scRNAseq Processing Workflows

Description

This document describes how to use some scanpy-based scRNAseq workflows on galaxy Australia.

The aim of these workflows is to handle the routine 'boring' part of single cell RNAseq data processing. It will produces an 'AnnData' object, which can then be used as a base for downstream analysis – either within galaxy or outside of it. AnnData is a standard format used by the 'scanpy' python package.

These workflows represent just one way of processing data for a 'typical' scRNAseq experiment – there are many other options!

This document describes 3 sub-workflows for processing single cell RNAseq data with scanpy

- **Load counts matrix:** link This workflow adds a sample name, which enables multi-sample analyses
- **Single cell QC:** link This workflow generates some basic QC plots and applies filtering
- **Single cell QC to basic processing:** link This generates a UMAP, does clustering and calculates cluster marker genes.

For single sample experiments, there is a streamlined workflow that runs all 3 sub-workflows all at once

- **Single sample workflow:** link This workflow loads counts matrix, does some basic processing, suitable for a single sample.

These workflows are all available on galaxy australia.

Analysis overview

1. Start with fastq files
2. Cellranger|starSOLO will align the reads to the genome, and make a table of the number of times each gene is counted per cell.
3. Perform some basic QC on the counts matrix, and filter out 'cells' that have too little RNA counts or too much mitochondrial gene content
4. Run some basic single cell analyses: Normalisation, PCA, UMAP, clustering and identify cluster markesr

5. The resulting AnnData object can be analysed further. See 'Next steps'

Example output

When run in full, these workflows produce the following main outputs

- A processed AnnData file, which contains gene expression and annotation information ready for downstream analysis.
- Tables of 'marker' gene information - to aid detemination of cell types present in the experiment
- Two report summaries.
    - Single cell QC report (example) : This shows QC metrics at the cell level, to evaluate filtering thresholds and data quality.
    - Single cell basic processing report (example) : Some basic UMAP, clustering and cluster marker results to begin an analysis with.

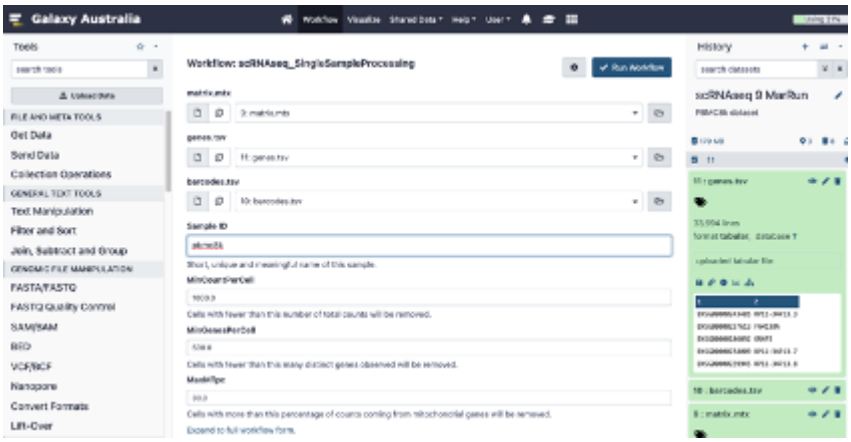
User guide

Running a single sample workflow

When there is only a single biological sample in a study, there is a streamlined workflow.

- Single sample workflow

1. Input *will be* fastq files.
2. Search for the *scRNAseq Single sample workflow* under 'workflow' and run it. You'll be prompted to customise any filtering parameters, and choose a sensible name for the biological sample.



3. This pipeline should take a [few minutes/few hours] to run.

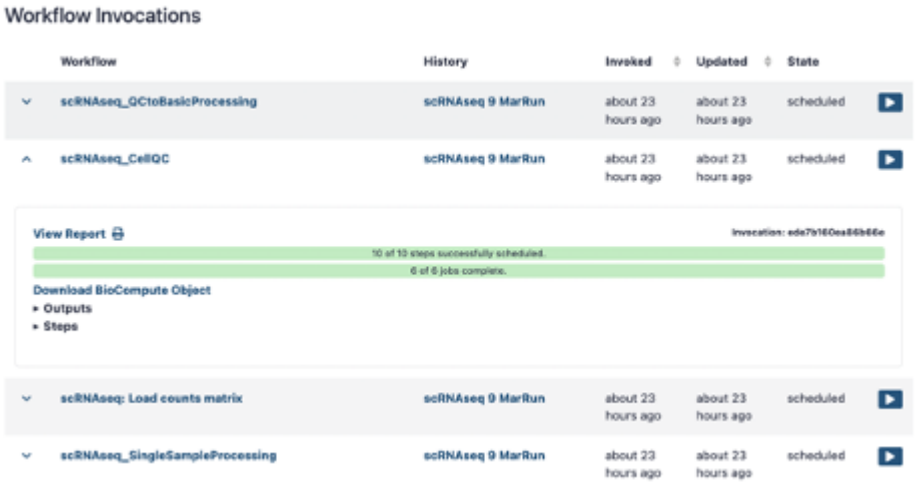
4. Return to galaxy, and look up the run, or invocation, of this workflow:

User Menu > Workflow Invocations

This brings up the history of workflow invocations. This particular workflow runs a number of sub-workflows. The **Cell QC** and **QC to Basic Processing** subworkflows produce reports, which can be viewed

shared and saved.

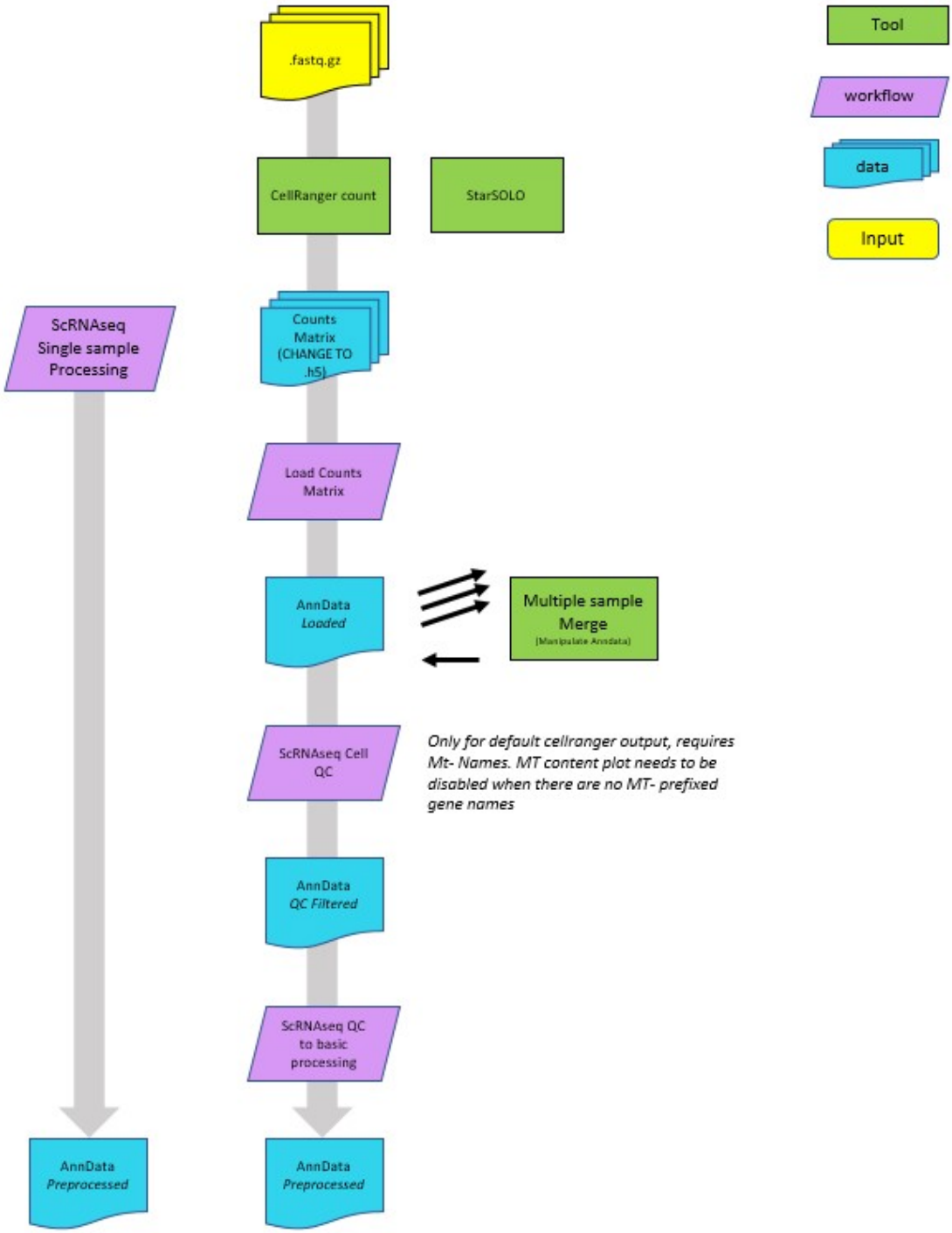
- Cell QC : Generates the cell level QC plots
- QC to Basic Processing: Shows the subsetquent UMAP, clustering and marker information. Includes links to download the processed AnnData object for downstream work



Running a multi-sample workflow

With multi-sample experiments, each sample is loaded independently and then combined. The overall method is the same, but the QC and processing steps are run separately.

WORKFLOWS FOR SCRNASEQ

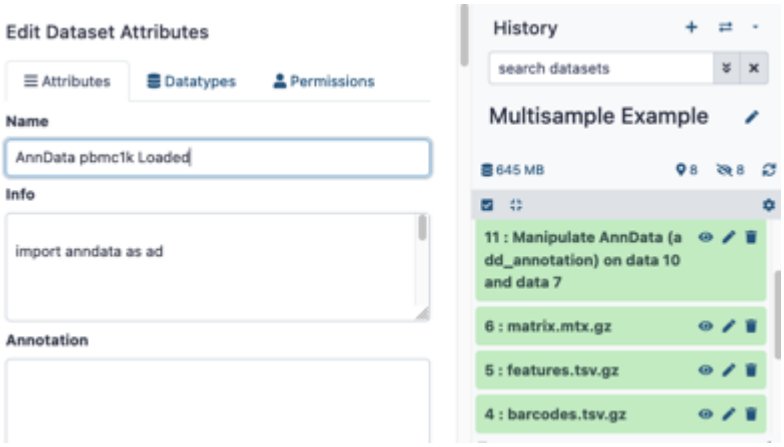


1. Upload the raw data for one sample. Then run The ‘scRNAseq: Load counts matrix’ workflow – this will prompt you for a sample name that will be used throughout.


```
**IF USING the 3 file input, describe how to change barcodes.tsv to
tabular, and note that features == genes, and that .tsv.gz will work fine.
**
```

This part of the workflow will load the counts matrix into an anndata object, and then adds an extra column in the metadata called ‘sample’. This means the sample information can be tracked when multiple samples are combined.

The AnnData object that it produces in your history will probably be named something like ‘Manipulate AnnData (add_annotation) on data 20 and data 17’. You may choose to rename this object via the ‘edit attributes’ option in the history panel, so its easier to find later.



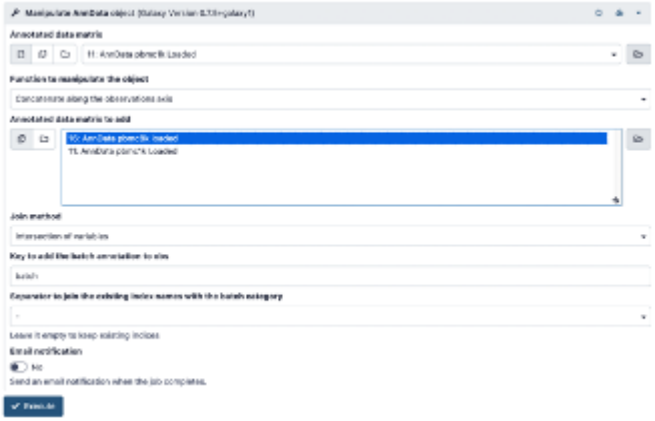
2. In the same history, repeat for all other samples.

3. Next, join all samples with the ‘Manipulate AnnData object’ Tool (search on the tools pane on the left).

This tool can do several different operations – listed under ‘Function to manipulate the object’, but we want the default; “Concatenate along the observation axis”. This combines cells (observations) from multiple sample runs.

Choose the AnnData object of one of your samples in the ‘Annotated data matrix’ dropdown. Then, choose the rest of your samples under ‘Annotated matrix to add’. Use ctrl-select / option-select to highlight multiple samples.

Note: Be careful not to select the sample in the 'Annotated data matrix' dropdown again – else it will be joined to itself! In this example there is only two samples, so only pbmc8k is selected to be added to pbmc1k.



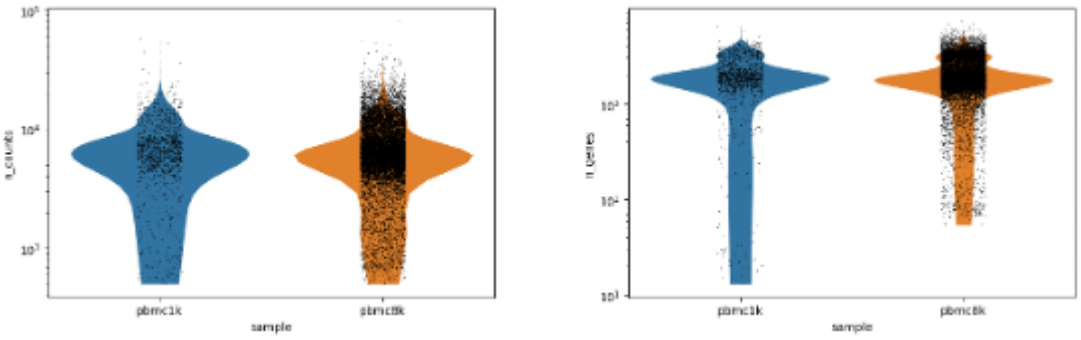
A combined anndata object will created in your history.

4. Next, run the **scRNAseq Cell QC** workflow on your combined anndata object.

This workflow plots some basic cell-level QC thresholds, and applies the QC thresholds to produce a filtered Anndata object.

5. Once it finishes running, view the report (Go to User menu > Invocations to find it).

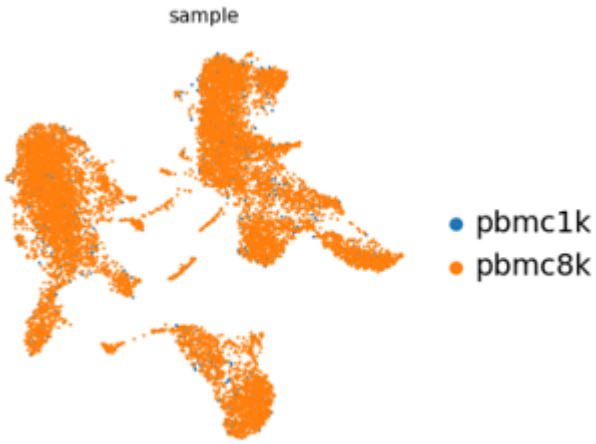
You'll notice you can see each sample plotted separately in the QC plots. You may elect to rerun with tweaked thresholds (e.g. higher minimum counts threshold) once you've seen this output.



6. If you are happy with the filtering thresholds, you can launch the next workflow, **scRNAseq QC to Basic Processing** to do some routine single cell calculations. It only asks for the filtered Anndata Object (typically the last AnnData in your history, with a name ending in 'Filtered Cells Anndata'.)

7. This takes a few minutes to run. Once finished, return to the invocations page to see the QC to Basic Processing report, as per the single sample workflow.

The first umap now shows the different samples that make up the data.



Next steps

The AnnData object generated is ready for analysis! Options include

- CellXgene : CellXgene is a tool for browsing and exploring single cell data. It use the AnnData object, and all of the annotation stored within it (expression, clusters, sample names). See CellXgene documentation, and a user guide. This can be launched within galaxy.
- Scanpy : Scanpy is a python based suite of tool for working with single cell data.
    - Many scanpy functions are available within galaxy (and are used within this workflow). Explore the galaxy scRNAseq tutorials for more information
    - Or, you download your AnnData object and work with the scanpy toolkit directly, using python on your computer or other server. This can provide more fine-grained control of your analysis. There are many tutorials to work from.

Note that there are toolkits other than scanpy (e.g. Seurat, SingleCellExperiment objects) which may not be directly compatible without conversions.

Background and Tutorials

For more general information about **single cell RNAseq processing on galaxy**; there are some excellent tutorials to be found here on the galaxy training website scRNA section. The workflow implemented here is heavily influenced by the Clustering 3kPBMCs with Scanpy tutorial

More general information on **using galaxy** can be found on the galaxy training website

There are many general resources online about the princials of single cell analysis. The Scanpy preprocessing and clustering tutorial may be of particular use becuase it describes the scanpy methods used in this workflow. Even if you don't use the python code, it works through and explains many of the plots these workflows generate.

License(s)

Acknowledgements/citations/credits

The workflow implemented here is heavily influenced by the Clustering 3kPBMCs with Scanpy tutorial