flowType: Phenotyping Flow Cytometry Assays

Nima Aghaeepour

January 4, 2019

naghaeep@gmail.com

Contents

1 Licensing **1**

2 Introduction **1**

3 Installation **2**

4 Loading the Library **2**

5 Running flowType **2**

6 Deciding on a cutoff **5**

7 Cross-sample Analysis **6**

1 Licensing

2 Introduction

This document demonstrates the functionality of the flowType package for phenotyping FCM assays. flowType uses a simple threshold, Kmeans, flowMeans or flowClust to partition every channel to a positive and a negative cell population. These partitions are then combined to generate a set of multi-dimensional phenotypes. For more details on how this package can be used in a complete analysis pipeline, please refer to the manuscript that describes analysis of 466 HIV^+ patients [1, 3]. Further examples are available through [2, 4, 5].

3 Installation

flowType relies on the boost libraries. See www.boost.org for installation instructions. In Mac OS X you should be able to use homebrew (http://mxcl.github.com/homebrew/): *#brew install boost*

4 Loading the Library

We start by loading the library (for installation guidlines see the Bioconductor website).

```
> library(flowType)
> data(DLBCLExample)
```

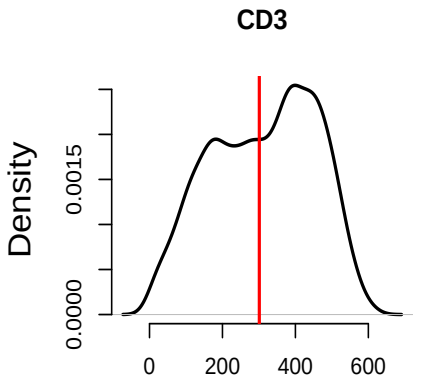
5 Running flowType

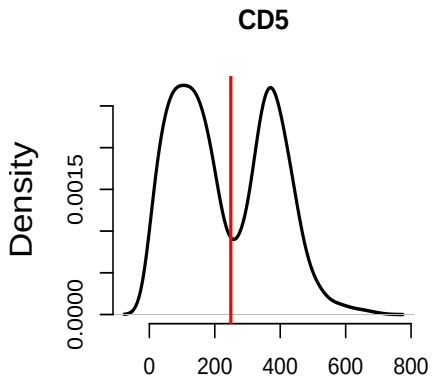
We will use the $PropMarkers$ and $MFIMarkers$ arrays for measuring cell proportions and MFIs, respectively. Cell proportion of a given cell population is the number of cells in that population divided by the total number of cells:

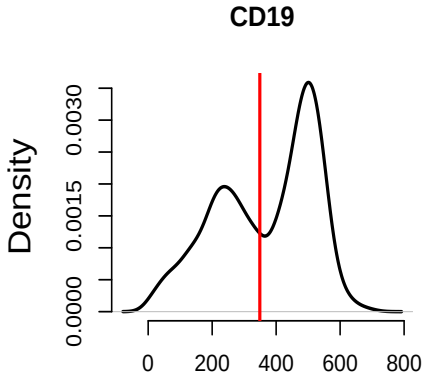
```
> PropMarkers <- 3:5
> MFIMarkers <- PropMarkers
> MarkerNames <- c('FS', 'SS','CD3','CD5','CD19')
> Res <- flowType(DLBCLExample, PropMarkers, MFIMarkers, 'kmeans', MarkerNames);
```

We can look at the single-dimensional partitions:

```
> plot(Res, DLBCLExample);
```

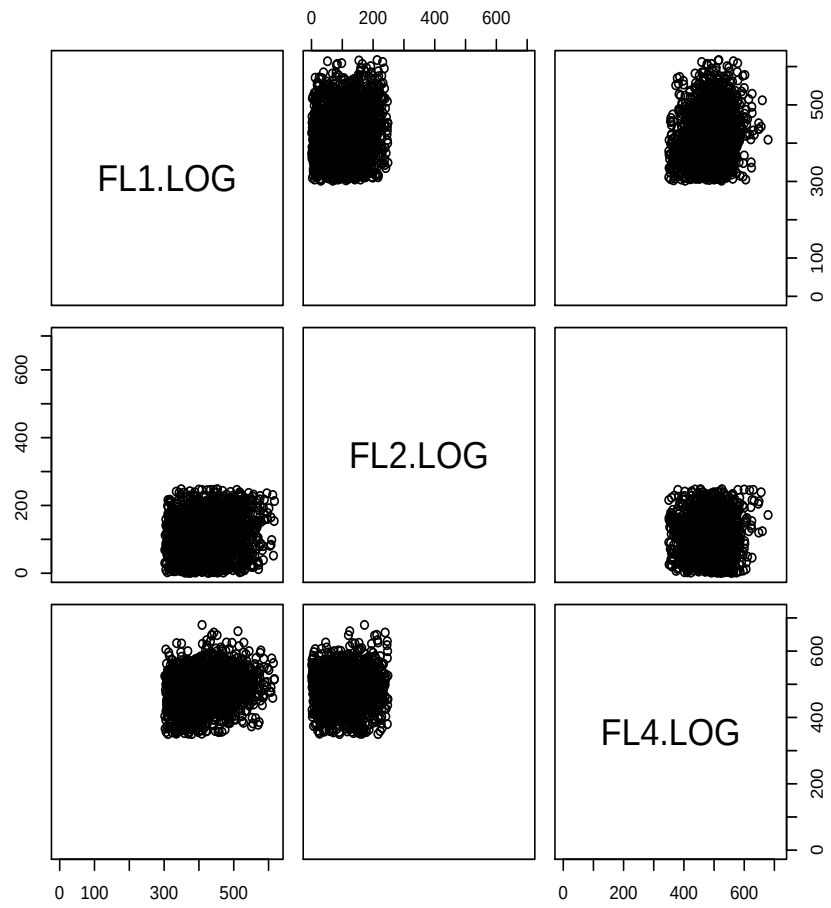






And we can plot a specific cell population:

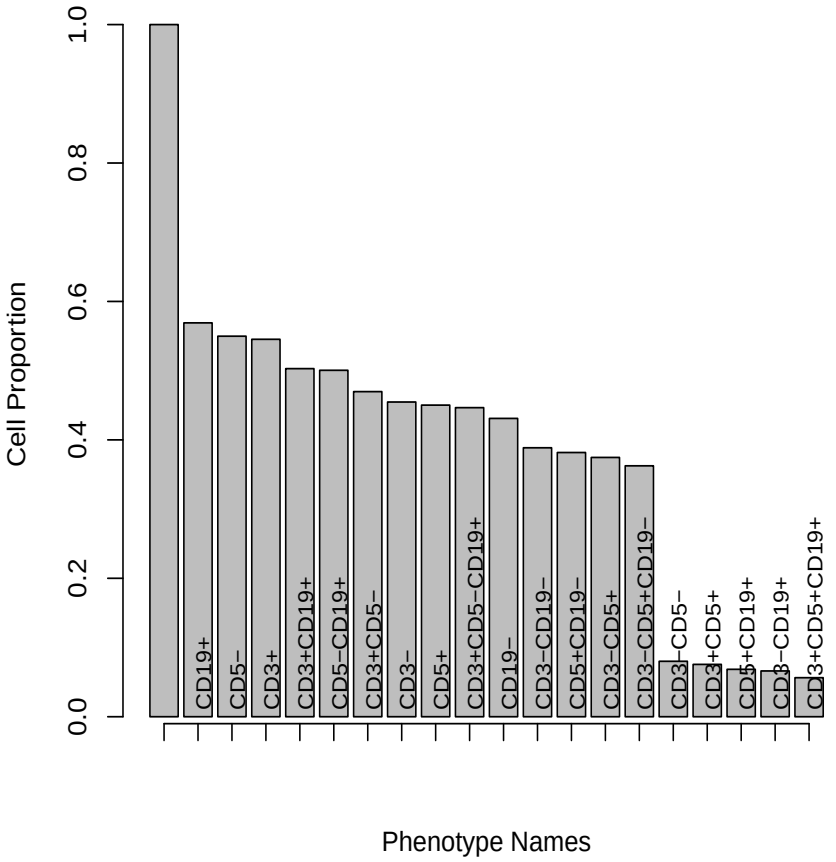
```
>   plot(Res, "CD3+CD5-CD19+", Frame=DLBCLExample);
```



Next we will plot the 20 largest phenotypes. The first phenotype includes all of the cells.

```
> MFIs=Res@MFIs;
> Proportions=Res@CellFreqs;
> Proportions <- Proportions / max(Proportions)
> names(Proportions) <- unlist(lapply(Res@PhenoCodes,
+                       function(x){return(decodePhenotype(
+                       x,Res@MarkerNames[PropMarkers],
+                       Res@PartitionsPerMarker))}))
> rownames(MFIs)=names(Proportions)
> index=order(Proportions,decreasing=TRUE)[1:20]
> bp=barplot(Proportions[index], axes=FALSE, names.arg=FALSE)
> text(bp+0.2, par("usr")[3]+0.02, srt = 90, adj = 0,
```

```
+      labels = names(Proportions[index]), xpd = TRUE, cex=0.8)
> axis(2);
> axis(1, at=bp, labels=FALSE);
> title(xlab='Phenotype Names', ylab='Cell Proportion')
```



These phenotypes can now be analyzed using a predictive model (*E.g.,* classification or regression).

6 Deciding on a cutoff

For high-dimensional data (anything more than 15 or so markers), it is often necessary for memory constraint reasons to only find phenotypes made up of less than a certain number of markers at once. We have provided a convenience function to determine how many phenotypes a certain number of markers and

thresholding strategy will yield for a given cutoff. Say you wanted to know for 34-colour mass cytometry data how many phenotypes you would get with a cutoff of 10 and using 2 partitions per marker:

```
> calcNumPops(rep(2,34), 10)
```

```
[1] 166562607753
```

You can also get an idea of where a good cutoff would be by plotting different cutoffs. If you decided that 10^6 were enough phenotypes, then you could read off that 4 would be a good choice:

```
> plot(log10(sapply(1:10, function(x){calcNumPops(rep(2,34), x)})), ylab='Cell types (log10)
```

7 Cross-sample Analysis

This document demonstrates the functionality of the flowType package for performing cross-sample analysis. The dataset used here is provided by the Scott laboratory of the Simon Fraser University and Spina laboratory of University of California, San Diego. This analysis is performed as a proof of principle and is not a complete analysis of this dataset. The data is transformed, compensated, and the lymphocytes are manually gated. The flowFrames have been downsampled to 1000 events.

The dataset consists of 19 HIV$^+$ and 13 normal subjects. Raw FCS files are available. The meta-data is stored in a matrix (which consists of FCS filename, tube number, and patient label). In this example, we are interested in the second tube only.

```
> library(flowType)
> data(HIVMetaData)
> HIVMetaData <- HIVMetaData[which(HIVMetaData[,'Tube']==2),];
```

We convert the subject labels so that HIV$^+$ and normal subjects are labeled 2 and 1, respectively.

```
> Labels=(HIVMetaData[,2]=='+')+1;
```

Load the data and run flowType:

```
> library(sfsmisc);
> library(flowCore);
> data(HIVData)
> PropMarkers <- 5:10
> MFIMarkers <- PropMarkers
> MarkerNames <- c('Time', 'FSC-A','FSC-H','SSC-A','IgG','CD38','CD19','CD3',
+                  'CD27','CD20', 'NA', 'NA')
> ResList <- fsApply(HIVData, 'flowType', PropMarkers, MFIMarkers, 'kmeans', MarkerNames);
```

Extract all cell proportions from the list of flowType results and normalize them by the total number of cells:

```
> All.Proportions <- matrix(0,3^length(PropMarkers),length(HIVMetaData[,1]))
> rownames(All.Proportions) <- unlist(lapply(ResList[[1]]@PhenoCodes,
+                         function(x){return(decodePhenotype(
+                         x,ResList[[1]]@MarkerNames[PropMarkers],
+                         ResList[[1]]@PartitionsPerMarker))}))
> for (i in 1:length(ResList)){
+   All.Proportions[,i] = ResList[[i]]@CellFreqs / ResList[[i]]@CellFreqs[
+                       which(rownames(All.Proportions)=='')]
+ }
```

We use a t-test to select the phenotypes that have a significantly different mean across the two groups of patients (FDR=0.05). Remember that in real world use-cases the assumptions of a t-test must be checked or a resampling-based alternative (e.g., a permutation test) should be used. P-value correction for multiple testing (e.g., bonferonni's method) and sensitivity analysis (e.g., bootstrapping) are also necessary.

```
> Pvals <- vector();
> EffectSize <- vector();
> for (i in 1:dim(All.Proportions)[1]){
+   if (length(which(All.Proportions[i,]!=1))==0){
+     Pvals[i]=1;
+     EffectSize[i]=0;
+     next;
+   }
+   temp=t.test(All.Proportions[i, Labels==1], All.Proportions[i, Labels==2])
+   Pvals[i] <- temp$p.value
+   EffectSize[i] <- abs(temp$statistic)
+ }
> Selected <- which(Pvals<0.05);
> print(length(Selected))

[1] 173
```

179 phenotypes have been selected. After P-value adjustment, only 5 of them remain in the list:

```
> Selected <- which(p.adjust(Pvals)<0.05);
> library(xtable)
> MyTable=cbind(rownames(All.Proportions)[Selected], format(Pvals[Selected],
+   digits=2), format(p.adjust(Pvals)[Selected],digits=3),
+   format(rowMeans(All.Proportions[Selected,]), digits=3))
> colnames(MyTable)=c('Phenotype', 'p-value', 'adjusted p-value', 'cell frequency')
> print(xtable(MyTable, caption='The selected phenotypes, their p-values, adjusted p-values,
+ caption.placement = "top");
```

Table 1: The selected phenotypes, their p-values, adjusted p-values, and cell frequencies

	Phenotype	p-value	adjusted p-value	cell frequen
IgG-CD27+	IgG-CD27+	4.8e-05	0.0322	0.248
IgG-CD38-CD27+	IgG-CD38-CD27+	4.8e-05	0.0321	0.239
IgG-CD19-CD27+	IgG-CD19-CD27+	4.8e-05	0.0322	0.247
IgG-CD27+CD20-	IgG-CD27+CD20-	4.8e-05	0.0322	0.247
IgG-CD38-CD19-CD27+	IgG-CD38-CD19-CD27+	4.9e-05	0.0324	0.237
IgG-CD38-CD27+CD20-	IgG-CD38-CD27+CD20-	4.8e-05	0.0322	0.237
IgG-CD19-CD27+CD20-	IgG-CD19-CD27+CD20-	4.7e-05	0.0318	0.246
IgG-CD38-CD19-CD27+CD20-	IgG-CD38-CD19-CD27+CD20-	4.8e-05	0.0322	0.237

References

[1] Nima Aghaeepour, Pratip K Chattopadhyay, Anuradha Ganesan, Kieran O'Neill, Habil Zare, Adrin Jalali, Holger H Hoos, Mario Roederer, and Ryan R Brinkman. Early immunologic correlates of hiv protection can be identified from computational analysis of complex multivariate t-cell flow cytometry assays. *Bioinformatics*, 28(7):1009–1016, 2012.

[2] Nima Aghaeepour, Greg Finak, Holger Hoos, Tim R Mosmann, Ryan Brinkman, Raphael Gottardo, Richard H Scheuermann, et al. Critical assessment of automated flow cytometry data analysis techniques. *Nature methods*, 10(3):228–238, 2013.

[3] Nima Aghaeepour, Adrin Jalali, Kieran O'Neill, Pratip K Chattopadhyay, Mario Roederer, Holger H Hoos, and Ryan R Brinkman. Rchyoptimyx: Cellular hierarchy optimization for flow cytometry. *Cytometry Part A*, 81(12):1022–1030, 2012.

[4] Fiona E Craig, Ryan R Brinkman, Stephen Ten Eyck, and Nima Aghaeepour. Computational analysis optimizes the flow cytometric evaluation for lymphoma. *Cytometry Part B: Clinical Cytometry*, 2013.

[5] Federica Villanova, Paola Di Meglio, Margaret Inokuma, Nima Aghaeepour, Esperanza Perucha, Jennifer Mollon, Laurel Nomura, Maria Hernandez-Fuentes, Andrew Cope, A Toby Prevost, et al. Integration of lyoplate based flow cytometry and computational analysis for standardized immunological biomarker discovery. *PloS one*, 8(7):e65485, 2013.