Shan Tang
Kyoko Yamaguchi
Project proposal
STAT6500

Our dataset consists of 528 patient samples for Head and Neck Squamous Cell Carcinoma (HNSC) with matched clinical annotations and methylation profiles from the Genomic Data Commons (GDC, https://portal.gdc.cancer.gov/). After cleaning and applying mean imputation to fill in missing values, a matrix of the top 1000 most variably expressed CpGs were retained for further analysis. Because there were multiple entries in the clinical annotations table for one patient which needed to be formatted to have unique rownames, it was wrangled such that we retained the information from the follow-up date that was furthest from the study start date.

Our aims for this project are to **1) build a predictive/grouping model using this methylation data**, and **2) investigate the association between methylation-based patient clusters and other clinical features such as survival or tumor grade.**

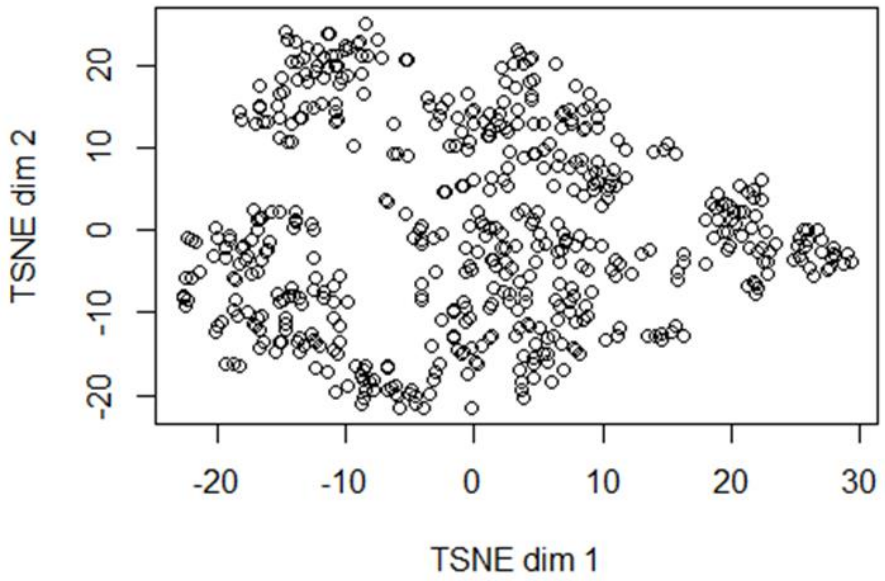
The first major part of our project is clustering. We will apply different clustering methods such as K-means clustering (kmeans), K-medoids clustering (pam), hierarchical clustering (hclust), model-based clustering (mclust), and spectral clustering (specc, Spectrum). We will select the clustering algorithm that gives the greatest visual agreement with T-SNE and UMAP projections. If feasible, we may also choose to create an artificial dataset with similar distributions and dimensions to our data, assign our “gold standard” theoretical clusters, and apply these methods to choose the clustering method that agrees most with the "gold standard" cluster assignments using the Adjusted Rand Index (ARI).

After finalizing our clustering scheme, it may be a good idea to examine any clinical factors that may influence patient clusters. We will likely narrow our search from all variables in our clinical table to only those containing 10 or less unique levels. This will be important to keep in mind, as some patient clusters may be influenced by factors like age group, sex, tumor grade, etc. which are already known to be predictors of survival.

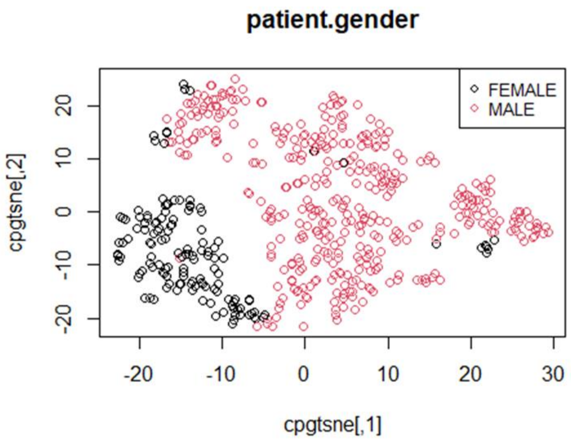
The latter half of our project will focus on survival analysis. A Kaplan-Meier curve generated from columns “event” and “time-to-event" can be divided with our cluster groupings, and a Log-rank test could be performed to test whether at least one of these curves is different from the others. Post-hoc tests will need to be performed as necessary to find which cluster is most influential to overall survival (OS). Similar univariate analyses will be conducted with variables from the clinical annotations table containing less than 6 unique levels, as too many categories would be difficult to visualize any separation between the Kaplan-Meier curves.

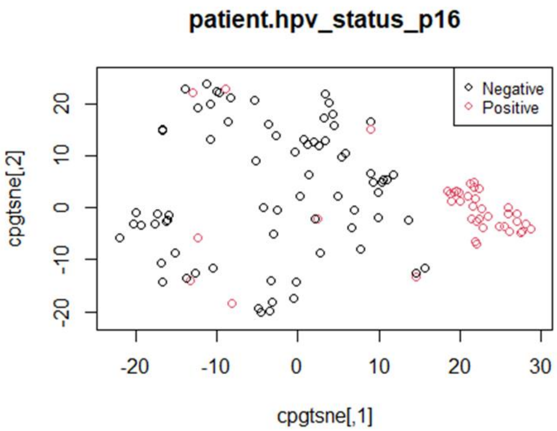
Cluster membership can also be used as independent predictors for a multivariate Cox proportional hazards model with which we could feed in multiple predictors to predict OS. It would be important to investigate whether membership to any cluster(s) is important to OS. It would also be important, as stated earlier, to identify if any clusters are confounded with variables that are already known predictors of OS. Stepwise variable selection using AIC or BIC, or regularized Cox regression (glmnet::glmnet with family=”cox”) could be used to reduce the number of predictors to only those that are impactful to the multivariate Cox proportional hazards model.

Below are some key figures and numerical summaries relevant to our questions.



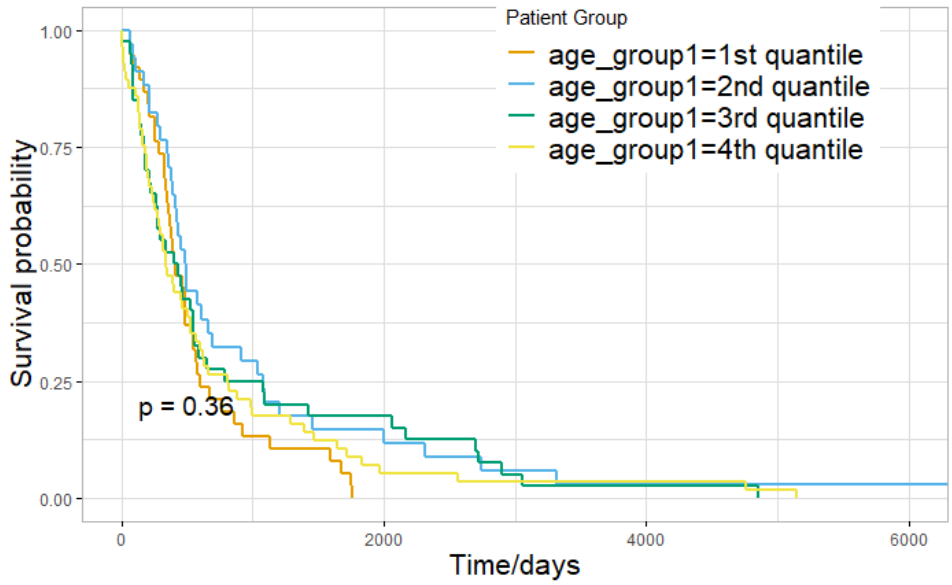
Here is an uncolored T-SNE plot, which appears to have 4 or 5 clusters, at least from visual inspection. Possible things that can influence the separation of patient samples could be found by overlaying categorical information from the clinical annotations table as colors on top of the T-SNE.





For example, the cluster we see on the bottom left appears to be defined by females, and the small rightmost cluster of patients appears to be defined by HPV positivity. This means that even if the bottom left cluster and the rightmost cluster happen to be associated with OS, they are also confounded by other variables that may be known predictors of OS, so we may not want to place them in the same model.

Preliminary survival analyses show that certain clinical factors postulated to influence OS are not: for example, when we divide patients into 4 age groups based on quartiles, we do not see any differences in OS.

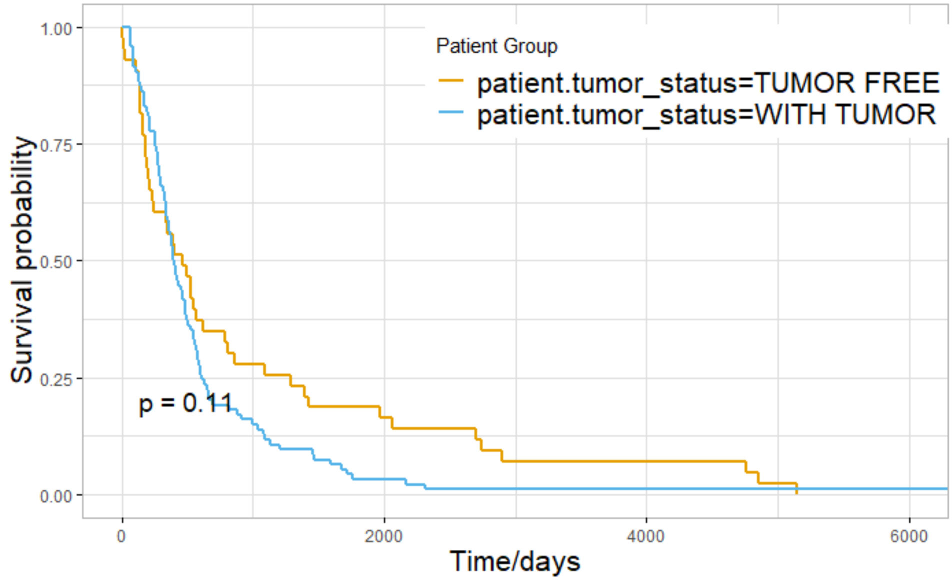


This, however, may be due to the fact that the quartiles span a narrow range in this cohort.

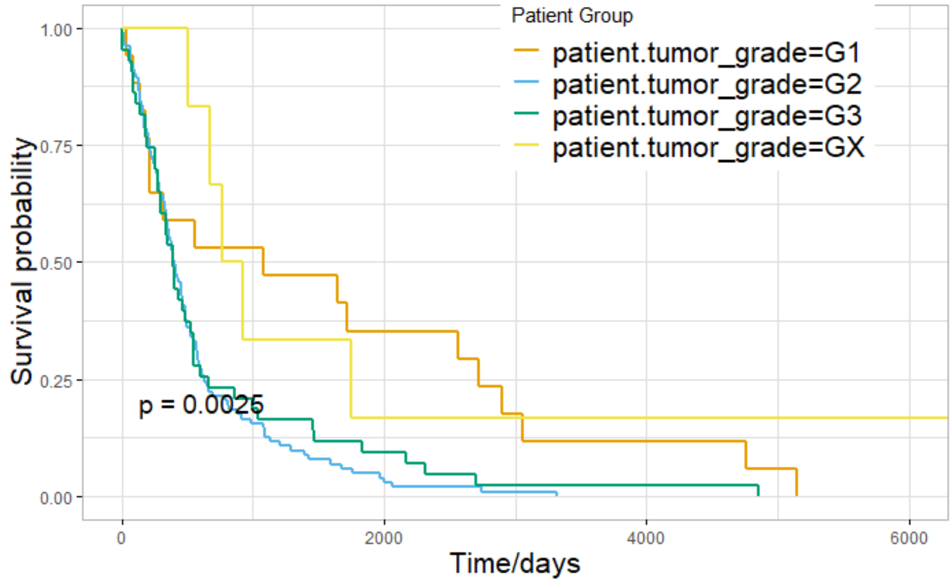
```
summary(clin_sub$patient.age_year)
```

```
##     Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    20.00   53.35   61.14   61.44   69.17   90.06       1
```

Here, we see that the K-M curves separated by tumor status are not quite statistically separable.



However, we do see that the tumor grade produces curves that are statistically different from each other:



In a similar fashion, we would need to investigate many clinical features that may be associated with survival outcome.