***Determining Biomarkers of Human Prostate Cancer Tumors***

**ABSTRACT**

Our investigation consists primarily of evaluating a set of tumor and non tumor tissues associated with prostate cancer in an individual. Our group utilized several bioinformatics tools and analyses in order to reach a deeper understanding of how prostate cancer may manifest itself genetically in an individual in addition to finding specific biomarkers that may prove beneficial for the purpose of providing greater care for those affected. Furthermore, the samples utilized varied by their proximity to the main tumorous site, in addition to their presence of cancerous tissue. Locating and understanding the genes that are associated with the cancerous tissue would allow future investigations to better treat the progression and manifestation of prostate cancer in humans. Due to the lack of understanding surrounding prostate cancer proliferation throughout the body and the weak genetic documentation of specific genes regarding its presence in humans, our group decided to expand in this direction of research.

Our main finding throughout the experiment was the presence of two strongly differentially expressed genes in contrast to the rest of the expressed samples, and the major presence of a marker known as COL1A1 which is known to promote metastasization of certain cancers, not including prostate cancer. This is an interesting point to consider as similar investigations may proceed in the future in regards to better understanding the genetic bases of prostate cancer in humans. The use of differential gene expression analysis and unsupervised gene clustering in our investigation helped produce a group of genes of interest that provide a good direction for future research into the field of cancer genomics and bioinformatics. Furthermore, the use of a wide set of analyses in our methods helps ensure that we are able to obtain and evaluate as much data as possible in order to better understand cancer through the use of bioinformatics. The overall design of our project can be considered an effective way of approaching preliminary investigations of biomarkers for an array of tissue samples across an array of pathologies. The aforementioned breadth of analyses help narrow down many of the more significant areas that should be investigated for further understanding and eventually treating diseases such as cancer.

**INTRODUCTION**

Our scientific question was to see if we could identify prostate cancer in humans based on analysis of a tissue sample. Our first method to answer this question was through differential analysis of gene expression ([Yuanli Zuo et al,2019](https://www.frontiersin.org/articles/10.3389/fgene.2019.01018/full); [Frank C. Cackowski et al,2019](https://onlinelibrary.wiley.com/doi/abs/10.1002/pros.23896?casa_token=ZORFiXjQb50AAAAA:IPvZTSsoVoz7FFV8wB_7W2r3vT5TgWV6qO6Oa8hlthcjtZSlJN540lzg8Gti-qG6m7pfvKbF5Y0kh3_3); [Xuechao Wan et al, 2016](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5312399/); [John R Prensner et al, 2011](https://www.nature.com/articles/nbt.1914?message-global=remove&page=9)) . We compiled the data into a counts matrix where each column was a sample and each row held the number of occurrences of a specific gene in that sample. We then conducted a differential analysis to see which genes were statistically significant in distinguishing between a tumor sample and a non-tumor sample. In addition, we used different methods of enrichment analysis to see if enriched genes could determine the type of tissue sample. We also used a variety of clustering methods to see if we could determine if tumor samples were similar to each other in a way we could analyze. We tried to answer our question through gene expression differential analysis such as RNA sequencing ([Heike Theimeyer et al, 2021](https://www.mdpi.com/1422-0067/22/21/11481); [Anastasia S. Nikitina et al, 2017](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5464844/); [Shancheng Ren et al, 2012](https://www.nature.com/articles/cr201230); [Gordon M. Cann et al, 2012](https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0049144)) and high throughput sequencing ([W.-S. CHENG et al, 2014](http://europeanreview.org/wp/wp-content/uploads/3504-3510.pdf)) to try to provide ourselves with better insight on the ability to identify tumor tissues based on certain genes. We also tried to compare our results with different RNA-SEQ researches on different cancers to see if our results were similar to previous findings ([Binsheng He et al, 2020](https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0049144)).

Ultimately, several of our analyses were fruitful in providings significant biomarkers related to the identification of prostate cancer in human tissue samples. However, their exact applications across many other samples were limited in consistency, as observed through our clustering analyses and statistical calculations. This implies that although we cannot broadly apply these biomarkers to all prostate cancers in regard to identifying those cells/genes, we can focus our analyses further on the significant biomarkers that were exposed. The information that was gathered allows us a more clear understanding of where to take the analyses of these biomarkers in the future, in addition to being able to craft a thorough approach to the bioinformatic analyses of cancer tissues.

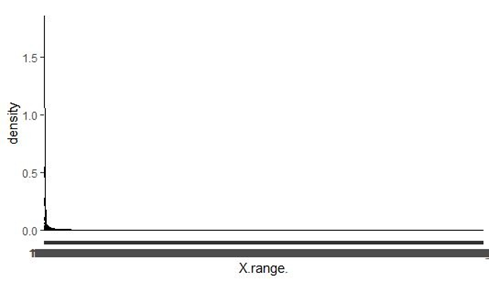
**METHODS**

All of our code and analyses are available for public view/use at our Github repository:<https://github.com/SebastianTirado/Bioinformatics-Group>

Our GEO data came as readings from patients with prostate cancer as well as control patients who did not have cancer. It was formatted as a collection of csv files which recorded if certain strings of amino acids were present within a gene in a sample. Samples were divided into 4 categories: tumor, involved, distant, and benign. Tumor, involved, and distant samples were all taken from patients with prostate cancer with different levels of distance to the tumor tissue. Benign tissues were taken from patients without cancer to act as non-malignant control samples. There were about 35,000 genes that were recorded in this data.

DIFFERENTIAL EXPRESSION:

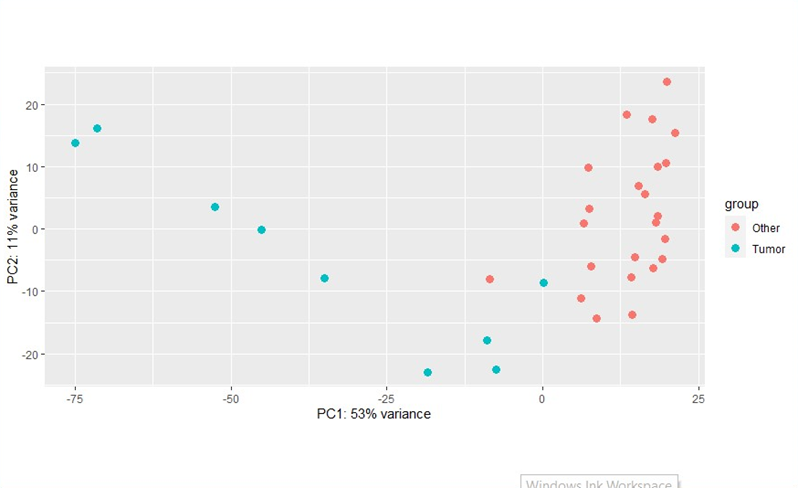
In order to perform analysis on our data, we first had to organize the data that we had into a format that we could utilize. We did this by combining all 32 of the samples we had into an expression matrix that displayed the total counts of all genes of interest within the samples. Once we loaded the data into R, we generated a density plot to display per-gene expression ranges.



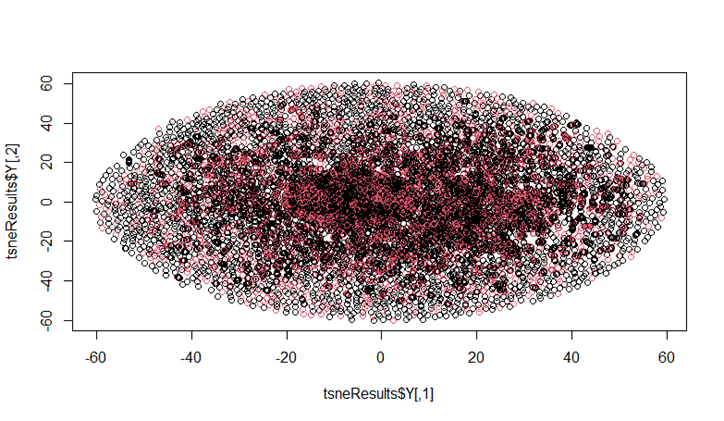
The density plot was made by subtracting the lowest count value for each gene by the highest count for each gene. The density plot of the gene expression ranges has an extremely long tail. This means that while most genes have a relatively narrow band of count values, some genes are found thousands of times in some samples while almost none in others.

Once our data was organized, the next step was to perform differential expression analysis on our samples. We performed differential expression analysis on our tissue samples through the use of the DESeq2 package in R. Using this package, we formatted our data into a DESeqDataSet and performed the analysis using the DESeq() wrapper function. ​

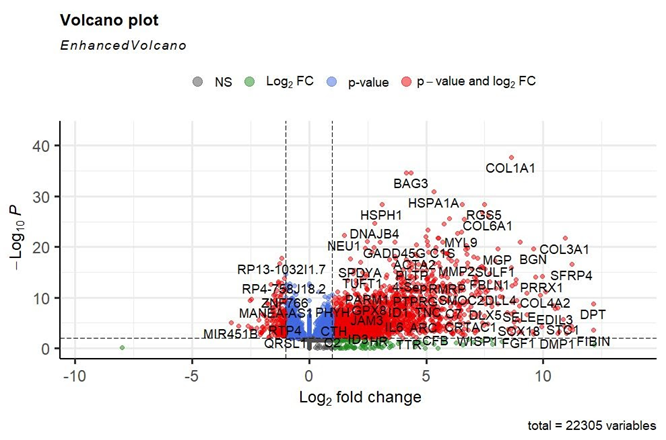
To visualize the results of the differential expression analysis, we generate a series of plots:



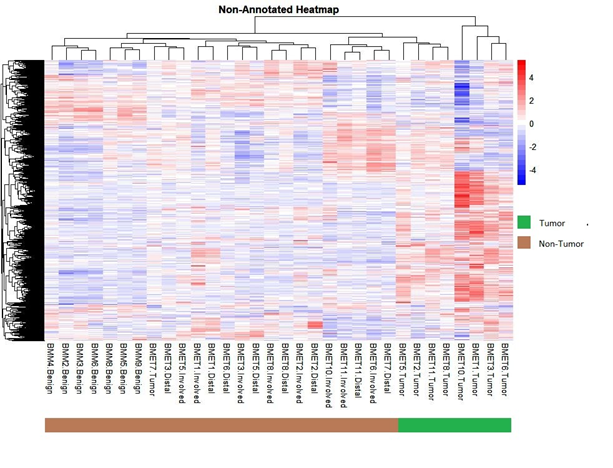
The first plot shown above is the PCA plot and was created using the DESeq2 object on our labelled sample data. The PCA plot uses principal components to show patterns in the data. As can be seen in this plot, there is a separation between the Tumor and non-Tumor groups, which implies there is a difference in gene expression between them.



Next, we created this t-SNE plot using the Rtsne library and running our expression matrix through the associated Rtsne function. After plotting the data and separating by tumor vs. non-tumor groups, we found that the tumor samples were highly clustered near the center while the non-tumor groups trended to the outskirts of the graph. This graph is similar to the PCA-plot in that it shows that the tumor samples tend to be clustered together (in the center of the t-SNE plot and to the right in the PCA plot). However, they differ in that the t-SNE plot uses significantly more data points in its calculation. Also, the shapes of the plots are drastically different in that the t-SNE plot is very rounded while the PCA plot has more of a parabolic shape.



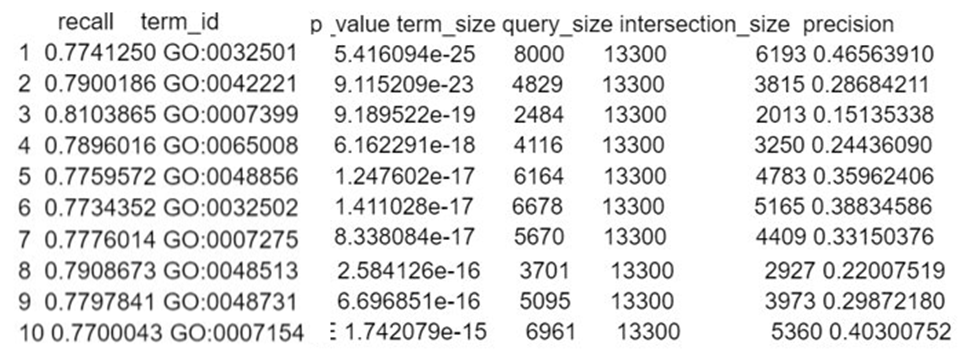
The volcano plot was created using the DESeq2 library to run a differential analysis between the tumor and non-tumor groups and EnhancedVolcano to create the plot. It shows a comparison of the statistical significance of the genes expressed in the groups to show which ones are most important in distinguishing the groups. After playing around with the pCutoff value, we found that the value recommended by the article gave the most intelligible results. An interesting result is that the gene COL1A1 seems to have a very high significance in the tumor tissues. A quick Google search shows that a study found this gene to be highly correlated with breast cancer metastasization.



Finally, we created a heatmap of our data using the DESeq2 module and the pheatmap library, which shows the expression levels for different genes for the different samples. The only parameters we used were to cluster the like groups together so all the tumor samples appear next to each other to make them easier to compare. The plot shows some clear differences in the gene expression in the two groups, although which exact genes are not clear due to the number of genes present.

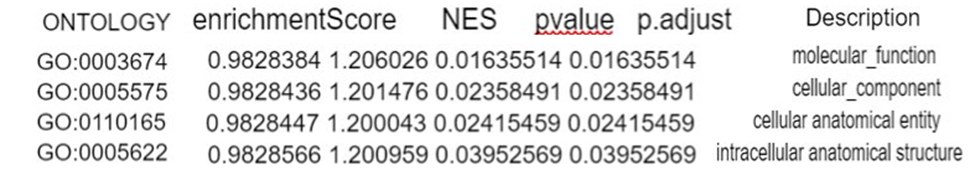
GENE SET ENRICHMENT ANALYSIS:

To further understand our data, we perform both Gene Set Enrichment Analysis (GSEA) and Disease Ontology Enrichment Analysis using a variety of methods. The first method we use to perform such analysis is the gProfiler2 package in R:



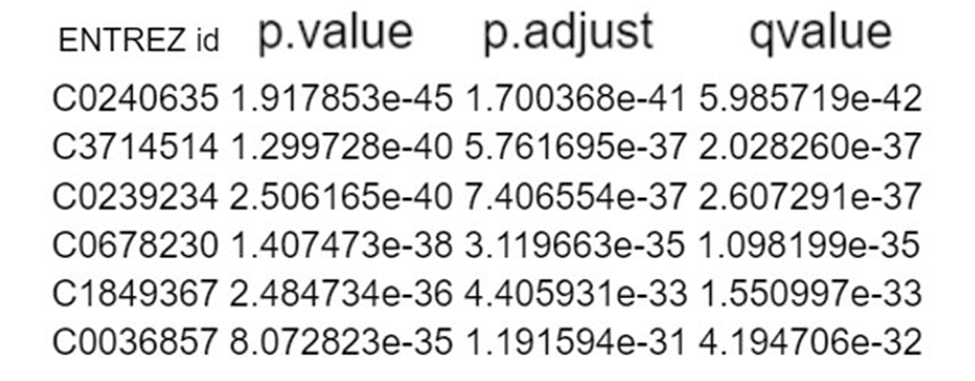
Using GProfiler2 (through the gost() function), we generated the table above which shows the 10 most statistically significantly enriched gene ontology terms and some statistical data to go along with the query.​ This shows that there are many genes which are extremely statistically significant, with a p-value with 19+ decimal places.

We then performed similar analysis using the ClusterProfiler R package:



Using ClusterProfiler (through the gseGO() function), we generated the above table which describes four sets of gene clusters and their corresponding descriptors: molecular\_function, cellular\_component, cellular anatomical entity, and intracellular anatomical structure. ​ Also provided is each sets pvalue, adjusted pvalue, enrichment score, and normalized enrichment score (NES). Something interesting about this data is that each of the four sets had a notably high enrichment score of about 0.983, and also that each of the four sets of data produced an NES that was statistically significant with all four pvalues falling below 0.05.

Next, we performed Disease Ontology Enrichment Analysis using the ClusterProfiler R package:

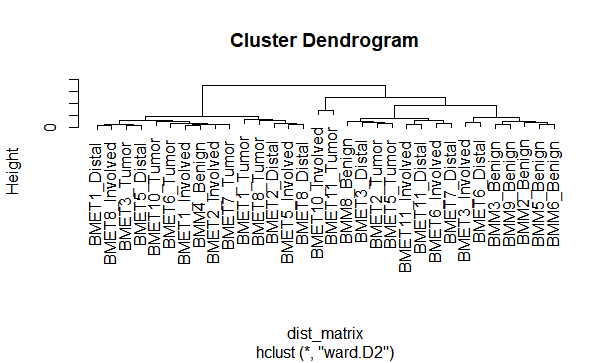


Using ClusterProfiler (through the enrichDGN() function), we generated the table above which shows the 6 most statistically significantly enriched genes and some statistical data to go along with the query.​ This process, again, shows that there are many genes which are extremely statistically significant, as we found that all 6 of these genes produced remarkably low p-values.​

CLUSTERING

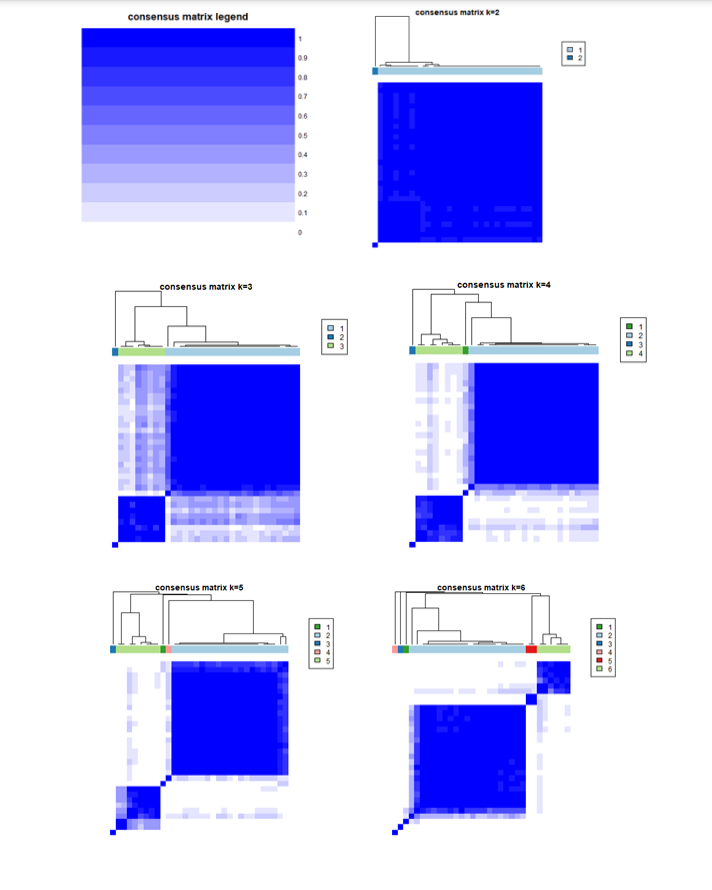
To define a more precise dataset for the clustering methods, lists of highly variable genes were necessary. We calculated the variance of each gene across the samples and ordered the gene table from most variable to least variable. Sublists of the 10, 100, 1000, 5000, and 10000 most variable genes were then created. This allowed us to analyze how the number of genes included in the clustering changed the final results. It is important to note that when running the clustering algorithms, we were dividing the samples into clusters. The goal was to see if comparing highly variable genes may show relationships between similar samples, specifically tumor and non-tumor samples.

The first method we used to analyze was the hclust R package. This package uses hierarchical clustering to create a dendrogram showing relationships between the samples. This package does not accept a k-value as input so some analysis is necessary to divide the final dendrogram into the final clusters. The dendrogram for the 1000 most variable genes is shown below.

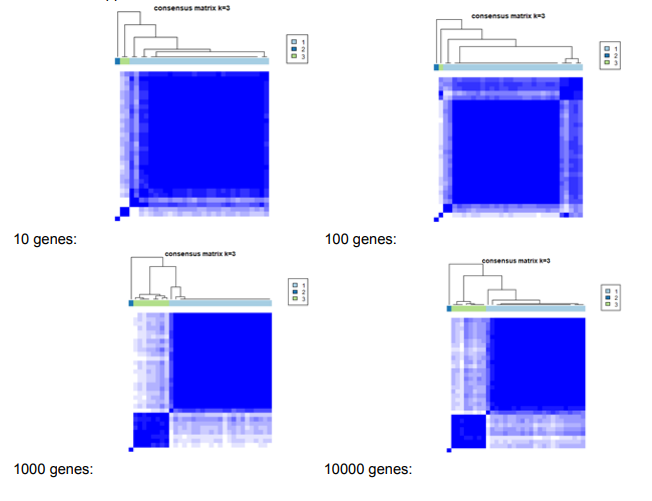


We decided 4 clusters to be the most appropriate k-value for this dendrogram. We also used the hclust package on the other sized lists of variable genes. We noticed that using the lists of size 10 and 100 have different clusterings, but any list of greater size has the same clustering. Therefore this is the clustering we used when conducting our statistical analysis later.

The next method we used was the Consensus Clustering Plus (CPP) package. What is unique about this clustering method is that a target k-value is not selected but rather a maximum reasonable k-value. The package will then perform clustering with k-values starting from 2 and increasing to the maximum. This allows analysis of the best k-value without separating the clusters ourselves like is the case with hclust. We decided that a k-value higher than 6 would not make sense for our data set and conducted the clustering with that value as the inputted maximum. The clusters for each k-value with 5000 genes are shown below.

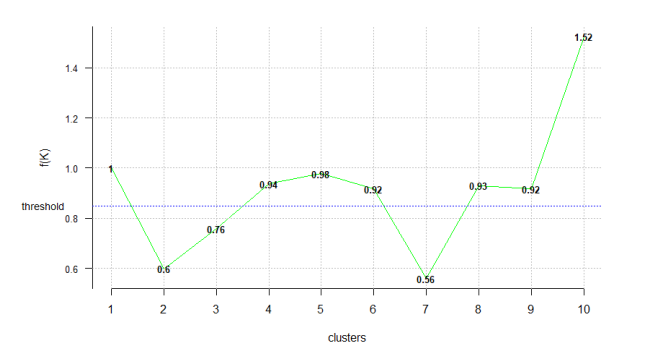


As seen above, the chosen k-value has a large effect on the distribution of the samples among the clusters. We found 3 to be the best k-value for this method. This is because when the k-value is smaller, one of the groups is so large that we learn little about the relationships between the samples. However, as the k-value increases the additional clusters are too small for us to be sure that the samples in them are actually alike. When the k-value is 3, we see the most informative distribution to analyze.

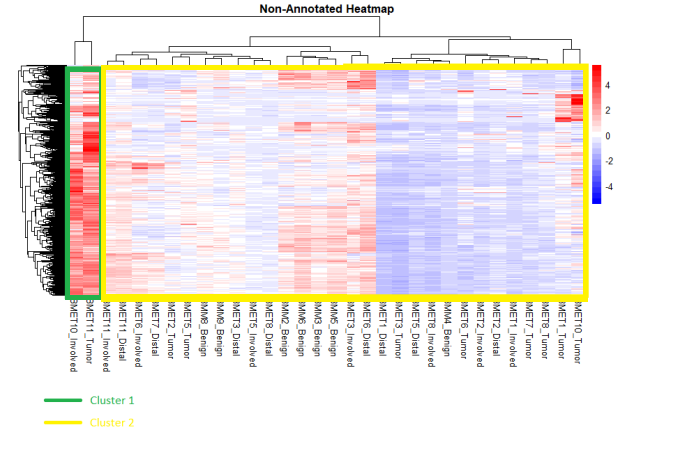


Additionally, we analyzed how the number of genes used in the clustering affects the results. As shown above, the results are similar to those we found when using hclust. With lists of 10 and 100 genes, the results are noisy and subpar. After a size of 1000 genes, the plots begin to look virtually identical. This suggests that using more genes is giving better results, or that after a certain point the genes do not occur enough to change the clustering results.

The next method we used was K Nearest Neighbors (KNN). This is a somewhat simpler method than those discussed previously that uses Euclidean distance to determine the clusters. This method requires the selection of a k, so first we had to determine which k-value or values would be best for our dataset. To do this, we used a pre-existing function f(K) that determines how well a k-value clusters the data. The lower this value is, the better the method was able to separate our samples into distinct groups.

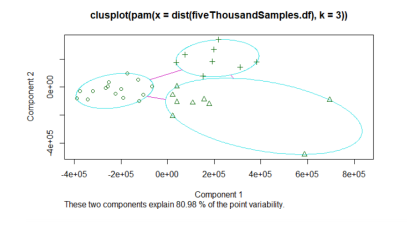


We used a standard threshold value and computed the f(K) value for the k-values from 1 to 10. As shown, the k-values of 2 and 7 were found to minimize the function best. While 7 actually had the best score, we decided it was not the ideal k to use. This was due to multiple factors. The first was the goal of our project. If we are working to use clustering to determine biomarkers of prostate tumor and non-tumor samples, 7 clusters is simply too many to extract meaningful data. In addition, we only had 4 types of samples (Tumor, Involved, Distal, and Benign), so using 7 clusters would likely extract features that are not especially relevant to the dataset.

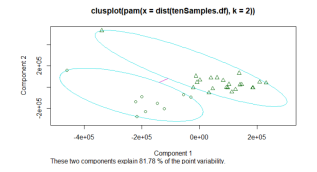


The heatmap with the KNN clustering is shown above. Using different amounts of genes was shown to have little effect on how the genes were split into their clusters, though there was some effect on the “heat” relationships for the genes. Visually, this data should that 3 or 4 may be valid k-values as well, but we chose to stay with the f(K) function’s results.

The final method we used was PAM clustering. This method is a more robust version of the k-means clustering algorithm, with a few improvements which generally give better results. PAM clustering requires the input of a k value so we first analyzed the plot to see which value might make the most sense.



As shown above, the principle component analysis displays the groups of samples based on two components. Despite the outlying data, we believe that a k-value of 3 works best for the data when using the list of 5000 genes. Using the different sized gene lists only made a difference in clustering for the list of size 10. As shown below, the plot looks quite different and a k-value of 2 is likely more appropriate to group the data.



STATISTICS

For each of the clustering methods, we created a 2xk table with each column representing a cluster and each row representing the number of tumor or non-tumor samples within that cluster. We then ran a chi-squared independence test on each table to determine if the clustering was statistically significant in separating the tumor and non-tumor samples from each other. This would allow us to examine whether highly variable genes are a useful biomarker for determining whether a tissue sample is a tumor.

After finding a p-value for each method, we needed to correct for the fact we conducted multiple tests, which increases the likelihood of Type I errors. We used the Holm-Bonferroni method and obtained new adjusted p-values for each method, which is what we examined to determine statistical significance.

**RESULTS**

Through the differential analysis, we found many genes that were extremely statistically significant when it came to determining if a sample was tumor tissue or not. Further investigation into these genes, and perhaps creating some sort of predictive model with them could certainly find an effective way of classifying tissue samples. Our results told us which genes were significant but not how to use this information or why this occurs. One of the strongest markers was COL1A1 which has been found to promote metastitization in bowel and breast cancers, but is usually not as strongly linked to prostate cancer. This could be a further area of investigation, as its effects are clearly linked to multiple forms of cancer.

Our GSEA methods gave us a variety of useful results. We found over 100 gene ontologies that were significantly enriched, with the top 10 having extremely small p-values. This gives us useful information about the functions of the differentially expressed genes and how their expression may play a part in the development of the disease. These GO terms are the subject of study already so we could use previous research to understand why these terms are being expressed. In addition we found many genes that are very enriched during our disease ontology enrichment analysis. Some of the most significant genes have p-values with over 40 decimal places, so they clearly are significant.

Our clustering methods did not find any sort of statistical significance. None of our clustered methods were found to have a p-value of less than .05 using a chi-squared test of independence. This means that clustering the samples based on highly variable genes was not an effective way of classifying them as it did not achieve much better results than random chance. Something to note is that two of our samples were strong outliers when compared to the others. This came up very clearly when analyzing a heatmap of our samples and they had an expression of genes that was very similar to each other but totally different to the other samples. They are also clearly present in the PAM cluster plot where they are off by themselves and far away from the other samples using the principle components. This may have been a confounding factor in our results as it affected how the clustering methods grouped the other samples. In addition, the fact that they formed their own cluster may have affected the p-values of the chi-squared tests. Further examination as to exactly why these samples were so common would instruct how to treat them relative to the other data.

Overall, our project was neither a resounding success or a total failure. We did not confirm our initial hypothesis that we would be able to use the genomic data to determine whether a tissue sample comes from a prostate cancer tumor or not, but that does not mean the project failed. It just means that in the future other teams may want to modify our methods or dataset to give a higher likelihood of success. We were also able to determine specific differentially expressed genes, enriched gene ontology terms, and enriched genes for disease ontology. All of these are useful results for further research into the relationship between genomic biomarkers and prostate cancer and give plenty of information to conduct further investigation. Thus, despite not confirming our hypothesis we consider this project a general success.

There are a few weaknesses in our project. One is the fact that we attempted to find biomarkers to distinguish only between tumor and non-tumor samples. However, we actually had four different types of samples which could have added additional information into our analysis. While distal, involved, and benign samples are all not tumors, that does not mean their relationships to cancer are all the same. Therefore, it may have been beneficial to examine these samples in a broader context and attempt to glean more information about different types of tissues.

A bioethics issue that may be related to our project and its result would be the possibility of health insurance providers biasing against genetics as a pre-existing condition. There are concerns people may be denied or charged extra for healthcare if health insurance companies get access to their genetic information. This information may show they are more likely to develop certain diseases during their lives. If our project were to reveal information about which genes were involved with the development or metastitization of prostate cancer, we may contribute to this sort of issue. Unfortunately, there is not much we could do to avoid this issue. We simply must trust that the concrete benefits of better understanding the mechanics of a common and deadly form of cancer outweigh the possible negative effects.

If we were to continue working on this project we would wish to further investigate analysis with more focus on the types of samples beyond tumor and non-tumor. We believe this may be more effective at identifying biomarkers. We would also investigate using different genes beyond the most variable, as perhaps other types of genes may play a larger role in cancer than we believed. In addition, we would propose incorporating a much more diverse and larger sample size of data to be involved. This would be for the purpose of not only aiding in the statistical significance of any findings, but also for providing a more uniform representation of humanity as it pertains to cancer diagnoses. Our sample size for this project was relatively small and therefore there may have been statistical noise that affected our results.

**CONCLUSION**

In conclusion we found there were several differentially expressed genes and enriched gene ontologies that distinguished tumor and non-tumor samples for prostate cancer. Although our results are not definitive, and the sample size is rather small (around 40 samples), we can conclude that this area of research is of high importance in fighting a particularly deadly cancer.

For future work, we found that there are many differentially expressed genes that should be further analyzed in terms of their relationship to the metatistization of prostate cancer. Specifically, we found that there are a few biomarkers that may suggest the presence of prostate cancer tumor tissue.

**REFERENCES**

* Yuanli Zuo et al, Transcriptome Analysis identifying Piwi-interacting RNAs as Prognostic Markers for Recurrence of Prostate Cancer, 2019.
* Heike Theimeyer et al, An RNA-Seq-based Framework for Characterizing Canine Prostate Cancer and Prioritizing Clinically Relevant BioMarker Candidate Genes, 2021
* Frank C. Cackowski et al, Detection and isolation of disseminated tumor cells in bone marrow of patients with clinically localized prostate cancer, 2019
* Binsheng He et al, A Neural Network Framework for Predicting the Tissue-of-Origin of 15 Common Cancer Types Based on RNA-Seq Data,2020
* Xuechao Wan et al, Identification of androgen-responsive lncRNAs as diagnostic and prognostic markers for prostate cancer, 2016
* W.-S. CHENG et al, Both genes and lncRNAs can be used as biomarkers of prostate cancer by using high throughput sequencing data, 2014
* Anastasia S. Nikitina et al, Novel RNA biomarkers of prostate cancer revealed by RNA-seq analysis of formalin-fixed samples obtained from Russian patients, 2017
* John R Prensner et al, Transcriptome sequencing across a prostate cancer cohort identifies *PCAT-1*, an unannotated lincRNA implicated in disease progression, 2011
* Shancheng Ren et al, RNA-seq analysis of prostate cancer in the Chinese population identifies recurrent gene fusions, cancer-associated long noncoding RNAs and aberrant alternative splicings,2012
* Gordon M. Cann et al, mRNA-Seq of Single Prostate Cancer Circulating Tumor Cells Reveals Recapitulation of Gene Expression and Pathways Found in Prostate Cancer, 2012

**References:**

1. <https://www.frontiersin.org/articles/10.3389/fgene.2019.01018/full>
2. <https://www.mdpi.com/1422-0067/22/21/11481>
3. <https://onlinelibrary.wiley.com/doi/abs/10.1002/pros.23896?casa_token=ZORFiXjQb50AAAAA:IPvZTSsoVoz7FFV8wB_7W2r3vT5TgWV6qO6Oa8hlthcjtZSlJN540lzg8Gti-qG6m7pfvKbF5Y0kh3_3>
4. <https://www.frontiersin.org/articles/10.3389/fbioe.2020.00737/full?report=reader>
5. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5312399/>
6. <http://europeanreview.org/wp/wp-content/uploads/3504-3510.pdf>
7. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5464844/>
8. <https://www.nature.com/articles/nbt.1914?message-global=remove&page=9>
9. <https://www.nature.com/articles/cr201230>
10. <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0049144>