Galaxy for Cut Site Detection Operation Manual

Division of Biochemical, National Institutes of Health Sciences

Index

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)

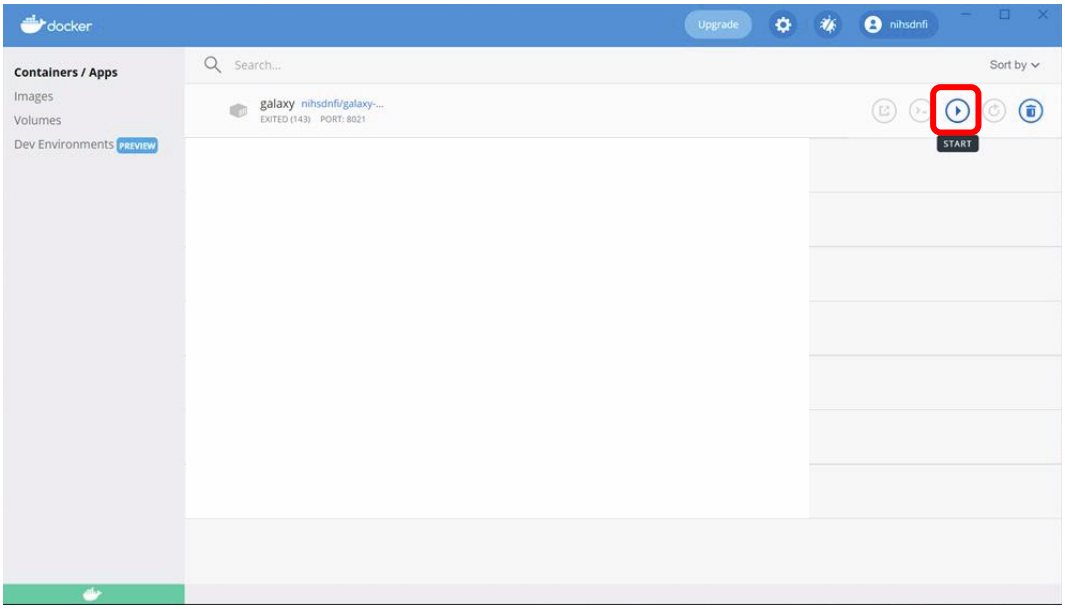
Create and start a Galaxy for Cut Site Detection container from docker image with the following docker run command at first time using.

```
$ 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).

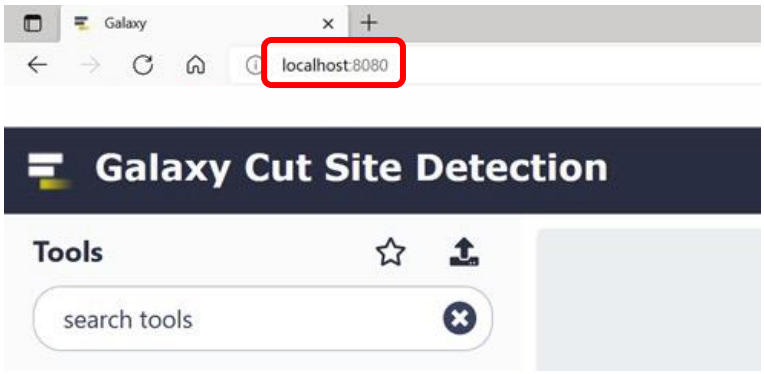
5. Start a Galaxy Cut Site Detection Docker container (for secondary use)

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

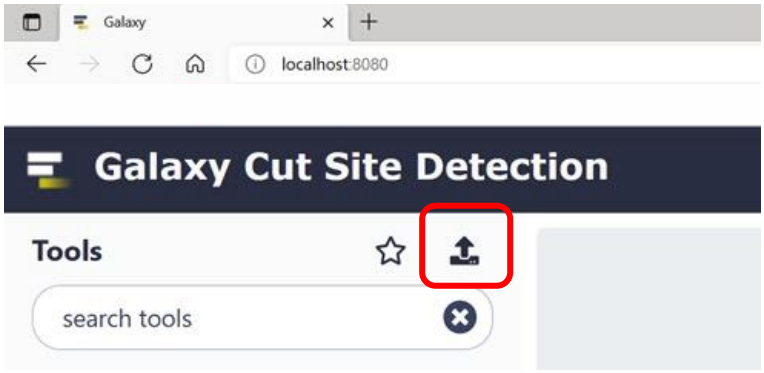
Log in from Login/Register. The default Public name /Password is as follows.

Public name: admin

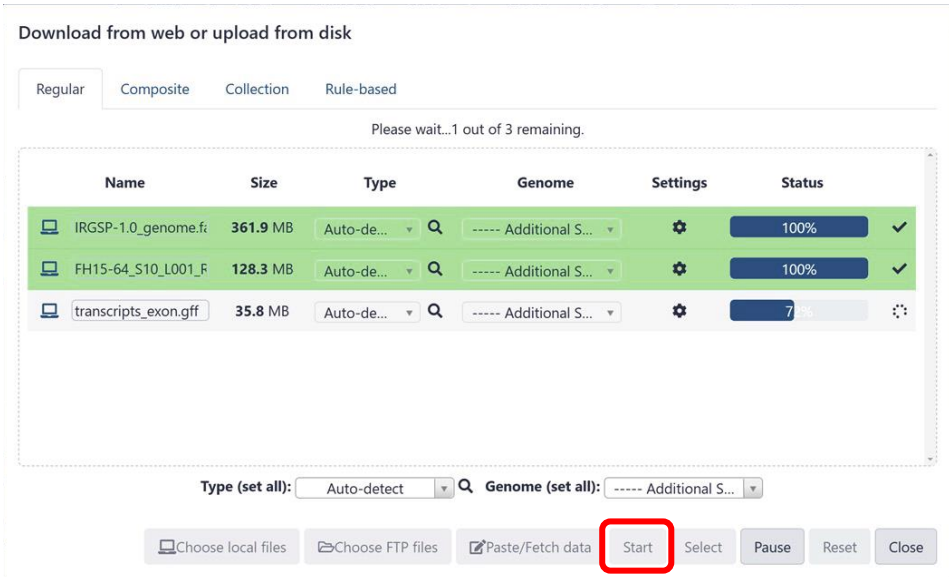
Password: password

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.

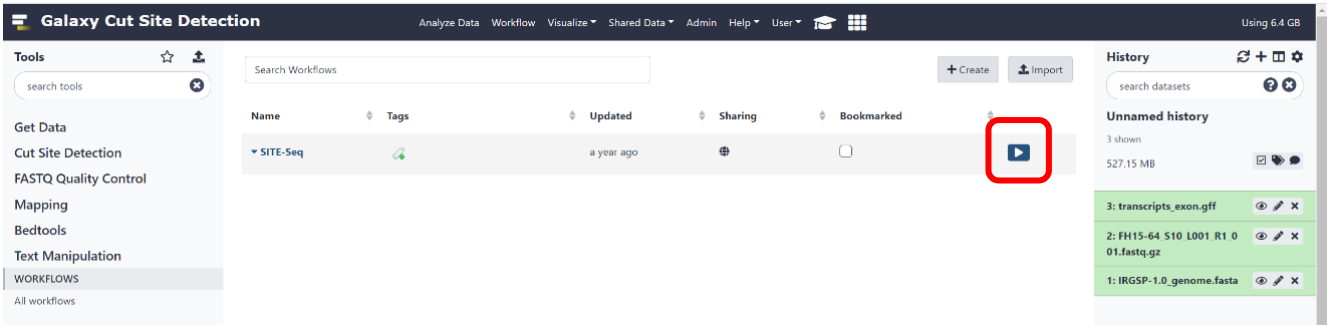


Drop the datasets into the upload window. Click "Start" button to upload the datasets.



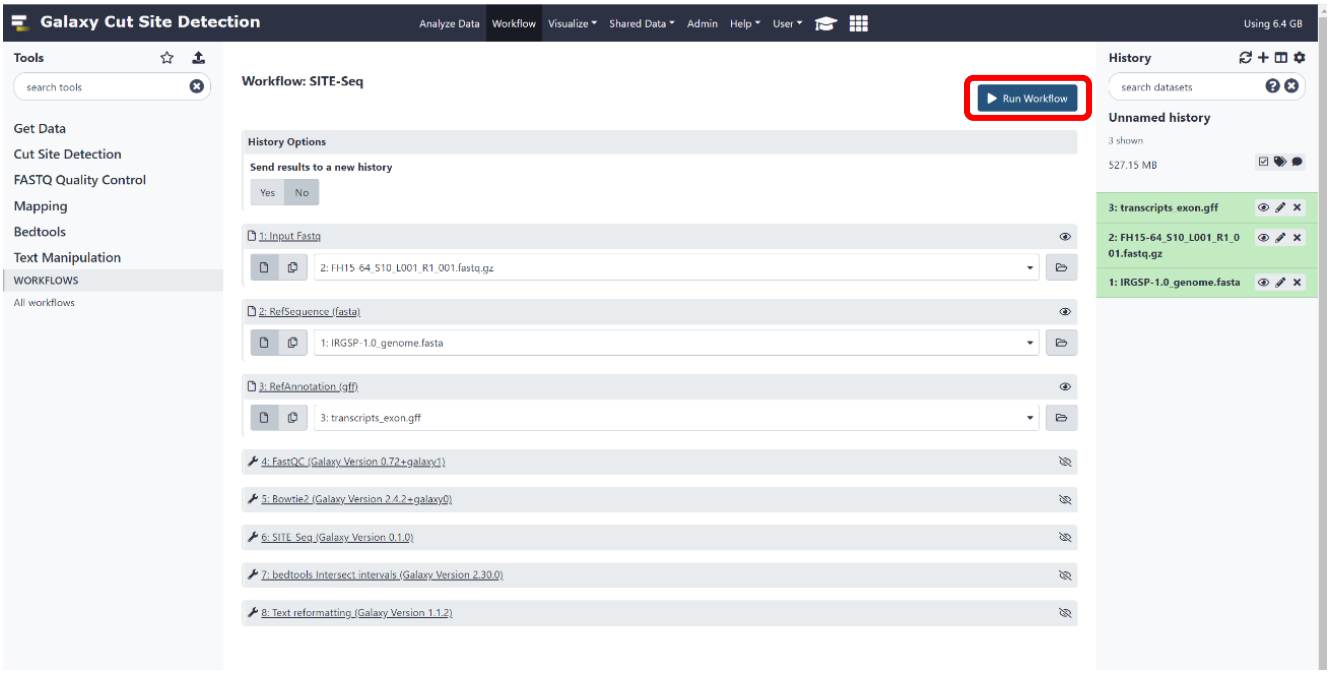
9. Execute analysis

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.



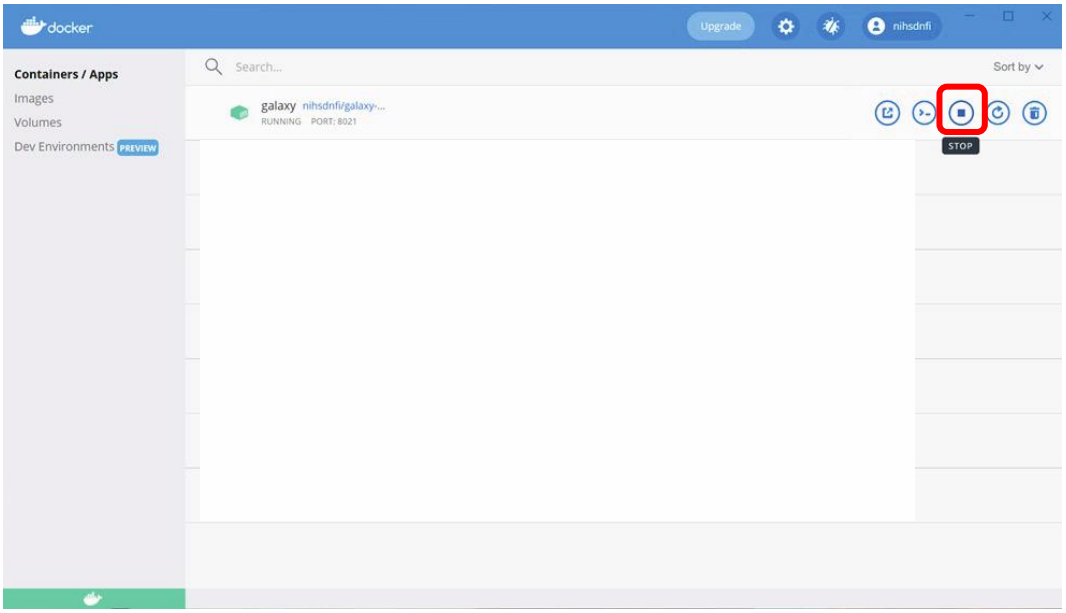
10. About output files

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

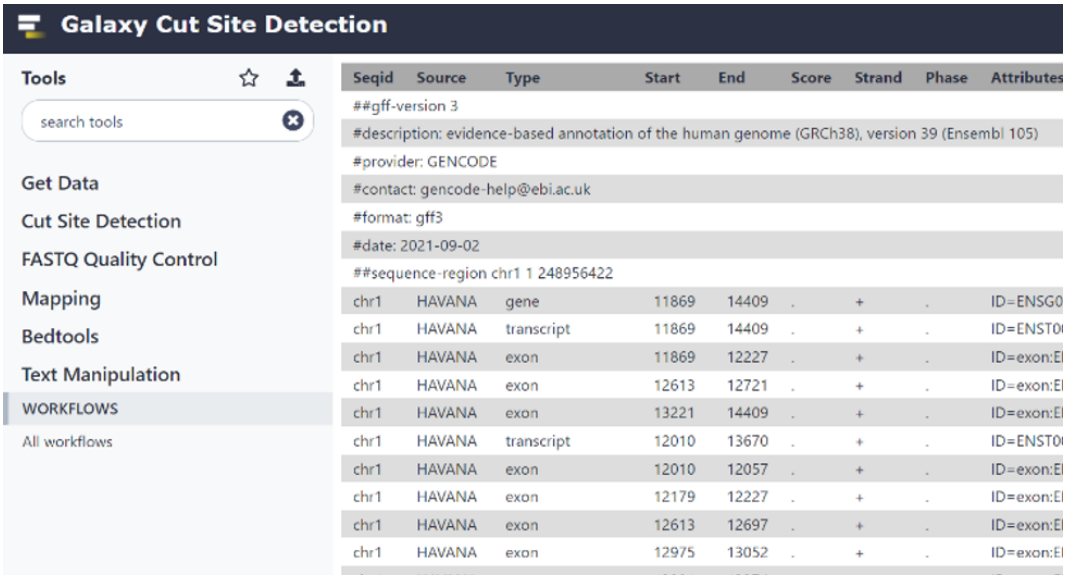
Terminate the Docker container with the "STOP" button of the Galaxy for Cut Site Detection container from Containers/Apps in the Docker Desktop window.

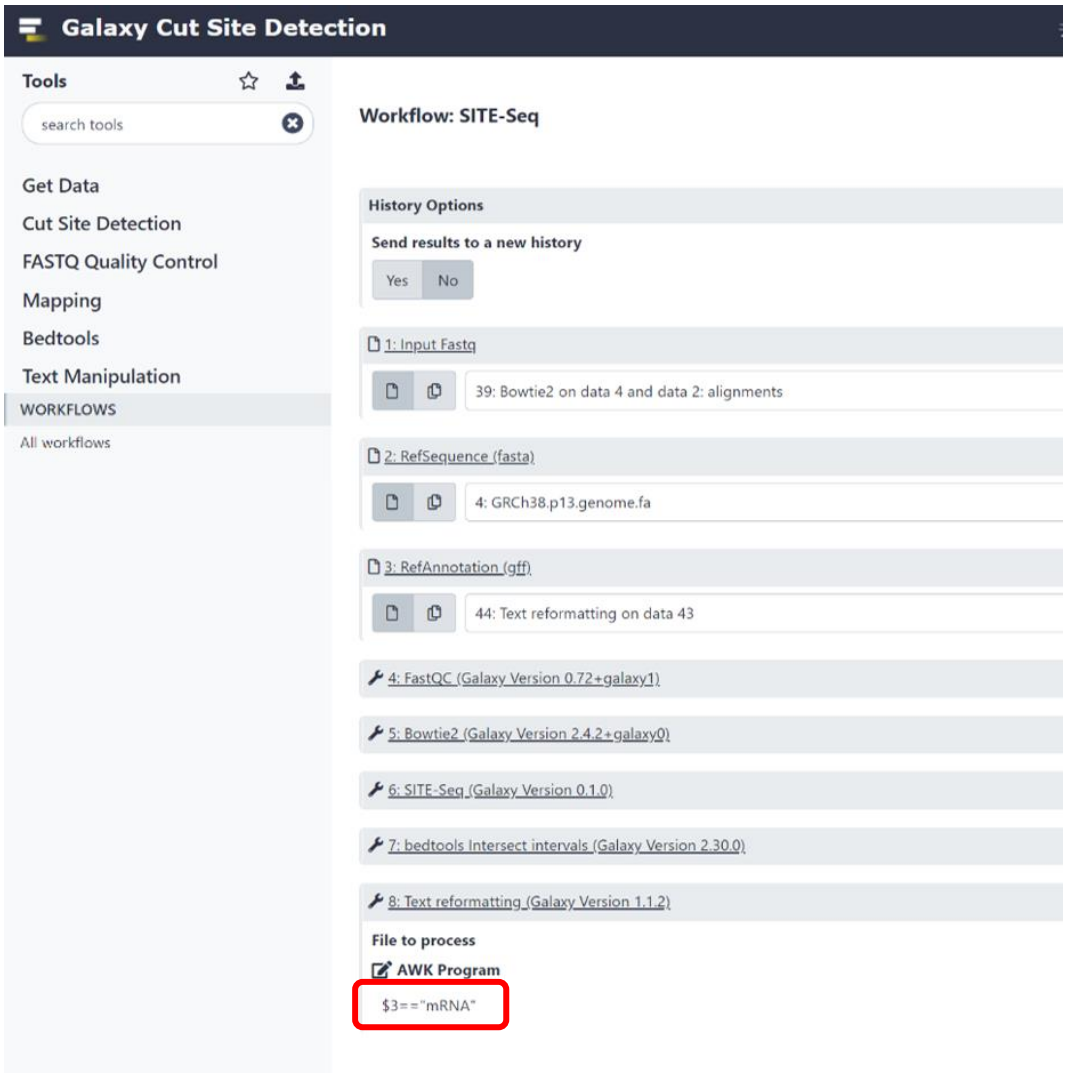


12. Q&A

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).





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.