
This item was submitted to [Loughborough's Research Repository](#) by the author.
Items in Figshare are protected by copyright, with all rights reserved, unless otherwise indicated.

Evaluation of flow cytometry automated data analysis platform performance for rare event detection

PLEASE CITE THE PUBLISHED VERSION

<http://www.uksb.org.uk/>

PUBLISHER

UK Society for Biomaterials

VERSION

AM (Accepted Manuscript)

LICENCE

CC BY-NC-ND 4.0

REPOSITORY RECORD

Cheung, Melissa, Jonathan Campbell, Julian Braybrook, Rob Thomas, and Jon Petzing. 2020. "Evaluation of Flow Cytometry Automated Data Analysis Platform Performance for Rare Event Detection". Loughborough University. <https://hdl.handle.net/2134/12574151.v1>.

POSTER FLASH: EVALUATION OF FLOW CYTOMETRY AUTOMATED DATA ANALYSIS PLATFORM PERFORMANCE FOR RARE EVENT DETECTION

Melissa Cheung^{1*}, Jonathan Campbell², Julian Braybrook², Rob Thomas¹, Jon Petzing¹.

¹Centre for Biological Engineering, Loughborough University, UK. ²LGC, Teddington, UK.

*Corresponding author: M.Cheung@lboro.ac.uk – PhD student (2nd year)

Introduction

Flow cytometry is a widely used technique for the detection and quantification of low frequency or rare cell populations at limits of detection (LOD) of 0.1% or lower, such as circulating tumour cells and haematopoietic stem cells. An increasing number of computational tools are available for automated gating of flow cytometry data. These tools have the capability to eliminate operator variation from traditional manual gating. Nevertheless, the potential for variation in data outputs arising from analysis of the same dataset using different algorithms means that research is needed to define the confidence in the use of automated data analysis software platforms for identifying rare cell populations.

Materials and Methods

Automated data analysis platforms using different mathematical principles have been investigated; hierarchical clustering, k-means clustering, self-organising map, density-based clustering, and model-based clustering methods. Two cluster and three cluster synthetic flow cytometry datasets have been developed with uniquely controlled rare cell events down to 0.0001%, among total events of up to 10^7 cells. The data analysis platform performances have been assessed based on a range of metrics including; accuracy, repeatability, and output pass/fail rates.

Results and Discussion

Whilst, for some data analysis platforms, the LOD of automated gating typically improve in percentage terms as the magnitude of total events increases, such improvement was often at the expense of prohibitive runtimes. Different LODs as a function of total cell events were observed. The k-means clustering method, when tested on default parameters, gave rare cell outputs with the least variation, but inadequate performance in terms of accuracy. In contrast the density-based clustering method was found to perform accurate rare cell count, but its repeatability and pass/fail performances did not rank as highly.

Conclusions

Careful consideration therefore needs to be given to derive a suitable range of metrics for judging the strengths and limitations of automated data analysis platforms.

Benchmarking automated flow cytometry data analysis software using synthetic datasets

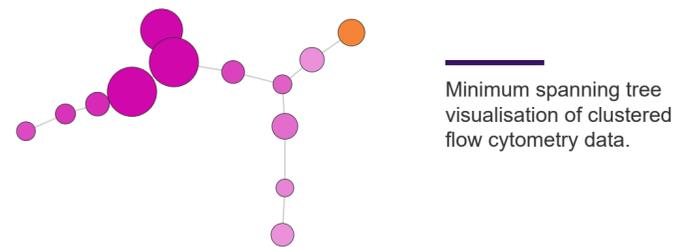
Melissa Cheung¹, Jonathan J Campbell², Julian Braybrook², Rob Thomas¹ and Jon Petzing¹

¹ Centre for Biological Engineering, Loughborough University, UK. ² LGC, Teddington, UK.

M.Cheung@lboro.ac.uk

Introduction

- Recent years have seen the development of many computational tools for automated flow cytometry data analysis.
- Different software will use different mathematical approaches to identify and quantify cells, so outputs can potentially vary when analysing the same dataset.
- To assess the performance and robustness of algorithms, live cell benchmarking datasets are used. However, these lack a 'ground truth' and often come from a narrow range of cell or disease models.
- Computer simulated flow cytometry datasets offer many advantages over their real-world counterparts; data properties can be highly controlled to objectively and systematically compare between software.



Aims

To define the confidence in flow cytometry automated data analysis software platforms in the context of cell therapy manufacturing.

Methods

We designed synthetic flow cytometry datasets, ran the datasets through automated software, then analysed the performance of resulting cell populations.



REFERENCES

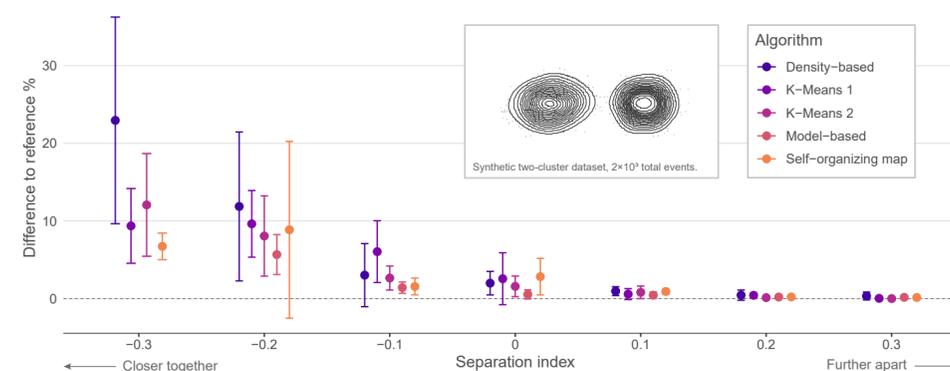
- Saeyes Y, Van Gassen S, Lambrecht BN. Computational flow cytometry: Helping to make sense of high-dimensional immunology data. *Nat Rev Immunol*. 2016;16(7):449–62.
- Weber LM, Robinson MD. Comparison of clustering methods for high-dimensional single-cell flow and mass cytometry data. *Cytom Part A*. 2016;89(12):1084–96.
- Aghaepour N, Finak G, Hoos H, Mosmann TR, Brinkman R, Gottardo R, et al. Critical assessment of automated flow cytometry data analysis techniques. *Nat Methods*. 2013 Mar 10;10(3):228–38.
- Qiu P. Toward deterministic and semiautomated SPADE analysis. *Cytom Part A*. 2017;91(3):281–9.

- Qian Y, Wei C, Eun-Hyung Lee F, Campbell J, Halliley J, Lee JA, et al. 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. *Cytom Part B Clin Cytom*. 2010;78B(S1):S69–82.
- Aghaepour N, Nikolic R, Hoos HH, Brinkman RR. Rapid cell population identification in flow cytometry data. *Cytom Part A*. 2011;79 A(1):6–13.
- Van Gassen S, Callebaut B, Van Helden MJ, Lambrecht BN, Demeester P, Dhaene T, et al. FlowSOM: Using self-organizing maps for visualization and interpretation of cytometry data. *Cytom Part A*. 2015;87(7):636–45.

Results

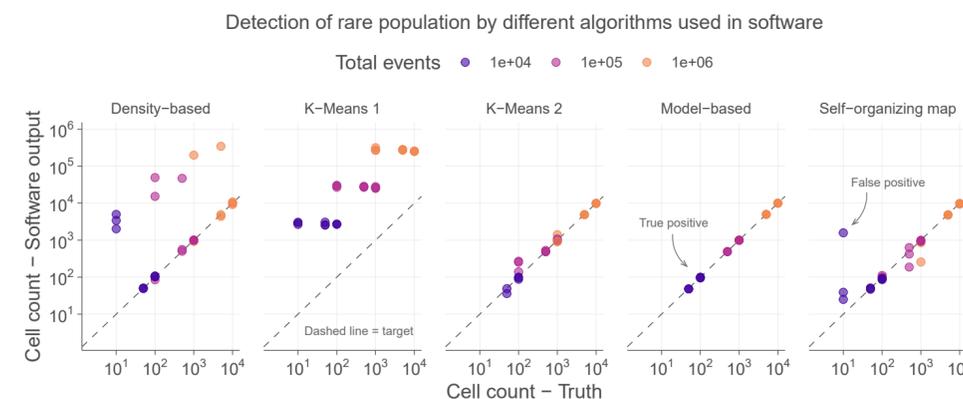
Distance improves performance

This graph shows how software have similar performance limitations as clusters move closer together. Points nearer to zero represent better accuracy; smaller $\pm 1SD$ error bars indicate better repeatability.



Rare subsets challenge software

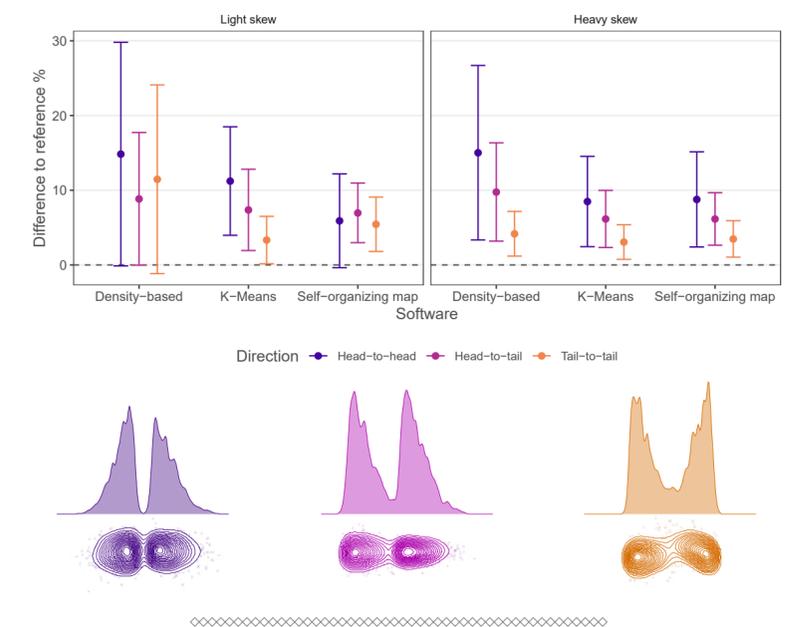
Certain software were more suited to rare cell population analysis. Most software had limits of detection (LOD) of 10^2 in 10^5 total events (0.1%). Increasing the total events increases the LOD in percentage terms.



Results

Skew orientation affects software performance

This plot shows that when using non-normally distributed clusters, the presence of skewed cell populations did not affect clustering performance, but the orientation of cluster skewness did.



Conclusion

- We have designed and used synthetic reference datasets to evaluate the initial performance characteristics of a number of popular flow cytometry data analysis software.
- This research demonstrates that benchmarking of flow cytometry automated software platforms will be possible with a high level of testing integrity using synthetic cluster dataset design leading to enhanced confidence in the data quality of cell characterisations.
- Consistently accurate rare cells detection at 10^2 in 10^5 events remains problematic for automated flow cytometry data analysis software.

ACKNOWLEDGEMENTS



CONTACT INFORMATION

Wolfson School of Mechanical, Electrical and Manufacturing Engineering
Loughborough University
Leicestershire LE11 3TU UK
www.lboro.ac.uk/departments/meme