Automated flow cytometric analysis across a large number of samples

Xiaoyi Chen

Université de Cergy-Pontoise

École Doctorale EM2C

Thèse de doctorat

Discipline: Mathématiques appliquées

présentée par

Xiaoyi Chen

Analyse des données de cytométrie en flux pour un grand nombre d'échantillons

dirigée par Bernard Chalmond et Benno Schwikowski

Soutenue le 06/10/2015 devant le jury composé de:

Bernard Chalmond	Univ. de Cergy-Pontoise	Directeur
Stéphane Girard	INRIA Rhône-Alpes	Rapporteur
Christine Graffigne	Univ. Paris Descartes	Rapporteur
Lars Rogge	Institut Pasteur	Examinateur
Benno Schwikowski	Institut Pasteur	CoDirecteur
Spencer Shorte	Institut Pasteur	Examinateur

Résumé

Cette thèse a conduit à la mise au point de deux nouvelles approches statistiques pour l'identification automatique de populations cellulaires en cytométrie de flux multiparamétrique, et ceci pour le traitement d'un grand nombre d'échantillons, chaque échantillon étant prélevé sur un donneur particulier. Ces deux approches répondent à des besoins exprimés dans le cadre du projet Labex "Milieu Intérieur"(ci-après étude MI). Dix panels cytométriques de 8 marqueurs ont été sélectionnés pour la quantification des populations principales et secondaires présentes dans le sang périphérique. Sur la base de ces panels, les données ont été acquises et analysées sur une cohorte de 1000 donneurs sains.

Tout d'abord, nous avons recherché une quantification robuste des principales composantes cellulaires du système immunitaire. Nous décrivons une procédure computationnelle, appelée FlowGM, qui minimise l'intervention de l'utilisateur. Le cœur statistique est fondé sur le modèle classique de mélange de lois gaussiennes. Ce modèle est tout d'abord utilisé pour obtenir une classification initiale, le nombre de classes étant déterminé par le critère d'information BIC. Après cela, une méta-classification, qui consiste en l'étiquetage des classes et la fusion de celles qui ont la même étiquette au regard de la référence, a permis l'identification automatique de 24 populations cellulaires sur quatre panels. Ces identifications ont ensuite été intégrées dans les fichiers de cytométrie de flux standard (FCS), permettant ainsi la comparaison avec l'analyse manuelle opérée par les experts. Nous montrons que la qualité est similaire entre FlowGM et l'analyse manuelle classique pour les lymphocytes, mais notamment que FlowGM montre une meilleure discrimination des sous-populations de monocytes et de cellules dendritiques (DC), qui sont difficiles à obtenir manuellement. FlowGM fournit ainsi une analyse rapide de phénotypes cellulaires et se prête à des études de cohortes.

A des fins d'évaluation, de diagnostic et de recherche, une analyse tenant compte de l'influence de facteurs, comme par exemple les effets du protocole, l'effet de l'âge et du sexe, a été menée. Dans le contexte du projet MI, les 1000 donneurs sains ont été stratifiés selon le sexe et l'âge. Les résultats de l'analyse quantitative faite avec FlowGM

ont été jugés concordants avec l'analyse manuelle qui est considérée comme l'état de l'art. On note surtout une augmentation de la précision pour les populations $CD16^{+}$ [1] et cDC1, où les sous-populations $CD14^{lo}CD16^{hi}$ et HLA-DR^{hi} cDC1 ont été systématiquement identifiées. Nous démontrons que les effectifs de ces deux populations présentent une corrélation significative avec l'âge. En ce qui concerne les populations qui sont connues pour être associées à l'âge, un modèle de régression linéaire multiple a été considéré qui fournit un coefficient de régression renforcé. Ces résultats établissent une base efficace pour l'évaluation de notre procédure FlowGM.

Lors de l'utilisation de FlowGM pour la caractérisation détaillée de certaines sous-populations présentant de fortes variations au travers des différents échantillons, par exemple les cellules T, nous avons constaté que FlowGM était en difficulté. En effet, dans ce cas, l'algorithme EM classique initialisé avec la classification de l'échantillon de référence est insuffisant pour garantir l'alignement et donc l'identification des différentes classes entre tous les échantillons. Nous avons donc amélioré FlowGM en une nouvelle procédure FlowGMP. Pour ce faire, nous avons ajouté au modèle de mélange, une distribution a priori sur les paramètres de composantes, conduisant à un algorithme EM contraint. Enfin, l'évaluation de FlowGMP sur un panel difficile de cellules T a été réalisée, en effectuant une comparaison avec l'analyse manuelle. Cette comparaison montre que notre procédure Bayésienne fournit une identification fiable et efficace des onze sous-populations de cellules T à travers un grand nombre d'échantillons.

[1] CD16 est le nom d'un molécule de cluster de différenciation présent à la surface de nombreuses cellules de l'immunité, donc $CD16^{+}$ indique les celluses portant CD16. Pareil pour CD14, cDC1, HLA-DR mentionnés ici. Nous allons discuter en détail les CD moléculaires dans Section 1.2.1.

Abstract

In the course of my Ph.D. work, I have developed and applied two new computational approaches for automatic identification of cell populations in multi-parameter flow cytometry across a large number of samples. Both approaches were motivated and taken by the LabEX "Milieu Intérieur" study (hereafter MI study). In this project, ten 8-color flow cytometry panels were standardized for assessment of the major and minor cell populations present in peripheral whole blood, and data were collected and analyzed from 1,000 cohorts of healthy donors.

First, we aim at robust characterization of major cellular components of the immune system. We report a computational pipeline, called FlowGM, which minimizes operator input, is insensitive to compensation settings, and can be adapted to different analytic panels. A Gaussian Mixture Model (GMM) - based approach was utilized for initial clustering, with the number of clusters determined using Bayesian Information Criterion. Meta-clustering in a reference donor, by which we mean labeling clusters and merging those with the same label in a pre-selected representative donor, permitted automated identification of 24 cell populations across four panels. Cluster labels were then integrated into Flow Cytometry Standard (FCS) files, thus permitting comparisons to human expert manual analysis. We show that cell numbers and coefficient of variation (CV) are similar between FlowGM and conventional manual analysis of lymphocyte populations, but notably FlowGM provided improved discrimination of "hard-to-gate" monocyte and dendritic cell (DC) subsets. FlowGM thus provides rapid, high-dimensional analysis of cell phenotypes and is amenable to cohort studies.

After having cell counts across a large number of cohort donors, some further analysis (for example, the agreement with other methods, the age and gender effect, etc.) are required naturally for the purpose of comprehensive evaluation, diagnosis and discovery. In the context of the MI project, the 1,000 healthy donors were stratified across gender (50% women and 50% men) and age (20-69 years of age). Analysis was streamlined using our established approach FlowGM, the results were highly concordant with the state-of-art

gold standard manual gating. More important, further precision of the CD16$^+$ [2] monocytes and cDC1 population was achieved using FlowGM, CD14^{lo}CD16^{hi} monocytes and HLA-DR^{hi} cDC1 cells were consistently identified. We demonstrate that the counts of these two populations show a significant correlation with age. As for the cell populations that are well-known to be related to age, a multiple linear regression model was considered, and it is shown that our results provided higher regression coefficient. These findings establish a strong foundation for comprehensive evaluation of our previous work.

When extending this FlowGM method for detailed characterization of certain subpopulations where more variations are revealed across a large number of samples, for example the T cells, we find that the conventional EM algorithm initiated with reference clustering is insufficient to guarantee the alignment of clusters between all samples due to the presence of technical and biological variations. We then improved FlowGM and presented FlowGMP pipeline to address this specific panel. We introduce a Bayesian mixture model by assuming a prior distribution of component parameters and derive a penalized EM algorithm. Finally the performance of FlowGMP on this difficult T cell panel with a comparison between automated and manual analysis shows that our method provides a reliable and efficient identification of eleven T cell subpopulations across a large number of samples.

[2] CD16 is the name of a cluster of differentiation molecule found on the surface of several cell types, thus CD16$^+$ indicates those cells carrying CD16. The same for CD14, cDC1, HLA-DR mentioned here. We will discuss with more details the CD molecular in Section 1.2.1

Acknowledgements

Thanks first to my supervisors, Bernard Chalmond and Benno Schwikowski, for letting me try and make mistakes, offering help whenever I needed it, and for teaching me to think and work like a scientist. I would like to thank the reviewer of this thesis, Stéphane Girard and Christine Graffigne, for sacrificing their time to read my manuscript and to evaluate this work. Thanks also to my committee members, Lars Rogge and Spencer Shorte, for their helpful advices over the years. My thanks go to the Laboratoire AGM (Analyse Géométrie Modélisation), University of Cergy-Pontoise, and LabEX MI project, Institut Pasteur for providing years of funding, which gave me opportunity and confidence to explore my interests. This work would not have been possible without the great people I had the pleasure to collaborate with. I would like to particularly mention Matthew L.Albert from the Institut Pasteur in Paris, scientific coordinator of the LabEX MI project, for his encouragement and affirmation of my work from a biologist's point of view. I am deeply grateful to Alejandra Urrutia, Milena Hasan, Valentina Libri and Benoit Beitz from Center for Human Immunology and Vincent Rouilly from Center for Bioinformatics of Institut Pasteur, for letting me joint their weekly meeting, introducing me to the new world of immunology and spending their valuable time on the interpretation of my results. Many thanks to all past and present members of the Systems Biology Lab for creating a wonderful community and giving me a lot of help and various ideas. Most importantly, I have to thank my family: my parents, my husband and parents-in-law, your support and love make me feel like I can do anything.

Xiaoyi Chen
September 30th, 2015

Contents

List of Figures

List of Tables

Glossary

ADC Analogue-to-Digital Conversion

AIDS Acquired Immune Deficiency Syndrome

BIC Bayesian Information Criterion

CD Cluster of Differentiation

CMV Cytomegalovirus

CV Coefficient of Variation

DC Dendritic Cell

EM Expectation Maximization

FCM Flow Cytometry

FCS Flow Cytometry Standard

FlowGM Flow cytometry analysis by Gaussian Mixture model

FlowGMP Flow cytometry analysis by Gaussian Mixture model Plus

FSC Forward Scatter

GMM Gaussian Mixture Model

HIV Human Immunodeficiency Virus

ILC Innate Lymphoid Cell

MAP Maximum A Posterior estimate

MFI Mean Fluorescence Intensity

MI project "Milieu Intérieur" project

MLE Maximum Likelihood Estimate

NK Natural Killer

PBMC Peripheral Blood Mononuclear Cell

PMN Polymorphonuclear leukocytes

SpA Ankylosing Spondylitis

SSC Side Scatter

Chapter 1

Introduction

1.1 Overview

One central task of immunology is to determine the number and function of various immune cells, to further the understanding of how humans defend themselves against numerous pathogenic microorganisms. As immune cells display a huge diversity, hundreds of subsets can be theoretically distinguished according to their different surface marker proteins, however, in practice, the detailed identification of the immune cell population is challenging. Flow cytometry is a key technology for characterization of the cellular component of the immune system. It allows simultaneous multi-parametric analysis of up to thousands of cells per second, which makes it a powerful tool and commonly used in basic research, clinical practice and clinical trials. Although recent technical advances have enabled comprehensive immunoprofiling of larger cohorts, computational techniques and data analysis tools capable of quantifying multiple cell types across many samples do not yet exist. This thesis presents a pipeline for automated flow cytometry (FCM) analysis across a large number of samples, including the development of a computational method for identification of numerous cell types, the evaluation of the automated flow cytometry analysis, and the improvement of the method for addressing a difficult situation.

With the goal of learning the basics of flow cytometry and sequentially drawing the computational modeling, this chapter supplies a brief introduction of FCM analysis. Each of the following aspects will be described: the principle, instruments, data, and applications. We will give an overview of the flow cytometry technology, then introduce the applications in immunology, highlighting the LabEX MI project, which provided the motivation of my work. Finally, we will discuss previous works in this field. Chapter 2 describes in detail a computational method for FCM data analysis across a large number of samples and cell types. Chapter 3 discusses the comprehensive evaluation study on 600 healthy

donors. Chapter 4 describes a difficult case for automated flow cytometry data analysis, introduces an improvement of the method, and discusses the validation. Finally, chapter 5 provides a brief discussion of the implications of the thesis work and suggests some steps for extending this research.

1.2 Determination of immunophenotypes

Different immune cells are distinct in form and function, but in basic science research and laboratory diagnosis, we could not identify them by observing the form or testing the function. Immunophenotyping is a technique used to quantify cell population by classifying cells on the basis of the proteins expressed on their surface.

1.2.1 Immunophenotype group

The protein located on the cell surface is a key feature for distinguishing cell populations of different types and functions. Each surface molecule which could be bound by two specific monoclonal antibodies[1] is then assigned a CD (cluster of differentiation) number. Cell populations are usually defined using a "+" or a "-" symbol to indicate whether a certain cell fraction expresses or lacks a CD molecule. Some populations can also be defined as ^{hi}, ^{int} or ^{low} (alternatively ^{bright}, ^{mid} or ^{dim}), indicating an overall level in CD expression. More than 300 surface molecules have been named. Based on what combination of CD molecules and how many of each CD molecule are present on the cell surface, very specific cell types are distinguished. This procedure is called as "immunophenotyping". An example is shown in Figure 1.1. The cells carrying CD45, noted as "$CD45^+$", correspond to leukocytes. Within this population, the cells which carry also CD3 molecules are identified as T cells ("$CD45^+CD3^+$"), and we could further classify T cells into "$CD4^+$" and "$CD8^+$" subpopulations, based on whether CD4 or CD8 are present on their surface. Similarly, "$CD45^+CD19^+$", "$CD45^+CD14^+$" and "$CD45^+CD56^+$" correspond to B lymphocytes, monocytes and NK cells, respectively.

[1]Monoclonal antibodies are monospecific antibodies that are made by identical immune cells, in contrast to polyclonal antibodies which are made from several different immune cells. Monoclonal antibodies bind to the same specific substance, they can then serve to detect that substance.

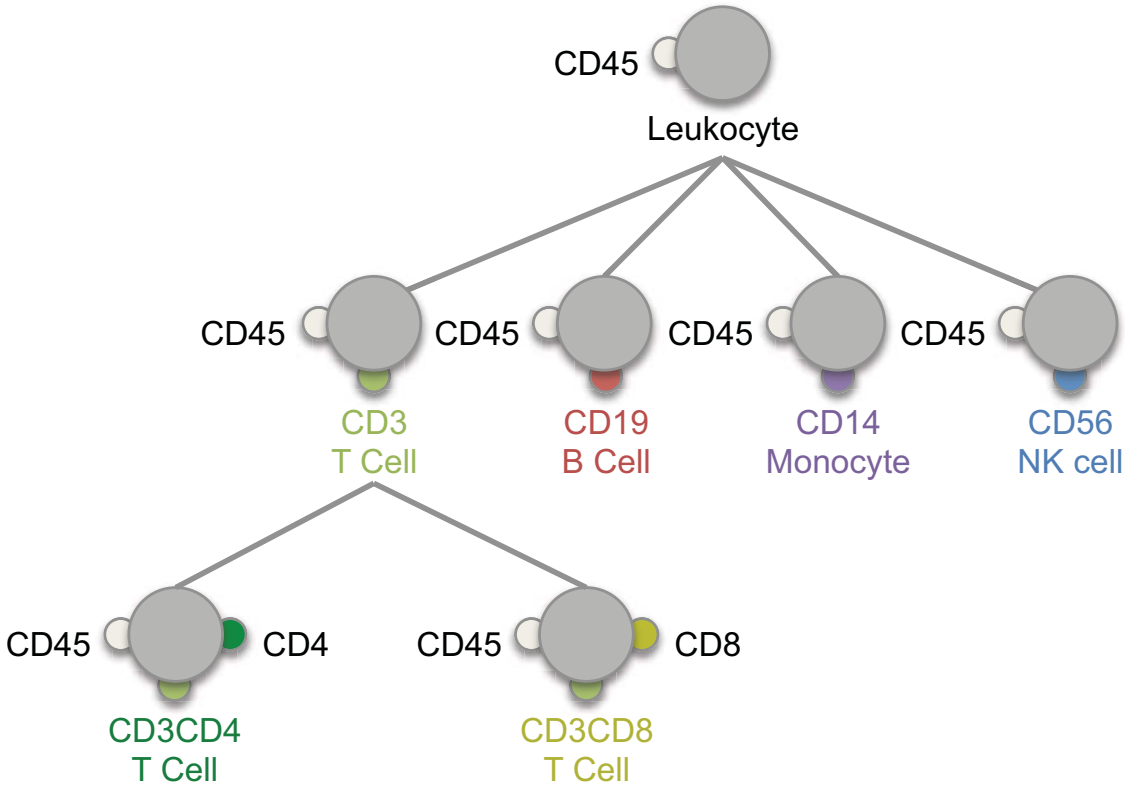

Figure 1.1: The cells carrying CD45, noted as "$CD45^+$", correspond to leukocytes, within this population, the cells which carry also CD3 molecules are identified as T cells("$CD45^+CD3^+$"), and we could further classify T cells into "$CD4^+$" and "$CD8^+$" subpopulations based on whether CD4 or CD8 are present on their surface. Similarly, "$CD45^+CD19^+$", "$CD45^+CD14^+$" and "$CD45^+CD56^+$" correspond to B lymphocytes, monocytes and NK cells respectively.

1.2.2 Flow cytometer

Immunophenotyping is commonly performed using a flow cytometer, which is a laser-based instrument capable of analyzing thousands of cells per second. By suspending cells in the fluid, passed one by one through the optical and electronic detectors and then measuring the photoelectric signals, flow cytometry enables to quantify different parameters of single cells and allows a high-speed, and multi-dimensional quantitative analysis.

The instrument described by Andew Moldavan in 1934 is generally acknowledged to be an early flow cytometer [Givan, 2011]. Following work by Coulter in 1953 disclosed the first impedance-based flow cytometry device [Coulter, 1956]. In 1968 the work by Fulwyler, Dittrich and Göhde led to significant changes in the overall design and resulted in a first fluorescence-based flow cytometry device, which was largely similar to today's

cytometers [Dittrich and Göhde, 1969]. Modern widely used flow cytometers have three main components [Biosciences, 2000]:

- The fluidics system transports cells in a stream of liquid droplets to the laser beam, so that they pass one by one through the light beam for sensing;
- The optics system consists of light sources and the beam collection system;
- The electronics system includes an Analogue-to-Digital Conversion (ADC) system and a data processing system.

In the flow cytometer, suspended cells are labeled with antibodies specific for cell surface CD molecules. The antibodies are tagged with fluorescent dyes of different colors. The labeled cells form a stream of droplets each containing one cell. A beam of laser light is directed onto the stream of fluid. A number of detectors are aimed at the point where the stream passes through the light beam: one in the line of the light beam (called Forward Scatter or FSC), one perpendicular to it (called Side Scatter or SSC) and several fluorescence detectors. The FSC reflects the cell volume, the SSC depends on the inner complexity of the cell (i.e., the shape of the nucleus, the amount and type of cytoplasmic granules or the membrane roughness), and the fluorescence detectors could detect fluorescence chemicals attached to the cell surface, which reflects the quantity of surface CD molecules. By analyzing the fluctuation in brightness at each detector, it is then possible to derive various types of information about the physical and chemical structure of each cell (Figure 1.2A).

For more than 30 years, the fluorescence-based technique of flow cytometry has been widely used by clinicians and researchers. From its early beginning, it has been associated with monoclonal antibodies to identify immunocompetent cells, to quantify changes in expression of surface determinants, and to separate cell subsets prior to the test of their functional properties.

From relative cell counting to absolute cell counting The main advantage of FCM is the identification and enumeration of cell subset in a mixed cell population, for example, the lymphocyte population consists, according to their surface markers, of T lymphocytes ($CD3^+$), B lymphocytes ($CD19^+$) and Natural Killer lymphocytes ($CD56^+CD3^-CD19^-$), and each of these populations could be further subdivided, e.g., within T lymphocytes, there are $CD4^+$ T lymphocytes, $CD8^+$ T lymphocytes, etc. In the past the cell counts of these subsets are mostly expressed as a proportion in a mixed cell population, and do not reflect the absolute number per unit volume of blood. But in the clinical diagnosis of some diseases, such as AIDS, the absolute count needs to be considered: for example,

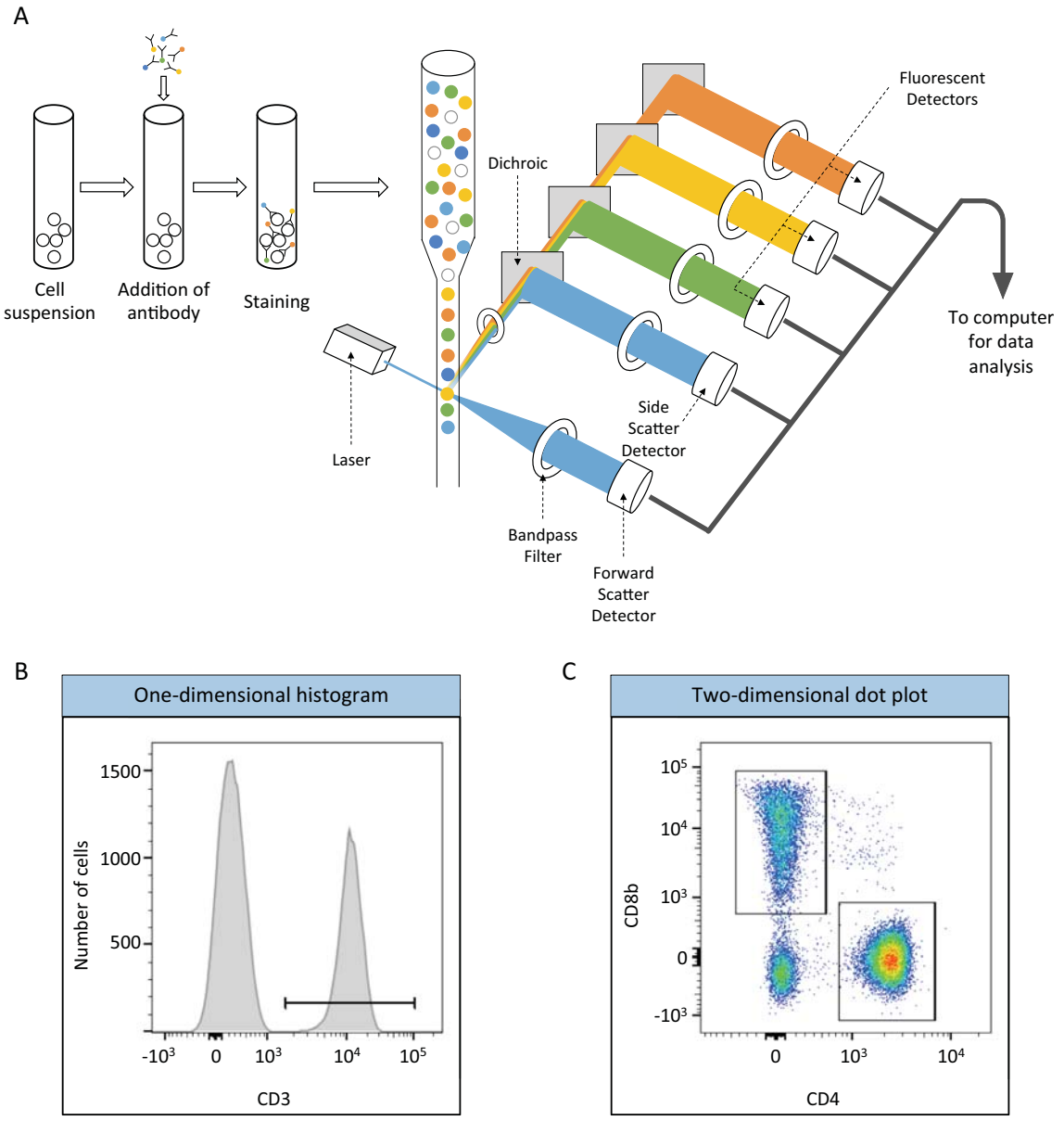

Figure 1.2: Schematic representation of a flow cytometer and flow cytometry data analysis (A) Flow cytometer (B) Flow cytometry data analysis with one-dimensional histogram: the left group of cells which express CD3 are selected. (C) Flow cytometry data analysis with two-dimensional scatter plot: the $CD4^+$ population and $CD8\beta^+$ population are identified by drawing rectangles containing a homogeneous group of cells.

in a healthy, HIV-negative adult, the CD4 count ranges from 500 and 1200 per μL of

blood, then it is recommended to start the treatment if one's CD4 count is around 350, and if it drops below 200, there is increased risk of serious infections. So it is all about the numbers - one's CD4 count is the most important indicator of how the immune system is working and deciding the best way and time to treat the HIV disease [U.S. Department of Health & Human Services, 2014].

From single color to multiple color fluorescence analysis In some work of the 1960s, like during the early stages of Kamentsky's studies on cell classification, people's experience led them to anticipate having to use multiple parameters to develop a discriminant function to identify abnormal cells [Shapiro, 2005]. With the advent of immunofluorescence technique, flow cytometry has developed rapidly from initial indirect immunofluorescence staining to four to six-colored fluorescence analysis. The increased number of lasers and detectors allows for labeling with multiple antibodies [O'Neill et al., 2013], and can identify a target population by their phenotypic markers with better precision. Recent technological advancements have enabled the flow cytometry measurement of 45 parameters on millions of cells.

1.2.3 Flow cytometry data

With the above knowledge, we could clarify several definitions, which is very important for describing formally the flow cytometry data and the problem later.

Although the number of parameters that can be measured at the same time has been increased, it is still limited, given that there are hundreds of CD unique clusters and subclusters that can be distinguished. In practice, one needs to choose appropriate antibodies, limited to the number of markers allowed by the instrument, to determine the differentiation of some cell type of interest. The combination of those chosen antibodies is commonly called a "panel".

As contemporary flow cytometers can measure objects other than cells, including bacteria, sperm, plankton, even viruses, etc [Givan, 2011], "particle" can be used as a more general term for any of the objects flowing through a flow cytometer. "Event" is a term that is used to indicate anything that has been interpreted by the instrument, rightly or wrongly, to be a single particle. For example, two cells close together may actually be detected as one event, called "doublet". Because most of the particles passed through cytometers and detected as events are in fact single cells, those words will be used here somewhat interchangeably.

Now we could define formally the flow cytometry dataset and the analysis problems.

Definition 1.2.1. We denote by N the number of events in a flow cytometry data sample, m the number of markers (or interchangeably number of antibodies, number of colors) in one panel, the data $\underline{\boldsymbol{x}} = \{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_i, \cdots, \boldsymbol{x}_N\}$ is N quantitative m-measurements, each $\boldsymbol{x}_i$ is an m-measurement of one event, and each measurement corresponds to one marker, then x_{ij} is the fluorescence intensity of the jth marker measured on the ith cell, $1 \leq i \leq N$, $1 \leq j \leq m$.

The data set here $\underline{\boldsymbol{x}}$ consists of N m-dimensional vectors, the dimension m is usually around 10, which is limited by the instrument. The number N of data points can be often on the order of 10^6, which is the number of cells in one blood sample.

Problem 1.2.1. We aim: (1) to partition $\underline{\boldsymbol{x}} = \{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_i, \cdots, \boldsymbol{x}_N\}$ into K groups, i.e. to associate for each vector $\boldsymbol{x}_i$, a scalar z_i, $1 \leq z_i \leq K$, indicating which group the ith cell belongs to; (2) to assign a cell type to each group, then obtain for each population the relevant statistics, which include cell counts n_k, cell proportions $\frac{n_k}{N}$, and mean fluorescence intensities (MFIs), which is $\frac{1}{n_k}\sum_{i:z_i=k}\boldsymbol{x}_i$; (3) for further diagnosis, discovery or other biomedical analysis.

The gold standard for the analysis of raw flow cytometry data has until now remained "hand gating" [Aghaeepour et al., 2013a] [Fienberg and Nolan, 2014], where the term "gating" means the extraction of regions in a series of one- or two-dimensional projections of the data $\underline{\boldsymbol{x}}$ containing homogeneous groups of cells. The regions can be intervals on a histogram, or rectangles, polygons or ellipses on a bi-variate scatter plots. Figure 1.2B shows a gate (an interval region of CD3 histogram) for $CD3^+$ population, and Figure 1.2C shows two gates (two rectangles on the CD4/CD8β scatter plot) for two different $CD3^+$ subpopulations.

It is obvious that the manual operation is laborious and subject to biased visual inspection and gate adjustment [Bashashati and Brinkman, 2009] [Lugli et al., 2010]. The concerns grow with increasing numbers m of measured phenotypic markers, as the number of two-dimensional projections that need to be analyzed grows quadratically with m. Moreover, there is a major limitation on that information critical for accurate gating may not be present in the selected (or any) two-dimensional projections, but in higher-dimensional space. As a conclusion, the manual analysis requires an inordinate amount of time and is error-prone, non-reproducible, non-standardized, and not open for re-evaluation, making it the most limiting aspects of flow cytometry technology [Bashashati and Brinkman, 2009].

1.3 Application in immunology

1.3.1 Immune system, immune cell and cell population identification

Immunology is the study of the physiological mechanisms that humans use to defend their bodies from invasion by other organisms. Numerous pathogenic microorganisms, including bacteria, viruses, some members of the fungi and so on, live everywhere around or inhabit healthy human bodies and have the potential to cause disease. Although the skin and mucosal surfaces form physical barriers that separate the body from its external environment and prevent most pathogens from gaining access to the cells and tissues of the body, when the barrier is breached, the immune system is then brought into play [Parham, 2009]. The cells of the immune system that are involved in defending the body against both infectious disease and foreign invaders are principally the white blood cells or leukocytes, and the tissue cells related to them. Different from any other tissue in the body, the cells of the immune system display a huge diversity and hundreds of subsets can be identified even within the same lineage, this could only be technically achieved through flow cytometry which allows the analysis of multiple surface and intracellular markers at the level of single cell [Lugli et al., 2010]. Determining the number and function of subsets of leukocyte is an important means to observe the immune status of organism.

1.3.2 MI project

My work was motivated by the MI project, which is a population-based study coordinated by Institut Pasteur, Paris. It is part of a larger French Governmental Initiative called Investissement d'Avenir - Laboratoire d'Excellence (LabEX). This project is named after the French physiologist Claude Bernard's concept of milieu intérieur and aims to establish the determinants of a healthy immune response by identifying genetic and environmental factors that contribute to the observed heterogeneity of immune responses. Restoring the 'personal' in medical care is a major challenge for medicine, and the driving vision of the project [Thomas et al., 2015]. These efforts will establish parameters for stratifying individuals within a population, thus making it possible to glean meaningful interpretation from measurements of stress-induced host response. In achieving this goal, the project will provide a foundation for defining perturbations in an individual's immune response, thus laying the foundation for personalized medicine.

In order to realize the promise of personalized medicine, one pending challenge is to

establish the boundaries of a healthy immune response, and to correlate this information with human genome variation and environmental factors. To that end, the MI Consortium initiated in September 2012 a 1000-person population-based study in order to assess the determinants of immunologic variance within the general healthy population "Genetic & Environmental Determinants Of Immune Phenotype variance: Establishing A Path Towards Personalized Medicine (ID-RCB Number: 2012-A00238-35)". A de novo study-group of 1,000 healthy individuals (1:1 sex ratio) stratified across five-decades of life (age 20-70) from French-descent and recruited as per a detailed medical questionnaire has been recruited between September 2012 and August 2013. Whole blood was collected for immunologic phenotyping, functional immune stimulation and preparation of DNA for genomic analysis. Fecal samples and nasal swabs were obtained for metagenomic studies. Punch biopsies of the skin were taken and primary fibroblast lines were generated for future mechanistic investigation (Figure 1.3).

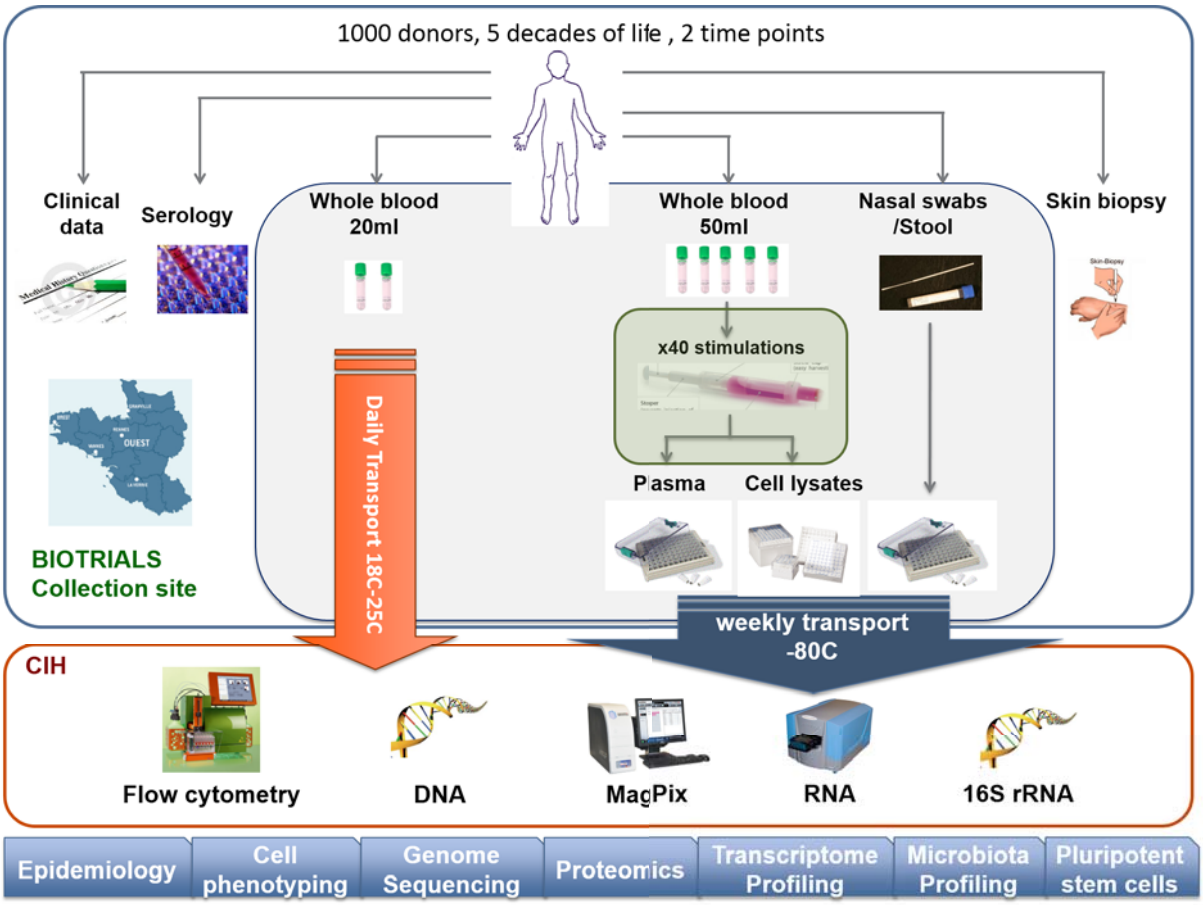

Figure 1.3: Schematic representation of clinical protocol for MI project: This figure was taken from the MI website: `http://www.milieuinterieur.fr`

Cell phenotyping constitutes one of the major datasets to be integrated into the data

warehouse, and as such efforts were made to standardize each step of the sample collection, technical procedures and data analysis. Each pre-analytical aspect of flow cytometry analysis, such as the selection of reagents, the instrumentation, semi-automated staining procedure, quality control and manual gating strategy, were described in [Hasan et al., 2015]. The flow cytometry data of the LabEX MI project consist of 10 different 8-color panels, concerning tens of different cell types. The manual analysis comprises two phases, the first phase is to define a gating strategy specifically for the given panel on selected reference samples, and the second one is to apply the gating strategy to each other cohort samples, visually verify and adjust the gate positions. In order to minimize bias introduced by subjective analysis by different individuals, one panel is commonly required to be analyzed by the same individual for all samples. Thus it could take months to complete the analysis for all cohort donors for one given panel. Moreover, the results obtained from two individuals could be sometime very different.

To address these difficulties, an automated, rapid and un-biased analysis approach for large cohort data set is required. It is critical to this LabEX MI project, and expectantly could be used in other population-based studies and clinical trial settings.

We can formally describe the new and more complicated problem as follows (more details, see Chapter 4 section 4.1):

Problem 1.3.1. Note D the number of samples for a given panel, the objective is to partition each sample $\underline{\boldsymbol{x}}^{(d)} = \{\boldsymbol{x}_1^{(d)}, \cdots, \boldsymbol{x}_i^{(d)}, \cdots, \boldsymbol{x}_{N^{(d)}}^{(d)}\}$ into K groups, $d = 1, 2, \cdots, D$, i.e. to associate with each vector $\boldsymbol{x}_i^{(d)}$ from dth sample, a scalar $z_i^{(d)}$, $1 \leq z_i^{(d)} \leq K$, indicating which group the ith cell belongs to, $1 \leq i \leq N^{(d)}$, and to align automatically cell populations across all samples, which means to ensure that the identical value of $z_i^{(d)}$ from different donor sample d is of the same cell type, such that the counts, proportions, MFIs or other statistics from different cohort samples are comparable.

1.4 Background of automated flow cytometry data analysis

In a statistical learning framework, the cell population identification problem can be viewed as a clustering problem. Clustering is the grouping of a set of objects into subsets or "clusters", such that those within each cluster are more closely related to one another than objects assigned to different clusters [Hastie et al., 2009]. Here we group cells into clusters, such that cells within each cluster are from the same immunophenotype group. There is no delicate objective function, thus a validation by comparison with reference solutions

on given examples, which are usually from manual gating analysis, or comprehensive evaluation are often considered.

1.4.1 Computational approaches for individual samples

A large number of computational approaches have been developed for automated cell population identification, most of which are focused on Problem 1.2.1, i.e. clustering a single sample $\underline{\boldsymbol{x}}$ or very few samples independently.

When considering such a large sample, connectivity-based clustering algorithms, like hierarchical clustering and graph-theoretic clustering methods, often fail. Hierarchical clustering approaches become infeasible and space-inefficient because such algorithms require the continuous storage of the whole dataset to maintain a similarity matrix of size $O(N^2)$ among all possible point pairs. The same argument also applies to graph-theoretic clustering methods, where all possible pairs of points have to be connected via arcs, in a way the set of points forms a fully connected graph, and an adjacency matrix of all pairwise distances has to be continuously stored during the clustering process [Lakoumentas et al., 2009]. Other typical clustering algorithms including density-based clustering, centroid-based clustering, distribution-based clustering, etc, have been discussed in FCM field.

Density clustering

In density-based clustering [Ester et al., 1996] [Kriegel et al., 2011], clusters are defined as connected dense regions in the data space. [Sugár and Sealfon, 2010] have described a unsupervised density contour clustering algorithm, called Misty Mountain, and [Ge and Sealfon, 2012] have developed FlowPeaks which combines K-means and local peaks searching by exploring the density function. This class of algorithms does not use a large similarity matrix, but often require manually calibrated smoothing parameter and threshold to define a "high density region", moreover it is expensive for high-dimensional datasets. Therefore, Misty Mountain and FlowPeaks are all limited to a validation in three or even fewer dimensions.

Centroid-based clustering

In centroid-based clustering, clusters are represented by a central vector, which may not necessarily be one of the data points. *K-means* clustering is the most typical algorithm of this class, which gives a formal definition as an optimization problem: find the K cluster

centers and assign the data points to the nearest cluster, such that the squared distances from the cluster are minimized. [Aghaeepour et al., 2011] have developed flowMeans for automated identification of cell populations in FCM based on *K-means* clustering. By modeling a single population with multiple clusters, it can identify also concave cell populations. [Wilkins et al., 2001] have discussed the *Fuzzy C-means* clustering method in FCM field, which is a soft version of *K-means* allowing each data point having a fuzzy degree of belonging to each cluster.

Distribution-based clustering

In distribution-based clustering, clusters are defined as points belonging most likely to the same distribution. One prominent method is known as mixture models, where the data set is usually modeled with a fixed number of distributions. Once the parameters of each multivariate distribution are estimated, every data point will be assigned to the component (cluster) with maximum posterior probability. FlowClust [Lo et al., 2008] [Lo et al., 2009] models cell populations using mixtures of t-distributions. FLAME [Pyne et al., 2009] uses a mixture of *skew-t*-distribution to make the model more flexible to skewed cell populations.

1.4.2 Consideration for large cohorts

During and after the initialization of our project, there are several works that treat the clustering Problem 1.3.1 in the context of large cohorts.

The recent X-Cyt approach [Hu et al., 2013] was designed explicitly to efficiently address the problem of a large number of samples. However, X-Cyt focuses on few very specific cell types and still requires the definition of a "partitioning schema", a series of mixture models whose sequence and parameters have to be manually configured and calibrated for each cell type of interest in any given analytic panel.

There are also several attempts addressing a large number of samples and the alignment issue with regard to phenotypic relevant clusters despite technical and biological variation. [Cron et al., 2013] and [Dundar et al., 2014] consider Bayesian mixture modeling and Markov chain Monte Carlo algorithms. The main limitation of these two approaches is that there are too many ad hoc parameters. Moreover it takes too long for the computation. [Pyne et al., 2014] improved FlowClust method by introducing a second layer

modeling. For all these three publications, FlowCAP [Aghaeepour et al., 2013a] competition datasets are used as the basis for parameter and method calibration, which is of a much smaller scale than our case.

Chapter 2

Method development - Automated flow cytometry analysis across a large number of samples and cell types (FlowGM)

This chapter is originally appeared in *Clinical Immunology 157(2): 249-260 (2015).* Xiaoyi Chen, Milena Hasan, Valentina Libri, Alejandra Urrutia, Benoit Beitz, Vincent Rouilly, Darragh Duffy, Etienne Patin, Bernard Chalmond, Lars Rogge, Lluis Quintana-Murci, Matthew L.Albert, Benno Schwikowski for The Milieu Intérieur Consortium

Highlights

- The novel FlowGM flow cytometry workflow targets a large number of samples
- Largely automated analysis with minimal operator guidance
- Quantification of 24 cell types across 115 samples and 4 panels
- Embedding of results in FCS files permits inspection and validation in FlowJo
- Validated performance is as good as, or even better than manual gating

2.1 Introduction

Flow cytometry is a key technology for the characterization of the cellular component of the immune system. Flow cytometers are able to simultaneously quantify different surface

markers of single cells, allowing the identification and quantification of different immune cell subpopulations. In recent years, improvements in measurement speed and experimental automation have enabled comprehensive immunoprofiling of larger cohorts [Orrù et al., 2013].

The gold standard for the analysis of raw flow cytometry data has until now remained "hand gating" (i.e., analysis through computer-assisted procedures for the classification of cells into single cell types using software tools such as FlowJo [Tree Star, 2014]). Each sample is analyzed by successively separating cell types by successive "gating" in a series of one- or two-dimensional projections. However, the manual operation is laborious and subject to biased visual inspection and gate adjustment. These concerns grow with increased numbers of measured phenotypic markers. Moreover, there is a major limitation on that information critical for accurate gating may not be present in the selected two-dimensional projections.

Here, we report a new method for analyzing multi-parametric flow cytometry, the need for which was motivated by the MI study. This project aims at defining the genetic and environmental determinants of variable immunologic phenotypes in a healthy population [Thomas et al., 2015]. Cell phenotyping constitutes one of the major datasets to be integrated into the data warehouse, and as such efforts were made to standardize each step of the sample collection, technical procedures and data analysis. A Companion paper highlights the pre-analytical, semi-automated measures put in place for labeling and data generation [Hasan et al., 2015]. This manuscript details the automated analytic workflow developed for the identification and analysis of 24 cell types across four 8-color cytometry panels.

Our work follows from a large number of computational approaches that have been developed for automated flow cytometry analysis. Recently, the FlowCAP study evaluated a range of approaches [Hu et al., 2013]. In all cases, however, the datasets used by these investigators were on a smaller scale than the ones in our study, in terms of samples studied (FlowCAP: up to 30 samples; here: 115 samples $\times$ 4 panels), and the number of events per experiment (FlowCAP: up to approximately 100,000 events; here: on average 300,000 events per FCS file). Due to these differences, we found that the top-ranked FlowCAP approaches were inadequate to address the needs of our data sets. For example, the ADICyt approach [Adinis, 2014] required more than 6 hours for the analysis of a single sample. The flowMeans software [Aghaeepour et al., 2011] was faster, but required manual assignment of cell types to each cluster in every single sample. The recent X-Cyt approach [Hu et al., 2013] was designed explicitly to efficiently address the problem of large numbers of samples. However, X-Cyt still requires the definition of a "partitioning scheme", a series of mixture models whose sequence and parameters have to be manually configured and calibrated for each cell type of interest in any given analytic panel.

To support the analysis of the MI cohort data set, we developed a novel high-dimensional

data analysis approach, which we refer to as FlowGM, utilizing fast algorithms that enable the standardized analysis of a large number of samples. We describe its application to two representative 8-color panels with up to 11 cell populations classified per panel. Its principal feature is that, after the definition of global parameters in a reference sample (i.e., a one-time manual assignment of cell type labels to clusters), it is possible to automatically position and identify cell populations across the entire dataset. This approach will enable analysis of our large healthy donor cohort.

2.2 Materials and Methods

2.2.1 Dataset

Four 8-color cytometry panels that are targeting major leukocyte populations across 115 individuals from different age groups and genders were designed to characterize the major immune cell populations (T cells, B cells, NK cells and monocytes), as well as subpopulations of T cells, dendritic cells (DC) and polymorphonuclear leukocytes (PMN). The standardized procedure for collection and treatment of the whole blood sample is described in [Hasan et al., 2015]. For each of the four panels, technical replicates performed by five parallel blood samples obtained from three donors ("repeatability" studies from [Hasan et al., 2015]) were generated to examine the robustness of the experimental and computational protocols.

2.2.2 FlowGM cluster model

The input to FlowGM is a set of m sets of N quantitative measurements ("events"), formally, m N-dimensional vectors. Clustering is based on a multivariate Gaussian Mixture Model (GMM) [McLachlan and Peel, 2004], which has the form

$$p(\underline{\boldsymbol{x}}|\theta) = \sum_{k=1}^{K} \alpha_k \mathcal{N}(\underline{\boldsymbol{x}}|\boldsymbol{\mu}_k, \Sigma_k)$$

A GMM thus corresponds to a set of K clusters, each described by a cluster weight α_k and an n-dimensional Gaussian (normal) probability distribution, whose parameters θ are its centroid $\boldsymbol{\mu}_k$, and its extent and orientation, Σ_k in m dimensions. The weight of each cluster corresponds to the proportion of all cells assigned to it. Gaussian mixture models have been used for flow cytometry, but a particularity of FlowGM is that several such clusters can be used to model cells of one type that may not adequately be modeled by a single normal distribution.

2.2.3 Clustering cells using Expectation Maximization (EM)

Starting from an initial configuration, the degree of fit between the clusters and the data is quantified by a likelihood function. Each stage of an iterative optimization process (Expectation maximization, EM) improves the likelihood in two steps [Dempster et al., 1977] [Bilmes et al., 1998]. In an E (Expectation) step, each event is assigned to (potentially, multiple) clusters whose location is close to the event. In an M (Maximization) step, the cluster parameters are optimized to fit the events assigned to it.

2.2.4 FlowGM workflow

Step 1: Define pre-processing parameters (manual)

To initialize automatic processing of Phase I, FlowGM requires the input of a few parameters, such as the choice of a reference sample, and the selection of potential pre-filtering and post-filtering parameters.

Step 2: Perform pre-filtering (automatic)

Automated pre-filtering helps eliminate noise (such as doublets) and/or "uninteresting" cells (i.e., Dump populations), which is of importance when the cell types of interest are rare. Two filters have been pre-configured: A doublet filter and a filter that eliminates cells that are negative relative to user-definable markers (based on two-component one- or two-dimensional GMMs). The filter eliminates the 95th percentile of the cluster corresponding to the "uninteresting" cells.

Step 3: Determine the number of clusters (automatic)

The number of clusters K used to model the reference sample is determined by minimizing the Bayesian Information Criterion (BIC) [Schwarz et al., 1978]. The BIC represents a tradeoff maximizing the degree of fit between the cluster model and the data on one hand (expressed by the likelihood $p(\underline{\boldsymbol{x}}|\theta)$), and, minimizing, on the other hand, model complexity (based on the number of clusters k):

$$BIC_K = -2\ \ln(p(\underline{\boldsymbol{x}}|\theta)) + P\ \ln(N)$$

where $P = K*m + K*m*(m+1)/2 + (K-1)$ is the number of parameters. Specifically, we choose K that minimizes the average of BIC_K under 20 EM runs starting with random initial configurations.

Step 4: Establish the reference clustering (automatic)

Once the number K of clusters has been determined, FlowGM determines 100 random initial configurations of K clusters as starting points, and performs clustering using Expectation Maximization, as described in Section 2.2.3. The resulting clustering with the highest likelihood is selected as the reference clustering in the second FlowGM phase.

Step 5: Label reference clusters with cell types (manual)

An operator defines the cell types of interest, and assigns one or more corresponding clusters to each such cell type (labeling). Thus, each cell type of interest corresponds to a set of clusters (meta-cluster).

Step 6: Perform post-filtering (automatic, optional)

This optional step offers the possibility to eliminate additional "uninteresting" events that remain in the clusters determined in Step 5 (analogous to a "dump" gate for conventional approaches and useful in focusing the clustering analysis). Two filters have been preconfigured: A dead cell filter (based on the Viability channel), and a "dump" filter that eliminates selected cells in specified meta-clusters. In both instances, the cells above or below a defined threshold are removed. This threshold is automatically determined as the 95th/99th percentile of a fitted one-dimensional Gaussian distribution of a reference population along a pre-defined channel. The reference population may either be the meta-cluster itself, or a negative control that has been removed in the pre-filtering (Step 2).

Step 7: Cohort samples: Pre-filter and cluster by adjusting labeled reference clustering (automated)

After the reference sample has been processed in Steps 1-5, FlowGM processes all other samples in a fully automated manner. Pre-filtering proceeds as described for the reference donor (Step 2). FlowGM then determines the clustering using EM, as described in Section 2.2.3, starting with the labeled reference clustering (from Step 5) as the initial configuration. Finally, post-filtering is applied (if selected), as in Step 6.

2.2.5 Visualization of the resulting clusters in FlowJo

One innovation incorporated into FlowGM included the embedding of labels for each cluster and meta-cluster as additional attributes (numerical identifiers) for each cell in

the FCS data file. This allows inspection of the different clusters in FlowJo [Tree Star, 2014] or other software that can analyze FCS data files.

2.2.6 Software implementation

FlowGM was implemented using Matlab and Statistics Toolbox Release 2012b [The MathWorks, 2014] and R (version 3.0.1) [Ihaka and Gentleman, 1996] flowCore package [Ellis et al., 2009]. The visualization graphs were prepared with FlowJo software version 9.7.5.

2.3 Results

2.3.1 FlowGM workflow

Motivated by the need for high-quality analysis of a large flow cytometry data set, we developed the novel, and largely automated FlowGM data analysis approach. Its computational high-dimensional clustering approach avoids the limitations inherent to analysis based on two-dimensional projections (Figure 2.1).

Experimental data is modeled as a mixture of normal distributions (See Material &

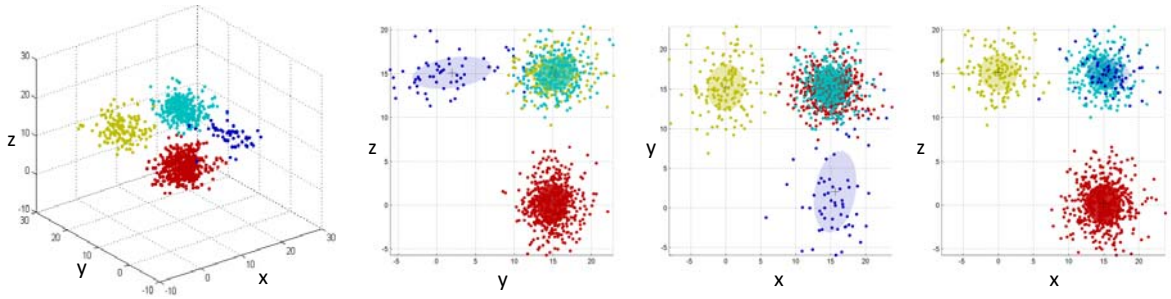

Figure 2.1: Four simulated clusters in 3D space that cannot be separated in any 2D projection.

Methods, Section 2.2.3) and employs Expectation Maximization (EM) to iteratively adapt model parameters (Figure 2.2 and see Section 2.2.3).

The overall operation of the FlowGM workflow can be understood on the basis of its similarities and differences relative to the current 'gold standard' manual FlowJo workflow (Figure 2.3). For both approaches, two phases can be distinguished. In the first phase, method parameters are calibrated on selected reference samples. In a second phase, all other samples are processed on the basis of the calibrated parameters. To be suitable for large cohort studies, FlowGM was designed to minimize the manual per-sample effort in the second phase.

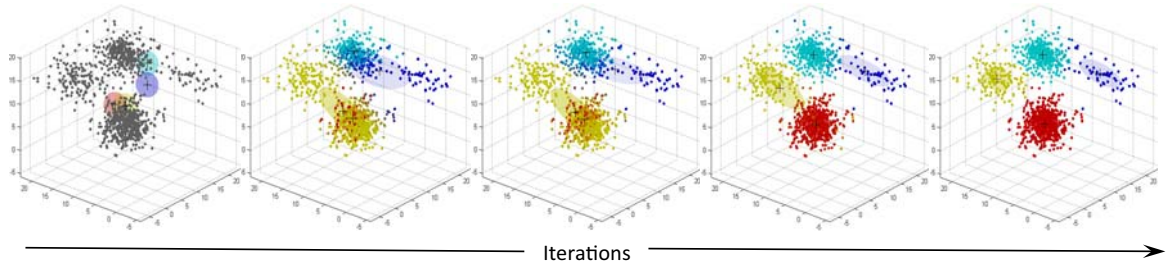

Figure 2.2: Illustration of the expectation-maximization (EM) clustering algorithm using Gaussian mixture model (GMM) clusters, when applied to this data. Points are colored according to their posterior likelihood, the ellipsoid reflects cluster shape, '+' indicates the cluster centroid, transparency of each ellipsoid reflects cluster weight. Five phases are shown: initial random parameter values, updated parameters after the first M-step, after two iterations, after ten iterations, and final solution.

2.3.2 Identification of the major cell lineages by FlowGM

We first applied FlowGM to the lineage panel dataset [Hasan et al., 2015]. Cells were stained with the markers CD45, CD3, CD4, CD8β, CD14, CD16, CD19, and CD56. Following the approach of the manual analysis by Hasan et al., we used forward and side scatter (FSC/SSC) solely to exclude doublets; the remainder of our data analysis is performed on the dimensions of the indicated eight markers. The number of events in the data files ranged from 106000 to 787000. After filtering out doublets, FlowGM estimated the optimal number of clusters K to be 36, using the BIC (see Materials & Methods, Section 2.2.4) on the reference donor (Figure 2.4).

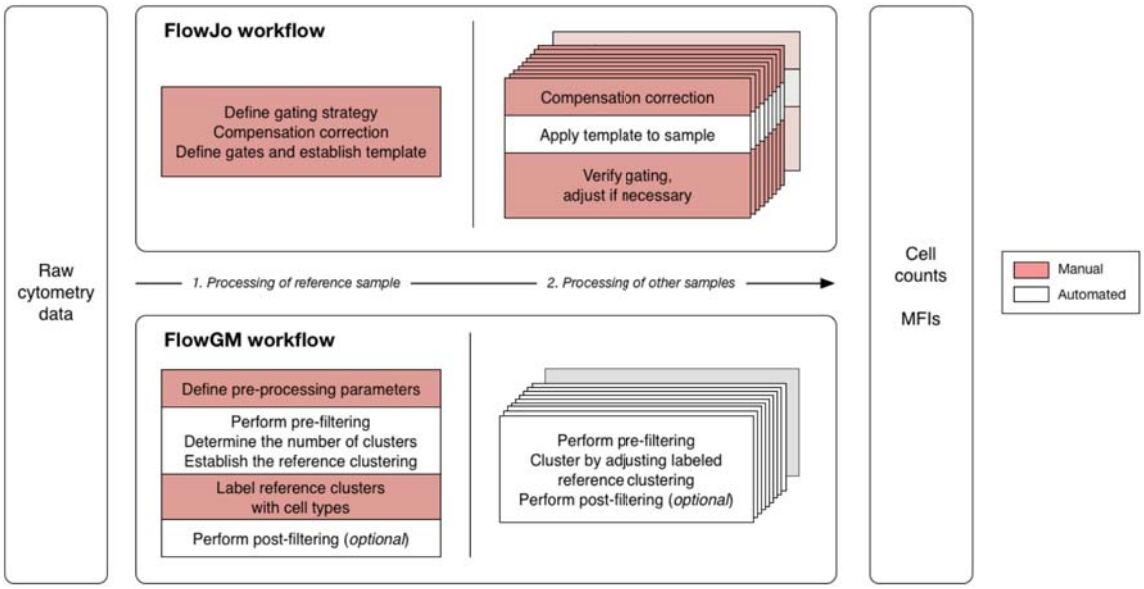

Figure 2.3: FlowJo and FlowGM workflows

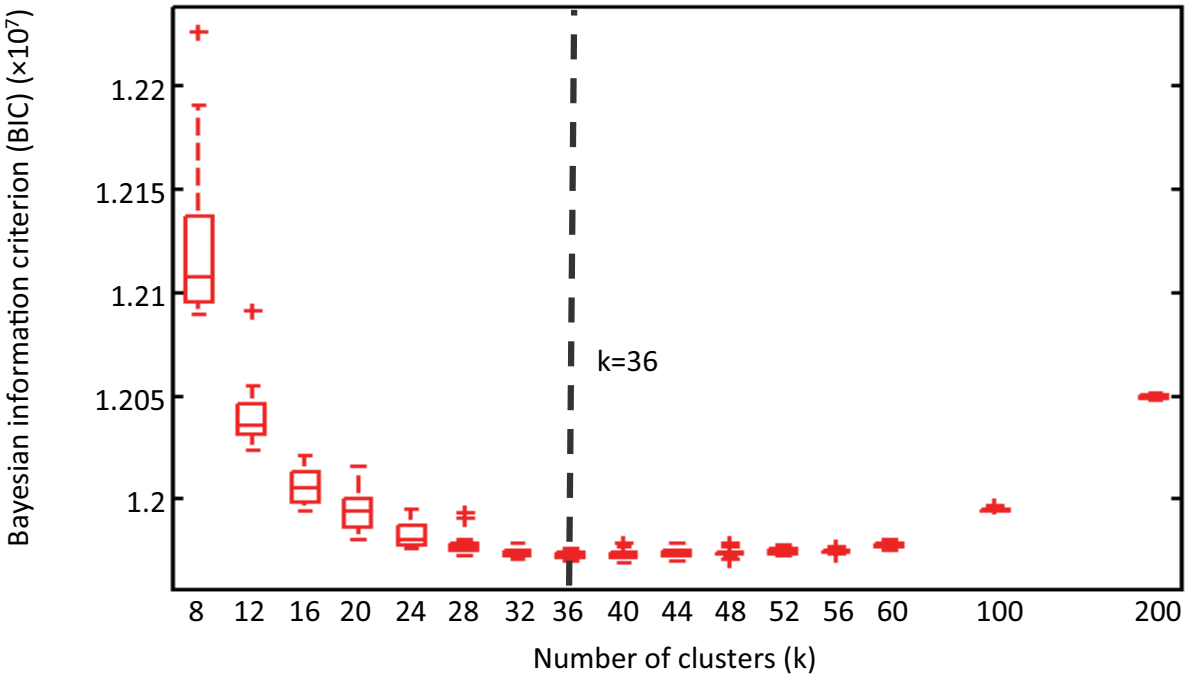

Figure 2.4: The number of clusters K is determined with the minimum average Bayesian Information Criterion (BIC) when evaluated on 20 random initial solutions for each choice of K. For the lineage panel, $K = 36$ is optimal.

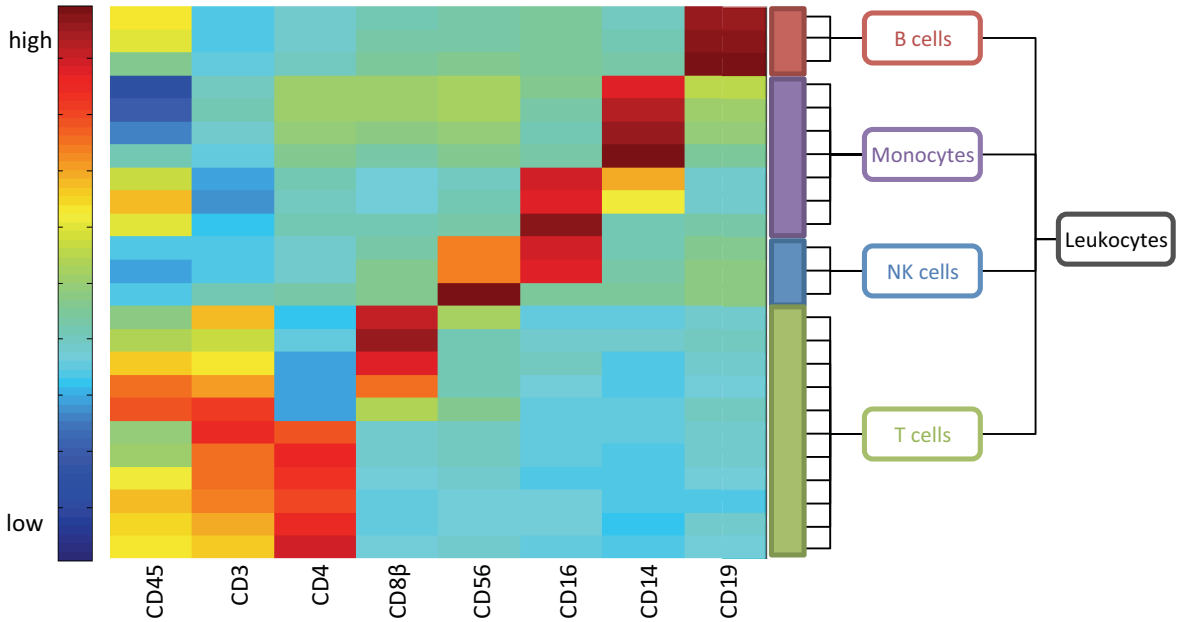

Figure 2.5: User-based aggregation of FlowGM clusters into meta-clusters for immune cell type characterization with cluster centroid heat map (normalized coordinates). B cells are identified as $CD19^{+}$, T cells are identified as $CD3^{+}$ with two subsets: $CD4^{+}$ (T-1) and $CD8\beta^{+}$ (T-2), NK cells are identified as $CD56^{+}$ with two subsets: $CD16^{hi}$ (NK-1) and $CD56^{hi}$ (NK-2), monocytes are identified as three subsets: $CD14^{hi}$ (Mono-1), $CD14^{hi}CD16^{hi}$ (Mono-2) and $CD14^{lo}CD16^{hi}$ (Mono-3). The manually assigned cell types are indicated on the right.

Once K was determined, FlowGM performed EM clustering 100 times, starting with different random initial configurations of k clusters. The clustering solution with the highest likelihood $p(\underline{\boldsymbol{x}}|\theta)$ constitutes the reference clustering, whose clusters were then manually labeled with the different cell types of interest (i.e., leukocyte subpopulations). The corresponding cluster centroids are represented as a heat map, with the assigned cell types indicated (Figure 2.5).

Note that only 24 of the 36 clusters corresponded to cell types of interest, and the color coding is chosen independently for each marker to resolve the entire spectrum of expression across these cell types (using the Matlab HeatMap function). For example, as $CD45^-$ cell populations were not of interest in this study, all selected cells were $CD45^+$ and as indicated by the normalization, the lowest and highest levels of CD45 expression were observed in monocytes and T cells, respectively (Figure 2.6).

To facilitate the understanding of our findings and permit user cross-validation, FlowGM

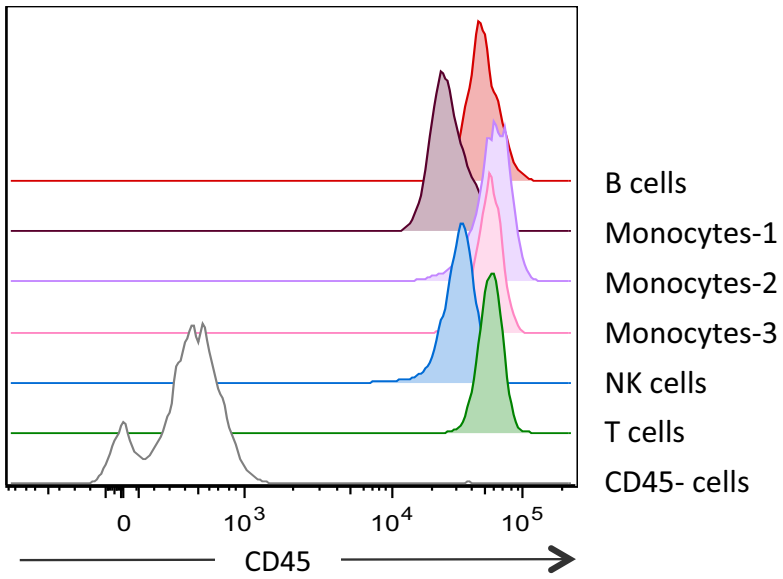

Figure 2.6: Distribution of CD45 intensity for different cell types of interest in the reference donor.

allows the embedding of cluster IDs and meta-cluster IDs as additional channels (designated "C-ID" and "MC-ID", respectively) into the FCS input file, permitting importation of all data into FlowJo (or other FCS-compatible software). FlowJo visualizations of the labeled FlowGM lineage clusters confirmed our GMM-based assignments (Figure 2.7).

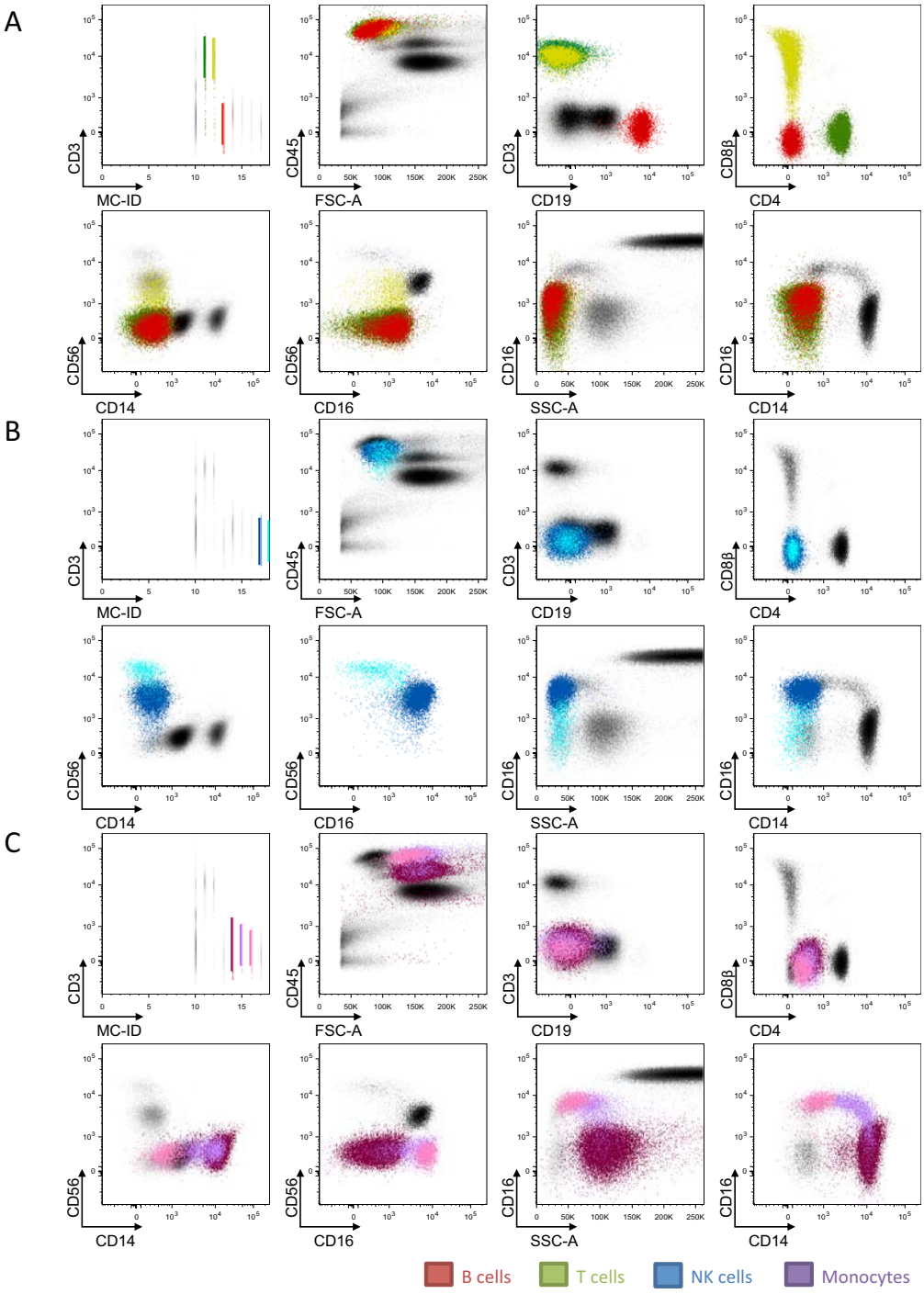

Figure 2.7: Visualization of labeled meta-clusters in FlowJo Cluster IDs are incorporated into the FlowJo input file. Shown are meta-clusters with all principal manual gating steps, starting with SSC-A/Meta-Cluster ID (MC-ID). (A) The identified $CD19^{+}$ B cells (red) and $CD4^{+}$ (green) and $CD8\beta^{+}$ (yellow) subsets of $CD3^{+}$ T cells. (B) $CD56^{hi}$ (light blue) and $CD16^{hi}$ (dark blue) NK cell sub-populations. (C) $CD14^{hi}$ monocytes (Mono-1, mauve), $CD14^{hi}CD16^{hi}$ monocytes (Mono-2, lavender) and $CD14^{lo}CD16^{hi}$ monocytes (Mono-3, light purple).

By gating on MC-ID to select one FlowGM meta-cluster, it is possible to view the clustered cells in 2D projections that correspond to manual gating strategies. FlowJo visualizations of all 36 FlowGM clusters are shown in Figure S1. Backgating is also possible: starting with manual gated data and examining where the captured events cluster in C-ID or MC-ID space (data not depicted).

2.3.3 Pre-filtering supports clustering of rare dendritic cell

We next evaluated the performance of the method on rare subsets of cells (< 1% of the total cell events). In addition to the elimination of doublets (Figure 2.8) early in the analysis, we identified the need for pre-filtering of cells considered by the user as uninteresting - similar to the use of a "Dump" gate - only in the case of FlowGM the procedure is automated and thus removes operator bias. Pre-filtering of the DC panel was based

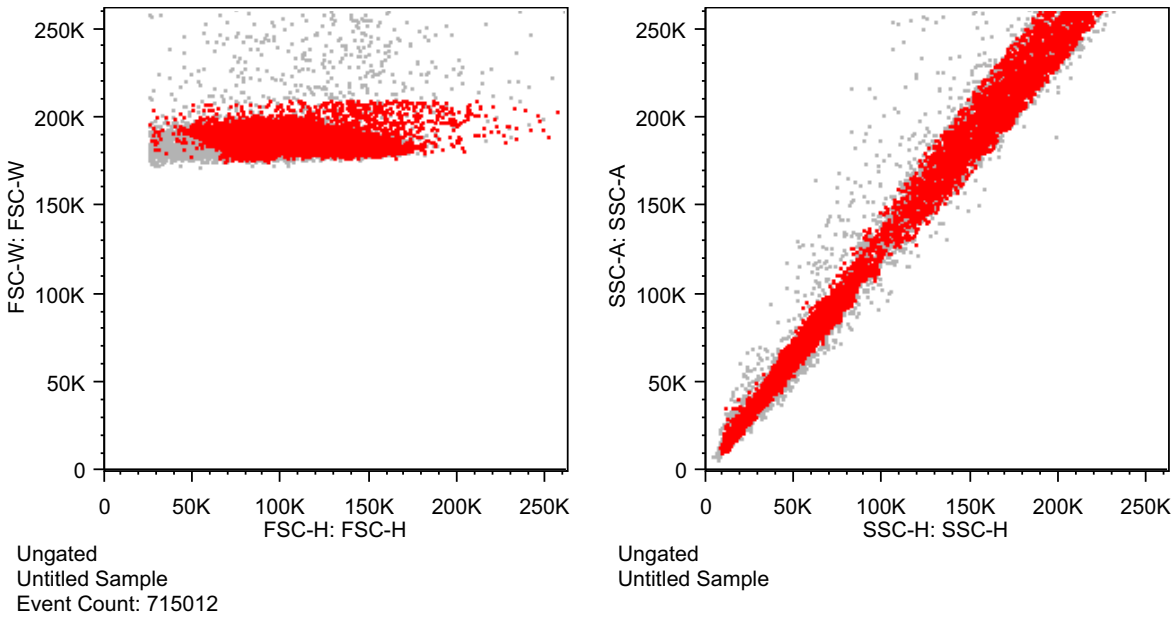

Figure 2.8: Pre-filtering of doublet using FSC/SSC

on a two-component, two-dimensional GMM that utilized data from CD14 and HLA-DR markers. Thresholds were automatically set at 95th percentiles of the CD14/HLA-DR double-negative population (represented by the red line, Figure 2.9). The resultant cells were investigated using the FCS embedding feature of FlowGM, and inspection of representative files revealed accurate retention of desired HLA-DR$^+$ and/or CD14$^+$ cells.

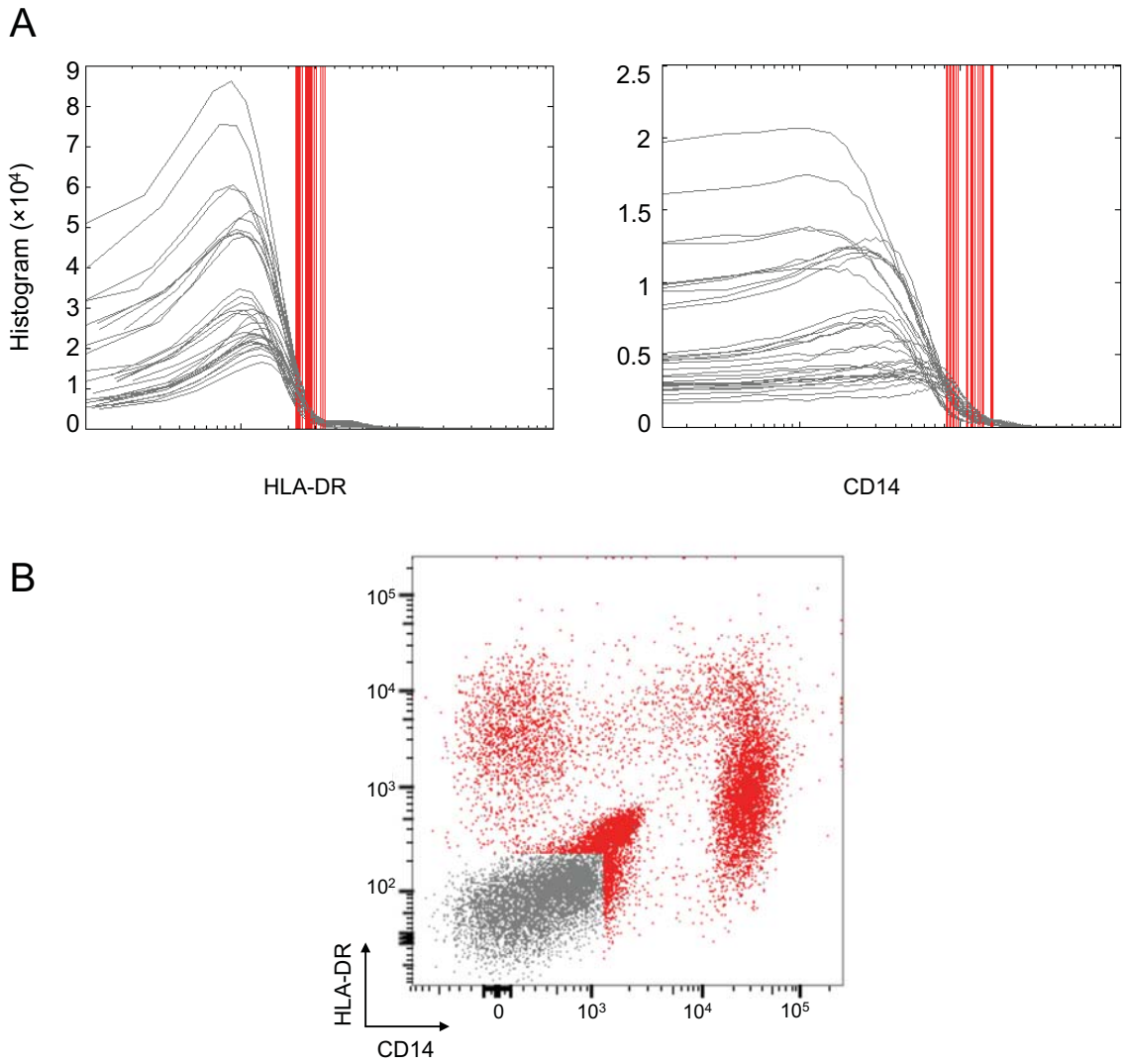

Figure 2.9: Pre-filtering for analysis of rare cell populations (A) Pre-filtering in dendritic cells (DC) by low expression of CD14 and HLA-DR. Red lines indicate the thresholds that were automatically determined using GMM. (B) Validation of pre-filtering using FlowJo visualization.

Next, we estimated K using the BIC and defined a clustering solution using data from a reference donor (Figure 2.10).

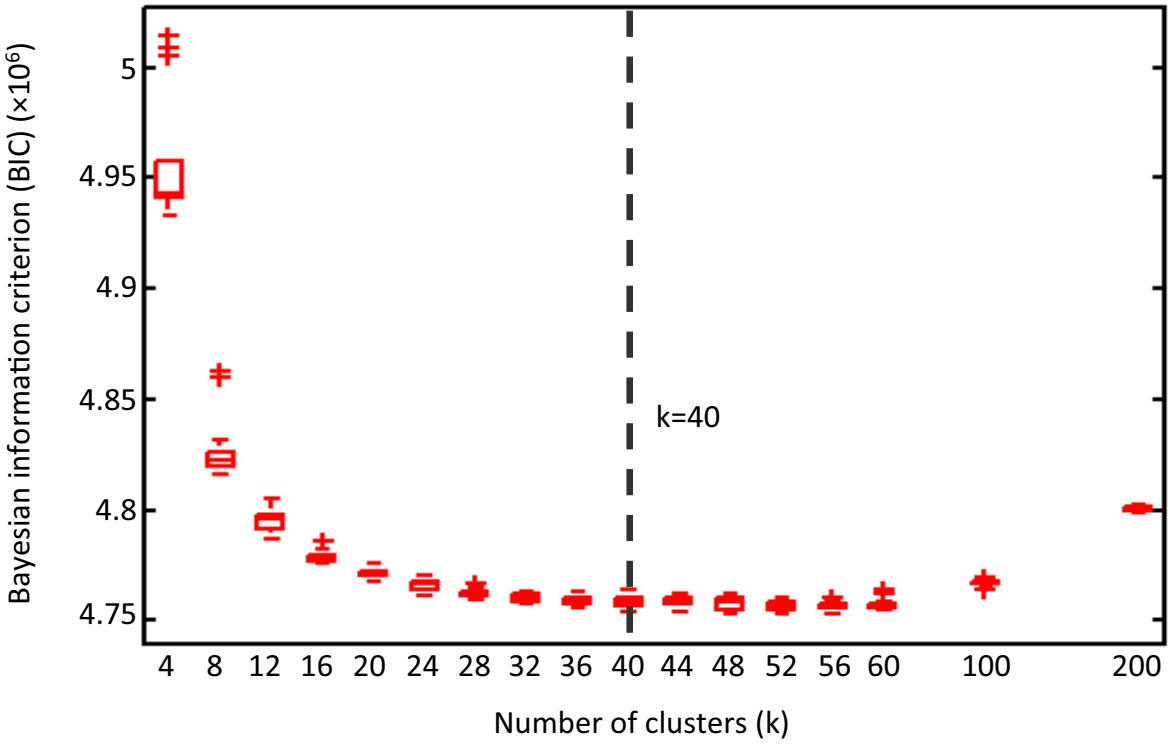

Figure 2.10: The number of clusters K is determined as the minimum average Bayesian information criterion (BIC) when evaluated on 20 random initial solutions for each choice of K. For the DC panel, $K = 40$ is optimal.

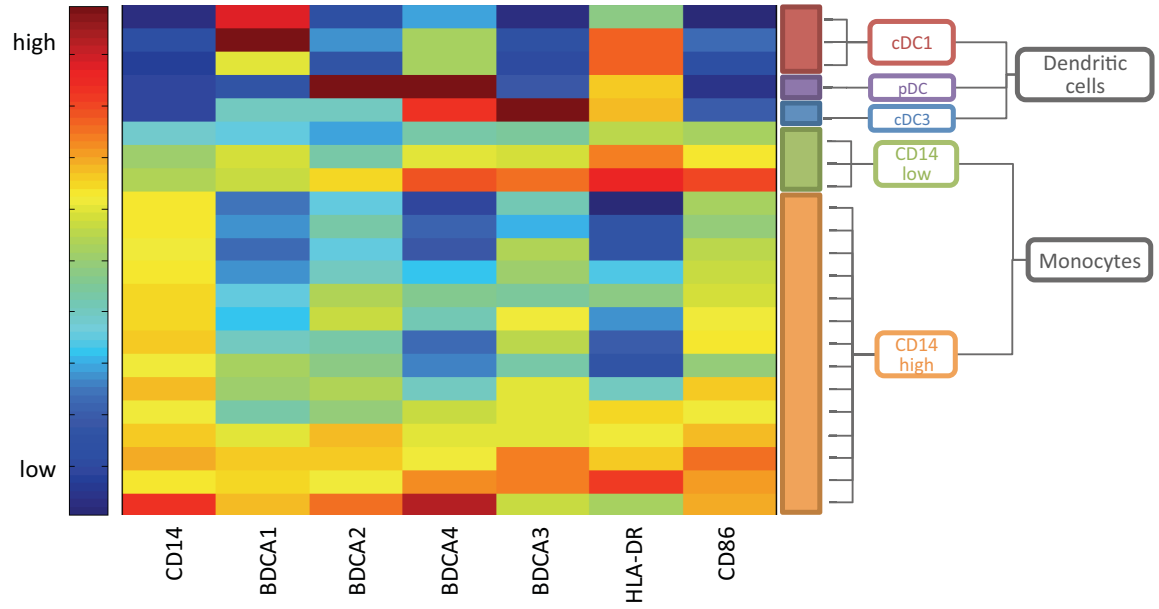

Figure 2.11: User-based aggregation of FlowGM clusters into meta-clusters for immune cell type characterization with cluster centroid heat map (normalized coordinates). Plasmacytoid dendritic cells (pDCs), BDCA-1$^+$ and BDCA-3$^+$ conventional dendritic cells (herein referred to as cDC1 and cDC3) were identified as BDCA4$^+$BDCA2$^+$ (CD304$^+$CD303$^+$), BDCA1$^+$ (CD1c$^+$) and BDCA3$^+$ (CD141$^+$), respectively. Monocytes could be identified as two subsets: CD14^{hi} and CD14^{lo}. The manually assigned cell types are indicated on the right.

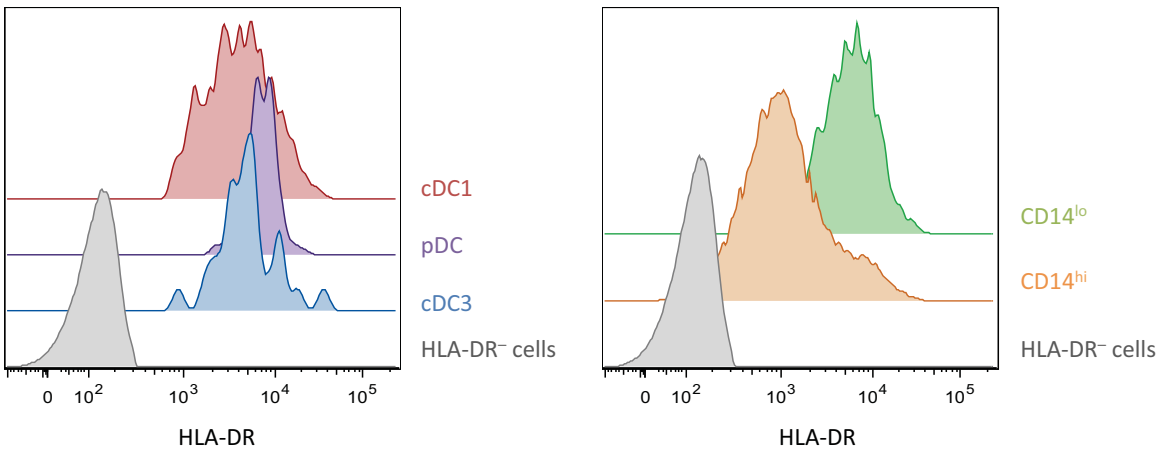

Figure 2.12: Distribution of HLA-DR intensity for different DC cells and subsets of monocytes in the reference donor.

Of the 40 clusters defined as the optimal fit, 22 were of interest and manual labeling of the meta-clustered data captured five myeloid cell subsets: cDC1, identified by their high BDCA2 MFI and low expression of CD14; pDCs, identified by the highest BDCA2 and BDCA4 MFIs; cDC3, identified by their expression of BDCA3; $CD14^{lo}$ monocytes, identified by the intermediate expression of CD14; and $CD14^{hi}$ monocytes, by the high CD14 MFI (Figure 2.11). Again, we highlight that the data represented in the heat map has been normalized, and in instances where all cell populations are positive for a given marker (i.e., HLA-DR), the normalization will scale values to span the range of marker expression. To illustrate the distributions of HLA-DR intensity, histogram plots for DCs and monocytes are shown (Figure 2.12).

Next, an initial post-filter removed dead cells from each meta-cluster, based on the Dump

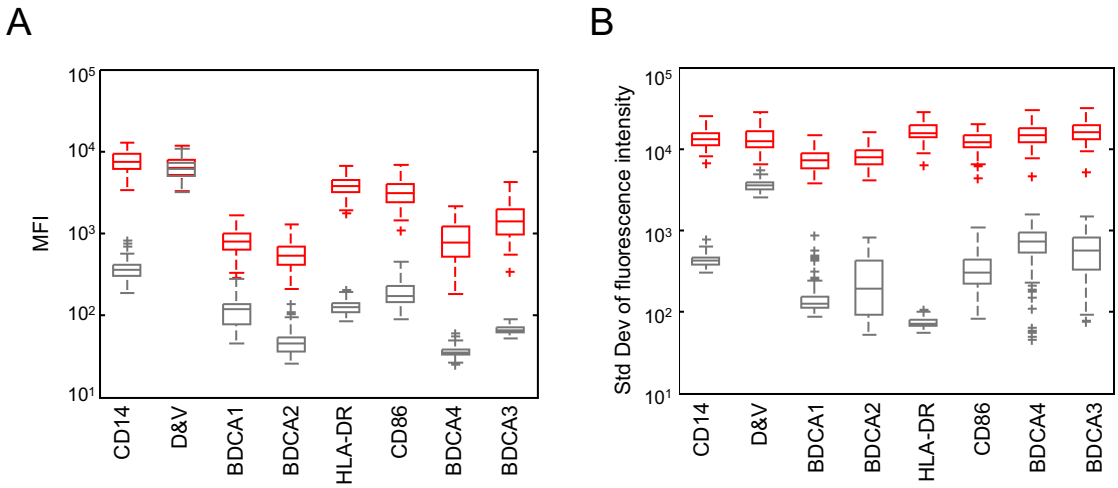

Figure 2.13: (A) MFI of filtered (grey) and remaining (red) cells. Pre-filtered cells display a lower MFI in all channels except Dump. (B) Standard deviation of fluorescence intensity for the same cell population. Filtered cells display less variation.

channel. A second post-filter removed cells from cDC1 and cDC3 populations based on expression of BDCA1 and BDCA3 respectively, of the CD14/HLA-DR double-negative population that was previously filtered out.

As a final validation step, we compared the level of marker expression between retained cells and events that were removed by the filtering process. Across all dimensions of the data set, we confirmed the efficacy of the pre-filtering approach (Figure 2.13).

Additional visual confirmation can be found in the FlowJo-projected data, where meta-clustered data is overlaid on the total cell events in a representative file (Figure 2.14).

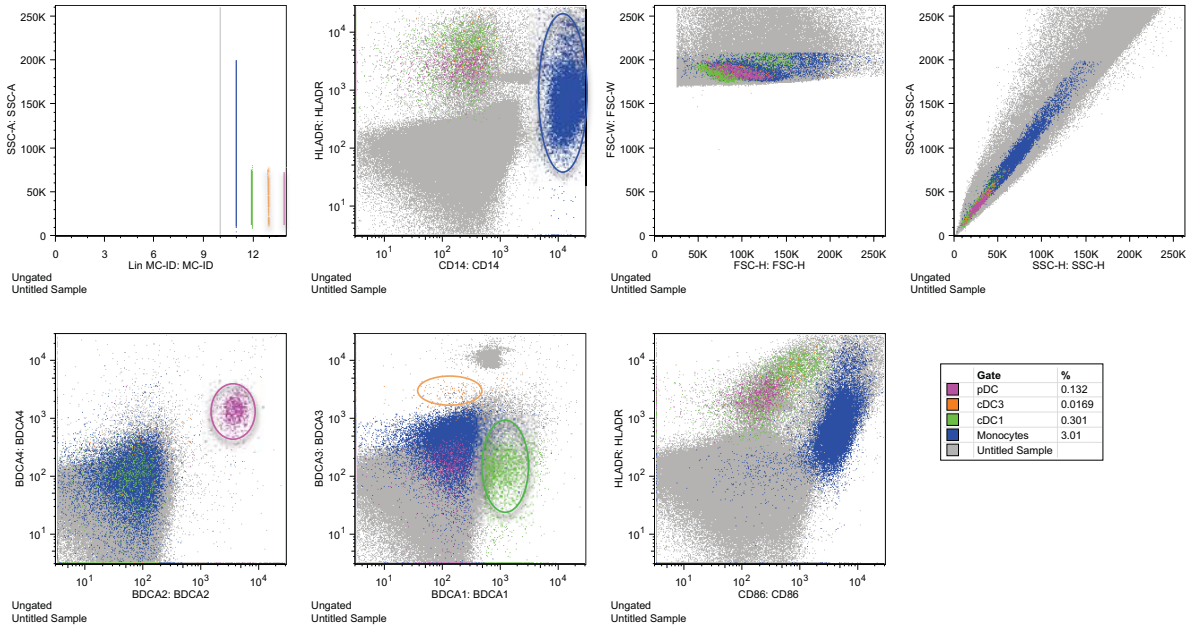

Figure 2.14: Visualization of assigned meta-clusters in FlowJo in DC panel. Cluster IDs are incorporated into the FlowJo input file. Shown are meta-clusters identified as pDC, cDC1 and cDC3.

2.3.4 FlowGM is robust to selection of reference donor and may be applied to uncompensated data

One potential concern with the FlowGM approach is the sensitivity of the clustering result to the choice of the reference sample in Step 1 (cf. Section 2.2.4). This is an important issue, as the resulting reference clustering will be used as the basis to cluster the data from all other samples. While practitioners may have a good intuition about which one of the input samples is "representative", the degree of sensitivity to this choice could, in principle, be large.

We therefore investigated whether a more representative reference clustering based on a larger group of samples would be needed. To this end, we constructed 11 different clusterings: the originally chosen reference clustering (which we denote here by 1*), and ten alternative reference clusterings $(1, \cdots, 10)$ of increasing complexity, which were obtained by selecting a series of 10 samples from randomly chosen donors, and then merging the samples $1, \cdots, d$ for each $d = 1, \cdots, 10$. Merging different samples without alignment can be expected to create reference clusterings that contain technical shifts, and thus could translate into significant variation in the clustering result.

For each possible pair of these 11 reference clusterings, we then determined the similarity of the two outcomes after clustering, using the F-measure [van Rijsbergen, 1979] (Figure

2.15 A). Notably, the F-measure values were close to 1, independently, for all pairs of reference clusterings, indicating that the different reference clusterings did not translate into significantly different clustering outcomes. The locations of the resulting cell types for the different reference clusterings were further represented in parallel coordinate plots (Figure 2.15 B). Except for the Mono-2 and Mono-3 populations, all coordinates match extremely well among the different reference clusterings across all dimensions. Together, these observations suggest that the choice of the initial reference clustering may not have a large impact on the resulting outcome.
We also investigated the impact of compensation. Routinely, automatic hardware compensation [Hasan et al., 2015] is employed. Here, we compare the results of our approach on the same input data in an uncompensated state; machine-compensated; or machine-compensated and FlowJo-corrected. The computational analysis on these three datasets were initialized with the re-estimated parameters from the reference clustering on machine-compensated data. The counts for three repeatability samples obtained from different dataset are shown (Figure 2.16), and indicate that FlowGM is insensitive to instrument compensation, and therefore resistant to potential compensation error in the context of large datasets.

2.3.5 Benchmarking of FlowGM demonstrates its reliability and utility

To directly compare FlowGM clusters to manually gated data sets, we first calculated, for each hand-gated cluster in the reference donor, the percentage of its events present in every other FlowGM cluster (Figure 2.17). The values indicated that, overall, the two approaches group events similarly. The one exception were monocytes, where FlowGM supported easy segregation of the $CD14^{hi}CD16^{hi}$ sub-population of monocytes (Mono-2) from $CD14^{lo}CD16^{hi}$ sub-monocytes (Mono-3), despite the lack of additional monocyte-specific markers (e.g., MCSF-1, CX3CR1, CCR2 [Cros et al., 2010]). We also studied the variability of manual and FlowGM-derived cell counts across the repeatability samples studied in Hasan et al. [Hasan et al., 2015] (Figure 2.18). We find that FlowGM results showed good agreement with the results from manual analysis. The slight bias for higher numbers from FlowGM may stem from the need for high-dimensional information to confidently assign certain events to cell types (as in the schematic example shown, Figure 2.1). Coefficients of variation (CVs), which represent variation of data analysis and experimental variation, were at similar levels, further indicating the high accuracy of FlowGM analysis. Absolute counts and CVs for the repeatability data from all four panels are provided (Table 2.1). The estimation of the number of clusters and the resulting

cluster positions, and assignments to cell types for the T cell and PMN panels are shown in Figure S5 and Figure S6 respectively. For the observed cell types, absolute counts were highly reproducible, with most CVs <15%. Compared to results of Hasan et al. [Hasan et al., 2015], the level of reproducibility of FlowGM was similar to the manual gating results across all four panels.

Finally, we used FlowGM-generated absolute cell counts of the lineage panel across 115 donors from the MI cohort [Thomas et al., 2015], comparing results to those obtained by manual gating. Again, results were highly concordant (Figure 2.19). The running time of the computational analysis for a single panel depends on the number n of measured events in each sample and the number k of clusters. For the panels analyzed here, the computation required 0.5 hours (DC panel) and $\sim$ 4 hours (lineage panels) on a standard laptop PC.

Table 2.1: Repeatability

	Donor†:	#1	#2	#3
Lineage	$CD4^+$ T cells	16870 (4.4)⋆	77306 (9.9)	28838 (4.4)
	$CD8\beta^+$ T cells	6986 (3.8)	19408 (10.6)	21416 (4.0)
	$CD19^+$ B cells	5983 (5.3)	23679 (9.8)	3325 (4.1)
	Monocytes	27615 (3.0)	42233 (11.0)	26894 (3.5)
	$CD14^{hi}CD16^{lo}$ mono	22269 (3.2)	34969 (10.9)	22872 (3.2)
	$CD14^{hi}CD16^{hi}$ mono	3196 (4.2)	3759 (11.7)	1436 (8.9)
	$CD14^{lo}CD16^{hi}$ mono	2058 (3.3)	3505 (10.9)	2907 (5.3)
	NK cells	9803 (5.1)	15989 (12.9)	12534 (4.0)
	$CD16^{hi}$ NK	8633 (4.9)	15424 (13.0)	11632 (3.7)
	$CD56^{hi}$ NK	1171 (7.4)	565 (11.2)	902 (8.9)
T cell	$CD4^+$ T cells	13172 (4.5)	64809 (16.4)	23450 (0.7)
	$CD4^+$ $T_{naïve}$	3043 (4.8)	23398 (13.8)	8961 (8.1)
	$CD4^+$ T_{CM}	8973 (4.4)	39350 (18.2)	13044 (3.6)
	$CD4^+$ T_{EM}	1044 (6.7)	3329 (18.3)	1250 (11.4)
	$CD8\beta^+$ T cells	5245 (5.7)	14847 (16.8)	15283 (3)
	$CD8\beta^+$ $T_{naïve}$	553 (8.2)	5692 (16.8)	5903 (2.3)
	$CD8\beta^+$ T_{CM}	2297 (6.2)	5737 (13.6)	5996 (7.7)
	$CD8\beta+$ T_{EM}	548 (10.2)	1181 (15)	1092 (21.2)
	$CD8\beta^+$ T_{EMRA}	717 (5.1)	1206 (46.8)	954 (16)
	$CD8\beta^+$ 27^{int}	1036 (8.7)	1096 (23.3)	1516 (11.7)
	$CD4{+}CD8\alpha^+$ T cells	153 (11.4)	770 (19.3)	539 (28)
DC	$CD14^+$ monocytes	25232 (12.2)	29764 (4.4)	21287 (8.4)
	pDC	304 (18.5)	409 (4.1)	438 (5.0)
	cDC1	2159 (12.1)	5188 (3.9)	1677 (10.4)
	cDC3	42 (30)	87 (16)	44 (8.1)
PMN	Neutrophils	96062 (14.3)	188428 (13.0)	119529 (12.0)
	Basophils	1751 (11.4)	5878 (7.2)	2323 (11.6)
	Eosinophils	10483 (13.2)	18539 (10.6)	22329 (6.2)

† Fresh blood samples from three healthy donors were divided in five aliquots each and immediately stained using four antibody panels.

⋆ Median absolute cell counts per 1mL of blood in five independent analysis is represented for each cell population, as well as the corresponding coefficient of variation(CV).

2.4 Discussion

The FlowGM flow cytometry approach was developed to address the need for fast, robust and high-quality analysis for the MI study. Our comprehensive validation study has shown that FlowGM has produced user-validated results whose quality is on par with, and in some cases, exceeds, the hand-gating approach. This is an exciting finding, as its simple computational approach does not require the expert knowledge and experience that is available to human operators. One important difference lies in the systematically higher number of events assigned to cell types by FlowGM, which suggests that the full dimensionality of the data, instead of two-dimensional views, allows to assign cells that are unassigned in manual two-dimensional analysis due to the lacking dimensionality and user-bias. Another facet of this fundamental difference may be the observed ability of FlowGM to segregate subpopulations of monocytes without the need for an additional specific marker. Notably, separation of $CD14^{lo}CD16^{hi}$ monocytes from NK cells and other cell populations was achieved by integrating information from all eight dimensions.
When comparing the design of FlowGM workflow to other computational clustering approaches, a characteristic difference lies in the choice to computationally model single cell types as mixtures of Gaussians, as opposed to single Gaussians, or other distributions, coupled with the incorporation of knowledge and experience of a human operator to define which clusters belong to the same cell type (referred to herein as meta-clusters). This design may constitute a 'sweet spot' in cytometry workflow design: A fast and efficient overall workflow, combined with a mathematical model that is flexible enough to model experimental data well, the solution of a hard core problem (the assignment of cell types to clusters) using operator intervention, and the limitation of this intervention to a single reference sample, as the transposition of this knowledge to all other samples can be automated with high accuracy.
The minimization of operator intervention means not only significant savings in terms of manual effort, but also the elimination of variability between different samples introduced by subjective decisions, and a considerable improvement in transparency and reproducibility of the path from the samples to the absolute and relative cell counts. Furthermore, the facility with which results are accessible for human inspection using conventional tools, and the relative simplicity of the FlowGM approach itself imply a high level of accessibility to non-specialists that - we believe - will continue to play an important role in the evolution of the approach.
We believe that the FlowGM workflow is applicable to most other flow cytometry datasets, and anticipate that the need for fast, robust, and high-quality analysis of large cytometry datasets will only increase. Adaptations of the method may be required for heterogeneous samples, in which no single reference sample may be representative for all others (e.g.,

disease populations). We believe that there are relatively straightforward approaches to extend FlowGM to automatically detect cases of inadequate fit, for example, through the introduction of additional reference donors (with recursive iteration of the manual Step 5). The increased availability of experimental datasets that have been acquired under standardized conditions may facilitate comparison and integration, which may lead to the necessary insights and technical developments to fully automate flow cytometry data analysis.

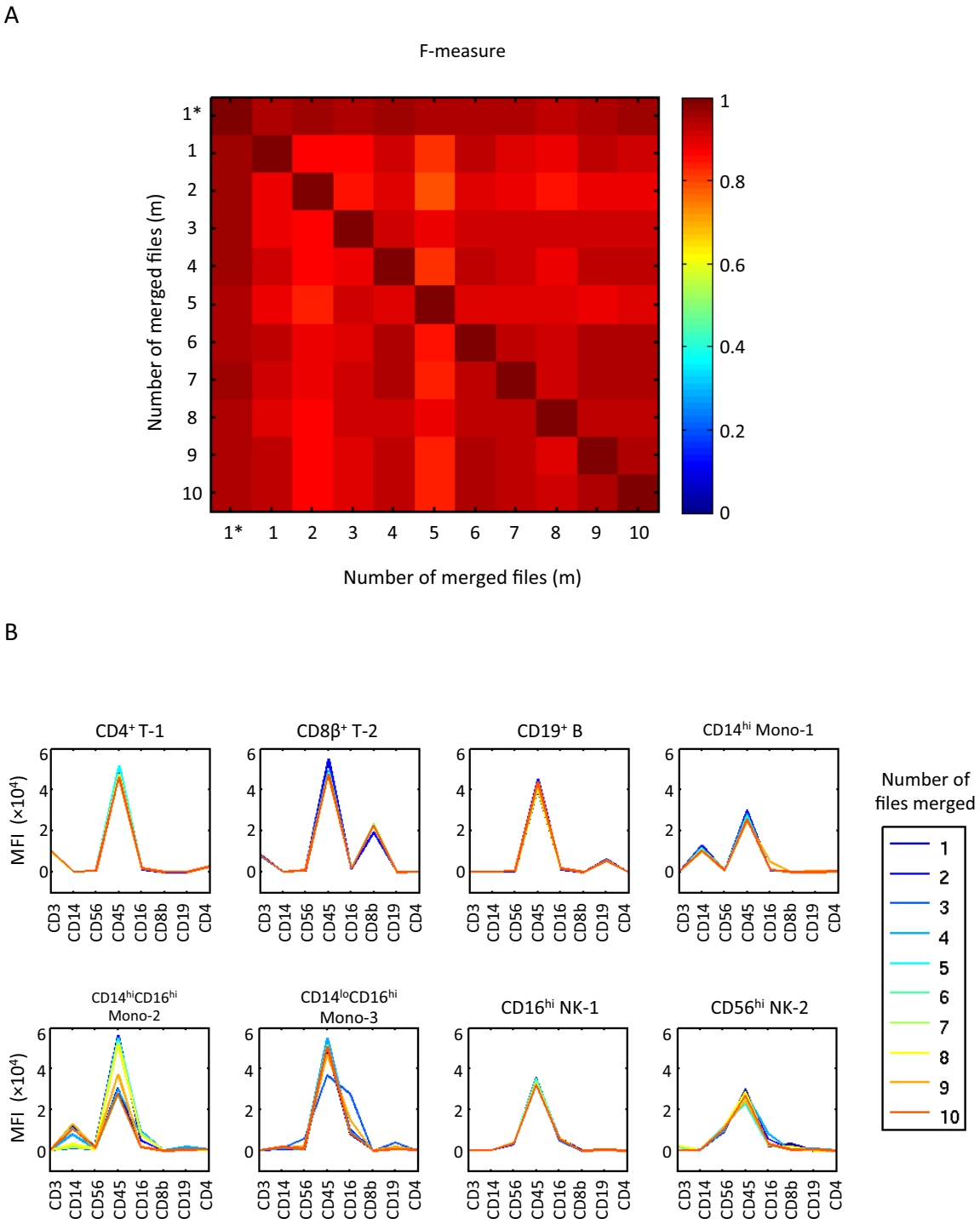

Figure 2.15: Different reference clusterings are generated by merging data from one to ten randomly selected donors; solutions are then applied to 115 cohort donors. (A) Pairwise average similarity (F-measure) of solutions over 115 cohort donors after using different reference clusterings. (B) Mean fluorescence intensity (MFI) of each identified cell population from different reference clusterings.

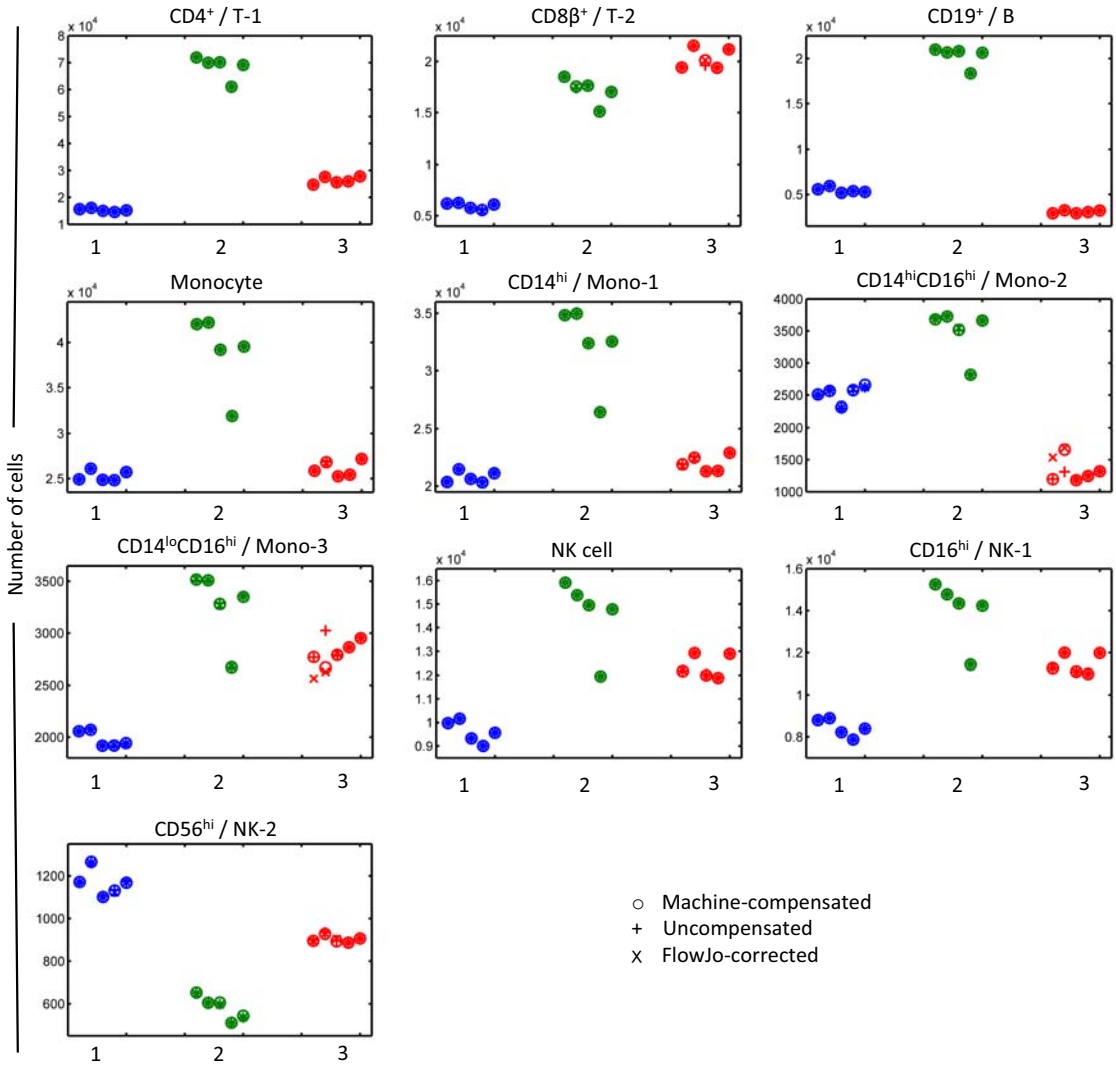

Figure 2.16: FlowGM counts of each cell type for three donors with five replicates, obtained from three different datasets: uncompensated, machine-compensated and FlowJo-corrected. The results suggests that FlowGM is insensitive to instrument compensation errors.

		FlowJo gates						
		CD4+	CD8+	CD19+	CD14^{hi} mono	CD16^{hi} mono	CD16^{hi} NK	CD56^{hi} NK
FlowGM meta-clusters	T - 1	99.8	0	0	0	0	0	0
	T - 2	0	98.7	0	0	0	0	0
	B	0	0	98.8	0	0	0	0
	Mono - 1	0	0	0	97.3	0	0	0
	Mono - 2	0	0	0	2.64	42.7	0	0
	Mono - 3	0	0	0	0	50.2	0	0
	NK - 1	0	0	0	0	3.8	96.2	1.5
	NK - 2	0	0	0	0	0	0.7	98.2
	Unmapped clusters	0.2	1.3	1.2	0.06	3.3	3.1	0.3

Figure 2.17: Comparison of manually counting and FlowGM analysis - Performance on reference donor: percentage of events in FlowJo cluster present in FlowGM clusters.

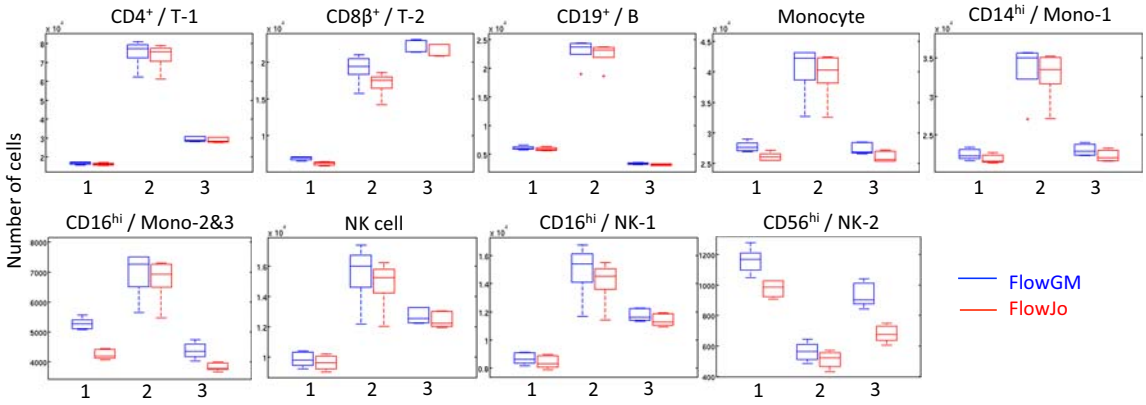

Figure 2.18: Comparison of manually counting and FlowGM analysis - Performance on repeatability data: counts of each cell type for three donors with five replicates. The FlowGM results show a comparable CV with manually counting.

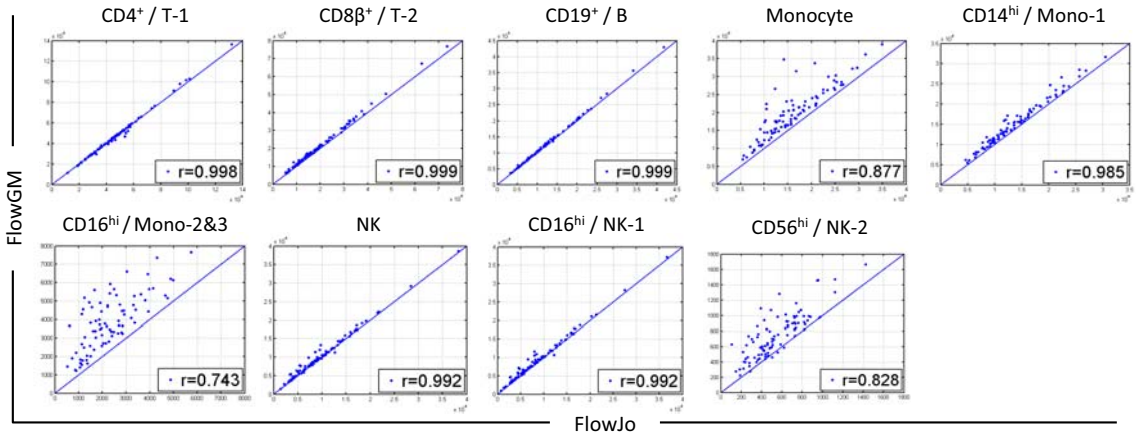

Figure 2.19: Comparison of manually counting and FlowGM analysis for lineage panel on $D = 600$ cohort donors. The X-axis is the counts from FlowJo, and the Y-axis is the counts from FlowGM. The counts obtained by these two methods highly agree on $D = 600$ cohort donors, with average correlation equal to 0.944 across all 9 identified population in lineage panel.

Chapter 3

Evaluation - A 600-person healthy donor study

This chapter was made possible by Alejandra Urrutia, Vincent Rouilly, Milena Hasan, Molly Ingersoll, Philip De Jager and Benno Schwikowski.

Highlights

- High concordance between FlowGM and FlowJo counts of 15 cell types across 600 samples
- Significant effect of age on FlowGM exclusively segregated $CD14^{lo}CD16^{hi}$ monocytes and HLA-DR^{hi} cDC1 cells

3.1 Introduction

The central task of flow cytometry analysis is to identify cell populations of interest and obtain the relevant statistics for each population, including cell counts, cell proportions, and mean fluorescence intensities (MFIs). These statistics are regularly used as diagnostic criteria for many diseases, such as acquired immune deficiency syndromes (AIDS) and hematological malignancies. They can also help in discovery from large cohort analysis of new correlates between immune cell populations and diseases [Jaye et al., 2012] [Aghaeepour and Brinkman, 2014].

As we discussed in previous chapters, the flow cytometry data can be analyzed manually or automatically. The conventional manual analysis is based on visualization in 1D histograms or 2D scatter plots of the data. The use of new techniques for sample preparation

and more stringent pre-analytical procedures have improved the reproducibility and the quality of the results. The main limitations of manual analysis are that the analysis is subjective and it takes too much time. During recent years, more than 20 computational methods have been developed based on statistical models to automate the identification of cell populations. Cells are assigned to different clusters according to their expression level of antibodies, and then related statistics are obtained. However, how to evaluate the results is always challenging.

In the framework of data mining, two types of considerations are commonly taken for the evaluation of clustering results. One is based on the data itself, where the standard of evaluation consists of the similarity within the clusters and the dissimilarity between different clusters. Its main limitation for cell population identification is that, the resulted cluster might not necessarily represent a cell type. Therefore, in the field of FCM analysis, people usually consider another type of evaluation, which is based on the comparison with known class label. Due to the lack of ground truth, a benchmark which is manually built by experts is often considered as a gold standard, although people have a consensus that it is not a gold standard, but only a reference. In the analysis of small scale, for example, in the study of FlowCAP [Aghaeepour et al., 2013b], the idea of constructing a gold standard is realized by taking a consensus of 8 manual operators, which can undoubtedly reduce the subjectivity, but is obviously impractical in real applications, especially for those involving large cohort.

In the context of the MI Consortium, four 8-color panels were standardized for assessment of the major and minor cell populations present in peripheral whole blood collections. Data was collected and analyzed for $D = 600$ healthy donors in parallel by human experts and with our newly established software that employs Gaussian Mixture Model (FlowGM). Major leukocyte population statistics were obtained from both methods, which allowed a comparison between manual and automated analysis in a large number of samples.

The 600 healthy donors were stratified across gender (50% men and 50% women) and age (120 individuals from each decade of life between 20 to 70 years), demographics, including as well metabolism score, CMV infection status, smoking history and other lifestyle factors were recorded, which made the further comprehensive evaluation possible. For example, age has been shown to have an important influence on certain immune cell populations: the decline in plasmacytoid dendritic cells (pDCs) numbers with age has been reported. This type of observation has important implications for understanding immunosenescence.

In this chapter, we focus on the lineage panel and the dentritic cell panel to examine the quality of FlowGM results and the level of agreement with manual results. Furthermore, we concentrate on the influence of age on cell populations. Using a linear regression model, we test whether previously reported correlations can be confirmed by either manual or automatic FlowGM analsis.

3.2 Materials and Methods

3.2.1 Data

The 600 healthy donors were selected based on a stringent inclusion and exclusion criteria [Thomas et al., 2015]. Whole blood samples were collected from each donor and were stained with a lineage panel targeting T cells, B cells, NK cells, monocytes and PMN, and a dendritic cell panel targeting cDC1, cDC3 and pDC subpopulations. Flow cytometry data were manually analyzed using FlowJo (version 9.0 Treestar), and automatically analyzed with FlowGM. The FlowJo and FlowGM counts were obtained for in total 14 cell populations across all 600 donors.

3.2.2 Standard statistical analysis of the countings

All the statistical analysis discussing in this chapter for the evaluation of the results is standard, but they are straightforward and routine in FCM study involving large cohorts, and the results are biologically valuable. For these reasons we will introduce briefly the related statistical tools and explain how they are adapted to our problems. In order to evaluate FlowGM counts, we first use the Pearson's correlation to quantify the agreement of counts from manual and automated methods, then we employ the classical linear regression model to analyze the aging effect on all identified cell populations.

• Let Y_1 and Y_2 be the random variables representing the counting of a particular cell type computed by the FlowJo and FlowGM, respectively, for the same experiment. For $D = 600$ donors, the occurrences of Y_1 and Y_2 are denoted $\underline{\mathbf{y_1}} = \{y_1^{(1)}, \cdots, y_1^{(D)}\}^T$ and $\underline{\mathbf{y_2}} = \{y_2^{(1)}, \cdots, y_2^{(D)}\}^T$. To quantify the eventual linear dependence between Y_1 and Y_2, we consider the classical Pearson's correlation:

$$corr(Y_1, Y_2) = \frac{\text{cov}(Y_1, Y_2)}{\sigma_{Y_1}\sigma_{Y_2}}.$$

For our D-sample, the empirical correlation is

$$r_{y_1,y_2} = \frac{\sum_{d=1}^{D}(y_1^{(d)} - \bar{y}_1)(y_2^{(d)} - \bar{y}_2)}{\sqrt{\sum_{d=1}^{D}(y_1^{(d)} - \bar{y}_1)^2}\sqrt{\sum_{d=1}^{D}(y_2^{(d)} - \bar{y}_2)^2}}, \tag{3.1}$$

where $\bar{y}_1$ and $\bar{y}_2$ are the empirical means. The correlation r is computed independently for each cell type.

• We are now interested in testing the effect of age on the counting in order to infer its effect on circulating immune cell populations.

Consider the random counts Y (Y_1 or Y_2) and the random variable A of a donor's age. We would like to verify whether there is a significant linear relationship between Y and A as it is suggested in Figure 3.4. We consider the linear model:

$$\begin{aligned} &Y = \beta_0 + \beta_1\, A + \epsilon, \\ &\text{where} \quad \mathbb{E}(\epsilon) = 0,\ \text{var}(\epsilon) = \sigma^2. \end{aligned} \tag{3.2}$$

The conventional hypothesis testing technique is used to study the significance of deviation from age-independence. We consider the null hypothesis that the coefficient β_1 is equal to zero, which means that the age has no effect on the count of this cell population:

$$\begin{aligned} &H_0 : \beta_1 = 0, \quad \text{which implies} \quad (\mathcal{M}_0) : Y = \beta_0 + \epsilon \\ &H_1 : \beta_1 \neq 0, \quad \text{which implies} \quad (\mathcal{M}_1) : Y = \beta_0 + \beta_1\, A + \epsilon \end{aligned} \tag{3.3}$$

For deciding between these two hypotheses, one uses the classical F statistics:

$$F = \frac{(RSS_0 - RSS_1)/p}{RSS_1/(D-p-1)} \quad \sim \quad \text{Fisher}(p, D-p-1),$$

where $p = 1$ is the number of parameters in model $(\mathcal{M}_0)$, and $RSS_0 = \|\mathbf{y} - \bar{\mathbf{y}}\mathbb{1}\|^2$ and $RSS_1 = \|\mathbf{y} - \hat{\mathbf{y}}\|^2$, are the residual sum of squares for the least squares fit of the model $(\mathcal{M}_0)$ and $(\mathcal{M}_1)$ respectively. The statistics F has a Fisher distribution only under H_0. The decision is made as follows: given a risk of type I (i.e. the probability to decide H_1 whereas H_0 is true), the Fisher distribution returns the critical value f_α such that $p\,[\,\text{Fisher}(1, D-2) > f_\alpha] = \alpha$. We have chosen $\alpha = 0.05$. Finally H_0 is rejected if the observed F is greater than f_α.

• Instead of performing two independent tests (for Y_1 and then for Y_2), one can perform a single test. In this case, Y represents both Y_1 and Y_2 as it is formalized by the following linear model:

$$Y = \beta_0 + \beta_1\, A + \beta_2\, M + \beta_3\, A\, M + \epsilon, \tag{3.4}$$

where the dependant variable M represents the method type: $M = 0$ for FlowJo and $M = 1$ for FlowGM:

$$\begin{aligned} M = 0 \quad &\Rightarrow \quad Y = \beta_0 + \beta_1 A + \epsilon, \\ M = 1 \quad &\Rightarrow \quad Y = [\,(\beta_0 + \beta_2) + (\beta_1 + \beta_3)A\,] + \epsilon. \end{aligned}$$

To test the influence of M and A on Y, we consider the null and alternative hypotheses

$$\begin{aligned} &H_0 : \beta_2 = \beta_3 = 0, \quad \text{which implies} \quad (\mathcal{M}_0) : Y = \beta_0 + \beta_1 A + \epsilon, \\ &H_1 : H_0 \text{ is not true, which implies} \quad (\mathcal{M}_1) : Y = \beta_0 + \beta_1\, A + \beta_2\, M + \beta_3\, A\, M + \epsilon, \end{aligned} \tag{3.5}$$

H_0 states that M has no effect on the cell counting in model (3.4). Similarly,

$$F = \frac{(RSS_0 - RSS_1)/q}{RSS_1/(2D - p - 1)} \quad \sim \quad \text{Fisher}(q, 2D - p - 1),$$

where $q = 2$ is the number of regression coefficients restricted to zero, $p = 3$ is the number of no zero regression coefficients, RSS_0 and RSS_1 are the residual sum of squares for the least squares fit of the model $(\mathcal{M}_0)$ and $(\mathcal{M}_1)$ respectively.

If H_0 is rejected, which means that the two methods give different results, we further test only for β_3 to check which method is better:

$$\begin{aligned} &H_0 : \beta_3 = 0 \\ &H_1 : \beta_3 > 0, \quad \text{i.e. FlowGM counts provide stronger effect on age.} \end{aligned} \tag{3.6}$$

This unilateral test is performed using the t-test, because it is symmetric around 0.

3.3 Results

3.3.1 Mean fluorescence intensities

Before comparing directly the cell counts obtained from manual method and FlowGM, we would like to examine the quality of cell population identification across 600 donors. The MFIs of all 8 markers for each cell type and each donor are shown in Figure 3.1. Every circle represents the MFI of one donor, different colors correspond different cell types. The descriptions of the populations (e.g., $CD3^+CD4^+$ for the first line) confirm our labeling (e.g. as T cells for the first line). The compactness of donor's MFI shows a high quality of the results.

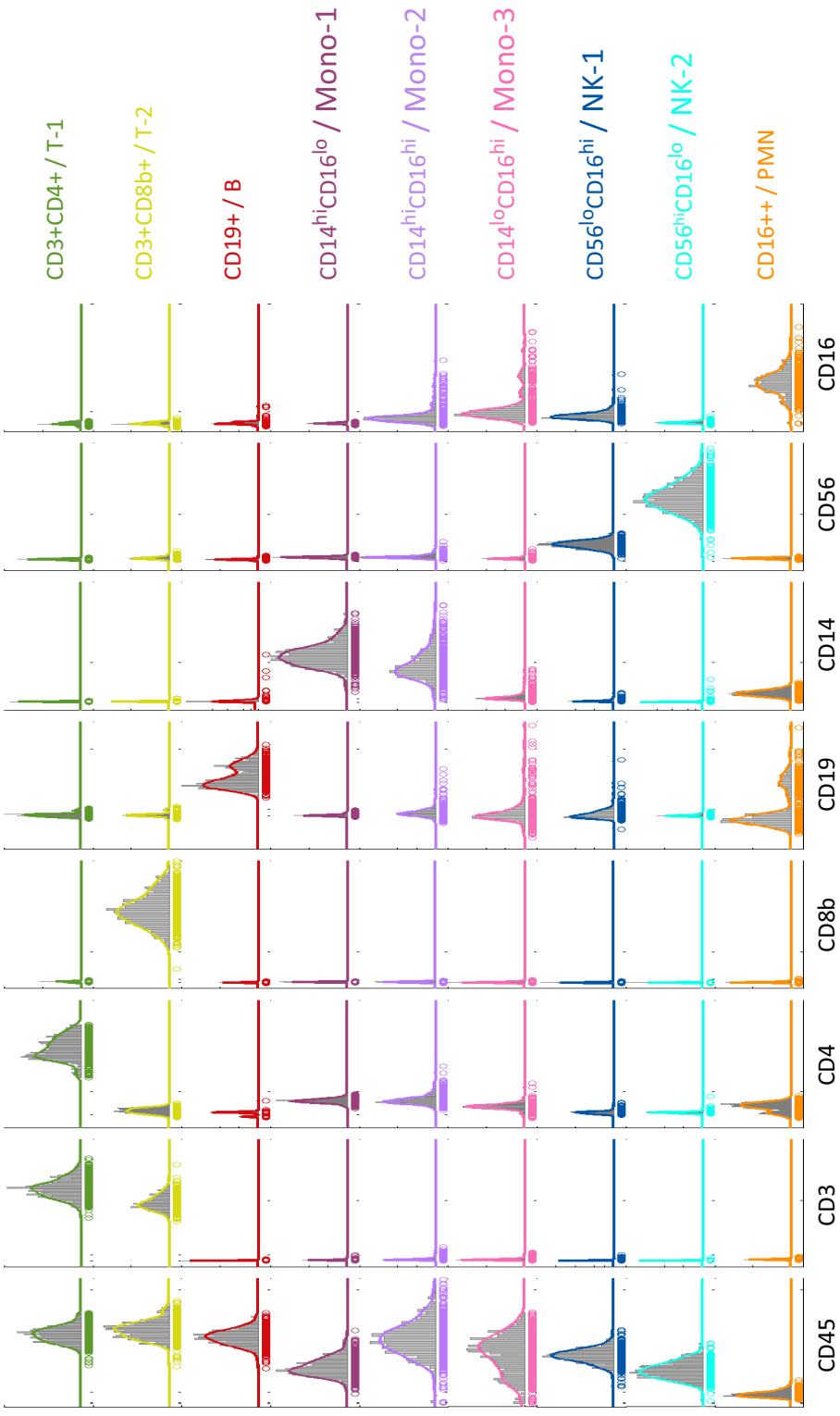

Figure 3.1: MFIs for each cell type and each donor. Every circle represent the MFI of one donor, different colors correspond to different cell types.

3.3.2 Counting correlation

We have shown at the end of Chapter 2, that the FlowGM and FlowJo counts correlation across 115 donors (Figure 2.19), here we extend our analysis to all 600 donors. Again, the results were highly concordant. Among all 14 identified cell populations (10 from the lineage panel and 4 from the DC panel), 11 of them have a correlation larger than 0.9 (Figure 3.2 and Figure 3.3), where the correlation r is computed by (3.1). More precisely, the average correlation $\bar{r} = 0.938$ for lineage panel and $\bar{r} = 0.864$ for DC panel.

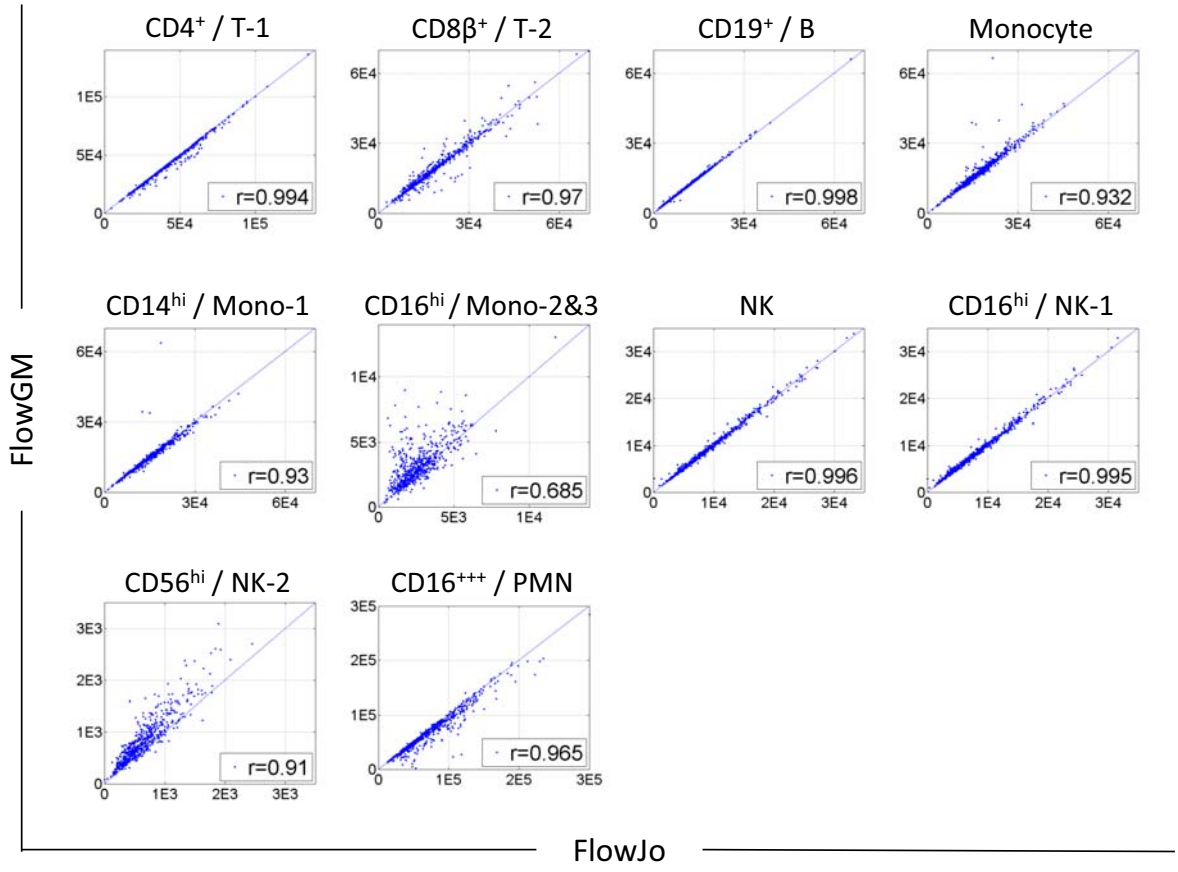

Figure 3.2: Comparison of manually counting and FlowGM analysis for lineage panel on $D = 600$ cohort donors. In this panel, we have identified 10 cell populations, for each of them, the X-axis is the counts from FlowJo, and the Y-axis is the counts from FlowGM. The counts obtained by these two methods highly agree on $D = 600$ cohort donors, with average correlation equal to 0.938 across all 10 identified cell populations.

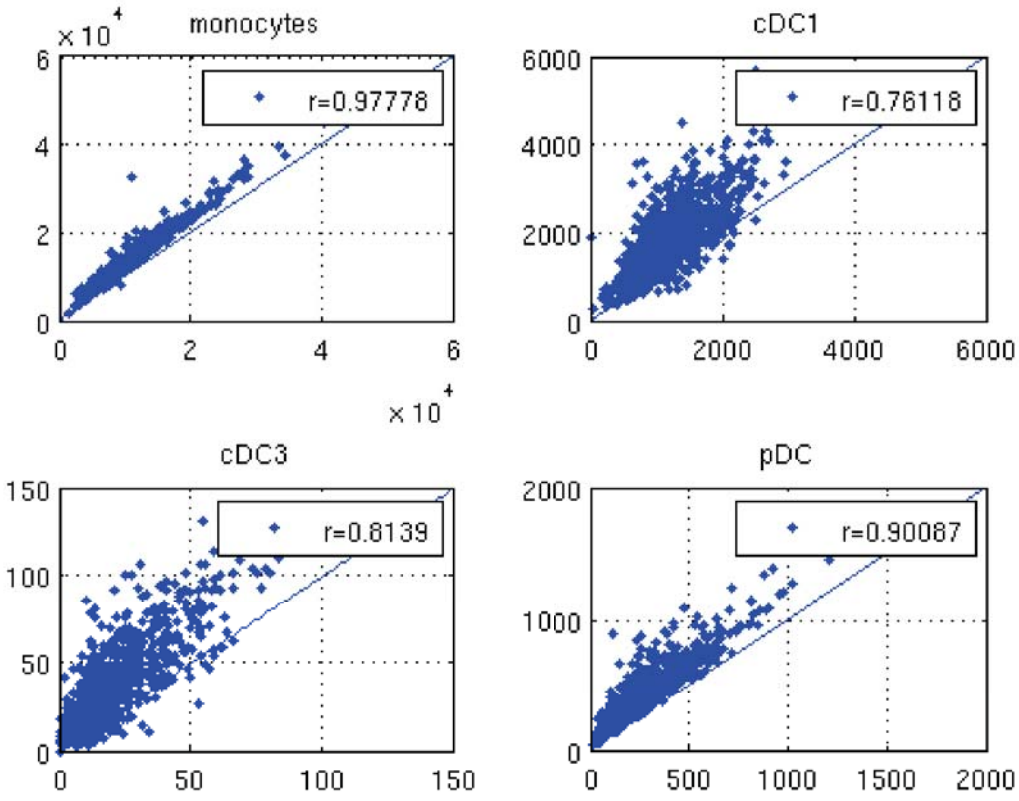

Figure 3.3: Comparison of manually counting and FlowGM analysis for lineage panel on $D = 600$ cohort donors. In this panel, we have identified 4 cell populations, for each of them, the X-axis is the counts from FlowJo, and the Y-axis is the counts from FlowGM. The counts obtained by these two methods highly agree on $D = 600$ cohort donors, with average correlation equal to 0.864 across all 4 identified cell populations.

3.3.3 Aging effect

Age is known to affect a number of different immune cell populations, both in terms of counts and function. We applied the test (3.3) to cell counts from manual and FlowGM analysis, we identified 12 cell types where the counts have a significant effect of age, and for 5 of them, which are $CD4^+$ T cells, $CD8\beta^+$ T cells, B cells, monocytes and $CD14^{hi}$ monocyte subpopulation, the counts from FlowJo and FlowGM are from the same linear regression model using statistic test (3.5). We plotted the FlowJo and FlowGM numbers of these cells against age and fitted a regression model separately for men (blue) and women (red) (Figure 3.4). The counts from FlowJo and FlowGM are from the same linear regression model, i.e. H_0 of test (3.5) is accepted for these 5 populations (Table 3.3.3).

Table 3.1: Decisions for multiple tests with significant level 0.05

	Cell populations	test (3.3), FlowJo [1]	test (3.3), FlowGM [2]	test (3.5) [3]	test (3.6) [4]
Lineage	$CD4^+$/ T-1	rejected	rejected	accepted	accepted
	$CD8\beta^+$/ T-2	rejected	rejected	accepted	accepted
	$CD19^+$/ B	rejected	rejected	accepted	accepted
	Monocyte	rejected	rejected	accepted	accepted
	$CD14^{hi}CD16^{lo}$/ Mono-1	rejected	rejected	accepted	accepted
	$CD16^{hi}$/ Mono-3	accepted	rejected	rejected	rejected
	NK	accepted	accepted	accepted	accepted
	$CD56^{lo}CD16^{hi}$/ NK-1	accepted	accepted	accepted	accepted
	$CD56^{hi}CD16^{lo}$/ NK-2	rejected	rejected	rejected	rejected
	$CD16^{+++}$/ PMN	rejected	rejected	rejected	accepted
DC	$CD14^{hi}$	rejected	rejected	accepted	accepted
	cDC1	rejected	rejected	rejected	rejected
	cDC3	accepted	rejected	rejected	rejected
	pDC	rejected	rejected	rejected	rejected

[1] Relevance of age on FlowJo counts, "rejected" means that $H_0 : \beta_1 = 0$ is rejected with the risque I and II equal to 0.05, which means that there is a significant linear relationship between FlowJo counts and age.

[2] Relevance of age on FlowGM counts, "rejected" means that $H_0 : \beta_1 = 0$ is rejected, which means that there is a significant linear relationship between FlowGM counts and age.

[3] The effect of the method on cell counts, "rejected" means that $H_0 : \beta_2 = \beta_3 = 0$ is rejected, i.e. FlowJo and FlowGM two methods give different counts with respect to the linear model.

[4] Aging effect importance, "rejected" means that $H_0 : \beta_3 = 0$ is rejected, i.e. FlowGM provides a stronger effect on age.

The exceptions include CD16^{hi} subpopulation of monocytes and dendritic cells. The discrepancies between FlowJo and FlowGM in the identification of monocyte subpopulations have been discussed in Chapter 2: FlowGM segregated the CD14^{hi}CD16^{hi} sub-population of monocytes (Mono-2) from CD14^{lo}CD16^{hi} sub-monocytes (Mono-3) despite the lack of additional monocyte-specific markers (e.g., MCSF-1, CX3CR1, CCR2), and the result was user-validated. In this study, we find that a clear decline with age is evident for CD14^{hi} subpopulation of monocytes (Mono-1) for both two methods, the manually identified CD16^{hi} subpopulation of monocytes (Mono-2&3) has no remarkable effect of age, however, the FlowGM identified CD14^{lo}CD16^{hi} sub-monocytes (Mono-3) showed a significant increase in the counts with increasing age (Figure 3.5). The immunology literature offers several possible explanations for this observations. First, levels of CD16^{hi} monocytes increase with new infection/inflammation [Pinke et al., 2013], thus aging people carry more CD16^{hi} monocytes than young people. Second, artherosclerosis increases CD16^{hi} levels [Tacke and Randolph, 2006] [Tacke et al., 2007], which is more common in the elderly. The compatibility of our observation with the literature suggests that our segregation, based on FlowGM, is valid. As for dendritic cells, it has been largely reported that the pDC count is decreasing with age. In our case, both methods give a confirmatory result. We evaluated whether manual or FlowGM analysis resulted in the stranger statistical evidence using (3.6). Figure 3.6A shows that the FlowGM result is less noisy, thus with stronger regression coefficient. More interesting, further precision of the cDC1 population is achieved using FlowGM, identifying three independent clusters that were segregated based on CD86 and HLA-DR expression. By focusing on cDC1 subsets, it was possible to demonstrate that the number of HLA-DR^{hi} cDC1 cells significantly decreases with age (Figure 3.6B).

We analyzed as well the effect of gender, CMV infection, smoking history, and other available information on more than 20 circulating immune cell populations. To have confidence in the stability of our phenotypes of circulating immune cells, we performed a second sampling on our healthy donors (Visit 2, one month after the initial analysis of Visit 1). Whole blood was collected, processed, analyzed, and cell counts were generated using FlowJo and FlowGM in exactly the same manner as for the first time. We also performed for all cell populations an outlier analysis to identify donors with counts more than 3 standard deviations from the mean. We revealed interesting outliers, individuals with $> 15\times$ the upper limit of normal for cDC1 or pDC numbers. We integrated our results with induced protein signatures and find that decreased pDC numbers in older individuals correlate to reduced levels of TLR7 and TLR9 induced cytokines. As the space constraints, detailed reports are not included here.

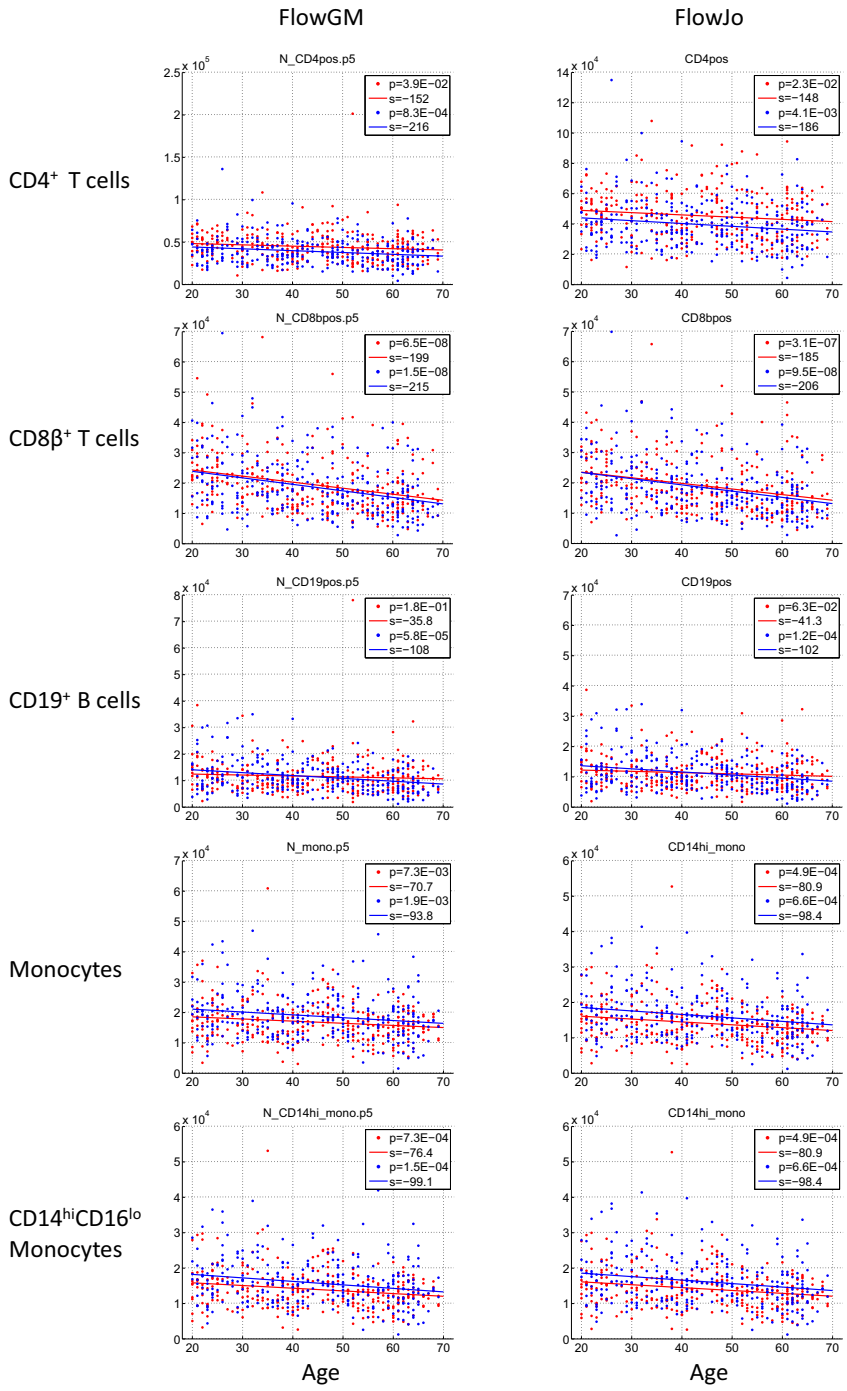

Figure 3.4: Effect of age on lymphocyte and monocyte populations. For CD4+ T cells, CD8β^+ T cells, B cells, monocytes and CD14^{hi} monocyte subpopulation, which correspond to the first 5 lines of Table 3.3.3, the FlowJo (right column) and FlowGM (left column) counts against age are plotted separately for men (blue) and women (red), where the X-axis is age and Y-axis is cell count. The performed regression model are plotted as well, the slope and p-value are given on the top-right of each plot, which suggest that there is a significant effect of age on the count of these cell populations.

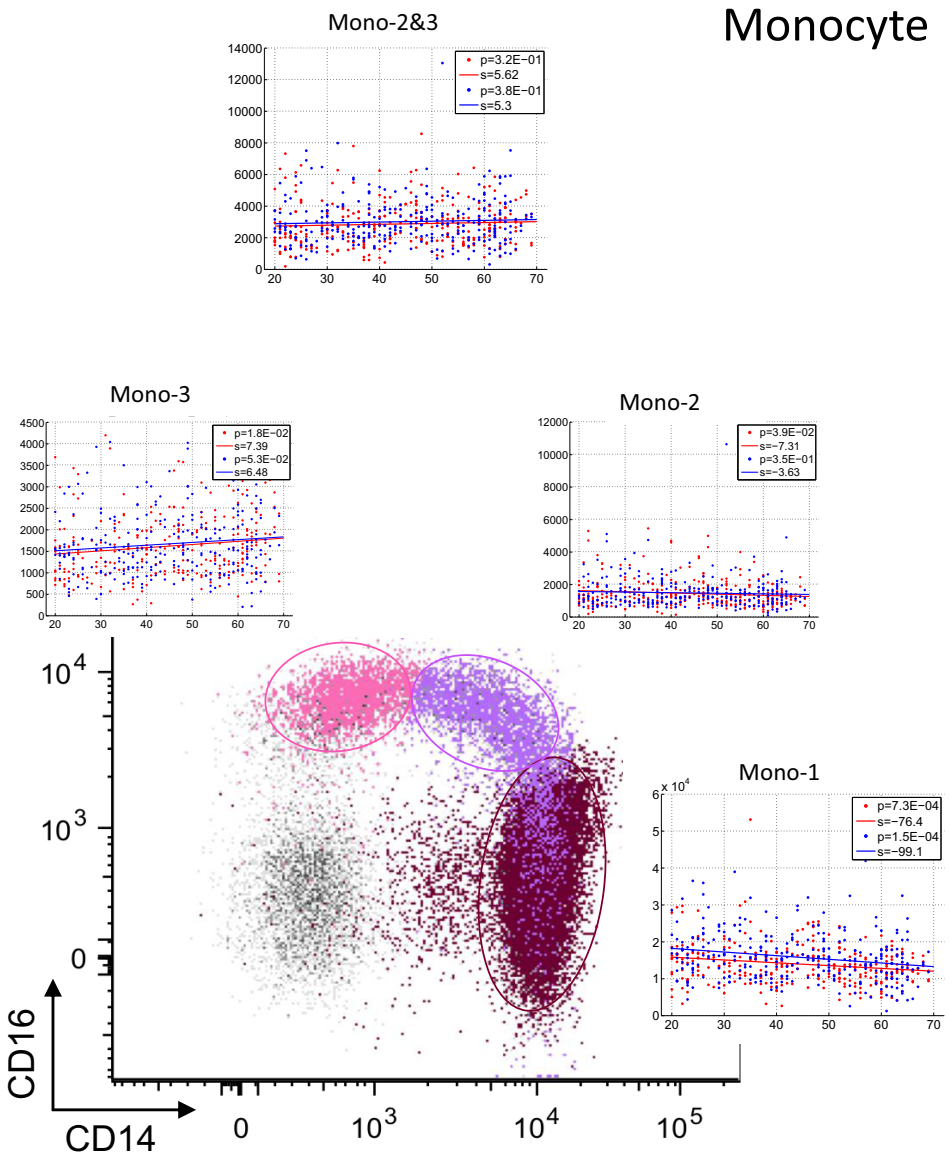

Figure 3.5: Aging effect on the counting of monocyte subpopulations. FlowGM identified CD14^{hi} monocytes (Mono-1, mauve) has an evident decline with age, CD14^{hi}CD16^{hi} monocytes (Mono-2, lavender) has no remarkable effect on age, and CD14^{lo}CD16^{hi} monocytes (Mono-3, light purple) showed a significant increase in the counts. On the other hand, manual analysis by FlowJo could not separate Mono-2 and Mono-3, and Mono-2&3 together shows no significant effect on age.

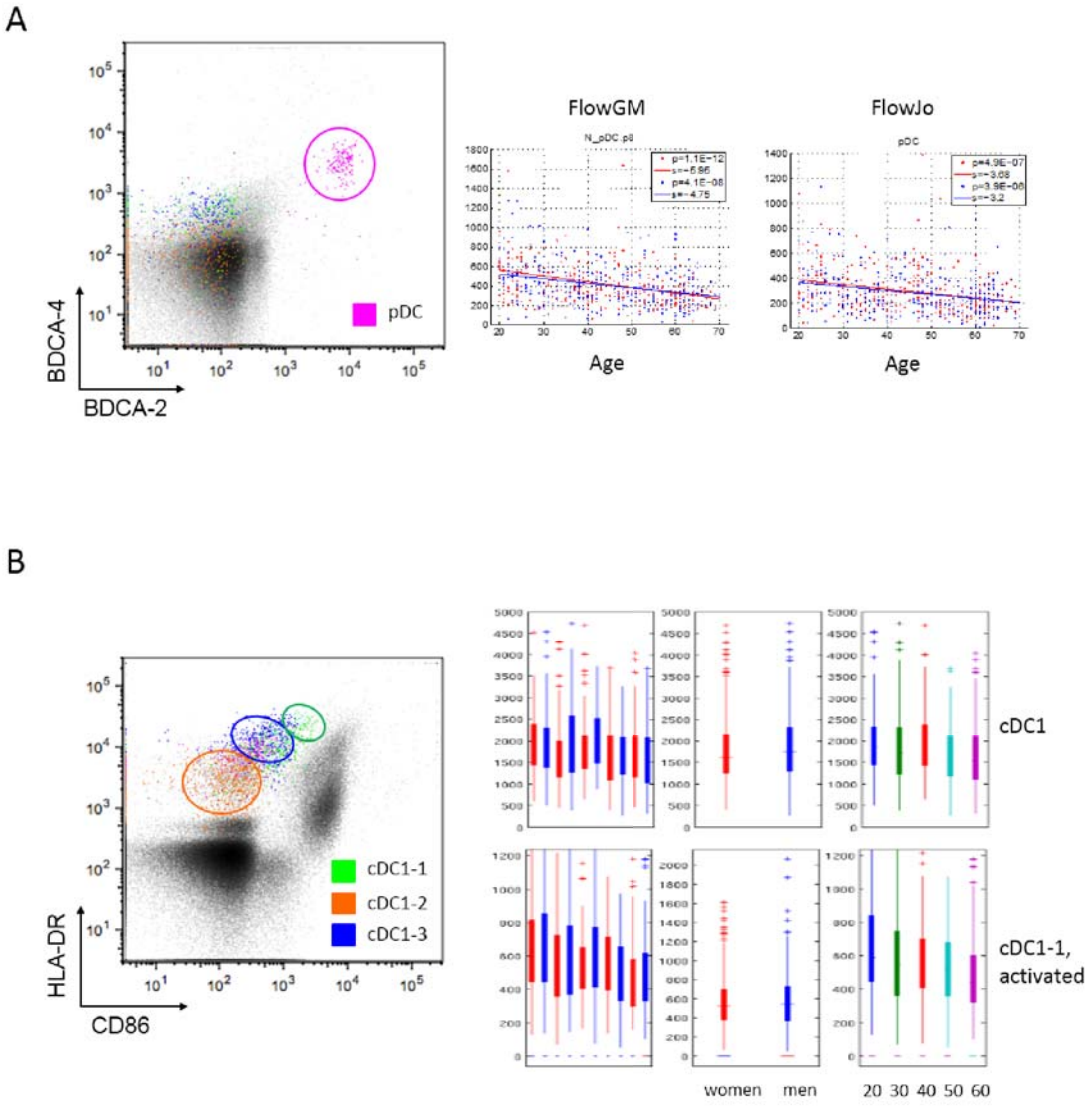

Figure 3.6: Aging effect on the counting of DC subpopulations (A) Both FlowGM and FlowJo methods give a pDC count significantly decreasing with age, for men (blue) and women (red). (B) Further precision of the cDC1 population is achieved using FlowGM, identifying three subpopulations (green, blue and orange) that were segregated based on HLA-DR and CD86 expression. The number of HLA-DR^{hi} cDC1 cells (cDC1-1, activated cDC1, green) show a significant decrease with age.

Chapter 4

Improvement of the FlowGM method by Bayesian Gaussian mixture modeling (FlowGMP)

Highlights

- The novel FlowGMP flow cytometry method targets a large number of samples where technological and biological variations are more relevant
- Largely automated and high quality analysis with consideration of prior information
- Quantification of 11 T cell subpopulations across 115 MI sample
- Validated performance is as good as, or even better than manual gating

4.1 Introduction and motivation

Let $\boldsymbol{X}$ denote a random vector in $\mathbb{R}^m$, and $\underline{\boldsymbol{x}} = \{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N\}$ denote a sample of N m-dimensional independent observations of $\boldsymbol{X}$, $N \gg m$. With the goal of clustering these observations, we consider a mixture model by assuming that each observation $\boldsymbol{x}_i$ is generated by a mixture of K multivariate distributions. Letting Z denote the latent variable indicating the mixture component that each observation belongs to, the probability distribution of an observation set $\underline{\boldsymbol{x}}$ can be written as

$$p_\theta(\underline{\boldsymbol{x}}) = \sum_{k=1}^{K} p(Z = k) p_{\theta_k}(\underline{\boldsymbol{x}} | Z = k),$$

where θ_k is the unknown parameter describing the distribution of the kth component, and $\theta = \{\theta_k\}_{k=1}^K$.

We aim to assign a component k to each observation $\boldsymbol{x}_i$, i.e. to estimate the occurrence z_i of Z by the value $\hat{z}_i$ that maximizes the posterior probability:

$$\hat{z}_i = \arg\max_k p_\theta(Z = k|\boldsymbol{x}_i). \tag{4.1}$$

By denoting $\breve{\zeta}_k$ the indicator function of the $k$the estimated cluster, this cluster is defined by:

$$\breve{\zeta}_k(\underline{\boldsymbol{x}}) = \{\boldsymbol{x}_i|\, \hat{z}_i = k\},$$

the clustering is then defined as

$$\mathcal{Z}(\underline{\boldsymbol{x}}) = \{\breve{\zeta}_1(\underline{\boldsymbol{x}}), \cdots, \breve{\zeta}_K(\underline{\boldsymbol{x}})\},$$

or simply if there is no confusion

$$\mathcal{Z} = \{\breve{\zeta}_1, \cdots, \breve{\zeta}_K\}.$$

The number of points in every cluster is denoted $n_k = |\breve{\zeta}_k| = \sum_i \mathbb{1}(\hat{z}_i = k)$.

We consider an application in flow cytometry, where the goal is the characterization of the cellular component of the immune system. In one flow cytometry experiment, we measure simultaneously the expression level of m antibodies for each of the N cells, then we cluster the N cells into different groups based on their m-measurements, and finally we assign a cell type to each group with the help of experts. This process is referred as "Cell population identification". In fact, in the analysis of flow cytometry data, which motivates this work, the observations come from multiple experiments. Let D be the number of experiments, and denote the set of the D samples by

$$\mathcal{X} = \{\underline{\boldsymbol{x}}^{(1)}, \cdots, \underline{\boldsymbol{x}}^{(D)}\}.$$

The observed sample $\underline{\boldsymbol{x}}^{(d)}$ of $\boldsymbol{X}$ generated by the dth experiment is a $N^{(d)}$-tuple of points in $\mathbb{R}^m$:

$$\underline{\boldsymbol{x}}^{(d)} = \{\boldsymbol{x}_1^{(d)}, \cdots, \boldsymbol{x}_{N^{(d)}}^{(d)}\},\ d = 1, \cdots, D.$$

Theoretically, this writing implies that each sample $\underline{\boldsymbol{x}}^{(d)}$ comes from the same random vector $\mathbf{X}$. In practice, this assumption is perturbed by exogenous factors.

Recent technical advances have encouraged the studies involving cohorts with large numbers D of patients, replicates, or experiments with different stimulation conditions [Pyne et al., 2014]. Therefore, the difficulty in the analysis of such large number of experiments lies in the presence of technical and biological variations. Technical variations arise from

varying experimental conditions, instrument settings, or laboratory operations. Biological variations are, for example, the age and gender effect as we discussed in chapter 3. In flow cytometry analysis, the size cluster n_k is of main importance since it is used for cell identification. The exogenous factors can perturb the clustering and therefore n_k too. Therefore, we must take into account the effects of these factors in our analysis.

Challenge The need to assign an immunological cell type implies that we have to not only to identify cell populations in an individual experiment or in several experiments independently, but we also have to "align" cell populations across all experiments, i.e., identify those populations corresponding to the same cell type. Figure 4.1 illustrates this problem. Once such an alignment is established for all populations, and we know the cell type of one population in one experiment, say, in experiment $d = 0$, then this allows us to infer the cell type in all other experiments.

Formally speaking, we deal with two specific properties: (1) the latent variable Z has

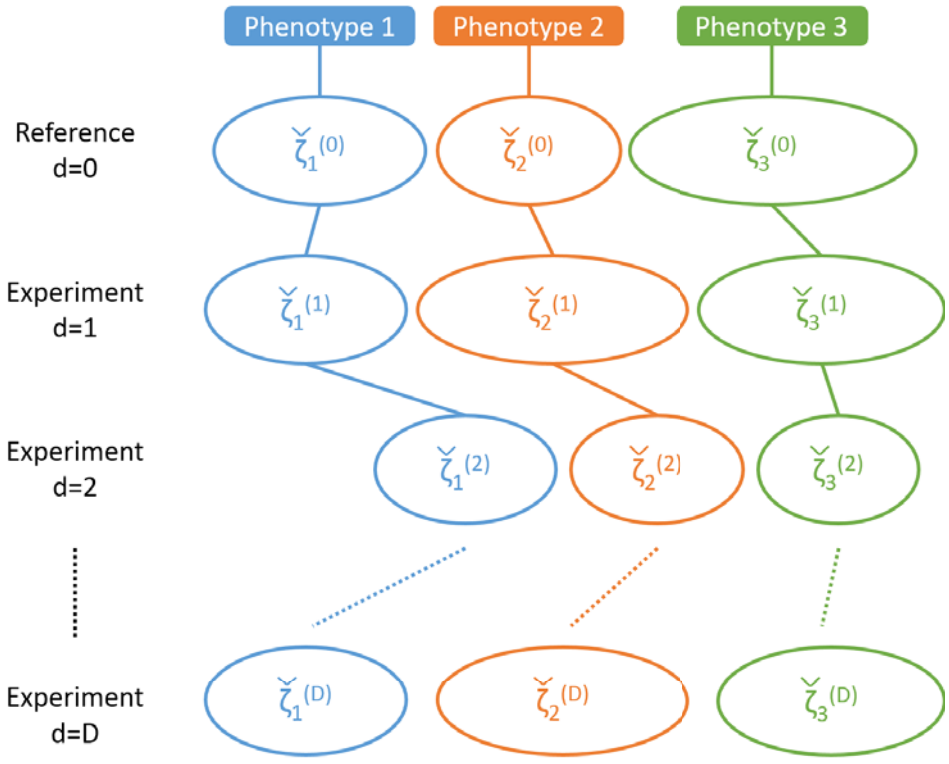

Figure 4.1: Alignment of clusters across experiments. Three clusters are identified in the reference experiment $d = 0$, presenting three different phenotypes. The phenotype of all other estimated clusters $\breve{\zeta}_k^{(d)}, d = 1, \cdots, D$ is the same, thanks to the alignment on the reference sample.

some common phenotypic interpretation across all experiments, which means that, for any $k = 1, \cdots, K$, each estimated cluster $\breve{\zeta}_k^{(d)}$ represents the same cell type, regardless of d; (2) the parameter $(\alpha_k^{(d)}, \theta_k^{(d)})$ depends on the experiment and varies from one to another due to the variations, which is in contradiction with the property (1).

Automated analysis methods therefore need to be able to partition observed data from each experiment into "proper" clusters, and to characterize at the same time the intrinsic variation systematically, where the "proper" quality refers to the "cell type" interpretation. It is shown that, for clustering analysis, an algorithm designed for some kind of models cannot perform well if the dataset contains a radically different set of models, or if the evaluation consists of a radically different criterion [Estivill-Castro, 2002]. Concrete to our case, the variation across all experiment datasets can be sometimes very large (for example, different age groups follow different models), moreover, the "cell type" interpretation, as the evaluation criteria, consists often much complicated situation than in a simple modeling. Therefore, a model which doesn't take into account of these considerations will fail.

There has been a number of approaches to address this situation in model-based clustering within the Bayesian framework [Pan and Shen, 2007]. In the field of flow cytometry, there are also several attempts considering a large number of samples and the alignment issue with regard to phenotypic relevant clusters despite technical and biological variation [Pyne et al., 2014] [Cron et al., 2013] [Dundar et al., 2014]. For these publications, FlowCAP [Aghaeepour et al., 2013a] competition datasets are commonly used as the basis for parameters and method calibration. Here, we present a mixture model-based approach for multi-parameter clustering analysis across a large number of samples, where the variations are prominent. It is motivated by the needs of the MI project [Thomas et al., 2015], where flow cytometry data is on a larger scale (thousands of samples) and has more a complicated structure than the FlowCAP datasets.

In the following part of this chapter, we propose in Section 4.2 a Bayesian mixture model that assumes a prior distribution on the component parameter θ_k with respect to the phenotypic (cell type) interpretation of each component. In fact, this prior distribution is limited to the class center $\boldsymbol{\mu}_k$ in θ_k. Then we derive the corresponding penalized EM algorithm. The computation is restricted to the Gaussian case. Finally, we discuss the application on MI flow cytometry analysis in Section 4.3.

4.2 Models and method

4.2.1 Conventional mixture model

Let $\mathbf{X}$ denote a random vector in $\mathbb{R}^m$, $\underline{\boldsymbol{x}} = \{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N\}$ denote N m-dimensional independent observations of $\boldsymbol{X}$, and z_i, taking value k in $\{1, \cdots, K\}$, which denotes the cluster labels. The mixture model has the form

$$\begin{aligned} p(\boldsymbol{x_i}|\phi) &= \sum_{k=1}^{K} p(Z_i = k) p(\boldsymbol{x_i}|Z_i = k) \\ &= \sum_{k=1}^{K} \alpha_k P_{\theta_k}(\boldsymbol{x_i}|Z_i = k)\,, \\ \text{where } \phi &= \{\alpha_k, \theta_k\}_{k=1}^{K}\,. \end{aligned} \tag{4.2}$$

The vector $\boldsymbol{\alpha} = \{\alpha_1, \cdots, \alpha_K\}$ denotes the mixing proportions. If the $\boldsymbol{x}_i$ are independent observations of $\mathbf{X}$, the log likelihood is

$$l(\phi|\underline{\boldsymbol{x}}) = \sum_{i=1}^{N} \log p(\boldsymbol{x}_i|\phi) = \sum_{i=1}^{N} \log(\sum_{k=1}^{K} \alpha_k P_{\theta_k}(\boldsymbol{x}_i|Z_i = k)),$$

and the conventional maximum likelihood estimate (MLE) is

$$\hat{\phi}_{ML}(\underline{\boldsymbol{x}}) = \arg\max_{\phi} l(\phi|\underline{\boldsymbol{x}}). \tag{4.3}$$

4.2.2 Bayesian mixture model

The challenge described in Section 4.1 appears when we consider a large set of experiments $\mathcal{X} = \{\underline{\boldsymbol{x}}^{(d)}\}_{1 \le d \le D}$, where D is the number of experiments. The identical value of $z_i^{(d)}$ through d must lead to the same phenotypic interpretation of the clusters, so that the counts $n_k^{(d)}$, the parameter $\theta_k^{(d)}$ and other statistics from different experiment are comparable, in order to allow the characterization of scientific variations in further studies as it will be detailed below.

With this consideration, we propose a second layer of model by assuming a prior distribution on every θ_k, i.e. for a given k, each θ_k is an occurrence of a probability distribution modeling fluctuation, as illustrated in Figure 4.1. In the Gaussian case, $\boldsymbol{x_i}|(Z_i = k) \sim \mathcal{N}(\boldsymbol{\mu}_k, V_k)$, then $\theta_k = \{\boldsymbol{\mu}_k, V_k\}$. We only assume $\boldsymbol{\mu}_k$ to be a random variable, and use the following prior distribution:

$$\boldsymbol{\mu}_k \sim \mathcal{N}(\boldsymbol{\tau}_k, \delta\Gamma_k),$$

where $\boldsymbol{\tau}_k$ defines the prior location of $\boldsymbol{\mu}_k$, Γ_k describes the shape and orientation of the distribution, and $\delta > 0$ controls the spread, which rules the importance of the prior. This

parameter is called *hyper-parameter*. No specific prior is assumed on other parameters Consequently, due to the focus on $\boldsymbol{\mu}_k$ and the independence assumption between the parameters, we have

$$p(\phi) = p(\boldsymbol{\alpha})\, p(\boldsymbol{\theta}) = p(\boldsymbol{\alpha})\, p(\underline{\boldsymbol{\mu}})\, p(V) = p(\underline{\boldsymbol{\mu}}) = \prod_{k=1}^{K} p(\boldsymbol{\mu}_k),$$

where $\underline{\boldsymbol{\mu}} = \{\boldsymbol{\mu}_1, \cdots, \boldsymbol{\mu}_K\}$. Within this Bayesian framework, the prior distribution leads the posterior probability distribution

$$p(\phi|\underline{\boldsymbol{x}}) = \frac{p(\underline{\boldsymbol{x}}|\phi)p(\phi)}{p(\underline{\boldsymbol{x}})} = \frac{p(\underline{\boldsymbol{x}}|\phi)p(\underline{\boldsymbol{\mu}})}{p(\underline{\boldsymbol{x}})}.$$

The log-posterior probability is then

$$l_p(\phi|\underline{\boldsymbol{x}}) = \sum_{i=1}^{N} \log p(\boldsymbol{x}_i|\phi)p(\underline{\boldsymbol{\mu}}) = \sum_{i=1}^{N} \log(\sum_{k=1}^{K} \alpha_k p_{\theta_k}(\boldsymbol{x}_i|z_i = k)) + \log p(\underline{\boldsymbol{\mu}}). \tag{4.4}$$

It can be interpreted as a log penalized likelihood whose penalization is $\log p(\boldsymbol{\mu})$. With the Gaussian prior assumption, the last term of (4.4) is

$$\log p(\underline{\boldsymbol{\mu}}) = \log \prod_{k=1}^{K} p(\boldsymbol{\mu}_k) = \sum_{k=1}^{K} (-\frac{1}{2}(\boldsymbol{\mu}_k - \boldsymbol{\tau}_k)'(\delta\Gamma_k)^{-1}(\boldsymbol{\mu}_k - \boldsymbol{\tau}_k) - \frac{m}{2}\log(2\pi) - \frac{1}{2}\log|(\delta\Gamma_k)|),$$

thus (4.4) could be written as

$$-l_p(\phi|\underline{\boldsymbol{x}}) = -l(\phi|\underline{\boldsymbol{x}}) + \sum_{k=1}^{K} (\frac{1}{2}(\boldsymbol{\mu}_k - \boldsymbol{\tau}_k)'(\delta\Gamma_k)^{-1}(\boldsymbol{\mu}_k - \boldsymbol{\tau}_k) + const.$$

The maximum a posterior estimate (MAP) is

$$\hat{\phi}_{\text{MAP}}(\underline{\boldsymbol{x}}) = \arg\max_{\phi} l_p(\phi|\underline{\boldsymbol{x}}).$$

When maximizing $l_p(\phi|\underline{\boldsymbol{x}})$ or minimizing $-l_p(\phi|\underline{\boldsymbol{x}})$, it is to minimize the negative log likelihood $-l(\phi|\underline{\boldsymbol{x}})$ penalized with a sum of Mahalanobis distances between each point $\boldsymbol{\mu}_k$ and its prior distribution $\mathcal{N}(\boldsymbol{\tau}_k, \delta\Gamma_k)$. The Mahalanobis distance between a point $\boldsymbol{\mu}_k$ and a distribution $\mathcal{N}(\boldsymbol{\tau}_k, \delta\Gamma_k)$ has the form

$$\|\boldsymbol{\mu}_k - \boldsymbol{\tau}_k\|^2_{\delta\Gamma_k} = \frac{1}{2}(\boldsymbol{\mu}_k - \boldsymbol{\tau}_k)'(\delta\Gamma_k)^{-1}(\boldsymbol{\mu}_k - \boldsymbol{\tau}_k) = \frac{1}{\delta}\|\boldsymbol{\mu}_k - \boldsymbol{\tau}_k\|^2_{\Gamma_k}.$$

Then

$$-l_p(\phi|\underline{\boldsymbol{x}}) = -\, l(\phi|\underline{\boldsymbol{x}}) + \frac{1}{\delta}\sum_{k=1}^{K} \|\boldsymbol{\mu}_k - \boldsymbol{\tau}_k\|^2_{\Gamma_k}.$$

Big value of δ means a large spread of the distribution of $\boldsymbol{\mu}_k$, which indicates a weak belief in the prior, therefore a small penalization; while small value of δ means a little spread, which can be interpreted as a strong belief in the prior, therefore a big penalization.

4.2.3 Penalized-EM algorithm

For a conventional mixture model (4.2) and maximum likelihood estimation (4.3), an expectation-maximization (EM) algorithm is commonly used. Starting from an initial guess for the parameters, each stage of an iterative optimization process improves the likelihood in two steps: in an E (Expectation) step, the expectation of log-likelihood is evaluated using the current estimate for the parameters, and in an M (Maximization) step, the parameters are optimized by maximizing the expected log-likelihood.

Since the EM algorithm will converge to a local optimum, the result is, in general, sensitive to the starting values of the parameters. This local property could be sometimes a defect for certain cases, but we can here take advantage of it to enable the automated alignment of clusters across all experiments. We propose to build a reference clustering that contains one cluster for each target component, and to then use the parameters of this reference clustering as an initialization for all experiments. A penalized EM algorithm is described here, which indexes each cluster in new experiment according to the reference, which allows automated alignment of clusters across all experiments.

Like in the standard EM algorithm for (4.3), we consider the expectation of the log penalized likelihood with respect to the random variables $(Z_i|X_i)$ whose distribution are $p(z_i|\boldsymbol{x}_i, \phi')$, ϕ' denoting a previous estimated parameter:

$$\begin{aligned} Q_p(\phi, \phi') =& \mathbb{E}_{Z|X,\phi'}\left[\log \prod_{i=1}^{N} p(\boldsymbol{x_i}, Z_i|\phi)p(\phi)\right] \\ =& \mathbb{E}_{Z|X,\phi'}\left[\sum_{i=1}^{N} \log[p(\boldsymbol{x_i}, Z_i|\phi)p(\phi)]\right] \\ =& \sum_{i=1}^{N} \mathbb{E}_{Z|X,\phi'}\left[\log[p(\boldsymbol{x_i}, Z_i|\phi)p(\phi)]\right] \\ =& \sum_{i=1}^{N}\sum_{k=1}^{K} p(Z_i = k|\boldsymbol{x_i}, \phi') \log(\alpha_k p(\boldsymbol{x_i}|\theta_k)p(\boldsymbol{\mu_k})), \end{aligned} \tag{4.5}$$

By denoting

$$p(Z_i = k|\boldsymbol{x_i}, \phi') = \gamma_{ik}^{'},$$

(4.5) could be written as

$$Q_p(\phi, \phi') = \sum_{i=1}^{N}\sum_{k=1}^{K} \gamma_{ik}^{'} \log \alpha_k + \sum_{i=1}^{N}\sum_{k=1}^{K} \gamma_{ik}^{'} \log p(\boldsymbol{x_i}|\theta_k) + \sum_{i=1}^{N}\sum_{k=1}^{K} \gamma_{ik}^{'}\delta \log p(\boldsymbol{\mu_k}).$$

Then the Bayesian estimate $\hat{\phi} = \{\hat{\alpha}_k, \hat{\theta}_k\}_{k=1}^K$ is the solution of

$$\left.\frac{\partial Q_p(\phi, \phi')}{\partial \alpha_k}\right|_{\phi=\hat{\phi}} = 0 \quad \text{and} \quad \left.\frac{\partial Q_p(\phi, \phi')}{\partial \theta_k}\right|_{\phi=\hat{\phi}} = 0.$$

As the penalization is independent of α_k, the re-estimation of α_k remains the same with that in standard situation. Now we get

$$\left.\frac{\partial Q_p(\phi, \phi')}{\partial \boldsymbol{\mu}_k}\right|_{\phi=\hat{\phi}} = 0 \quad \text{implies } \sum_i \gamma_{ik}' \hat{V}_k^{-1}(\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_k) - (\delta\Gamma_k)^{-1}(\hat{\boldsymbol{\mu}}_k - \boldsymbol{\tau}_k) = 0,$$

thus

$$\hat{\boldsymbol{\mu}}_k = (-\sum_i \gamma_{ik}' \hat{V}_k^{-1} + (\delta\Gamma_k)^{-1})^{-1}(-\sum_i \gamma_{ik}' \hat{V}_k^{-1}\boldsymbol{x}_i + (\delta\Gamma_k)^{-1}\boldsymbol{\tau}_k) \tag{4.6}$$

As we don't have any prior on V_k, the new estimate of V_k has the standard form [Bilmes et al., 1998], which depends to the new $\hat{\boldsymbol{\mu}}_k$:

$$\hat{V}_k = \frac{\sum_i \gamma_{ik}'(\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_k)'(\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_k)}{\sum_i \gamma_{ik}'}.$$

The difficulty here is that (4.6) depends to the new $\hat{V}_k$, and solving simultaneity the system on these two parameters is too complicate. An optimization of coordinate by coordinate [Friedman et al., 2007] within each EM iteration is then proposed here:

$$\text{Iteration (1):} \begin{cases} \hat{\boldsymbol{\mu}}_k^{(1)} \leftarrow (-\sum_i \gamma_{ik}'(V_k')^{-1} + (\delta\Gamma_k)^{-1})^{-1}(-\sum_i \gamma_{ik}'(V_k')^{-1}\boldsymbol{x}_i + (\delta\Gamma_k)^{-1}\boldsymbol{\tau}_k), \\ \hat{V}_k^{(1)} \leftarrow \dfrac{\sum_i \gamma_{ik}'(\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_k^{(1)})'(\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_k^{(1)})}{\sum_i \gamma_{ik}'}; \end{cases}$$

$$\text{Iteration (t+1):} \begin{cases} \hat{\boldsymbol{\mu}}_k^{(t+1)} \leftarrow (-\sum_i \gamma_{ik}'(\hat{V}_k^{(t)})^{-1} + (\delta\Gamma_k)^{-1})^{-1}(-\sum_i \gamma_{ik}'(\hat{V}_k^{(t)})^{-1}\boldsymbol{x}_i + (\delta\Gamma_k)^{-1}\boldsymbol{\tau}_k), \\ \hat{V}_k^{(t+1)} \leftarrow \dfrac{\sum_i \gamma_{ik}'(\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_k^{(t+1)})'(\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_k^{(t+1)})}{\sum_i \gamma_{ik}'}; \end{cases}$$

where ϕ' indicates the current guess of appropriate parameters in each EM iteration, and within this iteration, a second loop for coordinate optimization of estimated parameters is noted as $\hat{\phi}^{(t)}, t = 1, 2, \cdots$.

The penalized version of EM algorithm described here owns also the property: the likelihood $p_{\hat{\theta}}(\underline{\boldsymbol{x}})$ increases at each iteration of the EM algorithm. The proof is given in Appendix.

4.2.4 Model calibration

To estimate the hyper-parameter δ, we use a measure of similarity between two clusterings performed on a N-sample $\underline{\boldsymbol{x}} = \{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N\}$. The first clustering is here considered as

"ground truth", where the clusters are called "classes". These classes, denoted here by $\breve{c}_k(\underline{\boldsymbol{x}})$, represent a benchmark, which has been created by human experts. The second clustering, where clusters are denoted by $\breve{\zeta}_k(\underline{\boldsymbol{x}})$, are the output of the penalized-EM algorithm, which depend on δ. The similarity measure is therefore a real function of $\breve{c}_k$ and $\breve{\zeta}_k$, which is high if the clusters are similar to the classes, and close to zero otherwise.

Similarity measure for the comparison of two clusterings

Here we consider a similarity related to the F-measure [van Rijsbergen, 1979], which is well-known in the Data mining literature, and has proven to be successful to evaluate the performance of automated cell population identification algorithms [Rosenberg and Hirschberg, 2007] [Aghaeepour et al., 2011]. Our choice of similarity measure is motivated by the goal to align different flow cytometry samples.

Let $\mathcal{C} = \{\breve{c}_1, \cdots, \breve{c}_{K_1}\}$ be the set of human expert labeled classes, and $\mathcal{Z} = \{\breve{\zeta}_1, \cdots, \breve{\zeta}_{K_2}\}$ the set of clusters assigned by an algorithm, where K_1 and K_2 are total number of classes/clusters in $\mathcal{C}$ and $\mathcal{Z}$ respectively, and they are not forced to be the same.

Let us consider two conditional probabilities: $P(i,j) = p(Z = j|C = i)$ and $R(i,j) = p(C = i|Z = j)$, where C denotes the variable indicating the class that each observation belongs to. In the data mining literature, these probabilities are referred as "Precision" and "Recall", respectively. Their estimations are straightforward. By denoting $n_{i;j} = |\breve{c}_i \bigcap \breve{z}_j|$ the number of co-occurrences in class $\breve{c}_i$ and cluster $\breve{z}_j$, and $n_i = |\breve{c}_i|$, $n_j = |\breve{z}_j|$, then

$$\hat{P}(i,j) = \frac{n_{i;j}}{n_j}, \quad \hat{R}(i,j) = \frac{n_{i;j}}{n_i}$$

are the estimations of these probabilities. Figure 4.2 shows some examples of poor or imperfect similarities.

The F-measure between class $\breve{c}_i$ and cluster $\breve{\zeta}_j$ is defined as the harmonic mean of precision and recall:

$$f(i,j) = 2 \cdot \frac{\hat{P}(i,j) \cdot \hat{R}(i,j)}{\hat{P}(i,j) + \hat{R}(i,j)}, \tag{4.7}$$

Intuitively, high precision means that cluster $\breve{\zeta}_j$ includes more points in class $\breve{c}_i$ than those are not, while high recall means that cluster $\breve{\zeta}_j$ includes most of the points in class $\breve{c}_i$. Therefore, $f(i,j) \in [0,1]$, high value of $f(i,j)$ means a high similarity between class $\breve{c}_i$ and cluster $\breve{\zeta}_j$.

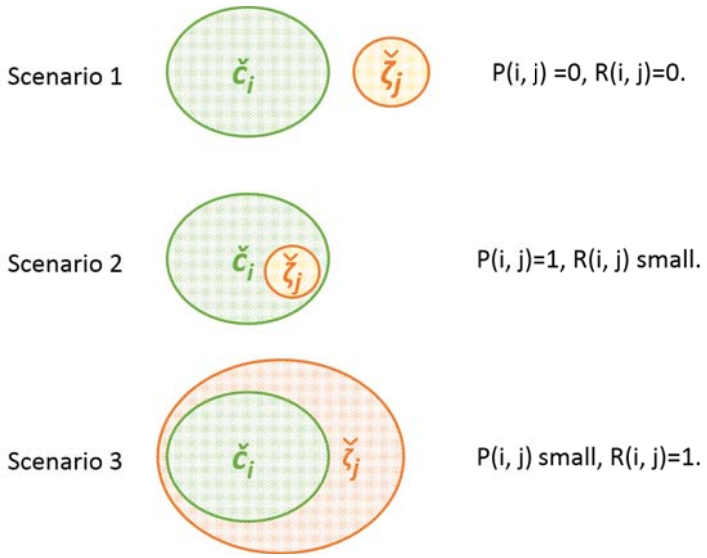

Figure 4.2: Comparison of a cluster with a class. The class $\breve{c}$, colored in green, is considered as a "ground truth", and the cluster $\breve{\zeta}$, colored in red, is the output of an algorithm. Scenario 1-3 show examples of pool similarities between class $\breve{c}$ and cluster $\breve{\zeta}$.

Then the F-measure between two clusterings $\mathcal{C}$ and $\mathcal{Z}$ is defined as a weighted sum of $f(i,j)$ between class $\breve{c}_i$ and its best matching cluster $\breve{\zeta}_j$ for all classes:

$$F(\mathcal{C},\mathcal{Z}) = \sum_{\breve{c}_i \in \mathcal{C}} \frac{n_i}{N} \max_{\breve{\zeta}_j \in \mathcal{Z}} f(i,j). \tag{4.8}$$

One can easily prove that $F \in [0,1]$. High value of $F(\mathcal{C},\mathcal{Z})$ means that all classes in $\mathcal{C}$ have a highly matching clusters in $\mathcal{C}$, taking into account of different contributions of class sizes.

For our application, we modify the F measure as follows. We assume that clusters $\mathcal{Z}$ are ordered corresponding to classes in $\mathcal{C}$, which is to say $K_1 = K_2$, and $i = j$ implies class $\breve{c}_i$ and cluster $\breve{z}_j$ correspond. Now, to deal with the alignment task, we define here the F_p-measure, which looks for the similarity of corresponding cluster instead of the best matching cluster for each class $\breve{c}_i$:

$$F_p(\mathcal{C},\mathcal{Z}) = \sum_{\breve{c}_i \in \mathcal{C}} \frac{n_i}{N} f(i,i). \tag{4.9}$$

Similarly, $F_p \in [0,1]$, and we have always $F_p(\mathcal{C},\mathcal{Z}) \leq F(\mathcal{C},\mathcal{Z})$. High value of $F_p(\mathcal{C},\mathcal{Z})$ means that each class $\breve{c}_i$ in $\mathcal{C}$ have a high similarity with clusters $\breve{\zeta}_i$ in $\mathcal{Z}$, taking into account of different contributions of class sizes.

In some applications, there may exist some rare but important populations. To address to this situation and to reveal the contribution of rare classes, one could consider an

unweighted F-measure by assuming that all classes contribute equally to the quality of clustering, irrespective of class size.

Estimation of the tuning parameter

Given δ, the penalized-EM algorithm described in Section 4.2.3 provides a clustering set $\mathcal{Z}^{(d)}[\delta]$ for each experiment d, thus permit a prediction of classes. If the ground truth $\mathcal{C}^{(d)}$ is available, the estimation of prediction error could be achieved through F-measure or F_p-measure as proposed in Section 4.2.4. Therefore, we have to estimate the tuning parameter δ.

The estimation procedure that we propose here bears some resemblance with conventional cross-validation. Consider a small subset of experiments $\mathcal{D} = \{\underline{\boldsymbol{x}}^{(1)}, \cdots, \underline{\boldsymbol{x}}^{(S)}\}$, $\mathcal{D} \subset \mathcal{X}$, where the ground truths $\{\mathcal{C}^{(1)}, \cdots, \mathcal{C}^{(S)}\}$ are available. Define the leave-one-out subsets $\mathcal{D}^{-s} = \mathcal{D}/\underline{\boldsymbol{x}}^{(s)}$, $s = 1, \cdots, S$. For a given s, $\mathcal{D}^{-s}$ is seen as a training data set and $\underline{\boldsymbol{x}}^{(s)}$ as a test data set, which is used to estimate the prediction error. In the conventional cross-validation procedure, all samples in $\mathcal{D}^{-s}$ constitute a unique sample. For every δ, this big sample is used to estimate a unique $\hat{\phi}(\delta)$, which is therefore common to all the samples in $\mathcal{D}^{-s}$. This estimate allows to define a partitioning of the space $\mathbb{R}^m$ with respect to the decision rule (4.1). In particular, this partitioning is tested on $\underline{\boldsymbol{x}}^{(s)}$ and the prediction error $\mathrm{ERR}(s, \hat{\phi}(\delta))$ is computed. Finally, the mean error $\overline{\mathrm{ERR}}(\hat{\phi}(\delta))$ is used to estimated δ.

In our context, the samples in $\mathcal{D}^{-s}$ cannot be gathered into a unique sample, since ϕ depends on each sample. In other words, we cannot estimate a unique partitioning of $\mathbb{R}^m$ for every δ. In fact, due to the availability of $\mathcal{C}^{(s)}$, our clustering task on $\mathcal{D}$ is a supervised problem, therefore, the prediction error can be revised. We propose the following learning/test procedure.

The clustering sets $\{\mathcal{Z}^{(1)}, \cdots, \mathcal{Z}^{(S)}\}$ are obtained independently on each data set $\boldsymbol{x}^{(s)}$ for every $\delta \in \Delta$, where Δ is a given finite set of values of δ. We write $\mathcal{Z}^{(s)}[\delta]$ these clusters. Instead to estimate the best partitioning of $\mathbb{R}^m$ for every δ, we directly estimate δ on $\mathcal{D}^{-s}$. The estimation of δ from the training set is given by maximization of the F_p-measure mean:

$$\hat{\delta}^{\mathcal{D}^{-s}} = \arg\max_{\delta \in \Delta} \frac{1}{S-1} \sum_{r \in \{1, \cdots, S\}/s} F_p(\mathcal{C}^{(r)}, \mathcal{Z}^{(r)}[\delta]),$$

which is the one in Δ performing the best mean similarity on the training set. The loss

function quantifying the mismatching on the test set $\underline{\boldsymbol{x}}^{(s)}$ is defined as

$$\text{loss}(s,\delta) = \; F(\mathcal{C}^{(s)}, \mathcal{Z}^{(s)}[\hat{\delta}^{\mathcal{D}^{-s}}]) \quad - \quad F_p(\mathcal{C}^{(s)}, \mathcal{Z}^{(s)}[\delta]),$$

where the first term is the best similarity on the test set $\underline{\boldsymbol{x}}^{(s)}$ achieved with $\hat{\delta}^{\mathcal{D}^{-s}}$, and the second term is the similarity with any δ, taking count of mismatching penalization. Then the estimate of tuning parameter δ is given by minimization of mean error:

$$\hat{\delta} = \arg\min_{\delta \in \Delta} \frac{1}{S} \sum_{s=1}^{S} \text{loss}(s,\delta). \tag{4.10}$$

4.3 Application on flow cytometry analysis

4.3.1 Background

Concrete to the MI project, a 8-color "T cell" panel was designed (m=8) to identify T cell subpopulations. For ones who are not familiar with flow cytometry, we would first like to recall what is a "panel". A panel is a combination of m antibodies selected to targeting certain cell types of interest, where m depends on the flow cytometer that are used in the experiment (for more details, see page 6 Section 1.2.3). Here in this panel, we combine CD3, CD4, CD8β, CD8α, CD27, CD45RA, CCR7 and HLA-DR to classify T cell subpopulations (Figure 4.3A). The CD3 marker is utilized to identify all T cells (CD3$^+$). Within CD3$^+$, two big groups of cells CD4$^+$ T cells and CD8β^+ T cells can be identified based on the expression of CD4 and CD8β. For each of the two groups, naive (T_{naive}), central memory (T_{CM}), effector memory (T_{EM}) and EMRA$^+$ (T_{EMRA}) T cells subsets can be characterized, utilizing the relative expression levels of CD27, CD45RA and CCR7 [Hasan et al., 2015]. By including CD8α, we were able to distinguish CD4$^+$CD8α^+ T cells. The early thymocytes that express neither CD4 nor CD8 are classed as double-negative (CD4$^-$CD8β^-) cells, and the next maturational stage where cells express both CD4 and CD8 are classed as double-positive (CD4$^-$CD8β^-) cells. The surface expression of HLA-DR is utilized to quantify the activation status of T cells. All the eight markers and related cell types are listed in Figure 4.3A and Figure 4.3B, respectively.

The related cell types in this panel can be represented by a dependency tree as shown in Figure 4.3C. Note that not every marker defines directly a cell type, for example, CCR7 is highly correlated with CD27, but in this study we use CD27 for identification of naive and central memory cells, and CCR7 only for control; and HLA-DR is a marker indicating the activation level. Thus, these two markers are not present in Figure 4.3B and Figure 4.3C.

The object of automatic analysis is to identify all these cell types. As we model this problem as a clustering problem, we only need to identify the cell types that are represented as leaves of the tree, then the cell types on an upper level of the hierarchy can be obtained directly. Therefore, we consider eleven cell types of interest (Figure 4.3B), which correspond eleven leaves in Figure 4.3C.

Note that $C = 11$ is the number of cell types of interest, but not the number of clusters K. For the same reasons discussed in Chapter 2, and similarly with FlowGM and many other mixture model based automatic flow cytometry analysis methods, we allow multiple gaussian clusters to describe one cell type. Therefore, K is larger than eleven, and to be determined later. Then the automatic cell population identification consists of partitioning cells presented in a blood sample into K clusters based on the measurements of the eight markers, and assigning one of the eleven cell types to each cluster.

The differentiation of naive T cells into effector and memory subsets represents one of the most fundamental facets of T cell mediated immunity [Appay et al., 2008], but the detailed characterization of the phenotype and function of distinct T cell sub-populations in human remains lacking of global consensus [Appay et al., 2008]. There are two reasons for this. First, the T cell population can be divided into distinct subsets based on their expression of diverse cell surface receptors, including the receptors involved in activation (e.g., CD45RA), costimulation(e.g., CD27) and some chemokine receptors (e.g. CCR7). Second, the expression levels of these receptors often cannot be defined simply as "positive/negative" or "high/low", but in a continuous fashion. Therefore, no clear boundaries exist between distinct subsets, and the existence of overlapping groups and regions of transition between cell types make the manual analysis, even for individual samples, difficult (Figure 4.4). When considering large cohorts, the phenotypes of distinct T cell sub-populations vary a lot in terms of expression levels of diverse receptors and population sizes with regard to their age, gender, infection history, activation of antigenic stimulation and many other aspects, all of which represent also major challenges for automated approaches.

4.3.2 Results

As flow cytometry analysis is intricate and difficult to understand without some basic knowledge of immune cell populations, we generate here a simulated data set, which is a simplification of the real data. In this section, we will first discuss the results on the simulation and then a subset of MI project T cell panel data.

A

	Fluorochrome	Marker
X_1	eF450	CD3
X_2	APC-H7	CD4
X_3	PE-Cy7	CD8β
X_4	PE	CD8α
X_5	PerCP-Ef710	CD27
X_6	FITC	CD45RA
X_7	APC	CCR7
X_8	V500	HLA-DR

B

Index	Cell type	Immunophenotype
	CD4+ T cells	CD3+CD4
c=1	CD4+ naïve	CD3+CD4+CD27+CD45RA+
c=2	CD4+ CM	CD3+CD4+CD27+CD45RA-
c=3	CD4+ EM	CD3+CD4+CD27-CD45RA-
c=4	CD4+ EMRA	CD3+CD4+CD27-CD45RA+
	CD8β+ T cells	CD3+CD8β+
c=5	CD8β+ naive	CD3+CD8β+CD27+CD45RA+
c=6	CD8β+ CM	CD3+CD8β+CD27+CD45RA-
c=7	CD8β+ EM	CD3+CD8β+CD27-CD45RA-
c=8	CD8β+ EMRA	CD3+CD8β+CD27-CD45RA+
c=9	CD4+CD8β+	CD3+CD4+CD8β+
c=10	CD4+CD8α+	CD3+CD4+CD8α+
c=11	CD4-CD8β-	CD3+CD4-CD8β-

C

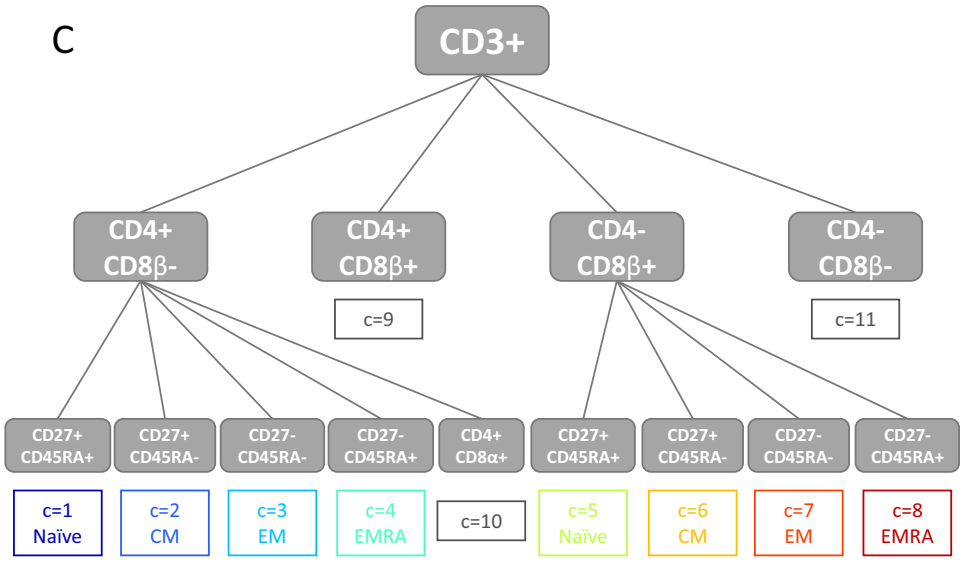

Figure 4.3: Markers and cell types of interest in the T cell panel. (A) shows the eight fluorochromes and markers. (B) shows the eleven target cell types and their immunophenotypes. (C) shows a partial panel tree with eleven cell types on the leaves for the MI T cell panel.

Simulated data

The dimension of the simulated data is set to four, with the markers CD4, CD8β, CD27 and CD45RA, which makes a simplified T cell panel (Figure 4.5). We use $C = 8$ latent classes, presenting CD4$^+$ and CD8β^+ naive, central memory, effector memory and

EMRA$^+$ populations. In other words, we select X_2, X_3, X_5 and X_6 from the real MI T cell panel, focus on the 8 first cell types, $c = 1, \cdots, 8$, and we omit the rest. Two flow cytometry samples, one reference sample and another special sample (referred as "donor 1") were simulated.

A synthetic example is shown in Figure 4.6. The eight cell populations are presented in eight colors with the same color code as in Figure 4.5. For both two samples, two groups of cells could be separated in CD4/CD8 projection, each group could be further separated into four sub-groups using CD27/CD45RA, which are CD27$^+$CD45RA$^+$ naive cells, CD27$^+$CD45R$^-$ central memory cells, CD27$^-$CD45RA$^-$ effector memory cells and CD27$^-$CD45RA$^+$ effector memory cells expressing CD45RA, which makes in total 8 cell populations. When we compare donor 1 with the reference sample, the CD4$^+$ populations express higher CD45RA, and the CD8$^+$ populations have a lower expression level on both CD27 and CD45RA, but as the subsets are identified with the relative expression levels, they could stilled be well separated visually. But when applying FlowGM [Chen et al., 2015], the standard gaussian mixture model on donor 1 data with the reference settings, a proportion of CD4$^+$ T naive cells are misclassified as CD8β^+ T naive cells, and within CD8β^+ population, the naive group is wrongly labeled as central memory, central memory is wrongly labeled as effector memory, and effector memory cells are included in EMRA$^+$.

By introducing a prior for cluster means, more clearly, by indicating a prior belief of the location and restricting the spread direction and size of each cluster centroid, FlowGMP is applied to donor 1 data. We allow the reference clusters spread larger on CD27 and CD45RA dimensions than on CD4 and CD8β, in order to avoid the misclassification of CD4$^+$ and CD8β^+ populations. For the identification of sub-populations, the prior belief of the location help the algorithm to capture correspondent population.

The outputs of FlowGMP on simulated donor 1 data after the first, fifth and ninth iteration are shown in Figure 4.7. Points are colored according to their posterior likelihood, the ellipsoid reflects the cluster shape, in a size of three times the standard deviation, covering over 90% of the total probability mass. We could see that the algorithm is very efficient, all the eight populations are correctly identified with only one single misclassification. It is even invisible from the figure, which is a CD8β^+ central memory cell (light red) labels as a CD8β^+ EMRA$^+$ cell (deep red), and is marked in the right bottom of the figure.

LabEX MI T cell panel data

Overview As shown in Figure 4.3, the eight-color cytometry panel targeting distinct T cell sub-populations across 100 healthy individuals from different age groups and genders was designed to characterize $CD4^+$ and $CD8\beta^+$ T_{naive}, T_{CM}, T_{EM} and T_{EMRA} subsets, as well as $CD4^+CD8\beta^+$, $CD4^+CD8\alpha^+$, and $CD4^-CD8\beta^-$ T cells.

The overall operation of the FlowGMP workflow is shown in Figure 4.8. The first phase is the same with FlowGM, where method parameters are calibrated on selected reference samples. In a second phase, the distributions of the centers of C cell types are esitmated on selected training samples, and then considered as the prior distribution of all corresponding clusters. In a third phase, all other cohort samples are processed on the basis of the calibrated parameters. Similar with FlowGM, this workflow as designed to minimize the manual effort in the cohort treatment phase.

Reference clustering First, we utilized the FCS data from the reference donor to build a reference clustering. The same pre-processing steps as described earlier for FlowGM [Chen et al., 2015] were taken. Here we repeat briefly these steps, which are identical for the reference donor and in the following for all other donors. After the elimination of doublets based on Forward Scatter (FSC) and Side Scatter (SSC), pre-filtering of the T cell panel was based on a two-component, one-dimensional GMM that utilized the measurement of the CD3 marker. Thresholds were automatically set at the 95th percentiles of $CD3^-$ populations. Next, we estimated the number of clusters K with the BIC. The optimal fit is $K = 32$, each of the 32 cluster mean, covariance matrix and proportion were determined. We then manually assigned a cell type to each cluster with the help of experts. Of the 32 clusters, 24 were of interest and thus mapped to eleven cell populations, called meta-clusters. Other 8 clusters were mapped to a meta-cluster 0 (dump). Explictly, $k = 1, \cdots, 32$ are mapped to $mk = 0, 1, \cdots, 11$ (mk for meta-cluster index), which correspond to dump and eleven cell types. The 24 cluster centroids are represented as a heatmap, which is shown in Figure 4.9, with their manually assigned cell type indicated on the right.

Choice of the prior The idea for the choice of the prior is to use the empirical distribution of each cell population of interest from a small subset of samples as a prior distribution of the corresponding clusters. We randomly selected a subset of S samples ($S = 10$) from LabEX MI T cell panel data, where FlowGM failed, to set the prior distribution for applying our proposed method FlowGMP. There are around a third to half

of those cohort samples, and FlowGM failed because the cluster positions, shapes or proportions are different with the reference. Therefore, in order to quantify the variation, we selected the "outliers" rather than the samples where FlowGM worked. Manual analysis was carefully performed by an expert, $C = 11$ cell populations were manually identified and visually confirmed in every sample of the subset, which is then considered as a "ground truth". The coordinates of all gates were recorded, and cell labels and important statistics, including population counts and mean fluorescent intensities (MFIs) were obtained. The cells that were not included in any class by manual gating, for example, outliers and biologically irrelevant populations, were labeled as $c = 0$, while other eleven interesting cell populations were labeled from $c = 1$ to $c = 11$, presenting eleven cell types as shown in Figure 4.3B. The manually gated $CD4^+$ and $CD8\beta^+$ sub-populations for each of the ten sample from the training set are shown in Figure 4.10, where each ellipsis corresponds to one of the ten samples, and the same colored ones represent the same cell type. With the manual identification for the ten selected samples, we obtain the empirical distribution of each cell population of interest.

Note that the manual gating yields the estimated distribution of the C class centers, but not that of the K cluster centers. The merging of K clusters into eleven meta-clusters allows one to one correspondence between meta-clusters and cell types.

For each cell population obtained from manual analysis, the mean and covariance matrix $\theta_c^{(d)} = \{\boldsymbol{\mu}_c^{(d)}, V_c^{(d)}\}$, $c = \{1, \cdots, 11\}$ were computed for each sample d, the empirical distribution of $\boldsymbol{\mu}_c$ was used as a prior distribution for each mapped cluster; all clusters that represented the same population share the same prior. For example, in Figure 4.9, the last two lines (in green) correspond to two of the 32 clusters, which are manually mapped to $CD4^+$ naive cells, then the empirical distribution of $CD4^+$ naive cell centers from the ten samples are used as the prior distribution for both of these two clusters. Similarly, the seven clusters just above (in orange) that are mapped to $CD4^+$ CM cells share the same prior. Note that the initial position, shape and size of each cluster are different, but as the clusters corresponding to a same cell type have the same prior, which means that they are iteratively built around τ_c, this constraint allows the cell type assignment and thus the alignment.

For other unmapped clusters, which are not shown in Figure 4.9, as we don't know which cell type they represent, or the cell types are not of our interest (dump), these populations are impossible to access, let alone their empirical distributions. Therefore the prior parameters are set to $\boldsymbol{\tau}_k = \boldsymbol{\mu}_k^{(0)}, \Gamma_k = V_k^{(0)}/100$, which come from the reference donor.

Estimation of tuning parameter using cross-validation The role of the tuning parameter δ is to balance likelihood and penalization terms, where a small value of δ indicates strong penalization. The F_p-measure between FlowGMP resulted meta-clusters and manual gating resulted classes for each donor under different values of δ are shown in Figure 4.11, and the estimate $\hat{\delta}$, determined by minimization of mean error (4.10), is shown in Figure 4.12. For $\Delta = \{10^{-3}, 10^{-2}, \cdots, 10^{5}\}$, $\hat{\delta} = 100$ is obtained.

Comparison of FlowGMP with FlowGM and manual gating analysis To evaluate the improvement of FlowGMP, we compare directly FlowGM and FlowGMP against manual analysis. The F_p-measures calculated for each donor in the training set were shown in Table 4.1. The values indicated that, overall, FlowGMP performed better than FlowGM.

Donor ID	FlowGM, F_p	FlowGMP, F_p
Donor 23	0.7735	0.8768
Donor 34	0.8344	0.8746
Donor 44	0.8828	0.8919
Donor 45	0.8625	0.8858
Donor 428	0.7978	0.8307
Donor 429	0.7708	0.8449
Donor 482	0.8262	0.8296
Donor 629	0.8045	0.8404
Donor 665	0.8250	0.8165
Donor D	0.8732	0.8795

Table 4.1: F_p-measure for FlowGM and FlowGMP approaches. FlowGMP results a high-quality clustering (improved F_p-measure compared to FlowGM).

The counts resulting from FlowGM and FlowGMP analysis were compared with manual gating analysis. Figure 4.13 shows that, in addition to the F_p-measure, the overall agreement of population counts is also improved for all cell types. The average correlation increases from 0.728 to 0.863, especially for CD8β+ memory cells, which were considered as the most big challenge of this data set.

Finally, we applied FlowGMP to 115 donors from MI cohort, using estimated tuning parameter from the training set. Absolute cell counts were compared with manual analysis. Again, the results were highly concordant with average correlation $r = 0.940$, compared to FlowGM where the average correlation was $r = 0.783$.

In terms of running time, FlowGMP is more efficient. Briefly, the computation required 8 minutes for 115 cohort donors on a standard laptop PC, while FlowGM required 24 minutes on the same data set and the same machine.

4.4 Conclusion

The FlowGMP approach was developed out of the original FlowGM approach to address the need for robust and high-quality analysis for the MI T cell panel study. Our evaluation study on our training data set has shown that FlowGMP has produced better identification of different cell populations. The comprehensive evaluation study on the cohort data set has shown that FlowGMP has produced a user-validated results. Using a prior on the cluster position constrains the local optimization in the EM algorithm, reduced the mislabeling of clusters, and made the automated analysis more efficient.

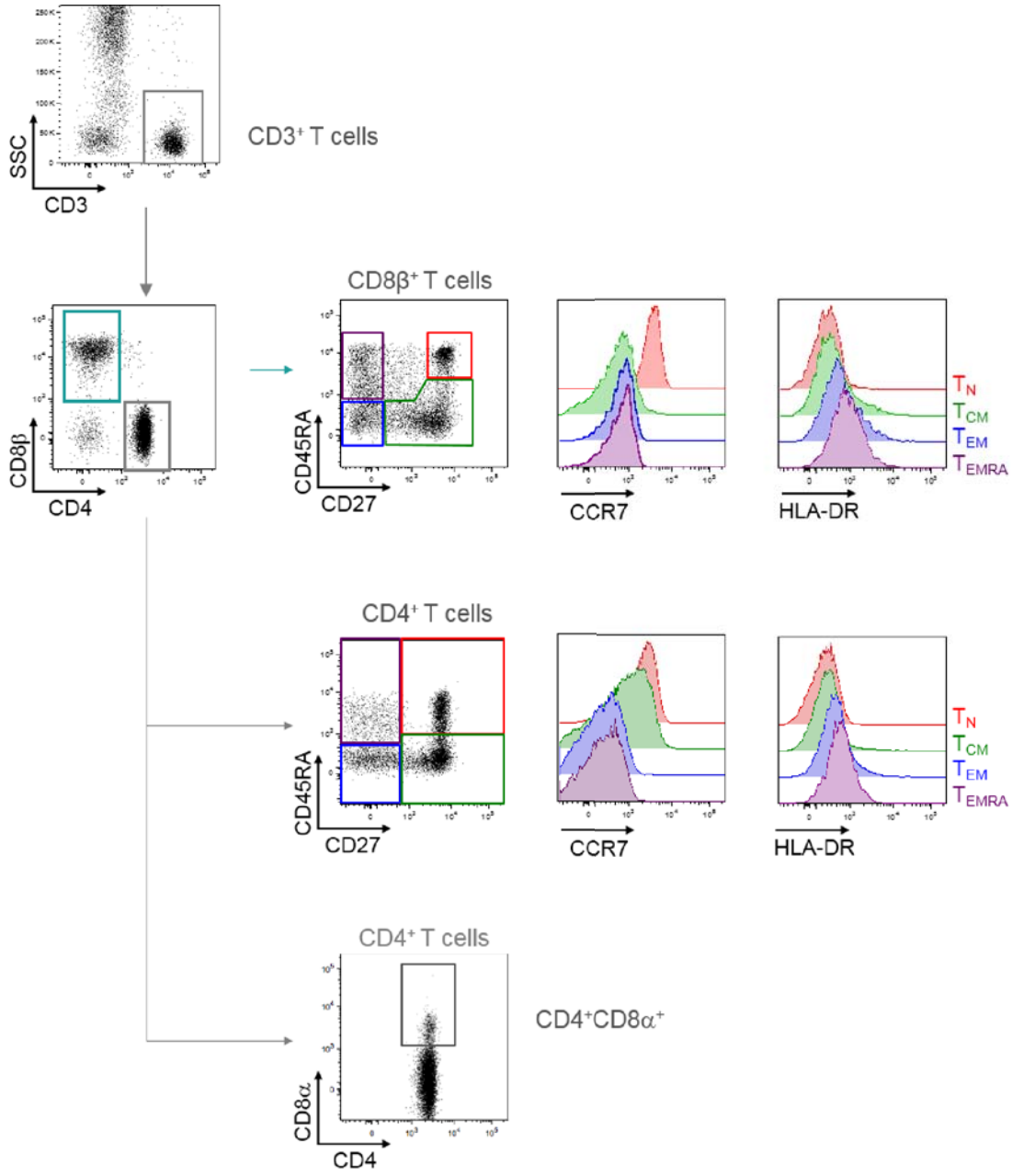

Figure 4.4: Manual gating strategy for the MI T cell panel. $CD3^+$ cells were identified. Subsequent phenotypic analysis identified $CD4^+$ and $CD8\beta^+$ T cell subsets based on CD4 and $CD8\beta$ expression, respectively. T_{naive} ($CD27^+CD45RA^+$ cells, and red shaded histograms), T_{CM} ($CD27^+CD45RA^-$ cells, and green shaded histograms), T_{EM} ($CD27^-CD45RA^-$ cells, and blue shaded histograms) and T_{EMRA} ($CD27^-CD45RA^+$ cells, and violet shaded histograms) were based on surface expression of CD27 and CD45RA. On each of the eight respective populations, CCR7 and HLA-DR expression was analyzed and plotted as a histogram. The $CD4^+CD8\beta^-$ cells expressing $CD8\alpha$ were identified (gray gate).

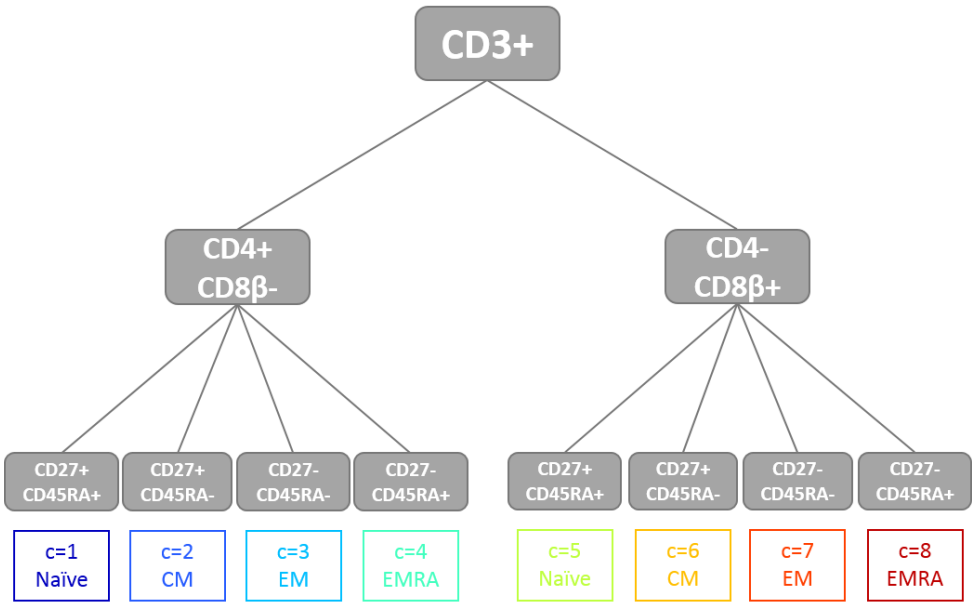

Figure 4.5: A partial panel tree with cell types on the leaves for the simulated data.

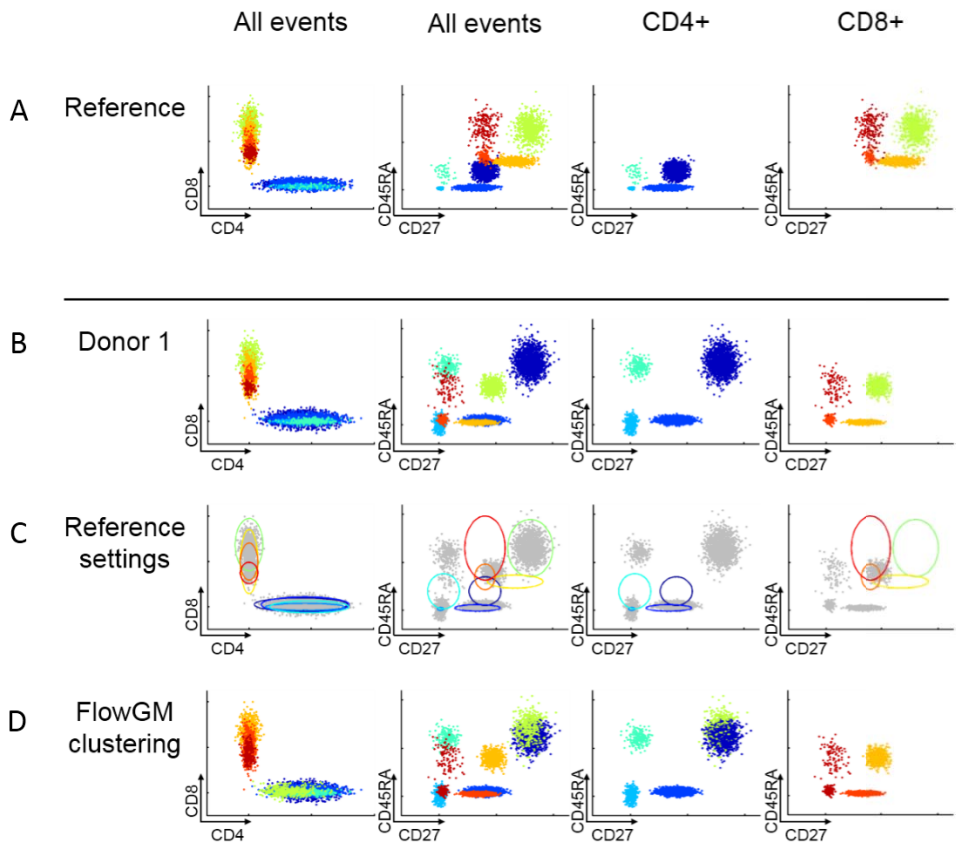

Figure 4.6: Simulated example showing the difficulties of cell population identification using the classical approach. Columns 1 and 2 show all events in CD4/CD8β and CD27/CD45RA projections, column 2 is split into column 3 and column 4 according to $CD4^+$ and $CD8\beta^+$, respectively, as it is indicated at the top of the figure. (A) shows the reference sample. Two groups of cells could be separated in CD4/CD8 projection, depicted as group "blue" and group "red-green". Each group could be further separated into four subsets using CD27/CD45RA, which are $CD27^+CD45RA^+$ naive cells, $CD27^+CD45RA^-$ central memory cells, $CD27^-CD45RA^-$ effector memory cells and $CD27^-CD45RA^+$ effector memory cells expressing RA. (B) shows an independent simulated sample, the colors indicate the "ground truth". (C) shows the simulated data and reference cluster centroids. (D) shows the use of reference clustering to classify events in the simulated sample. Results are described in the text. FlowGM clustered some $CD4^+$ naive cells as $CD8^+$ naive cells, and $CD8^+$ naive as CM, CM as EM, and EM cells are included in $EMRA^+$.

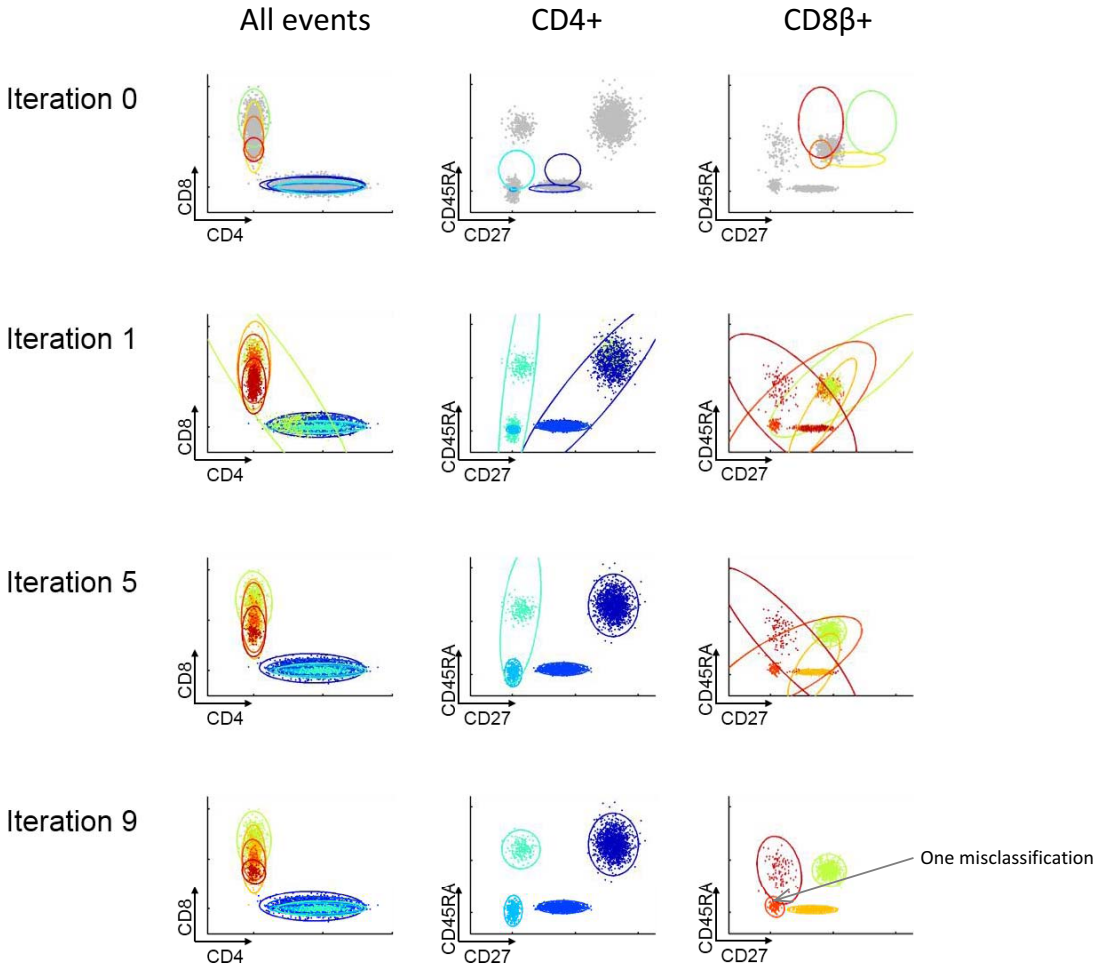

Figure 4.7: The results of FlowGMP on simulated donor 1 data after the first, fifth and ninth iteration. The first row (Iteration 0) corresponds to the first row in Figure 4.6. Points are colored according to their posterior likelihood, the ellipsoid reflects cluster shape, in a size of three times of the standard deviation, covering over 90% of the total probability mass.

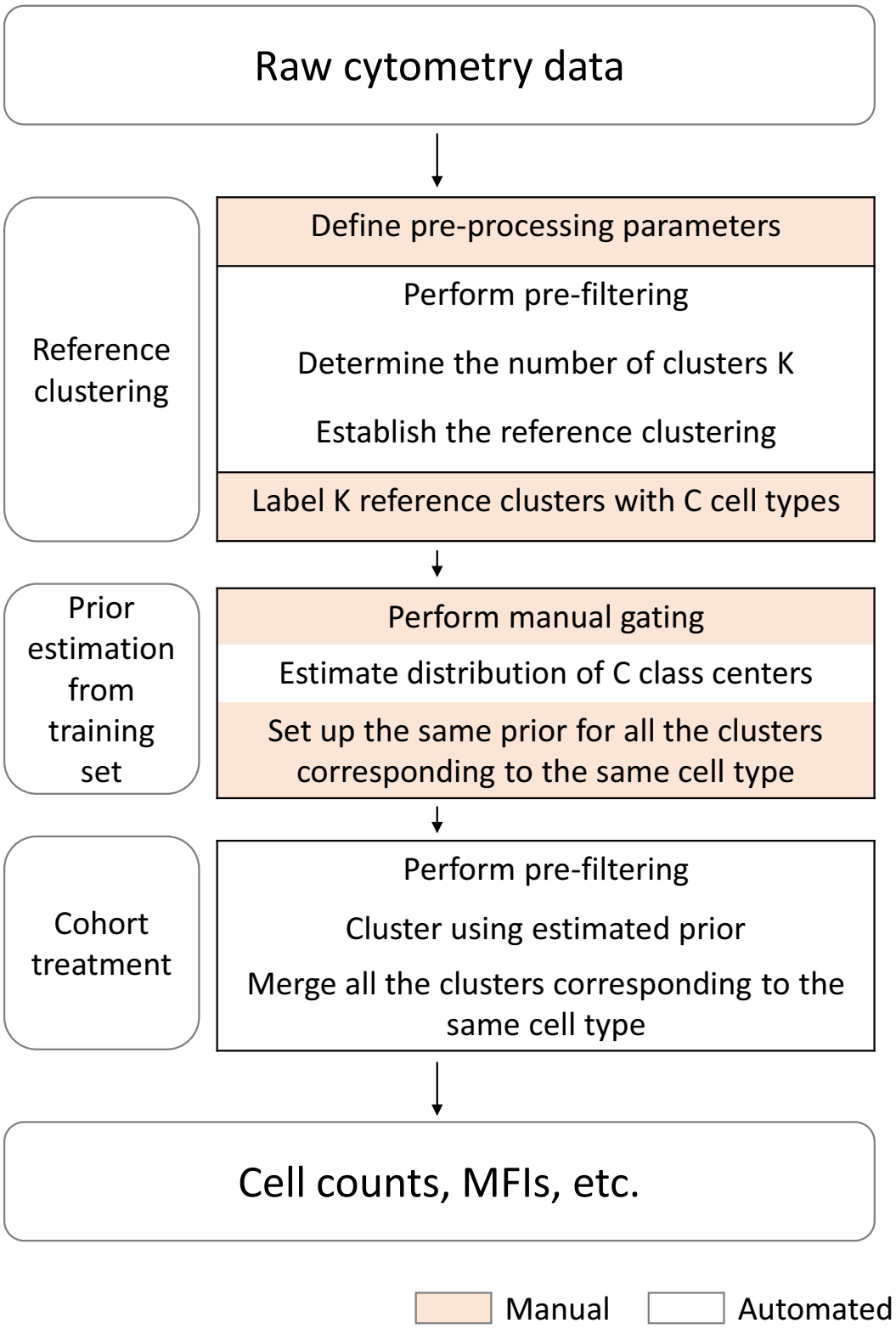

Figure 4.8: Overview of the FlowGMP workflow. Three phases can be distinguished. The manual steps are colored in light red, while the automated steps are in white.

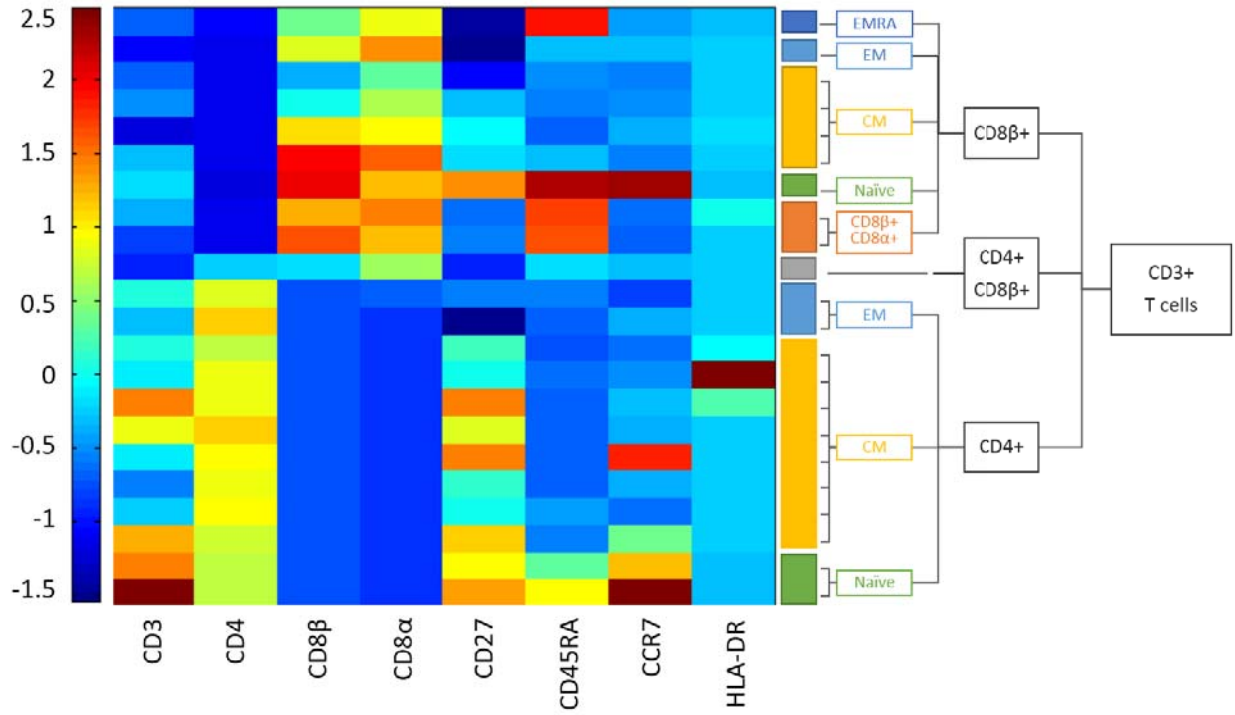

Figure 4.9: User-based aggregation of clusters into meta-clusters for immune cell type characterization with cluster centroid heat map (normalized coordinates). Each line corresponds to one cluster, and the manually assigned cell types are indicated on the right.

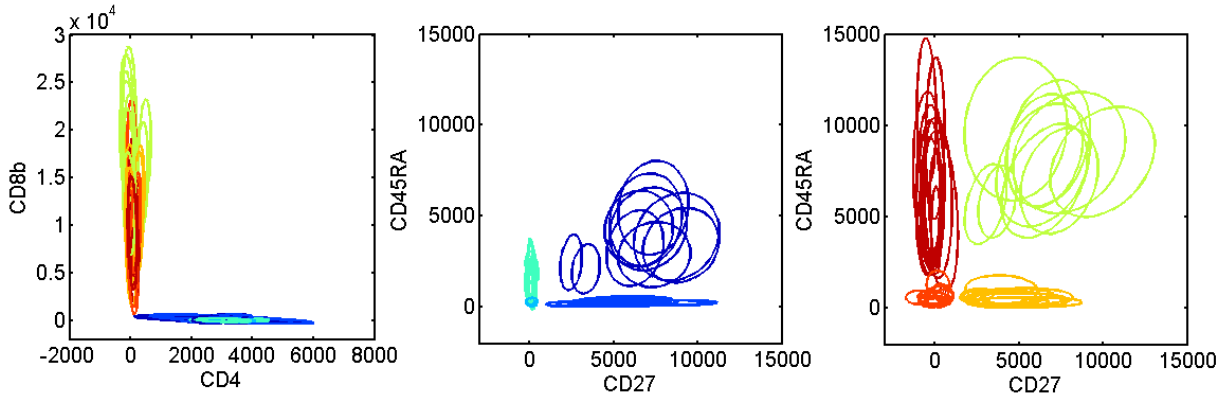

Figure 4.10: Manual gated populations for 10 samples of training set. Cell populations are colored the same as for the simulation. For every population, each ellipsis correspond to one of the 10 samples.

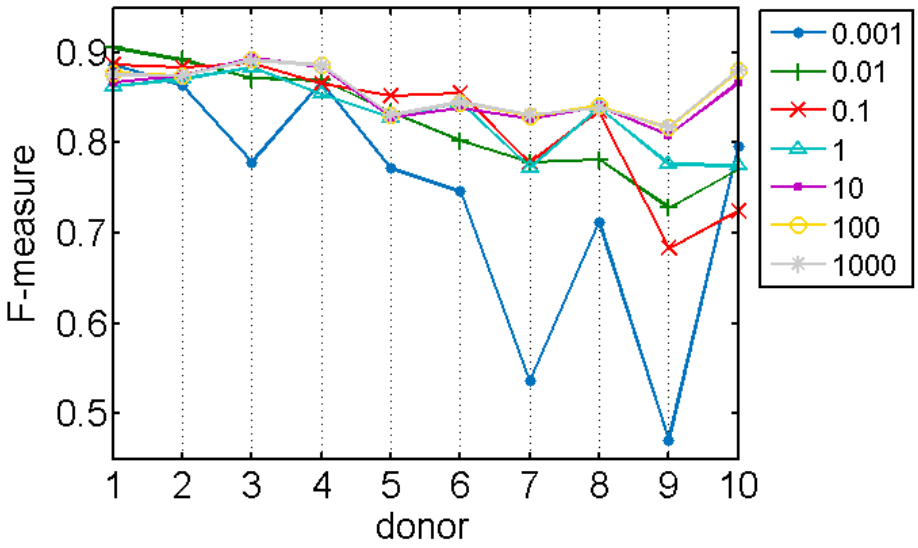

Figure 4.11: FlowGMP was applied on each sample of our training set with different values of δ. The resulting clustering assignment was compared with manual labeling. The relative F_p-measure is shown.

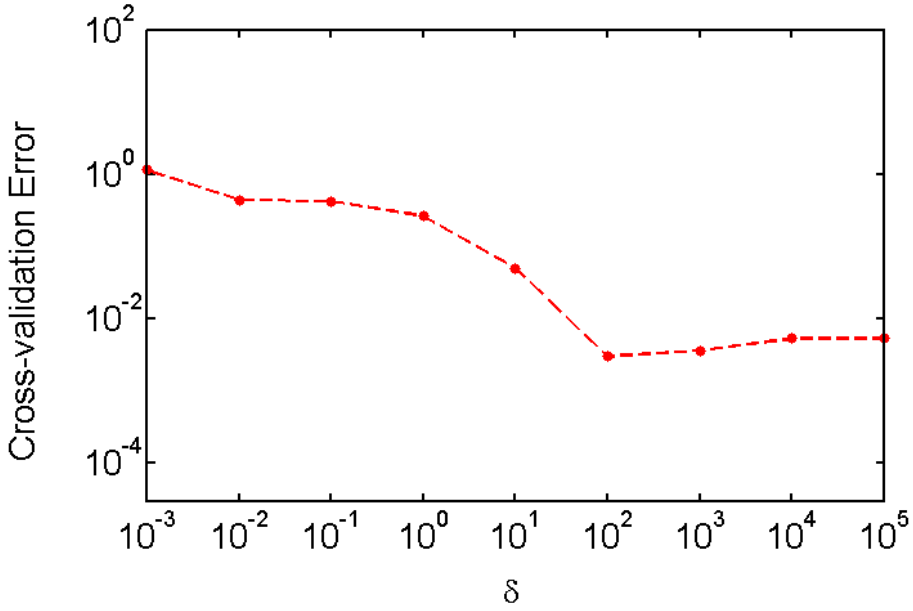

Figure 4.12: Cross-validation error for different δ. The estimate was obtained by minimization of cross-validation error, then the analysis was applied to cohort samples using $\hat{\delta} = 100$.

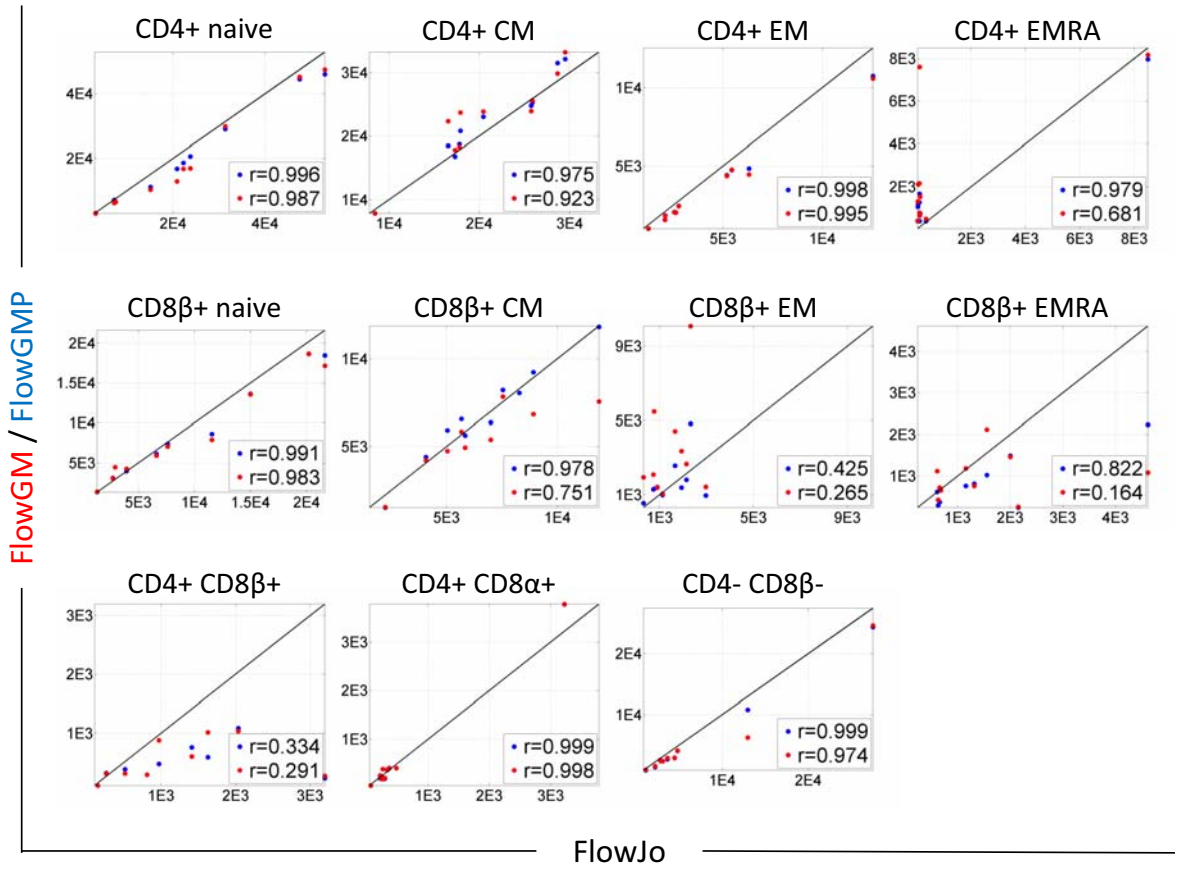

Figure 4.13: Comparison of manually counting with FlowGM and FlowGMP analysis on ten donors. In this T cell panel, we have identified eleven T cell subpopulations, for each of them, the X-axis is the counts from FlowJo, and the Y-axis is the counts from FlowGM (red) or FlowGMP (blue). The counts obtained by FlowGMP (average correlation 0.863) agrees better than FlowGM (average correlation 0.728) with manual gated data on these ten donors.

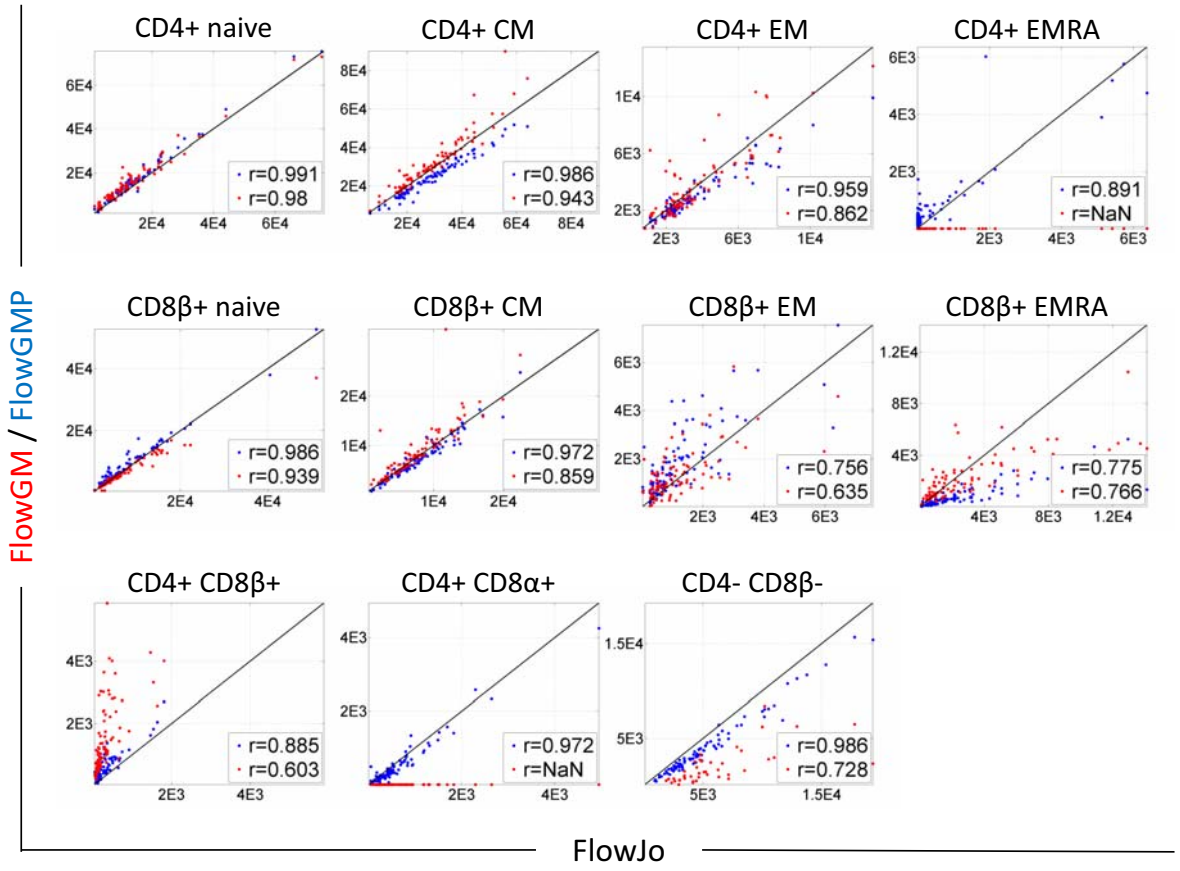

Figure 4.14: Comparison of manually counting with FlowGM and FlowGMP analysis on $D = 115$ cohort donors. In this T cell panel, we have identified 11 T cell subpopulations, for each of them, the X-axis is the counts from FlowJo, and the Y-axis is the counts from FlowGM (red) or FlowGMP (blue). The counts obtained by FlowGMP (average correlation 0.923) agrees better than FlowGM (average correlation 0.783) with FlowJo counts on 115 cohort donor.

Appendix

We know that, for standard EM algorithm, $L(\phi^{(s+1)}) \geq L(\phi^{(s)})$, where $L(\phi) = p_\phi(\underline{\boldsymbol{x}})$ is the likelihood, and s is the index of EM iteration. For our case, given $L_p(\phi) = \prod_{i=1}^{N} \left(p_\phi\left(\boldsymbol{x}_i\right) p(\phi)\right) = p_\phi(\underline{\boldsymbol{x}}) p(\phi)^N$, we would like to demonstrate that $L_p(\phi^{(s+1)}) \geq L_p(\phi^{(s)})$ holds true as well, which means that the likelihood increases at each iteration of the proposed EM algorithm.

Proof By noting $\hat{\phi} = \phi^{(s+1)}$, $\phi' = \phi^{(s)}$, $\hat{p} = p(\hat{\phi})^n$, $p' = p(\phi')^n$ and using Jensen's inequality, i.e. $f\left(\mathbb{E}(z)\right) \geq \mathbb{E}\left(f(z)\right)$ if f is a concave function, we thus obtain

$$
\begin{aligned}
\log \frac{L_p(\phi^{(s+1)})}{L_p(\phi^{(s)})} = \log \frac{p_{\hat{\phi}}(\underline{\boldsymbol{x}})\hat{p}}{p_{\phi'}(x)p'} &= \log \left[\frac{1}{p_{\phi'}(\underline{\boldsymbol{x}})p'} \sum_z p_{\hat{\phi}}(z, \underline{\boldsymbol{x}})\hat{p} \right] \\
&= \log \left[\sum_z \frac{p_{\phi'}(z|\underline{\boldsymbol{x}})}{p_{\phi'}(z, \underline{\boldsymbol{x}})p'} p_{\hat{\phi}}(z, \underline{\boldsymbol{x}})\hat{p} \right] \\
&= \log \mathbb{E}\left[\frac{p_{\hat{\phi}}(Z, \underline{\boldsymbol{x}})\hat{p}}{p_{\phi'}(Z, \underline{\boldsymbol{x}})p'} \middle| \phi' \right] \\
&\geq \mathbb{E}\left[\log \frac{p_{\hat{\phi}}(Z, \underline{\boldsymbol{x}})\hat{p}}{p_{\phi'}(Z, \underline{\boldsymbol{x}})p'} \middle| \phi' \right] \\
&= \mathbb{E}\left[\log p_{\hat{\phi}}(Z, \underline{\boldsymbol{x}})\hat{p} | \phi' \right] - \mathbb{E}\left[\log p_{\phi'}(Z, \underline{\boldsymbol{x}})p' | \phi' \right] \\
&= Q_p(\hat{\phi}, \phi') - Q_p(\phi', \phi')
\end{aligned}
$$

As we maximize Q_p in each step with $\hat{\phi} = \arg\max_{\hat{\phi}} Q_p(\phi, \phi')$, then $\log \frac{L_p(\phi^{(s+1)})}{L_p(\phi^{(s)})} \geq 0$, $L_p(\phi^{(s+1)}) \geq L_p(\phi^{(s)})$.

Chapter 5

Conclusions and future work

5.1 Discussion

This thesis presents two novel methods for flow cytometry data analysis across a large number of samples and large number of cell types. In Chapter 2, we described a workflow, named FlowGM, for automated cell population identification. We were able to reliably quantify 24 cell types across 115 MI samples and 4 panels. We showed intuitively and simply that our performance is on par with, or exceeding the quality of manual gating on a subset of 115 donors, and concluded that it is amenable to whole cohort studies.

In Chapter 3, I applied this method to all cohort data, and demonstrated the global agreement of the results with manual analysis in terms of cell counts. Correlating analysis results with donor age allowed other interesting conclusions. We discussed the effect of age on circulating immune cell populations, highlighting more particular discoveries of aging effects on newly identified $CD14^{lo}CD16^{hi}$ monocyte subpopulation and HLA-DR^{hi} activated cDC1 population using our FlowGM method. Correlations of other factors with gender, CMV infection, smoking history and metabolism score on all identified cell populations are also analyzed in exactly the same manner. We performed as well as "Visit 2" control analysis, cross-panel analysis and outlier analysis. Our results suggest that the systematic, high-quality analysis of cell counts, which our methodology enables, can be expected to create numerous opportunities for the discovery of new correlations in biological data.

In chapter 4, I dealt with a specific situation of cell population identification in flow cytometry analysis, where the presence of technical and biological variation made the automatic alignment of populations challenging. I developed a novel method FlowGMP, which was

able to identify cell populations across a large number of samples despite the presence of variability. It was demonstrated that FlowGMP reduced the mislabeling of clusters, and improved the efficiency of automated analysis on a training set of MI samples. I then applied this method to 115 cohort donors, and showed the improved agreement with manual analysis.

5.2 Future work

This thesis was mostly focused on exploratory analysis of FCM data in the context of the LabEX MI project. However, the pipeline presented here can be modified for other applications. During the course of my thesis, I applied this pipeline to other datasets or other research projects, for example, for studying CMV infections, for studying SpA patients before and after treatment, for identifying rare ILC populations etc. My collaboration experience with biologists and immunologists made me believe that it is indispensable to build a user-friendly interface of our proposed FCM analysis pipeline, and it is essential to keep working closely with the new technologies, to access complex datasets and problems, which permits continuously technology/data-driven improvement of computational tools.

Our experience of analyzing huge amounts of FCM data in various real biological problems made us believe that it will be interesting to be able to simultaneously characterize the dependencies between markers for a given population and identify the population with the consideration of different marker dependencies. This type of study could potentially provide some idea for integrating FCS data from different panels and help in panel design for better immunophenotyping.

Bibliography

[Adinis, 2014] Adinis (2009–2014). ADICyt, Adinis, Ltd., SK. `http://www.adinis.sk/en/`.

[Aghaeepour and Brinkman, 2014] Aghaeepour, N. and Brinkman, R. (2014). Computational analysis of high-dimensional flow cytometric data for diagnosis and discovery. In *High-Dimensional Single Cell Analysis*, pages 159–175. Springer.

[Aghaeepour et al., 2013a] Aghaeepour, N., Finak, G., Consortium, F., Consortium, D., Hoos, H., Mosmann, T. R., Brinkman, R., Gottardo, R., and Scheuermann, R. H. (2013a). Critical assessment of automated flow cytometry data analysis techniques. *Nature methods*, 10(3):228–238.

[Aghaeepour et al., 2013b] Aghaeepour, N., Finak, G., Hoos, H., Mosmann, T. R., Brinkman, R., Gottardo, R., Scheuermann, R. H., Consortium, F., Consortium, D., et al. (2013b). Critical assessment of automated flow cytometry data analysis techniques. *Nature methods*, 10(3):228–238.

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

[Appay et al., 2008] Appay, V., van Lier, R. A. W., Sallusto, F., and Roederer, M. (2008). Phenotype and function of human t lymphocyte subsets: Consensus and issues. *Cytometry Part A*.

[Bashashati and Brinkman, 2009] Bashashati, A. and Brinkman, R. R. (2009). A survey of flow cytometry data analysis methods. *Advances in Bioinformatics*.

[Bilmes et al., 1998] Bilmes, J. A. et al. (1998). A gentle tutorial of the em algorithm and its application to parameter estimation for gaussian mixture and hidden markov models. *International Computer Science Institute*, 4(510):126.

[Biosciences, 2000] Biosciences, B. (2000). Introduction to flow cytometry: A learning guide. *Manual Part*, (11-11032):01.

[Chen et al., 2015] Chen, X., Hasan, M., Libri, V., Urrutia, A., Beitz, B., Rouilly, V., Duffy, D., Patin, É., Chalmond, B., Rogge, L., et al. (2015). Automated flow cytometric analysis across large numbers of samples and cell types. *Clinical Immunology*, 157(2):249–260.

[Coulter, 1956] Coulter, W. H. (1956). High speed automatic blood cell counter and cell size analyzer. In *Proc Natl Electron Conf*, volume 12, pages 1034–1040.

[Cron et al., 2013] Cron, A., Gouttefangeas, C., Frelinger, J., Lin, L., Singh, S. K., Britten, C. M., Welters, M. J., van der Burg, S. H., West, M., and Chan, C. (2013). Hierarchical modeling for rare event detection and cell subset alignment across flow cytometry samples. *PLoS computational biology*, 9(7):e1003130.

[Cros et al., 2010] Cros, J., Cagnard, N., Woollard, K., Patey, N., Zhang, S.-Y., Senechal, B., Puel, A., Biswas, S. K., Moshous, D., Picard, C., et al. (2010). Human cd14 dim monocytes patrol and sense nucleic acids and viruses via tlr7 and tlr8 receptors. *Immunity*, 33(3):375–386.

[Dempster et al., 1977] Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society. Series B (methodological)*, pages 1–38.

[Dittrich and Göhde, 1969] Dittrich, W. and Göhde, W. (1969). [impulse fluorometry of single cells in suspension]. *Zeitschrift fur Naturforschung. Teil B: Chemie, Biochemie, Biophysik, Biologie*, 24(3):360–361.

[Dundar et al., 2014] Dundar, M., Akova, F., Yerebakan, H. Z., and Rajwa, B. (2014). A non-parametric bayesian model for joint cell clustering and cluster matching: identification of anomalous sample phenotypes with random effects. *BMC bioinformatics*, 15(1):314.

[Ellis et al., 2009] Ellis, B., Haaland, P., Hahne, F., Le Meur, N., and Gopalakrishnan, N. (2009). flowcore: basic structures for flow cytometry data. *R package version*, 1(0).

[Ester et al., 1996] Ester, M., Kriegel, H.-P., Sander, J., and Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *Kdd*, volume 96, pages 226–231.

[Estivill-Castro, 2002] Estivill-Castro, V. (2002). Why so many clustering algorithms: a position paper. *ACM SIGKDD Explorations Newsletter*, 4(1):65–75.

[Fienberg and Nolan, 2014] Fienberg, H. G. and Nolan, G. P. (2014). *High-Dimensional Single Cell Analysis*. Springer.

[Friedman et al., 2007] Friedman, J., Hastie, T., Höfling, H., Tibshirani, R., et al. (2007). Pathwise coordinate optimization. *The Annals of Applied Statistics*, 1(2):302–332.

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

[Givan, 2011] Givan, A. L. (2011). Flow cytometry: an introduction. In *Flow Cytometry Protocols*, pages 1–29. Springer.

[Hasan et al., 2015] Hasan, M., Beitz, B., Rouilly, V., Libri, V., Urrutia, A., Duffy, D., Cassard, L., Di Santo, J. P., Mottez, E., Quintana-Murci, L., et al. (2015). Semi-automated and standardized cytometric procedures for multi-panel and multiparametric whole blood immunophenotyping. *Clinical Immunology*, 157(2):261–276.

[Hastie et al., 2009] Hastie, T., Tibshirani, R., Friedman, J., Hastie, T., Friedman, J., and Tibshirani, R. (2009). *The elements of statistical learning*, volume 2. Springer.

[Hu et al., 2013] Hu, X., Kim, H., Brennan, P. J., Han, B., Baecher-Allan, C. M., Jager, P. L. D., Brenner, M. B., and Raychaudhuri, S. (2013). Application of user-guided automated cytometric data analysis to large-scale immunoprofiling of invariant natural killer t cells. *Proceedings of the National Academy of Sciences of the United States of America*, 110(47):19030–19035.

[Ihaka and Gentleman, 1996] Ihaka, R. and Gentleman, R. (1996). R: a language for data analysis and graphics. *Journal of computational and graphical statistics*, 5(3):299–314.

[Jaye et al., 2012] Jaye, D. L., Bray, R. A., Gebel, H. M., Harris, W. A., and Waller, E. K. (2012). Translational applications of flow cytometry in clinical practice. *The Journal of Immunology*, 188(10):4715–4719.

[Kriegel et al., 2011] Kriegel, H.-P., Kröger, P., Sander, J., and Zimek, A. (2011). Density-based clustering. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 1(3):231–240.

[Lakoumentas et al., 2009] Lakoumentas, J., Drakos, J., Karakantza, M., Nikiforidis, G. C., and Sakellaropoulos, G. C. (2009). Bayesian clustering of flow cytometry data for the diagnosis of b-chronic lymphocytic leukemia. *Journal of biomedical informatics*, 42(2):251–261.

[Lo et al., 2008] Lo, K., Brinkman, R. R., and Gottardo, R. (2008). Automated gating of flow cytometry data via robust model-based clustering. *Cytometry Part A*, 73(4):321–332.

[Lo et al., 2009] Lo, K., Hahne, F., Brinkman, R. R., and Gottardo, R. (2009). flowClust: a bioconductor package for automated gating of flow cytometry data. *BMC Bioinformatics*, 10.

[Lugli et al., 2010] Lugli, E., Roederer, M., and Cossarizza, A. (2010). Data analysis in flow cytometry: The future just started. *Cytometry Part A*, 77(7):705–713.

[McLachlan and Peel, 2004] McLachlan, G. and Peel, D. (2004). *Finite mixture models*. John Wiley & Sons.

[O'Neill et al., 2013] O'Neill, K., Aghaeepour, N., Špidlen, J., and Brinkman, R. (2013). Flow cytometry bioinformatics. *PLoS computational biology*, 9(12):e1003365.

[Orrù et al., 2013] Orrù, V., Steri, M., Sole, G., Sidore, C., Virdis, F., Dei, M., Lai, S., Zoledziewska, M., Busonero, F., Mulas, A., Floris, M., Mentzen, W. I., Urru, S. A., Olla, S., Marongiu, M., Piras, M. G., Lobina, M., Maschio, A., Pitzalis, M., Urru, M. F., Marcelli, M., Cusano, R., Deidda, F., Serra, V., Oppo, M., Pilu, R., Reinier, F., Berutti, R., Pireddu, L., Zara, I., Porcu, E., Kwong, A., Brennan, C., Tarrier, B., Lyons, R., Kang, H. M., Uzzau, S., Atzeni, R., Valentini, M., Firinu, D., Leoni, L., Rotta, G., Naitza, S., Angius, A., Congia, M., Whalen, M. B., Jones, C. M., Schlessinger, D., Abecasis, G. R., Fiorillo, E., Sanna, S., and Cucca, F. (2013). Genetic variants regulating immune cell levels in health and disease. *Cell*, 155(1):242–256.

[Pan and Shen, 2007] Pan, W. and Shen, X. (2007). Penalized model-based clustering with application to variable selection. *The Journal of Machine Learning Research*, 8:1145–1164.

[Parham, 2009] Parham, P. (2009). *The immune system*. Garland Science, third edition.

[Pinke et al., 2013] Pinke, K. H., Calzavara, B., Faria, P. F., do Nascimento, M., Venturini, J., and Lara, V. S. (2013). Proinflammatory profile of in vitro monocytes in the ageing is affected by lymphocytes presence. *Immun Ageing*, 10(10).

[Pyne et al., 2009] Pyne, S., Hu, X., Wang, K., Rossin, E., Lin, T.-I., Maier, L. M., Baecher-Allan, C., McLachlan, G. J., Tamayo, P., Hafler, D. A., et al. (2009). Automated high-dimensional flow cytometric data analysis. *Proceedings of the National Academy of Sciences*, 106(21):8519–8524.

[Pyne et al., 2014] Pyne, S., Wang, K., Irish, J., Tamayo, P., Nazaire, M.-D., Duong, T., Lee, S., Ng, S.-K., Hafler, D., Levy, R., et al. (2014). Joint modeling and registration of cell populations in cohorts of high-dimensional flow cytometric data. *PLoS One*, 9(7):e100334.

[Rosenberg and Hirschberg, 2007] Rosenberg, A. and Hirschberg, J. (2007). V-measure: A conditional entropy-based external cluster evaluation measure. In *EMNLP-CoNLL*, volume 7, pages 410–420.

[Schwarz et al., 1978] Schwarz, G. et al. (1978). Estimating the dimension of a model. *The annals of statistics*, 6(2):461–464.

[Shapiro, 2005] Shapiro, H. M. (2005). *Practical flow cytometry*. John Wiley & Sons.

[Sugár and Sealfon, 2010] Sugár, I. P. and Sealfon, S. C. (2010). Misty mountain clustering: application to fast unsupervised flow cytometry gating. *BMC Bioinformatics*, 11.

[Tacke et al., 2007] Tacke, F., Alvarez, D., Kaplan, T. J., Jakubzick, C., Spanbroek, R., Llodra, J., Garin, A., Liu, J., Mack, M., van Rooijen, N., et al. (2007). Monocyte subsets differentially employ ccr2, ccr5, and cx3cr1 to accumulate within atherosclerotic plaques. *Journal of Clinical Investigation*, 117(1):185.

[Tacke and Randolph, 2006] Tacke, F. and Randolph, G. J. (2006). Migratory fate and differentiation of blood monocyte subsets. *Immunobiology*, 211(6):609–618.

[The MathWorks, 2014] The MathWorks (1994–2014). MATLAB, The MathWorks, Inc., Torrance, US. `http://www.mathworks.com/`.

[Thomas et al., 2015] Thomas, S., Rouilly, V., Patin, E., Alanio, C., Dubois, A., Delval, C., Marquier, L.-G., Fauchoux, N., Sayegrih, S., Vray, M., et al. (2015). The milieu intérieur study–an integrative approach for study of human immunological variance. *Clinical Immunology*, 157(2):277–293.

[Tree Star, 2014] Tree Star (1995–2014). FlowJo: TreeStar, Inc., Ashland, OR. `http://www.flowjo.com/`.

[U.S. Department of Health & Human Services, 2014] U.S. Department of Health & Human Services (2014). CD4 count. `https://www.aids.gov/hiv-aids-basics/just-diagnosed-with-hiv-aids/understand-your-test-results/cd4-count/`.

[van Rijsbergen, 1979] van Rijsbergen, C. (1979). *Information Retrieval. 1979*. Butterworth.

[Wilkins et al., 2001] Wilkins, M. F., Hardy, S. A., Boddy, L., and Morris, C. W. (2001). Comparison of five clustering algorithms to classify phytoplankton from flow cytometry data. *Cytometry*, 44(3):210–217.