What are the genomic differences between long-term and short- or moderate-term survivors of high-grade serous ovarian cancer?

Contributors: Seraj Saboungi, Dillan Maraj, and Devin Rodriguez

**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. Individuals who do survive beyond this time, classified as long-term survivors, can help provide insight into genetic contributions to long term survival. Using a dataset composed of 63 patients with advanced-stage HGSC who survived more than 10 years after diagnosis and 68 short- or moderate-term (non-long term) survivors, we attempted to identify the genetic differences between the two groups to determine strong indicators for long-term survival of HGSC patients. Here we are unable to show that long-term survival of HGSC observed can be predicted through gene expression counts. Using various bioinformatics analysis techniques to characterize gene expression, we were not able to identify distinct differences between our control groups, and instead observed possible sample bias. Clusters formed in the process of our analysis did not align according to control groups. Previously, we believed that long-term survival would depend largely on the consistent presence or absence of particular genes in samples, but our source paper suggested alterations within genes were responsible for the differences. Our results demonstrate the importance of sample normalization in forming a verifiable conclusion.

**Introduction**

Ovarian cancer is one of the most common cancers to develop in women, with high grade serous ovarian cancer, or HGSC, especially proving to be a particularly dangerous form of the disease as it is responsible for around 70-80% of ovarian cancer deaths (2, 3, 4, 6, 7). Unfortunately, over the past two decades, we have seen little improvement in the survival rate of HGSC patients (2, 8). On average, the survival of HGSC patients is less than 5 years after diagnosis, specifically about 3-4 years; while long term survival is considered to be more than 10 years (1, 5).

We examined 68 long term survivors of HGSC and 63 non-long term survivors, whose gene samples were provided by GSE211669, an experiment studying the genomic landscape of long-term survivors of HGSC found in the GEO database, to see if there were any genetic differences between the two groups. This was done by applying various analysis methods such as clustering and differential expression at the gene count level.

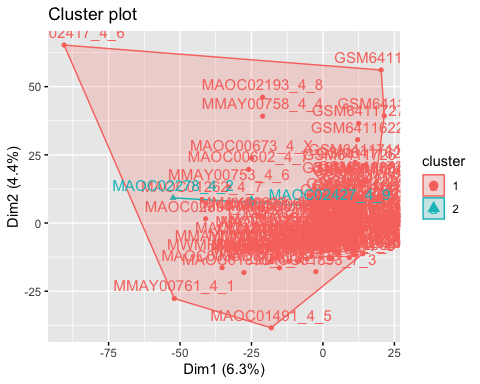
Upon completion of our experiment, we found that the analysis methods that we dealt with all assessed gene level expressions and not differential expression in allele level alterations; which the study that we sourced our data from referenced as being relevant to the question of long-term vs. short-term survival. This means that, unfortunately, the gene level expression did not provide clear delineation for clustering that resulted in identification/separation of control groups; not allowing us to answer our original question.

Treatment of HGSC is usually “non-individualized” and does not vary among cases; the treatment plan usually consists of undergoing surgery to remove the malignant tumor(s) before going through with platinum-based chemotherapy, a common form of chemotherapy used to treat cancer (9, 10). We hoped that the results of our experiment would aid in the development of specialized treatment for HGSC patients; however, we were disappointed to find that our results did not reflect any conclusive findings.

**Methods**

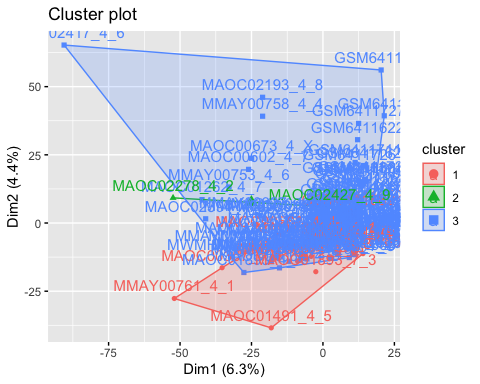
Clustering

We used three clustering algorithms, K-means, hclust, and ConsensusClusterPlus, to separate our data in clusters, which in turn clarified to us any similarities in gene level expression regarding long-term and short-/moderate-term survival within the samples; along with occurring extremities in our data.

k-means k=2 plot

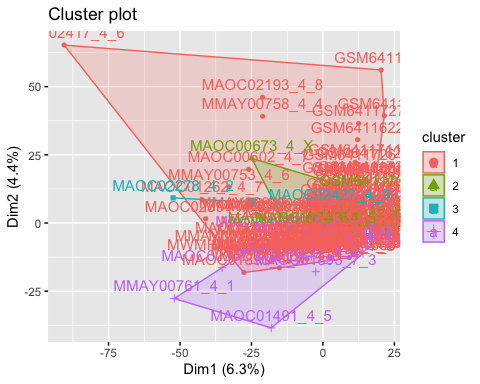
After creating a separate group consisting of the 5000 most variable genes from our metadata, we ran three clustering algorithms (k-means, hclust, and consensusclusterplus) that required us to select the number of clusters (k) to find. For each clustering algorithm, we tested out 3 different k values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

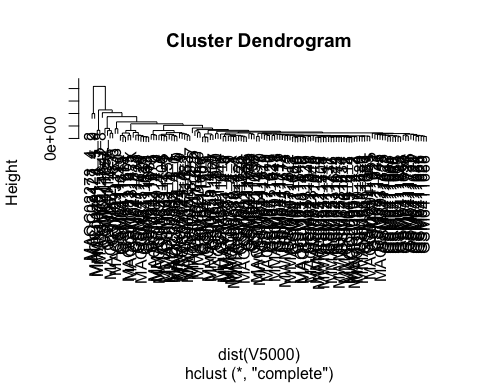
k-means k=3 plot

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.

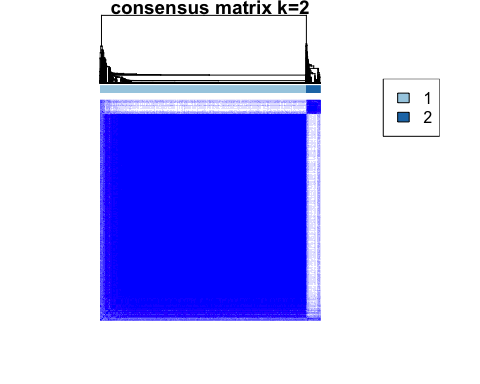
k-means k=4 plot



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.

hclust

This plot shows the clustering found using the hclust algorithm.

consensusclusterplus k=2

Graphical user interface

Description automatically generated with low confidenceconsensusclusterplus k=3

Graphical user interface

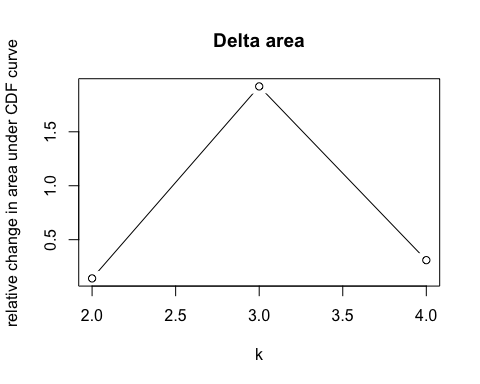
Description automatically generated with medium confidenceconsensusclusterplus k=4

Chart, line chart

Description automatically generated

Using k-means 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.

Differential Expression

We also ran differential expression analysis in our experiments, creating a PCA plot, alluvial diagram, volcano plot, and heatmap. We used a PCA plot to understand the similarities that can be found within our data. From our PCA plot we discovered a subgroup of long-term survivors that were more like short-term survivors then other long-term survivors. The alluvial diagram we created showed us the effect of using different numbers of genes (10, 100, 1000, & 10000) on cluster membership using K-means and ConsensusClusterPlus clustering algorithms.

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. The heatmap helped us find any changes in gene expression of the different survivors in our sample.

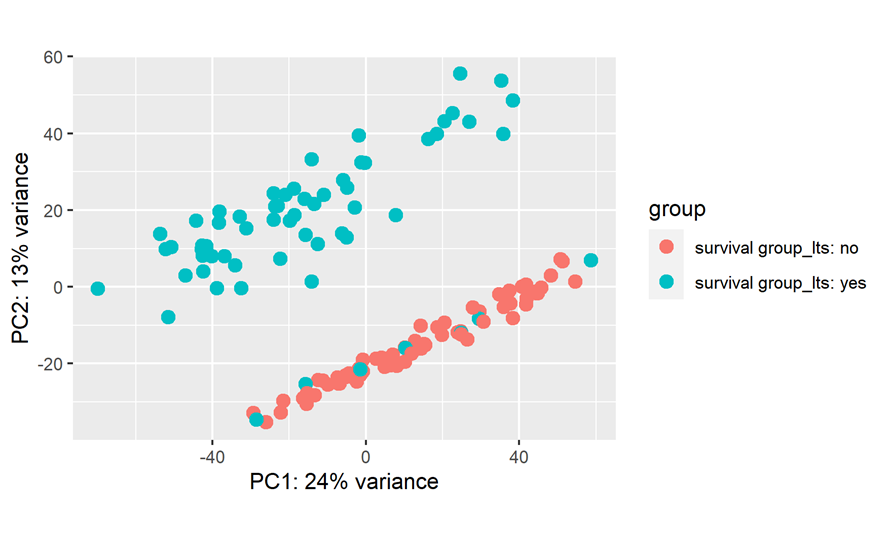


Figure 1 PCA Plot

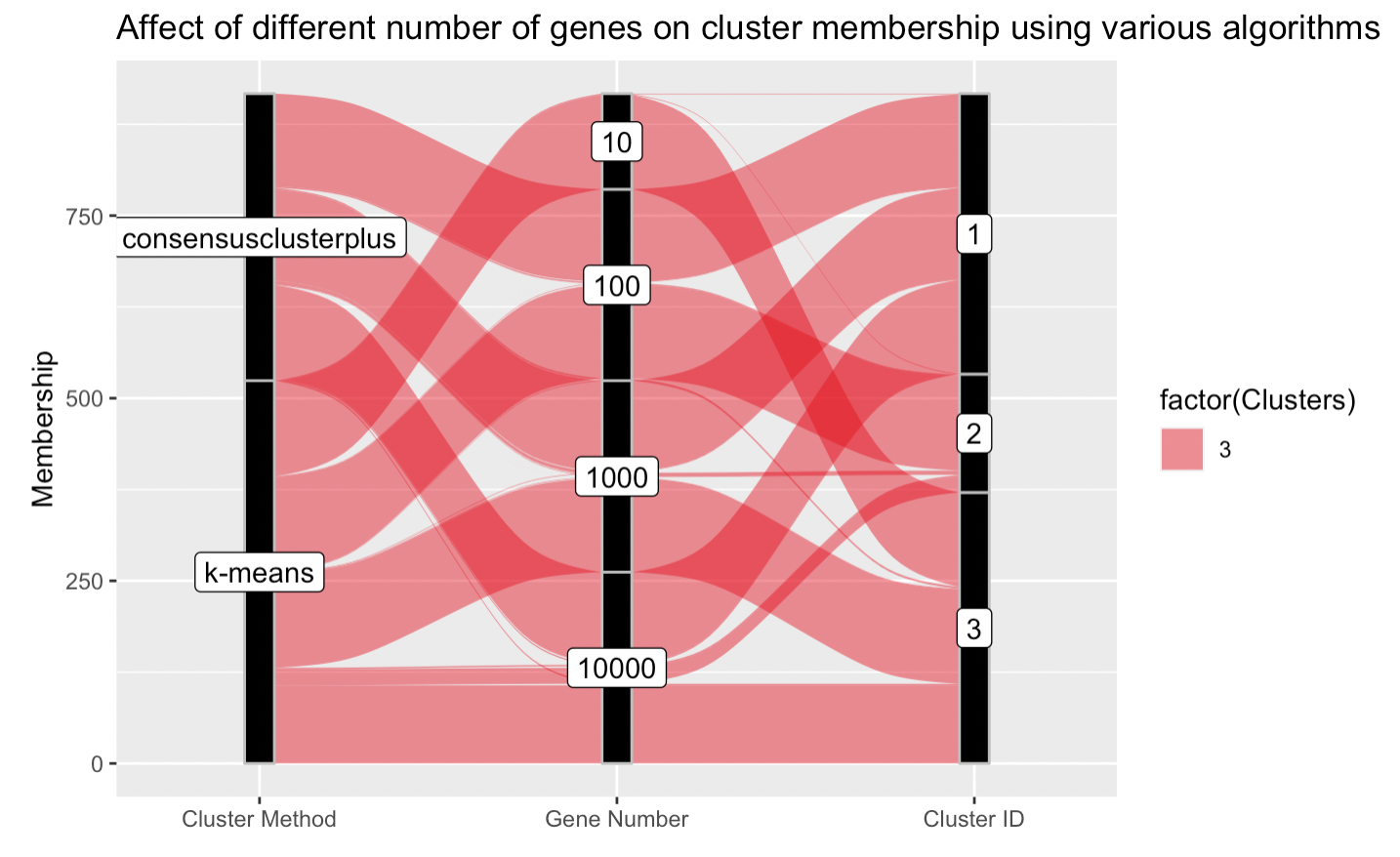


Figure 2 Alluvial Diagram

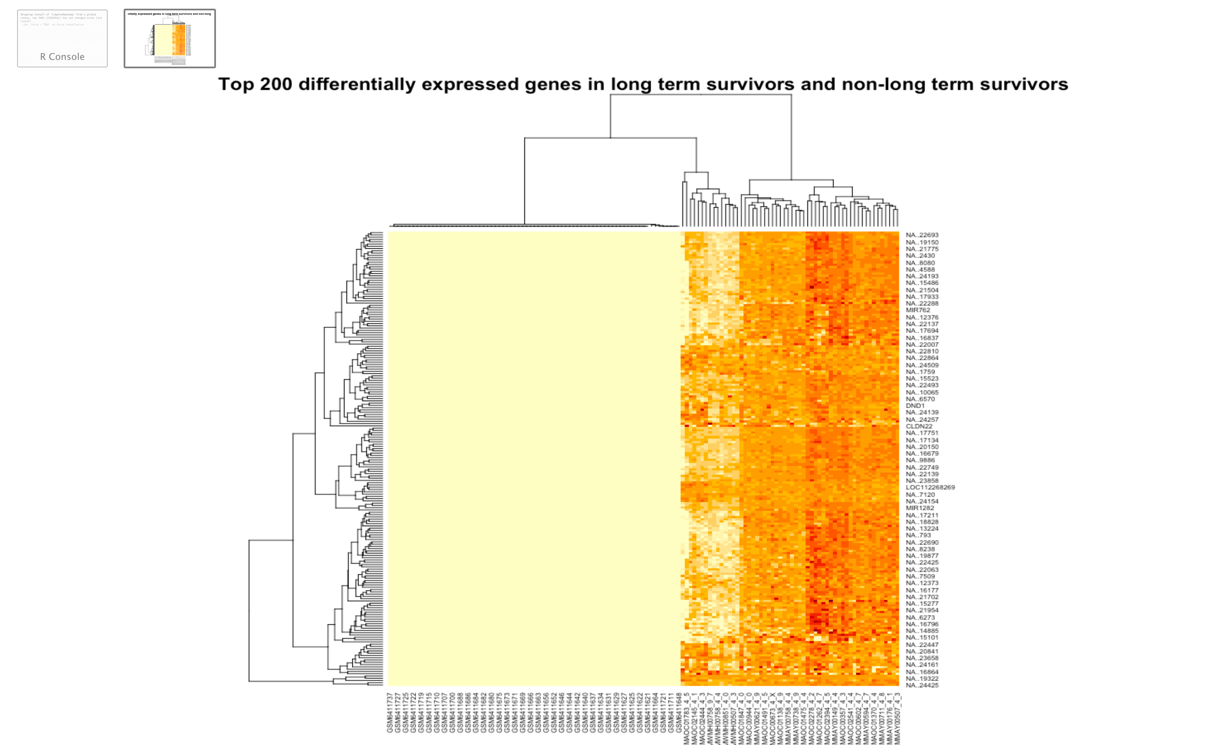


Figure 3 Heatmap

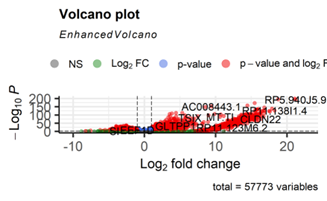


Figure 4 Volcano Plot

**Results**

Methods

**Clustering** – Our clustering results suggested weak grouping characteristics between our

two control groups. The results did not depict clear bases for distinction between the

two with regard to gene counts.

**Differential Expression** – While our differential expression analysis did produce two

distinct groups, those groups only partially represented our control groups, similar to

our PCA plot.

In the volcano plot, we are comparing statistical significance, or the p-value, (which is the y axis) to the fold change, or the magnitude/amount of change, (the x axis) to show us which of the 57,000+ gene expressions simultaneously have large fold changes and are also statistically significant. The gene expressions which fall in that category, meaning having large fold change and high statistical significance, are the more biologically significant genes, and can be found on the top right of the graph.

The heatmap exhibits a clear distinction between two primary groups of samples on the basis of the top 200 differential expressed genes in our data. However, sample membership in this chart does not coincide with the membership of our two control groups, long-term and non-long term survivors. The group on the right, with sample title beginning with M, is composed entirely of long term survivors. But the left group also contains long term survivors as well, although that would not be able to be inferred by the heatmap. The graph also seems to confirm the results of our PCA plot, which sees a small group in the long term survival group separate into a group sharing characteristics with the non-long term survival group as evidenced by our sample dendrogram.

**GSEA** – not completed

**Statistics** – not completed

Success

We were successful in determining that gene count is not the primary driver of long-term survival of HGSC. This result means that we can now move our attention and look for other contributors to long-term survival of HGSC patients.

Weakness

We were unable to find a way to effectively convert the gene identifiers, given in an augmented ensemble format, to HUGO gene names. Simply extracting Ensemble ID’s returned either an NA gene or multiple gene names. This issue cascaded into the necessary analysis techniques we were expected to use, namely Gene Set Enrichment Analysis, which relied on the presence of a single, defined gene name.

We did not give ourselves a sufficient amount of time to pursue the possibility that sample bias may be present, as evidenced by the strong difference in our heatmap data between the two separate sample groups in our data and not the two separate control groups.

Bioethics

Since our data came from an open source database we avoided the bioethical issue of privatizing research.

Future Work

If we were to continue working on this research in the future, we would want to do differential expression at an allele count level instead of a gene count level like we have done so far in this experiment. The study that our samples were pulled from found conclusive results from working on an allele level. We would like to find out if the methods we used in this experiment would produce more conclusive results if we performed them on the allele level.

**Conclusion**

Our results unfortunately did not answer our original question. 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 level which we conducted all the analysis methods of clustering and differential expression. 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. The differential expression analysis showed us that there was indeed a small group within the long term survivor sample that shared characteristics with short and moderate term survivors.

While we performed our experiment at the gene expression level, it would be better to analyze the data in allele level alterations; as allele alterations are located inside the gene, and are what the original GEO experiment suggested were the cause in the difference in HGSC survival.

**References**

(1) 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.

(2) 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.

(3) 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.

(4) 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.

(5) 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.

(6) 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.

(7) 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.

(8) 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.

(9) 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.

(10) 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.