flowFDA using flowFP Probability Binning Fingerprinting

Lieven Clement & Olivier Thas

February 21, 2014

Contents

1 Introduction

The `flowFDA` package can be used for analysing flow cytometry (FC) experiments with functional model based clustering, functional principal component analysis, functional discriminant analysis and to compare multivariate flowcytometry fingerprints across treatments.

Flow cytometry (FC) can generate fast fingerprints by characterizing the multivariate distribution of cellular features of single cells. We developed a statistical pipeline for classifying samples and for inferring on distributional changes induced by experimental factors. Our method consists of 1) Creating a quantitative fingerprint from the multivariate distribution, 2) Extracting informative fingerprint features by discriminant analysis, 3) Permutation tests for assessing differences across treatment groups in the reduced feature space and 4) Interpreting these differences in terms of changes in the multivariate FC distribution.

In this vignette we illustrate the same functionalities as in the main `flowFDA` vignette, but we replace the pairwise bivariate kernel density fingerprint by the probability binning fingerprint

implemented in the Bioconductor `flowFP` package (Holyst and Rogers, 2009). In the example data five treatments on bottled Evian water were considered in the experiment: control (c), 3h heat treatment (h3), 24h heat treatment (h24). 3h nutrient treatment (n3) and 24h nutrient treatment (n24). More details on the experiment can be found in the main `flowFDA` vignette and in De Roy et al (2012).

2 Importing Data

The package builds upon the Bioconductor package `flowCore` to import raw flowcytometric data files in R. We first have to load the `flowFDA` package and read the flowset. For importing the data we refer to the main vignette of `flowFDA`. Again, we start from the example data that is available in the `flowFDAExampleData` package.

```
> source("http://www.bioconductor.org/biocLite.R")
> biocLite("flowFDAExampleData")
```

It can be loaded by

```
> library(flowFDAExampleData)
> library(flowFDA)
> data(fset)
> data(group)
> param=c("SS Log","FL 1 Log","FL 3 Log")
> nGroup=nlevels(group)
> nSamp=length(fset)
> groupLevels=levels(group)
```

3 Constructing a flowBasis object using Probability Binning

A p-dimensional quantitative fingerprint is derived from the multivariate FCM distribution using the recursive probability binning (PB) algorithm for flow cytometry data that is implemented in the Bioconductor package `flowFP`. At the first level of the algorithm the population of the cells are divided into two bins. Then, each of the two "parent" bins is divided into two "daughter" bins, and so forth. The final number of bins, n_{bin}, is determined by the number of recursive subdivisions I, such that $n_{\text{bin}} = 2^I$. Note, that the algorithm constructs the bins in such a way that they contain a nearly equal number of cells from the pooled FC sample that were used to build the PB model. This provides an efficient representation of the structure in the multivariate data space by hyper-rectangular regions (bins) of varying size and shape. We pool the data of all samples together for constructing the PB model. The obtained PB model is then applied to each individual sample, which results in a feature vector of counts for each bin of the model. The bin counts for each sample are normalized by the total number of cells in the sample. The normalized bin counts are also referred to as the fingerprint. More details on the PB algorithm can be found in Roederer et al. (2001). Note, that the PB method has to be reconstructed if new samples are available. Hence, the PB fingerprint for a particular sample depends on the other samples that were analysed, simultaneously. This bivariate kernel density fingerprint implemented in the

main `flowFDA` vignette, however, does not depend upon other samples and do not have to be reconstructed when samples are added or removed from the analysis. With the PB approach all functionalities and interpretation plots of the `flowFDA` package can also be used.

The code below can be used to set up a flowBasis object using a PB fingerprint.

```
> fbasisPb=flowBasis(fset,param,nbin=128,probBin=TRUE)
> fbasisPb


flowBasis object
Probability Binning using 128 bins.

bivariate densities for channels
     [,1]       [,2]
[1,] "SS Log"   "FL 1 Log"
[2,] "SS Log"   "FL 3 Log"
[3,] "FL 1 Log" "FL 3 Log"
```

4 Data exploration

The PB fingerprints can be explored graphically. An example for third flowset can be generated using the code below. Bivariate projections for the probability binning approach used in De Roy et al. (2012) are given in 1. Note, that the bivariate kernel density (KD) approach provides a better interpretation than the PB approach. The multivariate bins are projected on all bivariate combinations of the flow channels that were used to build the PB model. Hence, they might overlap each other in the projection, which obscures the interpretation. Moreover, each bin was constructed so as to contain approximately an equal amount of cells in the pooled sample used to construct the model. Large bins, thus, coincide with a low density of cells. Hence, low density regions are often overemphasised in the graphs.

```
> par(mfrow=c(2,2))
> plot(fbasisPb,ask=FALSE,samples=3)
```

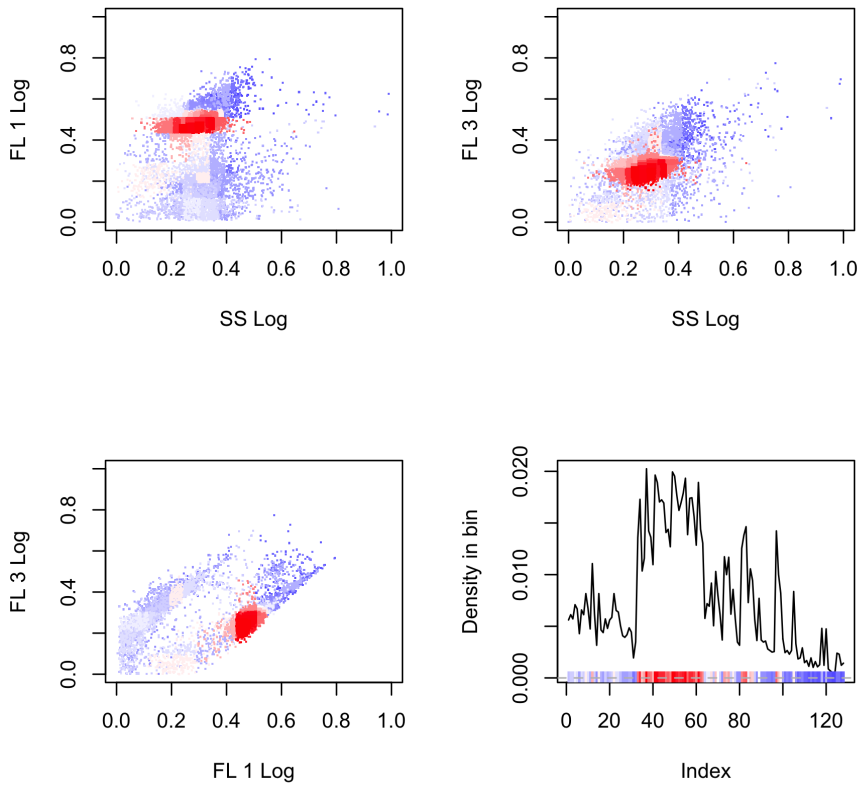

Figure 1: Plot of bivariate projection of probability binning fingerprint for third flow of the flowSet. In white regions cells are absent, the blue-to-red colour gradient corresponds to low-to-high density of cells. The bottom-right panel shows the fingerprint: normalized cell count for each of the 128 bins.

The fingerprints can also be averaged over several flows, i.e. for the flows belonging to the same group. An example of the graphical interpretation of the averaged fingerprints for the control group (c) and 24h nutrient treatment group (n24) are given in Figure 2.

```
> par(mfrow=c(2,4))
> plot(fbasisPb,ask=FALSE,samples=group==groupLevels[1],main=groupLevels[1])
> plot(fbasisPb,ask=FALSE,samples=group==groupLevels[4],main=groupLevels[4])
```

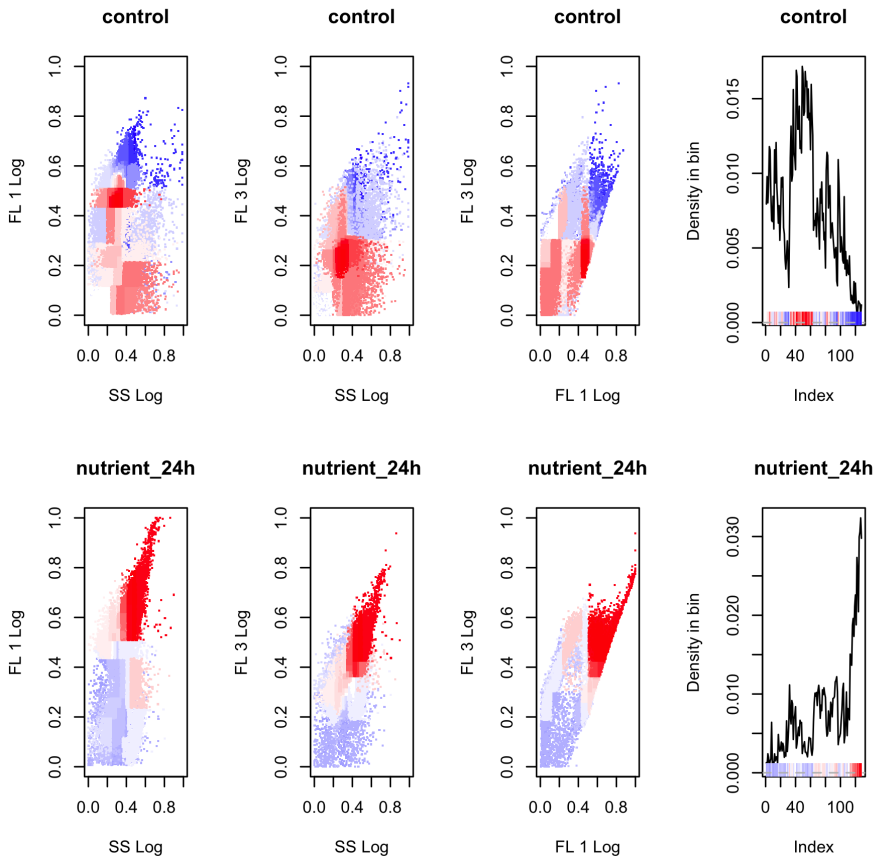

Figure 2: Plot of average PB fingerprint for flows belonging to the control (c) group and the 24h nutrient (n24) group using probability binning. c-samples on average have more cells at low FL1 and FL3 levels whereas n24-samples on average have a higher density at high SS, FL1 and FL3 intensities, which corresponds to larger cells with higher nucleic acids content and intact membranes. The difference can also be observed in the fingerprint plot right panels: c-samples have a larger density of cells for bins between 40-60 and less cells in bins with higher bin-numbers whereas an opposite pattern can be observed for n24-samples.

The average contrast of the fingerprints belonging to the 24h nutrient treatment group (n24) and control group (c) is visualised in Figure 3. A negative contrast is represented by a light-to-dark blue colour and a positive contrast is indicated with a light-to-dark red colour scheme. After n24 treatment, a lower density of cells are observed at low SS, FL1 and a FL3 values as compared to the c-group, i.e. blue region with a negative contrast. A part of the mass of the distribution shifted to higher SS, FL1 and FL3 values, red regions. The fingerprint contrast is also plotted and will be used for the interpretation in the downstream analysis: it shows that n24-samples on average have a lower cell density in bins with index 40-60 and have a higher density at bins with an index above 100 than c-samples. From the colours above the bin indices, it can be seen that the higher bin-index correspond to bins with higher SS, FL1 and FL3 values.

```
> par(mfrow=c(2,2))
> L=rep(0,length(group))
> L[group==groupLevels[1]]=-1/sum(group==groupLevels[1])
> L[group==groupLevels[4]]=1/sum(group==groupLevels[4])
> par(mfrow=c(2,2))
> plot(fbasisPb,L=L,set=which(L!=0),ask=FALSE,main=paste(groupLevels[4],"-",groupLevels[1],sep=""))
```

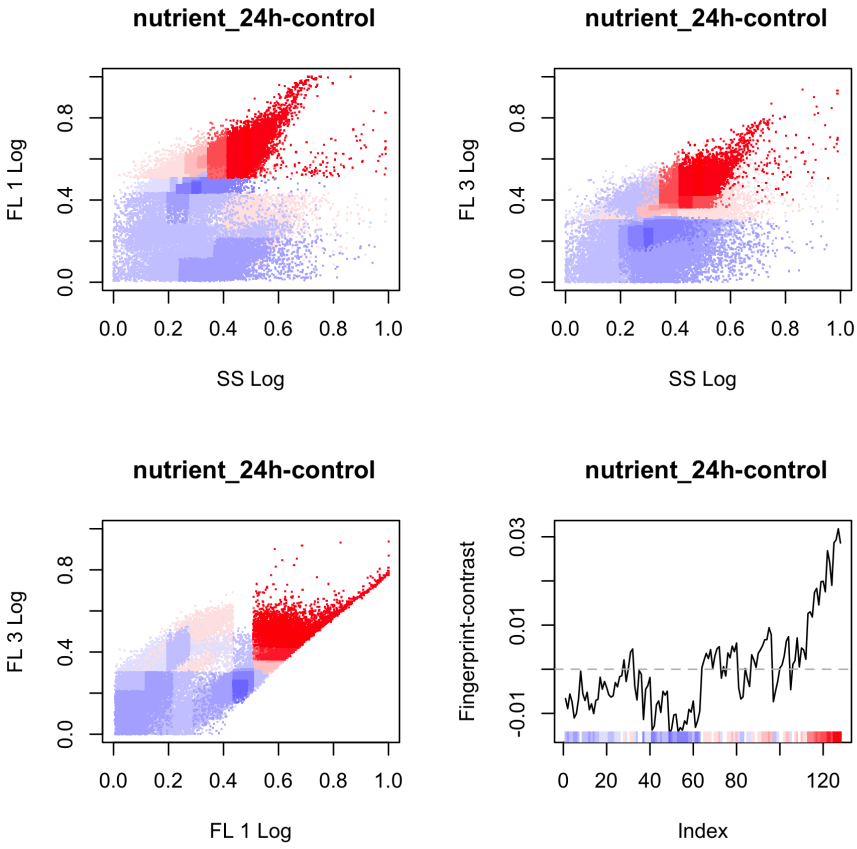

Figure 3: Plot of contrast between probability binning fingerprints belonging to the 24h nutrient treatment (n24) group and the control (c) group. n24-samples on average have a lower density of cells are at low SS, FL1 and FL3 blue region with a negative contrast, as compared to the control treatment. A part of the mass of the distribution shifted to higher SS, FL1 and FL3 red regions. The contrast of the average fingerprint with the 128 probability bins is given in the fourth panel. The fingerprint-contrast will be useful for the interpretation of the downstream analysis.

The fingerprints can also be explored by using principal component analysis and model based clustering. Details on the methods can be found in the main `flowFDA` vignette. We build upon the `mclust` package that relies on Gaussian mixture models. The model based clustering in this example is performed by using the first 9 principal components. They capture 95% of the variability in the original fingerprints. Scores for the first 2 principal components are given in Figure 4.

```
> #construct flowPca object with probability binning basis
> fPcaPb=flowPca(fbasisPb)
> #perform model based clustering,
> #use n PCs so as to capture at least 95 % of the variability
> nPca(fPcaPb)<-.95
> nPca(fPcaPb) #number of PCs used for model based clustering
```

```
[1] 9
```

```
> setClust(fPcaPb)<-Mclust(getPcaScore(fPcaPb,nPca(fPcaPb))) #Model based clustering
> cbind(as.character(getClustClass(fPcaPb)),as.character(group)) # cluster class labels and real gro
```

```
      [,1] [,2]
 [1,] "1"  "control"
 [2,] "1"  "control"
 [3,] "1"  "control"
 [4,] "1"  "control"
 [5,] "1"  "control"
 [6,] "1"  "control"
 [7,] "1"  "heat_24h"
 [8,] "1"  "heat_24h"
 [9,] "1"  "heat_24h"
[10,] "1"  "heat_24h"
[11,] "1"  "heat_24h"
[12,] "1"  "heat_24h"
[13,] "2"  "heat_3h"
[14,] "2"  "heat_3h"
[15,] "2"  "heat_3h"
[16,] "2"  "heat_3h"
[17,] "2"  "heat_3h"
[18,] "2"  "heat_3h"
[19,] "2"  "nutrient_24h"
[20,] "3"  "nutrient_24h"
[21,] "3"  "nutrient_24h"
[22,] "3"  "nutrient_24h"
[23,] "3"  "nutrient_24h"
[24,] "3"  "nutrient_24h"
[25,] "4"  "nutrient_3h"
[26,] "4"  "nutrient_3h"
[27,] "4"  "nutrient_3h"
[28,] "4"  "nutrient_3h"
[29,] "4"  "nutrient_3h"
[30,] "4"  "nutrient_3h"
```

```
> par(mfrow=c(1,2))
> plot(fPcaPb,groups=getClustClass(fPcaPb),main="Prob. Bin. (Clustering)")
> plot(fPcaPb,groups=group,main="Prob. Bin. (Treatment)")
> legend("topleft",legend=c("c","h24","h3","n24","n3"),pch=1:5,col=1:5)
```

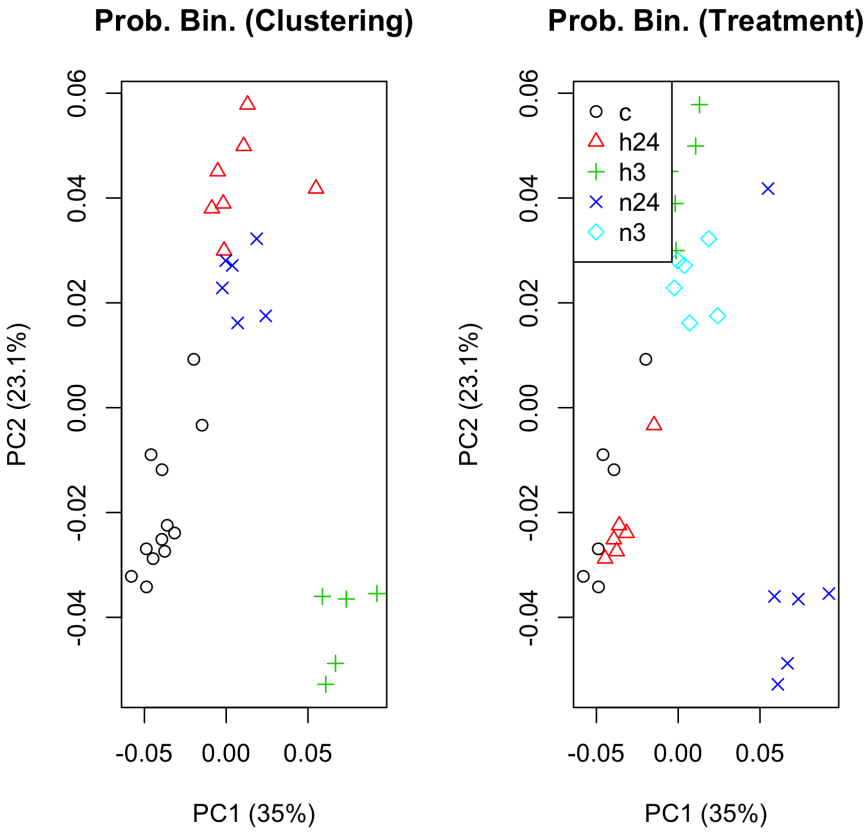

Figure 4: Scores for the first two principal components. In the left panel the samples are classified using model based clustering and in the right panel the samples are labelled according to the actual experimental factor.

The scores on the PC's can be interpreted in terms of the original bivariate distributions. They consist of linear combinations of the contrast between the sample- and average fingerprint over all samples. Some regions will contribute negatively to the PC score and others positively. The colours in the plot indicates the contribution to the PC score, which consists of the sum over all $n_{\text{bin}} = 128$ bins of the PB fingerprint. The bottom-right panel shows the contrast of the fingerprints along with bin-colours according to their contribution to the PC-scores. The interpretation of the score for sample 3 is given in the caption of Figure 5.

5 Discriminant Analysis and Classification

Here, we construct a `flowDa` object for supervised discrimination method between groups. Again, regularisation is provided by performing PCA first and adopting DA on the first few PC's that explain more that 95% of the variability in the fingerprint. In our application this corresponds to 9 PCs. Hence, we reduced the dimensionality of the problem from 128 bin-features to 6. More details on the procedure can be found in the main `flowFDA` vignette.

```
> #####Discriminant analysis for prob. bin.
> fDaPb=flowDa(fbasisPb,groups= group, nPca=.95)
> fDaPb


flowDa object
Probability Binning using 128 bins.

channels
     [,1]       [,2]
[1,] "SS Log"   "FL 1 Log"
[2,] "SS Log"   "FL 3 Log"
[3,] "FL 1 Log" "FL 3 Log"
```

```
> intSamples=3 #for the group average of first group set intSamples=which(group=groupLevels[1])
> layout(matrix(c(1,2,3,1,4,5),nrow=2,byrow=TRUE))
> par(pty="s")
> plot(fPcaPb,groups=group,main="Actual grouping")
> pcX=mean(getPca(fPcaPb)$x[intSamples,1])
> pcY=mean(getPca(fPcaPb)$x[intSamples,2])
> arrows(x0=pcX,x1=pcX,y0=-4,y1=pcY)
> intSamples=3 #for the group average of first group set intSamples=which(group=groupLevels[1])
>  #PCA is done after centering
>  # interpretation in terms of contrast to average bivariate density
>  #contrast between average bivariate density of intSamples vs overall average
> L=rep(-1/nSamp,nSamp)
> L[intSamples]=L[intSamples]+1/length(intSamples)
> plot(fPcaPb,fBasis=fbasisPb,disc=1,plotType="pcaCont",L=L,ask=FALSE,main="PC 1")
```

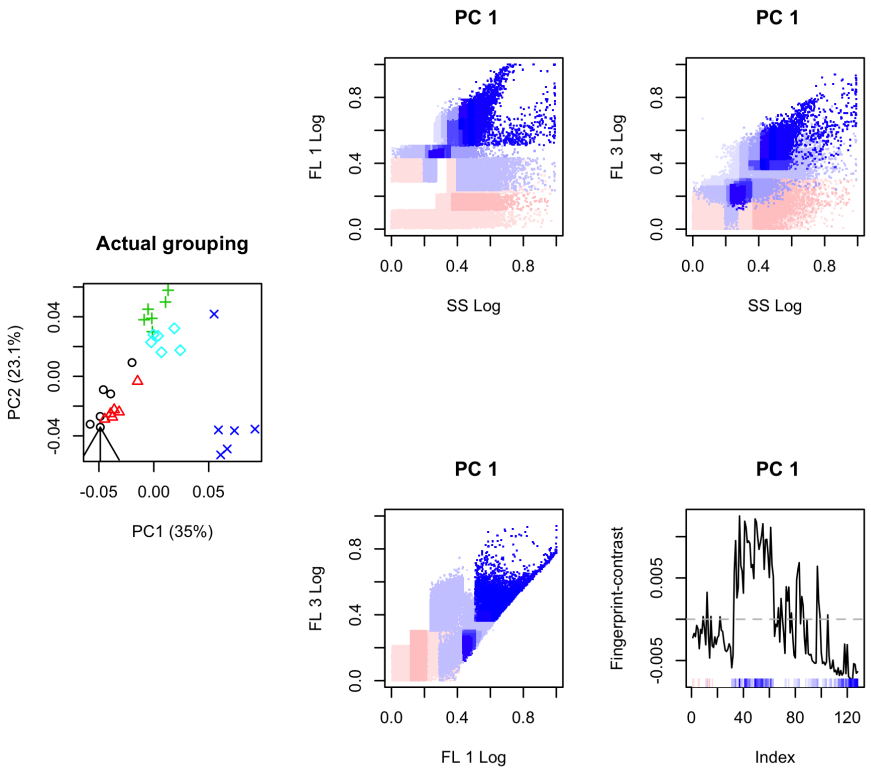

Figure 5: Scores for first two principal components along with an interpretation plot for the score of PC1 for sample 3 (indicated with arrow). The colour in the interpretation plot indicates negative (blue) or positive (red) or small (uncoloured) contributions. Overall the colour is neutral or blue, indicating that the score on the first PC will be negative. For a better interpretation, the fingerprint contrast is given (bottom right panel). They indicate bins for which the density of sample 3 is higher (positive contrast) or lower (negative fingerprint contrast) than the average fingerprint over all samples. The bin colour reflects the contribution to the first PC score. The low score on the first PC, thus, originates from a distribution of sample 3 that is higher than the average distribution around an SS of 0.3, FL1 of 0.5 and a FL3 of 0.3 (dark blue colour in bivariate projection plots and in bins with index 40-60 in fingerprint contrast) and because it has a lower density at higher SS, FL1 and FL3 (dark blue colour in bivariate projection plots and at bin indexes 120-128 in fingerprint contrast)

```
> par(mfrow=c(1,2))
> plot(fDaPb,groups=group,main="Prob. Bin. PCA",plotType="pcaPlot")
> plot(fDaPb,main="Prob. Bin. DA")
> legend("topleft",legend=c("c","h24","h3","n24","n3"),pch=1:5,col=1:5)
```

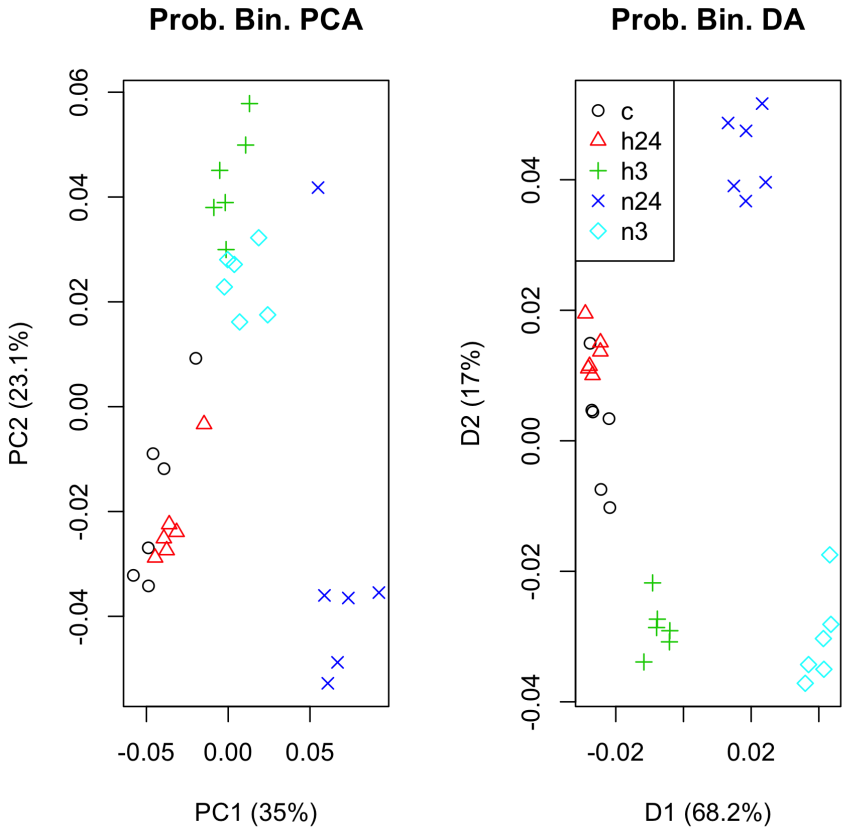

Figure 6: Comparison between clustering of treatments in the PCA space (unsupervised) and in the DA space (supervised).

6 Test for differences in the discriminant space

Pairwise permutation tests are used to assess differences between treatments. The tests are performed for each discriminant dimension separately so as to retain the interpretation feature. Significant tests can be interpreted in terms of the features of the original fingerprint. More details on the procedure can be found in the main `flowFDA` vignette.

```
> nPerm=100
> #Only 100 permutations are used
> #so as to restrict the computational burden when generating the vignette
> #nPerm=10000
> disc=1:2 #Test only in the space of first 2 discriminants
> fDaPb=flowDaTest(fDaPb,disc=disc,nPerm)
```

```
progress
```

Note, that only 100 permutations are adopted to reduce the computational burden when building the package on the Bioconductor suite. Hence, users have to uncomment the `#nPerm=10000` line for obtaining more reliable permutation p-values based on 10000 permutations.

Next, we will adjust the p-values for multiple testing (10 x ndisc tests).

```
> adjustedPvalues=pAdjustMx(getMpc(fDaPb)$pValuePerm)
> adjustedPvalues
```

```
                           D1   D2
heat_24h-control         1.00 0.45
heat_3h-control          0.00 0.00
nutrient_24h-control     0.00 0.00
nutrient_3h-control      0.00 0.00
heat_3h-heat_24h         0.08 0.00
nutrient_24h-heat_24h    0.05 0.00
nutrient_3h-heat_24h     0.00 0.00
nutrient_24h-heat_3h     0.00 0.00
nutrient_3h-heat_3h      0.00 1.00
nutrient_3h-nutrient_24h 0.00 0.00
```

Note, again that the adjusted p-values in this vignette are based on very few permutations to reduce the computational burden. Significant differences can be interpreted in the original space. This is illustrated for the first discriminant in Figure 7. The interpretation is given in the caption of the plot.

```
> groupLevels=levels(group)
> nSamp=nSet(fDaPb)
> L<-rep(0,nSamp)
> L[group==groupLevels[4]]<-1/sum(group==groupLevels[4])
> L[group==groupLevels[1]]<--1/sum(group==groupLevels[1])
> layout(matrix(c(1,2,3,1,4,5),nrow=2,byrow=TRUE))
> par(pty="s")
> plot(fDaPb)
> disc=1
> plot(fDaPb,fBasis=fbasisPb,L=L,ask=FALSE,plotType="discCont",disc=disc)
```

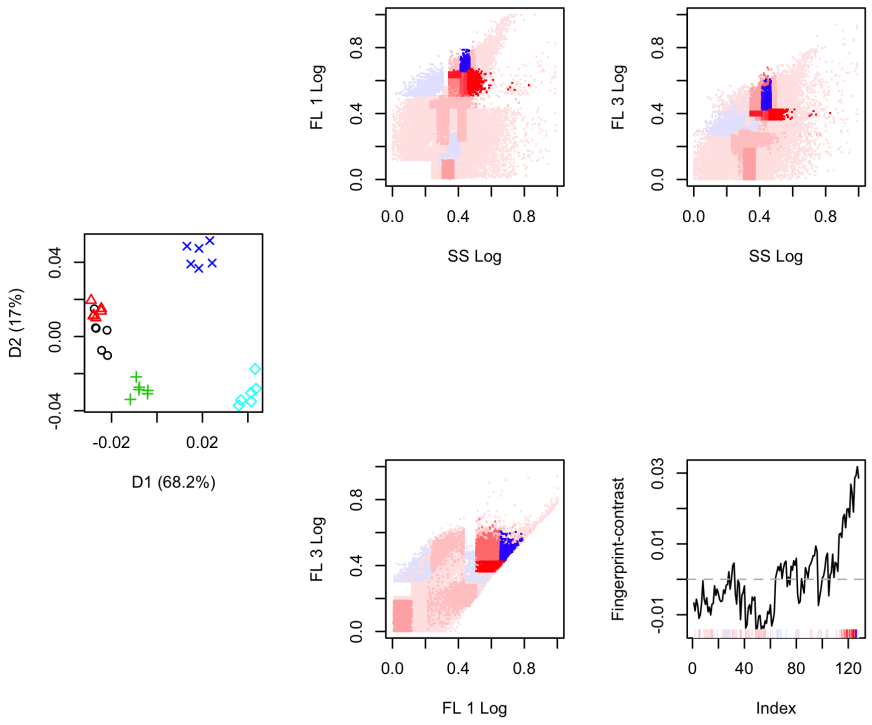

Figure 7: Interpretation of 24h nutrient and control (n24-c) contrast on the first discriminant (D1). The colours in the interpretation plots indicate the contribution of the region to the n24-c contrast on D1. Negative contributions are indicated from light-to-dark blue, positive contributions are coloured in light-to-dark red. The lower-right panel displays the original fingerprint-contrast. Note, that "large" contributions to the D1 score are linked to more cells (positive contrast) in the bins with index 118-128. The majority of these bins are linked to a positive contribution to the D1-contrast (except for bin-index 126 we observe a higher density combined with negative contribution (blue) to he n24-c D1 contrast). Overall, the D1-contrast is positive because of a density shift in n24 samples towards higher SS, FL1 and FL3 values, which corresponds to larger cells with higher nucleic acid content and intact membranes. Note, however, that the interpretation with the PB-approach is less straightforward than when using the default bivariate kernel density fingerprint (main `flowFDA` vignette).

7 Generating plots for multiple contrasts

Again, plots for multiple contrasts can be generated when using PB fingerprinting. We refer to the main `flowFDA` vignette for detailed code. The flowBasis, flowPca and flowDa object have to be replaced by objects generated with a PB fingerprint basis.

References

De Roy, K., Clement, L., Thas, O., Wang, Y., and Boon, N. (2012). Flow cytometry for fast microbial community fingerprinting. Water Research, 46 (3), 907-919.

Ellis, B., Haaland, P., Hahne, F., Le Meur, N. and Gopalakrishnan, N. (2009). flowCore: flowCore: Basic structures for flow cytometry data. R package version 1.26.3.

Holyst, H. and Rogers, W. (2009). flowFP: Fingerprinting for Flow Cytometry. R package version 1.18.0.

Roederer, M., Moore, W., Treister, A., Hardy, R. R. and Herzenberg, L. A. (2001). Probability binning comparison: a metric for quantitating multivariate distribution differences. Cytometry, 45(1):47-55.