MICTI Documentation

Release 0.0.3

Nigatu Ayele

Dec 16, 2018

Contents

1 Installation **3**

2 User's Guide **5**
- 2.1 Creat MICTI Object . . . 5
- 2.2 Data visualisation . . . 6
- 2.3 Clustering cells . . . 6
- 2.4 Cell-type marker genes . . . 7
- 2.5 significant cluster markers . . . 7
- 2.6 Gene-list enrichment analysis . . . 8

3 Tutorials **9**
- 3.1 Import MICTI module . . . 9
- 3.2 Import data . . . 9
- 3.3 Creating MICTI object for known cell-types . . . 11
- 3.4 Lower dimensional data visualization . . . 11
- 3.5 Creating MICTI object for clustering cells into pre-defined k clusters . . . 12

Recent advances in single-cell gene expression profiling technology have revolutionized the understanding of molecular processes underlying developmental cell and tissue differentiation, enabling the discovery of novel cell types and molecular markers that characterize developmental trajectories. Common approaches for identifying marker genes are based on pairwise statistical testing for differential gene expression between cell types in heterogeneous cell populations, which is challenging due to unequal sample sizes and variance between groups resulting in little statistical power and inflated type I errors.

We developed an alternative feature extraction method, Marker gene Identification for Cell Type Identity (MICTI), that encodes the cell-type specific expression information to each gene in every single cell. This approach identifies features (genes) that are cell-type specific for a given cell-type in heterogeneous cell population.

Contents:

CHAPTER 1

Installation

1. Obtain Python 3.5 and virturalenv.
2. Create a virtual environment somewhere on your disk, and then activate it.

```
$ virtualenv --no-site-packages --python=python3.5 micti_env
$ source micti_env/bin/activate
```

3. Download the source code and install the requirements.

```
$ pip install MICTI
```

   pip will install the following packages:

   - NumPy
   - SciPy
   - matplotlib
   - pandas
   - gprofiler
   - seaborn
   - scikit-learn

4. Import MICTI.

```
$from MICTI import *
```

CHAPTER 2

User's Guide

2.1 Creat MICTI Object

$MICTI(data,geneNames,cellNames,k=None,cluster_label=None,cluster_assignment=None, th=0,seed=None, ensembel=False, organisum="hsapiens")

2.1.1 Input

data

Input data as sparce or dense matrix where the rows are cells and the columns are genes

geneNames

List of gene names

cellNames

List of cell names

k

The number of clusters or cell types

cluster_label

List of cluster lablees /cell types names

cluster_assignment

An aaray of cluster assignment for each of cells

th

The treshold gene expression value to consider a certain gene is expressed or not

ensembel

A boolian value indicating the given gene name is ENSEBEL gene Id or not

organisum

The organisum where dataset belong eg. hsapiens or mmusculus

2.1.2 Output

The output is the MICTI object

2.2 Data visualisation

```
$MICTI.get_Visualization(dim=2,method="tsne")
```

2.2.1 Input

dim

The number of dimension for visualisation dim=2 or dim=3

method

The method used for low dimensional visualisation, method="PCA" or method="tsne"

2.2.2 Output

Returns none. Desplays the lower dimensional representation of the dataset

2.3 Clustering cells

```
$MICTI.cluster_cells(numberOfCluster, method="kmeans", maxiter=500)
```

2.3.1 Input

numberOfCluster

The number of clusters

method

The method used for clustering. There are two options, ie. method="kmeans" for kmeans clustering or method="GM" gaussian mixture model for clustering

maxiter

The maximum iteration that the k-means or Gaussian mixture algorithm takes in the clustering process.

2.3.2 Output

Returns None, assigning each cells into k clusters

2.4 Cell-type marker genes

```
$MICTI.marker_gene_FDR_p_value(clusterNo)
```

2.4.1 Input

clusterNo

The cluster number. Each clusters are identified by number. For example, if there are six clusters/cell-types, the cluster numbers are from 0-5.

2.4.2 Output

Returns a table with Z-score, p-value and FDR p-value for each of the genes.

2.5 significant cluster markers

```
$MICTI.get_markers_by_Pvalues_and_Zscore(clusterNo,threshold_pvalue=.01,
threshold_z_score=0)
```

2.5.1 Input

clusterNo

The cluster number. Each clusters are identified by number. For example, if there are six clusters/cell-types, the cluster numbers are from 0-5.

threshold_pvalue

The threshold FDR p-value. Genes/Markers with less than the threshold FDR p-value are selected.

threshold_z_score

The threshold Z-scores. Genes/markers with greater than the threshold z-score are selected.

2.5.2 Output

Returns a table with Z-score, p-value and FDR p-value of significantlly cell-type/cluster marker genes filtered by FDR Pvalue and Z-score.

2.6 Gene-list enrichment analysis

```
$MICTI.get_sig_gene_over_representation()
```

2.6.1 Input

None

2.6.2 Output

Returns a list with gene-list enrichment analysis result for each of cell-type/cluster marker genes

CHAPTER 3

Tutorials

We developed an alternative feature extraction method, Marker gene Identification for Cell Type Identity (MICTI), that encodes the cell-type specific expression information to each gene in every single cell. This approach identifies features (genes) that are cell-type specific for a given cell-type in heterogeneous cell population.

3.1 Import MICTI module

```
$from MICTI import *
```

3.2 Import data

We collected single-cell RNA-Seq dataset from six different immune cell types. We performed TPM normaization for each of samples.

```
$import pandas as pa
```

```
$datamatrix=pa.read_csv("dataset.txt", sep="\t", index_col="genes")
```

Genes	GSM2181361	GSM2181362	GSM2181363	GSM2180852	GSM2181258	GSM2181259	GSM2180849	GSM2181360	GSM2181364	GSM2180853
A1BG	0.000000	0.043549	0.054509	0.000000	0.000000	0.066542	0.605715	0.651164	0.095305	0.000000
A1CF	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
A2M	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
A2ML1	0.046830	0.071208	0.018045	0.000000	0.000000	0.023222	0.531418	0.050903	0.098627	0.000000
A4GALT	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
AAAS	39.244719	4.173193	28.947780	0.000000	67.050516	97.502654	0.000000	2.375844	88.972850	341.262077
AACS	0.623697	0.401357	0.362420	0.777686	0.270946	0.893264	0.860927	0.546757	1.002484	0.000000
AADACL3	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
AADAT	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
AAED1	8.078604	8.696563	6.825583	4.692559	0.904554	0.456029	6.191677	12.625448	11.592398	10.103919

More information about the samples can be found from the metadata information. Metadata information contains disease stages, tissue catagory, sample source and other important information about the sample/cell. From the metadata table we extracted cell types/sample source in order to classify our cells according to cell type.

```
$metadata=pa.read_csv("metadata.txt", sep="\t", index_col="SampleID")
```

SampleID	SubjectID	DiseaseCategory	TissueCategory	BamFileName	CellType	Description	DiseaseStage	DiseaseState	Ethnicity
GSM2181041	No Info	hematologic cancer	hematopoietic system	EGAX00001437341.bam	lymphoblast	processed data file = cell_line_FPKM.csv	No Info	chronic myeloid leukemia (CML)	No Info
GSM2181022	No Info	hematologic cancer	hematopoietic system	EGAX00001437284.bam	lymphoblast	processed data file = cell_line_FPKM.csv	No Info	chronic myeloid leukemia (CML)	No Info
GSM2181013	No Info	hematologic cancer	hematopoietic system	EGAX00001437257.bam	lymphoblast	processed data file = cell_line_FPKM.csv	No Info	chronic myeloid leukemia (CML)	No Info
GSM2180862	No Info	hematologic cancer	hematopoietic system	EGAX00001437608.bam	B cell	processed data file = cell_line_FPKM.csv	No Info	B-cell lymphoma	No Info
GSM2181258	No Info	hematologic cancer	hematopoietic system	EGAX00001439870.bam	B cell	processed data file = cell_line_FPKM.csv	No Info	B-cell lymphoma	No Info

Now we have cell-type information for each of our samples/cells from the metadata table. So we wanted to get markers for each of the cell-types using MICTI

```
$cell_type=list(metadata["CellType"])
```

```
$set(cell-type)
```

```
{'B cell',
'CD4+ memory T cell',
'CD8+ memory T cell',
'conventional dendritic cell',
'fibroblast',
'lymphoblast'}
```

```
$geneName=list(datamatrix.index)
```

```
$print(geneName[:10])
```

```
['A1BG', 'A1CF', 'A2M', 'A2ML1', 'A4GALT', 'AAAS', 'AACS', 'AADACL3', 'AADAT',
'AAED1']
```

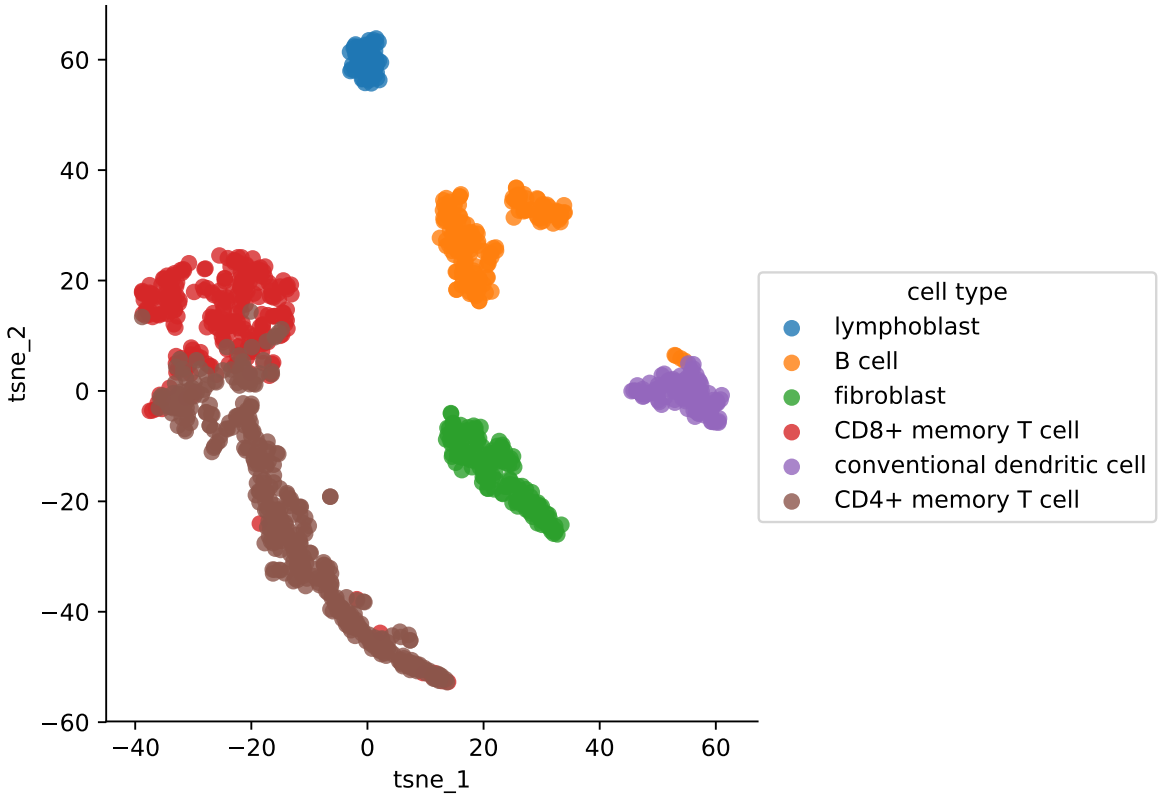
```
$cellName=list(datamatrix.columns)
```

3.3 Creating MICTI object for known cell-types

```
$mictiObject=MICTI(datamatrix, geneName, cellName, cluster_assignment=cell_type,
k=None, th=0, ensembel=False, organisum="hsapiens")
```

3.4 Lower dimensional data visualization

```
$mictiObject.get_Visualization(method="tsne")
```



3.4.1 Marker genes for each cluster

```
$mictiObject.marker_gene_FDR_p_value(0)
```

Genes	Z_scores	fdr	p_value
HLA-DRA	40.605319	0.000000e+00	0.000000e+00
MS4A1	40.199070	0.000000e+00	0.000000e+00
TUBB	15.099339	0.000000e+00	0.000000e+00
HLA-DPA1	14.701781	0.000000e+00	0.000000e+00
RPS18	61.131416	0.000000e+00	0.000000e+00

3.4.2 Marker genes for each cluster by P-value and Z-Score threshold

```
$mictiObject.get_markers_by_Pvalues_and_Zscore(1, threshold_pvalue=.01,
threshold_z_score=0)
```

Genes	Z_scores	fdr	p_value
CSF2	20.313988	0.000000e+00	0.000000e+00
IL2RG	12.560409	0.000000e+00	0.000000e+00
ATP9B	28.123272	0.000000e+00	0.000000e+00
HIST1H2BK	9.118146	0.000000e+00	0.000000e+00
PATL2	9.055203	0.000000e+00	0.000000e+00
CTLA4	8.523849	0.000000e+00	0.000000e+00
CCL20	11.984467	0.000000e+00	0.000000e+00
MAP3K14	32.571130	0.000000e+00	0.000000e+00
GZMB	17.080777	0.000000e+00	0.000000e+00
GPR171	10.677701	0.000000e+00	0.000000e+00

3.4.3 Enrichment analysis for identified marker genes

Get gene-over representation enrichmentlysis result for cel-type marker genes in all clusters of cell type

```
$enrechment_table=mictiObject.get_sig_gene_over_representation()
```

```
$enrechment_table[1]
```
#CD4+ cells

3.5 Creating MICTI object for clustering cells into pre-defined k clusters

In case, if the cell-type information for each cells is not known, we can perform unsupervided clustering to differentiate cells into predifined k clusters. Here, we use K-means and Gaussian mexture mode for clustering.

3.5.1 Creat MICTI object

```
$mictiObject_1=MICTI(datamatrix, geneName, cellName, cluster_assignment=None,
th=0, ensembel=False, organisum="hsapiens")
```

3.5.2 Cluster cells into k clusters

Cluster cells into k=6 clusters using Gaussian mixture model- method="GM", and k-means - method="kmeans"

```
$mictiObject_1.cluster_cells(6, method="GM", maxiter=10e3)
```

3.5.3 Marker genes per each cluster

#markers for the third cluster

```
$mictiObject_1.get_markers_by_Pvalues_and_Zscore(2, threshold_pvalue=.01,
threshold_z_score=0)
```

3.5.4 Gene-list Enrichment analysis for cluster marker genes

```
$enrechment_table=mictiObject_1.get_sig_gene_over_representation()
$enrechment_table[0]# Enrichment result for the first cluster
```