

Generating Matrix

1. Software Environment Dependencies

1.1 Manual installation

A) Python: Version 3.8 or above, with the following modules installed: cv2 (version 4.0 or later), matplotlib, seaborn, pandas, tifffile, Cellpose, Pyyaml, scikit-image.

B) STAR: Version 2.6.1d or later.

C) Perl: threads and threads::shared modules.

D) Requires the following R packages: seurat, dplyr, tibble, ggplot2, broom, purrr, cowplot, cluster, ggpubr, plotly, htmlwidgets, kableextra, htmltools, shiny, knitr, rmarkdown, optparse, getopt.

E) E) Pandoc.

Note: The dependent software needs to be added to the environment variable using **export** before running BSTMatrix.

1.2 Conda

You could follow the file *BSTMatrix_environment.rst* to set up the environment, or simply use the *environment.yaml* file after installing conda. To use the *.yaml* file, please run the command below.

```
conda env create -f environment.yaml
```

2. Input Data

A) Sequencing data: Paired-end sequencing *.fastq* file.

B) Reference genome: Genome sequence *.fasta* file, and the associated *.gtf* file (with exon in the 3rd column) or *.gff* file (with gene and exon in the 3rd column).

C) features.tsv file: generated by using the *.gtf* file with the command below.

```
perl ./tools/features_generate.pl -i xxx.gtf -o features.tsv.
```

D) STAR genome index files: generated by using the genome sequence file and the *.gtf* file as bellowed.

```
STAR --runThreadN 8 --runMode genomeGenerate --genomeDir star/ --genomeFastaFiles genome.fa --sjdbGTFfile gene.gtf.
```

E) Fluorescence decoding files and H&E images.

F) Fluorescence image (optional but required for cell segmentation analysis).

Configuration File (config.txt)

```

### Input sequencing data. Supports .gz format.
FQ1      /path/to/read_1.fq.gz
FQ2      /path/to/read_2.fq.gz

### Fluorescence decoding file.
FLU      /path/to/flu_info.txt          # Decoding file path.

### AllheStat.py setting.
HE       /path/to/HE.tif               # H&E image path.

# INSIDE  0                            # Whether to recognise the blank region in the image. 0 is no, and 1 is yes.

# GRAY    200                          # Gray scale setting for the H&E image analysis. The default is automatic setting.

### Cell segmentation analysis.
## If this analysis is chosen, the fluorescence image path, colour channel, and
path of the high-resolution H&E image should be provided.

CellSplit      True                    # Whether to perform cell segmentation analysis. True is yes, otherwise
the analysis is disabled.

Fluorescence   /path/to/fluorescence.tiff          # Fluorescence image path.

fluorescence_channel      0              # Fluorescence image colour channel, default is 0.

# FLGRAY    15                          # Gray scale setting for the fluorescent image analysis.
Default is automatic setting.

# cells_numpy /path/to/cells/npyfile          # Pre-existing cell segmentation result file .npy. If
provided, it will be used for the cell segmentation analysis.

# YAML      /path/to/cell_split/parameter/file    # File in .yaml format with parameters for
cell segmentation analysis. This is optional.

## Reference genome STAR setting
GenomeVer     xxx                      # Genome version information. This will be written in the output report.

INDEX         /path/to/STAR/index/dir/         # STAR reference genome index file.

GFF           /path/to/ref/gene/gff3/file       # Reference genome annotation file. File in .gtf format is
also acceptable.

## Reference genome features.tsv file
FEATURE       /path/to/features.tsv

## Output
OUTDIR        /path/to/result/dir/             # Output path.
PREFIX        outfile-prefix                  # Output file prefix.

### Program Parameters
## fastq2BcUmi

```

```

BCType      V2      # Barcode version type (usually V2 version).
BCThreads   8      # Number of threads.

## Umi2Gene
Sjdboverhang 100    # Value of -sjdboverhang parameter used during STAR
                    indexing, default is 100.
STARThreads  8      # Number of threads used in read alignment in STAR.

## Environment setting. If not provided, the system default path will be used.
Please add a “#” at the beginning of the lines if not set.

```

```

PYTHON  /path/to/python/dir/      # Path to Python.
Rscript  /path/to/Rscript/dir/    # Path to R.

```

Note: For plant tissue, the fluorescence image is not required for the cell segmentation analysis. So CellSplit should be set to “True” to enable the analysis, and the H&E image should be provided. In this case, please comment out the lines of “Fluorescence” and “fluorescence_channl” by adding a “#” character at the beginning of the lines.

1.4 Running guide

1.4.1 Process steps

The process consists of 8 steps, as outlined below:

- Step 1: Run **fastq2BcUmi** to identify barcodes and UMIs in fastq data.
- Step 2: Run **LinkBcChip** to decode the chip locations of each barcode.
- Step 3: Run **Umi2Gene** to align reads to the reference genome and obtain gene information for each UMI.
- Step 4: Run **MatrixMake** to generate the gene expression matrix.
- Step 5: Run **AllheStat** to process H&E images.
- Step 6: Run **cluster.R** for cluster analysis.
- Step 7: Run **CellSplit** to perform cell segmentation analysis.
- Step 8: Run **WebReport** to generate a web based report.

1.4.2 Command options

- **-c config.txt** Data configuration file.
- **-s** selection analysis steps to perform: set to 0 to run all eight steps. Or choose specific steps separated by commas.

Note:

1. When selecting 0 and performing cell segmentation analysis, the CellSplit parameter in the configuration file should be set to True.
2. For plant tissue, the fluorescence image is not required for the cell segmentation analysis. So CellSplit should be set to “True” to enable the analysis, and the H&E image should be provided.

1.4.3 Example Command Lines

```
./BSTMatrix -c config.txt -s 0
./BSTMatrix -c config.txt -s 1,2,3,4,5,6,7,8
./BSTMatrix -c config.txt -s 1,3
```

1.5 Description of Result Files

The directory structure and contents of the result files are as follows:

outdir/

└─ 01.fastq2BcUmi

```
├─ xxx.bc_dist
├─ xxx.bc_stat
├─ xxx.bc_umi_read.tsv
├─ xxx.bc_umi_read.tsv.id
├─ xxx.filter
├─ xxx.full_stat
├─ xxx.id_map
├─ xxx.qual.stat
├─ xxx.select_id
├─ xxx.stat
├─ xxx.umi
└─ xxx.umi_cor.info
```

Step 1: Directory for barcode and UMI detection from fastq file

Barcode detection and stats
Barcode detection and stats
Barcode type, UMI and read number statistics
Barcode type, and their UMI and read ID
Reads with fractional barcode
Read and UMI number for barcode types
File containing ID mappings
Read statistics
ID of reads with an intact barcode and UMI
Barcode detection stats
Barcode types and UMIs for each read
UMI correction information

└─ 02.LinkBcChip

```
├─ xxx.barcode_pos.tsv
├─ xxx.barcode.tsv
├─ xxx.flu.stat
├─ xxx.info
└─ xxx.null
```

Step 2: Directory for barcode location

Barcode location on the chip
Barcode type for the each chip
Barcode decoding statistics
Barcode spatial information
Unrecognized chip position information

└─ 03.Umi2Gene

```
├─ xxxAligned.sortedByCoord.out.bam
├─ xxx.cut0.fq
├─ xxxLog.final.out
├─ xxxLog.out
├─ xxxLog.progress.out
├─ xxx.map2gene
├─ xxxSJ.out.tab
├─ xxx_STARTtmp
├─ xxx.stat
├─ xxx.total.stat
└─ xxx.umi_gene.tsv
```

Step 3: Directory for gene expression information

STAR alignment bam file
Read2 sequences used in alignment
STAR alignment summary
STAR log file
STAR progress log file
Information of reads mapped to genes
Splice junction info from STAR
STAR temporary files
Preliminary alignment statistics
Alignment summary statistics
UMIs and genes for each barcode

<ul style="list-style-type: none"> 04.MatrixMake <ul style="list-style-type: none"> xxx.matrix.tsv xxx.matrix.tsv.filt xxx.select.bc_umi_read.tsv xxx.select.umi_gene.tsv xxx.select.umi_gene.tsv.filter xxx.sequencing_saturation.stat xxx.sequencing_saturation.png 	Step 4: Directory for expression matrix <ul style="list-style-type: none"> Gene expression matrix Filtered matrix file UMIs and read number for each barcode UMI and gene info for each barcode Filtered barcode and their gene info Sequencing saturation analysis Sequencing saturation graph
<ul style="list-style-type: none"> 05.AllheStat <ul style="list-style-type: none"> allhe <ul style="list-style-type: none"> he_roi_small.png he_roi.tif roi_heAuto.json stat.txt all_level_stat.txt BSTViewer_project <ul style="list-style-type: none"> cell_split cluster he_roi_small.png he.tif imgs level_matrix project_setting.json roi_groups subdata heAuto_level_matrix <ul style="list-style-type: none"> subdata level_matrix <ul style="list-style-type: none"> level_1 level_13 level_2 level_3 level_4 level_5 level_6 level_7 stat.txt umi_plot <ul style="list-style-type: none"> all_umi_count_small.png all_umi_count.tif roi_umi_count_small.png roi_umi_count.tif roi_umi_count_white_small.png roi_umi_count_white.tif 	Step 5: Directory for tissue expression analysis results <ul style="list-style-type: none"> Directory for tissue region information <ul style="list-style-type: none"> PNG image of identified H&E tissue regions TIFF image of identified H&E tissue regions JSON file with tissue region information Tissue region statistics Statistics for different resolution levels BSTViewer software input data directory <ul style="list-style-type: none"> Directory for cell segmentation data Empty directory PNG image of identified H&E tissue regions H&E image file Empty directory Directory for expression matrix at different resolution levels BSTViewer project JSON file Directory for tissue and H&E image JSON files Directory for expression matrix at different resolution levels in tissue regions Directory for expression matrix at different resolution levels in tissue regions Directory for expression matrix at different resolution levels in tissue regions Directory for expression matrix at different resolution levels for chips <ul style="list-style-type: none"> Directory for level 1 expression matrix Directory for level 13 expression matrix Directory for level 2 expression matrix Directory for level 3 expression matrix Directory for level 4 expression matrix Directory for level 5 expression matrix Directory for level 6 expression matrix Directory for level 7 expression matrix Tissue region analysis statistics Directory for UMI plot results <ul style="list-style-type: none"> PNG image of UMI count in chip regions TIFF image of UMI count in chip regions PNG image of UMI count in tissue regions TIFF image of UMI count in tissue regions PNG image with white background for UMI count in tissue regions TIFF image with white background for UMI count in tissue regions
<ul style="list-style-type: none"> 06.Cluster <ul style="list-style-type: none"> L13 	Step 6: Directory for clustering analysis results <ul style="list-style-type: none"> Directory for level 13 clustering results

├─ cluster.csv	Cluster results
├─ L13_cluster_files	Directory for cluster HTML appendix files
├─ L13_cluster.html	HTML image of clustering results
├─ L13_cluster.pdf	PDF image of clustering results
├─ L13_cluster.png	PNG image of clustering results
├─ L13_umap_clstr.pdf	Merged PDF image of UMAP results
├─ L13_umap_clstr.png	Merged PNG image of UMAP results
├─ L13_umap_files	Directory for UMAP HTML appendix files
├─ L13_umap.html	HTML image of UMAP results
├─ L13_umap.pdf	PDF image of UMAP results
├─ L13_umap.png	PNG image of UMAP results
├─ L3	Directory for level 3 clustering
├─ cluster.csv	Cluster result file
├─ L3_cluster_files	Directory for cluster HTML appendix files
├─ L3_cluster.html	HTML image of clustering results
├─ L3_cluster.pdf	PDF image of clustering results
├─ L3_cluster.png	PNG image of clustering results
├─ L3_umap_clstr.pdf	Merged PDF image of UMAP results
├─ L3_umap_clstr.png	Merged PNG image of UMAP results
├─ L3_umap_files	Directory for UMAP HTML appendix files
├─ L3_umap.html	HTML image of UMAP results
├─ L3_umap.pdf	PDF image of UMAP results
├─ L3_umap.png	PNG image of UMAP results
... ..	
├─ L7	Directory for level 7 clustering
├─ cluster.csv	Cluster result file
├─ L7_cluster_files	Directory for cluster HTML appendix files
├─ L7_cluster.html	HTML image of clustering results
├─ L7_cluster.pdf	PDF image of clustering results
├─ L7_cluster.png	PNG image of clustering results
├─ L7_umap_clstr.pdf	Merged PDF image of UMAP results
├─ L7_umap_clstr.png	Merged PNG image of UMAP results
├─ L7_umap_files	Directory for UMAP HTML appendix files
├─ L7_umap.html	HTML image of UMAP results
├─ L7_umap.pdf	PDF image of UMAP results
├─ L7_umap.png	PNG image of UMAP results
├─ 07.CellSplit	Step 7: Directory for cell segmentation analysis
├─ cell_split_result	Directory for cell segmentation results
├─ 0_0.npy	Local cell segmentation results
├─ 0_0_ori.tif	Local fluorescent image
├─ 0_0.tif	Local fluorescent image after cell recognition
... ..	
├─ 9500_9500.npy	Local cell segmentation results
├─ 9500_9500_ori.tif	Local fluorescent image
├─ 9500_9500.tif	Local fluorescent image after cell recognition
├─ all_barcode_num.txt	Cell barcode IDs
├─ all_outline.tif	Fluorescent image with added cell boundaries
├─ cell_color.tif	Recognized cell image file
├─ cellConts.json	Recognized cell JSON file
├─ cells.npy	Recognized cell NPY file
├─ colors.npy	Cell and color mapping file
├─ conts.tif	Cell segmentation tissue boundary

			fluorescence.tif	Tissue fluorescent image
			nucleus_color.tif	Recognized cell nucleus image
			nucleusConts.json	Recognized cell nucleus JSON
			nucleus.npy	Recognized cell nucleus NPY
			progress.txt	Progress percentage file
			SegtoBarcode.log	Log file
			cluster	Directory for clustering results
			cell_cluster_color_img.tif	Cell segmentation clustering image without legend in TIFF format
			cell_cluster_color_outline_img.tif	Cell segmentation clustering image with added cell boundaries in TIFF format
			cell_cluster_with_legend_img.png	Cell clustering image with legend in PNG format
			cell_cluster_with_legend_img_small.png	Low-resolution cell clustering image with legend in PNG format
			cell_cluster_with_legend_img.tif	Cell clustering TIFF image with legend
			cluster.csv	Clustering results
			cluster_cells_num.csv	Cluster cell number count
			clusters_colors.npy	Cluster category and color mapping results
			colors.npy	Cell and color mapping results
			legend.tif	Cluster legend
			marker_gene.csv	Marker gene information
			object.RDS	Seurat object from cell segmentation matrix
			UMAP.pdf	UMAP clustering results in PDF format
			UMAP.png	UMAP clustering results in PNG format
			images	Directory for cell segmentation related image
			fluorescence_cell_split.png	Fluorescent PNG image cell segmentation
			fluorescence_cell_split_small.png	Low-resolution fluorescent PNG image cell segmentation result
			fluorescence_cell_split.tif	Fluorescent TIFF image cell segmentation results
			fluorescence.png	Tissue fluorescent PNG image
			fluorescence_small.png	Low-resolution tissue fluorescent image in PNG format
			fluorescence.tif	Tissue fluorescent image in TIFF format
			he_cell_split.png	Tissue H&E staining cell segmentation PNG image
			he_cell_split_small.png	Low-resolution tissue H&E staining cell segmentation result in PNG format
			he_cell_split.tif	Tissue H&E staining cell segmentation result in TIFF format
			he_hr.tif	Tissue H&E image in TIFF format
			mtx	Directory for cell segmentation matrix results
			barcodes.tsv.gz	Cell barcode file
			cells_center.txt	Cell center location on chip
			cells_center.tif	Cell center image
			features.tsv.gz	Cell features file
			matrix.mtx.gz	Cell matrix file
			stat.xls	Cell statistics file
			08.WebReport	Step 8: Directory for web-based report
			src	Directory for web-based report source files
			xxx.filelist	List of files used for the web-based report
			xxx.stat.xls	Analyses summary
			xxx.rs_stat.xls	Analyses summary

	└─ xxx.html	Web-based report file
└─	xxx	Directory for original expression matrix results
	├─ barcode_pos.tsv	Barcode and corresponding chip position file
	├─ barcode.tsv	Barcode file
	├─ bc_umi_read.tsv.gz	UMIs and read counts for each barcode
	├─ features.tsv	Features file
	├─ matrix.tsv	Matrix file
	└─ umi_gene.tsv.gz	UMIs and genes corresponding to barcodes

Please note that "xxx" represents the specific file or directory names generated during the execution of the pipeline.