Using biomedical data to identify genetic variants that drive drug responses in Acute Myeloid Leukemia

Nicole Kauer

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

Master of Science

University of Washington Tacoma

2020

Committee:

Ka Yee Yeung-Rhee, Chair

Ling-Hong Hung

Wes Lloyd

Program Authorized to Offer Degree:
Computer Science and Systems

University of Washington Tacoma

Abstract

Using biomedical data to identify genetic variants that drive drug responses in Acute Myeloid Leukemia

Nicole Kauer

Chair of the Supervisory Committee:
Professor Ka Yee Yeung-Rhee
School of Engineering and Technology

Acute Myeloid Leukemia (AML) is a heterogeneous cancer of the blood that progresses quickly, with approximately 10,000 AML related deaths reported annually in the United States. Patients with AML tend to have genetic variations, which can significantly affect drug sensitivity and treatment outcomes. While massive amounts of big biomedical data have been generated to characterize AML, these genetic variations are not yet well-understood. Thus, the development of individualized approaches to AML therapy using these big data has great potential.

The promise of precision medicine is that knowledge of the genetic characteristics present within a cancer will enable better choices for therapy. In this thesis, we applied data science techniques to analyze AML biomedical data in collaboration with Dr. Pamela Becker, Hematology, UW-Seattle, with the goal of identifying genetic variants that drive drug responses in AML. We identified 30 novel gene-drug pairs with statistical significant responses. Additionally, both univariate and multivariate machine learning models were created, with multivariate feature selection via Bayesian Model Averaging. We found multivariate models to outperform univariate models in most cases. Additionally, an automated workflow for the analysis was created to allow for incremental additions of patient data over the course of Dr. Becker's study.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

GLOSSARY

AML: Acute Myeloid Leukemia, a cancer that affects an individual's blood.

DNA: Deoxyribonucleic acid; the material that comprises an individual's genetic makeup. Composed of nucleotides, denoted A, C, G, and T, that form gene base pairs. [27]

CODON: Groups of three nucleotides that encode the sequence of the synthesized protein [27]

GENETIC VARIANT: A variation within the DNA encoding of a gene; also known as a mutation. [27]

INDEL VARIANT: Insertions or deletions of nucleotides that may affect how genes are translated into proteins. [27]

IN-FRAME INDEL: An in-frame indel inserts or deletes a sequence which has a length that is a multiple of 3. This results in an insertion or deletion of a set of amino acids in the translated protein and a possible change of the codons at the site of deletion or the junction of the insertion. The rest of the gene sequence is still read in the same frame and the translated protein is unchanged after the indel. [27]

FRAMESHIFT INDEL: Occurs when the length of the insertion or deletion is not a multiple of 3. The translation reading frame of the gene sequence downstream of the indel is altered, resulting in wholesale changes in the amino acid sequence of the translated protein after the indel. In addition, frameshift indels often result in truncated proteins due to a premature stop codon arising from the change in reading frame. [27]

MISSENSE VARIANT: Altering of a single base pair, resulting in the encoding of a different amino acid when translating the gene into a protein. [20]

ACKNOWLEDGMENTS

I would like to thank Dr. Pamela Becker for providing the data and medical advice that was vital for this thesis. I would also like to thank my committee for their support and guidance. Not least, I wish to extend my heart-felt gratitude to Dr. Ka Yee Yeung for recruiting me to the BioSummer School, thus setting the course for my dream career.

DEDICATION

For those who believed in me when I did not believe in myself.

Chapter 1

INTRODUCTION

1.1 Background and Motivation

Acute Myeloid Leukemia (AML) is a cancer that affects blood cell development. AML starts in the bone marrow and may move to other organs [40]. While AML is a rare disease, nearly 21,500 new cases and 11,000 deaths were expected in 2019 [41]. AML patients often have a significant number of gene mutations, with thirteen that are the most common [6] and shown in Table 1.1. A DNA mutation is a genetic variation that changes the hereditary material of life [37]. These genetic abnormalities and their interactions create a wide variety of AML sub-types. As a result, AML is an extremely heterogeneous disease and remains one of the most difficult malignancies to treat. [6]

The prognosis for a patient can depend on many factors, one of the main being the type of gene mutations or combinations of gene mutations present. For example, a mutation in NPM1 has potential for a better outcome than having a mutation in both NPM1 and DNMT3A. [7]. For many years, a standard method to treat AML was to use a combination of chemotherapy drugs that depended on the patient's heart condition. Additional treatment was added if the patient had an FLT3 mutation or if there was CD33 protein present in the AML cells [39]. Figure 1.1 depicts this standard treatment as a flow chart. However, this standard treatment did not account for the variation in AML sub-types, nor for the individual differences in how medications affect people.

While the one-size-fits-all approach to treatment had proven helpful for at least a portion of AML patients, a personalized approach has the potential for better outcomes. Given the large variety in gene mutations across AML patients and the altered prognostic depending on these mutations, it is clear that the disease affects patients in different ways. Recent studies

Table 1.1: Most common AML mutated genes[6]

Mutated Gene	% of Patients	Mutated Gene	% of Patients
NPM1	25-35	CEBPA	6-10
RUNX1	5-15	FLT3-ITD	20
KIT	<5	NRAS	15
DNMT3A	18-22	ASXL1	5-17
IDH1	7-14	IDH2	8-19
TET2	7-25	KMT2A-PTD	5
TP53	8		

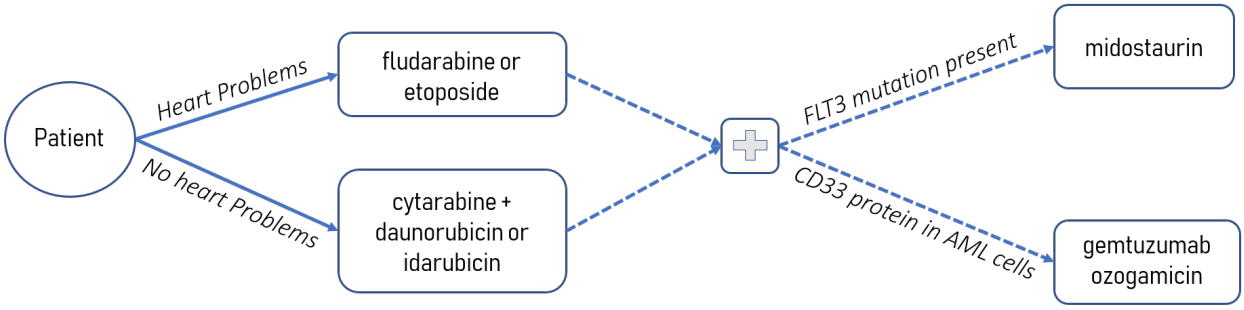

Figure 1.1: Flowchart of the treatment that was standard for many years [39]. Dotted lines represent the optional choices added to the standard treatment.

have shown that there is a correlation between higher drug sensitivity for some drug types and certain gene mutations [9]. In this thesis, we applied statistical methods and developed software tools to study the associations between gene variants and drug sensitivity. We aim to identify gene variants that drive this drug response using both uni-variate and multivariate analysis.

1.2 Contributions

My contributions include the identification of genetic variants that drive drug response in the patient samples, along with the development of a reproducible workflow for data processing and analysis. This will be accomplished through the following tasks:

- Univariate and multivariate analyses for identification of gene mutations that drive drug response.
- Comparison of our results to published work.
- Reproducible analytical workflows for processing AML data, from raw data to univariate and multivariate analyses.

Chapter 2

RELATED WORK

2.1 Mutation Targeting Drugs

Given that prognosis is often dependent on which gene mutations are present, some treatment drugs target these specific mutations. One of the most common mutations is FLT3-ITD [6]. The standard treatment for patients with this variant is midostaurin [39], a drug that inhibits tyrosine kinases, including FLT3 [8]. Table 2.1 shows examples of inhibitors and their corresponding target gene.

However, as previously discussed, different combinations of mutations, such as NPM1 with DNMT3A, can change a patient's prognosis [7]. Additionally, at least one recent study has found that drugs targeting specific mutations do not always result in the desired effect [9].

Table 2.1: Sample of AML drugs for inhibiting specific target mutations [9]

Drug	Target Inhibited
midostaurin, sorafenib	FLT3-ITD
cytosolic, ivosidenib	IDH1
cytosolic, enasidenib	IDH2
venetoclax	BCL2

2.2 Identifying Correlation Between Gene Mutations and Drug Response

One of the challenges in identifying gene mutations that drive drug response is the issue that genes do not act alone, but rather in combination with each other. This makes teasing out the gene mutation with the highest correlation to drug sensitivity difficult. Three recent groups have attempted to address this issue, one being the Leukemia and Lymphoma Society with their Beat AML initiative [9], the second being a collaboration with Dr. Pamela Becker and the University of Washington [25], and the third being a collaboration between the Sanford-Burham Medical Research Institute, Michigan State University, and Moores Cancer Center at the University of California San Diego (Kang et al.) [22]. Additionally, the National Cancer Institute (NCI) worked with pharmaceutical companies in the largest personalized medicine trial in the United States, NCI-MATCH [10, 2].

2.2.1 MERGE: An Algorithm for Incorporating Prior Knowledge to Determine Drivers of Drug Sensitivity

Drs. Su-In Lee, Pamela Becker and colleagues [25] identified correlations between drug sensitivity and gene expression using data generated in the Becker Lab at the University of Washington. They developed, MERGE, a machine learning algorithm that uses drug sensitivity data to learn weights for 5 variables that summarise a gene's data. The Becker Lab contributed genome-wide gene expression data and drug sensitivity data for 30 patients to use in this manuscript.

MERGE combines the rarity of a mutation, expression hubness, regulatory role, genomic copy number variation, and methylation to summarize a given gene. Expression hubness is a measure of how connected the gene is to others and was measured with an algorithm the group developed. Data for each of these came from literature. In vitro drug sensitivity for 53 drugs was incorporated to learn weights for each of these variables, with the weighted MERGE score determining the likelihood that a gene was the driver for drug sensitivity. Lee *et al.* used a False Discovery Rate (FDR) corrected critical p-value of 0.1 to identify

significant correlation between gene expression and drug sensitivity.

Overall, Lee *et al.* found that drug sensitivity for drugs that targeted specific genes did correlate with the differential expression for those same genes. The group also examined the effect of each variable on the final score. Not surprisingly, expression hubness had the highest influence on the merge score. Methylation, which decreases expression, had the second highest influence. Three of the categories showed little to no effect on the MERGE score: regulatory role, mutation rarity, and genomic copy variation, the difference in copies of a gene between individuals. One possible explanation for not seeing a distinct effect of regulatory role on drug sensitivity may have been the fact that expression hubness should have taken into account the regulatory role, thus counting this aspect twice.

2.2.2 Investigation of Personalized Medicine Validity

Kang *et al.*'s [22] research focused on the potential for personalized AML treatment. Their research was broken up into four sections along the spectrum of personalized medicine. In the first stage, they experimented with drug sensitivity in the lab. In the second stage, they created algorithms to find drugs that could work together, and narrowed down the list of potential drugs. In the third stage, they assayed combinations of drugs from stage two. In the final stage, they compared clinical drug sensitivity results to lab drug sensitivity.

In the initial phase of their research, Kang *et al.* used cell lines, human cells that have been grown in a laboratory, to test for drug sensitivity. More specifically, they used KG-1, an AML cell line, along with control cells to test drug sensitivity in terms of how well the drugs only targeted AML cells. This section of their research found that some drugs killed AML cells while leaving normal cells intact with as high a ratio as 100:1.

Kang *et al.* improved standard AML therapy by identifying drug combinations using a network model of AML intracellular signal transduction and literature data. Their approach showed that a three-drug combination could potentially work well together. The three drugs were PD0325901, a MEK inhibitor, Quizartinib, a FLT3 inhibitor, and Palbociclib, a CDK 4/6 inhibitor. In the third stage of their research, they performed 64 experiments with

different combinations of the three drugs on two patient cell lines. The two patient's cells reacted differently with the same combination of drugs, showing the benefit in personalized medicine.

Given the complexity of the human body, lab results do not always match the clinical results; in order to verify the usefulness of personalized medicine based on lab results, Kang *et al.* studied six AML patients undergoing standard treatment and compared the results to a lab experiment with the same standard treatment. The lab results were assessed 72 hours after treatment, whereas the clinical results in the form of residual blasts, a test that checks for leftover cancer cells, were taken at 28 days after initial treatment. Ultimately, they found a correlation of 0.99, showing that there is a high potential for lab results to directly correlate to how a patient's cells will react. Additionally, 72 hours was a quick turnaround for finding a better drug combination than the standard treatment for a given patient.

2.2.3 Beat AML Initiative

Beat AML is an initiative launched in 2013 by the Leukemia and Lymphoma Society (LLS) [26]. Beat AML aims to improve outcomes for AML patients by developing better treatments through understanding the genetic mutations that characterize the disease and targeting drugs to inhibit these mutations.

The Beat AML project is a collaboration between academic researchers and pharmaceutical companies. Massive datasets have been generated in this project. Specifically, Tyner *et al.* reported their findings on clinical and genomic data generated from a cohort of 672 tumor specimens in their key paper published in *Nature* in 2018 [9]. They built on data generated from The Cancer Genome Atlas' (TCGA) whole-genome sequencing for 200 AML patients [32] with their own lab data for 531 patients, spanning whole-exome sequencing, RNA sequencing, and drug sensitivity for 122 inhibitors. Tyner *et al.* found that there were common gene mutations across both sets of data, including many mutations that appear to work together.

Beat AML focused heavily on drug sensitivity, performing a wide variety of experiments.

Table 2.2: Selection of drug sensitivity experiments and results [9]

Drug Sensitivity Experiment	Results
Newly diagnosed AML versus relapsed AML	Relapsed showed less sensitivity.
Drug target versus target gene expression	Mixed results.
Top 20% versus bottom 20% (AUC)	78 of 122 drugs had significant gene expression signatures.
JAK (janus kinase) inhibitors versus target gene expression	BCOR + RUNX1 mutations correlated to increased sensitivity to all JAK inhibitors. BCOR + DNMT3A and BCOR + SRSF2 did not.

A selection of these experiments and some results can be found in Table 2.2. One of the interesting results was the comparison of four JAK (janus kinase) inhibitors to the gene expression for the genes targeted by these inhibitors. They found that a BCOR mutation co-occurring with a RUNX1 mutation correlated to higher drug sensitivity for all JAK inhibitors. However, if a BCOR mutation was accompanied by a mutation in DNMT3A or SRSF2, there was no correlation between drug sensitivity for the same inhibitors. This further shows that personalized, targeted therapy is promising.

2.2.4 NCI-MATCH

The National Cancer Institute (NCI) partnered with pharmaceutical companies to start the ongoing Molecular Analysis for Therapy Choice (NCI-MATCH) study in 2015, which aimed to match cancer patients to treatments, not by malignancy type, but by the genetic landscape of their tumor biopsies [2]. More than 10x the patients than the trial size attempted to join,

showing the demand for personalized medicine [10]. Ultimately, 6,000 patients were screened and tumor biopsies genetically sequenced with a targeted panel of 143 genes [10]. There were 24 arms of the trial, with specific genetic targets matched with drugs, 17 of which can be seen in Table 2.3 [2]. Patients were matched to drugs by which actionable variants were present in the tumor, with some tie-breaker rules in the event that there was more than one actionable variant [10]. 23% of patients screened matched to one of the drugs in the study [2]. Since the study began, the results for 3 arms of the study, I, H, and W, have been released and are summarized in Table 2.4. One thing to note is even though 6,000 patients were screened, approximately a quarter made it through to the study, who were then broken up into the 24 different arms. Of the arms released, the largest number of patients in a group was 65, which is a relatively small number of patients. Arm H had the best results, with the patient's cancer not progressing for 9.4 months. However, only 33% of tumors shrank. Arms I and W had worse results with 27% and 5% overall response rates. [24, 4, 38] The results thus far show that personalized cancer treatment should not be based on a single gene variant being present.

2.2.5 Comparison

All four groups agree that AML is a challenging disease to treat and that personalized drug treatment is the way of the future. There is still more work to be done to make this reality, however. Kang *et al.* [22] showed that gene expression and drug sensitivity were different between different patients, and that clinical therapy results and lab results do correlate. However, the sample size was small and the experiment would need to be repeated on a larger scale. The MERGE algorithm developed by Lee *et al.* [25] was a good start to narrowing down the genes to focus on, but was ultimately a feature selection method that would best be used with other machine learning algorithms in a personalized medicine workflow. Additionally, there was some discrepancy between the Lee *et al.* results when comparing drug sensitivity for specific inhibitors and the target gene expression. Lee *et al.* showed a correlation between these, but Tyner *et al.* had mixed results, with only some inhibitors being correlated. Tyner

Table 2.3: Arms of the NCI-MATCH study enrolling patients in 2018 [2]

Arm	Targeted genetic change	Drugs(s)
A	EGFR mut	Afatinib
C1	MET amp	Crizotinib
C2	MET ex 14 sk	Crizotinib
E	EGRF T790M	AZD9291
F	ALK transloc	Crizotinib
G	ROS1 transloc	Crizotinib
H	BRAF V6000	Dabrafenib + trametinib
L	mTOR mut	TAK-228
M	TSC1 or TSC2 mut	TAK-228
S2	GNAQ/GNA11 mut	Trametinib
T	SMO/PTCH1 mut	Vismodegib
U	NF2 loss	Defactinib
V	cKIT mut	Sunitinib
X	DDR2 mut	Dasatinib
Z1B	CCND1,2,3 amp	Palbociclib
Z1C	CDK4 or CDK6 amp	Palbociclib
Z1E	NTRK fusions	Larotrectinib (LOXO-101)

Table 2.4: Results from NCI-MATCH Arms H, I, W [24, 4, 38]

Arm	Targeted genetic change & Treatment Drug	# Patients	Results
H	BRAF V600E/K mut: Dabrafenib & trametinib	35	33% overall response 9.4 month median progression free survival
I	PI3KCA mut: Talelisib	65	No overall responses 27% 6 month progression free survival 67% tumors with co-occurring mutations
W	FGFR mut: AZD4547	50	5% overall response 17% 6 month progression free survival

et al. [9] showed that there was a relationship between gene expression and drug sensitivity, and most importantly, that there was significant evidence that combinations of mutations could be used to target therapy for an individual AML patient. NCI-MATCH found that the genetic landscape of a given tumor can vary by 10-15% [2], making it difficult to obtain the single best actionable variant. Additionally, results from 3 arms of the study show low overall response rates to the targeted drugs [24, 4, 38].

Chapter 3

METHODS

3.1 Data

3.1.1 Variant Data

The variant data was generated by a targeted next generation sequencing technology called MyAML^{TM} that identifies mutations in 194 genes known to be associated with AML [21]. The data provided by Dr. Becker's lab was broken down for each patient by variant type, and had additional statistics for each gene, including allele frequency (how frequently the variant occurred in the sample), sorting intolerant from tolerant (SIFT) and polymorphism phenotyping (PolyPhen). SIFT is a binary variable that predicts whether an amino acid substitution affects protein function [30] and the PolyPhen score represents the probability of the variant being damaging [17].

3.1.2 Drug Sensitivity Data

High throughput drug sensitivity was performed against a panel of 160 drugs. Patient cells were isolated in the lab and in vitro chemotherapy cytotoxicity testing was performed, with 8 concentrations for each drug. After a 4-day incubation, the cell survival was assessed. The data was plotted in Excel and fitted curves were derived from plots of survival versus drug concentrations. One widely-used variable derived from the fitted curve to measure the drug response is EC50, which is the drug concentration when the cells' survival rate is reduced to halfway between the baseline and maximum [18]. Dr. Becker's lab provided the drug sensitivity data after the EC50 values are computed. In our study, we use these EC50 values to represent a patient's response to a drug.

3.2 Processing

3.2.1 Variant Data

We focused on missense, inframe indel, and frameshift indel mutations. Indels are insertions or deletions of nucleotides (the basic base pair components of DNA) that may affect how genes are translated into proteins. Gene base pairs are translated into the amino acids that make up proteins in groups of three, called codons. An in-frame indel inserts or deletes a sequence which has a length that is a multiple of 3. This results in an insertion or deletion of a set of amino acids in the translated protein and a possible change of the codons at the site of deletion or the junction of the insertion. The rest of the gene sequence is still read in the same frame and the translated protein is unchanged after the indel. A frameshift indel occurs when the length of the insertion or deletion is not a multiple of 3. The translation reading frame of the gene sequence downstream of the indel is altered, resulting in wholesale changes in the amino acid sequence of the translated protein after the indel. In addition, frameshift indels often result in truncated proteins due to a premature stop codon arising from the change in reading frame [27]. Missense mutations are when a single base pair is altered, resulting in the encoding of a different amino acid when translating the gene into a protein [20]. With the advice from Dr. Becker, mutation data from MyAML was filtered with the following criteria:

- Missense, inframe and frameshift indels with allelic frequency >2.5%
- Missense with deleterious SIFT
- Missense with possibly or probably damaging PolyPhen

After applying these filters, we identified 98 genes with mutations across at least two patients.

3.2.2 Drug Sensitivity Data

In the AML drug sensitivity data, we aim to explore drug sensitivity across patients at a given cancer cell survival rate. We did this by focusing on the EC50 concentration, which is the drug concentration when the cancer cell survival rate is 50%. EC50 was calculated by Dr. Becker's lab via XLFit using the concentrations from the lab as the data points to fit the equation. Thresholds were applied to the EC50 data due to extreme outliers. EC50 above 10^{-4} M was considered drug resistance and set to a threshold of 0.1 M, while EC50 below 10^{-12} M was set to 10^{-12} M. The drug data was then scaled via a negative log transform. While all patients were tested with a panel of 160 drugs, the drugs on the panel had changed over time. Ultimately, we received data for 216 total drugs with 101 drugs that had data for all 69 patients. During uni-variate testing, patients without drug sensitivity data for a given drug were left out. Multivariate testing was only done on drugs that had data for all 69 patients. The scaled and transformed drug sensitivity data was used in all experiments and visualizations.

3.3 Visualization

3.3.1 Heatmaps

The heatmaps were created with the R package `ComplexHeatmaps` [13], with clustering by Euclidean distance. Missing drug sensitivity values in Figure 4.2 were set to 0 when clustering to get the row and column ordering. However, the missing values were left as NA and shown in light gray in the figure. Similarly, for the heatmap in Figure 4.5, the missing p-values and estimated mean differences were set to 1 and 0, respectively, only for the purpose of clustering. This clustering was used to order the rows and columns shown in the figure, but with light gray for the missing, NA, values. Additionally, the number of variants in Figure 4.1 were given a threshold of 4 to decrease the effect of outliers on the visual, as well as the number of genes reduced to only include those with at least 5 patients reported with the variant. Figures 4.7 and 4.9 are a little different in that they have annotations for either

rows, columns, or both. In the case of Figure 4.7, the number of non-NA values in the column or row, respectively, were used. For Figure 4.9, the mean of the row, not counting NA values, was used.

3.3.2 Bar Plot and Histograms

The drug sensitivity bar plot in Figure 4.4 was created with the `ggplot2` R package [46] as an `identity` bar plot, where input to the bar plot were the sums of non-NA drug sensitivity data for each patient across each drug. The histograms in Figure 4.11 were created with the `geom_histogram()` function from `ggplot2`. The input values were the differences between the average multivariate and univariate scores.

3.4 Univariate

3.4.1 Two-sided Paired T-Test

We aim to identify variants that are associated with the observed drug response. The Welch two-sided, two sample t-test was used, in which the variance of patients with the mutations and patients without the variants were assumed to be unequal. This test was used to see if the distribution between the patients with and without variants was the same, which would imply that the gene mutation did not affect drug sensitivity. The confidence interval was set at 0.95. For the t-test, the patient data was split between those who have the variant in a given gene and those who do not. The test was limited to only genes that had at least 2 patients but no more than 67 patients with variants. Tests were then run using the built-in R function `t.test()` [33], comparing the drug sensitivity of patients with variants against the patients without. Patients without drug sensitivity data for a given compound were left out of the test. The calculated p-values and estimated mean differences are shown in Figure 4.

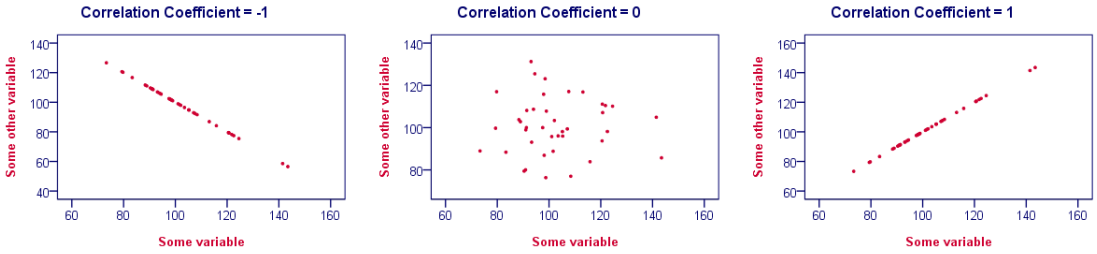

Figure 3.1: Example of datasets that are negatively correlated (left), positively correlated (right), and not correlated (middle). [45]

3.4.2 Pearson Correlation

Pearson correlation shows how well two variables are linearly associated. Figure 3.1 shows examples of what three datasets would look like if they were negatively, positively, or not correlated. The Pearson correlation between drug sensitivity and gene variants was calculated via the R function `cor()` from the `stats` package [33]. Two matrices with data for all 69 patients were used, one matrix with the $-log$(EC50 (M)) drug sensitivity values, and one binary matrix denoting the presence of a genetic variant. Each patient's drug sensitivity vector was correlated with each gene variant vector with the caveat that only data points which were pairwise complete between the two vectors were considered.

3.5 Multivariate

3.5.1 Bayesian Model Averaging

Bayesian Model Averaging (BMA) is an ensemble method generally used for feature selection. Unlike the univariate method described in Section 3.4 that ignores model uncertainty and uses a single gene mutation to predict the drug response, BMA accounts for the uncertainty about the best gene mutation to choose by averaging over multiple models (sets of potentially overlapping relevant gene mutations). In BMA, the predictions from multiple regression models are averaged and weighted by the posterior probability of each model. BMA is a

general framework that has been developed for linear regression, generalized linear regression, and survival analyses [14, 35, 36].

In this thesis, we focus on using linear regression to model the relationships between drug response and gene mutations. Hoeting *et al.* explained BMA for linear regression models in a 1999 tutorial [14]. Given k linear regression models, $M_0, M_1, \cdots, M_k$, Bayes factors can be calculated as a ratio of the probability of the data, D, given the two models using the equation [23],

$$B_{10} = pr(D|M_1)/pr(D|M_0).$$

The Bayes factor can be used to determine the overall posterior probability of a model given the data with the equation,

$$pr(M_k|D) = \frac{\alpha_k B_{k0}}{\sum_{r=0}^{K} \alpha_r B_{r0}},$$

where $\alpha_k = pr(M_k)/pr(M_0)$ for $k = 0, \cdots, K$. The posterior probability for a given gene, denoted as Δ, is then the weighted average over all models [14],

$$pr(\Delta|D) = \sum_{k=1}^{K} pr(\Delta|M_k, D) pr(M_k|D).$$

In this project, BMA was applied to select relevant gene mutations in the following two steps. First, `iterativeBMAlm()` from the `networkBMA` package [11, 48, 47, 28] was used. This algorithm takes the first 30 variables, and iteratively performs BMA using the `bicreg()` function from the `BMA` package [34], removing variables with the lowest posterior probability and adding new variables from the dataset until all variables have been tested. This first model was used to find the set of genes that had posterior probability greater than zero. Next, the `bicreg()` function was applied once again to the subset of selected genes to derive the final BMA linear regression model. The genes that had posterior probability >50% were then used in a linear regression model. In our empirical experiments, 101 drugs were individually tested. Drugs were chosen based on the number of patients that had drug sensitivity data. In this case, there were 101 drugs that had drug sensitivity data for all 69 patients. The predictor variables are the binary (0/1) patient-gene data matrix, where 1

indicated that a patient had a variant in the gene. The response variable was the continuous variable $-log$(EC50 (M)) representing drug response.

3.5.2 Modeling the Relationships between Drug response and Gene Mutations using Linear Regression

Linear regression takes the form of the linear equation

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + ...$$

where the β coefficients are calculated to minimize the squared error across all training data. In this case, the input variables, x_1, x_2, x_3, were the genes and y was the drug sensitivity value, $-log$(EC50 (M)). Linear regression models were created for each for the 101 drugs that had drug sensitivity data for all 69 patients. The gene data was a 0/1 patient-gene matrix, where 1 indicated the patient had a variant in the gene. Only genes that were selected by the BMA model as having posterior probability >50% for the given drug were used in the multivariate portion of the experiment described in the Experiment section below. The label values were derived from the continuous variable $-log$(EC50 (M)). 80% of patients were randomly sampled into the training set and the remaining 20% used as the test set. The `lm()` and `predict.lm()` functions from the `stats` package [33] were used to generate the model and predict the values for the test data. The β coefficients were also collected to determine whether a gene was more likely to indicate drug sensitivity ($\beta > 0$) or drug resistance ($\beta < 0$).

3.6 Experiment to Compare Univariate to Multivariate Analysis

The experiment used the t-test and final BMA model results, as well as linear regression to compare univariate analysis to multivariate analysis. The top 30 significant gene-drug pairs from the univariate t-test were chosen as the univariate dataset. The final BMA models for the 30 drugs in the univariate dataset were used to feature select genes with a posterior

probability threshold of 50% or greater. All 69 patients were used in the experiment, pseudo-randomly sampled into two groups, the training set and the testing set. 80% were used in the training set and 20% in the test set. Two linear regression models, one for univariate, and one for multivariate, were trained and tested on the same patient sets. Additionally, a baseline prediction for the test patients was calculated by taking the mean EC50 of the training data and using the base R function `jitter()` [33] to add a small amount of noise around the mean. Both models were then tested by predicting the EC50 value for the patient test set. The model predictions and the baseline prediction was scored as described in the Model Scoring subsection below. The experiment was repeated 5 times, re-sampling the patient data in between experiments. After all the experiments were finished the mean and standard deviation of the test scores were calculated.

3.6.1 Model Scoring

The linear regression models were scored using both correlation and root-squared-mean-error (RSME). Correlation of the true test set values to the predicted test values were done using both Pearson and Spearman correlation. The Spearman correlation is similar to the Pearson correlation as described in the section above. The difference between them is that the Spearman correlation first ranks the data values and then correlates the ranks. By correlating the ranks, there is more flexibility in the relationship, not requiring the two sets to be linearly related, but just that the two sets had similar rankings. The `cor()` function from the `stats` package [33] was used, specifying the type of correlation with the `method` parameter. RSME is the square root of the average squared error and shows how large the error is between the true test set values and predicted test set values. The RSME was calculated for the baseline prediction to show how well the models did compared to simply guessing.

Chapter 4
RESULTS AND DISCUSSION

4.1 Data Exploration

In this section, I explored the variant data and the drug sensitivity data using clustering and heatmaps described in Chapter 3.

4.1.1 Variant Data

Most patients had less than 4 variants in any given gene, with most patients having a single missense or indel in a gene. As seen in Figure 4.1, the common gene variants were in AKAP13, FLT3, KDM6B, KMT2B, KMT2C, MXRA5, NOTCH2, PKD1L2, SRRM2, TTBK1, ZBTB33, where at least 38 patients had a variant. 97 genes met the threshold criteria for the heatmap, which required a minimum of 5 patients with variants. Table 4.1 shows the variant frequency for the study population and the variant frequency for AML patients overall, as noted by Döhner *et al.* [6]. Out of the 13 most common genetic variants, the study population exceeded the expected frequency for 10 genes, with most patients having a FLT3 variant.

4.1.2 Drug Sensitivity Data

Drug sensitivity across all patients varied for most compounds, as shown in Figure 4.2. A subset of 15 drugs is shown in Figure 4.3. While many drugs reported an EC50 that met the threshold for drug sensitivity in most patients, few drugs had an active EC50 for all patients. Bortezomib is one example where most patients were sensitive to the drug, but one patient showed complete drug resistance. The tested drugs changed over time, which can be seen in Figure 4.2, where some patients do not have sensitivity data for a number of drugs. Figure

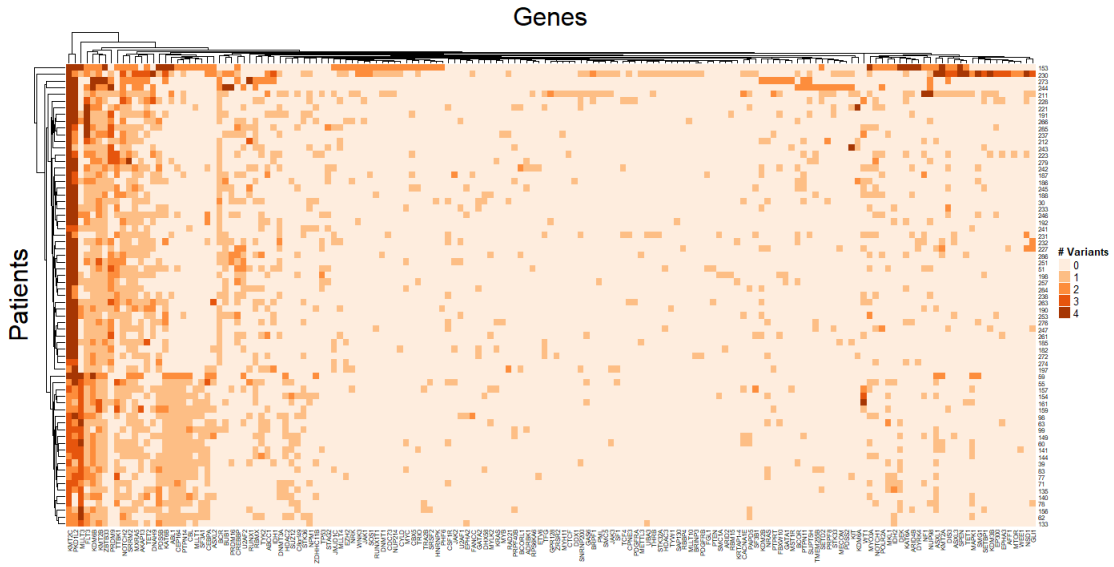

Figure 4.1: A visualization of the AML genetic variant data across all 69 patients, and filtered to only include genes with >5 patients with a variant. The genetic variant profiles of the patients were diverse, with 11 variants that were shared across most patients. The majority of patients had a variant in FLT3 and TET2.

Table 4.1: Comparison between population frequency for common genetic variants in AML [6] and the study patients

Gene	% AML Patients Overall [6]	% Study Patients
NPM1	25 - 35	19
CEBPA	6 - 10	36
RUNX1	5 - 15	19
FLT3	∼20	87
KIT	<5	10
NRAS	∼15	15
DNMT3A	18 - 22	29
ASXL1	5 - 17	27
IDH1	7 - 14	17
IDH2	8 - 19	13
TET2	7 - 25	52
KMT2A	5	13
TP53	∼8	14

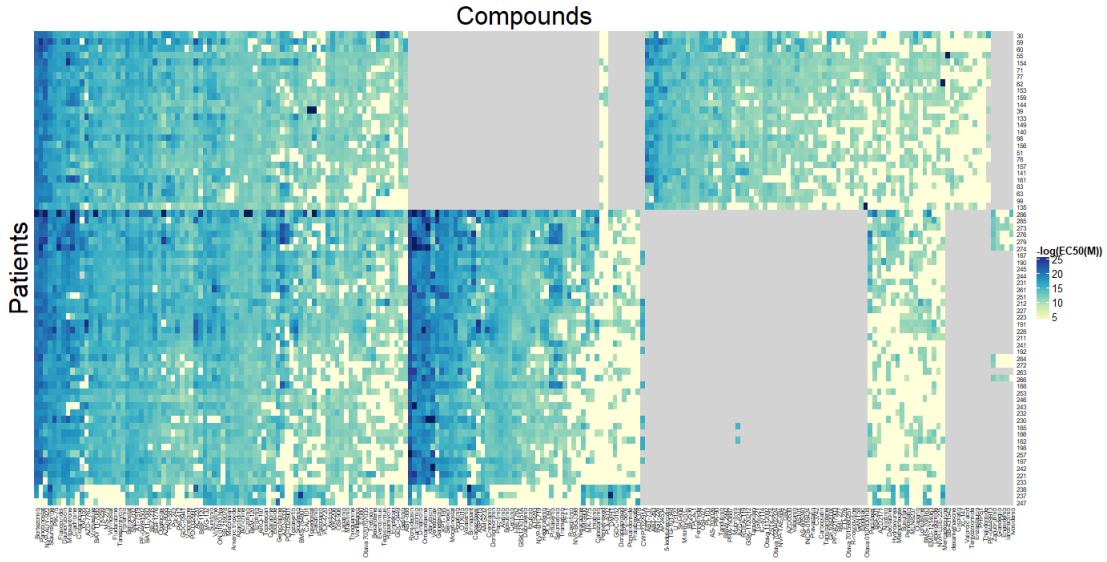

Figure 4.2: A visualization of the drug sensitivity in terms of scaled EC50 (M) across 215 drugs and 69 patients. Due to the scaling, the higher the value in the heatmap corresponds to higher drug sensitivity. This heatmap shows that, for any given drug, there were patient cells that reacted differently to it. In some cases, a set of patients had high drug sensitivity to a drug while a different set of patients had complete drug resistance for the same drug. Additionally seen in this heatmap is that not all patients were tested with the same drug. The gray indicates that there was no drug sensitivity data for a patient.

4.4 shows the number of patients with a common drug data set. This shows that there were 101 drugs with data for all 69 patients. The 101 drugs created the largest, complete data set to use for analysis.

4.2 Uni-variate Analyses

4.2.1 t-Test and Pearson Correlation

Figure 4.5 visualizes the t-test results in terms of p-values and estimated mean differences. Figure 4.6 shows the same data but with a subset of 15 drugs. There was a large number of p-values under 0.05, but many of these values had low estimated mean differences between patients with the genetic variant and those without. Due to this and the small sample size, we

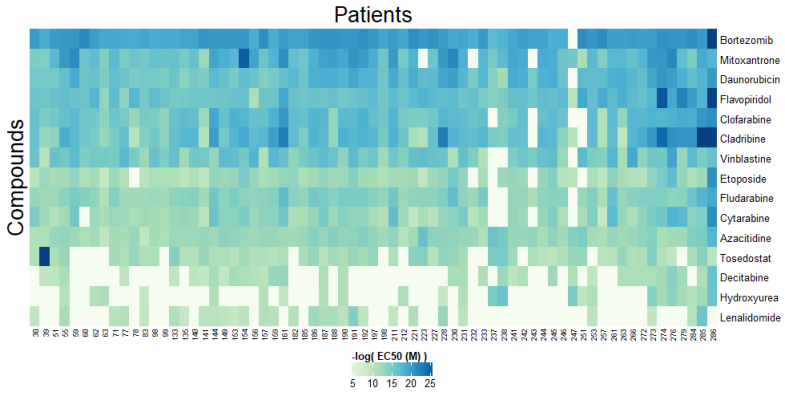

Figure 4.3: A subset of the drug sensitivity heatmap for 15 drugs across 69 patients. Due to the scaling, the higher the value in the heatmap corresponds to higher drug sensitivity. This figure shows that drugs had a similar effect for many patients, but not all. For example, Bortezomib had approximately the same drug sensitivity for all patients except 2. One of those patients had extreme drug sensitivity while the other had complete drug resistance.

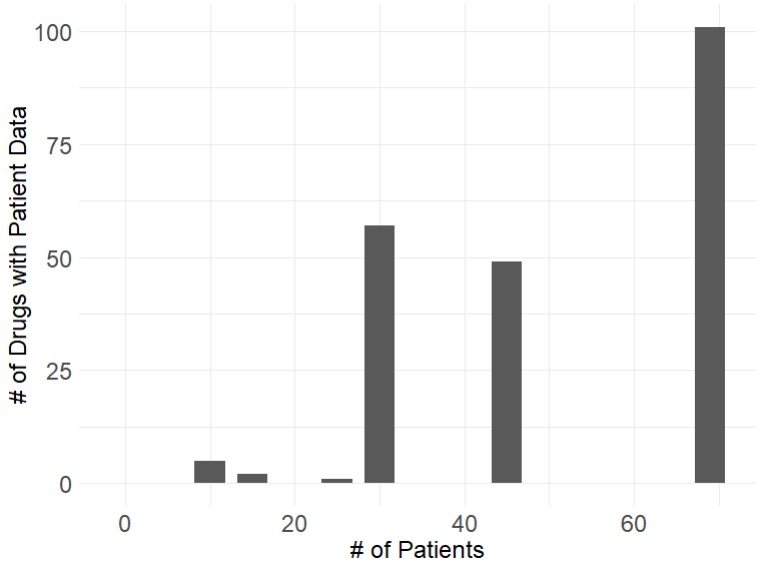

Figure 4.4: This bar chart shows the number of drugs that had patient data for n patients. 69 patients had drug sensitivity data for 101 drugs, which made for the largest number of drugs we could use to have the largest sample size of all 69 patients.

focused on gene-drug pairs with p-values $<1x10^{-4}$. Table 4.2 displays the most significant ($p < 5.0E-05$) gene-drug pairs found in the t-test, along with their estimated mean differences, the number of patients with the variant, and the Pearson correlation. Positive estimated mean differences implied that patients with the variant were more sensitive to the drug, while negative estimated mean differences implied drug resistance. These correlated with the Pearson correlation coefficients in that positive correlation meant drug sensitivity and negative correlation meant drug resistance. In Table 4.2, two gene-drug pairs, Acrichine with SF3A1 and CEP164 both showed drug resistance. Additionally, while this test was performed on all drugs, all significant relationships happened to be with drugs that had data for all 69 patients. A literature search of the top 30 significant gene - drug pairs, including a search for publications with both the gene and drug, as well as comparing the gene itself to drug information from DrugBank [31], came up empty. This implies that these results may point to novel genes for drug targets or that the pathways these genes lay in may prove useful to look more into.

4.3 Multivariate Analyses

4.3.1 Feature Selection with Bayesian Model Averaging

Bayesian Model Averaging was done on drugs for which data existed for all 69 patients, and the collected genes with posterior probability greater than 5% are shown in the Figure 4.7 heatmap. In total, 93 genes had non-zero posterior probabilities (`probne0 > 0`) for at least one drug. Many genes not only had high `probne0`, but 73 of the 93 genes had a 100% probability of having affected the drug sensitivity for at least one drug. Of these 73 genes, the following 18 genes had 100% `probne0` for at least 3 drugs: BCOR, CBL, CBX5, CSF1R, FBXW10, GATA2, GLI1, KAT6A, KDM6A, KMT2A, MLLT3, NRAS, RBMX, RUNX1, TET1, TP53, WT1. Interestingly of these genes, RUNX1, NRAS, KMT2A, and TP53 were also genes mentioned by Döhner *et al.* for being most common in AML and in the case of RUNX1 and TP53, associated with poor outcome [6]. The average number of genes chosen

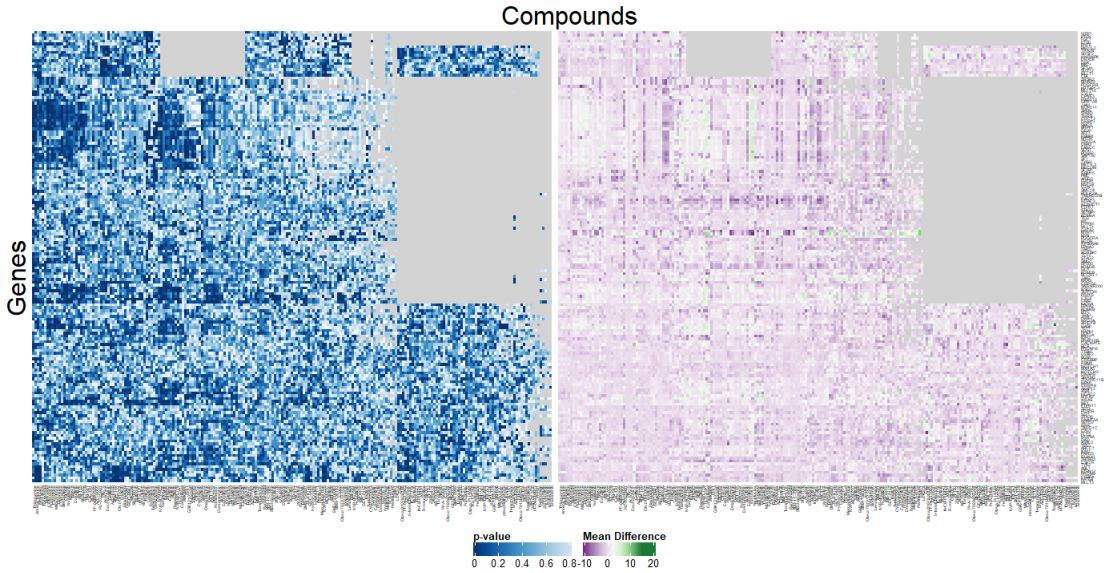

Figure 4.5: t-Test p-Value (left) and Estimated Mean Differences (right). Rows and columns were both clustered based on the p-value data, with the row and column ordering matched for the estimated mean difference heatmap. Darker colors indicate more significant p-values and higher estimated mean differences. Interestingly, lower p-values were overwhelmingly associated with estimated mean differences near zero, which can be seen by comparing the two dark blue patches in the upper level of the p-value heatmap to the similarly placed light colored patches in the estimated mean difference heatmap.

Table 4.2: Gene-Drug Pairs with Significant t-Test p-Values and their Pearson correlation

Gene - Drug Pair	p-Value	Est. Mean Diff.	# Variants	ρ
CBX5 - PLX-4720	2.1E-09	4.4	5	0.24
SUZ12 - MLN8237	5.3E-09	4.4	12	0.37
STAG2 - GDC-0449	5.5E-08	3.7	9	0.29
CBX5 - Vinblastine	2.8E-07	2.4	5	0.19
CBX5 - Pazopanib	4.0E-07	5.4	5	0.28
SUZ12 - Gefitinib	6.5E-07	3.4	12	0.31
CBX5 - Vincristine	7.9E-07	1.9	5	0.19
SUPT5H - Daunorubicin	4.6E-06	2.3	7	0.29
PTPN11 - Masitinib	7.7E-06	3.7	9	0.28
PTPN11 - Irinotecan	1.1E-05	3.6	9	0.26
PTPN11 - E7080	1.6E-05	4.3	9	0.30
NPM1 - AT-7519	1.9E-05	1.0	13	0.37
GAS6 - Bexarotene	2.6E-05	3.9	5	0.22
NPM1 - SNS-032	3.0E-05	1.3	13	0.28
PTPN11 - BIBF1120	3.2E-05	3.2	9	0.28
CBX5 - Thioguanine	3.2E-05	1.8	5	0.17
CBX5 - E7080	3.4E-05	4.0	5	0.22
CBX5 - PD0332991	3.5E-05	1.7	5	0.17
SUZ12 - Masitinib	4.0E-05	3.1	12	0.26
NRAS - PD-0325901	4.2E-05	6.1	10	0.39
NRAS - Pp-242	4.7E-05	2.5	10	0.27
SF3A1 - Acrichine	1.7E-05	-0.55	15	-0.38
CEP164 - Acrichine	2.6E-05	-0.54	27	-0.43

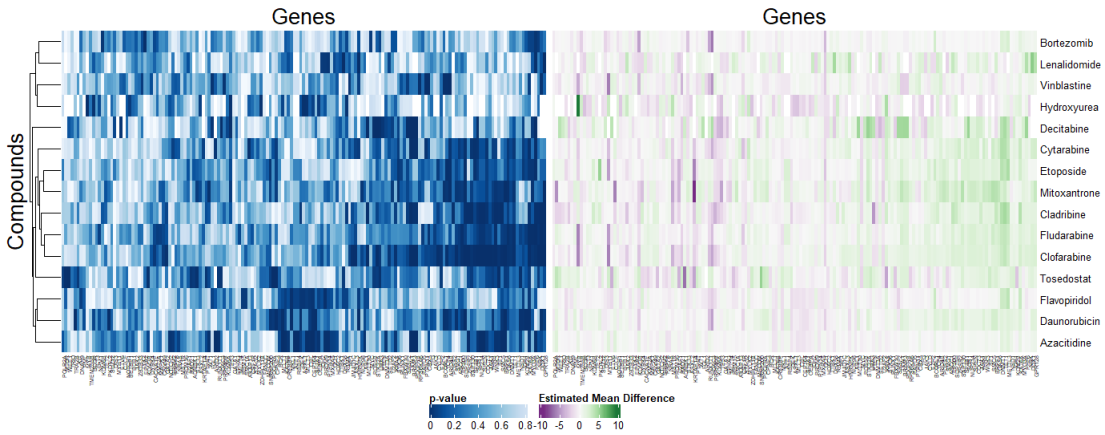

Figure 4.6: Subset of the t-test p-Value (left) and Estimated Mean Differences heatmaps (right). Rows and columns were both clustered based on the p-value data, with the row and column ordering matched for the estimated mean difference heatmap. Darker colors indicate more significant p-values and higher estimated mean differences. Most estimated mean differences were near zero. However, there was some overlap between high estimated mean differences and low p-values.

by BMA was 10.6, with a maximum of 20 genes for Nilotinib and a minimum of 3 genes for Rapamycin and SNS032. A closer look at the posterior probabilities for a subset of 15 drugs can be seen in 4.8. This subset shows the diversity of posterior probabilities for common compounds.

4.3.2 Linear Regression

Figure 4.9 shows the coefficient values for all the linear regression models, with the exception of the intercepts. Most coefficient values were between -1.7 and 1.5, with a mean of -0.09. The maximum was 3.3 and the minimum was -3.9. Figure 4.10 gives a closer look at the coefficients for 15 common compounds. Most of the coefficients were near 0. However, there were some outlier genes with significantly high or low coefficients that would imply having a higher impact on the drug sensitivity.

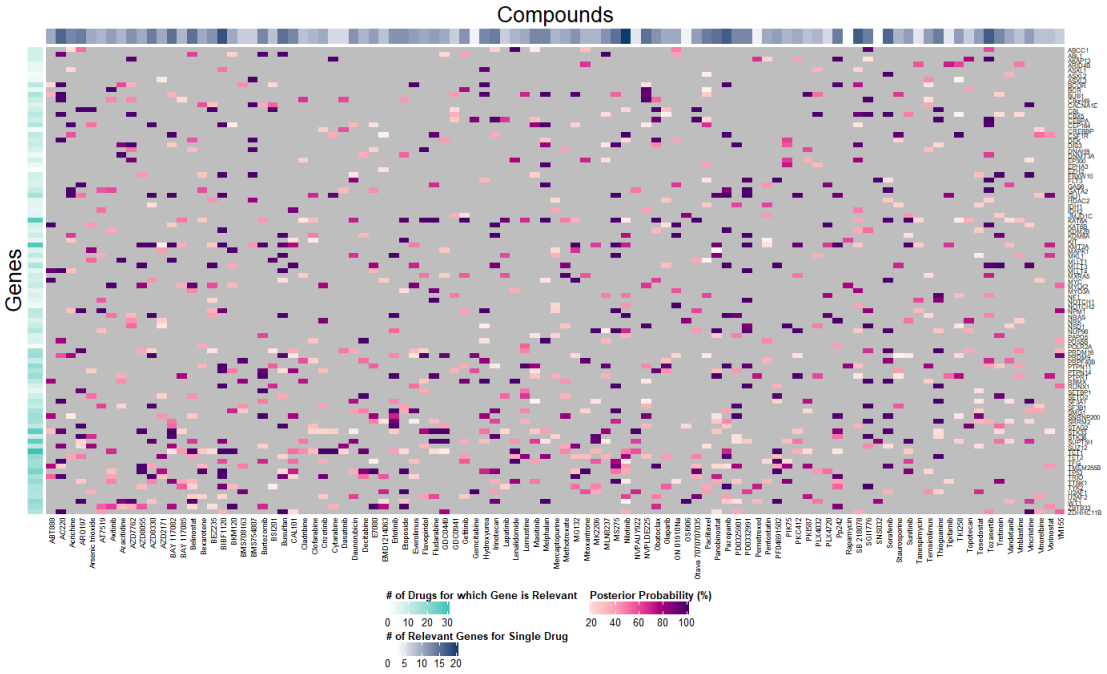

Figure 4.7: Genes with posterior probability greater than 5% as determined by Bayesian Model Averaging across all 69 patients. 101 drugs are represented. Darker purple means higher probability that the variant impacted drug sensitivity. Additionally, the annotations show how many genes were chosen to explain drug sensitivity for a specific compound (light green), and how many times a gene was chosen across all drugs (blue-gray). As expected, there were many genes with low posterior probability. However, there were also 73 genes with 100% posterior probability for at least one drug.

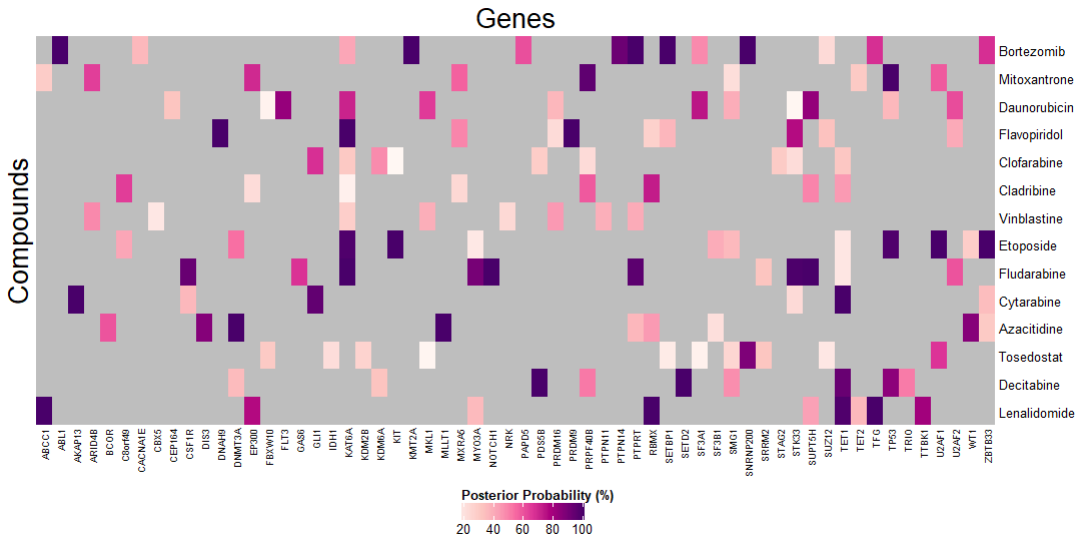

Figure 4.8: Subset of Figure 4.7 with 15 compounds. The figure shows genes with posterior probability greater than 5% as determined by Bayesian Model Averaging across all 69 patients. This subset better shows the diversity of posterior probabilities. Interestingly, drugs that had one gene with 100% posterior probability tended to have more than one gene with high or 100% posterior probability. Additionally, a gene with high or 100% probability for one drug did not correlate with having high posterior probability for all drugs. For example, STK33 had high probability for Fludarabine, but low probability for Daunorubicin.

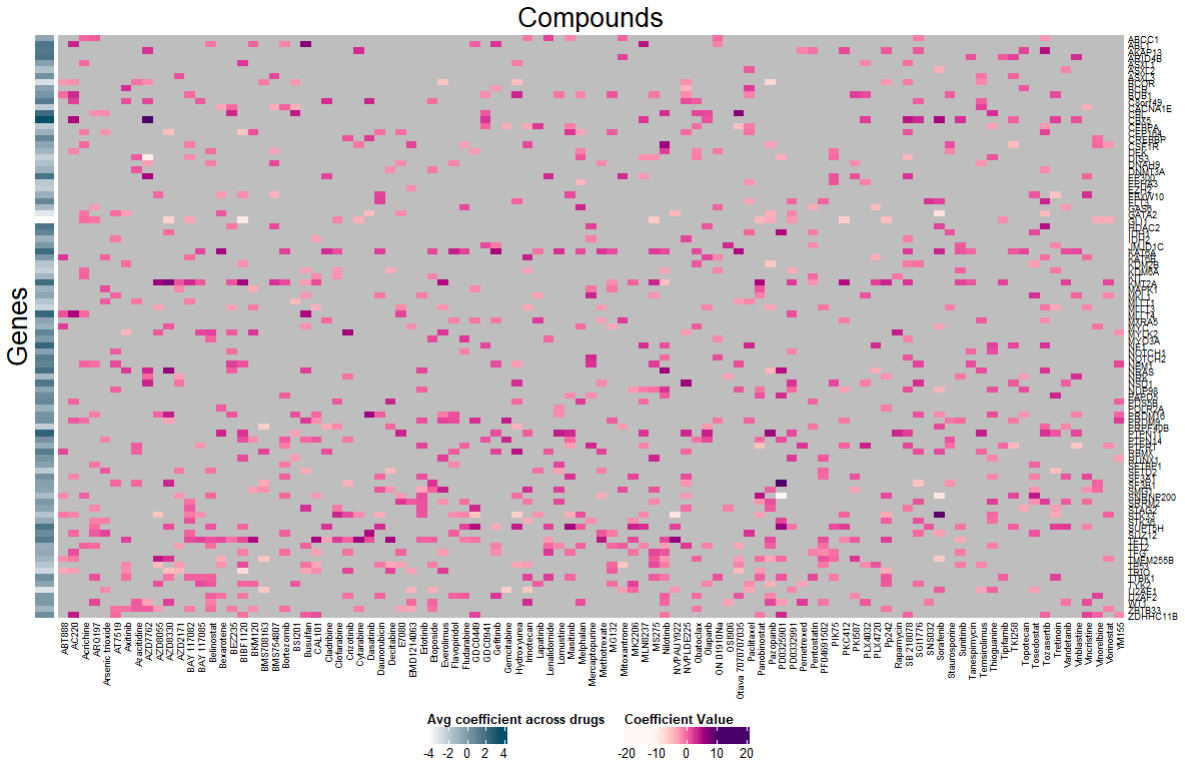

Figure 4.9: The coefficient values for each linear regression model, with the exception of the intercept. Additionally, the annotation shows the average of all coefficient values for a given gene across all drugs (dark green). Most genes had little impact on the linear regression model, with relatively low coefficients. Most models had a few genes with large coefficients, whether negative or positive, that had the greatest impact on the model.

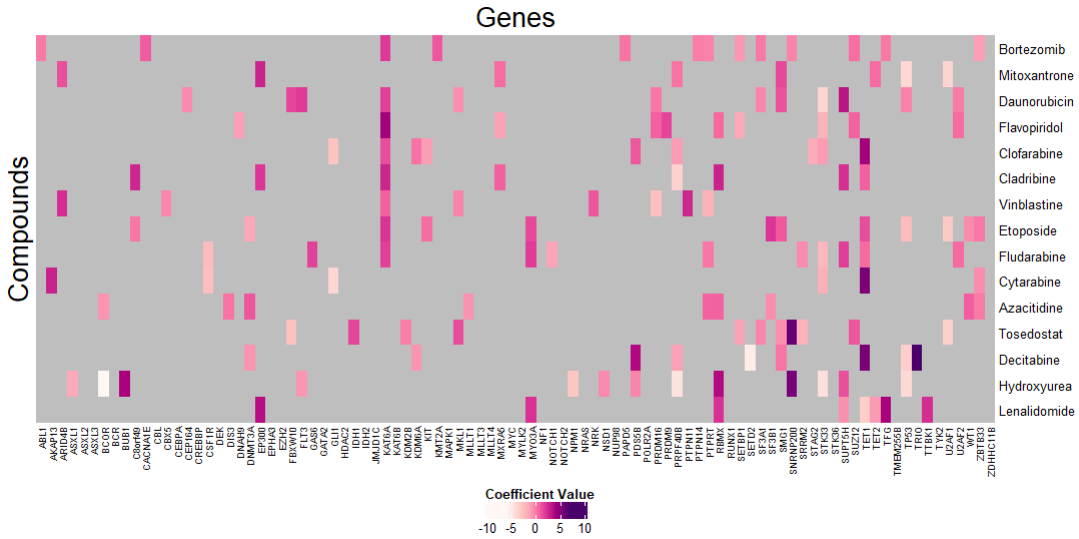

Figure 4.10: A subset of Figure 4.10 for 15 compounds showing the coefficient values for each linear regression model, with the exception of the intercept. Most coefficients were near zero, having a small effect on the overall model. However, some coefficients were large, indicating the gene should have a greater impact in the model. Additionally, there are both positive and negative coefficients, showing that some genes were potentially associated with lower drug sensitivity.

4.4 Comparison of Uni-variate to Multivariate

Tables 4.3 and 4.4 show the results of the univariate and multivariate models created for the top 30 significant p-value drugs in the t-test. See Tables A.1 and A.2 in Appendix A for the genes used in the multivariate model. Only 6 of the 30 models had the univariate gene as one of the genes in the multivariate model. In most cases, the multivariate model performed similarly or better for predicting the EC50 of the test patients. This is easier seen in Figure 4.11, where the differences between multivariate and univariate Spearman correlation and RSME are shown in histogram format. The results shown for the multivariate model used genes from the BMA step with 50% or greater posterior probability. Models were also tested with 0%, 5%, 25%, and 75% posterior probability thresholds. However, not all drugs had genes with greater than 50% posterior probability. Additionally, in most cases, high thresholds drastically reduced the number of genes. For example, in Table 4.3, it can be seen that Vinblastine, Bexarotene, and PD0332991 only have a single gene at 50% threshold, which puts these models in the same univariate category. The difference in the models for these 3 drugs lies in how the gene was chosen. For univariate, a simple t-test was used to find a significant relationship between the drug and the gene. For multivariate, the ensemble method BMA was used to find the gene with the greatest posterior probability for having an affect on the drug sensitivity. For 2 of 3 drugs, Vinblastine and Bexarotene, the gene chosen by multivariate analysis performed slightly better in the linear regression model. However, the gene chosen for PD0332991 by multivariate analysis performed significantly worse. This emphasizes the idea that targeting a single gene is the incorrect path. On the other end of the spectrum, drugs with many genes often performed worse than the univariate model, meaning the model was most likely over-fitting the data. Most models with 2 - 8 genes appeared to perform the best. Since the models used for feature selection were based on a single training set of randomly sampled patients, the multivariate analysis might be improved by using a similar method done for the scoring: run the feature selection BMA section 5 times and average the posterior probabilities of the genes found in each of the trials. Then

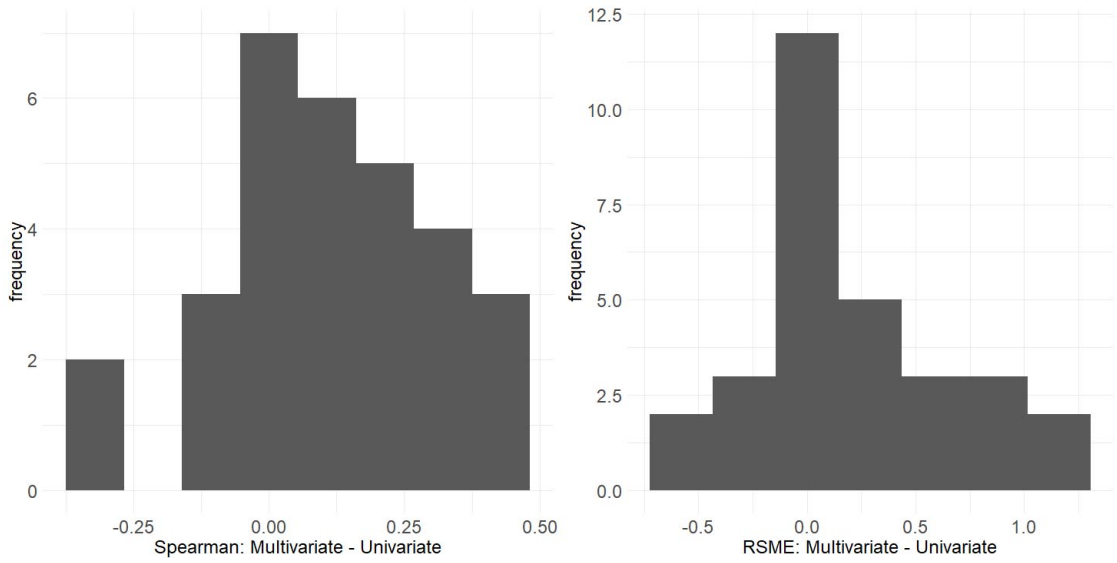

Figure 4.11: Histogram of the differences between Spearman coefficients (left) and RSME (right) comparing linear regression model predictions to true test values for 30 univariate and multivariate models. The top 30 gene-drug pairs with the lowest p-value in the t-test were selected for modeling. The univariate linear regression model used the single gene found significant in the t-test as a data point. The multivariate linear regression model used BMA to select genes with a threshold of >50% posterior probability as data points. In the majority of cases, the multivariate models performed similarly or better than univariate models.

use thresholds to find the top genes to include in the linear regression experiment.

Table 4.3: Univariate & Multivariate Linear Regression Model Correlation

	Uni			Multi	(50%)
Compound	r_s	r	n	r_s	r
PLX4720	0.45±0.05	0.36±0.08	4	0.46±0.07	0.53±0.046
MLN8237	0.35±0.04	0.33±0.03	8	0.63±0.04	0.72±0.035
GDC0449	0.36±0.05	0.31±0.039	2	0.54±0.11	0.45±0.09
Vinblastine	0.32±0.04	0.26±0.02	1	0.38±0.05	0.31±0.08
Pazopanib	0.48±0.07	0.38±0.06	11	0.46±0.14	0.42±0.09
Gefitinib	0.21±0.09	0.29±0.04	5	0.41±0.16	0.51±0.15
Vincristine	0.39±0.02	0.26±0.05	7	0.24±0.17	0.25±0.19
Daunorubicin	0.48±0.04	0.38±0.02	7	0.59±0.12	0.46±0.10
Masitinib	0.42±0.09	0.35±0.02	6	0.20±0.09	0.27±0.08
Irinotecan	0.40±0.05	0.35±0.06	7	0.59±0.10	0.54±0.08
E7080	0.27±0.07	0.28±0.04	5	0.58±0.04	0.58±0.05
AT7519	0.42±0.08	0.41±0.08	12	0.55±0.06	0.52±0.07
Bexarotene	0.37±0.13	0.44±0.09	1	0.39±0.04	0.36±0.06
SNS032	0.53±0.09	0.46±0.09	3	0.14±0.05	0.11±0.04
BIBF1120	0.27±0.06	0.27±0.07	11	0.52±0.07	0.67±0.03
Thioguanine	0.45±0.09	0.26±0.04	3	0.39±0.08	0.37±0.09
E7080	0.38±0.06	0.32±0.04	5	0.58±0.04	0.58±0.05
PD0332991	0.30±0.03	0.19±0.02	1	0.06±0.14	0.14±0.17
Masitinib	0.13±0.09	0.24±0.06	6	0.20±0.09	0.27±0.08
PD0325901	0.25±0.09	0.30±0.06	8	0.57±0.10	0.69±0.06
Pp242	0.34±0.06	0.32±0.05	9	0.41±0.10	0.47±0.08
Acrichine	0.45±0.07	0.40±0.07	8	0.65±0.08	0.63±0.12
Acrichine	0.55±0.16	0.51±0.17	8	0.65±0.08	0.63±0.12

Table 4.4: Univariate & Multivariate Linear Regression Model RSME

	Uni			Multi	(50%)
Compound	$RMSE_{baseline}$	$RSME$	n	$RMSE_{baseline}$	$RSME$
PLX4720	4.74	4.55±0.04	4	4.77	4.14±0.23
MLN8237	4.38	4.16±0.04	8	4.38	3.12±0.18
GDC0449	4.32	4.08±0.07	2	4.32	3.95±0.25
Vinblastine	2.98	2.81±0.09	1	2.94	2.78±0.42
Pazopanib	5.09	4.73±0.12	11	5.09	4.73±0.26
Gefitinib	4.40	4.23±0.05	5	4.42	3.68±0.39
Vincristine	2.25	2.17±0.09	7	2.26	2.46±0.34
Daunorubicin	2.79	2.65±0.15	7	2.77	2.50±0.61
Masitinib	4.49	4.25±0.06	6	4.46	4.77±0.34
Irinotecan	4.69	4.48±0.12	7	4.69	4.09±0.23
E7080	5.25	5.06±0.06	5	5.24	4.34±0.18
AT7519	1.16	1.06±0.06	12	1.17	0.98±0.12
Bexarotene	4.85	4.73±0.06	1	4.88	4.66±0.18
SNS032	1.08	0.92±0.08	3	1.07	1.38±0.12
BIBF1120	4.10	3.97±0.07	11	4.08	3.07±0.13
Thioguanine	2.74	2.68±0.11	3	2.73	2.59±0.60
E7080	5.26	5.02±0.06	5	5.23	4.34±0.18
PD0332991	2.68	2.64±0.15	1	2.70	2.71±0.34
Masitinib	4.49	4.38±0.06	6	4.51	4.77±0.34
PD0325901	5.42	5.23±0.09	8	5.45	4.01±0.26
Acrichine	0.73	0.68±0.10	8	0.74	0.57±0.07
Acrichine	0.75	0.65±0.10	8	0.76	0.57±0.07

Chapter 5

AUTOMATING THE ANALYSIS PIPELINE

5.1 Overview

We aim to automate all the analytical steps used in the analyses of the AML data described in this thesis. In addition, users can modify the parameters used in the analyses. This AML workflow was automated for two reasons, the first being that the results of this analysis should be reproducible to be good science. With the exception of the Bayesian model averaging and linear regression steps, the same data set will result in the same outcome on any computing environment capable of running Docker. Due to the pseudo-random sampling for the patient training and test sets, the results from these would be different each time unless a seed was set for the random selection. The second reason being that the Becker Lab is continuously generating new patient data that should be incorporated into the study. By automating the workflow, the new data simply has to be added to the current variant and drug sensitivity data and run through the pipeline to obtain the latest results.

The automation of the AML workflow shown in Figure 5.1 was done using the Biodepot-workflow-builder (Bwb) [16]. Bwb is a tool created by the bioinformatics research group at University of Washington Tacoma that allows for creation and execution of reproducible data analysis pipelines. Bwb runs in a Docker [19] container, consists of a drag-and-drop graphical user interface (GUI), and supports graphical output via a web server. The only requirements for a user to run the program is a web browser and having Docker installed on their local computer. Widgets, small elements of a workflow that complete a task, spin up a Docker container to handle the job. Since each task is a separate widget with its own Docker image, a workflow can aggregate tools written in multiple programming languages without placing a burden on the user to have their environment set up for each specific language.

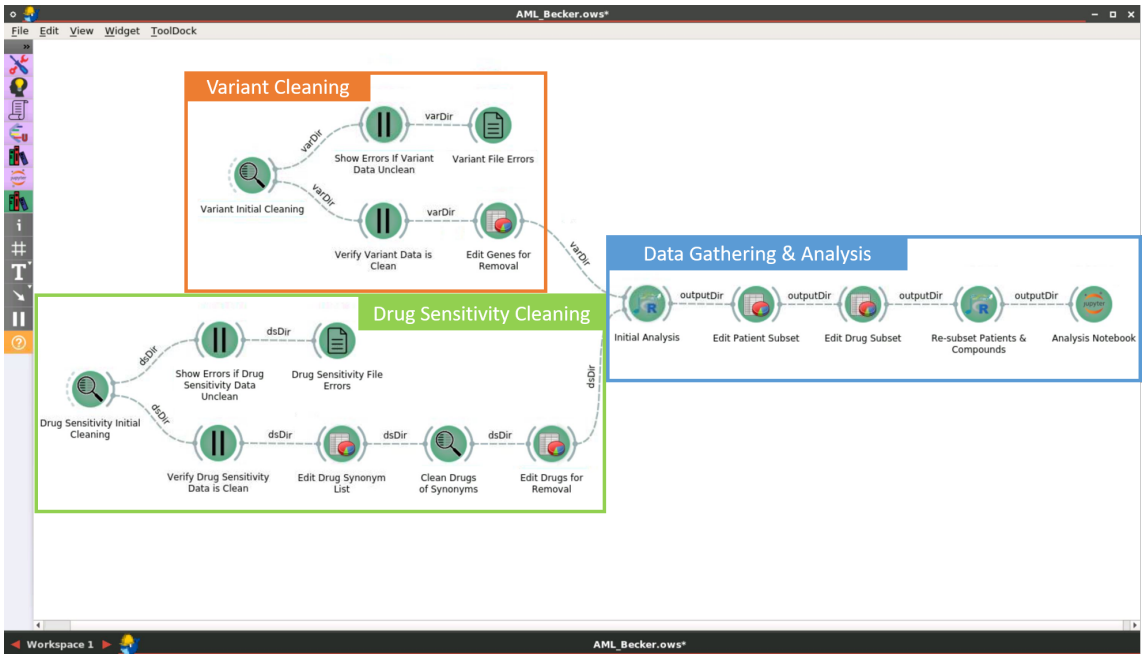

Figure 5.1: The full AML workflow created in Biodepot Workflow Builder, from cleaning to analysis.

5.2 Cleaning

A barrier to automated data analysis pipelines and data science, in general, is the lack of clean real world data. This was especially true for the data used in this thesis. To ensure the workflow would run correctly, it was vital to have data cleaning branches, one for the variant data, and one for the drug sensitivity data. These can be seen outlined in orange and green, respectively, in Figure 5.1, as well as zoomed in for Figures 5.2 and 5.3.

Both cleaning branches have a main task: ensure the files given meet the requirements needed for future tasks to succeed. The first widget in each branch takes in a single folder that contains its respective data files and checks for the following items in Table 5.1. Any file that does not meet the requirements is noted in an error log file written to disk that details each issue found. If this file exists, the 'gate' widgets block the path forward to the main analysis section, instead displaying the error log for the user to see the problems. The user is expected to fix the file problems to pass all the validation checks before continuing. By passing off the responsibility to fix the issues onto the user, this means that the data issues are being remedied by someone who knows about the data.

While both cleaning branches share a secondary task of filtering out unwanted compounds and genes, the drug sensitivity branch has an extra step. In this extra step, the user can go through all the drug compounds found and update a compound synonym file. This helps lump together compounds that have multiple names, such as generic versus brand, or misspelled names under a single name.

5.3 Filtering and Analysis

The filtering and analysis branch gathers the data, filters it based on criteria, allows for subsetting by patients and compounds, and ultimately produces a Jupyter Notebook with the analysis results. The branch can be seen outlined in blue in Figure 5.1 and zoomed in on Figure 5.4.

The filtering portion of the branch gathers the drug sensitivity and variant data, filters

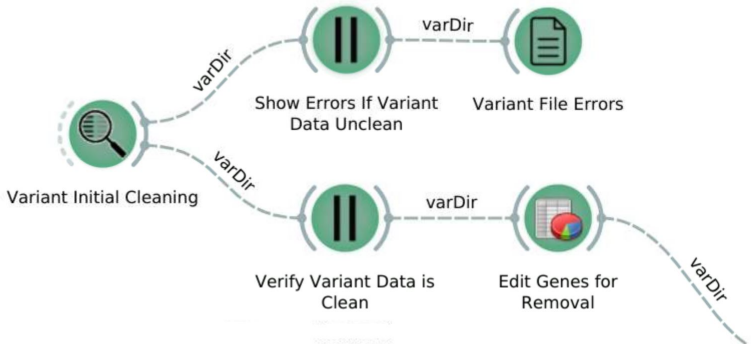

Figure 5.2: The variant cleaning section AML workflow created in Biodepot Workflow Builder

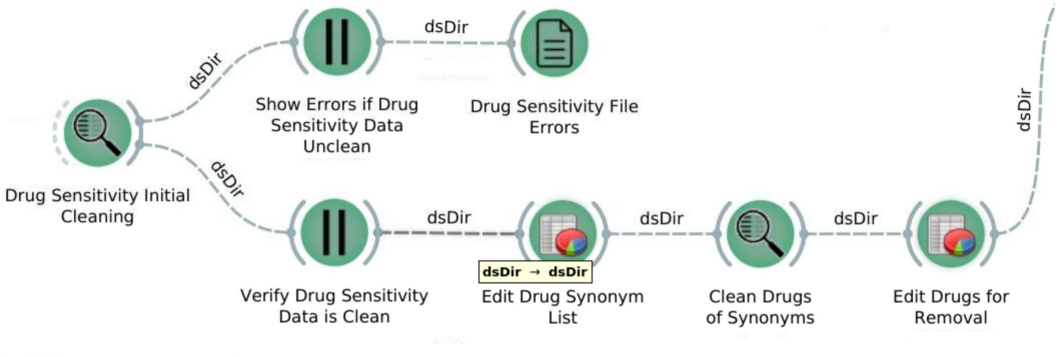

Figure 5.3: The drug cleaning section AML workflow created in Biodepot Workflow Builder

Table 5.1: File Cleaning Validation

Drug Sensitivity	Variant	Both
Patient number is in 'filename' column	Patient number is in 'Sample' column	Patient number is in filename Patient numbers in filename and column match
Excel sheet with at least one specified phrase exists	Excel sheets with 'missense' and 'indel' exist	
The columns 'Compound', 'EC50' or 'IC50', and 'AUC' (though not used) exist	The column 'Gene' exists	Data in columns is not missing

based on desired criteria, and outputs two matrices, one with the number of variants in a given gene and one with the EC50 compound values in $-log(\text{EC50})$ form for all patients. Only patients with both genetic and drug sensitivity data are included in the two matrices. The filtering criteria are options that can be input into the widget itself, as seen in Figure 5.5. If no options for SIFT or PolyPhen are chosen, then the data is only filtered by the intersect of patients with drug sensitivity data and variant data. This allows for flexibility of using data that does not include items such as PolyPhen and SIFT. Once the initial gathering and filtering of data is done, the user can then see the current patient and drug compound lists and choose which to include in the final analysis.

The analysis of the data is the final step of the workflow. After the data is cleaned, gathered, filtered, and subset, a Jupyter Notebook containing the analysis script is run and output for the viewer to see. Since the output is a Jupyter Notebook, this also allows the user to rerun any interesting portions, add or edit existing analysis, and more. The notebook contains all the code to create the heatmaps, histogram, t-test, correlation, Bayesian model

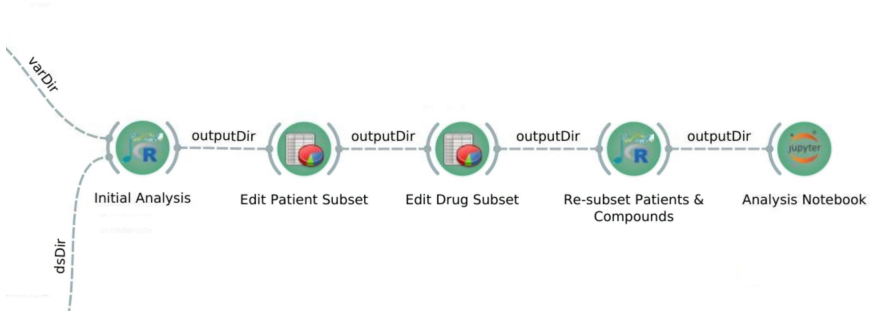

Figure 5.4: The data gathering and analysis section AML workflow created in Biodepot Workflow Builder

averaging, and linear regression analysis.

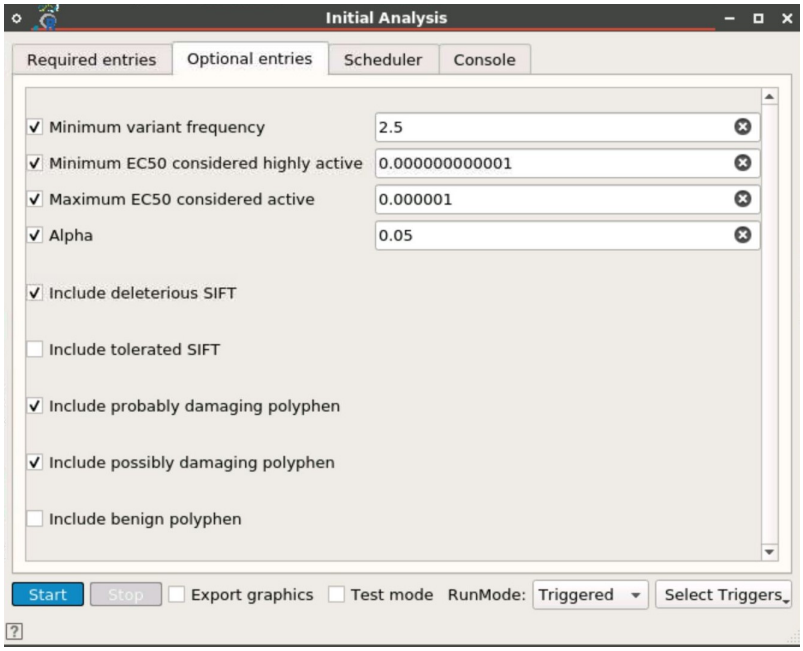

Figure 5.5: Allowed options for filtering the data with current defaults. If no SIFT or PolyPhen options are chosen, then the data is not filtered on these two categories.

Chapter 6

CONCLUSION

6.1 Conclusion

Overall, we found that multivariate analysis, comparing multiple genetic variants with drug sensitivity, to be better at predicting test patient sensitivity than univariate analysis, comparing a single genetic variant with drug sensitivity. We also found potentially novel relationships between genetic variants and target drugs. This work can be improved and expanded by separating the genes in the multivariate analysis into their respective pathways and comparing them to the known function of the target drug. Additionally, the experiment can be repeated with the availability of new patient data from incoming patients to Dr. Becker's lab, or by integrating big biomedical data from studies such as Beat AML. The automation process can be improved by allowing for the input of a single patient's mutation data and the output of the top drugs with potential for treatment. The most useful improvement to this study would be lab testing of the relationships found thus far to either confirm or deny the plausibility for treatment.

BIBLIOGRAPHY

[1] The Cancer Genome Atlas. About tcga, 2019.

[2] Lisa Barroilhet and Ursula Matulonis. The nci-match trial and precision medicine in gynecologic cancers. *Gynecologic Oncology*, 148(3):585–590, March 2018.

[3] Pamela Becker. Series gse107465, November 2017.

[4] Young Kwang Chae, Christos Vaklavas, Heather H. Cheng, Fangxin Hong, Lyndsay Harris, Edith P. Mitchell, James A. Zwiebel, Lisa McShane, Robert James Gray, Shuli Li, S. Percy Ivy, Sherry Singer Ansher, Stanley R. Hamilton, Paul M. Williams, James V. Tricoli, Carlos L. Arteaga, Barbara A. Conley, Peter J. O'Dwyer, Alice P. Chen, and Keith Flaherty. Molecular analysis for therapy choice (match) arm w: Phase ii study of azd4547 in patients with tumors with aberrations in the fgfr pathway. *Journal of Clinical Oncology*, 36(15):2503, May 2018.

[5] Merlise Clyde. Model averaging, 2005.

[6] Hartmut Döhner, Daniel J Weisdorf, and Clara D Bloomfield. Acute myeloid leukemia. *New England Journal of Medicine*, 373(1136), 2015.

[7] Hervé Dombret. Gene mutation and aml pathogenesis. *Blood*, 118(20), 2011.

[8] DrugBank. Midostaurin, 2019.

[9] JW Tyner et al. Functional genomic landscape of acute myeloid leukaemia. *Nature*, 562:526–531, 2018.

[10] Keith T Flaherty, Robert Gray, Alice Chen, Shuli Li, David Patton, Stanley R Hamilton, Paul M Williams, Edith P Mitchell, A John Iafrate, Jeffrey Sklar, Lyndsay N Harris, Lisa M McShane, Larry V Rubinstein, David J Sims, BS, Mark Routbort, Brent Coffey, MS, MBA, Tony Fu, MS, James A Zwiebel, Richard F Little, Donna Marinucci, Robert Catalano, Phar, Rick Magnan, BS, Warren Kibbe, Carol Weil, JD, James V Tricoli, Brian Alexander, Shaji Kumar, Gary K Schwartz, Funda Meric-Bernstam, Chih-Jian Lih, Worta McCaskill-Stevens, Paolo Caimi, Naoko Takebe, Vivekananda Datta, Carlos L Arteaga, Jeffrey S Abrams, Robert Comis, Peter J O'Dwyer, Barbara A Conley, and NCI-MATCH Team. The molecular analysis for therapy choice (nci-match) trial:

Lessons for genomic trial design. *JNCI: Journal of the National Cancer Institute*, January 2020.

[11] Chris Fraley, Wm. Chad Young, Ka Yee Yeung, and Adrian E. Raftery. t networkBMA: Regression-bsed network inference using bayesian model averaging, 2014.

[12] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22, 2010.

[13] Z. Gu, R. Eils, and M. Schlesner. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics*, 2016.

[14] Jennifer A. Hoeting, David Madigan, Adrian E. Raftery, and Chris T. Volinsky. Bayesian model averaging: A tutorial. *Statistical Science*, 14(4):382–417, 1999.

[15] Jennifer A Hoeting, David Madigan, Adrian E Raftery, and Chris T Volinsky. Bayesian model averaging: a tutorial. *Statistical science*, pages 382–401, 1999.

[16] Ling-Hong Hung, Jiaming Hu, Trevor Meiss, Alyssa Ingersoll, Wes Lloyd, Daniel Kristiyanto, Yuguang Xiong, Eric Sobie, and Ka Yee Yeung. Building containerized workflows using the biodepot-workflow-builder (bwb). *Cell Systems*, 9:508–514, 2019.

[17] Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, and Sunyaev SR. A method and server for predicting damaging missense mutations. *Nature Methods*, 7:248–9, April 2010.

[18] Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, and Sunyaev SR. Guidelines for accurate ec50/ic50 estimation. *Pharmaceutical Statistics*, 10:128–134, March 2011.

[19] Docker Inc. Docker, 2020.

[20] National Human Genome Research Institute. Missense mutation, 2019.

[21] invivoscribe. Myaml 194 targeted ngs gene panel, 2018-2020.

[22] Yunyi Kang, Trish Tran, Linda Zhang, Edward D Ball, Carlo Piermarocchi, and Giovanni Paternostro. Personalized drug combinations for the treatment of acute myeloid leukemia (aml) patients. *Blood: Abstracts and Meeting Program for the 56th ASH Annual Meeting*, 124(21), 2014.

[23] R.E. Kass and A.E. Raftery. Bayes factors. *Journal of the American Statistical Association*, 90:773–795, June 1995.

[24] Ian E. Krop, Opeyemi Jegede, Juneko E. Grilley-Olson, Josh David Lauring, Stanley R. Hamilton, James A. Zwiebel, Shuli Li, Lawrence Rubinstein, Austin Doyle, David R. Patton, Edith P. Mitchell, Carlos L. Arteaga, Barbara A. Conley, David Sims, Lyndsay Harris, Alice P. Chen, and Keith Flaherty. Results from molecular analysis for therapy choice (match) arm i: Taselisib for pik3ca-mutated tumors. *Journal of Clinical Oncology*, 36(15):101, May 2018.

[25] Su-In Lee, Safiye Celik, Benjamin A Logsdon, and et al. A machine learning approach to integrate big data for precision medicine in acute myeloid leukemia. *Nature Communications*, 9(42), 2018.

[26] Leukemia and Lymphoma Society. Lls is leading the offensive against acute myeloid leukemia, 2019.

[27] Ricki Lewis. *Human genetics : concepts and applications*. McGraw-Hill, New York, 9th ed. edition, 2010.

[28] Kenneth Lo, Adrian E. Raftery, Kenneth M. Dombek, Jun Zhu, Eric E. Schadt, Roger E. Bumgarner, and Ka Yee Yeung. Integrating external biological knowledge in the construction of regulatory networks from time-series expression data. *BMC Systems Biology*, 6:101, 2012.

[29] Julienne M Mullaney, Ryan E Mills, W Stephen Pittard, and Scott E Devine. Small insertions and deletions (indels) in human genomes. *Human Molecular Genetics*, 19(R2), 2010.

[30] Pauline C. Ng and Steven Henikoff. Sift: predicting amino acid changes that affect protein function. *Nucleic Acids Research*, 31:3812–3814, July 2003.

[31] Canada Institutes of Health Research. Drugbank, 2020.

[32] National Cancer Institute GDC Data Portal. Tcga-laml, 2018.

[33] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2019.

[34] Adrian Raftery, Jennifer Hoeting, Chris Volinsky, Ian Painter, and Ka Yee Yeung. *BMA: Bayesian Model Averaging*, 2019. R package version 3.18.11.

[35] A.E. Raftery. Bayesian model selection in social research (with discussion). *Sociological Methodolgy*, 25:111–196, 1995.

[36] A.E. Raftery, D Madigan, and C Volinksky. Accounting for model uncertainty in survival analysis improves predictive performance (with discussion). *Bayesian Statistics*, 5:323–349, 1995.

[37] Jane B. Reece, Lisa A. Urry, Michael L. Cain, Steven A. Wasserman, Peter V. Minorsky, and Robert B. Jackson. *Biology*. Pearson Education, Inc, 2011.

[38] April K.S. Salama, Shuli Li, Erin Renee Macrae, Jong-In Park, Edith P. Mitchell, James A. Zwiebel, Helen X. Chen, Robert James Gray, Lisa McShane, Lawrence Rubinstein, David Patton, Paul M. Williams, Stanley R. Hamilton, Deborah Kay Armstrong, Barbara A. Conley, Carlos L. Arteaga, Lyndsay Harris, Peter J. O'Dwyer, Alice P. Chen, and Keith Flaherty. Dabrafenib and trametinib in patients with tumors with braf v600e/k mutations: Results from the molecular analysis for therapy choice (match) arm h. *Journal of Clinical Oncology*, 37(15):3002, May 2018.

[39] American Cancer Society. Typical treatment of acute myeloid leukemia (except apl), 2019.

[40] American Cancer Society. What is acute myeloid leukemia?, 2019.

[41] American Cancer Society. What is acute myeloid leukemia?, 2019.

[42] GraphPad Support. 50% of what? how exactly are ic50 and ec50 defined?, 2010.

[43] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

[44] Statistics How To. Lasso regression: Simple definition, 2019.

[45] SPSS Tutorials. *Pearson Correlation Coefficient - Quick Introduction*.

[46] Hadley Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016.

[47] Ka Yee Yeung, Kenneth M. Dombek, Kenneth Lo, John E. Mittler, Jun Zhu, Eric E. Schadt, Roger E. Bumgarner, and Adrian E. Raftery. Construction of regulatory networks using expression time-series data of a genotyped population. *Proceedings of the National Academy of Sciences*, 108(48):19436–19441, 2011.

[48] Ka Yee Yeung, Chris Fraley, Adrian E. Raftery, and Wm. Chad Young. Uncovering gene regulatory relationships using networkBMA, 2012.

Appendix A

UNIVARIATE AND MULTIVARIATE MODEL GENES

Table A.1: Genes in Univariate and Multivariate Models

Compound	Univariate	Multivariate (50% Threshold)
PLX4720	CBX5	DNMT3A, GAS6, NRAS, SETBP1
MLN8237	SUZ12	ABL1, CBX5, DIS3, KAT6B, PAPD5, SF3B1, SMG1, TYK2
GDC0449	STAG2	BCOR, CREBBP
Vinblastine	CBX5	ARID4B
Pazopanib	CBX5	CBX5, GATA2, KAT6A, KDM2B, KIT, PDS5B, SNRNP200, TFG, TMEM255B, TP53, TYK2
Gefitinib	SUZ12	ABL1, CREBBP, KAT6A, NSD1, TFG
Vincristine	CBX5	CACNA1E, GAS6, GATA2, MLLT3, NUP98, PRDM16, SF3B1
Daunorubicin	SUPT5H	ARID4B, CEP164, CSF1R, GLI1, KMT2A, MXRA5, TYK2
Masitinib	PTPN11	DNMT3A, GATA2, MYLK2, STK33, SUPT5H, ZBTB33
Irinotecan	PTPN11	CREBBP, DNMT3A, FLT3, KDM6A, PRDM9, SMG1, TP53
E7080	PTPN11	HDAC2, KAT6A, MKL1, TP53, U2AF2
AT7519	NPM1	ABCC1, GATA2, IDH1, KMT2A, NOTCH1, NOTCH2, NPM1, PRDM16, STK36, SUPT5H, TET1, U2AF2
Bexarotene	GAS6	KAT6A
SNS032	NPM1	CSF1R, NUP98, SUPT5H
BIBF1120	PTPN11	ASXL1, ASXL2, CEP164, GLI1, KDM2B, KMT2A, NPM1, PTPN11, SUZ12, TRIO, ZBTB33
Thioguanine	CBX5	DEK, NRAS, TET1

Table A.2: Genes in Univariate and Multivariate Models

Compound	Univariate	Multivariate (50% Threshold)
E7080	CBX5	HDAC2, KAT6A, MKL1, TP53, U2AF2
PD0332991	CBX5	CSF1R
Masitinib	SUZ12	DNMT3A, GATA2, MYLK2, STK33, SUPT5H, ZBTB33
PD0325901	NRAS	DIS3, IDH1, NRAS, PDS5B, PRDM9, SF3B1, SNRNP200, SRRM2
Pp242	NRAS	GLI1, KDM6A, KMT2A, MXRA5, NRAS, PAPD5, SRRM2, SUZ12, TET1
Acrichine	SF3A1	CEP164, GATA2, GLI1, NF1, SETD2, TET1, TET2, TP53
Acrichine	CEP164	CEP164, GATA2, GLI1, NF1, SETD2, TET1, TET2, TP53