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
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.
LIST OF TABLE

Table 2.1: Comparing cost and outcome of different treatment strategies. ........................................... 11
Table 3.1: Patient characteristics in different demographic, prognostic and genotypic groups..... 24
Table 4.1: Genetic signature genes of T-ALL. ......................................................................................... 38
Table 4.2: Genetic signature genes of TEL-AML1. ............................................................................... 39
Table 4.3: Genetic signature genes of Hyperdiploid>50. ................................................................. 40
Table 4.4: Top 20 up-regulated probe sets............................................................................................. 44
Table 4.5: Top 20 down-regulated probe sets...................................................................................... 45
Table 4.6: Top 20 GO terms for the up-regulated probe sets.............................................................. 46
Table 4.7: Top 20 GO terms for the down-regulated probe sets.......................................................... 47
Table 4.8: Significant pathways for the differentially expressed probe sets between D8 and D0. 48
Table 4.9: Significant biological functions for the differentially expressed probe sets between D8
and D0...................................................................................................................................................... 49
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............................... 76
Table 5.2: ASD between the D0 and D15 samples. Relapses are highlighted with Underline.
Extremely slow responders are highlighted in Italic............................................................................. 77
Table 5.3: ASD between the D0 and D33 samples. Relapses are highlighted with Underline.
Extremely slow responders are highlighted in Italic............................................................................. 78
Table 5.4: ESD between the D0 and D8 samples. Relapses are highlighted with Underline.
Extremely slow responders are highlighted in Italic............................................................................. 79
Table 5.5: ESD between the D0 and D15 samples. Relapses are highlighted with Underline. Extremely slow responders are highlighted in Italic.................................................................80
Table 5.6: ESD between the D0 and D33 samples. Relapses are highlighted with Underline. Extremely slow responders are highlighted in Italic.................................................................81
Table 5.7: ESR between the D0 and D8 samples. Relapses are highlighted with Underline. Extremely slow responders are highlighted in Italic.................................................................82
Table 5.8: ESR between the D0 and D15 samples. Relapses are highlighted with Underline. Extremely slow responders are highlighted in Italic.................................................................83
Table 5.9: ESR between the D0 and D33 samples. Relapses are highlighted with Underline. Extremely slow responders are highlighted in Italic.................................................................84
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..........................................................................................................................................................89
Table 6.1: Patient characteristics of our AML dataset.................................................................................................................................95
Table 6.2: ASD and ESD of GSS-AML. Relapses are highlighted in the table. .................................................................98
Table A.1: Drug responsive genes of T-ALL subtype.................................................................................................................................104
Table A.2: Drug responsive genes of TEL-AML1 subtype.................................................................107
Table A.3: Drug responsive genes of Hyperdiploid>50 subtype.................................................................109
Table A.4: Drug responsive genes of E2A-PBX1 subtype.................................................................112
Table A.5: Drug responsive genes of BCR-ABL subtype.................................................................114
Table A.6: Drug responsive genes of MLL subtype.................................................................116
Table A.7: Drug responsive genes of other subtypes.................................................................119
LIST OF FIGURE

Figure 1.1: The number of annually published GEP datasets in GEO depository at NCBI from 2001 to 2010. .................................................................................................................................................. 2

Figure 1.2: A comprehensive overview of childhood ALL diagnosis and prognosis. ................. 6

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. ........................................ 12

Figure 2.2: Affymetrix GeneChip, reproduced from Affymetrix (Santa Clara, CA, USA). .......... 14

Figure 2.3: GeneChip hybridization, reproduced from Affymetrix (Santa Clara, CA, USA). ..... 15

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. ......................................................... 26

Figure 3.2: The batch effects of our GEPs. The 4 clusters correspond to the 4 batches in Figure 3.1 by color. ............................................................................................................................................ 26

Figure 3.3: An example of quantile normalization, reproduced from Bolstad et al. 2003........... 29

Figure 3.4: The process of quantile normalization. ............................................................................. 29

Figure 3.5: The gene expression distributions after quantile normalization. The black bold curve in the middle is the reference distribution................................................................. 31

Figure 3.6: GEPs after the batch effects removing. ............................................................................. 31

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. 34

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. ................................................................. 36

Figure 4.3: The top biological network, cancer, inflammatory response, and cell-to-cell signaling and interaction.................................................................................................................................................. 51

Figure 4.4: The second top biological network, inflammatory response, cell death, and cell-to-cell signaling and interaction.................................................................................................................................................. 52

Figure 4.5: The third top biological network, cancer, respiratory disease, and cellular development........................................................................................................................................................................ 53

Figure 4.6: The fourth top biological network, cell-to-cell signaling and interaction, tissue development, and cellular movement. ................................................................................................................................. 54

Figure 4.7: The fifth top biological network, cancer, gastrointestinal disease, and cell cycle........ 55

Figure 4.8: The global GSS model and its variance distribution. (a) The global GSS model. (b) The variance contained in top PCs........................................................................................................................................................................ 57

Figure 4.9: SJCRH samples in the global GSS model........................................................................................................ 58

Figure 4.10: DCOG samples in the global GSS model........................................................................................................ 58

Figure 4.11: DCOG2 samples in the global GSS model........................................................................................................ 59

Figure 4.12: COALL samples in the global GSS model........................................................................................................ 59

Figure 4.13: MILE-Diagnose samples in the global GSS model........................................................................................................ 60

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........................................................................................................................................................................ 62
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.

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

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

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

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

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.

Figure 5.1: Genetic status shifting distance.

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

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.
<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALL</td>
<td>Acute Lymphoblastic Leukemia</td>
</tr>
<tr>
<td>AML</td>
<td>Acute Myeloid Leukemia</td>
</tr>
<tr>
<td>CCR</td>
<td>Continuous Complete Remission</td>
</tr>
<tr>
<td>DT</td>
<td>Decision Tree</td>
</tr>
<tr>
<td>FDR</td>
<td>False Discovery Rate</td>
</tr>
<tr>
<td>GEP</td>
<td>Gene Expression Profiling</td>
</tr>
<tr>
<td>GO</td>
<td>Gene Ontology</td>
</tr>
<tr>
<td>GOEAST</td>
<td>Gene Ontology Enrichment Analysis</td>
</tr>
<tr>
<td>GSS</td>
<td>Genetic Status Shifting</td>
</tr>
<tr>
<td>IPA</td>
<td>Ingenuity Pathway Analysis</td>
</tr>
<tr>
<td>MAS5.0</td>
<td>Affymetrix Microarray Suite 5.0</td>
</tr>
<tr>
<td>MRD</td>
<td>Minimal Residual Disease</td>
</tr>
<tr>
<td>NB</td>
<td>Naïve Bayes</td>
</tr>
<tr>
<td>NBM</td>
<td>Normal Bone Marrow</td>
</tr>
<tr>
<td>Abbreviation</td>
<td>Full Form</td>
</tr>
<tr>
<td>--------------</td>
<td>-----------</td>
</tr>
<tr>
<td>PC</td>
<td>Principal Component</td>
</tr>
<tr>
<td>PCA</td>
<td>Principal Component Analysis</td>
</tr>
<tr>
<td>PCR</td>
<td>Polymerase Chain Reaction</td>
</tr>
<tr>
<td>RMA</td>
<td>Robust Multiple-Array Average</td>
</tr>
<tr>
<td>ROC</td>
<td>Receiver operating characteristic</td>
</tr>
<tr>
<td>SAM</td>
<td>Significance Analysis of Microarrays</td>
</tr>
<tr>
<td>SVM</td>
<td>Support Vector Machine</td>
</tr>
<tr>
<td>TP</td>
<td>Time Point</td>
</tr>
</tbody>
</table>
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).
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.
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.

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

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.
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.
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, $\chi^2$ 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 $\chi^2$ 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%. 
CHAPTER 2 RELATED WORK

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 (≥1.16). Molecular screening for TEL-AML1, BCR-ABL,
Table 3.1: Patient characteristics in different demographic, prognostic and genotypic groups.

<table>
<thead>
<tr>
<th>Category</th>
<th>Frequency</th>
<th>Percentage, %</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>RACE</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chinese</td>
<td>42</td>
<td>43.7</td>
</tr>
<tr>
<td>Malay</td>
<td>38</td>
<td>39.6</td>
</tr>
<tr>
<td>Indian &amp; Other</td>
<td>16</td>
<td>16.7</td>
</tr>
<tr>
<td><strong>SEX</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>46</td>
<td>47.9</td>
</tr>
<tr>
<td>Female</td>
<td>50</td>
<td>52.1</td>
</tr>
<tr>
<td><strong>AGE</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1-9</td>
<td>70</td>
<td>72.9</td>
</tr>
<tr>
<td>&lt;1 or &gt;9</td>
<td>26</td>
<td>27.1</td>
</tr>
<tr>
<td><strong>LINEAGE</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B-lineage</td>
<td>84</td>
<td>87.5</td>
</tr>
<tr>
<td>T-lineage</td>
<td>12</td>
<td>12.5</td>
</tr>
<tr>
<td><strong>WHITE BLOOD CELL</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt;50,000</td>
<td>67</td>
<td>69.8</td>
</tr>
<tr>
<td>≥50,000</td>
<td>29</td>
<td>30.2</td>
</tr>
<tr>
<td><strong>SUBTYPE</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TEL-AML1</td>
<td>26</td>
<td>26.8</td>
</tr>
<tr>
<td>BCR-ABL</td>
<td>5</td>
<td>5.2</td>
</tr>
<tr>
<td>E2A-PBX1</td>
<td>4</td>
<td>4.2</td>
</tr>
<tr>
<td>Hyperdiploid&gt;50</td>
<td>12</td>
<td>12.5</td>
</tr>
<tr>
<td>Others</td>
<td>50</td>
<td>51.5</td>
</tr>
<tr>
<td><strong>Day-8 RESPONSE</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Good</td>
<td>79</td>
<td>82.3</td>
</tr>
<tr>
<td>Poor</td>
<td>16</td>
<td>16.7</td>
</tr>
<tr>
<td>Missing</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td><strong>Day-33 MINIMAL RESIDUAL DISEASE</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt;0.01%</td>
<td>49</td>
<td>51.0</td>
</tr>
<tr>
<td>0.01-0.1%</td>
<td>24</td>
<td>25.0</td>
</tr>
<tr>
<td>0.1-1%</td>
<td>16</td>
<td>16.7</td>
</tr>
<tr>
<td>≥1%</td>
<td>5</td>
<td>5.2</td>
</tr>
<tr>
<td>Missing</td>
<td>2</td>
<td>2.1</td>
</tr>
<tr>
<td><strong>OUTCOME</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Remission</td>
<td>81</td>
<td>84.4</td>
</tr>
<tr>
<td>Relapse</td>
<td>13</td>
<td>13.4</td>
</tr>
<tr>
<td>Death</td>
<td>2</td>
<td>2.1</td>
</tr>
</tbody>
</table>
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 >50×10^9 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, \( p = 1.5 \times 10^{-11} \); D8 to D15, \( p = 1.1 \times 10^{-19} \); D15 to D33, \( p = 2.2 \times 10^{-19} \)).

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.

<table>
<thead>
<tr>
<th>Probe Set</th>
<th>Gene Symbol</th>
<th>Gene Title</th>
<th>Fold Change</th>
<th>p Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>213060_s_at</td>
<td>CHI3L2</td>
<td>chitinase 3-like 2</td>
<td>49.56</td>
<td>1.06E-22</td>
</tr>
<tr>
<td>210116_at</td>
<td>SH2D1A</td>
<td>SH2 domain protein 1A</td>
<td>53.47</td>
<td>1.16E-17</td>
</tr>
<tr>
<td>205674_x_at</td>
<td>FXYD2</td>
<td>FXYD domain containing ion transport regulator 2</td>
<td>11.07</td>
<td>4.98E-16</td>
</tr>
<tr>
<td>216705_s_at</td>
<td>ADA</td>
<td>adenosine deaminase</td>
<td>3.59</td>
<td>6.88E-14</td>
</tr>
<tr>
<td>202760_s_at</td>
<td>PALM2-AKAP2</td>
<td>PALM2-AKAP2 readthrough transcript</td>
<td>5.14</td>
<td>1.37E-13</td>
</tr>
<tr>
<td>217147_s_at</td>
<td>TRAT1</td>
<td>T cell receptor associated transmembrane adaptor 1</td>
<td>22.34</td>
<td>4.90E-13</td>
</tr>
<tr>
<td>202747_s_at</td>
<td>ITM2A</td>
<td>integral membrane protein 2A</td>
<td>18.60</td>
<td>9.24E-12</td>
</tr>
<tr>
<td>203238_s_at</td>
<td>NOTCH3</td>
<td>Notch homolog 3 (Drosophila)</td>
<td>7.98</td>
<td>1.13E-11</td>
</tr>
<tr>
<td>202746_at</td>
<td>ITM2A</td>
<td>integral membrane protein 2A</td>
<td>15.67</td>
<td>1.23E-10</td>
</tr>
<tr>
<td>211071_s_at</td>
<td>MLLT11</td>
<td>myeloid/lymphoid or mixed-lineage leukemia (Drosophila); translocated to, 11</td>
<td>2.94</td>
<td>1.00E-09</td>
</tr>
<tr>
<td>205484_at</td>
<td>SIT1</td>
<td>signaling threshold regulating transmembrane adaptor 1</td>
<td>11.25</td>
<td>1.42E-09</td>
</tr>
<tr>
<td>215030_at</td>
<td>GRSF1</td>
<td>G-rich RNA sequence binding factor 1</td>
<td>2.42</td>
<td>1.44E-09</td>
</tr>
<tr>
<td>209604_s_at</td>
<td>GATA3</td>
<td>GATA binding protein 3</td>
<td>12.04</td>
<td>2.10E-08</td>
</tr>
<tr>
<td>206533_at</td>
<td>CHRNA5</td>
<td>cholinergic receptor, nicotinic, alpha 5</td>
<td>3.27</td>
<td>3.04E-08</td>
</tr>
<tr>
<td>204529_s_at</td>
<td>TOX</td>
<td>thymocyte selection-associated high mobility group box</td>
<td>4.91</td>
<td>3.15E-08</td>
</tr>
<tr>
<td>219408_at</td>
<td>PRMT7</td>
<td>protein arginine methyltransferase 7</td>
<td>2.67</td>
<td>1.44E-07</td>
</tr>
<tr>
<td>204639_at</td>
<td>ADA</td>
<td>adenosine deaminase</td>
<td>2.63</td>
<td>4.94E-07</td>
</tr>
<tr>
<td>204530_s_at</td>
<td>TOX</td>
<td>thymocyte selection-associated high mobility group box</td>
<td>3.64</td>
<td>1.41E-06</td>
</tr>
<tr>
<td>206460_at</td>
<td>AJAP1</td>
<td>adherens junctions associated protein 1</td>
<td>8.07</td>
<td>3.15E-06</td>
</tr>
<tr>
<td>219660_s_at</td>
<td>ATP8A2</td>
<td>ATPase, aminophospholipid transporter-like, class I, type 8A, member 2</td>
<td>3.92</td>
<td>9.56E-06</td>
</tr>
</tbody>
</table>
### Table 4.2: Genetic signature genes of TEL-AML1.

<table>
<thead>
<tr>
<th>Probe Set</th>
<th>Gene Symbol</th>
<th>Gene Title</th>
<th>Fold Change</th>
<th>p Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>213317_at</td>
<td>CLIC5</td>
<td>chloride intracellular channel 5</td>
<td>176.36</td>
<td>6.53E-38</td>
</tr>
<tr>
<td>213558_at</td>
<td>PCLO</td>
<td>piccolo (presynaptic cytomatrix protein)</td>
<td>74.41</td>
<td>2.07E-23</td>
</tr>
<tr>
<td>218804_at</td>
<td>ANO1</td>
<td>anoctamin 1, calcium activated chloride channel</td>
<td>22.53</td>
<td>9.83E-20</td>
</tr>
<tr>
<td>205952_at</td>
<td>KCNK3</td>
<td>potassium channel, subfamily K, member 3</td>
<td>28.77</td>
<td>8.64E-18</td>
</tr>
<tr>
<td>203611_at</td>
<td>TERF2</td>
<td>telomeric repeat binding factor 2</td>
<td>7.41</td>
<td>1.92E-16</td>
</tr>
<tr>
<td>204914_s_at</td>
<td>SOX11</td>
<td>SRY (sex determining region Y)-box 11</td>
<td>65.05</td>
<td>2.88E-16</td>
</tr>
<tr>
<td>203038_at</td>
<td>PTPRK</td>
<td>protein tyrosine phosphatase, receptor type, K</td>
<td>48.06</td>
<td>5.37E-16</td>
</tr>
<tr>
<td>204915_s_at</td>
<td>SOX11</td>
<td>SRY (sex determining region Y)-box 11</td>
<td>30.62</td>
<td>1.59E-15</td>
</tr>
<tr>
<td>201911_s_at</td>
<td>FARP1</td>
<td>FERM, RhoGEF (ARHGEF) and pleckstrin domain protein 1 (chondrocyte-derived)</td>
<td>7.75</td>
<td>4.02E-15</td>
</tr>
<tr>
<td>204913_s_at</td>
<td>SOX11</td>
<td>SRY (sex determining region Y)-box 11</td>
<td>50.85</td>
<td>1.23E-14</td>
</tr>
<tr>
<td>206591_at</td>
<td>RAG1</td>
<td>recombination activating gene 1</td>
<td>8.34</td>
<td>5.39E-11</td>
</tr>
<tr>
<td>218820_at</td>
<td>C14orf132</td>
<td>chromosome 14 open reading frame 132</td>
<td>7.86</td>
<td>1.68E-08</td>
</tr>
<tr>
<td>209101_at</td>
<td>CTGF</td>
<td>connective tissue growth factor</td>
<td>14.76</td>
<td>8.10E-08</td>
</tr>
<tr>
<td>205267_at</td>
<td>POU2AF1</td>
<td>POU class 2 associating factor 1</td>
<td>2.96</td>
<td>9.92E-08</td>
</tr>
<tr>
<td>210432_s_at</td>
<td>SCN3A</td>
<td>sodium channel, voltage-gated, type III, alpha subunit</td>
<td>14.13</td>
<td>2.24E-07</td>
</tr>
<tr>
<td>32625_at</td>
<td>NPR1</td>
<td>natriuretic peptide receptor A/guanylate cyclase A (atriotriuretic peptide receptor A)</td>
<td>4.70</td>
<td>8.45E-07</td>
</tr>
<tr>
<td>211126_s_at</td>
<td>CSRP2</td>
<td>cysteine and glycine-rich protein 2</td>
<td>3.52</td>
<td>3.24E-06</td>
</tr>
<tr>
<td>214761_at</td>
<td>ZNF423</td>
<td>zinc finger protein 423</td>
<td>4.48</td>
<td>5.02E-06</td>
</tr>
<tr>
<td>203435_s_at</td>
<td>MME</td>
<td>membrane metallo-endopeptidase</td>
<td>5.20</td>
<td>2.31E-05</td>
</tr>
<tr>
<td>219686_at</td>
<td>STK32B</td>
<td>serine/threonine kinase 32B</td>
<td>4.60</td>
<td>2.45E-05</td>
</tr>
</tbody>
</table>
Table 4.3: Genetic signature genes of Hyperdiploid>50.

<table>
<thead>
<tr>
<th>Probe Set</th>
<th>Gene Symbol</th>
<th>Gene Title</th>
<th>Fold Change</th>
<th>p Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>201508_at</td>
<td>IGFBP4</td>
<td>insulin-like growth factor binding protein 4</td>
<td>5.21</td>
<td>2.49E-05</td>
</tr>
<tr>
<td>203063_at</td>
<td>PPM1F</td>
<td>protein phosphatase 1F (PP2C domain containing)</td>
<td>2.45</td>
<td>7.46E-05</td>
</tr>
<tr>
<td>206674_at</td>
<td>FLT3</td>
<td>fms-related tyrosine kinase 3</td>
<td>4.43</td>
<td>8.91E-05</td>
</tr>
<tr>
<td>214745_at</td>
<td>PLCH1</td>
<td>phospholipase C, eta 1</td>
<td>3.46</td>
<td>2.32E-04</td>
</tr>
<tr>
<td>218694_at</td>
<td>ARMCX1</td>
<td>armadillo repeat containing, X-linked 1</td>
<td>3.68</td>
<td>2.50E-04</td>
</tr>
<tr>
<td>208370_s_at</td>
<td>RCAN1</td>
<td>regulator of calcineurin 1</td>
<td>2.79</td>
<td>5.82E-04</td>
</tr>
<tr>
<td>201005_at</td>
<td>CD9</td>
<td>CD9 molecule</td>
<td>5.87</td>
<td>6.64E-04</td>
</tr>
<tr>
<td>202598_at</td>
<td>S100A13</td>
<td>S100 calcium binding protein A13</td>
<td>2.17</td>
<td>8.41E-04</td>
</tr>
<tr>
<td>207267_s_at</td>
<td>DSCR6</td>
<td>Down syndrome critical region gene 6</td>
<td>2.60</td>
<td>8.47E-04</td>
</tr>
<tr>
<td>41660_at</td>
<td>CELSR1</td>
<td>cadherin, EGF LAG seven-pass G-type receptor 1 (flamingo homolog, Drosophila)</td>
<td>2.91</td>
<td>1.05E-03</td>
</tr>
<tr>
<td>215263_at</td>
<td>ZXDA /// ZXDB</td>
<td>zinc finger, X-linked, duplicated A /// zinc finger, X-linked, duplicated B</td>
<td>1.72</td>
<td>1.05E-03</td>
</tr>
<tr>
<td>204462_s_at</td>
<td>SLC16A2</td>
<td>solute carrier family 16, member 2 (monocarboxylic acid transporter 8)</td>
<td>4.20</td>
<td>1.47E-03</td>
</tr>
<tr>
<td>206852_at</td>
<td>EPHA7</td>
<td>EPH receptor A7</td>
<td>4.56</td>
<td>1.51E-03</td>
</tr>
<tr>
<td>204454_at</td>
<td>LDOC1</td>
<td>leucine zipper, down-regulated in cancer 1</td>
<td>4.05</td>
<td>2.23E-03</td>
</tr>
<tr>
<td>214961_at</td>
<td>KIAA0774</td>
<td>KIAA0774</td>
<td>5.24</td>
<td>2.43E-03</td>
</tr>
<tr>
<td>209183_s_at</td>
<td>C10orf10</td>
<td>chromosome 10 open reading frame 10</td>
<td>3.60</td>
<td>2.71E-03</td>
</tr>
<tr>
<td>214156_at</td>
<td>MYRIP</td>
<td>myosin VIIA and Rab interacting protein</td>
<td>5.31</td>
<td>3.20E-03</td>
</tr>
<tr>
<td>211626_x_at</td>
<td>ERG</td>
<td>v-ets erythroblastosis virus E26 oncogene homolog (avian)</td>
<td>2.13</td>
<td>3.38E-03</td>
</tr>
<tr>
<td>213316_at</td>
<td>KIAA1462</td>
<td>KIAA1462</td>
<td>4.22</td>
<td>3.44E-03</td>
</tr>
<tr>
<td>212385_at</td>
<td>TCF4</td>
<td>transcription factor 4</td>
<td>2.81</td>
<td>3.60E-03</td>
</tr>
</tbody>
</table>
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 ($p = 1.99 \times 10^{-90}$), leukocyte activation ($p = 8.39 \times 10^{-21}$), hemoglobin complex ($p = 7.96 \times 10^{-18}$), and blood circulation ($p = 3.40 \times 10^{-15}$). The down-regulated terms include two categories. The first category involves the cell development and DNA synthesis, such as, cellular developmental process ($p = 2.40 \times 10^{-9}$), cell differentiation ($p = 4.53 \times 10^{-9}$), DNA packaging ($p = 7.69 \times 10^{-8}$), DNA binding ($p = 1.23 \times 10^{-7}$), and chromatin assembly ($p = 2.36 \times 10^{-7}$). The second category involves the negative regulation of apoptosis, such as, negative regulation of thymocyte apoptosis ($p = 2.55 \times 10^{-7}$), negative regulation of T cell
apoptosis ($p = 4.51 \times 10^{-7}$), negative regulation of lymphocyte apoptosis ($p = 7.19 \times 10^{-6}$), and negative regulation of mature B cell apoptosis ($p = 1.53 \times 10^{-5}$).

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 ($p = 4.17 \times 10^{-8}$), primary immunodeficiency signaling ($p = 2.04 \times 10^{-7}$), B cell development ($p = 2.34 \times 10^{-7}$), inflammatory response ($p = 3.51 \times 10^{-37}$), hematological system development and function ($p = 2.73 \times 10^{-21}$), and hematopoiesis ($p = 5.93 \times 10^{-12}$), 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 ($p = 1.82 \times 10^{-26}$), cellular growth and proliferation ($p = 1.66 \times 10^{-18}$), cellular development ($p = 7.38 \times 10^{-15}$), cellular assembly and organization ($p = 7.98 \times 10^{-6}$), DNA replication, recombination, and repair ($p = 1.41 \times 10^{-5}$), and gene expression ($p = 4.91 \times 10^{-4}$), are related to the cell development and DNA synthesis. Results, such as, cytotoxic T lymphocyte-mediated apoptosis of target cells ($p = 9.77 \times 10^{-5}$), and cell death ($p = 2.36 \times 10^{-13}$) 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.

<table>
<thead>
<tr>
<th>Probe Set ID</th>
<th>Gene Symbol</th>
<th>Gene Title</th>
<th>Fold Change</th>
</tr>
</thead>
<tbody>
<tr>
<td>205950_s_at</td>
<td>CA1</td>
<td>carbonic anhydrase I</td>
<td>7.395791184</td>
</tr>
<tr>
<td>205403_at</td>
<td>IL1R2</td>
<td>interleukin 1 receptor, type II</td>
<td>6.852324058</td>
</tr>
<tr>
<td>205997_at</td>
<td>ADAM28</td>
<td>ADAM metallopeptidase domain 28</td>
<td>6.708182719</td>
</tr>
<tr>
<td>214146_s_at</td>
<td>PPBP</td>
<td>pro-platelet basic protein (chemokine (C-X-C motif) ligand 7)</td>
<td>6.658931873</td>
</tr>
<tr>
<td>211372_s_at</td>
<td>IL1R2</td>
<td>interleukin 1 receptor, type II</td>
<td>6.466121916</td>
</tr>
<tr>
<td>203645_s_at</td>
<td>CD163</td>
<td>CD163 molecule</td>
<td>6.117093533</td>
</tr>
<tr>
<td>215049_x_at</td>
<td>CD163</td>
<td>CD163 molecule</td>
<td>6.087834242</td>
</tr>
<tr>
<td>205837_s_at</td>
<td>GYPA /// GYPB</td>
<td>glycophorin A (MNS blood group) /// glycophorin B (MNS blood group)</td>
<td>5.822822642</td>
</tr>
<tr>
<td>201110_s_at</td>
<td>THBS1</td>
<td>thrombospondin 1</td>
<td>5.745576019</td>
</tr>
<tr>
<td>211821_x_at</td>
<td>GYPA</td>
<td>glycophorin A (MNS blood group)</td>
<td>5.209992602</td>
</tr>
<tr>
<td>212768_s_at</td>
<td>OLFM4</td>
<td>olfactomedin 4</td>
<td>5.182515342</td>
</tr>
<tr>
<td>209555_s_at</td>
<td>CD36</td>
<td>CD36 molecule (thrombospondin receptor)</td>
<td>5.064436979</td>
</tr>
<tr>
<td>217388_s_at</td>
<td>KYNU</td>
<td>kynureninase (L-kynurenine hydrolase)</td>
<td>5.046071877</td>
</tr>
<tr>
<td>211560_s_at</td>
<td>ALAS2</td>
<td>aminolevulinate, delta-, synthase 2</td>
<td>5.008170532</td>
</tr>
<tr>
<td>206488_s_at</td>
<td>CD36</td>
<td>CD36 molecule (thrombospondin receptor)</td>
<td>4.920133289</td>
</tr>
<tr>
<td>215646_s_at</td>
<td>VCAN</td>
<td>Versican</td>
<td>4.917441199</td>
</tr>
<tr>
<td>210746_s_at</td>
<td>EPB42</td>
<td>erythrocyte membrane protein band 4.2</td>
<td>4.903890326</td>
</tr>
<tr>
<td>221731_x_at</td>
<td>VCAN</td>
<td>Versican</td>
<td>4.832463599</td>
</tr>
<tr>
<td>217418_x_at</td>
<td>MS4A1</td>
<td>membrane-spanning 4-domains, subfamily A, member 1</td>
<td>4.811721663</td>
</tr>
<tr>
<td>206390_x_at</td>
<td>PF4</td>
<td>platelet factor 4</td>
<td>4.801380171</td>
</tr>
</tbody>
</table>
Table 4.5: Top 20 down-regulated probe sets.

<table>
<thead>
<tr>
<th>Probe Set ID</th>
<th>Gene Symbol</th>
<th>Gene Title</th>
<th>Fold Change</th>
</tr>
</thead>
<tbody>
<tr>
<td>210487_at</td>
<td>DNTT</td>
<td>deoxynucleotidyltransferase, terminal</td>
<td>0.222299534</td>
</tr>
<tr>
<td>209035_at</td>
<td>MDK</td>
<td>midkine (neurite growth-promoting factor 2)</td>
<td>0.224017113</td>
</tr>
<tr>
<td>211341_at</td>
<td>POU4F1</td>
<td>POU class 4 homeobox 1</td>
<td>0.235935714</td>
</tr>
<tr>
<td>203434_s_at</td>
<td>MME</td>
<td>membrane metallo-endopeptidase</td>
<td>0.256940866</td>
</tr>
<tr>
<td>206660_at</td>
<td>IGLL1</td>
<td>immunoglobulin lambda-like polypeptide 1</td>
<td>0.258262938</td>
</tr>
<tr>
<td>215117_at</td>
<td>RAG2</td>
<td>recombination activating gene 2</td>
<td>0.259848868</td>
</tr>
<tr>
<td>207030_s_at</td>
<td>CSRP2</td>
<td>cysteine and glycine-rich protein 2</td>
<td>0.283315663</td>
</tr>
<tr>
<td>206067_s_at</td>
<td>WT1</td>
<td>Wilms tumor 1</td>
<td>0.304693921</td>
</tr>
<tr>
<td>214243_s_at</td>
<td>SERHL ///</td>
<td>serine hydrolase-like /// serine hydrolase-like 2</td>
<td>0.315189369</td>
</tr>
<tr>
<td>219740_at</td>
<td>VASH2</td>
<td>vasohibin 2</td>
<td>0.316965896</td>
</tr>
<tr>
<td>203435_s_at</td>
<td>MME</td>
<td>membrane metallo-endopeptidase</td>
<td>0.321395315</td>
</tr>
<tr>
<td>205755_at</td>
<td>ITIH3</td>
<td>inter-alpha (globulin) inhibitor H3</td>
<td>0.327868482</td>
</tr>
<tr>
<td>206591_at</td>
<td>RAG1</td>
<td>recombination activating gene 1</td>
<td>0.337911256</td>
</tr>
<tr>
<td>204165_at</td>
<td>WASF1</td>
<td>WAS protein family, member 1</td>
<td>0.339940607</td>
</tr>
<tr>
<td>219218_at</td>
<td>BAHCC1</td>
<td>BAH domain and coiled-coil containing 1</td>
<td>0.340474114</td>
</tr>
<tr>
<td>205795_at</td>
<td>NRXN3</td>
<td>neurexin 3</td>
<td>0.360422562</td>
</tr>
<tr>
<td>208950_s_at</td>
<td>ALDH7A1</td>
<td>aldehyde dehydrogenase 7 family, member A1</td>
<td>0.36198217</td>
</tr>
<tr>
<td>209983_s_at</td>
<td>NRXN2</td>
<td>neurexin 2</td>
<td>0.36527127</td>
</tr>
<tr>
<td>207426_s_at</td>
<td>TNFSF4</td>
<td>tumor necrosis factor (ligand) superfamily, member 4</td>
<td>0.365433513</td>
</tr>
<tr>
<td>213668_s_at</td>
<td>SOX4</td>
<td>SRY (sex determining region Y)-box 4</td>
<td>0.36606881</td>
</tr>
</tbody>
</table>
Table 4.6: Top 20 GO terms for the up-regulated probe sets.

<table>
<thead>
<tr>
<th>GO ID</th>
<th>Ontology</th>
<th>Term</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>GO:0002376</td>
<td>biological_process</td>
<td>immune system process</td>
<td>1.99E-90</td>
</tr>
<tr>
<td>GO:0006955</td>
<td>biological_process</td>
<td>immune response</td>
<td>4.58E-77</td>
</tr>
<tr>
<td>GO:0050896</td>
<td>biological_process</td>
<td>response to stimulus</td>
<td>3.70E-75</td>
</tr>
<tr>
<td>GO:0006952</td>
<td>biological_process</td>
<td>defense response</td>
<td>4.14E-67</td>
</tr>
<tr>
<td>GO:0005886</td>
<td>cellular_component</td>
<td>plasma membrane</td>
<td>1.23E-63</td>
</tr>
<tr>
<td>GO:0005488</td>
<td>molecular_function</td>
<td>Binding</td>
<td>9.26E-55</td>
</tr>
<tr>
<td>GO:0005576</td>
<td>cellular_component</td>
<td>extracellular region</td>
<td>1.05E-52</td>
</tr>
<tr>
<td>GO:0006950</td>
<td>biological_process</td>
<td>response to stress</td>
<td>4.48E-51</td>
</tr>
<tr>
<td>GO:0016020</td>
<td>cellular_component</td>
<td>Membrane</td>
<td>2.34E-45</td>
</tr>
<tr>
<td>GO:0044459</td>
<td>cellular_component</td>
<td>plasma membrane part</td>
<td>2.18E-42</td>
</tr>
<tr>
<td>GO:0005623</td>
<td>cellular_component</td>
<td>Cell</td>
<td>4.52E-41</td>
</tr>
<tr>
<td>GO:0044464</td>
<td>cellular_component</td>
<td>cell part</td>
<td>4.52E-41</td>
</tr>
<tr>
<td>GO:0004871</td>
<td>molecular_function</td>
<td>signal transducer activity</td>
<td>4.59E-40</td>
</tr>
<tr>
<td>GO:0060089</td>
<td>molecular_function</td>
<td>molecular transducer activity</td>
<td>4.59E-40</td>
</tr>
<tr>
<td>GO:0004872</td>
<td>molecular_function</td>
<td>receptor activity</td>
<td>8.74E-40</td>
</tr>
<tr>
<td>GO:0009611</td>
<td>biological_process</td>
<td>response to wounding</td>
<td>9.33E-37</td>
</tr>
<tr>
<td>GO:0005887</td>
<td>cellular_component</td>
<td>integral to plasma membrane</td>
<td>3.06E-35</td>
</tr>
<tr>
<td>GO:0005515</td>
<td>molecular_function</td>
<td>protein binding</td>
<td>4.29E-35</td>
</tr>
<tr>
<td>GO:0031226</td>
<td>cellular_component</td>
<td>intrinsic to plasma membrane</td>
<td>4.37E-35</td>
</tr>
<tr>
<td>GO:0002682</td>
<td>biological_process</td>
<td>regulation of immune system process</td>
<td>3.76E-34</td>
</tr>
</tbody>
</table>
Table 4.7: Top 20 GO terms for the down-regulated probe sets.

<table>
<thead>
<tr>
<th>GO ID</th>
<th>Ontology</th>
<th>Term</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>GO:0032502</td>
<td>biological_process</td>
<td>developmental process</td>
<td>2.89E-16</td>
</tr>
<tr>
<td>GO:0007275</td>
<td>biological_process</td>
<td>multicellular organismal development</td>
<td>1.03E-13</td>
</tr>
<tr>
<td>GO:0048731</td>
<td>biological_process</td>
<td>system development</td>
<td>8.56E-13</td>
</tr>
<tr>
<td>GO:0032501</td>
<td>biological_process</td>
<td>multicellular organismal process</td>
<td>1.67E-12</td>
</tr>
<tr>
<td>GO:0048856</td>
<td>biological_process</td>
<td>anatomical structure development</td>
<td>2.94E-12</td>
</tr>
<tr>
<td>GO:0005623</td>
<td>cellular_component</td>
<td>Cell</td>
<td>9.80E-11</td>
</tr>
<tr>
<td>GO:0044464</td>
<td>cellular_component</td>
<td>cell part</td>
<td>9.80E-11</td>
</tr>
<tr>
<td>GO:0048869</td>
<td>biological_process</td>
<td>cellular developmental process</td>
<td>2.40E-09</td>
</tr>
<tr>
<td>GO:0030154</td>
<td>biological_process</td>
<td>cell differentiation</td>
<td>4.53E-09</td>
</tr>
<tr>
<td>GO:0009987</td>
<td>biological_process</td>
<td>cellular process</td>
<td>5.46E-09</td>
</tr>
<tr>
<td>GO:0006323</td>
<td>biological_process</td>
<td>DNA packaging</td>
<td>7.69E-09</td>
</tr>
<tr>
<td>GO:0005488</td>
<td>molecular_function</td>
<td>binding</td>
<td>2.29E-08</td>
</tr>
<tr>
<td>GO:0071103</td>
<td>biological_process</td>
<td>DNA conformation change</td>
<td>3.17E-08</td>
</tr>
<tr>
<td>GO:0045112</td>
<td>biological_process</td>
<td>integrin biosynthetic process</td>
<td>4.29E-08</td>
</tr>
<tr>
<td>GO:0065007</td>
<td>biological_process</td>
<td>biological regulation</td>
<td>1.17E-07</td>
</tr>
<tr>
<td>GO:0050821</td>
<td>biological_process</td>
<td>protein stabilization</td>
<td>1.19E-07</td>
</tr>
<tr>
<td>GO:0003677</td>
<td>molecular_function</td>
<td>DNA binding</td>
<td>1.23E-07</td>
</tr>
<tr>
<td>GO:0016043</td>
<td>biological_process</td>
<td>cellular component organization</td>
<td>1.38E-07</td>
</tr>
<tr>
<td>GO:0031497</td>
<td>biological_process</td>
<td>chromatin assembly</td>
<td>2.36E-07</td>
</tr>
<tr>
<td>GO:0070244</td>
<td>biological_process</td>
<td>negative regulation of thymocyte apoptosis</td>
<td>2.55E-07</td>
</tr>
</tbody>
</table>
Table 4.8: Significant pathways for the differentially expressed probe sets between D8 and D0.

<table>
<thead>
<tr>
<th>Ingenuity Canonical Pathways</th>
<th>p value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Communication between Innate and Adaptive Immune Cells</td>
<td>4.17E-08</td>
</tr>
<tr>
<td>Primary Immunodeficiency Signaling</td>
<td>2.04E-07</td>
</tr>
<tr>
<td>B Cell Development</td>
<td>2.34E-07</td>
</tr>
<tr>
<td>Dendritic Cell Maturation</td>
<td>1.38E-06</td>
</tr>
<tr>
<td>Atherosclerosis Signaling</td>
<td>1.95E-06</td>
</tr>
<tr>
<td>IL-10 Signaling</td>
<td>1.17E-05</td>
</tr>
<tr>
<td>Hepatic Fibrosis / Hepatic Stellate Cell Activation</td>
<td>1.66E-05</td>
</tr>
<tr>
<td>Crosstalk between Dendritic Cells and Natural Killer Cells</td>
<td>1.95E-05</td>
</tr>
<tr>
<td>Systemic Lupus Erythematosus Signaling</td>
<td>4.47E-05</td>
</tr>
<tr>
<td>Graft-versus-Host Disease Signaling</td>
<td>6.61E-05</td>
</tr>
<tr>
<td>Cytotoxic T Lymphocyte-mediated Apoptosis of Target Cells</td>
<td>9.77E-05</td>
</tr>
<tr>
<td>TREM1 Signaling</td>
<td>0.000158</td>
</tr>
<tr>
<td>LXR/RXR Activation</td>
<td>0.000178</td>
</tr>
<tr>
<td>Autoimmune Thyroid Disease Signaling</td>
<td>0.000219</td>
</tr>
<tr>
<td>Allograft Rejection Signaling</td>
<td>0.000316</td>
</tr>
<tr>
<td>Role of Macrophages, Fibroblasts and Endothelial Cells in Rheumatoid Arthritis</td>
<td>0.000372</td>
</tr>
<tr>
<td>Altered T Cell and B Cell Signaling in Rheumatoid Arthritis</td>
<td>0.000457</td>
</tr>
<tr>
<td>T Helper Cell Differentiation</td>
<td>0.000457</td>
</tr>
<tr>
<td>Complement System</td>
<td>0.001047</td>
</tr>
<tr>
<td>Ascorbate and Aldarate Metabolism</td>
<td>0.001122</td>
</tr>
<tr>
<td>IL-6 Signaling</td>
<td>0.001479</td>
</tr>
<tr>
<td>Glycolysis/Gluconeogenesis</td>
<td>0.001479</td>
</tr>
<tr>
<td>Lipid Antigen Presentation by CD1</td>
<td>0.002455</td>
</tr>
<tr>
<td>Airway Pathology in Chronic Obstructive Pulmonary Disease</td>
<td>0.002512</td>
</tr>
<tr>
<td>Histidine Metabolism</td>
<td>0.002754</td>
</tr>
<tr>
<td>Granzyme A Signaling</td>
<td>0.00309</td>
</tr>
<tr>
<td>Type I Diabetes Mellitus Signaling</td>
<td>0.004169</td>
</tr>
<tr>
<td>Caveolar-mediated Endocytosis Signaling</td>
<td>0.004467</td>
</tr>
<tr>
<td>OX40 Signaling Pathway</td>
<td>0.00631</td>
</tr>
<tr>
<td>LPS/IL-1 Mediated Inhibition of RXR Function</td>
<td>0.008128</td>
</tr>
<tr>
<td>Nur77 Signaling in T Lymphocytes</td>
<td>0.008511</td>
</tr>
<tr>
<td>IL-8 Signaling</td>
<td>0.00912</td>
</tr>
<tr>
<td>CCR5 Signaling in Macrophages</td>
<td>0.009772</td>
</tr>
</tbody>
</table>
Table 4.9: Significant biological functions for the differentially expressed probe sets between D8 and D0.

<table>
<thead>
<tr>
<th>Category</th>
<th>p value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inflammatory Response</td>
<td>3.51E-37-1.4E-03</td>
</tr>
<tr>
<td>Infectious Disease</td>
<td>1.12E-31-2.19E-04</td>
</tr>
<tr>
<td>Respiratory Disease</td>
<td>1.12E-31-1.4E-03</td>
</tr>
<tr>
<td>Cell-To-Cell Signaling and Interaction</td>
<td>1.82E-26-1.4E-03</td>
</tr>
<tr>
<td>Hematological System Development and Function</td>
<td>2.73E-21-1.4E-03</td>
</tr>
<tr>
<td>Immune Cell Trafficking</td>
<td>7.56E-21-1.4E-03</td>
</tr>
<tr>
<td>Tissue Development</td>
<td>5.56E-19-1.4E-03</td>
</tr>
<tr>
<td>Cellular Growth and Proliferation</td>
<td>1.66E-18-1.13E-03</td>
</tr>
<tr>
<td>Cancer</td>
<td>3.54E-16-1.4E-03</td>
</tr>
<tr>
<td>Inflammatory Disease</td>
<td>6.95E-16-1.4E-03</td>
</tr>
<tr>
<td>Connective Tissue Disorders</td>
<td>9.75E-16-3.05E-04</td>
</tr>
<tr>
<td>Immunological Disease</td>
<td>9.75E-16-1.4E-03</td>
</tr>
<tr>
<td>Skeletal and Muscular Disorders</td>
<td>9.75E-16-3.05E-04</td>
</tr>
<tr>
<td>Cellular Development</td>
<td>7.38E-15-1.4E-03</td>
</tr>
<tr>
<td>Cellular Movement</td>
<td>5.77E-14-1.4E-03</td>
</tr>
<tr>
<td>Cell Death</td>
<td>2.36E-13-1.4E-03</td>
</tr>
<tr>
<td>Cell Signaling</td>
<td>4.21E-12-1.06E-03</td>
</tr>
<tr>
<td>Molecular Transport</td>
<td>4.21E-12-7.43E-04</td>
</tr>
<tr>
<td>Vitamin and Mineral Metabolism</td>
<td>4.21E-12-1.06E-03</td>
</tr>
<tr>
<td>Hematopoiesis</td>
<td>5.93E-12-9.7E-04</td>
</tr>
<tr>
<td>Cardiovascular System Development and Function</td>
<td>7.64E-12-8.95E-04</td>
</tr>
<tr>
<td>Dermatological Diseases and Conditions</td>
<td>8.62E-12-1.1E-05</td>
</tr>
<tr>
<td>Genetic Disorder</td>
<td>8.62E-12-1.4E-03</td>
</tr>
<tr>
<td>Cellular Function and Maintenance</td>
<td>1.03E-11-1.4E-03</td>
</tr>
<tr>
<td>Hematological Disease</td>
<td>1.96E-11-1.4E-03</td>
</tr>
<tr>
<td>Reproductive System Disease</td>
<td>5.13E-11-1.24E-03</td>
</tr>
<tr>
<td>Antigen Presentation</td>
<td>7.7E-09-4.1E-04</td>
</tr>
<tr>
<td>Cell-mediated Immune Response</td>
<td>1.39E-08-1.3E-03</td>
</tr>
<tr>
<td>Cellular Compromise</td>
<td>3.31E-08-2.22E-04</td>
</tr>
<tr>
<td>Cell Morphology</td>
<td>7.13E-08-1.4E-03</td>
</tr>
<tr>
<td>Gastrointestinal Disease</td>
<td>2.45E-07-8.49E-05</td>
</tr>
<tr>
<td>Lymphoid Tissue Structure and Development</td>
<td>2.47E-07-4.91E-04</td>
</tr>
<tr>
<td>Neurological Disease</td>
<td>1.12E-06-1.17E-03</td>
</tr>
<tr>
<td>Category</td>
<td>Log10(p)</td>
</tr>
<tr>
<td>--------------------------------------------------------------</td>
<td>----------------</td>
</tr>
<tr>
<td>Cardiovascular Disease</td>
<td>2.83E-06</td>
</tr>
<tr>
<td>Cellular Assembly and Organization</td>
<td>7.98E-06</td>
</tr>
<tr>
<td>Lipid Metabolism</td>
<td>8.5E-06</td>
</tr>
<tr>
<td>Small Molecule Biochemistry</td>
<td>8.5E-06</td>
</tr>
<tr>
<td>Post-Translational Modification</td>
<td>1.38E-05</td>
</tr>
<tr>
<td>DNA Replication, Recombination, and Repair</td>
<td>1.41E-05</td>
</tr>
<tr>
<td>Renal and Urological Disease</td>
<td>1.9E-05</td>
</tr>
<tr>
<td>Infection Mechanism</td>
<td>2.2E-05</td>
</tr>
<tr>
<td>Tumor Morphology</td>
<td>2.62E-05</td>
</tr>
<tr>
<td>Organismal Injury and Abnormalities</td>
<td>3.31E-05</td>
</tr>
<tr>
<td>Carbohydrate Metabolism</td>
<td>3.47E-05</td>
</tr>
<tr>
<td>Free Radical Scavenging</td>
<td>3.54E-05</td>
</tr>
<tr>
<td>Hypersensitivity Response</td>
<td>4.75E-05</td>
</tr>
<tr>
<td>Tissue Morphology</td>
<td>9.97E-05</td>
</tr>
<tr>
<td>Skeletal and Muscular System Development and Function</td>
<td>1.65E-04</td>
</tr>
<tr>
<td>Ophthalmic Disease</td>
<td>2.02E-04</td>
</tr>
<tr>
<td>Organismal Functions</td>
<td>2.1E-04</td>
</tr>
<tr>
<td>Organismal Development</td>
<td>2.24E-04</td>
</tr>
<tr>
<td>Antimicrobial Response</td>
<td>2.56E-04</td>
</tr>
<tr>
<td>Humoral Immune Response</td>
<td>2.56E-04</td>
</tr>
<tr>
<td>Renal and Urological System Development and Function</td>
<td>4.02E-04</td>
</tr>
<tr>
<td>Connective Tissue Development and Function</td>
<td>4.91E-04</td>
</tr>
<tr>
<td>Gene Expression</td>
<td>4.91E-04</td>
</tr>
<tr>
<td>Drug Metabolism</td>
<td>6.26E-04</td>
</tr>
<tr>
<td>Embryonic Development</td>
<td>7.53E-04</td>
</tr>
<tr>
<td>Metabolic Disease</td>
<td>8.95E-04</td>
</tr>
<tr>
<td>Nervous System Development and Function</td>
<td>9.56E-04</td>
</tr>
<tr>
<td>Hair and Skin Development and Function</td>
<td>1.24E-03</td>
</tr>
<tr>
<td>Protein Trafficking</td>
<td>1.4E-03</td>
</tr>
</tbody>
</table>
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 $\rightarrow$ D8 $\rightarrow$ D15 $\rightarrow$ 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 $\rightarrow$ D8 $\rightarrow$ D15 $\rightarrow$ D33 $\rightarrow$ 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.
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.
CHAPTER 4  GENETIC STATUS SHIFTING 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% ± 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.
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.
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.
CHAPTER 4  GENETIC STATUS SHIFTING MODEL

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

<table>
<thead>
<tr>
<th>Subtype: Hyperdiploid&gt;50</th>
<th>PC1</th>
<th>PC2</th>
<th>PC3</th>
<th>PC4</th>
<th>PC5</th>
<th>PC6</th>
<th>PC7</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Variance (MILE)</td>
<td>90.19%</td>
<td>1.96%</td>
<td>0.82%</td>
<td>0.73%</td>
<td>0.62%</td>
<td>0.52%</td>
<td>0.48%</td>
<td>95.32%</td>
</tr>
<tr>
<td>Variance (MASPORE)</td>
<td>77.12%</td>
<td>6.02%</td>
<td>1.18%</td>
<td>0.99%</td>
<td>0.50%</td>
<td>0.70%</td>
<td>0.39%</td>
<td>86.90%</td>
</tr>
</tbody>
</table>
Figure 4.17: The local GSS model of E2A-PBX1 subtype. (a) PC1 to PC2. (b) The variance contained in top PCs.
Figure 4.18: The local GSS model of BCR-ABL subtype. (a) PC1 to PC2. (b) The variance contained in top PCs.
Figure 4.19: The local GSS model of MLL subtype. (a) PC1 to PC2. (b) The variance contained in top PCs.
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 $50 \times 10^9$ 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.
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 \( s \rightarrow s' \).
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,

$$ \text{ASD}(\overline{ss'}) = \lvert \overline{ss'} \rvert. $$

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,

$$ \text{ESD}(\overline{ss'}) = \frac{\overline{ss'} \cdot \overline{dn}}{\lvert \overline{dn} \rvert}. $$

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,

$$ \text{ESR}(\overline{ss'}) = \frac{\overline{ss'} \cdot \overline{dn}}{\lvert \overline{dn} \rvert} / \frac{\overline{sn} \cdot \overline{dn}}{\lvert \overline{dn} \rvert} $$

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*.

<table>
<thead>
<tr>
<th>RANK</th>
<th>SAMPLE</th>
<th>ASD-D8</th>
<th>RANK</th>
<th>SAMPLE</th>
<th>ASD-D8</th>
<th>RANK</th>
<th>SAMPLE</th>
<th>ASD-D8</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>67_KL287</td>
<td>4.64</td>
<td>30</td>
<td>86_KL509</td>
<td>35.11</td>
<td>59</td>
<td>92_R332</td>
<td>55.25</td>
</tr>
<tr>
<td>2</td>
<td>97_R208</td>
<td>4.78</td>
<td>31</td>
<td>96_R202</td>
<td>36.65</td>
<td>60</td>
<td>93_R337</td>
<td>55.70</td>
</tr>
<tr>
<td>3</td>
<td>59_R220</td>
<td>4.30</td>
<td>32</td>
<td>51_KL461</td>
<td>37.44</td>
<td>61</td>
<td>26_KL369</td>
<td>57.13</td>
</tr>
<tr>
<td>4</td>
<td>11_R280</td>
<td>7.06</td>
<td>33</td>
<td>24_KL328</td>
<td>37.82</td>
<td>62</td>
<td>80_KL423</td>
<td>58.19</td>
</tr>
<tr>
<td>5</td>
<td>19_KL205</td>
<td>7.76</td>
<td>34</td>
<td>30_KL444</td>
<td>42.17</td>
<td>63</td>
<td>69_KL313</td>
<td>58.85</td>
</tr>
<tr>
<td>6</td>
<td>70_KL320</td>
<td>8.35</td>
<td>35</td>
<td>38_KL218</td>
<td>38.51</td>
<td>64</td>
<td>14_KKH19</td>
<td>59.57</td>
</tr>
<tr>
<td>7</td>
<td>56_KL464</td>
<td>8.62</td>
<td>36</td>
<td>47_R334</td>
<td>40.16</td>
<td>65</td>
<td>62_KKH22</td>
<td>60.80</td>
</tr>
<tr>
<td>8</td>
<td>20_KL274</td>
<td>9.99</td>
<td>37</td>
<td>74_KL383</td>
<td>40.85</td>
<td>66</td>
<td>01_KKH18</td>
<td>62.21</td>
</tr>
<tr>
<td>9</td>
<td>13_R410</td>
<td>11.94</td>
<td>38</td>
<td>57_KL535</td>
<td>40.99</td>
<td>67</td>
<td>75_KL385</td>
<td>62.85</td>
</tr>
<tr>
<td>10</td>
<td>41_KL441</td>
<td>12.25</td>
<td>39</td>
<td>10_R257</td>
<td>41.47</td>
<td>68</td>
<td>31_KL300</td>
<td>63.15</td>
</tr>
<tr>
<td>11</td>
<td>27_KL374</td>
<td>13.74</td>
<td>40</td>
<td>30_KL444</td>
<td>42.17</td>
<td>69</td>
<td>64_KKH29</td>
<td>64.44</td>
</tr>
<tr>
<td>12</td>
<td>35_R313</td>
<td>14.00</td>
<td>41</td>
<td>84_KL458</td>
<td>43.49</td>
<td>70</td>
<td>28_KL375</td>
<td>68.24</td>
</tr>
<tr>
<td>13</td>
<td>82_KL454</td>
<td>16.84</td>
<td>42</td>
<td>16_KKH21</td>
<td>43.55</td>
<td>71</td>
<td>32_R233</td>
<td>68.38</td>
</tr>
<tr>
<td>14</td>
<td>77_KL401</td>
<td>16.95</td>
<td>43</td>
<td>60_KKH30</td>
<td>43.62</td>
<td>72</td>
<td>34_R256</td>
<td>71.98</td>
</tr>
<tr>
<td>15</td>
<td>33_R247</td>
<td>19.22</td>
<td>44</td>
<td>83_KL457</td>
<td>43.88</td>
<td>73</td>
<td>08_KL456</td>
<td>73.10</td>
</tr>
<tr>
<td>16</td>
<td>55_KL419</td>
<td>19.37</td>
<td>45</td>
<td>66_KL247</td>
<td>44.46</td>
<td>74</td>
<td>68_KL304</td>
<td>73.27</td>
</tr>
<tr>
<td>17</td>
<td>39_KL395</td>
<td>19.79</td>
<td>46</td>
<td>50_KL360</td>
<td>45.17</td>
<td>75</td>
<td>18_KKH28</td>
<td>73.89</td>
</tr>
<tr>
<td>18</td>
<td>45_R194</td>
<td>23.01</td>
<td>47</td>
<td>37_R355</td>
<td>46.57</td>
<td>76</td>
<td>58_KL543</td>
<td>74.46</td>
</tr>
<tr>
<td>19</td>
<td>07_KL417</td>
<td>24.00</td>
<td>48</td>
<td>15_KKH20</td>
<td>46.78</td>
<td>77</td>
<td>36_R343</td>
<td>75.96</td>
</tr>
<tr>
<td>20</td>
<td>17_KKH27</td>
<td>24.58</td>
<td>49</td>
<td>94_R354</td>
<td>47.17</td>
<td>78</td>
<td>88_KL544</td>
<td>78.00</td>
</tr>
<tr>
<td>21</td>
<td>23_KL321</td>
<td>24.89</td>
<td>50</td>
<td>61_KKH13</td>
<td>48.70</td>
<td>79</td>
<td>12_R297</td>
<td>78.10</td>
</tr>
<tr>
<td>22</td>
<td>65_KL224</td>
<td>25.57</td>
<td>51</td>
<td>52_R252</td>
<td>48.97</td>
<td>80</td>
<td>72_KL377</td>
<td>80.71</td>
</tr>
<tr>
<td>23</td>
<td>29_KL439</td>
<td>26.93</td>
<td>52</td>
<td>73_KL381</td>
<td>49.52</td>
<td>81</td>
<td>87_KL522</td>
<td>81.06</td>
</tr>
<tr>
<td>24</td>
<td>43_KL536</td>
<td>28.38</td>
<td>53</td>
<td>48_R339</td>
<td>49.58</td>
<td>82</td>
<td>89_R245</td>
<td>85.64</td>
</tr>
<tr>
<td>25</td>
<td>90_R253</td>
<td>28.72</td>
<td>54</td>
<td>25_KL357</td>
<td>49.66</td>
<td>83</td>
<td>44_KL541</td>
<td>89.50</td>
</tr>
<tr>
<td>26</td>
<td>49_R432</td>
<td>29.27</td>
<td>55</td>
<td>85_KL485</td>
<td>51.39</td>
<td>84</td>
<td>63_KKH25</td>
<td>90.18</td>
</tr>
<tr>
<td>27</td>
<td>42_KL507</td>
<td>30.29</td>
<td>56</td>
<td>98_KL387</td>
<td>53.21</td>
<td>85</td>
<td>78_KL412</td>
<td>98.62</td>
</tr>
<tr>
<td>28</td>
<td>04_KL322</td>
<td>30.64</td>
<td>57</td>
<td>79_KL421</td>
<td>54.57</td>
<td>86</td>
<td>95_R431</td>
<td>105.60</td>
</tr>
<tr>
<td>29</td>
<td>99_KL416</td>
<td>31.70</td>
<td>58</td>
<td>05_KL354</td>
<td>54.59</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 5.2: ASD between the D0 and D15 samples. Relapses are highlighted with **Underline**. Extremely slow responders are highlighted in *Italic*.

<table>
<thead>
<tr>
<th>RANK</th>
<th>SAMPLE</th>
<th>ASD-D15</th>
<th>RANK</th>
<th>SAMPLE</th>
<th>ASD-D15</th>
<th>RANK</th>
<th>SAMPLE</th>
<th>ASD-D15</th>
<th>RANK</th>
<th>SAMPLE</th>
<th>ASD-D15</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>17_KL401</td>
<td>1.41</td>
<td>17</td>
<td>27_KL374</td>
<td>64.24</td>
<td>33</td>
<td>48_R339</td>
<td>79.94</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>97_R208</td>
<td>15.11</td>
<td>18</td>
<td>85_KL485</td>
<td>64.72</td>
<td>34</td>
<td>64_KKH29</td>
<td>80.13</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>56_KL464</td>
<td>25.51</td>
<td>19</td>
<td>80_KL423</td>
<td>66.69</td>
<td>35</td>
<td>28_KL375</td>
<td>82.79</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>76_KL398</td>
<td>27.47</td>
<td>20</td>
<td>99_KL416</td>
<td>66.79</td>
<td>36</td>
<td>87_KL522</td>
<td>84.89</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>96_R202</td>
<td>29.09</td>
<td>21</td>
<td>06_KL378</td>
<td>66.80</td>
<td>37</td>
<td>93_R337</td>
<td>87.84</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>68_KL391</td>
<td>17.64</td>
<td>22</td>
<td>61_KKH13</td>
<td>68.18</td>
<td>38</td>
<td>63_KKH25</td>
<td>92.24</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>67_KL287</td>
<td>41.60</td>
<td>23</td>
<td>74_KL383</td>
<td>69.77</td>
<td>39</td>
<td>14_KKH19</td>
<td>92.65</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>71_KL371</td>
<td>42.05</td>
<td>24</td>
<td>98_KL387</td>
<td>71.15</td>
<td>40</td>
<td>49_R432</td>
<td>93.43</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>50_KL360</td>
<td>47.50</td>
<td>25</td>
<td>84_KL458</td>
<td>71.19</td>
<td>41</td>
<td>94_R354</td>
<td>94.20</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>16_KKH21</td>
<td>51.87</td>
<td>26</td>
<td>18_KKH28</td>
<td>73.27</td>
<td>42</td>
<td>78_KL412</td>
<td>101.49</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>05_KL354</td>
<td>55.11</td>
<td>27</td>
<td>01_KKH18</td>
<td>74.20</td>
<td>43</td>
<td>37_R355</td>
<td>102.25</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>73_KL381</td>
<td>55.42</td>
<td>28</td>
<td>92_R332</td>
<td>74.38</td>
<td>44</td>
<td>36_R343</td>
<td>104.16</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>13_R410</td>
<td>58.63</td>
<td>29</td>
<td>52_R252</td>
<td>76.67</td>
<td>45</td>
<td>29_KL439</td>
<td>104.52</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>41_KL441</td>
<td>58.80</td>
<td>30</td>
<td>62_KKH22</td>
<td>78.08</td>
<td>46</td>
<td>20_KL274</td>
<td><strong>107.18</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>10_R257</td>
<td>62.54</td>
<td>31</td>
<td>86_KL509</td>
<td>79.17</td>
<td>47</td>
<td>95_R431</td>
<td>110.76</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>15_KKH20</td>
<td>63.69</td>
<td>32</td>
<td>88_KL544</td>
<td>79.66</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 5.3: ASD between the D0 and D33 samples. Relapses are highlighted with **Underline**. Extremely slow responders are highlighted in *Italic*.

<table>
<thead>
<tr>
<th>RANK</th>
<th>SAMPLE</th>
<th>ASD-D33</th>
<th>RANK</th>
<th>SAMPLE</th>
<th>ASD-D33</th>
<th>RANK</th>
<th>SAMPLE</th>
<th>ASD-D33</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>97_R208</td>
<td>9.73</td>
<td>20</td>
<td>66_KL247</td>
<td>72.91</td>
<td>39</td>
<td>83_KL457</td>
<td>87.30</td>
</tr>
<tr>
<td>2</td>
<td>96_R202</td>
<td>9.78</td>
<td>21</td>
<td>01_KKH18</td>
<td>73.20</td>
<td>40</td>
<td>82_KL454</td>
<td>87.42</td>
</tr>
<tr>
<td>3</td>
<td>76_KL398</td>
<td>22.02</td>
<td>22</td>
<td>18_KKH28</td>
<td>73.75</td>
<td>41</td>
<td>28_KL375</td>
<td>92.46</td>
</tr>
<tr>
<td>4</td>
<td>71_R249</td>
<td><strong>9.34</strong></td>
<td>23</td>
<td>84_KL458</td>
<td>74.46</td>
<td>42</td>
<td>63_KKH25</td>
<td>93.00</td>
</tr>
<tr>
<td>5</td>
<td>71_KL371</td>
<td>34.22</td>
<td>24</td>
<td>72_KL377</td>
<td>75.73</td>
<td>43</td>
<td>38_KL218</td>
<td>95.79</td>
</tr>
<tr>
<td>6</td>
<td>72_KL401</td>
<td><strong>9.41</strong></td>
<td>25</td>
<td>25_KL357</td>
<td>79.29</td>
<td>44</td>
<td>89_R245</td>
<td>95.80</td>
</tr>
<tr>
<td>7</td>
<td>50_KL360</td>
<td>47.48</td>
<td>26</td>
<td>88_KL544</td>
<td>79.96</td>
<td>45</td>
<td>49_R432</td>
<td>96.60</td>
</tr>
<tr>
<td>8</td>
<td>40_KL430</td>
<td>50.50</td>
<td>27</td>
<td>75_KL385</td>
<td>81.02</td>
<td>46</td>
<td>86_KL509</td>
<td>97.59</td>
</tr>
<tr>
<td>9</td>
<td>51_KL461</td>
<td>51.72</td>
<td>28</td>
<td>07_KL417</td>
<td><strong>81.33</strong></td>
<td>47</td>
<td>37_R355</td>
<td>98.98</td>
</tr>
<tr>
<td>10</td>
<td>67_KL287</td>
<td>52.87</td>
<td>29</td>
<td>87_KL522</td>
<td>81.46</td>
<td>48</td>
<td>55_KL419</td>
<td>100.11</td>
</tr>
<tr>
<td>11</td>
<td>42_KL507</td>
<td>56.30</td>
<td>30</td>
<td>90_R253</td>
<td><strong>81.50</strong></td>
<td>49</td>
<td>44_KL541</td>
<td>100.20</td>
</tr>
<tr>
<td>12</td>
<td>27_KL374</td>
<td>57.15</td>
<td>31</td>
<td>47_R334</td>
<td>81.52</td>
<td>50</td>
<td>20_KL274</td>
<td><strong>101.95</strong></td>
</tr>
<tr>
<td>13</td>
<td>60_KKH30</td>
<td>58.28</td>
<td>32</td>
<td>64_KKH29</td>
<td>81.77</td>
<td>51</td>
<td>29_KL439</td>
<td>102.15</td>
</tr>
<tr>
<td>14</td>
<td>85_KL485</td>
<td>60.39</td>
<td>33</td>
<td>62_KKH22</td>
<td>82.52</td>
<td>52</td>
<td>78_KL412</td>
<td>103.99</td>
</tr>
<tr>
<td>15</td>
<td>10_R257</td>
<td>62.15</td>
<td>34</td>
<td>19_KL205</td>
<td><strong>82.81</strong></td>
<td>53</td>
<td>14_KKH19</td>
<td>104.77</td>
</tr>
<tr>
<td>16</td>
<td>80_KL423</td>
<td>62.25</td>
<td>35</td>
<td>06_KL378</td>
<td>83.04</td>
<td>54</td>
<td>17_KKH27</td>
<td>106.20</td>
</tr>
<tr>
<td>17</td>
<td>11_R280</td>
<td><strong>64.21</strong></td>
<td>36</td>
<td>39_KL395</td>
<td><strong>83.29</strong></td>
<td>55</td>
<td>43_KL536</td>
<td>106.28</td>
</tr>
<tr>
<td>18</td>
<td>65_KL224</td>
<td>68.79</td>
<td>37</td>
<td>52_R252</td>
<td>86.10</td>
<td>56</td>
<td>30_KL444</td>
<td>110.44</td>
</tr>
<tr>
<td>19</td>
<td>69_KL313</td>
<td>72.80</td>
<td>38</td>
<td>08_KL456</td>
<td>86.44</td>
<td>57</td>
<td>95_R431</td>
<td>112.59</td>
</tr>
</tbody>
</table>
**CHAPTER 5 RELAPSE PREDICTION**

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

<table>
<thead>
<tr>
<th>RANK</th>
<th>SAMPLE</th>
<th>ESD-D8</th>
<th>RANK</th>
<th>SAMPLE</th>
<th>ESD-D8</th>
<th>RANK</th>
<th>SAMPLE</th>
<th>ESD-D8</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>KL464</td>
<td>-2.21</td>
<td>30</td>
<td>KL439</td>
<td>26.20</td>
<td>59</td>
<td>KL354</td>
<td>52.56</td>
</tr>
<tr>
<td>2</td>
<td>KL491</td>
<td>-4.31</td>
<td>31</td>
<td>R432</td>
<td>26.77</td>
<td>60</td>
<td>KL423</td>
<td>53.33</td>
</tr>
<tr>
<td>3</td>
<td>R247</td>
<td>-0.46</td>
<td>32</td>
<td>KL457</td>
<td>29.13</td>
<td>61</td>
<td>KL385</td>
<td>54.31</td>
</tr>
<tr>
<td>4</td>
<td>KL205</td>
<td>0.21</td>
<td>33</td>
<td>KL322</td>
<td>30.64</td>
<td>62</td>
<td>KL313</td>
<td>54.40</td>
</tr>
<tr>
<td>5</td>
<td>R194</td>
<td>2.13</td>
<td>34</td>
<td>KL383</td>
<td>31.52</td>
<td>63</td>
<td>R337</td>
<td>55.12</td>
</tr>
<tr>
<td>6</td>
<td>R128</td>
<td>2.52</td>
<td>35</td>
<td>KKH30</td>
<td>32.88</td>
<td>64</td>
<td>R337</td>
<td>55.35</td>
</tr>
<tr>
<td>7</td>
<td>R208</td>
<td>2.94</td>
<td>36</td>
<td>KL218</td>
<td>33.00</td>
<td>65</td>
<td>KL369</td>
<td>56.83</td>
</tr>
<tr>
<td>8</td>
<td>KL287</td>
<td>4.63</td>
<td>37</td>
<td>R257</td>
<td>33.39</td>
<td>66</td>
<td>KKH19</td>
<td>57.79</td>
</tr>
<tr>
<td>9</td>
<td>R202</td>
<td>5.14</td>
<td>38</td>
<td>KL430</td>
<td>33.55</td>
<td>67</td>
<td>KL375</td>
<td>58.21</td>
</tr>
<tr>
<td>10</td>
<td>R280</td>
<td>5.90</td>
<td>39</td>
<td>KL509</td>
<td>34.48</td>
<td>68</td>
<td>KKH22</td>
<td>60.55</td>
</tr>
<tr>
<td>11</td>
<td>KL374</td>
<td>6.31</td>
<td>40</td>
<td>KL461</td>
<td>34.83</td>
<td>69</td>
<td>KL300</td>
<td>60.80</td>
</tr>
<tr>
<td>12</td>
<td>KL320</td>
<td>6.64</td>
<td>41</td>
<td>KL421</td>
<td>35.43</td>
<td>70</td>
<td>KKH18</td>
<td>62.07</td>
</tr>
<tr>
<td>13</td>
<td>KL441</td>
<td>7.86</td>
<td>42</td>
<td>R354</td>
<td>35.66</td>
<td>71</td>
<td>R233</td>
<td>67.35</td>
</tr>
<tr>
<td>14</td>
<td>KL274</td>
<td>8.91</td>
<td>43</td>
<td>KL328</td>
<td>35.78</td>
<td>72</td>
<td>R256</td>
<td>67.95</td>
</tr>
<tr>
<td>15</td>
<td>KL454</td>
<td>9.53</td>
<td>44</td>
<td>R334</td>
<td>37.35</td>
<td>73</td>
<td>KL304</td>
<td>70.06</td>
</tr>
<tr>
<td>16</td>
<td>R410</td>
<td>11.45</td>
<td>45</td>
<td>KL448</td>
<td>38.26</td>
<td>74</td>
<td>KKH28</td>
<td>70.41</td>
</tr>
<tr>
<td>17</td>
<td>KL395</td>
<td>11.49</td>
<td>46</td>
<td>KL353</td>
<td>39.40</td>
<td>75</td>
<td>R343</td>
<td>71.95</td>
</tr>
<tr>
<td>18</td>
<td>KKH27</td>
<td>12.64</td>
<td>47</td>
<td>KL381</td>
<td>39.51</td>
<td>76</td>
<td>KL543</td>
<td>71.96</td>
</tr>
<tr>
<td>19</td>
<td>R313</td>
<td>13.76</td>
<td>48</td>
<td>KL247</td>
<td>39.69</td>
<td>77</td>
<td>KL456</td>
<td>72.39</td>
</tr>
<tr>
<td>20</td>
<td>KL419</td>
<td>18.60</td>
<td>49</td>
<td>KKH21</td>
<td>40.05</td>
<td>78</td>
<td>R297</td>
<td>74.21</td>
</tr>
<tr>
<td>21</td>
<td>KL416</td>
<td>19.83</td>
<td>50</td>
<td>KL458</td>
<td>40.12</td>
<td>79</td>
<td>KL544</td>
<td>75.25</td>
</tr>
<tr>
<td>22</td>
<td>KL321</td>
<td>20.37</td>
<td>51</td>
<td>KL355</td>
<td>41.68</td>
<td>80</td>
<td>KL377</td>
<td>76.58</td>
</tr>
<tr>
<td>23</td>
<td>KL360</td>
<td>20.80</td>
<td>52</td>
<td>KL357</td>
<td>42.14</td>
<td>81</td>
<td>KL541</td>
<td>79.45</td>
</tr>
<tr>
<td>24</td>
<td>R253</td>
<td>21.40</td>
<td>53</td>
<td>KKH20</td>
<td>43.35</td>
<td>82</td>
<td>KL522</td>
<td>79.91</td>
</tr>
<tr>
<td>25</td>
<td>KKH13</td>
<td>21.67</td>
<td>54</td>
<td>R252</td>
<td>45.31</td>
<td>83</td>
<td>R245</td>
<td>85.37</td>
</tr>
<tr>
<td>26</td>
<td>KL536</td>
<td>23.00</td>
<td>55</td>
<td>KL485</td>
<td>45.82</td>
<td>84</td>
<td>KKH28</td>
<td>85.67</td>
</tr>
<tr>
<td>27</td>
<td>KL224</td>
<td>23.30</td>
<td>56</td>
<td>R332</td>
<td>49.15</td>
<td>85</td>
<td>KL412</td>
<td>92.38</td>
</tr>
<tr>
<td>28</td>
<td>KL417</td>
<td>24.00</td>
<td>57</td>
<td>KL387</td>
<td>49.17</td>
<td>86</td>
<td>R431</td>
<td>104.57</td>
</tr>
<tr>
<td>29</td>
<td>KL507</td>
<td>24.00</td>
<td>58</td>
<td>R339</td>
<td>49.26</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 5.5: ESD between the D0 and D15 samples. Relapses are highlighted with **Underline**. Extremely slow responders are highlighted in *Italic*.

<table>
<thead>
<tr>
<th>RANK</th>
<th>SAMPLE</th>
<th>ESD-D15</th>
<th>RANK</th>
<th>SAMPLE</th>
<th>ESD-D15</th>
<th>RANK</th>
<th>SAMPLE</th>
<th>ESD-D15</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>97_R208</td>
<td>-12.33</td>
<td>17</td>
<td>13_R410</td>
<td>56.01</td>
<td>33</td>
<td>88_KL544</td>
<td>78.10</td>
</tr>
<tr>
<td>2</td>
<td>77_KL464</td>
<td>1.54</td>
<td>18</td>
<td>15_KKH20</td>
<td>58.17</td>
<td>34</td>
<td>48_R339</td>
<td>78.48</td>
</tr>
<tr>
<td>3</td>
<td>96_R202</td>
<td>5.90</td>
<td>19</td>
<td>10_R257</td>
<td>58.40</td>
<td>35</td>
<td>93_R337</td>
<td>81.68</td>
</tr>
<tr>
<td>4</td>
<td>56_KL464</td>
<td>13.45</td>
<td>20</td>
<td>84_KL458</td>
<td>60.75</td>
<td>36</td>
<td>28_KL375</td>
<td>81.79</td>
</tr>
<tr>
<td>5</td>
<td>19_R211</td>
<td>17.13</td>
<td>21</td>
<td>80_KL423</td>
<td>60.92</td>
<td>37</td>
<td>87_KL522</td>
<td>83.85</td>
</tr>
<tr>
<td>6</td>
<td>76_KL398</td>
<td>21.30</td>
<td>22</td>
<td>99_KL416</td>
<td>61.21</td>
<td>38</td>
<td>63_KKH25</td>
<td>86.06</td>
</tr>
<tr>
<td>7</td>
<td>71_KL371</td>
<td>23.56</td>
<td>23</td>
<td>06_KL378</td>
<td>61.58</td>
<td>39</td>
<td>49_R432</td>
<td>88.27</td>
</tr>
<tr>
<td>8</td>
<td>50_KL360</td>
<td>26.07</td>
<td>24</td>
<td>85_KL485</td>
<td>63.61</td>
<td>40</td>
<td>94_R354</td>
<td>88.83</td>
</tr>
<tr>
<td>9</td>
<td>67_KL287</td>
<td>29.96</td>
<td>25</td>
<td>64_KKH29</td>
<td>64.65</td>
<td>41</td>
<td>14_KKH19</td>
<td>92.64</td>
</tr>
<tr>
<td>10</td>
<td>27_KL374</td>
<td>43.64</td>
<td>26</td>
<td>18_KKH28</td>
<td>66.23</td>
<td>42</td>
<td>37_R355</td>
<td>92.94</td>
</tr>
<tr>
<td>11</td>
<td>61_KKH13</td>
<td>46.06</td>
<td>27</td>
<td>92_R332</td>
<td>67.53</td>
<td>43</td>
<td>29_KL439</td>
<td>95.71</td>
</tr>
<tr>
<td>12</td>
<td>16_KKH21</td>
<td>46.50</td>
<td>28</td>
<td>52_R252</td>
<td>68.87</td>
<td>44</td>
<td>78_KL412</td>
<td>96.95</td>
</tr>
<tr>
<td>13</td>
<td>73_KL381</td>
<td>53.48</td>
<td>29</td>
<td>98_KL387</td>
<td>69.83</td>
<td>45</td>
<td>36_R343</td>
<td>100.91</td>
</tr>
<tr>
<td>14</td>
<td>74_KL383</td>
<td>53.86</td>
<td>30</td>
<td>01_KKH18</td>
<td>74.10</td>
<td>46</td>
<td>20_KL274</td>
<td>104.35</td>
</tr>
<tr>
<td>15</td>
<td>05_KL354</td>
<td>54.14</td>
<td>31</td>
<td>62_KKH22</td>
<td>74.64</td>
<td>47</td>
<td>95_R431</td>
<td>110.53</td>
</tr>
<tr>
<td>16</td>
<td>41_KL441</td>
<td>54.81</td>
<td>32</td>
<td>86_KL509</td>
<td>76.20</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 5.6: ESD between the D0 and D33 samples. Relapses are highlighted with Underline. Extremely slow responders are highlighted in Italic.

<table>
<thead>
<tr>
<th>RANK</th>
<th>SAMPLE</th>
<th>ESD-D33</th>
<th>RANK</th>
<th>SAMPLE</th>
<th>ESD-D33</th>
<th>RANK</th>
<th>SAMPLE</th>
<th>ESD-D33</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>97_R208</td>
<td>-0.56</td>
<td>20</td>
<td>69_KL313</td>
<td>70.87</td>
<td>39</td>
<td>39_KL395</td>
<td>81.71</td>
</tr>
<tr>
<td>2</td>
<td>96_R202</td>
<td>0.34</td>
<td>21</td>
<td>18_KKH28</td>
<td>71.84</td>
<td>40</td>
<td>82_KL454</td>
<td>87.15</td>
</tr>
<tr>
<td>3</td>
<td>76_KL398</td>
<td>4.54</td>
<td>22</td>
<td>66_KL247</td>
<td>72.90</td>
<td>41</td>
<td>28_KL375</td>
<td>87.29</td>
</tr>
<tr>
<td>4</td>
<td>79_R287</td>
<td>18.14</td>
<td>23</td>
<td>84_KL458</td>
<td>72.93</td>
<td>42</td>
<td>63_KKH25</td>
<td>92.22</td>
</tr>
<tr>
<td>5</td>
<td>71_KL371</td>
<td>31.51</td>
<td>24</td>
<td>08_KL456</td>
<td>73.05</td>
<td>43</td>
<td>38_KL218</td>
<td>93.89</td>
</tr>
<tr>
<td>6</td>
<td>51_KL461</td>
<td>39.41</td>
<td>25</td>
<td>72_KL377</td>
<td>74.77</td>
<td>44</td>
<td>89_R245</td>
<td>94.81</td>
</tr>
<tr>
<td>7</td>
<td>27_KL406</td>
<td>41.74</td>
<td>26</td>
<td>07_KL417</td>
<td>76.35</td>
<td>45</td>
<td>86_KL509</td>
<td>95.31</td>
</tr>
<tr>
<td>8</td>
<td>40_KL430</td>
<td>46.33</td>
<td>27</td>
<td>83_KL457</td>
<td>76.72</td>
<td>46</td>
<td>17_KKH27</td>
<td>95.47</td>
</tr>
<tr>
<td>9</td>
<td>50_KL360</td>
<td>47.43</td>
<td>28</td>
<td>19_KL205</td>
<td>76.86</td>
<td>47</td>
<td>55_KL419</td>
<td>95.92</td>
</tr>
<tr>
<td>10</td>
<td>67_KL287</td>
<td>52.37</td>
<td>29</td>
<td>47_R334</td>
<td>78.70</td>
<td>48</td>
<td>49_R432</td>
<td>96.41</td>
</tr>
<tr>
<td>11</td>
<td>42_KL507</td>
<td>53.87</td>
<td>30</td>
<td>87_KL522</td>
<td>78.76</td>
<td>49</td>
<td>29_KL439</td>
<td>97.66</td>
</tr>
<tr>
<td>12</td>
<td>60_KKH30</td>
<td>56.61</td>
<td>31</td>
<td>90_R253</td>
<td>78.83</td>
<td>50</td>
<td>37_R355</td>
<td>98.84</td>
</tr>
<tr>
<td>13</td>
<td>27_KL374</td>
<td>57.11</td>
<td>32</td>
<td>88_KL544</td>
<td>79.25</td>
<td>51</td>
<td>44_KL541</td>
<td>99.89</td>
</tr>
<tr>
<td>14</td>
<td>85_KL485</td>
<td>59.75</td>
<td>33</td>
<td>75_KL385</td>
<td>79.26</td>
<td>52</td>
<td>78_KL412</td>
<td>99.97</td>
</tr>
<tr>
<td>15</td>
<td>80_KL423</td>
<td>61.74</td>
<td>34</td>
<td>25_KL357</td>
<td>79.27</td>
<td>53</td>
<td>20_KL274</td>
<td>100.55</td>
</tr>
<tr>
<td>16</td>
<td>10_R257</td>
<td>61.81</td>
<td>35</td>
<td>64_KKH29</td>
<td>80.39</td>
<td>54</td>
<td>14_KKH19</td>
<td>100.74</td>
</tr>
<tr>
<td>17</td>
<td>11_R280</td>
<td>63.24</td>
<td>36</td>
<td>62_KKH22</td>
<td>80.65</td>
<td>55</td>
<td>43_KL536</td>
<td>104.58</td>
</tr>
<tr>
<td>18</td>
<td>65_KL224</td>
<td>65.66</td>
<td>37</td>
<td>52_R252</td>
<td>81.43</td>
<td>56</td>
<td>30_KL444</td>
<td>106.99</td>
</tr>
<tr>
<td>19</td>
<td>01_KKH18</td>
<td>69.80</td>
<td>38</td>
<td>06_KL378</td>
<td>81.69</td>
<td>57</td>
<td>95_R431</td>
<td>111.91</td>
</tr>
</tbody>
</table>
Table 5.7: ESR between the D0 and D8 samples. Relapses are highlighted with **underline**. Extremely slow responders are highlighted in *italic*.

<table>
<thead>
<tr>
<th>RANK</th>
<th>SAMPLE</th>
<th>ESR-D8</th>
<th>RANK</th>
<th>SAMPLE</th>
<th>ESR-D8</th>
<th>RANK</th>
<th>SAMPLE</th>
<th>ESR-D8</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>56_KL464</td>
<td>-0.04</td>
<td>30</td>
<td>07_KL417</td>
<td>0.36</td>
<td>59</td>
<td>36_R343</td>
<td>0.69</td>
</tr>
<tr>
<td>2</td>
<td>77_KL401</td>
<td>0.04</td>
<td>31</td>
<td>30_KL444</td>
<td>0.37</td>
<td>60</td>
<td>69_KL313</td>
<td>0.72</td>
</tr>
<tr>
<td>3</td>
<td>33_R247</td>
<td>-0.01</td>
<td>32</td>
<td>86_KL509</td>
<td>0.38</td>
<td>61</td>
<td>21_KL300</td>
<td>0.75</td>
</tr>
<tr>
<td>4</td>
<td>19_KL205</td>
<td>0.00</td>
<td>33</td>
<td>74_KL383</td>
<td>0.38</td>
<td>62</td>
<td>04_KL322</td>
<td>0.75</td>
</tr>
<tr>
<td>5</td>
<td>97_R208</td>
<td>0.03</td>
<td>34</td>
<td>42_KL507</td>
<td>0.40</td>
<td>63</td>
<td>64_KKH29</td>
<td>0.75</td>
</tr>
<tr>
<td>6</td>
<td>45_R194</td>
<td>0.04</td>
<td>35</td>
<td>65_KL224</td>
<td>0.42</td>
<td>64</td>
<td>62_KKH22</td>
<td>0.76</td>
</tr>
<tr>
<td>7</td>
<td>78_R211</td>
<td>0.04</td>
<td>36</td>
<td>94_R354</td>
<td>0.42</td>
<td>65</td>
<td>57_KL535</td>
<td>0.77</td>
</tr>
<tr>
<td>8</td>
<td>96_R202</td>
<td>0.07</td>
<td>37</td>
<td>37_R355</td>
<td>0.42</td>
<td>66</td>
<td>15_KKH20</td>
<td>0.77</td>
</tr>
<tr>
<td>9</td>
<td>20_KL274</td>
<td>0.09</td>
<td>38</td>
<td>50_KL360</td>
<td>0.44</td>
<td>67</td>
<td>85_KL485</td>
<td>0.79</td>
</tr>
<tr>
<td>10</td>
<td>70_KL320</td>
<td>0.09</td>
<td>39</td>
<td>47_R334</td>
<td>0.46</td>
<td>68</td>
<td>73_KL381</td>
<td>0.81</td>
</tr>
<tr>
<td>11</td>
<td>11_R280</td>
<td>0.09</td>
<td>40</td>
<td>79_KL421</td>
<td>0.49</td>
<td>69</td>
<td>34_R256</td>
<td>0.83</td>
</tr>
<tr>
<td>12</td>
<td>67_KL287</td>
<td>0.10</td>
<td>41</td>
<td>51_KL461</td>
<td>0.52</td>
<td>70</td>
<td>68_KL304</td>
<td>0.83</td>
</tr>
<tr>
<td>13</td>
<td>27_KL374</td>
<td>0.10</td>
<td>42</td>
<td>25_KL357</td>
<td>0.53</td>
<td>71</td>
<td>32_R233</td>
<td>0.85</td>
</tr>
<tr>
<td>14</td>
<td>82_KL454</td>
<td>0.11</td>
<td>43</td>
<td>98_KL387</td>
<td>0.55</td>
<td>72</td>
<td>44_KL541</td>
<td>0.87</td>
</tr>
<tr>
<td>15</td>
<td>41_KL441</td>
<td>0.12</td>
<td>44</td>
<td>66_KL247</td>
<td>0.56</td>
<td>73</td>
<td>80_KL423</td>
<td>0.88</td>
</tr>
<tr>
<td>16</td>
<td>17_KKH27</td>
<td>0.13</td>
<td>45</td>
<td>10_R257</td>
<td>0.57</td>
<td>74</td>
<td>12_R297</td>
<td>0.89</td>
</tr>
<tr>
<td>17</td>
<td>39_KL395</td>
<td>0.14</td>
<td>46</td>
<td>84_KL458</td>
<td>0.58</td>
<td>75</td>
<td>08_KL456</td>
<td>0.89</td>
</tr>
<tr>
<td>18</td>
<td>13_R410</td>
<td>0.14</td>
<td>47</td>
<td>26_KL369</td>
<td>0.58</td>
<td>76</td>
<td>72_KL377</td>
<td>0.92</td>
</tr>
<tr>
<td>19</td>
<td>35_R313</td>
<td>0.18</td>
<td>48</td>
<td>52_R252</td>
<td>0.59</td>
<td>77</td>
<td>01_KKH18</td>
<td>0.93</td>
</tr>
<tr>
<td>20</td>
<td>55_KL419</td>
<td>0.19</td>
<td>49</td>
<td>48_R339</td>
<td>0.60</td>
<td>78</td>
<td>58_KL543</td>
<td>0.94</td>
</tr>
<tr>
<td>21</td>
<td>23_KL321</td>
<td>0.22</td>
<td>50</td>
<td>14_KKH19</td>
<td>0.61</td>
<td>79</td>
<td>89_R245</td>
<td>0.95</td>
</tr>
<tr>
<td>22</td>
<td>43_KL536</td>
<td>0.23</td>
<td>51</td>
<td>60_KKH30</td>
<td>0.62</td>
<td>80</td>
<td>05_KL354</td>
<td>0.95</td>
</tr>
<tr>
<td>23</td>
<td>90_R253</td>
<td>0.26</td>
<td>52</td>
<td>40_KL430</td>
<td>0.64</td>
<td>81</td>
<td>87_KL522</td>
<td>0.96</td>
</tr>
<tr>
<td>24</td>
<td>29_KL439</td>
<td>0.27</td>
<td>53</td>
<td>75_KL385</td>
<td>0.64</td>
<td>82</td>
<td>88_KL544</td>
<td>0.96</td>
</tr>
<tr>
<td>25</td>
<td>49_R432</td>
<td>0.29</td>
<td>54</td>
<td>28_KL375</td>
<td>0.67</td>
<td>83</td>
<td>95_R431</td>
<td>0.96</td>
</tr>
<tr>
<td>26</td>
<td>61_KKH13</td>
<td>0.31</td>
<td>55</td>
<td>16_KKH21</td>
<td>0.67</td>
<td>84</td>
<td>78_KL412</td>
<td>0.98</td>
</tr>
<tr>
<td>27</td>
<td>99_KL416</td>
<td>0.33</td>
<td>56</td>
<td>93_R337</td>
<td>0.68</td>
<td>85</td>
<td>18_KKH28</td>
<td>0.99</td>
</tr>
<tr>
<td>28</td>
<td>83_KL457</td>
<td>0.34</td>
<td>57</td>
<td>92_R332</td>
<td>0.68</td>
<td>86</td>
<td>63_KKH25</td>
<td>1.01</td>
</tr>
<tr>
<td>29</td>
<td>38_KL218</td>
<td>0.35</td>
<td>58</td>
<td>24_KL328</td>
<td>0.68</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 5.8: ESR between the D0 and D15 samples. Relapses are highlighted with Underline. Extremely slow responders are highlighted in Italic.

<table>
<thead>
<tr>
<th>RANK</th>
<th>SAMPLE</th>
<th>ESR-D15</th>
<th>RANK</th>
<th>SAMPLE</th>
<th>ESR-D15</th>
<th>RANK</th>
<th>SAMPLE</th>
<th>ESR-D15</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>97_R208</td>
<td>-0.13</td>
<td>17</td>
<td>41_KL441</td>
<td>0.85</td>
<td>33</td>
<td>88_KL544</td>
<td>0.99</td>
</tr>
<tr>
<td>2</td>
<td>77_KL434</td>
<td>0.66</td>
<td>18</td>
<td>84_KL458</td>
<td>0.88</td>
<td>34</td>
<td>93_R337</td>
<td>1.00</td>
</tr>
<tr>
<td>3</td>
<td>96_R202</td>
<td>0.08</td>
<td>19</td>
<td>64_KKH29</td>
<td>0.88</td>
<td>35</td>
<td>80_KL423</td>
<td>1.00</td>
</tr>
<tr>
<td>4</td>
<td>56_KL464</td>
<td>0.25</td>
<td>20</td>
<td>99_KL416</td>
<td>0.89</td>
<td>36</td>
<td>10_R257</td>
<td>1.00</td>
</tr>
<tr>
<td>5</td>
<td>19_R281</td>
<td>0.77</td>
<td>21</td>
<td>52_R252</td>
<td>0.90</td>
<td>37</td>
<td>87_KL522</td>
<td>1.00</td>
</tr>
<tr>
<td>6</td>
<td>50_KL360</td>
<td>0.55</td>
<td>22</td>
<td>18_KKH28</td>
<td>0.93</td>
<td>38</td>
<td>63_KKH25</td>
<td>1.01</td>
</tr>
<tr>
<td>7</td>
<td>67_KL287</td>
<td>0.62</td>
<td>23</td>
<td>92_R332</td>
<td>0.93</td>
<td>39</td>
<td>95_R431</td>
<td>1.01</td>
</tr>
<tr>
<td>8</td>
<td>74_KL383</td>
<td>0.65</td>
<td>24</td>
<td>49_R432</td>
<td>0.94</td>
<td>40</td>
<td>78_KL412</td>
<td>1.03</td>
</tr>
<tr>
<td>9</td>
<td>61_KKH13</td>
<td>0.66</td>
<td>25</td>
<td>62_KKH22</td>
<td>0.94</td>
<td>41</td>
<td>20_KL274</td>
<td>1.03</td>
</tr>
<tr>
<td>10</td>
<td>13_R410</td>
<td>0.69</td>
<td>26</td>
<td>28_KL375</td>
<td>0.94</td>
<td>42</td>
<td>94_R354</td>
<td>1.04</td>
</tr>
<tr>
<td>11</td>
<td>27_KL374</td>
<td>0.71</td>
<td>27</td>
<td>37_R355</td>
<td>0.95</td>
<td>43</td>
<td>15_KKH20</td>
<td>1.04</td>
</tr>
<tr>
<td>12</td>
<td>06_KL378</td>
<td>0.71</td>
<td>28</td>
<td>48_R339</td>
<td>0.96</td>
<td>44</td>
<td>85_KL485</td>
<td>1.10</td>
</tr>
<tr>
<td>13</td>
<td>71_KL371</td>
<td>0.72</td>
<td>29</td>
<td>36_R343</td>
<td>0.96</td>
<td>45</td>
<td>73_KL381</td>
<td>1.10</td>
</tr>
<tr>
<td>14</td>
<td>98_KL387</td>
<td>0.78</td>
<td>30</td>
<td>05_KL354</td>
<td>0.98</td>
<td>46</td>
<td>01_KKH18</td>
<td>1.11</td>
</tr>
<tr>
<td>15</td>
<td>16_KKH21</td>
<td>0.78</td>
<td>31</td>
<td>14_KKH19</td>
<td>0.98</td>
<td>47</td>
<td>76_KL398</td>
<td>1.55</td>
</tr>
<tr>
<td>16</td>
<td>86_KL509</td>
<td>0.84</td>
<td>32</td>
<td>29_KL439</td>
<td>0.99</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 5.9: ESR between the D0 and D33 samples. Relapses are highlighted with **Underline.** Extremely slow responders are highlighted in *Italic.*

<table>
<thead>
<tr>
<th>RANK</th>
<th>SAMPLE</th>
<th>ESR-D33</th>
<th>RANK</th>
<th>SAMPLE</th>
<th>ESR-D33</th>
<th>RANK</th>
<th>SAMPLE</th>
<th>ESR-D33</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>97_R208</td>
<td><strong>-0.01</strong></td>
<td>20</td>
<td>82_KL454</td>
<td>0.99</td>
<td>39</td>
<td>66_KL247</td>
<td>1.03</td>
</tr>
<tr>
<td>2</td>
<td>96_R202</td>
<td>0.00</td>
<td>21</td>
<td>55_KL419</td>
<td>0.99</td>
<td>40</td>
<td>30_KL444</td>
<td>1.03</td>
</tr>
<tr>
<td>3</td>
<td>90_K274</td>
<td><strong>0.21</strong></td>
<td>22</td>
<td>20_KL274</td>
<td><strong>0.99</strong></td>
<td>41</td>
<td>85_KL485</td>
<td>1.03</td>
</tr>
<tr>
<td>4</td>
<td>76_KL398</td>
<td>0.33</td>
<td>23</td>
<td>39_KL395</td>
<td><strong>0.99</strong></td>
<td>42</td>
<td>43_KL536</td>
<td>1.04</td>
</tr>
<tr>
<td>5</td>
<td>51_KL461</td>
<td>0.59</td>
<td>24</td>
<td>25_KL357</td>
<td>0.99</td>
<td>43</td>
<td>01_KKH18</td>
<td>1.05</td>
</tr>
<tr>
<td>6</td>
<td>72_KL401</td>
<td>0.61</td>
<td>25</td>
<td>38_KL218</td>
<td>1.00</td>
<td>44</td>
<td>89_R245</td>
<td>1.05</td>
</tr>
<tr>
<td>7</td>
<td>40_KL430</td>
<td>0.88</td>
<td>26</td>
<td>11_R280</td>
<td><strong>1.00</strong></td>
<td>45</td>
<td>86_KL509</td>
<td>1.05</td>
</tr>
<tr>
<td>8</td>
<td>83_KL457</td>
<td>0.88</td>
<td>27</td>
<td>50_KL360</td>
<td>1.01</td>
<td>46</td>
<td>84_KL458</td>
<td>1.06</td>
</tr>
<tr>
<td>9</td>
<td>72_KL377</td>
<td>0.90</td>
<td>28</td>
<td>37_R355</td>
<td>1.01</td>
<td>47</td>
<td>78_KL412</td>
<td>1.06</td>
</tr>
<tr>
<td>10</td>
<td>08_KL456</td>
<td>0.90</td>
<td>29</td>
<td>18_KKH28</td>
<td>1.01</td>
<td>48</td>
<td>52_R252</td>
<td>1.06</td>
</tr>
<tr>
<td>11</td>
<td>42_KL507</td>
<td>0.90</td>
<td>30</td>
<td>17_KKH27</td>
<td>1.01</td>
<td>49</td>
<td>10_R257</td>
<td>1.06</td>
</tr>
<tr>
<td>12</td>
<td>27_KL374</td>
<td>0.93</td>
<td>31</td>
<td>29_KL439</td>
<td>1.01</td>
<td>50</td>
<td>14_KKH19</td>
<td>1.07</td>
</tr>
<tr>
<td>13</td>
<td>06_KL378</td>
<td>0.94</td>
<td>32</td>
<td>28_KL375</td>
<td>1.01</td>
<td>51</td>
<td>60_KKH30</td>
<td>1.07</td>
</tr>
<tr>
<td>14</td>
<td>75_KL385</td>
<td>0.94</td>
<td>33</td>
<td>88_KL544</td>
<td>1.01</td>
<td>52</td>
<td>63_KKH25</td>
<td>1.08</td>
</tr>
<tr>
<td>15</td>
<td>69_KL313</td>
<td>0.94</td>
<td>34</td>
<td>80_KL423</td>
<td>1.02</td>
<td>53</td>
<td>44_KL541</td>
<td>1.09</td>
</tr>
<tr>
<td>16</td>
<td>87_KL522</td>
<td>0.94</td>
<td>35</td>
<td>62_KKH22</td>
<td>1.02</td>
<td>54</td>
<td>67_KL287</td>
<td>1.09</td>
</tr>
<tr>
<td>17</td>
<td>90_R253</td>
<td><strong>0.96</strong></td>
<td>36</td>
<td>19_KL205</td>
<td><strong>1.02</strong></td>
<td>55</td>
<td>64_KKH29</td>
<td>1.10</td>
</tr>
<tr>
<td>18</td>
<td>71_KL371</td>
<td>0.97</td>
<td>37</td>
<td>49_R432</td>
<td>1.03</td>
<td>56</td>
<td>07_KL417</td>
<td><strong>1.14</strong></td>
</tr>
<tr>
<td>19</td>
<td>47_R334</td>
<td>0.97</td>
<td>38</td>
<td>95_R431</td>
<td>1.03</td>
<td>57</td>
<td>65_KL224</td>
<td>1.17</td>
</tr>
</tbody>
</table>
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,

\[
\begin{align*}
\text{sensitivity} &= \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}} \\
\text{specificity} &= \frac{\text{True Negative}}{\text{True Negative} + \text{False Positive}}
\end{align*}
\]

The p values refer to the areas under the curves, calculated by MedCalc 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 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 5.3: Receiver operating characteristics of D8 GSS distance in D8 response prediction. (a) Extremely slow response. (b) Slow response.
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.

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

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 $50 \times 10^9$ 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 \times 77.78\% = 8 of them, while giving false alarm on 76 \times (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 \times (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 \sim 78\% sensitivity, the corresponding specificity is \frac{76 - 13}{76} = 83\%, which is far higher than MRD-D33’s 54.12\% specificity at the same sensitivity level\footnote{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.}.

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 \frac{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\%.\footnote{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.

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

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.

<table>
<thead>
<tr>
<th>Rank</th>
<th>SAMPLE</th>
<th>ASD</th>
<th>Outcome</th>
<th>Rank</th>
<th>Sample</th>
<th>ESD</th>
<th>Outcome</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>R318-D5</td>
<td>0.28</td>
<td>R</td>
<td>1</td>
<td>R318-D33</td>
<td>-11.03</td>
<td>R</td>
</tr>
<tr>
<td>2</td>
<td>KL473-D32</td>
<td>3.04</td>
<td>R</td>
<td>2</td>
<td>R318-D5</td>
<td>0.04</td>
<td>R</td>
</tr>
<tr>
<td>3</td>
<td>KL343-D36</td>
<td>4.33</td>
<td>R</td>
<td>3</td>
<td>KL473-D32</td>
<td>2.83</td>
<td>R</td>
</tr>
<tr>
<td>4</td>
<td>KL448-D17</td>
<td>8.11</td>
<td>R</td>
<td>4</td>
<td>KL343-D36</td>
<td>3.34</td>
<td>R</td>
</tr>
<tr>
<td>5</td>
<td>KL505-D14</td>
<td>10.61</td>
<td>R</td>
<td>5</td>
<td>KL448-D17</td>
<td>6.99</td>
<td>R</td>
</tr>
<tr>
<td>6</td>
<td>R474-D33</td>
<td>11.52</td>
<td>R</td>
<td>6</td>
<td>KL505-D14</td>
<td>10.33</td>
<td>R</td>
</tr>
<tr>
<td>7</td>
<td>R318-D33</td>
<td>20.10</td>
<td>R</td>
<td>7</td>
<td>R474-D33</td>
<td>11.31</td>
<td>R</td>
</tr>
<tr>
<td>8</td>
<td>R474-D60</td>
<td>25.67</td>
<td>R</td>
<td>8</td>
<td>R474-D60</td>
<td>25.62</td>
<td>R</td>
</tr>
<tr>
<td>10</td>
<td>KL505-D45</td>
<td>31.07</td>
<td>R</td>
<td>10</td>
<td>KL505-D45</td>
<td>31.04</td>
<td>R</td>
</tr>
<tr>
<td>11</td>
<td>KKH14-D36</td>
<td>35.61</td>
<td>R</td>
<td>11</td>
<td>KKH14-D36</td>
<td>35.61</td>
<td>R</td>
</tr>
<tr>
<td>12</td>
<td>KL448-D51</td>
<td>39.71</td>
<td>R</td>
<td>12</td>
<td>KL448-D51</td>
<td>39.67</td>
<td>R</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Probe Set ID</th>
<th>Gene Symbol</th>
<th>Gene Title</th>
<th>p Value</th>
<th>Fold Change (T-ALL/Normal)</th>
</tr>
</thead>
<tbody>
<tr>
<td>201416_at</td>
<td>SOX4</td>
<td>SRY (sex determining region Y)-box 4</td>
<td>1.06E-98</td>
<td>18.63</td>
</tr>
<tr>
<td>201417_at</td>
<td>SOX4</td>
<td>SRY (sex determining region Y)-box 4</td>
<td>6.55E-94</td>
<td>9.97</td>
</tr>
<tr>
<td>201029_s_at</td>
<td>CD99</td>
<td>CD99 molecule</td>
<td>2.29E-89</td>
<td>5.63</td>
</tr>
<tr>
<td>211071_s_at</td>
<td>MLLT11</td>
<td>myeloid/lymphoid or mixed-lineage leukemia; translocated to, 11</td>
<td>1.02E-85</td>
<td>12.52</td>
</tr>
<tr>
<td>204529_s_at</td>
<td>TOX</td>
<td>thymocyte selection-associated high mobility group box</td>
<td>6.36E-77</td>
<td>14.40</td>
</tr>
<tr>
<td>Probe ID</td>
<td>Gene Symbol</td>
<td>Description</td>
<td>Log2FoldChange</td>
<td>p-value</td>
</tr>
<tr>
<td>------------------</td>
<td>-------------</td>
<td>---------------------------------------------------------------------------------------------</td>
<td>---------------</td>
<td>---------</td>
</tr>
<tr>
<td>204636_at</td>
<td>COL17A1</td>
<td>collagen, type XVII, alpha 1</td>
<td>1.18E-74</td>
<td>0.04</td>
</tr>
<tr>
<td>204639_at</td>
<td>ADA</td>
<td>adenosine deaminase</td>
<td>1.16E-73</td>
<td>9.60</td>
</tr>
<tr>
<td>201418_s_at</td>
<td>SOX4</td>
<td>SRY (sex determining region Y)-box 4</td>
<td>2.25E-73</td>
<td>18.69</td>
</tr>
<tr>
<td>216705_s_at</td>
<td>ADA</td>
<td>adenosine deaminase</td>
<td>4.80E-73</td>
<td>12.62</td>
</tr>
<tr>
<td>213668_s_at</td>
<td>SOX4</td>
<td>SRY (sex determining region Y)-box 4</td>
<td>5.38E-73</td>
<td>35.16</td>
</tr>
<tr>
<td>202760_s_at</td>
<td>PALM2-AKAP2</td>
<td>PALM2-AKAP2 readthrough transcript</td>
<td>1.66E-70</td>
<td>10.26</td>
</tr>
<tr>
<td>213539_at</td>
<td>CD3D</td>
<td>CD3d molecule, delta (CD3-TCR complex)</td>
<td>1.60E-67</td>
<td>10.64</td>
</tr>
<tr>
<td>206390_x_at</td>
<td>PF4</td>
<td>platelet factor 4</td>
<td>2.53E-65</td>
<td>0.03</td>
</tr>
<tr>
<td>202242_at</td>
<td>TSPAN7</td>
<td>tetranspanin 7</td>
<td>2.57E-65</td>
<td>21.94</td>
</tr>
<tr>
<td>201028_s_at</td>
<td>CD99</td>
<td>CD99 molecule</td>
<td>1.64E-63</td>
<td>6.24</td>
</tr>
<tr>
<td>202759_s_at</td>
<td>AKAP2</td>
<td>A kinase (PRKA) anchor protein 2</td>
<td>2.59E-63</td>
<td>6.68</td>
</tr>
<tr>
<td>214997_at</td>
<td>SCAI</td>
<td>suppressor of cancer cell invasion</td>
<td>3.21E-63</td>
<td>8.51</td>
</tr>
<tr>
<td>213437_at</td>
<td>RUFY3</td>
<td>RUN and FYVE domain containing 3</td>
<td>4.41E-63</td>
<td>10.12</td>
</tr>
<tr>
<td>204173_at</td>
<td>MYL6B</td>
<td>myosin, light chain 6B, alkali, smooth muscle and non-muscle</td>
<td>5.87E-62</td>
<td>3.95</td>
</tr>
<tr>
<td>210116_at</td>
<td>SH2D1A</td>
<td>SH2 domain protein 1A</td>
<td>1.34E-61</td>
<td>15.79</td>
</tr>
<tr>
<td>203787_at</td>
<td>SSBP2</td>
<td>single-stranded DNA binding protein 2</td>
<td>1.40E-60</td>
<td>5.46</td>
</tr>
<tr>
<td>218641_at</td>
<td>LOC65998</td>
<td>hypothetical protein LOC65998</td>
<td>2.08E-60</td>
<td>4.18</td>
</tr>
<tr>
<td>209473_at</td>
<td>ENTPD1</td>
<td>ectonucleoside triphosphate diphosphohydrolase 1</td>
<td>4.14E-60</td>
<td>0.14</td>
</tr>
<tr>
<td>200983_x_at</td>
<td>CD59</td>
<td>CD59 molecule, complement regulatory protein</td>
<td>4.18E-60</td>
<td>0.13</td>
</tr>
<tr>
<td>214298_x_at</td>
<td>SEPT6</td>
<td>septin 6</td>
<td>4.23E-60</td>
<td>5.48</td>
</tr>
<tr>
<td>34726_at</td>
<td>CACNB3</td>
<td>calcium channel, voltage-dependent, beta 3 subunit</td>
<td>4.64E-60</td>
<td>8.38</td>
</tr>
<tr>
<td>55872_at</td>
<td>ZNF512B</td>
<td>zinc finger protein 512B</td>
<td>7.10E-60</td>
<td>4.49</td>
</tr>
<tr>
<td>218865_at</td>
<td>MOSC1</td>
<td>MOCO sulphrase C-terminal domain containing 1</td>
<td>8.48E-59</td>
<td>0.06</td>
</tr>
<tr>
<td>213666_at</td>
<td>SEPT6</td>
<td>septin 6</td>
<td>1.24E-58</td>
<td>6.65</td>
</tr>
<tr>
<td>221203_s_at</td>
<td>YEATS2</td>
<td>YEATS domain containing 2</td>
<td>1.33E-58</td>
<td>2.63</td>
</tr>
<tr>
<td>204530_s_at</td>
<td>TOX</td>
<td>thymocyte selection-associated high mobility group box</td>
<td>1.72E-58</td>
<td>9.50</td>
</tr>
<tr>
<td>Gene ID</td>
<td>Symbol</td>
<td>Description</td>
<td>Log2 Fold Change</td>
<td>p-value</td>
</tr>
<tr>
<td>--------------</td>
<td>---------</td>
<td>-------------------------------------------------------</td>
<td>------------------</td>
<td>-------------</td>
</tr>
<tr>
<td>220359_s_at</td>
<td>ARPP-21</td>
<td>cyclic AMP-regulated phosphoprotein, 21 kD</td>
<td>1.89E-58</td>
<td>33.33</td>
</tr>
<tr>
<td>210638_s_at</td>
<td>FBXO9</td>
<td>F-box protein 9</td>
<td>1.03E-57</td>
<td>0.36</td>
</tr>
<tr>
<td>202804_at</td>
<td>ABCC1</td>
<td>ATP-binding cassette, sub-family C (CFTR/MRP), member 1</td>
<td>2.95E-57</td>
<td>3.00</td>
</tr>
<tr>
<td>215307_at</td>
<td>ZNF529</td>
<td>zinc finger protein 529</td>
<td>4.16E-57</td>
<td>3.85</td>
</tr>
<tr>
<td>219036_at</td>
<td>CEP70</td>
<td>centrosomal protein 70kDa</td>
<td>6.23E-57</td>
<td>6.33</td>
</tr>
<tr>
<td>221810_at</td>
<td>RAB15</td>
<td>RAB15, member RAS oncogene family</td>
<td>6.38E-57</td>
<td>10.57</td>
</tr>
<tr>
<td>206656_s_at</td>
<td>C20orf3</td>
<td>chromosome 20 open reading frame 3</td>
<td>1.08E-56</td>
<td>0.28</td>
</tr>
<tr>
<td>208792_s_at</td>
<td>CLU</td>
<td>Clusterin</td>
<td>3.06E-56</td>
<td>0.18</td>
</tr>
<tr>
<td>202789_at</td>
<td>PLCG1</td>
<td>phospholipase C, gamma 1</td>
<td>9.60E-56</td>
<td>5.58</td>
</tr>
<tr>
<td>208791_at</td>
<td>CLU</td>
<td>Clusterin</td>
<td>9.71E-56</td>
<td>0.06</td>
</tr>
<tr>
<td>202671_s_at</td>
<td>PDXK</td>
<td>pyridoxal (pyridoxine, vitamin B6) kinase</td>
<td>1.38E-55</td>
<td>0.23</td>
</tr>
<tr>
<td>213048_s_at</td>
<td>---</td>
<td>---</td>
<td>1.53E-55</td>
<td>2.00</td>
</tr>
<tr>
<td>222344_at</td>
<td>---</td>
<td>---</td>
<td>1.70E-55</td>
<td>15.07</td>
</tr>
<tr>
<td>201321_s_at</td>
<td>SMARCC2</td>
<td>SWI/SNF related, matrix associated, actin dependent regulator of chromatin, subfamily c, member 2</td>
<td>2.66E-55</td>
<td>2.16</td>
</tr>
<tr>
<td>200631_s_at</td>
<td>SET</td>
<td>SET nuclear oncogene</td>
<td>2.89E-55</td>
<td>2.21</td>
</tr>
<tr>
<td>213430_at</td>
<td>RUFY3</td>
<td>RUN and FYVE domain containing 3</td>
<td>6.75E-55</td>
<td>6.05</td>
</tr>
<tr>
<td>210140_at</td>
<td>CST7</td>
<td>cystatin F (leukocystatin)</td>
<td>7.91E-55</td>
<td>0.09</td>
</tr>
<tr>
<td>218081_at</td>
<td>C20orf27</td>
<td>chromosome 20 open reading frame 27</td>
<td>1.30E-54</td>
<td>0.25</td>
</tr>
<tr>
<td>218005_at</td>
<td>ZNF22</td>
<td>zinc finger protein 22 (KOX 15)</td>
<td>2.20E-54</td>
<td>2.99</td>
</tr>
</tbody>
</table>
Table A.2: Drug responsive genes of TEL-AML1 subtype.

<table>
<thead>
<tr>
<th>Probe Set ID</th>
<th>Gene Symbol</th>
<th>Gene Title</th>
<th>p Value</th>
<th>Fold Change (TEL-AML1/Normal)</th>
</tr>
</thead>
<tbody>
<tr>
<td>203373_at</td>
<td>SOCS2</td>
<td>suppressor of cytokine signaling 2</td>
<td>4.60E-91</td>
<td>21.67</td>
</tr>
<tr>
<td>222146_s_at</td>
<td>TCF4</td>
<td>transcription factor 4</td>
<td>2.45E-74</td>
<td>17.04</td>
</tr>
<tr>
<td>212012_at</td>
<td>PXDN</td>
<td>peroxidasin homolog (Drosophila)</td>
<td>3.07E-73</td>
<td>39.64</td>
</tr>
<tr>
<td>32625_at</td>
<td>NPR1</td>
<td>natriuretic peptide receptor A/guanylate cyclase A (atrionatriuretic peptide receptor A)</td>
<td>4.74E-73</td>
<td>46.04</td>
</tr>
<tr>
<td>212387_at</td>
<td>TCF4</td>
<td>transcription factor 4</td>
<td>6.45E-72</td>
<td>17.30</td>
</tr>
<tr>
<td>203372_s_at</td>
<td>SOCS2</td>
<td>suppressor of cytokine signaling 2</td>
<td>6.01E-71</td>
<td>27.21</td>
</tr>
<tr>
<td>212386_at</td>
<td>TCF4</td>
<td>transcription factor 4</td>
<td>3.26E-70</td>
<td>14.14</td>
</tr>
<tr>
<td>203355_s_at</td>
<td>PSD3</td>
<td>pleckstrin and Sec7 domain containing 3</td>
<td>9.44E-70</td>
<td>100.22</td>
</tr>
<tr>
<td>203787_at</td>
<td>SSBP2</td>
<td>single-stranded DNA binding protein 2</td>
<td>7.72E-69</td>
<td>12.72</td>
</tr>
<tr>
<td>219686_at</td>
<td>STK32B</td>
<td>serine/threonine kinase 32B</td>
<td>1.06E-67</td>
<td>78.02</td>
</tr>
<tr>
<td>212013_at</td>
<td>PXDN</td>
<td>peroxidasin homolog (Drosophila)</td>
<td>2.09E-67</td>
<td>47.59</td>
</tr>
<tr>
<td>202806_at</td>
<td>DBN1</td>
<td>drebrin 1</td>
<td>1.88E-66</td>
<td>15.05</td>
</tr>
<tr>
<td>214761_at</td>
<td>ZNF423</td>
<td>zinc finger protein 423</td>
<td>2.54E-66</td>
<td>128.71</td>
</tr>
<tr>
<td>213891_s_at</td>
<td>TCF4</td>
<td>transcription factor 4</td>
<td>9.26E-66</td>
<td>19.04</td>
</tr>
<tr>
<td>54037_at</td>
<td>HPS4</td>
<td>Hermansky-Pudlak syndrome 4</td>
<td>2.48E-64</td>
<td>10.65</td>
</tr>
<tr>
<td>203611_at</td>
<td>TERF2</td>
<td>telomeric repeat binding factor 2</td>
<td>3.50E-64</td>
<td>26.92</td>
</tr>
<tr>
<td>207030_s_at</td>
<td>CSRP2</td>
<td>cysteine and glycine-rich protein 2</td>
<td>8.63E-64</td>
<td>104.54</td>
</tr>
<tr>
<td>204257_at</td>
<td>FADS3</td>
<td>fatty acid desaturase 3</td>
<td>1.01E-63</td>
<td>8.75</td>
</tr>
<tr>
<td>203753_at</td>
<td>TCF4</td>
<td>transcription factor 4</td>
<td>1.38E-63</td>
<td>12.47</td>
</tr>
<tr>
<td>208794_s_at</td>
<td>SMARCA4</td>
<td>SWI/SNF related, matrix associated, actin dependent regulator of chromatin, subfamily a, member 4</td>
<td>2.42E-63</td>
<td>5.07</td>
</tr>
<tr>
<td>218966_at</td>
<td>MYO5C</td>
<td>myosin VC</td>
<td>1.54E-62</td>
<td>11.96</td>
</tr>
<tr>
<td>Probe ID</td>
<td>Gene Symbol</td>
<td>Description</td>
<td>log2 fold change</td>
<td>p-value</td>
</tr>
<tr>
<td>-------------</td>
<td>-------------</td>
<td>------------------------------------------------------------------------------</td>
<td>-----------------</td>
<td>----------</td>
</tr>
<tr>
<td>212385_at</td>
<td>TCF4</td>
<td>transcription factor 4</td>
<td>1.72E-62</td>
<td>18.51</td>
</tr>
<tr>
<td>214728_x_at</td>
<td>SMARCA4</td>
<td>SWI/SNF related, matrix associated, actin dependent regulator of chromatin, subfamily a, member 4</td>
<td>1.03E-61</td>
<td>4.14</td>
</tr>
<tr>
<td>219938_s_at</td>
<td>PSTPIP2</td>
<td>proline-serine-threonine phosphatase interacting protein 2</td>
<td>2.11E-61</td>
<td>0.10</td>
</tr>
<tr>
<td>208056_s_at</td>
<td>CBFA2T3</td>
<td>core-binding factor, runt domain, alpha subunit 2; translocated to, 3</td>
<td>2.94E-61</td>
<td>12.31</td>
</tr>
<tr>
<td>218613_at</td>
<td>PSD3</td>
<td>pleckstrin and Sec7 domain containing 3</td>
<td>3.01E-61</td>
<td>42.49</td>
</tr>
<tr>
<td>213720_s_at</td>
<td>SMARCA4</td>
<td>SWI/SNF related, matrix associated, actin dependent regulator of chromatin, subfamily a, member 4</td>
<td>1.74E-60</td>
<td>5.06</td>
</tr>
<tr>
<td>218217_at</td>
<td>SCPEP1</td>
<td>serine carboxypeptidase 1</td>
<td>2.84E-60</td>
<td>0.20</td>
</tr>
<tr>
<td>211071_s_at</td>
<td>MLLT11</td>
<td>myeloid/lymphoid or mixed-lineage leukemia (trithorax homolog, Drosophila); translocated to, 11</td>
<td>6.64E-60</td>
<td>7.73</td>
</tr>
<tr>
<td>201015_s_at</td>
<td>JUP</td>
<td>junction plakoglobin</td>
<td>1.20E-59</td>
<td>31.23</td>
</tr>
<tr>
<td>206591_at</td>
<td>RAG1</td>
<td>recombination activating gene 1</td>
<td>2.03E-59</td>
<td>149.73</td>
</tr>
<tr>
<td>209153_s_at</td>
<td>TCF3</td>
<td>transcription factor 3 (E2A immunoglobulin enhancer binding factors E12/E47)</td>
<td>2.96E-59</td>
<td>6.60</td>
</tr>
<tr>
<td>219753_at</td>
<td>STAG3</td>
<td>stromal antigen 3</td>
<td>7.25E-59</td>
<td>28.24</td>
</tr>
<tr>
<td>209514_s_at</td>
<td>RAB27A</td>
<td>RAB27A, member RAS oncogene family</td>
<td>1.32E-58</td>
<td>0.13</td>
</tr>
<tr>
<td>212812_at</td>
<td>---</td>
<td>---</td>
<td>1.59E-58</td>
<td>6.74</td>
</tr>
<tr>
<td>210829_s_at</td>
<td>SSBP2</td>
<td>single-stranded DNA binding protein 2</td>
<td>2.30E-58</td>
<td>11.00</td>
</tr>
<tr>
<td>203910_at</td>
<td>ARHGAP29</td>
<td>Rho GTPase activating protein 2</td>
<td>3.32E-58</td>
<td>40.35</td>
</tr>
<tr>
<td>212382_at</td>
<td>TCF4</td>
<td>transcription factor 4</td>
<td>3.36E-58</td>
<td>18.39</td>
</tr>
<tr>
<td>210094_s_at</td>
<td>PARD3</td>
<td>par-3 partitioning defective 3 homolog (C. elegans)</td>
<td>6.05E-58</td>
<td>18.34</td>
</tr>
<tr>
<td>209035_at</td>
<td>MDK</td>
<td>midkine (neurite growth-promoting factor 2)</td>
<td>1.09E-57</td>
<td>35.46</td>
</tr>
<tr>
<td>202519_at</td>
<td>MLXIP</td>
<td>MLX interacting protein</td>
<td>2.04E-57</td>
<td>5.61</td>
</tr>
<tr>
<td>206398_s_at</td>
<td>CD19</td>
<td>CD19 molecule</td>
<td>2.52E-57</td>
<td>19.93</td>
</tr>
<tr>
<td>218988_at</td>
<td>SLC35E3</td>
<td>solute carrier family 35, member E3</td>
<td>4.02E-57</td>
<td>17.91</td>
</tr>
</tbody>
</table>
## Appendix A  Drug Responsive Gene

<table>
<thead>
<tr>
<th>UNIQID</th>
<th>Gene Symbol</th>
<th>Gene Title</th>
<th>p Value</th>
<th>Fold Change (Hyperdiploid&gt;50/Normal)</th>
</tr>
</thead>
<tbody>
<tr>
<td>202039_at</td>
<td>MYO18A</td>
<td>myosin XVIIIA</td>
<td>4.54E-57</td>
<td>7.10</td>
</tr>
<tr>
<td>211026_s_at</td>
<td>MGLL</td>
<td>monoglyceride lipase</td>
<td>5.82E-57</td>
<td>0.16</td>
</tr>
<tr>
<td>209199_s_at</td>
<td>MEF2C</td>
<td>myocyte enhancer factor 2C</td>
<td>7.34E-57</td>
<td>15.57</td>
</tr>
<tr>
<td>204849_at</td>
<td>TCFL5</td>
<td>transcription factor-like 5 (basic helix-loop-helix)</td>
<td>9.59E-57</td>
<td>20.16</td>
</tr>
<tr>
<td>213702_x_at</td>
<td>ASAH1</td>
<td>N-acylsphingosine amidohydrolase (acid ceramidase) 1</td>
<td>2.56E-56</td>
<td>0.24</td>
</tr>
<tr>
<td>209583_s_at</td>
<td>CD200</td>
<td>CD200 molecule</td>
<td>3.98E-56</td>
<td>18.12</td>
</tr>
<tr>
<td>210980_s_at</td>
<td>ASAH1</td>
<td>N-acylsphingosine amidohydrolase (acid ceramidase) 1</td>
<td>5.70E-56</td>
<td>0.20</td>
</tr>
</tbody>
</table>

Table A.3: Drug responsive genes of Hyperdiploid>50 subtype.
### APPENDIX A  DRUG RESPONSIVE GENE

<table>
<thead>
<tr>
<th>Affymetrix ProbeID</th>
<th>Gene Symbol</th>
<th>Description</th>
<th>Log2 Fold</th>
<th>p Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>212385_at</td>
<td>TCF4</td>
<td>transcription factor 4</td>
<td>8.54E-56</td>
<td>20.70</td>
</tr>
<tr>
<td>213891_s_at</td>
<td>TCF4</td>
<td>transcription factor 4</td>
<td>2.15E-55</td>
<td>22.00</td>
</tr>
<tr>
<td>208370_s_at</td>
<td>RCAN1</td>
<td>regulator of calcineurin 1</td>
<td>8.77E-55</td>
<td>12.26</td>
</tr>
<tr>
<td>201540_at</td>
<td>FHL1</td>
<td>four and a half LIM domains 1</td>
<td>1.24E-54</td>
<td>10.48</td>
</tr>
<tr>
<td>212013_at</td>
<td>PXDN</td>
<td>peroxidasin homolog (Drosophila)</td>
<td>1.28E-54</td>
<td>48.40</td>
</tr>
<tr>
<td>211026_s_at</td>
<td>MGLL</td>
<td>monoglyceride lipase</td>
<td>2.36E-54</td>
<td>0.17</td>
</tr>
<tr>
<td>204122_at</td>
<td>TYROBP</td>
<td>TYRO protein tyrosine kinase binding protein</td>
<td>7.91E-54</td>
<td>0.07</td>
</tr>
<tr>
<td>209365_s_at</td>
<td>ECM1</td>
<td>extracellular matrix protein 1</td>
<td>2.83E-53</td>
<td>31.43</td>
</tr>
<tr>
<td>203753_at</td>
<td>TCF4</td>
<td>transcription factor 4</td>
<td>1.10E-52</td>
<td>12.77</td>
</tr>
<tr>
<td>202806_at</td>
<td>DBN1</td>
<td>drebrin 1</td>
<td>3.48E-52</td>
<td>9.26</td>
</tr>
<tr>
<td>211275_s_at</td>
<td>GYG1</td>
<td>glycogenin 1</td>
<td>4.98E-52</td>
<td>0.14</td>
</tr>
<tr>
<td>206001_at</td>
<td>NPY</td>
<td>neuropeptide Y</td>
<td>5.06E-52</td>
<td>86.18</td>
</tr>
<tr>
<td>221766_s_at</td>
<td>FAM46A</td>
<td>family with sequence similarity 46, member A</td>
<td>8.25E-52</td>
<td>0.04</td>
</tr>
<tr>
<td>202908_at</td>
<td>WFS1</td>
<td>Wolfram syndrome 1 (wolframin)</td>
<td>9.64E-52</td>
<td>16.41</td>
</tr>
<tr>
<td>205786_s_at</td>
<td>ITGAM</td>
<td>integrin, alpha M (complement component 3 receptor 3 subunit)</td>
<td>2.31E-51</td>
<td>0.07</td>
</tr>
<tr>
<td>207030_s_at</td>
<td>CSRP2</td>
<td>cysteine and glycine-rich protein 2</td>
<td>2.54E-51</td>
<td>74.97</td>
</tr>
<tr>
<td>204232_at</td>
<td>FCER1G</td>
<td>Fc fragment of IgE, high affinity I, receptor for; gamma polypeptide</td>
<td>3.48E-51</td>
<td>0.05</td>
</tr>
<tr>
<td>205237_at</td>
<td>FCN1</td>
<td>ficolin (collagen/fibrinogen domain containing) 1</td>
<td>8.98E-51</td>
<td>0.02</td>
</tr>
<tr>
<td>202598_at</td>
<td>S100A13</td>
<td>S100 calcium binding protein A13</td>
<td>1.84E-50</td>
<td>7.66</td>
</tr>
<tr>
<td>219694_at</td>
<td>FAM105A</td>
<td>family with sequence similarity 105, member A</td>
<td>2.37E-50</td>
<td>0.09</td>
</tr>
<tr>
<td>217728_at</td>
<td>S100A6</td>
<td>S100 calcium binding protein A6</td>
<td>3.85E-50</td>
<td>0.12</td>
</tr>
<tr>
<td>204620_s_at</td>
<td>VCAN</td>
<td>Versican</td>
<td>4.57E-50</td>
<td>0.03</td>
</tr>
<tr>
<td>219686_at</td>
<td>STK32B</td>
<td>serine/threonine kinase 32B</td>
<td>5.47E-50</td>
<td>52.11</td>
</tr>
</tbody>
</table>
## APPENDIX A  DRUG RESPONSIVE GENE

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

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

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

<table>
<thead>
<tr>
<th>Gene Accession</th>
<th>Gene Symbol</th>
<th>Description</th>
<th>Log2 Fold Change</th>
<th>p-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>202806_at</td>
<td>DBN1</td>
<td>drebrin 1</td>
<td>2.00E-45</td>
<td>13.74</td>
</tr>
<tr>
<td>219694_at</td>
<td>FAM105A</td>
<td>family with sequence similarity 105, member A</td>
<td>2.76E-45</td>
<td>0.07</td>
</tr>
<tr>
<td>212387_at</td>
<td>TCF4</td>
<td>transcription factor 4</td>
<td>4.20E-45</td>
<td>15.81</td>
</tr>
<tr>
<td>203922_s_at</td>
<td>CYBB</td>
<td>cytochrome b-245, beta polypeptide</td>
<td>6.17E-45</td>
<td>0.08</td>
</tr>
<tr>
<td>201792_at</td>
<td>AEBP1</td>
<td>AE binding protein 1</td>
<td>6.73E-45</td>
<td>28.37</td>
</tr>
<tr>
<td>201061_s_at</td>
<td>STOM</td>
<td>stomatin</td>
<td>8.16E-45</td>
<td>0.10</td>
</tr>
<tr>
<td>201060_x_at</td>
<td>STOM</td>
<td>stomatin</td>
<td>1.13E-44</td>
<td>0.09</td>
</tr>
<tr>
<td>214761_at</td>
<td>ZNF423</td>
<td>zinc finger protein 423</td>
<td>1.45E-44</td>
<td>66.70</td>
</tr>
<tr>
<td>204173_at</td>
<td>MYL6B</td>
<td>myosin, light chain 6B, alkali, smooth muscle and non-muscle</td>
<td>1.62E-44</td>
<td>3.73</td>
</tr>
<tr>
<td>202788_at</td>
<td>MAPKAPK3</td>
<td>mitogen-activated protein kinase-activated protein kinase 3</td>
<td>4.10E-44</td>
<td>0.19</td>
</tr>
<tr>
<td>210829__s__at</td>
<td>SSBP2</td>
<td>single-stranded DNA binding protein 2</td>
<td>5.15E-44</td>
<td>9.81</td>
</tr>
<tr>
<td>212386_at</td>
<td>TCF4</td>
<td>transcription factor 4</td>
<td>5.86E-44</td>
<td>12.84</td>
</tr>
<tr>
<td>213891_s_at</td>
<td>TCF4</td>
<td>transcription factor 4</td>
<td>8.12E-44</td>
<td>17.33</td>
</tr>
<tr>
<td>212385_at</td>
<td>TCF4</td>
<td>transcription factor 4</td>
<td>4.90E-43</td>
<td>17.45</td>
</tr>
<tr>
<td>201416_at</td>
<td>SOX4</td>
<td>SRY (sex determining region Y)-box 4</td>
<td>5.79E-43</td>
<td>17.80</td>
</tr>
<tr>
<td>208370__s__at</td>
<td>RCAN1</td>
<td>regulator of calcineurin 1</td>
<td>8.54E-43</td>
<td>8.11</td>
</tr>
<tr>
<td>201506_at</td>
<td>TGFBI</td>
<td>transforming growth factor, beta-induced, 68kDa</td>
<td>1.29E-42</td>
<td>0.04</td>
</tr>
<tr>
<td>215806__x__at</td>
<td>TARP /// TRGC2</td>
<td>TCR gamma alternate reading frame protein /// T cell receptor gamma constant 2</td>
<td>1.71E-42</td>
<td>0.12</td>
</tr>
<tr>
<td>211987_at</td>
<td>TOP2B</td>
<td>topoisomerase (DNA) II beta 180kDa</td>
<td>1.77E-42</td>
<td>4.27</td>
</tr>
<tr>
<td>216920_s_at</td>
<td>TARP /// TRGC2</td>
<td>TCR gamma alternate reading frame protein /// T cell receptor gamma constant 2</td>
<td>3.18E-42</td>
<td>0.11</td>
</tr>
<tr>
<td>212599_at</td>
<td>AUTS2</td>
<td>autism susceptibility candidate 2</td>
<td>5.12E-42</td>
<td>16.11</td>
</tr>
<tr>
<td>212197_x_at</td>
<td>MPRIP</td>
<td>myosin phosphatase Rho interacting protein</td>
<td>6.12E-42</td>
<td>3.72</td>
</tr>
<tr>
<td>204949_at</td>
<td>ICAM3</td>
<td>intercellular adhesion molecule 3</td>
<td>8.04E-42</td>
<td>0.14</td>
</tr>
<tr>
<td>217763_s_at</td>
<td>RAB31</td>
<td>RAB31, member RAS oncogene family</td>
<td>8.30E-42</td>
<td>0.04</td>
</tr>
</tbody>
</table>
Table A.5: Drug responsive genes of BCR-ABL subtype.

<table>
<thead>
<tr>
<th>Probe Set ID</th>
<th>Gene Symbol</th>
<th>Gene Title</th>
<th>p Value</th>
<th>Fold Change (BCR-ABL/Normal)</th>
</tr>
</thead>
<tbody>
<tr>
<td>203373_at</td>
<td>SOCS2</td>
<td>suppressor of cytokine signaling 2</td>
<td>1.98E-130</td>
<td>22.03</td>
</tr>
<tr>
<td>203372_s_at</td>
<td>SOCS2</td>
<td>suppressor of cytokine signaling 2</td>
<td>2.59E-107</td>
<td>37.15</td>
</tr>
<tr>
<td>212012_at</td>
<td>PXDN</td>
<td>peroxidasin homolog (Drosophila)</td>
<td>1.06E-92</td>
<td>31.34</td>
</tr>
<tr>
<td>203355_s_at</td>
<td>PSD3</td>
<td>pleckstrin and Sec7 domain containing 3</td>
<td>1.22E-91</td>
<td>63.31</td>
</tr>
<tr>
<td>212013_at</td>
<td>PXDN</td>
<td>peroxidasin homolog (Drosophila)</td>
<td>2.73E-91</td>
<td>43.46</td>
</tr>
<tr>
<td>218966_at</td>
<td>MYO5C</td>
<td>myosin VC</td>
<td>4.12E-89</td>
<td>12.88</td>
</tr>
<tr>
<td>219686_at</td>
<td>STK32B</td>
<td>serine/threonine kinase 32B</td>
<td>8.71E-89</td>
<td>43.78</td>
</tr>
<tr>
<td>207030_s_at</td>
<td>CSRP2</td>
<td>cysteine and glycine-rich protein 2</td>
<td>1.14E-87</td>
<td>63.60</td>
</tr>
<tr>
<td>201540_at</td>
<td>FHL1</td>
<td>four and a half LIM domains 1</td>
<td>1.13E-84</td>
<td>9.77</td>
</tr>
<tr>
<td>201029_s_at</td>
<td>CD99</td>
<td>CD99 molecule</td>
<td>6.67E-83</td>
<td>5.56</td>
</tr>
<tr>
<td>209365_s_at</td>
<td>ECM1</td>
<td>extracellular matrix protein 1</td>
<td>1.16E-78</td>
<td>41.47</td>
</tr>
<tr>
<td>32625_at</td>
<td>NPR1</td>
<td>natriuretic peptide receptor A/guanylate cyclase A (atriothriuretic peptide receptor A)</td>
<td>5.30E-78</td>
<td>34.93</td>
</tr>
<tr>
<td>Gene ID</td>
<td>Gene Symbol</td>
<td>Description</td>
<td>Log2 Fold Change</td>
<td>P-Value</td>
</tr>
<tr>
<td>--------------</td>
<td>-------------</td>
<td>-----------------------------------------------------------------------------</td>
<td>------------------</td>
<td>----------</td>
</tr>
<tr>
<td>206398_s_at</td>
<td>CD19</td>
<td>CD19 molecule</td>
<td>1.78E-77</td>
<td>19.73</td>
</tr>
<tr>
<td>218613_at</td>
<td>PSD3</td>
<td>pleckstrin and Sec7 domain containing 3</td>
<td>7.17E-77</td>
<td>28.97</td>
</tr>
<tr>
<td>214761_at</td>
<td>ZNF423</td>
<td>zinc finger protein 423</td>
<td>1.67E-74</td>
<td>58.41</td>
</tr>
<tr>
<td>222146_s_at</td>
<td>TCF4</td>
<td>transcription factor 4</td>
<td>9.23E-74</td>
<td>8.28</td>
</tr>
<tr>
<td>201015_s_at</td>
<td>JUP</td>
<td>junction plakoglobin</td>
<td>3.41E-73</td>
<td>16.80</td>
</tr>
<tr>
<td>210487_at</td>
<td>DNTT</td>
<td>deoxynucleotidyltransferase, terminal</td>
<td>7.36E-73</td>
<td>46.00</td>
</tr>
<tr>
<td>211126_s_at</td>
<td>CSRP2</td>
<td>cysteine and glycine-rich protein 2</td>
<td>1.07E-71</td>
<td>22.32</td>
</tr>
<tr>
<td>203787_at</td>
<td>SSBP2</td>
<td>single-stranded DNA binding protein 2</td>
<td>1.52E-71</td>
<td>7.26</td>
</tr>
<tr>
<td>212387_at</td>
<td>TCF4</td>
<td>transcription factor 4</td>
<td>3.26E-70</td>
<td>8.71</td>
</tr>
<tr>
<td>209576_at</td>
<td>GNA11</td>
<td>guanine nucleotide binding protein (G protein), alpha inhibiting activity polypeptide 1</td>
<td>4.30E-69</td>
<td>11.56</td>
</tr>
<tr>
<td>213891_s_at</td>
<td>TCF4</td>
<td>transcription factor 4</td>
<td>5.75E-69</td>
<td>9.79</td>
</tr>
<tr>
<td>212386_at</td>
<td>TCF4</td>
<td>transcription factor 4</td>
<td>9.79E-69</td>
<td>7.49</td>
</tr>
<tr>
<td>204030_s_at</td>
<td>SCHIP1</td>
<td>schwannomin interacting protein 1</td>
<td>6.41E-68</td>
<td>18.37</td>
</tr>
<tr>
<td>212385_at</td>
<td>TCF4</td>
<td>transcription factor 4</td>
<td>1.38E-67</td>
<td>9.09</td>
</tr>
<tr>
<td>201028_s_at</td>
<td>CD99</td>
<td>CD99 molecule</td>
<td>3.05E-67</td>
<td>6.48</td>
</tr>
<tr>
<td>210829_s_at</td>
<td>SSBP2</td>
<td>single-stranded DNA binding protein 2</td>
<td>9.89E-67</td>
<td>7.43</td>
</tr>
<tr>
<td>202123_s_at</td>
<td>ABL1</td>
<td>c-abl oncogene 1, receptor tyrosine kinase</td>
<td>4.08E-66</td>
<td>3.82</td>
</tr>
<tr>
<td>204636_at</td>
<td>COL17A1</td>
<td>collagen, type XVII, alpha 1</td>
<td>4.95E-65</td>
<td>0.04</td>
</tr>
<tr>
<td>202945_at</td>
<td>FPGS</td>
<td>folylpolyglutamate synthase</td>
<td>5.12E-65</td>
<td>4.97</td>
</tr>
<tr>
<td>209679_s_at</td>
<td>SMAGP</td>
<td>small trans-membrane and glycosylated protein</td>
<td>8.96E-65</td>
<td>11.37</td>
</tr>
<tr>
<td>207655_s_at</td>
<td>BLNK</td>
<td>B-cell linker</td>
<td>2.78E-64</td>
<td>14.78</td>
</tr>
<tr>
<td>203753_at</td>
<td>TCF4</td>
<td>transcription factor 4</td>
<td>2.03E-63</td>
<td>6.79</td>
</tr>
<tr>
<td>205983_at</td>
<td>DPEP1</td>
<td>dipепtidase 1 (renal)</td>
<td>6.02E-63</td>
<td>51.95</td>
</tr>
<tr>
<td>212675_s_at</td>
<td>CEP68</td>
<td>centrosomal protein 68kDa</td>
<td>1.65E-62</td>
<td>5.97</td>
</tr>
<tr>
<td>209199_s_at</td>
<td>MEF2C</td>
<td>myocyte enhancer factor 2C</td>
<td>3.43E-62</td>
<td>8.59</td>
</tr>
</tbody>
</table>
Table A.6: Drug responsive genes of MLL subtype.

<table>
<thead>
<tr>
<th>Probe Set ID</th>
<th>Gene Symbol</th>
<th>Gene Title</th>
<th>p Value</th>
<th>Fold Change (MLL/Normal)</th>
</tr>
</thead>
<tbody>
<tr>
<td>203373_at</td>
<td>SOCS2</td>
<td>suppressor of cytokine signaling 2</td>
<td>4.33E-79</td>
<td>21.74</td>
</tr>
<tr>
<td>207030_s_at</td>
<td>CSRP2</td>
<td>cysteine and glycine-rich protein 2</td>
<td>1.19E-78</td>
<td>160.22</td>
</tr>
<tr>
<td>211066_x_at</td>
<td>PCDHGA1</td>
<td>protocadherin gamma subfamily A, 1</td>
<td>5.33E-68</td>
<td>20.08</td>
</tr>
<tr>
<td>203372_s_at</td>
<td>SOCS2</td>
<td>suppressor of cytokine signaling 2</td>
<td>6.82E-68</td>
<td>30.70</td>
</tr>
<tr>
<td>211126_s_at</td>
<td>CSRP2</td>
<td>cysteine and glycine-rich protein 2</td>
<td>7.61E-66</td>
<td>51.71</td>
</tr>
<tr>
<td>Gene Symbol</td>
<td>Description</td>
<td>Value</td>
<td>p Value</td>
<td></td>
</tr>
<tr>
<td>-------------------</td>
<td>-----------------------------------------------------------------------------</td>
<td>---------</td>
<td>---------</td>
<td></td>
</tr>
<tr>
<td>217963_s_at</td>
<td>NGFRAP1 nerve growth factor receptor (TNFRSF16) associated protein 1</td>
<td>1.39E-64</td>
<td>0.05</td>
<td></td>
</tr>
<tr>
<td>206398_s_at</td>
<td>CD19 CD19 molecule</td>
<td>9.28E-62</td>
<td>18.07</td>
<td></td>
</tr>
<tr>
<td>218865_at</td>
<td>MOSC1 MOCO sulphurase C-terminal domain containing 1</td>
<td>2.00E-61</td>
<td>0.04</td>
<td></td>
</tr>
<tr>
<td>206674_at</td>
<td>FLT3 fms-related tyrosine kinase 3</td>
<td>3.44E-61</td>
<td>22.89</td>
<td></td>
</tr>
<tr>
<td>209170_s_at</td>
<td>GPM6B glycoprotein M6B</td>
<td>6.41E-61</td>
<td>33.73</td>
<td></td>
</tr>
<tr>
<td>201874_at</td>
<td>MPZL1 myelin protein zero-like 1</td>
<td>9.04E-61</td>
<td>4.66</td>
<td></td>
</tr>
<tr>
<td>36553_at</td>
<td>ASMTL acetylserotonin O-methyltransferase-like</td>
<td>1.85E-60</td>
<td>4.63</td>
<td></td>
</tr>
<tr>
<td>209079_x_at</td>
<td>PCDHGA1 protocadherin gamma subfamily A, 1</td>
<td>3.90E-60</td>
<td>15.70</td>
<td></td>
</tr>
<tr>
<td>205717_x_at</td>
<td>PCDHGA1 protocadherin gamma subfamily A, 1</td>
<td>4.86E-60</td>
<td>19.07</td>
<td></td>
</tr>
<tr>
<td>210638_s_at</td>
<td>FBXO9 F-box protein 9</td>
<td>7.15E-60</td>
<td>0.32</td>
<td></td>
</tr>
<tr>
<td>201416_at</td>
<td>SOX4 SRY (sex determining region Y)-box 4</td>
<td>1.36E-59</td>
<td>14.92</td>
<td></td>
</tr>
<tr>
<td>208949_s_at</td>
<td>LGALS3 lectin, galactoside-binding, soluble, 3</td>
<td>2.09E-59</td>
<td>0.06</td>
<td></td>
</tr>
<tr>
<td>209167_at</td>
<td>GPM6B glycoprotein M6B</td>
<td>3.00E-59</td>
<td>27.85</td>
<td></td>
</tr>
<tr>
<td>204636_at</td>
<td>COL17A1 collagen, type XVII, alpha 1</td>
<td>3.87E-59</td>
<td>0.04</td>
<td></td>
</tr>
<tr>
<td>204214_s_at</td>
<td>RAB32 RAB32, member RAS oncogene family</td>
<td>4.45E-59</td>
<td>0.08</td>
<td></td>
</tr>
<tr>
<td>201060_x_at</td>
<td>STOM stomatin</td>
<td>1.08E-58</td>
<td>0.09</td>
<td></td>
</tr>
<tr>
<td>211178_s_at</td>
<td>PSTPIP1 proline-serine-threonine phosphatase interacting protein 1</td>
<td>1.23E-58</td>
<td>0.15</td>
<td></td>
</tr>
<tr>
<td>204069_at</td>
<td>MEIS1 Meis homeobox 1</td>
<td>1.57E-58</td>
<td>20.82</td>
<td></td>
</tr>
<tr>
<td>200983_x_at</td>
<td>CD59 CD59 molecule, complement regulatory protein</td>
<td>1.85E-58</td>
<td>0.09</td>
<td></td>
</tr>
<tr>
<td>221485_at</td>
<td>B4GALT5 UDP-Gal:betaGlcNAc beta 1,4-galactosyltransferase, polypeptide 5</td>
<td>3.69E-58</td>
<td>0.10</td>
<td></td>
</tr>
<tr>
<td>209168_at</td>
<td>GPM6B glycoprotein M6B</td>
<td>4.72E-58</td>
<td>12.15</td>
<td></td>
</tr>
<tr>
<td>215925_s_at</td>
<td>CD72 CD72 molecule</td>
<td>7.21E-58</td>
<td>30.98</td>
<td></td>
</tr>
<tr>
<td>204173_at</td>
<td>MYL6B myosin, light chain 6B, alkali, smooth muscle and non-muscle</td>
<td>8.64E-58</td>
<td>3.84</td>
<td></td>
</tr>
<tr>
<td>Probe ID</td>
<td>Description</td>
<td>Value1</td>
<td>Value2</td>
<td></td>
</tr>
<tr>
<td>--------------</td>
<td>--------------------------------------</td>
<td>--------</td>
<td>----------</td>
<td></td>
</tr>
<tr>
<td>209514_s_at</td>
<td>RAB27A</td>
<td>2.62E-57</td>
<td>0.13</td>
<td></td>
</tr>
<tr>
<td>202945_at</td>
<td>RAB27A</td>
<td>3.41E-57</td>
<td>4.78</td>
<td></td>
</tr>
<tr>
<td>215836_s_at</td>
<td>protocadherin gamma subfamily A, 1</td>
<td>1.27E-56</td>
<td>26.89</td>
<td></td>
</tr>
<tr>
<td>201417_at</td>
<td>SRY (sex determining region Y)-box 4</td>
<td>1.92E-56</td>
<td>7.02</td>
<td></td>
</tr>
<tr>
<td>210951_x_at</td>
<td>RAB27A</td>
<td>2.57E-56</td>
<td>0.12</td>
<td></td>
</tr>
<tr>
<td>202332_at</td>
<td>casein kinase 1, epsilon</td>
<td>3.80E-56</td>
<td>3.84</td>
<td></td>
</tr>
<tr>
<td>209199_s_at</td>
<td>myocyte enhancer factor 2C</td>
<td>4.29E-56</td>
<td>10.96</td>
<td></td>
</tr>
<tr>
<td>203355_s_at</td>
<td>pleckstrin and Sec7 domain containing 3</td>
<td>9.39E-56</td>
<td>24.44</td>
<td></td>
</tr>
<tr>
<td>209949_at</td>
<td>neutrophil cytosolic factor 2</td>
<td>9.55E-56</td>
<td>0.07</td>
<td></td>
</tr>
<tr>
<td>208302_at</td>
<td>histocompatibility (minor) HB-1</td>
<td>1.84E-55</td>
<td>26.02</td>
<td></td>
</tr>
<tr>
<td>218641_at</td>
<td>hypothetical protein LOC65998</td>
<td>2.05E-55</td>
<td>5.13</td>
<td></td>
</tr>
<tr>
<td>203759_s_at</td>
<td>B-cell CLL/lymphoma 7A</td>
<td>6.00E-55</td>
<td>9.49</td>
<td></td>
</tr>
<tr>
<td>208791_at</td>
<td>Clusterin</td>
<td>6.46E-55</td>
<td>0.04</td>
<td></td>
</tr>
<tr>
<td>201540_at</td>
<td>four and a half LIM domains 1</td>
<td>1.02E-54</td>
<td>7.27</td>
<td></td>
</tr>
<tr>
<td>210987_x_at</td>
<td>tropomyosin 1 (alpha)</td>
<td>1.13E-54</td>
<td>0.16</td>
<td></td>
</tr>
<tr>
<td>201875_s_at</td>
<td>myelin protein zero-like 1</td>
<td>2.27E-54</td>
<td>4.66</td>
<td></td>
</tr>
<tr>
<td>204639_at</td>
<td>adenosine deaminase</td>
<td>3.17E-54</td>
<td>6.79</td>
<td></td>
</tr>
<tr>
<td>206390_x_at</td>
<td>platelet factor 4</td>
<td>4.25E-54</td>
<td>0.02</td>
<td></td>
</tr>
<tr>
<td>205237_at</td>
<td>ficolin (collagen/fibrinogen domain containing) 1</td>
<td>4.44E-54</td>
<td>0.03</td>
<td></td>
</tr>
<tr>
<td>205786_s_at</td>
<td>integrin, alpha M (complement component 3 receptor 3 subunit)</td>
<td>4.54E-54</td>
<td>0.08</td>
<td></td>
</tr>
<tr>
<td>203922_s_at</td>
<td>cytochrome b-245, beta polypeptide</td>
<td>5.23E-54</td>
<td>0.10</td>
<td></td>
</tr>
<tr>
<td>218332_at</td>
<td>brain expressed, X-linked 1</td>
<td>1.76E-53</td>
<td>0.04</td>
<td></td>
</tr>
</tbody>
</table>
Table A.7: Drug responsive genes of other subtypes.

<table>
<thead>
<tr>
<th>Probe Set ID</th>
<th>Gene Symbol</th>
<th>Gene Title</th>
<th>p Value</th>
<th>Fold Change (Other/Normal)</th>
</tr>
</thead>
<tbody>
<tr>
<td>202837_at</td>
<td>TRAFD1</td>
<td>TRAF-type zinc finger domain containing 1</td>
<td>4.97E-170</td>
<td>0.05</td>
</tr>
<tr>
<td>202829_s_at</td>
<td>VAMP7</td>
<td>vesicle-associated membrane protein 7</td>
<td>1.06E-153</td>
<td>17.66</td>
</tr>
<tr>
<td>202830_s_at</td>
<td>SLC37A4</td>
<td>solute carrier family 37 (glucose-6-phosphate transporter), member 4</td>
<td>1.30E-152</td>
<td>0.09</td>
</tr>
<tr>
<td>202804_at</td>
<td>ABCC1</td>
<td>ATP-binding cassette, sub-family C (CFTR/MRP), member 1</td>
<td>2.56E-150</td>
<td>0.16</td>
</tr>
<tr>
<td>202825_at</td>
<td>SLC25A4</td>
<td>solute carrier family 25 (mitochondrial carrier; adenine nucleotide translocator), member 4</td>
<td>1.85E-144</td>
<td>0.06</td>
</tr>
<tr>
<td>202843_at</td>
<td>DNAJB9</td>
<td>DnaJ (Hsp40) homolog, subfamily B, member 9</td>
<td>4.02E-143</td>
<td>0.13</td>
</tr>
<tr>
<td>2028_s_at</td>
<td>E2F1</td>
<td>E2F transcription factor 1</td>
<td>2.23E-131</td>
<td>0.06</td>
</tr>
<tr>
<td>202866_at</td>
<td>DNAJB12</td>
<td>DnaJ (Hsp40) homolog, subfamily B, member 12</td>
<td>2.46E-131</td>
<td>5.42</td>
</tr>
<tr>
<td>202824_s_at</td>
<td>TCEB1</td>
<td>transcription elongation factor B (SIII), polypeptide 1 (15kDa, elongin C)</td>
<td>4.61E-129</td>
<td>9.89</td>
</tr>
<tr>
<td>202836_s_at</td>
<td>TXNL4A</td>
<td>thioredoxin-like 4A</td>
<td>2.36E-122</td>
<td>76.16</td>
</tr>
<tr>
<td>202899_s_at</td>
<td>SFRS3</td>
<td>splicing factor, arginine/serine-rich 3</td>
<td>1.22E-121</td>
<td>6.69</td>
</tr>
<tr>
<td>202882_x_at</td>
<td>NOL7</td>
<td>nucleolar protein 7, 27kDa</td>
<td>7.88E-120</td>
<td>3.18</td>
</tr>
<tr>
<td>202822_at</td>
<td>LPP</td>
<td>LIM domain containing preferred translocation partner in lipoma</td>
<td>7.67E-119</td>
<td>10.75</td>
</tr>
<tr>
<td>202865_at</td>
<td>DNAJB12</td>
<td>DnaJ (Hsp40) homolog, subfamily B, member 12</td>
<td>1.09E-113</td>
<td>0.08</td>
</tr>
<tr>
<td>320_at</td>
<td>PEX6</td>
<td>peroxisomal biogenesis factor 6</td>
<td>1.56E-111</td>
<td>0.15</td>
</tr>
<tr>
<td>212013_at</td>
<td>PXDN</td>
<td>peroxidasin homolog (Drosophila)</td>
<td>7.77E-105</td>
<td>37.44</td>
</tr>
<tr>
<td>202874_s_at</td>
<td>ATP6V1C1</td>
<td>ATPase, H+ transporting, lysosomal 42kDa, V1 subunit C1</td>
<td>1.37E-104</td>
<td>6.83</td>
</tr>
<tr>
<td>202887_s_at</td>
<td>DDIT4</td>
<td>DNA-damage-inducible transcript 4</td>
<td>2.41E-104</td>
<td>31.66</td>
</tr>
<tr>
<td>Probe</td>
<td>Description</td>
<td>Log2 Fold Change</td>
<td>qValue</td>
<td></td>
</tr>
<tr>
<td>-------------</td>
<td>-------------------------------------------------------</td>
<td>-----------------</td>
<td>--------</td>
<td></td>
</tr>
<tr>
<td>204636_at</td>
<td>COL17A1 collagen, type XVII, alpha 1</td>
<td>6.44E-102</td>
<td>0.04</td>
<td></td>
</tr>
<tr>
<td>201416_at</td>
<td>SOX4 SRY (sex determining region Y)-box 4</td>
<td>9.11E-102</td>
<td>14.56</td>
<td></td>
</tr>
<tr>
<td>32091_at</td>
<td>SLC25A44 solute carrier family 25, member 44</td>
<td>3.35E-100</td>
<td>7.09</td>
<td></td>
</tr>
<tr>
<td>203355_s_at</td>
<td>PSD3 pleckstrin and Sec7 domain containing 3</td>
<td>2.23E-99</td>
<td>38.67</td>
<td></td>
</tr>
<tr>
<td>201417_at</td>
<td>SOX4 SRY (sex determining region Y)-box 4</td>
<td>3.44E-98</td>
<td>7.59</td>
<td></td>
</tr>
<tr>
<td>32029_at</td>
<td>PDPK1 3-phosphoinositide dependent protein kinase-1</td>
<td>1.78E-96</td>
<td>4.95</td>
<td></td>
</tr>
<tr>
<td>202854_at</td>
<td>HPRT1 hypoxanthine phosphoribosyltransferase 1</td>
<td>3.89E-96</td>
<td>4.06</td>
<td></td>
</tr>
<tr>
<td>202834_at</td>
<td>AGT angiotensinogen (serpin peptidase inhibitor,</td>
<td>3.03E-95</td>
<td>0.02</td>
<td></td>
</tr>
<tr>
<td></td>
<td>clade A, member 8)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>212012_at</td>
<td>PXDN peroxidasin homolog (Drosophila)</td>
<td>8.22E-95</td>
<td>28.12</td>
<td></td>
</tr>
<tr>
<td>202810_at</td>
<td>DRG1 developmentally regulated GTP binding protein 1</td>
<td>1.24E-94</td>
<td>5.64</td>
<td></td>
</tr>
<tr>
<td>206398_s_at</td>
<td>CD19 CD19 molecule</td>
<td>5.20E-94</td>
<td>16.38</td>
<td></td>
</tr>
<tr>
<td>218865_at</td>
<td>MOSC1 MOCO sulphurase C-terminal domain containing 1</td>
<td>2.11E-91</td>
<td>0.04</td>
<td></td>
</tr>
<tr>
<td>202883_s_at</td>
<td>PPP2R1B protein phosphatase 2 (formerly 2A),</td>
<td>4.12E-91</td>
<td>0.16</td>
<td></td>
</tr>
<tr>
<td></td>
<td>regulatory subunit A, beta isoform</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>214761_at</td>
<td>ZNF423 zinc finger protein 423</td>
<td>2.99E-90</td>
<td>49.35</td>
<td></td>
</tr>
<tr>
<td>201015_s_at</td>
<td>JUP junction plakoglobin</td>
<td>3.24E-90</td>
<td>16.86</td>
<td></td>
</tr>
<tr>
<td>32088_at</td>
<td>BLZF1 basic leucine zipper nuclear factor 1</td>
<td>3.71E-88</td>
<td>0.07</td>
<td></td>
</tr>
<tr>
<td>206656_s_at</td>
<td>C20orf3 chromosome 20 open reading frame 3</td>
<td>1.01E-87</td>
<td>0.22</td>
<td></td>
</tr>
<tr>
<td>34726_at</td>
<td>CACNB3 calcium channel, voltage-dependent, beta 3</td>
<td>4.20E-87</td>
<td>7.82</td>
<td></td>
</tr>
<tr>
<td></td>
<td>subunit</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>213668_s_at</td>
<td>SOX4 SRY (sex determining region Y)-box 4</td>
<td>4.64E-82</td>
<td>24.25</td>
<td></td>
</tr>
<tr>
<td>202332_at</td>
<td>CSNK1E casein kinase 1, epsilon</td>
<td>6.16E-82</td>
<td>4.03</td>
<td></td>
</tr>
<tr>
<td>202844_s_at</td>
<td>RALBP1 ralA binding protein 1</td>
<td>3.58E-80</td>
<td>5.09</td>
<td></td>
</tr>
<tr>
<td>215543_s_at</td>
<td>LARGE like-glycosyltransferase</td>
<td>7.03E-80</td>
<td>15.06</td>
<td></td>
</tr>
<tr>
<td>203787_at</td>
<td>SSBP2 single-stranded DNA binding protein 2</td>
<td>1.05E-79</td>
<td>7.61</td>
<td></td>
</tr>
<tr>
<td>202880_s_at</td>
<td>CYTH1 cytohesin 1</td>
<td>1.07E-79</td>
<td>24.58</td>
<td></td>
</tr>
<tr>
<td>Gene ID</td>
<td>Gene Symbol</td>
<td>Description</td>
<td>p-value</td>
<td>Fold Change</td>
</tr>
<tr>
<td>----------------</td>
<td>-------------</td>
<td>------------------------------------------------------------------------------</td>
<td>---------</td>
<td>-------------</td>
</tr>
<tr>
<td>211031_s_at</td>
<td>CLIP2</td>
<td>CAP-GLY domain containing linker protein 2</td>
<td>1.62E-79</td>
<td>14.21</td>
</tr>
<tr>
<td>202855_s_at</td>
<td>SLC16A3</td>
<td>Solute carrier family 16, member 3 (monocarboxylic acid transporter 4)</td>
<td>3.68E-79</td>
<td>0.01</td>
</tr>
<tr>
<td>202879_s_at</td>
<td>CYTH1</td>
<td>Cytohesin 1</td>
<td>5.44E-79</td>
<td>0.06</td>
</tr>
<tr>
<td>202945_at</td>
<td>FPGS</td>
<td>Folatepolyglutamate synthase</td>
<td>1.14E-78</td>
<td>4.68</td>
</tr>
<tr>
<td>201425_at</td>
<td>ALDH2</td>
<td>Aldehyde dehydrogenase 2 family (mitochondrial)</td>
<td>1.51E-78</td>
<td>0.11</td>
</tr>
<tr>
<td>218613_at</td>
<td>PSD3</td>
<td>Pleckstrin and Sec7 domain containing 3</td>
<td>2.20E-78</td>
<td>16.58</td>
</tr>
<tr>
<td>201418_s_at</td>
<td>SOX4</td>
<td>SRY (sex determining region Y)-box 4</td>
<td>4.56E-78</td>
<td>12.66</td>
</tr>
<tr>
<td>216041_x_at</td>
<td>GRN</td>
<td>Granulin</td>
<td>8.07E-77</td>
<td>0.15</td>
</tr>
</tbody>
</table>
BIBLIOGRAPHY


