RELAPSE PREDICTION IN CHILDHOOD ACUTE LYMPHOBLASTIC LEUKEMIA BY TIME-SERIES GENE EXPRESSION PROFILING

DIFENG DONG

NATIONAL UNIVERSITY OF SINGAPORE

2011

RELAPSE PREDICTION IN CHILDHOOD ACUTE LYMPHOBLASTIC LEUKEMIA BY TIME-SERIES GENE EXPRESSION PROFILING

DIFENG DONG

(B. COMP., FUDAN UNIVERSITY)

A THESIS SUBMITTED FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

DEPARTMENT OF COMPUTER SCIENCE

NATIONAL UNIVERSITY OF SINGAPORE

2011

ACKNOWLEDGEMENT

First and foremost, I thank my mentor, Prof. Limsoon Wong, for investing huge amount of time in advising my doctoral work. His great support in both spirit and finance allows me to follow my own heart in research and to eventually complete this thesis.

I thank Dario Campana, Elaine Coustan-Smith, Shirley Kham, Yi Lu, and Allen Yeoh for sharing the invaluable data with me.

I thank my friends since college, Su Chen, Dong Guo, Hao Li, Bin Liu, Yingyi Qi, Brian Wang, Vicki Wang, Ning Ye, and Jay Zhuo, for spending good time with me.

I thank my friends, Yexin Cai, Jin Chen, Tsunghan Chiang, Kenny Chua, Mornin Feng, Zheng Han, Chuan Hock Koh, Xiaowei Li, Yan Li, Bing Liu, Guimei Liu, Yuan Shi, Donny Soh, Junjie Wang, Hugo Willy, Lu Yin, and Boxuan Zhai, for sharing happiness with me.

I thank my wife, Peipei, for whatever she has done for me. I would not be able to finish this thesis without her support.

SUMMARY

Childhood acute lymphoblastic leukemia (ALL) is the most common type of cancer in children. Contemporary management of patients with childhood ALL is based on the concept of tailoring the intensity of therapy to a patient's risk of relapse, thereby maximizing the opportunity of cure and minimizing toxic side effects. However, practical protocols of relapse prediction remain imperfect. A significant number of patients with good prognostic characteristics relapse, while some with poor prognostic features survive. There is a demand to improve relapse prediction.

High-throughput gene expression profiling (GEP) has been proved valuable in the diagnosis of childhood ALL. However, its application in relapse prediction falls short on 3 issues: 1) the lack of biological fundamental, 2) the improper selection of computational methodology, and 3) the limited clinical value.

The treatment of childhood ALL is a process to gradually remove the leukemic cells in a patient. GEPs are capable of capturing leukemic genetic signatures in patients. Thus, we hypothesize that a leukemic sample consists of a mixture of leukemic cells and normal cells, where the intensity of the leukemic genetic signature measured by GEP could be used to infer the proportion of leukemic cells in the sample. In addition, as early response is known to have a great prognostic value in childhood ALL, we further expect to perform relapse prediction by the rate of the reduction of leukemic cells during treatment.

To validate our hypothesis, for the first time, we generate time-series GEPs in a leukemia study. We demonstrate that the time-series GEPs are capable of mimicking the removal of

leukemic cells in patients during disease treatment. By modeling our data, we propose to predict the relapses based on the change of GEPs between different time points, which is called genetic status shifting (GSS).

Our relapse prediction results suggest the prognostic strength of GSS is superior to that of any other prognostic factors of childhood ALL, including minimal residual disease (MRD), which is considered as the most powerful relapse predictor among all biological and clinical features tested to date. In our study, GSS outperforms MRD for over 20% in the accuracy of relapse prediction.

In addition, we prove the validity of GSS and its prognostic strength in acute myeloid leukemia (AML), a disease with only 40% of patients survived in 5 years. Our results suggest a new method to improve the prognosis of AML, and thus, probably, to increase the cure rate.

CONTENTS

LIST OF TABLE

LIST OF FIGURE

LIST OF ABBREVIATION

ALL	Acute Lymphoblastic Leukemia
AML	Acute Myeloid Leukemia
CCR	Continuous Complete Remission
DT	Decision Tree
FDR	False Discovery Rate
GEP	Gene Expression Profiling
GO	Gene Ontology
GOEAST	Gene Ontology Enrichment Analysis
GSS	Genetic Status Shifting
IPA	Ingenuity Pathway Analysis
MAS5.0	Affymetrix Microarray Suite 5.0
MRD	Minimal Residual Disease
NB	Naïve Bayes
NBM	Normal Bone Marrow

PC	Principal Component
PCA	Principal Component Analysis
PCR	Polymerase Chain Reaction
RMA	Robust Multiple-Array Average
ROC	Receiver operating characteristic
SAM	Significance Analysis of Microarrays
SVM	Support Vector Machine
TP	Time Point

CHAPTER 1

INTRODUCTION

The emergence of high-throughput gene expression profiling (GEP) allows the measurement of the activity of tens of thousands of genes at once. In the past decade, gene expression analysis is one of the most activated research area in bioinformatics. According to the record of the Gene Expression Omnibus (GEO) repository at the National Center for Biotechnology Information (NCBI), the number of annually published GEP datasets has dramatically increased from 47 in 2001 to 7,079 in 2010 (Figure 1.1) (Edgar, Domrachev and Lash 2002).

The focus of gene expression analysis is cancer, including leukemia (Golub et al. 1999), lymphoma (Alizadeh et al. 2000), melanoma (Bittner et al. 2000), breast cancer (van 't Veer et al. 2002), and others. By exploring the whole genome, a researcher is able to select relevant genes to diagnose a disease (diagnosis) and to predict a disease outcome (prognosis).

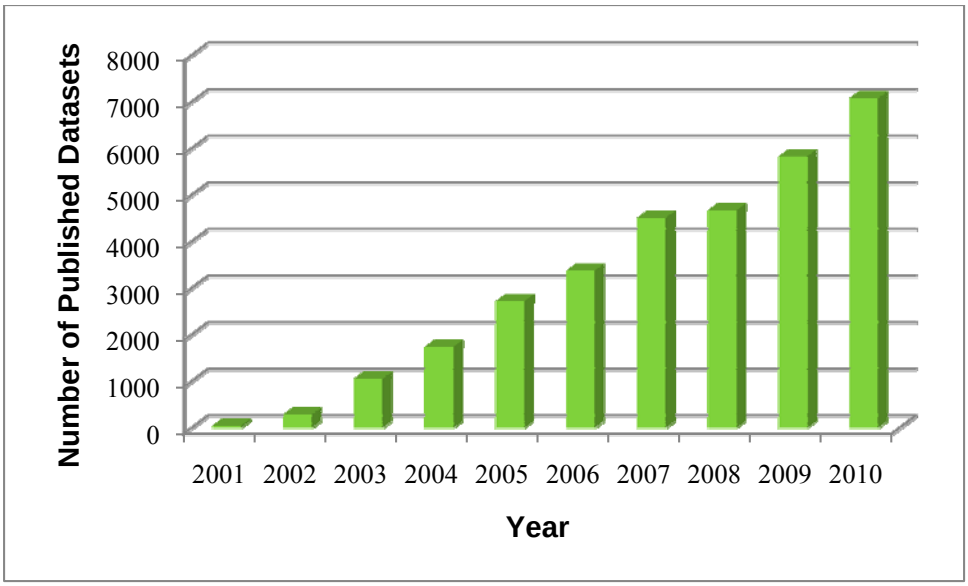


Figure 1.1: The number of annually published GEP datasets in GEO depository at NCBI from 2001 to 2010.

The application of gene expression analysis in the diagnosis of childhood acute lymphoblastic leukemia (ALL) is a successful story. In 2002, Yeoh and colleagues first demonstrate that GEPs can be used to accurately classify patients into 6 subtypes of childhood ALL (Yeoh et al. 2002). Their work is valuable, because the optimal treatment requires the accurate diagnostic subgroup to be upfront assigned to a patient to promise the correct intensity of therapy to be delivered to the patient to maximize the opportunity of cure and to minimize toxic side effects.

In this thesis, we present a recent study of time-series GEPs in childhood ALL. The purpose of the study is: 1) to understand cellular response to the treatment of childhood ALL, and 2) to improve the outcome prediction of the disease.

1.1 Motivation

1.1.1 Clinical Significance

ALL is diagnosed in around 4,000 persons in the United States every year, and two-thirds of them are children and adolescents, making ALL the most common cancer in these age groups (Pui and Evans 2006). ALL is a heterogeneous disease with many subtypes defined by chromosomal translocation. Common subtypes are T-ALL, TEL-AML1, BCR-ABL. E2A-PBX, MLL, and Hyperdiploid>50.

The disease outcome of ALL refers to the long-term event-free survival rate. The overall cure rate of ALL in children is nearly 80%, and about 45%-60% of adult patients have a favorable outcome (Pui and Evans 2006). The major reverse events of ALL are relapse, second malignancy, and death in remission, where relapse is the most common and concerned event (Pui et al. 2005).

Contemporary management of patients with childhood ALL is based on the concept of tailoring the intensity of therapy to a patient's risk of relapse, thereby maximizing the opportunity of cure and minimizing toxic side effects (Pui and Evans 2006, Pui et al. 2005, Pui, Robison and Look 2008). Typically, under treatment causes relapse and eventual death, while over treatment causes long-term damage in intelligence. Thus, to optimize disease outcome, it is important to accurately predict the risk of relapse in childhood ALL patients.

Practical risk classification protocols are based on a number of biological and clinical features, such as, age, blast count, DNA Index, chromosomal abnormality, early morphologic response, and minimal residual disease (MRD) (Pui et al. 2008, Smith et al. 1996, Schultz et al. 2007,

Borowitz et al. 2008). However, these protocols remain imperfect. A significant number of patients with good prognostic characteristics relapse, while some with poor prognostic features survive (Schultz et al. 2007, Sorich et al. 2008, Den Boer et al. 2003). There is a demand to improve relapse prediction.

1.1.2 Research Challenge

GEP is an emerging tool in leukemia diagnosis. The diagnosis of leukemia refers to 1) the confirmation of a leukemia case, and 2) the identification of the subtype of a leukemia case. A recent study, consisting of over 3,000 cases from 11 different laboratories, shows an approximately 95% accuracy in leukemia diagnosis, which has outperformed routine diagnostic methods (Haferlach et al. 2010). The cases of this study cover 6 subtypes of ALL, 6 subtypes of acute myeloid leukemia (AML), chronic lymphocytic leukemia, and chronic myelogenous leukemia, proving the general value of GEPs in leukemia diagnosis.

Nevertheless, the application of GEPs in the relapse prediction of childhood ALL is not very successful. Existing works identify discriminate genetic signatures between relapses and remissions from historical data, and subsequently use the identified signatures to predict new cases (Yeoh et al. 2002, Holleman et al. 2004, Bhojwani et al. 2008, Kang et al. 2010). However, these works fall short on 3 issues:

- **Biological fundamental**. The subtypes of ALL are defined by chromosomal translocation. Each kind of chromosomal translocation may cause a particular type of genetic duplication or deletion, leading to a distinct gene expression pattern from the

normal. Diagnosis by GEP is based on these abnormal gene expression patterns. However, the relationship between gene expression and relapse is still poorly understood. Published works try to explain the mechanisms of relapse by applying function or pathway enrichment analysis over the selected genes in their studies. However, very few of them are convincing and conclusive.

- **Computational methodology**. As illustrated in Figure 1.2, although from the view of clinical science, diagnosis and prognosis are distinctive, the computational toolset to be used are the same. The most commonly used method is supervised learning. Supervised learning makes predictions in new cases by optimizing the parameters of a computational model with historical training data. The predictions are only reliable when the sample size of the training data is large enough. Unfortunately, this is impractical in most GEP datasets. An improper application of supervised learning would cause the acquired parameters to be significantly biased to the batch effects of the training data, and result in prediction failures. In contrast, unsupervised learning targets on classifying cases in a dataset into several subgroups by evaluating the major variance of the data. This process is considered more resistant to the batch effects. It is worthwhile to mention that subtype-related leukemic genetic signatures can be identified by unsupervised learning. However, up to date, there is no reported genetic signature of relapse by unsupervised learning.
- **Clinical value**. MRD has the most prognostic strength among all biological and clinical features tested to date (Pui, Campana and Evans 2001). However, existing GEP studies do not show advantages in relapse prediction when compared to MRD as well as to other prognostic factors.

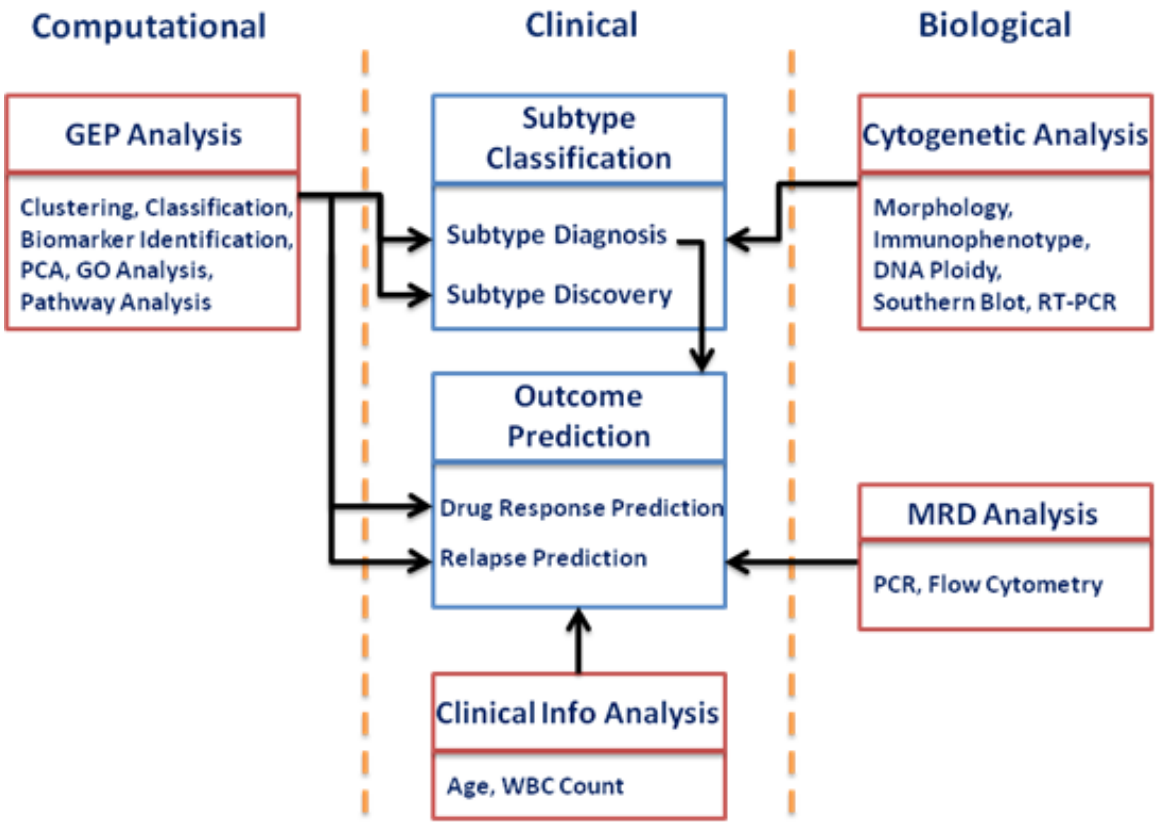


Figure 1.2: A comprehensive overview of childhood ALL diagnosis and prognosis.

1.2 Thesis Contribution

The treatment of childhood ALL is a process to gradually remove the leukemic cells in a patient. GEPs are capable of capturing leukemic genetic signatures in patients. Thus, we hypothesize that a leukemic sample consists of a mixture of leukemic cells and normal cells, where the intensity of the leukemic genetic signature measured by GEP could be used to infer the proportion of leukemic cells in the sample. In addition, as early response is known to have a great prognostic

value, we further expect to perform relapse prediction by the rate of the reduction of leukemic cells during treatment.

Specifically, we conclude our contributions as the following:

- We propose a new testable hypothesis for disease modeling and relapse prediction in childhood ALL.
- We generate the first time-series GEPs in leukemia. The data are collected at the time of diagnosis, and 8 days, 15 days and 33 days after the initial treatment, respectively.
- We confirm the validity of leukemic genetic signatures in our diagnostic GEPs, and demonstrate the dissolution of these signatures during disease treatment.
- We construct the global genetic status shifting (GSS) model based on our time-series GEPs to quantitatively describe the removal of leukemic cells.
- We construct the local GSS models for each of the 6 subtypes to quantitatively describe the removal of leukemic cells in each subtype.
- We design 3 metrics of GSS distance to calculate the rate of the reduction of leukemic cells during treatment, and we predict the relapses by GSS distance.
- We compare GSS-based relapse prediction to other practical prognostic protocols, and illustrate our method performs the best.
- We generate time-series GEPs of 8 AML patients. We validate the concept of GSS and its prognostic strength in this dataset.

1.3 Significance of the Work

We conclude the significances of our work as the following:

- To the best of our knowledge, we are the first to use time-series GEPs in a leukemia study. We have demonstrated that time-series GEPs are capable of mimicking the reduction of leukemic cells during disease treatment.
- To the best of our knowledge, we are the first to predict relapses by unsupervised learning, and the first to make predictions by time-series GEPs. Our relapse prediction results suggest the prognostic strength of GSS is superior to that of any other prognostic factors of childhood ALL, including MRD, which is considered as the most powerful relapse predictor among all biological and clinical features tested to date (Pui et al. 2001). In our study, GSS outperforms MRD for over 20% in the accuracy of relapse prediction.
- We have demonstrated that GSS and its prognostic strength are applicable to AML, a disease with only 40% of patients survived in 5 years (Colvin and Elfenbein 2003). Our results suggest a new method to improve the outcome prediction of AML, and thus, probably, to increase the cure rate.

1.4 Thesis Organization

Chapter 2 provides technical background for gene expression analysis and introduces related works to our study. Chapter 3 gives the details of our patients and the preprocessing of the time-series GEPs. Chapter 4 introduces the computational models constructed for mimicking the

leukemic cell removal. Chapter 5 predicts relapses and compares our method to other prognostic protocols. Chapter 6 validates GSS and its prognostic strength in AML. Chapter 7 summarizes our work and proposes some future works.

CHAPTER 2

RELATED WORK

2.1 Accomplishment of the Past

A successful application of gene expression analysis in childhood ALL is demonstrated by Yeoh and colleagues in 2002 (Yeoh et al. 2002). Childhood ALL has 6 known different subtypes with differing disease outcome. To avoid under treatment, which causes relapse and eventual death, or over treatment, which causes severe long-term side effects, accurate diagnostic subgroup must be assigned upfront so that the correct intensity of therapy can be delivered to ensure that a patient is accorded the highest chance for cure. Contemporary approaches to the diagnosis of childhood ALL use an extensive range of procedures that require multi-specialist expertise, generally unavailable in developing countries. Thus, although childhood ALL is a great success story of modern cancer therapy with survival rates of 75–80% in major advanced hospitals, it is still a fatal disease in developing countries with dismal survival rates of 5–20%.

Table 2.1: Comparing cost and outcome of different treatment strategies.

Treatment	Cost–new cases	Cost–relapses	Total cost
Low-intensity treatment for everyone	\$36K * 2000	\$150K * 1000	\$222M
Intermediate-intensity treatment for everyone	\$60K * 2000	\$150K * 200	\$150M and 50% of patients have side effects
High-intensity treatment for everyone	\$72K * 2000	\$0	\$144M and 90% of patients have side effects
Risk-stratified treatment; viz., low intensity to 50%, intermediate intensity to 40%, high intensity to 10%	\$36K * 1000 + \$60K * 800 + \$72K * 200	\$0	\$98M

As shown in Table 2.1, about 2,000 new cases of childhood ALL are diagnosed in ASEAN countries each year. About 50% of these cases need low-intensity therapy; 40% need intermediate-intensity; and 10% need high-intensity. Treatment for childhood ALL over 2 years for an intermediate-risk patient costs USD 60k; low-risk costs USD 36k; and high-risk costs USD 72k. Treatment for a relapse case costs USD 150k. As the less developed ASEAN countries generally lack the ability to diagnose the subtypes of their childhood ALL patients, the treatment for an intermediate-risk patient is conventionally applied for everyone, since it maximizes the expected benefit in such a situation.

The single-test platform based on gene expression analysis developed by Yeoh and colleagues has an over 96% accuracy in the subtype classification of childhood ALL patients (Yeoh et al. 2002). This can result in savings of USD 52M a year yet with better cure rates and much reduced side effects, as the correct intensity of therapy can be applied upfront.

In addition, Yeoh and colleagues demonstrate that gene expression analysis can be used in discovering new disease subtypes (Yeoh et al. 2002). In their study, they sample 327 childhood ALL patients, where over 60 of them cannot be categorized to any known subtypes. By

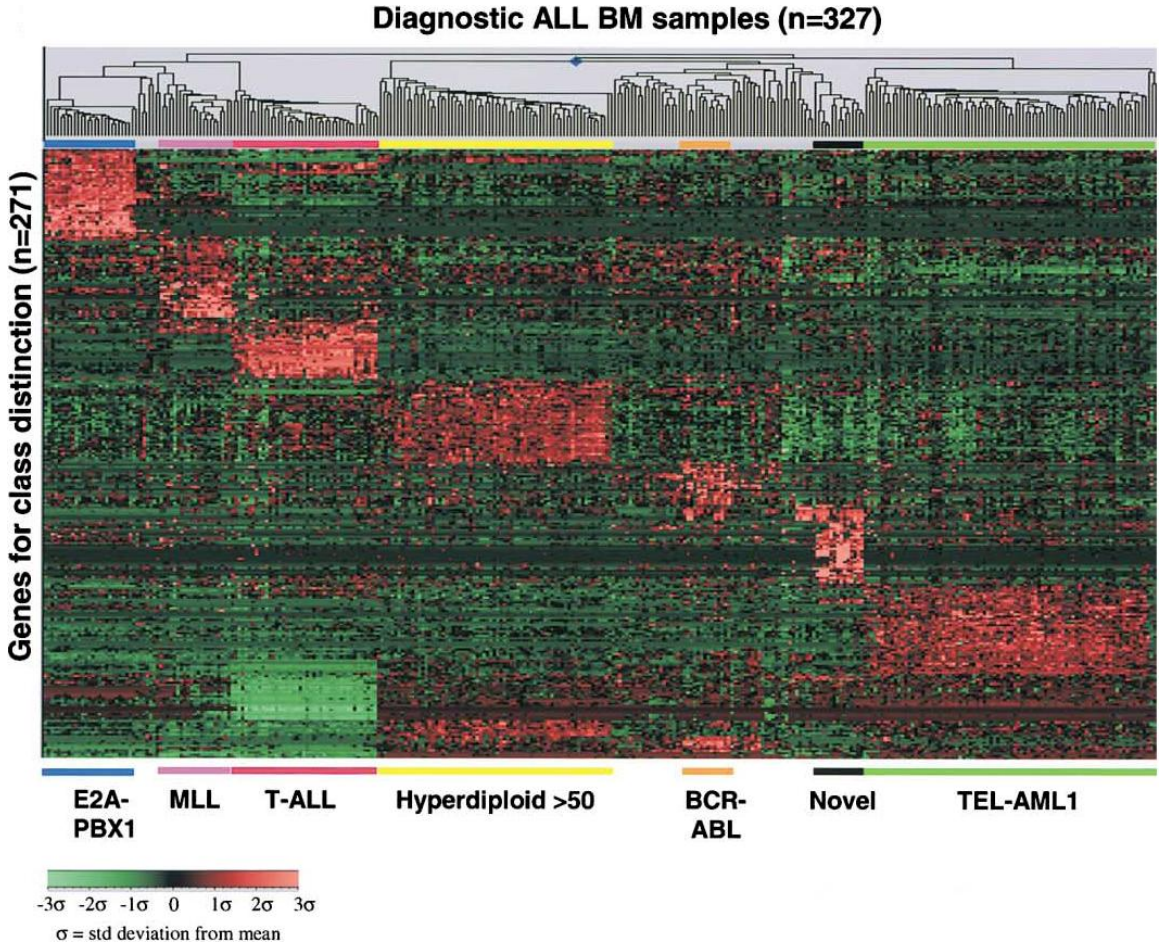


Figure 2.1: The subtype-related leukemic genetic signatures of childhood ALL. Each row is a probe set. Each column is a patient sample. The group of patients, labeled as "Novel", is the newly found subtype. The figure is reproduced from Yeoh et al. 2002.

biclustering analysis, they identify a subgroup, consisting of 14 samples with unknown subtype, shares a novel common distinguishing genetic signature (Figure 2.1). This novel subtype may be linked to lipoma-associated chromosomal translocation.

2.2 Gene Expression Profiling

Gene expression profiling (GEP) refers to the microarray technology, invented in the mid 1990s, that allows monitoring the activity of tens of thousands of genes simultaneously (Schena et al. 1995, Lockhart et al. 1996, Brown and Botstein 1999). Relative quantification of gene expression involves many steps including sample handling, messenger RNA (mRNA) extraction, *in-vitro* reverse transcription, labeling of complementary RNA (cRNA) with fluorescent sequences (probes) which are immobilized on solid surfaces, and the measurement of the intensity of the fluorescent signal which is emitted by the labeled target. The measured signal intensity per target is a measure of relative abundance of the particular mRNA species in the original biological sample (Scherer 2009).

Prevailing microarray platforms are Affymetrix (Santa Clara, CA, USA), Agilent Technologies (Santa Clara, CA, USA), Illumina (San Diego, CA, USA), and Roche Nimblegen (Madison, WI, USA). Even though each platform is designed by a slightly different method, the underlying mechanisms are the same.

To further elucidate the principle of microarray, Figure 2.2 illustrates the design of an Affymetrix GeneChip. The most comprehensive unit in a microarray is called a probe set. Typically, a gene consists of one or several probe sets, with each targeting a different transcriptional region. Each probe set contains about 20 different groups of probe pairs. In each probe pair, there are two typically synthesized 25-mer oligonucleotide probes. The one designed as an exact complement to its target sequence is called a perfect match. The other, designed as the same as the perfect match except for a mutation in the middle position, is called a mismatch.

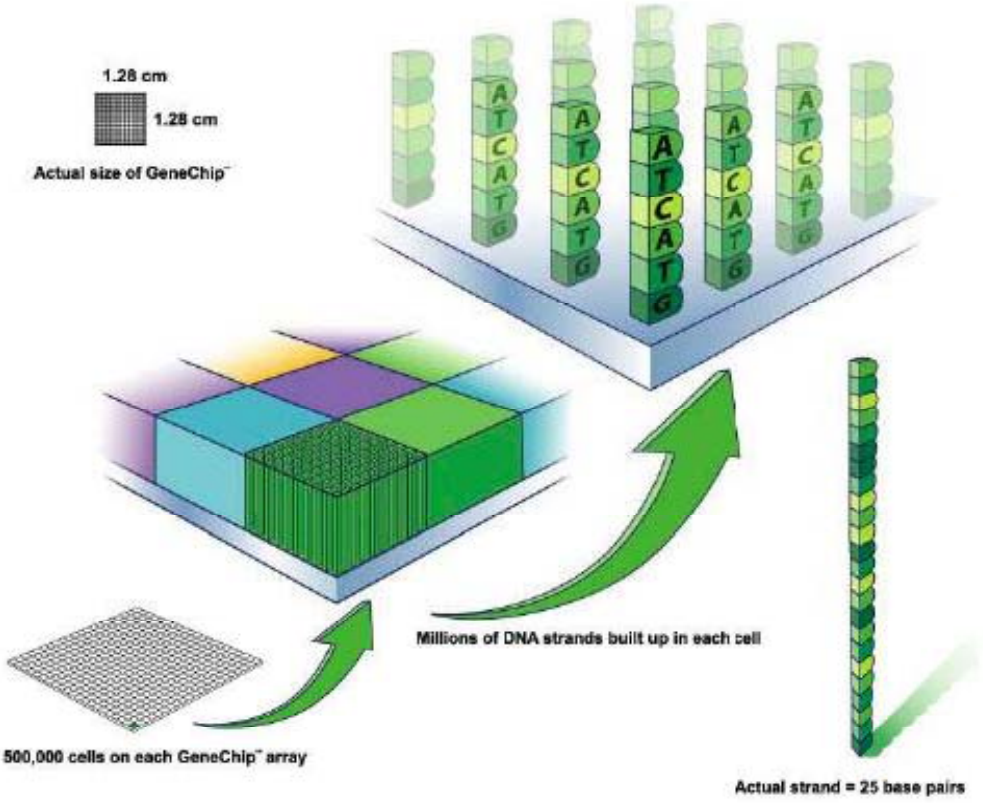


Figure 2.2: Affymetrix GeneChip, reproduced from Affymetrix (Santa Clara, CA, USA).

It is thus expected the perfect match to have a stronger binding affinity to the target sequence, rather than the paired mismatch. In practice, a perfect match is used to estimate the signal intensity, and a mismatch is used to estimate the background noise.

In experiments, long mRNA sequences are degraded into short segments, dyed with fluorescent molecules, and hybridized to a microarray. During the hybridization, once there is enough binding affinity between an mRNA segment and a probe, the mRNA segment will attach to the probe, and the fluorescent molecules on the mRNA segment will lighten its substrate.

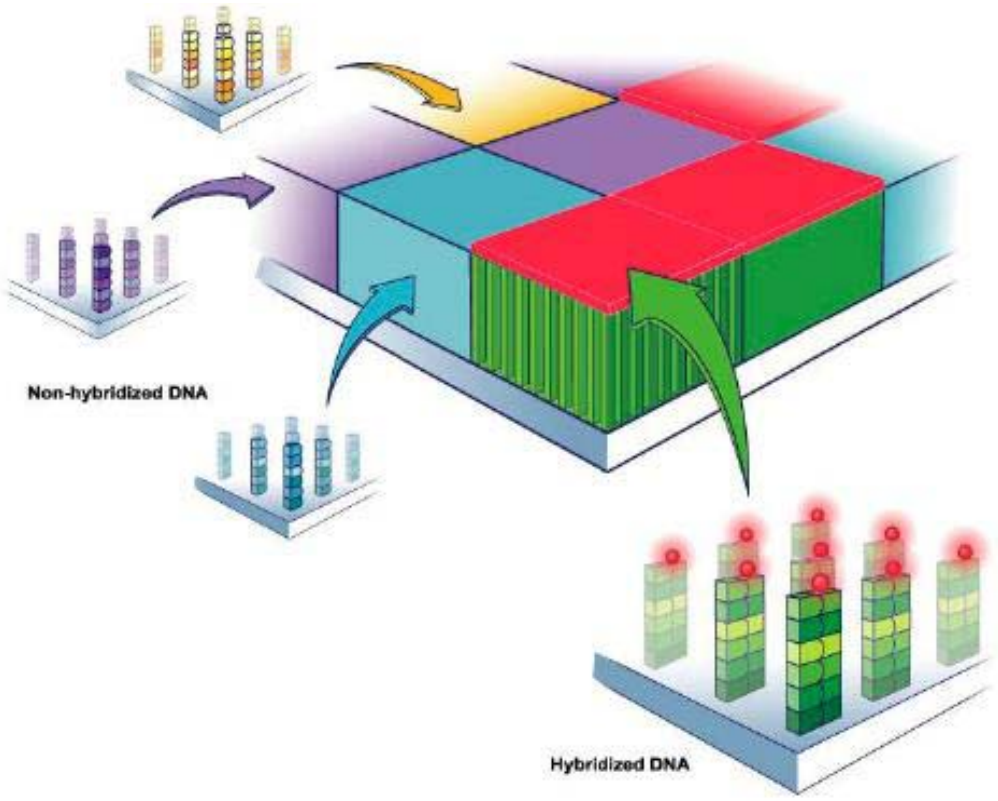


Figure 2.3: GeneChip hybridization, reproduced from Affymetrix (Santa Clara, CA, USA).

When a probe set has many lightened probes, it is considered as an expressed probe set. Figure 2.3 shows such an example. In general, the brighter the overall probe set is, the higher the expression level is.

To quantitatively assess gene expression values, a laser detector is used to scan the fluorescence intensity of each probe in a microarray and the result is saved into a .CEL file. An aggregative algorithm is then applied to each probe set to summarize the signal values of its

corresponding probes. The most popular aggregative algorithms are Affymetrix Microarray Suite 5.0 (MAS5.0) and Robust Multiple-Array Average (RMA) (Irizarry et al. 2003b).

MAS5.0 assumes that every microarray in a batch is independent. In addition to signal values, MAS5.0 also returns detection calls to indicate whether a probe set is present, marginally expressed, or absent. One disadvantage of MAS5.0 is its less sensitive to lowly expressed probe sets. According to the technical report supplied by Affymetrix, MAS5.0 randomly assigns small values to probe sets with "Absent" detection calls (Affymetrix). Recent studies indicate this random assignment strategy is a major source of systematic noise and batch effects (Pepper et al. 2007, Irizarry et al. 2003b, Scherer 2009).

In contrast, RMA makes up the weakness of MAS5.0 by estimating the background from the whole batch of microarrays. This improvement makes RMA much more sensitive to lowly expressed probe sets than MAS5.0 (Irizarry et al. 2003a, Irizarry, Wu and Jaffee 2006). However, the background correction of RMA is not applicable to microarrays hybridized in different machines or at different time. Theoretically, RMA amplifies the difference between different batches of experiments, and therefore refuses the possibility of combining datasets from different studies.

2.3 Subtype Classification

The main approach of leukemia diagnosis is supervised learning, as firstly illustrated by the classic paper of Golub and colleagues (Golub et al. 1999). To apply a supervised learning, GEPs of patients are collected and labeled according to the disease subtypes of the patients. The

analysis then proceeds in a framework of two main steps. In the first step, those genes that are most differentially expressed or most related to a specific subtype are identified. In the second step, a supervised learning algorithm is applied to the genes shortlisted in the first step to induce a classifier. The classifier is then used to predict the subtypes of new cases.

A wide variety of test statistics have been proposed for the first step to select relevant genes, which appears to be the more challenging of the two steps. Initially, classical test statistics such as the *t*-test, χ^2 test, and Wilcoxon rank sum test are used. As the number of genes far exceeds the number of samples in GEP datasets, more elaborate gene selection test statistics are also developed, such as rank products (Breitling and Herzyk 2005) and sparse logistic regression (Cawley and Talbot 2006), as well as techniques for assessing false discovery rates (Qiu and Yakovlev 2006). Integrated methods (Goh and Kasabov 2005, Liu, Li and Wong 2004), typically involving grouping genes with correlated expressions into bins and then selecting representatives from each bin, have also been used. One of the more interesting recent developments in gene selection techniques is to look for gene pairs with expression values that are highly correlated, instead of considering a single gene at a time (Olman et al. 2006). This is a reasonable technique because genes and their products generally function as a group in a specific pathway, and thus their expression values should be correlated.

In 1999, Golub and colleagues firstly propose the two-step framework and demonstrate its feasibility to classify AML and ALL by GEPs (Golub et al. 1999). Briefly, they first do neighborhood analysis to select genes that are uniformly high in one class and uniformly low in the other, and in the second step, they construct their class predictor by the weighted voting of the set of genes selected in the first step. Based on this framework, Golub and colleagues select 50

informative genes most closely correlated with AML-ALL distinction in 38 known samples (27 ALL and 11 AML) during the training stage. The built predictor is then tested in 34 new samples, where 29 of them get strong prediction with 100% accuracy.

This framework is then recruited to make predictions in 6 subtypes of childhood ALL by Yeoh and colleagues (Yeoh et al. 2002). Childhood ALL is a heterogeneous disease caused by chromosomal translocation. Each kind of chromosomal translocation is defined as a disease subtype. Specifically, there are 6 major subtypes, T-ALL, TEL-AML1, E2A-PBX1, BCR-ABL, MLL, and Hyperdiploid>50 (Pui and Evans 2006). Yeoh and colleagues first use the χ^2 statistics to select genes that are most associated with each of the 6 subtypes. They then use a support vector machine (SVM) to learn a classifier for the ALL subtypes from the selected genes. Their classifier achieves an exceedingly overall diagnostic accuracy of 96%. Later, their work is repeated by Ross and colleagues in the same patients but with a different microarray platform (Ross et al. 2003).

Another similar work is performed by Willenbrock and colleagues (Willenbrock et al. 2004). They classify childhood ALL into T-ALL and precursor B-ALL, where precursor B-ALL includes TEL-AML1, E2A-PBX1, BCR-ABL, Hyperdiploid>50 and MLL. Using the same framework, they select 50 most distinguishing genes to train a classifier by several different algorithms, including *k* nearest neighbor, nearest centroid and maximum likelihood. As a result, all of these methods reach 100% accuracy in both training (23 samples) and validation datasets (11 samples).

A recent study, consisting of over 3,000 leukemia cases from 11 different laboratories, shows an approximately 95% accuracy in the diagnosis of leukemia, which has outperformed routine

diagnostic methods (Haferlach et al. 2010). This work includes 6 subtypes of ALL, 6 subtypes of AML, chronic lymphocytic leukemia, and chronic myelogenous leukemia. Haferlach and colleagues follow the same framework as described previously. Specifically, they use the *t*-test to select top 100 differentially expressed probe sets and train an SVM for every pair of the subtypes. Finally, they combine all predictions by maximal voting.

2.4 Outcome Prediction

The two-step framework proposed by Golub and colleagues can be directly applied to predict disease outcome in childhood ALL. This mission is performed by changing the class label from subtype to outcome. There are two types of disease outcome, short-term response and long-term outcome. Short-term response refers to the level of the clearance of leukemic cells in a patient shortly after the initial treatment. Long-term outcome refers to long-time relapse-free survival.

Yeoh and colleagues are the first to predict relapses (Yeoh et al. 2002). They restrict their relapse prediction to only two subtypes, T-ALL and Hyperdiploid>50. For each subtype, they select differentially expressed probe sets between remissions and relapses by the *t*-test, and construct an SVM based on the selected probe sets to make predictions. As a result, they report 100% and 97% accuracy in the relapse prediction of T-ALL and Hyperdiploid>50, respectively.

The same strategy is later repeated by Willenbrock and colleagues in a study consisting of 10 relapses and 18 remissions (Willenbrock et al. 2004). To avoid methodological bias, they apply a panel of gene selection approaches and classifiers to predict the relapses. As a result, Willenbrock and colleagues report an overall accuracy over 75%.

However, both of these two works suffer from strong batch effects, as they use the whole dataset for gene selection, which causes the constructed classifiers to be over fitted to the datasets.

Bhojwani and colleagues identify a 47-probe-set classifier for relapse prediction (Bhojwani et al. 2008). However, the sensitivity of their classifier is only around 64% in the training data. It becomes even lower when the classifier is applied to independent validation datasets.

In a very recent work, Kang and colleagues propose a 38-gene-expression classifier to predict relapses (Kang et al. 2010). They validate their classifier in an independent cohort of 84 patients, where, however, about 50% of the relapses are wrongly predicted.

A second group of works select predictive genes of short-term response, and make use of these genes to predict long-term disease outcome. In practice, this strategy has been realized with different implementations in several different studies.

Holleman and colleagues first identify distinguishing genes between sensitive and resistant to each of the four tested drugs, prednisolone, vincristine, asparaginase, and daunorubicin, by applying the *t*-test to a cohort of 173 childhood ALL patients (Holleman et al. 2004). Then, they construct probabilistic classifiers to predict treatment response based on the genes selected in the first step for each of the four drugs. When a new patient comes, the patient's GEP will be evaluated by these classifiers to estimate the probability of being resistant to each of the four drugs. Finally, these probabilities are combined into a single indicator to predict the risk of relapse of the patient. To show the clinical significance of their method, Holleman and colleagues

validate their work in an independent cohort of 98 patients treated with the same drugs but in a different institute.

This work is later extended by Lugthart and colleagues, where they define cross-resistant and cross-sensitive to be globally resistant and sensitive to the same four drugs (Lugthart et al. 2005). Thereafter, differentially expressed genes are identified to discriminate cross-resistant and cross-sensitive patients. For each patient, the expression values of the selected differentially expressed genes are finally summed up as the indicator of the risk of relapse.

A similar work is carried out by Sorich and colleagues, where only one drug, methotrexate, is used in their study (Sorich et al. 2008).

2.5 Treatment Response Understanding

Some works investigate cellular response to disease treatment by comparing pre- and post-treatment GEPs. A typical process of treatment response understanding consists of two steps. In the first step, differentially expressed genes between pre- and post-treatment GEPs are selected. In the second step, the genes selected in the first step are performed hypergeometric test against the Gene Ontology (Ashburner et al. 2000) or pathway databases to identify the enriched biological processes and molecular functions.

Cheok and colleagues compare diagnostic GEPs and GEPs measured 1 day after treatment. They find drug responsive genes related to apoptosis, mismatch repair, cell cycle control and stress response (Cheok et al. 2003).

Tissing and colleagues compare GEPs of leukemic cells after an 8-hour exposure to glucocorticoids to that of unexposed cells. They identify MAPK pathways, NF-κB signaling and carbohydrate metabolism to be the most affected biological processes (Tissing et al. 2007).

Rhein and colleagues collect paired GEPs on diagnosis and 1 week after treatment. They find drug responsive mechanisms related to the inhibition of cell cycling, and increased expression of adhesion and cytokine receptors (Rhein et al. 2007).

Similar comparisons are conducted between diagnostic and relapsed GEPs to understand the mechanisms of relapse. Staal and colleagues use paired diagnosis-relapse GEPs to find that signaling molecules and transcription factors involved in cell proliferation and cell survival are highly up-regulated at relapse (Staal et al. 2003).

Beesley and colleagues generate GEPs from 11 pairs of diagnostic and relapsed samples, where they find genes of cell growth and proliferation are over expressed in the relapsed samples (Beesley et al. 2005).

Bhojwani and colleagues analyze GEPs in 35 matched diagnosis-relapse pairs and find significant difference in the expression of genes involved in cell-cycle regulation, DNA repair, and apoptosis between the diagnostic and relapsed samples (Bhojwani et al. 2006).

Staal and colleagues analyze 41 matched diagnosis-relapse pairs of ALL patients by GEP. They identify four major gene clusters corresponding to several pathways related to cell cycle regulation, DNA replication, recombination and repair, as well as B-cell development (Staal et al. 2010).

CHAPTER 3

PATIENT AND DATA PREPERATION

3.1 Patient Information

From July 2002 onwards, patients diagnosed as *de novo* childhood ALL are enrolled into the Malaysia-Singapore ALL 2003 trial (MASPORE) at 3 participating centers – National University Hospital (Singapore), University of Malaya Medical Center (Malaysia) and Subang Jaya Medical Center (Malaysia). We study 96 patients from MASPORE. Informed consent is obtained from all patients or their legal guardians in accordance with the Declaration of Helsinki. Both clinical and biological investigations are approved by the responsible review boards at all participating institutes.

Morphological assay and immunophenotyping are performed in the respective laboratories to diagnose subtypes of the patients. Hyperdiploid>50 is determined by either karyotyping or flow cytometry for DNA index ($\geq$1.16). Molecular screening for TEL-AML1, BCR-ABL,

Table 3.1: Patient characteristics in different demographic, prognostic and genotypic groups.

Category	Frequency	Percentage, %
RACE		
Chinese	42	43.7
Malay	38	39.6
Indian & Other	16	16.7
SEX		
Male	46	47.9
Female	50	52.1
AGE		
1-9	70	72.9
<1 or >9	26	27.1
LINEAGE		
B-lineage	84	87.5
T-lineage	12	12.5
WHITE BLOOD CELL		
<50,000	67	69.8
≥50,000	29	30.2
SUBTYPE		
TEL-AML1	26	26.8
BCR-ABL	5	5.2
E2A-PBX1	4	4.2
Hyperdiploid>50	12	12.5
Others	50	51.5
Day-8 RESPONSE		
Good	79	82.3
Poor	16	16.7
Missing	1	1
Day-33 MINIMAL RESIDUAL DISEASE		
<0.01%	49	51.0
0.01-0.1%	24	25.0
0.1-1%	16	16.7
≥1%	5	5.2
Missing	2	2.1
OUTCOME		
Remission	81	84.4
Relapse	13	13.4
Death	2	2.1

E2A-PBX1, and MLL fusions is performed by quantitative real-time PCR. Patient characteristics are summarized in Table 3.1.

3.2 Treatment Response

All patients are treated based on a modified ALL-BFM 2000 backbone and CCG augmented BFM regimen, which includes prednisolone as the major chemotherapeutic agent. High-risk patients (either age <1 or >9, or having leukocyte count $>50\times10^9$ per little at diagnosis) receive additional anthracyclines during the treatment.

The *in vivo* prednisolone response is defined on the day 8 of the treatment by the number of peripheral blood leukemic blasts persisting after a 7-day course of prednisolone treatment plus one intrathecal dose of methotrexate on the first day. The measurement of > 1,000 blasts/μL is considered as slow response. The measurement of > 10,000 blasts/μL is considered as extremely slow response. MRD is assessed on the day 33 by PCR.

3.3 Gene Expression Profiling and Data Preprocessing

Mononuclear cells are separated and harvested from bone marrow aspirates using Ficoll-Paque density gradient centrifugation. Total RNA is isolated using TRIzol reagent and hybridized to Affymetrix HG-U133A (day 0 (D0), n=22; day 8 (D8), n=22; day 15 (D15), n=0; day 33 (D33), n=0) and HG-U133 Plus2.0 (D0, n=74; D8, n=74; D15, n=52; D33, n=60) microarrays (Affymetrix, Santa Clara, CA).

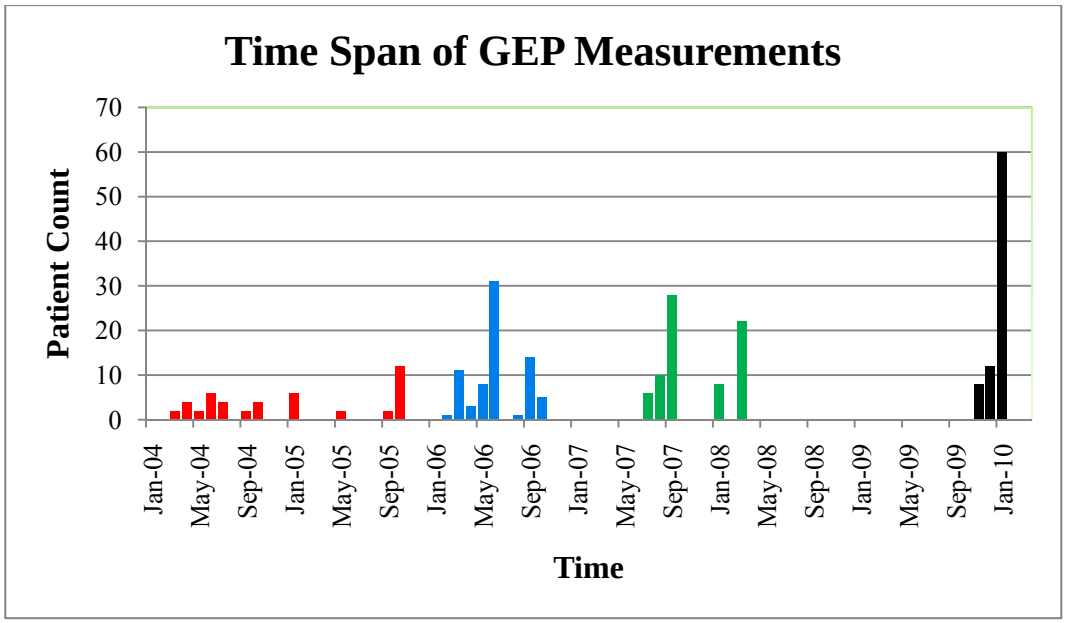


Figure 3.1: The time span of the GEP measurements. GEPs are assigned into four batches, marked with different colors, based on the time of measurement.

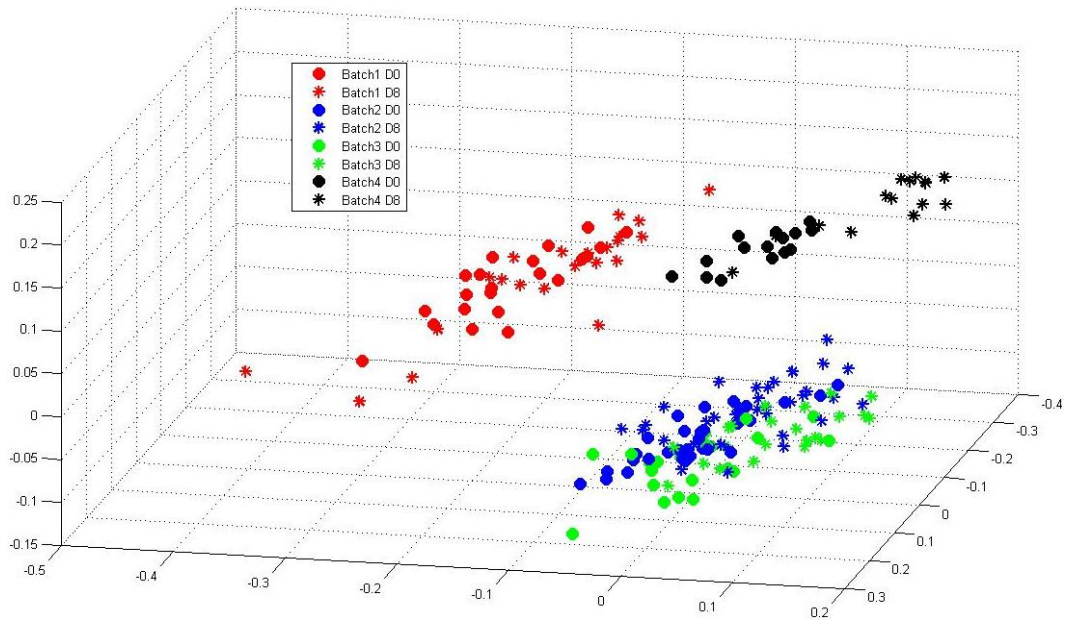


Figure 3.2: The batch effects of our GEPs. The 4 clusters correspond to the 4 batches in Figure 3.1 by color.

Considering two different microarray platforms are used in our study, we only interrogate signal values of the probe sets shared in both platforms by MAS5.0. Detection values ("Present", "Marginal" or "Absent") are determined by default parameters and signal values are scaled to a median of 500 in each microarray.

Although RMA is known to be more sensitive to low expressions, there are two reasons we use MAS5.0. First, HG-U133A contains a subset of probe sets of HG-U133 Plus 2.0. In order to make expressions of the two platforms compatible, we have to ignore the probe sets only in HG-U133 Plus 2.0. MAS5.0 allows us to mask the unusable probe sets, so we can achieve our purpose without any post-interrogation processing. However, RMA does not allow users to select a subset of probe sets during signal interrogation. If we use RMA, we have to remove the extra probe sets in HG-U133 Plus 2.0 after the interrogation and unify signal distributions of the two platforms. This is undesired, as any extra data manipulation could introduce extra systematic biases.

Second, RMA assumes the interrogated GEPs belonging to the same batch. However, as shown in Figure 3.1, the measurements of our GEPs span nearly 6 years. These GEPs should be considered as in different batches. In Figure 3.1, we assign our GEPs (D0 and D8) into 4 batches based on the time of the measurements. We then plot them into a 3-dimensional space calculated by principal component analysis (PCA). From Figure 3.2, we find that the 4 batches of samples actually form 4 distinct clusters, with each cluster following the same pattern of the separation between D0 and D8 samples. This result suggests that our data have significant batch effects, and thus RMA is not suitable to our data.

We design a three-step protocol to remove the batch effects. First, microarrays, whose scaling factor is larger than 20, are excluded, due to the over degradation of mRNA. As a result, 290 samples (D0, *n*=92; D8, *n*=90; D15, *n*=49; D33, *n*=59) are eligible for the next stage of data processing.

Second, MAS5.0 randomly assigns small signal values to "Absent" probe sets, which composes a major source of batch effects (Pepper et al. 2007, Affymetrix , Irizarry et al. 2003b, Scherer 2009). We thus remove lowly expressed probe sets. A probe set is retained only if it has a "Present" call in more than 30% of our samples at any of the 4 time points. As a result, 14,736 probe sets pass the filtration.

Finally, signal values of the remaining probe sets are transformed into 2-based logarithm scale and normalized by quantile normalization (Bolstad et al. 2003). Quantile normalization assumes each microarray to have the same signal distribution. It is a reasonable assumption, because the microarray technology is based on the assumption that the whole gene expressions of a sample follow a normal distribution (Slonim 2002). By performing quantile normalization, GEPs from different batches will be adjusted to follow the same distribution (Figure 3.3).

To perform quantile normalization, we first combine all probe sets in all microarrays as a reference distribution. For each microarray, we compute for each value, the quantile of that value in the distribution of the microarray. These quantiles are then transformed into the corresponding signal values according to the reference distribution. The whole process is shown in Figure 3.4, which can be described as:

$$x' = F_R^{-1}(F_i(x));$$

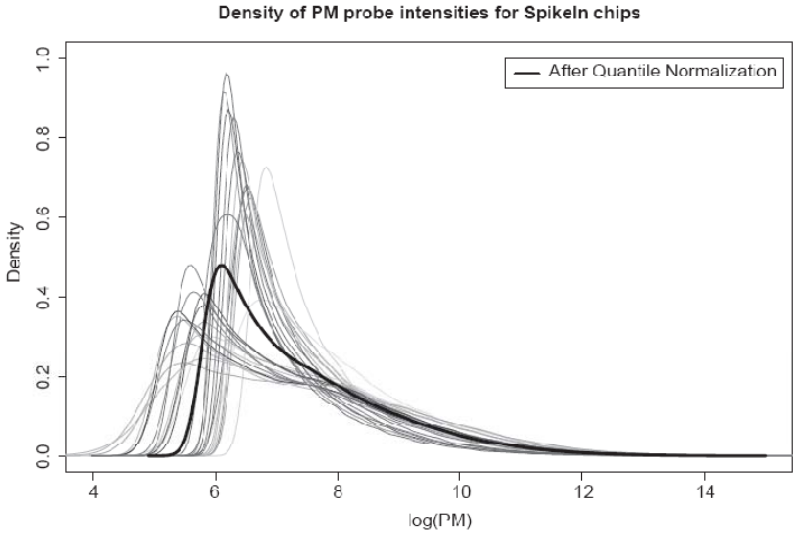


Figure 3.3: An example of quantile normalization, reproduced from Bolstad et al. 2003.

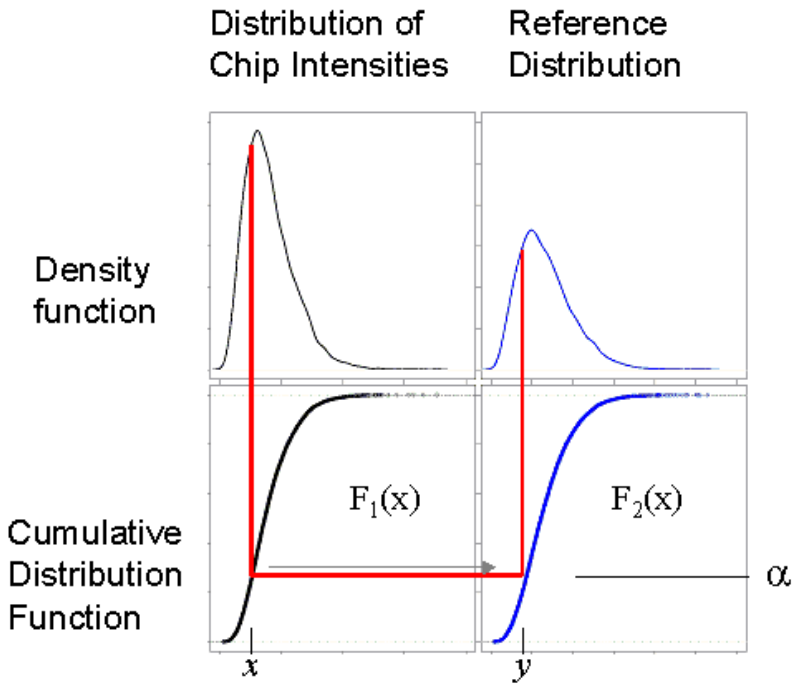


Figure 3.4: The process of quantile normalization.

where x is the original value and x' is the normalized value; F_i and F_R are the cumulative distribution function of the i-th microarray and the reference distribution, respectively.

Figure 3.5 shows the resulted distributions of our samples after quantile normalization. The curves are consistent with each other, and the curve of the reference distribution is close to a normal distribution.

Again, we plot our D0 and D8 samples by PCA in Figure 3.6. The result indicates that the batch effects we observe previously have been successfully removed.

3.4 Validation Dataset

Several published datasets are used to validate our study. They are the St. Jude Children's Research Hospital's dataset (SJCRH, *n*=132) (Ross et al. 2003), the Dutch Childhood Oncology Group's dataset (DCOG, *n*=107) (Den Boer et al. 2009), another Dutch Childhood Oncology Group's dataset (DCOG2, *n*=41) (Staal et al. 2010), the German Cooperative ALL's dataset (COALL, *n*=190) (Den Boer et al. 2009), and the Collaborative Microarray Innovations in Leukemia's dataset (MILE-Diagnose, *n*=750; MILE-NBM (normal bone marrow), *n*=73; MILE-AML, *n*=74) (Haferlach et al. 2010). All datasets are consistently processed by our GEP preprocessing protocol.

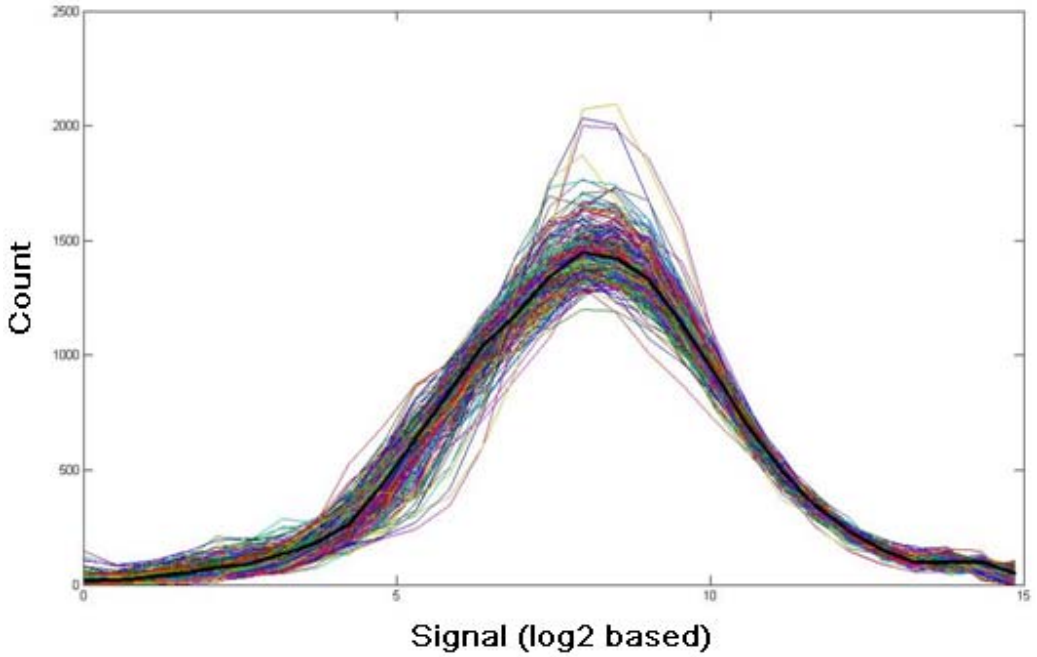


Figure 3.5: The gene expression distributions after quantile normalization. The black bold curve in the middle is the reference distribution.

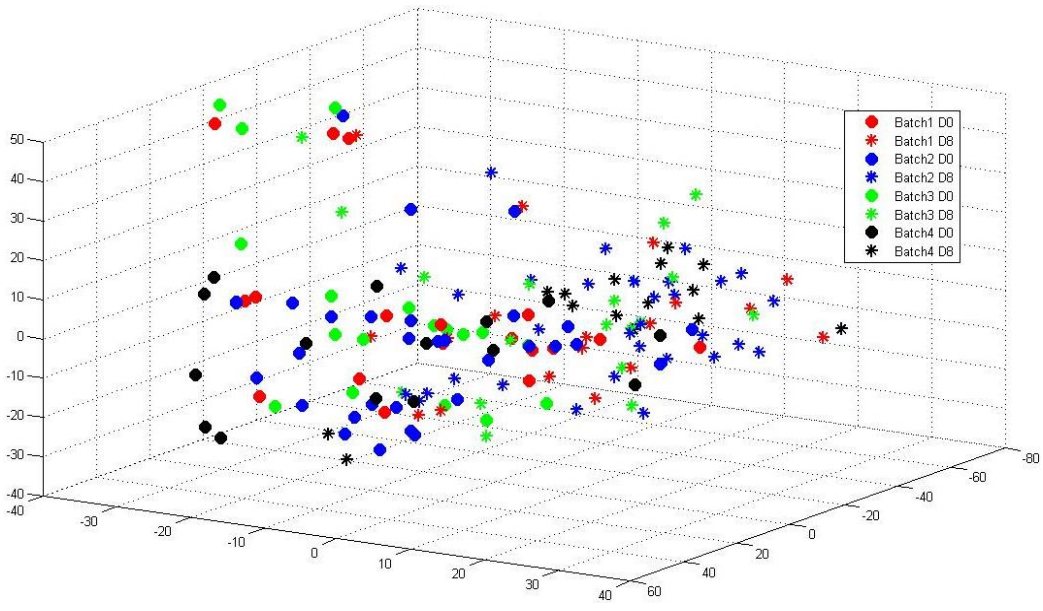


Figure 3.6: GEPs after the batch effects removing.

CHAPTER 4

GENETIC STATUS SHIFTING MODEL

4.1 Overview

The treatment of childhood ALL can be typically divided into two phases: 1) remission induction, and 2) consolidation therapy. The goal of remission induction is to eradicate more than 99% of the initial burden of leukemia cells in a patient and to restore normal hematopoiesis. Recent research indicates that 98% of patients can achieve a complete remission after the first stage of treatment (Pui and Evans 2006). When normal hematopoiesis is restored, patients in remission become candidates for consolidation therapy, which usually lasts for about four weeks. The purpose of consolidation therapy is to remove the remained leukemic cells in a patient and to prevent the patient from rapid relapse.

The treatment of childhood ALL is a process to gradually remove the leukemic cells in a patient. GEPs are capable of capturing leukemic genetic signatures in patients. Thus, we

hypothesize that a leukemic sample consists of a mixture of leukemic cells and normal cells, where the intensity of the leukemic genetic signature measured by GEP could be used to infer the proportion of leukemic cells in the sample. To validate this hypothesis, we generate time-series GEPs to investigate the relationship between GEPs and the removal of leukemic cells.

4.2 Unsupervised Hierarchical Clustering

Unsupervised hierarchical clustering creates a hierarchy of clusters, represented in a tree structure, called a dendrogram. The root of the tree is a single cluster containing all samples, and the leaves correspond to individual samples.

There are two important parameters in unsupervised hierarchical clustering, similarity and linkage. Similarity refers to the distance metric between two clusters. Euclid distance and Pearson's correlation are the common selections in gene expression analysis. Linkage specifies the way that similarity is calculated between two clusters. Candidates include single linkage, complete linkage, average linkage, and centroid linkage. Single linkage takes the similarity of the nearest samples between two clusters as the similarity of the two clusters. Complete linkage takes the similarity of the farthest samples between two clusters as the similarity of the two clusters. Average linkage averages the similarities of all possible pairs of samples between two clusters as the similarity of the two clusters. Centroid linkage calculates the centroid of each cluster first, and then calculates the distance between the two centroids as the similarity of the two clusters.

We apply unsupervised hierarchical clustering to our time-series GEPs. The algorithm is performed by Eisen's software, Cluster 3.0, with Pearson's correlation as the similarity and

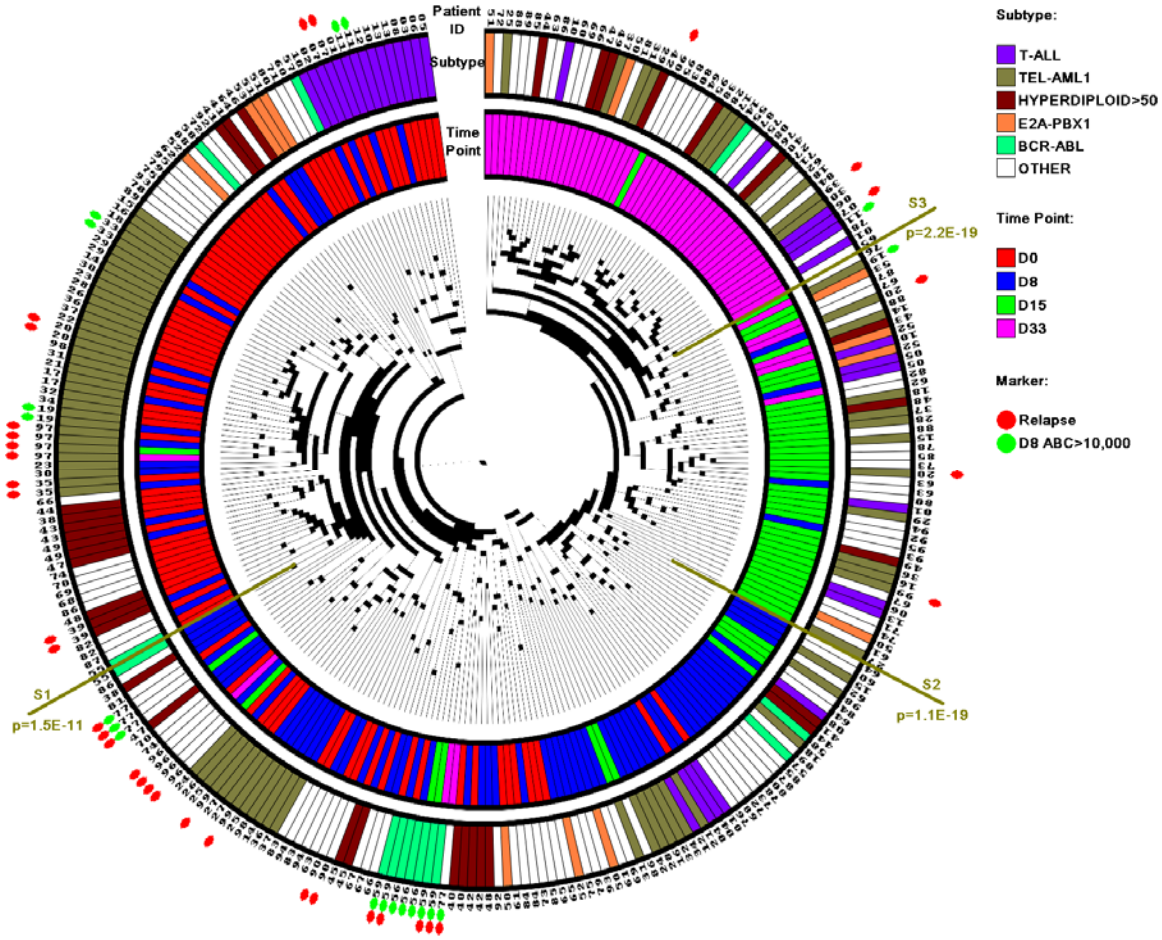


Figure 4.1: Unsupervised hierarchical clustering. The inner-loop units indicate the time points. The outer-loop units indicate the subtypes. Extremely slow responders (D8 blast count > 10,000 per μL) are marked in green. Relapses are marked in red. S1, S2 and S3 are the identified optimal boundaries to separate the samples of D0 and D8, D8 and D15, and D15 and D33, respectively.

complete linkage as the linkage (Eisen et al. 1998). To minimize the impact of systematic biases, which are mainly contained in low expressions, we only use top 10% of probe sets with the largest variance across the whole dataset (n = 1,474).

Figure 4.1 shows the result of unsupervised hierarchical clustering. In the figure, each sample corresponds to an inner-loop unit, indicating the time point, and an outer-loop unit, indicating the subtype. We emphasize two important observations. First, the samples collected on the same time point tend to be clustered together. We find our samples are organized in the order of D0 → D8 → D15 → D33 by unsupervised hierarchical clustering. To quantitatively describe the significance of the observation and to find out the optimal boundaries between the adjacent time points, we evaluate all possible positions by Fisher's exact test. As highlighted in Figure 4.1, our results suggest the separation between the time points is statistically significant (D0 to D8, $p = 1.5\times10^{-11}$; D8 to D15, $p = 1.1\times10^{-19}$; D15 to D33, $p = 2.2\times10^{-19}$).

Second, the diagnostic GEPs are clustered by subtype. Actually, this discovery has been reported before (Yeoh et al. 2002, Ross et al. 2003). Specifically, in the region of the D0 samples, there is a T-ALL cluster, a TEL-AML1 cluster, a Hyperdiploid>50 cluster and a BCR-ABL cluster. Moreover, we find that D8 GEPs are clustered by subtype as well. There are 2 TEL-AML1 clusters, a Hyperdiploid>50 cluster and a BCR-ABL cluster. However, when compared to D0 clusters, D8 clusters are smaller and sparser. This is probably due to the dilution of leukemic genetic signatures resulted from treatment. In contrast, we fail to identify any nontrivial subtype clusters in the region of D15 or D33 samples.

4.3 Genetic Signature Dissolution Analysis

The result of unsupervised hierarchical clustering suggests that leukemic genetic signatures are gradually removed during treatment. We thus design experiments to validate this hypothesis.

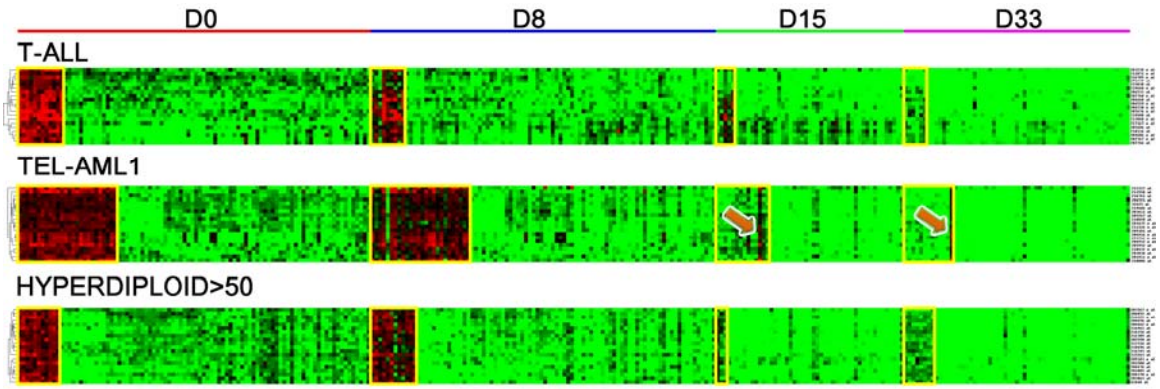


Figure 4.2: Leukemic genetic signatures are dissolved into the background during treatment. Red represents high expression. Green represents low expression. Yellow frames highlight the patients of the targeted subtype. The arrows indicate a relapse case.

We select leukemic genetic signature genes from the diagnostic GEPs of the 3 largest subtypes in our data (T-ALL, $n = 12$; TEL-AML1, $n = 26$; and Hyperdiploid>50, $n = 12$). For each subtype, we categorize the samples into two groups, belonging (Group 1) and not belonging (Group 2) to the subtype. We only consider a probe set if its expression in Group 1 is higher than in Group 2 (compared by the averaged expression of the two groups). We then calculate the *t*-statistics between the two groups, and select top 20 differentially expressed probe sets, ranked by the *p* value, as the leukemic genetic signature genes. The selected probe sets are listed in Table 4.1 to Table 4.3.

To examine the correctness of the leukemic genetic signature genes, we use them to predict the subtypes of samples of MILE-diagnosis. The sensitivity and specificity for T-ALL are 94.83% and 99.82%. The sensitivity and specificity for TEL-AML1 are 91.38% and 98.53%. The sensitivity and specificity for Hyperdiploid>50 are 85% and 97.13%. Thus, we show the identified leukemic genetic signature genes are reliable.

The result of genetic signature dissolution analysis is shown in Figure 4.2. From the figure, the signatures gradually dissolve into the background. A patient of TEL-AML1 subtype, who suffers from a relapse, shows a resistant signature during the course.

Table 4.1: Genetic signature genes of T-ALL.

T-ALL				
Probe Set	**Gene Symbol**	**Gene Title**	**Fold Change**	***p* Value**
213060_s_at	CHI3L2	chitinase 3-like 2	49.56	1.06E-22
210116_at	SH2D1A	SH2 domain protein 1A	53.47	1.16E-17
205674_x_at	FXYD2	FXYD domain containing ion transport regulator 2	11.07	4.98E-16
216705_s_at	ADA	adenosine deaminase	3.59	6.88E-14
202760_s_at	PALM2-AKAP2	PALM2-AKAP2 readthrough transcript	5.14	1.37E-13
217147_s_at	TRAT1	T cell receptor associated transmembrane adaptor 1	22.34	4.90E-13
202747_s_at	ITM2A	integral membrane protein 2A	18.60	9.24E-12
203238_s_at	NOTCH3	Notch homolog 3 (Drosophila)	7.98	1.13E-11
202746_at	ITM2A	integral membrane protein 2A	15.67	1.23E-10
211071_s_at	MLLT11	myeloid/lymphoid or mixed-lineage leukemia (trithorax homolog, Drosophila); translocated to, 11	2.94	1.00E-09
205484_at	SIT1	signaling threshold regulating transmembrane adaptor 1	11.25	1.42E-09
215030_at	GRSF1	G-rich RNA sequence binding factor 1	2.42	1.44E-09
209604_s_at	GATA3	GATA binding protein 3	12.04	2.10E-08
206533_at	CHRNA5	cholinergic receptor, nicotinic, alpha 5	3.27	3.04E-08
204529_s_at	TOX	thymocyte selection-associated high mobility group box	4.91	3.15E-08
219408_at	PRMT7	protein arginine methyltransferase 7	2.67	1.44E-07
204639_at	ADA	adenosine deaminase	2.63	4.94E-07
204530_s_at	TOX	thymocyte selection-associated high mobility group box	3.64	1.41E-06
206460_at	AJAP1	adherens junctions associated protein 1	8.07	3.15E-06
219660_s_at	ATP8A2	ATPase, aminophospholipid transporter-like, class I, type 8A, member 2	3.92	9.56E-06

Table 4.2: Genetic signature genes of TEL-AML1.

TEL-AML1				
Probe Set	**Gene Symbol**	**Gene Title**	**Fold Change**	***p* Value**
213317_at	CLIC5	chloride intracellular channel 5	176.36	6.53E-38
213558_at	PCLO	piccolo (presynaptic cytomatrix protein)	74.41	2.07E-23
218804_at	ANO1	anoctamin 1, calcium activated chloride channel	22.53	9.83E-20
205952_at	KCNK3	potassium channel, subfamily K, member 3	28.77	8.64E-18
203611_at	TERF2	telomeric repeat binding factor 2	7.41	1.92E-16
204914_s_at	SOX11	SRY (sex determining region Y)-box 11	65.05	2.88E-16
203038_at	PTPRK	protein tyrosine phosphatase, receptor type, K	48.06	5.37E-16
204915_s_at	SOX11	SRY (sex determining region Y)-box 11	30.62	1.59E-15
201911_s_at	FARP1	FERM, RhoGEF (ARHGEF) and pleckstrin domain protein 1 (chondrocyte-derived)	7.75	4.02E-15
204913_s_at	SOX11	SRY (sex determining region Y)-box 11	50.85	1.23E-14
206591_at	RAG1	recombination activating gene 1	8.34	5.39E-11
218820_at	C14orf132	chromosome 14 open reading frame 132	7.86	1.68E-08
209101_at	CTGF	connective tissue growth factor	14.76	8.10E-08
205267_at	POU2AF1	POU class 2 associating factor 1	2.96	9.92E-08
210432_s_at	SCN3A	sodium channel, voltage-gated, type III, alpha subunit	14.13	2.24E-07
32625_at	NPR1	natriuretic peptide receptor A/guanylate cyclase A (atrionatriuretic peptide receptor A)	4.70	8.45E-07
211126_s_at	CSRP2	cysteine and glycine-rich protein 2	3.52	3.24E-06
214761_at	ZNF423	zinc finger protein 423	4.48	5.02E-06
203435_s_at	MME	membrane metallo-endopeptidase	5.20	2.31E-05
219686_at	STK32B	serine/threonine kinase 32B	4.60	2.45E-05

Table 4.3: Genetic signature genes of Hyperdiploid>50.

Hyperdiploid>50				
Probe Set	**Gene Symbol**	**Gene Title**	**Fold Change**	***p* Value**
201508_at	IGFBP4	insulin-like growth factor binding protein 4	5.21	2.49E-05
203063_at	PPM1F	protein phosphatase 1F (PP2C domain containing)	2.45	7.46E-05
206674_at	FLT3	fms-related tyrosine kinase 3	4.43	8.91E-05
214745_at	PLCH1	phospholipase C, eta 1	3.46	2.32E-04
218694_at	ARMCX1	armadillo repeat containing, X-linked 1	3.68	2.50E-04
208370_s_at	RCAN1	regulator of calcineurin 1	2.79	5.82E-04
201005_at	CD9	CD9 molecule	5.87	6.64E-04
202598_at	S100A13	S100 calcium binding protein A13	2.17	8.41E-04
207267_s_at	DSCR6	Down syndrome critical region gene 6	2.60	8.47E-04
41660_at	CELSR1	cadherin, EGF LAG seven-pass G-type receptor 1 (flamingo homolog, Drosophila)	2.91	1.05E-03
215263_at	ZXDA /// ZXDB	zinc finger, X-linked, duplicated A /// zinc finger, X-linked, duplicated B	1.72	1.05E-03
204462_s_at	SLC16A2	solute carrier family 16, member 2 (monocarboxylic acid transporter 8)	4.20	1.47E-03
206852_at	EPHA7	EPH receptor A7	4.56	1.51E-03
204454_at	LDOC1	leucine zipper, down-regulated in cancer 1	4.05	2.23E-03
214961_at	KIAA0774	KIAA0774	5.24	2.43E-03
209183_s_at	C10orf10	chromosome 10 open reading frame 10	3.60	2.71E-03
214156_at	MYRIP	myosin VIIA and Rab interacting protein	5.31	3.20E-03
211626_x_at	ERG	v-ets erythroblastosis virus E26 oncogene homolog (avian)	2.13	3.38E-03
213316_at	KIAA1462	KIAA1462	4.22	3.44E-03
212385_at	TCF4	transcription factor 4	2.81	3.60E-03

4.4 Genetic Status Shifting Model

We present computational models to quantitatively describe the reduction of leukemic cells in patients during treatment, which are called genetic status shifting (GSS) models. In our results, we construct the global GSS model to include all our samples, and the local GSS models for each of the subtypes.

The construction of a GSS model consists of two steps. First, drug responsive genes are selected. Second, principal component analysis (PCA) is applied to the selected genes.

In gene expression analysis, PCA is a mathematical procedure that uses an orthogonal transformation to convert a set of expressions of possibly correlated genes into a set of values of uncorrelated variables called principal components (PCs). A PC is a linear combination of the original genes. In PCA, the resulted PCs are ranked by the contained data variance. Typically, although the number of PCs can be as many as the number of the original genes, the first several PCs include the most variance of a dataset.

4.4.1 Drug Responsive Gene

We identify drug responsive genes for the global GSS model by selecting differentially expressed probe sets between the D0 and D8 samples. The selection is taken with two actions. First, the *t*-test implemented in the Significance Analysis of Microarrays (SAM) is applied with the threshold of false discovery rate (FDR), $q < 0.0001$ (Tusher, Tibshirani and Chu 2001, Storey and

Tibshirani 2003). Second, any probe set to be considered as a differentially expressed probe set should show at least a 2-fold change (either up regulation or down regulation) in the averaged expression of the D8 samples when compared to that of the D0 samples.

A total number of 562 and 123 probe sets, representing 461 and 99 genes, are considered as up- and down-regulated differentially expressed probe sets by our criteria. Table 4.4 and 4.5 list the 20 most up- and down-regulated probe sets, ranked by the fold change, respectively.

To biologically understand drug responsive mechanisms, we evaluate the drug responsive genes by Gene Ontology Enrichment Analysis (GOEAST) (Zheng and Wang 2008), and Ingenuity Pathway Analysis (IPA, Ingenuity® Systems, www.ingenuity.com). For GOEAST, we use $p < 0.01$ as the threshold of selecting significant Gene Ontology (GO) terms. For IPA, we use $p < 0.01$ as the threshold of selecting significant pathways, biological functions and biological networks.

The 20 most significant GO terms, ranked by the p value, are listed in Table 4.6 and 4.7 for the up- and down-regulated probe sets, respectively. The up-regulated terms include biological functions related to the reconstruction of immune system and the restoration of normal hematogenesis, such, immune system process ($p = 1.99\times10^{-90}$), leukocyte activation ($p = 8.39\times10^{-21}$), hemoglobin complex ($p = 7.96\times10^{-18}$), and blood circulation ($p = 3.40\times10^{-15}$). The down-regulated terms include two categories. The first category involves the cell development and DNA synthesis, such as, cellular developmental process ($p = 2.40\times10^{-9}$), cell differentiation ($p = 4.53\times10^{-9}$), DNA packaging ($p = 7.69\times10^{-9}$), DNA binding ($p = 1.23\times10^{-7}$), and chromatin assembly ($p = 2.36\times10^{-7}$). The second category involves the negative regulation of apoptosis, such as, negative regulation of thymocyte apoptosis ($p = 2.55\times10^{-7}$), negative regulation of T cell

apoptosis ($p = 4.51\times10^{-7}$), negative regulation of lymphocyte apoptosis ($p = 7.19\times10^{-6}$), and negative regulation of mature B cell apoptosis ($p = 1.53\times10^{-5}$).

As to IPA, the up- and down-regulated probe sets are combined for the analysis. Table 4.8 lists the significant canonical signaling pathways. The significant biological functions are listed in Table 4.9. Both results are consistent with that of GOEAST. For example, communication between innate and adaptive immune cells ($p = 4.17\times10^{-8}$), primary immunodeficiency signaling ($p = 2.04\times10^{-7}$), B cell development ($p = 2.34\times10^{-7}$), inflammatory response ($p = 3.51\times10^{-37}$), hematological system development and function ($p = 2.73\times10^{-21}$), and hematopoiesis ($p = 5.93\times10^{-12}$), are related to the reconstruction of immune system and the restoration of normal hematogenesis. Another set of pathways and functions, such as cell-to-cell signaling and interaction ($p = 1.82\times10^{-26}$), cellular growth and proliferation ($p = 1.66\times10^{-18}$), cellular development ($p = 7.38\times10^{-15}$), cellular assembly and organization ($p = 7.98\times10^{-6}$), DNA replication, recombination, and repair ($p = 1.41\times10^{-5}$), and gene expression ($p = 4.91\times10^{-4}$), are related to the cell development and DNA synthesis. Results, such as, cytotoxic T lymphocyte-mediated apoptosis of target cells ($p = 9.77\times10^{-5}$), and cell death ($p = 2.36\times10^{-13}$) are related to the regulation of apoptosis.

In addition, the top 5 biological networks identified by IPA are shown in Figure 4.3 to 4.7. These networks are mainly related to cancer, inflammatory response, cell-to-cell signaling and interaction, cell death, cellular development, and cell cycle, which are consistent with the previous results of GOEAST and IPA.

Table 4.4: Top 20 up-regulated probe sets.

Probe Set ID	Gene Symbol	Gene Title	Fold Change
205950_s_at	CA1	carbonic anhydrase I	7.395791184
205403_at	IL1R2	interleukin 1 receptor, type II	6.852324058
205997_at	ADAM28	ADAM metallopeptidase domain 28	6.708182719
214146_s_at	PPBP	pro-platelet basic protein (chemokine (C-X-C motif) ligand 7)	6.658931873
211372_s_at	IL1R2	interleukin 1 receptor, type II	6.466121916
203645_s_at	CD163	CD163 molecule	6.117093533
215049_x_at	CD163	CD163 molecule	6.087834242
205837_s_at	GYPA /// GYPB	glycophorin A (MNS blood group) /// glycophorin B (MNS blood group)	5.822822642
201110_s_at	THBS1	thrombospondin 1	5.745576019
211821_x_at	GYPA	glycophorin A (MNS blood group)	5.209992602
212768_s_at	OLFM4	olfactomedin 4	5.182515342
209555_s_at	CD36	CD36 molecule (thrombospondin receptor)	5.064436979
217388_s_at	KYNU	kynureninase (L-kynurenine hydrolase)	5.046071877
211560_s_at	ALAS2	aminolevulinate, delta-, synthase 2	5.008170532
206488_s_at	CD36	CD36 molecule (thrombospondin receptor)	4.920133289
215646_s_at	VCAN	Versican	4.917441199
210746_s_at	EPB42	erythrocyte membrane protein band 4.2	4.903890326
221731_x_at	VCAN	Versican	4.832463599
217418_x_at	MS4A1	membrane-spanning 4-domains, subfamily A, member 1	4.811721663
206390_x_at	PF4	platelet factor 4	4.801380171

Table 4.5: Top 20 down-regulated probe sets.

Probe Set ID	Gene Symbol	Gene Title	Fold Change
210487_at	DNTT	deoxynucleotidyltransferase, terminal	0.222299534
209035_at	MDK	midkine (neurite growth-promoting factor 2)	0.224017113
211341_at	POU4F1	POU class 4 homeobox 1	0.235935714
203434_s_at	MME	membrane metallo-endopeptidase	0.256940866
206660_at	IGLL1	immunoglobulin lambda-like polypeptide 1	0.258262938
215117_at	RAG2	recombination activating gene 2	0.259848868
207030_s_at	CSRP2	cysteine and glycine-rich protein 2	0.283315663
206067_s_at	WT1	Wilms tumor 1	0.304693921
214243_s_at	SERHL /// SERHL2	serine hydrolase-like /// serine hydrolase-like 2	0.315189369
219740_at	VASH2	vasohibin 2	0.316965896
203435_s_at	MME	membrane metallo-endopeptidase	0.321395315
205755_at	ITIH3	inter-alpha (globulin) inhibitor H3	0.327868482
206591_at	RAG1	recombination activating gene 1	0.337911256
204165_at	WASF1	WAS protein family, member 1	0.339940607
219218_at	BAHCC1	BAH domain and coiled-coil containing 1	0.340474114
205795_at	NRXN3	neurexin 3	0.360422562
208950_s_at	ALDH7A1	aldehyde dehydrogenase 7 family, member A1	0.36198217
209983_s_at	NRXN2	neurexin 2	0.36527127
207426_s_at	TNFSF4	tumor necrosis factor (ligand) superfamily, member 4	0.365433513
213668_s_at	SOX4	SRY (sex determining region Y)-box 4	0.36606881

Table 4.6: Top 20 GO terms for the up-regulated probe sets.

GO ID	Ontology	Term	p-value
GO:0002376	biological_process	immune system process	1.99E-90
GO:0006955	biological_process	immune response	4.58E-77
GO:0050896	biological_process	response to stimulus	3.70E-75
GO:0006952	biological_process	defense response	4.14E-67
GO:0005886	cellular_component	plasma membrane	1.23E-63
GO:0005488	molecular_function	Binding	9.26E-55
GO:0005576	cellular_component	extracellular region	1.05E-52
GO:0006950	biological_process	response to stress	4.48E-51
GO:0016020	cellular_component	Membrane	2.34E-45
GO:0044459	cellular_component	plasma membrane part	2.18E-42
GO:0005623	cellular_component	Cell	4.52E-41
GO:0044464	cellular_component	cell part	4.52E-41
GO:0004871	molecular_function	signal transducer activity	4.59E-40
GO:0060089	molecular_function	molecular transducer activity	4.59E-40
GO:0004872	molecular_function	receptor activity	8.74E-40
GO:0009611	biological_process	response to wounding	9.33E-37
GO:0005887	cellular_component	integral to plasma membrane	3.06E-35
GO:0005515	molecular_function	protein binding	4.29E-35
GO:0031226	cellular_component	intrinsic to plasma membrane	4.37E-35
GO:0002682	biological_process	regulation of immune system process	3.76E-34

Table 4.7: Top 20 GO terms for the down-regulated probe sets.

GO ID	Ontology	Term	p-value
GO:0032502	biological_process	developmental process	2.89E-16
GO:0007275	biological_process	multicellular organismal development	1.03E-13
GO:0048731	biological_process	system development	8.56E-13
GO:0032501	biological_process	multicellular organismal process	1.67E-12
GO:0048856	biological_process	anatomical structure development	2.94E-12
GO:0005623	cellular_component	Cell	9.80E-11
GO:0044464	cellular_component	cell part	9.80E-11
GO:0048869	biological_process	cellular developmental process	2.40E-09
GO:0030154	biological_process	cell differentiation	4.53E-09
GO:0009987	biological_process	cellular process	5.46E-09
GO:0006323	biological_process	DNA packaging	7.69E-09
GO:0005488	molecular_function	binding	2.29E-08
GO:0071103	biological_process	DNA conformation change	3.17E-08
GO:0045112	biological_process	integrin biosynthetic process	4.29E-08
GO:0065007	biological_process	biological regulation	1.17E-07
GO:0050821	biological_process	protein stabilization	1.19E-07
GO:0003677	molecular_function	DNA binding	1.23E-07
GO:0016043	biological_process	cellular component organization	1.38E-07
GO:0031497	biological_process	chromatin assembly	2.36E-07
GO:0070244	biological_process	negative regulation of thymocyte apoptosis	2.55E-07

Table 4.8: Significant pathways for the differentially expressed probe sets between D8 and D0.

Ingenuity Canonical Pathways	***p* value**
Communication between Innate and Adaptive Immune Cells	4.17E-08
Primary Immunodeficiency Signaling	2.04E-07
B Cell Development	2.34E-07
Dendritic Cell Maturation	1.38E-06
Atherosclerosis Signaling	1.95E-06
IL-10 Signaling	1.17E-05
Hepatic Fibrosis / Hepatic Stellate Cell Activation	1.66E-05
Crosstalk between Dendritic Cells and Natural Killer Cells	1.95E-05
Systemic Lupus Erythematosus Signaling	4.47E-05
Graft-versus-Host Disease Signaling	6.61E-05
Cytotoxic T Lymphocyte-mediated Apoptosis of Target Cells	9.77E-05
TREM1 Signaling	0.000158
LXR/RXR Activation	0.000178
Autoimmune Thyroid Disease Signaling	0.000219
Allograft Rejection Signaling	0.000316
Role of Macrophages, Fibroblasts and Endothelial Cells in Rheumatoid Arthritis	0.000372
Altered T Cell and B Cell Signaling in Rheumatoid Arthritis	0.000457
T Helper Cell Differentiation	0.000457
Complement System	0.001047
Ascorbate and Aldarate Metabolism	0.001122
IL-6 Signaling	0.001479
Glycolysis/Gluconeogenesis	0.001479
Lipid Antigen Presentation by CD1	0.002455
Airway Pathology in Chronic Obstructive Pulmonary Disease	0.002512
Histidine Metabolism	0.002754
Granzyme A Signaling	0.00309
Type I Diabetes Mellitus Signaling	0.004169
Caveolar-mediated Endocytosis Signaling	0.004467
OX40 Signaling Pathway	0.00631
LPS/IL-1 Mediated Inhibition of RXR Function	0.008128
Nur77 Signaling in T Lymphocytes	0.008511
IL-8 Signaling	0.00912
CCR5 Signaling in Macrophages	0.009772

Table 4.9: Significant biological functions for the differentially expressed probe sets between D8 and D0.

Category	***p* value**
Inflammatory Response	3.51E-37-1.4E-03
Infectious Disease	1.12E-31-2.19E-04
Respiratory Disease	1.12E-31-1.4E-03
Cell-To-Cell Signaling and Interaction	1.82E-26-1.4E-03
Hematological System Development and Function	2.73E-21-1.4E-03
Immune Cell Trafficking	7.56E-21-1.4E-03
Tissue Development	5.56E-19-1.4E-03
Cellular Growth and Proliferation	1.66E-18-1.13E-03
Cancer	3.54E-16-1.4E-03
Inflammatory Disease	6.95E-16-1.4E-03
Connective Tissue Disorders	9.75E-16-3.05E-04
Immunological Disease	9.75E-16-1.4E-03
Skeletal and Muscular Disorders	9.75E-16-3.05E-04
Cellular Development	7.38E-15-1.4E-03
Cellular Movement	5.77E-14-1.4E-03
Cell Death	2.36E-13-1.4E-03
Cell Signaling	4.21E-12-1.06E-03
Molecular Transport	4.21E-12-7.43E-04
Vitamin and Mineral Metabolism	4.21E-12-1.06E-03
Hematopoiesis	5.93E-12-9.7E-04
Cardiovascular System Development and Function	7.64E-12-8.95E-04
Dermatological Diseases and Conditions	8.62E-12-1.1E-05
Genetic Disorder	8.62E-12-1.4E-03
Cellular Function and Maintenance	1.03E-11-1.4E-03
Hematological Disease	1.96E-11-1.4E-03
Reproductive System Disease	5.13E-11-1.24E-03
Antigen Presentation	7.7E-09-4.1E-04
Cell-mediated Immune Response	1.39E-08-1.3E-03
Cellular Compromise	3.31E-08-2.22E-04
Cell Morphology	7.13E-08-1.4E-03
Gastrointestinal Disease	2.45E-07-8.49E-05
Lymphoid Tissue Structure and Development	2.47E-07-4.91E-04
Neurological Disease	1.12E-06-1.17E-03

Cardiovascular Disease	2.83E-06-1.09E-03
Cellular Assembly and Organization	7.98E-06-1.4E-03
Lipid Metabolism	8.5E-06-7.33E-04
Small Molecule Biochemistry	8.5E-06-7.33E-04
Post-Translational Modification	1.38E-05-1.09E-03
DNA Replication, Recombination, and Repair	1.41E-05-1.95E-05
Renal and Urological Disease	1.9E-05-1.9E-05
Infection Mechanism	2.2E-05-2.2E-05
Tumor Morphology	2.62E-05-9.56E-04
Organismal Injury and Abnormalities	3.31E-05-1.4E-03
Carbohydrate Metabolism	3.47E-05-4.91E-04
Free Radical Scavenging	3.54E-05-1.3E-03
Hypersensitivity Response	4.75E-05-1.4E-03
Tissue Morphology	9.97E-05-9.56E-04
Skeletal and Muscular System Development and Function	1.65E-04-1.4E-03
Ophthalmic Disease	2.02E-04-1.4E-03
Organismal Functions	2.1E-04-2.1E-04
Organismal Development	2.24E-04-8.95E-04
Antimicrobial Response	2.56E-04-2.56E-04
Humoral Immune Response	2.56E-04-1.06E-03
Renal and Urological System Development and Function	4.02E-04-4.02E-04
Connective Tissue Development and Function	4.91E-04-9.82E-04
Gene Expression	4.91E-04-1.4E-03
Drug Metabolism	6.26E-04-6.26E-04
Embryonic Development	7.53E-04-1.4E-03
Metabolic Disease	8.95E-04-8.95E-04
Nervous System Development and Function	9.56E-04-9.56E-04
Hair and Skin Development and Function	1.24E-03-1.24E-03
Protein Trafficking	1.4E-03-1.4E-03

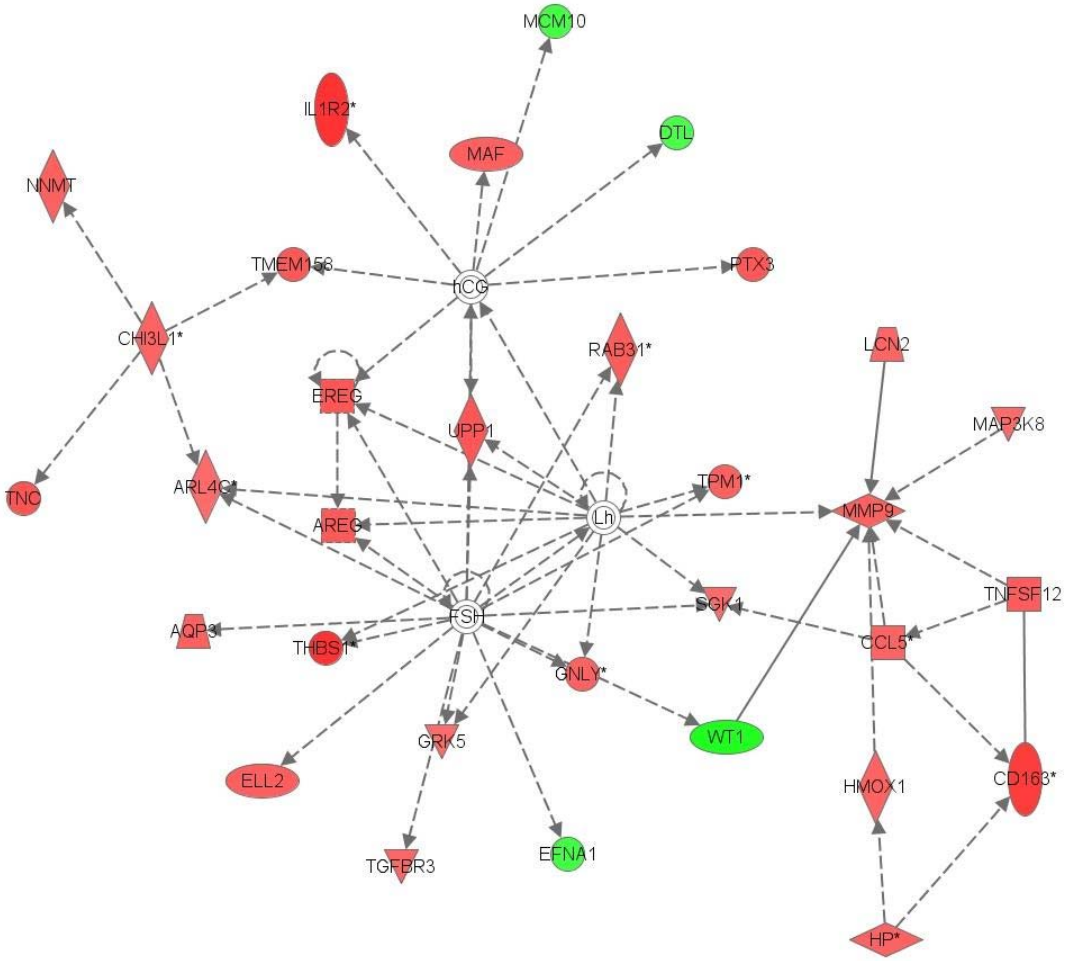


Figure 4.3: The top biological network, cancer, inflammatory response, and cell-to-cell signaling and interaction.

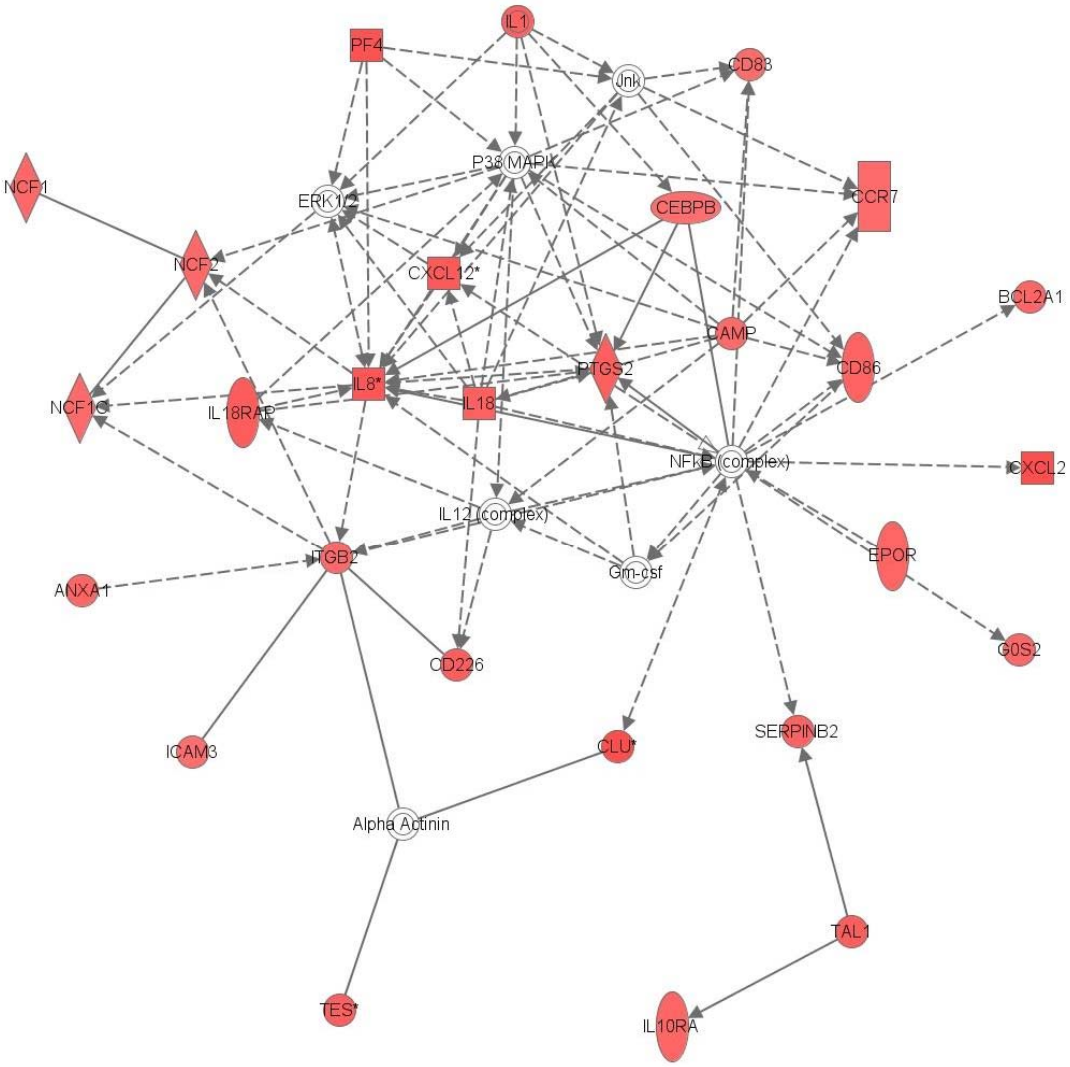


Figure 4.4: The second top biological network, inflammatory response, cell death, and cell-to-cell signaling and interaction.

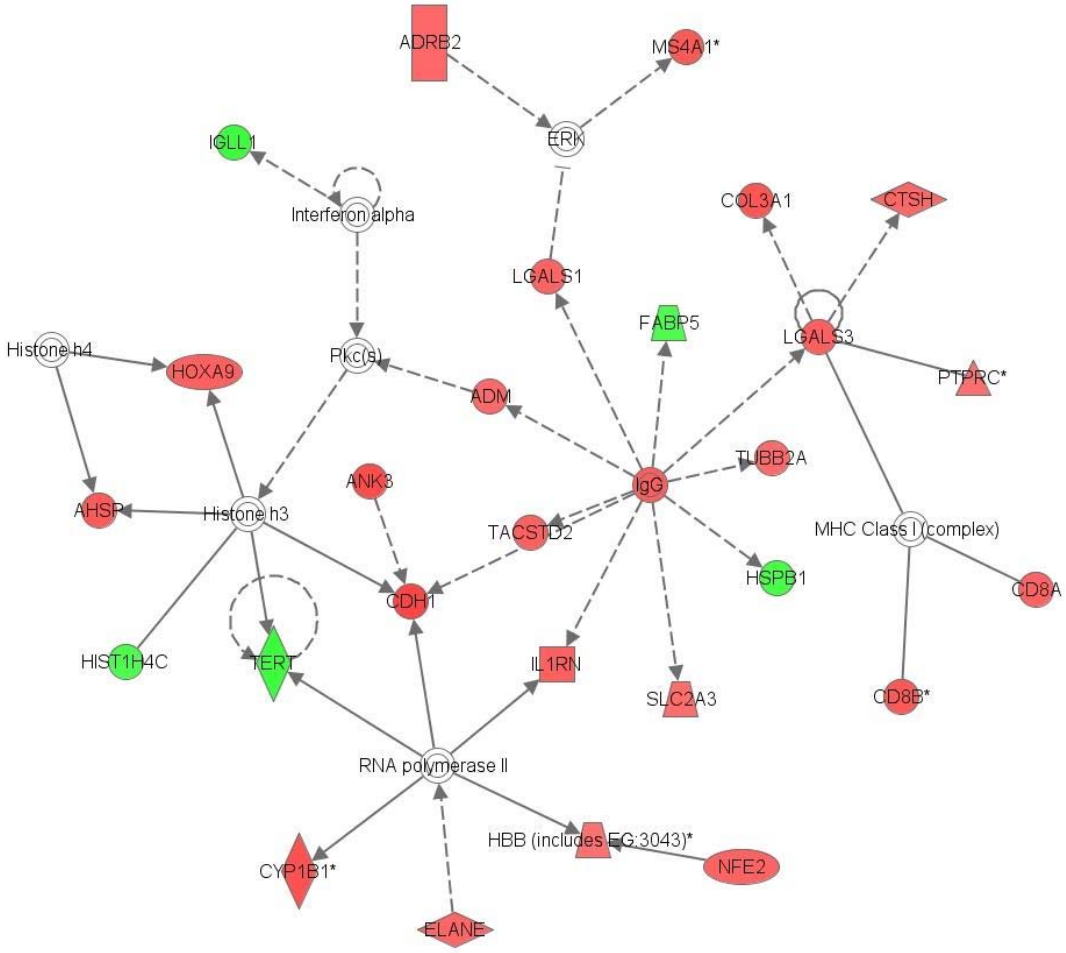


Figure 4.5: The third top biological network, cancer, respiratory disease, and cellular development.

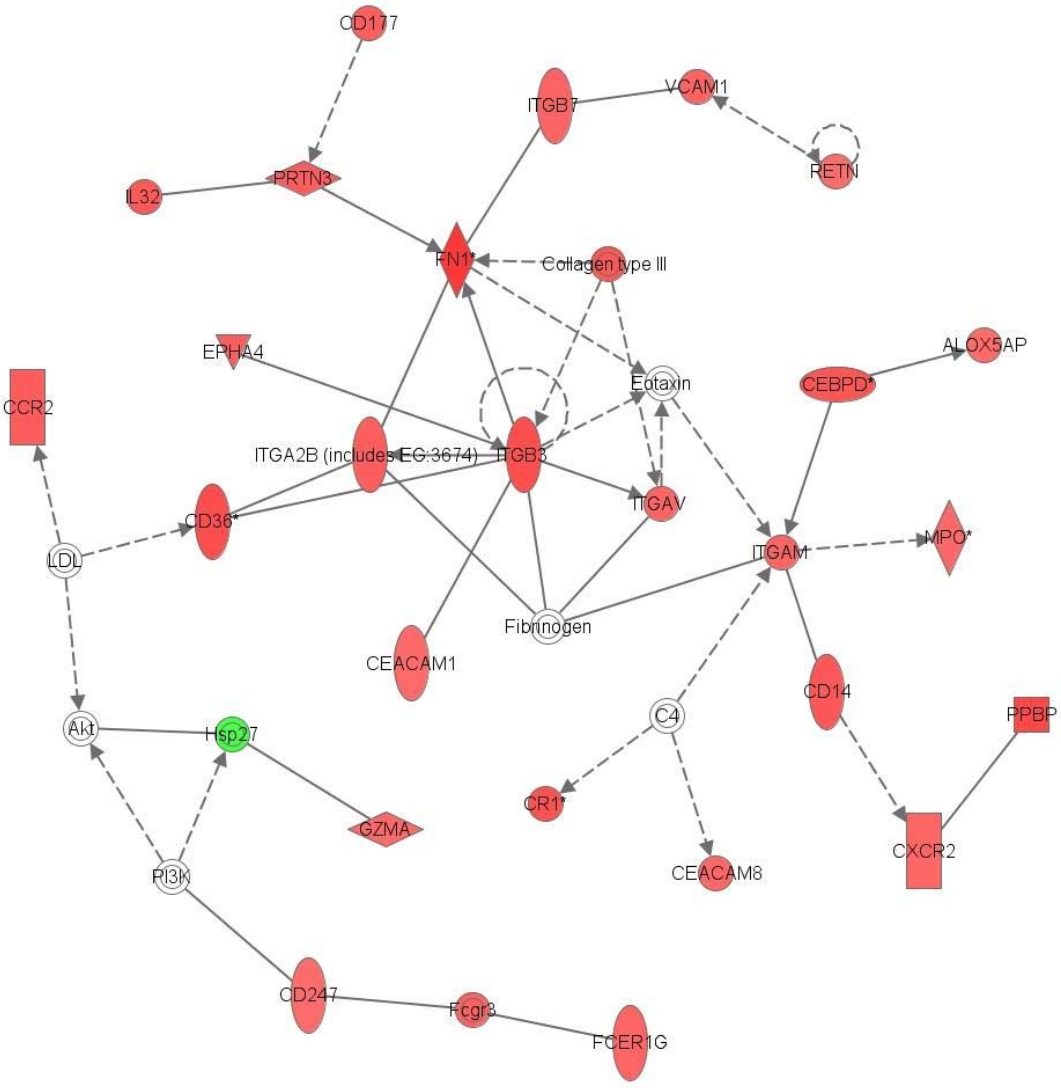


Figure 4.6: The fourth top biological network, cell-to-cell signaling and interaction, tissue development, and cellular movement.

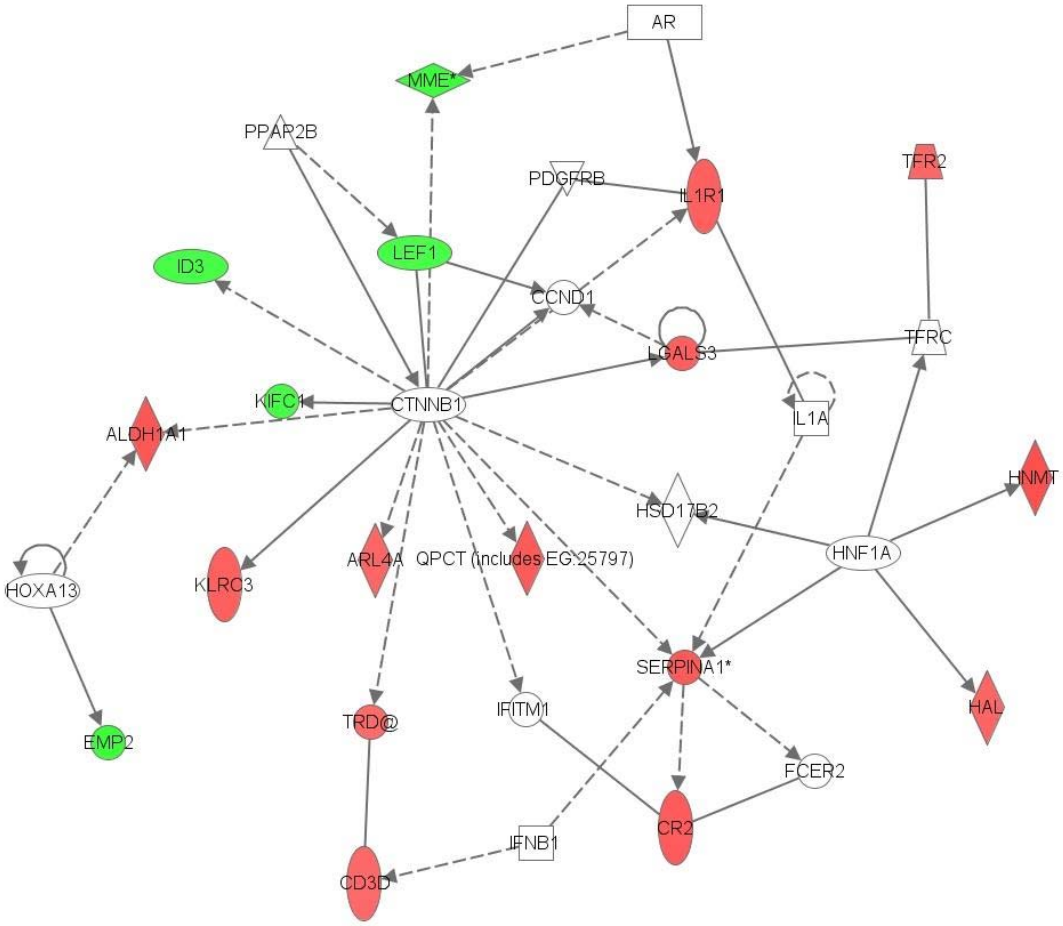


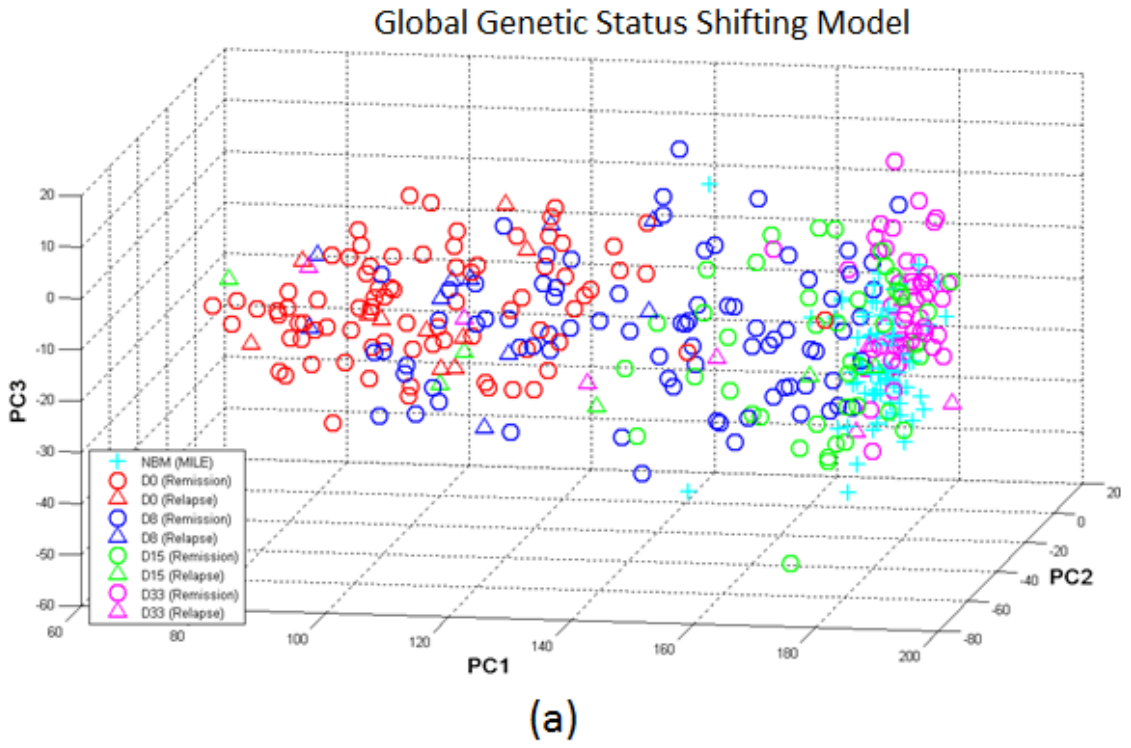
Figure 4.7: The fifth top biological network, cancer, gastrointestinal disease, and cell cycle.

4.4.2 Global Genetic Status Shifting Model

The global GSS model is constructed by applying PCA to the selected drug responsive genes. Figure 4.8 shows the global GSS model, determined by the first 3 PCs. In Figure 4.8, from left to right, samples are aligned in the order of D0 → D8 → D15 → D33. This observation is not unexpected, as we have learned it from the result of unsupervised hierarchical clustering. The first PC contains nearly 50% of the variance, probably reflecting the different loads of leukemic cells during the course.

We recruit a set of normal samples, MILE-NBM, to further evaluate the locations of our samples in the model. As shown in Figure 4.8, MILE-NBM samples collocate with the D33 samples, extending the transition pattern to D0 → D8 → D15 → D33 → Normal. This is an exciting discovery, because it suggests that the transition pattern we have observed in the global GSS model is meaningful, as it actually indicates the process of the removal of leukemic cells, by which patients eventually achieve remissions.

In addition, several diagnostic GEP datasets are compared to our model. The result of SJCRH is shown in Figure 4.9. The result of DCOG is shown in Figure 4.10. The result of DCOG2 is shown in Figure 4.11. The result of COALL is shown in Figure 4.12. The result of MILE-Diagnose is shown in Figure 4.13. In conclusion, samples of these datasets collocate well with our D0 samples.



(a)

PC	1	2	3	4	5	6	7	8	Total
Variance	49.08%	7.56%	5.19%	3.64%	2.10%	1.81%	1.61%	1.35%	72.34%

(b)

Figure 4.8: The global GSS model and its variance distribution. (a) The global GSS model. (b) The variance contained in top PCs.

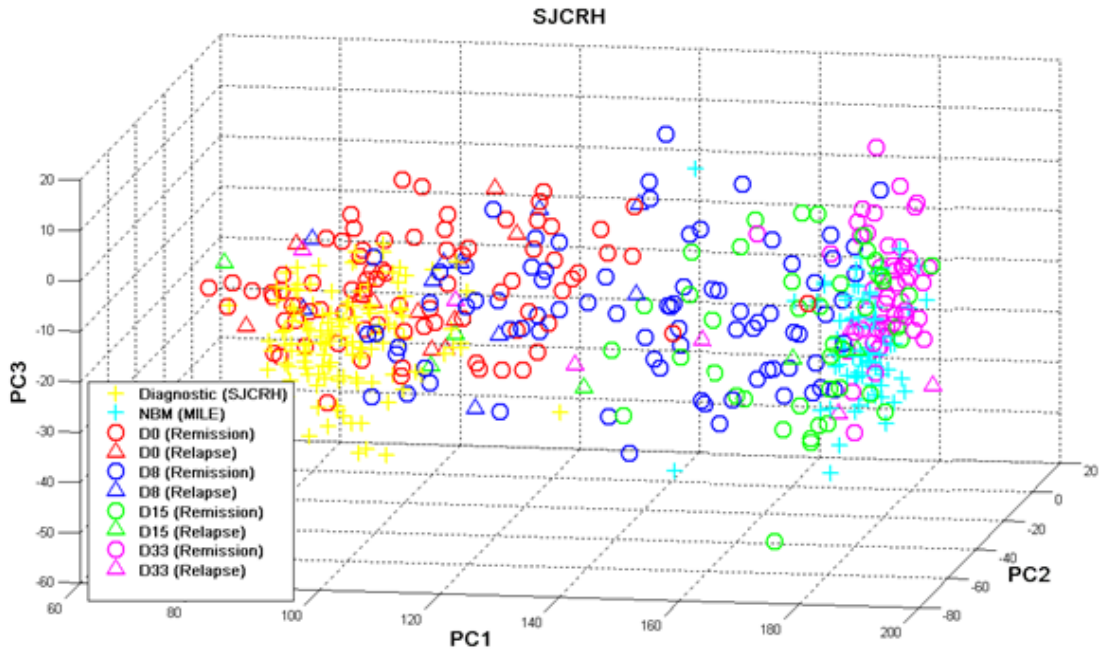


Figure 4.9: SJCRH samples in the global GSS model.

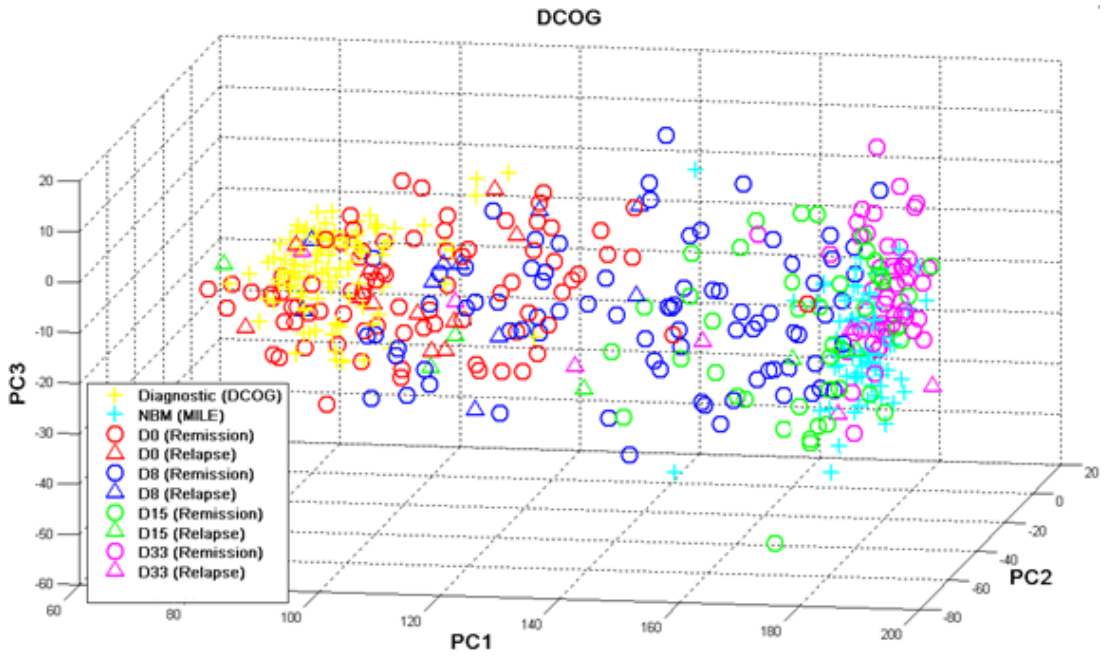


Figure 4.10: DCOG samples in the global GSS model.

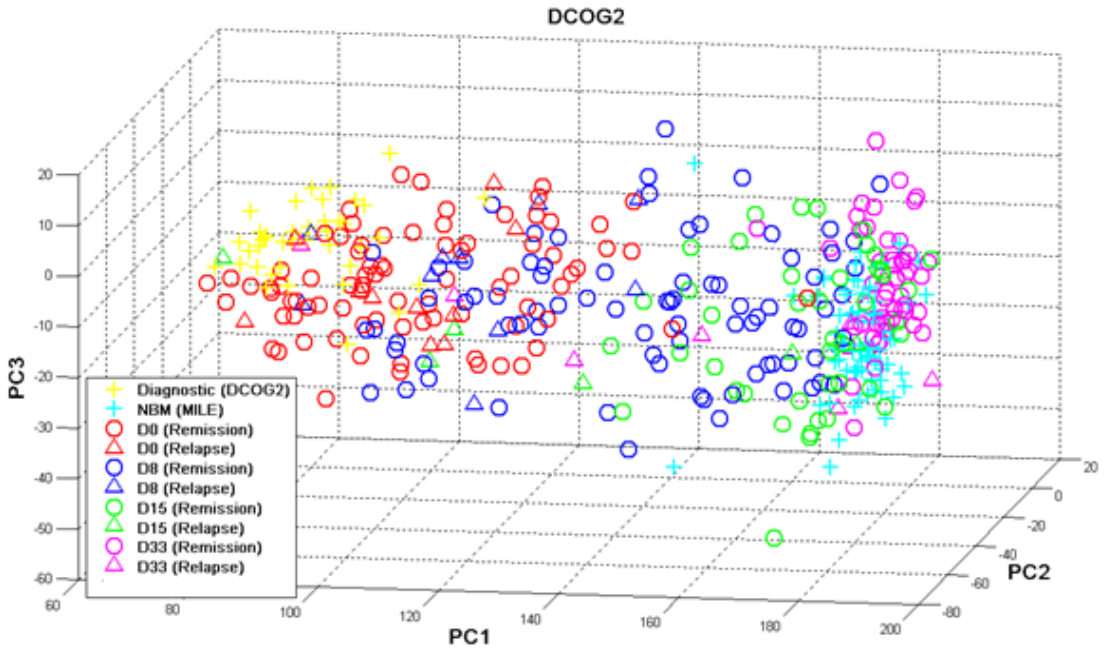


Figure 4.11: DCOG2 samples in the global GSS model.

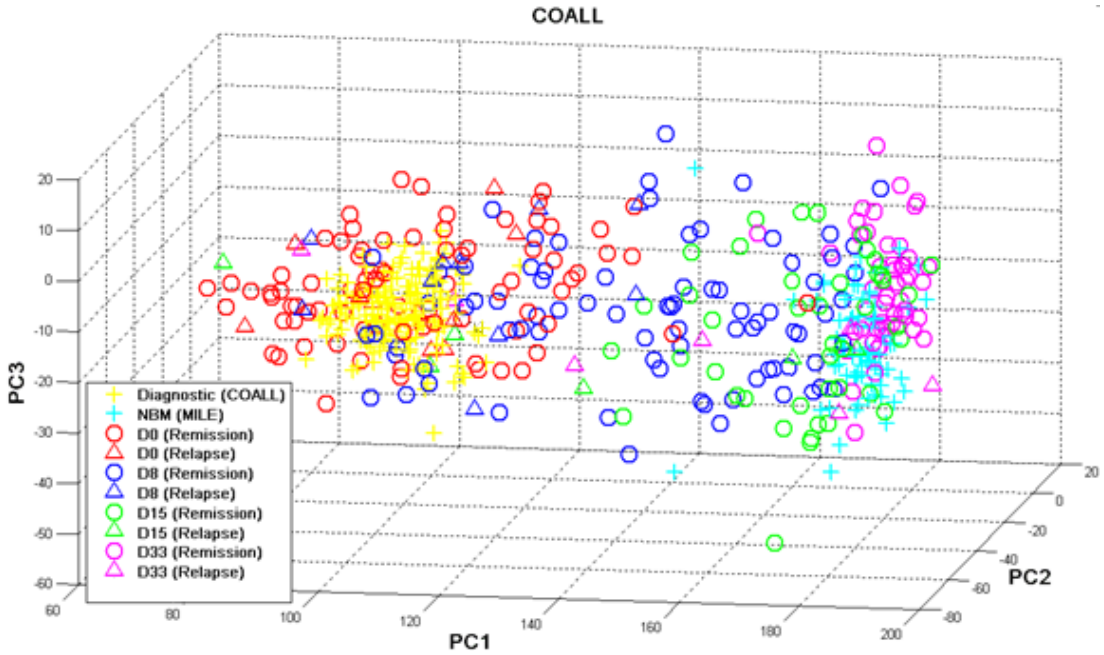


Figure 4.12: COALL samples in the global GSS model.

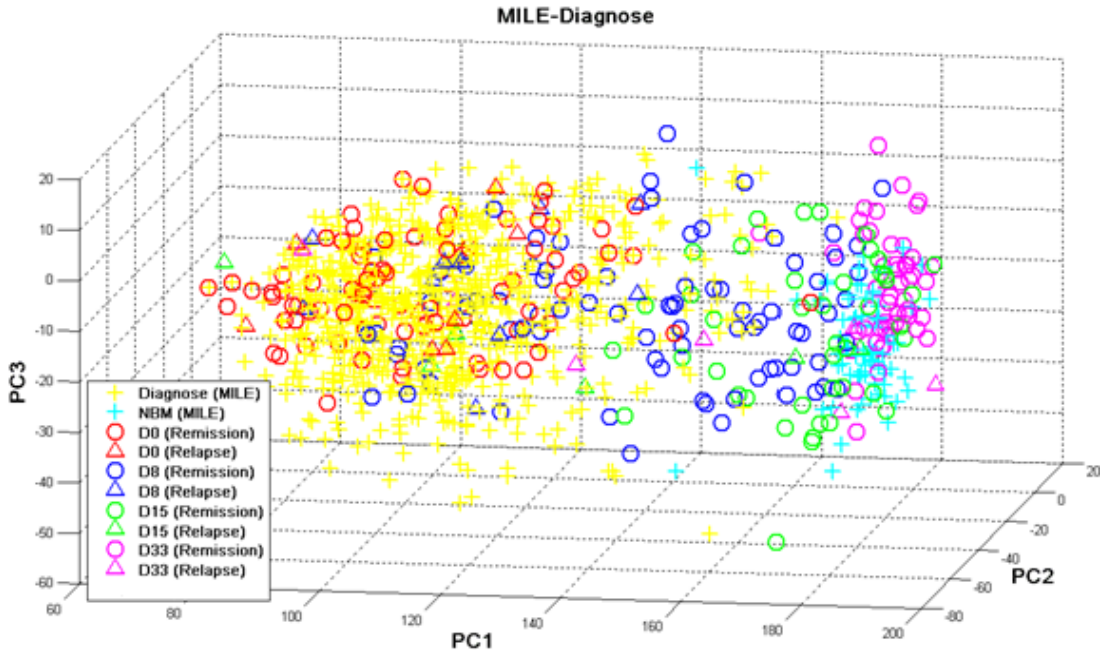
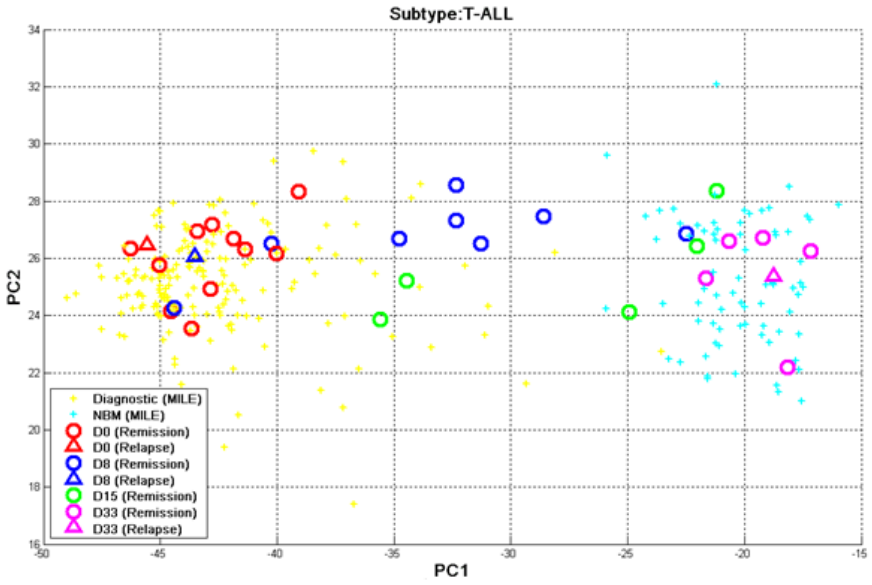


Figure 4.13: MILE-Diagnose samples in the global GSS model.

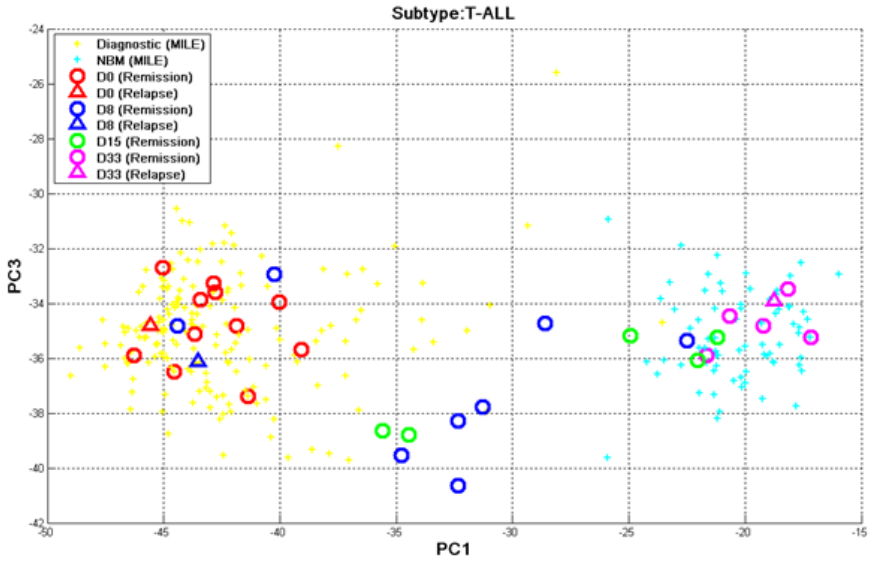
4.4.3 Local Genetic Status Shifting Model

ALL is a heterogeneous disease with many subtypes. We construct the local GSS models for each of the 6 subtypes of our data. However, since the number of samples of some subtypes is insufficient for drug responsive gene selection, we decide to construct the local models based on the MILE dataset, and then evaluate our samples in the constructed models. Specifically, we select top 50 differentially expressed probe sets between MILE-Diagnosis and MILE-NBM, ranked by the *p* value of the *t*-test, for each of the 6 subtypes and a group of other subtypes, as the drug responsive genes.

The identified drug responsive genes for each subtype are listed in Appendix A. The constructed local GSS models are shown in Figure 4.14 to 4.20. The same transition pattern from D0 to normal samples can be observed in these local models. For each local model, we calculate the variance contained in top PCs of our dataset as well as the MILE dataset. Interestingly, the variance contained in the first PC of the local models is much higher than that of the global model (Local: $69.60\% \pm 16.08\%$, Global: 49.08%). A possible explanation is that the subtype-based data stratification largely decreases the heterogeneity of our data, so most variance can be captured by the first PC already. Nevertheless, in the local model of other subtypes, the variance of the first PC is only 36.41%. It is probably because this group itself is a mixture of many rare subtypes, and the samples of this group could be very heterogeneous. Therefore, we use 3 PCs to show the local model of this group, where in other cases, 2 PCs are enough.



(a)

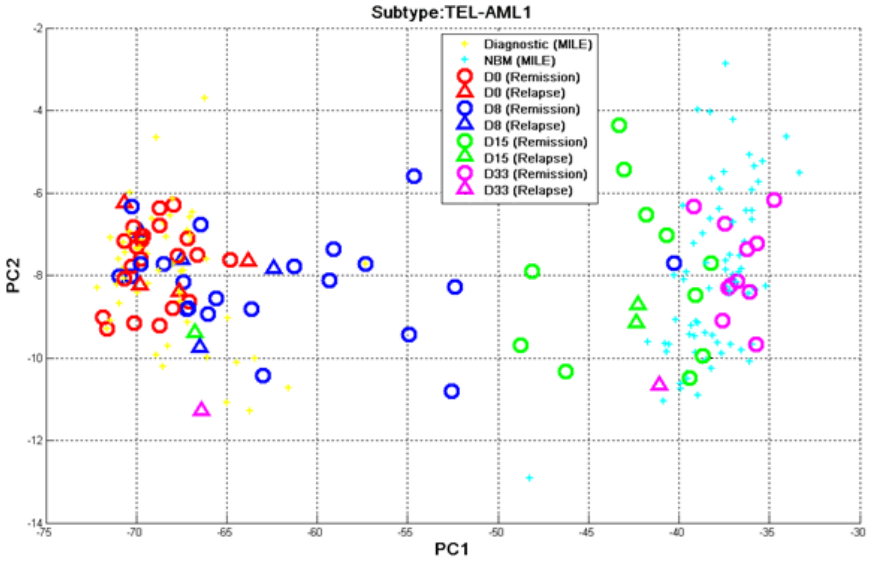


(b)

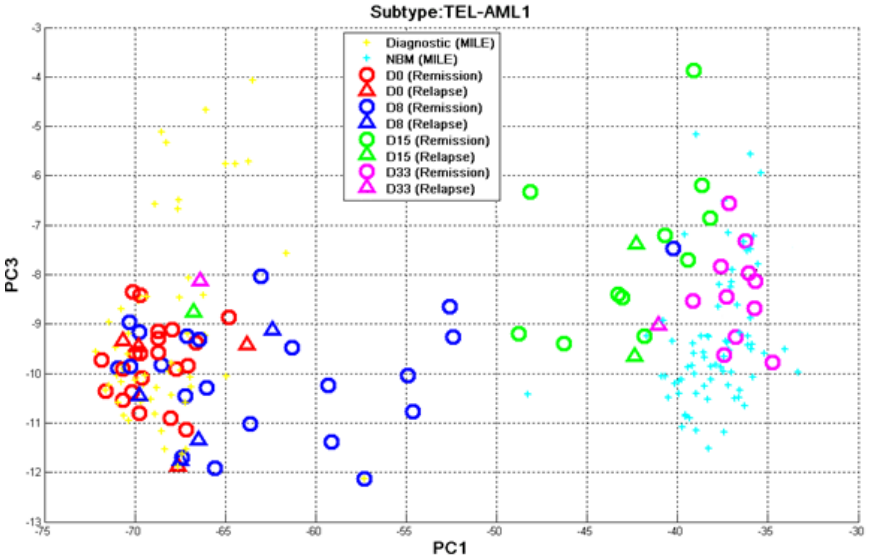
T-ALL	PC1	PC2	PC3	PC4	PC5	PC6	PC7	Total
Variance (MILE)	80.06%	2.34%	2.14%	1.56%	1.52%	1.37%	1.01%	90.00%
Variance (MASPORE)	70.83%	1.46%	2.69%	1.84%	2.28%	1.35%	2.12%	82.57%

(c)

Figure 4.14: The local GSS model of T-ALL subtype. (a) PC1 to PC2. (b) PC1 to PC3. (c) The variance contained in top PCs.



(a)

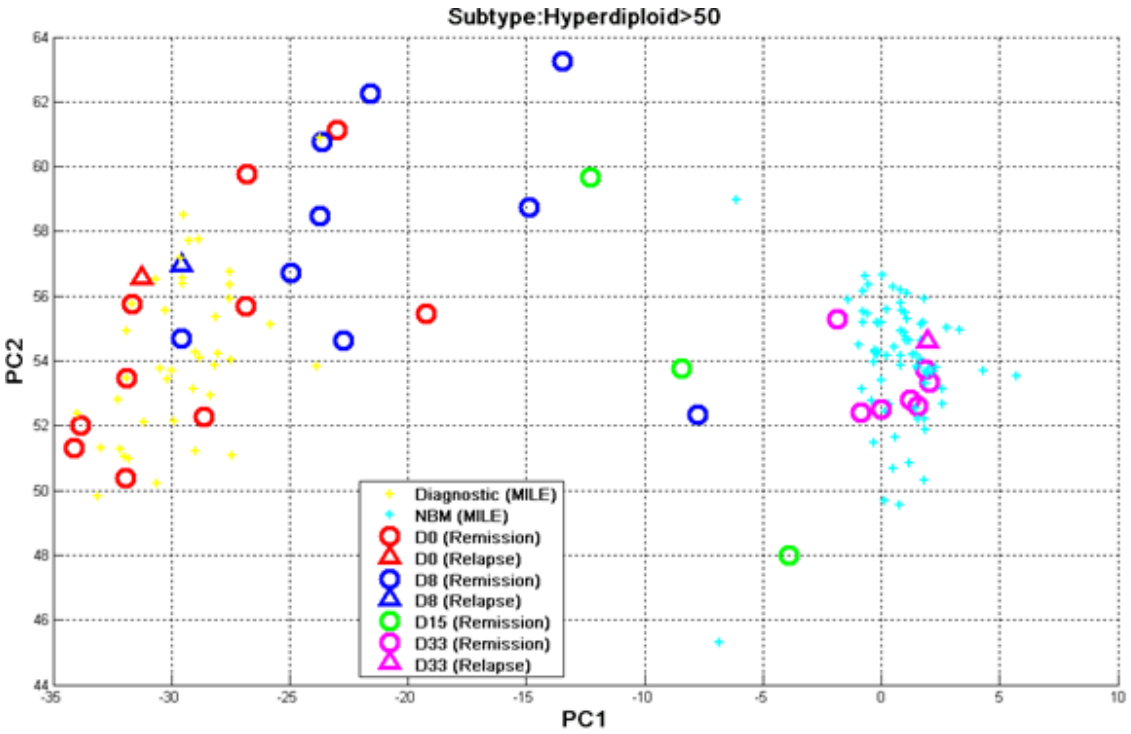


(b)

TEL-AML1	PC1	PC2	PC3	PC4	PC5	PC6	PC7	Total
Variance (MILE)	90.95%	1.20%	1.01%	0.72%	0.59%	0.56%	0.49%	95.52%
Variance (MASPORE)	86.03%	0.81%	0.96%	0.73%	0.85%	0.72%	0.41%	90.51%

(c)

Figure 4.15: The local GSS model of TEL-AML1 subtype. (a) PC1 to PC2. (b) PC1 to PC3. (c) The variance contained in top PCs.

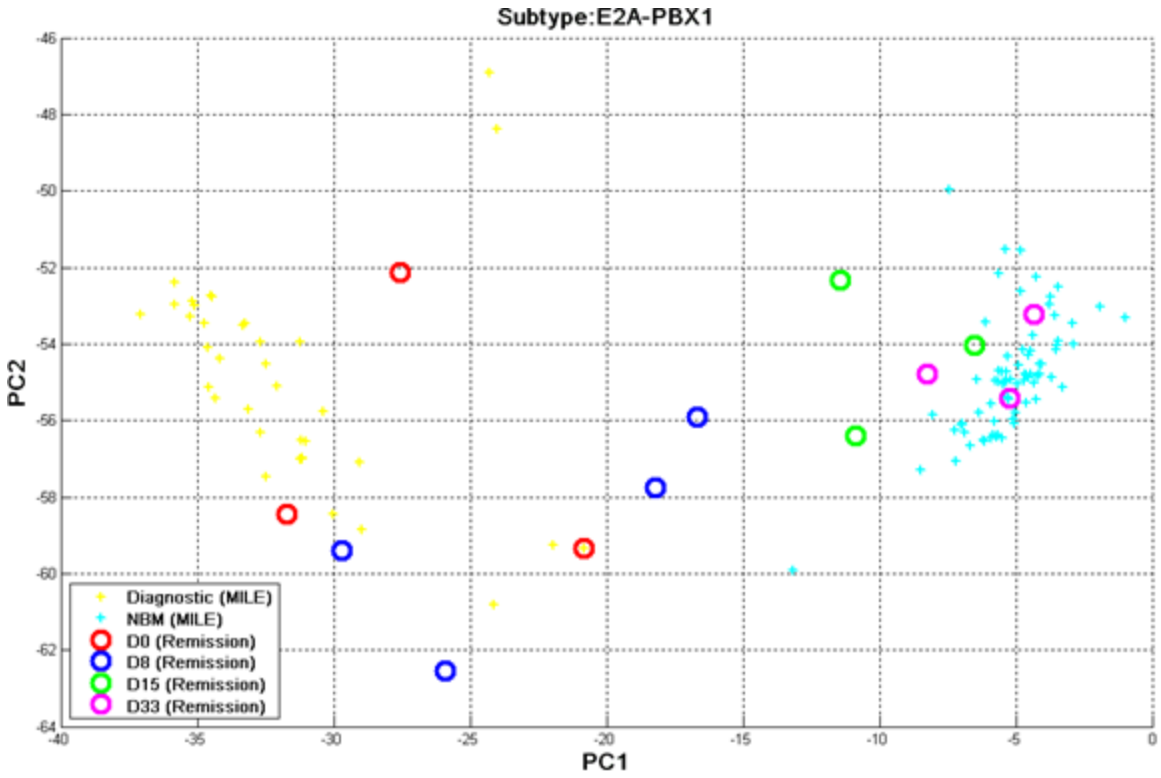


(a)

Hyperdiploid>50	PC1	PC2	PC3	PC4	PC5	PC6	PC7	Total
Variance (MILE)	90.19%	1.96%	0.82%	0.73%	0.62%	0.52%	0.48%	95.32%
Variance (MASPORE)	77.12%	6.02%	1.18%	0.99%	0.50%	0.70%	0.39%	86.90%

(b)

Figure 4.16: The local GSS model of Hyperdiploid>50 subtype. (a) PC1 to PC2. (b) The variance contained in top PCs.

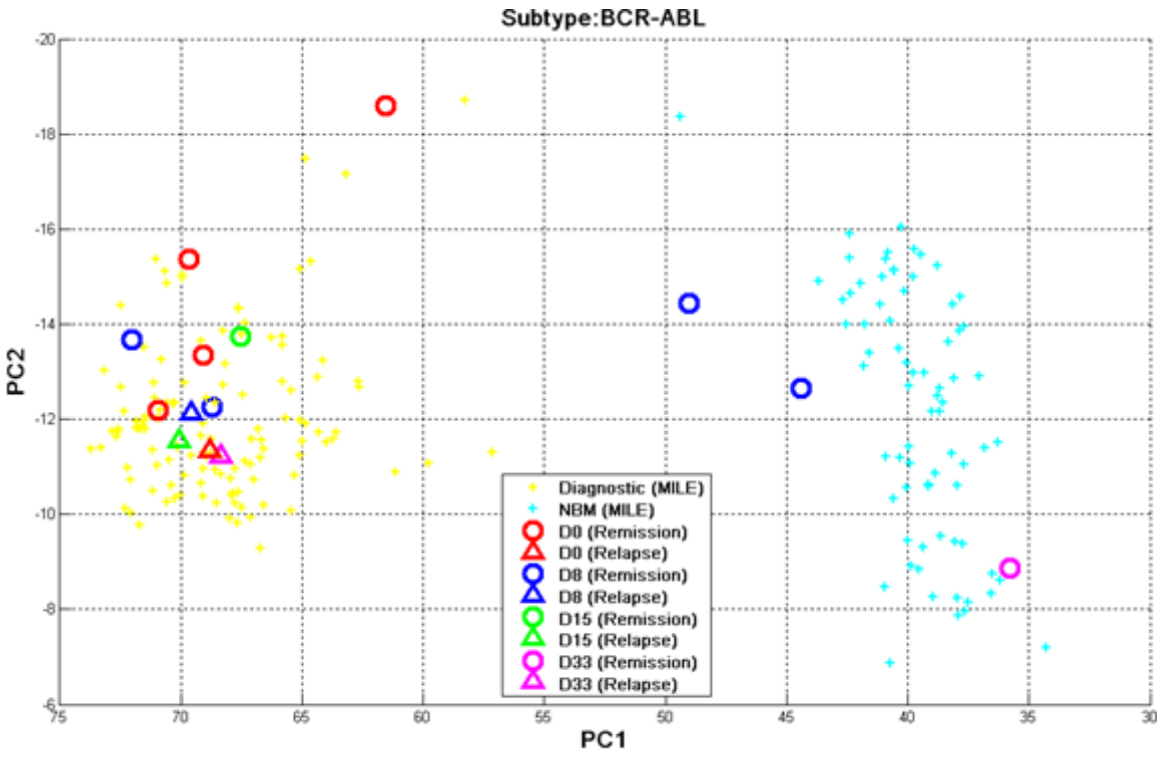


(a)

E2A-PBX1	PC1	PC2	PC3	PC4	PC5	PC6	PC7	Total
Variance (MILE)	89.39%	2.46%	1.15%	0.80%	0.61%	0.56%	0.55%	95.52%
Variance (MASPORE)	64.05%	6.53%	5.52%	1.77%	0.50%	0.48%	1.60%	80.45%

(b)

Figure 4.17: The local GSS model of E2A-PBX1 subtype. (a) PC1 to PC2. (b) The variance contained in top PCs.

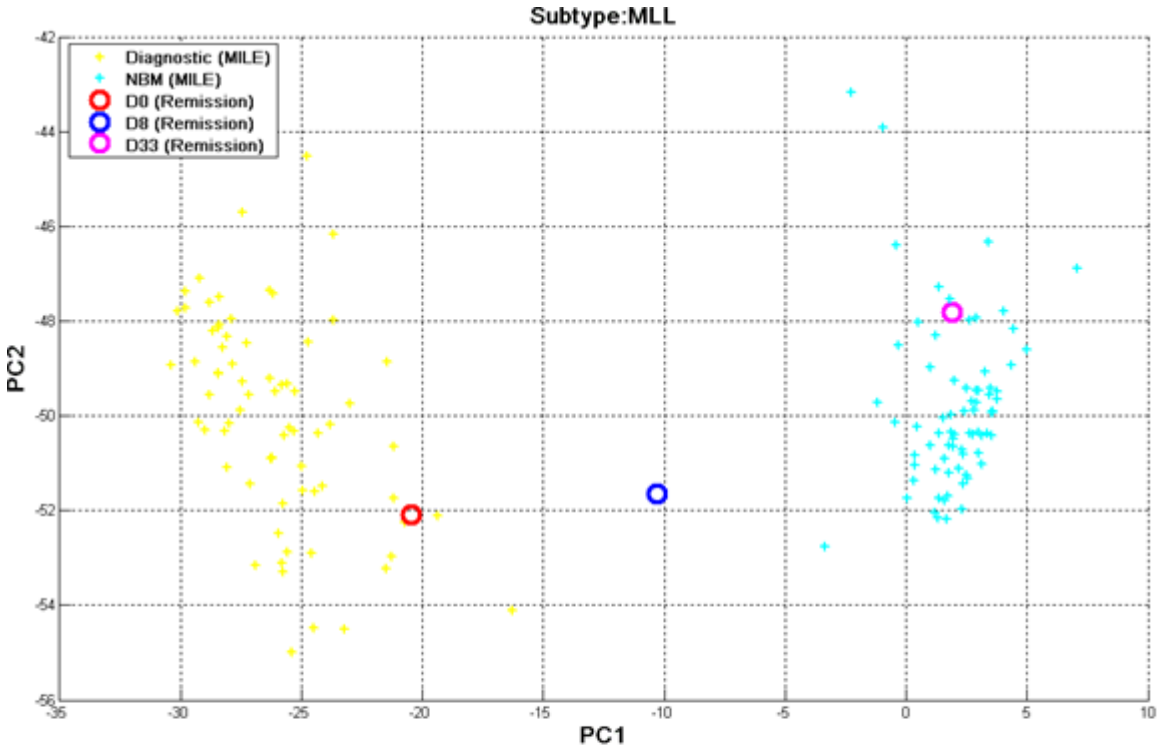


(a)

BCR-ABL	PC1	PC2	PC3	PC4	PC5	PC6	PC7	Total
Variance (MILE)	85.15%	2.01%	1.36%	0.99%	0.96%	0.84%	0.78%	92.09%
Variance (MASPORE)	76.75%	3.01%	2.56%	1.70%	0.64%	0.58%	0.97%	86.21%

(b)

Figure 4.18: The local GSS model of BCR-ABL subtype. (a) PC1 to PC2. (b) The variance contained in top PCs.

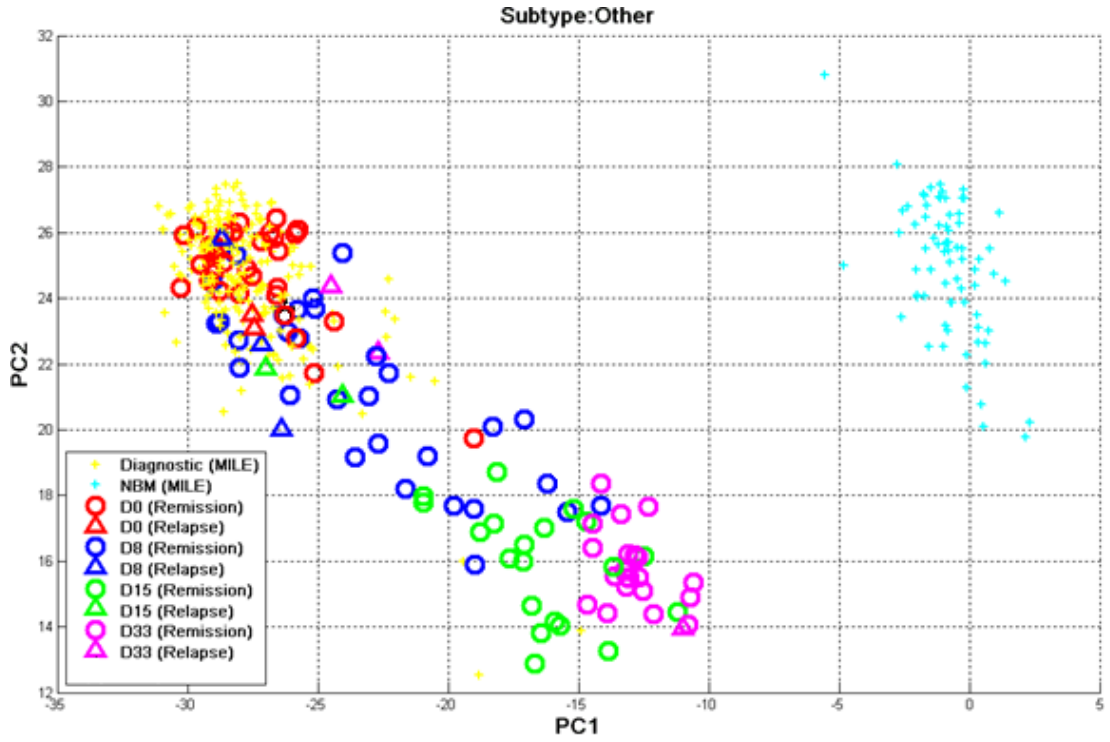


(a)

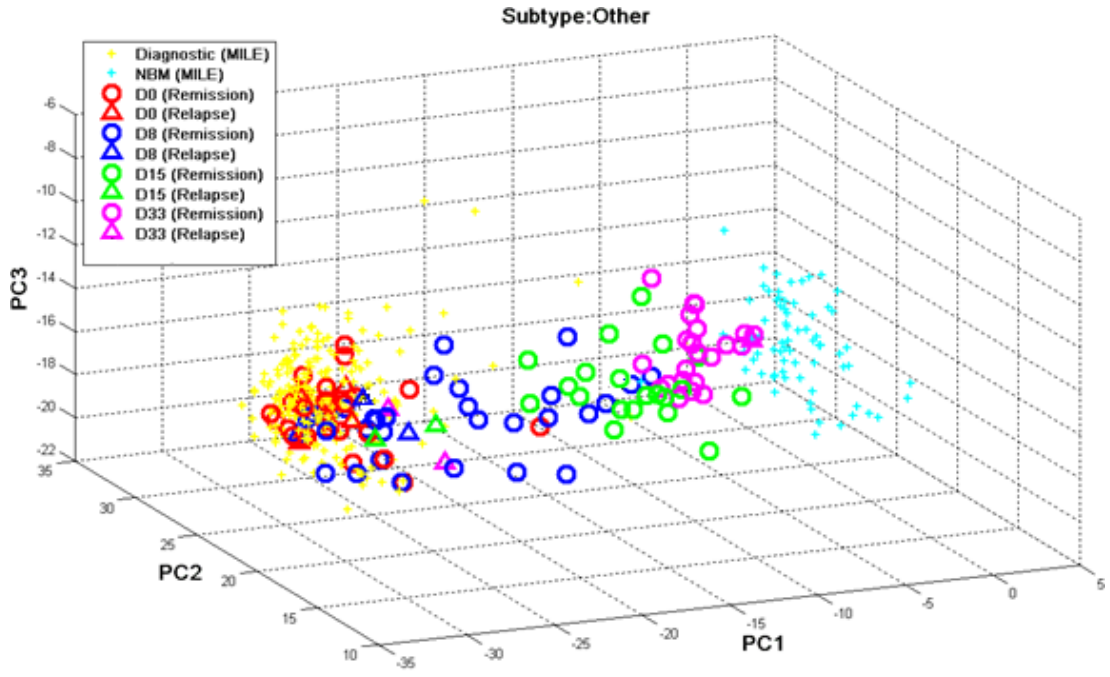
MLL	PC1	PC2	PC3	PC4	PC5	PC6	PC7	Total
Variance (MILE)	87.03%	1.69%	1.13%	1.05%	0.94%	0.80%	0.65%	93.29%
Variance (MASPORE)	76.02%	3.37%	0.60%	1.43%	1.60%	1.38%	0.68%	85.08%

(b)

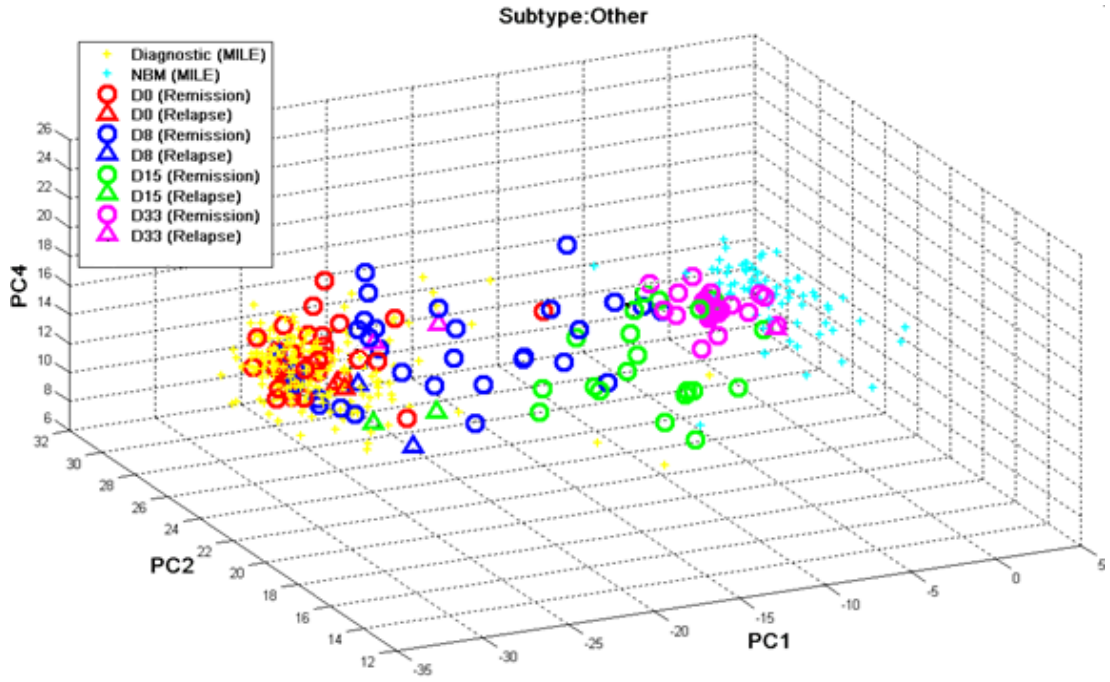
Figure 4.19: The local GSS model of MLL subtype. (a) PC1 to PC2. (b) The variance contained in top PCs.


Subtype:Other
PC2
PC1
Diagnostic (MILE)
NBM (MILE)
D0 (Remission)
D0 (Relapse)
D8 (Remission)
D8 (Relapse)
D15 (Remission)
D15 (Relapse)
D33 (Remission)
D33 (Relapse)

(a)


Subtype:Other
PC3
PC2
PC1
Diagnostic (MILE)
NBM (MILE)
D0 (Remission)
D0 (Relapse)
D8 (Remission)
D8 (Relapse)
D15 (Remission)
D15 (Relapse)
D33 (Remission)
D33 (Relapse)

(b)



(c)

Others	PC1	PC2	PC3	PC4	PC5	PC6	PC7	Total
Variance (MILE)	84.41%	1.76%	1.60%	1.16%	1.03%	0.79%	0.71%	91.46%
Variance (MASPORE)	36.41%	15.99%	5.32%	11.96%	2.08%	2.61%	2.18%	76.55%

(d)

Figure 4.20: The local GSS model of other subtypes. (a) PC1 to PC2. (b) PC1 to PC2 to PC3. (c) PC1 to PC2 to PC4. (d) The variance contained in top PCs.

4.5 Discussion

We hypothesize that a leukemic sample consists of a mixture of leukemic cells and normal cells, where the intensity of the leukemic genetic signature measured by GEP could be used to infer the proportion of leukemic cells in the sample. To validate this hypothesis, we generate time-series GEPs to investigate the relationship between GEPs and the removal of leukemic cells. We perform unsupervised hierarchical clustering and design genetic signature dissolution analysis. The results indicate that our samples are clustered by the time points. The samples of the same subtype are initially clustered together, but become scattered after treatment, which is later confirmed due to the removal of leukemic genetic signatures. In order to quantitatively describe the reduction of leukemic cells in patients during treatment, we construct the GSS models, and validate our models with several public available datasets. Our results suggest: 1) the patients achieve remissions eventually, and 2) the published diagnostic GEPs collocate well with our D0 samples in the constructed models.

We investigate cellular response to the treatment of childhood ALL by evaluating the drug responsive genes. As a result, we propose two mechanisms: 1) to induce the reconstruction of immune system and the restoration of normal hematogenesis, and 2) to suppress the negative regulation of apoptosis. The first mechanism is consistent with the philosophy of the treatment of childhood ALL, which supposes to replace the leukemic cells by newly generated normal cells in a patient. The second mechanism may explain how the leukemic cells are killed. Leukemic cells are propagated in a patient due to the lack of the proper regulation of apoptosis. The suppression

of the negative regulation of apoptosis can help to induce the apoptosis mechanisms and thus to suppress the propagation of leukemic cells.

CHAPTER 5

RELAPSE PREDICTION

5.1 Overview

In the previous chapter, we propose GSS model to mimic the removal of leukemic cells during treatment. Our models reveal that GEPs are sensitive to the load of leukemic cells in a patient. Nevertheless, we ask the question whether the constructed global GSS model can assist in the relapse prediction of childhood ALL.

Relapse prediction is important for the treatment of childhood ALL, since contemporary management of patients with childhood ALL requires patients to be upfront correctly assigned the risk of relapse. The risk-based approach allows children who historically remain in long-term remission to be treated with modest therapy and to be spared more intensive and toxic treatment, allowing children with a historically high chance of relapse to receive more intensive therapy that may increase their chance of cure.

A number of biological and clinical features have demonstrated prognostic value in childhood ALL. The National Cancer Institute classifies patients between 1 and 9 years of age and having a leukocyte count of less than 50×10^9 per liter at diagnosis as standard risk and the rest of patients as high risk (Smith et al. 1996, Pui et al. 2001). Cytogenesis-based risk assignment considers patients with BCR-ABL fusion, MLL rearrangement, and Hypodiploid<45 as high risk, patients with TEL-AML1 fusion and Hyperdiploid>50 as low risk, and the rest of patients as intermediate risk (Pui et al. 2008, Pui et al. 2009).

Early response to treatment, also known as minimal residual disease (MRD), which indicates the percentage of leukemic cells remained in a patient, has greater prognostic strength than does any other biologic or clinical features tested to date (Pui et al. 2001). An MRD level of less than 0.01% could reliably identify patients with an exceptionally good treatment outcome (Pui et al. 2001, Pui and Evans 2006). By contrast, patients with a level of 1% or more at the end of induction therapy or those with a level of 0.1% or more at late times have a very high risk of relapse (Pui et al. 2001, Pui and Evans 2006).

GEPs have been investigated for the value of prognosis as well. Holleman and colleagues identify 124 genes to predict relapses (Holleman et al. 2004). Bhojwani and colleagues identify a 24-probe-set genetic signature to predict Day-7 response, and a 47-probe-set genetic signature to predict relapses (Bhojwani et al. 2008).

In this chapter, we propose GSS distance to quantitatively describe the shifting between pre- and post-treatment samples in a GSS model. We predict relapses based on GSS distance, and compare its prognostic value with that of other clinical- and GEP-based methods.

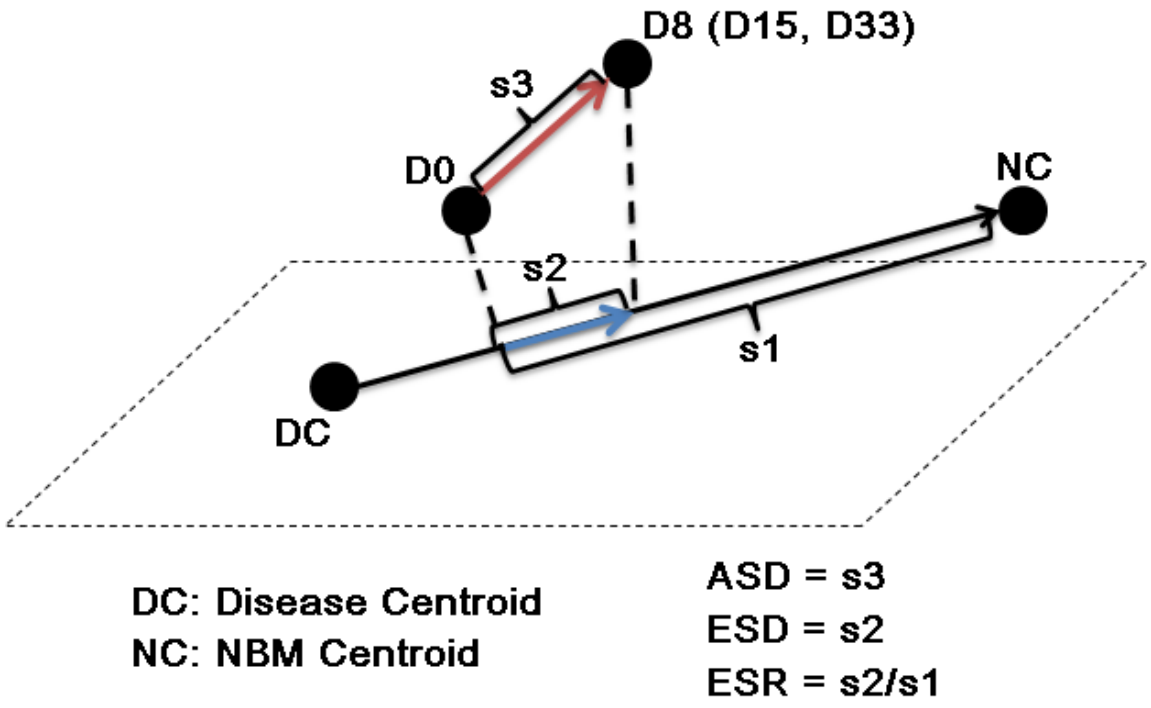


Figure 5.1: Genetic status shifting distance.

5.2 Genetic Status Shifting Distance

We think of a 3-dimensional GSS model as a 3-dimensional space defined by the 3 principal components of the model. A sample, a, can potentially be located anywhere in the space. The exact position of a is called a genetic status, denoted as $a(x, y, z)$. Given a pre-treatment genetic status $s(x, y, z)$ and a post-treatment genetic status $s'(x', y', z')$, a genetic status shifting (GSS) is defined as the vector from $s(x, y, z)$ to $s'(x', y', z')$, denoted as $\overrightarrow{ss'}$.

As shown in Figure 5.1, there are 3 metrics of GSS distance, absolute shifting distance (ASD), effective shifting distance (ESD), and effective shifting ratio (ESR).

ASD is defined as the Euclidean distance between the two genetic statuses of a GSS, formally,

$$ASD(\overrightarrow{ss'}) = \left\|\overrightarrow{ss'}\right\|.$$

ESD concerns not only the amount of a GSS, but the direction as well. It is defined as the projection of a GSS onto the direction from the centroid of pre-treatment samples to the centroid of normal samples. Formally, assuming the two centroids of pre-treatment and normal samples are $d(x_d, y_d, z_d)$ and $n(x_n, y_n, z_n)$, respectively,

$$ESD(\overrightarrow{ss'}) = \frac{\overrightarrow{ss'} \cdot \overrightarrow{dn}}{\left\|\overrightarrow{dn}\right\|}.$$

ESR further concerns the position of the pre-treatment sample of a GSS. It is defined as the ESD of a GSS divided by the Euclidean distance between the projection of the pre-treatment sample and the normal-sample centroid, formally,

$$ESR(\overrightarrow{ss'}) = \frac{\overrightarrow{ss'} \cdot \overrightarrow{dn}}{\left\|\overrightarrow{dn}\right\|} / \frac{\overrightarrow{sn} \cdot \overrightarrow{dn}}{\left\|\overrightarrow{dn}\right\|}.$$

We calculate ASD, ESD and ESR of our samples based on the global GSS model. ASD between the D0 and D8, D0 and D15, and D0 and D33 samples are shown in Table 5.1 - 5.3, respectively. ESD between the D0 and D8, D0 and D15, and D0 and D33 samples are shown in Table 5.4 - 5.6, respectively. ESR between the D0 and D8, D0 and D15, and D0 and D33 samples are shown in Table 5.7 – 5.9, respectively.

Table 5.1: ASD between the D0 and D8 samples. Relapses are highlighted with **Underline**. Extremely slow responders (D8 blast count > 10,000) are highlighted in *Italic*.

RANK	SAMPLE	ASD-D8	RANK	SAMPLE	ASD-D8	RANK	SAMPLE	ASD-D8
1	67_KL287	4.64	30	86_KL509	35.11	59	92_R332	55.25
2	**97_R208**	**4.78**	31	**96_R202**	**36.65**	60	93_R337	55.70
3	***59_R281***	***6.30***	32	51_KL461	37.44	61	26_KL369	57.13
4	*11_R280*	*7.06*	33	24_KL328	37.82	62	80_KL423	58.19
5	*19_KL205*	*7.76*	34	40_KL430	37.85	63	69_KL313	58.85
6	70_KL320	8.35	35	38_KL218	38.51	64	14_KKH19	59.57
7	*56_KL464*	*8.62*	36	47_R334	40.16	65	62_KKH22	60.80
8	**20_KL274**	**9.99**	37	74_KL383	40.85	66	01_KKH18	62.21
9	13_R410	11.94	38	57_KL535	40.99	67	75_KL385	62.85
10	41_KL441	12.25	39	10_R257	41.47	68	21_KL300	63.15
11	27_KL374	13.74	40	30_KL444	42.17	69	64_KKH29	64.44
12	**35_R313**	**14.00**	41	84_KL458	43.49	70	28_KL375	68.24
13	82_KL454	16.84	42	16_KKH21	43.55	71	32_R233	68.38
14	***77_KL401***	***16.85***	43	60_KKH30	43.62	72	34_R256	71.98
15	*33_R247*	*19.22*	44	83_KL457	43.88	73	08_KL456	73.10
16	55_KL419	19.37	45	66_KL247	44.46	74	68_KL304	73.27
17	**39_KL395**	**19.79**	46	50_KL360	45.17	75	18_KKH28	73.89
18	45_R194	23.01	47	37_R355	46.57	76	58_KL543	74.46
19	**07_KL417**	**24.00**	48	15_KKH20	46.78	77	36_R343	75.96
20	17_KKH27	24.58	49	94_R354	47.17	78	88_KL544	78.00
21	23_KL321	24.89	50	61_KKH13	48.70	79	12_R297	78.10
22	65_KL224	25.57	51	52_R252	48.97	80	72_KL377	80.71
23	29_KL439	26.93	52	73_KL381	49.52	81	87_KL522	81.06
24	43_KL536	28.38	53	48_R339	49.58	82	89_R245	85.64
25	**90_R253**	**28.72**	54	25_KL357	49.66	83	44_KL541	89.50
26	49_R432	29.27	55	85_KL485	51.39	84	63_KKH25	90.18
27	42_KL507	30.29	56	98_KL387	53.21	85	78_KL412	98.62
28	04_KL322	30.64	57	79_KL421	54.57	86	95_R431	105.60
29	**99_KL416**	**31.70**	58	05_KL354	54.59			

Table 5.2: ASD between the D0 and D15 samples. Relapses are highlighted with **Underline**. Extremely slow responders are highlighted in *Italic*.

RANK	SAMPLE	ASD-D15	RANK	SAMPLE	ASD-D15	RANK	SAMPLE	ASD-D15
1	*77_KL401*	*7.47*	17	27_KL374	64.24	33	48_R339	79.94
2	**97_R208**	**15.11**	18	85_KL485	64.72	34	64_KKH29	80.13
3	*56_KL464*	*25.51*	19	80_KL423	66.69	35	28_KL375	82.79
4	76_KL398	27.47	20	99_KL416	66.79	36	87_KL522	84.89
5	**96_R202**	**29.09**	21	06_KL378	66.80	37	93_R337	87.84
6	*59_R281*	*32.64*	22	61_KKH13	68.18	38	63_KKH25	92.24
7	67_KL287	41.60	23	74_KL383	69.77	39	14_KKH19	92.65
8	71_KL371	42.05	24	98_KL387	71.15	40	49_R432	93.43
9	50_KL360	47.50	25	84_KL458	71.19	41	94_R354	94.20
10	16_KKH21	51.87	26	18_KKH28	73.27	42	78_KL412	101.49
11	05_KL354	55.11	27	01_KKH18	74.20	43	37_R355	102.25
12	73_KL381	55.42	28	92_R332	74.38	44	36_R343	104.16
13	13_R410	58.63	29	52_R252	76.67	45	29_KL439	104.52
14	41_KL441	58.80	30	62_KKH22	78.08	46	**20_KL274**	**107.18**
15	10_R257	62.54	31	86_KL509	79.17	47	95_R431	110.76
16	15_KKH20	63.69	32	88_KL544	79.66			

Table 5.3: ASD between the D0 and D33 samples. Relapses are highlighted with **Underline**. Extremely slow responders are highlighted in *Italic*.

RANK	SAMPLE	ASD-D33	RANK	SAMPLE	ASD-D33	RANK	SAMPLE	ASD-D33
1	97_R208	9.73	20	66_KL247	72.91	39	83_KL457	87.30
2	96_R202	9.78	21	01_KKH18	73.20	40	82_KL454	87.42
3	76_KL398	22.02	22	18_KKH28	73.75	41	28_KL375	92.46
4	*59_R281*	*30.86*	23	84_KL458	74.46	42	63_KKH25	93.00
5	71_KL371	34.22	24	72_KL377	75.73	43	38_KL218	95.79
6	*77_KL401*	*44.26*	25	25_KL357	79.29	44	89_R245	95.80
7	50_KL360	47.48	26	88_KL544	79.96	45	49_R432	96.60
8	40_KL430	50.50	27	75_KL385	81.02	46	86_KL509	97.59
9	51_KL461	51.72	28	07_KL417	81.33	47	37_R355	98.98
10	67_KL287	52.87	29	87_KL522	81.46	48	55_KL419	100.11
11	42_KL507	56.30	30	90_R253	81.50	49	44_KL541	100.20
12	27_KL374	57.15	31	47_R334	81.52	50	20_KL274	101.95
13	60_KKH30	58.28	32	64_KKH29	81.77	51	29_KL439	102.15
14	85_KL485	60.39	33	62_KKH22	82.52	52	78_KL412	103.99
15	10_R257	62.15	34	*19_KL205*	*82.81*	53	14_KKH19	104.77
16	80_KL423	62.25	35	06_KL378	83.04	54	17_KKH27	106.20
17	*11_R280*	*64.21*	36	39_KL395	83.29	55	43_KL536	106.28
18	65_KL224	68.79	37	52_R252	86.10	56	30_KL444	110.44
19	69_KL313	72.80	38	08_KL456	86.44	57	95_R431	112.59

Table 5.4: ESD between the D0 and D8 samples. Relapses are highlighted with **Underline**. Extremely slow responders are highlighted in *Italic*.

RANK	SAMPLE	ESD-D8	RANK	SAMPLE	ESD-D8	RANK	SAMPLE	ESD-D8
1	*56_KL464*	*-2.21*	30	29_KL439	26.20	59	05_KL354	52.56
2	***77_KL401***	***-1.51***	31	49_R432	26.77	60	80_KL423	53.33
3	*33_R247*	*-0.46*	32	83_KL457	29.13	61	75_KL385	54.31
4	*19_KL205*	*-0.21*	33	04_KL322	30.64	62	69_KL313	54.40
5	45_R194	2.13	34	74_KL383	31.52	63	64_KKH29	55.12
6	***59_R281***	***2.50***	35	60_KKH30	32.88	64	93_R337	55.35
7	**97_R208**	**2.94**	36	38_KL218	33.00	65	26_KL369	56.83
8	67_KL287	4.63	37	10_R257	33.39	66	14_KKH19	57.79
9	**96_R202**	**5.14**	38	40_KL430	33.55	67	28_KL375	58.21
10	*11_R280*	*5.90*	39	86_KL509	34.48	68	62_KKH22	60.55
11	27_KL374	6.31	40	51_KL461	34.83	69	21_KL300	60.80
12	70_KL320	6.64	41	79_KL421	35.43	70	01_KKH18	62.07
13	41_KL441	7.86	42	94_R354	35.66	71	32_R233	67.35
14	**20_KL274**	**8.91**	43	24_KL328	35.78	72	34_R256	67.95
15	82_KL454	9.53	44	47_R334	37.35	73	68_KL304	70.06
16	13_R410	11.45	45	30_KL444	38.26	74	18_KKH28	70.41
17	**39_KL395**	**11.49**	46	57_KL535	39.40	75	36_R343	71.95
18	17_KKH27	12.64	47	73_KL381	39.51	76	58_KL543	71.96
19	**35_R313**	**13.76**	48	66_KL247	39.69	77	08_KL456	72.39
20	55_KL419	18.60	49	16_KKH21	40.05	78	12_R297	74.21
21	**99_KL416**	**19.83**	50	84_KL458	40.12	79	88_KL544	75.25
22	23_KL321	20.37	51	37_R355	41.68	80	72_KL377	76.58
23	50_KL360	20.80	52	25_KL357	42.14	81	44_KL541	79.45
24	**90_R253**	**21.40**	53	15_KKH20	43.35	82	87_KL522	79.91
25	61_KKH13	21.67	54	52_R252	45.31	83	89_R245	85.37
26	43_KL536	23.00	55	85_KL485	45.82	84	63_KKH25	85.67
27	65_KL224	23.30	56	92_R332	49.15	85	78_KL412	92.38
28	**07_KL417**	**24.00**	57	98_KL387	49.17	86	95_R431	104.57
29	42_KL507	24.00	58	48_R339	49.26			

Table 5.5: ESD between the D0 and D15 samples. Relapses are highlighted with **Underline**. Extremely slow responders are highlighted in *Italic*.

RANK	SAMPLE	ESD-D15	RANK	SAMPLE	ESD-D15	RANK	SAMPLE	ESD-D15
1	**97_R208**	**-12.33**	17	13_R410	56.01	33	88_KL544	78.10
2	***77_KL401***	***1.54***	18	15_KKH20	58.17	34	48_R339	78.48
3	**96_R202**	**5.90**	19	10_R257	58.40	35	93_R337	81.68
4	*56_KL464*	*13.45*	20	84_KL458	60.75	36	28_KL375	81.79
5	***59_R281***	***17.13***	21	80_KL423	60.92	37	87_KL522	83.85
6	76_KL398	21.30	22	**99_KL416**	**61.21**	38	63_KKH25	86.06
7	71_KL371	23.56	23	06_KL378	61.58	39	49_R432	88.27
8	50_KL360	26.07	24	85_KL485	63.61	40	94_R354	88.83
9	67_KL287	29.96	25	64_KKH29	64.65	41	14_KKH19	92.64
10	27_KL374	43.64	26	18_KKH28	66.23	42	37_R355	92.94
11	61_KKH13	46.06	27	92_R332	67.53	43	29_KL439	95.71
12	16_KKH21	46.50	28	52_R252	68.87	44	78_KL412	96.95
13	73_KL381	53.48	29	98_KL387	69.83	45	36_R343	100.91
14	74_KL383	53.86	30	01_KKH18	74.10	46	**20_KL274**	**104.35**
15	05_KL354	54.14	31	62_KKH22	74.64	47	95_R431	110.53
16	41_KL441	54.81	32	86_KL509	76.20			

Table 5.6: ESD between the D0 and D33 samples. Relapses are highlighted with **Underline**. Extremely slow responders are highlighted in *Italic*.

RANK	SAMPLE	ESD-D33	RANK	SAMPLE	ESD-D33	RANK	SAMPLE	ESD-D33
1	**97_R208**	**-0.56**	20	69_KL313	70.87	39	**39_KL395**	**81.71**
2	**96_R202**	**0.34**	21	18_KKH28	71.84	40	82_KL454	87.15
3	76_KL398	4.54	22	66_KL247	72.90	41	28_KL375	87.29
4	***59_R281***	***16.15***	23	84_KL458	72.93	42	63_KKH25	92.22
5	71_KL371	31.51	24	08_KL456	73.05	43	38_KL218	93.89
6	51_KL461	39.41	25	72_KL377	74.77	44	89_R245	94.81
7	***77_KL401***	***43.44***	26	**07_KL417**	**76.35**	45	86_KL509	95.31
8	40_KL430	46.33	27	83_KL457	76.72	46	17_KKH27	95.47
9	50_KL360	47.43	28	*19_KL205*	*76.86*	47	55_KL419	95.92
10	67_KL287	52.37	29	47_R334	78.70	48	49_R432	96.41
11	42_KL507	53.87	30	87_KL522	78.76	49	29_KL439	97.66
12	60_KKH30	56.61	31	**90_R253**	**78.83**	50	37_R355	98.84
13	27_KL374	57.11	32	88_KL544	79.25	51	44_KL541	99.89
14	85_KL485	59.75	33	75_KL385	79.26	52	78_KL412	99.97
15	80_KL423	61.74	34	25_KL357	79.27	53	**20_KL274**	**100.55**
16	10_R257	61.81	35	64_KKH29	80.39	54	14_KKH19	100.74
17	*11_R280*	*63.24*	36	62_KKH22	80.65	55	43_KL536	104.58
18	65_KL224	65.66	37	52_R252	81.43	56	30_KL444	106.99
19	01_KKH18	69.80	38	06_KL378	81.69	57	95_R431	111.91

Table 5.7: ESR between the D0 and D8 samples. Relapses are highlighted with **Underline**. Extremely slow responders are highlighted in *Italic*.

RANK	SAMPLE	ESR-D8	RANK	SAMPLE	ESR-D8	RANK	SAMPLE	ESR-D8
1	*56_KL464*	*-0.04*	30	**07_KL417**	**0.36**	59	36_R343	0.69
2	***77_KL401***	***-0.02***	31	30_KL444	0.37	60	69_KL313	0.72
3	*33_R247*	*-0.01*	32	86_KL509	0.38	61	21_KL300	0.75
4	*19_KL205*	*0.00*	33	74_KL383	0.38	62	04_KL322	0.75
5	**97_R208**	**0.03**	34	42_KL507	0.40	63	64_KKH29	0.75
6	45_R194	0.04	35	65_KL224	0.42	64	62_KKH22	0.76
7	***59_R281***	***0.04***	36	94_R354	0.42	65	57_KL535	0.77
8	**96_R202**	**0.07**	37	37_R355	0.42	66	15_KKH20	0.77
9	**20_KL274**	**0.09**	38	50_KL360	0.44	67	85_KL485	0.79
10	70_KL320	0.09	39	47_R334	0.46	68	73_KL381	0.81
11	*11_R280*	*0.09*	40	79_KL421	0.49	69	34_R256	0.83
12	67_KL287	0.10	41	51_KL461	0.52	70	68_KL304	0.83
13	27_KL374	0.10	42	25_KL357	0.53	71	32_R233	0.85
14	82_KL454	0.11	43	98_KL387	0.55	72	44_KL541	0.87
15	41_KL441	0.12	44	66_KL247	0.56	73	80_KL423	0.88
16	17_KKH27	0.13	45	10_R257	0.57	74	12_R297	0.89
17	**39_KL395**	**0.14**	46	84_KL458	0.58	75	08_KL456	0.89
18	13_R410	0.14	47	26_KL369	0.58	76	72_KL377	0.92
19	**35_R313**	**0.18**	48	52_R252	0.59	77	01_KKH18	0.93
20	55_KL419	0.19	49	48_R339	0.60	78	58_KL543	0.94
21	23_KL321	0.22	50	14_KKH19	0.61	79	89_R245	0.95
22	43_KL536	0.23	51	60_KKH30	0.62	80	05_KL354	0.95
23	**90_R253**	**0.26**	52	40_KL430	0.64	81	87_KL522	0.96
24	29_KL439	0.27	53	75_KL385	0.64	82	88_KL544	0.96
25	49_R432	0.29	54	28_KL375	0.67	83	95_R431	0.96
26	61_KKH13	0.31	55	16_KKH21	0.67	84	78_KL412	0.98
27	**99_KL416**	**0.33**	56	93_R337	0.68	85	18_KKH28	0.99
28	83_KL457	0.34	57	92_R332	0.68	86	63_KKH25	1.01
29	38_KL218	0.35	58	24_KL328	0.68			

Table 5.8: ESR between the D0 and D15 samples. Relapses are highlighted with **Underline**. Extremely slow responders are highlighted in *Italic*.

RANK	SAMPLE	ESR-D15	RANK	SAMPLE	ESR-D15	RANK	SAMPLE	ESR-D15
1	97_R208	-0.13	17	41_KL441	0.85	33	88_KL544	0.99
2	*77_KL401*	*0.02*	18	84_KL458	0.88	34	93_R337	1.00
3	96_R202	0.08	19	64_KKH29	0.88	35	80_KL423	1.00
4	*56_KL464*	*0.25*	20	99_KL416	0.89	36	10_R257	1.00
5	*59_R281*	*0.27*	21	52_R252	0.90	37	87_KL522	1.00
6	50_KL360	0.55	22	18_KKH28	0.93	38	63_KKH25	1.01
7	67_KL287	0.62	23	92_R332	0.93	39	95_R431	1.01
8	74_KL383	0.65	24	49_R432	0.94	40	78_KL412	1.03
9	61_KKH13	0.66	25	62_KKH22	0.94	41	20_KL274	1.03
10	13_R410	0.69	26	28_KL375	0.94	42	94_R354	1.04
11	27_KL374	0.71	27	37_R355	0.95	43	15_KKH20	1.04
12	06_KL378	0.71	28	48_R339	0.96	44	85_KL485	1.10
13	71_KL371	0.72	29	36_R343	0.96	45	73_KL381	1.10
14	98_KL387	0.78	30	05_KL354	0.98	46	01_KKH18	1.11
15	16_KKH21	0.78	31	14_KKH19	0.98	47	76_KL398	1.55
16	86_KL509	0.84	32	29_KL439	0.99			

Table 5.9: ESR between the D0 and D33 samples. Relapses are highlighted with **Underline**. Extremely slow responders are highlighted in *Italic*.

RANK	SAMPLE	ESR-D33	RANK	SAMPLE	ESR-D33	RANK	SAMPLE	ESR-D33
1	**97_R208**	**-0.01**	20	82_KL454	0.99	39	66_KL247	1.03
2	**96_R202**	**0.00**	21	55_KL419	0.99	40	30_KL444	1.03
3	***59_R281***	***0.25***	22	**20_KL274**	**0.99**	41	85_KL485	1.03
4	76_KL398	0.33	23	**39_KL395**	**0.99**	42	43_KL536	1.04
5	51_KL461	0.59	24	25_KL357	0.99	43	01_KKH18	1.05
6	***77_KL401***	***0.61***	25	38_KL218	1.00	44	89_R245	1.05
7	40_KL430	0.88	26	*11_R280*	*1.00*	45	86_KL509	1.05
8	83_KL457	0.88	27	50_KL360	1.01	46	84_KL458	1.06
9	72_KL377	0.90	28	37_R355	1.01	47	78_KL412	1.06
10	08_KL456	0.90	29	18_KKH28	1.01	48	52_R252	1.06
11	42_KL507	0.90	30	17_KKH27	1.01	49	10_R257	1.06
12	27_KL374	0.93	31	29_KL439	1.01	50	14_KKH19	1.07
13	06_KL378	0.94	32	28_KL375	1.01	51	60_KKH30	1.07
14	75_KL385	0.94	33	88_KL544	1.01	52	63_KKH25	1.08
15	69_KL313	0.94	34	80_KL423	1.02	53	44_KL541	1.09
16	87_KL522	0.94	35	62_KKH22	1.02	54	67_KL287	1.09
17	**90_R253**	**0.96**	36	*19_KL205*	*1.02*	55	64_KKH29	1.10
18	71_KL371	0.97	37	49_R432	1.03	56	**07_KL417**	**1.14**
19	47_R334	0.97	38	95_R431	1.03	57	65_KL224	1.17

5.3 Relapse Prediction

We predict relapses by GSS distance. The prediction is based on the assumption of the importance of early response (we consider GSS as early genetic response). Hypothetically, if a patient has a large GSS, the patient is supposed to respond well to treatment, and the patient's risk of relapse is low; on the contrary, if a patient has a very small or even a negative (only applicable to ESD and ESR) GSS, the patient is supposed to poorly respond to treatment, and the patient's risk of relapse is high.

Figure 5.2 shows the receiver operating characteristics (ROC) of the various measurements of GSS distance in relapse prediction, where,

$$\begin{cases} sensitivity = \dfrac{True\ Positive}{True\ Positive + False\ Negative} \\ specificity = \dfrac{True\ Negative}{True\ Negative + False\ Positive} \end{cases}.$$

The *p* values refer to the areas under the curves, calculated by MedCacl software, version 9.6.2.0 (MedCalc Software, Mariakerke, Belgium). Our results indicate GSS distance is very predictive of the relapses.

Among the 3 time points, D8 GSS distance performs the best (all three metrics with *p* value < 0.0001). We next ask whether D8 GSS distance can be used to predict D8 response. As introduced in Chapter 2, D8 response is defined based on the peripheral blood leukemic blasts. A measurement of > 1,000 blasts/μL is considered as a slow response, and a measurement of > 10,000 blasts/μL is considered as an extremely slow response.

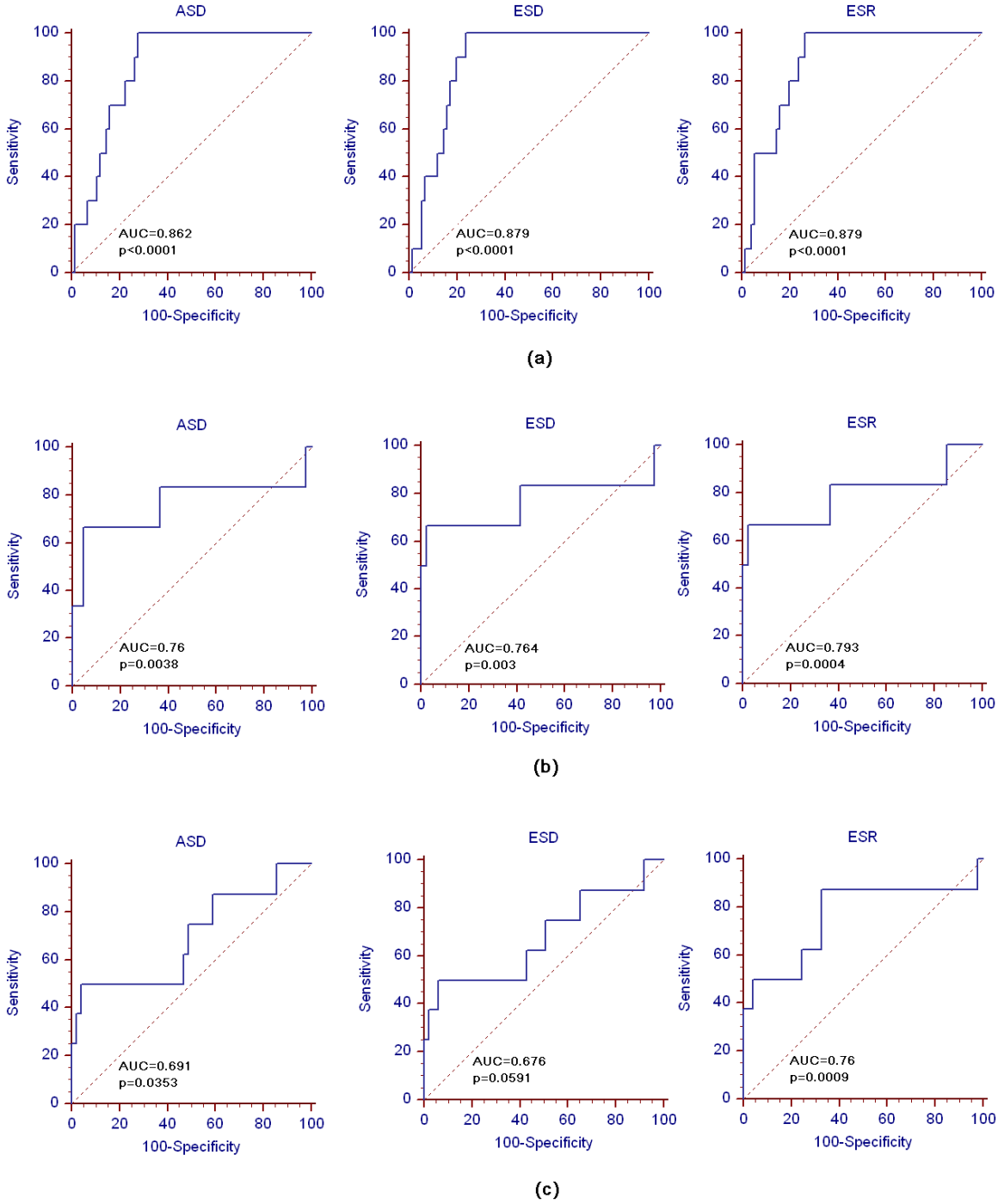


Figure 5.2: Receiver operating characteristics of GSS distance in relapse prediction. (a) D8 GSS distance. (b) D15 GSS distance. (c) D33 GSS distance.

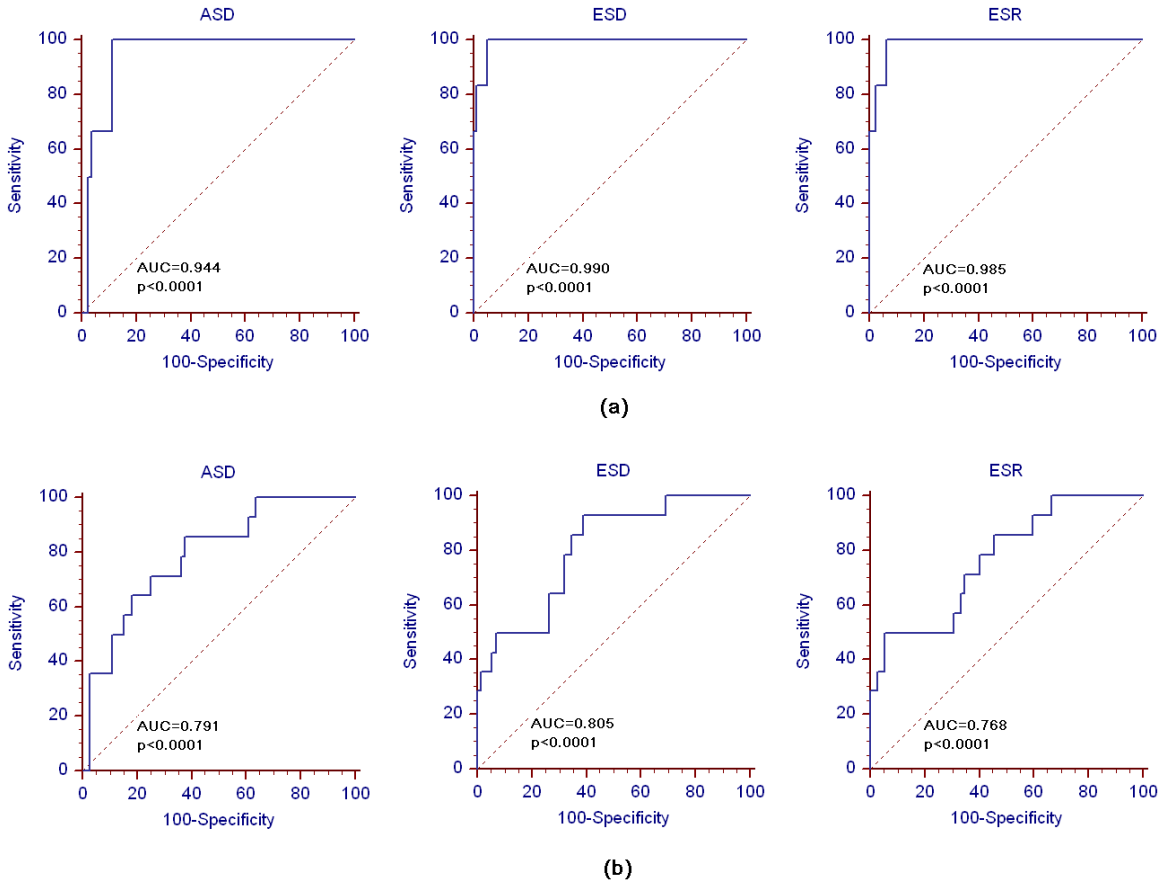


Figure 5.3: Receiver operating characteristics of D8 GSS distance in D8 response prediction. (a) Extremely slow response. (b) Slow response.

Figure 5.3 shows the ROCs of D8 GSS distance in D8 response prediction. Our results indicate D8 GSS distance is very predictive of D8 response. Especially, the prediction of extremely slow response is almost perfect (the area under the curve of ESD = 0.99).

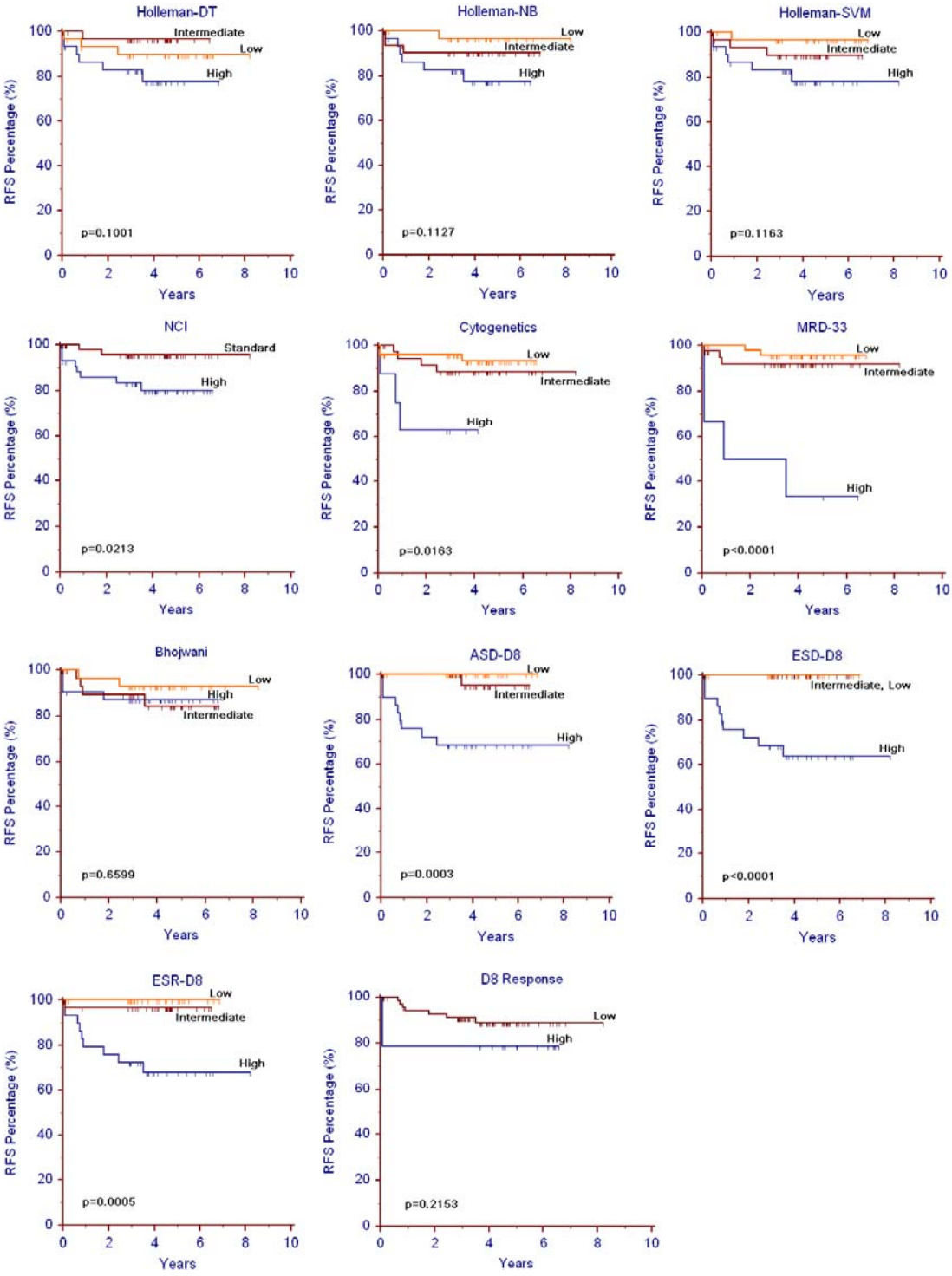


Figure 5.4: Relapse prediction results of various methods by Kaplan-Meier method.

Table 5.10: Comparison of relapse prediction performance among various methods. The performance is evaluated based on Figure 5.4, where high-risk patients are predicted as the relapses, and the rest of patients are predicted as the remissions. The best performer of each column is highlighted.

Method	Prognostic Feature	Sensitivity	Specificity	Accuracy
Holleman-DT	D0 GEP	60.00%	69.51%	64.76%
Holleman-NB	D0 GEP	60.00%	69.51%	64.76%
Holleman-SVM	D0 GEP	60.00%	69.51%	64.76%
Bhojwani	D0 GEP	20.00%	79.27%	49.63%
NCI	D0 GEP	80.00%	58.14%	69.07%
Cytogenetics	Diagnostic Cytogenetics	30.00%	**94.19%**	62.09%
MRD-D33	D33 MRD	77.78%	54.12%	65.95%
D8 Response	D8 Blast Count	30.00%	85.53%	57.76%
ASD-D8	D0 and D8 GEP	90.00%	73.68%	81.84%
ESD-D8	D0 and D8 GEP	**100.00%**	75.00%	**87.50%**
ESR-D8	D0 and D8 GEP	90.00%	73.68%	81.84%

We next compare D8-GSS-based relapse prediction with several other clinical- and GEP-based methods in our dataset. These protocols are described as the following:

- Holleman-DT: Proposed by Holleman and colleagues, 124 genes are used (Holleman et al. 2004). Decision tree is used as the classification model (not specified in the original paper). Patients are equally assigned into 3 risk groups based on the predicted combined drug resistance scores.
- Holleman-NB: Proposed by Holleman and colleagues, 124 genes are used (Holleman et al. 2004). Naïve Bayes is used as the classification model. Patients are equally assigned into 3 risk groups based on the predicted combined drug resistance scores.
- Holleman-SVM: Proposed by Holleman and colleagues, 124 genes are used (Holleman et al. 2004). Support vector machine is used as the classification model.

Patients are equally assigned into 3 risk groups based on the predicted combined drug resistance scores.

- Bhojwani: Proposed by Bhojwani and colleagues, 47 probe sets are used (Bhojwani et al. 2008).
- NCI: Proposed by the National Cancer Institute, patients between 1 and 9 years of age and having a leukocyte count of less than 50×10^9 per liter at diagnosis are assigned as standard risk. The rest of patients are assigned as high risk (Smith et al. 1996, Pui et al. 2001).
- Cytogenetics: Proposed by Pui and colleagues, patients with BCR-ABL fusion, MLL rearrangement, and Hypodiploid<45 are classified as high risk. Patients with TEL-AML1 fusion and Hyperdiploid>50 are classified as low risk. The rest of patients are classified as intermediate risk (Pui et al. 2008, Pui et al. 2009).
- MRD-D33: Proposed by Pui and colleagues, patients with D33 MRD < 0.01% are classified as low risk. Patients with D33 MRD > 1% are classified as high risk. The rest of patients are classified as intermediate risk (Pui et al. 2001, Pui and Evans 2006).
- D8 Response: Patients with D8 blast count > 1,000 are predicted as high risk. The rest of patients are predicted as low risk.
- ASD-D8: Patients are equally assigned into 3 risk groups based on D8 ASD.
- ESD-D8: Patients are equally assigned into 3 risk groups based on D8 ESD.
- ESR-D8: Patients are equally assigned into 3 risk groups based on D8 ESR.

Figure 5.4 compares the results of relapse prediction by Kaplan-Meier method, and Table 5.10 shows the corresponding performance evaluations. From Table 5.10, we see that ESD-D8 has 100% sensitivity at 75% specificity. This means it has both much better sensitivity and specificity than Hollerman-DT/NB/SVM, Bhojwani, NCI, and MRD-D33 at very large margin.

For example, MRD-D33 has 77.78% sensitivity at 54.12% specificity. Given that there are 10 relapses and 76 remissions, this means MRD-D33 can identify 10 × 77.78% = 8 of them, while giving false alarm on 76 × (1 – 54.12%) = 35 of the good patients. In contrast, ESD-D8 has 100% sensitivity at 75% specificity. This means ESD-D8 can identify all 10 poor patients, while giving false alarm on only 76 × (1 – 75%) = 19 of them. Clearly, ESD-D8 is far better than MRD-D33. In fact, looking at Table 5.4, we can easily compute that, when the score threshold of ESD-D8 is set at ~78% sensitivity, the corresponding specificity is (76 − 13) / 76 = 83%, which is far higher than MRD-D33's 54.12% specificity at the same sensitivity level[1].

To compare with D8 Response (based on D8 blast count), we note that D8 Response has sensitivity = 30% and specificity = 85.53%. Looking at Table 5.4, when the score threshold of ESD-D8 is set at 30% sensitivity, the corresponding specificity is (76 − 4) /76 = 94.74%, which is far better than D8 Response's 85.53%.

To compare with Diagnostic Cytogenetics, we note that cytogenetics has sensitivity = 30% and specificity = 94.19%. Looking at Table 5.4, when the score threshold of ESD-D8 is set at 30% sensitivity, the corresponding specificity is 94.74%, which is also better than cytogenetics's 94.19%.

[1] A widely accepted methodology for comparing two prediction systems is to first calibrate them to the same level of sensitivity and then compare their specificity.

Thus, we conclude ESD-D8 is superior to any other methods in comparison.

5.4 Discussion

The prognostic strength of GSS is not unexpected. As shown in Figure 4.1, unsupervised hierarchical clustering reveals that the post-treatment samples of a relapse (marked in red) or an extremely slow responder (marked in green) tend to be clustered together with the pre-treatment sample of the same patient ($p = 0.016$). This result suggests that a patient with poor outcome may have more resistant genetic characteristics to treatment, when compared to a patient with good outcome. In Figure 4.8, the global GSS model, we find the remissions shift generally further towards the normal samples than the relapses, especially in the D8 samples. In our results of relapse prediction, D8 GSS distance performs much better than D15 and D33 GSS distance. This observation suggests that early response may be more important than the result of remission induction in disease prognosis, which may explain why 98% of patients can achieve a complete remission after remission induction, while still nearly 20% of them relapse.

Both MRD and GSS value treatment response in relapse prediction. However, they are different. MRD only concerns the absolute number of leukemic cells after treatment, and it ignores the initial load of leukemic blasts in a patient. It allows a clinician to assess the risk of a patient without any diagnostic information. In contrast, GSS concerns the difference between pre- and post-treatment GEPs. It emphasizes the change, rather than the result, of treatment. We argue that, in philosophy, GSS is more close to the definition of the term, response, than MRD, and we have demonstrated GSS-based methods perform better than MRD-based method in our data.

Nevertheless, it is probably yet too early to declare that GSS is generally better than MRD in relapse prediction, as the prognostic value of MRD has been evaluated in tens of thousands of patients in the last twenty years.

In our results, ESD performs better than ASD and ESR. The explanation to that ESD is superior to ASD is straightforward, as ESD concerns the direction of a GSS, while ASD does not. However, it may be confusing that ESD performs better than ESR. A possible explanation is that the position of a pre-treatment sample in the global GSS model is not only decided by the initial load of leukemic blasts but the subtype of the patient as well. For example, in Figure 4.8, the pre-treatment samples of T-ALL are located closer to the normal samples than that of the rest subtypes. This difference is not attributed to the different level of the initial blast count, but to the difference between B and T lineage of the disease. We propose to solve this problem by constructing the local GSS models for each subtype. However, due to the limitation of the number of patients in our data, we are not able to make any conclusive comparison between ESD and ESR of the local models.

CHAPTER 6

PROOF OF CONCEPT – ACUTE MYELOID LEUKEMIA

6.1 Overview

Acute myeloid leukemia (AML) is characterized by a rapid growth of abnormal white blood cells in bone marrow, which thereafter inferences the growth and functioning of normal white blood cells (Lowenberg, Downing and Burnett 1999, Estey and Döhner 2006). Similar to ALL, the treatment of AML is generally composed of an induction phase and a consolidation phase. The first phase attempts to produce a complete remission, which is defined as a marrow with less than 5% of blast, a neutrophil count greater than 1,000, and a platelet count greater than 100,000 (Cheson et al. 2003). The second phase aims to prolong the remission achieved in the first phase (Estey and Döhner 2006). However, different from ALL, the overall 5-year survival rate of AML is only 40%, where relapse is the major reverse event (Colvin and Elfenbein 2003). Therefore, relapse prediction is critical to the treatment of AML.

Table 6.1: Patient characteristics of our AML dataset.

Sample	Stage	Age	Sex	Clinical Status	TP1	TP2	TP3	TP4
D318	M4	15.2	M	Relapse, Death	D0	D5	D33	
D474	M1	10.3	M	CCR	D0	--	D33	D60
KKH014	M2	3.9	F	CCR	D0	--	D36	
KL336	M2	11.3	F	CCR	D0	--	D31	
KL343	M3	2.9	M	Relapse, Death	D0	--	D36	
KL448	M3	2.6	F	CCR	D0	D17	--	D51
KL473	M7	2.7	M	Relapse, Death	D0	--	D32	
KL505	M3	7.4	F	CCR	D0	D14	--	D45

Since the treatment philosophy of AML is similar to that of ALL, we examine GSS and its prognostic value in an AML dataset as a proof of concept. The dataset consists of 20 samples from 8 AML patients at different time points. Table 6.1 shows the clinical characteristics of these patients.

6.2 Unsupervised Hierarchical Clustering

Affymetrix HG-U133 Plus2.0 microarrays (Affymetrix, Santa Clara, CA) are hybridized with the prepared specimens of our samples to generate time-series GEPs. Signal values are interrogated by MAS5.0. To reduce systematic batch effects, probe sets with "Present" calls in less than 50% of samples are removed. As a result, 25,408 probe sets are eligible for the next stage of analysis. Quantile normalization is applied to the whole dataset thereafter.

Unsupervised hierarchical clustering is performed by Eisen's software, Cluster 3.0, with Pearson's correlation and complete linkage as the parameters (Eisen et al. 1998). Figure 6.1 shows the resulted dendrogram.

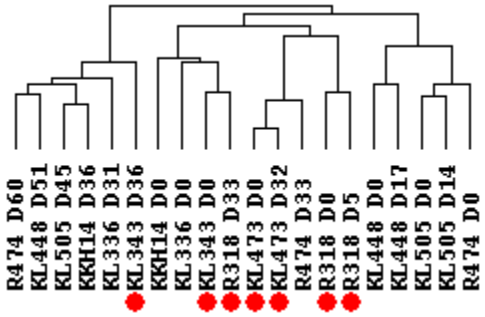


Figure 6.1: Unsupervised hierarchical clustering. The relapses are marked in the figure.

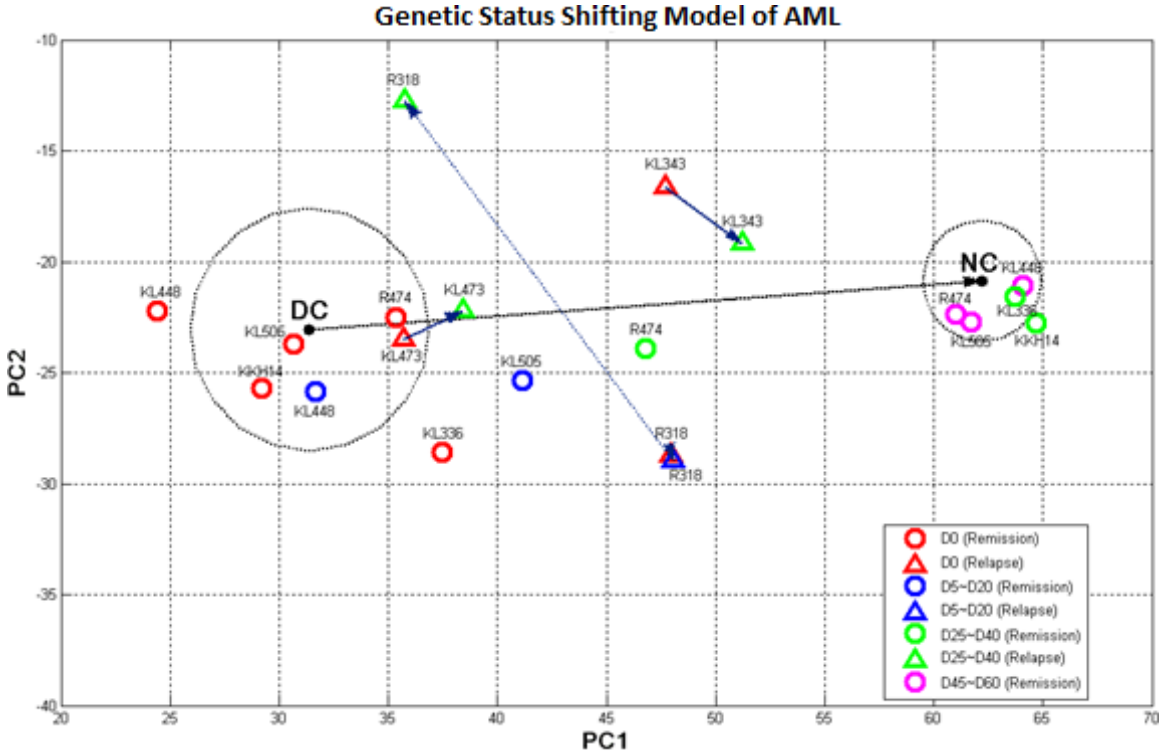


Figure 6.2: GSS-AML. The disease centroid (DC) and NBM centroid (NC) are calculated based on the samples of MILE-AML and MILE-NBM, respectively. The GSS of relapses are shown in the figure.

In Figure 6.1, the samples are organized with two major clusters separated by the time points. The cluster on the right is mainly composed of samples collected at early time points (Day < 25), and the cluster on the left consists of samples collected at later time points (Day > 25). Exceptions are R318_D33, KL473_D32, which are the relapses, and R474_D33, which is a slow responder to treatment, since a later sample of the same patient, R474_D60, is found to migrate to the left cluster.

6.3 Disease Status Shifting Model

We thereafter construct the GSS model of AML, denoted as GSS-AML, to validate the concept of GSS in AML. Considering our AML dataset is in a small scale, we use MILE-AML and MILE-NBM to construct GSS-AML, and then put our samples into the model.

Specifically, we analyze the two datasets by MAS5.0 and only retain probe sets with "Present" calls in all samples of either dataset. This results 7,760 probe sets eligible for the next stage of analysis. The two datasets are then combined and quantile normalization is applied to the combined dataset. We identify drug responsive genes by selecting top 100 differentially expressed probe sets between MILE-AML and MILE-NBM samples, ranked by the p value of the t-test, and GSS-AML is constructed based on the selected probe sets.

GSS-AML is shown in Figure 6.2. In the figure, most of the diagnostic samples are located nearby the disease centroid calculated from MILE-AML, while samples collected at later time points (Day > D25) are mostly located nearby the normal centroid calculated from MILE-NMB, and the rest samples are located between the two classes. Although the time points of the GEPs

Table 6.2: ASD and ESD of GSS-AML. Relapses are highlighted in the table.

Rank	SAMPLE	ASD	Outcome		Rank	Sample	ESD	Outcome
1	**R318-D5**	**0.28**	**R**		**1**	**R318-D33**	**-11.03**	**R**
2	**KL473-D32**	**3.04**	**R**		**2**	**R318-D5**	**0.04**	**R**
3	**KL343-D36**	**4.33**	**R**		**3**	**KL473-D32**	**2.83**	**R**
4	KL448-D17	8.11			**4**	**KL343-D36**	**3.34**	**R**
5	KL505-D14	10.61			5	KL448-D17	6.99	
6	R474-D33	11.52			6	KL505-D14	10.33	
7	**R318-D33**	**20.10**	**R**		7	R474-D33	11.31	
8	R474-D60	25.67			8	R474-D60	25.62	
9	KL336-D31	27.14			9	KL336-D31	26.65	
10	KL505-D45	31.07			10	KL505-D45	31.04	
11	KKH14-D36	35.61			11	KKH14-D36	35.61	
12	KL448-D51	39.71			12	KL448-D51	39.67	

are not synchronized in our dataset, the transition of genetic status from disease to normal is obvious. Thus, we claim the concept of GSS is valid in AML.

6.4 Relapse Prediction

We calculate ASD and ESD to predict the relapses. The results are shown in Table 6.2. Both metrics show a very promising value in the prediction. ESD outperforms ASD by capturing the negative GSS of R318-D33.

CHAPTER 7

CONCLUSION

7.1 Conclusion

GEP-based subtype classification of childhood ALL is a successful story of bioinformatics application in modern cancer research (Yeoh et al. 2002). By selecting genes exclusively expressing in each of the 6 disease subtypes, one can train a computational model to accurately classify the disease subtypes of new cases. This idea is later generalized to adult ALL and AML, and in both cases, GEP proves its value in disease diagnosis (Haferlach et al. 2010). A possible explanation to the success of the method is that chromosomal translocations caused abnormal gene expression patterns are reserved in disease subtypes, and they are catchable by high-throughput GEP technology.

Although GEP is valuable in disease diagnosis, its application in the relapse prediction of childhood ALL remains limited. Since contemporary management of patients with childhood

ALL tailors the intensity of therapy corresponding to a patient's risk of relapse, thereby maximizes cure and minimizes toxic side effects, it is crucial to accurately assign the risk of relapse upfront to optimize the treatment. Current risk assignment is based on a number of clinical and biological factors, such as, age, white blood cell count, DNA index, karyotype, recurrent translocations, early morphologic response and MRD, in which MRD is considered as the most predictive factor (Pui et al. 2001). Nevertheless, there are still about 20% of patients suffering from unpredicted relapses. For this reason, scientists are trying to discover new prognostic factors to improve the prediction of relapse by gene expression analysis. Several studies have been done to predict relapses based on diagnostic GEPs (Bhojwani et al. 2008, Kang et al. 2010, Holleman et al. 2004, Lugthart et al. 2005). However, there is little evidence to support their discoveries to be generalized to other studies, and the biological fundamental between the identified gene expression patterns and the relapses is still poorly understood.

We generate time-series GEPs to explore genetic response to disease treatment. This is the first time that time-series GEPs are used in a leukemia study. Through unsupervised hierarchical clustering and genetic signature dissolution analysis, we gain several interesting observations: 1) the samples collected at the same time point tend to be clustered together; 2) the samples of the early time points (D0 and D8) form several large subtype-related clusters, and this kind of clusters cannot be observed in the samples of late time points (D15 and D33); 3) the post-treatment samples of the relapses tend to be clustered with the pre-treatment samples of the same patients; 4) leukemic genetic signatures are gradually dissolved into the background during the process of treatment. These observations suggest that leukemic cells are gradually removed

during treatment, and the patients of different subtypes are eventually mixed up due to the removal of subtype-associated leukemic genetic signatures in GEPs.

We construct the global GSS model to quantitatively mimic the reduction of leukemic cells during the treatment of childhood ALL. As a result, the high-dimensional gene expression data are compressed into a 3-dimensional space, where each position in the space indicates a possible genetic status. In the global GSS model, diagnostic samples are observed to shift towards normal samples in the order of D0 → D8 → D15 → D33 → Normal. This observation is consistent with the result of unsupervised hierarchical clustering. In addition, the drug responsive genes we have identified for the construction of the global GSS model explain the fundamental of the genetic shift with two mechanisms: 1) the reconstruction of immune system and the restoration of normal hematogenesis, and 2) the suppression of the negative regulation of apoptosis.

We carry out our prediction of relapse by assuming the importance of early response to treatment. This is based on the hypothesis that if a patient is in low risk of relapse, the patient should be sensitive to disease treatment, and thus the post-treatment GEP should be different enough from the pre-treatment GEP. Practically, we introduce three GSS distance metrics, ASD, ESD, and ESR to calculate the difference between pre- and post-treatment genetic status. Our results suggest ESD-D8 has the best performance in relapse prediction, with an overall accuracy of 87.5%, when compared to the accuracy of several prevailing clinical and GEP-based protocols ranging from 62.1%-69.1%.

We evaluate our theory in an independent AML dataset consisting of 8 patients. Although AML and ALL are two different diseases, the treatment procedures of them are the same, both composed of an induction phase and a consolidation phase. However, the overall five-year

survival rate of AML is only 40%, which is much lower than that of ALL. Relapse is the major reverse event of AML. We construct GSS-AML to model our data. In the model, although the time points of the post-treatment samples are not synchronized, the pattern of GSS along treatment course can still be observed. Furthermore, both ASD and ESD show a very promising value in the relapse prediction of AML, where ESD outperforms ASD by capturing the negative shifting of R318-D33.

7.2 Future Work

We have demonstrated in this study that GSS-based method outperforms MRD-based method in the relapse prediction of childhood ALL. The essential of GSS-based relapse prediction is to consider the sensitivity of leukemic cells in a patient by calculating the difference between pre- and post-treatment GEPs of the patient. From the view of system, as we do not use the class labels (relapse vs. remission) in drug responsive gene selection and GSS model construction, the process of our prediction can be considered as an unsupervised process. Nevertheless, although GSS shows its advantage over MRD in relapse prediction in our study, it is probably yet too early to conclude GSS has a stronger prognostic strength than does MRD in general. This is because that MRD-based method has been evaluated in practice for over 20 years, while our method has only been tested in our own dataset. Thus, a very important future work is to test the validity of GSS-based relapse prediction in more cases. A new clinical trial, Malaysia-Singapore ALL 2010 trial, has been initiated for this purpose.

GSS distance is used only in univariate analysis of relapse prediction. The quantitative relationship between GSS distance and other clinical factors is still unknown. In addition to its independent prognostic strength, we are interested in knowing whether GSS distance could help to improve the current risk stratification system. A possible solution would exist in bivariate analysis with MRD, or in multivariate analysis with more clinical factors, such as age and white blood cell count. Generally, GEP is a very different information source from conventional clinical information, and we thus expect GSS distance to have complementary value in relapse prediction to sophisticated methods.

The ultimate purpose of our research is to look for new clinical solutions and to build applicable software to improve the treatment of childhood ALL. Since 2002, we have made several important breakthroughs in gene expression analysis of the disease. We are considering integrating these research discoveries into a practical software package. This software will be capable of disease diagnosis, subtype classification, subtype discovery, risk stratification and MRD detection. We hope our software would assist more clinicians in daily decision making and benefit more leukemia patients.

APPENDIX A

DRUG RESPONSIVE GENE

Table A.1: Drug responsive genes of T-ALL subtype.

Probe Set ID	Gene Symbol	Gene Title	*p* Value	Fold Change (T-ALL/Normal)
201416_at	SOX4	SRY (sex determining region Y)-box 4	1.06E-98	18.63
201417_at	SOX4	SRY (sex determining region Y)-box 4	6.55E-94	9.97
201029_s_at	CD99	CD99 molecule	2.29E-89	5.63
211071_s_at	MLLT11	myeloid/lymphoid or mixed-lineage leukemia; translocated to, 11	1.02E-85	12.52
204529_s_at	TOX	thymocyte selection-associated high mobility group box	6.36E-77	14.40

204636_at	COL17A1	collagen, type XVII, alpha 1	1.18E-74	0.04
204639_at	ADA	adenosine deaminase	1.16E-73	9.60
201418_s_at	SOX4	SRY (sex determining region Y)-box 4	2.25E-73	18.69
216705_s_at	ADA	adenosine deaminase	4.80E-73	12.62
213668_s_at	SOX4	SRY (sex determining region Y)-box 4	5.38E-73	35.16
202760_s_at	PALM2-AKAP2	PALM2-AKAP2 readthrough transcript	1.66E-70	10.26
213539_at	CD3D	CD3d molecule, delta (CD3-TCR complex)	1.60E-67	10.64
206390_x_at	PF4	platelet factor 4	2.53E-65	0.03
202242_at	TSPAN7	tetraspanin 7	2.57E-65	21.94
201028_s_at	CD99	CD99 molecule	1.64E-63	6.24
202759_s_at	AKAP2	A kinase (PRKA) anchor protein 2	2.59E-63	6.68
214997_at	SCAI	suppressor of cancer cell invasion	3.21E-63	8.51
213437_at	RUFY3	RUN and FYVE domain containing 3	4.41E-63	10.12
204173_at	MYL6B	myosin, light chain 6B, alkali, smooth muscle and non-muscle	5.87E-62	3.95
210116_at	SH2D1A	SH2 domain protein 1A	1.34E-61	15.79
203787_at	SSBP2	single-stranded DNA binding protein 2	1.40E-60	5.46
218641_at	LOC65998	hypothetical protein LOC65998	2.08E-60	4.18
209473_at	ENTPD1	ectonucleoside triphosphate diphosphohydrolase 1	4.14E-60	0.14
200983_x_at	CD59	CD59 molecule, complement regulatory protein	4.18E-60	0.13
214298_x_at	SEPT6	septin 6	4.23E-60	5.48
34726_at	CACNB3	calcium channel, voltage-dependent, beta 3 subunit	4.64E-60	8.38
55872_at	ZNF512B	zinc finger protein 512B	7.10E-60	4.49
218865_at	MOSC1	MOCO sulphurase C-terminal domain containing 1	8.48E-59	0.06
213666_at	SEPT6	septin 6	1.24E-58	6.65
221203_s_at	YEATS2	YEATS domain containing 2	1.33E-58	2.63
204530_s_at	TOX	thymocyte selection-associated high mobility group box	1.72E-58	9.50

220359_s_at	ARPP-21	cyclic AMP-regulated phosphoprotein, 21 kD	1.89E-58	33.33
210638_s_at	FBXO9	F-box protein 9	1.03E-57	0.36
202804_at	ABCC1	ATP-binding cassette, sub-family C (CFTR/MRP), member 1	2.95E-57	3.00
215307_at	ZNF529	zinc finger protein 529	4.16E-57	3.85
219036_at	CEP70	centrosomal protein 70kDa	6.23E-57	6.33
221810_at	RAB15	RAB15, member RAS onocogene family	6.38E-57	10.57
206656_s_at	C20orf3	chromosome 20 open reading frame 3	1.08E-56	0.28
208792_s_at	CLU	Clusterin	3.06E-56	0.18
202789_at	PLCG1	phospholipase C, gamma 1	9.60E-56	5.58
208791_at	CLU	Clusterin	9.71E-56	0.06
202671_s_at	PDXK	pyridoxal (pyridoxine, vitamin B6) kinase	1.38E-55	0.23
213048_s_at	---	---	1.53E-55	2.00
222344_at	---	---	1.70E-55	15.07
201321_s_at	SMARCC2	SWI/SNF related, matrix associated, actin dependent regulator of chromatin, subfamily c, member 2	2.66E-55	2.16
200631_s_at	SET	SET nuclear oncogene	2.89E-55	2.21
213430_at	RUFY3	RUN and FYVE domain containing 3	6.75E-55	6.05
210140_at	CST7	cystatin F (leukocystatin)	7.91E-55	0.09
218081_at	C20orf27	chromosome 20 open reading frame 27	1.30E-54	0.25
218005_at	ZNF22	zinc finger protein 22 (KOX 15)	2.20E-54	2.99

Table A.2: Drug responsive genes of TEL-AML1 subtype.

Probe Set ID	Gene Symbol	Gene Title	*p* Value	Fold Change (TEL-AML1/Normal)
203373_at	SOCS2	suppressor of cytokine signaling 2	4.60E-91	21.67
222146_s_at	TCF4	transcription factor 4	2.45E-74	17.04
212012_at	PXDN	peroxidasin homolog (Drosophila)	3.07E-73	39.64
32625_at	NPR1	natriuretic peptide receptor A/guanylate cyclase A (atrionatriuretic peptide receptor A)	4.74E-73	46.04
212387_at	TCF4	transcription factor 4	6.45E-72	17.30
203372_s_at	SOCS2	suppressor of cytokine signaling 2	6.01E-71	27.21
212386_at	TCF4	transcription factor 4	3.26E-70	14.14
203355_s_at	PSD3	pleckstrin and Sec7 domain containing 3	9.44E-70	100.22
203787_at	SSBP2	single-stranded DNA binding protein 2	7.72E-69	12.72
219686_at	STK32B	serine/threonine kinase 32B	1.06E-67	78.02
212013_at	PXDN	peroxidasin homolog (Drosophila)	2.09E-67	47.59
202806_at	DBN1	drebrin 1	1.88E-66	15.05
214761_at	ZNF423	zinc finger protein 423	2.54E-66	128.71
213891_s_at	TCF4	transcription factor 4	9.26E-66	19.04
54037_at	HPS4	Hermansky-Pudlak syndrome 4	2.48E-64	10.65
203611_at	TERF2	telomeric repeat binding factor 2	3.50E-64	26.92
207030_s_at	CSRP2	cysteine and glycine-rich protein 2	8.63E-64	104.54
204257_at	FADS3	fatty acid desaturase 3	1.01E-63	8.75
203753_at	TCF4	transcription factor 4	1.38E-63	12.47
208794_s_at	SMARCA4	SWI/SNF related, matrix associated, actin dependent regulator of chromatin, subfamily a, member 4	2.42E-63	5.07
218966_at	MYO5C	myosin VC	1.54E-62	11.96

212385_at	TCF4	transcription factor 4	1.72E-62	18.51
214728_x_at	SMARCA4	SWI/SNF related, matrix associated, actin dependent regulator of chromatin, subfamily a, member 4	1.03E-61	4.14
219938_s_at	PSTPIP2	proline-serine-threonine phosphatase interacting protein 2	2.11E-61	0.10
208056_s_at	CBFA2T3	core-binding factor, runt domain, alpha subunit 2; translocated to, 3	2.94E-61	12.31
218613_at	PSD3	pleckstrin and Sec7 domain containing 3	3.01E-61	42.49
213720_s_at	SMARCA4	SWI/SNF related, matrix associated, actin dependent regulator of chromatin, subfamily a, member 4	1.74E-60	5.06
218217_at	SCPEP1	serine carboxypeptidase 1	2.84E-60	0.20
211071_s_at	MLLT11	myeloid/lymphoid or mixed-lineage leukemia (trithorax homolog, Drosophila); translocated to, 11	6.64E-60	7.73
201015_s_at	JUP	junction plakoglobin	1.20E-59	31.23
206591_at	RAG1	recombination activating gene 1	2.03E-59	149.73
209153_s_at	TCF3	transcription factor 3 (E2A immunoglobulin enhancer binding factors E12/E47)	2.96E-59	6.60
219753_at	STAG3	stromal antigen 3	7.25E-59	28.24
209514_s_at	RAB27A	RAB27A, member RAS oncogene family	1.32E-58	0.13
212812_at	---	---	1.59E-58	6.74
210829_s_at	SSBP2	single-stranded DNA binding protein 2	2.30E-58	11.00
203910_at	ARHGAP29	Rho GTPase activating protein 29	3.32E-58	40.35
212382_at	TCF4	transcription factor 4	3.36E-58	18.39
210094_s_at	PARD3	par-3 partitioning defective 3 homolog (C. elegans)	6.05E-58	18.34
209035_at	MDK	midkine (neurite growth-promoting factor 2)	1.09E-57	35.46
202519_at	MLXIP	MLX interacting protein	2.04E-57	5.61
206398_s_at	CD19	CD19 molecule	2.52E-57	19.93
218988_at	SLC35E3	solute carrier family 35, member E3	4.02E-57	17.91

202039_at	MYO18A	myosin XVIIIA	4.54E-57	7.10
211026_s_at	MGLL	monoglyceride lipase	5.82E-57	0.16
209199_s_at	MEF2C	myocyte enhancer factor 2C	7.34E-57	15.57
204849_at	TCFL5	transcription factor-like 5 (basic helix-loop-helix)	9.59E-57	20.16
213702_x_at	ASAH1	N-acylsphingosine amidohydrolase (acid ceramidase) 1	2.56E-56	0.24
209583_s_at	CD200	CD200 molecule	3.98E-56	18.12
210980_s_at	ASAH1	N-acylsphingosine amidohydrolase (acid ceramidase) 1	5.70E-56	0.20

Table A.3: Drug responsive genes of Hyperdiploid>50 subtype.

UNIQID	**Gene Symbol**	**Gene Title**	***p* Value**	**Fold Change (Hyperdiploid>50/Normal)**
203373_at	SOCS2	suppressor of cytokine signaling 2	5.26E-73	17.77
222146_s_at	TCF4	transcription factor 4	3.19E-67	18.77
203372_s_at	SOCS2	suppressor of cytokine signaling 2	6.53E-63	23.96
32625_at	NPR1	natriuretic peptide receptor A/guanylate cyclase A (atrionatriuretic peptide receptor A)	2.15E-62	44.13
201005_at	CD9	CD9 molecule	2.19E-62	14.83
212387_at	TCF4	transcription factor 4	1.14E-59	18.22
212386_at	TCF4	transcription factor 4	1.95E-59	15.31
212012_at	PXDN	peroxidasin homolog (Drosophila)	2.05E-59	38.04
218694_at	ARMCX1	armadillo repeat containing, X-linked 1	8.95E-59	9.25
202039_at	TIAF1	TGFB1-induced anti-apoptotic factor 1	6.25E-56	9.36

212385_at	TCF4	transcription factor 4	8.54E-56	20.70
213891_s_at	TCF4	transcription factor 4	2.15E-55	22.00
208370_s_at	RCAN1	regulator of calcineurin 1	8.77E-55	12.26
201540_at	FHL1	four and a half LIM domains 1	1.24E-54	10.48
212013_at	PXDN	peroxidasin homolog (Drosophila)	1.28E-54	48.40
211026_s_at	MGLL	monoglyceride lipase	2.36E-54	0.17
204122_at	TYROBP	TYRO protein tyrosine kinase binding protein	7.91E-54	0.07
209365_s_at	ECM1	extracellular matrix protein 1	2.83E-53	31.43
203753_at	TCF4	transcription factor 4	1.10E-52	12.77
202806_at	DBN1	drebrin 1	3.48E-52	9.26
211275_s_at	GYG1	glycogenin 1	4.98E-52	0.14
206001_at	NPY	neuropeptide Y	5.06E-52	86.18
221766_s_at	FAM46A	family with sequence similarity 46, member A	8.25E-52	0.04
202908_at	WFS1	Wolfram syndrome 1 (wolframin)	9.64E-52	16.41
205786_s_at	ITGAM	integrin, alpha M (complement component 3 receptor 3 subunit)	2.31E-51	0.07
207030_s_at	CSRP2	cysteine and glycine-rich protein 2	2.54E-51	74.97
204232_at	FCER1G	Fc fragment of IgE, high affinity I, receptor for; gamma polypeptide	3.48E-51	0.05
205237_at	FCN1	ficolin (collagen/fibrinogen domain containing) 1	8.98E-51	0.02
202598_at	S100A13	S100 calcium binding protein A13	1.84E-50	7.66
219694_at	FAM105A	family with sequence similarity 105, member A	2.37E-50	0.09
217728_at	S100A6	S100 calcium binding protein A6	3.85E-50	0.12
204620_s_at	VCAN	Versican	4.57E-50	0.03
219686_at	STK32B	serine/threonine kinase 32B	5.47E-50	52.11

203355_s_at	PSD3	pleckstrin and Sec7 domain containing 3	6.14E-50	68.88
206674_at	FLT3	fms-related tyrosine kinase 3	1.01E-49	19.17
201012_at	ANXA1	annexin A1	1.33E-49	0.04
221773_at	ELK3	ELK3, ETS-domain protein (SRF accessory protein 2)	1.38E-49	10.06
218865_at	MOSC1	MOCO sulphurase C-terminal domain containing 1	1.41E-49	0.03
220088_at	C5AR1	complement component 5a receptor 1	1.64E-49	0.03
218988_at	SLC35E3	solute carrier family 35, member E3	3.05E-49	14.48
208438_s_at	FGR	Gardner-Rasheed feline sarcoma viral (v-fgr) oncogene homolog	3.60E-49	0.06
202788_at	MAPKAPK3	mitogen-activated protein kinase-activated protein kinase 3	4.41E-49	0.23
209696_at	FBP1	fructose-1,6-bisphosphatase 1	4.97E-49	0.10
218872_at	TESC	Tescalcin	7.53E-49	0.13
201360_at	CST3	cystatin C	8.21E-49	0.08
218005_at	ZNF22	zinc finger protein 22 (KOX 15)	8.97E-49	4.17
215543_s_at	LARGE	like-glycosyltransferase	1.93E-48	26.19
211429_s_at	SERPINA1	serpin peptidase inhibitor, clade A (alpha-1 antiproteinase, antitrypsin), member 1	2.32E-48	0.05
201425_at	ALDH2	aldehyde dehydrogenase 2 family (mitochondrial)	2.72E-48	0.14
214761_at	ZNF423	zinc finger protein 423	3.32E-48	49.50

Table A.4: Drug responsive genes of E2A-PBX1 subtype.

Probe Set ID	Gene Symbol	Gene Title	*p* Value	Fold Change (E2A-PBX1/Normal)
212012_at	PXDN	peroxidasin homolog (Drosophila)	1.89E-58	48.98
211404_s_at	APLP2	amyloid beta (A4) precursor-like protein 2	4.77E-55	0.15
212013_at	PXDN	peroxidasin homolog (Drosophila)	3.29E-53	62.97
222146_s_at	TCF4	transcription factor 4	2.50E-52	17.16
208690_s_at	PDLIM1	PDZ and LIM domain 1	3.08E-51	6.65
201425_at	ALDH2	aldehyde dehydrogenase 2 family (mitochondrial)	1.22E-49	0.10
208702_x_at	APLP2	amyloid beta (A4) precursor-like protein 2	2.63E-49	0.16
204674_at	LRMP	lymphoid-restricted membrane protein	2.98E-49	11.48
35974_at	LRMP	lymphoid-restricted membrane protein	1.21E-48	12.78
203787_at	SSBP2	single-stranded DNA binding protein 2	1.57E-48	10.32
218641_at	LOC65998	hypothetical protein LOC65998	2.37E-48	5.18
201005_at	CD9	CD9 molecule	2.43E-47	10.87
211071_s_at	MLLT11	myeloid/lymphoid or mixed-lineage leukemia (trithorax homolog, Drosophila); translocated to, 11	2.96E-47	9.07
204214_s_at	RAB32	RAB32, member RAS oncogene family	3.46E-47	0.08
204257_at	FADS3	fatty acid desaturase 3	4.25E-47	7.59
201719_s_at	EPB41L2	erythrocyte membrane protein band 4.1-like 2	1.06E-46	9.69
206398_s_at	CD19	CD19 molecule	1.12E-46	25.77
211178_s_at	PSTPIP1	proline-serine-threonine phosphatase interacting protein 1	1.30E-46	0.13
219938_s_at	PSTPIP2	proline-serine-threonine phosphatase interacting protein 2	3.86E-46	0.08
201417_at	SOX4	SRY (sex determining region Y)-box 4	5.94E-46	9.01
213358_at	KIAA0802	KIAA0802	9.06E-46	64.66
208949_s_at	LGALS3	lectin, galactoside-binding, soluble, 3	1.13E-45	0.07

202806_at	DBN1	drebrin 1	2.00E-45	13.74
219694_at	FAM105A	family with sequence similarity 105, member A	2.76E-45	0.07
212387_at	TCF4	transcription factor 4	4.20E-45	15.81
203922_s_at	CYBB	cytochrome b-245, beta polypeptide	6.17E-45	0.08
201792_at	AEBP1	AE binding protein 1	6.73E-45	28.37
201061_s_at	STOM	stomatin	8.16E-45	0.10
201060_x_at	STOM	stomatin	1.13E-44	0.09
214761_at	ZNF423	zinc finger protein 423	1.45E-44	66.70
204173_at	MYL6B	myosin, light chain 6B, alkali, smooth muscle and non-muscle	1.62E-44	3.73
202788_at	MAPKAPK3	mitogen-activated protein kinase-activated protein kinase 3	4.10E-44	0.19
210829_s_at	SSBP2	single-stranded DNA binding protein 2	5.15E-44	9.81
212386_at	TCF4	transcription factor 4	5.86E-44	12.84
213891_s_at	TCF4	transcription factor 4	8.12E-44	17.33
212385_at	TCF4	transcription factor 4	4.90E-43	17.45
201416_at	SOX4	SRY (sex determining region Y)-box 4	5.79E-43	17.80
208370_s_at	RCAN1	regulator of calcineurin 1	8.54E-43	8.11
201506_at	TGFBI	transforming growth factor, beta-induced, 68kDa	1.29E-42	0.04
215806_x_at	TARP /// TRGC2	TCR gamma alternate reading frame protein /// T cell receptor gamma constant 2	1.71E-42	0.12
211987_at	TOP2B	topoisomerase (DNA) II beta 180kDa	1.77E-42	4.27
216920_s_at	TARP /// TRGC2	TCR gamma alternate reading frame protein /// T cell receptor gamma constant 2	3.18E-42	0.11
212599_at	AUTS2	autism susceptibility candidate 2	5.12E-42	16.11
212197_x_at	MPRIP	myosin phosphatase Rho interacting protein	6.12E-42	3.72
204949_at	ICAM3	intercellular adhesion molecule 3	8.04E-42	0.14
217763_s_at	RAB31	RAB31, member RAS oncogene family	8.30E-42	0.04

205237_at	FCN1	ficolin (collagen/fibrinogen domain containing) 1	8.72E-42	0.04
36499_at	CELSR2	cadherin, EGF LAG seven-pass G-type receptor 2 (flamingo homolog, Drosophila)	1.46E-41	7.91
209574_s_at	C18orf1	chromosome 18 open reading frame 1	2.16E-41	9.48
209813_x_at	TARP	TCR gamma alternate reading frame protein	3.00E-41	0.11

Table A.5: Drug responsive genes of BCR-ABL subtype.

Probe Set ID	**Gene Symbol**	**Gene Title**	***p* Value**	**Fold Change (BCR-ABL/Normal)**
203373_at	SOCS2	suppressor of cytokine signaling 2	1.98E-130	22.03
203372_s_at	SOCS2	suppressor of cytokine signaling 2	2.59E-107	37.15
212012_at	PXDN	peroxidasin homolog (Drosophila)	1.06E-92	31.34
203355_s_at	PSD3	pleckstrin and Sec7 domain containing 3	1.22E-91	63.31
212013_at	PXDN	peroxidasin homolog (Drosophila)	2.73E-91	43.46
218966_at	MYO5C	myosin VC	4.12E-89	12.88
219686_at	STK32B	serine/threonine kinase 32B	8.71E-89	43.78
207030_s_at	CSRP2	cysteine and glycine-rich protein 2	1.14E-87	63.60
201540_at	FHL1	four and a half LIM domains 1	1.13E-84	9.77
201029_s_at	CD99	CD99 molecule	6.67E-83	5.56
209365_s_at	ECM1	extracellular matrix protein 1	1.16E-78	41.47
32625_at	NPR1	natriuretic peptide receptor A/guanylate cyclase A (atrionatriuretic peptide receptor A)	5.30E-78	34.93

206398_s_at	CD19	CD19 molecule	1.78E-77	19.73
218613_at	PSD3	pleckstrin and Sec7 domain containing 3	7.17E-77	28.97
214761_at	ZNF423	zinc finger protein 423	1.67E-74	58.41
222146_s_at	TCF4	transcription factor 4	9.23E-74	8.28
201015_s_at	JUP	junction plakoglobin	3.41E-73	16.80
210487_at	DNTT	deoxynucleotidyltransferase, terminal	7.36E-73	46.00
211126_s_at	CSRP2	cysteine and glycine-rich protein 2	1.07E-71	22.32
203787_at	SSBP2	single-stranded DNA binding protein 2	1.52E-71	7.26
212387_at	TCF4	transcription factor 4	3.26E-70	8.71
209576_at	GNAI1	guanine nucleotide binding protein (G protein), alpha inhibiting activity polypeptide 1	4.30E-69	11.56
213891_s_at	TCF4	transcription factor 4	5.75E-69	9.79
212386_at	TCF4	transcription factor 4	9.79E-69	7.49
204030_s_at	SCHIP1	schwannomin interacting protein 1	6.41E-68	18.37
212385_at	TCF4	transcription factor 4	1.38E-67	9.09
201028_s_at	CD99	CD99 molecule	3.05E-67	6.48
210829_s_at	SSBP2	single-stranded DNA binding protein 2	9.89E-67	7.43
202123_s_at	ABL1	c-abl oncogene 1, receptor tyrosine kinase	4.08E-66	3.82
204636_at	COL17A1	collagen, type XVII, alpha 1	4.95E-65	0.04
202945_at	FPGS	folylpolyglutamate synthase	5.12E-65	4.97
209679_s_at	SMAGP	small trans-membrane and glycosylated protein	8.96E-65	11.37
207655_s_at	BLNK	B-cell linker	2.78E-64	14.78
203753_at	TCF4	transcription factor 4	2.03E-63	6.79
205983_at	DPEP1	dipeptidase 1 (renal)	6.02E-63	51.95
212675_s_at	CEP68	centrosomal protein 68kDa	1.65E-62	5.97
209199_s_at	MEF2C	myocyte enhancer factor 2C	3.43E-62	8.59

201416_at	SOX4	SRY (sex determining region Y)-box 4	4.45E-62	9.83
212488_at	COL5A1	collagen, type V, alpha 1	1.15E-61	25.48
34726_at	CACNB3	calcium channel, voltage-dependent, beta 3 subunit	1.62E-61	8.00
1007_s_at	DDR1	discoidin domain receptor tyrosine kinase 1	4.15E-61	9.09
211031_s_at	CLIP2	CAP-GLY domain containing linker protein 2	7.14E-61	14.07
200983_x_at	CD59	CD59 molecule, complement regulatory protein	1.61E-60	0.18
205795_at	NRXN3	neurexin 3	2.69E-60	85.26
203354_s_at	PSD3	pleckstrin and Sec7 domain containing 3	2.96E-60	30.43
210638_s_at	FBXO9	F-box protein 9	3.86E-60	0.33
208690_s_at	PDLIM1	PDZ and LIM domain 1	9.12E-60	5.63
202242_at	TSPAN7	tetraspanin 7	1.28E-59	16.01
202598_at	S100A13	S100 calcium binding protein A13	4.21E-59	9.53
221286_s_at	MGC29506	hypothetical protein MGC29506	8.80E-59	7.97

Table A.6: Drug responsive genes of MLL subtype.

Probe Set ID	**Gene Symbol**	**Gene Title**	***p* Value**	**Fold Change (MLL/Normal)**
203373_at	SOCS2	suppressor of cytokine signaling 2	4.33E-79	21.74
207030_s_at	CSRP2	cysteine and glycine-rich protein 2	1.19E-78	160.22
211066_x_at	PCDHGA1	protocadherin gamma subfamily A, 1	5.33E-68	20.08
203372_s_at	SOCS2	suppressor of cytokine signaling 2	6.82E-68	30.70
211126_s_at	CSRP2	cysteine and glycine-rich protein 2	7.61E-66	51.71

217963_s_at	NGFRAP1	nerve growth factor receptor (TNFRSF16) associated protein 1	1.39E-64	0.05
206398_s_at	CD19	CD19 molecule	9.28E-62	18.07
218865_at	MOSC1	MOCO sulphurase C-terminal domain containing 1	2.00E-61	0.04
206674_at	FLT3	fms-related tyrosine kinase 3	3.44E-61	22.89
209170_s_at	GPM6B	glycoprotein M6B	6.41E-61	33.73
201874_at	MPZL1	myelin protein zero-like 1	9.04E-61	4.66
36553_at	ASMTL	acetylserotonin O-methyltransferase-like	1.85E-60	4.63
209079_x_at	PCDHGA1	protocadherin gamma subfamily A, 1	3.90E-60	15.70
205717_x_at	PCDHGA1	protocadherin gamma subfamily A, 1	4.86E-60	19.07
210638_s_at	FBXO9	F-box protein 9	7.15E-60	0.32
201416_at	SOX4	SRY (sex determining region Y)-box 4	1.36E-59	14.92
208949_s_at	LGALS3	lectin, galactoside-binding, soluble, 3	2.09E-59	0.06
209167_at	GPM6B	glycoprotein M6B	3.00E-59	27.85
204636_at	COL17A1	collagen, type XVII, alpha 1	3.87E-59	0.04
204214_s_at	RAB32	RAB32, member RAS oncogene family	4.45E-59	0.08
201060_x_at	STOM	stomatin	1.08E-58	0.09
211178_s_at	PSTPIP1	proline-serine-threonine phosphatase interacting protein 1	1.23E-58	0.15
204069_at	MEIS1	Meis homeobox 1	1.57E-58	20.82
200983_x_at	CD59	CD59 molecule, complement regulatory protein	1.85E-58	0.09
221485_at	B4GALT5	UDP-Gal:betaGlcNAc beta 1,4-galactosyltransferase, polypeptide 5	3.69E-58	0.10
209168_at	GPM6B	glycoprotein M6B	4.72E-58	12.15
215925_s_at	CD72	CD72 molecule	7.21E-58	30.98
204173_at	MYL6B	myosin, light chain 6B, alkali, smooth muscle and non-muscle	8.64E-58	3.84

209514_s_at	RAB27A	RAB27A, member RAS oncogene family	2.62E-57	0.13
202945_at	FPGS	folylpolyglutamate synthase	3.41E-57	4.78
215836_s_at	PCDHGA1	protocadherin gamma subfamily A, 1	1.27E-56	26.89
201417_at	SOX4	SRY (sex determining region Y)-box 4	1.92E-56	7.02
210951_x_at	RAB27A	RAB27A, member RAS oncogene family	2.57E-56	0.12
202332_at	CSNK1E	casein kinase 1, epsilon	3.80E-56	3.84
209199_s_at	MEF2C	myocyte enhancer factor 2C	4.29E-56	10.96
203355_s_at	PSD3	pleckstrin and Sec7 domain containing 3	9.39E-56	24.44
209949_at	NCF2	neutrophil cytosolic factor 2	9.55E-56	0.07
208302_at	HMHB1	histocompatibility (minor) HB-1	1.84E-55	26.02
218641_at	LOC65998	hypothetical protein LOC65998	2.05E-55	5.13
203795_s_at	BCL7A	B-cell CLL/lymphoma 7A	6.00E-55	9.49
208791_at	CLU	Clusterin	6.46E-55	0.04
201540_at	FHL1	four and a half LIM domains 1	1.02E-54	7.27
210987_x_at	TPM1	tropomyosin 1 (alpha)	1.13E-54	0.16
201875_s_at	MPZL1	myelin protein zero-like 1	2.27E-54	4.66
204639_at	ADA	adenosine deaminase	3.17E-54	6.79
206390_x_at	PF4	platelet factor 4	4.25E-54	0.02
205237_at	FCN1	ficolin (collagen/fibrinogen domain containing) 1	4.44E-54	0.03
205786_s_at	ITGAM	integrin, alpha M (complement component 3 receptor 3 subunit)	4.54E-54	0.08
203922_s_at	CYBB	cytochrome b-245, beta polypeptide	5.23E-54	0.10
218332_at	BEX1	brain expressed, X-linked 1	1.76E-53	0.04

Table A.7: Drug responsive genes of other subtypes.

Probe Set ID	Gene Symbol	Gene Title	*p* Value	Fold Change (Other/Normal)
202837_at	TRAFD1	TRAF-type zinc finger domain containing 1	4.97E-170	0.05
202829_s_at	VAMP7	vesicle-associated membrane protein 7	1.06E-153	17.66
202830_s_at	SLC37A4	solute carrier family 37 (glucose-6-phosphate transporter), member 4	1.30E-152	0.09
202804_at	ABCC1	ATP-binding cassette, sub-family C (CFTR/MRP), member 1	2.56E-150	0.16
202825_at	SLC25A4	solute carrier family 25 (mitochondrial carrier; adenine nucleotide translocator), member 4	1.85E-144	0.06
202843_at	DNAJB9	DnaJ (Hsp40) homolog, subfamily B, member 9	4.02E-143	0.13
2028_s_at	E2F1	E2F transcription factor 1	2.23E-131	0.06
202866_at	DNAJB12	DnaJ (Hsp40) homolog, subfamily B, member 12	2.46E-131	5.42
202824_s_at	TCEB1	transcription elongation factor B (SIII), polypeptide 1 (15kDa, elongin C)	4.61E-129	9.89
202836_s_at	TXNL4A	thioredoxin-like 4A	2.36E-122	76.16
202899_s_at	SFRS3	splicing factor, arginine/serine-rich 3	1.22E-121	6.69
202882_x_at	NOL7	nucleolar protein 7, 27kDa	7.88E-120	3.18
202822_at	LPP	LIM domain containing preferred translocation partner in lipoma	7.67E-119	10.75
202865_at	DNAJB12	DnaJ (Hsp40) homolog, subfamily B, member 12	1.09E-113	0.08
320_at	PEX6	peroxisomal biogenesis factor 6	1.56E-111	0.15
212013_at	PXDN	peroxidasin homolog (Drosophila)	7.77E-105	37.44
202874_s_at	ATP6V1C1	ATPase, H+ transporting, lysosomal 42kDa, V1 subunit C1	1.37E-104	6.83
202887_s_at	DDIT4	DNA-damage-inducible transcript 4	2.41E-104	31.66

204636_at	COL17A1	collagen, type XVII, alpha 1	6.44E-102	0.04
201416_at	SOX4	SRY (sex determining region Y)-box 4	9.11E-102	14.56
32091_at	SLC25A44	solute carrier family 25, member 44	3.35E-100	7.09
203355_s_at	PSD3	pleckstrin and Sec7 domain containing 3	2.23E-99	38.67
201417_at	SOX4	SRY (sex determining region Y)-box 4	3.44E-98	7.59
32029_at	PDPK1	3-phosphoinositide dependent protein kinase-1	1.78E-96	4.95
202854_at	HPRT1	hypoxanthine phosphoribosyltransferase 1	3.89E-96	4.06
202834_at	AGT	angiotensinogen (serpin peptidase inhibitor, clade A, member 8)	3.03E-95	0.02
212012_at	PXDN	peroxidasin homolog (Drosophila)	8.22E-95	28.12
202810_at	DRG1	developmentally regulated GTP binding protein 1	1.24E-94	5.64
206398_s_at	CD19	CD19 molecule	5.20E-94	16.38
218865_at	MOSC1	MOCO sulphurase C-terminal domain containing 1	2.11E-91	0.04
202883_s_at	PPP2R1B	protein phosphatase 2 (formerly 2A), regulatory subunit A, beta isoform	4.12E-91	0.16
214761_at	ZNF423	zinc finger protein 423	2.99E-90	49.35
201015_s_at	JUP	junction plakoglobin	3.24E-90	16.86
32088_at	BLZF1	basic leucine zipper nuclear factor 1	3.71E-88	0.07
206656_s_at	C20orf3	chromosome 20 open reading frame 3	1.01E-87	0.22
34726_at	CACNB3	calcium channel, voltage-dependent, beta 3 subunit	4.20E-87	7.82
213668_s_at	SOX4	SRY (sex determining region Y)-box 4	4.64E-82	24.25
202332_at	CSNK1E	casein kinase 1, epsilon	6.16E-82	4.03
202844_s_at	RALBP1	ralA binding protein 1	3.58E-80	5.09
215543_s_at	LARGE	like-glycosyltransferase	7.03E-80	15.06
203787_at	SSBP2	single-stranded DNA binding protein 2	1.05E-79	7.61
202880_s_at	CYTH1	cytohesin 1	1.07E-79	24.58

211031_s_at	CLIP2	CAP-GLY domain containing linker protein 2	1.62E-79	14.21
202855_s_at	SLC16A3	solute carrier family 16, member 3 (monocarboxylic acid transporter 4)	3.68E-79	0.01
202879_s_at	CYTH1	cytohesin 1	5.44E-79	0.06
202945_at	FPGS	folylpolyglutamate synthase	1.14E-78	4.68
201425_at	ALDH2	aldehyde dehydrogenase 2 family (mitochondrial)	1.51E-78	0.11
218613_at	PSD3	pleckstrin and Sec7 domain containing 3	2.20E-78	16.58
201418_s_at	SOX4	SRY (sex determining region Y)-box 4	4.56E-78	12.66
216041_x_at	GRN	Granulin	8.07E-77	0.15

BIBLIOGRAPHY

Affymetrix Expression Analysis Technical Manual.

Alizadeh, A. A., M. B. Eisen, R. E. Davis, C. Ma, I. S. Lossos, A. Rosenwald, J. C. Boldrick, H. Sabet, T. Tran, X. Yu, J. I. Powell, L. Yang, G. E. Marti, T. Moore, J. Hudson, L. Lu, D. B. Lewis, R. Tibshirani, G. Sherlock, W. C. Chan, T. C. Greiner, D. D. Weisenburger, J. O. Armitage, R. Warnke, R. Levy, W. Wilson, M. R. Grever, J. C. Byrd, D. Botstein, P. O. Brown & L. M. Staudt (2000) Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature,* 403**,** 503-511.

Ashburner, M., C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin & G. Sherlock (2000) Gene Ontology: tool for the unification of biology. *Nat Genet,* 25**,** 25-29.

Beesley, A. H., A. J. Cummings, J. R. Freitas, K. Hoffmann, M. J. Firth, J. Ford, N. H. de Klerk & U. R. Kees (2005) The gene expression signature of relapse in paediatric acute lymphoblastic leukaemia: implications for mechanisms of therapy failure. *Br J Haematol,* 131**,** 447-56.

Bhojwani, D., H. Kang, R. X. Menezes, W. Yang, H. Sather, N. P. Moskowitz, D. J. Min, J. W. Potter, R. Harvey, S. P. Hunger, N. Seibel, E. A. Raetz, R. Pieters, M. A. Horstmann, M. V. Relling, M. L. den Boer, C. L. Willman & W. L. Carroll (2008) Gene Expression Signatures Predictive of Early Response and Outcome in High-Risk Childhood Acute Lymphoblastic Leukemia: A Children's Oncology Group Study on Behalf of the Dutch Childhood Oncology Group and the German Cooperative Study Group for Childhood Acute Lymphoblastic Leukemia. *Journal of Clinical Oncology,* 26**,** 4376-4384.

Bhojwani, D., H. Kang, N. P. Moskowitz, D. J. Min, H. Lee, J. W. Potter, G. Davidson, C. L. Willman, M. J. Borowitz, I. Belitskaya-Levy, S. P. Hunger, E. A. Raetz & W. L. Carroll (2006)

Biologic pathways associated with relapse in childhood acute lymphoblastic leukemia: a Children's Oncology Group study. *Blood,* 108**,** 711-717.

Bittner, M., P. Meltzer, Y. Chen, Y. Jiang, E. Seftor, M. Hendrix, M. Radmacher, R. Simon, Z. Yakhini, A. Ben-Dor, N. Sampas, E. Dougherty, E. Wang, F. Marincola, C. Gooden, J. Lueders, A. Glatfelter, P. Pollock, J. Carpten, E. Gillanders, D. Leja, K. Dietrich, C. Beaudry, M. Berens, D. Alberts & V. Sondak (2000) Molecular classification of cutaneous malignant melanoma by gene expression profiling. *Nature,* 406**,** 536-40.

Bolstad, B. M., R. A. Irizarry, M. Åstrand & T. P. Speed (2003) A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics,* 19**,** 185-193.

Borowitz, M. J., M. Devidas, S. P. Hunger, W. P. Bowman, A. J. Carroll, W. L. Carroll, S. Linda, P. L. Martin, D. J. Pullen, D. Viswanatha, C. L. Willman, N. Winick & B. M. Camitta (2008) Clinical significance of minimal residual disease in childhood acute lymphoblastic leukemia and its relationship to other prognostic factors: a Children's Oncology Group study. *Blood,* 111**,** 5477-5485.

Breitling, R. & P. Herzyk (2005) Rank-based methods as a non-parametric alternative of the T-statistic for the analysis of biological microarray data. *Journal of Bioinformatics and Computational Biology,* 3**,** 1171-1190.

Brown, P. O. & D. Botstein (1999) Exploring the new world of the genome with DNA microarrays. *Nat Genet*.

Cawley, G. C. & N. L. C. Talbot (2006) Gene selection in cancer classification using sparse logistic regression with Bayesian regularization. *Bioinformatics,* 22**,** 2348-2355.

Cheok, M. H., W. Yang, C. H. Pui, J. R. Downing, C. Cheng, C. W. Naeve, M. V. Relling & W. E. Evans (2003) Treatment-specific changes in gene expression discriminate in vivo drug response in human leukemia cells. *Nat Genet,* 34**,** 85-90.

Cheson, B. D., J. M. Bennett, K. J. Kopecky, T. Büchner, C. L. Willman, E. H. Estey, C. A. Schiffer, H. Doehner, M. S. Tallman, T. A. Lister, F. Lo-Coco, R. Willemze, A. Biondi, W. Hiddemann, R. A. Larson, B. Löwenberg, M. A. Sanz, D. R. Head, R. Ohno & C. D. Bloomfield (2003) Revised Recommendations of the International Working Group for Diagnosis, Standardization of Response Criteria, Treatment Outcomes, and Reporting Standards for Therapeutic Trials in Acute Myeloid Leukemia. *Journal of Clinical Oncology,* 21**,** 4642-4649.

Colvin, G. A. & G. J. Elfenbein (2003) The latest treatment advances for acute myelogenous leukemia. *Med Health R I,* 86**,** 243-6.

Den Boer, M. L., D. O. Harms, R. Pieters, K. M. Kazemier, U. Göbel, D. Körholz, U. Graubner, R. J. Haas, N. Jorch, H. J. Spaar, G. J. L. Kaspers, W. A. Kamps, A. Van der Does-Van den Berg, E. R. Van Wering, A. J. P. Veerman & G. E. Janka-Schaub (2003) Patient Stratification Based on

Prednisolone-Vincristine-Asparaginase Resistance Profiles in Children With Acute Lymphoblastic Leukemia. *Journal of Clinical Oncology,* 21**,** 3262-3268.

Den Boer, M. L., M. van Slegtenhorst, R. X. De Menezes, M. H. Cheok, J. G. C. A. M. Buijs-Gladdines, S. T. C. J. M. Peters, L. J. C. M. Van Zutven, H. B. Beverloo, P. J. Van der Spek, G. Escherich, M. A. Horstmann, G. E. Janka-Schaub, W. A. Kamps, W. E. Evans & R. Pieters (2009) A subtype of childhood acute lymphoblastic leukaemia with poor treatment outcome: a genome-wide classification study. *Lancet Oncol,* 10**,** 125-134.

Edgar, R., M. Domrachev & A. E. Lash (2002) Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Research,* 30**,** 207-210.

Eisen, M. B., P. T. Spellman, P. O. Brown & D. Botstein (1998) Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences of the United States of America,* 95**,** 14863-14868.

Estey, E. & H. Döhner (2006) Acute myeloid leukaemia. *The Lancet,* 368**,** 1894-1907.

Goh, L. & N. Kasabov (2005) An integrated feature selection and classification method to select minimum number of variables on the case study of gene expression data. *Journal of Bioinformatics and Computational Biology,* 3**,** 1107-1136.

Golub, T. R., D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield & E. S. Lander (1999) Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring. *Science,* 286**,** 531-537.

Haferlach, T., A. Kohlmann, L. Wieczorek, G. Basso, G. T. Kronnie, M. C. Béné, J. De Vos, J. M. Hernández, W. K. Hofmann, K. I. Mills, A. Gilkes, S. Chiaretti, S. A. Shurtleff, T. J. Kipps, L. Z. Rassenti, A. E. Yeoh, P. R. Papenhausen, W. M. Liu, P. M. Williams & R. Foà (2010) Clinical Utility of Microarray-Based Gene Expression Profiling in the Diagnosis and Subclassification of Leukemia: Report From the International Microarray Innovations in Leukemia Study Group. *Journal of Clinical Oncology,* 28**,** 2529-2537.

Holleman, A., M. H. Cheok, M. L. den Boer, W. Yang, A. J. P. Veerman, K. M. Kazemier, D. Pei, C. Cheng, C. H. Pui, M. V. Relling, G. E. Janka-Schaub, R. Pieters & W. E. Evans (2004) Gene-Expression Patterns in Drug-Resistant Acute Lymphoblastic Leukemia Cells and Response to Treatment. *New England Journal of Medicine,* 351**,** 533-542.

Irizarry, R., B. Bolstad, F. Collin, L. Cope, B. Hobbs & T. Speed (2003a) Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Res,* 31**,** e15.

Irizarry, R., B. Hobbs, F. Collin, Y. Beazer-Barclay, K. Antonellis, U. Scherf & T. Speed (2003b) Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics,* 4**,** 249 - 264.

Irizarry, R., Z. Wu & H. Jaffee (2006) Comparison of Affymetrix GeneChip expression measures. *Bioinformatics,* 22**,** 789 - 794.

Kang, H., I. M. Chen, C. S. Wilson, E. J. Bedrick, R. C. Harvey, S. R. Atlas, M. Devidas, C. G. Mullighan, X. Wang, M. Murphy, K. Ar, W. Wharton, M. J. Borowitz, W. P. Bowman, D. Bhojwani, W. L. Carroll, B. M. Camitta, G. H. Reaman, M. A. Smith, J. R. Downing, S. P. Hunger & C. L. Willman (2010) Gene expression classifiers for relapse-free survival and minimal residual disease improve risk classification and outcome prediction in pediatric B-precursor acute lymphoblastic leukemia. *Blood,* 115**,** 1394-1405.

Liu, H., J. Li & L. Wong. 2004. Selection of patient samples and genes for outcome prediction. In *3rd International Computational Systems Bioinformatics Conference*, 382-392.

Lockhart, D. J., H. Dong, M. C. Byrne, M. T. Follettie, M. V. Gallo, M. S. Chee, M. Mittmann, C. Wang, M. Kobayashi, H. Norton & E. L. Brown (1996) Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nat Biotech,* 14**,** 1675-1680.

Lowenberg, B., J. R. Downing & A. Burnett (1999) Acute Myeloid Leukemia. *New England Journal of Medicine,* 341**,** 1051-1062.

Lugthart, S., M. H. Cheok, M. L. den Boer, W. Yang, A. Holleman, C. Cheng, C. H. Pui, M. V. Relling, G. E. Janka-Schaub, R. Pieters & W. E. Evans (2005) Identification of genes associated with chemotherapy crossresistance and treatment response in childhood acute lymphoblastic leukemia. *Cancer Cell,* 7**,** 375-386.

Olman, V., C. Hicks, P. Wang & Y. Xu (2006) Gene expression data analysis in subtypes of ovarian cancer using covariance analysis. *Journal of Bioinformatics and Computational Biology,* 4**,** 999-1014.

Pepper, S., E. Saunders, L. Edwards, C. Wilson & C. Miller (2007) The utility of MAS5 expression summary and detection call algorithms. *BMC Bioinformatics,* 8**,** 273.

Pui, C.-H., D. Campana & W. E. Evans (2001) Childhood acute lymphoblastic leukaemia: current status and future perspectives. *The Lancet Oncology,* 2**,** 597-607.

Pui, C.-H., D. Campana, D. Pei, W. P. Bowman, J. T. Sandlund, S. C. Kaste, R. C. Ribeiro, J. E. Rubnitz, S. C. Raimondi, M. Onciu, E. Coustan-Smith, L. E. Kun, S. Jeha, C. Cheng, S. C. Howard, V. Simmons, A. Bayles, M. L. Metzger, J. M. Boyett, W. Leung, R. Handgretinger, J. R. Downing, W. E. Evans & M. V. Relling (2009) Treating Childhood Acute Lymphoblastic Leukemia without Cranial Irradiation. *New England Journal of Medicine,* 360**,** 2730-2741.

Pui, C. H. & W. E. Evans (2006) Treatment of acute lymphoblastic leukemia. *N Engl J Med,* 354**,** 166-78.

Pui, C. H., D. Pei, J. T. Sandlund, D. Campana, R. C. Ribeiro, B. I. Razzouk, J. E. Rubnitz, S. C. Howard, N. Hijiya, S. Jeha, C. Cheng, J. R. Downing, W. E. Evans, M. V. Relling & M. Hudson

(2005) Risk of adverse events after completion of therapy for childhood acute lymphoblastic leukemia. *J Clin Oncol,* 23**,** 7936-41.

Pui, C. H., L. L. Robison & A. T. Look (2008) Acute lymphoblastic leukaemia. *Lancet,* 371**,** 1030-1043.

Qiu, X. & A. Yakovlev (2006) Some comments on instability of false discovery rate estimation. *Journal of Bioinformatics and Computational Biology,* 4**,** 1057-1068.

Rhein, P., S. Scheid, R. Ratei, C. Hagemeier, K. Seeger, R. Kirschner-Schwabe, A. Moericke, M. Schrappe, R. Spang, W. D. Ludwig & L. Karawajew (2007) Gene expression shift towards normal B cells, decreased proliferative capacity and distinct surface receptors characterize leukemic blasts persisting during induction therapy in childhood acute lymphoblastic leukemia. *Leukemia,* 21**,** 897-905.

Ross, M. E., X. Zhou, G. Song, S. A. Shurtleff, K. Girtman, W. K. Williams, H. C. Liu, R. Mahfouz, S. C. Raimondi, N. Lenny, A. Patel & J. R. Downing (2003) Classification of pediatric acute lymphoblastic leukemia by gene expression profiling. *Blood,* 102**,** 2951-9.

Schena, M., D. Shalon, R. W. Davis & P. O. Brown (1995) Quantitative Monitoring of Gene Expression Patterns with a Complementary DNA Microarray. *Science,* 270**,** 467-470.

Scherer, A. 2009. *Variation, Variability, Batches and Bias in Microarray Experiments: An Introduction*. John Wiley & Sons, Ltd.

Schultz, K. R., D. J. Pullen, H. N. Sather, J. J. Shuster, M. Devidas, M. J. Borowitz, A. J. Carroll, N. A. Heerema, J. E. Rubnitz, M. L. Loh, E. A. Raetz, N. J. Winick, S. P. Hunger, W. L. Carroll, P. S. Gaynon & B. M. Camitta (2007) Risk- and response-based classification of childhood B-precursor acute lymphoblastic leukemia: a combined analysis of prognostic markers from the Pediatric Oncology Group (POG) and Children's Cancer Group (CCG). *Blood,* 109**,** 926-935.

Slonim, D. K. (2002) From patterns to pathways: gene expression data analysis comes of age. *Nat Genet*.

Smith, M., D. Arthur, B. Camitta, A. J. Carroll, W. Crist, P. Gaynon, R. Gelber, N. Heerema, E. L. Korn, M. Link, S. Murphy, C. H. Pui, J. Pullen, G. Reamon, S. E. Sallan, H. Sather, J. Shuster, R. Simon, M. Trigg, D. Tubergen, F. Uckun & R. Ungerleider (1996) Uniform approach to risk classification and treatment assignment for children with acute lymphoblastic leukemia. *Journal of Clinical Oncology,* 14**,** 18-24.

Sorich, M. J., N. Pottier, D. Pei, W. Yang, L. Kager, G. Stocco, C. Cheng, J. C. Panetta, C. H. Pui, M. V. Relling, M. H. Cheok & W. E. Evans (2008) In Vivo Response to Methotrexate Forecasts Outcome of Acute Lymphoblastic Leukemia and Has a Distinct Gene Expression Profile. *PLoS Med,* 5**,** e83.

Staal, F. J. T., D. de Ridder, T. Szczepanski, T. Schonewille, E. C. E. van der Linden, E. R. van Wering, V. H. J. van der Velden & J. J. M. van Dongen (2010) Genome-wide expression analysis of paired diagnosis-relapse samples in ALL indicates involvement of pathways related to DNA replication, cell cycle and DNA repair, independent of immune phenotype. *Leukemia,* 24**,** 491-499.

Staal, F. J. T., M. van der Burg, L. F. A. Wessels, B. H. Barendregt, M. R. M. Baert, C. M. M. van den Burg, C. Van Huffel, A. W. Langerak, V. H. J. van der Velden, M. J. T. Reinders & J. J. M. van Dongen (2003) DNA microarrays for comparison of gene expression profiles between diagnosis and relapse in precursor-B acute lymphoblastic leukemia: choice of technique and purification influence the identification of potential diagnostic markers. *Leukemia,* 17**,** 1324-1332.

Storey, J. D. & R. Tibshirani (2003) Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences of the United States of America,* 100**,** 9440-9445.

Tissing, W. J. E., M. L. den Boer, J. P. P. Meijerink, R. X. Menezes, S. Swagemakers, P. J. van der Spek, S. E. Sallan, S. A. Armstrong & R. Pieters (2007) Genomewide identification of prednisolone-responsive genes in acute lymphoblastic leukemia cells. *Blood,* 109**,** 3929-3935.

Tusher, V. G., R. Tibshirani & G. Chu (2001) Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences of the United States of America,* 98**,** 5116-5121.

van 't Veer, L. J., H. Dai, M. J. van de Vijver, Y. D. He, A. A. M. Hart, M. Mao, H. L. Peterse, K. van der Kooy, M. J. Marton, A. T. Witteveen, G. J. Schreiber, R. M. Kerkhoven, C. Roberts, P. S. Linsley, R. Bernards & S. H. Friend (2002) Gene expression profiling predicts clinical outcome of breast cancer. *Nature,* 415**,** 530-536.

Willenbrock, H., A. S. Juncker, K. Schmiegelow, S. Knudsen & L. P. Ryder (2004) Prediction of immunophenotype, treatment response, and relapse in childhood acute lymphoblastic leukemia using DNA microarrays. *Leukemia,* 18**,** 1270-1277.

Yeoh, E. J., M. E. Ross, S. A. Shurtleff, W. K. Williams, D. Patel, R. Mahfouz, F. G. Behm, S. C. Raimondi, M. V. Relling, A. Patel, C. Cheng, D. Campana, D. Wilkins, X. Zhou, J. Li, H. Liu, C. H. Pui, W. E. Evans, C. Naeve, L. Wong & J. R. Downing (2002) Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling. *Cancer Cell,* 1**,** 133-43.

Zheng, Q. & X.-J. Wang (2008) GOEAST: a web-based software toolkit for Gene Ontology enrichment analysis. *Nucleic Acids Research,* 36**,** W358-W363.