Zachary Haiman & Isabel Shiratsubaki
BENG 213
6/13/2018

Gene Expression Analysis of The Cancer Genome Atlas

Abstract

The vast amount of cancer data and publications associated with the project make it difficult to overstate the value of the TCGA dataset[2], for with such large amounts of data available, several different types of analyses can be performed on subsets of the TCGA dataset. The purpose of this study was focused on determining whether the gene expression profile of a cancerous tissue sample can be used to predict the tumor's subtype. The chosen TCGA dataset contained 7803 unique cases of cancer, providing the gene expression data of 23369 genes associated with 7940 tumor samples. Dimensionality reduction, statistical testing of gene expression significance, and a predictive model using a random forest classifier were used to analyze the gene expression data. The results show that though there are several considerations that must be considered, it is possible to utilize gene expression data to accurately predict a tumor's subtype.

1. Introduction

The Cancer Genome Atlas (TCGA) project is a comprehensive and coordinated effort by the National Cancer Institute (NCI) and the National Human Genome Research Institute (NHGRI) to systematically explore a variety of cancers for the overarching purpose of improving medical diagnostics and treatments for cancer [1, 2]. The TCGA project began in 2006 with the collection of tissue samples for 33 different tumor types and is ending in Spring 2018 after its final publication of the integrative cross-platform analysis for the 33rd tumor type[1]. During that time frame, over 2.5 petabytes of data were generated to aid in the elucidation of the molecular basis of different cancers, formation of new classification systems, and identification of therapeutic targets[2]. Consequently, the vast amount of cancer data and publications associated with the project make it difficult to quantify the usage and impact of the overall dataset; however, it is difficult to overstate the value of the TCGA dataset[2]. With such large amounts of data available, several different types of analyses can be performed on subsets of the TCGA dataset.

Previous analyses conducted on the TCGA data for each individual cancer type include unsupervised clustering of data obtained through a variety of throughput technologies to identify recurrent somatic gene mutations, distinct molecular subtypes, DNA methylation profiles, and mRNA/miRNA expression profiles[3]. Some of the novel discoveries resulting from these analyses include the discovery of the remarkable similarity of colon and rectal non-hyper mutated carcinomas, a more clinically useful classification system for gastric adenocarcinoma subtypes, and several potential therapeutic targets to explore for personalized treatment options[2].

However, though each individual type has been thoroughly characterized and analyzed, there are several opportunities to utilize the knowledge of the genetic similarities and differences for the various subtypes of cancer. Therefore, the purpose of this study will be focused on determining whether the gene expression profile of a cancerous tissue sample can be used to predict the tumor's subtype.

2. Methods

2.1 Initial Data Visualization and Dimensionality Reduction

To narrow the scope of this study, a subset of normalized FPKM data for 18 different tumor types was utilized[4]. The dataset contained 7803 unique cases of cancer, providing the gene expression data of 23369 genes associated with 7940 tumor samples (Fig. 1).

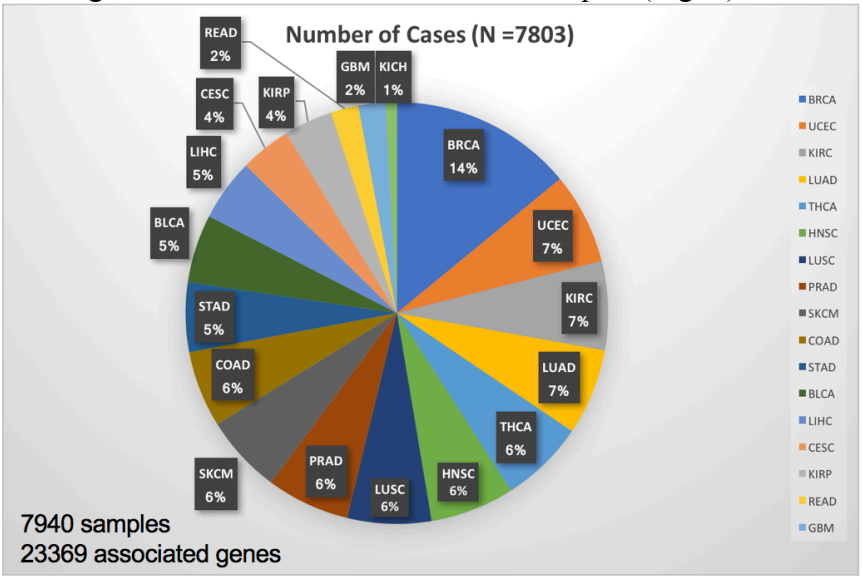

Figure 1: Overview of the dataset. Each entry contains the TCGA project name where the data for tissue sample was collected, processed, and analyzed. The resulting datasets were then made publically available as normalized FPKM expression data[4].

To better visualize potential features the data, PCA and t-SNE dimensionality reduction methods were utilized. Two different types of clustering methods were also applied to the t-SNE results by using the k-means and hierarchical algorithms. All dimensionality methods were run in R.

2.2 Statistical Testing of Gene Expression

To better understand the first and second principal components of the PCA analysis, the genes that contributed the most variance for each component was examined on the human protein atlas[4]. It was decided that the statistical tests to compare the gene expression average of the first principal component's top gene between the 18 subpopulations of each type of cancer and the global population of individuals. The mt.teststat function from the R-package *multitest* was utilized to run a two sample Welch t-test. The metrics used to control type I error were FWER and FDR. The significance level applied was 0.05.

2.3 Random Forest Classifier

Once dimensionality reduction had been used to determine principal components and the genes contributing the most variance to these components, it was hypothesized that these genes would be deemed important features by a classifier designed to predict a tumor sample's subtype based on its gene expression. To both test this hypothesis and determine whether the gene expression profile can be used to predict tumor type, a supervised learning algorithm that satisfied the following criteria was needed: 1) Must be able to handle numerical and categorical data; 2) Must be possible to statistically validate the model; 3) With the chosen dataset, an

algorithm that can easily implement bootstrapping is desirable; 4) No one class significantly dominates the other, therefore variability is a higher concern than bias. 5) Ideal if genes of interest could be used as predictors. Therefore, a random forest classification model was created for they are an easily interpretable white-box model with a logarithmic cost for training the decision trees. Additionally, the decision trees could be generated using random features, with genes of interest showing up more often as predictors.

To create the random forest classifier (RFC), an exhaustive parameter search was conducted using a StratifiedKFold cross-validation algorithm with 10 data partitions and determine the optimal parameters of the RFC. The feature selection for the RFC was determined using a linear support vector classification with an L2 normalization penalty, and the
Once the optimal parameters were obtained, the RCF was trained and tested using the full gene expression data, and an ROC curve and a confusion matrix were created to evaluate the model's performance.

3. Results and Discussion

3.1 Dimensionality Reduction

3.1.1 Principal Component Analysis

The results of PCA in a gene expression matrix of 7940 samples versus 23369 genes are given in Figure 2, with the individuals were labeled according to the type of tumor. Concentration ellipses were also added to the plot to better visualize the distribution of each type of tumor in the plot. The results show that the tumor types can be distinguished by gene expression data since the ellipses are not exactly overlapping over each other. The closer a specific tumor type ellipse is to another, the more they are similar. We noticed that COAD (Colon adenocarcinoma) ellipse is close to READ (Rectum adenocarcinoma) ellipse. These tumors are both types of intestinal cancer. The short distance between their ellipses might indicate high similarities between them regarding gene expression.

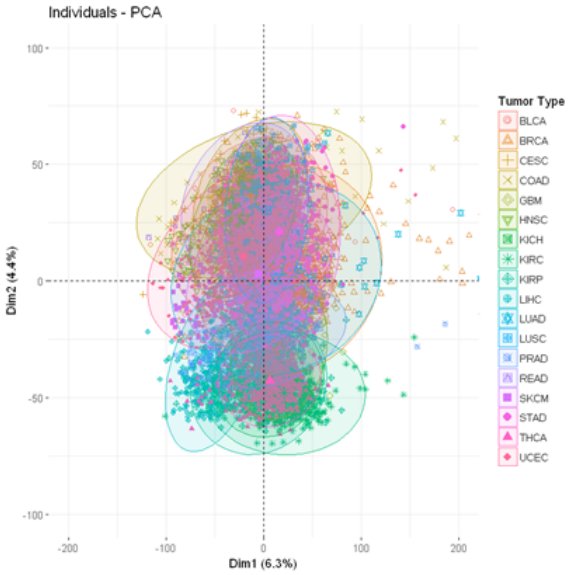

Figure 2: PCA results in gene expression dataset. Matrix: 7940 samples versus 23369 genes. The individuals were colored by tumor type. Concentration ellipses were added to better visualize distribution for each type of cancer.

We generated bar plots with the 10 genes that most contributed the most to variance in dimension 1 and 2 in PCA. The results show that CREB1 was the gene that contributed the most to the variance in the dataset in dimension 1 (Fig. 3). This gene is frequently associated with diseases such as Histiocytoma, a benign skin tumor, Angiomatoid Fibrous, a rare soft tissue tumor, and Melanoma of soft tissue.

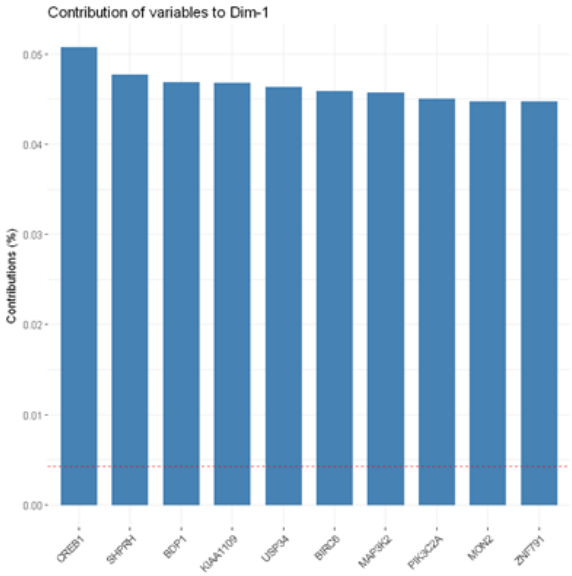

Figure 3: Top 10 genes of dimension 1 that contribute the most to the variance of the dataset.

It is known that that CREB1 is a cancer-related gene[5]. It is intracellular located, being responsible to encode a transcription factor which is a member of the leucine zipper family of DNA binding proteins. Its expression in normal tissue is associated with ubiquitous nuclear expression. The expression of CREB1 is usually lower in liver cancer[5]; therefore, the expression of this gene can be potentially used in prognostic of liver cancers (See Fig. A2 in the Appendix).

In dimension 2, it was observed that NCAPH was the top gene contributing to the variance in the dataset (Fig. 4). This gene is usually associated with Uterine Corpus Endometrial Carcinoma and Uterine Corpus Cancer[6].

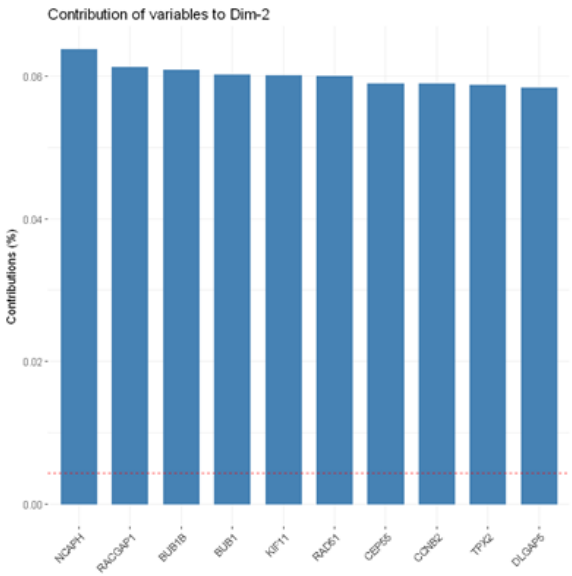

Figure 4: Top 10 genes of dimension 2 that contribute the most to the variance of the dataset.

NCAPH is found be an intracellular protein, and usually is expressed in normal tissues as cytoplasmic or nuclear expression in lymphoid reaction of cells in bone marrow and epithelial basal cells. The expression of NCAPH is low in many types of cancer[6]. Therefore, the expression of NCAPH can be a potential indicator in prognostics of many types of cancer (See Fig. A3 in the Appendix).

3.1.2 t-SNE Analysis

The results for t-SNE analysis are found in Fig. 5, with the individuals were colored by tumor type. The agglomeration of points in the top presents predominantly a bluish color. The occurrence of singular pink labeled points might due to the existence of repeated individuals in the dataset.

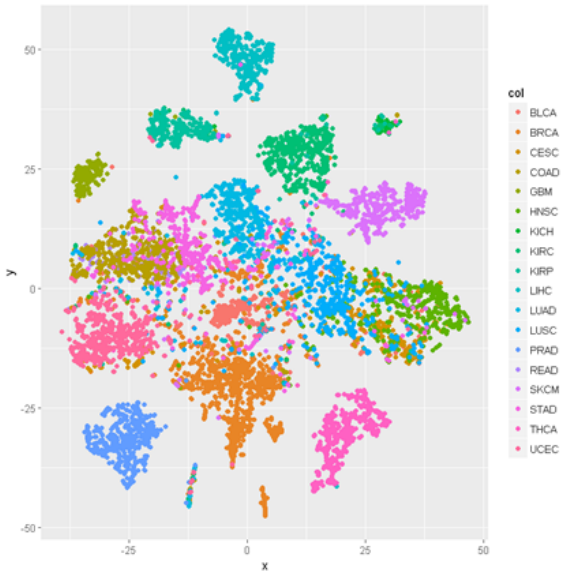

Figure 5: t-SNE results in gene expression dataset. Input matrix: 7940 samples versus 23369 genes. The individuals were colored by tumor type.

3.1.3 Clustering Analysis

The results of two different clustering methods in t-SNE results are given in Fig. 6. The results of each clustering method present similarities to each other. In both algorithms, the top agglomeration of points was identified as a cluster. There are some specific regions where the clustering methods do not agree. For instance, in the right plot (hierarchical method), the algorithm identified the green region cluster. The same region in the left plot was identified as two clusters by k-means method. The differences are due to the occurrence of multiple types of cancers in these regions, as it is shown in Fig. 5.

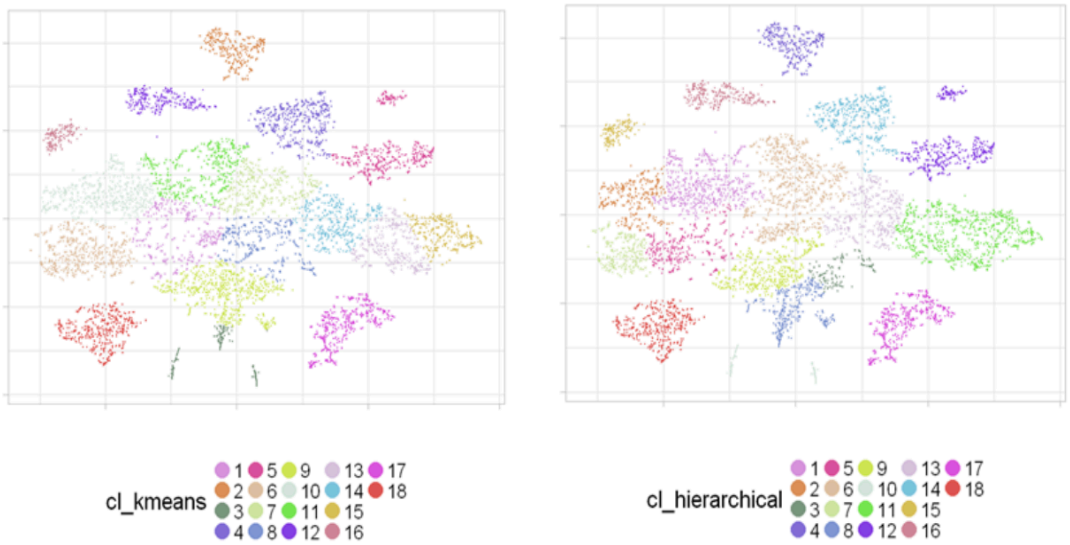

Figure 6: Clustering analysis using k-means (left) and hierarchical (right) methods applied to the t-SNE results. 18 clusters were set in each method.

3.2 Statistical Testing of Gene Expression

To examine the significance of CREB1 in the data, 18 statistical tests were performed to compare the gene expression average of the gene CREB 1 among the 18 subpopulations of tumor type and the global population. CREB1 was the gene found in PCA to be the one that contributes the most to the variance in the dataset in dimension 1.

Table 1: Number of rejections found in 18 statistical tests using different metrics to control type I error.

Metric to control type I error	Number of rejections	Number of acceptances
FWER	16	2
FDR	16	2

The number of rejections in Table 1 shows that the variance among the subpopulations regarding CREB1 expression is considerable. The cases where we got 2 acceptances were for the subpopulations of GBM (Glioblastoma multiforme) and THCA (Thyroid carcinoma) tumors. It shows that the average of CREB1 expression in those two types of tumors is like the CREB1 expression average in the global tumor type population. We also noticed that GBM (Glioblastoma multiforme) and THCA (Thyroid carcinoma) tumors present similar maximum, average and minimum values for CREB1 expression. We can also notice that their average expression of CREB1 is close to the global average expression of this gene[6] (See Fig. A4 in the Appendix).

3.3 Results from the Random Forest Classifier

The results of the parameter search with a StratifiedKFold cross-validation algorithm with 10 data partitions are shown in Table 2 and 3 in the Appendix. The accuracy score represents the mean accuracy of the cross-validation. From these results, the optimal parameters of the RGC were determined to be a forest of 40 trees, no limitation on the max tree depth, a gini impurity criterion, and a maximum feature limitation given as the square root of the total number of features.

Though the highest cross-validation accuracy of the model was obtained with no restrictions on the maximum depth of the trees, it is important to note that the trees generally had a depth of 10-12 branches, with only a few trees having more than 12. It is also important to note that quality of the accuracy score consistently increased with the number of trees in the forest as expected, for additional trees allow for a better consensus of important features and results.

Once the optimal parameters were determined, the model was trained and tested. A confusion matrix of the model's prediction versus the sample's actual tumor type was generated (Fig. 7), and the model's performance for each tumor type was evaluated using an ROC curve (Fig. 8).

Predicted Tumor Type / Actual Tumor Type	BLCA	BRCA	CESC	COAD	GBM	HNSC	KICH	KIRC	KIRP	LIHC	LUAD	LUSC	PRAD	READ	SKCM	STAD	THCA	UCEC
BLCA	103	0	0	0	0	9	0	0	0	0	0	0	0	0	0	0	0	0
BRCA	0	320	1	0	0	1	0	0	0	0	0	0	0	0	3	0	0	0
CESC	3	1	80	2	0	1	0	0	0	0	0	0	0	0	0	0	0	7
COAD	0	0	0	144	0	0	0	0	0	0	0	0	0	2	0	1	0	0
GBM	0	0	0	0	45	0	0	0	0	0	0	0	0	0	1	0	0	0
HNSC	1	0	5	0	0	131	0	0	0	0	0	5	0	0	0	0	0	0
KICH	0	0	0	0	0	0	14	1	0	0	0	0	0	0	1	0	0	0
KIRC	1	0	0	0	0	0	4	158	5	0	0	0	0	0	0	0	0	0
KIRP	2	0	0	0	0	0	2	6	76	0	0	0	0	0	0	0	0	0
LIHC	1	0	0	0	0	0	0	0	0	110	0	0	0	0	2	0	0	0
LUAD	2	0	0	0	0	0	0	0	0	0	146	1	0	0	0	2	1	0
LUSC	2	3	0	0	0	3	0	0	0	0	10	129	0	0	0	0	0	0
PRAD	0	0	0	0	0	0	0	0	0	0	0	0	150	0	0	0	0	0
READ	0	0	0	61	0	0	0	0	0	0	0	0	0	1	0	0	0	0
SKCM	0	0	0	0	0	0	0	0	0	0	2	0	0	0	165	0	0	0
STAD	1	0	0	0	0	2	0	0	0	0	0	0	0	0	1	123	0	0
THCA	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	150	0
UCEC	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	175

Figure 7: Confusion matrix of the model's predictions vs. the actual tumor type.

Examination of the confusion matrix revealed that the model's ability to distinguish between colon (COAD) and rectal (READ) adenocarcinomas. This behavior can be explained by the genetic similarities between colon and rectal adenocarcinomas in cases of non-hyper mutated cancers[2].

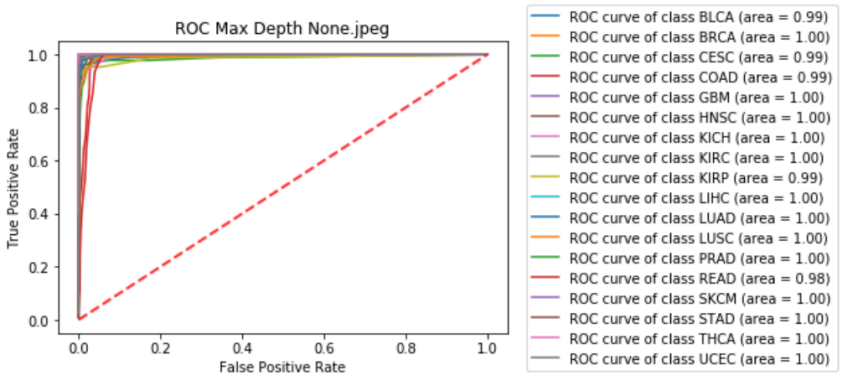

Figure 8: ROC curves of the model's performance for each class.

Gene	Relative Importance
PAX8	0.009582
NKX2-1	0.009213
XBP1	0.008257
CREB3L4	0.007682
SFTPA2	0.006937
EFHD1	0.006916
NAPSA	0.006242
MSX1	0.005924
SFTPA1	0.005545
AZGP1	0.005104
ARMCX2	0.005020
SFTA3	0.004850
SOX17	0.004831

Figure 9: Genes of the highest relative importance in the Random Forest Classifier. The relative importance is the percentage of importance as determined by the ensemble of trees in the Random Forest Classifier.

The ROC curves seen in Fig. 8 shows that each class has near perfect area under the curve (AUC) score, indicating that the model is likely to be overfitting the data. To assess reasons why this may be occurring, the most important genes utilized in the RFC were examined (Fig. 9). PAX8 was determined to be critical in the role of tissue/organ formation during embryonic development in addition to regulating several genes involved in thyroid hormone production[7]. NKX2-1 encodes for a homeobox protein that functions as a transcription factor and is particularly involved in the formation and function of the brain, lungs, and thyroid[7]. XBP1 is an X-Box protein that regulates gene expression for proper functioning of the immune system and cellular stress response[7]. Additional research into the genes listed in Fig. 9 that several of these genes are involved in tissue formation, allowing for the possible conclusion that the model is predicting cancer's subtype using the genetic profile of the tissue, regardless whether it is cancerous or healthy. Consequently, this means that although it is possible to predict a cancer's subtype based on the gene expression profile, genetic factors involved in determining the type of tissue may confuse the results. Therefore, additional filtering of the 20,000+ featured genes is necessary improve the generalizability of the model, and ensure that gene importance is determined based on its contribution to the tumor's formation.

4. Summary/Future Directions

TCGA is a rich source of datasets for cancer studies. The goal of this project was to work with gene expression data from TCGA and study the differences and similarities among 18 tumor types. PCA and t-SNE were run to reduce the dataset dimensionality and provide a better visualization of it. From PCA, we obtained a top 10 list of the genes that contribute the most to the dataset variance for each dimension of PCA. After consulting The Human Protein Atlas website and TCGA study results, we found that among the top 20 genes, 10 from each dimension of PCA, CREB1 is a cancer related gene. To do further analysis regarding CREB1 expression among the 18 types of cancer, we run 18 statistical tests. The high number of rejections confirm that each tumor type presents different average values of CREB1 expression compared to the global CREB1 expression average. We run RFC to find which are the top genes to classify the

tumor types. Our predicted and actual values for tumor type in READ and COAD cases showed that the supervised learning model had problems to distinguish those two types of cancer. This indicates that READ and COAD tumors, both gastrointestinal cancers, are very similar regarding gene expression. Future directions of our study would be creating a specialized RCF model to predict gastrointestinal cancers subtypes. In order to do that, a reduction of the current gene list in our dataset to a more specific set would improve the RCF ability to correctly classify gastrointestinal tumors. Next, it will be possible to get a top gene list of the most important genes from the new RCF model to classify the specific dataset. Finally, a next step would be run statistical tests for the top gene list.

References

1.The Cancer Genome Atlas Home Page. *The Cancer Genome Atlas - National Cancer Institute*(2018). at <https://cancergenome.nih.gov/>

2.Hutter, C. & Zenklusen, J. The Cancer Genome Atlas: Creating Lasting Value beyond Its Data. *Cell* 173, 283-285 (2018).

3. Tomczak, K., Czerwińska, P. & Wiznerowicz, M. Review The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. *Współczesna Onkologia* 1A, 68-77 (2015).

4.GEO Accession viewer. *Ncbi.nlm.nih.gov* (2018). at <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM1536837>

5.Expression of CREB1 in cancer - Summary - The Human Protein Atlas. *Proteinatlas.org*(2018). at <https://www.proteinatlas.org/ENSG00000118260-CREB1/pathology>

6.Expression of NCAPH in cancer - Summary - The Human Protein Atlas. *Proteinatlas.org*(2018). at <https://www.proteinatlas.org/ENSG00000121152-NCAPH/pathology#gene_information>

7.Reference, G. Genes. *Genetics Home Reference* (2018). at <https://ghr.nlm.nih.gov/gene/>

Appendix

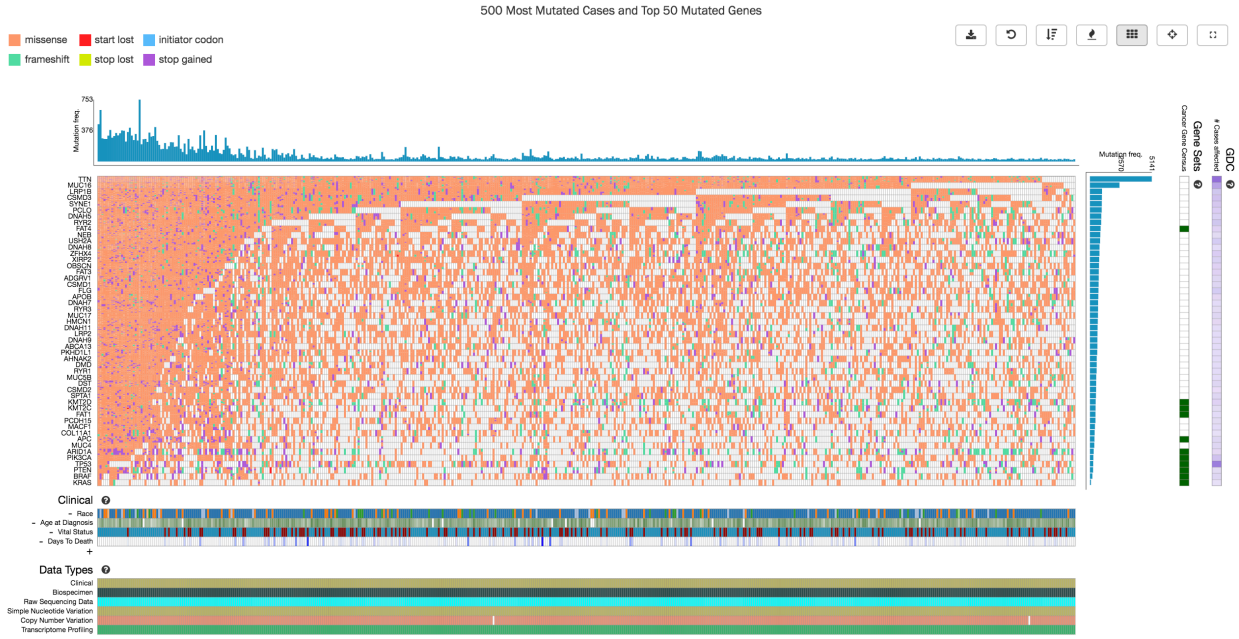

Figure A1: An overview of the properties of a subset of the data representing the top 50 mutated genes (y-axis) for the 500 most mutated cases (x-axis). The heatmap demonstrates the type of of mutation to occur (e.g. missense, frameshift, etc.) in that gene in the sample. The heatmap is clustered according to gene mutation frequency across samples and overall number of mutations per sample. Below the heatmap is a breakdown of the clinical metadata associated with each tumor, and below the clinical data is a summary of the available data types for that sample.

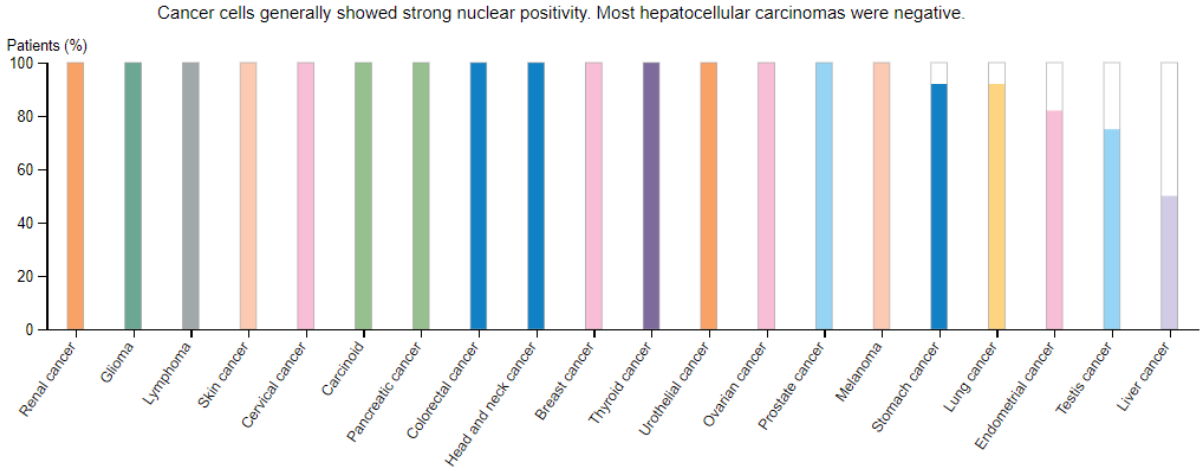

Figure A2: CREB1 protein expression summary in different types of cancer[5].

.

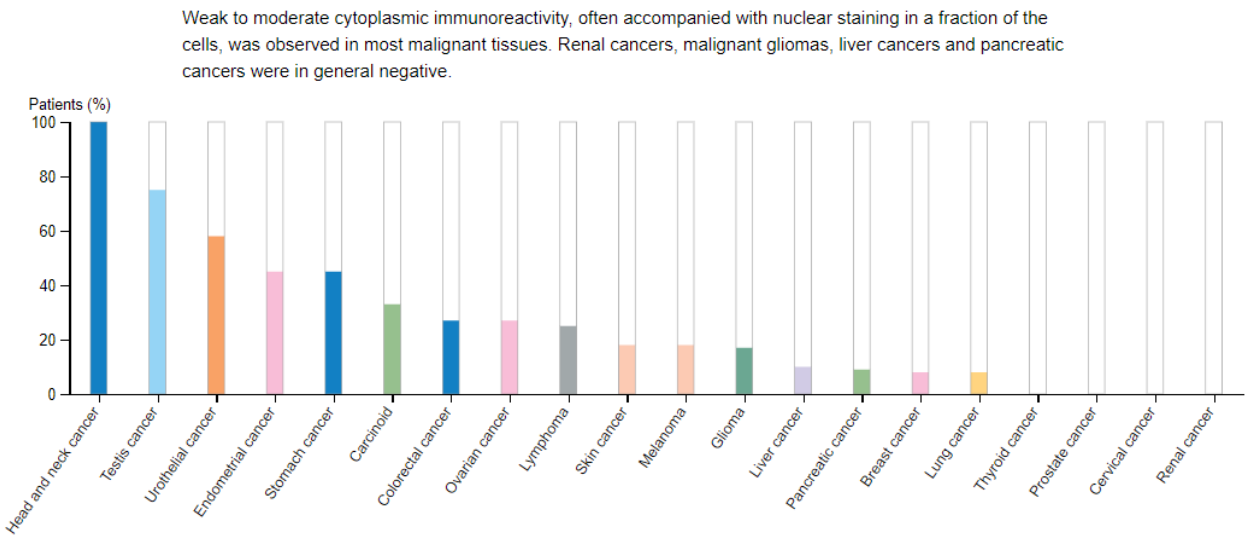

Figure A3: NCAPH protein expression summary in different types of cancer[6].

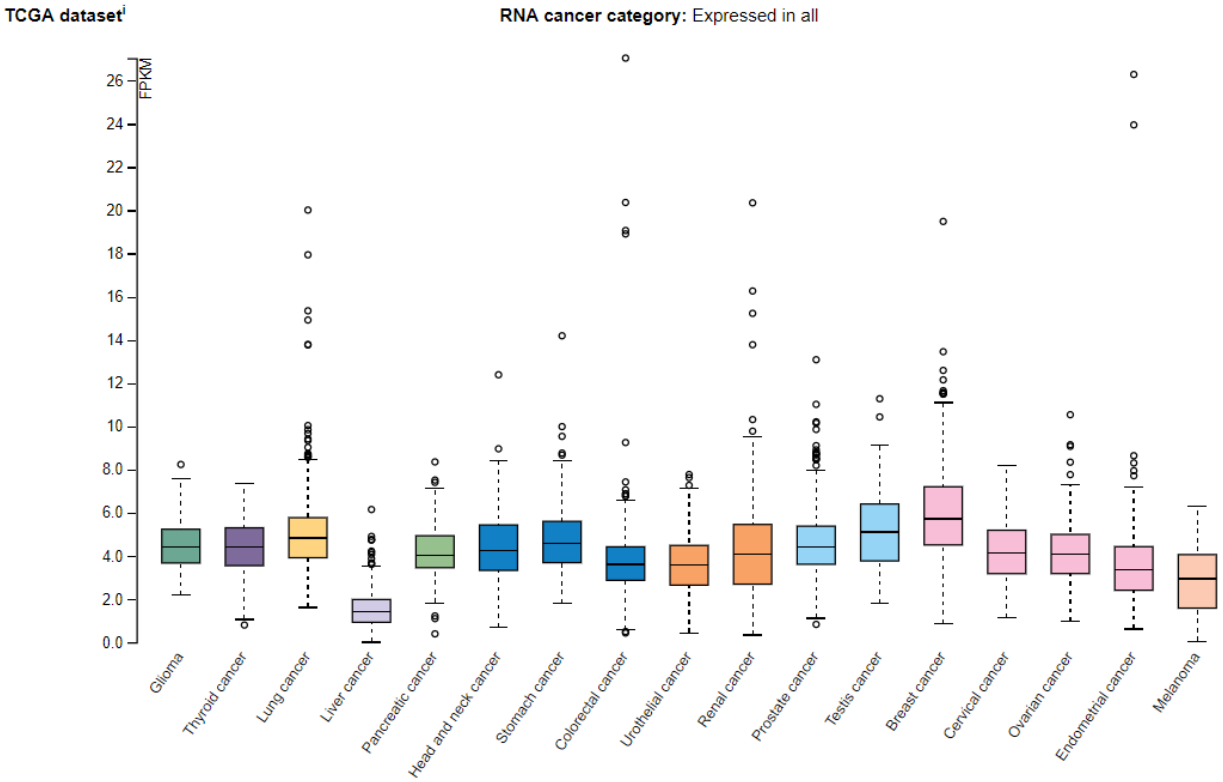

Figure A4: Gene expression values of CREB1 for each type of cancer[5].

Table 2: Cross-Validation Mean Accuracy Scores for various parameters of the Random Forest Classifier. The optimal parameters with the best accuracy score is bold.

Number of Trees	Max Depth	Max Features	Classification Criterion	Mean Accuracy Score (STD)	Number of Trees	Max Depth	Max Features	Classification Criterion	Mean Accuracy Score (STD)
5	3	log2	gini	0.570529 (0.025696)	5	12	log2	gini	0.848687 (0.003909)
10	3	log2	gini	0.676862 (0.018728)	10	12	log2	gini	0.892227 (0.009353)
20	3	log2	gini	0.702051 (0.017036)	20	12	log2	gini	0.915437 (0.004646)
40	3	log2	gini	0.736956 (0.029838)	40	12	log2	gini	0.926053 (0.005683)
5	3	sqrt	gini	0.639079 (0.047339	5	12	sqrt	gini	0.889349 (0.003867)
10	3	sqrt	gini	0.718964 (0.025523)	10	12	sqrt	gini	0.916517 (0.002067)
20	3	sqrt	gini	0.775639 (0.023693)	20	12	sqrt	gini	0.929831 (0.005611)
40	3	sqrt	gini	0.788413 (0.013427)	40	12	sqrt	gini	0.936128 (0.005381)
5	6	log2	gini	0.793271 (0.019619)	5	None	log2	gini	0.841130 (0.008732)
10	6	log2	gini	0.852105 (0.015860)	10	None	log2	gini	0.893487 (0.007836)
20	6	log2	gini	0.880533 (0.008953)	20	None	log2	gini	0.911119 (0.004608)
40	6	log2	gini	0.895106 (0.004846)	40	None	log2	gini	0.928392 (0.003288)
5	6	sqrt	gini	0.837531 (0.016714)	5	None	sqrt	gini	0.888629 (0.002232)
10	6	sqrt	gini	0.874775 (0.013792)	10	None	sqrt	gini	0.917416 (0.003480)
20	6	sqrt	gini	0.890068 (0.007954)	20	None	sqrt	gini	0.930191 (0.003160)
40	6	sqrt	gini	0.901943 (0.008025)	**40**	**None**	**sqrt**	**gini**	**0.943325 (0.003329)**

Number of Trees	Max Depth	Max Features	Classification Criterion	Mean Accuracy Score (STD)	Number of Trees	Max Depth	Max Features	Classification Criterion	Mean Accuracy Score (STD)
5	3	log2	entropy	0.619467 (0.024193)	5	12	log2	entropy	0.846168 (0.006518)
10	3	log2	entropy	0.680641 (0.024421)	10	12	log2	entropy	0.892227 (0.005621)
20	3	log2	entropy	0.677042 (0.029000)	20	12	log2	entropy	0.917596 (0.004485)
40	3	log2	entropy	0.727600 (0.019000)	40	12	log2	entropy	0.924793 (0.007150)
5	3	sqrt	entropy	0.638539 (0.037541)	5	12	sqrt	entropy	0.889709 (0.005649)
10	3	sqrt	entropy	0.724721 (0.016962)	10	12	sqrt	entropy	0.919935 (0.005854)
20	3	sqrt	entropy	0.737316 (0.017998)	20	12	sqrt	entropy	0.932350 (0.004131)
40	3	sqrt	entropy	0.743433 (0.018413)	40	12	sqrt	entropy	0.938647 (0.003785)
5	6	log2	entropy	0.819359 (0.016953)	5	None	log2	entropy	0.841850 (0.011435)
10	6	log2	entropy	0.868118 (0.011186)	10	None	log2	entropy	0.885031 (0.008202)
20	6	log2	entropy	0.891328 (0.005510)	20	None	log2	entropy	0.916337 (0.008030)
40	6	log2	entropy	0.901044 (0.003372)	40	None	log2	entropy	0.924073 (0.002700)
5	6	sqrt	entropy	0.880533 (0.009625)	5	None	sqrt	entropy	0.889888 (0.005535)
10	6	sqrt	entropy	0.900504 (0.002806)	10	None	sqrt	entropy	0.924433 (0.005551)
20	6	sqrt	entropy	0.914897 (0.003228)	20	None	sqrt	entropy	0.931450 (0.003687)
40	6	sqrt	entropy	0.923714 (0.005009)	40	None	sqrt	entropy	0.941166 (0.003270)