User's guide to flowMap

Chiaowen Joyce Hsiao[1*], Yu Qian[2], Richard H. Scheuermann[2]

[1] Center for Bioinformatics and Computational Biology (CBCB) and
and Applied Mathematics and Scientific Computation,
University of Maryland, College Park, US;
[2] Department of Informatics, J. Craig Venter Institute (JCVI), San Diego, US
[*]joyce.hsiao1 (at) gmail.com

August 30, 2014

Abstract

flowMap is a stand-alone software for mapping cell populations in multiple flow cytometry (FCM) data. A core issue in cell population matching is the ability of the algorithm to accurately quantify similarity between cell populations differing in proportion, size, and levels of expression markers. Our algorithm implements the Friedman-Rafskty (FR) statistic to compare similarity between cell populations across multiple flow cytometry samples. Users can generate a matrix of FR statistic, which can then be used to determine cell population groups across samples. The method can be incorporated in any standard flow cytometry sample processing pipeline at any stage where comparison of cell populations is required. We demonstrated the ability of the FR statistic in scenaiors of biological and technical differences between flow cytometry samples in C. Hsiao, M. Liu, R. Stanton, M. McGee, Y. Qian, and R. H. Scheuermann: Mapping cell populations in flow cytometry data for cross-sample comparison using the Friedman-Rafsky Test (2014) [1].

Contents

1 Introduction

In this vignette, we show how to use *flowMap* to compare cell populations cell populations across flow cytometry samples, and to visualize the results.

2 Data prepartion

flowMap input accepts any flow cytometry sample files with identified cell populations. Thus, users can use *flowMap* at any step of a FCM workflow to quantify similarity of cell populations. In order to use flowMap as a downstream analysis tool to compare phenotypes, the input FCM sample files need to have been preprocessed for debris filtering, transformation, and marker expression alignment.

Each FCM sample input needs to be in the matrix form of numeric values. Rows correspond to the events (cells) and the columns correspond to marker expression measurements. The last column of the sample matrix is usually named as `id`, serving as cell population membership index containing numeric values.

2.1 Example data

Here's a flow cytometry sample in *txt* format. There are 9 cell populations identified in the sample from 20,000 events measured in 4 feature markers (CD14,CD23,CD3,CD19,id). `id` index the cell population membership of each event from 1, 2, 3 through 9.

```
sam1 <- read.table(system.file("extdata/sample.txt",package="flowMap"),
                   header=T)
str(sam1)
```

```
## 'data.frame': 20000 obs. of  5 variables:
##  $ CD14: int  186 116 287 148 115 146 18 0 171 173 ...
##  $ CD23: int  0 272 370 111 198 178 53 0 290 338 ...
##  $ CD3 : int  216 232 349 576 481 553 577 91 263 333 ...
##  $ CD19: int  198 175 288 104 217 269 0 42 129 254 ...
##  $ id  : int  1 4 8 7 7 7 7 1 4 8 ...
```

```
table(sam1$id)
```

```
##
##    1    2    3    4    5    6    7    8    9
## 1641  809  330 2363 3422  943 7380 2788  324
```

3 FR statistic to quantify similarity between cell populations

The Friedman-Rafsky statistic is based on the minimum spanning tree (MST) algorithm which computes the extent to which the two cell populations overlap in their shared feature space [2]. Because the runtime of the MST algorithm is quadratic in the number of events, we devised a downsampling scheme to estimate the FR statistics in a single cell population pair comparison. First, for any cell population pair comparison, the events are combined to form `pooled data`. Next, samples containing the same number of events (default: 200) are taken from the pooled data. Each event in the pooled data may be sampled more than once. Key idea is to maintain a constant ratio of events from the two cell populations across samples. In each of the random samples, a MST is identified, followed by the FR statistic computation. The estimated FR statistic for each cell population pair comparison is based on the median of the FR statistics across the random samples.

Below are examples of MST finding when comparing two FCM samples. *Sample 1* and *Sample 2* each contains 9 cell populations.

```
sam1 <- read.table(system.file("extdata/sample.txt",package="flowMap"),
                   header=T)
sam2 <- read.table(system.file("extdata/sample.txt",package="flowMap"),
                   header=T)
table(sam1$id)
```

```
##
##    1    2    3    4    5    6    7    8    9
## 1641  809  330 2363 3422  943 7380 2788  324
```

```
table(sam2$id)
```

```
##
##    1    2    3    4    5    6    7    8    9
## 1641  809  330 2363 3422  943 7380 2788  324
```

3.1 Visualize the minimum spanning tree

3.1.1 Example 1

When events with different cell population membership are distant from each other, or in other words, events with the same membership congregate, the FR statistic determines the two cell populations to be dissimilar from each other. CP1 from Sample 1 is compared with CP3 from Sample 2 in the MST below. 100 events is sampled from the pooled data combinig events from the two cell populations, with the ratio of the cell population membership kept the same as that in the pooled data before sampling.

```
mat1 = sam1[sam1$id==1,]
mat2 = sam2[sam2$id==3,]

# combine events from the two cell populations to
# make pooled data
mat = rbind(mat1,mat2)

# sample 100 events from the pooled data
sampleSize = 100

# among the 100 events, sample events from the two cell populations
# such that the ratio of the cell population membership is the same
# as that in the pooled data
nn1 = round(sampleSize*table(mat$id)[1]/nrow(mat))
nn2 = round(sampleSize*table(mat$id)[2]/nrow(mat))
submat = rbind(mat1[sample(nrow(mat1),nn1),],mat2[sample(nrow(mat2),nn2),])
colnames(submat)[5] = "sam"

# plot MST of the 100 events
g1 = makeFRMST(submat)
par(mar=c(0,0,0,0))
plot(g1$g,vertex.label.cex=0.01,
     layout=layout.fruchterman.reingold(g1$g))
```

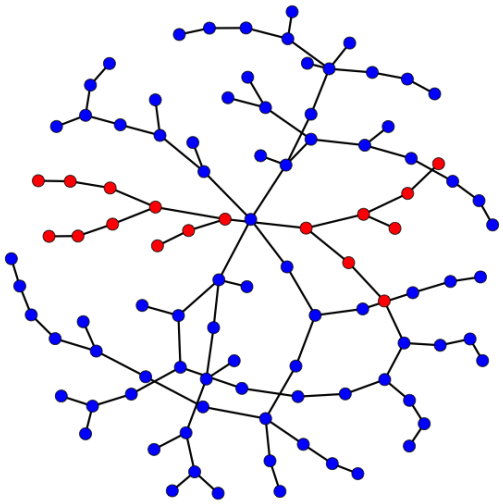

Figure 1: **Example 1: MST of sample events from two dissimilar cell populations** Blue nodes belong to Sample 1 CP1, and red nodes belong to Sample 2 CP 3.

3.1.2 Example 2

This is another example of two cell populations being dissimilar from each other. Events in the pooled data tend to congregate on the tree with events of the same cell population membership.

```
mat1 = sam1[sam1$id==4,]; mat2 = sam2[sam2$id==5,]
mat = rbind(mat1,mat2)
sampleSize = 100
nn1 = round(sampleSize*table(mat$id)[1]/nrow(mat))
nn2 = round(sampleSize*table(mat$id)[2]/nrow(mat))
submat = rbind(mat1[sample(nrow(mat1),nn1),],mat2[sample(nrow(mat2),nn2),])
colnames(submat)[5] = "sam"
g1 = makeFRMST(submat)
par(mar=c(0,0,0,0))
plot(g1$g,vertex.label.cex=0.01,layout=layout.fruchterman.reingold(g1$g))
```

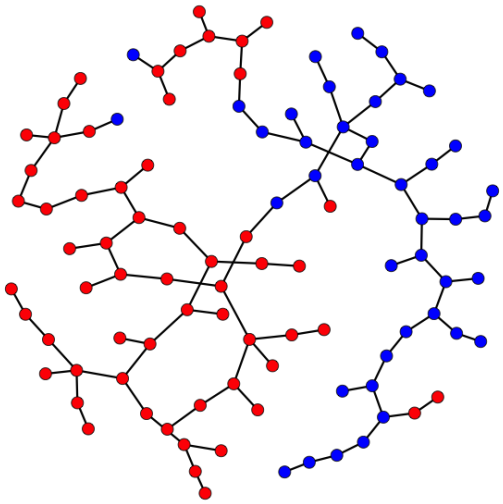

Figure 2: **Example 2: MST of sample events from two dissimilar cell populations** Blue nodes belong to sample 1 CP4, and red nodes belong to sample 2 CP5.

3.1.3 Example 3

When two cell populations share a similar feature space, events of different cell population membership tend to distribute evenly on the tree. Below is an example of Sample 1 CP6 compared with Sample 2 CP6.

```
mat1 = sam1[sam1$id==6,]
mat2 = sam2[sam2$id==6,]
mat1$id=1; mat2$id=2
mat = rbind(mat1,mat2)
sampleSize = 100
nn1 = round(sampleSize*table(mat$id)[1]/nrow(mat))
nn2 = round(sampleSize*table(mat$id)[2]/nrow(mat))
submat = rbind(mat1[sample(nrow(mat1),nn1),],mat2[sample(nrow(mat2),nn2),])
colnames(submat)[5] = "sam"
g1 = makeFRMST(submat)
par(mar=c(0,0,0,0))
plot(g1$g,vertex.label.cex=0.01,layout=layout.fruchterman.reingold(g1$g))
```

4 Mapping cell populations across FCM samples

flowMap directly computes the similarity betweeen cell populations across FCM samples and provides results in a table format. As a proof-of-concept, we compared a FCM sample against itself.

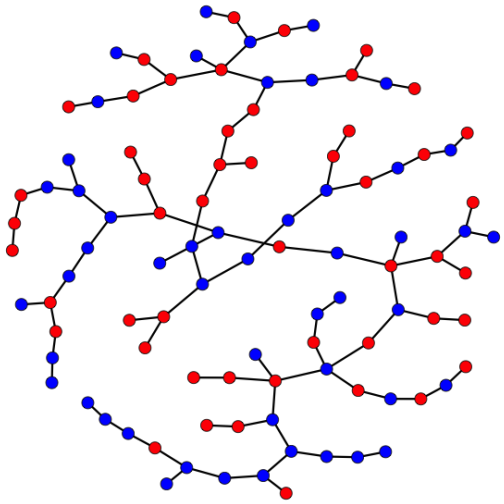

Figure 3: **Example 3: MST of sample events from two similar cell populations** Blue nodes belong to sample 1 CP6, and red nodes belong to sample 2 CP6.

4.1 Import data

The example data contains 9 identified cell populations.

```
sam1 <- read.table(system.file("extdata/sample.txt"
            ,package="flowMap"),header=T)
sam2 <- read.table(system.file("extdata/sample.txt"
            ,package="flowMap"),header=T)
table(sam1$id)
```

```
##
##    1    2    3    4    5    6    7    8    9
## 1641  809  330 2363 3422  943 7380 2788  324
```

```
table(sam2$id)
```

```
##
##    1    2    3    4    5    6    7    8    9
## 1641  809  330 2363 3422  943 7380 2788  324
```

4.2 Parameter setting in FR statistic computation

In order to optimize runtime, we devised a downsampling scheme to estimate FR statistics. The central idea is to draw random samples from the pooled data combining events in a cell population comparison and to estimate the FR statistics of the pooled data from the random samples. Parameters required are: number of random samples (`ndraws`; Default: 200), size of each random sample (`sampleSize`; Default: 200), sampling method (`sampleMethod`; Default: proportional), and number of processing cores (`ncores`; Default; maximum number of cores available in the computing environment). The

parallel computing function in *flowMap* is built upon the *doParallel* package.

The number of random samples and the size of each random samples determine the precision (variability of the FR statistics across random samples) and the accuracy (deviation of the sample FR statistics from the true FR statistic) of the estimated FR statistic when comparing any two cell populations. As illustrated in Hsiao et al., (2014), the ranks of cell population pairs remain the same when increasing the number of random samples to 500 or when increasing the size of the random sample to 500 events. Users are advised to use the default paramter setting when mapping cell populations when mapping cell populations.

4.3 Compute FR statistics

To compare any two FCM samples, users can use `getFRest` to obtain a matrix of estimated FR statistics comparing any two cell populations across samples. Rows and columns in the result matrix corresponds to the first and the second input sample in the `getFRest` function. For example, the (2,1) entry in the result matrix corresponds to the FR statistic comparing Sample 1 CP2 and Sample 2 CP1. The (4,3) entry in the result matrix corresponds to the FR statistic comparing Sample 1 CP4 and Sample 2 CP3.

```
res1 = getFRest(sam1,sam2,sampleMethod="proportional",sampleSize=100,
                ndraws=100,estStat="median",ncores=NULL)
res1@ww
```

```
##        1     2      3      4     5     6     7      8     9
## 1  0.202 -9.34 -7.627 -9.130 -8.50 -9.55 -7.64 -9.578 -7.41
## 2 -9.334  0.00 -8.937 -8.824 -7.65 -9.86 -6.09 -8.356 -9.04
## 3 -7.589 -8.86  0.404 -6.566 -5.84 -8.78 -3.71 -5.850 -8.88
## 4 -9.135 -8.81 -6.634 -0.202 -7.99 -9.09 -8.55 -9.053 -6.61
## 5 -8.581 -7.65 -5.796 -7.888  0.00 -8.39 -9.29 -8.356 -5.80
## 6 -9.548 -9.86 -8.780 -8.971 -8.36  0.00 -6.34 -8.617 -8.65
## 7 -7.555 -6.04 -3.706 -8.519 -9.29 -6.32  0.00 -8.850 -3.71
## 8 -9.590 -8.33 -5.983 -9.051 -8.22 -8.66 -8.87  0.202 -5.94
## 9 -7.430 -8.97 -8.882 -6.566 -5.80 -8.66 -3.71 -6.044  0.00
```

```
library(gplots)
par(mar=c(0,0,0,0))
heatmapCols <- colorRampPalette(c("red","yellow","white","blue"))(50)
heatmap.2(res1@ww,trace="none",col=heatmapCols,symm=FALSE,dendrogram="none",
          Rowv=FALSE,Colv=FALSE,xlab="Sample 2",ylab="Sample 1")
```

Users can also extract p-values of the FR statistics in the slot `pNorm`. Heatmap of p-values shows that the p-values of the matched cell population pairs (diagonal entries) are clearly separated from the p-values of the mismatched cell population pairs (off-diagonal entries).

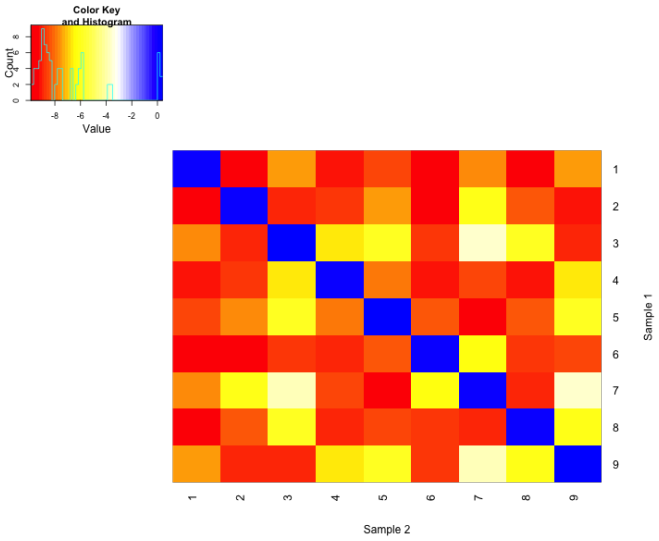

Figure 4: **Heatmap comparing cell populations between Sample 1 and Sample 2** The matched populations (in the diagonal) are quantified with significantly smaller FR statistic values than the mismatched populations (correspond to off-diagonal entries). The (i,j) cell contains the FR statistic of comparing the Sample 1 i-th cell population with Sample 2 j-th cell population

```
library(gplots)
par(mar=c(0,0,0,0))
heatmapCols <- colorRampPalette(c("red","yellow","white","blue"))(50)
heatmap.2(res1@pNorm,trace="none",col=heatmapCols,symm=FALSE,dendrogram="none",
          Rowv=FALSE,Colv=FALSE,xlab="Sample 2",ylab="Sample 1")
```

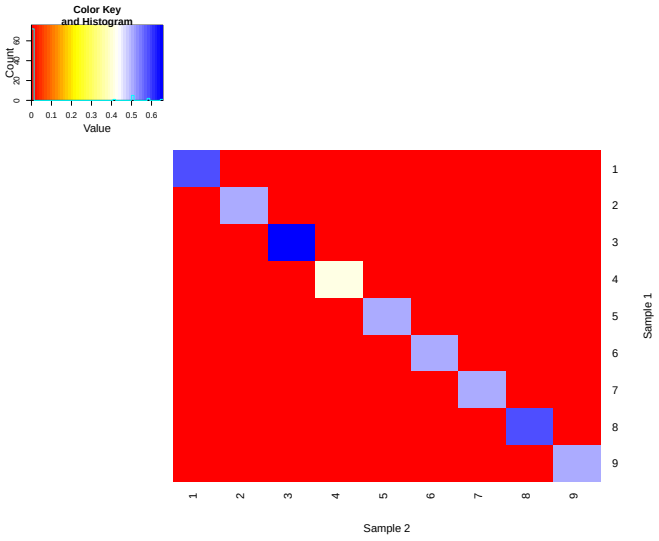

Figure 5: **Heatmap of the FR statistics p-values comparing Sample 1 with Sample 2** The p-values of the matched populations (in the diagonal) are significantly larger than the mismatched cell populations.

```
hist(res1@pNorm,xlab="log10 p-value histogram",main="")
```

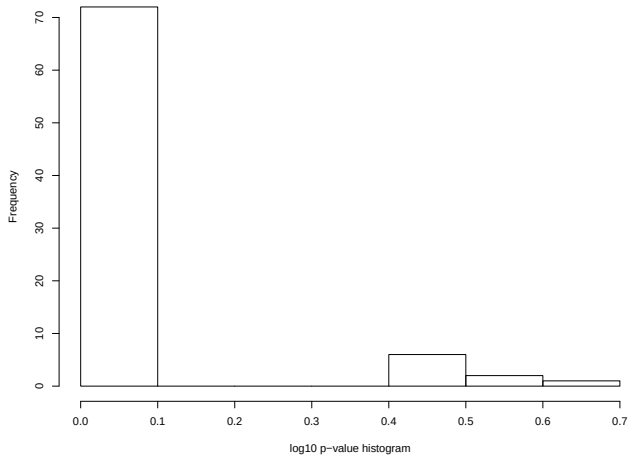

Figure 6: **Histogram of the FR statistics p-values mapping cell populations between Sample 1 with Sample 2** The bimodal distribution of the log10 p-values indicates a clear gap beteween p-values of matched cell population pairs versus p-values of mistmatched cell population pairs. The matched pairs have significantly larger p-values (p¿.4) than the mismatched pairs (p¡.01).

4.4 Generate a multi-sample similarity matrix for clustering

(makeDistmat) generates a complete similarity matrix containing FR statistics that is useful for multiple sample comparisons. Using the above Sample 1 versus Sample 2 example, user can obtain a 18-by-18 matrix for hierarchical clustering.

```
resMulti = makeDistmat(samples=list(sam1,sam2),sampleSize=100,ndraws=100)
require(gplots)
par(mar=c(0,0,0,0))
heatmapCols <- colorRampPalette(c("red","yellow","white","blue"))(50)
heatmap.2(resMulti$distmat,trace="none",col=heatmapCols,symm=FALSE,dendrogram="none",
          Rowv=FALSE,Colv=FALSE)
```

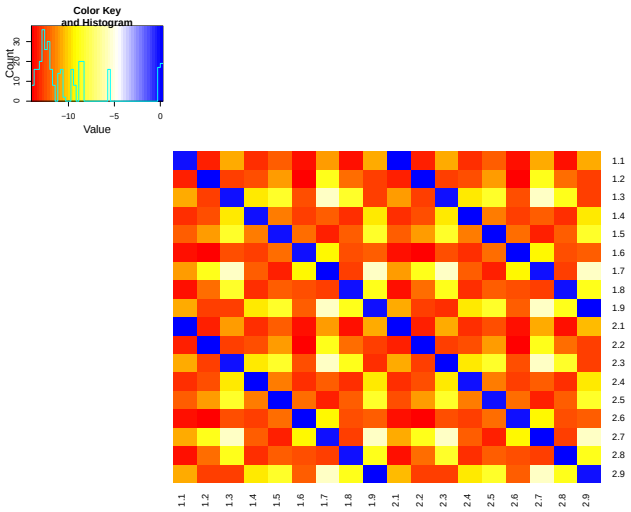

Figure 7: **Heatmap of the Sample 1 versus Sample 2 comparison similarity matrix** Row and column labels 1.x and 2.y denote Sample 1 cell populations and Sample 2 cell populations, respectively.

5 SessionInfo

The last part of this vignette calls for the function `sessionInfo`, which reports the computing environment in the session, including the R version number and all the packages used. Users should check their computing environment for consistency with this document as a first step in evaluating the errors/issues while using the *flowMap*.

- R version 3.1.1 (2014-07-10), `x86_64-apple-darwin13.1.0`
- Locale: `en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8`
- Base packages: base, datasets, graphics, grDevices, methods, parallel, stats, utils
- Other packages: abind 1.4-0, ade4 1.6-2, doParallel 1.0.8, flowMap 1.99.2, foreach 1.4.2, ggplot2 1.0.0, gplots 2.14.1, igraph 0.7.1, iterators 1.0.7, knitr 1.6, Matrix 1.1-4, reshape2 1.4, scales 0.2.4
- Loaded via a namespace (and not attached): BiocStyle 1.2.0, bitops 1.0-6, caTools 1.17, codetools 0.2-9, colorspace 1.2-4, compiler 3.1.1, digest 0.6.4, evaluate 0.5.5, formatR 1.0, gdata 2.13.3, grid 3.1.1, gtable 0.1.2, gtools 3.4.1, highr 0.3, KernSmooth 2.23-12, lattice 0.20-29, MASS 7.3-34, munsell 0.4.2, plyr 1.8.1, proto 0.3-10, Rcpp 0.11.2, stringr 0.6.2, tools 3.1.1

```
## list()
```

References

[1] Chiaowen Hsiao, Menyga Liu, Rick Stanton, Monnie McGee, Yu Qian, and Richard H Scheuermannn. Mapping cell populations in flow cytometry data for cross-sample comparison using the Friedman-

Rafsky Test. Submitted for publication, 2014.

[2] Jerome H Friedman and Lawrence C Rafsky. Multivariate generalizations of the Wald-Wolfowitz and Smirnov two-sample tests. *The Annals of Statistics*, 7(4):697–717, 1979.