High-throughput flow cytometry data and how to load, transform and visualise data and gate populations in Bioconductor (R)

Ulrike Naumann

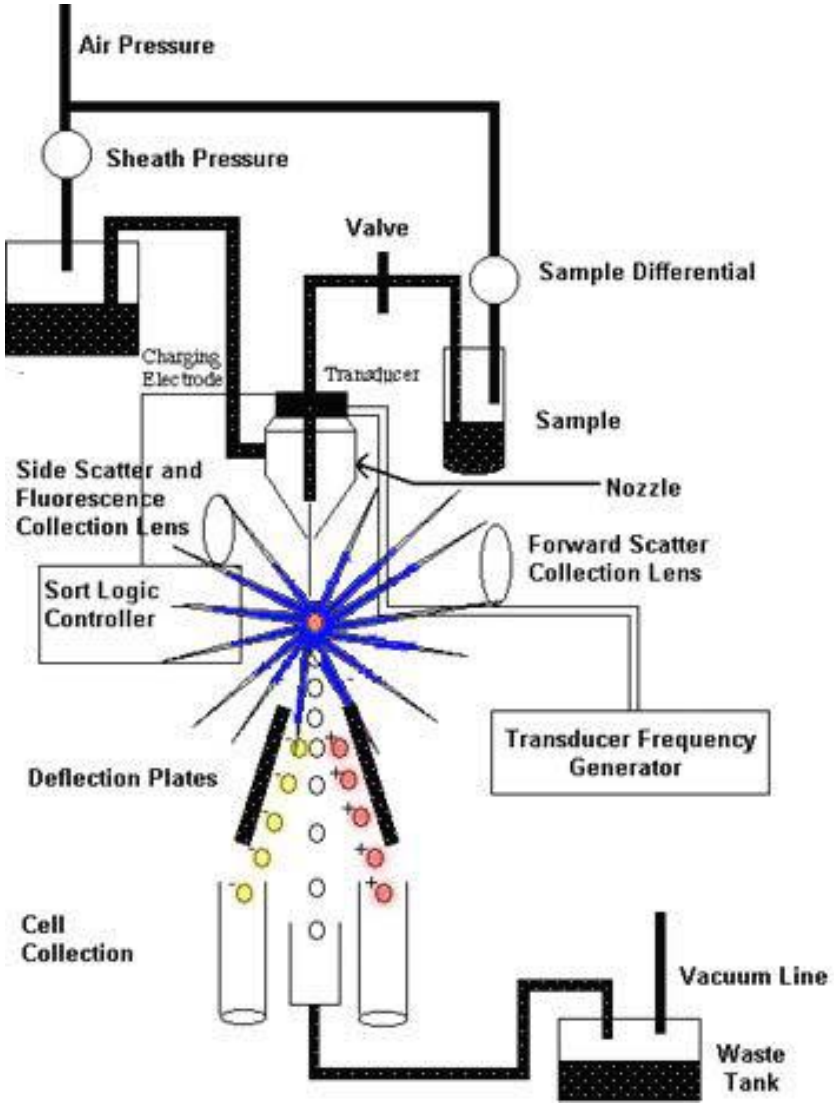
Ulrike.Naumann1@gmail.com

Content

- What is flow cytometry?
- Bioconductor
- Software packages in Bioconductor for flow cytometry data
- Read data into R
- Transform the data, take out extremes
- Steps in the analysis process
- Example of output
- How to summarise the output visually and interpret it
- References

Flow Cytometry (FCM)

- Flow Cytometry is a technique for counting, examining and sorting microscopic particles suspended in a stream of fluid.



Flow Cytometry

- Cells have been stained with monoclonoal antibodies.
- Several Detectors are aimed at the point where the stream of fluid passes through a light beam and pick up the reflected/scattered light
- The detectors differ in what they measure:
  FSC (Forward Scatter) detector measures correlates with the cell volume and SSC (Side Scatter) measures correlate with the inner complexity of the particle

High-throughput flow cytometry

- 21st Century technology
- large numbers of flow cytometric samples can be processed and analysed in a short period of time
- Challenge: high-dimensional complex data. Manual analysis time-consuming, subjective and error-prone
  → Development of automatic gating methods

Software for Analysis of Flow Cytometry Data

- Beckman Coulter Kaluza Software
- Cellular Symphony Flow Cytometry Software
- Millipore Flow Cytometry Software
- SPICE Data Mining & Visualization Software
- WEASEL Flow Cytometry Software
- …

Today I'm only going to explain briefly how Flow Cytometry Data can be processed using **Bioconductor**

Bioconductor

- provides tools for the analysis and comprehension of high-throughput genomic data.
- uses the R statistical programming language, and is open source and open development

Installation of basic Bioconductor packages in R:

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

> biocLite()

> biocLite("flowCore ")

> biocLite("curvHDR")

- install curvHDR package over the menue. This will also install the following packages: ‘abind’ ‘akima’ ‘magic’ locfit’ ‘ash’ ‘mvtnorm’ ‘feature’ ‘geometry’ ‘hdrcde’ ‘ks’ ‘misc3d’ ‘ptinpoly ‘rgl ‘

Software packages in Bioconductor for flow cytometry data

- These packages use standard FCS files, including infrastructure, utilities, visualization and semi-automated gating methods for the analysis of flow cytometry data.
- flowCore: Basic structures for flow cytometry data
  flowViz: Visualization of flow cytometry
  flowQ: Quality control for flow cytometry
  flowStats: Statistical methods for the analysis of flow cytometry data
  flowUtils: Utilities for flow cytometry
  flowFP: Fingerprint generation of flow cytometry data, used to facilitate the application of machine learning and data mining tools for flow cytometry.
  flowTrans: Profile maximum likelihood estimation of parameters for flow cytometry data transformations.
  iFlow: Tool to explore and visualize flow cytometry

Algorithms for clustering flow cytometry data are found in these packages:

- flowClust: Robust model-based clustering using a t-mixture model with Box-Cox transformation.
- flowMeans: Identifies cell populations in Flow Cytometry data using non-parametric clustering and segmented-regression-based change point detection.
- flowMerge: Merging of mixture components for model-based automated gating of flow cytometry data using the flowClust framework.
- SamSPECTRAL: Given a matrix of coordinates as input, SamSPECTRAL first builds the communities to sample the data points.

- A typical workflow using the packages flowCore, flowViz, flowQ and flowStats is described in detail in flowWorkFlow.pdf. The data files used in the workflow can be downloaded from here.

Load the Data

Loading data is a complex step, that can take a long time.

Flow cytometry experiments typically involve data from

- several patients
- several time points
- a number of antibody stain combinations

1. Understand the structure of the data

2. Read in the data. We decided to create for each patient a separate R workspace file (.Rdata)

Format, and organising the data to read it into R

Time of this presentation is too short to give a detailed account so I will go only into some issues we encountered.

1. The available data is a time series. The number of days differ for each participant.
2. The days available differ for each study participant!
3. 10 different Antibody-combinations were used in our data.

Transform the data, take out extremes

- Determine minima and maxima of each flow cytometry sample. Remove recordings that accumulate on the boundaries (usually upper boundaries)
- possibly transform samples to reduce their skewness $x_{new} = \sinh^{-1}(x) = \log\left(x + \sqrt{x^2 + 1}\right)$

Gating Flow Cytometry Data with curvHDR - Steps in the analysis process

curvHDR – package in R/Bioconductor for gating FCM data and displaying gates

Functions: curvHDRfilter and plot

The most important parameters of the function curvHDRfilter are the dataset x, HDRlevel, growthFac and signifLevel

x	a numerical vector or a matrix or data frame having 1-3 columns.
HDRlevel	number between 0 and 1 corresponding to the level of the highest density region within each high curvature region.
growthFac	growth factor parameter. High curvature regions are grown to have 'volume' growthFac times larger than the original region.
signifLevel	number between 0 and 1 corresponding to the significance level for curve region determination.

Steps in the analysis process

Defaults:

HDRlevel = 0.1

growthFac = 5^(d/2) where d is the dimension of the input data;

signifLevel = 0.05

Start with trial values of signifLevel and growthFac and HDRlevel, set at defaults.

Example:

```
xBiva <- cbind(c(rnorm(1000,-2),rnorm(1000,2)),
               c(rnorm(1000,-2),rnorm(1000,2)))
gate2a <- curvHDRfilter(xBiva)
```

Example of output

Output:

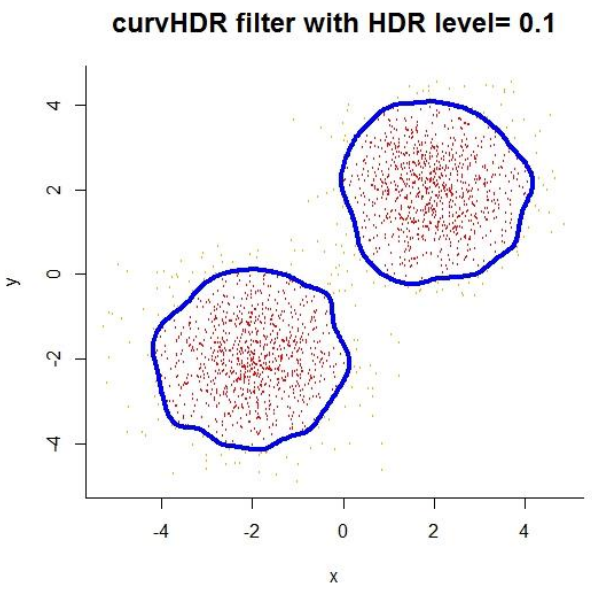
data	the input data (for use in plotting).
insideFilter	logical variable indicating the rows of the input data matrix corresponding to points inside the curvHDR filter.
polys	the curvHDR filter. Depending on the dimension d this is a list of intervals (d=1), polygons (d=2) or polyhedra (d=3).
HDRlevel	highest density region level

xBiva[gate2a$insideFilter,] contains the data inside of the gate.

Example of Output

We can visualise the output of curvHDRfilter with the plot command.

plot(gate2a)



In our actual dataset, we combined respectively 3 gates to obtain our final combined gates – a 2-dim gate on (FSC,SSC), a 1-dim gate Ab3, a 2-dim gate on (Ab1,Ab2). Our intention was to match the results of an already existing analysis (Brinkman 2007).

We selectively tested out a number of choices for the parameters HDRlevel, growthFac and signifLevel. Applying the same set of these 3 parameters globally for the entire dataset provided good results.

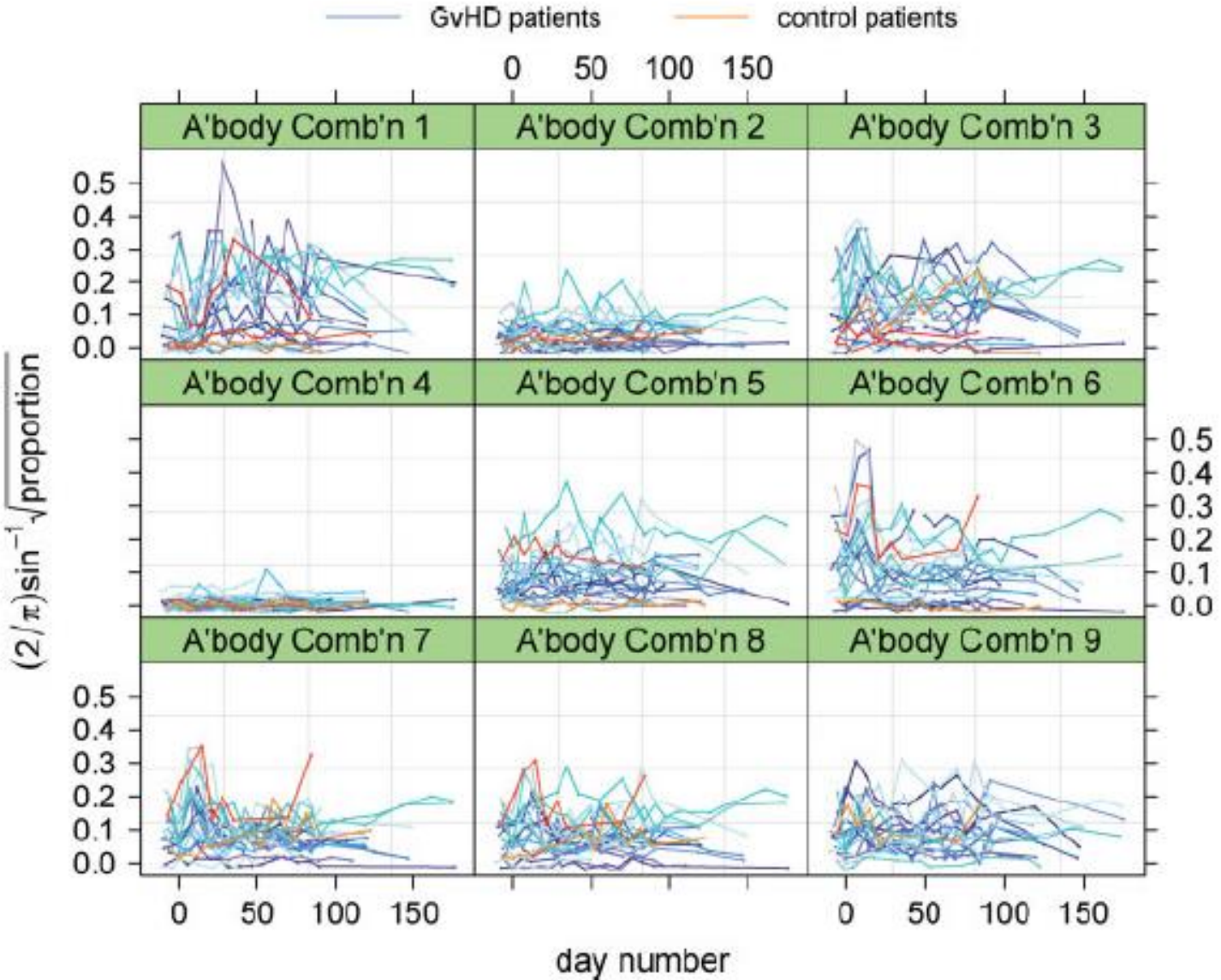
How to summarise the output visually and interpret it

We want obtain for each patient and each patient day and each anti-body combination, a summary of the data that passed through the gate.

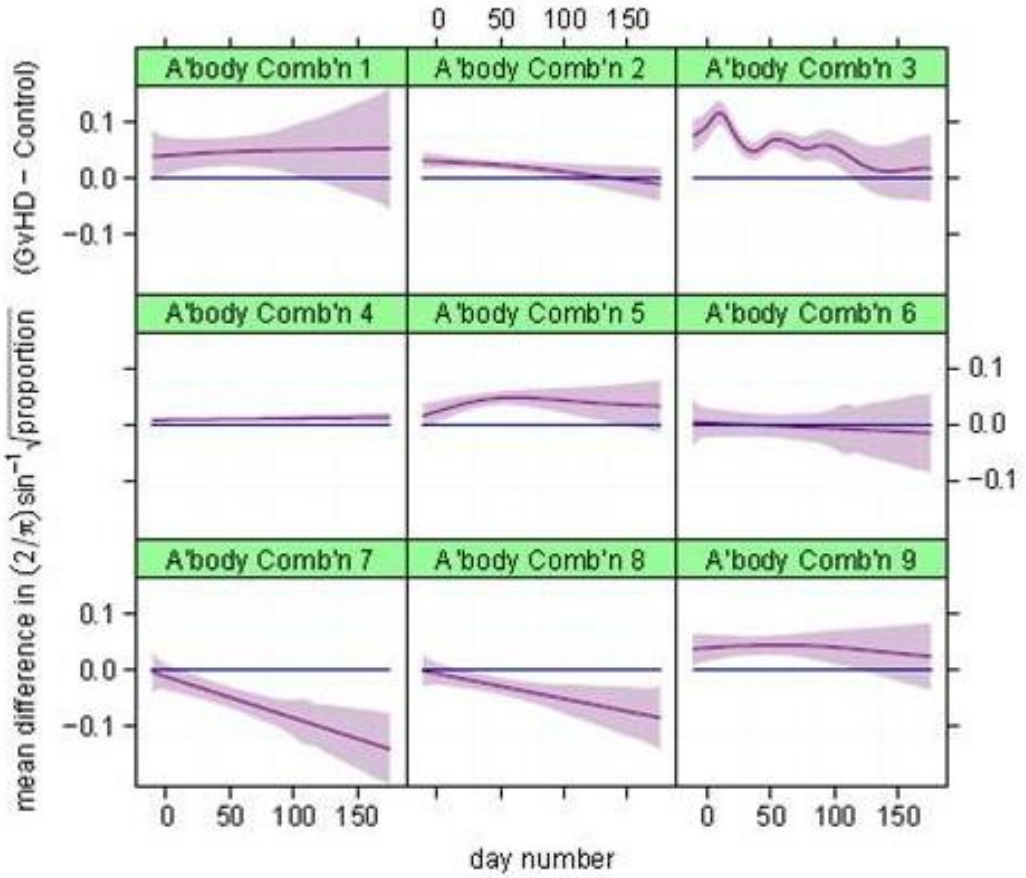
In curvHDR:

- get the proportion of gated cells for each patient, each day and each AB combination
  example: length(xBiva[gate2a$insideFilter,1])
- Apply variance stabilising transformation $y_{new} = \left(2/\pi\right)\sin^{-1}\left(\sqrt{y}\right)$ on the proportion data

Summary using Lattice graphics - ggplot



Findings – Signatures for Graft-versus-Host Disease



- Estimated contrast curves (cellular signatures) arising from fitting the longitudinal data.
- The shading around each curve corresponds to approximate point-wise 95% confidence intervals.

References

Brinkman, R.R., Gasparetto, M., Lee, S.-J.J., Ribickas, A.J., Perkins, J., Janssen, W., Smiley, R. and Smith, C. (2007). High-content flow cytometry and temporal data analysis for defining a cellular signature of graft-versus-host disease. Biology of Blood and Marrow Transplantation, 13, 691–700.

Ellis, B., Haaland, P., Hahne, F., Le Meur, F. & Gopalakrishnan, N. (2009). flowCore 1.10.0. Basic structures for flow cytometry data Bioconductor package. http://www.bioconductor.org.

Shapiro, H.M. (2003). Practical Flow Cytometry, 4th Edition. John Wiley & Sons, New York.

Lo, K., Brinkman, R.R. & Gottardo, R. (2008). Automatic gating of flow cytometry data via robust model-based clustering. Cytometry Part A, 321–332

References

Naumann U. & Wand M.P. (2009). Automation in high-content flow cytometry screening. *Cytometry Part A*, (2009), 75A, 789-797.

Naumann U., Luta G. & Wand M.P. (2009). The curvHDR method for gating flow cytometry samples. *BMC Bioinformatics*, (2010), 11:44, 1-13.

Gentleman R., Carey V., Huber W., Irizarry R., Dutoit S., (2005). Bioinformatics and Computational Biology Solutions Using R and Bioconductor

Gentleman R., (2008). R Programming for Bioinformatics

Aghaeepour N., Finak G., The FlowCAP Consortium, The DREAM Consortium, Hoos H., Mosmann T. R., Brinkman R., Gottardo R. & Scheuermann R. H., (2013). Critical assessment of automated flow cytometry data analysis techniques. Nature Methods (Advanced Online Publication)