Classifying Leukemia Patients Based on DNA Microarray Data

Veronica Peng
Product Design
tpeng24

Allan Li
Mathematics
shilun

Sarah Tran
Computer Science
sdtran

1. Introduction

Our project aims to diagnose whether a patient has acute myeloid leukemia (AML) or acute lymphoblastic leukemia (ALL) based on their genes. Leukemia is a group of cancers that affect blood cells [1]. The most common form of leukemia in children is ALL, and the most common form of acute leukemia in adults is AML [1].

ALL is a type of cancer that develops in immature lymphocytes, a type of white blood cell [2]. There are two main types of lymphocytes: B cells and T cells. B cells can mature into plasma cells, which produce antibodies or become memory B cells, which are important to secondary immune responses [2]. T cells have a diverse set of functions, including directly killing virus-infected cells [2]-[3].

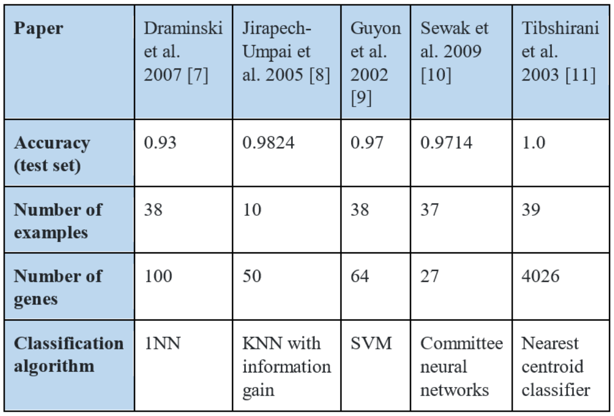
In 2015, over two million people had leukemia, and leukemia caused over 350,000 deaths [4]. Acute forms of leukemia are faster progressing and typically more aggressive than chronic forms, and if left untreated, could be fatal in a few months, so early diagnosis is especially important [1], [5]. Currently, cytochemistry and lineage phenotypes are usually used in combination to identify which type of leukemia the patient has. Both methods are subject to certain limitations and even when used in combination, may not yield a reliable diagnosis.

For example, there is a rare and aggressive type of leukemia (4% in all leukemia patients) called mixed-phenotype acute leukemia (MPAL), and lineage plasticity is a hallmark feature of MPAL [6]. Lineage plasticity can be tested for in cell cultures, but not directly in a clinical setting [6]. Preijers et al. conducted a 15-year study on checking the reliability of flow cytometry assays in diagnosing leukemia, and found that many technicians perform the assay incorrectly, leading to inaccurate results [18].

A correct diagnosis is key to determine a patient's likely response to certain treatments. We are thus interested in increasing the diagnostic accuracy by exploring the possibility of diagnosing leukemia based on the patient's genes.

2. Related Work

Surveying the literature, we find that there have been a wide array of techniques used to classify leukemia patients in the Golub dataset, some of which are novel, as in the committee neural networks in Sewak et al. 2009 [10]. The models in the literature vary in the number of input features, ranging from 27 to 4,026. We aim to improve on this work by selecting fewer features for a more parsimonious model with less of a potential to overfit, yet still have comparable accuracy.



Paper	Draminski et al. 2007 [7]	Jirapech-Umpai et al. 2005 [8]	Guyon et al. 2002 [9]	Sewak et al. 2009 [10]	Tibshirani et al. 2003 [11]
Accuracy (test set)	0.93	0.9824	0.97	0.9714	1.0
Number of examples	38	10	38	37	39
Number of genes	100	50	64	27	4026
Classification algorithm	1NN	KNN with information gain	SVM	Committee neural networks	Nearest centroid classifier

Figure 1: A review of the classification methods used on the Golub dataset, as well as the number of features used in the model.

4 out of the 6 papers we reviewed also published which features their model found important. We looked at the top 10 most important features (although [9] only showed the top 7 features, and [22] showed the top 4 features). Only 3 out of 27 genes were re-selected across methods (zyxin, CD33, and elastase), and genes were only selected at most twice over the four papers.

Data from microarrays also inherently present challenges. Although there is a large amount of data generated from a microarray, there is no formal guideline for standardizing the data in the literature, although some journals such as *Nature* and *Cell* independently use the MIAME standard [12].

There are also difficulties replicating the data, since there are many small changes in data collection that can influence the data, such as whether RNA was extracted once and then aliquoted, or extracted multiple times [13]. Finally, there is growing concern about the ability to store microarray data, as the technology improves to become more efficient and generate data at a higher resolution [14].

3. Dataset and Features

We will use the Golub dataset, which has the expression levels of 7,129 genes from 72 subjects [15]. Bone marrow and peripheral blood samples were collected from the patients and assayed with HGU6800 chips.

The dataset has been split into a training and test dataset, which contain the data from 38 and 34 patients respectively. 29% of the training set represents AML patients, while 42% of the test data set are AML patients.

We will also use the Chiaretti et al. 2004 dataset [16]. The dataset has 128 subjects with ALL along with the expression levels of 12,625 genes assayed using HGU95aV2 gene chips. We will try to classify patients with ALL into two subcategories: ALL affecting B cells and ALL affecting T cells.

4. Methods

For the baseline, we sampled 10 random genes without replacement and used those as input to the logistic regression model. We did this 100 times to get an estimate of the average classification accuracy of AML and ALL using 10 random genes.

We then did feature selection to select the top 10 most informative genes using two-sample Student's t-tests. A significance level of less than 0.0001 indicates that the expression level of the gene differed between AML and ALL patients. Because there were many features, we also performed the Benjamini-Hochberg procedure to control the false discovery or type I error rate. The procedure is to sort the p-values from smallest to largest. Let *m* be the number of null hypotheses test. If the i-th p-value is less than α * i / m, then it is considered significant [17].

Because the alpha level was 0.0001 and the number of features was 7,129, on average we expect less than 0.0001 * 7,129 or less than one of the null genes will be considered significant.

The 10 genes with the lowest p-values were selected. We then decomposed those features using PCA to reduce collinearity, taking the first component. Then these features will be used in logistic regression to classify the patients as having AML or ALL. We performed 10-fold cross-validation on all models.

To test the robustness of our methods on microarray data, we also classified ALL subtypes (B cell and T cell) using the Chiaretti dataset.

Since the Golub and Chiaretti datasets have many more features than data points, the models may overfit. We experimented with lasso regression to avoid this using the L1 norm and to reduce the dimensionality, computing a regularization path for each classification task. Using the accuracy from cross-validation (CV), we selected a regularization strength that minimized CV error plus one standard error. This is a standard technique when selecting a model, as the addition of one standard error makes feature selection more parsimonious [19].

L1 is more efficient at selecting features than L2 regularization if there are many more irrelevant features than there are training examples [20]. The sample complexity grows at a logarithmic rate relative to the number of irrelevant features, while L2 grows linearly [20]. L1 regularization induces a sparser selection of features than L2, since it imposes on ordering on when features enter and leave the model, driving more coefficient values to 0 [21].

We also wanted to help the audience better understand our predictions by visualizing the data. Specifically, we generated heatmaps on how certain genes relate to different types of leukemia by biclustering. We used hierarchical agglomerative cluster analysis with Euclidean distance as the metric. We also checked the stability of the AML/ALL clusters using consensus clustering. This involved selecting the top 100 most informative genes, randomly selecting 15 at each iteration for 500 iterations to use to cluster the subjects, and averaging the number of times each subject appeared in the same cluster as another subject

5. Results and Discussion

The baseline logistic regression model used to classify AML/ALL patients (Golub dataset) using 10 random genes had a mean accuracy on the test set of 67%, with a standard deviation of 9%.

Using two-sample t-test, we selected 10 genes: *RAB1A, M29037, M84605, calpain 2, N-WASP, U50146*, HG1614*, progesterone receptor membrane component, integrin subunit alpha 7,* and *elastase*. All associated p-values were significant (p < 0.0001). Without the Benjamini–Hochberg (BH) procedure, we estimated 305 genes to differ significantly between the two groups. With, we estimated 187 genes.

The accuracy of the logistic regression on the test set with the 10 features was 100% on the test and training set. Using

10-fold cross-validation, the mean accuracy was 95.7%; the highest was 100% and the lowest 85.7%.

We then used PCA to transform the data associated with the 10 genes. We used only the first component, which accounted for 78% of the variance in the dataset. The accuracy of the logistic regression with these transformed features was 100% on the training set and 97% on the test set. Using 10-fold cross-validation, the mean accuracy was 92%; the highest was 100% and the lowest 57%. The decrease in accuracy may be due to the fact that the genes selected were not closely related to each other; they had a wide variety of functions related to controlling appetite, membrane traffic, and breaking down tissue, and more. Interestingly, genes RAB1A, N-WASP, and integrin subunit alpha 7 have functions related to the cell membrane [23]-[25], and HG1614 affects cell division [26]. These results are intuitive given that we are interested in classifying two types of cancer.

Only one of the 10 genes selected using t-tests were reported as important genes in other papers (elastase). However, as we stated in the related works section, only 3 out of 27 genes across the papers we reviewed were selected as important more than once.

As mentioned in the methods section, we used the above strategy in another classification task related to leukemia and microarray data. We did this on the Chiaretti dataset, classifying ALL subtypes into B cell-type and T cell-type. The baseline had an accuracy of 89.5% on the training set 15.8% on the test set. Without PCA, and taking the top 10 features based on p-values, there was a 100% accuracy on the training set and 96.9% accuracy on the test set. With PCA, and only taking the top 5 features, there was a 100% accuracy on the test and training set, and the mean accuracy using 10-fold CV was 98.8%. Using t-tests, we rejected 857 null hypotheses before the BH procedure, and 306 after.

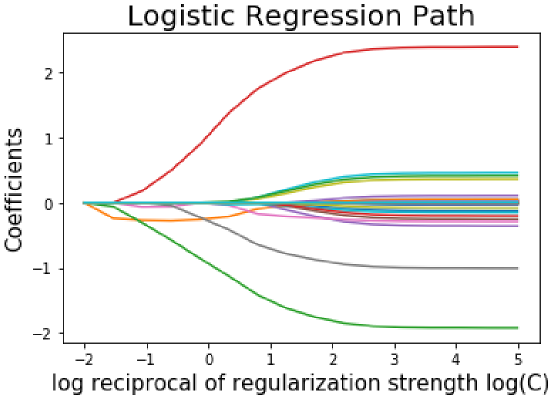


Figure 2: Logistic regression models on the Golub dataset. Regularization path using an L1 penalty.

We had planned to use more sophisticated classification models such as Gaussian discriminant analysis, but unexpectedly, logistic regression did well. The model on classifying ALL and AML types of leukemia based on the Golub dataset was comparable to more complex models in the literature using the same dataset (see figure 1); their classification rates on the test dataset ranged from 93% to 98% using a moderate amount of features, and one paper reported 100% accuracy using more than 4,000 genes.

To account for overfitting on the Golub dataset (figure 2), we used L1 regularization. The model we selected had a regularization strength of 2.78. It achieved 100% accuracy on the training set, and 95.5% accuracy on the test set. 45 out of 50 features were eliminated, leaving only CD44, CETN2, SUMO1, M11119,* and L17325* (asterisks represent genes that do not have a formal name, and so are represented by their GenBank ID). CD44 is involved in tumor metastasis and has been widely studied as a marker of carcinogenesis [27]. SUMO1 helps mediate transcriptional regulation and apoptosis [28], and CETN2 is a part of the centrosome, which is important for cell division [29].

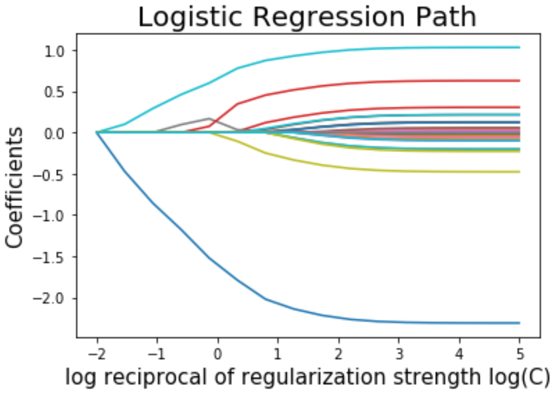


Figure 3: Logistic regression models on the Chiaretti dataset. Regularization path using an L1 penalty.

The regularization strength for the model we selected was 21.54, a relatively high constant. Nevertheless, the training and test accuracy was 100%. In the model, 48 out of 50 features were zeroed out. The features left were the genes that encoded for solute carrier family 25 (aspartate/glutamate carrier) member 13 (SLC25A13) and 2,4-dienoyl CoA reductase 1, mitochondrial (DECR1). Both are involved in mitochondrial function [30]-[31].

Then, to visualize and understand the relationship between the genes and patient classification, we used hierarchical agglomerative cluster analysis (HAC). We used Euclidean distance as the metric. In the figures below, the dark regions represent a low expression of the genes, while the light regions represent a high expression.

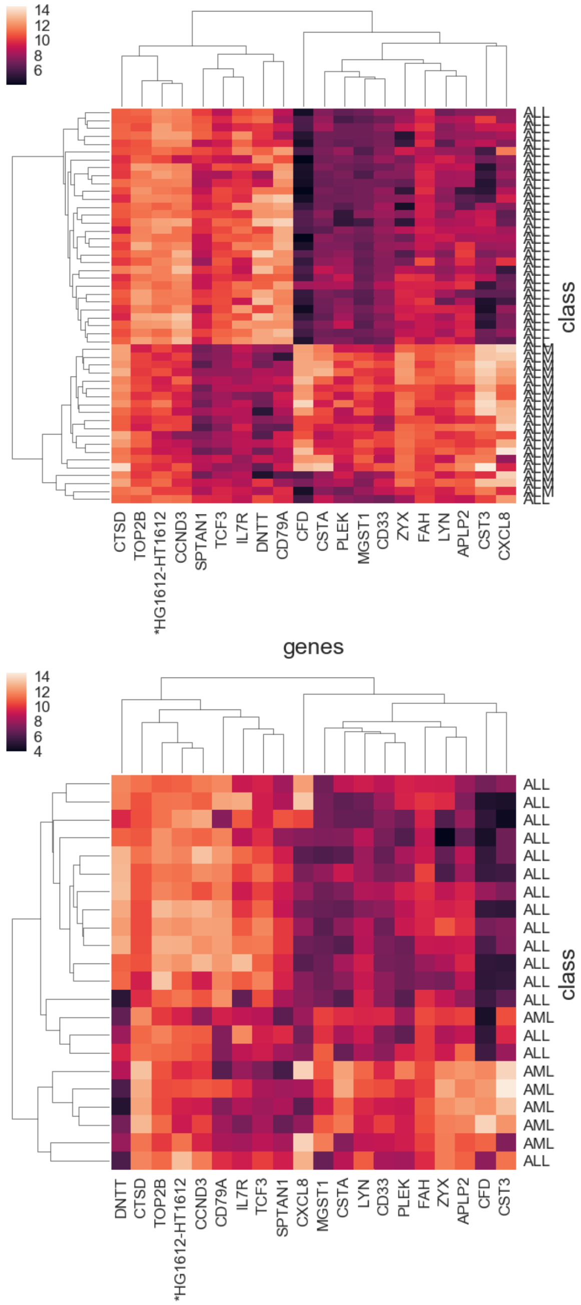


Figure 4: Hierarchical clustering on the Golub dataset. Clustering on the training set is on the top, while the clustering results on the test set is on the bottom.

We will first discuss the AML/ALL clustering on the Golub dataset. We see in the cluster map on the training set that the clusters are clearly delineated. Looking at the dendrogram for the classes, we see that if we cut the tree so that we have only two clusters, HAC cleanly separates ALL patients from AML patients. Out of 36 examples, only one patient was misclassified (i.e., a subject with ALL was put into the AML cluster).

In the training set, genes CFD, CSTA, PLEK, MGST1, CD33, ZYX, FAH, LYN, APLP2, CST3, and CXCL8 were highly expressed in patients with AML relative to those with ALL, and genes SPTAN1, TCF3, IL7R, DNTT, and CD79A were more highly expressed in ALL over AML patients. CFD had an especially strong decrease in expression in AML patients, and similarly SPTAN1 with ALL patients. Additionally, the gene MGST1 had an evenly low expression in ALL subjects and a higher expression in AML subjects. CFD has a role in suppressing infections, and SPTAN1 is involved in cell cycle regulation [32]-[33]. MGST1 mediates inflammation [34]. Given that cell cycle dysregulation is an important part of cancer, and that white blood cells are important for the immune system, the fact that these genes are important in classifying leukemia is intuitive.

In the test set, 4 out of 36 examples were misclassified. All misclassifications resulted from ALL subjects appearing in the AML cluster. If we split the AML cluster into two smaller clusters, we see that 3 out of 4 ALL misclassifications appear in the same sub-cluster. We also see that gene CFD, identified in the training set as a highly discriminative feature that was less expressed in ALL subjects, also had a low level of expression in all misclassified ALL subjects. It seems that clustering those 4 subjects into the AML cluster was highly motivated by the gene DNTT, which had an especially low expression in the "AML" cluster. DNTT is expressed in pre-B and pre-T lymphocytes [35].

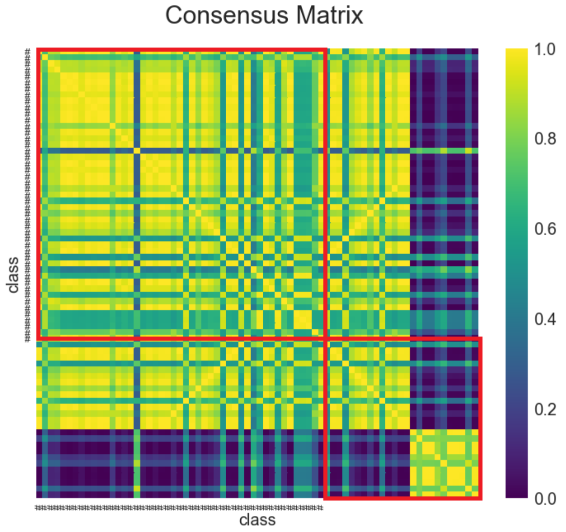
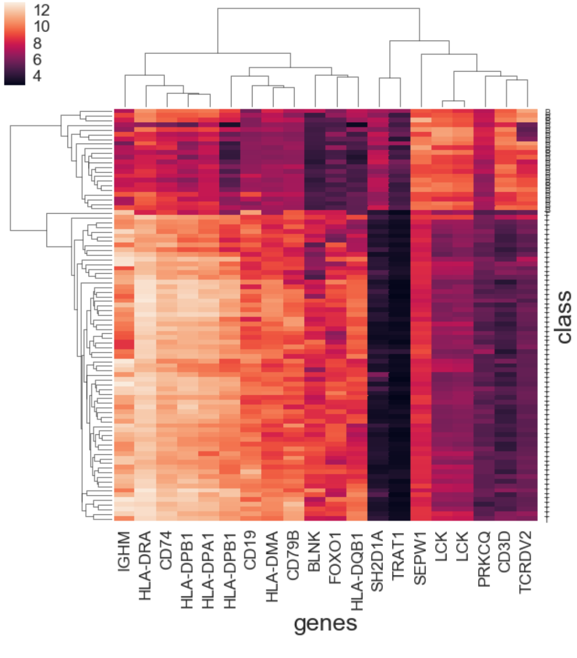


Figure 5: The hash symbol represents ALL patients. Lighter regions represent high cluster fidelity, i.e., these two patients were often clustered together. The color bar represents the proportion over 500 iterations.

Since the model did not perform as well on the test set in clustering ALL and AML subtypes, we checked for cluster stability using cluster ensembles. As seen in the upper red box, ALL patients were very often in the same cluster, which suggests a degree of heterogeneity. In contrast, AML patients seemed to fall into two subclasses which were dissimilar to each other. As suggested by other papers, this recapitulates the need for more stringent statistical tests for microarray data, as the high number of features increases the number of false positives [36]-[37].

We will now look at the ALL subtype classification task using the Chiaretti dataset (figure 6). In both training set and the test set, both B cell-type and T cell-type ALL are cleanly classified. The model generalizes well to the test set because the training set is representative of the test set, as one can see from the visualization, since the cluster maps look extremely similar.

In both sets, SH2D1A and TRAT1 are expressed more in patients with B cell-type than T cell-type ALL, and similarly IGHM, HLA-DRA, CD74, HLA-DPB1, HLA-DPA1, CD19, HLA-DMA, CD79B, BLNK, FOXO1, and HLA_DQB1 in the reverse direction. The human leukocyte antigen (HLA) system is involved in immune response, and particularly in presenting antigens to T cells [38].



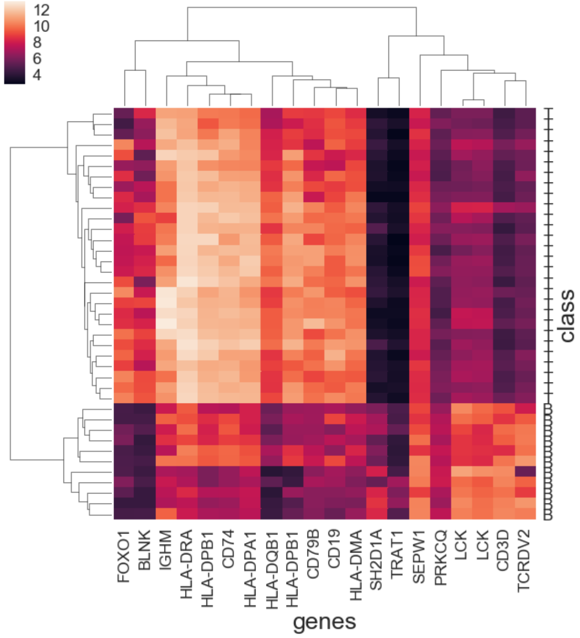


Figure 6: Hierarchical clustering on the Chiaretti dataset. Left represents training set, right represents test set. LCK and HLA-DPB1 are represented twice because there were two DNA probes that assessed each of their expressions.

6. Conclusion

In this project, we presented conceptually simple but well-performing models to increase the diagnostic accuracy in two aspects based on patient's gene: 1) differentiating two subtypes of leukemia, ALL and AML, and 2) differentiating two subtypes of ALL, one affecting B cells and the other affecting T cells. We conducted two-sample Student's t-tests and PCA to pre-process the data and avoid overfitting since there are many more features than data points. We applied 1) logistic regression, and 2) lasso regression, a variation of logistic regression that automates the feature selection to make predictions. For both algorithms, we also performed cross-validation to reduce false discovery. In order to visualize the relationship between genes expression and leukemia classification, we used hierarchical agglomerative cluster analysis (HAC) to generate heatmaps. Our algorithm yields high accuracy rate for the prediction and reveals previously unreported key indicator genes for leukemia subtypes. In the future, we may apply our algorithms on other gene classification problems including the following directions: 1) identifying what genes leads to high risk for certain diseases such as Alzheimer and Parkinson, 2) classifying whether a person has certain traits such as flat feet or the ability to have perfect pitch based on genetic information.

Contributions

Sarah Tran:

- Modify the project proposal
- Write the method and preliminary experiments section in the project milestone
- Write the related work, dataset and features, methods, and results and discussion section in the final report
- Implement the cross validation, consensus cluster, and Benjamini–Hochberg procedure

Allan Li

- Search for datasets
- Write the method and intended experiments sections in the project proposal
- Write the outline of methods, and results and discussion section for the final report
- Implement the baseline, the t-test, PCA, HAC, logistic regression and lasso regression

Veronica Peng

- Search for datasets
- Write the motivation section in the project proposal
- Write the next steps section in the project milestone
- Write the conclusion section in the final report
- Design and print the poster

References

[1] "Leukemia." *National Institute of Health*, n.d., https://www.cancer.gov/types/leukemia. Accessed 8 June 2019.

[2] Kierszenbaum, Abraham L., and Laura Tres. *Histology and Cell Biology: an introduction to pathology E-Book*. Elsevier Health Sciences, 2015.

[3] Janeway, Charles A., et al. *Immunobiology: the immune system in health and disease*. Vol. 7. London: Current Biology, 1996.

[4] Vos, Theo, et al. "Global, regional, and national incidence, prevalence, and years lived with disability for 310 diseases and injuries, 1990–2015: a systematic analysis for the Global Burden of Disease Study 2015." *The Lancet* 388.10053 (2016): 1545-1602.

[5] Beghi, E., and G. Logroscino, eds. *Tumors in adolescents and young adults*. Karger Medical and Scientific Publishers, 2016.

[6] Charles, Nathan J., and Daniel F. Boyer. "Mixed-phenotype acute leukemia: diagnostic criteria and pitfalls." *Archives of pathology & laboratory medicine* 141.11 (2017): 1462-1468.

[7] Dramiński, Michał, et al. "Monte Carlo feature selection for supervised classification." *Bioinformatics* 24.1 (2007): 110-117.

[8] Jirapech-Umpai, Thanyaluk, and Stuart Aitken. "Feature selection and classification for microarray data analysis: Evolutionary methods for identifying predictive genes." *BMC bioinformatics* 6.1 (2005): 148.

[9] Guyon, Isabelle, et al. "Gene selection for cancer classification using support vector machines." *Machine learning* 46.1-3 (2002): 389-422.

[10] Sewak, Mihir S., Narender P. Reddy, and Zhong-Hui Duan. "Gene expression based leukemia sub-classification using committee neural networks." *Bioinformatics and biology insights* 3 (2009): BBI-S2908.

[11] Tibshirani, Robert, et al. "Class prediction by nearest shrunken centroids, with applications to DNA microarrays." *Statistical Science* 18.1 (2003): 104-117.

[12] Constans, Aileen. "State of the Microarray: Challenges and Concerns with Microarrays." *The Scientist*, 10 Feb 2003, https://www.the-scientist.com/technology-profile/state-of-the-microarray-challenges-and-concerns-with-microarrays-52124. Accessed 8 June 2019.

[13] Churchill, Gary A. "Fundamentals of experimental design for cDNA microarrays." Nature genetics 32.4s (2002): 490.

[14] Guzzi, Pietro Hiram, and Mario Cannataro. "Challenges in microarray data management and analysis." *2011 24th International Symposium on Computer-Based Medical Systems (CBMS)*. IEEE, 2011.

[15] Golub, Todd R., et al. "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring." *Science* 286.5439 (1999): 531-537.

[16] Chiaretti, Sabina, et al. "Gene expression profile of adult T-cell acute lymphocytic leukemia identifies distinct subsets of patients with different response to therapy and survival." *Blood* 103.7 (2004): 2771-2778.

[17] "False Discovery Rate." *Columbia University Mailman School of Public Health*, n.d, https://www.mailman.columbia.edu/research/population-health-methods/false-discovery-rate. Accessed 8 June 2019.

[18] Preijers, Frank WMB, et al. "Fifteen years of external quality assessment in leukemia/lymphoma immunophenotyping in The Netherlands and Belgium: A way forward." *Cytometry Part B: Clinical Cytometry* 90.3 (2016): 267-278.

[19] Krstajic, Damjan, et al. "Cross-validation pitfalls when selecting and assessing regression and classification models." *Journal of cheminformatics* 6.1 (2014): 10.

[20] Ng, Andrew Y. "Feature selection, L1 vs. L2 regularization, and rotational invariance." *Proceedings of the twenty-first international conference on Machine learning*. ACM, 2004.

[21] McCaffrey, James. "Test Run - L1 and L2 Regularization for Machine Learning." *Microsoft Magazine*, Feb 2015, https://msdn.microsoft.com/en-us/magazine/dn904675.aspx. Accessed 8 June 2019.

[22] Subramani, Prabakaran, Rajendra Sahu, and Shekhar Verma. "Feature selection using Haar wavelet power spectrum." *BMC bioinformatics* 7.1 (2006): 432.

[23] Asaoka, Rin, et al. "Arabidopsis RABA1 GTPases are involved in transport between the trans-Golgi network and the plasma membrane, and are required for salinity stress tolerance." *The Plant Journal* 73.2 (2013): 240-249.

[24] Rohatgi, Rajat, et al. "The interaction between N-WASP and the Arp2/3 complex links Cdc42-dependent signals to actin assembly." *Cell* 97.2 (1999): 221-231.

[25] Ziober, B. L., et al. "Alternative extracellular and cytoplasmic domains of the integrin alpha 7 subunit are differentially expressed during development." *Journal of Biological Chemistry* 268.35 (1993): 26773-26783.

[26] McGill, Gaël G., et al. "Bcl2 regulation by the melanocyte master regulator Mitf modulates lineage survival and melanoma cell viability." *Cell* 109.6 (2002): 707-718.

[27] Basakran, Nawwaf S. "CD44 as a potential diagnostic tumor marker." *Saudi medical journal* 36.3 (2015): 273.

[28] “SUMO1.” *National Center for Biotechnology Information*, 8 June 2019, https://www.ncbi.nlm.nih.gov/gene/7341. Accessed 8 June 2019.

[29] “CETN2 Gene.” *GeneCards*, n.d., https://www.genecards.org/cgi-bin/carddisp.pl?gene=CETN2. Accessed 8 June 2019.

[30] “SLC25A13 gene.” *National Institute of Health: Genetics Home Reference*, 28 May 2019, https://ghr.nlm.nih.gov/gene/SLC25A13. Accessed 8 June 2019.

[31] “DECR1 gene.”*National Institute of Health: Genetics Home Reference*, 28 May 2019, https://ghr.nlm.nih.gov/gene/DECR1. Accessed 8 June 2019.

[32] “CFD.” *National Center for Biotechnology Information*, 3 June 2019, https://www.ncbi.nlm.nih.gov/gene/1675. Accessed 8 June 2019.

[33] Metral, Sylvain, et al. "αII-Spectrin is critical for cell adhesion and cell cycle." *Journal of Biological Chemistry* 284.4 (2009): 2409-2418.

[34] “MGST1.” *National Center for Biotechnology Information*, 8 June 2019, https://www.ncbi.nlm.nih.gov/gene/4257. Accessed 8 June 2019.

[35] “DNTT Gene.” *GeneCards*, n.d., https://www.genecards.org/cgi-bin/carddisp.pl?gene=DNTT. Accessed 8 June 2019.

[36] Park, Peter J., et al. "A permutation test for determining significance of clusters with applications to spatial and gene expression data." *Computational statistics & data analysis* 53.12 (2009): 4290-4300.

[37] Monti, Stefano, et al. "Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data." *Machine learning* 52.1-2 (2003): 91-118.

[38] Taylor, Craig J., Eleanor M. Bolton, and J. Andrew Bradley. "Immunological considerations for embryonic and induced pluripotent stem cell banking." *Philosophical Transactions of the Royal Society B: Biological Sciences* 366.1575 (2011): 2312-2322.