Stockholms universitet

Quality control and analysis of flow cytometry data

Amos Thairu

Masteruppsats i matematisk statistik
Master Thesis in Mathematical Statistics

Masteruppsats 2012:6
Matematisk statistik
Juni 2012

www.math.su.se

Matematisk statistik
Matematiska institutionen
Stockholms universitet
106 91 Stockholm

Stockholm University

Mathematical Statistics
Stockholm University
Master Thesis **2012:6**
http://www.math.su.se

Quality control and analysis of flow cytometry data

Amos Thairu*

June 2012

Abstract

Data captured from new generation flow cytometers are characterised by increasing volume and complexity. For example a single specimen may give rise to a data set of about 1 million rows. However, analysis of these data is usually performed using user-interactive software which makes it time consuming and highly affected by operator experience. The main tasks involved in flow cytometry data analysis include visualization of scatter plots and extraction and summarization of cell sub-populations. In an attempt to solve the problems associated with the user-interactive methods, various software developments have recently emerged. In this project statistical methods are implemented in R and Bioconductor to automate the analysis of flow cytometry data collected from a HIV research study. The results are compared with those obtained by an expert analysing the same data using the traditional methods. The R functions developed are applied to an analysis of specimens taken from a HIV-infected infant at birth and at 3 months of age and changes in cell populations are described. The R code developed would be useful for laboratory scientists for automating some of the steps in the analysis of flow cytometry data, thereby offering significant savings in time and more reproducible results.

*Postal address: Mathematical Statistics, Stockholm University, SE-106 91, Sweden. E-mail: mbuguatea@yahoo.com. Supervisor: Niklas Norén.

Acknowledgements

I would like to express my gratitude to all those who gave me the possibility to complete this thesis.

A special thanks to my supervisors Professor Marie Reilly at Karolinska Institute and Niklas Noren at Stockholm University, whose help and advice helped me in writing of this thesis.

I would also like to thank Jennifer Slyker at Harborview Medical Center, Seattle for her advice and assistance.

Contents

1 Introduction

1.1 Biological background

The cells of the immune system protect the body against diseases in various ways: macrophages and neutrophils engulf the bacteria or viruses, B cells generate antibodies that bind to infectious agents and natural killer cells can kill cells infected with certain viruses. B cells and natural killer cells are members of a category of white blood cells called lymphocytes, which also contains T-cells. Some T-cells (cytotoxic lymphocytes) destroy foreign invaders while others, the helper T-cells, assist other immune cells. T-cells produce a protein called a receptor on their surface and these receptors are important in recognizing foreign antigens. T-cells can be identified among other lymphocyte types by the presence of these special receptors. The helper T-cells produce a surface protein called CD4 while the cytotoxic T-cells produce a surface protein called CD8[14].

CD4 has become well known because of its role in HIV infection. HIV attaches itself to CD4, through which it invades the cell leading to a progressive reduction in the number of T cells expressing CD4. For this reason "CD4 count" is a commonly used laboratory test useful in monitoring the immunologic status of patients with HIV infection.

Cells expressing CD4 or CD8 have numerous other markers on their surface and the types of these markers depend on the role of the cell in the immune system. Besides identifying the cell types, markers can be used to identify the state that a cell is in at a given time. These states include proliferation, differentiation, memory and senescence. Proliferation refers to cellular reproduction. Differentiation is the process by which newly formed cells become different from each other and get assigned distinct functions. Senescence is the phenomenon by which cells age, lose the ability to divide further and die. Some T-cells that have taken part in the immune response during an earlier infection have the ability to react quickly when they see the same infection again. These are called 'memory T-cells'[16]. There are more than 250 identified types of markers[15] but in this project interest is focused on CD3, CD4, CD8, CD27, CD28, CD45RA, CD56, CD57, CD71 and CCR7. This choice of markers is due to the properties that we wish to describe (see table 1).

Marker	Property defined
CD3	T cell specific antigen
CD4	Major T cell subset
CD8	Major T cell subset
CD56	NK cell lineage
CD71	proliferation
CD28	Differentiation
CD27	Differentiation
CCR7	Lymphoid homing
CD45RA	Memory
CD57	Senescence

Table 1: *Markers and their corresponding phenotype description.*

1.2 Flow cytometry

Flow cytometry is an effective way of observing the physical properties of cells by staining the various cell surface markers. It has become an important technique in clinical research especially for studies of the immune system of patients. A specimen is placed in the flow cytometer and the cells are passed through a beam of light where their morphological properties are measured using the principle of light scattering and fluorescence. Cells are first stained with antibodies and visualised by the excitement of fluorescent antibody labels. The antibodies are chosen to bind to the cell surface markers of interest so that the fluorescent intensity of the corresponding label gives a measure of the amount of the surface markers present. Other properties measured for each cell in the specimen include the size and granularity (i.e. internal complexity). All these characteristics are determined using an optical-to-electronic coupling system that records how the cell or particle scatters incident laser light and emits fluorescence[1].

After the data has been generated by the flow cyometer it is saved as Flow Cytometry Standard (.FCS) files. Data captured from new generation flow cytometers are characterised by increasing volume and complexity. For instance, a single workstation can process up to 1000 samples per day each containing hundreds of thousands of cells[2]. For each cell, simultaneous measurements are made for the labelled markers and for the size and granu-

larity. Thus a 7-colour cytometer produces data on nine parameters.

1.3 Objectives

This project was motivated by a HIV research study conducted in Kenya. In the study the immune cells of HIV-infected and HIV-exposed infants were investigated using flow cytometry assays run on specimens collected from these infants at different time points during their first year of life.

The traditional analysis of such data focuses on the visualization of scatter plots, and the gating and summarizing of sub-populations of cells. Using current standard methods, these operations are performed manually and thus the results can be highly affected by personnel experience. In addition, the analysis is very time consuming and even a skilled operator could spend 20-30 minutes identifying, gating and saving the cell sub-populations in a single sample.

The main objective of this project is to develop tools to help laboratory personnel in the capture of good quality, reproducible summaries from flow cytometry data and to enhance the analysis of cell populations in these data. The specific aims are to:

- Extract sub-populations of cells by automated gating
- Perform a quality assessment of the gating technique applied
- Generate meaningful graphical summaries of the multi-dimensional data

2 Materials and methods

Peripheral blood mononuclear cells (PBMC) were isolated from the blood of HIV-infected and HIV-exposed uninfected infants. HIV-exposed uninfected infants refer to those who are not infected but are at risk of getting infected by their mothers through breast milk. Specimens were collected at birth, 1, 3, 6, 9, and 12 months of age. In this project we will develop and illustrate our software tools using the specimens that were collected on a single infant at birth and at three months.

2.1 Markers and fluorochromes

Analysis of flow cytometry data is aimed at identifying and quantifying subpopulations of lymphocytes based on their physical properties and cell surface markers. The physical properties are indicated by the Forward-scattered light (FSC) and the side-scattered light (SSC). In particular, FSC is proportional to cell-surface area (size) while SSC is proportional to cell granularity (internal complexity). T-cells are characterised by CD3, with helper T-cells also expressing CD4 and cytotoxic T-cells expressing CD8. Specific cellular properties and function of CD4 and CD8 lymphocytes were of interest in this project and these are recognised by specific markers. For example, cells that are differentiating will express CD27 and CD28 while memory cells will express CD45RA (see tab. 1).

The higher the fluorescence intensity measurement from a given fluorochrome (label), the higher the expression of the corresponding marker on the cell. The 7-colour cytometer used here to generate the data had a capacity of allowing at most seven simultaneous markers. Since we need to measure CD3, CD4 and CD8 in order to identify the lymphocytes of interest, this leaves only four labels for identifying subpopulations of these lymphocytes. Thus the antibody markers were organized in two panels where each panel measured a specific combination of seven markers that enabled cell properties of interest to be measured (see table 2).

Panel A			Panel C		
Fluorochrome	antibody	Panel A defines	Fluorochrome	antibody	Panel C defines
Pacific Blue	CD3	T cell specific antigen	Pacific Blue	CD3	T cell specific antigen
APC-Cy7	CD4	Major T cell subset	APC-Cy7	CD4	Major T cell subset
PE-Cy7	CD8	Major T cell subset	PE-Cy7	CD8	Major T cell subset
PE-Cy5	CD56	NK cell lineage	PE-Cy5	CD56	NK cell lineage
FITC	CD28	Differentiation	FITC	CD57	Senescence
PE	CD27	Differentiation	PE	CCR7	Lymphoid homing
APC	CD45RA	Memory subset	APC	CD71	proliferation subset

Table 2: *Antibody panels used in the analysis of infant T cells.*

2.1.1 Isotype controls

Ideally, the antibodies used for staining the cells target the receptors on the cell surface. However, these antibodies might bind through non-specific protein interactions with cellular molecules (proteins, lipids and carbohydrates) thus producing fluorescence above that resulting from specific binding[13]. Moreover, fluorescence might be due to cell autofluorescence, thus there is need to distinguish specifically bound fluorescent molecules and resolve them above other nonspecific background signals[5]. For this reason, 2 panels of antibodies were designed as "isotype controls". These were antibodies directed at mouse antigens and had no specificity for the cells in question. Thus they were not expected to bind to human antigens and served as negative controls. Using the same fluorescent markers as described in table 2 these antibodies were introduced into a specimen from a normal healthy control, and the 99th percentiles of the various markers used to define a threshold. The labels for each of the phenotype defining antibodies was matched to an isotype control antibody (see table 3) and thus the isotype antibodies were organised in two panels corresponding to the experimental antibodies.

Fluorochrome	Experimental antibody	Corresponding isotype antibody
Panel A		
FITC	CD28	IGG1
PE	CD27	IGG1
APC	CD71	IG1
Panel C		
FITC	CD57	IGM
PE	CCR7	IGG2
APC	CD45RA	IG2

Table 3: *Experimental antibodies matched to isotype antibodies for determination of thresholds. The fluorochromes are used to label the corresponding markers.*

2.2 Data files and data structure

The data files used in this project are stored in a single folder. A total of six files are used for analysis: two panels for the cord blood specimen, two panels for the month 3 specimen and two isotype control panels (see table 4).

Specimen/tube ID (.fcs files)	Patient source
Patient specimens	
cord PANEL A CTL-B1-081.fcs	Infant cord blood at birth
cord PANEL C CTL-B1-081.fcs	Infant cord blood at birth
m3 PANEL A CTL-B1-081.fcs	Infant blood at 3 months
m3 PANEL C CTL-B1-081.fcs	Infant blood at 3 months
Isotype controls	
ISO I	Healthy donor- panel A
ISO II	Healthy donor- panel C

Table 4: *Flow cytometry files used to store the data.*

Each file contains data in the form of a matrix, with each row representing a single cell so that the number of rows is the total number of cells in the tube. The nine columns record the forward scatter, side scatter and the seven fluorescence intensity measurements for the markers in the panel. For example in the cord blood specimen for panel A, approximately 1 million cells were collected and stained with 7 antibodies. As a result a raw data matrix with approximately 1 million rows and 9 columns was generated. The first five rows of the data are shown in table 4.

	FSC	SSC	CD28-FITC	CD27-PE	CD56-PE-Cy5	CD8-PE-Cy7	CD3-PB	CD45RA-APC	CD4-APC-Cy7
row 1	111	31	4.2950	5.9112	3.6365	3.8296	4.2854	5.7408	5.9915
row 2	115	32	3.0037	6.0592	3.6365	3.8991	3.7025	5.7797	5.5880
row 3	116	30	3.0037	6.0185	4.5028	2.8716	7.9724	5.8319	5.5629
row 4	114	30	3.5317	6.5851	3.6039	3.1561	3.0928	6.0320	5.6005
row 5	109	31	3.3461	5.7279	4.4026	3.8991	3.8124	6.4241	5.3182

Table 5: *Extract of data showing intensity measures recorded for each cell.*

2.3 Gating

Analysis of flow cytometry data involves displaying the data on a sequence of plots and estimating the percentages of various subpopulations identified from the plot. The usual method used for this analysis is progressive reduction of the raw data into subsets using bivariate scatter plots. In the example in figure 1(a), cells are first displayed according to two parameters, FSC and SSC, allowing a sub-population to be selected (gated) according to the size and granularity of the cells. The cells in the gate in figure 1(a) have large size and low granularity and thus correspond to lymphocytes. This subset of cells is extracted and displayed according to two additional parameters, CD3 and CD8, in figure 1(b) where the CD8+ T-cells are identified as those with high CD3 and CD8. The CD4 and CD8 cells are progressively plotted in two dimensional scatter plots until the investigator is able to describe the properties of interest[1]. Figure 1(c) and figure 1(d) shows clusters of CD27 and CD28 positive cells, corresponding to differentiating cells.

There are some problems associated with this approach. Firstly, the precision of gating is dependent upon the skill of the investigator and the resolution of the computer display. Secondly, it is difficult to reproduce the same results on a different occasion. Some automatic gating algorithms aimed at making gating less biased have been developed but most software tools still require user interaction.

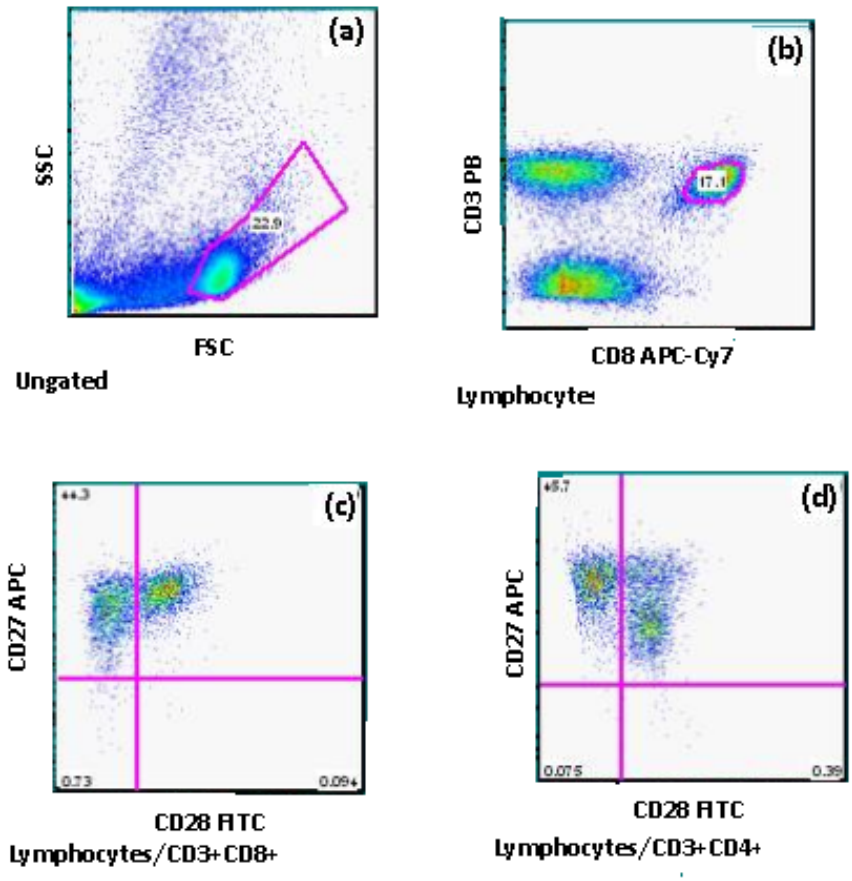


Figure 1: *Using gating to demonstrate lymphocyte phenotype.(a) Select lymphocytes based on size and granularity. (b) Select T-cell subsets based on expression of CD3 and CD8.(c) and (d) Subsequent investigation of the phenotype of T-cell subsets.*

2.4 Cluster analysis

The goal of cluster analysis is to classify a collection of objects into subsets, such that those within the same cluster are more similar to one another than they are to those assigned to different clusters. A clustering method groups the objects according to the definition of a similarity measure, e.g. Euclidean distance in the case of K-means clustering[3]. Various methods such as K-means and mixture models are used to cluster data sets.

2.4.1 K-means cluster analysis

K-means clustering is a method for finding clusters in a set of unlabeled data by minimization of the sum of the squared distances between the points in a cluster and the cluster mean[3]. Given a set of observations $(x_1, x_2, \ldots, x_n)$, one chooses the desired number of clusters, say k and the K-means procedure iteratively moves the cluster centers to minimize the total within cluster variance and to maximize the between cluster variance[6]. In other words the n observations are moved in and out of the groups (clusters) until the best partition into k clusters is achieved. This involves minimizing the objective function;

$$\underset{x}{argmin} \sum_{i=1}^{k} \sum_{xj \in s_i} (x_j - \mu_i)^2 \quad (k \leq n)$$

where S_i denotes cluster i and μ_i is the mean of the points in S_i. The iterative procedure begins by randomly selecting the initial centroids, that are then refined by repeatedly assigning points to their closest centroids. The centroids are then recomputed based on these assignments[6]. Immune cells fall into two main categories with regard to some of the markers measured i.e. FSC, CD3, CD4 and CD8. Thus k-means cluster analysis based on these measures should enable us to extract these categories (groups) from our data.
The K-means clustering algorithm used in our analysis is from Hartigan and Wong[9].

2.4.2 Bivariate normal gates

Data plotted according to two markers might exhibit an elliptical cloud with a few cells isolated from this cloud. Usually it is of interest to extract the cells that are within the cloud and discard the isolated ones, which are regarded as outliers. The distribution of the cells according to such markers can be assumed to be normal and thus a robust normal fit can be used to select cells within an appropriate number of standard deviations[4].

2.4.3 Mixture models

Mixture models are commonly applied in medical research to identify groups in datasets[11]. The Gaussian mixture model is a widely used model-based clustering technique that has been found to give good results in different areas such as biology[7]. This method is based on the assumption that each cluster has a multivariate normal distribution. However, in

many cases data are not quite normally distributed and there may be outliers. In such cases the normal mixture models might not give a good description of the data. The presence of outliers may lead to exaggeration of the number of clusters. To overcome these problems, robust methods based on adding components to the normal distribution or using the t-distribution can been used as an alternative[8, 10]. A robust method that has been used to handle these two issues is the one based on the multivariate t distributions with the following power transformation[8, 7]. A simpler model

$$L(\varphi|y_1,\ldots,y_n) = \prod_{i=1}^{n}\sum_{g=1}^{G} w_g \phi_p({y_i}^{(\lambda_g)}|\mu_g,\Sigma_g,v_g).\,|J(y_i;\lambda_g)|$$

where w_g is the probability that observation y_i belongs to the component g and $\sum_{g=1}^{G} w_g = 1$, $\phi_p(.|\mu_g,\Sigma_g,v_g)$ is the multivariate t distribution with mean μ_g (p-dimensional), covariance matrix $v_g(v_g-2)^{-1}\Sigma_g$ and v_g degrees of freedom. The transformation of y_i is $y_1^{(\lambda_g)}$ and λ_g is the Box-Cox parameter. Lastly, the Jacobian generated by this transformation is: $|J(y_i;\lambda_g)|$. To estimate the parameters: $\theta = (w_g,\mu_g,\Sigma_g,v_g,\lambda_g)$. Maximul likelihood method is used to fit the mixture model and the parameters are estimated using the Expectation Maximization (EM) algorithm. To determine the optimal number of components to have in the model various models with different number of components are fit and model selection is performed using the Bayesian Information Criterion[8, 11]. The likelihood equation might have multiple roots due to local maxima, and for this reason the EM algorithm needs to be initialized by randomly dividing the data into a number of groups. The number of groups correspond to the number of components g. An effective way of performing this is to make a number, say 10, of parallel random partitions and then run a few EM iterations. The initial partition is choosen as the one generating the highest likelihood value[8, 7, 11].

2.5 Hypothesized change in cell populations

Between birth (cord blood) and month 3 the following changes in the distribution of CD27, CD28 and CD45RA sub-populations were expected:

- Decrease in percentage of CD45RA (in both CD4 and CD8)
- Increase in percentage of CD8 T cells (CD8 proliferation and CD4 death)

- Decrease in CD27+CD28+ (naive cells) concurrent with increase in CD27-CD28+ (intermediate cells) and CD27-CD28- (late cells) in CD4 subset
- Increase in CD27+CD28- (intermediate cells) and CD27-CD28- (late cells) in CD8 subset

In general, changes in the distribution of these markers is expected to be of a larger magnitude in the CD8 cells than in the CD4 cells.

2.6 Cell sub-populations studied

Cell populations were identified successively where each one was extracted as a subset of the preceding population, based on one or two of the variables: FSC, SSC, CD3, CD4 and CD8. For our automated extraction we followed slightly different steps than those in figure **??**(a), (b) for the identification of T-cells (see figure 2). The FSC values were first examined in order to identify cells with high FSC, the "FSC+" subset, and this involved the use of cluster analysis. These cells were extracted and examined for the "CD3 PB" marker values, again using cluster analysis to extract the "CD3+" subset. The "CD3+" subset was assessed simultaneously for the levels of CD3 and the side scatter (SSC) and a bivariate normal distribution was fitted to identify the T-cells. Having extracted the T-cells, the next step was to categorize them as either "CD4+" or "CD8+" based on the CD4 and the CD8 markers respectively.

In the final stages of analysis the CD4 and CD8 positive cells were assessed for expression of various markers (CD27,CD28,CD45RA,CD57, CD71 and CCR7) to determine their phenotype. This required the use of the isotype controls as described in section 2.1.1. The isotype controls were processed by gating on FSC, SSC and then on CD4 and CD8. 99th percentiles of the intensity of isotype antibody markers were calculated and applied as cut offs on the specimens to define cells that were positive for a single marker, or to split a scatter plot into four quadrants based on the thresholds of two markers.

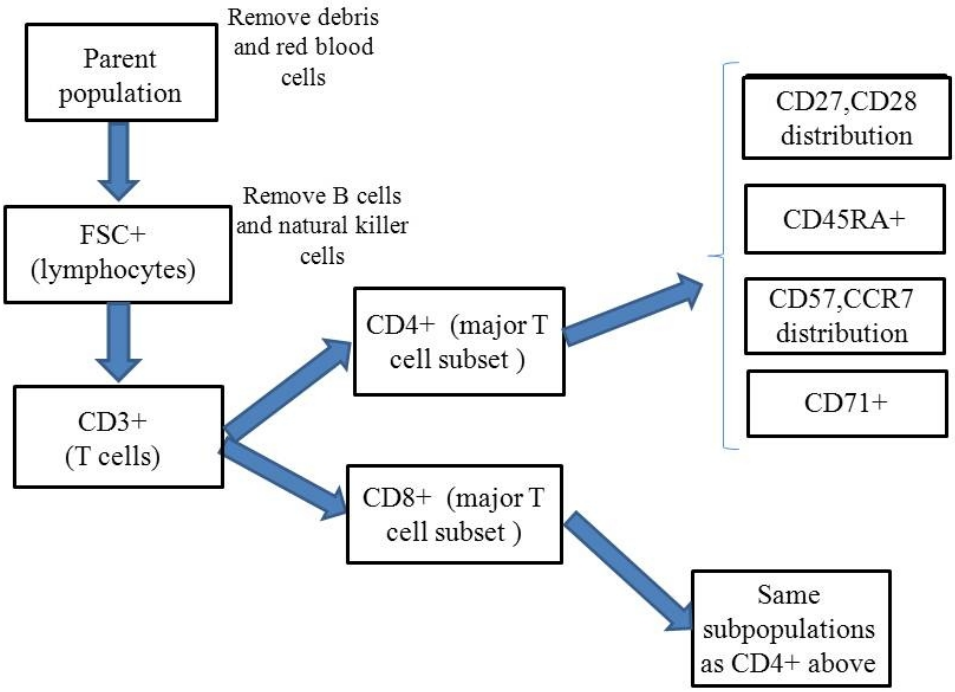


Figure 2: *Stages in the extraction of cell populations and subpopulations.*

2.7 Computation

Analysis was done in R version 2.10.1, using R commands and the flowCore[12], prada[2] and FlowClust[7] packages from Bioconductor. Custom functions were written using these commands in order to implement the specific steps required for our analysis. FCS files were imported into R using the function read.FCS{flowCore} and expression data was read from the FCS files using the function exprs{flowCore}. For cluster analysis the function kmeans{stats} was used. The fitNorm2 and plotNorm2 functions in prada were used to fit and plot the bivariate normal distributions. In the fitNorm2 function the minimum covariance determinant estimator of location and scatter was implemented to estimate the mean and covariance matrix. The FlowClust package in which the t-mixture model and the EM algorithm are implemented was used to perform model based cluster analysis.

3 Results

3.1 Cell sub-populations extracted

3.1.1 Cell sub-populations extracted using k-means

Different cell populations (such as lymphocytes, red blood cells and debris) can be distinguished by their size as indicated by their forward scatter (FSC). Lymphocytes were identified as those cells whose FSC values were above the lower threshold of the upper cluster while values below this threshold were assumed to be red blood cells and debris (see figure 3(a)).

T lymphocytes express CD3 unlike B cells and natural killer cells and were thus identified by the level of the CD3 marker. Thus the lymphocytes obtained at the previous stage were analyzed for their values of the "CD3 PB" marker and again two clusters were identified. The upper cluster, consisting of the "CD3+" cells was extracted and passed to the next stage of analysis (see figure 3(b)).

On assessing the CD3+ subset simultaneously for the levels of the "CD3 PB" and the side scatter(SSC) intensity, the bivariate distribution of these two markers showed a dense cloud of cells with some other cells isolated on the right hand side (see figure 3(c)). These isolated cells were assumed to be monocytes which are known to exhibit high SSC values, attributed to their granular structure that scatters light to a considerable extent compared to other cells. A bivariate normal distribution was fitted to the data by estimation of its center and covariance matrix and data points within a scale factor of three standard deviations from the mean of the distribution were selected (see figure 3(d)). Thus monocytes that had a low probability density in this distribution were discarded. The T-cells extracted were further categorized as either CD4+ or CD8+ cells, by applying cluster analysis to CD4 and CD8 levels and extracting the upper clusters to yield the CD4+ and CD8+ subsets (see figure 4).

Assessment of the CD4+ and CD8+ T-cell subsets for the expression of the markers of interest (CD27, CD28, CD45RA, CD57, CD71 and CCR7) yielded proportions of positive and negative sub-populations (see figure 6 to 9).

3.1.2 Cell sub-populations extracted using mixture model

The mixture model was used to estimate the proportions of CD3+, CD4+ and CD8+ populations for the month 3 panel A sample. The first step is to extract lymphocytes based on the expression of FSC and SSC as displayed in figure 10 (a). Four clusters are identified in this mixture as shown in figure 10 (b). Cluster 2 corresponds to the lymphocyte population which is passed on to the next stage and analysed according to CD3 and CD56 expression to extract the T-cells as shown in figure 10 (c). Three clusters are identified and the cluster with the high CD3 and low CD56 expression corresponds to the T-cell population. Finally, the T-cells are analysed according to CD4 and CD8 to obtain the T-cell subsets as shown in figure 10 (d). At this stage it is assumed that there are two clusters that are the CD4 and the CD8 T-cells. The Bayesian Information Criterion is used to determine the optimal number of clusters in the models involved in the various stages of extraction of subpopulations. A plot of BIC against the number of clusters is shown in figure 11.
The estimated proportions of the various sub-populations for the month 3 panel A specimen are displayed in table 6. These proportions match those found by the k-means approach.

Population	Manual	K-Means	Mixture Model
CD3+	53,304(24.33)	66,328(30.27)	61,484(28.06)
CD4+	25,069(47)	33,761(56.50)	32,629(53.07)
CD8+	21,247(39.9)	25,290(41.86)	22,869(37.19)

Table 6: *Comparison of sizes (proportions in parenthesis) of cell populations defined manually versus and those estimated using K-Means and mixture model clustering approaches. Comparison is done for the month 3 Panel A sample.*

3.2 Proportions of Subpopulations at different time points

There were remarkable changes in the proportions of the various sub-populations identified at birth and at 3 months. A summary of these changes is displayed in table 7 and table 8. There was a decrease in CD4 cells while the CD8 cells increased. This is expected since a decrease in CD4 cells is concurrent to an increase in CD8 cells. These CD4 and CD8 cells were extracted and analysed further according to their expression of the CD45RA, CD57, CD71, ccr7, CD27 and CD28 markers. There is a decrease in CD45RA+ cells in both CD4 and CD8 populations with a larger change among the CD8 cells (see figure 6). Expression of

CD27 and CD28 decreased between the two time points in both CD4 and CD8 populations with a larger change among the CD8 cells (see figure 7). As seen in the figure the clusters shifts to the left and towards the bottom of the graph. An increase in CD57+CCR7- cells in the CD4 and CD8 populations (higher for the CD8 cells compared to the CD4 cells) is notable. At birth there are quite a few cells at the upper quadrants but this is seen to increase at month 3 especially at the upper left quadrant (see figure 8). A mild change is seen in the expression of CD71 for both t-cell subsets (see figure 9).
These changes in general indicate important changes in the immunological status between birth and at three months. CD8 proliferation (increase) and CD4 death are consistent with the changes caused by HIV and CMV infection.

Panel A

	Birth		3 months		
Population	N	% of gate	N	% of gate	Change in prop.
total cells	1000000		219097		
FSC+	311921	31.2	131497	60.0	28.2
CD3+	92583	29.7	68382	52.0	22.3
CD4+	62135	79.2	35227	56.5	-22.7
CD45ra+	49787	80.22	25251	74.81	-5.41
CD27+CD28+	57943	93.37	30007	88.9	-4.47
CD27-CD28+	274	0.44	493	1.46	1.02
CD27+CD28-	3821	6.16	2447	7.25	1.09
CD27-CD28-	22	0.04	805	2.39	2.35
CD8+	14599	18.6	25755	41.3	22.7
CD45ra+	13840	94.98	13858	54.81	-40.17
CD27+CD28+	11197	76.84	6837	27.04	-49.8
CD27-CD28+	126	0.86	430	1.7	0.84
CD27+CD28-	3178	21.81	12268	48.52	26.71
CD27-CD28-	70	0.48	5751	22.74	22.26

Table 7: *Data summary of cell populations and sub-populations for panel A.*

Panel C

	Birth		3 months		
Population	N	% of gate	N	% of gate	Change in prop.
total cells	899205		212387		0
FSC+	307840	34.2	135447	63.8	29.6
CD3+	99223	32.2	70104	51.8	19.6
CD4+	68204	80.5	37105	57.6	-22.9
CD71+	1224	1.81	505	1.42	-0.39
CD57+CCR7+	304	0.45	281	0.79	0.34
CD57-CCR7+	38111	56.3	14453	40.66	-15.64
CD57+CCR7-	216	0.32	1784	5.02	4.7
CD57-CCR7-	29057	42.93	19026	53.53	10.6
CD8+	12316	14.5	25515	39.6	25.1
CD71+	314	2.57	259	1.03	-1.54
CD57+CCR7+	87	0.71	134	0.53	-0.18
CD57-CCR7+	3764	30.71	1369	5.46	-25.25
CD57+CCR7-	175	1.43	9603	38.28	36.85
CD57-CCR7-	8172	67.01	13976	55.71	-11.3

Table 8: *Data summary of cell populations and sub-populations for panel C.*

3.3 Descriptive analysis of cell populations

Various stages involved in the progressive subsetting of the cell populations and and the changes remarkable in the cell populations are displayed in the plots below. Detailed explanations of the graphs are done in section 3.1 and section 3.2.

3.3.1 Extraction of T-cell subsets

The initial stages of the automated subsetting from the raw cells up to the T-cell subsets is displayed in figure 3 and 4. The process is explained in detail in section 3.1.1.

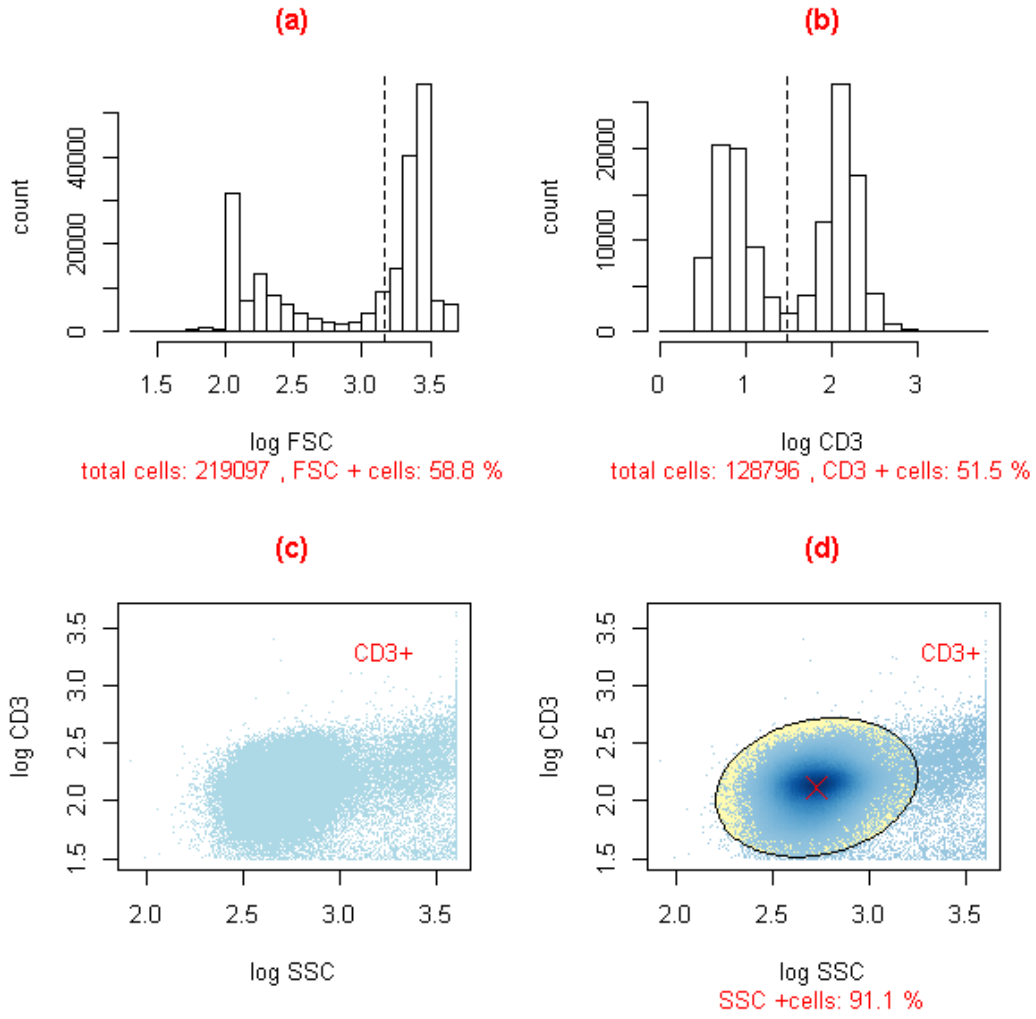


Figure 3: *Graphical representation of the initial stages of the automated subsetting process. Subsetting starts from the raw cells until the T-cell population has been extracted. Horizontal broken lines in (a) and (b) indicate the thresholds d to separating the clusters identified by K-means cluster analysis. The ellipse in (d) is obtained by fitting a bivariate normal distribution.*

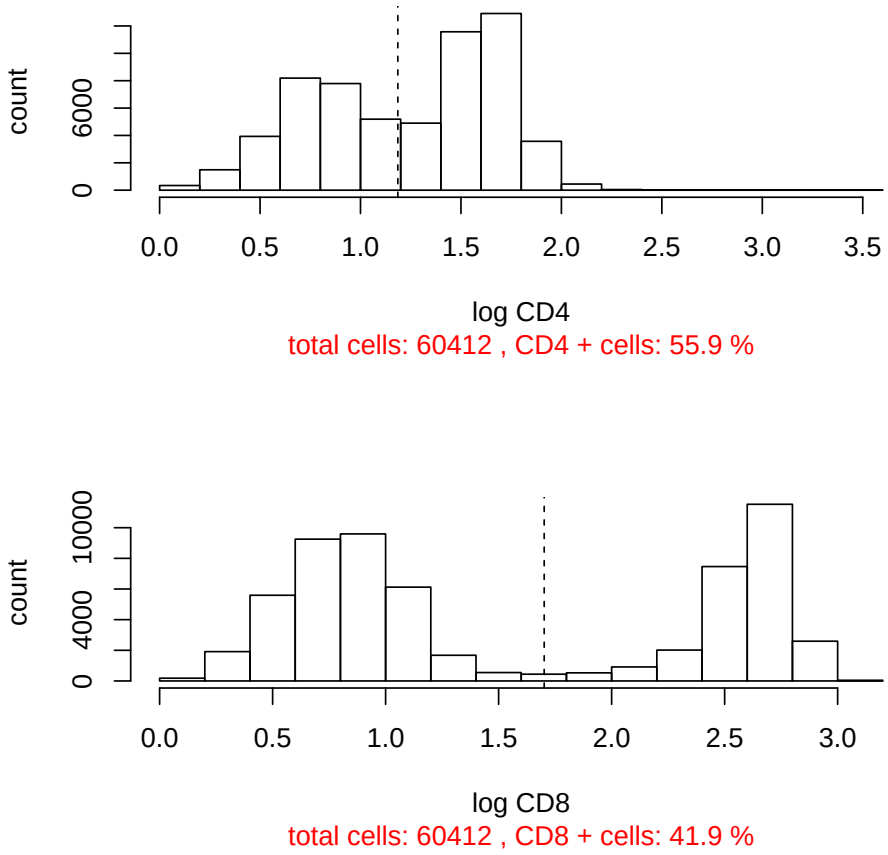


Figure 4: *Extraction of T-cell subsets. The upper cluster in each plot is defined as the CD4+ and CD8+ cells respectively.*

3.3.2 Determination of thresholds from isotype controls

As explained in section 2.1.1 and 2.6, two panels of antibodies were designed as isotype controls to served as negative controls. The plots on the left at figure 5 below shows derermination of thresholds from the IGG1-PE and the IG1-APC isotype controls. 99th percentiles of the expression of CD27 and CD45RA markers are computed and applied on the t-cell subsets to determine CD4+CD27+ and CD8+CD45RA+ populations respectively. This is further shown in figure 6 and figure 7.

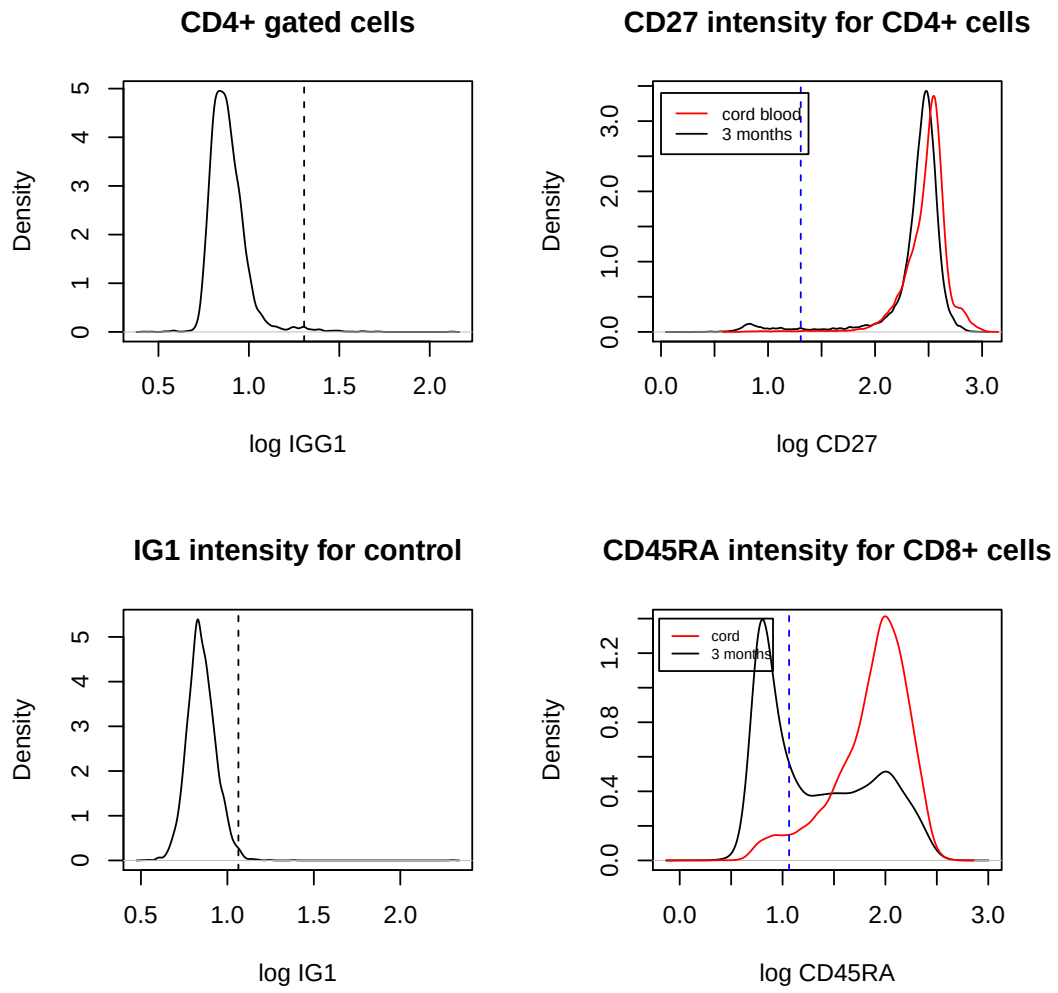


Figure 5: *Determination of thresholds from isotype controls to use in extracting the final cell subpopulations. The two thresholds displayed here represent the 99th percentile of CD27 and CD45 calculated on the isotype controls and they mark the level above which fluorescence intensity can be considered specific. The controls were processed by gating on FSC, SSC and then CD4+ and CD8+.*

3.3.3 Description of CD4 and CD8 subsets

With the computed thresholds assessment of the CD4+ and CD8+ T-cell subsets was done for the expression of the markers of interest:CD27, CD28, CD45RA, CD57, CD71 and CCR7, yielding proportions of positive and negative sub-populations. These proportions are displayed in figure 6 to 9 and the changes in the populations between birth and month

3 is explained in detail in section 3.2.

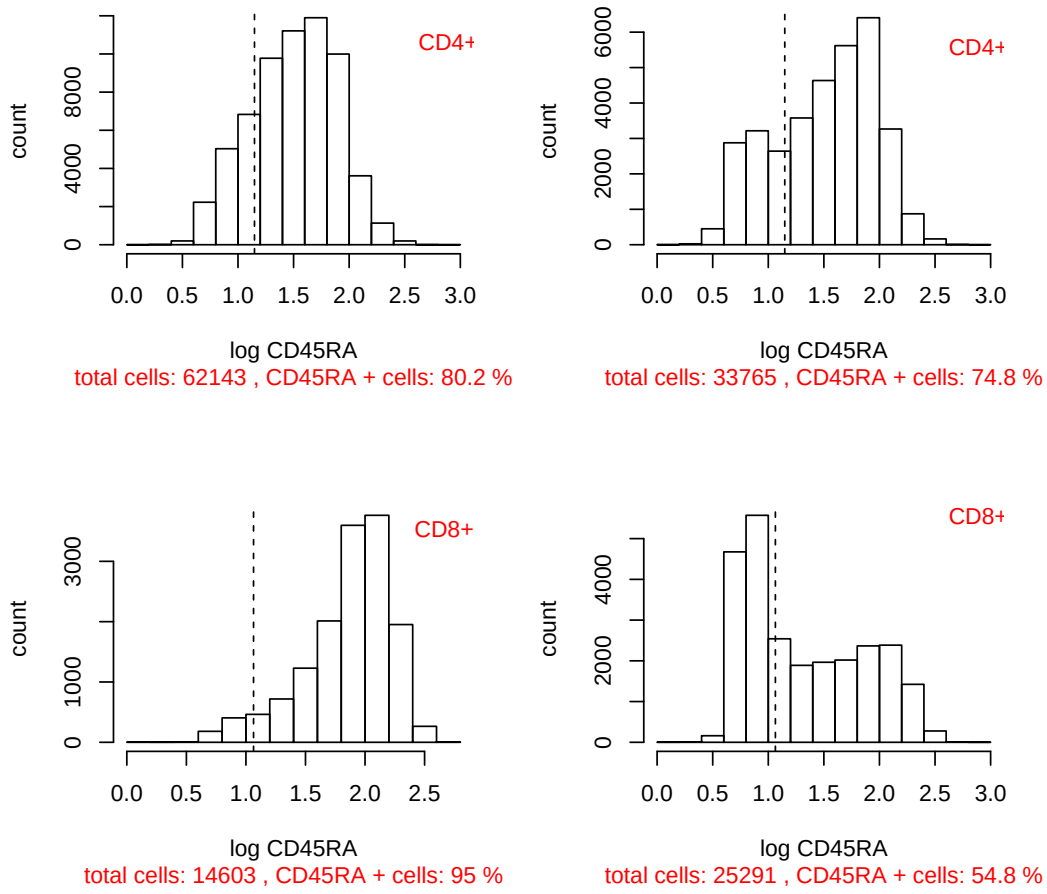


Figure 6: *CD45RA expression for the CD4+ and CD8+ subpopulations at birth (left hand side) and at month 3 (right hand side). The dashed vertical lines are the cut offs calculated as the 99th percentiles of the intensity of isotype antibody markers which are applied on the specimens to define CD 45RA+ cells.*

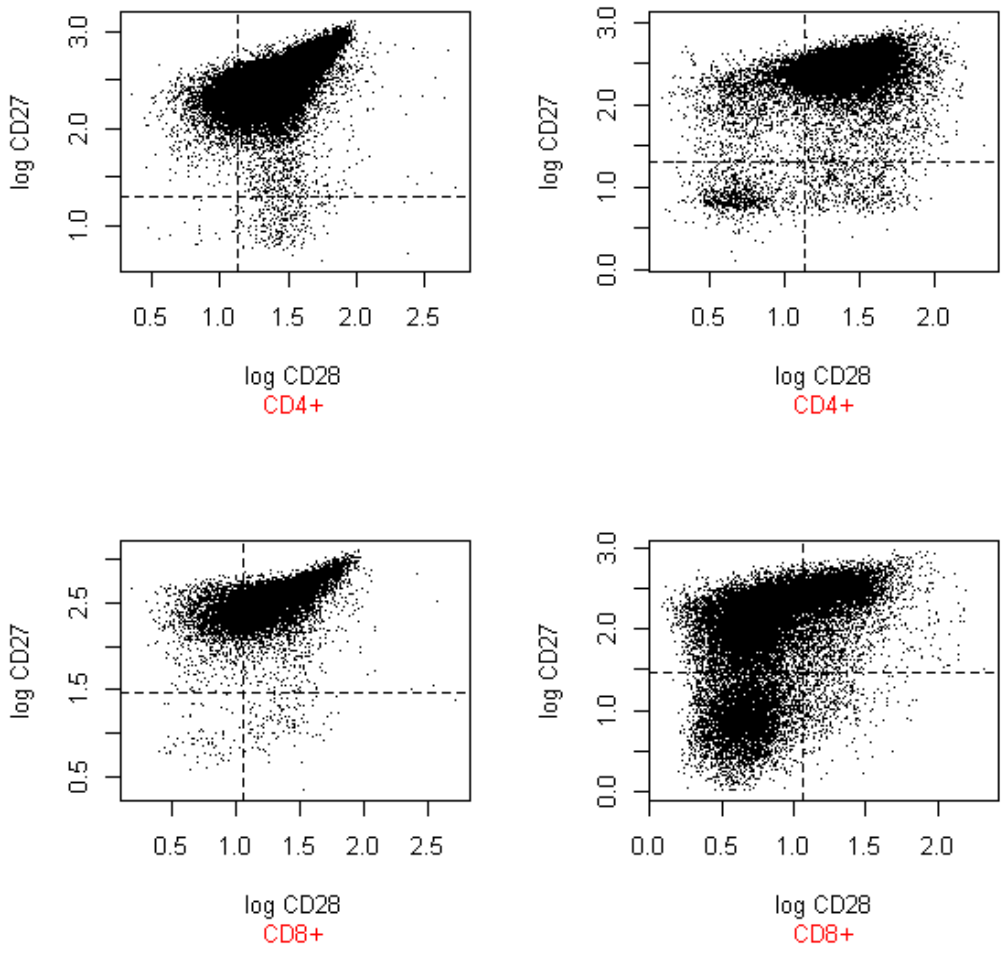


Figure 7: *CD27 and CD28 expression for the CD4+ and CD8+ subpopulations at birth and at 3 months. The dashed vertical lines are the cut offs calculated as the 99th percentiles of the intensity of isotype antibody markers which are applied on the specimens to define CD27+ and CD28+ cells. The clusters shifts to the left and towards the bottom of the graph.*

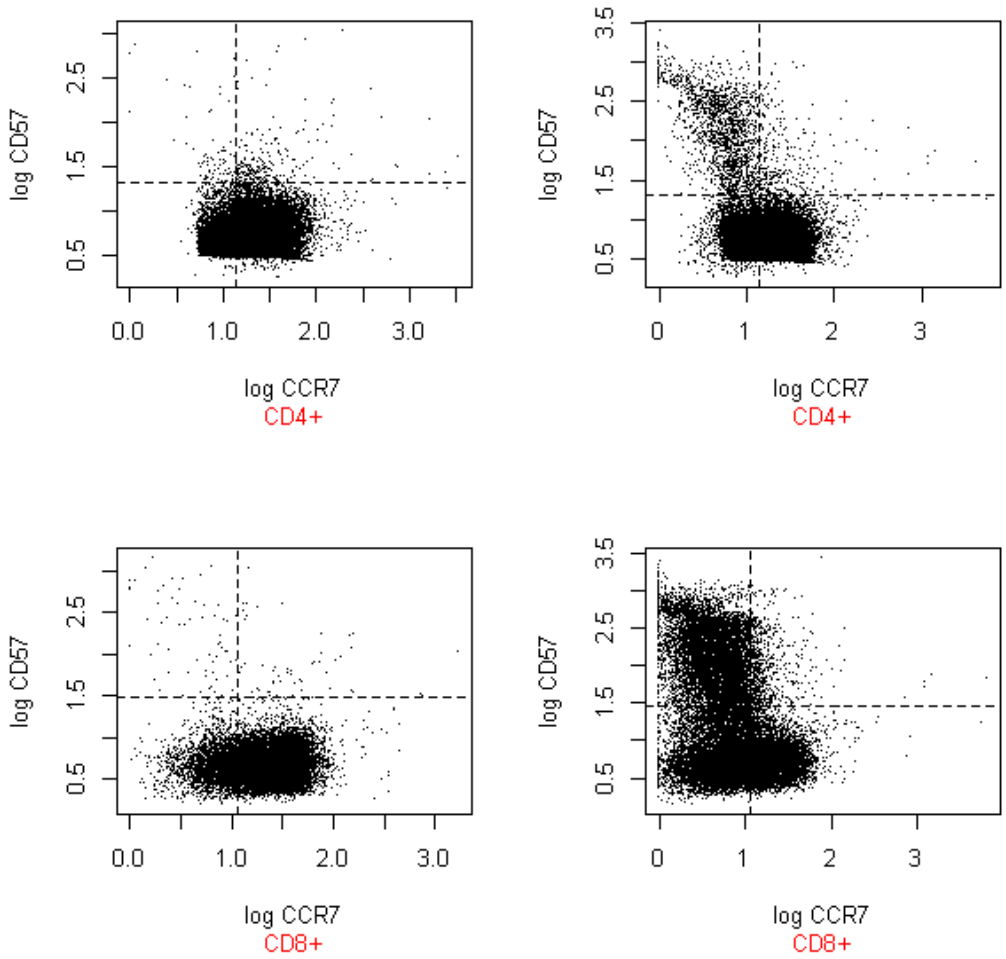


Figure 8: *CD57 and ccr7 expression for the CD4+ and CD8+ subpopulations at birth and at 3 months. The dashed vertical lines are the cut offs calculated as the 99th percentiles of the intensity of isotype antibody markers which are applied on the specimens to define CD57+ and ccr7+ cells. An increase at the upper left quadrant is remarkable.*

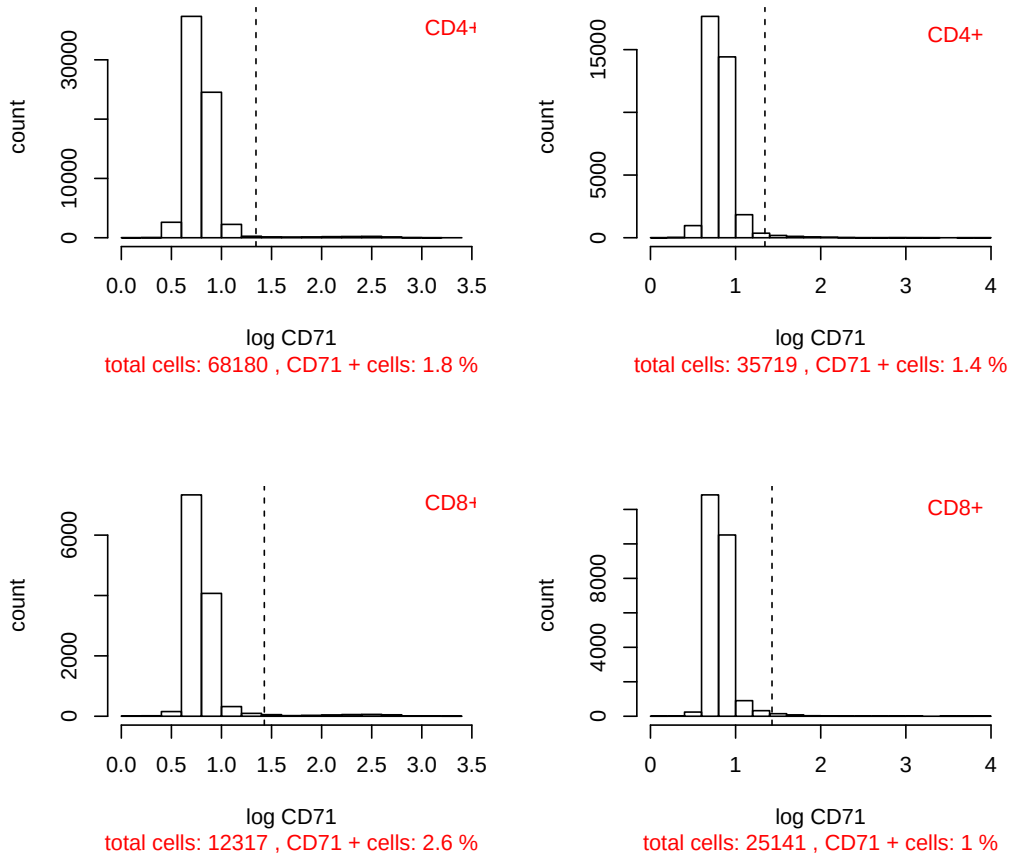


Figure 9: *CD71 expression for the CD4+ and CD8+ subpopulations at birth and at 3 months. The dashed vertical lines are the cut offs calculated as the 99th percentiles of the intensity of isotype antibody markers which are applied on the specimens to define CD71+ cells. A slight change between the two time points can be seen.*

3.3.4 Extraction of sub-populations using the mixture model

The mixture model was used to estimate the proportions of CD3+, CD4+ and CD8+ populations for the month 3 panel A sample. This is explained in detail in section 3.1.2.

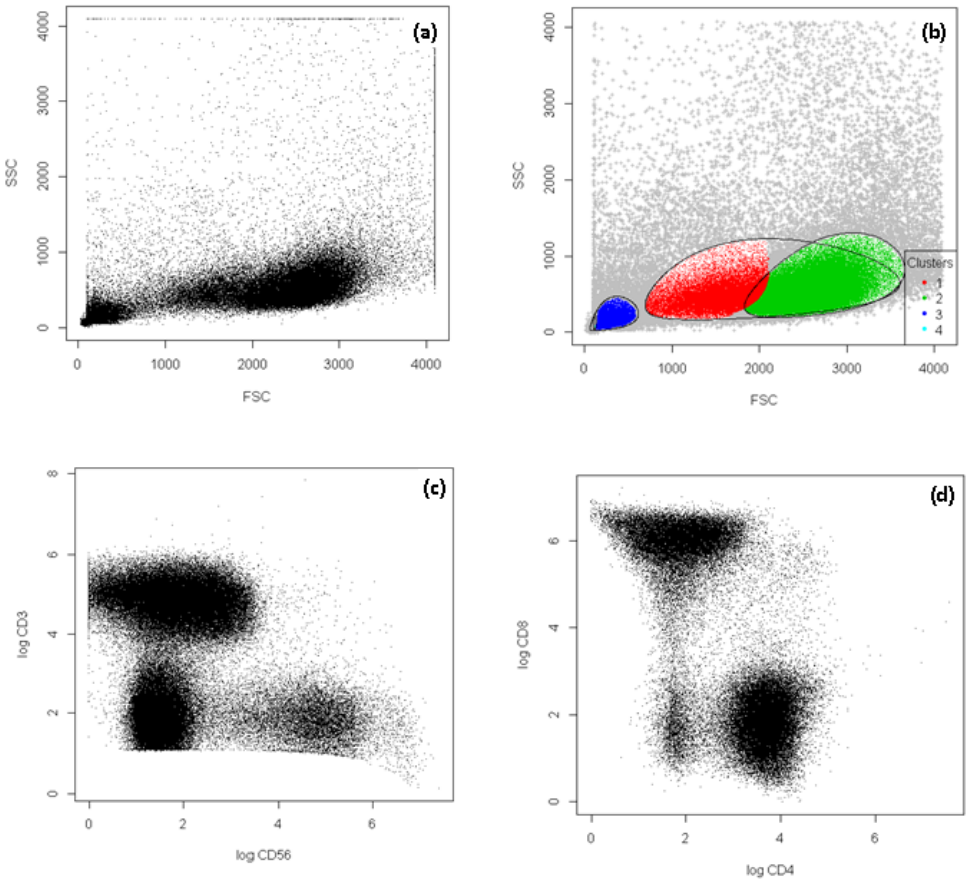


Figure 10: *Extraction of sub-populations using the mixture model approach. Subsetting starts from the raw cells whose distribution according to FSC and SSC is shown in (a). Plot (b) shows the clusters representing the lymphocytes (cluster 2) and the other components. The lymphocyte subpopulations: T-cells, B-cells and natural killer cells, can be seen as the 3 clusters in (c). (d) is the plot of one of the clusters from (c), in particular the T-cell sub-population.*

3.4 Assessing the precision of estimates (Bootstrap)

Bootstrap analysis of proportions and thresholds obtained by k-means cluster analysis is perfomed to get a sense of how much variability there is in the findings. Sampling from the empirical distribution is done with replacement to generate 200 bootstrap replicates with sample size equal to that of the original sample. The result is displayed in Table 9. The estimated standard errors and bias are low indicating a high precision of estimates.

Statistic	Original value	Bias	Std. error
Proportion of FSC+	0.59	-0.000165930	0.000454
Threshold for FSC+	3.16	0.001869281	0.001996
Proportion of CD3+	0.51	-1.068932e-05	0.000772
Threshold for CD3+	1.48	-4.385876e-04	0.001438
Proportion of CD4+	0.56	0.0002504172	0.001519
Threshold for CD4+	1.19	-0.0002024251	0.004060
Proportion of CD8+	0.42	-5.575494e-05	0.001477
Threshold for CD8+	1.70	1.450713e-03	0.002599

Table 9: *Bootstrap analysis of proportions and thresholds calculated for the month 3 panel A sample.*

3.5 Comparison of the manual and the automated procedures

	Manual	Automated	
	N(% of parent gate)	N(% of parent gate)	Diff. in prop
Cord Blood			
Cells acquired	1,000,000	1,000,000	
CD3+	63,438(6.34)	92,599(9.26)	2.92
CD4+	48,888(77.7)	62,129(79.2)	1.5
CD45RA+	38,483(78.7)	49,822(80.24)	1.54
CD27+CD28+	47,677(97.5)	57,975(93.37)	-4.13
CD27-CD28+	99(0.2)	271(0.44)	0.24
CD27+CD28-	1,107(2.26)	3,823(6.16)	3.9
CD27-CD28-	5(0.01)	22(0.04)	0.03
CD8+	10,036(15.8)	14,597(18.6)	2.8
CD45RA+	9,457(94.2)	13,850(94.99)	0.79
CD27+CD28+	8,069(80.4)	11,206(76.86)	-3.54
CD27-CD28+	0(0)	126(0.86)	0.86
CD27+CD28-	1,953(19.5)	3,179(21.8)	2.3
CD27-CD28-	14(0.14)	69(0.47)	0.33

Table 10: *Comparison of proportions of cell populations defined manually versus automatically. Sub-gates are indicated by indentations below parent gates. Comparison is done for the cord blood Panel A sample.*

	Manual	Automated	
	(N)% of parent gate	% of parent gate	Diff. in prop
Month 3			
Cells acquired	219,097	219,097	
CD3+	53,304(24.33)	66,328(30.27)	5.94
CD4+	25,069(47)	33,761(56.5)	9.5
CD45RA+	19,573(78.1)	25,268(74.82)	-3.28
CD27+CD28+	24,251(96.7)	30,024(88.9)	-7.8
CD27-CD28+	274(1.09)	494(1.46)	0.37
CD27+CD28-	215(0.86)	2,449(7.25)	6.39
CD27-CD28-	329(1.31)	805(2.38)	1.07
CD8+	21,247(39.9)	25,290(41.86)	1.96
CD45RA+	10,528(49.6)	13,868(54.82)	5.22
CD27+CD28+	5,331(25.1)	6,846(27.06)	1.96
CD27-CD28+	223(1.05)	431(1.7)	0.65
CD27+CD28-	10,806(50.9)	12,269(48.5)	-2.4
CD27-CD28-	4,887(23)	5,752(22.74)	-0.26

Table 11: *Comparison of proportions of cell populations defined manually versus automatically. Sub-gates are indicated by indentations below parent gates. Comparison is done for the month 3 Panel A sample.*

In general the proportions resulting from the manual and the automated procedures concurr with differences less than 6%. However the proportions obtained for CD4+ CD27+CD28- CD27+CD28+ have larger differences as 9.5, 6.4 and -7.8 respectively.

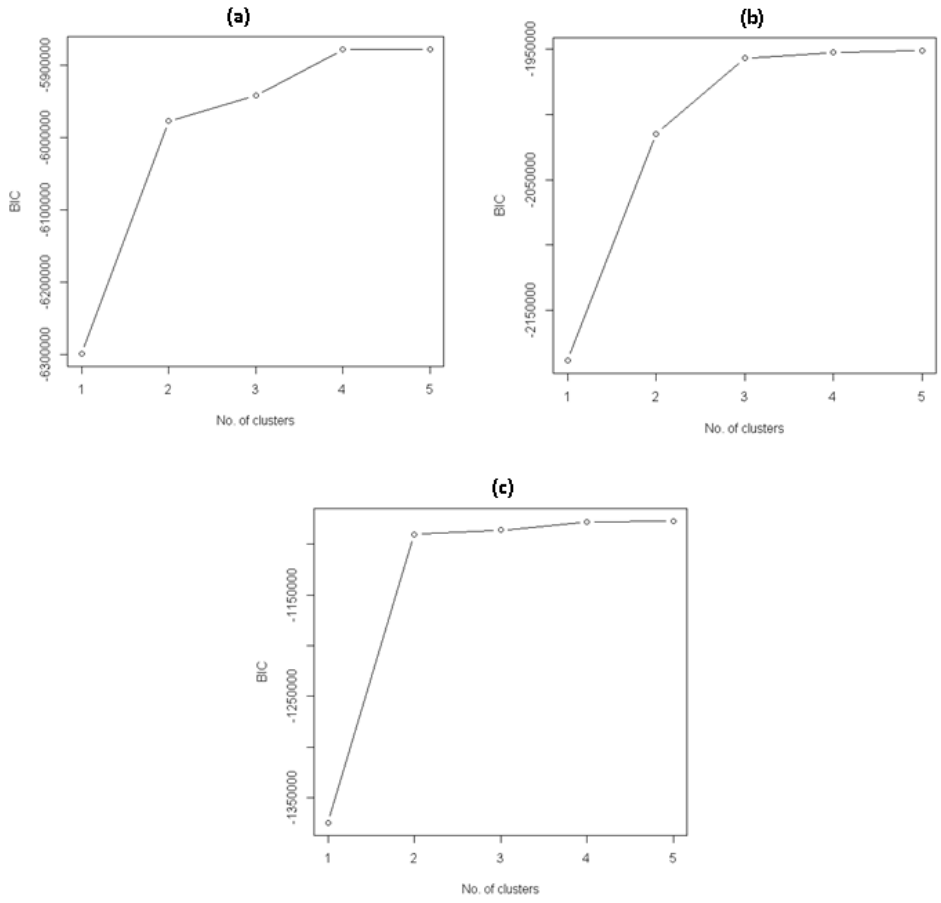
3.6 Conclusion

The automated gating method developed in this thesis has proved to be capable of performing gating efficiently and producing meaningful graphical summaries. The functions developed to perform these tasks (see section B in the appendix) can be combined into an R package that would aid laboratory scientists in producing quick results from flow cytometry data. In addition automation will yield reproducible results that reduce dependence on personnel experience. This could be particularly helpful in cases where large data sets are to be analysed and in which manual gating would be tedious to employ. The automated method can be extended to handle more that two dimensions at the same time when it is neccesary to fully appreciate the high dimensionality in the data. This would be an advantage over the manual method in that the human eye would not easily visualize data in three dimensions.

The results obtained using this automated method agree closely with those obtained by an expert analysing the same data set. The expert has good experience in analysing such specimens manualy and has sound understanding of the underlying biology of the experiment. We believe that the functions developed could easily be generalised to investigations of cells in settings other than in the HIV context for which it was developed.

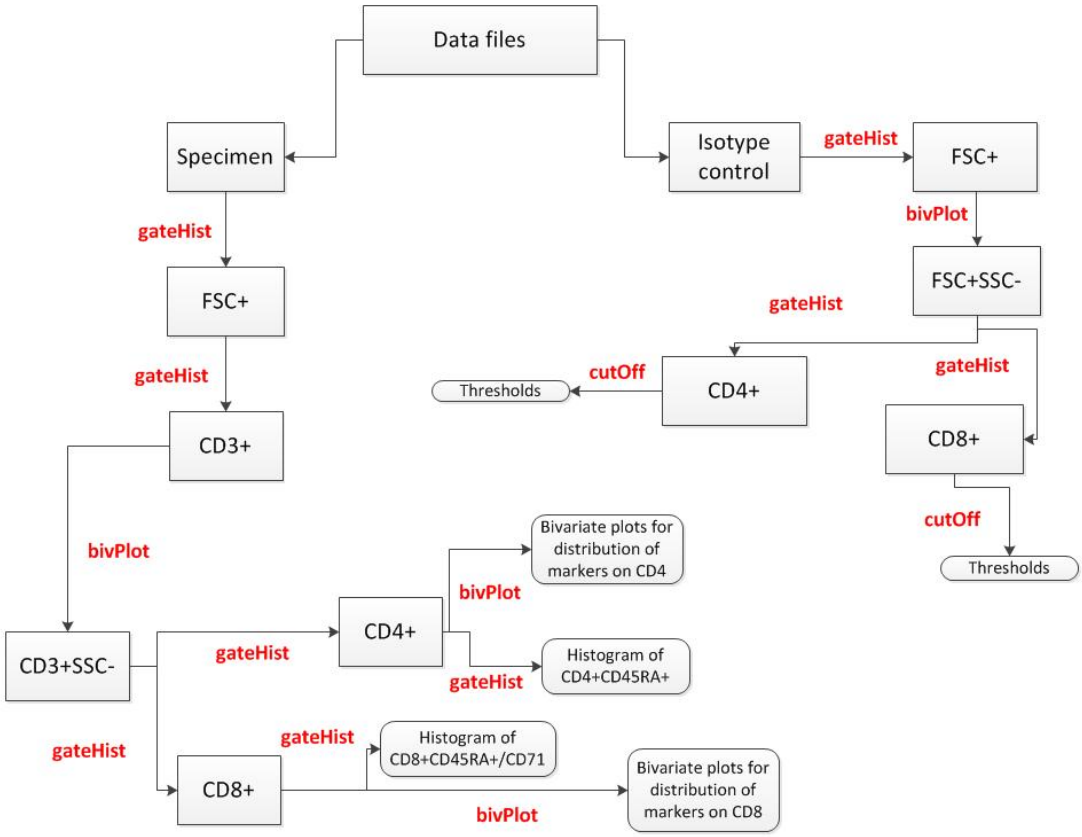
A further study that that applies the tools developed in this project and uses more specimens from the HIV study would be neccesary to investigate the population proportions that had relatively larger differences obtained using the manual and automated method. At the same time the performance of the two k-means and the mixture model approaches could be compared. The R implementation of the k-means algorithm, kmeans, is quite fast. Running the the algorithm on the large data sets on an ordinary pc takes about 4 seconds. However the k-means approach did not perform well in identifying subsets in more than one dimension for this data set and for this reason it was applied on one variable at a time. The mixture model on the other hand performed well in analysis with two variables and this could be an advantage in that it appreciates the more than one dimensionality of the data.

A supplementary figures



A. 11: *Determination of the optimal number of clusters using the Bayesian Information Criterion. The flattening of the BIC curve indicates that the optimal number of clusters has been reached. Plot (a) indicates that there are four clusters in the raw cells, (b) corresponds to the lymphocyte population and (c) corresponds to the T-cell population.*

B Functions created in R



A. 12: *Flow chart of functions created to gate and produce graphical summaries.*

gateHist

Input: data to subset, marker to gate on, threshold from isotype control (in case of the phenotype description stage)

Options: gate using K-means or based on threshold calculated from isotype control, log or linear scale, extract upper or lower cluster

Output: Histogram with a threshold, subset extracted

bivPlot

Input: data to subset, markers to gate on, thresholds from isotype control (in case of the phenotype description stage)

Options: display thresholds on the plot, log or linear scale, fit ellipse (bivariate normal distribution)

Output: scatter plot, fitted ellipse, subset extracted

cutOff

Input: isotype control data set, isotype control marker

Output: threshold calculated from isotype control

C List of references

References

[1] BD Biosciences, *"Introduction to Flow Cytometry: A Learning Guide by BD Biosciences", http://www.cancer.umn.edu/exfiles/research/fcintro.pdf.* Manual Part Number: 11-11032-01, April 2000.

[2] Nolwenn Le Meur, Florian Hahne, *"Analyzing Flow Cytometry Data with Bioconductor", Rnews, vol. 6, no. 5, pp. 2732, 2006.*

[3] Trevor Hastie, Robert Tibshirani, Jerome Friedman, *"The Elements of Statistical Learning: Data Mining, Inference and Prediction",* 2nd ed, Springer Verlag 2009. pg. 460, 501, 502, http://www-stat.stanford.edu/ tibs/ElemStatLearn/

[4] Florian Hahne, Dorit Arlt, Mamatha Sauermann, Meher Majety, Annemarie Poustka, Stefan Wiemann and Wolfgang Huber, *"Statistical methods and software for the analysis of highthroughput reverse genetic assays using flow cytometry readouts"*, 2006.

[5] Victoria L. Mosiman, Bruce K. Patterson, Luis Canterero, Charles L. Goolsby, *"Reducing Cellular Autofluorescence in Flow Cytometry: An In Situ Method"*, 1997.

[6] Thomas hill, Pawel Lewicki, *"Statistics, methods and applications, A comprehensive method for science, industry and data mining",* 1st ed, StatSoft inc., 2006, pg.121,122

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

[8] Kenneth Lo, Raphael Gottardo, *"Flexible mixture modeling via the multivariate t distribution with the Box-Cox transformation: an alternative to the skew-t distribution",* Stat Comput (2012) 22:3352.

[9] Hartigan, J. A. and Wong, M. A. (1979) *"A K-means clustering algorithm",* Applied Statistics 28, 100108.

[10] Peel D, GJ McLachlan, *"Robust mixture modelling using the t distribution",* Stat Comput 2000, 10(4):339-348.

[11] GJ McLachlan, SU Chang, *"Mixturemodelling for cluster analysis",* StatisticalMethods in Medical Research 2004; 13: 347-361.

[12] Florian Hahne, Nolwenn LeMeur, Ryan R Brinkman, Byron Ellis, Perry Haaland, Deepayan Sarkar, Josef Spidlen, Errol Strain, Robert Gentleman, *"flowCore: a Bioconductor package for high throughput flow cytometry",* BMC Bioinformatics 2009, 10:106 doi:10.1186/1471-2105-10-106.

[13] BD Biosciences, *http://www.bdbiosciences.com/support/resources/protocols/isotope_control.jsp.*

[14] Biology Reference, *http://www.biologyreference.com/Se-T/T-Cells.html.* Aug 11, 2010.

[15] Immunity blogspot, *http://cellular-immunity.blogspot.com/2007/12/cd.html.* 22 Dec 2010.

[16] Wikipedia encyclopedia, *http://en.wikipedia.org/wiki/Memory_T_Cells.* 15 May 2012.