On the Identification of Associations between Flow Cytometry Data, Systemic Sclerosis and Cancer

by

Hongtai Huang

A dissertation submitted to Johns Hopkins University in conformity

with the requirements for the degree of Doctor of Philosophy

Baltimore, Maryland

2015

ABSTRACT

This work seeks to develop reliable biomarkers of disease activity, progression and outcomes through the identification of significant associations between high-throughput flow cytometry data and a scleroderma clinical phenotype – initially, interstitial lung disease (ILD) - which is the leading cause of morbidity and mortality in Systemic Sclerosis (SSc). A specific aim of the work involves developing a clinically useful screening tool (hereafter a filter). Such a filter could yield accurate assessments of disease state such as the risk or presence of SSc-ILD, the activity of lung involvement and the possibility to respond to therapeutic intervention. Ultimately this instrument should facilitate a refined stratification of SSc patients into clinically relevant subsets at the time of diagnosis and subsequently during the course of the disease, preventing bad outcomes from disease progression or unnecessary treatment side effects. This role could involve a scenario in which an SSc patient passes the presumptive (FVCstpp) test for ILD, but the filter indicates that their flow cytometry (FC) profile is consistent with ILD. In such a case, a physician might: 1) increase frequency of testing to detect early development of ILD; 2) implement more sophisticated diagnostic procedures (e.g., high resolution

chest CT scan - HRCT) to confirm the presence of ILD; and 3) consider prophylactic disease modifying treatments. Note that the intention of this research is not to develop screening tools that merely aim at predictive accuracy, but to produce methods that also contribute to the understanding of disease mechanisms. Having used ILD as phenotype, subsequent analyses in this thesis used different phenotypes: antiTopoisomerase (ATA), antiCentromere Anti Nuclear Antibodies (these antibodies are most strongly associated with diffuse and limited systemic sclerosis respectively) and cancer. This research was based on clinical and peripheral blood flow cytometry data (**I**mmune **R**esponse **I**n **S**cleroderma, IRIS) from consented patients followed at the Johns Hopkins Scleroderma Center.

Methods. The methods utilized in the work involve: (1) data mining (Conditional Random Forests - CRF) to identify subsets of FC variables that are highly effective in classifying ILD patients; (2) Gene Set Enrichment Analysis (GSEA) to further refine FC subsets; (3) stochastic simulation and Classification and Regression Trees (CART) to design, test and validate ILD filters; and (4) Stepwise Generalized Linear Model (GLM) regression and Drop-in-Deviance testing to identify minimal size, best performing models for predicting ILD status from both FC and selected clinical variables.

Results. IRIS flow cytometry data provides useful information in assessing the ILD status of SSc patients. Our hybrid analysis approach proved successful in predicting SSc patient ILD status with a high degree of success (out-of-sample > 82%; training data set 79 patients, validation data set 40 patients). Pre-partitioning patients into groups using CART significantly increased validation performance to 95% successful ILD identification. When the phenotype was Cancer, FC subsets, created through ranked Student t Test scores and point-wise GLM were statistically significant ($p < 0.05$) using GSEA. After applying Stepwise GLM on the CRF FC subsets, four FC variables were observed to be highly associated with Cancer in SSc patients. An ILD-Cancer GSEA intercomparison was made (use the best ILD FC set with cancer as the phenotype, and vice-versa) showed that GSEA results were highly phenotype-specific. Other phenotypes including ATA and ACA were also analyzed and found to be statistically significantly associated with certain subset of FC variables, but with different FC set sizes (38 and 6 respectively) based on the CRF-GSEA-Stepwise GLM algorithm.

In future research, HRCT confirmation of patient ILD status will be a critical next step in developing additional confidence with our approach (and

the appropriateness of an 80% FVCstpp threshold for presumptive ILD determination).

Advisor: Dr. J. Hugh Ellis, Department of Geography and Environmental Engineering, Johns Hopkins University

Readers: Dr. Thomas A. Burke, Department of Health Policy & Management, Bloomberg School of Public Health, Johns Hopkins University

Dr. Benjamin F. Hobbs, Department of Geography and Environmental Engineering, Johns Hopkins University

ACKNOWLEDGEMENTS

This doctoral study is not only an academic pursuit but also a journey of self-improvement. Along the way, I received considerable help and support.

First and foremost, my appreciation is dedicated to my advisor, Dr. J. Hugh Ellis, for his belief in me, his understanding, altruistic support and constant care. I still remember that in the fall of 2011 when I was still a master's student applying for PhD program, he decided to financially support me to be a research assistant without even working with me previously. In the past three years, not only did he teach me tremendous amounts of academic knowledge, but he also showed me how to be a decent and honorable man in a Canadian way. He is my provider, my teacher, my role model and my academic father. It has been an honor and pleasure to work with him.

I would like to thank our collaborators in the Division of Rheumatology, JHMI. Dr. Anthony Rosen's first class academic guidance extraordinarily opened my mind and his generous financial support that made this work possible. Drs. Francesco Boin and Andrea Fava helped me

unravel a large number of medical puzzles and graciously provided me with the IRIS data sets along with Tara Guhr, Raffaello Cimbro and Luca Magnani's assistance. Their contributions are essential to this research.

I am also very grateful for the kindness, encouragement and support of my mentors Drs. Alan Stone, Benjamin Hobbs, Erica Schoenberger, Thomas Burke and Ciprian Craniceanu.

Last but not the least, many thanks are given to my Sangha friends, especially Chogden, Gyaltsen, Matthew, Menla, Anne and Yenny, my family, my friends, particularly Angie, Alex, Ben, Hank, Hongru, Fengwei, Jianyin, Joe, Keith, Liang, Lin, Matt, M.T., Neil, Saamrat, Sean, Stephanie, Valerie, Venkat, and the Ellis family. It is my good fortune to know all of you. Your kind support has made me a better person.

To Guan Shi Yin (Avalokiteshvara)

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

A. INTRODUCTION

A.1 Systemic Sclerosis

Systemic Sclerosis is an autoimmune disorder - it is "*a condition that occurs when the immune system mistakenly attacks and destroys healthy body tissue*" (Goronzy & Weyand, 2007; Siegel & Lipsky, 2009). SSc can have severe effects, principal among them extensive fibrosis, vascular alterations and autoantibody response (Gabrielli, Avvedimento & Krieg, 2009; Boin & Rosen, 2007). SSc is classified into limited and diffuse forms depending on the extent of skin involvement. Both subsets can manifest progression to visceral organ involvement, e.g., lungs, heart, gastrointestinal tract and kidneys (Harris & Rosen, 2003). Although this type of classification identifies distinct clinical phenotypes, it remains inadequate to fully capture the spectrum and heterogeneity of SSc clinical manifestations (Gabrielli, Avvedimento & Krieg, 2009). Limited cutaneous SSc often manifests as CREST Syndrome (**C**alcinosis, **R**aynaud's phenomenon, **E**sophageal dysfunction, **S**clerodactyly and **T**elangiectasias; Winterbauer, 1964). Both SSc types can become life-threatening,

Interstitial lung disease (ILD) is a major cause of death in SSc patients. In ILD, sections of lung tissue become hardened and scarred and thus lose function. Lung transplantation is often not an option for patients with severe ILD (De Cruz & Ross, 2013).

particularly pulmonary fibrosis (or Interstitial Lung Disease, ILD) which is an important cause of morbidity and frequent cause of death in SSc patients (Steen, 1998). The essential nature of ILD is: "*the majority of SSc-ILD patients show replacement of the normal lung parenchyma with inflamed and fibrotic tissue, which is ineffective for gas exchange*" (Luo & Xiao 2011; Harrison et al. 1990). Additionally, SSc patients are more susceptible than the general population to other severe diseases including a variety of malignancies (Shah & Rosen, 2011).

As is the case with other autoimmune disorders, there are no curative therapies but only treatments aimed at halting progression towards end-stage disease. Due to limited knowledge about the role of autoimmunity in the pathogenesis of SSc, conventional treatments such as anti-inflammatory and immunosuppressant therapies are typically poorly effective (Boin & Rosen, 2007).

There are three main obstacles preventing a full understanding of SSc and the development of effective targeted therapies. First, there is extreme heterogeneity in clinical manifestations among different SSc patients. The disease course is highly variable in terms of onset, timing, intensity of symptoms, patterns of organ involvement and response to therapy. It has

been suggested that susceptibility to SSc varies in accordance with certain demographic factors such as gender, race and age (Chifflot et al., 2008). For example, high female-to-male ratios were consistently reported, with males developing in general more severe disease. Non-Caucasian patients and in particular African Americans tend to have an earlier onset of SSc, a more aggressive clinical course and higher mortality (Gelber et al. 2013). Schachna et al. (2003) identified "*increasing age at scleroderma onset as a risk factor for pulmonary arterial hypertension (PAH)*" and Perez-Bocanegra et al. (2010) discovered that there exist "*differences in SSc clinical features and survival*" for different age groups.

The second major challenge derives from the occult nature of early immune effector's pathways and the complex interaction of multiple humoral or cellular mediators, making the identification of the key drivers of clinical phenotypes difficult. Subsumed within this challenge is the difficulty in measuring and characterizing immune response (Whitfield et al. 2003; Chung and Utz 2004; Warrington et al. 2006; Boin et al. 2008; Salamunić 2010). Several levels of evidence support the involvement of the immune system, particularly during early stages of SSc or at the time of disease activity within specific target tissues (e.g., lungs). Nonetheless, the

relationship between a sustained immune response and the progression toward different clinical outcomes remains poorly understood. In addition, despite convincing in-vitro data linking innate and adaptive immunity to aberrant collagen synthesis and endothelial cell dysfunction, no reliable and accurate measure of the ongoing immune response has been defined in-vivo in SSc patients (Salamunić 2010; Boin et al. 2008; Cracowski et al. 2001). Pathologic studies on SSc patients with early lung disease showed that fibrosis is preceded by the presence of a mixed interstitial inflammatory infiltrate spilling into the alveolar spaces (alveolitis) composed mainly of macrophages, lymphocytes (notably T cells), granulocytes and other accessory cells (Harrison et al. 1990). As the disease progresses, deposition of collagen and thickened alveolar walls substitute air spaces with less evidence of inflammation. It is plausible that with early detection and treatment of lung inflammation, SSc patients may avoid progression to severe pulmonary fibrosis (Varga, 2014). T lymphocytes seem to have a central role and are required for initiation and propagation of the fibrotic lung insult. In SSc patients with alveolitis, T cell counts are increased in the pulmonary interstitium on lung biopsies and in Bronchoalveolar lavage (BAL) fluids. CD8+ T cells with an activated phenotype predominate and correlate with more severe pulmonary fibrosis (Yamadori et al. 2000; Luzina

et al. 2003). Previous studies have shown that increased frequency of circulating T cells exhibiting a "polarized" phenotype (i.e., T cells manifesting specific patterns of cytokine secretion) are significantly associated in SSc patients with the presence of pulmonary fibrosis and lung disease progression (Boin et al 2008; Truchetet et al. 2010). Despite all this evidence, it remains unclear how T cells contribute with their unique features and function to the pathogenesis of SSc at different times along disease progression, which is the third obstacle in understanding SSc. Elucidating the close temporal as well as biological relationship that exists between abnormal immune activation and the clinical manifestations present in SSc may allow for identification of novel and specific cellular as well as molecular probes to monitor disease activity, predict with accuracy clinical outcomes and ultimately design novel disease-specific therapeutic strategies.

This thesis will therefore concentrate on different subsets of T cells.

Potential participants in SSc pathogenesis include: (1) auto-antigens (Rosen & Casciola-Rosen, 1999, 2009; Casciola-Rosen, Anhalt & Rosen, 1994); (2) antigen presenting cells (APCs) (Leon et al., 2000; Alexander & Wahl, 2011); (3) Interleukin 2 (IL-2) (Burroughs et al., 2006, 2008; Isaeva & Osipove, 2009a,b); (4) Type I interferon (Hall & Rosen, 2010); (5) T

lymphocytes or T cells (Leon et al., 2000, 2004; Chao, et al., 2004; Burroughs et al., 2006, 2008; Carneiro et al., 2005, 2007; Isaeva & Osipove, 2008; Baltcheva, 2010; Alexander & Wahl, 2011; Velez de Mendizabal et al., 2011; and Saeki & Iwasa, 2009, 2010). A possible mechanism for autoimmunity is shown in Figure A-1 which illustrates a series of interrelated biological events.

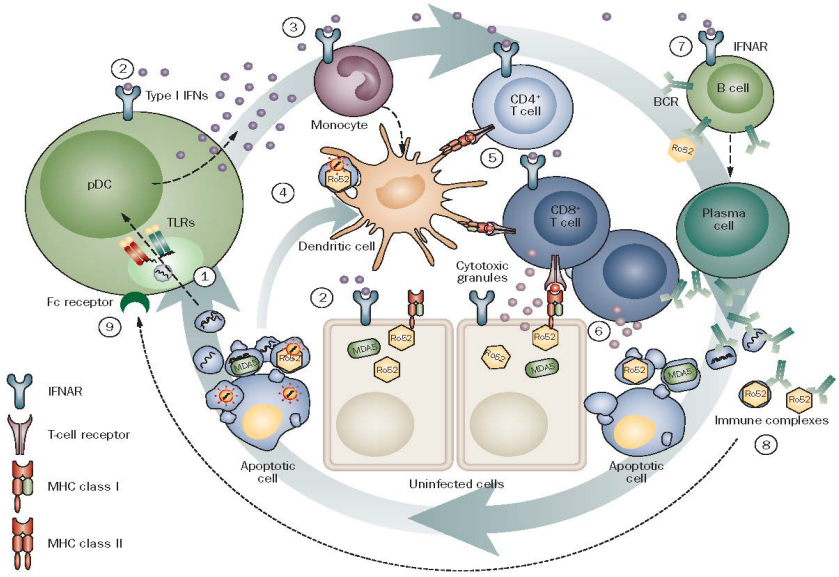


Figure A-1. A Possible Mechanism for Autoimmunity

Source: Hall & Rosen, 2010.

After plasmacytoid Dendritic Cells (pDCs) recognize virus in the form of nucleic acid in apoptotic debris, a large amount of type I IFN is rapidly produced in response, which triggers three processes: (1) "*self-amplification of the Toll Like Receptor (TLR) pathway in pDCs*"; (2) the

state of the target cells becomes antiviral; and (3) monocytes are differentiated and Dendritic Cells (DCs) activated. DCs will "*process and present self and viral antigens derived from dying cells*" and activate autoreactive CD4+ and CD8+ T cells for which survival would be promoted by type I IFN receptor signals. These signals also "*enhance the cytotoxic activity of Cytotoxic T Lymphocytes (CTLs) which eliminate uninfected host cells, expressing large quantities of autoantigens via the granzyme B pathway*". Remnants of dying cells are consumed by DCs and "*presented for recognition by T cells in a self-amplifying loop*" (Hall & Rosen, 2010).

A.2 Literature Review - Methods

Different quantitative approaches have been used to investigate SSc for purposes of prediction and explanation. In particular, various methods have been applied in order to: 1. better understand the mechanism of systemic autoimmunity in general; 2. provide improved insights concerning the SSc patients population in terms of epidemiological characteristics; 3. identify association between certain biomarkers and clinical manifestation or measurements of SSc; and 4. make predictions of disease outcomes. Examples of each are summarized in the literature review of this thesis.

They fall into two major categories: mechanistic modeling and data-based (or data-driven) analysis.

A.2-1 Mechanistic Models

Although the pathogenesis of SSc is not yet fully known, it has been suggested that all systemic autoimmune diseases may share common underlying mechanisms (Wahren-Herlenius & Dorner, 2012). Most mechanistic models of autoimmunity are built using ordinary differential equations. Waniewski and Waniewski & Prikrylova (1988) mathematically described autotolerance and autoimmunity considering the effect of plasmapheresis and immunosuppression. Nevo et al. (2004) presented a spatio-temporal model based on the concept of "comprehensive immunity" which views autoimmunity as a "special case" of immunity. They suggested that autoimmunity provides a protective mechanism for the Central Nervous System (CNS), e.g., preventing the CNS from degenerating into a more chaotic state. Leon et al. (2000, 2004) proposed a cross-regulation model that implies a "bi-stable" state – autoimmunity or tolerance – a state in which effector T cells and regulatory T cells coexist in a balanced manner. It was argued that there exist tradeoffs between the risk of autoimmunity and reactivity of the system. The role of Antigen Presenting Cells (APC) was

emphasized in the interaction between cytotoxic T cells and regulatory T cells. Burroughs et al. (2006) argued that cytokines such as IL-2 are actively involved in autoimmunity and therefore should be included in mathematical models. Leon & Garcıa-Martınez (2011) explicitly include IL-2 into their cross-regulation model. Iwami et al. (2007) developed a mathematical model based on a personal immune response function and a target cell growth function. Alexander and Wahl (2011) emphasized the importance of including professional[1] APC in modeling the mechanism of autoimmune diseases, which is consistent with the theory offered by Hall & Rosen (2010) on the self-amplifying nature of type I IFN production and tissue damage in systemic autoimmunity. Velez de Mendizabal et al. (2011) used the T cell cross-regulation model to analyze the "*relapsing-remitting dynamics*" of Multiple Sclerosis. Saeki & Iwasa (2009, 2010) used a fitness function to explain "*the advantage of having regulatory T cells*" and identified an optimal number of regulator T cells.

[1] Professional Antigen Presenting Cells are APC that produce a co-stimulatory signal that activates T cells. In our case, the co-stimulatory molecule is MHC II (Major Histocompatibility Complex II). pAPC includes Dendritic Cells, Macrophages and certain B cells. (Child, 2006)

A.2-2 Data-Driven Models

Traditional approaches for analyzing SSc integrate data from biological measurements, physician insight, clinical experience and other sources and thus are not purely data driven (Mathai et al., 2010; Mathian et al., 2012). Data-driven analysis has gradually entered the arena of autoimmunity research including SSc. Statistical methods have been applied to three levels of data: population, genetic and cellular. At the population level, the European League Against Rheumatism (EULAR) Scleroderma Trials And Research (EUSTAR) group collected the "*Minimal Essential Data Set*" (MEDS) with 3656 SSc patients from 102 centers and 30 countries (Distlar et al. 2009). They found that the association between autoantibody status and clinical manifestations of SSc was stronger than that between autoantibody and SSc subtypes (diffuse vs. limited) on a cross-sectional multivariate analysis. EUSTAR (2013) also concluded that "*pulmonary fibrosis, PAH and cardiac causes, accounted for the majority of deaths in SSc*" based on 5860 SSc patients. Another interdisciplinary registry of 1483 patients in Germany was established to better detect SSc patients with various disease manifestations (Hunzelmann et al., 2008). Recently, Merkel et al. (2012) performed an individual patient meta-analysis based on

629 diffuse SSc patients. Razykov et al. (2013) identified "*the association between sociodemographic and clinical variables and pruritus*" using multiple logistic regression based on 959 patients. Pruritus was determined to be statistically associated with "*the degree of skin involvement and gastrointestinal system involvement*".

Microarray data has been a major focus of SSc analysis on a genetic level. Symbolic Discriminant Analysis was used to make classifications and predictions of autoimmune disease using DNA gene expression data collected in peripheral blood mononuclear cells (PBMC) from 12 control individuals and 16 patients with either rheumatoid arthritis (RA) or systemic lupus erythematosus (SLE) (Moore et al., 2002). Gene expressions of skin biopsies from four diffuse SSc patients and four normal volunteers were analyzed using hierarchical clustering (Whitfield et al., 2003) while genetic programming, an extension of genetic algorithms, was employed to identify "*differences in patterns of gene expression of skin biopsies*" from control and case groups of SSc patients (Paul & Iba, 2006). Microarray data of a molecular phenotype (combined proteome and transcriptome) from non-obese diabetic (NOD) mice was used to explore pathways of autoimmune

Data driven research analyses on Scleroderma are fairly recent. SSc data driven analysis using flow cytometry data is novel.

diabetes using two-way analysis of variance (ANOVA), k-mean clustering and principal component analysis (PCA) (Gerling et al., 2006). Duan et al. (2008) analyzed the gene expression of purified monocytes and T lymphocytes from 18 female SSc patients and 11 healthy female control subjects in an attempt to gain insights into the pathogenetic mechanisms of SSc. They suggested that "*leukocytes respond to cytokine [messenger RNA (mRNA)] locally in the vessels*". Lindahl et al. (2013) identified "*a strongly suppressed interferon-stimulated gene program in fibroblasts from fibrotic lung*" using microarray profiling.

Advanced techniques for acquiring data on a cellular level include: enzyme-linked immunosorbent assay (ELISA) which uses antibodies and color change to identify different entities (Ashihara et al., 2011); MACS MicroBeads Column (Miltenyi Biotec, 2014) which is a type of cell separation and culturing method; and flow cytometry (§B.1). In order to examine the association between PAH biomarkers, IFN-regulated gene expression and "*alternative activation pathways of SSc*", Christmann et al. (2011) analyzed: (1) experimental data of Peripheral Blood Mononuclear Cells (PBMCs) isolated using a MACS MicroBeads Column (Miltenyi Biotec, 2014); (2) microarray data of IFN-regulated and "PAH biomarker"

genes; (3) plasma measurement of Interleukin-13 (IL-13); and (4) IL-4 concentrations and flow cytometry data of CD14 and other cells. Rank correlation (Spearman's coefficient) and the paired Wilcoxon signed ranks test were also used to analyze flow cytometry data of Treg cells and ELISA measurement of TGF-β and IL-10 respectively to evaluate the role of Treg cells in SSc patients. Linear regression was applied to investigate the relations between severity score and activity index of SSc and case-control ratio of Treg cell count (Slobodin et al., 2010).

A review of the literature showed that few studies investigating SSc pathogenesis have been solely driven by quantitative data. A study on how certain SSc clinical phenotypes are associated with a group of cellular level biomarkers purely based on quantitative analysis can be useful in further understanding SSc mechanisms. This research therefore mainly addresses the quantitative association between certain SSc phenotypes and group of FC variable as a whole, namely FC set, and examines the clinical utility of the identified FC set based on statistical analyses.

A.3 Dissertation Outline

There are six more chapters after the Introduction in this dissertation. Chapter B provides detailed description of the flow cytometry and the IRIS

data set. Chapter C illustrates the main methodologies used. Chapter D presents results associated with different approaches and phenotypes. Chapter E includes discussions of the results, followed by conclusions (Chapter F) and suggestions for future research (Chapter G).

B. DATA

B.1 Flow Cytometry

Flow cytometry (FC) is a powerful tool used to analyze multiple characteristics of individual cells within heterogeneous populations (Shapiro, 2003; Picot et al., 2012). Through more than seven decades of innovation (Perfetto et al., 2004; Picot et al., 2012) flow cytometry has proven to be exceedingly useful in biological and medical studies, especially in the field of immunology (Hedley et al., 1983; Nicoletti et al., 1991; Vermes et al., 2000; Raja et al., 2013).

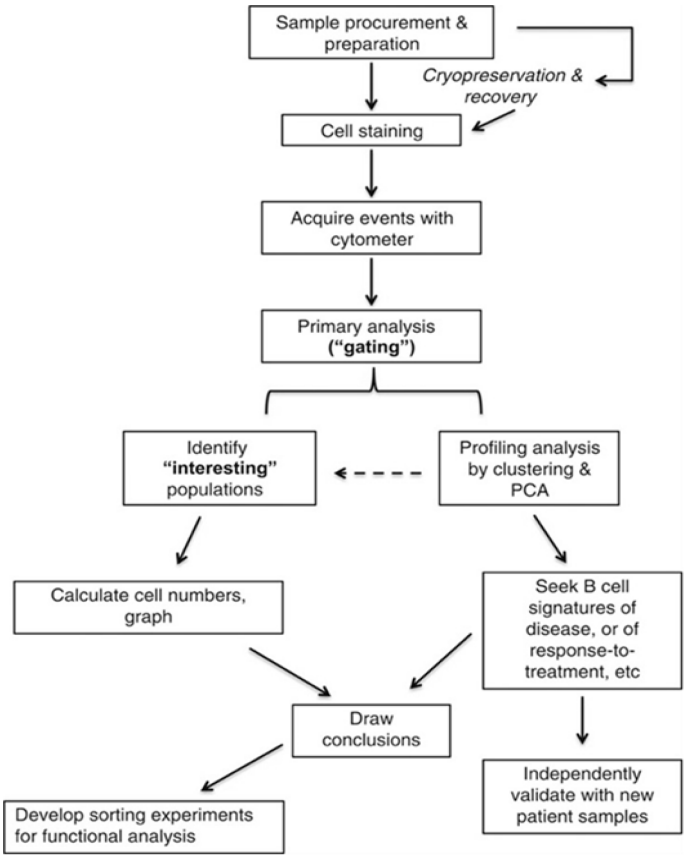


Figure B-1. Flow Cytometry Processes Sequence

The sequence of processes that constitute flow cytometry is shown in Figure B-1.

As shown below in Figure B-2, flow cytometer instruments have the following components: (1) the fluidics system; (2) laser; (3) optics; (4) detectors; and (5) the electronics and computer system. The fluidics system aligns cells (one at a time) using hydro-dynamic focusing (Lee et al., 2001). Individual cells are then excited with a laser beam that causes either forward or side scatter. The scattered light from each cell is then directed by optics to detectors that generate signals. A dedicated computing system then analyzes and converts these signals into useful statistics regarding characteristics of each cell (http://media.invitrogen.com.edgesuite.net/tutorials/4Intro_Flow/player.html).

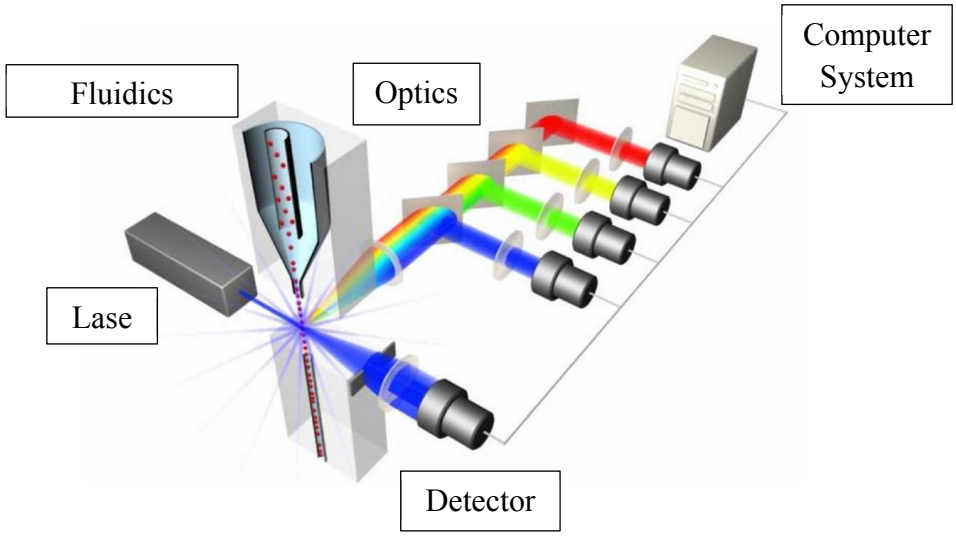


Figure B-2. Components of Flow Cytometer Instrument

Source: http://media.invitrogen.com.edgesuite.net/tutorials/4Intro_Flow/player.html

B.2 JHU Data Set

The data set used in this work (**I**mmune **R**esponse **I**n **S**cleroderma, IRIS, 2013) was provided by the Division of Rheumatology of the JHU School of Medicine. It is based on anonymous human subjects. Currently there are 158 SSc patients in total and each has 190 features grouped according to: 1. general background knowledge, e.g., age, sex and race; 2. clinical data such as presence of lung disease and skin severity score; 3. serology that indicates the presence and type of autoantibodies; 4. pulmonary function tests; 5. echocardiograms; 6. medications; and 7. T cell flow cytometry data.

The original 116 T cell flow cytometry variables[2] contained in the data set fall into 4 functional panels: memory, activation, polarization and traffic (data in this panel are particularly pertinent to skin and lung T cell migration). Each functional panel has different T cell subsets that are connected to each other through a hierarchical structure. An example of the memory panel is given in Figure B-3. All of the child nodes are expressed as percent of the parent node. Specific biological definitions of each acronym in the figure can be found in Table B-1.

[2] FC expressions are the values of FC variables. In the literature, these terms are sometimes used interchangeably.

This IRIS data set has been occasionally updated at times which is a normal course of events as procedures (gating) are refined, new patients are added and additional blood testing performed. Six versions of the data set exist so far, labelled by their dates: Jun. 21st, 2012, Aug. 2nd, 2012, Jan. 29th, 2013, Feb. 14th, 2013, Jul. 18th, 2013 and Mar. 8th, 2013. The version of the data set will be denoted below as IRISMMDDYY, e.g., IRIS071813 means data set updated on Jul. 18th, 2013. Although results based on different data sets can differ, these differences have consistently been inconsequential.

The hierarchical structure and FC variable definitions for the remaining three panels are provided in Appendix I.1.

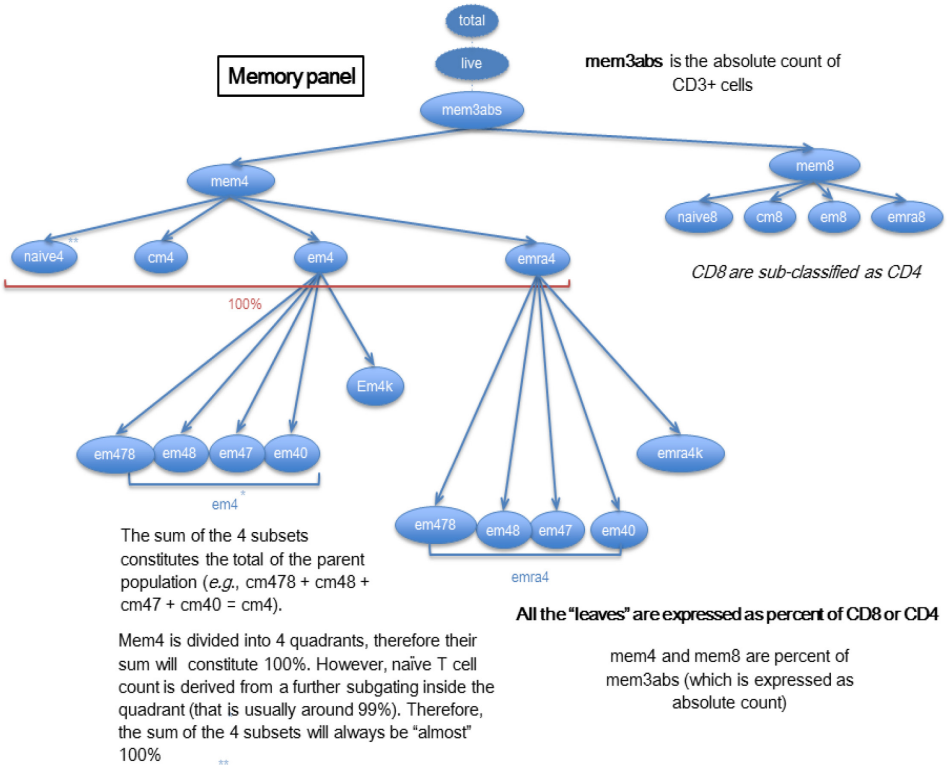


Figure B-3. Memory Panel Hierarchy

Source: Dr. Andrea Fava, Division of Rheumatology, JHMI

Table B-1 Memory Panel T cell Subset Definitions

Code	Subset Definition
mem3abs	CD3+ absolute count
mem4	CD3+/CD4+/CD8-
mem8	CD3+/CD4-/CD8+
memratio48	ratio of mem4 to mem8
naive4	CD3+/CD4+/CD8-/CD8-/CD45RA+/CCR7+/CD27+/CD28+ (Naïve T cells)
cm4	CD3+/CD4+/CD8-/CD8-/CD45RA-/CCR7+ (Central memory)
emra4	CD3+/CD4+/CD8-/CD8-/CD45RA+/CCR7- (Terminally differentiated "effector memory CD45RA+" cells)
em4	CD3+/CD4+/CD8-/CD8-/CD45RA-/CCR7- (effector memory)
emra478	CD3+/CD4+/CD8-/CD45RA+/CCR7-/CD27+/CD28+
emra47	CD3+/CD4+/CD8-/CD45RA+/CCR7-/CD27+/CD28-
emra48	CD3+/CD4+/CD8-/CD45RA+/CCR7-/CD27-/CD28+
emra40	CD3+/CD4+/CD8-/CD45RA+/CCR7-/CD27-/CD28-
cm478	CD3+/CD4+/CD8-/CD45RA-/CCR7+/CD27+/CD28+
cm48	CD3+/CD4+/CD8-/CD45RA-/CCR7+/CD27-/CD28+
em478	CD3+/CD4+/CD8-/CD45RA-/CCR7-/CD27+/CD28+
em47	CD3+/CD4+/CD8-/CD45RA-/CCR7-/CD27+/CD28-
em48	CD3+/CD4+/CD8-/CD45RA-/CCR7-/CD27-/CD28+
em40	CD3+/CD4+/CD8-/CD45RA-/CCR7-/CD27-/CD28-

cd4k	CD3+/CD4+/CD8-/CD57+
emra4k	CD3+/CD4+/CD8-/CD45RA+/CCR7-/CD57+
em4k	CD3+/CD4+/CD8-/CD45RA-/CCR7-/CD57+
naive8	CD3+/CD4-/CD4-/CD8+/CD45RA+/CCR7+/CD27+/CD28+ (Naïve T cells)
cm8	CD3+/CD4-/CD4-/CD8+/CD45RA-/CCR7+ (Central memory)
emra8	CD3+/CD4-/CD8+/CD45RA+/CCR7- (Terminally differentiated "effector memory CD45RA+" cells)
em8	CD3+/CD4-/CD8+/CD45RA-/CCR7- (effector memory)
emra878	CD3+/CD4-/CD8+/CD45RA+/CCR7-/CD27+/CD28+
emra87	CD3+/CD4-/CD8+/CD45RA+/CCR7-/CD27+/CD28-
emra88	CD3+/CD4-/CD8+/CD45RA+/CCR7-/CD27-/CD28+
emra80	CD3+/CD4-/CD8+/CD45RA+/CCR7-/CD27-/CD28-
cm878	CD3+/CD4-/CD8+/CD45RA-/CCR7+/CD27+/CD28+
em878	CD3+/CD4-/CD8+/CD45RA-/CCR7-/CD27+/CD28+
em87	CD3+/CD4-/CD8+/CD45RA-/CCR7-/CD27+/CD28-
em88	CD3+/CD4-/CD8+/CD45RA-/CCR7-/CD27-/CD28+
em80	CD3+/CD4-/CD8+/CD45RA-/CCR7-/CD27-/CD28-
cd8k	CD3+/CD4-/CD8+/CD57+
emra8k	CD3+/CD4-/CD8+/CD45RA+/CCR7-/CD57+
em8k	CD3+/CD4-/CD8+/CD45RA-/CCR7-/CD57+

C. METHODOLOGY

C.1 Overview

In the following sections, several methods are described that all essentially have the same purpose: to reduce the dimensionality of the data. The presumption made here is basically that not all 116 (or later, 112) flow cytometry variables are likely to be useful in either predicting the ILD status of SSc patients or gaining a better understanding of the etiology and pathogenesis of systemic sclerosis and its connections to selected phenotypes, mostly notably, ILD and cancer. Later, I also performed preliminary analyses on other phenotypes including Anti-topoisomerase I antibodies (ATA, or anti-Scl-70 antibodies) and anti-centromere antibodies (ACA).

The rest of the chapter starts with introducing the traditional statistical approach Principal Component Analysis (PCA) which is well known in reducing data dimensionality (§C.2). It was found that the resultant principal components cannot be biologically interpreted. Therefore, results of PCA were not used for the rest of this research. Next is description of four non-parametric data mining methods (Classification And Regression Tree, Random Forest, Conditional Random Forest and Support Vector Machines)

and related model performance evaluation procedures (§C.3). In Section §C.4, I described Gene Set Enrichment Analysis (GSEA) algorithm and how it was adapted in this research including its algorithm and permutation test. Having identified the best FC subset using data mining methods and GSEA, there were two analysis directions – drawing statistical inference and making predictions. The former is the combination of Generalized Linear Regression Model (GLM) and stepwise variable selection algorithm which will be illustrated in Section §C.5. The latter is the randomized filter design (introduced in §C.6) which essentially is a screening tool used to differentiate SSc patients with ILD from those without. Figure C-1 gives an illustration of the relations among the last three methods mentioned.

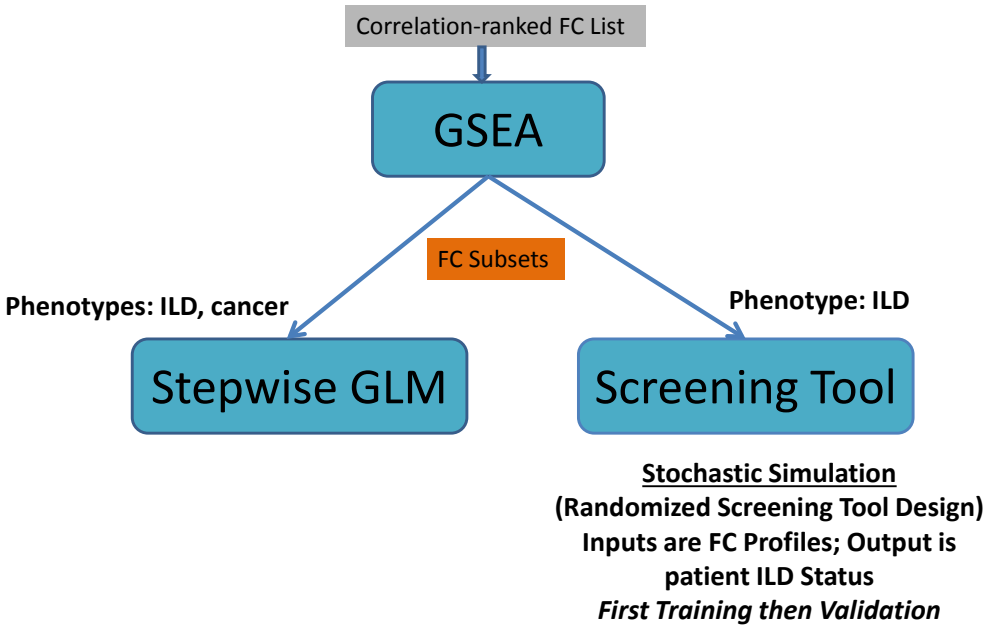


Figure C-1. Two Analysis Directions

C.2 Principal Component Analysis

Principal Component Analysis (PCA) is a mathematical procedure widely used in reducing data dimensionality without loss of useful information from original data set. In specific, PCA will create a set of new variables, namely Principal Components, that are linear combinations of the original variables (Pearson, 1901; Ramsey & Schafer, 2012). The coefficients of the original variables that are associated with each principal component are called loadings. Variables have large variance will tend to have large loadings (Wold, Esbensen, & Geladi, 1987). The principal components are closely related to each other. The first principal component will be constructed to have the largest variance and the second principal component will be established in the same way with one extra constraint that it will be perpendicular to the first one. Other principal components will be established similarly. For example, the third principal component will also have the largest variance and have to be orthogonal to the plane where the first and second principal component locate. All the created principal components will therefore be uncorrelated even if the original variables are correlated (Wold, Esbensen, & Geladi, 1987; Abdi & Williams, 2010; Ramsey & Schafer, 2012). The purpose is to define a few dimensions (the

first few PCs) that capture most of the variance, are uncorrelated, and might have an interpretation. Factor analysis (Child, D., 2006) is a set of procedures that further manipulates those components (including orthogonal and non-orthogonal rotations) in order to increase their interpretability.

In this research, interpretation of the principal components can be difficult in that we found that the linear combinations of FC variables unfortunately contained little understandable biological meaning. In the next two sections (§C.3 and §C.4), I will present a new method that is combination of data mining approach and Gene Set Enrichment Analysis. This method can maintain interpretability of the results while reducing data dimensionality. In future research, Factor Analysis could be applied with the same goal.

C.3 Data Mining and Partitioning Methods

Three major schools of statistical and data mining methods have been proven useful for medical diagnosis (Kononenko, 2001) including: 1) statistical pattern recognition method such as naïve Bayesian Classifiers (Domingos & Pazzani, 1997); 2) artificial neural networks (Bishop, 1995); and 3) inductive learning of symbolic rules such as the decision trees method (Breiman et al., 1984). In that the IRIS data set used in this study is featured

by high correlated predictor variables issue, naïve Bayesian Classifiers was not used because of its assumption that all the features in the data set of interest should be independent.

Although artificial neural networks can yield high predictive accuracy in medical diagnosis (Khan et al., 2001; Dreiseitl & Ohno-Machado, 2002), it has been found that "the network comes to clinical closure based on the settings of all variables in a pattern and that the impact of a single variable cannot be taken out of the context of a pattern" (Baxt, 1992). Lately, there was discovery that neural network may be useful in presenting information concerning contribution of each variable for estimating the response variables but with the condition that interpretation of model parameters can only be verified externally (Olden et al., 2004). This model was not adopted in this research because of its limitations in making statistical inference such as extracting variable importance information.

After removing observations whose FC profile was not complete (i.e., containing missing values), the number of observations is smaller than the number of covariates, resulting in the problem known as "Large P small N". Therefore, only non-parametric methods such as decision trees models were considered because parametric approaches tend to overfit the data set. Later,

in this research, after the dimensionality of the data set was reduced, parametric methods such as logistic regression model were revisited. In this section, I will describe the models that were evaluated in this research and the criteria used to determine the best model.

C.3-1 Models Implemented

Classification And Regression Trees (CART) (Breiman et al., 1984) is a modeling approach for classification and regression. Classification refers to the situation where the response is binary or categorical. Its regression interpretation is used here, that is, using predictor variables (flow cytometry expression) to predict a continuous (0,1) response (the probability of having ILD or Cancer). CART is a non-parametric procedure (there is no reliance upon data distribution) comprised of a sequence of recursive tests, with the outcome of a current Test determining the specifics of the next Test and terminated by stopping criteria. The first Test is to identify which FC variable is most important in accurately predicting ILD status and the value of that variable (from among all the values in the data set). There exist different metrics for importance depending upon whether CART is used for classification or regression. The following equations and corresponding

description regarding CART are from (Hastie et al., 2009). The measure of node impurity used here is residual sum of squares (RSS):

$$RSS = \sum \left(y_i - f\left(x_i \right) \right)^2$$

For each SSc patient *i (i = 1, 2, ..., N)* there are *p* FC expressions $x_i = \left(x_{1i}, x_{i2}, ..., x_{ip} \right)$ and a binary response y_i (ILD status). The FC data are partitioned into M regions $R_1, R_2, \ldots, R_M$. Response is modeled as a constant in each region

$$f\left(x \right) = \sum_{m=1}^{M} c_m I\left(x \in R_M \right)$$

The best $\widehat{c_m}$ is the average of y_i in region R_m:

$$\hat{c}_m = ave\left(y_i \mid x_i \in R_m \right)$$

which, from an implementation perspective, is not a helpful result because identifying the best binary partition on the basis of minimum sum of squares is, in general, computationally infeasible. The recourse is to follow a greedy algorithm: Starting with all of the data, a splitting variable *j* and split point *s* are selected which defines the half-planes

$$R_1\left(j, s \right) = \left\{ X \mid X_j \le s \right\} \quad and \quad R_2\left(j, s \right) = \left\{ X \mid X_j \ge s \right\}$$

Next, the splitting variable *j* and split point *s* are found that solve

$$\min_{j,s}\left[\min_{c_1}\sum_{x_i\in R_1(j,s)}\left(y_i-c_1\right)^2+\min_{c_2}\sum_{x_i\in R_2(j,s)}\left(y_i-c_2\right)^2\right]$$

For any choice *j* and *s*, the inner minimization is solved by

$$c_1=ave\left(y_i\mid x_i\in R_1\left(j,s\right)\right)\quad and\quad c_2=ave\left(y_i\mid x_i\in R_2\left(j,s\right)\right)$$

Having found the best split value *s* for FC variable *j*, the data are partitioned into two regions with the splitting process repeated for each region. The process is then repeated for all of the remaining regions. An additional complication is how large to grow the tree (equivalently, how many splits to perform). A large tree may over-fit the data whereas a small tree may fail to capture important structure in the data. The balance between these two extremes is achieved through validation (see § D.6).

Graphically, this gives rise to a tree-like structure shown below. We can see in Figure C-2 that the FC expression act4103 at value 1.525 was identified as most important (it is associated with the greatest decrease in node impurity). SSc patients whose act4103 expression is less than 1.525 are split to the left branch, those with act4013 expression greater than or equal to 1.525 are directed to the right. The process is repeated (it is recursive) with the next most important variable identified as memem4 at value 17.65, and so on.

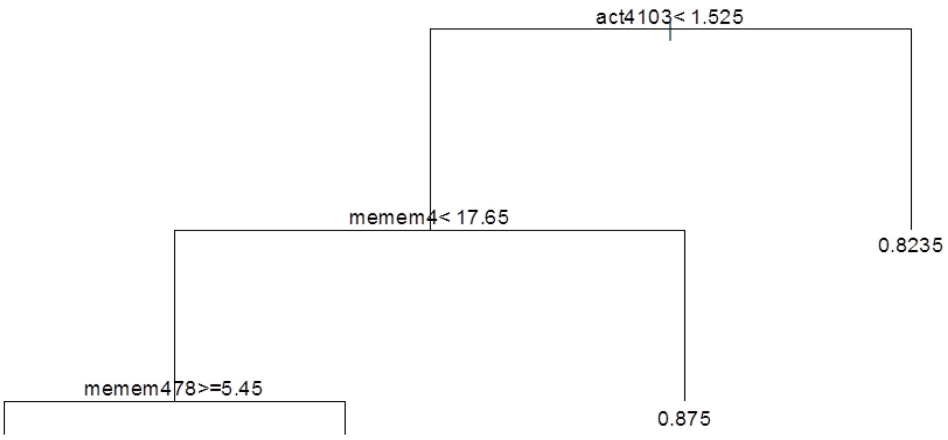


Figure C-2. Graphical Presentation of CART

Thus, the main elements of CART are (Nisbet et al. 2009): rules for splitting data at a node based on the value of one variable; stopping rules for deciding when a branch is terminal and can be split no more; and finally, a prediction for the target variable in each terminal node.

The stopping rules involve two considerations: (1) instances where subsequent splitting is impossible, i.e., a node contains only one patient or a node is pure (all patients are ILD or non-ILD); and (2) a pre-specified stopping criterion (our stopping criterion was less than 20 patients in a node, which is the default setting in R package "rpart" (Therneau et al., 2014)).

In comparison with traditional regression, CART has advantageous attributes beyond being independent of data distribution: (1) CART is relatively insensitive to outliers in the input variables; (2) Stopping rules can

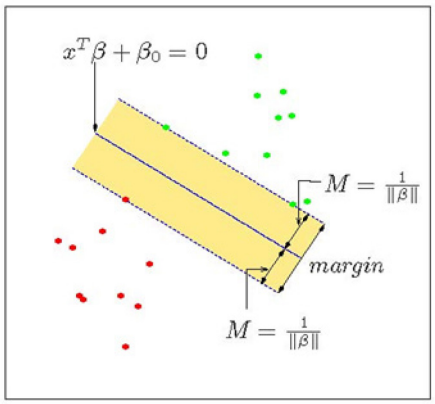
be relaxed to over-fit the data. The training tree can then be pruned back to a level that maximizes validation performance; and, (3) CART can re-use variables in different parts of the tree and possibly uncover complex interdependencies between sets of variables. (Nisbet et al. 2009)

The Random Forest (RF) modeling approach (Breiman, 2001) involves an ensemble of many regression or classification trees. Each tree in RF differs from CART in the following respects: 1. random (bootstrap) sampling of the original data is used to create training subsets (as opposed to using the entire data set); and 2. the splitting variables at each node in a tree are randomly chosen from a subset of covariates as opposed to the pool of all covariates. The output from an RF model is the average of the performance from all of the regression trees generated. (Breiman, 2001)

The Conditional Random Forest (CRF) modeling approach is similar to RF in that it is also an ensemble of trees - but with the following modification. The variable selection process is separated from the splitting criteria and involves a hypothesis testing procedure. The null global hypothesis - all stimulus variables are independent of the response - is tested by examining the partial hypotheses that each stimulus variable is independent of the response. Only when the null global hypothesis is

rejected does the variable selection process continue. This modification enforces the condition that each predictor variable selected as a splitting variable in each tree to be strongly associated with response variable through hypotheses testing under an unbiased conditional inference framework (Hothorn et al., 2006). This process exploits the discriminatory power of predictor variables and is especially important with our IRIS data set in which numerous covariates within the same panel are highly correlated. The Variable Importance Measures (VIMs), quantitative measurements of relative importance among predictor covariates, of RF can be unstable and suffer from "*correlation bias*" due to the effects of predictor variable correlation: 1. VIMs are not necessary connected to discriminatory power of stimulus variables; 2. the size of the group of correlated variables is consequential (Gregorutti et al., 2013; Toloşi & Lengauer, 2011); and, 3. VIMs do not "*directly reflect the coefficients in the generating model*" (Nicodemus et al., 2010). Strobl et al. (2008) showed that VIMs based on the conditional permutation scheme described above better match the coefficients associated with greatest predictor discriminatory power and that VIM stability was improved over that of the unconditional importance approach.

Support Vector Machines (SVM) (Cortes & Vapnik, 1995) is a binary linear classifier which takes predictor data as input; the output is a prediction function. In this application, flow cytometry data are the inputs with ILD status (0 or 1) the prediction output. FC data are represented as points in space, with prediction arranged ("mapped") into categories (0 = non-ILD; 1= ILD) separated by as large a distance as possible (i.e., the margin as shown in Figure C-3). Additional FC variables follow the same mapping that the model is trained on, with prediction being which category they are assigned. Extensions to nonlinear partitioning are accomplished by expanding the predictor variable space through so-called kernel functions (Hastie et al., 2009; Kecman, 2005).



Source: Hastie et al., 2009

Figure C-3. Support Vector Machine

C.3-2 Evaluation Criteria

Based on different purposes (either making diagnosis or predictions), statistical models were evaluated by various methods (Cook, N. R., 2008). In this thesis, I focused on predictive accuracy and goodness-of-fit given their appropriateness for the statistical methods evaluated.

C.3-2-1 Predictive Accuracy

50 Times holdout analysis was used to examine the out-of-bag (OOB) predictive accuracy of all the above-mentioned models. Each time of the holdout analysis was essentially an unequal size two-fold cross validation (Kohavi, 1995) with the modification that each time the training data set was created (by randomly subsampling approximately 90% of the original data) the remaining 10% was used as the test data set for validation. Because five models (CART, RF, CRF, SVM and mean-only model) were examined in total, it created 10 simultaneous hypothesis tests. In order to hold an overall confidence level of 95% for the combined set of hypothesis tests, each test could be regarded as significant if its p-value is below 5×10^{-3}, based on Bonferroni correction (Dunn, 1961) for multiple hypothesis tests.

Alternatively, Leave-One-Out Cross Validation (LOOCV) (Kohavi, 1995) was also used to evaluate predictive accuracy of different models. It is a special case of cross validation in which only one observation will be held out at a time and the remaining data will be used as training data. This process will be repeated for N times (N being the number of observations). Eventually, all the estimated responses will be compared with the observed values of responses.

The Receiver Operating Characteristics curve (ROC) (Zweig & Campbell, 1993) also served as a tool for evaluating model performance. The Receiver Operating Characteristic curve is a graphical means of assessing binary classifier performance. It is a graph of the fraction of true positives out of the actual positives (TPR = true positive rate, generally known as sensitivity – plotted on the ordinate) against the fraction of false positives out of the actual negatives (FPR = false positive rate which equals one minus specificity – plotted on the abscissa) for varying discriminant thresholds (Zweig and Campbell, 1993).

C.3-2-2 Goodness of Fit

The whole data set was used to construct the best candidate models without holding any observation. The model would then generate continuous estimation of response variable for each observation, which would become potential cut points for converting continuous response into binary values. ROC curves based on these cut points values were then plotted to indicate goodness of fit for the model of interest.

After the best model was found, Variable Importance Measures (VIMs) would be extracted from the best model. Statistical significance of the top certain number of most important variables as a group will be examined

using an algorithm named Gene Set Enrichment Analysis (GSEA), which will be illustrated in the next section.

C.4 Gene Set Enrichment Analysis

C.4-1 GSEA Algorithm

A gene set is a manipulatable number of genes in a typically very long DNA sequence (Yang et al. 2010). The key attribute here (that was discovered early on in genomics) is that analyses involving the study of only one gene at a time were of limited usefulness. What was needed was to examine sets of genes, with the determination of gene set size a critical issue. This in turn led to Gene Set Enrichment Analysis (GSEA) (Mootha et al. 2003; Subramanian et al., 2005). The idea of coordinated, multiple FC expression movement seemed applicable to this work, thus gene sets became FC sets. An example of the GSEA algorithm can be found in the "Methods" section of (Subramanian et al., 2005).

The following will describe how the GSEA algorithm was adapted to this research. Three main components of the GSEA algorithm are the ranked list, gene set of interest and random walk. First, any suitable metric that can measure the correlation between a given phenotype and gene expressions

can constitute the ranked list (Subramanian et al., 2005). In our analysis, correlation coefficients between flow cytometry variables and a specific phenotype (e.g., ILD) are computed and ranked. The second component - gene set of interest (FC set in this study) - can be determined in differing ways including literature-based information, biological guidance from experts and identification via model inference. In this research, the Variable Importance List from the best model Conditional Random Forest (CRF) (Strobl et al., 2009) was used to identify FC sets. Third, after the FC set is identified, a random walk is performed using ranked correlation coefficients between response ILD and all flow cytometry variables. The process involves moving the FC set down the ranked list from top to bottom and recording the running sum for each step. If a variable in the ranked list is encountered that is in the FC set, the following quantity is added to the running sum:

$$\sqrt{\frac{N-G}{G}}$$

otherwise add:

$$-\sqrt{\frac{G}{N-G}}$$

where G is size of the FC set and N is the total number of FC variables (116, or later 112). Below in Figure C-4 is a schematic representation of GSEA.

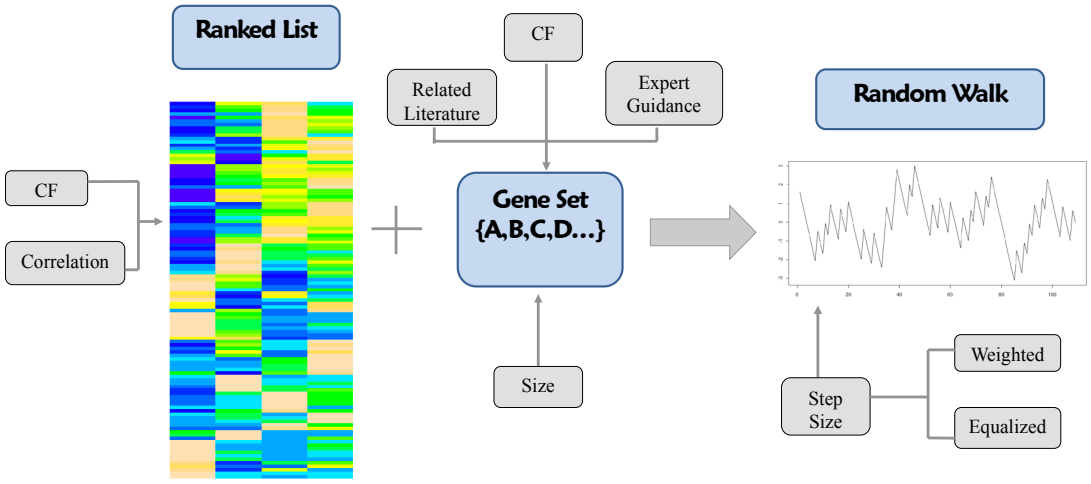


Figure C-4. Schematic Representation of GSEA Algorithm

Having obtained the random walk, the maximum deviation from zero (absolute value), namely Enrichment Score (ES) will be recorded in order to evaluate the degree of enrichment.

C.4-2 Permutation Test

To examine the statistical significance level of ES*, a permutation test (Subramanian et al., 2005) for 10,000 times will be performed. Basically, the response values of all the subjects in the data matrix will be shuffled while remaining part of the data will stay the same in each permutation. After each shuffling, a new data set will be generated and the ranked list, i.e., sorted Pearson correlation coefficients, will be recalculated. The same FC set

identified previously will go through the newly calculated ranked list, a new random walk and corresponding ES will therefore be obtained. For 10,000 permutations, there will be 10,000 values of ES. The p-value of this permutation test is the total number of ES that are at least as large as the observed ES*, divided by 10,000.

If a group of variables, or a set, is statistically significant, it means that the corresponding ES of the random walk will be a relative large number, and only very few simulated ES in the permutation test will be higher than the observed ES, i.e., the p-value of permutation test is smaller than prescribed significance level such as 0.05.

If the permutation test of the GSEA algorithm indicates that the FC set is statistically significant, there were two analysis directions – drawing statistical inference and making predictions. In the following two sections, I will present details regarding these analysis directions, starting with randomized filter design that is a novel screening techique differentiating SSc patients with ILD from those without based on the best FC set, followed by stepwise GLM, a tool used to draw statistical inference with respect to the association between the responses and FC variables.

C.5 Randomized Filter Design and Testing for ILD vs. no-ILD Classification

By using the new method combining data mining and GSEA, data dimensionality can be reduced. Essentially, a group of FC variables would be identified. In this research, it was found that none of the models implemented has high predictive accuracy when response variable is ILD (see §D.2). Therefore, the first analysis direction was to create a generalizable method and practical tool for assessing patient's ILD status given their FC data. The original motivation derived from the observation that assigning thresholds to individual FC expressions and applying these thresholds (in the form of a filter) had success in identifying SSc patient's ILD status. In our filter design, a patient is declared ILD if any of their FC expressions is above a positive or below a negative standardized threshold in the FC set.

Mathematically, the logic of a filter can be described as the follows.

Let's define the following parameters:

K – number of components, or FC variables in a filter;

I – number of patients;

FC_{ki} – binary variable indicates whether the i^{th} patient's k^{th} FC expression is above an upper threshold U_i or below a lower threshold L_i ;

Z_i - binary variable indicating whether the i^{th} patient has ILD or not. We have:

$$FC_{ki} = \begin{cases} 1, & if\ FC_{ki} > U_i\ or\ FC_{ki} < L_i \\ 0, & otherwise \end{cases}$$

$$\sum_{k=1}^{K} FC_{ki} \leq \mathrm{M}^* Z_i$$

where, M is a large number. If Z_i equals 1, then the patient has ILD and else otherwise.

We experimented with different metrics to assess filter performance: (1) the ratio of number of predicted ILD patients to the sum of correctly predicted ILD and incorrectly predicted no-ILD patients; (2) the ratio of the number of predicted ILD patients to the true number of ILD patients (i.e., the True Positive Rate); (3) The product of (1) and (2) (which penalizes filters with good ILD prediction but poor no-ILD prediction); and (4) The fraction of total misclassified patients (the Overall Misclassification Rate, OMR) that equally weights both forms of misclassification. We decided on a two-level metric (OMR with TPR used break ties if necessary) because we had discovered that in some situations, best filters were not unique (Examples could be found in §D.6).

Standardized threshold deviates were computed using the FC expression ranges from the IRIS data. Thresholds were randomly generated using a uniform generator (R Core Team, 2013) which is an efficient way to explore a large unknown parameter space due to simplicity of coding. The design process is computationally challenging in that we have *N* FC variables with which to construct filters, but no a priori knowledge of how many variables and which variables should be included in any particular filter. A conservative but computationally expensive approach would involve full combinatorial expansion, that is, we would construct filters comprised of $\binom{N}{1}$, $\binom{N}{2}$, … , $\binom{N}{N}$ FC variables, which represents a very large number with increasing *N.* Adding to the computational challenge, stochastic simulation is performed many times for each filter realization. Being completely random in nature (i.e., we have no biological or other guidance to suggest which variable subsets and respective thresholds are likely to perform well) it follows that the vast majority of filters we create will perform poorly in a validity test (their Overall Misclassification Rate will be high in a validity test).

C.6 Stepwise GLM

Having identified a group of FC variables, the data dimension reduced and therefore the 'Large P small N' issue no longer held. Parametric method especially Stepwise General Linear Model analysis was used to draw statistical inferences and also further reduce the dimensionality of the data. After a full model is fitted, a stepwise algorithm in both directions (forward and backward) was applied to find the "best" model using the Akaike Information Criterion (AIC) (Akaike, 1981):

$$AIC = -2LL + 2p$$

where, LL is the maximum log likelihoods and p is the number of parameters. AIC was chosen for the reason of obtaining a model with good fit but as smaller number of parameters as possible.

Starting with a full model (many parameters) a backward algorithm is first used - remove one variable at a time (i.e., make the model smaller) to determine whether AIC decreases. When AIC no longer decreases by a prescribed amount, we then declare the current model as best. Forward algorithm is the opposite. Start with a small model, say only the intercept term is involved, then one variable at a time is added to see how AIC changes.

The stepwise algorithm (Hastie et al., 2009; Ramsey & Schafer, 2012) combines forward and backward algorithms. No restrictions are made regarding which direction to move. Adding a variable or removing a variable are acceptable as long as AIC decreases by a prescribed (threshold) amount.

C.6-1 Drop-in-Deviance test

Consider two models: one with more parameters (Full model, denoted F) and the other with fewer parameters (Reduced model, denoted R). By defining Deviance as the following:

$$D = C - 2LL$$

where, D means Deviance, C is constant and LL is the maximum log-liklihood,

Deviance_F will in general be less than Deviance_R because a model with more parameters will tend to fit the data better. Deviance approximately follows a Chi-square distribution with degree of freedom equal to the difference of the number of parameters between the two models (Nelder & Wedderbu, 1972). The conventional goodness-of-fit Test is a special case of Drop-in-Deviance in which the full model is “saturated” with n parameters (n being the sample size). An alternative way of showing

goodness-of-fit of the model is to plot an in-sample ROC curve and calculate the Area Under Curves (AUC). These ROC curves will be in-sample because the stepwise GLM model will not be used for making predictions but drawing statistical inferences. Therefore, they are different from those out-of-sample ROC curves.

We therefore have adapted and integrated three approaches, including CRF, GSEA and stepwise GLM, to define a group of statistically significant variables (therefore reducing the dimensionality of the data set). Recall that the output of CRF – the Variable Importance List, is used as input to GSEA, which in turn yields the best FC set associated with highest ES. This set of FC variables serve as the input covariates for the stepwise GLM of which the fitted coefficients are instrumental in interpreting the underlying biological significance of different subsets of T cells in the occurrence and development of SSc-ILD.

In the next chapter, results associated with the above-mentioned methodology will be shown.

D. RESULTS

In this chapter, the PCA results will be presented first (§D.1) which shows that an inability to interpret the created principal components is an issue. Therefore, these PCA results were not further utilized. In Section §D.2, performance indices, such as predictive accuracy and goodness-of-fit, of the four data mining models were compared. The variable importance measures were extracted from the best model and became input to the Gene Set Enrichment Analysis (GSEA) of which the results are presented in Section §D.3. Results regarding robustness of the GSEA algorithm are presented in Section §D.4. Based on the FC sets identified in Section §D.3, on one hand, randomized design filters were constructed to differentiate SSc patients with ILD from those without. Details with respect to refining the filters in order to reduce misclassification rates are shown in Section §D.5, and validation results of the best filters found are presented in §D.6. On the other hand, stepwise GLM and partial dependence plots (PDP) analysis using CRF are used to draw statistical inference from the identified FC set, whose results are included in Sections §D.7 and §D.8.

Through Sections §D.2 to §D.8, the phenotype of interest was ILD. In Section §D.9, the same procedures (CRF-GSEA-Stepwise GLM/PDP) were

applied to another phenotype ‘Cancer’. In order to examine inter-relationship between phenotype ILD and Cancer, results of GSEA inter-comparison of these two are presented in Section §D.10.

Next, three different methods of determining FC sets were evaluated including, biological information (§D.11), Student’s t-test statistics (§D.12) and p-values from point-wise GLM (§D.13).

This chapter closes in Section §D.14 with results of CRF-GSEA-stepwise GLM analysis using other phenotypes such as Anti-topoisomerase I antibodies (ATA, or anti-Scl-70 antibodies) and anti-centromere antibodies (ACA).

D.1 Predictor variable correlation - PCA

Predictor variable correlation exists due in part to the hierarchical relations among FC variables in each panel of the data set. Without consideration of input variable correlation, results can be adversely affected. Consider for example the estimated p-value of the GSEA test (a developed statistical test described in section §C.4 *Gene Set Enrichment Analysis*). Variable correlation has the effect of making the p-value smaller than what it really is because highly correlated variables tend to cluster towards certain

area in a ranked list, thus producing a nonconservative result. A traditional method for dealing with colinearity between variables is to use Principal Components Analysis (PCA). Application of PCA to the IRIS data showed that the first 10 principal components (PC) explained 57% of the total variance and more than 30 components are needed to explain 90% of the total variance of the 112 variables. Note that these results are based on the FC variable correlation matrix as opposed to the original FC expressions to avoid scaling inconsistencies.

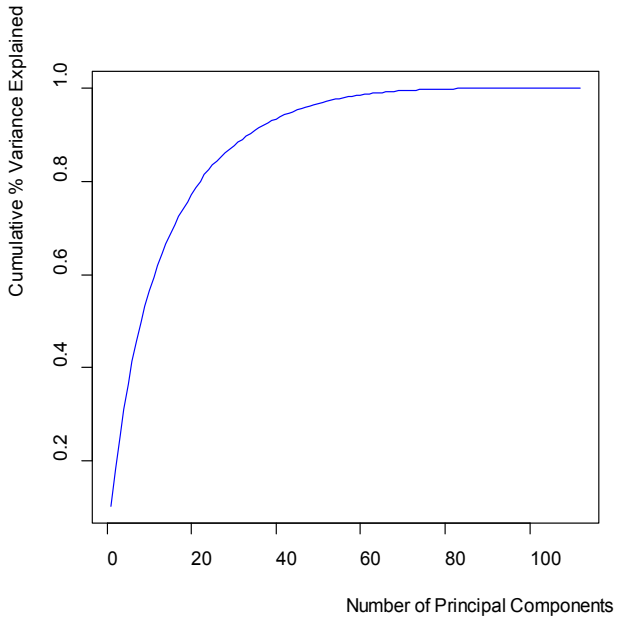


Figure D-1. Cumulative Variance Explained for Different Number of PC

Table D-1 Cumulative Variance Explained for Selected PC Number

PC number	10	22	33	71
Cumulative Variance Explained	57%	80%	90%	100%

This suggests that PCA can reduce the dimension of the data set, but the degree of reduction did not outperform the hybrid CRF-GSEA approach because the number of principal components are larger than the number of variables in the best FC set(results are shown in §D.3). Moreover, the principal component variables (weighted FC expressions) are difficult to interpret clinically or biologically, in that one PC is a linear combination of more than 100 FC variables of which each is associated with different loadings. The loadings of each PC can be found in Appendix I.4. In that PCA was not useful in terms of interpretability, no further analysis concerning PCA were performed and other unsupervised machine learning methods such as factor analysis were not attempted as well.

In the next section, I will present results regarding model performance of the data mining methods evaluated in this work.

D.2 Data Mining Model Performance

Five classification methods were tested (Classification and Regression Trees (CART), Random Forests (RF), Conditional Random Forests (CRF), Support Vector Machines (SVM) and a mean-only model) using 112 FC expressions as predictor variables and ILD as response. The mean-only model simply uses the mean value of the response in the training data as

future prediction. Mean Absolute Errors (MAEs) and Mean Squared Errors (MSEs) are the two measurements of the comparison. They were calculated based on the following formula:

$$MAE = \frac{1}{n}\sum_{i}^{n}\left|\hat{y}_i - y_i\right|$$

$$MSE = \frac{1}{n}\sum_{i}^{n}\left(\hat{y}_i - y_i\right)^2$$

where, n = 50, $\hat{y}_1$ is the estimate of i^{th} response variable, and y_i is the actual value of the i^{th} response variable. Table D-2 shows the mean MAE and mean MSE of the five approaches.

Table D-2 Mean MAE & Mean MSE of 50 Times Holdout Analysis for Data Mining Approaches

	mean MAE	mean MSE
CART	0.471	0.356
RF	0.482	0.246
CRF	0.492	0.248
SVM	0.502	0.253
mean	0.502	0.254

When comparing predictive accuracy between two models, a one-sided two samples Student's t-tests was performed to examine whether the mean of the two corresponding vectors of MAEs or MSEs are equivalent. RF and CRF perform best (by a small amount) but their differences in MAE and especially MSE are not statistically significant (p-value = 0.0996 for

MAE and p-value = 0.785 for MSE). The mean-only result confirmed our understanding that this statistical estimation problem is very flat (i.e., no single FC variable or small subset of variables is highly associated with ILD status). From their Receiver Operating Characteristic (ROC) curves performance (Figure D-2a) RF, SVM and CRF emerged as the most effective classifiers. All consistently yielded AUC (Area Under Curve) values of greater than 0.95. Thus, RF, SVM and CRF fit the data well and exhibited high true positive rates (out of the positives) and few false positives (out of the true negatives) , but this is training performance. As shown in Figure D-2b, all four models were fairly poor out-of-sample (i.e., out-of-bag, OOB) classifiers based on the Leave-One-Out Cross Validation (LOOCV). The OOB AUCs for all models evaluated are between 0.5 and 0.6. The highest AUC is associated with RF (0.57) while the lowest associated with SVM (0.53). This is the reason why we created screening tools via stochastic simulation.

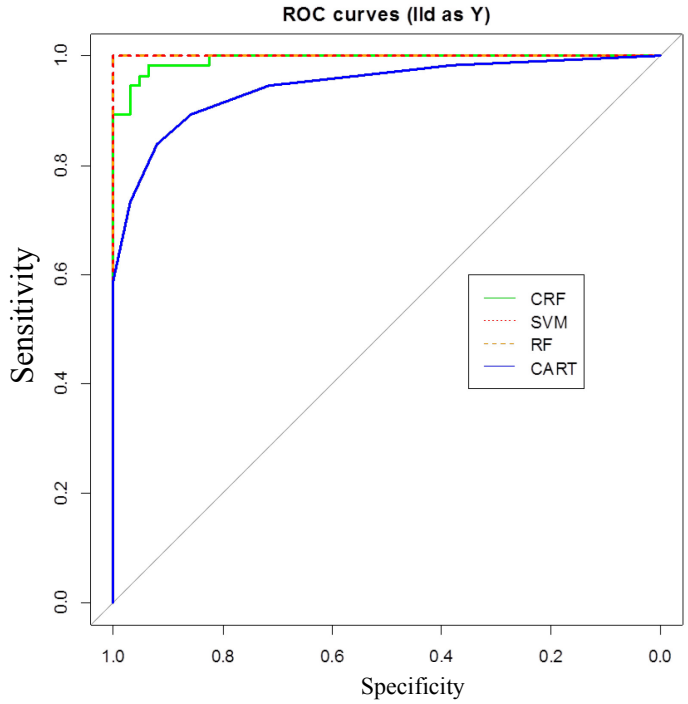


Figure D-2a. Goodness-of-Fit ROC Curves for Various Data Mining Methods (Phenotype is ILD)

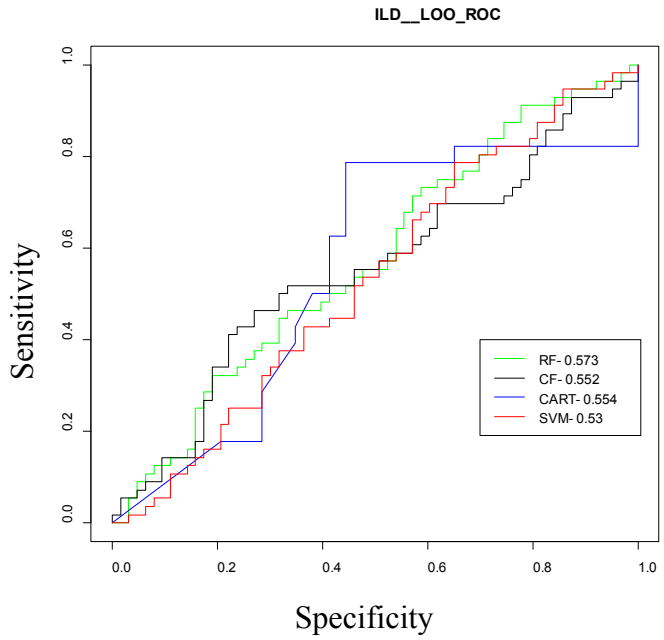


Figure D-2b. Leave-One-Out Cross Validation (LOOCV) ROC Curves for Various Data Mining Methods (Phenotype is ILD)

Conditional Random Forest (CRF) was eventually chosen over RF and SVM for several reasons. First, the permutation computing scheme for variable importance measures (VIMs) in CRF provides a "*more fair means of comparison that can help identify the truly relevant predictor variables*" (Strobl, 2008; Strobl, Hothorn, & Zeileis, 2009). Specifically, it enforces the requirement that each predictor variable that is selected as a split variable in each tree must be strongly associated with response variables (through hypotheses testing under an unbiased conditional inference framework) (Hothorn, Hornik, & Zeileis, 2006). This is a robust way of enhancing the discriminant power of a predictor variable and is particularly useful for our IRIS data set in which a significant number of FC expressions are highly correlated. In contrast, the VIMs of RF were unstable and suffered from "*correlation bias*" due to effects related to predictor correlation, including: (1) VIMs were not necessarily aligned with the discriminant power of the stimulus variable; (2) the size of the group of correlated variables has a pronounced effect (Gregorutti, B Michel, & Saint-Pierre, 2013; Toloşi & Lengauer, 2011); and, (3) VIMs did not "*directly reflect the coefficients in the generating model*" (Nicodemus, Malley, C Strobl, & Ziegler, 2010). Using the conditional importance measure, Strobl et al. (2008) showed that VIMs based on the conditional permutation scheme better reflect the pattern

of the coefficients associated with predictor discriminant power and the variability was "*lower than that of the unconditional importance within each level of mtry*" (mtry is the parameter in R specifying the number of covariates randomly selected to split the node in each tree of the RF model). In this research, it was discovered that the conditional permutation scheme exerted almost no influence on the VIM output. Also, the computational burden of executing the conditional permutation scheme was particularly expensive in that the computing time increased exponentially as the number of observations increased. Eventually, VIM information was extracted from CRF without conditional permutation scheme. Second, the ROC curves of the fitted CRF, RF and SVM models suggested that RF and SVM might be overfitting. CRF misclassified 5 patients out of 79 whereas RF and SVM had 100% predictive accuracy. Third, the ES based on the variable importance list drawn from CRF were always larger than those of RF regardless of configuration settings, including the number of trees and mtry. The results are shown in Table D-3.

Table D-3 Enrichment Scores of RF vs. CRF

mtry	ntree	RF	CRF
5	1000	21.77	23.43
11	1000	20.76	22.34

To the best of the author's knowledge, no existing predictive methods could yield well predictive accuracy when the response variable is ILD. However, as to the randomized filter design, promising results were obtained in terms of predictive accuracy of this screening tool whose validation results will be presented later in Section § D.6.

Recall that the Gene Set Enrichment Analysis (GSEA) was designed and adapted in this study to capture combined effects of a group of FC variables. In the following sections, performance of GSEA (in §D.3) and robustness of this algorithm (in §D.4) will be demonstrated.

D.3 GSEA Performance

At this point in the analyses the IRIS data set changed and consisted of the original set of patients in IRIS071813, modified by: (1) removing the four dropped FC variables; (2) updating against IRIS030814; (3) adding missing fvcstpp/ILD data. From 112 FC variables, we found via the CRF-GSEA algorithm 27 FC variables to be the most important in differentiating SSc patients with ILD from those without. Those variables are shown in Table D-4.

Table D-4 Most Important Variables Identify via CRF-GSEA algorithm (Phenotype is ILD)

pol8ccr5	pol8ccr5cxcr3neg	mememm4	pol8ccr5cxcr3	act4103
act425lo	mememm478	act8103	mememm8	mememra87
act425103	traff4cxcr6	memnaive4	act425tot	memcm878
act410371	pol4ccr6	mememm878	mememra478	memcm4
memcm478	traff8cxcr4	mememra4	act425hladr	memcm8
act4103hladr	mememm48			

Below is the random walk for FC27. Figure D-3 shows that there is a peak between the 1st ranked FC variable and the 40th FC variable, and a valley when the ranking of FC variable was around 90. The former, in genomics research, is called up-regulated and the latter down-regulated. It means increase or decrease in gene or FC expression.

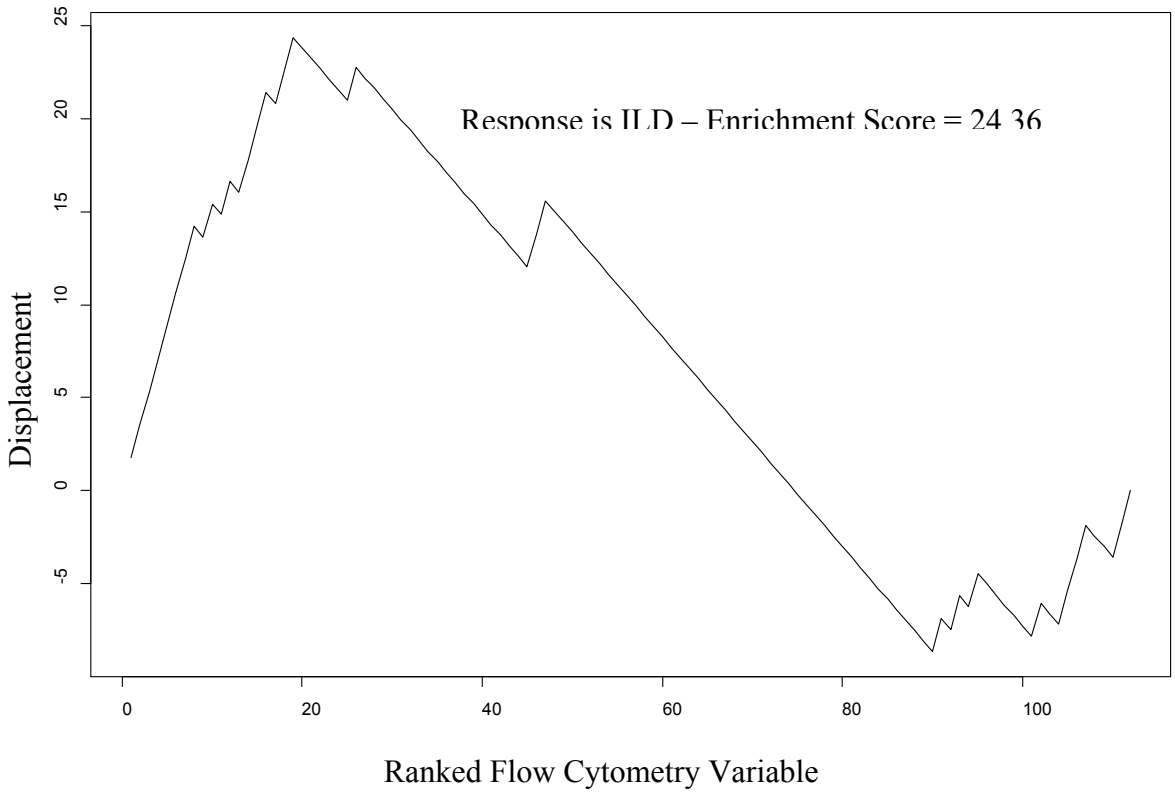


Figure D-3. Random Walk that Results from FC Set Comprised by Top 27 Most Important Variables

Figure D-3 shows that the highest ES is associated with FC set size 27. Figure D-4 is the full set of GSEA results that identified the best performing set as FC27. The corresponding ES statistical significance levels are presented in Figure D-5. The estimated p-values were calculated based on permutation test described in §C4-2. It can be seen in Figure D-5 that the estimated p-values of the FC set were smaller than 0.01 when the FC set size is between 15 and 40.

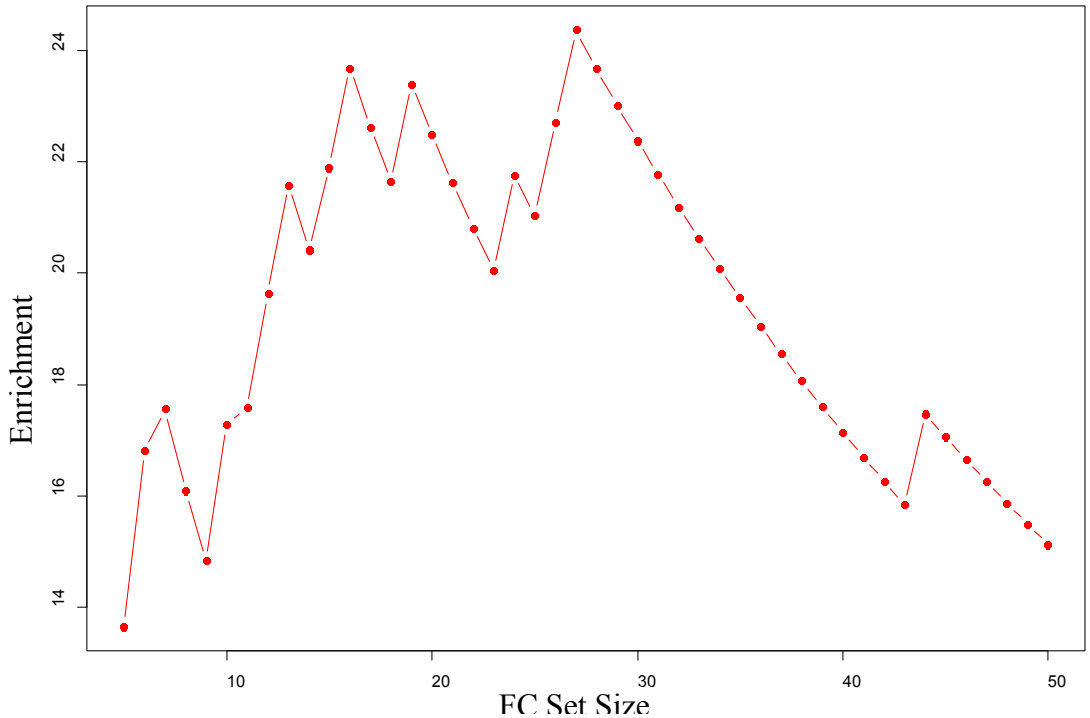


Figure D-4. Enrichment Scores of GSEA for Different FC Set Sizes (Phenotype is ILD)

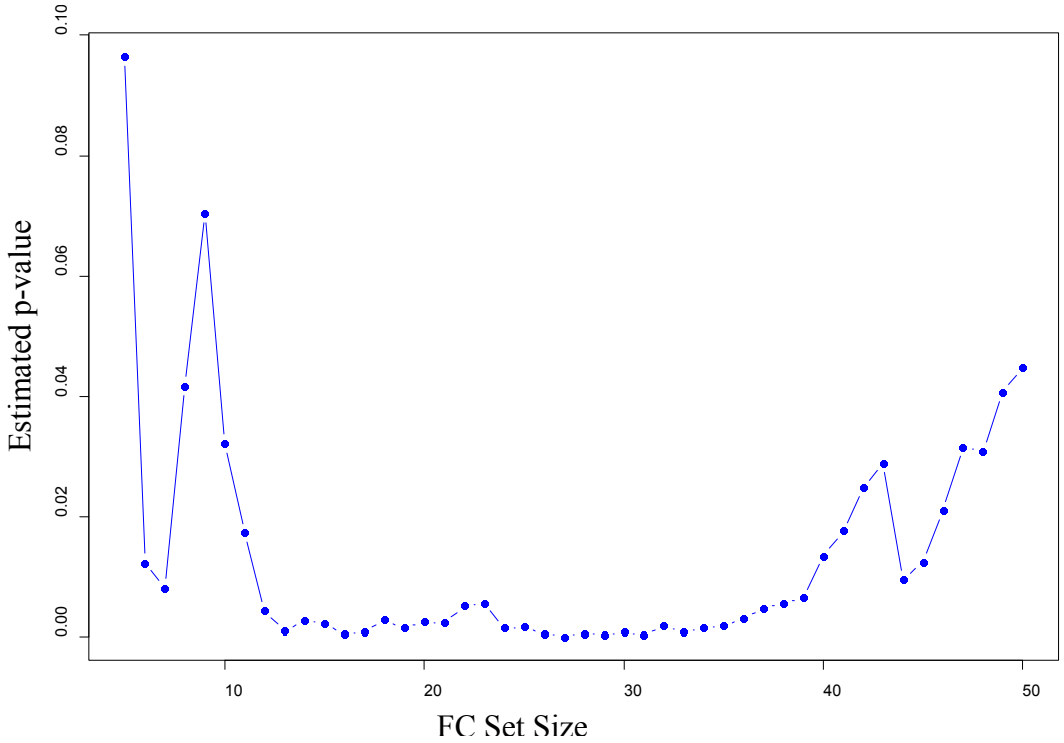


Figure D-5. P-values associated with Enrichment Scores for Different FC Set Sizes

The results in this section indicate that the GSEA algorithm can identify the best set size of a group of FC variables based on the Enrichment Scores (ES). When the response variable was ILD, the best FC set size was 27 because its associated ES was the largest. This FC set of 27 variables is also considered as statistically significant based on the permutation test. In Section §D.4, I will present results with respect to robustness of the GSEA algorithm.

D.4 Robustness of the GSEA algorithm

D.4-1 Random Walk with differing FC set sizes

All FC sets created to date (including the best performing – FC27) were based on the top-ranked variables in the CRF variable importance (VI) list. To examine the validity and robustness of this procedure, random walks were plotted using different FC set selection criteria. Shown in Figure D-6 is the random walk that results from an FC set comprised of the top 5 FC expressions in the VI list.

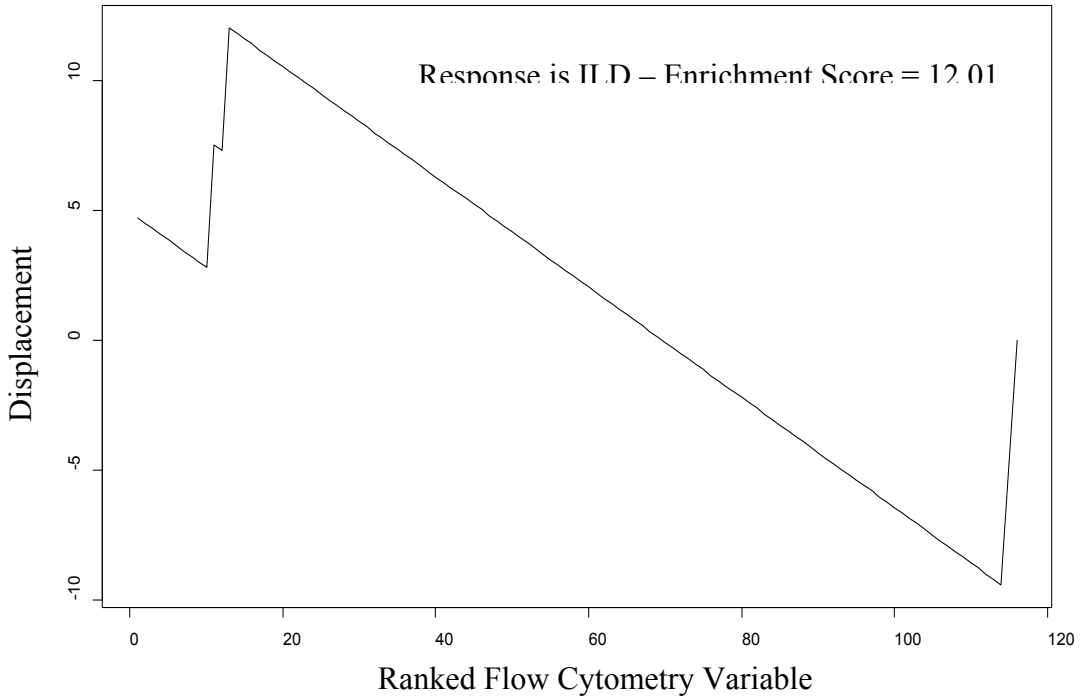


Figure D-6. Random Walk with Top 5 Most Important Variables

In Figure D-6, it shows that the random walk was enriched in both ends (displayed as a peak and a valley) but have a relatively low ES if the FC set contains the top 5 most important variables. In addition, the maximum deviation from zero, or the supremum, is positive value, which means that this random walk is up-regulated. Note that enrichment score is the absolute value of the maximum deviation from zero and therefore it is always positive. When the FC set size was increased to 10, similar enrichment structure of the random walk showed up and ES increases substantially (see Figure D-7).

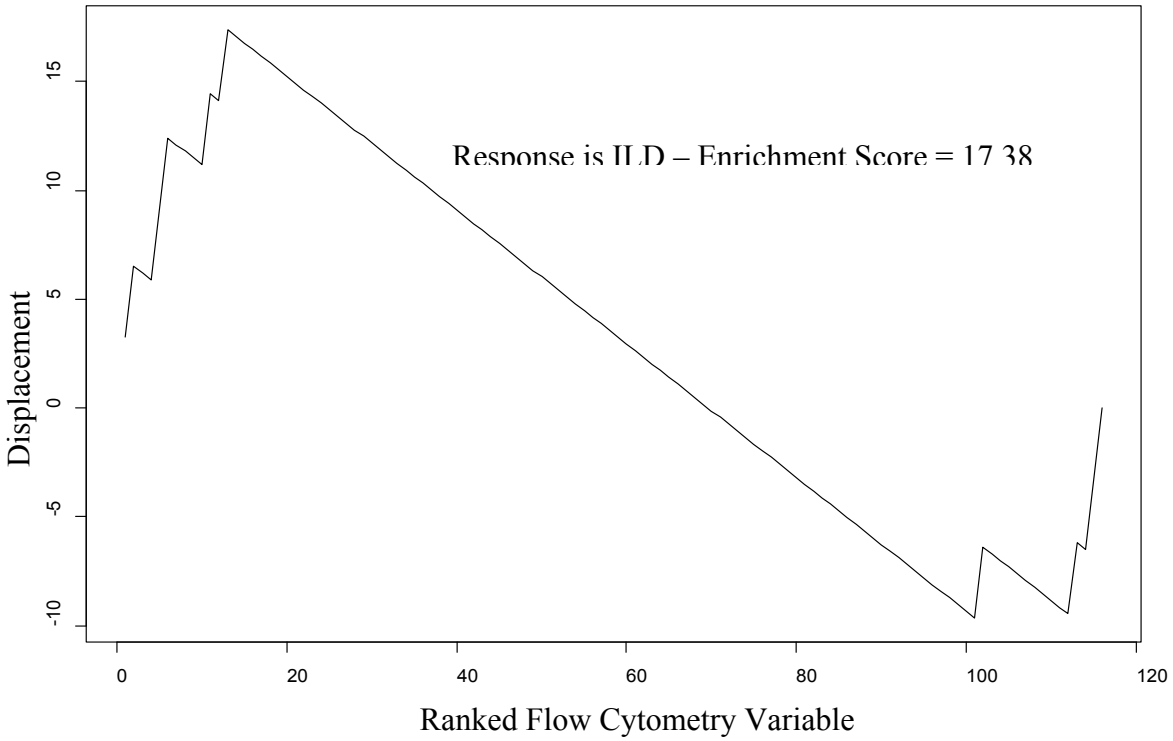


Figure D-7. Random Walk with Top 10 Most Important Variables

Keep increasing FC set size in this situation will yields similar enrichment structure of a random. Figure D-8 shows the random walk with FC set size 50 and the random walk is up-regulated as well. The Enrichment Score did not increase substantially in that adding more FC variables to the FC set does not necessarily contribute the increase of the maximum deviation from zero.

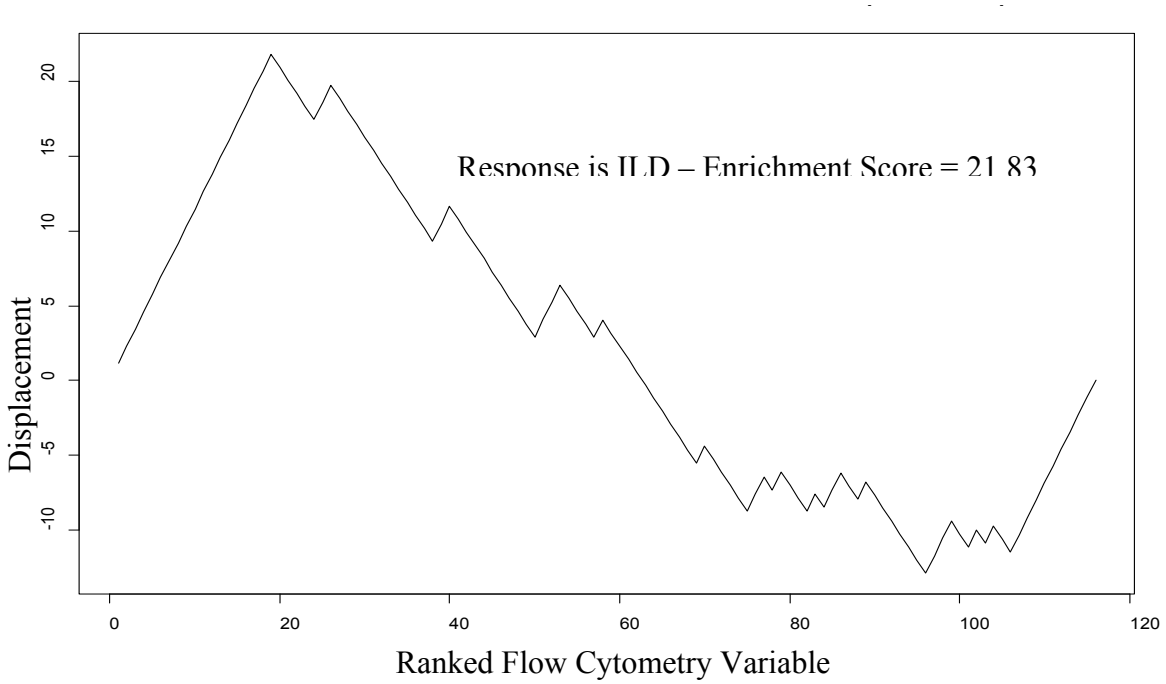


Figure D-8. Random Walk with Top 50 Most Important Variables

So far, the FC set is all determined by choosing the FC variables from the top of the variable importance measure (VIM) list. Next, the GSEA performance was examined when the FC set was determined by choosing certain sequence of the variable importance list as opposed to choosing variables from the top. For example, a FC set can be determined by choosing from the 20^{th} to the 30^{th} variables in the VIM list. Shown in Figure D-9 is the random walk that resulted from an FC set comprised of the 20^{th} to 30^{th} variables in the VIM list. It can be seen that the enrichment structure switch to down-regulation. In other words, the associated maximum deviation becomes negative and its absolute value, or the ES, decreased.

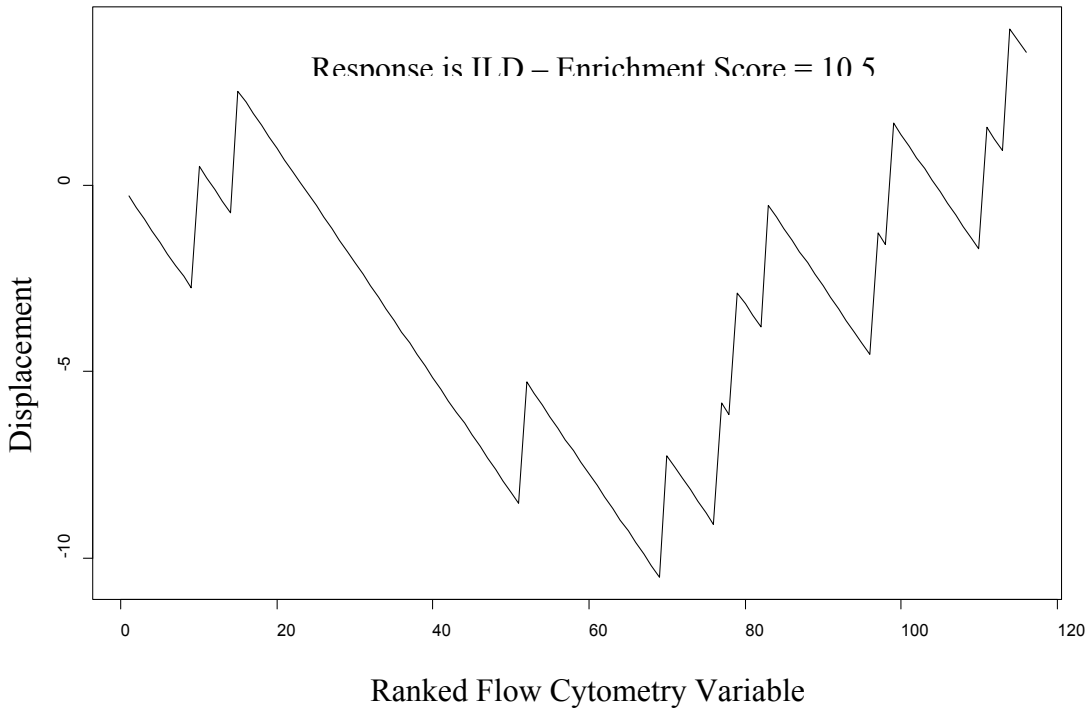


Figure D-9. Random Walk with Top 20th to 30th Most Important Variables

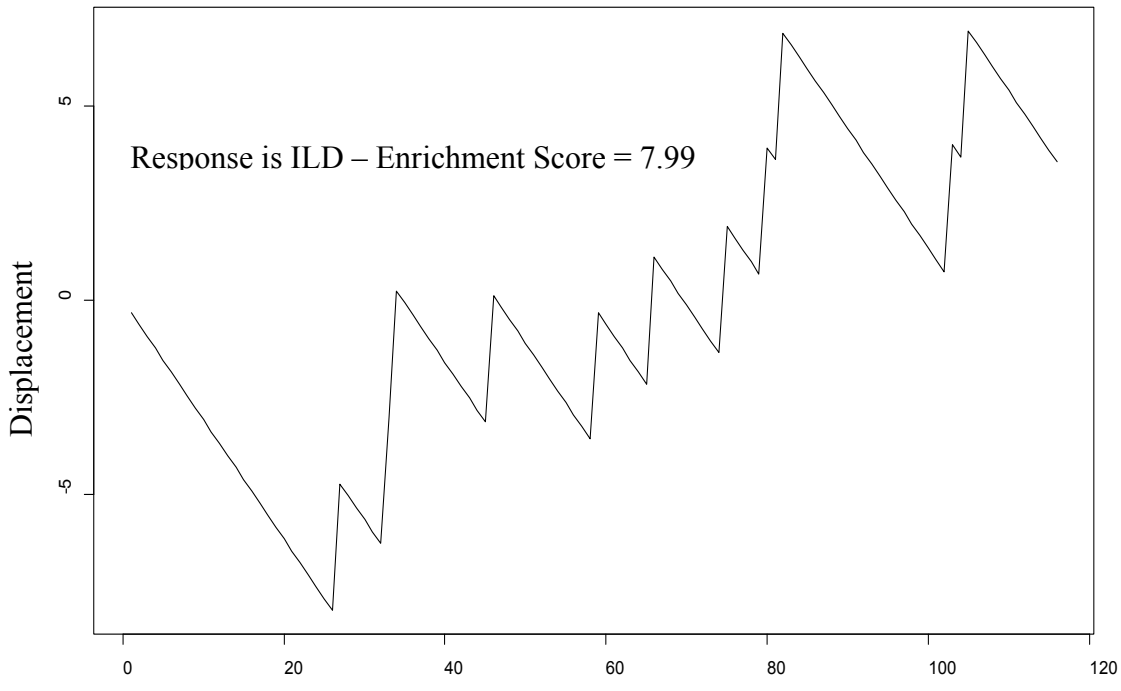


Figure D-10. Random Walk with Bottom 10 Most Important Variables

Figure D-10 shows the random walk using an FC set comprised of the bottom 10 variables of the VI list which are considered by Conditional Random Forest the least important variables. Almost no enrichment structure can be found in this setting, and the ES is less than 8.

In sum, the GSEA algorithm is robust when the size of the FC set differs. In specific, the random walk is always up-regulated with large enrichment score if the FC set is determined by choosing FC variables from the top of the variable importance list. In the next sub-section, I will examine robustness of the GSEA permutation test.

D.4-2 Robustness of the GSEA Permutation Test

Results in this section were based on data set IRIS071813. When the number of permutations varied ranging from 1000 to 10,000,000, the estimated p-value of the GSEA test changed only slightly (between 0.019 and 0.0262) - all significant at level 0.05. This implies that the GSEA significance level is robust to permutation number. Considering both reliability and computational efficiency, 10,000 was chosen as the permutation number for GSEA.

For the next several sections (from §D.5 to §D.8), I will present results regarding the two directions analysis mentioned in §C.1.

D.5 Refinements in Filter Design

In this section, I describe two methods of refining the filters design procedure - – threshold tightening and pre-partitioned FC classification, so as to reduce overall misclassification rate (OMR). The corresponding improvements of the filters performance associated with each method in terms of OMR will also be presented. Because of the superior improvements by using the second method, details of the associated best filters regarding filter components and associated thresholds are also explored and shown later in this section. Results in this section were based on data set IRIS071813.

The first method refers to a procedure in which the ranges of random thresholds for each variable in the FC set were narrowed through an examination of well performing filters. Consider the case of 10,000 random filter trials, from which there emerged 200 filters for which the smallest Overall Misclassification Rate, or OMR, was about 0.24. It means that if there are 100 patients then we have 24 patients misclassified by the random filter. The random threshold ranges for each FC expression was updated for

subsequent analyses using the FC thresholds that produced these 200 filters. Threshold tightening alone improved filter performance OMR to about 0.16. Improved OMR results with threshold refinement are due to our randomization procedure wasting less time evaluating bad-performing filters.

The second approach, Pre-partitioning using CART (with pruning at the 2nd level split producing three groups), was used to classify the data set into *N* subgroups with *N* being the number of end branches (Figure D-12). Random filter design was then applied to each subgroup. This method substantially improved filter performance. At the time, the lowest global OMR found considering all groups was 0.09 (see Table D-5).

Table D-5 Classification Statistics of Three Prepartioned Groups

Group	Feature	SSc	# of NO ILD	# of ILD	Best OMR	Misclassified ILD	Misclassified NO ILD
A	traff4ccr10 < 1.265	11	1	10	0	0	0
B	traff4ccr10 ⩾ 1.265 & mememra478 < 0.595	58	43	15	0.13	3	5
C	traff4ccr10 ⩾ 1.265 & mememra478 ⩾ 0.595	31	12	19	0.03	1	0

OVERALL OMR = (3+5+1)/100 = 0.09

In one of the more recent experiments using all CART-partitioned groups (10) the filter-based approach successfully classified 124 of 125 SSc patients (OMR = 0.008). This extremely low OMR, however, needs a strong caveat: it may be a result of overfitting (Cawley & Talbot, 2010; Forsyth et al., 1994) that is, decreases in training data set error are accompanied by increases in validation data set error as the extent of pre-partitioning increases. Shown below in Figure D-11 are four pruning levels (the blue lines) for CRF-informed analyses involving 125 SSc patients.

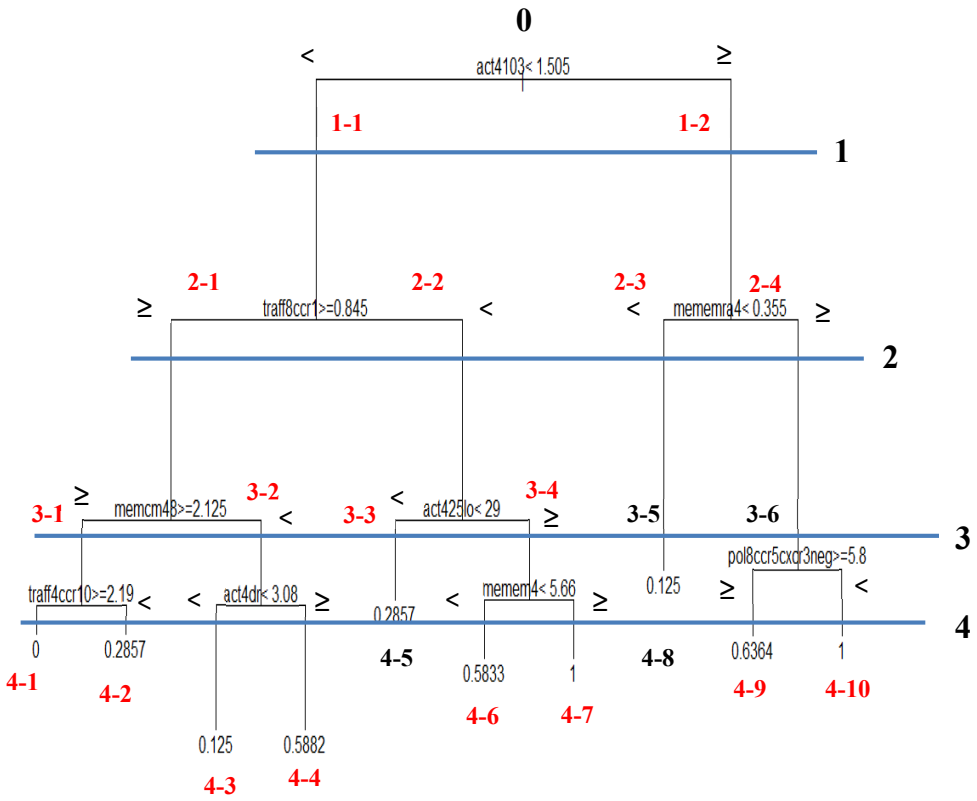


Figure D-11. Four Pruning Levels of CART Pre-partitioning

The randomly generated threshold bounds for each FC variable were derived from the training set data. We then addressed the issue of which subsets of variables (from the set of 27 "best" variables identified through GSEA) should be used to construct filters. Experiments were conducted to see whether full combinatorial expansion was necessary, that is, did we need to evaluate filters with $\binom{27}{1}$ through $\binom{27}{27}$ components (134,217,727 possible combinations of variables; recall also that for each combination, many random threshold realizations are generated). Our testing showed that the best performing filters (those with the lowest OMR) were consistently comprised of at least three and no more than six FC variables, therefore all subsequent analyses involved filters comprised of $\binom{27}{3}$ through $\binom{27}{6}$ (i.e., 2,925 + 17,550 + 80,730 + 296,010 = 397,215) combinations of FC variables.

The best performing filter at this stage of the analyses had an overall misclassification rate of 18.98% (15 patients misclassified out of 79). We next performed pre-partitioning using CART to classify the data set into subgroups (Figure D-11). Random filter design was then performed for each subgroup (there are 14 shown in total in Figure D-11, but six are child nodes whose parents have OMR = 0; these six node ID's are shown in black). This

method substantially improved filter performance. The pre-partitioned OMR results are shown in Table D-6 (MC = Misclassified).

Table D-6 OMR Results for All Pre-partitioned Levels

CART Node	# ILD	# NoILD	# MC	# ILD MC	# NoILD MC	OMR
Level 0	38	41	15	11	4	0.1878
1-1	24	38	11	9	2	0.1774
1-2	14	3	0	0	0	0.0
Level 1	38	41	11	9	2	0.1392
2-1	17	37	9	8	1	0.1667
2-2	7	1	0	0	0	0.0
Level 2	38	41	9	8	1	0.1139
3-1	1	18	1	0	1	0.0526
3-2	16	19	5	2	3	0.1429
Level 3	38	41	6	2	4	0.0759
4-2	3	13	1	0	1	0.0625
4-3	13	6	0	0	0	0.0
Level 4	38	41	1	1	1	0.0127
Parent Node > Child Node; 1-2 > 2-3 > 3-4 > 4-5; 2-2 > 3-3 > 4-4; 3-1 > 4-1						

Pre-partitioning had a large effect on OMR performance, reinforcing our expectation that CART is a highly effective classifier and that sub-groups of patients identified through CART are, in a sense, relatively easier to correctly classify. But we are mindful not to overstate this result, for the sub-groups with the highest OMR values also had the most patients. To continue this line of thought, consider an extreme situation: 79 one-patient "groups" with a best filter created for each. The resulting training filter's OMR performance will be perfect but operationally useless, both in validation and clinically. Excepting the fourth level (where only one patient

of 79 was misclassified) there was considerable more misclassification of ILD than no-ILD patients. Moreover, the result that only one patient was misclassified at the fourth level of pre-partitioning is strongly suggestive of overfitting (which is addressed below in §D-6 Validation).

Following in Table D-7 are details of the best training filters for no pre-partitioning (Level 0) and all lower levels. The FC variable names are shown as are the corresponding standardized random threshold deviates.

Table D-7 Details of FC Variables in Different Prepartioned Levels

Variable/Node	0	1-1	1-2	2-1	2-2	3-1	3-2	4-2	4-3
act4103	0.57				-1.58				
act425lo		2.10	-1.60		-1.63		1.82		
act425tot	2.29		0.07		-0.24			2.47	
act8103									1.17
act810371	1.43	2.44	0.68	2.13			1.28		2.09
memem4	1.05	0.98			-0.59				
memem478					-0.41	1.85			-1.07
memem48			-0.27			2.34	1.31	1.81	0.70
memcm478									-0.26
memcm4k						3.12			
memem8				-1.26					
memem878	-1.23	-0.92							
mememra4								1.88	
mememra4k						2.82			
mememra478								3.46	
pol8th17			1.84				-0.68	1.26	
pol8th1th2ratio				1.64		1.67			
pol4ccr6		-1.33		-1.36					
traff4ccr3	1.88			2.31			0.61		

Of the full set of 27 “best FC set” variables, 19 appeared in the highest performing training filters. Activation and polarization variables

have the highest representation. Only three variables, act425tot, act810371 and mememm4 appeared in all Levels (0-4). Pre-partitioning had very pronounced effects on active variables within and across levels. In Level 2 for example, node 2-2 FC thresholds are all negative (cutoff thresholds for those variables less than their means). For all levels, active variable sets for all nodes are typically very different.

D.6 Validation of Randomized Filter Design

Having achieved in-sample success in filter design and calibration, the next critical step is validation. FC profiles of SSc patients whose FC expressions were not involved in filter design (i.e., out of sample analysis) will be processed using our best performing filters. In this section, results two protocols (A and B) of validation test will be presented. The main difference of these two validation protocols lies in the way the training data was generated of which the details will be described below.

D.6-1 Validation Test A

D.6-1-1 Test A Protocol

The training data set consists of the original set of patients in IRIS071813, modified by: (1) removing the four dropped FC variables (2)

updating against IRIS030814. The resulting set contains data on 129 patients (61 ILD; 68 NO_ILD) and 20 "new" validation patients were provided (6 ILD; 14 NO_ILD), namely IRIS041314. There are fewer patients to work with as FC set size increases due to missing data.

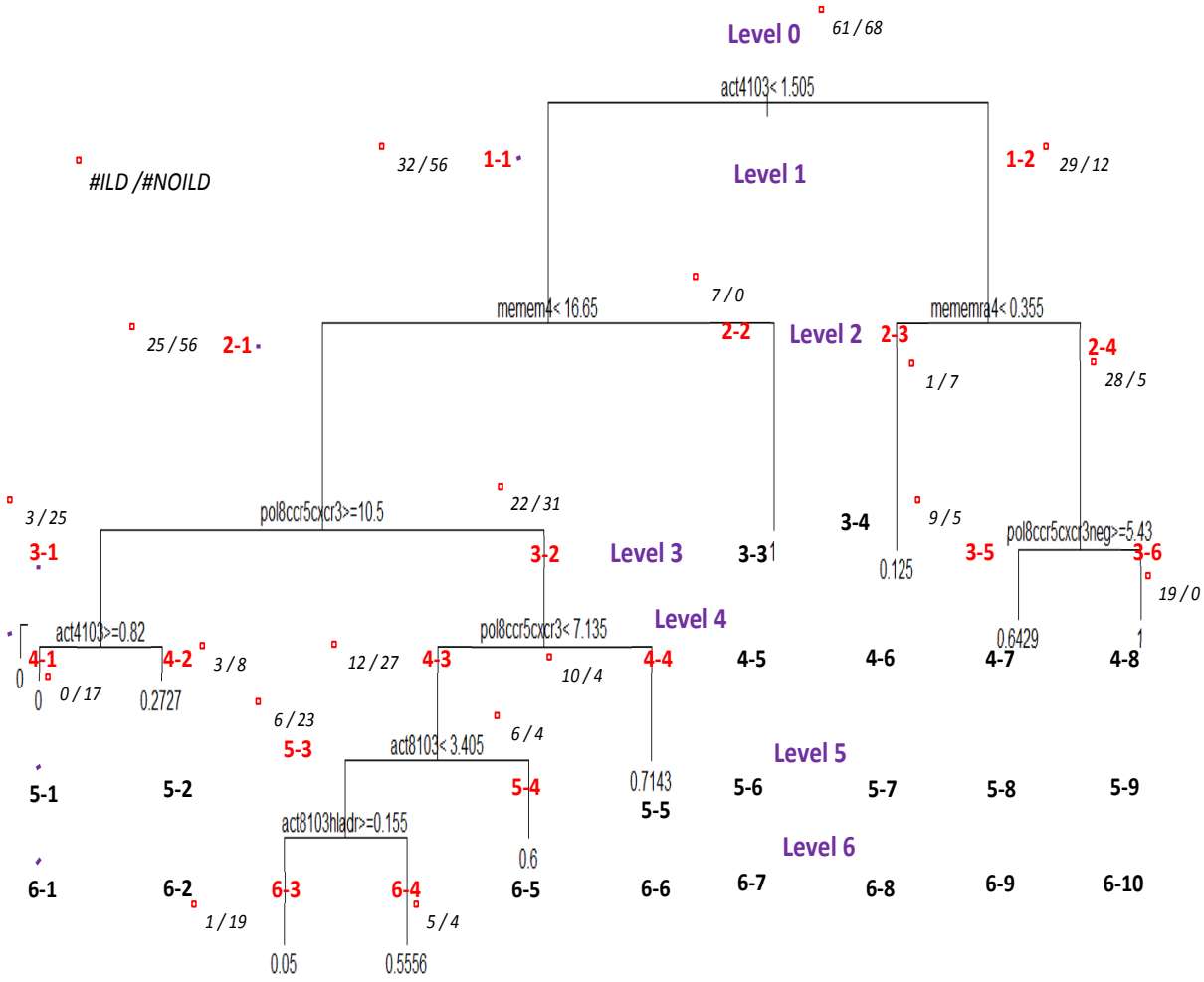


Figure D-12. Six Pruning Levels of CART Pre-partitioning (Protocol A)

Through pre-partitioning, it is possible to reduce training set error (Overall Misclassification Rate) but likely at the expense of increasing validation set error (i.e., overfitting) visualized as:

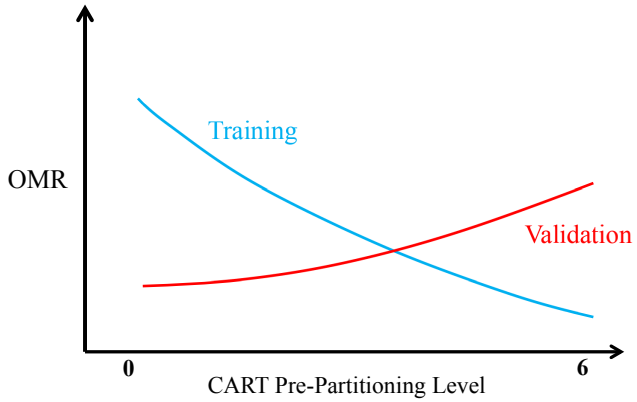


Figure D-13. Tradeoff between OMR and CART Pre-partitioning Level

D.6-1-2 Validation Test A Results

With no CART pre-partitioning (Level 0) the screening tool correctly classified the ILD status of 80% of the validation patients. In specific, 4 out of 6 ILD and 12 out of 14 NO ILD were correctly predicted. Level 1 results are the same as Level 0. At CART pre-partitioning Level 2 the screening tool correctly classified the ILD status of 90% of the validation patients. Details of the results are shown in Table D-8. The combined Level 2 OMR is 0.1.

Table D-8 Classification Statistics for Level 2 (Protocol A)

FC level	FC21_2-1	FC21_2-2	FC21_2-3
actual ILD	3	3	0
correctly predicted ILD	2	3	0
actual NO ILD	10	3	1
correctly predicted NO_ILD	10	2	1
OMR	0.077	0.167	0

D.6-2 Validation Test B

D.6-2-1 Test B Protocol

A different protocol was developed to expose our methodology to a potentially more demanding validation test. It involves the following procedures: (1) Combining original and "new" patients into one data set and thoroughly shuffling them; (2) Randomly selecting a training set and a validation set of given proportions (66.7%-33.3%). These proportions derive from published guidance and testing of our data following Dobbin & Simon (2011); (3) 120 random selections were performed involving 38,750,400,000 screening tool design and testing stochastic simulations.

There are 79 patients (38 ILD and 41 NO ILD) in the training data set and 40 patients (18 ILD and 22 NO ILD) in the validation data set.

D.6-2-2 Validation Test B Results

The best training filters were validated using FC data from 40 patients not used in training. Without pre-partitioning, the overall correct ILD classification rate was 82.5 % (7 patients misclassified out of 40). Pre-partitioning the validation patients (using the CART-derived variables and splitting levels developed for the training data) increases correct validation classification to 95% after two levels of pruning (2 patients misclassified out

of 40). This indicates that overfitting was occurring for the deepest pre-partitioning level (the training and validation curves cross) as figure D-14 shows. Notable is the similar OMR performance between training and validation for no pre-partitioning (Level 0).

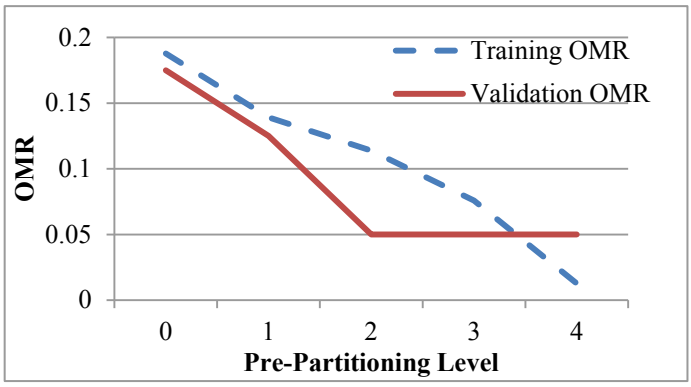


Figure D-14. Training & Validation OMR at different Prepartioning Level

Table D-9 gives the details of the best validation filters. Note that the best training filters are not the best validation filters.

Table D-9 Details of Flow Cytometry Variables in the Best Validation Filters in Different CART Pre-partitioning Levels

Variable/Node	0	1-1	1-2	2-1	2-2	3-1	3-2	4-2	4-3
act4103			-0.08		-1.21				
act425lo	1.26		1.30			0.65		1.85	
act425tot			1.04		-1.45	0.48		1.89	1.62
act8103	3.23				0.06				
act810371		2.69		2.43			1.05		
mememm4	3.59	1.54							

mememm478								0.65
mememm48	3.94					1.83		
memcm478					-1.33	1.05		
mememm8		-0.87	0.23	-0.87				
mememm878							-0.97	2.05
mememra4				1.14		5.22	1.00	
mememra478				4.78				
mememra4k		6.50						
pol8ccr4								
pol8th2					1.95	-0.68		
pol8th17			2.84					2.48
pol8x3r4ratio		2.60		1.68	-1.28			
pol8th1th2ratio				1.10				
traff4ccr3					0.29		0.91	-0.11

Note that each column contains information of the best filter in that particular level.

Only 20 of the 27 “best FC set” variables were used in the best performing validation filters. All act4103 variables that were active in establishing ILD status had thresholds below their mean. On occasion we see thresholds that correspond to extremely high FC expressions (e.g., 5.22 standard deviations above the mean for mememra4; 6.50 for mememra4k).

D.7 Generalized Linear Regression Model Results

Results in this section were based on data set IRIS041314 (defined in §B-2). The estimated coefficients of the best GLM found based on the stepwise model selection algorithm is shown in Table D-10. The p-value of Goodness-of-Fit–Test (Pr($\chi^2_{13} > 128.58$)) is approximately 0.15, which is larger than 0.05. It indicates that there is no evidence that the model is inadequate.

Table D-10 Estimated Coefficients of Stepwise GLM

	Mean	Standard Error	z value	Pr(>|z|)
(Intercept)	-0.95	1.32	-0.72	4.70E-01
pol8ccr5cxcr3neg	-0.12	0.04	-2.81	**4.89E-03**
pol8ccr5cxcr3	0.03	0.02	2.00	**4.50E-02**
mememe8	-0.25	0.10	-2.42	**1.54E-02**
mememra87	0.43	0.18	2.43	**1.53E-02**
memcm878	0.29	0.12	2.53	**1.16E-02**
memcm4	-0.03	0.02	-2.00	**4.54E-02**
type	1.05	0.47	2.22	**2.64E-02**
scl70_ab	0.94	0.46	2.03	**4.25E-02**
dd1symptom_y	0.07	0.03	2.28	**2.25E-02**
Residual deviance: 128.58 on 113 degrees of freedom				

Note: Bold values are statistically significant at 5% significance level or lower

A Drop-in-Deviance Test comparing the Null model (or Intercept only model) with current model was performed with Drop-in-Deviance equal to 39.58 (=168.16 - 128.58) and d.f. equal to 9 (=122-113). The corresponding p-value is smaller than 10^{-6}, which suggests that the null hypothesis that coefficients of all variables in the model are equal zero is rejected. In other

words, the variables in stepwise GLM (pol8ccr5cxcr3neg, pol8ccr5cxcr3, mememe8, mememra87, memcm878, memcm4, type, scl70_ab, dd1symptom_y) are all statistically significant.

Table D-11 95% Confidence Interval of the Estimated Coefficients of Stepwise GLM

	Estimate	**2.50%**	**97.50%**
pol8ccr5cxcr3neg	-0.124	-0.217	-0.043
pol8ccr5cxcr3	0.033	0.002	0.066
mememe8	-0.252	-0.472	-0.061
mememra87	0.434	0.105	0.814
memcm878	0.294	0.081	0.542
memcm4	-0.034	-0.068	-0.002
type	1.050	0.142	2.009
scl70_ab	0.937	0.041	1.863
dd1symptom_y	0.074	0.013	0.141

Depending on the sign of coefficient for each FC variable, certain variables are negatively associated with the odds (not probability) of having ILD. As an example, holding the other variables constant, the odds of having ILD will change by a multiplicative factor of 0.8836 (or, exp (-0.12374)) with one unit increase of pol8ccr5cxcr3neg. In other words, the odds of having ILD will be 11.67% smaller with one unit increase of pol8ccr5cxcr3neg.

To further explore the relative importance of the group of clinical variables v.s. the 6 FC variables, two reduced GLMs (clinical-only model and FC-only model) were fitted and compared with the stepwise GLM.

Table D-12 Estimated Coefficients of Reduced GLMs

Clinical-only Model				
	Mean	Std. Error	z value	Pr(>|z|)
(Intercept)	-1.90	0.46	-4.12	**3.82E-05**
type	1.27	0.43	2.96	**3.12E-03**
scl70_ab	0.93	0.41	2.25	**2.46E-02**
dd1symptom_y	0.08	0.03	2.69	**7.22E-03**
Residual deviance: 147.79 on 119 degrees of freedom				
FC-only Model				
	Mean	Std. Error	z value	Pr(>|z|)
(Intercept)	1.07	1.10	0.97	3.32E-01
pol8ccr5cxcr3neg	-0.11	0.04	-2.73	**6.30E-03**
pol8ccr5cxcr3	0.03	0.02	2.06	**3.94E-02**
memem8	-0.27	0.09	-2.88	**3.99E-03**
mememra87	0.47	0.17	2.82	**4.79E-03**
memcm878	0.32	0.11	2.99	**2.77E-03**
memcm4	-0.05	0.02	-2.94	**3.32E-03**
Residual deviance: 141.56 on 116 degrees of freedom				

Table D-13 Drop-in-Deviance-Test Comparing Stepwise GLM and Reduced GLMs

	Drop-in-Deviance	Drop-in-d.f.	p-value
Clinical-only Model	19.21	6	**3.83E-03**
FC-only Model	12.98	3	**4.68E-03**

In Table D-13, both of the estimated *p*-values of the Drop-in-Deviance-Test are smaller than 0.005, which implies that there is strong statistical evidence that the 3 clinical variables and the six FC variables individually and as groups are highly associated with the odds of having ILD. In particular, with six FC variables, the residual deviance of stepwise GLM model decreases by a statistically significant amount (p-value < 0.005).

D.7-1 Case-Influence Statistics

During the process of searching for a statistical model that fits the data set well, it is important to examine the individual influence of selected data. Case-influence statistics are mathematical measurements used to characterize such individual influence (Ramsey & Schafer, 2012). Many of them have been used to further improve fitted models (Belsley, Kuh and Welsch (1980), Cook and Weisberg (1982), Williams (1987), Fox (1997, 2002)). Four different kinds of statistics have been utilized in this research: leverages, standardized residuals and Studentized residuals, Cook's distances.

(1) Leverage measures the distance between the explanatory variable value of a case (in this research, a case means a patient) and the mean of the explanatory variables value of all cases. The leverage of the i^{th} case is:

$$h_i = \frac{(x_i - \bar{x})^2}{\sum (x - \bar{x})^2} + \frac{1}{n}$$

where, x_i is the explanatory variable value of the i^{th} case, $\bar{x}$ is the average of all x, and n is the sample size.

(2) Standardized residual is the residual of a case divided by the variance. Its formula is the following:

$$r_i^* = \frac{\hat{\varepsilon}_i}{\hat{\sigma}\sqrt{1-h_i}}$$

where, $\hat{\varepsilon}_i$ is the deviation or residual of the i^{th} case, $\hat{\sigma}$ is the estimated standard deviation from the fit, and h_i is the leverage of the i^{th} case.

(3) Studentized residual is similar to standardized residual except using a different definition of variance in which the calculation does not include the case of interest.

$$t_i = \frac{\hat{\varepsilon}_i}{\hat{\sigma}_{(-i)}\sqrt{1-h_i}}$$

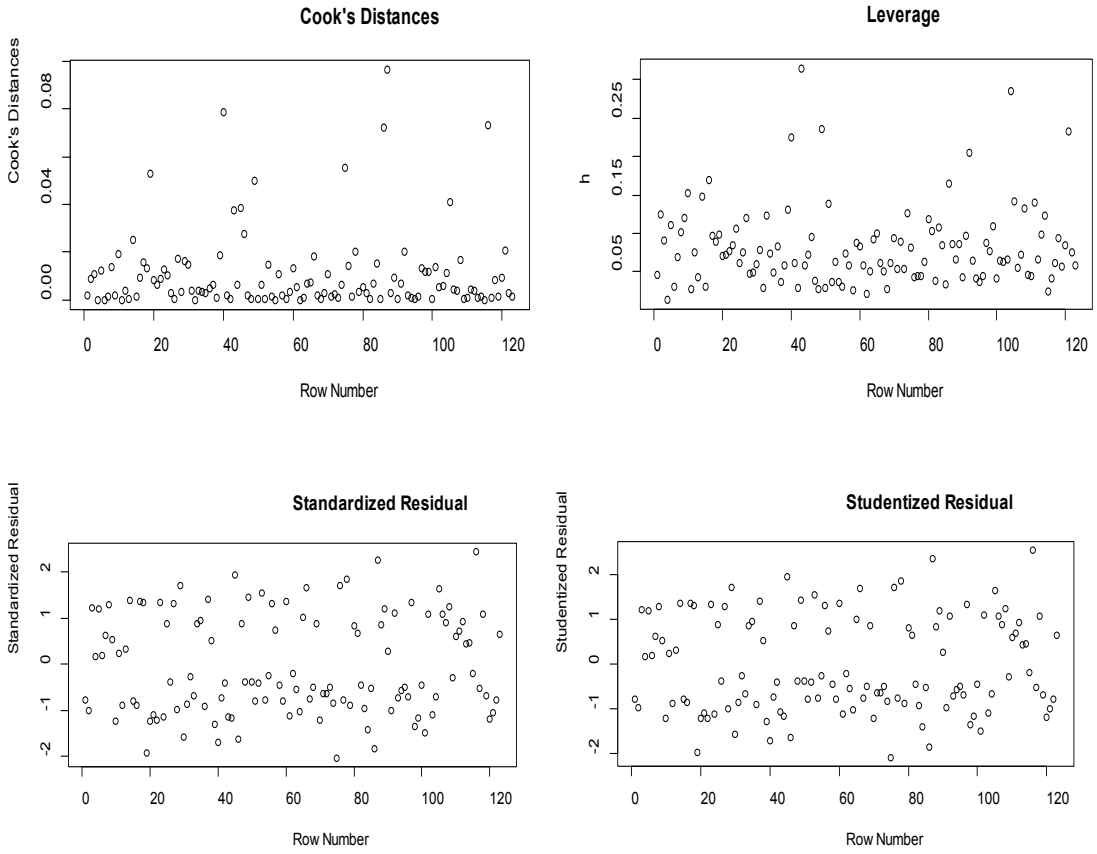
(4) Cook's Distance measures the effect of omitting the i^{th} case upon the estimated overall regression coefficients, using the following formula:

$$D_i = \sum_{j=1}^{n} \frac{\left(\hat{Y}_{j(i)} - Y_j\right)^2}{p\hat{\sigma}^2}$$

where, $\hat{Y}_j$ is the j^{th} fitted value based on a fit model using the entire data set, $\hat{Y}_{j(i)}$ is the j^{th} fitted value in a fit excluding the i^{th} case from the data set, p is the number of parameters and $\hat{\sigma}$ is the estimated standard deviation from the fit. As shown in Figures D-15, few patients had Cook's distances and leverage that were distinct from the majority. The Studentized residuals and

standardized residuals plots also suggest that no outliers were found based on stepwise GLM.

Figure D-15. Diagnostics Statistics Based on GLM (Phenotype is ILD)



D.8 Partial Dependence Analyses

Results in this section were based on data set IRIS041314. A Partial Dependence Plot is a tool to estimate the marginal effects of a subset of explanatory variables (usually less than 3) upon a response, accounting for the effects of all other FC variables on that response (Hastie et al., 2009). The method used is CRF. The PDP procedure is:

- Select a FC variable in the FC set
- Sort the expressions of that variable in the IRIS data set
- Replace all expressions of that variable with the first expression in the sorted list; No other FC expressions in the data set are changed
- Perform CRF with that newly created IRIS matrix (only one column has changed and all of that column's row entries are the same); the CRF result is probability of having ILD
- Repeat for all the other expressions in the sorted list; plot completed
- Select another FC variable in the FC set

Figure D-16. Partial Dependence Plots for FC Variables (Phenotype is ILD)

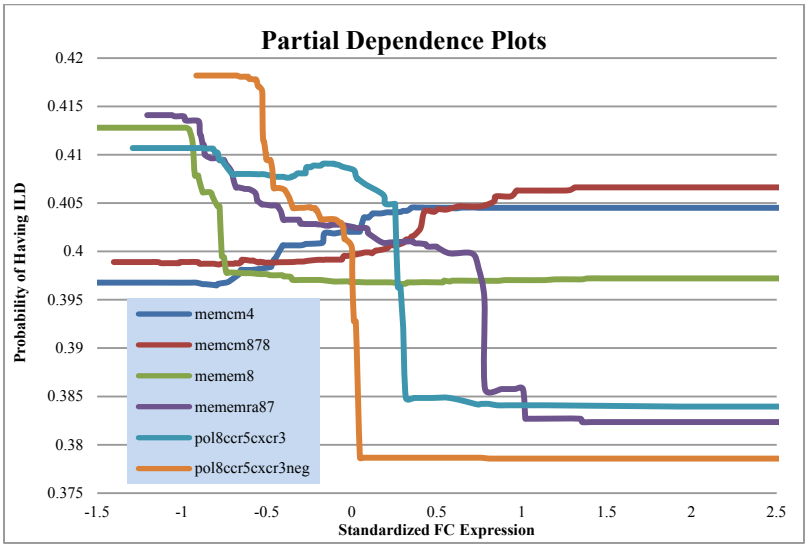


Figure D-16 shows that as the values of FC variables memcm878 and memcm4 increase, the probability of having ILD becomes larger. The remaining four FC variables had the opposite effect on the probability of having ILD. Increasing trends for the two memory panel variables were limited between probability 0.395 and 0.41, while the ranges of declining trends were relatively larger. FC variable mememl8 has the smallest decrease (from about 0.415 to 0.395). Mememra87 and two polarization variables reduced from 0.42 to 0.375.

D.9 Phenotype as Cancer

So far, the phenotype of interest has been ILD. In this section, the same procedures (CRF-GSEA-Stepwise GLM/PDP) were applied to another phenotype 'Cancer' in order to identify a group of flow cytometry variables as a whole that are strongly associated with cancer in SSc patients.

D.9-1 CRF-GSEA

Using the IRIS071813 data set (Section §B-2), CRF classification was first performed. Through Gene Set Enrichment Analysis, the relationship between FC set size and Enrichment Score was examined. Next, the FC variables that comprise the FC sets were identified.

Following are some preliminary results, beginning with an in-sample CRF ROC curve and continuing through FC set size experiments, associated p-values, the GSEA Random Walk and finally a table of FC expressions in FC set sizes 5 to 50. In Figure D-17, it shows that the AUC of the in-sample ROC curve associated with CRF was 1, which suggests that the CRF model fitted the data set very well.

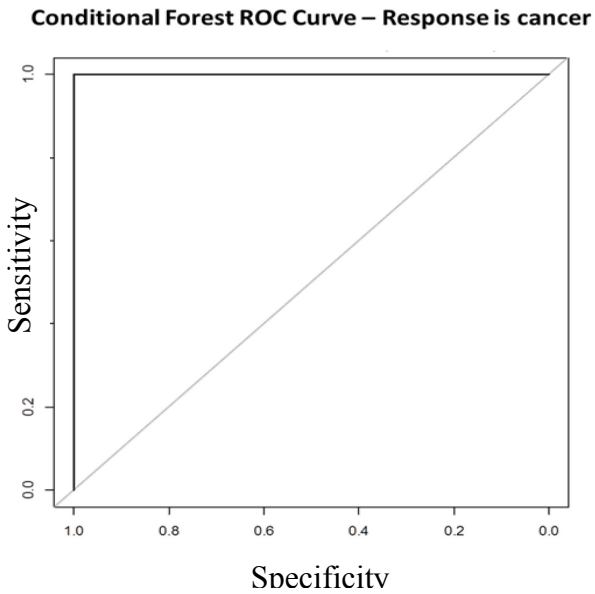


Figure D-17. ROC Curve for CRF (Phenotype is Cancer)

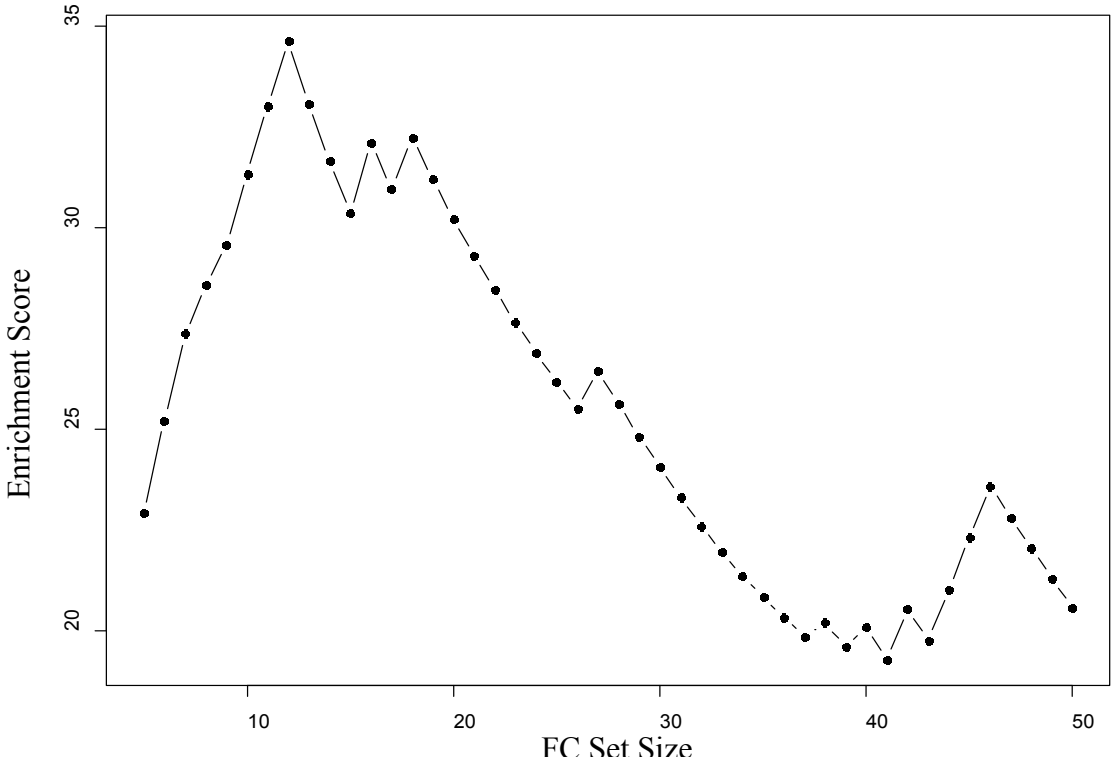


Figure D-18. Enrichment Scores of GSEA for Different FC Set Sizes (Phenotype is Cancer)

Figure D-18 shows that the highest ES is associated with FC set size 12. ES increases as FC set size becomes larger from set size 5 to set size 12 until it reaches the peak 34.6, but displays a sharp declining trend with greater FC set size. The estimated p-values associated with each ES are all below 10^{-4} when FC set size is smaller than 30 (see Figure D-19) but become less significant afterward. However, they are all below the significance level 0.05.

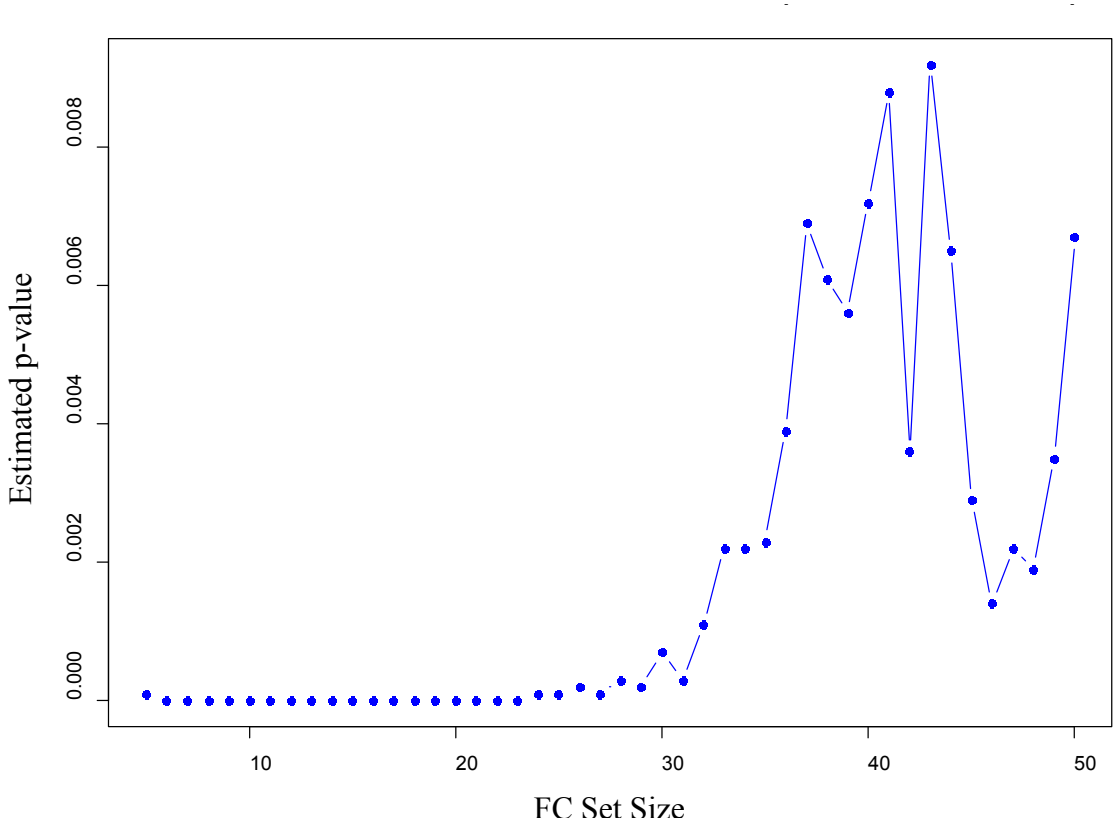


Figure D-19. P-values associated with Enrichment Scores of GSEA for Different FC Set Sizes (Phenotype is Cancer)

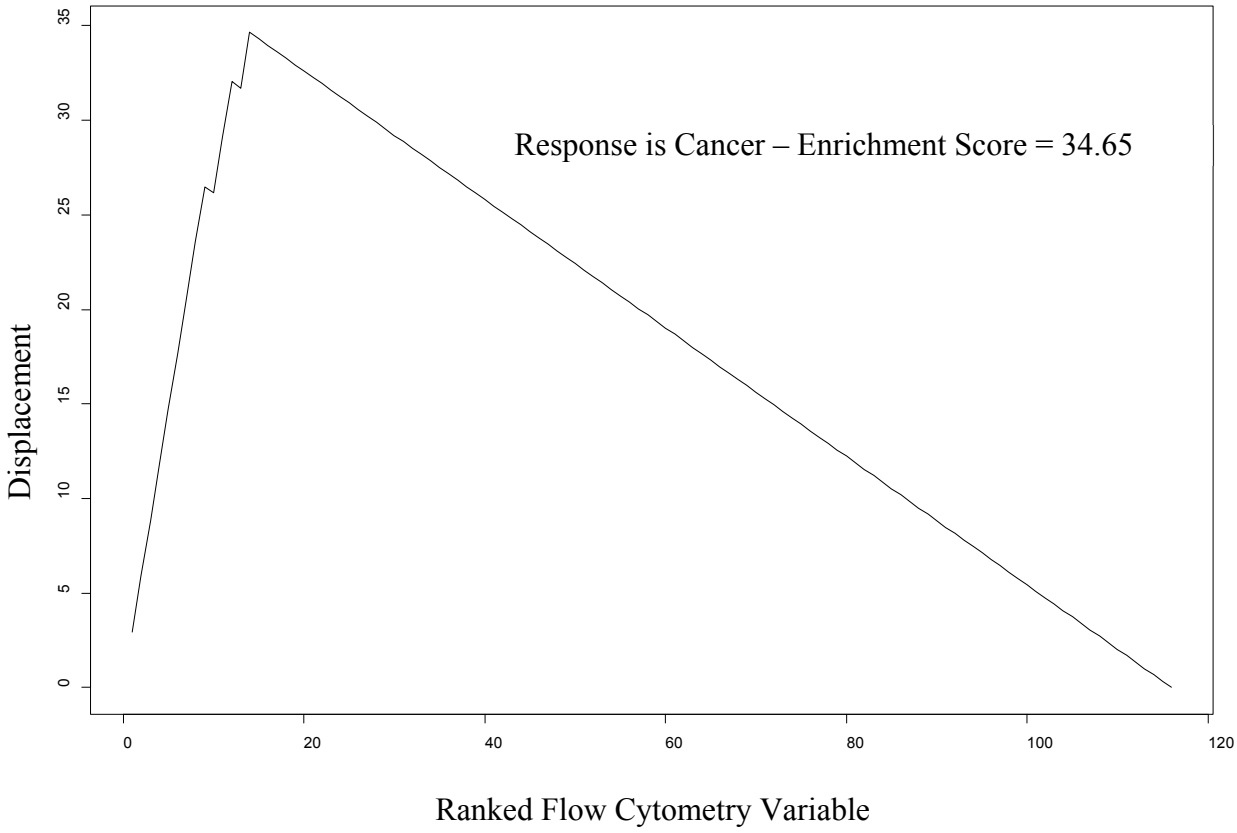


Figure D-20. Random Walk that Results from FC Set Comprised by Top 12 Most Important Variables

D.9-2 Stepwise GLM

All the results shown from now to the end of this section were based on data set IRIS041314. A logistic binomial linear regression model with 12 FC as covariates was fit to estimate the binary outcome CANCER, followed by the stepwise variable selection algorithm. The results are shown in Table D-14.

Table D-14 Estimated Coefficients of Stepwise GLM (Phenotype is Cancer)

	Mean	Standard Error	z value	Pr(>|z|)
(Intercept)	-4.63	1.21	-3.82	**1.33E-04**
act4103hladr	3.05	1.09	2.80	**5.07E-03**
traff4ccr3	0.66	0.33	2.02	**4.39E-02**
act825	0.10	0.04	2.49	**1.28E-02**
pol4th2	-0.19	0.08	-2.27	**2.35E-02**
Residual deviance: 66.302 on 134 degrees of freedom				

From the goodness-of-fit Test (p=0.9999998; a large p value indicates that the model is adequate) and Drop-in-Deviance test comparing the current model with the Null model ($p < 10^{-5}$), we know that stepwise GLM is sufficient to estimate response CANCER.

Table D-15 95% Confidence Interval of the Estimated Coefficients of Stepwise GLM

	Estimate	2.50%	97.50%
act4103hladr	3.053	0.983	5.384
traff4ccr3	0.658	0.043	1.350
act825	0.103	0.024	0.189
pol4th2	-0.189	-0.377	-0.046

Table D-16 95% C.I. of the Exponentiated Estimated Coefficients (of Stepwise GLM) Subtracted 100%

	Estimate	2.50%	97.50%
act4103hladr	20.182	1.672	216.823
traff4ccr3	0.930	0.044	2.859
act825	0.108	0.024	0.208
pol4th2	-0.173	-0.314	-0.045

Among the 4 FC variables, pol4th2 was negatively associated with the odds of having cancer, i.e., holding the other variables constant, the odds of having cancer will be 17.3% smaller with one unit increase of pol4th2. The other three variables are positively associated with the odds of having cancer, among which variable act4103hladr had the strongest association – approximately 20 times higher with one unit increase in the FC expression.

Four FC variables were discovered to be highly associated with phenotype 'Cancer'. Stepwise GLM and Partial Dependence Plot were useful in drawing statistical inference from the identified FC set.

Note:
act4103hladr = CD3+/CD4+/CD8-/CD103+/HLADR+
traff4ccr3 = CD3+/CD4+/CD8-/CCR3+
act825 = CD3+/CD4-/CD8+/CD25+
pol4th2 = CD3+/CD4+/CD8-/CXCR3-/CCR4+/CCR6-

The ROC curve for this model was also plotted (see Figure D-21). The Area Under Curve (AUC) is 0.8688, which supports the conclusion that stepwise GLM fits the data set reasonably well.

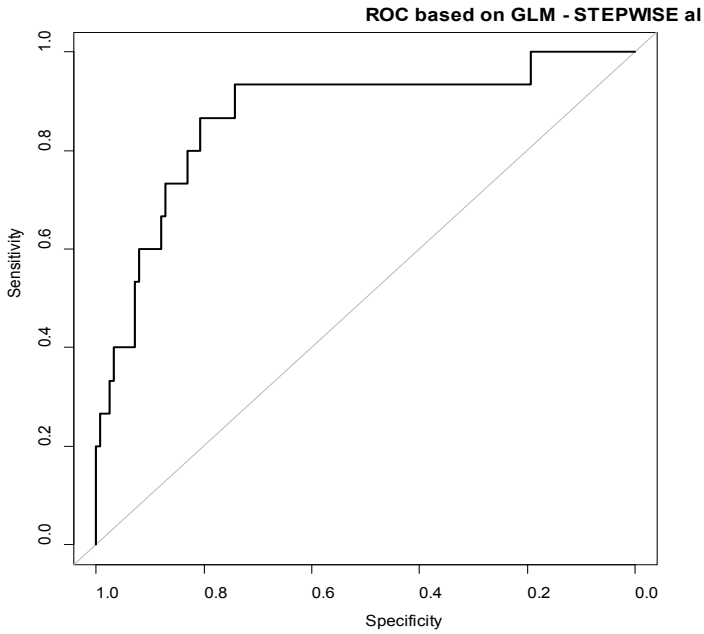


Figure D-21. In-sample ROC Curve for Stepwise GLM

Stepwise GLM yielded the result that only 4 FC variables remain. Each is statistically significant at the 5% level. To further explore relationships among these covariates, a matrix of scatter plots between pairs of these variables is presented in Figure D-22. Two different symbols are used to represent patients with (red triangle) and without (black square) cancer diagnosis. We find in Figure D-22 that variable pol4th2 is particularly intriguing in this context. In the last column of the matrix of scatter plots, the red triangles cluster cohesively toward the left while the black dots spread more toward the bottom.

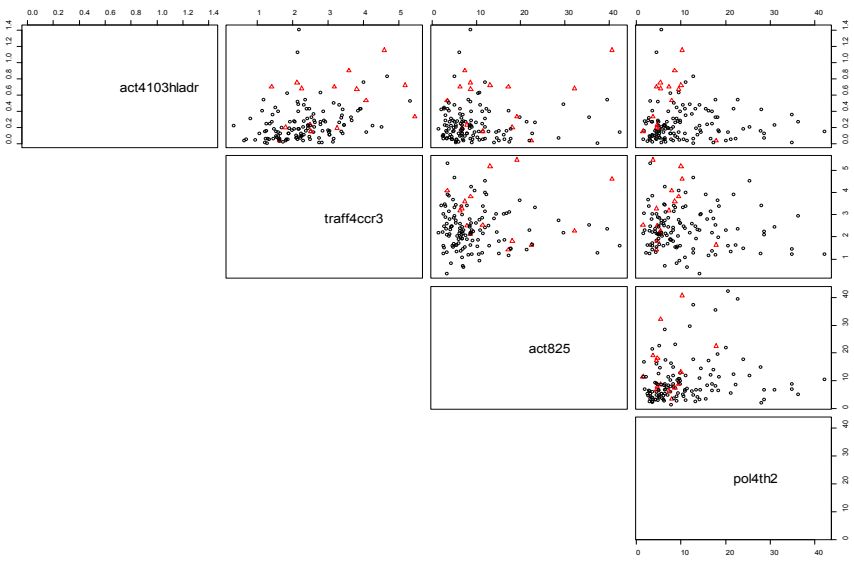


Figure D-22. Matrix of Scatter Plots for 4 FC Variables in GLM

D.9-3 Diagnostic Statistics

Based on four case influence statistics described in Section §D. 7-1, two patients (ID:2202, 3083) are considered as isolated observations (identified by the blue arrows in the figures). Note that when the response variable was ILD, no isolated observations were found. To determine their influence, a stepwise GLM was implemented using the data set without these two isolated cases. Table D-17 gives the output.

Figure D-23. Diagnostics Statistics Based on GLM (Phenotype is Cancer)

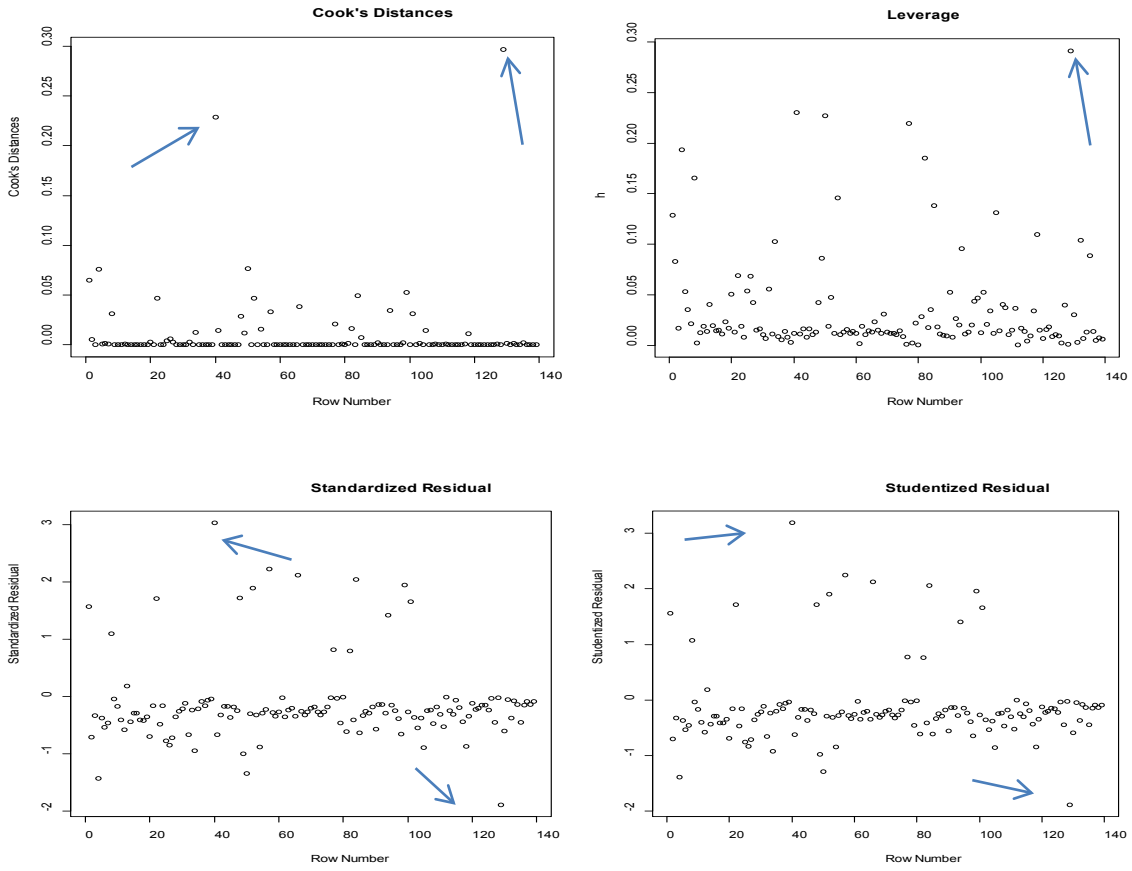


Table D-17 Estimated Coefficients of Stepwise GLM

	Mean	Standard Error	z value	Pr(>|z|)
(Intercept)	-4.81	1.43	-3.37	**7.63E-04**
act4103hladr	5.52	1.63	3.39	**6.89E-04**
traff4ccr3	0.67	0.38	1.78	7.50E-02
act825	0.11	0.05	2.10	**3.53E-02**
pol4th2	-0.35	0.13	-2.70	**7.04E-03**
Residual deviance: 50.893on 132 degrees of freedom				

Compared with output of the stepwise GLM with full data set, the coefficients of variable act4103hladr (from 3.05 to 5.52) changed enormously while the others less so. Variable traff4ccr3 became less

significant but the p-values associated with act4103hladr and pol4th2 decreased. The p-value of the corresponding Drop-in-Deviance test is 8.54e-06, which highly suggests that these 4 FC variables are statistically significant. The goodness-of-fit Test (§C.5-1) gives a p-value that is almost 1. It indicates that there is no evidence for the inadequacy of the fitted model. For these four FC variables, Student's t-tests were undertaken with the null hypothesis that the mean difference between FC variable expressions is zero for SSc cancer and SSc non-cancer patients. Shown in Table D-18, zero was not included in the 95% C.I. of the mean difference between non-cancer patients and cancer patients when the FCs of interest were act4103hladr and traff4ccr3. It suggests that the act4103hladr and traff4ccr3 expression differs between cancer patients and non-cancer patients with a 95% confidence interval.

Table D-18 95% C.I. for 4 FC Variables in GLM based on Student t-tests

FC Variable	mean difference	95% C.I.	
act4103hladr	-0.290357	-0.31972	-0.26099
traff4ccr3	-0.8182419	-1.31324	-0.32324
act825	-5.3151505	-40.6785	30.04821
pol4th2	3.8572312	-31.115	38.82946

Next, stepwise GLM without CRF-GSEA using the full data set was fitted to evaluate the influence of the CRF-GSEA procedure. The results (see Table D-19) show that 16 FC variables out of 112 remain but none of them

are statistically significant. The implication here is that our hybrid CRF-GSEA procedure is highly effective in selecting important FC sets.

Table D-19 Details of the Fitted GLM Using Full Data set

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-2821.09	226601.20	-0.01	0.99
mememra48	2253.91	217791.60	0.01	0.992
memcm8	4892.27	401141.20	0.01	0.99
memem8	-167.93	13631.79	-0.01	0.99
memcm878	-4901.00	402153.80	-0.01	0.99
memcm87	-4531.66	394922.70	-0.01	0.991
memcm88	-4544.56	388574.60	-0.01	0.991
memcm80	-5152.71	396416.50	-0.01	0.99
memem878	172.73	14162.81	0.01	0.99
memem87	268.68	21355.31	0.01	0.99
memem80	153.21	12911.43	0.01	0.991
pol4cxcr3	20.50	1873.06	0.01	0.991
pol8x3r4ratio	18.85	1590.05	0.01	0.991
act425lo	15.50	1380.24	0.01	0.991
act4103hladr	567.84	51596.43	0.01	0.991
act82571	446.11	51474.60	0.01	0.993
act86971	-450.96	45674.70	-0.01	0.992

D.9-4 Partial Dependence Analyses

The PDP plots in Figure D-24 show that, with the exception of pol4th2, as all four remaining FC variables from stepwise GLM increased, the probability of having cancer in SSc patients increases.

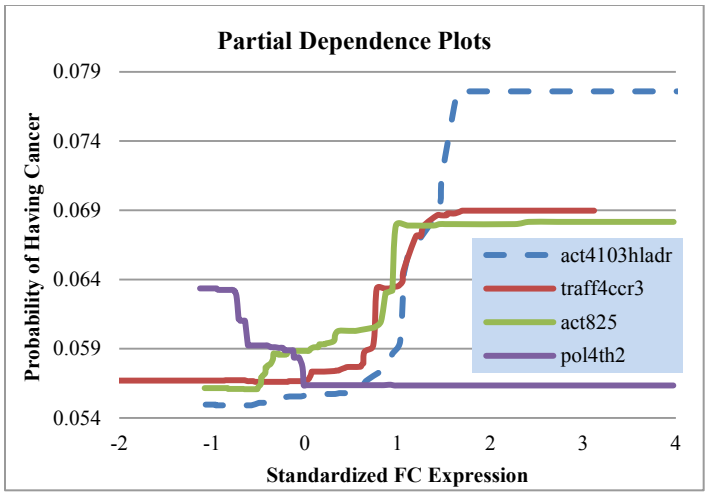


Figure D-24. Partial Dependence Plots for FC Variables (Phenotype is Cancer)

In particular, act4103hladr had the largest effect: the probability of having cancer ascended from approximately 0.054 to 0.079 as the standardized act4103hladr expression increased from around -1 to 1.7. On the contrary, the cancer probability decreased from around 0.063 to 0.056 when standardized pol4th2 increased from about -1 to 0.

Two different angles for the same 3D PDP with pol4th2, act825 and cancer probability are shown in Figure D-25. It was an estimated surface in which cancer probability grew with combination of decreasing pol4th2 and increasing act825.

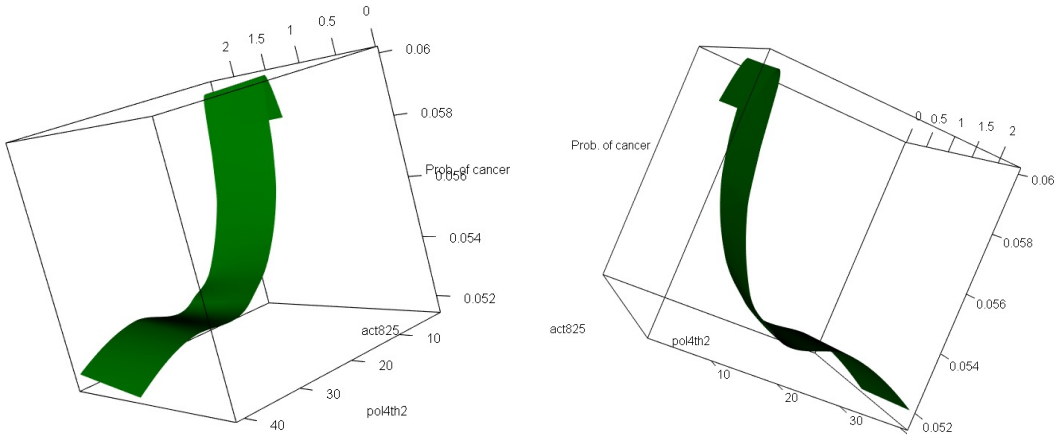


Figure D-25. 3D PDP with Two FC Variables

Another issue was revisited: if the variables in the FC set are highly correlated, how would that affect the final results and interpretation? When using cancer as phenotype based on 071813 data set, the FC set includes 4 highly correlated variables: one mother node (act4103) and three child nodes (act4103hladr, act410371, act425103). To examine the effect of variable correlation and robustness of PDP, 3 child variables from the FC set containing the parent act4103 were removed before the PDP of act4103 was created. Figure D-27 indicates that no significant change occurred.

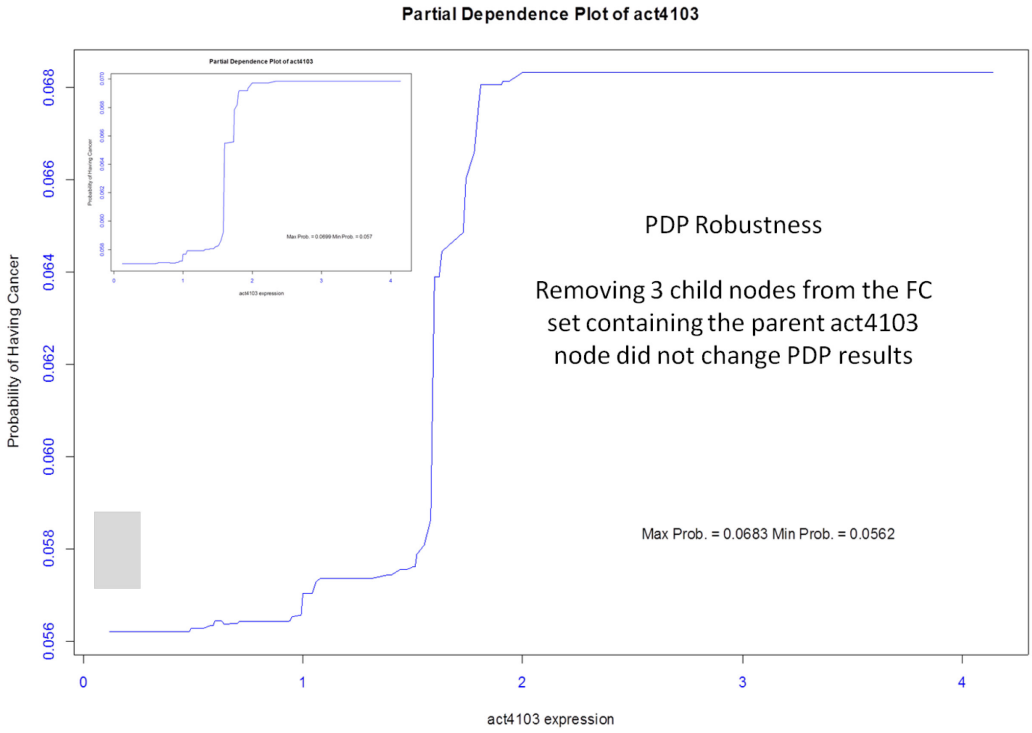


Figure D-26. PDP of Act4103 Before and After Removing 3 Child Nodes

The results in this section are all related to phenotype 'Cancer'. In the next section, results of how the phenotypes ILD and Cancer are related in terms of GSEA will be presented.

D.10 ILD – Cancer GSEA Intercomparison

In order to examine inter-relationship between phenotype ILD and Cancer, results of GSEA inter-comparison of these two are shown here.

Results in this section were based on data set IRIS071813. GSEA performance of the cancer FC set with ILD as the phenotype and vice-versa

were examined. The results were immediately obvious – GSEA performance in these experiments was not statistically significant. When ILD is phenotype using the cancer set, the ES is always smaller than 12 and very often below 10 regardless of FC set size (see Figure D-27). The associated estimated p-values are all above 0.3 which indicates a low level of statistical significance (show in Figure D-28). A similar situation occurs for cancer as the phenotype with the ILD FC set: no ES is above 12 and all estimated p-values are over 0.3 (See Figures D-29 and D-30). The performance of these random walks was completely different from what was seen previously when using the same phenotype as the basis for the establishment of the ranked list and FC set. The motivation for performing these experiments was to examine whether overlapping FC sets exist for both ILD and Cancer phenotypes.

Not only did these results indicate that there is little similarity shared between the two phenotypes in SSc patients, but they also highlight the specificity of different phenotypes. More detailed discussion can be found in section §E.8.

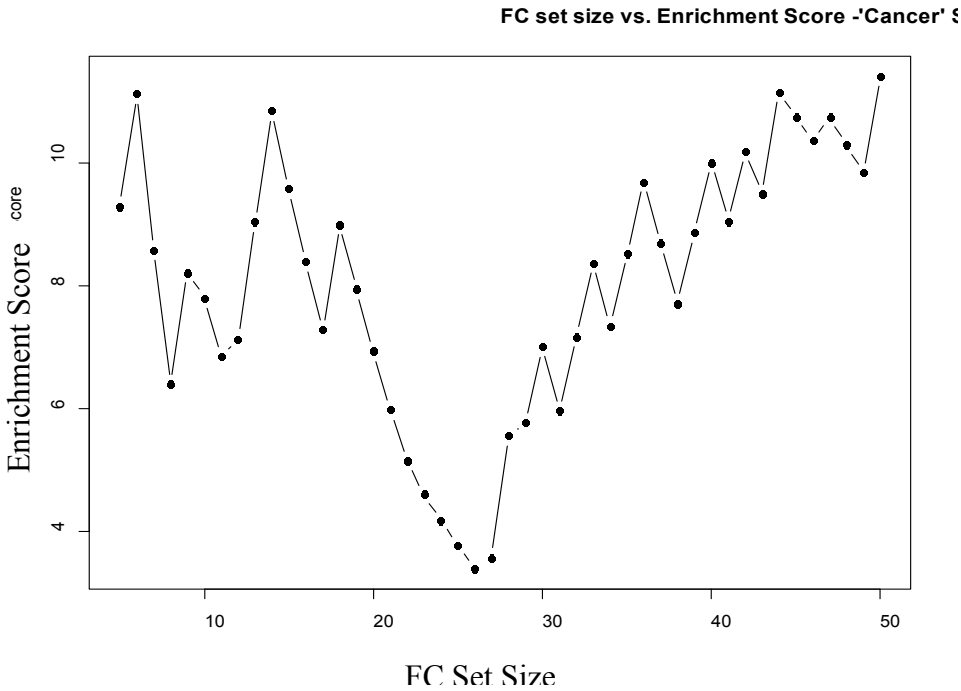


Figure D-27. Enrichment Scores of GSEA for Different FC Set Sizes (CANCER Set – ILD is phenotype)

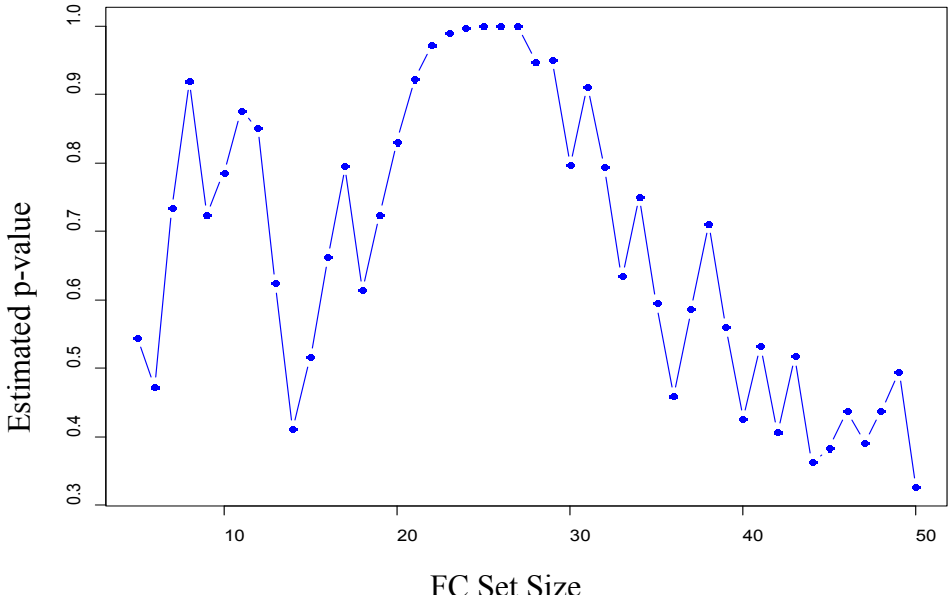


Figure D-28. P-values associated with Enrichment Scores of GSEA for Different FC Set Sizes (CANCER Set – ILD is phenotype)

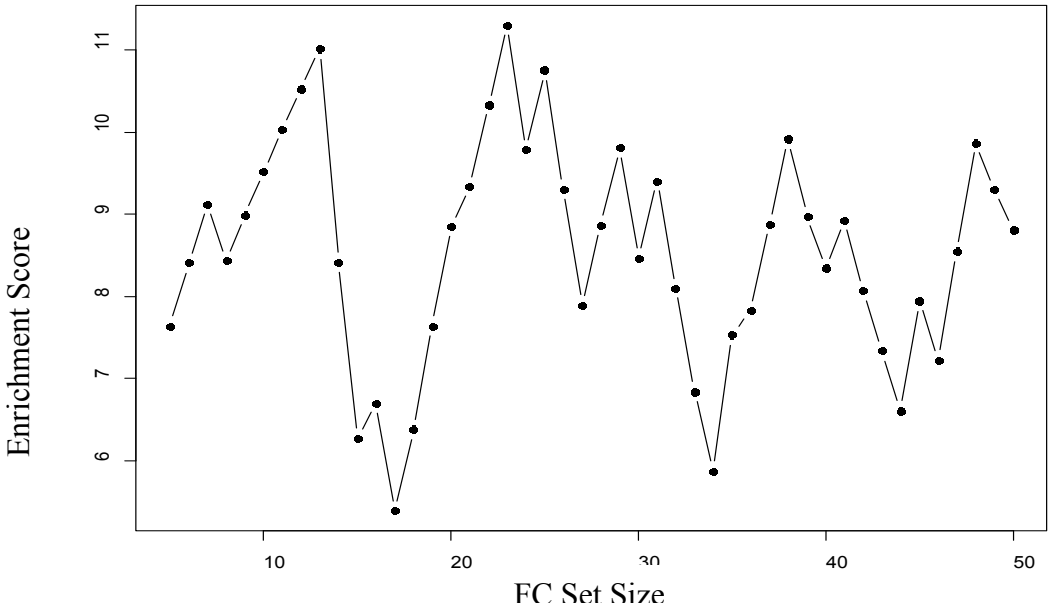


Figure D-29. Enrichment Scores of GSEA for Different FC Set Sizes (ILD Set –CANCER is phenotype)

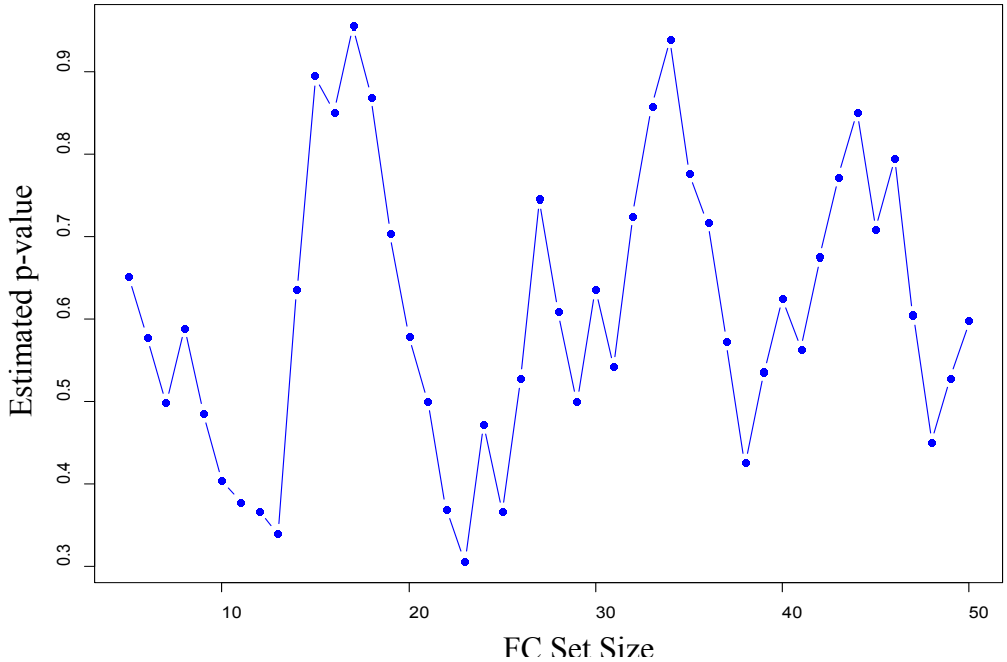


Figure D-30. P-values associated with Enrichment Scores of GSEA for Different FC Set Sizes (ILD Set –CANCER is phenotype)

In the following three sections, three different methods of determining FC sets were evaluated including biological information (§D.11), Student's t-tests statistics (§D.12) and p-values from point-wise GLM (§D.13).

D.11 Bio-informed FC Sets

Another modification for creating FC sets abandoned the CRF-informed VI list altogether and instead used two alternative sets (Table D-20) comprised of possibly important markers that were identified by our medical colleagues:

Table D-20 Biological Informed FC Sets List

CD4	CD8
pol4th1	act8103
pol4th2	pol8cxcr3
pol4th17	pol8ccr4
pol4th1th17	pol8ccr6
act425lo	act8dr
act425hi	traff8cxcr6
act4dr	traff8ccr10
traff4ccr10	memnaive8
traff4cxcr6	mem8cmefratio
memnaive4	mem8ememraratio
mem4cmefratio	memcd8k

Here is the random walk for the CD4 FC set (Figure D-31). No significant enrichment structure can be observed and the corresponding enrichment score is relatively low.

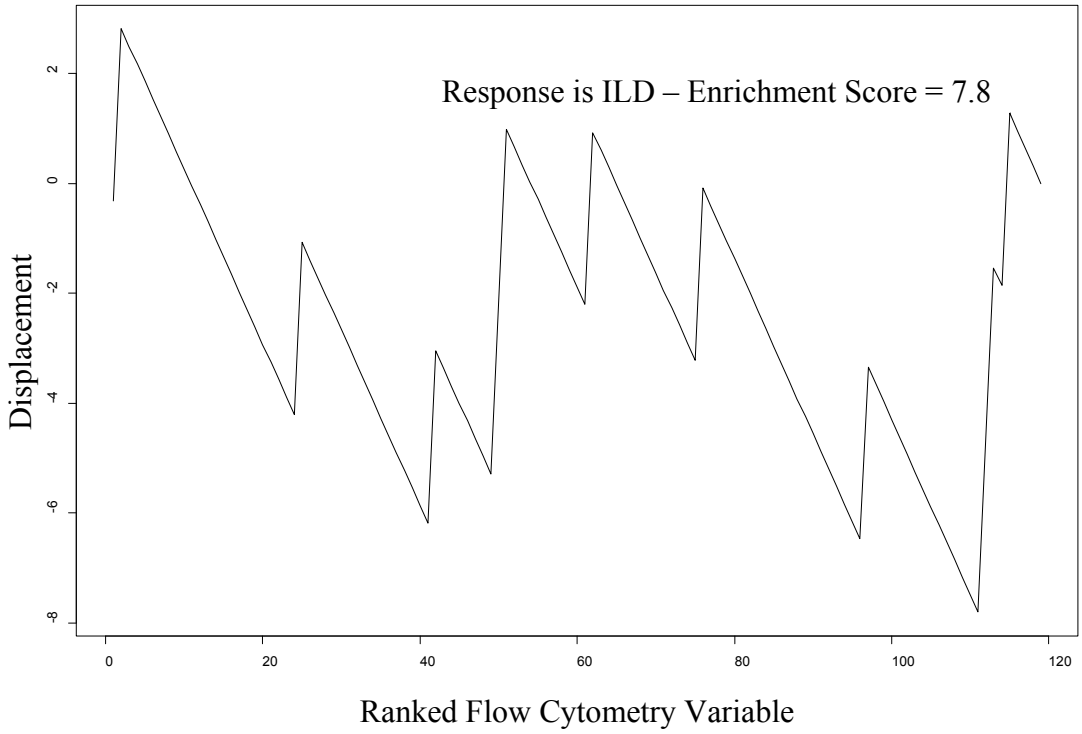


Figure D-31. Random Walk that Results from CD4 Bio-informed FC Set

Figures D-32 and D-33 show the corresponding GSEA results and significance levels for FC set sizes 3 to 11. No FC sets had statistically significant enrichment scores ($p > 0.25$). CD8 results were comparable.

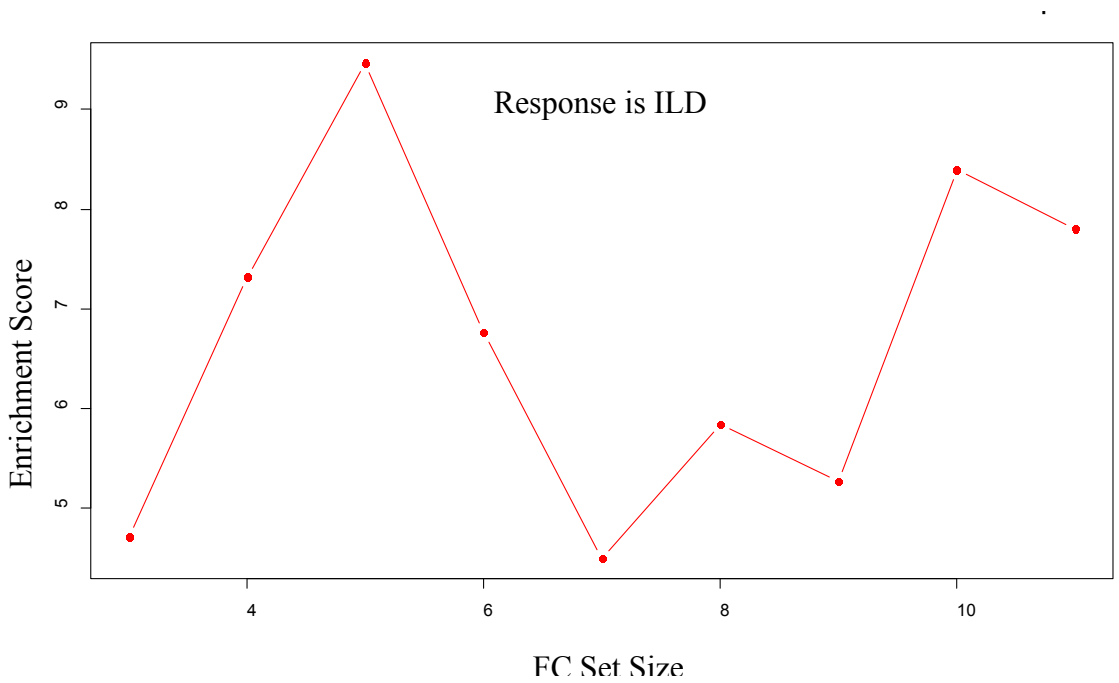


Figure D-32. Enrichment Scores of GSEA for Different FC Set Sizes (CD4 Bio-informed FC Set)

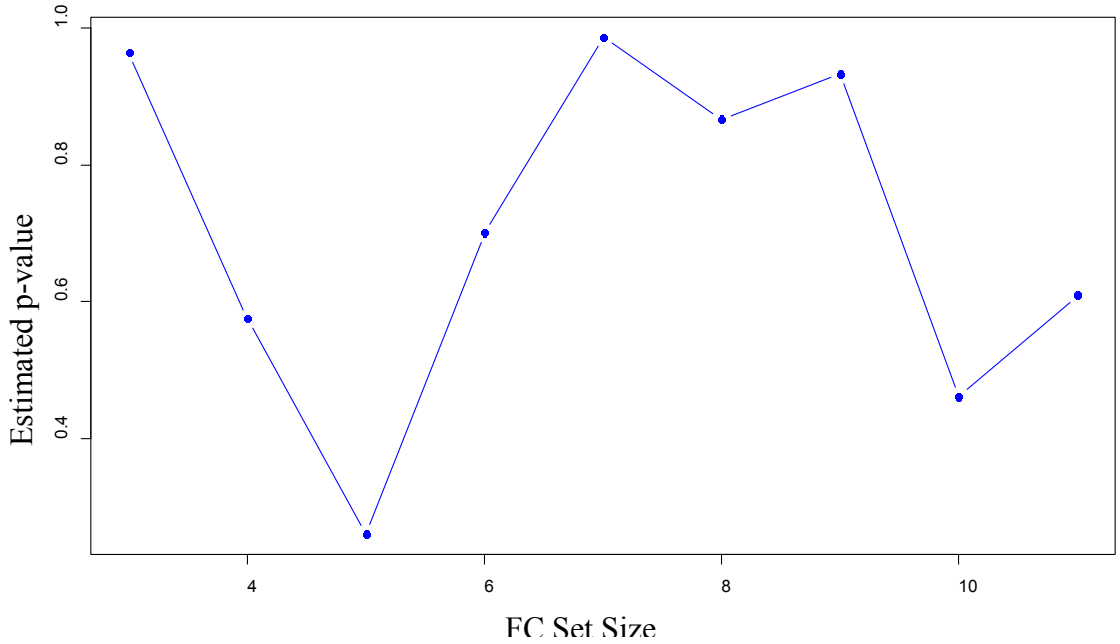


Figure D-33. P-values associated with Enrichment Scores of GSEA for Different FC Set Sizes (CD4 Bio-informed FC Set)

D.12 Student's t-tests Based FC sets

Two samples Student's t-tests for act4103hladr and traff4ccr3 expressions strongly suggested that there was differences between the cancer patients group and non-cancer patients group (at the 95% confidence level) (see §D.9-3). Motivated by these encouraging results, t statistics was used as the basis to construct FC sets. The followings are specific procedures of obtaining the Student's t-tests based FC set.

1. Perform t Tests for all 112 FC variables between cancer patients and non-cancer group.

2. Extract all the FC variables whose 95% C.I. excluded zero. Here, only 17 FC variables remained.

3. Ranked the 17 FC variables based on absolute signal-to-noise ratios. The signal here is the mean difference between cancer and non-cancer and the noise is the pooled sample variance.

4. Used the ranked 17 FC variables as input for GSEA.

It showed that the best FC set size is 12.

Table D-21 Comparison between Ranked t-test Set and CRF VI List

Ranked t-test Set	Absolute Signal-to-Noise Ratio	CRF
act4103hladr	5.34	act4103hladr
act410371	3.48	act410371
mememra48	2.16	traff8cxcr5
act425103	1.44	traff4ccr3
act810369	1.12	act4103
act4103	1.1	act425103
memcm88	1.03	act82571
act82571	0.99	act825
traff4ccr3	0.89	pol4th2
cd4cd8ratioLOG	0.87	memcm88
mememra88	0.83	memem4
act810371	0.7	traff8ccr10
memcm87	0.7	
memcm80	0.67	
act469hladr	0.62	
memcm40	0.55	
memem47	0.55	

In Table D-21, the t-test set based on ranked absolute signal-to-noise ratio is presented, so is the CRF FC set. The ones highlighted in yellow are those appears in both FC sets. Shown in Figure D-34, the ES for t Test based FC sets achieved its maximum at FC set size 12. It can also be discerned that the ES did not increase monotonically. The estimated p-values for permutation tests were all smaller than 0.05 regardless of FC set size. This suggests that the t Test based FC sets are statistically significant as a group.

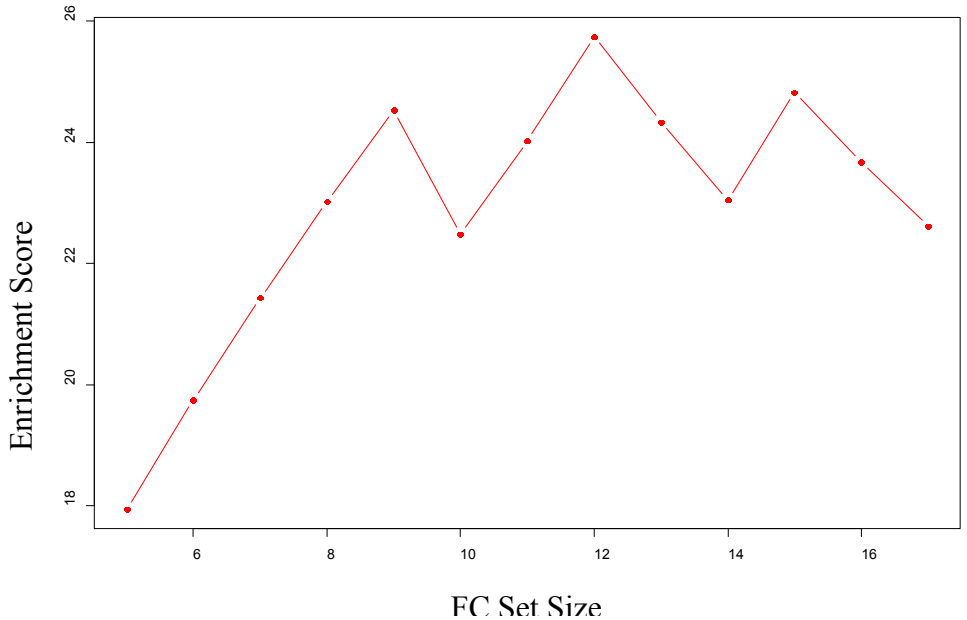


Figure D-34. Enrichment Scores of GSEA for Different Ranked t-tests FC Set Sizes (Cancer is Phenotype)

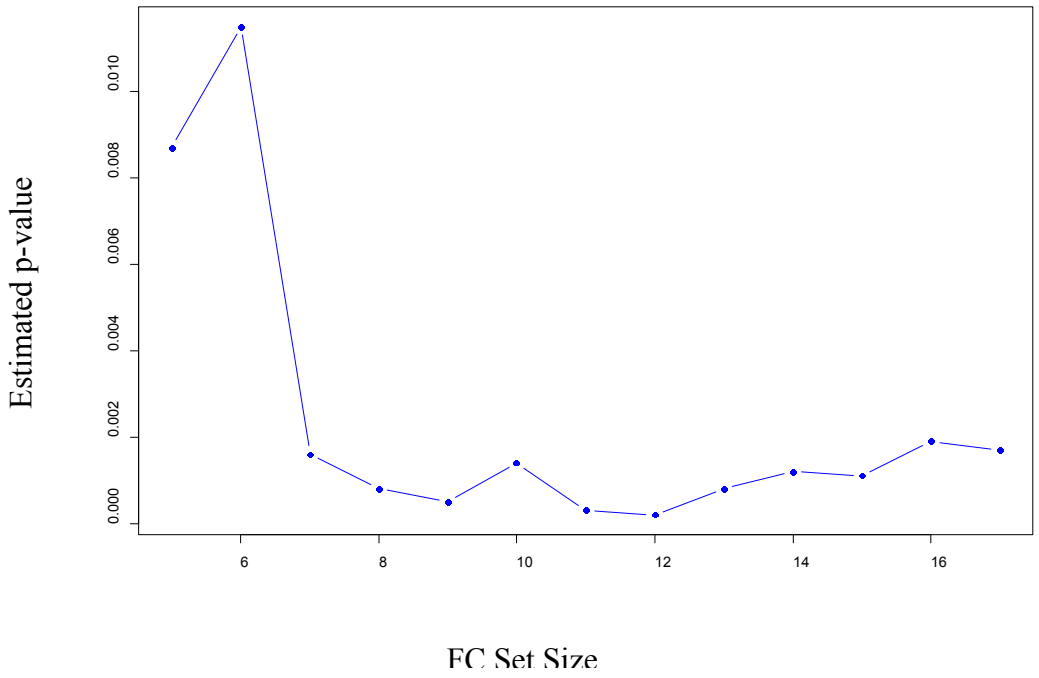


Figure D-35. P-values associated with Enrichment Scores of GSEA for Different Ranked t-tests FC Set Sizes (Cancer is Phenotype)

D.13 p-value Based FC Sets

D.13-1 Motivation and Procedure

Inspired by the success achieved through drawing statistical inference from stepwise GLM (§C.5), FC sets based on the p-value associated with FC variable in GLM were constructed and used as input to GSEA. The analysis procedures are:

- Construct a ranked FC list based on GLM p-value
- Use FC sets as input for GSEA
- Plot ES vs. FC set size

D.13-2 Phenotype is ILD

We first used ILD as the phenotype for our p-value derived FC set analyses. Figure D-36 shows a comparison of our original and new lists for creating FC sets. The three FC variables in red are those appear both in the p-value based FC set and stepwise GLM. FC variables showing up in both CRF list and p-value based FC set are highlighted in yellow. Red slope lines were used to indicate FC variables that have high ranking in CRF list but lower ranking in p-value based FC list, while green is to the opposite. Gray lines indicate no ranking changes for certain variables such as pol8ccr5 and

mememm4. The green horizontal line separates FC variable whose p-value is smaller than 0.05 from others.

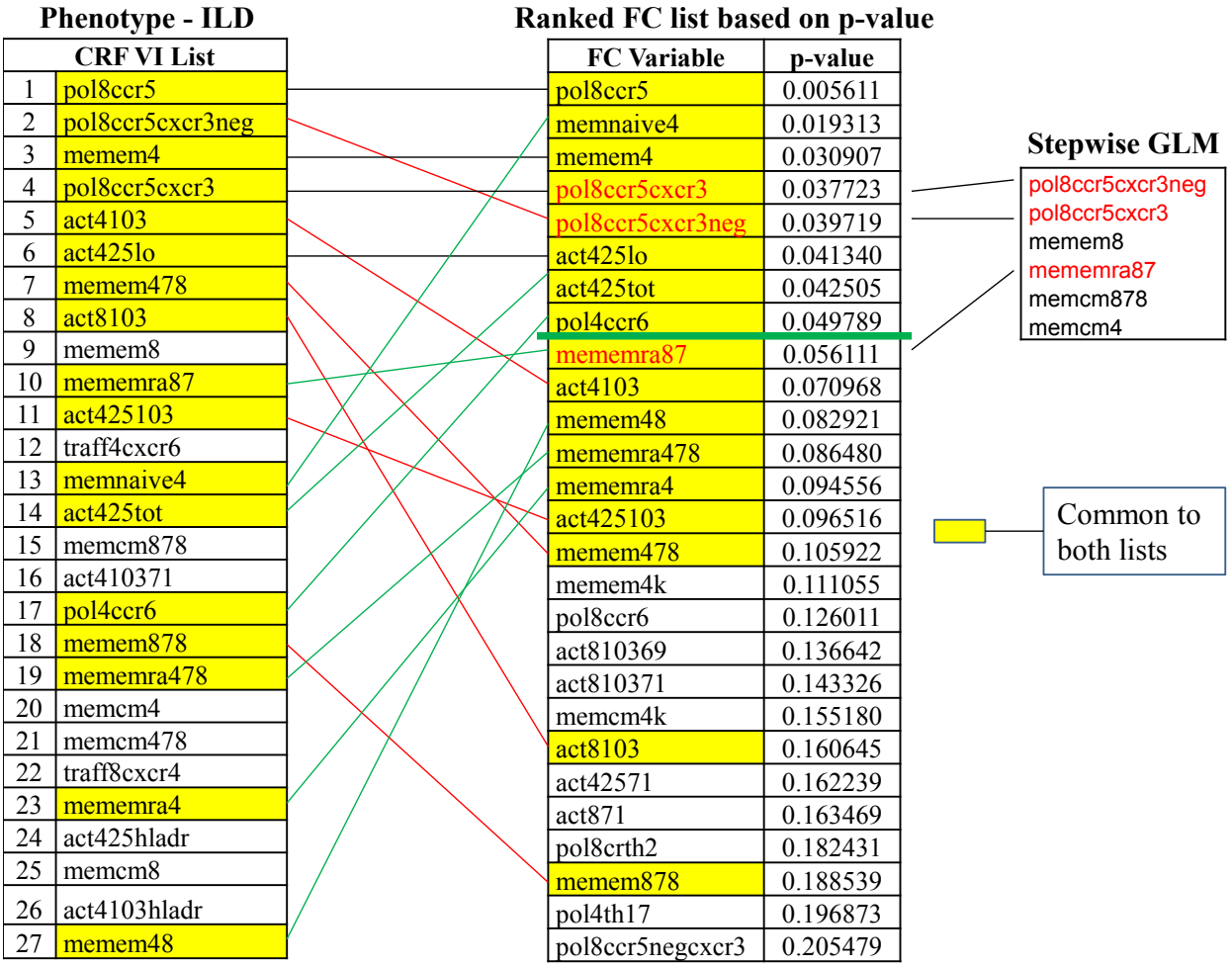


Figure D-36. Comparison between CRF Variable Importance List and P-value based FC list

The FC set size experiment results are shown in Figure D-37.

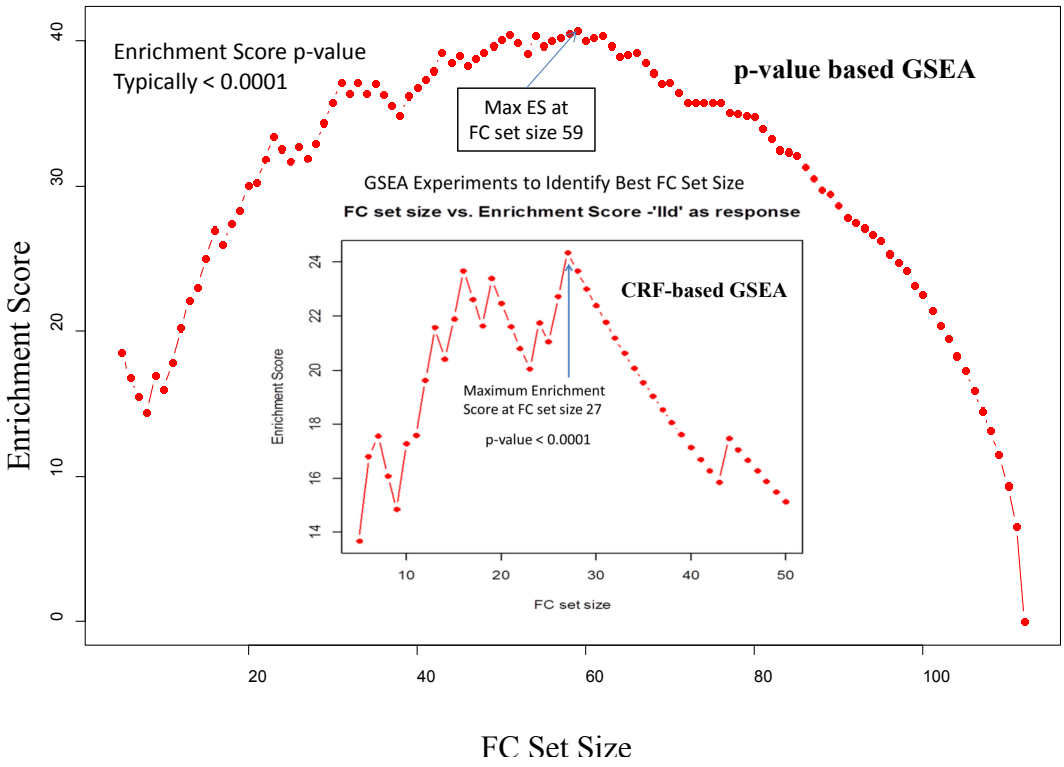


Figure D-37. Comparison between P-value based GSEA and CRF based GSEA in Enrichment Scores (Phenotype is ILD)

When the phenotype is ILD, comparing the 27 most important FC identified by CRF with the sorted list of FC based on p-value from GLM, it can be found that 15 of FC variables appeared in both two lists. However, the order is completely different. There are only 10 FC were statistically significant at significance level 0.05 based on GLM. Several FC variables, such as mememm8 and traff4cxcr6, were highly valued by CRF but not considered to be relatively important based on the above-mentioned p-values. On the other hand, 4 FC (act425hladr, act410371, memcm878 and memcm8) were included in the FC set but their p-values were not statistically significant. The 10 statistically significant FC variables based on GLM were all included in the FC set identified through CRF-GSEA algorithm. In addition to these 10, the FC set also contains variables that had large p-values. This implies that the FC set based on CRF-GSEA algorithm does not simply include all the variables that have generally linear and significant relations with the response. It also contained variables that will contribute to the increase of ES, but whose association with ILD cannot be captured by GLM.

These results suggest that FC variables appear consistently in FC sets constructed in different ways should be emphasized in that they may contain more useful information in explaining the response variables.

D.13-2-1 Comparison based on ROC

I further compared our two methods for creating FC sets by examining ROC performance. The procedure is:

- Using the 15 overlapped variables as the "base" group (Figure D-36, highlighted in yellow), calculate the in-sample ROC and Area Under Curve (AUC) associated with the model, denoted as AUC0
- Add one of the remaining 12 variables, construct the model again with 16 FC variables (15+1), record the AUC, denoted as AUCi
- Calculate the difference between AUCi and AUC0
- repeat for the other remaining variables

When GLM was used to draw statistical inference (it also helps reduce the dimension of the data set in this case), the results shown in Table D-22 were obtained:

Table D-22 Increase in AUC associated with each FC using GLM

GLM	15 overlapped FC		
AUC0	0.8033		
Adding one of the following FC, the AUC becomes:			
	AUC		AUC-AUC0
act8103	0.804476	act8103	0.00113
mememm8	0.804476	mememm8	0.00113
traff4cxcr6	0.800859	traff4cxcr6	-0.00249
memcm878	0.804476	memcm878	0.00113
act410371	0.806058	act410371	0.002712
mememm878	0.804476	mememm878	0.00113
memcm4	0.80425	memcm4	0.000904
memcm478	0.803798	memcm478	0.000452
traff8cxcr4	0.812387	traff8cxcr4	0.009042
act425hladr	0.811483	act425hladr	0.008137
memcm8	0.803798	memcm8	0.000452
act4103hladr	0.803798	act4103hladr	0.000452

It can be seen in Table D-22 that two of the variables have relatively larger contribution to AUC (highlighted in red) but the addition of traff4cxcr6 caused decrease in AUC (highlighted in blue).

When the model used was CRF, we obtained:

Table D-23 Increase in AUC associated with each FC using CRF

CRF	15 overlapped FC		
AUC0_CRF	0.9394		
Adding one of the following FC, the AUC becomes:			
	AUC		AUC-AUC0
act8103	0.9401	act8103	0.000678
mememm8	0.941456	mememm8	0.002034
traff4cxcr6	0.944846	traff4cxcr6	0.005425
memcm878	0.94349	memcm878	0.004069
act410371	0.940778	act410371	0.001356
mememm878	0.939195	mememm878	-0.00023
memcm4	0.941004	memcm4	0.001582
memcm478	0.941004	memcm478	0.001582

traff8cxcr4	0.941682	traff8cxcr4	0.00226
act425hladr	0.941908	act425hladr	0.002486
memcm8	0.943942	memcm8	0.004521
act4103hladr	0.939421	act4103hladr	0

For both models, none of the 12 non-overlapped FC variables contributed substantially in terms of AUC. This is likely due to the result that AUC of the base model was already large, especially when CRF was used. The 12 non-overlapped FC variables selected by CRF were not selected because of their contribution to goodness-of-fit measured by AUC.

D.13-3 Phenotype is Cancer

The analyses including CRF-GSEA and point-wise GLM were repeated but with cancer as the phenotype.

D.13-3-1 Comparison between Ranked lists

Phenotype - Cancer

	CRF VI List
1	act4103hladr
2	act410371
3	traff8cxcr5
4	traff4ccr3
5	act4103
6	act425103
7	act82571
8	act825
9	pol4th2
10	memcm88
11	mememm4
12	traff8ccr10

FC Variable	P-value
act4103hladr	0.000362451
act82571	0.003062673
traff4ccr3	0.003625922
act410371	0.004275550
act4103	0.006118499
act425103	0.010310364
traff8cxcr5	0.023792671
act825	0.024301760
memcm88	0.057465886
traff8cxcr4	0.074113375
pol4th2	0.086708193
traff8ccr10	0.094178295

Stepwise GLM

act4103hladr
traff4ccr3
act825
pol4th2

17th In the CRF VI List

p-value = 0.6075

Common to both lists

Figure D-38. CRF Variable Importance List and P-value based FC List (Cancer is Phenotype)

If "cancer" is the phenotype, almost all of the important FCs identified by CRF are statistically significant in the point-wise GLMs except one - mememm4. Another FC, traff8cxcr4, had a low p-value but was not included in the FC set based on the CRF-GSEA approach. 11 out of 12 variables in the FC set were considered statistically significant by GLM, which suggests that their relations with "cancer" are identified by GLM. These results are consistent with the significant outputs of the stepwise GLM using 12 FC to estimate cancer. Following are the results of the FC set size experiments.

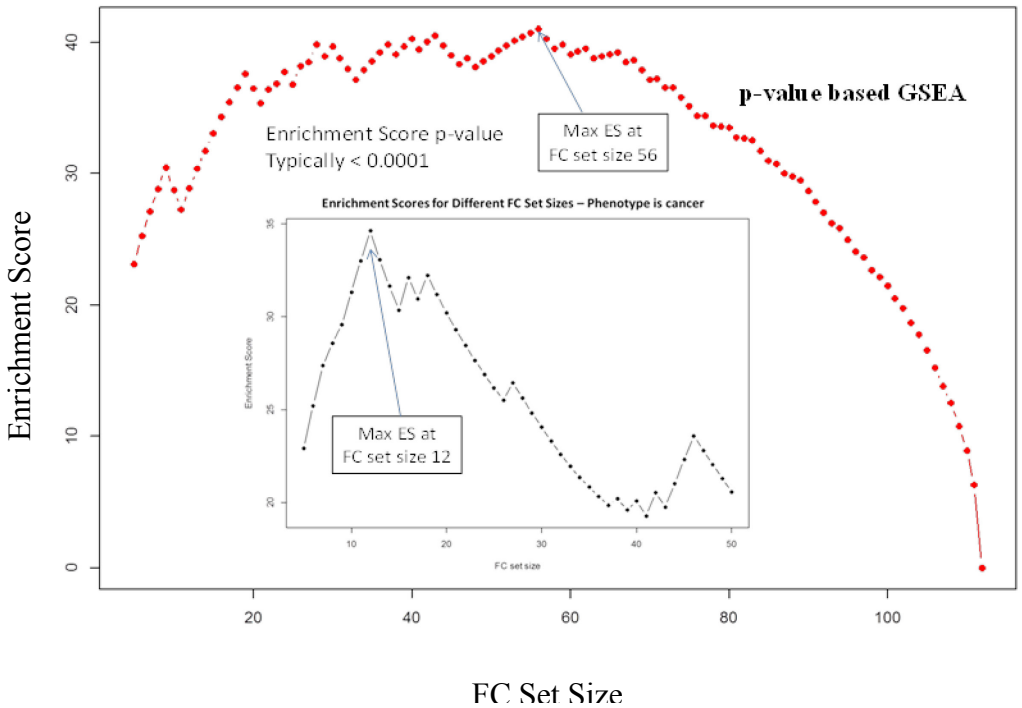


Figure D-39. Comparison between P-value based GSEA and CRF based GSEA in Enrichment Scores (Cancer is Phenotype)

D.14 Other Phenotypes

The associations between SSc and various autoantibodies, especially Anti-topoisomerase I antibodies (ATA, or anti-Scl-70 antibodies) and anti-centromere antibodies (ACA) have been discussed. Dick et al., (2002) found that "*patients with both autoantibodies ATA and ACA often have diffuse scleroderma and show immunogenetic features of both antibody defined subsets of SSc*". Later, ACA was observed to be positive mostly in SSc patients who have CREST symptoms (Miyawaki, et al., 2005) and suggesting the occurrence of limited skin involvement (Castro, Jimenez, & Jefferson, 2010). Lota & Renzoni (2012) discovered that ATA and ACA were linked to pulmonary fibrosis and pulmonary hypertension respectively, and the presence of anti-Scl70 antibodies was believed to be indicator of higher risk for diffuse skin involvement and SSc lung disease (Castro, Jimenez, & Jefferson, 2010). Such connections may lead to new hypotheses on biological pathways and clinical relevance. Two clinical variables: Scl70_ab (Anti-topoisomerase) and ACA (Anti-Centromere Antibodies) were used as phenotype with FC as explanatory variables for CRF-GSEA analysis. Results shown in this section were based on data set IRIS041314.

D.14-1 Scl70_ab

When using Scl70_ab as phenotype variable, the ES peaked at 23.57 which was associated with FC set size 38, and reach its minimum when FC set size is 60 (see Figure D-40). With regard to p-values of permutation tests, they were all statistically significant when FC set size is smaller than 60 at significance level 0.05 (See Figure D-41).

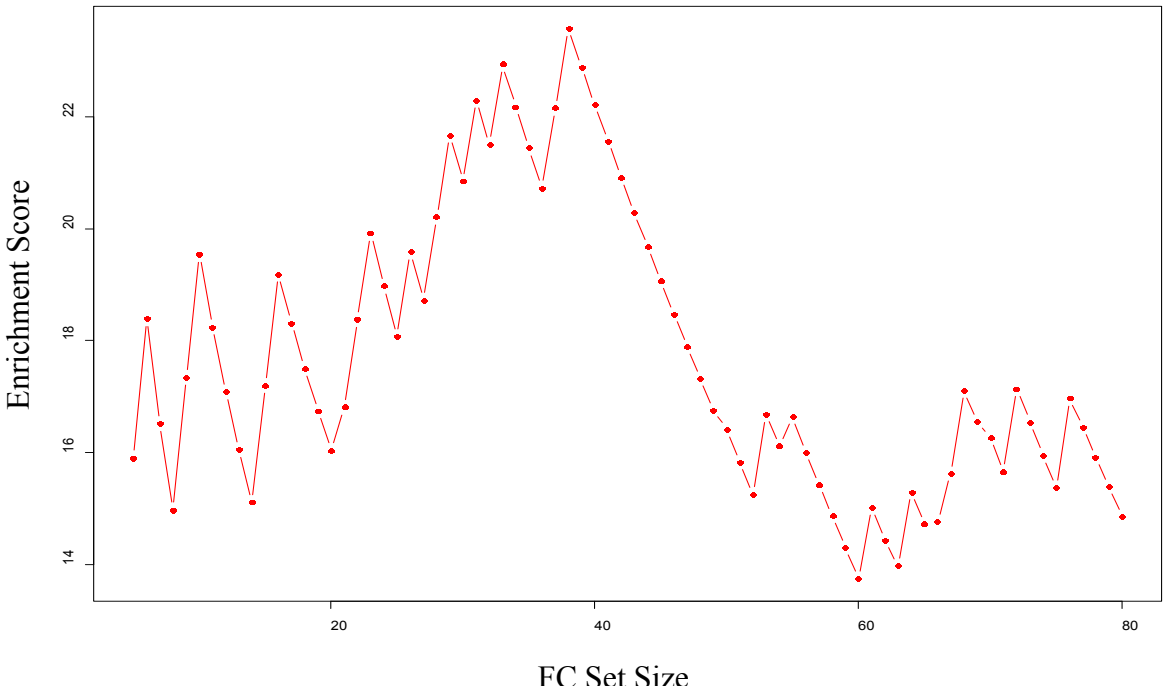


Figure D-40. Enrichment Scores for Different FC Set Sizes (Phenotype is Scl70_ab)

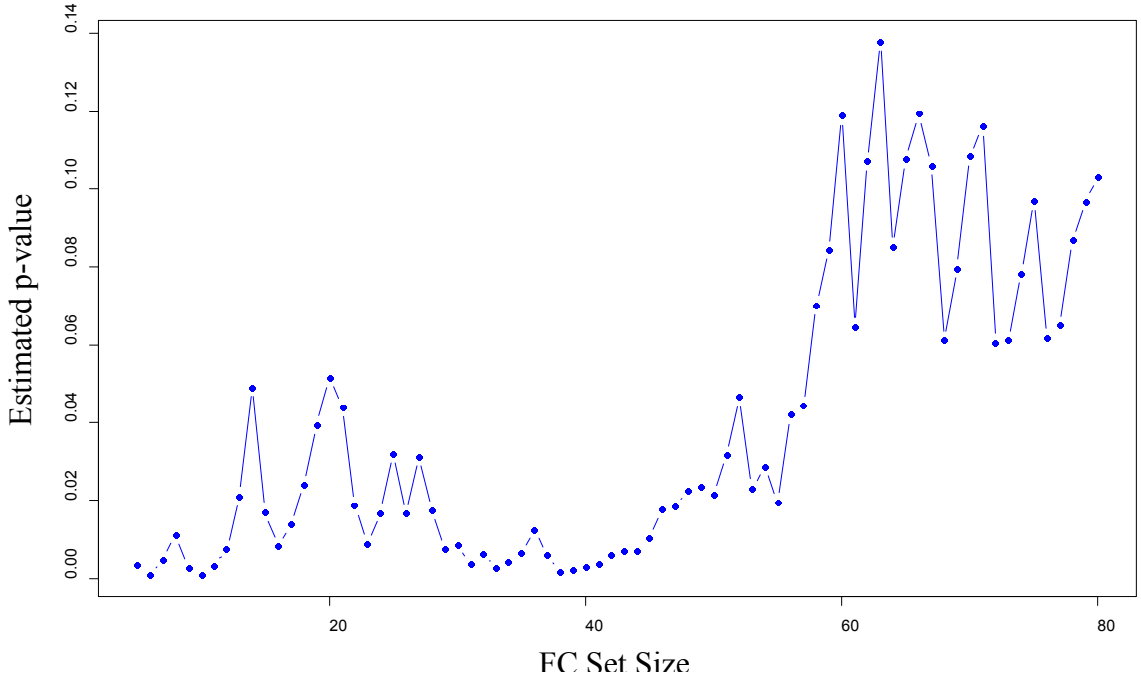


Figure D-41. P-values associated with Enrichment Scores for Different FC Set Sizes (Phenotype is Scl70_ab)

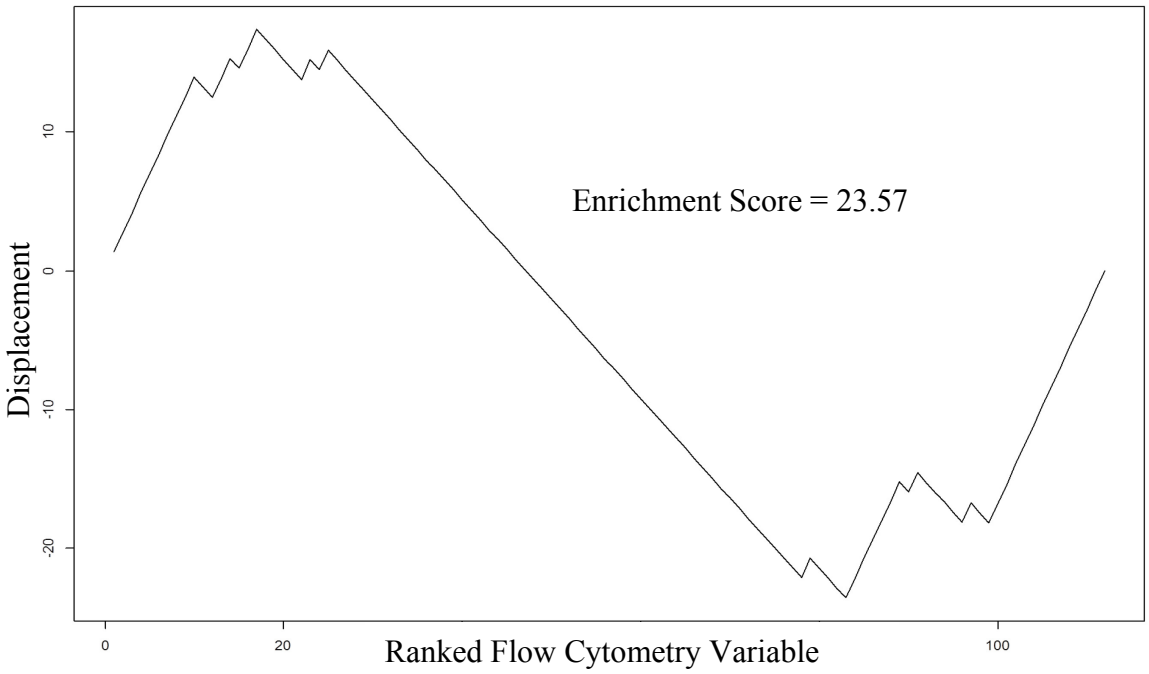


Figure D-42. Random Walk that Results from FC Set Comprised by Top 38 Most Important Variables (Phenotype is Scl70_ab)

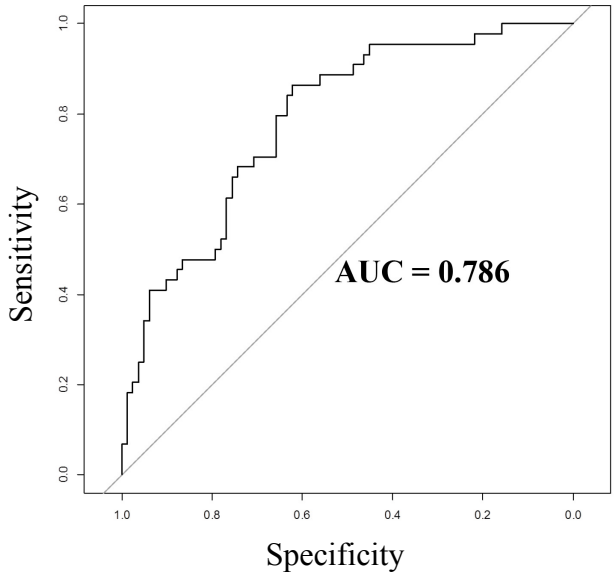


Figure D-43. In-sample ROC Curve for Stepwise GLM (Phenotype is Scl70_ab)

Thirty-eight FC covariates were used as input to fit a logistic binomial linear regression model with the two-directional stepwise variable selection algorithm estimating the binary outcome Scl_70ab (see Table D-24).

Table D-24 Estimated Coefficients of Stepwise GLM (Phenotype is Scl70_ab)

	Estimate	**Std. Error**	**z value**	**Pr(>|z|)**
(Intercept)	-4.84	3.24	-1.5	1.34E-01
act82571	-1.32	0.39	-3.35	**8.15E-04**
memnaive4	-0.05	0.02	-2.96	**3.12E-03**
pol4	0.15	0.07	2.07	**3.83E-02**
cd4cd8ratioLOG	-7.42	3.19	-2.33	2.00E-02
pol8cxcr3	0.03	0.01	1.97	**4.88E-02**
Residual deviance: 129.31 on 120 degrees of freedom				

The Drop-in-Deviance test comparing the current stepwise GLM model with a null model (with intercept only) implies that these five variables are statistically significant (p-value = 2.71×10^{-6}). The p-value of

the goodness-of-fit Test was 0.265, suggesting that there is no evidence that the model is inadequate.

Although a multivariate GLM will be used eventually for statistical inference, it is still interesting to examine which FC variable contributes most to explain the response. To evaluate how much each FC variable in the stepwise GLM contributes to goodness-of-fit, GLMs were fitted using only one FC at a time to estimate the response ILD (point-wise GLMs). Given that five FC variables were included in the stepwise GLM, there are five point-wise GLMs in total (Details in Table D-25). The corresponding AUC values of the in-sample ROC curves based on the point-wise GLMs are presented in Table D-26. Individually, the point-wise GLM with only act82571 fit the data set best in terms of AUC (=0.6757206). The fit was worst with pol4 (=0.5767738) among all the five point-wise GLMs.

Table D-25 Details of Point-wise GLM (Phenotype is Scl70_ab)

Pointwise GLM				
	Estimate	**Std. Error**	**z value**	**Pr(>|z|)**
(Intercept)	0.57	0.38	1.49	1.36E-01
act82571	-1.08	0.33	-3.23	**1.24E-03**
	Estimate	**Std. Error**	**z value**	**Pr(>|z|)**
(Intercept)	0.53	0.61	0.87	3.85E-01
memnaive4	-0.02	0.01	-1.95	5.16E-02
	Estimate	**Std. Error**	**z value**	**Pr(>|z|)**

(Intercept)	1.11	1.07	1.04	3.01E-01
pol4	-0.03	0.02	-1.63	1.03E-01

	Estimate	**Std. Error**	**z value**	**Pr(>|z|)**
(Intercept)	0.25	0.41	0.60	5.51E-01
cd4cd8ratioLOG	-1.69	0.73	-2.30	**2.12E-02**

	Estimate	**Std. Error**	**z value**	**Pr(>|z|)**
(Intercept)	-1.79	0.74	-2.41	**1.58E-02**
pol8cxcr3	0.02	0.01	1.65	9.90E-02

Table D-26 AUC based on pointwise GLM using FC variable from the stepwise GLM (Phenotype is Scl70_ab)

	AUC
act82571	0.676
memnaive4	0.610
pol8cxcr3	0.609
cd4cd8ratioLOG	0.605
pol4	0.577

Next, the cumulative effects of FC variables in the GLM were examined. The five FC were ranked based on their point-wise AUC values. GLM models were fitted adding one FC at one time starting with act82571. The AUC of GLMs gradually increased from approximately 0.6757 to 0.786, shown in Table D-27. Details of these accumulative GLMs are presented in Table D-28.

Table D-27 GLM with accumulative FC sets (Phenotype is Scl70_ab)

	Accumulative AUC
act82571	0.676
memnaive4	0.733
pol8cxcr3	0.761
cd4cd8ratioLOG	0.767
pol4	0.786

Table D-28 Details of Accumulative GLMs (Phenotype is Scl70_ab)

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.57	0.38	1.49	1.36E-01
act82571	-1.08	0.33	-3.23	**1.24E-03**

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	2.45	0.85	2.88	**4.02E-03**
act82571	-1.23	0.36	-3.48	**5.05E-04**
memnaive4	-0.04	0.01	-2.54	**1.11E-02**

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	1.27	1.04	1.22	2.21E-01
act82571	-1.26	0.37	-3.44	**5.84E-04**
memnaive4	-0.04	0.02	-2.84	**4.56E-03**
pol8cxcr3	0.03	0.01	2.03	**4.21E-02**

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	1.56	1.08	1.44	1.50E-01
act82571	-1.21	0.37	-3.29	**9.91E-04**
memnaive4	-0.04	0.02	-2.48	**1.32E-02**
pol8cxcr3	0.03	0.01	2.09	**3.62E-02**
cd4cd8ratioLOG	-1.16	0.84	-1.38	1.67E-01

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-4.84	3.24	-1.50	1.34E-01
act82571	-1.32	0.39	-3.35	**8.15E-04**
memnaive4	-0.05	0.02	-2.96	**3.12E-03**
pol8cxcr3	0.03	0.01	1.97	**4.88E-02**
cd4cd8ratioLOG	-7.42	3.19	-2.33	**2.00E-02**
pol4	0.15	0.07	2.07	**3.83E-02**

D.14-2 ACA

The analyses above were repeated with ACA as the phenotype. ES peaked at 24.9 which was associated with FC set size 6 and non-monotonically decreased arriving at its minimum when FC set size is 79 (see Figure D-44). The p-values of permutation tests were smaller than 0.05 when FC set size is smaller than 25, shown in Figure D-45.

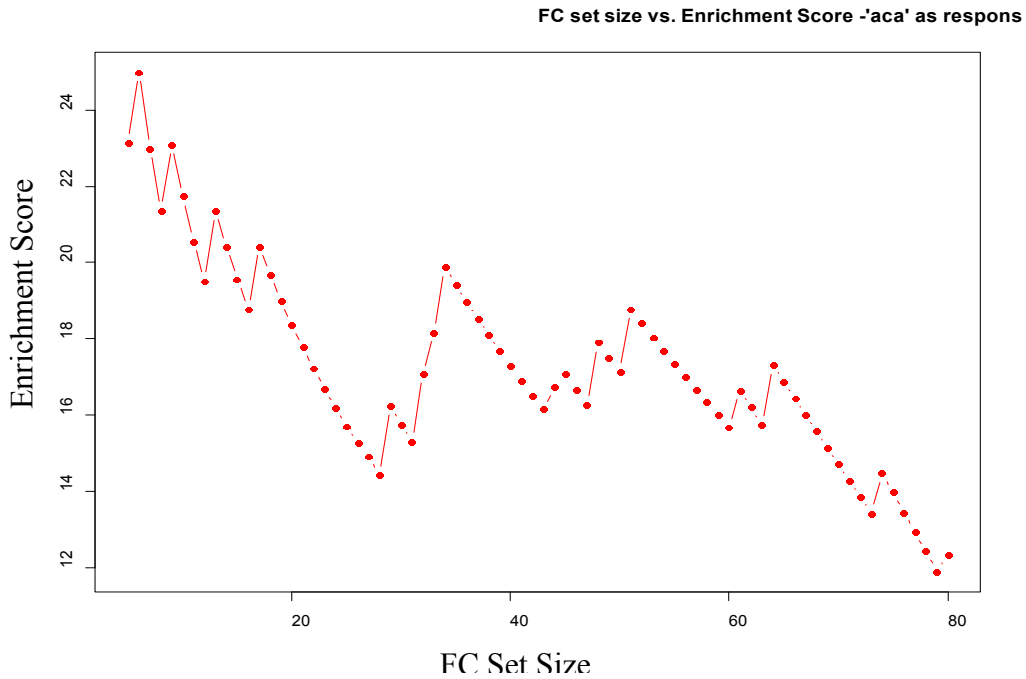


Figure D-44. Enrichment Scores of GSEA for Different FC Set Sizes (Phenotype is ACA)

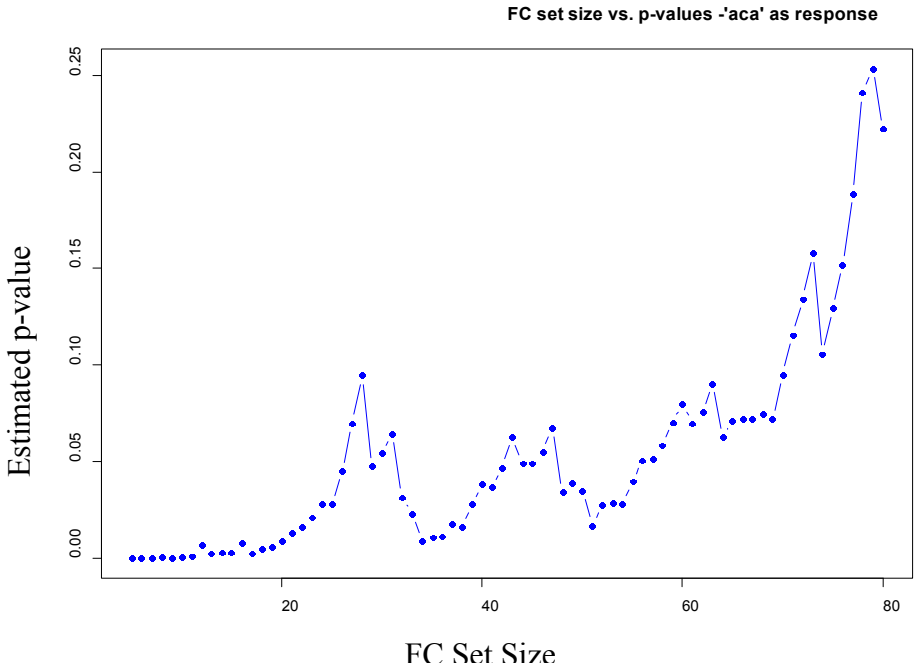


Figure D-45. P-values associated with Enrichment Scores for Different FC Set Sizes (Phenotype is ACA)

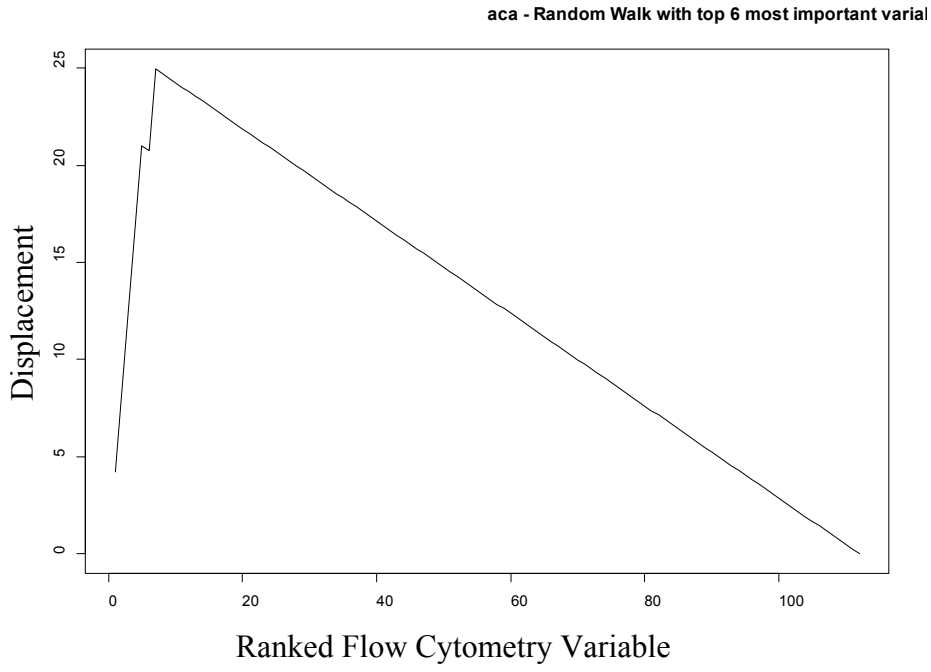


Figure D-46. Random Walk that Results from FC Set Comprised by Top 6 Most Important Variables (Phenotype is ACA)

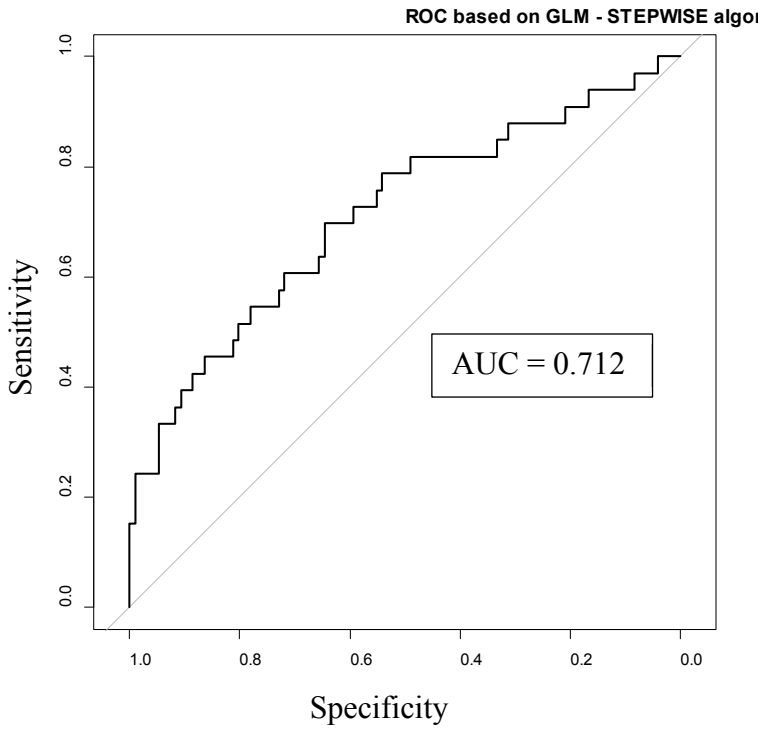


Figure D-47. In-sample ROC Curve for CRF (Phenotype is ACA)

GLM with logistic link function was fitted using input of 6 FC variables identified as the most important through CRF-GSEA algorithm to estimate the binary response ACA. Two-directional stepwise variable selection algorithm was then applied afterward. Only two of the six FCs remained in the final model (see details in Table D-29).

Table D-29 Stepwise GLM with ACA as phenotype

	Estimate	**Std. Error**	**z value**	**Pr(>|z|)**
(Intercept)	-2.22	0.38	-5.83	**5.40E-09**
pol8crth2	0.12	0.05	2.70	**7.03E-03**
mememm88	0.29	0.13	2.18	**2.90E-02**
Residual deviance: 125.84 on 126 degrees of freedom				

The p-value of goodness-of-fit Test is 0.487, which suggests that there is no evidence that the model is inadequate. On the other hand, the p-value of Drop-in-Deviance test is 2.95E-05, which is significantly smaller than 0.05. The null hypothesis that none of the two remaining FC variables (pol8crth2 and memem88) are statistically significant is rejected.

E. DISCUSSION

In this chapter, the data mining methods used will be discussed (§E.1), followed by evaluation of the possibility of using other data mining models (§E.2). In section §E.3, different aspects of the Gene Set Enrichment Analysis (GSEA) algorithm will be focused, including potential improvements that can be made in the future. I will then discuss the randomized filter design procedure (§ E.4) and its potential clinical value in medical practice (§E.5).

Biological meaning and interpretation of the defined FC set associated with ILD will be explored in Section §E.6. Issues regarding standardized Forced Vital Capacity (FVC) that was used to define ILD will be discussed in §E.7, and the importance of phenotype specificity will be highlighted in Section §E.8. This chapter ends with discussion on the two tools used to draw statistical inferences - stepwise GLM and Partial Dependence Plot (PDP) analysis.

E.1 Data Mining

The data mining techniques used (CART, RF, CRF and SVM) cannot outperform the mean-only model by a statistically significant amount in classifying the ILD status of SSc patients given their FC data as predictors.

However, their in-sample training performance showed that these models fit the data set well. CRF was eventually chosen based on its published performance with correlated predictor variables and testing with the IRIS data set. More specifically, the permutation computing scheme for variable importance measure (VIMs) in CRF can provide a "*more fair means of comparison that can help identify the truly relevant predictor variables*" (Strobl et al., 2008).

E.2 Other Data Mining Methods

Characteristics of the data set of interest ultimately determine what type of models to be used. The IRIS data set used in this research is featured by 'large P small N' and highly correlated predictors issue. If the fundamental structure of the data set of interest is altered, especially the correlation structure of the FC variables, other tree-based models such as Random Forests, could possibly perform better. Moreover, other data mining methods such as GLM (that are not suitable for "large P small N" data sets) could prove effective provided that sample size is greatly expanded. But there also exist other approaches that have not yet been evaluated. For example, Bayesian confidence propagation neural network (BCPNN) was shown to be effective in classification in medical sciences (Orre et al., 2000;

Lisboa et al., 2003). Also, the selection of phenotype can significantly affect model performance. We discovered this when using pah_45 as phenotype. The resulting classification and GSEA performance was poor.

In this work, CRF was chosen as a non-parametric method due to its strength in handling highly correlated predictor issue featuring the IRIS data set. CRF is recommended to be used in data sets with similar issues.

E.3 Gene Set Enrichment Analysis

The ES values associated with the FC sets formed from the group of variables at the top CRF variable importance list were consistently statistically significant based on the GSEA permutation test. The corresponding random walks were up-regulated. There exists a unique FC set size associated with an Enrichment Score supremum. Larger FC set sizes did not necessarily lead to greater enrichment scores. If all the variables in the FC set were highly correlated, the corresponding random walks would have even stronger enrichment and the estimated p-value of the GSEA test will be underestimated.

As mentioned earlier, having determined the FC set, two directions analysis were performed – making predictions and drawing statistical inference. Understanding relationship between variables in the FC set will be

useful in drawing statistical inference. The best FC sets that we found consisted of variables from different panels, thus indirectly rejecting the scenario that all the variables in the FC set were highly correlated with each other (variables within the same panels have higher correlation than those from different panels). Occasionally, however, two or more variables in the best FC sets were highly correlated. In this case, confounder effects (multiple predictor variables and the response are pairwise correlated with each other) must be considered (Hennekens et al., 1987; Mickey & Greenland, 1989). Several possible options are available to cope with confounder effects. The first option is to manually remove one of the predictor variables, which is imprudent in that the variable removed may have strong biological or clinical relevance. Another option is to combine predictor variables (forming a composite variable such as the ratio of two variables). A common practice in epidemiology is to stratify the data by one of the confounders and then give an overall estimate combining individual strata-specific predictions (Hoggart et al., 2003). This method works well for categorical predictor variables, but certain thresholds must be established for continuous variables depending on the density and range of the variable.

E.3-1 GSEA Robustness

In our "top-down" analyses (form FC sets beginning at the top of the ranked correlation list in GSEA), all random walks were up-regulated with strong enrichment scores. To test the robustness of this result, we instead constructed FC sets beginning at some variable considerably down the ranked list, and in some cases, formed FC sets beginning at the bottom of the ranked list. Results in Section §D.4-1 indicate that the regulation direction of the random walks were robust. Significant enrichment structure of random walk will always result in larger ES which was associated with statistically significant p-value in permutation test. GSEA performance was also robust to the number of shuffling in the permutation test (see Section §D.4-2).

E.3-2 The GSEA Ranked List

Thus far, the ranked lists of the GSEA test were based on sorted correlation coefficients. This is the original and most common approach (Subramanian et al., 2008), but others exist, for example, the absolute value of correlation coefficients could serve as the foundation of ranked list. It follows in that case that all random walks would be up-regulated (see Appendix §I.4). This could result from the situation that all the FCs with high correlation coefficient clustered toward the same area in the ranked list.

It is another way of constructing ranked list and could possibly be applied in condition that the sign of correlation coefficient does not matter.

E.3-3 GSEA – FC Set Determination

In the original genomics applications, gene sets for GSEA were formed through biological insight (our rough equivalent might be effector's pathways). A reasonable alternative here is to define FC sets based on FC variables considered to be potentially important biomarkers. Two such FC sets (CD4-based and CD8-based) were created and tested through GSEA (see Section §D.11). Performance was poor when compared with the hybrid CRF-GSEA approach, but this exercise was very limited in scope.

In the future, if certain pathway of autoimmune disorder is hypothesized, the group of variables involved in the pathway can become input of the GSEA-FC set and tested with permutation test, in order to examine the statistical significance level of the assumed pathway.

E.3-4 GSEA – Permutation Test

Gene Set Analysis (GSA) in general can be divided into two major schools based on the permutation schemes used in their statistical tests: class label randomization and gene randomization (Luo et al., 2009). The GSEA

algorithm in this research, SAFE (Barry et al., 2005) and SAM- CS (Dinu et al., 2007) belongs to the first category, while PAGE (Kim & Volsky, 2005), T-profiler (Boorsma et al., 2005) and Random Set (Newton et al., 2007) are classified as gene-randomization. An important factor in assessing the suitable approach is sample size. Our data set contains 119 patients. This sample size may not "*allow rigorous evaluation of significance levels by permuting the class labels*" (Subramanian et al., 2008), however, our results were indeed robust with respect to the number of permutations used to estimate p-values (see Section §D.4). The alternative approach (gene-randomization) is to permute FC variables, but this will lead to non-conservative significance levels, i.e., smaller p-values resulting in more false positives, because this approach does not account for stimulus variable correlation (Subramanian et al., 2008).

In sum, the GSEA algorithm used in this research can help determine a group of important variables with regard to certain phenotype variable. Its performance is robust. With certain modification in the three components of GSEA such as FC set and ranked list, this algorithm can be applied in many more fields along with other variable settings.

E.4 Randomized Filter Design

Randomized filter design is a novel mathematical tool used for ILD classification. Its non-parametric nature allows application in many other settings. In the current design, the predictor variables are continuous and the response binary, however, with fairly simple modification, response, for example, could be categorical.

While training filters performed well when classifying the entire data set (Overall Misclassified Rate (OMR) = 0.1898), the added step of pre-partitioning patients using CART (and finding best training filters specific to each CART group) significantly improved filter performance. The splitting criterion used in CART appears to be a good starting point for identifying subpopulation of SSc patients. Our training filter experiments consistently resulted in best filters having three to six components with five most often. We currently have no mathematical or biological explanation why this should be the case.

Predictive accuracy of this screening tool was promising in this study. However, its computation was rather expensive. It will be rather useful if this tool can be modified such that less computing time is needed in the future.

E.5 Clinical Value of Screening Tool

Validation was successful (Section §D.6) with a correct classification rate of 82.5% for the entire validation data set (40 patients), increasing to 95% with CART pre-partitioning. There exists a reasonable balance between training and validation error. Given FC data as input, the prediction of patient ILD status with our filter approach is essentially instantaneous and can be accomplished through a variety of software implementations. Ultimately this instrument should facilitate a refined stratification of SSc patients into clinically relevant subsets at the time of diagnosis and subsequently during the course of the disease, preventing bad outcomes from disease progression or unnecessary treatment side effects. This role could involve a scenario in which an SSc patient passes the presumptive test for ILD, but the filter indicates that their flow cytometry (FC) profile is consistent with ILD. In such a case, a physician might: 1) increase frequency of testing to detect early development of ILD; 2) implement more sophisticated diagnostic procedures (e.g., high resolution chest CT scan) to confirm the presence of ILD; and 3) consider prophylactic disease modifying treatments (e.g., cyclophosphamide, corticosteroids, interferons (White, 2003)).

We would expect that as more data becomes available, filter performance will continue to improve. This suggests another potentially important role for our approach in better understanding the progression of disease. We posit the scenario in which FC profile characteristics may change with disease progression and that these changes could be captured – reflected in changes in filter design and performance. Procedurally, we contemplate a procedure in which certain FC variables and their expressions are used as a basis for partitioning patients into disease progression states, with corresponding state-specific filter designs.

The potential clinical value of the screening tool is encouraging.

E.6 Biological Interpretation

In FC27 (see Section §D.3, Table D-4), the two FC variables (pol8ccr5cxcr3neg, pol8ccr5cxcr3) identify Type 1 helper (Th1) polarized CD8 T cells. Informed by our medical colleagues, the first (lacking CXCR3) is “protective”; the second (CXCR3) is a “risk factor” for ILD. CXCR3 is a chemokine receptor which has been shown to direct inflammatory cells inside target tissue and drives acute inflammation (synovial tissue in rheumatoid arthritis, liver in autoimmune hepatitis, etc.). The variables mememem8, mememra87, memcm878 and memcm4 belong to the T cell

memory subset. It appears that ILD status is associated with a shift of the CD8 T cells towards the activated effector memory/terminally differentiated state. This is in keeping with the pro-inflammatory polarized status observed.

E.7 Issues Regarding FVCstpp

FVCstpp is imprecise (e.g., poorly performed maneuvers) and uncertain (variable from test to test) (Alhamad et al., 2001; Enright, 2003; Hegewald & Crapo, 2010; Miller, 2005; Pierce, 2005; Petty & Enright, 2003; Hankinson, 1999) which presents several questions involving: (1) the accuracy of the current FVCstpp cutoff in accurately assessing ILD status; (2) the effect of the cutoff in filter performance; and (3) a corollary, whether the patients that are misclassified in training and validation have FVCstpp measurements in the so-called "gray area". Further to (3) Table E-1 shows FVCstpp values from the IRIS data set that are close to the ILD cutoff (80 +/- 5%). The results are inconclusive: nine patients in this group of twenty-two are among the misclassified patients in training and validation. The entries shaded in blue are training misclassifications; those in red are for validation misclassifications.

Table E-1 FVCstpp values from the IRIS data set that are close to the ILD cutoff (80 +/- 5%)

FVCstpp	
79.2	82.99
77.7	82.66
78.29	82.7
75.93	82.3
77.55	83.53
75.33	82.9
79.95	81.41
75.53	81.22
75.38	81.5
76.78	84.05
77.02	80.94

Issue (1) points to the need for High Resolution Computing Tomography (HRCT) for interstitial lung disease (ILD) confirmation (Moore et al., 2013; Pandey et al., 2010; Zompatori et al., 2013).

E.8 Phenotype Specificity (ILD vs. Cancer)

Using different phenotypes as response variable resulted in distinctive outcomes including the best FC sets, direction of random walk, ES and model performance. In this study, the results associated with "cancer" as phenotype were found to be more striking than other phenotypes (see Section §D.9).

GSEA performance of the cancer FC set with ILD as the phenotype and vice-versa were examined. The results were not statistically significant – small ES and large estimated p-value. Although all the patients of interest were diagnosed as having SSc, yet the two phenotypes (cancer and ILD) are discovered to be related to different biomarkers and therefore distinct FC expressions. For example, as mentioned in section *A.2 Systemic Scleorsis*, previous studies have shown that increased frequency of circulating T cells exhibiting a "polarized" phenotype –the Polarization panel in the IRIS data set- are significantly associated in SSc patients with the presence of pulmonary fibrosis and lung disease progression (Boin et al 2008; Truchetet et al. 2010). However, it remains unknown whether such association still holds when the phenotype is cancer from the perspective of medical sciences.

Regarding methodology, the inferior performance of the ILD-Cancer GSEA intercomparison experiments greatly suggests that the two components of GSEA algorithm (ranked list and FC set) should be always associated with the same phenotype in order to obtain reasonable and interpretable results, unless there is a significant number of overlapping "genes" across different sets. Subramanian et al. (2008) performed similar experiments using the Boston gene set in the Michigan lung cancer data sets, and the Michigan gene set in the Boston data set, in an attempt to draw

biological insight by examining overlapping genes. In our research, a similar analysis indicates that very few FC variables were overlapping between the ILD set and Cancer set (see Section §D.10).

E.9 Statistical Inference

E.9-1 Stepwise GLM

When the phenotype was ILD, the p-value of the Goodness-of-fit Test for the stepwise GLM was approximately 0.15, which suggests that there was no evidence that the model is inadequate. However, this p-value was not large enough (close to 1) to be considered as statistically significant, therefore statistical inference from this model should be made cautiously, particularly the interpretation of the coefficients of each FC in the model. This under-fit eventually becomes one of the reasons why the outputs of this model are inconsistent with those of the Partial Dependence Plot for CRF which will be discussed in the next section. However, several Drop-in-Deviance Tests regarding this stepwise GLM strongly suggest that the six FC variables and the three clinical variables (type, scl70_ab, dd1symptom_y) are of significant importance in estimating the response variable ILD. In addition, none of the four diagnostic statistics indicate that there were

isolated cases that have either high leverages or outliers regarding the fitted GLM.

If using cancer as the phenotype, the stepwise GLM fitted the data set very well (p-value of the Goodness-of-fit Test was almost 1). All 4 FC variables were statistically significant at 5% level. Among these variables, pol4th2, or CD3+/CD4+/CD8-/CXCR3-/CCR4+/CCR6-, was negatively associated with the odds of having cancer and especially important in differentiating patients with cancer from those without. The other three variables are positively associated with the odds of having cancer, among which variable act4103hladr had the strongest association. These results were consistent with the PDP outputs below. In addition, based on Cook's distance, standardized residuals and Studentized residuals, two patients (ID:2202, 3083) were identified as isolated observations. These two cases heavily influenced the coefficients of variable act4103hladr (from 3.05 to 5.52). Lastly, without CRF-GSEA procedure, the fitted stepwise GLM based on the full data set did not fit the data set well – none of the remaining FC variables were statistically significant.

GLM with a logistic link function assuming binomial distribution of the responses followed by stepwise algorithm has been widely used in

explaining dichotomous response variable using multiple continuous or categorical predictor variables (Hirzel et al., 2001; Ethier et al., 2008; Karlsson et al., 2010;). It was used in this research after the data dimension was reduced and it turned out to be rather useful in terms of drawing statistical inferences.

Statistical inferences directly drawn from the stepwise GLM regarding coefficients of the FC variables should be of caution. It is true that stepwise GLM may have certain drawbacks such as potential exclusion of important variables but inclusion of noise variables (Derksen and Keselman, 1992) and Type 1 error inflation issues (Mundry & Nunn, 2009). However, in this work, the starting GLM was a full model, i.e., including all predictor variables at the beginning, therefore no exclusion of variables was made. Also, the statistical significance of the group of important variables was determined by the Drop-in-Deviance tests, instead of simply relying on significance level of Z-score Wald tests on an individual basis. The p-values of Z-score Wald tests were not used as the basis of determining statistical significance of multiple variables as a group, because these individual p-values tend to change when the number of input variables varies. Note that before the GLM had been fitted, a GSEA test was performed to guarantee that the group of FC variables was statistically significant as an entity. Lastly, the intention of

applying stepwise GLM is to present one possible approach for extending the results from CRF-GSEA and hopefully shed lights on future medical research direction.

One interesting finding was that without CRF-GSEA procedure, the stepwise GLM based on original full data set had low level of goodness-of-fit; but with CRF-GSEA, its fit was considerably improved. This result may originate from that stepwise GLM itself was insufficient to identify important variables, especially when the data set suffers from "Large P small N" issue. Another related explanation could be the predictor correlation issue which in this case was handled to some degree by the hypothesis testing structure in CRF.

E.9-2 Partial Dependence Analysis

For ILD as phenotype, as the values of FC variables memcm878 and memcm4 increased, the probability of having ILD became larger. The remaining four FC variables had the opposite trend with regard to the association with the probability of having ILD. Comparing these results with the estimated coefficients of stepwise GLM, only a moderate level of consistency can be found regarding whether the FC variables exert positive

or negative impacts on the probability of having ILD. The most likely reason for this result is that stepwise GLM did not fit the data set very well.

When cancer was the phenotype, as three of four FC variables increased, the estimated probability of having cancer in SSc patients also increased when using PDP. The exception was pol4th2. The 3D PDP with pol4th2, act825 and cancer probability indicated that cancer probability grew with a combination of decreasing pol4th2 and increasing act825. These PDP results are consistent with the stepwise GLM outputs. These PDP results strongly suggest that act4103hladr, traff4ccr3, act825 are positively associated with the probability of having cancer; its association with pol4th2 is negative.

Partial dependence plots are useful in diagnosing the dependence of a response variable on the joint values of stimulus variables (Hastie et al., 2009). However, visualization with high dimensions can be difficult to visualize and interpret. PDP results are useful for yielding insights into the marginal effect of individual covariates, especially when the model of interest is classification or regression trees models (Elith et al., 2008).

F. CONCLUSIONS

1. The combination of CRF and GSEA algorithm can determine statistically significant FC sets associated with phenotypes ILD and cancer. This approach is robust in terms of statistical significance of the GSEA permutation test.

2. Randomized filter design is an effective approach in differentiating patients with ILD from those without.

3. Stepwise GLM in conjunction with Partial Dependence Plots could be useful in drawing statistical inference from the outputs of CRF-GSEA algorithm.

4. The specificity of phenotypes will directly impact the performance of GSEA algorithm.

5. The construction of FC sets can be based on different methods.

6. When ILD is the phenotype, variables included in p-value based FC set are the most important ones for the following reasons: a. results based on CRF VI list, filter set list and p-value based FC set were more statistically significant than bio-informed FC set; b. all the variables included in the p-value based FC set were included in CRF VI list and filter set list. These

variables are: pol8ccr5, memnaive4, mememm4, pol8ccr5cxcr3, pol4ccr6, pol8ccr5cxcr3neg, act425lo, act425tot, mememra87, act4103, memem48, mememra478, mememra4, act425103 and memem478.

If Cancer is the phenotype, the variables listed below are attached with more importance because they showed up in CRF VI list, p-value based FC set and T Test based FC set. These variables are: act4103hladr, act410371, traff4ccr3, act4103, act425103, act82571 and memcm88.

Details of the best FC set identified by different methods can be found in Appendix I.2.

G. FUTURE RESEARCH

The major focus thus far has been on ILD and cancer, driven by the seriousness of these diseases, but there are other phenotypes associated with SSc that too have important clinical outcomes, for example, elevated Right Ventricular Systolic Pressure (rsvp) and depressed Spirometry Diffusion Capacity (dlco). Principal Components Analysis was eventually not used because of the limitation of interpretability. Other factor analysis such as Sparse Factor Analysis (SPA) (Carvalho et al., 2008; Engelhardt & Stephens, 2010) that may be able to overcome the interpretability issue can be explored.

Data mining approaches followed by GSEA can be applied to these phenotypes. For those phenotypes that share common or highly relevant biological pathways, a GSEA intercomparison could be used to examine the overlaps in order to yield biological insights. Other data mining methods could be evaluated in addition to the five (CART, RF, CRF, SVM, GLM) we considered. Different antibodies associated with SSc can also be used as the phenotype. More ways of determining significant FC sets can be explored.

In the future, if certain pathway of autoimmune disorder is hypothesized, the group of variables involved in the pathway can become

input of the GSEA-FC set and tested with permutation test, in order to examine the statistical significance level of the assumed pathway.

The randomized screening tool can be modified such that less computing time is needed.

H. REFERENCES

Alexander, HK and LM Wahl. 2011. “Self-Tolerance and Autoimmunity in a Regulatory T cell Model” , *Bulletin of Mathematical Biology* 73 (1) (January): 33–71. doi:10.1007/s11538-010-9519-2. http://www.ncbi.nlm.nih.gov/pubmed/20195912.

Alhamad, E. H., Lynch III, J. P., & Martinez, F. J. (2001). Pulmonary function tests in interstitial lung disease: what role do they have?. Clinics in chest medicine, 22(4), 715-750.

Akaike, H. (1981). Citation Classic - a New Look at the Statistical-Model Identification. Current Contents/Engineering Technology & Applied Sciences(51), 22-22.

Ashihara Y, KY, RM Nakamura. 2011. “Immunoassay and Immunochemistry”, in: McPherson RA, Pincus MR, eds., Henry's Clinical Diagnosis and Management by Laboratory Methods, 22nd ed. Philadelphia, Pa: Saunders Elsevier; 2011:chap 44.

Baltcheva, I. 2010. “Mathematical Modeling of T cell Experimental Data”, Ph.D. thesis, Ecole Polytechnique Federale de Lausanne, Lausanne, Switzerland.

Barry, W. T., Nobel, A. B., & Wright, F. A. (2005). Significance analysis of functional categories in gene expression studies: a structured permutation approach. Bioinformatics, 21(9), 1943-1949.

Baxt, W. G. (1992). Analysis of the Clinical-Variables Driving Decision in an Artificial Neural Network Trained to Identify the Presence of Myocardial-Infarction. Annals of Emergency Medicine, 21(12), 1439-1444. doi: Doi 10.1016/S0196-0644(05)80056-3

Belsley, D. A., Kuh, E. and Welsch, R. E. (1980) Regression Diagnostics. New York: Wiley.

Bishop, C. M. (1995). Neural networks for pattern recognition.

Boin, F, U De Fanis, SJ Bartlett, FM Wigley, A Rosen and V Casolaro. 2008. “T cell Polarization Identifies Distinct Clinical Phenotypes in Scleroderma Lung Disease”, *Arthritis and Rheumatism* 58 (4) (April): 1165–74. doi:10.1002/art.23406. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2662772&tool=pmcentrez&rendertype=abstract.

Boin, F and A Rosen. 2007. “Autoimmunity in Systemic Sclerosis: Current Concepts”, *Current Rheumatology Reports* 9 (2) (May): 165–72. http://www.ncbi.nlm.nih.gov /pubmed/17502048.

Boorsma, A., Foat, B. C., Vis, D., Klis, F., & Bussemaker, H. J. (2005). T-profiler: scoring the activity of predefined groups of genes using gene expression data. Nucleic acids research, 33(suppl 2), W592-W595.

Breiman, L.1984. Classification and Regression Trees, New York : Chapman & Hall.

Breiman, L. 2001. “Random Forests”, *Machine Learning* 45 (1) (January): 5–32. doi:10.1186/1478-7954-9-29. http://www.ncbi.nlm.nih.gov/pubmed/21849043.

Burroughs, NJ, BMP Mendes de Oliveira and AA Pinto. 2006. “Regulatory T cell Adjustment of Quorum Growth Thresholds and the Control of Local Immune Responses”, *Journal of Theoretical Biology* 241 (1): 134–41. doi:10.1016/j.jtbi.2005.11.010. http://www.ncbi.nlm.nih.gov /pubmed/16403532.

Burroughs, N. J., BMPM Oliveira, AA Pinto and HJT Sequeira. 2008. "Sensibility of the quorum growth thresholds controlling local immune responses", *Mathematical and Computer Modelling*, 47(7), 714-725.

Carneiro, J, K Leon, I Caramalho, C van den Dool, R Gardner, V Oliveira, M-L Bergman, et al. 2007. "When Three Is Not a Crowd: a Crossregulation Model of the Dynamics and Repertoire Selection of Regulatory CD4+ T cells", *Immunological Reviews* 216 (April): 48–68. doi:10.1111/j.1600-065X.2007.00487.x. http://www.ncbi.nlm.nih.gov/pubmed/17367334.

Carneiro, J, T Paixão, D Milutinovic, J Sousa, K Leon, R Gardner and J Faro. 2005. "Immunological Self-Tolerance: Lessons from Mathematical Modeling", *Journal of Computational and Applied Mathematics* 184 (1) (December): 77–100. doi:10.1016/j.cam.2004.10.025. http://linkinghub .elsevier.com/retrieve/pii/S0377042705000713.

Carvalho, C. M., Chang, J., Lucas, J. E., Nevins, J. R., Wang, Q., & West, M. (2008). High-Dimensional Sparse Factor Modeling: Applications in Gene Expression Genomics. Journal of the American Statistical Association, 103(484), 1438-1456. doi: 10.1198/016214508000000869

Casciola-Rosen, LA, G Anhalt and A Rosen. 1994. “Autoantigens Targeted in Systemic Lupus Erythematosus Are Clustered in Two Populations of Surface Structures on Apoptotic Keratinocytes”, *The Journal of Experimental Medicine* 179 (4) (April 1): 1317–30. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2191465&tool=pmcentrez&rendertype=abstract.

Castro, S. V, Jimenez, S. A., & Jefferson, T. (2010). Biomarkers in systemic sclerosis R eview, *4*, 133–147.

Cawley, GC and NLC Talbot. 2010. “On Over-fitting in Model Selection and Subsequent Selection Bias in Performance Evaluation”, *Journal of Machine Learning Research* 11, 2079-2107

Chan, C, RI Lechler and JT George. 2004. “Tolerance Mechanisms and Recent Progress”, *Transplantation Proceedings* 36 (2 Suppl) (March): 561S–569S. doi:10.1016/j.transproceed. 2004.01.019. http://www.ncbi.nlm.nih.gov/pubmed/15041406.

Chao, DL, MP Davenport, S Forrest and AS Perelson. 2004. “A Stochastic Model of Cytotoxic T cell Responses”, *Journal of Theoretical Biology* 228 (2) (May 21): 227–40. doi:10.1016/j.jtbi.2003.12.011. http://www.ncbi.nlm.nih.gov/pubmed/15094017.

Chifflot, H, B Fautrel, C Sordet, E Chatelus and J Sibilia. 2008. “Incidence and Prevalence of Systemic Sclerosis: a Systematic Literature Review”, *Seminars in Arthritis and Rheumatism* 37 (4) (February): 223–35. doi:10.1016/j.semarthrit.2007.05.003. http://www.ncbi.nlm.nih.gov/pubmed/17692364.

Child, D. (2006). The Essentials of Factor Analysis Retrieved from http://books.google.com/books?hl=en&lr=&id=rQ2vdJgohH0C&oi=fnd&pg=PR7&dq=The+essentials+of+factor+analysis&ots=mW2lNScM7J&sig=13ryj2qBedWCGah3jdftKpmASrg#v=onepage&q=The%20essentials%20of%20factor%20analysis&f=false

Christmann, RB, E Hayes, S Pendergrass, C Padilla, G Farina, AJ Affandi, ML Whitfield, HW Farber and R Lafyatis. 2011. “Interferon and Alternative Activation of Monocyte/macrophages in Systemic Sclerosis-Associated Pulmonary Arterial Hypertension”, *Arthritis and Rheumatism* 63 (6) (June): 1718–28. doi:10.1002/art.30318. http://www.ncbi.nlm.nih.gov/pubmed/21425123.

Chung, L and PJ Utz, 2004. “Antibodies in Scleroderma: Direct Pathogenicity and Phenotypic Associations”, *Current Rheumatology Reports* 2004, 6:156–163

Cook, N. R. (2008). Statistical evaluation of prognostic versus diagnostic models: Beyond the ROC curve. Clinical Chemistry, 54(1), 17-23. doi: DOI 10.1373/clinchem.2007.096529

Cook, R. D. and Weisberg, S. (1982) Residuals and Influence in Regression. London: Chapman and Hall.

Cracowski, J-L, C Marpeau, PH Carpentier, B Imbert, M Hunt, F Stanke-Labesque and G Bessard. 2001. "Enhanced in vivo lipid peroxidation in scleroderma spectrum disorders", *Arthritis & Rheumatism*, 44: 1143–1148. doi: 10.1002/1529-0131(200105)44:5<1143::AID-ANR196>3.0.CO;2-#

De Cruz, S and D Ross. 2013. "Lung transplantation in patients with scleroderma", *Curr Opin Rheumatol*. 2013 Nov;25(6):714-8. doi: 10.1097/01.bor.0000434670.39773.a8.

Derksen, S., and H. J. Keselman. 1992. Backward, forward and stepwise automated subset selection algorithms: frequency of obtaining authentic and noise variables. British Journal of Mathematical and Statistical Psychology 45:265–282.

Dick, T., Mierau, R., Bartz-Bazzanella, P., Alavi, M., Stoyanova-Scholz, M., Kindler, J., & Genth, E. (2002). Coexistence of antitopoisomerase I and

anticentromere antibodies in patients with systemic sclerosis. Annals of the Rheumatic Diseases, 61(2), 121–7.

Dinu, I., Potter, J. D., Mueller, T., Liu, Q., Adewale, A. J., Jhangri, G. S., ... & Yasui, Y. (2007). Improving gene set analysis of microarray data by SAM-GS. BMC bioinformatics, 8(1), 242.

Distler, JHW, Y Allanore, J Avouac, R Giacomelli, S Guiducci, F Moritz, A Akhmetshina, et al. 2009. "EULAR Scleroderma Trials and Research Group Statement and Recommendations on Endothelial Precursor Cells", *Annals of the Rheumatic Diseases* 68 (2) (February): 163–8. doi:10.1136/ard.2008.091918. http://www.ncbi.nlm.nih.gov/pubmed/18653485.

Dobbin, K. K., & Simon, R. M. (2011). Optimally splitting cases for training and testing high dimensional classifiers. Bmc Medical Genomics, 4. doi: 10.1186/1755-8794-4-31

Domingos, P., & Pazzani, M. (1997). On the optimality of the simple Bayesian classifier under zero-one loss. Machine Learning, 29(2-3), 103-130. doi: Doi 10.1023/A:1007413511361

Dreiseitl, S., & Ohno-Machado, L. (2002). Logistic regression and artificial neural network classification models: a methodology review. Journal of

Biomedical Informatics, 35(5-6), 352-359. doi: Doi 10.1016/S1532-0464(03)00034-0

Duan, H, J Fleming, DK Pritchard, LM Amon, J Xue, HA Arnett, G Chen, et al. 2008. "Combined Analysis of Monocyte and Lymphocyte Messenger RNA Expression with Serum Protein Profiles in Patients with Scleroderma", *Arthritis and Rheumatism* 58 (5) (May): 1465–74. doi:10.1002/art.23451. http://www.ncbi.nlm.nih.gov/pubmed/18438864.

Dunn, O. J. (1961). Multiple Comparisons among Means. Journal of the American Statistical Association, 56(293), 52-&. doi: Doi 10.2307/2282330

Elith, J., Leathwick, J. R., & Hastie, T. (2008). A working guide to boosted regression trees. Journal of Animal Ecology, 77(4), 802-813.

Engelhardt, B. E., & Stephens, M. (2010). Analysis of population structure: a unifying framework and novel methods based on sparse factor analysis. PLoS Genet, 6(9), e1001117. doi: 10.1371/journal.pgen.1001117

Enright, P. L. (2003). How to make sure your spirometry tests are of good quality. Respiratory care, 48(8), 773-776.

Ethier, A. L. M., Scheuhammer, A. M., & Bond, D. E. (2008). Correlates of mercury in fish from lakes near Clyde Forks, Ontario, Canada. Environmental pollution,154(1), 89-97.

Forbes, SA, N Bindal, S Bamford, C Cole, CY Kok, D Beare, M Jia, et al. 2011. "COSMIC: Mining Complete Cancer Genomes in the Catalogue of Somatic Mutations in Cancer", *Nucleic Acids Research* 39 (Database issue) (January): D945–50. doi:10.1093/nar/gkq929. pubmedcentral.nih.gov/articlerender.fcgi?artid=3013785&tool=pmcentrez &rendertype=abstract.

Forsyth, RS, DD Clarke and RLWright. 1994. "Overfitting Revisited: An Information Theoretic Approach to Simplifying Discrimination Trees", *J. Expt. Theor. Artif. Intell*. 6, 289-302.

Fox, J. (1997) Applied Regression, Linear Models, and Related Methods. Sage.

Gabrielli, A, EV Avvedimento and T Krieg. 2009. "Scleroderma", *The New England Journal of Medicine*, May 7;360(19):1989-2003. doi: 10.1056/NEJMra0806188.

Gelber, AC, RL Manno, AA Shah, A Woods, EN Le, F Boin, LK Hummers and FM Wigley. 2013. "Race and association with disease manifestations and mortality in scleroderma: a 20-year experience at the Johns Hopkins Scleroderma Center and review of the literature", *Medicine*, 92(4), 191-205.

Gerling, IC, S Singh, NI Lenchik, DR Marshall and J Wu. 2006. "New Data Analysis and Mining Approaches Identify Unique Proteome and Transcriptome Markers of Susceptibility to Autoimmune Diabetes", *Molecular & Cellular Proteomics : MCP* 5 (2) (February): 293–305. doi:10.1074/mcp.M500197-MCP200. http://www.ncbi.nlm.nih.gov/pubmed/16227630.

Gershon, ND. 1992. "The Role of Visualization in Biomedicine", *Engineering in Medicine and Biology Society*, 1992 14th Annual International Conference of the IEEE, Vol. 7, 2881-2882.

Goronzy JJ, CM Weyand. 2007. "The innate and adaptive immune systems", In: Goldman L, D Ausiello, eds. *Cecil Medicine*. 23rd ed. Philadelphia, Pa: Saunders Elsevier; 2007: chap 42.

Gregorutti, B., Michel, B., & Saint-Pierre, P. (2013). Correlation and variable importance in random forests. arXiv preprint arXiv:1310.5726.

Guikema, SD, SM Quiring and S‐R Han. "Prestorm estimation of hurricane damage to electric power distribution systems", *Risk Analysis* 30.12 (2010): 1744-1752.

Hall, JC and A Rosen. 2010. "Reviews Type I Interferons : Crucial Participants in Disease Amplification in Autoimmunity" *Nature Reviews Rheumatology* 6 (1): 40–49. doi:10.1038/nrrheum.2009.237. http://dx.doi.org/10.1038/nrrheum.2009.237.

Hankinson, J.L.. 1999. *Office spirometry: does poor quality render it impractical?* Chest, 1999. **116**(2): p. 276–277.

Harris, ML and A Rosen. 2003. "Autoimmunity in Scleroderma: The Origin, Pathogenetic Role and Clinical Significance of Autoantibodies", *Current Opinion in Rheumatology* 15 (6) (November): 778–84. http://www.ncbi.nlm.nih.gov/pubmed/14569210.

Harrison NK, R J McAnulty, PL Haslam, CM Black and GJ Laurent .1990. "Evidence for protein oedema, neutrophil influx and enhanced collagen

production in lungs of patients with systemic sclerosis", *Thorax*, 45: 606-610.

Hastie, T., Tibshirani, R., Friedman, J., Hastie, T., Friedman, J., & Tibshirani, R. (2009). The elements of statistical learning (Vol. 2, No. 1). New York: Springer.

Hedley, DW, ML Friedlander, W Taylor, CA Rugg and EA Musgrove. 1983. "Method for Analysis of Cellular DNA Content of Paraffin-Embedded Pathological Material Using Flow Cytometry", *Journal of Histochemistry & Cytochemistry* 31 (11) (November 1): 1333–1335. http://jhc.sagepub.com/lookup/doi/10.1177/31.11.6619538.April): 570–8. doi:10.1002/acr.20416. http://www.ncbi.nlm.nih.gov/pubmed/21452268.

Hegewald MJ and RO Crapo. 2010. "Pulmonary function testing", In: Mason RJ, Broaddus VC, Martin TR, et al, eds., Murray and Nadel's Textbook of Respiratory Medicine, 5th ed. Philadelphia, Pa: Saunders Elsevier; 2010:chap 24.

Hennekens, C. H., Buring, J. E., & Doll, R. (1987). Epidemiology in medicine (Vol. 255, No. 304, pp. 246-252). S. L. Mayrent (Ed.). Boston: Little, Brown.

Hirzel, A. H., Helfer, V., & Metral, F. (2001). Assessing habitat-suitability models with a virtual species. Ecological modelling, 145(2), 111-121.

Hoggart, C. J., Parra, E. J., Shriver, M. D., Bonilla, C., Kittles, R. A., Clayton, D. G., & McKeigue, P. M. (2003). Control of confounding of genetic associations in stratified populations. The American Journal of Human Genetics, 72(6), 1492-1504.

Hothorn, T., Hornik, K., & Zeileis, A. (2006). Unbiased recursive partitioning: A conditional inference framework. Journal of Computational and Graphical statistics, 15(3), 651-674.

Hunzelmann, N, E Genth, T Krieg, W Lehmacher, I Melchers, M Meurer, P Moinzadeh, et al. 2008. "The Registry of the German Network for Systemic Scleroderma: Frequency of Disease Subsets and Patterns of Organ Involvement", *Rheumatology (Oxford, England)* 47 (8) (August): 1185–92. doi:10.1093/rheumatology/ken179.

http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2468885&tool=pmcentrez&rendertype=abstract.

IRIS - Immune Response In Scleroderma dataset. (2013). http://www.hopkinsscleroderma.org/ research/ translational-research/

Isaeva, O and VA Osipov. 2009a. "Different Strategies for Cancer Treatment: Mathematical Modelling", *Computational and Mathematical Methods in Medicine* 10 (4): 253–272. doi:10.1080/17486700802536054. http://www.hindawi.com/journals/cmmm/2009/621782/.

———. 2009b. "Modelling of Anti-Tumour Immune Response: Immunocorrective Effect of Weak Centimetre Electromagnetic Waves", *Computational and Mathematical Methods in Medicine* 10 (3): 185–201. doi:10.1080/17486700802373540. http://www.hindawi.com/journals/cmmm /2009/251617/.

Iwami, S, Y Takeuchi, Y Miura, T Sasaki and T Kajiwara. 2007. "Dynamical Properties of Autoimmune Disease Models: Tolerance, Flare-up, Dormancy", *Journal of Theoretical Biology* 246 (4) (June 21):

646–59. doi:10.1016/j.jtbi.2007.01.020. http://www.ncbi.nlm.nih.gov/pubmed/17343876.

Jiang, Z., & Gentleman, R. (2007). Extensions to gene set enrichment.Bioinformatics, 23(3), 306-313.

Joseph, CG, E Darrah, AA Shah, AD Skora, LA Casciola-Rosen, FM Wigley, F Boin, A Fava, C Thoburn, I Kinde, Y Jiao, N Papadopoulos, KW Kinzler, B Vogelstein and A Rosen. 2013. "Association of the Autoimmune Disease Scleroderma with an Immunologic Response to Cancer", *Science,* 10: Vol. 343 no. 6167 pp. 152-157, *DOI:* 10.1126/science.1246886

Karlsson, H., Henningsson, P., Bäckman, J., Hedenström, A., & Alerstam, T. (2010). Compensation for wind drift by migrating swifts. Animal Behaviour, 80(3), 399-404.

Khan, J., Wei, J. S., Ringner, M., Saal, L. H., Ladanyi, M., Westermann, F., Berthold, F., Schwab, M., Antonescu, C. R., Peterson, C., & Meltzer, P. S. (2001). Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. Nature Medicine, 7(6), 673-679. doi: Doi 10.1038/89044

Kim, S. Y., & Volsky, D. J. (2005). PAGE: parametric analysis of gene set enrichment. BMC bioinformatics, 6(1), 144.

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. Paper presented at the 14th international joint conference on Artificial intelligence.

Kononenko, I. (2001). Machine learning for medical diagnosis: history, state of the art and perspective. Artificial Intelligence in Medicine, 23(1), 89-109. doi: Doi 10.1016/S0933-3657(01)00077-X

Krzywinski, M, J Schein, I Birol, J Connors, R Gascoyne, D Horsman, SJ Jones and MA Marra. 2009."Circos: An Information Aesthetic for Comparative Genomics", *Genome Research* 19 (9) (September): 1639–45. doi:10.1101/gr.092759.109. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2752132&tool=pmcentrez&rendertype=abstract.

Lee, G-B, C-I Hung, B-J Ke, G-R Huang, B-H Hwei and H-F Lai. 2001. "Hydrodynamic Focusing for a Micromachined Flow Cytometer", *Journal of Fluids Engineering,* ASME Transactions, V. 123, 672-679.

León, K, R Peréz, A Lage and J Carneiro. 2000. "Modelling T cell-Mediated Suppression Dependent on Interactions in Multicellular Conjugates", *Journal of Theoretical Biology* 207 (2) (November 21): 231–54. doi:10.1006/jtbi.2000.2169. http://www.ncbi.nlm.nih.gov/pubmed/11034831.

León, K, J Faro, A Lage and J Carneiro. 2004. "Inverse Correlation Between the Incidences of Autoimmune Disease and Infection Predicted by a Model of T Cell Mediated Tolerance", *Journal of Autoimmunity* 22 (1) (February): 31–42. doi:10.1016/j.jaut.2003.10.002. linkinghub.elsevier.com/retrieve/pii/S0896841103001719.

Leon, K and K Garcia-Martinez. 2011. Mathematical Models and Immune Cell Biology, Edited by Carmen Molina-París and Grant Lythe. New York, NY: Springer New York. doi:10.1007/978-1-4419-7725-0. springerlink.com/index/10.1007/978-1-4419-7725-0.

Lindahl, GE, CJW Stock, X Shi-Wen, P Leoni, P Sestini, SL Howat, G Bou-Gharios, et al. 2013. "Microarray Profiling Reveals Suppressed Interferon Stimulated Gene Program in Fibroblasts from Scleroderma-Associated Interstitial Lung Disease", *Respiratory Research* 14 (1) (January): 80. doi:10.1186/1465-9921-14-80.

http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3750263&tool=pmcentrez&rendertype=abstract.

Lisboa, P. J., Wong, H., Harris, P., & Swindell, R. (2003). A Bayesian neural network approach for modelling censored data with an application to prognosis after surgery for breast cancer. Artificial intelligence in medicine, 28(1), 1-25.

Loeliger, E, CL Nehaniv and AJ Munro. 2012. "Heat-Maps and Visualization for Heterogeneous Biomedical Data Based on Information Distance Geometry, Information Processing in Cells and Tissues", *Lecture Notes in Computer Science* Volume 7223, 2012, pp 183-187

Lota, H. K., & Renzoni, E. A. (2012). Circulating biomarkers of interstitial lung disease in systemic sclerosis. Int J Rheumatol, 2012, 121439. doi:10.1155/2012/121439

Luo, W., Friedman, M. S., Shedden, K., Hankenson, K. D., & Woolf, P. J. (2009). GAGE: generally applicable gene set enrichment for pathway analysis.BMC bioinformatics, 10(1), 161.

Luzina, IG, SP Atamas, R Wise, FM Wigley, J Choi, HQ Xiao and B White. 2003. "Occurrence of an activated, profibrotic pattern of gene

expression in lung CD8+ T cells from scleroderma patients“, *Arthritis & Rheumatism*, 48(8), 2262-2274.

Lyons, AB, CR Parish. 1994. “Determination of Lymphocyte Division by Flow Cytometry”, *Journal of Immunological Methods* 171 (1) (May 2): 131–7. http://www.ncbi.nlm.nih.gov/pubmed/8176234.

Mammen, AL, LA Casciola-Rosen, JC Hall, L Christopher-Stine, AM Corse and A Rosen. 2009. “Expression of the Dermatomyositis Autoantigen Mi-2 in Regenerating Muscle”, *Arthritis and Rheumatism* 60 (12) (December): 3784–93. doi:10.1002/art.24977. http://www. pubmedcentral.nih. gov/articlerender.fcgi? artid=2828900&tool=pmcentrez&rendertype=abstract.

Manno, R L and FM Wigley. 2011. Geriatric Rheumatology, Edited by Yuri Nakasato and Raymond L. Yung: 275–285. doi:10.1007/978-1-4419-5792-4. http://www.springerlink.com /index/10.1007/978-1-4419-5792-4.

Manno, RL, FM Wigley, AC Gelber and LK Hummers. 2011a. “Late-Age Onset Systemic Sclerosis”, *The Journal of Rheumatology* 38 (7) (July): 1317–25. doi:10.3899/jrheum.100956.

http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3136880&tool=pmcentrez &rendertype=abstract.

Manno, RL and FM Wigley. 2011b. Scleroderma in the Elderly Population. In Geriatric Rheumatology (pp. 275-285). Springer New York.

Mathai, SK, M Gulati, X Peng, TR Russell, AC Shaw, AN Rubinowitz, LA Murray, et al. 2010. "Circulating Monocytes from Systemic Sclerosis Patients with Interstitial Lung Disease Show an Enhanced Profibrotic Phenotype", *Laboratory Investigation; a Journal of Technical Methods and Pathology* 90 (6) (June): 812–23. doi:10.1038/labinvest.2010.73. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3682419&tool=pmcentrez& rendertype=abstract.

Mathian, A, C Parizot, K Dorgham, S Trad, L Arnaud, M Larsen, M Miyara, et al. 2012. "Activated and Resting Regulatory T Cell Exhaustion Concurs with High Levels of Interleukin-22 Expression in Systemic Sclerosis Lesions", *Annals of the Rheumatic Diseases* 71 (7) (July): 1227–34. doi:10.1136/annrheumdis-2011-200709. http://www.ncbi.nlm.nih.gov/pubmed/22696687.

Mayes, MD. 2003. “Scleroderma Epidemiology” *Rheumatic Disease Clinics of North America* 29 (2) (May): 239–254. doi:10.1016/S0889-857X(03)00022-X. http://linkinghub.elsevier.com /retrieve/pii/S0889857X0300022X.

Merico D, R Isserlin, O Stueker, A Emili, and GD Bader .2010. “Enrichment Map: A Network-Based Method for Gene-Set Enrichment *Visualization and Interpretation*”, *PLoS ONE* 5(11): e13984. doi:10.1371/journal.pone.0013984

Merkel, PA, NP Silliman, PJ Clements, CP Denton, DE Furst, MD Mayes, JE Pope, RP Polisson, JB Streisand and JR Seibold. 2012. “Patterns and Predictors of Change in Outcome Measures in Clinical Trials in Scleroderma: An Individual Patient Meta-Analysis of 629 Subjects with Diffuse Cutaneous Systemic Sclerosis”, *Arthritis and Rheumatism* 64 (10) (October): 3420–9. doi:10.1002/art.34427. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3357459 &tool=pmcentrez&rendertype=abstract.

Mickey, R. M., & Greenland, S. (1989). The impact of confounder selection criteria on effect estimation. American journal of Epidemiology, 129(1), 125-137.

Miller, M. R. (2005). Standardisation of spirometry. Eur Respir J, 26(2), 319-38.

Miltenyi Biotech. 2014. https://www.miltenyibiotec.com/en/products-and-services/macs-cell-separation/macs-technology/microbeads_dp.aspx

Miyawaki, S., Asanuma, H., Nishiyama, S., & Yoshinaga, Y. (2005). Clinical and serological heterogeneity in patients with anticentromere antibodies. The Journal of Rheumatology, 32(8), 1488–94. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/16078324

Moore, JH, JS Parker, NJ Olsen and TM Aune. 2002. "Symbolic Discriminant Analysis of Microarray Data in Autoimmune Disease", *Genet Epidemiol.*, 69 (March): 57–69. doi:10.1002/gepi.01117.

Mootha VK, CM Lindgren, KF Eriksson, A Subramanian, S Sihag, J Lehar, P Puigserver, E Carlsson, M Ridderstråle, E Laurila, N Houstis, MJ Daly, N Patterson, JP Mesirov, TR Golub, P Tamayo, BM Spiegelman, ES Lander, JN Hirschhorn, D Altshuler, LC Groop. 2003. "*PGC-1α Responsive Genes Involved in Oxidative Phosphorylation are Coordinately Downregulated in Human Diabetes", Nature Genetics* 34(3):267-73.

Mundry, R., & Nunn, C. L. (2009). Stepwise model fitting and statistical inference: turning noise into signal pollution. The American Naturalist, 173(1), 119-123.

Murphy, K. (2012). Janeway's immunobiology (8th ed.). London and New York: Garland Science.

Nelder, J. A., & Wedderbu.Rw. (1972). Generalized Linear Models. Journal of the Royal Statistical Society Series a-General, 135(3), 370-&. doi: Doi 10.2307/2344614

Nevo, U, I Golding, AU Neumann, M Schwartz and S Akselrod. 2004. "Autoimmunity as an Immune Defense Against Degenerative Processes: a Primary Mathematical Model Illustrating the Bright Side of Autoimmunity", *Journal of Theoretical Biology* 227 (4) (April 21): 583–92. doi:10.1016/j.jtbi.2003.11.031. http://www.ncbi.nlm.nih.gov/pubmed/15038992.

Newton, M. A., Quintana, F. A., Den Boon, J. A., Sengupta, S., & Ahlquist, P. (2007). Random-set methods identify distinct aspects of the enrichment signal in gene-set analysis. The Annals of Applied Statistics, 85-106.

Nicodemus, K. K., Malley, J. D., Strobl, C., & Ziegler, A. (2010). The behaviour of random forest permutation-based variable importance measures under predictor correlation. BMC bioinformatics, 11(1), 110.

Nicoletti, I, G Migliorati, MC Pagliacci, F Grignani and C Riccardi. 1991. "A Rapid and Simple Method for Measuring Thymocyte Apoptosis by Propidium Iodide Staining and Flow Cytometry", *Journal of Immunological Methods* 139 (2) (June 3): 271–9. http://www.ncbi.nlm.nih.gov /pubmed/1710634.

Nisbet, R., John Elder IV, Gary Miner. (2009). Handbook of statistical analysis and data mining applications.

Olden, J. D., Joy, M. K., & Death, R. G. (2004). An accurate comparison of methods for quantifying variable importance in artificial neural networks using simulated data. Ecological Modelling, 178(3-4), 389-397. doi: DOI 10.1016/j.ecolmodel.2004.03.013

Orre, R., Lansner, A., Bate, A., & Lindquist, M. (2000). Bayesian neural networks with confidence estimations applied to data mining. Computational Statistics & Data Analysis, 34(4), 473-493.

Pandey, A. K., Wilcox, P., Mayo, J. R., Sin, D., Moss, R., Ellis, J., Borwn, J. & Leipsic, J. (2010). Predictors of pulmonary hypertension on high-resolution computed tomography of the chest in systemic sclerosis: a retrospective analysis. Canadian Association of Radiologists Journal, 61(5), 291-296.

Paul, TK and H Iba. 2006. "Classification of Scleroderma and Normal Biopsy Data and Identification of Possible Biomarkers of the Disease", *Computational Intelligence and Bioinformatics and Computational Biology*, CIBCB '06. 2006 IEEE Symposium on , vol., no., pp.1,6, 28-29 Sept. doi: 10.1109/CIBCB.2006.330951

Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. Philosophical Magazine, 2(7-12), 559-572.

Pérez-Bocanegra C, R Solans-Laqué, CP Simeón-Aznar, M Campillo, V Fonollosa-Pla and M Vilardell-Tarrés. 2010. "Age-Related Survival and Clinical Features in Systemic Sclerosis Patients Older or Younger Than 65 at Diagnosis", *Rheumatology* (Oxford, England) 49 (6) (June): 1112–7. doi:10.1093/rheumatology/keq046. http://www.ncbi.nlm.nih.gov/pubmed/20223816.

Pérez Campos D, M Estevez Del Toro, A Pena Casanovas , PP Gonzalez Rojas , L Morales Sánchez and AR Gutiérrez Rojas . 2012. "Are high doses of prednisone necessary for treatment of interstitial lung disease in systemic sclerosis?", *Reumatology Clin*, 2012 Mar-Apr;8(2):58-62. doi: 10.1016/j.reuma.2011.11.006. Epub 2012 Feb 7.

Perfetto, SP, PK Chattopadhyay and M Roederer. 2004. "Seventeen-Colour Flow Cytometry: Unravelling the Immune System", *Nature Reviews. Immunology* 4 (8) (August): 648–55. doi:10.1038/nri1416. http://www.ncbi.nlm.nih.gov/pubmed/15286731.

Petty, T.L. and P.L. Enright. 2003.Simple Office Spirometry for Primary Care Practitioners. NLHEP and AlphaMedia, Inc..

Picot, J, CL Guerin, C Le Van Kim and Ch M Boulanger. 2012. "Flow Cytometry: Retrospective, Fundamentals and Recent Instrumentation", *Cytotechnology* 64 (2) (March): 109–30. doi:10.1007/s10616-011-9415-0. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3279584&tool=pmcentrez&rendertype=abstract.

Pierce, R .2005. "Spirometry: An essential clinical measurement", *Australian Family Physician,* 34 (7): 535–539.

Qian, Y, C Wei, FE-H Lee, J Campbell, J Halliley, JA Lee, J Cai, et al. 2010. "Elucidation of Seventeen Human Peripheral Blood B-Cell Subsets and Quantification of the Tetanus Response Using a Density-Based Method for the Automated Identification of Cell Populations in Multidimensional Flow Cytometry Data", *Cytometry. Part B, Clinical Cytometry* 78 Suppl 1 (May) (January): S69–82. doi:10.1002/cyto.b.20554. http://www.pubmedcentral.nih.gov /articlerender.fcgi? artid=3084630&tool= pmcentrez &rendertype=abstract.

R Core Team. 2013. "R: A language and environment for statistical computing", R Foundation for Statistical Computing, Vienna, Austria, http://www.R-project.org/

Raja, K, M Raja, M Plasil, L Rihova, J Pelcova, Z Adam and R Hajek. 2013. "Flow Cytometry-Based Enumeration and Functional Characterization of CD8 T Regulatory Cells in Patients with Multiple Myeloma Before and After Lenalidomide Plus Dexamethasone Treatment", *Cytometry B Clin Cytom*, Jul 3. doi: 10.1002/cytob.21109.

Ramsey, F., & Schafer, D. (2012). The statistical sleuth: a course in methods of data analysis. Cengage Learning.

Razykov, I, B Levis, M Hudson, M Baron and BD Thombs. 2013. "Prevalence and Clinical Correlates of Pruritus in Patients with Systemic Sclerosis: An Updated Analysis of 959 Patients", *Rheumatology (Oxford, England)* 52 (11) (November): 2056–61. doi:10.1093/rheumatology/ket275. http://www.ncbi.nlm.nih.gov/pubmed/23946437.

Reynolds HY. 2011. "Respiratory structure and function: mechanisms and testing", In: Goldman L, Schafer AI, eds. *Cecil Medicine*, 24th ed. Philadelphia, PA: Saunders Elsevier; 2011:chap 85.

Rosen, A and L Casciola-Rosen. 1999. "Autoantigens as Substrates for Apoptotic Proteases: Implications for the Pathogenesis of Systemic Autoimmune Disease" *Cell Death and Differentiation* 6 (1) (January): 6–12. doi:10.1038/sj.cdd.4400460. http://www.ncbi.nlm.nih.gov/pubmed/10200542.

———. 2009. "Autoantigens in Systemic Autoimmunity: Critical Partner in Pathogenesis", *Journal of Internal Medicine* 265 (6) (June): 625–31. doi:10.1111/j.1365-2796.2009.02102.x. http://www.ncbi.nlm.nih.gov/pubmed/19493056.

Saeki, K and Y Iwasa. 2009. “Advantage of Having Regulatory T Cells Requires Localized Suppression of Immune Reactions”, *Journal of Theoretical Biology* 260 (3) (October 7): 392–401. doi:10.1016/j.jtbi.2009.06.020. http://www.ncbi.nlm.nih.gov/pubmed/19563814.

———. 2010. “Optimal Number of Regulatory T Cells”, *Journal of Theoretical Biology* 263 (2) (March 21): 210–8. doi:10.1016/j.jtbi.2009.11.012. http://www.ncbi.nlm.nih.gov/pubmed/19961861.

Salamunić, I. 2010. “Laboratory diagnosis of autoimmune diseases – new technologies, old dilemmas”, *Biochemia Medica* 2010;20(1):45-56. http://dx.doi.org/10.11613/BM.2010.006

Sanz, I and FEH Lee. 2010. “B cells as therapeutic targets in SLE”, *Nature Reviews Rheumatology*, 6(6), 326-337.

Schachna, L, FM Wigley, B Chang, B White, RA Wise and AC Gelber. 2003. “Age and Risk of Pulmonary Arterial Hypertension in Scleroderma”, *Chest* 124 (6) (December): 2098–104. http://www.ncbi.nlm.nih.gov/pubmed/14665486.

Shah, AA and A Rosen. 2011. “Cancer and Systemic Sclerosis: Novel Insights into Pathogenesis and Clinical Implications”, *Current Opinion in Rheumatology* 23 (6) (November): 530–5. doi:10.1097/BOR.0b013e32834a5081. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3373179&tool=pmcentrez&rendertype=abstract.

Shapiro HM. 2003. Practical flow cytometry, Wiley, New Jersey (USA): Fourth edition edn; 2003.http://books.google.com/books?hl=en&lr=&id=JhSyimPKuJwC&oi=fnd&pg=PR7&dq=Shapiro+practical+flow+cytomery&ots=OPF_FSNONA&sig=8gv-8h5KDfXDckDDWDjg5gZIfcY#v=onepage&q&f=false

Siegel RM, PE Lipsky.2009. Autoimmunity, In: Firestein GS, Budd RC, Harris Ed, et al, eds. Kelley's Textbook of Rheumatology, 8th ed. Philadelphia, Pa: Saunders Elsevier; 2009:chap 15.

Slobodin, G, M Sheikh Ahmad, I Rosner, R Peri, M Rozenbaum, A Kessel, E Toubi and M Odeh. 2010. “Regulatory T Cells (CD4(+)CD25(bright)FoxP3(+)) Expansion in Systemic Sclerosis Correlates with Disease Activity and Severity”, *Cellular Immunology*

261 (2) (January): 77–80. doi:10.1016/j.cellimm.2009.12.009. http://www.ncbi.nlm.nih.gov/pubmed/20096404.

Steen, VD. 1998. “Clinical Manifestations of Systemic Sclerosis”, *Seminars in Cutaneous Medicine and Surgery* 17 (1) (March): 48–54. www.ncbi.nlm.nih.gov/pubmed/9512107.

Strobl, C, A-L Boulesteix, T Kneib, T Augustin and A Zeileis. 2008. “Conditional Variable Importance for Random Forests”, *BMC Bioinformatics* 9 (January): 307. doi:10.1186/1471-2105-9-307. http://www.pubmedcentral.nih.gov/articlerender. fcgi?artid=2491635&tool=pmcentrez &rendertype=abstract.

Strobl, C, T Hothorn and A Zeileis. 2009. “Party on ! A New , Conditional Variable Importance Measure for Random Forests Available in the Party Package Party on !”, *The R Journal of Animal Ecology* (050).

Subramanian, A, P Tamayo, VK Mootha, S Mukherjee and BL Ebert. 2005. “Gene Set Enrichment Analysis : A Knowledge-Based Approach for Interpreting Genome-Wide”, *PNAS*, vol. 102 no. 4315545–15550, doi: 10.1073/pnas.0506580102

Therneau, T., Atkinson, B., Ripley, B., & Ripley, M. B. (2014). R Package 'rpart'.

Toloşi, L., & Lengauer, T. (2011). Classification with correlated features: unreliability of feature ranking and solutions. Bioinformatics, 27(14), 1986-1994.

Truchetet, M-E, NC Brembilla, E Montanari and C Chizzolini. 2010. "T cell Subsets in Scleroderma Patients", *Expert Review of Dermatology* 5 (4) (August): 403–415. doi:10.1586/edm.10.33. http://www.expert-reviews.com/doi/abs/10.1586/edm.10.33.

Tufte, ER. 2001. The Visual Display of Quantitative Information, Cheshire, Conn.

Varga, J .2014. http://www.uptodate.com/contents/prognosis-and-treatment-of-interstitial-lung-disease-in-systemic-sclerosis-scleroderma#H1

Vélez de Mendizábal, N, J Carneiro, RV Solé, J Goñi, J Bragard, I Martinez-Forero, S Martinez-Pasamar, et al. 2011. "Modeling the Effector - Regulatory T Cell Cross-Regulation Reveals the Intrinsic Character of Relapses in Multiple Sclerosis", *BMC Systems Biology* 5 (1) (January): 114. doi:10.1186/1752-0509-5-114.

http://www.pubmedcentral.nih.gov/articlerender.
fcgi ?artid=3155504&tool=pmcentrez&rendertype=abstract.

Vermes, I, C Haanen and C Reutelingsperger. 2000. “Flow Cytometry of Apoptotic Cell Death”, *Journal of Immunological Methods* 243 (1-2) (September 21): 167–90. http://www.ncbi.nlm.nih.gov/pubmed/10986414.

Wahren-Herlenius, M and T Dorner. 2012. “Immunopathogenic mechanisms of systemic autoimmune disease", *The Lancet*, Volume 382, Issue 9894, 31 August–6 September 2013, Page 744

Waniewski, J and D Prikrylová. 1988. “Autoimmunity and Its Therapy: Mathematical Modelling”, Immunology Letters 18 (1) (May): 77–80. ncbi.nlm.nih.gov/pubmed/3378833.

Warrington, KJ, U Nair, LD Carbone, AH Kang and AE Postlethwaite. 2006. “Characterisation of the immune response to type I collagen in scleroderma”, Arthritis Research & Therapy, 8:R136 (doi:10.1186/ar2025)

White, B. (2003). Interstitial lung disease in scleroderma. Rheumatic Disease Clinics of North America, 29(2), 371-390.

Whitfield, ML, DR Finlay, J I Murray, O G Troyanskaya, J-T Chi, A Pergamenschikov, TH McCalmont, PO Brown, D Botstein and M Kari Connolly. 2003. "Systemic and Cell Type-Specific Gene Expression Patterns in Scleroderma Skin", *Proceedings of the National Academy of Sciences of the United States of America* 100 (21) (October 14): 12319–24. doi:10.1073/pnas.1635114100. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=218756&tool =pmcentrez&rendertype=abstract.

Williams, D. A. (1987) Generalized linear model diagnostics using the deviance and single case deletions. Applied Statistics 36, 181–191.

Winterbauer RH. 1964. "Multiple telangiectasia, Raynaud's phenomenon, sclerodactyly and subcutanious calcinosis: a syndrome mimicking hereditary hemorrhagic telangiectasia", *Bulletin of the Johns Hopkins Hospital* 114: 361–83.

Yamadori, I, J Fujita, H Kajitani, S Bandoh, M Tokuda, Y Yang, ... and T Ishida. 2000. "Lymphocyte subsets in lung tissues of non-specific interstitial pneumonia and pulmonary fibrosis associated with collagen vascular disorders: correlation with CD4/CD8 ratio in bronchoalveolar lavage", *Lung*, 178(6), 361-370.

Yang, Y, EJ Kort, N Ebrahimi, Z Zhang and BT The. 2010. “Dual KS: Defining Gene Sets with Tissue Set Enrichment Analysis”, *Cancer Informatics* 2010:9 1-9 doi: 10.4137/CIN.S2892

Zompatori, M., Leone, M. B., Giannotta, M., Galiè, N., Palazzini, M., Reggiani, M. L. B., Bono, L. & Pollini, G. S. (2013). Pulmonary hypertension and systemic sclerosis: the role of high-resolution computed tomography. La radiologia medica, 118(8), 1360-1372.

Zweig, MH and G Campbell. “Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine”, *Clinical Chemistry,* 39.4 (1993): 561-577.

I. APPENDIX

I.1 Appendix - FC Variable Panels

Figure I-1. Hierachical Structure of Activation Panel

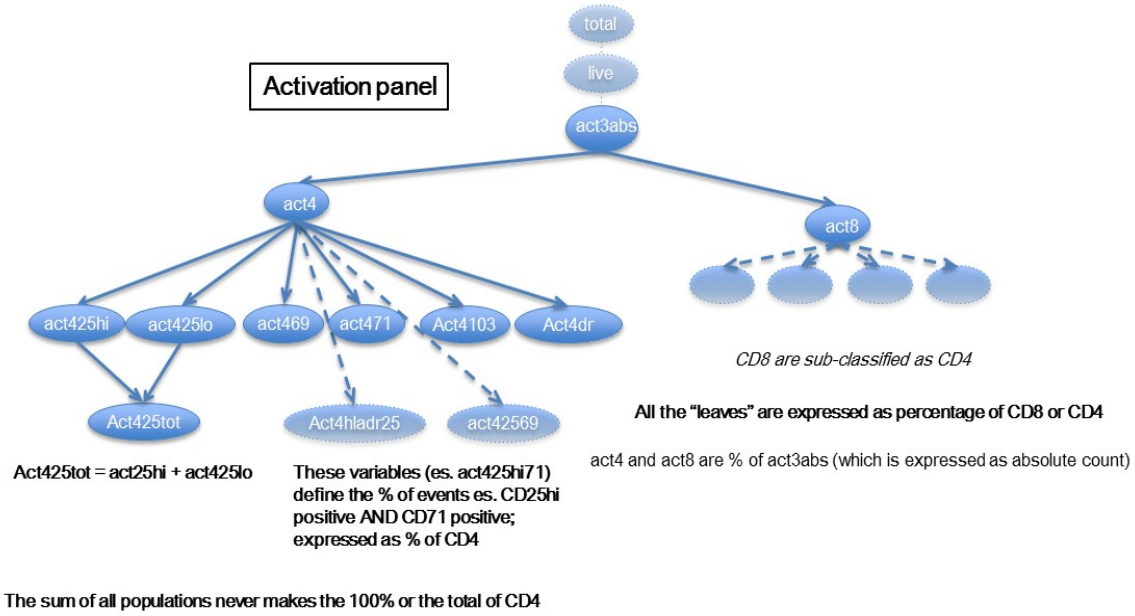


Figure I-2. Hierarchical Structure of Polarization Panel

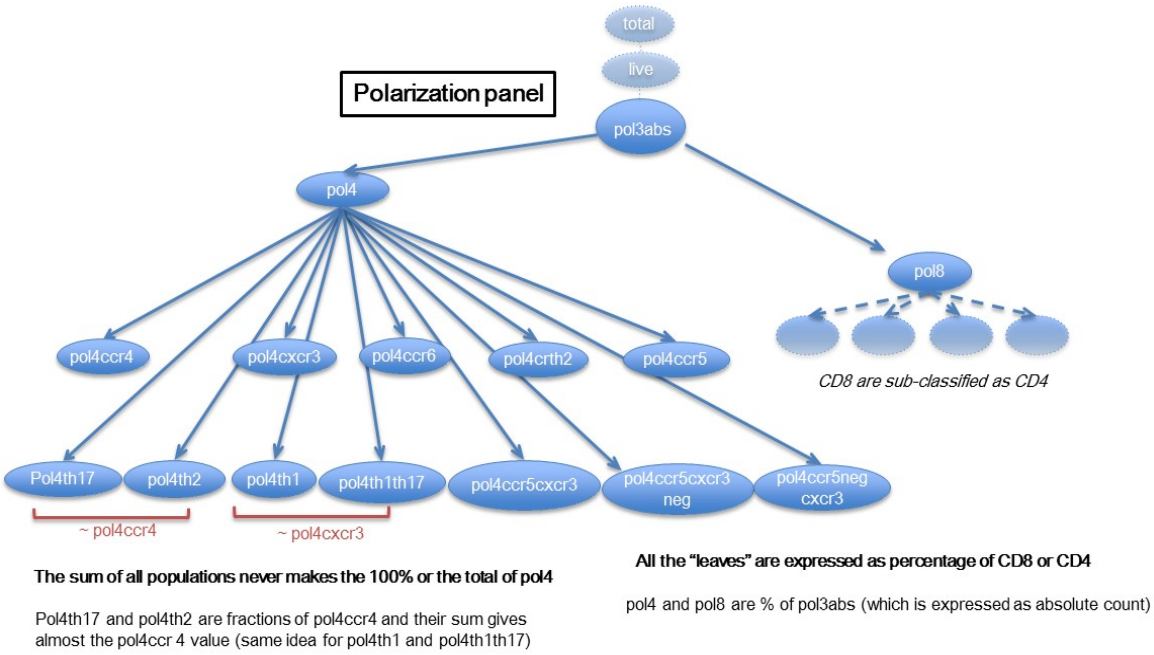
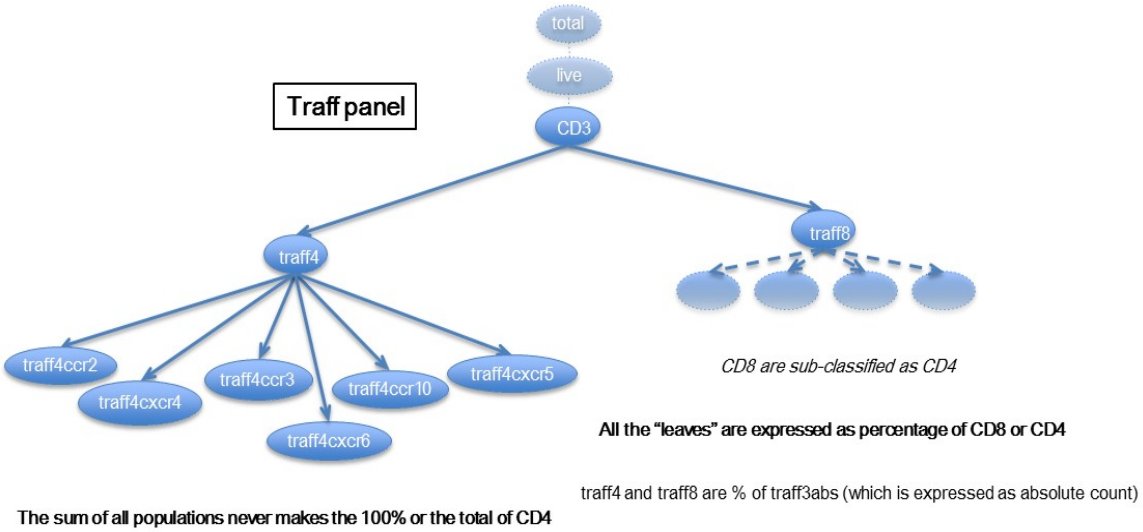


Figure I-3. Hierarchical Structure of Traffic Panel



I.2 Appendix FC sets identified by CRF-GSEA algorithm

Table I-1 FC sets identified by CRF-GSEA algorithm (Phenotype is ILD)

CRF FC Set	Filter Set	Bioinformed FC Set	P-value Based FC Set
pol8ccr5	act4103	pol4th1	pol8ccr5
pol8ccr5cxcr3neg	pol8ccr5	pol4th2	memnaive4
mememm4	act425tot	pol4th17	mememm4
pol8ccr5cxcr3	act425lo	pol4th1th17	pol8ccr5cxcr3
act4103	pol8th17	act425lo	pol8ccr5cxcr3neg
act425lo	mememra4	act425hi	act425lo
mememm478	act8103	act4dr	act425tot
act8103	pol8ccr5cxcr3neg	traff4ccr10	pol4ccr6
mememm8	pol8ccr4	traff4cxcr6	mememra87
mememra87	act425103	memnaive4	act4103
act425103	pol8th1th2ratio	mem4cmefratio	mememm48
traff4cxcr6	act810371	act8103	mememra478
memnaive4	mememm4	pol8cxcr3	mememra4
act425tot	pol8x3r4ratio	pol8ccr4	act425103
memcm878	mememm478	pol8ccr6	mememm478
act410371	mememra4k	act8dr	
pol4ccr6	pol8th2	traff8cxcr6	
mememm878	mememm878	traff8ccr10	
mememra478	mememra478	memnaive8	
memcm4	mememm48	mem8cmefratio	
memcm478	act4dr	mem8ememraratio	
traff8cxcr4	mememm8	memcd8k	
mememra4	memcm478		
act425hladr	memcm4k		
memcm8	act8dr		
act4103hladr	traff4ccr3		
mememm48	pol4ccr6		

Table I-2 FC sets identified by CRF-GSEA algorithm (Phenotype is Cancer)

CRF FC Set	p-value Based FC Set	t-tests Based FC Set
act4103hladr	act4103hladr	act4103hladr
act410371	act82571	act410371
traff8cxcr5	traff4ccr3	mememra48
traff4ccr3	act410371	act425103
act4103	act4103	act810369
act425103	act425103	act4103
act82571	traff8cxcr5	memcm88
act825	act825	act82571
pol4th2	memcm88	traff4ccr3
memcm88	traff8cxcr4	cd4cd8ratioLOG
mememm4	pol4th2	mememra88
traff8ccr10	traff8ccr10	act810371
		memcm87
		memcm80
		act469hladr
		memcm40
		mememm47

Table I-3 FC sets identified by CRF-GSEA algorithm (Phenotype is Scl70_ab)

CRF FC Set
act82571
pol8ccr6
memnaive4
memcm4
act86925
traff8ccr3
pol8cxcr3
pol8
pol8th17
act869hladr
memcm478
pol4ccr5cxcr3
pol4ccr5
pol4
cd4cd8ratioLOG
pol8ccr5negcxcr3
pol8th1th17
act86971
pol4cxcr3
act869
act810325
memcm47
act8hladr25
pol8th1
memcm48
pol8crth2
pol4th1th17
memcm88
memem48
pol4th17
act871
act46971
act8hladr71

Table I-4 FC sets identified by CRF-GSEA algorithm (Phenotype is ACA)

CRF FC Set
pol8crth2
memcm88
memem88
pol8th17
act825
pol4crth2

I.3 Appendix - PCA Loading Matrix

Table I-5 PCA Loading Matrix for all FC Variables (First 10 Principal Components)

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6	Comp.7	Comp.8	Comp.9	Comp.10
pol4	-0.114	0.134	-0.128	0.015	0.041	0.115	-0.098	0.152	-0.055	0.019
pol8	0.076	-0.131	0.139	-0.034	-0.081	-0.142	0.056	-0.166	0.037	-0.055
memnaive4	-0.193	0.027	0.031	0.092	-0.132	0.052	-0.018	0.161	-0.106	-0.005
memcm4	0.151	-0.057	-0.046	-0.210	-0.012	-0.041	-0.035	-0.092	0.132	-0.111
mememra4	0.050	0.014	0.025	0.171	0.142	0.129	0.019	-0.119	-0.020	0.214
memem4	0.101	0.038	0.001	0.112	0.220	-0.072	0.076	-0.122	-0.011	0.114
mememra478	-0.056	0.026	-0.004	0.158	0.098	0.026	0.000	-0.090	0.011	0.193
mememra47	0.087	0.068	0.024	0.026	0.105	0.009	0.030	-0.076	-0.072	-0.008
mememra48	0.077	0.012	0.044	0.089	0.076	0.098	0.065	0.014	0.069	0.216
mememra40	0.149	-0.018	0.049	0.076	0.101	0.181	0.021	-0.082	-0.055	0.098
memcm478	0.141	-0.055	-0.056	-0.203	-0.028	-0.046	-0.054	-0.100	0.128	-0.142
memcm47	-0.006	-0.008	0.000	-0.060	0.071	-0.086	0.124	0.018	-0.064	0.115
memcm48	0.096	-0.025	0.033	-0.120	0.069	0.034	0.069	0.023	0.102	0.135
memcm40	0.116	-0.054	0.061	0.053	0.061	0.034	0.079	-0.037	-0.069	-0.007
memem478	0.021	0.056	-0.063	0.125	0.159	-0.104	0.026	-0.129	0.037	0.114
memem47	0.096	0.016	0.052	-0.024	0.106	-0.098	0.109	-0.064	0.034	0.019
memem48	0.106	0.035	0.040	0.008	0.205	-0.044	0.142	-0.020	0.068	0.135
memem40	0.155	-0.040	0.108	0.038	0.121	0.060	0.047	-0.046	-0.157	-0.024
memcm4k	0.172	-0.041	0.108	0.049	0.144	0.101	0.051	-0.048	-0.135	0.027
mememra4k	0.140	-0.011	0.057	0.066	0.103	0.171	0.004	-0.079	-0.070	0.094
memem4k	0.156	-0.033	0.106	0.030	0.142	0.059	0.061	-0.034	-0.149	-0.015
memnaive8	-0.220	-0.026	-0.019	0.005	-0.087	0.036	0.032	-0.207	-0.029	0.012
memcm8	0.083	0.133	-0.129	-0.213	0.108	0.101	0.008	0.004	0.034	-0.047

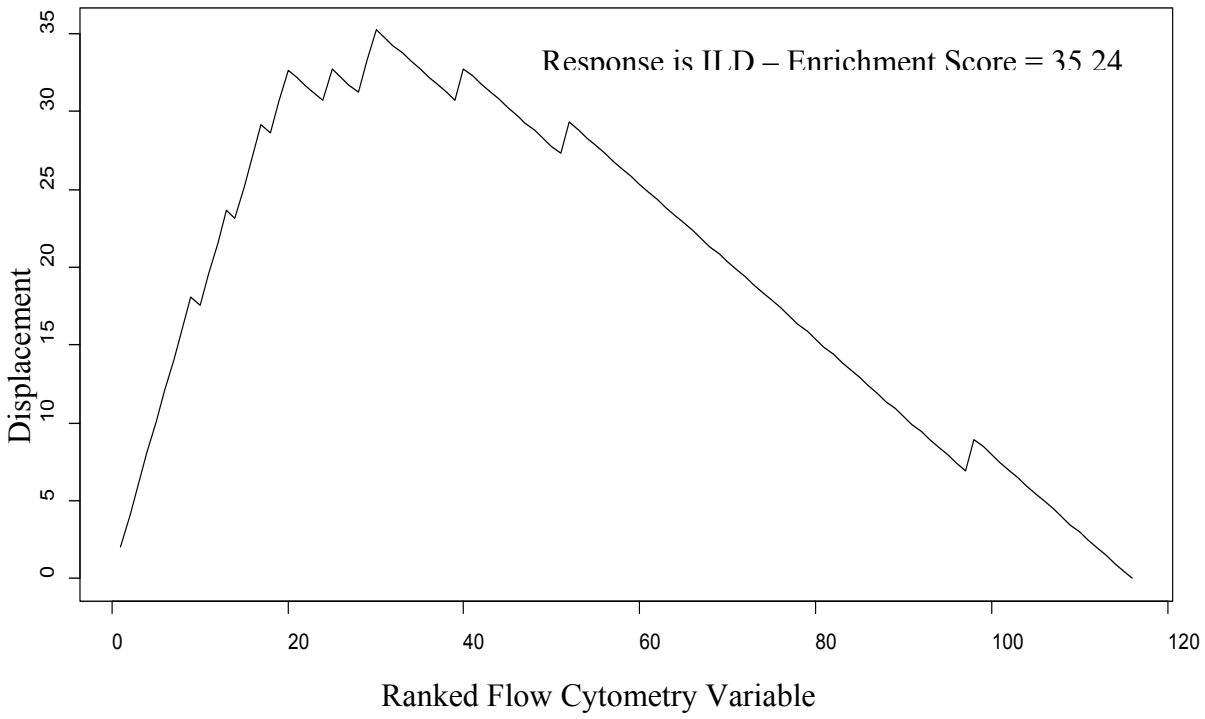
mememra8	0.124	-0.060	0.190	0.089	-0.023	0.062	-0.050	0.076	0.067	0.092
memem8	0.157	0.006	-0.059	0.050	0.077	-0.172	-0.008	0.229	-0.036	-0.073
mememra878	-0.025	0.059	-0.012	0.139	0.042	-0.056	-0.067	-0.054	0.093	0.202
mememra87	0.060	-0.008	0.117	0.014	-0.010	-0.025	-0.103	0.048	0.117	-0.004
mememra88	0.095	0.036	0.063	0.013	0.052	0.012	0.019	0.034	0.087	0.175
mememra80	0.129	-0.081	0.188	0.067	-0.035	0.089	-0.018	0.085	0.025	0.056
memcm878	0.078	0.129	-0.130	-0.216	0.104	0.095	0.011	0.004	0.040	-0.050
memcm87	0.058	0.105	-0.096	-0.030	0.074	0.105	-0.077	-0.007	0.004	-0.012
memcm88	0.112	0.144	-0.077	-0.158	0.108	0.091	0.005	0.004	-0.034	0.012
memcm80	0.101	0.051	0.007	0.021	0.107	0.159	-0.018	0.009	-0.118	0.025
memem878	0.092	0.051	-0.176	0.048	0.036	-0.205	-0.047	0.123	0.056	0.044
memem87	0.100	-0.010	0.051	0.007	0.021	-0.089	0.004	0.105	0.023	-0.169
memem88	0.114	0.112	-0.083	-0.022	0.087	-0.105	-0.001	0.062	0.034	0.043
memem80	0.107	-0.072	0.108	0.032	0.064	-0.004	0.040	0.202	-0.141	-0.136
memcd8k	0.147	-0.120	0.159	0.049	-0.012	0.059	-0.037	0.176	-0.041	-0.077
mememra8k	0.127	-0.099	0.168	0.055	-0.048	0.105	-0.075	0.087	0.033	0.031
memem8k	0.110	-0.099	0.088	0.022	0.032	-0.016	0.022	0.205	-0.105	-0.170
pol4ccr4	0.109	0.167	0.026	0.096	-0.077	-0.150	0.032	-0.143	-0.030	-0.115
pol4ccr5	0.013	-0.025	-0.060	0.008	0.053	-0.204	0.118	-0.057	-0.268	0.089
pol4ccr6	0.128	-0.092	-0.083	0.006	-0.011	-0.090	-0.095	-0.196	-0.062	-0.105
pol4crth2	0.109	-0.047	-0.013	-0.044	-0.064	-0.122	0.063	-0.109	0.007	-0.054
pol4cxcr3	0.060	-0.115	-0.079	-0.156	0.155	-0.121	0.139	-0.016	0.113	-0.016
pol4ccr5cxcr3	-0.007	-0.067	-0.086	-0.052	0.089	-0.177	0.165	0.025	-0.177	0.102
pol4ccr5cxcr3neg	0.037	0.045	-0.011	0.078	-0.016	-0.159	-0.004	-0.132	-0.269	0.038
pol4ccr5negcxcr3	0.071	-0.102	-0.053	-0.155	0.138	-0.062	0.088	-0.028	0.202	-0.061
pol4th1	0.016	-0.138	-0.059	-0.183	0.175	-0.027	0.114	0.077	0.135	0.065
pol4th1th17	0.021	-0.170	-0.107	-0.081	0.059	-0.093	0.012	-0.078	0.010	-0.075
pol4th2	0.048	0.221	0.061	0.113	-0.109	-0.111	0.060	-0.066	0.002	-0.065

pol4th17	0.147	0.005	-0.022	0.064	-0.048	-0.075	-0.136	-0.199	-0.120	-0.122
pol8ccr4	0.065	0.234	0.030	0.051	-0.064	-0.124	0.075	-0.038	-0.018	-0.109
pol8ccr5	0.078	-0.048	-0.150	0.082	-0.019	-0.185	-0.064	0.220	-0.073	0.067
pol8ccr6	0.087	0.032	-0.181	0.037	-0.034	-0.143	-0.156	0.078	-0.040	0.039
pol8crth2	0.107	0.117	-0.039	-0.138	0.034	0.004	0.051	-0.029	-0.111	-0.089
pol8cxcr3	-0.150	-0.104	-0.111	0.054	0.083	-0.052	0.178	-0.057	-0.018	-0.092
pol8ccr5cxcr3	0.021	-0.111	-0.132	0.035	0.032	-0.121	0.083	0.193	-0.092	0.009
pol8ccr5cxcr3neg	0.096	0.043	-0.093	0.089	-0.061	-0.158	-0.183	0.137	-0.017	0.093
pol8ccr5negcxcr3	-0.172	-0.067	-0.066	0.044	0.077	-0.006	0.158	-0.141	0.018	-0.103
pol8th1	-0.152	-0.195	-0.103	-0.010	0.109	0.041	0.086	-0.042	-0.030	-0.001
pol8th1th17	-0.024	-0.082	-0.172	0.009	0.050	-0.049	-0.104	0.064	0.016	0.035
pol8th2	0.092	0.254	0.075	-0.023	-0.062	-0.052	-0.011	-0.039	-0.026	-0.033
pol8th17	0.115	0.108	-0.102	-0.013	-0.022	-0.076	-0.141	0.001	-0.128	-0.001
cd4cd8ratioLOG	-0.089	0.144	-0.129	0.020	0.073	0.136	-0.074	0.154	-0.036	0.041
pol8th1th2ratio	-0.128	-0.250	-0.089	0.000	0.078	0.058	0.013	-0.017	-0.011	0.053
pol8x3r4ratio	-0.121	-0.149	-0.045	-0.038	0.055	0.062	-0.068	-0.072	-0.036	0.023
cd8r5th2ratio	-0.012	-0.074	-0.054	0.080	0.006	-0.101	-0.085	0.177	-0.029	0.082
cd4r5th2ratio	-0.062	0.031	-0.032	0.041	0.100	-0.094	0.017	0.036	-0.184	0.114
act425hi	0.018	-0.034	0.011	-0.029	-0.082	-0.107	0.023	-0.095	0.027	0.060
act425lo	0.044	-0.013	-0.006	0.075	-0.092	0.047	-0.154	-0.190	-0.017	-0.047
act425tot	0.047	-0.019	-0.004	0.071	-0.105	0.030	-0.151	-0.206	-0.012	-0.037
act469	0.066	-0.005	-0.124	0.226	-0.068	0.073	0.193	0.005	0.121	-0.094
act471	0.032	-0.123	-0.101	0.052	-0.083	-0.011	-0.025	-0.080	-0.178	0.036
act4103	0.123	-0.115	-0.014	-0.016	-0.152	-0.036	-0.006	-0.045	0.090	0.192
act4dr	0.028	0.092	0.045	-0.147	-0.192	0.058	0.180	0.027	-0.051	0.142
act825	0.116	0.150	-0.090	-0.150	0.016	0.083	0.005	-0.052	-0.053	-0.015
act869	0.047	0.037	-0.156	0.201	-0.016	0.051	0.138	0.053	0.066	-0.154
act871	0.083	-0.111	-0.128	0.030	-0.038	0.176	0.070	-0.009	-0.073	-0.016

act8103	0.041	0.075	-0.151	-0.003	0.047	0.055	-0.174	0.025	0.127	0.121
act8dr	-0.007	0.075	0.012	-0.142	-0.167	0.104	0.167	0.048	-0.145	0.083
act42569	0.079	-0.016	-0.124	0.222	-0.087	0.082	0.166	-0.029	0.107	-0.085
act42571	0.030	-0.130	-0.099	0.063	-0.091	-0.002	-0.068	-0.124	-0.184	0.030
act425103	0.110	-0.105	-0.029	-0.012	-0.161	-0.043	-0.025	-0.093	0.068	0.167
act425hladr	0.032	0.081	0.038	-0.126	-0.211	0.067	0.140	0.002	-0.052	0.118
act46971	0.086	-0.023	-0.153	0.169	-0.124	0.036	0.161	0.025	0.106	-0.041
act469hladr	0.084	-0.009	-0.108	0.078	-0.178	0.077	0.168	0.048	0.124	0.091
act410371	0.095	-0.171	-0.077	-0.007	-0.147	-0.022	-0.027	-0.003	0.054	0.153
act4103hladr	0.107	-0.078	-0.057	-0.037	-0.175	-0.032	0.017	-0.020	0.117	0.167
act471hladr	0.058	-0.050	-0.082	-0.069	-0.218	-0.006	0.130	0.049	-0.115	0.187
act82571	0.093	-0.007	-0.202	-0.092	-0.020	0.147	-0.017	0.008	-0.116	-0.006
act86925	0.047	0.118	-0.149	0.065	0.020	0.101	0.053	-0.046	-0.016	-0.068
act86971	0.096	-0.017	-0.145	0.153	-0.022	0.115	0.186	0.006	0.054	-0.096
act869hladr	0.037	0.038	-0.114	0.107	-0.119	0.088	0.231	0.091	0.016	-0.111
act810325	0.040	0.110	-0.153	0.020	0.025	0.055	-0.154	-0.047	0.077	0.093
act810369	0.055	0.028	-0.164	0.121	0.002	0.094	0.005	-0.044	0.135	-0.056
act8103hladr	-0.016	0.084	-0.118	0.070	0.023	-0.009	-0.025	-0.076	-0.009	0.122
act810371	0.074	-0.065	-0.155	-0.004	0.007	0.098	-0.172	-0.018	0.063	0.050
act8hladr25	0.022	0.120	-0.045	-0.169	-0.095	0.066	0.114	-0.022	-0.087	0.057
act8hladr71	0.001	-0.098	-0.091	-0.073	-0.092	0.099	0.048	0.016	-0.221	-0.026
traff4ccr2	0.038	-0.025	0.040	-0.103	-0.115	-0.083	0.083	0.005	-0.025	0.104
traff4ccr3	0.137	-0.069	0.004	-0.062	-0.127	0.027	-0.089	-0.027	0.059	-0.054
traff4cxcr4	-0.081	0.163	0.061	0.036	-0.022	-0.152	0.053	0.031	0.092	0.048
traff4cxcr5	0.087	-0.130	-0.045	0.095	0.018	-0.035	-0.119	-0.055	-0.001	-0.131
traff4cxcr6	0.017	0.040	0.076	-0.058	-0.067	-0.040	0.028	-0.005	0.106	0.027
traff8ccr2	0.022	0.024	-0.014	-0.048	-0.087	-0.014	0.046	0.032	-0.146	0.066
traff8ccr3	-0.001	-0.020	-0.067	-0.105	0.000	0.127	-0.113	0.005	-0.086	-0.040

traff8cxcr4	-0.149	0.139	-0.033	0.007	0.030	-0.125	0.096	-0.137	0.048	0.022
traff8cxcr5	-0.004	-0.051	-0.122	-0.057	-0.019	0.102	-0.069	-0.058	-0.147	-0.042
traff8cxcr6	0.050	0.018	0.015	0.040	0.008	-0.003	0.103	0.099	0.023	-0.118
traff4ccr10	0.049	-0.013	-0.022	-0.054	-0.048	-0.081	0.049	0.014	-0.005	0.091
traff8ccr10	0.029	-0.041	-0.062	-0.055	-0.027	-0.019	-0.002	0.062	-0.001	0.085

I.4 Random Walk based on Absolute Value of Ranked list



VITA

Hongtai Huang was born and raised in Swatow, China on January 20th, 1987. He received his bachelor degree in Environmental Sciences in Sun Yat-Sen University (SYSU) in 2009. He was a research assistant at the Institute of Environmental Sciences at SYSU in 2010. From the fall of 2010 to the spring of 2011, he served as a student consultant at Johns Hopkins University (JHU). In December 2011, he received his masters degree in Environmental Economics and Management at JHU. His masters research is on Mississippi river delta diversion project design. He coauthored a paper "Cost analysis of water and sediment diversions to optimize land building in the Mississippi River delta" which won the 2013 Water Resources Research Editor's Choice Award. He was recruited as a doctoral student at the department of Geography and Environmental Engineering at JHU in June 2012. His major research interests involve environmental decision making, environmental health and autoimmune disorders.

He was also a mentor for four students in the Women In Sciences and Engineering (WISE) program from 2011 to 2013.