Linking TCGA's to Lung Cancers
Clustering Lung Cancer Patients by Genomic Information

Haoda Li

February 11, 2020

Summary

This report examines whether lung cancer patients' with different survival time is correlated with distinct genomic subgroups. Three distinct clusters are formed through dimension reduction and community finding and showed different survival curves. The report further considers the the association among different observed clinical factors and their genomic features, while the associations were not found for gender, age, and cancer's tumor stage. This report concluds that the genomic subgroups can be used to predict the patients' survival curve.

1. Introduction

Human genes are sequences of tiny piece chemicals that are often encoded as T, C, G, and A's. With the examinations on such sequences of letters, the scientists are able to find clues of how cancers happen. However, the gene sequences are as long as millions of letters, hence are almost unable to interpret. One possible solution is to reduce the dimensions and build clusters so that each cluster show different biological features.

The report will focus on 424 lung cancer patients and their 60483 gene sequence expressions from TCGA [1] and find distinct gene subgroups that have different survival curves. In addition, clinical features such as age, gender, and tumor stage and examined. We aim to find whether the gene group plays a unique role in predicting a patient's survival.

2. Methodology

Overall, the data is divided into 2 parts. The gene scanning and the clinical data. For the gene data, because of the extremely large dimensionality, we need to first reduce the dimension. Then, we choose clusters from different clustering algorithms and evaluate how well the clustering performs on the patients' survival data. Afterwards, other clinical features are then incorporated and we try to build regression models that can explain the relationship among different features and the survival time.

2.1. Clustering on Gene Data

2.1.1.Preprocessing

The first difficulty is that the data points positive and extremely right skewed. As shown in the plot, half of the data points are close to 0 and the other half are arbitrarily large. Therefore, we normalize the data points and take a log one plus transformation so that the data. After the normalization, the data is less right skewed and numerically smaller, which makes further processing easier.

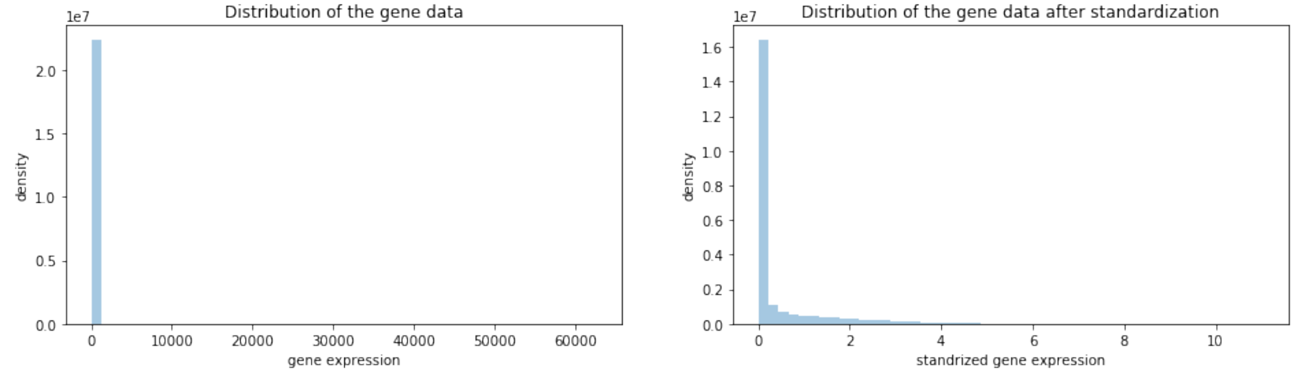

Fig 1. Distributions of the gene data points

Also, as shown in Fig 2, note that the data is close to spare. 80% of the data entries is close to zero. Therefore, we can use principle component analysis to reduce the dimension. PCA can then preserve the most variations and the lower dimensions is easier to be interpreted and analyzed. After performing PCA, we take the first 50 components, which preserves the majority of the variance. As shown in Fig 2, the data entries are dense and the first few dimensions have much variations.

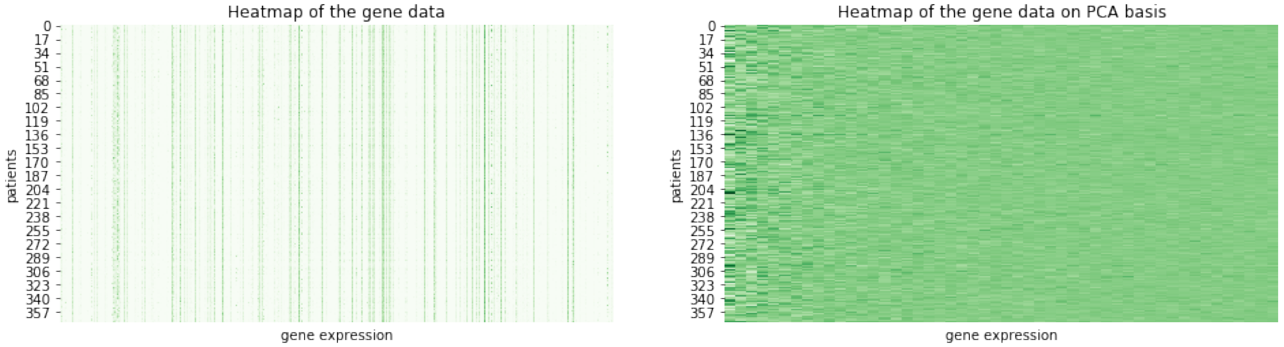

Fig 2. Heatmaps of the gene data

2.1.2. Clustering

Among many clustering algorithms, we choose to use Louvain algorithm. Louvain algorithm [2] are adapted from network theorem and has great performance on clustering gene expressions. Louvain algorithm works by first constructing a neighborhood graph , which is to assign edges based on Euclidean distance among genes. Then, Louvain algorithm tends to find "communities", which are subgroups of genes that are closed and have many connections in the neighborhood graph. Compare to distance based clustering algorithm such as k-nearest neighbors, Louvain algorithm tends to capture the internal connections based on the number of points within some distance. Rather than finding hyper-balls shaped clusters by kNN, Louvain algorithm can find various clusters with different shapes.

2.2. Other Factors from Clinical Data

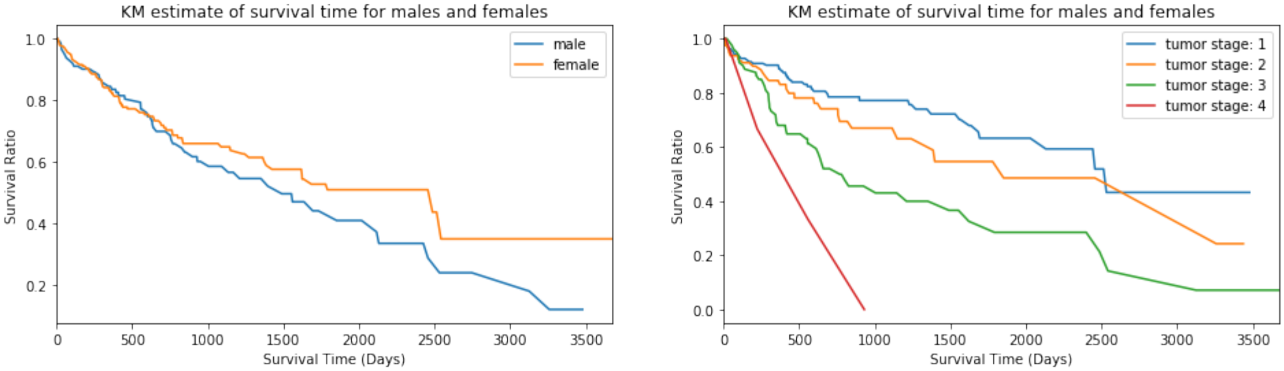

Fig 3. KM estimate for different factors

Taking consideration of other clinical factors, including age at the diagnosis, tumor stage, and gender. Naturally, the aged people are more likely to be diagnosed with cancers. Which is reflected in our dataset, as the majority of the patients are centered around 60. However, due to the low variations of age in our dataset, evaluating the relationship between age and survival time is hard. In addition, tumor stages definitely have decisive impact on the survival time, as shown in Fig 5. Finally, there is no evidence that gender is related to survival, as shown in the survival curve. Also, note that the number of observations between

2.3. Models

From the investigations above, two Cox regression models are proposed. The additive model considers the effect of gene clusters, age, and tumor stage on survival time. The interaction model, in addition, consider the interactions of gene clusters vs. age and gene clusters vs. tumor stage. The models are given as

$$h_1 = \text{baseline} \cdot \exp(\beta_1 \overline{\text{cluster}} + \beta_2 \overline{\text{tumor}} + \beta_3 \overline{\text{age}})$$

$$h_2 = \text{baseline} \cdot \exp(\beta_1 \overline{\text{cluster}} + \beta_2 \overline{\text{tumor}} + \beta_3 \overline{\text{age}} + \beta_4 \overline{\text{cluster:tumor}} + \beta_5 \overline{\text{cluster:age}})$$

where $\overline{\text{group}}$ are zero-centered observations. We use the interactive model to assess whether the gene clusters is correlated with age or tumor stage. On the other words, whether the gene groups is a unique predictor that can provide additional information than clinical observations.

3. Results

The clusters of genes are plotted on a 2D UMAP[3] basis. UMAP is a non-linear dimension reduction algorithm that approximates a manifold from the neighborhood graph. UMAP tends to give a more intuitive visualization of the distance and connections. We also show the dendrogram to see the hierarchical clusters.

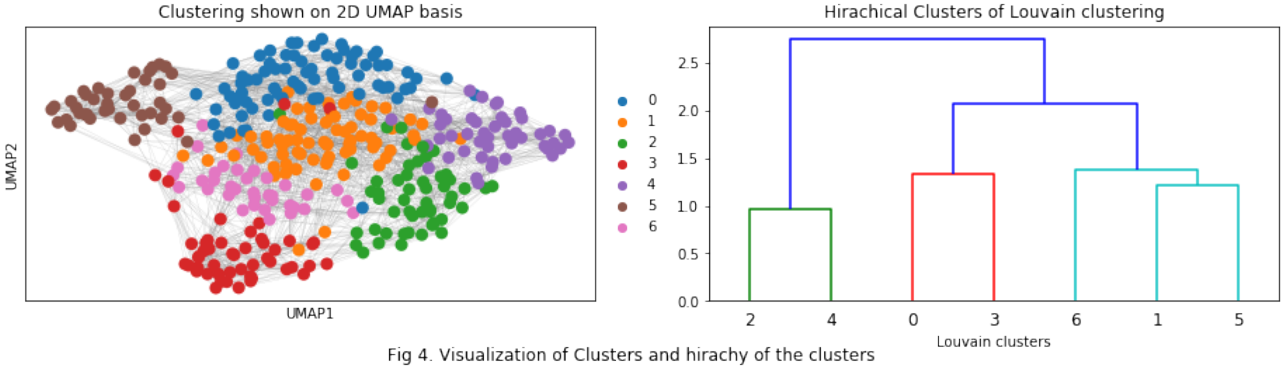

Fig 4. Visualization of Clusters and hirachy of the clusters

The performance of our clusters on survival time is evaluated through Kaplan-Meier estimate. However, from the plot, we see some of the clusters from the Louvain algorithm have very similar similar survival curve, hence we will use 3 clusters from hierarchical clusters at level 2 of the dendrogram.

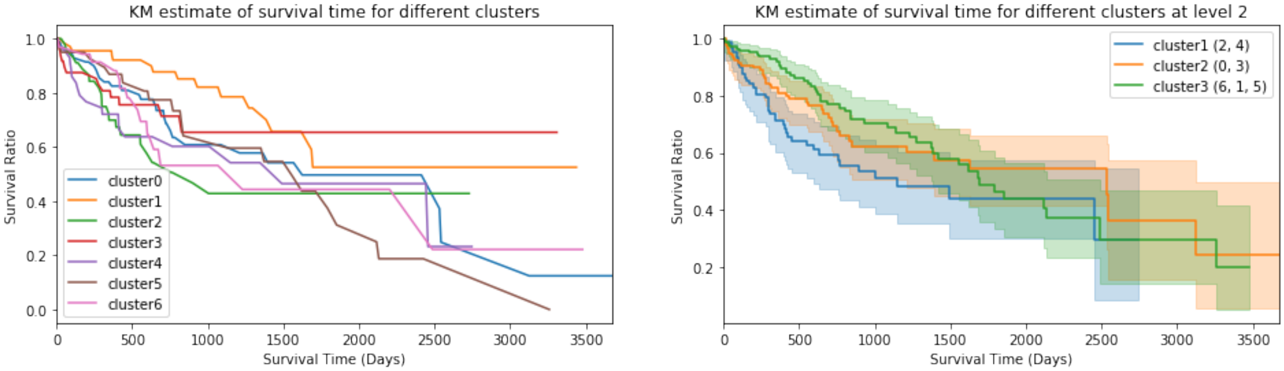

Fig 5. Survival curves of the clusters

Then, we fit the regression models based on the clusters and the coefficients of the fitted models are shown below.

name	coef. of model1	p-value of model1	coef. of model2	p-value of model2
louvain_h=2	-0.489939	0.0387053	-1.77848	0.145783
louvain_h=3	-0.568266	0.0148153	-1.19727	0.271095
tumor_stage=2	0.359658	0.153736	0.78168	0.110364
tumor_stage=3	0.990963	4.89213e-06	1.55286	7.68451e-05
tumor_stage=4	1.85683	0.00215931	1.84658	0.0138706
age_at_index	0.0133238	0.0740811	-0.00365337	0.772061
louvain_h=2:tumor_stage=2			-0.373705	0.56589
louvain_h=3:tumor_stage=2			-0.737301	0.246352
louvain_h=2:tumor_stage=3			-0.922433	0.0938969
louvain_h=3:tumor_stage=3			-0.603801	0.261471
louvain_h=2:tumor_stage=4			-0.241636	0.850306
louvain_h=3:tumor_stage=4				
louvain_h=2:age_at_index			0.0294332	0.131746
louvain_h=3:age_at_index			0.0179933	0.30297

The p-values of all interaction terms are extremely large. In addition, the overall p-values in model 2 are much larger than model 1. This observation hints multicollinearity problem in model 2. Therefore, we conclude that there is no clues of interactions for cluster vs. tumor stage and cluster vs. age. In addition, the independence of cluster vs. age and cluster vs. age is hinted in Fig 6, where the distributions are similar among cluster groups.

For coefficients in model 1, there is some evidence that the clusters is associated with survival time. Although this association is much weaker than tumor stage. Because the p-value is still close to 0.05 and the model does not fit quite well, this clustering may still be insignificant.

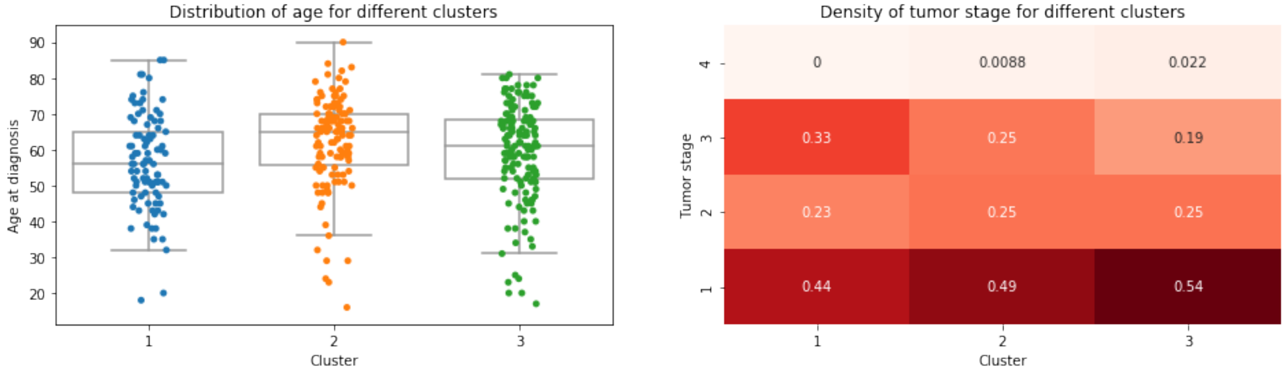

Fig 6. The association for different factors and gene clusters

4. Conclusion

In conclusion, the gene subgroups that found by PCA dimension reduction and Louvain algorithm provide some clues of survival times for the patients with lung cancer. In addition, the gene clustering information is likely to be independent of the observed clinical information, such as age, gender, and tumor stage. Therefore, measuring the genes of the patients can provide unique information for the treatment plan.

However, we have to notice that the Cox regression fit is not quite good, and the coefficients of the clusters are just below 0.05. Because the dimensionality of genes is much greater than the number of observations, this model may not be generalized to other cancer cases. Further validations and testings should be conducted on other cases to measure the performance of this clustering procedure. In addition, these gene clusters can correlate with some clinical factors that is not examined, and the clusters may not have unique and independent impact on the severity of the cancer. We need to consider more clinical factors that may be correlated with the gene clusters.

References

[1] John N Weinstein, Eric A Collisson, Gordon B Mills, Kenna R Mills Shaw, Brad A Ozenberger, Kyle Ellrott, Ilya Shmulevich, Chris Sander, Joshua M Stuart, Cancer Genome Atlas Research Network, et al. The cancer genome atlas pan-cancer analysis project. *Nature genetics*, 45(10):1113, 2013.

[2] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008(10):P10008, 2008.

[3] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction, 2018.