demuxEM Documentation

Release 0.1.7

Bo Li, Joshua Gould, Yiming Yang

Mar 11, 2022

CONTENTS

1 Version 0.1.7 *October 05, 2021* 3

2 Version 0.1.6 *May 28, 2021* 5

3 Version 0.1.5 *September 16, 2020* 7

4 Version 0.1.4 *July 15, 2020* 9

5 Version 0.1.1 *April 9, 2020* 11

6 Version 0.1.0 *April 8, 2020* 13

Index 19

demuxEM is a module on analyzing cell-hashing/nucleus-hashing data. It's used by Cumulus in demultiplexing step.

CHAPTER

ONE

VERSION 0.1.7 *OCTOBER 05, 2021*

Filter cell barcodes with 0 hashtag counts in estimate_background_probs

CHAPTER

TWO

VERSION 0.1.6 *MAY 28, 2021*

Fix the bug when the input multi-modal object contains non-RNA modality data.

CHAPTER

THREE

VERSION 0.1.5 *SEPTEMBER 16, 2020*

Add barplot showing percentage of RNA barcodes with HTO counts.

CHAPTER

FOUR

VERSION 0.1.4 *JULY 15, 2020*

Adapt to *pegasusio* 0.2.5.

CHAPTER

FIVE

VERSION 0.1.1 *APRIL 9, 2020*

Adapt to *pegasusio* 0.1.3.

CHAPTER

SIX

VERSION 0.1.0 *APRIL 8, 2020*

Initial release under new name **demuxEM**.

6.1 Installation

demuxEM is published on PyPI as a Python package, and you can simply install it via `pip`:

```
pip install demuxEM
```

6.2 Use `demuxEM` as a command line tool

If you have data generated by cell-hashing or nucleus-hashing, you can use `demuxEM` as a command line tool to demultiplex your data. Type:

```
demuxEM -h
```

to see the usage information:

```
Usage:
        demuxEM [options] <input_raw_gene_bc_matrices_h5> <input_hto_csv_file> <output_
↪name>
        demuxEM -h | --help
        demuxEM -v | --version
```

- Arguments:

  input_raw_gene_bc_matrices_h5 Input raw RNA expression matrix in 10x hdf5 format.

  input_hto_csv_file Input HTO (antibody tag) count matrix in CSV format.

  output_name Output name. All outputs will use it as the prefix.

- Options:

  -p <number>, -\-threads <number> Number of threads. [default: 1]

  -\-genome <genome> Reference genome name. If not provided, we will infer it from the expression matrix file.

  -\-alpha-on-samples <alpha> The Dirichlet prior concentration parameter (alpha) on samples. An alpha value < 1.0 will make the prior sparse. [default: 0.0]

-\-min-num-genes <number> We only demultiplex cells/nuclei with at least <number> of expressed genes. [default: 100]

-\-min-num-umis <number> We only demultiplex cells/nuclei with at least <number> of UMIs. [default: 100]

-\-min-signal-hashtag <count> Any cell/nucleus with less than <count> hashtags from the signal will be marked as unknown. [default: 10.0]

-\-random-state <seed> The random seed used in the KMeans algorithm to separate empty ADT droplets from others. [default: 0]

-\-generate-diagnostic-plots Generate a series of diagnostic plots, including the background/signal between HTO counts, estimated background probabilities, HTO distributions of cells and non-cells etc.

-\-generate-gender-plot <genes> Generate violin plots using gender-specific genes (e.g. Xist). <gene> is a comma-separated list of gene names.

-h, -\-help Print out help information.

- Outputs:

  output_name_demux.zarr RNA expression matrix with demultiplexed sample identities in Zarr format.

  output_name.out.demuxEM.zarr DemuxEM-calculated results in Zarr format, containing two datasets, one for HTO and one for RNA.

  output_name.ambient_hashtag.hist.pdf Optional output. A histogram plot depicting hashtag distributions of empty droplets and non-empty droplets.

  output_name.background_probabilities.bar.pdf Optional output. A bar plot visualizing the estimated hashtag background probability distribution.

  output_name.real_content.hist.pdf Optional output. A histogram plot depicting hashtag distributions of not-real-cells and real-cells as defined by total number of expressed genes in the RNA assay.

  output_name.rna_demux.hist.pdf Optional output. This figure consists of two plots. The first one is a horizontal bar plot depicting the percentage of RNA barcodes with at least one HTO count. The second plot is a histogram plot depicting RNA UMI distribution for singlets, doublets and unknown cells.

  output_name.gene_name.violin.pdf Optional outputs. Violin plots depicting gender-specific gene expression across samples. We can have multiple plots if a gene list is provided in '–generate-gender-plot' option.

- Examples:

```
demuxEM -p 8 --generate-diagnostic-plots sample_raw_gene_bc_matrices.h5 sample_hto.
↪csv sample_output
```

6.3 API

demuxEM can also be used as a python package. Import demuxEM by:

```
import demuxEM
```

6.3.1 Demultiplexing

estimate_background_probs(hashing_data[, ...])	For cell-hashing data, estimate antibody background probability using KMeans algorithm.
demultiplex(rna_data, hashing_data[, ...])	Demultiplexing cell/nucleus-hashing data, using the estimated antibody background probability calculated in `demuxEM.estimate_background_probs`.
attach_demux_results(input_rna_file, rna_data)	Write demultiplexing results into raw gene expression matrix.

demuxEM.estimate_background_probs

demuxEM.**estimate_background_probs**(*hashing_data*, *random_state=0*)

For cell-hashing data, estimate antibody background probability using KMeans algorithm.

Parameters

- **hashing_data** (`UnimodalData`) – Annotated data matrix for antibody.
- **random_state** (`int`, optional, default: `0`) – Random seed set for reproducing results.

Return type `None`

Returns

- `None`
- **Update ``hashing_data``** (*Filtered cell barcodes with 0 counts*)
- Update `hashing_data.uns` –
  - `hashing_data.uns["background_probs"]`: estimated antibody background probability.

Example

```
>>> estimate_background_probs(hashing_data)
```

demuxEM.demultiplex

demuxEM.**demultiplex**(*rna_data*, *hashing_data*, *min_signal=10.0*, *alpha=0.0*, *alpha_noise=1.0*, *tol=1e-06*, *n_threads=1*)

Demultiplexing cell/nucleus-hashing data, using the estimated antibody background probability calculated in `demuxEM.estimate_background_probs`.

Parameters

- **rna_data** (`UnimodalData`) – Data matrix for gene expression matrix.
- **hashing_data** (`UnimodalData`) – Data matrix for HTO count matrix.
- **min_signal** (`float`, optional, default: `10.0`) – Any cell/nucleus with less than `min_signal` hashtags from the signal will be marked as `unknown`.
- **alpha** (`float`, optional, default: `0.0`) – The Dirichlet prior concentration parameter (alpha) on samples. An alpha value < 1.0 will make the prior sparse.
- **alpha_noise** (`float`, optional, default: `1.0`) – The Dirichlet prior concenration parameter on the background noise.
- **tol** (`float`, optional, default: `1e-6`) – Threshold used for the EM convergence.
- **n_threads** (`int`, optional, default: `1`) – Number of threads to use. Must be a positive integer.

Returns

- `None`
- Update `data.obs` –
  - `data.obs["demux_type"]`: Demultiplexed types of the cells. Either `singlet`, `doublet`, or `unknown`.
  - `data.obs["assignment"]`: Assigned samples of origin for each cell barcode.
  - `data.obs["assignment.dedup"]`: Only exist if one sample name can correspond to multiple feature barcodes. In this case, each feature barcode is assigned a unique sample name.

Examples

```
>>> demultiplex(rna_data, hashing_data)
```

demuxEM.attach_demux_results

demuxEM.**attach_demux_results**(*input_rna_file*, *rna_data*)

Write demultiplexing results into raw gene expression matrix.

Parameters

- **input_rna_file** (`str`) – Input file for the raw gene count matrix.
- **rna_data** (`UnimodalData`) – Processed gene count matrix containing demultiplexing results

Return type `MultimodalData`

Returns

- `MultimodalData`

- *A multimodal data object.*

Examples

```
>>> data = attach_demux_results('raw_data.h5', rna_data)
```

6.4 Contact us

demuxEM is maintained by Cumulus team. If you have any questions, please feel free to contact us via Cumulus Support Google Group.

INDEX

A

attach_demux_results() (*in module demuxEM*), 16

D

demultiplex() (*in module demuxEM*), 16

E

estimate_background_probs() (*in module demuxEM*), 15