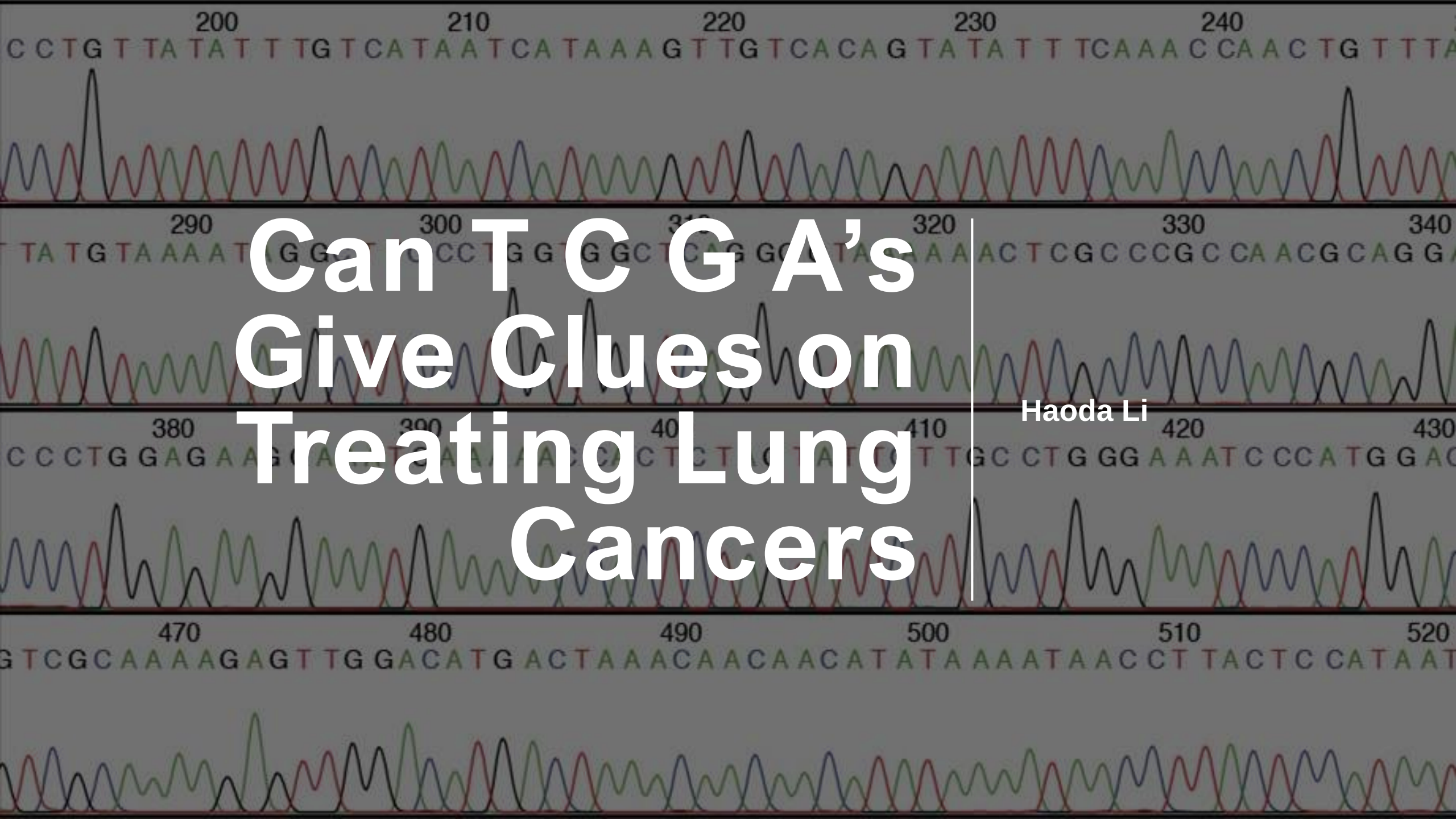


Can T C G A's Give Clues on Treating Lung Cancers

Haoda Li

Problem Definition

Is there any **genomic subgroups** that differentiates **patients with lung cancer** in survival times. And, if exists such subgroups, how are the groups related to other factors.

Genomic Subgroups

The patterns of each patient's RNA sequencing are described as a "dictionary" of genes. We aim to make clusters such that each cluster of patients show different characteristics.

Patients with Lung Cancer

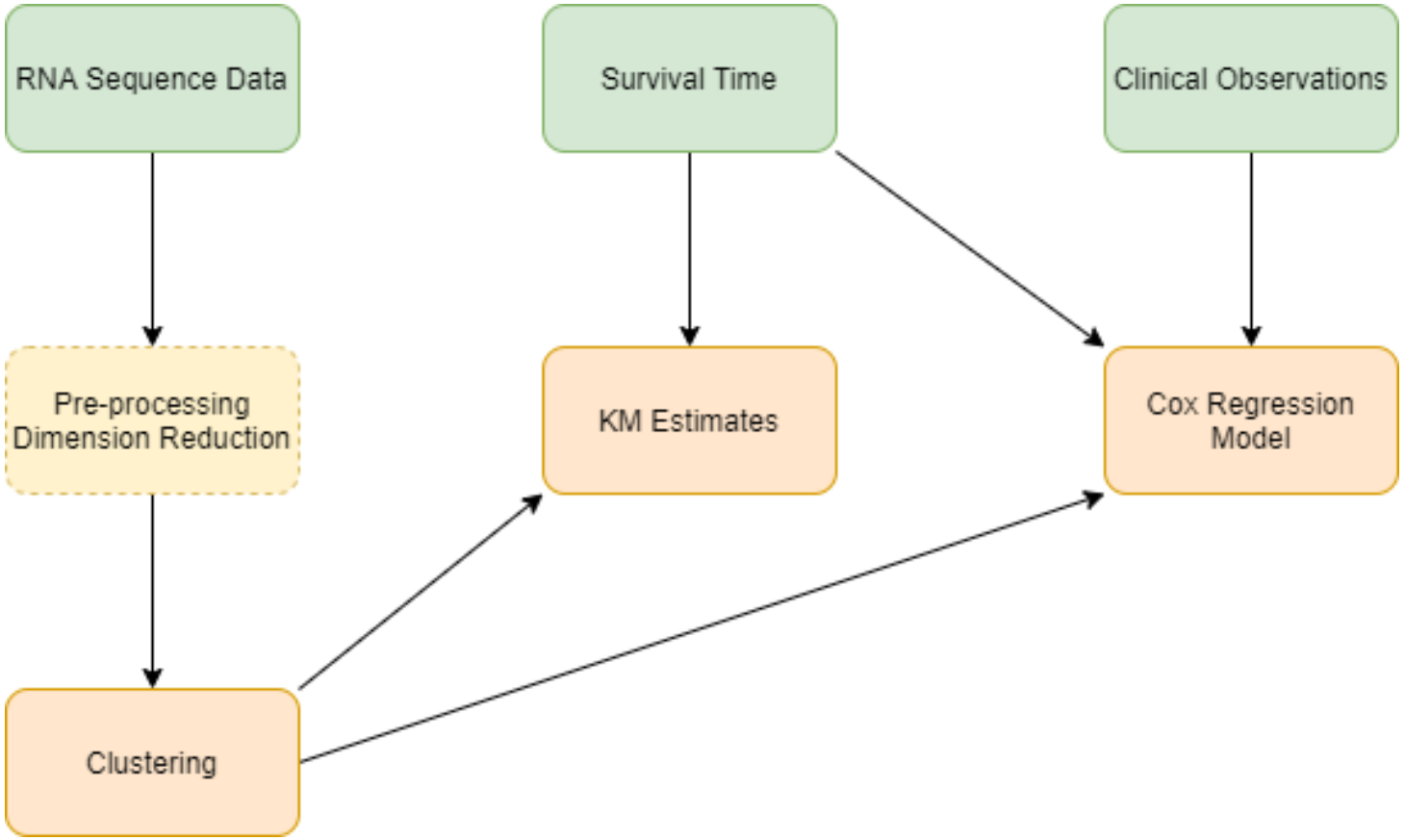
A subgroup of 424 cases of specific types of lung cancer obtained from TCGA Project[1].

Difficulties

Much higher dimension (60483 of different genes) than observations (424 patients).
The "optimal" subgroup is not defined.

[1] The results here are in whole or part based upon data generated by the TCGA Research Network: https://www.cancer.gov/tcga

WORKFLOW


RNA Sequence Data
Survival Time
Clinical Observations
Pre-processing
Dimension Reduction
KM Estimates
Cox Regression
Model
Clustering

Clustering Genomic Data

Pre-processing

Because our data is quite right skewed, standardizing and taking log transformation (plus 1 to avoid log 0)

Dimension Reduction

Using PCA to to reduce to 50 dimension as our data is quite sparse.

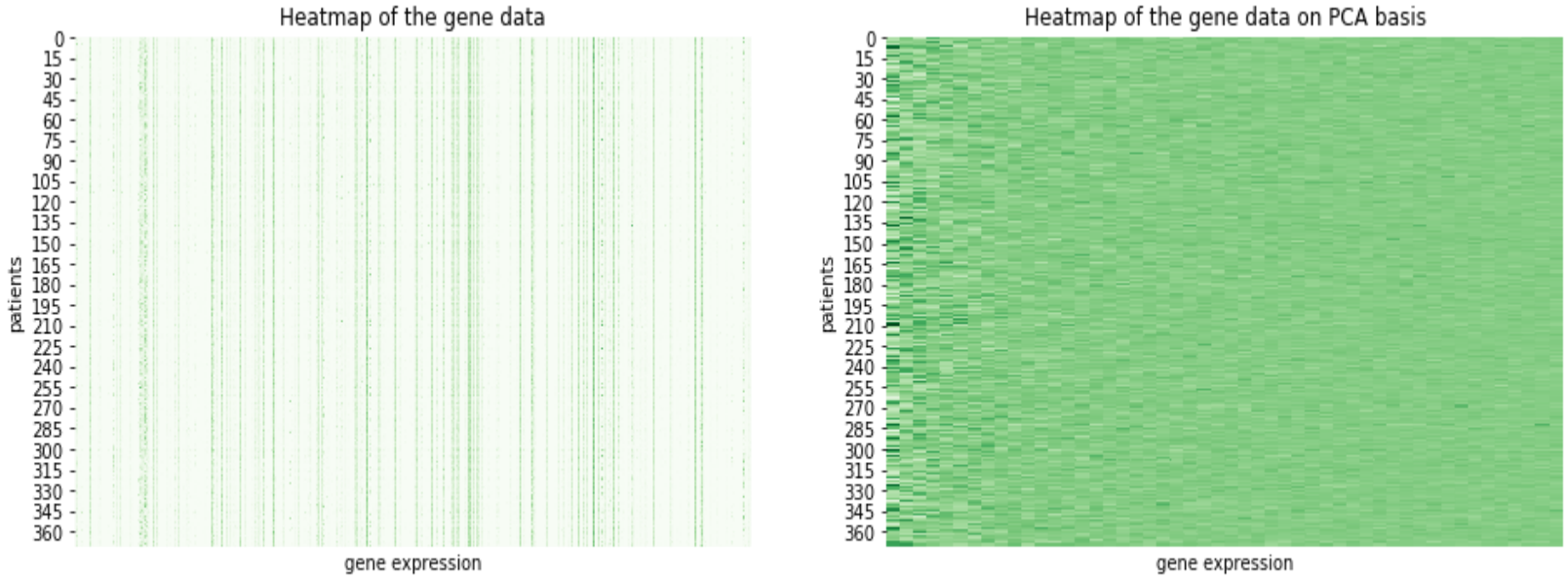


Fig 2. Heatmaps of the gene data

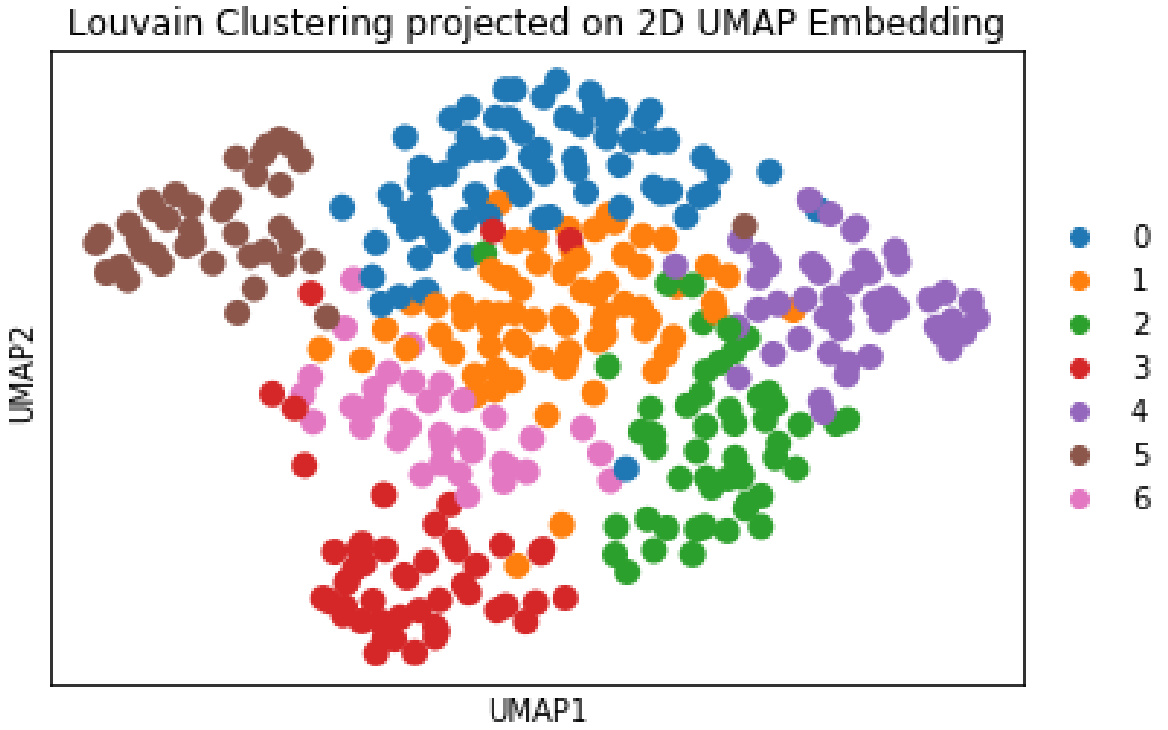
[1] McInnes, L., Healy, J., & Melville, J. (2018). Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.

Clustering Genomic Data (Cont.)

Neighboring
Building “neighborhood” (connections) based on distance.

Louvain Clustering [1]
Adapted from Network theory, make communities (clusters) based on the number of neighbors.

Compare to kMeans, neighboring and Louvain Algorithm focus on the common features across groups rather than simply distance and the number of clusters is not specified.



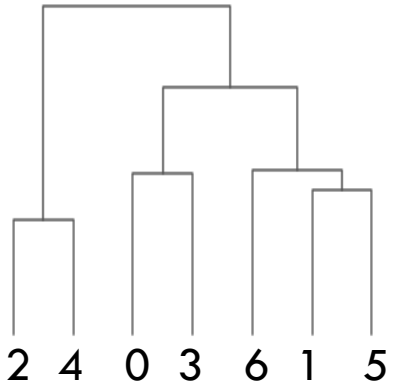
[1] Blondel, V. D., Guillaume, J. L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, *2008*(10), P10008.

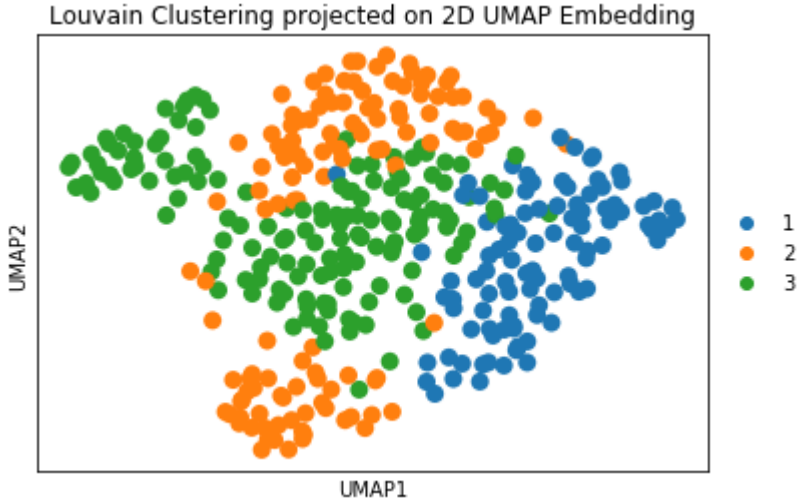
Survival Analysis on Subgroups

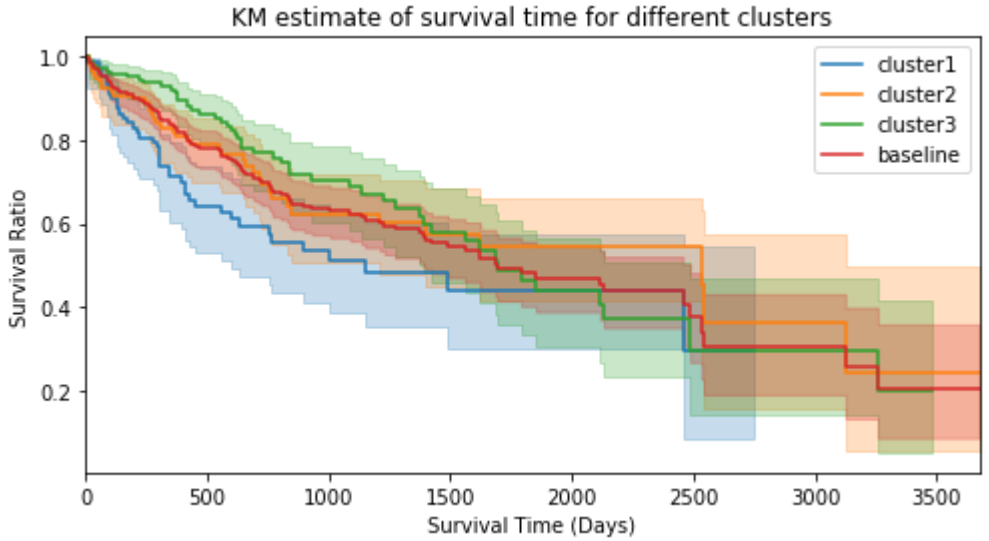
We consider a censoring event when patients that are alive and is at their day of the last report.

Using Cox Regression, there are weak evidence (p=0.13) that the 7 clusters have different survival time.

Based on the hierarchical clusters provided by dendrogram, we only consider 3 clusters. Then, there are some evidence (p=0.02) that the 3 clusters have different survival time.







Different Factors on Survival Time

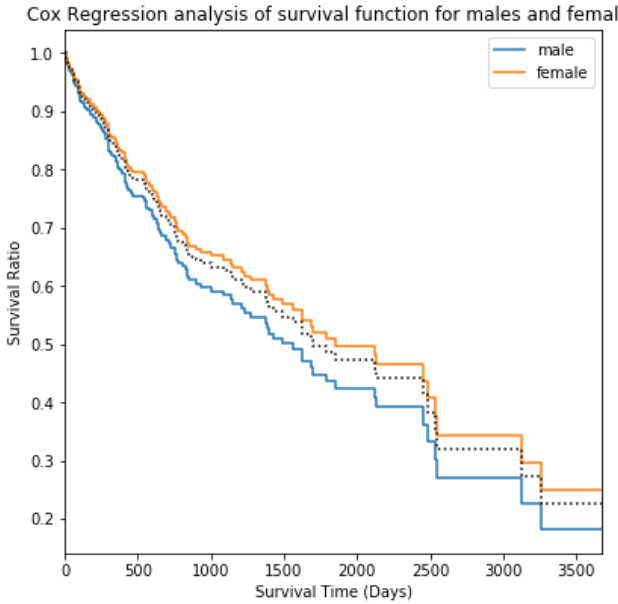
Gender

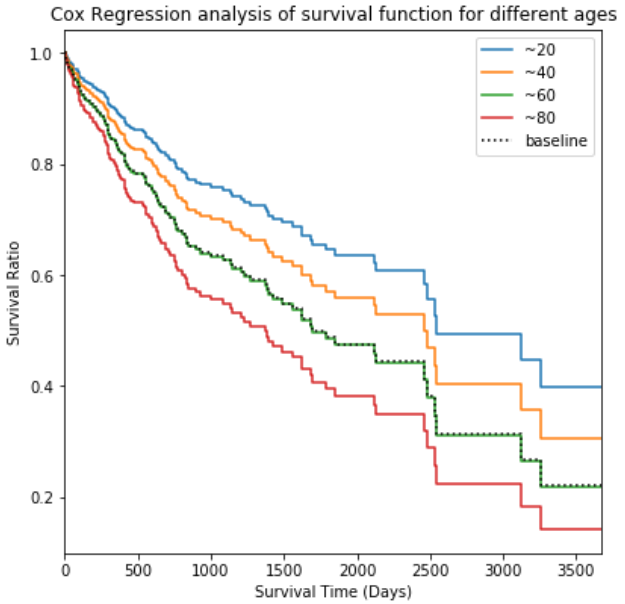
There’s no evidence that gender has impact on survival time.

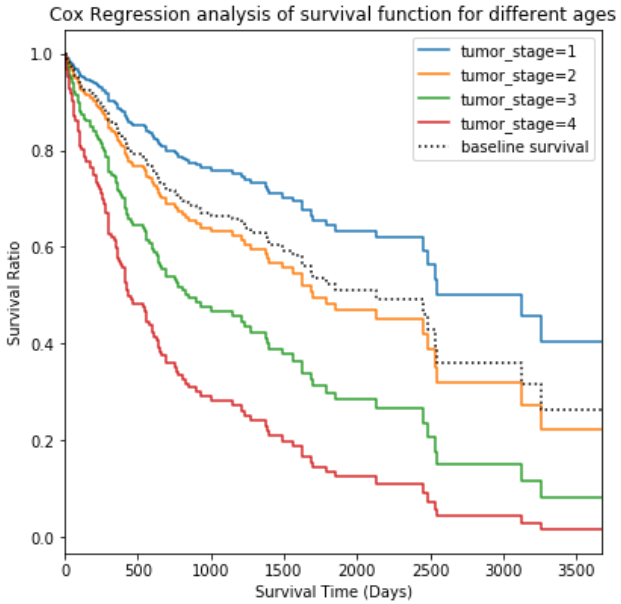
Age and Tumor Stage

Most patient’s ages are centered around 60.

Naturally, older patients and patients suffer later tumor stages have lower survival time.







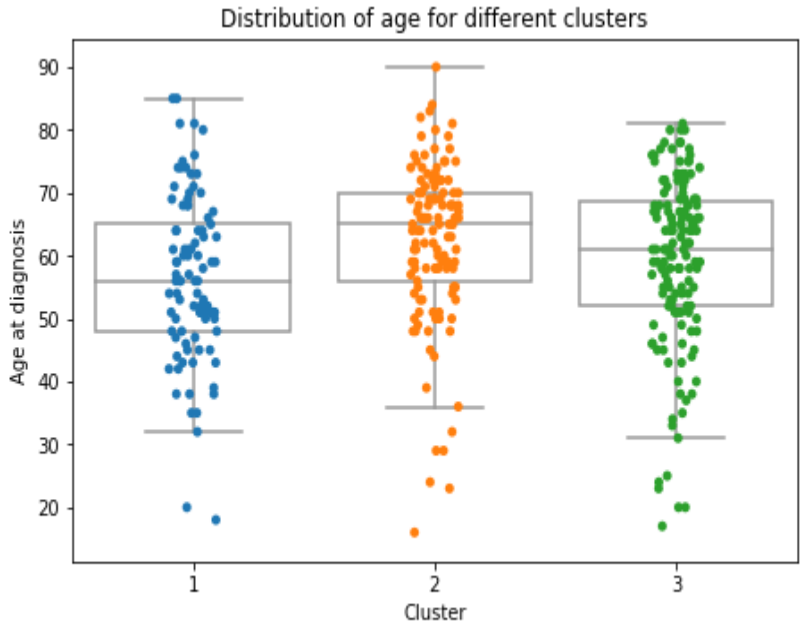
Different Factors on Genomic Clusters

Tumor Stage

There is some weak evidence that tumor stage is associated with the genomic clusters.

Age

There is some evidence that age is associated with genomic clusters while the association is too weak.



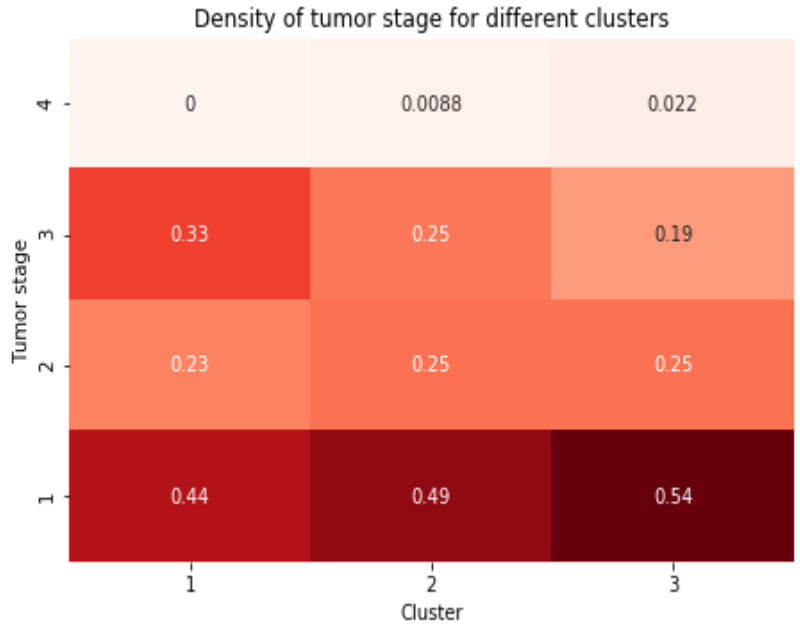


Fig 6. The association for different factors and gene clusters

Model and Results

We use Cox regressions with GOF tests on whether there are interactions between the clusters vs. age and clusters vs. tumor stages.

Model 1 (Additive Model)

Surv(duration, status) ~
factor(louvain) + factor(tumor stage) + age

Model 2 (Interaction Model)

Surv(duration, status) ~
factor(louvain)∗factor(tumor stage)
+factor(louvain)∗age

P-value for most of the interaction terms are ≥ 0.15, which provides strong evidence that there is no association on gene clusters vs. age and gene clusters vs. tumor stage.

Parameter	coefficient	P-value
Cluster = 2	-0.490	0.038
Cluster = 3	-0.568	0.015
Tumor stage = 2	0.360	0.15
Tumor stage = 3	0.99	4.9×10^{-6}
Tumor stage = 4	1.86	2.16×10^{-3}
Age	0.013	0.075

Coefficients for additive model

Conclusions

The genomic clusters have is associated with lung cancer. We may use such gene information to predict how severe the lung cancer is.

However, this genomic predictor does not seem to associated with overserved clinical factors.