Predicting Cancer Types from Transcriptome Profiles

Brittany Gomez
Department of Computational Biology
Carnegie Mellon University
Pittsburgh, PA 15213
bgomez@andrew.cmu.edu

Caroline Springer
Department of Computational Biology
Carnegie Mellon University
Pittsburgh, PA 15213
cjspring@andrew.cmu.edu

Yiqing (Melody) Wang
Department of Computational Biology
Carnegie Mellon University
Pittsburgh, PA 15213
yiqingwa@andrew.cmu.edu

Abstract

Early detection and accurate diagnosis is critically important to patient survival. By utilizing transcriptomic data in parallel with conventional cancer detection techniques, the odds of early detection can be improved. The close proximity of the kidney and the adrenal gland increase the difficulty of accurate diagnosis and is thus an important area for improvement in accurate modelling based on transcriptome data. In this paper, we investigate the possibility of trascriptome-based diagnosis by applying several machine learning techniques to available transcriptome data from pediatric patients with kidney or adrenal gland cancer to improve the ability to distinguish between the two types of cancer and better understand the differentiating factors as they relate to gene expression.

1 Introduction

Traditionally, cancer diagnosis is conducted using different types of noninvasive scans, such as CT, PET, ultrasound, and MRI, and biopsy during surgery, during which the tumor cell morphologies are investigated under the microscope [1]. Transcriptomic analysis offers a more accurate alternative that can diagnose cancer earlier [2], thereby greatly increasing the chances of cure and paving the way to personalized medicine. This project explores different machine learning methods that can be applied to predict cancer type from transcriptome profiles, specifically between kidney (or renal) cancer and adrenal gland cancer. Given the adrenal glands' anatomical and physiological proximity to the kidneys, a classification between the two cancer types would speed up diagnosis and therefore effective treatment.

The effectiveness of transcriptome-based diagnosis has been demonstrated before. Whole-exome, genome, and transcriptome sequencing have been used to assist in identifying genes that are overexpressed in neuroblastoma, a deadly cancer with a 50% survival rate [3]. More recently studies have examined single-cell transcriptomes to categorize low to high-risk neuroblastomas, and hierarchical clustering maximized by silhouette scores has been used to recognize subtypes that led to age dependent prognosis [4]. Another single-cell transcriptomic study investigated the origins of cancer in adult and children kidney tumors using a form of k-means clustering, which was then used to determine optimal principal components after cross-validation, from which a connection was found between mesenchymal cells during development and childhood renal tumors [5]. Overall, substantial effort has been made to predict cancer prognosis or study cancer types and subtypes using transcriptome data. This project aims to extend the applications to classification of cancer types, thus cancer diagnosis.

Preprint. Under review.

The dataset used was downloaded from the GDC Data Portal, provided by the Therapeutically Applicable Research to Generate Effective Treatments (TARGET) studies funded by the National Cancer Institute. It contains HTSeq gene expression (transcriptome profiling) levels of primary pediatric untreated tumors from kidney and adrenal gland cancer, normalized to Fragments Per Kilobase of transcript per Million mapped reads upper quartile (FPKM-UQ). There are 218 kidney cancer samples and 65 adrenal gland cancer samples, each containing expression levels of 60483 genes.

In this paper, three models are trained on the data to classify adrenal and renal cancers. It is also possible to gauge how well each performs and identify key features of the dataset itself. Naive Bayes was selected as the sole generative model to examine underlying structures of both types of cancer. Logistic regression aided in investigating the linear separability of the data, while the neural network discerned whether more information could be gleaned from a nonlinear model.

2 Methods

Given the binary nature of the data, each method applies a two class classifier. Let $Y = [y_1, \ldots, y_m]$ such that y_i is an element of $[0, 1]$ where $y_i = 0$ indicates adrenal gland cancer and $y_i = 1$ indicates kidney cancer in m patients. Let $X = x_1, \ldots, x_n$ be the expression levels of n genes, with x_j being a vector in $\mathbf{R}^m$ where $j \in [1, \ldots, n]$. $x_j^{(i)}$ is the expression level of the jth gene for the ith sample. For the selected dataset there are 283 samples with normalized gene expression data for 60,483 genes, out of which 80% of the samples are used to train the models, and 20% of the samples are used to test the performance of the models. The three models implemented calculate the probability of a given sample being adrenal cancer or renal cancer based on gene expression data and outputs the more likely label.

Precision is calculated as such: $\frac{TruePositives}{TruePositives+FalsePositives}$.

Recall is calculated as such:$\frac{TruePositives}{TruePositives+FalseNegatives}$.

F1 is calculated as such: $2 * \frac{precision*recall}{precision+recall}$

2.1 Naive Bayes

Naïve Bayes (NB) is a parametric method and a generative supervised learning model. Before training the model, there are a few pre-processing steps. First, the expression levels in the entire dataset are mean centered. Then, the genes that have zero expression levels for all samples are deleted. Finally, a small value $\epsilon = 1 * 10^{-10}$ was added to all gene expression levels that equal 0. ϵ is chosen to be small enough that it will not have a significant impact on the dataset. Then, the dataset is split into training and testing datasets as described before.

Now the model is trained on the training dataset. A Naïve Bayes model does not directly learn the distribution of $P(Y|X, \theta)$. Rather, the model estimates it by learning parameters of $P(Y)$ and $P(X_j|Y)$ for $j \in [1, \ldots, n]$ based on (1). $P(Y)$ is defined to follow a Bernoulli distribution, with one parameter π (3). π is estimated to be the number of samples with Y = 1 divided by the total number of samples in the training dataset. $P(X_j|Y)$ is defined to follow a Gaussian distribution (2), with parameters μ_j and σ_j^2. μ_j and σ_j^2 are estimated through maximum likelihood estimation based on the kidney and adrenal training dataset, which are equivalent to the empirical mean and the empirical variance. All genes that resulted in a 0 standard deviation were eliminated from training and test dataset because it causes a division by zero error in (2).

During the testing phase, the standard practice is to calculate $argmax_y P(Y = y) \prod_{j=1}^{n} P(X_j|Y = y)$ to get the more likely label for a given test sample. However, to avoid a $P(X_j|Y)$ from becoming 0 after multiplying many small probabilities and making the product 0, the natural log of this was taken, namely we followed (4). Finally, the precision, recall, and F1 score are calculated.

$$P(Y|X_1, \ldots, X_n) = \frac{P(Y) \prod_{j=1}^{n} P(X_j|Y)}{P(X_1, \ldots, X_n)} \tag{1}$$

$$P(X_i|Y) = \frac{1}{\sigma_y\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x_i - \mu_y}{\sigma_y}\right)^2\right) \tag{2}$$

$$P(Y = y) = \pi^y(1 - \pi)^{1-y} \tag{3}$$

$$\hat{y} = argmax_y(lnP(Y = y) + \sum_{j=1}^{n} lnP(X_j|Y = y)) \tag{4}$$

2.2 Logistic Regression

Logistic regression is a parametric statistical model used to classify samples. Identifying relationships between disease and predictive features is a common application of logistic regression with cancer-specific use often relying on gene expression data as the features. While it can be generalized to multiclass classification, the task focused on in this paper is binary classification.

The model is defined by $P(Y|Z, \theta)$ where Z is equal to mean centered and normalized X and θ denotes the learned parameters, or coefficients, of the logistic regression function. Mean centering and normalizing is defined by $z_{i,j} = (x_{i,j} - \mu_j)/\sigma_j$ where μ_j, σ_j are the mean and variance of the jth column of X. Optimization of the model is achieved by minimizing the binary cross entropy using mini batch gradient descent where the selected batch size is 32. The descent optimization is iterated 100 times with a learning rate of 0.01. The output of the model is the probability of kidney cancer, $P(y = 1|Z, \theta)$. The probability of adrenal cancer can be directly inferred from this value as $P(y = 1|Z, \theta) + P(y = 0|Z, \theta) = 1$.

2.3 Neural Network

A fully-connected feedforward neural network is implemented using Tensorflow packages. It is a supervised model. A neural network is built on unprocessed data, termed "Neural Network without data preprocessing" in Table 1. Another neural network is built on preprocessed data, termed "Neural Network with data preprocessing" in Table 1. Preprocessing refers to the procedure in which all genes with 0 expression levels across all samples are deleted from training and test datasets. Without preprocessing, the neural network has an input layer with 60483 nodes, equal to the total number of genes. With preprocessing, the input layer has only 58233 nodes. Both neural networks have the same hidden layers and output layers. There are five fully connected hidden layers, with each consecutive layer having half the number of nodes of the previous layer, starting from 1024 nodes, and ending with the last hidden layer having 64 nodes. All hidden layers have tanh activation functions. The output layer is a fully connected layer with one node, representing the probability of $y = 1$. The activation function of the output layer is sigmoidal.

2.4 Data Visualization

On the full data, K-means clustering was carried out with variable cluster sizes, ranging from 1 to 11, resembling the method proposed by Bedoya-Reina et. al [4]. For each cluster size, the objective loss function and the silhouette score, a popular clustering metric, are calculated [4]. The silhouette score is a metric that compares inter and intra cluster distance. A value closer to 1 signifies good separation between clusters and a value closer to -1 indicates poor clustering. The objective value vs. cluster size was plotted (Figure 4A). The cluster labels were used to label a tSNE and PCA plot (Figure 5). The class labels were also used to plot tSNE and PCA plots to compare (Figure 4B and C). All tSNE plots were ran with parameters in Table 2. The first two principle components from PCA are shown in the graphs, which together represent approximately 91% of variability within the total data.

On just the kidney data the same methods were employed as above K-means clustering, a tSNE, and a PCA were plotted (Figure 6). The K-means clustering was done using variable cluster sizes of 1 to 11. The objective value vs. cluster size was plotted, and the highest silhouette score is recorded (Table 3). A tSNE was generated with the same parameters as listed in Table 2.

2.5 Identification of Most Differentially Expressed Genes

Given the high performance of our models, it is worthwhile to investigate which genes are most differentially expressed between adrenal and renal cancer, perhaps being key features that help train a

Table 1: Model performance across methods.

Method	F1 Score	Precision	Recall
Naive Bayes	0.97	1	0.92
Logistic Regression	0.98	1	0.97
Neural Network without data preprocessing	0.96	1	0.93
Neural Network with data preprocessing	0.99	1	0.98

great model. For each gene, a t-test is performed between the expression levels from adrenal cancer samples and renal cancer samples, and the p-value is corrected for multiple hypothesis, meaning that each p value is multiplied by the total number of genes (equivalent to dividing the cutoff value α by the total number of genes). Then, the genes with the five lowest p-values are selected. All of the five p-values are below 0.05 after multiple hypothesis correction (from the lowest to the highest: 6.95e-100, 8.856e-090, 1.94e-089, 2.54e-088, 8.51e-087]).

3 Results

All methods held out 20% of the data as a testing set that was not used in the training of the models. The performance of each method is detailed in Table 1. ROC curves and precision recall curves (PRC) are produced for logistic regression and neural networks. PRC has been shown to be more informative than ROC curves when there is class imbalance [6]. Because there are more kidney cancer samples than adrenal cancer samples, PRC is a better reflection of the model performance.

3.1 Naive Bayes results

Naïve Bayes was chosen as a complement to Logistic Regression given both are parametric supervised learning, but Naïve Bayes is generative where Logistic Regression is a discriminative model. Naïve Bayes learns $P(X, Y|\theta)$ to obtain $P(Y|X, \theta)$, whereas with Logistic Regression $P(Y|X, \theta)$ is the complete model. Ideally, Naïve Bayes would be used for independent or conditionally independent data. However, it has been shown that Naïve Bayes still works well even for weakly independent data.

The results showed a perfect score on precision, or a comparison of true positives to all positives. NB produced a recall of 0.916, showing some false negatives predicted. The F1 score was 0.9565, combining an assessment of both metrics. The precision results point to a perfect classification of positives, which means that no adrenal sample was classified as a kidney sample. This hints at a strong classification for kidney transcriptomic data. The recall however tells another story with a higher rate of false negatives, this means some kidney samples were classified as adrenal.

3.2 Logistic regression results

Logistic regression performed with an F1-score of 0.98. As with Naive Bayes, the model achieved a 1.0 precision score indicating it did not incorrectly diagnose adrenal gland cancer as kidney cancer. The recall of 0.97 shows that it did misclassify kidney cancer as adrenal cancer in some cases. Given the strength of predictability of the logistic regression, much of the data appears to be linearly separable. Figure 1 shows the characteristic curve of the data as well as the precision recall curve.

3.3 Neural network results

Similar to the other two models, neural networks with or without data preprocessing both have achieved high performance at classifying adrenal and renal cancer correctly. The precision is 1, as is the case in the other two models, signifying perfect classification on samples that are predicted to be kidney cancer (Table 1). The recall improved after deleting genes that have 0 expression levels across all samples, leading to a perfect precision recall curve and an F1 score of 0.99. Therefore, data that provides no useful information can create noise and reduce model performance, highlighting the importance of data preprocessing.

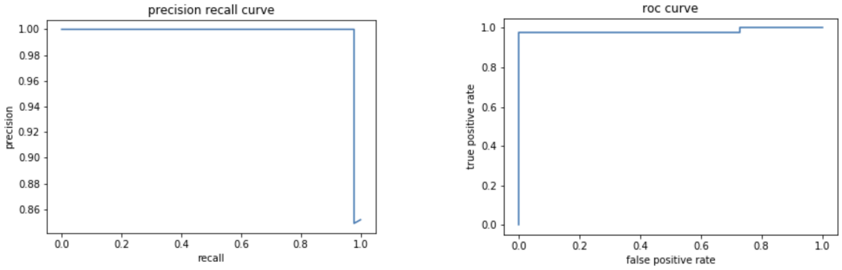

Figure 1: Characteristic curve and precision-recall for logistic regression model

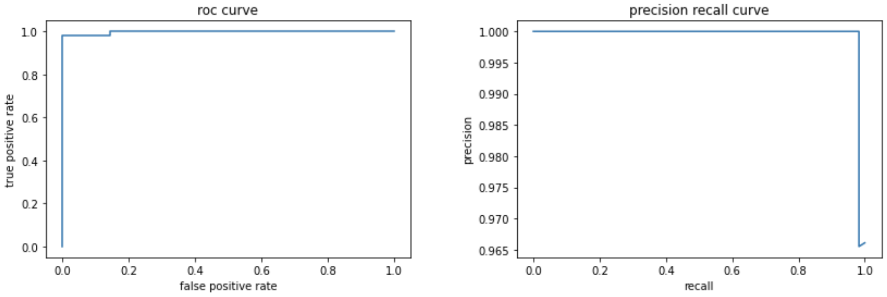

Figure 2: Characteristic curve and precision-recall for the neural network without data preprocessing

3.4 Data visualization results

Post model analysis, the dive into the unsupervised clustering seemed able to explain the high performance of our models. From figure 4A, the objective function experiences the biggest reduction when the cluster size increased to 2, suggesting that the best cluster size is 2. Similarly, the cluster size that has the highest silhouette score is 2 (Table 3). It seems that 2 should be an obvious choice, separating adrenal and kidney cancer samples into different clusters. Yet when we further explored the data using low dimensional visualizations such as tSNE and PCA with each sample color coded by its class label (Figure 4B and C), it was not convincing that K-means clustered the samples in a way that corresponds to the class labels. As it can be seen from Figure 4C, a portion of the kidney samples overlap with all adrenal samples, but there is another group of kidney samples off to the side.

This encouraged further examination with the cluster labels instead of the class labels as shown in Figure 5. As suspected, K-means distinguishes the two subtypes of kidney samples, rather than between kidney and adrenal samples. In other words, there was a subphenotype disguising itself as the binary class labels. This is evident in the high similarities between figures 4 and 5. The adrenal samples are all consumed within the larger cluster label while a small kidney cancer subphenotype is

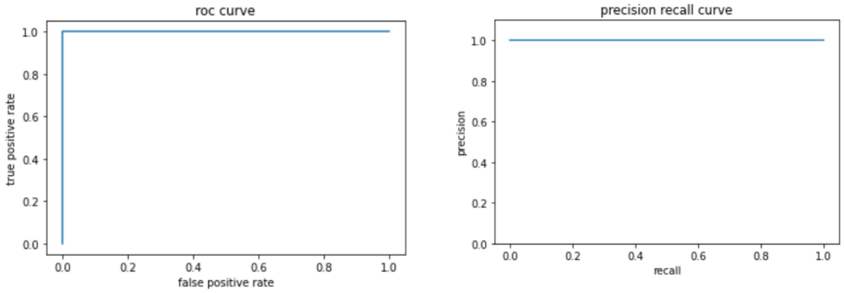

Figure 3: Characteristic curve and precision-recall for the neural network with data preprocessing

Table 2: Parameters for all tSNE plots.

Number of Components	Perplexity	Learning Rate
2	30	300

Table 3: the cluster size with the highest silhouette scores and the scores for all samples and just kidney samples.

	Cluster Size with the Highest Silouette Score	Highest Silhouette Score
All Samples	2	0.356
Kidney Only	2	0.398

observed in both the PCA and tSNE (Figure 5). It is noteworthy that even though the general local structures are preserved between the two tSNE graphs, they appear different. This is because tSNE is not deterministic and thus should be used primarily for data visualization purposes.

To support or deny this subphenotype hypothesis further, all adrenal data was removed. The results from the objective loss graph, tSNE, and PCA are show in Figure 6. The objective loss function looks almost identical to the one in Figure 4A. Thus, the predicted best cluster size of 2 is no surprise. It is notable that the silhouette score improves once the adrenal samples are removed further coalescing around the 2 possible phenotypes (Table 3). When exploring the tSNE plot, it is evident that there is a small ball of points located away from the larger amount of data (Figure 6B). This is similar to what is displayed in the tSNE plot from Figure 5B. Once again, when assessing the PCA plot the cluster labels are shown to be largely separated (Figure 6C), which mimics what is seen from the all samples cluster labeled PCA (Figure 5A).

It is puzzling that even though the samples do not sort themselves into clusters that perfectly correspond to their true labels, all our models perform exceptionally well on the classification task. A possible explanation for it is that the adrenal samples are **still** different enough from the kidney samples, such that the models can pick up on the differences and perform well. It is simply the case that differences within kidney samples are more prominent. Perhaps it is helpful to imagine a scenario as such: the models can confidently classify a test sample as either kidney subphenotype 1, kidney subphenotype 2, or adrenal.

3.5 Identification of Most Differentially Expressed Genes

The top five most differentially expressed genes are shown in Figure 7. IPO9, CTHRC1, KIAA1949 show unfavorable renal cancer prognosis, and AAR2 is detected in all cancer, whereas PDZD4 is detected in many [7], validating that the top most differentially expressed genes are related to cancer in general if not specifically renal cancer.

4 Conclusions

All our models have high precision, meaning that when the models encounter an adrenal sample, they will not wrongly assign it to be kidney. However, the recall shows that there are a handful of kidney samples that are wrongly classified as adrenal. That being said, the overall performance of all of the models is high, achieving F1 scores greater than 0.96.

Through clustering and visualizing the data in different ways, we are confident to report that there exists a subphenotype of pediatric kidney cancer based on transcriptome data. This is an interesting discovery, and future endeavors should focus on whether the subphenotypes correlate with different prognosis or presentations of kidney cancers. Analysis on genome sequencing or medical imaging data can also be carried out to confirm the existence of kidney cancer subphenotypes.

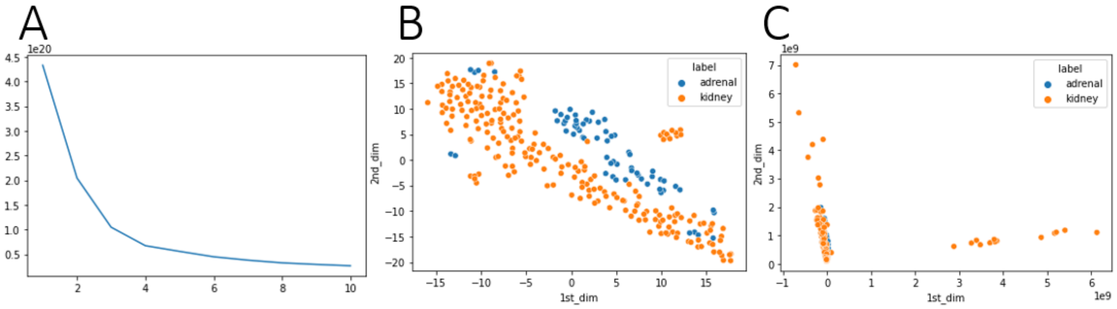

Figure 4: All samples shown in analysis, plots labeled with class labels. A) Objective loss function vs cluster size. B) tSNE plot with class labels. C) PCA plot with class labels.

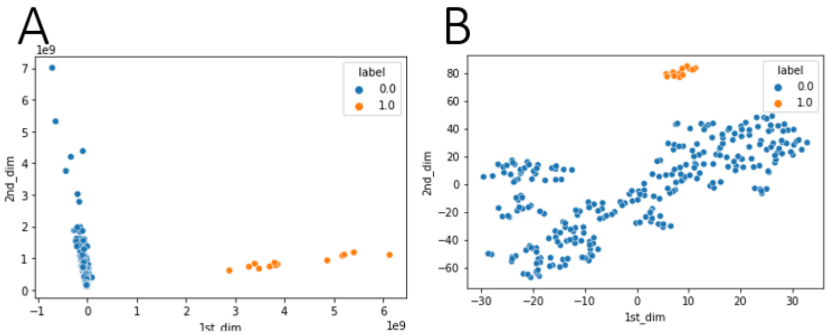

Figure 5: All samples shown in analysis, plots labeled with cluster labels. A) PCA plot with cluster labels. B) tSNE plot with cluster labels.

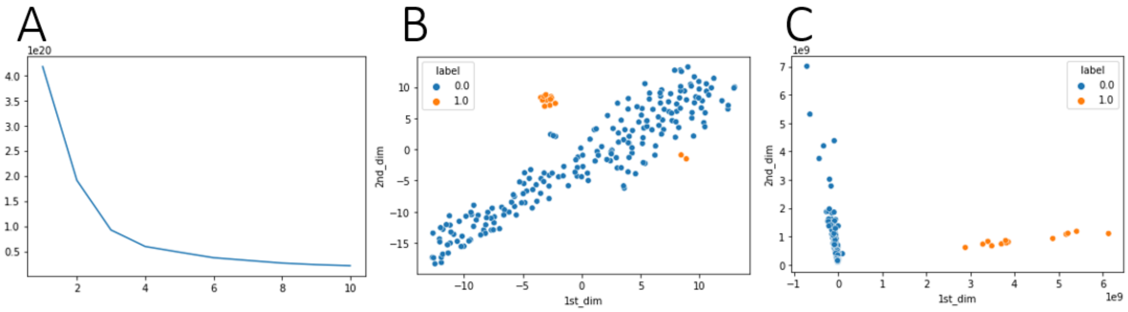

Figure 6: Only kidney samples shown in analysis, plots labeled with cluster labels. A) Objective loss function vs cluster size. B) tSNE plot with cluster labels. C) PCA plot with cluster labels.

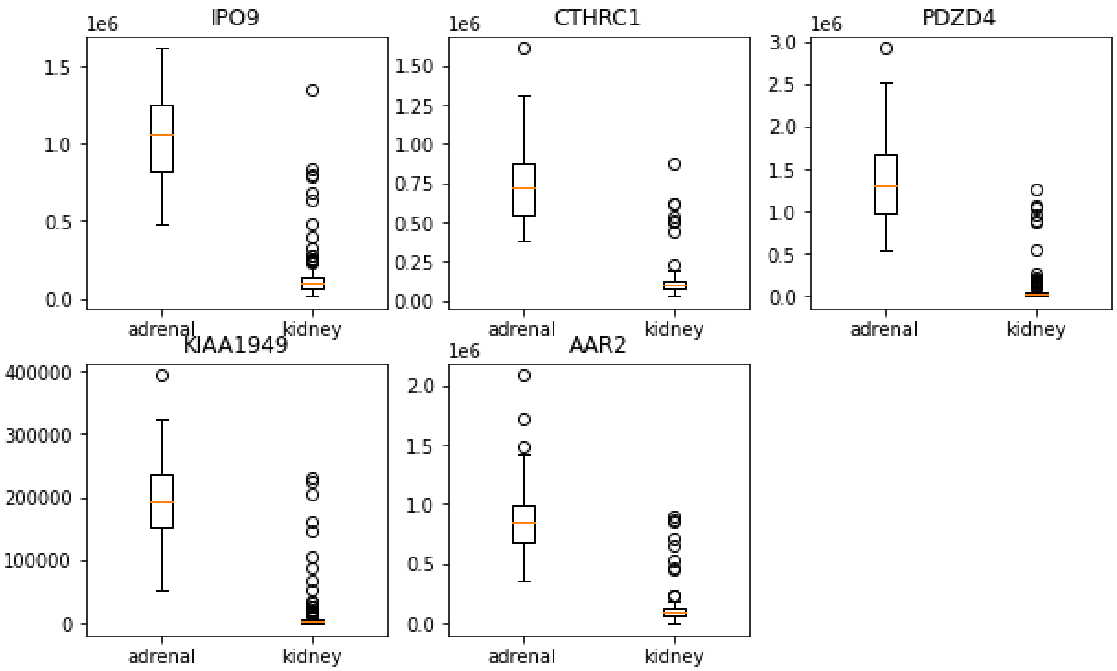

Figure 7: Top five most differentially expressed genes.

References

[1] Mayo Foundation for Medical Education and Research. (2021, April 27). Cancer. Mayo Clinic. Retrieved March 14, 2022, from https://www.mayoclinic.org/diseases-conditions/cancer/diagnosis-treatment/drc-20370594

[2] Sager, M., Yeat, N. C., Pajaro-Van der Stadt, S., Lin, C., Ren, Q., & Lin, J. (2015). Transcriptomics in cancer diagnostics: Developments in technology, clinical research and Commercialization. Expert Review of Molecular Diagnostics, 15(12), 1589–1603. https://doi.org/10.1586/14737159.2015.1105133

[3] Pugh TJ et. al. The genetic landscape of high-risk neuroblastoma. Nat Genet. 2013 Mar;45(3):279-84. doi: 10.1038/ng.2529. Epub 2013 Jan 20. PMID: 23334666; PMCID: PMC3682833.

[4] Bedoya-Reina, O.C., Li, W., Arceo, M. et al. Single-nuclei transcriptomes from human adrenal gland reveal distinct cellular identities of low and high-risk neuroblastoma tumors. Nat Commun 12, 5309 (2021). https://doi.org/10.1038/s41467-021-24870-7

[5] Young, M.D., Mitchell, T.J., Custers, L. et al. Single cell derived mRNA signals across human kidney tumors. Nat Commun 12, 3896 (2021). https://doi.org/10.1038/s41467-021-23949-5

[6] Saito, Takaya, and Marc Rehmsmeier. "The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets." PloS one vol. 10,3 e0118432. 4 Mar. 2015, doi:10.1371/journal.pone.0118432

[7] Human protein atlas — Uhlén M et al., Tissue-based map of the human proteome. Science (2015) PubMed: 25613900 DOI: 10.1126/science.1260419