Tumor Classification using Gene Expression using the *RNA-Seq* gene expression levels measured by *Illumina – HiSeq* Platform

Donna I. Baret

University of the Philippines – BGC
Professional Master's in Data Science (Analytics)
December 2019

Abstract

In this study we're going to classify human tumors and determine the sequence of genes that affect its manifestation using the RNA-Seq gene expression levels measured by *illumina HiSeq* platform. This will enable us to determine gene sequences or genes that acts as drivers for tumor growth. Determining factors affecting tumor manifestation and classifying them will help in the research of determining genetic mutations and the possibility of silencing the gene to prevent cancer, in which is an emerging method for cancer therapy: simultaneous multiplexed Gene Editing Technology (National Research Council of Science & Technology, 2019). We conducted Principal Component Analysis and reduced the data to 4 principal components. The genes *gene_3439, gene_19153, gene_7421, gene_1858, gene_19159* and *gene_220* provided the highest variability in the 1st principal component. On the other hand, the second principal component has genes *gene_3439, gene_1915, gene_7421, gene_1858, gene_19159, gene_220* with the highest loading. Genes *gene_3439, gene_19153, gene_7421, gene_1858, gene_19159, gene_220* have the highest loadings for rhe third principal. Lastly, the fourth principal component has genes *gene_2318, gene_5709, gene_3645, gene_18810, gene_17316, gene_6594* with the highest loading. Using these four principal components, we trained the data to create Naïve Bayes classifier which resulted to a model with a 99.58% level of accuracy. This means, that a highly dimensional data such as gene expression can be condensed into a smaller dimensional one using principal component analysis in order to classify different types of cancer for easier interpretation. Moreover, using the principal component projections for model building in classifying cancer also led to accurate results.

I. Introduction

Cancer is a collection of related disease caused by an uncontrolled growth of abnormal cells in the body. It is caused by old cells not dying and growing out and forming a mass of tissue more commonly known as tumors. Moreover, it is a genetic disease most commonly affected by how genes control cell growth and division and tend to affect three types of genes: Proto-Oncogenes, Tumor Suppressor Genes, and DNA repair genes. Cancer manifestation can be both genetic or caused by environmental factors (Cancer Treatment Centers of America, 2015).

A study on the comprehensive identification of mutational cancer driver genes across tumor types found out that out of 291 high-confidence cancer driver genes on 3205 tumors from 12 cancer types, 16 sustained mutations in one tumor type. (David Tamborero, 2013)

In this paper, we explore how genes or set of gene alterations affect the manifestation of cancer cells and explore how to classify them. We're going to use Principal Component Analysis (PCA) as a data reduction procedure on *our RNA-Seq (HiSeq) PANCAN* data set which is a data set of patients having different types of Tumor: BRCA, KIRC, COAD, LUAD and PRAD and use hierarchical clustering to determine which gene expressions affect cancer manifestations. To add to that, we're going to build a Naïve Bayes classifier in order to classify the tumors using a test data set.

Learning how different gene expressions affect the type of cancer manifestation will help us determine if a patient is predisposed to cancer. Moreover, learning how to classify them will help in the research of determining gene mutations and the possibility of silencing it to conduct preventive measures in order to avoid them.

II. Review of Related Literature

Despite many studies on identifying mutations on human cancers, it is still a challenge to distinguish cancer driver genes. (Collin J. Tokheim, 2016) defined a cancer driver gene as one that increases the mutation of the cell. Cancer tends to be affected by three types of genes: Proto-Oncogenes – most involved in cell growth and division, if genes are altered, may allow the cell to survive when they should not have been; Tumor Suppressor Genes – also involved in growth and division but an alteration causes the cells to divide uncontrollably; and DNA Repair Genes – when mutated, tends to create mutations on other types as well. (Cancer Treatment Centers of America, 2015)

Moreover, the most common types of cancer by cell type where they begin are -

i. Carcinoma which are formed by epithelial cells, which are the cells that cover the inside and outside surfaces of the body

ii. Sarcoma which form in bone and soft tissues, including muscle, fat, blood vessels, lymph vessels, and fibrous tissue (such as tendons and ligaments).

iii. Leukemia - begins in the blood-forming tissue of the bone marrow are called leukemias. These cancers do not form solid tumor

iv. Lymphoma - lymphoma is cancer that begins in lymphocytes (T cells or B cells). These are disease-fighting white blood cells that are part of the immune system. In lymphoma, abnormal lymphocytes build up in lymph nodes and lymph vessels, as well as in other organs of the body.

v. Multiple Myeloma - which is cancer that begins in plasma cells, another type of immune cell.

A. Genomics and Cancer

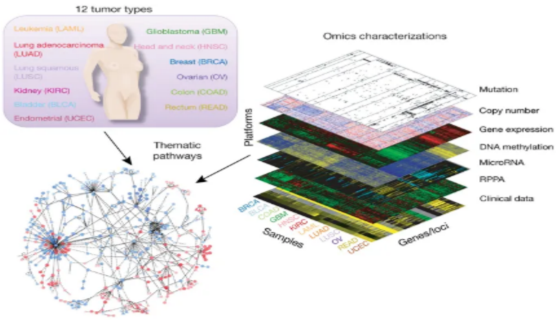
As humans live, DNA in human cells is exposed to mutagens that causes cell mutations. A study by (Biao Luo, 2008) shows that knowing more about the molecular mechanisms can improve the prevention, diagnosis and treatment of cancer. In their study they identified the genes tagged for growth and related the phenotypes in different cancer cells where they identified known and putative oncogenes. They are EGFR, KRAS, MYC, BCR-ABL, MYB, CRKL, and CDK4 which was identified as essential for cancer cell proliferation and also altered in human cancers and they mutated genes involved in the response of cancer cells .With this, they used parallel genetic screening strategy to easily identify genes that drive cancer and participate in biological processes

Another study by (Laura E. MacConaill, 2009) predicted the clinical outcome tumor by detecting critical cancer gene mutations and showed how genotyping and validation algorithm results to a robust tumor mutation profiling. Another study by (Fahriye Gemci, 2017) used medical data used a Naïve Bayes Algorithm on Gene Expression Cancer RNA-Seq Data Set to Classify the tumor type. On the other hand, (Ashton C. Berger, 2018) performed an analysis that identified molecular features characteristics of gynecological tumors. They found out that subtypes with high leukocyte infiltration, a marker for immune response and genes Gene-lncRNA interaction network of ESR1, DKC1, and lncRNAs TERC, NEAT1, and TUG1 are also found significant in tumor existence. This only shows how cancer manifestations can be classified by identifying different gene mutations.

B. TCGA Pan-Cancer Project

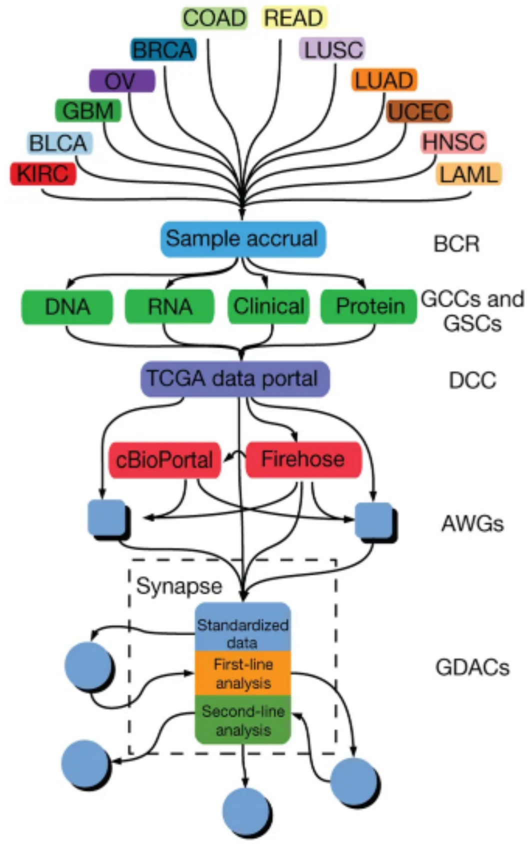
The TCGA Pan-Cancer Project: the hope is that investigations across tumor type such as the Pan-Cancer project will ultimately inform clinical decision-making.

A paper by (Chang, 2013) presented availability of the data set enabled new patterns of genomic drivers. The availability of the data set enabled the identification of new patterns of genomic drivers. It also allowed to explore computational approaches that enabled identification of frequently mutated gene by leveraging cross-tumor principles of replication timing and gene expression correlated with background mutation rates now enable the identification of frequently mutated genes which resulted to better classification rates. Moreover, there was an improved ability to distinguish 'driver' from 'passenger' aberrations by identifying multiple signals of positive selection. A tissue-associated pattern has been established in order to determine the rate and timing of the duplication events of the whole-genome. New classes of mutations, such as those in chromatin-remodeling genes has been identified as cancer drivers identified because the data set enabled analysts to collect less frequent events across tumor types, integrate event types such as mutations, copy number changes and epigenetic silencing, combine multiple algorithms to identify predicted drivers, and aggregate genes using gene networks and pathways.



The TCGA Pan-Cancer project assembled data from thousands of patients with primary tumors occurring in different sites of the body, covering 12 tumor types (top left) including glioblastoma multiformae (GBM), lymphoblastic acute myeloid leukemia (LAML), head and neck squamous carcinoma (HNSC), lung adenocarcinoma (LUAD), lung squamous carcinoma (LUSC), breast carcinoma (BRCA), kidney renal clear-cell carcinoma (KIRC), ovarian carcinoma (OV), bladder carcinoma (BLCA), colon adenocarcinoma (COAD), uterine cervical and endometrial carcinoma (UCEC) and rectal adenocarcinoma (READ). Six types of omics characterization were performed creating a 'data stack' (right) in which data elements across the platforms are linked by the fact that the same samples were used for each, thus maximizing the potential of integrative analysis. Use of the data enables the identification of general trends, including common pathways (bottom left), revealing master regulatory hubs activated (red) or deactivated (blue) across different tissue types.

Figure 1. Integrated data set for comparing and contrasting multiple tumor types. Reprinted from *Chang, K., Creighton, C., Davis, C. et al. The Cancer Genome Atlas Pan-Cancer analysis project. Nat Genet 45, 1113–1120 (2013) doi:10.1038/ng.2764*



Data were collected by the Biospecimen Collection Resource (BCR) from 12 different tumor types and characterized on 6 major platforms by the Genome Characterization Centers and Genomic Sequencing Centers (GCCs and GSCs). Data sets were deposited in the TCGA Data Coordination Center (DCC) from which they were then distributed to the Broad Institute's Firehose and the Memorial Sloan-Kettering Cancer Center's cBioPortal for various automated processing pipelines. Analysis Working Groups (AWGs) conducted focused analyses on individual tumor types. Results from the DCC, Firehose and AWGs were collected and stored in Sage Bionetworks' Synapse database system to create a data freeze. Genome data analysis centers (GDACs) accessed and deposited both data and results through Synapse to coordinate distributed analyses.

Figure 2. Data coordination for the Pan-Cancer TCGA project. Reprinted from *Chang, K., Creighton, C., Davis, C. et al. The Cancer Genome Atlas Pan-Cancer analysis project. Nat Genet 45, 1113–1120 (2013) doi:10.1038/ng.2764*

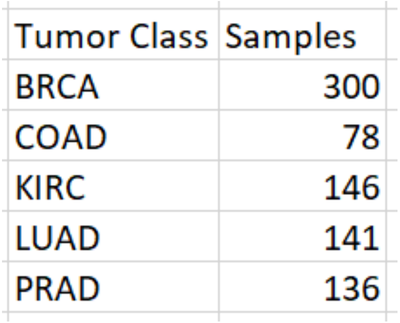
III. Scope and Limitations

In this study, we’re only going to investigate 5 tumor classes namely prostate adenocarcinoma (PRAD), lung adenocarcinoma (LUAD), breast carcinoma (BRCA), kidney renal clear-cell carcinoma (KIRC), and colon adenocarcinoma (COAD) in the Gene Expression Cancer RNA-Seq data set and will look into 20,532 gene sequences and their expression levels and 801 patients’ samples

IV. Methodology

A. Data Understanding

Table 1. Distribution of Samples and Tumor Type



Tumor Class	Samples
BRCA	300
COAD	78
KIRC	146
LUAD	141
PRAD	136

a. Data Preparation

(Ying Lu, n.d.) study indicated that feature selection is important preprocessing technique in data mining because of this led to improved classification accuracy of most classifiers this is due to the data set’s high dimensionality, very small data set size and because most genes are not related to cancer classification.

On this paper, we trimmed the data set down to the genes which has a total expression of more than 50 and a standard deviation of more than then since genes that are not very variable do not contribute much to the distances between patients. (Akalin, 2019)

B. Data Mining Analytics/Business Process Discovery

a. Principal Component Analysis

Because of many gene expression, we have a high dimensional data, we will conduct a Principal Component Analysis on our data in order to lessen computational work-load as well as in order to introduce a better fit or generalization. Moreover, we can generate more meaningful distance measure that aid in exploratory data analysis

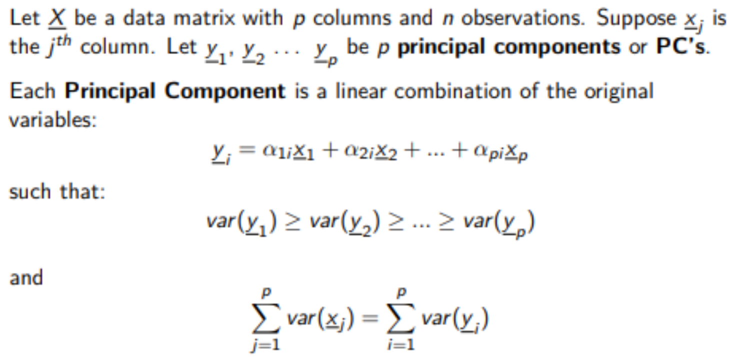
PCA does dimension reduction by linearly combining variables

1. If there are p variables $x_1, .., x_p$, PCA lets you find k p linear combinations of x's that contain as much information as the original p variables

2. Information is measured in terms of variance. PCA derives linear combinations that maximizes variance

3. Tendency to combine correlated variables together

The procedure can be summarized as:



Let $\underline{X}$ be a data matrix with p columns and n observations. Suppose $\underline{x}_j$ is the j^{th} column. Let $\underline{y}_1$, $\underline{y}_2$ ... $\underline{y}_p$ be p **principal components** or **PC's**.

Each **Principal Component** is a linear combination of the original variables:

$$\underline{y}_i = \alpha_{1i}\underline{x}_1 + \alpha_{2i}\underline{x}_2 + ... + \alpha_{pi}\underline{x}_p$$

such that:

$$var(\underline{y}_1) \geq var(\underline{y}_2) \geq ... \geq var(\underline{y}_p)$$

and

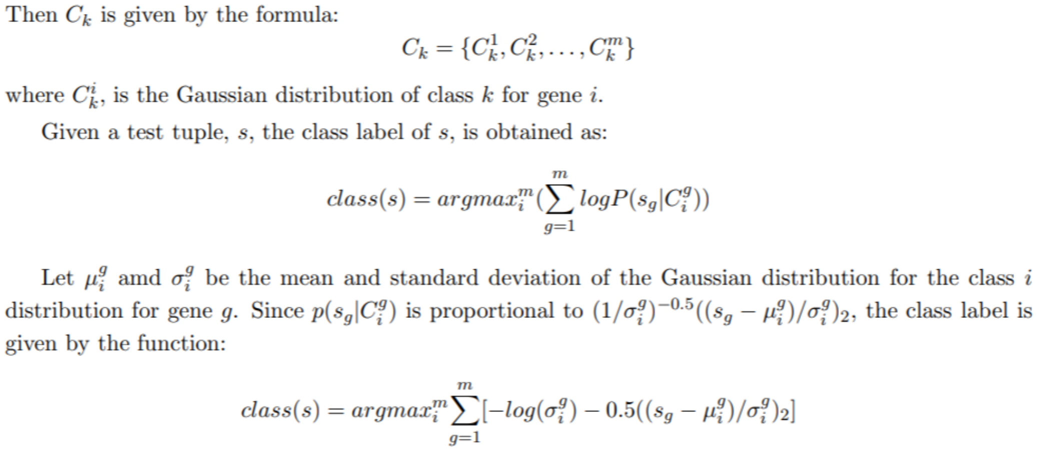
$$\sum_{j=1}^{p} var(\underline{x}_j) = \sum_{i=1}^{p} var(\underline{y}_i)$$

b. Hierarchical Clustering

Hierarchical clustering is one of the most common clustering algorithms wherein the relationship between different data points can be seen. This is done by joining small clusters to each other based on intercluster distance in which eventually becomes a tree structure. In this paper, we will use the Ward's minimum variance method which aims to find compact, spherical clusters by selecting clusters to merge based on the change in the cluster variances. The clusters are merged if the increase in the combined variance over the sum of the cluster specific variances is minimum compared to alternative merging operation.(Akalin, 2019)

c. Probabilistic Induction: Naive Bayes (NB) Method

A study by (Andrew D. Keller, 2000) used Naïve Bayes algorithm for gene classification. They modeled each class as a set of Gaussian distribution one for each gene in the training sample set. They then set K as number of classes and each class Ck is modeled using a set of Gaussian distributions.



Then C_k is given by the formula:

$$C_k = \{C_k^1, C_k^2, \ldots, C_k^m\}$$

where C_k^i, is the Gaussian distribution of class k for gene i.

Given a test tuple, s, the class label of s, is obtained as:

$$class(s) = argmax_i^m(\sum_{g=1}^{m} logP(s_g|C_i^g))$$

Let μ_i^g amd σ_i^g be the mean and standard deviation of the Gaussian distribution for the class i distribution for gene g. Since $p(s_g|C_i^g)$ is proportional to $(1/\sigma_i^g)^{-0.5}((s_g - \mu_i^g)/\sigma_i^g)_2$, the class label is given by the function:

$$class(s) = argmax_i^m \sum_{g=1}^{m}[-log(\sigma_i^g) - 0.5((s_g - \mu_i^g)/\sigma_i^g)_2]$$

After preprocessing the data and conducting a data reduction procedure using PCA, we are a going to split our data and train a part of the data set and create a classifier model using a Naïve Bayes.

IV. Results

A. Principal Component Analysis

We will choose the first four principal components

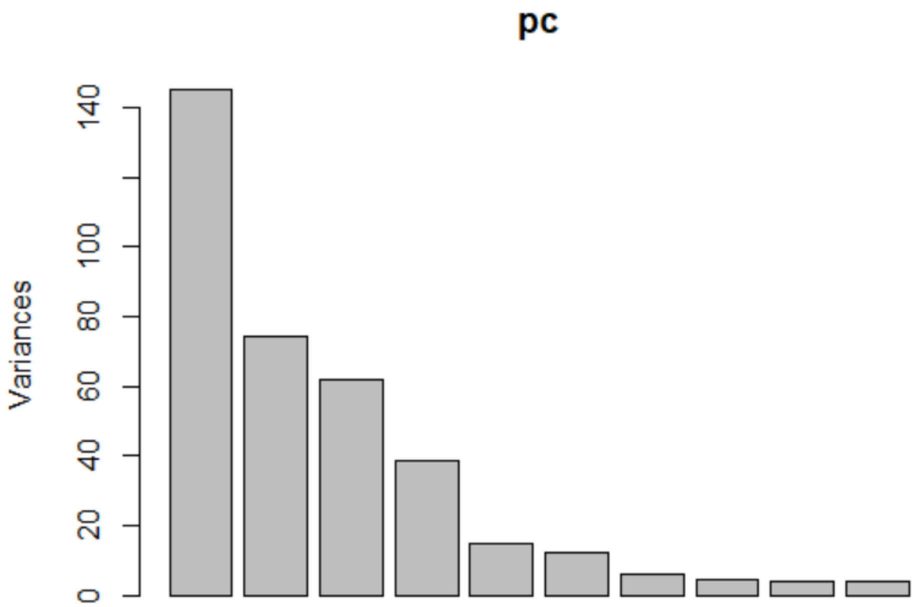
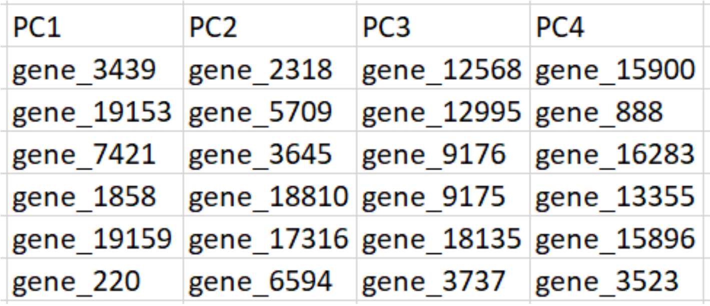


Figure 3. Scree Plot of the Principal Component Based on their Scree Plot

As we can see on figure 3, there is a sharp decrease of variance after the fourth principal component, therefore we will choose the first four principal components for our analysis.

a. Projection of Loadings

Table 2. Top six that have the highest loadings (in absolute value)



PC1	PC2	PC3	PC4
gene_3439	gene_2318	gene_12568	gene_15900
gene_19153	gene_5709	gene_12995	gene_888
gene_7421	gene_3645	gene_9176	gene_16283
gene_1858	gene_18810	gene_9175	gene_13355
gene_19159	gene_17316	gene_18135	gene_15896
gene_220	gene_6594	gene_3737	gene_3523

For the first principal component, the top genes with the highest loadings are genes: *gene_3439, gene_19153, gene_7421, gene_1858, gene_19159, gene_220*. On the other hand, the second principal component has genes *gene_3439, gene_1915, gene_7421,gene_1858,gene_19159 , gene_220* with the highest loading. Genes *gene_3439, gene_19153, gene_7421, gene_1858, gene_19159, gene_220* have the highest loadings for rhe third principal. Lastly, the fourth principal component has genes *gene_2318 gene_5709 gene_3645 gene_18810 gene_17316 gene_6594* with the highest loading.

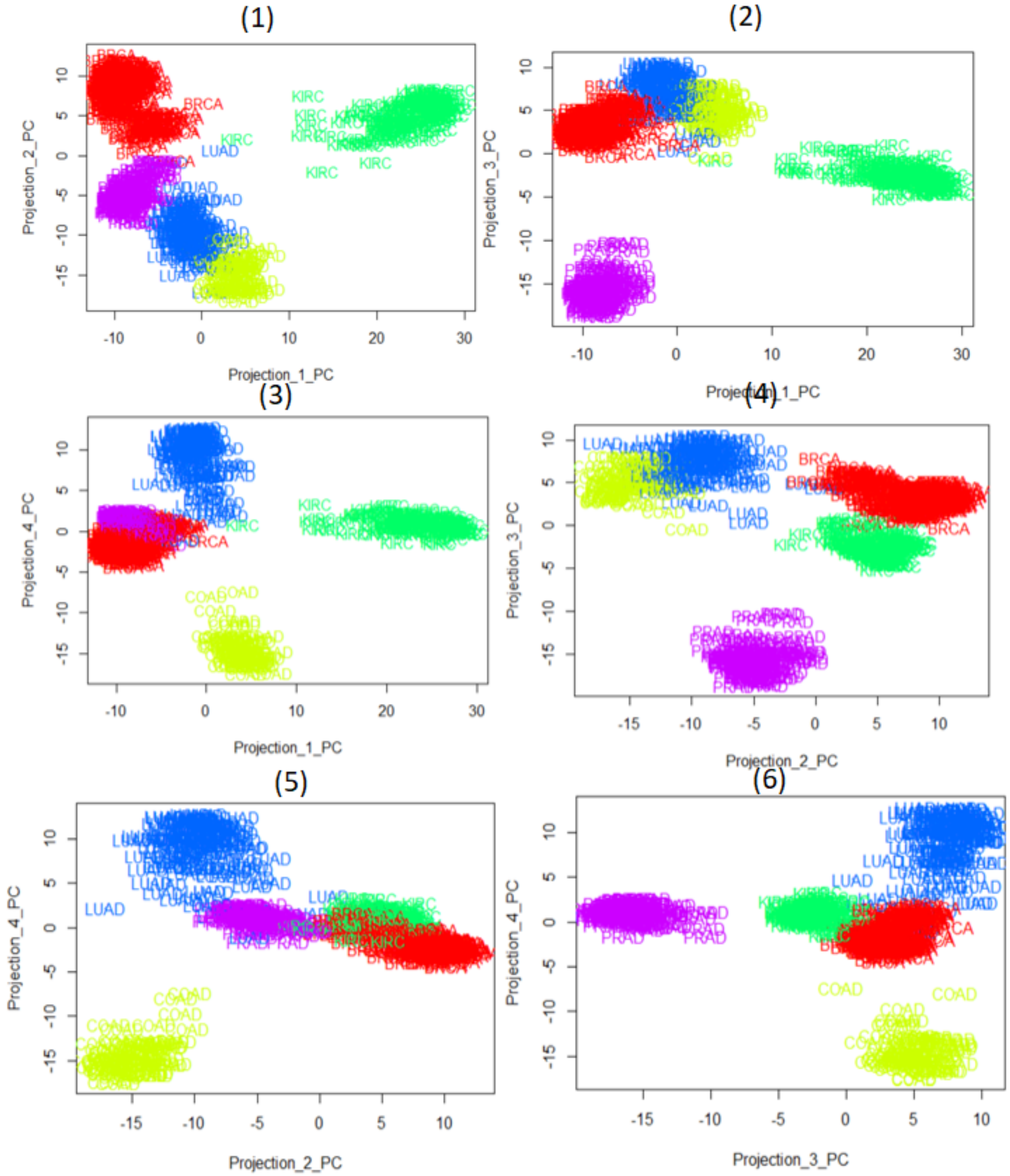


Figure 4. (1) Projection of PC1 and PC2 by Tumor Type (2) Projection of PC1 and PC3 by Tumor Type (3) Projection of PC1 and PC4 by Tumor Type (4) Projection of PC2 and PC3 by Tumor Type (5) Projection of PC2 and PC4 by Tumor Type (6) Projection of PC3 and PC4 by Tumor Type

The projection of PC1 and PC2 separates the tumor type best. Moreover, KIRC is separated really well on the combination of PC1 and PC2, PC1 and PC3, and PC1 and PC4. On the other hand, COAD is best separated on the PC1 and PC4, PC2 and PC4, PC3 and PC4 projections. Moreover, PRAD is separated best using PC1 and PC3, PC2 and PC3, and PC3 and PC4.

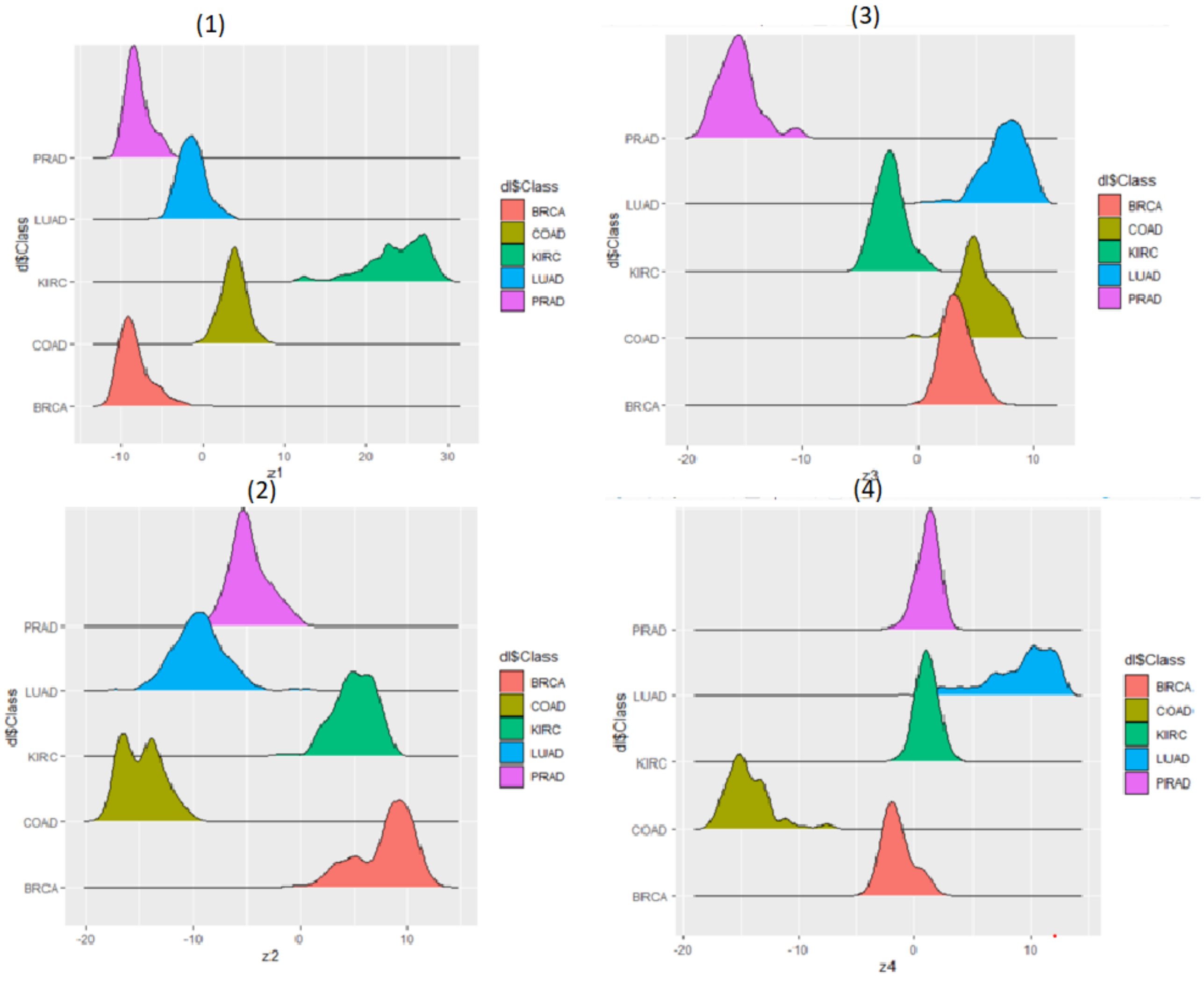


Figure 5. (1) Projection of PC1 (2) Projection of PC2 (3) Projection of PC3 (4) Projection of PC4.

Figure 5 (1) shows that the 1st principal component does a good job picking out KIRC among others, As shown on Figure 5(3), PC3 picks out COAD best.

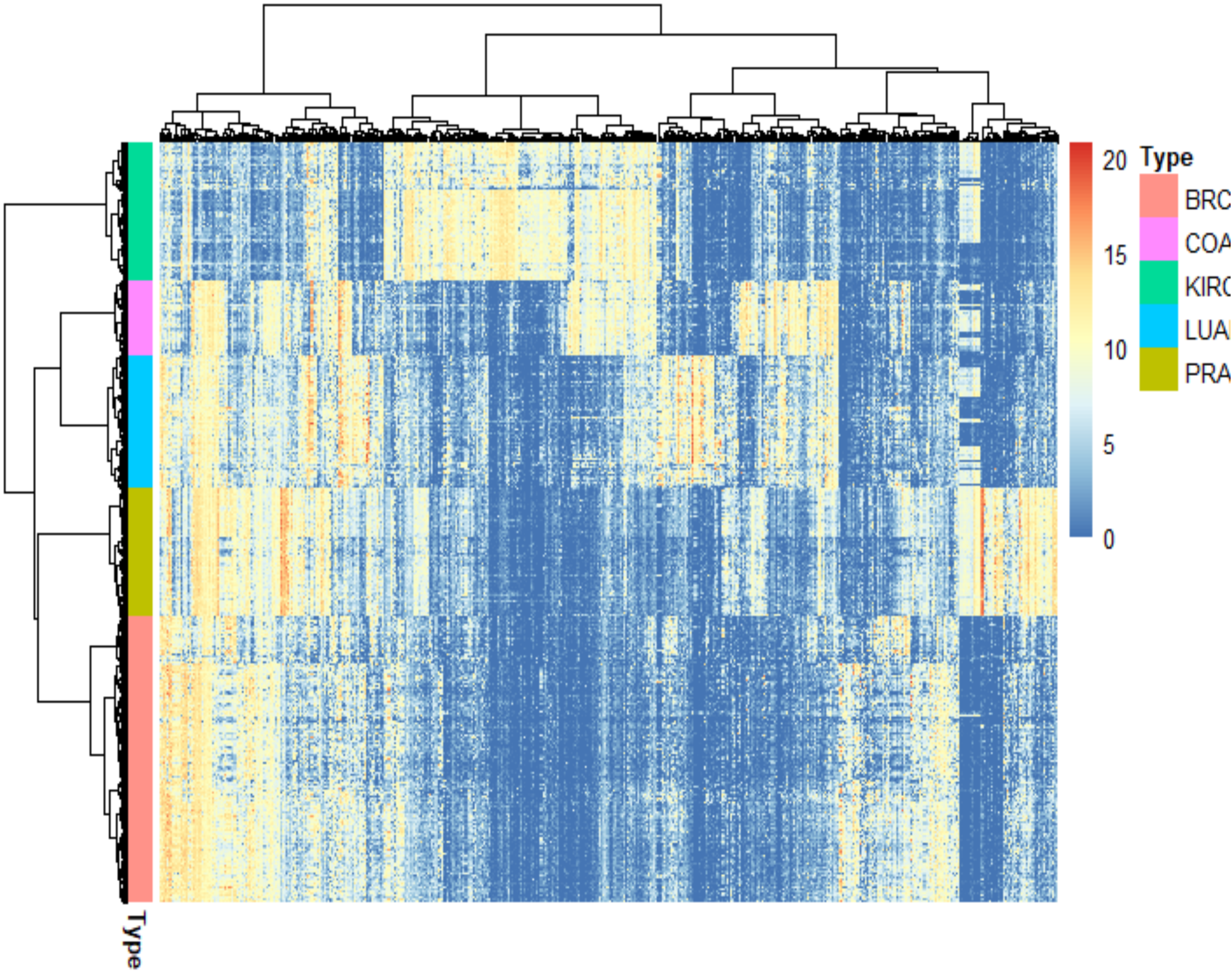
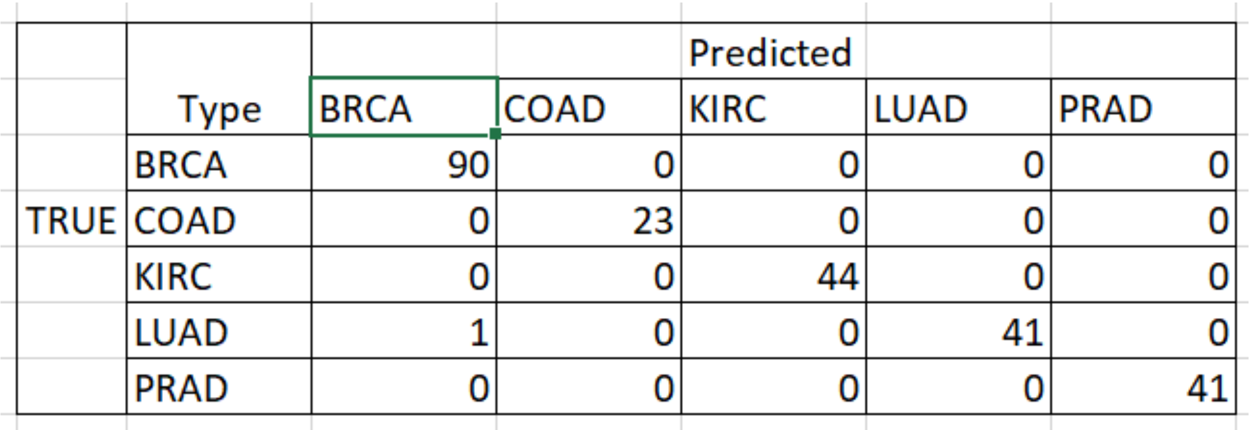


Figure 6. Heatmap of Each Cluster and Gene Expression Values

We can see on figure 6 that each tumor type has a distinct set of expression values. We can also see that the clusters distinguish between tumor type quite well.

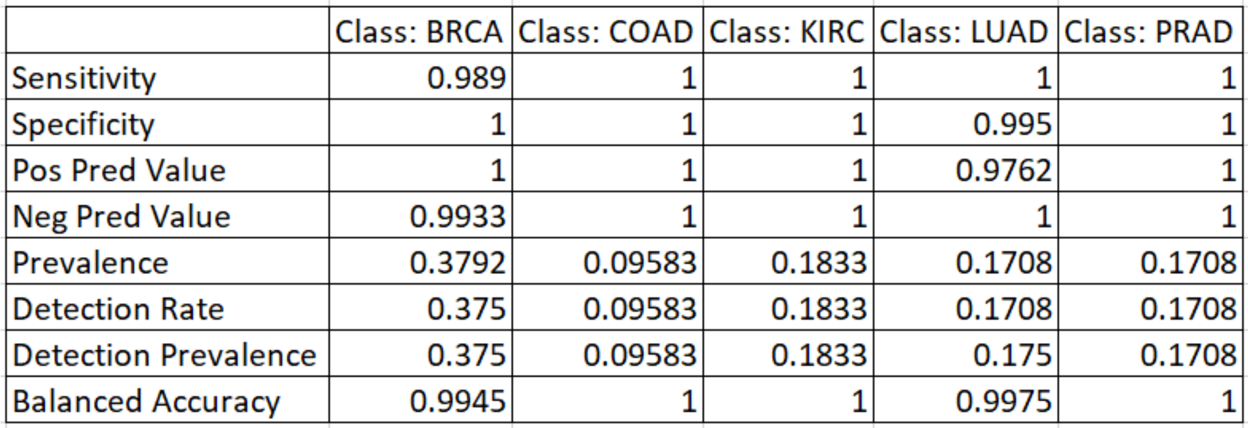
Table 3. Performance of the Classification Model on the Test Data



				Predicted		
	Type	BRCA	COAD	KIRC	LUAD	PRAD
	BRCA	90	0	0	0	0
TRUE	COAD	0	23	0	0	0
	KIRC	0	0	44	0	0
	LUAD	1	0	0	41	0
	PRAD	0	0	0	0	41

Almost all samples are correctly classified except one sample which predicted the tumor type to be BRCA instead of LUAD.

Table 4. Statistics by Class of the Classification Model on Test Data



	Class: BRCA	Class: COAD	Class: KIRC	Class: LUAD	Class: PRAD
Sensitivity	0.989	1	1	1	1
Specificity	1	1	1	0.995	1
Pos Pred Value	1	1	1	0.9762	1
Neg Pred Value	0.9933	1	1	1	1
Prevalence	0.3792	0.09583	0.1833	0.1708	0.1708
Detection Rate	0.375	0.09583	0.1833	0.1708	0.1708
Detection Prevalence	0.375	0.09583	0.1833	0.175	0.1708
Balanced Accuracy	0.9945	1	1	0.9975	1

.

Our model appears to be both accurate and sensitive except for classifying a true BRCA tumor.

V. Discussion/Conclusion

In this study we're classified human tumor and determined the sequence of genes that affect cancer manifestation using the RNA-Seq gene expression levels measured by *illumina HiSeq platform.* We also a explored some techniques such as PCA, Hierarchical Clustering, and Naïve Bayes Classifier to describe the gene expression behavior on different kinds of cancer. In this paper, we have shown that PCA can reduce the large set of gene expressions and condensed it into four principal components and still be capable of classifying the tumor type. For the first principal component, the top genes with the highest loadings are genes: *gene_3439, gene_19153, gene_7421, gene_1858, gene_19159, gene_220*. On the other hand, the second principal component has *genes gene_3439, gene_1915, gene_7421,gene_1858,gene_19159 , gene_220* with the highest loading. Genes *gene_3439, gene_19153, gene_7421, gene_1858, gene_19159, gene_220* have the highest loadings for rhe third principal. Lastly, the fourth principal component has genes gene*_2318 gene_5709 gene_3645 gene_18810 gene_17316 gene_6594* with the highest loading. This means that the genes that have a high loading on the principal components affect the condensed value most. From the principal components, it will be interesting to know how the genes with highest loadings per principal component are related to each other. This will be a good indicator of driver genes for tumor growth.

Moreover, using hierarchical clustering, we can see from the heatmaps that the gene expression levels are distinctively different from different types of tumor. This is a good indication which means that gene expression is different from different types of tumor growth and its measure is a good classifier for type of cancer. It is recommended to explore these genes, their level of expression and their gene sequences in order to determine which sets of genes are associated as well as which set of genes affects other types of manifestation on a patient's cell or body.

Going back to the principal component loadings, we have used this condensed variable to create a Naïve Bayes Classifier and found out that the model can classify the tumor type with a 99.58% accuracy. This means that condensing the variables has not foregone the accuracy of then classifier using Naïve Bayes,

Since we saw that condensing the variables into smaller ones using PCA gives an accurate result in classifying tumor types, this would help researchers simplify the analysis of complex and highly dimensional data such as gene expression. The use of Naïve Bayes classifier would also help researchers determine if a patient is predisposed to certain type of cancer.

Bibliography

Akalin, A. (2019). *Computational Genomics with R.* https://compgenomr.github.io/book/.

Andrew D. Keller, M. S. (2000). *Bayesian Classification of DNA Array Expression Data.* Seattle, Washington: Department of Computer Science and Engineering, University of Washington, Seattle, WA.

Ashton C. Berger, 1. A. (2018). A Comprehensive Pan-Cancer Molecular Study of Gynecologic and Breast Cancers. *Cell Press*, 690-705.

Biao Luo, H. W. (2008). Highly parallel identification of essential genes in cancer cells. *Proceeding of the National Academy of Sciences of the United States of America* (pp. 20380-20385). Proceedings of the National Academy of Sciences.

Cancer Treatment Centers of America. (2015, February 9). *What is Cancer.* Retrieved from National Cancer Institute: https://www.cancer.gov/about-cancer/understanding/what-is-cancer

Chang, K. C. (2013). The Cancer Genome Atlas Pan-Cancer analysis project. *Nat Genet 45*, 11113-1120.

Collin J. Tokheim, N. P. (2016). Evaluating the evaluation of cancer driver genes. *Proceedings of the National Academy of Sciences of the United States of America*, (pp. 14330-14335).

David Tamborero, A. G.-P.-L.-P.-B. (2013). *Comprehensive identification of mutational cancer driver genes across 12 tumor types.* Barcelona Spain: SCIENTIFIC REPORTS.

Fahriye Gemci, T. I. (2017). Tumor Type Detection. *International Conference on Engineering Technologies.* Konya, Turkey.

Laura E. MacConaill, C. D. (2009). Profiling Critical Cancer Gene Mutations in Clinical Tumor Samples. *PLOS|ONE*, 9.

National Research Council of Science & Technology. (2019, August 20). *A new path to cancer therapy: developing simultaneous multiplexed gene editing technology*. Retrieved from ScienceDaily: https://www.sciencedaily.com/releases/2019/08/190820081853.htm

Ying Lu, J. H. (n.d.). *Cancer Classification Using Gene Expression Data*. Retrieved from Knowledge Discovery and Data Mining, Database Systems: http://hanj.cs.illinois.edu/pdf/is03_cancer.pdf