

Docexpress –

A galaxy docker for estimation of Expression profiling in RNA-seq



SPEAKER: Ping-Heng Hsieh

DATE: 2018/12/07



LAB OF System Biology & Network Biology

中央研究院資訊科學研究所

@iis, Academia Sinica, TAIWAN

系統生物學與網路生物學實驗室

- » Introduction for Docexpress
- » Walkthrough – Estimate Expression Profiling
- » Introduction for Docmethyl

1.

INTRODUCTION FOR DOCEXPRESS

Introduction for Docexpress

4

Docexpress

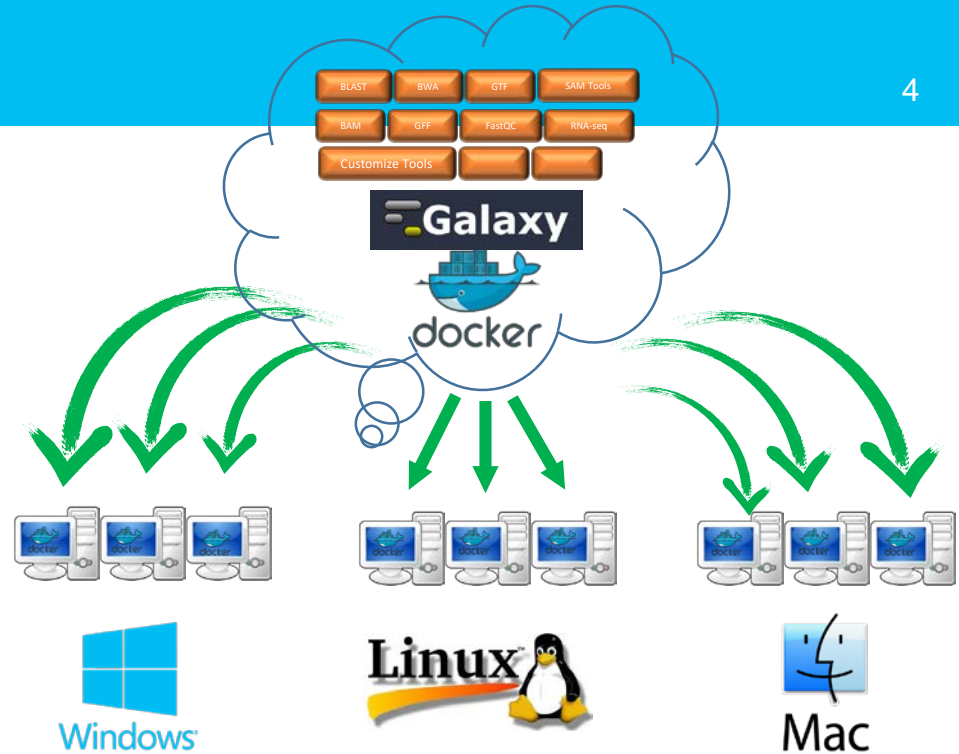
- » Galaxy + Docker

Galaxy

- » “Galaxy is an open source, web-based platform for data intensive biomedical research.” – by Galaxy

Docker

- » “Docker is an open platform for developing, shipping, and running applications.” – by Docker docs



Docexpress

- » For alleviating the burden on processing massive NGS data, we create the workflows based on galaxy/ Docker to **estimate expression profiling in RNA-seq.**
- » After the data pre-processing done, the expression profiling can be submitted to MOLAS, is a robust web application that can take gene expression data (FPKM/TPM) from different libraries as inputs, **map these expressed genes with build-in annotations for further analyses** and reveal biological meaning of the complex data in the intuitive interface.



1. Install Docker on your local system [[Ref.](#)]
2. Open a terminal or Use command mode
3. Following the “Install & Usage” on our docker hub [[link](#)]

- » Create data store directory “galaxy_guest”

```
md galaxy_guest #for windows command line
```

- » Pull the docexpress images

```
docker pull lsbnb/docexpress_fastqc
```

- » Run docexpress

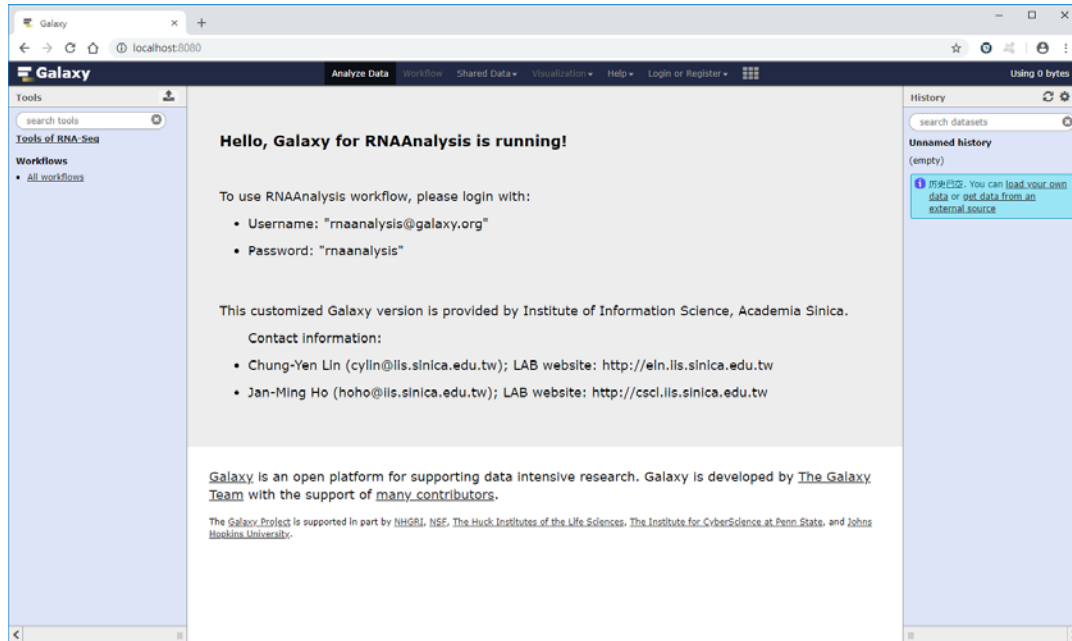
```
docker run -d -t -i -p 8080:80 -p 8021:21 -p 8022:22 -v  
%cd%/galaxy_guest:/root/galaxy/database/ftp/rnaanalysis@galaxy.org/  
lsbnb/docexpress_fastqc /bin/bash
```

- » Open the browser and input the address

“<http://localhost:8080/>”

Lunch succeeded!

8



The screenshot shows a web browser window with the URL localhost:8080. The page title is "Galaxy" and the navigation bar includes "Analyze Data", "Workflow", "Shared Data", "Visualization", "Help", and "Login or Register". The main content area displays a welcome message and login instructions for RNAAnalysis. The left sidebar shows "Tools" and "Workflows" sections. The right sidebar shows "History" and "Using 0 bytes".

Hello, Galaxy for RNAAnalysis is running!

To use RNAAnalysis workflow, please login with:

- Username: "rnaanalysis@galaxy.org"
- Password: "rnaanalysis"

This customized Galaxy version is provided by Institute of Information Science, Academia Sinica.

Contact information:

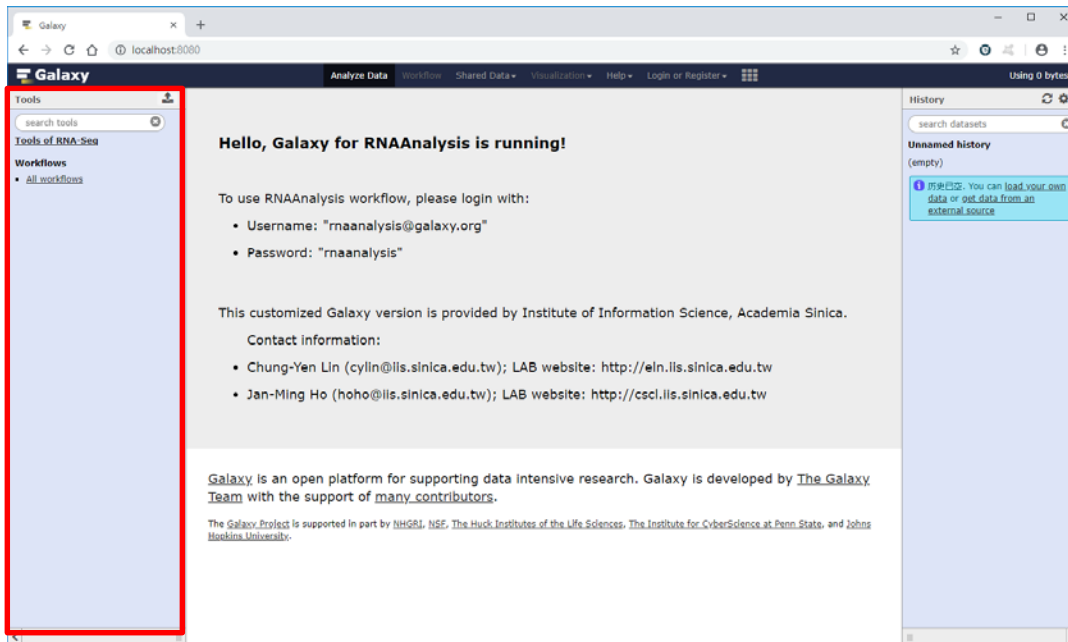
- Chung-Yen Lin (cylln@iis.sinica.edu.tw); LAB website: <http://ein.iis.sinica.edu.tw>
- Jan-Ming Ho (hoho@iis.sinica.edu.tw); LAB website: <http://csci.iis.sinica.edu.tw>

Galaxy is an open platform for supporting data intensive research. Galaxy is developed by [The Galaxy Team](#) with the support of [many contributors](#).

The Galaxy Project is supported in part by [NH&MRC](#), [NSF](#), [The Huck Institutes of the Life Sciences](#), [The Institute for CyberScience at Penn State](#), and [Johns Hopkins University](#).

Lunch succeeded!

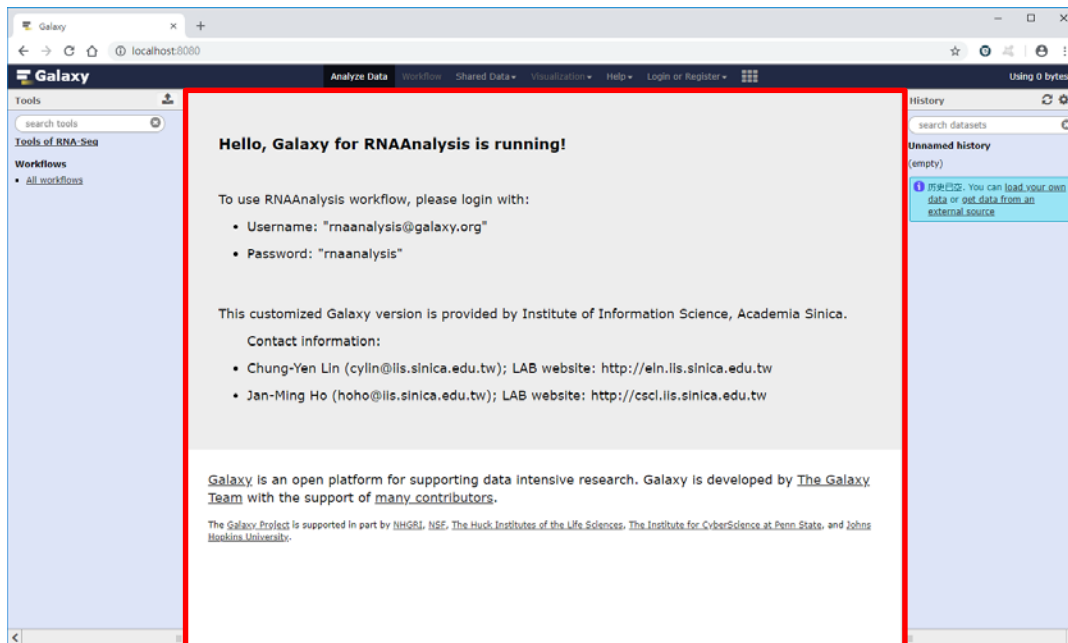
Tool panel: which contains all tools and workflows



Lunch succeeded!

10

Main : show the homepage and execution result...etc



The screenshot shows the Galaxy web interface in a browser window. The address bar displays 'localhost:8080'. The main content area, highlighted with a red border, contains the following text:

Hello, Galaxy for RNAAnalysis is running!

To use RNAAnalysis workflow, please login with:

- Username: "rnaanalysis@galaxy.org"
- Password: "rnaanalysis"

This customized Galaxy version is provided by Institute of Information Science, Academia Sinica.

Contact information:

- Chung-Yen Lin (cylln@iis.sinica.edu.tw); LAB website: <http://ein.iis.sinica.edu.tw>
- Jan-Ming Ho (hoho@iis.sinica.edu.tw); LAB website: <http://csci.iis.sinica.edu.tw>

Galaxy is an open platform for supporting data intensive research. Galaxy is developed by [The Galaxy Team](#) with the support of [many contributors](#).

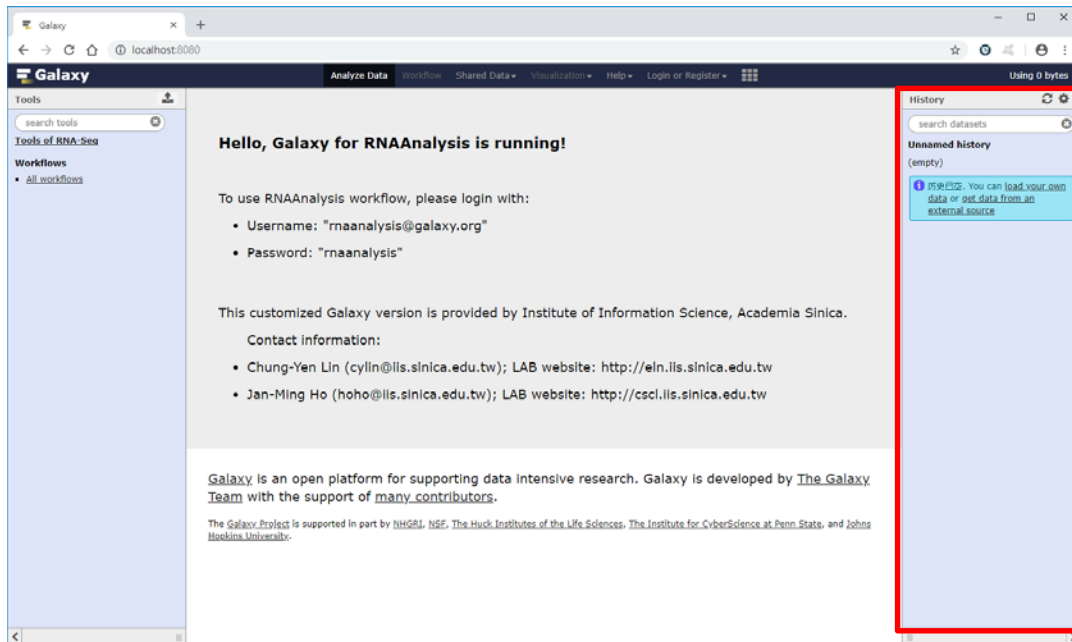
The Galaxy Project is supported in part by [NH&MRC](#), [NSF](#), [The Huck Institutes of the Life Sciences](#), [The Institute for CyberScience at Penn State](#), and [Johns Hopkins University](#).

The interface also shows a 'Tools' sidebar on the left and a 'History' sidebar on the right. A blue notification box in the history sidebar reads: '消息: You can load your own data or get data from an external source'.

Lunch succeeded!

History: contains all input and output data and result file.

Please get on Galaxy 101 for more details [[link](#)]



2.

WALKTHROUGH

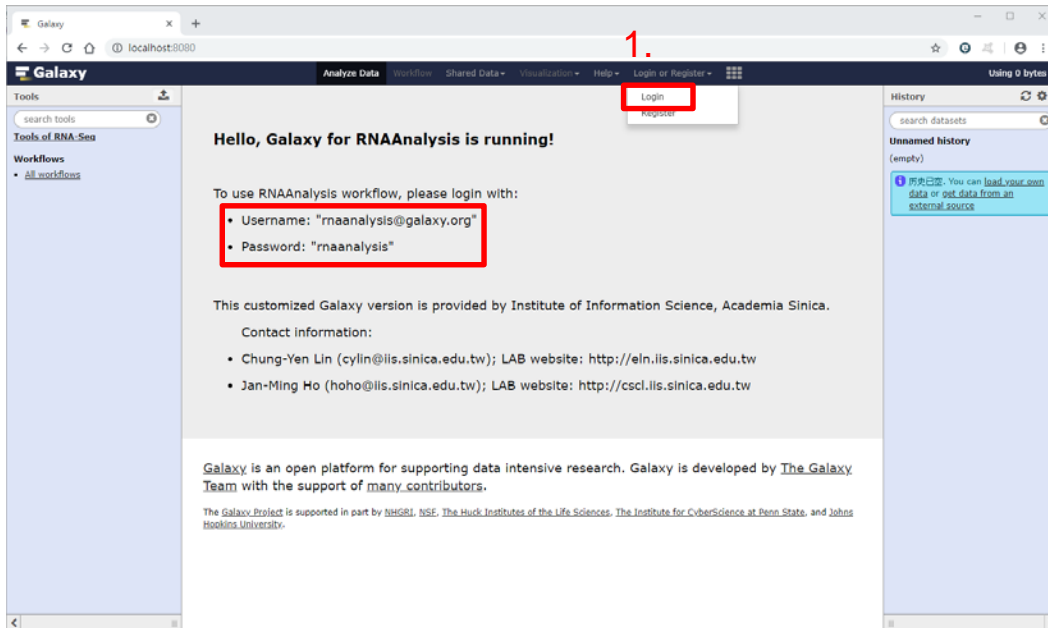
– Estimate Expression Profiling

Download test data

- » Download [link]
- » Put all test data into directory “galaxy_guest”

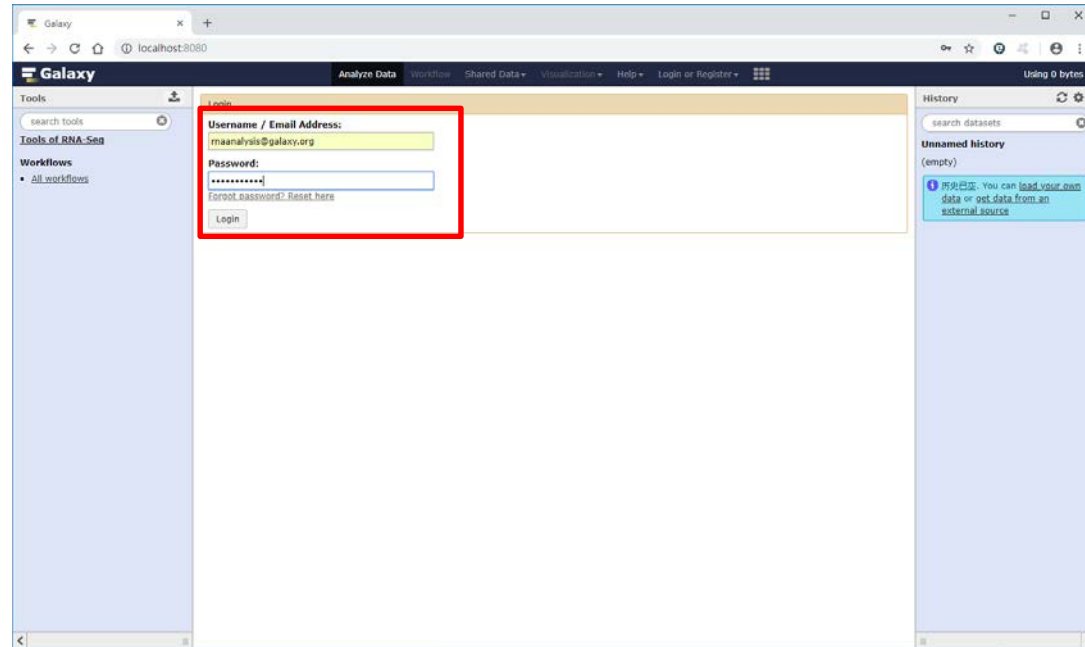
Login Galaxy

Login for full accessing all functions of Docexpress



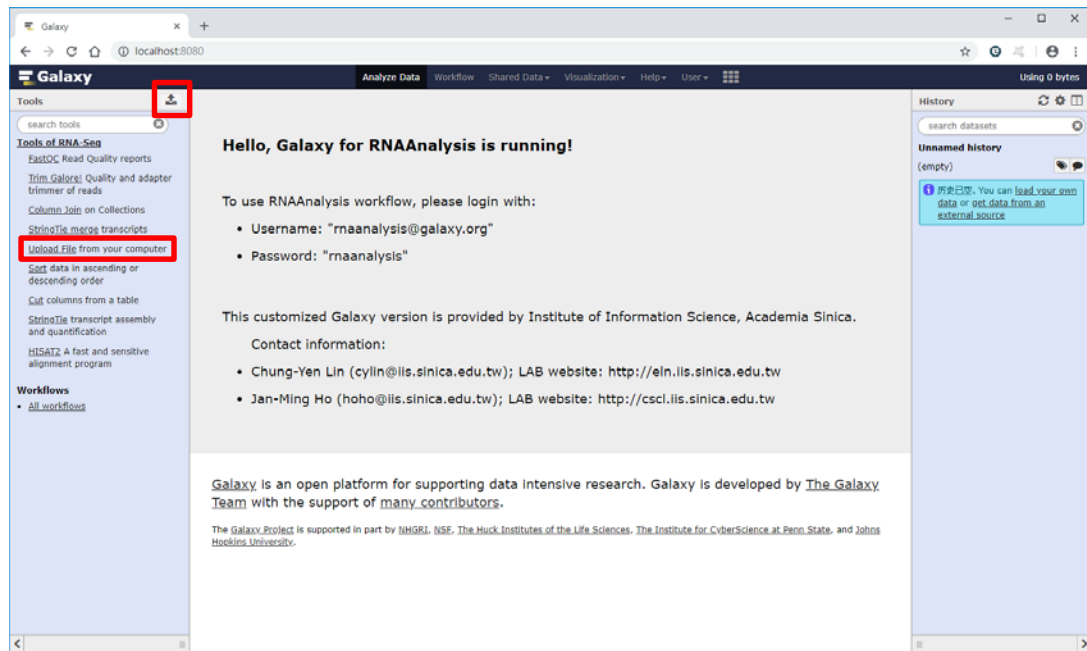
Login Galaxy

Login for full accessing all functions of Docexpress



Upload files

Two ways for calling the upload function

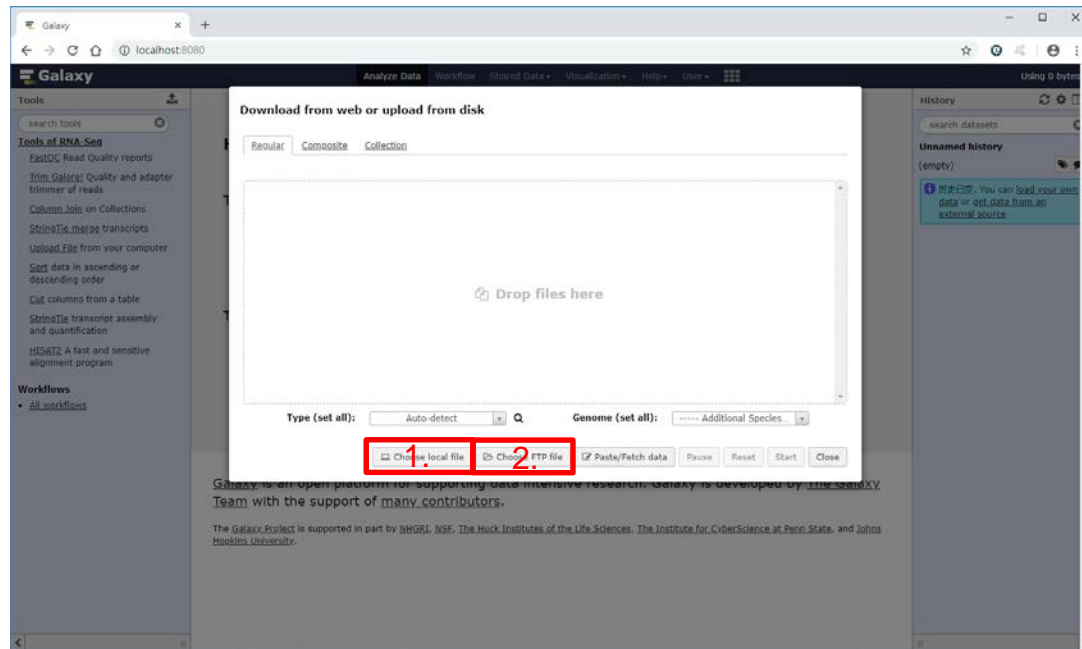


Upload files

17

Two ways for uploading data,

1. Single file < 2 GB
2. No limitation, depend on the capability of local system



Upload files

18

1. Choose FTP file
2. Check all files
3. Close selection window

The screenshot shows the Galaxy web interface with a dialog box titled "Download from web or upload from disk". The dialog box has three red boxes with numbers 1, 2, and 3. Box 1 is on the "Choose FTP file" button. Box 2 is on the "Name" column header of the file list. Box 3 is on the "Settings" column header of the file list.

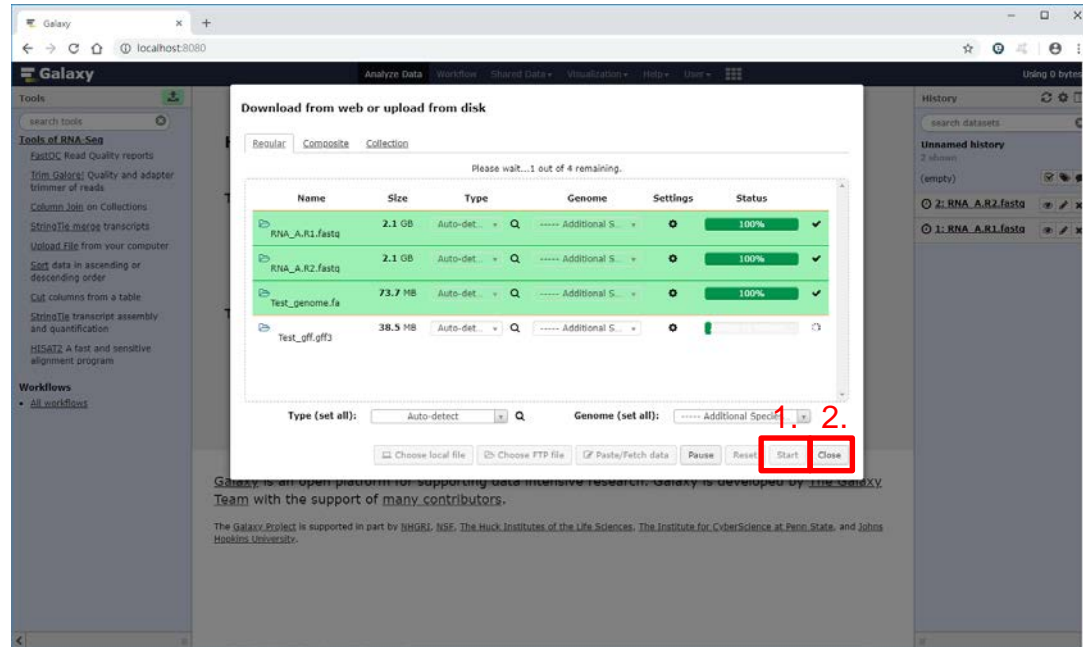
The dialog box contains the following table:

Name	Size	Type	Genome	Settings	Status
RNA_A.R1.fastq	2.1 GB	FASTQ			
RNA_A.R2.fastq	2.1 GB	FASTQ			
Test_genome.fa	73.7 MB	FASTA			
Test_gff.gff3	38.5 MB	GFF3			

The dialog box also has a "Choose local file" button, a "Choose FTP file" button (highlighted with a red box 1), a "Paste/Fetch data" button, and a "Start" button.

Upload files

1. Upload files
2. Close upload function



Run Workflow : 1-1: Transcript_FPKM

20

1. Choose “All workflows”
2. Choose “1-1: Transcription_FPKM”
3. Run the workflow

The screenshot displays the Galaxy web interface. On the left, the 'Tools' panel is visible, with 'All workflows' highlighted in red and labeled '1.'. The main area shows a table of 'Your workflows'. The first row, '1-1: Transcript_FPKM', is highlighted in red and labeled '2.'. A context menu is open over this row, with the 'Run' option highlighted in red and labeled '3.'. The table has columns for Name, Tags, Owner, # of Steps, Published, and Show in tools panel. The right sidebar shows a 'History' panel with a search for datasets and a list of datasets including '4: Test_off.off3', '2: Test_genome.fa', '2: RNA_A.R2.fasta', and '1: RNA_A.R1.fasta'.

Name	Tags	Owner	# of Steps	Published	Show in tools panel
1-1: Transcript_FPKM		You	7	No	<input type="checkbox"/>
		You	7	No	<input type="checkbox"/>
		You	7	No	<input type="checkbox"/>
1-6: Gene_TPM_Single_End		You	6	No	<input type="checkbox"/>
2: Expression_table_for_Molas		You	1	No	<input type="checkbox"/>

Run Workflow : 1-1: Transcript_FPKM

1. Choose RAN-seq data as the input

1.

Workflow: 1-1: Transcript_FPKM Run workflow

History Options

Send results to a new history

Yes No

1: Trim Galore! (Galaxy Version 0.4.3.1)

Is this library paired- or single-end?

Paired-end

Reads In FASTQ format

1: RNA_A.R1.fastq

Reads In FASTQ format

2: RNA_A.R2.fastq

Adapter sequences to be trimmed

Automatic detection

Trims 1 bp off every read from its 3' end.

false

Remove N bp from the 3' end of read 1

Empty.

Remove N bp from the 3' end of read 2

Empty.

Trim Galore! advanced settings

Use defaults

RRBS specific settings

Use defaults (no RRBS)

Job Post Actions

Rename output 'trimmed_reads_pair1' to '#(input_mate1|basename).trimmed'. Rename output 'trimmed_reads_pair2' to '#(input_mate2|basename).trimmed'. Delete parent datasets of this step created in this workflow that aren't flagged as outputs. Hide

Run Workflow : 1-1: Transcript_FPKM

1. Choose RAN-seq data as the input
2. Select genome

2.

Workflow: 1-1: Transcript_FPKM Run workflow

output 'unpaired_reads_2'; Hide output 'unpaired_reads_1'; Hide output 'trimmed_reads_unpaired_collection'; Hide output 'trimmed_reads_paired_collection'.

2: HISAT2 (Galaxy Version 2.1.0)

Source for the reference genome

Use a genome from history

Select the reference genome

3: Test_genome.fa

Single-end or paired-end reads?

Paired-end

FASTA/Q file #1

Output dataset 'trimmed_reads_pair1' from step 1

FASTA/Q file #2

Output dataset 'trimmed_reads_pair2' from step 1

Specify strand information

Forward (FR)

Paired-end options

Use default values

Summary Options

Advanced Options

Job Post Actions

Hide output 'output_unaligned_reads_r'. Delete parent datasets of this step created in this workflow that aren't flagged as outputs. Hide output 'output_aligned_reads_f'. Hide output 'output_aligned_reads_r'. Hide output 'summary_file'. Hide output 'output_alignments'. Rename output 'output_alignments' to '#[input_1]basename'.bam'. Hide output 'output_unaligned_reads_f'.

3: FastQC_R1 (Galaxy Version 0.72)

Short read data from your current history

Output dataset 'trimmed_reads_pair1' from step 1

Contaminant list

Run Workflow : 1-1: Transcript_FFKM

1. Choose RAN-seq data as the input
2. Select genome
3. Select GFF/GTF file
4. Run workflow

Workflow: 1-1: Transcript_FPKM 4. Run workflow

5: StringTie (Galaxy Version 1.3.3.1)

Input mapped reads
Output dataset 'output_alignments' from step 2

Specify strand information
Unstranded

Select 'Forward (FR)' if your reads are from a forward-stranded library, 'Reverse (RF)' if your reads are from a reverse-stranded library, or 'Unstranded' if your reads are not from a stranded library. See Help section below for more information. Default: Unstranded

Use a reference file to guide assembly?
Use reference GTF/GFF3

Reference file
Use a file from history

GTF/GFF3 dataset to guide assembly
4: Test_gff.gff3

Use Reference transcripts only

true

Output files for differential expression?
Ballgown

Output coverage file?
false

Advanced Options

Job Post Actions
Hide output 'gene_counts'. Hide output 'intron_transcript_mapping'. Delete parent datasets of this step created in this workflow that aren't flagged as outputs. Hide output 'coverage'. Hide output 'output_gtf'. Rename output 'transcript_expression' to 'e (input_bam)basename()'. Hide output 'transcript_counts'. Hide output 'legend'. Hide output 'gene_abundance_estimation'. Hide output 'exon_expression'. Hide output 'transcript_expression'. Hide output 'exon_transcript_mapping'. Hide output 'intron_expression'.

6: Sort (Galaxy Version 1.0.3)

Run Workflow : 1-1: Transcript_FFKM

24

1. Choose RAN-seq data as the input
2. Select genome
3. Select GFF/GTF file
4. Run workflow

The screenshot shows the Galaxy web interface. A green notification box at the top center contains the following text: "Successfully invoked workflow 1-1: Transcript_FFKM. You can check the status of queued jobs and view the resulting data by refreshing the History pane. When the job has been run the status will change from 'running' to 'finished' if completed successfully or 'error' if problems were encountered." The History panel on the right side of the interface shows a list of workflow steps, with the first four steps highlighted in green: "1: RNA_A.R1.fasto", "2: RNA_A.R2.fasto", "3: Test_genome.fa", and "4: Test_gffL1". A red number "4." is positioned to the right of the history panel.

Run Workflow : 1-1: Transcript_FFKM

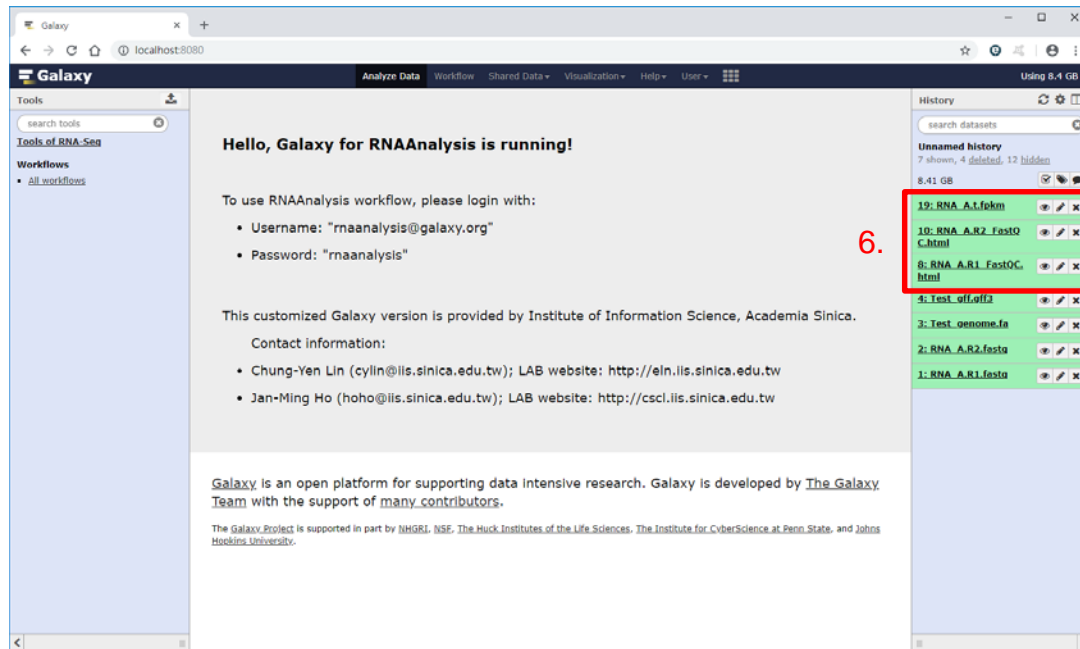
25

1. Choose RAN-seq data as the input
2. Select genome
3. Select GFF/GTF file
4. Run workflow
5. Repeat 1. - 4. until all RNA-seq data selected and execute the workflow

The screenshot displays the Galaxy web interface. At the top, the browser address bar shows the URL: localhost:8080/workflow/run?id=63cd3858d057a6d1. The main header includes the Galaxy logo and navigation tabs for Analyze Data, Workflow, Shared Data, Visualization, Help, and User. A green notification box in the center reads: "Successfully invoked workflow 1-1: Transcript_FFKM. You can check the status of queued jobs and view the resulting data by refreshing the History pane. When the job has been run the status will change from 'running' to 'finished' if completed successfully or 'error' if problems were encountered." On the left, the Tools panel shows a search for "Tools of RNA-Seq" and a list of workflows. On the right, the History pane shows a list of 13 jobs, with the first job, "1: RNA_A.R1.fasto", highlighted in green, indicating it is the current job being tracked.

Run Workflow : 1-1: Transcript_FFKM

1. Choose RAN-seq data as the input
2. Select genome
3. Select GFF/GTF file
4. Run workflow
5. Repeat 1. - 4. until all RNA-seq data selected and execute the workflow
6. Outputs of workflow,
(1) FastQC.html
(2) Samlpe.t.fpkm



Output of Workflow : 1-1: Transcript_FFKM

- Outputs of workflow,
(1) Sample.t.fpkkm
(2) FastQC.html

The screenshot shows the Galaxy web interface. The main area displays a table with two columns, labeled '1' and '2'. The table contains a list of transcript identifiers (TEU1.t1, TEU10.t1, etc.) and their corresponding FPKM values. A callout bubble points to the table with the text 'Download this file'. To the right, the 'History' panel shows a list of workflow outputs, including '19: RNA_A.1.fpkkm', '10: RNA_A.R2_FastQ C.html', '8: RNA_A.R1_FastQC.html', '4: Test_off.off3', '3: Test_genome.fa', '2: RNA_A.R2.fastq', and '1: RNA_A.R1.fastq'. A callout bubble points to the '19: RNA_A.1.fpkkm' entry with the text 'View the result in the main section'. A red box highlights the '19: RNA_A.1.fpkkm' entry in the history panel, and a callout bubble points to it with the text '(1)'.

1	2
TEU1.t1	0.000000
TEU10.t1	0.000000
TEU100.t1	0.000000
TEU1000.t1	0.000000
TEU10000.t1	116.277184
TEU10001.t1	0.000000
TEU10002.t1	130.300903
TEU10003.t1	24.849367
TEU10004.t1	0.000000
TEU10004.t2	0.000000
TEU10004.t3	0.000000
TEU10004.t4	0.000000
TEU10004.t5	2.875377
TEU10005.t1	179.968903
TEU10006.t1	8.956815
TEU10006.t2	0.000000
TEU10006.t3	0.000000
TEU10006.t4	0.000000
TEU10007.t1	0.000000
TEU10008.t1	5.959774
TEU10008.t2	0.000000
TEU10009.t1	334.145691
TEU1001.t1	0.000000
TEU10010.t1	0.000000
TEU10010.t2	0.000000
TEU10011.t1	786.189148
TEU10012.t1	84.993212
TEU10013.t1	10.293010
TEU10014.t1	101.603466
TEU10015.t1	177.307480
TEU10015.t2	0.000000
TEU10016.t1	3.924205
TEU10017.t1	0.000000
TEU10017.t2	0.000000
TEU10017.t3	0.000000
TEU10017.t4	0.000000
TEU10017.t5	0.000000
TEU10017.t6	0.000000

Output of Workflow : 1-1: Transcript_FFKM

- Outputs of workflow,
(1) Samlpe.t.fpkkm
(2) FastQC.html

RNA_A_R1_trimmed FastQC Report

FastQC Report
Wed 5 Dec 2018
RNA_A_R1_trimmed

Summary

- Basic Statistics
- Per base sequence quality
- Per tile sequence quality
- Per sequence quality scores
- Per base sequence content
- Per sequence GC content
- Per base N content
- Sequence Length Distribution
- Sequence Duplication Levels
- Overrepresented sequences
- Adapter Content

Basic Statistics

Measure	Value
Filename	RNA_A_R1_trimmed
File type	Conventional base calls

View the result in the main section

10: RNA_A_R1_FastQC.html

(2)

Run Workflow : 2: Expression_table_for_Molas

29

1. Choose “All workflows”
2. Choose “2:Expression_table_for_Molas”
3. Run the workflow

The screenshot shows the Galaxy web interface. On the left sidebar, under 'Tools of RNA-Seq', the 'Workflows' section has 'All workflows' highlighted with a red box and labeled '1.'. The main panel shows a table of workflows:

Name	Tags	Owner	# of Steps	Published	Show in tools panel
1-1: Transcript_FPKM		You	7	No	<input type="checkbox"/>
1-2: Gene_FPKM		You	7	No	<input type="checkbox"/>
1-3: Gene_TPM		You	7	No	<input type="checkbox"/>
1-4: Transcript_FPKM_Single_End		You	6	No	<input type="checkbox"/>
1-5: Gene_FPKM_Single_End		You	6	No	<input type="checkbox"/>
1-6: Gene_TPM_Single_End		You	6	No	<input type="checkbox"/>
2: Expression_table_for_Molas		You	1	No	<input type="checkbox"/>

A dropdown menu is open for the workflow '2: Expression_table_for_Molas', with the 'Run' button highlighted in red and labeled '3.'. The right sidebar shows a 'History' panel with a list of datasets.

Run Workflow : 2: Expression_table_for_Molas

30

1. Choose all .t.fpkm files with pressing the “Ctrl”
2. Run workflow

The screenshot shows the Galaxy web interface for running a workflow. The browser address bar shows the URL: localhost:8080/workflow/run?id=911dde3ddb677bcd. The workflow is titled "Workflow: 2: Expression_table_for_Molas".

On the left sidebar, under "Tools of RNA-Seq", the "Workflows" section is expanded to show "All workflows".

The main panel displays the workflow configuration for step 1: "FPKM_table_generate (Galaxy Version 0.0.2)". The "History Options" section includes a "Send results to a new history" toggle set to "No". The "Tabular files" section is highlighted with a red box and labeled "1.", showing a list of files: "36: RNA_B.t.fpkm", "19: RNA_A.t.fpkm", "4: Test_off.off3", and "3: Test_genome.fa (as tabular)".

Below the file list, there are options for "Number of Header lines in each item" (set to 1), "Fill character" (set to "."), and "Additional datasets to create" (Nothing selected). The "Job Post Actions" section is set to "Rename output 'tabular_output' to 'fpkm.table'".

On the right side, the "History" panel shows a list of datasets, including "26: RNA_B.t.fpkm", "27: RNA_B.R2_FastQC.html", "25: RNA_B.R1_FastQC.html", "21: RNA_B.R2.fasta", "20: RNA_B.R1.fasta", "19: RNA_A.t.fpkm", "10: RNA_A.R2_FastQC.html", "8: RNA_A.R1_FastQC.html", "4: Test_off.off3", "3: Test_genome.fa", "2: RNA_A.R2.fasta", and "1: RNA_A.R1.fasta".

A red box labeled "2." highlights the "Run workflow" button in the top right corner of the workflow configuration panel.

Run Workflow : 2: Expression_table_for_Molas

1. Result: fpkm.table

The screenshot displays the Galaxy web interface. At the top, a green notification box states: "Successfully invoked workflow 2: Expression_table_for_Molas. You can check the status of queued jobs and view the resulting data by refreshing the History pane. When the job has been run the status will change from 'running' to 'finished' if completed successfully or 'error' if problems were encountered." The main area shows a list of workflow steps in the History pane on the right. The step "27: fpkm.table" is highlighted with a red box and a red number "1." next to it. Other steps include "16: RNA_B.L.fekm", "27: RNA_B.R2_FastQC.html", "25: RNA_B.R1_FastQC.html", "21: RNA_B.R2.fastq", "20: RNA_B.R1.fastq", "19: RNA_A.L.fekm", "10: RNA_A.R2_FastQC.html", "8: RNA_A.R1_FastQC.html", "4: Test_off.off", "3: Test_genome.fa", "2: RNA_A.R2.fastq", and "1: RNA_A.R1.fastq".

Run Workflow : 2: Expression_table_for_Molas

32

1. View and download the fpkm.table for MOLAS system

The screenshot shows the Galaxy web interface. The main panel displays a table with the following columns: #KEY, RNA_A.t.fpkm_2, and RNA_B.t.fpkm_2. The table contains data for various TEU keys, such as TEU1.11, TEU10.11, TEU100.11, etc. A red box highlights a file named '27: fpkm.table' in the history panel, with a callout bubble saying 'Download this file'. Another callout bubble says 'View the result in the main section'.

#KEY	RNA_A.t.fpkm_2	RNA_B.t.fpkm_2
TEU1.11	0.000000	0.000000
TEU10.11	0.000000	0.000000
TEU100.11	0.000000	0.000000
TEU1000.11	0.000000	0.000000
TEU10000.11	116.277184	88.075256
TEU10001.11	0.000000	0.000000
TEU10002.11	130.300903	108.687370
TEU10003.11	24.849267	0.000000
TEU10004.11	0.000000	0.000000
TEU10004.12	0.000000	1.210395
TEU10004.13	0.000000	0.000000
TEU10004.14	0.000000	0.000000
TEU10004.15	2.875577	4.879134
TEU10005.11	179.968903	953.210754
TEU10006.11	8.956815	0.000000
TEU10006.12	0.000000	0.000000
TEU10006.13	0.000000	11.704576
TEU10006.14	0.000000	0.000000
TEU10007.11	0.000000	0.000000
TEU10008.11	5.959774	0.000000
TEU10008.12	0.000000	0.000000
TEU10009.11	334.145691	194.510789
TEU1001.11	0.000000	0.000000
TEU10010.11	0.000000	0.000000
TEU10010.12	0.000000	0.000000
TEU10011.11	786.189148	721.523882
TEU10012.11	64.992012	99.557846
TEU10013.11	10.293010	21.677700
TEU10014.11	101.693466	147.884201
TEU10015.11	177.307480	170.098160
TEU10015.12	0.000000	0.000000
TEU10016.11	3.924205	0.000000
TEU10017.11	0.000000	0.000000
TEU10017.12	0.000000	0.000000
TEU10017.13	0.000000	0.000000

3.

INTRODUCTION FOR DOCMETHYL

DocMethyl

- » We packed a Docker container DocMethyl to deal with raw data processing, mapping, and methylation calling/ scoring to give the summary, **mtable**, of the whole genome methylation status by the gene.
- » Mtables are uploaded to the web server EpiMOLAS_web for **linking with gene annotation databases that enable rapid data retrieval and analyses.**



1. Install Docker on your local system [[Ref.](#)]
2. Open a terminal or Use command mode
3. Following the “Install & Usage” on our docker hub [[link](#)]

THANKS!

36

Any questions?



LAB OF System Biology & Network Biology

中央研究院資訊科學研究所 @iis, Academia Sinica, TAIWAN

系統生物學與網路生物學實驗室



1. Docker - <https://www.docker.com/>
2. Docker Install - <https://docs.docker.com/install/linux/docker-ce/ubuntu/>
3. Galaxy - <https://usegalaxy.org/>
4. Docexpress - <https://hub.docker.com/r/lisbnb/docexpress/>
5. DocMethyl - <https://hub.docker.com/r/lisbnb/docmethyl/>
6. Galaxy basic usage <https://galaxyproject.github.io/training-material/topics/introduction/tutorials/galaxy-intro-101/tutorial.html>