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 involve complex flow cytometry analysis.
- Number of biomarkers investigated in any single study is increasing.
- Traditional manual gating is impractical, subjective and error-prone.
- Current technologies in flow cytometry bioinformatics are 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.

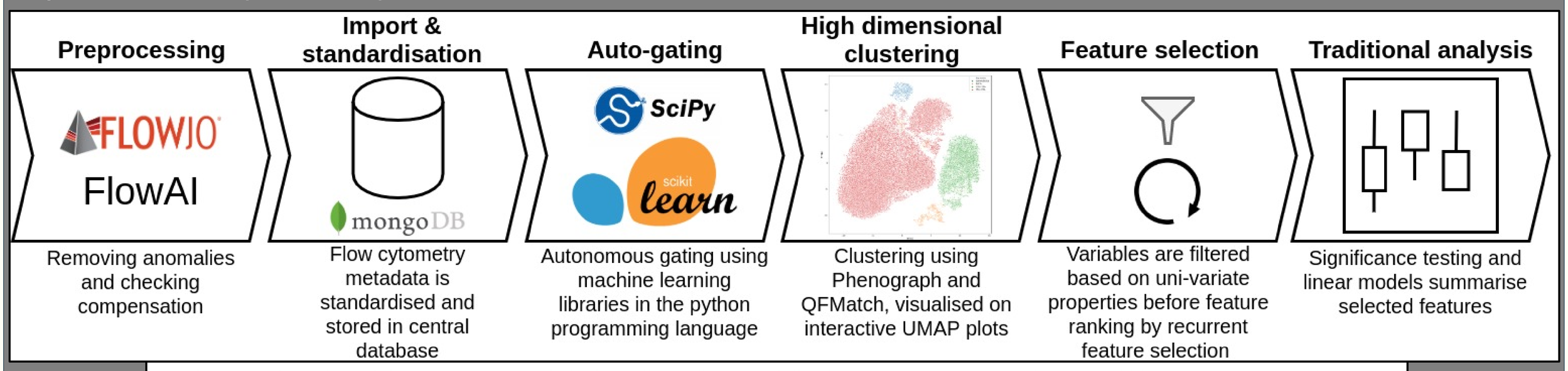

Figure 1. Overview of the Immunova analytical pipeline. All stages following 'preprocessing' are housed within the Immunova software

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

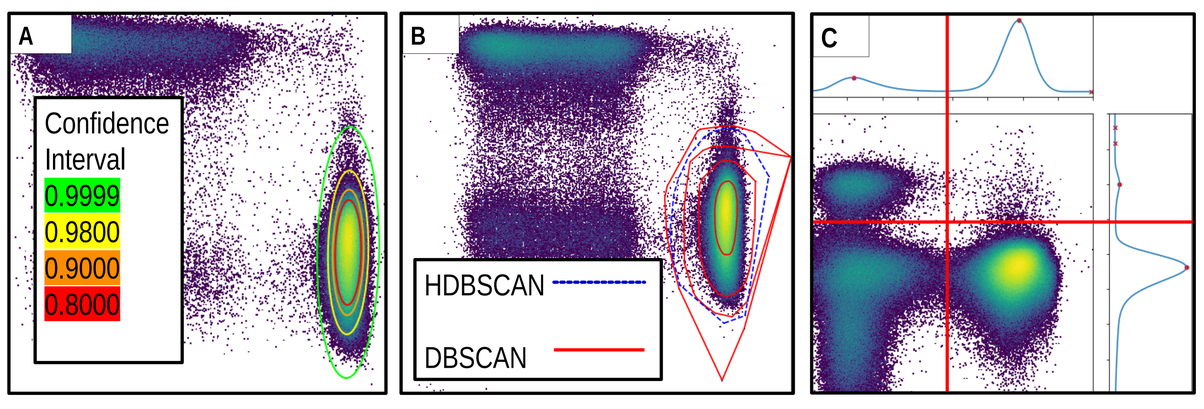

Figure 2. Overview of the four algorithms provided by Immunova for automated gating

Immunova supplies four algorithms for automated gating:

- **Gaussian mixture models (Fig 2A)** – probabilistic and influenced by a user supplied confidence interval.
- **DBSCAN & HDBSCAN (Fig 2B)** – density based clustering; DBSCAN is sensitive to choice of hyperparameters, HDBSCAN is less sensitive but is slower.
- **Density Threshold (Fig 2C) –** models the properties of 1-dimensional KDE using a peak finding algorithm and local minima calculation.

Autonomous gating matches the performance of a human expert

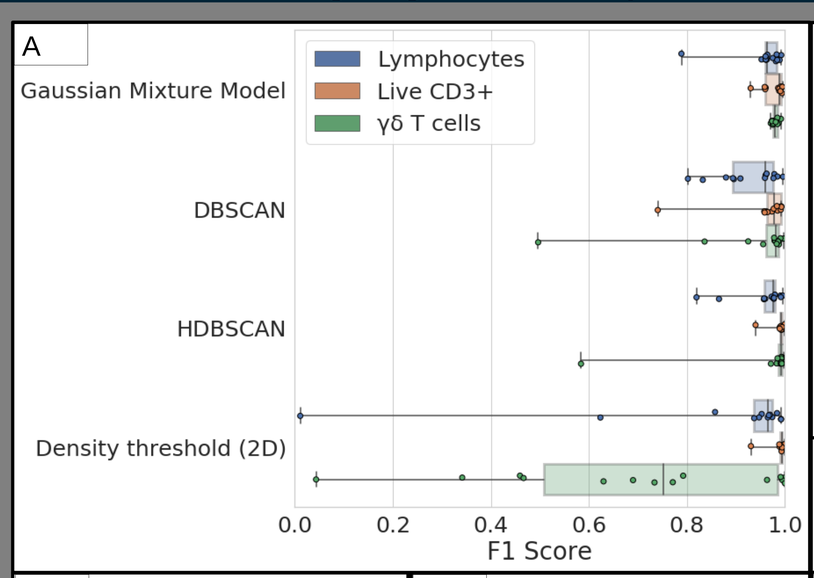

Figure 3. Comparison of autonomous gating algorithms to manual gating of lymphocytes (B), live CD3⁺ cells (C), and γδ T cells (D). Box plot (A) shows performance by F1 score across 14 separate flow cytometry experiments. Solid line in 2D histograms (B,C,D) show autonomous gate and dotted line manual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

Manual

Autonomous

- 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, whereas Gaussian Mixture Model and HDBSCAN consistently give good performance

Automated gating is combined with high dimensional clustering

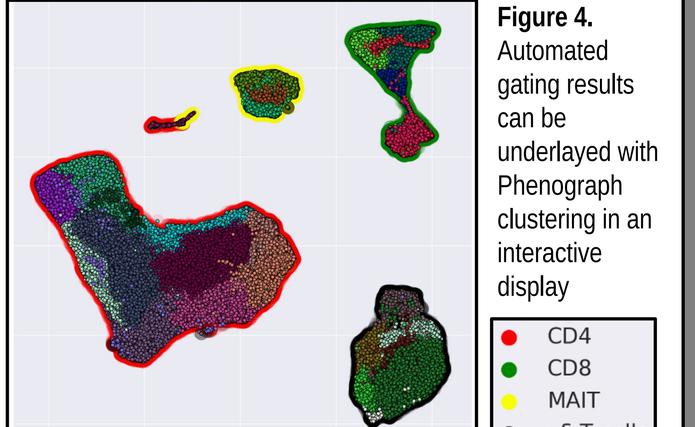

Figure 4. Automated gating results can be underlayed with Phenograph clustering in an interactive display

- Contrast automated gating to **Phenograph clustering** to reveal possible biases in gating strategy
- Immunova provides an **interactive display** to explore high dimensional plots such as Fig. 4
- Merge clustering results between patients using **QFMatch**

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

This research was supported by the Cardiff University School of Medicine Studentship (RJB) and MRC project grant MR/N023145/1. Special thanks to Oliwia Michalak, Alexander Greenshields-Watson, and John Pulford for discussion and critical review of this work.