Adventures in computational immunology: a novel approach to analyse high dimensional flow cytometry data

Ross J Burton[1], Simone Cuff[1], Peter Ghazal[2], Matthew Morgan[1,3], Andreas Artemiou[4], and Matthias Eberl[1,2]

1. Division of Infection and Immunity, School of Medicine, Cardiff University Heath Park, Cardiff, CF14 4XN
2. Systems Immunity Research Institute, School of Medicine, Cardiff University Heath Park, Cardiff, CF14 4XN
3. Cardiff & Vale University Health Board, Heath Park, Cardiff, CF14, 4XN
4. School of Mathematics, Cardiff University, Cardiff, CF24 4AG

Introduction

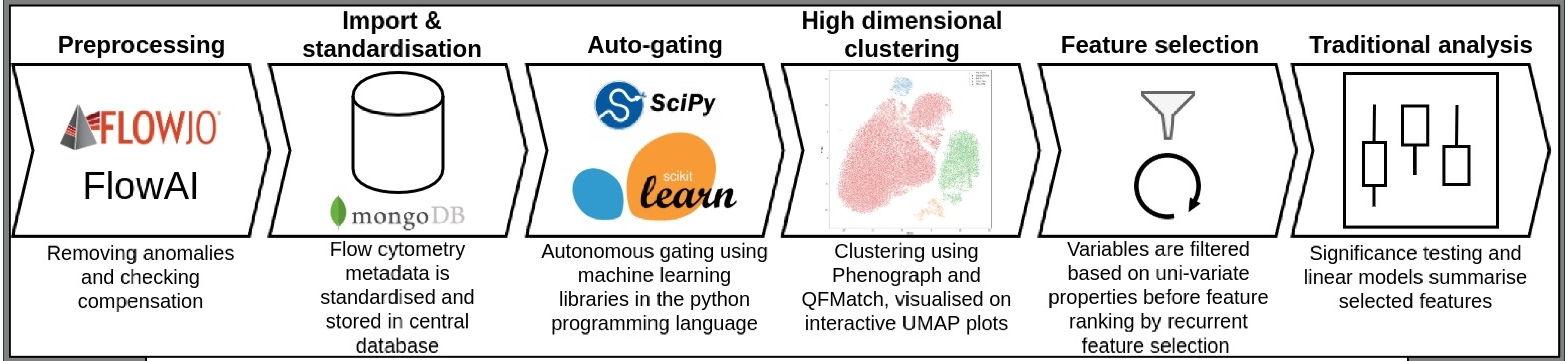
Clinical studies investigating the immune response in disease often involve complex flow cytometry analysis. The number of biomarkers investigated in any single study is increasing making traditional manual gating impractical. Furthermore this practice is both subjective and error-prone. In the past decade there has been a great effort to resolve this using techniques from the fields of bioinformatics and machine learning, however they remain inaccessible to the wider immunological community and do not address issues such as data management. Here we introduce Immunova, an analytical pipeline that aims to:

- Manage, standardise and store single cell, assay, and experimental meta-data.
- Automate traditional 'gating' by means of data-driven machine learning algorithms.
- Visualise, extract, and then select variables from high dimensional data that are significant to a clinical/experimental end-point.

Methods: Developing Immunova

Immunova is open source software built using the Python programming language. It's analytical steps are summarised in Figure 1.

- Python programming as opposed to alternatives such as R focuses on code readability and is considered "beginner friendly", making our solution more accessible.
- Central to it's design is a Document-Based Database; unlike tabular structures, data is stored in JSON format providing improved performance and greater flexibility.
- Gating has the advantage of interpretability but high-dimensional clustering in unbiased in the populations it uncovers. Immunova facilities both techniques for the generation of variables from flow cytometry data.
- Summary statistics from cell populations are filtered based on variability and then ranked according to their contribution to predicting a clinical/experimental endpoint of interest.



Preprocessing — Removing anomalies and checking compensation

Import & standardisation — Flow cytometry metadata is standardised and stored in central database

Auto-gating — Autonomous gating using machine learning libraries in the python programming language

High dimensional clustering — Clustering using Phenograph and QFMatch, visualised on interactive UMAP plots

Feature selection — Variables are filtered based on uni-variate properties before feature ranking by recurrent feature selection

Traditional analysis — Significance testing and linear models summarise selected features

Figure 1. Overview of the Immunova analytical pipeline. All stages following 'preprocessing' are housed within the Immunova software

Machine learning replicates manual gating and is 'data-driven'

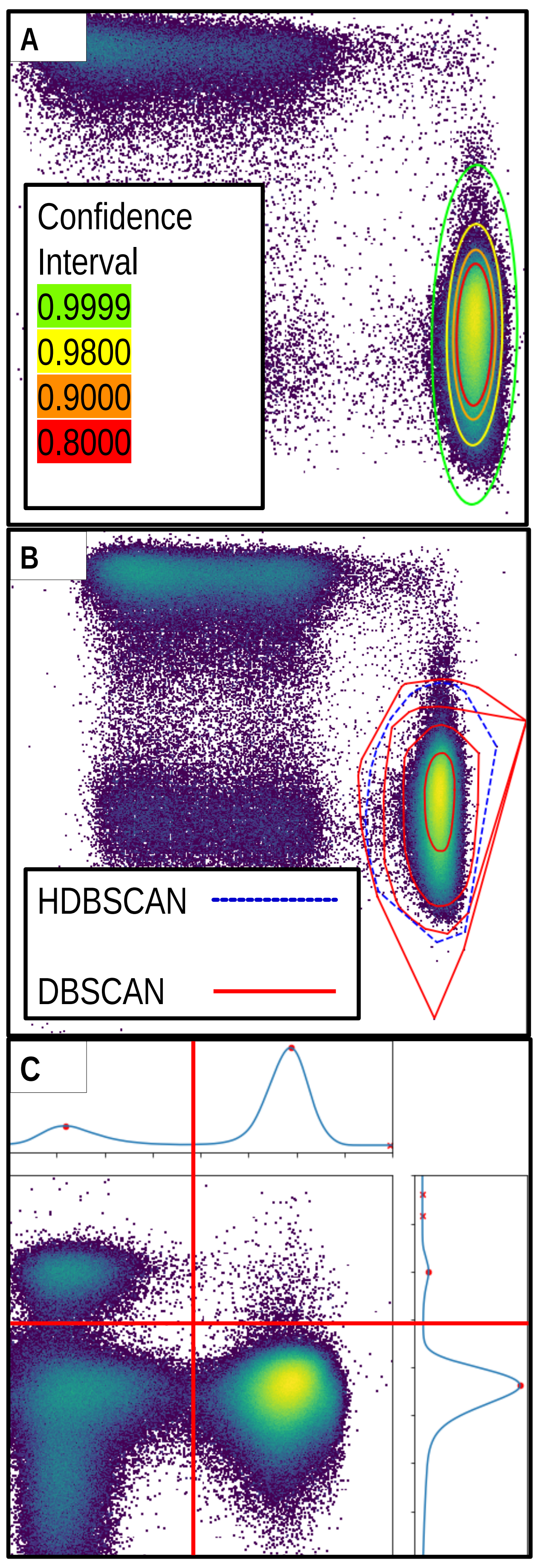


Figure 2. Overview of the four algorithms provided by Immunova for automated gating

Immunova provides four algorithms that can be used to build a custom gating template:

Gaussian Mixture Model (Fig. 2A)

- Probabilistic model that assumes the underlying data is derived from a finite number of gaussian distributions
- Suited to well defined populations
- User provides an estimated target centroid and a confidence interval (CI) that defines the 'tightness' of the resulting gate; Fig. 2A shows the result of choosing different CI

DBSCAN & HDBSCAN (Fig. 2B)

- Detection of clusters as areas of high density amongst regions of low density
- DBSCAN requires that the user specifies an estimated the maximum distance in which two cells would be classified as neighbours
  - This is both difficult and sensitive to subtle changes; red line in Fig. 2B shows variation with different input values
- HDBSCAN (dotted blue line Fig. 2B) does not require neighbour distance hyperparameter

Density Threshold (Fig. 2C)

- Identifies thresholds that segment populations using properties of the kernel density estimate in 1-dimensional space
- Red points in Fig 2C show identified peaks and red crosses are peaks excluded due to low intensity
- Threshold determined as the local minima between two density maxima

Autonomous gating matches the performance of a human expert

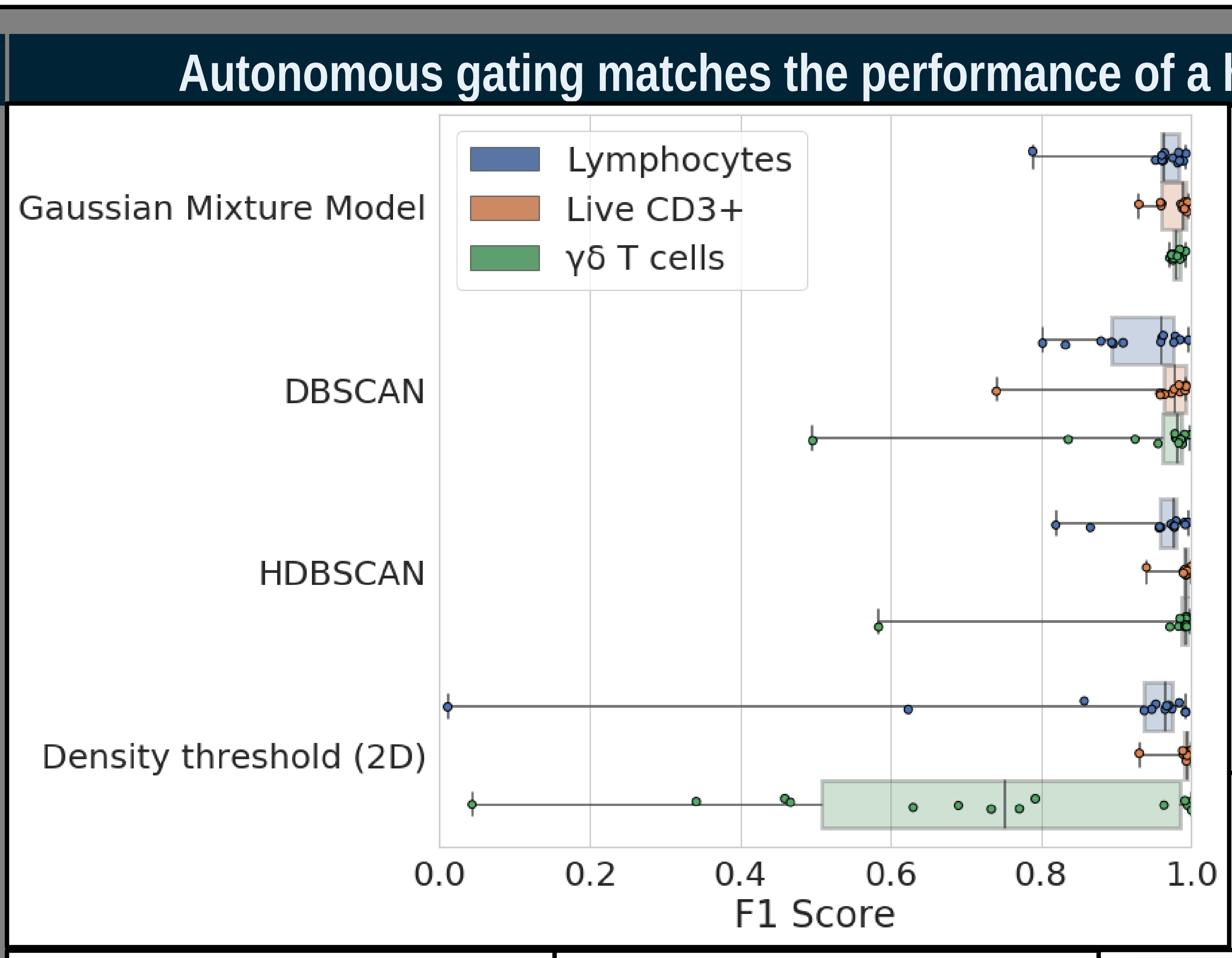
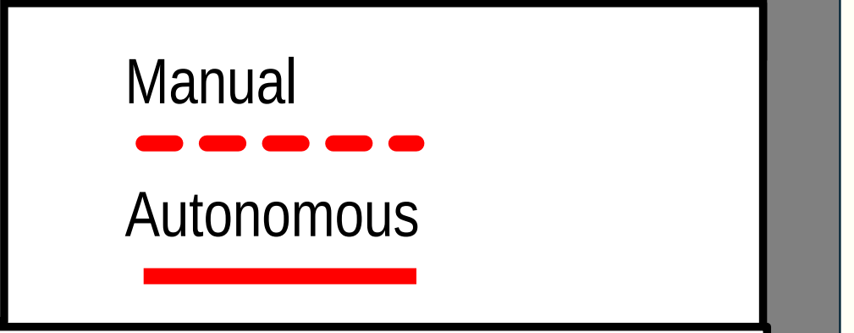
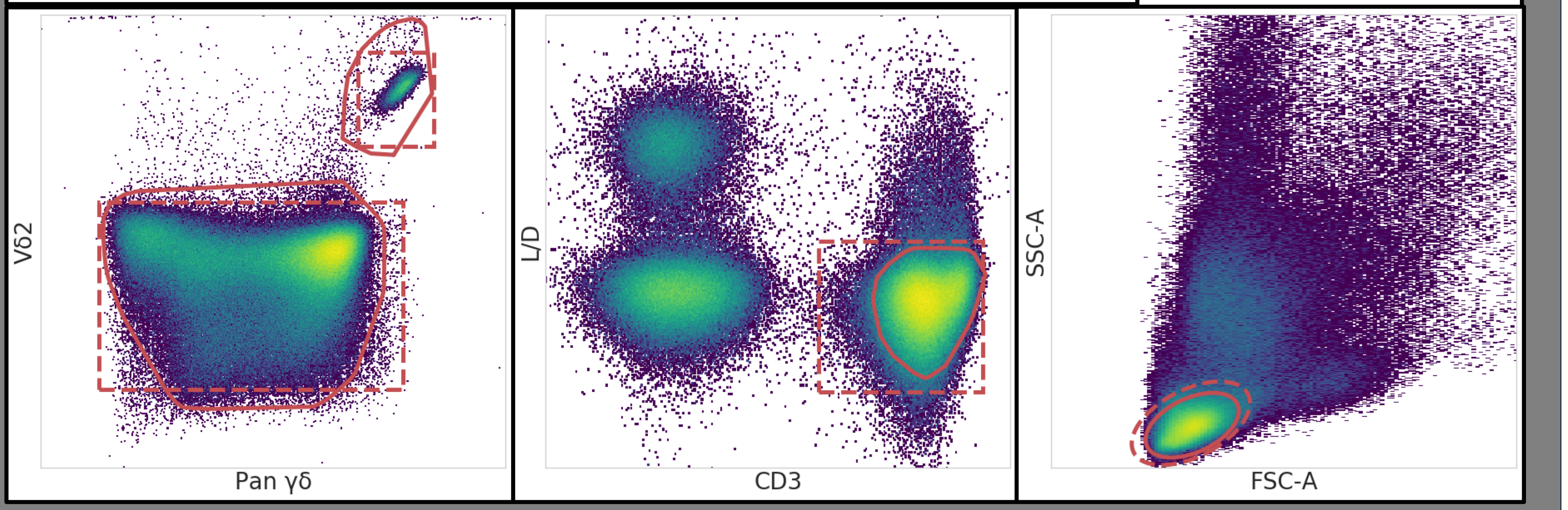


Figure 3. Comparison of autonomous gating algorithms to manual gating of lymphocytes (B), live CD3⁺ cells (C), and γδ T cells (D). Box plot (A) shows performance by F1 score across 14 separate flow cytometry experiments. Solid line in 2D histograms (B,C,D) show autonomous gate and dotted line manual gate by human expert.





- To establish a proof-of-concept, autonomous gating algorithms have been compared to manual gating by a human expert.
- Fig. 3 demonstrates that automated gating algorithms can replicate human identification of cell populations including rare subsets such as γδ T cells.
- Some algorithms provide greater performance than others e.g. the density threshold algorithm is not suitable for identifying γδ T cells because of ambiguity regarding the selection of two density maxima, whereas Gaussian Mixture Model and HDBSCAN consistently give good performance

Conclusions & Future Work

- Immunova is nearing the end of the development stage.
- Currently being applied to in-house datasets.
- We hypothesise that autonomous gating will reduce inter-sample variation when compared to manual gating.
- Results from Immunova will be contrasted against traditional analysis for validation.
- We believe that Immunova will provide an accessible alternative in flow cytometry analysis by introducing techniques from machine learning and data science into the immunological workflow.

Acknowledgements

We would like to thank the entire team in the Moser/Eberl group at Cardiff University for their help and contribution to this work. Special thanks to Oliwia Michalak, Alexander Greenshields-Watson, and John Pulford for discussion and critical review of this work.