Machine Learning Methods for Flow Cytometry Analysis and Visualization

2018

Emily Sassano
University of Central Florida

MACHINE LEARNING METHODS FOR MULTIPARAMETER FLOW CYTOMETRY ANALYSIS AND VISUALIZATION

by

EMILY SASSANO
B.S. University of Central Florida, 2007
M.S. University of Central Florida, 2013

A dissertation submitted in partial fulfilment of the requirements
for the degree of Doctor of Philosophy
in the Department of Computer Science
in the College of Engineering & Computer Science
at the University of Central Florida
Orlando, Florida

Summer Term
2018

Major Professor: Sumit K. Jha

ABSTRACT

Flow cytometry is a popular analytical cell-biology instrument that uses specific wavelengths of light to profile heterogeneous populations of cells at the individual level. Current cytometers have the capability of analyzing up to 20 parameters on over a million cells, but despite the complexity of these datasets, a typical workflow relies on subjective labor-intensive manual sequential analysis. The research presented in this dissertation provides two machine learning methods to increase the objectivity, efficiency, and discovery in flow cytometry data analysis.

The first, a supervised learning method, utilizes previously analyzed data to evaluate new flow cytometry files containing similar parameters. The probability distribution of each dimension in a file is matched to each related dimension of a reference file through color indexing and histogram intersection methods. Once a similar reference file is selected the cell populations previously classified are used to create a tailored support vector machine capable of classifying cell populations as an expert would. This method has produced results highly correlated with manual sequential analysis, providing an efficient alternative for analyzing a large number of samples.

The second, a novel unsupervised method, is used to explore and visualize single-cell data in an objective manner. To accomplish this, a hypergraph sampling method was created to preserve rare events within the flow data before divisively clustering the sampled data using singular value decomposition. The unsampled data is added to the discovered set of clusters using a support vector machine classifier, and the final analysis is displayed as a minimum spanning tree. This tree is capable of distinguishing rare subsets of cells comprising of less than 1% of the original data.

I dedicate this to my constant companion and most purfect friend, my sweet MeMurton.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABBREVIATIONS

APC Antigen Presenting Cell

CD Cluster of Differentiation

CV Coefficient of Variation

DC Dendritic cell

FCS Flow Cytometry Standard

FSC Forward scatter

HTS High Throughput System

ICCS Intracellular Cytokine Staining assay

IQR Interquartile range

LVS Live Vaccine Strain

mAbs Monoclonal antibodies

MHC Major histocompatibility complex

PBMC Peripheral Blood Mononuclear Cells

PCA Principal Component Analysis

SSC Side scatter

SVD Singular Value Decomposition

SVM Support Vector Machine

T_H T helper cell

CHAPTER 1: INTRODUCTION

Over the last century and a half, our understanding of the function and structure of the human immune system has grown tremendously. However, in the last 50 years, this knowledge has exploded due to the development of high dimensional flow cytometry. Flow cytometry is a popular analytical cell-biology technique that uses specific wavelengths of light to profile cell suspensions on a single cell basis. Most other analytical techniques are only capable of a population level analysis, while flow cytometry is capable of analyzing up to 20 parameters on a single cell. This high-resolution dataset arguably makes flow cytometry the most powerful tool to study cell phenotypes, identify antigen-specific cells, tumor-specific cells, understanding cell physiology, and developing new vaccines and therapeutics. However, there are drawbacks.

Despite the complexity of cytometry datasets, a typical workflow relies on subjective labor-intensive manual sequential analysis. In this method, one-dimensional histograms or two-dimensional dot plots of the data are generated. Within these plots, the researcher visually identifies the populations of interest. A gate (polygon) is drawn, simply using a mouse on the computer screen, to encompass the cells that require further analysis. While there are a plethora of different computational methods that can produce similar results none fully replicate the manual gating that has been the standard for over half a century. Because of this investigators still rely on manual methods of data analysis to keep statistics consistent over decades of historical experiments. We developed a histogram matched support vector machine gating method to overcome this.

Histogram matched support vector machine, a supervised learning method, utilizes the vast amount of previously analyzed data to analyze new flow cytometry files containing similar parameters. The probability distribution of each dimension in a file to be analyzed is matched to each related dimension of a reference file through color indexing and histogram intersection methods. Once a similar

reference file is selected the cell populations previously classified are used to create a tailored support vector machine capable of classifying cell populations as an expert analyst would. This method has produced results that are highly correlated with those of manual sequential analysis, providing an efficient alternative for analyzing a large number of samples.

Illustrating the usefulness of the histogram matching support vector methodology a cell viability analysis algorithm is presented. Using the light scattering properties of the cells (forward and side scatter) measured by the flow cytometer, live cell events can be classified and viable cells per mL are calculated. This method counted viable cells with a correlation coefficient of 0.97 and less than a 20% difference compared to manual cell viability counting. This method was also the most reproducible when compared to three commercially available instruments with less than a 5% coefficient of variation.

Another drawback of high-dimensional flow cytometry data analysis is the lack of data exploration. Manual gating is performed on only one or two dimensions at a time, limiting the ability to explore every cell population over all dimensions fully. This restricts the possibility of discovering new cell phenotypes. Current visual methods that aim to explore and visualize the dataset rely on user-defined parameters such as the number of cell populations. These methods also use sampling or clustering methods that introduce stochastic variation into the results making them unreproducible from run to run. The second machine learning method presented aims to address these problems.

A novel unsupervised method is used to explore and visualize single-cell data in an objective manner that is free of user-defined parameters. A variety of exploratory gating techniques have been investigated in the past but have seen slow acceptance due to memory resource restrictions, consistency of results, and the need to rely on user-defined parameters. The goal of this method was to allow for this exploration of data to be processed on a typical desktop computer. To accomplish this, the proposed method uses a hypergraph sampling method to preserve rare events within the

flow cytometry data file before clustering the sampled events using singular value decomposition as a method to divisively cluster cell populations.

Demonstrating the usefulness of this method files were analyzed from an *in vitro* culture system aimed to generated antigen-specific human CD4+ T cell responses. It was possible to differentiate and visualize the heterogeneity of effector CD4+ T helper cells responding to three different vaccine conditions. Th1 and Th17 T helper cells can be visualized within different datasets. These rare cell populations comprised less than 2% of the cells collected in a $300,000$ event cytometry file. This novel method should provide a more reliable and standardized approach for the analysis of today's high-resolution flow cytometry datasets.

The remainder of the document is organized as followed. Chapter 2 gives background information. Briefly, the human immune system will be introduced, describing the different cell populations, cell types, and cellular markers that will be discussed throughout this document. The technique of flow cytometry including components, traditional data analysis, and common computational methods will be addressed. Chapter 3 will focus on a supervised learning method called Histogram Matching Support Vector Machine Gating to capture cell populations as a trained analysis would. Chapter 4 presents an unsupervised exploratory flow cytometry analysis method able to discover rare subsets of cells within a dataset visually. Finally, we will take a look at the overall results of both machine learning methods and explore a few open problems for future work.

CHAPTER 2: BACKGROUND

Immunology

The immune system is a collection of molecules, cells, and tissues that work together to prevent and eradicate infections from foreign invaders, such as bacteria, viruses, and parasites. The immune system can be separated into two distinct lines of defense, innate and adaptive immunity. Innate immunity is the first to assault an intruder and offers initial protection from pathogens. Epithelial, phagocytes, natural killer cells, plasma proteins, stomach acidity, and enzymes are all players of the innate immune system poised to prevent an attack. Adaptive immunity is the second line of defense against infection. It develops more slowly, but it is capable of a tailored specific attack. The cells of the adaptive immune system called lymphocytes can specifically recognize at least one billion different foreign substances called antigens. It is this group of cells that is often the subject of flow cytometry analysis.

There are two types of adaptive immunity called humoral immunity and cellular immunity, mediated by distinct cell populations in defending against extracellular and intracellular pathogens respectively. Humoral immunity is mediated by proteins called antibodies or immunoglobulins produced by B lymphocytes. These proteins are capable of neutralizing and eliminating microbes and toxins. B lymphocytes mature in the bone marrow expressing a unique membrane-bound immunoglobulin generated from a random rearrangement of a series of gene segments. If a naïve B cell is fortunate enough to encounter a foreign antigen capable of binding its immunoglobulin receptor the signal produced initiates its activation. Once activated, the lymphocyte divides rapidly and differentiates into an antibody-secreting effector cell. The antibodies generated by these B cells are very efficient in eliminating extracellular threats, but they do not have access to microbes that live and divide within cells.

Defense against intracellular microbes is the territory of cell-mediated immunity and is mediated by T lymphocytes. T cells, like B cells, also arise in the bone marrow but migrate to the thymus to mature. T cells consist of two distinct subpopulations; T cytotoxic and T helper cells characterized by surface cluster differentiation (CD) glycoproteins. T cytotoxic cells display CD8 glycoprotein on their surface and exhibit cell-killing or cytotoxic activity. Cells infected with intracellular microbes are killed by $CD8^+$ T cytotoxic cells to eliminate reservoirs of infection. Identified by the expression of the CD4 glycoprotein on their surface, T helper or T_H cells, are involved in activating macrophages to kill phagocytosed microbes, in the activation and growth of $CD8^+$ T cells, and in B cell activation. Most T cells have specific receptors that can only recognize foreign antigen bound to cell membrane proteins called major histocompatibility complex (MHC). Naïve T cells recognize the antigen-MHC complex on a professional antigen presenting cells (APC), such as dendritic cells or B cells. These cells express MHC class II molecules on their surface and can deliver the costimulatory signal necessary for $CD4^+$T cell activation [1, 26, 35].

In a primary immune response, some proliferating B and T cells migrate into a primary lymphoid follicle where they continue to divide and form a germinal center. Proliferating B cells comprise the majority of the germinal center lymphocytes with antigen-specific T cells making up approximately 10%. With T cell help, B cells undergo critical modifications, including somatic hypermutation (altering the variable regions of B cells), affinity maturation (selecting cells with high affinity for an antigen), and isotype switching (forming antibodies of one of the five different classes). The surviving B cells of a germinal center reaction then differentiate into plasma cells or memory cells. Plasma cells secrete antibody at a high rate and eventually migrate to the bone marrow. Memory B cells are long-lived cells that do not emit antibody but are prime for a secondary attack against the same antigen [1, 26, 35]. In addition to their critical role in the adaptive immune system B cells are also used to generate the specific antibodies needed for flow cytometry.

T cells become activated in response to being presented with peptide antigens by MHC molecules

expressed on the surface of APCs. When T cells become activated, they divide and produce proteins that modulate the immune system called cytokines. The patterns of cytokines a T cell produces can be used to differentiate them into different lineages. T_H cells, typically express CD4 and can be grouped into at least four lineages. These lineages include T_H1, T_H2, T_H17, and Treg [17,35,63]. The first two major lineages discovered by Mosmann and Coffman were the T_H1 and T_H2 subsets [34]. T_H1 cells can be identified by their high level of IFN-γ as well as moderate amounts of IL-2 and TNF-α production. T_H1 cells are produced to protect against intracellular infections. T_H2 cells do not produce IFN-γ but tend to produce IL-4, IL-5, TNF-α and to a lesser degree IL-2. The Th17 lineage was not described until 2003 and are characterized by IL-17 production [24, 41]. While these cytokine production profiles are good basic classifications, it has been shown that the heterogeneity of cytokine patterns within each T_H lineage is tremendous and that differentiation into one lineage dose not preclude the T cell from acquiring the ability to produce other T_H specific cytokines [39]. Cytokine production along with the expression of CD154 or CD40 ligand is associated with the antigen-specific activation of $CD4^+$ T cells. CD154 binds to CD40 on the surface of APCs, a critical cell to cell interaction for the development of T cell effector function. The combination of CD154 and cytokines production by T cells during activation is an essential area of research for vaccine development. Polyfunctional $CD4^+$ T cells, a subset of antigen-specific $CD4^+$ cells, simultaneously produce multiple effector cytokines and CD154 in response to activation have repeatedly been linked to the positive clinical outcome of vaccine trials [16, 19, 40, 53]. Identifying and visualizing the differences within the small population of effector T cells is an important step in further correlating *in vitro* results to clinical trial outcomes.

Throughout this document many of these cell types will be revisited within the context of an *in vitro*, taking place outside of a living organism, human cell culture. While detailed knowledge of immune cell activation, antibody production, and germinal center formation is not critical for the continued understanding of this text, it is essential to have an appreciation of the broad diversity of

different cell types of the immune system and the sophisticated technology that we use to categorize them.

Flow Cytometry

This popular analytical cell-biology technique uses specific wavelengths of light to profile heterogeneous populations of cells at the individual level. Most other analytical methods are capable only of measuring samples on a population level, highlighting flow cytometry as a unique and indispensable tool in research and clinical applications. Flow cytometers are used ubiquitously in biomedical research labs (immunology, cancer biology, neurobiology, molecular biology, microbiology), diagnostic laboratories (HIV/AIDS, transplant, tumor immunology), medical engineering (protein engineering, nanoparticles), and marine and plant biology [13, 18, 38, 44].

Modern flow cytometers consist of three main components; the fluidics system (sheath), optics (light source lasers, optical path filters, and light collection detectors), and electronics (detector signal processing, computer interface, data storage, and output). Once cells have been labeled with specific monoclonal antibodies that have been conjugated to fluorochromes. The fluidics system takes a suspension of single cells and hydrodynamically focuses it within a stream of fluid called the sheath. This steam of fluid then intersects the path of the lasers within the optic system. Detectors built to register specific wavelengths of light collect the light emitted from the fluorescent particles that label specific physical or chemical characteristics of a cell. These signals are then amplified and covered to a digital signal that a computer can display for analysis. The emitted light recorded from each cell creates a pattern of singles unique for different cell types based on what cellular marker each fluorochrome-antibody combination refers to.

While the hardware and reagents associated with this technology have undergone rapid advance-

ments over the last several decades, the software used to analyze these dense multidimensional data sets has not seen the same advancement. In addition to the slower development of analysis tool, there has been an even slower rate to adopt these techniques stifling this technology from reaching its full potential.

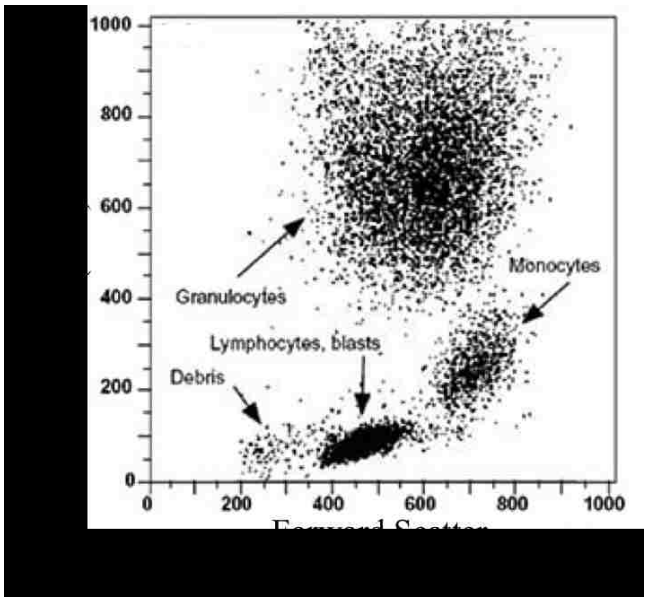


Figure 2.1: Dot plot of Forward (FSC) versus Side Scatter (SSC). Each dot in this plot represents a single cell that has been characterized by its size and granularity. These light scattering characteristics alone can classify multiple cell populations.

Scatter Parameters

Included in the possible 20 or more parameters a flow cytometer is capable of analyzing are simple light scattering properties of the cell. The ratio between the size of the cell and the laser's wavelength alters the scatter of light before reaching the detector. A cell larger than the laser's wavelength will show a higher intensity of scattering compared to a call smaller than the wavelength of the laser. Light scatter measured in the same path as the laser is referred to as forward

scatter (FSC). FCS is correlative to the diameter of the cell. The second light scatter measurement is taken as a 90° angle to the laser and is called side scatter (SSC). SSC gives information about the homogeneity or granularity of the cell. Intracellular components of the cell cause light to refract increasing the SSC intensity [55]. By using these two intrinsic characteristics of a cell, it is possible to classify many cell types of the immune system. Figure 2.1 illustrates this.

Antibodies and Fluorochrome Conjugation

Cell types such as CD4 and CD8 cells have been discussed previously but what does 'CD' mean? CD is short for Cluster of Differentiation. Cluster of differentiation or sometimes called cluster designation or classification determinant is a protocol used to identify cell surface markers to aid in the phenotyping of immune cells. The discovery of CD markers specific for cell function has been a primary force in pushing for the increase in flow cytometry dimensionality. CD markers function for the cells in numerous ways, often as a receptor or ligand. A receptor is a protein molecule that receives chemical signals to initiate some form of cellular response. A receptor only binds with ligands of a specific structure, similar to how a lock only accepts a specifically shaped key. When a ligand of the corresponding structure to a receptor binds, it activates or inhibits some biochemical pathway. A ligand is a substance that joins with a biomolecule to serve a biological purpose. This identification of receptors and ligands associated with cellular functions (cell signaling, cell adhesion, cell activation, cell inhibition) lead to CD nomenclature we use today. Established in 1982 to create a universal system of Human Leukocyte Differentiation Antigens (HLDA) classification [11, 14], this method uses monoclonal antibodies (mAbs) generated against specific epitopes of receptors and ligands of cells. An epitope is an area of an antigen that is recognized by the immune system. In the context of flow cytometry, this pertains to the ability of an antibody produced by a B cell to bind to a specific epitope or portion of the target of interest. Currently, more than 370 CD markers have been identified [64]. A CD marker is identified and

assigned a number designation once two specific mAb are shown to bind the molecule. CD marker identification has been a crucial part of the flow cytometry technology. The knowledge of what makers are or are not present on cells of specific function allows for immunophenotyping.

B cells, a component of the adaptive immune system discussed previously, can create proteins called antibodies to millions of different epitopes with a broad range of specificity. It is this adaptive ability that allows mammals to fight off and resist infection and is also the basis for immunodiagnostic assays such as flow cytometry. B cells produce immunoglobulin in five classes: IgG, IgM, IgA, IgE, and IgD with IgG being used most frequently in flow cytometry and other immune-based assays. The IgG immunoglobulin is composed of two types of protein chains: light and heavy. Each IgG molecule consists of two identical light, and heavy chains joined with disulfide bridges. The two chains can be further broken down into variable and constant regions. The variable domains of the light chain and heavy chain form the antigen binding site. When an animal's immune system detects and antigenic substance, the production of polyclonal antibodies is the result. These antibodies can bind to multiple epitopes of an antigen to offer better detection of the pathogen by the immune system. However, this ability is a hindrance for immune assays due to the wide range of binding specificities. To overcome this non-specificity and to reduce the background of assays monoclonal antibodies were developed. Monoclonal antibodies, in contrast, are the product of one B cell clone. The antibodies produced by a particular clone will have the same amino acid sequence in the variable regions of the antigen binding site. Therefore every antibody will behave identically in an immuno-based assay [18, 35, 55].

Conjugating, or attaching, particles that illuminate under specific wavelengths to monoclonal antibodies is what gives flow cytometry the power to phenotype individual cells of a heterogeneous mixture. A fluorochrome is a molecule that absorbs light energy of a specific wavelength and then reemits light of a longer wavelength. When a photon of energy from the laser of the flow cytometer hit a fluorescent molecule, an electron of the fluorochrome is promoted from its ground

energy state (S_0) to a higher unstable energy state (S_1). In this brief process, the electron loses some of the absorbed energy as heat (vibrational energy) as it begins to fall back to (S_0). As the electron returns completely to (S_0), light is emitted at a longer wavelength than was absorbed [30]. The difference in the wavelength of light causing excitation and the wavelength of light emitted from the fluorochrome is called the Stokes shift. Fluorochromes are engineered to have a specific Stokes shift to utilize one laser but emit at wavelengths different enough to be detected as independent signals by the detectors [55]. It is crucial to select fluorochromes whose excitation and emission spectra work optimally together with the particular laser/detector combinations of the flow cytometer. The use of multiple lasers each surrounded by many detectors able to record the emission of these chemically engineered fluorochromes is what gives flow cytometers the ability to measure multiple parameters on or in a single cell.

Data Pre-processing

Data output from the Flow Cytometer is organized into a file termed Flow Cytometry Standard (FCS) data file. The FCS file format has been developed and maintained by the International Society for Advancement of Cytometry (ISCA) [6]. The raw data within the FCS file is saved as a two-dimensional array. Each event (cell) forms the rows and with their raw fluorescence and scatter data represented as floating point or double precision floating point values in the columns. A second two-dimensional array is also stored within this data file representing the fluorescence spillover matrix needed for compensation.

All of the following pre-processing steps outlined in the following subsections have been applied to the flow cytometry data presented within this document.

Compensation

Each fluorochrome molecule used in flow cytometry has an excitation and an emission spectra. The excitation spectrum is the range of wavelengths that will cause the molecule to emit light. The emission spectrum is the range of wavelengths of this emitted light. It would be wonderful if every fluorochrome's excitation and emission spectra both consisted of a very narrow range of the light; however, fluorochrome chemistry is not perfect. Fluorochromes can be excited by multiple wavelengths and have emission picked up by multiple detectors causing spectral overlap. Fluorochrome combinations can be optimized to minimize this spectral overlap, but when using a high dimensional staining panel fluorochromes will often contribute signal on more than only their intended detector. To combat this, it is vital that with each analysis a series of control samples stained with only one of the fluorochromes used in the analysis is run on the cytometer. This quantifies unintended signal contributions from each fluorochrome on every detector. From this data, a matrix of relative spectral overlaps of each fluorochrome can be calculated. This data is typically stored within the FCS data file as a spillover matrix. This matrix specifies the values to calculate the compensated data from the raw fluorescence values held in the main data array. Each cell recorded by the cytometer is then multiplied by the inverse of this spectral overlap matrix to obtain its corrected emission estimates for each fluorochrome [48, 50]. However, cells with none or very low emission levels can be corrected to have negative fluorescence calculations. It is this characteristic of flow cytometry data that necessitates the need for data transformation.

Data Transformation

Most flow cytometry applications that use parameters beyond light scatter, FCS and SSC, require some method of data transformation to display cell populations adequately. After compensation is applied to account for spectral overlap, it is possible for values at the low end of the fluores-

cence range to become negative. Due to this a simple $\log$ transformation cannot accommodate these negative or zero values, and cell populations with low values appear squished to the axis. To display these cell populations accurately, a subset of biexponential functions called a Logicle scaling function has been developed for use in flow cytometry [42]. This transformation allows negative values and those close to zero to be displayed in the linear range [18, 37, 42, 56]. This method of data transformation is available and implemented as the default transformation function on analysis software such as FlowJo, Bioconductor, and Cytobank [21, 27].

Doublet Discrimination

Doublet events are the consequence of two cells becoming physically attached to each other, or two cells passing through the cytometer so close together that they are processed as a single event. When this happens with two cells of different phenotypes a population of cells can be produced that becomes misleading when interpreting the data. Therefore, doublets should be removed to generate the most accurate and sensitive analysis. The optical detectors that are used to detect fluorescence are called photomultiplier tubes (PMTs). PMTs read the light scatter of a cell or the light emitted from a fluorochrome. The PMT voltage can be increased or decreased to adjust for cell size or fluorescence intensity. A voltage pulse is processed by a detector for every cell that flows through the cytometer and is defined by area (A), height (H), and width (W). W is the time that the cell took to go through the laser's path and is proportional to the cell's size and the duration of the signal but is not impacted by PMT voltage. H is the intensity of the signal and is impacted by the PMT voltage. By using $A = H \times W$ doublets can be discriminated by detecting discrepancies between H, W, and A [62]. Seen in the first dot plot of figure 2.2 doublets (events outside of the polygon gate) have approximately double the A while H is the same.

Data Standardization

While data standardization is not a necessity in manual flow gating, it is generally required when using computational and machine learning methods. Light scattering parameter values typically fall on a scale of zero to $250,000$ while fluorochrome values, after compensation, will typically be on a scale exceeding these boundaries. For example, the average range of the fluorochrome parameters scale in figure 2.2 is $-2,000$ to $265,000$. However, even the scales of the fluorochrome parameters can differ due to PMT voltage settings, fluorochromes used, and compensation. Having the scale of all data parameters standardized ensures that every parameter is viewed equally in subsequent computational steps such as support vector machine classification and singular value decomposition. Because in singular value decomposition we are interested in what components contribute most to the variance of the data z-score shown in equation 2.1 was selected over other methods of standardization such as Min-Max scaling. The z-score of a single value x is calculated by finding the mean, μ, and the standard deviation, σ, of all data points for a parameter. This standardization is done for every parameter of the dataset.

$$z = \frac{x - \mu}{\sigma} \tag{2.1}$$

Manual Sequential Gating

The most popular method of flow cytometry analysis is manual sequential gating or hierarchical gating. This method uses one-dimensional histograms or two-dimensional dot or contour plots to visualize the data. From these plots, a researcher manually identifies populations of interest. A gate (polygon) is drawn, using a mouse on the computer screen, to encompass cells that require further analysis. Figure 2.2 shows an example gating strategy used in a ten-parameter FCS file

to analyze the $CD4^+$ T cell population's cytokine production profile. During this analysis, the researcher must draw on previous experience and knowledge of cellular markers and function to choose what parameters will best define the final cell population of interest during each hierarchal step of the analysis. Due only to the subjectivity of this analysis method, variation in the final quantitative cell population statistics has been shown to be between $17 - 44\%$ [7].

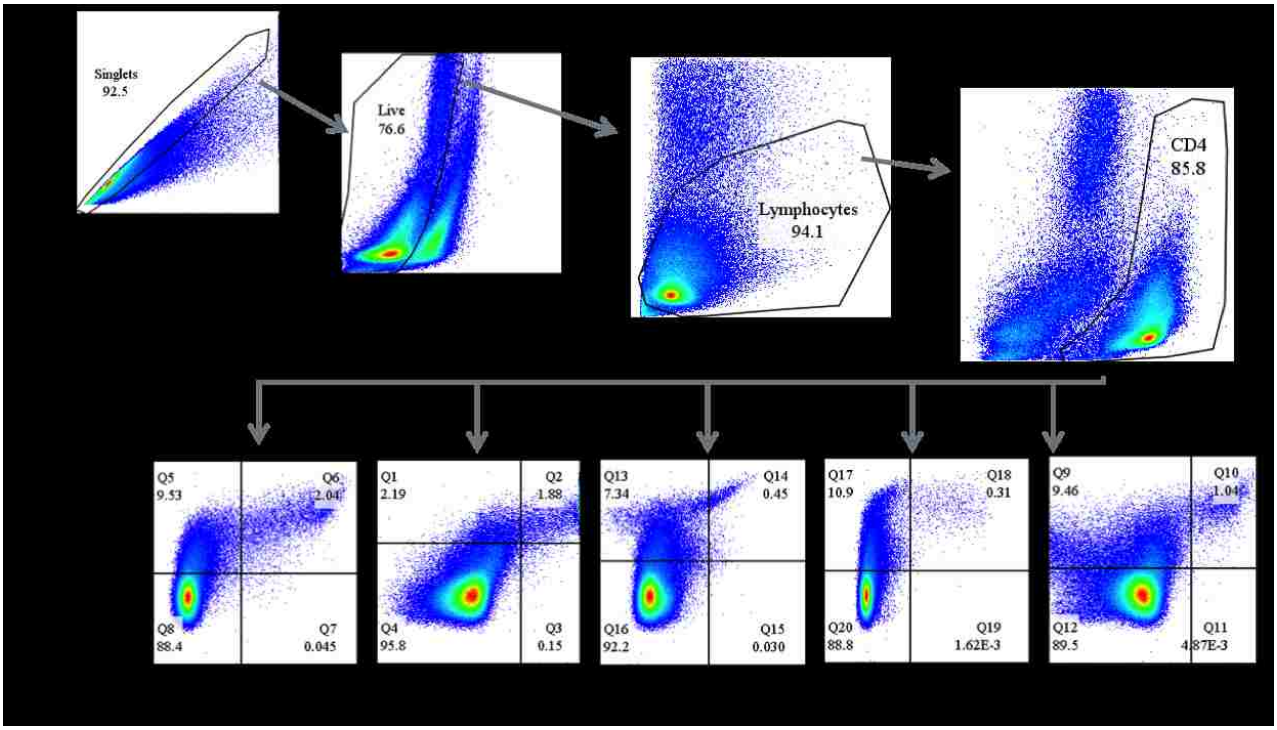


Figure 2.2: Manual analysis path is describing the gating strategy of a ten-parameter FCS file. The black polygon gates where drawn manually. The events captured within these gates are then selected to move to the next hierarchical gating step with the final goal of visualizing what cytokines are being produced by live, activated, $CD4^+$ T cells. These effector T cells are seen in the double positive quadrants of the second row of dot plots.

In addition to being a main contributing factor of experimental variability, manual gating can not truly explore all possible event populations within a high dimensional data set. If two-dimensional plots are used to explore a ten-dimensional data set, there will be 45 possible combinations for

gating the fist population alone. Every subsequent analysis step offers the same number of combinations to explore. The number of theoretical analysis pathways grows quickly to reach thousands of possibilities even with a modest number of sequential analysis steps. Seen in figure 2.3 are $CD4^+$ T cells displayed using Cytokine vs. $CD154$ dot plots. Using this method of analysis and visualization, it is impossible to determine what T cells are simultaneously making what combination of cytokines. This highlights the need for computational exploratory analysis methods as means of novel cell population discovery.

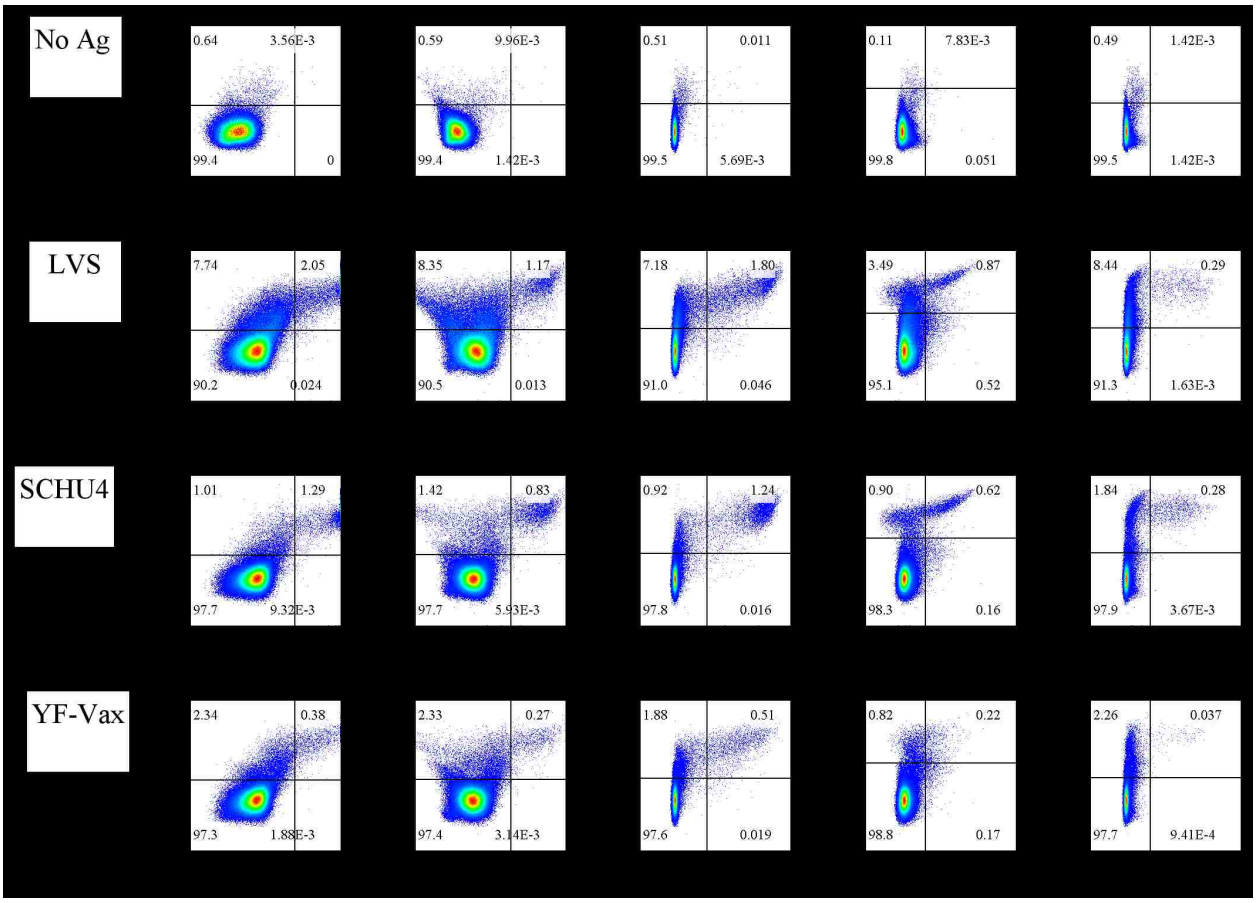


Figure 2.3: By only using dot plots it is not possible to display T cell ploy-functionality.

Current Computational Analysis Methods

Technological advances in lasers, detectors, monoclonal antibodies, and fluorochromes have allowed for the generation of flow cytometry data many orders of magnitude larger than in the past. Despite great analytical advancements, the adoption of computational analysis methods has been slow. While the numbers of parameters that can be measured on a signal cell have been increasing the primary method of analysis used in research is still simple manual sequential gating. This iterative process plots cells as two-dimensional scatterplots and based on experience, intuition, and a little bit of luck, the user defines a population of cells to be the focus of further analysis. This is done iteratively until the anticipated cell population is discovered. Manual analysis has been shown as the major source of variability in flow cytometry analysis. In a published study analyzing the causes of variability of flow cytometry, cells were identically cultured and stained and distributed to 15 laboratories to be acquired on their cytometers and for the resulting data to be analyzed. The coefficient of variation (CV) between the laboratories was 17-44% [7]. With this amount of variation coming from identical samples, it should be apparent why standardized automated analysis techniques need to be adopted. Despite such a clear need, the lack of easily accessible, executable, and understandable automated analysis methods has made researchers reluctant to switch practices.

Clustering

Clustering methods seek to identify a natural grouping of objects or with flow cytometry data, cells that have some shared combination of characteristics. Identifying a cluster is not a simple task due to that no definitive definition for a cluster even exists. Just as seen in a manual analysis of flow cytometry data, one person's or algorithm's interpretation of what defines a cluster can differ from another's.

K Means

The K means algorithm was the first in the literature to be used for the automated analysis of flow cytometry data [36]. While this method is easy to implement, quick to classify items into populations, and used with promising results in many fields, its application in flow cytometry analysis is limited. In its most simplistic implementation, it is limited in that the number of clusters must be specified by the user and the clusters discovered are limited to spherical populations. FlowPeaks and flowMean modified the traditional K means algorithm to allow the cluster number to be discovered and to create non-spherical cell populations [4, 20].

FLOw Clustering without K (FLOCK) uses a density-based clustering method to identify cell populations similar to K means. FLOCK uses a grid-based partitioning method to identify the densities of data points within the multidimensional data space of the FCS file [3,46]. This method has seen greater use and is now part of the Immunology Database and Analysis Portal (ImmPort: http://www.immport.org). It is the main FCM analysis software developed in the context of the Bioinformatics Integration Support Contract (BISC) by the NIH National Institute of Allergy and Infectious Diseases [12].

Model-Based Methods

The vast majority of automated gating approaches rely on modeling the data by using some parametric representation of its distribution. This is typically done with an unsupervised learning algorithm such as expectation maximization to fit the data to the model parameters. The methods follow a similar methodology of assuming the data is a collection of a finite number of populations (many times with this number input by the user), and that the data within each population can be defined using a standard statistical model. The use of a Gaussian mixture model (GMM) in

conjunction with an expectation-maximum (EM) algorithm is a conventional model-based clustering approach that has shown advantages over manual gating. Algorithms that use a model-based method to identify cell populations include FLAME, flowClust, flowMerge, flowGM, immunoClust and SWIFT [31,45].

Visualization

Methods that aid in both visualization and population discovery have been the most well-adopted among flow cytometry computational methods. Often still used in parallel with manual gating, these methods have aided in the discovery of new cell types and have helped us better understand of the heterogeneity of immune cells.

SPICE

SPICE is an analysis software developed for the quantitative analysis of polychromatic flow cytometry data and is used in conjunction with manual sequential analysis [49]. Despite its reliance on manual gating, this method is among the top methods to visualize and quantify phenotypic profiles within cell populations. Using basic Boolean logic (AND, NOT) over several markers of a cell population during manual gating analysis, it is possible to determine the co-expression of these markers. This method is often used in vaccine research and development to quantify polyfunctional $CD4^+$ or $CD8^+$ T cells, due to their presence correlating positively to protection after vaccination [17,40,53]. Figure 2.4 shows piecharts generated by SPICE from manually gated FCS files with Boolean logic gating. Here $CD4^+$ T cell cultures are analyzed for effector function. The slices of the pie charts represent the number of different cytokines (1-5) T cells are producing, and the arcs identify the specific cytokines. These charts display the same data as in figure 2.3 but with the ability to visualize T cell polyfunctionality.

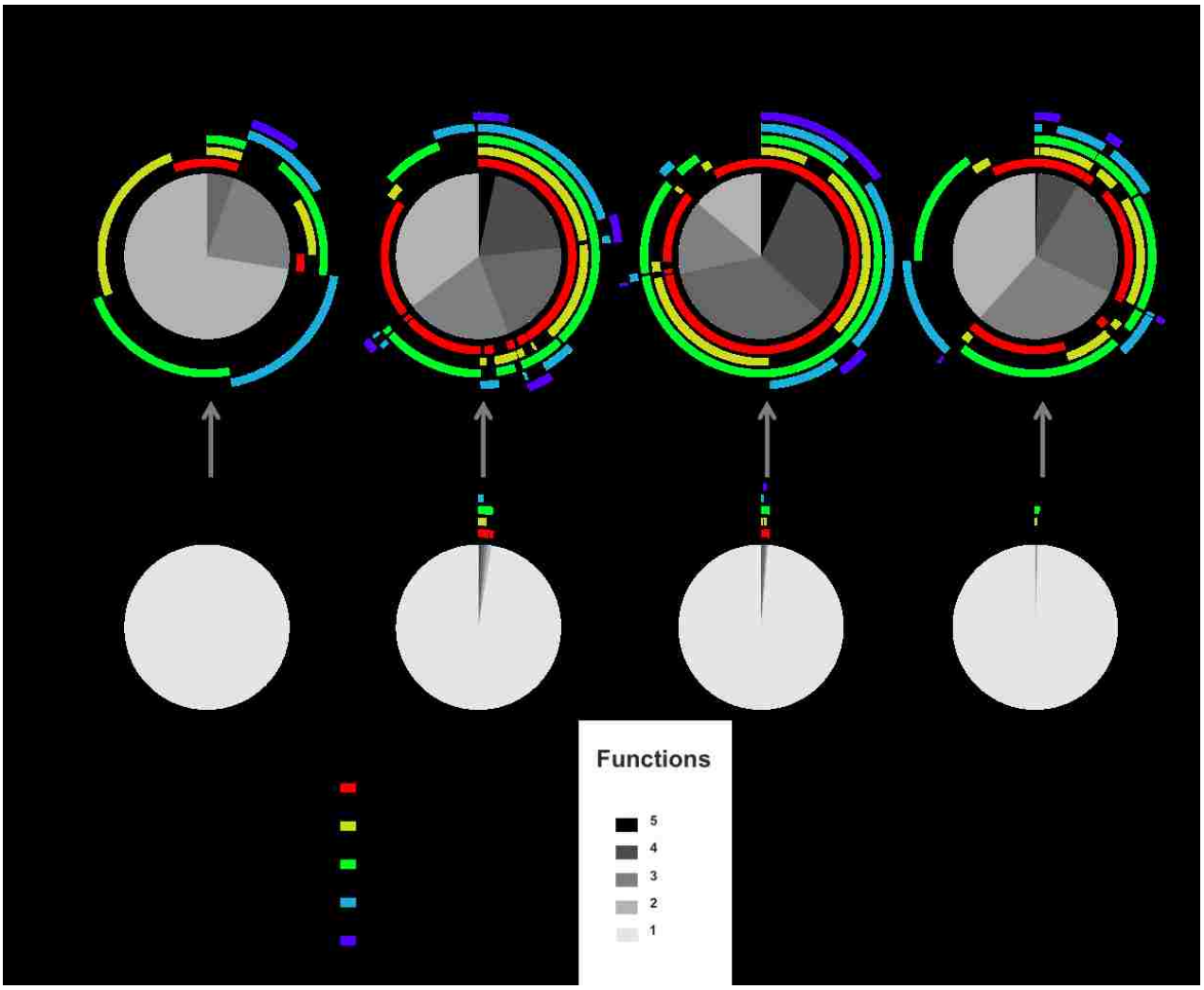


Figure 2.4: Graphical output from the SPICE analysis software. The pie charts in the second row represent the total $CD4^{+}CD154^{+}$ T cell population of each culture condition. The pie charts in the first row show the small fraction of these T cells which are also producing cytokines.

SPADE

Spanning-tree progression analysis of density-normalized events or SPADE is an unsupervised visualization method that converts multidimensional single-cell data down to a two-dimensional network of a user-defined number of interconnected cell populations. This method contains four

analysis steps. First, SPADE samples the dataset using user-defined thresholds for the cell density within the multidimensional space that identifies an outlier verses a cell population or target density. Cells that fall below the specified outlier density are not sampled, cells falling between outlier and target densities are all sampled, and those with a local density above the target density are randomly sampled to meet the target density. Next, a hierarchical amalgamative clustering method is used to iteratively group similar cells (based on distance matrix) into clusters until the number of clusters matches that which the user has defined. Third, a minimum spanning tree is created from the discovered clusters based on their median marker values. Finally, SPADE places all the cell from the original data into the cell cluster that its nearest neighbor belongs to [47]. Figure 2.5 shows the SPADE tree representation of the same data manually gated in figure 2.2. The target and outlier density was set so $10,000$ events would be sampled from the LVS FCS file, and the number of clusters was set to 100.

viSNE

viSNE currently is the most widely used method to visualize flow cytometry data. Using a technique called t-stochastic neighbor embedding (t-SNE), data points in high dimensional space are assigned a new position within a two or three-dimensional space [61]. In brief, the computational steps include [61]:

1. Uniform random sampling of $6,000$ - $12,000$ cells
2. Calculation of a pairwise distance matrix in high-dimensional space
3. Transformation of the distance matrix to a similarity matrix based on the probability that X_i will be a neighbor with X_j
4. Random mapping of data points into a two or three-dimensional space calculating its simi-

larity matrix

5. Minimization of the divergence between the high and low probability distributions using gradient descent to generate the final two or three-dimensional mapping

Figure 2.6 shows viSNE's representation of the same data manually gated in figure 2.2. These data plots were generated using Cytobank, a subscription cloud-based flow cytometry analysis tool, implementation of the viSNE algorithm [27]. The FCS data was gated previous to this analysis to exclude doublets, dead cells, and debris. The number of events selected for uniform random sampling was set at $10,000$ from the remaining $272,030$ of the FCS file.

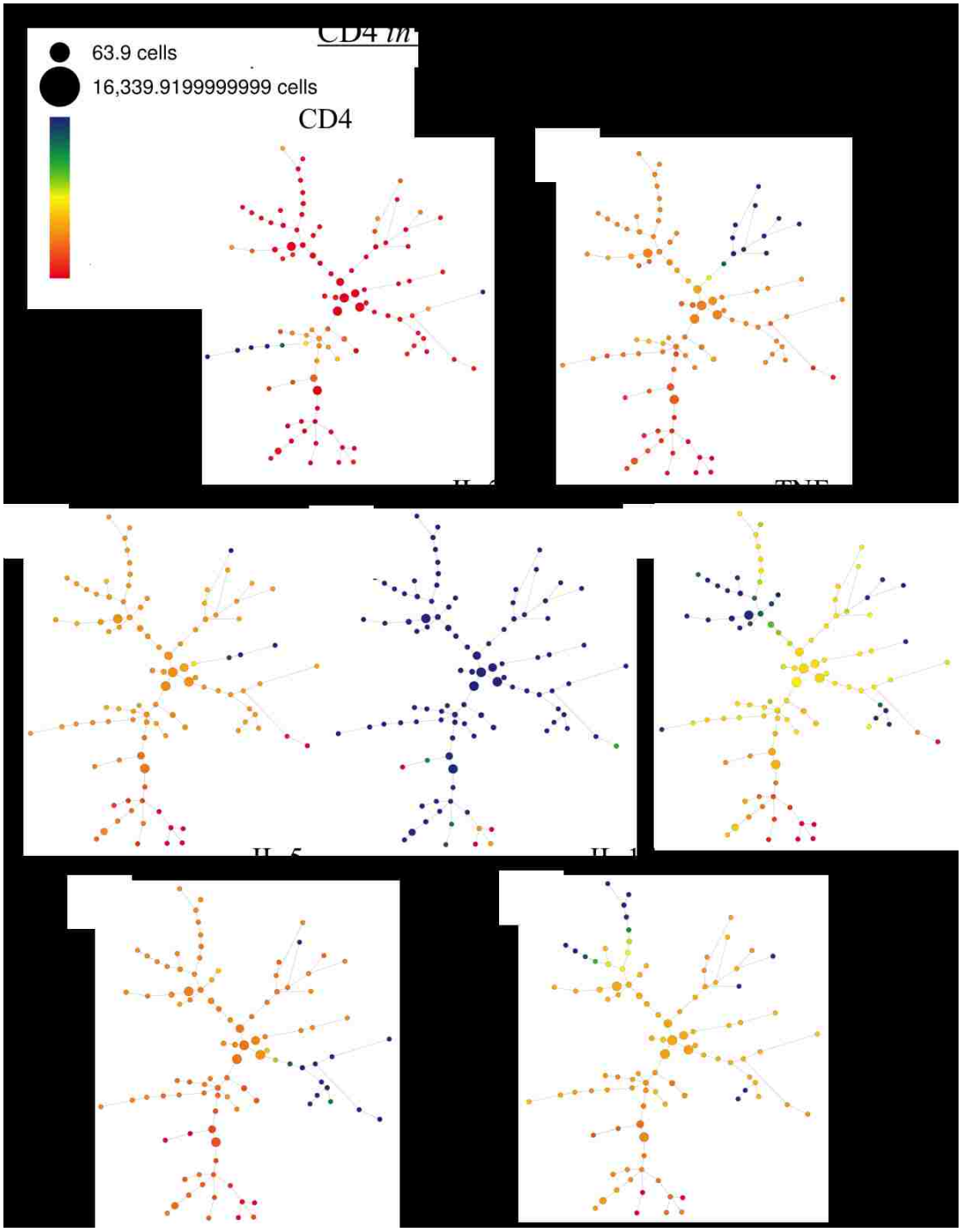


Figure 2.5: Graphical output from the SPADE algorithm implemented in Cytoscape. The target and outlier density was set so $10,000$ events were sampled from the LVS FCS file, and the number of clusters was set to 100. Data was gated previous to this analysis to exclude doublets, dead cells, and debris.

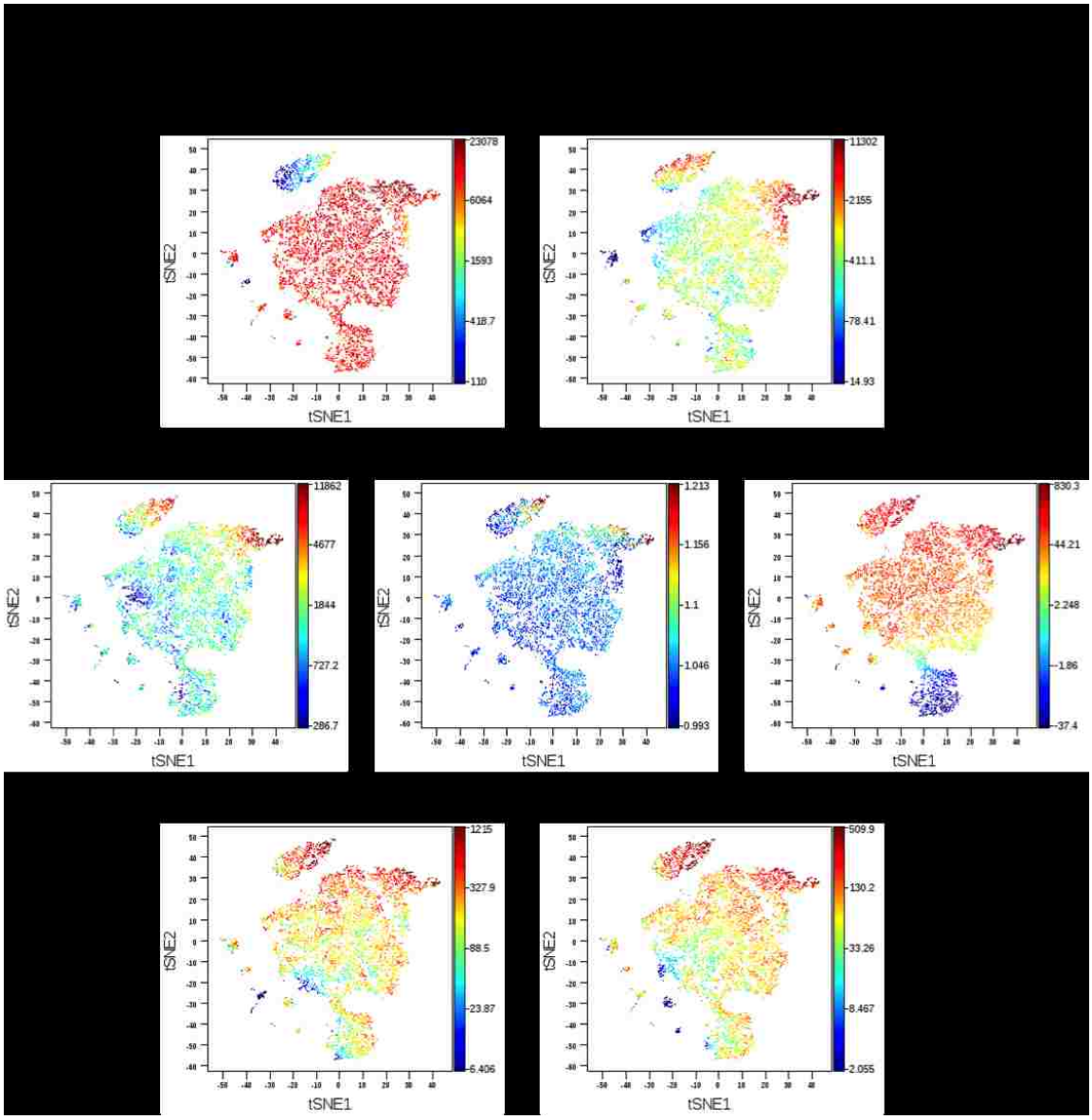


Figure 2.6: Graphical output from the viSNE algorithm implemented in Cytoscape using $10,000$ sampled data points of the LVS *in vitro* culture FCS file. viSNE uses a uniform random sampling of FCS data in high dimensional and maps it to a two-dimensional space. The cluster dot plots that are produced are colored to display areas marker intensity. Data was gated previous to this analysis to exclude doublets, dead cells, and debris.

CHAPTER 3: HISTOGRAM MATCHED SUPPORT VECTOR MACHINE

Many computational methods aim to discover cell populations within flow cytometry data. Variations of K means clustering and model-based methods are the most commonly used. While they have been shown to be able to identify cell populations more consistently than manual gating they still have seen little adoption in clinical and research labs. The populations discovered by these methods do not faithfully recapitulate the populations that a researcher aims to detect. Over decades of flow cytometry use both clinical and research scientists have developed specific flow cytometry staining panels and systematic hierarchical gating paths to discover cell populations and interpret the outcome of a diagnosis or experiment. When using computational methods, these populations, and therefore their population statistics are no longer comparable to historical data. Learning how to interpret this new information, correlating the results to past outcomes, and validating the experimental procedure is too much of an undertaking for many research labs and would impose even further restrictions on clinical work. Because of this much of flow cytometry gating is still done manually despite it presenting a bottleneck in analysis and the potential of added variability.

This histogram matched support vector machine method aims to use expertly gated data to create support vector machines that can gate flow cytometry data and produce the same population statistics as manual gating. Using this method, the data generated can be directly interpreted to a clinical or experimental outcome just as it has been for decades. As a practical method to test the histogram matched support vector machine we use a flow cytometer's light scatter properties (FSC, SSC) to extract live and dead cell counts from culture samples. This will not only fulfill the goal of testing the accuracy of population gating using a histogram matched support vector machine, but also will automate another manual, tedious and error-prone process, microscopic cell counting.

Introduction

Cell counts for viability is a fundamental measurement made in many biological experiments. Its accuracy is imperative when correlating live cell numbers to parameters of a biological function. Despite the many automated cell counting methods available today, manual counting using a hemocytometer, a specialized microscope slide, is still the most commonly used method [2, 57]. A cell culture suspension is injected into space between the slide and coverslip and relies on the analyst's ability to evaluate a cell's attributes, usually in the presence of a stain such as Trypan Blue. Trypan Blue exclusion method is based on the principle that an intact cell membrane surrounding a living cell is capable of excluding the dye while a compromised cell membrane of a non-viable cell allows the dye to enter causing a blue appearance. This type of manual evaluation is time-consuming which prohibits a large number of samples to be analyzed at one time. Manual counting methods have also been shown to be subject to inter and intra-user variation of 15 and 35% respectively. Figure 3.1 shows the variation of three trained researchers in our laboratory counting three replicates of the same four samples using Trypan Blue.

Many automated cell counting instruments are commercially available. Examples of such systems are the Luna (Logos BioSystems, Annandale, VA), the Cellometer (Nexcelom Bioscience, Lawrence, MA), and the Celigo (Nexcelom Bioscience). In general, these automated instruments consist of a camera and image analysis software to detect and count viable cells. The Luna and Cellometer use Trypan Blue on a proprietary slide requiring the user to focus a camera before imaging the slide to evaluate cell viability. The Celigo also uses imaging as a means for detecting live/dead cells but has the additional capability of using fluorescent viability dye to aid in their detection on a 96 well plate. These automated methods work very well for cell cultures that display uniform morphology such as established cell lines, but when used to count heterogeneous longterm primary *in vitro* cultures they have difficulty generating consistent and accurate counts. Histogram

matched support vector machine method can count viable cells without the use of viability dye with a correlation coefficient to Trypan Blue of over 0.90 as well as increasing the counting consistency.

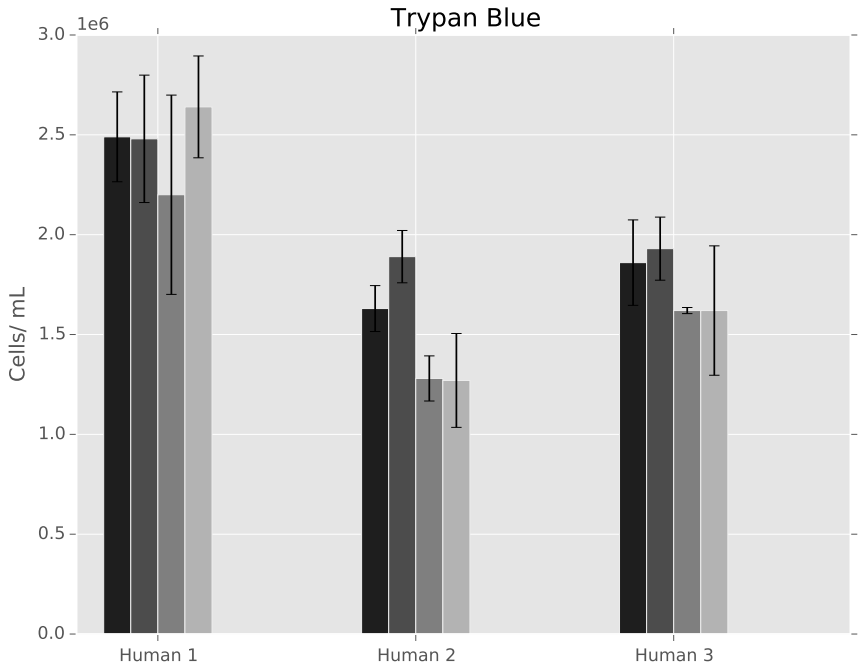


Figure 3.1: Three trained laboratory analysis counted four separate long-term *in vitro* human cell cultures in triplicate. The inter-user variation was approximately 15%, and the intra-user variation was 35%.

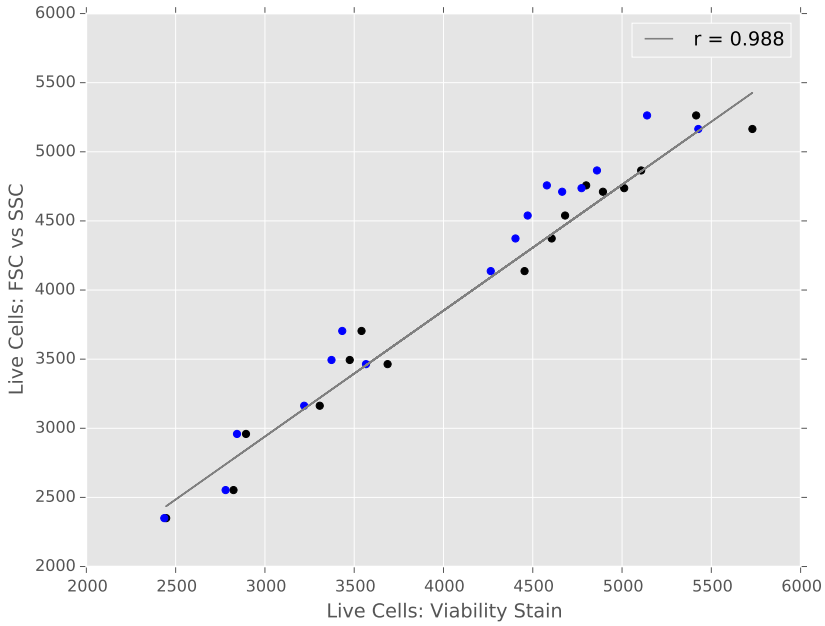
Scatter Properties vs. Live Dead Stain

Although it is established that dead and apoptotic cells decrease in forward scatter and increase in side scatter compared to viable cells we wanted to verify that these changes in light scattering properties of the cell could discriminate between live and dead populations with enough accuracy for our purpose. To test this, we stained 12 day old *in vitro* lymphocyte cell cultures with a

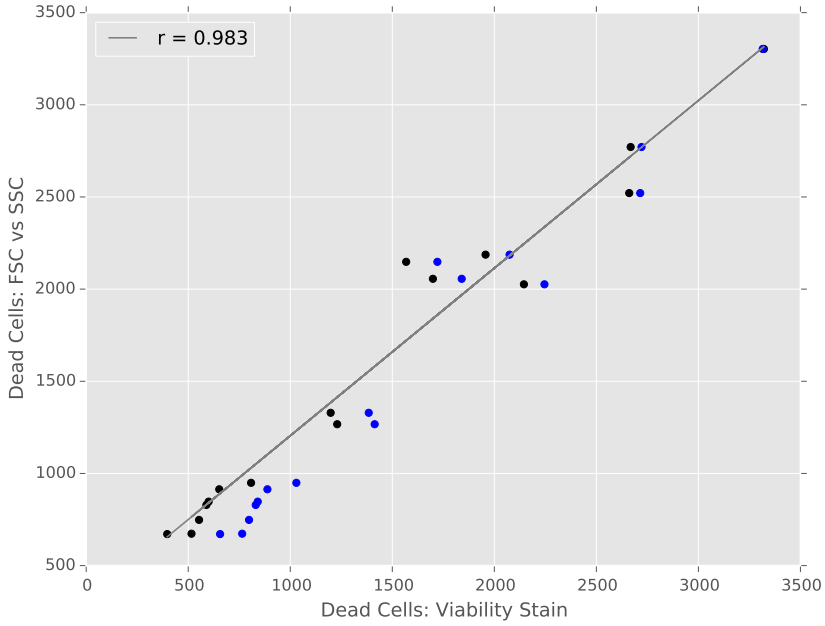
membrane permeability dead cell apoptosis kit containing PO-PRO and 7-Aminoactinomycin D (7-AAD) (Invitrogen, Carlsbad, CA). These reagents stain apoptotic and dead cells respectively. Sixteen longterm *in vitro* primary cell cultures were stained in duplicate and acquired on a BD Frortessa. Using Flow-Jo analysis software live and dead cells were manually separated from each other on the premise of being stained with both Po-Pro and 7-AAD. A separate manual analysis on the same FCS files, gated live and dead cells based on expected dead/apoptotic FCS and SSC patterns without using the viability stains. Cell counts for each were exported from FlowJo and compared. Data in figure 3.2 shows for both live and dead cell population the correlation was greater than 0.98, demonstrating that a viability stain is not required for the accurate discrimination of live and dead cells. Once it was determined that a viability stain was not needed to exclude live from dead lymphocytes the next step was to overcome the fluctuations in scattering profiles due to different cell morphology brought on by different culture conditions, activation states, and donor to donor variation.

Training Data

For the initial training set, six different stimulation conditions were used to generate a variety of activation states within a human *in vitro* assay. These different activation states translated into a variety of forward and side scatter profiles. 150 FCS files containing forward and side scatter data were generated using 25 donors over six simulation conditions: no stimulation, CEF peptide, seasonal Flu vaccine, PHA/PMA, PWM, and IL-2 addition. Fifteen donor's results over the six stimulation conditions were gated for live, dead, and debris populations to be used as the initial training files for the SVM classifier. This gave the initial run of the algorithm 90 possible forward and side scatter profiles to select from when choosing the optimal FCS file for the creation of the classifier.



(a) Live correlation



(b) Dead correlation

Figure 3.2: Sixteen separate twelve day old human *in vitro* lymphocyte cultures were stained for viability and acquired on a flow cytometer in duplicate. The resulting FCS files were gated using the viability dye as well as using forward, and side scatter.

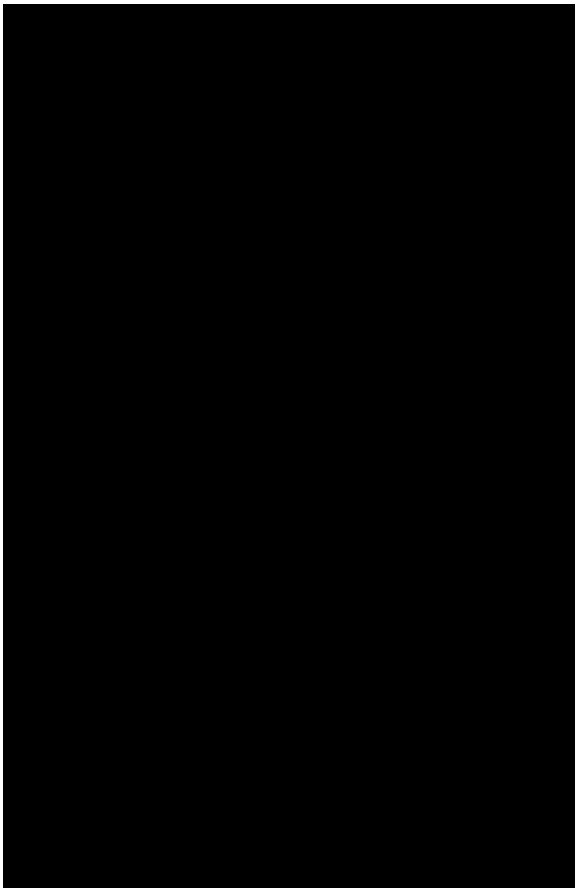


Figure 3.3: Process chart outlines the creation of the SVM vector list from a gated bank of FCS files.

Optimal Support Vector Machine Selection Using Histogram Matching

Due to routine analysis of datasets using the same gating patterns to separate cell populations by visual boundaries it seems probable that a SVM would be a good fit flow cytometry classification. However, SVM have not seen much utilization in flow cytometry analysis. One reason may be that populations size, shape, density, and location can vary significantly from sample to sample. Population variation is due to the phenotypic difference in the cells, variations in sample preparations, stimulation conditions, and other laboratory differences. If a SVM was given a set of training data

to generalize all possible FCS files the chances are that it would not be a proper fit to gate every file presented. To overcome this, a simple image matching technique is used to select the most similar SVM (previously gated FCS training file) to the current sampling being counted.

Selecting Number of Bins

Histograms are an intuitive non-parametric density estimator. While simple, the estimation of a sample's probability density is highly dependent on the choice of the number of bins and in the case of uniform bin-width histograms the bin width. The width of the bins must be sufficiently small to capture all major features of the data but also be large enough to ignore the small fluctuations in the data. There is no best answer to the number of bins to use to generate a histogram. Depending on the data distribution a variety of bin widths may be appropriate to accurately create an estimation of the probability distribution, for this work when a histogram is needed as an estimation of the probability distribution the Freedman-Diaconis rule shown in equation 3.1 is used to determine the bin width and therefore the number of bins. Where IQR is the interquartile range of the dataset x, and n is the number of data points in the set.

$$h = 2\frac{IQR(x)}{n^{1/3}} \tag{3.1}$$

Creating the SVM Vector List

For each FCS file in the training set, histograms descriptive of the forward and side scatter distributions are created. The proportion of events within each bin of these histograms is saved as a vector. An overview of this step can be seen in figure 3.3. This procedure is performed for the setup of the initial, and subsequently each time a new FCS file is added to the bank of files that can

be used to create a SVM.

Histogram Matching and Color Indexing

Color indexing is a computer vision technique developed in 1991 by Swain and Ballard for quick visual skills to allow robots to react in real time to their environment [58]. It was shown that using color histograms as a representation of an image was an efficient and accurate method of image matching. Here a method based on this process is used to select the optimal SVM classifier for the discrimination of live from dead cells in long-term cell cultures. With traditional color indexing, a color space is a specific color axis; red, green, or blue for example. A color histogram is created by breaking one color axis into a set of discrete bins and counting the number of pixels that fall into each of these distinct bins. Discretizing the FCS and SSC values from an absolute minimum and maximum, a histogram analogous to that of one created for a color axis can be created. To identify the scatter 'image' that most closely matches the scatter plot from the FCS file to be counted, a method to analyze this similarity called histogram intersection is used. This intersection value is the number of events from the model (reference file) that has corresponding events in the same scatter axis as the sample file. Given a pair of histograms representing FSC, for example, each containing n bins the intersection of the histograms is: $\sum_{j=1}^{n} min(I_j, M_j)$ [58, 59]. Because all flow files could and probably do have different event numbers and distribution statistics to compute the percentage match the intersection value is normalized by the number of events in the model histogram (equation 3.2), resulting in the final similarity or intersection score.

$$H(I, M) = \frac{\sum_{j=1}^{n} min(I_j, M_j)}{\sum_{j=1}^{n} M_j} \tag{3.2}$$

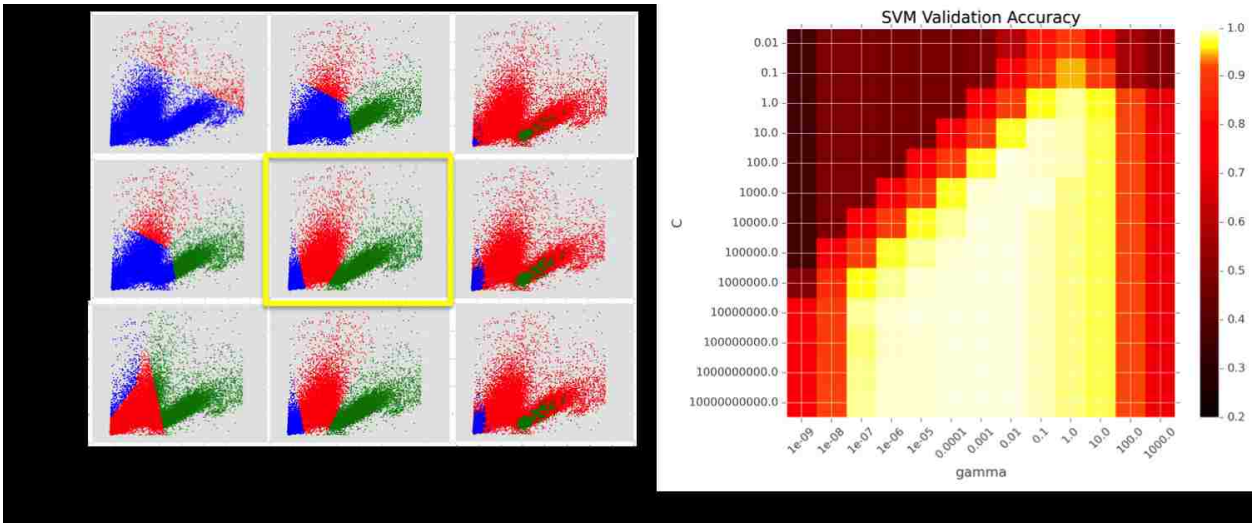


Figure 3.4: The dot plots show live (green), dead (red), and debris (blue) gates for one FCS SVM file using varying γ and C parameters. The heat map displays average classification accuracy over the 90 FCS file in the original training set over exponentially spaced γ and C parameters.

SVM Creation

Using scikit-learn, a machine learning Python package, we trained a non-linear SVM using a radial basis function (RBF) Kernel, $\exp(-\gamma||x-x^{'}||^2)$ [15,43]. To determine the most appropriate values for γ of the RBF and C, a parameter used in all SVM kernels to justify a smooth decision boundary over misclassification to avoid overtraining, we conducted a cross-validation study of exponentially spaced parameter values. Figure 3.4 shows the results of this study. Using the average best parameters over the 90 FCS training files a γ value of 0.001 and a C of $10,000$ was used for all subsequent classifications.

Histogram Mismatch Threshold

To assess the gating accuracy and to determine the threshold of similarity where the SVM classifier differs enough from the data to be classified that an accurate count is no longer possible we ran the remaining ten donors FCS files using the best, worst, and three randomly selected classifiers on each sample. 300 images depicting the classification (live, dead, debris) were generated and analyzed to determine if the populations classified by the SVM were accurate or not. Figure 3.5 shows what was classified as a correct or incorrect gate as well as the distribution of similarity scores. It was discovered that any score above a 0.63 similarity always resulting in correct gating. This became the minimum threshold required to constitute a file match.

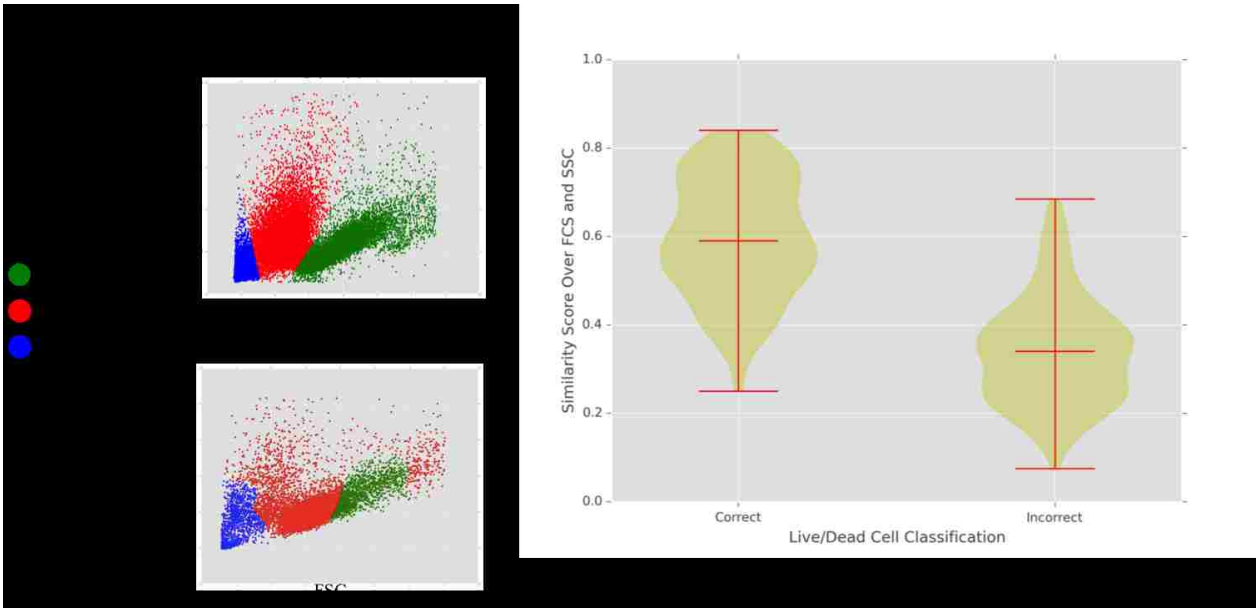


Figure 3.5: Ten donors FCS files were gated for live, dead and debris using the best, worst, and three randomly selected classifiers. From the 300 FSC vs. SSC plots generated the gating was classified as correct or incorrect if all three populations were captured correctly. A similarity score of 0.63 or higher was associated with always gating a sample correctly.

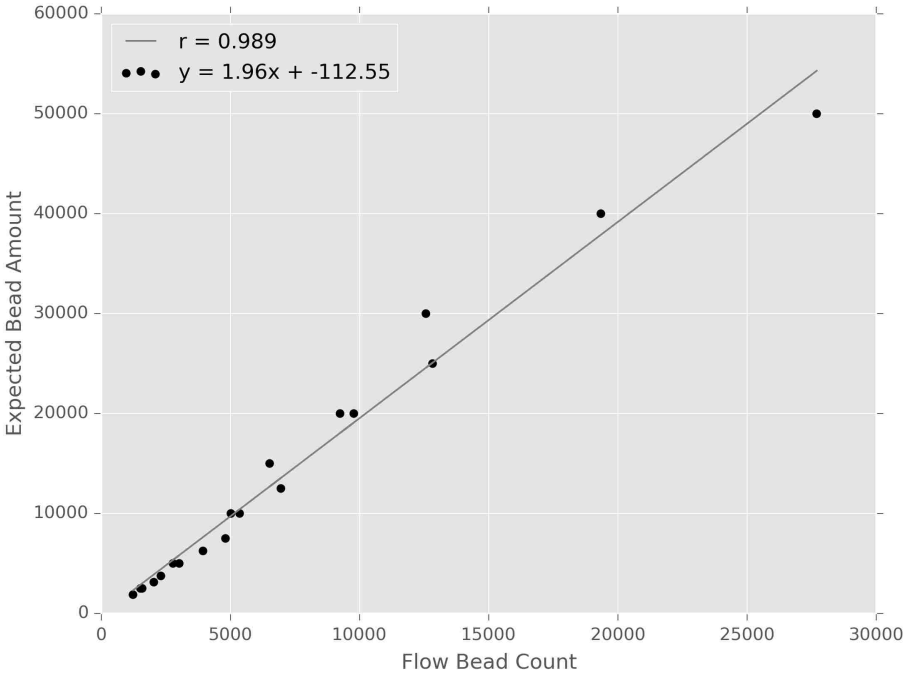


Figure 3.6: A dilution series of calibrated bead samples were run on the cytometer. The resulting FCS files were analyzed for bead counts to calculate the precise volume in μL analyzed the cytometer.

Bead Calibration

The high throughput system (HTS) on some flow cytometers allows the user to enter acquisition volume (μL) and speed (μL/second) that will be used to run the sample through the cytometer's fluidics and laser systems. Typically this setting is used to ensure a slow enough flow rate for cells to pass one by one through the cytometer's fluidics system. In this application, however, this flow rate must be finely tuned to produce an accurate cell count. To determine the exact volume in μL that passes through the cytometer's fluidics and detector systems during sample acquisition we used AccuBeads (Hamilton Thorne Inc, Beverly, MA) which have been verified by the manufacturer to contain $2.5X10^6$ beads / mL. Three separate $1:2$ serial dilutions were performed in triplicate

starting at $2.5X10^6$, $2.0X10^6$, and $1.5X10^6$ beads/ml over four dilutions each. This corresponded to a bead range per mL of $2.5X10^6$ to $93,750$. The HTS was set to acquired 20μL of each sample at a flow rate of 2.5μL/sec. After the samples were run on the cytometer, the beads were manually gated in FlowJo and their counts exported. Figure 3.6 shows the beads gated fro the FCS files and the known bead amount are strongly correlated with a r value greater than 0.98. A dilution series similar to this was used with every counting run on the cytometer to ensure accuracy, with the fitted linear function serving to calculate exact cells per mL.

Results

This cell counting method was tested using 11 long-term *in vitro* human cell cultures. Each sample was counted in triplicate by a laboratory analysts and the histogram matching SVM. The average counts calculated by each method are plotted against each other in figure 3.7. The counts were strongly correlated with a r value of 0.931.

Six separate long-term *in vitro* cell cultures were analyzed in triplicate using Trypan Blue, Luna, Cellometer, or the Celigo with and without viability stain. The average coefficient of variation was calculated for all six sampled over all the counting methods. The flow histogram matching SVM had the least variation over the six samples tested.

Discussion

As new FCS files are counted using this method it is possible that a file will not find a match within the specified similarity threshold. If this occurs, the user can be prompted that the discovered forward and side scatter profile of this sample is unique enough to warrant using its manually gated file to create another SVM. Expanding the database in this fashion can be done over time as

new cell types or new culture conditions cause a change in the forward and side scatter parameters that have been previously seen.

While what presented here is an analysis in two dimensions (FSC and SSC), this application can be applied to any flow cytometry analysis where expertly gated training data is available. Future work will use this same process over a hierarchical gating scheme to identify a greater diversity of cell populations over many more dimensions. Preliminary work has been done to identify thirteen populations over eight dimensions with results similar to manual gating.

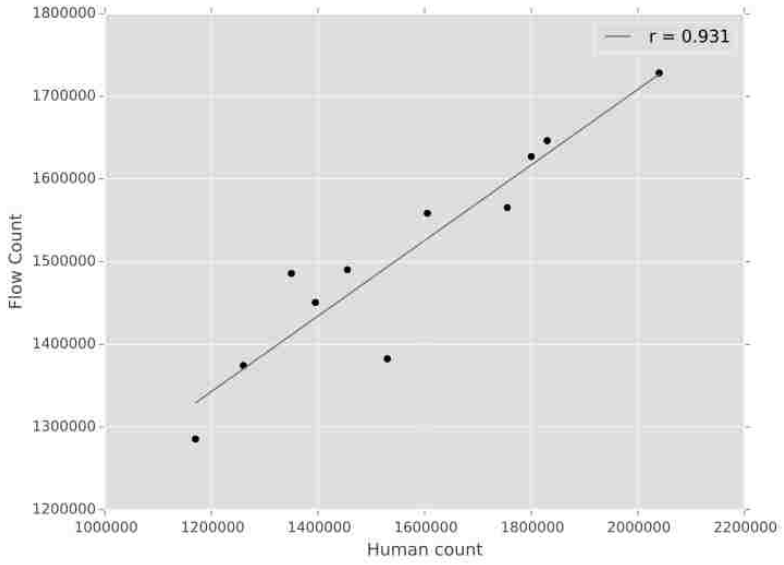


Figure 3.7: Counts from 11 *in vitro* cell cultures were analyzed for viable cells per mL using the histogram matching support vector machine method. These cell counts are compared to an average of three analysis using the Trypan Blue exclusion method. A strong correlation between to the methods was seen.

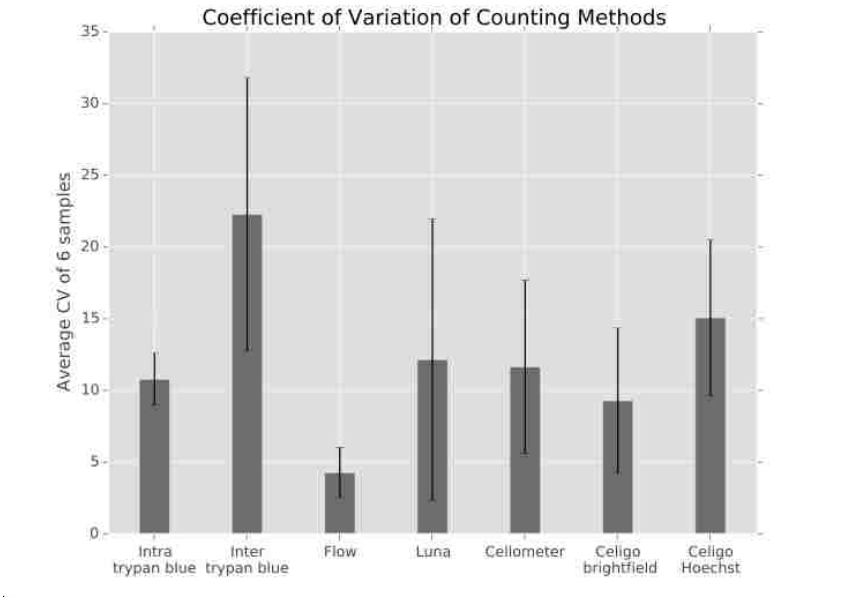


Figure 3.8: Histogram matching SVM flow cytometry method had the lowest variation among all the cell counting methods tested.

CHAPTER 4: EXPLORATORY ANALYSIS: DIVISIVE CLUSTER DISCOVERY AND VISUALIZATION OF ADAPTIVE ANTIGEN SPECIFIC T HELPER CELL RESPONSE

This method contains four modules. First, a hypergraph sampling method is used to reduce the number of data points in a FCS file while preserving rare and possibly essential events. Using random sampling as opposed to a method such as this could result in the loss of this subset of the data that may be essential in the determining the outcome of the diagnosis/experiment. Next, singular value decomposition (SVD) is used to deconvolute the subset of FCS data from the first module. The three matrices produced by SVD, (U,Σ,V^T) are used as a mechanism to split the data into related clusters divisively. This clustering method uses the direction of the most prominent leading vector of the U matrix to group the sampled FCS data into meaningful clusters. This divisive clustering step is halted after reaching a designated threshold based on the amount of information the user requests to capture based on the values contained in the Σ matrix. Due to hypergraph sampling, rare events are not overpowered by the abundant events within the dataset, and therefore can be distinguished into their own populations. The third step uses the sampled subset of data along with the newly defined clusters to create, train, and test a support vector machine (SVM). Using this SVM, the events that were not selected to be used in the second and third computational steps are classified into one of the discovered clusters. The final fourth module uses the median value of each parameter for each discovered cell clusters to form a minimum spanning tree. Finally, an XML file for trees visualization in the open source complex network analysis tool Cytoscape is created. Using Cytoscape the nodes of the network can be annotated and colored to dynamically reflect a markers relative intensity. Within this network, one can visualize antigen-specific polyfunctional $CD4^+$ T helper cells generated by an *in vitro* human immune culture system. Figure 4.1 gives a general overview of the computational modules of this method.

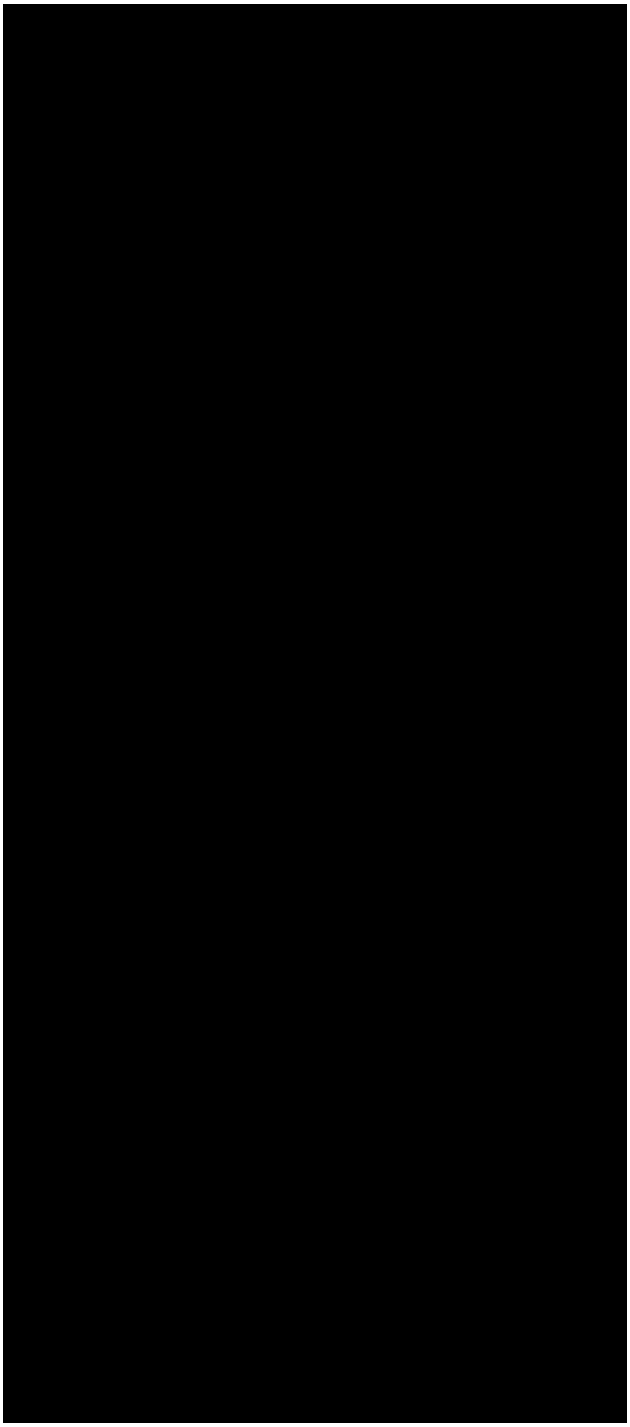


Figure 4.1: Overview of the computational modules for this unsupervised exploratory network analysis.

In Vitro Generation of Antigen-Specific Responses

Donor PBMC Isolation

Donor peripheral blood mononuclear cells (PBMC) were collected from healthy donors using leukapheresis. PBMCs were isolated from this enriched apheresis product using a Ficoll-Paque PLUS(GE Healthcare Bio-Sciences, Piscataway, NJ) density gradient [8, 33]. The interface of PBMCs was removed, washed, and cryopreserved in IMDM media (Lonza, Walkersville, MD) containing DMSO (Sigma-Aldrich, St. Louis, MO), and autologous serum.

Cytokine-Derived Dendritic Cells

Dendritic cells (DC) were prepared as previously described [33, 52]. Briefly, monocytes are purified from cryopreserved PBMCs by positive CD14 magnetic bead separation (Miltenyi Biotec, Auburn, CA). The separated $CD14^+$ cells were cultured in X-VIVO 15 media (Lonza) for six days in the presence of 100ng/mL GM-CSF (R&D Systems, Minneapolis, MN) and 25ng/mL IL-4 (R&D Systems). After Incubation the DCs were either left untreated (no antigen control), pulsed with a $1 : 250$ dilution of the commercially available Yellow Fever Vaccine YF-VAX®(sanofi pasteur), pulsed with 1μg/mL of killed *Francisella tularensis* SCHU4, or invected with the live attenuated investigational Live Vaccine Strain (LVS) of *Francisella tularensis* at a $1 : 10$ bacteria to DC ratio. DCs were left to incubate with antigen overnight.

$CD4^+$ T Cell Stimulation

Autologous $CD4^+$ T cells were purified from cryopreserved PBMCs by negative magnetic bead selection (Miltenyi Biotec). The isolated $CD4^+$ T cells were then cultured with either untreated,

YF pulsed, SCHU4 pulsed, or LVS invected DCs at a ratio of $60:1$, $CD4^+$ T cells to DCs. These co-cultures were left to incubate for 12 days at $37°C$ and 5% CO_2 in X-VIVO 15 media (Lonza). Following incubation, the cultures were harvested and evaluated for effector T cell activity using an intracellular cytokine staining assay (ICCS). This method uses autologous DCs prepared as previously described to restimulate any effector T cells that were generated during the 12-day co-culture. The T cells and target DCs were cultured for 6 hours in the presence of 1μg/mL of brefeldin A (Sigma-Aldrich) to inhibit protein transport in the Golgi apparatus and cause an accumulation of intracellular proteins. After incubation, cells were labeled with a Live/Dead Fixable Stain (Invitrogen, Carlsbad, CA), permeabilized with cytofix/cytoperm and permwash (BD Biosciences, San Jose, CA) and labeled with antibodies specific for human $CD4$ (SK3), $CD154$ (TRAP1), IFNγ (B27), TNFα (MAb11), IL-2 (MQ1-17H12), IL-5 (TRFK5), and IL-$17A$ (N49-653) (eBioscience, San Diego, CA). The samples were then acquired on a BD LSRFortessa (BD Biosciences).

Flow Cytometry Data Preparation

In preparation for computational analysis, the FCS files were pre-processed as described in 2. The files were also gated using the analysis software FlowJo (TreeStar, CA) to remove doublets, dead cells, and debris. This gating removes artifacts that due to non-specific binding can interfere with further analysis.

Hypergraph Sampling

The first module of this unsupervised exploratory analysis samples a flow dataset in a manner to preserve rare events over all dimensions/parameters. Flow cytometry files can contain anywhere from thousands to millions of events, each being individually described by three to 20 parameters.

With datasets of this size and dimensionality, computation time and memory requirements for analysis can quickly become prohibitive, even when using a method of modest computational complexity. For example, for a one million event file, a distance matrix describing the relationship of every data point to one another using four bytes per entry would approach four terabytes of memory using a conventional adjacency matrix. While a sparse matrix or linked list could be used to represent this data structure and limit the memory used, such a representation would still be impractical to be used on a typical desktop computer. To overcome this memory constraint a representative sample of events that ensures the preservation of rare events is necessary.

Due to the nature of flow cytometry data, it is not in the best interest of the analysis to merely take a random sample. In many experiments, the reason the cell sample is of interest to be analyzed using flow cytometry is to identify and phenotype a rare subset of the data. This subset, although small, represents a very informative set in many experiments. These rare events in our hands and in published experimental results are often seen at a rate of one in a $1,000$ to one in $10,000$ events. If the rare events are present in the sample at a rate of one to $10,000$ within a FCS file of $1,000,000$ events, the chances of gathering all 100 events that compose this population are improbable with random sampling. To generate a subset that contains rare events as well as those in abundance, the density of events surrounding each cell in the d dimensional space is needed. This, however, this distance calculation is not practical. To alleviate this need in computational power and memory a hypergraph sampling method was developed. The graph constructed here is in the context of graph theory, the study of mathematical structures that are used to model pairwise relationships between objects, or in our case flow cytometry events. A graph in this context is made up of vertices or nodes (V) and connected by edges (E). A hypergraph is then a generalization of this type of graph where a single edge can connect any number of vertices.

Hypergraph Creation

The first step in the creation of the sampling hypergraph is to place a limit on the maximum number of edges that can be used to represent the flow dataset. The hyperedges of the graph represent a portion of the data's probability distribution at a specific area of the multidimensional space. Once the graph is created the weights of its edges, or the number of flow cytometry events contained in each edge is directly used to extrapolate the events to use in the sampled dataset. If a relevant maximum edge number is not selected it could be possible for every event within the file to be contained by a unique edge. However will not help in distinguishing the rare from abundant events. On the other extreme, all events in the file could be connected using a single edge. This leads to each event being viewed to fall within the same area of multidimensional flow data space and will lead us to essentially taking a random sample of the dataset. To ensure that neither of these two extremes occur we limit the possible number of edges that can be created during the hypergraph sampling based on the average number of events that need to be seen to detect a rare event. With the dataset explored in this document (polyfunctional $CD4^+$ T cells) typically a ratio of $1,000$ common events to one rare ($CD4^+CD154^+Cytokine^+$) is seen. While $1,000 : 1$ ratio is quite common in the literature rare event rates of $10,000 : 1$ to even $1,000,000 : 1$ have been reported and this parameter should be tuned for the estimated rarity of the cell population(s) of interest. For the purpose of our analysis however the rare event ratio of $1,000 : 1$ will be used as it best describes our dataset.

The rare event ratio is used to calculate the position and volume within the multidimensional flow space a hyperedge occupies. If r is the number of events needed to statistically have a chance of detecting a rare event and d is the number of parameters (markers) describing the events of the FCS file we calculate b, the number of discrete value ranges per parameter using equation 4.1. In the dataset presented seven parameters ($CD4$, $CD154$, IFNγ, TNFα, IL-2, IL-5, and IL-17) are

used to create the hypergraph. These parameters with the $1000:1$ rare event rate requires three discrete value ranges per parameter. If the number of parameters is high in relation to proportion of rare events, for example 11 parameters measured on a $100,000:1$ rare event estimate, limiting the number of value ranges per parameter at less than 3 then a minimum constant of 3 data ranges is used per dimension. Even with data sets of lower dimensions (5 or fewer) 3 bins has been tested to yield expectable sampling results that translate into meaningful populations at the conclusion of the exploratory analysis.

$$b = \lceil (\sqrt[d]{r}) \rceil \tag{4.1}$$

The boundaries of each discrete value range are calculated independently for each parameter just as in the creation of a simple one dimensional histogram. Each event is processed on each dimension to determine what value range for each dimension best describes it. An events combination of data value ranges corresponds to the hyperedge that will contain it. Events are added to the hypergraph either by creating a new hyperedge if one has not yet been generated for the events unique combination of ranges or it is placed into an existing hyperedge. Once all events within the FCS file are placed within their descriptive hyperedge the dataset can now be sampled so the rare events are preserved.

Event Sampling

Fortunately, flow cytometry data is not uniformly distributed within its data space. This is the basis of why the most simplistic form of flow analysis, sequential manual gating, is so popular and what we leverage when using hypergraph sampling. Once all needed hyperedges are created, and all events have been placed within the edge most descriptive to its place within the data space,

a representative subset of the original data can be extracted. Edges with a low weight, or few events, contain rare events while edges with a higher weights contain events present in greater abundance within the dataset. A higher proportion of hyperedges contain very few events and a low proportion of hyperedges contain a large number of events. The hyperedges containing a low number of events represent areas of the multidimensional space where events are scarce. These events could be the rare cells of interest that despite their low frequency are crucial to analysis. In the example presented here these events maybe antigen specific polyfunctional effector $CD4^+$ T cells. On the other extreme, hyperedges containing a high number of events represent a very common cell type, in context of this analysis, $CD4^+$ T cells that are not responding to the given antigen. The number of events per hyperedge therefore generally follows a negative exponential distribution, an example of which can be seen in figure 4.2. In calculating what constitutes a rare event and at what threshold edge weight do all events within the hyperedge need to be sampled we fit the data to the exponential probability function, $\lambda \exp(-\lambda x)$,using least-squares fit. This method aims to minimize the sum of squared residuals, or the difference between an observed value and the fitted value given by the model. Once the exponential function is estimated we can use the computed value of λ with Tukey's criteria for anomalies to calculate at what weight do hyperedges begin to contain an anomalous number of events [32, 60]. Tukey's criteria for anomalies is a method commonly used in box plots where the interquartile range is used as a measure to describe the extent to which a distribution is spread. An outlier is considered a data point outside the IQR by one and a half times the IQR. Therefor we use equation 4.2 to determine the cut off where outliers, heavily weighted hyperedges begin or in the context of this flow cytometry data where common events are grouped together. Hyperedges at or under this calculated threshold, t, have all their events sampled. Hyperedges above this threshold must have the number of events they contribute the final subset reduced to this number. While random events could be sampled per edge this would add in a stochastic element to the algorithm leading to the possibility of generating alternate exploratory networks in the final analysis. To chose the same subset of events every this algorithm

is run a hyperedge over the event limit t is analyzed. Within the subset of events contained within it the most variable parameter is calculated and t events are sampled uniformly along this parameter.

$$t = \left\lceil \frac{\ln(4)}{\lambda} + 1.5\frac{\ln(4)}{\lambda} \right\rceil = \lceil Q3 + 1.5 \times IQR \rceil \tag{4.2}$$

Sampling Results

To test hypergraph sampling a simulated flow cytometry 'like' data set was initially explored. Using a synthetic dataset gave us control over the frequency of the parameters expressed on each event as well as eliminated biological variation of the human *in vitro* data. The simulated data set was generated to contain $300,000$ events in three dimensions; P1, P2, and P3. Each dimension or parameter varied in frequency and co-expression. This was used to mimic flow cytometry data where a rare event must often be found by several hierarchical gating steps. Events could be 'positive' or 'negative' for a particular parameter. Positive values were generated randomly from a Gaussian distribution with a mean of 0.8 and a standard deviation of 0.2. Negative values were represented by values with a mean of 0 and a standard deviation of 0.2. Parameter P1 was used to simulate a common cell type. Fifty percent of the events were assigned to belong to a population positive for P1. Of the events selected to represent a negative P1 population 20% of these events at random were selected to be positive for P2. Finally 1% of those events who were both P1 negative and P2 positive were randomly selected to be P3 positive. Figure 4.3 shows the results of random sampling of this data versus hyperedge sampling. Hyperedge sampling, over the synthetic dataset faithfully retained the rare events. As shown by P2 vs P3 and P1 vs P3 dot plots. The small P3$^+$ population represents 0.1% of the dataset. This rare population is well preserved with hyperedge sampling while it is nearly completely lost when sampling the same number of random events. In addition to preserving this obvious rare population the events along the boundaries of

all populations are also preserved. In a subsequent analysis step a SVM is used to classify the cells not used in this sampling step into discovered clusters. Due to SVMs using the boundary events to create optimal decision boundaries between populations this feature of the sampling may also help with classification accuracy.

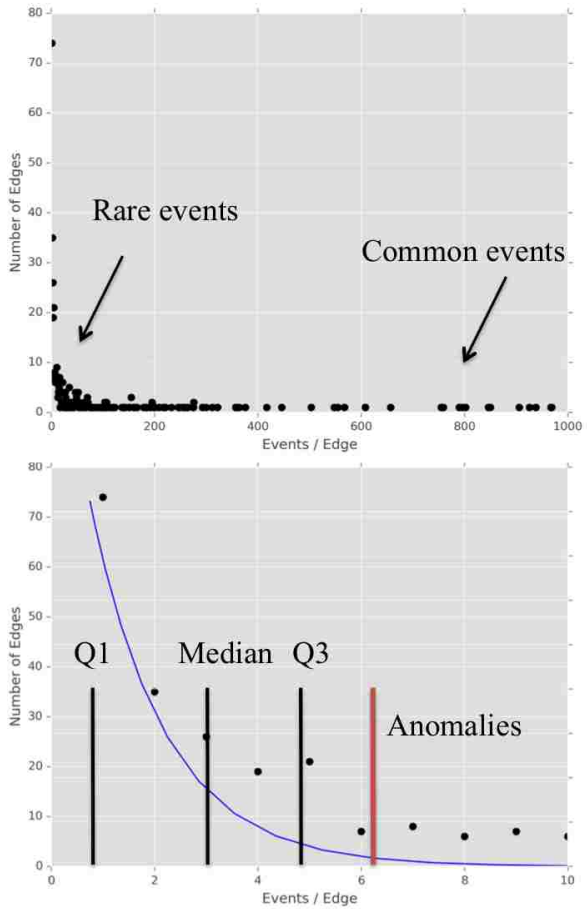


Figure 4.2: This is the edge weight topology of the hypergraph created during the analysis of LVS stimulated *in vitro* cell culture. The first graph shows all edges within the hypergraph. The second shows the edges containing rare events. In this analysis t from equation 4.2 equals seven. Meaning seven events are sampled at maximum from each edge of the hypergraph. Figure 4.4 shows the results of this sampling.

This same sampling method was then used with the human *in vitro* culture samples. This FCS data

was sampled over seven parameters: CD4, CD154, IFNγ, TNFα, IL-2, IL-5, and IL-17. Figure 4.4 shows the results of sampling the LVS culture's FCS file. The data shown is for CD154 and TNFα, typically a cytokine while still rare is produced in higher amounts, and IL-17 the least abundant cytokine produced in this dataset. While populations within flow cytometry data are known to deviate from a Gaussian distribution this method was still able to preserve these important effector T cells while reducing the amount of data by over 99%.

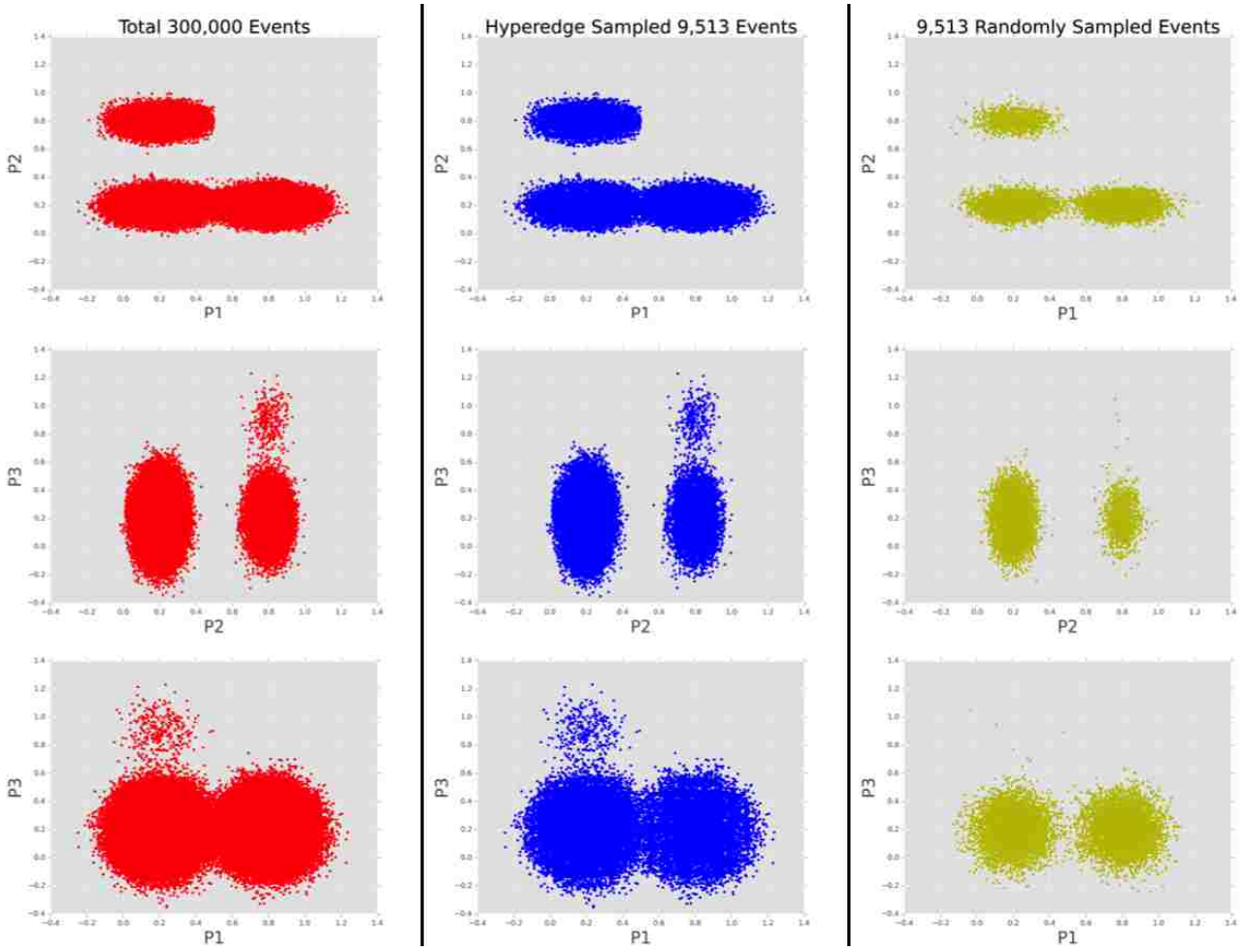


Figure 4.3: Example of sampling a simulated flow data set of $300,000$ events over three dimensions. The sampled data is approximately 3% of the original dataset while preserving the rare events that make up 0.1% of the data.

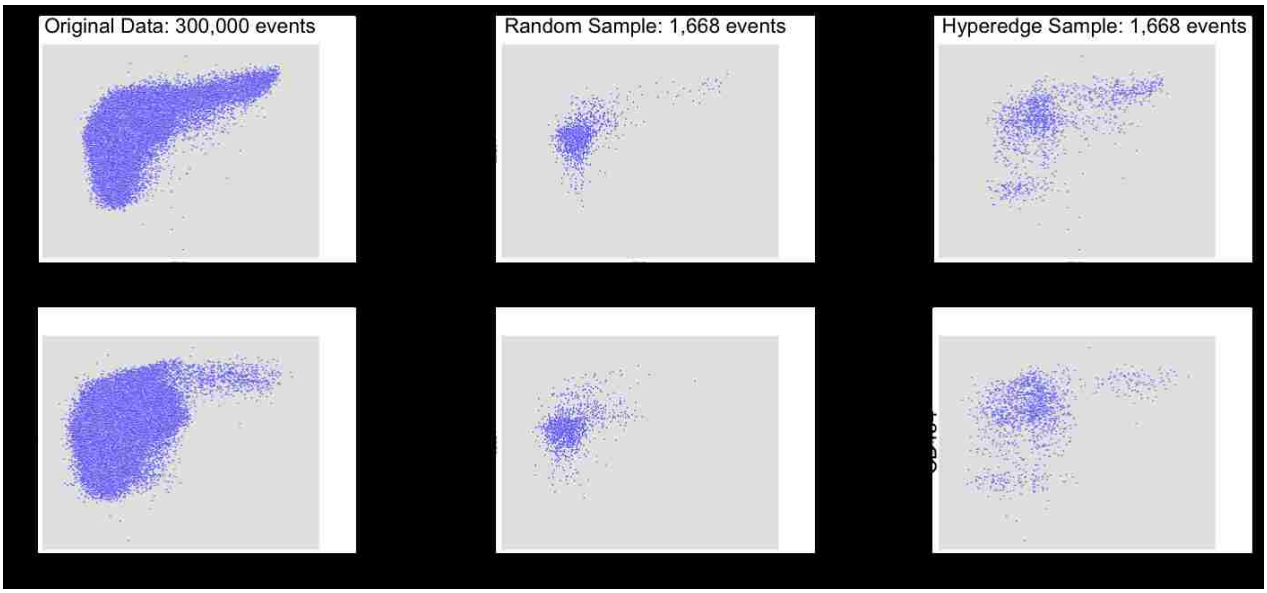


Figure 4.4: Comparing the original dataset over two parameters that illustrate the rare/critical events to this CD4 T cell analysis. Using random sampling we greatly reduce the $CD154^{+}TNF\alpha^{+}$ population and nearly eliminate the $CD154^{+}IL\text{-}17^{+}$ population. Using a hypergraph to sample the same number of events both of these populations remain present in the sampled data.

SVD Clustering

The second computational step uses singular value decomposition as an objective clustering algorithm to group related data points together without the use of a user-defined cluster number. Using SVD the sampled FCS data is divisively split to uncover cell populations in a top-down fashion. This method allows for the discovery of clusters by defining by variation in the data as opposed to being explicitly stated by the user.

Singular value decomposition (SVD) is a widely used linear algebra technique that takes a high dimensional and variable set of vectors and reduces it to a lower dimensional space where the structure of the original dat is preserved. SVD has been used in many fields such as neuroimaging,

complex systems, text reconstruction, sensory analysis, image compression, and genetics. SVD is the core to many statistical techniques including PCA , correspondence analysis, multidimensional scaling, and partial least squares. This technique can be used to achieve three main objectives:

1. Transform correlated variables into a set of uncorrelated ones that better expose the various relationships of the original data.
2. Order dimensions by increasing variation in the data set.
3. Finding the best approximation of the original data points while using fewer dimensions.

Every n cell/event within the cytometry file is represented by a vector of real numbers of length m, the number of parameters. This vector is comprised of the standardized fluorescence values from each parameter. The equation for SVD is the following $M = U\Sigma V^T$, where U is a $n \times m$ matrix, Σ is a $n \times m$ diagonal matrix, and V^T is a $m \times m$ matrix. The columns of U are the left singular vectors, uk, and form an orthonormal basis for the marker expression profiles of the events. The rows of V^T contain the elements of the right singular vectors, v_k, and form an orthonormal basis for the parameters. The elements of Σ are only non-zero on the diagonal, and are called the singular values, $\Sigma = diag(\sigma_1, \ldots \sigma_m)$. The ordering of the singular vectors is determined by high to low sorting of singular values, with the highest value in the upper left index of the Σ matrix. Left singular vectors, the columns of U, can be referred to as the eigen-Markers and the right singular vectors as the eigen-Events. For this application we are interested in the expression of the marker profiles of the sample to understand the relations among the cells.

Division Number

By setting the error $e(r)$ in equation 4.3, to the desired similarity and solving for k, the number of singular values along the diagonal of matrix Σ that allows you to reach the needed similarity,

cell populations can be found. This number of σ values, k, then represents the number of divisive iterative splits of matrix M that are needed. These iterative splits are done using the direction of the left singular vectors of U. Each column of the U matrix, u_k, contains n data points corresponding to an event from the original matrix M. The direction ($+$ or $-$) of $u_k 1$ to $u_k n$ directs what population that event will be in after iterative spilt k.

Depending on the desired similarity requested by the user k may correspond to using all principal components of the dataset. However, typically the top principal components are able to explain a large proportion of the original matrix data structure. As a visual example of how this method is applied to lossy image compression see figure 4.5. By using only 40 of the $1,100$ vectors of the original dataset the compressed image still retains 85.5% of the original data. For our example *in vitro* dataset, clustering was done using a 0.90 value for $e(r)$. In other words requesting that at least 90% of the original data was preserved in the final network. This led to using the top 6 singular values or $6 = k$ divisive splits.

$$e(r) = 1 - \sqrt{\sum_{1}^{k} \sigma_i^2 / \sum_{1}^{n} \sigma_i^2} \tag{4.3}$$

Classification

The third step of this exploratory analysis classifies the remaining data that was not sampled into one of the cell populations discovered during the preceding clustering step. Classification using a Support Vector Machine (SVM) produced a classification accuracy of over 95% on average during testing. This method uses the LIBSVM implementation of a non-linear SVM using the RBF kernel [15, 43].

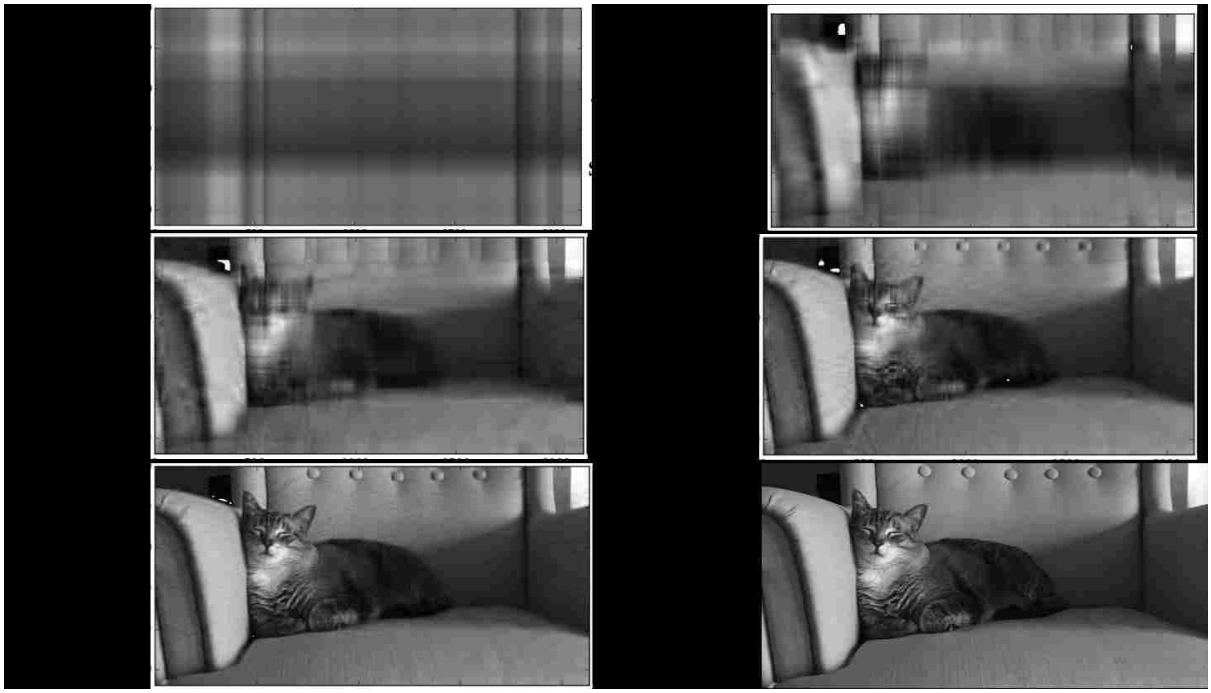


Figure 4.5: MeMurton demonstrates the use of SVD in the context of image compression. If every row of pixels is considered a vector the original image contains $1,100$ vectors. Using only the top vectors an image capturing 85.5% of the information contained in the original image can be constructed using only the top 40 vectors. This same idea is used when clustering flow cytometry data.

Other visualization methods either do not use the complete dataset such as viSNE where between $6,000$ to a maximum of $30,000$ uniformly sampled events are used for analysis and visualization [5]. Or in other methods the remaining data is reintroduced to in to the final visualization using K nearest neighbor algorithm, assigning an event of unknown classification to the cluster that its nearest neighbor belongs to [47]. During development of this algorithm we discovered first hand that simple distance measures that serve us well in three-dimensional space, such as Euclidean distance, often yield sub-optimal results when used in higher dimensional space. This is not a new finding. The expression 'Curse of dimensionality' was coined by Richard E. Bellman in 1961 to refer to the fact that many algorithms that worked well in low dimensions become in-

tractable when the input is high-dimensional [10]. This is because when dimensionality increases the volume of the space that the data points exist in increases so fast (exponentially) that the data points that inhabit this space become very sparse [9, 10]. This 'curse' may be what seems to limit the classification accuracy of K nearest neighbor algorithm on our dataset. To increase classification accuracy a non-linear SVM to classify the subset of data not sampled. Table 4.1 shows the classification accuracy of SVM classification compared to K nearest neighbor classification. The SVD clustered data was split with 80% used for training the the remaining 20% used for testing for all the methods shown. The accuracy shown is the average of 5 of these testing and training experiments using the LVS *in vitro* data. SVM classification outperformed K nearest neighbor classification by over 20%.

Table 4.1: Classification accuracy as a proportion of the testing set classified as correct according to where they were grouped using SVD clustering method. SVM using radial basis function with C of and gamma of compared with commonly used K nearest neighbor of two, five, seven, and ten nearest neighbors.

Clustering Accuracy		
Method	Mean	Std deviation
SVM	0.953	0.0056
K nearest 2	0.707	0.0118
K nearest 5	0.745	0.011642
K nearest 7	0.754	0.1203
K nearest 10	0.750	0.012505

Visualization

Once all events are assigned to their corresponding clusters the median of each parameter for each cluster is calculated. This measurement is used to calculate a distance matrix. A distance matrix is a squared matrix $N \times N$ in size, where N is the number of elements in the set, in this case the discovered clusters. Each cluster of cells represents a node of the minimum spanning tree with

each cluster represented by the median of mean value for each cell marker, therefore each cluster is represented by a m length vector, where m is the number of parameters per file. The distance to each cluster is calculated and a fully connected network with edge weights corresponding to the Euclidean distance is used for edge weights. A minimum spanning tree is then constructed using NetworkX's implementation of Kruskal's algorithm [22]. The network is then exported using NetworkX to an XML file that open source complex network tool Cytoscape can read and display [22, 54]. The minimum spanning trees are rendered with arbitrary node angles and edge lengths; therefore the positions of the nodes does not impact the meaning of the network itself. The node size however is representative to the proportion of events found to fall within that cluster from the cytometry file, and the color or the nodes is normalized for marker intensity. Nodes are colored according to the magnitude of the difference in their median responses relative the files being compared. This effectively eliminates the subjectivity of manual classification and improves the resolution of the heat map. When comparing multiple flow files as in the case in the figure. The fluorescence intensities are normalized over all the files in order for the color to represent the overall intensity of each marker across all files. Networks created to by this method to describe the *in vivo* cell cultures are seen in figures 4.6 and 4.7.

Results

The nodes of the minimum spanning trees seen in figure 4.6 display areas in red to orange where the high values a marker was found. The areas outlined in small dashes indicate $CD4^+$ areas of the network. In the second column of networks the populations outlined by the large dash show $CD4^+$ T cells that are also expressing $CD154$. Because $CD154$ is critical for the development of T cell effector function, these areas are where we should focus to visualize cytokine production in response to the antigen treatments [19]. Figure 4.7 retains the $CD4^+CD154^+$ annotations of

the networks but displays marker intensity based on the five cytokines profiled. The no antigen control as expected has very few nodes of intensity for any cytokine. SCHU4 however displays a variety of effector T cell function. IFNγ, IL-2, and TNFα can be seen being co-produced in many nodes indicating a T_H1 response. Of further interest is that two nodes show high expression of IL-17 while also being negative for the production of IL-2. IL-2 has been shown to inhibit the differentiation of T cells into T_H17 indicating that what we are detecting within these networks is truly representative of T cell activation and differentiation [28, 29].

Another interesting thing to note is that within the no antigen network the populations are much more uniform in size than any of the networks produced by a stimulated culture. This could potentially indicate CD4 proliferation and differentiation into effector cells. While this is a speculative hypothesis it sparks and interesting idea for future research into different topological features of the networks and how they could be quantified to add further value to this exploratory analysis pathway. What could the number of nodes at a set similarity threshold tell from sample to sample? What does the node size, homo or heterogeneity, and branching within cell populations tell us about immune activity?

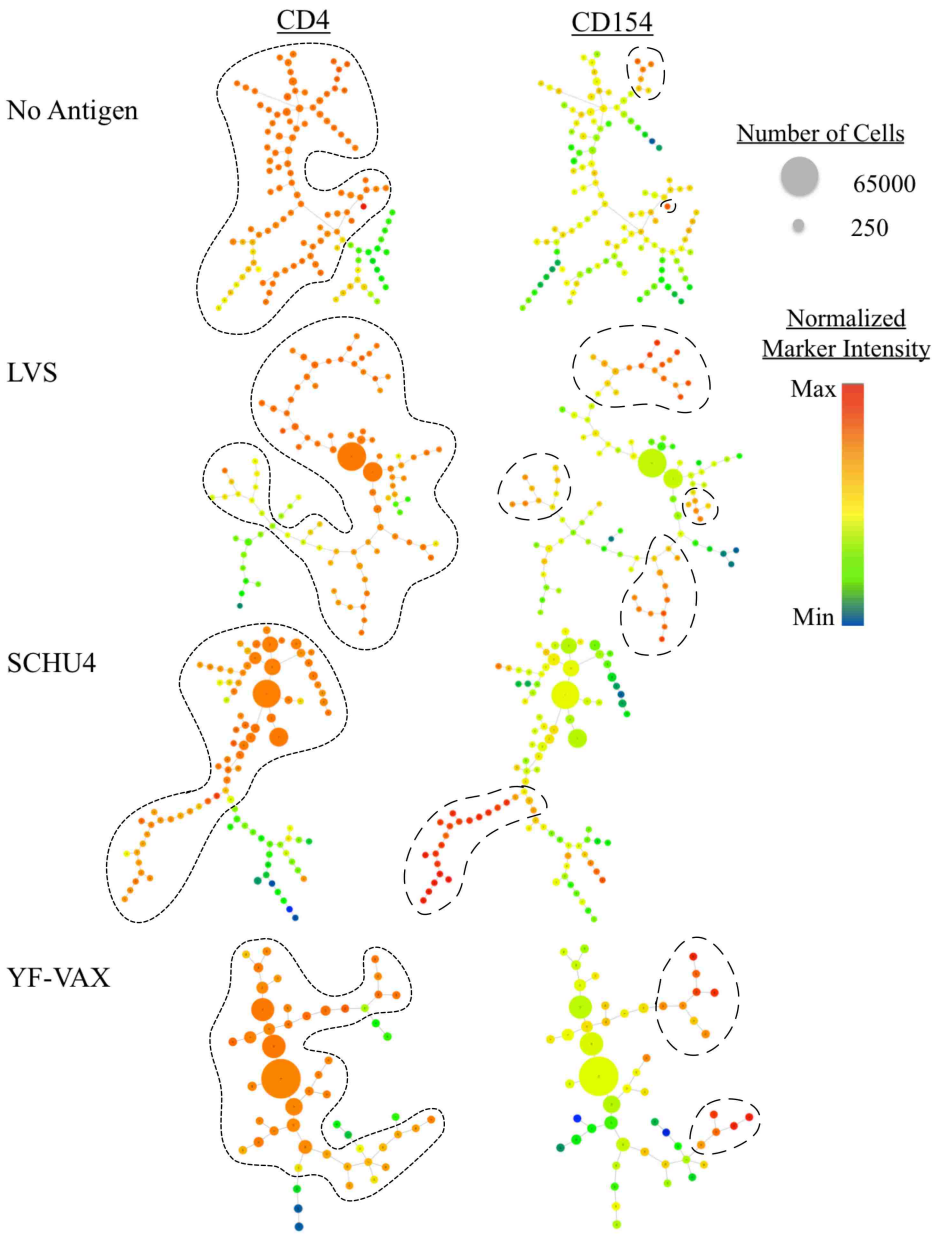


Figure 4.6: Areas of the network indicated by the small dash lines in the first column show $CD4^+$ T cells. Those indicated by the larger dashed lines in the second column indicate activated T cells responding to antigen.

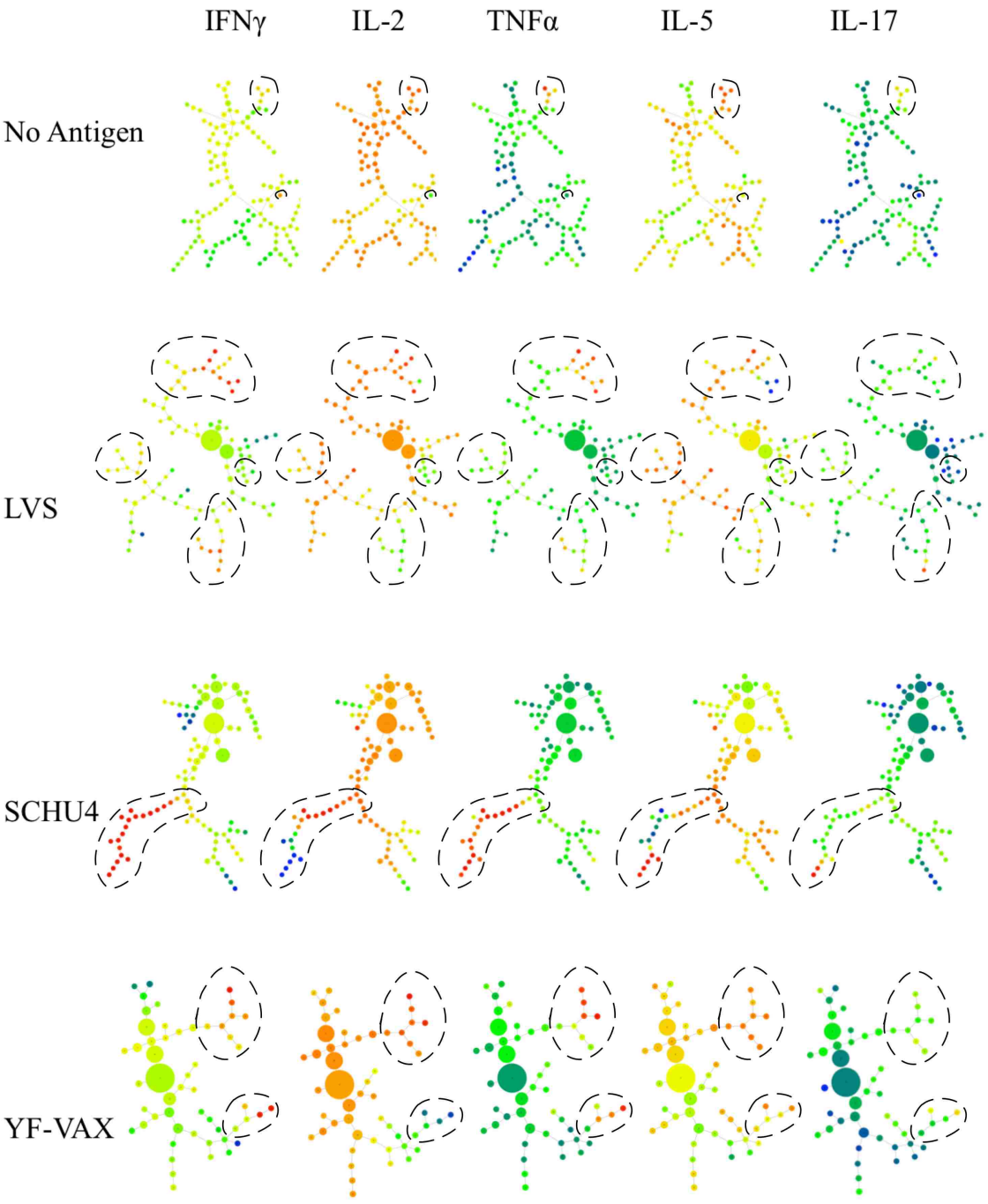


Figure 4.7: Within the networks areas of different T_H cell subsets can be found.

CHAPTER 5: CONCLUSION

Histogram Matching Support Vector Machine Gating

The histogram matching support vector machine (HMSVM) gating of flow cytometry data proved to be both accurate for the quantification of live cells per volume and exhibited the least variation among the commercially available cell counters tested. With a correlation coefficient of greater than 0.98 with manual counting and a coefficient of variation between counts of approximately 5%, this method has proved to be a highly valuable asset to the lab. This procedure saves staff-hours as well as adding consistency to experimental results. To date, this method has been used within our lab to count thousands of samples, normalizing cultures for CD4 and CD8 ICCS polyfunctional T cell readouts. However, there are two open problems within this analysis method that I would like to address. The first is to determine if the histogram mismatch threshold discovered gating live and dead lymphocytes using FCS and SSC holds true for other two-dimensional gates. The second is to create an appropriate method to select the best FCS files to create a gate and remove those not needed to reduce the time to search for the best match.

Despite these open questions, preliminary work had been done using a hierarchical implementation of this method on a multi-dimensional quality control PBMC dataset. Because this method is intended to be used on experiments that have a set flow cytometry staining panel as well as an established analysis gating pathway, this group of experiments seemed perfect for this application. Within our laboratory over $1,000$ human leukapheresis donations have been processed over the last 16 months. To ensure that when these stocks of cells are needed for experiments that they will produce valid experimental results a quality control experiment is conducted on every sample. This experimental protocol involves thawing a cryopreserved PBMC donation and labeling them with CFSE, a fluorescent staining dye to monitor cell proliferation. Next, four stimulation

conditions: no treatment, PMA/PHA, PHA, or Cytostim are added to the PBMCs, and they are incubated for five days. At the end of the incubation, the four samples are stained for flow cytometry evaluation using PO-PRO to asses viability, CD4, CD19, CD8, to analyze CD4 T_H cell, B cell, CD8 cytotoxic cell proportions, and CD25 to identify cellular activation. The analysis of the resulting four FCS data files is processed by a similar hierarchical gating scheme as seen in figure 2.2. Briefly, lymphocytes are gated, then live lymphocytes are selected from this population for further analysis. From the live lymphocyte population $CD4^+$, $CD8^+$, and $CD19^+$ cells are isolated into distinct populations. Each lymphocyte population is then analyzed for increases in CD25 indicating activation and decreases in CFSE showing cell proliferation. An example of the HMSVM results of one donor's PBMC QC assay is shown in figure 5.1. Once this analysis including the dot plots and populations statistics is complete, the researcher must use this information to classify if the donation is fit for use in further experiments or not. This pass or fail classification in itself is also a somewhat subjective process. The researcher looks at cell viability, the proportions of lymphocytes as well as comparing their lack of response in the no antigen control to the activation and proliferation in the cultures containing stimulus.

The final goal in implementing HMSVM gating for a multidimensional dataset such as this will be to use another classification method to output the end result of an analysis. In this case a quality pass or fail. Preliminary testing using 50 pass and 50 fail PBMC QC donations were run using the analysis method that produced the results in figure 5.1. A decision tree was trained using the population statistics from these 100 QC experiments using the Python package scikit-learn [43]. The testing set comprised of 30 new passing samples and 20 new failing QC experiments. The results from this classification were very encouraging with an 82% classification accuracy compared to the researcher's classification. Of the 18% or 9 donations that were classified incorrectly 3 were classified falsely as failing QC and 6 were classified falsely as passing QC.

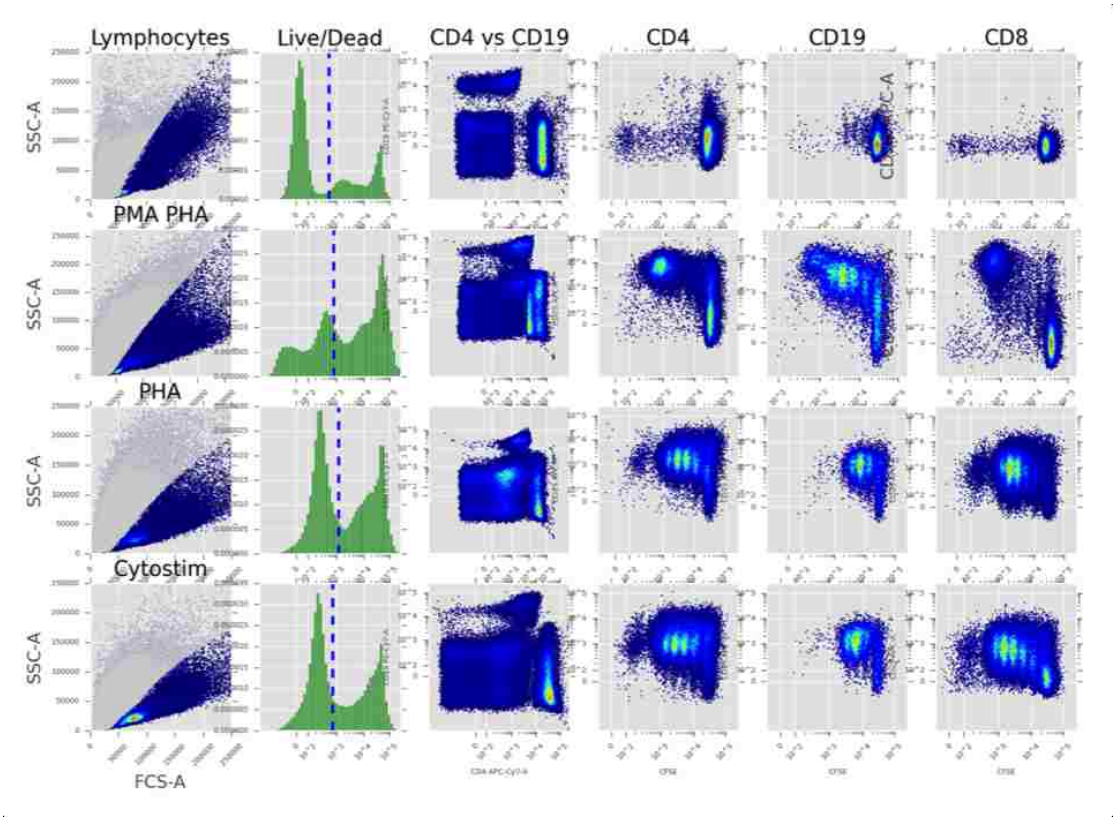


Figure 5.1: Example of a PBMC quality control analysis generated using HMSVM gating. Each row of plots shows a stimulation condition: No antigen control, PMA/PHA, PHA, and Cytostim. From these four sample's flow cytometry dot plots a researcher decides if a donation passes or fails quality inspections. We aim to automate the gating process as well as the final pass or fail classification.

Exploratory Network

Leveraging the nature of flow cytometry data, that events cluster together and are not uniformly distributed, we could sample the data quickly to preserve rare events. Using a hypergraph, rare events were distinguished from common events, and the distribution of hyperedge weights produced the sampled subset of FCS data. Then using the U and Σ matrices of SDV the events could be clustered without the need of a user-defined cluster number. The clustered data was then used

to create, train, and validate a SVM to classify the remaining data from the FCS file not selected in the sampling step. Finally, a minimum spanning tree was created where the visualization of T_H subsets was possible. Comparing the normalized marker intensities over networks created using different vaccine treatments we were able to see that LVS and SCHU4 were capable of generating a T_H 17 response in the donor tested while YF-VAX did not for example, however, what else can we learn from these networks? Additional research is needed to determine what the topology of these networks can teach us about the cell cultures. For example, the no antigen control sample in figure 4.7 has nodes much more uniform in size throughout the entire network compared to the samples treated with vaccine or antigen. This could potentially indicate the lack of CD4 proliferation and differentiation in the no antigen culture. The vaccine and antigen-stimulated cultures, in contrast, have large $CD4^{+}CD154^{-}$ nodes which could indicate naive CD4 expansion. While this is a speculative hypothesis, it sparks an interesting avenue of future research. Running this analysis over many more files an average network topology could be discovered leading to a possible classification method of vaccine immunogenicity for example.

The second problem for additional focus is how to use multiple FCS files to create one descriptive network. It would be simple to append many FCS files together, creating a large FCS file of multiple donors over the same culture condition for analysis using this algorithm. However, this would only produce valid results if the samples were run using the same antibody-fluorochrome lots, cytometer settings, and specific staining protocol. Small variations in the experimental setup, not to mention biological variations, could cause a shift in populations and lead to a network that was not truly representative of the average of the files. For this type of network to be implemented normalization in the position of populations would need to occur before the algorithm was run. There has been some work on this in the past. Methods include gaussNorm and fdaNorm both available in the open source toolkit Bioconductor [21, 23]. These use density estimates of the data to match relevant population peaks between samples and transform the data to minimize the

distance between these landmarks [23]. These methods were tested using the polyfunctional CD4 *in vitro* dataset shown previously. Both methods tended to view the rare subset of T cells as outliers moving the populations closer to the median value for all cytokine markers, effectively eliminated them from the dataset. We have explored our own methods of population alignment, with the most successful being an implementation of derivative dynamic time warping [25, 51], but much more work is needed in this area to faithfully normalize population peaks within the data space without losing essential data.

LIST OF REFERENCES

[1] Abul K Abbas, Andrew HH Lichtman, and Shiv Pillai. *Basic immunology: functions and disorders of the immune system*. Elsevier Health Sciences, 2012.

[2] Marlene Absher. Hemocytometer counting. In *Tissue Culture*, pages 395–397. Elsevier, 1973.

[3] Nima Aghaeepour, Greg Finak, FlowCAP Consortium, DREAM Consortium, Holger Hoos, Tim R Mosmann, Ryan Brinkman, Raphael Gottardo, and Richard H Scheuermann. Critical assessment of automated flow cytometry data analysis techniques. *Nat Methods*, 10(3):228–38, Mar 2013.

[4] Nima Aghaeepour, Radina Nikolic, Holger H Hoos, and Ryan R Brinkman. Rapid cell population identification in flow cytometry data. *Cytometry Part A*, 79(1):6–13, 2011.

[5] El-ad David Amir, Kara L Davis, Michelle D Tadmor, Erin F Simonds, Jacob H Levine, Sean C Bendall, Daniel K Shenfeld, Smita Krishnaswamy, Garry P Nolan, and Dana Pe'er. visne enables visualization of high dimensional single-cell data and reveals phenotypic heterogeneity of leukemia. *Nat Biotechnol*, 31(6):545–52, Jun 2013.

[6] Lora W Barsky, Michele Black, Matthew Cochran, Benjamin J Daniel, Derek Davies, Monica DeLay, Rui Gardner, Michael Gregory, Desiree Kunkel, Joanne Lannigan, et al. International society for advancement of cytometry (isac) flow cytometry shared resource laboratory (srl) best practices. *Cytometry Part A*, 89(11):1017–1030, 2016.

[7] Ali Bashashati and Ryan R Brinkman. A survey of flow cytometry data analysis methods. *Adv Bioinformatics*, page 584603, 2009.

[8] D.R. Drake III B.C Schanen. A novel approach for the generation of human dendritic cells from blodd monocytes in the absence of exogenous factors. *J. Immunol. Methods*, 335:53, 2008.

[9] Richard E Bellman. *Dynamic programming*. Courier Corporation, 2013.

[10] Richard E Bellman. *Adaptive control processes: a guided tour*. Princeton university press, 2015.

[11] A Bernard and L Boumsell. Human leukocyte differentiation antigens. *Press medicale (Paris, France: 1983)*, 13(38):2311–2316, 1984.

[12] Sanchita Bhattacharya, Sandra Andorf, Linda Gomes, Patrick Dunn, Henry Schaefer, Joan Pontius, Patty Berger, Vince Desborough, Tom Smith, John Campbell, et al. Immport: disseminating data to the public for the future of immunology. *Immunologic research*, 58(2-3):234–239, 2014.

[13] Raul C Braylan. Impact of flow cytometry on the diagnosis and characterization of lymphomas, chronic lymphoproliferative disorders and plasma cell neoplasias. *Cytometry Part A*, 58(1):57–61, 2004.

[14] JKC Chan, CS Ng, and PK Hui. A simple guide to the terminology and application of leucocyte monoclonal antibodies. *Histopathology*, 12(5):461–480, 1988.

[15] Chih-Chung Chang and Chih-Jen Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

[16] Pratip K Chattopadhyay, Joanne Yu, and Mario Roederer. A live-cell assay to detect antigen-specific cd4+ t cells with diverse cytokine profiles. *Nature medicine*, 11(10):1113, 2005.

[17] Patricia A Darrah, Dipti T Patel, Paula M De Luca, Ross WB Lindsay, Dylan F Davey, Barbara J Flynn, Søren T Hoff, Peter Andersen, Steven G Reed, Sheldon L Morris, et al. Multifunctional t h 1 cells define a correlate of vaccine-mediated protection against leishmania major. *Nature medicine*, 13(7), 2007.

[18] Interpreting flow cytometry data: a guide for the perplexed. Interpreting flow cytometry data: a guide for the perplexed. *Interpreting flow cytometry data: a guide for the perplexed*, 7(7):681–685, 2006.

[19] Marco Frentsch, Olga Arbach, Dennis Kirchhoff, Beate Moewes, Margitta Worm, Martin Rothe, Alexander Scheffold, and Andreas Thiel. Direct access to cd4+ t cells specific for defined antigens according to cd154 expression. *Nature medicine*, 11(10):1118–1124, 2005.

[20] Yongchao Ge and Stuart C Sealfon. flowpeaks: a fast unsupervised clustering for flow cytometry data via k-means and density peak finding. *Bioinformatics*, 28(15):2052–2058, 2012.

[21] Bates D.M. Gentleman R.C., Carey V.J. Bioconductor: open software development for computational biology and bioinformatics. http://www.bioconductor.org/, 2004.

[22] Aric A. Hagberg, Daniel A. Schult, and Pieter J. Swart. Exploring network structure, dynamics, and function using NetworkX. In *Proceedings of the 7th Python in Science Conference (SciPy2008)*, Aug, 2008.

[23] Florian Hahne, Alireza Hadj Khodabakhshi, Ali Bashashati, Chao-Jen Wong, Randy D Gascoyne, Andrew P Weng, Vicky Seyfert-Margolis, Katarzyna Bourcier, Adam Asare, Thomas Lumley, Robert Gentleman, and Ryan R Brinkman. Per-channel basis normalization methods for flow cytometry data. *Cytometry A*, 77(2):121–31, Feb 2010.

[24] Laurie E Harrington, Robin D Hatton, Paul R Mangan, Henrietta Turner, Theresa L Murphy, Kenneth M Murphy, and Casey T Weaver. Interleukin 17–producing cd4+ effector t cells

develop via a lineage distinct from the t helper type 1 and 2 lineages. *Nature immunology*, 6(11):1123, 2005.

[25] Eamonn J Keogh and Michael J Pazzani. Derivative dynamic time warping. In *Proceedings of the 2001 SIAM International Conference on Data Mining*, pages 1–11. SIAM, 2001.

[26] Thomas J Kindt, Richard A Goldsby, Barbara A Osborne, and Janis Kuby. *Kuby immunology*. Macmillan, 2007.

[27] Nikesh Kotecha, Peter O Krutzik, and Jonathan M Irish. Web-based analysis and publication of flow cytometry experiments. *Current protocols in cytometry*, pages 10–17, 2010.

[28] Arian Laurence, Cristina M Tato, Todd S Davidson, Yuka Kanno, Zhi Chen, Zhengju Yao, Rebecca B Blank, Françoise Meylan, Richard Siegel, Lothar Hennighausen, et al. Interleukin-2 signaling via stat5 constrains t helper 17 cell generation. *Immunity*, 26(3):371–381, 2007.

[29] Wei Liao, Jian-Xin Lin, and Warren J Leonard. Interleukin-2 at the crossroads of effector responses, tolerance, and immunotherapy. *Liao, Wei and Lin, Jian-Xin and Leonard, Warren J*, 38(1):13–25, 2013.

[30] Jeff W Lichtman and José-Angel Conchello. Fluorescence microscopy. *Nature methods*, 2(12):910, 2005.

[31] Kenneth Lo, Florian Hahne, Ryan R Brinkman, and Raphael Gottardo. flowclust: a bioconductor package for automated gating of flow cytometry data. *BMC Bioinformatics*, 10:145, 2009.

[32] Robert McGill, John W Tukey, and Wayne A Larsen. Variations of box plots. *The American Statistician*, 32(1):12–16, 1978.

[33] Janice M Moser, Emily R Sassano, Del C Leistritz, Jennifer M Eatrides, Sanjay Phogat, Wayne Koff, and Donald R Drake. Optimization of a dendritic cell-based assay for the in vitro priming of naive human cd4+ t cells. *Journal of immunological methods*, 353(1):8–19, 2010.

[34] Timothy R Mosmann, Holly Cherwinski, Martha W Bond, Martin A Giedlin, and Robert L Coffman. Two types of murine helper t cell clone. i. definition according to profiles of lymphokine activities and secreted proteins. *The Journal of immunology*, 136(7):2348–2357, 1986.

[35] Kenneth M Murphy. *Janeway's immunobiology*. Garland Science, 2011.

[36] Robert F Murphy. Automated identification of subpopulations in flow cytometric list mode data using cluster analysis. *Cytometry*, 6(4):302–309, 1985.

[37] David Novo and James Wood. Flow cytometry histograms: transformations, resolution, and display. *Cytometry A*, 73(8):685–92, Aug 2008.

[38] Alberto Orfao, Francisco Ortuño, Maria de Santiago, Antonio Lopez, and Jesus San Miguel. Immunophenotyping of acute leukemias and myelodysplastic syndromes. *Cytometry A*, 58(1):62–71, Mar 2004.

[39] John J O'Shea and William E Paul. Mechanisms underlying lineage commitment and plasticity of helper cd4+ t cells. *Science*, 327(5969):1098–1102, 2010.

[40] Eleni Panagioti, Paul Klenerman, Lian N Lee, Sjoerd H Van Der Burg, and Ramon Arens. Features of effective t cell-inducing vaccines against chronic viral infections. *Frontiers in Immunology*, 9:276, 2018.

[41] Heon Park, Zhaoxia Li, Xuexian O Yang, Seon Hee Chang, Roza Nurieva, Yi-Hong Wang, Ying Wang, Leroy Hood, Zhou Zhu, Qiang Tian, et al. A distinct lineage of cd4 t cells

regulates tissue inflammation by producing interleukin 17. *Nature immunology*, 6(11):1133, 2005.

[42] David R Parks, Mario Roederer, and Wayne A Moore. A new "logicle" display method avoids deceptive effects of logarithmic scaling for low signals and compensated data. *Cytometry Part A*, 69(6):541–551, 2006.

[43] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[44] Carlos E Pedreira, Elaine S Costa, Quentin Lecrevisse, Jacques JM van Dongen, Alberto Orfao, EuroFlow Consortium, et al. Overview of clinical flow cytometry data analysis: recent advances and future challenges. *Trends in biotechnology*, 31(7):415–425, 2013.

[45] Saumyadipta Pyne, Xinli Hu, Kui Wang, Elizabeth Rossin, Tsung-I Lin, Lisa M Maier, Clare Baecher-Allan, Geoffrey J McLachlan, Pablo Tamayo, David A Hafler, et al. Automated high-dimensional flow cytometric data analysis. *Proceedings of the National Academy of Sciences*, 106(21):8519–8524, 2009.

[46] Yu Qian, Chungwen Wei, F Eun-Hyung Lee, John Campbell, Jessica Halliley, Jamie A Lee, Jennifer Cai, Y Megan Kong, Eva Sadat, Elizabeth Thomson, et al. Elucidation of seventeen human peripheral blood b-cell subsets and quantification of the tetanus response using a density-based method for the automated identification of cell populations in multidimensional flow cytometry data. *Cytometry Part B: Clinical Cytometry*, 78(S1):S69–S82, 2010.

[47] Peng Qiu, Erin F Simonds, Sean C Bendall, Kenneth D Gibbs, Jr, Robert V Bruggner, Michael D Linderman, Karen Sachs, Garry P Nolan, and Sylvia K Plevritis. Extracting

a cellular hierarchy from high-dimensional cytometry data with spade. *Nat Biotechnol*, 29(10):886–91, Oct 2011.

[48] Mario Roederer. Spectral compensation for flow cytometry: visualization artifacts, limitations, and caveats. *Cytometry*, 45(3):194–205, 2001.

[49] Mario Roederer, Joshua L Nozzi, and Martha C Nason. Spice: Exploration and analysis of post-cytometric complex multivariate datasets. *Cytometry Part A*, 79(2):167–174, 2011.

[50] Yvan Saeys, Sofie Van Gassen, and Bart N Lambrecht. Computational flow cytometry: helping to make semse of high-dimensional immunology data. *Nature Reviews Immunology*, 2016.

[51] Stan Salvador and Philip Chan. Toward accurate dynamic time warping in linear time and space. *Intelligent Data Analysis*, 11(5):561–580, 2007.

[52] Brian C Schanen, Anne S De Groot, L Moise, Matt Ardito, Elizabeth McClaine, William Martin, Vaughan Wittman, William L Warren, and Donald R Drake. Coupling sensitive in vitro and in silico techniques to assess cross-reactive cd4+ t cells against the swine-origin h1n1 influenza virus. *Vaccine*, 29(17):3299–3309, 2011.

[53] Robert A Seder, Patricia A Darrah, and Mario Roederer. T-cell quality in memory and protection: implications for vaccine design. *Nature Reviews Immunology*, 8(4):247, 2008.

[54] Paul Shannon, Andrew Markiel, Owen Ozier, Nitin S Baliga, Jonathan T Wang, Daniel Ramage, Nada Amin, Benno Schwikowski, and Trey Ideker. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome research*, 13(11):2498–2504, 2003.

[55] Howard M Shapiro. *Practical flow cytometry*. John Wiley and Sons, 2005.

[56] Josef Spidlen, Wayne Moore, David Parks, Michael Goldberg, Chris Bray, Pierre Bierre, Peter Gorombey, Bill Hyun, Mark Hubbard, Simon Lange, et al. Data file standard for flow cytometry, version fcs 3.1. *Cytometry Part A*, 77(1):97–100, 2010.

[57] Warren Strober. Trypan blue exclusion test of cell viability. *Current protocols in immunology*, 21(1):A–3B, 1997.

[58] Michael J Swain and Dana H Ballard. Color indexing. *International journal of computer vision*, 7(1):11–32, 1991.

[59] Michael J Swain and Dana H Ballard. Indexing via color histograms. In *Active Perception and Robot Vision*, pages 261–273. Springer, 1992.

[60] John W Tukey. *Exploratory data analysis*. Reading, Mass., 1977.

[61] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(2579-2605):85, 2008.

[62] Sonja Wulff, Kathleen Martin, Angela Vandergaw, Thomas Boenisch, Ian Brotherick, Terry Hoy, Jim Hudson, Christina Jespersgaard, Peter Lopez, Alberto Orfao, et al. *Guide to flow cytometry*. Dako Cytomation., 2006.

[63] Jinfang Zhu, Hidehiro Yamane, and William E Paul. Differentiation of effector cd4 t cell populations. *Annual review of immunology*, 28:445–489, 2009.

[64] Heddy Zola, Bernadette Swart, Ian Nicholson, Bent Aasted, Armand Bensussan, Laurence Boumsell, Chris Buckley, Georgina Clark, Karel Drbal, Pablo Engel, et al. Cd molecules 2005: human cell differentiation molecules. *Blood*, 106(9):3123–3126, 2005.