GateFinder: Projection-based Gating Strategy Optimization for Polychromatic and Mass Flow Cytometry Data

Nima Aghaeepour and Erin F. Simonds

January 25, 2018

naghaeep@gmail.com and erin.simonds@gmail.com

Contents

1 Licensing

2 Introduction

Exploratory analysis using polychromatic [6] and mass [5] flow cytometry together with modern computational tools (*e.g.,* [1–4,9]) often result in identification of complex cell populations that cannot be easily described using a limited number of markers. GateFinder attempts to identify a series of gates (*i.e.* polygon filters on 2-dimensional scatter plots) that can discriminate between a target cell population and other cells.

Briefly, the analysis consists of three steps:

1. Project the data points into all possible pairs of dimensions. Use robust statistics to exclude outliers [8]. Calculate a convex hull (a convex polygon around the remaining data points) [7].

2. Calculate F-measure values for all available gates. Select the best one.

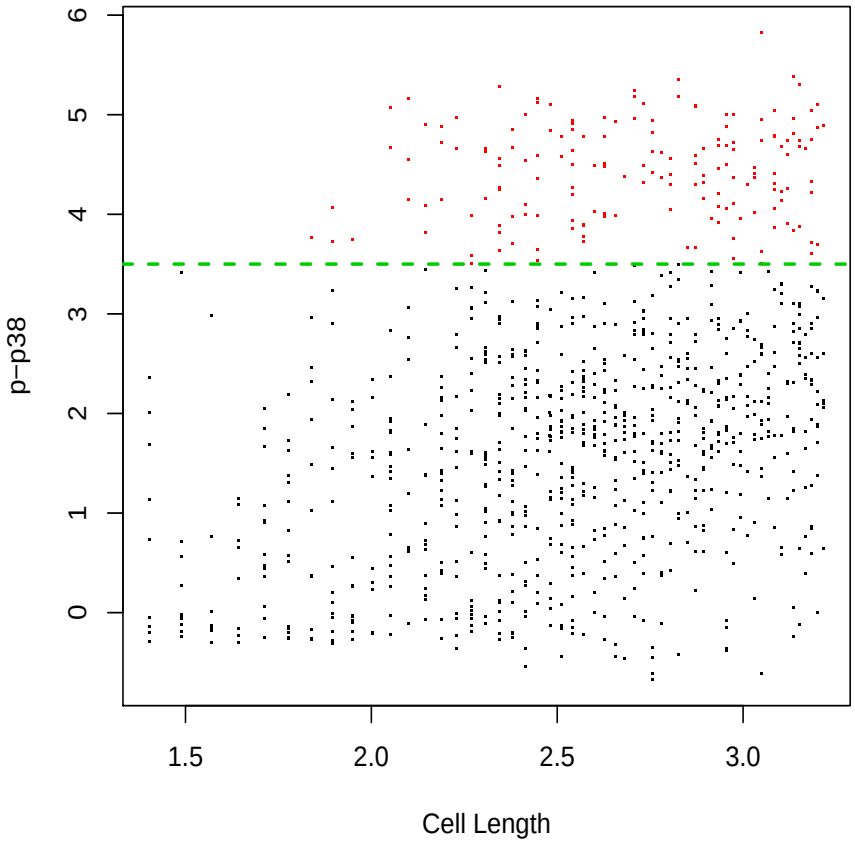
3. Depending on software configurations (see the *update.gates* parameter) either go to 1 or 2 unless the maximum number of iterations has been reached.

3 Basic Functionality

This example uses part of a publicly available bone marrow mass cytometry dataset [5]. In this specific subset of the dataset cells were stimulated by lipopolysaccharide (LPS) and the response was measured phosphorylation of p38 mitogen-activated protein kinase (p38 MAPK). A random subset of 1000 cells were selected for this analysis to comply with BioConductor's size and run time requirements. The optimal number of cells for GateFinder depends on the number of parameters in the search space, the size of the target population, and the desired purity. GateFinder expects transformed data. This dataset was previously transformed with the *arcsinhTransform()* function from the flowCore package, using parameters $a = 0$, $b = 0.2$, $c = 0$. Original analysis of the data revealed that the majority of the p38 MAPK response is in the CD11b+ monocytes. Here, we will use GateFinder to derive a specific gating strategy for the LPS-responsive cell population.

First, we select the target cell population by gating the phospho-p38 marker (dimension number 34) and selecting all cells with intensity greater than 3.5:

```
> library(GateFinder)
> library(flowCore)
> data(LPSData)
> targetpop <- (exprs(rawdata)[,34] > 3.5)
> plot(exprs(rawdata)[ , c(2,34)], pch='.', col=targetpop+1,
+      xlab='Cell Length', ylab='p-p38')
> abline(h=3.5, col=3, lwd=2, lty=2)
```

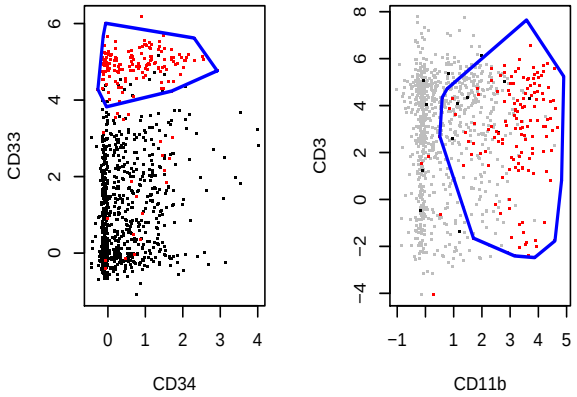


Next, we select the markers that should be considered for the gating strategy and run the core *GateFinder()* function:

```
> x=exprs(rawdata)[ , prop.markers]
> colnames(x)=marker.names[prop.markers]
> results=GateFinder(x, targetpop)
```

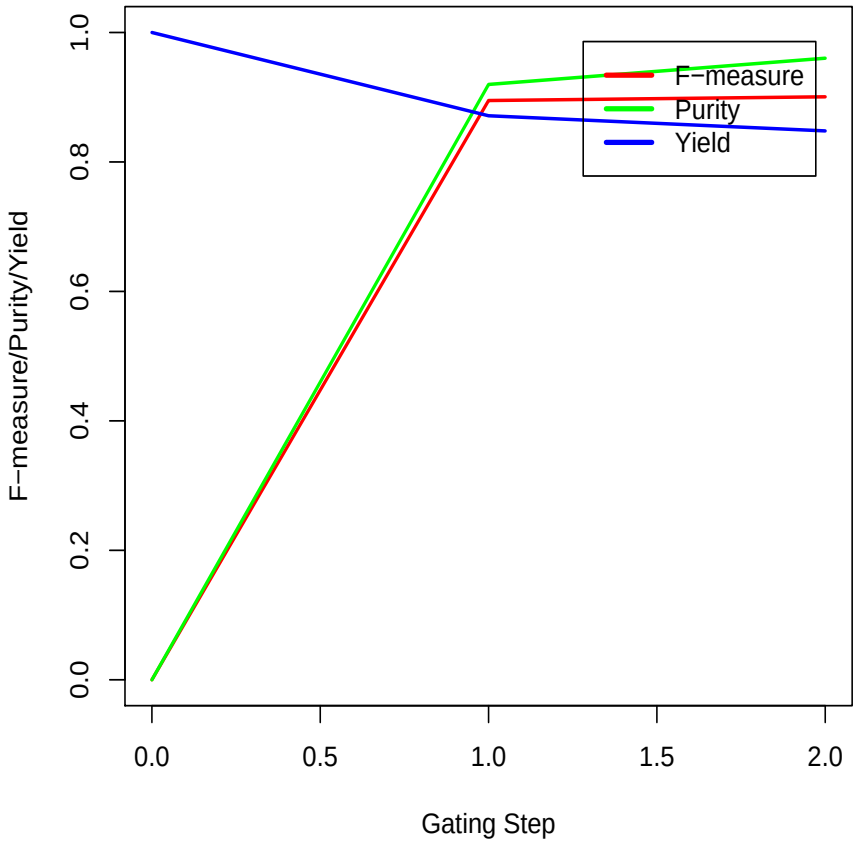
Now we can create a scatter plot of each gating step. GateFinder's *plot.GateFinder()* function accepts 4 arguments specifying the raw data, the output of the *GateFinder()* function, the layout of figure panels to assemble in the plot, and a logical mask specifying the target cells. The original target cells are highlighted in red. Gray cells were excluded in one of the previous gating steps. Black cells are cells that are not in the original target population. This analysis suggests that the target population is $CD33^{+}CD38^{+}CD11b^{+}CD123^{-}$.

```
> plot (x, results, c(2,3), targetpop)
```



We can also visualize the F-measure, precision (i.e., "purity"), and recall (i.e., "yield") of each step. As expected, making the gating more strict (by including more gating steps) increases the precision and decreases the recall of the gating strategy.

```
> plot(results)
```

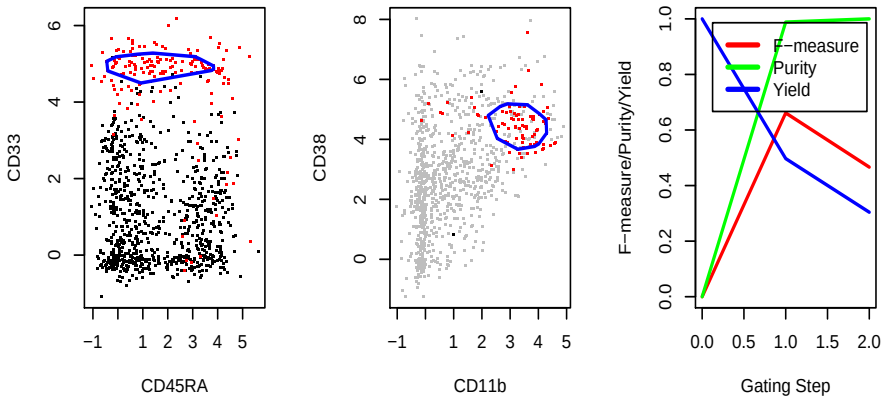


4 Advanced Parameters

GateFinder's functionality can be controled using two parameters: the *outlier.percentile* value controls the robustness of the convex hulls (polygon gates) to outliers and the *beta* value controls the relative impact of precision and recall on the F-measure calculations.

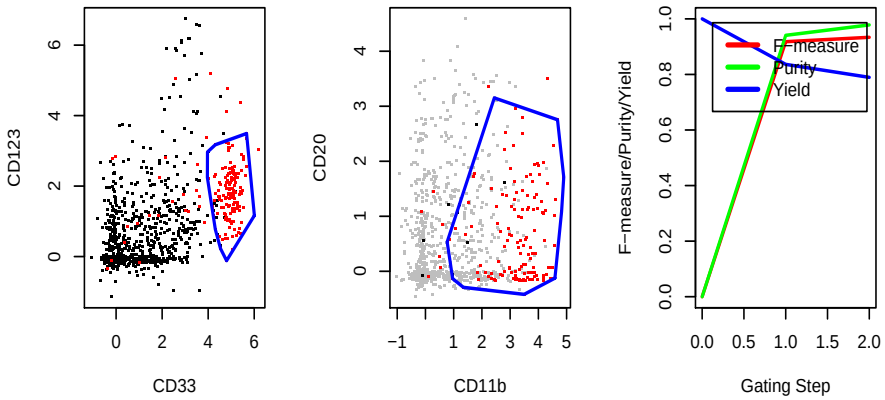
Higher values for the *outlier.percentile* parameter make the gates less strict (and therefore will increase precision and decrease recall):

```
> results=GateFinder(x, targetpop, outlier.percentile=0.5)
> plot (x, results, c(2,3), targetpop)
> plot(results)
```



Similarly, a *beta* value smaller than 1 increases in the impact of precision on the F-measure calculations. In the following calculations a value of 0.5 makes precision twice as important as recall. Therefore the algorithm modifies the gating strategy to increase the precision of the gating strategy. This is achieved by combining CD11b and CD33 in the very first gate at the cost of a decreased recall.

```
> results=GateFinder(x, targetpop, beta=0.5)
> plot (x, results, c(2,3), targetpop)
> plot(results)
```



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] El-ad David Amir, Kara L Davis, Michelle D Tadmor, Erin F Simonds, Jacob H Levine, Sean C Bendall, Daniel K Shenfeld, Smita Krishnaswamy, Garry P Nolan, and Dana Pe'er. visne enables visualization of high dimensional single-cell data and reveals phenotypic heterogeneity of leukemia. *Nature biotechnology*, 2013.

[5] Sean C Bendall, Erin F Simonds, Peng Qiu, D Amir El-ad, Peter O Krutzik, Rachel Finck, Robert V Bruggner, Rachel Melamed, Angelica Trejo, Olga I Ornatsky, et al. Single-cell mass cytometry of differential immune and drug responses across a human hematopoietic continuum. *Science*, 332(6030):687–696, 2011.

[6] Pratip K Chattopadhyay, David A Price, Theresa F Harper, Michael R Betts, Joanne Yu, Emma Gostick, Stephen P Perfetto, Paul Goepfert, Richard A Koup, Stephen C De Rosa, et al. Quantum dot semiconductor nanocrystals for immunophenotyping by polychromatic flow cytometry. *Nature medicine*, 12(8):972–977, 2006.

[7] William F Eddy. A new convex hull algorithm for planar sets. *ACM Transactions on Mathematical Software (TOMS)*, 3(4):398–403, 1977.

[8] Peter Filzmoser, Ricardo Maronna, and Mark Werner. Outlier identification in high dimensions. *Computational Statistics & Data Analysis*, 52(3):1694–1711, 2008.

[9] Peng Qiu, Erin F Simonds, Sean C Bendall, Kenneth D Gibbs Jr, Robert V Bruggner, Michael D Linderman, Karen Sachs, Garry P Nolan, and Sylvia K Plevritis. Extracting a cellular hierarchy from high-dimensional cytometry data with spade. *Nature biotechnology*, 29(10):886–891, 2011.