Clustering Methods for Flow Cytometry

Ultán Doherty

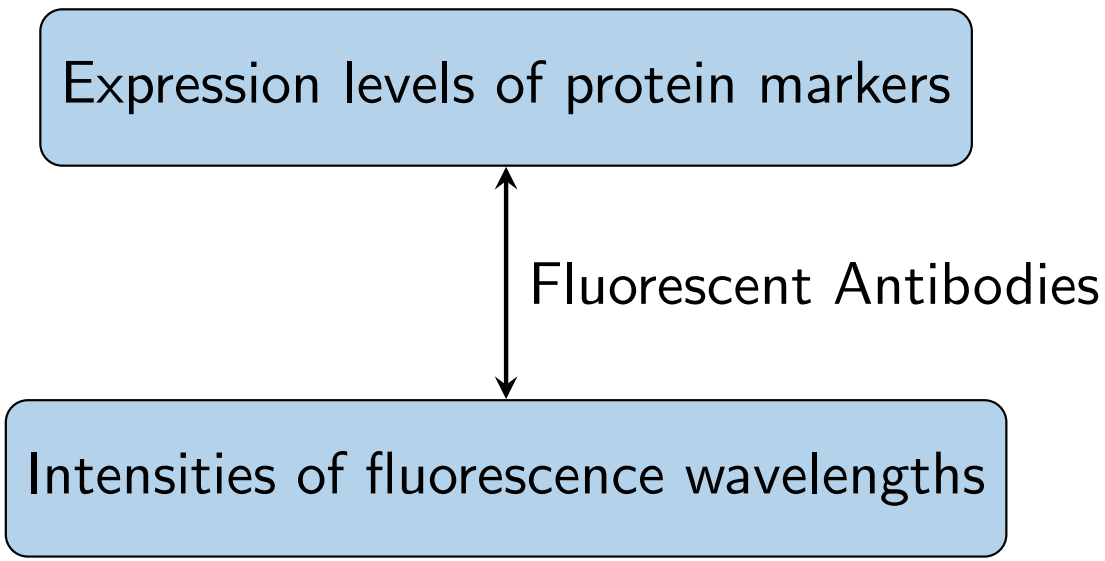
Supervisors: Arthur White & Rachel McLoughlin

Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin

IRISH RESEARCH COUNCIL
An Chomhairle um Thaighde in Éirinn

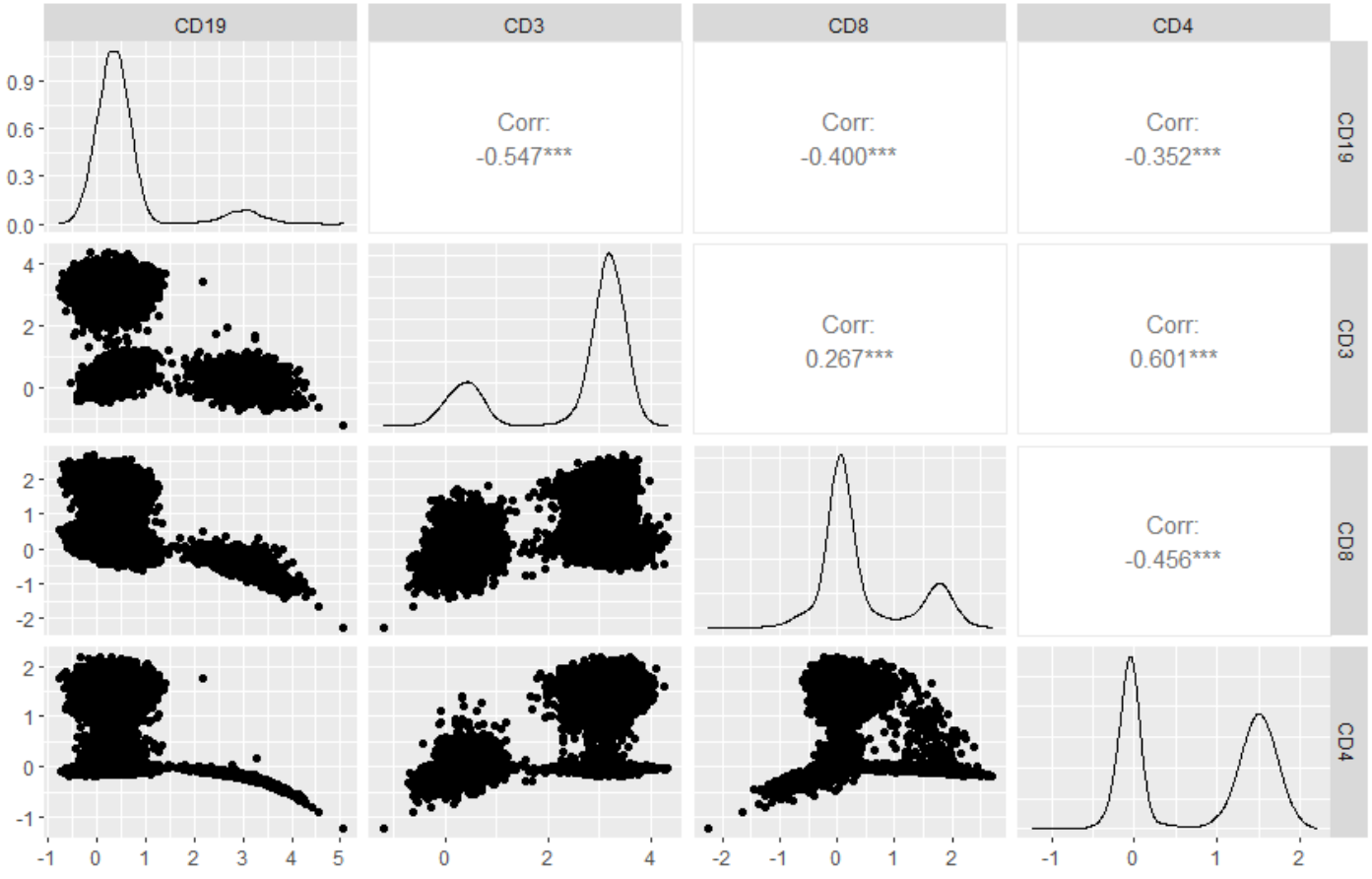
Principle of Flow Cytometry

- Flow Cytometry analyses large numbers of cells to determine their expression levels of protein markers.



- Antibodies bind to specific protein markers on cells.
- Fluorochromes emit unique fluorescence wavelengths.

Visualising Flow Cytometry Data

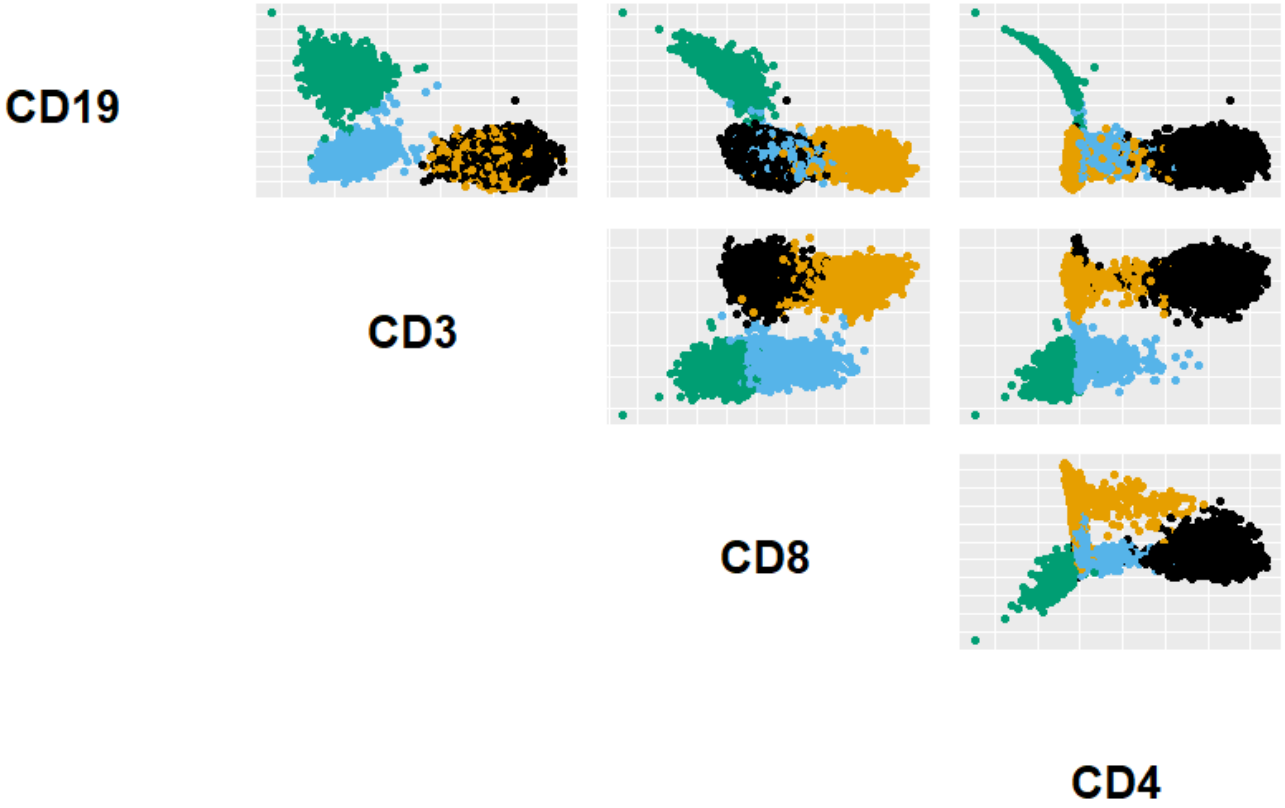


High CD19 $\Longrightarrow$ B-Cells

High CD3, High CD8 $\Longrightarrow$ Cytotoxic T-Cells

High CD3, High CD4 $\Longrightarrow$ Helper T-Cells

Cell Population Identification



High CD19	$\Longrightarrow$	B-Cells	= Green
High CD3, High CD8	$\Longrightarrow$	Cytotoxic T-Cells	= Orange
High CD3, High CD4	$\Longrightarrow$	Helper T-Cells	= Black

Sequential Manual Gating

1. Visualise the data in a two-dimensional plot.
2. Manually draw a boundary around the desired cells.
3. Subset the data using this boundary.
4. If necessary, repeat Steps 1-3 with only this subset.

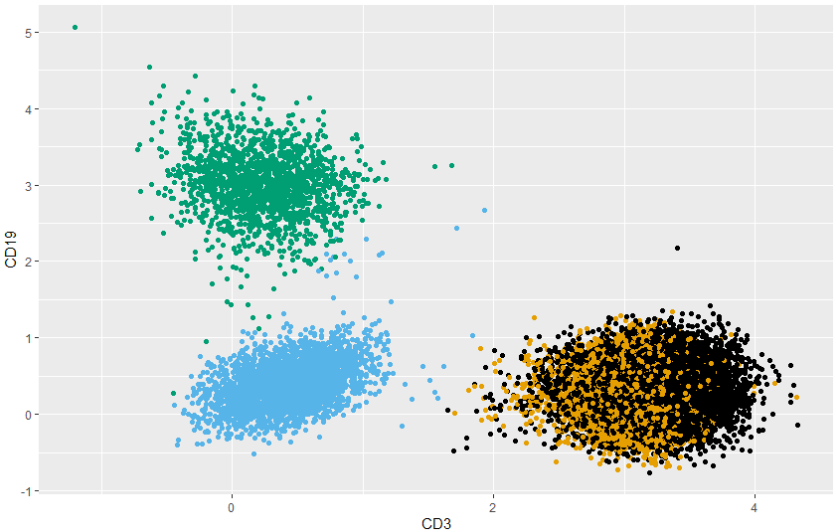


Figure: B vs T

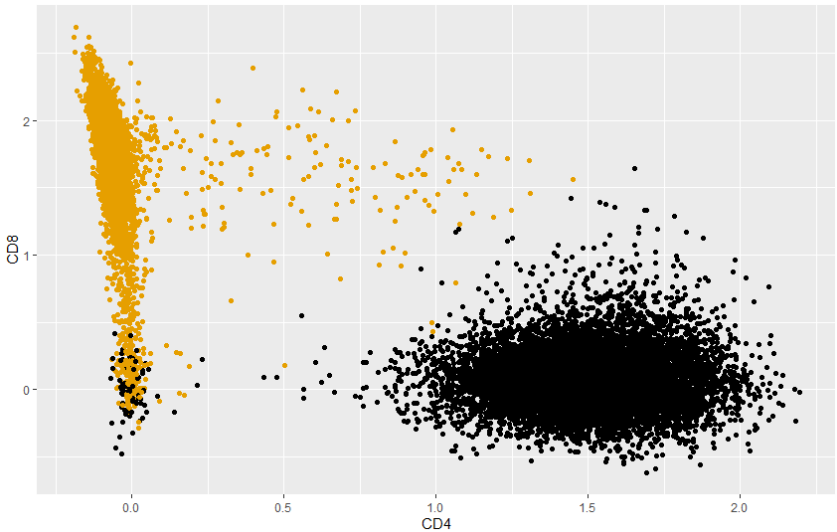


Figure: Cytotoxic T vs Helper T

Problems with Sequential Manual Gating

- Manually drawn boundaries are biased and irreproducible.
- The sequence identifies cells with known properties.
- 2D plots do not utilise the high-dimensionality of the data.
- Subsetting can exclude interesting populations.
- Gating is time-consuming and complex for high dimensions.

Automated Cell Population Identification

- Significant research has been carried out on developing automated population identification methods.
- However, a 2020 survey revealed that 53% of laboratories never use automated flow cytometry software to identify cell populations.
  (UK NEQAS Leucocyte Immunophenotyping survey; Cheung et al, 2021)

Difficulties with Flow Cytometry Data

- Manual gating is an unreliable benchmark for comparison.
- Populations can be skewed, heavy-tailed, and non-convex.
- Modern Flow Cytometers can measure 30-50 markers and Mass Cytometers can measure up to 100 markers.
- Each sample contains $\sim 10^5$ cells.
- Technical and biological variation exists between samples.
- Population identification must be compatible across multiple samples.

Stages of FCM Analysis

1a. Clustering & Meta-Clustering
- Identify cell populations / clusters in each sample.
- Match corresponding clusters in different samples.

1b. Joint Clustering
- Identify clusters in all samples simultaneously.

2. Regression
- Predict a clinical outcome based on cluster properties.

My Research

Current Work

- Literature Review:
  - Studying existing methods and review papers.

Future Work

- Mixed Membership Model:
  - Developing a population identification method.