Automated Identification and Comparison of Cell Populations from Multi-Dimensional Flow Cytometry Data

Yu “Max” Qian, Ph.D.

Division of Biomedical Informatics, Dept. of Clinical Sciences & Dept. of Pathology

Jan 17th, 2012

yu.qian@utsouthwestern.edu
Outline

- Motivation
- Methods and Results
- Evaluation and Discussion
- Summary
FCM instrumentation & reagents

A practical approach to multicolor flow cytometry for immunophenotyping

Nicole Baumgarth\textsuperscript{a,1}, Mario Roederer\textsuperscript{b,*}

\textsuperscript{a}Department of Genetics, Stanford University Medical School, Beckman Center B-007, Stanford, CA 94305-5318, USA
\textsuperscript{b}Department of Stomatology, UCSF, San Francisco, CA, USA

- BD LSR II: 19 parameters
- Amnis – Imaging flow cytometer
- BD LSR Fortessa: 20 parameters
- Cytof mass cytometer
Traditional Flow Cytometry Analysis

Goal - group together cells with similar characteristics

Traditional approach - manual gating 2D at a time

- Subjective
- Time-consuming
- Doesn’t handle overlapping distributions well
- Sensitive to slight difference in fluorescence intensity distributions between samples
- Requires at least one 2D plot that clearly segregates populations in question
Overlapping populations
Populations can be better segregated when looking at >2 markers each time; therefore, computation may be necessary when >3 markers need to be examined.
Solution: Clustering

• Assumption: Cells of the same population express ALL biological markers similarly
Design of flock: Flow cytometry clustering without $K$
Algorithm Procedure

1) Generate initial clusters
2) Normalize dimensions within clusters
3) Select dimensions for initial clusters
4) Partition and merge the initial clusters in their selected subspaces
5) Output the final clusters

2D example
Divide with hyper-grids
Find dense hyper-regions
Merge neighboring dense hyper-regions
Clustering based on region centers
Bin selection methods

Goal is to minimize the Mean Squared Error

\[ L(h(x), f(x)) = \int (h(x) - f(x))^2. \]

- Scott’s method
  \[ v_{scott} = 3.49 s N^{-1/3}. \]

- Stone’s method
  \[ K(v, M) = \frac{1}{v} \left( \frac{2}{N - 1} - \frac{N + 1}{N - 1} \sum_{m=1}^{M} \pi_i^2 \right) \]

- Knuth’s method, to maximize
  \[ N \log M + \log \Gamma \left( \frac{M}{2} \right) - M \log \Gamma \left( \frac{1}{2} \right) - \log \Gamma \left( N + \frac{M}{2} \right) + \sum_{k=1}^{M} \log \Gamma \left( n_k + \frac{1}{2} \right) + K \]
Density threshold selection

- Minimum description length

\[ \mu_s(i) = \left( \frac{\sum_{1 \leq j \leq i} x_j}{i} \right) \]

\[ \mu_d(i) = \left( \frac{\sum_{i+1 \leq j \leq \sigma} x_j}{(\sigma - i)} \right) \]

\[ L(i) = \log_2(\mu_s(i)) + \sum_{1 \leq j \leq i} \log_2(|x_j - \mu_s(i)|) + \log_2(\mu_d(i)) + \sum_{i+1 \leq j \leq \sigma} \log_2(|x_j - \mu_d(i)|) \]
Density Variability in High-Dimensional Data Space

Fix the number of bins and density threshold, and use a Gaussian simulator to simulate 2-d, ..., 10-d data with 2 Gaussian clusters

X-axis: Number of dimensions
Y-axis: Number of groups of adjacent hyper-regions

X-axis: Number of dimensions
Y-axis: Number of bins selected by Stone’s Method
Dimension Selection and Cluster Merging

1) 0-1 column-wise normalize each cluster

2) Select 3 dimensions for each cluster based on standard deviations (if number of dimensions < 3, all dimensions are used)

3) Partition a cluster into two, *if necessary* (this step can be optionally repeated)

4) 0-1 column-wise normalize each pair of partitions

5) Select 3 dimensions for each pair of partitions

6) Starting from the pair that are closest in the 3-dimensional space, merge a pair of partitions, *if necessary*

7) Repeat Steps 4) to 6) until there is no pair to merge
Merging/Partitioning Criteria

Propose the following linear-time approach to replace the slow \(O(N^2)\) nearest/mutual neighbor graph

Two partitions should not be merged

Two partitions should be merged
Identification of 17 B-cell populations in Human Blood
Overview of FLOCK-identified B-cell Populations in PBMC
FLOCK-identified Naïve B-cell Populations in PBMC
FLOCK-identified Double Negative B-cells and Plasmablasts

**F** Double Negative Memory B cells (CD27-, IgD-, CD38-)
- **DNM1** (IgG+)
- **DNM2** (IgG-)

**G** Plasmablasts (CD27high, IgD-, IgG-, CD38+)
- **PB**

- **B220**
- **CD24**
- **CD38**
- **IgG**
## Cell Population Summary Table

**Seventeen Peripheral Blood B Cell Subsets**

<table>
<thead>
<tr>
<th>Population</th>
<th>Color</th>
<th>CD27</th>
<th>IgD</th>
<th>IgG</th>
<th>CD38</th>
<th>CD24</th>
<th>B220</th>
<th>Proportion</th>
<th>Putative cell type</th>
</tr>
</thead>
<tbody>
<tr>
<td>N1</td>
<td>Magenta</td>
<td>−</td>
<td>+</td>
<td>−</td>
<td>−</td>
<td>+</td>
<td>+</td>
<td>4.69%</td>
<td>Naïve (CD38−)</td>
</tr>
<tr>
<td>N2</td>
<td>Gray</td>
<td>−</td>
<td>+</td>
<td>−</td>
<td>+</td>
<td>Int</td>
<td>+</td>
<td>48.94%</td>
<td>Naïve (CD38+)</td>
</tr>
<tr>
<td>N3</td>
<td>Purple</td>
<td>−</td>
<td>+</td>
<td>−</td>
<td>+</td>
<td>+</td>
<td>Low</td>
<td>4.41%</td>
<td>Naïve (CD38+B220low)</td>
</tr>
<tr>
<td>UM1</td>
<td>Darkred</td>
<td>+</td>
<td>+</td>
<td>−</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>1.55%</td>
<td>Unswitched memory (CD38+)</td>
</tr>
<tr>
<td>UM2</td>
<td>Salmon</td>
<td>+</td>
<td>+</td>
<td>−</td>
<td>−</td>
<td>+</td>
<td>+</td>
<td>0.94%</td>
<td>Unswitched memory (CD38−)</td>
</tr>
<tr>
<td>UM3</td>
<td>Darkblue</td>
<td>+</td>
<td>Int</td>
<td>−</td>
<td>+</td>
<td>+</td>
<td>Low</td>
<td>6.16%</td>
<td>IgDlow unswitched memory (CD38+)</td>
</tr>
<tr>
<td>UM4</td>
<td>Green</td>
<td>+</td>
<td>Int</td>
<td>−</td>
<td>−</td>
<td>+</td>
<td>Low</td>
<td>11.50%</td>
<td>IgDlow unswitched memory (CD38−)</td>
</tr>
<tr>
<td>GSM1</td>
<td>Grayishgreen</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>0.36%</td>
<td>Switching memory (IgD+IgG+CD38+)</td>
</tr>
<tr>
<td>GSM2</td>
<td>Yellow</td>
<td>+</td>
<td>−</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>Low</td>
<td>4.05%</td>
<td>Switched memory (CD38+)</td>
</tr>
<tr>
<td>GSM3</td>
<td>Blue</td>
<td>+</td>
<td>−</td>
<td>+</td>
<td>−</td>
<td>+</td>
<td>Low</td>
<td>4.40%</td>
<td>Switched memory (CD38−)</td>
</tr>
<tr>
<td>GNSM1</td>
<td>Cyan</td>
<td>+</td>
<td>−</td>
<td>−</td>
<td>−</td>
<td>+</td>
<td>Low</td>
<td>4.84%</td>
<td>IgD-IgG- memory</td>
</tr>
<tr>
<td>GNSM2</td>
<td>Darkgreen</td>
<td>+</td>
<td>−</td>
<td>−</td>
<td>−</td>
<td>+</td>
<td>Low</td>
<td>3.84%</td>
<td>IgD-IgG- memory</td>
</tr>
<tr>
<td>GNSM3</td>
<td>Teal</td>
<td>+</td>
<td>−</td>
<td>−</td>
<td>−</td>
<td>+</td>
<td>+</td>
<td>1.30%</td>
<td>IgD-IgG- memory</td>
</tr>
<tr>
<td>GNSM4</td>
<td>Orange</td>
<td>+</td>
<td>−</td>
<td>−</td>
<td>−</td>
<td>−</td>
<td>Low</td>
<td>0.51%</td>
<td>IgD-IgG- memory</td>
</tr>
<tr>
<td>DNSM1</td>
<td>Pink</td>
<td>−</td>
<td>−</td>
<td>+</td>
<td>−</td>
<td>−</td>
<td>+</td>
<td>0.85%</td>
<td>Double negative memory (IgG+)</td>
</tr>
<tr>
<td>DNSM2</td>
<td>Darkgray</td>
<td>−</td>
<td>−</td>
<td>−</td>
<td>−</td>
<td>−</td>
<td>+</td>
<td>0.91%</td>
<td>Double negative memory (IgG−)</td>
</tr>
<tr>
<td>PB</td>
<td>Red</td>
<td>High</td>
<td>−</td>
<td>−</td>
<td>High</td>
<td>−</td>
<td>Low</td>
<td>0.75%</td>
<td>Plasmablasts</td>
</tr>
</tbody>
</table>
FlowCAP COMPETITION
Research Article
Merging Mixture Components for Cell Population Identification in Flow Cytometry

Greg Flank, Ali Bashashati, Ryan Brinkman, and Raphael Gottardo

1Computational Biology Unit, Clinical Research Institute of Montreal, 150 Pine Avenue West, Montreal, QC, Canada
2Jefco Laser Laboratory, BC Cancer Research Centre, Vancouver, BC, Canada V6H 3Z3
3Department of Biochemistry, University of Montreal, Montreal, QC, Canada

Elucidation of Seventeen Human Peripheral Blood B-Cell Subsets and Quantification of the Tetanus Response Using a Density-Based Method for the Automated Identification of Cell Populations in Multidimensional Flow Cytometry Data

Yu Qian,1,2 Chungun Wei,1 F. Eun-Hyang Lee,3 John Campbell,4 Jessica Halliley,5 Jaimie A. Lee,6 Jennifer Cai,5 Y. Megen Kong,7 Eva Sada,8 Elizabeth Thomson,9 Patrick Dunn,10 Adam C. Lepine,11 Inzit J. Karandikar,12 Christopher M. Tipton,13 Tim Moomaw,14 Haidi Saha,15 and Richard H. Schuermann16

1Department of Pathology, University of Texas Southwestern Medical Center, Dallas, Texas, 75390
2Division of Biomedical Informatics, University of Texas Southwestern Medical Center, Dallas, Texas, 75390
3Department of Medicine, University of Rochester Medical School and Department of Rochester, New York, 14642
4Health Information Systems, Northeastern University, Boston, Massachusetts, 02115

Cytometry Part B 1471-2126 (2010) 11:569-582

Cytometry

Statistical Mixture Modeling for Cell Subtype Identification in Flow Cytometry

Cliburn Chan,1 Fang Feng,1 Janet Ottinger,2 David Foster,3 Mike West,4 Thomas B. Kepler5,6

1Cytometry Part B 1471-2126 (2010) 11:569-582

BMC Bioinformatics

flowClust: a Bioconductor package for automated gating of flow cytometry data

Kenneth Lo (c.lo@stat.ubc.ca)
Florian Hahne (fhahne@fccc.org)
Ryan R Brinkman (rbrinkman@bccrc.ca)
Raphael Gottardo (raphael.gottardo@ircc.qc.ca)

Data reduction for spectral clustering to analyze high throughput flow cytometry data

Habil Zare1, Parisa Shooshtari1, Arvind Gupta2, Ryan R Brinkman3,4

BMC Bioinformatics

Automated high-dimensional flow cytometry data analysis

Saurabhjyoty Pyne4, Xinli Hu1, Kui Wang1, Elizabeth Rossin1, Rongping Liu1, Lisa M. Maier1,2, Claire Baecher-Allan1, Geoffrey J. McLachlan1,4, Pablo Tamayo2, David A. Haffar1,2, Philip L. DeJager1,2, and Jill P. Mesirov2,3

1Broad Institute of MIT and Harvard, 7 Cambridge Center, Cambridge, MA 02142
2Department of Mathematics and Institute for systems Biology, University of Queensland, St. Lucia, Queensland, 4072, Australia
3Department of Applied Mathematics, National Chung Hsing University, Taichung 402, Taiwan
4Division of Molecular Immunology, Center for Neurologic Diseases, Brigham and Women’s Hospital and Harvard Medical School, 77 Avenue Louis Pasteur, Boston, MA 02115, and Partners Center for Personalized Genetic Medicine, Boston, MA 02115

Cytometry

Automated Gating of Flow Cytometry Data via Robust Model-Based Clustering

Kenneth Lo,1,2 Ryan Remy Brinkman,2 Raphael Gottardo1

Cytometry Part B 1471-2126 (2010) 11:569-582

BMC Bioinformatics

A Statistical Pattern Recognition Approach for Determining Cellular Viability and Lineage Phenotype in Cultured Cells and Murine Bone Marrow

John Quinn,1,2 Paul W. Fisher,3 Renold J. Capocassele,3 Ram Achuthanandan,3 Moshe Kam,1 Peter J. Bugelski,1,3 Leonid Irribaren1,3

Cytometry Part B 1471-2126 (2010) 11:569-582

Zare et al. BMC Bioinformatics 2010, 11:569
http://www.biomedcentral.com/1471-2105/11/569

METHODOLOGY ARTICLE
Open Access

Data reduction for spectral clustering to analyze high throughput flow cytometry data

Habil Zare1, Parisa Shooshtari1, Arvind Gupta2, Ryan R Brinkman3,4

METHODOLOGY ARTICLE
Open Access

Automated Gating of Flow Cytometry Data via Robust Model-Based Clustering

Kenneth Lo,1,2 Ryan Remy Brinkman,2 Raphael Gottardo1

METHODOLOGY ARTICLE
Open Access

The curvHDR method for gating flow cytometry samples

Ulrike Naumann1, George Luta2, Matthew P Wand3

BMC Bioinformatics

flowClust: a Bioconductor package for automated gating of flow cytometry data

Kenneth Lo (c.lo@stat.ubc.ca)
Florian Hahne (fhahne@fccc.org)
Ryan R Brinkman (rbrinkman@bccrc.ca)
Raphael Gottardo (raphael.gottardo@ircc.qc.ca)
Flow Cytometry: Critical Assessment of Population Identification Methods (FlowCAP)

Organizing Committee
Ryan Brinkman, British Columbia Cancer Agency
Raphael Gottardo, Fred Hutchinson Cancer Research Center
Tim Mosmann, University of Rochester
Richard H. Scheuermann, University of Texas Southwestern Medical Center
<table>
<thead>
<tr>
<th>Dataset</th>
<th># Samples</th>
<th># Events</th>
<th># Colors</th>
<th>Analyte-Reporter</th>
<th>Provider</th>
</tr>
</thead>
<tbody>
<tr>
<td>GvHD</td>
<td>12</td>
<td>14,000</td>
<td>4</td>
<td>CD4-FITC; CD8b-PE; CD3-PerCP; CD8-APC</td>
<td>BCCRC &amp; TreeStar</td>
</tr>
<tr>
<td>DLBCL</td>
<td>30</td>
<td>5,000</td>
<td>3</td>
<td>CD5-FITC; CD19-PE; CD3-Cy5</td>
<td>BCCRC</td>
</tr>
<tr>
<td>HSCT</td>
<td>30</td>
<td>10,000</td>
<td>4</td>
<td>CD45.2-APC; CD45.1-FITC; Ly65/Mac1-PE; Dead cells-PI</td>
<td>BCCRC</td>
</tr>
<tr>
<td>WNV</td>
<td>13</td>
<td>100,000</td>
<td>6</td>
<td>CD3-PECy5; CD4-PECy7; CD8-AF700; IFNγ-PEA; IL17-APC; Free amine-CFSE</td>
<td>McMaster</td>
</tr>
<tr>
<td>ND</td>
<td>30</td>
<td>17,000</td>
<td>10</td>
<td>CD56-Q605; CD8-AF700; CD45-PECy5; CD3/CD14-PECy7; Proprietary-FITC, PerCPCy5, PacificBlue, PacificOrange, APC, PE</td>
<td>Amgen</td>
</tr>
</tbody>
</table>
Automatically Predict Cluster Membership of Each Event (Cell)

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Rank Score</th>
<th>Total Runtime</th>
</tr>
</thead>
<tbody>
<tr>
<td>flowMeans</td>
<td>45.6</td>
<td>00:04:23:27</td>
</tr>
<tr>
<td>FLOCK</td>
<td>42.6</td>
<td>00:00:37:38</td>
</tr>
<tr>
<td>FLAME</td>
<td>42.1</td>
<td>00:05:31:12</td>
</tr>
<tr>
<td>SamSPECTRAL</td>
<td>41.8</td>
<td>00:07:21:44</td>
</tr>
<tr>
<td>MM&amp;PCA</td>
<td>27.8</td>
<td>00:00:04:35</td>
</tr>
<tr>
<td>FlowVB</td>
<td>26.8</td>
<td>03:02:23:09</td>
</tr>
<tr>
<td>MM</td>
<td>25.8</td>
<td>00:00:13:00</td>
</tr>
<tr>
<td>flowClust/Merge</td>
<td>24.2</td>
<td>10:13:00:00</td>
</tr>
<tr>
<td>FEK</td>
<td>19.5</td>
<td>00:15:25:00</td>
</tr>
<tr>
<td>CDP</td>
<td>18.2</td>
<td>00:01:48:06</td>
</tr>
<tr>
<td>SWIFT</td>
<td>15.5</td>
<td>05:23:24:30</td>
</tr>
</tbody>
</table>

*Table 6: Total runtimes (dd:hh:mm:ss) and rank scores for challenge 1*
Cross-Sample Comparison
FlowCAP2 Challenge 5

Before Population Mapping

tube1_arm1_071 73311_FL1vsFL4
tube1_arm2_071 83231_FL1vsFL4
tube1_arm3_071 63161_FL1vsFL4
tube1_arm4_071 73091_FL1vsFL4

After Population Mapping
Cross-Sample Comparison with FLOCK

Proportion change of PlasmaBlasts at different days with Tetanus study

FCM data from Sanz Lab, Univ. of Rochester
Visualization and UI
System Availability

• Immunology Database and Analysis Portal: http://www.immport.org

• http://immportflock.sourceforge.net
Discussion and Summary

• Manual gating can be problematic
  – May need to look at more than two dimensions each time to segregate populations
• Computational methods may be useful
• Evaluation criteria is still manual gating
  – Reproduce manual gating
  – Exploratory analysis
• Population interpretation, visualization, and databasing of experiments
• FlowCAP competition aims to provide method assessment and guide for end users
Acknowledgment

UT Southwestern
Richard Scheuermann
David Dougall
Megan Kong
Paula Guidry
Eva Sadat
Jamie Lee
Adam Seegmiller
Nitin Karandikar

Northrop Grumman
John Campbell
Yue Liu
Liz Thompson
Patrick Dunn
Jeff Wiser
Mike Atassi

Northrop Grumman

Rochester
Iñaki Sanz
Chungwen Wei
Eun Hyung Lee
Denise Kaminski
Tim Mosmann
Jessica Halliley
Chris Tipton

Univ. of Western Sydney
Quang Vinh Nguyen

BCCA
Ryan Brinkman

Supported by NIH N01 AI40076 (BISC)