

Galaxy for Cut Site Detection Operation Manual

Division of Biochemical, National Institutes of Health Sciences

Index

1. About Galaxy Cut Site Detection.....	2
2. Installation of Docker Desktop.....	2
3. Pull a Galaxy Cut Site Detection Docker image.....	2
4. Run a Galaxy Cut Site Detection Docker container (for first launch)	2
5. Start a Galaxy Cut Site Detection Docker container (for secondary use)	2
6. Access for Galaxy page.....	3
7. Login.....	3
8. Load files to analysis.....	3
9. Execute analysis.....	4
10. About output files.....	4
11. Termination of analysis.....	5
12. Q&A.....	5

1. About Galaxy for Cut Site Detection

Galaxy for Cut Site Detection is a Galaxy (<https://usegalaxy.org/>) based data analysis pipeline for off-target sequence such as SITE-Seq. Galaxy for Cut Site Detection identifies DSB sites with termination positions of aligned reads. Optionally, user can focus the cut sites only within the annotated region (e.g. exon). Please note that you must agree to the terms of the license as described on GitHub (<https://github.com/NIHS-DNFI/galaxy-cutsite-detection>) when you use this program.

2. Installation of Docker Desktop

Galaxy for Cut Site Detection requires Docker software. Download from Docker Desktop official site (<https://www.docker.com/products/docker-desktop>).

3. Pull a Galaxy Cut Site Detection Docker image

Pull a docker image of Galaxy for Cut Site Detection from our repository (<https://hub.docker.com/repository/docker/nihsdnfi/galaxy-cutsite-detection>).

```
$ docker pull nihsdnfi/galaxy-cutsite-detection:0.XX
```

User need to pull the latest TAG available at the time (e.g., 0.10).

4. Run a Galaxy Cut Site Detection Docker container (for first launch)

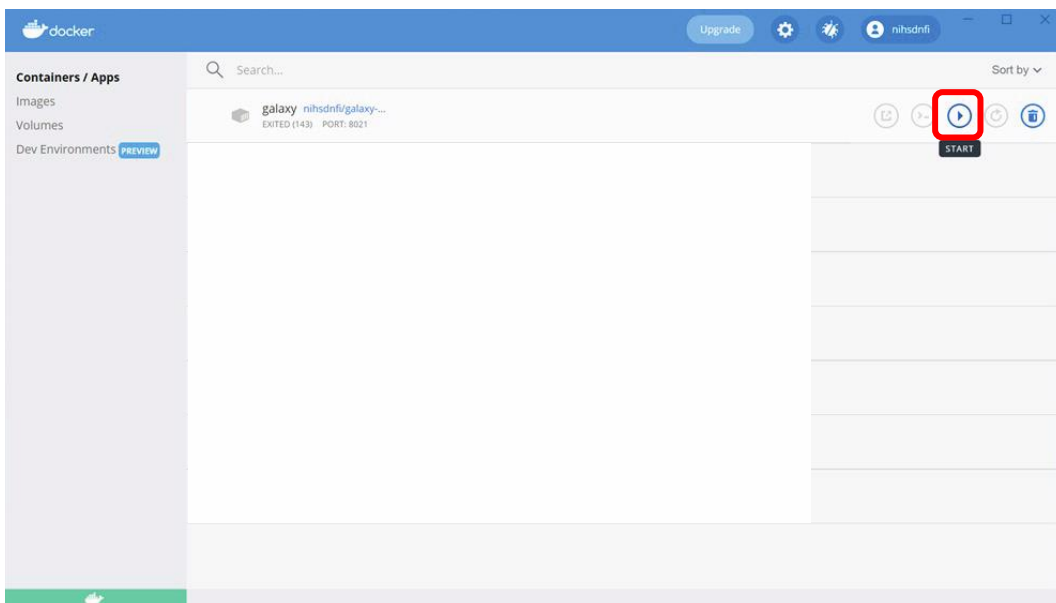
Create and start a Galaxy for Cut Site Detection container from docker image with the following docker run command at first time using.

```
$ docker run -d --name galaxy-csd -p 8080:80 nihsdnfi/galaxy-cutsite-detection:0.XX
```

Change to the latest TAG you have pulled (e.g., 0.10).

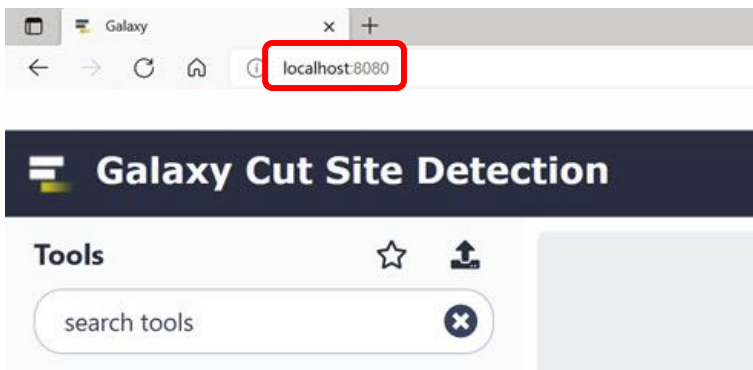
5. Start a Galaxy Cut Site Detection Docker container (for secondary use)

If you have already created the Galaxy for Cut Site Detection container, the above docker run command is not necessary; start the Docker container from Containers/Apps in the Docker Desktop window by clicking the "START" button.



6. Access for Galaxy page

After a few minutes running Docker container, you will be able to access Galaxy for Cut Site Detection via "localhost:8080" from web browser.



7. Login

Log in from Login/Register. The default Public name /Password is as follows.

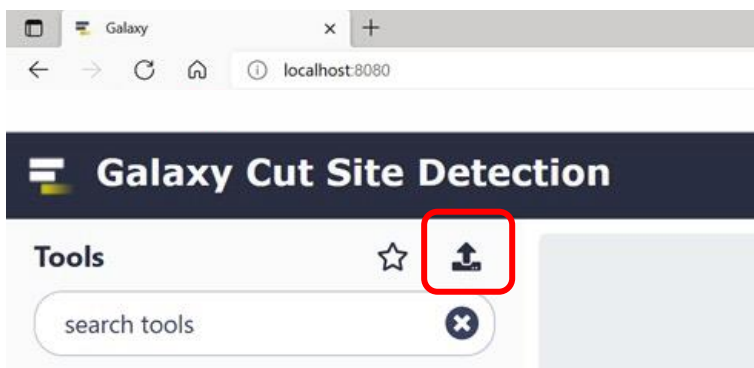
Public name: admin

Password: password

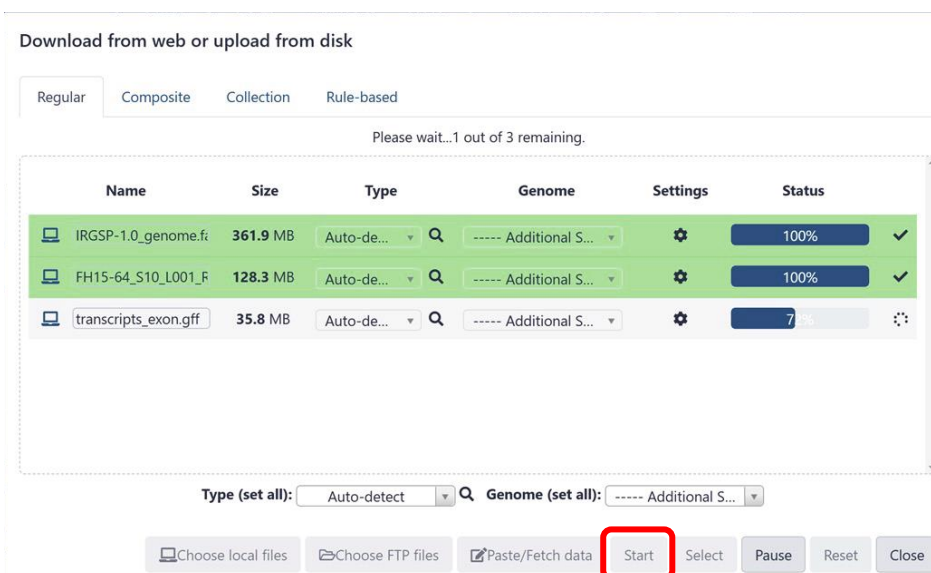
8. Upload the datasets

Upload the datasets (sequence data, reference genome, gene annotation) to analyze on Galaxy for Cut Site Detection.

Note that the sequencing data and analysis result never get known to us because the Galaxy for Cut Site Detection runs stand-alone.

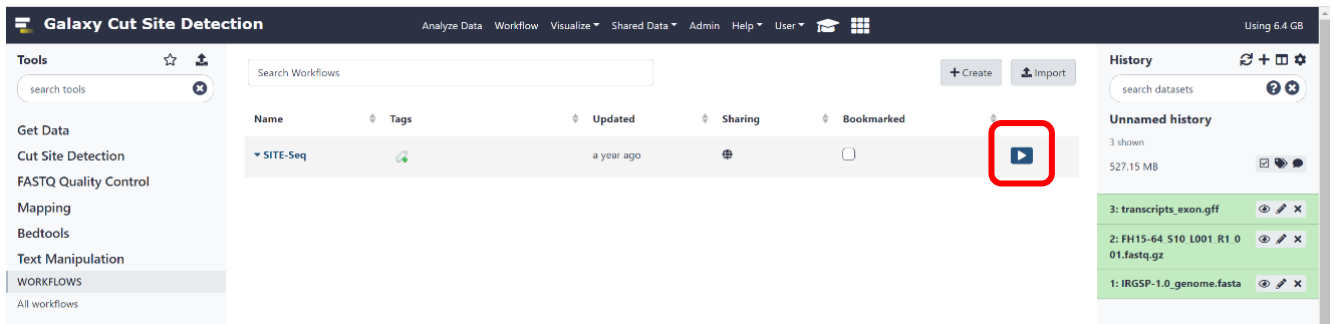


Drop the datasets into the upload window. Click "Start" button to upload the datasets.



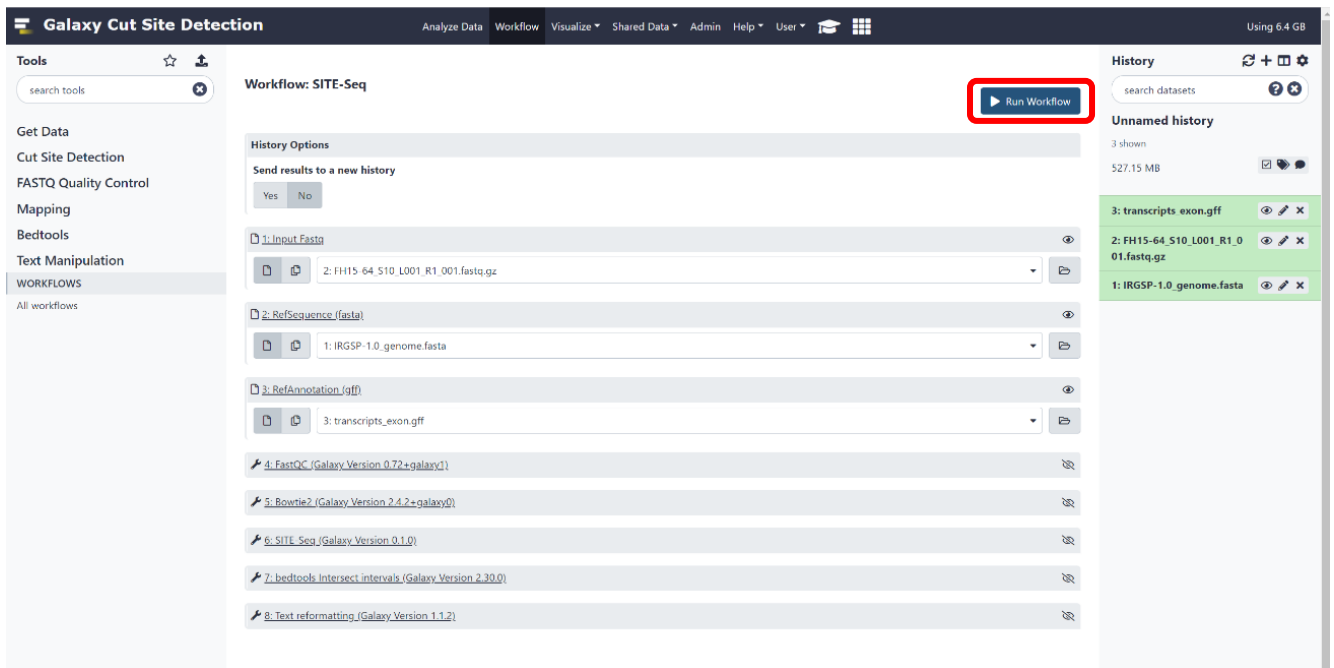
9. Execute analysis

Select Run Workflow for SITE-Seq from All workflows in the Tools menu.



Select the files to analysis for each attribute, and execute Run Workflow.

Users can change the cut site read threshold depending on sequence coverage.



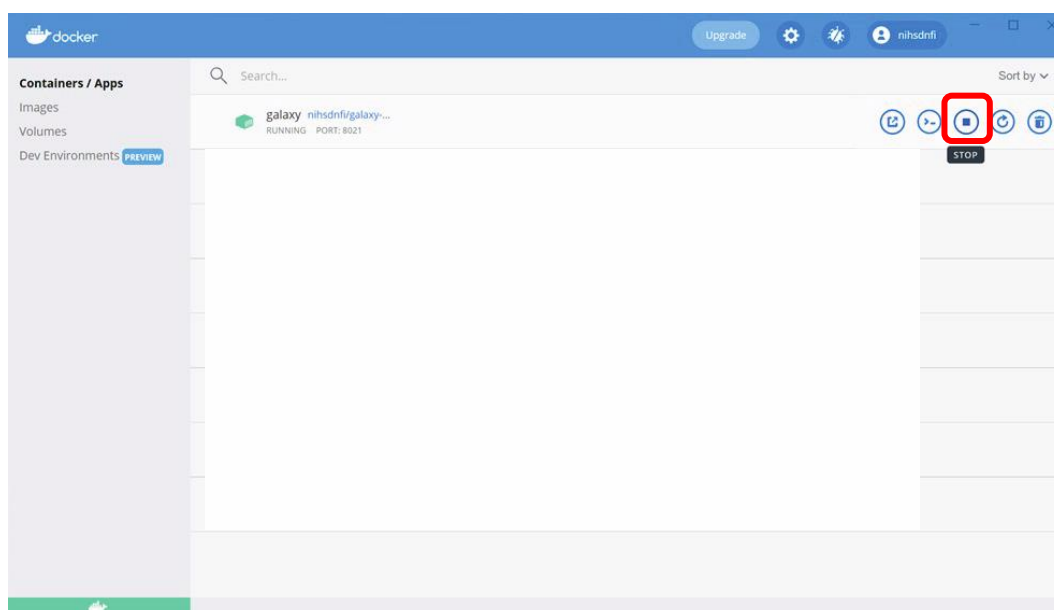
10. About output files

SITE-Seq on ~ (tabular format): The list of cut sites detected across the whole genome

Text reformatting ~: The list of cut site detected within the annotated region (e.g. exon).

11. Termination of analysis

Terminate the Docker container with the "STOP" button of the Galaxy for Cut Site Detection container from Containers/Apps in the Docker Desktop window.



12. Q&A

Q. A file without any cut site information (Text reformatting ~) in the annotation region is output.

A. Check the consistency between the analysis parameters and the annotation file to be used. The default analysis parameters extract cut sites in the region annotated as "mRNA," but should be changed according to the annotation file to be used. As an example, in the annotation file for human GRCh38, the third column contains "gene, transcript, exon..." (Upper figure). If you need only cut sites in exon, change AWK Program to "\$3=="exon", and if you need cut sites in intron or UTR, change to "\$3=="gene" (Under figure).

Galaxy Cut Site Detection										
Tools		Seqid Source Type Start End Score Strand Phase Attributes								
search tools		#gff-version 3								
		#description: evidence-based annotation of the human genome (GRCh38), version 39 (Ensembl 105)								
		#provider: GENCODE								
		#contact: gencode-help@ebi.ac.uk								
		#format: gff3								
		#date: 2021-09-02								
		#sequence-region chr1 1 248956422								
		chr1	HAVANA	gene	11869	14409	.	+	.	ID=ENSG0
		chr1	HAVANA	transcript	11869	14409	.	+	.	ID=ENST01
		chr1	HAVANA	exon	11869	12227	.	+	.	ID=exon:E
		chr1	HAVANA	exon	12613	12721	.	+	.	ID=exon:E
		chr1	HAVANA	exon	13221	14409	.	+	.	ID=exon:E
		chr1	HAVANA	transcript	12010	13670	.	+	.	ID=ENST01
		chr1	HAVANA	exon	12010	12057	.	+	.	ID=exon:E
		chr1	HAVANA	exon	12179	12227	.	+	.	ID=exon:E
		chr1	HAVANA	exon	12613	12697	.	+	.	ID=exon:E
		chr1	HAVANA	exon	12975	13052	.	+	.	ID=exon:E

Galaxy Cut Site Detection

Tools

search tools

Get Data
Cut Site Detection
FASTQ Quality Control
Mapping
Bedtools
Text Manipulation
WORKFLOWS
All workflows

Workflow: SITE-Seq

History Options
Send results to a new history

Yes
No

1: Input Fastq

39: Bowtie2 on data 4 and data 2: alignments

2: RefSequence (fasta)

4: GRCh38.p13.genome.fa

3: RefAnnotation (gff)

44: Text reformatting on data 43

4: FastQC (Galaxy Version 0.72+galaxy1)

5: Bowtie2 (Galaxy Version 2.4.2+galaxy0)

6: SITE-Seq (Galaxy Version 0.1.0)

7: bedtools Intersect intervals (Galaxy Version 2.30.0)

8: Text reformatting (Galaxy Version 1.1.2)

File to process

AWK Program

\$3=="mRNA"

Q. How can I analyze species for which there are no annotation files?

A. It is possible to get only cut site information of the whole genome. Create a decoy .gff file (only header information is required) and select it in RefAnnotation. bedtools intersect will generate an error, and analysis of Text reformatting ~ will be stopped, but an intermediated file containing cut site information of the whole genome (SITE-Seq on ~ (tabular format)) will be generated successfully.