Automated Analysis of High-Throughput Flow Cytometry Data from Hematopoietic Stem Cell Experiments

by

Sergio Adrián Cortés

B.Sc., The University of British Columbia, 2006

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

MASTER OF SCIENCE

in

The Faculty of Graduate Studies

(Bioinformatics)

THE UNIVERSITY OF BRITISH COLUMBIA

(Vancouver)

March 2010

Abstract

Flow cytometry (FCM) is a technology that allows the rapid quantification of physical and chemical properties of up to millions of cells in a sample. It is a technology commonly used in drug discovery, health research, medical diagnosis and treatment, and vaccine development. Recent technological advancements in optics and reagents allow the quantification of up to 21 parameters per cell and advancements in robotics allow the use of FCM as a high-throughput technology. Lagging in the development of FCM technologies is the data analysis component. Conventional analysis of FCM data is labour intensive, subjective, hard to reproduce, error prone and not standardized. Indeed, the traditional analysis represents one of the main bottlenecks for the future adoption of recent technological advancements in biomedical research and the clinical environment.

Here, an analysis framework developed for the automated analysis of FCM data derived from hematopoietic stem cell (HSC) transplant experiments using data generated in the Terry Fox Laboratory is presented. The data analysis pipeline developed aims to simplify approaches to analyze such data and generated automated tools for accurate analysis and quality control. The tool presented achieves equivalent results when compared to the traditional analysis, but avoids the traditional need for continuous user interaction.

Incorporated into the analysis pipeline, is a model to predict the repopulation outcome from the HSC transplant experiments. Because HSC purification strategies are typically below 50%, more than half of the mice transplanted with a single cell will not be repopulated. The repopulation prediction model showed a performance of correctly identifying 81% of the mice that did not showed a positive engraftment, while keeping the incorrect misclassification of positive engraftments below 5%.

Table of Contents

Abstract ii

Table of Contents iv

List of Tables vii

List of Figures viii

List of Abbreviations xi

Preface xii

Acknowledgements xiii

Dedication xv

1 Introduction 1

1.1 Flow Cytometry (FCM) 2

1.1.1 The Technology 3

1.1.2 FCM Bioinformatics 7

1.2 Hematopoietic Stem Cells (HSC) 11

1.2.1 Background 11

1.2.2 Murine HSC Markers and Isolation of HSC 13
1.2.3 Assessment of HSC from Mouse Bone Marrow by their *In Vivo* Repopulating Activity 14

2 Automated Analysis Pipeline 21
2.1 Methods 22
2.1.1 Clustering Techniques 22
2.1.2 Kullback-Leibler Divergence (KLD) 22
2.1.3 1 Dimensional High Density Region Analysis 24
2.2 Automated Gating Strategy 28
2.2.1 Data Preprocessing and Transformation 28
2.2.2 Identifying Viable Cells 29
2.2.3 Identifying Cell Types 34
2.3 Quality Control 38
2.3.1 Processes Description 41
2.4 Results and Assessment of the Automated Gating Pipeline . . 46
2.5 Implementation and Accessibility 48
2.6 Discussion 50

3 Repopulation Analysis 51
3.1 The Data 52
3.2 Methods 54
3.2.1 The Prediction Models 54
3.2.2 Classifier Selection 56
3.3 Results and Application 60

4 Discussion and Conclusion 64

Bibliography 66

List of Tables

1.1 HSC sorting strategies. 14
1.2 Description of staining cocktails. 17
1.3 Data used in the development and validation of the automated analysis pipeline 17

3.1 List of possible prediction outcomes 52
3.2 Data used for the repopulation prediction analysis of single cell HSC transplant experiments 53
3.3 Classification models examined 54
3.4 Repopulation prediction results on 341 single cell transplants . 63

List of Figures

1.1 Schematic diagram of a typical flow cytometer 4
1.2 Light scattering diagram 5
1.3 Fluorochrome emission 6
1.4 Fluorochrome emission distributions when excited with a 480 nm laser 7
1.5 Model of the hematopoietic developmental hierarchy 12
1.6 Flow cytometric profiles for isolating CD45^{mid}lin^{-}Rho^{-}SP cells 15
1.7 Gating strategy for selecting viable cells. 19
1.8 Manual gating strategy for selecting donor and recipient derived cell populations 20
1.9 Manual gating strategy for selecting lineage marker-positive cells 20

2.1 Automated pipeline analysis diagram 23
2.2 Flow cytometry profile showing recipient and donor derived populations 26
2.3 Identification of high density regions 27
2.4 First gating step in the identification of viable cells 31

2.5 Classification of homogeneous populations based of morphology parameters. 33
2.6 Locating landmarks for cell populations from positive control samples 1 36
2.7 Locating landmarks for cell populations from positive control samples 2 37
2.8 Identifying recipient and donor derived cell populations 39
2.9 Identification of $B220^+$ cells 40
2.10 Equivalence among shared parameters for sample tubes 43
2.11 Example of manual gates on mistakenly stained samples . . . 43
2.12 Quality control on donor derived cell output 45
2.13 Pipeline results comparison 47
2.14 Screen capture of the GenePattern module configuration site . 49

3.1 Repopulation prediction false negative rate as a function of the cutoff value 57
3.2 Repopulation prediction tradeoff between the false negative rate and the false positive rate 59
3.3 Performance comparison for different binary classification predictors in the repopulation problem 60
3.4 Repopulation probability at the 16 week post-transplant time point as a function of the donor-derived proportion of viable cells at the 8 week post-transplant time point 62

List of Abbreviations

APC	Allophycocyanin
ARC	Animal Resource Centre
BIC	Bayesian Information Criterion
CART	Classification and Regression Tree
ECDF	Empirical Cumulative Distribution Function
FACS	Fluorescent-Activated Cell Sorting
FCM	Flow Cytometry
FITC	Fluorescein isothiocyanate
FNR	False Negative Rate
FPR	False Positive Rate
FSC	Forward Scatter
GM	Set of analyzed myeloid cells, comprised of granulocytes and monocytes
HSC	Hematopoietic Stem Cells

KLD	Kullback-Leibler Divergence
lin^+	Lineage Positive
lin^-	Lineage Negative
PDF	Probability Density Function
PE	Phycoerythrin
PI	Propidium Iodide
PMT	Photomultiplier tube
QC	Quality Control
RBC	Red Blood Cell
Rho	Rhodamine
SD	Standard Deviation
SP	Side Population
SSC	Sideward Scatter
TNR	True Negative Rate
TPR	True Positive Rate
WBC	White Blood Cell

Preface

The CIHR/MSFHR Strategic Training Program in Bioinformatics at The University of British Columbia was designed to begin with three four-month rotation projects and to conclude with a Master thesis study. During the rotation projects I was exposed to several areas of bioinformatics research and gained experience with several techniques. My first rotation project was under the supervision of Dr. Raphael Gottardo and Dr. Michael Kobor, in this project, I was introduced to the field of chromatin structure modifications and their analysis by ChIP-on-chip experiments. During the second rotation I worked under the supervision of Dr. Artem Cherkasov in the field of cheminformatics. Here I gained experience in the drug discovery pipeline, both with computational techniques and with *in vivo* validation techniques of candidate computational hits. For the third and summer rotation, I had the opportunity to work with Dr. Cherkasov's colleague Dr. Mikhail Gelfand at the Institute for Information Transmission Problems in Moscow, Russia. In this project I gained experience with phylogenetics and comparative genomics analyses. The work presented on this thesis corresponds to the work undertaken during my Master thesis project under the supervision of Dr. Ryan Brinkman; hence, the work accomplished during the rotations projects is not represented here.

Acknowledgements

I feel very grateful to have had the opportunity to study in the bioinformatics program. During this time I have met admirable people who have inspired me to be a better student and a better person. Among these, I would like to specially thank:

My supervisor Dr. Ryan Brinkman. Ryan's support during this time has been invaluable. Thanks Ryan for your encouragement to pursue my ambitions, for always being ready to give your good advice and for always having the well-being of your students at the forefront.

Dr. Jennifer Bryan and Dr. Connie Eaves, my thesis committee members, for valuable comments during our meetings and for reading this work. My rotation supervisors: Dr. Raphael Gottardo, Dr. Michael Kobor, Dr. Artem Cherkasov and Dr. Mikhail Gelfand for providing the rotation projects I worked on and for their worthy guidance and teachings.

Members of the Team Mouse, specially Dr. Connie Eaves (again!), Dr. David Kent and Dr. Claudia Benz, for introducing me to the field of hematopoietic stem cells, for their constant support and help when required, and for allowing me to be part of the admirable research they accomplish, pivotal in furthering the understanding of this exciting field!

Members of the Brinkman lab: Nina Aghaeepour, Ali Bashashati, Melanie Courtot, Alireza Khodabakhshi, Kieran O'Neill, Parisa Shooshtari, Josef Špidlen and Habil Zare. They have all contributed to a productive and enjoyable learning environment and I thank them for that.

All the administrative staff at the BCCRC and GSC. Special thanks to Ms. Sharon Ruschkowski, the bioinformatics program coordinator, for her always prompt assistance.

The CIHR/MSFHR Strategic Training Program in Bioinformatics, and all the faculty members that make this program possible, for providing a great learning environment and to all its students, from whom I learned many valuable lessons.

My studies were possible thanks to the funding received from the CIHR/MSFHR Strategic Training Program in Bioinformatics and Genome BC (Technology Development Grant).

A mis padres y a mi hermano...

gracias por su entrega y su amor.

Siempre admiraré su lucha continua.

Chapter 1

Introduction

Flow cytometry (FCM) is a technology that allows the rapid quantification of physical and chemical properties of up to millions of cells in a sample. It is a technology commonly used in drug discovery, health research, medical diagnosis and treatment, and vaccine development. In particular, it is an essential tool in the area of stem cell research, the application of which is addressed in this thesis, where flow cytometry is used for the characterization of hematopoietic stem cells (HSC). Scientists routinely perform experiments where murine HSCs are isolated and their properties are studied by engrafting a single or a group of HSCs (referred to as the donor HSC(s)) to a mouse (the recipient). FCM is then used to measure the repopulation kinetics of the donor HSC(s) in the recipient mice. In a typical experiment, about 32 HSCs are transplanted and measurements of donor-derived cells in the recipient's blood are performed at up to five time points over a period of 36 weeks. Multiple experiments are running at one given time and researchers run their own set of experiments. The time required to analyze the samples increases to a significant amount, where experimentalists analyze over 100 samples per week. The conventional method to analyze flow cytometry data requires visual inspection and manual analysis of the data by the researcher, a process that is labour intensive, subjective, not reproducible, error prone

and not standardized.

I developed a data processing pipeline for the automated analysis of such HSC experiments. This works alleviates the efforts and time dedicated for the analysis of the data. The analysis framework presented matches the results obtained when the data is traditionally analyzed and provides quality control reports that facilitate the detection of experimental errors promptly. A description of the analysis pipeline is presented in Chapter 2.

A second component of this thesis is a meta analysis of all experiments performed from 2005 to 2009. The main objective of this exercise was to study the possibility of predicting which engraftments were not going to be of significance for the hypotheses being tested. As will be discussed, hundreds of assessments are currently performed in such experiments, of which it is known *a priori* that only small a proportion will be informative, but not which particular ones. Proper prediction of mice that will be positive would reduce costs, efforts and time when performing the experiments, and reduce the allocated space used for the animals in the animal resource centre (ARC). This analysis is presented in Chapter 3.

1.1 Flow Cytometry (FCM)

FCM began in the 1960's with the combination of emerging technologies: microscopy and electronics. The first commercialized flow cytometers were developed in the late 1960's by Van Dilla *et al.* and Dittrich and Göhde (Shapiro and Leif, 2003). In 1976, Herzenberg and Sweet (1976) reported on

the first instrument to combine fluorescence flow cytometry with cell sorting, commercialized as the FACS (fluorescence-activated cell sorting) by Becton-Dickinson. In the decades to come, flow cytometry continued to adapt newer technologies and advancements; today, flow cytometers are capable of measuring up to 18 fluorescence signals with the aid of recently developed quantum dots (Chattopadhyay et al., 2006).

1.1.1 The Technology

Typical FCM machines are capable of measuring four to six fluorescence signals using two lasers (Figure 1.1). Cells, which have previously been stained with fluorescent conjugated antibodies (*e.g.*, anti-CD5-PE for quantifying murine T cells or propidium iodide (PI) for quantifying DNA content or for cell cycle analysis) are introduced in the fluidics system of the flow cytometer. Hydrodynamic forces align the cells in tandem past the light sources (*i.e.*, the two lasers) at rates of up to 70,000 cells per second. Upon contact with the laser beams, the cells will scatter some of the light and also absorb it and then re-emit it. The scattered light is measured by at least two detectors: the forward scatter (FSC) detector measures the amount of light diffracted in the forward direction (see Figure 1.2) and this is proportional to cell size. The side scatter (SSC) detector measures the amount of light reflected by the cells and it is proportional to cell granularity or internal complexity. Light being absorbed and then re-emitted at lower frequencies is detected by an additional set of photomultiplier tubes (PMT), each of which is configured to measure a particular range of wavelengths with the use of

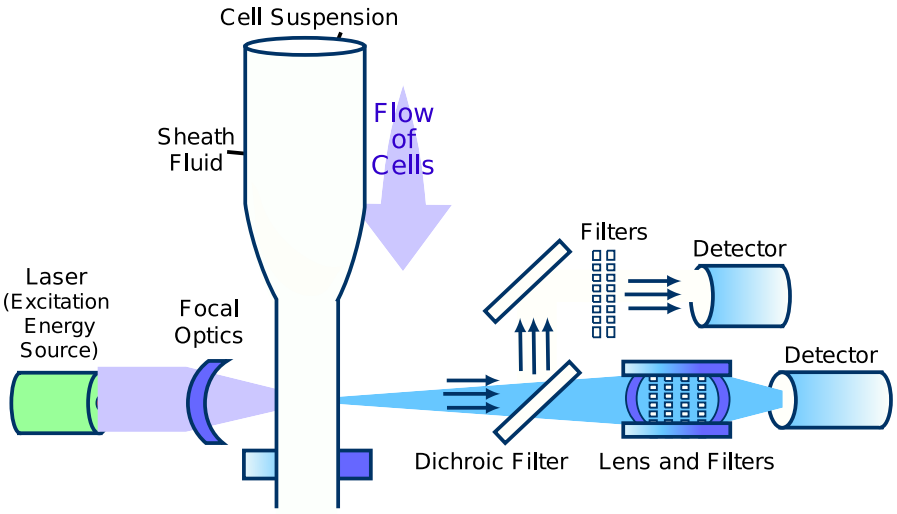


Figure 1.1: Schematic diagram of a typical flow cytometer. Liquid flow moves the suspended cells through the fluidics systems and past the lasers. When the laser beam makes contact with a cell the light gets either scattered or absorbed and then re-emitted. Scattered light is measured by the forward scatter (FSC) and sideward scatter (SSC) detectors. Emitted light (fluorescence) is collected by the photomultiplier tubes (PTM), each of which measures a particular range of wavelengths. (Image adapted from The Science Creative Quarterly (`http://www.scq.ubc.ca/`).)

filters. The fluorochrome-antibody pairs are designed to bind to particular proteins on the cell cell surface. When light makes contact with a cell (depicted in Figure 1.3), the fluorochrome absorbs the light and then re-emits the light (fluoresces) at lower frequency. The amount of fluorescence detected is proportional to the amount of protein in the cell (Shapiro and Leif, 2003).

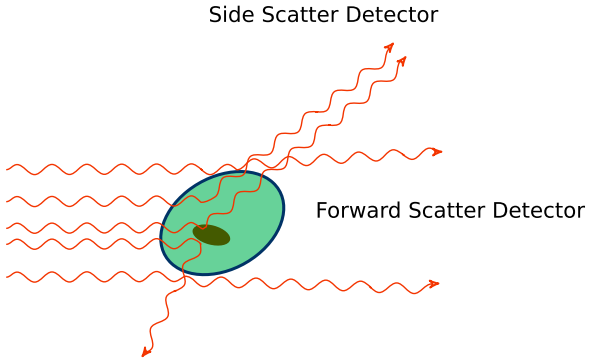


Figure 1.2: Light scattering from a cell. Light diffracted in the forward direction is proportional to cell size, while light reflected is proportional to cell granularity or internal complexity.

Compensation

The flow cytometer used to generate the data described here has two lasers and was configured to measure the emission of four fluorochromes. Each fluorochrome has a range of wavelengths at which it can absorb light and at which it can emit light. For example, Fluorescein isothiocyanate (FITC), Phycoerythrin (PE) and PI are all excited with the 488 nm laser but Allophycocyanin (APC) needs to be excited with a 630 nm laser. After a fluorochrome absorbs

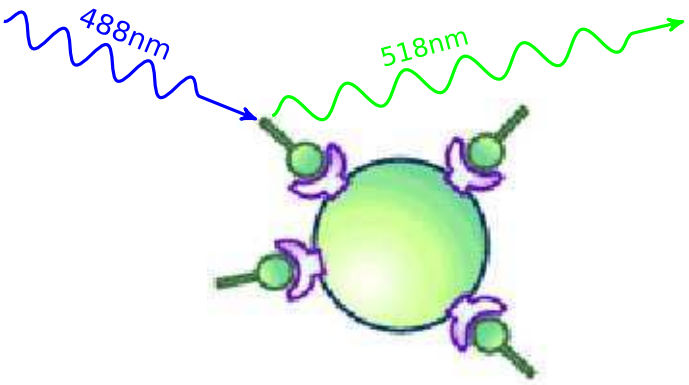


Figure 1.3: Fluorochrome emission. Cells are treated with a solution containing fluorochrome-conjugated antibodies which bind specifically to a cell surface marker (in purple) or other biological markers. The fluorochrome absorbs light (shined from a laser) and re-emits the light at a lower frequency. (Image adapted from The Science Creative Quarterly (`http://www.scq.ubc.ca/`).)

light, it re-emits the light at different wavelengths. The emission wavelength distributions are shown if Figure 1.4 for those fluorochromes excited with the 488 nm laser. For each fluorochrome, emission intensities are measure for a range of wavelengths with the aid of light filters. In Figure 1.4 the wavelength regions used to measure emission for the three fluorochromes are shown on top of their emission distributions. Since emission distributions overlap to some extent, detectors measuring emission intensities for one fluorochrome will detect the emissions of other fluorochromes as well. This is clearly seen in the region corresponding to the filter measuring the PE emission intensities (yellow colored). This detector will mainly measure PE emission intensity but will also detect some FITC and PI emissions. Compensation refers to the process of correcting the measured intensities by subtracting the overlapping emission from other fluorochromes.

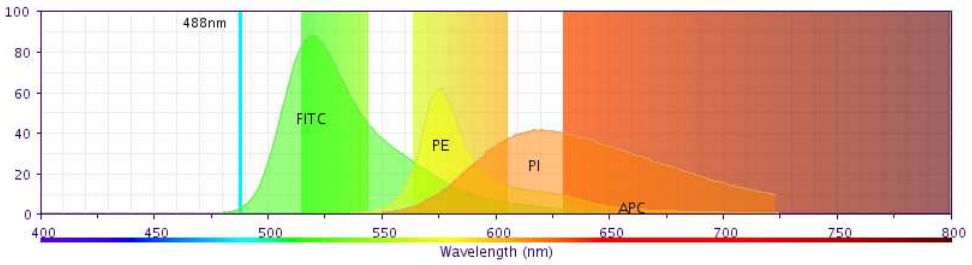


Figure 1.4: Fluorochrome emission distributions when excited with a 480 nm laser. Each fluorochrome emission is measured in the range of wavelengths represented by shaded areas behind the emission distribution. Image generated with Fluorescence Spectrum Viewer from BD Biosciences.

1.1.2 FCM Bioinformatics

The field of FCM bioinformatics provides the necessary technology for FCM data analysis; comprising of tools for data management, statistical analysis and interpretation of flow cytometry data. The past decades have seen dramatic advancements in the hardware components of flow cytometers (*e.g.*, lasers, amplifiers and detectors) and in the reagents and fluorochromes used (*e.g.*, monoclonal antibodies and quantum dots). Recent advancements in reagents and fluorochromes include quantum dots (Chattopadhyay et al., 2008) that extend the repertoire of available fluorochromes for flow cytometry. Quantum dots have narrower emission spectra than conventional fluorochromes and current reports describe cytometers capable of resolving 17 colours (Chattopadhyay et al., 2006). Lagging in the development and adaptation of FCM technologies is the data analysis component. Traditional manual analysis, which involves the identification of cell populations by manually drawing gates (*i.e.*, data filters) in one or two dimensional graphical repre-

sentation of the data, does not scale with newer technologies. A typical four colour experiment, identifying positive and negative populations, can generate $2^4 = 16$ potential populations for analysis. Newer technologies provide the potential to study $2^{17} = 131,072$ subpopulations, an immense task to complete with conventional analysis. This opens the possibility to use FCM as a non-hypothesis driven technology making plausible the discovery of unforeseen relationships that would not have been considered otherwise, such as the discovery of new T cell subtypes and their function (*e.g.*, see De Rosa et al. (2001)). Another reason for the development of flow cytometry data analysis tools originates with the traditional method to analyze FCM data. To analyze FCM data, the researcher needs to visually inspect the data and manually select the population boundaries. This methodology has several drawbacks. It is time-consuming for the researcher: the experiments mentioned in this thesis take at least an hour to analyze, and this can be required several times a week. Since the group boundaries are manually placed, the analysis is subjective to how confined the populations are defined by the researcher. Due to this subjectivity, this analysis is also non-reproducible, error prone, non standardized and not open for re-evaluation. For these reasons, in the past few years there has been an increasing interest to develop new data analysis techniques that will exploit the full potential of modern flow cytometers (Lizard, 2007) and that will provide standardized, reproducible and objective analyses.

Recent efforts have concentrated on clustering techniques; statistical methods for identifying homogeneous groups within FCM data. Indeed, in the past few years, FCM data analysis has been the motivation of new statisti-

cal clustering techniques, and most likely additional new methods will come. Clustering is a widely used method for numerous data analysis problems, and several groups have applied common techniques from other fields to analyze FCM data (*e.g.*, K-means (Murphy, 1985; Schut et al., 1993), Gaussian mixture models (Chan et al., 2008), and several classification techniques: classification and regression trees (Beckman et al., 1995), neural networks (Kothari et al., 1996) and support vector machines (Morris et al., 2001)). Also, the need for more adequate techniques has resulted in the development of new methods: such as the ones presented in the software packages flowClust (Lo et al., 2008), flowMerge (Finak et al., 2009) and Flame (Pyne et al., 2009), which all use a form of mixture models; probability binning (Roederer et al., 2001); and an adaptation of spectral clustering (Shooshtari et al., 2009). With all of these clustering techniques accumulating in the toolbox it is still unclear which, if any, is more appropriate for the analysis of FCM data. The FlowCAP project (FlowCAP, 2009) aims at providing a comparison and assessment framework for all of these methods, and methods to come. But until such conclusions are available, it remains up to the individual scientist to decide which method to apply.

Of the methods developed for FCM data analysis mentioned above, those using mixture models have caught the most interest from statisticians working in the field of FCM bioinformatics. Mixture models use the assumption that the data can be conceptualized as observations from a probability distribution function in a N dimensional space, where N can be the number of parameters measured by the flow cytometer (2 for the light scatter parameters and $N - 2$ for the number of fluorochromes measured). The different

methods differ on which distribution function and data transformation they use to model the data, which method is used to infer the model parameters and how the number of homogeneous groups is determined. The first mixture model method to be suggested was by Lo et al. (2008), which used a Student-t distribution with a Box-Cox transformation. Shortly after, Chan et al. (2008) proposed the Gaussian mixture model. The Student-t distribution is similar to the Gaussian distribution but it has wider tails: this allows the t distribution to be more robust to outliers, which are commonly seen in FCM data. The Box-Cox transformation incorporated into the methodology presented by Lo et al. (2008), allows the identification on non-elliptical cell populations, a limitation presented in the methodology by Chan et al. (2008). Inherited from the mixture model framework, these techniques require the assumption, *a priori* to learning the model parameters, of the number of homogeneous groups in the data, usually represented with the symbol K. Since the number of homogeneous groups, or cell populations, is not always known, and sometimes there is the interest to infer them from the data, there are several methods to compute the parameter K from the data. A popular method is to model the data with multiple values of K, say from 1:15, and based on the results, the best model is chosen using the criterion known as the Bayesian Information Criterion (BIC) (Schwarz, 1978). This criterion minimizes the tradeoff between the goodness of fit to the data and the complexity of the model.

Apart from clustering techniques for FCM data, several groups have concentrated their efforts in techniques for quality control of data from high-throughput flow cytometers and in data visualization. A lot of these efforts

have been materialized in open source software packages written in the statistical programming language R. Within this framework, there is a suite of packages that provide diverse functionality: data visualization in *flowViz* (Sarkar et al., 2008), normalization in *flowStats* (Hahne et al., 2009a), and quality control and data management in *flowCore*, *flowQ*, *plateCore* and *flowUtils* (Errol et al., 2009; Hahne et al., 2009b).

1.2 Hematopoietic Stem Cells (HSC)

1.2.1 Background

In an adult human, the hematopoietic system requires the production of 1.5 $\times 10^6$ blood cells every second (Bryder et al., 2006). At the top of the homeostatic control mechanism lies a rare population of cells with the properties of self-renewal and multipotentiality; these cells are called hematopoietic stem cells (HSC). The self-renewal property refers to the ability of a cell to maintain its undifferentiated state throughout the cell division cycle; in this manner creating identical copies of itself. This type of cell division is called a symmetric self-renewal cell division. The other type of self-renewal cell division is called asymmetric: in this case, a HSC will produce at least one daughter cell which has irreversibly reduced self-renewal potential. These daughter cells will thus differentiate to mature cells within a finite number of divisions. The multipotency property refers to the ability of HSCs to give rise to all cell types in the hematopoietic system through several rounds of asymmetric cell divisions (see Fig. 1.5).

The study of HSCs started more than 50 years ago (*e.g.*, Till and McCul-

loch (1961)) and since then, much effort has been placed in characterizing these cells in mice. A crucial and challenging step in the study of HSCs is their isolation, as they are found in the bone marrow at an extremely low frequency, $\sim 0.01\%$. No single set of biological markers exclusively identifies HSCs and it is not clear if all HSCs can be identified by a common set of molecular features (Kent et al., 2007). Determining the HSC purity of an isolated population of cells involves elaborate functional assays that test for stem cell properties of the cells.

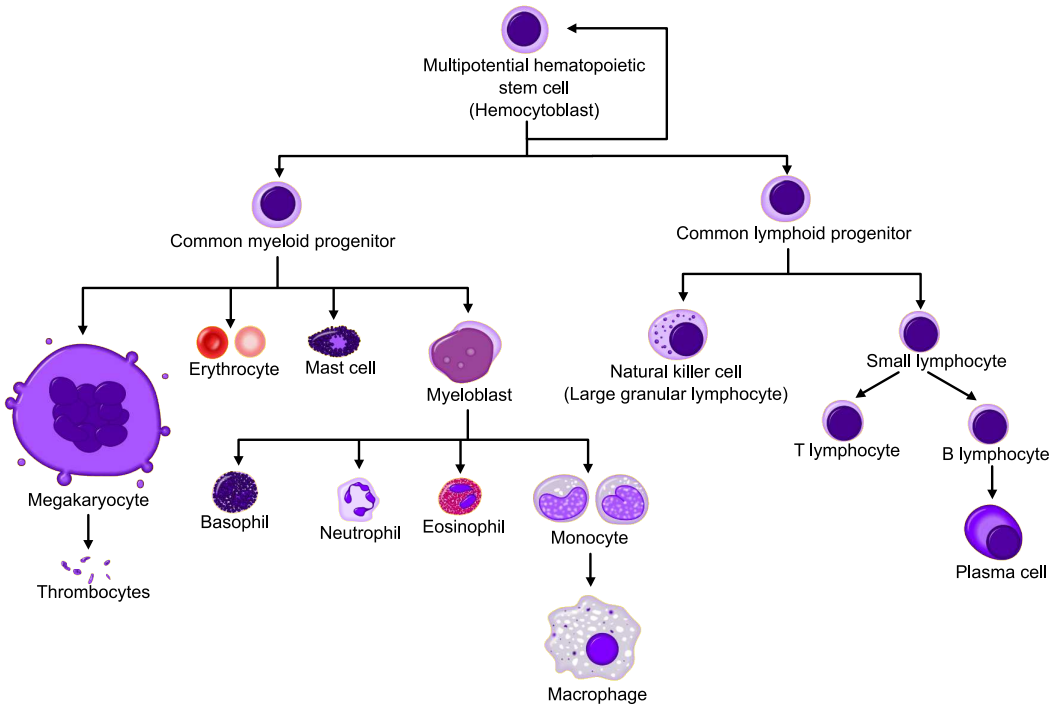


Figure 1.5: Model of the hematopoietic developmental hierarchy. Image from Wikimedia Commons by Mikael Häggström (`http://en.wikipedia.org/wiki/Haematopoiesis`).

1.2.2 Murine HSC Markers and Isolation of HSC

The phenotype of a cell can be characterized, in part, by the expression of biological markers, which can be quantified by flow cytometry. These markers can be, for example, extracellular or cytosolic proteins, or the ability to efflux a particular dye (*e.g.*, Hoechst). When designing a flow cytometry experiment, the experimentalist will choose a particular set of biological markers to quantify and/or isolate the cells of interest. Stem cells can be isolated by a flow cytometer with sorting capabilities, based on the marker-derived phenotype. The set of markers used, and their relative expression, is referred to as the sorting strategy. Several sorting strategies have been suggested for the purification of murine HSCs and to this date, this remains an active area of research. New markers for HSC purification are continually reported and new sorting strategies increase the HSC purity of the isolated cells. Sorting strategies for HSCs are a combination of positive selection for cell surface markers, negative selection for markers expressed on mature hematopoietic cells of different lineages or other non-stem cells, and selection of dye-effluxing side populations (SP) (*e.g.*, Hoechst 33342 and Rhodamine-123). HSCs have the ability to efflux these dyes via membrane transport pumps, which are highly active in HSCs relative to the other hematopoietic cells (Challen et al., 2009).

As an example, the CD45^{mid}lin^{-}Rho^{-}SP sorting strategy (Dykstra et al., 2006) was used to isolate some of the HSCs which were then used to generate much of the data used in this thesis. The CD45 antigen is an enzyme which is also known as protein tyrosine phosphatase, receptor type, C. This protein is a signalling molecule and is specifically expressed in hematopoietic

cells (NCBI Entrez Gene). lin^- stands for lineage-negative; meaning, there is no detectable expression of lineage markers. The cells are stained with a solution that contains monoclonal antibodies that, in combination, stain for most differentiated hematopoietic cells. The third element of the strategy refers to the ability of HSCs to efflux efficiently the dyes Hoechst 33342 and Rhodamine-123 which allows their selection as the negative population. Fig. 1.6 shows an example of a gating strategy for isolating HSCs. This strategy gives HSC purities of at least 25% (Dykstra et al., 2006). Other strategies used in Dr. Connie Eaves' lab are listed in Table 1.1.

Table 1.1: HSC sorting strategies.

Strategy	Reference
$CD45^{mid}lin^-Rho^-SP$	Dykstra et al. (2006)
$CD45^{mid}Rho^-EPCR^+$	Balazs and Mulligan (2003)
$CD45^{mid}EPCR^+CD48^-CD150^+$ (E-SLAM)	Kiel et al. (2005)

1.2.3 Assessment of HSC from Mouse Bone Marrow by their *In Vivo* Repopulating Activity

Retrospective functional assays are performed to assess the HSC purity and properties of the isolated cells. The cells (either a single isolated cell, or a group of them) are transplanted into recipient mice and the cell's ability to produce mature blood cells for prolonged periods of time *in vivo* is measured; thus, assessing for HSC activity in the input cells. The experimental proto-

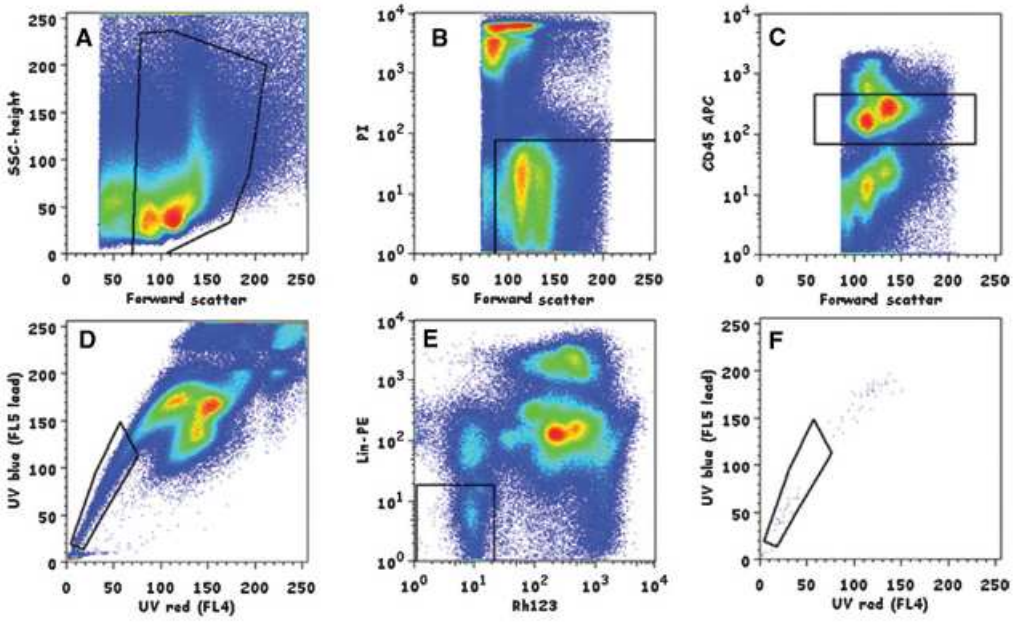


Figure 1.6: Flow cytometric profiles for isolating CD45^{mid}lin^{-}Rho^{-}SP. Stained bone marrow cells are first gated using FSC/SSC (A) and PI (B) to exclude debris, erythrocytes, dead cells, and cell clumps. Gates are then set around the CD45^{mid} (C), SP (D), and lin^{-}Rho (E) populations. (F) In combination, these five gates select for 0.004% of the original bone marrow cells. Image taken from Kent et al. (2007).

col, in full detail, can be found in Kent et al. (2007).

Briefly, for the data analyzed here (summarized in Table 1.3), HSCs were isolated from the bone marrow of healthy mice, single cells, were injected into a recipient mouse who has been sub-lethally irradiated to compromise its own production of blood cells. The number and proportions of myeloid (granulocytes/monocytes, GM) and lymphoid (B cell and T cell) cells were then measured with a flow cytometer 4, 8, 12, 16, and 24 weeks after transplantation. To distinguish the origin of the WBCs, donor and recipient mice congenic[1] on the *Ly5* (also known as CD45) locus were used. Thus, in a given experiment, the donor mouse would be CD45.1 positive and the recipient mouse CD45.2 positive, or vice-versa.

To analyze the WBC content of the recipient mouse, a sample of peripheral blood was collected and divided into three aliquots (also referred as tubes in flow cytometry). The cells in each tube were stained with Propidium Iodide (PI), a viability marker; two antibody-fluorochrome pairs for distinguishing donor and recipient CD45 (*Ly5*) allotypes: anti-CD45.1-allophycocyanin [APC] and anti-CD45.2-fluorescein isothiocyanate [FICT]; and then each tube of cells was stained with either one of anti-Ly6g-phycoerythrin [PE]/anti-Mac1-PE for myeloid (GM) cells, anti-B220-PE for B cells and anti-CD5-PE for T cells. Table 1.2 summarizes the content of the flow cytometry data.

Analysis of Peripheral Blood FCM Data

Flow cytometry data is typically manually analyzed using commercial software such as FlowJo (Tree Star Inc., Ashland, Oregon). FlowJo is an in-

[1]Refers to two strains of mice who are genetically identical except for one locus

Table 1.2: Description of staining cocktails.

Channel	Fluorochrome	Antibody	Function
FL1	FITC	anti-CD45.2	Donor/Recipient Derived Cells
		anti-Ly6g/Mac1	Granulocytes and Macrophages
FL2	PE	anti-B220	B cells
		anti-CD5	T cells
FL3	PI		Viability marker
FL4	APC	anti-CD45.1	Recipient/Donor Derived Cells

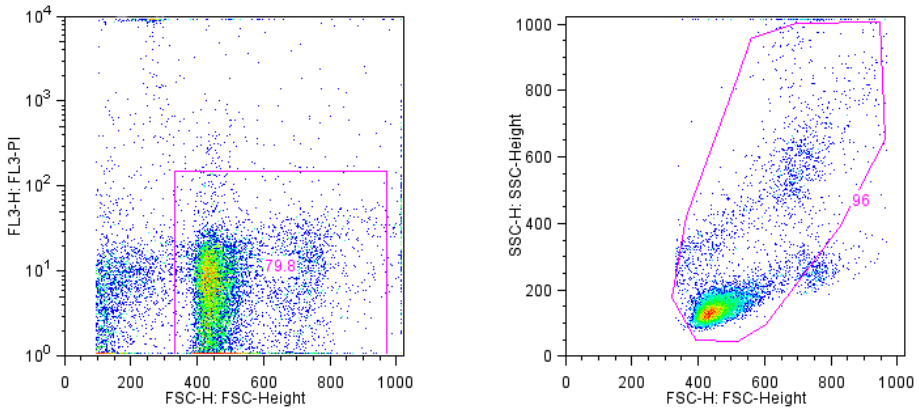
Table 1.3: Data used in the development and validation of the automated analysis pipeline

Phenotype of HSC	Experiment	Weeks	Number	Use	Reference
$CD45^{mid}lin^{-}Rho^{-}SP$				development	Dykstra et al. (2007)
E-SLAM	OC2908	16	15	development	communicated by Dr. Claudia Benz
E-SLAM	JL0307	4,8,16,24	27	development	communicated by Dr. David Kent
E-SLAM	SE1707	8,16,24	30	development	communicated by Dr. David Kent
E-SLAM	AP0609	8,16	35	validation	communicated by Dr. Claudia Benz
E-SLAM	AP2908	8	24	validation	communicated by Dr. Claudia Benz
E-SLAM	DE0407	4,16,24	48	validation	communicated by Dr. Claudia Benz
E-SLAM	FE0508	16	38	validation	communicated by Dr. Claudia Benz
E-SLAM	FE0509	8,16	38	validation	communicated by Dr. Claudia Benz
E-SLAM	FE1709	8,16,24	37	validation	communicated by Dr. Claudia Benz

teractive software with advanced data visualization capabilities. The user loads data into the workspace and visualizes the data in 2-dimensional scatter plots. Cell populations are identified visually by the user in the plots and cell populations are labelled and selected by drawing a boundary around them, which is referred as gating. This gate is then used to filter the population for subsequent analysis. When analyzing many files, the gates are usually determined for one file (usually a control) or taken from a template and then these gates are applied to the rest of the corresponding files. For the peripheral blood analysis referred here, three gating templates were determined for each staining cocktail: those detecting GM, B and T cells.

The first step of the analysis consisted of filtering out debris and dying or dead cells. The identification of dead cells or dying cells is commonly accomplished with the aid of a viability marker, in this case PI. It is important to remove these cells because they can show different patterns of fluorescence or non-specific antibody binding compared to intact viable cells. PI is a non-permeant dye that can penetrate the membrane of dying or dead cells, intercalating into DNA or RNA molecules (for this reason it can also be used for cell cycle analysis or quantifying DNA content) (Shapiro and Leif, 2003). Events with high PI intensity values are hence removed from further analysis. Before flow cytometry analysis, the samples are usually lysed to remove red blood cells (RBC), as a consequence, samples can contain a substantial amount of cellular fragments, referred to as debris. Debris events are easily identified by their relative lower FSC and SSC values.

The filtering of viable cells is accomplished in two gating steps. First, a gate is set with the FSC and PI values (Figure 1.7(a)) and secondly, a gate is



(a) (b)

Figure 1.7: Gating strategy for selecting viable cells.

set with the FSC and SSC values (Figure 1.7(b)). Events inside these two gates are considered viable cells. Once viable cells have been gated, the next steps consist of classifying each cell as a donor or recipient cell, and determining whether they are GM, B or T cells, depending on the tube being analyzed. This is done by selecting the cells expressing the corresponding antigen. Donor and recipient populations are identified by gating on the CD45.1 and CD45.2 intensity values (Fig. 1.8) and cell lineage is determined by the intensity values corresponding to the cell lineage antibody (Figure 1.9). A full detailed protocol of the analysis can be found in Kent et al. (2007).

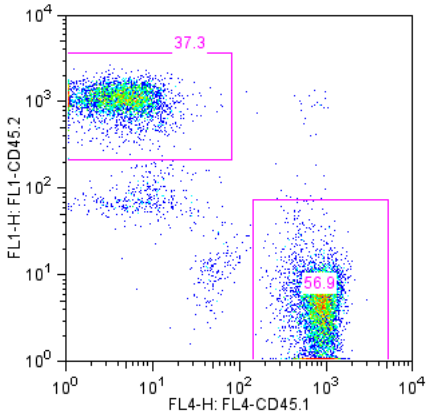
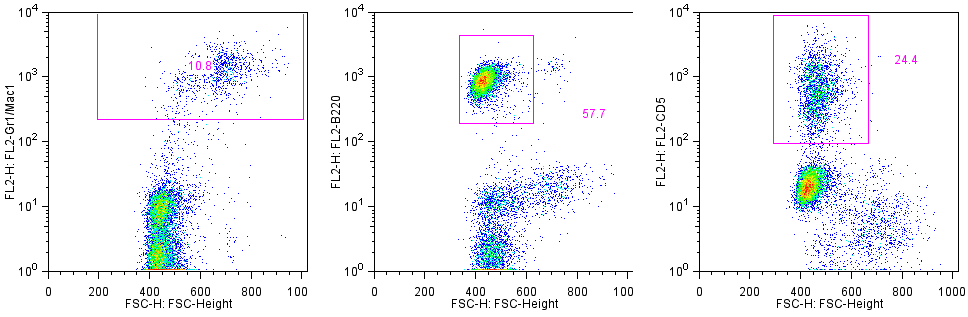


Figure 1.8: Manual gating strategy for selecting donor and recipient derived cell populations. Donor derived cells have the CD45.2 allotype and recipient derived cells have the CD45.1 allotype.



(a) (b) (c)

Figure 1.9: Manual gating strategy for selecting lineage marker-positive cells. Myeloid cells are Gr1/Mac1$^+$ (a), B cells are B220$^+$ (b) and T cells are CD5$^+$ (c).

Chapter 2

Automated Analysis Pipeline

The main objective of the research project described in this thesis was to develop a tool for the analysis of FCM data obtained from the peripheral blood analysis of the HSC transplant experiments listed in Table 1.3 and described previously in Section 1.2.3. As mentioned in the introduction, the traditional analysis of FCM data demands manual intervention at all time from the scientists, which constitutes a major burden and bottleneck in FCM data analysis. In this chapter, I describe the tool that was developed for the automated analysis of FMC data from the HSC transplant experiments.

The proposed analysis framework uses software being developed, in part, by members of the Terry Fox Laboratory (*e.g.*, flowCore (Hahne et al., 2009b) and flowClust (Lo et al., 2008)), is written in the statistical programming language R (R Development Core Team, 2009), and uses the flow cytometry analysis packages from the Bioconductor project (Gentleman et al., 2004). First, the automated gating strategy that I developed is described. This strategy mimics the manual gating strategy outlined in Section 1.2.3 and described in Kent et al. (2007), but the location of the gates are data driven and determined for each sample individually. Then, the quality control reports are described which were designed to identify potential experimental errors; abnormalities in cell lineage proportions, which can be caused by leukemia

in the recipient mice; and possible errors in the automated gating steps. The diagram in Figure 2.1 depicts the sequential steps in the pipeline and where quality of the data is assessed.

2.1 Methods

2.1.1 Clustering Techniques

The methods presented in Lo et al. (2008) were used for the identification of viable cells, since at its time of publication, it showed better performance at cell population identification that other available methods. Since its publication, a few other methods have been proposed using similar techniques but it is still not clear if they provide better results. The method chosen is available through the Bioconductor package *flowClust* (Lo et al., 2009) which makes its integration with other analyses ideal for a full automated analysis pipeline.

2.1.2 Kullback-Leibler Divergence (KLD)

For the purpose of statistical analysis, FCM data can be conceptualized as arising from a probability distribution function. A component of the quality control of flow cytometry data is the comparison of multiple data samples and quantifying the difference between the samples. A metric that allows this quantification is the Kullback-Leibler Divergence (Kullback, 1997), which is a non-parametric estimate of the distance between two probability distribution

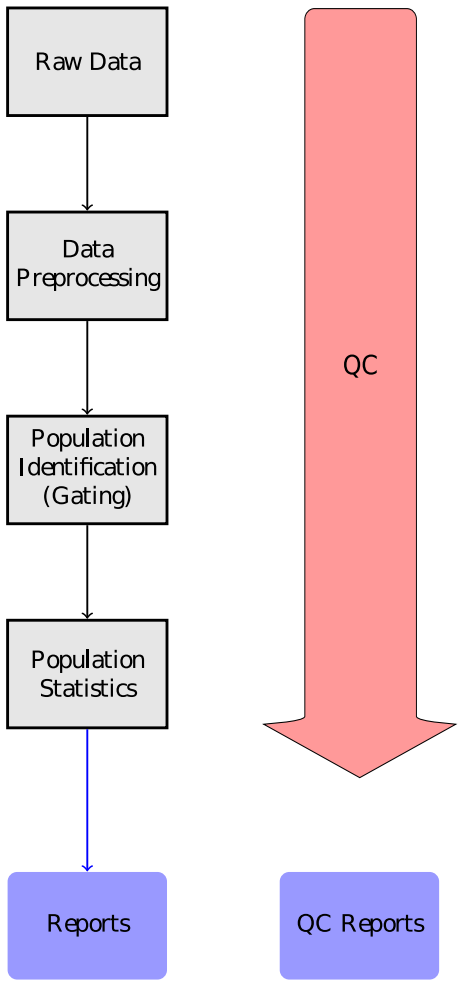


Figure 2.1: Automated pipeline analysis diagram

functions.

Given two probability densities, Q and P, the divergence of Q from P is defined to be:

$$KL(P||Q) = \sum_i P(i) log \frac{P(i)}{Q(i)}, \tag{2.1}$$

if the distributions are discrete, and

$$KL(P||Q) = \int_{-\infty}^{\infty} p(x) log \frac{p(x)}{q()} dx, \tag{2.2}$$

for continuous distributions. Note that the KLD is not a distance metric because it fails the condition of symmetry, *i.e.*, $KL(P||Q) \neq KL(Q||P)$, and does not satisfy the triangle inequality. For these reasons, it is commonly utilized as:

$$D_{KL}(P||Q) = KL(P||Q) + KL(Q||P), \tag{2.3}$$

which is symmetric.

The computation of the KLD has been implemented in the R package *bioDist* from Bioconductor (Gentleman et al., 2004). In the case of observed data arising from continuos distributions, the observed data is binned and the KL distance is computed from the discrete distributions obtained using Equation 2.1. The function *KLdist.matrix* bins the data.

2.1.3 1 Dimensional High Density Region Analysis

Flow cytometry experiments are commonly designed to test for the presence or absence of a particular biological marker. For example, in the experiments listed in Table 1.3, the experiment was designed to provide the data that will facilitate the deduction of which of the CD45 allotypes the cells expressed, whether it was CD45.1 or CD45.2. If a group of cells expressing the

CD45.1 allotype were present in the sample, then a positive population will be detected in the intensity values corresponding to the parameter measuring CD45.1. If there were no cells expressing CD45.1 in the sample, then only a negative population will be detected in the parameter measuring CD45.1 intensities, which corresponds to background emission of the conjugated fluorochrome.

The computational method described in this section was designed to automatically identify the presence of positive and negative populations, and also the case when the cells exhibit a non-specific expression of the antibody-fluorochrome pair, which emitted a mean intensity between the negative and the positive population. Figure 2.2 shows a typical flow cytometry profile of a sample with donor derived cells. In this case, the donor derived cells expressed the CD45.1 allotype and the recipient cells the CD45.2 allotype. There is a group of cells showing a dim expression of either of the CD45 allotypes, these cells expressed non-specific binding of the antibodies.

For each parameter value, populations were located by identifying high density regions in the intensity value distribution. Figure 2.3 shows the histogram of expression values for the detector measuring the CD45.2 antibody (for the same sample shown in Figure 2.2). The probability density function of these values is estimated using a kernel density estimate (Parzen, 1962), shown in orange. High density regions, shown in blue, are defined as those regions with significant curvature, where the second derivative of the kernel density estimate is negative (methodology described in Chaudhuri and Marron (1999)). Significance testing for identifying curvature regions is implemented in the R package *feature* (Duong et al., 2008).

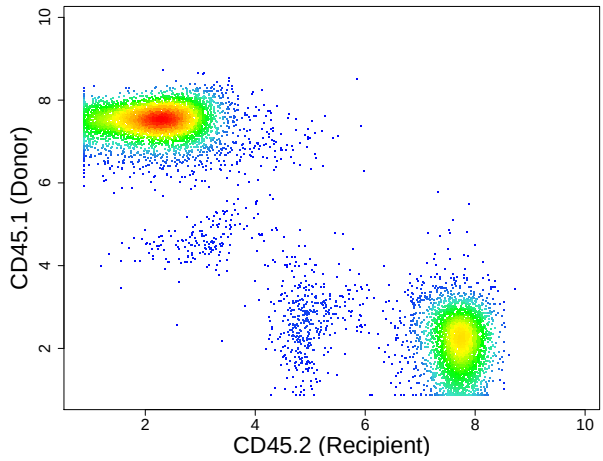


Figure 2.2: Flow cytometry profile showing recipient and donor derived populations. Donor derived cells express the CD45.1 allotype and recipient derived cells the CD45.2 allotype. Cells with dim expression of either antibody are cases of non-specific binding of the antibody.

After identification of high density regions, these are classified either by proximity to pre-specified landmarks or by known order of relative intensity. In this case, the first high density region (leftmost) corresponds to cells not expressing the CD45.2 allotype and showing background intensity, the middle region corresponds to those cells expressing non-specific binding of the CD45.2 antibody, and the rightmost high density region corresponds to cells expressing the CD45.2 allotype (*i.e.*, the recipient derived cells). The boundary of the populations are computed based of how stringent the cell population gates are desired, these can be determined as the site of lowest density between high density regions (marked with red dots in Figure 2.3) or any distance away from the mean of the region, where the distance is a function of the estimated standard deviation of the expression values of the cells cor-

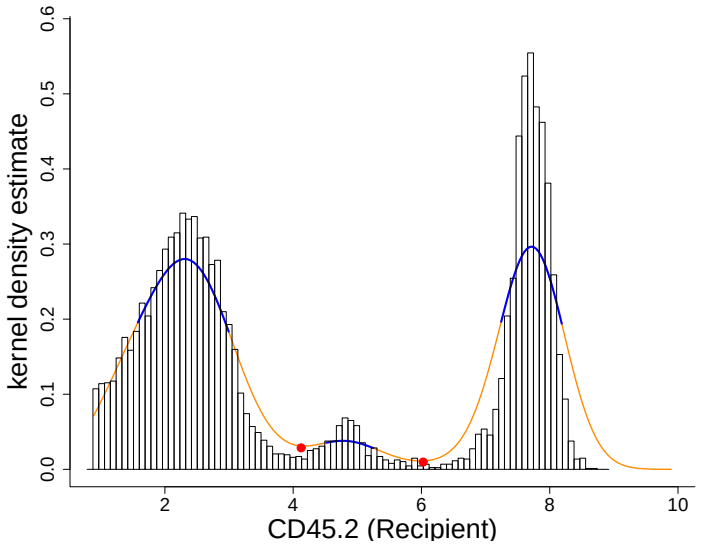


Figure 2.3: Identification of high density regions. The probability density estimate of expression values of the CD45.2 antibody is approximated using a kernel density estimate, shown in orange. High density regions are shown in blue and (•) mark the site of lowest density between high density regions.

responding to the high density region.

The method described above is similar to the function *rangeGate* from the Bioconductor *flowStats* package (Gentleman et al., 2004). This function uses the same methodology to identify high density regions but it only allows the identification of negative and positive populations. If more than two high density populations are found, these are merged to a combined negative or positive population.

2.2 Automated Gating Strategy

2.2.1 Data Preprocessing and Transformation

The fluorochrome intensity value measured by flow cytometers grows exponentially with respect to protein concentration in the cell, hence, intensity values are more-or-less log-normal distributed or a mixture of log-normal distributions (Parks et al., 2006). For this reason, expression values are usually transformed with a log-like function for visualization and analysis purposes. Using a logarithm transformation has the disadvantage that the function is not defined at zero or negative values, which can arise in flow cytometry data after compensation (Roederer, 2001). To properly transform values in all ranges of the data, the inverse hyperbolic sine transformation or its generalization (the Logicle transformation) (Parks et al., 2006) is commonly used. In this analysis, the inverse hyperbolic sine transformation was used on the fluorescence parameters of the data. This transformation is defined at zero and always returns positive values for positive valued arguments. Since this dataset contains no negative values there is no need to use the more extensible Logicle transformation. The FSC and SSC parameters were analyzed as measured by the detectors.

The second preprocessing step was the removal of those events at the lower and upper limit of the data range, also referred to as axis events. An expression value at the lower limit of the data range can occur because either there was no expression of the corresponding protein or the detector failed to measure a signal. Expression values at the upper limit occur when the signal intensity is greater than the measurable capacity of the detector, so

a maximum is recorded. In the preprocessing, axis events were removed for the following parameters: events with FSC and SSC values at the lower and upper limits, events with PI intensities in the upper limit, and events where the intensity values for both CD45.1 and CD45.2 measurements are in the lower limit. Removing axis events facilitated the accurate estimation of distribution parameters.

2.2.2 Identifying Viable Cells

In the manual analysis, three parameters were used for the identification of viable cells: PI intensity, and SSC and FSC values. PI was used to identify possible dead cells and FSC and SSC to discriminate debris and cell clumps. To identify the viable cells, the automated approach used the same parameters, statistics computed in the other parameters, and information extracted from positive control samples. The proportion of viable cells in the samples can range from 100% to about 30% of total events depending in the sample processing steps previous to the data aquisition (*e.g.*, how well RBC were lysed and the supernatant removed (Kent et al., 2007)) and how was the data stored (*e.g.*, sometimes the operator will remove events with low FSC values before writing the data to FCM files).

The first step was to eliminate events with high PI intensity. To do this, I used a technique which was used recurrently in the rest of the analysis pipeline: to identify high density regions and classify them appropriately. The range of values in the PI intensities usually showed one significant high-density region, in these cases the algorithm computed the mode of the distribution arising

from the observed events in the high-density region and estimated its variance. A threshold was set at 1.8 standard deviations and those events above the threshold were considered dead cells. The threshold was chosen with the aid of an experienced researcher in the field. If a sample contained a relatively high proportion of dead cells, the analysis resulted in two high-density regions, where the second one was located in the upper range of the data. This second high-density region was classified as a dead cell population. The split between the two populations was taken somewhere in between the two high density regions, in this case at 30% the distance between the two regions above the lower valued density region (see Figure 2.4).

The second step was to filter events corresponding to debris. Debris fragments are relatively small and were identified by their lower FSC and SSC values. The adopted approach was to identify populations based on cell morphology (homogeneous groups based in FSC and SSC parameters) using the methods implemented in the R package flowClust (Lo et al., 2009) (description in page 22). The data was clustered for the set $K \in [4, 15]$ and all parameters were left as default except for the parameter *level*, the threshold quantile used to call a point an outlier, which was set to 0.98. The best model was selected using the BIC criterion. Once the optimal model was obtained, the populations were classified as viable or debris. For each population obtained, high-density regions were computed in the intensity values for the expression of CD45.1 and CD45.2, and these were classified as being above or at background expression level (see Figure 2.5). If the proportion of events that fall above background in either of the CD45.1 or CD45.2 measurements were above 20% then the population was labelled as a viable cell

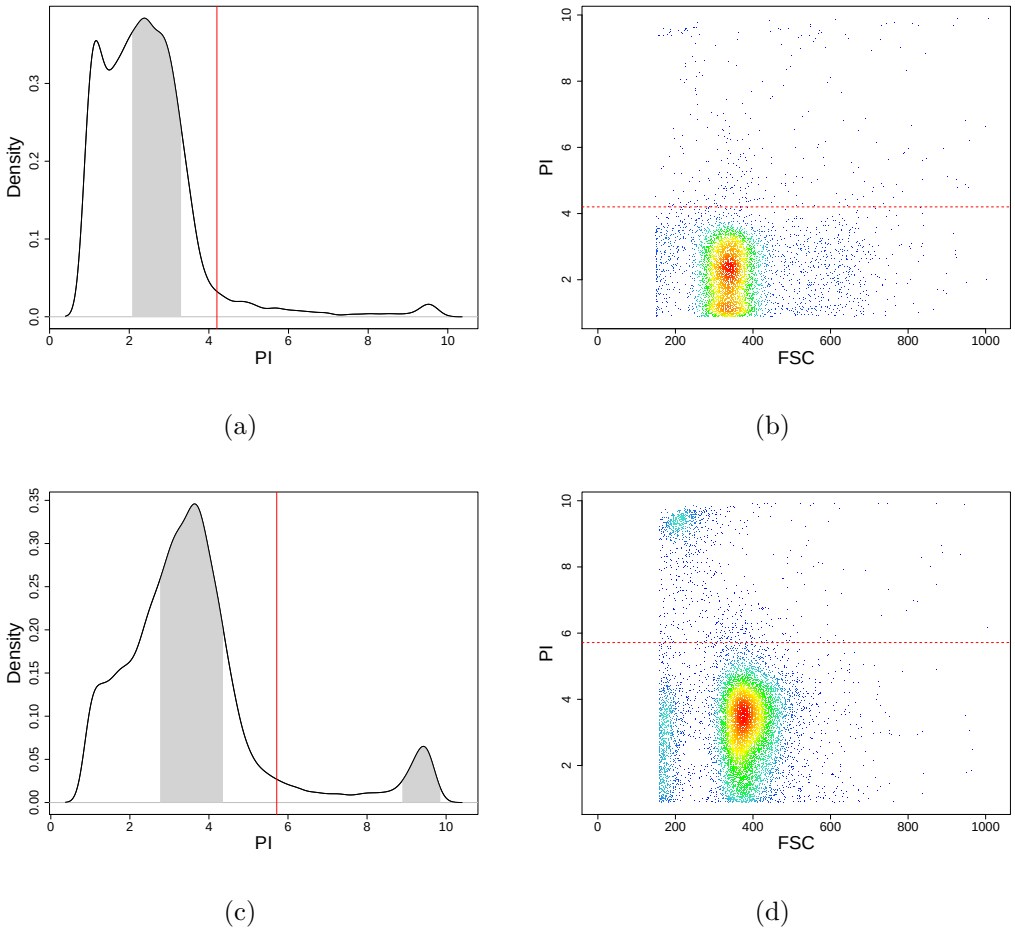
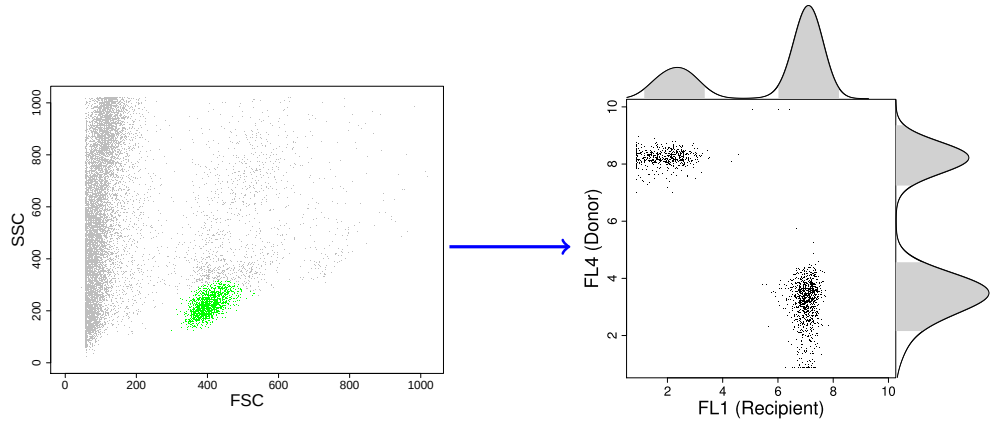


Figure 2.4: First gating step in the identification of viable cells. In the first sample, (a) and (b), only one high density region was identified (in gray), in the second sample, (c) and (d), two high density regions were identified. In the first case a gate was determined 1.8 sd from the mean of the high density region, for the second case a gate was determined 30% down the distance between the two regions. PI expression values are shown in the inverse hyperbolic sine scale.

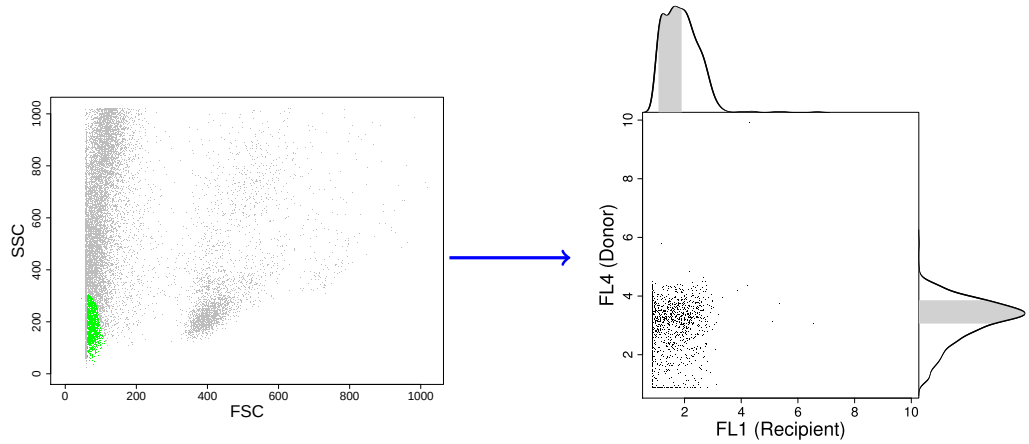
population. The rationale for adopting this approach was that a method was needed that would identify the populations throughout all experiments. The location and proportions of the debris and viable populations were not consistent throughout a series of experiments and sometimes the files were trimmed so that low FSC-H events are not written to file.

The method described above was accurate at gating the relevant events for the down-stream analysis but had the disadvantage of being computationally expensive. For a typical data sample, containing about 50,000 events, the above procedure took 9 minutes on an 8 core 2.93 GHz Intel machine with 12 Gb of RAM or 31 minutes on a dual core 2.16 GHz Powerbook Apple machine with 1 Gb of RAM. Repeating this procedure on the 90 samples that a typical experiment generates was inconvenient. To overcome this limitation, a statistical classifier was inferred for the identification of viable cells. The previous method described was used to gate three samples, the positive controls, and with the results obtained the classifier was inferred and applied to the rest of the data files. This procedure accelerated the identification of viables cells to less than a minute per data file.

Using SVM to learn the position of debris and dead cells. The approach adopted for this task was to use a SVM classifier. This procedure is valid under the assumption that all settings in the flow cytometer remain unchanged during data acquisition. If this is the case, the locations of viable cell events and debris events remain consistent through the set of samples analyzed collectively. After gating the positive controls with the above pro-



(a)



(b)

Figure 2.5: Classification of homogeneous populations based of morphology parameters. For each population returned by the clustering algorithm, high density regions were identified in the CD45.1 and CD45.2 intensity values. If population events were above background in either of the CD45.1 and CD45.2 parameters then the population was labelled as viable (a), otherwise the population was labelled as debris (b).

cedure, those events identified as viable cells were labelled as class 0 and those identified as dead cells or debris were labelled as class 1. The classifier was inferred with the function ksvm in the R package kernlab (Karatzoglou et al., 2004), using default parameters. The classifier was then applied to the rest of the samples, in this manner the gating procedure of viable cells was much faster.

2.2.3 Identifying Cell Types

After gating the viable cells, the next step was to infer, for each cell, the CD45 allotype and whether it was expressing one of the markers for Granulocyte/monocytes, B cells or T cells. From the positive controls, estimated mean CD45.1 and CD45.2 expression values, referred to as landmarks, for the donor- and recipient-derived cell populations, the non-specific binding populations (if present), and the background intensity were derived. These landmarks were then used to guide the classification of high density regions and computation of population boundaries in the rest of the samples.

First, since cells expressed only one of the CD45 allotypes, the background median intensity of fluorescence from one antibody (either CD45.1 or CD45.2) was derived from the intensity of fluorescence of this antibody in the population of cells expressing the opposite allotype. For example, Figure 2.6 shows the intensity values of the two allotypes in a positive control sample. Here both donor and recipient derived cell populations were present and their location was obtained by locating the high density region with the strongest fluorescence intensity in the respective antibody fluorescence in-

tensity values. The mean background fluorescence intensity for the CD45.2 antibody was obtained from the mean CD45.2 fluorescence intensity of the donor-derived cell population (*i.e.*, the $CD45.1^{+}CD45.2^{-}$ population). As can be seen from the kernel density estimates in the two axes in Figure 2.6, the curvature significance test failed to identify the non-specific binding populations. This occured because the relative cell count on these groups was much lower than the other cell groups. To accurately determine the landmark for these groups, another search for high density regions was performed but this time the background population was removed. Figure 2.7 shows the high density estimate results for the CD45.2 fluorescence intensity values after background removal. This time the desired landmarks were obtained, shown by blue dots. This was repeated for the CD45.1 fluorescence intensity values to obtain the full set of landmarks.

After the set of landmarks was identified, the same procedure was applied to the rest of the samples in the experiment, but this time high density regions were classified based on their proximity to the obtained landmarks. In Figure 2.8 a sample is shown which was weakly repopulated by the donor HSC and high density region estimation is shown for each of the CD45 allotype fluorescence intensity values in Figure 2.8(b) and 2.8(c). In the CD45.2 fluorescence intensity values, only one high density region was identified, which corresponds to the recipient-derived cell population (by proximity to the derived landmarks). The location of the population boundary was found by searching the flanking regions of the high density region for the place of minimal density before the gradient of the kernel density estimate starts to increase, depicted by a red dot in Figure 2.8(b). This procedure was repeated

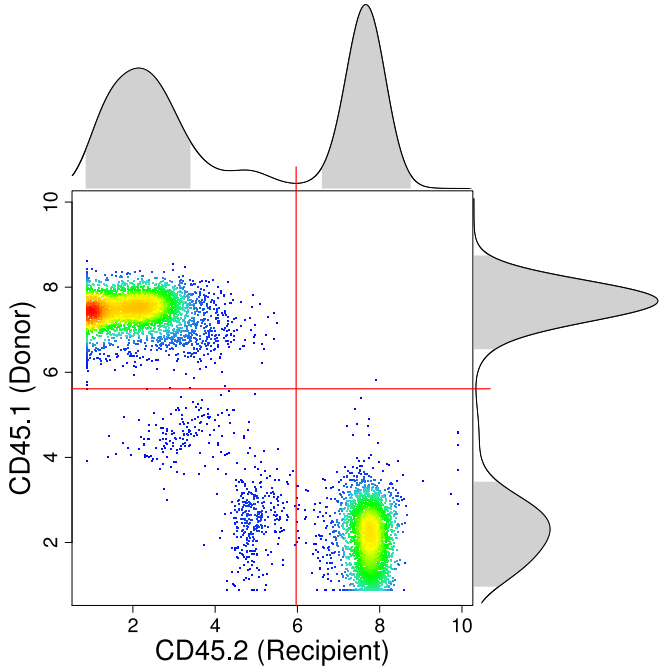


Figure 2.6: Locating landmarks for cell populations from positive control samples. The kernel density estimate is shown for each of the CD45 allotype intensity measurements on the corresponding axis. Red lines marks the location of lowest density at each kernel density estimate and the population boundary for the recipient derived cell population, on the x-axis, and the donor derived cell population, on the y-axis. The two cell populations in the double negative quadrant are referred as non-specific binding cell populations.

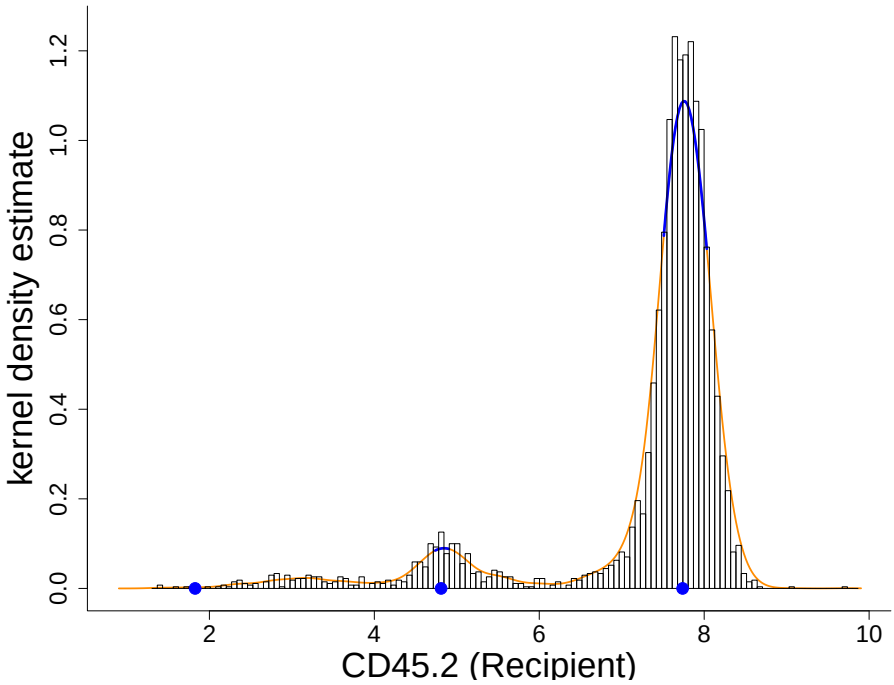


Figure 2.7: Locating landmarks for cell populations from positive control samples. Histogram and kernel density estimate of CD45.2 (Recipient) intensity values after removing donor derived cell population (background emission). Blue regions depict high density regions and (•) the derived landmarks.

for the CD45.1 fluorescence intensity values (see Figure 2.8(c)), in this case two high density regions were identified: the donor-derived cell population and the background emission population.

After the delineation of the recipient and the donor derived cell populations, these two were merged for the identification of the subset that expressed each cell type marker (*i.e.*, Ly6g/Mac1 for myeloid cells, B220 for B cells, and CD5 for T cells). This was achieved in an analogous manner as before, the high density regions were identified and proper separation was localized. Figure 2.9 shows the fluorescence intensity values of the combined recipient and donor-derived cells for the B220 B cell marker.

2.3 Quality Control

Quality control (QC) of raw data and processed data is essential when implementing high-throughput technologies and automated data analysis. Proper QC enables promptly detection of experimental errors, giving the experimentalist the option to restart the experiment before valuable samples are wasted or stop the experiment before time and reagents are employed ineffectively. The QC incorporated in the analysis pipeline was built using the framework provided by the Bioconductor package *flowQ* (Gentleman et al., 2004). This package provides the data structures, the scaffolding code to generate the HTML reports and some sample QC processes. All processes presented here were written for the purpose of this analysis pipeline.

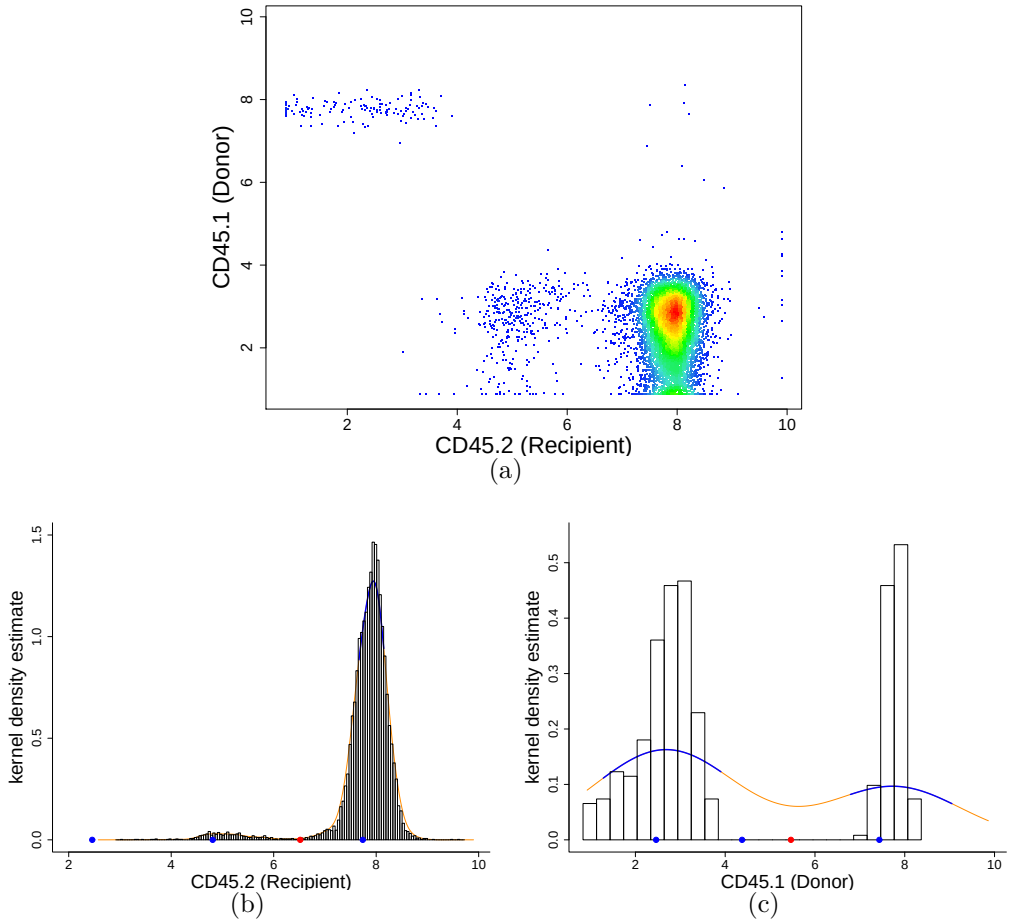


Figure 2.8: Identifying recipient and donor derived cell populations. (a) Sample where recipient and donor derived cell populations need to be localized. High density regions were identified and classified based of proximity to landmarks (●) for (b) the CD45.2 intensity values and (c) the CD45.1 intensity values. (●) marks population boundaries.

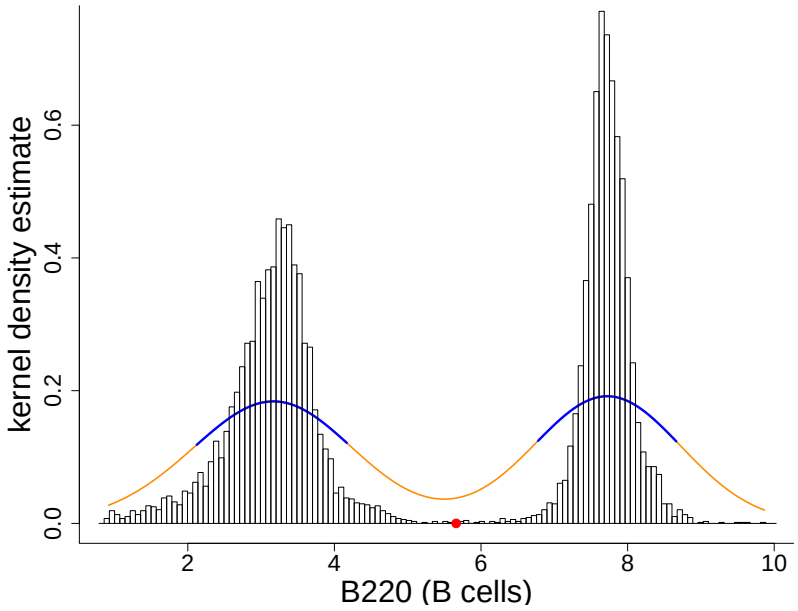


Figure 2.9: Identification of B220$^+$ cells. Histogram and kernel density estimate of B220 intensity values for the donor and recipient derived cells. (•) marks the boundary between B220$^-$ and B220$^+$.

2.3.1 Processes Description

Equivalence Among Tubes

Each mouse sample had been divided into three tubes and each tube had been stained with a corresponding antibody cocktail (see page 14). The distribution of cell types and cell morphology in the three tubes should therefor be equivalent because they were derived from a common source. Hence, the distribution of values in the channels measuring physical properties of the cells (*i.e.*, FSC and SSC) and expression of common antigens (*i.e.*, CD45.1, CD45.2 and PI) should be equivalent in the raw data.

The use of multiple tubes per sample with different staining cocktails is common practice in flow cytometry, and quality control of equivalence among shared features in the tubes has been addressed in the literature before (Le Meur et al., 2007). Le Meur et al. (2007) proposed comparing the empirical cumulative distribution function (ECDF), probability distribution function (PDF), or the boxplots either visually or by testing for outlying distributions with abnormal medians. This method has the disadvantage of assuming normality of the statistic compared (*e.g.*, medians), by using the proposed Grubbs' test for outlying detection, or that visual inspection is not always possible or desired. The method proposed here uses common techniques in probability theory for comparing distributions without the assumptions mentioned or the requirement of visual inspection. Comparing two distributions and computing a divergence (or distance) between the two is a well studied problem in probability theory; more specifically, in information theory. The Kullback-Leibler divergence, also known as the relative entropy of a proba-

bility distribution, is a nonparametric measure of the difference between two probability distributions (see Section 2.1.2 for a detailed description). Here, this metric was used to compare the distributions among triplets. For each feature and for each sample, pairwise distances were computed for the three tubes and the average was stored. Then, for each feature, all averages were compared and those with higher than normal average distances were flagged for inspection, using the generalized extreme studentized deviate (Rosner, 1983) outlier detection routine, available in the R package *parody*.

Figure 2.10 demonstrates the use of this QC process. Each row shows the ECDF for the shared parameters FSC, SSC, CD45.1 and CD45.2 in the three tubes for two difference mice. In the first row, equivalent ECDF curves were obtained for all parameters but in the second row, it can be seen that one tube had a different ECDF than the other two tubes for the CD45.1 and CD45.2 expression intensities. Inspecting the manual gates that were originally used to analyze this data verifies the finding and suggests that tube one was not properly stained (see Figure 2.11). This mishandling of reagents was properly detected by the QC analysis while it was unnoticed by the experimentalist.

Consistency Among Tubes in Cell Population Proportions

After cell population identification, the proportion of viable cells from total events in raw data and the proportion donor-derived cells from total viable cells should be consistent among the three data files per mouse sample. Not being the case could be a consequence of the gating analysis failing to identify and match the populations properly, or in the case of donor derived

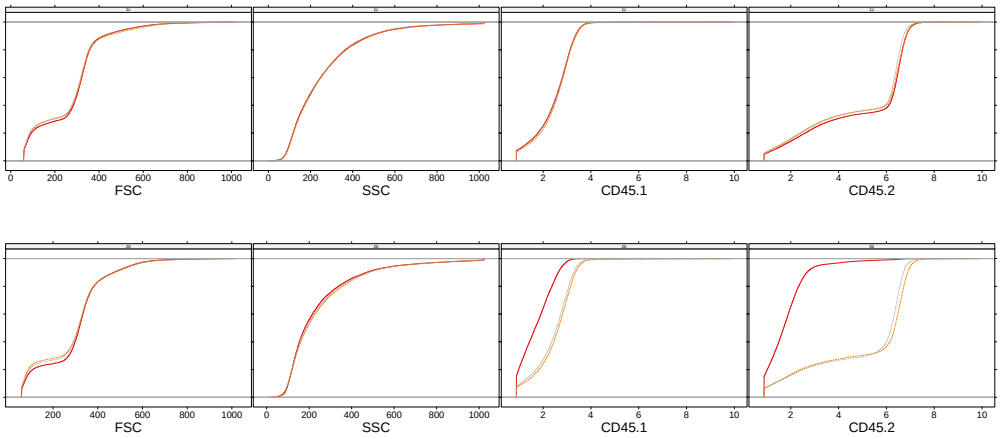
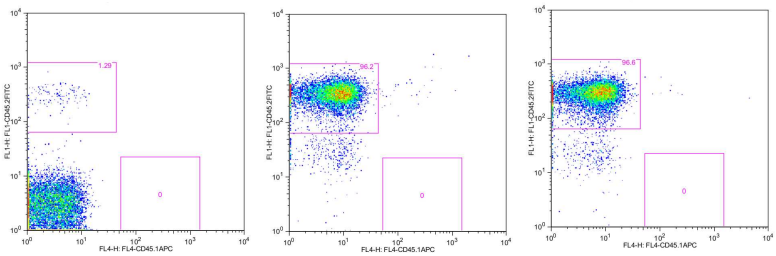


Figure 2.10: Equivalence among shared parameters for sample tubes. Equivalence was assessed by comparing the ECDF of each parameter in the three tubes from the same sample. Row one demonstrates the case when all parameters were equivalent among the three tubes. Second row shows the case where the expression intensities of CD45.1 and CD45.2 were not equivalent.

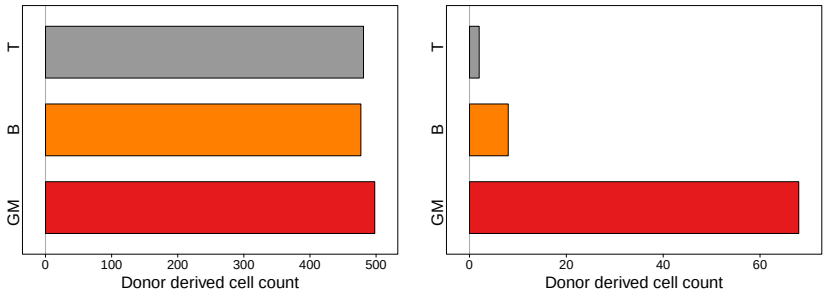


(a) (b) (c)

Figure 2.11: Example of manual gates on mistakenly stained samples. The aliquot containing the GM cocktail (a) was not properly stained and this resulted in erroneous calculation of cell populations through the manual analysis. Sample from experiment DE0407, mouse 2B and week 16 time point.

proportions, that one of the tubes was contaminated. In this procedure, samples were loaded into a 96 well plate in tandem, three wells were loaded per sample, where the corresponding well for the GM cocktail was loaded first and the corresponding well for the T cell cocktail was loaded at last. If the sample loading instrument was not properly cleaned, remains of the previously loaded well would have been deposited in the next well. In this manner, if a previous sample had a high proportion of donor-derived cells and the instrument was not properly cleaned after loading the T cell well, cells from this sample would be deposited in the GM well corresponding to the next sample. This contamination would result in an artificial inflation of the proportion of donor-derived cells detected in the GM tube.

I designed a QC process to detect occurrences where the proportion of viable cells and donor-derived cells was not equivalent in the three tubes corresponding to the same sample. The proportion of donor-derived cells in transplanted mice is an important factor in decision making in downstream analysis and hypothesis testing, hence the importance of accurate measurements and proper identification of possible experimental or analysis artifacts in the reported proportions. Figure 2.12 shows an example where the donor cell count for the second mouse, Figure 2.12(b), were unequal in the three tubes. In this case the donor cell count decreased from the GM to the T cell tube suggesting that proper washing of the instrument was not performed.



(a) (b)

Figure 2.12: Quality control on donor derived cell output. Samples were divided into three aliquots and absolute count of donor derived cells should be equivalent among these (a). Contaminated tubes from another sample resulted in disproportional donor derived cell counts in the wells (b).

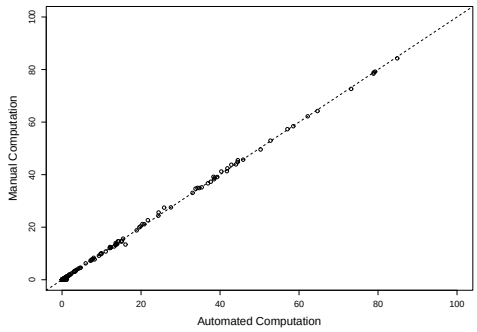
Control of WBC output

I designed and implemented a third QC process to detect cases of disproportional WBC output. For both donor-derived and recipient-derived cell populations, the proportions of GM, B and T cells should add to 100%; otherwise, this is a sign of improper staining or mistakenly annotated samples. When the sum of the proportion of GM, B and T cells is not $100\% \pm 15\%$ the samples were flagged for inspection. The graphics in the reports are also useful for monitoring the health of the recipient mouse. The output measured for the recipient mouse is the steady state output of hematopoietic cells (represents the average output of all its HSCs) and the proportions should be roughly 10-25% myeloid cells and 75-90% lymphoid cells, and within the lymphoid cells the proportions are slightly skewed towards the B cells. Because the mice were irradiated before donor HSC engraftment, there is always the risk that the recipient mouse develops a leukemia, which results in dispro-

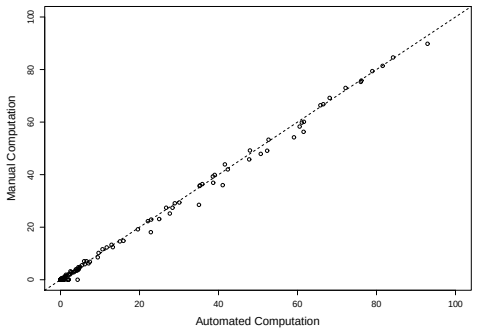
portional myeloid and lymphoid cell types.

2.4 Results and Assessment of the Automated Gating Pipeline

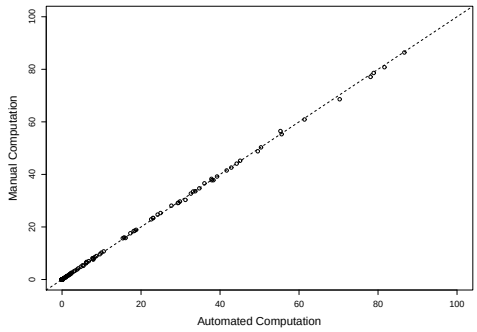
In downstream analyses, the proportion of donor-derived cells and the proportion of each cell lineage from donor and recipient cell populations are important for functional characterization and classification of HSC. Proper quantification of donor-derived cells is important for characterizing the HSC kinetics in the recipient mouse and for predicting repopulation potential at later time points. Accurate measurements of cell lineage proportions is important for characterizing lineage-biases of the transplanted HSC. In this section, measurements obtained with the automated analysis are compared to the same measurements obtained with the conventional manual analysis. For the set of most previous experiments, where data was collected after development of the automated gating pipeline (see Table 1.3), the comparison of relevant statistics is shown in Figure 2.13. All comparisons show high correlation, $\rho > 0.985$, and they are all significantly greater than zero (p-value $< 10^{-15}$). These comparisons demonstrate that the automated pipeline can analyze the data and provide equivalent results as when the data is analyzed with manual gates.



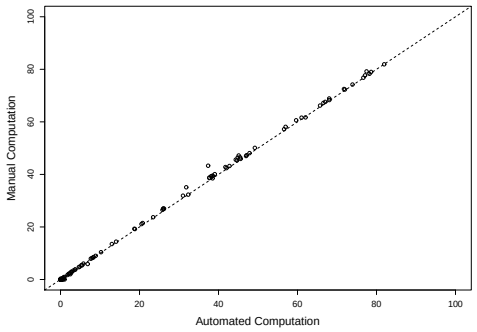
(a) Proportion of donor derived cells



(b) Proportion of donor derived GM cells



(c) Proportion of donor derived B cells



(d) Proportion of donor derived T cells

Figure 2.13: Pipeline results comparison

2.5 Implementation and Accessibility

To facilitate the more general use of the software presented here to the scientific community I have, with the aid of Dr. Josef Špidlen, created a GenePattern module (Reich et al., 2006). GenePattern is a software package which provides a comprehensive environment to access a repository of bioinformatics tools through their website `http://www.broad.mit.edu/genepattern`. The project aims to make the tools accessible to a community of users at all levels of computational experience. The focus of the available tools concentrate around those commonly used for genomics-type research; although our laboratory has a running interest in extending the availability to FCM data analysis tools. Currently, the module is available through an in-house installation of the GenePattern software and is accessible from any machine inside the BCCRC at the web address `http://bioinfosrv1.bccrc.ca:8000/gp`. To run the analysis, all the user needs to do is specify a directory containing the FCM data files from an experiment analysis time point and to annotate the experiment with the experiment name, the week time point and the donor-derived cell marker (*i.e.*, either CD45.1 or CD45.2). The software then reads the meta data in each FCM data file to identify positive controls, mouse identifiers and staining antibodies. See Figure 2.14 for a screen capture of the automated analysis pipeline configuration site. At completion, a list of files is returned which includes a spreadsheet file with all the cell population statistics and a HTML file with the quality control report.

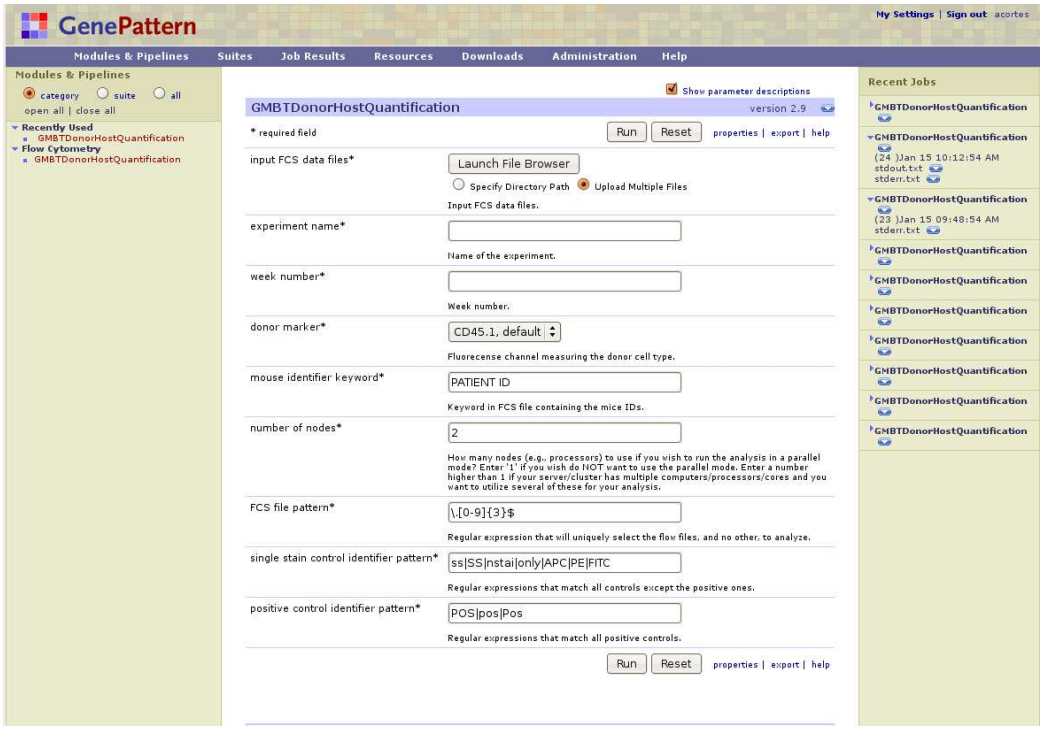


Figure 2.14: Screen capture of the GenePattern module configuration site.

2.6 Discussion

As mentioned in the introduction, there is a growing interest in developing tools for the automated analysis of flow cytometry data. Such tools will aid the analysis of data as laboratories adapt newer technologies that yield data faster and in greater quality and quantity. In the context of our collaborator's laboratory, I developed a tool that will facilitate the analysis of data in several ways. First, the results obtained through the automated analysis tool are equivalent to those obtained through the traditional manual analysis and this is obtained in an unbiased manner: it is common that the person doing the manual analysis changes during the course of an experiment and user effects are observed in the end results. This analysis framework offers comparable and reproducible results. Second, a methodology to assess the quality of the data is integrated into the automated analysis pipeline and HTML reports are provided for quick and easy inspection of the assessment. Quality control is performed on raw and on gated data to pinpoint potential problems originating in the experimental and/or automated gating procedures.

Chapter 3

Repopulation Analysis

The frequencies of HSC in the most purified populations currently investigated are generally below 50% (rates depend on sorting strategy and experimental procedure, see Section 1.2.2). This means that in single cell transplant experiments, most of the recipient mice will not be repopulated by the donor HSC. WBC readout is measured at 4, 8, 16 and 24 weeks post transplantation, where the 16 weeks and later time point readouts are used for subsequent analysis. I investigated whether the readout data before 16 weeks could be used to make the decision to euthanize those mice that would not be repopulated at the 16 weeks time point.

The euthanization of mice predicted not to show repopulation has the potential to increase the available space in the ARC, to reduce the cost in the care of the animals, and to reduce the time and reagents needed to analyze mice at later time points.

A major concern with the implementation of such predictions into an analysis pipeline is the possible elimination of mice that might have shown repopulation at later time points but were not predicted to do so (*i.e.*, false-negatives, see Table 3.1 for possible predictor outcomes). A false-positive would correspond to a false prediction of repopulation at the analysis time point. For the design of the predictor, special care was given to the false-negative rate.

Although it was also desired to minimize the false-positive rate, these errors are not detrimental since they would be identified at later time points, as it is currently done. The set of true negatives are those predictions that are correctly labelled as having no potential for repopulation; this is the set representing what is saved in terms of cost, space and time if these animal were sacrificed early in the experiment. Table 3.1 lists the prediction outcomes with the classifier terms associated with them.

In the literature, different groups use different criteria for calling a mouse repopulated (Kent et al., 2007). Here, the definition in Dykstra et al. (2006) was used: at least 1% chimerism at 16 weeks and at least 1% of each lineage at any time.

Table 3.1: List of possible prediction outcomes.

Classifier Outcome	Implication
True Positive (TP)	Correct repopulation prediction
True Negative (TN)	Correct nonrepopulation prediction
False Positive (FP)	Incorrect repopulation prediction
False Negative (FN)	Incorrect nonrepopulation prediction

3.1 The Data

Data was compiled from 341 single cell transplants from multiple experiments (source of the data listed on Table 3.2). Predictions were assessed at the 8 week time point since only the early experiments (prior to 2006) measured WBC output at 12 weeks. The quality of the data was assessed

Table 3.2: Data used for the repopulation prediction analysis of single cell HSC transplant experiments. comm.- Unpublished data communicated by source.

Phenotype of HSC	Experiments	Number	Reference
$CD45^{mid}lin^{-}Rho^{-}SP$	1004d, 1004e, 105a, 105b, 205 AL1305, AL2005	105	Dykstra et al. (2007)
E-SLAM	SE1707, SE1306, SE2506, JL0307	61	comm. by Dr. Dave Kent
E-SLAM	OC3007, DE0407, FE0508, MR0308 AP2908, NO2608, JA1507	87	comm. by Dr. Claudia Benz
E-SLAM	08FE12, 08JAN25 07NOV26, 08MAR12	88	comm. by Michael Copley

before analysis and those samples where donor-derived cell counts were not consistent among the three sample tubes or the proportion of GM, B and T cells from the total donor cells did not sum to $100 \pm 10\%$ (see Section 2.3 for explanation and discussion) were removed from the dataset. For each sample (corresponding to one mouse), 7 features were extracted: average donor contribution; the proportion of GM, B and T cells from the total donor population; and the relative proportion of donor derived GM, B and T cells out of the total GM, B and T cell populations, respectively. Exploratory data analysis indicated a high correlation, $\rho = 0.98$, between the average donor contribution and the proportion of donor derived B cells out of the total B cell count. Because of this high correlation, the second variable was removed from the analysis to avoid multicollinearities in the final multivariable model.

3.2 Methods

The repopulation prediction was formulated as a binary classification problem with two possible states: nonrepopulation and repopulation. Several binary classification predictors were compared, Table 3.3 lists the models examined.

Table 3.3: Classification models examined. LR - logistic regression; CART - classification and regression tree; SVM - support vector machine; ML - maximum likelihood; fss - forward stepwise selection.

Model	R Implementation	Ref.
LR with L2 regularization - fss	step.plr	(Park and Hastie, 2008)
LR with L2 regularization	plr	(Park and Hastie, 2008)
LR with ML and AIC	glm and stepAIC	(Venables et al., 1996)
SVM	ksvm	(Karatzoglou et al., 2004)
CART	tree	(Ripley, 2008)

3.2.1 The Prediction Models

For assessing the performance of the models, the data was first divided into two sets: a training set consisting of 80% of the data entries and the remaining 20% comprising the test set. Test and training sets were constructed so that the repopulation rates were equal in both sets; otherwise, samples were assigned randomly to each set. The training data was then used to fit

each of the models in Table 3.3 and the test data was used to asses the predictive power of the models. This procedure was then repeated 1,000 times to determine which of the models provided, on average, the best predictive model.

CART. Classification and Regression Tree (CART) is a popular classification approach in the medical sciences as it resembles the way medical decisions are made (Hastie et al. (2005)). The method recursively finds binary partitions of the feature space assigning a class to the final partitions; the partitions can then be described as a decision tree. The algorithm has been implemented in the R package *tree*. For each training set, a classification tree was inferred and then 10-fold cross-validation was used to determine the cost-complexity parameter to prune the tree.

Logistic Regression. Logistic regression (LR) is another popular classifier. As for CART, it has a simple interpretability. The model returns the odds of a mouse being repopulated by the donor HSC as a function of the features in the model. There exist several ways to fit the model and to select the features in the final model. Here, a maximum likelihood inference was used with and without L2 regularization. Model selection was determined with and without forward stepwise selection of the features, based on the AIC criterion (Venables et al., 1996). The R package *stepPlr* was used to infer the logistic regression model with L2 regularization.

SVM. Support vector machines (SVM) are a relative new classification method developed in the 1990s by Vapnik *et al.* (one of the introductory

references being Cortes and Vapnik (1995)). As opposed to LR, that determines a regression line through the data points, the SVM attempts to find a separation boundary to classify the input variables. It either finds the linear boundary within a number of input variables or it increases the dimensionality of the feature space by mathematically transforming the input variables and inferring the boundary within this higher dimensional feature space. To fit the model the R function *ksvm* in the package *kernlab* was used. Cross-validation was used to find appropriate model parameters (*i.e.*, the C-constant of the regularization term and the kernel function hyperparameter γ) to prevent overfitting, since no significant improvement was obtained, default parameters were used.

3.2.2 Classifier Selection

Each of the models returned a repopulation probability for each of the test set entries. The cutoff value is the probability at which any repopulation probability at or above this value the predictor will return a repopulation label. In one extreme, a cutoff value of 0 will label all entries as repopulated (this is analogous to the current situation where all animals are kept alive). In the other extreme, a cutoff value of 1 will label all entries as nonrepopulated, and the false discovery rate will be at its maximum. In this manner, as the cutoff value is increased, the false discovery rate increases too. Figure 3.1 depicts this trend, where the blue line is the average line of the 1,000 training/testing repetitions when using the prediction model LR_L2_step.

The chosen cutoff value represents a tradeoff between the false positive rate

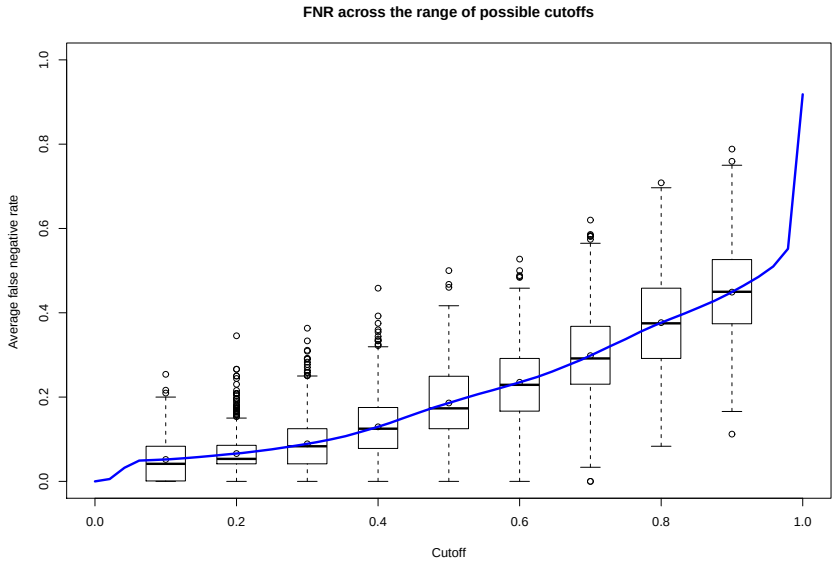


Figure 3.1: Repopulation prediction false negative rate as a function of the cutoff value for the model LR_L2_step. Blue line averages the 1,000 iterations and boxplots show the variation observed for particular cutoff values.

(FPR) and the false negative rate (FNR). FPR and FNR are defined as:

$$FPR = \frac{Num.\,of\,false\,positives}{Num.\,of\,negative} \tag{3.1}$$

and

$$FNR = \frac{Num.\,of\,false\,negatives}{Num.\,of\,positives}. \tag{3.2}$$

Figure 3.2 shows this relationship, where each point in the line represents a cutoff value (color coded) and the associated FNR and FPR are given by the x- and y-axis coordinates, respectively. The closer the line passes through the origin the better the predictor; where the curve going from (0,1) to (0,0) and then from (0,0) to (1,0) describes the perfect predictor. As mentioned above, in this analysis it was necessary to control the FNR. This was done by choosing a cutoff value that provided a permitted FNR and minimized the FPR. Here, a FNR of 0.05 was chosen to explore the potential of repopulation prediction, this rate allows a small FNR and a significant gain in the true negative rate (TNR). In a graph such as the one shown in Figure 3.2 the cutoff value of 0.05 is located at the point in the line with y-coordinate equal to the permitted FNR. For each training and testing iteration, the cutoff value was chosen as indicated and the FPR was recorded. Then for the 1,000 iterations the FPRs across the models tested were compared (Figure 3.3). The performance was not significantly different among the LR models (using the Mann-Whitney test), and these models performed significantly better than the SVM and CART models at a significant level of $p < 0.05$. With these results, the model using logistic regression with L2 regularization and forward step feature selection was chosen as the best model as it had the lowest median (FPR = 0.16) and minimized outlying FPRs. Outlying FPRs

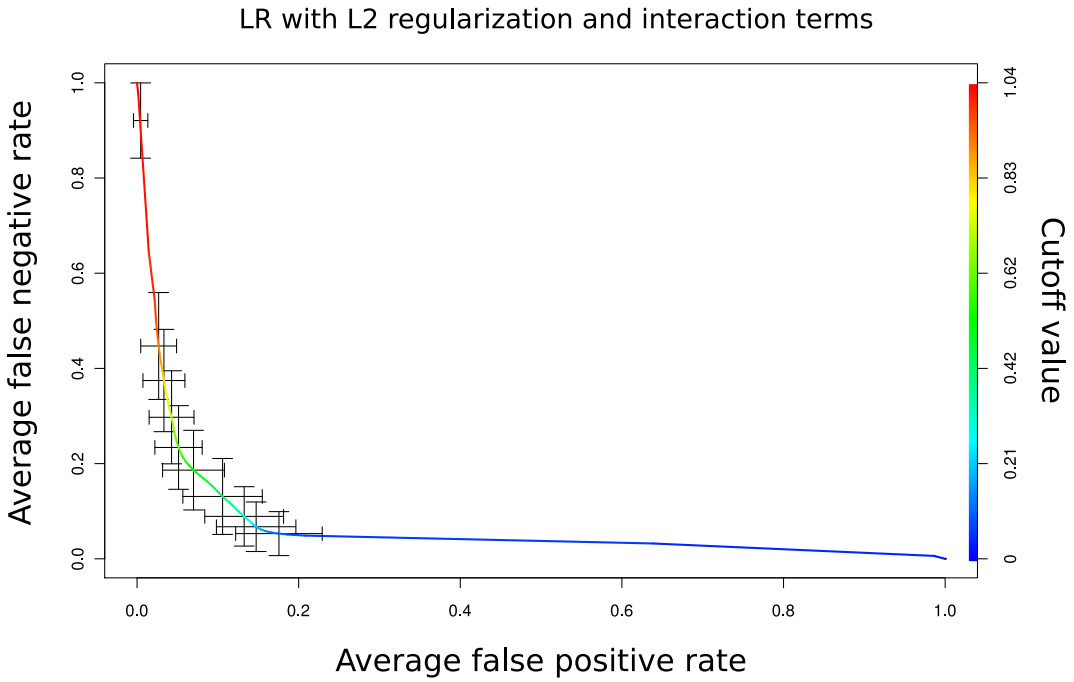


Figure 3.2: Repopulation prediction trade-off between the false negative rate and the false positive rate. Line shows the average of the 1,000 iterations. The cutoff value is color coded and error bars show ± 1 standard deviation.

was a sign of over fitting, depicted in Figure 3.3 as individual points above 1.5 the interquartile range at each boxplot.

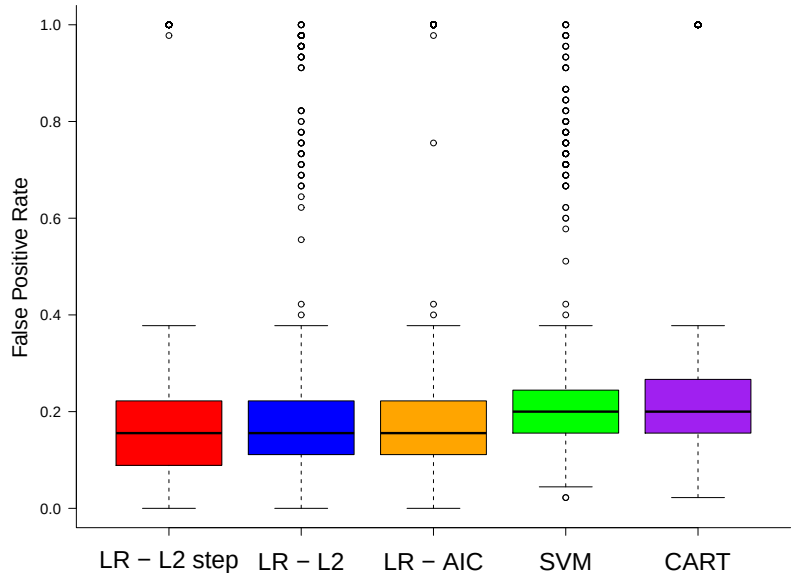


Figure 3.3: Performance comparison for different binary classification predictors in the repopulation problem. Cutoff values were selected so that the FNR was at the most 0.05. Performance is measured as the FPR. Performance of the SVM and CART models was significantly worse than the other models ($p < 0.05$) and performance among the LR models was not significantly different.

3.3 Results and Application

The final model was constructed using the logistic regression with L2 regularization and forward step feature selection methodology. This model contained two significant features and one significant interaction term. The two features in the final model were the proportion of donor-derived cells from

the viable cell population and the proportion of myeloid cells from the donor-derived cell population. The interaction term was among the proportion of donor-derived cells from the viable cell population and the relative proportion of donor derived cells from the myeloid cell population. The significance of these features agrees with the biological knowledge of the system. The proportion of donor-derived cells from the total viable cell population is a direct measure of the activity of the transplanted HSC. The higher the proportion of the donor derived cells the higher the repopulation probability, this is shown in Figure 3.4. On this figure, the red line depicts the cutoff value that was derived to obtain a FNR of 5%, which in this case corresponds to 1.58% of donor-derived cells when the other two features are held constant. The other two variables relate to the proportions of donor-derived myeloid cells. Myeloid cells have a shorter half-life than lymphoid cells; hence, their detected presence allows a better diagnostic that the engraftment was sustained, as opposed to the lymphoid cells. In the proposed model, the repopulation probability increased as the proportion of donor-derived myeloid cells increased, whenever the proportion of donor-derived cells from viable cells was kept constant.

If the model obtained had been used to predict the repopulation of all single cell transplant experiments compiled (341 in total), a FNR of 4.2% and a FPR of 14.9% would have been obtained. This means that, of the 120 transplants that did repopulate the recipient mice, 5 would have been classified as nonrepopulated and euthanized. In contrast, of the 221 transplants that did not repopulate the recipient mouse, 188 would have been correctly predicted and only 33 would have been kept alive. Table 3.4 summarizes these results.

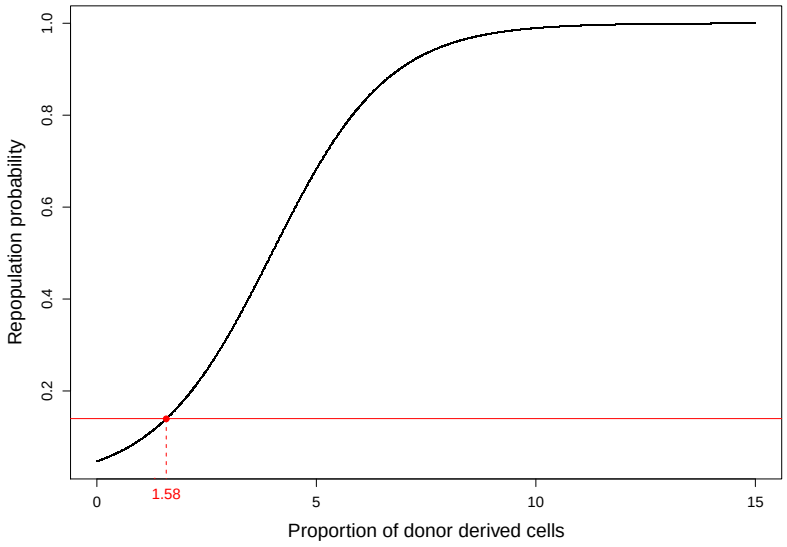


Figure 3.4: Repopulation probability at the 16 week post-transplant time point as a function of the donor-derived proportion of viable cells at the 8 week post-transplant time point, as predicted by logistic regression with L2 regularization and forward step feature selection. Horizontal red line depicts the cutoff value to assign a repopulation label to events above this line. All other model variables, myeloid proportion of donor derived cells and donor derived proportion of all myeloid cells, were kept constant at 50% and 25%, respectively.

This exercise demonstrates that predictive analysis elicits the possibility to liberate many of the resources that these experiments currently utilize. There is a tradeoff that has to be taken into consideration when evaluating the benefits of applying this predictive analysis, and that is between the potential loss of valuable information and the value and implications of the liberated resources. In the scenario described above, 5 mice would have been lost that could have provided valuable information, and the resources used by 188 mice would have been freed by euthanizing them at the 8 week time point. Maintaining these mice in the ARC costs the researchers $0.75 per day per cage (each cage housing up to 4 mice). As these 188 mice represent more than half of the transplanted mice, the cost of maintaining the mice after the 8 week time point could have been reduced by one half.

Table 3.4: Repopulation prediction results on 341 single cell transplants. Neg - nonrepopulation; Pos - repopulation.

		Labels	
		Neg	Pos
Prediction	Neg	188	5
	Pos	33	115

Chapter 4

Discussion and Conclusion

The significance of the work presented here is twofold. First, the analyses have yielded a validated tool that allows analysis of data generated from single-cell HSC transplant experiments. The results generated with the automated analysis pipeline proved equivalent to the results obtained with the conventional manual analysis. The automated analysis pipeline offers the advantage that the gating analysis of the data could become effortless for the experimentalist and the QC reports can provide the prompt detection of experimental errors in a rapid manner. Also, this analysis framework provides reproducible results, and an objective, open and extensible analysis.
The repopulation analysis demonstrates the power of proper data management and predictive analysis. Proper data management enables predictive analysis; in this case, a major bottleneck encountered in this project was compiling the data. Many experiments have been performed in the past years by more than one researcher and it took significant time and effort to identify the correspondence between raw data, experiment, mouse identifiers, and time point of the analysis. The results obtained through this analysis demonstrated that predicting the repopulation can significantly reduce the cost, time and effort these experiments utilize. As it is usually the case with predictive analyses there was the possibility of erroneously classifying

an event and it is ultimately up to the researcher to decide if the error rate is worth the saving. In the case presented here, for five nonrepopulation errors there were 188 mice that were correctly predicted not to repopulate. If the resources liberated by euthanizing these 188 mice had been re-allocated to new experiments, there would have been the potential to observe more than 5 positive repopulations. This new framework can thus speed the process of hypothesis testing.

Secondly, the results shown here provide a proof of concept that high-throughput flow cytometry data analysis is feasible. This analysis pipeline builds on the efforts of many scientists who have donated their work in open-source statistical software packages for FCM data analysis. All of these packages are now mature enough to assemble them into a full automated pipeline for the analysis of data derived from any kind of experiment generating significant amount of data and where the FCM data analysis is equivalently and routinely performed.

Bibliography

A.B. Balazs and R.C. Mulligan. A novel marker for hematopoietic stem cells. In *From Stem Cells to Therapy 2003, Keystone Meeting, Colorado Springs, CO, USA*, 2003.

R.J. Beckman, G.C. Salzman, and C.C. Stewart. Classification and regression trees for bone marrow immunophenotyping. *Cytometry Part B: Clinical Cytometry*, 20(3):210–217, 1995.

D. Bryder, D.J. Rossi, and I.L. Weissman. Hematopoietic stem cells: the paradigmatic tissue-specific stem cell. *American Journal of Pathology*, 169 (2):338, 2006.

G.A. Challen, N. Boles, K.Y.K. Lin, and M.A. Goodell. Mouse hematopoietic stem cell identification and analysis. *Cytometry Part A*, (1), 2009.

C. Chan, F. Feng, J. Ottinger, D. Foster, M. West, and T.B. Kepler. Statistical mixture modeling for cell subtype identification in flow cytometry. *Cytometry-Part A*, 73(8):693–701, 2008.

P.K. Chattopadhyay, D.A. Price, T.F. Harper, M.R. Betts, J. Yu, E. Gostick, S.P. Perfetto, P. Goepfert, R.A. Koup, S.C. De Rosa, et al. Quantum dot semiconductor nanocrystals for immunophenotyping by polychromatic flow cytometry. *Nature medicine*, 12(8):972–977, 2006.

P.K. Chattopadhyay, C.M. Hogerkorp, and M. Roederer. A chromatic explosion: the development and future of multiparameter flow cytometry. *Immunology*, 125(4):441, 2008.

P. Chaudhuri and JS Marron. SiZer for Exploration of Structures in Curves. *Journal of the American Statistical Association*, 94(447):807–808, 1999.

C. Cortes and V. Vapnik. Support-vector networks. *Machine learning*, 20 (3):273–297, 1995.

SC De Rosa, LA Herzenberg, and M. Roederer. 11-color, 13-parameter flow cytometry: identification of human naive T cells by phenotype, function, and T-cell receptor diversity. *Nature medicine*, 7(2):245, 2001.

T. Duong, A. Cowling, I. Koch, and MP Wand. Feature significance for multivariate kernel density estimation. *Computational Statistics and Data Analysis*, 52(9):4225–4242, 2008.

B. Dykstra, J. Ramunas, D. Kent, L. McCaffrey, E. Szumsky, L. Kelly, K. Farn, A. Blaylock, C. Eaves, and E. Jervis. High-resolution video monitoring of hematopoietic stem cells cultured in single-cell arrays identifies new features of self-renewal. *Proceedings of the National Academy of Sciences*, 103(21):8185, 2006.

S. Errol, H. Florian, R. Ryan, et al. Analysis of High-Throughput Flow Cytometry Data Using plateCore. *Advances in Bioinformatics*, 2009.

G. Finak, A. Bashashati, R. Brinkman, and R. Gottardo. Merging mixture components for cell population identification in flow cytometry. *Advances in Bioinformatics*, 2009.

FlowCAP. Flow Cytometry: Critical Assesment of Population Identification Methods, September 2009. URL `http://flowcap.flowsite.org/`.

R. Gentleman, V. Carey, D. Bates, B. Bolstad, M. Dettling, S. Dudoit, B. Ellis, L. Gautier, Y. Ge, J. Gentry, et al. Bioconductor: open software development for computational biology and bioinformatics. *Genome biology*, 5(10):R80, 2004.

F. Hahne, AH Khodabakhshi, A. Bashashati, CJ Wong, RD Gascoyne, AP Weng, V. Seyfert-Margolis, K. Bourcier, A. Asare, T. Lumley, et al. Per-channel basis normalization methods for flow cytometry data. *Cytometry. Part A: the journal of the International Society for Analytical Cytology*, 2009a.

F. Hahne, N. Le Meur, R.R. Brinkman, B. Ellis, P. Haaland, D. Sarkar, J. Spidlen, E. Strain, and R. Gentleman. flowCore: a Bioconductor package for high throughput flow cytometry. *BMC bioinformatics*, 10(1):106, 2009b.

T. Hastie, R. Tibshirani, J. Friedman, and J. Franklin. The elements of statistical learning: data mining, inference and prediction. *The Mathematical Intelligencer*, 27(2):83–85, 2005.

LA Herzenberg and RG Sweet. Fluorescence-activated cell sorting. *Scientific American*, 234(3):108, 1976.

Alexandros Karatzoglou, Alex Smola, Kurt Hornik, and Achim Zeileis. kernlab – an S4 package for kernel methods in R. *Journal of Statistical Software*, 11(9):1–20, 2004. URL `http://www.jstatsoft.org/v11/i09/`.

D. Kent, B. Dykstra, and C. Eaves. Isolation and assessment of long-term reconstituting hematopoietic stem cells from adult mouse bone marrow. *Current Protocols in Stem Cell Biology*, 2:4–2, 2007.

M.J. Kiel, O.H. Yilmaz, T. Iwashita, O.H. Yilmaz, C. Terhorst, and S.J. Morrison. SLAM family receptors distinguish hematopoietic stem and progenitor cells and reveal endothelial niches for stem cells. *Cell*, 121(7): 1109–1121, 2005.

R. Kothari, H. Cualing, and T. Balachander. Neural network analysis of flow cytometry immunophenotype data. *IEEE Transactions on Biomedical Engineering*, 43(8):803–810, 1996.

S. Kullback. *Information theory and statistics*. Dover publications Mineola, MN, 1997.

N. Le Meur, A. Rossini, M. Gasparetto, C. Smith, R.R. Brinkman, and R. Gentleman. Data quality assessment of ungated flow cytometry data in high throughput experiments. *Cytometry Part A*, (6), 2007.

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

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

K. Lo, F. Hahne, R.R. Brinkman, and R. Gottardo. flowClust: a Bioconductor package for automated gating of flow cytometry data. *BMC bioinformatics*, 10(1):145, 2009.

C.W. Morris, A. Autret, and L. Boddy. Support vector machines for identifying organismsa comparison with strongly partitioned radial basis function networks. *Ecological Modelling*, 146(1-3):57–67, 2001.

R.F. Murphy. Automated identification of subpopulations in flow cytometric list mode data using cluster analysis. *Cytometry Part B: Clinical Cytometry*, 6(4):302–309, 1985.

NCBI Entrez Gene. PTPRC protein tyrosine phosphatase, receptor type, C [*Homo sapiens*]. URL `http://www.ncbi.nlm.nih.gov/sites/entrez?Db=gene\&Cmd=ShowDetailView\&TermToSearch=5788`.

M.Y. Park and T. Hastie. Penalized logistic regression for detecting gene interactions. *Biostatistics*, 9(1):30, 2008.

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

E. Parzen. On estimation of a probability density function and mode. *The annals of mathematical statistics*, 33(3):1065–1076, 1962.

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

R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2009. URL `http://www.R-project.org`. ISBN 3-900051-07-0.

M. Reich, T. Liefeld, J. Gould, J. Lerner, P. Tamayo, and JP Mesirov. GenePattern 2.0. *Nature genetics*, 38(5):500, 2006.

B.D. Ripley. *Pattern recognition and neural networks*. Cambridge Univ Pr, 2008.

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

M. Roederer, W. Moore, A. Treister, R.R. Hardy, and L.A. Herzenberg. Probability binning comparison: a metric for quantitating multivariate distribution differences. *Cytometry*, 45(1):47–55, 2001.

B. Rosner. Percentage points for a generalized ESD many-outlier procedure. *Technometrics*, 25(2):165–172, 1983.

D. Sarkar, N. Le Meur, and R. Gentleman. Using flowViz to visualize flow cytometry data. *Bioinformatics*, 24(6):878, 2008.

T.C.B. Schut, B.G. De Grooth, and J. Greve. Cluster analysis of flow cytometric list mode data on a personal computer. *Cytometry*, 14(6):649–659, 1993.

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

H.M. Shapiro and R.C. Leif. *Practical flow cytometry*. Wiley-Liss New York, 2003.

P. Shooshtari, H. Zare, A. Gupta, and Brinkman R.B. Faithful Sampling for Spectral Clustering to Analyse High Throughput Flow Cytometry Data. *submitted to BMC Bioinformatics*, 2009.

JE Till and EA McCulloch. A direct measurement of the radiation sensitivity of normal mouse bone marrow cells. *Radiation Research*, pages 213–222, 1961.

W.N. Venables, B.D. Ripley, and WN Venables. *Modern applied statistics with S-Plus*. Springer New York, 1996.