Tutorial on how to use the single cell RNA seq Snap pipeline

2025-04-17

Contents

1 Introduction

1.1 Overview

This notebook offers guidelines and best practices for effectively using the single cell RNA seq Snap pipeline[1]. Please note that we will continue to maintain and update the pipeline with new analysis modules.

1.2 Prerequisites

We assume users have basic experience using Shell, git, and GitHub.

2 Getting Started

2.1 Accessing the Code

We recommend that users fork the `sc-rna-seq-snap` repository and then clone their forked repository to their local machine. Team members should use the stjude-dnb-binfcore[2] account, while others can use their preferred GitHub account. We welcome collaborations, so please feel free to reach out if you're interested in being added to the `stjude-dnb-binfcore` account.

2.2 Naming your fork

- Team members: Retain the pipeline name and append the project name at the end, e.g., `sc-rna-seq-snap-Vsx2_SE`.
- External collaborators: Keep the pipeline name as-is, unless you plan to use the repository for multiple projects. In that case, follow the same naming convention as team members.

2.3 Managing the privacy of your fork

Please be aware that forked repositories are public by default and will contain the code from the main branch of the `sc-rna-seq-snap` repository, which has been reviewed and tested.

No results will be published unless the project is under review or has been officially published.

Furthermore, we adhere to strict guidelines to ensure privacy and protect sensitive data. Sensitive file paths and patient-related data must never be published on GitHub.

2.4 Running the Code

1. Configure Your Parameters

Replace the project_parameters.Config.yaml file with your own file paths and parameters.

2. Navigate to an Analysis Module

Change to the relevant directory and run the desired shell script:

```
cd ./sc-rna-seq-snap/analyses/<module_of_interest>
```

3. Sync Your Fork

User needs to ensure that the main branch of the forked repository is always up to date with `stjude-dnb-binfcore/sc-rna-seq-snap:main`.

[1] https://github.com/stjude-dnb-binfcore/sc-rna-seq-snap
[2] https://github.com/stjude-dnb-binfcore

If your fork is behind the main repository (`stjude-dnb-binfcore/sc-rna-seq-snap:main`), sync it to ensure you have the latest updates. This will update the main branch of your project repo with the new code and modules (if any). This will add code and not break any analyses already run in your project repo.

When syncing your forked repository with the main repository, please be cautious of any changes made to the following files, as they are typically modified and specified for project data analysis:

- `project_parameters.Config.yaml`

Before pulling the latest changes, stash any modifications you have made to these files. This ensures that you won't accidentally overwrite your changes when syncing with the main repository.

Some useful git commands:

```
git branch
git checkout main
git config pull.rebase false

git status
git add project_parameters.Config.yaml
git commit -m "Update yaml"
```

Finally, `git pull` to get the most updated changes and code in your project repo. Please be mindful of any local changes in files in your project repo that you have done, e.g., `project_parameters.Config.yaml`. You will need to commit or stash (or restore) the changes to the yaml before completing the pull.

```
git pull
```

3 Data

3.1 Set up parameters for the project

Replace parameters in the `project_parameters.Config.yaml` with the correct file paths and parameters specific to your project. Ensure that you use the appropriate file paths and parameters for your data, such as assay type and condition. The default values are set for scRNA-seq data analysis without ambient RNA removal and doublet detection.

3.2 Preparing project metadata

The pipeline requires a TSV file containing essential metadata for cohort analysis. The file must be named `project_metadata.tsv`. It can include one or more samples, as long as it contains at least the following columns in this exact order: `ID`, `SAMPLE`, and `FASTQ`. Additional metadata columns can be added and arranged as needed by the user (though not required).

The file can be stored anywhere, but its filepath must be specified in the `project_parameters.Config.yaml` file.

For user convenience, an example project_metadata.tsv[3] file is provided.

3.3 Genome references

Our team at the Bioinformatics core at DNB maintains the following genome references: 1) human:`GRCh38`; (2) mouse: `GRCm39`, `mm10`, and `mm9`; and (3) dual index genomes: `GRCh38ANDGRCm39`,

[3] https://github.com/stjude-dnb-binfcore/sc-rna-seq-snap/blob/main/data/project_metadata

`GRCh38_mm10`, and `GRCh38_GFP_tdTomato`. Please submit an issue[4] to request the path to the reference genome of preference.

3.4 Cell type gene marker lists

Our team at the Bioinformatics core at DNB maintains the following cell type gene marker lists for cell type annotation: `mouse brain tissue`, `human adult retina tissue`, `human fetal retina tissue` and `mouse retina tissue`. Please submit an issue[5] to request the list of gene markers of preference.

4 Analysis

4.1 Analysis module

Please refer to the analysis_module/README.md[6] files for instructions on how to run the specific analysis module. These files contain the required parameters and necessary files to successfully execute the module's pipeline.

4.2 Container Overview

We have generated a Docker image that contains all tools, packages, and dependencies necessary to run the code and analyses modules. The environment is specifically configured for `Rstudio/R v4.4.0` and `Seurat v4.4.0`. For more details, please refer to the README.md[7].

4.3 CPU and Memory Resources

While we provide estimates for the computational resources required (based on 8 samples with approximately 50,000 cells), users may need to adjust memory settings based on cohort size and analysis requirements.

Important Considerations:

- Adjust memory requests according to the size of your cohort and specific analysis needs.
- For St. Jude users:
  - Refer to the Introduction to the HPCF cluster[8] for detailed guidance.
  - If you require more than 1 TB of memory, use the `large_mem` queue to ensure proper resource allocation.

5 Contact

Contributions, issues, and feature requests are welcome! Please feel free to check issues[9].

5.1 Authors

Antonia Chroni, PhD (@AntoniaChroni[10])

[4] https://github.com/stjude-dnb-binfcore/sc-rna-seq-snap/issues
[5] https://github.com/stjude-dnb-binfcore/sc-rna-seq-snap/issues
[6] https://github.com/stjude-dnb-binfcore/sc-rna-seq-snap/tree/main/analyses
[7] https://github.com/stjude-dnb-binfcore/sc-rna-seq-snap/blob/main/run-container/README.md#load-specific-version-of-singularity
[8] https://wiki.stjude.org/display/HPCF/Introduction+to+the+HPCF+cluster#IntroductiontotheHPCFcluster-queuesQueues:
[9] https://github.com/stjude-dnb-binfcore/trainings/issues
[10] https://github.com/AntoniaChroni

[11] https://www.stjude.org/
[12] https://opensource.org/license/bsd-2-clause