Cell-cycle assignment using Mixture Models

Tallulah Andrews

September 17, 2022

Introduction

```
> set.seed(1973)
> require("SingleCellExperiment")
> require("scater")
> require("CycleMix")
```

Genesets

This package provides some published cell-cycle marker gene sets for human (HGeneSets) and mouse (MGeneSets). Sources of the gene sets can be found in the manual e.g. `?HGeneSets`.

```
> names(HGeneSets)
```

```
[1] "Whitfield" "Tirosh"    "Macosko"   "Quiesc"
```

You can use your own geneset simply by creating a data.frame of the same format:

```
> head(MGeneSets$Cyclone)
```

```
                     Gene Stage Dir
ENSMUSG00000000001  Gnai3    G1  -1
ENSMUSG00000000028  Cdc45     S   1
ENSMUSG00000000355  Mcts1     S   1
ENSMUSG00000000486  Sept1     S   1
ENSMUSG00000000708  Kat2b     S   1
ENSMUSG00000000743 Chmp1a    G1  -1
```

The "Dir" column is used to weight the gene expression we have simplified this to positive (1) and negative (-1) markers, but any numeric value can be used, for instance fold change.

We can convert human cell-cycle genes to mouse orthologs using `biomaRt` with our provided wrappers. For instance if we want to use the quiescence markers in mouse we would convert them as so:

```
> require("biomaRt")
> #map <- downloadEnsemblData()
> mouse_g0 <- HGeneSets$Quiesc
> #mouse_g0$Gene <- mapGeneNames(map, mouse_g0$Gene, in.name="symbol", in.org="Hsap", out.
> mouse_g0 <- mouse_g0[mouse_g0$Gene != "",]
```

Gene sets can also be combined together easily:

```
> mouse_CC <- rbind(mouse_g0, MGeneSets$Cyclone)
```

Input Data

CycleMix requires a SingleCellExperiment object as input. For this tutorial we have provided some example data already formatted correctly. Crucially there must be a column of the rowData dataframe which contains gene IDs matching your marker gene table.

```
> head(rowData(Ex))
```

```
DataFrame with 6 rows and 1 column
                   feature_symbol
                         <factor>
ENSMUSG00000000001          Gnai3
ENSMUSG00000000028          Cdc45
ENSMUSG00000000049          Apoh
ENSMUSG00000000078          Klf6
ENSMUSG00000000126          Wnt9a
ENSMUSG00000000148          Baat1
```

You should also have a log-transformed normalized expression matrix in the SingleCellExperiment object, in our case this matrix is called "logcounts":

```
> names(assays(Ex))
```

```
[1] "counts"    "logcounts"
```

See the SingleCellExperiment or scater packages for details on creating and normalizing data in SingleCellExperiment objects.

Assigning Cells

We can now assign cell-cycle stages to our cells. For our example data we know all the cells are cycling thus won't include the quiescence markers.

```
> output <- classifyCells(Ex, MGeneSets$Cyclone)
> summary(factor(output$phase))
```

```
  G1  G2M None    S
  89   60   27   97
```

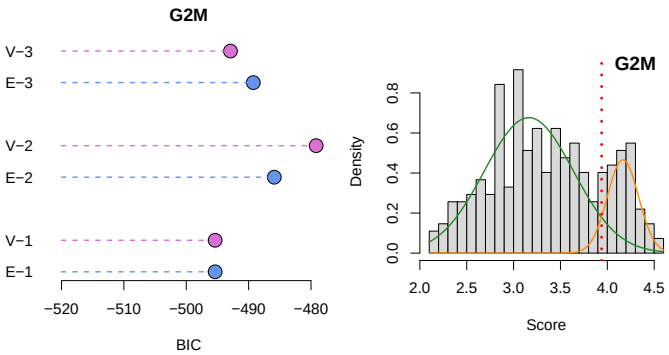
Our example data comes from staged cells so we can compare our assignments to the ground truth:

```
> table(factor(output$phase), Ex$cell_type1)
```

```
      G1 G2M  S
G1    64  13 12
G2M    0  60  0
None   6  19  2
S     25   4 68
```

We can also examine the gaussian mixture model fit to each cell-cycle stage:

```
> plotMixture(output$fit[["G2M"]], BIC=TRUE)
```



The plot on the left shows the 6 different models considered : mixtures of 1-3 gaussian distributions with equal (E) or different (V) variances. The BIC criterion was used to select the optimal model. The plot on the right shows the distribution of expression scores across all cells. The curves of the fitted distributions are plotted on top. The threshold for assigning cells to the stage (if applicable) is indicated with the red dotted line.