Multiclass Classification of Cancer Types using TCGA Datasets

Jakub Markiewicz

mjmarkiewicz@outlook.com

May 6, 2024

1. Introduction

Cancer is a disease of abnormal cell growth that spreads throughout the body, often being fatal if left undiagnosed. Gene expression could play a crucial role in the classification and early detection of cancer types. Genes are chains of nucleotides within Deoxyribonucleic Acid (DNA), with information to synthesize proteins through transcription. Gene expression therefore dictates the way this information is processed to assemble such a protein - some genes have high expression values in a named tissue, where other genes may have low expression values. Therefore, this genetic characteristic could correlate to the probability of certain cancerous cells being produced; low expression of certain genes inhibit the creation of suppressor proteins, which may lead to unregulated cell division [1].

Machine Learning (ML) algorithms have proven to be effective at understanding deep patterns within large datasets. Where gene expression provides a strong overview of the underlying biochemical processes within human cells, ML classifiers may efficiently train on thousands of its features (genes) to discriminate and classify by tumor type. **Bladder**, **lung**, **liver**, **prostate** and **colorectal** cancers are 5 common tumors observed in men [2]. With 3.2 billion base pairs in the human genome, novel ML approaches directly tackle the ”Big Data” problem within genomics, and therefore should be utilized to ensure early detection of cancer to improve chances at survival of patients.

This project aimed to demonstrate the implementation of such a multinomial classifier algorithm using the TCGA (*The Cancer Genome Atlas Program*) dataset in order to test for a gene-cancer type relationship. The report begins with an overview of the dataset in Chapter 2, the pipeline design (dimensionality reduction and algorithm training) in Chapter 3, and a performance evaluation of the learner in Chapter 4. Finally, a conclusion of the findings as well as ideas for future research can be found in Chapter 5.

2. Selecting & Exploring The TCGA Dataset

Cancer Tumor	**Total Samples**
Bladder (BLCA)	426
Lung (LUNG)	1129
Liver (LIHC)	423
Prostate (PRAD)	550
Colorectal (COADREAD)	434
Total	**2962**

Table I: TCGA dataset patient sample sizes for five cancer types.

This project was interested in creating a multiclass classifier for the bladder, lung, liver, prostate and colorectal cancers observed primarily in men. The relevant TCGA datasets were obtained from the UCSC Xena portal [3]. These gene profiles have been obtained via. IlluminaHiSeq 2000 RNA sequencing by the University of North Carolina TCGA genome characterization center. This high quality data samples expression values of 20531 genes within cancerous tissue for each patient. The total size of the five datasets is summarised by Table I.

Once this data was downloaded, it was then imported into Jupyter Notebook for the creation of the classifier pipeline. Using Pandas, the five gene expression profiles were set up as dataframes and subsequently transposed into a *sample x identifier* (gene name) format.

Classification labels were generated by encoding each sample with its relevant cancer type. To avoid later issues in training, a number was given for individual tissues rather than utilizing string labels. Moreover, bladder samples were encoded as 1, lung samples as 2, liver samples as 3, prostate samples as 4 and colorectal samples as 5.

Stratified sampling was then utilized to generate the training dataset as well as the testing dataset. The randomness of this sampling method allows for a greater representation of gene expressions in the training data which is useful for optimizing the generalisation ability of the algorithm. The datasets underwent a classic 70/30 split which is commonly utilized by similar studies and, in general, is agreed to be optimal for training an ML algorithm [4].

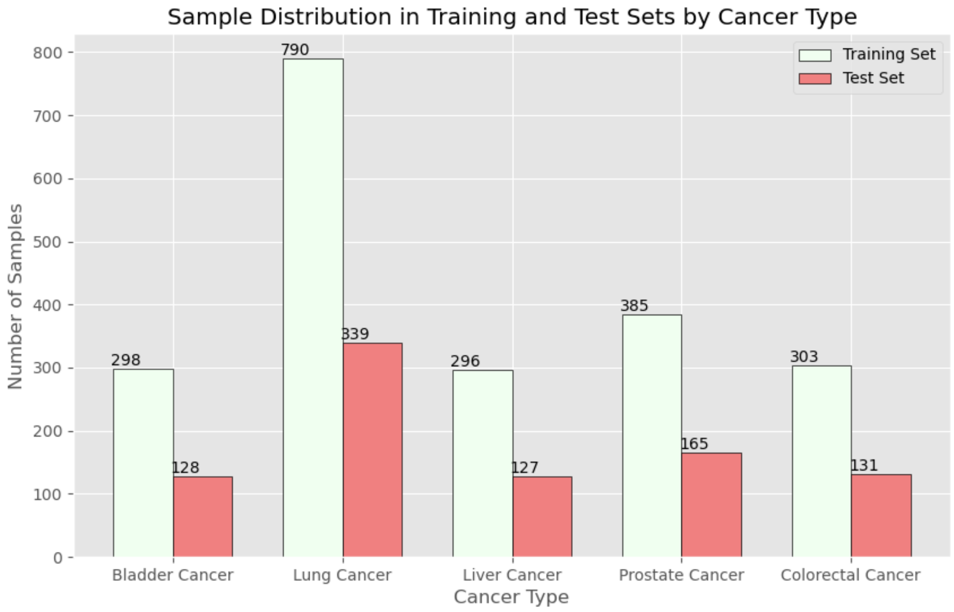


Figure 1: Bar chart showing distribution of training and test datasets after stratified sampling.

Fig 1 indicates that lung cancer exhibits a larger sample size of data relative to other cancer types. This may imply a sample bias, which may cause the learner to become overfitted by learning the intricacies of the lung gene expression profile. Fortunately, due to the large number of features for each sample, it was assumed that this would have a negligible effect on the overall performance of the algorithm.

The five training sets and five testing sets were then concatenated to yield one training set and one testing set, each with 5 of the encoded ground-truth labels. The training dataset containing 2072 samples and testing dataset containing 890 samples maintained the 70/30 split, ensuring a representative gene profile for pipeline design of the multiclass model.

TCGA data does not show additional significant biases which may may effect the accuracy and generalisation of the model.

3. Pipeline Design

3.1. Feature Selection

The feature space consists of 20531 gene expression values for each sample. In the training data, that is:

2072 samples × 20531 identifiers = 42.5 million total features

Although larger datasets generally improve the performance of learners, the compute required to process all the data becomes greater. It is useful to reduce the dimensions of the feature space in such situations to balance computational complexity with model performance. This can be done through statistical methods such as Principal Component Analysis (PCA). PCA reduces the feature space of the dataset while maintaining some designated level of data variance. This, however, can also be computationally intense - therefore it is practical to reduce the dataset through other means before applying PCA.

This was done by introducing a threshold variable to the gene expression profile. In gene expression data, it is common that most genes are off by having low expression values. It can be therefore assumed that such genes do not contribute to the formation of cancer cells as much as abnormally high expressed genes [5]. 9464 samples or 43% of the gene expression data falls under the threshold value of 0.15 across all cancer types. This method of gene filtration across all tumors ensures that this process affects all five datasets equally. Although highly expressed genes may contribute to the irregular synthesis of tumor cells, this assumption may indeed be limited by ignoring the relationship between abnormally low-expressed genes and its affect on transcription.

Nevertheless, the training data sample size was reduced to 11067 which almost halves the original dimensions. Subsequently, PCA was then applied on the dataset to further extract only the relevant genes. As a result, the training dataset was further reduced to 656 genes. Finally, the total feature space becomes:

2072 samples × 656 identifiers = 1.4 million total features

This is a 30x reduction in the feature space and allows for more efficient model training and performance analysis. The threshold and PCA variance parameters may be altered to reduce or increase the effect of feature selection in this stage of the pipeline.

3.2. Algorithm Training

Model selection was chosen primarily based on the remaining large feature space - it is common to observe overfitting for large datasets. Three learners were trained in order to achieve the classification task.

The first model trained was a **Lasso Regression** algorithm. In literature that takes a similar ML approach to genomic data, Lasso models are utilized to further reduce the feature space through a regularization/shrinkage process [6]. Lasso penalizes regression coefficients that do not contribute much to the classification outcome, thus reducing them to zero. Finally non-zero coefficients are extracted for model training. Often, Lasso is used in conjunction with previous pre-processing steps, however the output is directly used to train a logistic regression model.

The second model trained was a **Decision Tree** algorithm. Decision trees further tackle overfitting by controlling the depth parameter, while also allowing to extract the most relevant features (genes) that are used to dictate a classification outcome. This is incredibly useful, as it allows for further analysis of genes most responsible for a certain tumor.

Finally, an **SVM** model was also trained. SVM models are particularly effective at modelling non-linear relationships by obtaining the optimal hyperplane that discriminates between classes typically for binary classification problems. Scikit however allows for multi class implementation and often SVM models have great general accuracy across genomic datasets [7].

4. Performance Evaluation

The performance of the trained models was evaluated using standard metrics such as the confusion matrix, accuracy, precision, recall and F1 scores. The confusion matrix reports the number of correctly and incorrectly predicted cancer-types when the trained model is applied on the testing dataset. The model's predictions are compared to the ground-truth labels of the testing dataframe. Correct classifications are shown in the diagonal elements of the matrix. Incorrect classifications are shown in the other matrix elements and signify when the model predicts a cancer type when it is in fact another. The heatmap of Fig 2 shows good and almost equal classification accuracy across all three models.

It is possible to obtain the accuracy, precision, and F1 score for each individual class through the Scikit Python module. Finally, the overall performance of the model is dictated by the F1 score, which is averaged over all five classes. For all these metrics, the higher the result, the better the performance of the model.

The Lasso regression model obtained the best **F1 score of 0.99** - indicating that the model is capable of correctly classifying close to every single cancer type based on gene expression. This is aligned with values in literature with Lasso models scoring similar F1 values [6]. Although the Decision Tree model scored lower, it was possible to extract the

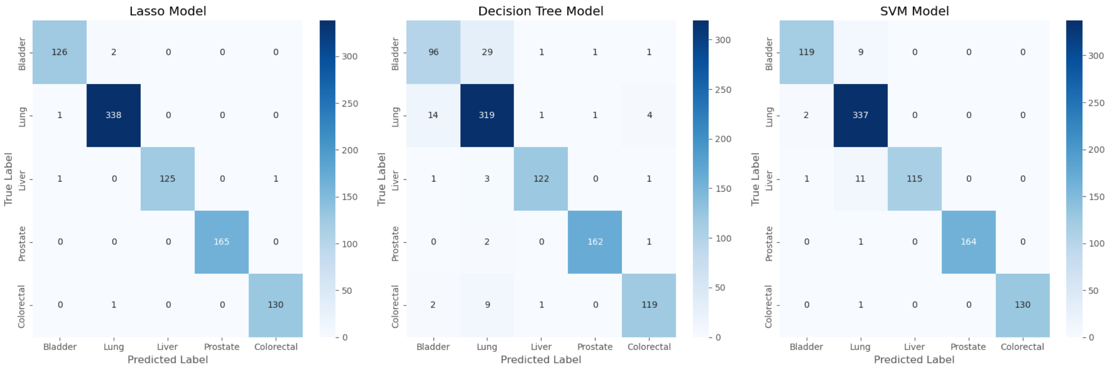
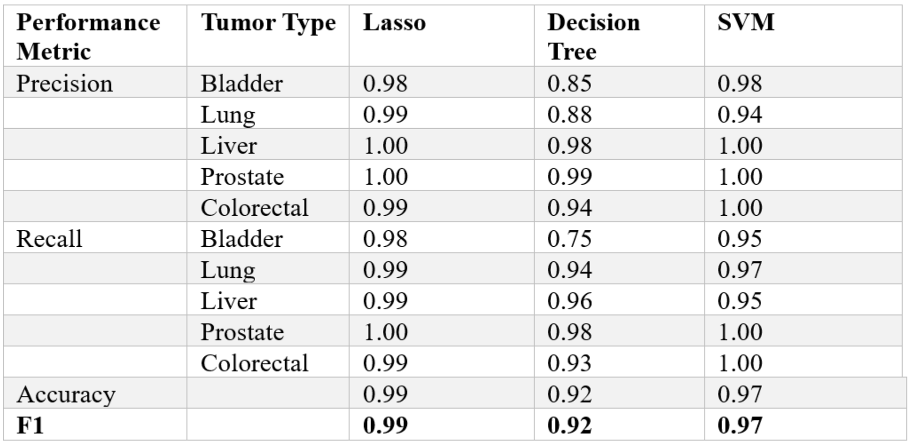


Figure 2: Confusion matrix heatmap for all three models. Left matrix shows the outcome for the Lasso model. The middle matrix is for the Decision Tree model. The right matrix is the SVM model.



Performance Metric	Tumor Type	Lasso	Decision Tree	SVM
Precision	Bladder	0.98	0.85	0.98
	Lung	0.99	0.88	0.94
	Liver	1.00	0.98	1.00
	Prostate	1.00	0.99	1.00
	Colorectal	0.99	0.94	1.00
Recall	Bladder	0.98	0.75	0.95
	Lung	0.99	0.94	0.97
	Liver	0.99	0.96	0.95
	Prostate	1.00	0.98	1.00
	Colorectal	0.99	0.93	1.00
Accuracy		0.99	0.92	0.97
F1		**0.99**	**0.92**	**0.97**

Figure 3: Table shows the predictive performance of the multiclass models.

most meaningful genes upon which its classifications were made; these top 30 genes are shown in Fig 4.

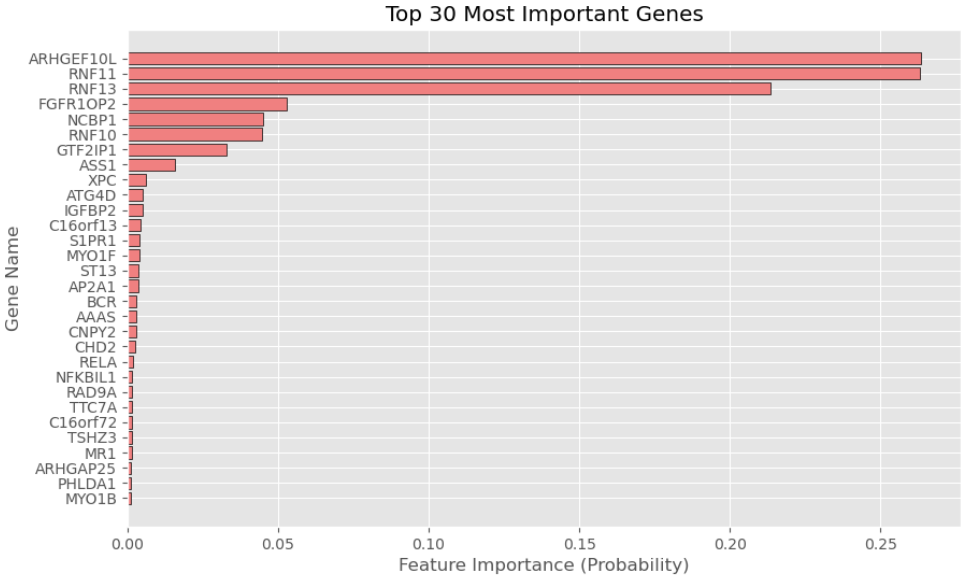


Figure 4: Bar chart shows which genes were most likely to predict a classification outcome.

These genes may be further investigated by comparing expression levels in cancerous tissue versus normal tissue - this could verify a relationship between certain genes being responsible for certain cancers. Identifying such a relationship has been the focus of cancer research, as it can provide target antigen sites for drug development such as personalized mRNA vaccines [8].

5. Conclusion

In this project, a multiclass ML approach was taken in order to classify the most common cancer types diagnosed in men by gene expression profiles gathered through RNA sequencing. TCGA datasets imported from UCSC Xena were selected to examine bladder, lung, liver, prostate and colorectal cancer genomic matrices. Data labels were generated by tumor type, and subsequently the datasets underwent stratified sampling to generate a 70/30 training to testing data split.

Due to the large 42.5 million feature space, dimension reduction methods were employed to extract the most meaningful gene expression values from the dataset. It was assumed that low expression genes could be filtered. A threshold value of 0.15 and PCA was employed to reduce the feature size to 1.4 million. Furthermore, three models were trained in mind of the remaining large feature space to prevent overfitting. First, a Lasso regression model was trained. Then, a decision tree model was trained with a low depth (=8) parameter. Finally, an SVM model was trained to examine for non-linear patterns. The Lasso model scored the best with an F1 score of **0.99**, being able to classify almost every cancer type correctly. It was found that the genes 'ARHGEF10L', 'RNF11' and 'RNF13' were most likely to predict an outcome.

Gene expression levels of the listed genes must be further investigated to verify the relation between gene expression and the development of tumor cells, this can be done through the TNMplot tool [9]. Additional figures listed in the appendix can be used as starting points for future research.

References

[1] G. M. Cooper, *The Cell: A Molecular Approach*, 2nd ed. Sunderland, MA: Sinauer Associates, 2000. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK9894/

[2] Office for National Statistics. (2017) Cancer registration statistics, england: 2017. [Online]. Available: https://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/conditionsanddiseases/bulletins/cancerregistrationstatisticsengland/2017

[3] Xena Functional Genomics Explorer. (Accessed 2024) Xena functional genomics explorer. [Online]. Available: https://xenabrowser.net/datapages/

[4] A. Gholamy, V. Kreinovich, and O. Kosheleva, "Why 70/30 or 80/20 relation between training and testing sets: A pedagogical explanation," 2018.

[5] M. Li, Q. Sun, and X. Wang, "Transcriptional landscape of human cancers," *Oncotarget*, vol. 8, no. 21, pp. 34 534–34 551, May 2017.

[6] M. Mohammed, H. Mwambi, and I. e. a. Mboya, "A stacking ensemble deep learning approach to cancer type classification based on TCGA data," *Sci Rep*, vol. 11, no. 1, p. 15626, 2021. [Online]. Available: https://doi.org/10.1038/s41598-021-95128-x

[7] S. Huang, N. Cai, P. P. Pacheco, S. Narrandes, Y. Wang, and W. Xu, "Applications of support vector machine (svm) learning in cancer genomics," *Cancer Genomics & Proteomics*, vol. 15, no. 1, pp. 41–51, 2018.

[8] N. Xie, G. Shen, W. Gao *et al.*, "Neoantigens: promising targets for cancer therapy," *Signal Transduction and Targeted Therapy*, vol. 8, p. 9, 2023. [Online]. Available: https://doi.org/10.1038/s41392-022-01270-x

[9] A. Bartha and B. Gyorffy, "TNMplot.com: A Web Tool for the Comparison of Gene Expression in Normal, Tumor and Metastatic Tissues," *International Journal of Molecular Sciences*, vol. 22, p. 2622, 2021.

A. Gene Expression Distributions

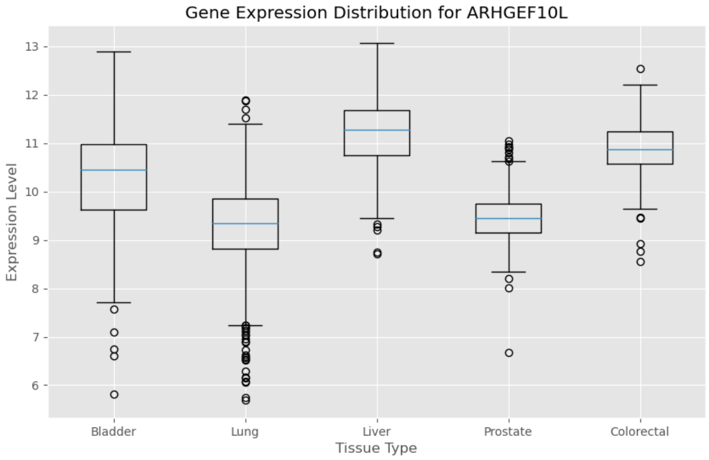


Figure 5: Box plot of gene expression distribution of gene ’ARHGEF10L’ in 5 tumor types.

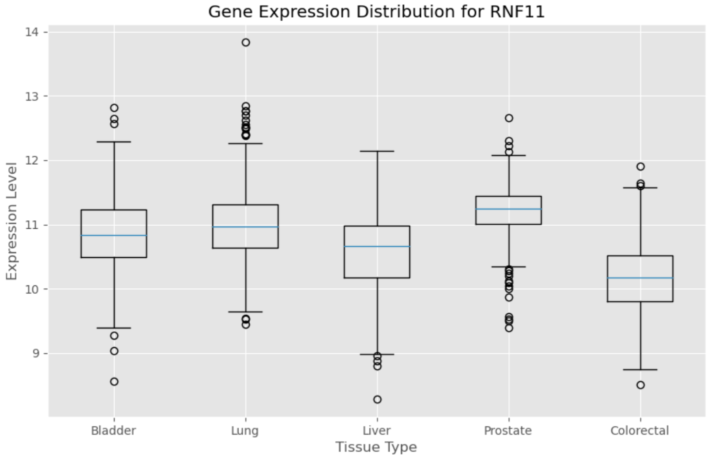


Figure 6: Box plot of gene expression distribution of gene ’RNF11’ in 5 tumor types.

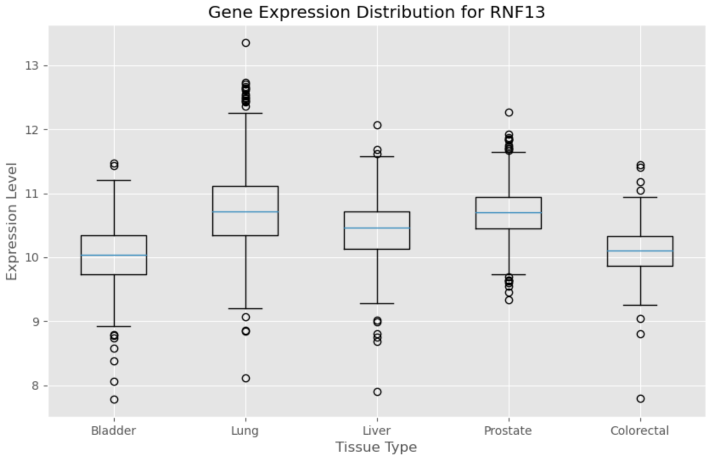


Figure 7: Box plot of gene expression distribution of gene ’RNF13’ in 5 tumor types.

B. K-means Clustering

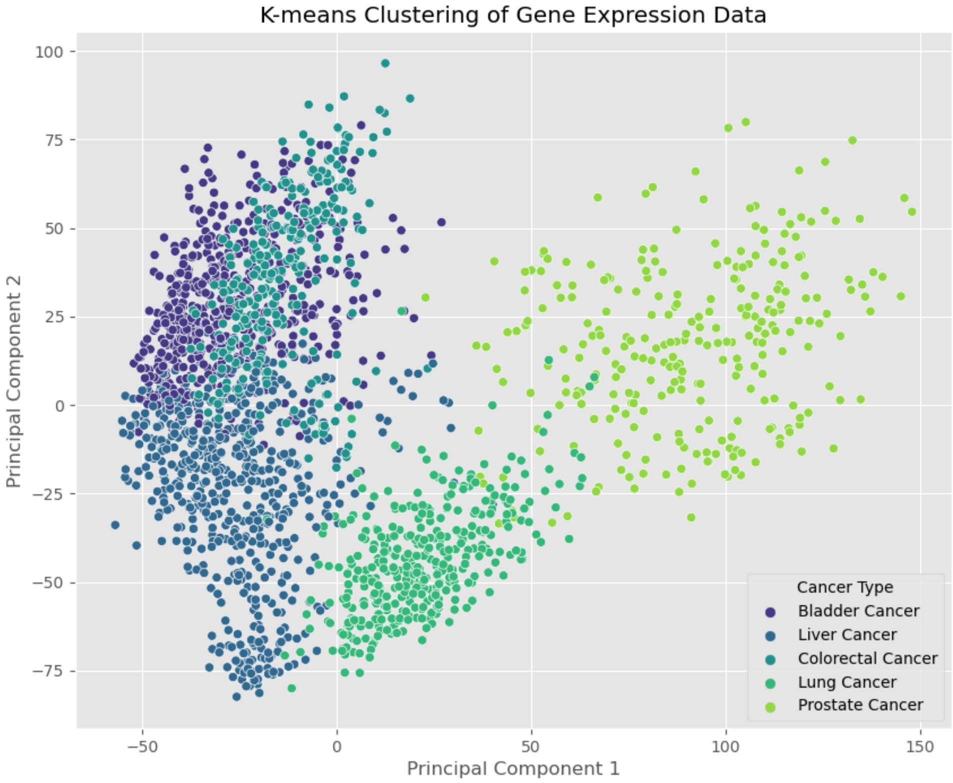


Figure 8: K-means clustering of different cancer types. It appears that bladder, liver, colorectal and lung cancers are heavily impacted by similar genes - whereas prostate cancer appears the least clustered. Gene distributions should be further investigated and compared to expression levels in normal tissue in order to locate such genes.