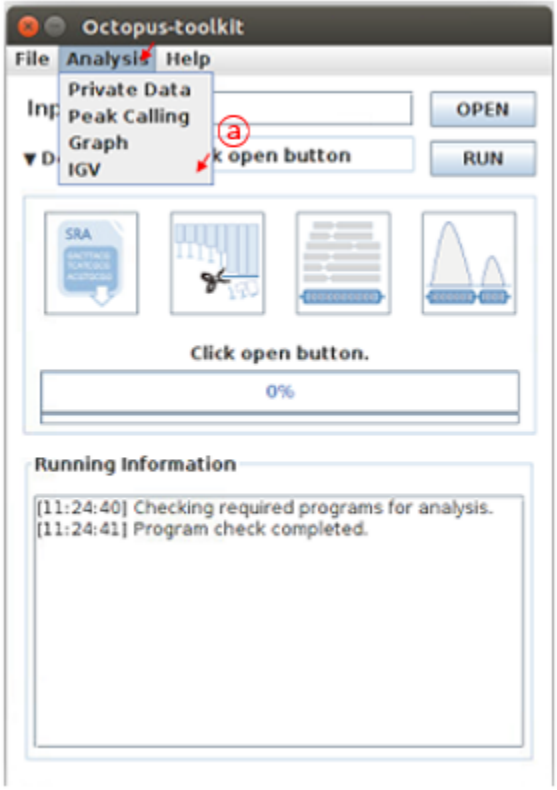


6-7.IGV

Note: `6-7.IGV` describes how to visualize genomes with data (bigWig files) via IGV.

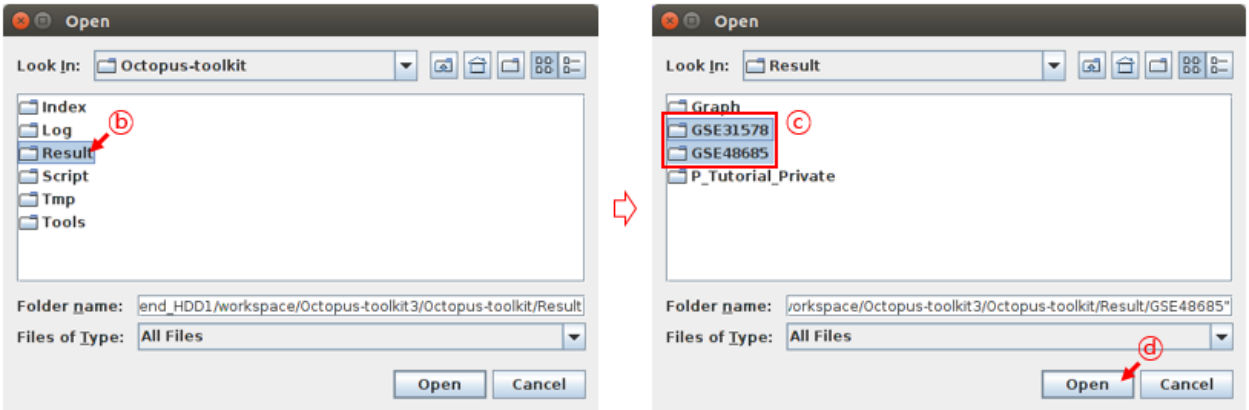
Octopus-toolkit generates bigWig files which can be visualized using Integrative Genomics Viewer(IGV).

First, open the Analysis tab and then, click the IGV function.



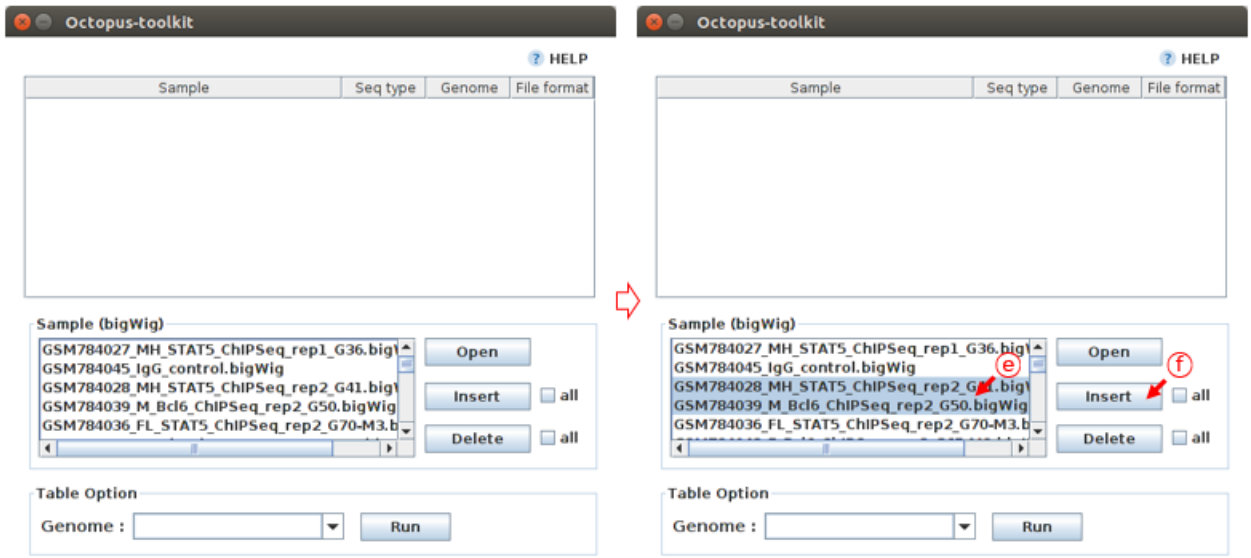
- A : Click the IGV in the Analysis menu bar.

You need to select appropriate studies (GSE directories) in the Result folder. For example, select the GSE48685 and GSE31578 directories. It will load all bigWig files in the selected directories.



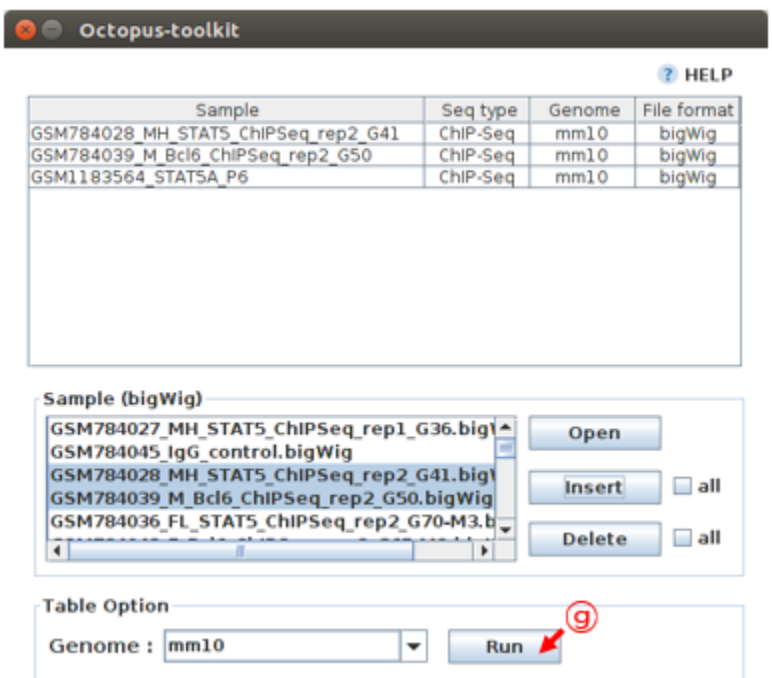
- B : Select the Result folder.
- C : Select the GSE48685 and GSE31578 folders.
- D : Click the Open button.

Let's say you select the following samples. You must select an appropriate genome for visualization. Obviously, you cannot load bigWig files from different genomes.



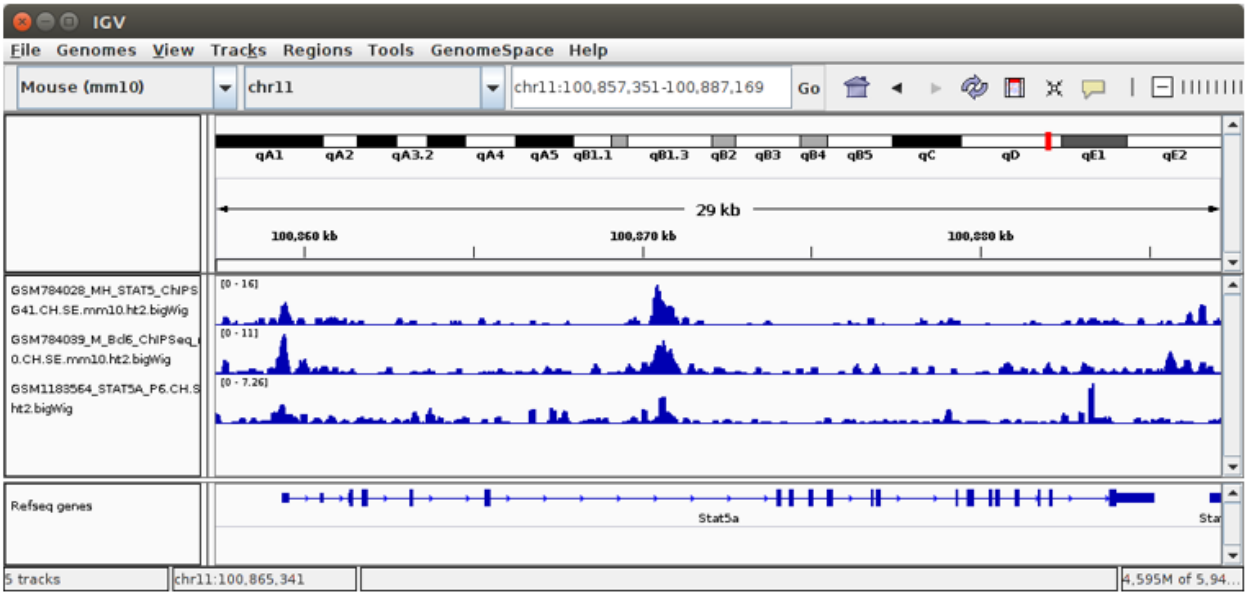
- E : Select samples.
- F : Click the Insert button.

Click the Run button to start the Graph analysis.



- G : Click the Run button.

Depending on the number and size of data, it may take some time to load those files onto the IGV. Please take your time.



6-8.User's custom adapter sequence(Trimming)

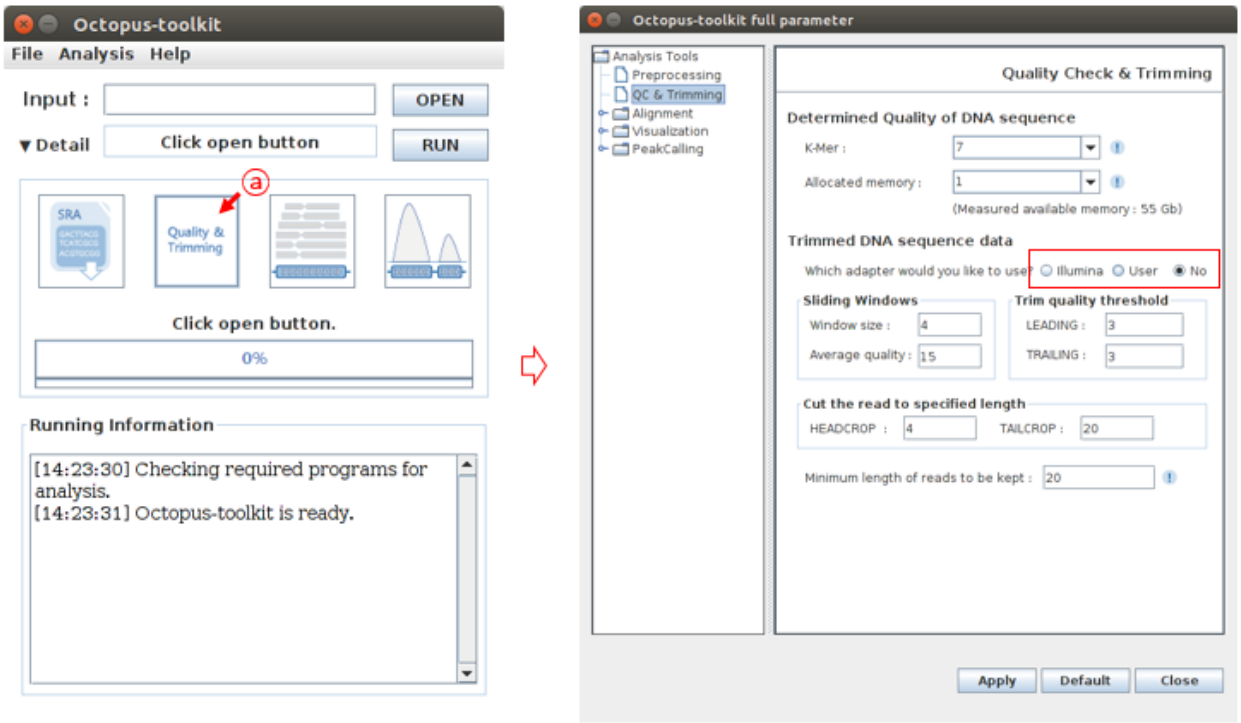
Note: `6-8.IGV` describes how to use a custom adapter sequence generated by ownself.

Typically, The user uses the adapter sequence provided Trimmomatic. But some users want to use custom adapter sequence generated by ownself.

First, The user should make a custom adapter sequence file. The format of files equals a single or multiple sequence file. (File name extension is `.fasta` and `.fa`) (`Custom_adapter.fasta`)

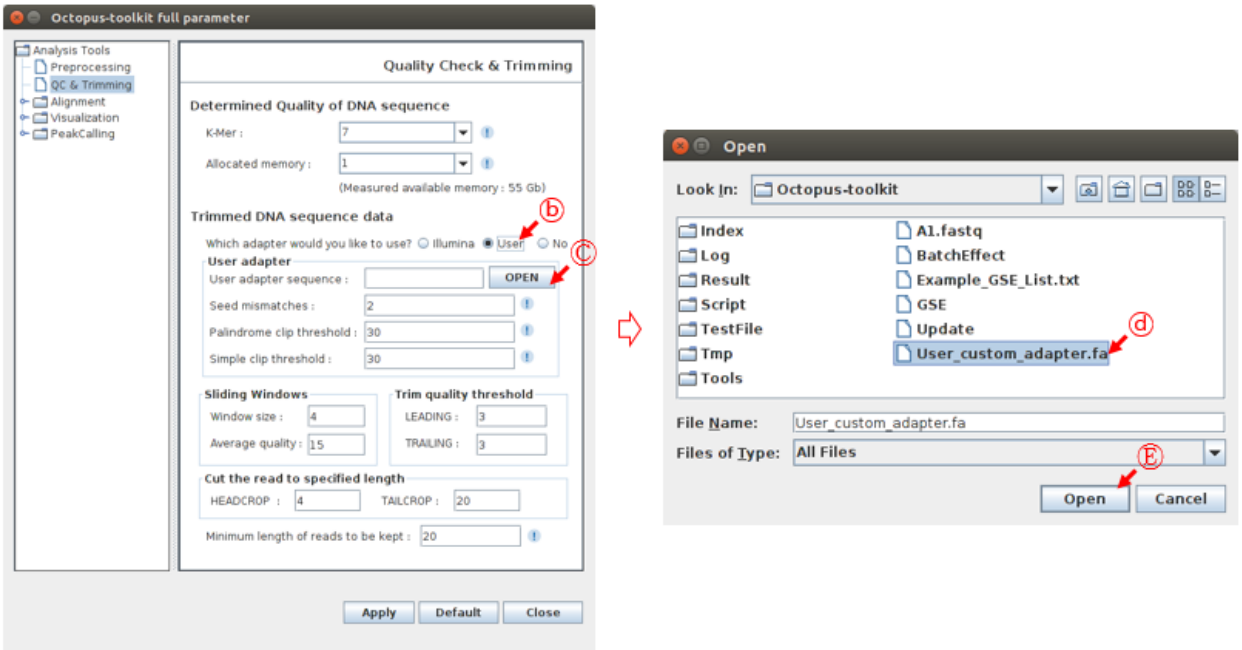
For more detailed information, please refer to the link below. (Trimmomatic : How to make the adapter fasta)

Second, click the Quality & Trimming button in Octopus-toolkit.



- A : Click the Quality & Trimming button.

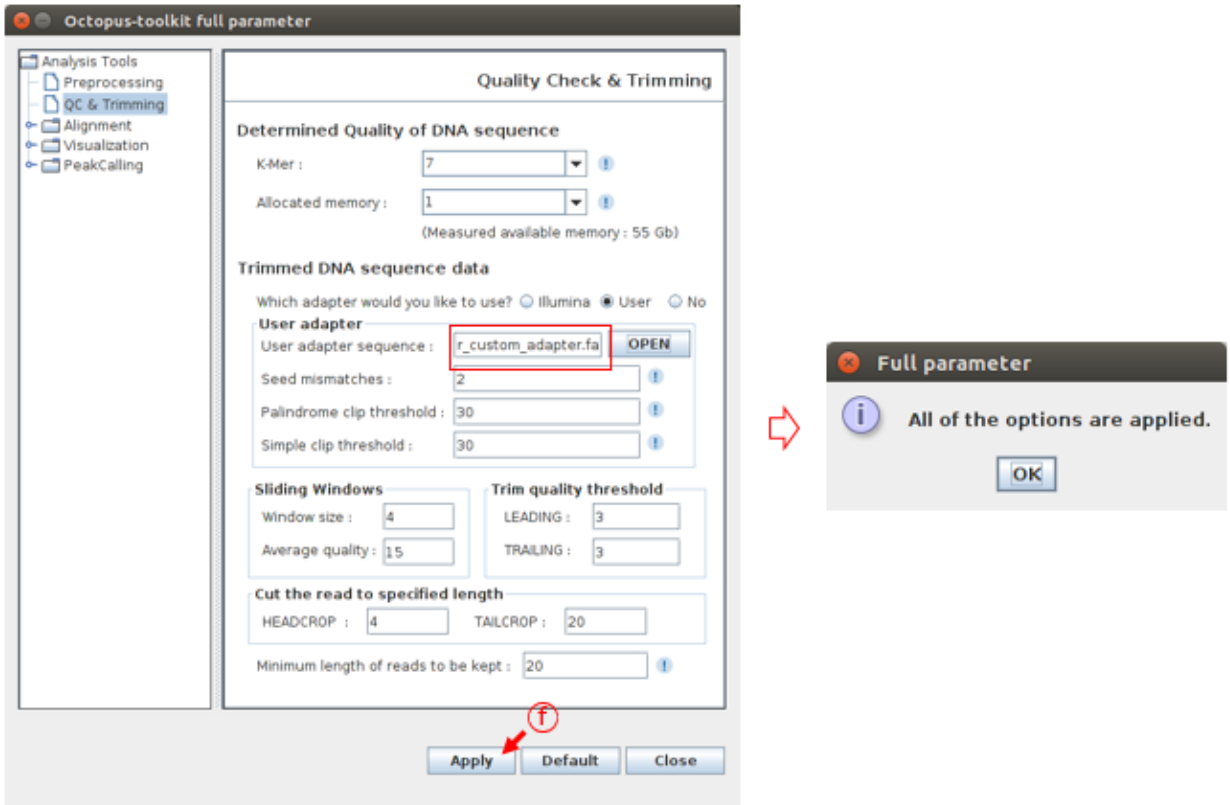
To open the custom adapter sequence file, Select a User radio button and click the Open button. You need to select your adapter sequence file in your computer.



- B : Click the User radio button.
- C : Click the Open button.
- D : Select the custom adapter sequence file generated by user.

- `E` : Click the Open button

Click the Apply button to apply the custom adapter sequence.



- `F` : Check the Apply button.

6-9.Motif analysis

Note: `6-9.Motif analysis` describes how to discover de novo and known motif using the output file of Octopus-toolkit : 6-1 ~ 6-5.

Octopus-toolkit is not supporting a motif analysis yet.

The user can analyze de novo and known motif using below command before to be completing development about motif analysis.

We will use a bed format file, which is generated by peak calling in Octopus-toolkit, for discovering motif.

Table 6: Test environment.

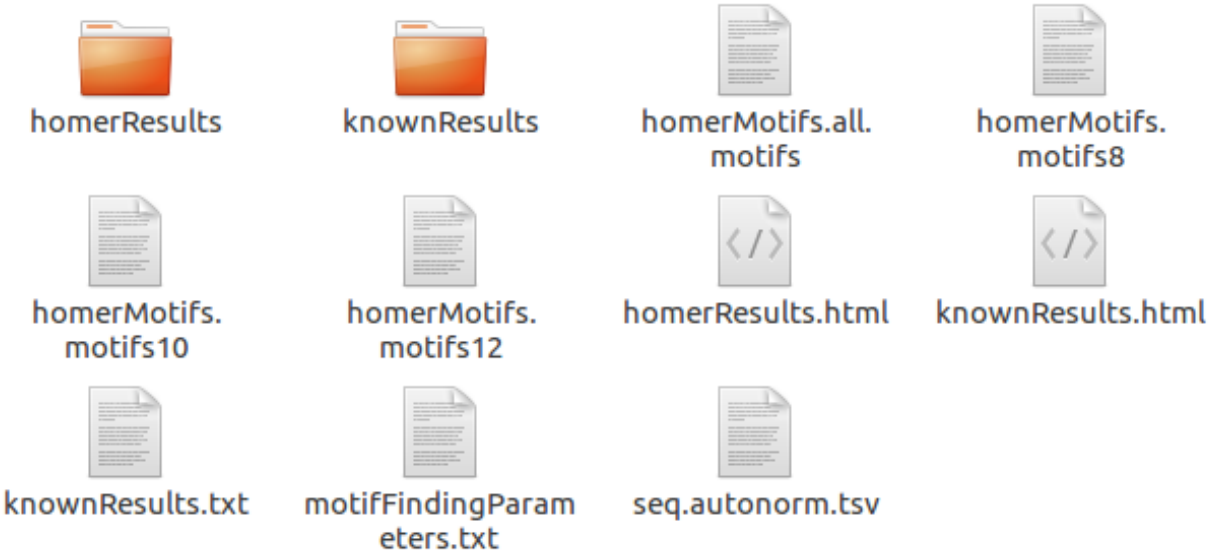
NO	command	Description
1	Pathway of Octopus-toolkit	/home/user_id/Octopus-toolkit/
2	user_id	octopus
3	Output of Octopus-toolkit	GSE48685
4	Input file like bed format file	05_Analysis/Bed/GSM1183564_STAT5A_P6.CH.SE.mm10.ht
5	Genome	mm10

- The `command` for `Motif` analysis:

```
// Add the Homer pathway
export Octopus_Homer="/home/user_id/Octopus-toolkit/Tools/Homer/bin"
export PATH=$PATH:$Octopus_Homer

cd /home/octopus/Octopus-toolkit/Result/GSE48685/
mkdir 06_Motif_output
/home/octopus/Octopus-toolkit/Tools/Homer/bin/findMotifsGenome.pl 05_Analysis/Bed/
↪GSM1183564_STAT5A_P6.CH.SE.mm10.ht2.bed mm10 06_Motif_output/
```

- The output of the Motif analysis.

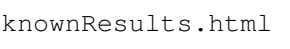


The output of Motif analysis provides a motif's letter-probability matrix, list of a detected motif, statistical value and best-matched gene symbol.

- `homerResults.html : De novo Motif`

Homer *de novo* Motif Results (06_Motif_output//)

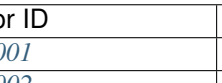
Known Motif Enrichment Results
Gene Ontology Enrichment Results
If Homer is having trouble matching a motif to a known motif, try copy/pasting the matrix file into STAMP
More information on motif finding results: HOMER | Description of Results | Tips
Total target sequences = 150014
Total background sequences = 150190
* - possible false positive

Rank	Motif	P-value	log P-pvalue	% of Targets	% of Background	STD(Bg STD)	Best Match/Details	Motif File
1		1e-668	-1.538e+03	7.09%	4.00%	50.5bp (58.0bp)	Stat3(Stat)/mES-Stat3-ChIP-Seq(GSE11431)/Homer(0.942) More Information | Similar Motifs Found	motif file (matrix)
2		1e-293	-6.747e+02	38.67%	34.14%	55.5bp (56.8bp)	Rbpj1(?)/Panc1-Rbpj1-ChIP-Seq(GSE47459)/Homer(0.785) More Information | Similar Motifs Found	motif file (matrix)
3		1e-231	-5.335e+02	42.69%	38.58%	55.1bp (57.9bp)	NFIA/MA0670.1/Jaspar(0.880) More Information | Similar Motifs Found	motif file (matrix)
4		1e-139	-3.217e+02	43.41%	40.21%	56.0bp (58.3bp)	RUNX1(Runt)/Jurkat-RUNX1-ChIP-Seq(GSE29180)/Homer(0.802) More Information | Similar Motifs Found	motif file (matrix)
5		1e-114	-2.639e+02	32.95%	30.22%	55.6bp (58.1bp)	Fox:Ebox(Forkhead,bHLH)/Panc1-Foxa2-ChIP-Seq(GSE47459)/Homer(0.711) More Information | Similar Motifs Found	motif file (matrix)
6		1e-102	-2.352e+02	26.99%	24.58%	54.9bp (57.0bp)	PB0165.1_Sox11_2/Jaspar(0.599) More Information | Similar Motifs Found	motif file (matrix)
7		1e-96	-2.233e+02	1.67%	1.07%	53.5bp (57.7bp)	NeuroD1(bHLH)/Islet-NeuroD1-ChIP-Seq(GSE30298)/Homer(0.621) More Information | Similar Motifs Found	motif file (matrix)

- `knownResults.html : known Motif`

Homer Known Motif Enrichment Results (06_Motif_output/)

Homer *de novo* Motif Results
Gene Ontology Enrichment Results
Known Motif Enrichment Results (txt file)
Total Target Sequences = 149985, Total Background Sequences = 150264

Rank	Motif	Name	P-value	log P-pvalue	q-value (Benjamini)	# Target Sequences with Motif	% of Targets Sequences with Motif	# Background Sequences with Motif	% of Background Sequences with Motif	Motif File	PDF
1		STAT5(Stat)/mCD4+-Stat5-ChIP-Seq(GSE12346)/Homer	1e-618	-1.425e+03	0.0000	9734.0	6.49%	5469.6	3.64%	motif file (matrix)	pdf
2		STAT1(Stat)/HelaS3-STAT1-ChIP-Seq(GSE12782)/Homer	1e-469	-1.082e+03	0.0000	8066.0	5.38%	4647.3	3.09%	motif file (matrix)	pdf
3		Stat3(Stat)/mES-Stat3-ChIP-Seq(GSE11431)/Homer	1e-378	-8.719e+02	0.0000	10427.0	6.95%	6832.7	4.55%	motif file (matrix)	pdf
4		STAT4(Stat)/CD4-Stat4-ChIP-Seq(GSE22104)/Homer	1e-336	-7.744e+02	0.0000	19630.0	13.09%	14922.6	9.93%	motif file (matrix)	pdf
5		Stat3+il21(Stat)/CD4-Stat3-ChIP-Seq(GSE19198)/Homer	1e-314	-7.236e+02	0.0000	14402.0	9.60%	10486.8	6.98%	motif file (matrix)	pdf
6		EHF(ETS)/LoVo-EHF-ChIP-Seq(GSE49402)/Homer	1e-179	-4.131e+02	0.0000	23596.0	15.73%	19798.5	13.17%	motif file (matrix)	pdf
7		ERG(ETS)/VCaP-ERG-ChIP-Seq(GSE14097)/Homer	1e-148	-3.418e+02	0.0000	27368.0	18.25%	23680.7	15.75%	motif file (matrix)	pdf

4.1.8 7.Error Code

7-1.Summary

Error ID	Description
Err001	Octopus-toolkit cannot access the web page.
Err002	Incorrect GEO accession number.
Err003	The experiment type cannot be handled with Octopus-toolkit.
Err004	The data cannot be processed.
Err005	Not enough disk space.
Err006	Related to each processing step.
Err007	Some analytics tools are not installed.
Err008	Incorrect password.
Err009	Octopus-toolkit can't read/write files from your computer.
Err010	Incorrect number of Paired-End data.

If you have any questions, Please contact us at Octopustoolkit@gmail.com

7-2.Detail

Err001

Octopus-toolkit attempts to access the NCBI server (National Center for Biotechnology Information) to obtain sample information.

If `your network connection is unstable`, or `the NCBI server is tempararily unavailable`, Octopus-toolkit cannot get information for GSE and/or GSM.

First, `check the network connection` of your computer. If it is ok, please `check the` NCBI and whether the server is operating normally.

If the above cannot solve the problem, the connection to the NCBI may be `timed out` due to unknown reasons. Please re-run Octopus-toolkit after some time (`temporary phenomenon`).

Err002

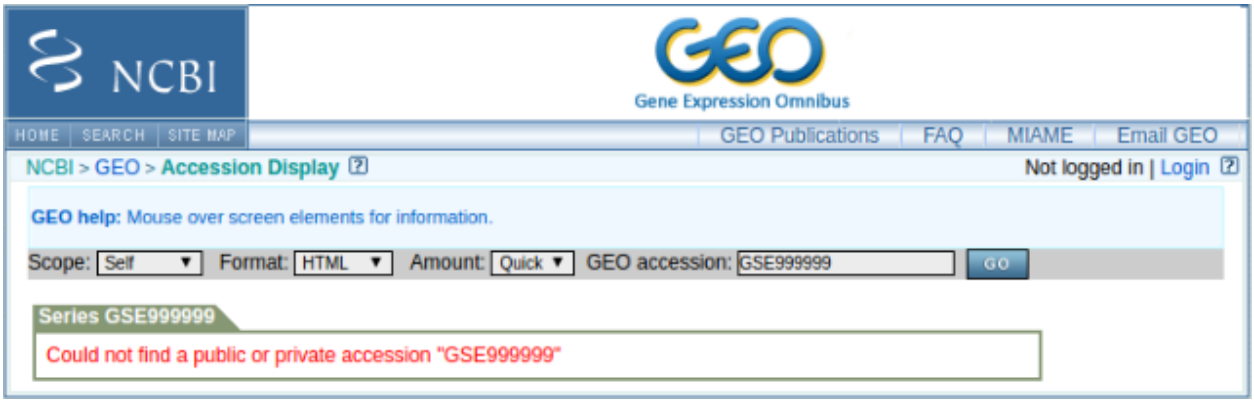
Octopus-toolkit obtains sample information from the GEO (gene expression omnibus) website.

- `GEO Accession Number`

```
A GSExxx is a unique GEO accession number assigned to a study.
A GSMxxx is a unique GEO accession number assigned to a sample. A single GSE␣
↪(study) can have a number of GSM (samples).
```

Octopus-toolkit can only process registered GSE or GSM ids in GEO. `Err002` occurs when you put `unregistered` accession ids or `misspelled accession ids`.

- Unregistered GSE id (Input : `GSE999999`)



- Misspelled or incorrect accession number (Input : `ChIP-Seq`)

Please `check the GEO accession number` whether it is registered in the GEO.

Err003

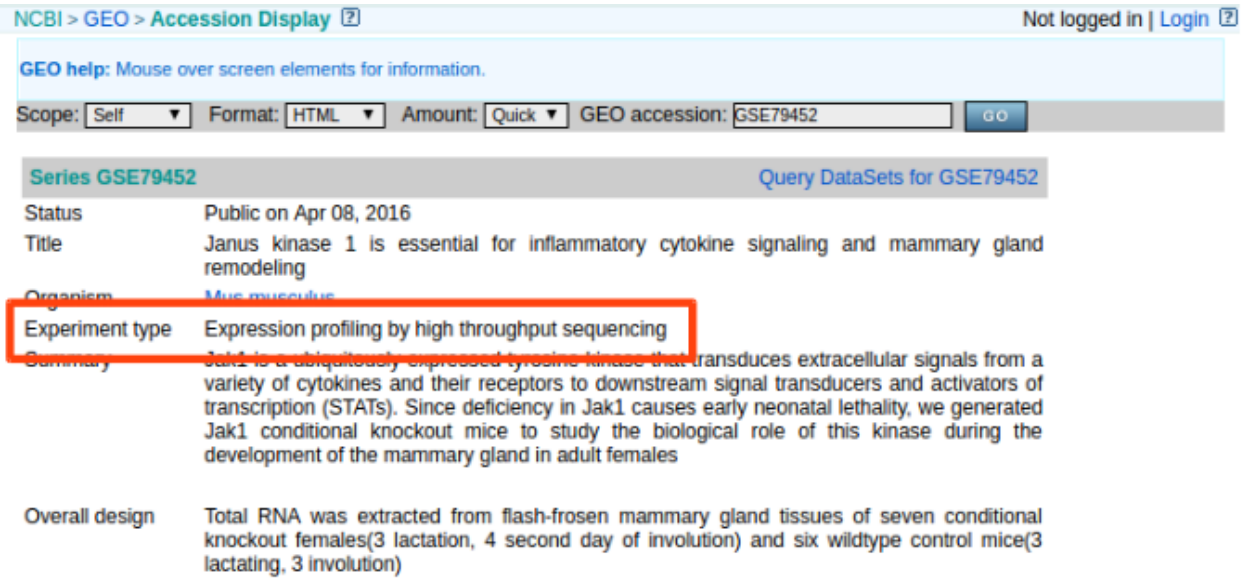
There are many different types of next-generation sequencing (NGS) data. As defined by NCBI (`NGS data` - Study type), genome binding/occupancy profiling by high throughput sequencing indicates ChIP-seq data.

Octopus-toolkit currently supports the following types of NGS data. Other NGS types will be skipped. `expression profiling by high throughput sequencing` (RNA-seq) `genome binding/occupancy profiling by high throughput sequencing` (ChIP-seq / MNase-seq / ATAC-seq / MeDIP-seq / DNase-seq)

(Other NGS types will be added later)

You can check `experiment type` of given GEO accession number through the website. (ex: GSE79452)

- Experiment Type



Err004

Not all data in the GEO can be processed with the Octopus-toolkit. Octopus-toolkit check the following information before the processing. `Organism`, `Library strategy`, `Instrument model`, and `FTP Address(SRA Experiment)`). (Important)

- DataSet for GSE79452 (Ex : GSE79452)

Sample GSM2095539 Query DataSets for GSM2095539

Status	Public on Apr 08, 2016
Title	Jak1_42186_Con_Lactation
Sample type	SRA
Source name	Mammary gland tissue
Organism	Mus musculus
Characteristics	tissue: Mammary gland developmental stage: Day 7 of Lactation
Extracted molecule	total RNA
Extraction protocol	Total RNA was extracted from flash-frozen mammary gland tissues of seven conditional knockout femalesand six wildtype control mice using the RNeasy Mini Kit (Qiagen). RNA samples were processed using the TruSeq RNA Sample kit and sequenced using a HiSeq2000 sequencer (Illumina).
Library strategy	RNA-Seq
Library source	transcriptomic
Library selection	cDNA
Instrument model	Illumina HiSeq 2000
Platform ID	GPL13112
Series (1)	GSE79452 Janus kinase 1 is essential for inflammatory cytokine signaling and mammary gland remodeling

Relations

BioSample	SAMN04571205
SRA	SRX1650911

Supplementary file	Size	Download	File type/resource
SRX/SRX165/SRX1650911		(ftp)	SRA Experiment

Raw data provided as supplementary file

Processed data is available on Series record

`Err004` is divided into the following four subcategories.

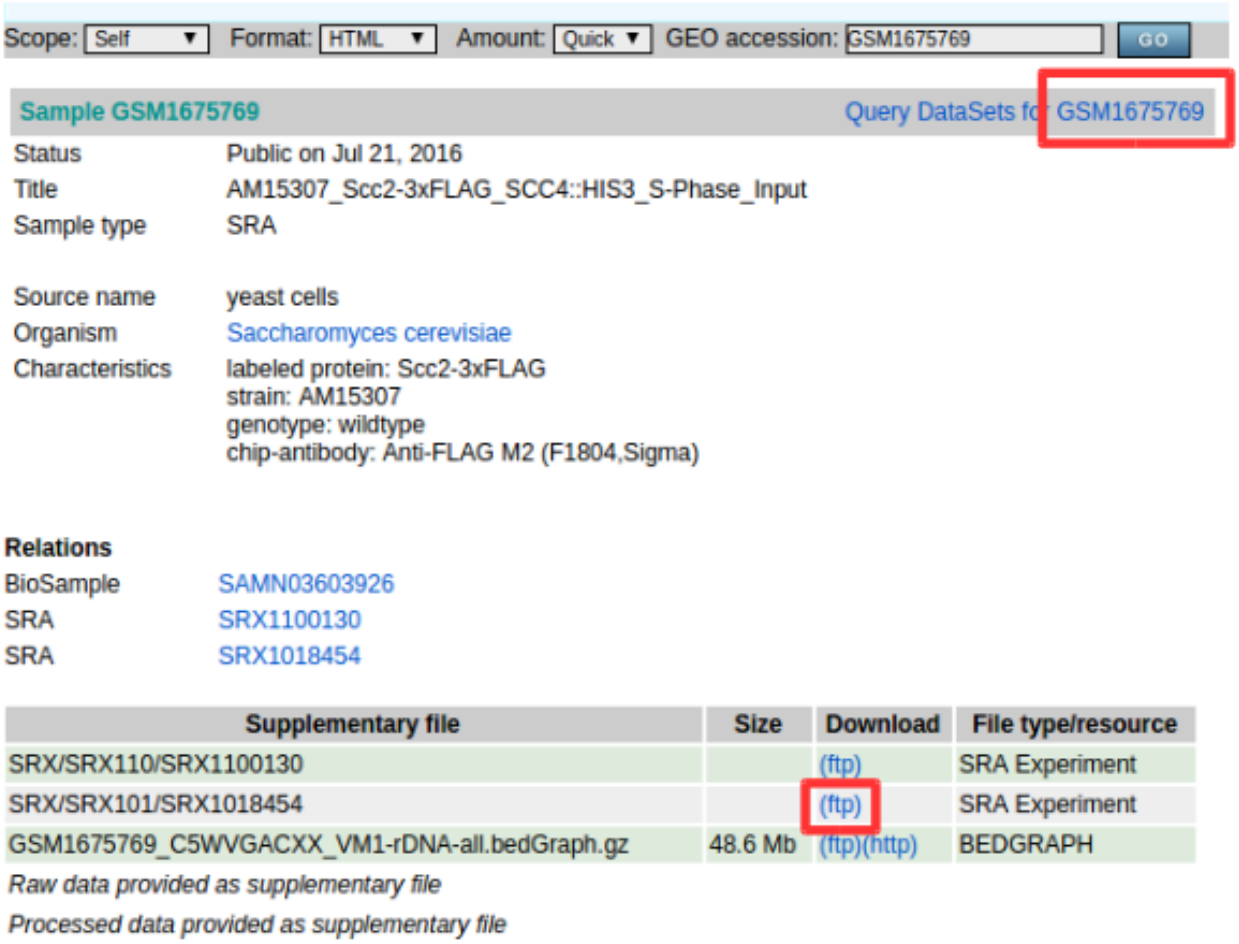
Sub Error ID	Description
`Err004-1`	The organism is not supported.
`Err004-2`	The experiment type is not supported (for example Exome-seq).
`Err004-3`	The instrument is not supported. Octopus-toolkit can only process data generated by Illumina instrument.
`Err004-4`	Raw data (.sra) is currently unavailable (probably newly registered data).

`Err004` is related to unsupported data by Octopus-toolkit. The following data is currently handled with Octopus-toolkit.

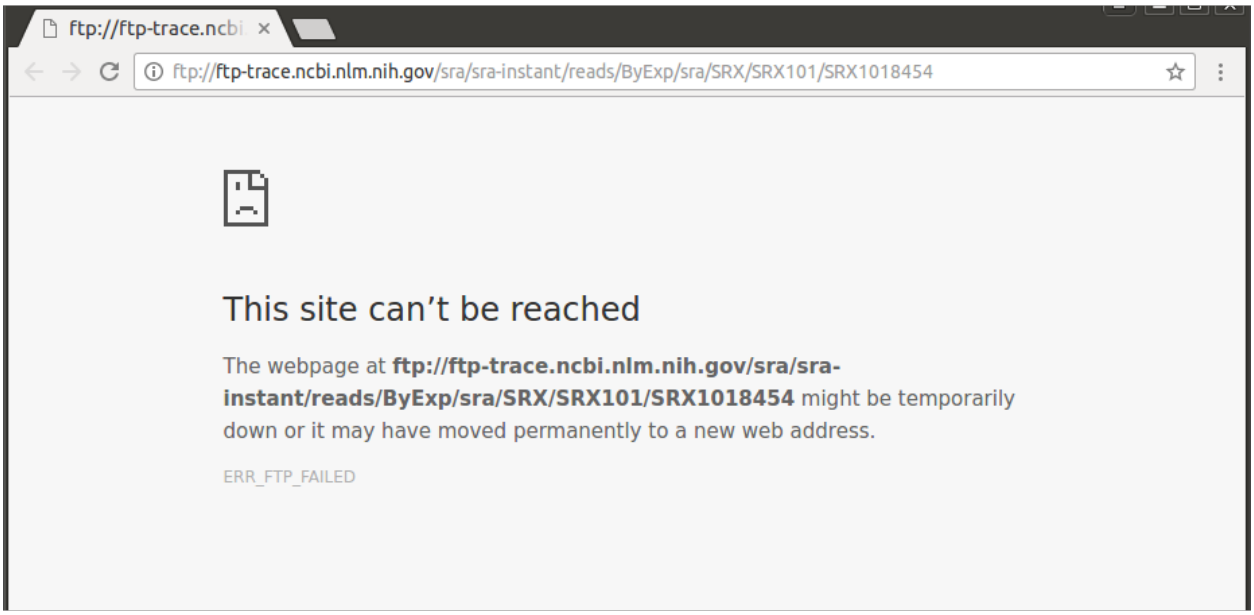
Type	Description
Organism	Homo sapiens, Mus musculus, Drosophila melanogaster, Saccharomyces cerevisiae, Canis lupus familaris, Arabidopsis thaliana, Danio rerio, Caenorhabditis elegans
Library Strategy	ChIP-Seq, RNA-Seq, MeDIP-Seq, ATAC-Seq, DNase-Seq, MNase-Seq
Instrument Model	Illumina GA/HiSeq/MiSeq (Illumina)

`Err004-4` indicates that data has been registered in the GEO, but the raw data (.sra) has not been released yet. Therefore, please check the availability of raw files.

- `Error004-4` example : GSM1675769



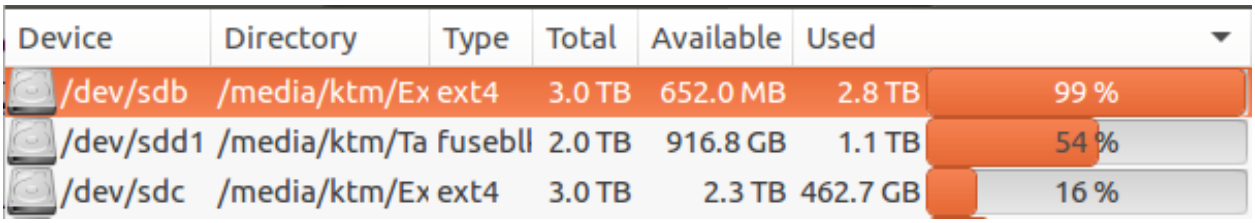
- No raw files (.sra).



Err005

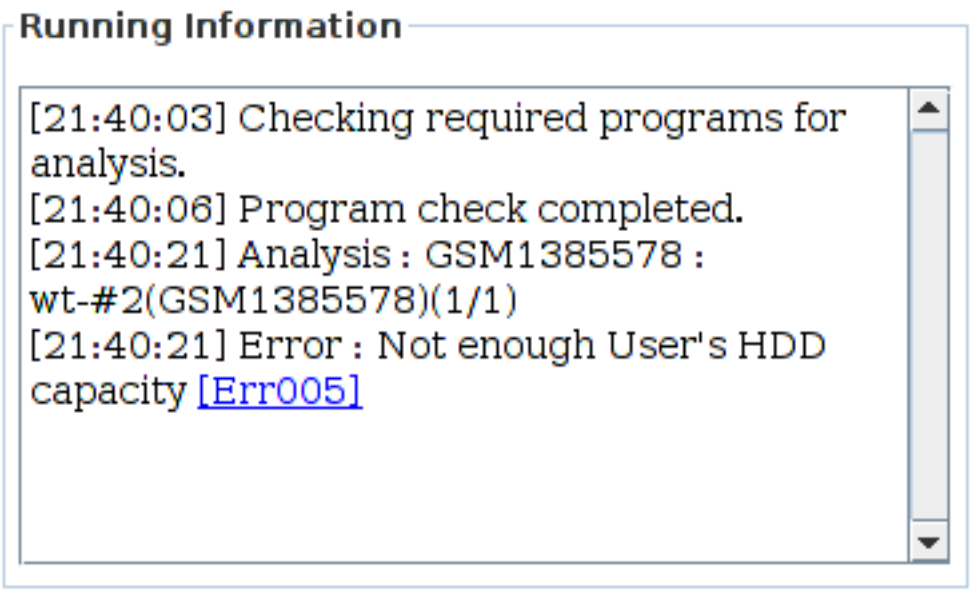
This error is related to disk space. To resolve this issue, obtain `enough free space` (more than 10Gb) and re-run the analysis.

- Check your hard disk space.



Device	Directory	Type	Total	Available	Used	
/dev/sdb	/media/ktm/Ex	ext4	3.0 TB	652.0 MB	2.8 TB	99 %
/dev/sdd1	/media/ktm/Ta	fusebll	2.0 TB	916.8 GB	1.1 TB	54 %
/dev/sdc	/media/ktm/Ex	ext4	3.0 TB	2.3 TB	462.7 GB	16 %

- Status window.



Err006

`Err006` is divided into six subcategories.

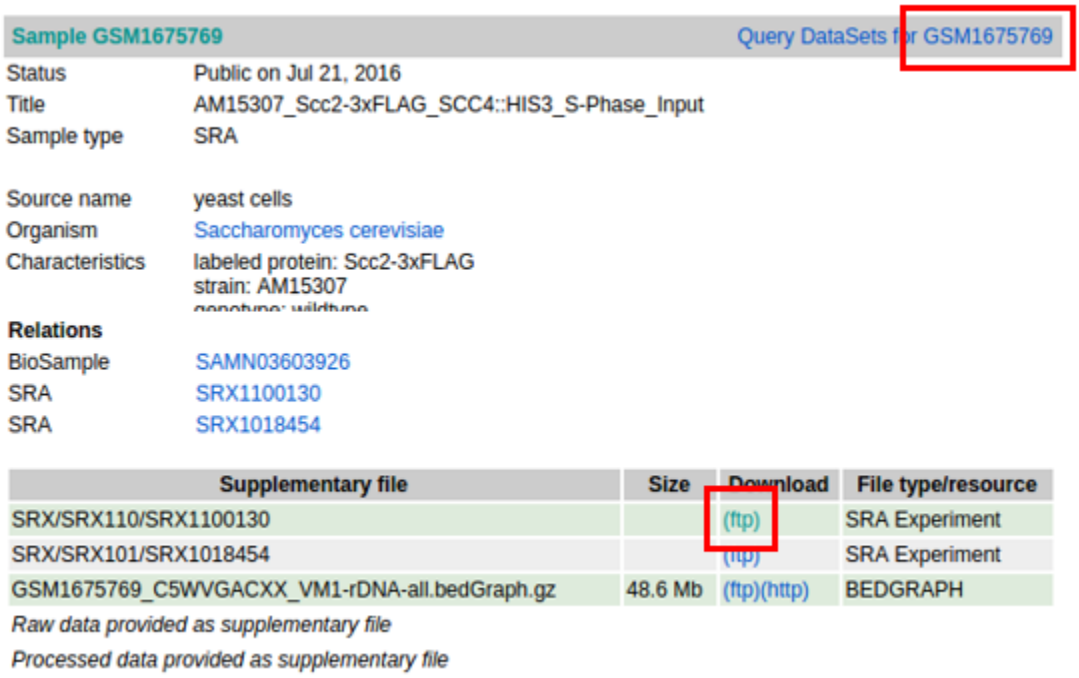
Sub Error ID	Description
Err006-1	Cannot `access` NCBI's FTP server.
Err006-2	File converting error from .sra to .fastq using `fastq-dump`.
Err006-3	Related to the .fastq file while checking the quality using `FastQC`.
Err006-4	No input file (.fastq) for `Trimming`.
Err006-5	Related to the `Mapping` step.
Err006-6	Related to the `Sorting` step (BAM file).

Err006-1

NCBI provides raw data of published sample through `FTP server` to user. If the NCBI homepage is working normally, you can extract the sample information, but if the FTP server does not work, you will not be able to download the data.

To solve this issue, you connect directly to the FTP server of NCBI.

- `Error006-1` example : GSM1675769



Sample GSM1675769 Query DataSets for GSM1675769

Status	Public on Jul 21, 2016
Title	AM15307_Scc2-3xFLAG_SCC4::HIS3_S-Phase_Input
Sample type	SRA
Source name	yeast cells
Organism	Saccharomyces cerevisiae
Characteristics	labeled protein: Scc2-3xFLAG strain: AM15307 genotype: wildtype

Relations

BioSample	SAMN03603926
SRA	SRX1100130
SRA	SRX1018454

Supplementary file	Size	Download	File type/resource
SRX/SRX110/SRX1100130		(ftp)	SRA Experiment
SRX/SRX101/SRX1018454		(ftp)	SRA Experiment
GSM1675769_C5WVGACXX_VM1-rDNA-all.bedGraph.gz	48.6 Mb	(ftp)(http)	BEDGRAPH

Raw data provided as supplementary file

Processed data provided as supplementary file

If you can connect to the FTP server, manually download the published sample.

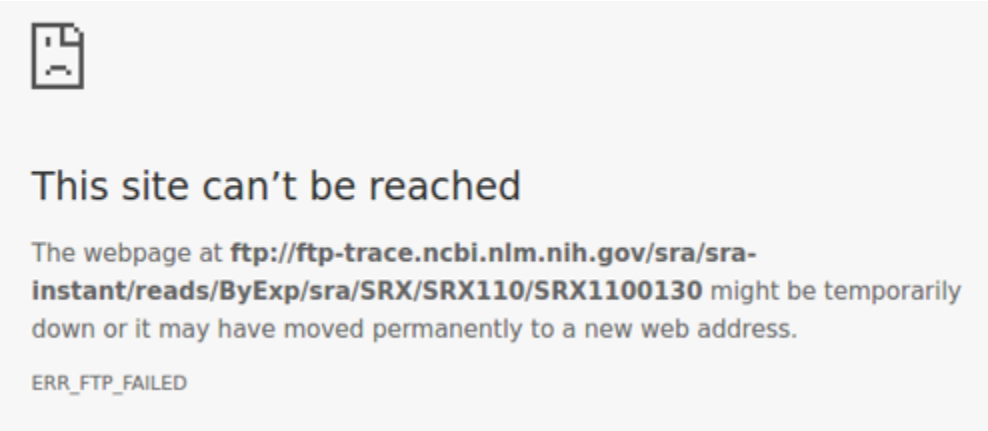
- NCBI Ftp server is running.(`Success`)

Index of /sra/sra-instant/reads/ByExp/sra/SRX/SRX110/SRX1100130

Name	Size	Date Modified
[parent directory]		
SRR2125091/		

If the server is closed or samples can not be downloaded, please contact the NCBI because it is an issue for the NCBI.

- NCBI Ftp server is closed.(`Fail`)



This site can't be reached

The webpage at **ftp://ftp-trace.ncbi.nlm.nih.gov/sra/sra-instant/reads/ByExp/sra/SRX/SRX110/SRX1100130** might be temporarily down or it may have moved permanently to a new web address.

ERR_FTP_FAILED

If the above method works normally, please try Octopus-toolkit again.

If you still have an `Err006-1` in the retrial, please contact us at the address below.

Contact us : Octopustoolkit@gmail.com

Err006-2

Raw data of samples downloaded from NCBI is compressed in `SRA format`. For NGS analysis, `SRA` file should be converted to `Fastq` format. The tool used in this step is `Fastq-dump`, a sub tool of `SRA-Toolkit`.

- `Input file` : Sequence Read Archive (Extension : `sra`)
- `Output file` : Short read sequence. (Extension : `fastq`)

`006-2` occurs when there is no or invalid SRA file, which is the input file for executing Fastq-dump.

This error may arise due to an abrupt disconnection during the previous downloading step of the raw data from FTP server, or raw data uploaded to NCBI may be broken.

You should check your `network status`, `free space` on your computer and try the analysis again.

If the above method does not work, please contact us at the address below.

Contact us : Octopustoolkit@gmail.com

Err006-3

`Err006-3` means that the input file(`Fastq`) for the `Quality Check` is invalid or an issue in the system itself during `Quality Check` using `FastQC`.

You should check fastq files on your computer and try the analysis again.

If the above method does not work, please contact us at the address below.

Contact us : Octopustoolkit@gmail.com

After successfully completing the `Quality Check` step, some problems may prevent `FastQC` from generating `Fastqc_data.txt`.

Octopus-toolkit extracts the encoding information of the sample from `fastqc_data.txt` among the outputs of `FastQC`. Therefore, if `Fastqc_data.txt` is not generated, it stores the encoding information of the latest samples. (`Sanger / Illumina 1.9`)

- `Err006-3` Encoding information:

Basic Statistics

Measure	Value
Filename	H3K4me1_ChIPSeq_mb.fastq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	11308561
Sequences flagged as poor quality	0
Sequence length	35-37
%GC	48

Err006-4

`Err006-4` occurs when there is no input file(`Fastq`) for `Trimming` step or when all reads are removed due to `bad quality`.

You should check fastq files on your computer and try the analysis again.

If the above method does not work, please contact us using address below.

Contact us : Octopustoolkit@gmail.com

If all reads are removed by `bad quality`, Octopus-toolkit will use the non-trimmed input file(`Fastq`) to proceed. (Next step : `Mapping`)

Err006-5

`Err006-5` may arise due to the following reasons.

- The input file (`non_trimmed Fastq`, `Trimmed Fastq`) does not exist.
- A large number of reads are trimmed due to `bad sequencing quality` or `high threshold used`.
- Too few mapped reads (Less than 2 MegaByte).

You should check your input file (`non-trimmed and trimmed fastq files`), `read count`, `file size after timming`.

Err006-6

`Err006-6`: BAM (mapped) file does not exist or the number of mapped reads is too small.

You should check `input file` and `BAM file`.

Err007

`Err007` is related to the installation step.

To use the Octopus-toolkit, your must follow the installation procedure completely: `Requirement(Err007-1)` and `analysis tools(Err007-2)`.

- *Requirement* : Library files must be installed.
- Analysis tools : Tools are installed automatically by Octopus-toolkit. If the installation procedure is interrupted, please remove the Octopus-toolkit directory and rerun it.

Octopus-toolkit download files from the `HOMER` website. `Err007` occurs if the website (http://homer.ucsd.edu/homer/) is unavailable, `Err007` can occur.

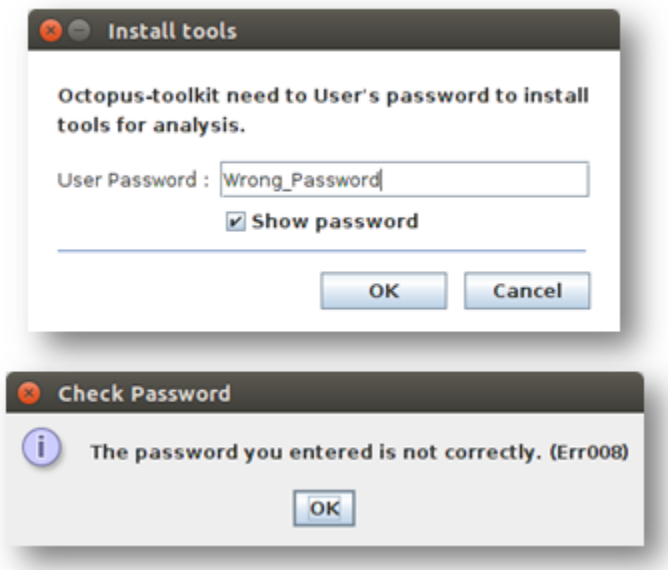
Err008

`Err008` is related to password issue.

- `Password` : You must enter your password once during the installtion step.

Please check your password and try again.

- When you enter incorrect password (Example : My password = ktm123)



Err009

`Err009` is related to script files generated by Octopus-toolkit. If this happens, please rerun it later.

Err010

`Err010` indicates that the number of files (paired-end sample) does not match when merging.

If there are several SRA files in one sample (GSM), Octopus-toolkit will merge them.

Paired-end data must have two files, Sample1_1.fastq and Sample1_2.fastq.

`Err010` occurs if any of these fails.

4.1.9 8.License

GNU GENERAL PUBLIC LICENSE

Version 3, 29 June 2007

Everyone is permitted to copy and distribute verbatim copies of this license document, but changing it is not allowed.

GPLv3 Link : https://www.gnu.org/copyleft/gpl.html

Preamble

The GNU General Public License is a free, copyleft license for software and other kinds of works.

The licenses for most software and other practical works are designed to take away your freedom to share and change the works. By contrast, the GNU General Public License is intended to guarantee your freedom to share and change all versions of a program–to make sure it remains free software for all its users. We, the Free Software Foundation, use the GNU General Public License for most of our software; it applies also to any other work released this way by its authors. You can apply it to your programs, too.

When we speak of free software, we are referring to freedom, not price. Our General Public Licenses are designed to make sure that you have the freedom to distribute copies of free software (and charge for them if you wish), that you receive source code or can get it if you want it, that you can change the software or use pieces of it in new free programs, and that you know you can do these things.

To protect your rights, we need to prevent others from denying you these rights or asking you to surrender the rights. Therefore, you have certain responsibilities if you distribute copies of the software, or if you modify it: responsibilities to respect the freedom of others.

For example, if you distribute copies of such a program, whether gratis or for a fee, you must pass on to the recipients the same freedoms that you received. You must make sure that they, too, receive or can get the source code. And you must show them these terms so they know their rights.

Developers that use the GNU GPL protect your rights with two steps: (1) assert copyright on the software, and (2) offer you this License giving you legal permission to copy, distribute and/or modify it.

For the developers' and authors' protection, the GPL clearly explains that there is no warranty for this free software. For both users' and authors' sake, the GPL requires that modified versions be marked as changed, so that their problems will not be attributed erroneously to authors of previous versions.

Some devices are designed to deny users access to install or run modified versions of the software inside them, although the manufacturer can do so. This is fundamentally incompatible with the aim of protecting users' freedom to change the software. The systematic pattern of such abuse occurs in the area of products for individuals to use, which is precisely where it is most unacceptable. Therefore, we have designed this version of the GPL to prohibit the practice for those products. If such problems arise substantially in other domains, we stand ready to extend this provision to those domains in future versions of the GPL, as needed to protect the freedom of users.

Finally, every program is threatened constantly by software patents. States should not allow patents to restrict development and use of software on general-purpose computers, but in those that do, we wish to avoid the special danger that patents applied to a free program could make it effectively proprietary. To prevent this, the GPL assures that patents cannot be used to render the program non-free.

The precise terms and conditions for copying, distribution and modification follow.

TERMS AND CONDITIONS

0. Definitions.

"This License" refers to version 3 of the GNU General Public License.

"Copyright" also means copyright-like laws that apply to other kinds of works, such as semiconductor masks.

"The Program" refers to any copyrightable work licensed under this License. Each licensee is addressed as "you". "Licensees" and "recipients" may be individuals or organizations.

To "modify" a work means to copy from or adapt all or part of the work in a fashion requiring copyright permission, other than the making of an exact copy. The resulting work is called a "modified version" of the earlier work or a work "based on" the earlier work.

A "covered work" means either the unmodified Program or a work based on the Program.

To "propagate" a work means to do anything with it that, without permission, would make you directly or secondarily liable for infringement under applicable copyright law, except executing it on a computer or modifying a private copy. Propagation includes copying, distribution (with or without modification), making available to the public, and in some countries other activities as well.

To "convey" a work means any kind of propagation that enables other parties to make or receive copies. Mere interaction with a user through a computer network, with no transfer of a copy, is not conveying.

An interactive user interface displays "Appropriate Legal Notices" to the extent that it includes a convenient and prominently visible feature that (1) displays an appropriate copyright notice, and (2) tells the user that there is no warranty for the work (except to the extent that warranties are provided), that licensees may convey the work under this License, and how to view a copy of this License. If the interface presents a list of user commands or options, such as a menu, a prominent item in the list meets this criterion.

1. Source Code.

The "source code" for a work means the preferred form of the work for making modifications to it. "Object code" means any non-source form of a work.

A "Standard Interface" means an interface that either is an official standard defined by a recognized standards body, or, in the case of interfaces specified for a particular programming language, one that is widely used among developers working in that language.

The "System Libraries" of an executable work include anything, other than the work as a whole, that (a) is included in the normal form of packaging a Major Component, but which is not part of that Major Component, and (b) serves only to enable use of the work with that Major Component, or to implement a Standard Interface for which an implementation is available to the public in source code form. A "Major Component", in this context, means a major essential component (kernel, window system, and so on) of the specific operating system (if any) on which the executable work runs, or a compiler used to produce the work, or an object code interpreter used to run it.

The "Corresponding Source" for a work in object code form means all the source code needed to generate, install, and (for an executable work) run the object code and to modify the work, including scripts to control those activities. However, it does not include the work's System Libraries, or general-purpose tools or generally available free programs which are used unmodified in performing those activities but which are not part of the work. For example, Corresponding Source includes interface definition files associated with source files for the work, and the source code for shared libraries and dynamically linked subprograms that the work is specifically designed to require, such as by intimate data communication or control flow between those subprograms and other parts of the work.

The Corresponding Source need not include anything that users can regenerate automatically from other parts of the Corresponding Source.

The Corresponding Source for a work in source code form is that same work.

2. Basic Permissions.

All rights granted under this License are granted for the term of copyright on the Program, and are irrevocable provided the stated conditions are met. This License explicitly affirms your unlimited permission to run the unmodified Program. The output from running a covered work is covered by this License only if the output, given its content, constitutes a covered work. This License acknowledges your rights of fair use or other equivalent, as provided by copyright law.

You may make, run and propagate covered works that you do not convey, without conditions so long as your license otherwise remains in force. You may convey covered works to others for the sole purpose of having them make modifications exclusively for you, or provide you with facilities for running those works, provided that you comply with the terms of this License in conveying all material for which you do not control copyright. Those thus making or running the covered works for you must do so exclusively on your behalf, under your direction and control, on terms that prohibit them from making any copies of your copyrighted material outside their relationship with you.

Conveying under any other circumstances is permitted solely under the conditions stated below. Sublicensing is not allowed; section 10 makes it unnecessary.

3. Protecting Users' Legal Rights From Anti-Circumvention Law.

No covered work shall be deemed part of an effective technological measure under any applicable law fulfilling obligations under article 11 of the WIPO copyright treaty adopted on 20 December 1996, or similar laws prohibiting or restricting circumvention of such measures.

When you convey a covered work, you waive any legal power to forbid circumvention of technological measures to the extent such circumvention is effected by exercising rights under this License with respect to the covered work, and you disclaim any intention to limit operation or modification of the work as a means of enforcing, against the work's users, your or third parties' legal rights to forbid circumvention of technological measures.

4. Conveying Verbatim Copies.

You may convey verbatim copies of the Program's source code as you receive it, in any medium, provided that you conspicuously and appropriately publish on each copy an appropriate copyright notice; keep intact all notices stating that this License and any non-permissive terms added in accord with section 7 apply to the code; keep intact all notices of the absence of any warranty; and give all recipients a copy of this License along with the Program.

You may charge any price or no price for each copy that you convey, and you may offer support or warranty protection for a fee.

5. Conveying Modified Source Versions.

You may convey a work based on the Program, or the modifications to produce it from the Program, in the form of source code under the terms of section 4, provided that you also meet all of these conditions:

a) The work must carry prominent notices stating that you modified it, and giving a relevant date.

b) The work must carry prominent notices stating that it is released under this License and any conditions added under section 7. This requirement modifies the requirement in section 4 to "keep intact all notices".

c) You must license the entire work, as a whole, under this License to anyone who comes into possession of a copy. This License will therefore apply, along with any applicable section 7 additional terms, to the whole of the work, and all its parts, regardless of how they are packaged. This License gives no permission to license the work in any other way, but it does not invalidate such permission if you have separately received it.

d) If the work has interactive user interfaces, each must display Appropriate Legal Notices; however, if the Program has interactive interfaces that do not display Appropriate Legal Notices, your work need not make them do so. A compilation of a covered work with other separate and independent works, which are not by their nature extensions of the covered work, and which are not combined with it such as to form a larger program, in or on a volume of a storage or distribution medium, is called an "aggregate" if the compilation and its resulting copyright are not used to limit the access or legal rights of the compilation's users beyond what the individual works permit. Inclusion of a covered work in an aggregate does not cause this License to apply to the other parts of the aggregate.

6. Conveying Non-Source Forms.

You may convey a covered work in object code form under the terms of sections 4 and 5, provided that you also convey the machine-readable Corresponding Source under the terms of this License, in one of these ways:

a) Convey the object code in, or embodied in, a physical product (including a physical distribution medium), accompanied by the Corresponding Source fixed on a durable physical medium customarily used for software interchange.

b) Convey the object code in, or embodied in, a physical product (including a physical distribution medium), accompanied by a written offer, valid for at least three years and valid for as long as you offer spare parts or customer support for that product model, to give anyone who possesses the object code either (1) a copy of the Corresponding Source for all the software in the product that is covered by this License, on a durable physical medium customarily used for software interchange, for a price no more than your reasonable cost of physically performing this conveying of source, or (2) access to copy the Corresponding Source from a network server at no charge.

c) Convey individual copies of the object code with a copy of the written offer to provide the Corresponding Source. This alternative is allowed only occasionally and noncommercially, and only if you received the object code with such an offer, in accord with subsection 6b.

d) Convey the object code by offering access from a designated place (gratis or for a charge), and offer equivalent access to the Corresponding Source in the same way through the same place at no further charge. You need not require recipients to copy the Corresponding Source along with the object code. If the place to copy the object code is a network server, the Corresponding Source may be on a different server (operated by you or a third party) that supports equivalent copying facilities, provided you maintain clear directions next to the object code saying where to find the Corresponding Source. Regardless of what server hosts the Corresponding Source, you remain obligated to ensure that it is available for as long as needed to satisfy these requirements.

e) Convey the object code using peer-to-peer transmission, provided you inform other peers where the object code and Corresponding Source of the work are being offered to the general public at no charge under subsection 6d. A separable portion of the object code, whose source code is excluded from the Corresponding Source as a System Library, need not be included in conveying the object code work.

A "User Product" is either (1) a "consumer product", which means any tangible personal property which is normally used for personal, family, or household purposes, or (2) anything designed or sold for incorporation into a dwelling. In determining whether a product is a consumer product, doubtful cases shall be resolved in favor of coverage. For a particular product received by a particular user, "normally used" refers to a typical or common use of that class of product, regardless of the status of the particular user or of the way in which the particular user actually uses, or expects or is expected to use, the product. A product is a consumer product regardless of whether the product has substantial commercial, industrial or non-consumer uses, unless such uses represent the only significant mode of use of the product.

"Installation Information" for a User Product means any methods, procedures, authorization keys, or other information required to install and execute modified versions of a covered work in that User Product from a modified version of its Corresponding Source. The information must suffice to ensure that the continued functioning of the modified object code is in no case prevented or interfered with solely because modification has been made.

If you convey an object code work under this section in, or with, or specifically for use in, a User Product, and the conveying occurs as part of a transaction in which the right of possession and use of the User Product is transferred to the recipient in perpetuity or for a fixed term (regardless of how the transaction is characterized), the Corresponding Source conveyed under this section must be accompanied by the Installation Information. But this requirement does not apply if neither you nor any third party retains the ability to install modified object code on the User Product (for example, the work has been installed in ROM).

The requirement to provide Installation Information does not include a requirement to continue to provide support service, warranty, or updates for a work that has been modified or installed by the recipient, or for the User Product in which it has been modified or installed. Access to a network may be denied when the modification itself materially and adversely affects the operation of the network or violates the rules and protocols for communication across the network.

Corresponding Source conveyed, and Installation Information provided, in accord with this section must be in a format that is publicly documented (and with an implementation available to the public in source code form), and must require no special password or key for unpacking, reading or copying.

7. Additional Terms.

"Additional permissions" are terms that supplement the terms of this License by making exceptions from one or more of its conditions. Additional permissions that are applicable to the entire Program shall be treated as though they were included in this License, to the extent that they are valid under applicable law. If additional permissions apply only to part of the Program, that part may be used separately under those permissions, but the entire Program remains governed by this License without regard to the additional permissions.

When you convey a copy of a covered work, you may at your option remove any additional permissions from that copy, or from any part of it. (Additional permissions may be written to require their own removal in certain cases when you modify the work.) You may place additional permissions on material, added by you to a covered work, for which you have or can give appropriate copyright permission.

Notwithstanding any other provision of this License, for material you add to a covered work, you may (if authorized by the copyright holders of that material) supplement the terms of this License with terms:

a) Disclaiming warranty or limiting liability differently from the terms of sections 15 and 16 of this License; or

b) Requiring preservation of specified reasonable legal notices or author attributions in that material or in the Appropriate Legal Notices displayed by works containing it; or

c) Prohibiting misrepresentation of the origin of that material, or requiring that modified versions of such material be marked in reasonable ways as different from the original version; or

d) Limiting the use for publicity purposes of names of licensors or authors of the material; or

e) Declining to grant rights under trademark law for use of some trade names, trademarks, or service marks; or

f) Requiring indemnification of licensors and authors of that material by anyone who conveys the material (or modified versions of it) with contractual assumptions of liability to the recipient, for any liability that these contractual assumptions directly impose on those licensors and authors. All other non-permissive additional terms are considered "further restrictions" within the meaning of section 10. If the Program as you received it, or any part of it, contains a notice stating that it is governed by this License along with a term that is a further restriction, you may remove that term. If a license document contains a further restriction but permits relicensing or conveying under this License, you may add to a covered work material governed by the terms of that license document, provided that the further restriction does not survive such relicensing or conveying.

If you add terms to a covered work in accord with this section, you must place, in the relevant source files, a statement of the additional terms that apply to those files, or a notice indicating where to find the applicable terms.

Additional terms, permissive or non-permissive, may be stated in the form of a separately written license, or stated as exceptions; the above requirements apply either way.

8. Termination.

You may not propagate or modify a covered work except as expressly provided under this License. Any attempt otherwise to propagate or modify it is void, and will automatically terminate your rights under this License (including any patent licenses granted under the third paragraph of section 11).

However, if you cease all violation of this License, then your license from a particular copyright holder is reinstated (a) provisionally, unless and until the copyright holder explicitly and finally terminates your license, and (b) permanently, if the copyright holder fails to notify you of the violation by some reasonable means prior to 60 days after the cessation.

Moreover, your license from a particular copyright holder is reinstated permanently if the copyright holder notifies you of the violation by some reasonable means, this is the first time you have received notice of violation of this License (for any work) from that copyright holder, and you cure the violation prior to 30 days after your receipt of the notice.

Termination of your rights under this section does not terminate the licenses of parties who have received copies or rights from you under this License. If your rights have been terminated and not permanently reinstated, you do not qualify to receive new licenses for the same material under section 10.

9. Acceptance Not Required for Having Copies.

You are not required to accept this License in order to receive or run a copy of the Program. Ancillary propagation of a covered work occurring solely as a consequence of using peer-to-peer transmission to receive a copy likewise does not require acceptance. However, nothing other than this License grants you permission to propagate or modify any covered work. These actions infringe copyright if you do not accept this License. Therefore, by modifying or propagating a covered work, you indicate your acceptance of this License to do so.

10. Automatic Licensing of Downstream Recipients.

Each time you convey a covered work, the recipient automatically receives a license from the original licensors, to run, modify and propagate that work, subject to this License. You are not responsible for enforcing compliance by third parties with this License.

An "entity transaction" is a transaction transferring control of an organization, or substantially all assets of one, or subdividing an organization, or merging organizations. If propagation of a covered work results from an entity transaction, each party to that transaction who receives a copy of the work also receives whatever licenses to the work the party's predecessor in interest had or could give under the previous paragraph, plus a right to possession of the Corresponding Source of the work from the predecessor in interest, if the predecessor has it or can get it with reasonable efforts.

You may not impose any further restrictions on the exercise of the rights granted or affirmed under this License. For example, you may not impose a license fee, royalty, or other charge for exercise of rights granted under this License, and you may not initiate litigation (including a cross-claim or counterclaim in a lawsuit) alleging that any patent claim is infringed by making, using, selling, offering for sale, or importing the Program or any portion of it.

11. Patents.

A "contributor" is a copyright holder who authorizes use under this License of the Program or a work on which the Program is based. The work thus licensed is called the contributor's "contributor version".

A contributor's "essential patent claims" are all patent claims owned or controlled by the contributor, whether already acquired or hereafter acquired, that would be infringed by some manner, permitted by this License, of making, using, or selling its contributor version, but do not include claims that would be infringed only as a consequence of further modification of the contributor version. For purposes of this definition, "control" includes the right to grant patent sublicenses in a manner consistent with the requirements of this License.

Each contributor grants you a non-exclusive, worldwide, royalty-free patent license under the contributor's essential patent claims, to make, use, sell, offer for sale, import and otherwise run, modify and propagate the contents of its contributor version.

In the following three paragraphs, a "patent license" is any express agreement or commitment, however denominated, not to enforce a patent (such as an express permission to practice a patent or covenant not to sue for patent infringement). To "grant" such a patent license to a party means to make such an agreement or commitment not to enforce a patent against the party.

If you convey a covered work, knowingly relying on a patent license, and the Corresponding Source of the work is not available for anyone to copy, free of charge and under the terms of this License, through a publicly available network server or other readily accessible means, then you must either (1) cause the Corresponding Source to be so available, or (2) arrange to deprive yourself of the benefit of the patent license for this particular work, or (3) arrange, in a manner consistent with the requirements of this License, to extend the patent license to downstream recipients. "Knowingly relying" means you have actual knowledge that, but for the patent license, your conveying the covered work in a country, or your recipient's use of the covered work in a country, would infringe one or more identifiable patents in that country that you have reason to believe are valid.

If, pursuant to or in connection with a single transaction or arrangement, you convey, or propagate by procuring conveyance of, a covered work, and grant a patent license to some of the parties receiving the covered work authorizing them to use, propagate, modify or convey a specific copy of the covered work, then the patent license you grant is automatically extended to all recipients of the covered work and works based on it.

A patent license is "discriminatory" if it does not include within the scope of its coverage, prohibits the exercise of, or is conditioned on the non-exercise of one or more of the rights that are specifically granted under this License. You may not convey a covered work if you are a party to an arrangement with a third party that is in the business

of distributing software, under which you make payment to the third party based on the extent of your activity of conveying the work, and under which the third party grants, to any of the parties who would receive the covered work from you, a discriminatory patent license (a) in connection with copies of the covered work conveyed by you (or copies made from those copies), or (b) primarily for and in connection with specific products or compilations that contain the covered work, unless you entered into that arrangement, or that patent license was granted, prior to 28 March 2007.

Nothing in this License shall be construed as excluding or limiting any implied license or other defenses to infringement that may otherwise be available to you under applicable patent law.

12. No Surrender of Others' Freedom.

If conditions are imposed on you (whether by court order, agreement or otherwise) that contradict the conditions of this License, they do not excuse you from the conditions of this License. If you cannot convey a covered work so as to satisfy simultaneously your obligations under this License and any other pertinent obligations, then as a consequence you may not convey it at all. For example, if you agree to terms that obligate you to collect a royalty for further conveying from those to whom you convey the Program, the only way you could satisfy both those terms and this License would be to refrain entirely from conveying the Program.

13. Use with the GNU Affero General Public License.

Notwithstanding any other provision of this License, you have permission to link or combine any covered work with a work licensed under version 3 of the GNU Affero General Public License into a single combined work, and to convey the resulting work. The terms of this License will continue to apply to the part which is the covered work, but the special requirements of the GNU Affero General Public License, section 13, concerning interaction through a network will apply to the combination as such.

14. Revised Versions of this License.

The Free Software Foundation may publish revised and/or new versions of the GNU General Public License from time to time. Such new versions will be similar in spirit to the present version, but may differ in detail to address new problems or concerns.

Each version is given a distinguishing version number. If the Program specifies that a certain numbered version of the GNU General Public License "or any later version" applies to it, you have the option of following the terms and conditions either of that numbered version or of any later version published by the Free Software Foundation. If the Program does not specify a version number of the GNU General Public License, you may choose any version ever published by the Free Software Foundation.

If the Program specifies that a proxy can decide which future versions of the GNU General Public License can be used, that proxy's public statement of acceptance of a version permanently authorizes you to choose that version for the Program.

Later license versions may give you additional or different permissions. However, no additional obligations are imposed on any author or copyright holder as a result of your choosing to follow a later version.

15. Disclaimer of Warranty.

THERE IS NO WARRANTY FOR THE PROGRAM, TO THE EXTENT PERMITTED BY APPLICABLE LAW. EXCEPT WHEN OTHERWISE STATED IN WRITING THE COPYRIGHT HOLDERS AND/OR OTHER PARTIES PROVIDE THE PROGRAM "AS IS" WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESSED OR IMPLIED, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE. THE ENTIRE RISK AS TO THE QUALITY AND PERFORMANCE OF THE PROGRAM IS WITH YOU. SHOULD THE PROGRAM PROVE DEFECTIVE, YOU ASSUME THE COST OF ALL NECESSARY SERVICING, REPAIR OR CORRECTION.

16. Limitation of Liability.

IN NO EVENT UNLESS REQUIRED BY APPLICABLE LAW OR AGREED TO IN WRITING WILL ANY COPYRIGHT HOLDER, OR ANY OTHER PARTY WHO MODIFIES AND/OR CONVEYS THE PROGRAM AS PERMITTED ABOVE, BE LIABLE TO YOU FOR DAMAGES, INCLUDING ANY GENERAL, SPECIAL, INCIDENTAL OR CONSEQUENTIAL DAMAGES ARISING OUT OF THE USE OR INABILITY TO USE THE PRO-

GRAM (INCLUDING BUT NOT LIMITED TO LOSS OF DATA OR DATA BEING RENDERED INACCURATE OR LOSSES SUSTAINED BY YOU OR THIRD PARTIES OR A FAILURE OF THE PROGRAM TO OPERATE WITH ANY OTHER PROGRAMS), EVEN IF SUCH HOLDER OR OTHER PARTY HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGES.

17. Interpretation of Sections 15 and 16.

If the disclaimer of warranty and limitation of liability provided above cannot be given local legal effect according to their terms, reviewing courts shall apply local law that most closely approximates an absolute waiver of all civil liability in connection with the Program, unless a warranty or assumption of liability accompanies a copy of the Program in return for a fee.

END OF TERMS AND CONDITIONS

How to Apply These Terms to Your New Programs

If you develop a new program, and you want it to be of the greatest possible use to the public, the best way to achieve this is to make it free software which everyone can redistribute and change under these terms.

To do so, attach the following notices to the program. It is safest to attach them to the start of each source file to most effectively state the exclusion of warranty; and each file should have at least the "copyright" line and a pointer to where the full notice is found.

> <one line to give the program's name and a brief idea of what it does.> Copyright (C) <year> <name of author>
>
> This program is free software: you can redistribute it and/or modify it under the terms of the GNU General Public License as published by the Free Software Foundation, either version 3 of the License, or (at your option) any later version.
>
> This program is distributed in the hope that it will be useful, but WITHOUT ANY WARRANTY; without even the implied warranty of MERCHANTABILITY or FITNESS FOR A PARTICULAR PURPOSE. See the GNU General Public License for more details.
>
> You should have received a copy of the GNU General Public License along with this program. If not, see <http://www.gnu.org/licenses/>.

Also add information on how to contact you by electronic and paper mail.

If the program does terminal interaction, make it output a short notice like this when it starts in an interactive mode:

> <program> Copyright (C) <year> <name of author>
>
> This program comes with ABSOLUTELY NO WARRANTY; for details type show w.
>
> This is free software, and you are welcome to redistribute it
>
> under certain conditions; type show c for details.

The hypothetical commands show w and show c should show the appropriate parts of the General Public License. Of course, your program's commands might be different; for a GUI interface, you would use an "about box".

You should also get your employer (if you work as a programmer) or school, if any, to sign a "copyright disclaimer" for the program, if necessary. For more information on this, and how to apply and follow the GNU GPL, see <http://www.gnu.org/licenses/>.

The GNU General Public License does not permit incorporating your program into proprietary programs. If your program is a subroutine library, you may consider it more useful to permit linking proprietary applications with the library. If this is what you want to do, use the GNU Lesser General Public License instead of this License. But first, please read <http://www.gnu.org/philosophy/why-not-lgpl.html>.

4.1.10 9.Help

If you have any questions, please email us to the following address: Octopustoolkit@gmail.com

When sending an email, please include the following text in the title.

- `Recommend`

```
[Question-Type] Title
```

9-1.Example

- [Octopus-toolkit] What is the latest version of Octopus-toolkit?
- [Error] How to resolve the Err001?
- [Question] Is the Octopus-toolkit capable of ChIP-Seq analysis?
- [Request] Can you add a previous mouse genome version (mm8)?
- [Etc] What is your name?

4.1.11 10.Octopus-toolkit for Windows(alpha version)

10-1.Development Environment

- Window version : 7
- Eclipse : Neon.1a Service Release(4.6.1)
- Language : Java Programming language (JDK1.8)
- Graphic User Interface(GUI) : Swing & Windowbuilder

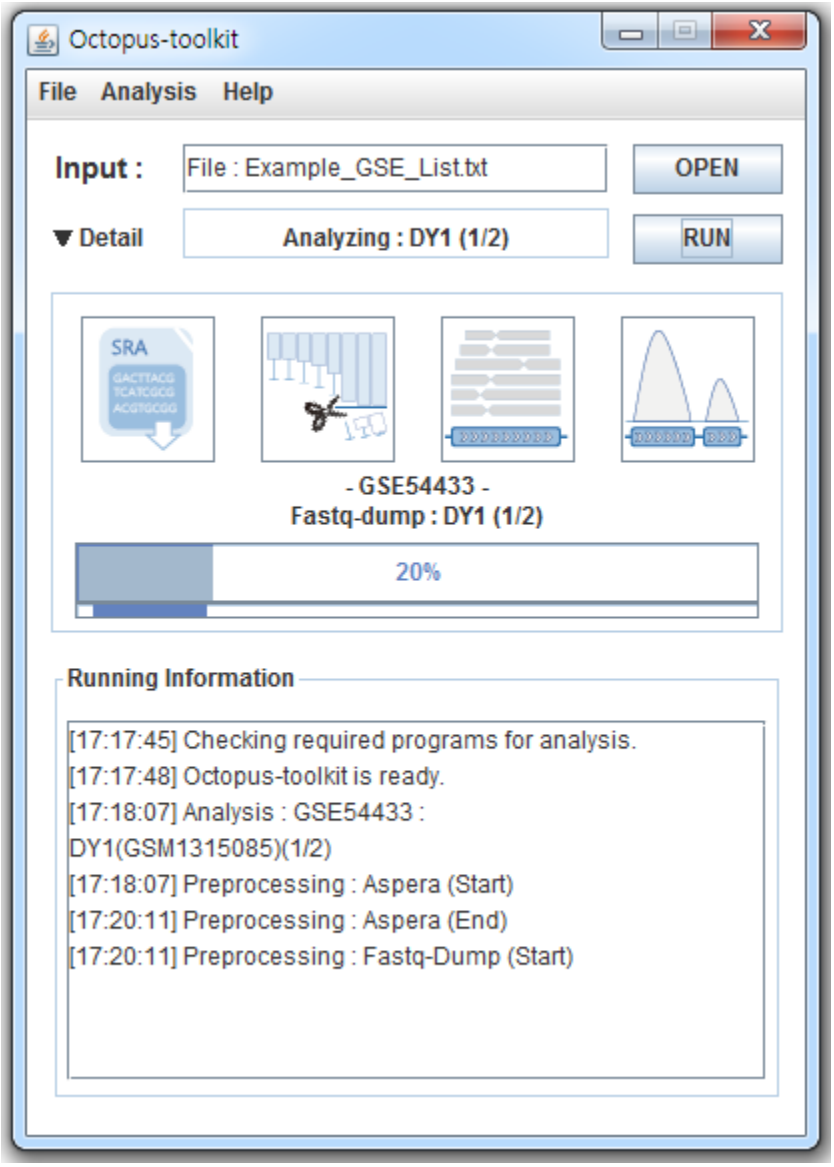
10-2. Requirement

To run the Octopus-toolkit, Java 8 (JDK, Java Development ToolKit) or higher must be installed on your computer.

10-3.Alpha test version

Octopus-toolkit Alpha version : (`Octopus-toolkit_win_test_version.exe`)

- Running window. (Test)



4.1.12 11. 3rd Party Tools

11-1.Version 2.2.0

Octopus-toolkit utilizes the following 3rd party tools during the process.

3rd party tool	Version	Function
`Aspera`	v3.7.2.141527	Download SRA files from NCBI
`SRAToolkit`	v2.9.2	Convert SRA files to Fastq files
`FastQC`	v0.11.5	Quality check for raw data
`Trimmomatic`	v0.36	Trimming for adapter sequence and portions of low-quality reads
`Hisat2`	v2.1.0	Indexing and Mapping to reference genome
`STAR`	v2.5.1	Indexing and Mapping to reference genome for RNA-Seq
`Homer`	v4.10.1	Create bigWig for visualization and Detect enriched regions by mapped reads
`Bwtool`, `libbeato`	v1.0	Calculate normalized values from bigWig files
`R`	v3.1	Draw the heatmap and line plot
`IGV`	v2.7.2	Explore the genome with processed data (bigWig files)
`Samtools`	v1.5	Sorting and Indexing the mapped reads