

Springboard Data Science Career Track

Classification of Acute Leukemias Using Gene Expression Profiles

Capstone Project I

Caitlin McDonold

April 2019

Table of Contents

- I. Introduction
- II. Overview of Dataset
 - i. Dataset Description
 - ii. Dataset Cleaning and Wrangling
- III. Exploratory Data Analysis

I. Introduction

One of the challenges of cancer treatment lies in determining which line of therapy will be most effective in treating the patient's particular type of cancer. A therapy regimen that works strongly against one type of tumor, specifically targeting and eradicating the malignant cells, may have no effect at all on a different tumor. As cancer therapies have become more targeted and specific, the need for accurate tools to classify tumors and discriminate between tumors based on their clinical and biological features has become paramount.

Traditionally, cancers have been classified based on where in the body they originated (i.e. lung cancer), and then further subdivided based on the patient's age, cell type, histology, and sometimes biomarkers such as hormone receptors. However, high-throughput technologies have yielded an unprecedented volume of new data on variations in DNA, RNA, or protein expression in many cancers. The scope and complexity of these new datasets requires the processing power of computer programs and algorithms to organize and analyze the data. By combining the abundance of DNA sequencing and gene expression data with machine learning algorithms, we may be able to discover new tumor subclasses and more accurately predict the therapeutic strategy that is most likely to be successful in achieving remission.

Acute leukemias are an example of morphologically similar tumors that have different clinical pathologies and can respond differently to clinical treatments. Both ALL and AML are cancers of the bone marrow and blood, though they originate in different cell types; ALL develops from immature lymphocytes, while AML develops in early forms of myeloid cells. The distinction between ALL and AML has been well established over years of careful scientific study and characterization, and successful treatment depends on distinguishing between ALL and AML, as they respond best to different chemotherapy agents. However, no single test is yet sufficient to distinguish between ALL and AML; instead, an experienced hematopathologist is needed to interpret results of multiple complex and highly specialized analyses, and misclassifications do sometimes occur. While treating ALL with the ideal chemotherapy regimen for AML can lead to remission (and vice versa), the success rates are greatly diminished.

The goal of this project is to use microarray data of bone marrow samples obtained from acute leukemia patients to train a classifier to distinguish between the two types of acute

leukemia. The trained classifier will then be used to classify new acute leukemia samples as either ALL or AML to aid doctors in diagnosis so that they may determine an appropriate chemotherapy regimen.

II. Overview of Dataset

i. Dataset Description

The dataset for this project comes from a proof-of-concept study published by Golub et. al., which demonstrated that new cases of AML and ALL could be classified by gene expression data gathered using DNA microarrays.¹ The initial ‘training’ dataset comprises gene expression data from 38 bone marrow samples obtained from acute leukemia patients; 27 samples are from ALL patients, and 11 are from AML patients. The samples were analyzed by Affymetrix microarrays containing probes from 6817 human genes. The independent dataset contains microarray data from an additional 34 samples from acute leukemia patients; these samples are much more diverse than those in the training dataset and include samples from peripheral blood rather than bone marrow, from a broader range of patients (both adults and children), and from different laboratories. Both the training and the independent datasets are publicly available on Kaggle for download (<https://www.kaggle.com/crawford/gene-expression/home>).

The training and independent datasets are stored in separate .csv files: ‘data_set_ALL_AML_independent.csv’ and ‘data_set_ALL_AML_independent.csv’. The files contain columns labeled ‘Gene Description’ and ‘Gene Accession Number,’ followed by a column for each tumor sample, labeled by number (i.e. ‘1’, ‘2’, etc.). The entries of the sample columns are the experimental values from the DNA microarrays; larger values indicate higher gene expression levels. Each sample column is followed by a column labeled ‘call’ – the entries in these columns are either ‘P’ (present) or ‘A’ (absent) that indicate whether the specific gene was expressed in that sample. Both files contain data for 7129 hybridization probes; most of these correspond to individual genes, although a handful serve as controls. The training dataset

¹ Golub, T.R. et. al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 286, 531-537 (1999).

contains tumor samples 1 through 38, while the independent dataset contains tumor samples 39 through 72.

A third .csv file, 'actual.csv,' contains two columns: 'patient' and 'cancer.' This file provides the labels for both the training and independent datasets; the 'patient' column contains entries from '1' to '72,' which correspond to the sample columns in the previous two .csv files, while the 'cancer' column labels each sample as either 'ALL' or 'AML.'

ii. Data Cleaning and Wrangling

The dataset was fairly clean and therefore easy to shape into useable form. I began by importing the training dataset into a pandas DataFrame, using the 'Gene Accession Number' column as the index. There were no missing or null values in the dataset, though I did need to remove extraneous columns and rows corresponding to the microarray control probes. As I am only interested in the actual gene expression values, the 'call' columns are unnecessary. I generated a list containing these column names (i.e. ['call.1', 'call.2', ...]), then passed this list to the drop() method to remove the columns.

The rows corresponding to assay controls are also not useful. Conveniently, I discovered that the control probes were at the head of the dataframe and contained the string 'AFFX' in the gene accession number (Affymetrix manufactured the microarray chips used). An additional control was missing the 'AFFX' string but was clearly labeled as a control in the 'Gene Description' column. I created a list containing the index values for these control probes and then removed them from the dataframe. I also dropped the 'Gene Description' column, as it is not useful at this stage. Having done this, I then mapped the gene expression columns as integers and sorted the index so that the patient samples were in order from 1 to 38. Finally, I transposed the dataframe so that the tumor samples (independent observations) formed the rows and the genes (the features) formed the columns.

Another critical step to enable exploratory data analysis was to label each tumor sample with the type of cancer. To do this, I read in the 'actual.csv' file containing the cancer labels for each sample, setting the patient column as the index. I was then able to merge the gene expression dataframe with the cancer key to label each sample with the appropriate form of

cancer in my dataframe. As the cancer key contained the independent samples as well as the training samples, I used an inner join so that only the samples that appeared in both the gene expression dataframe and the cancer key would be retained in the new dataframe.

III. Exploratory Data Analysis

I began my analysis by exploring the dataset for variations and patterns in the gene expression profiles of AML and ALL tumors. I first grouped the data by cancer type and calculated the mean and standard deviation for both the ALL and AML tumors for each gene and saved these calculations as a new dataframe, 'stats'. I was then able to visually explore the mean gene expression level across all genes for both types of cancer. Figure 1 shows a scatterplot of the mean expression level of the ALL samples versus the AML samples for each gene in the dataset. One clear outlier datapoint has large negative mean expression values for both cancer types; this datapoint corresponds to the gene SOX12, which is a transcription factor. Raw

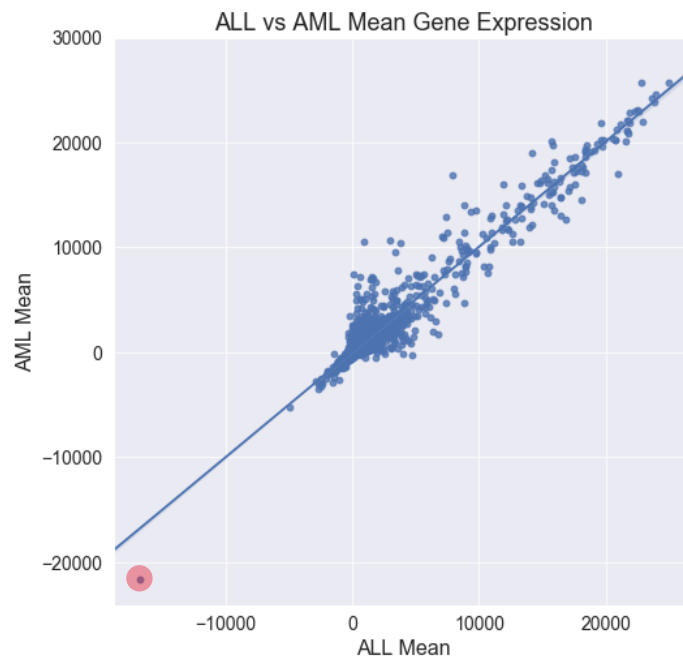


FIGURE 1: Mean gene expression level of ALL samples vs AML samples for each gene. Outlier datapoint for SOX12 gene expression is highlighted in red.

microarray data should be positive, though you can get negative values after data processing steps such as subtracting out background signal. I was unable to discover what processing and/or

transformations were applied to the raw dataset to yield the final values in the published dataset; however, it is likely that the extremely large negative values for SOX12 are either an error or an artifact of data processing that the original researchers used on the raw microarray data; therefore, this gene has been excluded from further analysis.

I next plotted the empirical cumulative distribution functions (ECDF) of the gene expression level for both the ALL and AML samples (Figure 2). As expected, both ECDF plots look similar for the different cancer types; only a small subset of the ~7000 genes tested in the microarrays are likely responsible for determining an ALL cell versus an AML cell. By zooming in on the ECDF plots, we can see that ~25% of the genes have negative expression levels (again, probably an effect of background subtraction on the raw dataset), ~50% of the genes (~3500 genes) have positive expression levels less than 500, ~12.5% have expression levels between 500 and 1000, and ~12.5% have expression levels greater than 1000 (Figure 3). The vast majority of the genes, over 70%, have low to moderate expression levels (between -500 to 500), while about 5% (~350 genes) have very high expression genes (over 4000).

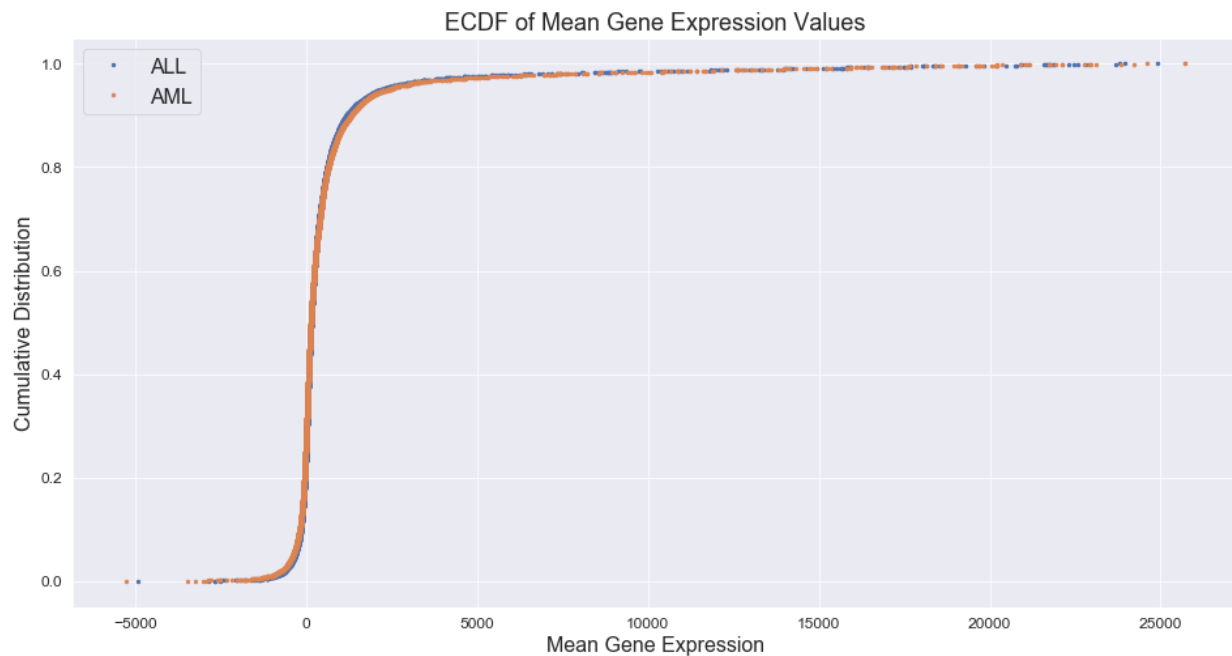


FIGURE 2: Empirical Cumulative Distribution Functions of gene expression levels for the ALL and AML samples.

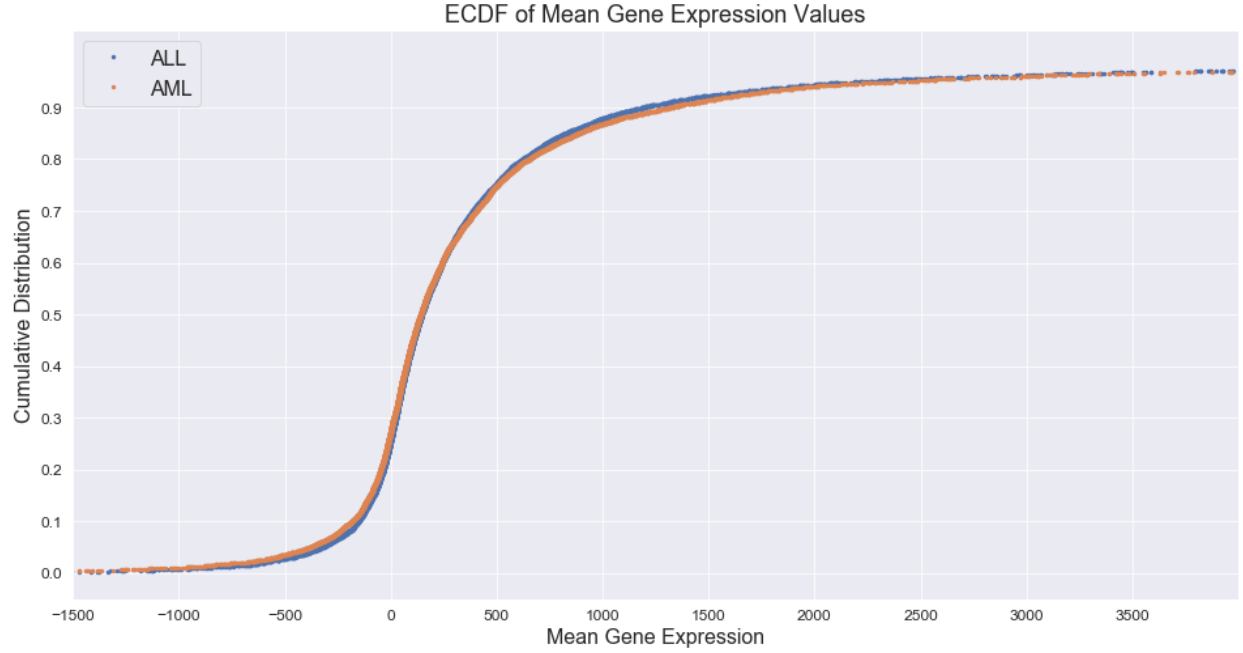


FIGURE 3: Zoom-in view of ECDFs for the ALL samples and the AML samples.

While most of the genes in the dataset have similar expression levels across all samples regardless of cancer type, a small subset of the genes should be more highly expressed in either the ALL or the AML samples. Identifying these differentially expressed genes will allow me to train classifiers using only the genes most highly correlated with the ALL/AML class distinction. The scatterplot of mean gene expression in ALL samples vs AML samples gives a rough picture of the number of genes that are differentially expressed between the two cancer variations; the points that lie above the main regression line are more highly expressed in the AML samples, while those that lie below the line are more highly expressed in the ALL samples. A rough estimate of correlation could be calculated using the ratio of mean gene expression in ALL samples over AML samples; however, the simple ratio of mean expression does not account for the spread of the individual samples for each cancer type. To better identify the genes that are most highly correlated with the ALL and AML samples, I need to account for both the mean expression level and the standard deviation of the mean for the ALL and AML samples. Therefore, I estimated the correlation between a gene and the cancer type by calculating the difference between the ALL and AML means divided by the sum of the standard deviation of the means:

$$\text{estimated correlation} = \frac{\mu_{ALL} - \mu_{AML}}{\sigma_{ALL} + \sigma_{AML}}$$

Genes that are more highly expressed in the ALL samples will have positive estimated correlation values, while genes that are more highly expressed in the AML samples will have negative estimated correlation values. A histogram of the correlation estimates for all genes shows that the correlation values are centered around 0 and are roughly normally distributed (Figure 4).

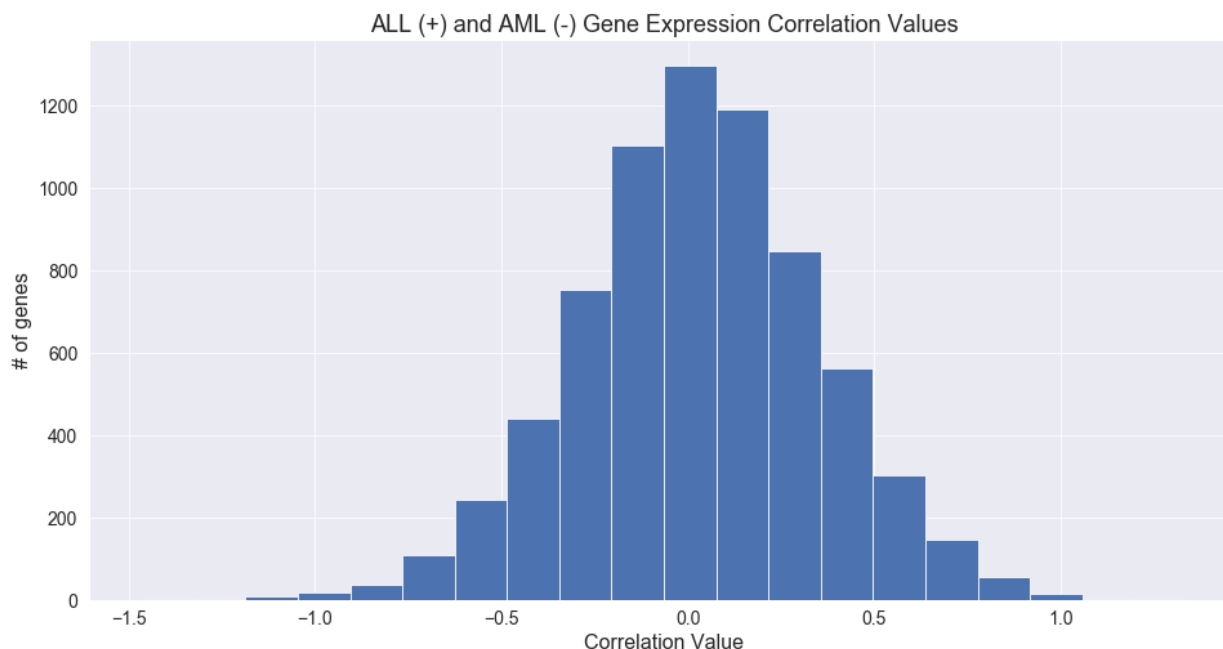


FIGURE 4: Histogram of estimated correlation values for all genes.

After sorting the genes by their correlation estimates, I was then able to pull out the 25 genes most correlated with each cancer type. The location of these genes on the scatterplot of mean gene expression in ALL v AML samples shows that many of genes correlated with cancer type have expression levels between 0 to 5000 (Figure 5). We can also see that the genes lie closer to the regression line at lower expression levels but further away as the expression level increases. Many of genes correlated with the AML samples have higher mean gene expression levels, and 8 of the 25 genes have expression levels over 5000. On the other hand, all of the genes correlated with the ALL samples have expression levels under 5000, and only 6 genes have expression levels over 2000 (Table 1).

I also generated ECDFs of the correlation estimates for both the ALL and the AML samples. As the sign of the correlation value indicates correlation with either ALL or AML, the ECDF of the correlation values for the ALL samples is simply the inverse of the ECDF for the



FIGURE 5: Location of the 50 genes with highest correlation estimates in the scatterplot of mean gene expression level of ALL samples vs AML samples.

AML samples. Thus, the inverse of the correlation estimates was used to calculate the ECDF of ALL cancer so that both the ALL and AML ECDFs could be plotted alongside one another (Figure 6). Examination of the ECDF plots for the correlation estimates show that ~100 genes

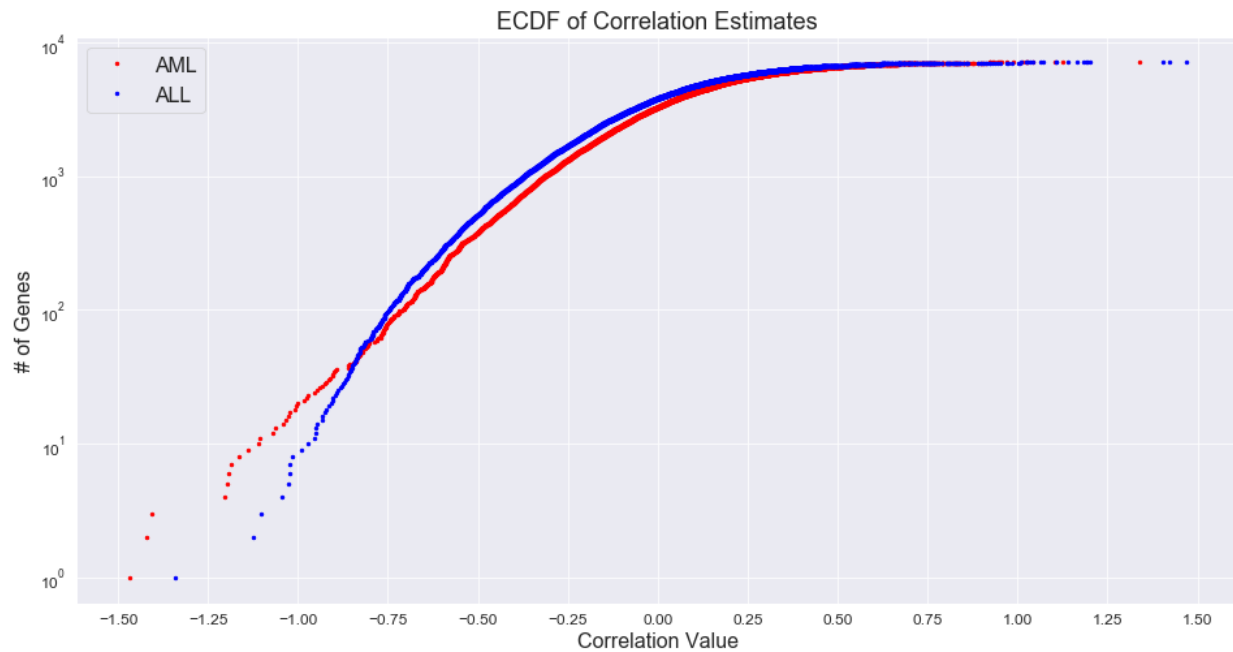


FIGURE 6: Empirical cumulative distribution function of correlation values for ALL and AML samples.

TABLE 1: Statistical summary information for top 50 genes with the highest correlation estimates

	ALL_mean	AML_mean	ALL_stdev	AML_stdev	mean_diff	std_sum	correlation
M55150_at	810.3	1836.27	338.18	360.89	-1025.98	699.06	-1.47
U50136_rna1_at	977.78	2562.18	324.69	789.75	-1584.4	1114.44	-1.42
X95735_at	349.89	3023.64	395.52	1506.46	-2673.75	1901.98	-1.41
M16038_at	375.37	1811.64	240.3	953.69	-1436.27	1193.99	-1.2
M23197_at	175.04	767.27	84.09	411.1	-592.24	495.19	-1.2
M84526_at	-177.41	3483.0	176.45	2896.84	-3660.41	3073.29	-1.19
Y12670_at	411.15	1294.45	212.37	533.21	-883.31	745.57	-1.18
U82759_at	241.37	850.73	239.56	283.87	-609.36	523.43	-1.16
D49950_at	75.37	243.91	54.69	93.18	-168.54	147.87	-1.14
M27891_at	144.44	7423.55	427.02	6135.52	-7279.1	6562.54	-1.11
X17042_at	1642.81	7108.73	1774.33	3167.84	-5465.91	4942.17	-1.11
U12471_cds1_at	152.22	287.18	67.4	58.76	-134.96	126.16	-1.07
U46751_at	1512.15	5291.27	534.44	3017.11	-3779.12	3551.55	-1.06
Y00787_s_at	892.22	10496.91	2052.56	7173.98	-9604.69	9226.54	-1.04
L08246_at	1066.56	3767.45	668.59	1941.52	-2700.9	2610.11	-1.03
M80254_at	33.74	428.64	106.58	277.66	-394.9	384.25	-1.03
M62762_at	1238.59	3070.55	689.8	1101.22	-1831.95	1791.03	-1.02
M81933_at	64.26	248.91	104.27	78.88	-184.65	183.15	-1.01
M96326_rna1_at	571.11	7117.55	527.23	5984.23	-6546.43	6511.46	-1.01
M28130_rna1_s_at	355.15	6298.91	939.0	4999.32	-5943.76	5938.32	-1.0
M63138_at	1779.26	5599.36	795.5	3088.19	-3820.1	3883.69	-0.98
M11147_at	7885.37	16838.09	3956.12	5236.04	-8952.72	9192.16	-0.97
M57710_at	262.85	3760.36	516.34	3084.56	-3497.51	3600.9	-0.97
M81695_s_at	548.19	1439.64	248.77	686.58	-891.45	935.35	-0.95
X85116_rna1_s_at	382.04	1842.73	329.23	1214.66	-1460.69	1543.89	-0.95
J05243_at	867.63	134.0	627.05	197.24	733.63	824.29	0.89
Z69881_at	2196.37	278.45	1883.6	269.31	1917.92	2152.91	0.89
U20998_at	1786.78	690.27	1007.18	216.46	1096.51	1223.64	0.9
X63469_at	315.04	88.0	172.58	79.25	227.04	251.82	0.9
D38073_at	913.41	324.73	473.19	178.77	588.68	651.95	0.9
U29175_at	1202.67	382.55	775.04	128.14	820.12	903.18	0.91
M91432_at	899.89	204.45	664.69	96.58	695.43	761.27	0.91
S50223_at	354.93	51.18	228.41	101.18	303.74	329.59	0.92
AF009426_at	67.85	-49.0	92.59	33.79	116.85	126.38	0.92
X15949_at	294.67	42.91	244.04	26.44	251.76	270.48	0.93
X52142_at	143.3	-70.73	141.72	87.62	214.02	229.35	0.93
Z15115_at	4213.0	1365.27	2518.89	493.94	2847.73	3012.83	0.95
M28170_at	210.11	-21.27	188.93	54.77	231.38	243.7	0.95
L47738_at	1251.74	192.73	897.85	216.42	1059.01	1114.27	0.95
U32944_at	1707.81	328.0	1296.54	148.49	1379.81	1445.02	0.95
M31523_at	1331.7	310.27	906.88	144.82	1021.43	1051.69	0.97
D26156_s_at	1306.0	578.82	520.5	214.69	727.18	735.2	0.99
U09087_s_at	180.89	46.91	106.37	25.72	133.98	132.09	1.01
M31211_s_at	553.48	154.64	287.57	102.4	398.85	389.97	1.02
L13278_at	127.93	0.55	93.87	30.56	127.38	124.43	1.02
X74262_at	1416.96	255.73	971.84	158.08	1161.24	1129.92	1.03
M92287_at	4029.48	1073.64	2352.36	481.48	2955.85	2833.83	1.04
U05259_rna1_at	3562.93	288.18	2726.43	242.03	3274.74	2968.46	1.1
X59417_at	4361.37	1138.18	2291.29	574.69	3223.19	2865.98	1.12
U22376_cds2_s_at	3863.3	702.36	1999.78	360.34	3160.93	2360.12	1.34

have correlation values of 0.75 or higher for both the ALL and AML samples.

Next, I went back to the original dataset to look at the spread of the gene expression data for each of the top 50 genes. For the following plots, I only looked at the gene expression data from the correlated cancer samples (Figures 7, 8). Overall, the genes in both cancer sets show a wide variety of expression levels, from genes with very low expression to those with quite high expression levels. However, as see in the scatterplot earlier, the more genes associated with AML have higher expression levels, and the maximum expression level is over twice as high as the maximum gene expression level in the ALL set. Indeed, if the top 50 genes were plotted

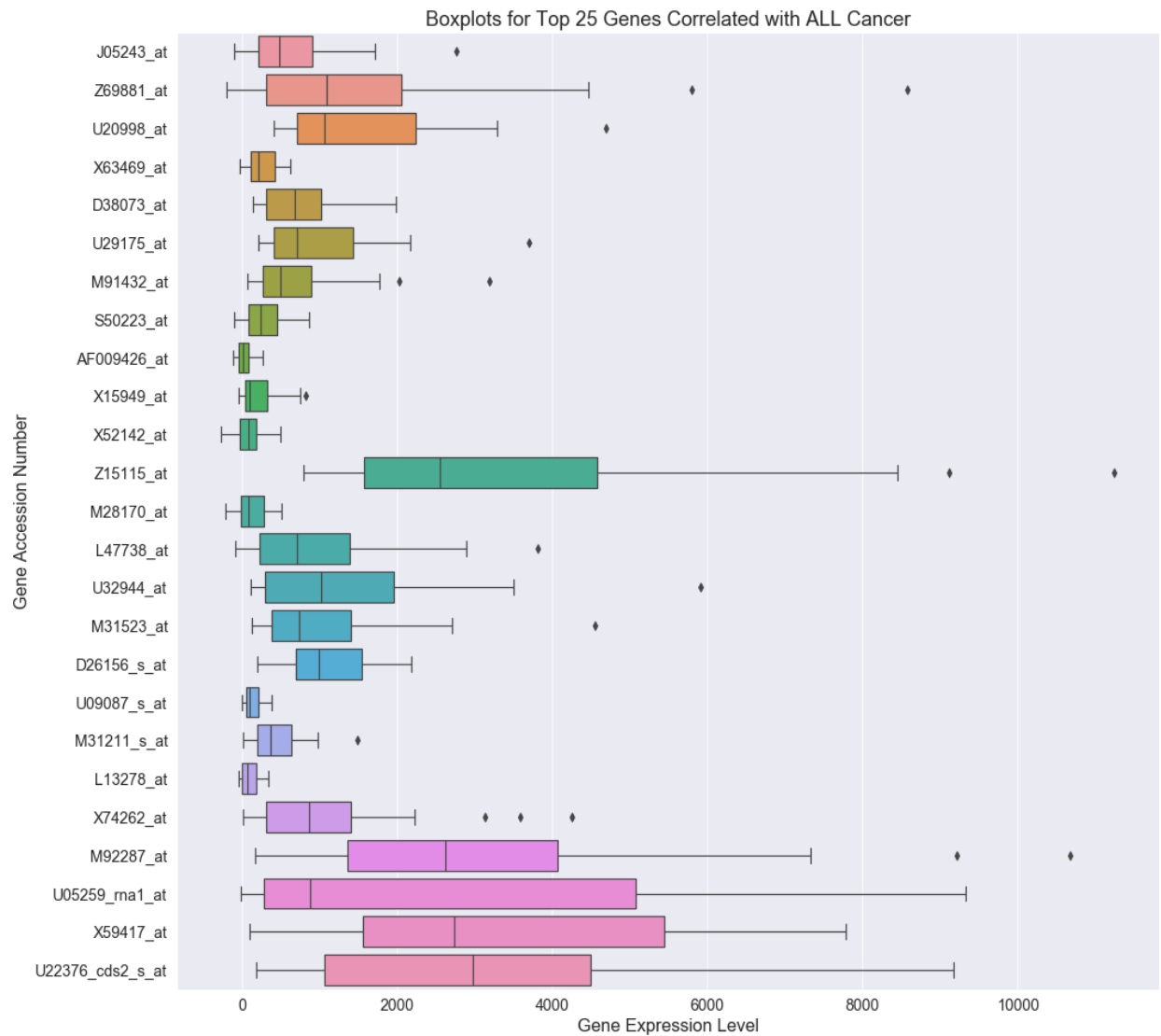


FIGURE 7: Boxplots of the gene expression data from the ALL samples for the top 25 genes correlated with the ALL samples.

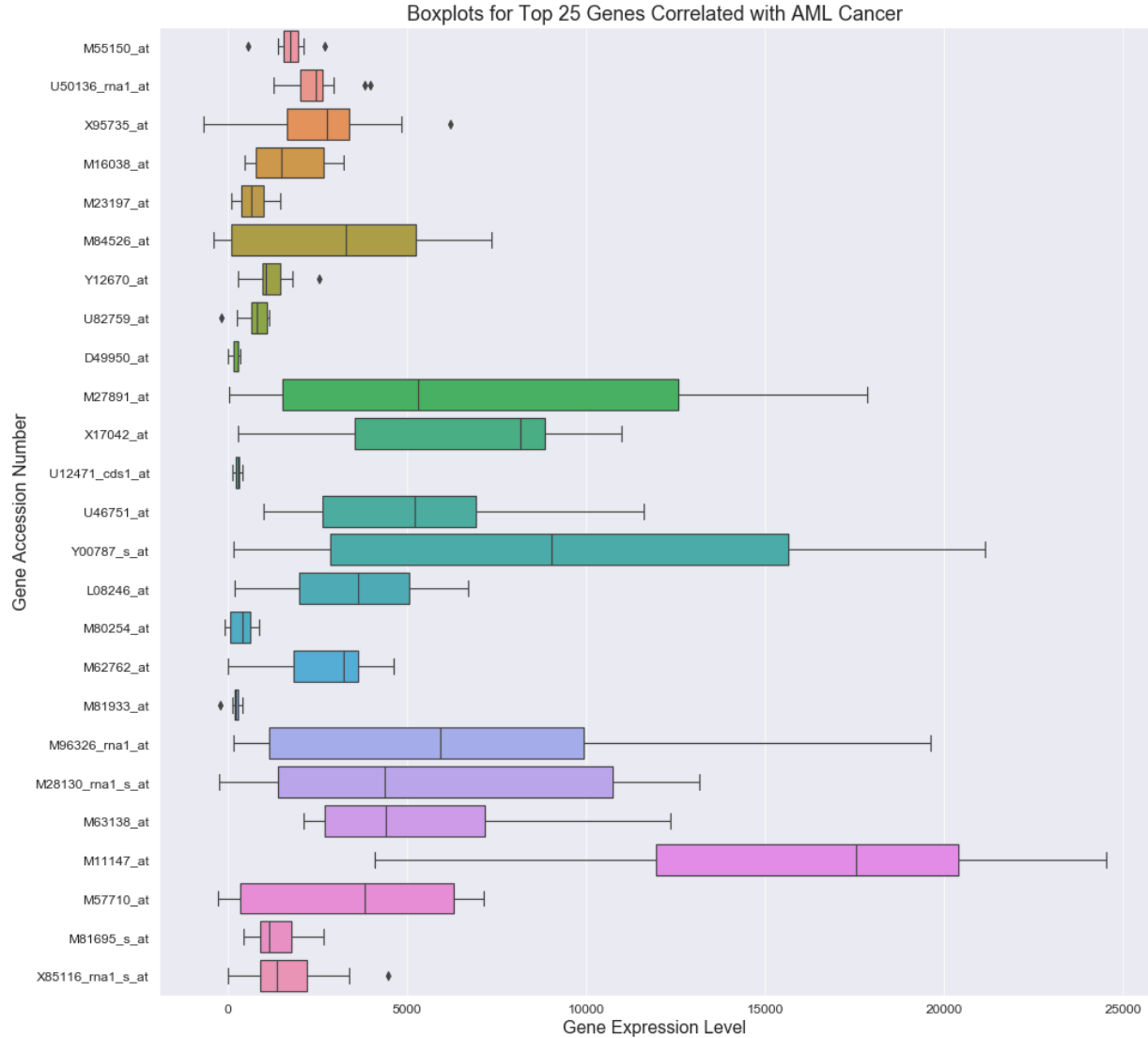


FIGURE 8: Boxplots of the gene expression data from the AML samples for the top 25 genes correlated with the AML samples.

together, the boxplots for all of the ALL-associated genes would fit into the first half of the plot. Another general feature is that the spread of the gene expression data increases as the expression level increases, suggesting that the data is logarithmically distributed and should be log-transformed to normalize data prior to further analysis. Finally, we can also see that about a third of the genes have between 1-3 outlier values. If necessary, these outlier values could be analyzed later on to refine the gene selection (i.e. feature selection) process.

One of the best ways to visualize gene expression data is by using a heatmap, which makes it very easy to see patterns in the gene expression levels across many samples. As the expression levels of the top 50 genes vary widely, each gene was first normalized using the

maximum and minimum expression levels so that patterns across the individual samples could be detected. As you can see in the heatmap below, there is a clear difference in the gene expression pattern between the ALL samples and AML samples (Figure 9). As expected, the top 25 genes are more highly expressed in the ALL samples compared to the AML samples, while the bottom 25 genes are more highly expressed in the AML samples.

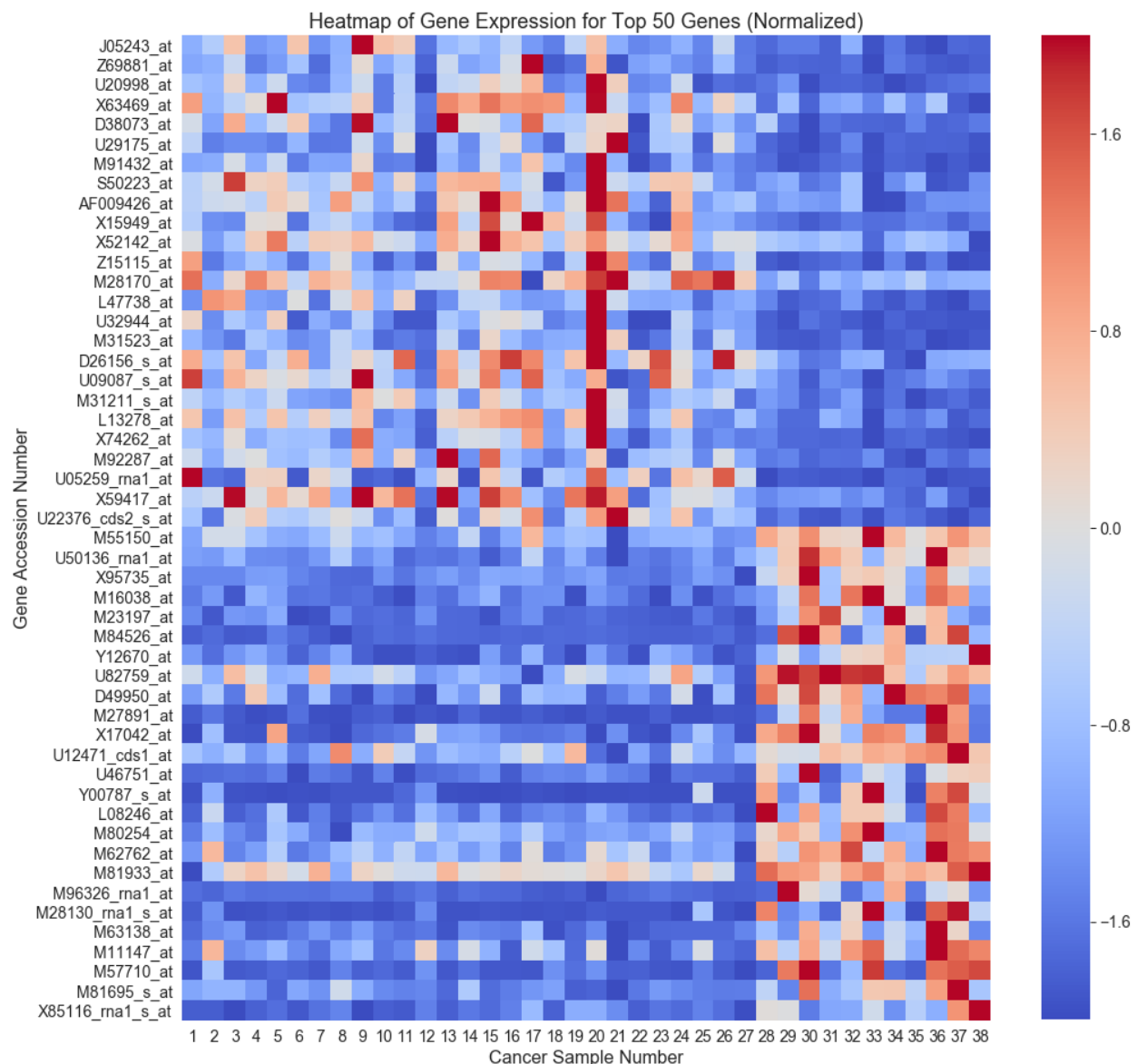


FIGURE 9: Heatmap of normalized gene expression data for the top 50 gene associated with the ALL/AML distinction. Samples 1-27 are from patients with ALL, while samples 28-38 are from patients with AML cancer. Red indicates upregulated (enhanced) gene expression, while blue indicates downregulated (suppressed) gene expression.

Finally, I computed p-values and confidence intervals around the correlation estimates for the top 50 genes. To do this, I tested the null hypothesis that the gene expression levels of the ALL and AML samples have identical probability distributions. The first step was to combine the ALL and AML gene expression data for each gene, generate permutation samples by randomly splitting the gene expression data into ALL and AML samples, and then calculate the correlation values on these permutation samples. Correlation estimates for 10,000 permutation samples were calculated per gene, and the p-values (the probability of the getting the observed correlation estimate if there is no difference between ALL and AML cancer) were then calculated by dividing the number of permutation samples that resulted in correlation estimates at least as extreme as the observed value by the total number of permutations performed. At most, only 1-2 permutations out of 10,000 yielded correlation estimates equal to or greater than the observed correlation estimate. Therefore, the observed p-values less than 0.001 for all of the 50 genes most associated with cancer type (Table 2). Bootstrap analysis was used to calculate confidence intervals for the correlation estimates for each of the top 50 genes. For these calculations, two bootstrap replicates were generated for each gene, one for the ALL samples and one for the AML samples; these bootstrap replicates were then used to calculate the resulting correlation value. For each gene, 10,000 correlation estimates were calculated, and the resulting distribution was used to determine the 95% confidence interval of the correlation estimate for that gene (Table 2).

TABLE 2: Probabilities and 95% confidence intervals for correlation estimates for top 50 genes with the highest correlation estimates.

	correlation	p-value	95% CI
M55150_at	-1.47	0.0	[-2.29 -1.13]
U50136_rna1_at	-1.42	0.0	[-2.23 -1.06]
X95735_at	-1.41	0.0	[-2.24 -1.09]
M16038_at	-1.2	0.0	[-1.96 -0.85]
M23197_at	-1.2	0.0	[-1.87 -0.87]
M84526_at	-1.19	0.0	[-2.21 -0.73]
Y12670_at	-1.18	0.0	[-1.87 -0.92]
U82759_at	-1.16	0.0	[-1.99 -0.75]
D49950_at	-1.14	0.0	[-2.22 -0.7]
M27891_at	-1.11	0.0	[-1.84 -0.77]
X17042_at	-1.11	0.0	[-2.2 -0.66]
U12471_cds1_at	-1.07	0.0	[-1.71 -0.75]
U46751_at	-1.06	0.0	[-1.99 -0.67]
Y00787_s_at	-1.04	0.0	[-1.89 -0.63]
L08246_at	-1.03	0.0	[-1.79 -0.65]
M80254_at	-1.03	0.0	[-1.8 -0.64]
M62762_at	-1.02	0.0	[-1.87 -0.62]
M81933_at	-1.01	0.0	[-1.43 -0.8]
M96326_rna1_at	-1.01	0.0	[-1.72 -0.67]
M28130_rna1_s_at	-1.0	0.0	[-1.77 -0.62]
M63138_at	-0.98	0.0	[-1.58 -0.7]
M11147_at	-0.97	0.0	[-1.69 -0.6]
M57710_at	-0.97	0.0001	[-1.89 -0.54]
M81695_s_at	-0.95	0.0	[-1.55 -0.61]
X85116_rna1_s_at	-0.95	0.0	[-1.5 -0.67]
J05243_at	0.89	0.0	[0.64 1.32]
Z69881_at	0.89	0.0	[0.69 1.3]
U20998_at	0.9	0.0001	[0.64 1.31]
X63469_at	0.9	0.0001	[0.62 1.32]
D38073_at	0.9	0.0	[0.57 1.5]
U29175_at	0.91	0.0	[0.68 1.34]
M91432_at	0.91	0.0	[0.72 1.34]
S50223_at	0.92	0.0001	[0.65 1.35]
AF009426_at	0.92	0.0	[0.67 1.31]
X15949_at	0.93	0.0	[0.69 1.28]
X52142_at	0.93	0.0	[0.71 1.34]
Z15115_at	0.95	0.0	[0.73 1.31]
M28170_at	0.95	0.0001	[0.63 1.45]
L47738_at	0.95	0.0	[0.72 1.34]
U32944_at	0.95	0.0	[0.72 1.38]
M31523_at	0.97	0.0	[0.75 1.48]
D26156_s_at	0.99	0.0001	[0.71 1.41]
U09087_s_at	1.01	0.0	[0.75 1.42]
M31211_s_at	1.02	0.0	[0.77 1.52]
L13278_at	1.02	0.0	[0.73 1.5]
X74262_at	1.03	0.0	[0.82 1.45]
M92287_at	1.04	0.0	[0.83 1.42]
U05259_rna1_at	1.1	0.0	[0.82 1.55]
X59417_at	1.12	0.0001	[0.85 1.56]
U22376_cds2_s_at	1.34	0.0	[1.04 1.87]