Distribution Agreement

Signature:

____________________________ ______________

John Andrew Pura Date

Development and Validation of a Computer-Aided Multidimensional Flow Cytometry Analysis Pipeline

By

John Andrew Pura

Master of Public Health

Department of Biostatistics and Informatics

____

Joel H. Saltz, MD, PhD

Committee Chair

Development and Validation of a Computer-Aided Multidimensional Flow Cytometry Analysis Pipeline

By

John Andrew Pura

Bachelor in Science in Engineering

Duke University

2008

Thesis Committee Chair: Joel H. Saltz, MD, PhD

An abstract of

A thesis submitted to the Faculty of the
Rollins School of Public Health of Emory University

in partial fulfillment of the requirements for the degree of
Master of Public Health
in Biostatistics

2012

Abstract

Development and Validation of a Computer-Aided Multidimensional
Flow Cytometry Analysis Pipeline

By John Andrew Pura

Abstract

Flow cytometry (FCM) is a popular technique, in basic and clinical research, for the high-throughput characterization of cellular properties. Modern FCM instruments are now capable of measuring up to 20 characteristics of an individual cell, resulting in a rich array of multidimensional information on hundreds of thousands of cells. Data analysis, however, remains a challenging aspect of FCM research; the ability to acquire large multidimensional datasets has outpaced the ability to accurately and reproducibly detect expected and novel cell populations and efficiently test biological hypotheses.

The main objective of this work is to develop and validate a mechanistic pipeline that addresses the challenges found in high-throughput, multidimensional FCM data analysis. My pipeline development employs state-of-the art biomedical informatics software that can be integrated with clinical outcomes.

The pipeline consists of five key steps: 1) data preprocessing, 2) automated gating, 3) automated labeling, 4) feature extraction, and 5) feature selection. A robust quality assessment step was implemented at every step of the pipeline to account for potential sources of systematic error prior to entering a subsequent step in the pipeline. Validation against human expert analyses showed good detection of expected cell populations. The integration of state-of-the art informatics techniques into a single pipeline shows great potential for scalability to hundreds of thousands of events and multiple dimensions.

In particular, this pipeline can be used to assess a transplant recipient's immune repertoire over time in single patients and in a cross-sectional patient population with diverse diagnostic and demographic characteristics. The main assumption in the proposed pipeline is that the mechanisms driving complications in organ transplantation intersect in a way that is both anticipatable and measurable.

A thesis submitted to the Faculty of the

Rollins School of Public Health of Emory University

in partial fulfillment of the requirements for the degree of
Master of Public Health
in Biostatistics

2012

Table of Contents

Chapter 1

Introduction

Flow cytometry (FCM) is a powerful tool for rapidly quantifying the chemical and physical properties of up to millions of cells stained with monoclonal antibodies conjugated to fluorescent dyes. Flow cytometers employ light sources (lasers) that probe a specific property of an individual cell based on light scattering and fluorescence, and detectors that measure each property. Today, advancements in FCM technologies, such as the ability to process thousands of cells per second (high-throughput) and multiple detectors and lasers, have made it suitable for detecting up to twenty properties [1, 2], resulting in a rich array of information on a large number of cell events.

The popularity of FCM in basic and clinical research has led to the discovery of interesting links between various pathologies and their underlying molecular mechanisms, and has helped to elucidate a network of relationships between different diseases. For example, FCM has helped to identify functional antigenic markers associated with stem-cells [3], graft-versus-host-disease [4], lymphoma subtypes [5], and early progression of human immunodeficiency virus (HIV) [6]. It has also been an important tool for immunophenotype (i.e. type of cell population) matching during organ transplantation [7].

In this chapter, I provide an overview of FCM technology, the current methodology for data analysis, and the challenges often encountered in data analysis as perceived by the FCM community. Furthermore, I propose a computer-aided analysis pipeline to address these challenges. In Chapter 2, I discuss in detail the individual components of the pipeline. Finally, chapter 3 presents the validation of this pipeline using patient FCM data from a clinical study investigating the effect of immunotherapy on ragweed-induced allergic rhinitis. Furthermore, I discuss the implications of this pipeline for exploratory of analysis of viral activation and T-cell repopulation in a longitudinal clinical study of renal transplant patients.

1.1 FCM Technology and Data Analysis

1.1.1 Overview

Research utilizing FCM technology generally employs the sequential strategy depicted in the schematic below (Figure 1). The process begins with sample preparation of the cells (1.1.2). This includes staining procedures to tag cells with fluorescence-antibody probes. Data is acquired using a flow cytometer which provides information for each cell (1.1.2). Data analysis focuses on identifying and classifying the cells through a cell extraction process known as gating (1.1.3). Finally, the results of the analyses can be associated with a set of biological hypothesis that drive the data collection.

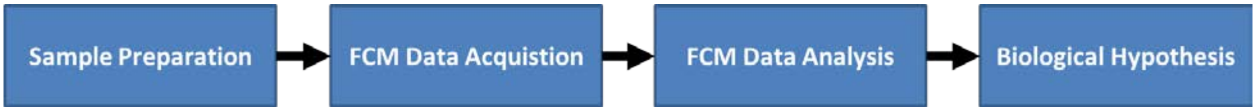


Figure 1. *Basic workflow for FCM analysis*. Using FCM in a research framework typically includes the above steps: 1) sample preparation (Section 1.1.2), 2) data acquisition via a flow cytometer (Section 1.1.2), 3) data analysis (Section 1.1.3), which includes cell

population extraction (gating), and 4) correlation of extracted populations to a set of biological hypotheses.

1.1.2 Sample Preparation and Data Acquisition

In FCM, cells expressing a set of desired cell surface or intracellular receptors (antigens) are initially suspended in a solution containing a cocktail of fluorochrome-conjugated antibodies. Fluorochromes (e.g. fluorescein isothiocyanate, FITC) are dyes, which absorb light at a certain wavelength, and re-emits light energy at different wavelengths. These can be specifically bound to a particular monoclonal antibody, such as anti-CD3, which identifies the CD3 antigen present on T-cells. After staining, the cell suspension is then forced into the path of laser beams, such that only one cell is illuminated at a single time (i.e. a cell event). As the cells pass through the lasers, detectors (channels) capture signals corresponding to the scattered and fluoresced light of each cell. Forward scatter channels (FSC) measure incident light scattering, which is proportional to the cell's size, while side scatter channels (SSC) measure orthogonal light scattering, which is proportional to the cell's granularity or internal complexity. Light that is absorbed and emitted by the cells is captured by additional fluorescent detectors, which are specifically configured for particular range of wavelengths. The fluorescence intensity per cell corresponds to the detection of fluorochrome-bound antibodies on the cell surfaces. Figure 2 shows a standard flow cytometer setup, including the sample input and detectors for data acquisition.

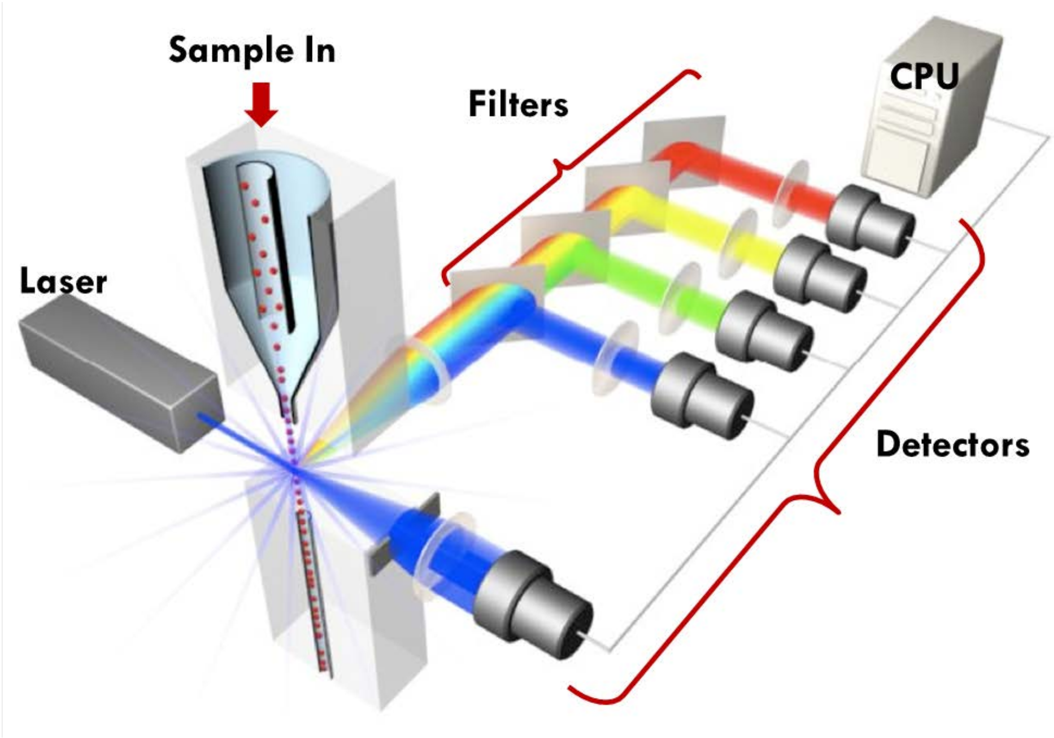


Figure 2. *Standard flow cytometer setup.* A suspension of stained cells is hydronamically focused within the flow cytometer. A laser beam excites the fluorescent dyes attached to a cell, which results in light re-emission at different wavelengths. Filters screen out specific wavelengths of re-emitted light, which is then passed into detectors that measure fluorescence intensity per cell. (Image from http://www.invitrogen.com).

In FCM sample preparation, it is a common practice to separate a single sample of peripheral blood from a patient into multiple tubes (also known as aliquots), which can then be treated with different antibody cocktails. Typically a certain set of antibodies in a cocktail is kept constant for in a multi-tube experiment, while others are adjusted in order to accommodate a variety of different cell populations or immunophenotypes (expression of antigens). Table 1 shows a typical panel used to describe the possible cell populations, which can be identified within a set of tubes in an experiment. Not included in the panel in Table 1 are the forward and side scatter channels, which are common to all tubes, but are not used to detect antigenic markers.

	Fluorochrome								
Tube	FITC	PE	PE-cy7	APC	Alexa700	APCCy7	V450	Pac Orange	Qdot655
Comps	-	CD20	CD197	HLA-DR	CD3	CD8	CD4	CD20	CD45RA
1	CD45RO	γδTCR	CD197	αβTCR	CD3	CD8	CD4	CD14/CD20	CD45RA
2	CD57	CD279	CD197	CD127	CD3	CD8	CD4	CD14/CD20	CD45RA
3	CD11a	CD279	CD197	CCR5	CD3	CD8	CD4	CD14/CD20	CD45RA
4	CD2	CD28	CD197	CD27	CD3	CD8	CD4	CD14/CD20	CD45RA
5	CD69	CD38	CD197	HLA-DR	CD3	CD8	CD4	CD14/CD20	CD45RA
6	CD28	CD137	CD197	-	CD3	CD8	CD4	CD14/CD20	CD45RA
7	CD103	CD31	CD197	PTK7	CD3	CD8	CD4	CD14/CD20	CD45RA
8	CD80	CD86	CD197	CD127	CD3	CD8	CD4	CD14/CD20	CD45RA

Table 1. *Panel for studying different cell types in a FCM experiment*. Fluorochromes are listed on the topmost row, and the antigens (e.g. CD20, HLA-DR, etc.) identified by each fluorochome are listed across a row for each tube. The first tube (Comps) is the compensation tube, which is used to subtract overlapping fluorescence signatures from different fluorochromes during data analysis. FSC and SSC information are excluded from the panel.

1.1.3 Data Analysis

The traditional approach to analyzing flow cytometry data is first identifying biologically similar groups of cells based on physical similarities or a unique expression of antigens, a process known as gating, and then comparing these groups of cells with a set of biological hypotheses. The commercial software, FlowJo (http://www.flowjo.com; Tree star, Ashland, Oregon) has been widely used in gating, visualization, and analysis of data from flow cytometry experiments. FlowJo use is primarily manual and visual, restricting the user to analyzing at most two dimensions simultaneously to isolate particular cell subsets**.** Here, we define a *dimension* to be a single channel that detects either a physical (e.g. FSC) or chemical (e.g. FITC) property of a cell. A panel, such as Table 1 above, can be used to detect all possible populations for a single tube. A typical gating strategy begins with visualizing one-dimensional histograms or two-dimensional dotplots, and proceeds with sequential delineation of cell population boundaries based on the light scattering properties (FSC and SSC), followed by fluorescent information. A hierarchical strategy is employed, by first finding regions of cells with less differentiated surface antigens (e.g. lymphocytes) and analyzing these mature cell subsets under additional projections of fluorescence dimensions to capture more differentiated cell subpopulations (e.g. effector memory T helper cells). Figure 3 shows a typical manual gating pipeline:

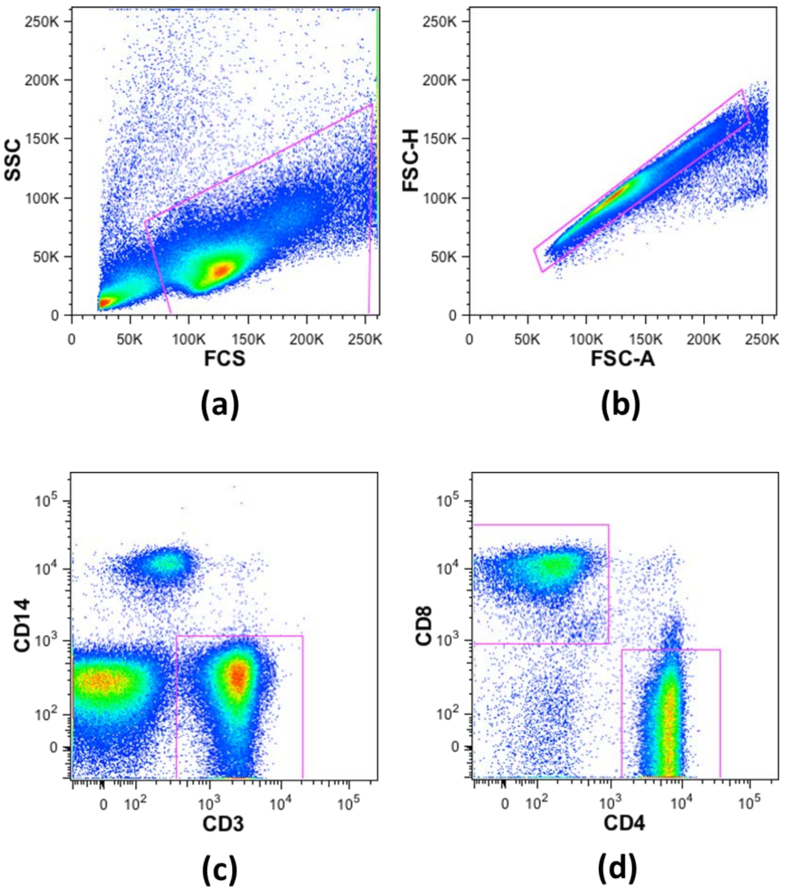


Figure 3. *Sequential manual gating using FlowJo software.* Sequential gating follows a hierarchical procedure that looks at one or 2-D plots at a single time. a) A typical gating analysis first examines a 2-D plot of the the scatter channels to isolate the lymphocyte and monocyte population. b) Singlets are then separated from doublets by examining the FSC-H vs FSC-A channels. c) CD3+CD14- cells are then identified. d) CD4+ and CD8+ populations can then be isolated form the CD3 cells.

Unfortunately, the ability to rapidly collect multidimensional FCM data and the increase in applications for such technology has exceeded the ability of commercially available tools, such as FlowJo, and researchers to efficiently discover and test biological and clinical hypotheses in large FCM data [8]. Consequently, FCM data analysis has been a major bottleneck in research. The following section discusses the challenges often faced by researchers in the analysis of FCM data.

1.2 Challenges in Data Analysis

While it has been effectively used in smaller FCM data sets with three to five dimensions, the traditional manual, qualitative approach may not be feasible for the large data sets produced by high-throughput, multidimensional FCM experiments. Visualization and analysis of multiple dimensions is difficult, laborious, and subject to the users' expertise. More importantly, multidimensional relationships and features may be missed during two-dimensional sequential gating. Several other challenges related to multidimensional FCM analysis have also been recognized by the FCM community, regarding quality assessment, population gating and labeling, and feature selection [1, 9].

1.2.1 Quality Assessment

Quality assessment (QA) is an integral step in any high-throughput FCM analysis framework, as it allows the researcher to potentially identify, investigate and remove any measurement variability from any experimental, non-biological sources [9, 10]. For example, not accounting for changes in instrumentation settings or sample preparation may lead to erroneous conclusions regarding cell population identification and tracking.

Furthermore, the detection of multivariate outliers or boundary events is an important step prior to gating and statistical analysis of the data. These events may arise from unusual cell events such as non-viable cells (e.g. debris, red blood cells), doublets or those reaching the detection limits of the flow cytometer along any subset of fluorescence channels. Outliers may significantly alter the gating results as they may overestimate the number of cell populations, particularly in exploratory analyses, where the goal is to discover novel cell populations [9].

Reliable QA processes should also be able to ensure consistency of cell event distribution among tubes with shared antigen profiles (e.g. Alexa700 and V450, which detect CD3 and CD4, respectively as seen in Table 1). This is especially true for multi-tube experiments in which each tube contains cell events derived from a single sample.

1.2.2 Population Gating

A critical bottleneck in FCM data analysis lies in gating populations. Manual gating has been successfully applied to numerous studies analyzing FCM data, and has been useful in cases where an expert's contextual information is needed to account for confounding biological and non-biological factors during gating or for gating rare cell populations. However, applying it to a high-throughput paradigm is ultimately difficult due to several well-recognized drawbacks [11, 12]: 1) the choice of gate regions is subjective and based on an analyst's experience. Informative multivariate features may be missed during the gating process, and the resulting variability among analysts may lead to irreproducible gating; and 2) the current process is time-consuming and labor-intensive. In addition, with multidimensional FCM data, it may be of interest to the researcher to identify novel populations for biomarker discovery, rather than define a set of known populations.

1.2.3 Population Labeling

After gating target cell populations (often performed using clustering methods) it is often desired to label and compare cell populations across samples, subjects or timepoints. While labeling may be accounted for in the gating step, reliable tracking and comparison of labeled populations remains a challenging in the analysis of FCM data. The labeling step is usually manually done and based on cell locations or mean intensities [13]. This is problematic, as a gating that is valid for one sample or timepoint, may not be valid for an alternate sample or timepoint. Standardization of FC samples, such in a multi-tube experiment, is one way to ensure that similar cell events are well resolved for better classification of clusters. This can be accomplished through data normalization, which matches the distributions across samples. However, such a procedure may normalize out biologically motivated differences across samples. Therefore, a robust checking mechanism should also be integrated into the population labeling/matching procedure.

1.2.4 Feature Selection

Another recognized challenge in FCM data analysis is the selection of appropriate features (e.g. cell population frequencies, antigenic markers, antigenic ratios) for predictive analyses and biologically-driven hypotheses. It is often a goal in FCM research to find the most informative set of FCM features that relate to a set of clinical outcomes. Inadequate feature selection may complicate diagnoses of certain diseases, such as lymphoma, where a high number of antigenic markers are used to differentiate between lymphoma sub-types [5].

1.3 Proposed Solution – A Computer-Aided Analysis Pipeline

To address the challenges presented in this chapter, I propose an informatics approach to the development of a computer-aided analysis pipeline for FCM data with the following components: quality assessment, data preprocessing, automated gating, automated labeling, feature extraction and feature selection. Quality assessment will be embedded at every step of the pipeline to identify upstream sources of non-biological errors. It is comprised of error checking procedures to avoid inconsistent results in downstream processes. The goal of data preprocessing is to facilitate automated gating and labeling procedures through proper data scaling and removal of biological and statistical outliers. Automated gating focuses on clustering the multidimensional space into biologically meaningful groups. Labeling assigns cell expression profiles to each cluster by analyzing the fluorescence information of each cluster. Feature extraction analyzes statistical features of each cluster, such as mean fluorescent intensity, cluster size, or shape. Finally, the goal of feature selection is to select informative features that are associated with particular clinical outcomes.

1.4 Organization of Thesis

The remainder of this thesis focuses on the development, validation, and application of the FCM data analysis pipeline. Chapter 2 discusses the development of the computer-aided pipeline and describes the software used for quality assessment, data management, and data analysis. Chapter 3 presents validation of the pipeline using real data from a clinical study on seasonal rhinitis. Automated results are compared to human-generated results in the validation. Finally, I present the driving biological project of organ transplantation, and how multidimensional FCM analysis can be used to investigate cross-sectional and longitudinal patterns in cell types relating to the risk of rejection and viral infection, and cell repopulation.

Chapter 2

Introduction

The ability to acquire large multidimensional datasets has outpaced the ability to efficiently classify cell types and assign clinical correlates with regard to their implications for risk. Furthermore, manual FCM analysis focuses on at most two dimensions simultaneously, potentially missing biological information from other dimensions. This highly qualitative and visual approach also has limited ability to accurately and reproducibly detect expected as well as novel cell populations. As a result biomedical informatics (BMI) software, dedicated to FCM data, have been developed to transition to a statistical-based and quantitative approach that can leverage the entire multidimensional space, analyze the data in a reasonable amount of time, and reduce variability in results.

One such BMI initiative is the Bioconductor Project (http://www.bioconductor.org) [21], an open software environment built on R statistical programming language (http://www.R-project.org). A suite of state-of-the-art R packages are available in the Bioconductor and the Comprehensive R Archive Network (CRAN) repository for a wide range of FCM analysis uses: data management in *flowCore* and *flowFlowJo*, statistical processing in *flowStats*, quality control in *flowFP* and *flowQ*, data visualization in *flowViz*, gating (known as clustering in a BMI context) in *flowClust/flowMerge*, *flowMeans*, and *SamSPECTRAL*, immunophenotyping using *flowType*, and feature selection using *FeaLect*.

Several other algorithms have also been proposed and presented through the FlowCAP project (http://flowcap.flowsite.org), an international collaboration to develop and objectively test computational tools for identifying cell populations in FCM data. For the past two years, challenges have been held to compare these tools against manual findings, and assess their performance in predicting clinical outcomes.

2.1 Overview of Proposed Pipeline

As discussed in Chapter 1, research endeavors utilizing FCM traditionally follows a sequence of steps from sample preparation and data acquisition to data analysis and relating these analyses to biologically-motivated hypotheses (Figure 1). In this work, I focus primarily on the data analysis block, which is a widely recognized bottleneck in clinical research. This limitation is particularly pronounced for large, multidimensional data. Current methods rely largely on the user's experience and are inefficient for volumes of data. Manual gating of FCM data is time-consuming and can be limiting for exploratory analyses or diagnostic studies. Furthermore, the lack of adequate quality control at each step of the analysis can potentially waste valuable time and resources and lead to spurious conclusions.

Despite the growing number of biomedical informatics tools for automated processing of multidimensional FCM data, there is little published information regarding integration, evaluation and validation of the ability of these tools in a clinical context [6, 9, 22-24]. **The main objective of this work is to develop and validate a mechanistic pipeline that addresses the challenges found in high-throughput, multidimensional FCM data analysis.** Furthermore, my pipeline development employs state-of-the art BMI software that can be integrated with clinical outcomes. Validation against expert analyses ensures proper functionality with respect to the biological problem.

The pipeline shown in Figure 4 is divided into two main parts. **Quality assessment** on the right is performed after each of the main steps on the left. To accomplish QA, I utilize a cytometric fingerprinting algorithm to identify informative regions in the data. These informative regions or "fingerprints" can then be used to compare distributions of cell events or quality of gating. **Data preprocessing** is accomplished through transformation of the scatter and fluorescence channels, quantile normalization of fluorescence channels, and the removal of non-viable cells and doublets using spectral clustering. **Automated clustering and labeling** are accomplished using k-means partitioning of individual fluorescence channels and combining these partitions into multidimensional clusters. **Feature extraction** is performed through statistical analysis of cluster properties. Finally, **feature selection** is accomplished through a robust classifier based on the popular regularization technique, LASSO.

R statistical programming software available in the Bioconductor project facilitated the processing of the high-throughput FCM analysis pipeline**.** All processes are tightly integrated through the Bioconductor framework, which allows for such a pipeline to be easily merged with clinical ontologies.

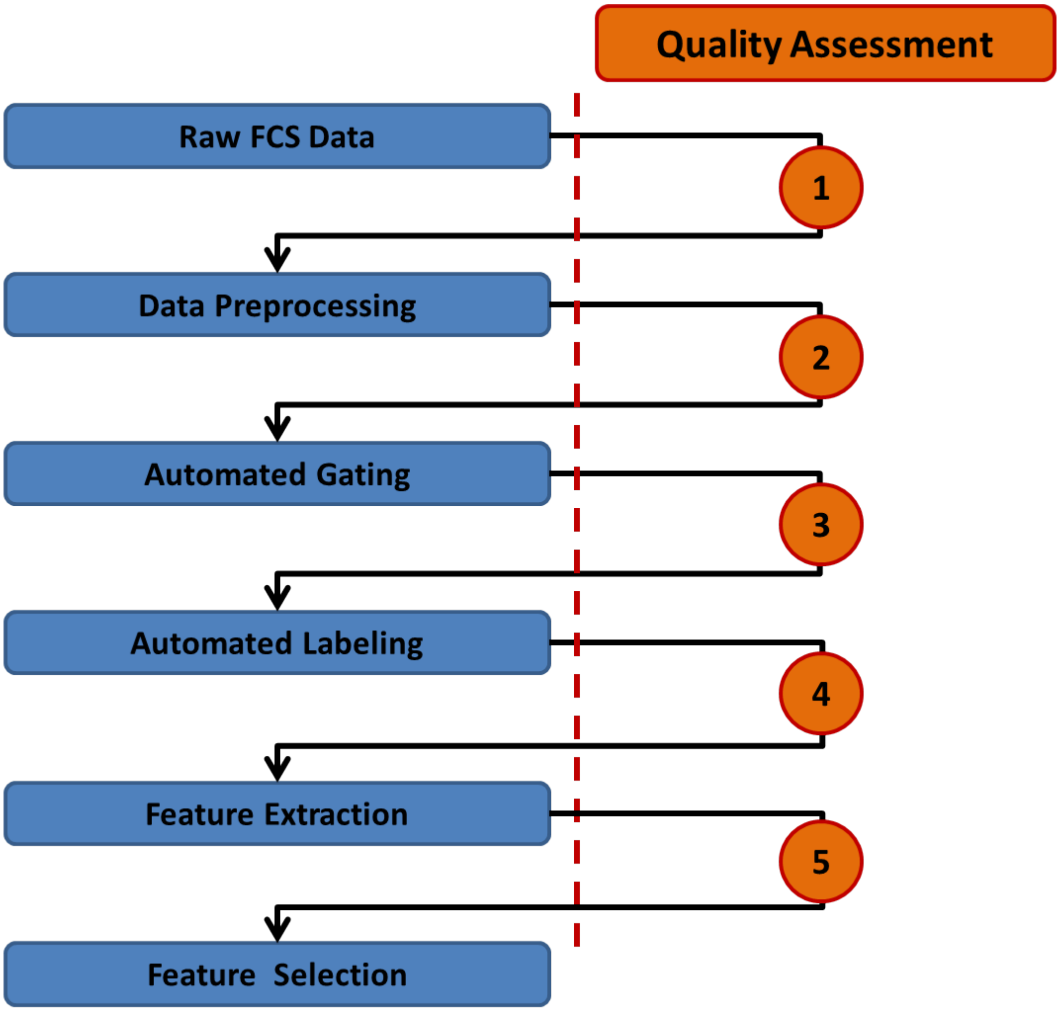
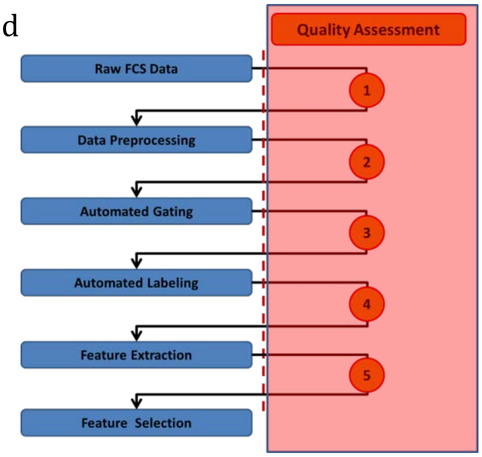


Figure 4. *Proposed FCM analysis pipeline*. The left panel shows the main computational components – from raw FCM data to feature selection. The output of each component is passed to quality assessment prior to entering successive stages of the pipeline. Quality assessment serves as intermediate steps, which controls for: 1) compensation, boundary/margin events, consistency across tubes in raw data; 2) consistency across tubes for viable cells, 3) and 4) biological relevance of desired populations, and 5) statistics of cell population proportions (features)

2.2 Pipeline Components

2.2.1 Quality Assessment



Quality assessment (QA) was incorporated throughout the analysis pipeline, using the available functions in the *flowQ* and *flowFP* packages. QA processes serve as intermediate steps between the main components, and serve to identify upstream, non-biological sources (e.g. experimental setup, computed processing) of error. After **data compensation** and **transformation** of fluorescence channels, boundary and margin events were visualized using *flowQ* and potentially removed for specific channels. **Boundary/margin events** are considered potential statistical outliers and arise from weak signals, weak protein expression, or saturated signal intensity. A large number of boundary events may be due to poor compensation, flow cytometer settings, or drifts in cell events. Boundary and margin events were removed from any subsequent analyses as they may significantly affect downstream QA and preprocessing. A diagnostic report (Figure 5) was created and channels that have at least 3% boundary/margin events are flagged. This cutoff was selected based on visual inspection of univariate and bivariate plots of the data.

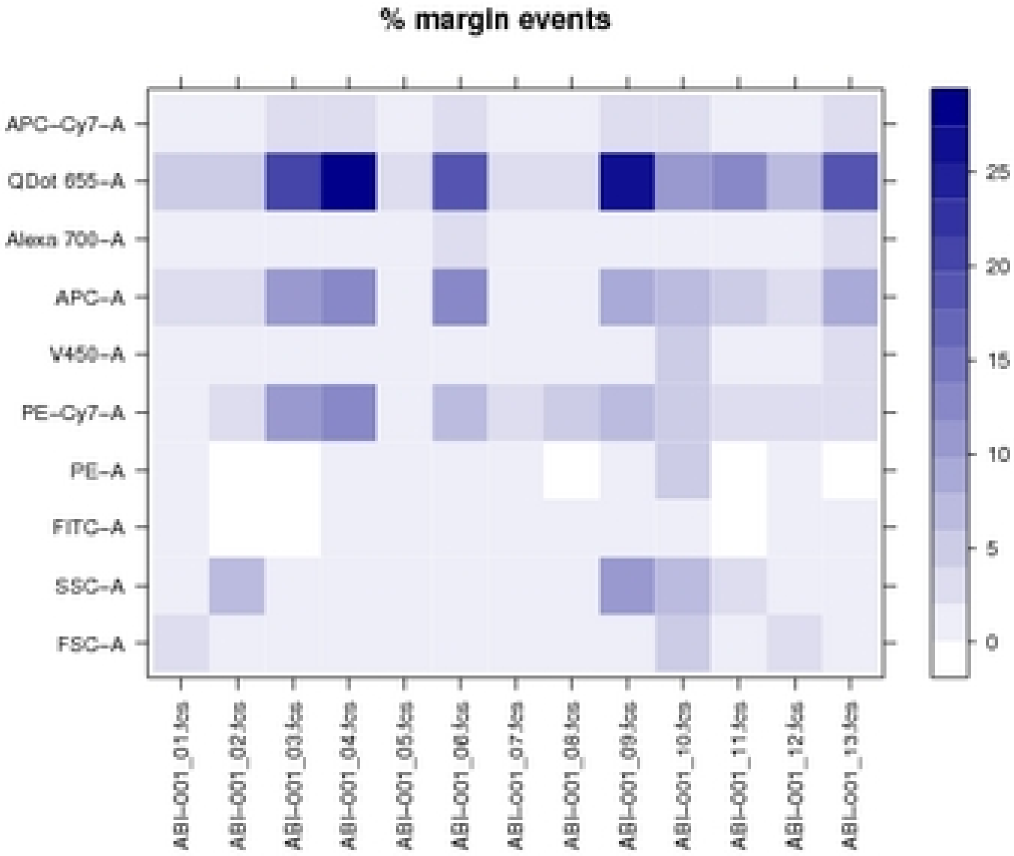


Figure 5. *Boundary/Marginal cell event quality assessment*. The above panel shows percentage of marginal cell events for each channel (vertical axis) and each tube (horizontal axis) for a multi-tube experiment. The color bar on the right shows percentage level of cells, and is useful for easy visualization of marginal events. The panel shows that the channels associated with QDot 655 , APC, and PE-Cy7 show high concentrations of margin events, which should be further investigated and potentially removed.

Consistency of cell event distribution across tubes was also assessed. Since each patient's sample at each timepoint had been aliquotted into separate tubes, I expect the distributions of cells that share morphological characteristics (measured by FSC and SSC) and specific cell surface antigens (e.g. CD3, CD4, CD45RA) to be similar across aliquots. I also expect more robust comparisons between tubes processed and stained in parallel, and analyzed sequentially on the same instrument at the same timepoint.

This assumption was checked using the function *flowFP*, which takes in a set a multi-tube experiments and a set of channels, which can be specified by the user. *flowFP* uses the probability binning (PB) algorithm [25] to recursively subdivide the multidimensional space into a set of hyper-rectangular regions (bins) with an equal number of events. The process is initialized by applying the PB algorithm to a training set that is randomly sampled from the data, thus creating a model of the multidimensional space. Figure 6 shows a two-dimensional projection of the model along the FSC and SSC channels. The algorithm first finds the channel with the highest variance, divides the population at the median of the channel, and continues this subdivision process for several recursions, as specified by the user. The model is then reapplied to the remainder of the data, resulting in feature vectors, or "fingerprints" that describe the multivariate probability distribution for each tube. Based on visual inspection of the resulting fingerprints, the algorithm was run for a total of 5 recursions, resulting in 2^5=32 bins. This recursive number resulted in the most appropriate resolution of the one-dimensional projection of the fingerprint values (Figure 7). The standard deviation of the fingerprint values for each tube provides a comparative measure of the degree of similarity between each tube and the norm of all the tubes in the experiment.

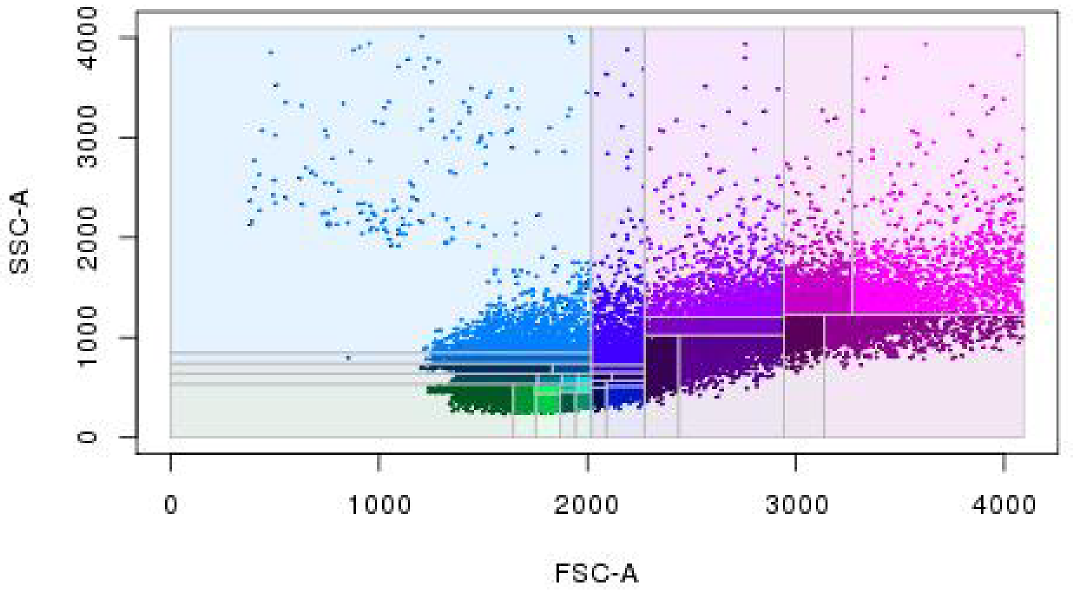
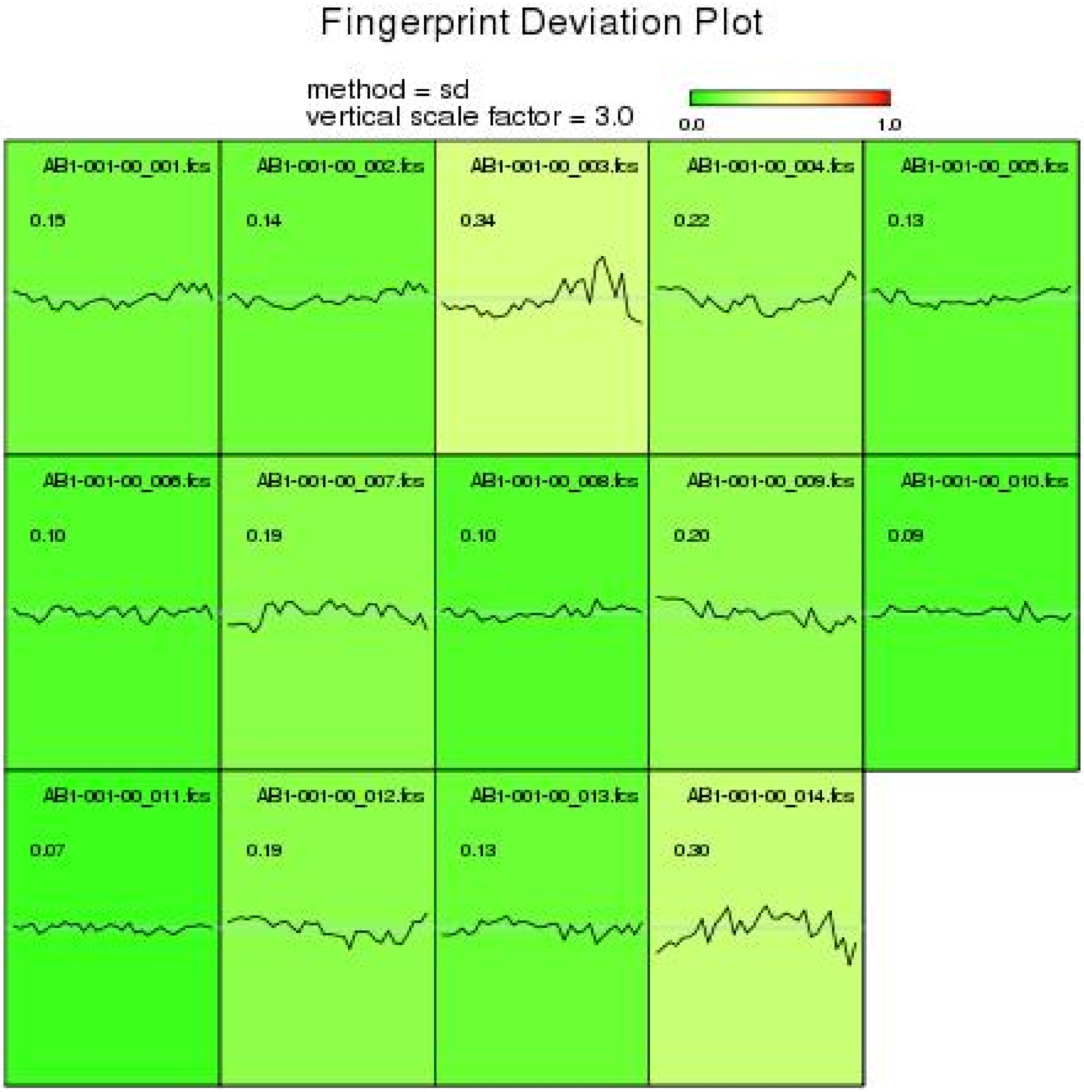
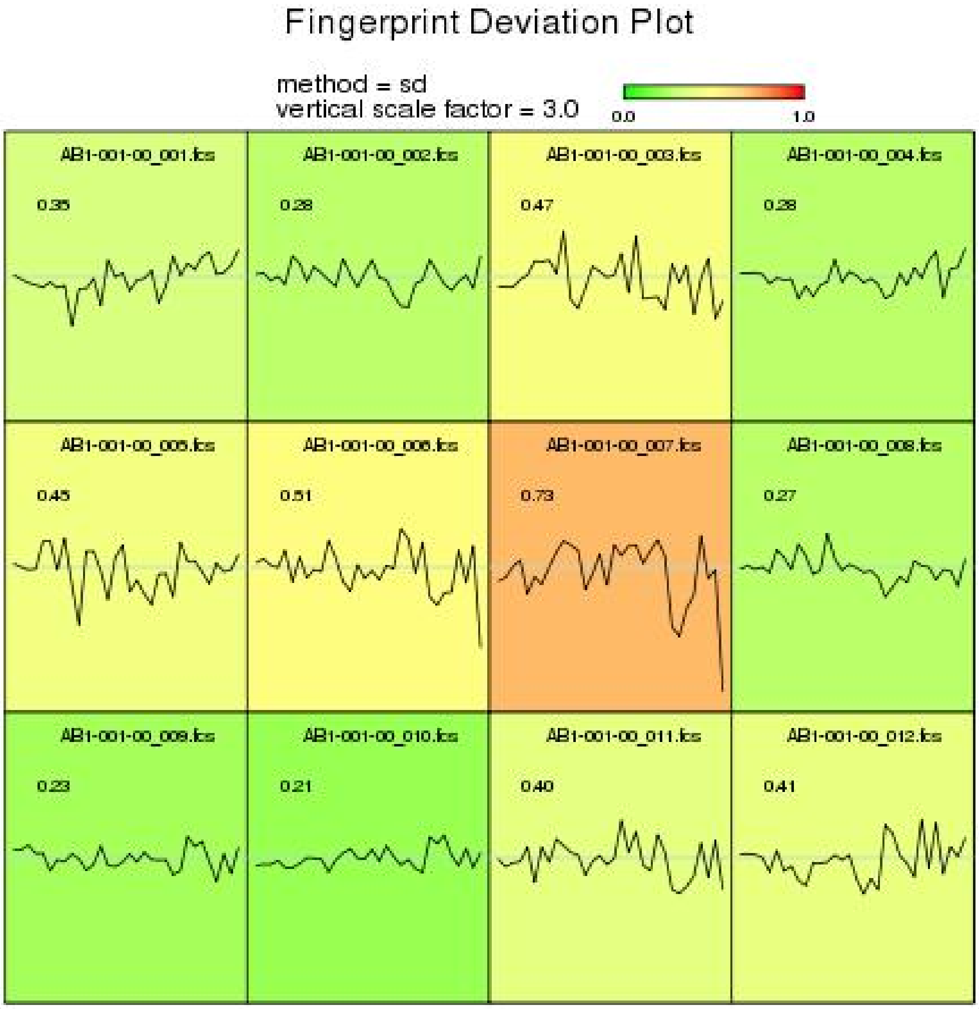


Figure 6. *Multidimensional Cytometric Fingerprint Regions*. The two-dimensional (SSC-A vs. FSC-A) space shown above for a single tube is subdivided into rectangular bins of approximately equal events for calculating the multidimensional cytometric fingerprints. This visualization process can be used to view bins in tubes that contain excess events corresponding to cell location drifts or the presence of new populations.

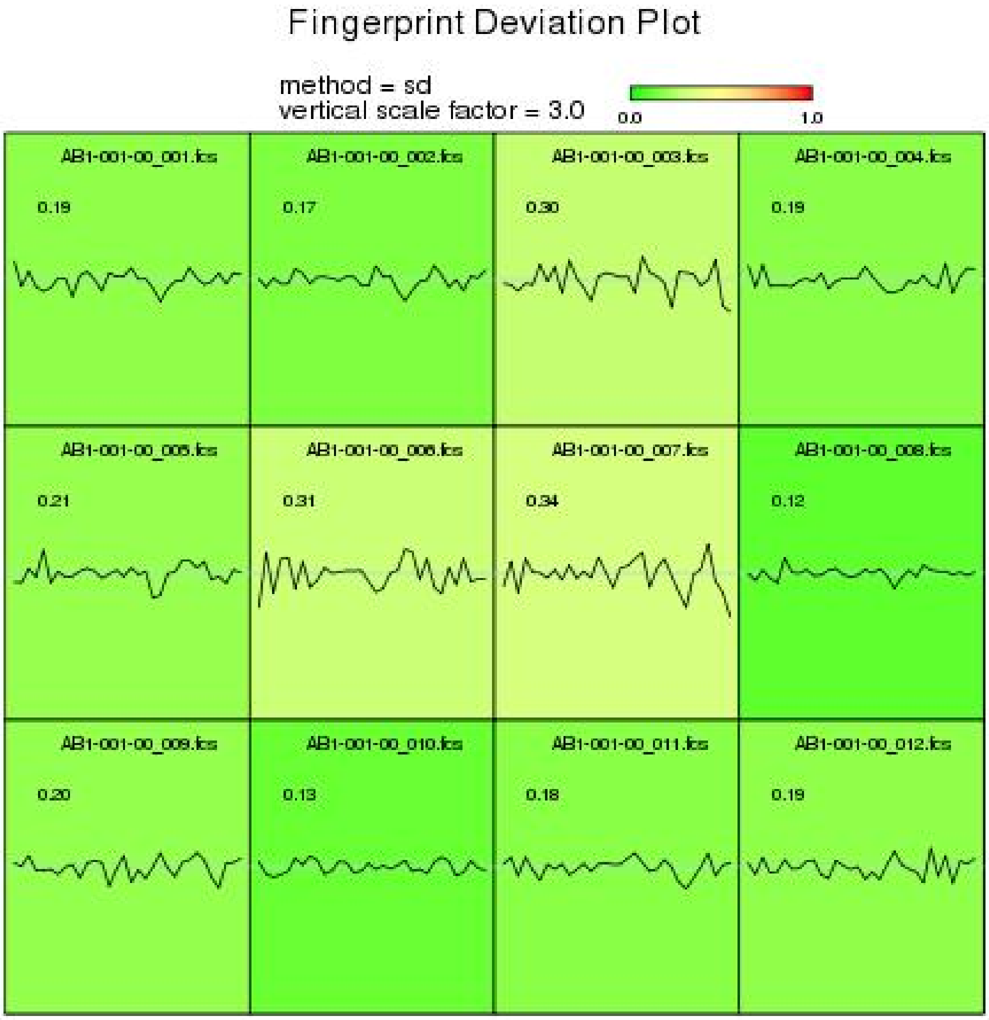
Checking for consistency of cell events across tubes was also useful for assessing whether **normalization** was needed. Tubes sharing certain antigenic markers should have very little variation in the plots produced by *flowFP*. This check can be performed by specifying the channels involved in gating out these markers. Aberrations can be readily detected in the QA plots produced by *flowFP*, as shown in Figure 7b. Panel plots are created to assess the magnitude of the gating deviation.


Fingerprint Deviation Plot
method = sd
vertical scale factor = 3.0
0.0
1.0
AB1-001-00_001.fcs
0.15
AB1-001-00_002.fcs
0.14
AB1-001-00_003.fcs
0.34
AB1-001-00_004.fcs
0.22
AB1-001-00_005.fcs
0.13
AB1-001-00_006.fcs
0.10
AB1-001-00_007.fcs
0.19
AB1-001-00_008.fcs
0.10
AB1-001-00_009.fcs
0.20
AB1-001-00_010.fcs
0.09
AB1-001-00_011.fcs
0.07
AB1-001-00_012.fcs
0.19
AB1-001-00_013.fcs
0.13
AB1-001-00_014.fcs
0.30

(a)


Fingerprint Deviation Plot
method = sd
vertical scale factor = 3.0
0.0
1.0
AB1-001-00_001.fcs
0.35
AB1-001-00_002.fcs
0.28
AB1-001-00_003.fcs
0.47
AB1-001-00_004.fcs
0.28
AB1-001-00_005.fcs
0.45
AB1-001-00_006.fcs
0.51
AB1-001-00_007.fcs
0.73
AB1-001-00_008.fcs
0.27
AB1-001-00_009.fcs
0.23
AB1-001-00_010.fcs
0.21
AB1-001-00_011.fcs
0.40
AB1-001-00_012.fcs
0.41

(b)



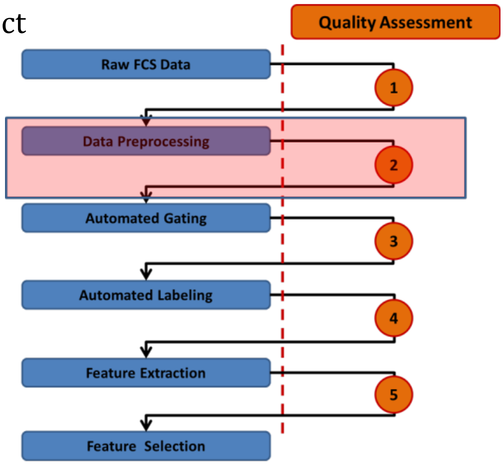
(c)

Figure 7. *flowFP QA panel plots.* Panel plots can be used to quickly assess consistency of gating across tubes or distribution of cell events via the color map. a) The top panel shows a multi-tube experiment that has good consistency across tubes for specific channels. b) The panel plot can be easily scanned for aberrations, such as Tube 7. The high number of yellow and red plots in this multi-tube experiment suggests that normalizeation may be needed across the shared channels. c) The problem in b) is resolved through normalization, as seen in the *flowFP* panel plot.

2.2.2 Data Preprocessing

2.2.2.1 Compensation and Boundary Event Removal

Raw FCM data was first **compensated** to correct measured intensities by removing the overlapping emissions shared between fluorochromes. This was accomplished by applying the inverse of the spillover matrix to the raw data. Boundary events were examined and removed for the FSC and SSC channels, as determined from QA.



2.2.2.2 Normalization and Transformation

In order to better resolve cell populations and reduce the influence of asymmetry and heterogeneity issues in the data, transformations along the fluorescence and scatter channels were performed (Figure 8) [26]. Additionally, normalization of the data along was performed on a per-channel basis [27] to remove effects arising from technical, non-biological variation and to facilitate downstream matching of cell populations. As suggested in Finak, et al. [26], fluorescence and scatter channels were analyzed separately, with the data normalized on the scale it would be visualized. Consequently, scatter channel data was normalized prior to transformation, and fluorescence channel data was normalized after transformation.

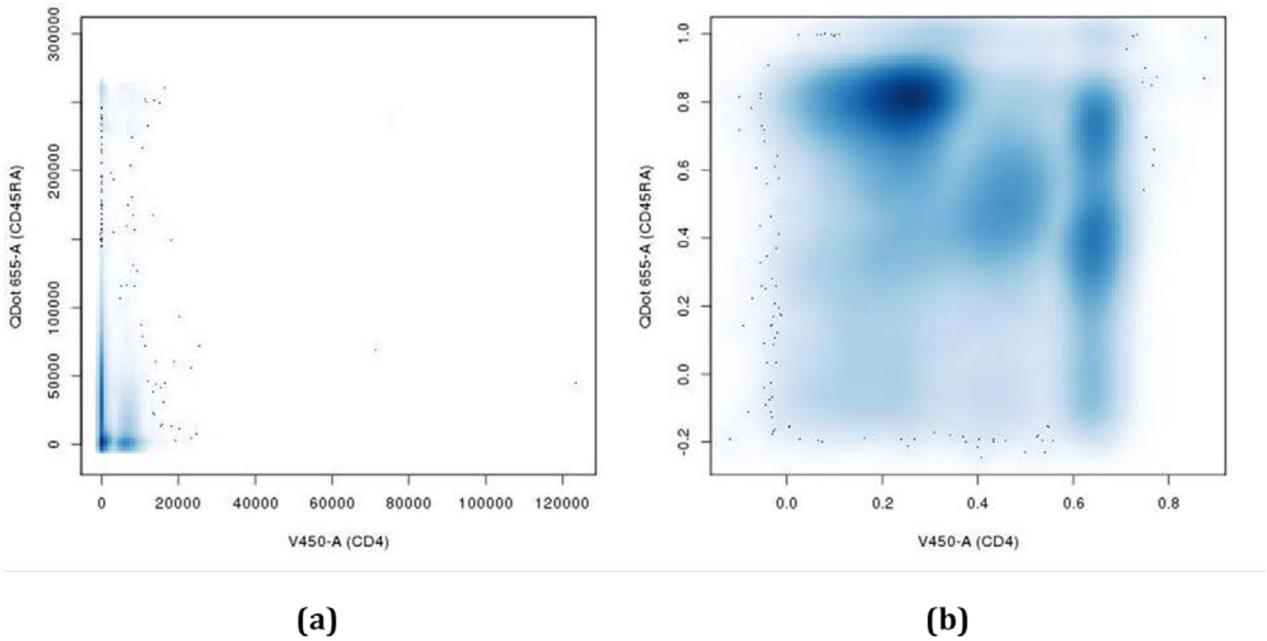


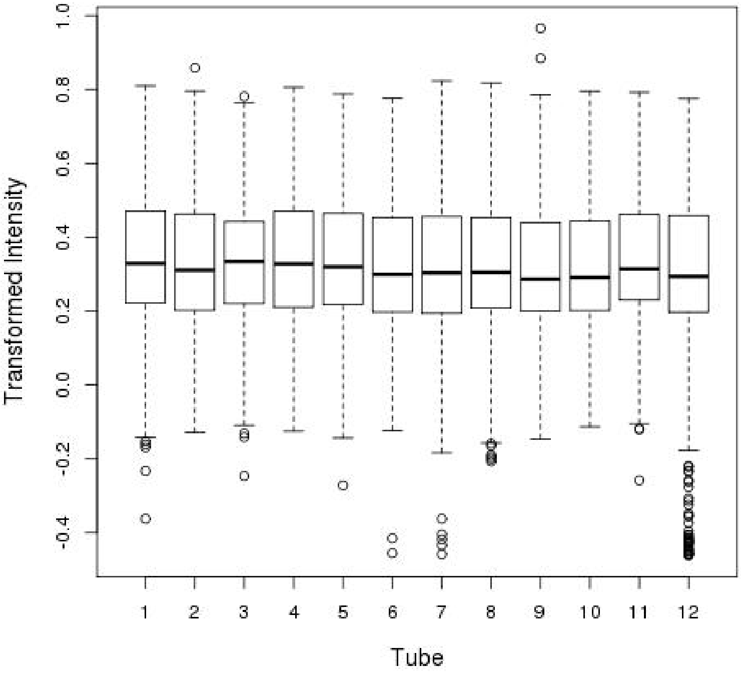
Figure 8. *Logicle transformation of compensated data*. The two-dimensional plots above correspond to CD45RA vs. CD4. a) Compensated data before transformation. b) Logicle transformation allows cell events to be better visualized for downstream gating/clustering purposes.

Due to the exponential relationship between fluorochrome intensities and protein concentration in the cell, a **logicle (biexponential) transformation** was used along the fluorescence channels [28]. The presence of negative values in the data due to compensation also warranted the use of the logicle scaling function, *S(x)*, given by Eq. 1:

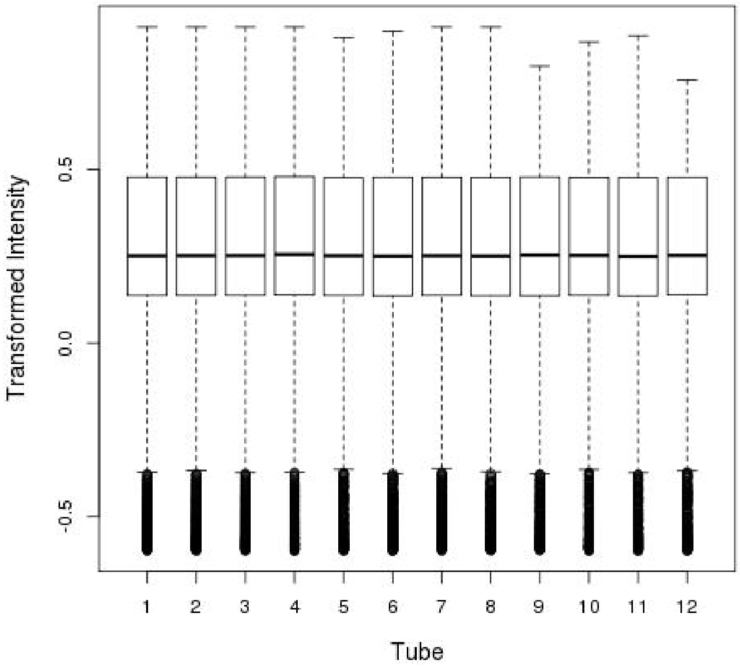
$$S(x;w) = Te^{-(m-w)}\left(e^{x-w} - p^2 e^{-\frac{x-w}{p}} + p^2 - 1\right) \quad \text{for } x \geq w \qquad \text{(Eq. 1)}$$

The parameter, w, gives the range of linearized data about zero, in decades; m represents the range of the display, in decades; p is a constant, such that, $w = 2p\frac{\ln p}{p+1}$; and T represents the maximum value of the display. Based on the data, the values, w = 0.5, m = 4.5, and T = 262144, resulted in good segregation of the data. For the scatter channels, a linear transformation was used, such that the display ranged from 0 to 4095.

Normalization of FCM data for a single multi-tube experiment was accomplished on a per-channel basis using **quantile normalization**, a technique that is widely used in microarray analysis, and also successfully applied to FCM analyses in the absence of normalization beads [29, 30]. Quantile normalization is a technique that matches the empirical distributions of fluorescence intensities for selected channels that identify the same set of antigens across all tubes (Figure 9). For each tube in a single multi-tube experiment, the untransformed (scatter channels) or transformed (fluorescence channels) fluorescence intensities of selected channels are sorted. For a single channel, the sorted intensities are then replaced with the mean of the intensities across all the tubes. Because not all tubes contained the same number of events, missing values were induced. Based on the assumption that these values were missing at random, missing values in each tube were replaced with the median intensities of the selected channels. Again, QA helped to make a decision on whether normalization is appropriate for the data. In FCM data, it is often difficult to ascertain if a variation is due to technical problems or real biological differences. Blindly normalizing the data may potentially remove useful information by normalizing out real biological differences between two samples.


Transformed Intensity
1.0
0.8
0.6
0.4
0.2
0.0
-0.2
-0.4
1
2
3
4
5
6
7
8
9
10
11
12
Tube

(a)


Transformed Intensity
0.5
0.0
-0.5
1
2
3
4
5
6
7
8
9
10
11
12
Tube

(b)

Figure 9. *Quantile Normalization of Shared Tubes*.. a) Transformed data before normalization. b) Quantile normalization matches distributions across tubes and allows for even comparison and consistent gating processes

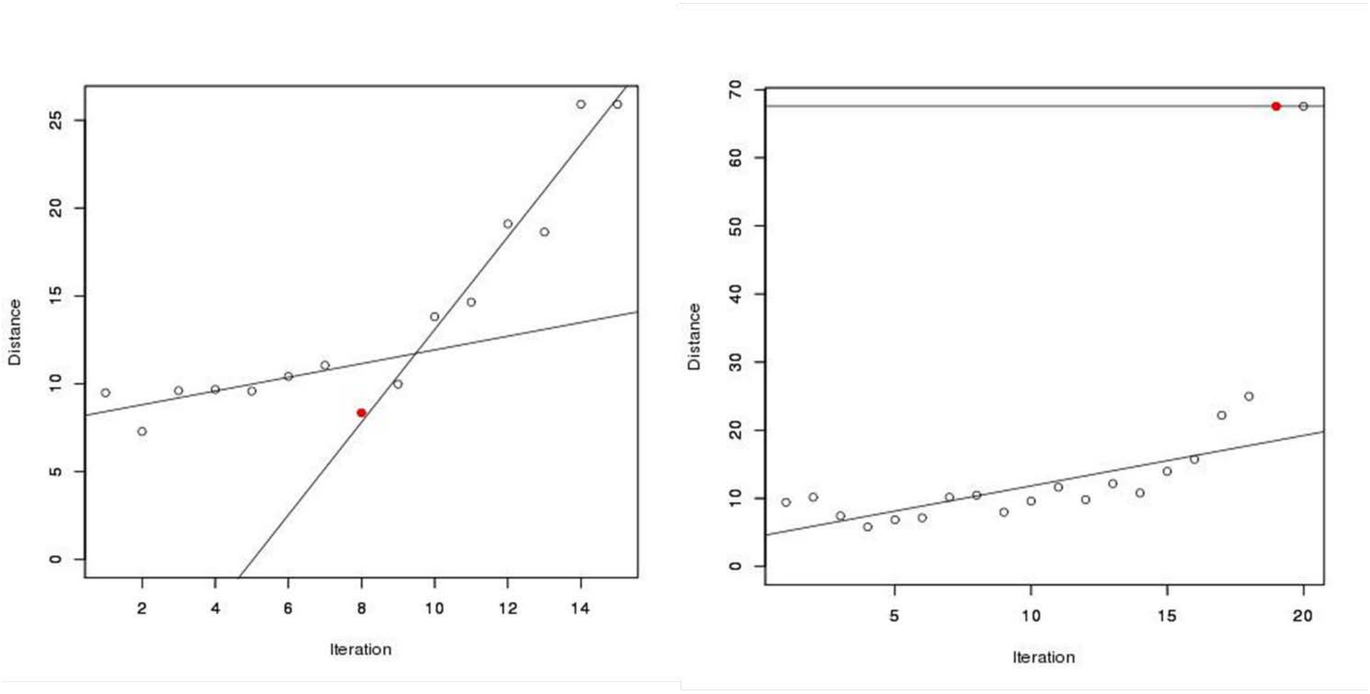
2.2.2.3 Non-Viable Cell Removal

After normalization and transformation of the compensated FCM data, **non-viable** cells, such as cell debris or red-blood cells, were removed. The removal of these cells is important for accurate, reproducible analyses. Non-viable cell events are traditionally identified through staining with propidium iodide (PI) and gating on events with high PI intensity; however, the available FCM data in this study did not employ this procedure. Instead, non-viable cells were removed via clustering in the FSC and SSC dimensions for each tube in a multi-tube experiment. Non-viable cells are defined as events with low FSC and SSC values.

Initially, *flowMeans*, a model-based clustering method that extends k-means, was used to cluster non-viable cells [31]. Unlike traditional k-means, the user does not need to predefine the number of clusters in *flowMeans* and the cluster shapes are not spherical. Often it is not possible to predefine the number clusters due to sample intervariability in FCM data [31]. Instead, a reasonable maximum based on the kernel density estimation of eigenvectors of the data was used to initialize the data. Furthermore, a small uniform noise was added to each event to avoid singularity issues that may arise from transformation. The algorithm first creates overlapping clusters, which are later merged based on a symmetric derivation of the scale-invariant Mahalanobis distance metric:

$$D(\boldsymbol{X},\boldsymbol{Y}) = \min\begin{cases}\sqrt{(\overline{\boldsymbol{X}}-\overline{\boldsymbol{Y}})\cdot S_{\boldsymbol{X}}^{-1}\cdot(\overline{\boldsymbol{X}}-\overline{\boldsymbol{Y}})^{T}}\\ \sqrt{(\overline{\boldsymbol{X}}-\overline{\boldsymbol{Y}})\cdot S_{\boldsymbol{Y}}^{-1}\cdot(\overline{\boldsymbol{X}}-\overline{\boldsymbol{Y}})^{T}}\end{cases} \qquad \text{(Eq. 2)}$$

In Eq. 2, $\boldsymbol{X}$ and $\boldsymbol{Y}$ represent two cluster populations and $S_{\boldsymbol{X}}$ and $S_{\boldsymbol{Y}}$, represent the covariance matrix of clusters $\boldsymbol{X}$ and $\boldsymbol{Y}$, respectively. Finally, a piecewise linear regression change-point algorithm (Figure 10) estimated the optimal number of subpopulation based on a break point that minimized the error of the regression model. A leave-one-out cross-validation strategy was used to obtain parameter values (i.e. maximum number of cluster = 20 and initial number of cluster solutions = 10).



(a) **(b)**

Figure 10. *Changepoint detection algorithm in flowMeans.* a) Correct changepoint detected the right number of clusters corresponding to the iteration number (red dot). b) Incorrect changepoint detected using cross-validated parameters.

Ultimately, I found that the *flowMeans* algorithm was not yielding stable results when only the scatter channels were specified for multidimensional clustering. The changepoint detection algorithm did not always provide the correct number of clusters, given the parameters obtained from cross-validation (Figure 10). As a result, the algorithm did not adequately separate non-viable cells in these cases, often creating multiple subpopulations in this compartment or merging non-viable cells with lymphocyte populations (medium FSC and low SSC values) (Figure 11). Prior to cell population identification and labeling, it is important to remove as much as possible influence from non-viable cells. Therefore for this final stage of data preprocessing, a more robust alternative was explored.

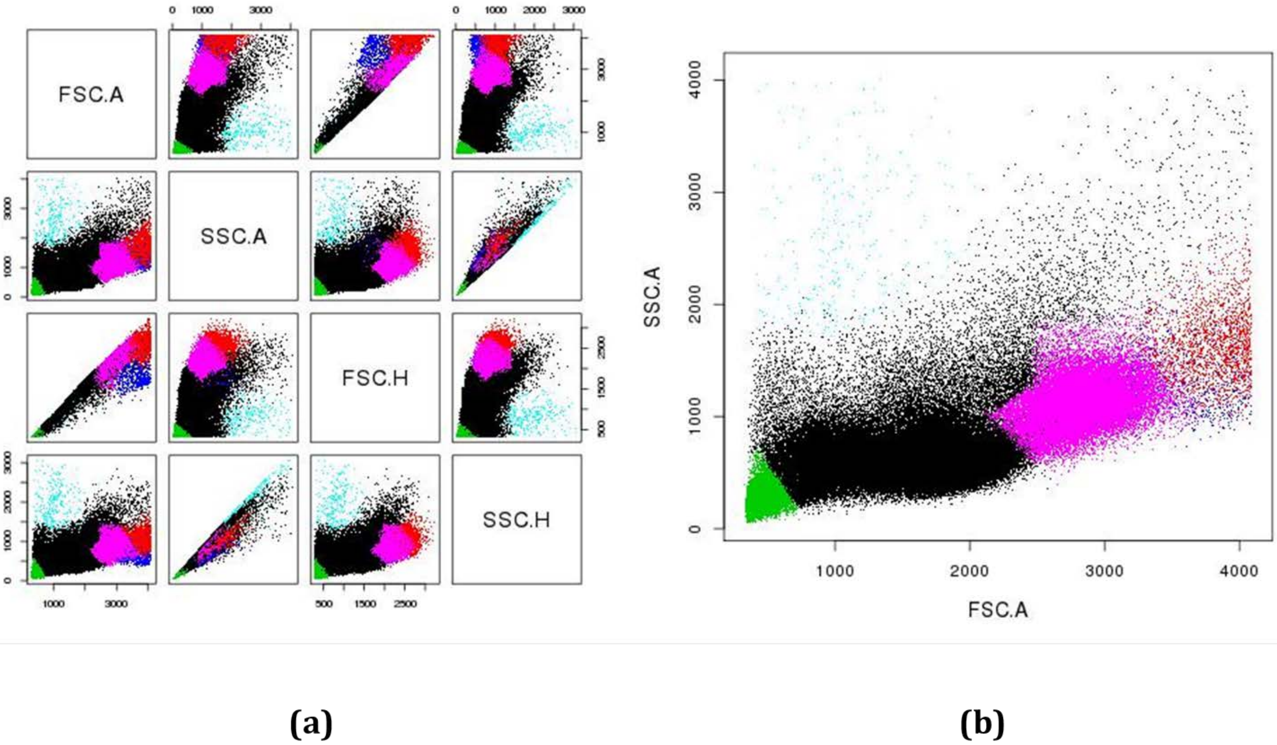


Figure 11. *Inadequate separation of non-viable cells using flowMeans.* The *flowMeans* algorithm poorly segregated non-viable cells in several samples. The above example corresponds to the incorrect changepoint detection seen in Figure 10b. a) Multivariate view of the clusters. Green cells are considered non-viable (low FSC and low SSC). However, as seen in b) the algorithm greatly underestimates the non-viable population by incorrectly identifying some non-viable cell events as part of the lymphocyte population (black).

Spectral clustering is an unsupervised, graph-based technique that has been applied to many biological datasets and has low sensitivity to the required parameters (i.e. a scaling parameter, σ that controls edge weights and the separation factor) [32].

Furthermore, it is robust to noise, outliers, and cluster shape. A computationally effective algorithm, called *SamSPECTRAL*, has been implemented in Bioconductor framework [33]. The algorithm requires a matrix of data points (i.e. a single tube of FCM cells). To mitigate the expensive computational cost associated with traditional spectral clustering, *SamSPECTRAL* first performed a data reduction scheme based on faithful sampling, which produces a nearly uniform set of representative vertices (data points) and edges that preserve density information of the original data. *SamSPECTRAL* then builds a graph, with vertices corresponding to data points (i.e. cell events), and edges connecting all pairs of vertices. Edges are weighted based on a similarity criterion between any two vertices. After normalizing the adjacency matrix of the graph, the algorithm then computes the eigenspace of the normalized graph in order to automatically determine the number of clusters based on the "knee point" of the eigenvalue curve (similar to the change-point detection of *flowMeans*. A modified Markov Clustering (MCL) algorithm [34] partitioned the graph instead of the traditional spectral clustering method. Finally, k-means is used to partition the graph into a set of clusters. The post-processing stage of the algorithm combines biological knowledge of FCM data (i.e. similarities between vertices are higher in regions with higher densities and tend to exist at the center of cell populations) with the resulting clusters. Figure 12 diagrams the *SamSPECTRAL* process in the successful partitioning of non-viable and viable cells in a single tube, using a scaling parameter of σ=100 and a separation factor of 1.2, as well as the channels, FSC-A, FSC-H, SSC-A, and SSC-H as input for the multidimensional clustering algorithm. The parameter values were selected based on a leave-one-out cross validation strategy.

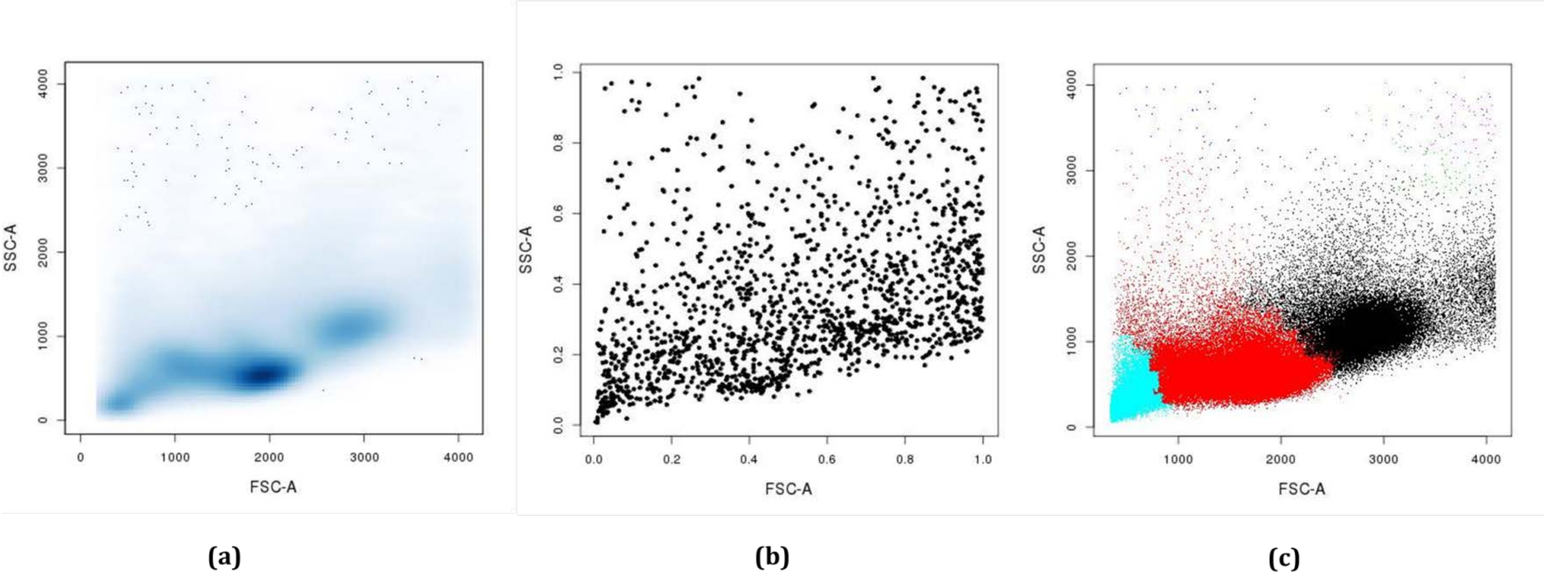


Figure 12. *Faithful sampling and spectral clustering of FCM data using SamSPECTRAL.* For comparison with *flowMeans*, the results above are obtained from the same tube data used to generate Figures 11 and 12. a) Faithful sampling successfully preserved dense areas in the data. b) *SamSPECTRAL* clustering provided better separation of non-viable cells (cyan), without over extracting the lymphocyte population (red).

At a cost of computational running time, *SamSPECTRAL* was able to resolve non-viable cells by considering only the scatter channels in the multidimensional clustering algorithm. For a typical FCM sample with 50,000 events, *SamSPECTRAL* had a running time of 2.5 minutes, compared to a run time of 30 seconds for *flowMeans*. Based on the clustering results, I defined non-viable cells as clusters whose centers had transformed FSC values less than a determined cutoff point (Figure 13). Proper normalization of the scatter channels was needed to set this cutoff, as cells may drift slightly over time. In other datasets, this cutoff may need to be set at the discretion of the user. However, the robust performance of *SamSPECTRAL* should ensure that populations are well-separated at this step to specify a reasonable exclusion criterion. Non-viable events were ultimately excluded from subsequent steps in the analysis pipeline.

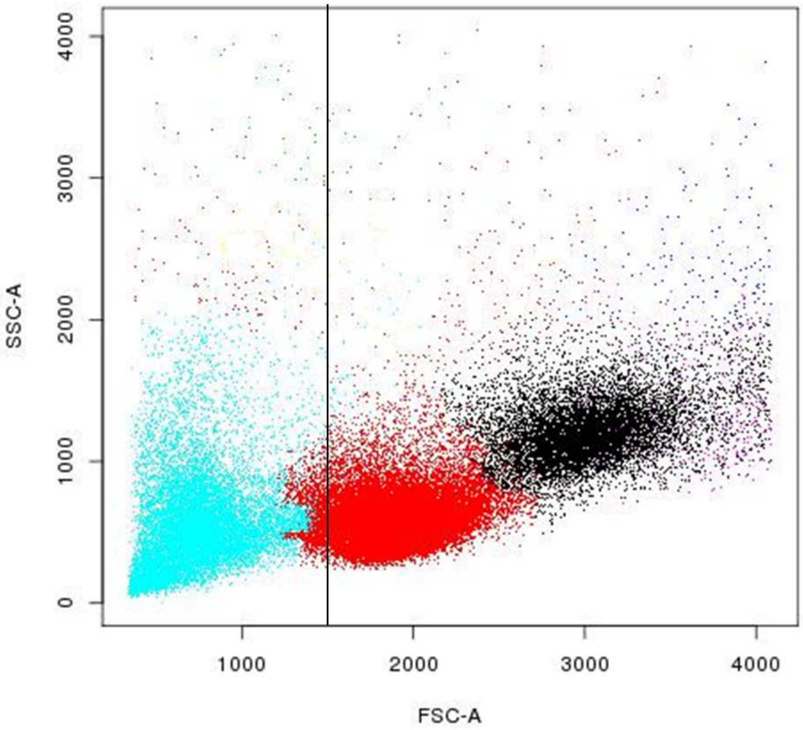


Figure 13. *Removal of non-viable cells.* The *SamSPECTRAL* algorithm facilitated identification and removal of non-viable cells using the scatter channels (FSC-A, FSC-H, SSC-A, and SSC-H) as input for multidimensional clustering. Non-viable cells corresponded to clusters that have mean FSC-A value less than 1500.

2.2.2.4 Doublet Removal

In addition to non-viable cells, **doublets** were excluded by clustering in the scatter channels. Doublets often arise when two cells are considered as a single event, as they pass through the laser. This is particularly problematic, as the aggregate phenotypic expression on both cells may be erroneously considered a type of cell subpopulation. In practice, doublets are gated by viewing the data in the area versus height plot of a scatter channel (e.g. FSC or SSC). On the area versus height plot, these events are often characterized by a broadened height signal and reside above a narrow band (on-axis) of densely clustered events. Using this property, *SamSPECTRAL* was used to remove these off-axis events using the same parameters for non-viable cell clustering (i.e. scaling parameter of σ=100 and a separation factor of 1.2) (Figure 14). A post-processing step removed the smallest cluster corresponding to the doublet events. This heuristic was based on the observation that the densely populated areas will be clustered together as the largest clusters, while the sparse doublet events will be considered together, forming the smallest cluster.

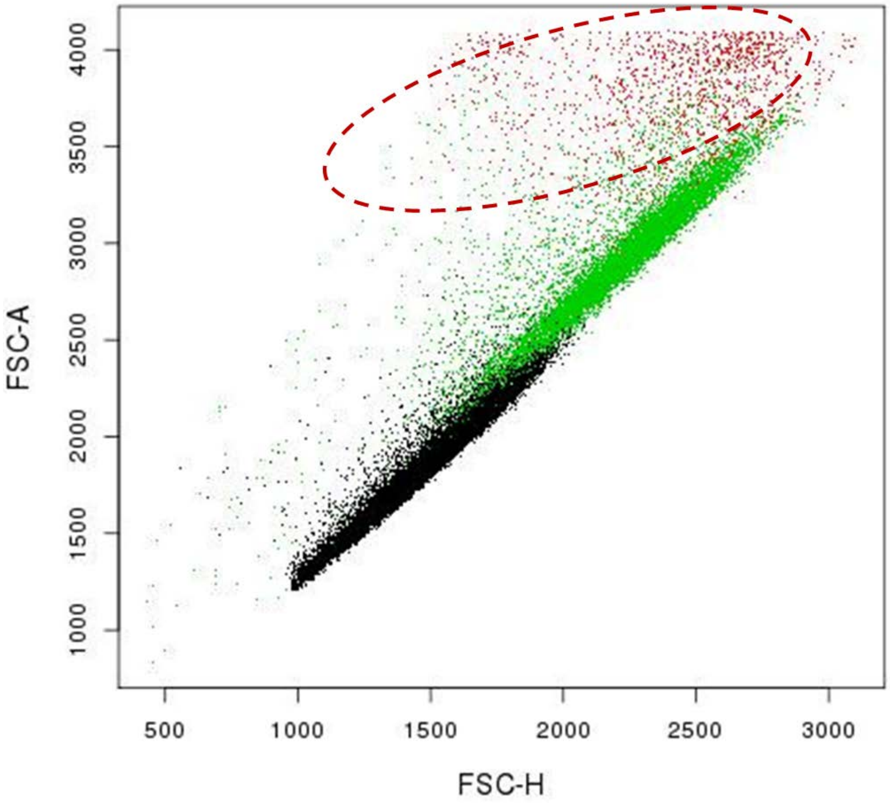
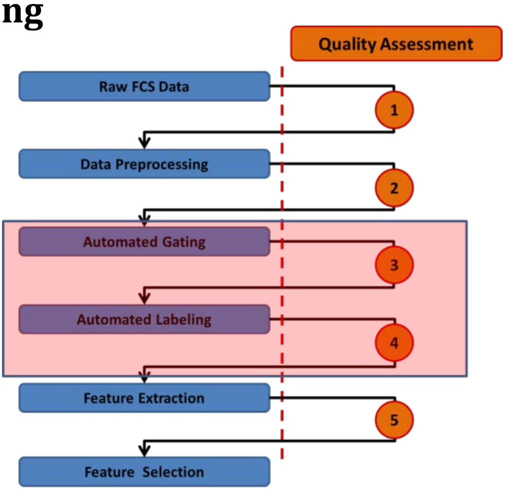


Figure 14. *Doublet Gating Procedure.* Doublets (dotted ellipse) were identified using *SamSPECTRAL* clustering, by specifying the scatter channels (FSC-A, FSC-H, SSC-A, SSC-H). The smallest cluster on the off-diagonal axis was considered the doublet population.

2.2.3 Cell Population Identification and Labeling



After data preprocessing, the Bioconductor package, *flowType* [6] was used to cluster and label cell populations. I hypothesized that using the k-means algorithm in *flowType* to partition each channel into a set of negative and positive populations, and then merging the results to generate multidimensional immunophenotypes can lead to an efficient clustering and labeling strategy. The assumption is that each antigenic marker identified by each channel is either expressed (positive) or unexpressed (negative). Furthermore, using *flowType* assumes that *bright* and *dim* marker expressions, which arise from fluorescence-spilling between channels and become difficult to accurately quantify, are resolved in other dimensions. *flowType* takes in a single tube experiment and a set of channels, which I have defined as all fluorescence channels available in the data. Figure 15 shows successful partitioning of the channels into negative and positive populations.

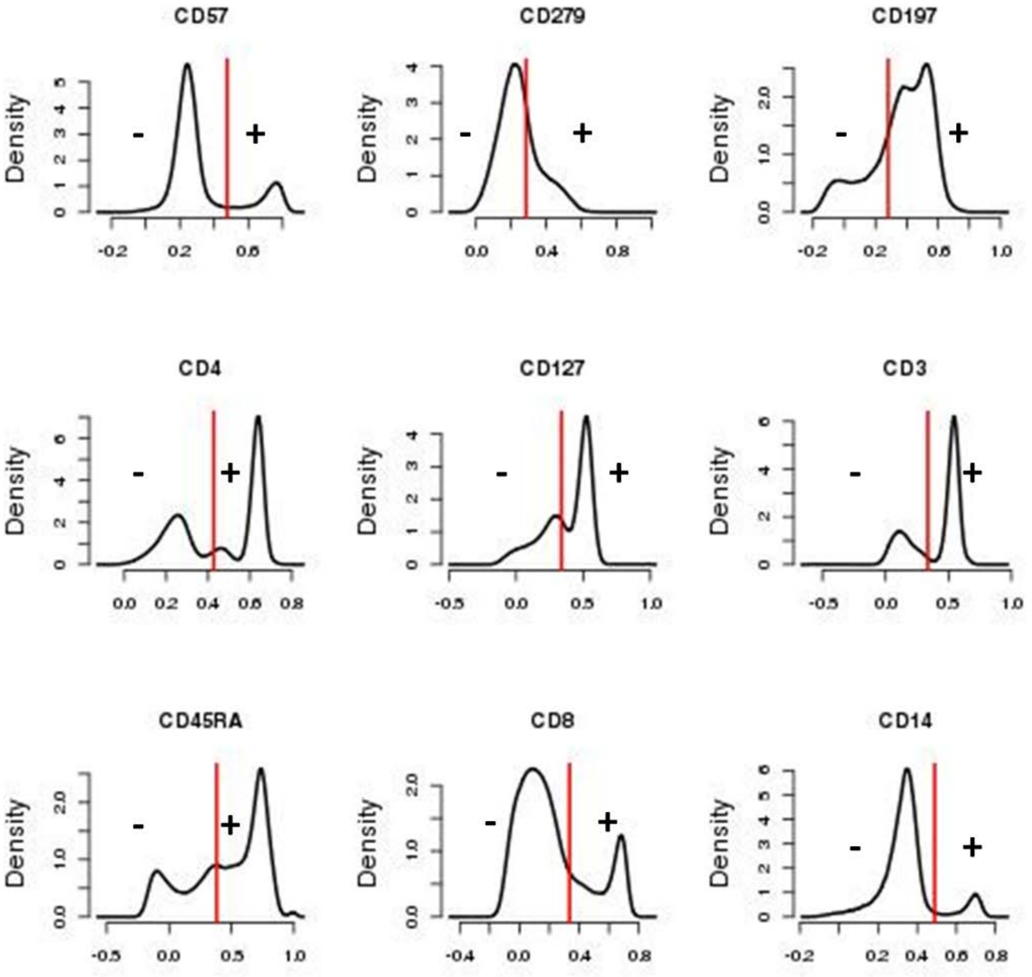
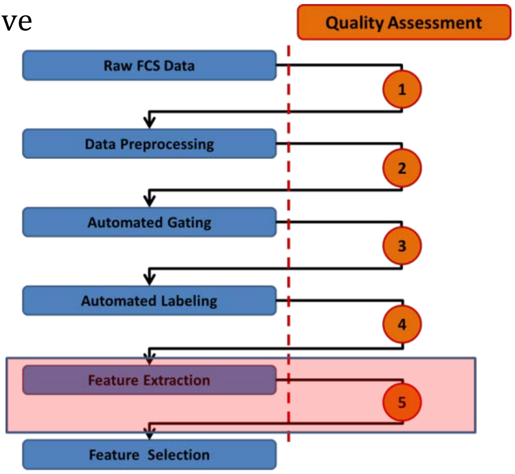


Figure 15. *Single-dimensional partitions using flowType.* The *flowMeans* algorithm successfully divided univariate distributions into positive and negative cell populations for a 9-color FCM experiment.

Bipartitions are then combined for all channels resulting in 2^n distinct cell immunophenotypes that correspond to different antigenic expressions (e.g. CD3+CD20-CD4-CD8-). However, because some cell populations are subsets of other populations (e.g. CD3+CD20-CD4+CD8-CD197+CD45RA+ is a subset of CD3+CD20-CD4+CD8-), the total number of cell immunophenotypes that can be extracted using *flowType* is increased to 3^n, because a marker can be considered neutral or not present in the immunophenotype.

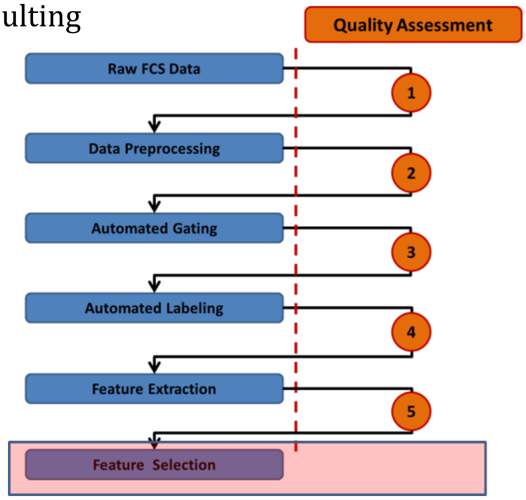
2.2.4 Feature Extraction

flowType was also used to display a list of relative cell frequencies corresponding to each immunophenotype. Similar to Aghaeepour, et al. [6], the total number of resulting immunophenotypes from the *flowType* output was further reduced to a set of homogeneous groups using linkage hierarchical clustering [35] (*hclust* function in R *stats* package) of these cell frequencies. The final set of immunophenotypes from the hierarchical clustering can then be further reduced via the feature selection procedure of the pipeline.



2.2.5 Feature Selection

It is often the case that the number of resulting features or classifiers from the analysis (e.g. immunophenotypes) may greatly exceed the number of available training instances (e.g. patients in the study) (i.e. "large *p* and small *n*"), resulting in model overfitting and poor predictive performance. To prevent overfitting in regression of multidimensional features, regularization techniques, such as Lasso[36] and its variant, elastic nets [37], have been proposed to introduce regularization parameters or weights that penalize features



that are irrelevant. In the proposed study, I leverage an existing package, called *FeaLect*, which presents a robust classifier for feature selection based on Lasso. I hypothesize that this feature scoring procedure can yield robust results for exploratory and predictive analyses.

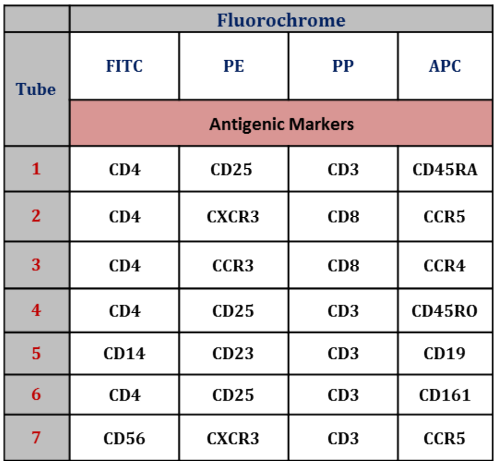
Chapter 3

3.1 Validation of the Analysis Pipeline

The innovation in the proposed work lies in the application of biomedical informatics techniques to create an analysis pipeline that can reliably anticipate and measure cell populations in the clinical domain of organ transplantation. The validation strategy carried out in this chapter relies on comparison with human-driven gating. Traditionally, human-driven gating is used to capture known cell populations belonging to the immune repertoire of interest. Therefore, it is useful in determining the biological and clinical relevance of automated results, and evaluating reproducibility of automated analyses. I define my pipeline as validated if the features (e.g. cell population frequencies) from the automated clustering are statistically similar to human validation. Successful validation can then be used to justify extension of the analytical framework to exploratory analyses for detecting novel and uncommon cell populations. In this chapter I present the validation results of the proposed pipeline as a proof-of-concept that high-throughput, multidimensional flow cytometry analysis is feasible for application in organ transplantation studies.

3.1.1 Dataset

The FCM validation dataset was obtained from the FlowCAP data repository (https://www.immport.org/immportWeb/display.do?content=FlowCap) under Challenge 5. The dataset originated from a clinical study [38] investigating the effect of immunotherapy (placebo, omalizumab alone, rush immunotherapy alone, and omalizumab and rush immunotherapy combined) on ragweed-induced allergic rhinitis in 145 patients over the four treatment arms. Peripheral blood mononuclear cells were isolated in blood samples over the various timepoints of the study. Samples were aliquotted into 7 tubes and stained with 4 different fluorescent dyes. FSC and SSC parameters were also specified and common to all the tubes. Each tube contained approximately 12,000 to 20,000 cells. The following panel in Table 2 summarizes the staining profile. In this validation, only the results of the first treatment arm (34 patients) (immunotherapy with anti-IgE) at timepoint 5 are analyzed with respect to the human-gated results. A single, expert observer performed manual analysis, thus eliminating potential interobserver variability.



	Fluorochrome			
Tube	FITC	PE	PP	APC
	Antigenic Markers			
1	CD4	CD25	CD3	CD45RA
2	CD4	CXCR3	CD8	CCR5
3	CD4	CCR3	CD8	CCR4
4	CD4	CD25	CD3	CD45RO
5	CD14	CD23	CD3	CD19
6	CD4	CD25	CD3	CD161
7	CD56	CXCR3	CD3	CCR5

Table 2. *Seven-tube, 4-color staining panel for validation dataset.*

3.1.2 Methods

Using the analysis pipeline, the data was preprocessed prior to automated gating, labeling and feature extraction. The QA process ensured that the necessary steps of compensation, non-viable cell and doublet removal, and consistency were met prior to running downstream analyses. Parameters for cytometric fingerprinting and spectral clustering were determined from a randomly selected training set of 10 patients. Selected parameters were then applied generally to the remaining 24 patients. Labeling analyses were performed on the lymphocyte compartment, which was automatically clustered with *SamSPECTRAL* using the scatter channel dimensions, a scaling parameter of σ=300, a separation factor of 0.9, and by specifying three clusters corresponding to granulocytes, monocytes and lymphocytes. Lymphocytes were defined as the cluster with centers located in the lower left quadrant of the SSC-H vs. FSC-H plot (Figure 16).

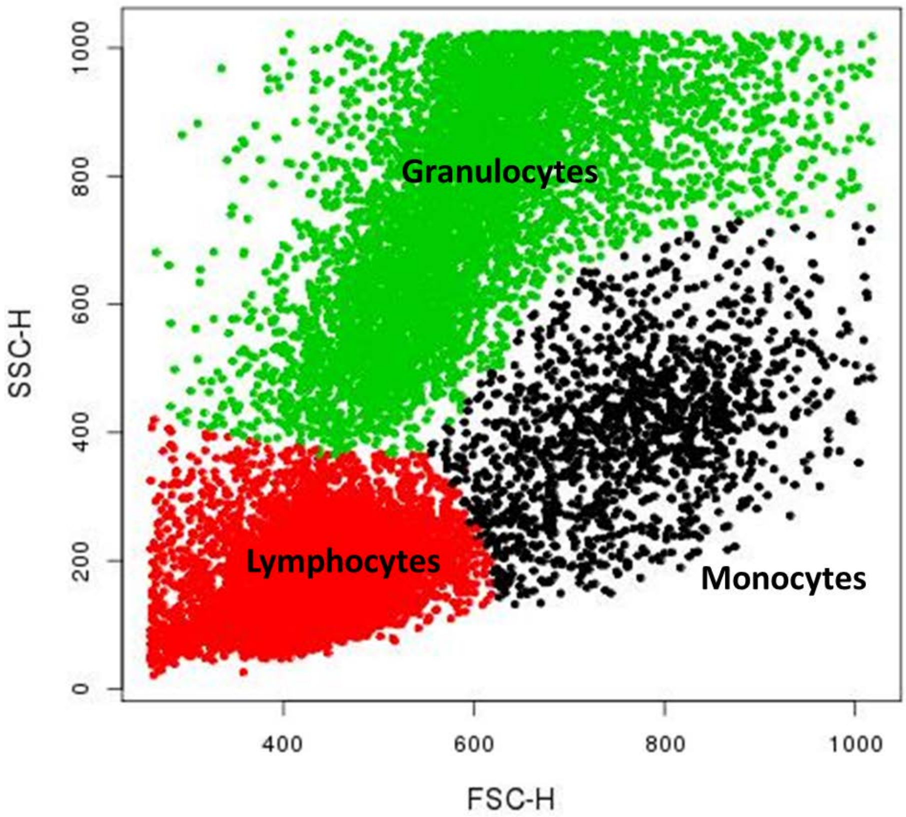


Figure 16. *Lymphocyte clustering in validation data. SamSPECTRAL* successfully divided the cell event space into the three main cell populations- lymphocytes (red), granulocytes (green) and monocytes (black). The natural biological relative locations of these populations were used to extract only the lymphocytes

After clustering the lymphocyte compartment, each tube was further analyzed using *flowType* to label all possible multidimensional immunophenotypes. Proportions of cell populations, expressed as frequency of the population divided by the total number of cells in the sample, were obtained for each immunophenotype and compared to human-gating results.

3.1.3 Statistical Analyses

First, I aimed to evaluate the extracted lymphocyte populations, which was assessed statistically by performing a repeated measure factor ANOVA test to measure the intrapatient variation of lymphocyte proportions across all tubes. One major assumption that must be checked is the consistency of cell distributions across tubes containing the same sample of blood from a given timepoint. Evaluating this assumption gives an idea of the pipeline's ability to reproducibly gate shared populations between tubes in an experiment. Using the standard ANOVA to test consistency was not appropriate as it fails to model the dependency between each tube, therefore violating the independence assumption of the ANOVA test. The sphericity, (i.e equality of variances of the differences in between the various tubes) assumption of the repeated measure factor ANOVA was determined using Mauchly's test, which formulates a test-statistic based on the sample covariance matrix. The F-test statistic is corrected using the Greenhouse-Geisser ε if sphericity is violated.

Secondly, my validation strategy examined the agreement between manual and automated results using Bland-Altman (BA) analysis [39], in which the differences (defined in the plots to be the automated results subtracted from the manual results) in measured proportions for a particular subpopulation is plotted against the average proportion obtained from the two methods. BA analyses were carried out for abundant and rare cell populations in the lymphocyte compartment. The correlation coefficient measured the linear relationship between the manual and automated results for a single cell population.

All values reported are mean ± standard deviation, unless otherwise indicated. Statistical significance is considered at the 0.05 level.

3.1.4 Results

3.1.4.1 Lymphocyte Extraction

The repeated measure factor ANOVA test did not support the null hypothesis of consistency of lymphocyte cell events across tubes (unadjusted F = 3.52, df = 6, p=0.0029), while Mauchly's test showed violation of the sphericity assumption (W = 0.09, χ^2 = 73.95, p < 0.001). Afterm the Greenhouse-Geisser ε (0.523) correction for sphericity, the adjusted p-value remained significant (p=0.026), which suggests intrapatient differences in the proportion means among the tubes. However, as seen in Figure 17, there is a strong, significant linear relationship (r = 0.91, p<0.001) between the manually-determined and automatically-extracted lymphocyte population proportions, and the matched points fall roughly on the identity line, suggesting good accuracy. The BA plot of lymphocytes (Figure 18a) also shows generally good agreement between the manual and automated gating process. The automated gating method has an average bias of about 5% lower compared to the manual gating, suggesting that the automated pipeline tends to slightly over-extract the lymphocyte population. All, but one data point, fall within the 95% limits of agreement (-15% to +5%).

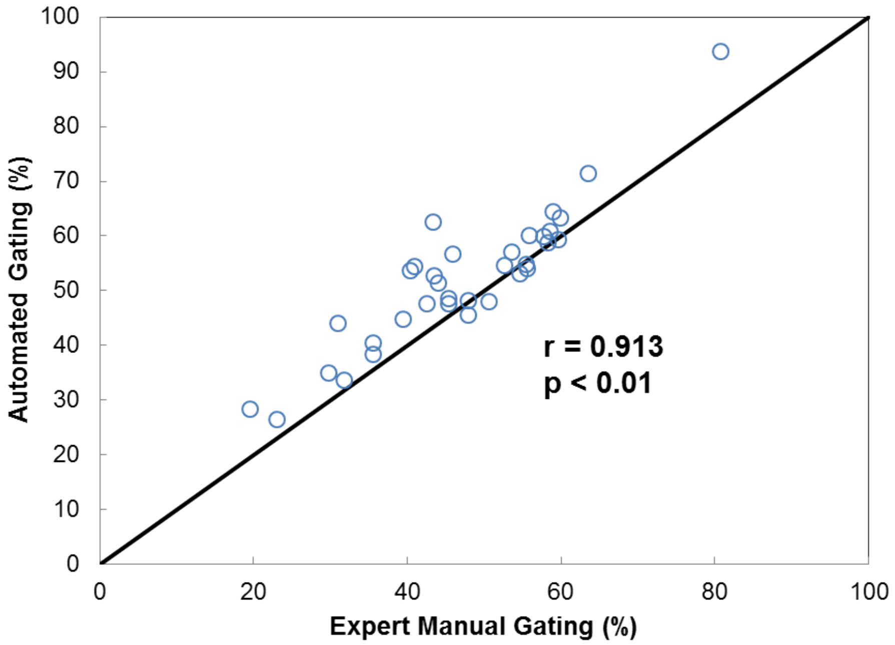


Figure 17. *High correlation between manual and automated gating methods.* The automated pipeline resulted in significant and strong linear relationship with manual gating results. Cell proportions for lymphocytes are plotted above. The black line represents the identity line.

3.1.4.1 Agreement Assessment

Bland-Altman analyses for a diverse set of cell populations are shown in Figures 19 and 20. Populations range from abundant immunophenotypes, such as lymphocytes and CD3+CD4+lymphocytes to more uncommon and rare immunophenotypes, such as CD3-CD4-CD25-CD45RO- All plots showed good agreement between manual and automated analyses, with biases less than 3.5%, and nearly all observations contained within the 95% limits of agreement. In all, but the measurement for CD3-CD56-CCR5-CXCR3-, the automated algorithm overextracted the cells. The best agreement was observed for

CD3+CD4+CD25+CD161 (Figure 19), with bias < 0.5%. The correlation coefficient for this population was only moderate, but highly significant (r=0.75, p<0.01).

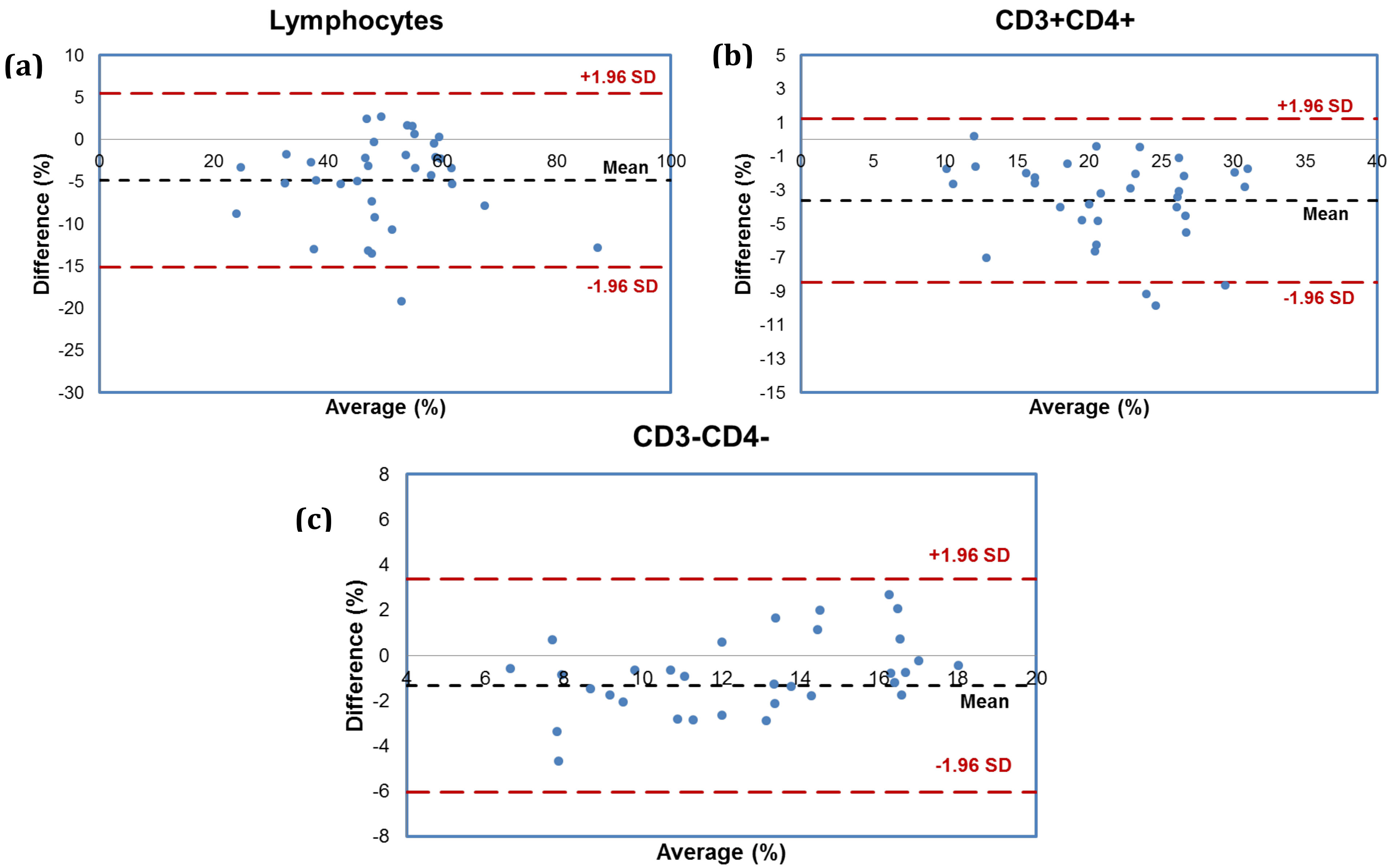


Figure 18. *Bland-Altman analyses for intermediate-frequency and rare cell populations.* Differences are expressed as manual minus automated

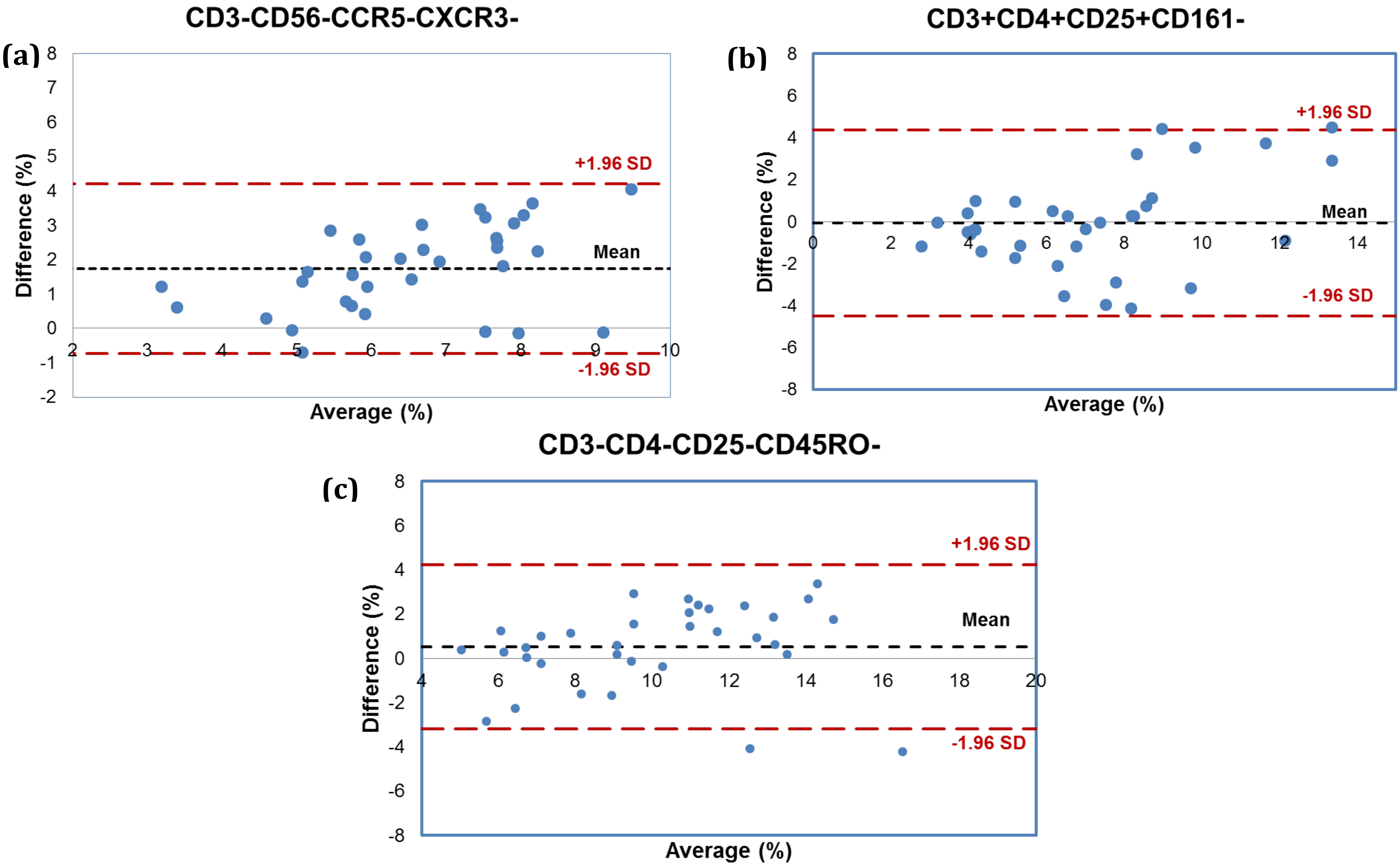


Figure 19. *Bland-Altman analyses for intermediate-frequency and rare cell populations.* Differences are expressed as manual minus automated

3.1.5 Discussion

The results of this validation study suggest a multidimensional, automated analysis pipeline can be used in lieu of the manual gating traditionally employed in FCM data analysis. Although the repeated measure factor ANOVA test resulted in rejecting the assumption of consistency among extracted lymphocyte events, the average variation across tubes for all patients is only 3%, which may be clinically acceptable for most studies. Furthermore, this error is similar to the 2% average intraobserver variation due to the manual analysis of a single individual. This error is likely to be much higher if multiple individuals or an inexperienced individual gated the data; manual, visual based cell population identification is often a highly subjective task and depends on the analyzer's expertise. Nevertheless, there was very significant, strong correlation and good agreement between the two methods for extracting lymphocytes.

The validation study also shows extensibility of the pipeline to reliably measure not just abundant cells, but even intermediate-frequency and rare populations in the data. From BA analyses of several, diverse cell populations, the rare populations often had the lowest bias between manual and automated methods. . Theses biases were also within 5%, which may not be clinically important. Correlations between the two methods for these populations were also highly significant. These findings have important implications for exploratory analyses, where one may be interested in finding all possible cell populations associated with a particular clinical outcome.

Human validation is only the first step in determining the biological and clinical relevance of the analysis pipeline. Here, I have shown that for expected cell populations, the automated methods show good agreement with manual results. For novel or unexpected cell populations, a prospective experimental validation can be performed. The aim is to first purify these novel populations using FCM and then test their ability to mediate certain biological processes as anticipated from clinical relationships.

This study is not without limitations. While *flowType* provided a good initial step for labeling similar clusters across tubes, the algorithm at the moment only uses hard constraints to define positive and negative expressions of an antigen, with the assumption that less discernible expression levels, such as *dim* and *bright*, for an antigen, will be resolved in other channels. The use of a hard constraint may have resulted in the consistent overextraction of cell populations when compared to the manual results. An alternative would be to extend *flowType* to incorporate more than 2 populations for a single dimension or a fuzzy k-means implementation where cells can belong to more than one population.

Overall, this validation study has shown that the results from the automated pipeline are as reproducible and accurate as the traditional manual approach. Potentially, this can lead to more efficient FCM data analysis and lead to quicker tests of biological hypotheses in clinical studies. I aim to use this pipeline to further analyze the functional characteristics of a transplant recipient's immune repertoire with respect to several clinical outcomes.

3.2 Ongoing and Future Work

Ongoing and future work aims to leverage the proposed FCM analysis pipeline to study two main objectives associated with renal transplant patients enrolled in an ongoing clinical trial to study the safety and effectiveness of a new investigational drug combination consisting of alemtuzumab, belatacept, and sirolimus. 1) The first objective aims to characterize immune repertoire these patients who have all undergone BK virus reactivation. 2) A second objective aims to characterize T-cell repopulation in these patients receiving alemtuzumab depletional induction and belatacept/sirolimus maintenance therapy.

3.3 Biological Motivation - Organ Transplantation

3.3.1 Significance

The adverse side-effects of immunosuppressive treatments in organ transplantation constitutes an important clinical problem due to the markedly lower survival rate in transplant patients compared to age-matched healthy controls [14], and the associations of immunosuppressive drugs with long-term infection and comorbidities of vascular diseases [15-18]. Immunosuppressive drugs are administered to transplant recipients to reduce the risk of acute organ rejection by impairing the helper, activation, proliferation, and effector functions of T-cells. This reduction of acquired immunity, however, may also result in higher susceptibility to viral infection. Therefore, immunologic monitoring is an important part of the clinical workup for transplant individuals.

3.3.2 Intrinsic Risk Profile

Decreased acquired immunity is further compounded by an individual's intrinsic risk profile (Figure 4), which is a function of age, environmental factors, and past exposure to pathogens. Over time, environmental antigen exposure provokes a patient's T-cell repertoire to be memory enriched, resulting in a decreasing risk of primary viral infection, and concomitant increasing risk of allograft rejection due to the effect of heterologous alloreactivity [19]. During aging, persistent exposure to environmental challenges drives the T-cell repertoire into one dominated by exhausted and senescent immunophenotypes. This exposure increases the risk of viral reactivation, lessens risk of rejection, but increases risk of immunosuppressive drug toxicity. The dynamic changes across the T-cell repertoire are highly individualistic. Therefore, understanding these phenotypic and functional changes may require a personalized approach that considers these fluctuations with regards to a patient's clinical outcomes and risk for rejection and infection.

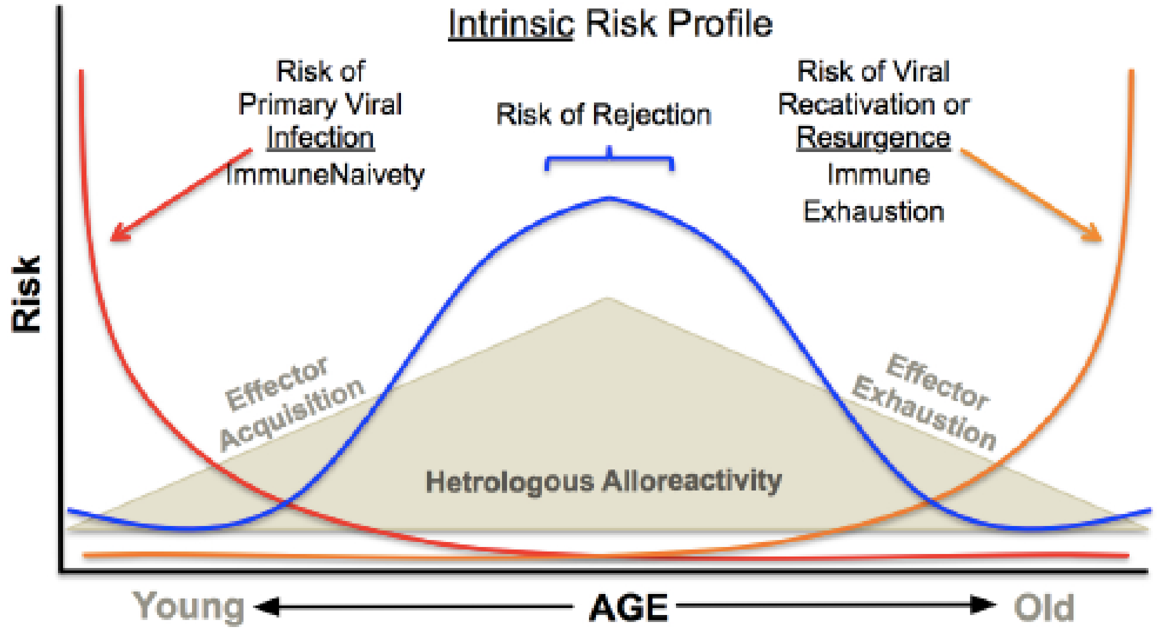


Figure 20. *Intrinsic risk profile.* Owing to heterologous alloreactivity, an individual's intrinsic risk for rejection or opportunistic infection is a function of age, environmental factors, and past pathogen exposure. (Image courtesy of Dr. Alan Kirk)

3.4 Role of Multidimensional FCM Analysis

Characterizing a transplant recipient's immune repertoire is critical to understanding the mechanisms behind immunologic transplant failure and potential viral infection. Mutidimensional FCM is frequently used to quantify the immune repertoire during the course of transplantation and recovery. Based on the information in Section 3.3, I define immune repertoire to be the diverse group of cells involved in the mediation of risk of rejection in organ transplantation or risk of opportunistic viral infection. These cells predominantly include naïve T-cells, which are activated upon initial exposure to foreign substances, and proliferate and expand to effector and memory T-cell populations [20]. As previously mentioned, chronic exposure to viral infection and age-related decline of the thymus function also leads to exhausted and senescent T-cell types. FCM The data analysis pipeline can be applied to assess a transplant recipient's immune repertoire over time in single patients and in a cross-sectional patient population with diverse diagnostic and demographic characteristics. The main assumption in using such a pipeline is that the mechanisms driving complications in organ transplantation intersect in a way that is both anticipatable and measurable.

3.5 Study Descriptions

3.5.1 Patient Cohort

The patient cohort currently consists of 19 patients, aged 18 and older, undergoing non-HLA-identical living donor renal transplantation. Patients receive an infusion of alemtuzumab prior to transplantation as a T-cell depletion therapy, followed by a combination of belatacept and sirolimus for immune maintenance therapy post-transplant. To date, a total of 19 patients have been recruited and followed-up to five years.

3.5.2 Dataset

Peripheral blood mononuclear cells (PBMCs) were obtained from eighteen of the nineteen renal transplant patients, for 5-10 time points, during a period of 24-36 months.

Each timepoint consisted of 14 individual tubes. For each tube, 8 fluorescent channels were used to measure the expression of cell surface receptors, in addition to 4 scatter channels (2 each corresponding to the area and height pulse of FSC and SSC channel). Therefore, each tube yielded 12-dimensional data with 100,000-300,000 cells.

3.5.3 BK Virus Study

The BK virus (BKV) is a widespread pathogen commonly affecting renal transplant recipients. Aggressive immunosuppressive therapies have been known to impair BKV-specific immunity, predisposing an individual to viral reactivation. Among transplant cases with reactivated BKV, 1-10% potentially develops BKV-associated nephropathy, leading to progressive renal injury and acute allograft rejection [40, 41]. BKV is often asymptomatic and detected in the blood through PCR and through histological examination of the kidneys [42]. To date, there is no standard treatment for BKV.

Using BKV as a driving biological project, I aim to identify and quantify longitudinal patterns (e.g. shape, count, rate of growth) of *all possible* cell populations. Comparing the degree of immunosuppression in BKV patients with non-BKV patients could also yield interesting information regarding immune repertoire unique to BKV and risk for BKV-associated nephropathy. The feature selection portion, utilizing a variant of the regularization procedure, Lasso, will be important in determining which subset of features can best predict and explain viral activation.

3.5.4 T-cell Repopulation Study

Numerous phenotypes are being discovered in relation to maturation, exhaustion, and senescence, which naturally occur as an immune system expands and contracts to meet environmental needs. In the clinical study, patients are administered the drug, alemtuzumab, which impairs T-cell functionality. It is expected that T-cell activation will burst following depletion. However, the kinetics and functional characteristics following this event are poorly understood. Using the automated analysis pipeline, this study will quantify patterns among T-cell activation states upon repopulation and compare longitudinal trends with baseline (pre-transplant). Findings will be compared to a control group (e.g. patients administered alternative immunosuppressive regimens).

Concluding Remarks

The work proposed in this thesis is significant in two aspects: 1) it presented the development of an objective quantitative framework for analyzing high-throughput multidimensional FCM data, and 2) a validation study showed that the results of the analysis pipeline have good agreement with traditional, manual methods.

The pipeline consisted of five key steps: 1) data preprocessing, 2) automated gating, 3) automated labeling, 4) feature extraction, and 5) feature selection. A robust quality assessment step was implemented at every step of the pipeline to account for potential source of systematic error prior to entering a subsequent step in the pipeline. My approach to this pipeline was aimed at addressing the challenges facing the FCM data analysis bottleneck. The integration of state-of-the art informatics techniques into a single pipeline shows great potential for scalability to hundreds of thousands of events and multiple channel dimensions.

Secondly this pipeline showed good performance against human validation. Cell populations obtained from a clinical study were similar to those obtained with the automated pipeline. Furthermore, results were reproducible and accurate.

Future studies will examine hypotheses with respect to the driving biological problem of risk of rejection and viral activation in organ transplantation. Studies will investigate longitudinal patterns of the immune repertoire of these patients as it relates to their clinical characteristics.

References

1. Lugli, E., M. Roederer, and A. Cossarizza, *Data Analysis in Flow Cytometry: The Future Just Started.* Cytometry Part A, 2010. **77A**(7): p. 705-713.

2. Chattopadhyay, P.K., et al., *Quantum dot semiconductor nanocrystals for immunophenotyping by polychromatic flow cytometry.* Nature medicine, 2006. **12**(8): p. 972-977.

3. Preffer, F. and D. Dombkowski, *Advances in Complex Multiparameter Flow Cytometry Technology: Applications in Stem Cell Research.* Cytometry Part B-Clinical Cytometry, 2009. **76B**(5): p. 295-314.

4. Ratajczak, P., et al., *Th17/Treg ratio in human graft-versus-host disease.* Blood, 2010. **116**(7): p. 1165-1171.

5. Ribeiro, A., et al., *EUS-guided fine-needle aspiration combined with flow cytometry and immunocytochemistry in the diagnosis of lymphoma.* Gastrointestinal Endoscopy, 2001. **53**(4): p. 485-491.

6. Aghaeepour, N., et al., *Early immunologic correlates of HIV protection can be identified from computational analysis of complex multivariate T-cell flow cytometry assays.* Bioinformatics, 2012. **28**(7): p. 1009-16.

7. Talbot, D., *Flow cytometric crossmatching in human organ transplantation.* Transpl Immunol, 1994. **2**(2): p. 138-9.

8. Lizard, G., *Flow Cytometry analyses and bioinformatics: Interest in new softwares to optimize novel technologies and to favor the emergence of innovative concepts in cell research.* Cytometry Part A, 2007. **71A**(9): p. 646-647.

9. Bashashati, A. and R.R. Brinkman, *A Survey of Flow Cytometry Data Analysis Methods.* Advances in bioinformatics, 2009. **2009**: p. 584603.

10. Keeney, M., D. Barnett, and J.W. Gratama, *Impact of standardization on clinical cell analysis by flow cytometry.* J Biol Regul Homeost Agents, 2004. **18**(3-4): p. 305-12.

11. Achuthanandam, R., et al., *Sequential univariate gating approach to study the effects of erythropoietin in murine bone marrow.* Cytometry A, 2008. **73**(8): p. 702-14.

12. Chan, C., et al., *Statistical mixture modeling for cell subtype identification in flow cytometry.* Cytometry A, 2008. **73**(8): p. 693-701.

13. Jeffries, D., et al., *Analysis of flow cytometry data using an automatic processing tool.* Cytometry Part A, 2008. **73A**(9): p. 857-867.

14. Dobbels, F., et al., *Growing pains: non-adherence with the immunosuppressive regimen in adolescent transplant recipients.* Pediatr Transplant, 2005. **9**(3): p. 381-90.

15. Kainberger, F., et al., *[Renal osteodystrophy: the spectrum of the x-ray symptoms in modern forms of kidney transplantation and long-term dialysis therapy].* Rofo, 1992. **157**(5): p. 501-5.

16. Salvatierra, O., Jr., M. Millan, and W. Concepcion, *Pediatric renal transplantation with considerations for successful outcomes.* Semin Pediatr Surg, 2006. **15**(3): p. 208-17.

17. London, N.J., et al., *Risk of Neoplasia in Renal-Transplant Patients.* Lancet, 1995. **346**(8972): p. 403-406.

18. Kasiske, B.B., CM., *Cardiovascular risk factors associated with immunosuppression in renal transplantation.* Transplantation Reviews, 2002. **16**(1): p. 1-21.

19. Adams, A.B., et al., *Heterologous immunity provides a potent barrier to transplantation tolerance.* J Clin Invest, 2003. **111**(12): p. 1887-95.

20. Jameson, S.C. and D. Masopust, *Diversity in T cell memory: an embarrassment of riches.* Immunity, 2009. **31**: p. 859-71.

21. Gentleman, R.C., et al., *Bioconductor: open software development for computational biology and bioinformatics.* Genome Biology, 2004. **5**(10).

22. Bashashati, A., et al., *A pipeline for automated analysis of flow cytometry data: preliminary results on lymphoma sub-type diagnosis.* Conference proceedings : ... Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Conference, 2009. **2009**: p. 4945-8.

23. Costa, E.S., et al., *A new automated flow cytometry data analysis approach for the diagnostic screening of neoplastic B-cell disorders in peripheral blood samples with absolute lymphocytosis.* Leukemia, 2006. **20**(7): p. 1221-30.

24. Shulman, N., et al., *Development of an automated analysis system for data from flow cytometric intracellular cytokine staining assays from clinical vaccine trials.* Cytometry Part A, 2008. **73A**(9): p. 847-856.

25. Roederer, M., et al., *Probability binning comparison: a metric for quantitating multivariate distribution differences.* Cytometry, 2001. **45**(1): p. 47-55.

26. Finak, G., et al., *Optimizing transformations for automated, high throughput analysis of flow cytometry data.* BMC bioinformatics, 2010. **11**: p. 546.

27. Hahne, F., et al., *Per-channel basis normalization methods for flow cytometry data.* Cytometry. Part A : the journal of the International Society for Analytical Cytology, 2010. **77**: p. 121-31.

28. Parks, D.R., M. Roederer, and W.A. Moore, *A new "Logicle" display method avoids deceptive effects of logarithmic scaling for low signals and compensated data.* Cytometry Part A, 2006. **69A**(6): p. 541-551.

29. Dendrou, C.A., et al., *Fluorescence intensity normalisation: correcting for time effects in large-scale flow cytometric analysis.* Adv Bioinformatics, 2009: p. 476106.

30. Bolstad, B.M., et al., *A comparison of normalization methods for high density oligonucleotide array data based on variance and bias.* Bioinformatics, 2003. **19**(2): p. 185-193.

31. Aghaeepour, N., et al., *Rapid cell population identification in flow cytometry data.* Cytometry. Part A : the journal of the International Society for Analytical Cytology, 2011. **79**: p. 6-13.

32. Luxburg, U.v., *A tutorial on spectral clustering.* Statistics and Computing, 2007. **17**(4): p. 395–416.

33. Zare, H., et al., *Data reduction for spectral clustering to analyze high throughput flow cytometry data.* BMC bioinformatics, 2010. **11**.

34. Dongen, S.V., *Graph clustering via a discrete uncoupling process.* SIAM Journal on Matrix Analysis and Applications, 2008. **30**(1): p. 121-141.

35. Everitt, B., S. Landau, and M. Leese, *Cluster analysis*. 4th ed2001, London. New York: Arnold ; Oxford University Press. viii, 237 p.

36. Tibshirani, R., *Regression shrinkage and selection via the Lasso.* Journal of the Royal Statistical Society Series B-Methodological, 1996. **58**(1): p. 267-288.

37. Zou, H. and T. Hastie, *Regularization and variable selection via the elastic net.* Journal of the Royal Statistical Society Series B-Statistical Methodology, 2005. **67**: p. 301-320.

38. Casale, T.B., et al., *Omalizumab pretreatment decreases acute reactions after rush immunotherapy for ragweed-induced seasonal allergic rhinitis.* J Allergy Clin Immunol, 2006. **117**(1): p. 134-40.

39. Bland, J.M. and D.G. Altman, *Statistical methods for assessing agreement between two methods of clinical measurement.* International Journal of Nursing Studies, 2010. **47**(8): p. 931-936.

40. Schachtner, T., et al., *BK Virus-Specific Immunity Kinetics: A Predictor of Recovery From Polyomavirus BK-Associated Nephropathy.* American journal of transplantation : official journal of the American Society of Transplantation and the American Society of Transplant Surgeons, 2011. **11**: p. 2443-52.

41. Dall, A. and S. Hariharan, *BK virus nephritis after renal transplantation.* Clinical journal of the American Society of Nephrology : CJASN, 2008. **3 Suppl 2**: p. S68-75.

42. Randhawa, P.S., et al., *Immunoglobulin G, A, and M responses to BK virus in renal transplantation.* Clinical and vaccine immunology : CVI, 2006. **13**: p. 1057-63.