Computational Exploratory Analysis of High-Dimensional Flow Cytometry Data for Diagnosis and Biomarker Discovery

by

Nima Aghaeepour

B. Sc, University of Tehran, 2003

A THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF

Doctor of Philosophy

in

THE FACULTY OF GRADUATE STUDIES
(Bioinformatics)

The University Of British Columbia
(Vancouver)

December 2012

Abstract

Flow Cytometry (FCM) is widely used to investigate and diagnose human disease. Although high-throughput systems allow rapid data collection from large cohorts, manual data analysis can take months. Moreover, identification of cell populations can be subjective, and analysts rarely examine the entirety of the multidimensional dataset (focusing instead on a limited number of subsets, the biology of which has usually already been well-described). Thus, the value of Polychromatic Flow Cytometry (PFC) as a discovery tool is largely wasted.

In this thesis, I will present three computational tools that once merged together provide a complete pipeline for analysis and visualization of FCM data: (1) a clustering algorithm for identification of homogeneous groups of cells (cell populations); (2) a set of statistical tools for identifying immunophenotypes (based on the cell populations) that are correlated with an external variable (*e.g.,* a clinical outcome); (3) a tool for identifying the most important parent populations that can best describe a set of related immunophenotypes. In addition to technical advancements, this pipeline represents a conceptual advance that allows a more powerful, automated, and complete analysis of complex flow cytometry data than previously possible. As a side product, this pipeline allows complex information from PFC studies to be translated into clinical or resource-poor settings, where multiparametric analysis is less feasible. I demonstrated the utility of this approach in a large ($n = 466$), retrospective, 14-parameter PFC study of early HIV infection, where we identified three T-cell subsets that strongly predicted progression to AIDS (only one of which was identified by an initial manual analysis).

Before and during the development of this pipeline, a wide range of computational tools for analysis of FCM data were published. However, guidance for end

users about appropriate use and application of these methods is scarce. The Flow Cytometry: Critical Assessment of Population Identification Methods (FlowCAP) is a highly collaborative project for evaluation of these computational tools using real-world datasets. The FlowCAP results presented here will help both computational and biological scientists to better develop and use advanced bioinformatics pipelines.

Preface

A version of chapter 2 has been published (*N. Aghaeepour, R. Nikolic, H. Hoos, and R. Brinkman. Rapid cell population identification in flow cytometry data. Cytometry Part A, 79(1): 6-13, 2011*). I designed, implemented, and evaluated the methodology and contributed to writing the manuscript. Three other co-authors contributed to the design of the methodology as well as to writing and editing of the manuscript.

A version of chapter 3 has been published (*N. Aghaeepour, P. K. Chattopadhyay, A. Ganesan, K. ONeill, H. Zare, A. Jalali, H. H. Hoos, M. Roederer, and R. R. Brinkman. Early Immunologic Correlates of HIV Protection can be Identified from Computational Analysis of Complex Multivariate T-cell Flow Cytometry Assays. Bioinformatics, 28(7):1009-1016, 2012*). I implemented and evaluated the methodology and contributed to designing the study and writing the manuscript. Pratip Chattopadhyay, the co-lead author of the manuscript, produced the dataset, interpreted the results and contributed to designing the methodology and writing the manuscript. Eight other co-authors contributed to designing the methodology, producing the data, analyzing the results, as well as to writing and editing of the manuscript.

A version of chapter 4 has been accepted for publication (*N. Aghaeepour, A. Jalali, K. ONeill, P. Chattopadhyay, M. Roederer, H. H.H., and R. Brinkman. Rchy-Optimyx: cellular hierarchy optimization for flow cytometry. Accepted, Cytometry Part A, 2012*). I designed the study and contributed to its implementation and evaluation. I also contributed to writing the manuscript. Adrin Jalali, the co-lead author of the manuscript, contributed to designing and implementing the methodology. Five other co-authors contributed to supervising the project as well as to

writing and editing of the manuscript.

A version of chapter 5 has been accepted for publication (*N. Aghaeepour, G. Finak, D. Dougall, A. Hadj-Khodabakhshi, P. Mah, G. Obermoser, J. Spidlen, I. Taylor, S. A. Wuensch, J. Bramson, C. Eaves, A. P. Weng, E. S. F. III, K. Ho, T. Kollmann, W. Rogers, S. D. Rosa, B. Dalal, A. Azad, A. Pothen, A. Brandes, H. Bretschneider, R. Bruggner, R. Finck, R. Jia, N. Zimmerman, M. Linderman, D. Dill, G. Nolan, C. Chan, F. E. Khettabi, K. ONeill, M. Chikina, A. Gupta, P. Shooshtari, H. Zare, P. L. D. Jager, M. Jiang, J. Keilwagen, J. M. Maisog, P. Majek, J. Vilcek, T. Manninen, H. Huttunen, P. Ruusuvuori, M. Nykter, G. J. McLachlan, K. Wang, I. Naim, G. Sharma, R. Nikolic, S. Pyne, Y. Qian, P. Qiu, J. Quinn, A. Roth, R. Norel, G. Stolovitzky, P. Meyer, J. Saez-Rodriguez, M. Bhattacharjee, M. Biehl, P. Bucher, K. Bunte, B. D. Camillo, S. Dimitrieva, J. Grau, I. Grosse, S. Posch, N. Guex, J. Keilwagen, M. Kursa, B. Liu, M. Maienschein-Cline, T. Manninen, G. J. McLachlan, K. Wang, S. Pyne, P. Qiu, P. S. Seifert, M. Strickert, J. M. G. Vilar, H. Hoos, T. Mosmann, R. Gottardo, R. Brinkman, and R. H. Scheuermann. Critical Assessment of Cell Population Identification Techniques for Flow Cytometry Data: Results of FlowCAP. Accepted, Nature Methods, 2012)*. I designed the methodology and contributed to its implementation and evaluation. I also contributed to writing the manuscript. This project received significant support from Greg Finak (the second author of the manuscript). The project was also supported by the FlowCAP organizing committee and the Dialogue for Reverse Engineering Assessments and Methods (DREAM) initiative as well as to algorithm developers and data providers from various research groups (the full list of the 91 co-authors and their contributions is provided in the manuscript).

The *flowType-FeaLect* entry in FlowCAP-II (chapter 5) is joint work with Habil Zare (a Ph.D. student in the Brinkman Lab at the time). We both contributed equally to this work.

Table of Contents

List of Tables

List of Figures

Glossary

AIC Akaike Information Criterion

AIDS Acquired Immunodeficiency Syndrome

BIC Bayesian Information Criterion

CDF Cumulative Distribution Function

CI Confidence Interval

CPHR Cox Proportional Hazards Regression

DLBCL Diffuse Large B-Cell Lymphoma

DREAM Dialogue for Reverse Engineering Assessments and Methods

EM Expectation Maximization

FACS Fluorescence-activated Cell Sorting

FCM Flow Cytometry

FITMAN Flow Immunophenotyping Technical Meetings

FOCIS Federation of Clinical Immunology Societies Federation of Clinical Immunology Societies

FSC Forward Scatter

GMM Gaussian mixture model

GVHD Graft versus Host Disease

HAART Highly Active Anti-Retroviral Therapy

HIPC Human Immuno Phenotyping Consortium

HIV Human Immunodeficiency Virus

ICL Integrated Complete Likelihood

KDE Kernel Density Estimation

OMIPS Optimized Multicolor Immunophenotyping Panels

PFC Polychromatic Flow Cytometry

SIV Simian Immunodeficiency Virus

SSC Side Scatter

SWR Scale-free Weighted Ratio

Acknowledgements

This work wouldn't have been possible without the support of my supervisors, Ryan Brinkman and Holger Hoos as well as my thesis committee - Paul Pavlidis, Steven Jones, and Matias Salibian-Barrera. Much of the methodology presented in this work were designed in a collaboration with Mario Roederer's group. The FlowCAP project is a long term collaboration with Raphael Gottardo, Tim Mosmann, and Richard Scheuermann.

I would like to thank the *ISAC Scholar Program*, *University of British Columbia's Four Year Fellowship*, and the *CIHR/MSFHR Strategic Training Program in Bioinformatics for Health Research* for funding my graduate studies. In addition, on behalf of the co-authors of the four manuscripts that describe this work I would like to acknowledge the following grants and scholarships: NIH grant 1R01EB008400, NIH/N01AI40076, NIH/R01NS067305, NIH/RC2-GM093080, CCS #700374, an NSERC Discovery Grant held by Holger Hoos, NIAID Intramural Research Program; NIH/NIBIB grant EB008400, Michael Smith Foundation for Health Research Scholar Award to Ryan Brinkman, and the Terry Foundation and The Terry Fox Research Institute. The FlowCAP summits - held on the NIH campus, Bethesda, MD, United States, 2010 and 2011 - were generously sponsored by NIH/NIAID.

This work is dedicated to my partner, parents, and my brother as well as to Damon Lindelof and Masoud Hashemian for bringing hope, emotion, trust, and science to my life.

Chapter 1

Introduction

Flow Cytometry (FCM) is the primary tool for measurement of multiple markers (primarily surface or intra-cellular proteins) simultaneously on single cells in a high-throughput fashion. FCM enables investigators to divide cells to subsets based on their phenotype and/or function. This makes this technology a very powerful tool for exploratory analysis of cellular systems for designing diagnosis tests, identifying targets for therapies, and monitoring the progression of diseases. In addition, FCM's ability to isolate cell subsets based on their phenotype has made it a unique tool for *in vivo* and *in vitro* studies of homogeneous cell populations.

In FCM, cells are labelled with fluorescent markers and are then moved past a laser beam that excites the fluorochrome (a fluorescent molecule that can emit light upon excitation) one cell at a time. The light emitted from each individual cell is collected using a series of light and colour detectors. In addition to fluorescence intensity values for each marker, measurements also include Forward Scatter (FSC) and Side Scatter (SSC), which correlate to the cell size and granularity, respectively. Traditionally, FCM analysis has been a labour-intensive process [8]. New technological developments have made it possible to apply FCM in a high-throughput fashion, rendering data analysis a significant challenge.

The FCM hardware was mostly developed in the 1990s, data analysis tools capable of analyzing a large number of measurements per cell are not yet widely available [71]. For the recently developed mass cytometry [11] and single-cell gene-expression analysis instruments [22], even partial exploration of the high

multidimensional space through conventional manual analysis approaches is no longer feasible.

1.1 Cell Population Identification

Identification of homogenous groups of cells for further study (gating) is critical in analysis of FCM data. Most frequently, this is done by drawing polygons on series of bi-variate scatter plots produced from two dimensional projections of the data (*a.k.a.* manual gating). However, manual gating is subject to user variability [47, 108, 121] and is unsuitable for high-throughput data analysis [48]. In addition, the number of bi-variate plots that need to be analyzed grows exponentially by increasing the number of measurements per cell, rendering a complete manual analysis of even less than ten dimensions unfeasible.

Clustering is the problem of partitioning an unlabelled set of multi-dimensional vectors into groups (clusters) of "similar" points. This problem is similar to the population identification problem for FCM data in that in both cases the target is identification of homogeneous subsets of a multi-dimensional space. Over the past decades, an extensive amount of research has been dedicated to designing objective functions that, at least implicitly, define what constitutes a cluster and optimizing them (*i.e.*, finding the best clusters given an objective function)[51]. Here, I will discuss the main classes of clustering algorithms: 1) hierarchical, 2) spectral, 3) density-based, and 4) model-based clustering:

Hierarchical Clustering

Hierarchical clustering algorithms are based on the idea that successive clusters can be inferred from previously established clusters. For example, agglomerative (bottom up) hierarchical algorithms begin with each element as a separate cluster and merge them into successively larger clusters. The number of clusters (or an equivalent threshold) needs to be pre-identified for these algorithms.

One of the main drawbacks of hierarchical clustering algorithms is the amount of resources that they require. These algorithms use a similarity matrix. Creating and storing this matrix requires $O(n^2)$ time and memory, where n is the number of data points. Due to these time and memory requirements, hierarchical clustering

algorithms have not been successfully applied to FCM experiments [8].

Spectral Clustering

In spectral clustering, first a graph is produced in which every cell is considered a node and the length of the edges represent the multidimensional distance between the cells; then the graph is partitioned into different sub-graphs using objective functions from graph theory. One of the most prominent approaches is partitioning the graph into sub-graphs that minimize the normal cuts of each partition [127]. Spectral clustering algorithms (like hierarchical ones) work based on a similarity matrix and cannot be directly applied to large datasets. However, they can automatically select the number of clusters using their objective function (*e.g.*, the SamSPECTRAL algorithm [127] as discussed below).

Density-based Clustering

Density-based clustering is based on the assumption that a cluster is a region with high density [34]. This enables this class of algorithms to avoid using a large similarity matrix. Estimation of the density is usually performed using smoothing methods (*e.g.*, Kernel Density Estimation (KDE), probability binning, etc.[99]). However, this estimation becomes more challenging for high-dimensional datasets [29]. Due to time requirement issues, this approach is usually limited to three or fewer dimensions [29]. In addition, a successful density estimation usually relies on a user-defined "smoothing parameter" (*e.g.*, bandwidth or bin size).

Another challenge in this type of clustering is defining a "high density region". Definitions used in practice are usually based on a manually defined threshold over the estimated density function or its derivatives (see, *e.g.* [84]).

Model-based clustering

Model-based clustering has its roots in the fitting of Gaussian mixture model (GMM)s [70]. The most popular approach is to use Expectation Maximization (EM) to estimate the parameters of a multivariate normal distribution. Then, every point will be assigned to the component (cluster) with maximum posterior probability. Model selection criterions (*e.g.*, Bayesian Information Criterion (BIC), Akaike

Information Criterion (AIC), and Integrated Complete Likelihood (ICL)) can be used to estimate the correct number of clusters [70]. The K-means algorithm can be seen as a special case of GMMs with equal spherical variances. It is possible to construct more robust models using *t* and *skew-t* mixture models at the cost of higher time complexity (see Chapter 2 for more details).

1.1.1 Clustering Algorithms for Cell Population Identification

Several methods have been developed to use the clustering methodologies above to automate the gating process. *flowClust* [70] is a model-based clustering approach that models cell populations using a mixtures of *t*-distributions. Box-Cox transformations are used to remove potential skewness of each component of the mixture model. *flowMerge* [37] extends the flowClust algorithm by applying a cluster merging algorithm [10] to allow multiple components to model the same populations, enabling it to fit concave cell populations. *FLAME* [97] uses a mixture of skew-*t*-distributions to make the model more flexible to skewed cell populations. The *CDP* algorithm uses a GPU-based procedure for fitting Gaussian mixture models in parallel [20]. This can be potentially applied to either flowClust or FLAME for faster analysis. *curvHDR* [84], *FLOCK* [99], *Misty Mountain* [117], and *flowPeaks* [44] are non-parametric density-based approaches, and therefore are not limited to identifying cell populations based on shape but usually have different draw backs. For example, *curvHDR* models cell populations based on the curvature of the underlying distribution; however, it requires user-defined parameter values and cannot be applied to more than three-dimensional data. *SamSpectral* [127] uses an spectral clustering algorithm to find cell populations, including non-convex ones. To deal with the high time and memory requirements of the spectral clustering algorithm, SamSpectral finds cell populations based on representative sub-sampling of the data (also see *SWIFT* [83] for another alternative); however, this can potentially decrease the quality of the gating, as some biological information can be lost during the sampling. SamSpectral also requires user-defined parameter values for each data set of similar experiments.

1.1.2 Fast and Flexible Clustering for Cell Population Identification

Model-based clustering for identification of cell populations can be made more robust to noise using more complex statistical models. However, fitting these complicated models to the data takes longer, to the extent that would render these algorithms useless for larger datasets, even when using state-of-the-art computing clusters. Finak *et. al.* suggested that a post-clustering refining of the identified cell populations (to merge highly overlapping clusters into single cell populations) can improve the results[37]. I hypothesized that this post-clustering step can be much more important than the initial model fitting step, and that replacing the model fitting step with a simpler process can significantly speed up the cell population identification process.

K-means is a fast clustering algorithm that has been widely used in different areas over the past few decades. However, it requires the number of clusters to be pre-specified. This is not possible for most FCM use-cases. It also is very sensitive to the algorithms initialization and is limited to spherical clusters (not all cell population in FCM are spherical). In Chapter 2, I discuss a methodology that refines the clusters produced by K-means using a merging strategy based on a Gaussian mixture model for faster cell population identification without compromising the benefits of model-based clustering.

1.2 Immunophenotyping using FCM for Cross-Sample Exploratory Analysis

A primary use-case of FCM is exploratory analysis of the immune system for identification of immunophenotypes that correlate with a clinical outcome. These cell populations can then be used for diagnosis and monitoring purposes as well as for guiding the design of new therapies. However, manual exploration of high-dimensional datasets in addition to being subjective and error-prone is highly time consuming [75]. At as low as 6 measurements per cell, looking at all possible cell populations becomes very challenging. 13 color experiments are now common in clinical setting, and 40 to 100 dimensional studies in limited scales have also been reported [22]. For none of these a complete manual analysis can be envisioned due to exponential increase in the number of cell populations that can be analyzed and

the number of bi-variate scatter plots that need to be investigated.

Using the computational cell population identification algorithms described above is a natural choice for mining these multi-dimensional spaces. Several recent studies have reported such analysis ([9, 27, 129]). However, multi-dimensional cell population identification for exploratory analysis is associated with several complications. First, the cell populations need to be matched to each other across multiple samples. This process has proven to be subjective, often requiring input from human experts [98]. Second, this approach ignores the hierarchical nature of the cells involved in the immune system by assuming that every cell belongs to only one cell population. However, in presence of a larger number of markers, cell populations should be allowed to overlap (because certain marker combinations might provide partially redundant information) to enable the computational model to explore the exclusion of certain markers to determine if they are clinically relevant. Third, these algorithms do not incorporate the background knowledge of human experts to guide the identification of rare cell populations that cannot be automatically identified.

In Chapter 3, I will describe a methodology that combines several one dimensional cell populations to produce a large number of high-dimensional overlapping clusters. Due to the simple nature of the original one dimensional analysis, incorporating expert knowledge and matching the cell populations across multiple samples becomes very simple. The large number of overlapping cell populations increases the chance of a positive hit in exploratory analysis and reveals important information about the clinical relevance of the markers (this will be used as the basis of Chapter 4).

1.3 Characterization and Visualization of Immunophenotypes

The methodology described in Chapter 3 usually identifies a large number of highly overlapping immunophenotypes (*e.g.,* $CD4^{+}CD8^{-}$ cells are also included in the $CD4^{+}$ immunophenotype). In Chapter 4, I will describe a methodology for organizing these cell populations in a hierarchy, using their most important parent populations (as determined by the strength of the correlation with a clinical

outcome). This approach not only will better visualize the correlates of a clinical outcome, but also helps translate the complicated findings of high-dimensional assays to lower dimensions appropriate for clinical and/or highly regulated settings or for sorting of these populations for *invivo* and *invitro* studies.

1.4 Critical Assessment of Computational Pipelines for Analysis of FCM Data

In absence of public repositories and guidelines in scientific journals that would encourage the publication of FCM data, a very limited amount of high quality data is publicly available. Computational tools for FCM have frequently been tested on small datasets and evaluation of the results have usually relied on visual inspection, providing very limited information about the generalizability of the results and therefore the practical utility of the work in clinical and/or biological settings. In Chapter 5, I will discuss a highly collaborative project in which we evaluated a large number of computational methodologies on a wide range of real world FCM datasets. The use-cases include both cell population identification (identification of all cell populations, *e.g.* for exploratory analysis of the immune response to a vaccine) and sample classification (prediction of an external outcome, *e.g.* a disease outcome) under different settings.

Chapter 2

flowMeans: Rapid Cell Population Identification in Flow Cytometry Data

2.1 Introduction

With the advent of high-throughput FCM analysis, millions of cells can be analyzed for up to 40 markers per sample. For these experiments, the runtime of gating algorithms is a bottleneck of automated FCM data analysis pipelines [8]. The K-means clustering algorithm was the first automated data analysis approaches applied to FCM data [82]. Given a n vectors, $X = (X_1, X_2, ..., X_n)$, of length n, K-means aims to partition X into $K < n$ sets $S = S_1, S_2, ..., S_k$ so as to minimize the within-cluster sum of squares:

$$\underset{S}{\operatorname{argmin}} \sum_{i=1}^{K} \sum_{X_j \in S_i} ||X_j - c_i||^2, \tag{2.1}$$

where c_i is the centroid or center of S_i estimated by its mean value.

However, the adoption of K-means has been restricted, because it requires the number of populations to be pre-identified, it is sensitive to its initialization, and it is limited to modelling spherical cell populations. To estimate the number of

clusters, Pelleg et al. [91] and Hamerly et al. [50] extended basic K-means by using the Bayesian Information Criterion and a normality test, respectively. Voting-K-means [59] tries to achieve a good clustering by running the K-means algorithm with a number of different settings and combining the results using an ensemble clustering algorithm. However, the application of these algorithms for automated FCM data analysis has not been successful since the first two are sensitive to noise, and all three require user-defined parameter values [8, 127].

In this section, we present a new K-means-based clustering framework that addresses the initialization, shape limitation, and model-selection problems of K-means clustering, and can be applied to FCM data. We extended the flowMerge [41] approach by replacing the statistical model with a faster clustering algorithm. By introducing a new merging criterion, our approach finds non-convex cell populations, and we use a change point detection algorithm to estimate the number of clusters.

2.2 Materials and Methods

2.2.1 Initial Number of Clusters

The K-means clustering algorithm relies on users to define the number of clusters (K) to find. Using a predefined number of clusters for all FCM samples is not possible due to intersample variability. We solved this problem by automatically choosing K based on a reasonable maximum. The variants of the K-means algorithm discussed in the introduction try to estimate the exact number of clusters, and are not suitable for estimating the maximum number of clusters. Using the number of cells as the maximum is also not practical due to high runtime required for merging a large number of cells in FCM experiments (*e.g.*, commonly in the hundreds of thousands). Instead, we use the number of modes found individually in every eigenvector of the data. Using individual eigenvectors makes solving the mode-counting problem practical, but results in overlapping clusters (since some cell populations will be projected on more than one eigenvector and will be counted more than once). These overlapping clusters are later merged.

While many mode-detection algorithms are available, we used an approach

based on the work of Duong *et al*. [29] for mode detection using kernel density estimation, which has has previously proven to be successful on FCM data [84]. Formally, for a n vectors, $X = (X_1, X_2, ..., X_n)$, of length n sampled randomly from the density function f, the kernel density estimator $\hat{f}$ is defined to be the mean of n Gaussian kernel estimations:

$$\hat{f}(x) = \frac{\sum_{i=1}^{n} K(\frac{x-X_i}{h})}{n \cdot h}, \tag{2.2}$$

where h is the bandwidth selected using Scott's rule [111], and $K(\cdot)$ is the Gaussian kernel function:

$$K(x) = \frac{1}{\sqrt{2 \cdot \pi}} \cdot e^{-\frac{x^2}{\sqrt{2}}}. \tag{2.3}$$

The gradient of the estimator is:

$$\Delta\hat{f}(x) = \frac{2}{n \cdot h^2} \cdot \sum_{i=1}^{n} (x - X_i) \cdot K\left(\frac{x - X_i}{h}\right) \tag{2.4}$$

We then used a simultaneous significance test (based on Bonferroni's correction) to find the regions where the gradient is significantly different from zero [29]. Finally, the number of modes in the data is estimated by the number of times that the gradient changes from positive to negative for every one dimensional projection of the data on the eigenvectors. The K-means algorithm is then initialized with the total number of modes across all dimensions.

2.2.2 Merging

We solved the initialization problem at the cost of finding redundant clusters. To find the correct populations, these clusters must be merged. In addition, to capture non-spherical populations, we allow more than one cluster to model a single population (*i.e.*, nearby clusters are merged).

The merging procedure iterates between the following two steps until all of the points are merged to a single cluster: (1) calculate/update the distance between every pair of clusters; (2) identify and merge the closest pair of clusters.

Distance Metric

Given two populations $X = (x_1, x_2, ..., x_N)$ and $Y = (y_1, y_2, ..., y_M)$, we want to estimate the probability that the point (in this case, cell) y_i belongs to X. The closer y_i is to the center of X (*i.e.*, $\overline{X}$), the more likely it is to belong to X. However, the probability also depends on the dispersion of X. This can be estimated by the normalized Euclidean distance $\frac{\overline{X}-y_i}{S_X}$, where S_X is the sample standard deviation of X. In the multivariate case, the direction in which X is spread is also important, so the normalization term should be replaced by the covariance matrix. This results in a distance metric called the Mahalanobis distance. Formally, the Mahalanobis distance between X and y_i is defined as:

$$D(X, y_i) = \sqrt{(\overline{X} - y_i) \cdot S_X^{-1} \cdot (\overline{X} - y_i)^\top}, \tag{2.5}$$

where S_X is the covariance matrix of X.

Based on $D(x, y_i)$, we define a symmetric semi-metric (semi-distance) function between populations X and Y:

$$D(X, Y) = \min \begin{cases} \sqrt{(\overline{X} - \overline{Y}) \cdot S_X^{-1} \cdot (\overline{X} - \overline{Y})^\top} \\ \sqrt{(\overline{X} - \overline{Y}) \cdot S_Y^{-1} \cdot (\overline{X} - \overline{Y})^\top} \end{cases}. \tag{2.6}$$

Estimating the Number of Populations

As long as two clusters are overlapping (*i.e.*, model the same cell population), the distance between them will be very small, and these will be merged. After several merging steps, when the remaining clusters are well separated, the distance between the next clusters to be merged is significantly larger than the previous ones, indicating that these likely represent separate cell populations. We devised a segmented regression algorithm to detect the change point in the distance between the merged clusters. This algorithm divides the data to two subsets based on a given break point and fits a line to each of the subsets. The break point that minimizes the error of this model represents the number of clusters for which the clusters are well separated.

Formally, let N be the initial number of clusters, $i = (1, ..., N)$ the vector of

iteration numbers, $NC = N - i$ the vector of number of clusters at each iteration, and *Dist* the distance between the merged clusters at each iteration. The segmented regression model can be described with the following equation:

$$DistR_{(i,BP)} = \begin{cases} A_1 \cdot NC_{(i)} + B_1, \text{ if } NC_{(i)} < BP \\ A_2 \cdot NC_{(i)} + B_2, \text{ if } NC_{(i)} \geq BP \end{cases}, \tag{2.7}$$

where *DistR* is the vector of predicted values for *Dist*, *BP* is the break point at which we are expecting an abrupt change of the distance between clusters, and (A_1, B_1) and (A_2, B_2) are the slope and offset of the regression lines for the points before and after the break point, respectively. The least squares method must be applied separately to each segment to estimate the parameters of each line. Finally, the optimized break point BP_{opt} value that minimizes the sum of squared errors (SSE) can be found using exhaustive search over $BP \in \{2, 3, ..., \#Clusters - 1\}$:

$$BP_{opt} \in \underset{BP}{\operatorname{argmin}} \left(\sum_{i=1}^{N} \left(Dist_{(i)} - DistR_{(i,BP)} \right)^2 \right), \tag{2.8}$$

Figure 2.1 shows an example where the change point is in the solution with 6 clusters.

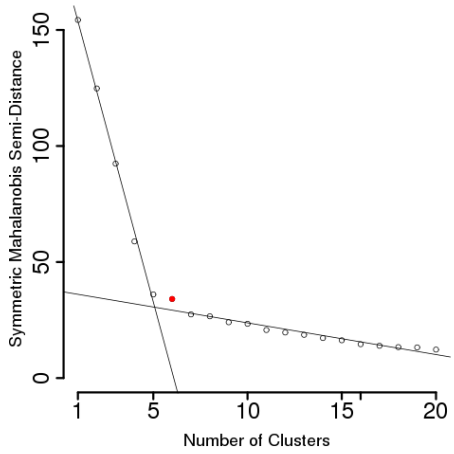


Figure 2.1: An example of finding the change point using segmented regression. The chosen solution (shown in red) consists of 6 populations.

2.2.3 Evaluation

We compared flowMeans to flowMerge and FLAME, the current state-of-the-art automated gating algorithms. acBIC and Scale-free Weighted Ratio (SWR) were used to determine the initial number of clusters for flowMerge and FLAME, respectively. The comparison was conducted using a computer running Ubuntu LTS 8.04 with a 3.2 GHz Intel Pentium CPU and 3 GB of RAM. For flowMeans and flowMerge, 10 random clustering solutions were used for initialization. To avoid model singularity issues caused by the data transformations, a small uniform noise was added to every event before the analysis by any of the algorithms. Convergence was determined using the default criteria of each software. flowMerge and FLAME both have optional free parameters that the user can use to adjust the behaviour of the algorithm (for example by specifying a threshold for the boundary events). We left these parameters at their default values to study the unsupervised performance of all three algorithms.

Our evaluation of the algorithms was based on comparison against manual analysis by human experts that was performed using a set of scatter plots of two dimensional projections of the data. While several metrics are available for comparison of clusterings [106], we used the F-measure, because it has proven to be successful for evaluation of the performance of automated gating algorithms [1]. Let n be the number of data points, C the set of membership labels assigned by the human expert, and K the set of membership labels calculated by the automated algorithm. The F-measure is formally defined as:

$$F(C,K) = \sum_{c_i \in C} \frac{|c_i|}{n} \cdot \max_{k_j \in K} F(c_i,k_j) \tag{2.9}$$

$$F(c_i,k_j) = \frac{2 \cdot R(c_i,k_j) \cdot P(c_i,k_j)}{R(c_i,k_j) + P(c_i,k_j)} \tag{2.10}$$

$$R(c_i,k_j) = \frac{n_{ij}}{|c_i|} \tag{2.11}$$

$$P(c_i,k_j) = \frac{n_{ij}}{|k_j|}, \tag{2.12}$$

where n_{ij} is the number of points with label $c_i \in C$ that are assigned to $k_j \in K$.

The points that the human expert had not included in the analysis (for example outliers or biologically irrelevant populations) were excluded before calculating the F-measure.

We measured the F-measure of every sample and reported the average as a single value representative of the distribution. While the average F-measure value helps to evaluate the overall performance of the algorithm across a dataset, it does not help in understanding how these algorithms differ in the analysis of individual samples. We therefore selected four cases where the F-measure values of one of these algorithms was significantly better than another algorithm for further visual illustration of the performance of each method.

Since FLAME's web-based interface does not provide CPU time measurement, all runtimes were measured as wall-clock time on our reference machine. However, we verified that for flowMeans, the difference between CPU time and wall-clock time never exceeded 200 milliseconds.

2.2.4 Datasets

We used two fully gated datasets to evaluate our approach:

Graft versus Host Disease (GVHD)

GvHD occurs after stem cell transplantation. This dataset is a collection 12 randomly selected peripheral blood samples (from 31 patients) analyzed for CD4, CD8b, CD3, and CD8 [15].

Diffuse Large B-Cell Lymphoma (DLBCL)

DLBCL is an aggressive lymphoma that can quickly spread to different parts of the body. Its diagnosis is usually performed via lymph node biopsy. The lymphoma dataset from the BC Cancer Agency consists of 30 randomly selected lymph node biopsies from patients seen between 2003 and 2008 [48]. These patients were histologically confirmed to have DLBCL. Cells were stained for three markers, CD3, CD19, and CD5.

Table 2.1: Comparison of F-measure of flowMeans, flowMerge, and FLAME.

Dataset	Mean F-measure (SD)			
	flowMeans Euclidean	flowMeans Mahalanobis	flowMerge	FLAME
GvHD	0.63(0.10)	0.84(0.07)	0.80(0.06)	0.68(0.13)
DLBCL	0.65(0.11)	0.92(0.04)	0.92(0.05)	0.59(0.14)

2.3 Results

Table 2.1 shows the average F-measure values for flowMerge, FLAME, flowMeans (using the symmetric Mahalanobis semi-distance function), and flowMeans-Euclidean (using an Euclidean distance function) against expert manual analysis. flowMeans and flowMerge performed similarly on both of the datasets, while FLAME had a lower F-measure. As can be seen from the Cumulative Distribution Function (CDF) plots shown in Figure 2.2, these averages are not distorted by the presence of outliers.

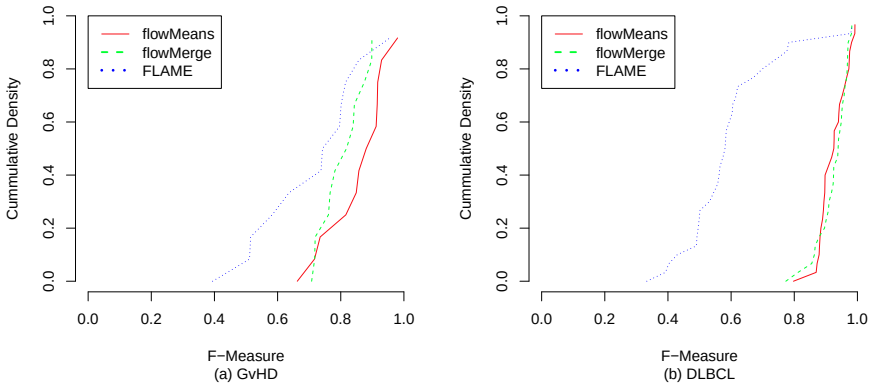


Figure 2.2: Cumulative Distribution of F-measure over different samples

Figure 2.3 shows the number of clusters identified by each of the algorithms and the manual analysis. For the GvHD dataset, the results obtained by flowMeans

are the closest to those from the manual analysis, followed by those from flowMerge. The number of clusters identified by FLAME are in a much larger interval. For the DLBCL dataset, again, the results obtained by flowMeans are the closest to those from the manual analysis, followed by those from flowMerge. The difference between the results of flowMerge and flowMeans is smaller in the DLBCL dataset. FLAME typically identifies a quite high number of clusters (10 on average).

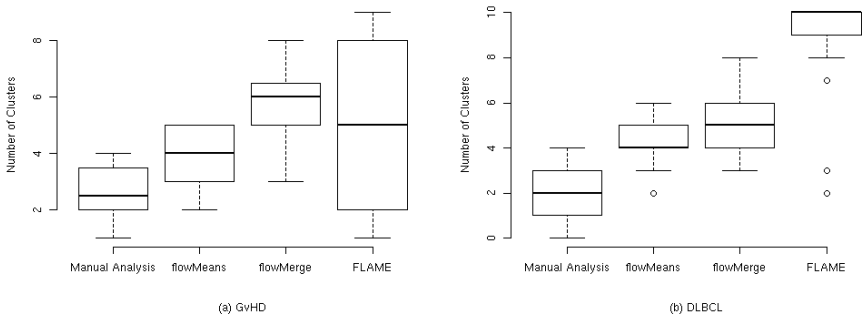


Figure 2.3: Boxplots of the number of clusters selected by manual analysis and the three algorithms for the (a) GvHD and (b) DLBCL datasets.

Table 2.2 shows that on average, the runtime of flowMeans was significantly lower than that of flowMerge and FLAME. We next examined whether this difference was due to the time requirement of the clustering method or the model-selection approach. Tables 2.1 and 2.2 show that while calculating the symmetric Mahalanobis semi-distance function increases the time requirement, replacing it with a simple Euclidean distance function decreases the accuracy of the identified populations to less than that obtained by the current state-of-the-art methods. Table 2.3 shows the runtime of the clustering algorithm used by each of these frameworks for identifying 10 clusters. This demonstrates that flowMeans' simpler clustering model is contributing to the lower runtime.Figure 2.4 shows the agreement between the F-measure of flowMeans and either flowMerge or FLAME. All F-measure values were in the interval $[0.5, 1]$ (shown in panels (a) and (b)), indicating that flowMerge and flowMeans perform similar to each other, even for outlier samples

in the correlation plots. The flagged sample in panel (c) shows the extreme case in which FLAME's performance might be closer to the manual gates than that of flowMeans. In this sample, flowMeans has identified an extra population, while FLAME has avoided this at the cost of not identifying one of the manually gated populations. Figure 2.4 panel (c) shows that the F-measure of these two algorithms is rather close while FLAME is slightly higher. However, in panel (d) (flowMeans' best case) FLAME did not perform equally well, since it found too many sub-populations.

Table 2.2: Comparison of Average Wall-Clock Runtime of flowMeans, flowMerge, and FLAME.

Dataset	Average Runtime (mm:ss)			
	flowMeans Euclidean	flowMeans Mahalanobis	flowMerge	FLAME
GvHD	00:17	00:28	15:34	18:41
DLBCL	00:13	00:21	11:40	15:35

The output of each algorithm for the four outlier samples (marked with red $X's$ in Figure 2.4) is shown in Figure 2.5.Panel (a) in Figure 2.5 shows the sample chosen in Figure 2.4 (a). In this sample, the performance of flowMerge is better than that of flowMeans, since flowMerge identified the four populations found by the human expert, while flowMeans found only three. Panel (b) of Figure 2.5 illustrates the two out of three biologically interesting populations found by flowMeans; we note that the remaining cluster is also missed by flowMerge, even though it identifies three additional populations. Similarly, panels (c) and (d) in Figure 2.5 show two other samples for which FLAME performed better than flowMeans and vice-versa.

2.4 Discussion

Model-based methods have proven to be successful in automating the FCM gating process [41]. However, the time-requirement of these methods represents a bottleneck in applying them to samples with millions of cells and tens of parameters. The application of simpler models to speed up the population

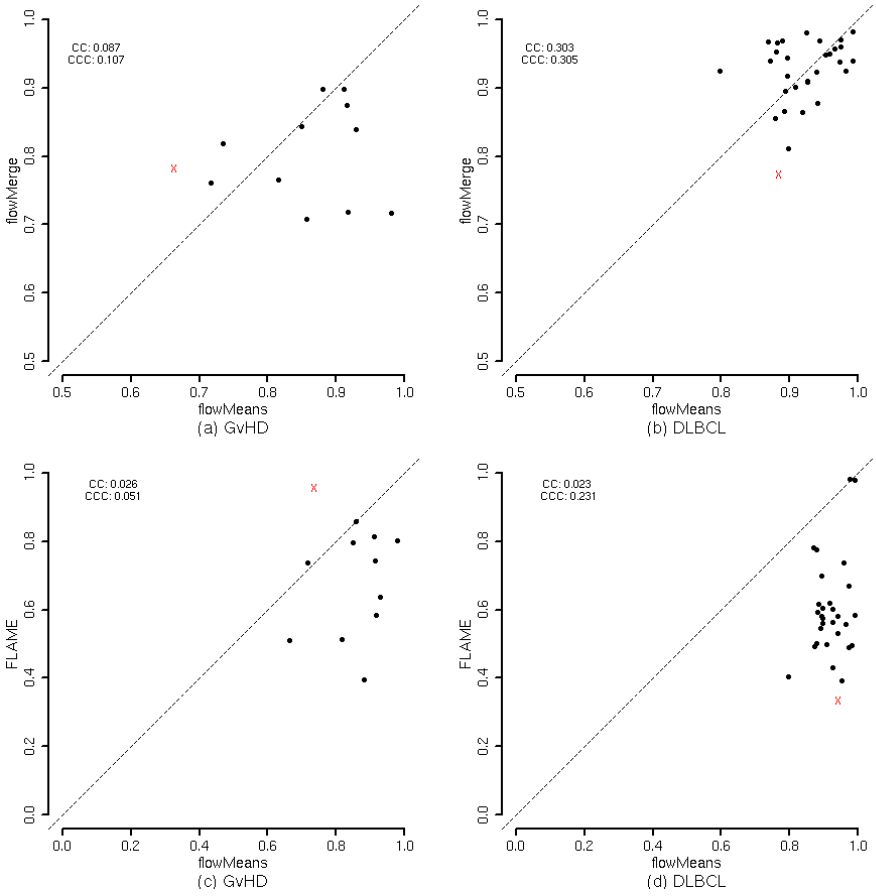


Figure 2.4: Agreement between F-measures of flowMeans and either flowMerge (a,b) or FLAME(c,d) on GvHD(a,c) and DLBCL(b,d) datasets. The cell populations for the samples indicated with red $X's$ in panels (a)-(d) are shown in respective panels in Figure 2.5. The correlation coefficient (CC) and concordance correlation coefficient (CCC) are shown as legends.

identification problem has not been successful, as these algorithms are limited by different factors (*e.g.*, reliance on user-defined parameters or specific shapes of populations). For example, while the K-means clustering algorithm (as a special case of GMMs with spherical variance constant across clusters) is quick compared

to other model based approaches, applying it to FCM data has not been successful, since it is limited to spherical cell populations and relies on pre-defined number of populations. A GMM can handle elliptical populations, but has a higher running time, since more iterations are required for fitting it to FCM data, which is generally quite noisy. *t* and *skew-t* mixture models are more flexible with respect to kurtosis and skewness, at the cost of further increasing the running time [127]. These models can use model selection criteria to estimate the number of populations; however, fitting multiple models compounds runtime requirements.

Since FCM cell populations are not elliptical, flowMerge allows more than one elliptical component to model the same population. We developed a similar framework to extend the K-means algorithm by merging the clusters that belong to the same population. Using the spherical model of the K-means algorithm, our framework has a significantly lower runtime compared to more flexible but computationally expensive statistical models (*e.g.*, a skew/t-mixture model). Improvements in processing time are an important consideration in high-throughput data production environments. Savings in runtime also increase as the number of measured parameters increases, as is the trend in FCM technology.

The use of more than one centroid to model the same population enabled our K-means based approach to find non-convex cell populations. However, the initial number of clusters needs to be determined before applying K-means. Choosing the correct number of clusters to initialize K-means is not critical, as long as the number selected is larger than the number of cell populations, since the extra (overlapping) clusters are later merged. We used the number of modes in the data (orthogonally projected on one-dimensional sub-spaces) as an upper bound for the number of clusters. Using one-dimensional projections of the data has the drawback of not finding populations that can only be identified in multiple dimensions. flowMeans addresses this problem, to some extent, by projecting the points on the eigenvectors (instead of individual markers) followed by multi-dimensional clustering. However, this can potentially be improved by designing a multi-dimensional procedure for finding a more accurate upper bound for the number of clusters. Regardless of the specific approach, an important advantage of flowMeans over the current model-based approaches is that it doesn't need to fit multiple models to estimate the correct number of clusters. This, along

with avoiding an expensive statistical model, resulted in a significantly improved running time (>20 times on average) compared to the current state-of-the-art model-based gating algorithms, without any decrease in accuracy.

We used the position and shape of clusters to identify candidate clusters for merging. Furthermore, We defined a symmetric Mahalanobis semi-distance function that takes the covariance of the clusters into account for calculating the distance between them. At every iteration of flowMeans, these Mahalanobis semi-distances need to be recalculated for the modified cluster. This recalculation procedure represents a bottleneck in the runtime of our framework. However, Tables 2.1 and 2.2 show that replacing it with an Euclidean distance function decreased the accuracy of the predicted populations. One possible approach to preserve accuracy and increase speed would be to use a covariance matrix updating procedure (see, e.g. [57]) to update the symmetric Mahalanobis semi-metric without recalculating it.

Our empirical evaluation was based on comparison against manual analysis. While a wide range of metrics are available for cluster evaluation, we used F-measure, because it has been shown to have a better performance in discriminating between the clustering solutions that are similar or different from the manual analysis [1]. The F-measure values show that flowMeans and flowMerge perform similarly, both on average and for individual samples.In spite of using a more flexible statistical model, FLAME usually has a lower F-measure. Figure 2.3 suggests that this might be due to the high number of populations that FLAME identifies. To further study the characteristics of these algorithms, we used the F-measure values to select four extreme case samples where the performance of the algorithms varies significantly for visual comparison. While visual comparison generally confirmed the F-measure values, it is important to note that due to the high dimensionality of the data, the margins of the populations could not be effectively visualized. Using human gates as the gold standard for comparison is also complicated, as human results can be subjective and highly variable [47, 108, 121]. For example, in Figure 2.5(d) it is not clear if the human has missed the green population found by flowMeans, has intentionally decided to merge it with the blue population, or has marked those cells as outliers. For cases similar to this, if a sample is critically important and the F-measure value alone cannot be

trusted, multi-dimensional visualization (*i.e.*, looking at different bi-variate plots as done in the back-gating procedure) can be used to check the margins using different dimensions. Visualizing cell populations in multiple dimensions remains an area for future improvement. This includes finding the dimensions (or combination of dimensions) that can effectively visualize the populations using feature selection and feature extraction strategies.

An implementation of flowMeans is publicly available as an R package through Bioconductor, a free, open source and open development software project for the analysis and comprehension of genomic data [45].

Table 2.3: Comparison of Average Runtime of the Clustering Algorithms used for each Framework for Identifying 10 Clusters.

Dataset	Average Runtime (mm:ss)			
	K-means (flowMeans)	Gaussian Mixture Model (flowMerge)	t Mixture Model (flowMerge)	skew-t Mixture Model (FLAME)
GvHD	00:07	04:26	05:37	07:36
DLBCL	00:05	03:31	04:07	05:51

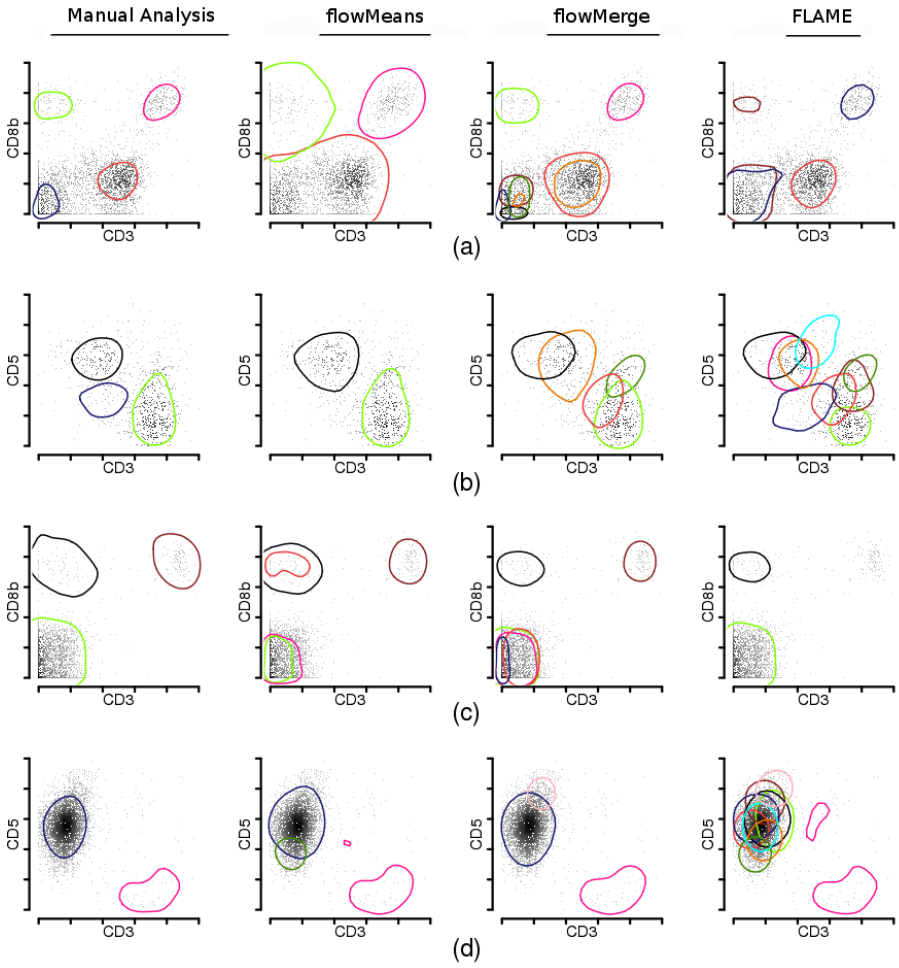


Figure 2.5: Panels (a)-(d) illustrate the cell populations found by flowMeans, flowMerge, and FLAME for the samples shown with red X's in respective panels in Figure 2.4. In this figure, the >90th percentiles of each cluster are visualized to make the boundaries more robust after projection to a two dimensional scatter plot. Therefore the populations might be different from the real distributions on the margins. The pink cluster in panel (d) is a multi-modal population with 2 high-density regions. In every panel, colors of each solution are matched with the solution with the maximum number of clusters.

Chapter 3

flowType: Immunophenotype Extraction for Flow Cytometry Data with Application to Identification of Immunologic Correlates of HIV Protection

3.1 Introduction

The immune response to infection, vaccination, or malignancy can be characterized by examining changes in the expression of a wide array of proteins expressed on leukocytes (either generally or on antigen-specific B- or T-cells). These proteins identify an enormous variety of cell types, and it is often not known which subsets of cells are clinically relevant. In some settings, the immunologically-relevant cell subset represents a small minority of the bulk cell population. Therefore, gross measurements taken from heterogeneous samples (as generally done with microarrays) may mask immunologically or clinically significant signals. This limitation can be overcome with Polychromatic Flow Cytometry (PFC) (>5 color), where protein expression can be assessed among a large number of cell subsets, at

the single cell level [24, 93].

The need for PFC is particularly apparent in studies of Human Immunodeficiency Virus (HIV), were the strongest cellular correlate of clinical outcome (CD4+ T-cell count) provides little help in identifying those individuals who would benefit from early initiation of Highly Active Anti-Retroviral Therapy (HAART) [16, 26, 60, 109]. Recent studies of Simian Immunodeficiency Virus (SIV) infection of nonhuman primates provide some guidance, demonstrating that the level of central memory T-cells may be a relevant predictor of the need for early therapy [65, 78, 122]. Similarly, a recent study of early HIV infection suggests the presence of long-lived T-cells during early infection correlates with long-term progression, as does the absence of proliferating cells [42]. Likewise, measurements of polyfunctional T-cells (simultaneously producing at least three of the following: IFNγ, IL2, CD107a, MIP1β and TNFα) are relevant in individuals whose disease progresses slowly [61, 123]. Importantly, enumeration of central memory, long-lived, proliferating, or polyfunctional cells requires PFC technology, since many markers are needed to discriminate each of these cell types from other populations of leukocytes.

Thus, it is evident that highly multiplexed approaches (such as PFC [11, 90]) are critical, at least as exploratory tools to identify potential correlates of pathogenesis; however, even though a PFC experiment collects data describing tens of thousands of cell subsets, only a small proportion of those can be reasonably queried against a clinical outcome. The choice of these subsets depends heavily on the investigator; therefore, important immunophenotypes that were not initially hypothesized to be important may be ignored [23]. Another challenge emerges when assessing the statistical rigor of findings from manual data analysis. Since the number of exploratory attempts at the analysis is rarely reported, adjustment for multiple comparisons is not usually performed. Multiple testing correction is complicated further when the choice of candidate cell populations for exploratory analysis is biased by the results of previous similar studies. A fourth challenge is the identification of the minimal set of markers that describe a clinically relevant cell type. Although thousands of immunophenotypes can be identified in a PFC experiment, it is not clear how many of these subsets represent functionally distinct cell populations. Moreover, for those cells that are clinically relevant, the exact

set of markers needed to identify that cell subset is rarely known. This is a particularly important problem, because it prevents the translation of results from PFC studies to more widespread use in clinical or resource-poor settings where complex instrumentation is often not available.

To address these problems, we developed a computational approach for identifying biomarkers in PFC data with clinical outcomes. Briefly, this approach first defines all possible immunophenotypes within a dataset and assesses the relationship between each and the clinical outcome. Importantly, the approach combines completely automated analysis of markers with some level of expert guidance to facilitate identification of rare subsets. Next, it reveals the minimal set of markers needed to identify the cell populations of interest. We demonstrate the utility of this approach by applying it to a dataset derived from a large retrospective study of individuals at the early stage of HIV infection. The dataset included a well-defined clinical outcome - time to Acquired Immunodeficiency Syndrome (AIDS) diagnosis or death, against which the frequency of each immunophenotype was correlated. We identified three groups of related T-cell subsets whose frequency during early infection had a statistically and clinically significant relationship with progression to AIDS. One of these groups was closely related to a cell population identified previously using standard manual approaches [42].

3.2 Materials and Methods

3.2.1 The Cohort

The HIV Natural History Study has collected clinical data on HIV-infected patients since 1985. Basic demographic characteristics of this dataset are described elsewhere [125]. We studied a subset of these subjects ($n = 466$) for which peripheral blood mononuclear cells (PBMCs) acquired within 18 months of seroconversion were available. The cohort included 135 death/AIDS events as defined by 1993 guidelines [19]. The date of the last follow-up or initiation of HAART was considered a censoring event. The immunologic and virologic characteristics of this subset were previously published [42].

3.2.2 Flow Cytometry Assays

Antibodies, staining procedures, and instrumentation were described previously [42]. Briefly, the staining panel enumerated various subsets of naïve and memory T-cells defined by CD3, CD4, CD8, CD45RO, CD27, CD28, CD57, CCR5, CCR7, CD127, and KI-67. CD14 and V-amine dye were used to exclude monocytes and dead cells, respectively. All study samples were treated the same way using methods common to the field (*i.e.*, gradient centrifugation of whole blood, isolation of PBMC, cryopreservation, and thawing). Therefore, the results presented are not confounded by sample manipulation, and are applicable to most of the settings in which HIV pathogenesis/vaccination studies are performed. On average $\approx 400\,000$ cells including $\approx 120\,000$ T-cells were measured (Figure 3.2).

3.2.3 Population Identification

Dead cells, doublets, and cellular debris were removed, and live T-cells were selected by manual gating as previously described [42]. The flowMeans algorithm was used for cell population identification within the T-cell compartment as described in Chapter 2. The software package, as well as the infrastructure for PFC data analysis [49] are available through Bioconductor [45]. flowMeans identified many clusters in the data and repeatedly merged adjacent ones based on the Mahalanobis distance between them until the desired number of clusters was reached. For each of the 10 markers in our data, flowMeans was used to identify a partition that divided the cells into a positive and a negative population (a movie demonstrating this partitioning is available online[1]). This was based on the assumption that the expression was either on or off (*i.e.,* there are two distinct cell populations). These 10 partitions could be combined in 2^{10} possible ways, resulting in 1024 cell populations. To allow exclusion of markers from subset identification (which later enabled us to identify the most clinically meaningful markers), each marker could be assigned a "neutral" value (*i.e.,* that marker was excluded from the clustering - see the Discussion section); thus, for any single subset, each marker could be negative, positive, or neutral (ignored). This increased the number of possible cell populations to 3^{10} (59 049). An example of all possible combinations

[1] http://www.youtube.com/watch?v=SDwub9PPN0Y

Table 3.1: Comparison of F-measure of flowMeans, flowMerge, and FLAME.

	Immunophenotype	p-value	p-value CI	Adjusted p-value	CPHR Coefficient	R^2	Cell Frequency
1	Ki-67$^+$CD127$^-$	2.7×10^{-08}	$(2.9\times10^{-15}, 2.1\times10^{-6})$	1×10^{-3}	19	0.069	0.01
2	CD45RO$^-$CD8$^+$CD57$^+$CCR5$^-$ CD27$^+$CCR7$^-$CD127$^-$	3.1×10^{-07}	$(1.5\times10^{-11}, 1.6\times10^{-2})$	1×10^{-2}	633	0.059	6×10^{-4}
3	CD28$^-$CD45RO$^+$CD57$^-$	5.6×10^{-7}	$(1.1\times10^{-11}, 2.6\times10^{-4})$	2e-02	12	0.056	5×10^{-2}

of gates (partitions) for two markers is shown in Figure 3.1(a). Notably, the Ki-67$^+$ population was rare ($< 5\%$ of the total number of cells), and could not be identified by flowMeans. Therefore, for this marker, historical negative controls provided a static gate to partition the cells. The appropriateness of gate was confirmed manually, by visual inspection of each participant's data.

3.2.4 Predictive Analysis

To measure the predictive power of each immunophenotype, a CPHR was used to calculate the correlation between the measured phenotypes' cell frequencies (the number of cells in that immunophenotype divided by the total number of T-cells) and the clinical outcome (survival time) [14]. Next, the immunophenotypes with a statistically significant correlation to the survival time were identified by the logrank test [52], after multiple testing correction using the Bonferroni method.

The sensitivity of the predictive power (measured by coefficient of determination (R^2) as the effect size of the logrank test) was determined using a bootstrapping procedure that tested the phenotypes of different subsets of the cohort [53]. Specifically, for a given vector S of subjects, a 95% Confidence Interval (CI) for the effect size can be calculated using the following procedure:

(1) Repeat for 10^4 times: from S, draw a uniform random sample of size $|S|$ with replacement, fit the CPHR model and record R^2.

(2) Report the $2.5th$ and $97.5th$ percentiles of the distribution of R^2 values from Step 1 as the lower and upper bounds of the CI, respectively.

Thus, if an immunophenotype was measured over 10^4 subsets of the cohort and every subject's probability of selection (as defined in Equation 3.1) $P_{selection} = 0.63$,

then in 95% of the trials the R^2 (and therefore the p-value) would have been within the range of the CI.

$$P_{selection} = 1 - \left(\frac{|S|-1}{|S|}\right)^{|S|} \approx 1 - \frac{1}{e} \approx 0.63 \tag{3.1}$$

3.2.5 Phenotype Extraction

Many of the cell populations identified were subsets of others (*e.g.,* $CD28^+CD45RO^-$ cells are also $CD28^+$), and therefore could be redundant. We used an approach known as complete linkage hierarchical clustering to find homogeneous groups of immunophenotypes that are similar to each other [35]. Let f_i, $i \in \{1,2,...,59049\}$ be the vector of cell frequencies across all subjects for immunophenotypes. For the hierarchical clustering, we used the distance function $dist_{i,j} = cor(f_i, f_j)$, where i and j are immunophenotype numbers, and cor is the Pearson correlation coefficient. The output of this procedure consists of several groups of immunophenotypes; however, the immunophenotypes in each group were highly correlated and likely to be subsets of the same parent cell type. Therefore, two additional steps were employed to identify the cell populations underlying these overlapping immunophenotypes.

Marker Selection

This step was designed to identify the markers that had a positive impact on the predictive power of a group of immunophenotypes. To investigate this, we let the *impact* of a marker be the absolute difference between a) the means of CPHR R^2 goodness-of-fit scores for the given groups of immunophenotypes and b) the scores after forcing that marker to be neutral. The impact value reflected the increase in the error of the CPHR model when that specific marker was excluded. To identify the markers with impacts significantly higher than zero, the same bootstrapping procedure described in the predictive analysis section was applied to given groups of immunophenotypes (see Figure 3.3). Combining these markers identified the candidate cell population representative of the immunophenotypes in the respective group.

Backward Marker Elimination

In the previous step, we selected the markers that, on average, had a positive impact on the predictions of the respective groups of immunophenotypes. The next step was to identify the markers that were redundant (*i.e.*, were uninformative in presence of others). For each immunophenotype, we sequentially removed markers starting with the one with lowest impact. At every step, the p-value of the logrank test was calculated and evaluated (false discovery date $= 0.05$ after adjustment). The last statistically significant cell population was selected. This cell population could define the immunophenotypes in the respective group with a minimum number of markers.

3.2.6 Sensitivity Analysis

The pipeline is an exploratory analysis tool that outputs a list of immunophenotypes (and not a multivariate predictive model). Therefore, cross-validation or holdout-validation (*i.e.* keeping a test-set) are not meaningfully applicable. Instead, we used the following bootstrapping procedure to assess the generalizability of the selected immunophenotypes to previously unseen data:

(1) Repeat for K times: from the given set of subjects, S, draw a uniform random sample of size $|S|$ with replacement, run the pipeline and record the selected immunophenotypes;

(2) Report the proportion of iterations in step (1) in which each immunophenotype was selected,

where K is the number of iterations, set manually by considering the amount of variation in the data and the computing resources available. To measure the sensitivity of the pipeline to different subsets of the cohort, this procedure detemines the proportion of trials on subsets of the subjects in which a given immunophenotype was selected by the pipeline. Like the previous bootstrapping step, it can be shown that the probability of every sample being included in the subset is 0.63. Therefore, phenotypes that are selected in a high proportion of trials (with different subject compositions of 37% on average) are not sensitive to variations within the cohort of subjects.

3.3 Results

3.3.1 Identification of Cell Subsets Related to Clinical Outcome

Cell populations were identified (as described in the Methods section) and the frequencies of the 59 049 immunophenotypes were calculated (Figure 3.1(a)). Next, these immunophenotypes were related to each patient's time to AIDS/death by CPHR analysis (Figure 3.1(b)). 101 of these immunophenotypes were revealed as candidate correlates of HIV disease progression by the predictive model; these were analyzed in two ways. First, we examined the correlations between cell frequencies using a clustered heat map, shown in Figure 3.1(c). The "correct" number of clusters (as in any other clustering algorithm) is subjective; our choice to use three groups is justified later in this section. Second, all 101 immunophenotypes were listed, using the order determined by the heatmap clustering (see Table 3.2). To make it easier to observe patterns among the immunophenotypes represented, the immunophenotype names are illustrated with a heat map in Figure 3.4. The dendrogram and the side-bar are identical to Figure 3.1(c). The immunophenotype names in Figure 3.4 are consistent with the clusters of immunophenotypes identified in Figure 3.1(c) based on correlation between cell frequencies. These figures show that closely correlated immunophenotypes have similar combinations of markers. This process allowed us to define the immunophenotypes that exhibited high correlation (*i.e.*, describe almost identical cell types).

Table 3.2: Statistically significant immunophenotypic correlates of survival of HIV$^+$ subjects are predicted by flowType. The p-values of the log rank tests, 95% confidence intervals calculated using bootstrapping, adjusted p-values using Bonferroni's method, coefficients and R^2s of the Cox proportional hazards regression models, and the frequency of the cells are provided as columns of the table.

#	Phenotype	p-value	p-value, CI	adjusted p-value	CPHR Coefficient	R^2	Cell Frequency
1	CD28-CD45RO+CD57-CCR5+	5.3e-07	(4.3e-14, 1.3e-02)	2e-02	20.5	0.056	0.03048
2	CD28-CD8+CD57-CD127-	2.5e-07	(2.3e-14, 3.8e-04)	1e-02	12.3	0.060	0.05975
3	CD28-CD45RO+CD57-CCR7-	5.1e-07	(2.3e-14, 6.1e-04)	2e-02	15.7	0.057	0.03829
4	CD28-CD45RO+CD4-CD57-	3.5e-07	(2.3e-14, 1.1e-03)	1e-02	13.2	0.058	0.04357
5	CD45RO+CD4-CD57-CD127-	2.7e-07	(1.2e-13, 7.1e-03)	1e-02	12.8	0.059	0.05062
6	CD28-CD45RO+CD57-CD127-	4.7e-08	(1.7e-14, 6.8e-04)	2e-03	16.0	0.067	0.03732
7	CD45RO+CD4-CD27-CD127-	4.4e-07	(5.8e-14, 1.1e-03)	2e-02	14.3	0.057	0.04830
8	CD28-CD45RO+CD57-	5.6e-07	(4.4e-14, 4.1e-04)	2e-02	12.4	0.056	0.05015

Table 3.2: Statistically significant immunophenotypic correlates of survival of HIV$^+$ subjects are predicted by flowType. The p-values of the log rank tests, 95% confidence intervals calculated using bootstrapping, adjusted p-values using Bonferroni's method, coefficients and R^2s of the Cox proportional hazards regression models, and the frequency of the cells are provided as columns of the table.

#	Phenotype	p-value	p-value, CI	adjusted p-value	CPHR Coefficient	R^2	Cell Frequency
9	CD45RO+CD4-CD127-	6.5e-07	(4.7e-15, 2.9e-03)	2e-02	9.6	0.056	0.07176
10	CD28-CD45RO+CD4-CD127-	3.1e-07	(0.0e+00, 5.7e-03)	1e-02	11.7	0.059	0.05300
11	CD28-CD45RO+CD57-CCR5+CD27-CCR7+CD127-	4.7e-07	(5.7e-14, 7.7e-03)	2e-02	171.4	0.057	0.00315
12	CD28-CD45RO+CD4-CD57-CCR5+CD27-CCR7+CD127-	4.5e-07	(1.8e-13, 3.9e-04)	2e-02	176.2	0.057	0.00294
13	CD28-CD57-CD127-	3.3e-07	(3.4e-15, 8.0e-03)	1e-02	8.0	0.058	0.12341
14	CD28-CD4-CD57-	8.8e-07	(2.2e-15, 2.9e-03)	3e-02	7.2	0.054	0.15525
15	CD57-CD27-CD127-	6.2e-08	(2.4e-14, 4.7e-03)	2e-03	9.5	0.065	0.12173
16	CD4-CD57-CD27-CD127-	4.7e-08	(4.2e-14, 3.3e-03)	2e-03	9.7	0.067	0.09721
17	CD28-CD57-CCR7-CD127-	2.8e-07	(9.7e-15, 1.0e-02)	1e-02	9.8	0.059	0.08417
18	CD28-CD4-CD57-CD127-	3.3e-08	(2.0e-12, 5.7e-04)	1e-03	9.1	0.068	0.10852
19	CD4-CD57-CCR7-CD127-	6.5e-07	(3.8e-15, 2.3e-03)	2e-02	8.8	0.056	0.09501
20	CD45RO-CD4-CD57+CCR5-CD27+CCR7-CD127-	6.1e-07	(1.2e-12, 2.6e-03)	2e-02	498.4	0.056	0.00097
21	CD28-CD45RO-CD4-CD57+CCR5-CD27+CCR7-CD127-	2.5e-07	(0.0e+00, 7.7e-03)	1e-02	561.2	0.060	0.00074
22	CD45RO-CD8+CD57+CCR5-CD27+CCR7-CD127-	1.2e-07	(4.6e-14, 3.3e-04)	5e-03	638.6	0.063	0.00068
23	CD45RO-CD8+CD4-CD57+CCR5-CD27+CCR7-CD127-	1.2e-07	(5.1e-14, 2.0e-03)	5e-03	638.6	0.063	0.00068
24	CD28-CD45RO-CD4-CD57+CCR5-CD27+CD127-	5.7e-07	(1.1e-13, 2.3e-03)	2e-02	298.3	0.056	0.00099
25	KI-67+CD28-CCR5+	1.0e-11	(2.9e-13, 2.8e-03)	4e-07	96.1	0.101	0.00547
26	KI-67+CD28-CCR5+CD27-	8.7e-12	(1.5e-14, 8.9e-04)	3e-07	115.3	0.102	0.00453
27	KI-67+CCR5+	1.3e-11	(2.4e-14, 7.0e-03)	5e-07	53.4	0.100	0.01192
28	KI-67+CD28+CD45RO+CD57-CCR7-CD127-	4.2e-09	(5.6e-16, 3.0e-03)	2e-04	241.3	0.077	0.00209
29	KI-67+CD45RO-CD4-CD27-CCR7-CD127-	1.2e-09	(2.0e-14, 4.4e-03)	4e-05	161.9	0.082	0.00297
30	KI-67+CD28-CD45RO-CD8-CD4-	5.0e-09	(2.9e-12, 1.7e-03)	2e-04	176.0	0.076	0.00225
31	KI-67+CD8-CD4-	8.1e-09	(6.1e-13, 4.5e-02)	3e-04	58.1	0.074	0.00738
32	KI-67+CCR5+CD27-CCR7-	2.0e-11	(3.8e-14, 6.0e-04)	8e-07	109.8	0.099	0.00532
33	KI-67+CD8-CCR5+CCR7-	1.3e-10	(3.1e-13, 2.0e-03)	5e-06	147.3	0.091	0.00392
34	KI-67+CD28-CD8-CCR5+CCR7+CD127-	2.6e-09	(1.6e-14, 1.1e-02)	1e-04	625.8	0.079	0.00061
35	KI-67+CD28+CD45RO+CD8+CD57-CD27+CCR7+	6.7e-07	(3.8e-13, 1.5e-03)	3e-02	585.4	0.055	0.00051
36	KI-67+CD28+CD45RO+CD8+CD4-CD57-CD27+CCR7+	6.7e-07	(1.1e-16, 4.7e-03)	3e-02	585.4	0.055	0.00051
37	KI-67+CD8+CD27-CCR7-CD127-	4.7e-11	(1.3e-13, 1.4e-03)	2e-06	141.3	0.095	0.00292
38	KI-67+CD8+CD4-CD27-CCR7-CD127-	4.7e-11	(1.3e-13, 1.3e-03)	2e-06	141.3	0.095	0.00292
39	KI-67+CD28-CD8+CD27-CCR7-CD127-	2.7e-11	(1.0e-13, 7.6e-04)	1e-06	164.5	0.097	0.00241
40	KI-67+CD28-CD8+CD4-CD27-CCR7-CD127-	2.7e-11	(2.7e-13, 1.4e-03)	1e-06	164.5	0.097	0.00241
41	KI-67+CD28-CD8+CCR7-CD127-	6.6e-11	(5.6e-14, 1.5e-02)	3e-06	132.9	0.094	0.00293
42	KI-67+CD28-CD8+CD4-CCR7-CD127-	6.6e-11	(1.2e-14, 8.4e-04)	3e-06	132.9	0.094	0.00293
43	KI-67+CD45RO+CD8+CD27-CCR7-	1.2e-09	(4.0e-12, 2.8e-03)	5e-05	143.6	0.082	0.00216
44	KI-67+CD45RO+CD8+CD4-CD27-CCR7-	1.2e-09	(1.0e-12, 1.2e-02)	5e-05	143.6	0.082	0.00216
45	KI-67+CD28-CD45RO+CD8+CD27-CCR7-	1.0e-09	(1.9e-15, 7.3e-04)	4e-05	188.5	0.082	0.00155
46	KI-67+CD28-CD45RO+CD8+CD4-CD27-CCR7-	1.0e-09	(1.7e-13, 2.0e-03)	4e-05	188.5	0.082	0.00155
47	KI-67+CD45RO+CD8+CD27-CD127-	7.1e-10	(1.2e-14, 6.8e-03)	3e-05	152.4	0.084	0.00221
48	KI-67+CD45RO+CD8+CD4-CD27-CD127-	7.1e-10	(3.4e-14, 1.5e-03)	3e-05	152.4	0.084	0.00221
49	KI-67+CD28-CD45RO+CD8+CD27-CD127-	5.0e-10	(6.0e-13, 3.1e-03)	2e-05	201.3	0.085	0.00163
50	KI-67+CD28-CD45RO+CD8+CD4-CD27-CD127-	5.0e-10	(4.6e-14, 2.7e-03)	2e-05	201.3	0.085	0.00163
51	KI-67+CD28-CD45RO+CD8+CD127-	1.0e-09	(1.2e-15, 3.2e-03)	4e-05	150.5	0.083	0.00222
52	KI-67+CD28-CD45RO+CD8+CD4-CD127-	1.0e-09	(1.5e-11, 3.6e-03)	4e-05	150.5	0.083	0.00222
53	KI-67+CD45RO+CD8+CD4-CD127-	2.2e-09	(2.8e-13, 2.1e-03)	9e-05	99.8	0.079	0.00362
54	KI-67+CD28-CD45RO+CD8+CD4-CCR7-	8.0e-09	(2.7e-12, 7.2e-04)	3e-04	133.6	0.074	0.00209
55	KI-67+CD28-CD45RO+CD57-CCR7+CD127-	5.9e-08	(4.0e-15, 4.5e-03)	2e-03	376.6	0.066	0.00075
56	KI-67+CD28-CD45RO+CD4-CD57-CCR7+CD127-	5.0e-08	(4.8e-13, 3.9e-03)	2e-03	409.6	0.066	0.00070
57	KI-67+CD57-CD27-CD127-	5.9e-10	(3.2e-14, 2.7e-03)	2e-05	44.9	0.085	0.00806
58	KI-67+CD28-CD27-CD127-	4.8e-10	(7.3e-15, 2.5e-03)	2e-05	50.6	0.086	0.00711
59	KI-67+CD4-CD127-	1.3e-10	(4.4e-16, 9.7e-03)	5e-06	37.1	0.091	0.01159
60	KI-67+CD28-CD127-	4.9e-10	(1.1e-12, 1.4e-03)	2e-05	41.4	0.086	0.00823
61	KI-67+CD4-CD27-	5.6e-09	(2.1e-14, 2.6e-03)	2e-04	28.6	0.075	0.01122
62	KI-67+CD28-CD4-CD27-	1.8e-09	(3.6e-13, 5.3e-03)	7e-05	40.2	0.080	0.00785
63	KI-67+CD27-CD127-	1.3e-09	(9.8e-15, 1.1e-03)	5e-05	33.0	0.082	0.01052
64	KI-67+CCR7-CD127-	6.5e-11	(1.4e-15, 9.6e-04)	2e-06	47.3	0.094	0.00947
65	KI-67+CD4-CD27-CCR7-	9.6e-11	(1.1e-16, 1.5e-03)	4e-06	52.1	0.092	0.00764
66	KI-67+CD4-CCR7-	1.7e-10	(3.0e-14, 1.0e-02)	7e-06	41.4	0.090	0.00987
67	KI-67+CD45RO+CD57-CCR7-	1.4e-09	(6.6e-13, 1.2e-03)	5e-05	49.6	0.081	0.00695
68	KI-67+CD45RO+CD57-CD27-CCR7-	9.1e-10	(8.6e-12, 2.5e-03)	3e-05	66.4	0.083	0.00505
69	KI-67+CD45RO+CD4-	2.0e-09	(8.0e-13, 2.5e-03)	8e-05	45.3	0.080	0.00851
70	KI-67+CD28-CD45RO+	1.3e-08	(1.2e-12, 2.4e-03)	5e-04	54.9	0.072	0.00525
71	KI-67+CD45RO+CD127-	1.1e-09	(4.4e-16, 1.5e-02)	4e-05	42.5	0.082	0.00834
72	KI-67+CD45RO+CD57-CD127-	2.9e-10	(1.5e-14, 6.4e-04)	1e-05	55.0	0.088	0.00719
73	KI-67+CD28-CD45RO+CD8+CD27-	9.2e-09	(2.6e-15, 2.3e-03)	4e-04	138.0	0.073	0.00201
74	KI-67+CD28-CD45RO+CD8+CD4-CD27-	9.2e-09	(1.0e-15, 4.6e-03)	4e-04	138.0	0.073	0.00201
75	KI-67+CD8+CD4-CD57-CD27-CD127-	1.9e-09	(5.9e-14, 7.0e-03)	7e-05	113.8	0.080	0.00274
76	KI-67+CD28-CD45RO+CD8+	9.3e-09	(5.9e-13, 1.4e-03)	4e-04	102.7	0.073	0.00279
77	KI-67+CD28-CD45RO+CD8+CD4-	9.3e-09	(0.0e+00, 1.6e-03)	4e-04	102.7	0.073	0.00279
78	KI-67+CD45RO+CD8+	2.1e-08	(6.9e-15, 6.8e-04)	8e-04	59.1	0.070	0.00512
79	KI-67+CD8+CCR7-	3.0e-08	(7.7e-13, 2.8e-03)	1e-03	49.5	0.068	0.00530
80	KI-67+CD8+CD27-CCR7-	8.3e-09	(1.0e-13, 3.6e-03)	3e-04	70.7	0.074	0.00377
81	KI-67+CD4-	2.8e-08	(1.0e-13, 2.3e-03)	1e-03	17.1	0.069	0.01627

Table 3.2: Statistically significant immunophenotypic correlates of survival of HIV$^+$ subjects are predicted by flowType. The p-values of the log rank tests, 95% confidence intervals calculated using bootstrapping, adjusted p-values using Bonferroni's method, coefficients and R^2s of the Cox proportional hazards regression models, and the frequency of the cells are provided as columns of the table.

#	Phenotype	p-value	p-value, CI	adjusted p-value	CPHR Coefficient	R^2	Cell Frequency
82	KI-67+CD28-CD4-	1.1e-08	(5.9e-14, 4.0e-03)	4e-04	26.7	0.073	0.00950
83	KI-67+CD127-	2.7e-08	(1.2e-12, 2.1e-03)	1e-03	19.1	0.069	0.01460
84	KI-67+CCR7-	8.4e-08	(3.4e-15, 2.3e-03)	3e-03	18.3	0.064	0.01311
85	KI-67+CD27-CCR7-	3.5e-08	(1.7e-13, 1.2e-03)	1e-03	25.2	0.068	0.00998
86	KI-67+CD45RO+CD27-	7.5e-07	(5.4e-13, 1.8e-03)	3e-02	24.0	0.055	0.00862
87	KI-67+CD45RO+CD57-	1.2e-07	(2.1e-13, 3.1e-03)	5e-03	22.9	0.062	0.01123
88	KI-67+CD4-CD57-	1.3e-08	(3.8e-15, 2.1e-03)	5e-04	25.3	0.072	0.01209
89	KI-67+CD28-CD4-CD57-	9.7e-09	(5.5e-12, 1.2e-03)	4e-04	37.7	0.073	0.00698
90	KI-67+CD57-CD127-	3.3e-09	(1.3e-13, 3.3e-03)	1e-04	28.1	0.078	0.01128
91	KI-67+CD45RO+CCR7-	4.2e-09	(7.8e-15, 2.5e-03)	2e-04	37.5	0.077	0.00819
92	KI-67+CD57-CCR7-	2.7e-08	(2.8e-13, 2.8e-03)	1e-03	26.6	0.069	0.01008
93	KI-67+CD57-CD27-CCR7-	1.2e-08	(4.9e-13, 2.6e-03)	5e-04	36.8	0.072	0.00762
94	KI-67+CD28-CCR7-	3.3e-09	(4.6e-14, 5.7e-03)	1e-04	37.7	0.078	0.00739
95	KI-67+CD28-CD27-CCR7-	3.3e-09	(2.6e-14, 6.5e-04)	1e-04	43.0	0.078	0.00647
96	KI-67+CD28-	1.9e-07	(4.0e-15, 2.7e-03)	7e-03	18.3	0.061	0.01053
97	KI-67+CD28-CD27-	7.1e-08	(1.5e-12, 8.6e-04)	3e-03	26.3	0.065	0.00874
98	KI-67+CD28-CD8-	8.3e-08	(5.5e-14, 2.5e-03)	3e-03	44.2	0.064	0.00523
99	KI-67+CD45RO+	8.9e-07	(1.9e-13, 2.5e-03)	3e-02	15.4	0.054	0.01343
100	KI-67+CD8+CD57-	1.1e-06	(4.4e-14, 3.1e-03)	4e-02	28.3	0.053	0.00648
101	KI-67+CD8+CD27-	6.4e-07	(2.3e-14, 1.1e-02)	2e-02	35.2	0.056	0.00560

Next, we identified the minimum set of markers necessary to describe each of the three groups of immunophenotypes. This helped define the clinically relevant cells using the simplest possible immunophenotype, which described the most general cell population of those measured. As described in the previous section, this process was carried out in two steps: 1) selection of the markers with a positive impact on the predictive power; 2) elimination of the redundant markers.

3.3.2 Impact of Individual Markers

For each immunophenotype group, we selected the markers that had a positive impact on the immunophenotype, as measured by the changes in mean effect size (Figure 3.1(d)). 95% confidence intervals were calculated using bootstrapping (over the patient cohort). Thus, for the three groups of immunophenotypes, the predictive power depended on the combination of different markers included in the measurements (Figure 3.1(d)). It is important to note that the impact value depends on the effect-size (R^2) of the original immunophenotypes in a given group. Different immunophenotype groups had different mean R^2 (and p-values);

therefore, impact values cannot be compared across multiple groups.

We used the impact value to confirm that the heat map clustered by frequency described three groups (and not two or four; Figures 3.5 and 3.6). With only two groups, a mix of positive and negative labels was observed, suggesting that the groups consisted of heterogeneous subpopulations. When the impact values for four groups were analyzed, two had very similar marker impacts, suggesting that we had bisected a single homogeneous cell population into two populations artificially. Finally, those markers with impacts significantly higher than zero, as indicated by the confidence intervals (Table 3.3), were selected as representatives of each phenotypic group, in order to define the most clinically relevant immunophenotype. By selecting markers that, on average, had a positive impact on the predictions of the respective groups of immunophenotypes, we narrowed down the list of potential immunophenotypes to three (Table 3.3).

Table 3.3: The representative immunophenotypes. The markers within Figure 1(d) with a positive impact on the predictive power were combined to form these immunophenotypes.

	Immunophenotype	p-value	p-value CI	Adjusted p-value	CPHR Coefficient	R^2	Cell Frequency
1	Ki-67^+CD4^-CCR5^+CD127^-	1.7×10^{-10}	$(0, 1.0\times10^{-5})$	6.5×10^{-6}	78	0.090	0.00704
2	CD45RO^-CD8^+CD4^-CD57^+ CCR5^-CD27^+CCR7^-CD127^-	1.2×10^{-7}	$(0, 7.7\times10^{-5})$	4.6×10^{-3}	639	0.063	0.00068
3	CD28^-CD45RO^+CD4^-CD57^- CD27^-CD127^-	6.5×10^{-8}	$(2.2\times10^{-16}, 1.9\times10^{-5})$	2.4×10^{-3}	22	0.065	0.02456

Marker Elimination

Next, we identified the markers that were uninformative in the presence of others. For each of the immunophenotype groups, we removed the markers one at a time, starting with the one with lowest impact, until only the marker with the highest impact remained. Figure 3.1(e) lists the p-values after every removal step. The first phenotypic group was originally described as Ki-67^+CD4^-CCR5^+CD127^- (Panel (a)). However, the iterative removal of markers only affected the p-value when CD4 and CCR5 were removed from the analysis, indicating that the relationship to disease progression in this immunophenotype is driven by Ki-67 and CD127. For the second phenotypic group, the p-value remains significant for a combination of eight markers (CD45RO^-CD8^+CD4^-CD57^+ CCR5^-CD27^+CCR7^-CD127^-). Finally, the representative immunophenotype of the third group was simpli-

fied from $CD28^{-}CD45RO^{+}CD4^{-}CD57^{-}$ $CD27^{-}CD127^{-}$ to $CD28^{-}CD45RO^{+}$ $CD57^{-}$. The most frequent cell population with a p-value higher than the threshold determined by multiple comparisons adjustment (*i.e.,* the statistically significant immunophenotype with minimum number of markers) was reported as the representative immunophenotype of the respective group (Table 3.1).

3.3.3 Confirmatory Analysis

We performed several experiments to confirm the results obtained by the pipeline. We manually identified $CD28^{-}CD45RO^{+}CD57^{-}$ cells using conventional methods (polygon gates on two scatter plots as demonstrated in Figure 3.7) and confirmed the relationship between frequencies of these cells and survival time ($p = 7 \times 10^{-6}$). This result is similar to that obtained with the automated pipeline ($p = 5 \times 10^{-7}$); any difference is likely due to minor variations in the data that cannot be captured using the manual analysis. A second confirmatory analysis was performed by using the three identified immunophenotypes to partition the patients into two groups by thresholding the cell frequencies; these groups had different survival patterns (Figure 3.1(f)), confirming the ability of the automated pipeline to identify clinically meaningful cell populations. Finally, the sensitivity of the automated pipeline was determined based on 100 bootstrap iterations, which required nearly 2000 CPU days. The immunophenotypes selected in the first and third groups were clearly dominant as demonstrated in Figure 3.7 panels (d), (e), and (f). However, the second phenotypic group could be labelled $CD4^{-}$ or $CD8^{+}$, according to this analysis. Importantly, these populations likely overlap significantly, as expression of CD4 and CD8 are usually mutually exclusive on T-cells in the peripheral blood. Thus, the $CD4^{-}$ label includes primarily $CD8^{+}$ T-cells [62].

3.4 Discussion

We described a computational approach to analyze a high dimensional clinical flow cytometry dataset that was previously investigated through laborious manual inspection. The findings from our analysis both replicate and extend the original analysis by human experts, revealing the T cell subsets and markers most highly correlated with HIV progression. The pipeline consists of five steps: 1) automated

identification of positive and negative populations for each marker, 2) quantification of subsets defined by every combination of markers, 3) identification of those cell subsets whose frequency is most highly associated with clinical outcome, 4) calculation of the impact of each individual marker, and 5) identification of the minimal set of markers needed to describe significant cell populations.

The first step in the pipeline delineates positive and negative populations for every channel. This step uses a clustering tool that was developed exclusively for PFC data [2]. Many such tools have been developed for identifying cell populations in a multidimensional setting, but several limitations have kept these algorithms from replacing manual analysis. Firstly, the use of these algorithms (as any other clustering tool) is highly subjective and complicated – often, the concept of what comprises a cluster/cell population is not well-defined. Clustering tools are also limited in their ability to find rare cell populations. Furthermore, meta-clustering of candidate clusters must be performed to identify clinically relevant immunophenotypes; however, for this, clusters must be linked to subjectively-defined categories of cells. It is also difficult to visualize and interpret results because clusters cannot be described using marker names. Lastly, biological information is rarely incorporated into the clustering process. The algorithm presented here overcomes these limitations by partitioning cells one marker at a time and by using combinations of the partitions to extract immunophenotypes/features for predictive analysis.

A potential shortcoming of this approach is the underlying assumption that every channel has only two well-separated cell populations (*i.e.*, expression is either on or off). However, some cellular proteins exhibit a continuum of expression across a cell population, with cells that lack expression, others with low levels of expression, and some with very high levels of expression. Furthermore, for some markers these differences are known to be biologically meaningful; CCR7 expression is high on naïve T-cells, but low on more differentiated central memory T-cells [42]. Thus, a potential limitation of our approach is that CCR7^{bright} and CCR7^{dim} cells would be classified as a single cell population, or conceivably, that the CCR7^{dim} would be grouped with the CCR7^{-}. To address this limitation, the pipeline could be modified to support automatic gating of more than two cell populations. This will become particularly important for bar-coded samples (where

dozens of different populations are represented by the bar-code [64]), although in this case the problem is lessened by having prior knowledge of the number of populations present. Nevertheless, because these cells differ in expression of other markers, the populations may be resolved when the complete phenotypic combinations using the rest of markers are created [107].

The second step lists all possible combinations of markers, and assesses the frequency of each immunophenotype within patient samples. By designating positive and negative populations for each of the 10 markers studied, 2^{10} (1024) terminal immunophenotypes were identified. Thus, every subset, defined by any combination of markers, was examined. However, this assumes that every marker is relevant to clinical outcome, which is unlikely. To examine immunophenotypes defined both by combinations of all markers, and by combinations of all subsets of markers, our algorithm allowed markers to be neutral. It is thus possible to measure the frequency of each of the parent populations as well as the terminal ones. For example, our algorithm identified and quantified not only $CD4^{+}CD45RA^{-}CCR7^{+}Ki\text{-}67^{+}CD57^{-}CD27^{+}$ cells, but also cells in the $CD4^{+}CD45RA^{-}CCR7^{+}$ parent population (*i.e.*, $CD4^{+}CD45RA^{-}CCR7^{+}Ki\text{-}67^{N}CD57^{N}CD27^{N}$, where N marks the neutral state). This ability to allow neutral markers is important to discovery efforts, since it enables researchers to include markers in their experimental design without knowing ahead of time whether they are clinically relevant. This process resulted in the identification of 3^{10} (59 049) immunophenotypes, defined by all combinations of positive and negative populations over all combinations of the 10 markers.

The third step determines whether the frequency of each of these immunophenotypes is associated with the clinical outcome by CPHR and the log rank test. Because of the high number of candidate immunophenotypes, adjustment for multiple comparisons is critical. We chose the conservative approach of using Bonferroni's method, knowing that the level of false positives would be low, at the cost of some statistical power. Alternatively, less conservative approaches used in other high-dimensional biological assays [87] could be employed. At this step, the pipeline identified 101 phenotypes with a statistically significant relationship with the clinical outcome (time to AIDS/death).

However, since the second element of the algorithm allows for inclusion of

parent populations, some of the phenotypes identified are overlapping and highly correlated. To unravel relationships that are driven by parent populations from uniquely important cell subsets, the fourth step of our pipeline calculates the impact of each individual marker. This is determined by clustering the immunophenotypes based on the Pearson correlation between them, and then assuming that each cluster of immunophenotypes represents a single cell type, uniquely related to the clinical outcome. In the dataset presented here, we find three distinct populations of cells that predicted time to AIDS/death.

Finally, the fifth step of the pipeline simplifies the cell populations with the strongest relationship to clinical outcome by identifying the minimal set of markers that can be used to define them. Unlike subjective methods that are based on a researcher's assessment of which markers are important, this step is based on "impact" values calculated by the algorithm. One disadvantage of this method is that it is a greedy approach, capable of finding the subtractively minimal marker set, but potentially not the globally optimal markers. In future, graph theory [86] or graphical modeling tools could be developed both to visualize connections between the cell populations that affect clinical outcome, and to find globally optimized marker sets defining them. Nevertheless, even in its current form, the algorithm can distill the complexity of a multivariate data set into immunophenotypes that can be assessed in resource-poor or clinical settings.

The three cell populations defined by the algorithm included one closely related to the $CD8^+$ Ki-67^+ (proliferating) cells identified in previous analysis [42]. However, our computational pipeline showed that the presence of these cells in both the $CD4^+$ and $CD8^+$ T-cell compartment had predictive value. Moreover, the pipeline refined the definition of these cells to include only those that were lacking a receptor involved in homoeostatic proliferation ($CD127^-$). These cells may represent antigen-experienced memory and effector cells, proliferating in response to the immune activation that occurs during HIV infection. A second population identified by the algorithm was $CD45RO^-CD8^+CD57^+$ $CCR5^-CD27^+$ $CCR7^-CD127^-$. Interestingly, this cell type could not be defined by fewer markers (*i.e.*, it was not flagged as redundant by the backward elimination algorithm in step five, thus demonstrating the importance of multiparametric measurements. The immunophenotype of these cells is consistent with highly differentiated (termi-

nal) effector T-cells, which have re-expressed CD45RA -not measured- and CD27. Notably, these cells represent the polar opposite of naïve cells, which were found to have slight predictive power in the manual analysis. The number of markers necessary to define these cells likely reflects the expression of markers of terminal effector cells (like CD57) within other memory cell populations. Thus, the automated algorithm has honed in on the best possible definition of this cell type. Finally, the algorithm identified $CD28^{-}CD45RO^{+}CD57^{-}$ cells as clinically relevant. This population likely includes cells capable of strong effector function, which have not yet lost the ability to proliferate or differentiate. The biological function of these cells is not well understood, but the predictive value of this immunophenotype suggests that studies to further characterize these cells is necessary. In the future, cell ontology approaches may be developed to define a consistent nomenclature for the subsets identified in PFC analysis, particularly those that have unique clinical importance. Such efforts would facilitate our understanding of the underlying biology and would allow simpler meta analysis of data across studies [7, 113]. Following this direction, it will be possible to connect PFC studies to the existing efforts of system biologists [89].

Importantly, all three cell subtypes are rare after removing the redundant markers (Table 3.1); this highlights another major advantage of this pipeline over standard methods: manual or computational identification of rare cell subtypes is challenging [4, 28]. However, a large number of rare cell subtypes exist in the human immune system, and it is well established that rare cells play an important role in the immune system (*e.g.*, HIV [40], stem cell research [88], and cancer [130]).

We allowed the automated pipeline to search for clinically relevant subsets from the entire T-cells, rather than within only $CD4^{+}$ or $CD8^{+}$ T-cell compartments (as is typically done with standard methods). This approach has two advantages. First, it limits the preliminary gates that are needed to prepare the data, making the analysis easier and less susceptible to error or subjectivity. Second, some of the immunophenotypes identified may be relevant to both $CD4^{+}$ and $CD8^{+}$ T-cell biology, as is the case for immunophenotypes where the algorithm identified that the CD4 and CD8 markers are irrelevant. Given the stark differences between $CD4^{+}$ and $CD8^{+}$ T-cell biology in HIV (one cell type is infected and depleted,

Table 3.4: The identified phenotypes, projected into the Cytotoxic and Helper T-cell compartments.

	Phenotype	p-value	p-value CI p-value	adjusted	CPHR Coefficient	R^2	Cell Frequency
	Original Phenotypes:						
1	KI-67+CD127-	2.7e-08	(0.0e+00, 7.3e-06)	1e-03	19.1	0.06886	1e-02
2	CD45RO-CD8+CD57+CCR5-CD27+CCR7-CD127-	3.1e-07	(1.7e-13, 3.4e-03)	1e-02	633.0	0.05869	6e-04
3	CD28-CD45RO+CD57-	5.6e-07	(1.2e-14, 2.6e-04)	2e-02	12.4	0.05620	5e-02
	Projected to the Cytotoxic Compartment (CD8+CD4-):						
4	KI-67+CD8+CD4-CD127-	6.4e-08	(4.2e-14, 2.7e-05)	2e-03	43.6	0.06528	6e-03
5	CD45RO-CD8+CD4-CD57+CCR5-CD27+CCR7-CD127-	3.1e-07	(5.6e-14, 2.7e-03)	1e-02	633.0	0.05869	6e-04
6	CD28-CD45RO+CD8+CD4-CD57-	2.6e-06	(2.2e-10, 2.9e-03)	1e-01	15.3	0.04982	3e-02
	Projected to the Helper Compartment (CD8-CD4+):						
7	KI-67+CD8-CD4+CD127-	2.7e-04	(2.4e-12, 1.2e-03)	1e+01	31.9	0.03023	3e-03
8	CD45RO-CD8-CD4+CD57+CCR5-CD27+CCR7-CD127-	4.3e-01	(5.7e-03, 9.3e-01)	2e+04	-163.1	0.00144	5e-05
9	CD28-CD45RO+CD8-CD4+CD57-	6.3e-01	(1.4e-02, 9.4e-01)	2e+04	4.7	0.00054	7e-03

while the other expands), immunophenotypes that are clinically relevant and shared between the two compartments may be particularly interesting for future study. Table 3.4 shows the projection of these populations into the cytotoxic and helper populations. The table shows that the cytotoxic compartment has a stronger predictive power than the helper compartment, which confirms the findings of previous manual analysis [42]. In addition, similar results were reported in a recent comparison of these cells against other components of the immune system (*i.e.,* natural killer (NK) cells and B-cells) in SIV infection [31].

Although much of our effort was geared toward development of an computational pipeline, we embedded a number of opportunities for users to integrate their biological knowledge into the analysis, with the aim of producing a more robust system. For example, biological knowledge could be used to exclude irrelevant cells (*e.g.*, B-cells, dead cells and debris cells, and doublets); therefore, we allowed manual identification of live, $CD3^+$ T-cells. In addition, for low frequency populations (*e.g.*, Ki-67^+ cells), we offered the ability to set a threshold gate based on a negative control. Finally, the number of phenotype groups reported by the algorithm could be limited, based on the investigator's biological knowledge.

In summary, our pipeline allows the identification of a large number of rare populations associated with clinical outcome and then characterizes these cell types using only the most impactful markers. Although it was applied to an HIV dataset in this work, it can be used in its current form to analyze any PFC study, across a wide variety of disciplines (including but not limited to studying malaria, tuberculosis, autoimmune diseases and various blood cancer subtypes). In particular, this computational approach holds significant potential for: 1) detailed

exploratory analysis of the immune system (using a high number of markers to parse the cell populations), 2) analysis of large cohorts of subjects (*e.g.*, clinical studies and vaccine/drug trials), and 3) screening studies to identify appropriate marker panels for further clinical investigation.

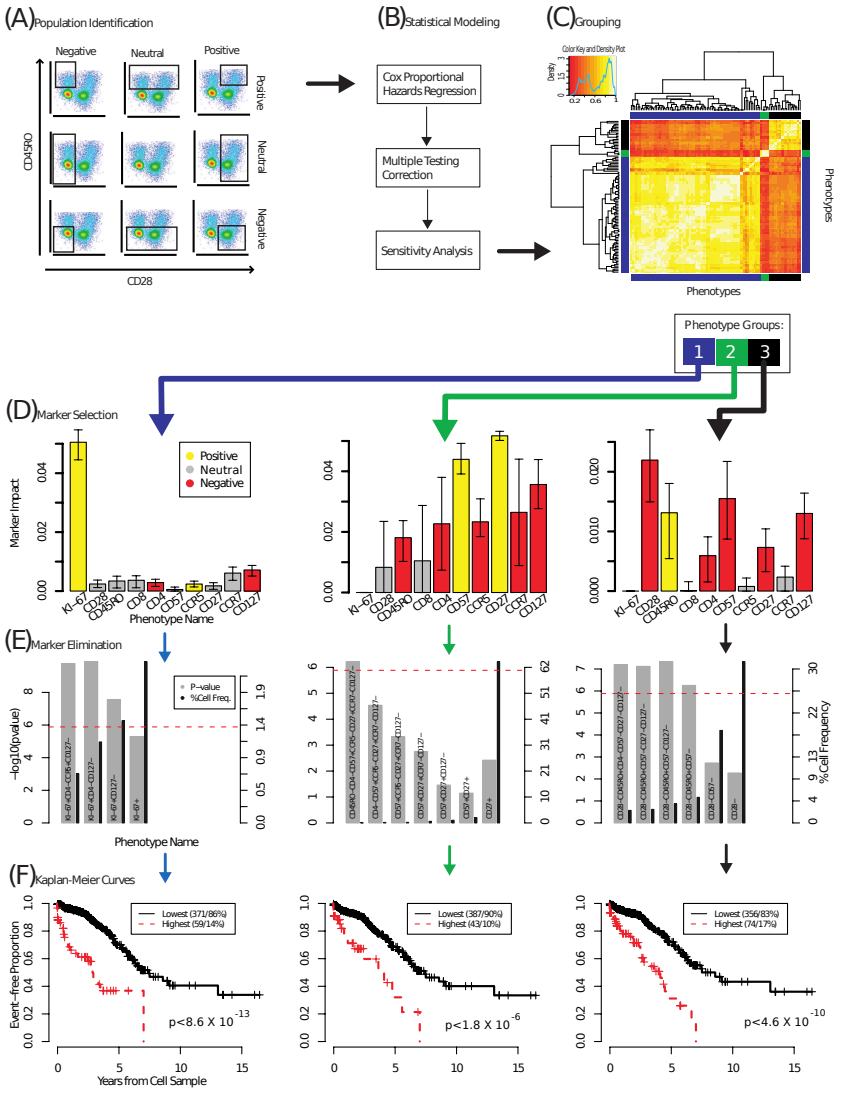


Figure 3.1: The computational pipeline for discovering correlates of HIV protection using PFC. (A) 59069 cell populations were identified for 466 patients; a CPHR model was used to select the immunophenotypes with significant predictive power; (C) the correlation between the immunophenotypes suggested 3 internally correlated groups, shown in the side-bar colors and circumscribed by the bright yellow squares on the diagonal; (D) each group was represented by a specific combination of markers. The markers that were consistently positive or negative across all immunophenotypes are colored yellow and red, respectively, the markers with a mix of positive and negative values are grey; (E) the redundant markers were removed without affecting the predictive power; (F) the resulting immunophenotypes were used to partition the patients to two groups with different survival patterns.

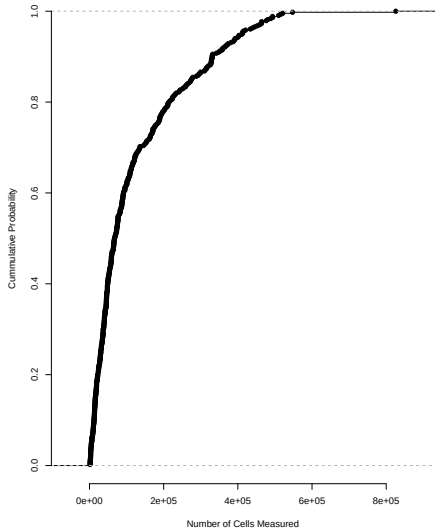


Figure 3.2: Empirical CDF of the number of T-cells measured for each sample. Minimum, maximum, mean and median of the distribution are 144, 825739, 123682, and 68095, respectively.

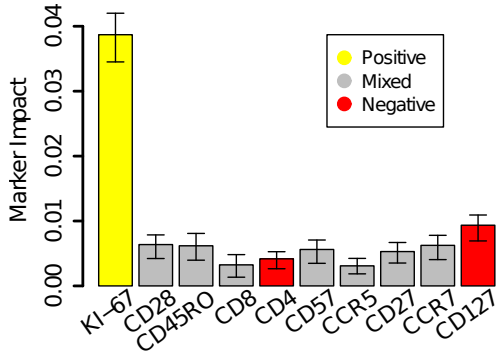


Figure 3.3: Bulk (over all phenotypes) measurement of the impact of each marker and the respective %95 confidence intervals.

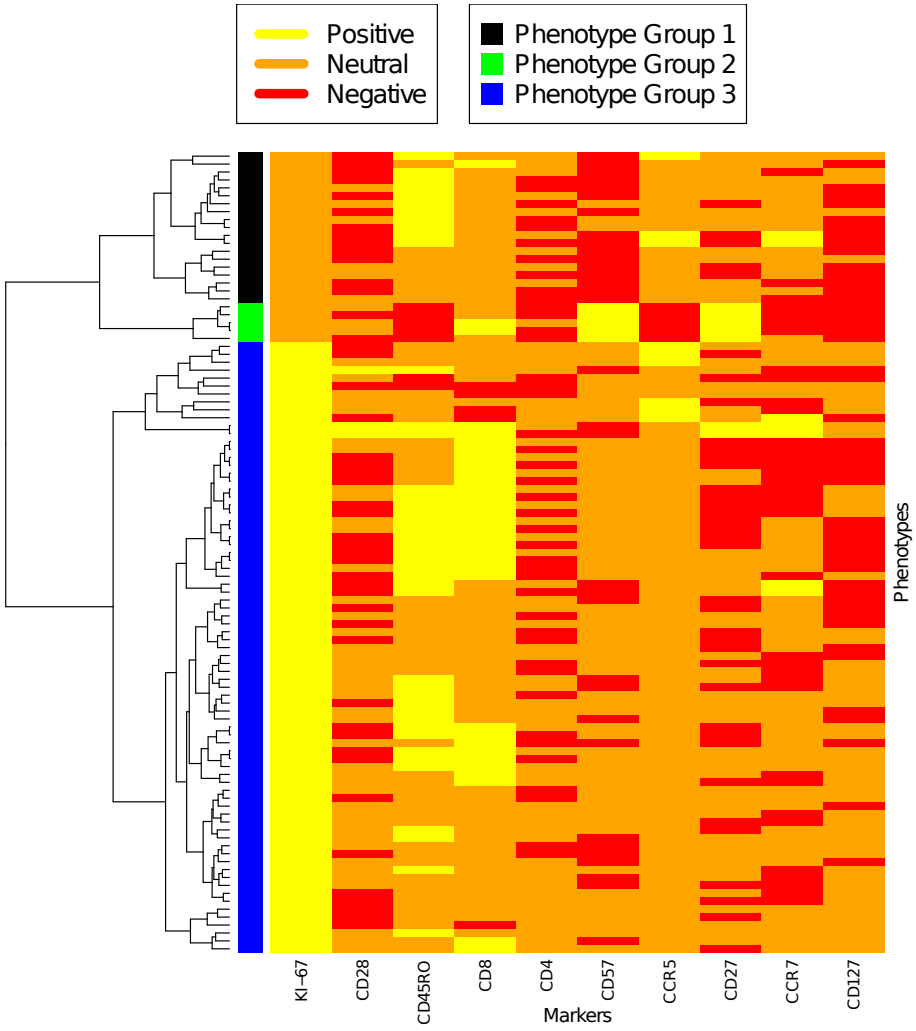


Figure 3.4: Hierarchical clustering of the statistically significant phenotypes based on the correlation between them. The phenotype names are replaced with a heatmap to make it easier to observe patterns. The colours denote the "state" of each marker (column) for each phenotype (row).

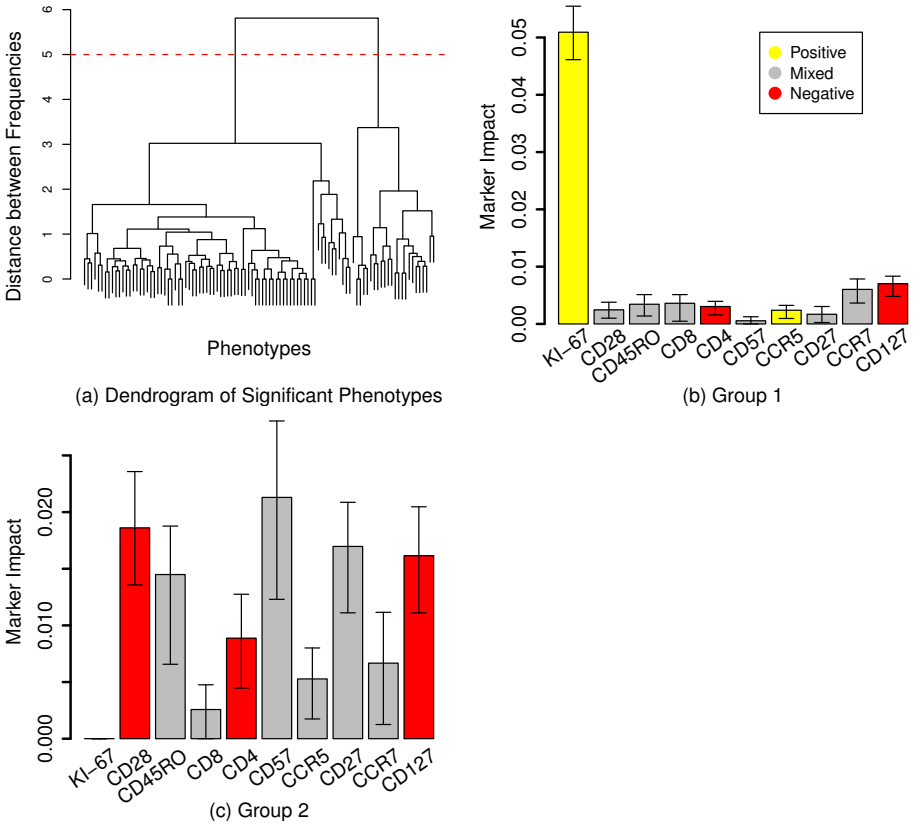


Figure 3.5: (a) Hierarchical clustering of phenotypes. The red dashed line shows the threshold which results in five groups of phenotypes, (b) and (c) the impact of each of the markers inside the groups of phenotypes.

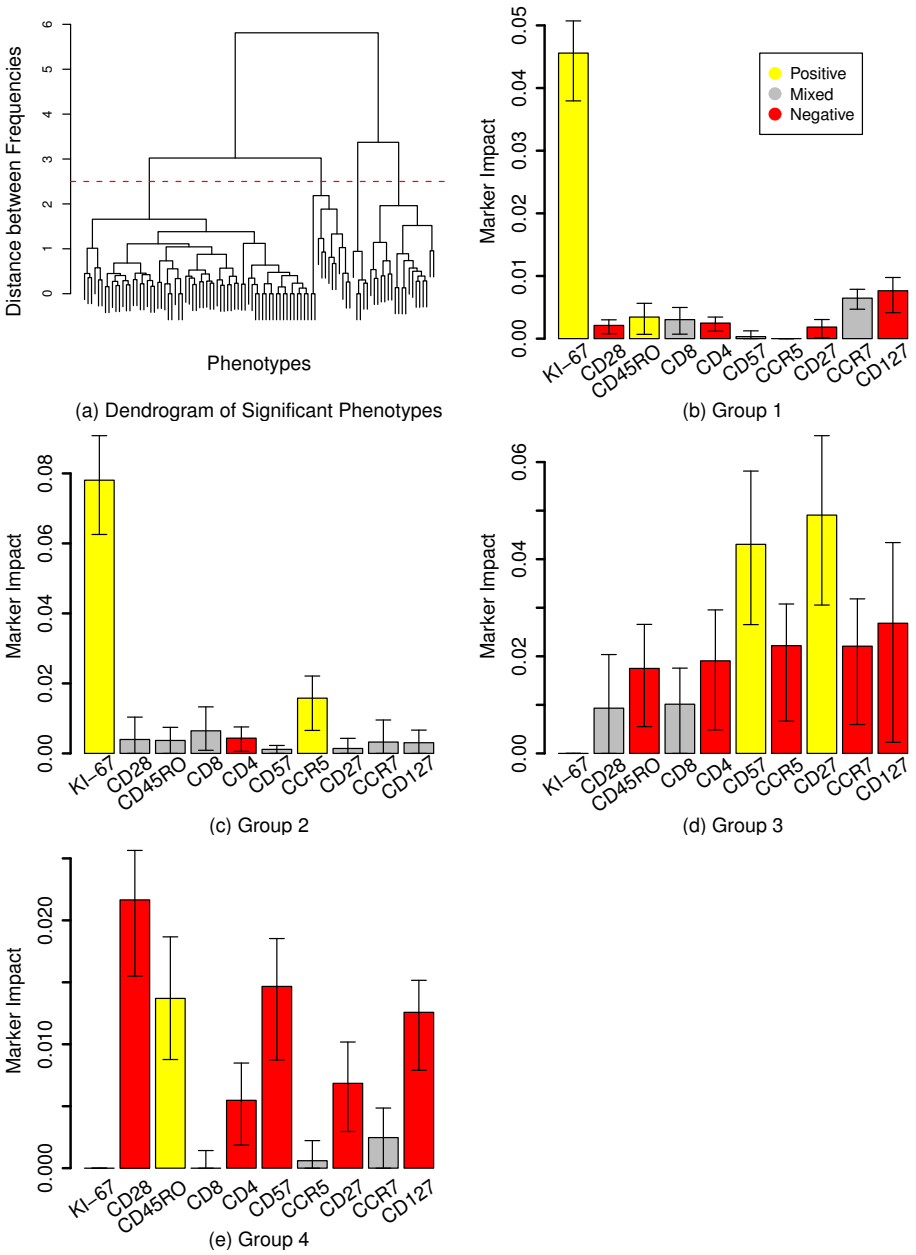


Figure 3.6: (a) Hierarchical clustering of phenotypes. The red dashed line shows the threshold which results in five groups of phenotypes, (b), (c), (d), and (e) the impact of each of the markers inside the groups of phenotypes.

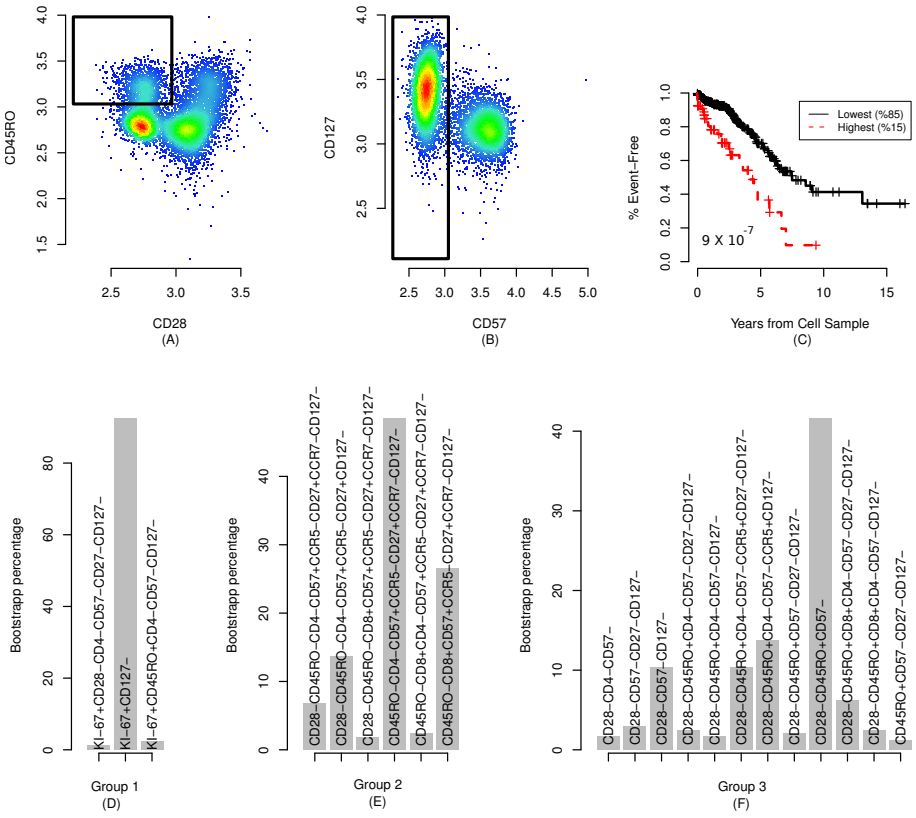


Figure 3.7: Confirmatory analysis. (A,B) The $CD28^{-}CD45RO^{+}CD57^{-}$ immunophenotype was identified by manual analysis of all samples. (C) Kaplan-Meier curves confirm the predictive power of the manually measured immunophenotype. (D,E, and F) The immunophenotypes originally selected by the pipeline were dominant in bootstrapping-based sensitivity analysis of the entire pipeline.

Chapter 4

RchyOptimyx: Cellular Hierarchy Optimization for Flow Cytometry

4.1 Introduction

Recent advances in FCM instrumentation and reagents have enabled high-dimensional analyses to identify large numbers of cell populations with potentially significant correlations to an external outcome (see, *e.g.,* 4). However, studies often fail to characterize the complex relationships between the markers involved in the identification of these cell populations. Revealing this information can provide additional insight into the biological characteristics of the populations identified. The choice of markers for new panels has been a source of ongoing debate, including efforts such as the Human Immuno Phenotyping Consortium (HIPC), the Federation of Clinical Immunology Societies Federation of Clinical Immunology Societies (FOCIS) sponsored Flow Immunophenotyping Technical Meetings (FITMAN), and the Optimized Multicolor Immunophenotyping Panels (OMIPS) articles [13, 25, 32, 39, 66, 75, 76, 81, 95, 104, 124, 131]. Understanding the relationships between the markers involved in identification of the target cell population and the characteristics of that cell population (*e.g.,* its correlation with a clinical

outcome) is fundamental to the design of effective marker panels. For example, one could use a high-dimensional flow or mass cytometry assay to measure a large list of candidate markers. However, this can result in parsing the cells into (*e.g.,* clinically) redundant subsets [12]. Excluding these redundancies (*e.g.*, markers less important for prediction of a clinical outcome) will result in a panel of the most clinically relevant markers.

High dimensional FCM data is usually analyzed using a laborious sequential manual analysis (see, *e.g.,* [43, 92]). However, manual gates provide little insight into the relative importance of each gate to the final results. For example, consider a six color assay with markers named 1 to 6. If the expression of each marker is considered to be on, off, or don't care (*e.g.,* markers named 1, 2, and 3 in phenotype 1^+2^-, respectively), a total of $3^6 = 729$ cell populations can be distinguished based on these markers. A given immunophenotype involving all six of these markers (*e.g.,* $1^+2^-3^+4^-5^+6^-$) can have $2^6 = 64$ parent populations (*e.g.,* 1^+, 1^+2^-).

Quantifying the relationship between the cell population of interest and these parent populations is fundamental to our understanding of the importance of the markers for different gating strategies. The order in which the gates are applied to the data is not important, as long as all of the gates are used (*i.e.,* sequential gating is commutative). However, to decrease the size of the marker panel, the relative importance of the gates should be determined. For example, the measurement of the phenotype mentioned above using only five colors requires the determination of the importance of each marker to identify and remove the least important one (*i.e.,* the identification of the parent population with five markers that is most similar to the original phenotype). This is further complicated by the fact that some cell populations can be identified using more than one combination of markers and gating strategy; therefore, each marker can be used in different positions in the gating hierarchy and can have different priorities, depending on the choice of the gating strategy. For example, the 3^+ gate is involved in both $1^+2^-3^+$ and $3^+4^-5^+$, both parents of the $1^+2^-3^+4^-5^+6^-$ phenotype described above. However, depending on the amount of redundancy between marker 3 and others, this marker can have different levels of importance for these two parent populations.

Another use–case for measuring the importance of the markers is the investi-

gation of a large number of closely related phenotypes (*e.g.*, those identified by bioinformatics pipelines) by identifying their common parent populations. Several computational tools have been developed for automated identification of cell populations (*e.g.,* [2, 11, 20, 38, 70, 85, 97, 99, 101, 117, 128]) and recent studies have used these tools to identify novel cell populations that correlate with clinical outcomes (*e.g.,* [3, 9, 27, 105, 129]). In addition, the results of the FlowCAP-II project (see Chapter 5) and also the results presented in Chapter 3 have shown that several algorithms can accurately and reproducibly identify cell populations correlated with external outcomes. However, these algorithms provide limited information regarding the importance of the markers involved in defining the cell populations [3, 21].

This situation is even more complicated than sequential manual gating, since most of these bioinformatics pipelines work based on multivariate classifiers, and as a result, more than one cell population can be responsible for the final predictions. Therefore, markers can have different relative importance in defining the multiple cell populations within the multivariate model. Quantifying the markers for each phenotype involved in the multivariate model can provide additional insight into the differences between closely related cell populations. For example, if two phenotypes $1^+2^-3^+4^-5^+$ and $1^+2^-3^+4^-6^+$ are identified as correlates of a disease, and if markers 5 and 6 (which are the only differences between them) are the least important markers for the former and latter phenotypes, respectively, then these two phenotypes are likely to correspond to the same cell population (as far as the correlation with the disease is concerned). However, if markers 5 and 6 are the most important for the phenotypes, these can correspond to two biologically different cell populations.

To address these problems, we developed RchyOptimyx, a computational tool that uses dynamic programming and optimization techniques from graph theory to construct a cellular hierarchy, providing the best gating strategies to identify target populations to a desired level of purity or correlation with a clinical outcome, using the simplest possible combination of markers.

4.2 Materials and Methods

Our methodology builds on the flowType pipeline described in Chapter 3. flowType comprehensively identifies cell populations defined by all possible gating strategies (hierarchies) in the data set using a partitioning strategy (*e.g.*, clustering algorithm like flowMeans [3]) and scores them by a statistical test (*e.g.*, the log rank test for difference in survival distributions). Given the list of all cell populations and their scores, RchyOptimyx uses a dynamic programming approach to find the best cellular hierarchy within a reasonable time (*i.e.* less than 2 minutes for 30 color data), as well as a number of best suboptimal hierarchies, to enable mining of the space of best gating strategies and purities for a given target cell population.

4.2.1 Terms and Definitions

Let $\mathscr{M}$ be the set of m markers of interest (*e.g.*, $\mathscr{M} = \{KI\text{–}67, CD28, CD45RO\}$), a single marker phenotype be a phenotype having only one marker (*e.g.*, $CD28^+$), a phenotype P be a set of single marker phenotypes (*e.g.*, $P = KI\text{–}67^+CD28^-$), and M be a phenotype of size m that involves all of the markers (*e.g.* $M = KI\text{–}67^+ CD28^-CD45RO^-$). The power set of M, $\mathscr{P}(M)$, is of size 2^m and contains every possible subset of M. The scoring function $S(.)$ assigns a score to each member of $\mathscr{P}(M)$, such that higher values are assigned to more important phenotypes (*e.g.*, those with a stronger correlation with a clinical outcome).

Given an arbitrary M, the directed acyclic graph (DAG) G_M has $m+1$ levels from 0 to m, each level i including every member of $\mathscr{P}(M)$ of size i. Node s is connected to node t with a directed edge (s,t) if and only if $|t| = |s|+1$ and the two associated sets of s and t differ only in one single phenotype marker (*i.e.,* t is an immediate parent of s). Let the weight of the edge (s,t) be $-S(t)$ (so that paths with maximum score can be found by searching for paths with minimum total weight).

The node with 0 markers is the root (or source) node, and the node with the complete set of markers is the sink node. A path from source to sink is called a hierarchy path, or simply a hierarchy. An example of graph G_M for $M = KI\text{–}67^+CD4^-CCR5^+CD127^-$ is illustrated in Figure 4.1.

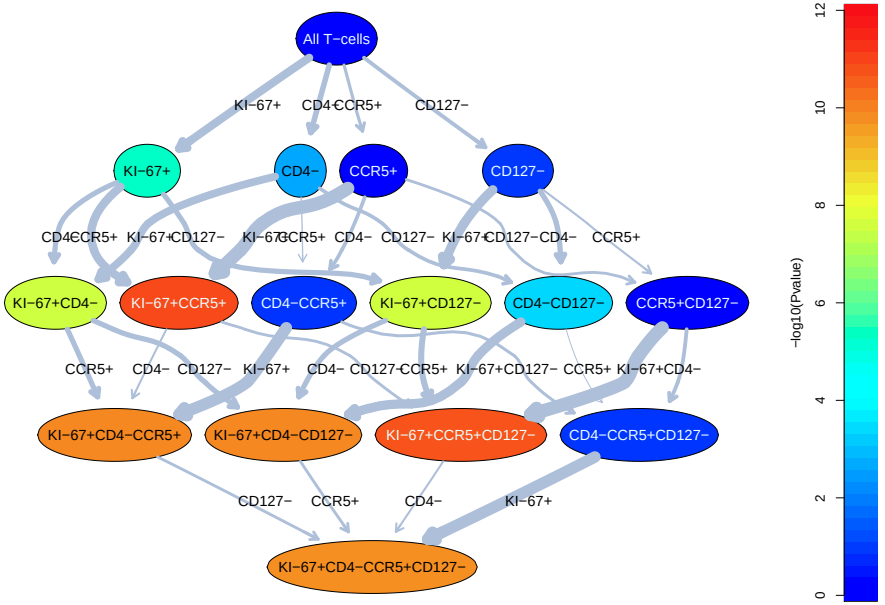


Figure 4.1: A complete cellular hierarchy for prediction of HIV clinical outcome using $KI67^{+}CD4^{-}CCR5^{+}CD127^{-}$ T-cells. The color of the nodes shows the significance of the correlation with the clinical outcome (p-value of the logrank test for the cox proportional hazards model) and the width of each edge (arrow) shows the amount of change in this variable between the respective nodes. The positive and negative correlation of each immunophenotype with with outcome can be shown by the arrow type leading to the node, however as all correlations are negative in this hierarchy, only one arrow type is shown.

The graph G_M has $|\mathscr{P}(M)| = 2^m$ nodes, one node for each parent phenotype of the phenotype of interest. The number of edges is equal to the number of markers (m), times the number of edges that have the specified marker. Each marker appears in 2^{m-1} nodes, therefore the number of edges is $m \times 2^{m-1}$.

Intuitively, comparing two hierarchies, the one which goes through nodes with higher score nodes is better. On our graphs, to ensure "better" hierarchies get lower scores, we define the weight of each path to be the score of the respective hierarchy. Modeling this intuition. The score of a hierarchy, thus, can be written as follows:

$$\begin{aligned} T(\mathscr{H}) &= \sum_{(s,t)\in E_{\mathscr{H}}} W(s,t) \\ &= \sum_{(s,t)\in E_{\mathscr{H}}} -S(t) \\ &= \sum_{t\in V_{\mathscr{H}}\setminus source} -S(t) \end{aligned} \tag{4.1}$$

in which $\mathscr{H}$ is the hierarchy, $E_{\mathscr{H}}$ is the set of edges of hierarchy $\mathscr{H}$, $V_{\mathscr{H}}$ is the set of vertices of same hierarchy, and *source* is the first node in the hierarchy.

4.2.2 Dynamic Programming to Identify the Best Hierarchy

For cell populations characterized by m markers, finding the best hierarchy by searching through all possible hierarchies would require time $O(m!)$, which is impractical for even moderately large m. To make this problem tractable using dynamic programming, we define *best total score* function $T^*(.)$, which computes the score of the best hierarchy leading to the given phenotype. $T^*(.)$ is defined recursively as follows:

$$T^*(P^k) = \begin{cases} -S(P^k) & \text{if } k = 1 \\ \min\{T^*(P^k \setminus P_i^k) - S(P^k) | i = 1,\ldots,k\} & \text{otherwise} \end{cases}, \tag{4.2}$$

where P^k is a cell population defined by k single marker phenotypes, and $P^k \setminus P_i^k$ is P^k with the i^{th} single marker phenotype removed. For example, if $P^3 = KI\text{–}67^+CD28^-CD45RO^+$, then $P^3 \setminus P_1^3 = CD28^-CD45RO^+$. In other words, there

is an edge from $P^k \setminus P_i^k$ to P^k in G_M, where P^k is a subset of M. Also note that $-S(P^k)$ is the weight of the edge $(P^k, P^k \setminus P_i^k)$ in G_M.

Using dynamic programming, we calculate the value of $T^*(.)$, iterating from level 0 to m on G_M. Calculating each node's score requires a number of constant–time operations equal to the number of edges entering the node. Therefore, the total number of operations is proportional to total number of edges ($m \times 2^{m-1}$), and the overall time complexity of our programming procedure for determining $T^*(.)$ values for all phenotypes in the graph is $O(m \times 2^{m-1})$. An illustration of the dynamic programming space for $m = 3$ as well as two paths in that space is shown in Figure 4.2.

4.2.3 Search for Near-Optimal Hierarchies

The hierarchy selected by the dynamic programming algorithm is the best gating strategy for a given cell population. However, we would also like to identify alternate gating strategies with slightly worse scores. To find these near-optimal paths, we reformulate the problem as identification of a desired number of minimum weight paths: In G_M, the minimum weight path from source to sink is the best hierarchy (identical to the one generated by dynamic programming). To generate additional, sub-optimal hierarchies, a list of the next minimum weight paths must also be generated. These paths can be identified using the method by Eppstein detailed in [33]. Briefly, this method uses the minimum spanning tree of G_M and computes a heap structure for each node; it then merges the heaps in an efficient way to construct a 4-heap data structure. Using this 4-heap and a given arbitrary number l (the number of desired paths), it generates l-minimum weight paths in time $O(e + v + l)$ for a DAG with e edges and v nodes (see Theorem 4 of [33] for details).

Hence, the time complexity of our algorithm can be calculated by plugging the number of edges and nodes into the time complexity of the l-minimum weight

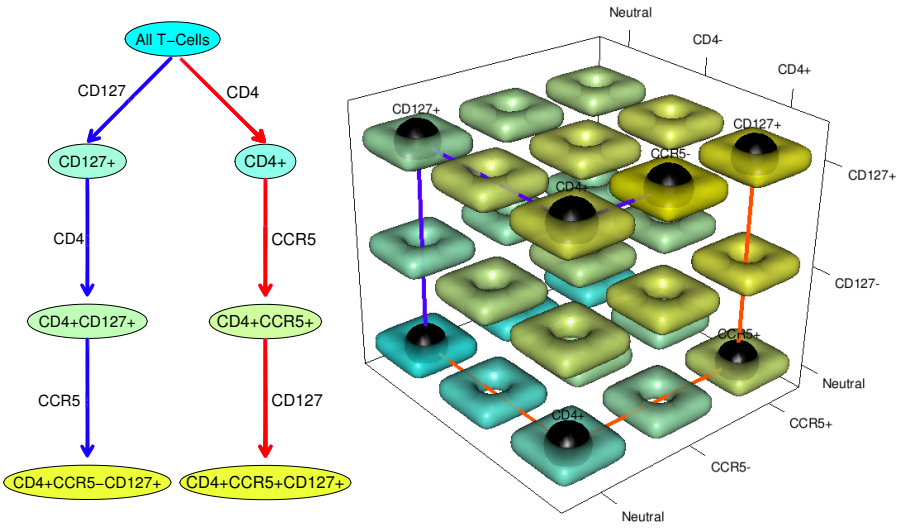


Figure 4.2: Dynamic programming algorithm for two cell populations defined by 3 markers. The best paths for each of the cell populations are shown in red and blue, respectively. As an example, the red path ends at $CD4^{+}CCR5^{+}CD127^{+}$. Three markers are available to be added. First, CD4 is added (changes from don't care to positive). Then, two options will be available for the next step (CD127 and CCR5). After selection of CCR5, only one option will be left for the final step (CD127). Therefore for three markers, $\frac{3\cdot(3-1)}{2} = 6$ comparisons were required. **Left:** A hierarchy for the two paths. The label of an edge is the name of the single marker phenotype that is the difference between its head set (s) and its tail set (t). **Right:** the dynamic programming space for the 3 markers. Black spheres mark the nodes in the dynamic programing space used by the two paths. The colors of the nodes on the left match that of the square tori on the right and correspond to the relative score of each cell population.

paths method:

$$\begin{aligned} O(e+v+l) &= O(m\times 2^{m-1}+2^{m}+l) \\ &= O(m\times 2^{m-1}+2\times 2^{m-1}+l) \\ &= O((m+2)\times 2^{m-1}+l). \end{aligned} \tag{4.3}$$

For example, the number of operations with our approach on a dataset with $m = 10$ markers would be $\approx 10^4$ compared to $\approx 3 \times 10^6$ for the exhaustive search approach. Our method therefore takes ≈ 0.23 CPU seconds vs ≈ 69 CPU seconds for exhaustive search, run under 64 bit Linux (version 3.3) on a $2.93GHz$ Intel Xeon CPU with sufficient memory (proportional to 2^M). For a phenotype involving $m = 20$ markers, these numbers increase to ≈ 1.2 CPU seconds vs $\approx 10^{11}$ CPU seconds (more than 4000 years), respectively. Even for a phenotype involving $m = 30$ markers, measured by a CyTOF assay (mass spectrometry-flow cytometry hybrid device [11, 22, 90]), RchyOptimyx remains feasible, with a runtime of ≈ 102 CPU seconds, while the brute-force method would take $\approx 10^{22}$ CPU seconds. The final output of RchyOptimyx is the corresponding subgraph of G_M that includes all calculated paths (*i.e.*, the optimized hierarchy, *e.g.*, Figure 4.3).

4.2.4 Datasets

We validated RchyOptimyx on two high-dimensional datasets, produced by mass and polychromatic flow cytometry.

Mass cytometry analysis of bone marrow cells from normal donors

In this dataset, 31 parameters were measured for mononuclear cells from a healthy human bone marrow (see [11] for details). We used the results of three assays on samples subject to *ex vivo* stimulation by IL7 (measured by pSTAT5), BCR (measured by pBLNK), and LPS (measured by p-p38) as well as an unstimulated control. 13 surface markers were included in the analysis: CD3, CD45, CD45RA, CD19, CD11b, CD4, CD8, CD20, CD34, CD33, CD123, CD38, and CD90. Singlets were gated manually, as described by Bendall *et al.,* [11].

Polychromatic flow cytometry analysis of HIV$^+$ patients

As described in Chapter 3, this dataset consists of 13 color PFC assays of 466 HIV$^+$ subjects enrolled in the Infectious Disease Clinical Research Program's HIV Natural History Study. Basic demographic characteristics of this dataset are described elsewhere [125]. Peripheral blood mononuclear cells stored within 18 months of the date of seroconversion were analyzed using PFC as described by

Ganesan *et al.* [42]. The cohort included 135 death/AIDS events, as defined by 1993 guidelines [19]. The date of the last follow-up or initiation of highly active anti-retroviral therapy (HAART) was considered a censoring event. CD14 and V-amine dye were used to exclude monocytes and dead cells, respectively, CD3 was used to gate T-cells. Using the staining panel and flowType, we enumerated various subsets of naive and memory T-cells, defined by CD4, CD8, CD45RO, CD27, CD28, CD57, CCR5, CCR7, CD127, and KI-67. Using a log rank test with Bonferroni's multiple test correction, we scored each subset (cell population) in terms of its correlation with HIV progression [3].

4.3 Results

4.3.1 Designing a Panel to Detect a Population Expressing an Intracellular Marker using Surface Markers

In this use–case, our goal was to identify cell populations that are affected by different stimulations in the mass cytometry dataset. We used flowType to identify a list of populations that had a high overlap with either the $IL3^+$, BCR^+, or LPS^+ populations (determined manually - see Figure 4.9). For each cell population, this value was calculated as the difference in its intersection with the $IL3^+$, BCR^+, or LPS^+ compartments between the stimulated and unstimulated sample. For example, for a given cell population CP, the overlap with $IL3^+$ was defined as:

$$Overlap^{IL3^+}(CP) = \left(\frac{\#\,\mathrm{IL3}^+ cells\ in\ CP}{\#\ cells\ in\ CP}\right)_{stim} - \left(\frac{\#\,\mathrm{IL3}^+ cells\ in\ CP}{\#\ cells\ in\ CP}\right)_{unstim} \tag{4.4}$$

The immunophenotypes with a high overlap, as identified by flowType, are listed in Tables 4.1, 4.2, and 4.3. These immunophenotypes were analyzed using RchyOptimyx (*e.g.*, Figure S1 for BCR) and then merged into a single graph, shown in Figure 4.5. This graph suggests that T-cells ($CD3^+$) followed by cytotoxic T-cells ($CD3^+CD4^+$) are the main parent populations affected by IL7 stimulation (panel A). As expected, BCR stimulation affected B-cells ($CD19^+$ $CD20^+CD3^-$), and LPS stimulation increased the proportion of $CD19^-CD33^+CD3^-$ cells (Panels B and C, respectively). These results are generally consistent with those reported by Bendall *et al.* (Figure 2 and panel C of

Figure 3 of [11]).

Table 4.1: The phenotypes with a high overlap with the BCR(pBLNK)$^+$ compartment as identified by flowType. The table includes the cell proportion of these immunophenotypes (second column) and the differences in the cell proportion of BCR(pBLNK)$^+$ cells in the stimulated and unstimulated assays (third column).

Phenotype Name	Cell Proportion	$BCR^+_{(stim-unstim)}$
CD19+CD4-CD8-CD34+CD20+CD123+CD38-CD3-	0.001	0.160
CD19+CD4-CD34+CD20+CD123+CD38-CD3-	0.001	0.160
CD19+CD4-CD34+CD20+CD123+CD3-	0.001	0.155

Table 4.2: The phenotypes with a high overlap with the IL7(pSTAT5)$^+$ compartment as identified by flowType. The table includes the cell proportion of these immunophenotypes (second column) and differences in the cell proportion of IL7(pSTAT5)$^+$ cells in the stimulated and unstimulated assays (third column).

Phenotype Name	Cell Proportion	$IL7^+_{(stim-unstim)}$
CD19-CD4+CD8+CD20+CD33+CD38-CD3+	0.008	0.364
CD19-CD4+CD8+CD20+CD33+CD3+	0.008	0.366
CD19-CD4+CD8+CD34+CD33+CD38-CD3+	0.008	0.366
CD19-CD4+CD8+CD34+CD33+CD3+	0.008	0.368
CD19-CD4+CD8+CD34+CD20+CD33+CD38-CD3+	0.006	0.399
CD19-CD4+CD8+CD34+CD20+CD33+CD3+	0.006	0.402
CD4+CD8+CD20+CD33+CD38-CD3+	0.011	0.365
CD4+CD8+CD20+CD33+CD3+	0.011	0.371
CD4+CD8+CD34+CD33+CD38-CD3+	0.011	0.366
CD4+CD8+CD34+CD33+CD3+	0.011	0.371
CD4+CD8+CD34+CD20+CD33+CD38-CD3+	0.008	0.399
CD4+CD8+CD34+CD20+CD33+CD3+	0.009	0.405
CD19+CD4+CD8+CD20+CD33+CD38-CD3+	0.003	0.364
CD19+CD4+CD8+CD20+CD33+CD3+	0.003	0.378
CD19+CD4+CD8+CD34+CD33+CD38-CD3+	0.003	0.359
CD19+CD4+CD8+CD34+CD33+CD3+	0.003	0.372
CD19+CD4+CD8+CD34+CD20+CD33+CD38-CD3+	0.002	0.397
CD19+CD4+CD8+CD34+CD20+CD33+CD3+	0.002	0.409

Table 4.3: The phenotypes with a high overlap with the LPS(p-p38)$^+$ compartment as identified by flowType. The table includes the cell proportion of these immunophenotypes (second column) and differences in the cell proportion of LPS(p-p38)$^+$ cells in the stimulated and unstimulated assays (third column).

Phenotype Name	Cell Proportion	$LPS^+_{(stim-unstim)}$
CD19-CD4-CD8-CD34-CD20-CD33+CD123-CD38-CD3-	0.008	0.474
CD19-CD4-CD8-CD34-CD20-CD33+CD123-CD3-	0.008	0.473
CD19-CD4-CD8-CD34-CD20-CD33+CD38-CD3-	0.009	0.466
CD19-CD4-CD8-CD34-CD20-CD33+CD3-	0.009	0.465
CD19-CD4-CD8-CD34-CD33+CD123-CD38-CD3-	0.022	0.460
CD19-CD4-CD8-CD34-CD33+CD123-CD3-	0.022	0.459
CD19-CD4-CD8-CD34-CD33+CD38-CD3-	0.022	0.452
CD19-CD4-CD8-CD34-CD33+CD3-	0.022	0.451
CD19-CD4-CD8-CD34-CD20+CD33+CD123-CD38-CD3-	0.013	0.450
CD19-CD4-CD8-CD34-CD20+CD33+CD123-CD3-	0.013	0.449
CD19-CD4-CD8-CD20-CD33+CD123-CD38-CD3-	0.023	0.453
CD19-CD4-CD8-CD20-CD33+CD123-CD3-	0.023	0.452
CD19-CD4-CD34-CD20-CD33+CD123-CD38-CD3-	0.011	0.456
CD19-CD4-CD34-CD20-CD33+CD123-CD3-	0.011	0.455
CD19-CD8-CD34-CD20-CD33+CD123-CD38-CD3-	0.012	0.462
CD19-CD8-CD34-CD20-CD33+CD123-CD3-	0.012	0.461
CD19-CD8-CD34-CD20-CD33+CD38-CD3-	0.012	0.454
CD19-CD8-CD34-CD20-CD33+CD3-	0.012	0.454
CD4-CD8-CD34-CD20-CD33+CD123-CD38-CD3-	0.011	0.462
CD4-CD8-CD34-CD20-CD33+CD123-CD3-	0.011	0.461
CD4-CD8-CD34-CD20-CD33+CD38-CD3-	0.011	0.454
CD4-CD8-CD34-CD20-CD33+CD3-	0.011	0.454
CD8-CD34-CD20-CD33+CD123-CD38-CD3-	0.015	0.450
CD8-CD34-CD20-CD33+CD123-CD3-	0.015	0.449

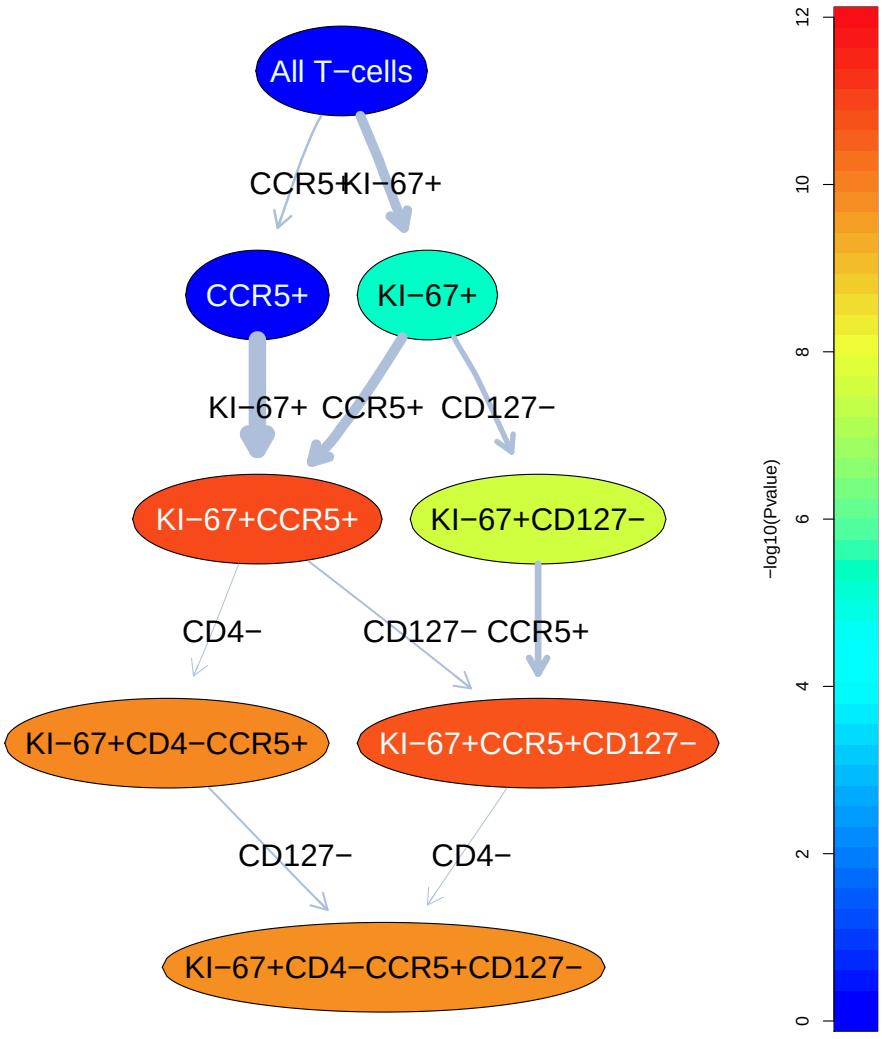


Figure 4.3: An optimized cellular hierarchy for prediction of HIV's clinical outcome using $KI67^{+}CD4^{-}CCR5^{+}CD127^{-}$ T-cells. The color of the nodes shows the significance of the correlation with the clinical outcome (p-value of the logrank test for the cox proportional hazards model) and the width of each edge (arrow) shows the amount of change in this variable between the respective nodes.

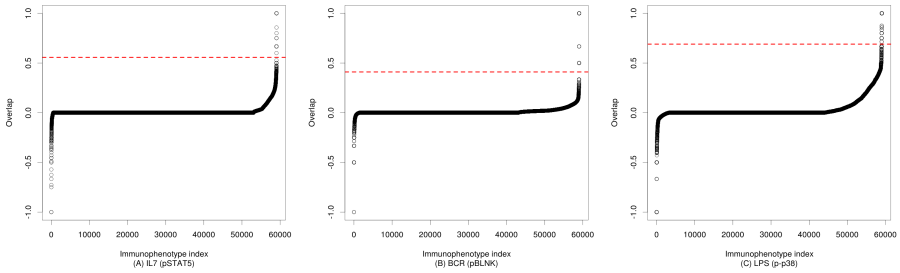


Figure 4.4: All immunophenotypes ordered by their overlap with the cell population of interest. The red dashed lines demonstrate the cutoffs used for selected the immunophenotypes with "high overlap".

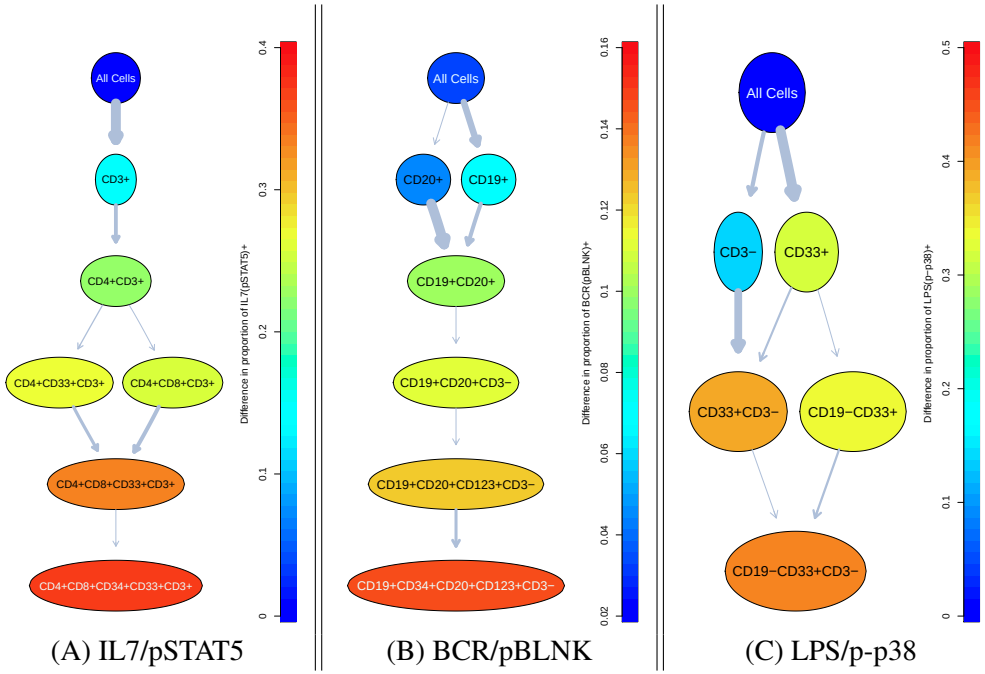


Figure 4.5: Three optimized hierarchies for identification of cell populations with maximum response to IL7, BCR, and LPS measured by pSTAT5, pBLNK, and p-p38, respectively. The colour of the nodes and the thickness of the edges indicates the proportion and change in proportion of cells expressing the intracellular marker of interest, respectively.

4.3.2 Simplifying Gating Strategies

Here we use RchyOptimyx to demonstrate an example of the use case of establishing a simpler combination of markers that can be used to identify a target population at a desired level of purity. For analysis of the PFC dataset, Ganesan *et al.* used a strict, but potentially redundant definition for naive T-cells, of $CD28^{+}CD45RO^{-}CD57^{-}CCR5^{-}CD27^{+}CCR7^{+}$, within the $CD3^{+}CD14^{-}$ compartment [42]. The purity of a given parent cell population (CP) of this target was defined as its mean purity for the strictly-defined naive T-cells:

$$Purity(CP) = \frac{\sum \frac{\#\, CD28^{+}CD45RO^{-}CD57^{-}CCR5^{-}CD27^{+}CCR7^{+}\ cells}{\#\, cells\ in\ CP}}{\#\, Samples} \tag{4.5}$$

Figure 4.6 shows the results of analysis with RchyOptimyx where a combination of only three markers ($CD45RO^{-}CCR5^{-}CCR7^{+}$) identified the strict naive T cell population to 95% purity (within the $CD3^{+}CD14^{-}$ compartment). The range of available purities and determination of an appropriate cutoff are experiment dependent (*e.g.*, on the range of available markers, biological question being researched).

4.3.3 Characterization of a Large Number of Immunophenotypes

In this section, we use RchyOptimyx to demonstrate an example of the use-case of summarizing a large list of immunophenotypes of interest (as identified by a bioinformatics pipeline) into a single hierarchy using their most important common parent populations.

In a previous study of the PFC dataset in Chapter 3, we identified 101 immunophenotypes (Table 3.2) in HIV^{+} patients that had a statistically significant correlation with HIV's progression [3]. The score of each population was calculated as $-log_{10}(p)$ where p was the p-value of the logrank test before adjustment for multiple testing (higher values represent a stronger correlation with the clinical outcome). The 101 immunophenotypes were analyzed using RchyOptimyx, and the resulting hierarchies were merged into a single graph (Figure 4.7). This graph indicated three groups of immunophenotypes that were significantly correlated with HIV's outcome (left, center, and right branches). The left branch consisted of $KI\text{-}67^{+}CD4^{-}$ $CCR5^{+}CD127^{-}$ T-cells. These cells were thought to

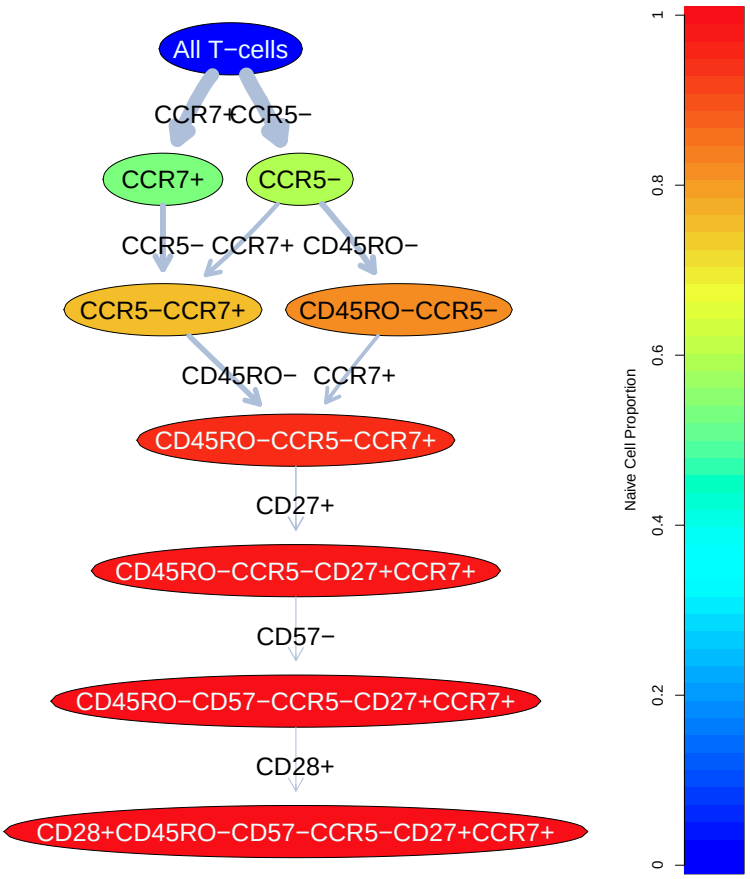


Figure 4.6: An optimized cellular hierarchy for identifying naive T-cells. The color of the nodes and the thickness of the edges shows the purity and change in purity of the original naive phenotype within the given cell population, respectively.

be statistical significant, mainly because they are long-lived ($CD127^-$) T-cells with high proliferation ($KI\text{-}67^+$). RchyOptimyx showed that the significance of this population is related to the $KI\text{-}67^+CCR5^+$ compartment and not $CD127^-$ (Figure 4.7, the left branch), as the CD127 marker is not needed to achieve the approximately the same score. This is in agreement with the results of two recent studies [46, 58]. The terminal node of the center branch consisted of seven

markers (CD45RO$^-$CD8$^+$CD57$^+$CCR5$^-$ CD27$^+$CCR7$^-$CD127$^-$). RchyOptimyx revealed that its most important parent population is CD8$^+$CCR7$^-$ CD127$^-$, with a weaker correlation with the clinical outcome. Finally, the right branch (CD28$^-$CD45RO$^+$CD4$^-$CD57$^-$CD27$^-$ CD127$^-$) suggests several parent populations with minimal overlap and strong correlation with the clinical outcome (*e.g.,* CD28$^-$ CD4$^-$ CD57$^-$ CD127$^-$ and CD45RO$^+$CD4$^-$CD127$^-$).

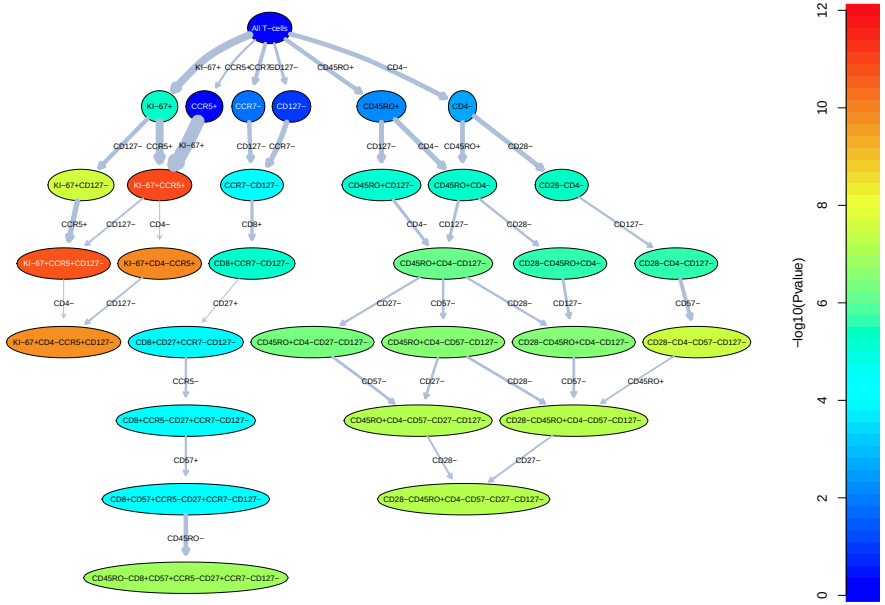


Figure 4.7: An optimized hierarchy for all three populations correlated with protection against HIV. The color of the nodes indicates the significance of the correlation with the clinical outcome (p-value of the logrank test for the cox proportional hazards model) and the width of each edge (arrow) indicates the amount of change in this variable between the respective nodes. The positive and negative correlation of each immunophenotype with outcome can be seen from the arrow type leading to the node, however, as all correlations are negative in this hierarchy, only one arrow type is shown.

4.4 Discussion

Sequential analysis of the markers involved in manual or automated identification of cell populations is fundamental to our understanding of the characteristics of the cell population. In sequential gating, the order in which the gates have been applied does not affect the final results. However, ordering the gates by their relative importance has two use-cases: 1) identifying a cell population of interest, using the smallest possible panel of markers; 2) summarizing a long list of closely related (and perhaps overlapping) immunophenotypes by identifying their most important common parent populations. However, increasing the number of markers quickly renders this approach unfeasible. (*e.g.*, Figure 4.8 for only six markers).

To address this challenge, we developed RchyOptimyx, a computational tool that automatically characterizes the complex findings of high dimensional exploratory FCM studies. RchyOptimyx sorts all parent populations of an immunophenotype of interest into hierarchies, and selects those hierarchies that are better able to maintain the characteristics of the immunophenotype of interest (*e.g.,* correlation with a clinical outcome). This reveals the best order in which markers can be excluded from an immunophenotype. RchyOptimyx uses dynamic programming and efficient tools from graph theory to make the problem tractable using the computing resources readily available in most laboratories.

Since most cells can be described using more than one combination of markers, there usually are several alternative cellular hierarchies associated with every population. RchyOptimyx is able to find all these "paths" and merge them into a single hierarchy, starting from "all cells", or any arbitrary point in a hierarchy, and finishing at the terminal population of interest. This reveals the relationships between different gating strategies and how they differentiate, and also facilitates the reproduction of high–dimensional exploratory studies using low–color instruments. The ability to suggest multiple panels is particularly important when designing new panels, because the choice of markers depends on a large number of external parameters including, but not limited to, reagents available through vendors, potential spectral overlaps, the instruments available, and budget limitations.

Another important use-case for RchyOptimyx is in the interpretation of

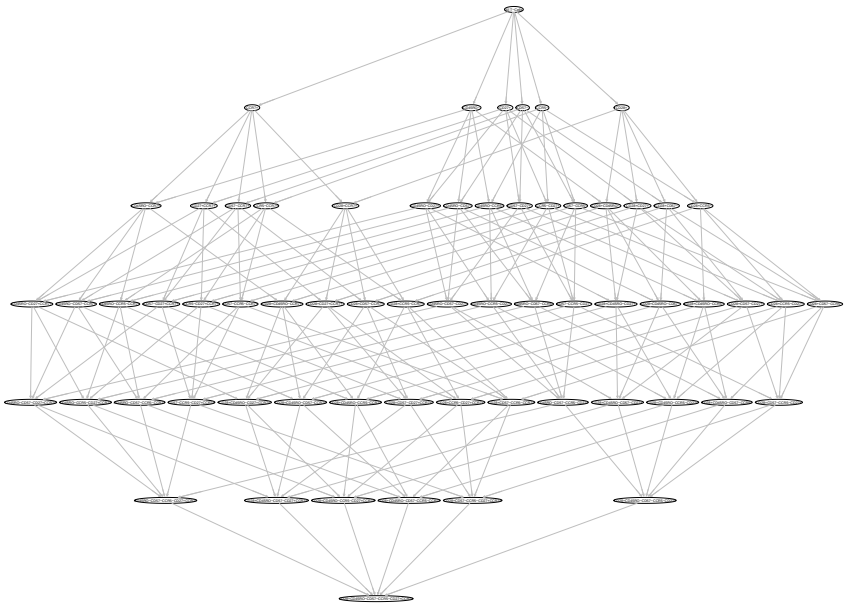


Figure 4.8: A complete cellular hierarchy for identifying naive T-cells. The colour of the nodes and the thickness of the edges have been removed to facilitate visualization of the complex graph.

the findings obtained from bioinformatics pipelines. While these pipelines have recently been very successful in identifying cell populations correlated with clinical outcomes, their results can be difficult fo understand for two reasons: 1) they usually rely on high-dimensional clustering of the data and therefore cannot propose gating strategies for reproduction of their results; 2) their predictive power often relies on a large list of immunophenotypes. Some of these immunophenotypes are closely related (*e.g.*, refer to close or overlapping cell populations) while others are not. RchyOptimyx addresses the first problem by suggesting optimized gating hierarchies for identification of these cell populations to a desired level of purity or correlation with clinical outcome. The latter problem is addressed by summarizing closely related immunophenotypes using their most important common parents.

In evaluating RchyOptimyx, we combined its functionality with the automated gating functionality provided by flowMeans and flowType. However, RchyOptimyx can be built upon the results of any cell population identification method, including manual analysis, provided all intermediate cell populations (*i.e.,* each layer, removing one marker at a time) from the cell population of interest up to the desired start of the hierarchy are provided to the algorithm.

We evaluated RchyOptimyx for three use-cases, using a small, but high–dimensional mass cytometry dataset and a clinical dataset of high-dimensional conventional FCM assays of 466 patients, previously analyzed by both manual and automated analysis. First, we constructed cellular hierarchies for identification of cells that were produced in response to different stimulations. This use-case represents the problem of designing panels of surface markers (primarily for sorting) for cells that can only be defined using their intra-cellular signature (possibly after proper stimulation). For example, plasmacytoid dendritic cell (PDC)s are known to express the toll-like receptor 9 (TLR9) in response to stimulation using CpG [63]. A large number of surface candidates were recently proposed for PDCs [18, 77, 110, 119]. An interesting direction to extend this work would be to measure all these markers in a single panel, subject to CpG stimulation (using appropriate controls) to design a panel of surface markers for PDCs. In this case, TLR9 could be used as the external variable for optimization.

Second, we demonstrated that RchyOptimyx can be used to simplify existing gating strategies (*e.g.,* the identification of naive T-cells previously defined using a complex panel of six markers to a 95% purity using only three). This proof-of-concept use-case is relevant when a subset of markers needs to be selected for reproduction of the results using fewer colors. For certain biological use–cases, purities higher than 95% can be required. For such use–cases, a larger number of markers for exclusion of non-naive T-cells should be included in the panel.

Third, we showed that RchyOptimyx, together with a complex bioinformatics pipeline, can analyze a large high-dimensional clinical dataset, to reveal correlates of a clinical outcome, hidden from previous manual and automated analysis of the same dataset. In addition, RchyOptimyx suggests the best gating strategies and marker panels for reproduction of these results in low-color settings. By identifying the best cellular hierarchies, RchyOptimyx allows the user to make an informed

decision about the trade-off between the number of markers and the significance of the correlation with the clinical outcome. This feature is particularly important in hypothesis generating studies that need to be further validated using large clinical studies.

For the third example, it is important to note that the correct measurement for the amount of correlation with a clinical outcome is an effect size (such as the root squared error of the estimated proportional hazard). However, such effect size does not provide any information about the significance of the correlation. As RchyOptimyx is intended to be a decision support tool, and in this case the decision is the degree to which a cell population can be generalized while maintaining the statistical significance of the correlation, we decided that the p-values of the log-rank tests are more appropriate for optimization of the hierarchies. To support this decision, we empirically investigated the differences between the p-values and effect sizes of the cox proportional hazard models (Figure 4.9) and concluded that these values are highly correlated (which is not surprising, given the large size of our cohort).

The concept of computationally extracting cellular hierarchies from FCM data has previously been introduced by the SPADE algorithm [11, 101]. SPADE generates a large number of multidimensional clusters and then connects them to each other using the distance between their mean/median fluorescence intensities. These are then manually annotated by biologists with domain knowledge. This makes SPADE useful for identification and visualization of a large number of clusters, particularly when expression of markers change gradually (*e.g.,* cell-cycle analysis and some intracellular studies). However, the hierarchies generated by SPADE are logically and conceptually different from those generated by RchyOptimyx and have different use cases. For example, the results of the mass cytometry dataset presented here are very close to results previously obtained from SPADE analysis. However, SPADE required manual annotation of the results by a human expert, using different plots demonstrating the expression of different surface markers and the intra–cellular marker of interest (Figure 2 and panel C of Figure 3 of [11]). More complicated relationships that involve several markers cannot be easily identified by these manual annotations. In addition, SPADE is limited in that the relationships between cell populations is exclusively defined

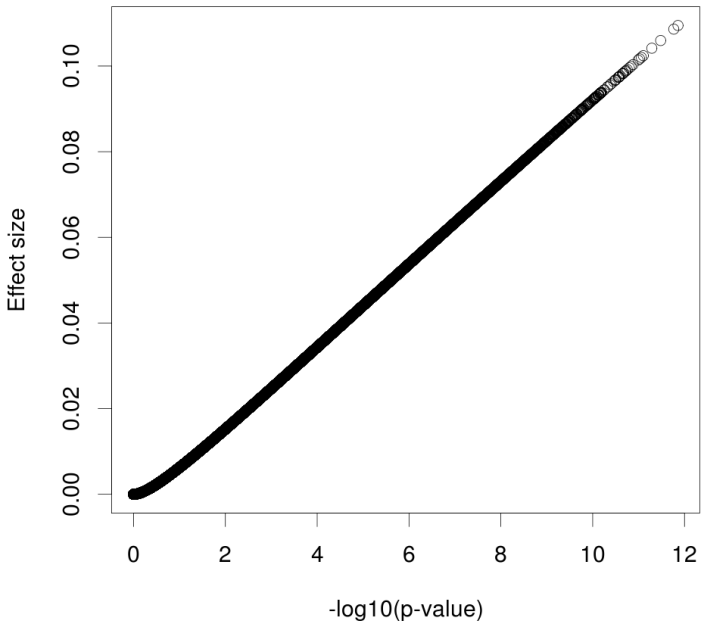


Figure 4.9: The correlation between effect sizes and p-values of the log rank tests for the cox proportional hazards models for each immunophenotypes. Pearson's correlation test: Correlation coefficient: 0.997, p-value $< 2.2e-16$.

using the multidimensional distances between them. However, two cell populations that are close to each other in the multidimensional space can be far in terms of specific markers (which can be the most important ones). The cellular hierarchies generated by RchyOptimyx are based on parent-child relationships, guided by an external variable (cell populations that have common parents with similar patterns of correlation with a clinical outcome or intracellular response to stimulation are grouped together). This enables RchyOptimyx to automatically annotate a large number of cell populations identified by other methods (*e.g.*, manual gating or SPADE) in terms of the importance of the markers involved and summarize them in a single hierarchy.

There are several directions in which this work can be extended. RchyOptimyx provides no information about the robustness of the hierarchies. Bootstrapping strategies could be used to produce confidence intervals for the tree structure and increase generalizability to previously unseen data [118]. Also, our current implementation of RchyOptimyx assumes that every marker can be partitioned into a positive and negative population. While the underlying theory does support additional (*e.g.*, dim, bright, or low) populations, parts of the software package would need to be modified to accommodate these cases.

Chapter 5

FlowCAP: Critical Assessment of Automated Flow Cytometry Data Analysis Techniques

5.1 Introduction

Beginning in 2007 there has been a renaissance in the development of computational methods for FCM data in an effort to overcome the continued limitations in manual gating-based analysis with successful results reported in each case (see, *e.g.*, Chapter 2, 3, and [2, 3, 5, 11, 27, 38, 44, 70, 85, 97, 99, 101, 102, 105, 116, 117, 128, 129]). However, it has been unclear how the results from these state–of–the–art approaches compared with traditional manual gating results in general, how they could be used to discover new cell populations of interest, and how these computation methods compared with each other, as every new algorithm was assessed using distinct datasets and evaluation methods. To address these shortcomings, members of the algorithm development, FCM user, and software and instrument vendor communities initiated the *Flow Cytometry: Critical Assessment of Population Identification Methods (FlowCAP)* project. The goal of FlowCAP is to advance the development of computational methods for the identification of cell populations of interest in FCM data by providing the means to objectively test and

compare these methods, and to provide guidance to the end user about how best to use these algorithms. Here we report the results from the first two FlowCAP-sponsored competitions, which evaluated the ability of automated approaches to address two important use cases - cell population identification and sample classification.

5.2 Cell Population Identification

5.2.1 Structure of the Challenges

Algorithms competed in four challenges for cell population identification:

1. **Completely Automated:** The goal of this challenge was to compare automated gating algorithms for exploratory analysis. Software used in this challenge either did not have any tuning parameters, or if there were tuning parameters, the values were fixed in advance and used across all datasets.

2. **Manually Tuned:** The goal of this challenge was to compare semi-automated gating algorithms that were permitted to have manually adjusted parameters (*i.e.*, participants were allowed to supply results from their algorithm with parameters tuned for individual datasets).

3. **Assignment of Cells to Populations with Pre-defined Number of Populations:** The goal was to compare the ability of the algorithms to assign correct labels to cells when the number of expected populations was known.

4. **Supervised Approaches Trained using Human-Provided Gates:** In this challenge, 25% of the files with manual gates (*i.e.*, membership labels) were provided to participants for training/tuning their algorithms for each dataset. The results were evaluated using the complete dataset.

Five datasets were used for these challenges (the markers evaluated are listed in Table 5.1):

1. **Diffuse Large B-cell Lymphoma (DLBCL)** The DLBCL dataset consists of data from 30 randomly selected lymph node biopsies from patients treated

at the British Columbia Cancer Agency between 2003 and 2008. These patients were histologically confirmed to have diffuse large B-cell lymphoma (DLBCL). This dataset was provided by the BCCRC[1].

2. **Symptomatic West Nile Virus (WNV)** Samples are human peripheral blood mononuclear cells from patients with symptomatic West Nile virus infection stimulated *in-vitro* with peptide pools representing different regions of the WNV polyprotein. This dataset was provided by McMaster University[2].

3. **Normal Donors (ND)** The investigators examined differences in the response of a variety of cell-types to various stimuli for a set of healthy donors. For the samples used here, the time-periods were relatively short, such that the surface cell-type markers would not be expected to change. The staining panel contains antibodies to surface markers and intracellular proteins. Note that these experiment were done with phosflow-fixed cells, and thus some of the populations are not as distinct or clean as would be seen with other processing methods. This dataset was provided by Amgen Inc.[3] .

4. **Hematopoietic Stem Cell Transplant (HSCT)** This dataset contains data from 30 randomly selected samples derived from hematopoietic stem cell transplant experiments done in the Terry Fox Laboratory. This dataset was provided by the BCCRC [4].

5. **Graft versus Host Disease (GvHD)** Twelve FCM samples for finding cellular signatures to predict or correlate with early detection of GvHD. This dataset was provided by the BCCRC and Treestar Inc.[5] and Treestar Inc.[6].

The following pre-processing steps were applied to these datasets before providing them to the participants: (1) compensation (to account for the overlap of emission spectra from antibody fluorescent labels); (2) transformation to linear

[1]Andrew P. Weng: aweng@bccrc.ca
[2]Jonathan Bramson: bramsonj@mcmaster.ca
[3]Hugh Rand: rand@amgen.com
[4]Connie Eaves: ceaves@bccrc.ca
[5]Ryan Brinkman: rbrinkman@bccrc.ca
[6]Jill Schoenfeld: jill@treestar.com

Table 5.1: Summary of the description of the datasets.

Dataset	#Samples	#Events	Analyte	Detector	Reporter	Provided By
GvHD	12	14,000	CD4	Anti-CD4	FITC	BCCRC
			CD8b	Anti-CD8b	PE	&
			CD3	Anti-CD3	PerCP	TreeStar
			CD8	Anti-CD8	APC	
DLBCL	30	5,000	CD3	Anti-CD3	CY5	BCCRC
			CD5	Anti-CD5	FITC	
			CD19	Anti-CD19	PE	
ND	30	17,000		Proprietary	FITC	Amgen
				Proprietary	PerCPCy5	
				Proprietary	PacificBlue	
				Proprietary	PacificOrange	
			CD56	Anti-CD56	Qdot605	
				Proprietary	APC	
			CD8	Anti-CD8	Alexa700	
				Proprietary	PE	
			CD45	Anti-CD45	PECy5	
			CD3/CD14	Anti-CD3/CD14	PECy7	
WNV	13	100,000	IFNγ	Anti-IFNγ	PEA	McMaster
			CD3	Anti-CD3	PECy5	
			CD4	Anti-CD4	PECy7	
			IL17	Anti-IL17	APC	
			CD8	Anti-CD8	AlexaFluor700	
			Free Amines	NA	CFSE	
HSCT	30	10,000	CD45.1	Anti-CD45.1	FITC	BCCRC
			Ly65/Mac1	Anti-Ly65/Mac1	PE	
			Dead Cells	NA	PI	
			CD45.2	Anti-CD45.2	APC	

space (to scale data appropriately for visualization); (3) pre-gating for removal of irrelevant cells (*e.g.,* dead cells as performed by the human analysts).

For these challenges, cell population membership defined by each algorithm was compared against cell population membership defined by manual gating performed by the data set provider in order to compare algorithm results with the current standard practice for FCM data analysis. The F-measure statistic (see the *Methods* section for a detailed description) was used for this comparison in order to consider both sensitivity and specificity of the automated method. An F-measure of 1.0 indicates perfect recapitulation of the manual gating result with no false positive or false negative cells.

5.2.2 Clustering F-measure

F-measure is the harmonic mean of the sensitivity and specificity of an algorithm. It can be written as $F = (2 \cdot Se \cdot Sp)/(Se + Sp)$, where Se (sensitivity) is the number of cells correctly assigned to a cluster divided by all the cells that should have been assigned to that cluster, and Sp (specificity) is the number of cells correctly assigned to a cluster divided by the total cells assigned to that cluster.

Given a correct set of reference clusters $C = \{c_1, c_2, ..., c_n\}$ and a clustering result $K = \{k_1, k_2, .., k_m\}$, the number of matches between combinations of C and K is a matrix, $M = [a_{ij}]$, where $i \in \{1, ..., n\}$ and $j \in \{1, ..., m\}$. Then $Se(c_i, k_j) = a_{ij}/|k_j|$ and $Sp(c_i, k_j) = a_{ij}/|c_i|$, where $|c_i|$ denotes the number of elements in c_i. The F-measure to compare one cluster to another is then $F(c_i, k_j) = 2 \cdot Se(c_i, k_j) \cdot Sp(c_i, k_j)/(Se(c_i, k_j) + Sp(c_i, k_j))$. To calculate the F-measure of an entire clustering result, for each cluster c_j in the reference, a set of F-measures against every predicted cluster k_j is calculated, and the largest F-measure (best match), normalized by the size of k_j is reported. The sum of these scores produced a total F measure, defined as $F(C, K) = \sum_{c_i \in C} \frac{|c_i|}{N} \cdot \max_{k_j \in K} \{F(c_i, k_j)\}$. F-measure values are always in the interval $[0, 1]$, with 1 indicating a perfect prediction. See [1] for a comparison of F-measure versus other metrics for evaluation of clustering algorithms.

While mean F-measures can be used to assess the performance of each of the algorithms on each dataset, the significance of the difference in the F-measures values must be be accounted for in order to truly rank the algorithms. Therefore, to measure how significant these differences were (*i.e.*, how sensitive they are to this specific set of samples), bootstrapping was used to compute 95% confidence intervals. Algorithms with overlapping CIs were subsequently considered tied (bolded in Table 5.3).

5.2.3 Rank score

To derive an overall ranking of the algorithms, we used their rank score calculated as the sum of fractional rankings of each algorithm across different datasets. Fractional ranking is based on the Borda count strategy [30]: For N algorithms, the top algorithm scored N points, the second one $N-1$ points, and so on. The last algorithm scored 1 point. The average number of points was used in case of ties (*i.e.*, overlapping CIs). For D datasets, rank score values are in the $[D, N \times D]$ interval; an algorithm that scored first in every dataset would have a rank equal to $N \times D$.

5.2.4 Algorithm Performance

FlowCAP received a total of 36 submissions from 14 research groups (Tables 5.2 and 5.3). Not all algorithms competed in all challenges. For example, supervised classification methods, like Radial SVM, require training data to establish classification rules, and therefore were not appropriate for Challenges 1–3. In each challenge, the submitted algorithms were sorted by their rank score (described in the *Methods* section). Many algorithms performed well in multiple challenges on multiple datasets, with F-measures exceeding 0.85. Some of the algorithms were always in the top group (*i.e.,* were not significantly different from the top algorithm), some were in the top group for some of the datasets, and some were never in the top group.

Table 5.2: Brief description of the methodologies used by the algorithms, their software platforms (if applicable), as well as citations.

Algorithm Name	Availability	Brief Description	Ref
Cell Population Identification			
ADICyt	Commercially Available	Hierarchical clustering and entropy–based merging	-
CDP	Python source–code	Bayesian non-parametric mixture models, calculated using massively parallel computing on GPUs	[20]
FLAME	R package	Multivariate finite mixtures of skew and heavy-tailed distributions	[97]
FLOCK	C source–code	Grid-based partitioning and merging	[99]
flowClust/Merge	Two R/BioC packages	t-mixture modeling and entropy-based merging	[38, 70]
flowKoh	R source–code	Self-organizing maps	-
flowMeans	R/BioC package	k-means clustering and merging using the Mahalanobis distance	[2]
FlowVB	Python source–code	t-mixture models using variational Bayes inference	-
L2kmeans	JAVA source–code	Discrepancy learning	[36]
MM, MM&PCA	Windows and Linux executable	Density-based misty mountain clustering	[117]
NMF-curvHDR	R source–code	Density-based clustering and non-negative matrix factorization	[85]
Radial SVM	MATLAB source–code	Supervised training of radial SVMs using example manual gates	[102]
SamSPECTRAL	R/BioC package	Efficient spectral clustering using density-based down-sampling	[128]
SWIFT	MATLAB source–code	Weighted iterative sampling and mixture modeling	[83]
Ensemble Clustering	R/CRAN package	Combines the results of all of the participating algorithms	[55, 56]
Sample Classification			
2DhistSVM	Pseudocode	2D histograms of all pairs of dimensions and support vector machines	-
admire-lvq	MATLAB source–code	1D features and learning vector quantization	-
biolobe	Pseudocode	k-means and correlation matrix mapping	-
daltons	MATLAB source–code	Linear discriminant analysis and logistic regression	-
EMMIXCYTOM	R source–code	Skew-t-mixture model and KullbackLeibler divergence	-
DREAM–A	Pseudocode	2&3D histograms and cross-validation of several classifiers	-
DREAM–B	Pseudocode	1D Gaussian mixtures and support vector machines	-
DREAM–C	Pseudocode	1D gating and several different classifiers	-
DREAM–D	Pseudocode	4D clustering and bootstrapped t-tests	-
FiveByFive	Pseudocode	1D histograms and support vector machines	-
flowBin	R package	High-dimensional cluster mapping across multiple tubes and support vector machines	-
flowCore-flowStats	R source–code	Sequential gating and normalization and a Beta-Binomial model	[49]
flowPeakssvm Kmeanssvm	R package	Kmeans and density-based clustering and support vector machines	[44]
flowType FeaLect	Two R/BioC packages	1D gates extrapolated to multiple dimensions and bootstrapped LASSO classification	[3, 129]
JKJG	JAVA source–code	1D Gaussian and logistic regression	-
PBSC	C source–code	Multi-dimensional clustering and cross sample population matching using a relative distance order	[99]
PRAMS	R source–code	2D Clustering and logistic regression	-
PramSpheres and CIHC	Pseudocode	Genetic algorithm and gradient boosting	-
RandomSpheres	Pseudocode	Hypersphere–based Monte Carlo optimization	-
SPADE, BCB	MATLAB, Cytoscape, R/BioC	Density-based sampling, kmeans clustering, and minimum spanning trees	[101]
SPCA+GLM	Pseudocode	1D probability binning and principal component analysis	-
SWIFT	MATLAB source–code	SWIFT clustering and support vector machines	[83]
team21	Python source–code	1D relative entropies	-
uqs	Pseudocode	Skew–t-mixture models and KullbackLeibler divergence	-

Table 5.3: Mean and 95 percent CIs for the F-Measures, Rank Scores, and runtimes of the cell population identification algorithms. In each dataset/challenge, the top algorithm (highest mean F-measure) and the algorithms with overlapping CIs with the top algorithm are bolded. Algorithms are sorted by rank score within each challenge (see methods for detailed description of the rank score). Runtime was calculated as time per CPU per sample.

	F-measure						Runtime	Rank
	GvHD	DLBCL	HSCT	WNV	ND	Mean	hh:mm:ss	Score
Challenge 1: Completely Automated								
ADICyt	**0.81 (0.72, 0.88)**	**0.93 (0.91, 0.95)**	**0.93 (0.90, 0.96)**	**0.86 (0.84, 0.87)**	**0.92 (0.92, 0.93)**	0.89	04:50:37	52
flowMeans	**0.88 (0.82, 0.93)**	**0.92 (0.89, 0.95)**	**0.92 (0.90, 0.94)**	**0.88 (0.86, 0.90)**	0.85 (0.76, 0.92)	0.89	00:02:18	49
FLOCK	**0.84 (0.76, 0.90)**	**0.88 (0.85, 0.91)**	0.86 (0.83, 0.89)	**0.83 (0.80, 0.86)**	0.91 (0.89, 0.92)	0.86	00:00:20	45
FLAME	**0.85 (0.77, 0.91)**	**0.91 (0.88, 0.93)**	**0.94 (0.92, 0.95)**	0.80 (0.76, 0.84)	0.90 (0.89, 0.90)	0.88	00:04:20	44
SamSPECTRAL	**0.87 (0.81, 0.93)**	0.86 (0.82, 0.90)	0.85 (0.82, 0.88)	0.75 (0.60, 0.85)	**0.92 (0.92, 0.93)**	0.85	00:03:51	39
MM&PCA	**0.84 (0.74, 0.93)**	0.85 (0.82, 0.88)	**0.91 (0.88, 0.94)**	0.64 (0.51, 0.71)	0.76 (0.75, 0.77)	0.80	00:00:03	29
FlowVB	**0.85 (0.79, 0.91)**	0.87 (0.85, 0.90)	0.75 (0.70, 0.79)	0.81 (0.78, 0.83)	0.85 (0.84, 0.86)	0.82	00:38:49	28
MM	**0.83 (0.74, 0.91)**	**0.90 (0.87, 0.92)**	0.73 (0.66, 0.80)	0.69 (0.60, 0.75)	0.75 (0.74, 0.76)	0.78	00:00:10	28
flowClust/Merge	0.69 (0.55, 0.79)	0.84 (0.81, 0.86)	0.81 (0.77, 0.85)	0.77 (0.74, 0.79)	0.73 (0.58, 0.85)	0.77	02:12:00	24
L2kmeans	0.64 (0.57, 0.72)	0.79 (0.74, 0.83)	0.70 (0.65, 0.75)	0.78 (0.75, 0.81)	0.81 (0.80, 0.82)	0.74	00:08:03	20
CDP	0.52 (0.46, 0.58)	0.87 (0.85, 0.90)	0.50 (0.48, 0.52)	0.71 (0.68, 0.75)	0.88 (0.86, 0.90)	0.70	00:00:57	19
SWIFT	0.63 (0.56, 0.70)	0.67 (0.62, 0.71)	0.59 (0.55, 0.62)	0.69 (0.64, 0.74)	0.87 (0.86, 0.88)	0.69	01:14:50	15
Ensemble Clustering	0.88	0.94	0.97	0.88	0.94	0.92	-	64
Challenge 2: Manually Tuned								
ADICyt	**0.81 (0.71, 0.89)**	**0.93 (0.91, 0.95)**	**0.93 (0.90, 0.96)**	**0.86 (0.84, 0.87)**	**0.92 (0.92, 0.93)**	0.89	04:50:37	34
SamSPECTRAL	**0.87 (0.79, 0.94)**	**0.92 (0.89, 0.94)**	0.90 (0.86, 0.93)	**0.85 (0.83, 0.88)**	**0.91 (0.91, 0.92)**	0.89	00:06:47	31
FLOCK	**0.84 (0.76, 0.90)**	0.88 (0.85, 0.91)	0.86 (0.83, 0.89)	**0.84 (0.82, 0.86)**	0.89 (0.87, 0.91)	0.86	00:00:15	23
FLAME	**0.81 (0.75, 0.87)**	0.87 (0.84, 0.90)	0.87 (0.82, 0.90)	**0.84 (0.83, 0.85)**	0.87 (0.86, 0.87)	0.85	00:04:20	23
SamSPECTRAL-FK	**0.87 (0.80, 0.94)**	0.85 (0.81, 0.89)	0.90 (0.86, 0.92)	0.76 (0.71, 0.81)	**0.92 (0.91, 0.93)**	0.86	00:04:25	23
CDP	**0.74 (0.67, 0.80)**	0.89 (0.86, 0.91)	0.90 (0.88, 0.92)	0.75 (0.71, 0.78)	0.86 (0.85, 0.88)	0.83	00:00:18	19
flowClust/Merge	0.69 (0.53, 0.78)	0.87 (0.85, 0.90)	**0.96 (0.94, 0.97)**	0.77 (0.75, 0.79)	0.88 (0.81, 0.91)	0.83	02:12:00	18
NMF-curvHDR	**0.76 (0.69, 0.82)**	0.84 (0.83, 0.86)	0.70 (0.67, 0.74)	0.81 (0.77, 0.84)	0.83 (0.83, 0.84)	0.79	01:39:42	13
Ensemble Clustering	0.87	0.94	0.98	0.87	0.92	0.91	-	41
Challenge 3: Assignment of Cells to Populations with Pre-defined Number of Populations								
ADICyt	**0.91 (0.84, 0.96)**	**0.96 (0.94, 0.97)**	**0.98 (0.97, 0.99)**			0.95	00:10:49	26.2
SamSPECTRAL	**0.85 (0.75, 0.93)**	**0.93 (0.91, 0.95)**	**0.97 (0.95, 0.98)**			0.92	00:02:30	26.2
flowMeans	**0.91 (0.84, 0.96)**	**0.94 (0.91, 0.96)**	0.95 (0.93, 0.96)			0.93	00:00:01	23.4
TCLUST	**0.93 (0.91, 0.96)**	**0.93 (0.91, 0.95)**	0.93 (0.90, 0.95)			0.93	00:00:40	23.4
FLOCK	**0.86 (0.79, 0.93)**	0.92 (0.89, 0.94)	**0.97 (0.95, 0.98)**			0.92	00:00:02	22.2
CDP	**0.85 (0.77, 0.92)**	**0.92 (0.89, 0.94)**	0.76 (0.72, 0.81)			0.84	00:00:21	16.9
flowClust/Merge	**0.88 (0.82, 0.93)**	0.90 (0.86, 0.94)	0.83 (0.79, 0.88)			0.87	00:49:24	15.9
FLAME	**0.85 (0.79, 0.91)**	0.90 (0.86, 0.93)	0.86 (0.82, 0.91)			0.87	00:03:20	15.9
SWIFT	**0.90 (0.84, 0.95)**	0.00 (0.00, 0.00)	0.88 (0.84, 0.92)			0.59	00:01:37	11.9
flowKoh	0.85 (0.80, 0.90)	0.85 (0.82, 0.88)	0.87 (0.84, 0.91)			0.86	00:00:42	9.5
NMF	0.74 (0.69, 0.78)	0.84 (0.80, 0.88)	0.80 (0.76, 0.84)			0.79	00:01:00	7.5
Ensemble Clustering	0.95	0.97	0.98			0.97	-	35.0
Challenge 4: Supervised Approaches Trained using Human-Provided Gates								
Radial SVM	**0.89 (0.83, 0.95)**	0.84 (0.80, 0.87)	**0.98 (0.96, 0.99)**	**0.96 (0.94, 0.97)**	**0.93 (0.92, 0.94)**	0.92	00:00:18	21
flowClust/Merge	**0.92 (0.88, 0.95)**	**0.92 (0.89, 0.94)**	**0.95 (0.92, 0.97)**	0.84 (0.82, 0.86)	0.89 (0.88, 0.90)	0.90	05:31:50	19
randomForests	**0.85 (0.78, 0.91)**	0.78 (0.74, 0.83)	0.81 (0.79, 0.83)	0.87 (0.84, 0.90)	**0.94 (0.92, 0.95)**	0.85	00:02:06	15
FLOCK	0.82 (0.77, 0.87)	**0.91 (0.89, 0.93)**	0.86 (0.76, 0.93)	0.86 (0.82, 0.89)	0.86 (0.77, 0.92)	0.86	00:00:05	13
CDP	0.78 (0.68, 0.87)	**0.95 (0.93, 0.97)**	0.75 (0.71, 0.78)	0.86 (0.84, 0.88)	0.83 (0.80, 0.86)	0.83	00:00:15	11
Ensemble Clustering	0.91	0.94	0.95	0.92	0.94	0.93	-	26

Allowing participants to tune algorithm parameters did not result in much improvement, as the highest overall F-measure did not increase (0.89 for both completely automated and manually tuned algorithms); only three of the six algorithms that participated in both Challenge 1 and Challenge 2 demonstrated an improvement in overall F measure, and these improvements were modest. In some cases the F-measures actually decreased after human intervention (*e.g.*, FLAME). In contrast, providing the number of cell populations sought in Challenge 3 made predictions more accurate for seven of the eight algorithms that participated in both Challenge 1 and Challenge 3, with five algorithms achieving overall F-measures greater than 0.9. In addition, providing a set of example results for algorithm training and parameter tuning in Challenge 4 improved the results of flowClust/Merge by 0.13. With example results for training, the Radial SVM approach outperformed the algorithms used in Challenge 1 in four of the five datasets. Taken together, these results suggest that estimating the correct number of cell populations (as defined by manual gates) remains a challenge for most automated approaches. Providing several examples as a training-set improves this situation. However, not many of the existing algorithms can support training-sets; hence, the low number of participants in Challenge 4.

The "Runtime" column of Table 5.3 shows the estimated runtimes per sample of the algorithms on single core CPUs or GPUs (for CDP only). Runtimes ranged from 1 second to more than 4 hours per sample. ADICyt, which had the highest rank score in the first three challenges, also required the longest runtimes. flowMeans, FLOCK, FLAME, SamSPECTRAL, and MM&PCA needed substantially shorter runtimes and still performed reasonably well in comparison with ADICyt. Note that, due to hardware and software differences, these numbers may not be precisely comparable; the information is provided here to give some sense of the differences in time requirements of these specific algorithm implementations.

5.2.5 Combining Predictions

Similar to other data analysis settings (see [126] for a review), combining results from different cell population identification methods provides improved accuracy

over any individual method. The last row of each challenge's section in Table 5.3 shows the results obtained by combining the results of all the submitted algorithms (Ensemble Clustering). For all four challenges, this ensemble method resulted in a higher overall F-measure and rank score than any of the individual algorithms (Table 5.3).

Methods

The consensus clustering problem is defined as follows: given a set of partitions (the ensemble), find a new partition P that minimizes the dissimilarity between P and participating partitions. A partition M is defined as a binary matrix with each column corresponding to a class label. The dissimilarity between a partition P and a partition element of the ensemble Q is defined as

$$d(P,Q) = \min_{\Pi} ||P - Q \times \Pi||_p$$

where $||\cdot||_p$ is the entry-wise p-norm. The permutation matrix provides a mapping between corresponding classes. For example given three observations x, y, z, one partition may label the observations as $x \in A, y \in B, z \in C$ and another may label the observations (with independent labels) as $y \in \alpha, x \in \gamma, c \in \gamma$. The partitions are in fact the same if we consider the classes as $A = \gamma, B = \alpha, C = \gamma$. The permutation matrix Π determines how the classes in P correspond to the classes in Q. When $p = 1$, the measure is known as the Manhattan distance. This distance can be calculated efficiently using linear programming methods. Once a dissimilarity measure is defined, in our case, the Manhattan distance with $p = 1$, we must solve the harder problem of finding the partition P^* that minimizes the distance for all of the partitions Q in the ensemble E.

$$P^* \in \operatorname*{argmin}_{P} \sum_{Q \in E} \min_{\Pi} ||P - Q\Pi||_1.$$

This is a known NP-hard problem (Multi-dimensional Assignment) so we used a heuristic method, described by Hornik [56], that provides approximate solutions for the consensus partition problem. The `clue` package [55] includes an implementationl of this heuristic.

Results

For all of the four challenges, this ensemble method resulted in a higher overall F-measure and rank score than any of the individual algorithms (Figure 5.2). In addition, ensemble clustering gave a higher F-measure for each of the individual datasets in each challenge, with only three exceptions in Challenge 4 (Figures 5.1, 5.2, and 5.3).

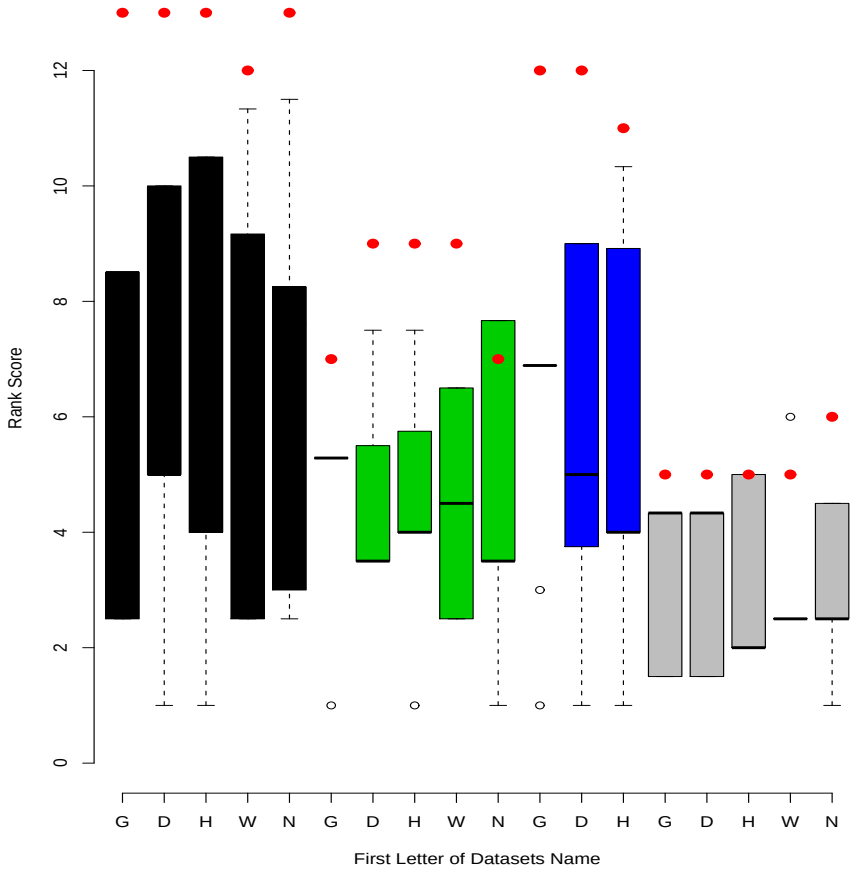


Figure 5.1: Rank scores of all individual algorithms (box plots) compared with the ensemble clustering (red dots) in each dataset and challenge.

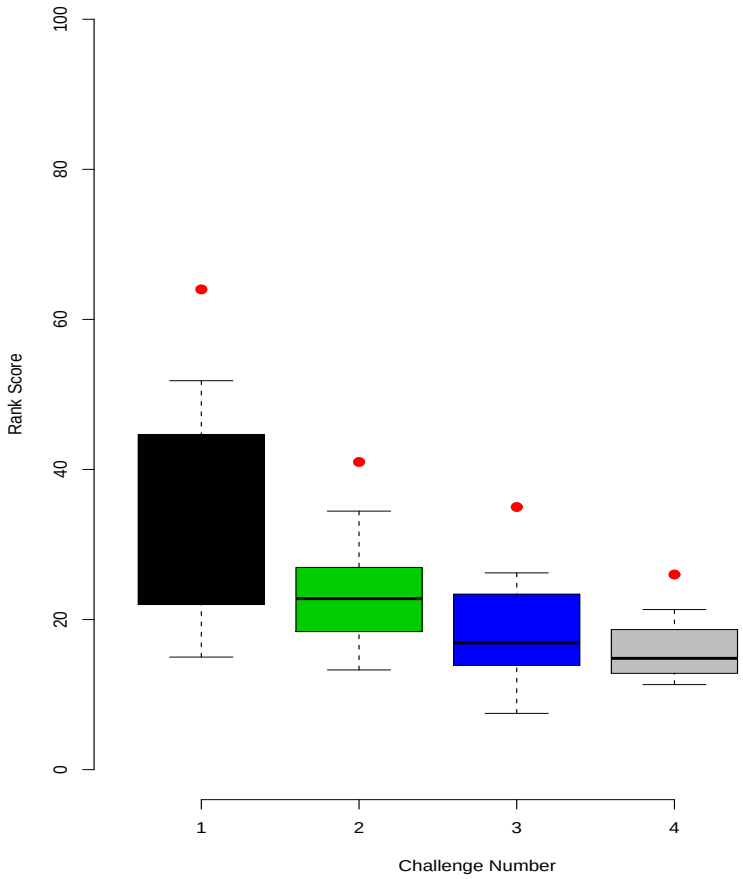


Figure 5.2: Rank scores of all individual algorithms (box plots) are compared with the ensemble clustering (red dots) across all challenges.

Ablation Analysis

We also investigated whether all of the algorithms are required to be included in the ensemble clustering to ensure a high F-measure. Figure 5.4 shows the change in F-measure as each algorithm was removed from the ensemble cluster in order of their relative contribution, with the algorithm contributing the least to the ensemble clustering results removed first. For example in Challenge 3, when

only 4 algorithms were included in the ensemble (*i.e.,* TCLUST, ADICyt, FLAME, and SWIFT), the F-measure was still close to 0.95. Including two more algorithms to the set resulted in a minor improvement, and after that, no improvements were observed. Similar patterns were observed in the other challenges.

Although the absolute order differed in the ablation analysis, algorithms with higher F-measures tended to be removed later (*i.e.*, they contributed mroe to the ensemble). For example, in Challenge 1 and 2 the top 2 algorithms were removed last. Interestingly, in the 8th iteration, where only 5 algorithms are left in the ensemble, the F-measure dropped dramatically indicating that even algorithms that individually perform rather poorly can contribute to a good ensemble result.

We also performed the ablation analysis in the reversed order (*i.e.,* the algorithm with maximum contribution was removed first). Figure 5.5 demonstrates the results. As expected, the algorithms with a higher F-measure tend to be excluded earlier (confirming that they have contributed more to the ensemble).

5.2.6 Results with Refined Manual Gates

Without detailed guidance on the goals of FlowCAP, the data providers tended to focus gating only on those populations of interest for their work and therefore provided incomplete population delineation in many cases. In addition, by relying on the single set of gates completed by the data providers, inconsistencies in manual gating by different analysts were not taken into account. To address these deficiencies, the HSCT and GvHD datasets were provided to eight individuals from five different institutions who were instructed to try and identify all cell populations discernible from the available data. These datasets were selected since they had the highest and lowest overall F-measures across all algorithms representing the best/worst cases for the algorithms.

A reference for evaluation of the algorithms was generated using a consensus of all manual gates using ensemble clustering was calculated. Consider the mean F-measure of each population in the consensus (across all of the manual gates). This score provides a measurement for the amount of agreement among human experts on every cell population in the consensus. Prevalent cell populations, in terms of both absolute cell count and proportions, tended to have higher F-measures. Rare

cell populations were more variable in classification consistency between manual gaters. However, the cell populations with a high score also included a wide range of cell frequencies (Figures 5.6, 5.7, and 5.8).

The consensus of manual gates was used to rank the algorithms. Comparison of the algorithms started by the cell population in the entire dataset with the best match across all manual gates and then gradually expanded to more cell populations with a weaker match across the human analysts (Figure 5.9). Including the cell populations with lower agreement across the human experts resulted in a gradual reduction in F-measures of both manual gates and algorithms, suggesting that certain populations were more difficult to resolve by both manual and automated analysis, especially for the GvHD dataset. However, the overall performance of algorithms for both datasets using these multiple sets of exhaustive gates was generally consistent with our initial results (Table 5.3).

As an alternative to the overall F-measures, the reference clusters were used in a per-population analysis to determine if certain cell populations were responsible for high or low F-measures for the algorithms. Human consensus results were matched across samples to the sample with maximum number of populations. Then, the human consensus for each sample was used as a reference for matching of the automated results of that sample. Pairwise F-measures between all algorithms and manual gates for the HSCT dataset are shown in Figure 5.10. The dendrograms were calculated using the complete-linkage clustering algorithm and the Euclidean distance between the F-measures [114]. These results can be used to identify cell populations that are responsible for high (or low) F-measures for further visual investigation. For example, cell population #3 of Figure 5.10 demonstrates high pairwise F-measures between all of the algorithms and manual gates, suggesting that this cell population was correctly identified by most of the algorithms and manual gates (Figure 5.11).

Panel B of Figure 5.12, however, represents a cell population that has only been identified by manual gating. Figure 5.13 shows that this population (colored in red) is generally identical to the cyan population in every channel but has a lower FSC. This emphasizes the importance of designing methodologies that can use background biological knowledge in the clustering process. In this case, the human experts used their knowledge about the scatter channels to partition these

cells into two different populations despite their similarity in every other channel.

The per-population analysis suggested that some algorithms had better matches with the manual analysis for each population, but importantly, the best-matching algorithms were not always the same for each population. This suggests that different algorithms may have different abilities to resolve populations depending on the exact structure of the data, which is not surprising given the wide range of strategies utilized by the different algorithms. This may also explain why the ensemble analysis matched the manual consensus more closely than any of the individual algorithms for all cell populations.

5.3 Sample Classification

5.3.1 Structure of the Challenges

Another important use case for FCM analysis is the use of biomarker patterns in FCM data for the purposes of sample classification. We assembled a benchmark of three datasets in which the subjects/samples were associated with an external variable that could be used as an independent measure of truth for sample classification. The benchmark consisted of three datasets for: (1) studying the effect of HIV exposure on 44 African infants using 6 tubes of 8 color assays (HIV-exposed *in utero*, but uninfected (HEU) vs. unexposed (UE)); (2) diagnosis of acute myeloid leukemia (AML) using 8 tubes of 5 color assays on 359 subjects provided by a reference laboratory (AML vs. non-AML); (3) discriminating between two antigen stimulation groups of post-HIV vaccine T-cells using two tubes of 8 color assays on 48 subjects (Gag-stimulated vs. Env-stimulated): For each dataset, half of the correct sample classifications were provided to the participants for training purposes. The other half of the data was used as an independent cohort for testing/validation. For the AML dataset, additional results where submitted through the DREAM (Dialogue for Reverse Engineering Analysis and Methods) [17, 79, 96, 115] initiative.

Challenge 1: HIV-Exposed-Uninfected versus Un-exposed

The goal of this challenge was to find cell populations that can be used to discriminate between HEU ($n = 20$) and UE ($n = 24$) infants. Blood samples were taken at 6 months after birth and were left unstimulated (for control) or stimulated with 6 Toll-like receptor molecules. In addition to raw FCS files, half of the subject labels were provided for training purposes. Algorithms had to use this data to label the rest of the samples. These labels were then used to evaluate your algorithms performance.

Challenge 2: Acute Myeloid Leukaemia

The goal of this challenge was to find cell populations that can be used to discriminate between AML positive ($n = 43$) and healthy donor ($n = 316$) patients. Peripheral blood or bone marrow aspirate samples were collected over a 1 year period using 8 tubes (tube #1 is an isotype control and #8 is unstained) with different marker combinations. In addition to raw FCS files, half of the subject labels were provided for training purposes. Your algorithm must use this data to label the rest of the samples. These labels will be used to evaluate your algorithms performance.

Challenge 3: Identification of Antigen Stimulation Group of Intracellular Cytokine Staining of Post-HIV Vaccine Antigen Stimulated T-cells.

The goal of this challenge was to correctly label the antigen stimulation group of post-HIV vaccine T-cells. The data set contains samples from 48 individuals (column pub-id in the metadata). Each individual received an experimental HIV vaccine. Samples were collected approximately 10 months later and T-cells challenged with two antigens *ENV-1-PTEG* and *GAG-1-PTEG*, column antigen in the metadata). The response of $CD4^+$ and $CD8^+$ T-cells was measured by flow cytometry for each of these groups. The cells were found to respond differently to the two antigen stimulations. This was essentially a classification challenge. For training purposes we provided data from 24 individuals within each group. The antigen stimulation label was provided (column antigen in the metadata). The testing data ($n = 24$) did not have an antigen stimulation group

label. Participants had to correctly identify the antigen stimulation group of the test data. The complete data set consisted of 240 FCS files. The data was compensated, transformed and partially gated (gated for singlets, live cells and lymphocytes). We note that the data set contained positive and negative controls (sebctrl, negctrl) which were not part of this challenge, and do not need to have an antigen group label assigned to them to complete the challenge. Only the metadata rows where the antigen code is missing had to be labelled correctly.

5.3.2 Classification F-measure

F-measure for classification is defined as the harmonic mean of sensitivity and specificity (the additional "matching" step for clustering F-measure is not required). Sensitivity was defined as $\frac{TP}{TP+FN}$ and specificity is defined as $\frac{TN}{TN+FP}$, where TP, TN, FP, and FN are true positives (*e.g.,* and AML predicted as AML), true negatives, false positives, and false negatives, respectively.

5.3.3 Algorithm Performance

A total of 43 submissions we received (as noted in Table 5.2). Fourteen of these submissions were through the DREAM project. The sensitivity, specificity, accuracy, and F-measure values on the testset (Table 5.4)show that for two of the datasets (AML and HIV Vaccine Trials Network (HVTN)) many algorithms were able to perfectly predict the external variables even under very conservative conditions (*i.e.*, using an independent test set as large as the trainingset). In the first challenge, despite mostly accurate predictions on the trainingset, none of the algorithms performed strongly on the testset.

5.3.4 Outlier Analysis

In all datasets, the misclassifications were uniformly distributed across the testsets, except for one sample from the AML dataset. This suggests that no systematic problem was causing the misclassifications. The only exception (sample #340 of the AML dataset) is illustrated in Figure 5.15(a). Visualization of the outlier FCM data against typical AML and non-AML subjects suggests that the outlier, like typical AML cases, had a sizable $CD34^+$ population. However, the forward

Table 5.4: Performance of algorithms in the sample classification challenges on the validation cohort. Not all algorithms participated in all challenges. Particularly, a large number of algorithms participated through the DREAM project that only included the AML dataset.

	Sensitivity	Specificity	Accuracy	**Fmeasure**	Sensitivity	Specificity	Accuracy	**Fmeasure**	Sensitivity	Specificity	Accuracy	**Fmeasure**
	Challenge 1: HEUvsUE				Challenge 2: AML				Challenge 3: HVTN			
FlowCAP												
2DhistsSVM	0.091	0.91	0.50	**0.17**	0.95	1.00	0.99	**0.97**				
BAD					1.00	1.00	1.00	**1.00**				
EMMIXCYTOM					0.95	0.99	0.99	**0.97**				
flowBin	0.000	0.91	0.45	**0.00**	0.30	1.00	0.92	**0.46**				
flowCore-flowStats	0.455	0.64	0.55	**0.53**					1.00	1.00	1.00	**1.00**
flowPeakssvm					1.00	1.00	1.00	**1.00**				
flowType	0.636	0.55	0.59	**0.59**	0.95	0.99	0.99	**0.97**	0.71	0.90	0.81	**0.80**
flowType-FeaLect	0.273	0.45	0.36	**0.34**	1.00	1.00	1.00	**1.00**	1.00	1.00	1.00	**1.00**
Kmeanssvm									1.00	1.00	1.00	**1.00**
PBSC	0.545	0.55	0.55	**0.55**	0.75	0.97	0.94	**0.85**	0.95	0.95	0.95	**0.95**
PRAMS									1.00	1.00	1.00	**1.00**
PramSpheres	0.364	0.36	0.36	**0.36**					0.90	0.90	0.90	**0.90**
RandomSpheres					0.95	0.99	0.99	**0.97**				
SORT					1.00	1.00	1.00	**1.00**				
SPADE					1.00	1.00	1.00	**1.00**	1.00	1.00	1.00	**1.00**
SWIFT	0.545	0.73	0.64	**0.62**					1.00	1.00	1.00	**1.00**
DREAM												
admire-lvq					1.00	1.00	1.00	**1.00**				
bcb					1.00	1.00	1.00	**1.00**				
biolobe					1.00	1.00	1.00	**1.00**				
cihc					0.95	1.00	0.99	**0.97**				
daltons					1.00	1.00	1.00	**1.00**				
DREAM–A					0.95	0.99	0.99	**0.97**				
DREAM–B					0.85	1.00	0.98	**0.92**				
DREAM–C					0.85	1.00	0.98	**0.92**				
DREAM–D					0.95	0.99	0.99	**0.97**				
fivebyfive					1.00	0.99	0.99	**1.00**				
jkjg					1.00	1.00	1.00	**1.00**				
SPCA+GLM					0.85	0.99	0.97	**0.91**				
team21					1.00	1.00	1.00	**1.00**				
uqs					0.95	1.00	0.99	**0.97**				

scatter values of this population overlap with normal lymphocytes (Figure 5.15 panels (B) to (G)). Obtaining additional information on this patient was not possible. However, independent analysis of the FCM assays by a hematopathologist suggested two possibilities that would explain why this patient was an outlier: The forward scatter (roughly proportional to the diameter of the cell) of the blasts was lower than that of the other AML patients. The size of leukemic blasts shows wide variations from patient to patient and even within a given patient, being medium to large in size in most [6], and very small "microblastic" cells in rare patients (*e.g.,* [72, 120]). The other possibility is that given the lower blasts frequency (16.7%), this patient may have been diagnosed with a high grade myelodysplasia (blasts 10-19%), a preleukemic condition, rather than AML, which requires a blast count of

$>$20% for diagnosis. Alternately, the patient may have AML by morphological blast count, but flow cytometry may be underestimating the blast frequency. This can result from hemodilution of the bone marrow specimen or presence of cell debris or unlyzed red blood cells [94].

5.3.5 Automated Methods Select Cell Populations Identified as Predictive Through Manual Analysis

Previous manual gating and analysis of the HVTN data identified the $CD4^+/IL2^+$ T–cell subpopulation as discriminative between Env- and Gag-stimulated samples, with the proportion of $CD4^+/IL2^+$ cells in the Env-stimulated samples systematically higher than the proportion of $CD4^+/IL2^+$ cells in the Gag- stimulated samples. This effect was not observed in manually gated placebo data, suggesting it is vaccine specific, and is consistent with the gp120 Env protein boost (the only protein component of this vaccine) given to individuals participating in this study. Interestingly, examination of the features selected by automated gating methods for prediction between Env- and Gag-stimulated samples revealed that, of the eight methods that could directly identify predictive features, four selected features containing the $CD4^+/IL2^+$ phenotype. Furthermore, the flowStats/flowCore method, which was designed to be directly comparable to the manual gating scheme, identified the exact same $CD4^+/IL2^+$ population as predictive of stimulation with accuracy comparable to that of manual gating and like manual gating, failed to detect the effect in the placebo group in a post–hoc analysis. The sample classifications using the $CD4^+IL2^+$ population gated manually were slightly less accurate than the automatic results obtained from the same population. Post-hoc examination of the data revealed that several of the control and stimulated samples in the data set were matched from different experimental runs, suggesting a possible run–specific effect. When these samples were filtered out of the analysis, manual gating was able to perform as accurately as the algorithms, suggesting that algorithmic approach was more robust to the technical variation.

5.4 Discussion

Two sets of benchmark FCM data were assembled through the FlowCAP project. These benchmark data sets were used to evaluate automated gating methods based on their ability to either recapitulate cell populations defined through manual gating by human experts, or their ability to classify samples based on external variables. Seventy-seven different computational pipeline/challenge combinations were evaluated through these efforts.

In the population identification challenges, pre-defined populations identified by human experts using traditional manual gating approaches were used as the current best practice for evaluating the performance of the current state–of–the–art automated gating algorithms for multi–dimensional FCM data. Although there was general agreement between populations identified by the top algorithms and the results from manual analysis, as illustrated by high F-measure values, it was not possible to identify a single top performing algorithm across all data sets.

In general, demonstrating superiority of a clustering method is difficult due to lack of a ground truth [103]. In the cell population identification challenges, populations identified using traditional manual gating by the data providers were used to establish the reference data for the initial comparison, since it represents the current best practice for the analysis of FCM data. However, manual gating is known to be subjective and potentially error-prone even in the hands of domain experts [73]. Therefore, to increase the robustness of the results, eight sets of additional manual gates by independent experts were produced. The GvHD and HSCT datasets (the datasets with the lowest and highest F-measures, respectively) were chosen for this experiment. The human experts were directed to perform ”exhaustive” manual gating (*i.e.,* attempt to identify as many cell populations as possible, subject to extensive back-gating). The results were generally consistent with those of the initial manual analysis. For example, the top four algorithms for the HSCT dataset were FLAME, ADICyt, flowMeans, and MM&PCA for both the initial and the refined manual gates.

For the GvHD dataset, there was significant disagreement between the algorithms as well as between the manual gates produced by different analysts. However, the results were still consistent with the original results with only minor vari-

ations. Per-population analysis of this dataset revealed cell populations that were merged by most of the algorithms to other cell populations with generally similar marker expression patterns but separated by the manual gates based on a subset of the markers. This emphasizes the importance of designing methodologies that can use background biological knowledge in the clustering process. In this case, human analysts used their knowledge about the scatter channels to partition these cells into two different populations despite their similarity in every other channel.

The mean F-measure values and rank scores showed that the combined predictions obtained by ensemble clustering were more accurate than the results from individual algorithms. This is particularly important for computational analysis, because in practice it may not be feasible to hire multiple experts to carry out multiple manual gating; however, it is realistically possible to run automated ensemble methods at minimal cost. The ablation analysis confirmed that increasing the number of algorithms in the ensemble resulted in improved predictions up to a certain point (perfect F-measure was never achieved). When algorithms with high scores were more frequent, the ensemble clustering performed better and was less sensitive to the exclusion of several of the algorithms (challenges 1 and 3 in contrast to 2 and 4). This suggests that having a number of good algorithms is necessary to obtain good ensemble results, but there might be a point after which adding more algorithms does not significantly improve the results. Similar results were observed using the refined manual gates. Particularly, when a large number of algorithms with high F-measures were available (the HSCT dataset and the top 50 most consistently identified populations in the GvHD dataset), the ensemble clustering out-performed the individual algorithms. When the individual algorithms were performing poorly (the remaining cell populations in the GvHD dataset), the ensemble clustering's performance decreased as well.

In the sample classification challenges external variables were used as a reference for evaluation of the algorithms. Many of the algorithms were able to achieve a high performance in discriminating between AML and non–AML and between Env- and Gag-stimulated samples suggesting that automated methods performed extremely well for these sample classification use cases. In the HEU vs. UE study, the algorithms were not able to correctly label the majority of the test/validation set. Manual analysis of this dataset by expert flow cytometry

analysts did not identify any statistically significant differences [112]. This is not surprising, since all samples were derived from HIV-negative newborns, with HEU samples from individuals that had been exposed to HIV *in utero* but were uninfected, and UE samples from individuals that had never been exposed. In a way, this negative example provides further support for the effectiveness of these automated approaches, since they did not generate positive classification results when none likely exist. In the AML challenge, one sample was identified as an outlier as it was misclassified by approximately half (12) of the methods, while most samples were misclassified by only one or two methods. This case was then compared to typical AML and non–AML cases by a clinical expert and was confirmed as a clinically outlier case, potentially with pre-leukemia. A post-hoc analysis of the HVTN dataset was performed to compare the features selected by manual gating against those selected by some of the automated gating methods. Additional confirmation for these findings was provided by comparison against placebo samples that were not available to the participants. The results of this post-hoc study demonstrated that automated gating can perform as well as manual gating, and even exceed the performance of manual gating in some cases, as evidenced by the ability of automated gating to maintain accuracy in presence of technical variations that affect manual gating results.

Every approach to automated flow cytometry published in the last five years, as well as several unpublished methods, participated in at least one of the FlowCAP challenges. Participation by the flow informatics community was not only widespread, it was also collaborative. This collaboration included the sharing of ideas, and the distribution of work to avoid unintended duplication of efforts. The development of flow informatics coincided with the expansion in the open source software philosophy, and this mindset has been widely adopted by the flow informatics community. This open access philosophy has most certainly contributed to the rapid maturation of these novel methods. One of the challenges of the second competition was organized in collaboration with the DREAM (Dialogue for Reverse Engineering Analysis and Methods) initiative [17, 79, 96, 115]. As FlowCAP does with the flow cytometry community, DREAM aims at nucleating the systems biology community around important computational biology problems. Given the growing use of flow cytometry data in systems

biology research, the collaboration between DREAM and FlowCAP was a natural and fruitful one.

Taken together, the data presented here suggest that the current state of the art FCM analysis algorithms perform very well. However, our ability to make stronger evaluations is limited by two specific shortcomings in the challenges. First, for sample classification, instead of probabilities of each subject belonging to each class, the participants were asked to provide discrete outputs (class names). This made it impossible to perform receiver operating characteristic (ROC) analysis and potentially decreased the robustness of the study. Second, for cell population identification, the data provided to the participants was pre-processed, which could potentially be a source of bias. For example, in some datasets the analysis was limited to the lymphocyte population, and other cells were manually excluded. In some cases, for consistency with manual gating, algorithms were forced to process data that was improperly transformed. This was disadvantageous to the algorithms when, for example, artifactual clusters of cells were introduced [80, 100]. For example, in some cases the log transformation was used rather than the logicle resulting in a large number of events on the axes [80].

We identified several challenges that remain to be addressed in the future: 1) Many of the algorithms evaluated in the sample classification challenges relied on matching cell populations across multiple samples. Several alternatives for this process have been proposed (*e.g.,* see [67, 97]), but the performances of population matching methods have never been compared objectively. 2) In retrospect, the data used for the sample classification challenges appeared to be overly challenging (HEU vs. UE) or overly simple (AML and HVTN) for the algorithms. The analysis of these algorithms should be extended to evaluation using datasets with correlation structures that can be more challenging for these algorithms to reveal their potential shortcomings in more details. 3) Our preliminary results (post-hoc analysis of HVTN) suggest that computational methods can outperform humans in handling technical variation. This needs to be investigated in more detail by providing benchmarks of cross–institutional datasets with standardized panels (*e.g.*, those produced by the human immunology project [75]) to design computational pipelines that are more robust to technical variation. 4) The runtimes in the cell population identification challenges were measured using different

hardware and software environment. While this provides some estimate of the time requirements, direct comparison was not possible. In the sample classification challenges, the situation was further complicated by having separate training and testing procedures which often included visual exploration of the data by the algorithm developers. In future challenges, we intend to address this problem by introducing standardized interfaces and data-formats between the participating software and the evaluation pipeline so that the evaluation can be performed in a unified hardware/software setting. In addition to providing an objective comparison of time requirements, this will also facilitate independent reproduction of the results.

5.4.1 Availability

The display items presented here can be fully reproduced using the scripts provided on the FlowCAP website[7]. Annotated raw data using MIFlowCyt descriptions [68] is available through a public repository sponsored by the International Society for Advancement of Cytometry (FlowRepository.org) using the following experiment IDs: FR-FCM-ZZY2 (GvHD), FR-FCM-ZZYY (DLBCL), FR-FCM-ZZY3 (WNV), FR-FCM-ZZY6 (HSCT), FR-FCM-ZZYZ (ND), FR-FCM-ZZZU (HEUvsUE), FR-FCM-ZZYA (AML), and FR-FCM-ZZZV (HVTN).

[7] http://flowcap.flowsite.org/codeanddata

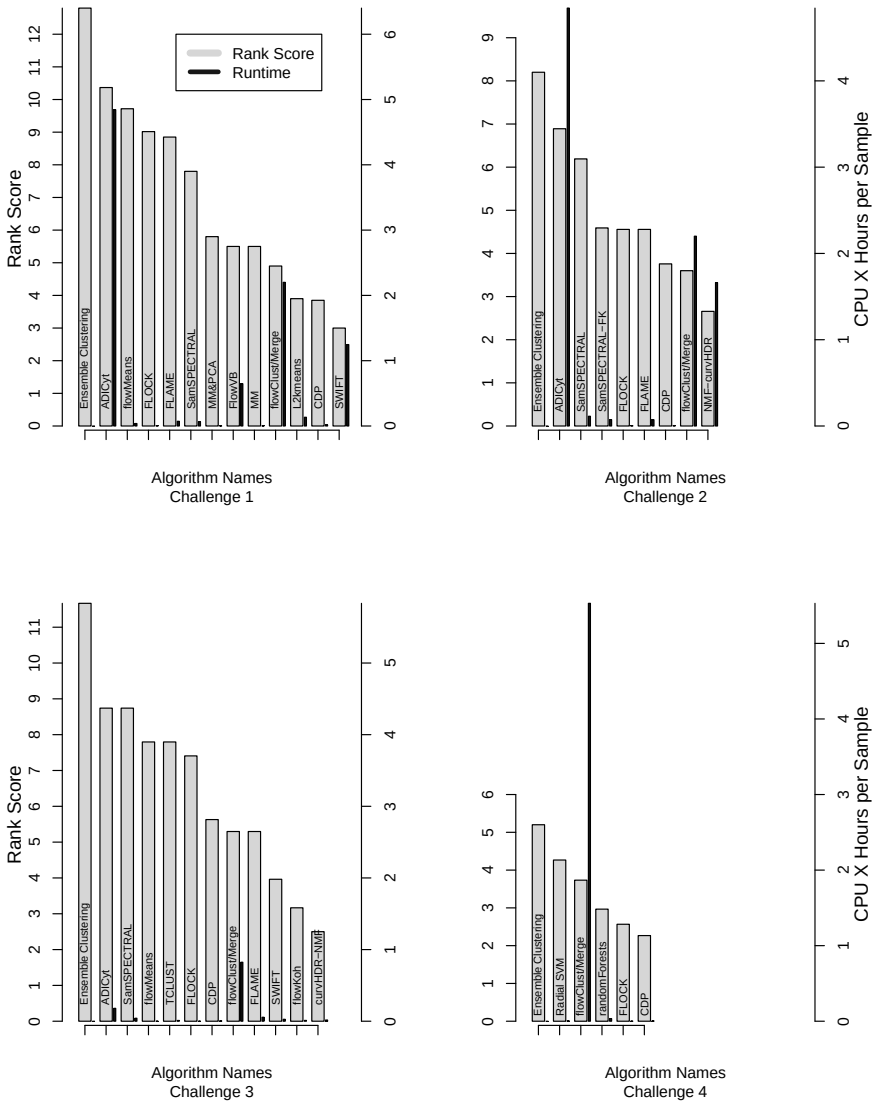


Figure 5.3: Rank scores and runtimes (per CPU per sample) for each algorithm/challenge. The runtime of the ensemble clustering methods is not included, but it would be close to the sum of the runtimes of all other algorithms.

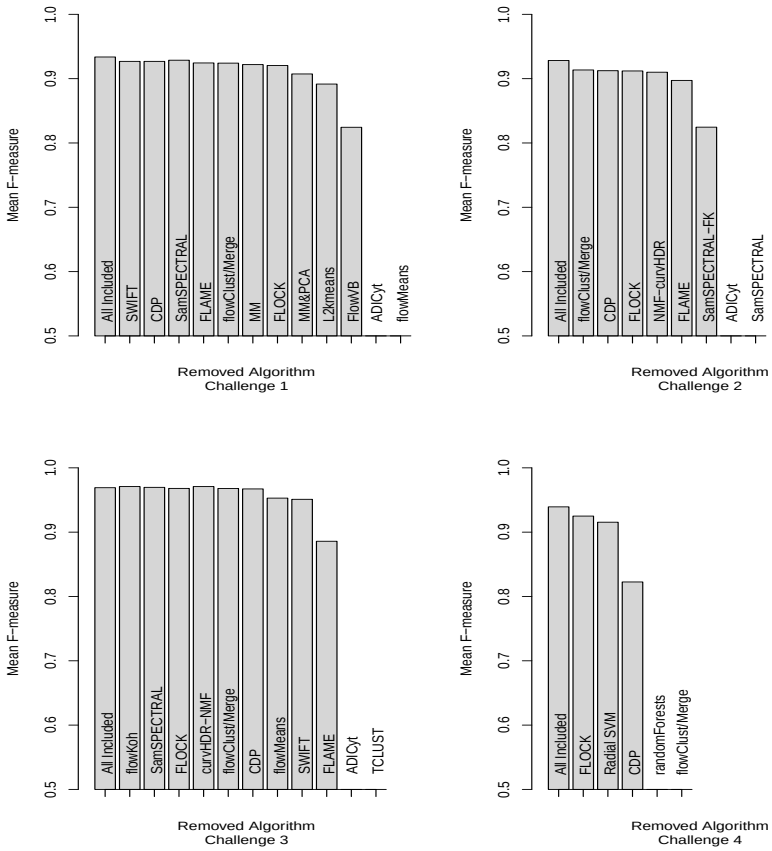


Figure 5.4: Ablation analysis results. The algorithm are listed in order of impact, from lowest to highest, on the F-measure value for each challenge, and the respective F-measure of the combined predictions indicated on the y-axis. Ensemble clustering for less than 3 algorithms is undefined for the CLUE package, therefore, the last two steps (where 2 and 1 algorithms are left, respectively) are not shown in this figure.

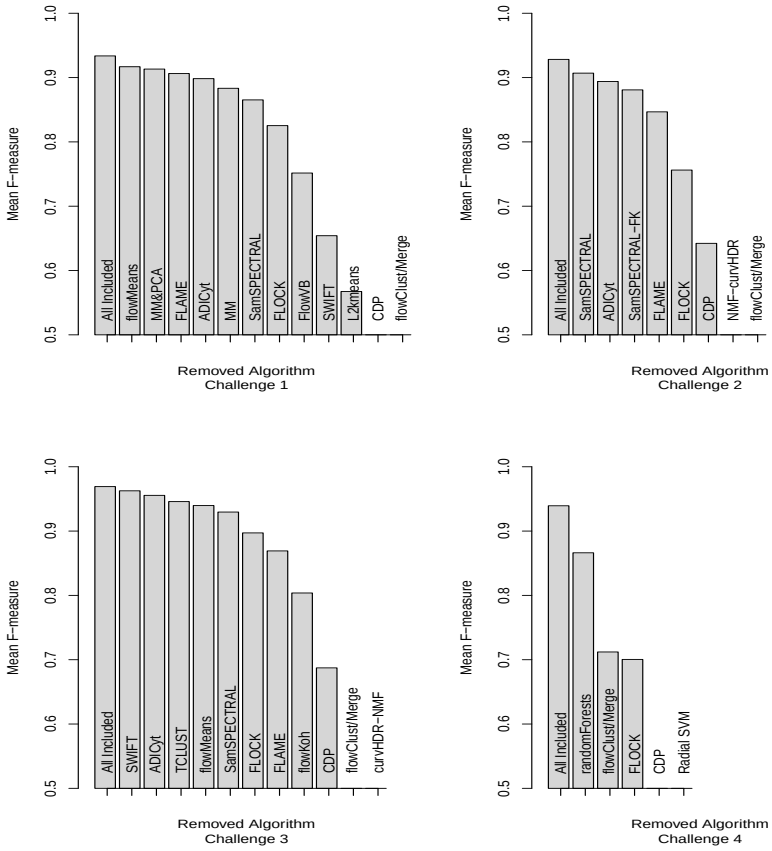


Figure 5.5: Reversed Ablation analysis results. The algorithm with maximum contribution at each step of the ablation analysis (for each challenge) and the respective F-measure of the combined predictions are listed from highest to lowest. Ensemble clustering for less than 3 algorithms is undefined for the CLUE package. Therefore, the last two steps (where 2 and 1 algorithms are left, respectively) are not shown in this figure.

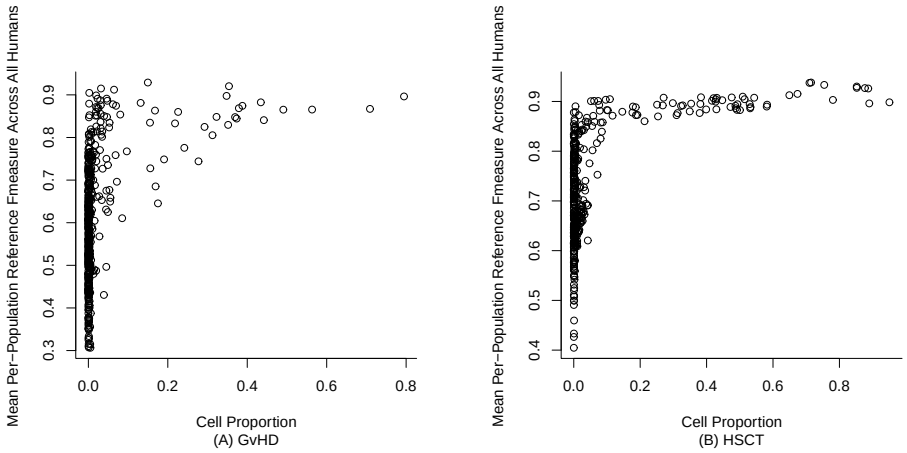


Figure 5.6: Correlation between F-measure value and cell population size. These plots show the average F-measures versus the size of the cell population across the samples in the two datasets for all eight sets of manual gates. Generally, these data suggest that there is a stronger consensus among humans when the cell population is larger. Agreement among independent human gaters can also be found for some small cell populations but not for others.

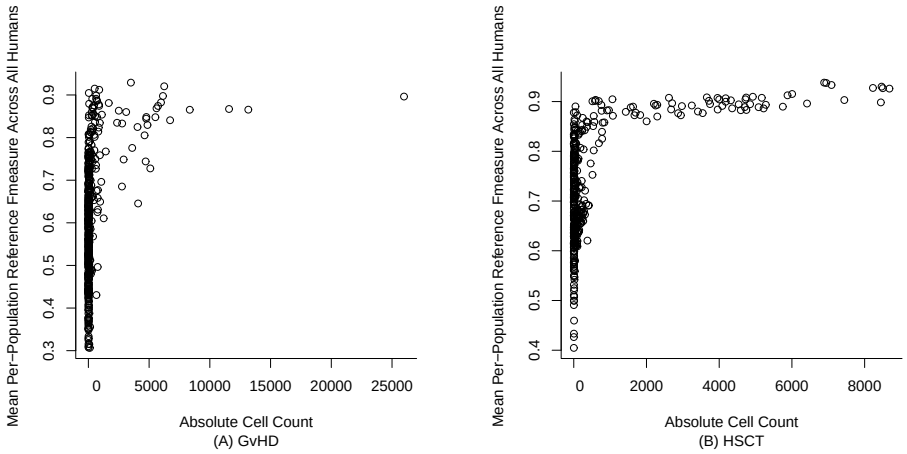


Figure 5.7: Same as Figure 5.6 using absolute cell count instead of cell proportion.

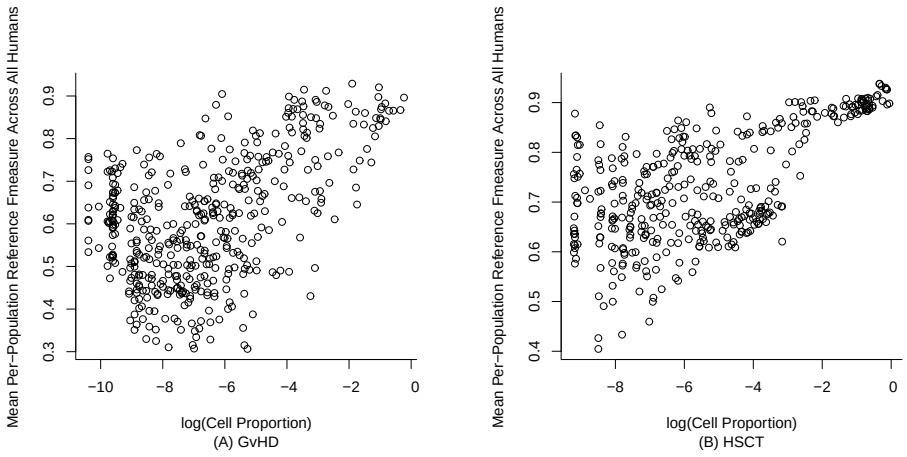


Figure 5.8: Same as Figure 5.6 on a log scale.

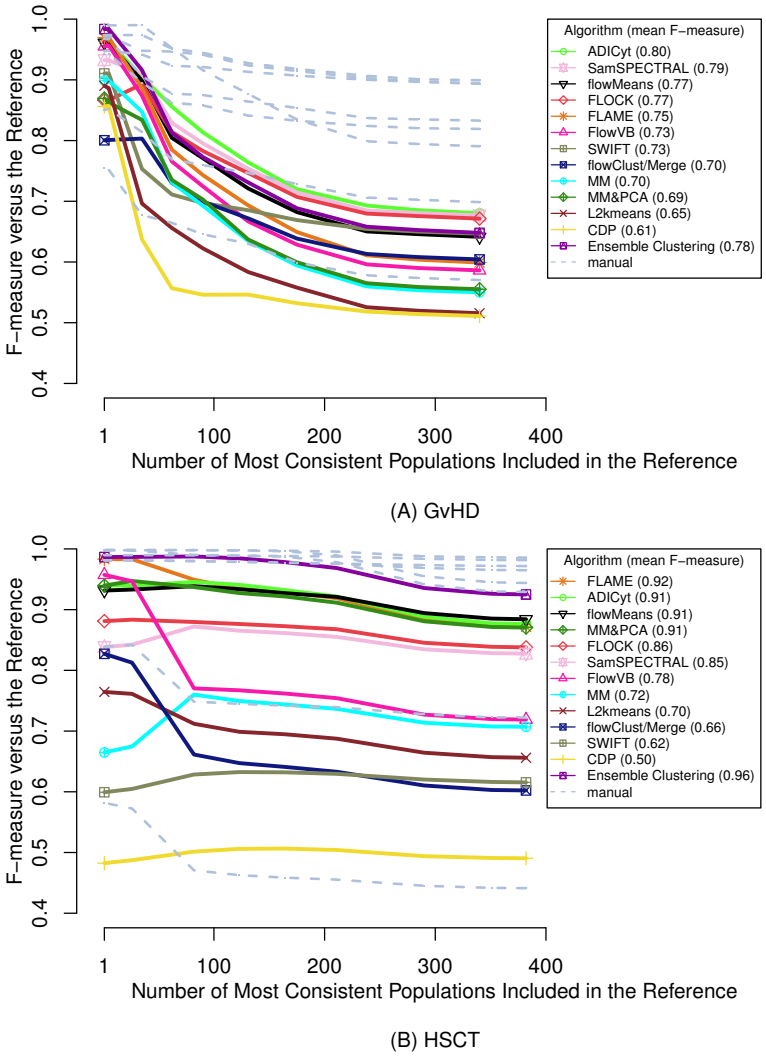


Figure 5.9: Comparison of algorithms and manual gates using the consensus of humans expert manual gates. For the (A) GvHD and (B) HSCT datasets, the few reference populations that match all of the manual gates strongly (left) resulted in high F-measure values. Adding more cell populations with lower consistency among manual gates decreased the F-measures gradually.

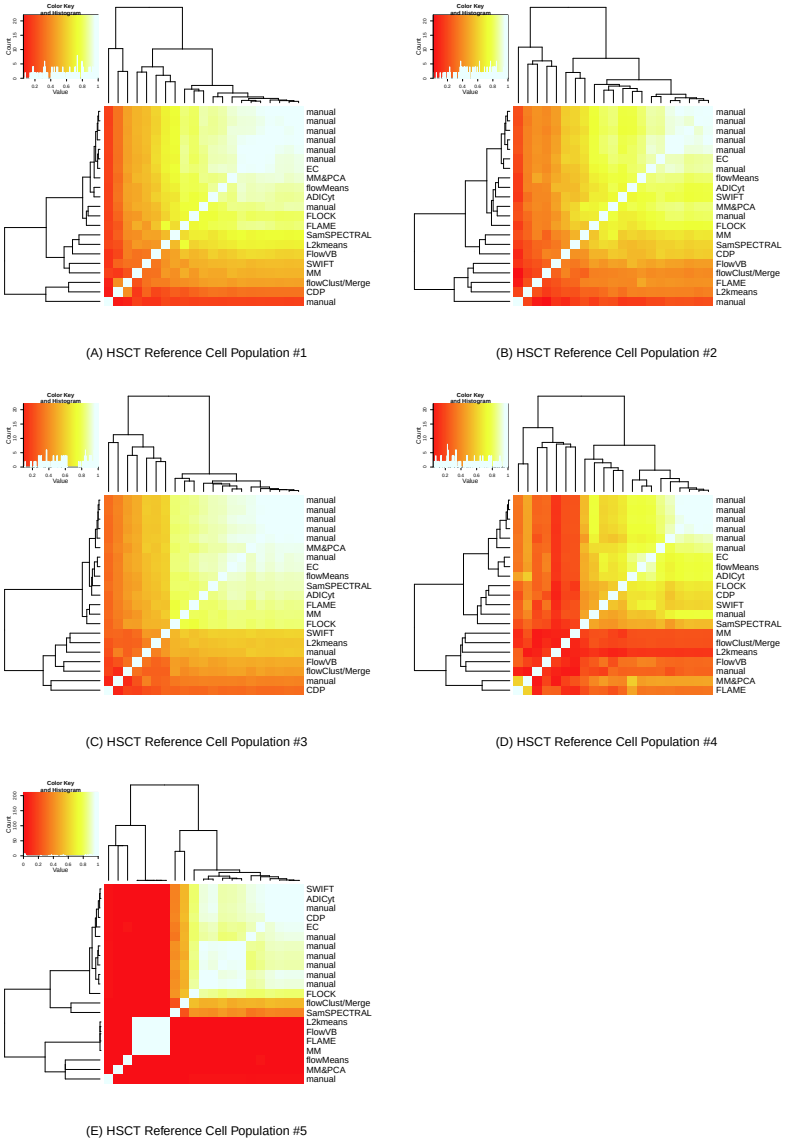


Figure 5.10: Per population pair-wise comparisons of manual gating and algorithm results. Average F-measures of all pairs of results for the cell populations across all samples in the HSCT dataset was determined (*i.e.*, one heatmap for every cell population in the reference). The manual gate consensus for each sample was used as a reference for matching of the automated results of that sample. Pair-wise F-measures between all algorithms and manual gates for the HSCT dataset are shown. The dendrogram groups the algorithms/manual gates based on the similarities between their pair-wise F-measures.

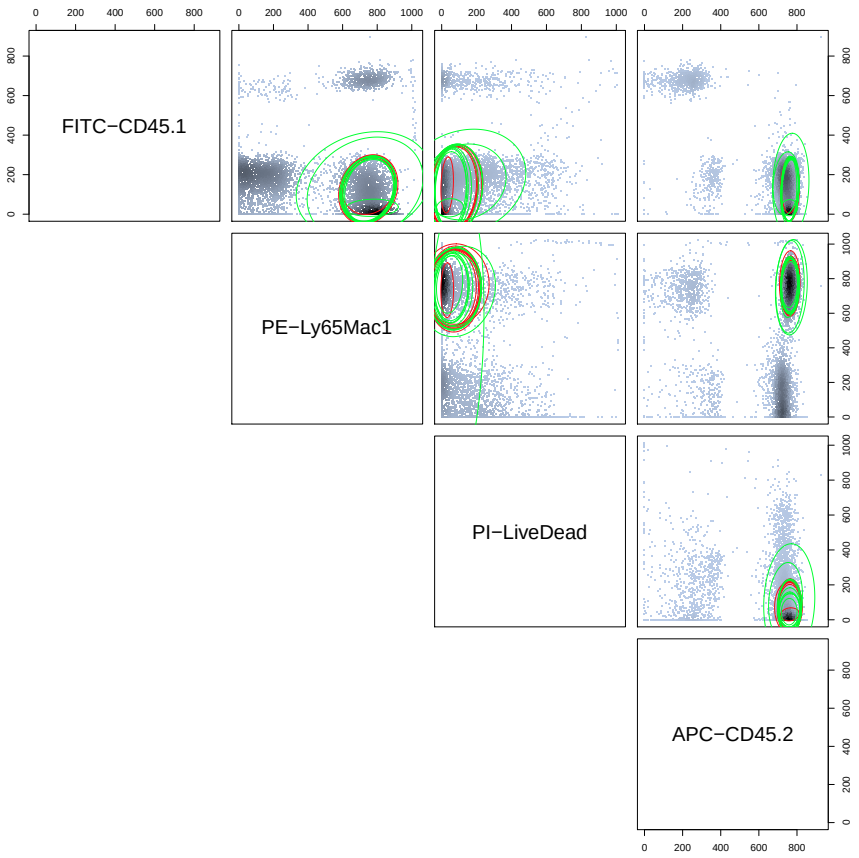
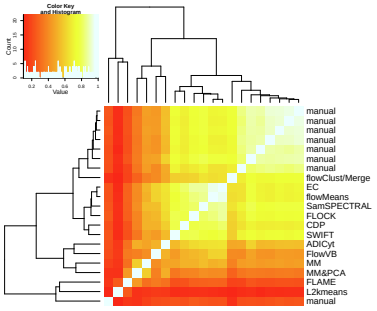
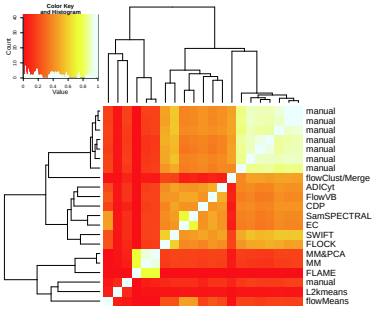


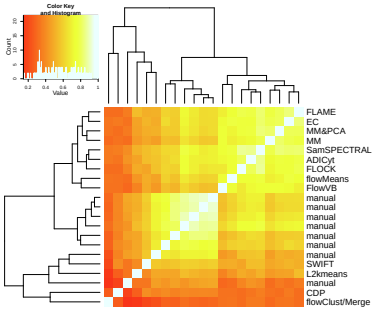
Figure 5.11: Scatter plot of Sample 26 of the HSCT dataset (the sample with maximum number of reference cell populations) for the third population for which a relatively high agreement between all algorithms and manual gates have been observed (Figure 5.10, Panel C). In this plot, algorithm results are partitioned with green ellipses, and manual gating results are partitioned with red ellipses.



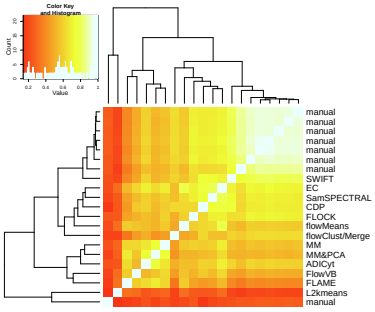
(A) GvHD Reference Cell Population #1



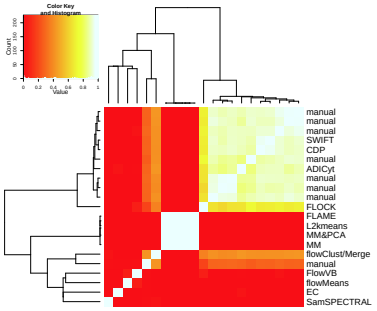
(B) GvHD Reference Cell Population #2



(C) GvHD Reference Cell Population #3



(D) GvHD Reference Cell Population #4



(E) GvHD Reference Cell Population #5

Figure 5.12: Similar to Figure 5.10 for the GvHD dataset.

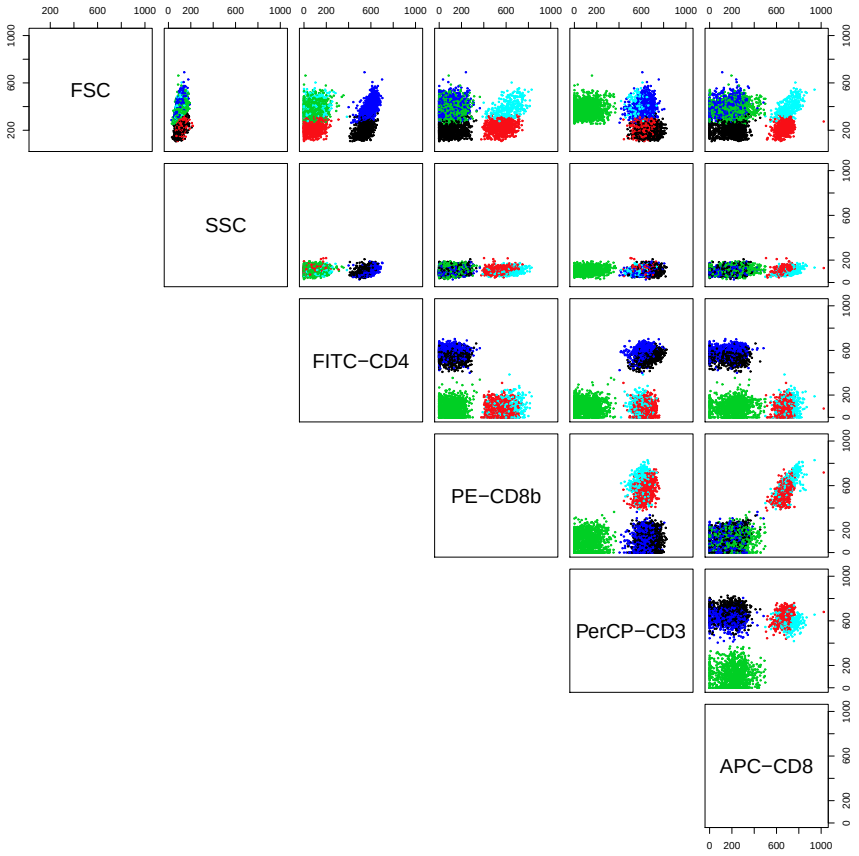


Figure 5.13: Scatter plot of Sample 1 of the GvHD dataset (the sample with maximum number of reference cell populations). Colors are as follow (can be matched to the panels of Figure 5.12): 1-black, 2-red, 3-green, 4-blue, and 5-cyan. The red population has been consistently missed by all of the algorithms and consistently identified by most of the manual gates (Figure 5.12 Panel B). The only major difference between the red and the cyan population is in the forward scatter channel (FSC.H).

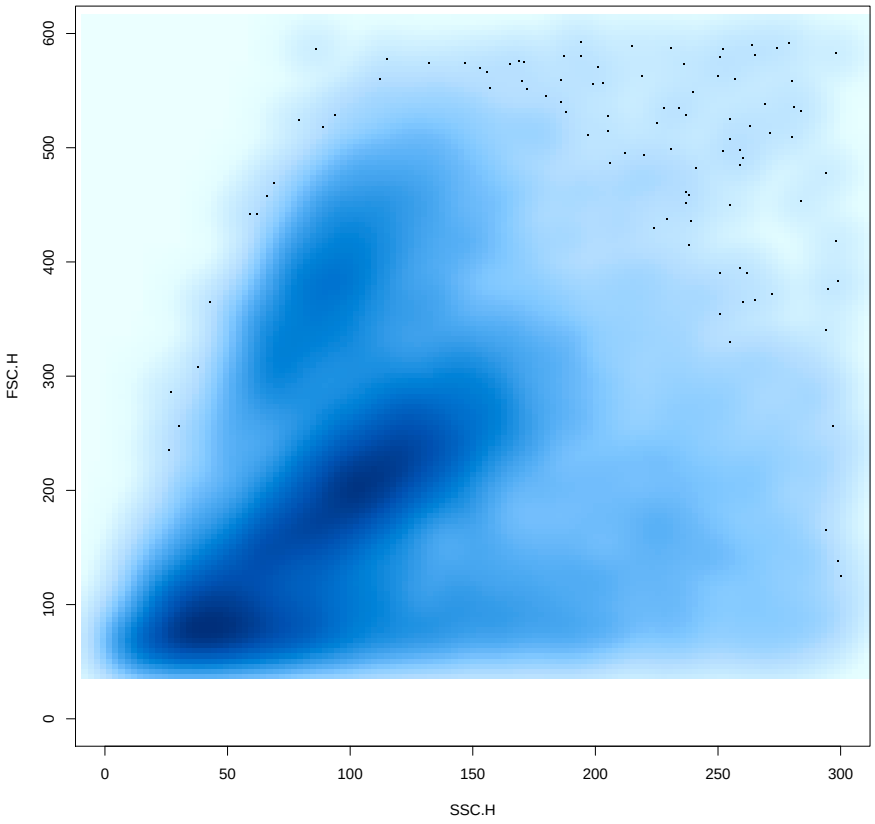


Figure 5.14: Forward and side scatters of the sample visualized in Figure 5.13 to confirm the existence of two different cell populations (red and cyan). Deadcells (low FSC.H) have been manually removed)

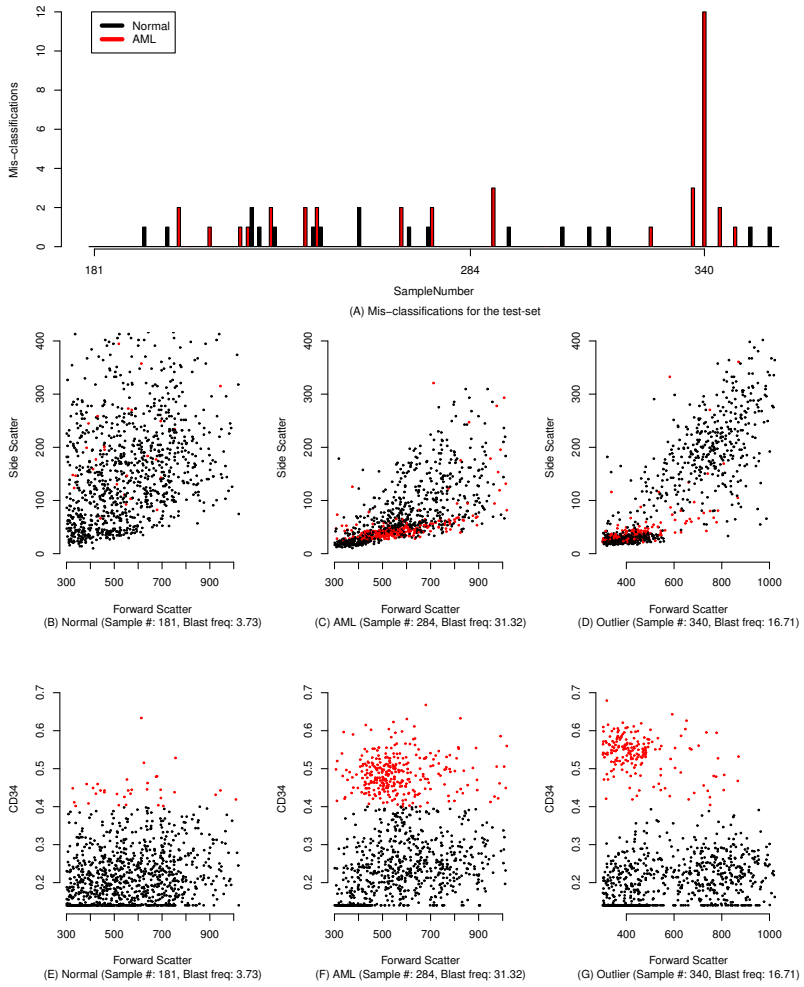


Figure 5.15: Outlier AML subject, detected by the algorithms. (A) Total number of misclassifications for each sample in the test-set (samples #180 #359) of the AML dataset is presented. Sample #340 was frequently misclassified. FSC/SSC (B-D) and FSC/CD34 (E-G) scatter plots of representative Normal (B & E) and AML (C & F) samples and the outlier (D & G) are shown, with the $CD34^+$ cells highlighted in red (B) to (G). Cell proportions of the $CD34^+$ population are reported as Blast freq. percentages. The outlier sample appears to be different from typical AML and normal samples in terms of both the frequency of $CD34^+$ cells and the MFI of forward scatter.

Chapter 6

Conclusions and Future Work

6.1 Summary

High-dimensional flow cytometry is routinely used for exploratory analysis of the immune system. However, in absence of proper data analysis methods, hypothesis driven manual gating has been used for exploring a limited number of immunophenotypes [71]. Thus, the value of these high-content technologies has been largely wasted. While several computational pipelines existed [8], due to several issues discussed in Chapters 2, 3, and 4 (including high time requirements, lack of mechanisms for incorporating background biological knowledge and dependence on subjective cluster matching) their application to real world datasets were extremely limited [71].

In this work, I developed a computational pipeline for exploratory analysis of high-dimensional FCM assays. The first step of the pipeline is a cell population identification algorithm that combined the low time-complexity of K-means clustering with the robustness of Gaussian mixture models to automatically identify non-convex cell populations. The original K-means algorithm [69] requires the number of clusters to be pre-identified, is very sensitive to the initialization strategy, and is limited to spherical cell populations. More robust statistical mixture models were used to address these issues at the cost of a significantly higher runtime (to the extend that analysis of tens of parameters measured across millions of cells was unfeasible). In Chapter 2, I presented

results suggesting that a combination of K-means clustering and post-processing using a statistical model (called flowMeans) can be as accurate as statistical mixture modeling, yet significantly faster. One of the limitations of flowMeans is the sensitivity of the final results to the initial number of clusters for K-means. While this sensitivity is much lower than the sensitivity to the actual number of cell populations or the initialization of K-means cluster centers, it can still be problematic for some use-cases. Since the development of flowMeans, at least one new cell population identification algorithm has been published [44], and more will be published in the future. However, due to the subjective nature of the cell population identification problem, at least one free parameter has to exist for each of these algorithm to enable them to adjust to the requirements of the user. As discussed in the next section, increasing the number of free parameters that control different aspects of the cell populations and then automatically optimizing them is a very interesting direction for future work.

Exploratory analysis of high-content FCM assays of large cohorts using multi-dimensional clustering algorithms is limited by several factors: the cell populations identified by these tools need to be matched across several samples extracted from different sources (*e.g.,* patients) in a subjective manner, incorporating human experts' knowledge for identification of rare cell populations is difficult, and little information will be provided about the contributions of different markers to the final results. In Chapter 3, I proposed a new cell population identification strategy based on combining one-dimensional partitions for production of multidimensional cell populations. Human experts' knowledge of specific cell populations can be easily incorporated into the one dimensional analysis and the meta-clustering problem will be significantly easier to solve as cell populations are being matched in one dimension at a time. In addition, when combining single dimensional gates to generate multidimensional immunophenotypes, this approach considered all possible combinations including those that do not involve some of the markers. This enabled our statistical analysis to study the effect of each marker on the characteristics of a cell population of interest.

One of the main limitations of this approach is the assumption that the markers can be analyzed independently. Biological relationships can exist between these markers that would challenge this assumption. More importantly, spectral

overlap between the fluorochromes conjugated to the antibodies can decrease the independence of the measurements. To avoid these problems, proper panel design, quality control, and compensation for spectral overlap is necessary for this pipeline. Another limitation of this approach is the time and memory requirement of analyzing a large number of markers. However, this can be controlled by limiting the depth of the analysis (the maximum number of markers included in an immunophenotype) depending on the availability of computational resources.

While considering all combinations of cell populations allowed a more complete analysis than previously possible, it also resulted in a very large hit-list of potentially interesting immunophenotypes (*e.g.,* 101 in the study described in Chapter 3). Because we allowed exclusion of certain markers, these cell populations often overlapped in multiple complicated ways. For example, $CD4^+CD8^-$ cells were also $CD4^+$. They also were very likely to have a significant overlap with $CD3^+CD4^+$. In Chapter 4, I described the last step of the pipeline that characterizes the immunophenotypes in terms of the markers involved, and organizes them in a hierarchy using their most important (in terms of correlation with an external outcome) parent population. This approach not only provided a better visualization of the results, but also helped control the trade-off between the number of markers required for measurement of an immunophenotype and the strength of the correlation with the clinical outcome. This is particularly important in settings where the complicated instruments for high-dimensional assays are not available including in poor countries (particularly important for TB and HIV), highly regulated clinical settings, and for identification of targets for new therapies.

This pipeline was primarily applied to a large dataset of 466 HIV^+ subjects. PBMCs were extracted at the time of infection and were analyzed by a 14 color panel including 13 surface markers and KI-67. The final clinical outcome of the patients (time to AIDS, death, or initiation of HAART) was also available. The goal of the study was to find cell populations that could predict the clinical outcome. The dataset was previously analyzed manually, resulting in identification of two cell populations with a modest correlation with the clinical outcome. Using the pipeline described above, not only we reproduced these two cell populations, but also identified a hit-list of 101 immunophenotypes correlated with the outcome. After analysis of the overlapping sets these were narrowed down to three main

hierarchies of cell populations with statistically significant correlations with the clinical outcome much stronger than those identified manually.

Before the development of this pipeline, for most new use-cases the development of a new pipeline (or extensive customization of existing pipelines) was necessary in most practical settings (*e.g.,* [9]). The pipeline presented here, however, is very robust and flexible. For example, it can work with different types of clinical or biological outcomes, different statistical tests, and any clustering algorithm. In fact, currently it is being used extensively in the Brinkman lab for analysis of a wide range of datasets, including HIV (in different settings in collaboration with different groups), several subtypes of Leukemia and Lymphoma, Tuberculosis, Parkinson's disease, kidney transplantation rejectors, and different inflammatory diseases. In most cases, these studies resulted in identification of novel cell populations missed by previous manual analysis by the labs that produced these datasets. In others, computational analysis was used as a preliminary experiment to guide the manual gating strategy.

In most bioinformatics works on clinical data, designing a classifier that can successfully label different clinical outcomes is one of the most challenging step in exploratory analysis. For example, in gene expression analysis, once the important features used by the classifier have been identified, a clear hit-list of differentially expressed genes will be available for gene ontology analysis and confirmatory studies. For analysis of FCM data, similarly, computational efforts have been mostly focused on development of multi-variate models for cross-sample studies (in addition to cell population identification compared against manual gates). However, due to the hierarchical nature of cell types, identification of one or more cell populations that can discriminate between groups of patients is insufficient for interpretation of the results. The pipeline presented in this work, to the best of my knowledge, is the only pipeline that focuses on characterization of the identified cell populations in terms of the markers involved.

However, for cases where an accurate classification is more important than identification of specific immunophenotypes, this pipeline can be used for multivariate supervised learning. For example, I used a classifier to combine the predictive power of the single dimensional immunophenotypes to produce a more robust multivariate model. The classifier used was a linear model with $L1$-constrains on

the weights. FeaLect, a feature selection methodology developed in the Brinkman group, was used for selecting the best immunophenotypes for the multivariate model [129]. FeaLect is a wrapper function for the linear model (*i.e.,* selects the features based on combinations of their predictive powers as opposed to filter functions that can only look at one feature at a time). As described in Chapter 5, this pipeline resulted in perfect classifications of every single sample in both the HVTN and AML datasets in FlowCAP-II.

FlowCAP is a highly collaborative project with two main goals: 1) to provide guidance to the end users regarding the proper use of computational tools for analysis of FCM data; 2) to identify the shortcomings of existing tools to facilitate the development of new approaches by the informatics community. Chapter 5 includes the results of the first two FlowCAP competitions in 2010 and 2011.

In FlowCAP-I we focused on comparison of algorithms for identification of cell populations. The evaluation was performed against the results of the current best practice - cell populations identified by expert human analysts. Five different datasets were used for the evaluation. We found that manual tuning of the free parameters of the algorithms by the developers does not necessarily result in an improvement. In fact, in some cases, after human intervention the similarities to manual analysis decreased. We also found that generally, providing the expected number of cell populations improved the results. However, this information is usually not available in exploratory settings. Also, our results suggested that providing a small subset of the manual gates to some of the algorithms as a training-set can improve the accuracies significantly. Finally, we found that a consensus of fully automated algorithms produced by an ensemble clustering algorithm out performed every single one of the individual algorithms over a wide range of datasets.

To investigate the sensitivity of these results to our human expert analysis, we recruited eight additional analysts to manually gate parts of our dataset. The consensus of these manual gates (again, produced using the ensemble clustering algorithm) was used as our refined reference cell populations. Evaluation of the algorithms using this new reference confirmed our initial results with only minor variations.

While the comparison against manual gating has been helpful in demonstrating

the practical utility of these algorithms, it is important to note that it will never be a good gold standard for evaluation of these algorithms. Clustering (and cell population identification) is a subjective and ill-defined problem. For example, a very important rare cell population for one application can be considered noise in another. Increasing the number of manual gates improves the robustness of these evaluations but still penalizes the algorithms in cases where they have performed better than the manual gates (*e.g.,* algorithms might be penalized for identification of a small cell population that has been missed by the majority of the expert analysts). The only solution to this problem is an indirect evaluation using an external biological or clinical outcome.

In the second FlowCAP competition, we therefore focused on evaluation of computational pipelines in prediction of external variables. The project consisted of three binary classification challenges based on real-world datasets. In each case, the dataset were randomly and uniformly divided into a training- and a test-set. The external variable was provided to the participants only for the training-set and the test-set was used as an independent validation cohort.

Overall, the participating computational sample classification methods performed stronger than expected. For example, one of the dataset included PBMCs from a post-HIV-vaccination study. The goal of the challenge was to identify T-cell population that could discriminate between two antigen stimulation groups (Env and Gag). A large fraction of the algorithms were able to classify these samples perfectly. These results were surprising since previous manual analysis of the same dataset achieved a lower accuracy. Further inspection of the dataset revealed a technical bias in some of the assays which was contributing to the lower performance of the manual analysis. Exclusion of these samples resulted in a perfect classification by the manual analysts. These results suggest that computational methods can match, and in some cases exceed, the ability of expert humans in exploratory analysis of FCM data.

While FlowCAP is not the first project to report superiority of computational analysis of FCM data in comparison to manual analysis, it is unique in four aspects: First, the evaluation was performed by an independent group, ensuring that all participants had equal access to the data. Second, a wide range of dataset representing different real-world use-cases were used in the evaluation

process. Third, the correct answers were provided to the participants only after the submission of the final results to minimize the effect of over-tuning of parameters. Fourth, the highly collaborative design of the project guaranteed the quality of the submitted results as they were produced by the groups that originally developed the respective software.

6.2 Future Work

This thesis was mostly focused on exploratory analysis of clinical data. However, the pipeline presented here can be modified for a wide range of use-cases, including diagnosis, marker panel design, and guiding Fluorescence-activated Cell Sorting (FACS)-based sorting strategies:

The pipeline presented here can be used for design of accurate diagnosis tests using multivariate classifiers. Some preliminary results were provided through the *flowType-FeaLect* pipeline in FlowCAP-II, but much remains to be done in the future. Particularly, numerous free parameters throughout the pipeline could be optimized using automated parameter tuning approaches for a higher classification accuracy [54].

For cell population identification, ensemble clustering algorithms specifically optimized for FCM data remain to be designed, implemented, and tested. The matching of cell populations across multiple samples, especially in presence of technical variations [74] from multi-center studies which are becoming increasingly more popular, is another important subject that can improve the quality of computational methods for FCM data. Finally, as FlowCAP-I suggested, cell population identification algorithms that can learn from examples provided by human experts can be significantly more accurate than unsupervised algorithms, particularly when specific cell populations are of interest. However, very few algorithms in FlowCAP were able to learn from manual gating examples, and there probably is room for further improvements.

Another potential use-case for this pipeline is for designing marker panels. Traditionally, FCM marker panels are designed based on the hypothesis of the study and previously produced results from the literature. For example, for studying T cells in HIV^+ patients, based on previous biological knowledge,

markers like CD3, CD4 and CD8 are very widely used. This is no longer feasible, due to the exploratory nature of modern FCM analysis of complex cellular systems, such as cellular signalling. High-dimensional FCM together with RchyOptimyx can be used to design low-color panels guided by high-content experiments. We believe this approach will be particularly useful for cell sorting applications. For example, sorting based on intracellular markers for further *in vivo* or *in vitro* studies is not currently possible. However, high-dimensional FCM and RchyOptimyx can be used to design panels of surface markers for sorting specific cell populations, guided by intracellular signatures. Several preliminary examples using a mass cytometry dataset were provided in Chapter 4.

The FlowCAP project is an on-going work with several paths actively being considered for future competitions. First, we will try to analyze more sample classification datasets with more challenging clinical outcomes and a higher number of dimensions and cells. Second, several other aspects of FCM data analysis should be explored by FlowCAP including cross-sample cell population matching, identification of specific cell populations, and data analysis in presence of technical variation from multi-center studies. The long term plan for FlowCAP is to convert it to a real-time and online resource for both computational and biological scientists to access real datasets and find suitable software tools, respectively.

To understand the pathogenesis of malignancies, the function of different cellular phenotypes must be analyzed. For more complex cellular systems, such as those involved in cancer, a very wide range of markers must be measured for every single cell. In addition, automated high-throughput FCM will enable us to perform several high-dimensional assays per sample. Although I provide a strong pipeline for analysis of these datasets, these new technologies will produce datasets even more complex than those discussed in this work in terms of the number of markers, cells, patients, and time points. Only upon further improvement of these computational tools these technologies can be used to their full potential, for example, to characterize a wide range of drug effects on live cells for designing personalized therapeutic strategies.

Bibliography

[1] N. Aghaeepour, A. H. Khodabakhshi, and R. R. Brinkman. An empirical study of cluster evaluation metrics using flow cytometry data. Whistler, British Columbia, Canada, December 2009. Clustering Theory Workshop, Neural Information Processing Systems (NIPS). http://clusteringtheory.org/papers/empiricalmetrics.pdf. → pages 13, 20, 76

[2] N. Aghaeepour, R. Nikolic, H. Hoos, and R. Brinkman. Rapid cell population identification in flow cytometry data. *Cytometry Part A*, 79(1): 6–13, 2011. ISSN 1552-4930. → pages 35, 49, 72, 78

[3] N. Aghaeepour, P. K. Chattopadhyay, A. Ganesan, K. O'Neill, H. Zare, A. Jalali, H. H. Hoos, M. Roederer, and R. R. Brinkman. Early Immunologic Correlates of HIV Protection can be Identified from Computational Analysis of Complex Multivariate T-cell Flow Cytometry Assays. *Bioinformatics*, 28(7):1009–1016, 2012. → pages 49, 50, 56, 62, 72, 78

[4] S. Altschuler and L. Wu. Cellular heterogeneity: do differences make a difference? *Cell*, 141(4):559–563, 2010. → pages 38

[5] A. Azad, S. Pyne, and A. Pothen. Matching phosphorylation response patterns of antigen-receptor-stimulated T cells via flow cytometry. *BMC Bioinformatics*, 13(Suppl 2):S10, 2012. → pages 72

[6] B. Bain. *Blood cells: A practical guide*. Wiley Online Library, fourth edition, 2006. → pages 89

[7] J. Bard, S. Rhee, and M. Ashburner. An ontology for cell types. *Genome Biology*, 6(2):R21, 2005. → pages 38

[8] A. Bashashati and R. Brinkman. A Survey of Flow Cytometry Data Analysis Methods, 2009. → pages 1, 3, 8, 9, 108

[9] A. Bashashati, N. Johnson, A. Khodabakhshi, M. Whiteside, H. Zare, D. Scott, K. Lo, R. Gottardo, F. Brinkman, J. Connors, et al. B cells with high side scatter parameter by flow cytometry correlate with inferior survival in diffuse large b-cell lymphoma. *American Journal of Clinical Pathology*, 137(5):805–814, 2012. → pages 6, 49, 111

[10] J. Baudry, A. Raftery, G. Celeux, K. Lo, and R. Gottardo. Combining mixture components for clustering. *Journal of Computational and Graphical Statistics*, 19(2):332–353, 2010. → pages 4

[11] S. Bendall, E. Simonds, P. Qiu, E. Amir, P. Krutzik, R. Finck, R. Bruggner, R. Melamed, A. Trejo, O. Ornatsky, et al. Single-cell mass cytometry of differential immune and drug responses across a human hematopoietic continuum. *Science*, 332(6030):687, 2011. → pages 1, 24, 49, 55, 57, 69, 72

[12] S. Bendall, G. Nolan, M. Roederer, and P. Chattopadhyay. A deep profiler's guide to cytometry. *Trends in Immunology*, 2012. → pages 48

[13] A. Biancotto, P. Dagur, J. Chris Fuchs, M. Langweiler, and J. Philip McCoy Jr. OMIP-004: In-depth characterization of human T regulatory cells. *Cytometry Part A*, 81:360–361, 2011. → pages 47

[14] N. Breslow. Analysis of survival data under the proportional hazards model. *International Statistical Review/Revue Internationale de Statistique*, pages 45–57, 1975. → pages 27

[15] R. R. Brinkman, M. Gasparetto, S. J. Lee, A. J. Ribickas, J. Perkins, W. Janssen, R. Smiley, and C. Smith. High-content flow cytometry and temporal data analysis for defining a cellular signature of graft-versus-host disease. *Biology of blood and marrow transplantation : Journal of the American Society for Blood and Marrow Transplantation*, 13(6):691–700, Jun 2007. → pages 14

[16] R. Burgoyne and D. Tan. Prolongation and quality of life for HIV-infected adults treated with highly active antiretroviral therapy (HAART): A balancing act. *Journal of Antimicrobial Chemotherapy*, 61(3):469–474, 2008. ISSN 0305-7453. → pages 24

[17] A. Califano, M. Kellis, and G. Stolovitzky. Preface: Recomb systems biology, regulatory genomics, and dream 2011 special issue. *Journal of Computational Biology*, 19(2):101–101, 2012. → pages 86, 93

[18] W. Cao. Molecular characterization of human plasmacytoid dendritic cells. *Journal of Clinical Immunology*, 29(3):257–264, 2009. → pages 68

[19] K. Castro, J. Ward, L. Slutsker, J. Buehler, H. Jaffe, R. Berkelman, and J. Curran. Revised classification system for HIV infection and expanded surveillance case definition for AIDS among adolescents and adults. *MMWR Recomm Rep*, 41:1–19, 1992. → pages 25, 56

[20] C. Chan, F. Feng, J. Ottinger, D. Foster, M. West, and T. Kepler. Statistical mixture modeling for cell subtype identification in flow cytometry. *Cytometry Part A*, 73(8):693–701, 2008. → pages 4, 49, 78

[21] C. Chan, L. Lin, J. Frelinger, V. Hérbert, D. Gagnon, C. Landry, R. Sékaly, J. Enzor, J. Staats, K. Weinhold, et al. Optimization of a highly standardized carboxyfluorescein succinimidyl ester flow cytometry panel and gating strategy design using discriminative information measure evaluation. *Cytometry Part A*, pages 1126–1136, 2010. → pages 49

[22] P. Chattopadhyay and M. Roederer. Cytometry: Today's technology and tomorrow's horizons. *Methods*, Feb 2012. → pages 1, 5, 55

[23] P. Chattopadhyay, C. Hogerkorp, and M. Roederer. A chromatic explosion: the development and future of multiparameter flow cytometry. *Immunology*, 125(4):441–449, 2008. ISSN 1365-2567. → pages 24

[24] P. Chattopadhyay, J. Melenhorst, K. Ladell, E. Gostick, P. Scheinberg, A. Barrett, L. Wooldridge, M. Roederer, A. Sewell, and D. Price. Techniques to improve the direct ex vivo detection of low frequency antigen-specific CD8+ T cells with peptide-major histocompatibility complex class I tetramers. *Cytometry Part A*, 73(11):1001–1009, 2008. ISSN 1552-4930. → pages 24

[25] P. Chattopadhyay, M. Roederer, and D. Price. OMIP-002: Phenotypic analysis of specific human CD8+ T-cells using peptide-MHC class I multimers for any of four epitopes. *Cytometry Part A*, 77(9):821–822, 2010. → pages 47

[26] J. Conway and D. Coombs. A stochastic model of latently infected cell reactivation and viral blip generation in treated hiv patients. *PLoS Computational Biology*, 7(4):e1002033, 2011. → pages 24

[27] E. Costa, C. Pedreira, S. Barrena, Q. Lecrevisse, J. Flores, S. Quijano, J. Almeida, M. del Carmen García-Macias, S. Bottcher, J. Van Dongen,

et al. Automated pattern-guided principal component analysis vs expert-based immunophenotypic classification of b-cell chronic lymphoproliferative disorders: a step forward in the standardization of clinical immunophenotyping. *Leukemia*, 24(11):1927–1933, 2010. → pages 6, 49, 72

[28] S. De Rosa, L. Herzenberg, L. Herzenberg, and M. Roederer. 11-color, 13-parameter flow cytometry: identification of human naive T cells by phenotype, function, and T-cell receptor diversity. *Nature Medicine*, 7(2): 245–248, 2001. → pages 38

[29] T. Duong, A. Cowling, I. Koch, and M. Wand. Feature significance for multivariate kernel density estimation. *Computational Statistics and Data Analysis*, 52(9):4225–4242, 2008. → pages 3, 10

[30] C. Dym, W. Wood, and M. Scott. Rank ordering engineering designs: pairwise comparison charts and Borda counts. *Research in Engineering Design*, 13(4):236–242, 2002. ISSN 0934-9839. → pages 76

[31] M. Elemans, R. Thiébaut, A. Kaur, and B. Asquith. Quantification of the Relative Importance of CTL, B Cell, NK Cell, and Target Cell Limitation in the Control of Primary SIV-Infection. *PLoS computational biology*, 7 (3):e1001103, 2011. → pages 39

[32] M. Eller and J. Currier. OMIP-007: Phenotypic analysis of human natural killer cells. *Cytometry Part A*, 2012. → pages 47

[33] D. Eppstein. Finding the k shortest paths. *SIAM J. Comput.*, 28(2): 652–673, 1998. → pages 53

[34] M. Ester, H. Kriegel, J. Sander, and X. Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proc. KDD*, volume 96, pages 226–231, 1996. → pages 3

[35] B. Everitt, S. Landau, and M. Leese. *Cluster Analysis*, volume 4. Arnold, London, 2001. ISBN 9780340761199. → pages 28

[36] E. K. F. The L2 Discrepancy Framework to Mine High-Throughput Screening Data for Targeted Drug Discovery: Application to AIDS Antiviral Activity Data of The National Cancer Institute. SIAM, 2006. http://www.siam.org/meetings/sdm06/workproceed/Scientific Datasets/Elkhettabi.pdf. → pages 78

[37] G. Finak, A. Bashashati, R. Brinkman, and R. Gottardo. Merging mixture components for cell population identification in flow cytometry. *Advances in Bioinformatics*, 2009. → pages 4, 5

[38] G. Finak, A. Bashashati, R. Brinkman, and R. Gottardo. Merging mixture components for cell population identification in flow cytometry. *Advances in Bioinformatics*, v09, 2009. → pages 49, 72, 78

[39] K. Foulds, M. Donaldson, and M. Roederer. OMIP-005: Quality and phenotype of antigen-responsive rhesus macaque T cells. *Cytometry Part A*, pages 360–361, 2012. → pages 47

[40] B. Franz, K. F. May, G. Dranoff, and K. Wucherpfennig. Ex vivo characterization and isolation of rare memory B cells with antigen tetramers. *Blood*, 118:348–357, Jul 2011. → pages 38

[41] F. G, B. A, B. R, and G. R. Merging mixture model components for improved cell population identification in high throughput flow cytometry data. *Advances in Bioinformatics*, page to appear, 2009. → pages 9, 17

[42] A. Ganesan, P. K. Chattopadhyay, T. M. Brodie, J. Qin, W. Gu, J. R. Mascola, N. L. Michael, D. A. Follmann, M. Roederer, C. Decker, T. Whitman, S. Tasker, A. Weintrob, G. Wortmann, M. Zapor, M. Landrum, V. Marconi, J. Okulicz, N. Crum-Cianflone, M. Bavaro, H. Chun, R. V. Barthel, A. Johnson, B. Agan, N. Aronson, W. Bradley, G. Gandits, L. Jagodzinski, R. O'Connell, C. Eggleston, and J. Powers. Immunologic and virologic events in early HIV infection predict subsequent rate of progression. *J. Infect. Dis.*, 201:272–284, Jan 2010. → pages 24, 25, 26, 35, 37, 39, 56, 62

[43] L. Gattinoni, E. Lugli, Y. Ji, Z. Pos, C. Paulos, M. Quigley, J. Almeida, E. Gostick, Z. Yu, C. Carpenito, et al. A human memory t cell subset with stem cell-like properties. *Nature Medicine*, pages 1290–1297, 2011. → pages 48

[44] Y. Ge and S. Sealfon. flowPeaks: a fast unsupervised clustering for flow cytometry data via K-means and density peak finding. *Bioinformatics*, 2012. → pages 4, 72, 78, 109

[45] R. C. Gentleman, V. J. Carey, D. M. Bates, B. Bolstad, M. Dettling, S. Dudoit, B. Ellis, L. Gautier, Y. Ge, J. Gentry, K. Hornik, T. Hothorn, W. Huber, S. Iacus, R. Irizarry, F. Leisch, C. Li, M. Maechler, A. J. Rossini, G. Sawitzki, C. Smith, G. Smyth, L. Tierney, J. Y. H. Yang, and J. Zhang.

Bioconductor: Open software development for computational biology and bioinformatics. *Genome Biology*, 5:R80, 2004. URL http://genomebiology.com/2004/5/10/R80. → pages 21, 26

[46] S. Gordon, B. Cervasi, P. Odorizzi, R. Silverman, F. Aberra, G. Ginsberg, J. Estes, M. Paiardini, I. Frank, and G. Silvestri. Disruption of intestinal CD4+ T cell homeostasis is a key marker of systemic CD4+ T cell activation in HIV-infected individuals. *The Journal of Immunology*, 185(9): 5169, 2010. → pages 63

[47] J. Gratama, J. Kraan, M. Keeney, V. Granger, and D. Barnett. Reduction of variation in T-cell subset enumeration among 55 laboratories using single-platform, three or four-color flow cytometry based on CD45 and SSC-based gating of lymphocytes. *Cytometry Part B: Clinical Cytometry*, 50(2):92–101, 2002. → pages 2, 20

[48] F. Hahne, A. Khodabakhshi, A. Bashashati, C. Wong, R. Gascoyne, A. Weng, V. Seyfert-Margolis, K. Bourcier, A. Asare, T. Lumley, et al. Per-channel basis normalization methods for flow cytometry data. *Cytometry Part A*, 77(2):121–131, 2009. → pages 2, 14

[49] F. Hahne, N. LeMeur, R. Brinkman, B. Ellis, P. Haaland, D. Sarkar, J. Spidlen, E. Strain, and R. Gentleman. flowCore: a Bioconductor package for high throughput flow cytometry. *BMC Bioinformatics*, 10(1):106, 2009. ISSN 1471-2105. → pages 26, 78

[50] G. Hamerly and C. Elkan. Learning the K in k-means. *Advances in Neural Information Processing Systems*, 17:281–288, 2004. → pages 9

[51] J. Handl, J. Knowles, and D. B. Kell. Computational cluster validation in post-genomic data analysis. *Bioinformatics (Oxford, England)*, 21(15): 3201–3212, Aug 1 2005. → pages 2

[52] D. Harrington. Linear rank tests in survival analysis. *Encyclopedia of biostatistics*, 2005. → pages 27

[53] T. Hesterberg, D. Moore, S. Monaghan, A. Clipson, and R. Epstein. Bootstrap methods and permutation tests. *Introduction to the Practice of Statistics*, 47(4):1–70, 2005. ISSN 0040-1706. → pages 27

[54] H. Hoos. Programming by optimization. *Communications of the ACM*, 55 (2):70–80, 2012. → pages 114

[55] K. Hornik. A CLUE for CLUster Ensembles. *Journal of Statistical Software*, 14(12), September 2005. URL http://www.jstatsoft.org/v14/i12/. → pages 78, 81

[56] K. Hornik and W. Bohm. Hard and Soft Euclidean Consensus Partitions. *Data Analysis, Machine Learning and Applications*, pages 147–154, 2008. → pages 78, 81

[57] C. Igel, T. Suttorp, and N. Hansen. A computational efficient covariance matrix update and a (1+ 1)-CMA for evolution strategies. In *Proceedings of the 8th annual conference on Genetic and evolutionary computation*, page 460. ACM, 2006. → pages 20

[58] H. Jaspan, L. Liebenberg, W. Hanekom, W. Burgers, D. Coetzee, A. Williamson, F. Little, L. Myer, R. Coombs, D. Sodora, et al. Immune activation in the female genital tract during hiv infection predicts mucosal cd4 depletion and hiv shedding. *Journal of Infectious Diseases*, 204(10): 1550–1556, 2011. → pages 63

[59] L. Kaufman and P. Rousseeuw. *Finding groups in data: an introduction to cluster analysis*. Wiley New York, 1990. → pages 9

[60] M. Kitahata, S. Gange, A. Abraham, B. Merriman, M. Saag, A. Justice, R. Hogg, S. Deeks, J. Eron, J. Brooks, et al. Effect of early versus deferred antiretroviral therapy for HIV on survival. *New England Journal of Medicine*, 360(18):1815–1826, 2009. ISSN 0028-4793. → pages 24

[61] R. Klausner, A. Fauci, L. Corey, G. Nabel, H. Gayle, S. Berkley, B. Haynes, D. Baltimore, C. Collins, R. Douglas, et al. Enhanced: The need for a global HIV vaccine enterprise. *Science*, 300(5628):2036, 2003. → pages 24

[62] B. Korber, M. LaBute, and K. Yusim. Immunoinformatics comes of age. *PLoS Computational Biology*, 2(6):e71, 2006. → pages 34

[63] A. Krug, A. Towarowski, S. Britsch, S. Rothenfusser, V. Hornung, R. Bals, T. Giese, H. Engelmann, S. Endres, A. Krieg, et al. Toll-like receptor expression reveals CpG DNA as a unique microbial stimulus for plasmacytoid dendritic cells which synergizes with CD40 ligand to induce high amounts of IL-12. *European Journal of Immunology*, 31(10): 3026–3037, 2001. → pages 68

[64] P. Krutzik and G. Nolan. Fluorescent cell barcoding in flow cytometry allows high-throughput drug screening and signaling profiling. *Nature Methods*, 3(5):361–368, 2006. → pages 36

[65] D. Kuhrt. *SIV infection results in detrimental phenotypic and functional alterations of the naive and memory B cell compartments that are initiated during acute infection*. PhD thesis, School of Medicine, University of Pittsburgh, 2010. → pages 24

[66] L. Lamoreaux, R. Koup, and M. Roederer. OMIP-009: Characterization of antigen-specific human T-cells. *Cytometry Part A*, 2012. → pages 47

[67] G. Lee, W. Finn, and C. Scott. Statistical file matching of flow cytometry data. *Journal of Biomedical Informatics*, 2011. → pages 94

[68] J. Lee, J. Spidlen, K. Boyce, J. Cai, N. Crosbie, M. Dalphin, J. Furlong, M. Gasparetto, M. Goldberg, E. Goralczyk, et al. MIFlowCyt: the minimum information about a Flow Cytometry Experiment. *Cytometry Part A*, 73(10):926–930, 2008. → pages 95

[69] S. Lloyd. Least squares quantization in pcm. *Information Theory, IEEE Transactions on*, 28(2):129–137, 1982. → pages 108

[70] K. Lo, R. Brinkman, and R. Gottardo. Automated gating of flow cytometry data via robust model-based clustering. *Cytometry Part A*, 73(4):321–332, 2008. → pages 3, 4, 49, 72, 78

[71] E. Lugli, M. Roederer, and A. Cossarizza. Data analysis in flow cytometry: the future just started. *Cytometry Part A*, 77(7):705–713, 2010. ISSN 1552-4930. → pages 1, 108

[72] A. Maddox, M. Keating, J. Trujillo, A. Cork, E. Youness, M. Ahearn, K. McCredie, and E. Freireich. Philadelphia chromosome-positive adult acute leukemia with monosomy of chromosome number seven: a subgroup with poor response to therapy. *Leukemia Research*, 7(4):509–522, 1983. → pages 89

[73] H. Maecker, A. Rinfret, P. D'Souza, J. Darden, E. Roig, C. Landry, P. Hayes, J. Birungi, O. Anzala, M. Garcia, et al. Standardization of cytokine flow cytometry assays. *BMC Immunology*, 6(1):13, 2005. → pages 91

[74] H. Maecker, J. McCoy, M. Amos, J. Elliott, A. Gaigalas, L. Wang, R. Aranda, J. Banchereau, C. Boshoff, J. Braun, et al. A model for harmonizing flow cytometry in clinical trials. *Nature Immunology*, 11(11): 975–978, 2010. → pages 114

[75] H. T. Maecker, J. P. McCoy, and R. Nussenblatt. Standardizing immunophenotyping for the Human Immunology Project. *Nature Reviews Immunology*, 12:191–200, 2012. → pages 5, 47, 94

[76] Y. Mahnke and M. Roederer. OMIP-001: Quality and phenotype of Ag-responsive human T-cells. *Cytometry Part A*, 77(9):819–820, 2010. → pages 47

[77] T. Marafioti, J. Paterson, E. Ballabio, K. Reichard, S. Tedoldi, K. Hollowood, M. Dictor, M. Hansmann, S. Pileri, M. Dyer, et al. Novel markers of normal and neoplastic human plasmacytoid dendritic cells. *Blood*, 111(7):3778–3792, 2008. → pages 68

[78] J. Mattapallil, D. Douek, B. Hill, Y. Nishimura, M. Martin, and M. Roederer. Massive infection and loss of memory CD4 T cells in multiple tissues during acute SIV infection. *Nature*, 434(7037):1093–1097, 2005. → pages 24

[79] P. Meyer, L. Alexopoulos, T. Bonk, A. Califano, C. Cho, A. de la Fuente, D. de Graaf, A. Hartemink, J. Hoeng, N. Ivanov, et al. Verification of systems biology research in the age of collaborative competition. *Nature Biotechnology*, 29(9):811–815, 2011. → pages 86, 93

[80] W. Moore and D. Parks. Update for the logicle data scale including operational code implementations. *Cytometry Part A*, 2012. → pages 94

[81] D. Murdoch, J. Staats, and K. Weinhold. OMIP-006: Phenotypic subset analysis of human T regulatory cells via polychromatic flow cytometry. *Cytometry Part A*, 81A:281–283, 2012. → pages 47

[82] R. Murphy. Automated identification of subpopulations in flow cytometric list mode data using cluster analysis. *Cytometry Part A*, 6(4):302–309, 2005. → pages 8

[83] I. Naim, S. Datta, G. Sharma, J. Cavenaugh, and T. Mosmann. SWIFT: Scalable weighted iterative sampling for flow cytometry clustering. *Proc. IEEE Intl. Conf. Acoustics Speech and Sig. Proc.*, pages 509–512, 2010. → pages 4, 78

[84] U. Naumann and M. Wand. Automation in high-content flow cytometry screening. *Cytometry Part A*, 75:789–797, 2009. → pages 3, 4, 10

[85] U. Naumann, G. Luta, and M. Wand. The curvHDR method for gating flow cytometry samples. *BMC bioinformatics*, 11(1):44, 2010. ISSN 1471-2105. → pages 49, 72, 78

[86] C. Needham, J. Bradford, A. Bulpitt, and D. Westhead. A primer on learning in Bayesian networks for computational biology. *PLoS Comput Biol*, 3(8):e129, 2007. → pages 37

[87] W. Noble. How does multiple testing correction work? *Nature Biotechnology*, 27(12):1135–1137, 2009. ISSN 1087-0156. → pages 36

[88] F. Notta, S. Doulatov, E. Laurenti, A. Poeppl, I. Jurisica, and J. Dick. Isolation of single human hematopoietic stem cells capable of long-term multilineage engraftment. *Science*, 333(6039):218, 2011. → pages 38

[89] P. Nurse. Systems biology: Understanding cells. *Nature*, 424(6951): 883–883, 2003. ISSN 0028-0836. → pages 38

[90] O. Ornatsky, D. Bandura, V. Baranov, M. Nitz, M. Winnik, and S. Tanner. Highly multiparametric analysis by mass cytometry. *Journal of Immunological Methods*, 361(6030):1–20, 2010. → pages 24, 55

[91] D. Pelleg and A. Moore. X-means: Extending K-means with efficient estimation of the number of clusters. In *Proceedings of the Seventeenth International Conference on Machine Learning table of contents*, pages 727–734. Morgan Kaufmann Publishers Inc. San Francisco, CA, USA, 2000. → pages 9

[92] S. Perfetto, P. Chattopadhyay, and M. Roederer. Seventeen-colour flow cytometry: unravelling the immune system. *Nature Reviews Immunology*, 4 (8):648–655, 2004. → pages 48

[93] S. Perfetto, P. Chattopadhyay, L. Lamoreaux, R. Nguyen, D. Ambrozak, R. Koup, and M. Roederer. Amine reactive dyes: an effective tool to discriminate live and dead cells in polychromatic flow cytometry. *Journal of Immunological Methods*, 313(1-2):199–208, 2006. ISSN 0022-1759. → pages 24

[94] J. Peters and M. Ansari. Multiparameter flow cytometry in the diagnosis and management of acute leukemia. *Archives of Pathology & Laboratory Medicine*, 135(1):44–54, 2011. → pages 90

[95] F. Preijers, E. Huys, and B. Moshaver. OMIP-010: A new 10-color monoclonal antibody panel for polychromatic immunophenotyping of small hematopoietic cell samples. *Cytometry Part A*, 2012. → pages 47

[96] R. Prill, D. Marbach, J. Saez-Rodriguez, P. Sorger, L. Alexopoulos, X. Xue, N. Clarke, G. Altan-Bonnet, and G. Stolovitzky. Towards a rigorous assessment of systems biology models: the dream3 challenges. *PloS ONE*, 5(2):e9202, 2010. → pages 86, 93

[97] S. Pyne, X. Hu, K. Wang, E. Rossin, T. Lin, L. Maier, C. Baecher-Allan, G. McLachlan, P. Tamayo, D. Hafler, et al. Automated high-dimensional flow cytometric data analysis. *Proceedings of the National Academy of Sciences*, 106(21):8519, 2009. → pages 4, 49, 72, 78, 94

[98] S. Pyne, X. Hu, K. Wang, E. Rossin, T.-I. Lin, L. M. Maier, C. Baecher-Allan, G. J. McLachlan, P. Tamayo, D. A. Hafler, P. L. D. Jager, and J. P. Mesirov. Automated high-dimensional flow cytometric data analysis. *Proc Natl Acad Sci U S A*, 2009. → pages 6

[99] Y. Qian, C. Wei, F. Eun-Hyung Lee, J. Campbell, J. Halliley, J. Lee, J. Cai, Y. Kong, E. Sadat, E. 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. → pages 3, 4, 49, 72, 78

[100] Y. Qian, Y. Liu, J. Campbell, E. Thomson, Y. Kong, and R. Scheuermann. FCSTrans: An open source software system for FCS file conversion and data transformation. *Cytometry Part A*, 2012. → pages 94

[101] P. Qiu, E. Simonds, S. Bendall, K. Gibbs Jr, R. Bruggner, M. Linderman, K. Sachs, G. Nolan, and S. Plevritis. Extracting a cellular hierarchy from high-dimensional cytometry data with spade. *Nature Biotechnology*, 29: 886–891, 2011. → pages 49, 69, 72, 78

[102] J. Quinn, P. Fisher, R. Capocasale, R. Achuthanandam, M. Kam, P. Bugelski, and L. Hrebien. A statistical pattern recognition approach for determining cellular viability and lineage phenotype in cultured cells and murine bone marrow. *Cytometry Part A*, 71(8):612–624, 2007. ISSN 1552-4930. → pages 72, 78

[103] D. Rocke, T. Ideker, O. Troyanskaya, J. Quackenbush, and J. Dopazo. Papers on normalization, variable selection, classification or clustering of

microarray data. *Bioinformatics*, 25(6):701, 2009. ISSN 1367-4803. → pages 91

[104] M. Roederer and A. Tárnok. OMIPsOrchestrating multiplexity in polychromatic science. *Cytometry Part A*, 77(9):811–812, 2010. → pages 47

[105] M. Roederer, J. Nozzi, and M. Nason. Spice: Exploration and analysis of post-cytometric complex multivariate datasets. *Cytometry Part A*, 79(2): 167–174, 2011. → pages 49, 72

[106] A. Rosenberg and J. Hirschberg. V-measure: A conditional entropy-based external cluster evaluation measure. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning(EMNLP-CoNLL), Prague, Czech Republic*, pages 410–420, 2007. → pages 13

[107] F. Sallusto, D. Lenig, R. Förster, M. Lipp, and A. Lanzavecchia. Two subsets of memory T lymphocytes with distinct homing potentials and effector functions. *Nature*, 402:34–38, 1999. → pages 36

[108] C. Satoh, K. Dan, T. Yamashita, R. Jo, H. Tamura, and K. Ogata. Flow cytometric parameters with little interexaminer variability for diagnosing low-grade myelodysplastic syndromes. *Leukemia Research*, 32(5): 699–707, 2008. → pages 2, 20

[109] P. Sax and L. Baden. When to Start Antiretroviral TherapyReady When You Are? *New England Journal of Medicine*, 360(18):1897–1899, 2009. ISSN 0028-4793. → pages 24

[110] P. Schuster, N. Donhauser, K. Pritschet, M. Ries, S. Haupt, N. Kittan, K. Korn, and B. Schmidt. Co-ordinated regulation of plasmacytoid dendritic cell surface receptors upon stimulation with herpes simplex virus type 1. *Immunology*, 129(2):234–247, 2010. → pages 68

[111] D. Scott. *Multivariate density estimation: theory, practice, and visualization*. Wiley-Interscience, 1992. → pages 10

[112] A. Slogrove, B. Reikie, S. Naidoo, C. De Beer, K. Ho, M. Cotton, J. Bettinger, D. Speert, M. Esser, and T. Kollmann. Hiv-exposed uninfected infants are at increased risk for severe infections in the first year of life. *Journal of Tropical Pediatrics*, 2012. → pages 93

[113] B. Smith, M. Ashburner, C. Rosse, J. Bard, W. Bug, W. Ceusters, L. Goldberg, K. Eilbeck, A. Ireland, C. Mungall, et al. The OBO Foundry: Coordinated evolution of ontologies to support biomedical data integration. *Nature Biotechnology*, 25(11):1251–1255, 2007. → pages 38

[114] T. Sørensen. A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on danish commons. *Biol. skr.*, 5:1–34, 1948. → pages 85

[115] G. Stolovitzky, R. Prill, and A. Califano. Lessons from the dream2 challenges. *Annals of the New York Academy of Sciences*, 1158(1): 159–195, 2009. → pages 86, 93

[116] M. Suchard, Q. Wang, C. Chan, J. Frelinger, A. Cron, and M. West. Understanding GPU programming for statistical computation: Studies in massively parallel massive mixtures. *Journal of Computational and Graphical Statistics*, 19(2):419–438, 2010. ISSN 1061-8600. → pages 72

[117] I. Sugár and S. Sealfon. Misty Mountain clustering: application to fast unsupervised flow cytometry gating. *BMC Bioinformatics*, 11:502, 2010. → pages 4, 49, 72, 78

[118] R. Suzuki and H. Shimodaira. Pvclust: an r package for assessing the uncertainty in hierarchical clustering. *Bioinformatics*, 22(12):1540–1542, 2006. → pages 71

[119] M. Swiecki and M. Colonna. Unraveling the functions of plasmacytoid dendritic cells during viral infections, autoimmunity, and tolerance. *Immunological Reviews*, 234(1):142–162, 2010. → pages 68

[120] C. Tecimer, B. Loy, and A. Martin. Acute myeloblastic leukemia (m0) with an unusual chromosomal abnormality:: Translocation (1; 14)(p13; q32). *Cancer Genetics and Cytogenetics*, 111(2):175–177, 1999. → pages 89

[121] M. Van Blerk, M. Bernier, X. Bossuyt, B. Chatelain, J. D Hautcourt, C. Demanet, L. Kestens, D. Van Bockstaele, T. Crucitti, and J. Libeer. National external quality assessment scheme for lymphocyte immunophenotyping in Belgium. *Clinical Chemistry and Laboratory Medicine*, 41:323–330, 2003. → pages 2, 20

[122] R. Veazey, M. DeMaria, L. Chalifoux, D. Shvetz, D. Pauley, H. Knight, M. Rosenzweig, R. Johnson, R. Desrosiers, and A. Lackner. Gastrointestinal tract as a major site of CD4+ T cell depletion and viral replication in SIV infection. *Science*, 280(5362):427, 1998. → pages 24

[123] Y. Voronin, A. Manrique, and A. Bernstein. The future of hiv vaccine research and the role of the global hiv vaccine enterprise. *Current Opinion in HIV and AIDS*, 5(5):414, 2010. → pages 24

[124] C. Wei, J. Jung, and I. Sanz. OMIP-003: Phenotypic analysis of human memory B cells. *Cytometry Part A*, 79:894–896, 2011. → pages 47

[125] A. Weintrob, A. Fieberg, B. Agan, A. Ganesan, N. Crum-Cianflone, V. Marconi, M. Roediger, S. Fraser, S. Wegner, and G. Wortmann. Increasing age at HIV seroconversion from 18 to 40 years is associated with favorable virologic and immunologic responses to HAART. *JAIDS Journal of Acquired Immune Deficiency Syndromes*, 49(1):40, 2008. ISSN 1525-4135. → pages 25, 55

[126] P. Yang, H. Yang, B. Zhou, Y. Zomaya, et al. A Review of Ensemble Methods in Bioinformatics. *Current Bioinformatics*, 5(4):296–308, 2010. ISSN 1574-8936. → pages 80

[127] H. Zare, P. Shooshtari, A. Gupta, and R. Brinkman. Data reduction for spectral clustering to analyze high throughput flow cytometry data. *BMC Bioinformatics*, 11(1):403, 2010. ISSN 1471-2105. URL http://www.biomedcentral.com/1471-2105/11/403. → pages 3, 4, 9, 19

[128] H. Zare, P. Shooshtari, A. Gupta, and R. Brinkman. Data reduction for spectral clustering to analyze high throughput flow cytometry data. *BMC Bioinformatics*, 11(1):403, 2010. → pages 49, 72, 78

[129] H. Zare, A. Bashashati, R. Kridel, N. Aghaeepour, G. Haffari, J. Connors, R. Gascoyne, A. Gupta, R. Brinkman, and A. Weng. Automated analysis of multidimensional flow cytometry data improves diagnostic accuracy between mantle cell lymphoma and small lymphocytic lymphoma. *American Journal of Clinical Pathology*, 137(1):75–85, 2012. → pages 6, 49, 72, 78, 112

[130] L. Zimmerlin, V. S. Donnenberg, and A. D. Donnenberg. Rare event detection and analysis in flow cytometry: bone marrow mesenchymal stem cells, breast cancer stem/progenitor cells in malignant effusions, and pericytes in disaggregated adipose tissue. *Methods Mol. Biol.*, 699: 251–273, 2011. → pages 38

[131] C. Zuleger and M. Albertini. OMIP-008: Measurement of Th1 and Th2 cytokine polyfunctionality of human T cells. *Cytometry Part A*, 2012. → pages 47