htseq-count-cluster Documentation

Release 1.3

Shaurita Hutchins, Robert Gilmore

Feb 25, 2022

Contents:

A cli wrapper for running htseq's `htseq-count` on a cluster.

View a project overview at our Datasnakes site.

CHAPTER 1

Install

```
pip install HTSeqCountCluster
```

CHAPTER 2

Features

- For use with large datasets (we've previously used a dataset of 120 different human samples)
- For use with SGE/SGI cluster systems
- Submits multiple jobs
- Command line interface/script
- Merges counts files into one counts table/csv file
- Uses `accepted_hits.bam` file output of `tophat`

2.1 Examples

2.1.1 Run htseq-count-cluster

After generating bam output files from tophat, instead of using HTSeq's `htseq-count`, you can use our `htseq-count-cluster` script. This script is intended for use with clusters that are using pbs (qsub) for job monitoring.

Our default `htseq-count` command is `htseq-count -f bam -s no file.bam file.gtf -o htseq.out`. This command does not take into account any strandedness (`-s no`) for the input bamfiles (`-f bam`) and uses the default `union` mode. For the default mode `union`, only the aligned read determines how the read pair is counted.

```
htseq-count-cluster -p path/to/bam-files/ -f samples.csv -g genes.gtf -o path/to/
↪cluster-output/
```

Argument	Description	Required
-p	This is the path of your .bam files. Presently, this script looks for a folder that is the sample name and searches for an accepted_hits.bam file (tophat output).	Yes
-i	You should have a csv file list of your samples or folder names (no header).	Yes
-g	This should be the path to your genes.gtf file.	Yes
-o	This should be an existing directory for your output counts files.	Yes
-e		

This script uses logzero so there will be color coded logging information to your shell.

A common linux practice is to use `screen` to create a new shell and run a program so that if it does produce output to the stdout/shell, the user can exit that particular shell without the program ending and utilize another shell.

Help message output for `htseq-count-cluster`

```
usage: htseq-count-cluster [-h] -p INPATH -f INFILE -g GTF -o OUTPATH
                              [-e EMAIL]

This is a command line wrapper around htseq-count.

optional arguments:
  -h, --help            show this help message and exit
  -p INPATH, --inpath INPATH
                        Path of your samples/sample folders.
  -f INFILE, --infile INFILE
                        Name or path to your input csv file.
  -g GTF, --gtf GTF     Name or path to your gtf/gff file.
  -o OUTPATH, --outpath OUTPATH
                        Directory of your output counts file. The counts file
                        will be named.
  -e EMAIL, --email EMAIL
                        Email address to send script completion to.

*Ensure that htseq-count is in your path.
```

2.1.2 Merge output counts files

In order to prep your data for `DESeq2`, `limma` or `edgeR`, it's best to have 1 merged counts file instead of multiple files produced from the `htseq-count-cluster` script. We offer this as a standalone script as it may be useful to keep those files separate.

```
merge-counts -d path/to/cluster-output/
```

Help message for `merge-counts`

```
usage: merge-counts [-h] -d DIRECTORY

Merge multiple counts tables into 1 counts .csv file.

Your output file will be named:  merged_counts_table.csv
```

(continues on next page)

(continued from previous page)

```

optional arguments:
  -h, --help            show this help message and exit
  -d DIRECTORY, --directory DIRECTORY
                        Path to folder of counts files.
```

CHAPTER 3

ToDo

- [ ] Monitor jobs.
- [ ] Enhance wrapper input for other use cases.
- [ ] Add example data.

CHAPTER 4

Maintainers

Shaurita Hutchins | @sdhutchins |
Rob Gilmore | @grabear |

CHAPTER 5

Help

Please feel free to open an issue if you have a question/feedback/problem or submit a pull request to add a feature/refactor/etc. to this project.

CHAPTER 6

Citation

*Simon Anders, Paul Theodor Pyl, Wolfgang Huber;***HTSeq—a Python framework to work with high-throughput sequencing data***, Bioinformatics, Volume 31, Issue 2, 15 January 2015, Pages 166–169,*https://doi.org/10.1093/bioinformatics/btu638

CHAPTER 7

Indices and tables

- genindex
- modindex
- search