Project Report

Abstract

Serous ovarian cancer is the most common type of ovarian cancer. Less than half of all patients with advanced-stage high-grade serous ovarian cancers (HGSC) survive more than five years post-diagnosis, but individuals who can be classified as long-term survivors can provide insight into genetic contributions to long term survival. Using a dataset comprised of 63 patients with advanced-stage, HGSC who survived more than 10 years after diagnosis and 68 short- or moderate-term survivors, we attempted to determine the genetic differences in the two groups. Here we show that HGSC long-term survival observed in our sample data cannot be predicted simply through gene expression counts. Using various bioinformatics analysis techniques to characterize gene expression, we were not able to identify consistent bases for clustering that were proportional to the ratio that samples existed in between our control groups, long-term and non-long-term survivors. In other words, clear groupings or clusters of samples into these groups were not found, suggesting the importance of gene level characteristics in long-term survival; whereas previously, we believed that long-term survival would depend largely presence or absence of particular genes in samples. Our results demonstrate the nuance in genetic expression that characterizes long term survival.

Introduction

1. What data did you use? Provide a general description

The data in this study is comprised of 131 samples, humans with advanced-stage high-grade serous ovarian cancer (HGSC), of which 75 were taken from another study (GSE209964, “Gene count level transcriptomic data of high grade serous ovarian cancer”). 63 of the 131 were long-term survivors of HGSC, meaning that they have survived for more than 10 years after their diagnosis; while 68 samples were short- or moderate-term survivors (moderate-term meaning having survived between 5-10 years after diagnosis, and short-term meaning having survived less than 5 years after diagnosis). These 131 samples are RNA sequences of human, or homo sapien, genes.

2. What question did you seek to answer, and how did you approach that (hint:

assignments 2 and 3 were looking to answer your question).

Our scientific question was: What are the genomic differences between long-term and short- or moderate-term survivors of high-grade serous ovarian cancer? We approached this question using total RNA sequencing (RNA-seq) technique. In approaching the analysis of our data, we used many bioinformatic techniques such as diagrams and various clustering plots and processes. To better understand similarities found in our data, we created a PCA plot and graphed other clustering algorithms, namely K-means, hclust, and ConsensusClusterPlus. These plots separated our data in clusters, which in turn clarified to us any similarities in gene expression regarding long-term and short-/moderate-term survival within the samples; along with occurring extremities in our data. Another plot created that showed the statistical significance of the genes was a volcano plot. The volcano plot helped us quickly identify whether certain gene expressions in long-term survivors are more biologically significant than those found in short- and moderate-term survivors. A heatmap was also created to help us find any changes in gene expression of the different survivors in our sample. We ran enrichment analyses that visibly identified statistically over-represented or under-represented genes within long-term survivors in our sample in comparison to short- and moderate-term survivors. In the end, we performed a chi-squared test to come up with a conclusion regarding our hypothesis. This statistical test compared the results that we observed from our analysis of the data with the results that we expected to occur.

3. Find at least 10 references supporting your hypothesis (hint: look at the original

publication that your data was used in)

Berchuck, Andrew, et al. “Patterns of Gene Expression That Characterize Long-Term Survival in Advanced Stage Serous Ovarian Cancers.” Clinical Cancer Research, vol. 11, no. 10, 16 May 2005, pp. 3686–3696., https://doi.org/10.1158/1078-0432.ccr-04-2398.

Berns, Els M.J.J., and David D. Bowtell. “The Changing View of High-Grade Serous Ovarian Cancer.” Cancer Research, vol. 72, no. 11, 31 May 2012, pp. 2701–2704., https://doi.org/10.1158/0008-5472.can-11-3911.

Bowtell, David D., et al. “Rethinking Ovarian Cancer II: Reducing Mortality from High-Grade Serous Ovarian Cancer.” Nature Reviews Cancer, vol. 15, no. 11, 23 Oct. 2015, pp. 668–679., https://doi.org/10.1038/nrc4019.

Feng, Zheng, et al. “Homologous Recombination Repair Gene Mutations Show No Survival Benefits in Chinese High-Grade Serous Ovarian Cancer Patients.” Annals of Translational Medicine, vol. 9, no. 5, Mar. 2021, pp. 364–364., https://doi.org/10.21037/atm-20-5136.

Hoppenot, Claire, et al. “Who Are the Long-Term Survivors of High Grade Serous Ovarian Cancer?” Gynecologic Oncology, vol. 148, no. 1, Jan. 2018, pp. 204–212., https://doi.org/10.1016/j.ygyno.2017.10.032.

Lancaster, Johnathan M., et al. “Gene Expression Patterns That Characterize Advanced Stage Serous Ovarian Cancers.” Journal of the Society for Gynecologic Investigation, vol. 11, no. 1, Jan. 2004, pp. 51–59., https://doi.org/10.1016/j.jsgi.2003.07.004.

Lisio, Michael-Antony, et al. “High-Grade Serous Ovarian Cancer: Basic Sciences, Clinical and Therapeutic Standpoints.” International Journal of Molecular Sciences, vol. 20, no. 4, 22 Feb. 2019, p. 952., https://doi.org/10.3390/ijms20040952.

Patch, Ann-Marie, et al. “Whole–Genome Characterization of Chemoresistant Ovarian Cancer.” Nature, vol. 521, no. 7553, 28 May 2015, pp. 489–494., https://doi.org/10.1038/nature14410.

Siamakpour-Reihani, Sharareh, et al. “Differential Expression of Immune Related Genes in High-Grade Ovarian Serous Carcinoma.” Gynecologic Oncology, vol. 156, no. 3, Mar. 2020, pp. 662–668., https://doi.org/10.1016/j.ygyno.2019.12.019.

Vilming Elgaaen, Bente, et al. “Global MIRNA Expression Analysis of Serous and Clear Cell Ovarian Carcinomas Identifies Differentially Expressed Mirnas Including Mir-200c-3P as a Prognostic Marker.” BMC Cancer, vol. 14, no. 1, 11 Feb. 2014, https://doi.org/10.1186/1471-2407-14-80.

Wilkerson, Matthew D., and D. Neil Hayes. "ConsensusClusterPlus: a class discovery tool with confidence assessments and item tracking." *Bioinformatics* 26.12 (2010): 1572-1573.

Yoshihara, Kosuke, et al. “Gene Expression Profile for Predicting Survival in Advanced-Stage Serous Ovarian Cancer across Two Independent Datasets.” PLoS ONE, vol. 5, no. 3, 12 Mar. 2010, https://doi.org/10.1371/journal.pone.0009615.

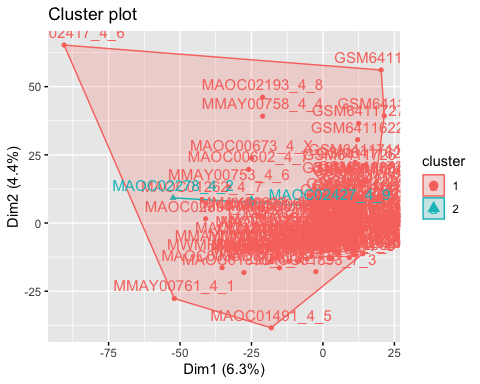
Methods

1.Create a bullet point for each method you used in all assignments. Group these by type

rather than by assignment. For example, only make 1 section for each enrichment

analysis method and note ALL applications of this method in your assignments.

​​​

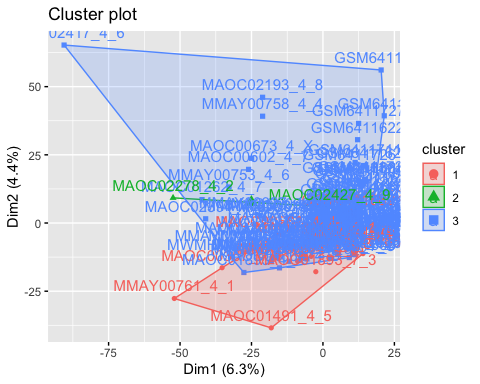


k-means k=2 plot

After creating a separate group consisting of the 5000 most variable genes from our metadata, we ran three clustering algorithms (k-means, hclust, and consensusclusterplus) that required us to select the number of clusters (k) to find. For each clustering algorithm, we tested out 3 different k values (k=2, 3, & 4).

The above graph shows the clustering found by the k-means algorithm using k=2 clusters. You can see that the 2 cluster groups are nested, meaning that one cluster is inside the other. This was interesting because we expected the 2 clusters to be separated from one another.

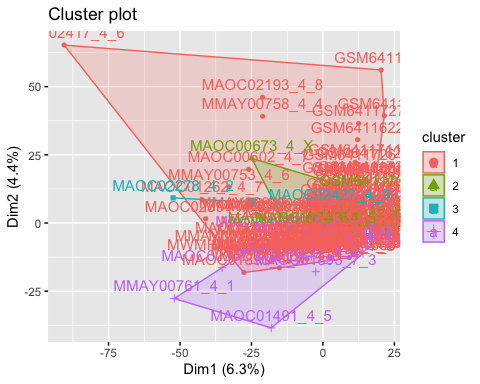
fviz\_cluster(k3, V5000)



k-means k=3 plot

This cluster plot shows the clustering found by the k-means algorithm using k=3 clusters. You can see that the nesting of clusters 2 and 3 remain, but what makes this plot different than the k=2 plot is that cluster 1 is just slightly overlapping cluster 3, with the majority of it being separated.

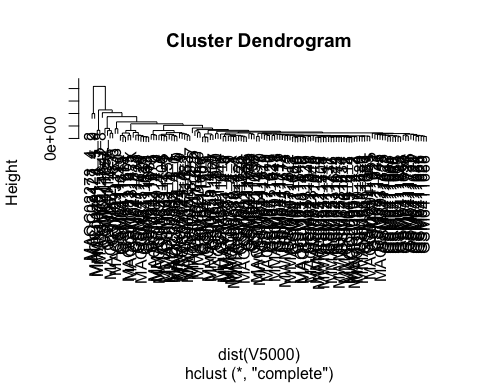
fviz\_cluster(k4, V5000)



k-means k=4 plot

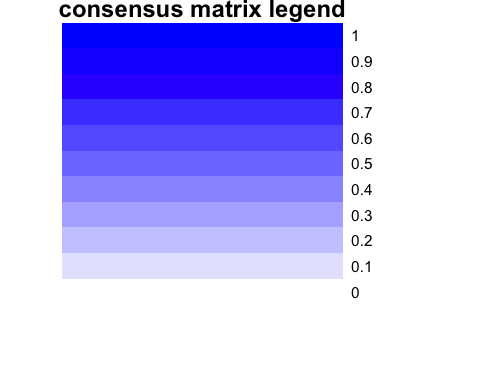
This plot shows the clustering found by the k-means algorithm using k=4 clusters. You can see that clusters 2 and 3 are fully nested within cluster 1. This plot has a 4th cluster that is again mainly separated but has a slight overlap onto cluster 1.

***##hierarchical clustering***  
library(stats)  
  
hclust <- hclust(dist(V5000), method = "complete", members = NULL)  
plot(hclust)

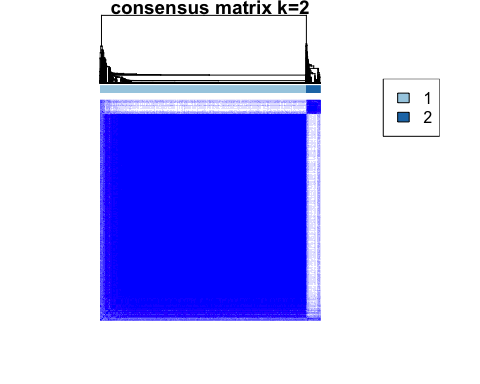


hclust

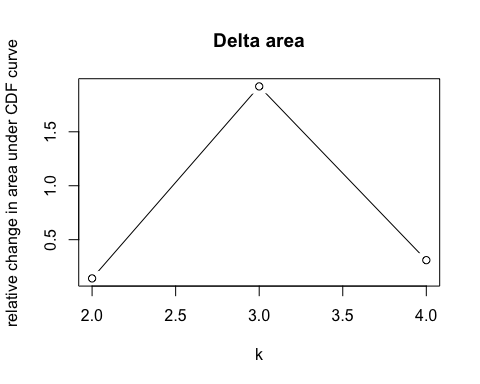
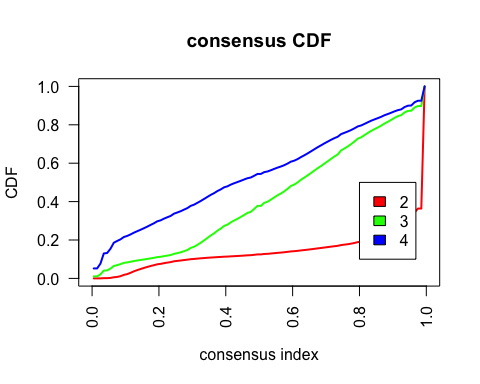
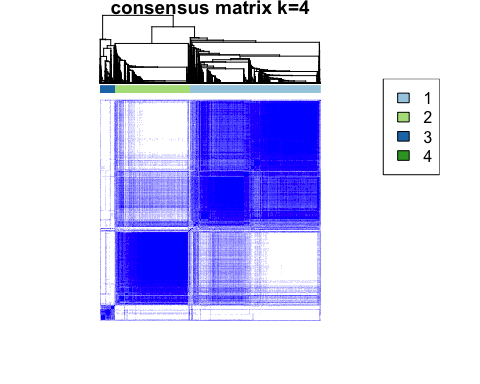
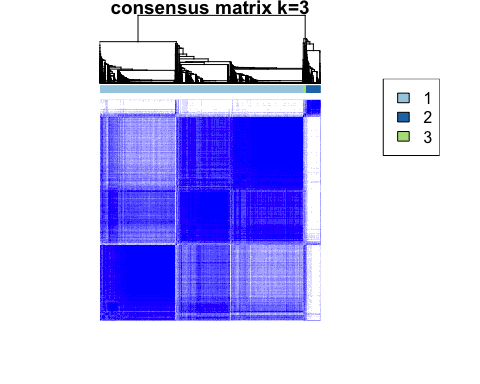
This plot shows the clustering found using the hclust algorithm



## clustered



## clustered

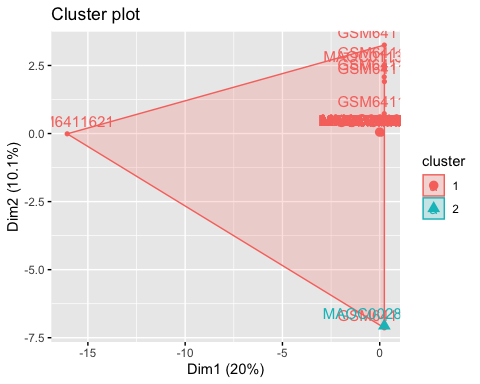
 ##d Using kmeans and an initial cluster number of 2, one small cluster is nested within a cluster that was much larger in terms of area and membership. With k=3, the small nested cluster persists, but the larger cluster is split into two. As in the previous case, there is a cluster with a membership that is much larger than the other clusters. The same pattern can be oberved for k=4.

The hierarchical clustering function doesn’t appear to have chosen a k or been determined by any input for k. Instead it llooks like the clustering resolved down to each individual sample while providing a metric for cluster proximity with its height axis. In this case, the three or four primary clusters still maintain an extreme inequality in terms of sample membership, similar to our kmeans clustering results.

Using consensus clustering, our delta area graph no appreciable difference in consensus beyond k-3. The consensus clustering graph shows a majority of membership in one group with a smaller group and an even smaller group next to that. With an increase to a cluster number of 4, the appearance of a larger group emerges. This group is possibly analogous to the nested group that appeared in the k=4 kmeans clustering.

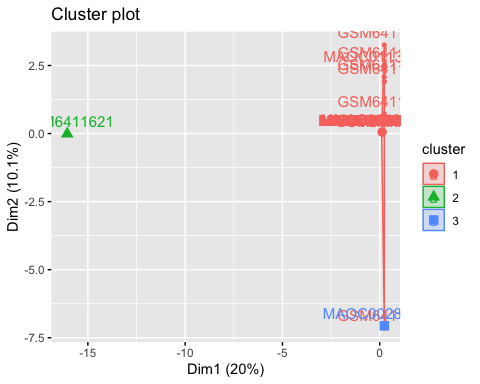
In each of the methods, the proportion of cluster membership does not seem to reflect that of the control groups for our experiment, which is roughly half and half.

##(e) rerun each clustering method for 10, 100, 1000 and 10000 genes



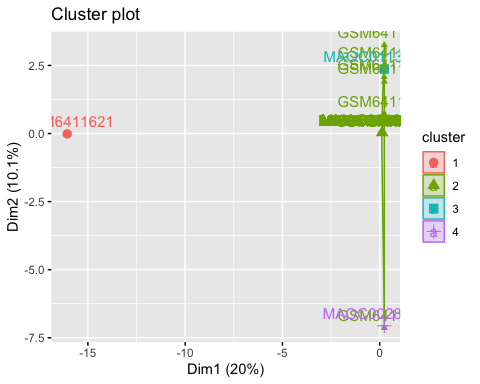
Kmeans plot with k=2. The second cluster is a single sample. We’re not sure why this wouldn’t cluster with the point its closest to in the red cluster. The area the larger cluster covers suggests its uniformly spread but really there are only two samples within it that stray from a dense patch in the upper right corner.

fviz\_cluster(k3\_10, V10)

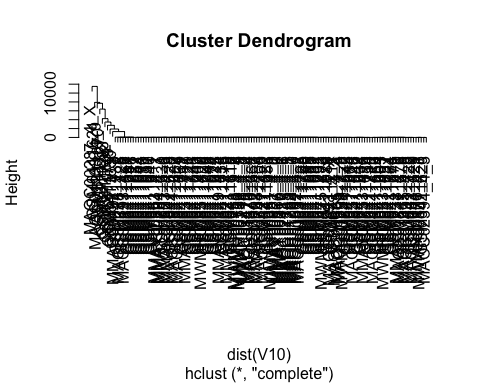


Our kmeans graph with k=3. The outlier on the left in the larger cluster forms the new cluster and maintains similar characteristics to the previous plot with k=2.

fviz\_cluster(k4\_10, V10)



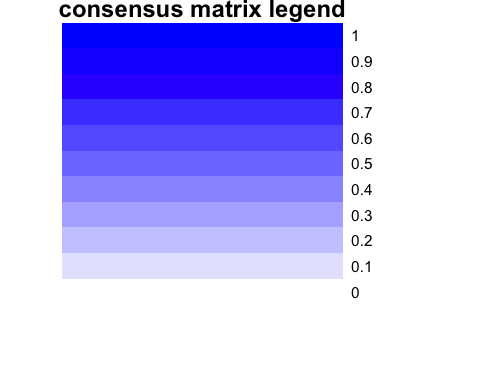
hclust10 <- hclust(dist(V10), method = "complete", members = NULL)  
plot(hclust10)



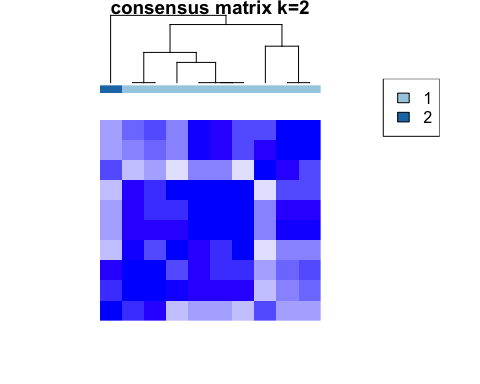
v10matrix <-data.matrix(V10, rownames.force = NA)  
results10 <- ConsensusClusterPlus(v10matrix,maxK=4,reps=50,pItem=0.8,pFeature=1,title='title',clusterAlg="hc",distance="pearson",plot="screen")

## end fraction

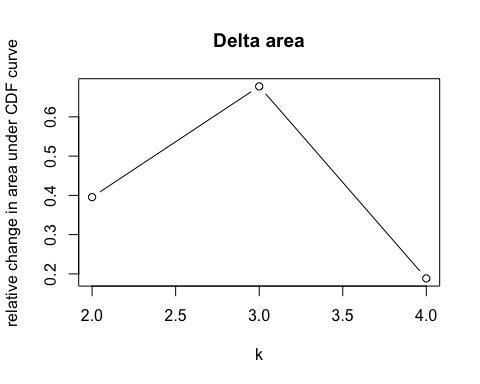
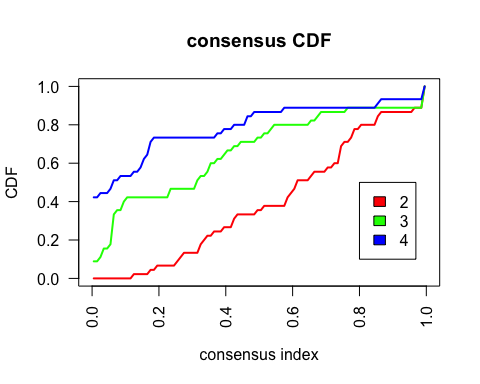
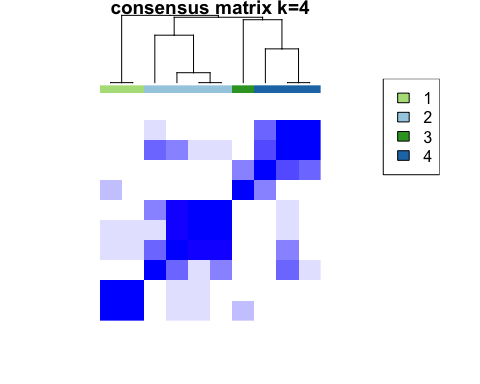
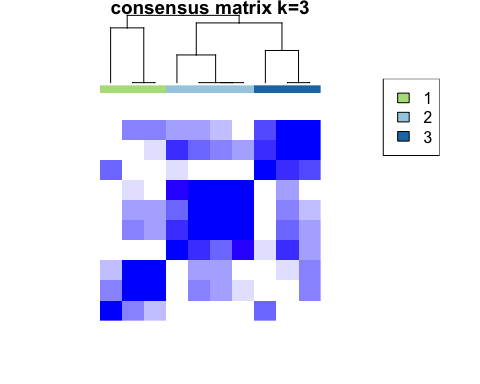
## clustered



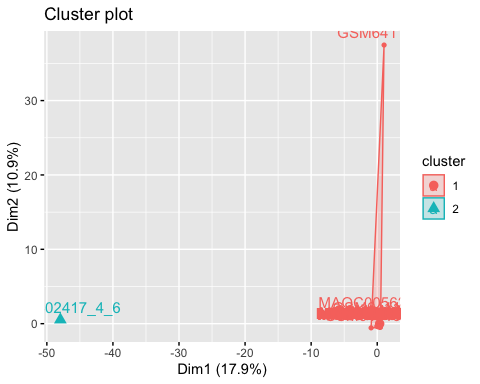
## clustered



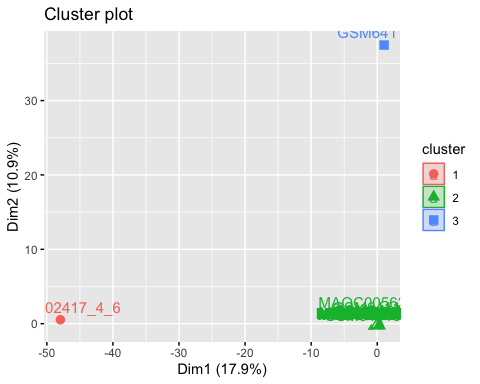
## clustered



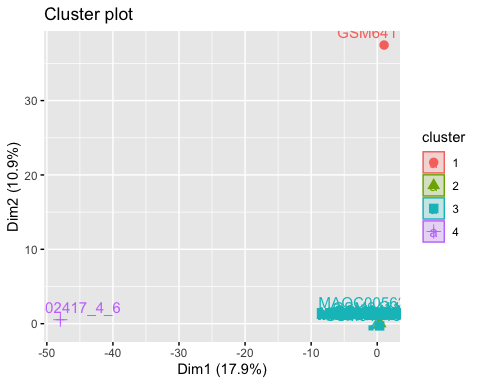
***##100***  
k2\_100 <- kmeans(V100, centers = 2)  
k3\_100 <- kmeans(V100, centers = 3)  
k4\_100 <- kmeans(V100, centers = 4)  
  
fviz\_cluster(k2\_100, V100)



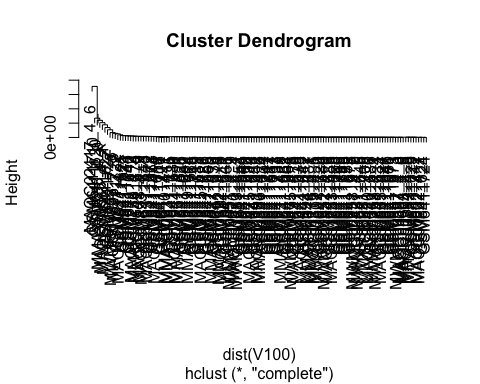
fviz\_cluster(k3\_100, V100)



fviz\_cluster(k4\_100, V100)



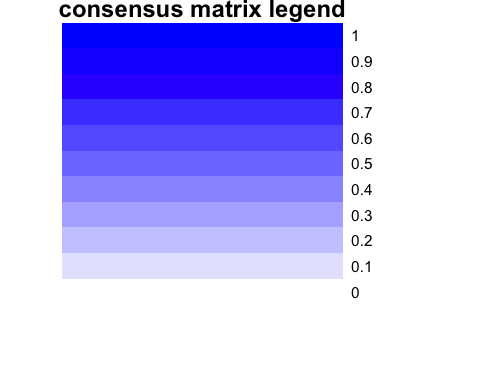
hclust100 <- hclust(dist(V100), method = "complete", members = NULL)  
plot(hclust100)



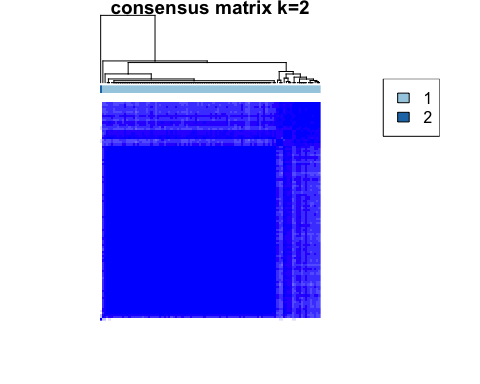
v100matrix <-data.matrix(V100, rownames.force = NA)  
results100 <- ConsensusClusterPlus(v100matrix,maxK=4,reps=50,pItem=0.8,pFeature=1,title='title',clusterAlg="hc",distance="pearson",plot="screen")

## end fraction

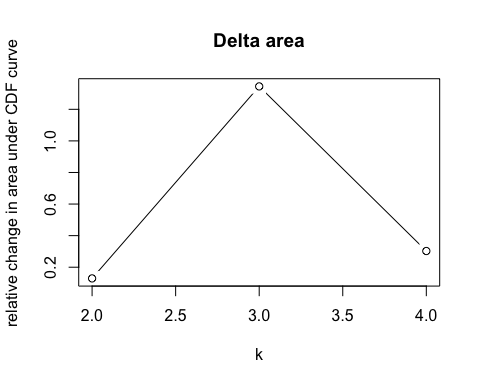
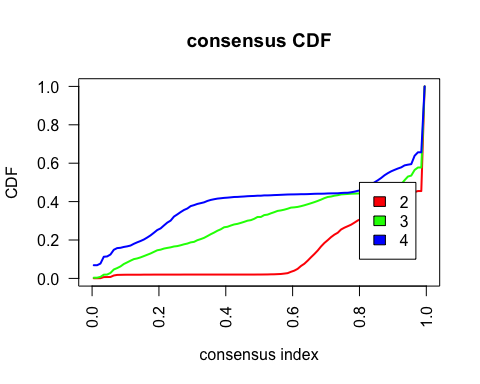
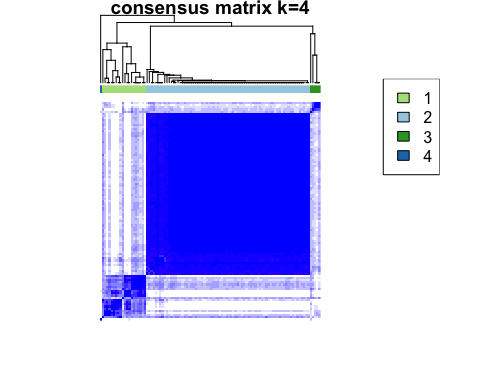
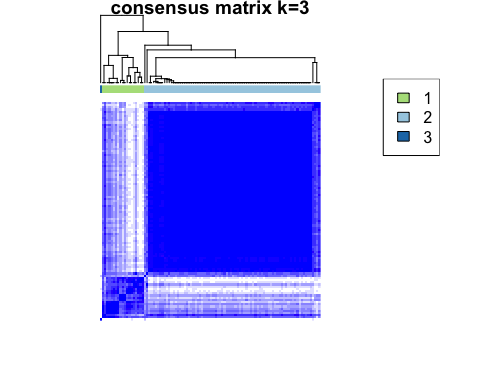
## clustered



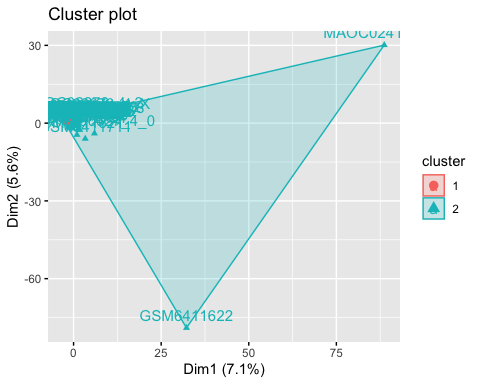
## clustered



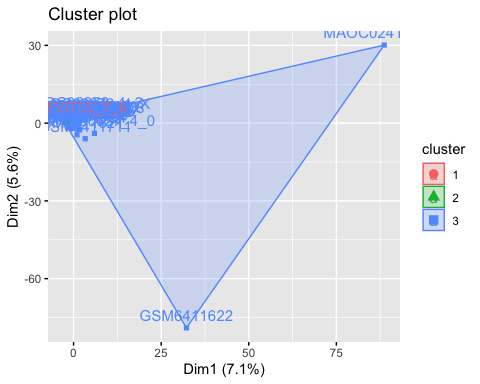
## clustered



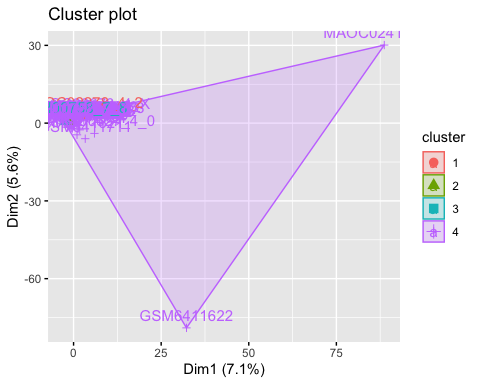
*#1000*  
k2\_1000 <- kmeans(V1000, centers = 2)  
k3\_1000 <- kmeans(V1000, centers = 3)  
k4\_1000 <- kmeans(V1000, centers = 4)  
  
fviz\_cluster(k2\_1000, V1000)



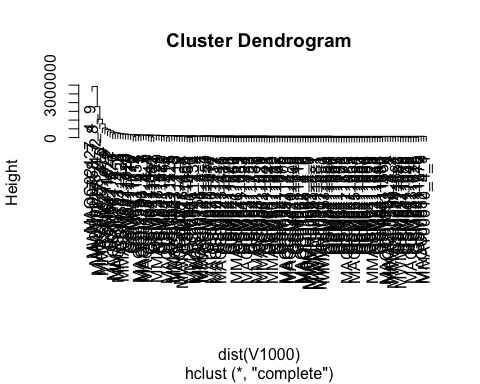
fviz\_cluster(k3\_1000, V1000)



fviz\_cluster(k4\_1000, V1000)



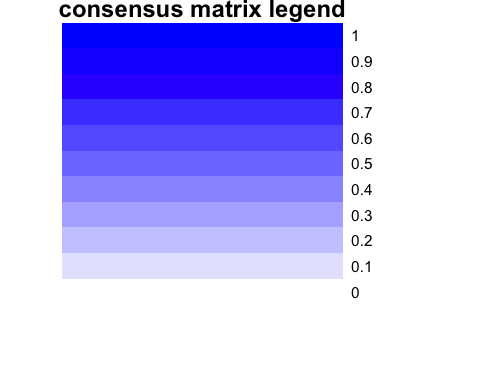
hclust1000 <- hclust(dist(V1000), method = "complete", members = NULL)  
plot(hclust1000)



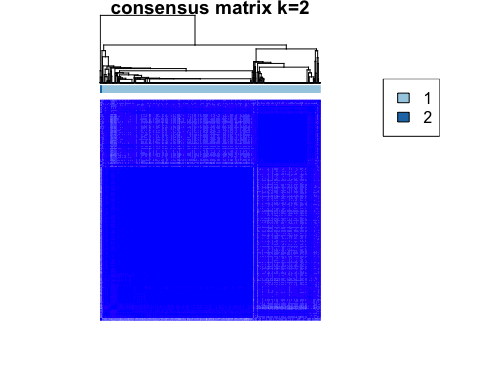
v1000matrix <-data.matrix(V1000, rownames.force = NA)  
results1000 <- ConsensusClusterPlus(v1000matrix,maxK=4,reps=50,pItem=0.8,pFeature=1,title='title',clusterAlg="hc",distance="pearson",plot="screen")

## end fraction

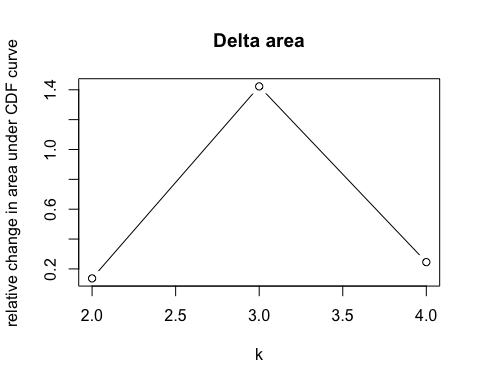
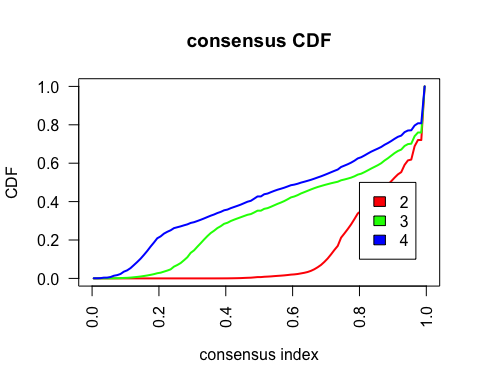
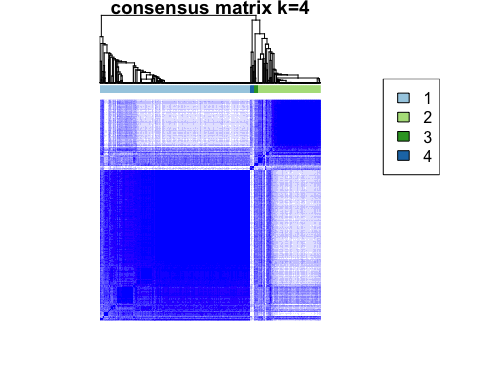
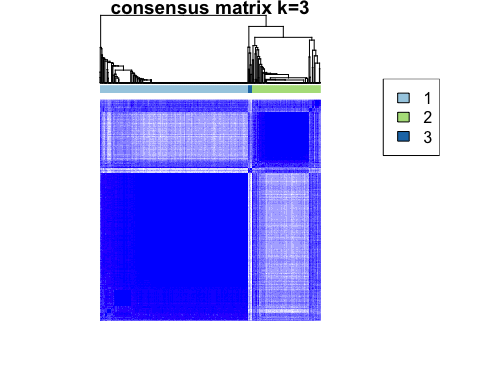
## clustered



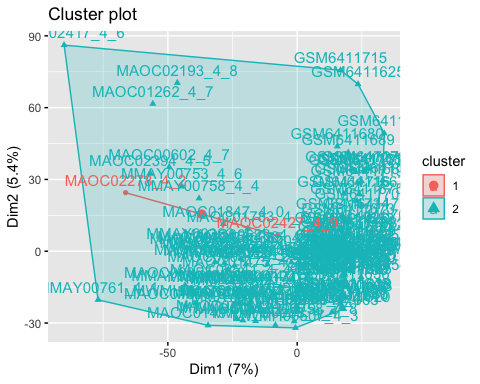
## clustered



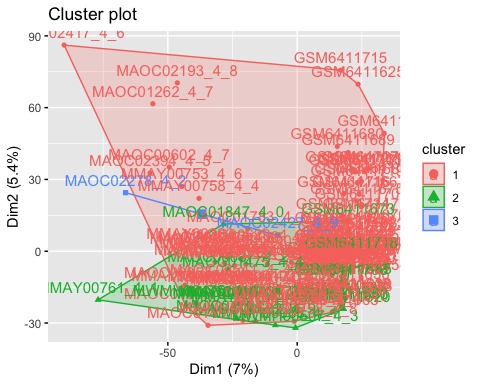
## clustered



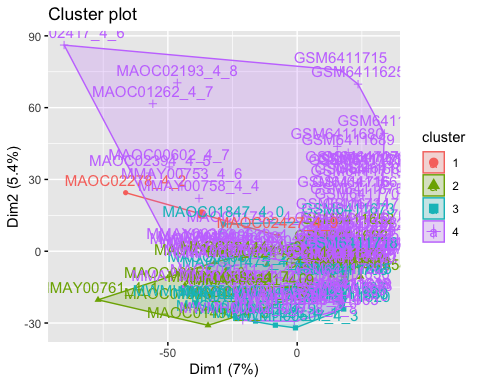
*#10000*  
k2\_10000 <- kmeans(V10000, centers = 2)  
k3\_10000 <- kmeans(V10000, centers = 3)  
k4\_10000 <- kmeans(V10000, centers = 4)  
  
fviz\_cluster(k2\_10000, V10000)



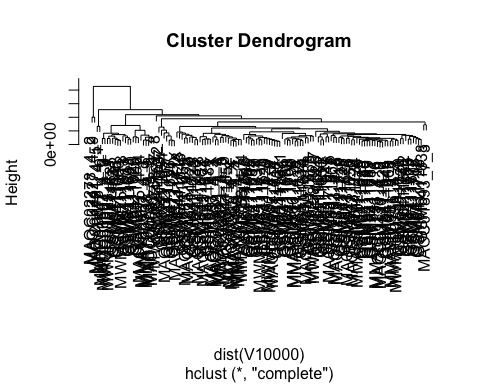
fviz\_cluster(k3\_10000, V10000)



fviz\_cluster(k4\_10000, V10000)



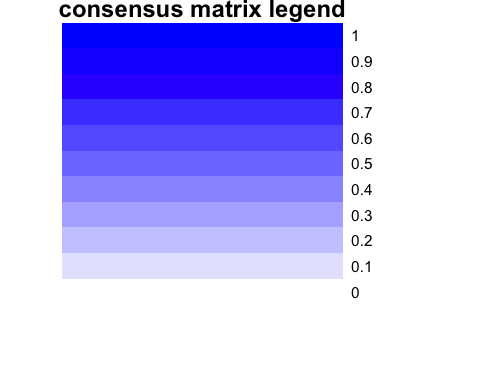
hclust10000 <- hclust(dist(V10000), method = "complete", members = NULL)  
plot(hclust10000)



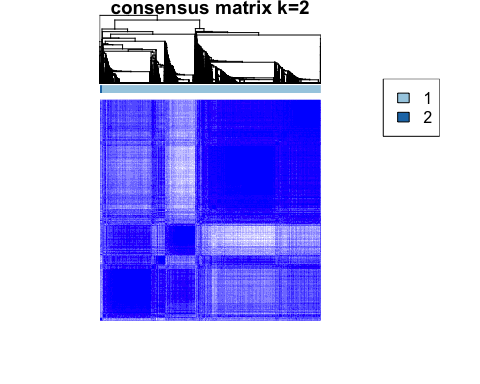
v10000matrix <-data.matrix(V10000, rownames.force = NA)  
results10000 <- ConsensusClusterPlus(v10000matrix,maxK=4,reps=50,pItem=0.8,pFeature=1,title='title',clusterAlg="hc",distance="pearson",plot="screen")

## end fraction

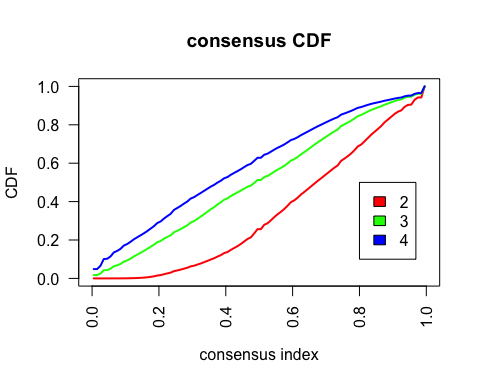
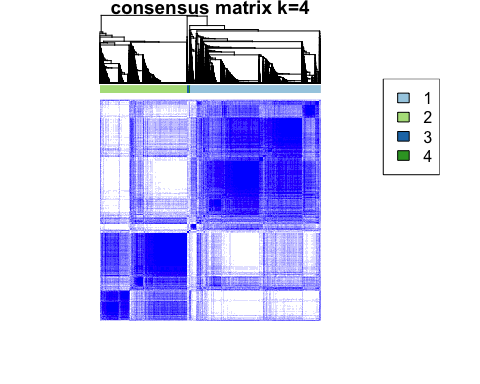
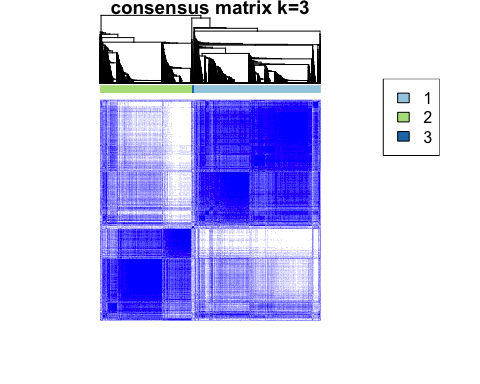
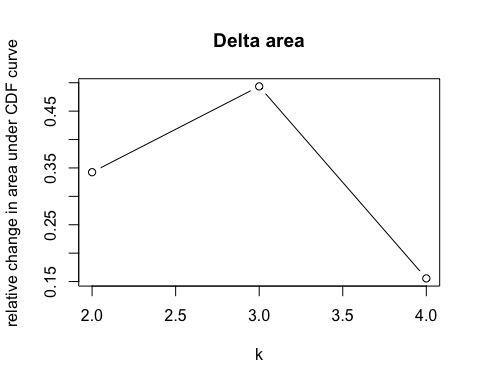
## clustered



## clustered



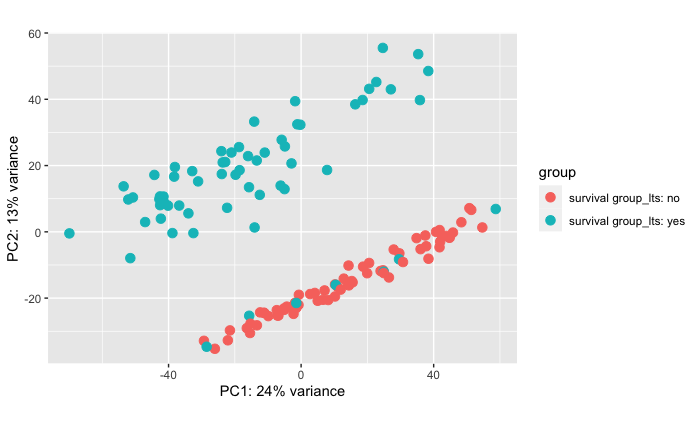
## clustered

 #

***##breaks w/5000 genes, so we used 1000 instead***

Diagram

Description automatically generated



2.Github Repository: <https://github.com/Drod0317/BioinformaticsProject.git>

Results

1. Were you able to answer your original question? Add 1-3 bullet points of why/why not

We were not able to answer our original question.

* The analysis methods that we dealt with all assessed gene level expressions and not differential expression in allele level alterations, which the study that we sourced our data from referenced as being relevant to the question of long-term vs. short-term survival.
* Gene level expression did not provide clear delineation for clustering that resulted in identification/separation of control groups.

2. What else did you find? Add 1-3 bullets of different unexpected findings

* There was a subgroup of long-term survivors that were more similar to short-term survivors in our PCA plot

Conclusion

1. Write a draft of your conclusion. This section should reiterate your hypothesis

(question from assignment 1) and interpret your findings from each assignment.

While there is genetic difference between long-term and short-term survivors of High Grade Serous Ovarian Cancer (HGSC), the differences that delineate either group are not observed at the gene expression level. The clustering analysis we conducted did not return clusters that corresponded to our control groups, suggesting gene count is not the primary driver of long-term survival of HGSC.