UCSF center for systems & synthetic biology
an NIGMS national systems biology center
www.systemsbiology.ucsf.edu

RNA-seq pipeline user guide

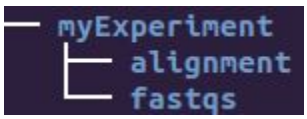
(Example with the drop seq pipeline) ~ Paul Rivaud UCSF 2015

Folder organization and format

In order to run your data through the pipeline, you need to log in onto the lab server. If you do not know the login and password, please contact:

- Sisi sisi.chen1@gmail.com
- Graham gheimberg@gmail.com

You need to create a tree as shown below in the `~/RNAseq_files/` folder:



The alignment folder will receive the SAM file outputted by Bowtie2. The fastqs folder must contain your R1 and R2 **compressed** files (.gz format). If you do not know how to copy the files from your computer on the server, you can open a terminal on your computer and use the following commands:

```
> cd path_to_folder/
> scp ./my_R1.fastq.gz ./my_R2.fastq.gz
server:~/RNAseq_files/myExperiment/fastqs/
```

Launching the pipeline run via a command line

You must now go to the following folder:
`~/TITAN_pipeline/pipeline_code/drop_tac/`

There are two Python files in this folder. You should not modify `drop_titan_html.py`, unless you know what you are doing. Nevertheless, it is highly recommended to copy the code folder into `~/TITAN_pipeline/pipeline_code/` if you want to make serious modifications.

You should modify the `drop_titan_params.py` file. It enables you to modify options regarding the pipeline run you are about to launch:

Primary options

- `dir_path_fastqs` (string): The path to the folder containing your compressed fastq files.

- `dir_path_alignment` (string): The path to the alignment folder that will contain the SAM file.

- `reference_genome` (string): The path to the reference genome / transcriptome files you want to use (no extension needed in the path because of the multiple files).

- `fasta_path` (string): The path to the fasta file used to create the transcriptome. This enables the pipeline to always create a similar gene list for experiments with the same species.

Detailled options (options that need insight on the expected data sctructure)

- `str_search` (string): String to be found in the first three characters of R1 (binders). For example 'TAC'. If nothing is expected, just put two simple quotes `''`.

- `tac_length` (integer): Length of the binder. If no binder was used in the experiment, put 0.

- `umi_length` (integer): Length of the Unique Molecular Identifiers.

- `barcode_length` (integer): Length of the cell barcodes.

- `tso` (string): String that dismisses read if found in read sequence.

- `occ_threshold` (integer): Minimum number of reads for a cell to be in the computed gene expression matrix

- `processors` (integer): Number of processors used by Bowtie2 during the alignment

- `bowtie2_options` (list): List containing various options for Bowtie2

Once you have set the options to your liking, you can launch the run via one of the following command lines:

- Pipeline run with duration displayed at the end of the run.
  ```
  > time python drop_titan_html.py
  ```

- Pipeline run with interactive option (gives you access to the variables after the run via a Python prompt)
  ```
  > python -i drop_titan_html.py
  ```

Results

The results files are in your ~/RNAseq_files/myExperiment/ folder. You can download them onto your computer via the following command lines, after opening a new terminal on your computer:

```
> cd path_to_save_folder/
> scp server:~/RNAseq_files/myExperiment/myExperiment_matrix.txt ./
> scp server:~/RNAseq_files/myExperiment/myExperiment_report.html ./
```

If you break the pipeline, do not worry too much, shit happens, but please tell someone in charge !