Totalseq analysis Pipeline – User Guide

Contents

[Summary 1](#_Toc106627403)

[Hardware requirements 1](#_Toc106627404)

[Software requirements 1](#_Toc106627405)

[Supported operating systems 1](#_Toc106627406)

[Running the Pipeline 2](#_Toc106627407)

[Running the Pipeline 2](#_Toc106627408)

[Pipeline output 3](#_Toc106627409)

[Downstream Analyses 3](#_Toc106627410)

# Summary

This document describes requirements and instructions for installing and running Totalseq analysis Pipeline on a Linux server.

# Hardware requirements

These minimum hardware requirements are based on test Pipeline runs with the 12 GB FASTQ files with 7 antibody tags:

* **CPU:** 16 CPU cores​
* **Memory:** 16 GB​ (32 GB if RNA-seq data is being analyzed)
* **Disk space:** 500 GB​

Note that hardware requirements are data-specific, with larger files requiring more resources.

# Software requirements

## Supported operating systems

This script is likely to support any operating system that supports the following software packages:

* Kallisto, and
* Python 3.7 or above.

Currently, only Linux (CentOS 7) has been tested.

# Running the Pipeline

This pipeline allows the analyses of following types of features: protein or protein + RNA.

## Running the Pipeline

The BEN-seq analysis pipeline performs quantification through benseq\_cli\_frontend.py. The main usage of the script is:

python benseq\_cli\_frontend.py -r1 [RNA\_R1\_FASTQs separated by space] -r2 [RNA\_R2\_FASTQs separated by space] -t [Kallisto transcriptome .idx] -a1 [ADT\_R1\_FASTQs separated by space] -a2 [ADT\_R2\_FASTQs separated by space] -f [ADT feature CSV file] -o output\_folder

| Argument | Definition |
| --- | --- |
| -a1, --adt\_r1\_fastq | Space-separated paths of first-read FASTQ file from the ADT libraries.  |
| -a2, --adt\_r2\_fastq | Space-separated paths of second-read FASTQ file from the ADT libraries. |
| -f, --feature-reference | A comma-separated feature reference file in the same format as the input for --feature-ref in Cell Ranger count. This can be generated on the BioLegend website using the Barcode Look-Up tool (<https://www.biolegend.com/totalseq> ) and export the selected markers using the “Export only selected items in Cell Ranger Format”. |
| -r1, --rna\_r1\_fastq | Space-separated paths of first-read FASTQ file from the RNA libraries.  |
| -r2, --rna\_r2\_fastq | Space-separated paths of second-read FASTQ file from the RNA libraries. |
| -t, --transcriptome\_index | A Kallisto .idx index for the species used in RNA. |
| -I, --output\_directory | The directory where the results should be written to. |

The following is an example to demonstrate its arguments:

python benseq\_cli\_frontend.py -r1 Sample\_RNA\_L001\_R1.fastq.gz Sample\_RNA\_L002\_R1.fastq.gz -r2 Sample\_RNA\_L001\_R2.fastq.gz Sample\_RNA\_L002\_R2.fastq.gz -a1 Sample\_ADT\_L001\_R1.fastq.gz Sample\_ADT\_L002\_R1.fastq.gz -a2 Sample\_ADT\_L001\_R2.fastq.gz Sample\_ADT\_L002\_R2.fastq.gz -t transcriptome.idx -f totalseqCellRangerExperiment.csv -o output\_folder

The example above assumes a sample called “Sample” with both RNA and ADT libraries, and sequenced in two Illumina lanes. The RNA data is to be quantified with the index file transcriptome.idx, the ADT data is to be quantified using the features in totalseqCellRangerExperiment.csv, and the results of both will be written to output\_folder.

The RNA steps will only be run when -r1, -r2 and -t are populated; similarly, the ADT steps will only be run when -a1, -a2 and -f are populated.

### Pipeline output

Depending on whether the ADT and the RNA quantification pipelines are ran, the pipeline creates the subdirectories ADT and/or RNA in the output folder. Both directories have the same structure, differing only on the nature of the quantified features:

|  |  |
| --- | --- |
| File name  | Description  |
| abundance.h5 | A HDF5 binary file containing run info, abundance estimates, bootstrap estimates, and transcript length information length. |
| abundance.tsv | A tab-separated text representation of abundance.h5. |
| run\_info.json  | A JSON file containing run statistics. |

For more details regarding these files, or Kallisto in general, please visit the authors’ website at <https://pachterlab.github.io/kallisto> .

This script also generates its log as output\_folder/ben\_seq.log.

### Downstream Analyses

In our experience the bulk ADT data requires different dispersion from the RNA data, even from the same sample. Thus, while we have successfully applied standard RNA-seq differential expression methods on bulk ADT data, it is recommended that the ADT data to be analyzed separately from the RNA data.