Discovering Epigenetic Factors of Pediatric Cancer Types Using Clustering and Neural Networks

By: Ria Ghosh, Briana Nguyen, Julie Reyes, Sarala Sharma, Jane Zhao

Abstract

This study explores the mechanisms governing pediatric cancers, with a focus on understanding the role of epigenetic factors in disease initiation, progression, and therapeutic response. A comprehensive analysis is conducted using k-means and neural networks on 5 types of pediatric cancer: osteosarcoma (OS), neuroblastoma (NBL), acute myeloid leukemia (AML), Wilms Tumor (WT), and acute lymphoblastic leukemia (ALL), which involve distinct epigenetic mechanisms in cancer progression.

Data collection involved obtaining raw RNA counts for the mentioned cancer types from TARGET, which is in The Cancer Genome Atlas (TCGA), resulting in nearly 1000 samples. Samples that excluded the lower 25% of gene expressions were filtered out, focusing on significantly expressed genes for each cancer type. K-means clustering and neural networks were applied to analyze the gene expression data, identifying the specific patterns associated with each cancer type. The accuracy of the clustering method ranged from 82.5% to 100%, demonstrating its effectiveness in classifying cancer types. Out of the 2 analysis methods, the neural network scored higher than k-means clustering, with an average accuracy of 99.49%. High accuracy rates indicate confidence in predicting cancer types based on gene expression data models.

Overall, this research contributes significantly to our understanding of pediatric cancer and paves the way for innovative strategies aimed at improving clinical outcomes and patient care. Clustering and neural networks can be used as potential biomarkers for early detection and targeted therapy, tailoring to each individual's patient's conditions with high accuracy.

Introduction

With only 1% of the human genome containing protein-coding genes, it is crucial to understand how the few protein-coding genes interact to produce a diverse array of proteins, especially regarding cancer-controlling epifactors. These factors can activate oncogenes or suppress tumor suppressor genes, pivotal in cancer initiation, progression, and response to treatment. The genes coding for epi factors that bind to tumor suppressors and oncogenes influence the cancer subtypes that are observed in clustering. Understanding the interaction of protein-coding genes and their role in producing proteins is crucial, particularly in pediatric cancers. These cancers involve distinct epigenetic mechanisms that can activate oncogenes or suppress tumor suppressor genes, influencing cancer progression and treatment response. Investigating the epigenetic factors in these pediatric cancers can provide insights into their unique characteristics.

The study aims to analyze pediatric cancer types (OS, NBL, AML, WT, ALL) using k-means clustering and neural networks to understand complex genetic and epigenetic interactions. This clustering allows for the identification of specific patterns associated with each cancer type, potentially uncovering novel biomarkers for early detection and targeted therapy. The neural network analyzes complex patterns within gene expression data, enabling the identification of specific cancer subtypes with higher accuracy than traditional clustering methods. These cancers present unique challenges due to their occurrence in children's developing bodies, making the understanding of their molecular profiles critical for developing targeted treatments.

Methodology

Data Collection

Raw RNA counts for 5 pediatric cancer types (OS, NBL, AML, WT, ALL) studied in the paper were obtained from TARGET using the TCGAbiolinks package on R and filtered by 700 epi-factor genes, resulting in almost 1000 samples.

Filtering

Using R, the raw counts were converted from character to numeric values. The mean expression level for each gene was calculated across the dataset, followed by a filtering process that excluded the lower 25% of gene expressions, focusing the analysis on genes most heavily expressed across the collective cancer-type dataset.

K-Means Clustering

Each observation represents a patient with a set of gene read levels. We log 2 transformed, z-scored, and arranged data into a matrix of cancer counts to perform k-means clustering. We created a data frame called cancer_counts_t, a transpose of the original matrix, containing read levels of 986 filtered genes for 525 patients. Performing t-SNE dimensionality reduction on the matrix with a random seed set to 66 allowed us to plot two-dimensional clusters for the most influential dimensions of gene expression. The dataset was arranged into clusters using the random seed set to 43, 45, or 47 to test the resemblance of clusters to the five pediatric cancer types categorizing each patient. The accuracy of our clusters was evaluated by calculating the percentage of patients in cancer that belong to the dominating cancer type of each cluster. Assigning each cluster to one cancer allows for prediction based on epi factor gene expression, unlike assigning each cancer to its most popular cluster. However, as visible in the Table 1 bar graphs, clustering on classification causes redundancy and bias toward more represented clusters. To subside this bias, we included the option to cluster with 7 centers to increase sensitivity to smaller cancers like OS, which succeeded in clustering with seed 45 as seen in Table 1.

Neural Network

A neural network model was created to predict pediatric cancer types based on the epi-factor gene expression counts and used to further analyze the gene expression data of each cancer. The data was initially preprocessed by label encoding the cancer types into numerical variables (Table 2), normalizing the gene expression data within the range (0, 1), and splitting the data into training, validation, and testing sets. The model (Figure 2) was then created using the keras Sequential model with a total of 4 layers: 1 input layer with 525 neurons and a “relu” activation function, 2 hidden layers with a total of 55 neurons, and the “relu” activation functions, and 1 output layer with 5 neurons and a “softmax” activation function. The number of hidden neurons was also based on the value, $\sqrt{number\ of\ input\ neurons\ \ *\ number\ of\ \ output\ layer\ neurons}$. After, the model was compiled using the keras Sparse Categorical Cross Entropy loss function and the Adam optimizer with a learning rate of 0.0001 and fitted using the training and validation data. Model performance was then evaluated by comparing accuracy and loss and feature importance by comparing feature weights.

Results

K-means clusters plotted revealed variable classifications and thus accuracies for different random seed initializations. Accuracy was evaluated by the percentage of clusters that belonged to the dominating cancer type. For the centroid initialization with seed 43, we saw cancers represented with accuracies above 80% for all clusters. For centroid initializations of 45, all cancers were represented, however, misclassification was high for clusters 1, 3, and 5 out of five clusters, and clusters 1 and 3 out of seven clusters. Due to the increased sensitivity to the OS cancer type among seven clusters, accuracies increased. For the centroid initialization with seed 47, only ALL, AML, and NBL patients were represented, with the largest misclassifications apparent in cluster 4 for both plots. However, the greatest misclassification occurred for cancer type NBL in cluster 4. For example, according to Table 1 which displays K-means accuracy for seeds 43, 45, and 47, the accuracy of clustering data still depends on centroid initialization. The kmeans() function for five clusters with seed 45 particularly shows the between-cluster sum of squares variance of 46.0% of total variance to be less than within-cluster variance of 54.0%, which is not ideal for prediction purposes. However, each cluster has its particular variance, with certain clusters corresponding to having the highest percentages of within-cluster variance. For example, the plot of seed 45 variance of five clusters in Table 1 F shows the cluster classifying NBL patients to possess 12.5% of total variance. Furthermore, the variance of clusters also depends on seed initialization, as seen in the plot of seed 43 variance of five clusters, which has a lower variance for ALL, NBL, and OS dominant clusters than the general variance in seed 45 variances. The dissonance between cluster accuracy and cluster variance for seed 45 of five clusters shows that cluster variance made the majority of variance, and the contribution to the variance of a misclassified cluster is not obvious from accuracy.

The neural network model evaluation showed 99.68% accuracy for training data, 99.49% for testing data, 98.73% for validation data, and 99.49% for total data. In the model accuracy plot (Figure 3), the training data and validation curves were similar with very high values, but with a slight difference when the Epoch value is 25. In the model loss plot (Figure 3), the training and validation curves had low values and minimal gaps. From the Top Gene-Feature weights plot (Figure 4), all the top genes have a weight well above the median of around 2.3. ANOVA (Table 4) was done for gene expression of JDP2, ZNF711, and APBB1 (top 3 genes) against the 5 cancer types. All three of these genes had very low p-values (2.2e-59, 9.1e-123, 9.3e-291).

Discussion

Our work addressed one of the guiding questions considered in the paper “How does the epi factor landscape of adult tumors compare with pediatric tumors?” Unlike the analysis of cancer subtypes for survival outcomes, our analysis focuses on clustering to identify types. Considering the clustering results of K-means clustering, accuracy is contingent on random initialization and the number of clusters, with poorly spaced centroids and too few clusters causing different categories to be clustered. Reflecting on the result of the Neural Network, the minimal gap between the training and validation curves suggests that the model's performance remains consistent across different datasets, further supporting its generalizability and stable predictive performance. Additionally, the top gene features, JDP2, ZNF711, and APBB1 had extremely low p-values, indicating that their gene expressions are significantly different and should be further investigated to learn more about the mechanisms responsible for the development of each cancer. By highlighting the role of epigenetic factors, it opens up new avenues for research into how these factors contribute to cancer initiation, progression, and response to therapy.

References

Cheng, M.W., Mitra, M. & Coller, H.A. Pan-cancer landscape of epigenetic factor expression predicts tumor outcome. Commun Biol 6, 1138 (2023). Accessed 8 Mar. 2024 https://doi.org/10.1038/s42003-023-05459-w

“BiomaRt, Bioconductor R Package.” E!Ensembl, useast.ensembl.org/info/data/biomart/biomart_r_package.html. Accessed 8 Mar. 2024.

Memon, Quratulain. “How to Build a Simple Neural Network Using Keras.” Educative, www.educative.io/answers/how-to-build-a-simple-neural-network-using-keras. Accessed 8 Mar. 2024.

Morgan Morgan, Martin Obenchain, et al. “Summarized Experiment for Coordinating Experimental Assays, Samples, and Regions of Interest.” Bioconductor, 5 Jan. 2023, bioconductor.org/packages/devel/bioc/vignettes/SummarizedExperiment/inst/doc/SummarizedExperiment.html.

“TCGAbiolinks: Downloading and Preparing Files for Analysis.” Bioconductor, 24 Oct. 2023, bioconductor.org/packages/release/bioc/vignettes/TCGAbiolinks/inst/doc/download_prepare.html

“Keras Documentation: Models API.” Keras, keras.io/2.15/api/models/. Accessed 8 Mar. 2024.

Appendix

Table 1.

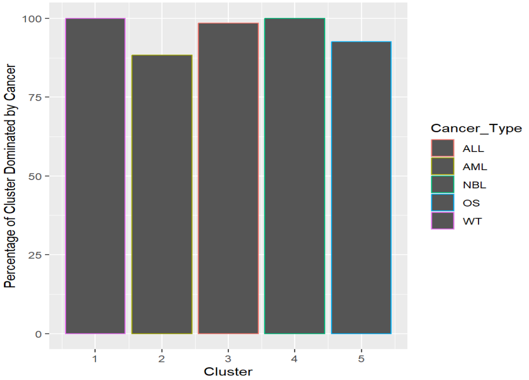
<u>K-Means Clustering Accuracies</u>

Table 1. K-means clustering accuracies for different seeds

A.

```
 Cluster Cancer_Type count total_count Percentage
       1          WT   130         130  100.00000
       2         AML   296         335   88.35821
       3         ALL   261         265   98.49057
       4         NBL   161         161  100.00000
       5          OS    88          95   92.63158
```

seed 43
5 clusters

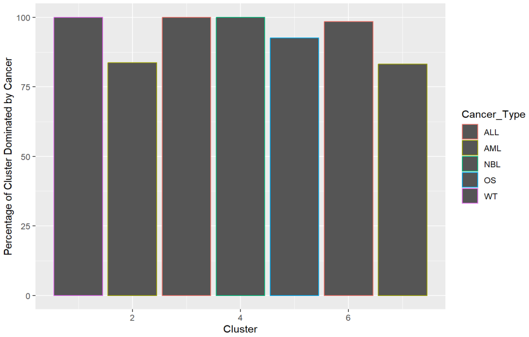


WT-AML-ALL-NBL-OS

B.

```
 Cluster Cancer_Type count total_count Percentage
       1          WT   130         130  100.00000
       2         AML   159         190   83.68421
       3         ALL   109         109  100.00000
       4         NBL   161         161  100.00000
       5          OS    88          95   92.63158
       6         ALL   132         134   98.50746
       7         AML   139         167   83.23353
```

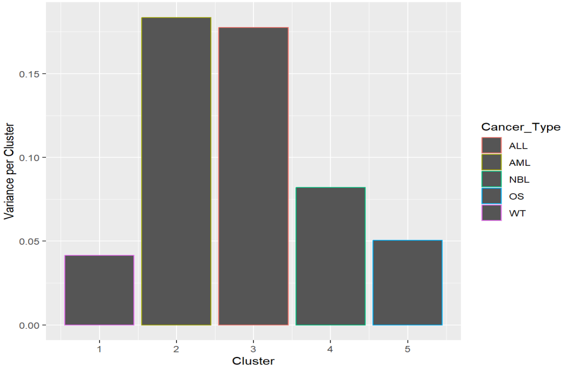
seed 43
7 clusters



WT-AML-ALL-NBL-OS-ALL-AML

C

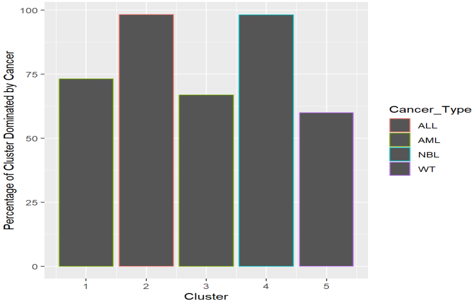
```
Cluster   per_cluster_variance
   1          0.04153088
   2          0.18341771
   3          0.17752420
   4          0.08201001
   5          0.05054999
```



D.

```
 Cluster per_cluster_variance
       1     0.11130059
       2     0.11596953
       3     0.10517324
       4     0.08519536
       5     0.12192667
```

seed 45
5 clusters

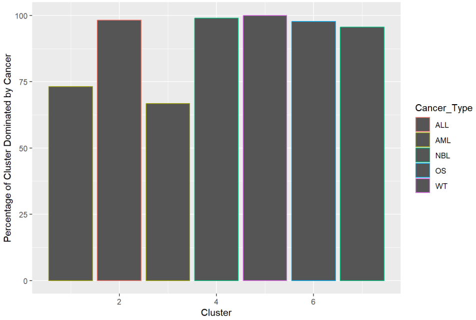


AML-ALL-AML-NBL-WT

E.

```
 Cluster Cancer_Type count total_count Percentage
       1         AML   158         216   73.14815
       2         ALL   173         176   98.29545
       3         AML   139         208   66.82692
       4         NBL    97          98   98.97959
       5          WT   130         130  100.00000
       6          OS    88          90   97.77778
       7         NBL    65          68   95.58824
```

seed 45
7 clusters

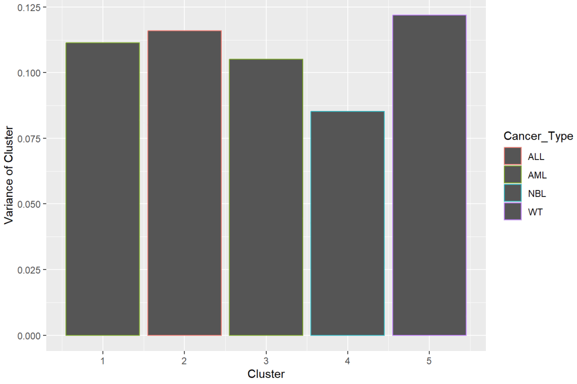


AML-ALL-AML-NBL-WT-OS-NBL

F.

```
 Cluster   per_cluster_var
       1      0.11130059
       2      0.11596953
       3      0.10517324
       4      0.08519536
       5      0.12192667
```

seed 45
5 clusters variance

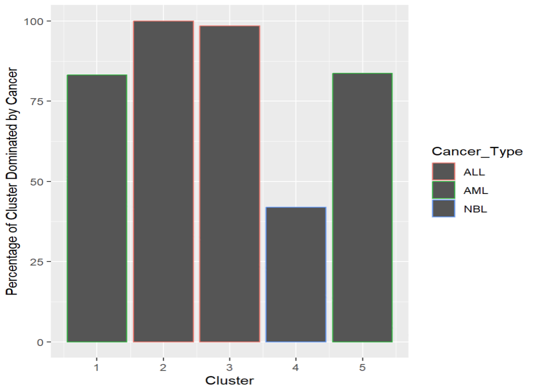


AML-ALL-AML-NBL-WT

G.

```
 Cancer_Type Cluster count total_count Percentage
         ALL       3   132         300         44
         AML       5   159         300         53
         NBL       4   162         162        100
          OS       4    88          88        100
          WT       4   136         136        100
```

seed 47
5 clusters

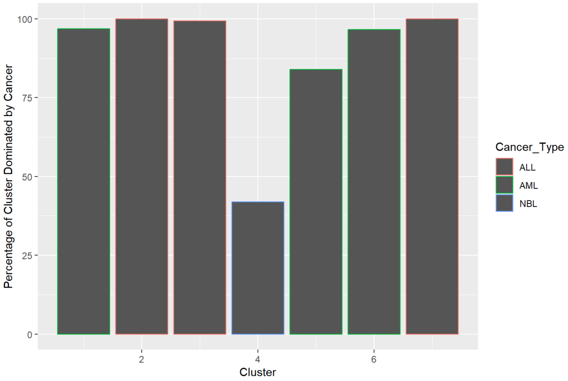


AML-ALL-ALL-NBL-AML

H.

```
 Cluster Cancer_Type  count  total_count  Percentage
1       1          AML    61           63   96.82540
2       2          ALL    30           30  100.00000
3       3          ALL   161          162   99.38272
4       4          NBL   162          386   41.96891
5       5          AML   125          149   83.89262
6       6          AML   113          117   96.58120
7       7          ALL    79           79  100.00000
```

seed 47
 7 clusters



AML-ALL-ALL-NBL-AML-AML-ALL

Figure 1.

T-SNE Clustering for Seed 43 K-Means Clustering

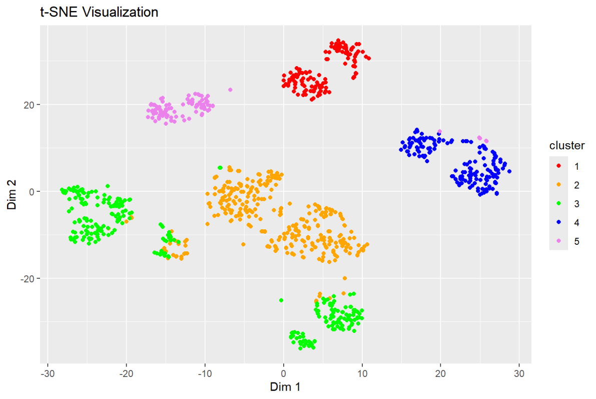
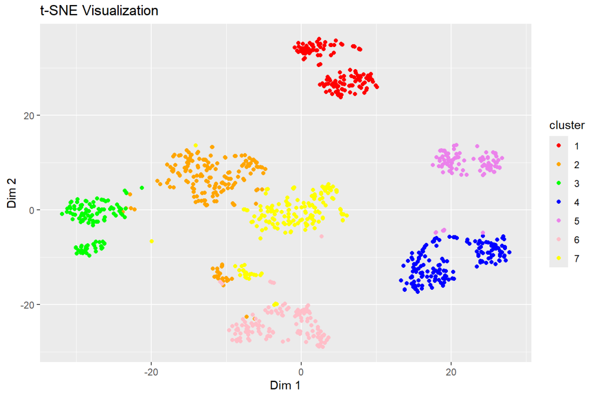
5 Clusters Seed 7	
7 Clusters	

Table 2.

Neural Network Cancer Label Encoder Values

Cancer Type	Numeric Label
Acute Lymphoblastic Leukemia (ALL)	0
Acute Myeloid Leukemia (AML)	1
Neuroblastoma (NBL)	2
Osteosarcoma (OS)	3
Wilms Tumor (WT)	4

Figure 2.

Neural Network Diagram Visualization

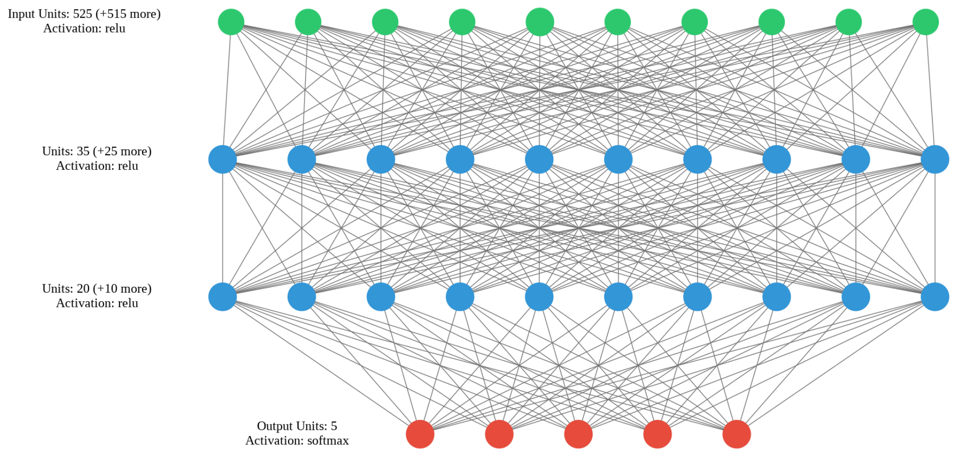


Figure 3.

Neural Network Model Accuracy and Loss

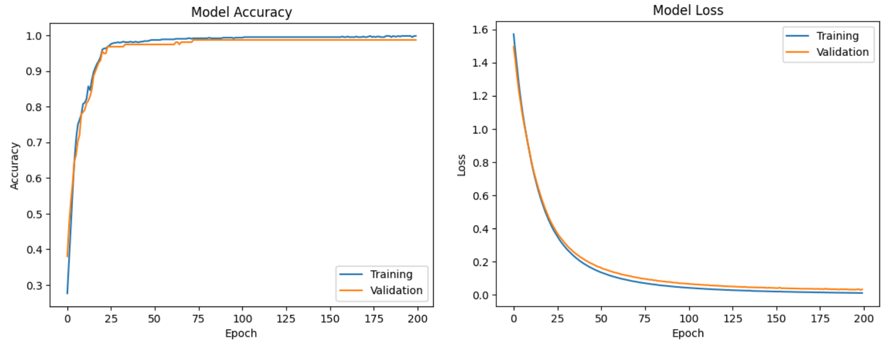


Table 3.

Neural Network Evaluation Accuraries

Dataset	Accuracy
Training	99.68%
Testing	99.49%
Validation	98.73%
Overall	99.49%

Figure 4.

Gene Feature Weights

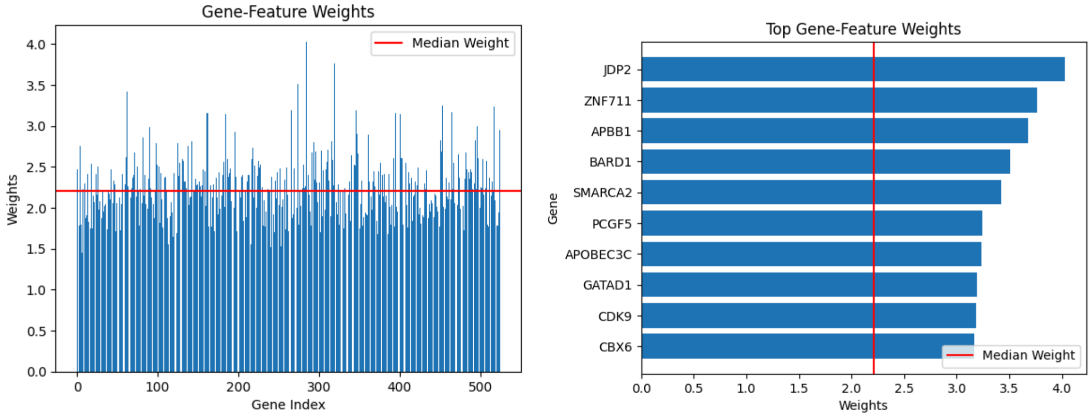
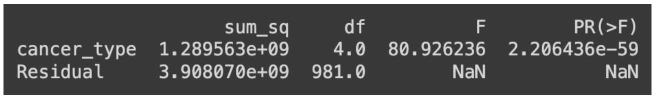
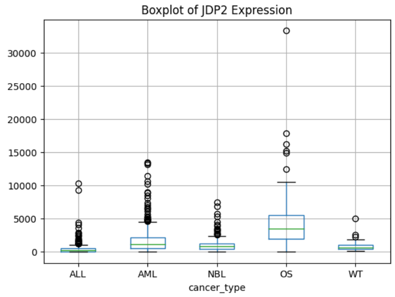
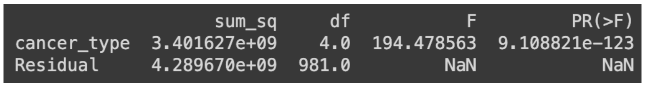
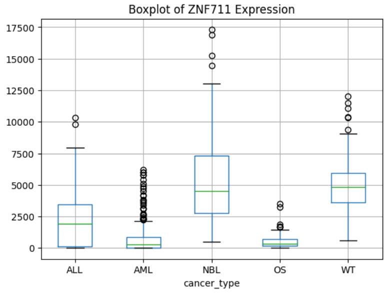
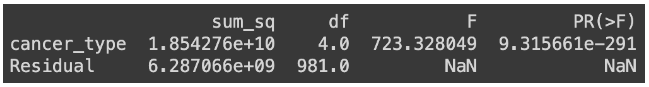
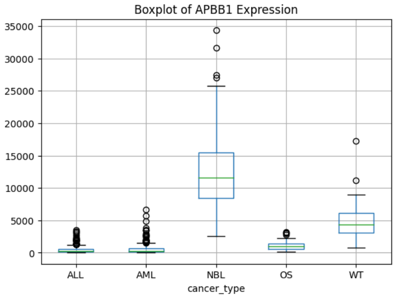


Table 4.

ANOVA of JDP2, ZNF711, and APBB1 Expression Based on Cancer Type

Gene	ANOVA	Boxplot
JDP2	 sum_sq df F PR(>F) cancer_type 1.289563e+09 4.0 80.926236 2.206436e-59 Residual 3.908070e+09 981.0 NaN NaN	
ZNF71 1	 sum_sq df F PR(>F) cancer_type 3.401627e+09 4.0 194.478563 9.108821e-123 Residual 4.289670e+09 981.0 NaN NaN	
APBB1	 sum_sq df F PR(>F) cancer_type 1.854276e+10 4.0 723.328049 9.315661e-291 Residual 6.287066e+09 981.0 NaN NaN	

Reference 1: Source Code (https://github.com/rghosh1353/cancer_epifactors)

- Folder for each step of the study, code in either Python or R