JAAHM
Journal of Advanced Analytics in Healthcare Management

Deciphering Gene Expression Patterns to Differentiate Among Leading Cancer Types

Arturo Chavez, Dimitris Koutentakis
Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology

Youzhi Liang,
Department of Mechanical Engineering, Massachusetts Institute of Technology

Sonali Tripathy,
Sloan School of Management, Massachusetts Institute of Technology

Jie Yun
Department of Civil and Environmental Engineering, Massachusetts Institute of Technology

Submitted: 13 June 2019
Accepted: 27 November 2019
Published: 05 January 2020

Abstract

While individual cancers have been extensively researched in terms of prognostic genes, comprehensive studies comparing these across different cancer types remain scarce. Proper cancer classification into subtypes is pivotal for accurate diagnosis and effective treatment strategies. This study delves into gene co-expression networks across five cancer types using patient-to-patient correlation network analysis and Weighted Gene Correlation Network Analysis (WGCNA), utilizing data from UC Irvine. We conduct a thorough comparison of network characteristics such as degree, centrality, and betweenness for each cancer type. Additionally, we employ multinomial logistic regression to pinpoint a crucial subset of genes. Our research provides insights into the unique and overlapping gene expression patterns among various cancer types.

I. INTRODUCTION

Cancer describes a collection of diseases that share some common characteristics, particularly unregulated cell growth. They also vary widely in terms of mortality rate, treatment options, and prevalence in the population. Accurate diagnosis of cancer type is essential to decide treatment options, therapy and prognoses. However, some cancers are difficult to distinguish based on a single test[1]. Gene expression microarray technology provides precise information for cancer prognosis and treatment and has been used to categorize cancers into subgroups[2]. Current classification methods include nearest prototype classifier by defining subset of genes that best characterize each class[3], supervised classification algorithms to identify gene expression signature, and the use of combined algorithms[4].

These methods have experienced moderate success, so clearly the methods are identifying relevant statistical differences in the tumor types in order to classify them correctly. Digging one level deeper, we are interested to explore the statistical differences between tumors, tying them to phenotype differences in disease outcomes. Using this data-driven approach, we aim to understand the variation within tumors of the same type as well as the consistent differentiating features that distinguish each tumor type.

Tumor Classification Background

Khan et al.1 explore the use of neural network classification models to classify cancer subtypes taking cDNA expression data as the input. Their analysis was specific to small blue-cell tumors (SBCTs) which can be further classified into neuroblastoma (NB), rhabdomyosarcoma (RMS), non-Hodgkin lymphoma (NHL) and the Ewing family of tumors (EWS). Correct classification into subtypes is critical to selecting the optimal course of treatment. Usual methods for tumor classification often use spectroscopy, but SBCTs are challenging to classify visually. There have been many attempts to use geneexpression data to aid in classification, but so far none have been proven to be effective in identifying cancers that belong to several categories.

They started with a panel of 6567 genes from which to find meaningful features. In order to make the dimensionality of the data more manageable, they eliminated genes that had expression levels below a threshold. With the remaining 2308 genes, they performed PCA to further reduce the dimensionality, taking the largest 10 components which accounted for 63% of the variation. After training on these features, their model was able to fit all of the 63 samples from their training set. To identify the most important genes, the authors altered each of the locations to measure the overall classifications sensitivity to that gene. After identifying the most important genes, the authors performed multidimensional scaling (MDS) to visualize the clear separation between cancers.

When they tested the models ability to classify new samples, they were pleased to be able to classify all the cancer types correctly. Unfortunately, they were unable to reach the level of 95% confidence level in the diagnosis that they were targeting. This highlights the challenge of using machine learning methods in the medical field, since clinical use needs highly reliable AI systems. This study motivates further work of this kind with other disease types and larger data sets. Also, this result of a reasonably successful classification method motivates our analysis of the statistical properties of the different tumor gene expression profiles, from which these classifiers form decision boundaries.

Network Methods Background

Specifying features in genetics is a challenge because there are often complicated interactions between genes. To understand these relationships researchers have used network models. Juan A. Botia et al.5 analyzed 1126 genes relating to 25 subtypes of Mendelian neurological disease defined by Genomics England (March 2017) together with 154 gene-specific features capturing genetic variation, gene structure and tissue-specific expression and co-expression. He developed a technique to identify the gene mutations that can lead to a neurological disorders. Random samples were selected with no disease association to develop decision tree models for each subtype. Within the disorder subtypes, network models were used to improve the predictive power.

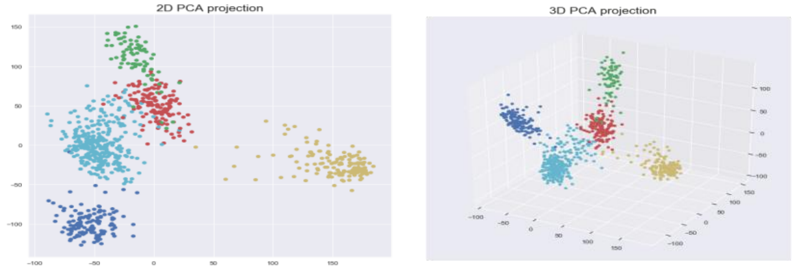
Another instance of network approaches in genomes, Yang et al.6 applied the weighted correlation network analysis (WGCNA) method to construct a gene coexpression network. In this study, they primarily investigated the prognostic genes that distinguish between cancers. They investigated these genes with three distinct levels of depth properties: specific genes, gene modules, and the system holistically. At the gene level, they found that network properties could distinguish prognostic genes from other genes. More specifically, using Fisher's exact test, they were able to conclude that prognostic genes tend not be hubs in the co-expression network. On the gene modules level, they discovered that prognostic genes are enriched significantly. Third, on the system level, some prognostic modules are conserved across tumour types.

II. PRELIMINARY METHODS

Dataset Description

The dataset is provided by University of California at Irvine and is located here. The data includes 801 samples, each with 20,532 gene positions. Each sample vector contains the RNA-Seq gene expression levels. Each sample in the dataset corresponds to a particular tumor type. Every sample is one of five types: breast invasive carcinoma (BRCA), kidney renal clear cell carcinoma (KIRC), colon adenocarcinoma (COAD), lung adenocarcinoma (LUAD), and prostate adenocarcinoma (PRAD).

Understanding the data



(a) 2D PCA data projection (b) 3D PCA data projection

A preliminary analysis of our data is shown in Fig. 1a. Not unusual in the genomics setting, we run into the curse of dimensionality, making our 801×20,532-dimensional matrix difficult to visualize. We performed Principal Component Analysis on our data set and plotted the projection of our data on the 2 principal components with the largest corresponding eigenvalues in Figure 1a, and the projection of our data on the 3 principal components with the larges eigenvalues in Figure 1b. Prior to performing the eigenvector decomposition, we preprocess our data by subtracting the column mean from the each entry. The result is the matrix X with dimension 801×20,532 with each column having mean 0. Taking the eigenvector decomposition we get X = V ΛV T, where Λ is a diagonal matrix of the eigenvalues (sorted such that the largest eigenvalue is in the top left) and V has the corresponding eigenvectors as its columns. Taking the first d columns of V, we get Td = XVd, where T has dimension n × d.

We compute T2 and T3 and plot the results. In both of the plots, each point was colored according to what type of cancer it represents. This further validates our intuitions that the each cancer type has particular features that distinguish it from the others.

Variance of Tumor Types in Reduced Dimension

We aim to understand how the various tumor types differ, both in statistical and phenotypic terms. The previous PCA results show that projecting the samples onto the first two or three principal components lead to a reasonably clean separation. Interestingly, some cancer types appear to be clustered more tightly together in this lower dimensional space, while others appear to be more loosely dispersed. Also, it is interesting to note which pairs of tumor types appear closer together in this space. To quantify both of these notions, we fit a Gaussian mixture model to the PCA-transformed points. Using T2 from the previous section, we fit a five Gaussian mixture that appears to closely approximate the true labeling of the points. Sampling from a Gaussian mixture can be thought of as a two step process. First, it involves sampling from a multinomial distribution with parameters π (similar to an unfair dice). The result of the first step determines which Gaussian to sample from in the second step. Therefore, the conditional probability of the coordinates of a sample, given it is a particular cancer type c, is distributed according to$\mathcal{N}(\mu_c, \sigma_c^2)$. The second step is simply to sample from that Gaussian. Gaussian mixture models are fit using the expectation-maximization (EM) algorithm, where the objective is to maximize the loglikelihood of generating the training data. Fitting this model results in the parameters π,μ,Σ for each type of cancer.

Assuming that this fit is reasonable, we can quantify the notions of homogeneity within a tumor type by inspecting the covariance matrix of the Gaussian corresponding to that cluster of samples. Because we are interested in the variance along the axis of the principal components, we constrain the Gaussians to be oriented along those axes (forcing the covariance matrices to be diagonal). We get the following results where the vector is ordered [LUAD, PRAD, KIRC, COAD, BRCA]:

$$\pi^T = \begin{pmatrix} 0.23 & 0.16 & 0.18 & 0.08 & 0.35 \end{pmatrix}$$

$$\mu_L = \begin{pmatrix} -1.60 \\ 51.24 \end{pmatrix}, \mu_P \begin{pmatrix} -54.91 \\ -100.11 \end{pmatrix}, \mu_K = \begin{pmatrix} 151.45 \\ -23.29 \end{pmatrix}$$

$$\mu_C = \begin{pmatrix} -19.50 \\ 120.05 \end{pmatrix}, \mu_B = \begin{pmatrix} -47.47 \\ -1.81 \end{pmatrix}$$

$$\Sigma_L = \text{Diag}\begin{pmatrix} 233.83 \\ 451.96 \end{pmatrix}, \Sigma_P = \text{Diag}\begin{pmatrix} 186.51 \\ 160.41 \end{pmatrix},$$

$$\Sigma_K = \text{Diag}\begin{pmatrix} 594.35 \\ 235.63 \end{pmatrix}, \Sigma_C = \text{Diag}\begin{pmatrix} 92.20 \\ 224.27 \end{pmatrix},$$

$$\Sigma_B = \text{Diag}\begin{pmatrix} 201.54 \\ 536.06 \end{pmatrix}.$$

The model was able to fit the data reasonably well, which can be seen in Figure 2. For example, the weights π are nearly the same distribution as the true labels which are: 0.$18\ \ 0.17\ \ 0.18\ \ 0.10\ \ 0.37)$. Interestingly, we see that

KIRC has the largest variance in the first principal component and BRCA has the largest variance in the second principal component. Overall it appears that KIRC is the most dispersed cancer type, since it has largest variance on average over the 2-D space. This can be evidenced in Figure 2, seeing that there are several samples that are more than 2 standard deviations away from the cluster mean. COAD and PRAD are the most concentrated, suggesting that those samples are more homogeneous.

Note, that our model is not a perfect fit. Comparing our fitted GMM model to the 2-D projection of the data in Figure 1a, we see that some COAD samples are found within the LUAD cluster. This GMM model does not explain this feature of the data, motivating the use of other methods to understand the gene expression profiles of these cancer types.

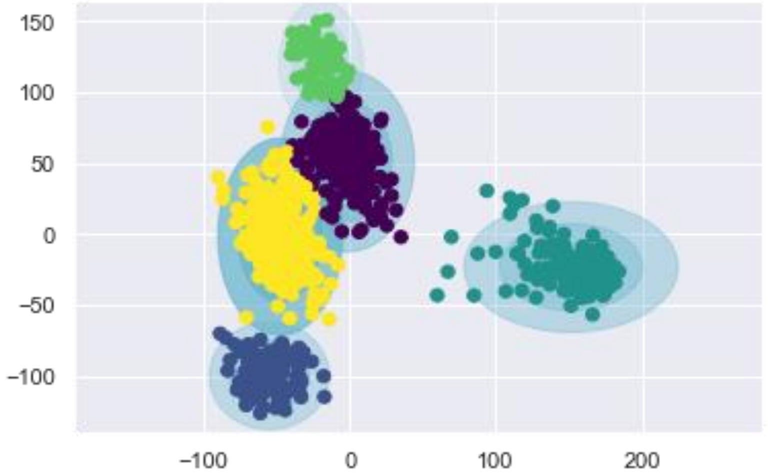


FIG. 2: Above is the visualization of the gaussian mixture model fit to explain the distribution of the samples in the 2 dimensional space specified by the first principal components. PRAD is blue, LUAD is purple, BRCA is yellow, KIRC is teal, and COAD is green.

Stochastic Neighbor Embedding

In addition to PCA, we perform t-distributed stochastic neighbor embedding to our data. tSNE is a probabilistic approach to place objects from high-dimensional space into low-dimensional space so as to preserve the identity of the neighbors. Prior to tSNE, stochastic neighbor embedding (SNE) was proposed, which used the same general approach by placed a Gaussian on each object in high-dimensional space. This resulted in the "crowding problem" where many points would be mapped together in the center. To overcome this problem, Hinton et al.7 proposed tSNE which has larger tails and a steeper drop moving away from the mean (within close range). Both methods are fit by minimizing the KL divergence between the low and high dimensional probabilities of picking a particular neighbor. Intuitively, this method keeps "nearby" points in high dimension close to each other in low dimensional space, while keeping separated points relatively far apart in the low dimensional space. In this case tSNE is able to separate the tumor types with high precision (notably better than PCA). This result supports our intuition that different cancer types are statistically distinguishable. For the rest of the paper, we aim to characterize those statistical differences more precisely.

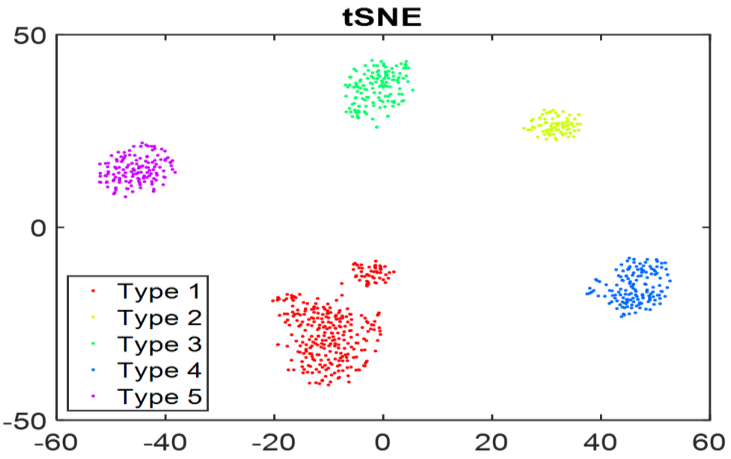


FIG. 3: tSNE on our data gives the following wellseparated clusters

Hierarchical Clustering

Hierarchical clustering of gene expression is a popular mechanism to cluster genes with similar expression patterns together. This clustering mechanism involves calculation of distance between two gene vectors to find the similarity between them. The dendrogram was sliced at a height of 370 to find five clusters in particular. Figure 4 demonstrates the samples clustered into five clusters where each cancer type is majorly clustered into just one cluster. There seems to be one of the clusters that consists of more than one cancer type, signifying that in some patients the distance between the gene vectors is close enough. These cancer types are BRCA, LUAD and COAD. From figure 2 also it could be seen that these three cancer types are close to each other.

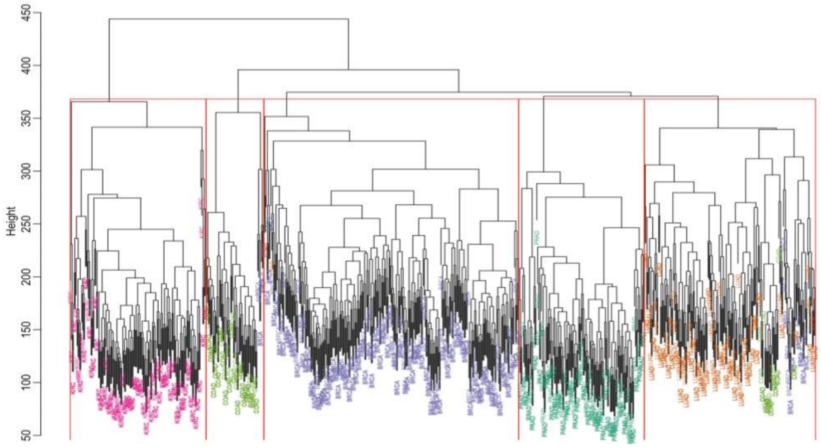


FIG. 4: Hierarchical clustering of the gene data set where pink is KIRC, green is COAD, purple is BRCA, teal is PRAD, and orange is LUAD.

III. NETWORK ANALYSIS

A. Patient-to-Patient Correlation Network

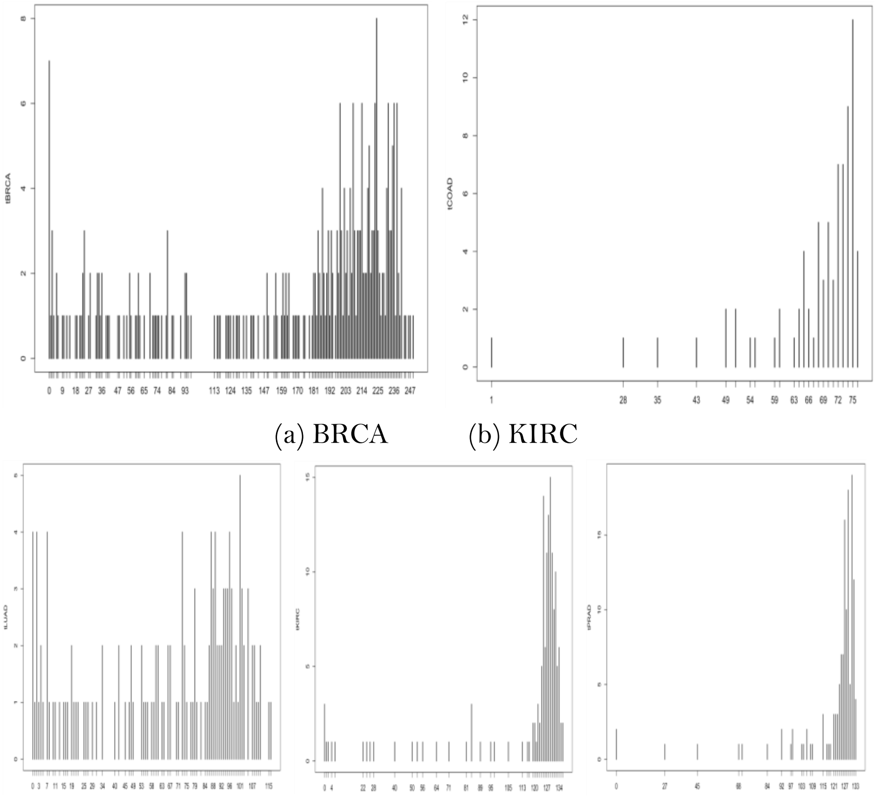
In order to understand the relationships between the samples in our data set, we constructed a network with each sample representing a node. The edges between S samples are determined by the level of the correlation between the G × 1 dimensional gene expression vectors. We start with our data matrix A which is S × G. Our correlation coefficients are defined as,

$$\rho_{i,j} = \frac{\sum_{k=1}^{G}(A_{i,k} - \mu_i)(A_{j,k} - \mu_j)}{\sigma_i \sigma_j} \quad .$$

Given the correlation ρi,j between gene expression vector for sample i with the gene expression vector for sample j, we define a threshold value, drawing an edge between sample i and sample j if the correlation is statistically significant. We determine whether a correlation coefficient is significant using the Fisher transformation, which converts the distribution of Pearson's correlation coefficients to a normal distribution. This transformation takes the following form:

$$Z_{i,j} = \frac{1}{2}\ln(\frac{1+\rho_{i,j}}{1-\rho_{i,j}})$$

Using the transformed correlation coefficients we can obtain a p-value from the Z-score, since they correspond to the normal distribution. We chose the 5% significance level to draw our edges in this graph. Figure 5 shows the degree distributions of the networks created by the mechanism discussed above for each cancer type. The degree distributions are left skewed suggesting that there are many high degree nodes among all the 801 patients. This also helps us conclude that the change in gene expression levels is highly similar for patients within one cancer type. Furthermore, the centrality measures are summarized in Table I. These measures suggest that there are certain patients that are central to the network corresponding to the cancer type, and thus are representative samples. We expect that adding more patients to the network would change the degree distributions and centralities of each patient. Then, depending on the degree measures of the new patient relative to patients close to this new patient who were already present in the network, we should be able to classify these patients into groups that would require similar therapies.



(a) BRCA (b) KIRC

(c) LUAD (d) COAD (e) PRAD

FIG. 5: Degree distribution of network for each cancer type

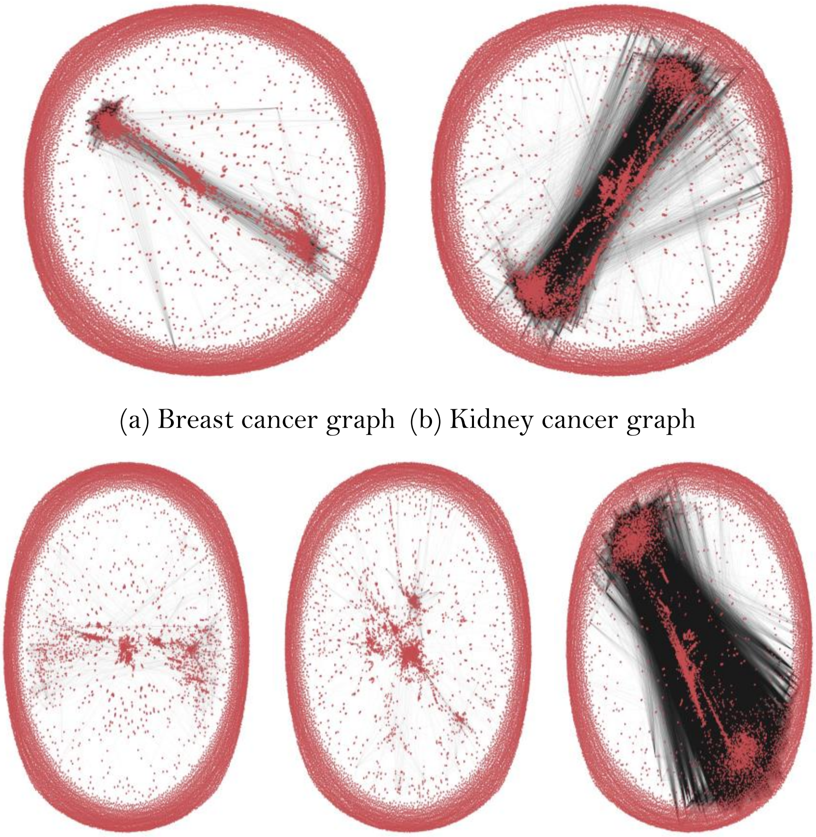
Cancer	Degree	Eigenvector	Pagerank
PRAD	34,158,275,390	34,158,275,390	34
BRCA	99	111	99
LUAD	229	229	229
KIRC	423,591	376,591	376
COAD	26,237,264,665	665	237,264,665

TABLE I: Table summarizing nodes with max. Centralities.

B. Weighted Gene Co-expression Network Analysis

A commonly used technique to analyze such data sets is to create a Weighted Gene Co-expression Network8,9. This is a graph which has genes as nodes and the edge is given between two nodes represents the correlation between the two nodes that the edge joins. In order to build such a network, we start by first splitting our data set based on cancer type and then proceed with the correlation computation as described above.

For the matrix containing the gene expressions for each subset of our data, we compute the Pearson correlation matrix, as shown above and then use that as our preliminary adjacency matrix. Once we have computed the matrix, we build the network by adding all the nodes, but only draw edges if the correlation is above a value of $\rho X,Y > 0.8$. This threshold was chosen based on Fisher exact test leads to a significance level of around 5%. Additionally, it yields graphs that are sparse enough to visualize, though also dense enough that will allow us to make accurate computations. The networks resulting from this method are shown in Figure 6. It is worth noting here that creating such large networks, plotting and computing the centrality measures proved to be very computationally intensive, as they all had above 20,000 nodes and between 40,000 and 200,000 edges. Even when using a powerful server (courtesy of the MIT Math Dep.), the algorithms took hours to run for each of the networks.



(a) Breast cancer graph (b) Kidney cancer graph

(c) Lung cancer (d) Colon cancer (e) Prostate cangraph graph cer graph

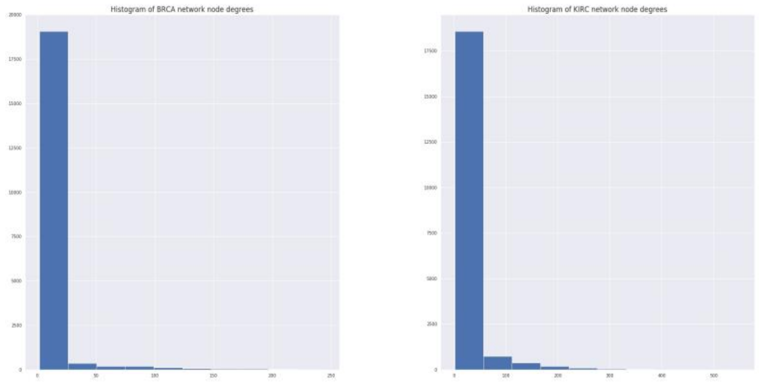
FIG. 6: WGCNs for each cancer type

It is interesting to note that even when only plotting the edges above the 0.8 correlation threshold, the graphs seem very dense. This is partially caused by the fact that some genes are naturally correlated and would be connected in the graph anyways. A way to go around this would be to use the partial correlation matrix as the adjacency matrix instead. The partial correlation would effectively condition on the rest of the genes, resulting in a more sparse network. However computing the partial correlation proved to be much harder computationally, or even impossible.

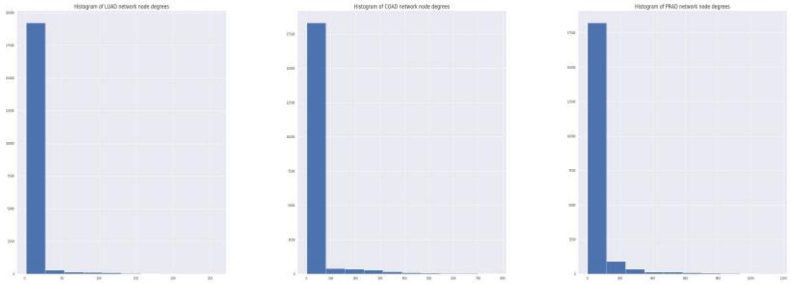
Once having the graphs, we first looked at the graph statistics. The basic statistics are summarized in Table II. Furthermore, we have plotted the histograms of the degree distributions in Figure 7. We can see that that the distributions seem to follow the power law. There seem to be many nodes that have low degrees.

Network descriptions			
Cancer Type	# Nodes	# Edges	Avg. Degree
BRCA	20,259	43,475	4.28
KIRC	20,262	70,219	6.93
LUAD	20,251	58,963	5.82
COAD	20,227	201,408	19.91
PRAD	20,252	171,008	16.89

TABLE II: Basic statistics of the networks



(a) BRCA (b) KIRC



(c) LUAD (d) COAD (e) PRAD

FIG. 7: Degree histogram for each WGCN

After creating those graphs, we computed some centrality measures, such as betweenness centrality, degree centrality and pagerank centrality. The results we got are summarized in tables Table III through Table VII, where we can see the gene numbers that ranked higher for each of the centrality measures we computed.

PRAD centralities			
Order	Degree	Pagerank	Betweenness
1	14,974	1,671	3,068
2	6,799	15,985	5,177
3	14,643	13,761	9,525
4	5,177	19,487	19,322
5	11,709	13,119	9,427

TABLE III: Centralities of PRAD WGCN

LUAD centralities			
Order	Degree	Pagerank	Betweenness
1	19,819	10,462	17,124
2	19,582	7,749	13,269
3	19,196	11,394	7,502
4	18,922	19,401	11,432
5	18,918	9	10,982

TABLE IV: Centralities of LUAD WGCN

BRCA centralities			
Order	Degree	Pagerank	Betweenness
1	15,512	14,974	19,862
2	14,376	17,430	4,749

3	3,356	16,274	14,974
4	8,355	19,847	20,355
5	1,139	715	1,511

TABLE V: Centralities of BRCA WGCN

KIRC centralities			
Order	Degree	Pagerank	Betweenness
1	6,799	1,363	15,147
2	2,111	18,173	19,309
3	6,022	2,124	17,805
4	3,267	1,298	5,330
5	17,791	3,913	13,650

TABLE VI: Centralities of KIRC WGCN

COAD centralities			
Order	Degree	Pagerank	Betweenness
1	19,375	12,509	1,213
2	6,259	12,402	16,556
3	18,822	5,280	16,463
4	3,997	4	5,198
5	19,862	15,139	713

TABLE VII: Centralities of COAD WGCN

From this analysis, we can see what the most ”important” genes are for each cancer type, based on their centralities. The genes with the highest centralities will be the most prominent in patients with the respective type of cancer, producing an outsize effect on the overall gene expression. We could then try to map each of the gene numbers to the actual gene name by ordering the gene sequence and finding the gene corresponding to each index number. From there we could research the function of that gene. We expect the function of the genes with the highest centrality in each cancer type to be somehow related to that organ in the body.

Furthermore, it is interesting to see that it is not the case that the top 5 genes are the same in each centrality measures. This happens because each centrality measure is computed differently and will lead to different results. Moreover, there is a very large amount of genes many of whom have very similar values for their centrality scores which means that even though one gene ranking highly in one centrality measure could have a very high centrality score in a different measure, it might still not make it in the “top 5”.

Using the list of the gene names and mapping that to the index given to us, we can find what gene name each gene index number corresponds to. Then we can search on the National Center for Biotechnology Information (NCBI), we can find what exactly each gene does and where it is most expressed. For example, the gene with the highest pagerank centrality in LUAD (lung cancer) is gene #10,462 which corresponds to gene ”MACF1 23499”. NCBI tells us that this gene ”encodes a large protein which is a member of a family of proteins that form bridges between different cytoskeletal elements”. Furthermore when we see in general this gene is mostly expressed in lung tissue, as shown in figure 8. Furthermore gene #17,124 (highest betweenness in lung cancer) or ”SPEN 23013” is a transcriptional repressor, which would make sense to have high centrality that regulates multiple genes expression that related to cancer.

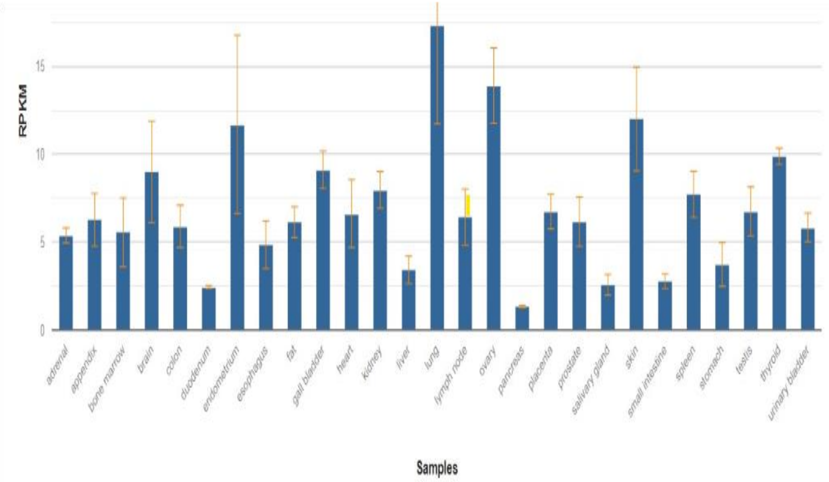


FIG. 8: Gene expression comparisonAdditionally, the gene with the highest degree centrality in colon cancer (VILL 50853), shows most expression in the stomach area, intestines and colon, as shown in Figure 9.

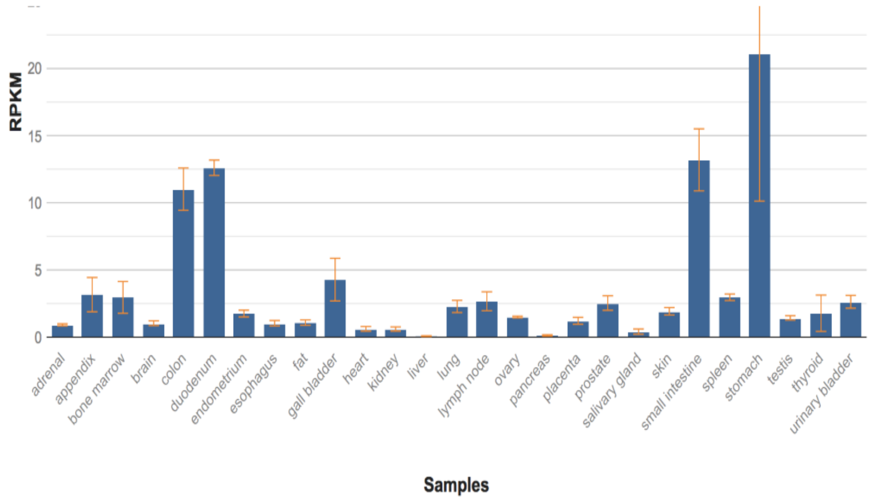


FIG. 9: Gene expression comparison

IV. CLUSTERS COMPARISON

A. Identifying the Critical Subset of Genes

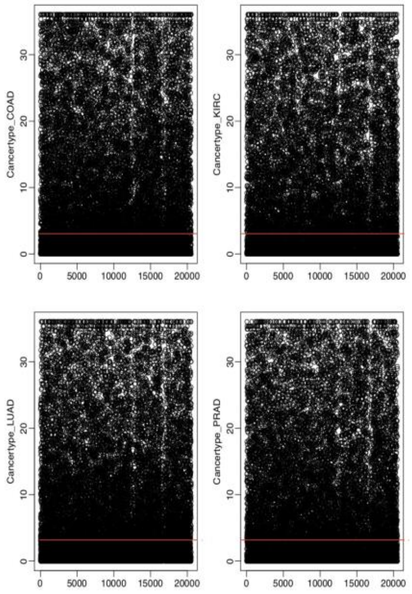
We fit a multinomial logistic regression model to classify our data, estimating coefficients for each gene. Analyzing these coefficients, we can determine whether that gene is a statistically significant determinant of a particular cancer type. Multinomial logistic regression is the generalization of logistic regression to multiple categories. Since we do not have normal samples in our dataset, we use BRCA samples to indicate a baseline, since it is the plurality of our samples. Fitting this model, we get both coefficients and standard errors, and each number corresponding to a model equation. For example, the first row (COAD) for the gene# is regressed to the equation.

$$ln(\frac{P(cancertype = COAD)}{P(cancertype = BRCA)}) = \beta_0 + \beta \cdot gene\#$$

The way to interpret this regression is that β means one unit increase in gene# is associated with the decrease of probability in being COAD instead of BRCA in the amount of β. To be more specific, the ratio of the probability of choose one outcome category over the probability of choose the baseline category is the righthand side linear equation exponentiated.

$$\frac{P(cancertype = COAD)}{P(cancertype = BRCA)} = a \cdot e^{\beta \cdot gene\#}$$

Thus, β are relative risk ratios for a unit change of predictor variable. Since we have 801 samples data, this should provide a reasonably accurate estimate through regression. We also got standard deviation from the regression processes (for the coefficient) P-values were calculated according to t-tests H0 : β = 0 vs.HA : β 6= 0. After we got the P-values for all cancer types based on breast cancer over 20532 genes. We chose the significance level to be 0.005 and delete the genes with P-values below this threshold in all 4 cancer types. The further analysis is based on this small dataset. The −log(p) vs. gene# were plotted (known as Manhattan plots) for each type of cancer.



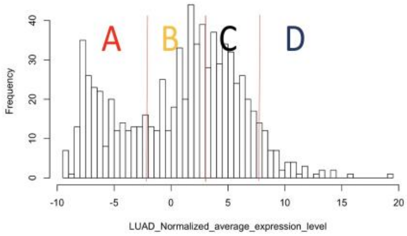


FIG 11:Histogramofgenesexpressionlevelsin Adown- Bnormal- Cover-

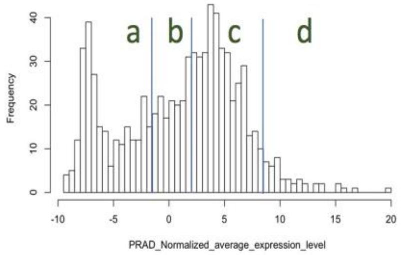


FIG. 10: Log(p) vs. Gene# of 4 cancer types COAD,KIRC, LUAD, PRAD. Red line (log(p)=2.3) is the threshold. Genes below this threshold were removed.

By choosing a threshold of p=0.005, log(p)=2.3. We removed data below the red line in the plot. However, according to the plots there are still lots of genes to be analyzed. Interestingly, there are two narrow blank spaces shown in all the plots and those parts may suggest that P value are all large, and the cancer types has no relationship with those genes. Thus, those genes can either be genes related to this cancer (and are similar regulated in all cancer types) or they can be genes unrelated to this cancer (similar expressed for all people). This method helped to reduce the set to 1075 genes, which we used for the following analysis.

B. Clustering by Gene Expression Levels

Now with the smaller dataset, we want to analyze the expression level of genes among different cancer types, specifically we focused on LUAD and PRAD. They are chosen since they have similar sample size. Before any further analysis, the gene expressions were normalized according to the average and standard deviation of that specific gene expression in all cancer types.

The expression levels distributed according to the histograms in LUAD, PRAD.

FIG. 12: Histogram of genes expression levels in PRAD. a down-, b normal-, c over-, d highly expressed.

According to the graph, we think the expression could be grouped into 5 groups, group A: (-10,-2), group B: (-2,2), group C: (2, 8), group D: more than 8, with other NA values to be group 0. This is consistent with the down-regulated genes (compared with other cancer types), normal-expressed genes, slightly over-expressed genes, highly over-expressed genes. (This could be changed according to the tissue type to make it more biological meaningful.)

C. Conserved of Gene Expression Levels in LUAD and PRAD

According to the groups, an adjacency matrix is generated with 1 at aij if both gene i and gene j are in same group, otherwise 0. This is then used to generate network. We expected to see all the genes in same group would be fully connected with each other, and a function were used to separate the connected components apart. The networks of LUAD and PRAD were shown as following.

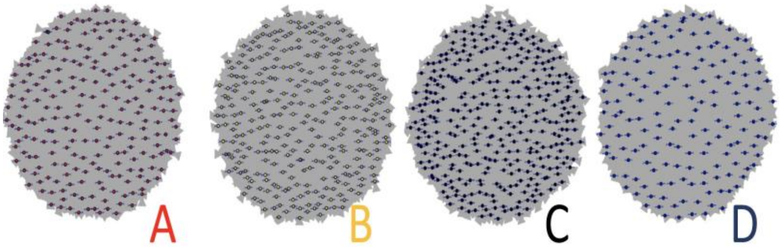


FIG. 13: LUAD Network with gene in same label(A B C D see Histogram) fully connected.

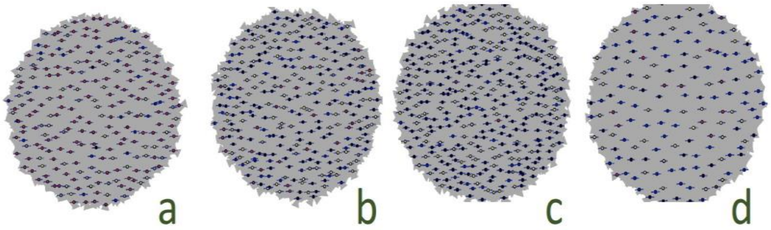


FIG. 14: PRAD Network with gene in same label(a b c d see Histogram) fully connected with color of the nodes consistent with LUAD labels

The nodes are colored according to the expression level in LUAD cancer: group A as red, B as yellow, C as black and D as blue (consistent with their labels in LUAD histogram). The graphs suggest that in general, the genes expressed highly in LUAD are distributed evenly in PRAD, and especially group B and group C are mixed evenly. Probably because the threshold choose is not significant. However, there are some preservation of the expression patterns.

V. DISCUSSION AND FURTHER WORK

Rank	Type	New Cases	%	Deadliest
1	Lung	2,093,876	12.3	1
2	Breast	2,088,849	12.3	3
3	Colorectal	1,800,977	10.6	2
4	Prostate	1,276,106	7.5	5
5	Stomach	1,033,701	6.1	n/a

TABLE VIII: Global cancer incidence10 where % refers to the percent of new cases of cancer diagnosed in the US in 2018 and deadliest refers to the ranking for that cancer in causing the most deaths in 2018

Cancer is one of the most significant public health challenges, particularly in the developed world. In this project, we examined 4 of the top 5 (and 5 of the top 15) cancer types in terms of new cases diagnosed in 2018, evidenced by Table VIII. In addition to being prevalent, the cancer types studied here correspond to 4 of the top 5 cancer types contributing to deaths in America. The prevalence of datasets and computational tools has revolutionized nearly all fields of science, particularly biology. Transferring successful models from the statistical modeling literature to this dataset has allowed us to validate existing scientific conclusions and identify areas which warrant further study. We were pleased to find that the cancer types can be clustered into groups using out-ofthe-box approaches for dimensionality reduction. Since each cancer type is different in many ways, it is reassuring to see those differences reflected in our statistical approach. Other portions of our report highlight areas that could be worth exploring further from the biomedical perspective. For example, in section 2.3, it was seen that the 4 most prevalent and deadly cancer types appeared clustered more closely together than to KIRC. It would be interesting to explore how this matches the intuition of oncologists, who might have a sense of which cancer types are more similar to each other. Through our network approaches we were able to identify genes of interest. We are optimistic that these centrality measure of our network correspond to biological insight and that these network approaches can serve as a spotlight to help guide researchers to study potentially high impact areas of the genome in a more efficient manner.

VI. CONCLUSION

In this report we were able to summarize the similarities and differences of 801 samples of 5 cancer types from a dataset generated by UC Irvine. After performing an exploratory analysis, we were surprised to see that the gene expression profiles could be easily clustered. This motivated further analysis to characterize interactions between patients and genes that were indicative of biological differences between the cancer types. To characterize these relationships we constructed networks: one that represented the relationships between patients and another that aimed to characterize the relationships between genes. Using standard network analysis measures (such as centrality statistics) we highlighted genes that appear to be highly influential for each cancer type, such as MACF1 23499 for LUAD and VILL 50853 for COAD. Both these genes appear to plausible genes involved with these cancer types, validating elements of our approach. These networks approaches and expression analyses applied gene expression data aim to motivate for future work to understand the biological implications of standard statistical measures in gene expression profiles.

VII. REFERENCES

1J. Khan, J. S. Wei, M. Ringner, L. H. Saal, M. Ladanyi, F. Westermann, F. Berthold, M. Schwab, C. R. Antonescu, C. Peterson, et al., Nature medicine 7, 673 (2001).

2Y. Lee and C.-K. Lee, Bioinformatics 19, 1132 (2003).

3R. Tibshirani, T. Hastie, B. Narasimhan, and G. Chu, Proceedings of the National Academy of Sciences 99, 6567 (2002). 4A. K. Dubey, U. Gupta, and S. Jain, Asian Pac J Cancer Prev 16, 4237 (2015).

5J. A. Botia, S. Guelfi, D. Zhang, K. D'Sa, R. Reinolds, D. Onah, E. M. McDonagh, A. Rueda-Martin, A. Tucci, A. Rendon, et al., bioRxiv , 288845 (2018).

6Y. Yang, L. Han, Y. Yuan, J. Li, N. Hei, and H. Liang, Nature communications 5, 3231 (2014).

7L. v. d. Maaten and G. Hinton, Journal of machine learning re- 8T. F. Fuller, A. Ghazalpour, J. E. Aten, T. A. Drake, A. J. Lusis, search 9, 2579 (2008). and S. Horvath, Mammalian Genome 18, 463 (2007).

9P. Langfelder and S. Horvath, BMC bioinformatics 9, 559 (2008).

10F. Bray, J. Ferlay, I. Soerjomataram, R. L. Siegel, L. A. Torre, and A. Jemal, CA: a cancer journal for clinicians 68, 394 (2018).