BIOINFORMATICS APPLICATIONS NOTE Vol. 31 no. 4 2015, pages 606–607
doi:10.1093/bioinformatics/btu677

Systems biology

Advance Access publication October 16, 2014

flowDensity: reproducing manual gating of flow cytometry data by automated density-based cell population identification

Mehrnoush Malek[1,†], Mohammad Jafar Taghiyar[1,†], Lauren Chong[3], Greg Finak[2], Raphael Gottardo[2] and Ryan R. Brinkman[1,*]

[1]Terry Fox Laboratory, BC Cancer Agency Research Centre, Vancouver, BC V5Z 1L3, Canada, [2]Vaccine and Infectious Disease Division, Fred Hutchinson Cancer Research Center, Seattle, WA 98109-1024, USA and [3]Bioinformatics Training Program, University of British Columbia, Vancouver, BC V5Z 4S6, Canada

Associate Editor: Jonathan Wren

ABSTRACT

Summary: flowDensity facilitates reproducible, high-throughput analysis of flow cytometry data by automating a predefined manual gating approach. The algorithm is based on a sequential bivariate gating approach that generates a set of predefined cell populations. It chooses the best cut-off for individual markers using characteristics of the density distribution. The Supplementary Material is linked to the online version of the manuscript.

Availability and implementation: R source code freely available through BioConductor (http://master.bioconductor.org/packages/devel/bioc/html/flowDensity.html.). Data available from Flow Repository.org (dataset FR-FCM-ZZBW).

Contact: rbrinkman@bccrc.ca

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on June 18, 2014; revised on August 29, 2014; accepted on October 12, 2014

1 INTRODUCTION

Flow cytometry (FCM) is the predominant technique used to identify and quantify the widest variety of cell types. However, it is widely recognized that FCM data analysis, primarily in the identification of clusters of homogenous cells in high dimensional space, has become a rate-limiting step in application of the technology. Researchers currently use a manually intensive and subjective process of serial inspection of one or two characteristics (dimensions) at a time (a process termed 'gating'). There is widespread demand for the development of software tools as the ability to organize, analyze and exchange FCM data is lagging far behind the ability to run samples, to the detriment of health research (O'Neill *et al*., 2013). One reason that end-users are still relying on manual analysis is the absence of a systematic way to formulate and transfer their expert knowledge to automated software in order to provide satisfactory results on hard to gate cell populations.

Although unsupervised clustering algorithms have been developed for FCM data, they often fail to replicate a human expert's gating results due to their generalized, global approach and often return as a result of a number of clusters that cannot be interpreted readily without more sophisticated cluster matching methods (Aghaeepour *et al*., 2013).

Supervised approaches have also focused on setting global parameters that are difficult to appropriately set for all clusters and requiring tuning of parameters that are not linked to researchers' biological understanding of the data (Hu *et al*., 2013; Zare *et al*., 2010). To address these issues, we developed flowDensity, an automated gating approach that emulates an expert's sequential 2D gating strategy to identify predefined cell populations using a sequential bivariate gating algorithm. The innovation of our approach is to use customized threshold calculations for different cell subsets, based on expert knowledge of hierarchical gating order and 1D density estimation. Using this approach, cell populations can be defined once using a gating strategy as a guide, which also enables clusters to be easily compared across samples.

2 APPROACH

flowDensity estimates the region around cell populations using characteristics of the marker density distribution (e.g. the number, height and width of peaks and the slope of the distribution curve). Parameters can be adjusted on a population-specific basis when extra information is given by a user (e.g. desired percentile cut-off, number of standard deviations from the peak). To pick the proper threshold for each marker within each sample, flowDensity first finds all the peaks ($p = 1, 2, \ldots n$) in the density distribution.

(1) If $p = 1$, the position of threshold can be determined by tracking the slope of density for a drastic change. Percentiles, standard deviation and FMO controls can be used when given as extra information.

(2) If $p = 2$, the position of threshold is the minimum intersection point between the two peaks on the density curve, unless forced to use methods above (e.g. percentiles, slope changes).

(3) If $p \geq 3$, for each peak, it calculates the height of the peak h and its distance from the next adjacent peak d and calculates the score $\frac{d}{h}$. It then finds the peak corresponding to the maximum of all computed scores and chooses this peak and its next adjacent and goes to the case $p = 2$.

*To whom correspondence should be addressed.
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

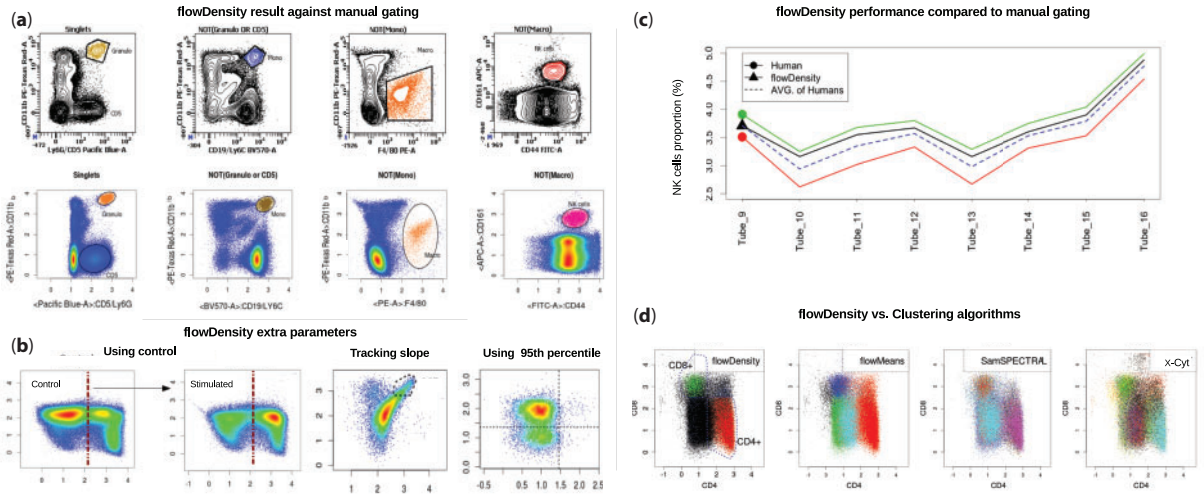

Fig. 1. (**a**) Comparison of flowDensity result to manual analysis. (**b**) flowDensity parameters for handling specific cases. (**c**) Comparison of flowDensity performance in terms of variability with expert's manual gating. (**d**) Comparison of flowDensity to state-of-the-art clustering algorithms

2.1 Gating rare cell populations

To address the challenge of identifying rare cell populations, flowDensity determines whether the slope of the curve of the density distribution along sections of the curve drops below a threshold. In rare cases where the slope varies slowly, a percentile of the density distribution (default 90th) is used as the threshold. If the spread of the density distribution is mostly around the mean (i.e. the standard deviation is small relative to the mean), then slope tracking tends to returns better results than percentile (Fig. 1b). If neither of these techniques are able to set a proper threshold, the peak value plus a multiplier of the standard deviation is chosen as the threshold. However, the user can also modify this decision by setting parameters specifically for the challenging cell populations.

2.2 Utilizing control samples

flowDensity can accept control data for each channel used in gating (Fig. 1b). When a control sample is included, the gating threshold is calculated based on the control population.

3 RESULTS AND DISCUSSION

flowDensity's completely automated results match that of expert users when it was possible to set cell population boundaries in a data-driven manner (Fig. 1a). For those cell populations where additional information is required to set the cell population boundaries according to those chosen by the user, thresholds are set based on the information (Fig. 1b). When no additional information is given or the population of interest is really small (~0.04% for a typical FCM data with 300 000 events), flowDensity might fail to identify a proper gating due to misleading kernel estimation.

flowDensity performs well in the range of human variation and is close to the average of humans (Fig. 1c). We also compared the flowDensity result with state-of-the-art supervised and unsupervised FCM analysis tools. Only flowDensity identified the correct populations, even though the number of clusters was given as input and parameters were extensively adjusted (Fig. 1d and Supplementary Material). This success lies in the approach, which does not identify all populations by examining all dimensions simultaneously like typical clustering algorithms. Rather it can be seen as a sequential algorithm that automates a 2D traditional gating scheme in a data-driven manner highly suited to targeted investigations such as clinical trials identifying predefined cell populations. The catch is customization, does have to be performed once for each marker panel which is simplified since flowDensity is integrated into the opencyto framework (Finak *et al*., 2014).

ACKNOWLEDGEMENTS

Thanks to Kelly Lundsten for helpful discussion on using FMO controls, Jonathan Bramson for control samples, Hervé Luche for mouse data and comments and Mike Jiang for software review.

Funding: This work was supported by National Institutes of Health (R01 EB008400), the Human Immunology Consortium (U19 AI089986) and National Science and Engineering Research Council.

Conflict of interest: none declared.

REFERENCES

Aghaeepour,N. *et al.* (2013) Critical assessment of automated flow cytometry data analysis techniques. *Nat. Methods*, **10**, 228–238.

Finak,G. *et al.* (2014) PLOS computational biology OpenCyto: an open source infrastructure for scalable, robust, reproducible, and automated end-to-end flow cytometry data analysis. *PLoS Comput. Biol.*, **10**, e1003806.

Hu,X. *et al.* (2013) Application of user-guided automated cytometric data analysis to large-scale immunoprofiling of invariant natural killer T cells. *Proc. Natl Acad. Sci. USA*, **110**, 19030–19035.

O'Neill,K. *et al.* (2013) Flow cytometry bioinformatics. *PLoS Comput. Biol.*, **9**, e1003365.

Zare,H. *et al.* (2010) Data reduction for spectral clustering to analyze high throughput flow cytometry data. *BMC Bioinformatics*, **11**, 403.