Application of Machine Learning Techniques to Acute Myeloid Leukemia

Qi WEI

A thesis
submitted in partial fulfillment of the
requirements for the degree of

Master of Science in Computer Science and Systems

University of Washington

2016

Committee:

Ka Yee Yeung

Ling Hong Hung

Martine De Cock

Program Authorized to Offer Degree:
Institute of Technology - Tacoma

University of Washington

Abstract

Application of Machine Learning Techniques to Acute Myeloid Leukemia

Qi WEI

Chair of the Supervisory Committee:
Dr. Ka Yee Yeung
Institute of Technology - Tacoma

This thesis is inspired by the position paper "Predictive, personalized, preventive, participatory (P4) cancer medicine" [1]. The basic concept of P4 medicine was "The right patient with the right drug at the right dose at the right time.". In other words, the goal is to tailor medical treatment to the individual characteristics, needs, and preferences of a patient during all stages of care, including prevention, diagnosis, treatment, and follow-up [1]. In this thesis, we used Acute Myeloid Leukemia (AML) as our case study because if untreated, AML progresses rapidly and is typically fatal within weeks or months, and also because genomic data were available. It has also been shown that AML is associated with gene mutations [2], and hence, genomic approaches have the potential to contribute to this heterogeneous cancer.

We applied machine learning algorithms to build predictive models using biomedical data profiling AML patients. Specifically, in chapter 1 we introduced the problem and related works of our projects and in chapter 2 we introduced background knowledge of all the algorithms and technologies being used. In chapter 3, we report the identification of 24-gene signature predictive of the relapse of low-risk AML patients. These 24 genes could be used to distinguish a future patient will be relapse or non-relapse. Our findings in chapter 3 were derived by mining gene expression data of AML patients generated from microarray technology and next generation sequencing technology. We would like to note

the limitations of this "personalized medicine" approach: further clinical evidence and trials would be needed to elucidate the underlying biological mechanisms. In chapter 4, we applied correlation analyses to high-throughput drug sensitivity data to identify gene mutations that could be potential candidates to explain patients' responses to AML drugs. In chapter 5, we concluded our projects and given an overview of possible future works. In this thesis, we focused on AML as our case study. However, our methods could be applicable to other diseases for which data are available.

ACKNOWLEDGMENTS

The author wishes to express sincere appreciation to University of Washington, Tacoma. Special thanks to Prof. Dr. Ka Yee Yeung, Prof. Dr. Martine De Cock, and Dr. Ling Hong Hung for your sincere help and instructions.

TABLE OF CONTENTS

Page

List of Figures iv

List of Tables vi

Chapter 1: Introduction 1
1.1 Personalized Medicine 1
1.2 Big biological data 1
1.3 Machine learning methods and biomarker discovery 3
1.4 Computational drug discovery 4

Chapter 2: Methodology And Experimental Design 6
2.1 Overview 6
2.2 Data sources 6
2.2.1 Pediatric AML gene expression data 6
2.2.2 AML drug sensitivity data 7
2.2.3 AML Gene Mutation data 8
2.3 Normalization and Data Processing Methods 8
2.3.1 Variance-stabilizing transformation (VST) 8
2.3.2 Interpolation of eight concentrations on drug sensitivity data 9
2.3.3 Mutation Analysis by MyAML 9
2.3.4 Pairwise Correlation Coefficients 10
2.3.4.1 Pearson's correlation coefficient 10
2.3.4.2 Rank-based Spearman's correlation 10
2.4 Visualization Methods 11
2.4.1 Principal Component Analysis (PCA) 11
2.4.2 Heat map 11
2.5 Machine Learning algorithms 11

2.5.1 Univariate feature ranking method: Between group sum of square to within group sum of square ratio (BSS/WSS) . . . 11
2.5.2 Multivariate feature selection method: Least absolute shrinkage and selection operator (LASSO) . . . 12
2.5.3 Iterative Bayesian Model Averaging (iBMA) . . . 12
2.5.4 Synthetic Minority Over-sampling Technique (SMOTE) . . . 12
2.5.5 Random forest . . . 12
2.5.6 Support Vector Machine (SVM) . . . 13
2.6 Assessment of classification accuracy . . . 13
2.6.1 Contingency table . . . 13
2.6.2 Balanced Accuracy (BAC) . . . 14
2.6.3 Area under the Receiver Operator Characteristic Curve (AUROC or AUC) . . . 14
2.6.4 Precision, recall and F1 . . . 14
2.6.5 Cross Validation . . . 15
2.6.6 Independent test data . . . 15

Chapter 3: Gene signatures predictive of relapse in low risk pediatric Acute Myeloid Leukemia patients . . . 16
3.1 Background: Acute Myeloid Leukemia (AML) . . . 16
3.1.1 Our contributions on this project . . . 16
3.2 Results on feature selection . . . 18
3.2.1 Univariate feature ranking . . . 18
3.2.2 Multivariate feature selection . . . 19
3.2.3 Visualizing the 24-gene signature in reduced dimensions . . . 20
3.3 Comparing Microarray and RNA sequencing data . . . 21
3.4 Assessment methods . . . 26
3.4.1 Classification . . . 26
3.5 Validation . . . 27
3.5.1 Cross validation using the microarray data . . . 27
3.5.2 Independent assessment using the RNA sequencing (RNAseq) data . 29

Chapter 4: Personalized Medicine for Acute Myeloid Leukemia (AML): Correlating drug sensitivity and mutation data . . . 35

4.1 Introduction: Correlating drug sensitivity and mutation data 35
4.2 Analysis of drug sensitivity data 36
4.2.1 Heatmap visualization 37
4.3 Correlation analyses of drug sensitivity data and mutation data 43
4.3.1 Heatmaps on mutation data 43

Chapter 5: Conclusions 48
5.1 Gene signatures predictive of relapse in AML patients 48
5.2 Correlating drug sensitivity and mutation data 48
5.3 Future work and contributions 48

LIST OF FIGURES

Figure Number Page

3.1 An overview of our method. 17

3.2 AUC and ACC of SVM when doing 10 fold 10 run cross validation on top n BSS/WSS ranked genes. Both AUC and ACC peaked at around 400 univariate ranked gene sets. 18

3.3 24 Selected genes and their regression coefficients given by LASSO using the AML microarray data. 19

3.4 Projection of the 24-gene signature onto the first three principle component space using the microarray data. Each data point represented one of the 119 low-risk AML patients. The red dots and blue dots represented relapse and non-relapse patients respectively. 21

3.5 A scatter plot comparing the variance stabilizing transformed (VST) RNAseq data (y-axis) and RMA normalized microarray data (x-axis) across the 24 signature genes and 52 overlapping patients. 22

3.6 A heatmap showed distribution of the microarray data across our 24 signature genes and 52 AML patients. These 52 patients were profiled in both microarray and RNAseq data from TARGET. The green represented the highly expressed genes and the red represented the lowly expressed genes. Those blue and yellow bars, on the left side showed the distribution of relapse and non-relapse patient samples. 24

3.7 A heatmap showed distribution of RNAseq data normalized by variance stabilizing transformation. Our 24 signature genes on the overlapping set of 52 patients were shown. 25

3.8 A heatmap showed distribution of RNAseq data normalized by variance stabilizing transformation before applied Linear regression based on Microarray data. 30

3.9 A heatmap showed distribution of RNAseq data normalized by variance stabilizing transformation after applied Linear regression based on Microarray data. 31

4.1 Personalized cancer therapy is a treatment strategy centered on the ability to predict which patients are more likely to respond to specific cancer therapies. 35
4.2 A heatmap of Drug Chemosensitivity across all 160 drugs and 24 patients at concentration 10^{-6}. 38
4.3 A heatmap of Drug Chemosensitivity across all 160 drugs and 24 patients at concentration 10^{-7}. 39
4.4 A Heatmap of Drug Chemosensitivity across all 160 drugs and 24 patients at concentration $3 * 10^{-7}$ 40
4.5 A heatmap of Drug Chemosensitivity across common leukemia drugs and 24 patients at concentration 10^{-6}. 41
4.6 A heatmap of Drug Chemosensitivity across common leukemia drugs and 24 patients at concentration 10^{-7}. 42
4.7 A heatmap of Drug Chemosensitivity across common leukemia drugs and 24 patients at concentration $3 * 10^{-7}$. 42
4.8 A heatmap of genes containing missense mutations that were present in at least 2 patient samples. 44
4.9 A heatmap of computed p-values between drug sensitivity and mutation at concentration 10^{-7} using the Pearson's correlation. 45
4.10 A heatmap of computed p-values between drug sensitivity and mutation at concentration 10^{-7} using Pearson's correlation (significant pairs). 46

LIST OF TABLES

Table Number Page

2.1 A Contingency table showed relationship between actual and predicted class 14

3.1 Cross validation results with all 119 patients using microarray data 28

3.2 Cross validation results with all 52 patients using microarray data 28

3.3 Cross validation results with all 103 patients using RNAseq data. 32

3.4 Cross validation results with all 52 patients using RNAseq data. 33

3.5 Results based on trained models with all 119 patients with microarray data and independent test with overlapped 52 patients with RNA-seq data. 34

Chapter 1
INTRODUCTION

1.1 Personalized Medicine

Personalized medicine is often described as providing "the right patient with the right drug at the right dose at the right time." Generally speaking, personalized medicine (also known as precision medicine) may be thought of as the tailoring of medical treatment to the individual characteristics, needs, and preferences of a patient during all stages of care, including prevention, diagnosis, treatment, and follow-up [3].

In a seminal paper titled "Predictive, personalized, preventive, participatory (P4) cancer medicine"[1], Drs. Lee Hood and Stephen Friend described a new discipline of medicine as well as potential problems of this discipline. A key aspect of personalized medicine is the detection and characterization of diseases and cancer. In particular, the identification of biomarkers, which are biological molecules that can be used to give information about the molecular characteristics of a tumor, is a promising area of cancer research. There is extensive literature on the design and translational impact of biomarkers, such as papers dealing with limitations of current biomarker development methodologies and regulatory and reimbursement policies and practices [4], blood-based strategies to detect and monitor cancer [5], compensatory measures to restore or improve anticancer immune responses[6], modeling for prediction of drug resistance[7], and comprehensive biomarker study registry[8].

1.2 Big biological data

Since year 2008, we have entered an era of big data explosion [9]. Scientists in various disciplines, including astronomers, physicists and biologists, faced difficulties and challenges to solve, normalize, filter and process petabytes of data [10].

With the development of technologies like high-throughput genomics and next generation sequencing, the size of genomic data being generated is doubled every year [11]. Many computational scientists participated in bioinformatics research and applied data science techniques to big biomedical data. In addition, there were unique barriers in biology data mining: biological big data generated from different experiments using separate biological principles and designs from multiple laboratories and universities were extremely heterogeneous. Therefore, data analyses methods and tools being used to interpret these data became crucial.

Microarrays have been used to find and illustrate patterns of gene expressions, variations in genomic DNA (e.g "SNP chips"), and bindings between transcription factors and DNA[12]. In chapter 3 of this thesis, "Gene signatures predictive of relapse in low risk pediatric Acute Myeloid Leukemia patients", microarrays technology were used to quantify levels of gene transcripts (mRNA). The microarray data generated at the Fred Hutchinson Cancer Research Center used the well-established Affymetrix microarrays. The Affymetrix Gene Chip arrays used multiple short oligo probes across each gene and transcript to measure the activity level.

Despite that useful results had been generated in the literature using the microarray technology, there were two major drawbacks: the first one was the incomplete nature of gene annotations, and the second one was limitations on microarray density[13]. When applying microarray technology, data scientists should consider signal-to-noise ratios, which varied between different probes. Due to these limitations, next-generation sequencing such as RNA sequencing (RNA-seq) can be used as an alternative technology to profile gene expression levels in recent years.

RNA sequencing technology was a relatively new method to transcriptome profiling with deep-sequencing technologies[14]. An advantage of RNA-seq technology, when compared to microarrays was the capability to identify previously uncharacterized mRNA isoforms and new classes of non-coding RNAs[15].

In this thesis, we applied data analyses and machine learning techniques to many different

types of biological data, including both microarray data and RNA sequencing data (Chapter 3), drug sensitivity data and mutation data generated by next generation sequencing technology (Chapter 4).

1.3 Machine learning methods and biomarker discovery

According to a report published in 2006 by Jemal et al., there were more than 1.4 million new cases of cancer with a fatality rate over 50%, which represented near one fourth of total death cases in the United States every year[16]. Therefore the importance of early stage cancers' detection and identification was increasing over time. One possible solution was the discovery of novel biomarkers for cancer using modern biological methods and genome-wide data, which could infer notable signatures and patterns for each class of cancer [17].

Genome-wide technology such as microarrays and next generation sequencing measured gene expression patterns, and these gene expression data could be used to classify tumours into clinically relevant subgroups (e.g.[18], [19], [20], and many others). New tools have been devised for disease recurrence and treatment response prediction. In the meantime, new perspectives into various oncogenic pathways and metastatic progression have also been proposed[21].

Many gene selection methods had been proposed for gene expression data. For example, Guyon *et al.* used support vector machine and recursive feature elimination to derive a small set of genes classifying cancer gene expression data[22]. As another example, Pirooznia *et al.* conducted an empirical study applying a variety of classification and feature selection algorithms to eight gene expression datasets and concluded that feature selection methods are essential [23]. Ensemble machine learning methods had been applied to applications in cancer classification. For example, Tan *et al.* showed that bagged and boosted decision trees usually outperformed single decision trees[24]. Boosting methods were also adopted for classifying cancer gene expression data in ([19]; [25]).

Many computational methods had already been developed for the analyses of cancer genomics

data. For example, Yeung and colleagues previously developed methods to integrate gene expression data with expert knowledge and predicted functional relationships using iterative Bayesian model averaging to chronic myeloid leukemia [26]. As another example, Huang *et al.* developed a pathway-based prognosis prediction model, which summarized pathway-based risk measurements using the pathway dysregulation score (PDS)[27].

1.4 Computational drug discovery

A general standard for new drug approval process of Food and Drug Administration (FDA)[28] started with drug companies sending FDA's Center for Drug Evaluation and Research (CDER) the evidence from the tests they conducted themselves to prove the drug is safe and effective for its intended use. A new drug could be marketed only if it has been thoroughly investigated in clinical trials. The results of these numerous trials were then submitted in a New Drug Application to the FDA. The FDA sent these results to a committee for review. The FDA then took its report and decided whether to approve the drug as safe and effective or not. Only after the FDA reached its final verdict, marketing of the drug could finally take place[29]. The process could take as long as 10-20 years. Compared to the average survival period of cancer patients, it took decades to embrace a new effective drug.

Take Genasense® as an example, which had been developed by a small pharmaceutical company to combat Metastatic melanoma and chronic lymphocytic leukemia (CLL). Genasense® research dates back to 1997, when the Journal of Pharmacology and Experimental Therapeutics reported on a study in mice that looked at its effect on the bcl-2 protein[30]. Subsequent studies had already confirmed the activity of Genasense®.

In the February 2007 issue of the Journal of Clinical Oncology, a study conducted at 100 centers around the world and involving patients with advanced chronic lymphocytic leukemia (CLL) was reported. The findings showed that 17 percents of patients in the Genasense® plus chemo arm achieved a complete or partial response, as opposed to 7 percents of patients who were treated only with chemotherapy[31]. The FDA still did not approve it (at year 2010), ignoring compelling evidence of the drug's efficacy.

The shortcomings of such a lengthy approval process were as follows: 1.Those clinical trials themselves could take many years to complete. 2. It could take decades after a promising cancer drug has been discovered before the FDA finally approve it[29].

Nowadays traditional drug developments, such as random screening, chance discovery, and high-throughput screening (HTS)[32], had begun to take precedence as means for finding novel therapeutics. They are all brute force approaches relying on automation to screen large numbers of molecules in search for those that elicit the desired biological response. These random, trial and error processes normally only focus on a single drug/molecule structure at a time[33]. Therefore, these shortcomings motivated the development of computational drug discovery methods, which required significantly less preparation time. Researchers have the capability to perform computer-aided drug design (CADD)[34] studies while traditional pipeline was being prepared and implemented.

Since 1981, for over thirty years computational drug discovery had played a major role in the development of therapeutically important small molecules. These computational methods were capable of increasing the hit rate of novel drug compounds. In a typical drug discovery process, CADD was usually used for three major purposes: 1. filtering large compound libraries into smaller sets of predicted active compounds that can be tested experimentally; 2. guiding the optimization of lead compounds; 3. designing novel compounds, by "growing" starting molecules one functional group at a time or by piecing together fragments into novel chemotypes[33].

Chapter 2

METHODOLOGY AND EXPERIMENTAL DESIGN

2.1 Overview

Our overarching goal was to discover predictors for cancer and drug sensitivity by applying machine learning methods to big data in biology. We will use Acute Myeloid Leukemia (AML)[35] as our case study. Acute myeloid leukemia (AML) is a heterogeneous blood cancer, which is a cancer of the myeloid line of blood cells, characterized by the rapid growth of abnormal white blood cells that accumulate in the bone marrow and interfere with the production of normal blood cells. AML progresses rapidly and is typically fatal within weeks or months if left untreated. The treatment of AML consists primarily of chemotherapy, and is divided into two phases: 1. induction and 2. consolidation therapy. The goal of the first phase is to achieve a complete remission by reducing the number of leukemic cells to an undetectable level; the goal of the second phase is to eliminate any residual undetectable disease and achieve a cure.

It has been hypothesized that gene mutations could be triggered by the progression of AML [2]. Therefore, the discovery of biomarkers from AML genomics data is promising. Our overarching goals in the projects described in this thesis are: 1. to identify gene signatures predictive of relapse in low-risk AML patients (see Chapter 3); and 2. to adapt patients' therapy based on drug sensitivity data and gene mutation data.

2.2 Data sources

2.2.1 Pediatric AML gene expression data

The pediatric AML gene expression data were part of the Therapeutic Applicable Research to Genetic Effective Therapy (TARGET) Initiative [36] generated in the Meshinchi Lab at

the Fred Hutchinson Cancer Research Center. The TARGET AML Initiative provided gene expression data, whole genome sequencing data as well as clinical data across a couple hundred AML patients. In particular, we would use both the microarray and RNAseq gene expression data across different cohorts in this project.

We used a subset of AML microarray data consisting of 119 low risk pediatric patients in the training step to identify a 24-gene signature, and validated our 24-gene signature using RNAseq data consisting of 77 low risk patients in the testing step. Patients with AML are often divided into risk groups (such as standard-risk, high-risk, or very high-risk), with more intensive treatment given to higher risk patients. Patients with low risk have less death rate and fewer simultaneous problems.

The microarray data were generated using Affymetrix Exon arrays and normalized using Robust Multi-array Average (RMA) [37]. The normalized microarray data consisted of roughly 54,000 probe sets. We mapped the Affymetrix gene cluster identifiers to Entrez Gene IDs, and filtered out probe sets that are not mapped to any Entrez Gene IDs. The resulting microarray gene expression dataset consisted of 32 thousand probe sets and 119 low-risk AML patients. Our goal was to identify genes predictive of relapse among low-risk AML patients. RNAseq was a comparatively newly developed approach for transcriptome profiling, which used deep-sequencing technologies. The Illumina platform was an established precise measurement of transcript levels compared to the microarray technology [38].

2.2.2 AML drug sensitivity data

Most AML patients either relapse or fail to respond to initial therapy. In collaboration with Professor Pamela Becker, M.D., Ph.D., in the Department of Hematology at University of Washington Seattle, we aim to develop individualized approaches to AML therapy. AML drug sensitivity data generated from the Becker Lab measured the level of cytotoxicity of 24 patient samples in response to approximately 160 drugs over various concentration [39]. Among these 160 AML drugs profiled in this drug sensitivity data, of which 56 are FDA approved and 104 are investigational. The raw drug sensitivity data provided us with the

survival rate (%) at 8 different concentrations for each drug and each patient. We aim to define molecular information that might better predict response to conventional or novel therapies [39].

2.2.3 AML Gene Mutation data

The mutation data provided by the Becker Lab was generated by MyAML™. MyAML™ used next generation sequencing (NGS) to analyze the 3' and 5' UTR and exonic regions of 196 genes, which known to have recurrent mutations in AML and potential genomic breakpoints within known somatic gene fusion breakpoints known to be associated with AML. The MyAML panel includes the coding an noncoding exons. Analysis for the 36 genes involved in the majority of chromosomal translocations found in AML is also provided. Fragmented genomic DNA (around 3.4Mb) was captured with a customized probe design, and sequenced with 300bp paired end reads on an Illumina MiSeq instrument to an average depth of coverage bigger than 1000x. Dr. Becker and colleagues identified single nucleotide variants (SNVs), insertion/deletions (indels), inversions and translocations were identified, annotated, characterized, and allelic frequencies from the mutation data.

In this thesis, we focused on mis-sense mutations (non-synonymous point mutation), which referred to a change in one amino acid in a protein, arising from a point mutation in a single nucleotide. Mis-sense mutation is a type of non-synonymous substitution in a DNA sequence. We then sought to correlate the mutation data with the chemotherapy sensitivity data.

2.3 Normalization and Data Processing Methods

2.3.1 Variance-stabilizing transformation (VST)

In applied statistics, variance-stabilizing transformation (VST) technology was a transformation of raw data in order to simplify considerations in graphical exploratory data analysis or to allow the application of simple regression-based or analysis of variance techniques [40, 41].

We used the VST implementation in an R package called "DEseq" from the Bioconductor project to normalize the RNA-seq data.

2.3.2 Interpolation of eight concentrations on drug sensitivity data

In the AML drug sensitivity data, we aim to explore drug sensitivity across patients at a given concentration. However, the concentrations for which the drug is applied were not identical across different patients and different drugs. Therefore, for simplicity, we assumed all survival percentages were linear within a close range, and used linear interpolation to interpolate percentage survival at concentrations of $10^{-6}, 10^{-7}, 3*10^{-7}$ M. linear interpolation at a given concentration x requires that there are measurements above x and measurements below x. In the case that either is absent, we assigned the value of unknown (i.e. NA) to the given drug and patient. In addition, we converted all negative percentage survival to zeros, and all 100+ % to 100%. By definition, percentages that are negative and above 100 have no realistic meaning.

2.3.3 Mutation Analysis by MyAML

The Becker Lab found that most commonly mutated genes in AML are all targeted with an average depth of coverage of 975x (range = 417x to 1370x). They also accurately detected known mutations, including missense and nonsense mutations in FLT3, DNMT3A, IDH1, IDH2, KIT, NRAS, KRAS, and TP53. The Becker Lab has accurately detected mutations with allelic frequencies as low as 2.5%, with 95% reproducibility even. We were provided the mutation data after Dr. Becker and her colleagues performed these quality control (QC) steps on the data. Therefore no independent quality control was conducted by our team.

2.3.4 Pairwise Correlation Coefficients

2.3.4.1 Pearson's correlation coefficient

We used Pearson's correlation coefficient (also Pearson's product-moment correlation coefficient), to measure the linear correlation between two variables X and Y. The outcome was a continuous value between -1 and +1, where +1 represented total positive correlation, 0 represented no correlation, and 1 represented total negative correlation[42]. Correlation is widely used as a measure of the degree of linear dependence between two variables. The most commonly used measure of association is Pearson's product moment correlation coefficient, often denoted as r[43]. Given N measurements $(X_1, Y_1), \ldots, (X_N, Y_N)$ of two variables X and Y measured on each of N individuals, we define:

$$\bar{X} = \frac{1}{N} \sum_{i=1}^{N} X_i \tag{2.1}$$

$$\bar{Y} = \frac{1}{N} \sum_{i=1}^{N} Y_i \tag{2.2}$$

$$r = \frac{\sum_{i=1}^{N} \left(X_i - \bar{X}\right) \left(Y_i - \bar{Y}\right)}{\sqrt{\sum_{i=1}^{N} \left(X_i - \bar{X}\right)^2} \sqrt{\sum_{i=1}^{N} \left(Y_i - \bar{Y}\right)^2}} \tag{2.3}$$

2.3.4.2 Rank-based Spearman's correlation

The Spearman's rank correlation coefficient (also Spearman's rho), often denoted by the Greek letter ρ, is a non-parametric measure of statistical dependence between two variables [44]. In contrast to the Pearson correlation coefficient, the Spearman's correlation coefficient is defined using the rankings of the two variables of interest [45]. Specifically, given a sample of size N, the N measurements X_i, Y_i are converted to ranks x_i, y_i. We define $d_i = x_i$ - y_i as the difference between ranks. The Spearman's correlation ρ is defined as:

$$\rho = 1 - \frac{6 \sum_{i=1}^{N} d_i^2}{N(N^2 - 1)} \tag{2.4}$$

2.4 Visualization Methods

2.4.1 Principal Component Analysis (PCA)

We used Principal component analysis (PCA) as a dimension reduction tool[46], which used an orthogonal transformation to convert a set of observations of possibly correlated variables into "principal components" (a set of values of linearly uncorrelated variables). The number of principal components must be less than or equal to the number of original variables. We use PCA to find and visualize low dimension projections based on maximal variability.

2.4.2 Heat map

We used heat map as a graphical tool to illustrate the drug sensitivity data and mutation data where the individual values (continuous for drug sensitivity, discrete for mutation) contained in a matrix are represented as colors (such as red versus green). Heat maps can also be used as 2-dimensional displays of clustering results.

2.5 Machine Learning algorithms

2.5.1 Univariate feature ranking method: Between group sum of square to within group sum of square ratio (BSS/WSS)

Yeung *et al.* previously proposed to use the BSS/WSS (between group sum of square to within group sum of square ratio) as a univariate pre-processing step for feature selection [47]. The intuition was that features that were highly distinct between different classes and had relatively high similarity within the same group, will be given a higher rank. The following equation showed the mathematical definition of BSS/WSS. For a gene j, let D_ij denoted the expression level of gene j under sample i, D_kj the average expression level of gene j over samples in class k and D_j the average expression level of gene j over all samples.

$$\frac{BSS(j)}{WSS(j)} = \frac{\sum_i \sum_k I(Y_i = k)(\bar{D_{kj}} - \bar{D_{.j}})^2}{\sum_i \sum_k I(Y_i = k)(D_{ij} - \bar{D_{kj}})^2} \tag{2.5}$$

2.5.2 Multivariate feature selection method: Least absolute shrinkage and selection operator (LASSO)

LASSO is a widely used multivariate variable selection method in many applications [48]. It minimizes the usual sum of squared errors, with a penalty term on the absolute values of the regression coefficients [49]. LASSO is implemented in the glmnet R package[50].

2.5.3 Iterative Bayesian Model Averaging (iBMA)

The Iterative Bayesian Model Averaging (BMA) algorithm is a variable selection and classification algorithm developed for the classification of 2-class microarray data [51]. We used the iBMA implementation in R called "iterativeBMA" available from the Bioconductor project.

2.5.4 Synthetic Minority Over-sampling Technique (SMOTE)

The Synthetic Minority Over-sampling Technique (SMOTE) is an over-sampling approach, in which the minority class is over-sampled by creating "synthetic" examples rather than over-sampling with replacement [52]. SMOTE is a technique commonly used to handle classification tasks in which the number of positive and negative training samples are unbalanced. In SMOTE, the over-sampling of minority class was achieved by assigning synthetic examples to each of the minority class sample. We used the "unbalanced" R package available from the Comprehensive R Archive Network (CRAN) (https://cran.r-project.org/).

2.5.5 Random forest

Random forest is an ensemble learning method for classification. It works by building a multitude of decision trees over training examples and then generating a mean regression of the individual trees [53]. Random forest is implemented in the "randomForest" R package (https://cran.r-project.org/web/packages/randomForest/index.html). In this thesis, we used the default parameters of random forest in our experiments (i.e. $ntree = 500$, *corr.bias* = FALSE). *ntree* is a parameter to define number of trees to grow. The value of *ntree*

should not be set to too small a number, to ensure that every input row gets predicted at least a few times. *corr.bias* is a parameter to define whether performing bias correction for regression or not. In this thesis, we decide not to use this correction.

2.5.6 Support Vector Machine (SVM)

SVM is a supervised learning model which represents positive and negative training examples as points in higher dimensional space such that the positive and negative examples are separated [54]. SVM is implemented in the "e1071" R package (https://cran.r-project.org/web/packages/e1071/index.html). We used the default parameters of e1071 in our experiments (i.e. $kernel =$ "radial", $gamma = 1/$(data dimension)). *gamma* is is a free parameter, in this thesis we just used 1 over the data dimension from both microarray and RNA-seq data. *kernel* is a parameter to define the the kernel used in training and predicting.

2.6 Assessment of classification accuracy

We defined a binary response variable Y representing the clinical outcome. Specifically, we defined $Y = 0$ for non-relapsed patients and $Y = 1$ for relapsed patients. We evaluated the performance of our classifiers using two measures: the BAC (balanced accuracy), AUC (also AUROC, Area under the Receiver Operator Characteristic Curve) and F1 score (also F-score or F-measure). We used a contingency table to compare the response variable Y to our predictions.

2.6.1 Contingency table

We used a contingency table approach to evaluate classification accuracy. Table 2.1 illustrated the following quantities: TP (true positives), TN (true negatives), FP (false positive) and FN (false negatives). Specifically, a patient was considered a TP if the response variable Y is 1 (relapsed) and the predicted class was also 1. Similarly, a false positive was identified if the response variable Y is 0 (non-relapsed) and the predicted class was mislabeled to 1, a

false negative was identified if the response variable Y is 1 (non-relapsed) and the predicted class was mislabeled to 0. Finally, a patient was considered a TN if the response variable Y is 0 (non-relapsed) and the predicted class was also 0.

	Predicted class =1	Predicted class =0
Y=1 (relapsed)	TP (True Positive)	FN (False Negative)
Y=0 (non-relapsed)	FP (False Positive)	TN (True Negative)

Table 2.1: A Contingency table showed relationship between actual and predicted class

2.6.2 Balanced Accuracy (BAC)

The Balanced Accuracy (BAC) score was defined by Equation 2.6, where P is the number of relapsed patients and N is the number of non-relapsed patients.

$$BalancedAccuracy : BAC = \frac{1}{2}(\frac{TP}{P} + \frac{TN}{N}) \tag{2.6}$$

2.6.3 Area under the Receiver Operator Characteristic Curve (AUROC or AUC)

The AUC studied the trade-off between the numbers of TP versus the numbers of FP over varying threshold probabilities. In general, a good predictive model should yield a high AUC. We used the R package "pROC" [55] by Robin *et al.* to compute the AUC.

2.6.4 Precision, recall and F1

Precision (P) is defined as the number of true positives divided by the sum of true positives and false positives (see Equation 2.7). Recall (R) is defined as the number of true positives divided by the sum of true positives and false negatives (see Equation 2.8). The F1 score is the harmonic mean of precision P and recall R (see Equation 2.9).

$$Precision : P = \frac{TP}{TP + FP} \tag{2.7}$$

$$Recall : R = \frac{TP}{TP + FN} \tag{2.8}$$

$$F1score : F_1 = 2\frac{PR}{P + R} \tag{2.9}$$

2.6.5 Cross Validation

Cross validation (CV), or M-fold cross-validation of prediction, is to divide the training set into M non-overlapping subsets of roughly equal size. The idea is to randomly divide the training data into M parts, build predictive models using $(M-1)$ parts of the training data, and then apply the predictive models to the remaining subset to obtain an estimate of the prediction error. This process is repeated M times for each of the M subsets, and the CV error is computed as the average of the errors over these M folds. These M-fold CV are typically repeated many times. In the case of $M = n$ (this special case is called leave-one-out CV), each observation (sample) in the training set is left out in turn [56].

In the computational assessment in this thesis, we applied 10-fold cross-validation across different machine learning algorithms (such as Support Vector Machine, Random Forest). We selected 10-fold as recommended by Ambroise *et al.* [56].

2.6.6 Independent test data

We also used independent test data consisting of independent patient samples generated using a different technology to validate the prediction accuracy of inferred signatures.

Chapter 3

GENE SIGNATURES PREDICTIVE OF RELAPSE IN LOW RISK PEDIATRIC ACUTE MYELOID LEUKEMIA PATIENTS

3.1 Background: Acute Myeloid Leukemia (AML)

With the advent of predictive, personalized, preventive, participatory (4P) cancer medicine [1], new analytical tools that employ systems-based approaches to diseases have been proposed. In this Chapter, we built predictive models using both microarray and RNAseq gene expression data in Acute Myeloid Leukemia (AML) and develop diagnostic tools that will facilitate the tailoring of therapy strategy for each individual patient.

Extensive genome-wide data, including epigenetic, expression and sequence data are publicly available for AML [57]. These data could potentially be used to identify potential targets for therapeutic intervention of AML [58]. Many predictive gene signatures suffered from sampling biases due to limited numbers of training samples, and subsequently did not guarantee comparable performance on independent cohorts of patients ([59, 60, 61]), gene expression data still have the potential to contribute to the development of personalized medicine [62].

3.1.1 Our contributions on this project

In this study, we used the pediatric AML data from the NIH TARGET Initiative (Therapeutic Applicable Research to Genetic Effective Therapy, https://ocg.cancer.gov/programs/target). Due to the heterogeneity of AML patients[63], we confined our study to a subset of low-risk patients with cytogenetic type inversion 16 (inv(16)) and translocation (8, 21) (t(8;21)). These patients were selected by our collaborators with clinical expertise in AML and were considered to be low risk from a clinical perspective.

Here, we presented a novel approach combining univariate and multivariate variable selection

methods that identified a 24-gene signature predictive of relapse in pediatric AML patients. We showed that individual genes could not effectively classify relapse among these patients. However, our 24-gene signature derived from a combination of univariate and multivariate techniques could separate the relapse versus the non-relapse patients using cross validation on the microarray data. Specifically, we used microarray data to build predictive models and subsequently validated our models using independent RNAseq data. Figure 3.1 showed an overview of our method.

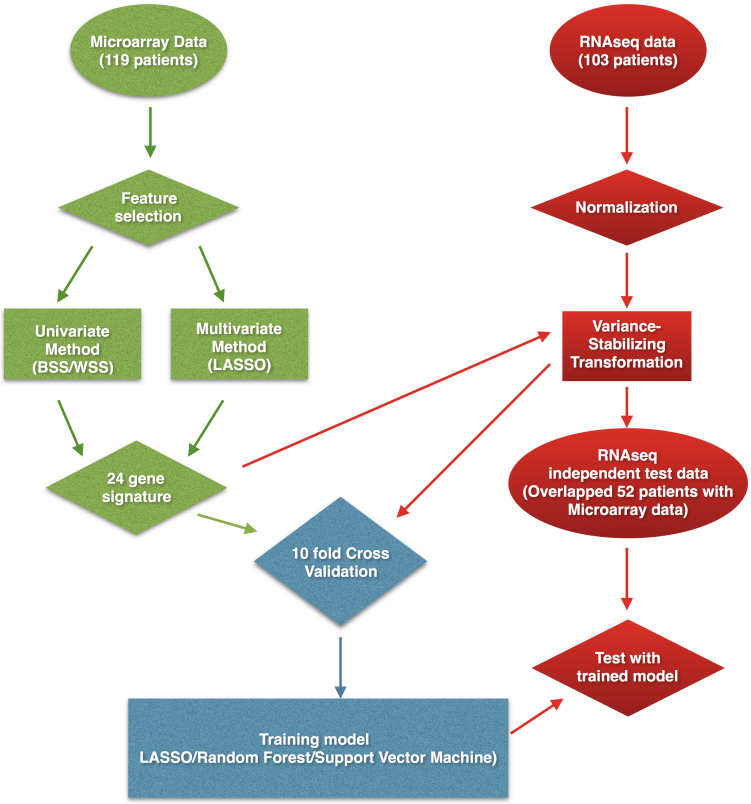

Figure 3.1: An overview of our method.

3.2 Results on feature selection

3.2.1 Univariate feature ranking

We ranked all the genes in the microarray data using the BSS/WSS (described in Chapter 2), then used 10-fold cross validation to find the most suitable subset of top-ranked features. Figure 3.2 showed a plot of the AUC and ACC, defined by the area under the ROC (AUC, also known as AUROC) and the percentage of correct prediction (ACC) over number of all cases respectively, using support vector machine (SVM) as the classification method. We found in our later experiments that these top 400 BSS/WSS genes contained the most predictive subset of genes.

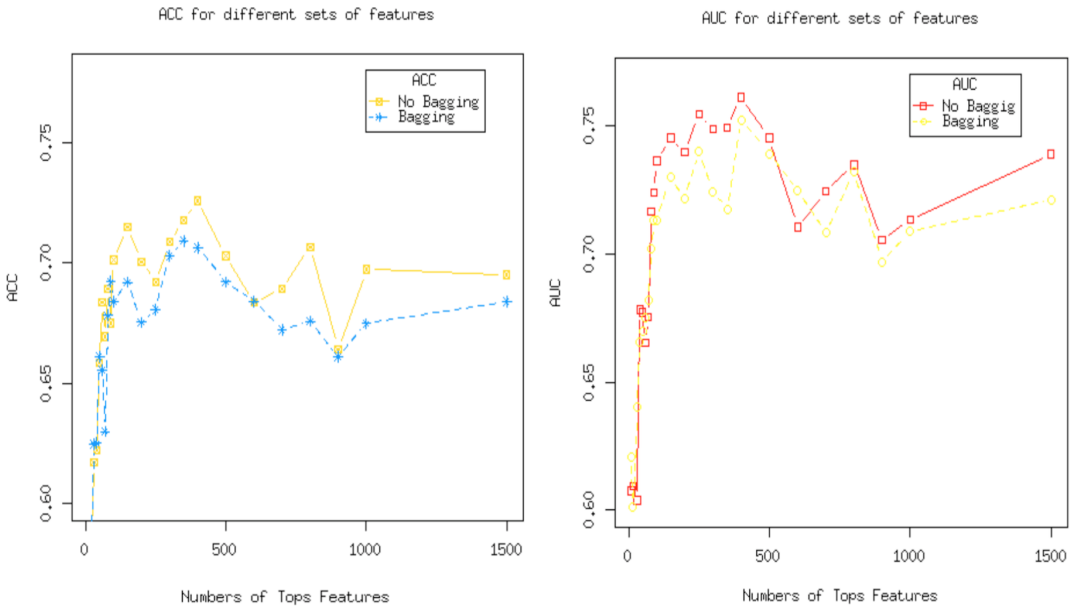

Figure 3.2: AUC and ACC of SVM when doing 10 fold 10 run cross validation on top n BSS/WSS ranked genes. Both AUC and ACC peaked at around 400 univariate ranked gene sets.

3.2.2 Multivariate feature selection

After we obtained the univariate ranked subsets of features, we applied LASSO (Least Absolute Shrinkage and Selection Operator) [49], a multivariate variable selection method, to further reduce the number of selected genes (variables). Figure 3.3 showed the 24 selected genes, corresponding Entrez Gene IDs and regression coefficients from LASSO.

	Gene Symbol	Entrez Gene ID	Coefficient
1	HIST1H2BB	3018	-0.095496
2	MAGEA6	4105	-0.025902
3	C6orf120	387263	0.2172099
4	MINOS1	440574	0.0042863
5	PPP1R27	116729	-0.056971
6	F2RL1	2150	0.0099738
7	RASGEF1C	255426	-0.097773
8	TINCR	257000	0.1424246
9	IFNL2	282616	0.0033698
10	IHH	3549	-0.048717
11	NEFL	4747	0.953549
12	GSKIP	51527	-0.26119
13	TMEM74B	55321	0.37289
14	TASP1	55617	0.000307
15	SLURP1	57152	-0.13955
16	PURA	5813	0.036951
17	SELK	58515	0.126837
18	RBP1	5947	0.008297
19	RPLP0	6175	0.004103
20	SCARNA16	677781	0.379951
21	SURF2	6835	-0.57031
22	AAGAB	79719	-0.11551
23	CCNH	902	-0.05768
24	ARHGEF17	9828	-0.03122

Figure 3.3: 24 Selected genes and their regression coefficients given by LASSO using the AML microarray data.

3.2.3 Visualizing the 24-gene signature in reduced dimensions

In order to visualize the separation between the relapse versus non-relapse patients across the 24-gene signature, we projected our 119 patient samples onto the subspace constructed by the first three principle components of these 24 genes. We performed principal component analysis (PCA) using the R function "princomp". Figure 3.4 showed a screenshot from a three dimensional visualization generated using ggobi [64]. The red dots and blue dots represented relapse and non-relapse patients respectively. We could see the clear separation when we visualizing our data in this subspace representing the 24-gene signature. Principal component analysis facilitated us to get the general tendency of the data points yet on a lower dimensional and keeping the most information about all the original features.

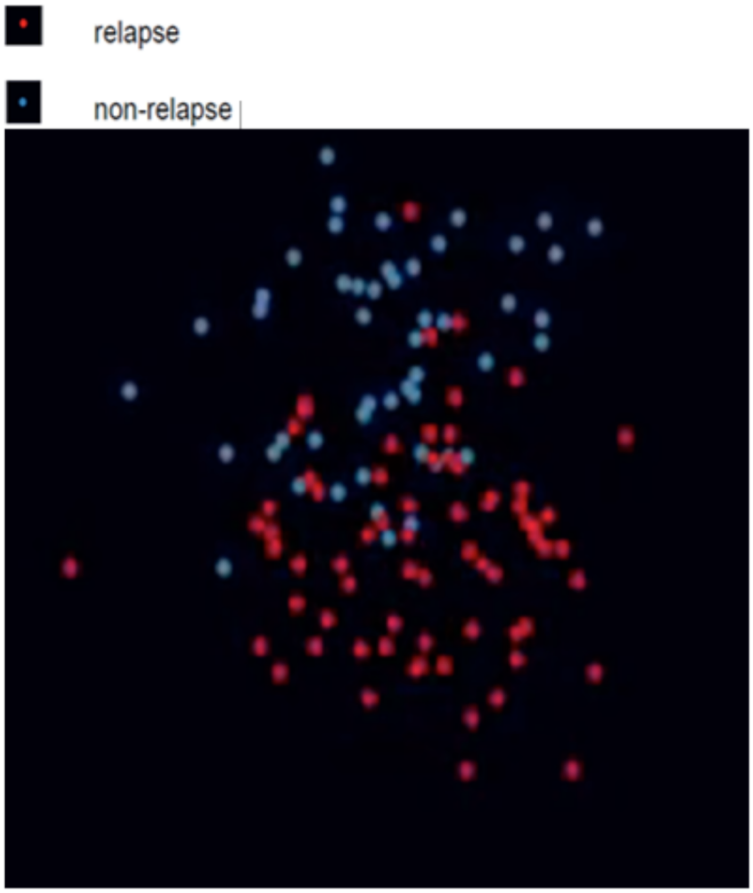

Figure 3.4: Projection of the 24-gene signature onto the first three principle component space using the microarray data. Each data point represented one of the 119 low-risk AML patients. The red dots and blue dots represented relapse and non-relapse patients respectively.

3.3 Comparing Microarray and RNA sequencing data

We experimented with different normalization approaches for the RNAseq (RNA sequencing) data, since it was an essential step for analyzing RNAseq gene expression data [65]. Specifically, we used the Variance Stabilizing Transform (VST) as implemented in the DESeq (differential expressed genes sequencing) Bioconductor package[66]. This normalization approach estimated the size factor, by dividing each sample according to the geometric means of the transcript counts, and used a variance stabilizing transformation (VST) to estimate

the mean-dispersion relationship of data. DESeq also trimmed the lower and upper portions of the data, by log fold changes to minimize the log-fold changes. Our exploratory studies showed that VST yielded normalized gene expression values more consistent with microarray gene expression data when compared to other normalized measures such as RPKM (Reads Per Kilobase of transcript per Million mapped reads). Our observation was in line with other papers in the literature ([65]; [67]).

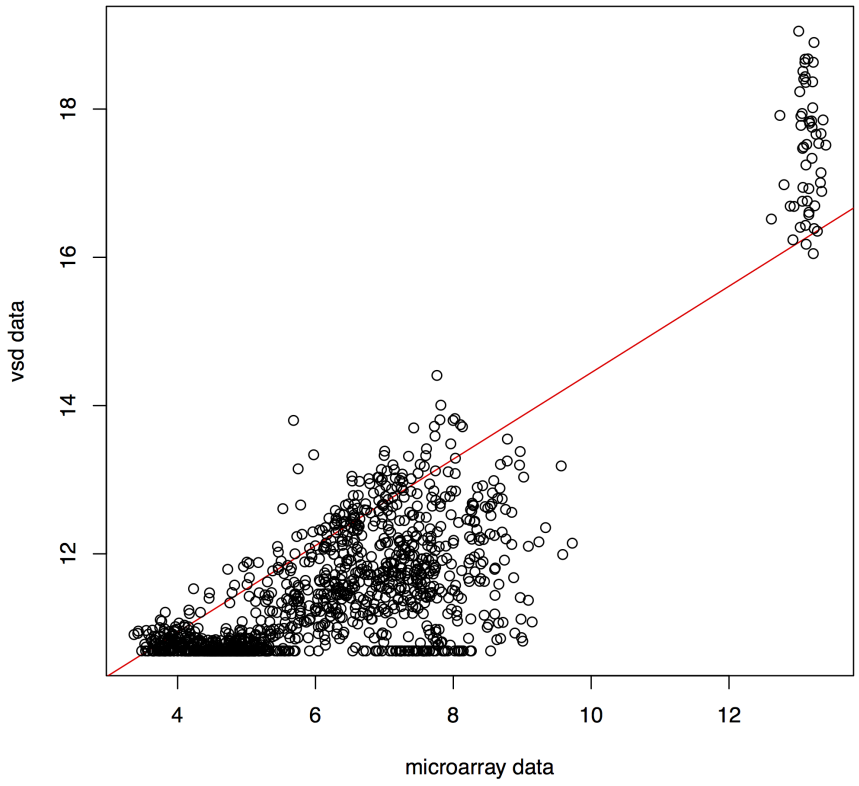

Figure 3.5: A scatter plot comparing the variance stabilizing transformed (VST) RNAseq data (y-axis) and RMA normalized microarray data (x-axis) across the 24 signature genes and 52 overlapping patients.

There were 52 AML patients for which both microarray and RNAseq measurements were profiled. We compared the distributions of microarray data and RNAseq data across these overlapping patients. Figure 3.5 showed a scatter plot comparing the VST normalized RNAseq data to the RMA normalized microarray data across the 24-gene signature in the 52 patients profiled on both microarray and RNAseq. We observed that our normalized RNAseq and microarray were generally concordant, except at extreme low or high expression levels. This effect is expected since extreme values are likely out of the dynamic range of the technology and that thresholding may have been applied. For example, genes like RPLP0 (HGNC: 10371 Entrez Gene: 6175 Ensembl: ENSG00000089157) were highly expressed in both the microarray and RNAseq expression data. While genes like TMEM74B (HGNC: 15893 Entrez Gene: 55321 Ensembl: ENSG00000125895) were lowly expressed in both microarray and RNAseq data. On the other hand, some of the genes such as SLURP1 (HGNC: 18746 Entrez Gene: 57152 Ensembl: ENSG00000126233) were lowly expressed in the microarray data, but expressed at a median level in the RNA-Seq data. Saw Figure 3.6 and Figure 3.7 for detailed comparisons. To summarize, we observed that our signature genes exhibit consistent expression levels in the microarray and RNAseq data.

Figure 3.6 showed the distribution of the microarray data across our 24 signature genes and 52 AML patients. These 52 patients were profiled in both microarray and RNAseq data from TARGET. The green represented the highly expressed genes and the red represented the lowly expressed genes. Those blue and yellow bars, on the left side showed the distribution of relapse and non-relapse patient samples.

Figure 3.7 showed the distribution of the RNAseq data normalized by variance stabilizing transformation. Our 24 signature genes on the overlapping set of 52 patients were shown. As shown in both Figure 3.6 and Figure 3.7, except for the outlier gene RPLP0, the remaining 23 genes clearly separated into three subgroups based on clustering, which increased our confidence in finding biomarkers to distinguish patients between relapse and non-relapse. As mentioned before, for the gene RPLP0 we still need further data from clinical experiments. Yet, we could not illustrate the uniqueness of this specific gene.

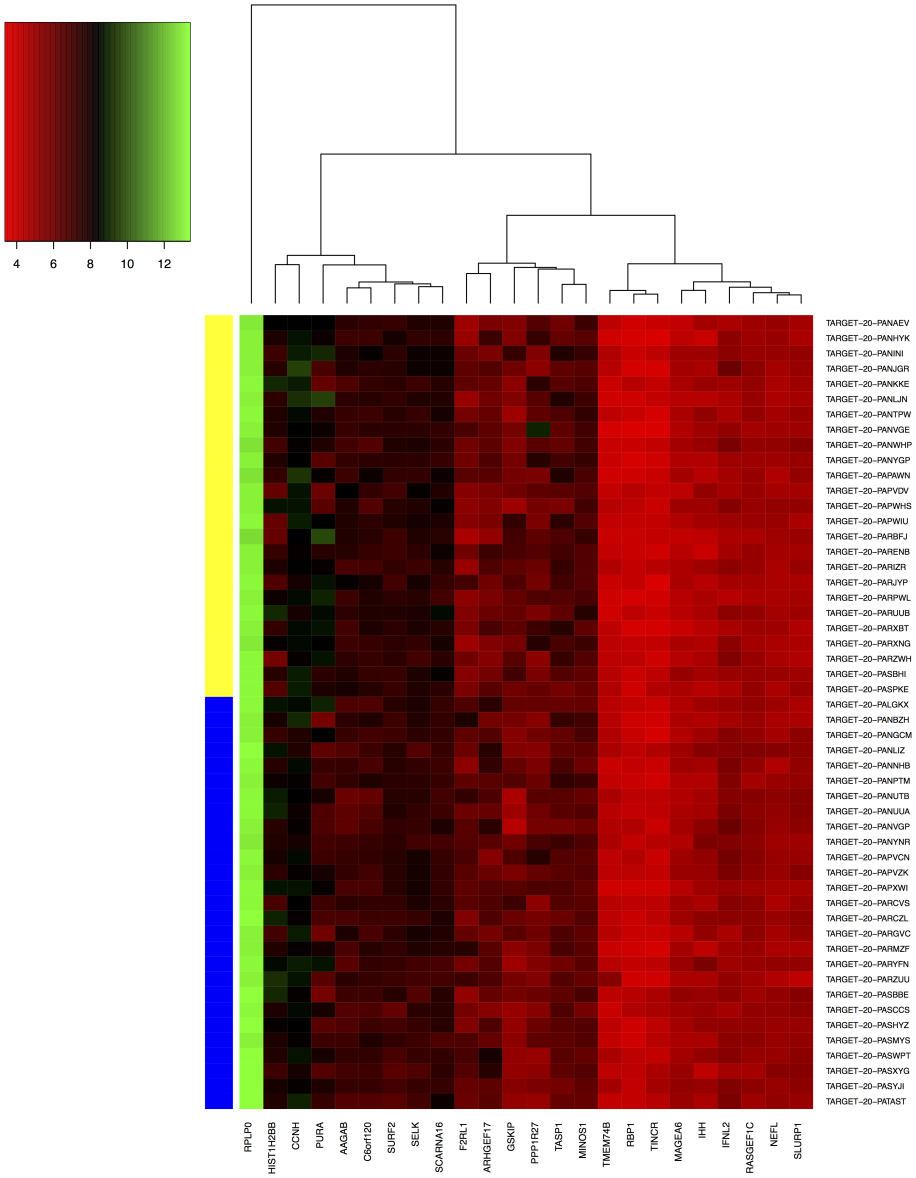

Figure 3.6: A heatmap showed distribution of the microarray data across our 24 signature genes and 52 AML patients. These 52 patients were profiled in both microarray and RNAseq data from TARGET. The green represented the highly expressed genes and the red represented the lowly expressed genes. Those blue and yellow bars, on the left side showed the distribution of relapse and non-relapse patient samples.

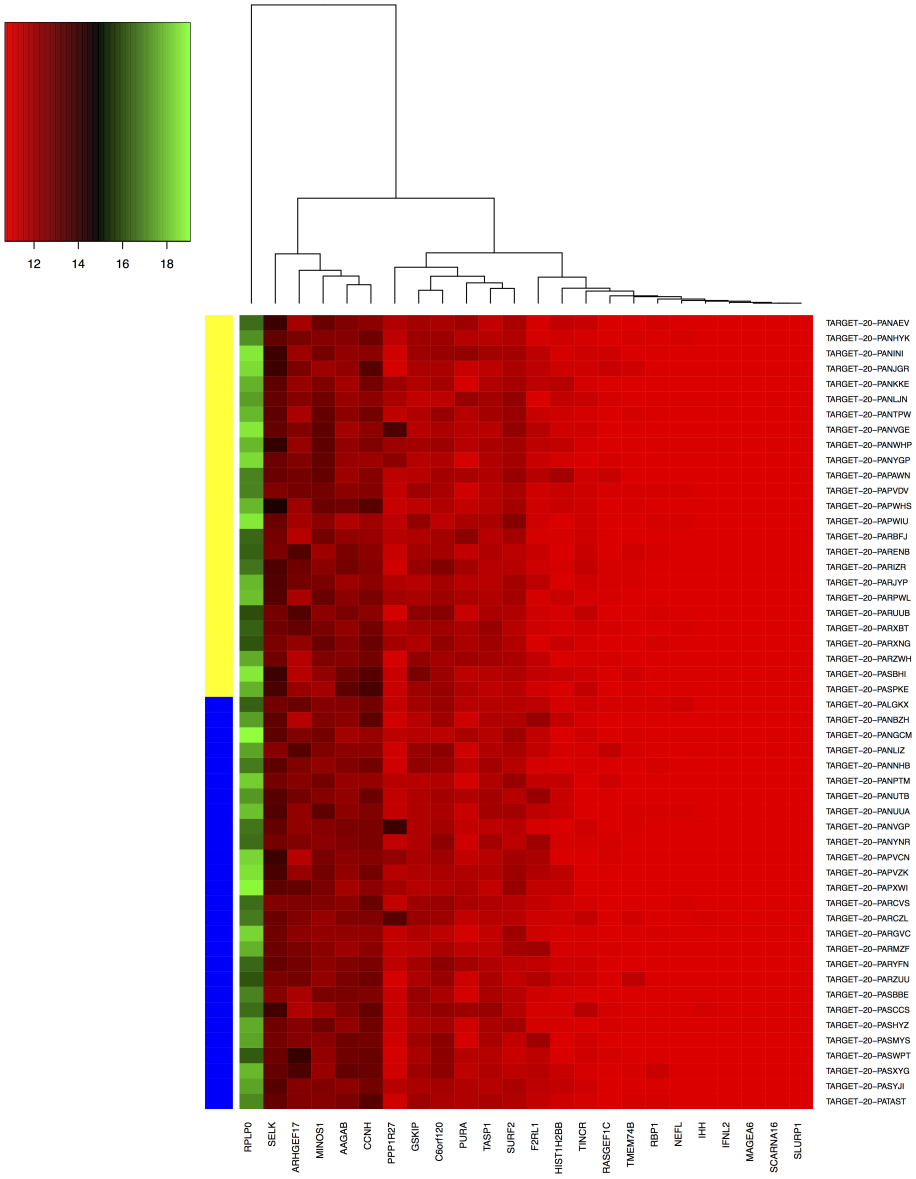

Figure 3.7: A heatmap showed distribution of RNAseq data normalized by variance stabilizing transformation. Our 24 signature genes on the overlapping set of 52 patients were shown.

3.4 Assessment methods

3.4.1 Classification

In our experiments, we adopted an ensemble approach, combining LASSO[49], random forest ([53]; [68]) and support vector machine (SVM)[69]. We used bagging[70] across 10-fold cross validation on different classifiers. Bagging aggregated multiple predictions generated from different sampled subsets of the original training data. It has been shown that bagging can yield substantial improvement in prediction accuracy [70]. In this work, we experimented with averaging and majority voting when aggregating predictions from different classification algorithms. One of the issues with our data was the unbalanced class size. In order to cope with this, we applied SMOTE (Synthetic Minority Technology) [52].

LASSO [49] was used in our study for feature selection. It was a L1 penalized linear regression, with a penalization on the scarcity of predicting feature sets. The objective function of LASSO is convex and had a global optimal. This could guarantee to have the same best feature sets when we reached the global optimal. On top of the selected gene signature, we used two classifiers: random forest as implemented in the "Random Forest" R package and SVM with the radial basis function kernel as implemented in the "e1071" R package. Random forest ([53];[68]) was an ensemble learning method, which aggregated multiple versions of decision trees. In another words, instead of randomly sampling over cases, random forest randomly selected features to construct different decision trees. SVM constructed a hyper plane or set of hyper planes in a high dimensional space such that these hyper planes should achieve the largest separation margin between the two classes. The radial basis function kernel was the most commonly used kernel for performing classification tasks with SVM[69]. Table 3.1, Table 3.2, Table 3.3, Table 3.4, and Table 3.5 showed the predicted results from three different classifiers, the first column was the class label and the rest are the results from lasso, random forest, and SVM.

3.5 Validation

3.5.1 Cross validation using the microarray data

Using the 24 signature genes identified by a combination of univariate and multivariate methods, followed by an ensemble classification method, we achieved high prediction accuracy in cross validation using the microarray data. When we used SVM with the radial basis function kernel, we had achieved an AUC of 0.94. The reason behind this strong predictive power of the model could be seen from Figure 3.4 that projected the 119 data points representing AML low-risk patients on a 3-dimensional space formed by the first three principle components of the 24 signature genes. The clear separation of the two classes (relapsed vs. non-relapsed) of most of the patient samples cased attributes to the selected 24 signature genes. Table 3.1 showed the detailed prediction accuracy in terms of AUC and BAC for different classification methods using all the 119 patients' microarray data from the experiments. In particular, LASSO yielded AUC = 0.95 and BAC=87%, SVM yielded AUC=0.94 and BAC=84%, random forest yielded AUC=0.87 and BAC=80% from Table 3.1.

Table 3.2 showed the detailed prediction accuracy in terms of AUC and BAC for different classification methods using the 52 patients that were profiled in both the microarray data and RNAseq data. We observed similar performance: LASSO yielded AUC=0.85% and BAC=84%, SVM yielded AUC=0.95 and BAC=88%, random forest yielded AUC=0.87 and BAC=77%. To summarize, SVM with the radial basis function kernel showed the strongest predictive capability in 10-fold CV across these 52 patients.

Method	Timing	Result
LASSO (Baseline)	28.53s	AUC = 0.945 BAC = 0.867 F1 Score = 0.845
Random forest	16.85s	AUC = 0.868 BAC = 0.792 F1 Score = 0.753
SVM with the radial basis function kernel	1.256s	AUC = 0.940 BAC = 0.859 F1 Score = 0.837

Table 3.1: Cross validation results with all 119 patients using microarray data

Method	Timing	Result
LASSO (Baseline)	17.74s	AUC = 0.851 BAC = 0.844 F1 Score = 0.857
Random forest	5.9928s	AUC = 0.871 BAC = 0.767 F1 Score = 0.786
SVM with the radial basis function kernel	0.893s	AUC = 0.950 BAC = 0.884 F1 Score = 0.889

Table 3.2: Cross validation results with all 52 patients using microarray data

3.5.2 Independent assessment using the RNA sequencing (RNAseq) data

We evaluated the effectiveness of our 24-gene signatures and classification scheme using RNAseq data consisting of 103 low-risk AML patients. We divided our RNAseq datasets into a training set consisting of 52 patients and an independent test set consisting of 51 patients. These 52 patients were profiled in both the microarray data and RNAseq data.

In Figure 3.4, we observed that the data distribution of the RMA normalized microarray data and variance stabilizing transformed (VST) RNAseq data have clear separated groups, with different means. In order to validate the predictive models inferred using our microarray data, we applied linear regression to re-scale the RNAseq data. Specifically, we fitted a linear regression using the microarray data across the 24-gene signature as the response and the RNAseq data as independent variables. We then used the regression coefficients to transform the RNAseq data before applying the predictive models inferred from the microarray data. Figure 3.8 displayed the data scale before and Figure 3.9 displayed the data scale after the transformation.

Both Figure 3.8 and Figure 3.9 have the same pattern and clusters, the only difference is the scale of legend on the upper left corner. This result was expected and encouraged. We wanted to keep all the information in the original RNA-seq data untouched, only altering the scaling of data.

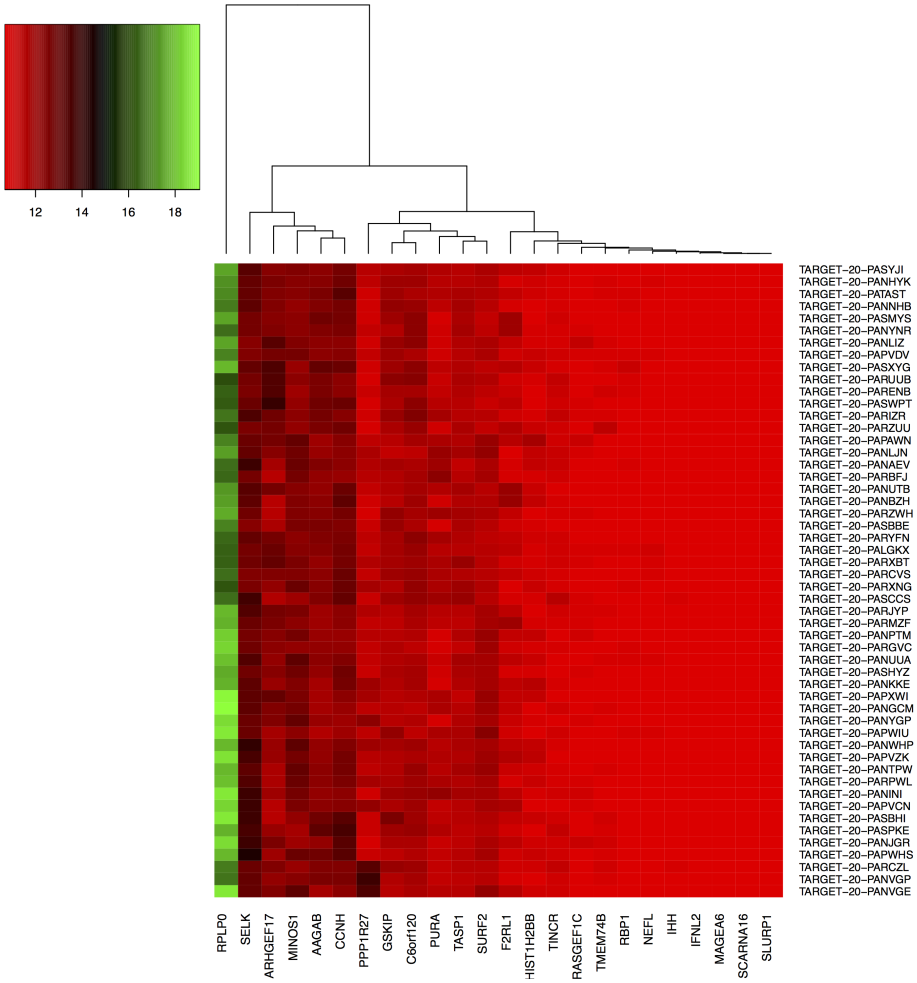

Figure 3.8: A heatmap showed distribution of RNAseq data normalized by variance stabilizing transformation before applied Linear regression based on Microarray data.

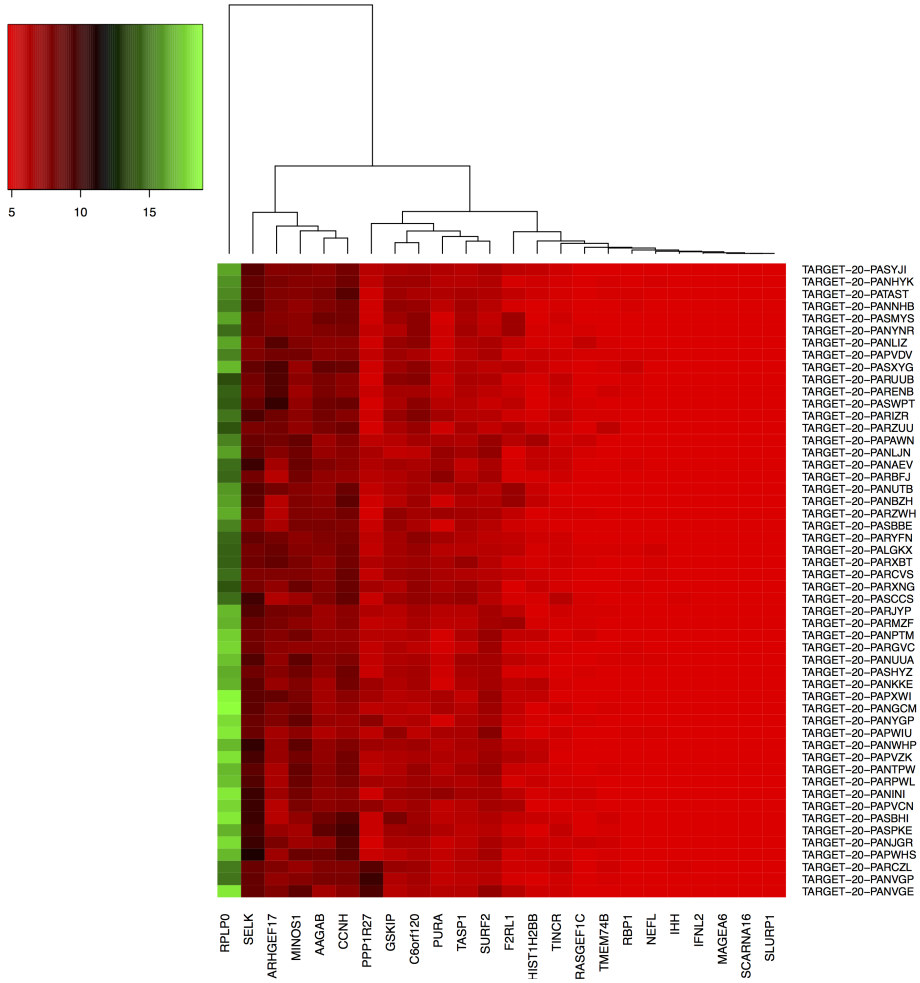

Figure 3.9: A heatmap showed distribution of RNAseq data normalized by variance stabilizing transformation after applied Linear regression based on Microarray data.

Table 3.3 showed the detailed prediction accuracy in terms of AUC and BAC for different classification methods using all 103 patients' RNAseq data from the experiments. Table 3.4 showed the detailed prediction accuracy in terms of AUC and BAC for different classification methods using only 52 patients' RNAseq data from the experiments.

We used Table 3.3 and Table 3.4 to show that RNA-seq data also has relatively high prediction accuracy with the 24-gene signature. In particular, as shown in Table 3.3, LASSO yielded AUC=0.72, SVM yielded AUC= 0.80, and random forest yielded AUC=0.77. We

observed that the prediction accuracy shown in Table 3.4 were lower relative to the corresponding results using the microarray data from Table 3.1 and Table 3.2.

Method	Timing	Result
LASSO (Baseline)	19.72s	AUC = 0.715 BAC = 0.500 F1 Score = 0.778
Random forest	13.79s	AUC = 0.771 BAC = 0.593 F1 Score = 0.726
SVM with the radial basis function kernel	1.176s	AUC = 0.802 BAC = 0.524 F1 Score = 0.730

Table 3.3: Cross validation results with all 103 patients using RNAseq data.

Method	Timing	Result
LASSO (Baseline)	1.16s	AUC = 0.715 BAC = 0.629 F1 Score = 0.689
Random forest	6.170s	AUC = 0.729 BAC = 0.669 F1 Score = 0.712
SVM with the radial basis function kernel	0.877s	AUC = 0.688 BAC = 0.612 F1 Score = 0.655

Table 3.4: Cross validation results with all 52 patients using RNAseq data.

We applied our classification scheme to the transformed RNAseq data. Table 3.5 showed the classification results in terms of AUC and BAC for different classification methods using the RNAseq data consisting of 52 overlapped patients as the test data. Table 3.5 showed that LASSO yielded AUC=0.62, SVM yielded AUC=0.54, and random forest yielded AUC=0.71. When comparing with the results in Table 3.4, there were small difference between them. We observed that in the independent test result of SVM with radial basis function kernel, the F1 score was "NaN", which stands for not a number. The explanation is that the F1 score is computed by both precision and recall, which are highly related to TP (True positive). When TP is 0 (all the positive samples were mislabeled), the formula of F1 score would output "NaN".

Method	Timing	Result
LASSO (Baseline)	0.356s	AUC = 0.618 BAC = 0.500 F1 Score = 0.684
Random forest	0.201s	AUC = 0.711 BAC = 0.582 F1 Score = 0.712
SVM with the radial basis function kernel	0.019s	AUC = 0.543 BAC = 0.500 F1 Score = NaN

Table 3.5: Results based on trained models with all 119 patients with microarray data and independent test with overlapped 52 patients with RNA-seq data.

Chapter 4

PERSONALIZED MEDICINE FOR ACUTE MYELOID LEUKEMIA (AML): CORRELATING DRUG SENSITIVITY AND MUTATION DATA

4.1 Introduction: Correlating drug sensitivity and mutation data

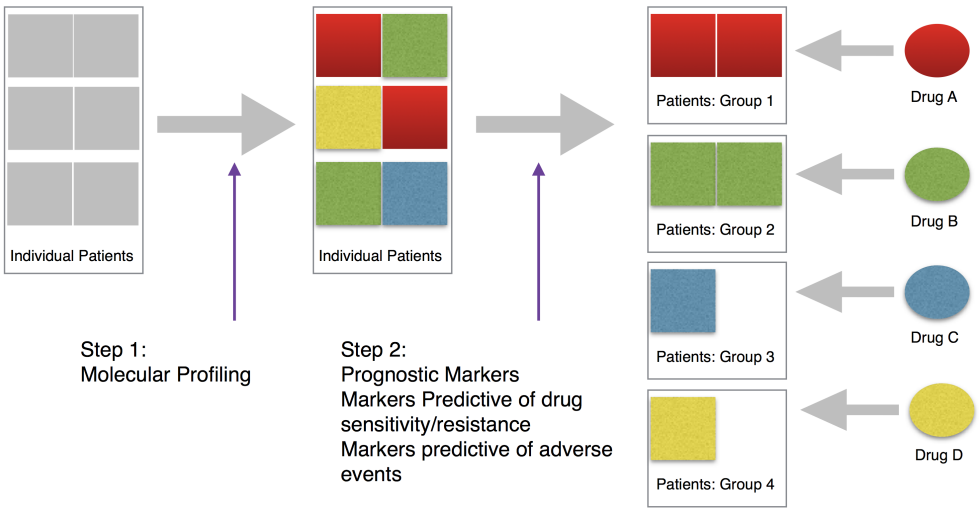

Figure 4.1: Personalized cancer therapy is a treatment strategy centered on the ability to predict which patients are more likely to respond to specific cancer therapies.

The central dogma of molecular biology states that DNA makes RNA and RNA makes proteins. In other words, sequence variations drive differences in phenotypes. In the case of personalized medicine, we expect DNA changes drive drug sensitivity response, individual prognosis and response to therapy. In addition, patient genetic factors can be associated with drug metabolism, drug response and drug toxicity [4]. Therefore, we could combine individual tumor molecular profiles, and other patient characteristics with our knowledge and

resources related to specific drugs to determine optimum individualized therapy options.
In this chapter, we seek to define molecular information that might better predict response to conventional or novel therapies. Specifically, we correlated drug sensitivity and mutation data provided to us from the Becker Lab, and identified 129 genes containing missense mutations with an allelic frequency >5% across at least 2 patients.
Our work was inspired by Klco *et al.*, who used a genomic approach to identify prognostic risk factors for adult AML patients. Specifically, they performed whole-genome or exome sequencing on 71 AML patients who were treated with standard induction therapy [71]. They reported heterogeneity in the mutations and chromosomal translocations associated AML from their sequencing data. They also found several mutations that correlated with prognosis, yet the prediction of chemotherapy drug sensitivity based on genomic data still has room for improvement.
Instead of using whole-genome or exome sequencing on samples[71], our data was generated by MyAML^{TM}. In order to identify gene mutations that drive drug sensitivity, we computed the Pearson's and Spearman's correlations between pairs of genes containing missense mutations and the in vitro cytotoxicity response across the 24 patients. The correlation analyses revealed significant associations (p= 0.006 to 0.04) between indel mutations in three genes and chemosensitivity to drugs commonly used in AML such as cladribine, clofarabine, cytarabine, daunorubicin, etoposide, fludarabine and mitoxantrone. Similarly, significant associations ($p < 0.05$) were identified between missense mutations in 5 genes and chemosensitivity to these drugs.

4.2 Analysis of drug sensitivity data

In this chapter, we used two major datasets. The first one is the AML drug sensitivity data as described in section 2.2.2. The data was generated from the Becker Lab measuring the level of cytotoxicity of 24 patient samples in response to approximately 160 drugs over various concentrations [72]. We interpolated the drug sensitivity data using the procedure described in Section 2.3.2. The other data is the AML gene mutation data described in

section 2.2.3. This dataset was generated by the Becker Lab using MyAML^{TM}, which used next genration sequencing (NGS) to analyze the 3' and 5' UTR and exonic regions of 196 genes known to have recurrent mutations in AML.

4.2.1 Heatmap visualization

Figure 4.2, Figure 4.3, Figure 4.4 showed the heatmaps of the interpolated drug sensitivity at three different concentrations across all the 160 drugs and 24 patients. The entries in the heatmap represented percentage of survival (see legend on the upper left corner). The patient samples are represented in the columns, and the drugs are represented as rows. The green color indicates survival close to 100%, which suggested that the give drug did not kill the cells. Conversely, the red color indicates that many cells are killed by the given drug.
Based on advice from Dr. Becker, we focused on the following drugs that are currently used to treat AML patients, including: Azacitidine, Clofarabine, Cladribine, Tosedostat, Cytarabine, Daunorubicin, Decitabine, Etoposide, Flavopiridol, Fludarabine, Vinblastine sulfate, Bortezomib, Hydroxyurea, Lenalidomide, Mitoxantrone, and Dexamethasone. Figure 4.5, Figure 4.6, and Figure 4.7 showed the heatmaps of the interpolated drug sensitivity at three different concentrations across all the drugs selected by Dr. Becker and 24 patients.
As mentioned before, we started by visualize the overall pictures on the three concentrations of drug sensitivity data. Figure 4.2 and Figure 4.4 showed that at concentrations $10^{-6}, and 3 * 10^{-7}$ M, only the bottom one-third area were in red color, which indicated cell actually dead after applying drugs. Figure 4.3 showed that at concentration 10^{-7} M (i.e. the lowest concentration), fewer drugs were effective based on our patients' samples.

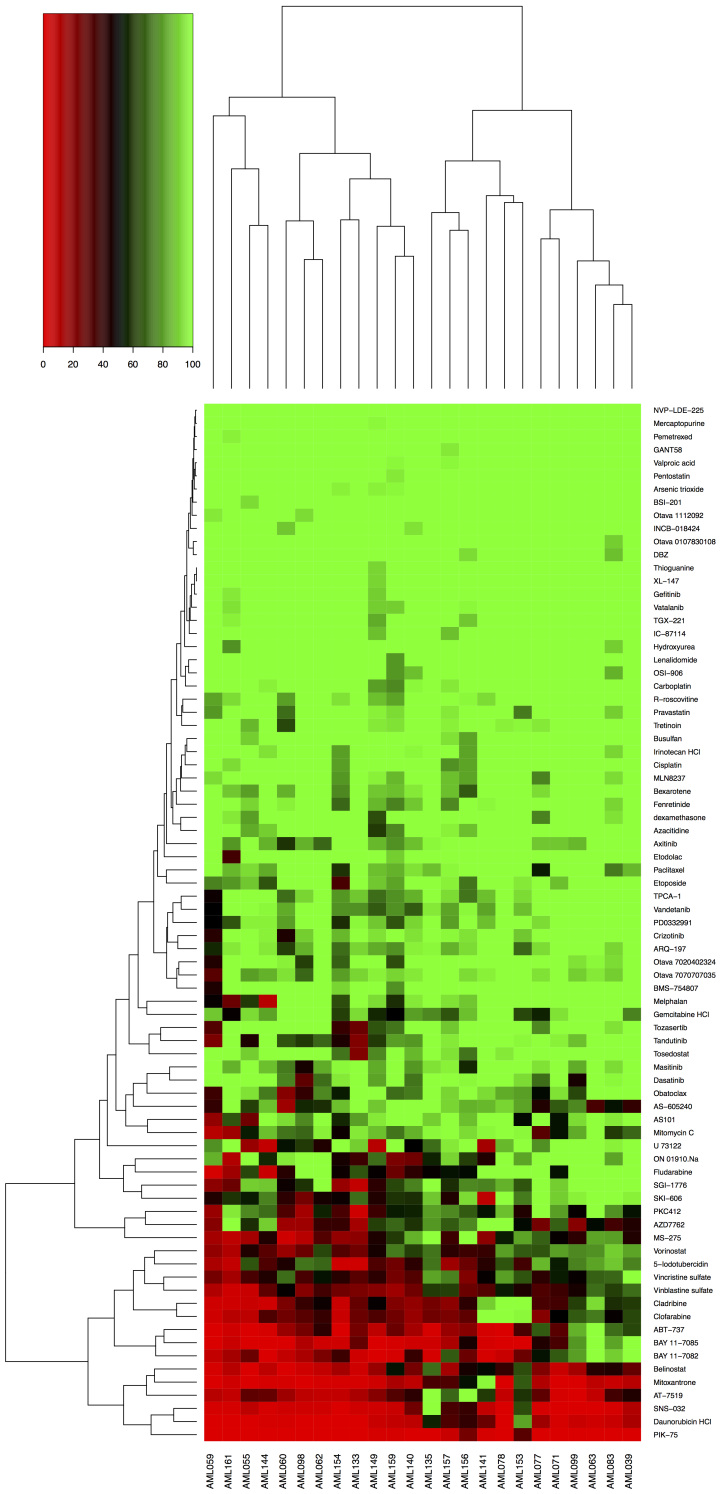

Figure 4.2: A heatmap of Drug Chemosensitivity across all 160 drugs and 24 patients at concentration 10^{-6}.

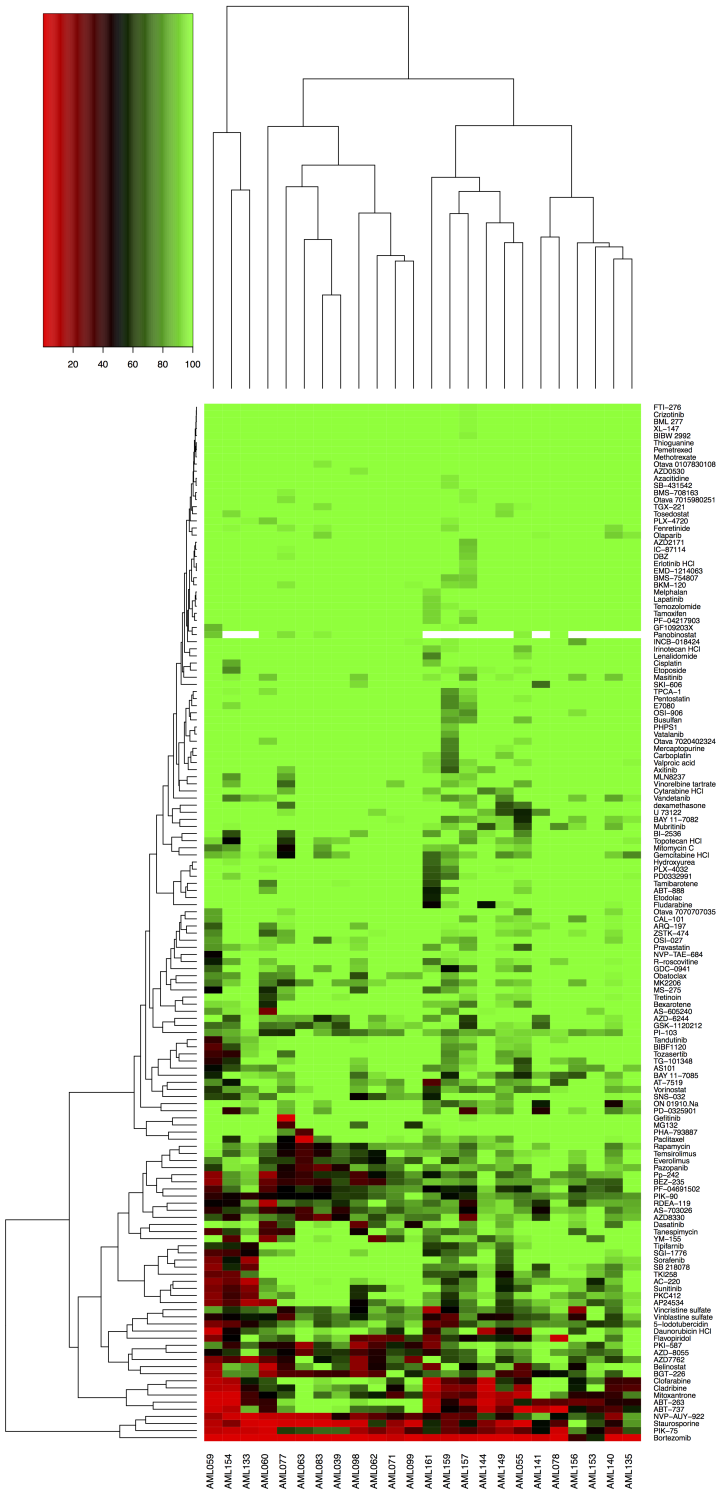

Figure 4.3: A heatmap of Drug Chemosensitivity across all 160 drugs and 24 patients at concentration 10^{-7}.

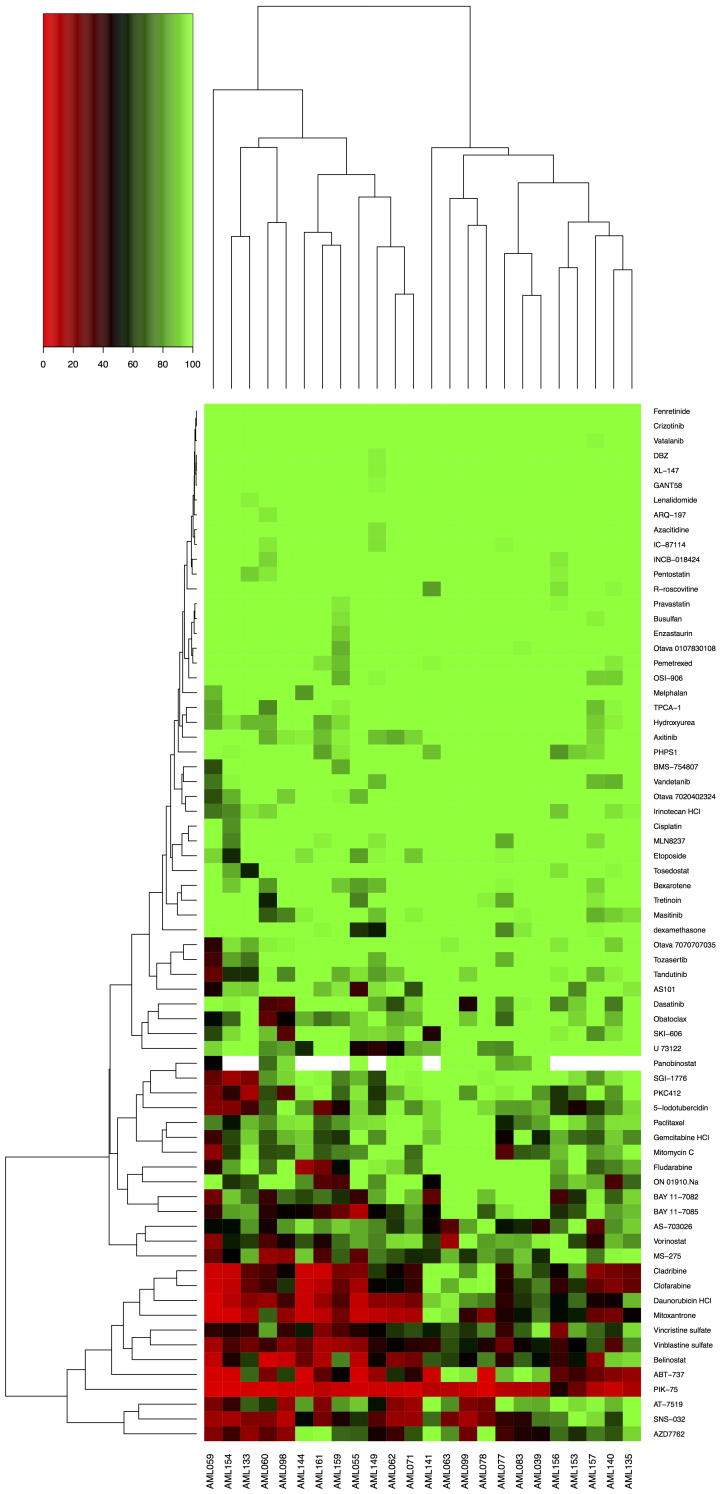

Figure 4.4: A Heatmap of Drug Chemosensitivity across all 160 drugs and 24 patients at concentration $3 * 10^{-7}$

As shown in Figure 4.5, Figure 4.6, and Figure 4.7, we observed that there are five drugs (Clofarabine, Cladribine, Daunorubicin, Vinblastine sulfate, and Mitoxantrone) that showed effectiveness across all three concentrations and over half of patients.

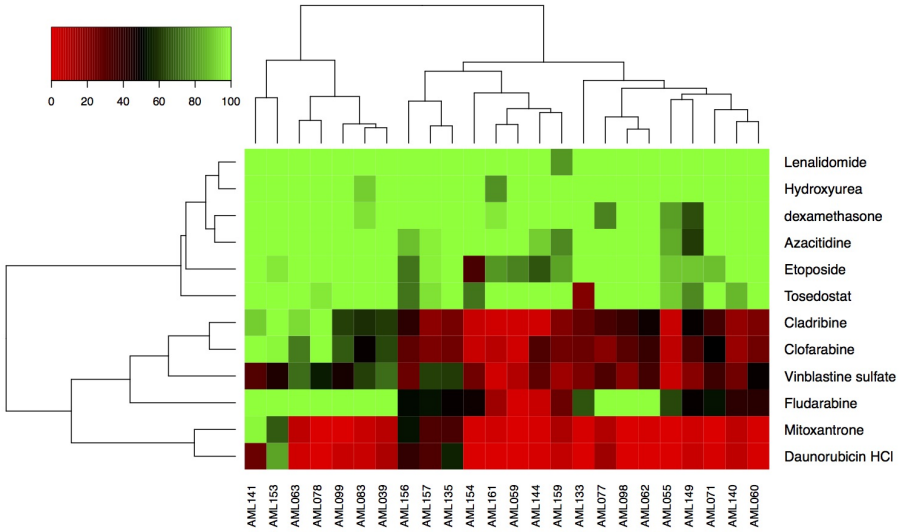

Figure 4.5: A heatmap of Drug Chemosensitivity across common leukemia drugs and 24 patients at concentration 10^{-6}.

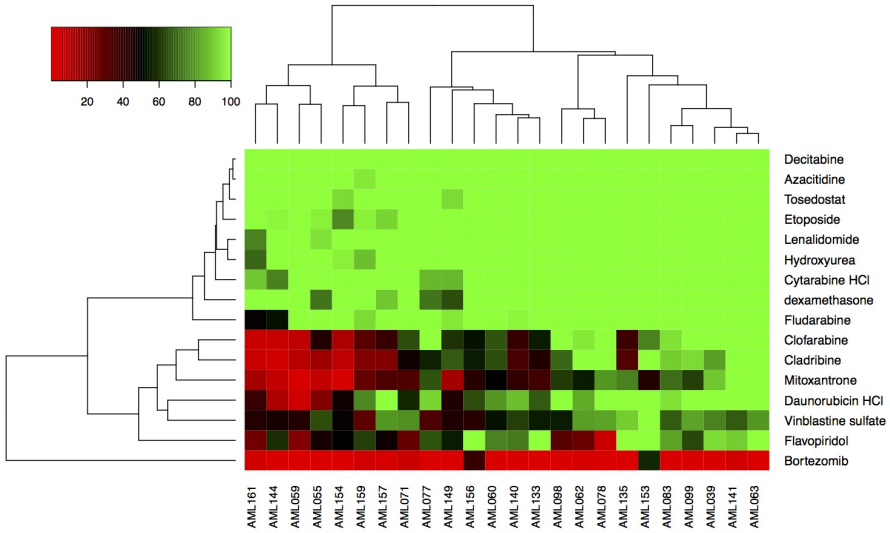

Figure 4.6: A heatmap of Drug Chemosensitivity across common leukemia drugs and 24 patients at concentration 10^{-7}.

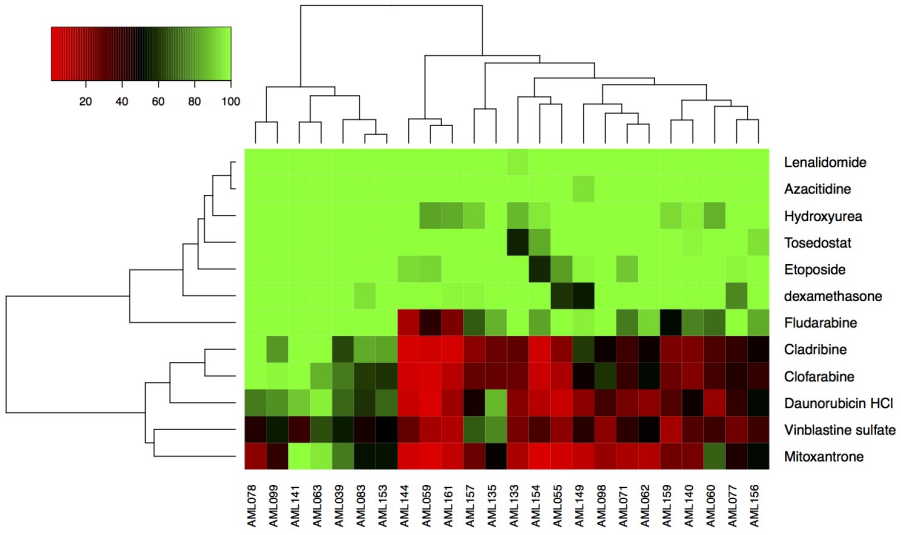

Figure 4.7: A heatmap of Drug Chemosensitivity across common leukemia drugs and 24 patients at concentration $3 * 10^{-7}$.

4.3 Correlation analyses of drug sensitivity data and mutation data

4.3.1 Heatmaps on mutation data

Figure 4.8 is a visualization of the AML gene mutation data. In this heatmap, blue indicated the presence of the mutation, yellow indicated that the mutation was not present. We filtered out missense mutations with less than 5% frequency and also those mutations that occurred in only one of the patients. After applying both filters, 150 genes were included. From Figure 4.8, we observed that approximately half of these mutation genes showed mutations across most of the 24 patients.

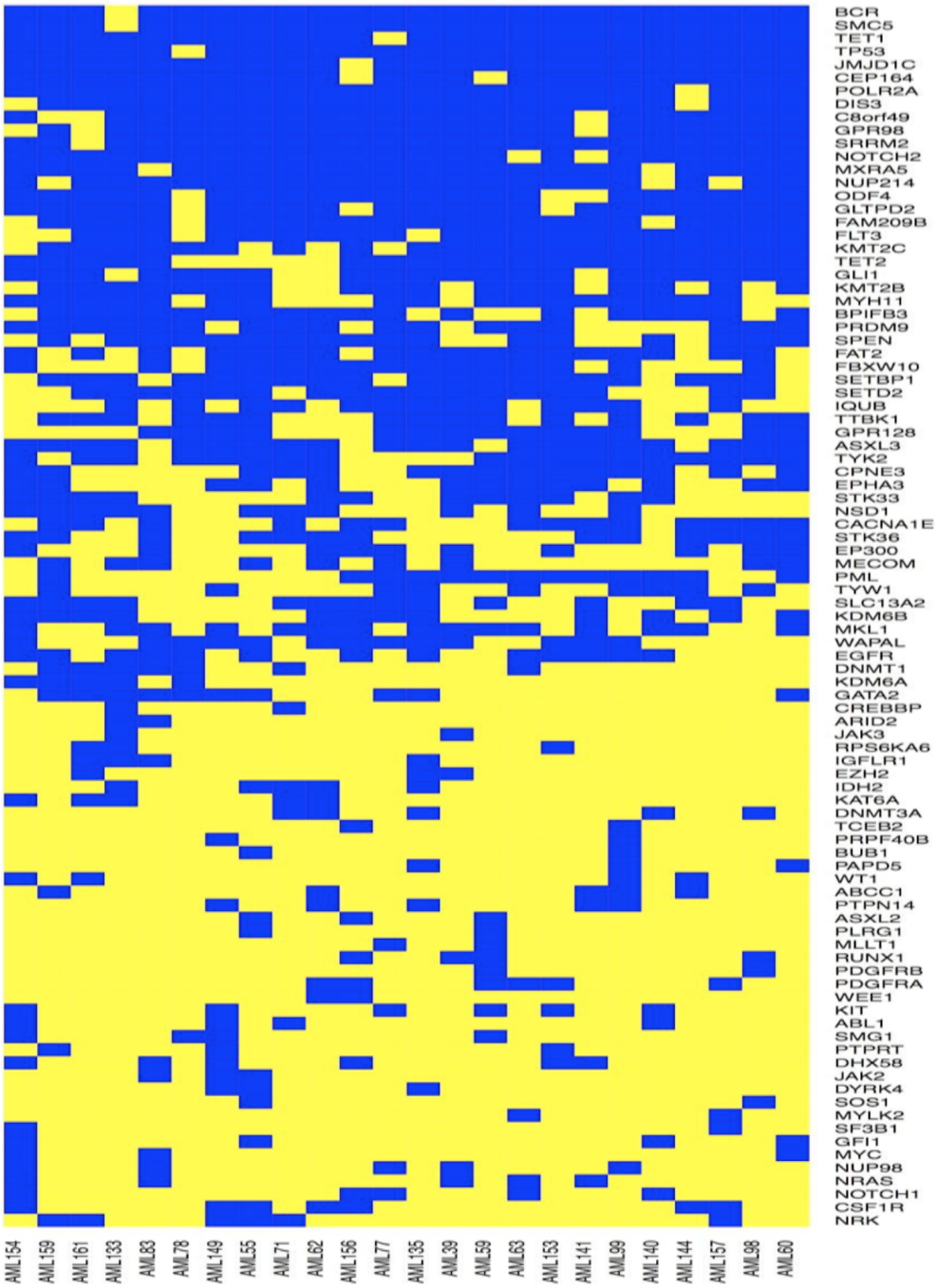

Figure 4.8: A heatmap of genes containing missense mutations that were present in at least 2 patient samples.

Next, we applied pairwise correlation coefficients methods, including Pearson's correlation coefficient described in Section 2.3.4.1, and rank-based Spearman's correlation as described in Section 2.3.4.2 to identify missense mutations that drive the observed drug sensitivity. Figure 4.9 showed the p-values from the correlations between drug sensitivity and mutation. We used a p-value threshold of 0.005 to identify significant (mutation, drug) pairs.

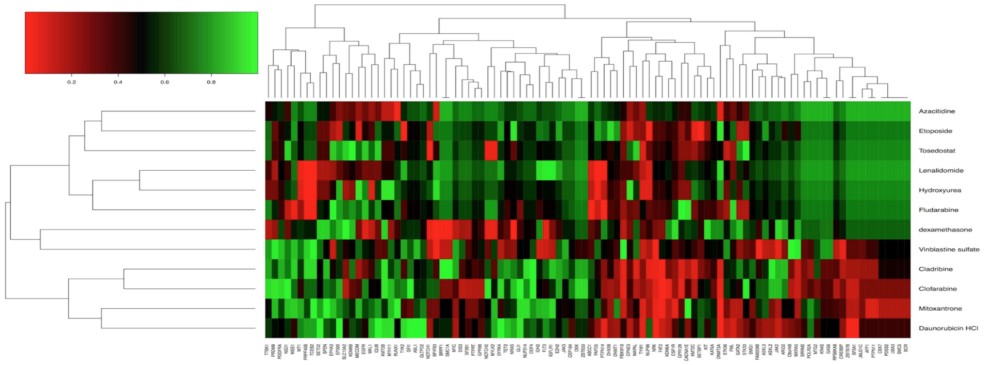

Figure 4.9: A heatmap of computed p-values between drug sensitivity and mutation at concentration 10^{-7} using the Pearson's correlation.

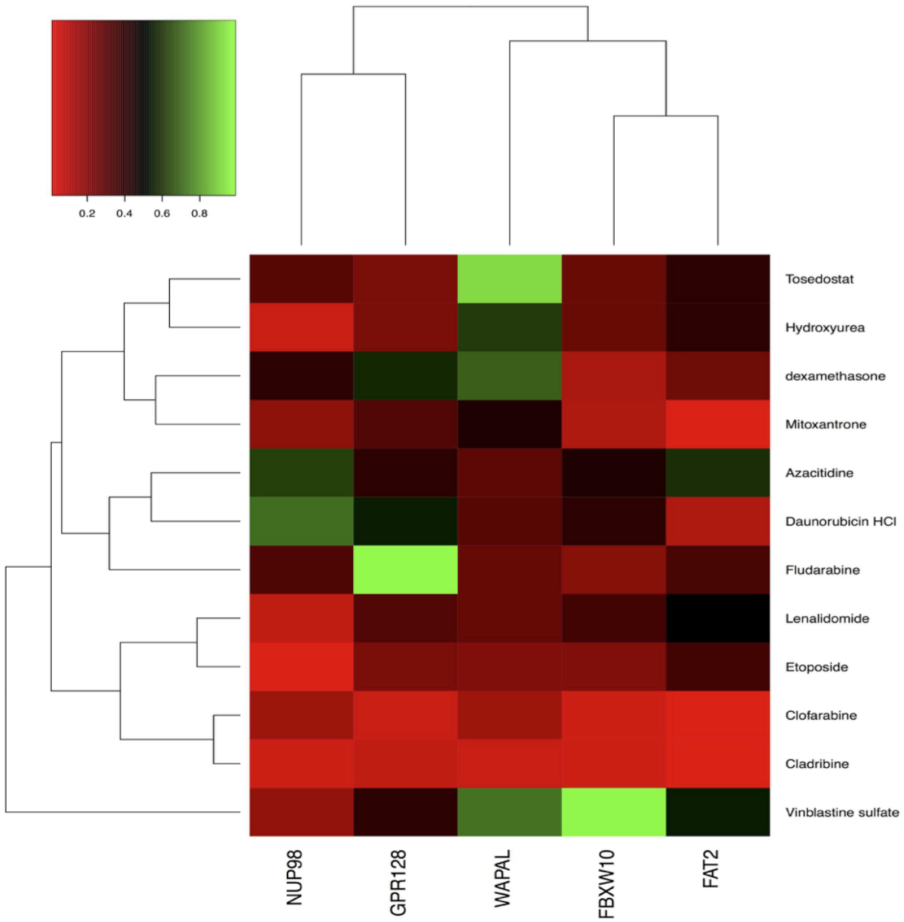

Figure 4.10: A heatmap of computed p-values between drug sensitivity and mutation at concentration 10^{-7} using Pearson's correlation (significant pairs).

After computing all three concentrations based on pearson's and spearsman's correlation, we have found 39 missense mutation genes that are consistent with the findings from Klco *et al.* [71]. The full list is: DNMT3A, IDH1, FLT3, NRAS, IDH2, PTPN11, GATA2, KRAS, TP53, KIT, RUNX1, SRSF2, TET2, CACNA1E, SETBP1, SF3B1, SOS1, DHX32, EGFR, EPHA2, EPHA3, EZH2, FANCC , GLI1, GPR128, JMJD1C, MYC, NOTCH1, OBSCN, PDS5B, PKD1L2, PRDM16, PTPN14, SMC3, TET1, U2AF1, WEE1, WNK4, and WT1. We have also found missense mutation genes that are not mentioned in the findings from Klco *et al.* [71]. For example, we have found mutation gene "NOTCH2", which showed

significant correlations with two of the selected drugs: 0.0274 with Tosedostat, and 0.0504 with Dexamethasone. These were promising (gene, drug) pairs, because both our and Klco *et al.* [71] observed that "NOTCH1" is a mutation gene. Both "NOTCH1" and "NOTCH2" genes are in Notch (DSL) proteins, which are a family of transmembrane proteins with repeated extracellular EGF domains and the notch (or DSL) domains. It is likely that "NOTCH1" and "NOTCH2" share biological features. As another example, we found a mutation gene "FAT2", which showed significant correlations with three selected drugs: 0.012 with Clofarabine, 0.015 with Cladribine, and 0.018 with Mitoxantrone. As shown in Figure 4.1, these (gene, drug) pairs we have found could serve as candidate predictors of drug sensitivity or resistance.

Chapter 5

CONCLUSIONS

5.1 Gene signatures predictive of relapse in AML patients

In this project, we have found a 24-gene signature which can be used to distinguish relapse versus non-relapse low-risk AML patients in the reduced dimensional space using microarray data. These 24 genes demonstrated high prediction accuracy on both the microarray data and RNAseq data. We built predictive models using three machine learning algorithms (LASSO, SVM, random forest) using microarray data, and then validated them using independent RNAseq data. We achieved the highest accuracy using support vector machine (SVM) using these 24 signature genes in 10-fold cross validation. We also noted potential over-fitting in using all samples in the microarray data to select these signature genes. Subsequently, we validated our models using RNAseq data across 52 patients that were profiled in both the microarray and RNAseq data.

5.2 Correlating drug sensitivity and mutation data

Personalized data derived from a targeted genomic assay and in vitro chemotherapy sensitivity testing of individual patient AML samples will likely lead to innovation in treatment, identification of novel targeted agents, and improved outcomes. However, we will need additional experiments and clinical trials to confirm our computationally derived hypotheses between individual mutations on gene sequence and drug sensitivity.

5.3 Future work and contributions

We applied statistic learning procedures to high-throughput data in order to interpret clinical outcomes. We confined ourselves to a subset of low risk AML patients. This subset of patients

is considered to exhibit relatively homogeneous gene expression data, thus more suitable for the inference of statistical learning models. However, as larger numbers of patient samples and more detailed clinical data are collected, we hope to develop some more generic models, which can cope with the massive gene expression data regardless of its homogeneity or heterogeneity. The recent uprising of deep neuron networks might be an intriguing direction for given quantitative interpretation for clinical empirical data.

In our current project, we computed correlations between drug sensitivity data, and mutation data consisting of 194 known genes related to AML. One possibility to extend our work is to integrate our analyses with additional data sources, including the genome-wide next generation sequence data published by Klco *et al.* [71] and the AML gene expression data published by The Cancer Genome Atlas (TCGA). We can also extend our work to other cancers for which data are publicly available in TCGA.

BIBLIOGRAPHY

[1] Leroy H. ; Stephen H. F. "Predictive, personalized, preventive, participatory (P4) cancer medicine". In: *Nature Reviews Clinical Oncology* 8 (2011), p. 184.

[2] Smith F. O. "Personalized medicine for AML?" In: *Blood* 116 (2010), pp. 2622–3.

[3] Waltz E. "FDA tows personalized line: Food and Drug Administration's Paving the Way for Personalized Medicine: FDA's Role in a New Era of Medical Product Development". In: *Nature Biotechnology* 32 (2014), p. 10.

[4] Ignacio I. W.; Juri G. G.; Jörg J. J.; Suzanne E. D. ; Roy S. H. "Methodological and practical challenges for personalized cancer therapies". In: *Nature Reviews Clinical Oncology* 8 (2011), pp. 135–141.

[5] Samir M. H.; Christina S. B. ; Olli K. "Emerging molecular biomarkers—blood-based strategies to detect and monitor cancer". In: *Nature Reviews Clinical Oncology* 8 (2011), pp. 142–150.

[6] Laurence Z.; Oliver K. ; Guido K. "Immune parameters affecting the efficacy of chemotherapeutic regimens". In: *Nature Reviews Clinical Oncology* 8 (2011), pp. 151–160.

[7] Marco A. P.; Elena T.; Tiziana N.; Sabrina P. ; Silvana P. "Targeted therapy in GIST: in silico modeling for prediction of resistance". In: *Nature Reviews Clinical Oncology* 8 (2011), pp. 161–170.

[8] Fabrice A.; Lisa M. M.; Stefan M.; David F. R.; Douglas G. A.; Jorge S. R.; Daniel F. H. ; Lajos P. "Biomarker studies: a call for a comprehensive biomarker study registry". In: *Nature Reviews Clinical Oncology* 8 (2011), pp. 171–176.

[9] Vivien M. "Biology: The big challenges of big data". In: *Nature* 498 (2013), p. 255.

[10] Mattmann C. A. "A vision for data science: to get the best out of big data, funding agencies should develop shared tools for optimizing discovery and train a new breed of researchers". In: *Nature* 493 (2013), p. 473.

[11] "Bioinformatics; Reports on Bioinformatics from EMBL-EBI Provide New Insights". In: *Computer Weekly News* (2013), p. 161.

[12] Andrej Y. Y.; Klebanov L. B.; Lev B.; Daniel G. "Statistical methods for microarray data analysis methods and protocols". In: *Humana Press : Springer* (2013), pp. 1–13.

[13] Licatalosi D. D. ; Darnell R. B. "RNA processing and its regulation: global insights into biological networks.(APPLICATIONS OF NEXT-GENERATION SEQUENCING)". In: *Nature Reviews Genetics* 11 (2010), p. 75.

[14] Zhong W. ; Mark G. ; Michael S. "RNA-Seq: a revolutionary tool for transcriptomics". In: *Nature Reviews Genetics* 10 (2009), p. 57.

[15] Guttman M. et al. "Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals". In: *Nature* 458 (2009), 223–227.

[16] Jemal A. ; Siegel R. ; Ward E. ; Murray T.; Xu J. ; Thun M. J. "Cancer statistics, 2007". In: *CA: a cancer journal for clinicians* 57 (2007), pp. 43–66.

[17] Kulasingam V. ; Diamandis E. P. "Strategies for discovering novel cancer biomarkers through utilization of emerging technologies". In: *Nature Clinical Practice Oncology* 5 (2008), p. 588.

[18] Nguyen D. V. ; Rocke D. M. "Tumor classification by partial least squares using microarray gene expression data". In: *Bioinformatics* 18 (2002), pp. 39–50.

[19] Dettling M. "BagBoosting for tumor classification with gene expression data". In: *Bioinformatics* 20 (2004), pp. 3583–3593.

[20] Gokmen Z. ; Ferhan E. ; Ahmet O. "Bagging Support Vector Machines for Leukemia Classification". In: *International Journal of Computer Science Issues* 9 (2012), pp. 355–358.

[21] Christos S. ; Martine J. P. "Taking gene-expression profiling to the clinic: when will molecular signatures become relevant to patient care?" In: *Nature Reviews Cancer* 7 (2007), p. 545.

[22] Guyon I. ; Weston J. ; Barnhill S. ; Vapnik V. "Gene Selection for Cancer Classification using Support Vector Machines". In: *Machine Learning* 46 (2002), pp. 389–422.

[23] Pirooznia M. ; Yang J. Y. ; Yang M. Q. ; Deng Y. "A comparative study of different machine learning methods on microarray gene expression data". In: *BMC Genomics* 9 (2008), S13–S13.

[24] Tan A. C. ; Gilbert D. "Ensemble machine learning on gene expression data for cancer classification". In: *Applied bioinformatics* 2 (2003), S75–83.

[25] Dettling M. ; Bhlmann P.R. "Boosting for tumor classification with gene expression data". In: *Bioinformatics* 19 (2003), pp. 1061–1069.

[26] Yeung K. Y. ; Gooley T. A. ; Zhang A.; Raftery A. E.; Radich J. P. ; Oehler V. G. "Predicting relapse prior to transplantation in chronic myeloid leukemia by integrating expert knowledge and expression data". In: *Bioinformatics* 28 (2012), pp. 823–830.

[27] Sijia H.; Cameron Y.; Travers C.; Herbert Y.; Lana X. G. "A Novel Model to Combine Clinical and Pathway-Based Transcriptomic Information for the Prognosis Prediction of Breast Cancer". In: *PLOS* 10 (2014).

[28] *Food and Drug Administration*. accessed on 2015/03/13. URL: `http://en.wikipedia.org/wiki/Food_and_Drug_Administration`.

[29] William F.; Donna P. "Life-Saving Cancer Drugs Not Approved by the FDA". In: *Life Extension Magazine* (2007).

[30] Raynaud F.I.; Orr R.M.; Goddard P.M.; et al. "Pharmocokinetics of G3139, a phosphorothioate oligodeoxyneucleotide antisense to bcl-2, after intravenous administration or continuous subcutaneous infusion to mice". In: *J Pharmacol Exp Ther* 281 (1997), pp. 420–427.

[31] O'Brien S.; Moore J.O.; Boyd T.E.; et al. "Randomized Phase III Trial of Fludarabine Plus Cyclophosphamide With or Without Oblimersen Sodium (Bcl-2 antisense) in Patients With Relapsed or Refractory Chronic Lymphocytic Leukemia". In: *J Clin Oncol* (2007).

[32] *High-throughput screening*. accessed on 2015/03/13. URL: `http://en.wikipedia.org/wiki/High-throughput_screening`.

[33] Sliwoski G. ; Kothiwale S. ; Meiler J. ; Lowe E. W. "Computational methods in drug discovery". In: *Pharmacological reviews* 66 (2014), pp. 334–395.

[34] *computer-aided drug design*. accessed on 2015/03/13. URL: `http://en.wikipedia.org/wiki/Drug_design`.

[35] *Acute myeloid leukemia*. accessed on 2015/01/27. URL: `http://en.wikipedia.org/wiki/Acute_myeloid_leukemia`.

[36] *TARGET/TCGA*. accessed on 2015/01/29. URL: `https://tcga-data.nci.nih.gov`.

[37] Bolstad B. M ; Irizarry R. A ; Strand M ; Speed T. P. "A comparison of normalization methods for high density oligonucleotide array data based on variance and bias". In: *Bioinformatics* 19 (2003), pp. 185–193.

[38] Zhong W. ; Mark G. ; Michael S. "RNA-Seq: a revolutionary tool for transcriptomics". In: *Nature Reviews Genetics* 10 (2009), p. 57.

[39] Yeung K. Y.; Blau C.A.; Oehler V.G.; Lee S.I.; Miller C.P.; Chien S.; Martins T.; Estey E.H; Becker P.H. "Personalized Approach To Treatment of Acute Myeloid Leukemia Using a High-Throughput Chemosensitivity Assay". In: *Blood* 122 (2013), p. 21.

[40] Brian E. *The Cambridge Dictionary of Statistics*. 1998 Cambridge, UK ; New York : Cambridge University Press.

[41] Huber W.; von H. A.; Sueltmann H. ; Poustka A.; Vingron M. "Parameter estimation for the calibration and variance stabilization of microarray data". In: *Statistical Applications in Genetics and Molecular Biology* 2 (2003), pp. 2752–2758.

[42] Peter Y. C.; Paula M. P. *Correlation : parametric and nonparametric measures.* ©2002 Thousands Oaks, Calif. : Sage Publications.

[43] Puth M.T. ; Neuhäuser M. ; Ruxton G. D. "Effective use of Pearson's product-moment correlation coefficient". In: *Animal Behaviour* 94 (2013), pp. 183–189.

[44] Puth M.T. ; Neuhäuser M. ; Ruxton G. D. "Spearman's rank correlation coefficient". In: *BMJ : British Medical Journal* 349 (2014).

[45] Rutherford A. "Research design and statistical analysis". In: *British Journal of Mathematical Statistical Psychology* 58 (2005), pp. 187–189.

[46] W. N. ; B. D. Ripley Venables. *Modern Applied Statistics with S.* Springer-Verlag., pp. 312–315.

[47] Yeung K. Y. ; Bumgarner R. E. ; Raftery A. E. "Bayesian model averaging: development of an improved multi-class, gene selection and classification tool for microarray data". In: *Bioinformatics* 21 (2005), pp. 2394–2402.

[48] Tibshirani R. "THE LASSO METHOD FOR VARIABLE SELECTION IN THE COX MODEL". In: *Statistics in Medicine* 16 (1997), pp. 385–395.

[49] Tibshirani R. "Regression Shrinkage and Selection via the Lasso". In: *Journal of the Royal Statistical Society* 58 (1996), pp. 267–288.

[50] Friedman J. ; Hastie T. ; Tibshirani R. "Regularization paths for generalized linear models via coordinate descent". In: *Journal of Statistical Software* 33 (2010), pp. 1–22.

[51] Yeung K. Y. ; Gooley T. A. ; Zhang A. ; Raftery A. E. ; Radich J. P. ; Oehler V. G. "Predicting relapse prior to transplantation in chronic myeloid leukemia by integrating expert knowledge and expression data". In: *Bioinformatics* 28 (2012), pp. 823–830.

[52] Chawla N. V. ; Bowyer K. W. ; Hall L. O. ; Kegelmeyer W. P. "SMOTE: Synthetic Minority Over-sampling Technique". In: *Journal Of Artificial Intelligence Research* 16 (2002), pp. 321–357.

[53] Breiman L. "Random Forests". In: *Machine Learning* 45 (2001), pp. 5–32.

[54] *Support vector machine*. accessed on 2015/01/30. URL: `http://en.wikipedia.org/wiki/Support_vector_machine`.

[55] Robin X. ; Turck N. ; Hainard A. ; Tiberti N. ; Lisacek F. ; Sanchez Jean C. ; Müller M. "pROC: an open-source package for R and S+ to analyze and compare ROC curves". In: *BMC Genomics* 12 (2011), p. 77.

[56] Ambroise C. ; Mclachlan G. J. "Selection Bias in Gene Extraction on the Basis of Microarray Gene-Expression Data". In: *Proceedings of the National Academy of Sciences of the United States of America* 99 (2002), pp. 6562–6566.

[57] Cancer Genome Atlas Research Network. "Genomic and Epigenomic Landscapes of Adult De Novo Acute Myeloid Leukemia". In: *The New England Journal of Medicine* 368 (2013), pp. 2059–2074.

[58] Tarlock K. ; Meshinchi S. "Pediatric acute myeloid leukemia: biology and therapeutic implications of genomic variants". In: *Pediatric clinics of North America* 62 (2015), pp. 75–93.

[59] Bullinger L. ; Döhner K. ; Bair E. ; Fröhling S. ; Schlenk R.F ; Tibshirani R. ; Döhner H. ; Pollack J. R. "Use of Gene-Expression Profiling to Identify Prognostic Subclasses in Adult Acute Myeloid Leukemia". In: *The New England Journal of Medicine* 350 (2004), pp. 1605–1616.

[60] Ein D.L. ; Kela I. "Outcome signature genes in breast cancer: is there a unique set?" In: *Bioinformatics* 21 (2005), pp. 171–178.

[61] Sahar B. W. ; Boer J. M ; Beverloo H. B. ; Moorhouse M. J ; Van D. S. P. J. ; Löwenberg B. ; Delwel R. Valk P. J. M. ; Verhaak R. G. W. ; Beijen M. A. ; Erpelinck C. A. J. ; Van D. K. "Prognostically Useful Gene-Expression Profiles in Acute Myeloid Leukemia". In: *The New England Journal of Medicine* 350 (2004), pp. 1617–1628.

[62] Chibon F. "Cancer gene expression signatures - the rise and fall?" In: *European journal of cancer (Oxford, England : 1990)* 49 (2013), pp. 2000–9.

[63] Bacher U. ; Schnittger S. ; Haferlach T. "Molecular genetics in acute myeloid leukemia". In: *Current opinion in oncology* 22 (2010), pp. 646–55.

[64] Lawrence M. ; Wickham H. ; Cook D. ; Hofmann H. ; Swayne D. "Extending the GGobi pipeline from R". In: *Computational Statistics* 24 (2009), pp. 195–205.

[65] Bullard J. H. ; Purdom E. ; Hansen K. D. ; Dudoit S. "Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments". In: *BMC Bioinformatics* 11 (2010), pp. 94–94.

[66] Anders S. ; Huber W. "Differential expression analysis for sequence count data". In: *Genome Biology* 11 (2010), R106–R106.

[67] Dillies M. A. ; Rau A. ; Aubert J. ; Hennequet A. C. ; Jeanmougin M. ; Servant N. ; Keime C. ; Marot G. ; Castel D. ; Estelle J. ; Guernec G. ; Jagla B. ; Jouneau L. ; Laloë D. ; Le Gall C.; Schaëffer B. ; Le C. S. ; Guedj M.; Jaffrézic F. "A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis". In: *Briefings in Bioinformatics* 14 (2013), pp. 671–683.

[68] Prasad A. M. ; Iverson L. R. ; Liaw A. "Newer Classification and Regression Tree Techniques: Bagging and Random Forests for Ecological Prediction". In: *Ecosystems* 9 (2006), pp. 181–199.

[69] Cortes C. ; Vapnik V. "Support-vector networks". In: *Machine Learning* 20 (1995), pp. 273–297.

[70] Breiman L. "Bagging Predictors". In: *Machine Learning* 24 (1996), pp. 123–140.

[71] Klco J. M. ; Link D. C. ; Radich J. P. ; Westervelt P. ; Mardis E. R.. ; Graubert T. A. ; Walter M. J. ; Welch J. S. ; Ozenberger B. A. ; Kulkarni S. ; Larson D. E. ; Demeter R. T. ; Magrini V. ; Fronick C. ; Miller C. A. ; Heath S. E. ; Ketkar K. S. ; Lamprecht T.L. ; Baty J. ; Griffith M.; Petti A. ; Spencer D. H. ; Griffith O. L. ; Dong S. ; Hundal J. ; Gue S. C. ; Fulton R. ; O'Laughlin M. ; Dipersio J. F. ; Helton N. M. ; Wilson R. K. ; Wartman L. D. ; Christopher M. ; Ley T. J. ; Duncavage E. J. ; Payton J.

E. "Association between mutation clearance after induction therapy and outcomes in acute myeloid leukemia". In: *The Journal of the American Medical Association* 811 (2015), pp. 321–357.

[72] *Meshinchi Lab*. accessed on 2015/01/29. URL: `http://www.seattlechildrens.org/research/about/`.