

Rapid Cell Population Identification in FCM Data

Nima Aghaeepour¹, R. Nikolic^{1,2}, H. H. Hoos³, R. R.
Brinkman^{1,4}

¹Terry Fox Laboratory, BC Cancer Agency, Vancouver, British Columbia, Canada

²Department of Statistics, University of Oxford, Oxford, United Kingdom

³Department of Computer Science, University of British Columbia, British
Columbia, Canada

⁴Department of Medical Genetics, University of British Columbia, British
Columbia, Canada

FlowCAP-I

Sorry, Nima is still waiting for a visa



Automated Clustering Methods are Increasing in Complexity

- K-means¹
- Gaussian Mixture Models²
- t Mixture Models³
- skew-t Mixture Models⁴

More complex models can better capture complexity of flow data

¹R.F. Murphy. Automated identification of subpopulations in flow cytometric list mode data using cluster analysis. *Cytometry Part A*, 6(4):302:309, 2005.

²Chan, C. and Feng, F. and Ottinger, J. and Foster, D. and West, M. and Kepler, T.B. Statistical mixture modeling for cell subtype identification in flow cytometry. *Cytometry Part A*, 73(8):693:701, 2008

³K. Lo, R.R. Brinkman, and R. Gottardo. Automated gating of flow cytometry data via robust model-based clustering. *Cytometry Part A*, 73:321:332, 2008.

⁴S. Pyne, X. Hu, K. Wang, E. Rossin, T.I. Lin, L.M. Maier, C. Baecher-Allan, G.J. McLachlan, P. Tamayo, and D.A. Hafler. Automated high-dimensional flow cytometric data analysis. *Proceedings of the National Academy of Sciences*, 106(21):8519, 2009.

Increasing Model Complexity Increases Runtime

Comparison of Run Time of the Clustering Algorithms to Identify 10 Clusters.

Dataset	Average Runtime (mm:ss)			
	K-means	Gaussian Mixture Model	t Mixture Model	skew-t Mixture Model
GvHD	00:07	04:26	05:37	07:36
DLBCL	00:05	03:31	04:07	05:51

K-means is fast but:

- Relies on pre-defined number of populations
- Is limited to spherical populations

flowMerge⁵:

- Look for well-separated populations instead of the best fit
- Allow more than one component to model the same population

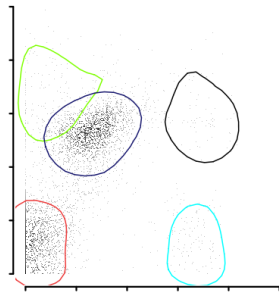
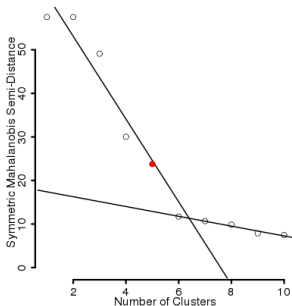
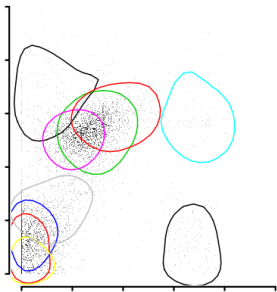
Our contribution:

- With a merging step, accuracy of the model is not necessary
- A simple model (*i.e.*, K-means) can be used to decrease the runtime

⁵G. Finak, A. Bashashati, R. Brinkman, and R. Gottardo. Merging mixture components for cell population identification in flow cytometry. *Advances in Bioinformatics*, 2009.

flowMeans: Overfitting to Overcome K-means Limitations

- Fit K-means using the “maximum number of clusters”
- Repeat until one cluster is left:
 - Calculate the distance between clusters
 - Merge the two closest clusters
- Find the change-point in which the clusters are well-separated



- Distance between clusters measured by Mahalanobis distance
 - Takes into account distance and spread of clusters
- Upper bound for the number of clusters:
 - Number of modes in 1D projections of the data on the principal components.
 - Using *feature* (R/BioConductor package from Matt Wand's group)

- Diffuse Large B-Cell Lymphoma (DLBCL)⁶:
 - 30 samples
 - Analyte: *CD3*, *CD5*, *CD19*

- Graft versus Host Disease (GvHD)⁷:
 - 12 samples
 - Analyte: *CD4*, *CD8b*, *CD3*, *CD8*

⁶ F. Hahne, A.H. Khodabakhshi, A. Bashashati, C.J. Wong, R.D. Gascoyne, A.P. Weng, V. Seyfert-Margolis, K. Bourcier, A. Asare, T. Lumley, et al. Per-channel basis normalization methods for flow cytometry data. *Cytometry Part A*, 77(2):121:131, 2009.

⁷ R. R. Brinkman, M. Gasparetto, S. J. Lee, A. J. Ribickas, J. Perkins, W. Janssen, R. Smiley, and C. Smith. High-content flow cytometry and temporal data analysis for defining a cellular signature of graft-versus-host disease. *Biology of blood and marrow transplantation : Journal of the American Society for Blood and Marrow Transplantation*, 13(6):691:700, Jun 2007.

Critical Assessment of Methods

- Accurate

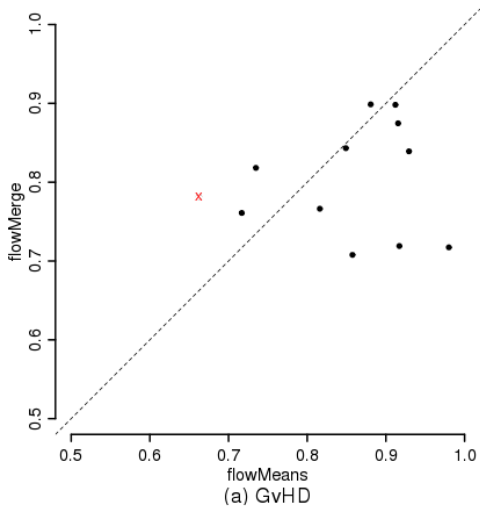
Dataset	Mean F-Measure ⁸ (SD)		
	flowMeans	flowMerge	FLAME
GvHD	0.84(0.07)	0.80(0.06)	0.68(0.13)
DLBCL	0.92(0.04)	0.92(0.05)	0.59(0.14)

- Fast

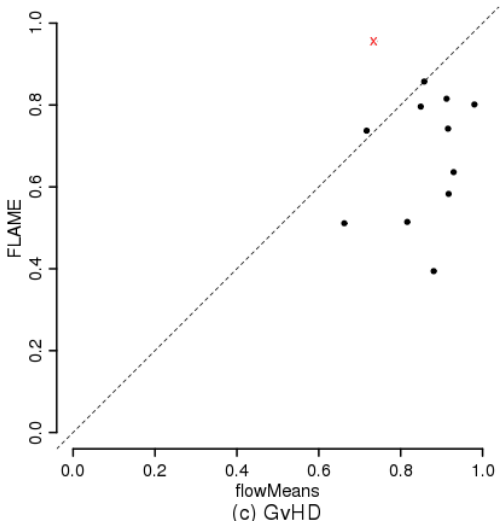
Dataset	Average Runtime (mm:ss)		
	flowMeans	flowMerge	FLAME
GvHD	00:28	15:34	18:41
DLBCL	00:21	11:40	15:35

⁸ N. Aghaeepour, A.H. Khodabakhshi, and R.R. Brinkman. An Empirical Study of Cluster Evaluation Metrics using Flow Cytometry Data. In Theoretical Clustering workshop, Advances in Neural Information Processing Systems, Volume 1, 2009.

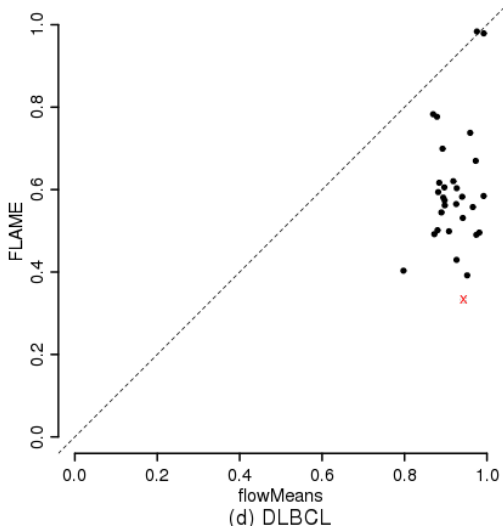
F-measure: flowMerge vs. flowMeans (GvHD)

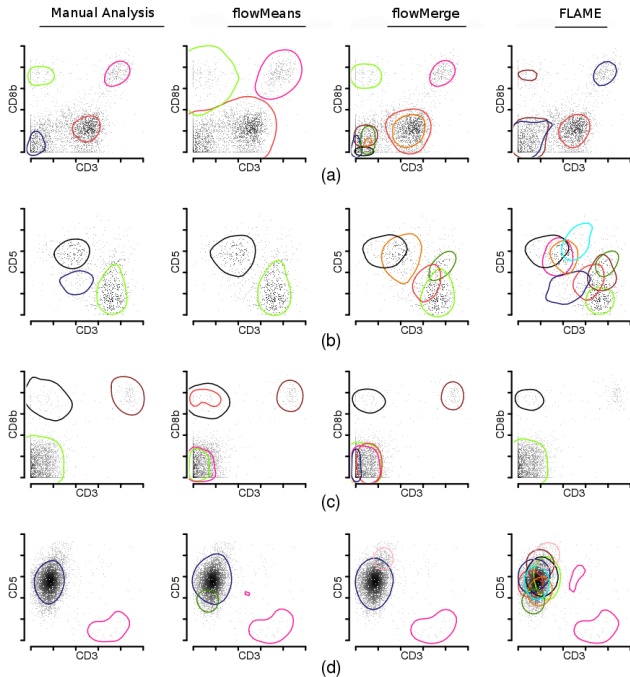


F-measure: FLAME vs. flowMeans (GvHD)



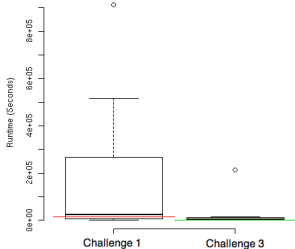
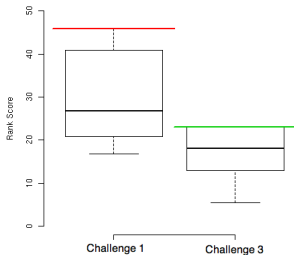
F-measure: FLAME vs. flowMeans (DLBCL)





Where we can do better

- flowMeans is slow in challenge 1 (Automated clustering)
 - Overall rank score: first
 - Ranked runtime: fourth (00:04:23:27)
- flowMeans is fast in challenge 3. (Automated; given k)
 - Overall rank score: first-ish (three-way tie)
 - Ranked runtime: second (00:00:01:03)
- The part of the method that has to decide on the number of clusters is slow and can be improved.



- Estimating the initial number of clusters without $1D$ projections.
- Update covariances for the Mahalanobis distances instead of recalculating.

- flowMeans: a framework that enables the application of K-means to FCM data
- In presence of a merging step, a simpler (and faster) clustering model can be used to speed up automated gating
- Our empirical evaluation showed that flowMeans can match the accuracy of the state of the art and is faster
- Paper will be out soon
- flowMeans source code & vignette is publicly available through Bioconductor:
<http://bioconductor.org/packages/devel/bioc/html/flowMeans.html>

Acknowledgements

- Brinkman lab** Josef Spidlen, Alireza Khodabakhshi, Parisa Shooshtari, Habil Zare, Kieran O'Neill
IRCM Greg Finak, Raphael Gottardo
FCCC Nishant Gopalakrishnan, Thomas Lumley
BCCA Andrew Weng, Randy Gascoyne, Nathalie Johnson
USW Matt Wand's group.
Funding NIH/NIBIB EB008400, MSFHR, UBC, CIHR/MSFHR-STPB

