PREDICTION OF GRAFT-VERSUS-HOST DISEASE BASED ON SUPERVISED TEMPORAL ANALYSIS ON HIGH-THROUGHPUT FLOW CYTOMETRY DATA

by

Shang-Jung Lee
B.Sc., Queen's University, 2003

THESIS SUBMITTED IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE

in

The Faculty of Graduate Studies
(Bioinformatics)

UNIVERSITY OF BRITISH COLUMBIA

APRIL 2007

ABSTRACT

Despite recent advancements in human immunogenetics, graft-versus-host disease (GvHD) continues to be the major and potentially fatal complication of hematopoietic stem cell transplantations affecting up to 80% of transplant patients [1]. Very little is known regarding the pathophysiologic mechanisms behind the manifestation of either acute or chronic GvHD. Diagnosis and treatment assessment are often hindered as they rely primarily on ambiguous clinical symptoms, such as tissue inflammation. It is likely that the outcome for patients diagnosed with GvHD could be improved if they were treated in a pre-emptive fashion, before the development of full-scale clinical symptoms.

Using flow cytometry high content screening [2], 123 subsets of immune cells were identified from blood samples taken at multiple time points from 31 patients who underwent allogenic bone marrow transplantations. I assembled a novel analysis pipeline specifically designed to process this high-throughput clinical flow cytometry dataset. The pipeline included a novel quality assurance test [3] and temporal classification via functional linear discriminant analysis [4]. Temporal patterns of multiple immune cell abundances both after the transplantation and around the acute GvHD diagnosis were screened for potential discriminative power for either acute or chronic GvHD.

Among many potential discriminative patterns: higher proportion values in immune cell with $CD3^{+}CD4^{+}CD8\beta^{+}$ phenotype were found in acute GvHD patients (21), compared to the patients unaffected by GvHD (3), between zero and 120 days post-transplant. I also generated a list of recommendations for an extended study designed to validate the current findings. The global approach of the high-throughput flow cytometry technique and the novel temporal analysis pipeline, implemented according to the list of recommendations would be beneficial in

elucidating pathophysiologic mechanisms of complex immunologically based diseases including GvHD.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

aGvHD	Acute graft-versus-host disease
ALL	Acute lymphoblastic leukemia
AML	Acute myeloid leukemia
APC	Allophycocyanin
BMT	Bone marrow transplantation
br	Bright (in FCM gating)
CD	Cluster of differentiation
CE-MS	Capillary electrophoresis coupled mass spectrometry
cGvHD	Chronic graft-versus-host disease
CLL	Chronic lymphoblastic leukemia
CML	Chronic myeloid leukemia
DP	Double positive
ECDF	Empirical cumulative distribution function
EM	Expectation maximization
FC-HCS	Flow cytometric high content screening
FCM	Flow cytometry
FCS	Flow cytometry standard
FITC	Fluorescein isothiocyanate
FSC	Forward scatter
GvHD	Graft-versus-host disease (refers to both acute and chronic GvHD)
HIV	Human immunodeficiency virus
HLA	Human leukocyte antigen
HSCT	Hematopoietic stem cell transplantation
int	Intermediate (in FCM gating)
LOOCV	Leave-one-out cross-validation

MDS	Myelodysplasia
MHC	Major histocompatibility complex
MNC	Mononuclear cell
MPD	Myeloproliferative disorder
MUD	Matched unrelated donor
NHL	Non-Hodgkin's lymphoma
NK	Natural killer (cells)
PE	Phycoerythrin
PerCP	Peridinin chlorophyll protein
QA	Quality assurance
rest/act	resting or activate states (of T cells)
SELDI-TOF	Surface-enhanced laser desorption ionization time-of-flight
SIB	Sibling donor
SSC	Side scatter
SVMs	Support vector machines
TCR	T cell receptors

LIST OF SYMBOLS

Y_{ij}	A set of observed value from patient j and class i (j = 1…J; i = 1…I)
S_{ij}	B-spline matrix
S_X	B-spline matrix for test data x
λ_0	Global base value
$\Lambda\alpha_i$	Class signal
γ_{ij}	Individual signal variation
ε_{ij}	Random experiment error
$\hat{\alpha}_x$	Linear discriminant value
weight	Weights of the difference between the test data and the global base value ($(\Lambda^T S_X^T \Sigma_X^{-1} S_X \Lambda)^{-1} \Lambda^T S_X^T \Sigma_X^{-1}$)
α	Significance level or tolerance of a type I error

LIST OF EQUATIONS

PREFACE

Graduate study in the CIHR/MSFHR Strategic Training Program in Bioinformatics at the University of British Columbia begins with three four-month rotation projects and concludes with a master thesis study. Please note that only the master project was included in this thesis in order to have a connected thesis framework. My three rotation projects with Dr. Peter M. Lansdorp & Dr. Ryan Brinkman, Dr. Artem Cherkasov, and Dr. Robert Hancock are not directly related to my master study and are therefore absent from this thesis. Nonetheless, I gained a significant part of my knowledge in genome analysis, drug target identification, microarray analysis, etc. through the rotation projects.

ACKNOWLEDGEMENT

I would like to thank everyone who helped or advised me with this project. Especially, I would like to extend my whole-hearted appreciation to:

My supervisor **Dr. Ryan Brinkman** who gave me the opportunity to study this innovative project and the freedom to explore many different analysis methods. In particular, I would like to thank Ryan for his continuous support through my illnesses.

Dr. Clay Smith and **Dr. Maura Gasparetto** for imparting invaluable knowledge of the graft-versus-host disease and flow cytometry. **Dr. Marco Marra** for his advice throughout the project. **Dr. Colleen Nelson** for her support and insight into the statistical validation of this study. Also, **Ben Smith** for his help on the static sample size calculation. I would also like to thank **Dr. Robert Gentleman** and **Dr. Nolwenn Le Meur** for their work on the flow cytometry quality assurance test ; **Simon Dablemont** for his MATLAB scripts for the functional linear discriminant analysis; and **James Wagner** for his help on SVMs.

My program committee members: **Dr. Marco Marra** (senior supervisor), **Dr. David Baillie**, and **Dr. Fiona Brinkman** for their guidance especially at the beginning of my graduate studies. Also, my rotation project supervisors: **Dr. Peter M. Lansdorp, Dr. Ryan Brinkman, Dr. Artem Cherkasov**, and **Dr. Robert Hancock.**

Administrative staff including **Ms. Sharon Ruschkowski** from the Bioinformatics program and **Ms. Monica Deutsch** from the UBC genetics program for their assistances.

Fellow students in the UBC genetics program, the Bioinformatics training program, and the BC Cancer Research Centre. Special thanks to **Debra Fulton**, **Evette Haddad**, and **Alison Meynert** for their friendship and brainstorming sections.

This study was funded by the CIHR/MSFHR Strategic Training Program in Bioinformatics and the British Columbia Transplant Foundation.

DEDICATION

This work is dedicated to my parents Shiaou-Cheng Lee and Mu-Tzu Tsou, for their support. Great opportunities like this master project were only made possible because of their insight and hard work in bringing me to Canada.

CHAPTER 1 INTRODUCTION

Hundreds of bone marrow transplantations (BMT) are performed in Canada each year. Despite numerous technical advances, graft versus host disease (GvHD) continues to be a major complication of hematopoietic stem cell transplantations (HSCT) [1, 5] with a maximum 90% fatality rate for severe GvHD [5-7]. Presently, there is no test to diagnose the disease definitively, nor standardized assessment for monitoring response to treatment. Therefore, it is imperative to develop more reliable and precise tests for predicting and diagnosing GvHD. In the present study, large scale immune cell population data obtained from a high-throughput flow cytometry (FCM) technique (section 1.1), were screened for their potential GvHD (section 1.2) predictive power by a novel temporal analysis pipeline (section 1.3). Finally, principles of sample size calculation are described in section 1.4.

1.1 Flow Cytometry

The first flow cytometer, an integration of the flow system and the static microscope, was developed by Wallace Coulter in 1954 to count red blood cells. Today, flow cytometers can separate and count almost any type of biological or non-biological particle by combining its light scattering properties, which provide an indication of particle size and shape, as well as the presence of specific fluorescence markers or fluorochromes.

In FCM, cells are typically labelled with antibody-conjugated fluorochromes that are used to detect the presence of cell surface proteins. The labelled cells are then suspended in sheath fluid and flow past the excitation light source, usually a laser, through a narrow tube one cell at a time. A detector measures the light emitted from the sample and the intensity of the light can then be used as an indication of, for example, the presence or absence of a fluorochrome. In the late

1970's and early 1980's, clinical applications of FCM rapidly developed in response to the emergence of the human immunodeficiency virus (HIV) [8]. Since then, several advancements in antibodies, fluorochromes, and resonance fluorescence techniques now allow researchers to count and sort an exact population of particles via sequential gating based on their physical or chemical characteristics.

Gating is a procedure for FCM data where cells with common measurement intensities are grouped together. This is performed by either identifying a particular group of cells or separating the entire cell population based on a one or two parameters display. In sequential gating, multiple markers can be utilized to identify a particular subset of particles. An example of the FCM sequential gating is shown in Figure 1.1. First, forward and side scatter (FSC and SSC) contour graphs (Figure 1.1a) were used to distinguish live cells (34%) and dead cells by their unique characteristic size and granularity. The population of live cells can be further divided using different cluster of differentiation (CD) markers. CDs generally represent cell-surface antigens. Different immune cell lineages and functions can be identified using different combination of the CD markers. In this case, the live cells can be further divided using CD3-fluorchorme intensity (Figure 1.1b) and then CD44 and CD25 (Figure 1.1 c & d). At the end, 68.8% and 31.2% live cells are with ($CD3^+$) and without ($CD3^-$) the CD3 surface marker respectively. These two populations can be further divided into subpopulations of $CD25^+CD44^+$, $CD25^+CD44^-$, etc.

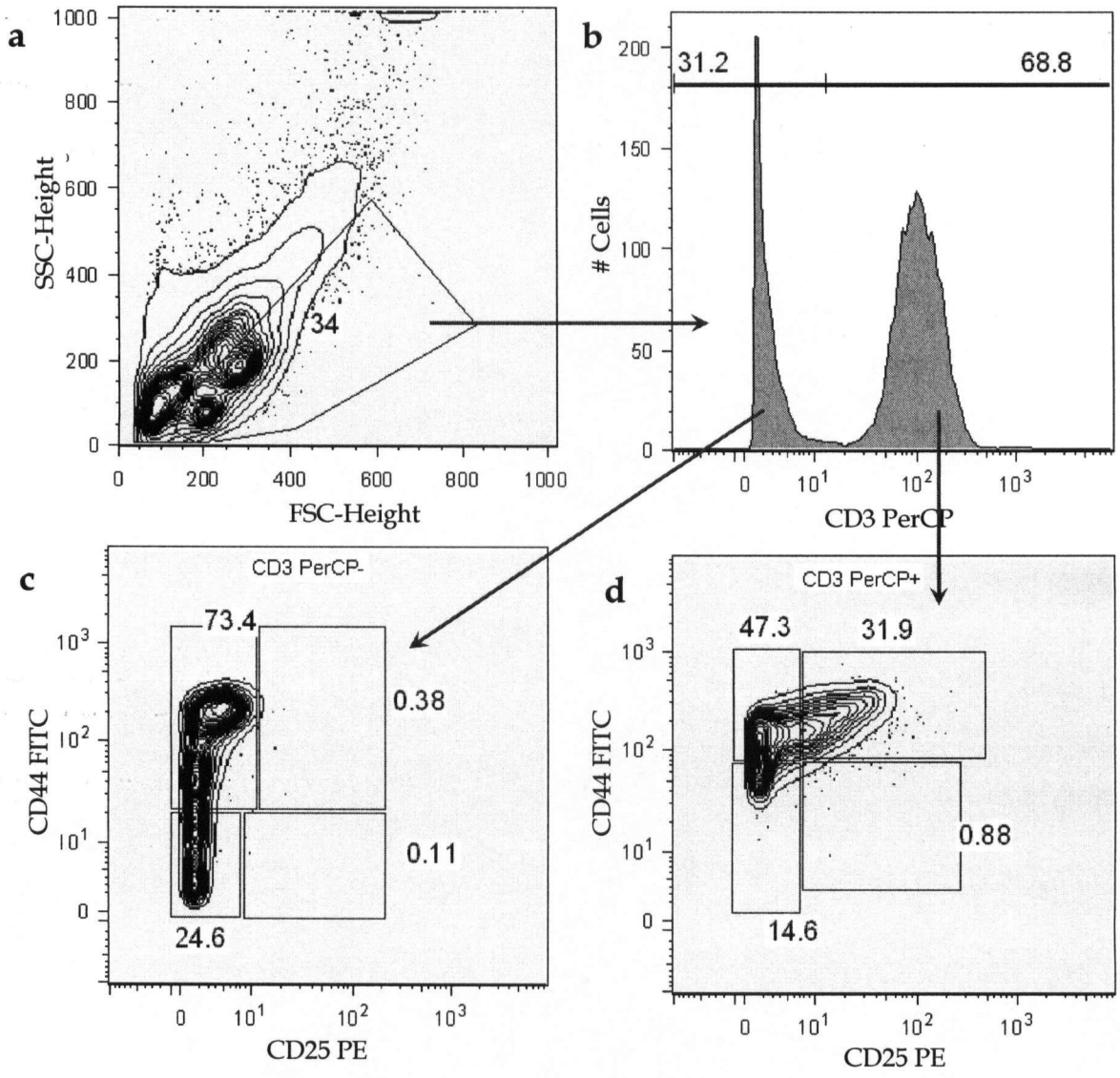


Figure 1.1 An example of sequential gating in FCM displayed in contour or histogram

Multiparametric FCM data analysis is an essential technique in immunophenotyping. Multiple antibodies and fluorochromes can be used to identify specific immune cell lineages. Major clinical uses of FCM include the diagnosis and monitoring of leukemia and lymphoma [9, 10], the evaluation of peripheral blood hematopoietic stem cell grafts [11], and the quantitation of $CD4^+$

versus $CD8^+$ T cells in blood to monitor HIV infection and to assess the treatment performance [8].

FCM high content screening (FC-HCS) [2], a high throughput FCM method, was developed by automating the staining and sample analyses using robotic devices. The technique is robust and can process up to a thousand samples per day. Using this technique, large FCM datasets with complexities similar to genomic techniques such as microarrays can be obtained relatively simply. The FC-HCS technique has many advantages over the conventional manual flow cytometric assays. First, only a few thousand cells are required for analysis. Consequently, replication and various experimental designs can be achieved from each sample collection. As this technique is almost entirely automated, mistakes in handling and staining large numbers of cell samples are minimized. These advantages dramatically enhance both the efficiency and the reproducibility of the high-throughput flow cytometric assays.

1.2 Graft versus host disease

GvHD occurs following allogeneic HSCT when immune cells in the graft attack the recipient's tissues. Very little is known about this potentially fatal disease [5] and for many that survive, the result is a significant decrease in quality of life [6, 7, 12, 13]. GvHD is the major limitation for broader application of HSCT which is the only curative treatment for many hematopoietic disorders [1].

GvHD occurs in two distinct forms, acute (aGvHD) and chronic GvHD (cGvHD). Here the term GvHD refers to both forms. GvHD requires the following three conditions to occur [14]: The graft contains enough immunologically competent cells; Antigens present in the recipient are different from those present in

the donor; and The recipient is incapable of mounting an effective immune response to destroy the graft.

1.2.1 Acute graft versus host disease

Manifestations of aGvHD can be described in three phases [1, 15-17], summarized in Figure 1.2. In phase one, preparative treatments such as chemotherapy or radiotherapy damage host tissues that subsequently secrete inflammatory cytokines. During phase two, the donor's T cell pathway is activated when it recognizes foreign recipient's antigens presented by host antigen-presenting cells. The donor's T cells proliferate and differentiate into effector cells. Finally, in phase three, Th1 inflammatory T cells' differentiation leads to the activation of cytotoxoic T cells, which in turn release a variety of inflammatory cytokines. This cytokine dysregulation results in skin, liver and gastrointestinal tract tissue damages.

aGvHD typically occurs within the first 100 days following the HSCT, usually between 14 and 42 days post-transplant [15]. The diagnosis and the subsequent grading of aGvHD usually involve skin and histopathologic examinations. However, a wide range of unrelated illnesses such as the basal cell necrosis, viral infection, and epidermolysis often exhibit similar symptoms and complicate the early diagnosis of aGvHD [18]. When an aGvHD diagnosis is made, it can be graded into four different levels based on the extent of tissue damage [16, 17]. The most important risk factor for developing aGvHD after a HSCT procedure is the degree of histoincompatibility in the human leukocyte antigen (HLA) between patient and donor [1]. Other aGvHD risk factors include increased age of donor and mismatched gender [1, 19].

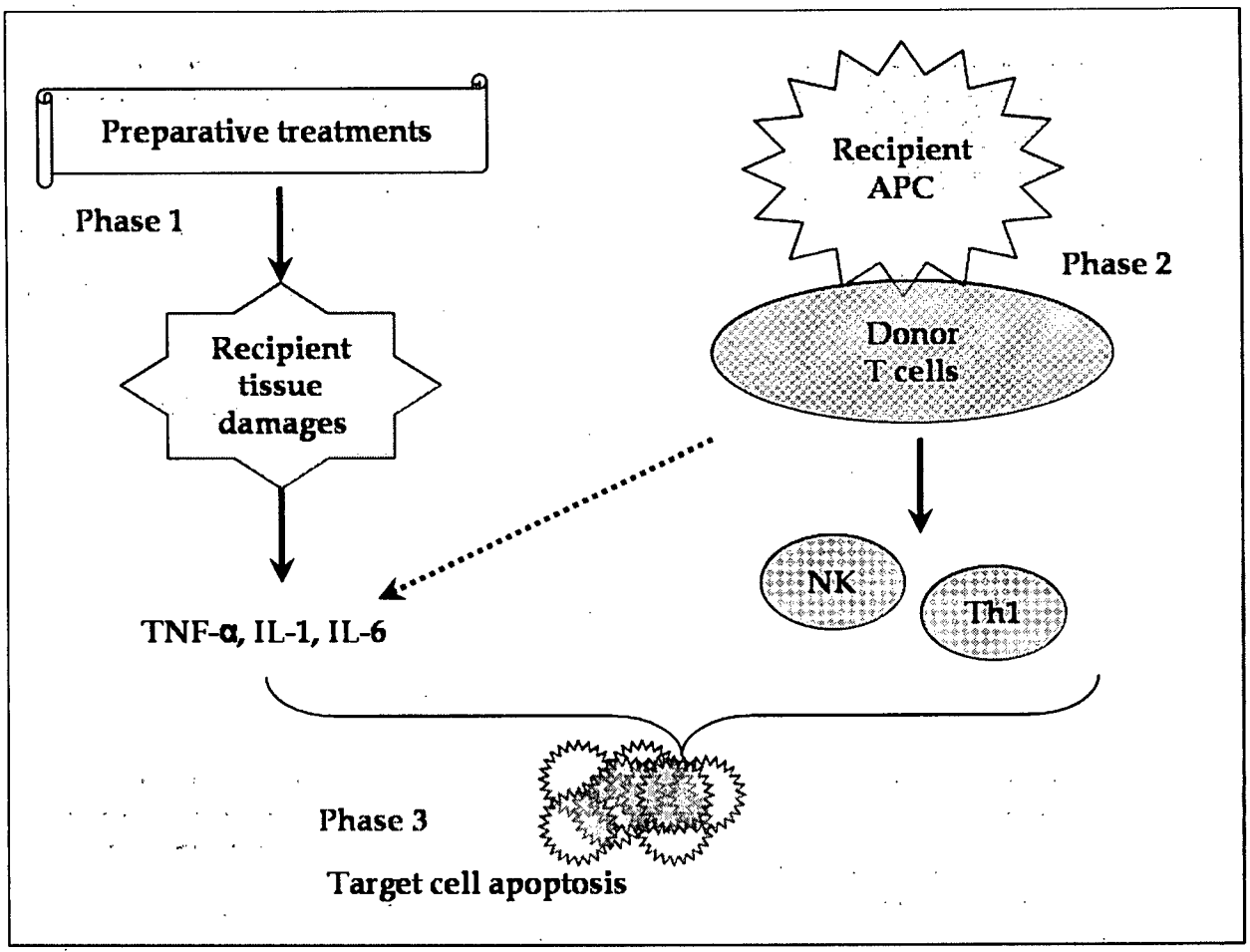


Figure 1.2 Pathophysiologic mechanism of aGvHD (adapted from Couriel *et al* [17])

Many immune cell populations have been identified as aGvHD mediators particularly through animal models and *ex vivo* graft treatment studies. They include the major (MHC) and minor histocompatibility complexes, dendritic cells, T cells, nature killer (NK) cells, macrophages, and cytokines [1]. The most prominent mediator is donor T cells [20]. T cell depleted BMT has been shown to reduce the occurrence of aGvHD significantly. However T cell depletion is rarely applied due to its severe side effects including increased rate of graft failure, prolonged

immunosuppressive state resulting in increased likelihood of fatal infections, and higher relapse rate [21-26].

Previous attempts to build a predictive model using $CD3^+$ T cells usually comprised small numbers of patients and exhibited conflicting results. Even though T cell depletion studies have demonstrated the importance of T cells in aGvHD development, many studies could not establish a significant correlation in the $CD3^+$, $CD3^+CD4^+$ or $CD3^+CD8^+$ T cells patterns (in either proportion or absolute number) to the onset of aGvHD [27, 28]. However, one study comparing nine moderate or severe aGvHD and 15 non-GvHD patients demonstrated significant correlation between the changes of three T cell subtypes ($CD4^+CD25^+$, $CD4^+CD69^+$, and $CD4^+CD134^+$) to the development of aGvHD [29]. Another study in humans demonstrated significant correlation between the rapid increase (>50%) of donor T cell chimerism and the development of moderate or severe aGvHD [30]. NK cells are also one of the known aGvHD mediators [1]. However, the exact NK cells population and their functions are not well defined. Some studies suggest NK cells contribute to tissue damage during aGvHD via secreting pro-inflammatory cytokines [31, 32] while others suggest that NK cells suppress GvHD effects [33, 34].

1.2.2 Chronic graft versus host disease

cGvHD affects 30-80% of patient surviving six months or longer after their HSCT procedure [35] and is the leading cause of non-relapse deaths. The pathophysiologic mechanism of cGvHD remains poorly defined despite numerous studies. Researchers have suggested the participation of both autoreactive and alloreactive T cells in the manifestation of cGvHD because the symptoms resemble autoimmune diseases. The development of cGvHD (Figure 1.3) might be the result of autoreactive T cells escaping negative selection in the damaged thymus caused by

the preparative treatments or aGvHD [36]. The resulting Th2 $CD4^+$ helper T cells facilitate synthesis of autoantibodies by host B cells [37].

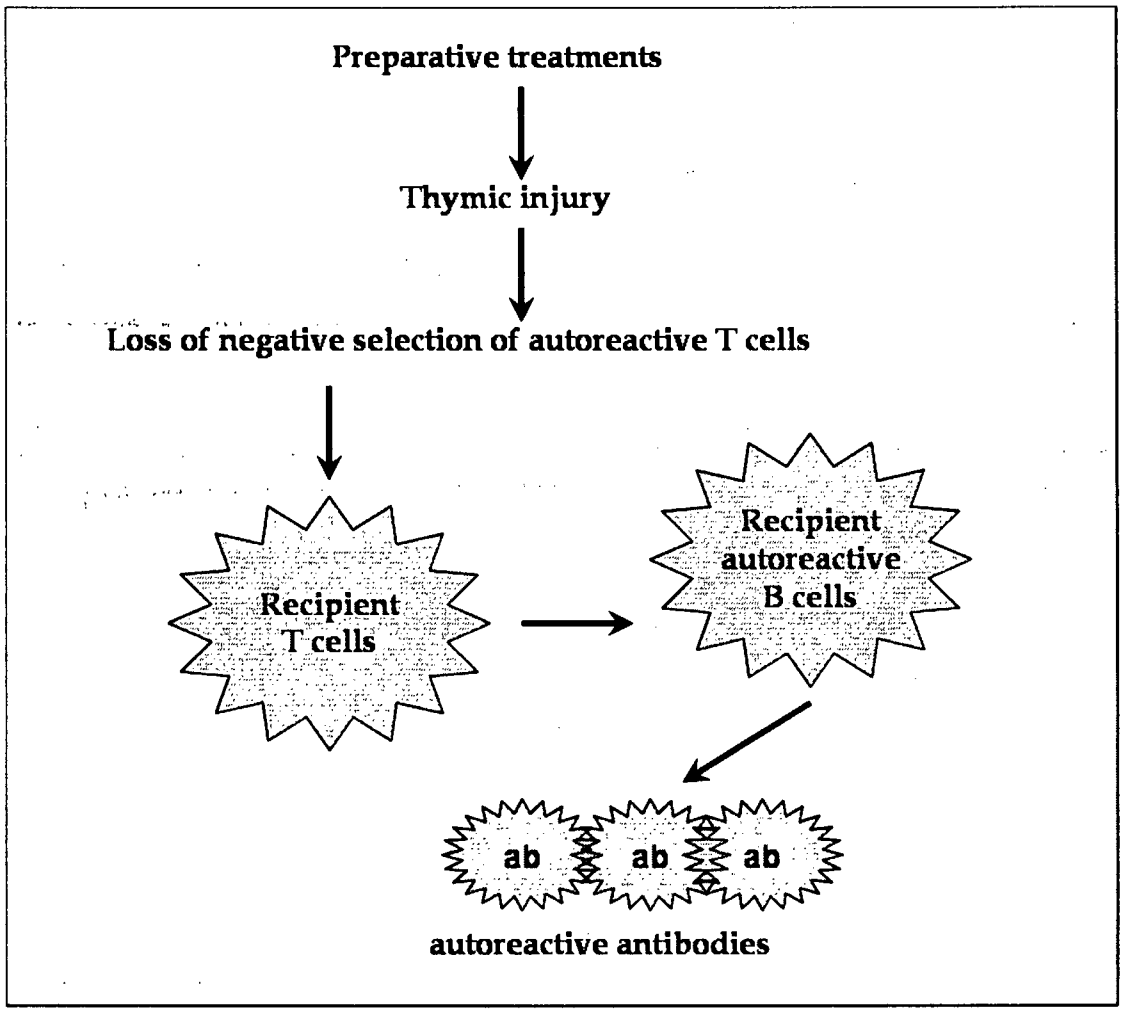


Figure 1.3 Pathophysiologic mechanism of cGvHD (adapted from Iwasaki *et al* [37].)

cGvHD usually occurs approximately four months after transplantation [38]. Similar to the diagnosis of aGvHD, cGvHD diagnostic methods are based on ambiguous clinical symptoms that involve skin and multiple internal organs.

cGvHD is usually differentially diagnosed apart from aGvHD and bacterial infections by at least one unique cGvHD symptom rather than the timing of the onset [37, 39].

cGvHD is graded into either limited or extensive disease based on the extent of skin tissue and internal organ damage. An alternative classification system is based on the cGvHD diagnosis time relative to the aGvHD status. Progressive cGvHD evolves directly from aGvHD and is associated with the most severe prognosis. Quiescent-type cGvHD with an intermediate prognosis occurs after an aGvHD free period. Finally, *de novo* cGvHD occurs without a prior history of aGvHD and has a better prognosis [37, 39]. The greatest risk factor associated with cGvHD is the prior incidence of aGvHD. The risk of developing cGvHD is more than ten times higher in patients with prior aGvHD [35]. Other factors include those common to aGvHD, such as the age of the patient and the degree of transplant histoincompatibility [39].

The known mediators of cGvHD include interleukin-18, T cells, and B cells [37]. Researchers have speculated that T cells are also the main mediator and effector cell type for the development of cGvHD. However, a recent randomized-trial study of T cell depletion contradicted previous findings [25, 26, 35, 40] and concluded that T cell depletion did not significantly reduce the incidence or the severity of cGvHD [21]. Attempts to build a predictive model using T cells or T cell subsets have resulted in conflicting or incomparable results. One study [41] demonstrated an insignificant correlation between the changes in $CD4^+$ and $CD8^+$ T cells and the onset of cGvHD. Another study utilizing both FCM and intracellular staining demonstrated a potential correlation between IL-4 producing $CD8^+$ T cells and cGvHD development. Other similar studies have focused on $CD34^+$ cells and suggested the importance of graft composition. However, they did not observed any significant correlation between any cell subset and the onset of cGvHD [27, 42-

45]. A pilot study of limited number of patients (six cGvHD and nine controls) focused on regulatory T cells with a $CD25^{high}$ phenotype and observed a significant increase of $CD4^{+}CD25^{high}$ T cells associated with the onset of cGvHD.

1.3 Temporal analyses

In comparison to the conventional static or multivariate analyses, temporal analysis is the most efficient analysis approach for the study of biological phenomena occurring over time [46]. In static analyses, values from a single fixed time point or the relationship between two fixed time points are examined. In multivariate analyses, values from multiple time points are examined as independent variables. Only in temporal analyses, values from multiple time points are examined as a single entity, thus conserving the continuity and dynamic of time.

Other main advantages of temporal analyses are that they are generally more tolerant to missing values and non-uniform sampling rate, the two most prominent challenges in a clinical dataset. On the other hand, the major challenge in designing a time-course experiment is the sampling rate. If the experiment is under-sampled, temporal aggregation may occur [47]. Oversampling is not favourable because of the cost. There is no standard sampling rate as it is specific to the biological phenomenon under investigation and the instrumental error rate [47]. Other experimental and computational challenges in a temporal analysis were previously reviewed by Ramsay and Silverman [48].

1.3.1 Temporal analysis for flow cytometry data

The popularity of time-course studies has already prompted the development of temporal versions of many conventional statistic analysis methods. Examples of these include algorithms for analysis of variance [49, 50], functional principal

component analysis [48, 51], clustering [52-59], and classification [4, 60-63]. Most temporal analysis algorithms were designed for or tested on microarray data. In some cases, the algorithms are not applicable to FCM data. The most noteworthy difference in the analyses of a microarray dataset versus a FCM dataset is the underlying assumption. Many microarray analyses are based on the assumptions that gene expression values follow a normal distribution and most do not change. These assumptions fit well with the whole genome approach of microarrays. The same assumptions have no standing in FCM data where only known cell populations are measured from a limited and biased selections of antibody-fluorochromes, and manual sequential gating. Furthermore, results from sequential gating overlap in their targeted immune cell subpopulations. Thus, FCM data are potentially dependent and correlated. To the best of my knowledge, no study has been done on the distribution of individual or overall immune cell population changes. As a result, availability of temporal algorithms suitable for FCM data analyses is further limited. Below in sections 1.3.2 to 1.3.4, I have summarized the common temporal analysis procedures: time-series data representation, pre-processing, and classification, employed in the pipeline I developed in response to the shortcomings of existing analysis methodologies.

1.3.2 Representing temporal data

The first step into a temporal analysis is to transform the time-series data consisting of a set of discrete values at multiple time points into one or more functions. The purpose of this transformation step is to represent the data as coefficients in a formula. The most common way of representing a non-periodic time-series data is the B-spline [48, 64].

A B-spline is a linear combination of a basis function. Two parameters involved in a B-spline shape are basis function order (n-1) and a knot placement.

Generally, these parameters were selected to ensure adaptability of a B-spline to the original data pattern. If n is two, a B-spline is built on combinations of linear basis functions between each knot. The spline dictates smoothness between the two basis functions on each side of a knot. The order of the basis function is also determined by the degree of the derivative function to be analyzed. For an example, a cubic B-spline (n=4) will ensure smoothness and the availability of up to the second derivative for further analyses. In a B-spline, knots designate the beginning of a new basis function where a change in the pattern is available. Subsequently, knots are placed around regions where complex variation is expected. By specifying the location and the number of knots, one can enforce regions with complex variation, ensure tolerance to non-uniform sampling time, and induce smoothing. Presently, there is no standard for the basis order or the knot positions.

The B-spline, represented as coefficient values in a matrix, is flexible to fit large numbers of data points and allows relatively easy implantation of various calculations. Other data representation techniques include: P-spline, polynomial function, exponential basis, power basis, step-function basis [65] and the Fourier basis for periodic data [48, 64]. In this study, I utilized B-splines, the most robust representation of time-series data and investigated how to build a B-spline that best reflects the raw data pattern.

1.3.3 Data pre-processing - Smoothing & Registration

When time-series datasets are transformed into one or more combinations of functions via methods such as B-splines, the resulting pattern is automatically smoothed. The purpose of performing an additional smoothing procedure is to minimize fluctuations in the pattern that might be motivated by random experimental errors instead of the underlying biological phenomena. The

commonly used smoothing methods: least square, roughness penalty, and positive smoothing methods, are briefly described below.

The basic aim of these smoothing methods is to determine the balance between goodness of fit to the intrinsic or external pattern and amount of information lost. For the least squares smoothing method, patterns are changed throughout the available time range in order to minimize sum of squared errors in fitting a simulated model with normally distributed and independent residues. For the roughness penalty method, variances among the patterns are decreased throughout. The amount of smoothing is unbiased and is controlled by the user specified parameter λ. The positive smoothing method modifies every pattern by enforcing a logarithmic property, thus adding positive constraint throughout. Overall, there is no standard degree of smoothing by any method, and as a result, this increases the complexity of long time-series data analyses.

Another form of smoothing where random experimental errors are estimated and removed is the signal-plus-noise model. Essentially, a set of observed values Y_{ij} from sample j in class i are divided into global base value λ_0, class signal $\Lambda\alpha_i$, individual signal variation γ_{ij} and individual experimental errors ε_{ij} (Equation 1.1). The parameters can be estimated via algorithms such as the Expectation Maximization (EM) algorithm.

$$Y_{ij} = \lambda_0 + \Lambda\alpha_i + \gamma_{ij} + \varepsilon_{ij}$$

Equation 1.1 Signal plus noise model

Registration in a temporal analysis refers to stretching and shrinking the time index of each observed data to fit an overall time-series data pattern [48]. This step is often necessary because the phenomena being measured may not follow the linear

time scale the data. Registration is particularly important for long time-series data and clinical data in order to synchronize different patient response times. Examples of registration methods are the landmark and continuous fitting criterion [48].

Landmark registration is biased as it depends on prior information. First, a minimum of two landmarks for the two ends of each time-series data are identified. More landmarks can be identified based on specified patterns or prior information such as disease diagnosis or combinations of both. There must be an equal number of landmarks specified for each set of time-series data. The landmark registration algorithm then transforms the time axis so that corresponding landmarks in the time-series dataset are comparable [48, 66]. Continuous fitting or global registration is unbiased and aims to minimize the least square value between the registered patterns and their means. At each iteration, amplitude differences between the patterns and their mean are minimized by modifying the time scale [48]. Other registration methods include shift registration, which applies a constant shift to the time index and warping function, which combines registration and smoothing.

1.3.4 Classification

Classification algorithms analyzing time course data can be categorized into two approaches. The first approach utilizes conventional multivariate analyses such as principal component analysis [67, 68], singular value decomposition [69], correlation analysis [70], and support vector machines [71]. These algorithms omit the time dependency of the time course data. The second approach includes the time dependency in the time course data. Thus, the second approach is generally considered more efficient in time course study and applicable to study with missing values and non-uniform sampling rate.

Algorithms categorized in the second approach include nonparametric curves discrimination [72], functional linear discriminant analysis (FLDA) [4], mixture functional discriminant analysis [73], predictive modular neural networks [60], etc. Among all these classification algorithms, only FLDA was designed specifically for sparsely sampled datasets. Therefore, FLDA [4] is deemed the most suitable temporal classification algorithm for the analysis of the present clinical dataset.

FLDA is a B-spline based method. Similar to the static linear discriminant analysis, it provides an easily interpretable classifier. In the static linear discriminant analysis [74], the classification of test data can be made via multiplying weight values (b_1, b_2,... b_m) with test data values (x_1, x_2, ... x_m) from the corresponding parameter (Equation 1.2). These weight values are determined for each parameter using a training dataset with multiple and independent parameters. The absolute value of these weight values also represents how strongly each test data will be accounted for in the classifier.

$$Group = a + b_1 x_1 + b_2 x_2 ... + b_m x_m$$

Equation 1.2 Static linear discriminant classification

For time-series datasets, FLDA builds a classifier by estimating the signal-plus-noise model (Equation 1.1 and Figure 1.4) using a training dataset where the first three parameters (global base value, class signal, and individual signal variation) are denoted by the B-spline matrix S_{ij}. In the FLDA classifier, weight values (Equation 1.3) are determined for a set of sampled time points using variables estimated in the signal plus noise model (Equation 1.1). Classification is made by multiplying the difference of the test data with the corresponding global base values λ_0, to the weight values at the sampled time points (Equation 1.4). The

polarity of the linear discriminant value $\hat{\alpha}_x$ is used to determine the classification of a test data into one of the two groups. In a FLDA classifier, large absolute weight values are assigned to time points where there is large separation between the estimated class signals. As a result, small differences between test data and the global base values at those time points will be accounted more heavily than differences at other time points in the overall classification (Equation 1.4).

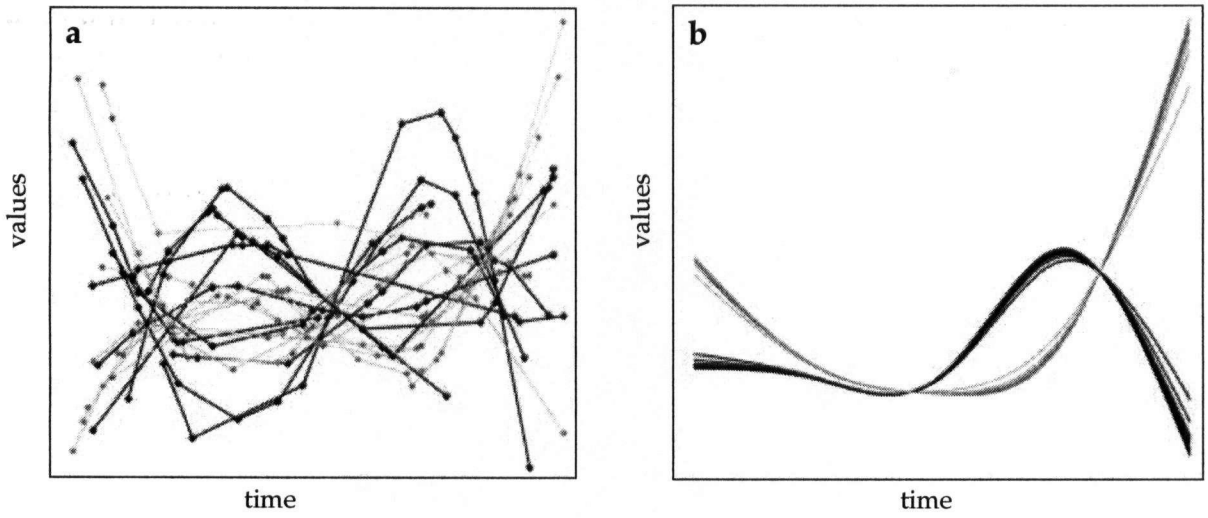


Figure 1.4 An example of the FLDA signal plus noise training from the raw data (panel a) to the estimated signals (panel b), adapted from James and Hastie [4]

$$weight = (\Lambda^T S_X^T \Sigma_X^{-1} S_X \Lambda)^{-1} \Lambda^T S_X^T \Sigma_X^{-1}$$

Equation 1.3 FLDA weight values at specified time points

$$\hat{\alpha}_x = weights \cdot (X - S_X \lambda_0)$$

Equation 1.4 Functional linear discriminant value

Validation techniques for classifiers can be categorized into four groups: external test set, resubstitution, bootstrap, and cross-validation. The external test set is the best validation technique because it provides unbiased error estimation by validating the classifier using a new dataset with prior knowledge of class assignment. Unfortunately, the external test set validation is usually impractical in studies with a small sample size. The other three groups of validation techniques utilize the same dataset for both training and validation of classifiers. Resubstitution is a method where the same training dataset is used as the test dataset and it usually underestimates the classifier error considerably [75]. Similar to resubstitution, bootstrap repeatedly re-analyzes a subset of the training dataset by selecting profiles with replacement. K-fold cross-validation also repeatedly re-analyzes a subset of the training dataset but without replacement. Error is estimated by k training datasets, each time leaving a subset of the original dataset as the testing dataset. If k is set to the size of a dataset, then leave-one-out cross-validation (LOOCV) is performed and a single data point is used as the testing dataset each time. Studies have shown that bootstrap technique generally results in biased error estimation with small variance while the cross-validation results in less biased estimation with large variance [76].

1.4 Sample size calculations

Sample size calculation or power analysis estimates the certainty of detecting an effect, which is inversely proportional to the probability of a false negative (a type II error) result. The estimated power depends on the tolerance of type I errors (significance level, α) and the data variance. In the case of a pilot project, sample size calculation may be used to determine how many samples are needed for a future study in order to achieve a certain power level. Generally, sample size calculation consists of four steps [77]:

1. Specify α
2. Specify hypothesis-testing procedure
3. Sampling of the original dataset to create simulated datasets of different sizes
4. Estimate power of the analysis based on multiple stimulated datasets

Most of the sample size calculations vary with their choice of hypothesis testing and sampling methods. Most include assumption of normal or known distributions. Power analysis by location shift [78] is entirely nonparametric and incorporates the average X & Y method for a conservative power estimation. It is a bootstrap based method where multiple stimulated datasets from the empirical cumulative distribution function (ECDF) are compared with the Wilcoxon test. It considers variances from the two original datasets separately and determines the overall power as the average of the power estimated from the two original datasets.

1.5 Thesis goals

Previously, high-throughput methods have proven useful in probing unknown diseases [79, 80]. High-throughput FCM has never been applied to the study of GvHD because of the technical difficulties of FCM were only resolved with the recent development of FC-HCS. As manifestations of GvHD are based on the immune system, it was thought a high-throughput analysis on immune cell changes in the blood following allogeneic HSCT might prove to be successful in predicting the onset of GvHD and elucidating their mechanisms. The main hypothesis of the present study was:

Onset of aGvHD or cGvHD can be predicted by identifying patterns of cellular markers in peripheral blood mononuclear cells (PBMCs) via FC-HCS.

It is suspected that there are multiple immune cells and pathways involved in GvHD disease manifestation [81]. The global approach used in this study should be beneficial in the further elucidation of the disease. The main goal of the present study was to develop a bioinformatics analysis pipeline that can analyze high-throughput clinical FCM data and if possible identify immune cell populations that may be used in a diagnosis of either aGvHD or cGvHD. The specific aims were:

1. Assemble a suitable temporal analysis pipeline to process the high-throughput FCM dataset
2. Identify one or more immune cell populations with potential discriminate power for either aGvHD or cGvHD
3. Construct diagnostic models for aGvHD and cGvHD
4. Recommend an analysis methodology for an extended study.

CHAPTER 2 PATIENTS AND METHODS

2.1 Overview

One hundred and twenty-three subsets of PBMCs were obtained by FC-HCS using samples taken from 31 patients (Table 2.1) at multiple time points. The quality of the dataset was assessed and suspicious outliers removed. This dataset was then separated based on patients' GvHD diagnoses and analyzed by a temporal classification algorithm. In order to verify the hypothesis of the present study, temporal patterns of immune cell populations' abundances that appeared to correlate with the onset of either aGvHD or cGvHD were identified and visually inspected. Finally, sample size calculations were performed based on the top classifiers in order to estimate statistical power of the current and future studies.

2.2 Study patients

Thirty-one patients who received HLA matched BMT from either sibling (SIB) or matched-unrelated donors (MUD) were enrolled at the Moffitt Cancer Center with the approval of the institutional review board. On average, there were 14 (±3) samples per patient, collected approximately every ten days (±14). Samples were collected from 0 to 16 days (average 6 ± 4 days) before the transplantation and until 49 to 400 days (average 125 ± 81 days) after the transplantation. This was a heterogeneous dataset. Among the 31 patients, there were seven different underlying hematopoietic disorders (Table 2.1) and at least four different types of pre-transplant treatments (data not shown). Twenty-one patients were diagnosed with aGvHD on average 36 days (± 18 days) post-transplant. Seven of these aGvHD patients were later diagnosed with cGvHD from 98 to 446 days post-transplant. The diagnosis and grading of GvHD were performed using previously published criteria [82]. Details of the stem cell source, GvHD diagnosis time, and maximum GvHD grades are summarized in Appendix A.

Table 2.1 Characteristics of the 31 patients recruited for the study.

Characteristics	Subtypes	Incidence (% of total population)
GvHD		
	aGvHD	21 (68%)
	aGvHD and survived	9/21 (29%)
	aGvHD then died or withdrew from the study	5/21 (16%)
	Progressive or quiescent-type cGvHD	7/21 (23%)
	non-GvHD	7 (23%)
	non-GvHD with records past 100 days post-transplant	4/7 (13%)
	non-GvHD died or withdrew before 100 days post-transplant	3/7 (10%)
	De novo cGvHD	3 (10%)
Underlying disorders		
	AML	11 (35%)
	MDS	1 (3%)
	MDS-AML	3 (10%)
	CML	5 (16%)
	NHL	7 (23%)
	MPD	1 (3%)
	CLL	2 (6%)
	ALL	1 (3%)
Donor-recipient relationship		
	SIB	17 (55%)
	MUD	7 (23%)
	unknown	7 (23%)
Total		31

2.3 Sample preparations and flow cytometry high content screening

Blood samples were obtained both pre- and post-transplantation on an approximate weekly basis. PBMCs were isolated using Ficoll-Hypaque technique. The samples were divided into ten aliquots in 96 well plates. Each aliquot was stained with four different antibodies out of the total 25 (Table 2.2) used in the present study. The four antibodies used per group were attached with different fluorochromes and the combinations of antibodies-fluorochromes were designed to target different immune cells (Table 2.3). Six aliquots named '1Activation', '2Activation', '3Activation', 'resting/activate (rest/act) T helper', 'rest/act T suppressor', and 'T cells' targeted subsets of T cells and their functional states. The other four aliquots targeted myeloid cells, B cells, NK cells, and T cell receptor (TCR) via aliquots so named.

Depending on the sample number and frequency, one or more 96-well plates were used for each patient. Samples were usually plated one row per aliquot and ordered in columns by their sampled time. These 96 well plates were stained with antibodies and then analyzed using multi-parameter FCM as part of the FC-HCS technique previously described [2]. Batch gating analysis of the FCM was performed using FlowJo software (Tree Star, Inc, Oregon) on one- or two-dimensional plots to generate abundance values for maximum 123 subsets of immune cells for each sample (Appendix B). The sample preparations and the FCM gating were previously performed by the Moffitt Cancer Center and Dr. Maura Gasparetto (BC Cancer Research Centre).

Table 2.2 Annotated functions and selected literature references on the 25 cell surface antigens used.

Gene Name(s)	Functions	Literature
CD2	Activation of T and NK cells	[83]
CD3	Known to be involved in phase II of acute GvHD	[84]
CD4	Regulation of interleukin-2 biosynthesis; T-cell differentiation; Known mediator in GvHD	[85-87]
CD5	Cell proliferation and recognition	[88, 89]
CD8	Know to be involved in phase II of acute GvHD	[84]
CD8β	T-cell activation, MHC class I binding	[90, 91]
CD10	Also known as common acute lymphoblastic leukemia; marks early lymphoid progenitor cells	[92]
CD14	Cell surface receptor linked signal transduction; inflammatory response	[93]
CD15	Neutrophil adhesion	[94]
CD16	Immune response	[95]
CD19	B cells marker	[96]
CD20	B cells activation; immune responses; signal transduction	[97]
CD22	Cell adhesion; antimicrobial humoral response	[98, 99]
CD25	Marker for strong or prolonged antigen stimulation	[96]
CD33	Cell adhesion	[100]
CD44	Cell adhesion	[101]
CD45	Lymphocytes activation	[102]
CD45RA	T cells in resting state	[103]
CD45RO	T cells in activating state	[103]
CD56	NK cells marker	[96]
CD69	Early T cell activation antigen, acute graft rejection	[104]
CD122	Cytokine receptor	[96]
CD134	Tumor necrosis factor receptor superfamily, marks activated $CD4^+$ cells	[105]
TCRab	T cell activation	[96, 106]
TCRgd	T cell activation	[96]

Table 2.3 The combinations of antibody - fluorochromes used in each of the 10 aliquots available per sample.

Aliquot #	Aliquot name	FITC	PE	PerCP	APC
1	Myeloids	CD15	CD45	CD14	CD33
2	T cells	CD4	CD8β	CD3	CD8
3	NK cells	CD16	CD2	CD3	CD56
4	B cells	CD10	CD20	CD19	CD22
5	TCR	TCRab	TCRgd	CD3	CD5
6	1Activation	CD44	CD25	CD3	CD69
7	2Activation	CD4	CD134	CD3	CD8
8	3Activation	CD4	CD122	CD3	CD8
9	rest/act T helper	CD45RA	CD45RO	CD3	CD4
10	rest/act T suppressor	CD45RA	CD45RO	CD3	CD8

2.4 Temporal analysis pipeline

A temporal analysis pipeline consisting of three steps was assembled specifically for the high-throughput clinical FCM dataset (Figure 2.1). Step one involved a quality assurance (QA) test in two parts. The purpose of this QA test was to identify values motivated by experimental errors. Step two involved the data transformation via a PERL script. Finally, step three involved the temporal classification via FLDA. The resulting classifiers were ranked based on their potential discriminative power for the onset of either aGvHD or cGvHD.

Step 1: Flow Cytometry Quality Control

Flow Cytometry

↓

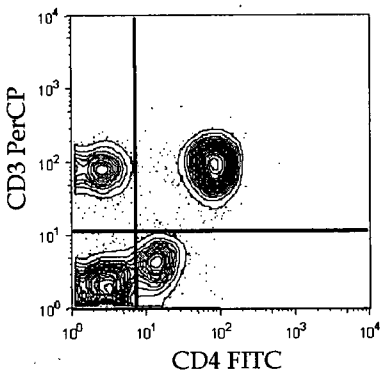
Flow cytometry data

↓

Quality assurance test

↓

Manual gating and data extraction



Step 2: Data transformation

Cell lineage abundances proportion to the PBMCs

	Markers	
	CD3+	CD3+ CD4br
Patient # 1 day 0	30	25
Patient #1 day 7	10	10
Patient #1 day 14	47	40

As is

↓

Proportion dataset (%)

	Time	
	Day 0	Day 7
Patient #1	30	10
Patient #2	6	39
Patient #3	27	52

Mononuclear cells concentration

↓

Concentration dataset (mm^3)

	Time	
	Day 0	Day 7
Patient #1	2.1	0.2
Patient #2	10.5	5.07
Patient #3	1.6	7.3

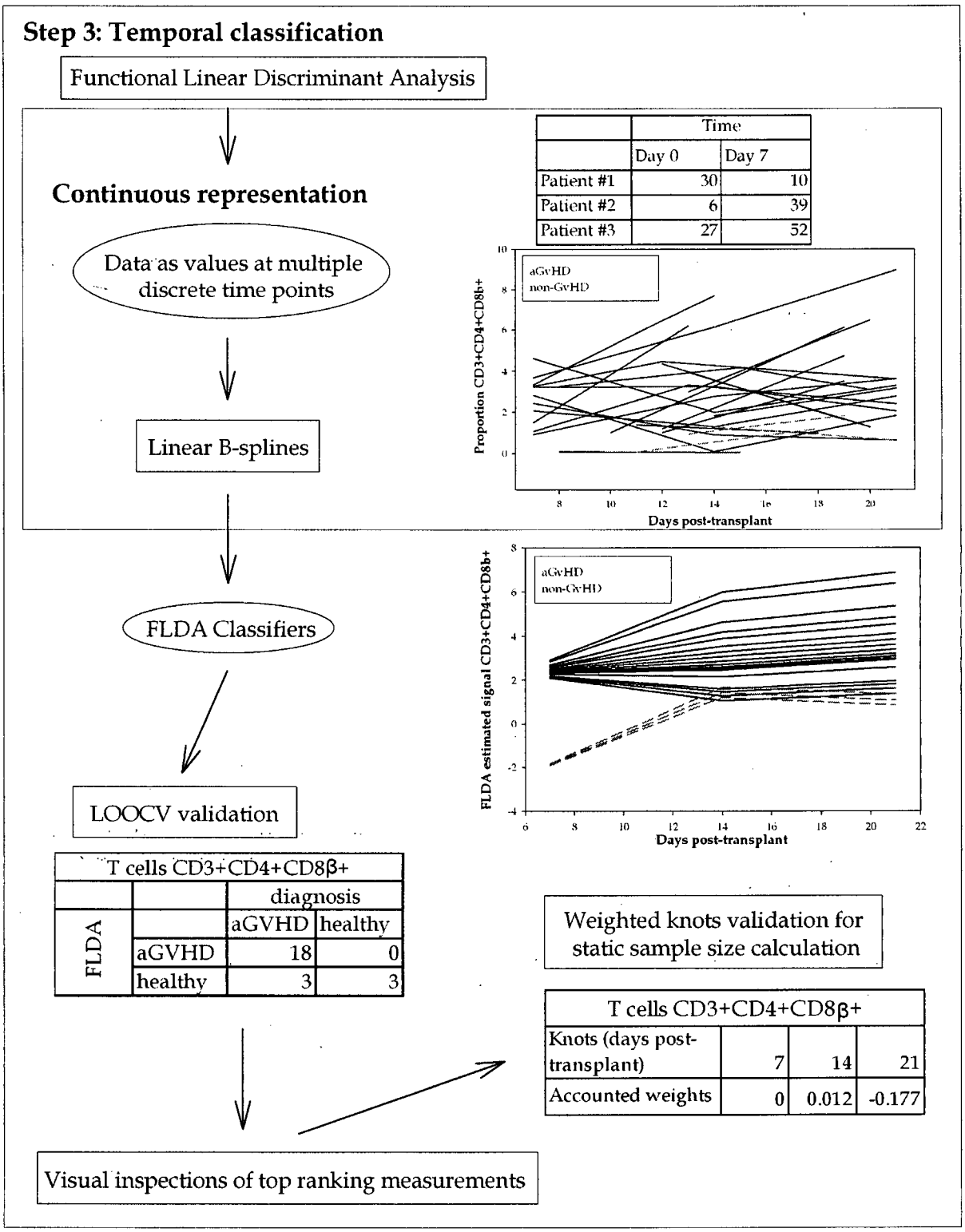


Figure 2.1 Temporal analysis pipeline designed for the high-throughput clinical FCM dataset.

2.4.1 Quality Assurance

The basic assumption for the main QA test was that distributions from common light scatter intensities of cells in different aliquots of the same sample should be similar [3]. Outliers were identified through visual inspection of ECDF, density plots and box plots. Part one of the QA test was performed on ungated data by Dr. Le Meur (Fred Hutchinson Cancer Centre) where the QA assumption was tested on intensities of the FSC and SSC measurements for all cells. Raw flow cytometry standard (FCS) files from a FACSCalibur (Becton Dickinson (BD), San Jose, CA) were obtained and analyzed in R via the rflowcyt package [107].

In part two, I tested the QA assumption based on the intensities of the FSC, SSC and CD3-PerCP antibody-fluorochrome for $CD3^+$ and $CD3^-$ populations separately. FCS files of the gated $CD3^+$ and $CD3^-$ populations were exported from FlowJo. Excess keywords in the FCS files were removed via a PERL script (fixFCS.pl, Appendix C) to generate a file format compatible with the rflowcyt package. Unlike the QA test on ungated data where up to ten aliquots were available per sample, there were only five or seven aliquots available for the QA test of gated data. Consequently, it was more difficult to identify outliers visually. In order to retain most of the limited data for the subsequent classification analysis, only obvious and singular outliers were identified. Criteria for outlier identification in the QA test on gated data were:

1. One outlier per sample
2. The outlier pattern must be visually different from all other aliquots
3. The outlier pattern cannot be visually explained by the observed general variations.

Under these criteria, outliers were identified and all their associated sub-gates were removed from subsequent analyses. Data with putative outliers that did not fit the above criteria were retained. Finally, all outliers and unusual patterns were mapped

back to the original plating chart in order to investigate the distribution of outliers on the 96-well plate.

2.4.2 B-spline parameters evaluation

The effects of two B-spline parameters: basis order and knot placement were tested using a time-series data from patient #2 between 0 and 13 weeks post-transplant. This patient was selected because of its uniform sampling rate and a single missing value at week one. The effects of these parameters on the overall fit between the resulting B-spline and this data were evaluated and were used as models in determining the optimal B-spline parameters for the dataset. However, because of the sampling rate disparities and the massive numbers of values available, this data may not be representative of the entire dataset.

First, the effects of different basis orders were examined using three B-splines created with two, three or four basis order creating linear, quadratic, and cubic basis functions. Knot placement of one knot for every sampled time point was used for all three B-splines. Secondly, the effects of different knot placements were examined with four B-splines consisting of linear basis functions. The four knot placements, with decreasing knot frequency were:

1. A weekly knot placement including one knot at week one post-transplant when patient information was not available
2. Knots at every sampled time points (no knot at week 1)
3. A bi-weekly knot placement covering the entire time range (0, 2, 4, 6, 8, and 13 weeks post-transplant)
4. A tri-weekly knot placement covering the entire time range (0, 3, 6, 9, and 13 weeks post-transplant)

2.4.3 Data transformation

Step two in the temporal analysis pipeline (Figure 2.1) involved data transformations via a PERL script (viz_days.pl; Appendix D). The 123 gated immune cell abundances were exported to text files using FlowJo software. The FCM data were then combined with immune cell concentration data and transformed into a proportion dataset and a concentration dataset. The proportion dataset contained all 123 subsets of immune cells; each corresponding to the proportion of cells (proportion of either the total PBMCs or total $CD3^+$ cells) in the gate. The mononuclear cell (MNC) concentration values (mm^3) were obtained separately using different samples taken from the same group of patients at multiple time points. The concentration dataset was obtained by multiplying each proportion value with the MNC concentration of samples taken at the closest date. Both datasets were tested because they may contribute different insights into immune responses.

The PERL script viz_days.pl (Appendix D) also rearranged the file layout and the time scale. Originally, data was recorded as the number of days after the BMT. Viz_days.pl combined the known aGvHD diagnosis date, BMT date, and the sampled time points to modify the time scale from days post-transplant into days from the aGvHD diagnosis. For patients unaffected by aGvHD, the average date of aGvHD diagnosis observed in the current dataset (36 days post-transplant) was used as the synchronization event. The non-GvHD patient data were transformed so they could be compared to the aGvHD patient data. Thus, patients' responses were synchronized by two events resulting in two time scales in days post-transplant and days from aGvHD diagnosis. The PERL script also excerpted three parts of the data for time ranges representing patterns right after BMT, and before and after aGvHD manifestation. Consequently, results derived from these three time ranges should be useful in elucidating the onset, manifestation, and progression of GvHD. In the end, three separate dataset of different time ranges were obtained:

1. 7 to 21 days post-transplant
2. 21 to 0 days prior to aGvHD diagnosis
3. 0 to 21 days from aGvHD diagnosis

2.4.4 Temporal classification

In step three of the temporal analysis pipeline (Figure 2.1), different combinations of GvHD and non-GvHD patient groups (Table 2.1) were analyzed using FLDA for both the proportion and concentration datasets. The first comparison was between the 21 aGvHD and the 4 non-GvHD patients. This comparison was intended to identify temporal patterns from one or more subsets of immune cells that could predict aGvHD reliably and precisely prior to the manifestation of clinical symptoms, or elucidate pathophysiologic pathway of aGvHD during the clinical manifestation of aGvHD. Supplementary comparisons including 17 Grade II-IV aGvHD vs. 4 non-GvHD patients and 12 Grade III-IV aGvHD vs. 4 non-GvHD patients were also performed. The second comparison was between seven patients diagnosed with both aGvHD and cGvHD and nine patients diagnosed with only aGvHD. This comparison was intended to identify temporal patterns from one or more subsets of immune cell that are predictive of progressive or quiescent-type cGvHD either after the BMT or during the manifestation of aGvHD.

A PERL script (FLDA_MATLAB.pl; Appendix E) read in the specified data and outputted necessary MATLAB (MathWorks, Inc. Boston) commands to build a FLDA classifier for each subset of immune cells and each patient group comparison. The PERL script also acted as a filter to omit data with fewer than three available sampled time points per patients in each of the selected time ranges, or fewer than three available patients per group. Because of missing values from the sampled time point and limited number of available aliquots, not all the identified immune cell

populations and patients were included in each analysis. The qualified data were then analyzed via the FLDA analyses with a linear B-spline and a weekly knot placement.

LOOCV was performed on the FLDA classifiers. The validation results were used to rank the FLDA classifiers and their corresponding subsets of immune cells as the values were directly proportional to the potential discriminative power of the temporal patterns. Top ranking classifiers were then inspected visually via time plots of the FLDA estimated signals and the raw data in the analyzed and extended time ranges.

2.5 Static sample size calculation

A static sample size calculation pipeline (Figure 2.2) was implemented in the R package 'PALS' (Power Analysis by Location Shift) based on the location shift hypothesis [78]. The analysis was performed on values from the top FLDA ranking immune cell populations closest to the time point where the class signal separation was the greatest based on the adjusted weight values (section 2.6). The purpose of this analysis was to estimate statistical power of the present and future studies.

Briefly, in the sample size calculation (Figure 2.2), simulated datasets were generated from random samplings of two ECDFs corresponding to the two groups of observed values. The simulated datasets were then analyzed using the Wilcoxon test for statistical significance ($\alpha \leq 0.1$). This was repeated 10,000 times to estimate the power of the study. Each ECDF was used to simulate data representing both groups. The average of the power from each ECDF was obtained. In the interest of time, an upper and lower limit of 100 and 0 was set for the random sampling from the proportion dataset.

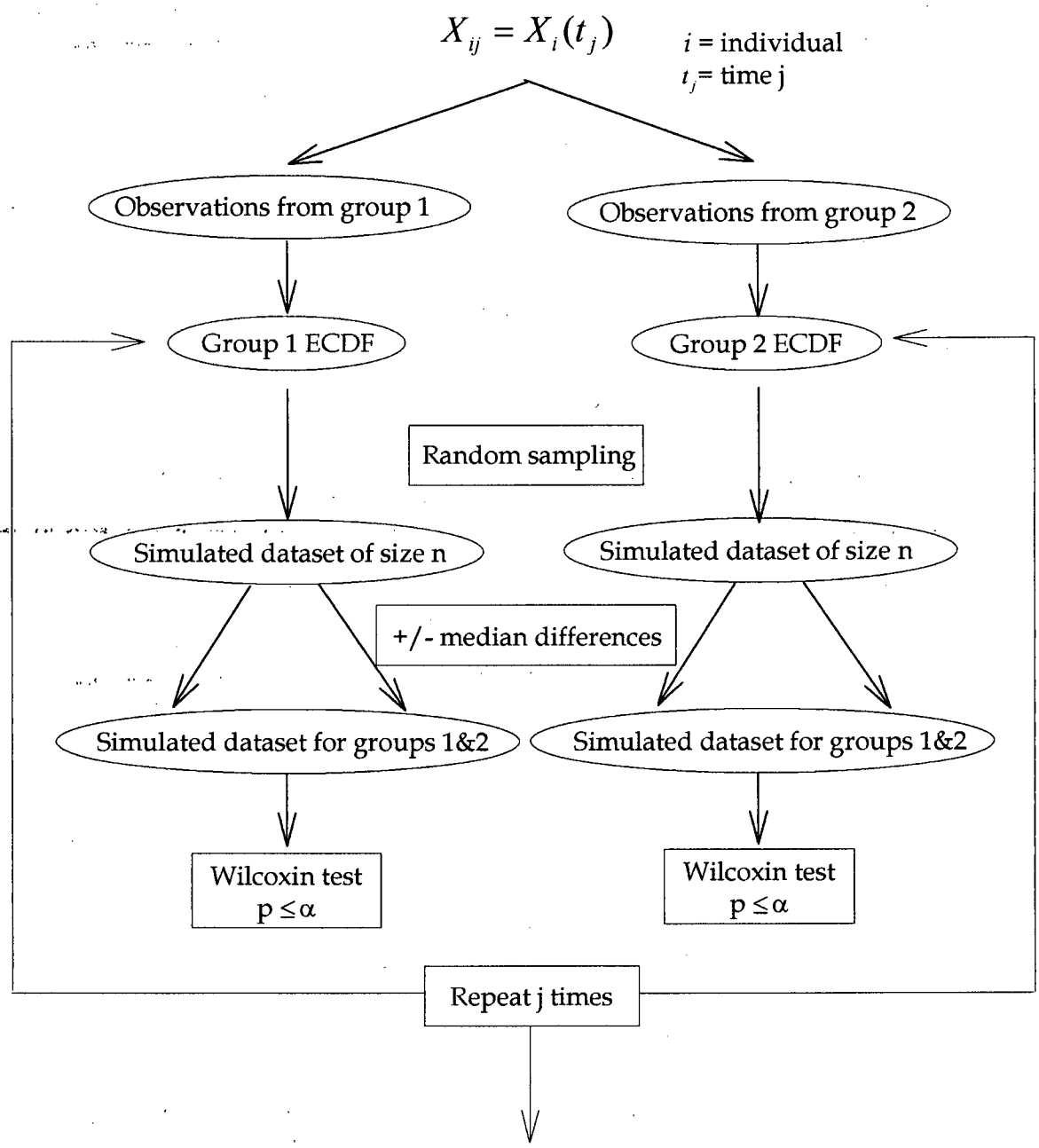


Figure 2.2 Static sample size calculation pipeline.

For the first comparison, between 21 aGvHD and three out of four non-GvHD patients, observed values from the immune cells $CD3^+CD4^+CD8\beta^+$ taken closest to 21 days post-transplant were used. For the second comparison, between seven aGvHD & cGvHD, and nine aGvHD only patients, observed values from the immune cells $CD3^+TCRab^+CD5^+TCRgd^+$ taken closest to seven days prior to the aGvHD diagnosis were used. Various simulated dataset sizes were used for both comparisons. However, sizes of aGvHD simulated data were two times larger than the non-GvHD simulated data sizes in order to imitate the aGvHD manifestation rate in the BMT patients. On the other hand, equal sizes were assigned between the aGvHD & cGvHD and aGvHD simulated datasets.

2.5.1 Weight values in the functional linear discriminant analysis classification

In a FLDA classifier, large absolute weight values are assigned to time points where there are large separation between the estimated class signals (Equations 1.3 and 1.4). For the static sample size calculation, weight values were determined at each of the weekly knots originally used in the FLDA analysis (section 2.4.4). The reliability of the weight values were accounted for by multiplying the weight value with the ratio of the corresponding total number of observed values and the total number of expected values. In the range between half the knot interval away from each knot on both sides, the number of expected and observed values from the class with the least number of patients was noted in order to obtain the most conservative estimations.

A hypothetical example of accounted weight values is described using a FLDA classifier built using fabricated samples from 21 aGvHD and three non-GvHD patients taken between 7 and 21 days post-transplant. Weight values for the weekly knots at 7, 14, and 21 days post-transplant were assumed to be 2, 0.5 and 3. Sample availability for the three non-GvHD patients were assumed to be two values at

seven days post-transplant, three at 14 days post-transplant and one at 21 days post-transplant. In a weekly sampled rate, one value was expected for every patient and every week. As a result, the accounted weight values were determined to be 4/3, 0.5 and 1 at each knot. Due to the lack of available values for the smaller non-GvHD patient group (between 18 and 21 days post-transplant), the estimated class separation at 21 days post-transplant was not reliable. By taking the actual number of values available around each knot into account, the greatest and the most reliable class separation was at 7 days post-transplant.

CHAPTER 3 RESULTS - QUALITY ASSURANCE AND B-SPLINE PARAMETERS

3.1 Quality assurance on ungated data

From the QA test on ungated data, two outliers corresponding to aliquots 'Myeloids' and '3Activation' were identified in the FSC intensity density plots for patient #6 (Figure 3.1). One of the two outliers (aliquot 'Myeloids') was also identified in the ECDF plots (data not shown). Box plots failed to depict details in the distributions while most differences were observed in the FSC distribution, compared to the SSC ones [3].

3.2 Quality assurance on gated data

3.2.1 Singular outliers

In the QA test on gated data, outliers such as aliquot '3Activation' from patient #6's samples taken at 76 days post-transplant (Figures 3.2 and 3.3) were selected using the criteria outlined in section 2.4.1,. In total, 29 aliquots (< 0.4% of the dataset) were identified as visually significant outliers (Table 3.1) and removed from the dataset. While the outlier '3Activation' can be easily identified in the FSC and SSC intensities density plots (Figures 3.2 and 3.3), the same aliquot would not have been identified as an outlier due to general variations observed in the density plot of CD3-PerCP intensity (Figure F.1). Consequently, CD3-PerCP was not used in the outlier identification. Results from the CD3-PerCP density plots and their potential role in gate quality control are described in Appendix F.

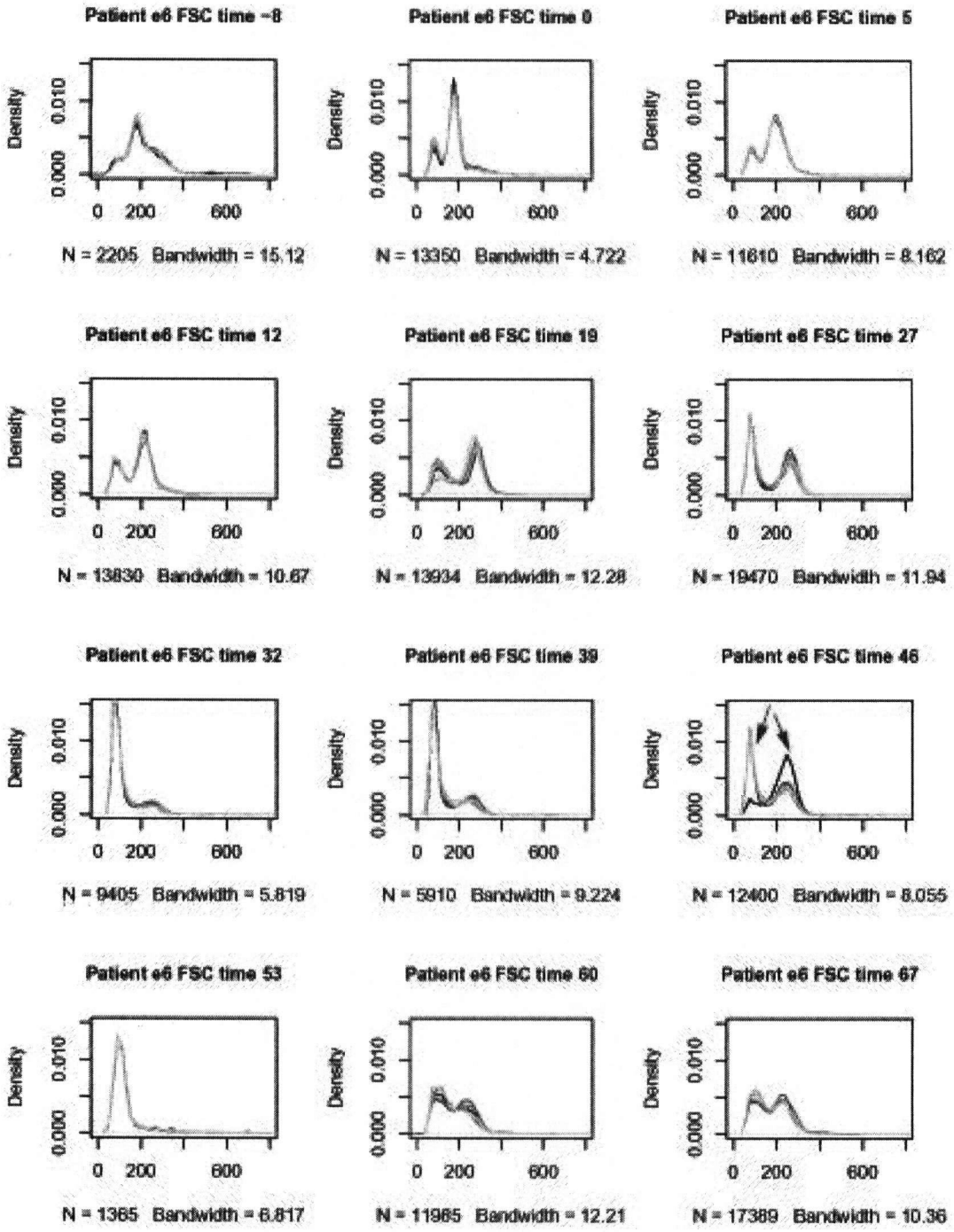


Figure 3.1 Density plots of the FSC intensity of different aliquots of samples taken at 12 different time points (adopted from [3]). At day 46, the two red arrows show distributions corresponding to aliquots 'leukocyte' and '3Activation' are substantially different from other aliquots.

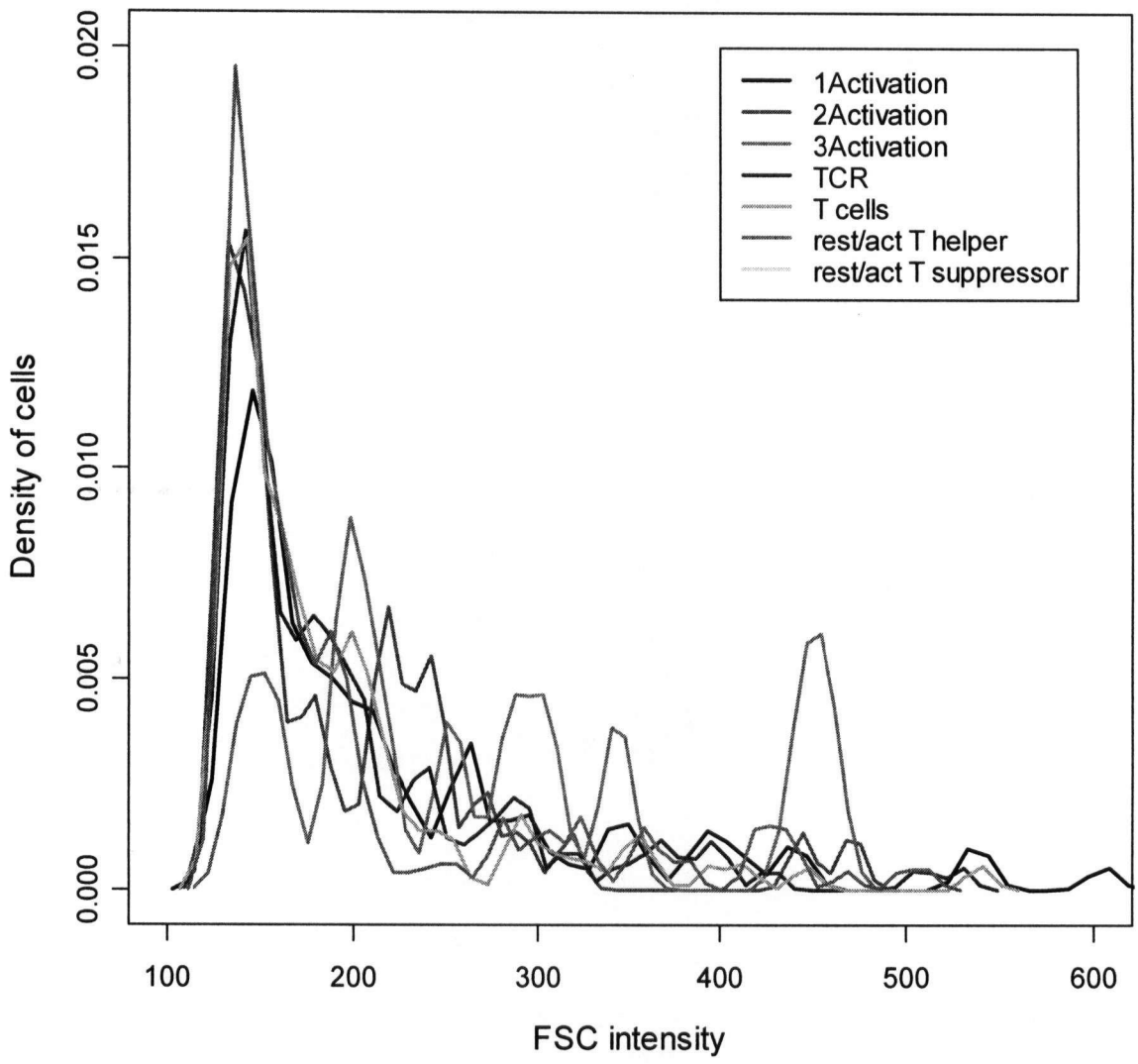


Figure 3.2 Density plot of the FSC intensity using $CD3^+$ cell population from seven aliquots of patient #6's 76 days post-transplant sample. Aliquot '3Activation' was identified as a visual outlier.

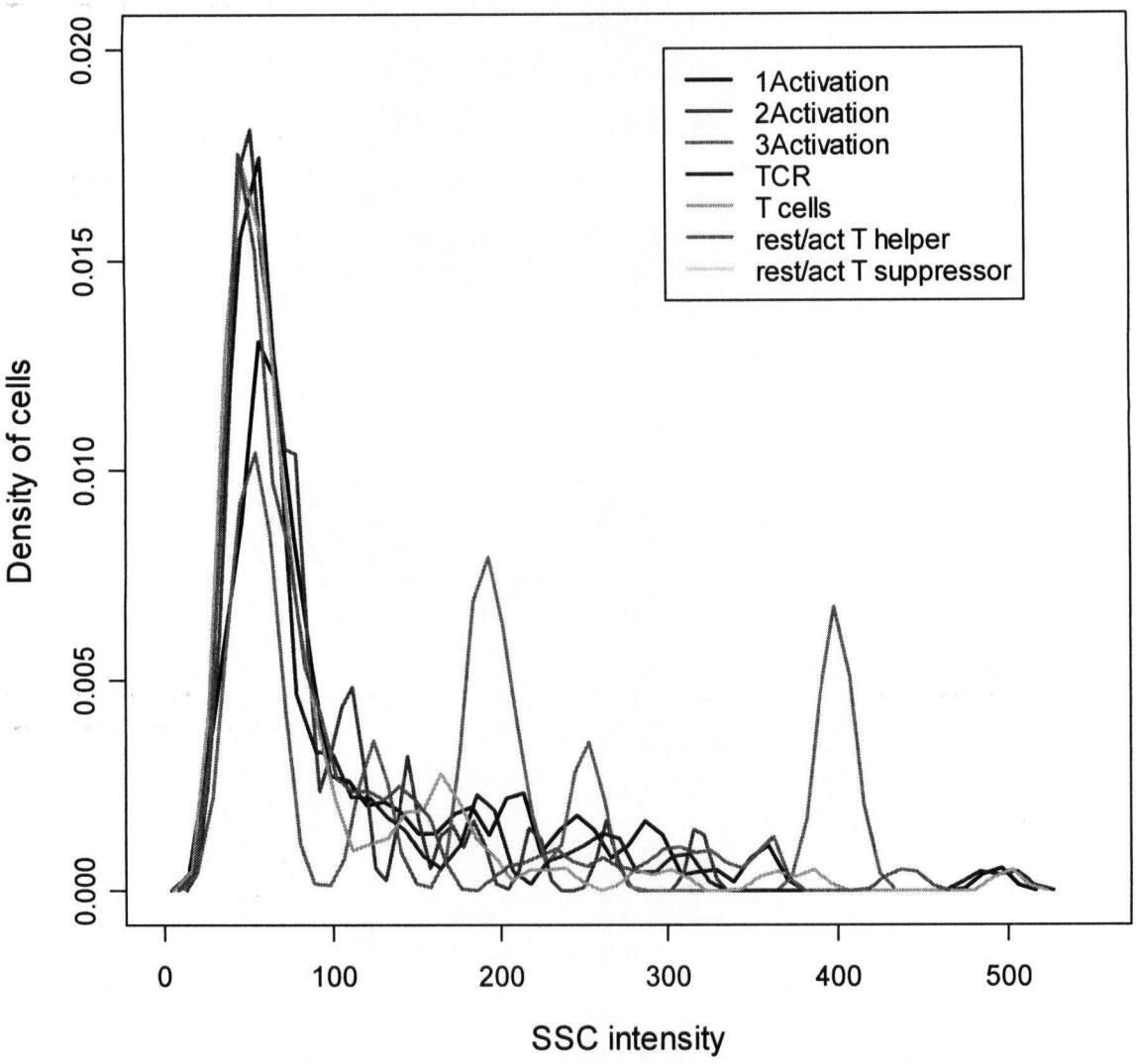


Figure 3.3 Density plot of the SSC intensity using $CD3^+$ cell population from seven aliquots of patient #6's 76 days post-transplant sample. Aliquot '3Activation' was identified as a visual outlier.

Table 3.1 Outliers identified in the QA test on gated data.

Patient #	Cell population	Outlier aliquot	Time point (days post-transplant)
p3	$CD3^-$	2Activation	14
	$CD3^+$	2Activation	14
	$CD3^-$	3Activation	0
p4	$CD3^-$	T cells	81
	$CD3^+$	T cells	81
	$CD3^+$	TCR	32
p6	$CD3^+$	3Activation	76 and 83
p7	$CD3^+$	TCR	35
p9	$CD3^+$	TCR	32
p13	$CD3^+$	TCR	20
p14	$CD3^+$	TCR	21
p17	$CD3^+$	TCR	34, 41, and 55
p18	$CD3^+$	1Activation	-6, 27, 34, and 41
	$CD3^-$	T cells	0
p19	$CD3^-$	T cells	28 and 38
p20	$CD3^+$	TCR	28
p23	$CD3^-$	T cells	28
p25	$CD3^+$	TCR	7 and 21
p31	$CD3^+$	TCR	21, 35, and 70

An example of an outlier and its representation in the density and ECDF plots is shown in Figures 3.4 and 3.5. Among the five available aliquots from patient #4's sample taken at 81 days post-transplant, aliquot 'T cells' exhibited a shift in the intensity while maintaining similar shape. In this case, evidence of this outlier was more prominent in the density plot (Figure 3.4), compared to its corresponding ECDF plot (Figure 3.5).

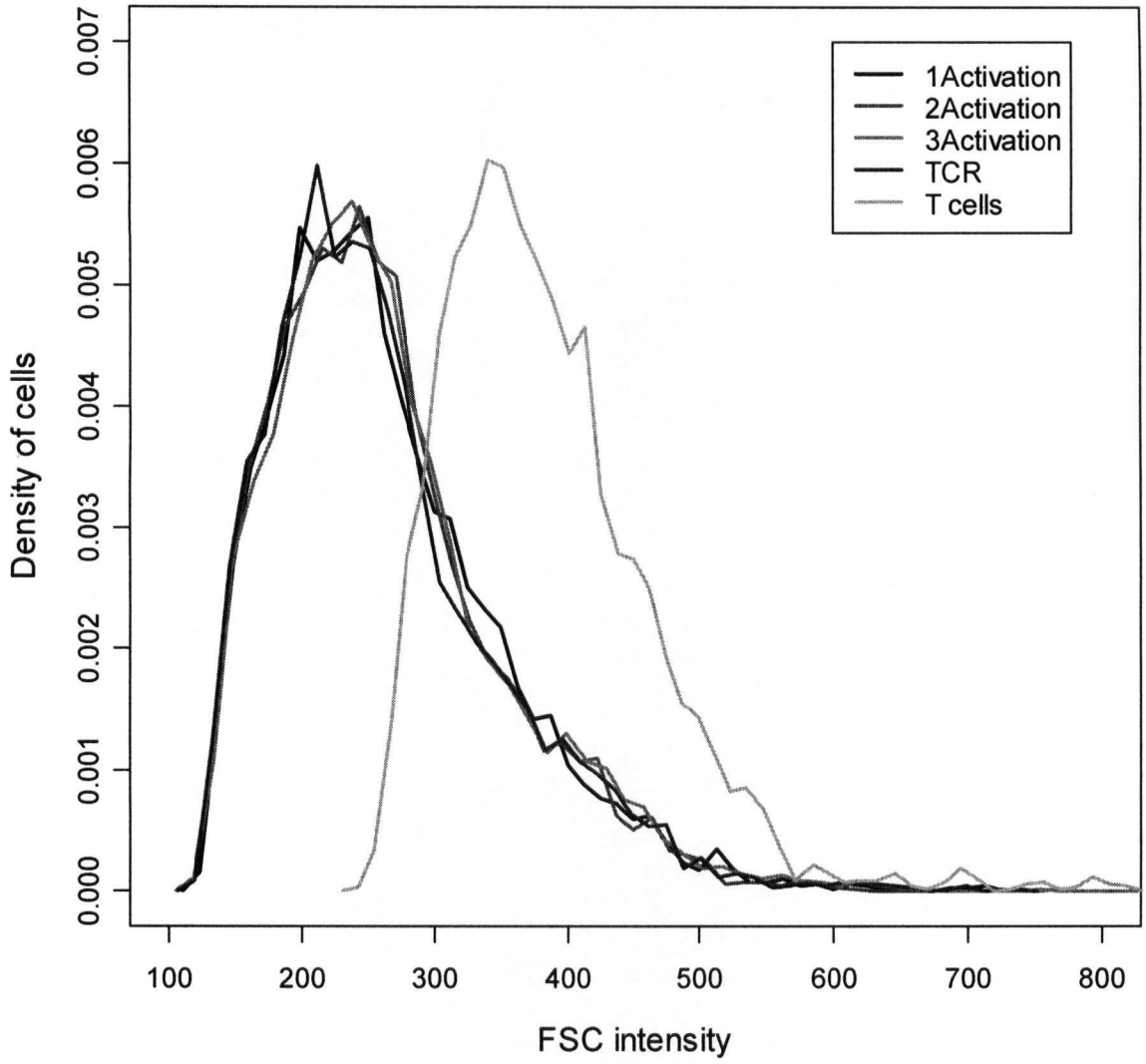


Figure 3.4 Density plot of the FSC intensity using CD3- cell population from five aliquots of patient #4's 81 days post-transplant sample. Aliquot 'T cells' was identified as a visual outlier.

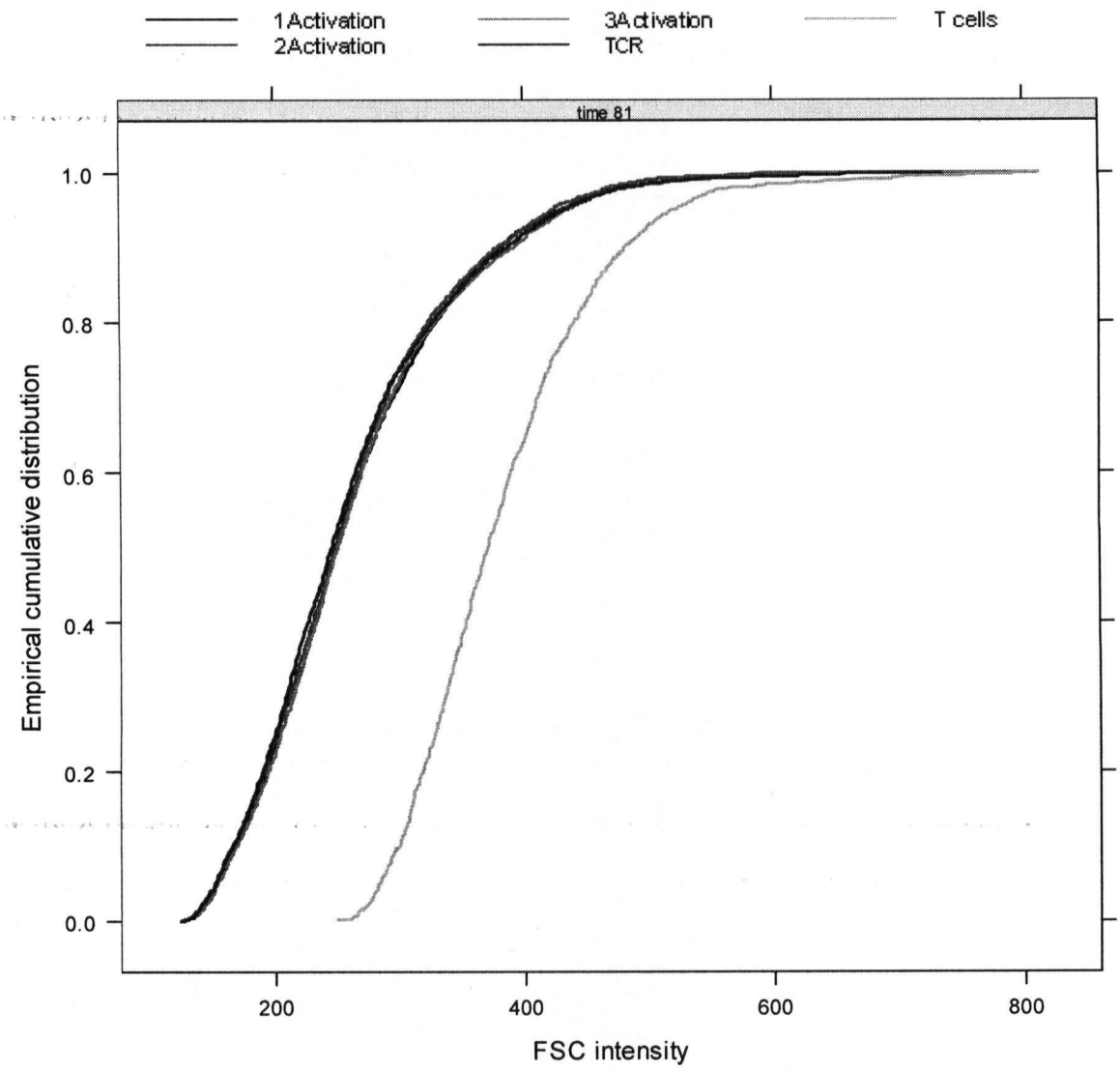


Figure 3.5 ECDF plot of the FSC intensity using CD3- cell population from five aliquots of patient #4's 81 days post-transplant sample. Aliquot 'T cells' was identified as a visual outlier.

3.2.2 Unusually large variations among aliquots

Among all the density plots of FSC and SSC intensities, there were 15 occurrences (Table 3.2) of unusually large variations among the available aliquots. These aliquots (1.4% of the dataset) were removed from the dataset. An example of this trend is shown using the density plot of the $CD3^-$ cell population from patient #28's sample at 14 days post-transplant (Figure 3.6). Although most density plots were mono- or bi-modal and relatively smooth, these 15 samples exhibited rapid polymodal distribution in both FSC and SSC intensity plots. The unusually large variations were also observed in the corresponding ECDF plots; however, the pattern was less apparent without details in the polymodal shape (data not shown). Upon visualization of the FCM data, less live cells were present in some of the aliquots identified (aliquots taken at 53 days post-transplant) with this unusually large variations compared to aliquots from sample taken at different time point (27 days post-transplant) (Figure 3.7).

Table 3.2 Cell populations and samples where $CD3^+$ or $CD3^-$ cell population exhibited unusual variations among the available aliquots.

Patient #	Cell population	Time point (days post-transplant)
p4	$CD3^+$	0
p6	$CD3^+$	46
	$CD3^-$	53
p9	$CD3^-$	6
p10	$CD3^-$	6
p15	$CD3^-$	7
p20	$CD3^-$	7
	$CD3^+$	49, 56, and 63
p26	$CD3^-$	1, 7, and 14
p28	$CD3^-$	14
p29	$CD3^-$	0

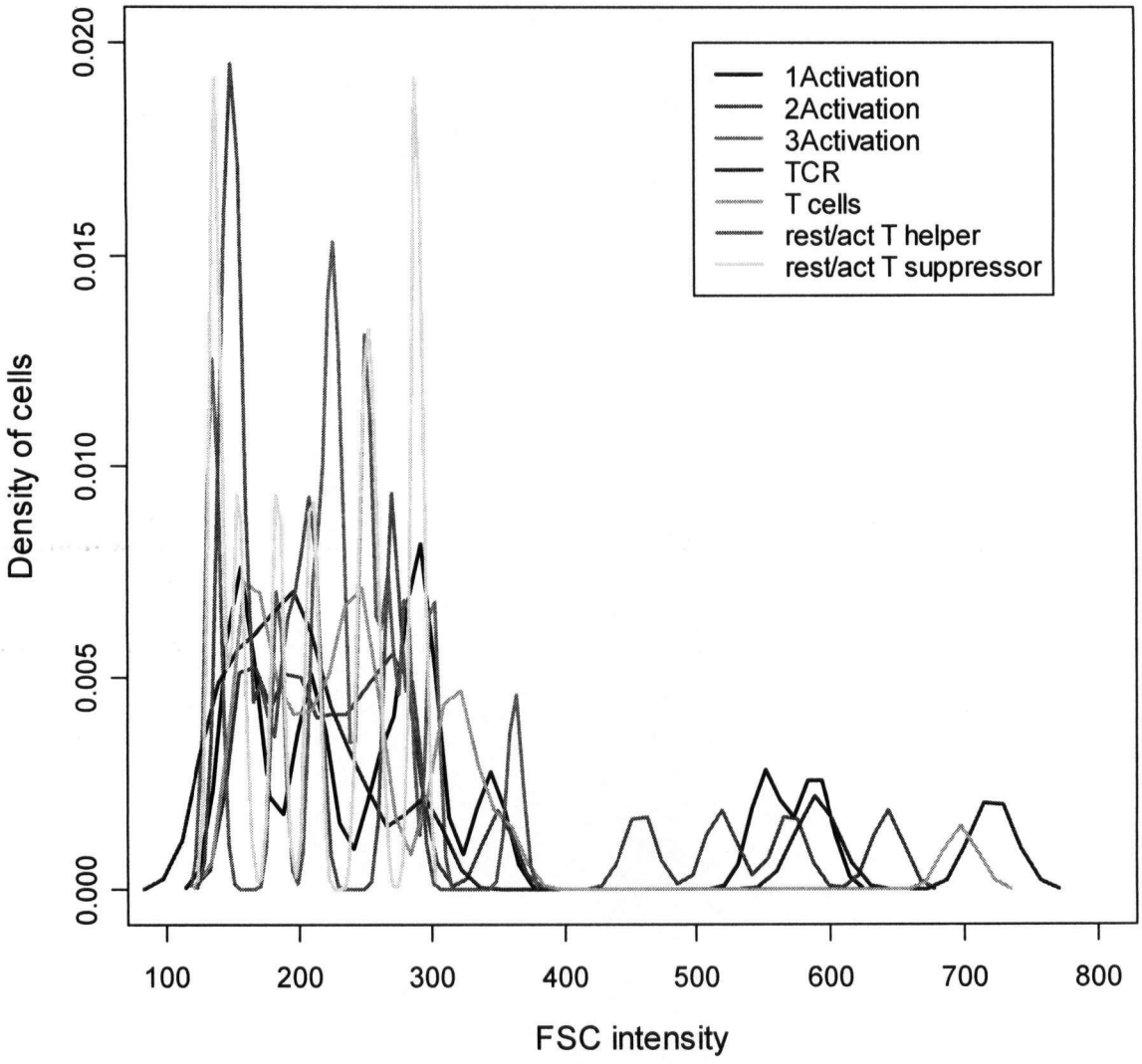


Figure 3.6 Density plot of the FSC intensity using CD3- cell population from seven aliquots of patient #28's 14 days post-transplant sample. All aliquots exhibited great variations from each other. Similar observations also occur in 15 other samples.

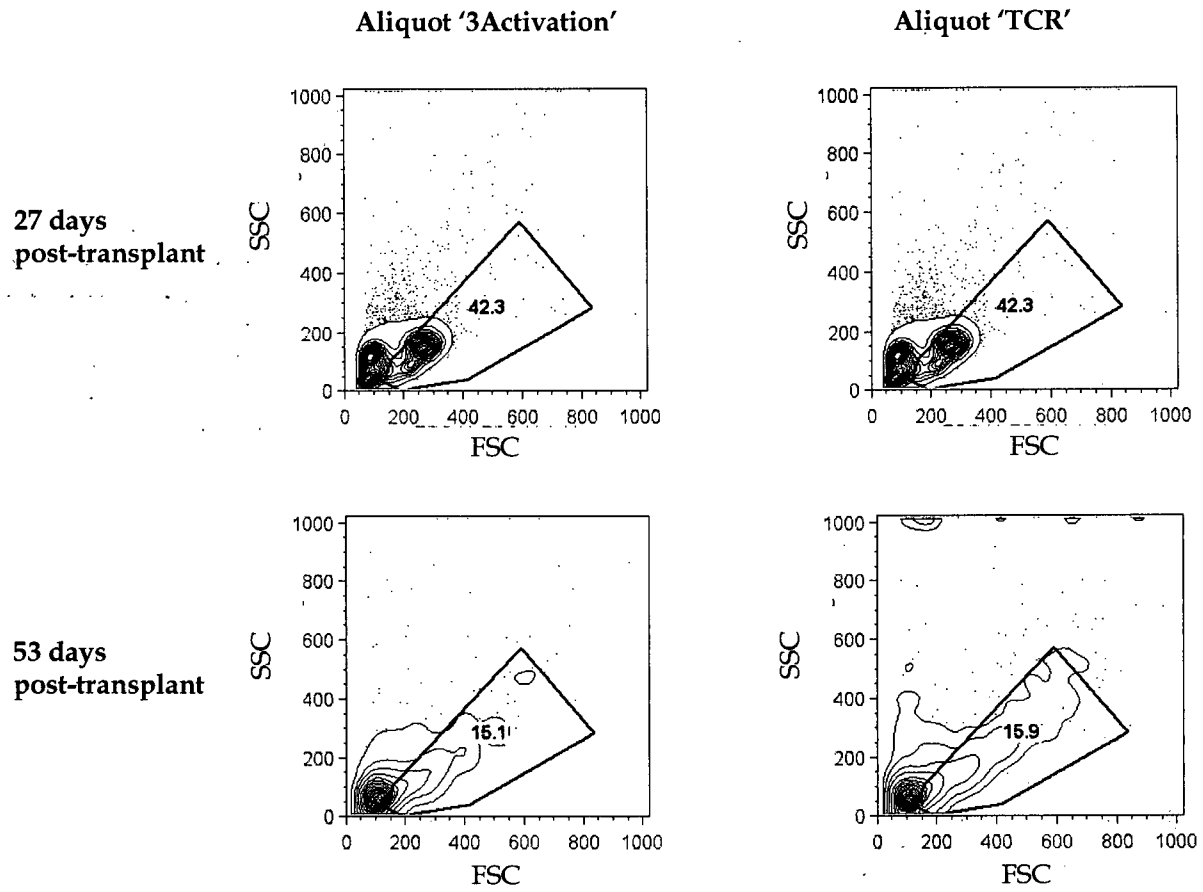


Figure 3.7 FCM contour graphs of FSC vs. SSC from patient #6, aliquots 'TCR' and '3Activation' from samples taken at 27 and 53 days post-transplant.

3.2.3 Repeated outlier conditions

The last unusual pattern I found were repeated 'rest/act T helper' and 'rest/act T suppressor' outliers. There were 33 cell populations where there were two distinct aliquot clusters (Table 3.3). In all cases, the 'rest/act T helper' and the 'rest/act T suppressor' aliquots exhibited similar pattern and formed one cluster whereas all other available aliquots formed another. This trend was most frequent in patients #6's and #7's samples. An example is shown with patient #7's sample taken at the day of BMT. In the $CD3^-$ cell population density plot (Figure 3.8), both shape and intensity were different between the two clusters: i. 'rest/act T helper' and 'rest/act T suppressors'; and ii. '1Activation', '2Activation', '3Activation', 'TCR' and 'T cells'. Relatively small variations were observed within each cluster.

Table 3.3 Cell populations and samples where the two aliquots rest/act T helper and rest/act T suppressor exhibited similar pattern within and different pattern compared to all other available aliquots.

Patient #	Cell population	Time point (days post-transplant)
p6	$CD3^-$	0, 5, 19, 27, 32, 39, 46, 60, and 67
	$CD3^+$	60 and 67
p7	$CD3^-$	-4, 0, 7, 21, 28, 35, 49, 56, 63, 70, and 77
p8	$CD3^-$	19, 33, 42, 49, 54, and 61
p9	$CD3^-$	-6, 55, and 62
p19	$CD3^-$	0
	$CD3^+$	77
p21	$CD3^+$	21

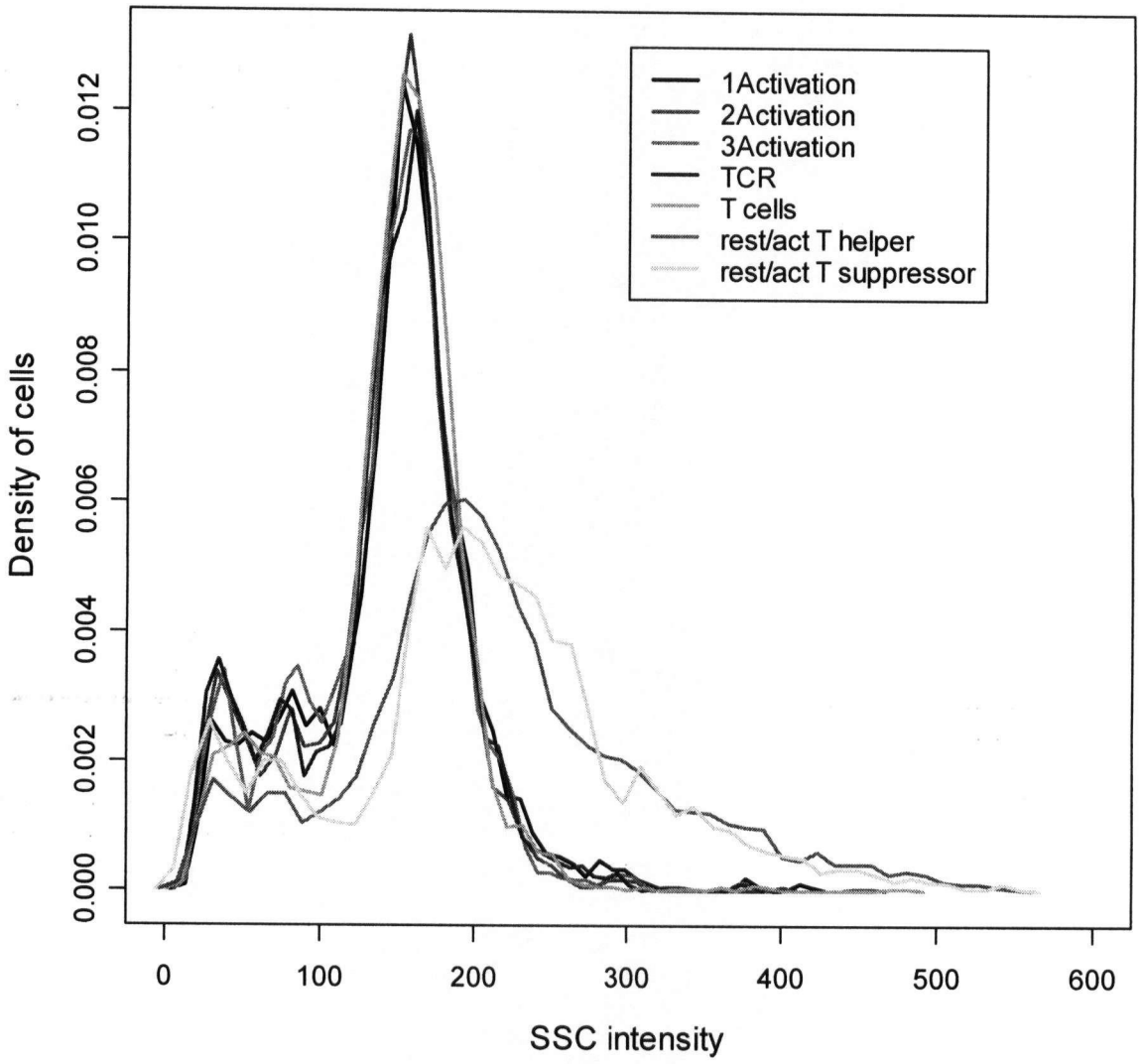


Figure 3.8 Density plot of the SSC intensity using $CD3^-$ cell population from seven aliquots of patient #7's sample taken at the day of BMT. Aliquots 'rest/act T helper' and 'rest/act T suppressor' exhibited different pattern than all other aliquots.

3.2.4 Outlier distributions on the 96-well plate

Distributions of all outliers and unusual patterns on the 96-well plate were investigated. The plating for samples from patient # 6 is shown as an example (Table 3.4). The two outliers, both from aliquot '3Activation', were from sample taken at 76 and 83 days post-transplant and were found to be platted next to each other in column at the left-hand corner of the second plate (Table 3.4). Unusually large variations were observed among all ten aliquots from samples taken at 46 and 53 days post-transplant (Table 3.2). Most of these aliquots were plated in ninth and tenth columns of the first plate and top of seventh and ninth columns of the second plate (Table 3.4). Furthermore, for all but one sample taken between 0 and 67 days post-transplant, two aliquots 'rest/act T helper' and 'rest/act T suppressor' exhibited similar pattern to each other while being completely different to other aliquots (Table 3.3). These aliquots were plated on different plate - the two rest/act aliquots were plated on the second plate while most of the other aliquots were plated on the first plate (Table 3.4).

There were many outliers observed from aliquots close or next to each other in column as there were a trend of cluster of time points when the same aliquot were identified as outliers at multiple time points (Table 3.1). Among the 29 outliers, 20 were aliquots 'TCR' or 'T cells' and 13 of which were identified from samples taken between 20 and 40 days post-transplant (Table 3.1). In many cases, these outliers were mapped to aliquots plated in the middle of a plate (Table 3.4). Similar to patient # 6's, many of the rest/act aliquots differences were observed when these aliquots were plated in a separate plate from most of the other aliquots. These trends could generally be observed from other patients' samples (data not shown).

Table 3.4 Plating order for patient #6 with samples taken at multiple time points on two plates. Aliquots identified as outliers and unusually variations are labelled with shaded areas.

Plate #1

Plate Rows	Aliquots	1	2	3	4	5	6	7	8	9	10	11	12
A	Myeloids	-8	0	5	12	19	27	32	39	46	53	60	67
B	T cells	-8	0	5	12	19	27	32	39	46	53	60	67
C	NK cells	-8	0	5	12	19	27	32	39	46	53	60	67
D	B cells	-8	0	5	12	19	27	32	39	46	53	60	67
E	TCR	-8	0	5	12	19	27	32	39	46	53	60	67
F	1Act Marker	-8	0	5	12	19	27	32	39	46	53	60	67
G	2Act Marker	-8	0	5	12	19	27	32	39	46	53	60	67
H	3Act Marker	-8	0	5	12	19	27	32	39	46	53	60	67

Plate #2

Plate Rows	Aliquots	1	2	3	4	5	6	7	8	9	10	11	12
A	Myeloids	76	83	90	176		-8	46	-8	46			
B	T cells	76	83	90	176		0	53	0	53			
C	NK cells	76	83	90	176		5	60	5	60			
D	B cells	76	83	90	176		12	67	12	67			
E	TCR	76	83	90	176		19	76	19	76			
F	1Act Marker	76	83	90	176		27	83	27	83			
G	2Act Marker	76	83	90	176		32	90	32	90			
H	3Act Marker	76	83	90	176		39	176	39	176			
							Helper		Suppressor				

3.3 B-spline parameters

The effects of the B-spline basis order and knot placements were evaluated using data from patient #2 with a missing observation at week one. First, B-splines were built with one knot at every sampled time point and three different basis orders (Figure 3.9). Although all three B-splines followed the general patterns exhibited by the raw data (red dots) by visual inspection, the B-spline with basis order two best reflected the raw data. Even though no knot was placed at week one, fitting a B-spline with basis order three imposed quadratic function between the two knots at week zero and weeks two. As a result, there was a discrepancy between the B-spline with basis order of three and the raw data pattern. A similar discrepancy was also observed between the raw data and B-spline fitted with basis order four, most evidently between five and six weeks post-transplant (Figure 3.9).

Secondly, B-splines were built with linear basis order and four different knot placements with decreasing knots interval. The B-splines becomes smoother and further away from the actual raw data pattern as the knot frequency decreased (Figure 3.10). Another noticeable feature was the behaviour of each spline at week one where no observed value was available. A knot at week one resulted in an imputed B-spline pattern at either side of the knot based on the trends of the previous basis function. As a result, the imputation created discrepancy from the raw data pattern (Figure 3.10).

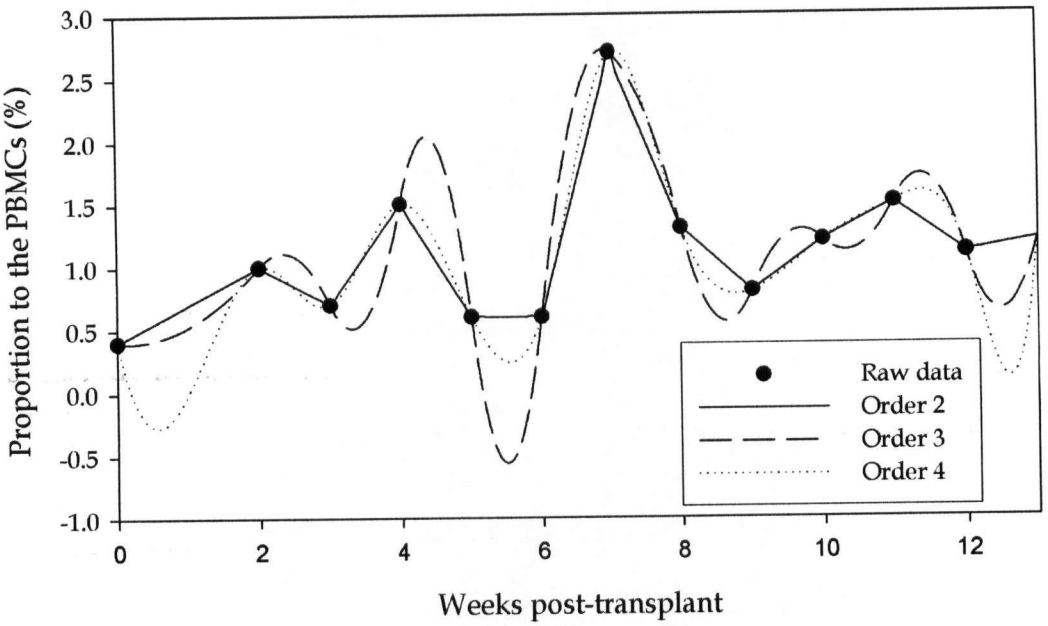


Figure 3.9 B-splines with knots located at every available time point and orders two, three or four fitting into the raw data.

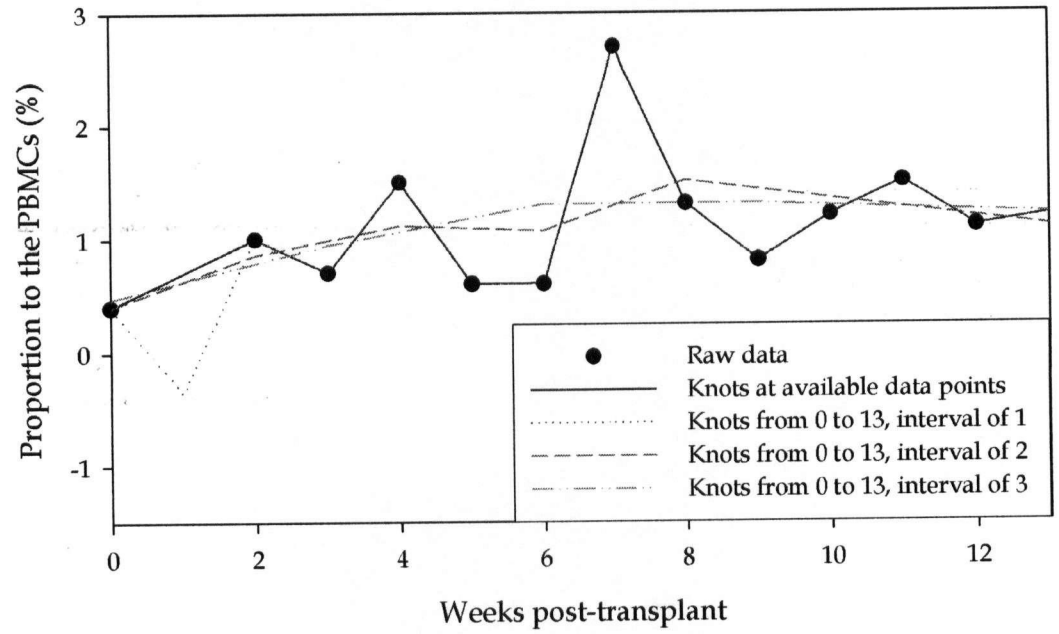


Figure 3.10 B-spline with order two and different distribution of knots fitting into the raw data.

CHAPTER 4 RESULTS - TOP RANKING CLASSIFIERS

In order to identify patterns of immune cell abundances that correlate to the onset of aGvHD and cGvHD, the temporal analysis pipeline was performed on qualified subsets of immune cells comparing between samples taken from the aGvHD and the non-GvHD patients, and between samples taken from seven aGvHD & cGvHD and nine aGvHD only patients respectively. Top ranking classifiers with potential discriminative patterns predicting the onset of aGvHD and cGvHD are described in sections 4.1 and 4.2.

4.1 Classifiers for the onset of acute graft versus host disease

Patient #17, a non-GvHD patient, was omitted from the FLDA analysis due to lack of available data within the selected time ranges. However, these data, if available, were included in the raw data time plots. Only top ranking classifiers from the proportion dataset using samples taken between 7 and 21 days post-transplant are described below (Table 4.1). All others are described in Appendix G. The complete validation results for all subsets of immune cells in each time range are listed in Tables H.1 - H.3 for the proportion dataset and Tables H.4 - H.6 for the concentration dataset. The time range after BMT (7 to 21 days post-transplant) was selected to exclude the day of BMT and 21 to 28 days post-transplant when the aGvHD diagnosis rate rapidly increased (Figure 4.1) so the top classifiers may be used for aGvHD prediction.

Table 4.1 Validation results for the top ranking subsets of immune cells and their related cell populations from the FLDA classification with different subsets of aGvHD vs. the non-GvHD patients using samples taken between 7 and 21 days post-transplant. (nd = not done due to lack of data).

Immune cells	Aliquot	aGvHD		Grade II-IV aGvHD		Grade III-IV aGvHD	
		Sensitivity	Specificity	Sensitivity	Specificity	Sensitivity	Specificity
$CD2^{dim}$ $CD16^{+}$ $CD56^{+}$ $CD3^{-}$	NK cells	90%	100%	82%	100%	92%	67%
$CD3^{+}CD4^{+}CD8\beta^{+}$	T cells	86%	100%	82%	100%	92%	100%
$CD3^{+}$CD4int	2Activation	81%	100%	82%	100%	83%	100%
$CD3^{+}CD4^{+}CD8\beta^{+}CD8^{+}$	T cells	71%	100%	76%	100%	83%	100%
$CD3^{+}$	1Activation	90%	33%	94%	33%	92%	33%
$CD3^{+}$	2Activation	86%	33%	94%	33%	92%	33%
$CD3^{+}CD4^{+}$	rest/act T helper	nd	nd	nd	nd	nd	nd
$CD3^{+}CD8\beta^{dim}CD8^{-}$	T cells	90%	0%	82%	67%	83%	67%
$CD3^{+}CD8\beta^{+}CD4^{-}$	T cells	81%	33%	76%	33%	75%	33%
$CD3^{+}CD8^{+}CD8\beta^{-}$	T cells	81%	33%	76%	33%	83%	33%
$CD3^{+}CD4^{+}CD8\beta^{-}$	T cells	90%	33%	100%	33%	100%	0%
$CD3^{+}CD8\beta^{+}CD8^{+}$	T cells	81%	33%	76%	33%	75%	33%

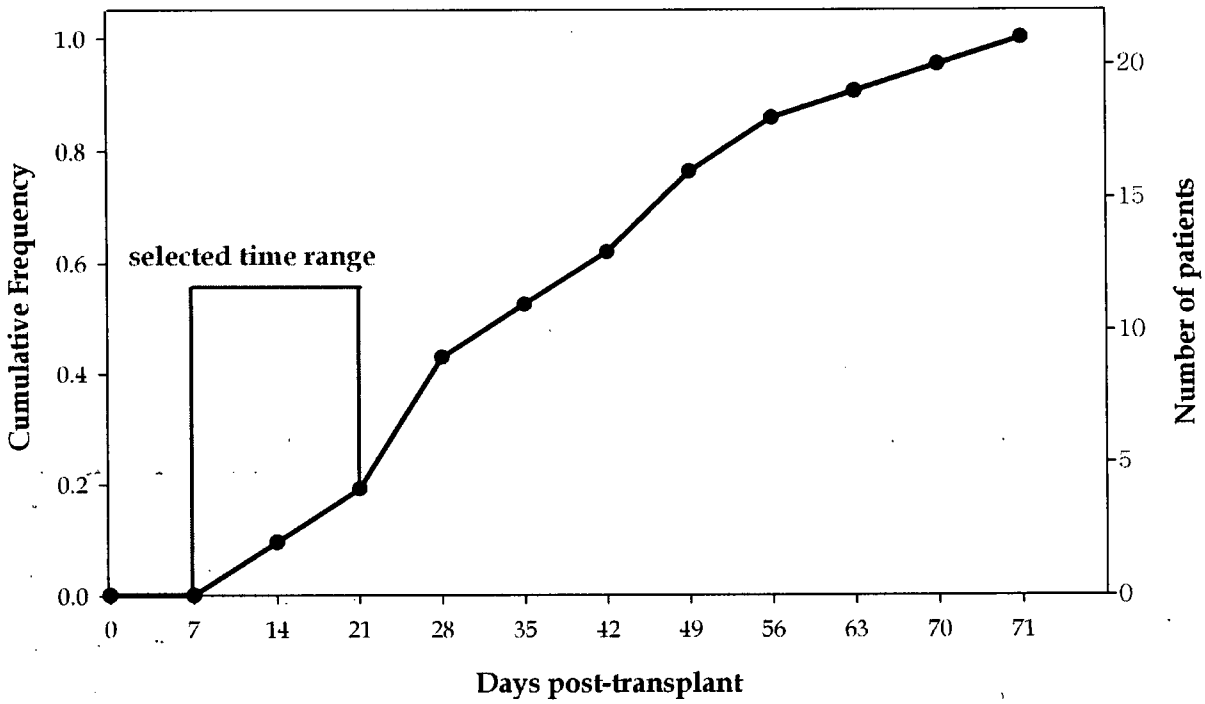


Figure 4.1 Cumulative distribution of the aGvHD diagnosis days post-transplant with the selected time range between 7 and 21 days post-transplant labelled.

4.1.1 Inconsistent classifier by missing values

The FLDA classifier built on the immune cells $CD2^{dim}CD16^{+}CD56^{+}CD3^{-}$ was estimated to have the highest sensitivity and specificity (Table 4.1). The FLDA estimated signals exhibited a very clear separation between the aGvHD and the non-GvHD patients at seven days post-transplant (Figure 4.2a). However, the separation around seven days post-transplant between the aGvHD and the non-GvHD patients was not observed in the raw data time plot of the same time range because there were no data available from the non-GvHD patients between seven and ten days post-transplant (Figure 4.2b). In the extended raw data time plot, the proportion values from two out of three non-GvHD patients were as high as the values from most aGvHD patients (Figure 4.3). Unlike all other top ranking classifiers described below, this subset of immune cells did not display a consistent pattern in its extended raw data time plot.

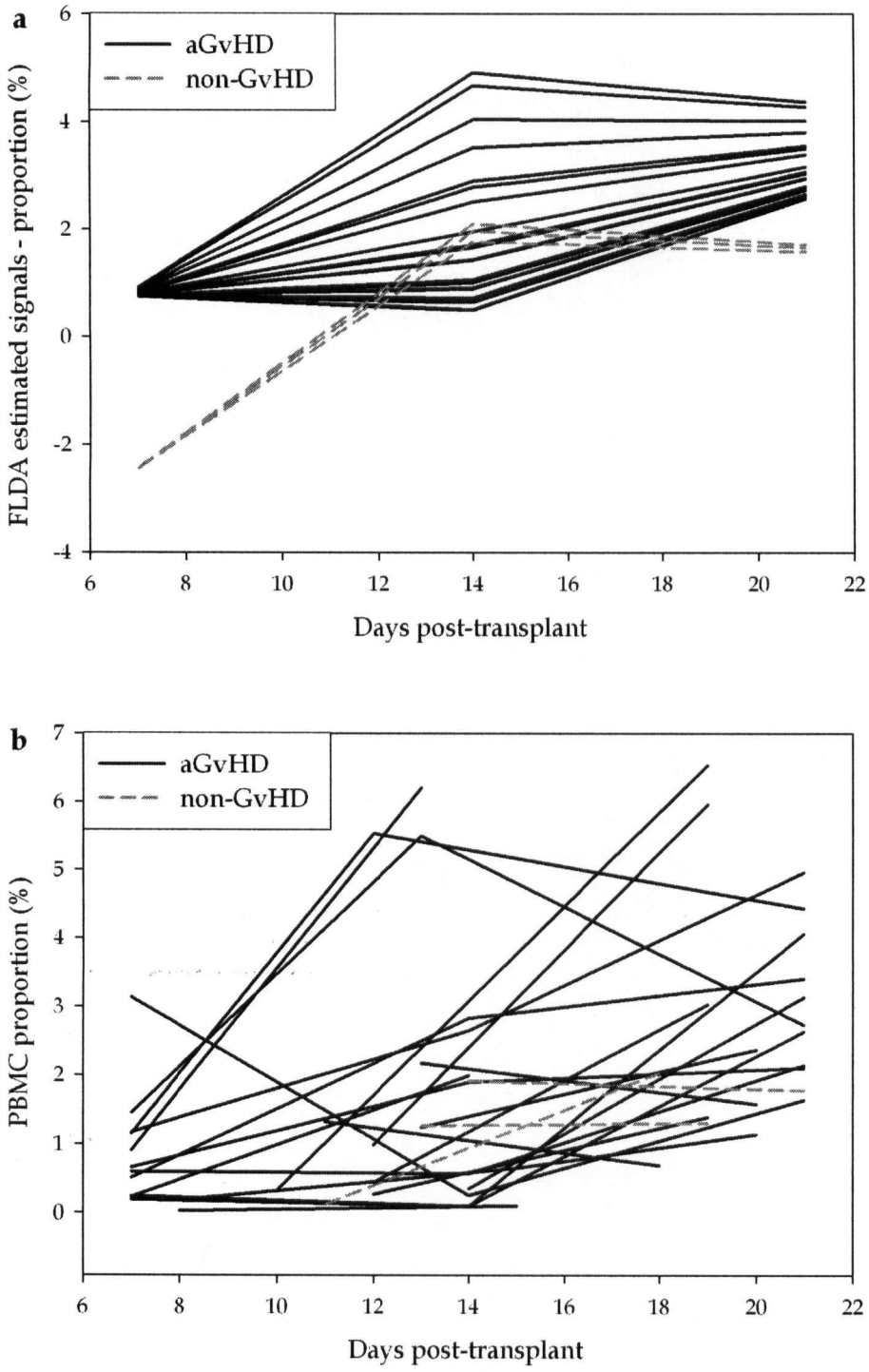


Figure 4.2 Time plots of the FLDA estimated signals (panel a) and the raw data (panel b) based on samples taken between 7 and 21 days post-transplant for the immune cells $CD2^{dim}CD16^{+}CD56^{+}CD3^{-}$ in proportion to PBMC.

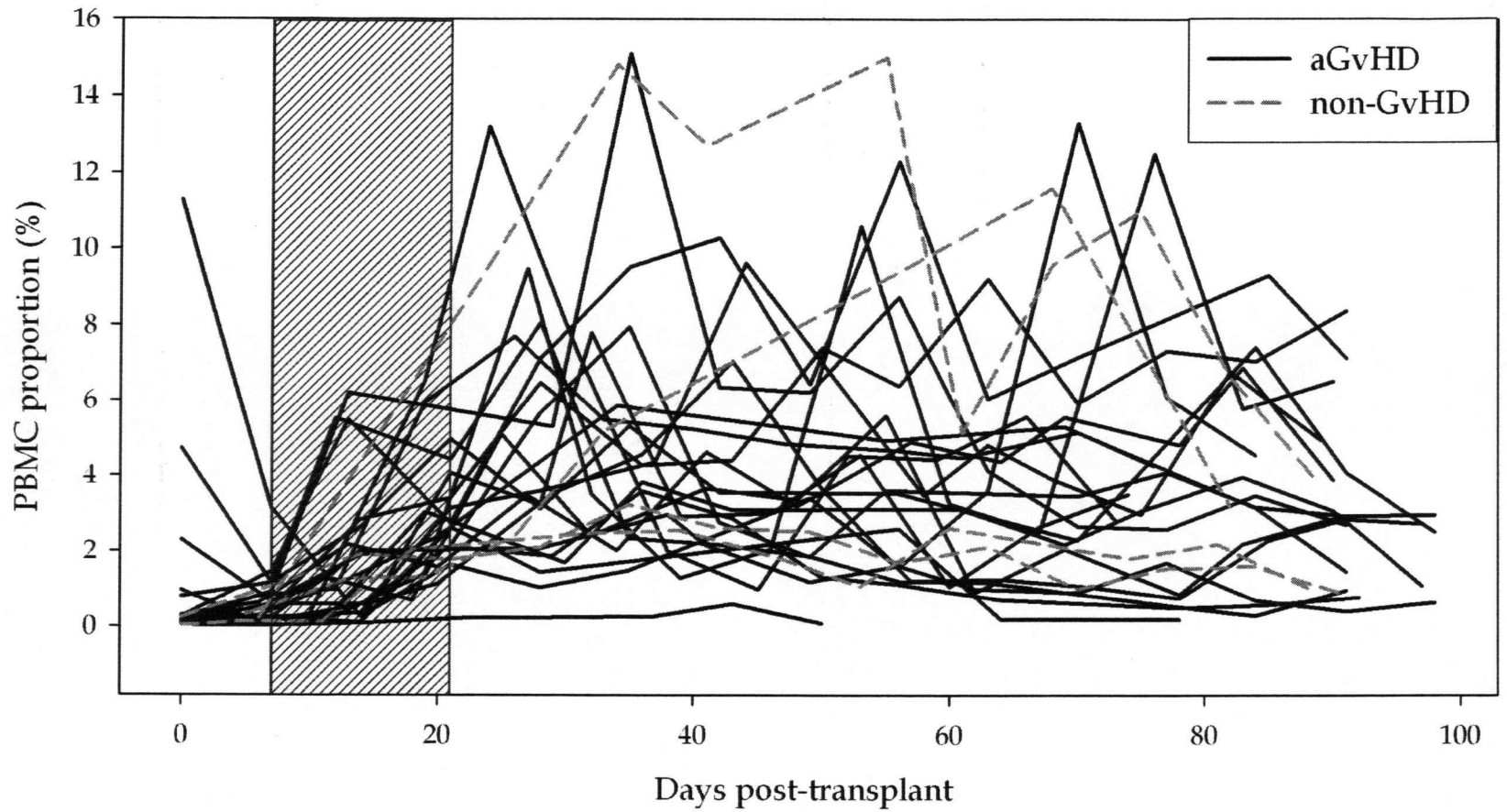


Figure 4.3 Raw data time plot for immune cells $CD2^{dim}CD16^{+}CD56^{+}CD3^{-}$ in proportion to PBMC based on samples taken between 0 and 100 days post-transplant. The purpled striped box indicates the time range where data was analyzed via FLDA.

4.1.2 $CD3^+CD4^+CD8\beta^+(CD8^+)$

The FLDA classifier built from the immune cells $CD3^+CD4^+CD8\beta^+$ was identified as one of the top ranking classifiers with the estimated sensitivity and specificity both higher than 70% in two time ranges: 7 to 21 days post-transplant (Table 4.1) and 21 and 0 days prior to aGvHD diagnosis (Table H.2). Estimated sensitivity and specificity increased in the supplementary comparisons between moderate or severe aGvHD and non-GvHD patients (Table 4.1). The FLDA estimated signals time plot (Figure 4.4) displayed a pattern of higher PBMC proportion values from the aGvHD patients, compared to values from the non-GvHD patients. A similar pattern was also observed in the FCM data in contour graphs between CD4 and CD8β intensities (Figure 4.5).

In the extended raw data time plot (Figure 4.6), all but one aGvHD patient had higher values and greater fluctuation, compared to the non-GvHD patients, within the time range from 0 to 120 days post-transplant. Patient #25, who was diagnosed with grade I aGvHD at 44 days post-transplant, had $CD3^+CD4^+CD8\beta^+$ proportion values lower than 0.5% from 0 to 50 days post-transplant. There were two sudden increases in the $CD3^+CD4^+CD8\beta^+$ proportion for patient #6's samples taken at 53 and 90 days post-transplant (Figure 4.6). They were the results of minimal amounts of viable cells in the aliquots (data not shown). Similar incidences were observed in the immune cells $CD3^+CD4^{int}$ described in section 4.1.3.

A new subpopulation was gated within the immune cells $CD3^+CD4^+CD8\beta^+$ to obtain abundance readings for a new immune cell population - $CD3^+CD4^+CD8\beta^+CD8^+$ (Figure 4.7). The FLDA classifier from this new subset of immune cells had an estimated 71% sensitivity and 100% specificity (Table 4.1), and displayed a similar pattern in both the raw data and FLDA signal time plots (Figure 4.8) to its parent population. All other related immune cell populations that were positive in only one of the CD4 or CD8/CD8β markers had a lower estimated

sensitivity and specificity (Table 4.1) and did not exhibit discriminative pattern between the two patient groups (Figure 4.12).

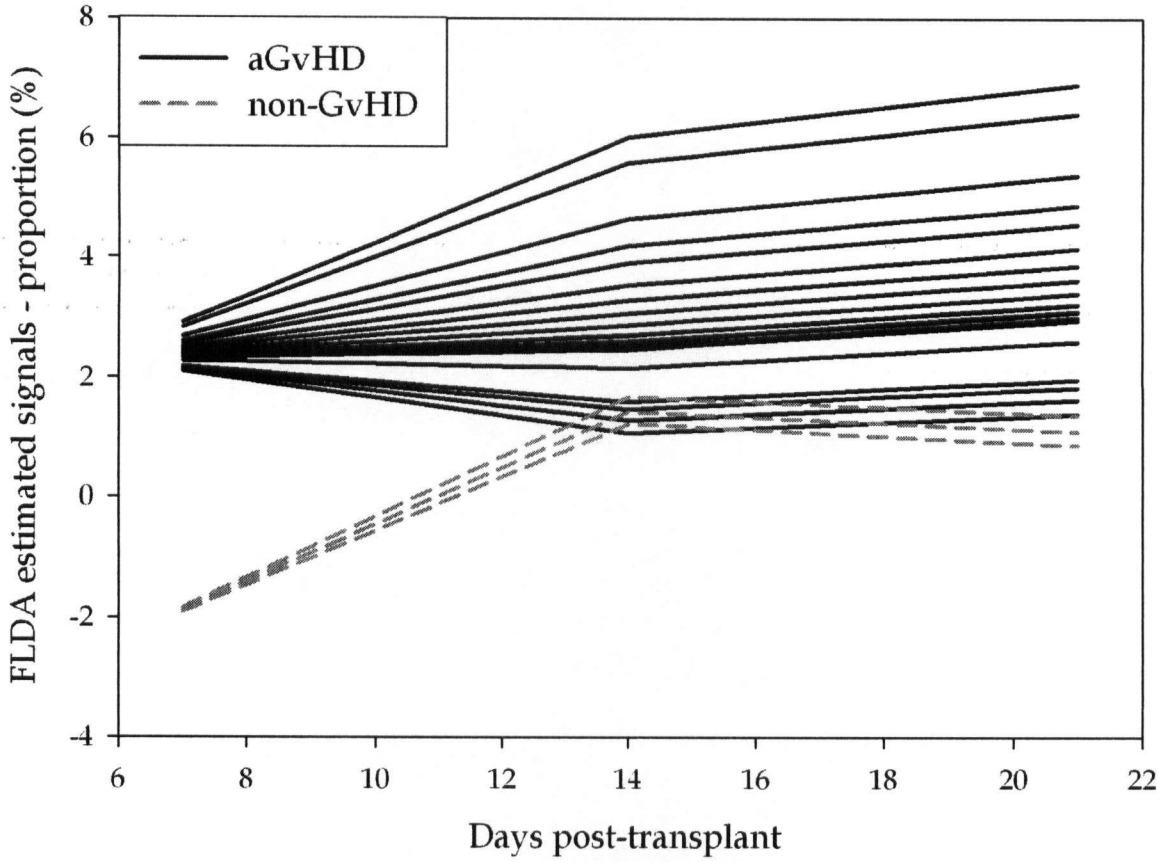


Figure 4.4 FLDA estimated signals time plot based on samples taken between 7 and 21 days post-transplant for immune cells $CD3^+CD4^+CD8\beta^+$ in proportion to PBMC.

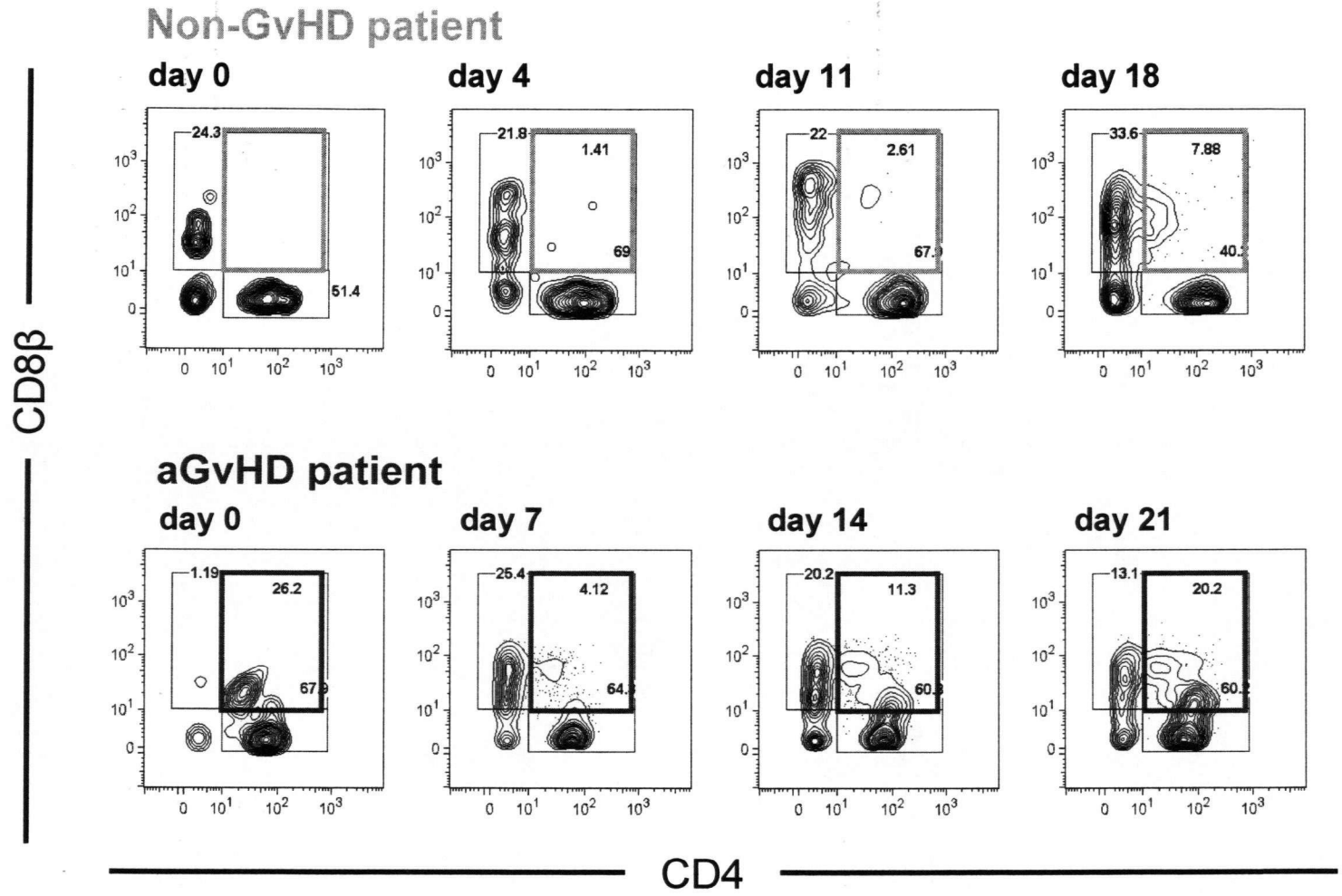


Figure 4.5 FCM contour graphs of transformed CD4 and CD8β marker measurements for a non-GvHD patient (#4) and aGvHD patients (#27) between zero and three weeks post-transplant. The $CD3^+CD4^+CD8\beta^+$ population is gated within the double positive gate.

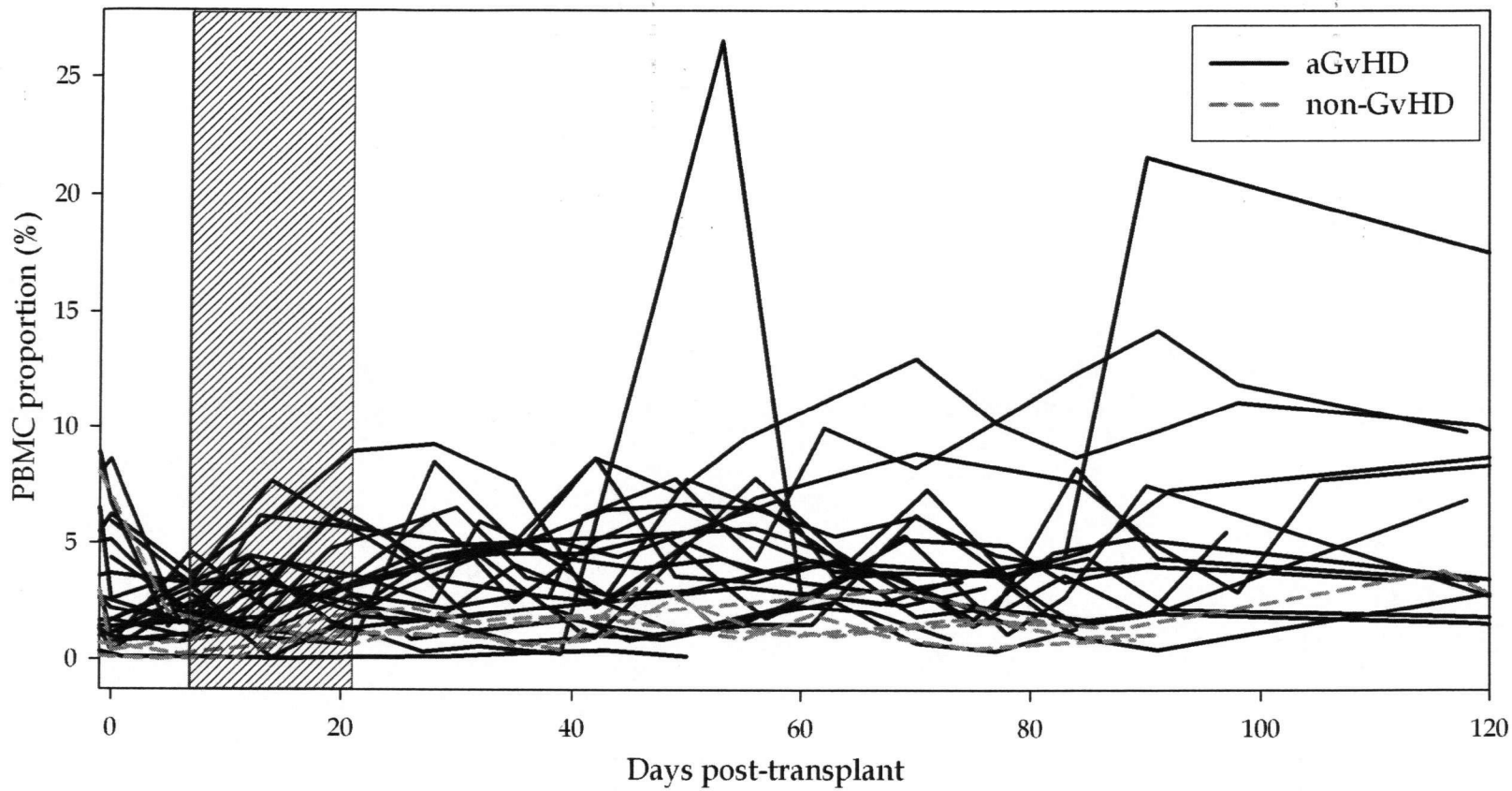


Figure 4.6 Raw data time plot for immune cells $CD3^+CD4^+CD8\beta^+$ in proportion to PBMC, based on samples taken between 0 and 120 days post-transplant. The purpled striped box indicates the time range where data was analyzed via FLDA.

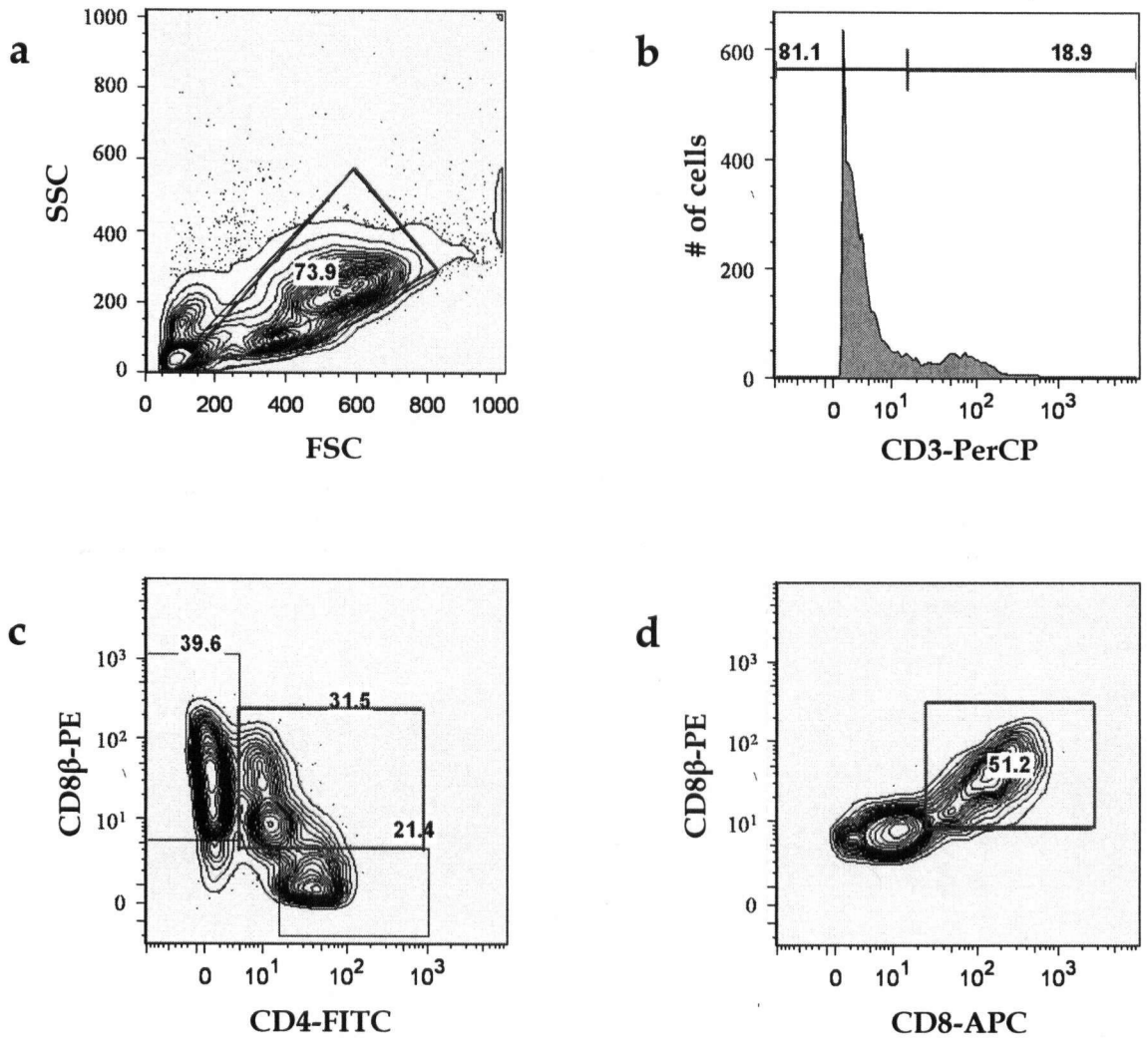


Figure 4.7 An example of sequential gating of the existing cell population $CD3^+CD4^+CD8\beta^+$ (red gates, panels a, b, and c) to identify a new immune cell population $CD3^+CD4^+CD8\beta^+CD8^+$ (panel d).

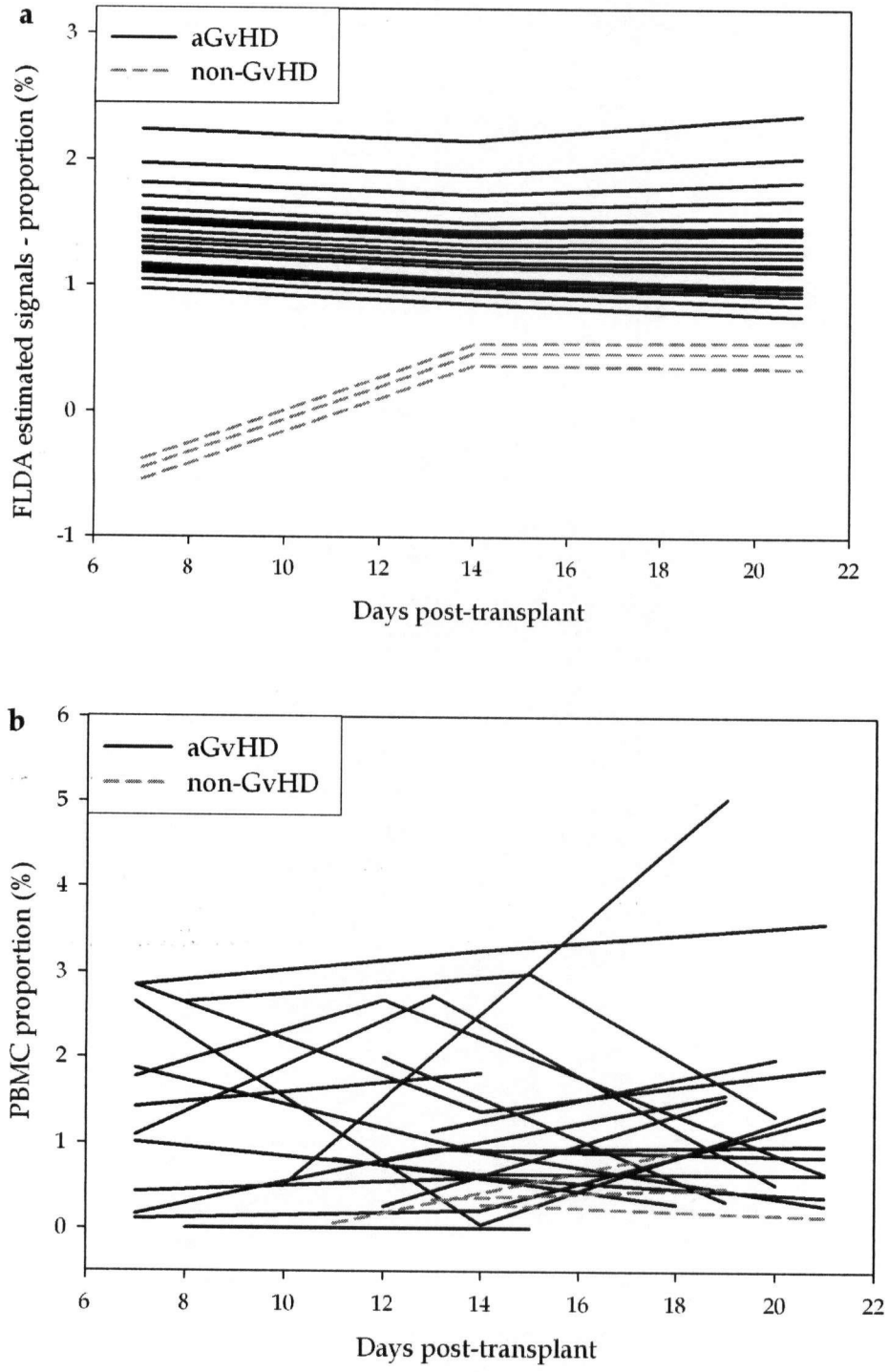


Figure 4.8 Time plots of the FLDA estimated signals (panel a) and the raw data (panel b) based on samples taken between 7 and 21 days post-transplant for the new immune cell population $CD3^+CD4^+CD8\beta^+CD8^+$ in proportion to PBMC.

4.1.3 $CD3^+CD4^{int}$

The FLDA classifier built using the immune cells $CD3^+CD4^{int}$ (aliquot '2Activation') had an estimated 71% sensitivity and 100% specificity (Table 4.1). The time plots of FLDA estimated signals (Figure 4.9a) and raw data (purple stripped area, Figure 4.9b) exhibited similar patterns to that of the immune cells $CD3^+CD4^+CD8\beta^+$ (Figure 4.4). In the FLDA estimated signals time plot, the aGvHD patients had higher proportion values of this subset of immune cells, compared to the non-GvHD patients, and the main separations were found around 7 and 14 days post-transplant. This pattern persisted in the raw data time plot from 0 to 100 days post-transplant (Figure 4.9b).

There were also two peaks from patient #6's samples at 53 and 90 days. At 39 days post-transplant, the proportion value was 1%. It increased to 26% at 53 days post-transplant and returned to 2.6% at 60 days post-transplant. After a relatively flat pattern between 60 and 83 days post-transplant, the value increased again to 25% at 90 days post-transplant. Similar peaks from patient #6 were also observed in the immune cells $CD3^+CD4^+CD8\beta^+$ (Figure 4.6). The corresponding FCM data from samples taken around the aforementioned time points were examined. Aliquot 'T cells' from samples taken at 53 and 90 days post-transplant exhibited very different pattern with less live cells within the gate in both the FSC-SSC scatter plot and CD3-PerCP histogram, when compared to sample taken before (46 days post-transplant) or after (60 days post-transplant) the sudden peaks (Figure 4.10).

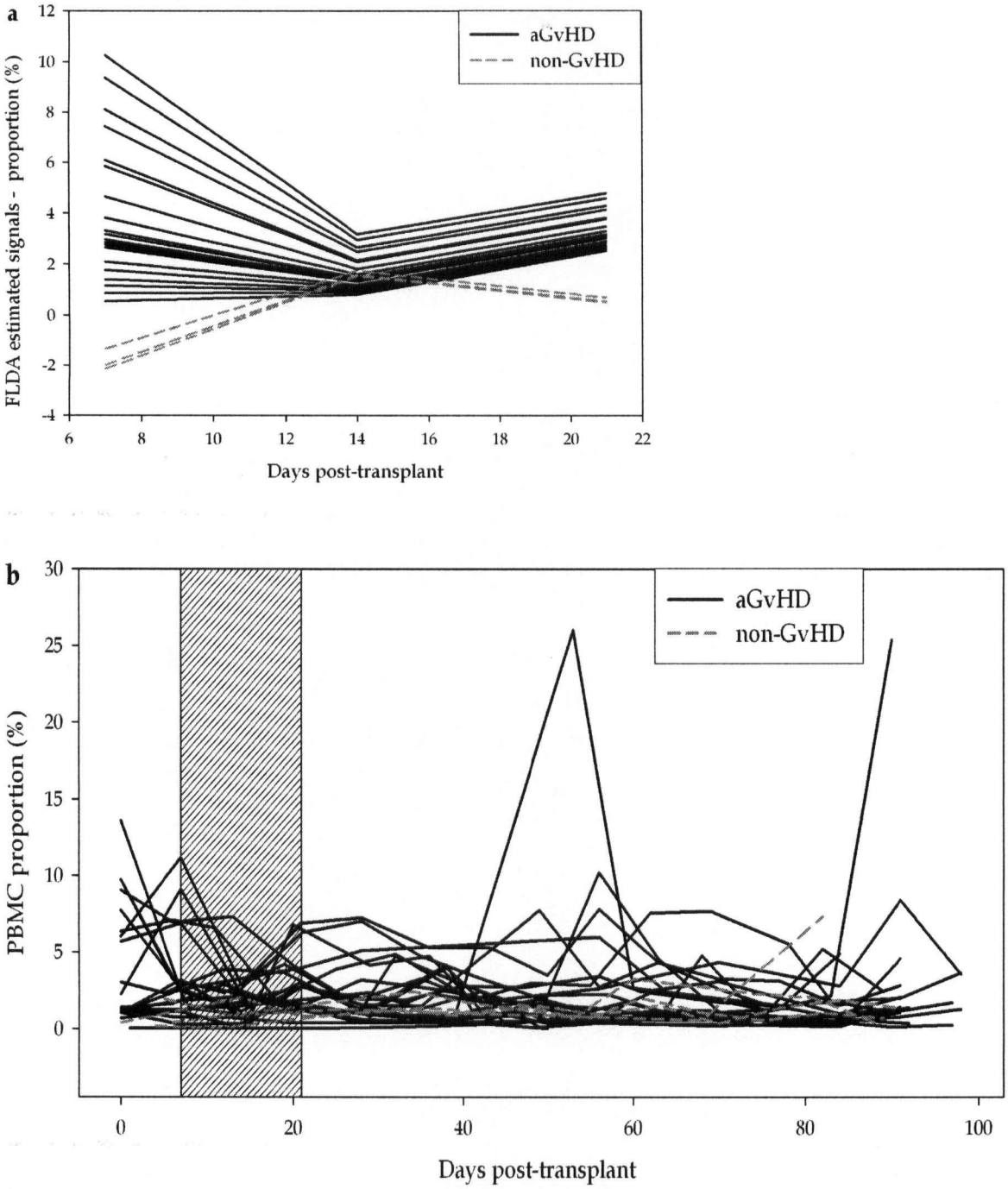


Figure 4.9 Time plot of the FLDA estimated signals (panel a) based on samples taken between 7 and 21 days post-transplant and time plot of the raw data (panel b) based on samples taken between 0 and 100 days post-transplant for the immune cells $CD3^{+}CD4^{int}$ in proportion to PBMC (aliquot '2Activation'). The purpled striped box indicates the time range where data was analyzed via FLDA.

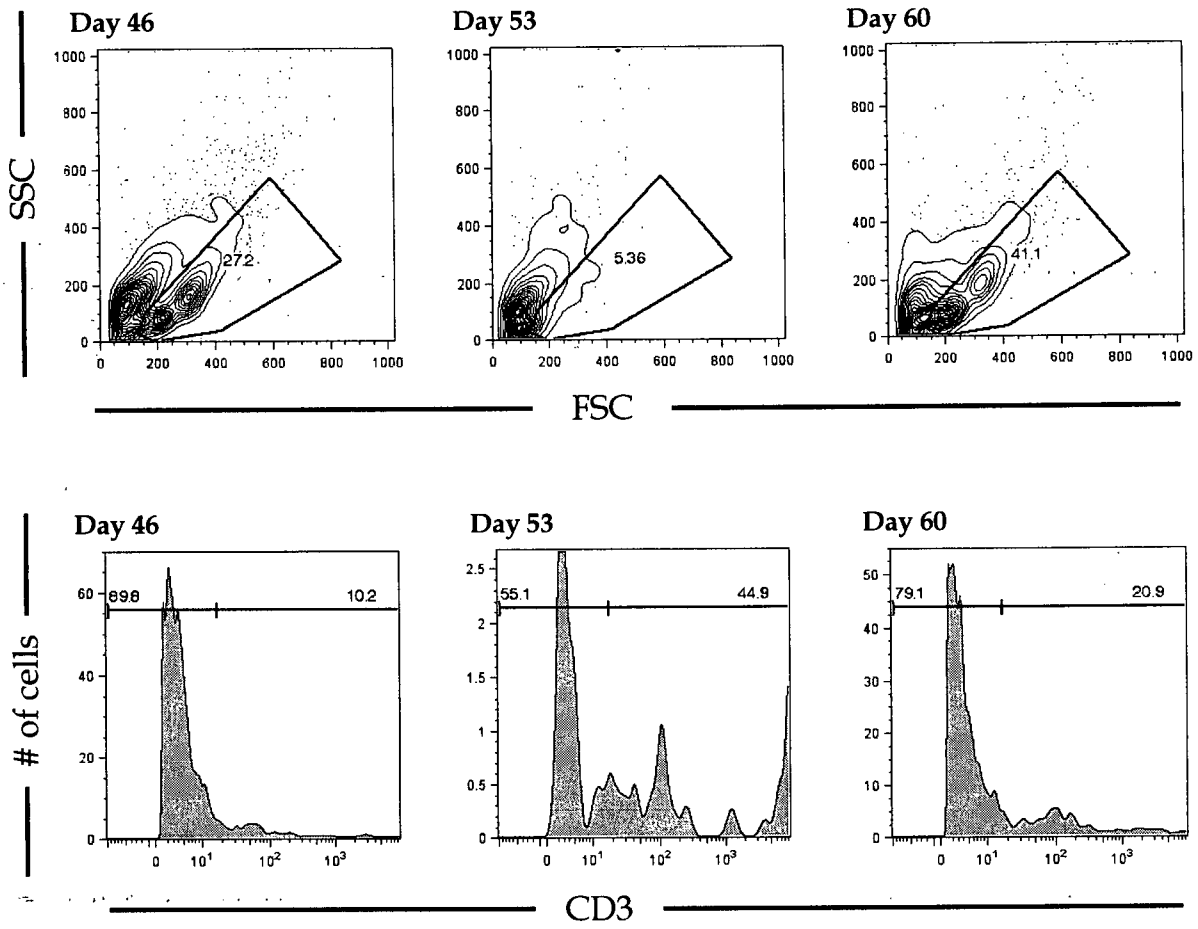


Figure 4.10 FCM data in scatter plot of FSC vs. SSC and histogram of CD3-PerCP intensity from patient #6, aliquot 'T cells' from samples taken at 45, 53, and 60 days post-transplant.

It was also noted that two other subsets of immune cells: $CD3^+$ and $CD3^+CD4^+$, representing immune cell populations closely related to cells with the phenotype $CD3^+CD4^{int}$ did not exhibit discriminative patterns between the aGvHD and the non-GvHD patients. Multiple readings from the $CD3^+$ immune cell population all had approximately 86% sensitivity but only 33% specificity (Table 4.1). In the time plot of $CD3^+$ immune cell population (Figure 4.11), the proportion values were high from both the aGvHD and non-GvHD patients. The subset of immune cells $CD3^+CD4^{int}$ was not analyzed via FLDA because of insufficient data. Regardless, its raw data time plot did not exhibit discriminative pattern (Figure 4.12).

All four subsets of immune cells: $CD3^+CD4^+CD8\beta^+$ (Figure 4.6), $CD3^+CD4^{int}$ (Figure 4.9), $CD3^+$ (Figure 4.11), and $CD3^+CD4^+$ (Figure 4.12) exhibited a rapid decrease in their proportion values between 7 and 21 days post-transplant followed by an increase. A common trend was observed in the four aforementioned subsets of immune cells and was more apparent in the latter two. However, it should be noted that this trend was present from most immune cell populations identified in the present study (data not shown).

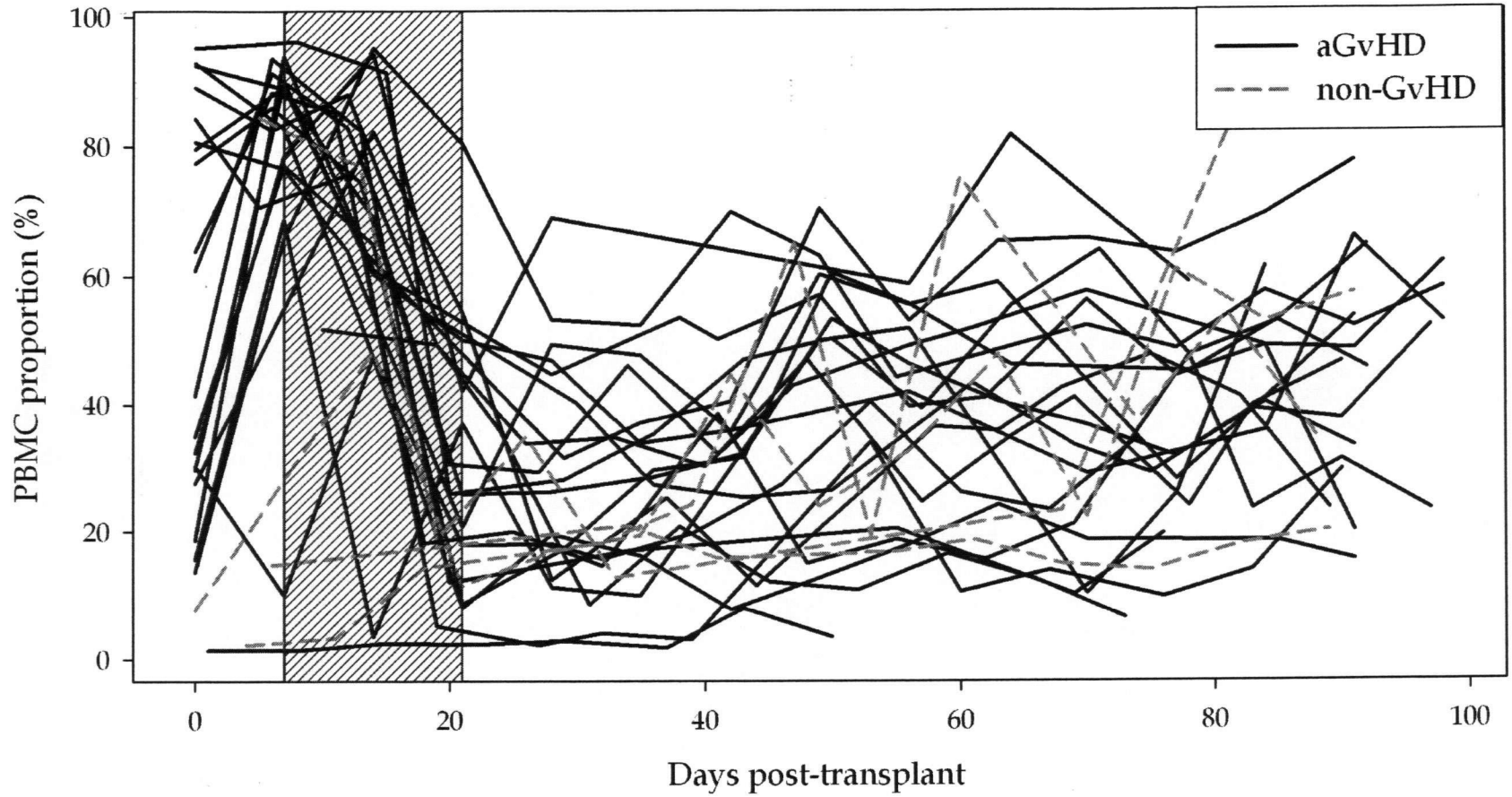


Figure 4.11 Raw data time plot for immune cells CD3+ (aliquot '1Activation') in proportion to PBMC based on samples taken between 0 and 100 days post-transplant. The purpled striped box indicates the time range where data was analyzed via FLDA

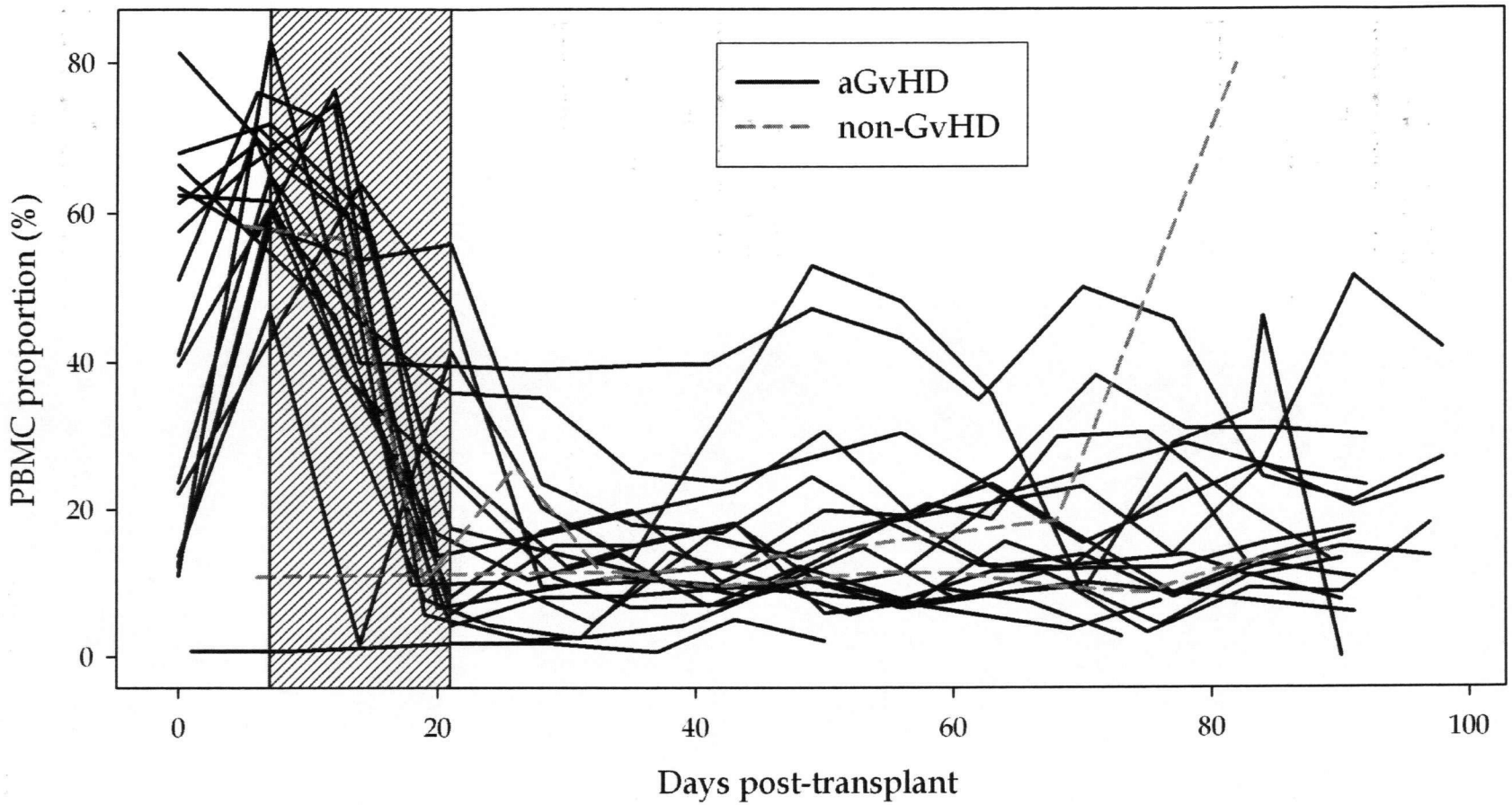


Figure 4.12 Raw data time plot for immune cells $CD3^{+}CD4^{+}$ (aliquot 'rest/act T helper') in proportion to PBMC based on samples taken between 0 and 100 days post-transplant. The purpled striped box indicates the time range where data was analyzed via FLDA.

4.1.4 Static sample size analysis

The FLDA classifier built using the immune cells $CD3^+CD4^+CD8\beta^+$ was the best classifier with consistent pattern observed in both the FLDA estimated signals and the raw data time plots (Figures 4.4 and 4.6). Values obtained closest to the 21 days post-transplant, when the accounted absolute weight value was largest, were used for the static sample size calculation. Even though the FLDA weight value was the largest at seven days post-transplant, the group separation observed around 21 days post-transplant were deemed more reliable because there was no available data from non-GvHD patients between seven and ten days post-transplant. Different sizes of the simulated aGvHD and non-GvHD datasets were tested. The present study compared data between 21 aGvHD and three non-GvHD patients and had an estimated 29% power at 90% confidence level. The unbalanced risk of aGvHD developments among HSCT patients severely compromised the analytical power. In order to achieve a study with 82% power at 90% confidence level, approximately 38 aGvHD and 18 non-GvHD patients will be required (Table 4.2).

Table 4.2 Estimated power of study via the static sample size calculation using $CD3^+CD4^+CD8\beta^+$ proportion values from samples taken closest to 21 days post-transplant.

aGvHD patients required	Non-GvHD patients required	Power estimated from aGvHD ($\alpha \leq 0.1$)	Power estimated from non-GvHD ($\alpha \leq 0.1$)	Average power ($\alpha \leq 0.1$)
21	3	29%	48%	39%
20	10	49%	77%	63%
30	15	62%	92%	77%
38	18	69%	95%	82%
40	20	73%	96%	85%
42	21	73%	97%	85%
46	23	77%	98%	87%
48	24	77%	99%	88%
50	25	79%	99%	89%
52	26	81%	99%	90%

4.2 Classifiers for the onset of chronic graft versus host disease

Only top ranking classifiers from the proportion dataset using samples taken between 21 and 0 days prior to aGvHD diagnosis (Table 4.3) are described below. All others are described in Appendix I. The complete validation results for all subsets of immune cells in each time range are listed in Table J.1 – J.3 for the proportion dataset and Tables J.4 – J.6 for the concentration dataset.

4.2.1 Inconsistent classifiers by pattern outlier

Even though there were more FLDA classifiers with high sensitivity and specificity for the onset of cGvHD compared to aGvHD (Chapter 4), only a fraction of the top ranking classifiers exhibited comparable patterns in both FLDA estimated signals and raw data time plots. From the time range of 21 to 0 days prior to aGvHD diagnosis, all the subsets of immune cells with putative discriminative patterns in their raw data time plots exhibited opposite FLDA signal patterns between groups. All other top classifiers were deemed inconsistent due to the presence of pattern outliers (Table 4.3).

The classifier built using the immune cells $45RA^{+}CD3^{+}$ had an estimated 71% sensitivity and 86% specificity. However there was no clear separation between most of the individual FLDA estimated signals (Figure 4.13a). Only two patients (#6 and #12) had proportion values above 30% between 20 and 7 days prior to aGvHD diagnosis (Figure 4.13b). These values caused the overall FLDA global base values (cross dots, Figure 4.13a) to rise thus separating the two groups.

Table 4.3 Validation results for the top ranking subsets of immune cells from the FLDA classification between the aGvHD & cGvHD and GvHD only patients using samples taken between 21 and 0 days prior to aGvHD diagnosis.

Immune cells	Aliquot	Sensitivity	Specificity	Accuracy	Pattern types
$CD45^{+}CD33^{-}CD15^{+}CD14^{-}$	Myeloids	71%	100%	88%	opposite FLDA signals
$45ROCD3^{-}CD4^{dim}$	rest/act T helper	86%	86%	86%	
$45RACD3^{-}$	rest/act T suppressor	86%	86%	86%	
$CD3^{-}$	3Activation	71%	89%	81%	
$45RACD3^{-}CD4^{dim}$	rest/act T helper	86%	71%	79%	
$45RACD3^{-}$	rest/act T helper	71%	86%	79%	
$CD3^{-}$	rest/act T helper	86%	71%	79%	
$CD3^{-}$	rest/act T suppressor	71%	86%	79%	
$CD3^{-}CD8^{-}$	rest/act T suppressor	71%	86%	79%	
$CD3^{-}$	2Activation	71%	78%	75%	
$CD3^{-}$	T cells	71%	78%	75%	
$CD3^{+}$	rest/act T helper	71%	71%	71%	
$CD3^{+}$	rest/act T suppressor	71%	71%	71%	
$45RACD3^{+}$	rest/act T helper	71%	86%	79%	pattern outlier
$CD4^{dim}$	rest/act T helper	86%	71%	79%	
$45RACD3^{+}$	rest/act T suppressor	71%	86%	79%	
$CD3^{-}44^{+}25^{-}$	1Activation	71%	78%	75%	
$CD3^{-}CD4^{dim}$	3Activation	71%	78%	75%	

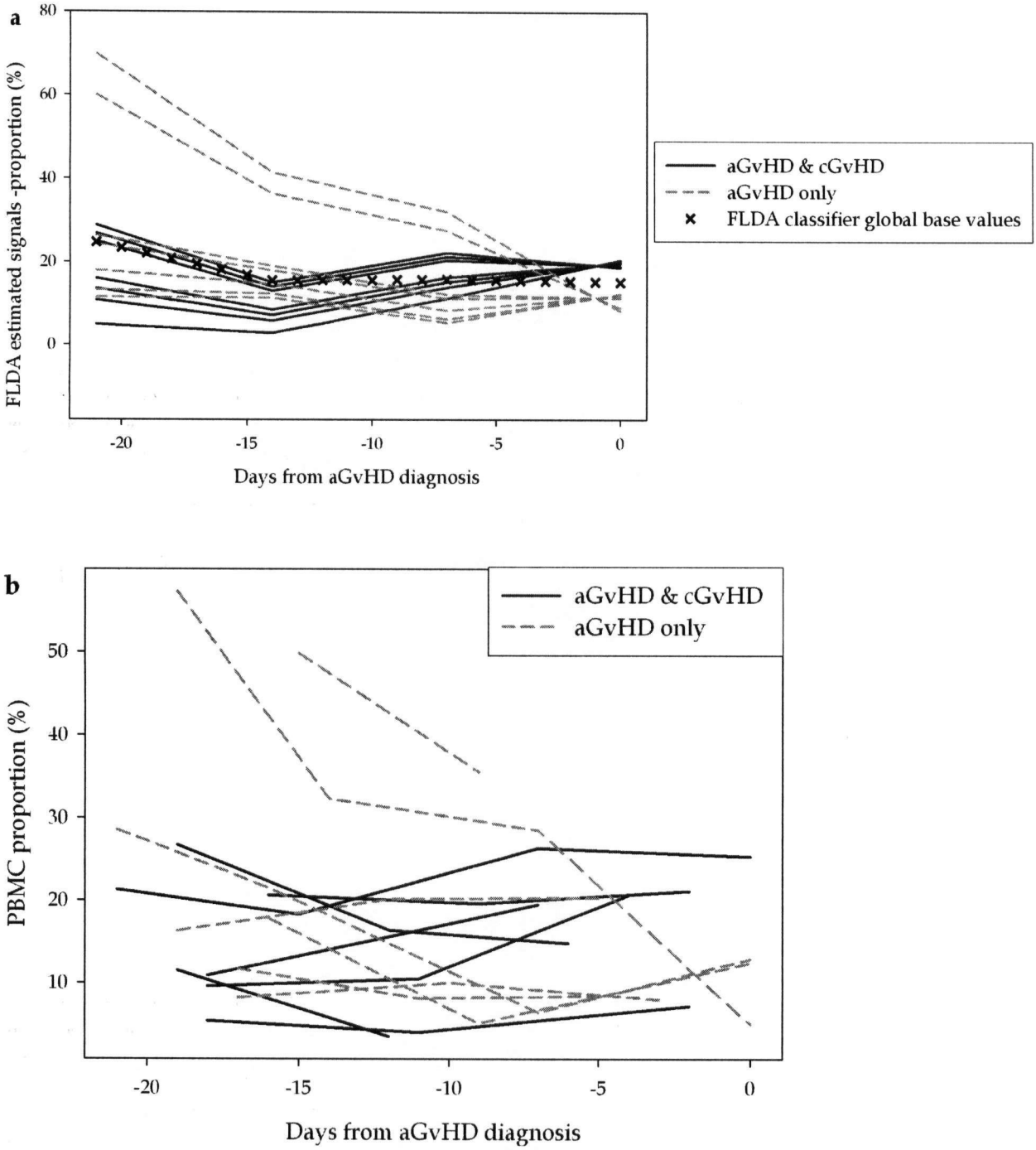


Figure 4.13 Time plot of the FLDA estimated signals (panel a) and raw data (panel b) based on samples taken between 21 and 0 days prior to aGvHD diagnosis for the immune cells $45RA^{+}CD3^{+}$ in proportion to PBMC (%).

4.2.2 Opposite estimated signals between groups

All 13 subsets of immune cells exhibiting consistent patterns between their FLDA estimated and raw data time plots, displayed exactly opposite FLDA signal patterns between the two patient groups. The top two classifiers exhibiting this pattern were $CD45^{+}CD33^{-}CD15^{+}CD14^{-}$ and $45RO^{+}CD3^{-}CD4^{dim}$. The FLDA signals between the two patients groups were the exact opposite of each other (Figures 4.14a and 4.15a). However, this pattern could not be easily identified in the local or extended raw data time plots for either subset of immune cells (Figures 4.14b and 4.15b).

4.2.3 Static sample size analysis

The FLDA classifier built using the immune cells $45RO^{+}CD3^{-}CD4^{dim}$ based on samples taken between 21 and 0 days prior to aGvHD diagnosis had the highest sensitivity (86%) and second highest specificity (86%) among the consistent top ranking (Table 4.3). In this case, the largest and the most reliable group separation were determined to be around 7 days prior to aGvHD diagnosis. Consequently, values obtained closest to that time were used for the static sample size calculation with equal sizes for aGvHD & cGvHD and aGvHD only simulated datasets (Table 4.4). The present study with seven aGvHD & cGvHD and nine aGvHD only patients had an estimated 50% power at 90% confidence level. In order to achieve a study with 81% power at 90% confidence level, approximately 23 aGvHD & cGvHD and 23 aGvHD only patients will be required.

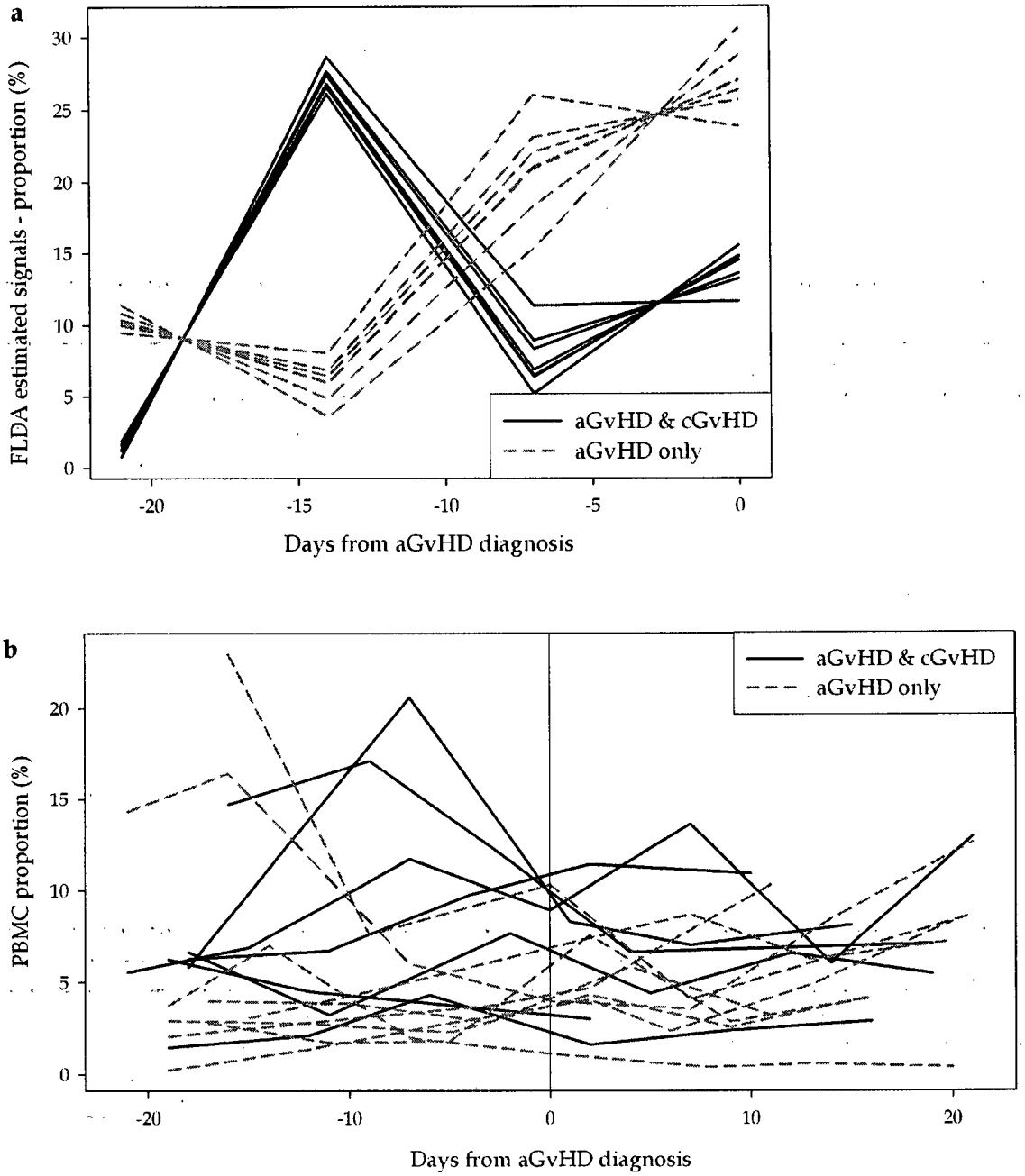


Figure 4.14 Time plot of the FLDA estimated signals (panel a) based on samples taken between -21 and 0 days from aGvHD and time plot of the raw data (panel b) based on samples taken between -21 and 21 days from aGvHD diagnosis for the immune cells $CD45^{+}CD33^{-}CD15^{+}CD14^{-}$ in proportion to PBMC. The aGvHD diagnosis day is labelled at day 0.

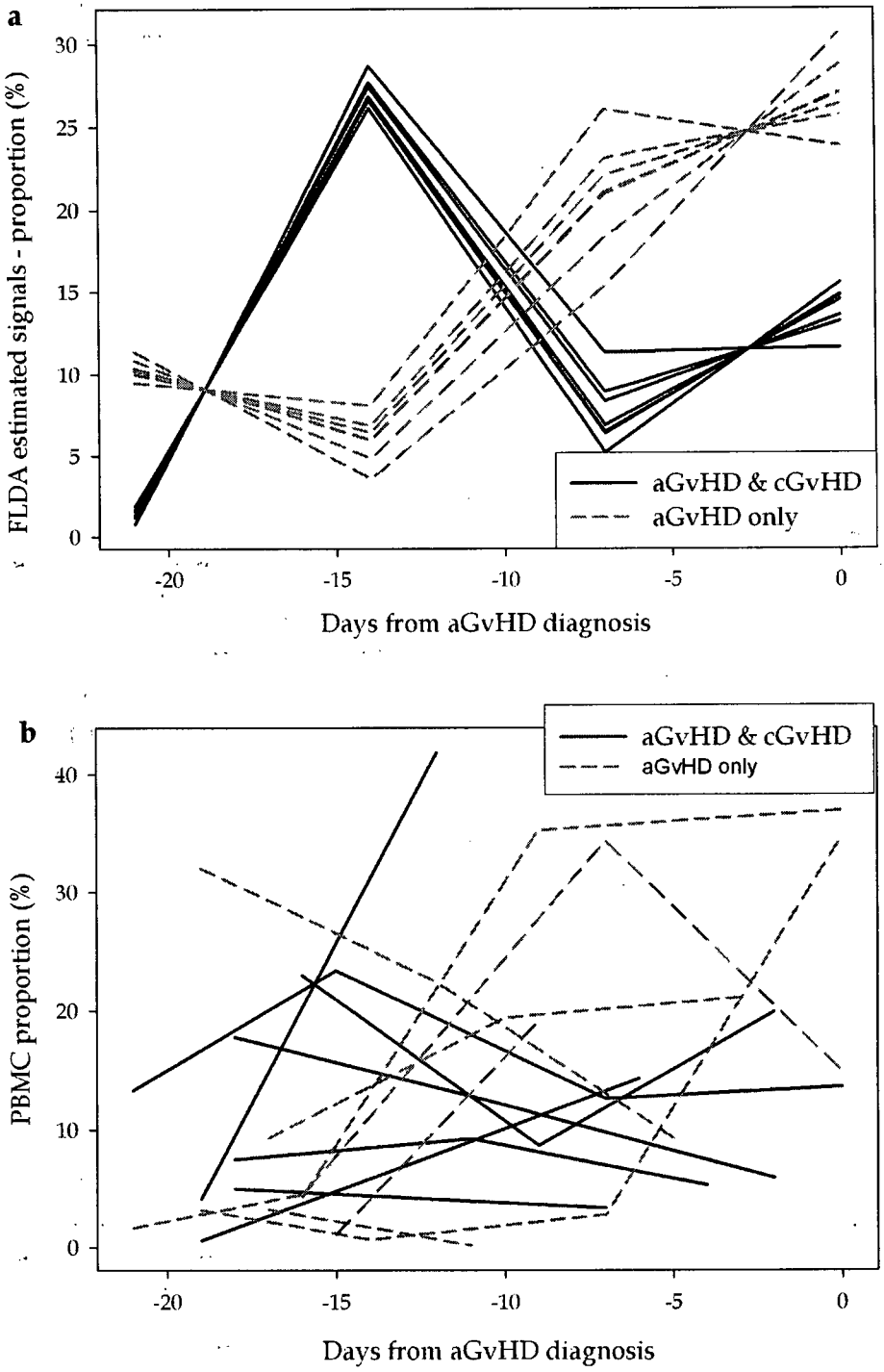


Figure 4.15 Time plot of the FLDA estimated signals (panel a) and raw data (panel b) based on samples taken between 21 and 0 days prior to aGvHD diagnosis for the immune cells 45RO^{+}CD3^{-}CD4^{dim} in proportion to PBMC (%).

Table 4.4 Estimated power of study via the static sample size calculation using $45RO^{+}CD3^{-}CD4^{dim}$ proportion values from samples taken closest to 7 days prior to aGvHD diagnosis.

aGvHD & cGvHD patients required	aGvHD only patients required	Power estimated from aGvHD & cGvHD ($\alpha \leq 0.1$)	Power estimated from aGvHD only ($\alpha \leq 0.1$)	Average power ($\alpha \leq 0.1$)
7	9	67%	34%	50%
10	10	78%	35%	56%
15	15	91%	49%	70%
20	20	97%	58%	77%
23	23	98%	63%	81%
25	25	99%	68%	83%
30	30	100%	74%	87%
35	35	100%	79%	90%
40	40	100%	84%	92%
45	45	100%	88%	94%
50	50	100%	91%	95%
60	60	100%	95%	97%

CHAPTER 5 DISCUSSION

For many patients diagnosed with hematopoietic disorders, HSCT is the only curative treatment [1]. However, the risk of developing fatal GvHD makes it the major limiting factor for broader application of the HSCT procedure [1]. Currently, there is no definitive diagnosis method, standard for treatment or treatment assessment, and very little understanding on the disease's pathophysiologic mechanism.

High-throughput genomic experiments have been useful in elucidating many diseases or conditions [79, 80, 108, 109]. Previous microarray studies have suggested multiple gene expression patterns associated with the onset of GvHD, however none were found to be statistically significant [110-114]. Proteomic methods such as surface-enhanced laser desorption ionization time-of-flight (SELDI-TOF) [115] and capillary electrophoresis coupled mass spectrometry (CE-MS) [116] have been utilized in studying GvHD [117, 118]. Both were pilot studies and further work is needed to link the peptides identified with known proteins in order to infer their role in the immune system and GvHD manifestation. Compared to genomic methods such as microarrays, proteomic methods and FCM have the advantage of visualizing physical characteristics of cells such as protein functions directly.

The main hypothesis of the present study was that one or more immune cell populations with differential temporal patterns that correlate to the onset of either aGvHD or cGvHD could be identified and potentially be used to predict the disease. The present dataset had the complexity of a microarray data with a large number of immune cell population abundances that were screened for their potential discriminative powers for either aGvHD or cGvHD. The present study was a pilot project with the main objective of assembling a temporal analysis pipeline for the high-throughput clinical FCM dataset. To the best of my knowledge, there is no

existing temporal analysis pipeline purposely designed for large-scale FCM data. Consequently, the majority of the discussion is devoted to experimental and analytical difficulties of the present study and corresponding improvements for a future one. In sections 5.1 to 5.3, obstacles from each step of the analysis pipeline (Figure 2.1): QA, data transformation, and temporal classification are discussed. Then possible predictive models and pathophysiologic mechanism for aGvHD and cGvHD are examined in section 5.4. A list of specific recommendations to improve the efficiency of future studies where current GvHD models will be validated is discussed in section 5.5.

5.1 Quality assurance

QA is an essential step in the analysis of any high-throughput dataset [119-121], probably more so in the case of clinical data with limited samples. The assumption of the QA test used in this study was that distributions of common intensities from different aliquots of the same sample should be similar [3]. Two aliquots were identified as outliers in the QA test on ungated data (Figure 3.1). Whereas 29 aliquots were identified as outliers in the QA on gated $CD3^+$ and $CD3^-$ live MNC populations (Table 3.1). Among the outliers, I observed both intensity shift (Figures 3.4 and 3.5), density or ECDF shape difference (Figures 3.2 and 3.3), or the combination of both (Figure 3.6). A simple intensity shift might indicate a different concentration of reagents during the staining procedure in the well corresponding to the outlier aliquot [3]. However, sources for other distribution differences were less understood. While further study is required to investigate the precise causes of outliers and unusual trends discussed below, they indicated potential complications with the FC-HCS technique [2]. At the end, based on the QA test results, approximately 1.8% of the dataset were removed from subsequent analyses because the differences observed might not be biologically but experimentally motivated.

5.1.1 Quality assurance on ungated and gated data

Analyzing both ungated and gated data each had their own advantages. Ungated data offered QA assessment without the interference of manual gating. On the other hand, the QA test on gated data provided an assessment of the gate quality. In addition, the QA test on the gated data provided an assessment of the population data that were used in the subsequent FCM analyses. There was no overlap of outliers identified in the two parts of the QA test.

For QA visualization, FSC was observed to have more informative patterns than SSC in the QA test of ungated data [3]; while both FSC and SSC displayed similar patterns and were both useful in QA test of gated data (Figures 3.2 and 3.3). FSC and SSC, which are strongly influenced by cell size and granularity respectively, are often used together in FCM gating to identify and exclude dead cells from further analyses (Figure 1.1a). Dead cells and debris that were excluded in the gated data have a very broad SSC intensity range and overlap with the relatively narrow SSC intensity range from MNC (Figure 5.1). Thus, the FSC intensity was more informative than the SSC intensity in the ungated data because more variations were observed from different cell types. Many unusual patterns might only be visible after removal of dead cells and debris via the gating procedure. Visualization using the CD3 intensity was proven to be the least informative in outlier identification as more variations were expected and observed because of the sensitivity of the antibody and the limited number of aliquots available (Figure F.1).

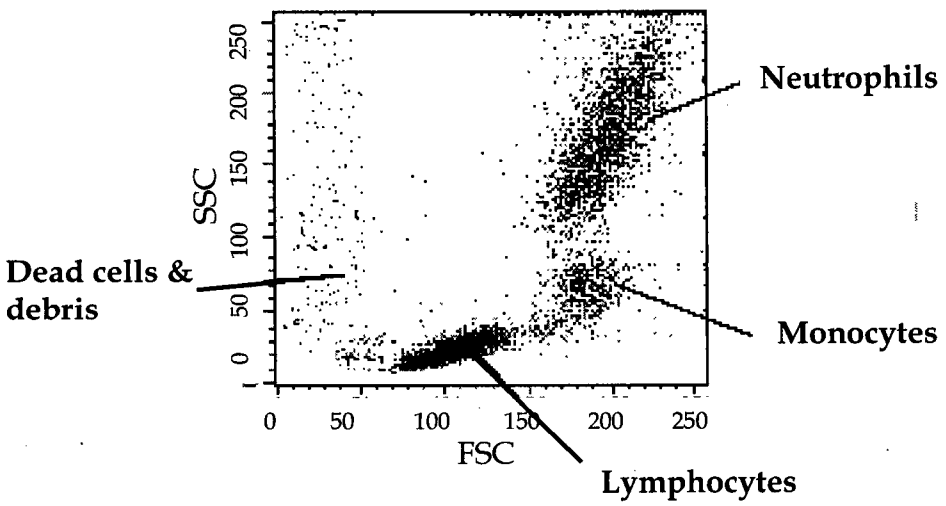


Figure 5.1 A pictorial example of FSC vs. SSC dot plot from a normal peripheral blood sample (adapted from [122]).

Outliers and other unusual patterns were frequently found in either $CD3^+$ or $CD3^-$ population but rarely both (Table 3.1). These observations could be related to the fact that $CD3^+$ and $CD3^-$ gates represented two different immune cell populations. The gated data only included live PBMCs that were divided into $CD3^+$ and $CD3^-$ populations. CD3 and TCR are exclusively expressed on 70% to 80% peripheral blood T cells [123]. Thus, $CD3^+$ and $CD3^-$ populations represented T cells and non-T cells among the PBMC populations. Future studies are needed to determine why these two cell populations behave differently and if certain cell population is more prone to experimental errors.

The trends observed among the outliers (Table 3.1) and unusual patterns (Tables 3.2 and 3.3) indicated possible non-random plating effects. Many outliers were mapped to aliquots plated close to each other or cluster of aliquots in the middle of the plate (Table 3.4). These trends potentially indicate: 1. Improper washing leading to false reading from cluster of wells and wells in the middle; 2.

Contamination affecting multiple wells next to each other; 3. Edge drying causing false readings from wells at the edges; 4. Different reagent or cell concentrations among wells; and 5. Different logarithmic compensations. The unusual pattern from the two rest/act aliquots (Table 3.3 and Figure 3.8) which were often mapped to a separate plate may also suggest noticeable differences in readings from different plates or effects of different sample storage time. Further examination of the FCM gates from the FSC-SSC contour graphs (Figure 3.7) indicate that the occurrences (Table 3.2) of unusually large variations (Figure 3.6) among all aliquots could result from interference of dead cells or a minimal amount of viable cells in patient samples. From a sample with a minimal amount of viable cells, the proportion of any subgates may be incorrect because there are not enough cells in the sample to represent the overall population properly.

5.1.2 Quality assurance via raw data time plots

Raw data time plots used in the visualization of FLDA classification may also be used as an additional QA test. Biologically, it is impossible to have an abrupt increase in either PBMC proportion or cell concentration such as the two peaks observed from patient #6 at 53 and 90 days post-transplant (Figures 4.6 and 4.9b). Upon visual inspection of the gated FCM data (Figure 4.10), I discovered that these abrupt increases were the result of an experimental error likely from a minimal amount of viable cells in the FSC-SSC gate. While the QA test via raw data time plots could be very useful in identifying experimental errors, it would require long time-series data. In addition, implementation of this QA test to large-scale data would require further studies on the rate of immune responses to establish a threshold for the rate of increase from a biological standpoint.

5.1.3 Robustness of the flow cytometry high content screening technique

Unfortunately, not all the trends mentioned above were always consistent in their distribution in the plates. There was only enough evidence to suggest possible plating effects but not to confirm it. Further studies are required to investigate the robustness of the FC-HCS technique [2], to elucidate the precise causes of these outliers, and to improve the present QA test procedure. Preferably, a larger quantity of samples from healthy individuals would provide a larger number of aliquots for outlier identification and a lower likelihood for occurrences of minimal viable cells to be used for future studies. Frequencies of outliers observed in different antibody-fluorochrome intensity, different location within a plate and between plates could be used to validate the current results. Furthermore, a larger number of aliquots may be used to determine the overall experimental variations among aliquots. Statistical tests such as the analysis of variance and visualizations such as box plots [124] in addition to the existing visualization methods for the outlier identification could potentially identify bias caused by the current manual visualization. Fortunately, some of the potential causes for these outliers such as difference in reagent concentrations and different sample storage time between plates can be easily avoided with an organized experiment design and a smooth instrumental pipeline. In addition, a simple procedure of random plating as discussed in section 5.5.1 may be used to combat effects of these potential plating effects.

5.2 Data issues

5.2.1 Patients

The present dataset is consisted of a heterogeneous group of patients (Table 2.1). Two patient grouping comparisons using prior GvHD diagnosis knowledge were selected to train FLDA classifiers. The comparisons were also designed to conserve the study population where the main factor was the onset of aGvHD or cGvHD.

The first patient group comparison between aGvHD and non-GvHD patients was devised to identify subsets of immune cells with patterns that correlate with the onset of aGvHD. All 21 patients who were diagnosed with aGvHD were included. However, only four out of seven patients not affected by aGvHD or cGvHD were included (Table 2.1). Three patients who were not diagnosed with aGvHD prior to their death before 100 days post-transplant were omitted from the analyses. This strict selection was chosen because I can only be certain that patients would not have developed aGvHD if there were information available past 100 days post-transplant, when most aGvHD diagnoses were made. The 100 days post-transplant is generally recognized as the cut-off for aGvHD diagnosis; however, it is possible to diagnose aGvHD after 100 days post-transplant [39]. Please be noted that one of the remaining four non-GvHD patients was often omitted in the FLDA analysis because of lack of data.

The second patient group comparison between aGvHD & cGvHD and aGvHD only patients was devised to identify subsets of immune cells with patterns that correlate with the onset of cGvHD which occurred weeks or months after the diagnosis of aGvHD. Among the 21 patients who were diagnosed with aGvHD, seven patients were later diagnosed with cGvHD and were included in the aGvHD & cGvHD dataset. However, only nine out of 14 were considered as aGvHD only patients because I could not be sure of patients who died or withdrew from the study after their aGvHD diagnoses (n=5) that they would not have developed cGvHD. *De novo* cGvHD patients were not considered. While diagnoses for both aGvHD and cGvHD are not definitive, a retrospective study in cGvHD diagnosis was performed by Vogelsang and colleagues in 2001 and found 25% misdiagnoses on active cGvHD [125].

Errors from incorrect patient groups from either false cut off diagnosis time or misdiagnoses [35, 125] could be exaggerated in the present study due to the

limited number of patient available and cause inconsistent classifiers. These exaggerated inconsistent classifiers may be avoided with an external test dataset with an adequate number of patients. However, the sensitivity and specificity of the new diagnostic model created using a dataset with potential misdiagnoses will be limited by the accuracy of the present diagnostic methods. Tolerance to misdiagnoses is discussed further in section 5.5.6.

5.2.2 Sampling time ranges

The three time ranges were selected to present patterns before and during the full clinical manifestation of aGvHD. These patterns were in turn analyzed via FLDA in order to identify immune cell populations that can predict either aGvHD or cGvHD. I decided that the time range most suitable for predicting the onset of aGvHD was between 7 and 21 days post-transplant. Classifiers found in this time range should be useful in predicting aGvHD because only four out of the total 21 aGvHD patients were diagnosed prior to 21 days post-transplant (Figure 4.1). The aGvHD diagnosis rate for the present study was comparable with previous studies where most aGvHD diagnosis is made within the first 100 days and most prominently between 14 and 42 days post-transplant [15]. The other two time ranges: 21 to 0 days prior to aGvHD diagnosis and 0 to 21 days post-aGvHD diagnosis were selected to reflect patterns occurring immediately before and after the aGvHD diagnosis. Molecular changes leading to or result of aGvHD may contribute to cGvHD manifestation at a later date as cGvHD may be a continuation of aGvHD [36]. For predicting the onset of cGvHD, the time range before the aGvHD diagnosis was selected because predictions would not be confounded by different aGvHD treatments. All three time ranges were purposely designed to be short in order to avoid loss of synchronization and smoothing requirements.

5.2.3 Proportion and concentration flow cytometry datasets

Both the proportion and concentration datasets were tested because they might contribute different insights into the immune responses. Previous GvHD studies have used both proportion (either to PBMC or chimerism) [30] and concentration values [27, 28]. However, more errors and thus more inconsistent classifiers were expected in the concentration datasets because different samples sometimes taken at different time were used to estimate the immune cell concentration. These errors could be avoided for future studies with a coordinated sample quantity standard.

5.3 Temporal analysis

Static analyses using rates of immune cell population changes from patients at multiple time points were performed. The rates of changes were extensively screened by a combination of dimension reduction via between group analysis [126] and hierarchical clustering via hierarchical ordered partitioning and collapsing hybrid [127]. However, the static approach failed to analyze the current dataset properly because of missing values, lack of synchronization events, and diverse patient response time (Table 2.1, Appendix A). Because of these shortcomings, I undertook a temporal approach for the present study. While temporal analysis has been suggested to be more efficient in analyzing biological process occurring across time [46], there are a limited number of available algorithms. During my temporal analysis investigation, I encountered three main challenges in adapting a suitable temporal analysis method for the current dataset:

1. Tolerance for missing values and non-uniform sampling time
2. Short vs. long time-series data
3. Limited number of samples

Using an excerpt of the current dataset, and combinations of basis order and knot placements, I determined that B-spline with a linear basis and weekly knot placements was most reflective to the raw data pattern (Figures 3.9 and 3.10). A B-spline fitted with basis order two best reflected the actual raw data especially for short time-series dataset such as the one used in the present study (Figure 3.9). Weekly knot placement was selected to fit the weekly sampled dataset because flexible knot placement was not compatible with the FLDA algorithm. While discrepancies between B-spline and raw data patterns were minimized, they could still exist and be exaggerated in a short time-series dataset with various sampling rates and missing values among the study patient population.

Similar to most of the existing temporal algorithms, FLDA was intended for long time-series data with more than eight time points [128]. Yet it was difficult to analyze a time range longer than three weeks (assumed one time points per week) without the possible loss of synchronization. In addition, usage of a long time-series data in the present clinical dataset with diverse patients' response time would require potentially biased smoothing and registration procedures. As a pilot study, short time-series data were purposely selected. However, short time-series data highlighted effects of missing values (Figure 4.2) and pattern outliers (Figure 4.13) resulting in inconsistent FLDA classifiers. While LOOCV might over-estimate the classification accuracy [76], it does reflect, to a certain degree, the overall stability of the classifiers. Unfortunately, the influence of pattern outliers to the FLDA global base values was still observed (Table 4.3 and Figure 4.13). There are many possible causes for these visually extreme pattern outliers from either the proportion or the concentration dataset and they may be remedied by improvements discussed in section 5.5.

Slightly different LOOCV results and FLDA classifier patterns from redundant readings such as the $CD3^+$ immune cell population (Table 4.1 and

Appendices H & J) demonstrated the instability of the FLDA classification with the limited number of patients available to the present study. These errors may be remedied by an external validation with large and separate testing dataset as proposed for future studies (Section 5.5). Ideally, continuous discriminative patterns between two groups of patients are preferred. However, visually clear discriminative pattern spanning a few day are sufficient to be identified by FLDA.

5.4 Predicting the onset of graft versus host disease

The biological motivation behind this study was to identify subsets of immune cells that may be used as molecular predictors of either aGvHD or cGvHD before the full clinical manifestation. All the top ranking classifiers and their corresponding subsets of immune cells might serve as potential GvHD diagnostic markers even if they do not correspond to known immune cell populations. Without knowing their function in the immune system, one limitation is that these subsets of immune cells could not be used to elucidate GvHD pathophysiologic mechanism. Lack of correction for multiple testing resulting in possible incorrect classifiers should be noted with the findings discussed below which must be validated via a future study (section 5.5).

5.4.1 Acute graft versus host disease

All the consistent top ranking classifiers for aGvHD were based on the proportion dataset (Tables 4.1 and H1-H3). The three top ranking classifiers from the concentration dataset were inconsistent due to missing values and pattern outliers (Appendix G). This was expected because there were more errors in the concentration dataset. Interestingly, all but two top ranking classifiers from the proportion dataset, target $CD3^+$ T cells or T cell subsets (Table 4.1). Apart from the inconsistent classifier $CD2^{dim}CD16^+CD56^+CD3^-$ based on samples taken between 7

and 21 days post-transplant (Figure 4.2), the only top ranking classifier targeting non-T cells ($CD3^-$) were $CD3^-$ and $CD2^{dim}CD16^+CD56^-CD3^-$ based on samples taken between 0 and 21 days post-aGvHD diagnosis. All the $CD3^+$ and its subsets (Table 4.1) displayed similar patterns with higher PBMC proportion values and greater fluctuation in the aGvHD patients when compared to the non-GvHD patients (Figures 4.4, 4.6, 4.8, 4.9, G.1, G.4, and G.5). Because all viable cells were divided into $CD3^+$ and $CD3^-$ cell populations (Figure 4.7), for the proportion dataset the $CD3^-$ cell population displayed the exact opposite pattern with higher PBMC proportion values from the non-GvHD patients (Figures G.2 and G.3). Higher proportion of $CD3^+$ immune cells in the PBMC represents higher numbers of T cells in the peripheral blood that could be the result of inflammatory response toward the 'foreign' host tissues.

Even though there is no precedent on a B-spline temporal pattern as predictive model for GvHD; the observation of higher proportions of T cells after HSCT in the aGvHD patients is comparable with other studies [27-29]. The current findings combined with other previous GvHD studies suggest that GvHD is a complex disease. While T cells' critical involvement in aGvHD (Figure 1.2) is proven by significantly less aGvHD occurrences in T cell depleted BMT [20-26], the exact subset of T cells with predictive pattern is yet to be identified.

The most persistent correlation to the onset of aGvHD was observed from the immune cells $CD3^+CD4^+CD8\beta^+$ and its subpopulation $CD3^+CD4^+CD8\beta^+CD8^+$ (Table 4.1). These two subsets of immune cells were higher and had greater fluctuation in the aGvHD patients, compared to the non-GvHD patients after BMT (Figures 4.4 and 4.8). This pattern was found to persist until 120 days post-transplant (Figure 4.6). FCM data in the contour graphs (Figure 4.5) confirmed the FLDA results. Interestingly, none of the related $CD3^+$ immune cell populations with the presence of CD4 or CD8/CD8β but not both exhibited similar pattern or had high estimated

sensitivity and specificity (Table 4.1). A future study with sufficient power (section 5.5.2) will need to determine the validity of the classifiers $CD3^+CD4^+CD8\beta^+(CD8^+)$ as predictors of aGvHD.

The two subsets of immune cells $CD3^+CD4^+CD8\beta^+(CD8^+)$ target cell populations that co-express CD4, $CD8\alpha\beta$ heterodimers and $CD\alpha\alpha$ homodimers. These specific phenotypes might contain an unusual group of double positive (DP) T cells and putatively suggest that the key T cell subtype for the prediction and development of aGvHD could be this unusual T cell subset. This also explains why the $CD3^+$ and $CD3^-$ immune cell populations were not identified as a top classifier based on samples taken between 7 and 21 days post-transplant (Table J.1). Large $CD3^+$ proportion values were observed from both patient groups right after BMT (Figure 4.11) could be the result of residual recipient T cells which are know to survive the preparative treatments [129, 130]. If so, there would only be a minimal impact on the DP T cells because of its low abundance and may not exist in the recipient prior to the BMT procedure [131-133].

The most prominent theory on T cell maturation suggests that T cell maturation is limited to thymus [133] (Figure 5.2). After the intense screening for the MHC restriction and self-tolerance, more than 95% immature DP T cells are killed via apoptosis. The remaining cells develop into mature single positive T cells (either $CD3^+CD4^-CD8^+$ or $CD3^+CD4^+CD8^-$) and are exported into peripheral blood. Consequently, DP T cells are not normally expected to occur in peripheral blood. However, this distinction was contradicted by many reports of peripheral DP T cells in humans [134-140]. The proportion values of DP T cells observed in the present dataset from the non-GvHD patients agreed with previous studies that most healthy individuals had less than 3% peripheral DP T cells [134, 137]. Increased DP T cells have been previously observed in older individuals [138, 139] and individuals with viral infections [135, 140].

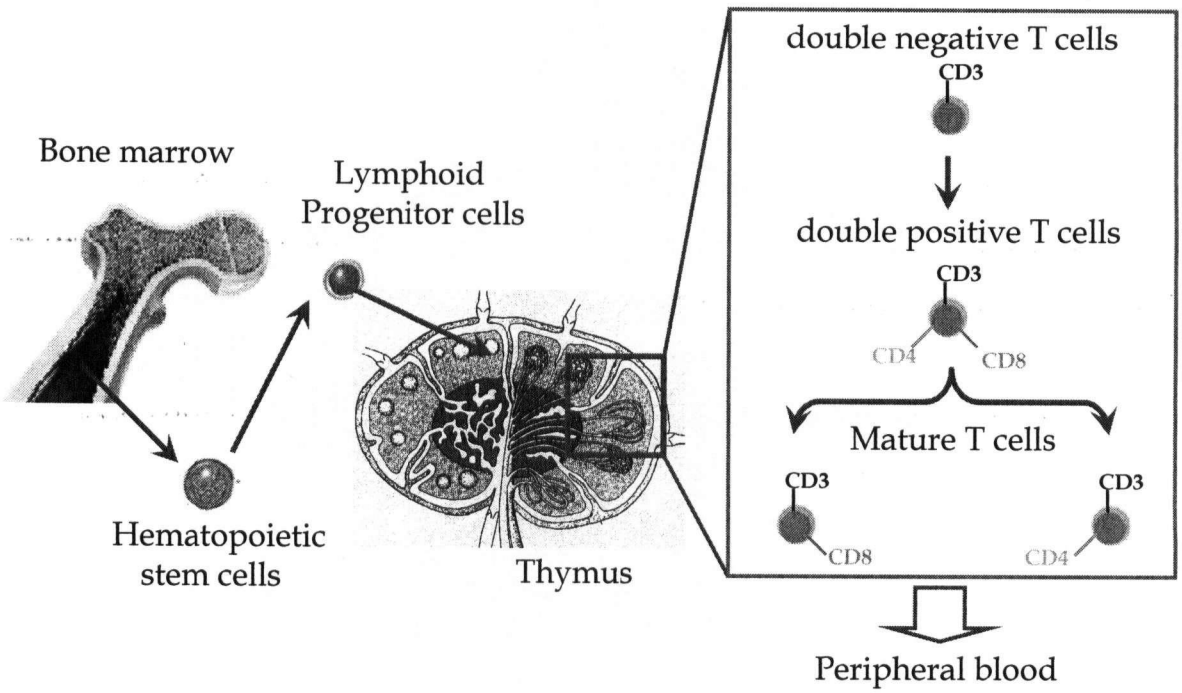


Figure 5.2 T cells development and maturation.

The origin and function of DP T cells are still not understood. Two DP T cell pathways have been proposed [131]: premature release from thymus and extrathymic maturation [141-143]. While premature release of DP T cells from thymus is more likely in a HSCT patient where thymus damages from either the preparative treatments or aGvHD have been reported [37]; the DP T cell population observed in the present study (Figure 4.7) appears to express lower levels of CD4 than typical immature thymocytes [144]. Thus, it is more likely that the DP T cells observed are mature antigen specific cells of extrathymic origin [145] and may play a role in the aGvHD manifestation. DP T cells may consist of two or more functional subgroups [135, 146, 147]. Consequently, future studies are needed to define the activation and differentiation status of the DP T cell population using additional markers (section 5.5.4).

The $CD2^{dim}CD16^{+}CD56^{-}CD3^{-}$ classifier, though targeting non-T cells, exhibited a similar pattern with higher PBMC proportion values from the aGvHD patients, compared to the non-GvHD patients between 0 and 21 days post-aGvHD diagnosis (Figure G.3). The combination of $CD3^{-}$ and $CD16^{+}$ exclusively targets NK cells [148]. However, previous studies on NK cells only distinguished two major NK cell subsets, both usually associated with $CD2^{+}$ or CD2br: $CD56^{br}$ and $CD56^{dim}$ [149, 150]. The subset of immune cells $CD2^{dim}CD16^{+}CD56^{-}CD3^{-}$ most likely targeted a NK cell subset similar to the highly dysfunctional NK subset $CD56^{-}CD16^{+}$ detected in HIV patients [151]. In vitro functional study of NK cell subset $CD56^{-}CD16^{+}$ [151] suggested that expansion of $CD56^{-}$ NK cells cause impaired NK cell function with lower cytotoxoic activity and cytokines production. Presently, there is no existing study on the $CD56^{-}$ NK cells and its possible role in GvHD development.

Another unknown cell type with the $CD3^{+}CD4^{int}$ phenotype (Figures 4.9 and G.4) also exhibited a similar pattern to $CD3^{+}$ cells based on samples taken between 0 and 21 days post-aGvHD diagnosis (Table H.3). The closest known T cell subtype with a similar phenotype is that of helper T cells ($CD3^{+}CD4^{+}$). Their main function in the immune response is to secrete cytokines responsible for proliferation and differentiation of T cells [133]. In the present study, $CD3^{+}CD4^{+}$ temporal patterns at any time range were not found to correlate with the onset of aGvHD (Figure 4.12). Further study is required to determine if $CD3^{+}CD4^{int}$ cells are a distinct immune cell population and their functions in the immune systems.

5.4.2 Acute graft versus host disease prediction model using $CD3^{+}CD4^{+}CD8\beta^{+}$

The FLDA classifier built using immune cells $CD3^{+}CD4^{+}CD8\beta^{+}$ and samples taken between 7 and 21 days post-transplant, had the highest sensitivity (86%) and specificity (100%) among the consistent classifiers. Classification of a new patient with sampled time points at 7, 14, and 21 days post-transplant, can be made using

the following model (Figure 5.3). Based on Equation 1.4, linear discriminant value can be calculated with Equation 5.1.

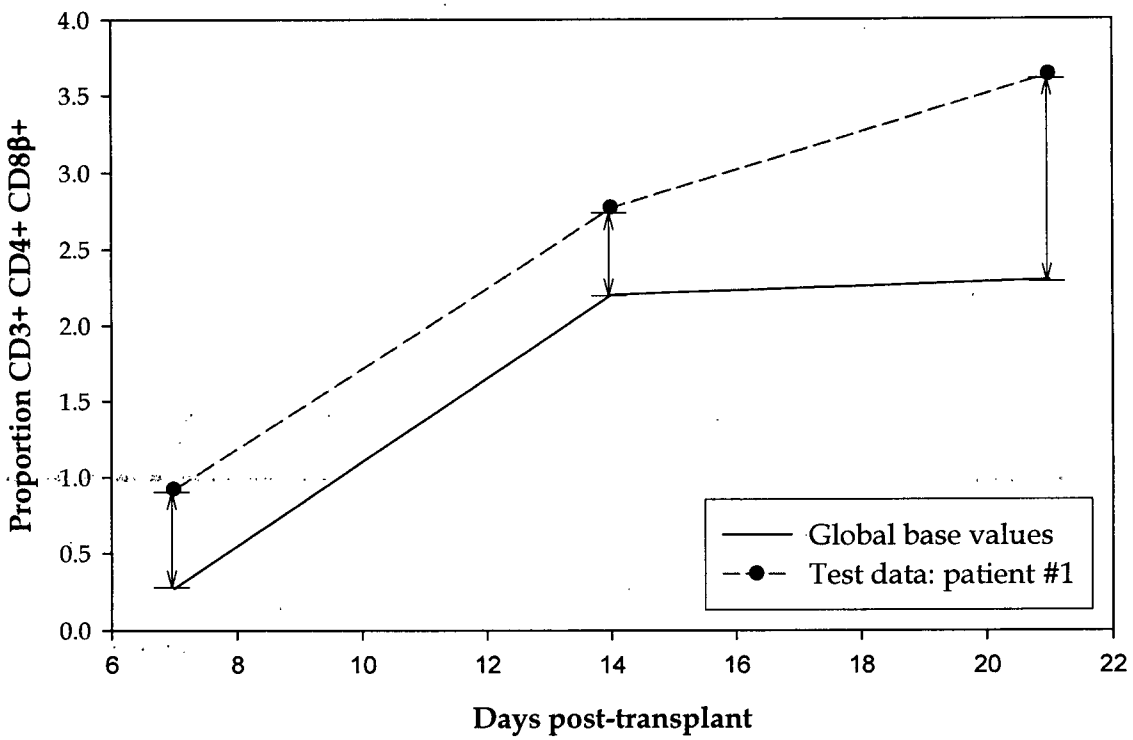


Figure 5.3 An example of FLDA classification using immune cells $CD3^+CD4^+CD8\beta^+$ in proportion to PBMC

$$\hat{\alpha}_X = -1.0823 \quad 0.0123 \quad -0.1767 \cdot (X - \begin{matrix} 0.2718 \\ 2.2034) \\ 2.3000 \end{matrix}$$

Equation 5.1 The aGvHD prediction formula for patient data sampled at 7, 14, and 21 days post-transplant

In a resubstitution example, patient# 1 with observed values $X = \begin{matrix} 0.92 \\ 2.77 \\ 3.63 \end{matrix}$ had an estimated linear discriminant value of -0.9. Based on the linear classification rule, patient #1 who was diagnosed with aGvHD at 26 days post-transplant and with $\hat{\alpha}_x$ smaller than zero, is classified into the aGvHD class, a true positive (Figure 5.3). The detail calculation of the weight values is available in Appendix K.

5.4.3 Chronic graft versus host disease

None of the consistent top classifiers for cGvHD exhibited patient group separation as clearly as the top ranking classifiers for aGvHD (section 5.4.1). Among the 13 (eight unique) FLDA classifiers that exhibited the opposite FLDA signal pattern (Table 4.3), none was comparable to prior cGvHD studies (Figure 1.3). None of these discriminative patterns was observed after aGvHD diagnosis probably because of different patient responses to various treatments (Table J.3).

During the FLDA analysis process, random experimental errors from each sample were estimated and removed in the final FLDA classification. This could be the reason why these FLDA signal patterns exhibiting the opposite signal pattern (Figures 4.14a, 4.15a, and I.2a) could not be easily identified in the corresponding raw data time plots (Figures 4.14b, 4.15b, and I.2b). Another plausible explanation is an over-correction from FLDA, which could be amplified because of the limited number of patients available. However, the frequent occurrence of this opposite FLDA signals pattern between the patient groups suggested potential cGvHD diagnosis markers that will require further investigations.

All the classifiers exhibiting the opposite FLDA signals pattern are built from subsets of immune cells representing heterogeneous T cell ($CD3^+$) and non-T cell ($CD3^-$) subsets. The one common CD marker among all these immune cell subsets: $CD45^+CD33^-CD15^+CD14^-$, $45ROCD3^-CD4^{dim}$, $45RACD3^-$, $45RACD3^-CD4^{dim}$, and $45RACD3^-$ (Table 4.1) is CD45 (RO/RA). CD45 is one of the major accessory molecules in immune response and functions as a protein tyrosine phosphatases [152]. The relationship between these immune cell subsets and cGvHD manifestation is not known.

The classifier built from immune cells $CD3^+CD4int$, based on samples taken between 0 and 21 days from aGvHD diagnosis, was identified as one of the top ranking classifier for cGvHD (Table J.3). The same subset of immune cells was also identified as one of the top ranking classifiers for aGvHD (Table 4.1). Here the PBMC proportion values for $CD3^+CD4^{int}$ were generally higher in the aGvHD only patients compared to the aGvHD & cGvHD patients (Figure I.3). Like the classifier using $CD3^+CD4^{int}$ for aGvHD prediction, the relationships between this unknown cell population with the $CD3^+CD4^{int}$ phenotype and the development of cGvHD is not yet defined.

5.4.4 Chronic graft versus host disease prediction model using $45RO^+CD3^-CD4^{dim}$

The FLDA classifier built using immune cells $45RO^+CD3^-CD4^{dim}$ in proportion to PBMC and samples taken between 21 and 0 days prior to aGvHD diagnosis, had the highest estimated 86% sensitivity and 86% specificity (Table 4.3), excluding the inconsistent classifiers. Classification of a new patient with sampled time points at 21, 15, 7, and 0 days prior to aGvHD diagnosis can be made using the following model (Figure 5.4). Based on Equation 1.4, linear discriminant value can

be calculated by multiplying the new values with the determined weight values (at each time point) (Equation 1.3):

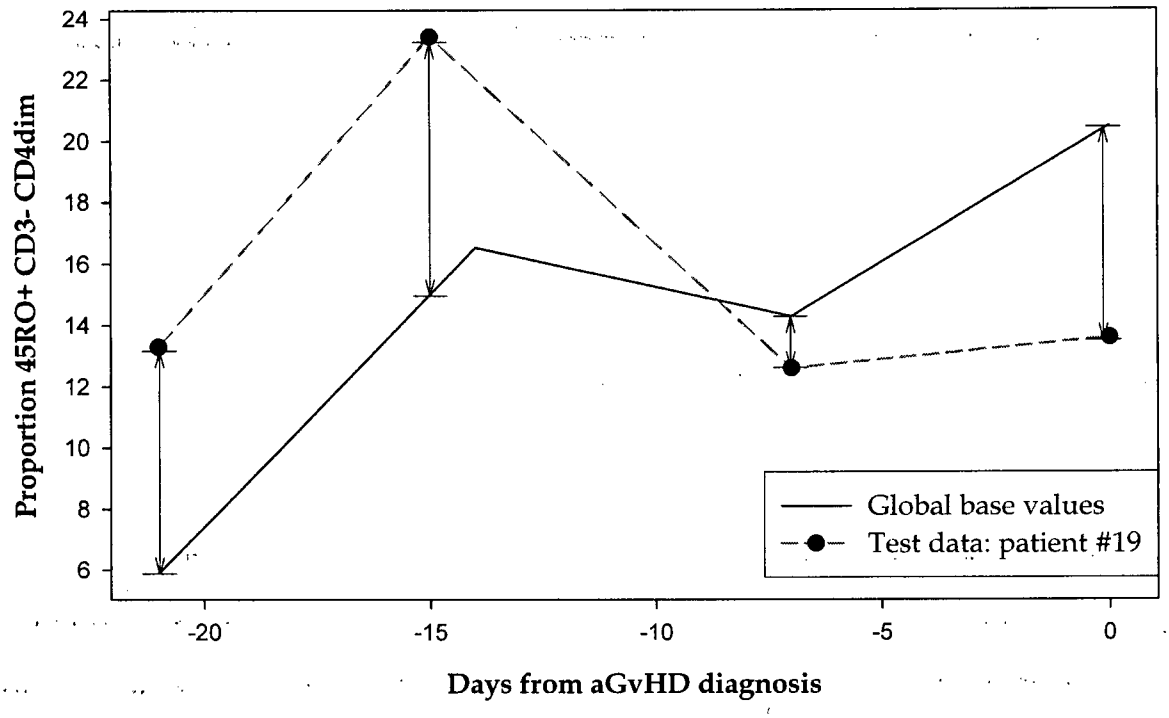


Figure 5.4 An example of FLDA classification using immune cells $45RO^{+}CD3^{-}CD4^{dim}$ in proportion to PBMC.

$$\hat{\alpha}_X = 0.0762 \quad -0.1436 \quad 0.1191 \quad 0.1091 \cdot \left(X - \begin{pmatrix} 5.8992 \\ 15.0097 \\ 14.2864 \\ 20.4889 \end{pmatrix} \right)$$

Equation 5.2 The cGvHD prediction formula for patient data sampled at 21, 15, 7 and 0 days prior to aGvHD diagnosis

In a resubstitution example, patient #19 with observed values $X = \begin{matrix} 13.3 \\ 23.4 \\ 12.6 \\ 13.6 \end{matrix}$ has an estimated linear discriminant value of -1.59. Based on the linear classification rule, patient #19 who was diagnosed with both aGvHD and cGvHD and with a negative $\hat{\alpha}_x$, is classified into the aGvHD & cGvHD class, a true positive (Figure 5.4). The detail calculation of the weight values is available in Appendix L.

5.5 Recommended improvements

The main objectives of the present pilot study were to assemble a novel temporal analysis pipeline for the high-throughput clinical FCM data and recommend improvements in preparation for future studies. While I have demonstrated the applicability of the analysis pipeline (Figure 2.1), there are seven practical and two tentative improvements needed to achieve better efficiency and power for future studies.

5.5.1 Random plating

The first recommendation for experiment procedures is random plating. The results of the QA test on the current dataset presented possible plating effects (Table 3.4). While further analysis (section 5.1.3) is required to elucidate the plating effects, random plating [153] will aid in minimizing the likelihood that changes observed are due to plating arrangements. For example, if samples taken prior to BMT are always plated in the first two columns, then it will not be clear if changes observed from these samples are from biological changes or the edge drying effect.

5.5.2 Patient recruitment

The second recommendation is to increase patient recruitment in order to achieve a sufficient power. The estimated power to detect any specific change for this pilot study was understandably low. In the comparison between aGvHD and non-GvHD patients using the immune cells $CD3^{+}CD4^{+}CD8\beta^{+}$, the analysis was estimated to have 29% power at 90% confidence level (Table 4.2). In the comparison between aGvHD & cGvHD and aGvHD only patients, using the immune cells $45RO^{+}CD3^{-}CD4^{dim}$, the analysis was estimated to have 50% power at 90% confidence level (Table 4.4).

Based on the present data, there was 68% chance of the recruited patients developing aGvHD and 13% chance of patients not affected aGvHD including early withdraws and fatality rate before 100 days post-transplant. This unbalanced number of aGvHD and non-GvHD patients could partially be the result of biased patient recruitments. Generally, patients with higher risks for disease are more inclined to enrol in studies [154]. Among the recruited aGvHD patients, there was 33% chance of developing cGvHD and 43% chance of being free of cGvHD including early withdraws and fatality rate. Overall, I estimate that 100 HSCT patients should result in 68 aGvHD and 13 non-GvHD cases; and 22 aGvHD & cGvHD and 29 aGvHD only cases. This will support an analysis with approximately 80% power at 90% confidence level for both patient group comparisons (Tables 4.2 and 4.4). This increased patient recruitment will also improve tolerance to the normality assumption in the FLDA. In addition, sample collection can be organized in a future study so the MNC concentration and immune cell proportions may be determined using the same sample in order to minimize errors in the concentration dataset. Another set of 100 HSCT patients would allow external validation (section 5.5.7).

5.5.3 Sampling rate

The third recommendation is to increase sampling rate immediately before and after BMT. In the present study, patients were sampled weekly. Multiple potential immune cell populations exhibiting discriminative patterns that may predict both aGvHD and cGvHD manifestations could be found following the BMT, between 7 and 21 days post-transplant (Tables H.1 and J.1).

The ideal sampling rate capturing immune cell population changes is daily. Flow cytometry is capable of capturing changes as small as 0.1% in the sample population [155]. In animal models, average daily turnover rates of T cells, B cells and NK cells under viral infections are 2, 3, and 3% [156]. The T cells' response to viral infection in mice can be detected in one to two days post-infection, reaching maximum by five to six days post-infection [157]. It may not be possible to establish a long-term rapid sampling rate for future studies. However, frequent sampling within the first two or three weeks of BMT, when patients might still be available in the hospital, may yield an informative dataset. The temporal analysis pipeline (Figure 2.1) requires a minimum of two samples per patient. The sampling rate can be non-uniform because of the robustness of the pipeline. Aside from the increased sampling rate around BMT, efforts should be made to obtain samples for the ends of the selected time range. Although the analysis pipeline was designed for clinical data with missing values and non-uniform sampling time, missing values still affected eligibility of the dataset to be included in the temporal analysis.

5.5.4 Additional markers

The fourth recommendation is to include markers specifically for the identification of the DP T cells and separation between host and donor origin immune cells. From the pilot study, I have found that immune cell populations $CD3^+CD4^+CD8\beta^+$ and $CD3^+CD4^+CD8\beta^+CD8^+$ exhibited a pattern of higher PBMC

proportion values and greater fluctuation from the aGvHD patients, when compare to the non-GvHD patients (Figures 4.6 and 4.7). Marker such as CD1a [134, 158] may be incorporated to distinguish thymocytes and mature T cells. Additional marker such as CD69, CD56, CD38, CD27, CD28, CD134, CXCR3, and CD62L will help to determine the exact origin and functional phenotype of the DP T cells and facilitate the efforts of validating current findings.

Additional experimental methods to separate immune cells of donor or recipient origins may also be necessary. The apparent DP T cell population was identified as a potential aGvHD marker from its pattern between 7 and 21 days post-transplant. During this time, the donor and the residual recipient immune cell chimerism has been documented in both human [129, 130] and mouse [159] models. Separation of immune cells' origin will aid in elucidating functions of T cells and T cell subset and their roles in the GvHD manifestation. Furthermore, the separation may also be useful in validating patterns of possible immune reconstruction (Figures 4.11 and 4.12).

5.5.5 Additional statistic tests

There are also three recommendations to improve the current analytical procedure (Figure 2.1). The first recommendation is the addition of statistical tests to the manual QA test and the FCM gating procedure. In the present study, outliers were identified from the QA test solely based on visual inspection. Conventional statistical tests such as analysis of variance and box plots to the current QA test may help to eliminate some biases. However, these tests are more efficient in identifying differences in distribution shifts instead of distribution shapes. Statistical tests such as the functional arbitrary covariance tests of shape [160] may be tested on its sensitivity to FCM QA testing using known samples (section 5.1.3) or simulated data. In this study, the FCM gating was performed manually by one or two -parameters

visualization with prior biological knowledge. These manual visual analyses were subjective and time consuming. Efforts have been made to improve gating efficiency and robustness. A recently developed feature-guided clustering algorithm [161] might be applicable in both QA and gating of high-throughput FCM dataset.

5.5.6 Graft versus host disease grades

The second recommendation for analytical improvement is the addition of GvHD grade in the analysis in order to accommodate GvHD misdiagnoses. At present, aGvHD and cGvHD diagnoses are ambiguous especially for mild forms of aGvHD (grade I) and cGvHD (limited). There are many reports on GvHD grading schemes [162-172] and their uncertain reproducibility [173, 174]. While the reproducibility might be remedied by a clinical algorithm [175], it will not decrease misdiagnoses.

Many previous aGvHD studies [29, 30] omitted patients diagnosed with grades I or II aGvHD from analyses so as to avoid interferences from misdiagnoses. This option was attempted for the present study resulting in similar or higher predictive powers from the top classifiers (Table 4.1). For future studies, I propose an addition of fuzzy clustering algorithm [176, 177] or mixture model based classification [178] to the temporal analysis pipeline in order to accommodate GvHD grades and misdiagnoses. It is important to predict not only the development of GvHD but also its severity. Many studies have suggested that due to the beneficial graft versus leukemia effects, only moderate or severe GvHD should be treated [154].

5.5.7 External validation

The third analytical recommendation is the implementation of external validation, which would only be possible if there are enough patients recruited to separate into a training dataset and a testing dataset. Two sets of 100 HSCT patients as the training and testing dataset for FLDA are recommended. Another set of 100 HSCT patients may be required for the multiparametric approached described below. Currently, LOOCV, which over-estimates classifier accuracy, is used to validate and rank FLDA classifiers without correction for multiple testing.

5.5.8 Multiparametric approach

The first tentative recommendation is an additional multiparametric analysis. Previous [81] and current studies have suggested a very complex GvHD manifestation. Presently, temporal classifiers from different subsets of immune cells were interpreted individually as there is no multiparametric temporal analysis algorithm available. However, preliminary results from Support Vector Machines (SVMs) analyses on the linear discriminant values from multiple FLDA classifiers indicated that predictive powers of these classifiers could be combined to achieve a better accuracy.

A SVM defines the best linear separating hyperplane between different classes of the training dataset projected into a high dimensional space. In the preliminary analysis, linear discriminant values from FLDA classifiers predicting the onset of aGvHD were obtained through resubstitution. Linear discriminant values representing weighted distances between the test data and the classifier, were then normalized to the range {-1, 1} using Equation 5.3 in preparation for the SVM analysis. Correlation-based feature selection method [179] was performed in Weka [180] to select a subset of the temporal classifiers by comparing the individual predictive power of the classifiers and the degree of redundancy between them. The

three top ranking classifiers selected based on LOOCV sensitivity and specificity were among the 11 classifiers selected using this feature selection method. Normalized linear discriminant values from these 11 classifiers were visually different between the aGvHD and non-GvHD patients (Figure 5.5). Individually, the best LOOCV estimated accuracy among these 11 classifiers was 86% sensitivity and 100% specificity. LOOCV estimated accuracy from the SVM [181] classifier of all 11 classifiers was 100% sensitivity and 100% specificity. Albeit resubstitution for the FLDA classifiers and LOOCV for the SVM classifier could result in a severe over-estimation in the final SVM's accuracy; the preliminary results suggest the applicability of SVM to combine predictive powers of multiple FLDA classifiers.

$$\hat{\alpha}'_X = \begin{cases} \hat{\alpha}_X < 0, \hat{\alpha}_X / \min(\hat{\alpha}) \\ \hat{\alpha}_X > 0, \hat{\alpha}_X / \max(\hat{\alpha}) \end{cases}$$

Equation 5.3 Normalization function for the linear discriminant values

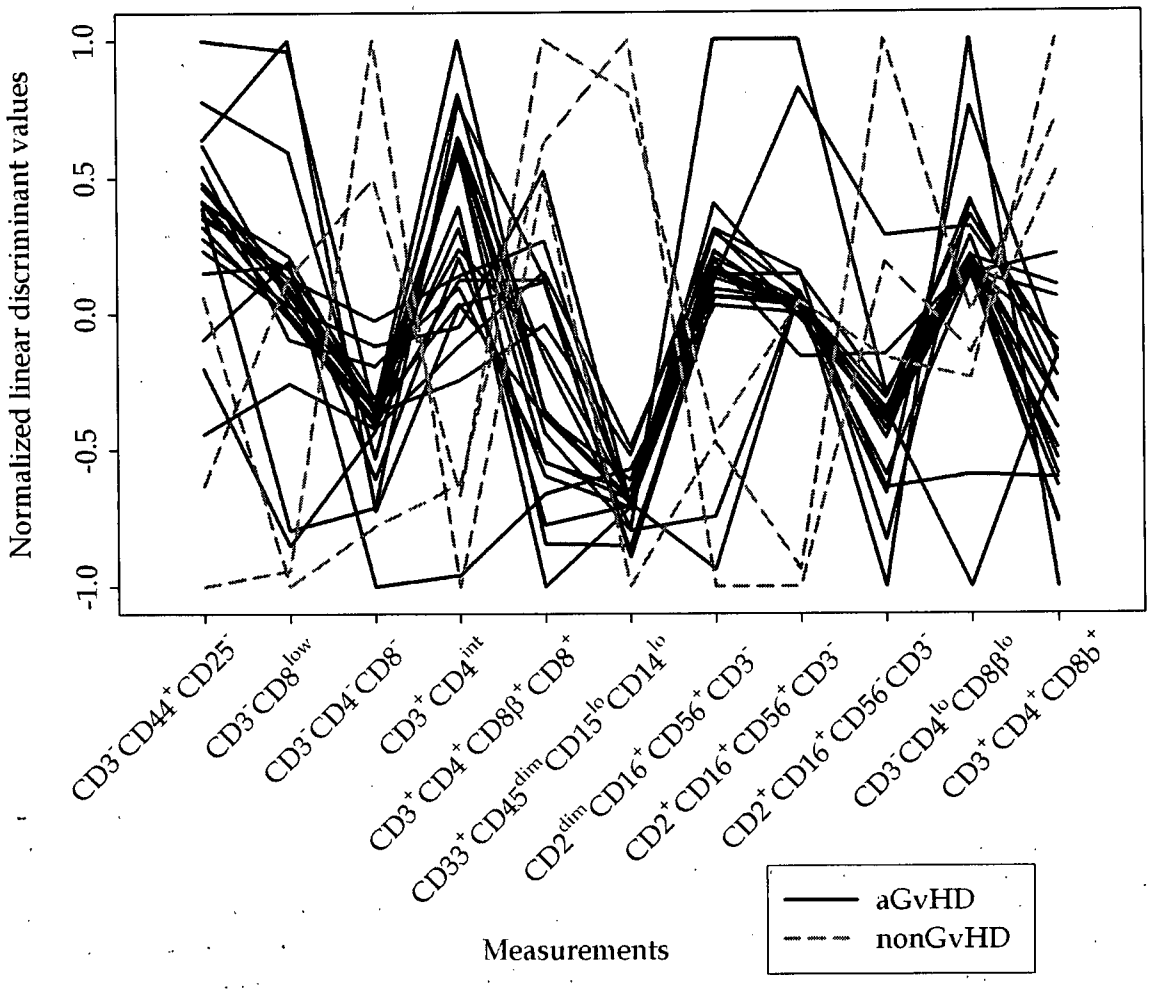


Figure 5.5 Parallel coordinates plot of the normalized linear discriminant values from the 11 FLDA classifiers selected via the correlation-based feature selection method.

5.5.9 Long time series analysis

The second tentative recommendation is an evaluation of additional long time series analyses. A long time series analysis would utilize the maximum amount of data and could be useful in elucidating the GvHD pathophysiologic mechanism occurring over time at different rates among patients. Also, most spline-

based methods including FLDA may perform more efficiently on long time series data [128]. Even though on average 15 weeks of data were available from each patient in the present dataset, I found the risk of desynchronization and needs for biased smoothing and registration procedures outweighed the benefits of a long time series analysis for this pilot study. However, for future studies, it might be possible to perform long time series analysis if detailed patient information such as GvHD progression can be incorporated into the registration procedure [48].

5.6 Conclusion

This pilot project achieved its objectives. The temporal analysis pipeline (Figure 2.1) was designed and implemented on the high-throughput clinical FCM data. Results of the QA test identify potential experiment errors. The screening of the current limited dataset by the temporal pipeline identified several potential aGvHD and cGvHD diagnosis markers including rare forms of T cells. In the present study, the most promising pattern was immune cells with $CD3^{+}CD4^{+}CD8\beta^{+}$ ($CD8^{+}$) phenotype which had higher proportion values and greater fluctuation from the aGvHD patients, compared to the non-GvHD patients (Figures 4.4 and 4.6). Multiple unknown immune cell subsets including $45RO^{+}CD3^{-}CD4^{dim}$ (Table 4.3) exhibited opposite FLDA estimated signal patterns (Figures 4.14 and 4.15) between the aGvHD & cGvHD and the aGvHD only patients.

While there was a high risk of false positives in the classification due to the limited number of available patients and errors from multiple testing, the current results demonstrated the applicability of the temporal analysis pipeline to the high-throughput clinical FCM data and the applicability of SVMs to combine multiple temporal classifiers' predictive powers. They also demonstrated the benefits of large scaled FCM study and temporal analysis. Large scale FCM study possibly combined with automatic gating process [161] would eliminate biases from prior knowledge

and could be very useful in elucidating GvHD. For instance, the DP T cells were never purposely included in other studies because they were not expected to exist based on the known T cell maturation mechanism (Figure 4.11).

Potential problems from the experimental and analytic procedures were identified and seven potential improvements recommended. They were:

1. Random plating
2. Increase patient recruitment, ideally two sets of 100 HSCT patients for training and testing purposes respectively
3. Increase sampling rate especially after the BMT procedure
4. Addition of markers targeting differentiation and function status of T cells
5. Addition of statistic tests to both the QA test and the FCM gating procedure to the existing visualization methods
6. Including GvHD grades in the temporal analysis in order to accommodate GvHD diagnosis errors
7. External validation for classifiers

As expected, none of the classifiers yielded significant correlation to the onset of either aGvHD or cGvHD. A future study made more efficient by these recommendations will be required to validate the current findings.

BIBLIOGRAPHY

1. Gilliam AC: **Update on Graft versus Host Disease.** *Progress in Dermatology* 2004, **123**:251-257.
2. Gasparetto M, Gentry T, Sebti S, O'Bryan E, Nimmanapalli R, Blaskovich MA, Bhalla K, Rizzieri D, Haaland P, Dunne J *et al*: **Identification of compounds that enhance the anti-lymphoma activity of rituximab using flow cytometric high-content screening.** *Journal of Immunological Methods* 2004, **292**(1-2):59.
3. Le Meur N, Rossini AJ, Gasparetto M, Smith C, Brinkman RR, Gentleman RC: **Quality Assessment of Ungated Flow Cytometry data in High Throughput experiments.** *Cytometry* 2007, **In press.**
4. James GM, Hastie TJ: **Functional linear discriminant analysis for irregularly sampled curves.** *Journal of the Royal Statistical Society Series B* 2001, **63**(3):533-550.
5. Reddy P: **Pathophysiology of acute graft-versus-host disease.** *Hematological Oncology* 2003, **21**:149-161.
6. Syrjala KL, Chapko MK, Vitaliano PP, Cummings C, Sullivan KM: **Recovery after allogeneic marrow transplantation: prospective study of predictors of long-term physical and psychosocial functioning.** *Bone Marrow Transplantation* 1993, **11**:319-327.
7. Duell T, van Lint MT, Ljungman P, Tichelli A, Socie G, Apperley J, Weiss M, Cohen A, Nekolla E, Kolb HJ: **Health and functional status of long-term survivors of bone marrow transplantation.** *Annals of Internal Medicine* 1997, **126**(3):184-192.
8. Mandy FF: **Twenty-five years of clinical flow cytometry: AIDS accelerated global instrument distribution.** *Cytometry* 2004, **58A**(1):55-56.
9. Orfao A, Ortuno F, de Santiago M, Lopez A, San Miguel J: **Immunophenotyping of acute leukemias and myelodysplastic syndromes.** *Cytometry* 2004, **58A**(1):62-71.
10. Braylan RC: **Impact of flow cytometry on the diagnosis and characterization of lymphomas, chronic lymphoproliferative disorders and plasma cell neoplasias.** *Cytometry* 2004, **58A**(1):57-61.
11. Keeney M, Gratama JW, Sutherland DR: **Critical role of flow cytometry in evaluating peripheral blood hematopoietic stem cell grafts.** *Cytometry* 2004, **58A**:72-75.
12. Sutherland HJ, Fyles GM, Adams G, Hao Y, Lipton JH, Minden MD, Meharchand JM, Atkins H, Tejpar I, Messner HA: **Quality of life following bone marrow transplantation: a comparison of patient reports with population norm.** *Bone Marrow Transplantation* 1997, **19**:1129-1136.
13. Socie G, Stone JV, Wingard JR, Weisdorf D, Henslee-Downey PJ, Bredeson C, Cahn J-Y, Passweg JR, Rowlings PA, Schouten HC *et al*: **Long-Term Survival and Late Deaths after Allogeneic Bone Marrow Transplantation.** *N Engl J Med* 1999, **341**(1):14-21.

14. Billingham RE: **The biology of graft-versus-host reactions**. *Harvey Lectures* 1966, **62**:21-78.
15. Goker H, Haznedaroglu IC, Chao NJ: **Acute graft-vs-host disease: pathobiology and management**. *Experimental Hematology* 2001, **29**:259-277.
16. Baron F, Storb R: **Allogeneic hematopoietic cell transplantation as treatment for hematological malignancies: a review**. *Springer Seminars in Immunopathology* 2004, **26**(1-2):71-94.
17. Couriel D, Caldera H, Champlin R, Komanduri K: **Acute graft-versus-host disease: pathophysiology, clinical manifestations, and management**. *Cancer* 2004, **101**(9):1936-1946.
18. Johnson ML, Farmer ER: **Graft-versus-host reactions in dermatology**. *Journal of the American Academy of Dermatology* 1998, **38**(3):369-392.
19. Klingemann HG, Storb R, Fefer A, Deeg HJ, Appelbaum FR, Buckner CD, Cheever MA, Greenberg PD, Stewart PS, Sullivan KM: **Bone marrow transplantation in patients aged 45 years and older**. *Blood* 1986, **67**:770-776.
20. Barrett AJ, Rezvani K, Solomon S, Dickinson AM, Wang XN, Stark G, Cullup H, Jarvis M, Middleton PG, Chao NJ: **New Developments in Allotransplant Immunology**. *Hematology* 2003:350-371.
21. Pavletic SZ, Carter SL, Kernan NA, Henslee-Downey J, Mendizabal AM, Papadopoulos E, Gingrich R, Casper J, Yanovich S, Weisdorf D: **Influrnece of T cell depletion on chronic Graft-Versus-Host Disease: results of a multi-center randomized trial in unrelated marrow donor transplantation**. *Blood* 2005, **106**(9):3308-3313.
22. Ichiki Y, Bowlus CL, Shimoda S, Ishibashi H, Vierling JM, Gershwin ME: **T cell immunity and graft-versus-host disease (GVHD)**. *Autoimmunity Reviews* 2006, **5**:1-9.
23. Kernan NA, Bartsch G, Ash RC, Beatty PG, Champlin R, Filipovich A, Gajewski J, Hansen JA, Henslee-Downey J, McCullough J *et al*: **Analysis of 462 Transplantations from Unrelated Donors Facilitated by the National Marrow Donor Program**. *N Engl J Med* 1993, **328**(9):593-602.
24. Marmont AM, Horowitz MM, Gale RP, Sobocinski K, Ash RC, van Bekkum DW, Champlin RE, Dicke KA, Goldman JM, Good RA: **T-cell depletion of HLA-identical transplants in leukemia**. *Blood* 1991, **78**(8):2120-2130.
25. Bacigalupo A, Lamparelli T, Bruzzi P, Guidi S, Alessandrino PE, di Bartolomeo P, Oneto R, Bruno B, Barbanti M, Sacchi N *et al*: **Antithymocyte globulin for graft-versus-host disease prophylaxis in transplants from unrelated donors: 2 randomized studies from Gruppo Italiano Trapianti Midollo Osseo (GITMO)**. *Blood* 2001, **98**(10):2942-2947.
26. Hale G, Zhang M-J, Bunjes D, Prentice HG, Spence D, Horowitz MM, Barrett AJ, Waldmann H: **Improving the Outcome of Bone Marrow Transplantation by Using CD52 Monoclonal Antibodies to Prevent Graft-Versus-Host Disease and Graft Rejection**. *Blood* 1998, **92**(12):4581-4590.

27. Baron F, Maris MB, Storer BE, Sandmaier BM, Panse JP, Chauncey TR, Sorror M, Little M-T, Maloney DG, Storb R *et al*: **High doses of transplanted CD34+ cells are associated with rapid T-cell engraftment and lessened risk of graft rejection, but not more graft-versus-host disease after nonmyeloablative conditioning and unrelated hematopoietic cell transplantation.** *Leukemia* 2005.
28. Storb R, Prentice R, Buckner CD, Clift RA, Appelbaum FR, Deeg J, Doney K, Hansen JA, Mason M, Sanders J *et al*: **Graft-versus-host disease and survival in patients with aplastic anemia treated by marrow grafts from HLA-identical siblings. Beneficial effect of a protective environment.** *N Engl J Med* 1983, **308**(6):302-307.
29. Paz Morante M, Briones J, Canto E, Sabzevari H, Martino R, Sierra J, Rodriguez-Sanchez JL, Vidal S: **Activation-associated phenotype of CD3+ T cells in acute graft-versus-host disease**. *Clinical and Experimental Immunology* 2006, **145**(1):36-43.
30. Jaksch M, Uzunel M, Remberger M, Sundberg B, Mattsson J: **Molecular monitoring of T-cell chimerism early after allogeneic stem cell transplantation may predict the occurrence of acute GVHD grades II-IV**. *Clinical Transplantation* 2005, **19**(3):346-349.
31. Ferrara J, Guillen FJ, van Dijken PJ, Marion A, Murphy GF, Burakoff SJ: **Evidence that large granular lymphocytes of donor origin mediate acute graft-versus-host disease**. *Transplantation* 1989, **47**(1):50-54.
32. Filep JG, Baron C, Lachance S, Perreault C, Chan JS: **Involvement of nitric oxide in target-cell lysis and DNA fragmentation induced by murine natural killer cells**. *Blood* 1996, **87**(12):5136-5143.
33. Asai O, Longo DL, Tian ZG, Hornung RL, Taub DD, Ruscetti FW, Murphy WJ: **Suppression of graft-versus-host disease and amplification of graft-versus-tumor effects by activated natural killer cells after allogeneic bone marrow transplantation**. *Journal of Clinical Investigation* 1998, **101**(9):1835-1842.
34. Klingemann HG: **Relevance and potential of natural killer cells in stem cell transplantation**. *Biology of Blood and Marrow Transplantation* 2000, **6**(2):90-99.
35. Vargas-Diez E, Garcia-Diez A, Marin A, Fernandez-Herrera J: **Life-threatening graft-vs-host disease**. *Clinics in Dermatology* 2005, **23**(3):285-300.
36. Kansu E: **The Pathophysiology of Chronic Graft-versus-Host disease**. *International Journal of Hematology* 2004, **79**(3):209-215.
37. Iwasaki T: **Recent Advances in the Treatment of Graft-Versus-Host Disease**. *Clinical medicine and research* 2004, **2**(4):243-252.
38. Lee SJ, Klein JP, Barrett AJ, Ringden O, Antin JH, Cahn J-Y, Carabasi MH, Gale RP, Giralt S, Hale GA *et al*: **Severity of chronic graft-versus-host disease: association with treatment-related mortality and relapse**. *Blood* 2002, **100**:406-452.
39. Higman MA, Vogelsang GB: **Chronic graft versus host disease**. *British Journal of Haematology* 2004, **125**(4):435-454.

40. Hale G, Jacobs P, Wood L, Fibbe WE, Barge R, Novitzky N, Toit C, Abrahams L, Thomas V, Bunjes et a: **CD52 antibodies for prevention of graft-versus-host disease and graft rejection following transplantation of allogeneic peripheral blood stem cells**. *Bone Marrow Transplantation* 2000, **26**(1):69-76.
41. Komatsuda M: **Changes of lymphocyte subsets in leukemia patients who received allogenic bone marrow transplantation**. *Acta medica Okayama* 1991, **45**(4):257-265.
42. Remberger M, Ringden O, Blau I-W, Ottinger H, Kremens B, Kiehl MG, Aschan J, Beelen DW, Basara N, Kumlien G *et al*: **No difference in graft-versus-host disease, relapse, and survival comparing peripheral stem cells to bone marrow using unrelated donors**. *Blood* 2001, **98**(6):1739-1745.
43. Zaucha JM, Gooley T, Bensinger WI, Heimfeld S, Chauncey TR, Zaucha R, Martin PJ, Flowers MED, Storek J, Georges G *et al*: **CD34 cell dose in granulocyte colony-stimulating factor-mobilized peripheral blood mononuclear cell grafts affects engraftment kinetics and development of extensive chronic graft-versus-host disease after human leukocyte antigen-identical sibling transplantation**. *Blood* 2001, **98**(12):3221-3227.
44. Mohty M, Bilger K, Jourdan E, Kuentz M, Michallet M, Bourhis JH, Milpied N, Sutton L, Jouet JP, Attal M *et al*: **Higher doses of CD34+ peripheral blood stem cells are associated with increased mortality from chronic graft-versus-host disease after allogeneic HLA-identical sibling transplantation**. *Leukemia* 2003, **17**(5):869-875.
45. Perez-Simon JA, Diez-Campelo M, Martino R, Sureda A, Caballero D, Canizo C, Brunet S, Altes A, Vazquez L, Sierra J *et al*: **Impact of CD34+ cell dose on the outcome of patients undergoing reduced-intensity-conditioning allogeneic peripheral blood stem cell transplantation**. *Blood* 2003, **102**(3):1108-1113.
46. Bar-Joseph Z: **Analyzing time series gene expression data**. *Bioinformatics* 2004, **20**(16):2493-2503.
47. Bay SD, Chrisman L, Pohorille A, Shrager J: **Temporal aggregation bias and inference of casual regulatory networks**. *Journal of Computational Biology* 2004, **11**(5):971-985.
48. Ramsay JO, Silverman BW: **Functional data analysis**, Second edn. New York: Springer; 2005.
49. Cuevas A, Febrero M, Fraiman R: **An anova test for functional data**. *Computational statistics and data analysis* 2004, **47**:111-122.
50. Park T, Yi S-G, Lee S, Lee SY, Yoo D-H, Ahn J-I, Lee Y-S: **Statistical tests for identifying differentially expressed genes in time-course microarray experiments**. *Bioinformatics* 2003, **19**(6):694-703.
51. Yao F, Muller H-G, Wang J-L: **Functional Data Analysis for Sparse Longitudinal Data**. *Journal of the American Statistical Association* 2005, **100**(470):577-590.

52. Liu X, Muller H-G: **Modes and clustering for time-warped gene expression profile data.** *Bioinformatics* 2003, **19**(15):1937-1944.
53. Balasubramaniyan R, Hullermeier E, Weskamp N, Kamper J: **Clustering of gene expression data using a local shape-based similarity measure.** *Bioinformatics* 2005, **21**(7):1069-1077.
54. Ernst J, Nau GJ, Bar-Joseph Z: **Clustering short time series gene expression data.** *Bioinformatics* 2005, **21**(Suppl 1):i159-i168.
55. Liu H, Tarima S, Borders AS, Getchell TV, Getchell ML, Stromberg AJ: **Quadratic regression analysis for gene discovery and pattern recognition for non-cyclic short time-course microarray experiments.** *Bioinformatics* 2005, **6**(1):106-122.
56. Bar-Joseph Z, Gerber GK, Gifford DK, Jaakkola TS, Simon I: **Continuous Representations of Time-Series Gene Expression Data.** *Journal of Computational Biology* 2003, **10**(3-4):341-356.
57. Ben-Dor A, Shamir R, Yakhini Z: **Clustering gene expression patterns.** *Journal of Computational Biology* 1999, **6**(3/4):281-297.
58. Azuaje F: **Clustering-based approaches to discovering and visuzlising microarray data patterns.** *Briefing in Bioinformatics* 2003, **4**(1):31-42.
59. Luan Y, Li H: **Clustering of time-course gene expression data using a mixed-effects model with B-splines.** *Bioinformatics* 2003, **19**(4):474-482.
60. Kehagias A, Petridis V: **Predictive Modular Neural Networks for Time Series Classification.** *Neural networks* 1997, **10**(1):31-49.
61. Mendez MA, Hodar C, Vulpe C, Gonzalez M, Cambiazo V: **Discriminant analysis to evaluate clustering of gene expression data.** *FEBS letters* 2002, **522**:24-28.
62. Hall P, Poskitt DS, Presnell B: **A Functional Data - Analytic Approach to Signal Discrimination.** *Technometrics* 2001, **43**(1):1-9.
63. Muller H-G: **Functional Modelling and Classification of Longitudinal Data.** *Scandinavian Journal of Statistics* 2005, **32**(2):223-240.
64. deBoor C: **A Practical Guide to Splines**, Revised Edition edn. New York: Springer; 2001.
65. Hastie T, Tibshirani R, Friedman J: **The Elements of Statistical Learning.** New York: Springer; 2001.
66. Ramsay JO, Li X: **Curve registration.** *Journal of the Royal Statistical Society Series B* 1998, **60**(2):351-363.
67. Raychaudhuri S, Stuart JM, Altman RB: **Principal components analysis to summarize microarray experiments: application to sporulation time series.** In: *Pacific Symposium on Biocomputing: 2000; Singapore*: World Scientific; 2000: 455-466.
68. Li KC, Yan M, Yuan SS: **A simple statistical model for depicting the cdc15-synchronized yeast cell-cycle regulated gene expression data.** *Statistica Sinica* 2002, **12**(1):141-158.

69. Alter O, Brown PO, Botstein D: **Singular value decomposition for genome-wide expression data processing and modeling.** *PNAS* 2000, **97**(18):10101-10106.
70. Kruglyak S, Tang H: **A New Estimator of Significance of Correlation in Time Series Data.** *Journal of Computational Biology* 2001, **8**(5):463-470.
71. Brown MPS, Grundy WN, Lin D, Cristianini N, Sugnet CW, Furey TS, Ares MJ, Haussler D: **Knowledge-based analysis of microarray gene expression data by using support vector machines.** *PNAS* 2000, **97**(1):262-267.
72. Ferraty F, Vieu P: **Curves discrimination: a nonparametric functional approach.** *Computational statistics and data analysis* 2003, **44**:161-173.
73. Gui J, Li H: **Mixture Functional Discriminant Analysis for Gene Function Classification Based on Time Course Gene Expression Data.** In: *Joint Statistical Meetings: 2003; San Francisco, California*; 2003.
74. Ripley BD: **Pattern Recognition and Neural Networks**: Cambridge University Press; 1996.
75. Braga-Neto UM, Hashimoto R, Dougherty ER, Nguyen DV, Carroll RJ: **Is cross-validation better than resubstitution for ranking genes?** *Bioinformatics* 2004, **20**(2):253-258.
76. Braga-Neto UM, Dougherty ER: **Is cross-validation valid for small-sample microarray classification?** *Bioinformatics* 2004, **20**(3):374-380.
77. van Belle G, Fisher LD, Heagerty PJ, Lumley T: **Biostatistics: A Methodology for the Health Sciences**, 2 edn. New Jersey: Wiley; 2004.
78. Collings BJ, Hamilton MA: **Estimating the power of the two-sample wilcoxon test for location shift.** *Biometrics* 1988, **44**:847-860.
79. Johnston-Wilson NL, Bouton CM, Pevsner J, Breen JJ, Torrey EE, Yolken RH: **Emerging technologies for large-scale screening of human tissues and fluids in the study of severe psychiatric disease.** *International Journal of Neuropsychopharmacology* 2001, **4**(1):83-92.
80. Alizadeh AA, Eisen MB, Davis RE, Ma C, Lossos IS, Rosenwald A, Boldrick JC, Sabet H, Tran T, Yu X *et al*: **Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling.** *Nature* 2000, **403**(6769):503-511.
81. Jaksch M, Mattsson J: **The Pathophysiology of Acute Graft-Versus-Host Disease.** *Scandinavian Journal of Immunology* 2005, **61**(5):398-409.
82. Gratwohl A, Hermans J, Apperley J, Arcese W, Bacigalupo A, Bandini G, di Bartolomeo P, Boogaerts M, Bosi A, Carreras E: **Acute graft-versus-host disease: grade and outcome in patients with chronic myelogenous leukemia. Working Party Chronic Leukemia of the European Group for Blood and Marrow Transplantation.** *Blood* 1995, **86**(2):813-818.
83. Crawford K, Stark A, Kitchens B, Sternheim K, Pantazopoulos V, Triantafellow E, Wang Z, Vasir B, Larsen CE, Gabuzda D *et al*: **CD2 engagement induces dendritic cell activation: implications for immune surveillance and T-cell activation.** *Blood* 2003, **102**(5):1745-1752.

84. Sykes M, Abraham VS: **The mechanism of IL-2-mediated protection against GVHD in mice. II. protection occurs independently of NK/LAK cells.** *Transplantation* 1992, **53**(5):1063-1070.
85. Chao NJ: **Graft-versus-host disease: the viewpoint from the donor T cell.** *Biology of Blood and Marrow Transplantation* 1997, **3**(1):1-10.
86. Miceli MC, von Hoegan P, Parnes JR: **Adhesion versus coreceptor function of CD4 and CD8: role of the cytoplasmic tail in coreceptor activity.** *PNAS* 1991, **88**(7):2623-2627.
87. Seong RH, Chamberlain JW, Parnes JR: **Signal for T-cell differentiation to a CD4 cell lineage is delivered by CD4 transmembrane region and/or cytoplasmic tail.** *Nature* 1992, **356**(6371):718-720.
88. Jones NH, Clabby ML, Dialynas DP, Huang HJ, Herzenberg LA, Strominger JL: **Isolation of complementary DNA clones encoding the human lymphocyte glycoprotein T1/Leu-1.** *Nature* 1986, **323**(6086):346-349.
89. van de Velde H, von hoegan I, Luo W, Parnes JR, Thielemans K: **The B-cell surface protein CD72/Lyb-2 is the ligand for CD5.** *Nature* 1991, **351**(6328):662-665.
90. Allam A, Illges H: **Calyculin A inhibits expression of CD8alpha but not CD4 in human peripheral blood T cells.** *Immunobiology* 2000, **202**(4):353-362.
91. Lin RS, Rodriguez C, Veillette A, Lodish HF: **Zinc is essential for binding of p56(lck) to CD4 and CD8alpha.** *Journal of Biological Chemistry* 1998, **473**(49):32878-32882.
92. Correale P, Tagliaferri P, Camera A, Caraglia M, Del Vecchio L, De Laurentis M, Pinto A, Rotoli B, Bianco AR: **CD10/common acute lymphoblastic leukemia-associated antigen and adhesion factor expression is predictive for lymphokine-activated killing sensitivity of adult B-lineage acute lymphoblastic leukemia.** *The Year in Immunology* 1993, **7**:90-95.
93. Gupta D, Kirkland TN, Viriyakosol S, Dziarski R: **CD14 is a cell-activating receptor for bacterial peptidoglycan.** *Journal of Biological Chemistry* 1996, **271**(38):23310-23316.
94. Kerr MA, Stocks SC: **The role of CD15-(Le(X))-related carbohydrates in neutrophil adhesion.** *Histochem J* 1992, **24**(11):811-826.
95. Anderson P, Calgiuri M, O'Brien C, Manley T, Ritz J, Schlossman SF: **Fc gamma receptor type III (CD16) is included in the zeta NK receptor complex expressed by human natural killer cells.** *PNAS* 1990, **87**(6):2274-2278.
96. Cruse JM, Lewis RE: **Atlas of Immunology**. Boca Raton: CRC press; 1999.
97. Shirakawa T, Li A, Dubowitz M, Dekker JW, Shaw AE, Faux JA, Ra C, Cookson WO, Hopkin JM: **Association between atopy and variants of the beta subunit of the high-affinity immunoglobulin E receptor.** *Nature Genetics* 1994, **7**(2):125-129.

98. Wilson GL, Najfeld V, Kozlow E, Menniger J, Ward D, Kehrl JH: **Genomic structure and chromosomal mapping of the human CD22 gene**. *Journal of Immunology* 1993, **150**(11):5013-5024.
99. Crocker PR, Mucklow S, Bouckson V, McWilliam A, Willis AC, Gordon S, Milon G, Kelm S, Bradfield P: **Sialoadhesin, a macrophage sialic acid binding receptor for haemopoietic cells with 17 immunoglobulin-like domains**. *the EMBO journal* 1994, **13**(19):4490-4503.
100. Vitale C, Romagnani C, Falco M, Ponte M, Vitale M, Moretta A, Bacigalupo A, Moretta L, Mingari MC: **Engagement of p75/AIRM1 or CD33 inhibits the proliferation of normal or leukemic myeloid cells**. *PNAS* 1999, **96**(26):15091-15096.
101. Shtivelman E, Bishop JM: **Expression of CD44 is repressed in neuroblastoma cells**. *Molecular and Cellular Biology* 1991, **11**(11):5446-5453.
102. Juretic E, Gagro A, Vukelic V, Petrovecki M: **Maternal and neonatal lymphocyte subpopulations at delivery and 3 days postpartum: increased coexpression of CD45 isoforms**. *American Journal of Reproductive Immunology* 2004, **52**(1):1-7.
103. Matto M, Nuutinen UM, Ropponen A, Myllykangas K, Pelkonen J: **CD45RA and RO Isoforms Have Distinct Effects on Cytokine- and B-Cell-Receptor-Mediated Signalling in Human B cells**. *Scandinavian Journal of Immunology* 2005, **61**(6):520-528.
104. Posselt AM, Vincenti F, Bedolli M, Lantz M, Roberts JP, Hirose R: **CD69 expression on peripheral CD8 T cells correlates with acute rejection in renal transplant recipients**. *Transplantation* 2003, **76**(1):190-195.
105. Stuber E, Neurath M, Calderhead D, Fell HP, Stober W: **Cross-linking of OX40 ligand, a member of the TNF/NGF cytokine family, induces proliferation and differentiation in murine splenic B cells**. *Immunity* 1995, **2**(5):507-521.
106. Pieters RHH, Punt P, Bol M, van Dijken JM: **The thymus atrophy inducing organotin compound DBTC stimulates TCRab-CD3 signaling in immature rat thymocytes**. *Biochem Biophys Res Commun* 1995, **214**(552-558).
107. Rossini AJ, Wan JY, Moodie Z: **rflowcyt: Statistical tools and data structures for analytic flow cytometry**. In., 1.0.1 edn; 2005: R package.
108. Huang E, West M, Nevins JR: **Gene expression profiling for prediction of clinical characteristics of breast cancer**. *Recent progress in hormone research* 2003, **58**:55-73.
109. Cheok MH, Yang W, Pui C-H, Downing JR, Cheng C, Naeve CW, Relling MV, Evans WE: **Treatment specific changes in gene expression discriminate in vivo drug response in human leukemia cells**. *Nature Genetics* 2003, **34**(1):85-90.
110. Miura Y, Thoburn CJ, Bright EC, Arai S, Hess AD: **Regulation of OX40 gene expression in graft-versus-host disease**. *Transplantation Proceedings* 2005, **37**(1):57-61.

111. Imamura M, Tsutsumi Y, Miura Y, Toubai T, Tanaka J: **Immune Reconstitution and Tolerance after Allogeneic Hematopoietic Stem Cell Transplantation**. *Hematology* 2003, **8**(1):19-26.
112. Ju XP, Xu B, Xiao ZP, Li JY, Chen L, Lu SQ, Huang ZX: **Cytokine expression during acute graft-versus-host disease after allogeneic peripheral stem cell transplantation**. *Bone Marrow Transplantation* 2005.
113. Ichiba T, Teshima T, Kuick R, Misek DE, Liu C, Takada Y, Maeda Y, Reddy P, Williams DL, Hanash SM *et al*: **Early changes in gene expression profiles of hepatic GVHD uncovered by oligonucleotide microarrays**. *Blood* 2003, **102**(2):763-771.
114. Bradley LM, Dalton DK, Croft M: **A direct role of IFN-gamma in regulation of Th1 cell development**. *J Immunol* 1996, **157**(4):1350-1358.
115. Maggie Merchant SRW: **Recent advancements in surface-enhanced laser desorption/ionization-time of flight-mass spectrometry**. *Electrophoresis* 2000, **21**(6):1164-1177.
116. Kaiser T, Hermann A, Kielstein JT, Wittke S, Bartel S, Krebs R, Hausadel F, Hillmann M, Golovko I, Koester P: **Capillary electrophoresis coupled to mass spectrometry to establish polypeptide patterns in dialysis fluids**. *Journal of Chromatography A* 2003, **1013**(1-2):157-171.
117. Srinivasan R, Daniels J, Fusaro V, Lundqvist A, Killian JK, Geho D, Quezado M, Kleiner D, Rucker S, Espina V: **Accurate diagnosis of acute graft-versus-host disease using serum proteomic pattern analysis**. *Experimental Hematology* 2006, **34**(6):796-801.
118. Kaiser T, Kamal H, Rank A, Kolb H-J, Holler E, Ganser A, Hertenstein B, Mischak H, Weissinger EM: **Proteomics applied to the clinical follow-up of patients after allogeneic hematopoietic stem cell transplantation**. *Blood* 2004, **104**(2):340-349.
119. Brazma A: **On the Importance of Standardisation in Life Sciences**. *Bioinformatics* 2001, **17**(2):113-114.
120. Chicurel M: **Bioinformatics: Bringing it all together technology feature**. *Nature* 2002, **419**(6908):751-757.
121. Boguski MS, McIntosh MW: **Biomedical informatics for proteomics**. *Nature* 2003, **422**(6928):233-237.
122. **Introductory Core Operation Course** [http://www.med.umich.edu/flowcytometry/InitialTraining/index.htm]
123. Lanier LL, Philips JP, Philips JH: **Correlation of cell surface antigen expression on human thymocytes by multi-color flow cytometric analysis: implications for differentiation**. *Journal of Immunology* 1986, **137**(8):2501-2507.
124. Dudoit S, Yang YH: **Bioconductor R packages for exploratory analysis and normalization of cDNA microarray data**. In: *The Analysis of Gene Expression Data: Methods and Software*. Edited by Parmigiani G, Garrett ES, Irizarry R, Zeger SL. New York: Springer; 2002.

125. Jacobsohn DA, Montross S, Anders V, Vogelsang GB: **Clinical importance of confirming or excluding the diagnosis of chronic graft-versus-host disease.** *Bone Marrow Transplantation* 2001, **28**(11):1047-1051.
126. Culhane AC, Perriere G, Considine EC, Cotter TG, Higgins DG: **Between-group analysis of microarray data.** *Bioinformatics* 2002, **18**(12):1600-1608.
127. **Hierarchical Ordered Partitioning and Collapsing Hybrid (HOPACH)** [http://www.stat.berkeley.edu/~laan/]
128. Bar-Joseph Z, Gerber G, Simon I, Gifford DK, Jaakkola TS: **Comparing the continuous representation of time-series expression profiles to identify differentially expressed genes.** *PNAS* 2003, **100**(18):10146-10151.
129. Butturini A, Seeger RC, Gale RP: **Recipient immune-competent T lymphocytes can survive intensive conditioning for bone marrow transplantation.** *Blood* 1986, **68**(4):954-956.
130. Bertheas MF, Lafage M, Levy P, Blaise D, Stoppa AM, Viens P, Mannoni P, Maraninchi D: **Influence of mixed chimerism on the results of allogeneic bone marrow transplantation for leukemia.** *Blood* 1991, **78**(11):3103-3106.
131. Zuckermann FA: **Extrathymic CD4/CD8 double positive T cells.** *Veterinary Immunology and Immunopathology* 1999, **72**:55-66.
132. Kelly K, Pilarski L, Shortman K, Scollay R: **CD4+ CD8+ cells are rare among in vitro activated mouse or human T lymphocytes.** *Cellular immunology* 1988, **2**:414-424.
133. Abbas AK, Lichtman AH: **Cellular and Molecular Immunology**, 5 edn. Philadelphia, PA: Saunders; 2003.
134. Blue ML, Daley JF, Levine H, Schlossman SF: **Coexpression of T4 and T8 on peripheral blood T cells demonstrated by two-color florescence flow cytometry.** *Journal of Immunology* 1985, **134**:2281-2286.
135. Ortolani C, Forti E, Radin E, Cibin R, Cossarizza A: **Cytofluorometric identification of two populations of double positive (CD4+, CD8+) T lymphocytes in human peripheral blood.** *Biochem Biophys Res Commun* 1993, **191**(2):601-609.
136. Kay NE, Bone N, Hupke M, Dalmasso AP: **Expansion of a Lymphocyte Population Co-Expressing T4 (CD4) and T8 (CD8) Antigens in the Peripheral Blood of a Normal Adult Male.** *Blood* 1990, **75**(10):2024-2029.
137. Patel SS, Wacholtz MC, Duby AD, Thiele DL, Lipsky PE: **Analysis of the functional capabilities of CD3+CD4-CD8- and CD3+CD4+CD8+ human T cell clones.** *Journal of Immunology* 1989, **143**:1108-1117.
138. Pawelec G, Adibzadeh M, Pohla H, Schaudt K: **Immunosenescence: Aging of the immune system.** *Immunology Today* 1995, **16**(9):420-422.
139. Colombatti A, Doliana R, Schiappacassi M, Argentini C, Tonutti E, Feruglio C, Sala P: **Age related persistent clonal expansions of CD28(-) cells: phenotypic and molecular TCR analysis reveals both CD4(+) and CD4(+)CD8(+) cells with identical CDR3 sequences.** *Clinical Immunology and Immunopathology* 1998, **89**(1):61-70.

140. Weiss L, Roux A, Garcia S, Demouchy C, Haeffner-Cavaillon N, Kazatchkine MD, Gougeon ML: **Persistent expansion, in a human immunodeficiency virus-infected person, of V beta-restricted CD4+CD8+ T lymphocytes that express cytotoxicity-associated molecules and are committed to produce interferon-gamma and tumor necrosis factor-alpha.** *The Journal of Infectious Diseases* 1998, **178**(4):1158-1162.
141. Blue ML, Daley JF, Levine H, Craig K, Scholssman SF: **Biosynthesis and surface expression of T8 by peripheral blood T4+ cells in vitro.** *Journal of Immunology* 1986, **137**:1202-1207.
142. Paliard X, de Waal Malefijt R, de Vries JE, Spits H: **Interleukin-4 mediates CD8 induction on human CD4+ T cell-clones.** *Nature* 1988, **335**:642-644.
143. Brod SA, Purvee M, Benjamin D, Hafler DA: **Frequency analysis of CD4-CD8+ T cells cloned with IL-4.** *Cellular immunology* 1990, **125**:426-436.
144. Jimenez E, Sacedon R, Vicente A, Hernandez-Lopez C, Zapata AG, Varas A: **Rat Peripheral CD4+CD8+ T Lymphocytes Are Partially Immunocompetent Thymus-Derived Cells That Undergo Post-Thymic Maturation to Become Functionally Mature CD4+ T Lymphocytes.** *J Immunol* 2002, **168**(10):5005-5013.
145. Nascimbeni M, Shin E-C, Chiriboga L, Kleiner DE, Rehermann B: **Peripheral CD4+CD8+ T cells are differentiated effector memory cells with antiviral functions.** *Blood* 2004, **104**(2):478-486.
146. Prince HE, Golding J, York J: **Characterization of circulating CD4+ CD8+ Lymphocytes in Healthy Individuals Prompted by Identification of a Blood Donor with a Markedly Elevated Level of CD4+ CD8+ Lymphocytes.** *Clinical and Diagnostic Laboratory Immunology* 1994, **1**(5):597-605.
147. Tonutti E, Sala P, Feruglio C, Yin Z, Colombatti A: **Phenotypic Heterogeneity of Persistent Expansions of CD4+CD8+ T Cells.** *Clinical Immunology and Immunopathology* 1994, **73**(3):312-320.
148. Barclay AN, Birkeland ML, Brown MH, Beyers AD, Davis SJ, Somoza C, Williams AF: **The Leucocyte Antigen FactsBook.** London: Academic Press Limited; 1993.
149. Farag SS, Caligiuri MA: **Human natural killer cell development and biology.** *Blood Reviews* 2006, **20**(3):123-137.
150. Cooper MA, Fehniger TA, Caligiuri MA: **The biology of human natural killer-cell subsets.** *Trends in Immunology* 2001, **22**(11):633-640.
151. Mavilio D, Lombardo G, Benjamin J, Kim D, Follman D, Marcenaro E, O'Shea MA, Kinter A, Kovacs C, Moretta A *et al*: **Characterization of CD56-/CD16+ natural killer (NK) cells: A highly dysfunctional NK subset expanded in HIV-infected viremic individuals.** *PNAS* 2005, **102**(8):2886-2891.
152. Wood GS, Szwejbka P, Schwandt A: **Human Langerhans Cells Express a Novel Form of the Leukocyte Common Antigen (CD45).** 1998, **111**(4):668-673.
153. Bittner ML, Butow R, DeRisi J, Diehn M, Eberwine J, Epstein CB, Glynne R, Grimmond S, Ideker T, Kacharmina JE *et al*: **Expression Analysis of RNA.** In:

DNA Microarrays: A Molecular Cloning Manual. Edited by Bowtell D, Sambrook J, vol. 1st. New York: Cold Spring Harbor Laboratory Press; 2003: 102-288.

154. Martin PJ, Nash RA: **Pitfalls in the Design of Clinical Trials for Prevention or Treatment of Acute Graft-versus-Host Disease**. *Biology of Blood and Marrow Transplantation* 2006, **12**(1, Supplement 2):31-36.
155. Owen RG, Rawstron AC: **Minimal residual disease monitoring in multiple myeloma: flow cytometry is the method of choice**. *British Journal of Haematology* 2005, **128**(5):732-733.
156. De Boer RJ, Mohri H, Ho DD, Perelson AS: **Turnover Rates of B Cells, T Cells, and NK Cells in Simian Immunodeficiency Virus-Infected and Uninfected Rhesus Macaques**. *J Immunol* 2003, **170**(5):2479-2487.
157. De Boer RJ, Oprea M, Antia R, Murali-Krishna K, Ahmed R, Perelson AS: **Recruitment Times, Proliferation, and Apoptosis Rates during the CD8+ T-Cell Response to Lymphocytic Choriomeningitis Virus**. *J Virol* 2001, **75**(22):10663-10669.
158. Schlossman SF, Boumsell L, Gilks W, Harlan JM, Kishimoto T: **Leucocyte Typing V**. New York, U.S.A.: Oxford University Press; 1995.
159. Choi EY, Christianson GJ, Yoshimura Y, Jung N, Sproule TJ, Malarkannan S, Joyce S, Roopenian DC: **Real-time T-cell profiling identifies H60 as a major minor histocompatibility antigen in murine graft-versus-host disease**. *Blood* 2002, **100**(13):4259-4264.
160. James GM, Sood A: **Performing Hypothesis Tests on the Shape of Functional Data**. *Computational statistics and data analysis* 2006, **50**(7):1774-1792.
161. Zeng QT, Pratt JP, Pak J, Ravnic D, Huss H, Mentzer SJ: **Feature-guided clustering of multi-dimensional flow cytometry datasets**. *Journal of Biomedical Informatics*, **In Press, Corrected Proof**.
162. Glucksberg H, Storb R, Fefer A, Buckner CD, Neiman PE, Clift RA, Lerner KG, Thomas ED: **Clinical manifestations of graft-versus-host disease in human recipients of marrow from HL-A-matched sibling donors**. *Transplantation* 1974, **18**(4):295-304.
163. Przepiorka D, Weisdorf D, Martin P, Klingemann HG, Beatty P, Hows J, Thomas ED: **1994 Consensus Conference on Acute GVHD Grading**. *Bone Marrow Transplantation* 1995, **15**(6):825-828.
164. Rowlings PA, Przepiorka D, Klein JP, Gale RP, Passweg JR, Henslee-Downey J, Cahn J-Y, Calderwood S, Gratwohl A, Socie G *et al*: **IBMTR Severity Index for grading acute graft-versus-host disease: retrospective comparison with Glucksberg grade**. *British Journal of Haematology* 1997, **97**(4):855-864.
165. Lerner KG, Kao CM, Storb R, Buckner CD, Clift RA, Thomas ED: **Histopathology of graft-vs.-host reaction (GvHR) in human recipients of marrow from HL-A-matched sibling donors**. *Transplantation Proceedings* 1974, **6**(4):367-371.

166. Sale GE: **Pathology and recent pathogenetic studies in human graft-versus-host disease.** *Survey and synthesis of pathology research* 1984, **3**(3):235-253.
167. Sale GE, Lerner KG, Barker EA, Shulman HM, Thomas ED: **The skin biopsy in the diagnosis of acute graft-versus-host disease in man.** *Am J Pathol* 1977, **89**(3):621-635.
168. Sviland L, Pearson AD, Green MA, Baker BD, Eastham EJ, Reid MM, Hamilton PJ, Proctor SJ, Malcolm AJ: **Immunopathology of early graft-versus-host disease--a prospective study of skin, rectum, and peripheral blood in allogeneic and autologous bone marrow transplant recipients.** *Transplantation* 1991, **52**(6):1029-1036.
169. Atkinson K, Munro V, Vasak E, Biggs J: **Mononuclear cell subpopulations in the skin defined by monoclonal antibodies after HLA-identical sibling marrow transplantation.** *British Journal of Dermatology* 1986, **114**(2):145-160.
170. Snover DC, Weisdorf SA, Ramsay N, McGlave P, Kersey J: **Hepatic graft versus host disease: a study of the predictive value of liver biopsy in diagnosis.** *Hepatology* 1984, **4**(1):123-130.
171. Shulman HM, Sharma P, Amos D, Fenster LF, McDonald GB: **A coded histologic study of hepatic graft-versus-host disease after human bone marrow transplantation.** *Hepatology* 1988, **8**(3):462-470.
172. Epstein RJ, McDonald GB, Sale GE, Shulman HM, Thomas ED: **The diagnostic accuracy of the rectal biopsy in acute graft-versus-host disease: a prospective study of thirteen patients.** *Gastroenterology* 1980, **78**(4):764-771.
173. Atkinson K, Horowitz MM, Biggs J, Gale RP, Rimm AA, Bortin MM: **The clinical diagnosis of acute graft-versus-host disease: a diversity of views amongst marrow transplant centers.** *Bone Marrow Transplantation* 1988, **3**(1):5-10.
174. Martino R, Romero P, Subirá M, Bellido M, Altés A, Sureda A, Brunet S, Badell I, Cubells J, Sierra J: **Comparison of the classic Glucksberg criteria and the IBMTR Severity Index for grading acute graft-versus-host disease following HLA-identical sibling stem cell transplantation.** *Bone Marrow Transplantation* 1999, **24**(3):283-287.
175. Weisdorf DJ, Hurd D, Carter S, Howe C, Jensen LA, Wagner J, Stablein D, Thompson J, Kernan NA: **Prospective grading of graft-versus-host disease after unrelated donor marrow transplantation: a grading algorithm versus blinded expert panel review.** *Biology of Blood and Marrow Transplantation* 2003, **9**(8):512-518.
176. Vinterbo SA, Kim E-Y, Ohno-Machado L: **Small, fuzzy and interpretable gene expression based classifiers.** *Bioinformatics* 2005, **21**(9):1964-1970.
177. Dunn J: **Well separated clusters and optimal fuzzy partitions.** *Journal Cybernet* 1974, **4**:95-104.
178. Hastie T, Tibshirani R: **Discriminant Analysis by Gaussian Mixtures.** *Journal of the Royal Statistical Society Series B* 1996, **58**:158-176.

179. Hall MA: **Correlation-based Feature Subset Selection for Machine Learning.** *PhD dissertation.* Hamilton, New Zealand: University of Waikato; 1999.
180. Witten IH, Frank E: **Data Mining: Practical machine learning tools and techniques**, Second edn. San Francisco: Elsevier; 2005.
181. Platt JC: **Fast Training of Support Vector Machines using Sequential Minimal Optimization**. In: *Advances in Kernel Methods - Support Vector Learning*. Edited by Schoelkopf B, Burges CJC, Smola AJ. Cambridge, Massachusetts: MIT Press; 1998: 185-208.

APPENDICES

Appendix A. Patient information on maximum GvHD grade, GvHD diagnosis in days post-transplant and patient-donor relationship

Patient #	Max aGvHD grade	aGvHD post-transplant	cGvHD post-transplant	Donor-patient relationship	Comments
1	3	26		MUD	Last contact 187 days post-transplant
2	0			SIB	
3	4	23		MUD	Expired 61 days post-transplant
4	0			SIB	Expired 278 days post-transplant
5	3	59		SIB	
6	3	19		SIB	
7	3	39		SIB	Expired 89 days post-transplant
8	0		122	SIB	
9	3	43	211	SIB	
10	1	11		MUD	
11	1	68	273	SIB	
12	3	22		SIB	
13	3	48		SIB	
14	2	28		MUD	Relapsed
15	2	19	98	SIB	
16	2	10		MUD	Expired 74 days post-transplant
17	0			SIB	Relapsed
18	0			SIB	Expired 54 days post-transplant
19	2	77	446	SIB	
20	0			SIB	Expired 55 days post-transplant
21	3	54	294	MUD	
22	3	32	223	SIB	
23	3	22		SIB	Last contact < 100 days post transplant
24	3	37		SIB	
25	1	44		SIB	Expired 89 days post-transplant
26	0		117	SIB	

Patient #	Max aGvHD grade	aGvHD post-transplant	cGvHD post-transplant	Donor-patient relationship	Comments
27	2	31		SIB	
28	1	51	177	MUD	
29	0			SIB	Expired 97 days post-transplant
30	0		104	SIB	
31	0			SIB	Last contact 109 days post-transplant; Relapsed

Appendix B. List of the subsets of immune cells from each of the ten aliquots

Aliquots	Immune cells
1 Activation	SSC,FSC/CD3 PerCP^{+}
	SSC,FSC/CD3 PerCP^{+}/CD44^{-}CD25^{-}
	SSC,FSC/CD3 PerCP^{+}/CD44^{-}CD25^{+}
	SSC,FSC/CD3 PerCP^{+}/CD44^{+}CD25^{+}
	SSC,FSC/CD3 PerCP^{+}/CD44^{+}CD25^{+}/CD69^{+}
	SSC,FSC/CD3 PerCP^{+}/CD44^{+}CD25^{-}
	SSC,FSC/CD3 PerCP^{-}
	SSC,FSC/CD3 PerCP^{-}/CD44^{+}CD25^{+}
	SSC,FSC/CD3 PerCP^{-}/CD44^{+}CD25^{+}/CD69^{+}
	SSC,FSC/CD3 PerCP^{-}/CD44^{-}CD25^{-}
	SSC,FSC/CD3 PerCP^{-}/CD44^{+}CD25^{-}
2 Activation	SSC,FSC/CD3 PerCP^{+}
	SSC,FSC/CD3 PerCP^{+}/CD4^{br}
	SSC,FSC/CD3 PerCP^{+}/CD4^{int}
	SSC,FSC/CD3 PerCP^{+}/CD8^{dim}
	SSC,FSC/CD3 PerCP^{+}/CD8^{br}
	SSC,FSC/CD3 PerCP^{-}
	SSC,FSC/CD3 PerCP^{-}/CD4^{dim}
	SSC,FSC/CD3 PerCP^{-}/CD4^{-}CD8^{-}
	SSC,FSC/CD3 PerCP^{-}/CD8^{low}
3 Activation	SSC,FSC/CD3 PerCP^{+}
	SSC,FSC/CD3 PerCP^{+}/CD4^{br}
	SSC,FSC/CD3 PerCP^{+}/CD4^{int}
	SSC,FSC/CD3 PerCP^{+}/CD8^{dim}
	SSC,FSC/CD3 PerCP^{+}/CD8^{br}
	SSC,FSC/CD3 PerCP^{-}
	SSC,FSC/CD3 PerCP^{-}/CD4^{dim}
	SSC,FSC/CD3 PerCP^{-}/CD4^{-}CD8^{-}
	SSC,FSC/CD3 PerCP^{-}/CD8^{low}
	SSC,FSC/CD3 PerCP^{-}/CD8^{low}/CD122^{hi}
B cells	SSC,FSC/CD20^{+}
	SSC,FSC/CD22^{+}
	SSC,FSC/CD22^{+}CD20^{+}
	SSC,FSC/CD20^{+}CD19^{+}

Aliquots	Immune cells
Myeloids	SSC,FSC/CD33^{+}CD45^{+}
	SSC,FSC/CD33^{+}CD45^{+}/CD15^{+}CD14^{+}
	SSC,FSC/CD33^{+}CD45^{dim}
	SSC,FSC/CD33^{+}CD45^{dim}/CD15^{+}CD14^{+}
	SSC,FSC/CD33^{+}CD45^{dim}/CD15^{low}CD14^{low}
	SSC,FSC/CD33^{+}CD45^{dim}/CD15^{+}CD14^{-}
	SSC,FSC/CD45^{+}CD33^{-}
	SSC,FSC/CD45^{+}CD33^{-}/CD15^{+}CD14^{-}
NK cells	SSC,FSC/CD2^{-}CD16^{+}
	SSC,FSC/CD2^{-}CD16^{+}/CD56^{+}CD3^{-}
	SSC,FSC/CD2^{-}CD16^{+}/CD3^{+}CD56^{-}
	SSC,FSC/CD2^{-}CD16^{+}/CD56^{-}CD3^{-}
	SSC,FSC/CD2^{dim}CD16^{+}
	SSC,FSC/CD2^{dim}CD16^{+}/CD56^{+}CD3^{-}
	SSC,FSC/CD2^{dim}CD16^{+}/CD3^{+}CD56^{-}
	SSC,FSC/CD2^{dim}CD16^{+}/CD56^{-}CD3^{-}
	SSC,FSC/CD2^{+}CD16^{+}
	SSC,FSC/CD2^{+}CD16^{+}/CD56^{+}CD3^{-}
	SSC,FSC/CD2^{+}CD16^{+}/CD3^{+}CD56^{-}
	SSC,FSC/CD2^{+}CD16^{+}/CD56^{-}CD3^{-}
	SSC,FSC/CD2^{+}CD16^{-}
	SSC,FSC/CD2^{+}CD16^{-}/CD56^{+}CD3^{-}
	SSC,FSC/CD2^{+}CD16^{-}/CD3^{+}CD56^{-}
	SSC,FSC/CD2^{+}CD16^{-}/CD56^{-}CD3^{-}
T cells	SSC,FSC/CD3 PerCP^{+}
	SSC,FSC/CD3 PerCP^{+}/CD4^{+}CD8β^{-}
	SSC,FSC/CD3 PerCP^{+}/CD4^{+}CD8β^{+}
	SSC,FSC/CD3 PerCP^{+}/CD4^{+}CD8β^{+}/CD8^{+}
	SSC,FSC/CD3 PerCP^{+}/CD4^{+}CD8β^{+}/CD8^{+} (proportion of CD3^{+} cells)
	SSC,FSC/CD3 PerCP^{+}/CD8β^{+}CD4^{-}
	SSC,FSC/CD3 PerCP^{+}/CD8^{+}CD8β^{-}
	SSC,FSC/CD3 PerCP^{+}/CD8β^{+}CD8^{+}
	SSC,FSC/CD3 PerCP^{+}/CD8β^{+}CD8^{low}
	SSC,FSC/CD3 PerCP^{+}/CD8β^{dim}CD8^{-}
	SSC,FSC/CD3 PerCP^{-}
	SSC,FSC/CD3 PerCP^{-}/CD4^{low}CD8β^{low}
	SSC,FSC/CD3 PerCP^{-}/CD8^{+}CD8β^{-}

Aliquots	Immune cells
T cells	SSC,FSC/CD3 PerCP^{-}/CD8β^{dim}CD8^{-}
rest/activate T helper	SSC,FSC/CD3 PerCP^{+}
	SSC,FSC/CD3^{+}CD4^{+}
	SSC,FSC/CD3^{+}CD4^{-}
	SSC,FSC/45ROCD3^{+}
	SSC,FSC/45ROCD3^{+}/CD4^{+}
	SSC,FSC/45ROCD3^{+}/CD4^{-}
	SSC,FSC/45ROCD3^{+}/CD4^{low}
	SSC,FSC/45RACD3^{+}
	SSC,FSC/45RACD3^{+}/CD4^{+}
	SSC,FSC/45RACD3^{+}/CD4^{-}
	SSC,FSC/45RACD3^{+}/CD4^{low}
	SSC,FSC/CD3 PerCP^{-}
	SSC,FSC/CD4^{dim}
	SSC,FSC/CD3^{-}CD4^{-}
	SSC,FSC/45ROCD3^{-}
	SSC,FSC/45ROCD3^{-}/CD4^{dim}
	SSC,FSC/45RACD3^{-}
	SSC,FSC/45RACD3^{-}/CD4^{dim}
rest/activate T suppressor	SSC,FSC/CD3 PerCP^{+}
	SSC,FSC/CD3^{+}CD8^{+}
	SSC,FSC/CD3^{+}CD8^{-}
	SSC,FSC/45ROCD3^{+}
	SSC,FSC/45ROCD3^{+}/CD8^{+}
	SSC,FSC/45ROCD3^{+}/CD8^{-}
	SSC,FSC/45ROCD3^{+}/CD8^{low}
	SSC,FSC/45RACD3^{+}
	SSC,FSC/45RACD3^{+}/CD8^{+}
	SSC,FSC/45RACD3^{+}/CD8^{-}
	SSC,FSC/45RACD3^{+}/CD8^{low}
	SSC,FSC/CD3 PerCP^{-}
	SSC,FSC/CD8^{+}CD3^{-}
	SSC,FSC/CD3^{-}CD8^{-}
	SSC,FSC/45ROCD3^{-}
	SSC,FSC/45RACD3^{-}
	SSC,FSC/45RACD3^{-}/CD8^{+}

Aliquots	Immune cells
TCR	SSC,FSC/CD3 PerCP^{+}
	SSC,FSC/CD3 PerCP^{+}/TCRab^{+}CD5^{+}
	SSC,FSC/CD3 PerCP^{+}/TCRab^{+}CD5^{+}/TCRab^{+}TCRgd^{+}
	SSC,FSC/CD3 PerCP^{+}/TCRgd^{+}CD5^{+}
	SSC,FSC/CD3 PerCP^{+}/TCRab^{+}CD5^{-}
	SSC,FSC/CD3 PerCP^{+}/TCRab^{+}CD5^{-}/TCRab^{+}TCRgd^{+}
	SSC,FSC/CD3 PerCP^{+}/TCRab^{+}CD5^{-}/TCRgd^{-}CD5^{-}
	SSC,FSC/CD3 PerCP^{-}
	SSC,FSC/CD3 PerCP^{-}/CD5^{+}
	SSC,FSC/CD3 PerCP^{-}/TCRab^{+}CD5^{+}
	SSC,FSC/CD3 PerCP^{-}/TCRab^{+}CD5^{-}
	SSC,FSC/CD3 PerCP^{-}/TCRab^{+}TCRgd^{-}/CD5^{+}
	SSC,FSC/CD3 PerCP^{-}/TCRab^{+}CD5^{-}/TCRab^{+}TCRgd^{+}
	SSC,FSC/CD3 PerCP^{-}/TCRab^{+}CD5^{-}/TCRgd^{-}CD5^{-}
	SSC,FSC/CD3 PerCP^{-}/TCR^{+}CD5^{+}
	SSC,FSC/CD3 PerCP^{-}/TCRab^{+}TCRgd^{-}/CD5^{-}

**the '/' indicates each level of the sequential gating.

Appendix C. PERL script fixFCS.pl for enforcing FCS file compatibility from FlowJo into rflowcyt

```
#!usr/bin/perl

#fixFCS_v0.7.pl
#Written by Shang-Jung (Jessica) Lee
#BC Cancer Research Centre
#Last updated: December 13, 2006
#Maintainer: Jessica Lee <jlee@bccrc.ca>

#Please be noted that Immune cell populations and measurements were used interchangeably in the
PERL codes/documentation
#This PERL script reads in the FCS files from FlowJo (Tree Star, Inc, Oregon)
#It then creates a new FCS files with necessary modifications to be successfully read into R via
rflowcyt
#NOTE: information on experiment details, samples labels may be lost!!
#Folder and files are selected based on its names.  User can modify this selection in the regular
expression located below the comment "######## USER MODIFY HERE"
#This script will also have updated header with new bytes information

#Tested on FCS version 2.0 exported from FlowJo version 6.3.4
#Tested with rflowcyt version 1.4.0 on R (windows 2.3.0)
#On Windows XP (Pentium 4 CPU, 1.00GB of RAM), it takes less than 1 minute to search through 500
files and modify/create 200 files.
#Please report all bugs and suggestions to <jlee@bccrc.ca>


use warnings;
use strict;
use File::Find;
use Storable;
use Getopt::Long;
use bytes ();
```

```
############
### MAIN ###
############
#opens log file to record status and errors
open (OUTFILElog, ">>fixFCS.log") or die ("Cannot open output file: $!");
print (OUTFILElog "\n\nSTART TIME: " . scalar localtime() . "\n");
#selects folder with the FCS files to be modified
&SELECT_FOLDER();
close (OUTFILElog) or die ("Cannot close output file: $!");



#################################################################
#sub SELECT_FOLDER
#selects one or all subfolders within the current location based on user specification
#calls subroutine SELECT_FILES internally

sub SELECT_FOLDER {
    print ("Subfolder name or \'all\' for all subfolders (based on the default selection criteria,
p#): ");
    chomp(my $userFolder = <STDIN>);

    if ($userFolder =~ m/all$/i){ #select all folders
     my @folderNames;
     find sub {push @folderNames, $File::Find::name if -d}, '.';
     foreach my $folderPos (0..$#folderNames){

         ####### USER MODIFY HERE for selecting folder/file
         if ($folderNames[$folderPos] =~ m|\./(p[\d]+)|){
          &SELECT_FILES ("$1");
         }
     }
    }
    else{ #select specific folder
```

```
      if (-d $userFolder){
          &SELECT_FILES ("$userFolder");
      }
      else{
          print OUTFILElog "END PROGRAM: cannot find folder $userFolder\n";
          die ("Cannot find folder $userFolder: $!");
      }
    }
} #sub SELECT_FOLDER



#################################################################
#sub SELECT_FILES
#selects the correct FCS files based on its naming scheme
#calls subroutine FIX internally
#INPUT: name and location (optional) of the subfolder where FCS files are located

sub SELECT_FILES {
    my $patientFolder = shift(@_);
    my @FCSfiles;
    #find all files in folder..
    find sub {push @FCSfiles, $File::Find::name}, "./$patientFolder";
    my %fixed;
    my %toBeFix;
    #in order to save time, skip any FCS file that was already fixed (ie a corresponding FCS file
with the modified data exists (+ "_fixed"))
    foreach my $currentFile (@FCSfiles){
     if (!($currentFile =~ m/\._/)){
          if ($currentFile =~ m/(.+)_fixed\.fcs/){ ####### USER MODIFY HERE for selecting /
excluding files
           $fixed{"$1"} = "$currentFile";
          }
          elsif($currentFile =~ m/(.+)\.fcs/){
           $toBeFix{"$1"} = "$currentFile";
          }
```

```
      }
    }
    foreach my $key (keys %toBeFix){
      if (!(exists($fixed{$key}))){
          print (OUTFILElog "fixing: $toBeFix{$key}\n");
          &FIX ("$toBeFix{$key}");
      }
      else{
          print (OUTFILElog "FIXED: $toBeFix{$key}\n");
      }
    } #foreach file
} #sub SELECT_FILES



################################################################
#sub FIX
#removes the unwanted keywords in the FCS file
#updates bytes information in the header
#creates a new FCS file with necessary modificaiton
#INPUT: name and location (optional) of the FCS file

sub FIX {
    my $currentFile = shift(@_);
    my $newFileName = "$currentFile";
    $newFileName =~ s/\.fcs/_fixed.fcs/; ####### USER MODIFY HERE for naming scheme
    my $keywords;

    my $temp;
    #reading in the BINARY file
    open (INFILE, "<:raw", "$currentFile") or die ("Cannot open input file: $!");
    binmode (INFILE);
    until (eof INFILE){
      $temp .= <INFILE>;
```

```
}

#remove $FIL (not necessary)
#if ($entireText =~ m|\$FIL.+\.fcs\\\$NEXTDATA|){
  #$entireText =~ s|\$FIL.+\.fcs\\\$NEXTDATA|\\\$NEXTDATA|;
  #print (OUTFILElog "remove \$FIL\n");
#}

#remove $BTIM...BD$NPAR....BD$P1N...
if ($temp =~ m|(\\\$DATATYPE\\.{1,2})\\\$BTIM\\.+\\BD\$NPAR.+\\BD\$P1N.+(\\\$P1N)|){
  print (OUTFILElog "remove \$BTIM...BD\$NPAR\n");
  $temp =~ s|(\\\$DATATYPE\\.{1,2})\\\$BTIM\\.+\\BD\$NPAR.+\\BD\$P1N.+(\\\$P1N)|$1$2|;
}
#remove $BEGINDATA (not necessary)


  #determine the old byte infomration
  my %oldBytes;
  my $newHeader;
  if ($segments[0]=~ m/(FCS\d\.\d)(.+)$/){
      $newHeader = $1;
      $oldBytes{"original"} = "" . $2;
      $oldBytes{"start keyword"} = substr($oldBytes{"original"},0,12);
      $oldBytes{"end keyword"} = substr($oldBytes{"original"},12,8);
      $oldBytes{"start data"} = substr($oldBytes{"original"},20,8);
      $oldBytes{"end data"} = substr($oldBytes{"original"},28,8);
      $oldBytes{"s0s0"} = substr($oldBytes{"original"},36);
      #make sure that the digits between the old byte and the new byte information are the
same
      if ((length($oldBytes{"start keyword"})-length($bytes{"start keyword"})) >=0){
        for (1..(length($oldBytes{"start keyword"})-length($bytes{"start keyword"}))){
            $newHeader .= " ";}
        $newHeader .= $bytes{"start keyword"};
      }
      else{print (OUTFILElog "Error: over size limit\n"); return();}
```

```
        if((length($oldBytes{"end keyword"})-length($bytes{"end keyword"}))>=0){
          for (1..(length($oldBytes{"end keyword"})-length($bytes{"end keyword"}))){
              $newHeader .= " ";}
          $newHeader .= $bytes{"end keyword"};
        }
        else{print (OUTFILElog "Error: over size limit\n"); return();}

        if ((length($oldBytes{"start data"})-length($bytes{"start data"}))>=0){
          for (1..(length($oldBytes{"start data"})-length($bytes{"start data"}))){
              $newHeader .= " ";}
          $newHeader .= $bytes{"start data"};
        }
        else{print (OUTFILElog "Error: over size limit\n"); return();}

        if((length($oldBytes{"end data"})-length($bytes{"end data"}))>=0){
          for (1..(length($oldBytes{"end data"})-length($bytes{"end data"}))){
              $newHeader .= " ";}
          $newHeader .= $bytes{"end data"};
        }
        else{print (OUTFILElog "Error: over size limit\n"); return();}
        $newHeader .= $oldBytes{"s0s0"};

        #replace old header with the new one
        open (OUTFILE, ">$newFileName") or die ("Cannot open output file: $!");
        binmode(OUTFILE);
        print (OUTFILE "" . $newHeader . $spaces40 . $segments[1] . $spaces40 . $segments[2]);
        close (OUTFILE) or die ("Cannot close output file: $!");
      }
    }
    else {
      print (OUTFILElog "ERROR: Cannot locate header in the FCS file ($currentFile)\n");
      return();
    }
} #sub FIX
```

Appendix D. PERL script viz_days.pl for flow cytometry data transformation

```
#!usr/bin/perl
use strict;
use warnings;
use File::Find;
use Storable;
use Getopt::Long;

#viz_days.pl
#Written by Shang-Jung (Jessica) Lee
#BC Cancer Research Centre
#Last updated: July 14, 2006
#Maintainer: Jessica Lee <jlee@bccrc.ca>

#Please be noted that Immune cell populations and measurements were
used interchangeably in the PERL codes/documentation
#This PERL script reads in files containing flow cytometry data and
clinical data
##acute GvHD diagnosis time in days post-transplant from file
"GvHD_days_p31.txt" via subroutine GVHD_DAY
##flow cytometry data files (for each patient, each available
aliquot) in the specified subfolder via subroutine FILES
##sampling time points for each patient in file
"sampling_time_p31.txt" via subroutine SAMPLING_TIME
##MNC values estiamted from different samples of the same patient
population from files "JL_MNC.txt" via subroutine READ_MNC

#It combines these files and user specified information such as
excerpt time range
#New files are created grouping samples from patients taken at
specific time range into individual file for each available
measurement in subroutine 'visualization'

#make a subfolder named 'visualization' if it does not exist
if (!-d ".\\\\visualization"){mkdir ".\\\\visualization" or die
("Cannot make subfolder visualization");}
my $log = ".\\visualization\\log_viz.txt"; #lot file

#Pre-specified parameters
#average aGvHD diagnosis in days post-transplant, used in the data
transformation of non-GvHD data from days post-transplant into days
from aGvHD diagnosis
my $averageGVHD = 36;
#input files:
#GVHD diagnosis day
my $gvhd_diagnosis_inputFile = "GvHD_days_p31.txt";
if (-e $gvhd_diagnosis_inputFile){die ("Cannot find file:
$gvhd_diagnosis_inputFile");}
#sampling time points
my $sampling_inputFile = "sampling_time_p31.txt";
```

```
if (-e $sampling_inputFile){die ("Cannot find file:
$sampling_inputFile");}
#mnc values
my $mnc_inputFile = "JL_MNC.txt";
if (-e $mnc_inputFile){die ("Cannot find file: $mnc_inputFile");}


################
### sub MAIN ###
################
my $reference;
#user specified option: time in days from transplantation or aGvHD
diagnosis
GetOptions('r|reference=s'=>\$reference);
if(!$reference || !($reference =~ m/transplant|gvhd/i)){
    die ("Usage: perl visualization.pl -r <post - \"transplant\" or
\"gvhd\"");
}

#open lot file to record status and errors
open (OUTFILElog, ">$log") or die ("Cannot open output file: $!");

#obtain data from files by calling the various subroutines
#read in raw flow cytometry data file as exported from FlowJo
my @temp = &FILES();
my %data = %{$temp[0]};  #data
my %ann = %{$temp[1]};  #measurement names
#read in sampling time poits for each patient in days post-
transplant
my %samplingTime = %{&SAMPLING_TIME()};
#read in GvHD diagnosis in days post-transplant
my %GVHDdays = %{&GVHD_DAY()};
#read in the mnc values
my %MNC = %{&READ_MNC()};

#############################
#change time scale:
#if user choose acute GvHD diagnosis as a point of reference
(instead of the transplantation), changes the days in sampling time,
data, and mnc
#time is originally recorded in days post-transplant
if ($reference =~ m/gvhd/i){
    print "changing sampling time and data to reflect time post-
aGVHD\n";
    my %tempData;
    my %tempMNC;
    foreach my $tempPatient (keys %GVHDdays){
     #get the GvHD diagnosed day (days post-transplant) for each
patient
     #if the patient was never diagnosed with GvHD (GVHDday = 0),
average GvHD day which is set at the beginning of the script is used
     my $gvhd = 0 + $GVHDdays{$tempPatient};
```

```
      if ($gvhd ==0){
          $GVHDdays{$tempPatient} = 0 + $averageGVHD;
          $gvhd = 0 + $averageGVHD;
      }

      #change the day in samplingTime
      if(exists($samplingTime{$tempPatient})){
          foreach (0..$#{$samplingTime{$tempPatient}}){
            $samplingTime{$tempPatient}[$_] =
$samplingTime{$tempPatient}[$_] - $gvhd;
          }}
      else{print (OUTFILElog "###Cannot find the following patient
in samplingTime: $tempPatient\n");}

      #change the day in data
      if(exists($data{$tempPatient})){
          foreach my $tempGroup (keys %{$data{$tempPatient}}){
            foreach my $tempMeasurement (keys
%{$data{$tempPatient}{$tempGroup}}){
                foreach (keys
%{$data{$tempPatient}{$tempGroup}{$tempMeasurement}}){

      $tempData{$tempPatient}{$tempGroup}{$tempMeasurement}{0+($_ -
$gvhd)} = 0 + $data{$tempPatient}{$tempGroup}{$tempMeasurement}{$_};
                }}}
      }
      else{print (OUTFILElog "###Cannot find the following patient
in data: $tempPatient\n");}

      #change the day in MNC
      if(exists($MNC{$tempPatient})){
          foreach my $MNCday (keys %{$MNC{$tempPatient}}){
            $tempMNC{$tempPatient}{0+($MNCday-$gvhd)} =
0+$MNC{$tempPatient}{$MNCday};
          }}
      else{print (OUTFILElog "Cannot find the following patient in
MNC: $tempPatient\n");}

    } #foreach patient
    %data = %tempData;
    %MNC=%tempMNC;
}

########################################
#get MNC sampling day for each patient into the array
%MNC{patient}{"array"}
foreach my $patientToArray (keys %MNC){
    foreach my $dayToArray (keys %{$MNC{$patientToArray}}){
      push (@{$MNC{$patientToArray}{"array"}}, 0 + $dayToArray);
    }
    @{$MNC{$patientToArray}{array}} = sort {$a<=>$b}
@{$MNC{$patientToArray}{array}};}
```

```
########################################
#user input time range (post-transplant or post-aGVHD)
my @rangeDays;
print ("Specify time range (interger in DAYS) separated by \',\'.
Leave this empty for the maximum available time range: ");
chomp(my $input = <STDIN>);
my @userSpecifyRange;

if ($input){
    @userSpecifyRange = split(",", $input);
    $rangeDays[0] = 0 + $userSpecifyRange[0];
    $rangeDays[1] = 0 + $userSpecifyRange[1];}
else{
    #determine the earliest and the latest day in sampling time
    my $earliest = 100;
    my $latest = -100;
    foreach (keys %samplingTime){
      foreach (@{$samplingTime{$_}}){
          if ($earliest > $_){$earliest = 0 + $_;}
          if ($latest < $_){$latest = 0 + $_;}
      }
    }
    $rangeDays[0] = $earliest;
    $rangeDays[1] = $latest;
}

################################
#MAIN PRINT OUT
foreach my $group (keys %ann){
    foreach my $measurement (keys %{$ann{"$group"}}){
      my $tempGroup = "" . $group;
      $tempGroup =~ s/\s|-//g;
      my $tempMeasurement = "" . $measurement;
      $tempMeasurement =~ s~/|\&~~g;
      $tempMeasurement =~ s/\s/_/g;
      $tempMeasurement =~ s/\+/plus/g;
      $tempMeasurement =~ s/-/minus/g;
      #subfolder visualization, convert possible minus sign (from
the time range) to "minus"
      my $subfolder1 = ".\\visualization\\asis_" . $reference .
"_d" . $rangeDays[0] . "_d" . $rangeDays[1]; $subfolder1 =~ s/-
/minus/g;
    my $subfolder2 = ".\\visualization\\mnc_" . $reference . "_d" .
$rangeDays[0] . "_d" . $rangeDays[1]; $subfolder2 =~ s/-/minus/g;
      #make the subfolder if they did not exist already
      if (!-d $subfolder1){mkdir $subfolder1 or die "Cannot make
subfolder $subfolder1"};
      if (!-d $subfolder2){mkdir $subfolder2 or die "Cannot make
subfolder $subfolder2"};
      ###HAVE NOT implemented to delete all existing files in the
subfolder
```

```
     my $fileAsis =  "$subfolder1\\" . $tempGroup . "_" .
$tempMeasurement . ".txt";
     my $fileMNC =  "$subfolder2\\" . $tempGroup . "_" .
$tempMeasurement . ".txt";
     open (OUTFILEAsis, ">$fileAsis") or die ("Cannot open output
file ($fileAsis): $!");
     open (OUTFILEMNC, ">$fileMNC") or die ("Cannot open output
file($fileMNC): $!");
     print (OUTFILEAsis "time"); print (OUTFILEMNC "time");

     foreach (sort {$a<=>$b} keys %samplingTime){ #print header
(patients)
         print (OUTFILEAsis "\tp$_"); print (OUTFILEMNC "\tp$_");
     }
     print (OUTFILEAsis "\n"); print (OUTFILEMNC "\n");

     foreach my $currentDay ($rangeDays[0]..$rangeDays[1]){
         print (OUTFILEAsis "$currentDay"); print (OUTFILEMNC
"$currentDay");
         foreach my $patient (sort {$a<=>$b} keys %samplingTime){

           my $currentProportion;
           if(exists($data{$patient}) &&
exists($data{$patient}{$group}) &&
exists($data{$patient}{$group}{$measurement}) &&
exists($data{$patient}{$group}{$measurement}{$currentDay})){
               $currentProportion = 0 +
$data{$patient}{$group}{$measurement}{$currentDay};
           }
           #matching MNC value at the closest sampling time to the
current day
           my @closest = &CLOSEST("$currentDay",
\@{$MNC{$patient}{array}});
           my $currentMNC;
           if (exists($MNC{$patient}) &&
exists($MNC{$patient}{$closest[1]})){
               $currentMNC = 0 + $MNC{$patient}{$closest[1]};
           }
           if (defined($currentProportion)){
               print (OUTFILEAsis "\t" . $currentProportion);
               if (defined($currentMNC)){
                 print (OUTFILEMNC "\t" . ($currentProportion *
$currentMNC));
               }
               else {
                 print (OUTFILEMNC "\t");
               }
           }
           else{
               print (OUTFILEAsis "\t");
               print (OUTFILEMNC "\t");
           }
```

```
        } #patient
        print (OUTFILEAsis "\n"); print (OUTFILEMNC "\n");
      } #current week
      close (OUTFILEAsis) or die ("Cannot close output file: $!");

      close (OUTFILEMNC) or die ("Cannot close output file: $!");
    } #measurement
} #group
close (OUTFILElog) or die ("Cannot close output file: $!");

#############################################################
#sub GVHD_DAY
#read in GVHD diagnosis day (post-transplant) from the specified
file $gvhd_diagnosis_inputFile
#return hash %aGvHD
##$aGVHD{patient number (number only)} => aGVHD diagnosis day (zero
if the patient was not diagnosed with aGVHD)

sub GVHD_DAY{
    my $input = "$gvhd_diagnosis_inputFile";

    #storing parsed GvHD diagnosis day data into hash %aGvHD
    my %aGVHD;

    print "performing subroutine GVHD_DAY...\n";
    open (INFILE, "$input") or die ("Cannot open input file: $!");
    until (eof INFILE){
      chomp(my $newText = <INFILE>);
      #first row contains patient number and second row is the aGVHD
diagnosis day (post-transplant)
      my @values = split("\t", $newText);
      if(@values){
          my $patient;
          if ($values[0] =~ m/([\d]+)/){
            my $temp = $1;
            $temp =~ s/^0+//g; #remove any zero at the beginning
            $patient = 0 + $temp;
          }
          else {
            print (OUTFILElog "###CANNOT find patient number in
GVHD_DAY: @values\n");
            die("Cannot find patient number in GVHD_DAY\n");
          }
          if($values[1] =~ m/([\d]+)/){
            $aGVHD{"$patient"} = 0 + $1;
          }
          else{
            print (OUTFILElog "###Cannot find aGVHD diagnosis day in
GVHD_DAY for patient $patient from @values\n");
            die ("Cannot find aGVHD diagnosis day in GVHD_DAY\n");
          }}}
    close (INFILE) or die ("Cannot close input file: $!");
```

```
    return(\%aGVHD);
}
#################################################################
#sub FILES
#read in the raw flow cytometry data files in the user-specified
subfolder
#flow cytometry data files were exported from flowJo
#flow cytometry data file naming scheme:
##the name of the file indicate patient number and aliquot name
##E#'patient number' 'patient initial' 'year'-'aliquot name'.txt
#flow cytometry data file format:
##first row includes sampling time in days post-transplant
##first column indicates the measurement name
#assumptions:
##same measurement (or comparable measurements) have the same name
##did NOT assume that the measurement were listed in any order
#return two hashes: %data and %ann
##%data
###$data{patient number (number only)}{measurement group
name}{measurement name}{time in day post-transplant} => acutal
measurement in % from FlowJo
##%ann
###$ann{measurement group name}{measurement name}

sub FILES{
    my (%data, %ann);
    print ("performing subroutine FILES....\n");

    #prompt for name of the subfolder
    print ("Specify folder name containg the raw data files: ");
    chomp (my $dataFolder = <STDIN>);
    if (!-d "$dataFolder"){
     die ("INVALID folder name entered!\n");
    }

    #find the all files in the specified subfolder
    my @fileNames;
    find sub {push @fileNames, $File::Find::name if !-d},
".\\$dataFolder";
    foreach my $file (@fileNames){
     #derive patient number and aliquot name from the file name
     my ($patient, $group);
     if ($file =~ m|E\#([\d]+)\s[\w]*\s[\d]*(.+)\.txt$|){
         $patient = 0 + $1;
         $group = "$2"; $group =~ s/\.jo-1//; $group =~ s/^-//;

         open (INFILE, "$file") or die ("Cannot open input file:
$!");
         print (OUTFILElog "Reading file: $file\n");

         #header with measurement names in the first column ([1,1]
is always "sample")
```

```
          chomp(my $header = <INFILE>);
          my @titles = split("\t", $header);
          foreach (0..$#titles){
           my $measurement = "$titles[$_]";
           #clean up the measurement name
           $measurement =~ s|^\"*\SSC,, FSC/||;
           $measurement =~ s|,Freq. of Parent\"*$|ofParent|;
           $measurement =~ s|,Freq. of,SSC, FSC\"*$|ofLiveCells|;
           $measurement =~ s|,Freq\..of,CD3.+erCP\+*$|ofTcells|;
           $measurement =~ s|[\s]PerCP||;
           $measurement =~ s|\s||g;
           $titles[$_] = "$measurement";
          }
          until (eof INFILE){
           chomp(my $text = <INFILE>);
           my @values = split("\t", $text);
           if (@values){
               #sampling day post-transplant in the first row (d# or
#d)
               my $day;
               if ($values[0] =~ m/(-*[\d]+)d/ | $values[0] =~ m/d(-
*[\d]+)/){
                $day = 0 + $1;
               }
               else{
                print (OUTFILElog "CANNOT FIND: $values[0]\n");
               }
               foreach my $count (1..$#values){
                if($values[$count] =~ m/[\d]/){ #could be empty

$data{"$patient"}{"$group"}{"$titles[$count]"}{"$day"} = 0 +
$values[$count];
                    $ann{"$group"}{"$titles[$count]"}++;
                }
               }
           }

          } #until

          close (INFILE) or die ("Cannot close input file: $!");
      } #patient number and measurement group name from file name
      else{
          print (OUTFILElog "###CANNOT find patient number and/or
measurement group name from the file name $file\n");
      }
     } #foreach file
     if (!%data || !%ann){die ("No data/annotation");}
     return (\%data, \%ann);
}


##############################################################
```

```
#sub SAMPLING_TIME
#read in the sampling time for each patients (raw data is not used
becuase not all measurements from one patients are available on all
the time points etc)
#input file: sampling_time_p31.txt
#fill in hash data complex: %samplingTime => saved as
samplingTime.hash
#@samplingTime{patient number, 1-31} => sorted (from small to large)
sampling time (day post transplant)
#return \%samplingTime
sub SAMPLING_TIME {
    my $samplingTimeFileName = "samplingTime.hash";
    my $input = "$sampling_inputFile";
    my %samplingTime;
    print "performing subroutine SAMPLING_TIME.....\n";
    open (INFILE, "$input") or die ("Cannot open input file: $!");
    until (eof INFILE){
      chomp (my $newText = <INFILE>);
      my @values = split ("\t", $newText);
      if (@values){
          #first row is the patient numbers
          my $patient;
          if ($values[0] =~ m/([\d]+)/){
            my $temp = $1;
            $temp =~ s/^0+//g; #get rid of extra zero in front of the
patient number(01->1)
            $patient = 0 + $temp;
          }
          else {
            print (OUTFILElog "###CANNOT find patinet number in
SAMPLING_TIME: @values\n");
            die ("Cannot find patient number!");
          }
          foreach my $count (1..$#values){
            if ($values[$count] =~ m/[\d]/){push
@{$samplingTime{"$patient"}}, 0 + $values[$count];}
          }
          @{$samplingTime{"$patient"}} = sort {$a <=> $b}
@{$samplingTime{"$patient"}};
      }}
    close (INFILE) or die ("Cannot close output file: $!");
    return(\%samplingTime);
}


#################################################################
#sub READ_MNC
#read in the mnc values from the specified file $mnc_inputFile
#return hash %MNC
##format: $MNC{patient number}{sampling time in days post-transplant}
=> mnc value (in mm3)
sub READ_MNC {
```

```
    my %MNC;
    open (INFILE, "$mnc_inputFile") or die ("Cannot open input file:
$!");
    my $title = <INFILE>;
    until (eof INFILE){
      chomp(my $newText = <INFILE>);
      my @cols = split ("\t", $newText);
      #cols[0] => patient number
      #cols[1] => sample date
      #cols[2] => MNC value
      #cols[3] => BMT date
      #cols[4] => days post-transplant
      #$MNC{patient number}{days post-transplant} = MNC value
      if ($cols[2] && $cols[0]){ #if both patient number and MNC
value exist
          $MNC{0+$cols[0]}{0 + $cols[4]} = 0 + $cols[2];
      }}
    close (INFILE) or die ("Cannot close input file: $!");
    return (\%MNC);
}


#################################################################
#sub CLOSEST
#INPUT: a target value and an array
#finds the value in the array that is closest to the target value
#returns two values: position of the closest value inside the array
and the actual closest value
####################
### sub CLOSEST ###
####################
sub CLOSEST {
    my $target = shift(@_);
    my @array = @{shift(@_)};
    if ($array[$#array] <= $target){
      return("$#array", "$array[$#array]");
    }
    elsif($array[0] >= $target){return("0", "$array[0]");}
    else{
      foreach my $position (0..$#array){
          my $element = $array[$position];
          if ($element == $target){return("$position", "$element");}
          elsif($element>$target){
            my $fromLarge = abs($element-$target);
            my $fromSmall = abs($target-$array[$position-1]);
            if($fromLarge<$fromSmall){return("$position",
"$element");}
            else{return(($position-1), $array[$position-1]);}}}}}
```

Appendix E. PERL script FLDA_MATLAB.pl for creating MATLAB commands performing FLDA analysis

```
#!usr/bin/perl
use warnings;
use strict;
use File::Find;
use Storable;
use Getopt::Long;
use Tie::File;
use POSIX;


#FLDA_MATLAB.pl
#Written by Shang-Jung (Jessica) Lee
#BC Cancer Research Centre
#Last updated: August 21, 2006
#Maintainer: Jessica Lee <jlee@bccrc.ca>

#Please be noted that Immune cell populations and measurements were
used interchangeably in the PERL codes/documentation
#this script read in the text files (each file represent different
measurement/population) prepared from the viz_days.pl
#it then outputted the necessary MATLAB commands to perform FLDA
classification to each measurements (that qualified, see filter
below).
#FLDA or functional linear discriminant analysis:
## James, G.M. and Hastie, T.J. (2001) Functional linear
discriminant analysis for irregular sampled curves.  Journal of the
Royal Statistical Society, Series B. 63(3): 533-550.
## FLDA was implemented in MATLAB by Simon Dablemont
<Dablemont@dice.ucl.ac.be>
## for everyting related to FLDA (ie setting different parameters
such as grid, B-spline order and knots, please refer to the
published paper and manuals available in Dr. Gareth James' website
<http://www-rcf.usc.edu/~gareth/>

#user is able to select:
##1.  where the data are located
##2.  which population to analyze (by chosing the approprite file)
or all files in the specified folder
##3.  FLDA parameter: grid range (a time range that covers all the
data selected)
##4.  FLDA parameter: grid interval
##5.  FLDA parameter: B-spline order (norder)
##6.  FLDA parameter: number of B-spline knots (which will be placed
uniformly covering the grid
##7.  Different pre-set patient comparisons

##############################################
```

```
#user inputs for data and FLDA parameters
#then checks that inputs (mostly format) are correct

#specify subfolder name where the data are located
print ("Specify the subfolder name: ");
chomp(my $folder = <STDIN>);
print ("Specify the file name or \"all\" for all files in the
specified subfolder: ");
chomp(my $fileName = <STDIN>);
my $userFile = ".\\$folder\\$fileName";
#check if the specified subfolder and file exists
if (!(-e $userFile) && $userFile =~ m/^all$/i) {die "Cannot find
input file: $userFile";}
#grid range (or time range)
print ("Specify grid range(#,#): ");
chomp(my $userInput_grid = <STDIN>);
my @grid = split(",", $userInput_grid);
#check if the correct grid format is used
if (!(scalar(@grid) == 2)){die ("Incorrect grid: $userInput_grid");}
print ("Specify grid interval: ");
chomp(my $by = <STDIN>);
#check if grid is given as number
if ($by =~ m/\D/){die ("Incorrect grid interval: $by");}
#B-spline basis order and knot number
print ("Specify norder and nbreaks (#,#): ");
chomp(my $userInput_orderBreaks = <STDIN>);
my @orderBreaks = split(",", $userInput_orderBreaks);
#check if correct order breaks format is used
if(!(scalar(@orderBreaks) == 2)){die ("Incorrect order and breaks:
$userInput_orderBreaks");}
#select patient comparison
print ("Specify the group membership comparison to use\n");
print ("'1' for aGVHDcGVHD(7) vs. aGVHDlived(9) vs. healthy4(4)\n'2'
for aGVHD(21) vs. healthy4(4)\n");
print ("'3' for aGVHDcGVHD(7) vs. aGVHDlived(9)\n'4' for aGVHD(21)
vs. non aGVHD(7)\n'5' for aGvHDcGvHD (7) vs. aGvHD (14)\n");
print ("group membership comparison: ");
chomp(my $comparison = <STDIN>);
check if the correct comparision number is given
if (!($comparison == 1 || $comparison == 2 || $comparison ==3 ||
$comparison == 4 || $comparison == 5)){die ("Invalid comparison
choice");}

#MAY 31, 2006
## leave-one-out cross-validation does not work for comparision
between more than two classes
if ($comparison == 1){die ("Leave-one-out cross-validation does not
work for comparision between mroe than two classes");}

#initialized @compareGroups based on the comparison chosen
#########YOU CAN MODIFY/CREATE NEW COMPAREGROUPS BY ADDING HERE
my @compareGroups;
```

```
if ($comparison ==1){ #'1' for aGVHDcGVHD(7) vs aGVHDlived(9) vs
healthy4(4)
    @compareGroups = (["aGVHDcGVHD"],
                      ["aGVHDlived"],
                      ["healthy4"],);}
elsif($comparison ==2){ #'2' for aGVHD(21) vs healthy4(4)\n")
    @compareGroups = (["aGVHDcGVHD", "aGVHDlived", "aGVHDdied"],
                      ["healthy4"],);}
elsif($comparison ==3){ #'3' for aGVHDcGVHD(7) vs aGVHDlived(9)
    @compareGroups = (["aGVHDcGVHD"],
                      ["aGVHDlived"],);}
elsif($comparison==4){ #'4' for aGVHD(21) vs non aGVHD(7)\n")
    @compareGroups = (["aGVHDcGVHD", "aGVHDlived", "aGVHDdied"],
                      ["healthy4", "healthyDied"],);}
elsif($comparison==5){ #5' for aGvHDcGvHD (7) vs. aGvHD all (14)
    @compareGroups = (["aGVHDcGVHD"],
                      ["aGVHDlived", "aGVHDdied"],);}
else{die ("###ERROR: incorrect comparison chosen");}



#open OUTFILEs
#outfiles are created within the specified subfolder
#file names includes comparison number, order, and kntos
#An MATLAB code file (.m) and log file (.log) are created
open (OUTFILE, ">.\\\\$folder\\\\FLDA_comparison$comparison" .
"_order$orderBreaks[0]" . "_breaks$orderBreaks[1]" . ".m") or die
("Cannot open output file: $!");
open (OUTFILElog, ">.\\\\$folder\\\\FLDA_comparison$comparison" .
"_order$orderBreaks[0]" . "_breaks$orderBreaks[1]" . ".log") or die
("Cannot open output file: $!");
#create subfolder "data" and "images" if they are not already
existed! (These folders are required for FLDA)
if (!-d ".\\\\$folder\\\\data"){mkdir ".\\\\$folder\\\\data" or die
("Cannot make subfolder data");}
if (!-d ".\\\\$folder\\\\images"){mkdir ".\\\\$folder\\\\images" or
die ("Cannot make subfolder images");}


#determine which files to be processed
#the selected file's name must matched the preset naming scheme
#'aliquot name'_'measurement name'.txt
#exclude files from the subfolders data and image (which have a very
similar naming scheme)
if ($userFile =~ m/all$/i){
    my @fileNames;
    find sub {push @fileNames, $File::Find::name if !-d},
".\\$folder";
    foreach my $f (@fileNames){
      if ($f =~ m|$folder[\\/]+(.+)\.txt$| && !($f =~
m~[\\/]data|images[\\/]~)){
          &MAIN("$f");
```

```
      }
    }
}
else{
    if ($userFile =~ m|$folder[\\/]+(.+)\.txt$|){
      &MAIN("$userFile");
    } #if
    else{
     die ("File $userFile does not have the correct naming scheme");
    }
}

close (OUTFILE) or die ("Cannot close output file: $!");
close (OUTFILElog) or die ("Cannot close output file: $!");



#################
### sub MAIN ###
#################
sub MAIN {
    #class information
    ####################YOU CAN MODIFY/CREATE NEW PATIENT GROUPS
HERE
    my %class;
    $class{"aGVHDcGVHD"}{"p9"} = "p9";   $class{"aGVHDcGVHD"}{"p11"}
= "p11";   $class{"aGVHDcGVHD"}{"p15"} = "p15";
$class{"aGVHDcGVHD"}{"p19"} = "p19";   $class{"aGVHDcGVHD"}{"p21"} =
"p21";   $class{"aGVHDcGVHD"}{"p22"} = "p22";
$class{"aGVHDcGVHD"}{"p28"} = "p28";
    $class{"aGVHDlived"}{"p1"} = "p1";   $class{"aGVHDlived"}{"p5"}
= "p5";   $class{"aGVHDlived"}{"p6"} = "p6";
$class{"aGVHDlived"}{"p10"} = "p10";   $class{"aGVHDlived"}{"p12"} =
"p12";   $class{"aGVHDlived"}{"p13"} = "p13";
$class{"aGVHDlived"}{"p14"} = "p14";   $class{"aGVHDlived"}{"p24"} =
"p24";   $class{"aGVHDlived"}{"p27"} = "p27";
    $class{"aGVHDdied"}{"p3"} = "p3";   $class{"aGVHDdied"}{"p7"} =
"p7";   $class{"aGVHDdied"}{"p16"} = "p16";
$class{"aGVHDdied"}{"p23"} = "p23";   $class{"aGVHDdied"}{"p25"} =
"p25";
    $class{"healthy4"}{"p2"} = "p2";   $class{"healthy4"}{"p4"} =
"p4";   $class{"healthy4"}{"p17"} = "p17";
$class{"healthy4"}{"p31"} = "p31";
    $class{"denovocGVHD"}{"p8"} = "p8";
$class{"denovocGVHD"}{"p26"} = "p26";   $class{"denovocGVHD"}{"p30"}
= "p30";
    $class{"healthyDied"}{"p18"} = "p18";
$class{"healthyDied"}{"p20"} = "p20"; $class{"healthyDied"}{"p29"} =
"p29";

    my $inFile = shift(@_); #name of the current processed file
```

```
    print (OUTFILElog "\nprocessing: $inFile\n");
    print (OUTFILE "%%%MEASUREMENT: $inFile%%%\nclc\nclear
all\nclose all\n");
    my $measurement = "$1";
    #read data from file in the specific folder
    my %data = %{&READTABLE("$inFile")};
    #load individual patient's data into variables in matlab
    #$omitPatients{$patient} = "reason", only include patients with
less than 2 values and more than 1
    #not patients without any value because they are not included in
%data
    my %omitPatients = %{&LOAD_DATA(\%data, \@grid, $by)};


    #check if any of the patient is not in the intput file
    #if not, delete the patient in %class and inlucde the patient in
%omitPatients
    foreach my $tempGroup (keys %class){
     foreach (keys %{$class{$tempGroup}}){
         if (!(exists($data{$_}))){
          delete $class{"$tempGroup"}{"$_"};
          $omitPatients{"$_"} = "is not in the input file";
         }
         if (exists($omitPatients{$_})){
          delete $class{"$tempGroup"}{"$_"};
         }
     }
    }
    foreach (keys %omitPatients){
     print (OUTFILElog "OMIT: $_ is omitted $omitPatients{$_}\n");
    }

    print (OUTFILE "%omit patients: " . join (", ",
keys(%omitPatients)) . "\n");


   #determine if there is enough data to perform FLDA, if not, skip
to the next file
    my @chkNumDataResults = &CHK_NUM_DATA(\%class, \@compareGroups,
\%data);
    if ($chkNumDataResults[1] =~ m/NO,(.+)/){
     print (OUTFILElog "SKIP: $inFile is ignored because$1\n");
     return ();
    }
    my %acceptedPatientsPerGroup = %{shift (@chkNumDataResults)};


############################
#PERFORMING FLDA WITH DATA
```

```
    #group data based on the specified groups in %class, then based
on the comparison type chosen, further group the data to fit the
FLDA format
    &GROUP_DATA(\%class, \@compareGroups,
\%acceptedPatientsPerGroup);
    #initialized flda parameters
    &FLDA_PARAMETERS(\@grid, $by, \@orderBreaks);
    #Running FLDA and writing data, parameters and results into text
file
    &FLDA("$measurement", "$comparison", \@orderBreaks);
    #group data for leave one out validation
    my $validationFileName = &LEAVEONEOUT(\@compareGroups, \%class,
\%acceptedPatientsPerGroup, "$comparison", \@orderBreaks,
"$measurement");
##############################END PERFORMING FLDA WITH DATA


############################
#read existing FLDA results from the current measurement from
subfolder data
#NOT IMPLEMENTED WHEN YOU ARE PERFORMING FLDA ANALYSIS WITH DATA via
subroutine GROUP_DATA, FLDA_PARAMETERS, and FLDA

#    &READ_FLDA_RESULTS ("$measurement", "$comparison",
@orderBreaks);
    #determine knots time index
#    my %knotRanges = %{&KNOTS_POSITIONS (\@grid, \@orderBreaks)};
############################END

    #determine how many values were observed per each compared
gruops of patients.  This is used to represnt how reliable a FLDA
analysis is.
    my %valuePerKnot = %{&VALUE_PER_KNOT(\%data, \%class,
\@compareGroups, \@grid, \@orderBreaks)};
    #determine weights
    #table output with weights and its reliability
    &WEIGHT_ON_KNOTS (\@grid, \@orderBreaks, "$comparison",
"$measurement", \%valuePerKnot, "$validationFileName");

} #sub MAIN



####################
### SUB READTABLE ###
####################
#INPUT:
##1) file name ($)
#OUTPUT:
##1) \%currentData
### $currentData{p#}{time in #} -> value
```

```
#FUNCTIONS:
## read in the specified tab-deliminted data text file
## input data file naming scheme: "group name"_"measurement
name".txt
## file format:
### columns -> patient (identified by patient number)
### row -> time in weeks
### values are actual (average) values of that measurement from the
patient at that time range.
sub READTABLE{
    my %currentData;
    my $inFileName = shift (@_);
    open (INFILE, "$inFileName") or die ("Cannot open input file:
$!");
    chomp(my $titleText = <INFILE>);
    my @titles = split("\t", $titleText);
    until (eof INFILE){
      chomp(my $text = <INFILE>);
      my @values = split("\t", $text);
      foreach my $pos (1..$#titles){
          if (defined($values[$pos]) && $values[$pos] =~ m/[\d]/){
            #$currentData{patient}{time} = measured proportion value
            $currentData{$titles[$pos]}{$values[0]} = 0 +
$values[$pos];}
      }
    }
    close (INFILE) or die ("Cannot close input file: $!");
    return (\%currentData);
}


######################
### SUB LOAD_DATA ###
######################
#INPUT:
##1) \%currentData (from sub READTABLE)
##2) @grid (grid begins at $grid[0] and ends at $grid[1])
##3) the interval of the grid
#OUTPUT:
##1) \%omitPatients
### $omitPatients{p#} -> numbers of values available
#FUNCTIONS:a
## print to OUTFILE
## commands to load individual patient's data
### p#.y = a vector containing values for patient #
### p#.timeindex = a vector containing the time index of the patient
# relative to the specified grid
### p#.curve = a vector of zeros with equal length to p#.y
### determine which patient (if number of available values is <2) is
omitted
sub LOAD_DATA {
    #omit patient if there are less than 2 values available
```

```
    my %omitPatients;
    my %currentData = %{shift(@_)};
    my @currentGrid = @{shift(@_)};
    my $currentBy = shift(@_);
    print (OUTFILE "userGrid = [$currentGrid[0]" . ":$currentBy" .
":$currentGrid[1]]\';\n");
    foreach my $patient (keys %currentData){
     my (@y, @timeindex);
     foreach my $time (sort {$a<=>$b} keys
%{$currentData{$patient}}){
         push (@y, 0 + $currentData{$patient}{$time});
         push (@timeindex, (($time - $currentGrid[0])/$currentBy)
+ 1));
     } #foreach time
     print (OUTFILE "$patient" . ".y = [" . join (", ", @y) .
"]\';\n");
     print (OUTFILE "$patient" . ".timeindex = [" . join (", ",
@timeindex) . "]\';\n");
     print (OUTFILE "$patient" . ".curve = ones(length($patient" .
".y), 1);\n\n");
     if (scalar(@y) < 2){$omitPatients{"$patient"} = "for having
less the three available values (" . scalar(@y) . ")";}
    } #foreach patient
    return(\%omitPatients);
} #sub


#########################
### sub CHK_NUM_DATA ###
#########################
#INPUT:
##%class (from MAIN)
##@compareGroups (global)
##%data (from READTABLE)
#OUTPUT:
##\%acceptedPatientsNum (number of accepted patients per group
##"enough" or "NO" to indicate if there is enough data to run FLDA
#FUNCTIONS:
##determine if there is enough patients to run flda
###pateints with less than 2 values available is omitted
###There must be at least 3 patients included in each class
##determine if there is enough time point to fit the nbreaks
specified
##ie if nbreaks is 4, there must be at least one patient with 4
available data points
sub CHK_NUM_DATA {
    my %currentClass = %{shift(@_)};
    my @compareGroups = @{shift(@_)};
    my %currentData = %{shift(@_)};

    #determine how many qualified patients there are in each group
    #patient is omitted if there are less than 2 values available
```

```
    my %acceptedPatientsNum;
    foreach my $group (keys %currentClass){
      foreach my $okPatients (keys %{$currentClass{"$group"}}){
          $acceptedPatientsNum{"$group"} ++;
      }
    }

    #determine how many qualified patients they are in each class
    my $maxNumData = 0;
    foreach my $numGroup (0..$#compareGroups){ #each class
      my $groupNumCheck = 0;
      foreach (@{$compareGroups[$numGroup]}){ #groups within class
          if(exists($acceptedPatientsNum{"$_"})){$groupNumCheck =
$groupNumCheck + $acceptedPatientsNum{"$_"};} #there are instances
when the whole group of patient is missing (so it won't be in
%acceptedPatientsNum
          foreach my $patient (keys %{$currentClass{"$_"}}){
            my $numDataPerPatient = 0;
            foreach my $time (keys %{$currentData{"$patient"}}){
                $numDataPerPatient ++;
            }
            if ($numDataPerPatient > $maxNumData){$maxNumData = 0 +
$numDataPerPatient;}
          }
      }
      if ($groupNumCheck < 3){return (\%acceptedPatientsNum, "NO,
less than 3 available patients in a class");}
    }
    if ($maxNumData < $orderBreaks[1]){return (\%acceptedPatientsNum,
"NO, nbreaks $orderBreaks[1] > max time points $maxNumData");}
    return (\%acceptedPatientsNum, "enough");
} #sub


######################
### sub GROUP_DATA ###
######################
#INPUT:
##1) \%class, class information in a hash
### $class{group/class}{p#} => p#
##2) comparison chosen
##3) %omitPatients from sub LOAD_DATA
#OUTPUT:
## none
#FUNCTIONS:
## print to OUTFILE
## group each patient data into the pre-specified gruop
### group.y = [p#.y', p#.y']';
### group.timeindex = [p#.timeindex', p#.timeindex']';a
### group.curve = [p#.curve'+1, p#.curve'+2]';
### group.num -> number of available patients in that group
```

```
## further group the grouped data based on the comparison chosen,
into format sutable for FLDA
### class = [ones(group.num, 1) + increment];
### curve = [group.curve' + $increment];
### timeindex = [group.timeindex'..];
### class = [group.y',...];
### data.y = y;
### data.timeindex = timeindex
### data.curve = curve
### data.class = class
sub GROUP_DATA {
    my %currentClass = %{shift(@_)};
    my @compareGroups = @{shift(@_)};
    my %acceptedPatientsNum = %{shift(@_)};
    foreach my $group (keys %acceptedPatientsNum){
     print (OUTFILE "$group" . ".num = " .
(0+$acceptedPatientsNum{"$group"}) . ";\n");
     my (@groupTimeindex, @groupY, @groupCurve);
     my $n = -1;
     foreach (keys %{$currentClass{"$group"}}){
         $n ++;
         push (@groupTimeindex, "$_" . ".timeindex\'");
         push (@groupY, "$_" . ".y\'");
         push (@groupCurve, "$_" . ".curve\'+ $n");
     }
     #group data into the pre-specified groups
     print (OUTFILE "$group" . ".timeindex = [" . join (", ",
@groupTimeindex) . "]\';\n");
     print (OUTFILE "$group" . ".y = [" . join (", ", @groupY) .
"]\';\n");
     print (OUTFILE "$group" . ".curve = [" . join (", ",
@groupCurve) . "]\';\n");
    } #each group

    #further group the grouped data into format sutiable for FLDA
    #group the grouped data based on the comparison chosen
    #first print out matlab commands for each of the following
variables: class, curve, timeindex, and y
   #then print out matlab commands to combine the above variables
into variable data (ie data.y, data.class, etc)
    my (@tempY, @tempClass, @tempCurve, @tempTimeindex);
    my $increment = 0;
    foreach my $numGroup (0..$#compareGroups){
     foreach (@{$compareGroups[$numGroup]}){
         push (@tempY, "$_" . ".y\'");
         push (@tempTimeindex, "$_" . ".timeindex\'");
         push (@tempClass, "ones($_" . ".num, 1)\' + $numGroup");
         push (@tempCurve, "$_" . ".curve\' + $increment");
         if (!(exists($acceptedPatientsNum{$_}))){
          print "###ERROR: cannot find group $_ in accepted
patients number\n";
          die("cannot find group $_ in accepted patients number");
```

```
            }
            $increment = $increment + $acceptedPatientsNum{$_};
            #individual patient's class
            foreach (keys %{$currentClass{$_}}){
              print (OUTFILE "$_" . ".class = " . ($numGroup + 1) .
";\n");
            }
        }
    }
    print (OUTFILE "data.class = [" . join (", ", @tempClass) .
"]\';\n");
    print (OUTFILE "data.curve = [" . join (", ", @tempCurve) .
"]\';\n");
    print (OUTFILE "data.timeindex = [" . join(", ",
@tempTimeindex) . "]\';\n");
    print (OUTFILE "data.y = [" . join(", ", @tempY) . "]\';\n");
} #sub


############################
### sub FLDA_PARAMETERS ###
############################
#INPUT:
##1) @currentGrid (grid begins at $grid[0] and ends at $grid[1])
##2) the interval of the grid
##3) @currentOrderBreaks -> $currentOrderBreaks[0] = order,
$currentOrderBreaks[1] = number of breaks;
#OUTPUT:
## none
#FUNCTIONS:
## print to OUTFILE
## initialized all the necessary FLDA parameters such as:
### userGrid, nbreaks, norder, nbasis, q, G, pert, p, h, tol, maxit
## commands to check p, q and h value making sure that they are
within range
sub FLDA_PARAMETERS{
    my @currentGrid = @{shift(@_)};
    my $currentBy  = shift(@_);
    my @currentOrderBreaks = @{shift (@_)};
    print (OUTFILE "\%nbreaks: number of
breaks\nnbreaks=$currentOrderBreaks[1];\n");
    print (OUTFILE "\%norder: order of the spline
(degree+1)\nnorder=$currentOrderBreaks[0];\n");
    print (OUTFILE "\%nbasis: number of basis
functions\nnbasis=nbreaks+norder-2;\n");
    print (OUTFILE "\%q: dimensionof the spline basis (q-2 equally
spaced knots)\nq=nbasis;\n");
    print (OUTFILE "\%G: number of
cluster\nG=length(unique(data.class));\n");
    print (OUTFILE "\%pert: small adjustment(ridge
regression)\npert=0.1;\n");
```

```
    print (OUTFILE "\%p: rank constraint on the gammas !!
p<=q\np=1;\n");
    print (OUTFILE "\%h: dimension of alpha !! h <= min(p, G-1) G=
number of clusters\nh=1;\n");
    print (OUTFILE "\%minimum relative change for loops (log
likelihood or sum of squares)\ntol = 0.001;\n");
    print (OUTFILE "\%maximum number of iterations\nmaxit=50;\n");
    print (OUTFILE "if p>q\nfprintf(\'error on p >q (Nb of basis) q
= %3i, p = %3i \\n\',q,p)\nreturn\nend\n");
    print (OUTFILE "max_h = min(p,G-1);\nif h > min(p,G-
1)\nfprintf(\'error on h > min(p,K-1)\\th=%3i\\tmin(p,G-
1)=%3i\\n\',h,max_h)\nreturn\nend\n\n");

} #sub



##################
### sub FLDA ####
##################
#INPUT:
##1) name of the current measurement
##2) number of the current comparison
##3) @currentOrderBreaks -> $currentOrderBreaks[0] = order,
$currentOrderBreaks[1] = number of breaks;
#OUTPUT:
## none
#FUNCTIONS:
## print to OUTFILE
## matlab commands for running the fldafit and fldapred using the
previously initiailzed parameters and data
## matlab commands to print the data, fitting parameters, and
prediction results into individual text files
sub FLDA {
    my $currentMeasurement = shift(@_);
    my $currentComparison = shift(@_);
    my @currentOrderBreaks = @{shift(@_)};

    #FLDA
    print (OUTFILE "[flda.parameters, flda.vars, flda.S, flda.FullS,
flda.likenew] = ...\n");
    print (OUTFILE "fldafit(data, norder, nbreaks, h, p, pert, maxit,
userGrid, tol);\n");
    print (OUTFILE "[flda.Calpha, flda.alphahat, flda.classpred,
flda.distance] = ...\n");
    print (OUTFILE "fldapred(flda.parameters, flda.vars, flda.S,
flda.FullS, flda.likenew, data);\n\n");

    #count the error rate
    print (OUTFILE "\%class1 = data.class == 1;\n\%class2 =
data.class == 2;\n");
```

```
    print (OUTFILE "\%error.TP = sum(flda.classpred(class1) ==
1);\n");
    print (OUTFILE "\%error.FN = sum(flda.classpred(class1) ==
2);\n");
    print (OUTFILE "\%error.FP = sum(flda.classpred(class2) ==
1);\n");
    print (OUTFILE "\%error.TN = sum(flda.classpred(class2) ==
2);\n");

    #print out error rate to file "error...txt"
    my $dlmwriteFile = "\%\'.\\error_comparison$currentComparison" .
"_Order$currentOrderBreaks[0]" . "Breaks$currentOrderBreaks[1]" .
".txt\'";
    my $dlmwriteParameter = "\%\'-append\', \'newline\', \'pc\',
\'delimiter\', \'\'";

    #print into text files
    print (OUTFILE "dlmwrite(\'.\\data\\$currentMeasurement" .
"_comparison$currentComparison" . "_Order$currentOrderBreaks[0]" .
"Breaks$currentOrderBreaks[1]" . "_class.txt\', data.class,
\'\\t\')\n");
    print (OUTFILE "dlmwrite(\'.\\data\\$currentMeasurement" .
"_comparison$currentComparison" . "_Order$currentOrderBreaks[0]" .
"Breaks$currentOrderBreaks[1]" . "_curve.txt\', data.curve,
\'\\t\')\n");
    print (OUTFILE "dlmwrite(\'.\\data\\$currentMeasurement" .
"_comparison$currentComparison" . "_Order$currentOrderBreaks[0]" .
"Breaks$currentOrderBreaks[1]" . "_timeindex.txt\', data.timeindex,
\'\\t\')\n");
    print (OUTFILE "dlmwrite(\'.\\data\\$currentMeasurement" .
"_comparison$currentComparison" . "_Order$currentOrderBreaks[0]" .
"Breaks$currentOrderBreaks[1]" . "_y.txt\', data.y, \'\\t\')\n");

    print (OUTFILE "dlmwrite(\'.\\data\\$currentMeasurement" .
"_comparison$currentComparison" . "_Order$currentOrderBreaks[0]" .
"Breaks$currentOrderBreaks[1]" . "_lambdazero.txt\',
flda.parameters.lambdazero, \'\\t\')\n");
    print (OUTFILE "dlmwrite(\'.\\data\\$currentMeasurement" .
"_comparison$currentComparison" . "_Order$currentOrderBreaks[0]" .
"Breaks$currentOrderBreaks[1]" . "_Lambda.txt\',
flda.parameters.Lambda, \'\\t\')\n");
    print (OUTFILE "dlmwrite(\'.\\data\\$currentMeasurement" .
"_comparison$currentComparison" . "_Order$currentOrderBreaks[0]" .
"Breaks$currentOrderBreaks[1]" . "_alpha.txt\',
flda.parameters.alpha, \'\\t\')\n");
    print (OUTFILE "dlmwrite(\'.\\data\\$currentMeasurement" .
"_comparison$currentComparison" . "_Order$currentOrderBreaks[0]" .
"Breaks$currentOrderBreaks[1]" . "_Theta.txt\',
flda.parameters.Theta, \'\\t\')\n");
    print (OUTFILE "dlmwrite(\'.\\data\\$currentMeasurement" .
"_comparison$currentComparison" . "_Order$currentOrderBreaks[0]" .
```

```
"Breaks$currentOrderBreaks[1]" . "_sigma.txt\',
flda.parameters.sigma, \'\\t\')\n");
    print (OUTFILE "dlmwrite(\'.\\data\\$currentMeasurement" .
"_comparison$currentComparison" . "_Order$currentOrderBreaks[0]" .
"Breaks$currentOrderBreaks[1]" . "_D.txt\', flda.parameters.D,
\'\\t\')\n");
    print (OUTFILE "dlmwrite(\'.\\data\\$currentMeasurement" .
"_comparison$currentComparison" . "_Order$currentOrderBreaks[0]" .
"Breaks$currentOrderBreaks[1]" . "_gamma.txt\', flda.vars.gamma,
\'\\t\')\n");

    print (OUTFILE "dlmwrite(\'.\\data\\$currentMeasurement" .
"_comparison$currentComparison" . "_Order$currentOrderBreaks[0]" .
"Breaks$currentOrderBreaks[1]" . "_Calpha.txt\', flda.Calpha,
\'\\t\')\n");
    print (OUTFILE "dlmwrite(\'.\\data\\$currentMeasurement" .
"_comparison$currentComparison" . "_Order$currentOrderBreaks[0]" .
"Breaks$currentOrderBreaks[1]" . "_alphahat.txt\', flda.alphahat,
\'\\t\')\n");
    print (OUTFILE "dlmwrite(\'.\\data\\$currentMeasurement" .
"_comparison$currentComparison" . "_Order$currentOrderBreaks[0]" .
"Breaks$currentOrderBreaks[1]" . "_classpred.txt\', flda.classpred,
\'\\t\')\n");
    print (OUTFILE "dlmwrite(\'.\\data\\$currentMeasurement" .
"_comparison$currentComparison" . "_Order$currentOrderBreaks[0]" .
"Breaks$currentOrderBreaks[1]" . "_distance.txt\', flda.distance,
\'\\t\')\n");

    print (OUTFILE "dlmwrite(\'.\\data\\$currentMeasurement" .
"_comparison$currentComparison" . "_Order$currentOrderBreaks[0]" .
"Breaks$currentOrderBreaks[1]" . "_S.txt\', flda.S, \'\\t\')\n");
    print (OUTFILE "dlmwrite(\'.\\data\\$currentMeasurement" .
"_comparison$currentComparison" . "_Order$currentOrderBreaks[0]" .
"Breaks$currentOrderBreaks[1]" . "_FullS.txt\', flda.FullS,
\'\\t\')\n\n");
} #sub



########################
### sub LEAVEONEOUT ###
########################
#INPUT:
##1) @compareGroups ($compareGroups[0..#1][0..#2]=> group name) #1
is the number of groups to compared and #2 indicates how many
subgroup group #1 is consists of
##2) %class information
##3) %acceptedPatientsPerGroup ($acceptedPatientsPerGroup{group name}
=> number of patients in the group
#OUTPUT:
##1) name of the validation file
#FUNCTIONS:
```

```
## print to file
## FLDA commands to assemble leave-one-out data based the the
previously specified class and comparison information
## Then run fldafit on the training dataset (dataset -1 patient) and
fldapred on the determined parameters and the one patient data
sub LEAVEONEOUT {
    my @compareGroups = @{shift(@_)};
    my %currentClass = %{shift(@_)};
    my %acceptedPatientsPerGroup = %{shift(@_)};
    my $currentComparison = shift (@_);
    my @currentOrderBreaks = @{shift(@_)};
    my $currentMeasurement = shift (@_);
    my %leavePatients;
    my $increament = 0;

    print (OUTFILE
"validation.TP=0;\nvalidation.FN=0;\nvalidation.FP=0;\nvalidation.TN
=0;\n");
    #create leave one out data
    LEAVECLASS: foreach my $leaveClass (keys %currentClass){
     my $leaveClassInCompare = 0;
     foreach (0..$#compareGroups){
         foreach (@{$compareGroups[$_]}){
          if ($_ =~ m/^$leaveClass$/){$leaveClassInCompare ++;}
         }
     }
     if ($leaveClassInCompare == 0){next LEAVECLASS;}
     foreach my $leavePatient (keys %{$currentClass{$leaveClass}}){
         #assemble the class data - the leave patient
         my (@leaveClassY, @leaveClassTimeindex, @leaveClassCurve,
@leaveClassClass);
         my $leaveN = -1;
         foreach my $notLeavePatient (keys
%{$currentClass{$leaveClass}}){
          if (!($notLeavePatient =~ m/^$leavePatient$/)){
              $leaveN ++;
              push (@leaveClassY, "$notLeavePatient" . ".y\'");
              push (@leaveClassCurve, "$notLeavePatient" .
".curve\' + $leaveN");
              push (@leaveClassTimeindex, "$notLeavePatient" .
".timeindex\'");
          }
         }
         print (OUTFILE "tempCurve = [" . join(", ",
@leaveClassCurve) . "]';\n");
         my (@tempY, @tempTimeindex, @tempClass, @tempCurve);
         my $increment = 0;
         foreach my $numGroup (0..$#compareGroups){

          foreach my $group (@{$compareGroups[$numGroup]}){
              if ($group =~ m/^$leaveClass$/){
               push (@tempY, @leaveClassY);
```

```
                push (@tempTimeindex, @leaveClassTimeindex);
                push (@tempCurve, "(tempCurve + $increment)\'");
                push (@tempClass, "ones($group" . ".num -1, 1)\' +
$numGroup");
                $increment = $increment +
$acceptedPatientsPerGroup{$group} -1;
              }
              else{
                push (@tempY, "$group" . ".y\'");
                push (@tempTimeindex, "$group" . ".timeindex\'");
                push (@tempClass, "ones($group" . ".num, 1)\' +
$numGroup");
                push (@tempCurve, "($group" . ".curve +
$increment)'");
                $increment = $increment +
$acceptedPatientsPerGroup{$group};
              } #else
            } #foreach my $group
          } #foreach $numGroup

          #commands to bulid the leavep# data
          print (OUTFILE "leave$leavePatient" . ".class = [" .
join(", ", @tempClass) . "]\';\n");
          print (OUTFILE "leave$leavePatient" . ".curve = [" .
join(", ", @tempCurve) . "]\';\n");
          print (OUTFILE "leave$leavePatient" . ".timeindex = [" .
join(", ", @tempTimeindex) . "]\';\n");
         print (OUTFILE "leave$leavePatient" . ".y = [" . join(", ",
@tempY) . "]\';\n");

          #FLDA commands
          print (OUTFILE "[leave$leavePatient" . ".parameters,
leave$leavePatient" . ".vars, leave$leavePatient" . ".S,
leave$leavePatient" . ".FullS, leave$leavePatient" . ".likenew]
= ...\n");
          print (OUTFILE "fldafit(leave$leavePatient, norder,
nbreaks, h, p, pert, maxit, userGrid, tol);\n");
          print (OUTFILE "[leave$leavePatient" . ".Calpha,
leave$leavePatient" . ".alphahat, leave$leavePatient" . ".classpred,
leave$leavePatient" . ".distance] = ...\n");
          print (OUTFILE "fldapred(leave$leavePatient" .
".parameters, leave$leavePatient" . ".vars, leave$leavePatient" .
".S, leave$leavePatient" . ".FullS, leave$leavePatient" . ".likenew,
$leavePatient);\n");

          #determine the correctness
          print (OUTFILE "if ($leavePatient" . ".class == 1) &&
($leavePatient" . ".class == leave$leavePatient" . ".classpred)\n");
          print (OUTFILE "validation.TP=validation.TP+1;\nend\n");
          print (OUTFILE "if ($leavePatient" . ".class == 1) &&
($leavePatient" . ".class ~= leave$leavePatient" . ".classpred)\n");
          print (OUTFILE "validation.FN=validation.FN+1;\nend\n");
```

```
        print (OUTFILE "if ($leavePatient" . ".class == 2) &&
($leavePatient" . ".class ~= leave$leavePatient" . ".classpred)\n");
        print (OUTFILE "validation.FP=validation.FP+1;\nend\n");
        print (OUTFILE "if ($leavePatient" . ".class == 2) &&
($leavePatient" . ".class == leave$leavePatient" . ".classpred)\n");
        print (OUTFILE "validation.TN=validation.TN+1;\nend\n");


     } #foreach $leavePatient
    } #foreach $leaveClass

    #print out leave-one-out cross-validation result
    my $dlmwriteFile =
"\'.\\validation_comparison$currentComparison" .
"_Order$currentOrderBreaks[0]" . "Breaks$currentOrderBreaks[1]" .
".txt\'";
    my $dlmwriteParameter = "\'-append\', \'newline\', \'pc\',
\'delimiter\', \'\'";
    print (OUTFILE "dlmwrite($dlmwriteFile, \'$currentMeasurement\',
$dlmwriteParameter)\n");
    print (OUTFILE "dlmwrite($dlmwriteFile, validation.TP,
$dlmwriteParameter)\n");
    print (OUTFILE "dlmwrite($dlmwriteFile, validation.FN,
$dlmwriteParameter)\n");
    print (OUTFILE "dlmwrite($dlmwriteFile, validation.FP,
$dlmwriteParameter)\n");
    print (OUTFILE "dlmwrite($dlmwriteFile, validation.TN,
$dlmwriteParameter)\n\n\n");
    return ("$dlmwriteFile");
}



##############################
### sub READ_FLDA_RESULTS ###
##############################
#INPUT:
#1. measurement name
#OUTPUT: none
#FUNCTIONS:
#read in the FLDA results written in subfolder data
#restore all the variables created during the FLDA process
sub READ_FLDA_RESULTS {
    print (OUTFILE "%read in all FLDA parameters back from subfolder
'data'\n");
    my $currentMeasurement = shift (@_);
    my $partialFileName = "$currentMeasurement" . "_comparison" .
shift(@_) . "_Order" . shift(@_) . "Breaks" . shift(@_);

    #print MATLAB command speciify the current measurement
    print (OUTFILE "measurement = \'$partialFileName\';\n");
    #data
```

```
    print (OUTFILE "[data.class] = dlmread([\'.\\data\\\',
measurement, \'_class.txt\'],\'\\t\');\n");
    print (OUTFILE "[data.curve] = dlmread([\'.\\data\\\',
measurement, \'_curve.txt\'],\'\\t\');\n");
    print (OUTFILE "[data.timeindex] = dlmread([\'.\\data\\\',
measurement, \'_timeindex.txt\'],\'\\t\');\n");
    print (OUTFILE "[data.y] = dlmread([\'.\\data\\\', measurement,
\'_y.txt\'],\'\\t\');\n");
    #flda.parameters
    print (OUTFILE "[flda.parameters.lambdazero] =
dlmread([\'.\\data\\\', measurement,
\'_lambdazero.txt\'],\'\\t\');\n");
    print (OUTFILE "[flda.parameters.Lambda] =
dlmread([\'.\\data\\\', measurement, \'_Lambda.txt\'],\'\\t\');\n");
    print (OUTFILE "[flda.parameters.alpha] = dlmread([\'.\\data\\\',
measurement, \'_alpha.txt\'],\'\\t\');\n");
    print (OUTFILE "[flda.parameters.Theta] = dlmread([\'.\\data\\\',
measurement, \'_Theta.txt\'],\'\\t\');\n");
    print (OUTFILE "[flda.parameters.sigma] = dlmread([\'.\\data\\\',
measurement, \'_sigma.txt\'],\'\\t\');\n");
    print (OUTFILE "[flda.parameters.D] = dlmread([\'.\\data\\\',
measurement, \'_D.txt\'],\'\\t\');\n");
    #other FLDA variables
    print (OUTFILE "[flda.vars.gamma] = dlmread([\'.\\data\\\',
measurement, \'_gamma.txt\'],\'\\t\');\n");
    print (OUTFILE "[flda.S] = dlmread([\'.\\data\\\', measurement,
\'_S.txt\'],\'\\t\');\n");
    print (OUTFILE "[flda.FullS] = dlmread([\'.\\data\\\',
measurement, \'_FullS.txt\'],\'\t\');\n");
    print (OUTFILE "[flda.Calpha] = dlmread([\'.\\data\\\',
measurement, \'_Calpha.txt\'],\'\\t\');\n");
    print (OUTFILE "[flda.alphahat] = dlmread([\'.\\data\\\',
measurement, \'_alphahat.txt\'],\'\\t\');\n");
    print (OUTFILE "[flda.classpred] = dlmread([\'.\\data\\\',
measurement, \'_classpred.txt\'],\'\\t\');\n");
    print (OUTFILE "[flda.distance] = dlmread([\'.\\data\\\',
measurement, \'_distance.txt\'],\'\\t\');\n");
    print (OUTFILE "\n");
} #sub read flda results


############################
### sub VALUE_PER_KNOT ###
############################
#INPUT:
#1. \%data
#2. \%class
#3. \@compareGroups
#4. \@currentGrid
#5. \@currentOrderBreaks
#OUTPUT:
```

```
#\%valuePerKnot
#$valuePerKnot{$knot#}{$class#}{expected} => expected number of
values
#$valuePerKnot{$knot#}{$class#}{observed} => observed number of
values
#only include class# that has the smallest observed number of value
for that knot
#FUNCTIONS
sub VALUE_PER_KNOT {
    my %currentData = %{shift(@_)};
    my %currentClass = %{shift(@_)};
    my @currentCompareGroups = @{shift(@_)};
    my @currentGrid = @{shift(@_)};
    my @currentOrderBreaks = @{shift(@_)};
    my $halfInterval = floor((($currentGrid[1] -
$currentGrid[0])/($currentOrderBreaks[1]-1))/2);


    my %tempValuePerKnot;
    my %valuePerKnot;


    for (my $pos = $currentGrid[0]; $pos <= $currentGrid[1]; $pos +=
(($currentGrid[1] - $currentGrid[0])/($currentOrderBreaks[1]-1))){
      print OUTFILElog "GRID: $pos\n";
      foreach my $numClass (0..$#currentCompareGroups){
          foreach my $group (@{$currentCompareGroups[$numClass]}){
            foreach my $patient (keys %{$currentClass{$group}}){
                foreach my $time (keys %{$currentData{$patient}}){
                  print OUTFILElog "TIME: $time\n";
                  if ($time <= $pos + $halfInterval && $time >= $pos
- $halfInterval){
                      $tempValuePerKnot{$pos}{$numClass}{"observed"}
++;
                      print OUTFILElog "adding; knot $pos from clas
$numClass\n";
                  }
                  else{
                      $tempValuePerKnot{$pos}{$numClass}{"observed"}
+= 0;
                  }
                }
                $tempValuePerKnot{$pos}{$numClass}{"expected"}++;
            } #foreach knot
          } #foreach patient in current class
      } #foreach group in current compare groups
    } #foreach compared class in current compare groups

    foreach my $printKnot (sort {$a<=>$b} keys %tempValuePerKnot){
      my $smallestObserved = 100;
      my $smallestObservedClass;
      foreach my $printClass (0..$#currentCompareGroups){
```

```
          if ($tempValuePerKnot{$printKnot}{$printClass}{"observed"}
<= $smallestObserved){
            $smallestObserved = 0 +
$tempValuePerKnot{$printKnot}{$printClass}{"observed"};
            print OUTFILElog "new small observed from knot:
$printKnot is
$tempValuePerKnot{$printKnot}{$printClass}{observed}\n";
            $smallestObservedClass = 0 + $printClass;
          }
          else{
            print OUTFILElog "wrong: knot $printKnot from class
$printClass has " .
$tempValuePerKnot{$printKnot}{$printClass}{"observed"} . "\n";
          }
      }
      $valuePerKnot{$printKnot}{"observed"} = $smallestObserved;
      print OUTFILElog "smallest observed at knot $printKnot is
$smallestObserved from class $smallestObservedClass\n";
      $valuePerKnot{$printKnot}{"expected"} =
$tempValuePerKnot{$printKnot}{$smallestObservedClass}{"expected"};
    }

    return (\%valuePerKnot);
} #sub



############################
### sub WEIGHT_ON_KNOTS ###
############################
#INPUT:
#1. \@grid
#2. \@orderBreaks
#3. $comparison
#4. $measurement
#5. \%valuePerKnot
#OUTPUT: none
#FUNCTIONS:
#print out MATLAB commands needed to determine weight using the
knots distribution
#print
sub WEIGHT_ON_KNOTS {
    my @currentGrid = @{shift(@_)};
    my @currentOrderBreaks = @{shift(@_)};
    my $currentComparison = shift (@_);
    my $currentMeasurement = shift (@_);
    my %currentValuePerKnot = %{shift(@_)};
    my $dlmwriteFile = shift(@_);
    print (OUTFILE "currentTimeIndex = int32([1:(($currentGrid[1]-
$currentGrid[0])/($currentOrderBreaks[1]-1)):($currentGrid[1]-
$currentGrid[0]+1)]\');\n");
    print (OUTFILE "Sij = flda.FullS(currentTimeIndex, :);\n");
```

```
    print (OUTFILE "N = 1;\n");
    print (OUTFILE "h = size(flda.parameters.alpha, 2);\n");
    print (OUTFILE "K = size(flda.parameters.alpha, 1);\n");
    print (OUTFILE "Calpha = zeros(N,h,h);\n");
    print (OUTFILE "n = length(currentTimeIndex);\n");
    print (OUTFILE "Sigma = flda.parameters.sigma * eye (n) + Sij *
flda.parameters.Theta * diag(flda.parameters.D) *
flda.parameters.Theta\' * Sij\';\n");
    print (OUTFILE "InvCalpha = flda.parameters.Lambda\' * Sij\' *
inv(Sigma) * Sij * flda.parameters.Lambda;\n");
    print (OUTFILE "Calpha(1, :, :) = inv(InvCalpha);\n");
    print (OUTFILE "[u,v,w] = size(Calpha);\n");
    print (OUTFILE "Cpart = reshape(Calpha(1,:,:),v,w);\n");
    print (OUTFILE "Weights = Cpart * flda.parameters.Lambda\' *
Sij\' * inv(Sigma);\n");
    print (OUTFILE "\%dlmwrite ($dlmwriteFile,
\'$currentMeasurement\', \'-append\', \'newline\', \'pc\',
\'delimiter\', \'\')\n");
    print (OUTFILE "dlmwrite ($dlmwriteFile, Weights, \'-append\',
\'newline\', \'pc\', \'delimiter\', \'\\t\')\n");
    my @printObserved;
    my @printExpected;
    foreach my $sortedKnot (sort {$a<=>$b} keys
%currentValuePerKnot){
     push (@printObserved,
$currentValuePerKnot{$sortedKnot}{"observed"});
     push (@printExpected,
$currentValuePerKnot{$sortedKnot}{"expected"});

    }
    print (OUTFILE "dlmwrite ($dlmwriteFile, [" . join(", ",
@printObserved) . "], \'-append\', \'newline\', \'pc\',
\'delimiter\', \'\\t\')\n");
    print (OUTFILE "dlmwrite ($dlmwriteFile, [" . join(", ",
@printExpected) . "], \'-append\', \'newline\', \'pc\',
\'delimiter\', \'\\t\')\n");
    print (OUTFILE "\n\n");

} #sub
```

Appendix F. QA on gated data using CD3 as the common intensity

The general variations observed in many CD3-PerCP density plots (Figure F.1) prevent their use as a QA test for the dataset. However, density plots of CD3-PerCP intensity were screened for gate quality control. An example of $CD3^-$ gate is shown Figure F.2 where small peaks with the CD3-PerCP intensity higher than 200 may indicate inclusion of $CD3^+$ cells in the $CD3^-$ gate.

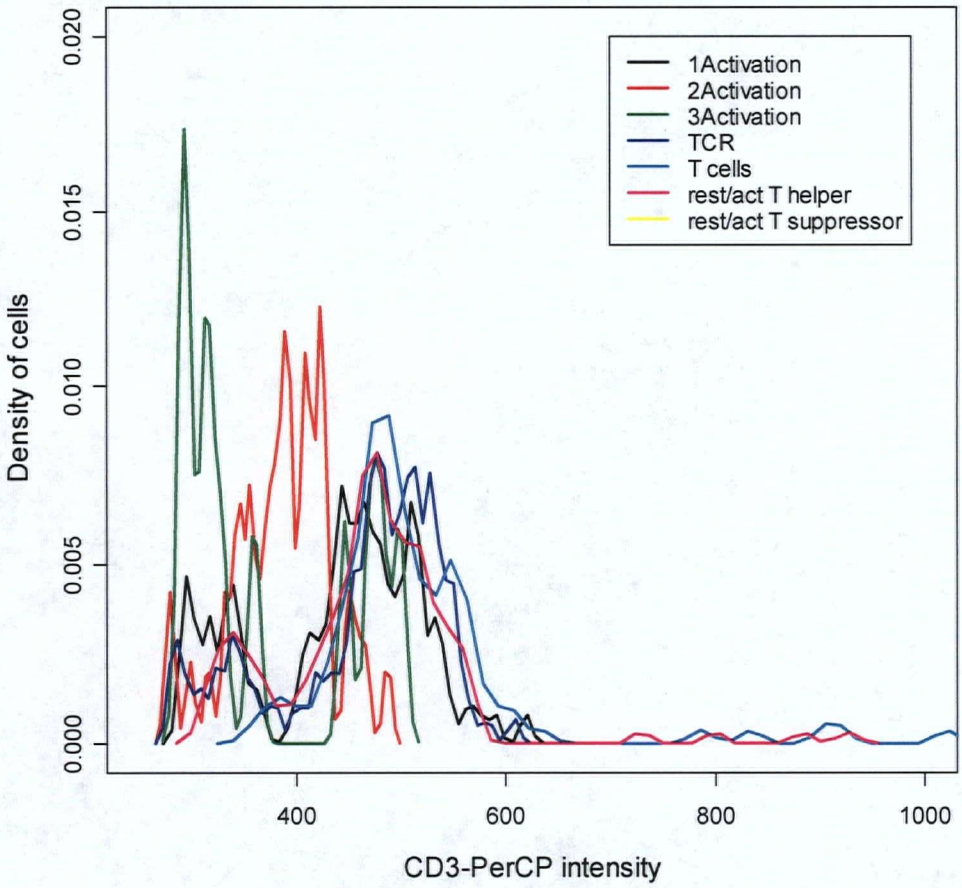


Figure F.1 Density plot of the CD3-PerCP intensity using $CD3^+$ cell population from seven aliquots of patient #6's 76 days post-transplant sample. There is no visible outlier.

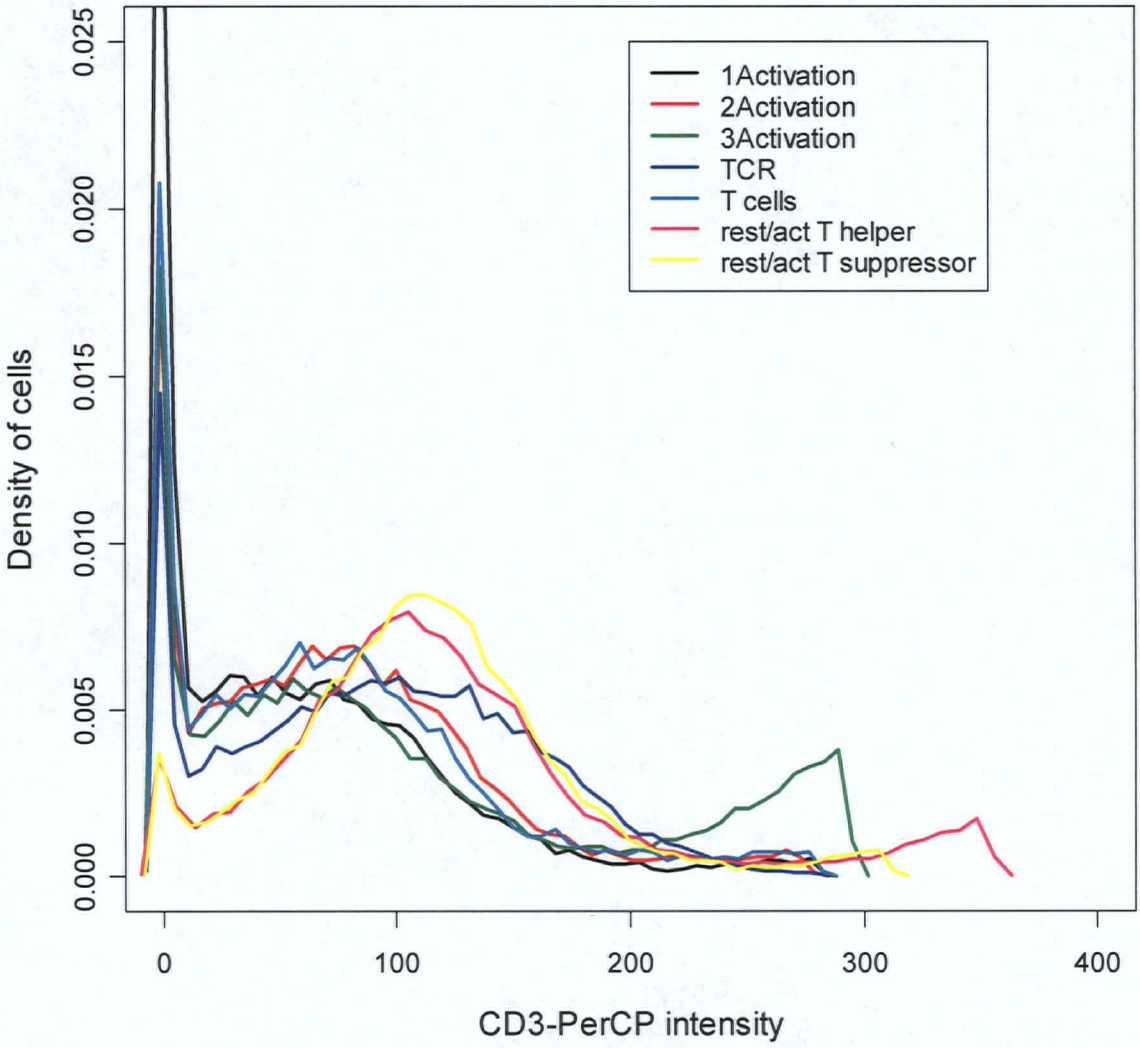


Figure F.2 Density plot of the CD3-PerCP intensity using $CD3^+$ cell population from seven aliquots of patient #6's -6 days post-transplant sample shown as an example of gate quality control.

Appendix G. Other top ranking classifiers for the onset of aGvHD

In the FLDA analysis of the proportion dataset using samples taken between 21 and 0 days prior to aGvHD diagnosis, there were six unique subsets of immune cells with an estimated sensitivity and specificity both higher than 70% (Table H.2). They included the immune cells $CD3^+CD4^+CD8\beta^+$ and $CD3^+CD4^+CD8\beta^+CD8^+$, previously identified as the top ranking classifiers based on samples taken between 7 and 21 days post-transplant (Table 4.1). All the $CD3^+$ and related subsets of immune cells exhibited the same pattern whereas the $CD3^-$ immune cell population exhibited the opposite pattern.

The $CD3^+$ and its related subsets of immune cells such as $CD3^+CD44^-CD25^-$ exhibited a pattern similar to that observed between aGvHD and non-GvHD patients from immune cells $CD3^+CD4^+CD8\beta^+$ between 7 and 21 days post-transplant. Time plots of the immune cells $CD3^+CD44^-CD25^-$ (Figure G.1) are shown as examples. In the FLDA estimated signals time plot for the immune cells $CD3^+CD44^-CD25^-$ (Figure G.1a), the aGvHD patients had higher signals than the non-GvHD patients did. From the raw data time plot from -21 to 21 days from aGvHD diagnosis (Figure G.1b), there was a consistent pattern in the raw data within the same time range. However, this pattern did not carry over after aGvHD was diagnosed.

The $CD3^-$ immune cell population (two readings from aliquots '1Activation' and '2Activation') exhibited a pattern opposite to the $CD3^+$immune cell population. In the FLDA estimated signals time plot, the aGvHD patients had lower signals than the non-GvHD patients did (Figure G.2a). A consistent pattern was also observed in the raw data time point within the same time range (Figure G.2b).

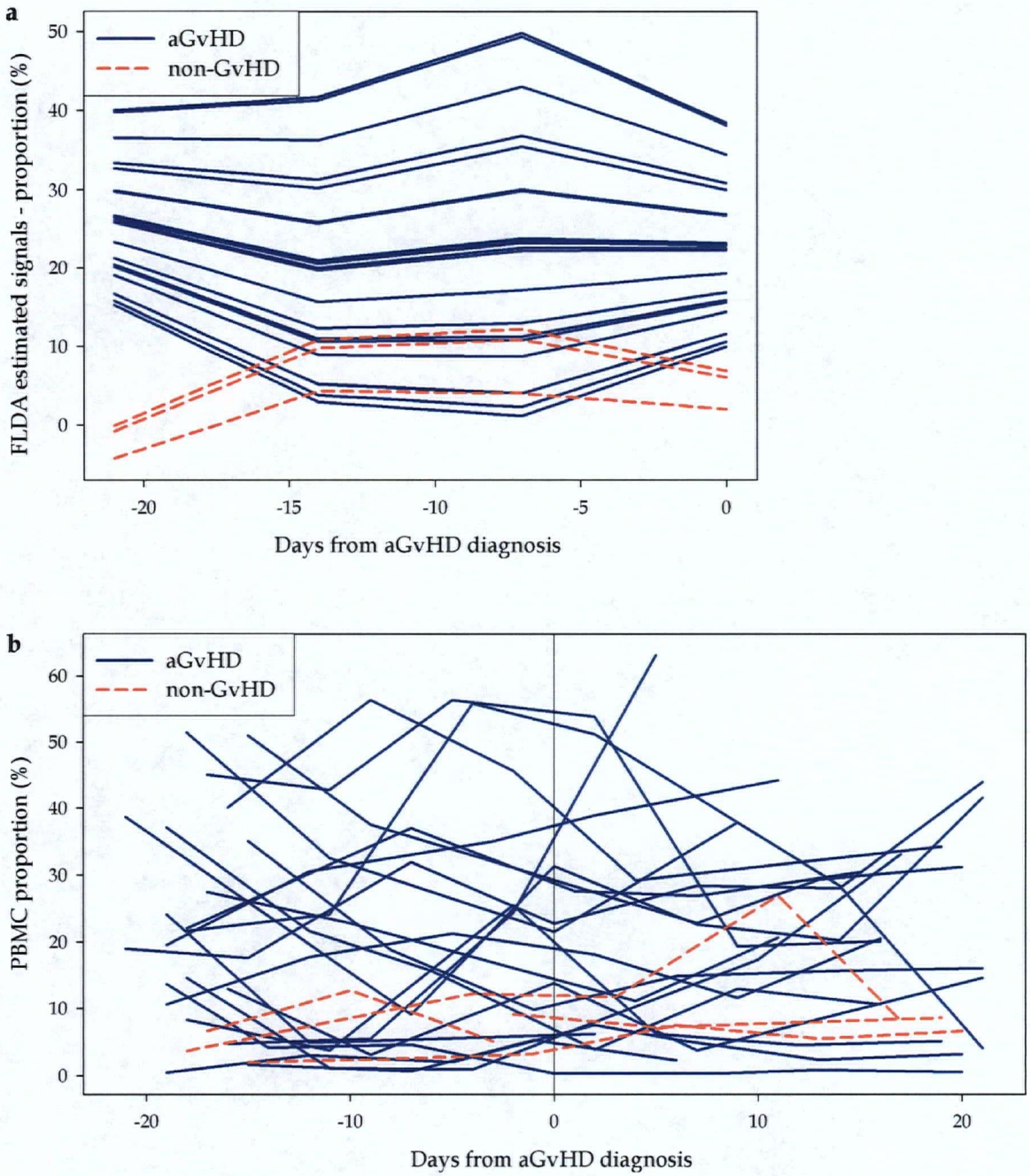


Figure G.1 Time plot of the FLDA estimated signals (panel a) based on samples taken between -21 and 0 days from aGvHD and time plot of the raw data (panel b) based on samples taken between -21 and 21 days from aGvHD diagnosis for the immune cells $CD3^{+}CD44^{-}CD25^{-}$ in proportion to PBMC. The aGvHD diagnosis day is labelled at day 0.

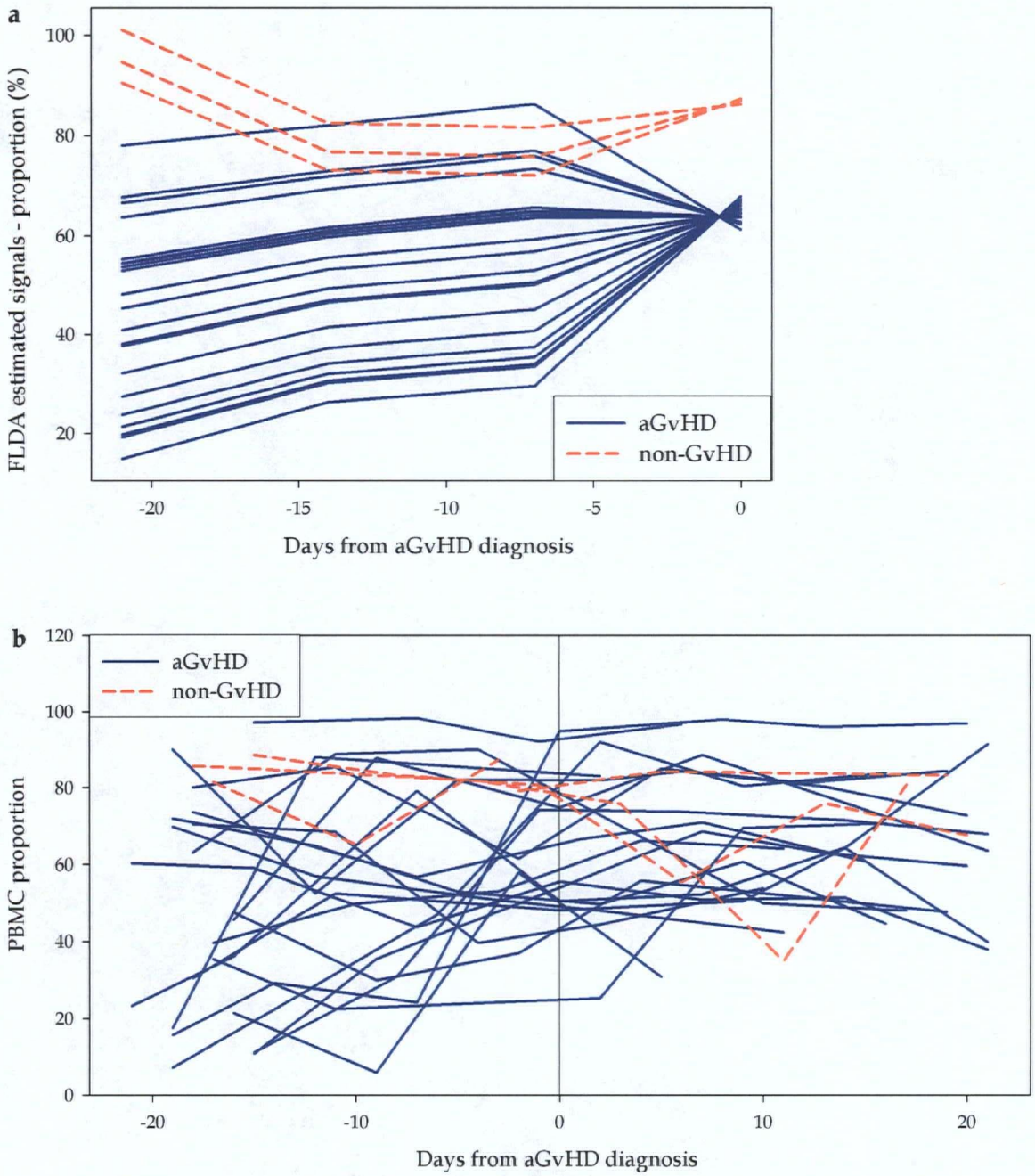


Figure G.2 Time plot of the FLDA estimated signals (panel a) based on samples taken between -21 and to 0 days from aGvHD and time plot of the raw data (panel b) based on samples taken between -21 and to 21 days from aGvHD diagnosis for the immune cells CD3 – (aliquot '1Activation') in proportion to PBMC. The date of aGvHD diagnosis is labelled as day 0.

In the FLDA analysis of the proportion dataset using samples taken between 0 and 21 days from aGvHD diagnosis, only three classifiers were found to have sensitivity and specificity both higher than 70% (Table H.3). They were $CD2^{dim}CD16^{+}CD56^{-}CD3^{-}$, $CD3^{+}CD4^{int}$ (from aliquot '3Activation') and $CD3^{+}CD4^{+}CD8\beta CD8^{+}$ in proportion to the $CD3^{+}$ cells (not PBMC). All three classifiers exhibited similar patterns to that of the $CD3^{+}$ T cells described in the previous section.

The FLDA classifier built from immune cells $CD2^{dim}CD16^{+}CD56^{-}CD3^{-}$ using samples taken between 0 and 21 days from aGvHD diagnosis had an estimated 78% sensitivity and 100% specificity. The FLDA estimated signals time plot (Figure G.3a) displayed a pattern of higher signals from the aGvHD patients compared to the non-GvHD patients, which was consistent with its corresponding raw data time plot (Figure G.3b). However, this pattern was not observed before aGvHD diagnosis (Figure G.3b).

The FLDA classifier built from immune cells $CD3^{+}CD4^{int}$ (from aliquot '3Activation') using samples taken between 0 and 21 days from aGvHD diagnosis had an estimated 72% sensitivity and 100% specificity. The FLDA estimated signals time plot (Figure G.4a) displayed a pattern of higher signals from the aGvHD patients compared to the non-GvHD patients. The separation between the two groups of patients was smaller than the one observed in the FLDA estimated signals for the immune cells $CD3^{+}CD4^{+}CD8\beta^{+}$ based on samples taken between 7 and 21 days post-transplant (Figure 4.4). Nevertheless, this pattern was consistent with its corresponding raw data time plot (Figure G.4b). A similar pattern was also observed in the raw data time plot before the aGvHD diagnosis, outside the analyzed time range. However, FLDA classifier using the same subset of immune cells based samples taken between 21 and 0 days prior to aGvHD diagnosis had only an estimated 57% sensitivity and 67% specificity (Table H.2).

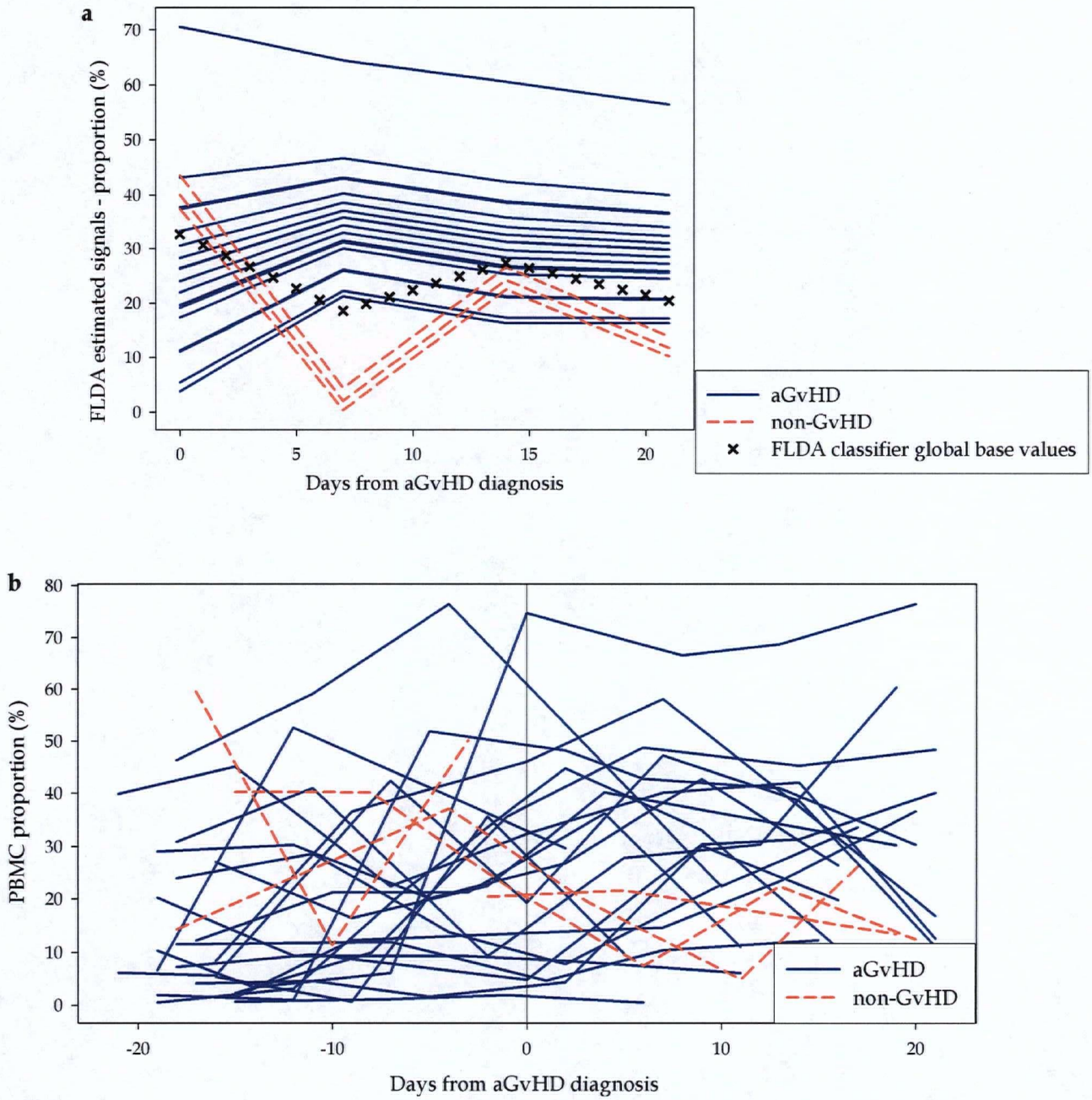


Figure G.3 Time plot of the FLDA estimated signals (panel a) based on samples taken between 0 and 21 days from aGvHD and time plot of the raw data (panel b) based on samples taken between -21 and 21 days from aGvHD diagnosis for the immune cells $CD2^{dim}CD16^{+}CD56^{-}CD3^{-}$ in proportion to PBMC. The date of aGvHD diagnosis is labelled as day 0.

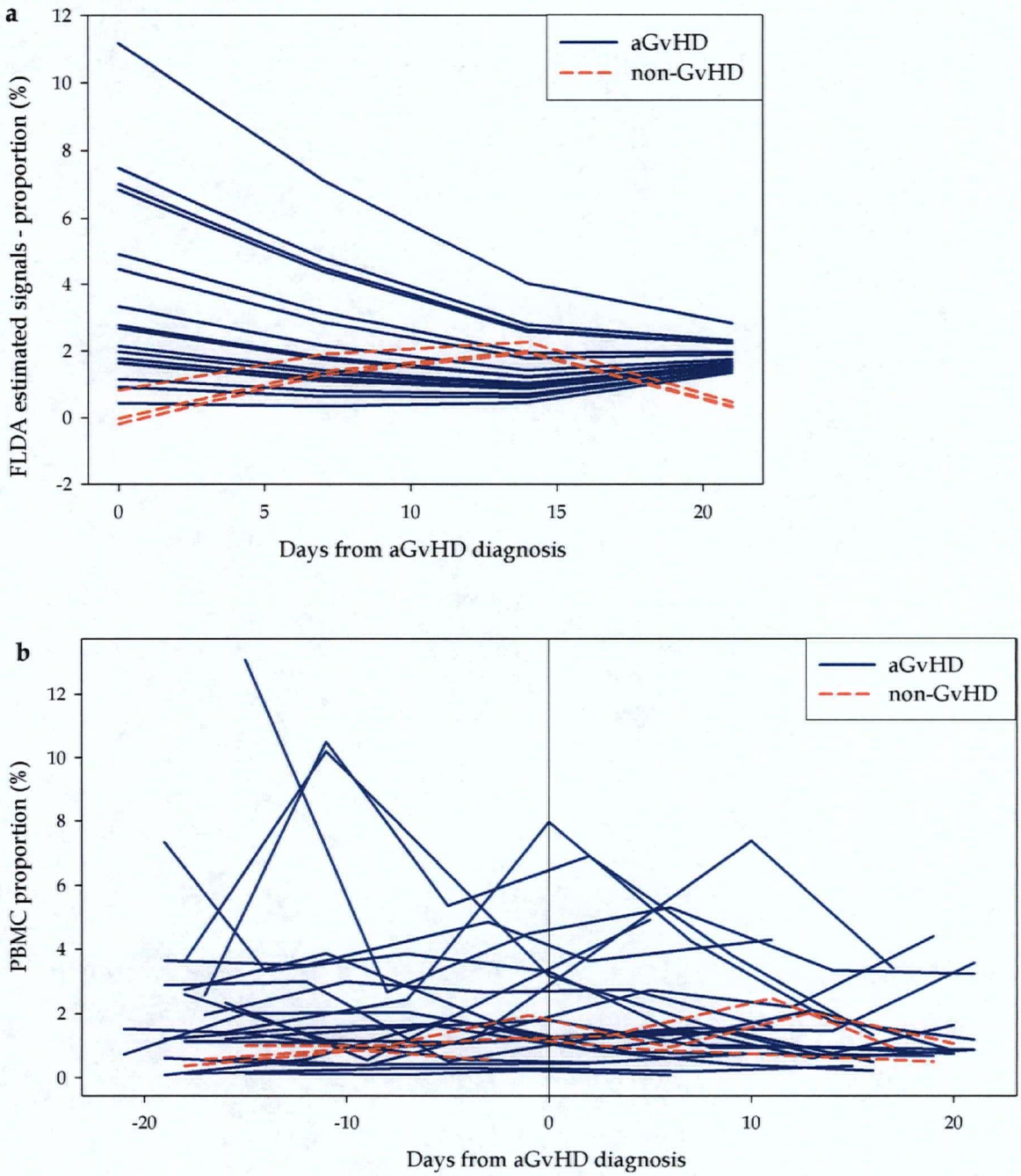


Figure G.4 Time plot of the FLDA estimated signals (panel a) based on samples taken between 0 and 21 days from aGvHD and time plot of the raw data (panel b) based on samples taken between -21 and 21 days from aGvHD diagnosis for the immune cells $CD3^{+}CD4^{int}$ (aliquot '3Activation') in proportion to PBMC. The date of aGvHD diagnosis is labelled as day 0.

The FLDA classifier built from the proportion of the immune cells $CD3^+CD4^+CD8\beta^+CD8^+$ relative to the total $CD3^+$ cells (instead of the usual PBMCs) using samples between 0 and 21 days from aGvHD diagnosis had an estimated 72% sensitivity and 100% specificity. Like most of classifiers previously described, it exhibited a pattern where both FLDA signals and the raw $CD3^+$ cells proportion were higher from the aGvHD patients, compared to the non-GvHD patients (Figure G.5). Even though the immune cell abundance was recorded in proportion to $CD3^+$ cells, it exhibited a similar pattern to $CD3^+CD4^+CD8\beta^+CD8^+$ in proportion to PBMC (Figure 4.8).

In the FLDA analysis of the concentration dataset using samples taken from all three time ranges, there were only three classifiers with their estimated sensitivity and specificity both higher than 70% (Tables H.4 - H.6). Overall, there was very little correlation between the classifiers accuracies from the proportion and concentration datasets ($r = 0.02$). The top ranking classifiers from the concentration dataset were:

1. .$CD2^+CD16^+$, based on samples taken between 7 and 21 days post-transplant (data not shown)
2. $CD3^-CD44^+CD25^+$, based on samples taken between 21 and 0 days prior to aGvHD diagnosis (data not shown)
3. $CD45^+CD33^-$, based on samples taken between 21 and 0 days prior to aGvHD diagnosis (Figure G.6)

These classifiers were all inconsistent due to pattern outliers as described in details in Chapter 4.

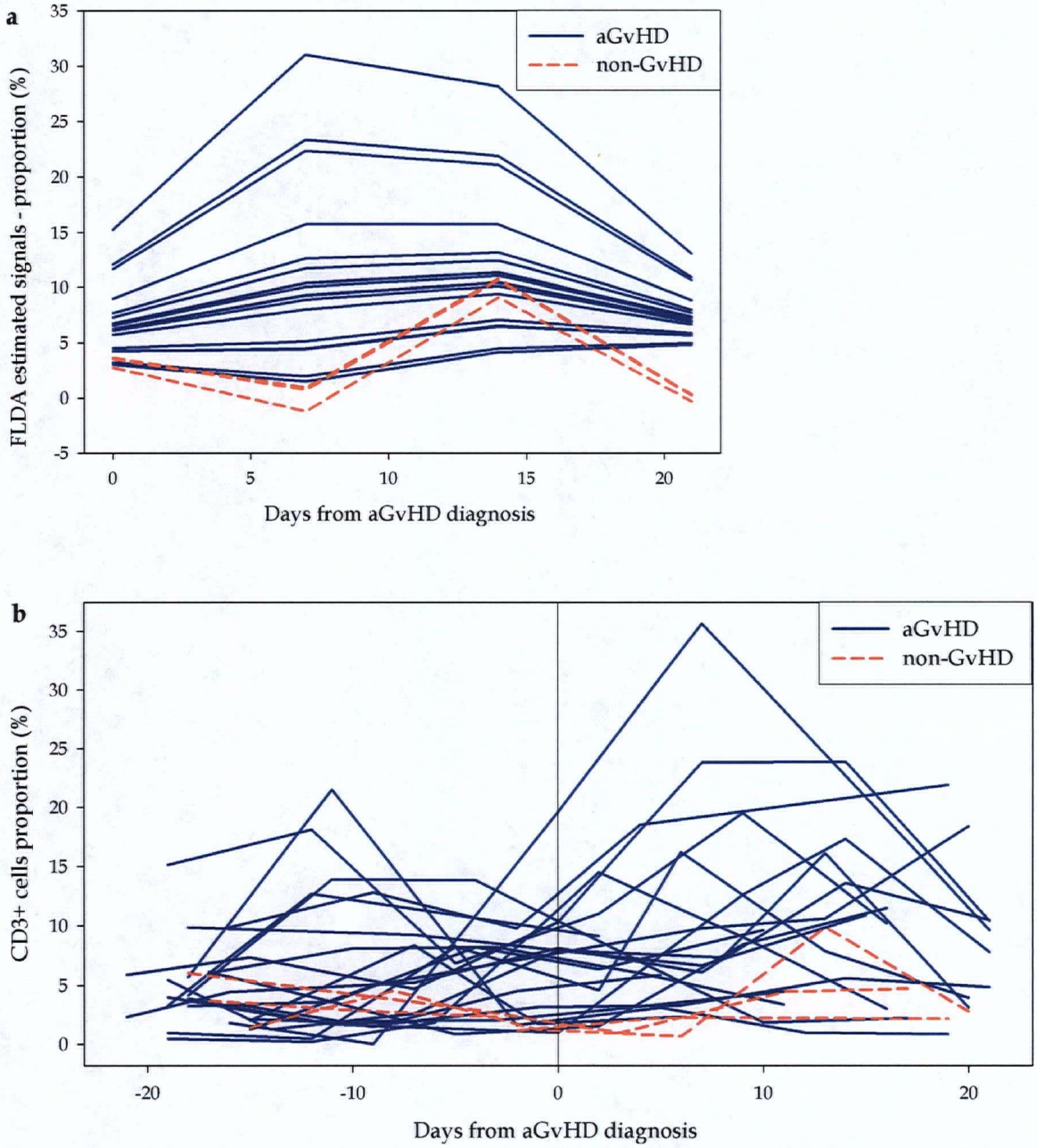


Figure G.5 Time plot of the FLDA estimated signals (panel a) based on samples taken between 0 and 21 days from aGvHD and time plot of the raw data (panel b) based on samples taken between -21 and 21 days from aGvHD diagnosis for the new subset of immune cells $CD3^{+}CD4^{+}CD8\beta^{+}CD8^{+}$ in proportion to $CD3^{+}$ cell population. The aGvHD diagnosis day is labelled at day 0.

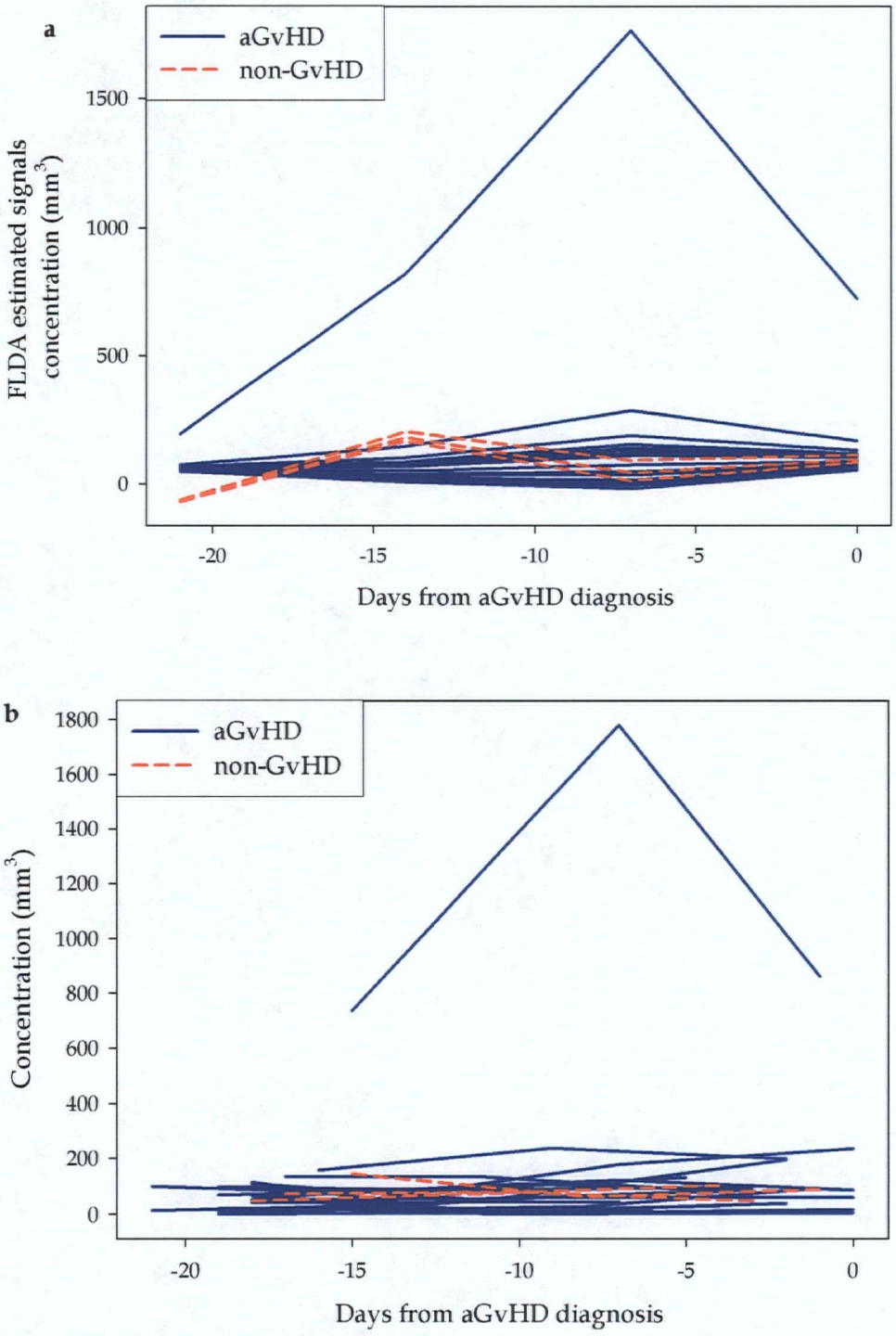


Figure G.6 Time plots of the FLDA estimated signals (panel a) and the raw data (panel b) based on samples taken between 21 and 0 days prior to aGvHD diagnosis for the immune cells $CD45^{+}CD33^{-}$ in concentration (mm^3).

Appendix H. Summaries of LOOCV results for the FLDA analyses between aGvHD and non-GvHD patients

Table H.1 Validation results for qualified subsets of immune cells in proportion to PBMC (%) from the FLDA classification between aGvHD and non-GvHD patients using samples taken from 7 to 21 days post-transplant.

Immune cells	Aliquot	Sensitivity (%)	Specificity (%)	Accuracy (%)
$CD3^-CD44^+CD25^-$	1Activation	86	33	79
$CD3^-CD44^+CD25^+CD69^+$		100	0	85
$CD3^-CD44^+CD25^+$		94	0	80
$CD3^-CD44^-CD25^-$		67	67	67
$CD3^-$		81	0	71
$CD3^+CD44^-CD25^+$		57	33	54
$CD3^+CD44^+CD25^+CD69^+$		24	67	29
$CD3^+CD44^+CD25^+$		81	0	71
$CD3^+CD44^-CD25^-$		81	33	75
$CD3^+$		90	33	83
$CD3^-CD4^{dim}$	2Activation	86	0	75
$CD3^-CD8^{low}$		57	0	50
$CD3^-CD4^-CD8^-$		95	33	88
$CD3^-$		86	0	75
$CD3^+CD4^{br}$		86	33	79
$CD3^+CD4^{int}$		81	100	83
$CD3^+CD8^{br}$		81	33	75
$CD3^+CD8^{dim}$		67	0	58
$CD3^+$		86	33	79

Immune cells	Aliquot	Sensitivity (%)	Specificity (%)	Accuracy (%)
$CD3^-CD4^{dim}$	3Activation	81	0	71
$CD3^-CD8^{low}$		62	0	54
$CD3^-CD4^-CD8^-$		95	33	88
$CD3^-$		86	0	75
$CD3^+CD4^{br}$		86	33	79
$CD3^+CD4^{int}$		76	67	75
$CD3^+CD8^{br}$		81	33	75
$CD3^+CD8^{dim}$		81	33	75
$CD3^+$		90	33	83
$CD22^+CD20^+$	B cells	95	0	83
$CD22^+$		100	0	88
$CD33^+CD45^{dim}CD15^{low}CD14^{low}$	Myeloids	100	67	96
$CD33^+CD45^{dim}CD15^+CD14^-$		62	0	54
$CD33^+CD45^{dim}CD15^+CD14^+$		67	0	58
$CD33^+CD45^{dim}$		81	0	71
$CD33^+CD45^+CD15^+CD14^+$		90	0	79
$CD33^+CD45^+$		95	0	83
$CD45^+CD33^-CD15^+CD14^-$		67	0	58
$CD45^+CD33^-$		86	0	75
$CD2^{dim}CD16^+CD3^+CD56^-$	NK cells	81	33	75
$CD2^{dim}CD16^+CD56^+CD3^-$		90	100	92
$CD2^{dim}CD16^+CD56^-CD3^-$		90	0	79
$CD2^{dim}CD16^+$		90	0	79
$CD2^-CD16^+CD3^+CD56^-$		86	33	79
$CD2^-CD16^+CD56^+CD3^-$		81	33	75
$CD2^-CD16^+CD56^-CD3^-$		67	33	62

Immune cells	Aliquot	Sensitivity (%)	Specificity (%)	Accuracy (%)
$CD2^{-}CD16^{+}$	NK cells	52	67	54
$CD2^{+}CD16^{-}CD3^{+}CD56^{-}$		86	33	79
$CD2^{+}CD16^{-}CD56^{+}CD3^{-}$		71	0	62
$CD2^{+}CD16^{-}CD56^{-}CD3^{-}$		71	0	62
$CD2^{+}CD16^{-}$		90	33	83
$CD2^{+}CD16^{+}CD3^{+}CD56^{-}$		67	67	67
$CD2^{+}CD16^{+}CD56^{+}CD3^{-}$		71	0	62
$CD2^{+}CD16^{+}CD56^{-}CD3^{-}$		95	33	88
$CD2^{+}CD16^{+}$		76	0	67
$CD3^{-}CD4^{low}CD8\beta^{low}$	T cells	90	0	79
$CD3^{-}CD8\beta^{dim}CD8^{-}$		90	0	79
$CD3^{-}CD8^{+}CD8\beta^{-}$		67	0	58
$CD3^{-}$		86	0	75
$CD3^{+}CD4^{+}CD8\beta^{-}$		90	33	83
$CD3^{+}CD4^{+}CD8\beta^{+}$		86	100	88
$CD3^{+}CD8\beta^{dim}CD8^{-}$		90	0	79
$CD3^{+}CD8\beta^{+}CD4^{-}$		81	33	75
$CD3^{+}CD8\beta^{+}CD8^{low}$		57	0	50
$CD3^{+}CD8\beta^{+}CD8^{+}$		81	33	75
$CD3^{+}CD8^{+}CD8\beta^{-}$		81	33	75
$CD3^{+}$		90	33	83
$CD3^{+}CD4^{+}CD8\beta^{+}CD8^{+}$		71	100	75
$CD3^{+}CD4^{+}CD8\beta^{+}CD8^{+}$ (proportion of$CD3^{+}$ cells)		48	33	46

Immune cells	Aliquot	Sensitivity (%)	Specificity (%)	Accuracy (%)
$CD3^-$	TCR	86	0	75
$CD3^-CD5^-TCRab^+TCRgd^-$		76	0	67
$CD3^-CD5^-TCRab^+$		86	0	75
$CD3^-CD5^-TCRab^+TCRgd^+$		50	0	43
$CD3^-CD5^-TCRab^+TCRgd^-$		85	0	74

Table H.2 Validation results for qualified subsets of immune cells in proportion to PBMC (%) from the FLDA classification between aGvHD and non-GvHD patients using samples taken between 21 and 0 days prior to aGvHD diagnosis.

Immune cells	Aliquot	Sensitivity (%)	Specificity (%)	Accuracy (%)
$CD3^-CD44^+CD25^-$	1Activation	90	33	83
$CD3^-CD44^+CD25^+CD69^+$		76	33	70
$CD3^-CD44^+CD25^+$		88	33	80
$CD3^-CD44^-CD25^-$		57	33	54
$CD3^-$		71	100	75
$CD3^+CD44^-CD25^+$		62	33	58
$CD3^+CD44^+CD25^+CD69^+$		57	100	62
$CD3^+CD44^+CD25^+$		43	0	38
$CD3^+CD44^-CD25^-$		76	100	79
$CD3^+$		71	100	75
$CD3^-CD4^{dim}$	2Activation	71	0	62
$CD3^-CD8^{low}$		67	0	58
$CD3^-CD4^-CD8^-$		62	0	54
$CD3^-$		71	100	75
$CD3^+CD4^{br}$		62	67	62
$CD3^+CD4^{int}$		57	100	62
$CD3^+CD8^{br}$		76	100	79
$CD3^+CD8^{dim}$		43	33	42
$CD3^+$		71	100	75

Immune cells	Aliquot	Sensitivity (%)	Specificity (%)	Accuracy (%)
$CD3^{-}CD4^{dim}$	3Activation	71	0	62
$CD3^{-}CD8^{low}$		71	0	62
$CD3^{-}CD4^{-}CD8^{-}$		67	0	58
$CD3^{-}$		67	100	71
$CD3^{+}CD4^{br}$		62	67	62
$CD3^{+}CD4^{int}$		57	67	58
$CD3^{+}CD8^{br}$		67	100	71
$CD3^{+}CD8^{dim}$		38	0	33
$CD3^{+}$		67	100	71
$CD22^{+}CD20^{+}$	B cells	90	0	79
$CD22^{+}$		81	0	71
$CD33^{+}CD45^{dim}CD15^{low}CD14^{low}$	Myeloids	90	33	83
$CD33^{+}CD45^{dim}CD15^{+}CD14^{-}$		95	33	88
$CD33^{+}CD45^{dim}CD15^{+}CD14^{+}$		81	0	71
$CD33^{+}CD45^{dim}$		90	0	79
$CD33^{+}CD45^{+}CD15^{+}CD14^{+}$		76	0	67
$CD33^{+}CD45^{+}$		71	0	62
$CD45^{+}CD33^{-}CD15^{+}CD14^{-}$		81	0	71
$CD45^{+}CD33^{-}$		71	0	62
$CD2^{dim}CD16^{+}CD3^{+}CD56^{-}$	NK cells	43	67	46
$CD2^{dim}CD16^{+}CD56^{+}CD3^{-}$		52	33	50
$CD2^{dim}CD16^{+}CD56^{-}CD3^{-}$		76	0	67
$CD2^{dim}CD16^{+}$		67	0	58
$CD2^{-}CD16^{+}CD3^{+}CD56^{-}$		90	33	83
$CD2^{-}CD16^{+}CD56^{+}CD3^{-}$		86	33	79
$CD2^{-}CD16^{+}CD56^{-}CD3^{-}$		38	33	38

Immune cells	Aliquot	Sensitivity (%)	Specificity (%)	Accuracy (%)
$CD2^{-}CD16^{+}$	NK cells	38	0	33
$CD2^{+}CD16^{-}CD3^{+}CD56^{-}$		67	67	67
$CD2^{+}CD16^{-}CD56^{+}CD3^{-}$		95	0	83
$CD2^{+}CD16^{-}CD56^{-}CD3^{-}$		67	0	58
$CD2^{+}CD16^{-}$		86	33	79
$CD2^{+}CD16^{+}CD3^{+}CD56^{-}$		52	67	54
$CD2^{+}CD16^{+}CD56^{+}CD3^{-}$		90	33	83
$CD2^{+}CD16^{+}CD56^{-}CD3^{-}$		86	67	83
$CD2^{+}CD16^{+}$		76	0	67
$CD3^{+}CD4^{+}CD8\beta^{+}CD8^{+}$	T cells	71	100	75
$CD3^{+}CD4^{+}CD8\beta^{+}CD8^{+}$ (proportion of $CD3^{+}$ cells)		57	100	62
$CD3^{-}CD4^{low}CD8\beta^{low}$		76	0	67
$CD3^{-}CD8\beta^{dim}CD8^{-}$		90	67	88
$CD3^{-}CD8^{+}CD8\beta^{-}$		86	0	75
$CD3^{-}$		67	100	71
$CD3^{+}CD4^{+}CD8\beta^{-}$		57	67	58
$CD3^{+}CD4^{+}CD8\beta^{+}$		67	100	71
$CD3^{+}CD8\beta^{dim}CD8^{-}$		48	100	54
$CD3^{+}CD8\beta^{+}CD4^{-}$		67	100	71
$CD3^{+}CD8\beta^{+}CD8^{low}$		81	67	79
$CD3^{+}CD8\beta^{+}CD8^{+}$		67	100	71
$CD3^{+}CD8^{+}CD8\beta^{-}$		71	100	75
$CD3^{+}$		67	100	71

Immune cells	Aliquot	Sensitivity (%)	Specificity (%)	Accuracy (%)
$CD3^-$	TCR	76	100	79
$CD3^-CD5^-TCRab^+TCRgd^-$		86	0	75
$CD3^-CD5^-TCRab^+$		76	33	71
$CD3^-CD5^-TCRab^+TCRgd^+$		80	0	70
$CD3^-CD5^-TCRab^+TCRgd^-$		65	33	61

Table H.3 Validation results for qualified subsets of immune cells in proportion to PBMC (%) from the FLDA classification between aGvHD and non-GvHD patients using samples taken between 0 and 21 days from aGvHD diagnosis.

Immune cells	Aliquot	Sensitivity (%)	Specificity (%)	Accuracy (%)
$CD3^-CD44^+CD25^-$	1Activation	72	33	67
$CD3^-CD44^+CD25^+CD69^+$		62	67	63
$CD3^-CD44^+CD25^+$		81	0	68
$CD3^-CD44^-CD25^-$		67	0	57
$CD3^-$		94	0	81
$CD3^+CD44^-CD25^+$		44	33	43
$CD3^+CD44^+CD25^+CD69^+$		56	100	62
$CD3^+CD44^+CD25^+$		78	0	67
$CD3^+CD44^-CD25^-$		72	33	67
$CD3^+$		94	33	86
$CD3^-CD4^{dim}$	2Activation	78	33	71
$CD3^-CD8^{low}$		83	0	71
$CD3^-CD4^-CD8^-$		94	0	81
$CD3^-$		94	0	81
$CD3^+CD4^{br}$		89	0	76
$CD3^+CD4^{int}$		67	100	71
$CD3^+CD8^{br}$		50	67	52
$CD3^+CD8^{dim}$		94	33	86
$CD3^+$		94	0	81

Immune cells	Aliquot	Sensitivity (%)	Specificity (%)	Accuracy (%)
$CD3^{-}CD4^{dim}$	3Activation	72	67	71
$CD3^{-}CD8^{low}$		61	0	52
$CD3^{-}CD4^{-}CD8^{-}$		100	0	86
$CD3^{-}$		94	0	81
$CD3^{+}CD4^{br}$		83	0	71
$CD3^{+}CD4^{int}$		72	100	76
$CD3^{+}CD8^{br}$		61	67	62
$CD3^{+}CD8^{dim}$		89	33	81
$CD3^{+}$		94	0	81
$CD22^{+}CD20^{+}$	B cells	94	67	90
$CD22^{+}$		94	0	81
$CD33^{+}CD45^{dim}CD15^{low}CD14^{low}$	Myeloids	72	33	67
$CD33^{+}CD45^{dim}CD15^{+}CD14^{-}$		44	33	43
$CD33^{+}CD45^{dim}CD15^{+}CD14^{+}$		33	100	43
$CD33^{+}CD45^{dim}$		33	100	43
$CD33^{+}CD45^{+}CD15^{+}CD14^{+}$		89	0	76
$CD33^{+}CD45^{+}$		56	33	52
$CD45^{+}CD33^{-}CD15^{+}CD14^{-}$		100	0	86
$CD45^{+}CD33^{-}$		72	0	62
$CD2^{dim}CD16^{+}CD3^{+}CD56^{-}$	NK cells	83	0	71
$CD2^{dim}CD16^{+}CD56^{+}CD3^{-}$		50	0	43
$CD2^{dim}CD16^{+}CD56^{-}CD3^{-}$		78	100	81
$CD2^{dim}CD16^{+}$		78	33	71
$CD2^{-}CD16^{+}CD3^{+}CD56^{-}$		94	33	86
$CD2^{-}CD16^{+}CD56^{+}CD3^{-}$		72	0	62
$CD2^{-}CD16^{+}CD56^{-}CD3^{-}$		89	33	81

Immune cells	Aliquot	Sensitivity (%)	Specificity (%)	Accuracy (%)
$CD2^{-}CD16^{+}$	NK cells	89	33	81
$CD2^{+}CD16^{-}CD3^{+}CD56^{-}$		89	33	81
$CD2^{+}CD16^{-}CD56^{+}CD3^{-}$		89	33	81
$CD2^{+}CD16^{-}CD56^{-}CD3^{-}$		67	0	57
$CD2^{+}CD16^{-}$		89	0	76
$CD2^{+}CD16^{+}CD3^{+}CD56^{-}$		50	67	52
$CD2^{+}CD16^{+}CD56^{+}CD3^{-}$		83	33	76
$CD2^{+}CD16^{+}CD56^{-}CD3^{-}$		94	0	81
$CD2^{+}CD16^{+}$		28	0	24
$CD3^{+}CD4^{+}CD8\beta^{+}CD8^{+}$	T cells	67	67	67
$CD3^{+}CD4^{+}CD8\beta^{+}CD8^{+}$ (proportion of $CD3^{+}$ cells)		72	100	76
$CD3^{-}CD4^{low}CD8\beta^{low}$		78	67	76
$CD3^{-}CD8\beta^{dim}CD8^{-}$		78	0	67
$CD3^{-}CD8^{+}CD8\beta^{-}$		72	0	62
$CD3^{-}$		94	33	86
$CD3^{+}CD4^{+}CD8\beta^{-}$		89	0	76
$CD3^{+}CD4^{+}CD8\beta^{+}$		72	67	71
$CD3^{+}CD8\beta^{dim}CD8^{-}$		56	100	62
$CD3^{+}CD8\beta^{+}CD4^{-}$		56	33	52
$CD3^{+}CD8\beta^{+}CD8^{low}$		83	0	71
$CD3^{+}CD8\beta^{+}CD8^{+}$		56	67	57
$CD3^{+}CD8^{+}CD8\beta^{-}$		72	33	67
$CD3^{+}$		100	0	86

Immune cells	Aliquot	Sensitivity (%)	Specificity (%)	Accuracy (%)
$CD3^-$	TCR	94	0	81
$CD3^-CD5^-TCRab^+TCRgd^-$		89	0	76
$CD3^-CD5^-TCRab^+$		67	33	62
$CD3^-CD5^-TCRab^+TCRgd^+$		39	33	38
$CD3^-CD5^-TCRab^+TCRgd^-$		72	0	62

Table H.4 Validation results for qualified subsets of immune cells in concentration (mm^3) from the FLDA classification between aGvHD and non-GvHD patients using samples taken from 7 to 21 days post-transplant.

Immune cells	Aliquot	Sensitivity (%)	Specificity (%)	Accuracy (%)
$CD3^-CD44^+CD25^-$	1Activation	76	33	71
$CD3^-CD44^+CD25^+CD69^+$		100	0	85
$CD3^-CD44^+CD25^+$		100	0	85
$CD3^-CD44^-CD25^-$		43	67	46
$CD3^-$		76	67	75
$CD3^+CD44^-CD25^+$		52	67	54
$CD3^+CD44^+CD25^+CD69^+$		81	67	79
$CD3^+CD44^+CD25^+$		71	0	62
$CD3^+CD44^-CD25^-$		43	67	46
$CD3^+$		43	67	46
$CD3^-CD4^{dim}$	2Activation	67	67	67
$CD3^-CD8^{low}$		67	33	62
$CD3^-CD4^-CD8^-$		81	33	75
$CD3^-$		71	67	71
$CD3^+CD4^{br}$		81	67	79
$CD3^+CD4^{int}$		33	100	42
$CD3^+CD8^{br}$		43	67	46
$CD3^+CD8^{dim}$		33	67	38
$CD3^+$		52	67	54

Immune cells	Aliquot	Sensitivity (%)	Specificity (%)	Accuracy (%)
$CD3^{-}CD4^{dim}$	3Activation	67	67	67
$CD3^{-}CD8^{low}$		71	33	67
$CD3^{-}CD4^{-}CD8^{-}$		86	33	79
$CD3^{-}$		71	67	71
$CD3^{+}CD4^{br}$		86	67	83
$CD3^{+}CD4^{int}$		43	67	46
$CD3^{+}CD8^{br}$		52	67	54
$CD3^{+}CD8^{dim}$		57	33	54
$CD3^{+}$		57	67	58
$CD22^{+}CD20^{+}$	B cells	62	100	67
$CD22^{+}$		95	33	88
$CD33^{+}CD45^{dim}CD15^{low}CD14^{low}$	Myeloids	100	33	92
$CD33^{+}CD45^{dim}CD15^{+}CD14^{-}$		71	33	67
$CD33^{+}CD45^{dim}CD15^{+}CD14^{+}$		71	0	62
$CD33^{+}CD45^{dim}$		71	0	62
$CD33^{+}CD45^{+}CD15^{+}CD14^{+}$		76	33	71
$CD33^{+}CD45^{+}$		76	67	75
$CD45^{+}CD33^{-}CD15^{+}CD14^{-}$		81	67	79
$CD45^{+}CD33^{-}$		81	67	79
$CD2^{dim}CD16^{+}CD3^{+}CD56^{-}$	NK cells	38	100	46
$CD2^{dim}CD16^{+}CD56^{+}CD3^{-}$		43	67	46
$CD2^{dim}CD16^{+}CD56^{-}CD3^{-}$		71	67	71
$CD2^{dim}CD16^{+}$		71	67	71
$CD2^{-}CD16^{+}CD3^{+}CD56^{-}$		90	0	79
$CD2^{-}CD16^{+}CD56^{+}CD3^{-}$		95	0	83
$CD2^{-}CD16^{+}CD56^{-}CD3^{-}$		76	33	71

Immune cells	Aliquot	Sensitivity (%)	Specificity (%)	Accuracy (%)
$CD2^{-}CD16^{+}$	NK cells	76	33	71
$CD2^{+}CD16^{-}CD3^{+}CD56^{-}$		52	33	50
$CD2^{+}CD16^{-}CD56^{+}CD3^{-}$		100	33	92
$CD2^{+}CD16^{-}CD56^{-}CD3^{-}$		76	33	71
$CD2^{+}CD16^{-}$		90	0	79
$CD2^{+}CD16^{+}CD3^{+}CD56^{-}$		52	100	58
$CD2^{+}CD16^{+}CD56^{+}CD3^{-}$		90	0	79
$CD2^{+}CD16^{+}CD56^{-}CD3^{-}$		86	67	83
$CD2^{+}CD16^{+}$		76	100	79
$CD3^{+}CD4^{+}CD8\beta^{+}CD8^{+}$	T cells	48	67	50
$CD3^{+}CD4^{+}CD8\beta^{+}CD8^{+}$ (proportion of $CD3^{+}$ cells)		48	33	46
$CD3^{-}CD4^{low}CD8\beta^{low}$		76	67	75
$CD3^{-}CD8\beta^{dim}CD8^{-}$		76	67	75
$CD3^{-}CD8^{+}CD8\beta^{-}$		71	67	71
$CD3^{-}$		76	67	75
$CD3^{+}CD4^{+}CD8\beta^{-}$		81	67	79
$CD3^{+}CD4^{+}CD8\beta^{+}$		43	0	38
$CD3^{+}CD8\beta^{dim}CD8^{-}$		38	0	33
$CD3^{+}CD8\beta^{+}CD4^{-}$		52	100	58
$CD3^{+}CD8\beta^{+}CD8^{low}$		90	33	83
$CD3^{+}CD8\beta^{+}CD8^{+}$		52	100	58
$CD3^{+}CD8^{+}CD8\beta^{-}$		38	67	42
$CD3^{+}$		57	67	58

Immune cells	Aliquot	Sensitivity (%)	Specificity (%)	Accuracy (%)
$CD3^-$	TCR	76	67	75
$CD3^-CD5^-TCRab^+TCRgd^-$		86	67	83
$CD3^-CD5^-TCRab^+$		67	67	67
$CD3^-CD5^-TCRab^+TCRgd^+$		75	0	65
$CD3^-CD5^-TCRab^+TCRgd^-$		75	67	74

Table H.5 Validation results for qualified subsets of immune cells in concentration (mm^3) from the FLDA classification between aGvHD and non-GvHD patients using samples taken between 21 and 0 days prior to aGvHD diagnosis.

Immune cells	Aliquot	Sensitivity (%)	Specificity (%)	Accuracy (%)
$CD3^-CD44^+CD25^-$	1Activation	90	0	79
$CD3^-CD44^+CD25^+CD69^+$		59	67	60
$CD3^-CD44^+CD25^+$		82	100	85
$CD3^-CD44^-CD25^-$		62	67	62
$CD3^-$		81	33	75
$CD3^+CD44^-CD25^+$		67	0	58
$CD3^+CD44^+CD25^+CD69^+$		95	67	92
$CD3^+CD44^+CD25^+$		95	0	83
$CD3^+CD44^-CD25^-$		57	67	58
$CD3^+$		62	33	58
$CD3^-CD4^{dim}$	2Activation	76	0	67
$CD3^-CD8^{low}$		62	0	54
$CD3^-CD4^-CD8^-$		76	67	75
$CD3^-$		81	33	75
$CD3^+CD4^{br}$		81	33	75
$CD3^+CD4^{int}$		38	67	42
$CD3^+CD8^{br}$		48	67	50
$CD3^+CD8^{dim}$		81	33	75
$CD3^+$		67	33	62

Immune cells	Aliquot	Sensitivity (%)	Specificity (%)	Accuracy (%)
$CD3^{-}CD4^{dim}$	3Activation	71	0	62
$CD3^{-}CD8^{low}$		62	0	54
$CD3^{-}CD4^{-}CD8^{-}$		76	67	75
$CD3^{-}$		81	33	75
$CD3^{+}CD4^{br}$		86	0	75
$CD3^{+}CD4^{int}$		57	67	58
$CD3^{+}CD8^{br}$		43	67	46
$CD3^{+}CD8^{dim}$		76	67	75
$CD3^{+}$		71	67	71
$CD22^{+}CD20^{+}$	B cells	76	0	67
$CD22^{+}$		81	0	71
$CD33^{+}CD45^{dim}CD15^{low}CD14^{low}$	Myeloids	100	33	92
$CD33^{+}CD45^{dim}CD15^{+}CD14^{-}$		90	33	83
$CD33^{+}CD45^{dim}CD15^{+}CD14^{+}$		86	0	75
$CD33^{+}CD45^{dim}$		86	33	79
$CD33^{+}CD45^{+}CD15^{+}CD14^{+}$		81	0	71
$CD33^{+}CD45^{+}$		76	0	67
$CD45^{+}CD33^{-}CD15^{+}CD14^{-}$		95	67	92
$CD45^{+}CD33^{-}$		71	100	75
$CD2^{dim}CD16^{+}CD3^{+}CD56^{-}$	NK cells	24	67	29
$CD2^{dim}CD16^{+}CD56^{+}CD3^{-}$		76	0	67
$CD2^{dim}CD16^{+}CD56^{-}CD3^{-}$		86	0	75
$CD2^{dim}CD16^{+}$		76	0	67
$CD2^{-}CD16^{+}CD3^{+}CD56^{-}$		95	33	88
$CD2^{-}CD16^{+}CD56^{+}CD3^{-}$		95	0	83
$CD2^{-}CD16^{+}CD56^{-}CD3^{-}$		81	67	79

Immune cells	Aliquot	Sensitivity (%)	Specificity (%)	Accuracy (%)
$CD2^{-}CD16^{+}$	NK cells	81	67	79
$CD2^{+}CD16^{-}CD3^{+}CD56^{-}$		81	67	79
$CD2^{+}CD16^{-}CD56^{+}CD3^{-}$		100	33	92
$CD2^{+}CD16^{-}CD56^{-}CD3^{-}$		76	0	67
$CD2^{+}CD16^{-}$		86	67	83
$CD2^{+}CD16^{+}CD3^{+}CD56^{-}$		62	33	58
$CD2^{+}CD16^{+}CD56^{+}CD3^{-}$		81	33	75
$CD2^{+}CD16^{+}CD56^{-}CD3^{-}$		86	33	79
$CD2^{+}CD16^{+}$		86	67	83
$CD3^{+}CD4^{+}CD8\beta^{+}CD8^{+}$	T cells	48	100	54
$CD3^{+}CD4^{+}CD8\beta^{+}CD8^{+}$ (proportion of $CD3^{+}$ cells)		52	67	54
$CD3^{-}CD4^{low}CD8\beta^{low}$		90	0	79
$CD3^{-}CD8\beta^{dim}CD8^{-}$		95	33	88
$CD3^{-}CD8^{+}CD8\beta^{-}$		67	67	67
$CD3^{-}$		81	33	75
$CD3^{+}CD4^{+}CD8\beta^{-}$		86	33	79
$CD3^{+}CD4^{+}CD8\beta^{+}$		43	67	46
$CD3^{+}CD8\beta^{dim}CD8^{-}$		43	33	42
$CD3^{+}CD8\beta^{+}CD4^{-}$		48	100	54
$CD3^{+}CD8\beta^{+}CD8^{low}$		100	33	92
$CD3^{+}CD8\beta^{+}CD8^{+}$		48	100	54
$CD3^{+}CD8^{+}CD8\beta^{-}$		48	33	46
$CD3^{+}$		67	67	67

Immune cells	Aliquot	Sensitivity (%)	Specificity (%)	Accuracy (%)
$CD3^-$	TCR	81	33	75
$CD3^-CD5^-TCRab^+TCRgd^-$		81	67	79
$CD3^-CD5^-TCRab^+$		95	33	88
$CD3^-CD5^-TCRab^+TCRgd^+$		90	0	78
$CD3^-CD5^-TCRab^+TCRgd^-$		95	33	87

Table H.6 Validation results for qualified subsets of immune cells in concentration (mm^3) from the FLDA classification between aGvHD and non-GvHD patients using samples taken between 0 and 21 days from aGvHD diagnosis.

Immune cells	Aliquot	Sensitivity (%)	Specificity (%)	Accuracy (%)
$CD3^-CD44^+CD25^-$	1Activation	72	33	67
$CD3^-CD44^+CD25^+CD69^+$		50	67	53
$CD3^-CD44^+CD25^+$		62	0	53
$CD3^-CD44^-CD25^-$		67	33	62
$CD3^-$		61	100	67
$CD3^+CD44^-CD25^+$		39	33	38
$CD3^+CD44^+CD25^+CD69^+$		33	100	43
$CD3^+CD44^+CD25^+$		72	0	62
$CD3^+CD44^-CD25^-$		78	33	71
$CD3^+$		89	33	81
$CD3^-CD4^{dim}$	2Activation	61	67	62
$CD3^-CD8^{low}$		56	0	48
$CD3^-CD4^-CD8^-$		67	33	62
$CD3^-$		44	100	52
$CD3^+CD4^{br}$		100	33	90
$CD3^+CD4^{int}$		56	67	57
$CD3^+CD8^{br}$		78	33	71
$CD3^+CD8^{dim}$		94	33	86
$CD3^+$		94	33	86

Immune cells	Aliquot	Sensitivity (%)	Specificity (%)	Accuracy (%)
$CD3^{-}CD4^{dim}$	3Activation	67	67	67
$CD3^{-}CD8^{low}$		72	0	62
$CD3^{-}CD4^{-}CD8^{-}$		67	33	62
$CD3^{-}$		44	100	52
$CD3^{+}CD4^{br}$		100	33	90
$CD3^{+}CD4^{int}$		83	67	81
$CD3^{+}CD8^{br}$		78	33	71
$CD3^{+}CD8^{dim}$		94	33	86
$CD3^{+}$		94	33	86
$CD22^{+}CD20^{+}$	B cells	94	0	81
$CD22^{+}$		100	0	86
$CD33^{+}CD45^{dim}CD15^{low}CD14^{low}$	Myeloids	22	67	29
$CD33^{+}CD45^{dim}CD15^{+}CD14^{-}$		50	33	48
$CD33^{+}CD45^{dim}CD15^{+}CD14^{+}$		22	100	33
$CD33^{+}CD45^{dim}$		28	100	38
$CD33^{+}CD45^{+}CD15^{+}CD14^{+}$		50	67	52
$CD33^{+}CD45^{+}$		56	67	57
$CD45^{+}CD33^{-}CD15^{+}CD14^{-}$		100	0	86
$CD45^{+}CD33^{-}$		94	33	86
$CD2^{dim}CD16^{+}CD3^{+}CD56^{-}$	NK cells	83	67	81
$CD2^{dim}CD16^{+}CD56^{+}CD3^{-}$		50	0	43
$CD2^{dim}CD16^{+}CD56^{-}CD3^{-}$		44	100	52
$CD2^{dim}CD16^{+}$		44	100	52
$CD2^{-}CD16^{+}CD3^{+}CD56^{-}$		89	0	76
$CD2^{-}CD16^{+}CD56^{+}CD3^{-}$		78	33	71
$CD2^{-}CD16^{+}CD56^{-}CD3^{-}$		83	67	81

Immune cells	Aliquot	Sensitivity (%)	Specificity (%)	Accuracy (%)
$CD2^{-}CD16^{+}$	NK cells	83	67	81
$CD2^{+}CD16^{-}CD3^{+}CD56^{-}$		100	33	90
$CD2^{+}CD16^{-}CD56^{+}CD3^{-}$		83	0	71
$CD2^{+}CD16^{-}CD56^{-}CD3^{-}$		100	33	90
$CD2^{+}CD16^{-}$		100	33	90
$CD2^{+}CD16^{+}CD3^{+}CD56^{-}$		61	33	57
$CD2^{+}CD16^{+}CD56^{+}CD3^{-}$		72	0	62
$CD2^{+}CD16^{+}CD56^{-}CD3^{-}$		89	0	76
$CD2^{+}CD16^{+}$		67	33	62
$CD3^{+}CD4^{+}CD8\beta^{+}CD8^{+}$	T cells	83	33	76
$CD3^{+}CD4^{+}CD8\beta^{+}CD8^{+}$ (proportion of $CD3^{+}$ cells)		56	100	62
$CD3^{-}CD4^{low}CD8\beta^{low}$		39	100	48
$CD3^{-}CD8\beta^{dim}CD8^{-}$		78	33	71
$CD3^{-}CD8^{+}CD8\beta^{-}$		67	0	57
$CD3^{-}$		44	100	52
$CD3^{+}CD4^{+}CD8\beta^{-}$		100	33	90
$CD3^{+}CD4^{+}CD8\beta^{+}$		72	33	67
$CD3^{+}CD8\beta^{dim}CD8^{-}$		61	67	62
$CD3^{+}CD8\beta^{+}CD4^{-}$		78	33	71
$CD3^{+}CD8\beta^{+}CD8^{low}$		83	0	71
$CD3^{+}CD8\beta^{+}CD8^{+}$		72	67	71
$CD3^{+}CD8^{+}CD8\beta^{-}$		94	33	86
$CD3^{+}$		94	33	86

Immune cells	Aliquot	Sensitivity (%)	Specificity (%)	Accuracy (%)
$CD3^-$	TCR	50	100	57
$CD3^-CD5^-TCRab^+TCRgd^-$		61	67	62
$CD3^-CD5^-TCRab^+$		44	100	52
$CD3^-CD5^-TCRab^+TCRgd^+$		44	100	52
$CD3^-CD5^-TCRab^+TCRgd^-$		50	100	57

Appendix I. Other top ranking classifiers for the onset of cGvHD

Many top ranking classifiers designed to predict or elucidate the onset and progression of cGvHD exhibited inconsistent patterns compared to its raw data patterns. An example of the inconsistent classifiers was shown in Section 5.1. In the FLDA analysis of the concentration dataset using samples taken between 7 and 21 days post-transplant, only one type of pattern: a sudden increase from aGvHD only patients, was observed. The FLDA classification built from the subset of immune cells $45RA^{+}CD3^{+}CD8^{low}$ in cell concentration (Figure I.1) was used as an example of this pattern. The classifier had an estimated 86% sensitivity and 71% specificity (Table J.4). The FLDA estimated signals from the aGvHD only patients increased at 15 days post-transplant and became higher than the aGvHD & cGvHD patients around 21 days post-transplant (Figure I.1a). This pattern was consistent with the raw data plotted from 0 to 100 days post-transplant (purpled striped area, Figure I.1b). In the extended raw data time plot, four out of the seven available non-GvHD patient datasets suddenly increased around 15 to 55 days post-transplant (Figure I.1b). Similar patterns were also observed from other classifiers such as $CD3^{-}TCRab^{+}CD5^{+}$ and $CD2^{dim}CD16^{+}CD3^{+}CD56^{-}$ (data not shown) but with a lower estimated sensitivity and specificity (Table J.4).

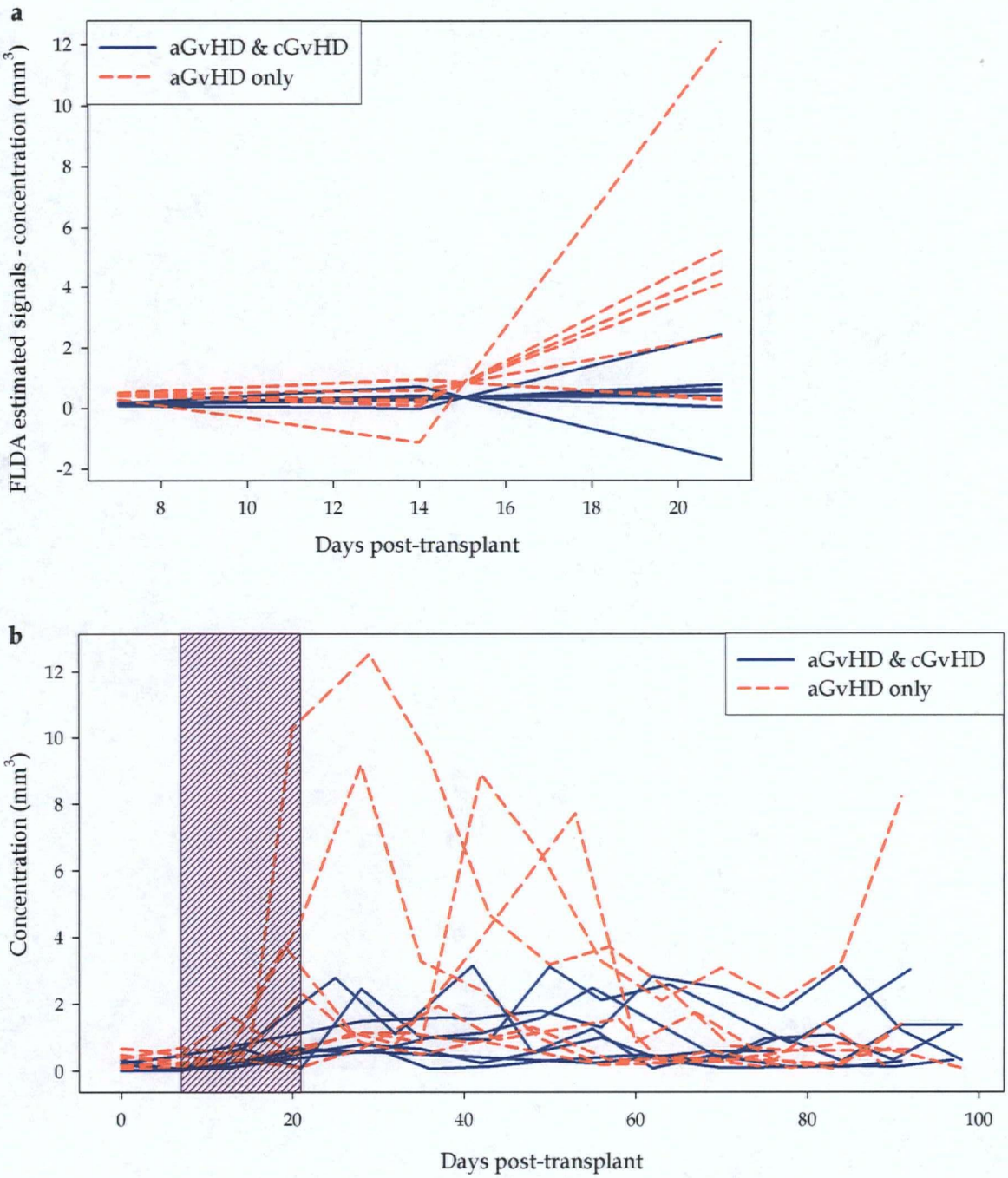


Figure I.1 Time plot of the FLDA estimated signals (panel a) based on samples taken between 7 and 21 days post-transplant and time plot of the raw data (panel b) based on samples taken between 0 and 100 days post-transplant for the immune cells $45RA^{+}CD3^{+}CD8^{low}$ in proportion to PBMC (%). The purple striped box indicates the time range where data was analyzed via FLDA.

In the FLDA analysis of the concentration dataset using samples taken between 21 and 0 days prior to aGvHD diagnosis, only one subset of immune cells exhibited a consistent classifier exhibiting opposite FLDA signal pattern. The top classifier was $45RA^{+}CD3^{-}CD4^{dim}$ (Figure I.2). The FLDA classifier had an estimated 86% sensitivity and 71% specificity (Table J.5). Its FLDA signals were the opposite between the patients groups (Figure I.2a). However, this pattern could not be easily identified in the local or extended raw data time plots for either subset of immune cells (Figure I.2b).

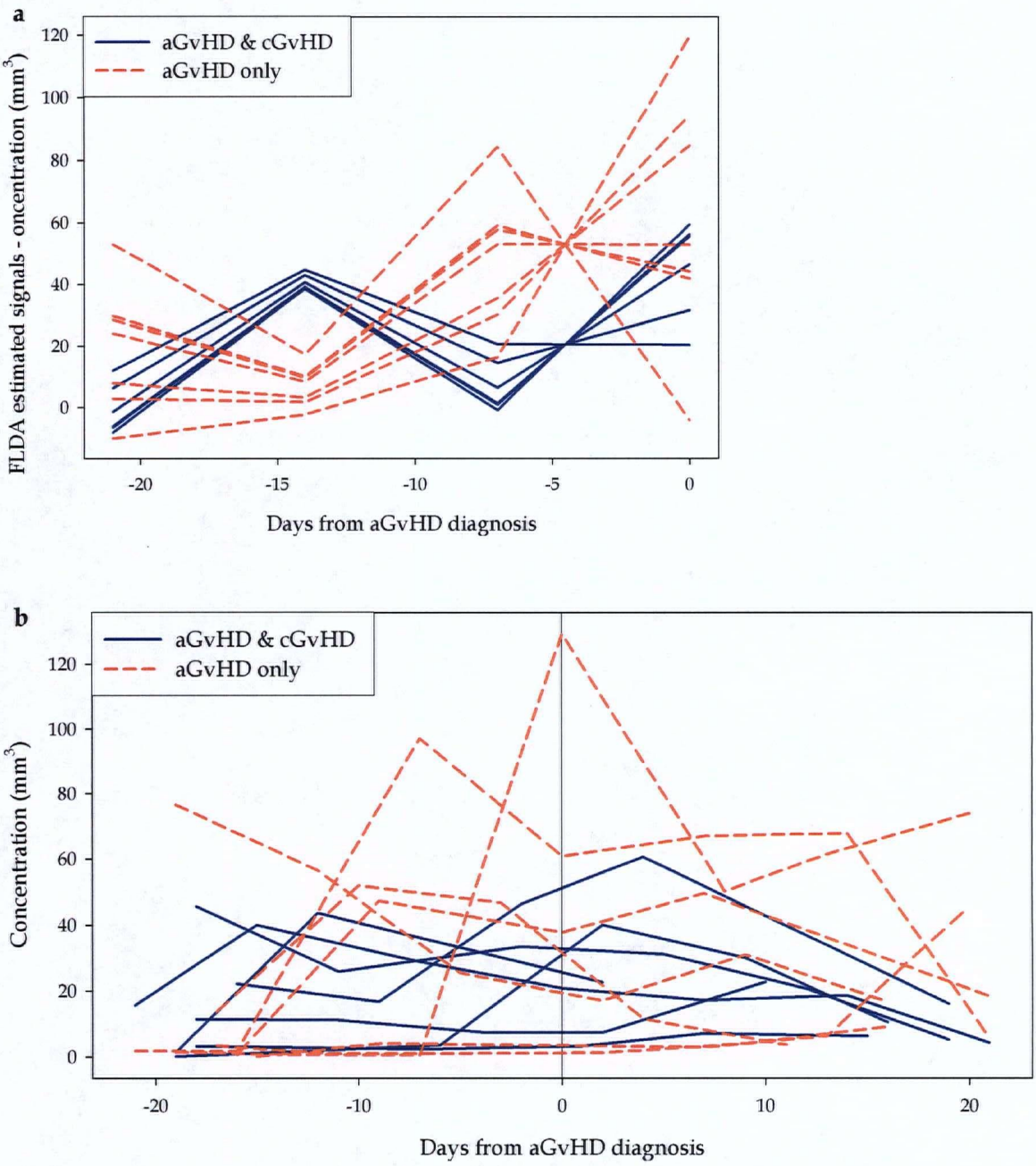


Figure I.2 Time plot of the FLDA estimated signals (panel a) based on samples taken between -21 and 0 from aGvHD diagnosis and time plot of the raw data (panel b) based on samples taken between -21 and 21 days from aGvHD diagnosis for the immune cells $45RA^{+}CD3^{-}CD4^{dim}$ in concentration (mm^3). The date of aGvHD diagnosis is labelled as day 0.

In the FLDA analysis of the proportion dataset using samples taken between 0 and 21 days post-aGvHD diagnosis, the FLDA classifier built from the immune cells $CD3^{+}CD4^{int}$ (aliquot '2Activation') had a pattern of higher values from the aGvHD only patients (Figure I.3). The classifier predicting the onset of cGvHD had an estimated 83% sensitivity and 89% specificity (Table J.3). The same subset of immune cells was also identified as top ranking classifier in the comparison between aGvHD and non-GvHD patients (section 4.1.3). In the FLDA estimated signals time plot (Figure I.3a), proportion values from the aGvHD only patients started with higher values at the beginning of the analyzed time range and steadily decreased, while the values from the aGvHD & cGvHD patients increased. In the raw data time plot from -21 to 21 days from aGvHD diagnosis (Figure I.3b), values from the aGvHD patients were generally higher across time points, when compared to the aGvHD & cGvHD patients.

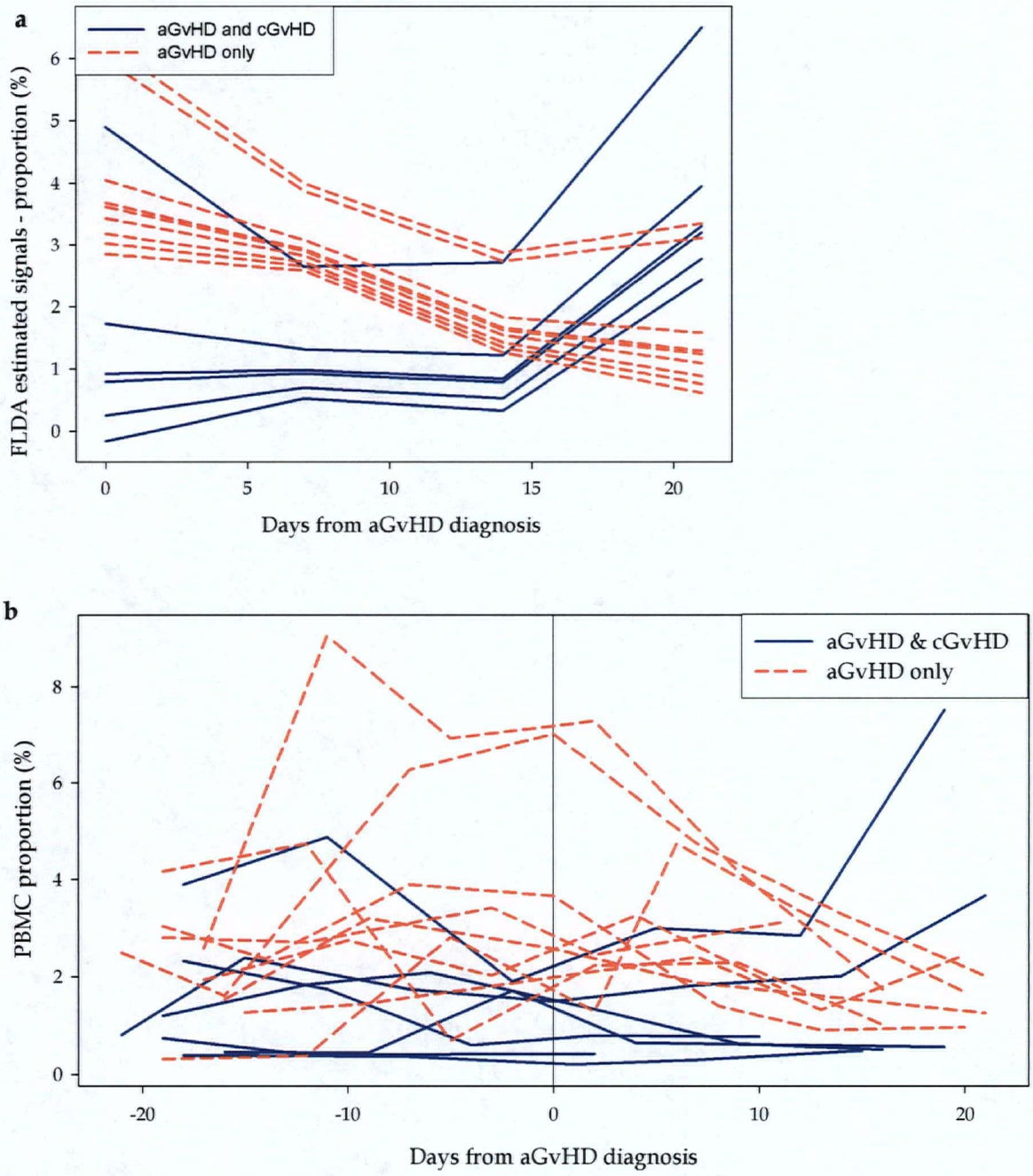


Figure I.3 Time plot of the FLDA estimated signals (panel a) based on samples taken between 0 and 21 days from aGvHD diagnosis and time plot of the raw data (panel b) based on samples taken between -21 and 21 days from aGvHD diagnosis for the immune cells $CD3^{+}CD4int$ (aliquot '2Activation') in proportion to PBMC (%). The date of aGvHD diagnosis is labelled as day 0.

Appendix J. Summaries of LOOCV results for the FLDA analyses between aGvHD & cGvHD and aGvHD only patients

Table J.1 Validation results for qualified subsets of immune cells in proportion to PBMC (%) from the FLDA classification between aGvHD & cGvHD and aGvHD only patients using samples taken from 7 to 21 days post-transplant.

Immune cells	Aliquot	Sensitivity (%)	Specificity (%)	Accuracy (%)
$CD3^-CD44^+CD25^-$	1Activation	43	67	56
$CD3^-CD44^+CD25^+CD69^+$		100	43	71
$CD3^-CD44^+CD25^+$		86	43	64
$CD3^-CD44^-CD25^-$		71	56	62
$CD3^-$		29	22	25
$CD3^+CD44^-CD25^+$		57	67	62
$CD3^+CD44^+CD25^-$		57	22	38
$CD3^+CD44^+CD25^+CD69^+$		86	67	75
$CD3^+CD44^+CD25^+$		86	44	62
$CD3^+CD44^-CD25^-$		14	56	38
$CD3^+$		43	11	25
$CD3^-CD4^{dim}$	2Activation	57	67	62
$CD3^-CD8^{low}$		57	22	38
$CD3^-CD4^-CD8^-$		86	44	62
$CD3^-$		43	33	38
$CD3^+CD4^{br}$		57	56	56
$CD3^+CD4^{int}$		86	67	75
$CD3^+CD8^{br}$		43	56	50

Immune cells	Aliquot	Sensitivity (%)	Specificity (%)	Accuracy (%)
$CD3^{+}CD8^{dim}$	2Activation	57	56	56
$CD3^{+}$		29	0	12
$CD3^{-}CD4^{dim}$	3Activation	57	67	62
$CD3^{-}CD8^{low}$		43	22	31
$CD3^{-}CD4^{-}CD8^{-}$		86	44	62
$CD3^{-}$		29	22	25
$CD3^{+}CD4^{br}$		86	67	75
$CD3^{+}CD4^{int}$		71	44	56
$CD3^{+}CD8^{br}$		43	67	56
$CD3^{+}CD8^{dim}$		71	44	56
$CD3^{+}$		29	22	25
$CD20^{+}CD19^{+}$	B cells	57	56	56
$CD22^{+}CD20^{+}$		29	67	50
$CD22^{+}$		71	33	50
$CD33^{+}CD45^{dim}CD15^{low}CD14^{low}$	Myeloids	29	44	38
$CD33^{+}CD45^{dim}CD15^{+}CD14^{-}$		57	11	31
$CD33^{+}CD45^{dim}CD15^{+}CD14^{+}$		71	78	75
$CD33^{+}CD45^{dim}$		14	11	12
$CD33^{+}CD45^{+}CD15^{+}CD14^{+}$		29	44	38
$CD33^{+}CD45^{+}$		14	44	31
$CD45^{+}CD33^{-}CD15^{+}CD14^{-}$		29	78	56
$CD45^{+}CD33^{-}$		57	11	31
$CD2^{dim}CD16^{+}CD3^{+}CD56^{-}$	NK cells	57	78	69
$CD2^{dim}CD16^{+}CD56^{+}CD3^{-}$		71	56	62
$CD2^{dim}CD16^{+}CD56^{-}CD3^{-}$		29	67	50
$CD2^{dim}CD16^{+}$		29	44	38

Immune cells	Aliquot	Sensitivity (%)	Specificity (%)	Accuracy (%)
$CD2^-CD16^+CD3^+CD56^-$	NK cells	86	89	88
$CD2^-CD16^+CD56^+CD3^-$		71	33	50
$CD2^-CD16^+CD56^-CD3^-$		57	44	50
$CD2^-CD16^+$		43	44	44
$CD2^+CD16^-CD3^+CD56^-$		57	33	44
$CD2^+CD16^-CD56^+CD3^-$		14	0	6
$CD2^+CD16^-CD56^-CD3^-$		29	44	38
$CD2^+CD16^-$		71	44	56
$CD2^+CD16^+CD3^+CD56^-$		71	44	56
$CD2^+CD16^+CD56^+CD3^-$		29	44	38
$CD2^+CD16^+CD56^-CD3^-$		57	33	44
$CD2^+CD16^+$		71	44	56
$45RA^+CD3^-CD4^{dim}$	rest/act T helper	0	0	0
$45RA^+CD3^-$		0	14	7
$45RA^+CD3^+CD4^{low}$		57	43	50
$45RA^+CD3^+CD4^-$		71	71	71
$45RA^+CD3^+CD4^+$		14	43	29
$45RA^+CD3^+$		57	57	57
$45RO^+CD3^-CD4^{dim}$		14	43	29
$45RO^+CD3^-$		29	57	43
$45RO^+CD3^+CD4^{low}$		43	71	57
$45RO^+CD3^+CD4^-$		86	71	79
$45RO^+CD3^+CD4^+$		57	43	50
$45RO^+CD3^+$		29	43	36
$CD3^-$		43	29	36
$CD3^+CD4^-$		57	57	57

Immune cells	Aliquot	Sensitivity (%)	Specificity (%)	Accuracy (%)
$CD3^+CD4^+$	rest/act T helper	57	43	50
$CD3^+$		29	57	43
$CD4^{dim}$		14	14	14
$CD3^-CD4^-$		29	43	36
$45RA^+CD3^-CD8$	rest/act T suppressor	0	14	7
$45RA^+CD3^-$		0	0	0
$45RA^+CD3^+CD8^{low}$		57	71	64
$45RA^+CD3^+CD8^-$		29	57	43
$45RA^+CD3^+CD8^+$		71	43	57
$45RA^+CD3^+$		57	57	57
$45RO^+CD3^-$		29	57	43
$45RO^+CD3^+CD8^{low}$		71	57	64
$45RO^+CD3^+CD8^-$		29	57	43
$45RO^+CD3^+CD8^+$		57	43	50
$45RO^+CD3^+$		14	43	29
$CD3^-$		29	29	29
$CD3^+CD8^-$		43	57	50
$CD3^+CD8^+$		71	43	57
$CD3^+$		14	43	29
$CD8^+CD3^-$		57	0	29
$CD3^-CD8^-$		29	43	36
$CD3^+CD4^+CD8\beta^+CD8^+$	T cells	57	56	56
$CD3^+CD4^+CD8\beta^+CD8^+$ (proportion of $CD3^+$ cells)		43	44	44
$CD3^-CD4^{low}CD8\beta^{low}$		29	78	56
$CD3^-CD8\beta^{dim}CD8^-$		29	67	50
$CD3^-CD8^+CD8\beta^-$		43	22	31

Immune cells	Aliquot	Sensitivity (%)	Specificity (%)	Accuracy (%)
$CD3^-$	T cells	43	33	38
$CD3^+CD4^+CD8\beta^-$		57	56	56
$CD3^+CD4^+CD8\beta^+$		57	33	44
$CD3^+CD8\beta^{dim}CD8^-$		57	78	69
$CD3^+CD8\beta^+CD4^-$		43	67	56
$CD3^+CD8\beta^+CD8^{low}$		100	22	56
$CD3^+CD8\beta^+CD8^+$		43	67	56
$CD3^+CD8^+CD8\beta^-$		29	11	19
$CD3^+$		14	0	6
$CD3^-CD5^+$	TCR	57	50	54
$CD3^-$		29	22	25
$CD3^-CD5^-TCRab^+TCRgd^-$		57	44	50
$CD3^-CD5^-TCRab^+$		0	56	31
$CD3^-CD5^-TCRab^+TCRgd^+$		43	44	44
$CD3^-CD5^-TCRab^+TCRgd^-$		14	44	31
$CD3^-TCR^+CD5^+$		71	67	69
$CD3^+$		57	22	38
$CD3^+CD5^-TCRab^+$		100	11	50
$CD3^+CD5^-TCRab^+TCRgd^+$		71	11	38
$CD3^+CD5^-TCRab^+TCRgd^-$		100	11	50
$CD3^+CD5^+TCRab^+$		71	38	53
$CD3^+CD5^+TCRab^+TCRgd^+$		57	25	40
$CD3^+CD5^+TCRgd^+$		71	38	53

Table J.2 Validation results for qualified subsets of immune cells in proportion to PBMC (%) from the FLDA classification between aGvHD & cGvHD and aGvHD only patients using samples taken between 21 and 0 days prior to aGvHD diagnosis.

Immune cells	Aliquot	Sensitivity (%)	Specificity (%)	Accuracy (%)
$CD3^-CD44^+CD25^-$	1Activation	71	78	75
$CD3^-CD44^+CD25^+CD69^+$		71	29	50
$CD3^-CD44^+CD25^+$		86	29	57
$CD3^-CD44^-CD25^-$		14	11	12
$CD3^-$		71	67	69
$CD3^+CD44^-CD25^+$		71	56	62
$CD3^+CD44^+CD25^-$		57	11	31
$CD3^+CD44^+CD25^+CD69^+$		86	33	56
$CD3^+CD44^+CD25^+$		29	33	31
$CD3^+CD44^-CD25^-$		57	56	56
$CD3^+$		71	56	62
$CD3^-CD4^{dim}$	2Activation	71	67	69
$CD3^-CD8^{low}$		43	44	44
$CD3^-CD4^-CD8^-$		57	33	44
$CD3^-$		71	78	75
$CD3^+CD4^{br}$		57	44	50
$CD3^+CD4^{int}$		86	44	62
$CD3^+CD8^{br}$		43	89	69
$CD3^+CD8^{dim}$		71	44	56
$CD3^+$		71	56	62

Immune cells	Aliquot	Sensitivity (%)	Specificity (%)	Accuracy (%)
$CD3^{-}CD4^{dim}$	3Activation	71	78	75
$CD3^{-}CD8^{low}$		43	33	38
$CD3^{-}CD4^{-}CD8^{-}$		57	33	44
$CD3^{-}$		71	89	81
$CD3^{+}CD4^{br}$		57	56	56
$CD3^{+}CD4^{int}$		86	56	69
$CD3^{+}CD8^{br}$		57	89	75
$CD3^{+}CD8^{dim}$		71	33	50
$CD3^{+}$		71	56	62
$CD20^{+}CD19^{+}$	B cells	43	22	31
$CD22^{+}CD20^{+}$		43	100	75
$CD22^{+}$		86	44	62
$CD33^{+}CD45^{dim}CD15^{low}CD14^{low}$	Myeloids	86	22	50
$CD33^{+}CD45^{dim}CD15^{+}CD14^{-}$		57	56	56
$CD33^{+}CD45^{dim}CD15^{+}CD14^{+}$		71	11	38
$CD33^{+}CD45^{dim}$		71	33	50
$CD33^{+}CD45^{+}CD15^{+}CD14^{+}$		71	56	62
$CD33^{+}CD45^{+}$		71	67	69
$CD45^{+}CD33^{-}CD15^{+}CD14^{-}$		71	100	88
$CD45^{+}CD33^{-}$		86	56	69
$CD2^{dim}CD16^{+}CD3^{+}CD56^{-}$	NK cells	86	56	69
$CD2^{dim}CD16^{+}CD56^{+}CD3^{-}$		57	56	56
$CD2^{dim}CD16^{+}CD56^{-}CD3^{-}$		57	78	69
$CD2^{dim}CD16^{+}$		71	56	62
$CD2^{-}CD16^{+}CD3^{+}CD56^{-}$		71	33	50
$CD2^{-}CD16^{+}CD56^{+}CD3^{-}$		71	56	62

Immune cells	Aliquot	Sensitivity (%)	Specificity (%)	Accuracy (%)
$CD2^-CD16^+CD56^-CD3^-$	NK cells	71	56	62
$CD2^-CD16^+$		57	56	56
$CD2^+CD16^-CD3^+CD56^-$		57	56	56
$CD2^+CD16^-CD56^+CD3^-$		14	44	31
$CD2^+CD16^-CD56^-CD3^-$		29	0	12
$CD2^+CD16^-$		71	67	69
$CD2^+CD16^+CD3^+CD56^-$		86	22	50
$CD2^+CD16^+CD56^+CD3^-$		57	56	56
$CD2^+CD16^+CD56^-CD3^-$		57	56	56
$CD2^+CD16^+$		57	11	31
$45RA^+CD3^-CD4^{dim}$	rest/act T helper	86	71	79
$45RA^+CD3^-$		71	86	79
$45RA^+CD3^+CD4^{low}$		86	43	64
$45RA^+CD3^+CD4^-$		57	86	71
$45RA^+CD3^+CD4^+$		57	57	57
$45RA^+CD3^+$		71	86	79
$45RO^+CD3^-CD4^{dim}$		86	86	86
$45RO^+CD3^-$		86	57	71
$45RO^+CD3^+CD4^{low}$		57	43	50
$45RO^+CD3^+CD4^-$		57	86	71
$45RO^+CD3^+CD4^+$		57	57	57
$45RO^+CD3^+$		57	43	50
$CD3^-$		86	71	79
$CD3^+CD4^-$		57	71	64
$CD3^+CD4^+$		57	57	57
$CD3^+$		71	71	71

Immune cells	Aliquot	Sensitivity (%)	Specificity (%)	Accuracy (%)
$CD4^{dim}$	rest/act T suppressor	86	71	79
$CD3^{-}CD4^{-}$		43	57	50
$45RA^{+}CD3^{-}CD8$		43	29	36
$45RA^{+}CD3^{-}$		86	86	86
$45RA^{+}CD3^{+}CD8^{low}$		43	29	36
$45RA^{+}CD3^{+}CD8^{-}$	rest/act T suppressor	71	57	64
$45RA^{+}CD3^{+}CD8^{+}$		57	86	71
$45RA^{+}CD3^{+}$		71	86	79
$45RO^{+}CD3^{-}$		71	43	57
$45RO^{+}CD3^{+}CD8^{low}$		29	43	36
$45RO^{+}CD3^{+}CD8^{-}$		71	57	64
$45RO^{+}CD3^{+}CD8^{+}$		57	86	71
$45RO^{+}CD3^{+}$		57	71	64
$CD3^{-}$		71	86	79
$CD3^{+}CD8^{-}$		71	57	64
$CD3^{+}CD8^{+}$		57	100	79
$CD3^{+}$		71	71	71
$CD8^{+}CD3^{-}$		43	29	36
$CD3^{-}CD8^{-}$		71	86	79
$CD3^{+}CD4^{+}CD8\beta^{+}CD8^{+}$	T cells	43	44	44
$CD3^{+}CD4^{+}CD8\beta^{+}CD8^{+}$ (proportion of $CD3^{+}$ cells)		57	44	50
$CD3^{-}CD4^{low}CD8\beta^{low}$		57	56	56
$CD3^{-}CD8\beta^{dim}CD8^{-}$		57	67	62
$CD3^{-}CD8^{+}CD8\beta^{-}$		14	22	19
$CD3^{-}$		71	78	75
$CD3^{+}CD4^{+}CD8\beta^{-}$		71	56	62

Immune cells	Aliquot	Sensitivity (%)	Specificity (%)	Accuracy (%)
$CD3^+CD4^+CD8\beta^+$	T cells	14	44	31
$CD3^+CD8\beta^{dim}CD8^-$		57	67	62
$CD3^+CD8\beta^+CD4^-$		43	78	62
$CD3^+CD8\beta^+CD8^{low}$		100	22	56
$CD3^+CD8\beta^+CD8^+$		43	89	69
$CD3^+CD8^+CD8\beta^-$		43	56	50
$CD3^+$		57	56	56
$CD3^-CD5^+$	TCR	43	33	38
$CD3^-$		71	67	69
$CD3^-CD5^-TCRab^+TCRgd^-$		29	44	38
$CD3^-CD5^-TCRab^+$		86	67	75
$CD3^-CD5^-TCRab^+TCRgd^+$		43	67	56
$CD3^-CD5^-TCRab^+TCRgd^-$		86	67	75
$CD3^-TCR^+CD5^+$		71	67	69
$CD3^+$		57	56	56
$CD3^+CD5^-TCRab^+$		86	0	38
$CD3^+CD5^-TCRab^+TCRgd^+$		57	33	44
$CD3^+CD5^-TCRab^+TCRgd^-$		86	11	44
$CD3^+CD5^+TCRab^+$		29	62	47
$CD3^+CD5^+TCRab^+TCRgd^+$		86	50	67
$CD3^+CD5^+TCRgd^+$		57	75	67

Table J.3 Validation results for qualified subsets of immune cells in proportion to PBMC (%) from the FLDA classification between aGvHD & cGvHD and aGvHD only patients using samples taken between 0 and 21 days from aGvHD diagnosis.

Immune cells	Aliquot	Sensitivity (%)	Specificity (%)	Accuracy (%)
$CD3^-CD44^+CD25^-$	1Activation	33	44	40
$CD3^-CD44^+CD25^+CD69^+$		83	43	62
$CD3^-CD44^+CD25^+$		83	43	62
$CD3^-CD44^-CD25^-$		50	44	47
$CD3^-$		50	44	47
$CD3^+CD44^-CD25^+$		67	22	40
$CD3^+CD44^+CD25^-$		83	44	60
$CD3^+CD44^+CD25^+CD69^+$		33	56	47
$CD3^+CD44^+CD25^+$		0	44	27
$CD3^+CD44^-CD25^-$		67	56	60
$CD3^+$		50	44	47
$CD3^-CD4^{dim}$	2Activation	17	33	27
$CD3^-CD8^{low}$		17	33	27
$CD3^-CD4^-CD8^-$		33	56	47
$CD3^-$		17	44	33
$CD3^+CD4^{br}$		17	22	20
$CD3^+CD4^{int}$		83	89	87
$CD3^+CD8^{br}$		17	67	47
$CD3^+CD8^{dim}$		67	33	47
$CD3^+$		0	44	27

Immune cells	Aliquot	Sensitivity (%)	Specificity (%)	Accuracy (%)
$CD3^{-}CD4^{dim}$	3Activation	17	22	20
$CD3^{-}CD8^{low}$		33	44	40
$CD3^{-}CD4^{-}CD8^{-}$		33	22	27
$CD3^{-}$		17	56	40
$CD3^{+}CD4^{br}$		17	33	27
$CD3^{+}CD4^{int}$		67	78	73
$CD3^{+}CD8^{br}$		17	67	47
$CD3^{+}CD8^{dim}$		50	22	33
$CD3^{+}$		17	56	40
$CD20^{+}CD19^{+}$	B cells	33	78	60
$CD22^{+}CD20^{+}$		17	89	60
$CD22^{+}$		33	56	47
$CD33^{+}CD45^{dim}CD15^{low}CD14^{low}$	Myeloids	50	78	67
$CD33^{+}CD45^{dim}CD15^{+}CD14^{-}$		33	78	60
$CD33^{+}CD45^{dim}CD15^{+}CD14^{+}$		33	89	67
$CD33^{+}CD45^{dim}$		33	89	67
$CD33^{+}CD45^{+}CD15^{+}CD14^{+}$		0	22	13
$CD33^{+}CD45^{+}$		0	22	13
$CD45^{+}CD33^{-}CD15^{+}CD14^{-}$		33	44	40
$CD45^{+}CD33^{-}$		17	44	33
$CD2^{dim}CD16^{+}CD3^{+}CD56^{-}$	NK cells	83	44	60
$CD2^{dim}CD16^{+}CD56^{+}CD3^{-}$		67	44	53
$CD2^{dim}CD16^{+}CD56^{-}CD3^{-}$		33	44	40
$CD2^{dim}CD16^{+}$		33	56	47
$CD2^{-}CD16^{+}CD3^{+}CD56^{-}$		83	22	47
$CD2^{-}CD16^{+}CD56^{+}CD3^{-}$		33	33	33

Immune cells	Aliquot	Sensitivity (%)	Specificity (%)	Accuracy (%)
$CD2^-CD16^+CD56^-CD3^-$	NK cells	33	33	33
$CD2^-CD16^+$		17	44	33
$CD2^+CD16^-CD3^+CD56^-$		33	22	27
$CD2^+CD16^-CD56^+CD3^-$		0	67	40
$CD2^+CD16^-CD56^-CD3^-$		67	56	60
$CD2^+CD16^-$		50	67	60
$CD2^+CD16^+CD3^+CD56^-$		50	67	60
$CD2^+CD16^+CD56^+CD3^-$		50	22	33
$CD2^+CD16^+CD56^-CD3^-$		50	56	53
$CD2^+CD16^+$		50	56	53
$45RA^+CD3^-CD4^{dim}$	rest/act T helper	33	14	23
$45RA^+CD3^-$		50	43	46
$45RA^+CD3^+CD4^{low}$		17	43	31
$45RA^+CD3^+CD4^-$		33	57	46
$45RA^+CD3^+CD4^+$		0	43	23
$45RA^+CD3^+$		33	57	46
$45RO^+CD3^-CD4^{dim}$		33	14	23
$45RO^+CD3^-$		33	43	38
$45RO^+CD3^+CD4^{low}$		0	43	23
$45RO^+CD3^+CD4^-$		0	57	31
$45RO^+CD3^+CD4^+$		33	14	23
$45RO^+CD3^+$		50	43	46
$CD3^-$		50	43	46
$CD3^+CD4^-$		50	71	62
$CD3^+CD4^+$		17	57	38
$CD3^+$		50	43	46

Immune cells	Aliquot	Sensitivity (%)	Specificity (%)	Accuracy (%)
$CD4^{dim}$	rest/act T helper	17	14	15
$CD3^{-}CD4^{-}$		17	43	31
$45RA^{+}CD3^{-}CD8$	rest/act T suppressor	33	57	46
$45RA^{+}CD3^{-}$		33	43	38
$45RA^{+}CD3^{+}CD8^{low}$		83	86	85
$45RA^{+}CD3^{+}CD8^{-}$		17	14	15
$45RA^{+}CD3^{+}CD8^{+}$		50	71	62
$45RA^{+}CD3^{+}$		50	86	69
$45RO^{+}CD3^{-}$		17	43	31
$45RO^{+}CD3^{+}CD8^{low}$		83	57	69
$45RO^{+}CD3^{+}CD8^{-}$		50	29	38
$45RO^{+}CD3^{+}CD8^{+}$		50	86	69
$45RO^{+}CD3^{+}$		67	57	62
$CD3^{-}$		67	43	54
$CD3^{+}CD8^{-}$		33	43	38
$CD3^{+}CD8^{+}$		33	71	54
$CD3^{+}$		67	43	54
$CD8^{+}CD3^{-}$		33	43	38
$CD3^{-}CD8^{-}$		67	43	54
$CD3^{+}CD4^{+}CD8\beta^{+}CD8^{+}$	T cells	0	33	20
$CD3^{+}CD4^{+}CD8\beta^{+}CD8^{+}$ (proportion of $CD3^{+}$ cells)		0	11	7
$CD3^{-}CD4^{low}CD8\beta^{low}$		50	44	47
$CD3^{-}CD8\beta^{dim}CD8^{-}$		33	56	47
$CD3^{-}CD8^{+}CD8\beta^{-}$		17	33	27
$CD3^{-}$		33	44	40
$CD3^{+}CD4^{+}CD8\beta^{-}$		50	44	47

Immune cells	Aliquot	Sensitivity (%)	Specificity (%)	Accuracy (%)
$CD3^+CD4^+CD8\beta^+$	T cells	0	44	27
$CD3^+CD8\beta^{dim}CD8^-$		17	56	40
$CD3^+CD8\beta^+CD4^-$		17	56	40
$CD3^+CD8\beta^+CD8^{low}$		100	22	53
$CD3^+CD8\beta^+CD8^+$		17	56	40
$CD3^+CD8^+CD8\beta^-$		67	11	33
$CD3^+$		33	44	40
$CD3^-CD5^+$	TCR	33	33	33
$CD3^-$		0	56	33
$CD3^-CD5^-TCRab^+TCRgd^-$		50	67	60
$CD3^-CD5^-TCRab^+$		33	56	47
$CD3^-CD5^-TCRab^+TCRgd^+$		100	33	60
$CD3^-CD5^-TCRab^+TCRgd^-$		17	67	47
$CD3^-TCR^+CD5^+$		100	33	67
$CD3^+$		0	56	33
$CD3^+CD5^-TCRab^+$		83	67	73
$CD3^+CD5^-TCRab^+TCRgd^+$		67	56	60
$CD3^+CD5^-TCRab^+TCRgd^-$		83	44	60
$CD3^+CD5^+TCRab^+$		50	75	64
$CD3^+CD5^+TCRab^+TCRgd^+$		50	62	57
$CD3^+CD5^+TCRgd^+$		17	50	36

Table J.4 Validation results for qualified subsets of immune cells in concentration (mm^3) from the FLDA classification between aGvHD & cGvHD and aGvHD only patients using samples taken from 7 to 21 days post-transplant.

Immune cells	Aliquot	Sensitivity (%)	Specificity (%)	Accuracy (%)
$CD3^-CD44^+CD25^-$	1Activation	29	67	50
$CD3^-CD44^+CD25^+CD69^+$		71	29	50
$CD3^-CD44^+CD25^+$		71	14	43
$CD3^-CD44^-CD25^-$		86	33	56
$CD3^-$		57	22	38
$CD3^+CD44^-CD25^+$		57	33	44
$CD3^+CD44^+CD25^-$		57	44	50
$CD3^+CD44^+CD25^+CD69^+$		86	33	56
$CD3^+CD44^+CD25^+$		57	33	44
$CD3^+CD44^-CD25^-$		86	33	56
$CD3^+$		71	22	44
$CD3^-CD4^{dim}$	2Activation	57	44	50
$CD3^-CD8^{low}$		57	11	31
$CD3^-CD4^-CD8^-$		86	22	50
$CD3^-$		57	33	44
$CD3^+CD4^{br}$		43	22	31
$CD3^+CD4^{int}$		86	44	62
$CD3^+CD8^{br}$		71	22	44
$CD3^+CD8^{dim}$		71	33	50
$CD3^+$		71	22	44

Immune cells	Aliquot	Sensitivity (%)	Specificity (%)	Accuracy (%)
$CD3^{-}CD4^{dim}$	3Activation	43	44	44
$CD3^{-}CD8^{low}$		57	33	44
$CD3^{-}CD4^{-}CD8^{-}$		86	22	50
$CD3^{-}$		57	33	44
$CD3^{+}CD4^{br}$		43	22	31
$CD3^{+}CD4^{int}$		86	44	62
$CD3^{+}CD8^{br}$		71	33	50
$CD3^{+}CD8^{dim}$		86	33	56
$CD3^{+}$		71	33	50
$CD20^{+}CD19^{+}$	B cells	57	22	38
$CD22^{+}CD20^{+}$		86	33	56
$CD22^{+}$		71	22	44
$CD33^{+}CD45^{dim}CD15^{low}CD14^{low}$	Myeloids	29	78	56
$CD33^{+}CD45^{dim}CD15^{+}CD14^{-}$		29	0	12
$CD33^{+}CD45^{dim}CD15^{+}CD14^{+}$		57	33	44
$CD33^{+}CD45^{dim}$		71	0	31
$CD33^{+}CD45^{+}CD15^{+}CD14^{+}$		29	56	44
$CD33^{+}CD45^{+}$		29	56	44
$CD45^{+}CD33^{-}CD15^{+}CD14^{-}$		57	44	50
$CD45^{+}CD33^{-}$		57	22	38
$CD2^{dim}CD16^{+}CD3^{+}CD56^{-}$	NK cells	86	56	69
$CD2^{dim}CD16^{+}CD56^{+}CD3^{-}$		57	33	44
$CD2^{dim}CD16^{+}CD56^{-}CD3^{-}$		29	33	31
$CD2^{dim}CD16^{+}$		43	33	38
$CD2^{-}CD16^{+}CD3^{+}CD56^{-}$		100	67	81
$CD2^{-}CD16^{+}CD56^{+}CD3^{-}$		71	44	56

Immune cells	Aliquot	Sensitivity (%)	Specificity (%)	Accuracy (%)
$CD2^-CD16^+CD56^-CD3^-$	NK cells	71	22	44
$CD2^-CD16^+$		71	22	44
$CD2^+CD16^-CD3^+CD56^-$		43	0	19
$CD2^+CD16^-CD56^+CD3^-$		57	22	38
$CD2^+CD16^-CD56^-CD3^-$		57	33	44
$CD2^+CD16^-$		57	0	25
$CD2^+CD16^+CD3^+CD56^-$		71	33	50
$CD2^+CD16^+CD56^+CD3^-$		57	22	38
$CD2^+CD16^+CD56^-CD3^-$		57	33	44
$CD2^+CD16^+$		71	44	56
$45RA^+CD3^-CD4^{dim}$	rest/act T helper	57	43	50
$45RA^+CD3^-$		57	43	50
$45RA^+CD3^+CD4^{low}$		57	71	64
$45RA^+CD3^+CD4^-$		71	57	64
$45RA^+CD3^+CD4^+$		43	71	57
$45RA^+CD3^+$		57	71	64
$45RO^+CD3^-CD4^{dim}$		71	29	50
$45RO^+CD3^-$		86	29	57
$45RO^+CD3^+CD4^{low}$		71	57	64
$45RO^+CD3^+CD4^-$		86	57	71
$45RO^+CD3^+CD4^+$		71	43	57
$45RO^+CD3^+$		86	43	64
$CD3^-$		71	29	50
$CD3^+CD4^-$		100	43	71
$CD3^+CD4^+$		71	43	57
$CD3^+$		71	43	57

Immune cells	Aliquot	Sensitivity (%)	Specificity (%)	Accuracy (%)
$CD4^{dim}$	rest/act T helper	71	29	50
$CD3^{-}CD4^{-}$		71	43	57
$45RA^{+}CD3^{-}CD8$	rest/act T suppressor	57	29	43
$45RA^{+}CD3^{-}$		57	29	43
$45RA^{+}CD3^{+}CD8^{low}$		86	71	79
$45RA^{+}CD3^{+}CD8^{-}$		57	71	64
$45RA^{+}CD3^{+}CD8^{+}$		57	43	50
$45RA^{+}CD3^{+}$		43	71	57
$45RO^{+}CD3^{-}$		86	43	64
$45RO^{+}CD3^{+}CD8^{low}$		86	57	71
$45RO^{+}CD3^{+}CD8^{-}$		100	43	71
$45RO^{+}CD3^{+}CD8^{+}$		86	43	64
$45RO^{+}CD3^{+}$		100	43	71
$CD3^{-}$		71	29	50
$CD3^{+}CD8^{-}$		86	43	64
$CD3^{+}CD8^{+}$		57	14	36
$CD3^{+}$		86	43	64
$CD8^{+}CD3^{-}$		57	29	43
$CD3^{-}CD8^{-}$		71	29	50
$CD3^{+}CD4^{+}CD8\beta^{+}CD8^{+}$	T cells	71	33	50
$CD3^{+}CD4^{+}CD8\beta^{+}CD8^{+}$ (proportion of $CD3^{+}$ cells)		57	22	38
$CD3^{-}CD4^{low}CD8\beta^{low}$		86	33	56
$CD3^{-}CD8\beta^{dim}CD8^{-}$		43	56	50
$CD3^{-}CD8^{+}CD8\beta^{-}$		57	11	31
$CD3^{-}$		57	33	44
$CD3^{+}CD4^{+}CD8\beta^{-}$		71	33	50

Immune cells	Aliquot	Sensitivity (%)	Specificity (%)	Accuracy (%)
$CD3^+CD4^+CD8\beta^+$	T cells	71	44	56
$CD3^+CD8\beta^{dim}CD8^-$		71	44	56
$CD3^+CD8\beta^+CD4^-$		57	22	38
$CD3^+CD8\beta^+CD8^{low}$		86	44	62
$CD3^+CD8\beta^+CD8^+$		71	33	50
$CD3^+CD8^+CD8\beta^-$		71	33	50
$CD3^+$		71	33	50
$CD3^-CD5^+$	TCR	71	33	54
$CD3^-$		57	33	44
$CD3^-CD5^-TCRab^+TCRgd^-$		57	11	31
$CD3^-CD5^-TCRab^+$		71	33	50
$CD3^-CD5^-TCRab^+TCRgd^+$		57	22	38
$CD3^-CD5^-TCRab^+TCRgd^-$		71	33	50
$CD3^-TCR^+CD5^+$		100	67	85
$CD3^+$		57	11	31
$CD3^+CD5^-TCRab^+$		100	33	62
$CD3^+CD5^-TCRab^+TCRgd^+$		71	44	56
$CD3^+CD5^-TCRab^+TCRgd^-$		100	11	50
$CD3^+CD5^+TCRab^+$		57	12	33
$CD3^+CD5^+TCRab^+TCRgd^+$		71	38	53
$CD3^+CD5^+TCRgd^+$		71	0	33

Table J.5 Validation results for qualified subsets of immune cells in concentration (mm^3) from the FLDA classification between aGvHD & cGvHD and aGvHD only patients using samples taken between 21 and 0 days prior to aGvHD diagnosis.

Immune cells	Aliquot	Sensitivity (%)	Specificity (%)	Accuracy (%)
$CD3^-CD44^+CD25^-$	1Activation	71	78	75
$CD3^-CD44^+CD25^+CD69^+$		86	29	57
$CD3^-CD44^+CD25^+$		100	29	64
$CD3^-CD44^-CD25^-$		57	22	38
$CD3^-$		57	33	44
$CD3^+CD44^-CD25^+$		57	22	38
$CD3^+CD44^+CD25^-$		14	11	12
$CD3^+CD44^+CD25^+CD69^+$		14	56	38
$CD3^+CD44^+CD25^+$		29	44	38
$CD3^+CD44^-CD25^-$		43	56	50
$CD3^+$		0	33	19
$CD3^-CD4^{dim}$	2Activation	71	78	75
$CD3^-CD8^{low}$		57	33	44
$CD3^-CD4^-CD8^-$		29	22	25
$CD3^-$		71	33	50
$CD3^+CD4^{br}$		14	44	31
$CD3^+CD4^{int}$		86	33	56
$CD3^+CD8^{br}$		29	67	50
$CD3^+CD8^{dim}$		57	0	25
$CD3^+$		43	56	50

Immune cells	Aliquot	Sensitivity (%)	Specificity (%)	Accuracy (%)
$CD3^-CD4^{dim}$	3Activation	71	78	75
$CD3^-CD8^{low}$		57	33	44
$CD3^-CD4^-CD8^-$		29	22	25
$CD3^-$		71	33	50
$CD3^+CD4^{br}$		14	56	38
$CD3^+CD4^{int}$		71	56	62
$CD3^+CD8^{br}$		29	67	50
$CD3^+CD8^{dim}$		57	22	38
$CD3^+$		43	67	56
$CD20^+CD19^+$	B cells	71	11	38
$CD22^+CD20^+$		57	100	81
$CD22^+$		71	11	38
$CD33^+CD45^{dim}CD15^{low}CD14^{low}$	Myeloids	86	22	50
$CD33^+CD45^{dim}CD15^+CD14^-$		57	33	44
$CD33^+CD45^{dim}CD15^+CD14^+$		86	11	44
$CD33^+CD45^{dim}$		86	11	44
$CD33^+CD45^+CD15^+CD14^+$		43	56	50
$CD33^+CD45^+$		57	67	62
$CD45^+CD33^-CD15^+CD14^-$		29	67	50
$CD45^+CD33^-$		0	33	19
$CD2^{dim}CD16^+CD3^+CD56^-$	NK cells	86	11	44
$CD2^{dim}CD16^+CD56^+CD3^-$		43	44	44
$CD2^{dim}CD16^+CD56^-CD3^-$		71	44	56
$CD2^{dim}CD16^+$		71	44	56
$CD2^-CD16^+CD3^+CD56^-$		86	11	44
$CD2^-CD16^+CD56^+CD3^-$		86	33	56

Immune cells	Aliquot	Sensitivity (%)	Specificity (%)	Accuracy (%)
$CD2^-CD16^+CD56^-CD3^-$	NK cells	86	22	50
$CD2^-CD16^+$		86	22	50
$CD2^+CD16^-CD3^+CD56^-$		29	78	56
$CD2^+CD16^-CD56^+CD3^-$		71	67	69
$CD2^+CD16^-CD56^-CD3^-$		57	22	38
$CD2^+CD16^-$		14	44	31
$CD2^+CD16^+CD3^+CD56^-$		29	44	38
$CD2^+CD16^+CD56^+CD3^-$		71	33	50
$CD2^+CD16^+CD56^-CD3^-$		57	56	56
$CD2^+CD16^+$		43	11	25
$45RA^+CD3^-CD4^{dim}$	rest/act T helper	86	71	79
$45RA^+CD3^-$		86	14	50
$45RA^+CD3^+CD4^{low}$		86	29	57
$45RA^+CD3^+CD4^-$		43	71	57
$45RA^+CD3^+CD4^+$		14	0	7
$45RA^+CD3^+$		14	29	21
$45RO^+CD3^-CD4^{dim}$		71	57	64
$45RO^+CD3^-$		71	57	64
$45RO^+CD3^+CD4^{low}$		86	14	50
$45RO^+CD3^+CD4^-$		43	57	50
$45RO^+CD3^+CD4^+$		43	43	43
$45RO^+CD3^+$		0	0	0
$CD3^-$		71	43	57
$CD3^+CD4^-$		14	57	36
$CD3^+CD4^+$		43	57	50
$CD3^+$		14	57	36

Immune cells	Aliquot	Sensitivity (%)	Specificity (%)	Accuracy (%)
$CD4^{dim}$	rest/act T helper	71	29	50
$CD3^{-}CD4^{-}$		43	29	36
$45RA^{+}CD3^{-}CD8$	rest/act T suppressor	86	29	57
$45RA^{+}CD3^{-}$		86	14	50
$45RA^{+}CD3^{+}CD8^{low}$		71	14	43
$45RA^{+}CD3^{+}CD8^{-}$		14	14	14
$45RA^{+}CD3^{+}CD8^{+}$		57	43	50
$45RA^{+}CD3^{+}$		43	43	43
$45RO^{+}CD3^{-}$		100	43	71
$45RO^{+}CD3^{+}CD8^{low}$		71	29	50
$45RO^{+}CD3^{+}CD8^{-}$		71	29	50
$45RO^{+}CD3^{+}CD8^{+}$		29	57	43
$45RO^{+}CD3^{+}$		29	14	21
$CD3^{-}$		71	43	57
$CD3^{+}CD8^{-}$		71	43	57
$CD3^{+}CD8^{+}$		29	57	43
$CD3^{+}$		14	14	14
$CD8^{+}CD3^{-}$		86	43	64
$CD3^{-}CD8^{-}$		71	43	57
$CD3^{+}CD4^{+}CD8\beta^{+}CD8^{+}$	T cells	29	56	44
$CD3^{+}CD4^{+}CD8\beta^{+}CD8^{+}$ (proportion of $CD3^{+}$ cells)		29	56	44
$CD3^{-}CD4^{low}CD8\beta^{low}$		100	44	69
$CD3^{-}CD8\beta^{dim}CD8^{-}$		86	56	69
$CD3^{-}CD8^{+}CD8\beta^{-}$		57	44	50
$CD3^{-}$		71	44	56
$CD3^{+}CD4^{+}CD8\beta^{-}$		29	44	38

Immune cells	Aliquot	Sensitivity (%)	Specificity (%)	Accuracy (%)
$CD3^+CD4^+CD8\beta^+$	T cells	14	44	31
$CD3^+CD8\beta^{dim}CD8^-$		29	44	38
$CD3^+CD8\beta^+CD4^-$		29	67	50
$CD3^+CD8\beta^+CD8^{low}$		71	44	56
$CD3^+CD8\beta^+CD8^+$		29	67	50
$CD3^+CD8^+CD8\beta^-$		43	56	50
$CD3^+$		29	67	50
$CD3^-CD5^+$	TCR	14	50	31
$CD3^-$		71	44	56
$CD3^-CD5^-TCRab^+TCRgd^-$		43	67	56
$CD3^-CD5^-TCRab^+$		100	33	62
$CD3^-CD5^-TCRab^+TCRgd^+$		86	33	56
$CD3^-CD5^-TCRab^+TCRgd^-$		100	33	62
$CD3^-TCR^+CD5^+$		100	17	62
$CD3^+$		29	56	44
$CD3^+CD5^-TCRab^+$		86	22	50
$CD3^+CD5^-TCRab^+TCRgd^+$		86	33	56
$CD3^+CD5^-TCRab^+TCRgd^-$		71	11	38
$CD3^+CD5^+TCRab^+$		29	50	40
$CD3^+CD5^+TCRab^+TCRgd^+$		43	50	47
$CD3^+CD5^+TCRgd^+$		43	0	20

Table J.6 Validation results for qualified subsets of immune cells in concentration (mm^3) from the FLDA classification between aGvHD & cGvHD and aGvHD only patients using samples taken between 0 and 21 days from aGvHD diagnosis.

Immune cells	Aliquot	Sensitivity (%)	Specificity (%)	Accuracy (%)
$CD3^-CD44^+CD25^-$	1Activation	50	11	27
$CD3^-CD44^+CD25^+CD69^+$		100	29	62
$CD3^-CD44^+CD25^+$		100	29	62
$CD3^-CD44^-CD25^-$		50	78	67
$CD3^-$		83	33	53
$CD3^+CD44^-CD25^+$		33	11	20
$CD3^+CD44^+CD25^-$		67	44	53
$CD3^+CD44^+CD25^+CD69^+$		100	33	60
$CD3^+CD44^+CD25^+$		100	33	60
$CD3^+CD44^-CD25^-$		83	44	60
$CD3^+$		67	44	53
$CD3^-CD4^{dim}$	2Activation	50	22	33
$CD3^-CD8^{low}$		50	67	60
$CD3^-CD4^-CD8^-$		100	22	53
$CD3^-$		100	33	60
$CD3^+CD4^{br}$		17	11	13
$CD3^+CD4^{int}$		100	22	53
$CD3^+CD8^{br}$		67	78	73
$CD3^+CD8^{dim}$		83	33	53
$CD3^+$		67	44	53

Immune cells	Aliquot	Sensitivity (%)	Specificity (%)	Accuracy (%)
$CD3^{-}CD4^{dim}$	3Activation	33	22	27
$CD3^{-}CD8^{low}$		67	56	60
$CD3^{-}CD4^{-}CD8^{-}$		83	22	47
$CD3^{-}$		100	44	67
$CD3^{+}CD4^{br}$		17	22	20
$CD3^{+}CD4^{int}$		100	22	53
$CD3^{+}CD8^{br}$		67	78	73
$CD3^{+}CD8^{dim}$		83	33	53
$CD3^{+}$		67	56	60
$CD20^{+}CD19^{+}$	B cells	0	56	33
$CD22^{+}CD20^{+}$		17	67	47
$CD22^{+}$		83	44	60
$CD33^{+}CD45^{dim}CD15^{low}CD14^{low}$	Myeloids	50	89	73
$CD33^{+}CD45^{dim}CD15^{+}CD14^{-}$		50	67	60
$CD33^{+}CD45^{dim}CD15^{+}CD14^{+}$		50	89	73
$CD33^{+}CD45^{dim}$		50	78	67
$CD33^{+}CD45^{+}CD15^{+}CD14^{+}$		67	44	53
$CD33^{+}CD45^{+}$		33	33	33
$CD45^{+}CD33^{-}CD15^{+}CD14^{-}$		83	89	87
$CD45^{+}CD33^{-}$		83	78	80
$CD2^{dim}CD16^{+}CD3^{+}CD56^{-}$	NK cells	100	44	67
$CD2^{dim}CD16^{+}CD56^{+}CD3^{-}$		67	33	47
$CD2^{dim}CD16^{+}CD56^{-}CD3^{-}$		83	11	40
$CD2^{dim}CD16^{+}$		83	22	47
$CD2^{-}CD16^{+}CD3^{+}CD56^{-}$		100	44	67
$CD2^{-}CD16^{+}CD56^{+}CD3^{-}$		50	33	40

Immune cells	Aliquot	Sensitivity (%)	Specificity (%)	Accuracy (%)
$CD2^-CD16^+CD56^-CD3^-$	NK cells	83	33	53
$CD2^-CD16^+$		83	44	60
$CD2^+CD16^-CD3^+CD56^-$		50	56	53
$CD2^+CD16^-CD56^+CD3^-$		17	78	53
$CD2^+CD16^-CD56^-CD3^-$		83	33	53
$CD2^+CD16^-$		67	67	67
$CD2^+CD16^+CD3^+CD56^-$		33	56	47
$CD2^+CD16^+CD56^+CD3^-$		33	44	40
$CD2^+CD16^+CD56^-CD3^-$		50	78	67
$CD2^+CD16^+$		67	44	53
$45RA^+CD3^-CD4^{dim}$	rest/act T helper	83	14	46
$45RA^+CD3^-$		100	43	69
$45RA^+CD3^+CD4^{low}$		83	43	62
$45RA^+CD3^+CD4^-$		67	71	69
$45RA^+CD3^+CD4^+$		0	43	23
$45RA^+CD3^+$		50	43	46
$45RO^+CD3^-CD4^{dim}$		67	29	46
$45RO^+CD3^-$		50	14	31
$45RO^+CD3^+CD4^{low}$		50	29	38
$45RO^+CD3^+CD4^-$		50	86	69
$45RO^+CD3^+CD4^+$		50	43	46
$45RO^+CD3^+$		67	57	62
$CD3^-$		100	43	69
$CD3^+CD4^-$		83	71	77
$CD3^+CD4^+$		17	29	23
$CD3^+$		83	57	69

Immune cells	Aliquot	Sensitivity (%)	Specificity (%)	Accuracy (%)
$CD4^{dim}$	rest/act T helper	67	14	38
$CD3^{-}CD4^{-}$		100	57	77
$45RA^{+}CD3^{-}CD8$	rest/act T suppressor	50	71	62
$45RA^{+}CD3^{-}$		100	43	69
$45RA^{+}CD3^{+}CD8^{low}$		67	57	62
$45RA^{+}CD3^{+}CD8^{-}$		33	43	38
$45RA^{+}CD3^{+}CD8^{+}$		83	71	77
$45RA^{+}CD3^{+}$		83	57	69
$45RO^{+}CD3^{-}$		67	29	46
$45RO^{+}CD3^{+}CD8^{low}$		100	57	77
$45RO^{+}CD3^{+}CD8^{-}$		67	29	46
$45RO^{+}CD3^{+}CD8^{+}$		67	100	85
$45RO^{+}CD3^{+}$		67	57	62
$CD3^{-}$		100	43	69
$CD3^{+}CD8^{-}$		33	14	23
$CD3^{+}CD8^{+}$		83	86	85
$CD3^{+}$		67	57	62
$CD8^{+}CD3^{-}$		67	71	69
$CD3^{-}CD8^{-}$		83	29	54
$CD3^{+}CD4^{+}CD8\beta^{+}CD8^{+}$	T cells	33	67	53
$CD3^{+}CD4^{+}CD8\beta^{+}CD8^{+}$ (proportion of $CD3^{+}$ cells)		50	44	47
$CD3^{-}CD4^{low}CD8\beta^{low}$		83	44	60
$CD3^{-}CD8\beta^{dim}CD8^{-}$		67	22	40
$CD3^{-}CD8^{+}CD8\beta^{-}$		50	56	53
$CD3^{-}$		100	44	67
$CD3^{+}CD4^{+}CD8\beta^{-}$		17	11	13

Immune cells	Aliquot	Sensitivity (%)	Specificity (%)	Accuracy (%)
$CD3^+CD4^+CD8\beta^+$	T cells	33	33	33
$CD3^+CD8\beta^{dim}CD8^-$		33	44	40
$CD3^+CD8\beta^+CD4^-$		50	78	67
$CD3^+CD8\beta^+CD8^{low}$		100	11	47
$CD3^+CD8\beta^+CD8^+$		50	78	67
$CD3^+CD8^+CD8\beta^-$		83	11	40
$CD3^+$		33	67	53
$CD3^-CD5^+$	TCR	50	67	58
$CD3^-$		100	44	67
$CD3^-CD5^-TCRab^+TCRgd^-$		83	67	73
$CD3^-CD5^-TCRab^+$		67	22	40
$CD3^-CD5^-TCRab^+TCRgd^+$		83	44	60
$CD3^-CD5^-TCRab^+TCRgd^-$		67	11	33
$CD3^-TCR^+CD5^+$		100	50	75
$CD3^+$		67	67	67
$CD3^+CD5^-TCRab^+$		100	33	60
$CD3^+CD5^-TCRab^+TCRgd^+$		100	56	73
$CD3^+CD5^-TCRab^+TCRgd^-$		83	11	40
$CD3^+CD5^+TCRab^+$		100	50	71
$CD3^+CD5^+TCRab^+TCRgd^+$		50	75	64
$CD3^+CD5^+TCRgd^+$		50	75	64

Appendix K. FLDA classification model for the onset of aGvHD

The FLDA classifier built using immune cells $CD3^+CD4^+CD8\beta^+$ and samples taken between 7 and 21 days post-transplant, had the highest sensitivity (86%) and specificity (100%) among the consistent classifiers.

The unknown parameters in the signal plus noise model (Equation 1.1) were estimated using the training dataset via the EM algorithm. The training dataset is consists of observed values Y_{ij} included 21 aGvHD and 3 non-GvHD patients with samples taken between 7 and 21 days post-transplant. Linear B-splines with weekly knot placement were used to model the observed data. At the end, the observed values were divided into different elements:

1. $\lambda_0 = \begin{matrix} -6.6980 \\ -2.4241 \\ -0.6519 \end{matrix}$ for each knot;

2. Class signals $\Lambda\alpha_i$, $\Lambda = \begin{matrix} 1.7458 \\ -0.4414 \\ 0.9158 \end{matrix}$ for each knot and $\alpha_i = \begin{matrix} -2.5267 \\ 2.5267 \end{matrix}$ for each class

3. A B-spline matrix denoting these first three parameters for each j (columns) representing each knot (j = 7, 14, 21) and each i (rows) representing each time unit (j = 7, 8, 9, ... 21) (values were rounded to two decimal place).

$$
S_{ij} = \begin{matrix}
-0.15 & 0.42 & -0.44 \\
-0.18 & 0.36 & -0.33 \\
-0.21 & 0.3 & -0.21 \\
-0.24 & 0.24 & -0.1 \\
-0.27 & 0.18 & 0.02 \\
-0.3 & 0.12 & 0.13 \\
-0.33 & 0.06 & 0.25 \\
-0.36 & 0 & 0.36 \\
-0.33 & -0.06 & 0.25 \\
-0.30 & -0.12 & 0.13 \\
-0.27 & -0.18 & 0.02 \\
-0.24 & -0.24 & -0.1 \\
-0.21 & -0.3 & -0.21 \\
-0.18 & -0.36 & -0.33 \\
-0.15 & -0.42 & -0.44
\end{matrix}
$$

4.. For the test data p1 with samples taken at 7, 14, and 21 days post-transplant:

$$
S_X = \begin{matrix}
-0.15 & 0.42 & -0.44 \\
-0.36 & 0 & 0.36 \\
-0.15 & -0.42 & -0.44
\end{matrix}
$$

Weight values can be determined using the estimated parameters from the FLDA classifier via Equation 1.3.

$weigh = -1.0823 \quad 0.0123 \quad -0.1767$, for each sampled time point.

Global base values $S_X \lambda_0 = \begin{matrix} 0.2718 \\ 2.2034 \\ 2.3000 \end{matrix}$

5. Classification of p1 can be made using Equation 1.4. If the linear discriminant value is negative, new data will be classified into the aGvHD patient group and vice versa for non-GvHD patient group.

For example, for a new patient with values $X = \begin{matrix} 0.92 \\ 2.77 \\ 3.63 \end{matrix}$ from samples taken at 7, 14, and 21 days post-transplant, the linear discriminant value $\hat{\alpha}_x = weight \cdot (X - S_X \lambda_0)$ is calculated to -0.9. The new patient is classified into the aGvHD class ($\hat{\alpha}_x \prec 0$).

Appendix L. FLDA classification model for the onset of cGvHD

The FLDA classifier built using immune cells $45RO^{+}CD3^{-}CD4^{dim}$ in proportion to PBMC and samples taken between 21 and 0 days prior to aGvHD diagnosis, had the highest estimated 86% sensitivity and 86% specificity (Table 4.1), excluding the inconsistent classifiers.

The unknown parameters in the signal plus noise model (Equation 1.1) were estimated using the training dataset via the EM algorithm. The training dataset is consists of observed values Y_{ij} included 7 aGvHD & cGvHD and 7 aGvHD only patients with samples taken between 21 and 0 prior to aGvHD diagnosis. Linear B-splines with weekly knot placement were used to model the observed data. At the end, the observed values were divided into different elements:

1. $\lambda_0 = \begin{matrix} -66.4930 \\ -10.1525 \\ -16.8377 \\ -13.1379 \end{matrix}$ for each knot;

2. Class signals $\Lambda\alpha_i$, $\Lambda = \begin{matrix} 0.1042 \\ -7.3447 \\ -4.2339 \\ 2.8252 \end{matrix}$ for each knot and $\alpha_i = \begin{matrix} -3.0568 \\ 3.0568 \end{matrix}$ for each class

3. The first three parameters are denoted by the specified B-spline matrix for each j (columns) representing each knot (j = -21, -14, -7, and 0) and each i (rows) representing each time unit (j = -21, -19, -18, -17, ... and 0) (values were rounded to two decimal place).

$$S_{ij} = \begin{matrix} -0.09 & 0.2 & -0.43 & 0.4 \\ -0.12 & 0.21 & -0.34 & 0.3 \\ -0.14 & 0.22 & -0.26 & 0.2 \\ -0.17 & 0.23 & -0.18 & 0.1 \\ -0.19 & 0.24 & -0.10 & 0 \\ -0.22 & 0.25 & -0.02 & -0.1 \\ -0.24 & 0.26 & 0.06 & -0.2 \\ -0.27 & 0.27 & 0.14 & -0.29 \\ -0.27 & 0.19 & 0.14 & -0.21 \\ -0.27 & 0.12 & 0.14 & -0.13 \\ -0.27 & 0.04 & 0.14 & -0.04 \\ -0.27 & -0.04 & 0.14 & 0.04 \\ -0.27 & -0.12 & 0.14 & 0.13 \\ -0.27 & -0.19 & 0.14 & 0.21 \\ -0.27 & -0.27 & 0.14 & 0.29 \\ -0.24 & -0.26 & 0.06 & 0.2 \\ -0.22 & -0.25 & -0.02 & 0.1 \\ -0.19 & -0.24 & -0.1 & 0 \\ -0.17 & -0.23 & -0.18 & -0.1 \\ -0.14 & -0.22 & -0.26 & -0.2 \\ -0.12 & -0.21 & -0.34 & -0.3 \\ -0.09 & -0.2 & -0.43 & -0.4 \end{matrix}$$

4. For the test data p19 with samples taken at 21, 15, 7, and 0 days prior to aGvHD diagnosis:

$$S_X = \begin{matrix} -0.09 & 0.2 & -0.43 & 0.40 \\ -0.24 & 0.26 & 0.06 & -0.20 \\ -0.27 & -0.27 & 0.14 & 0.29 \\ -0.09 & -0.20 & -0.43 & -0.4 \end{matrix}$$

Weight values specific to these sampled time points can be determined using the estimated parameters from the FLDA classifier via Equation 1.3.

$weight = 0.0762 \quad -0.1436 \quad 0.1191 \quad 0.1091$, for each sampled time point.

Global base values $S_X \lambda_0 = \begin{matrix} 5.8992 \\ 15.0097 \\ 14.2864 \\ 20.4889 \end{matrix}$

5. Classification of p19 can be made using Equation 1.4. If the linear discriminant value is negative, new data will be classified into the aGvHD & cGvHD patient group and vice versa for aGvHD only patient group.

For example for a new patient with values $X = \begin{matrix} 13.3 \\ 23.4 \\ 12.6 \\ 13.6 \end{matrix}$ from samples taken at 21, 15, 7 and 0 days prior to aGvHD diagnosis, the linear discriminant value $\hat{\alpha}_x = weight \cdot (X - S_X \lambda_0)$ is calculated to -1.59. The new patient is classified into the aGvHD & cGvHD group ($\hat{\alpha}_x \succ 0$).