Assignment 3: Gene Expression Analysis

Names: Liam Lalonde, Syed Shahzaib Ali

BINF*6970

28th March 2021

Problem 1: Tissue Gene Expression Analysis

Figure 2 contains 2 plots of the principal component analysis where data set points were identified by both tissue type and the tissue status. The first plot is of unscaled data which shows formation of clusters is clear where we have the formation of 3 clusters. The first cluster only includes liver tissues. The second cluster involves 4 different tissue types: kidney, placenta, colon and endometrium, and the third clusters includes: cerebellum and hippocampus. The formation of clusters is clear especially within the first and third cluster as kidney cells have been grouped with no confusion in discrimination and the cerebellum and hippocampus have been grouped as they are both parts of the brain. The second plot includes scaled data where the kidney tissues have been grouped with the liver tissue samples along with different kinds of tissue forming a very unclear cluster.

There does not seem to be abnormalities with the first(liver) and third cluster(brain) in the unscaled plot in-terms of the patient condition as the tissues from normal patients are grouped and do not compose of any gene expression values that belong to tumor tissues of the same type. What is interesting is that the expression of genes in the normal kidney tissue seems to resemble very much the expression of genes in the first trimester of placenta organs and cancerous colon tissue. Because PC1 accounts for 33.04% of variance if we consider the difference between cluster 1 and cluster2, and cluster2 and cluster3, the tissues involved in-terms of gene expression are fundamentally very different. This is also bringing to my next point that the considering PC1 cluster1 and cluster2 are very much close. With interest particularly on normal kidney cells and normal liver cells, you might observe these data points are closer thus concluding that gene expression in kidney tissue is much similar regarding gene expression in liver tissue. It is also of no surprise considering the constant communication the kidney and liver are in, within the complex organism's body.

Given the PCA plot in figure 3 where the samples were considered as variables. There seem to no clear discrimination between groups or the formation of clusters. All the data points seem to be clumped into one giant cluster. Even though a large amount of variance is explained by PC1 84.69% as compared to PCA plots in figure 2 the pattern of gene expression is not discriminated especially if gene expression of different tissues from different patients' status is involved.

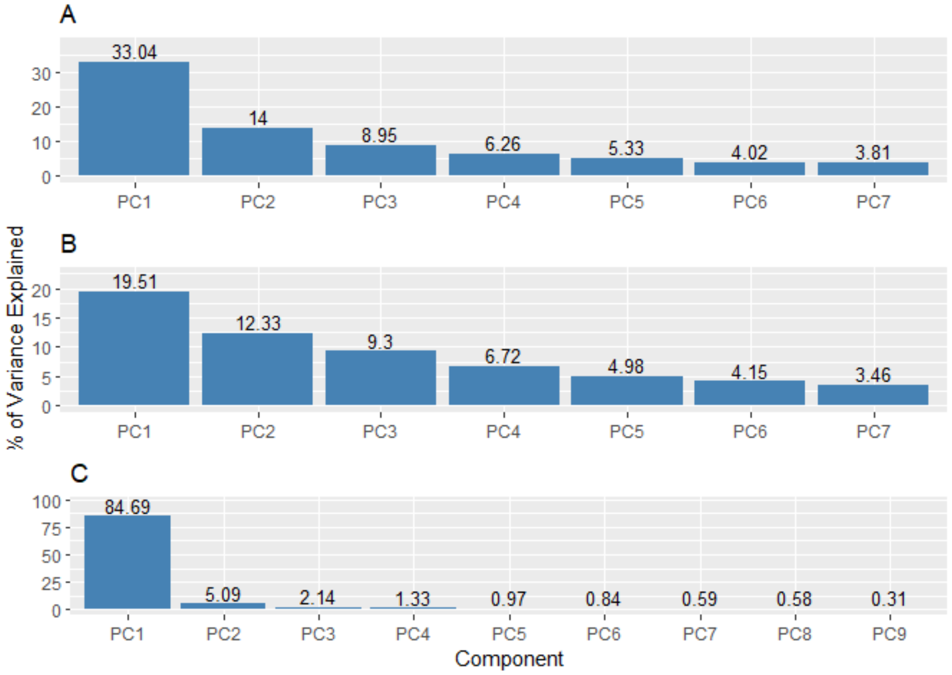


Fig 1. Scree plot showing variance explained by each PC for (A) Unscaled, and (B) Scaled. (C) When samples were considered variables.

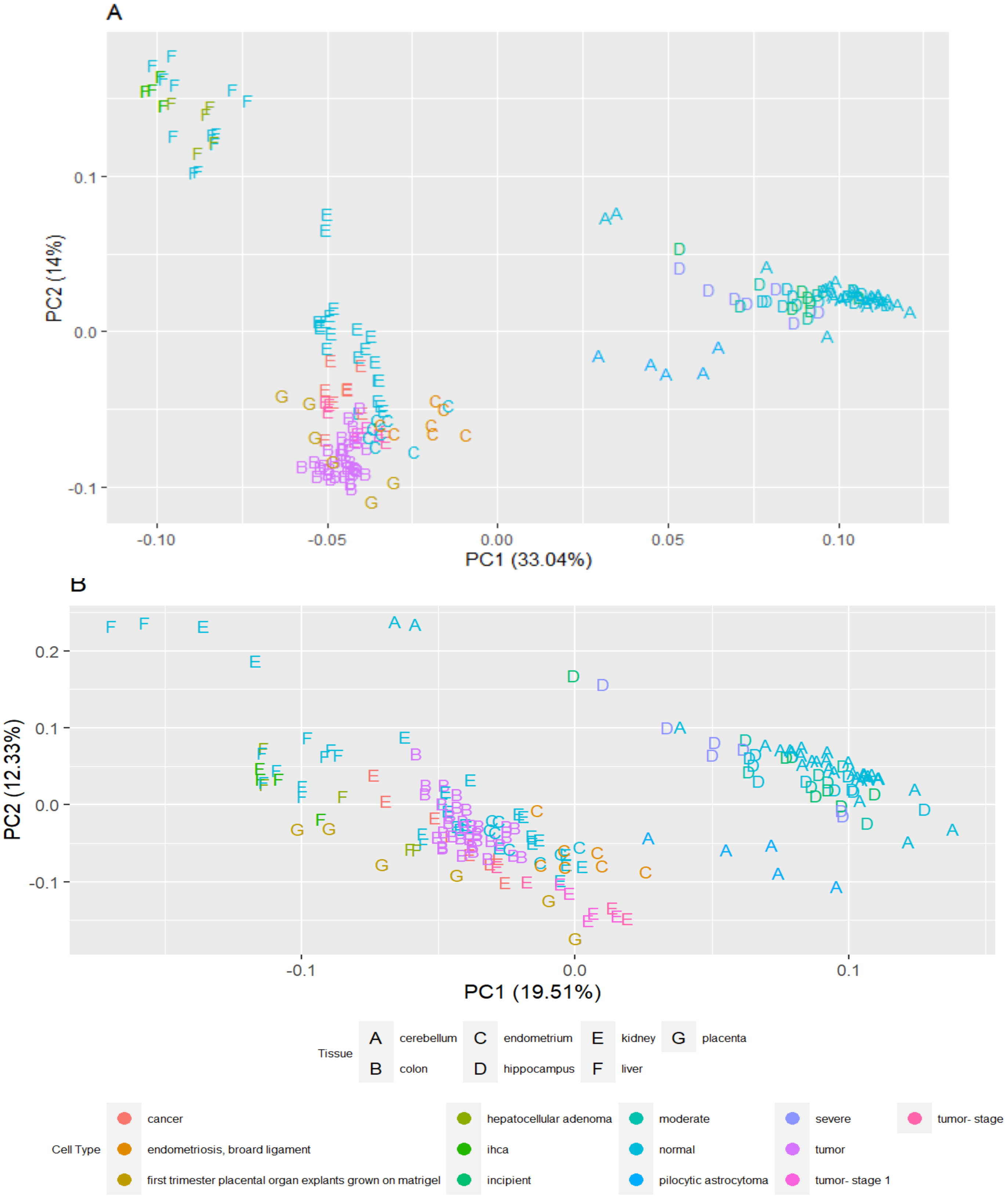


Fig 2. PCA analysis of gene expression data grouped by cell type, and tissue type. (A) Unscaled data. (B) Scaled data.

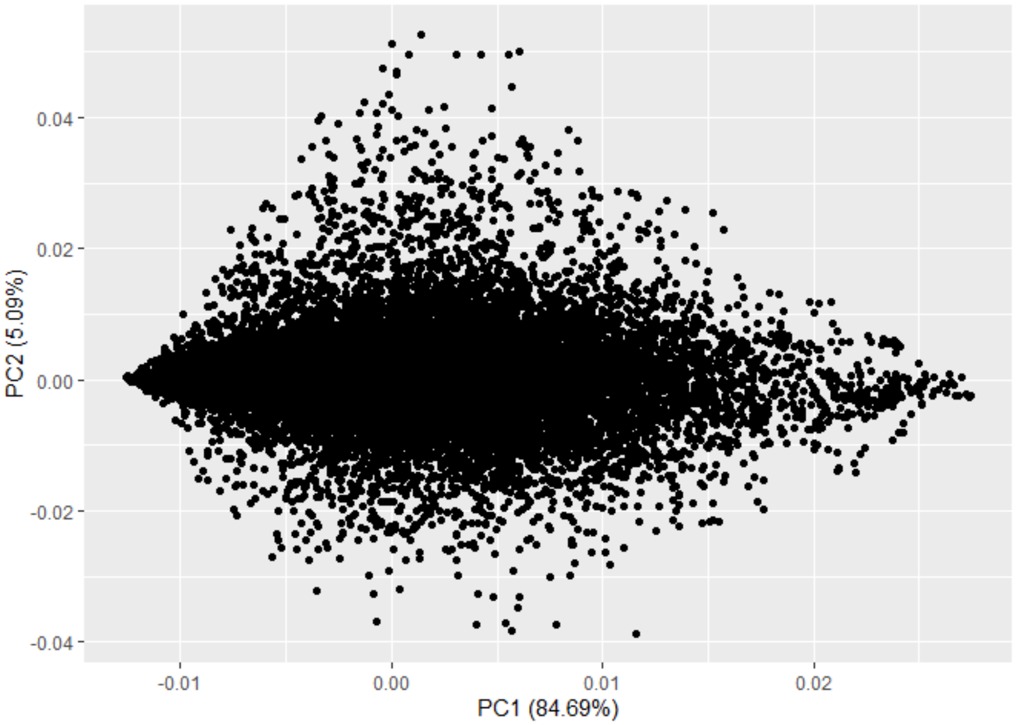


Fig 3. PCA analysis of gene expression data when samples were considered variables.

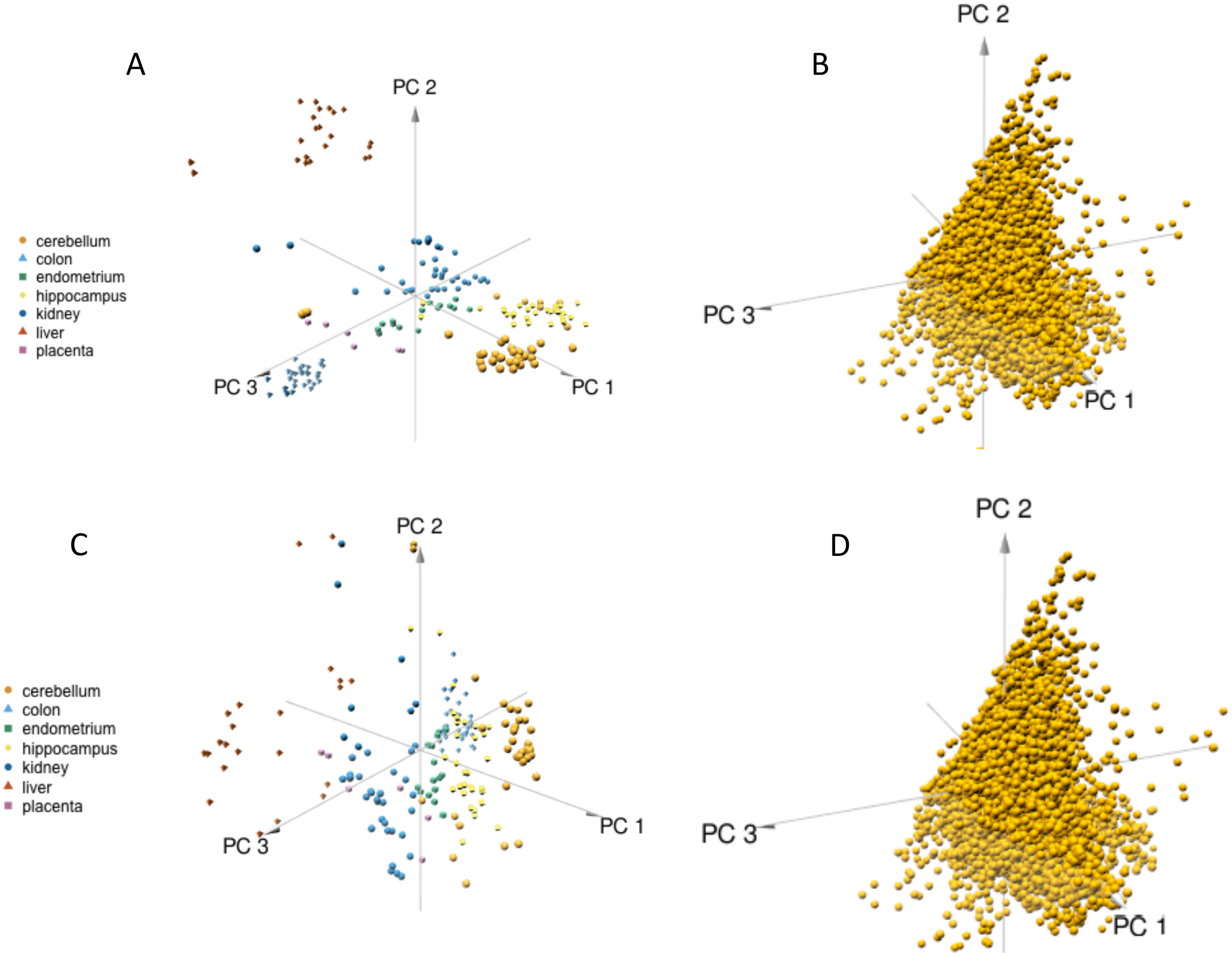


Figure 4: shows 3d PCA plots. A) 3d plot for unscaled data. B) 3d plot for unscaled samples as variables. C) 3d plot for scaled data. C) 3d plot for scaled samples as variable data. Considering the formation of clusters, we have selected unscaled as main model for problem 1

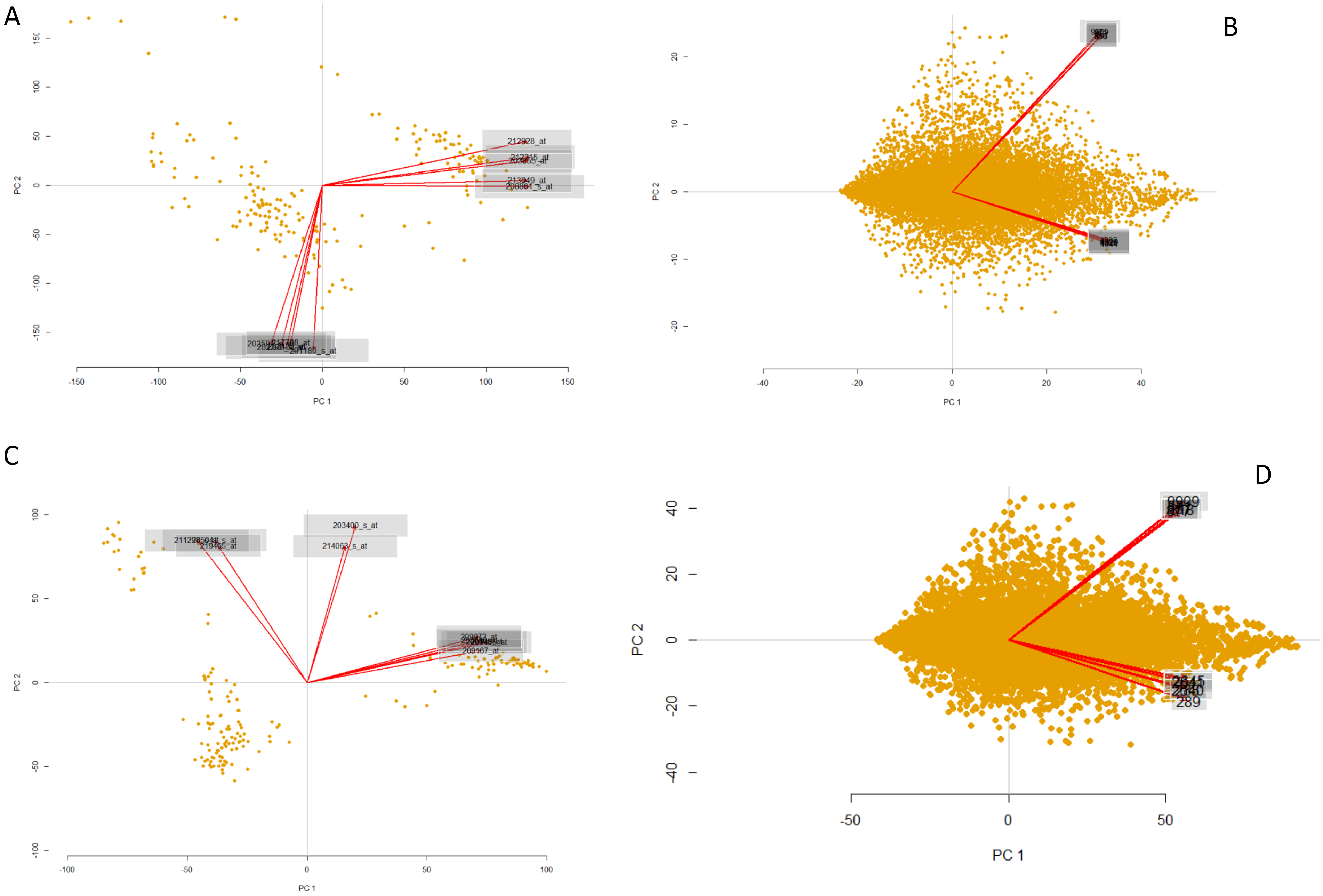


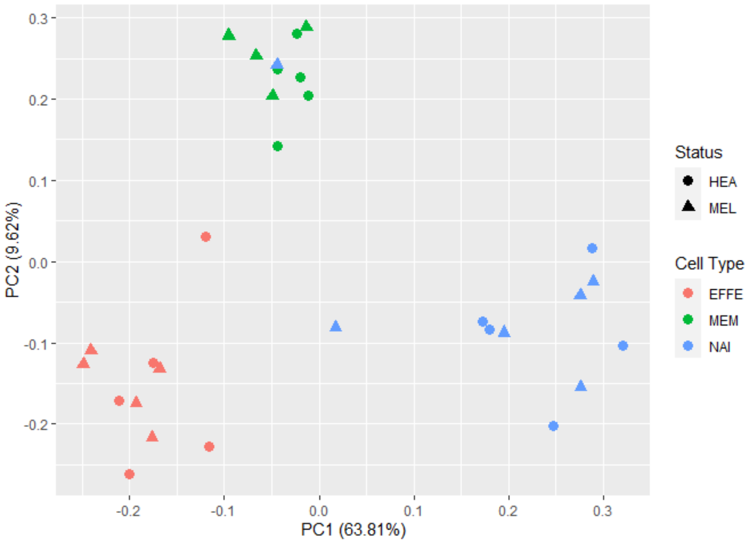
Figure 5: Shows biplots for both unscaled and scaled data. A) shows biplot for scaled tissue samples. B) shows biplot for scaled samples as variables. C) biplot for unscaled tissue samples D) biplot for unscaled samples as variables

Problem 2: T-Cell Gene Expression Analysis

T-cells are a type of lymphocyte cell that plays a vital role in the adaptive immune response. There are responsible for alerting the immune system and conducting the appropriate response if and when the pathogen previously the organism was exposed to appears again in the body. There are different types of T-cells which are memory T-cells, naïve T cells and effector T cells. The naïve T-cells are a kind of preliminary cell which when activated further differentiate into memory and effector T-cells. The purpose of Holmes et al 2005 study shed light on the action of this differentiation within T-cells and our statistical analysis will further explore the characteristics.

In figure 1 below a plot of principal component analysis has been shown. In the plot the types of points are indicated by condition of the cell, meaning status and also the type of T-cell they are. As shown, there is formation of three clusters and in all 3 clusters, the formation is clear because the different T-cell types have been grouped properly. Of all the 30 points only 3 seemed to sway from their designated clusters indicating that groupings are very accurate. Now in this data set gene expression of both melanoma cells and healthy T cells have been involved but if we just consider the status itself there seem to be no clear groupings involved indicating that the gene expression of melanoma cells is not much different from their designated healthy cells.

Considering there is no significant difference between the gene expression of healthy and cancerous cells this conclusion is not much different from the Holmes et al study. The lack of difference is explained by the study which imposes on the fact that Memory T-cells are not uniform within regard to gene expression and thus only differentiate into subtypes based on the expression of a receptor CCR7 present on the surface. Furthermore, when the cells differentiate, they fall into an intermediate state which themselves do not provide much of a difference between the expression of normal and melanoma cells. Finally, in the principal component plot because the PCA explains 63.8% variance and the effector cells are much closer to the cluster of memory cells and vice versa it supports the conclusion of Holmes et al 2015 study that memory cells differentiate into effector cells and at a rapid rate than it does with naïve cells.



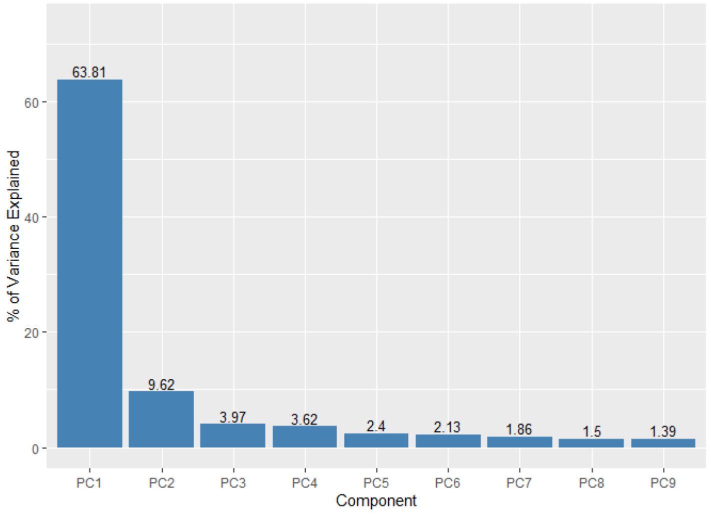


Fig 1. Shows the principal component plot where data points are identified by cell type and status of cell. The cells are T cells which are part of the immune system. ***Fig 2.*** Scree plot showing variance explained by each PC.

Problem 3: T-Cell Gene Expression Clustering Analysis

In order to confirm the results of Holmes et al 2015 study further analysis has been done. The figure below shows a visualization matrix and also a biplot where status of 3 different T-cells have been shown and the loadings have been indicated by genes. The loading indicates that there a few genes that seem to have the most influence in the T-cell differentiation. The visualization matrix shows that there are similarities between the effector cells and memory cells regardless of if it is from a melanoma organism or healthy one. The naïve cells also have a similarity to memory cells and significantly much more than effector cells. The naïve cells on the other hand have a high dissimilarity to effector cells. The genes that are revealed to have a high influence on variance are: X13492, X11495, X14831, X2698, X14844, X16348, X1135, and X17992.

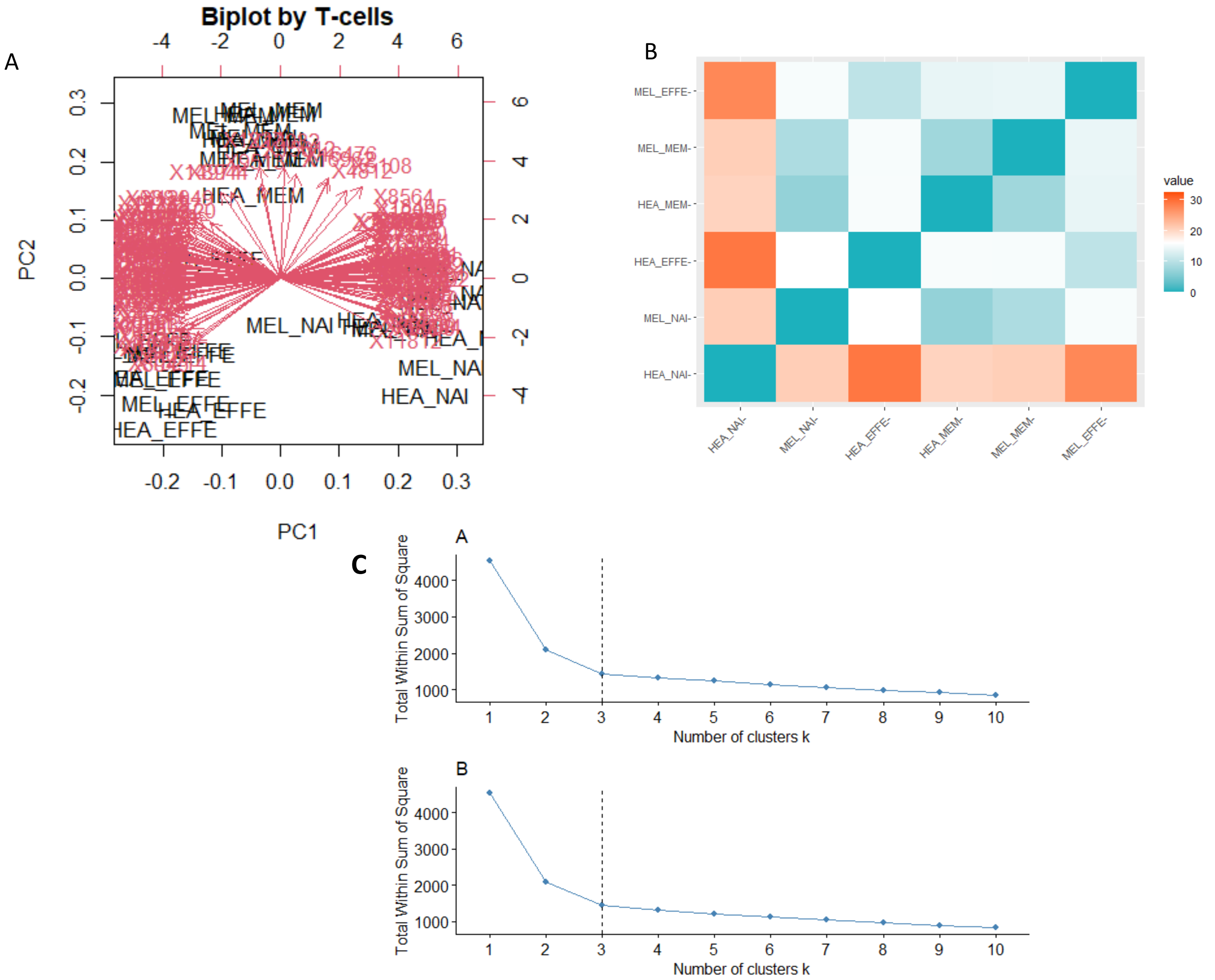


Fig 1a. Biplot where the T-cells and their status have been plotted. The red arrows indicate which genes perhaps play the most part. ***Fig 1b.*** Similarity Matrix of T-cells Status. Showing how similar each type of cell is to the other. ***Fig 1c Shows*** cluster validation function using (A) pam, and (B) kmeans. The validation is based on "Total Within Sum of Square" meaning how much variation is within the cluster. Ideally, we do not want to have too much variation within a single cluster. However, if you choose cluster number such as 10 even though sum of square is little the clusters itself won't differ to much externally thus it will not present a good formation. As 3 is ideal we have used 3 as appropriate number for non-hierarchical clusters.

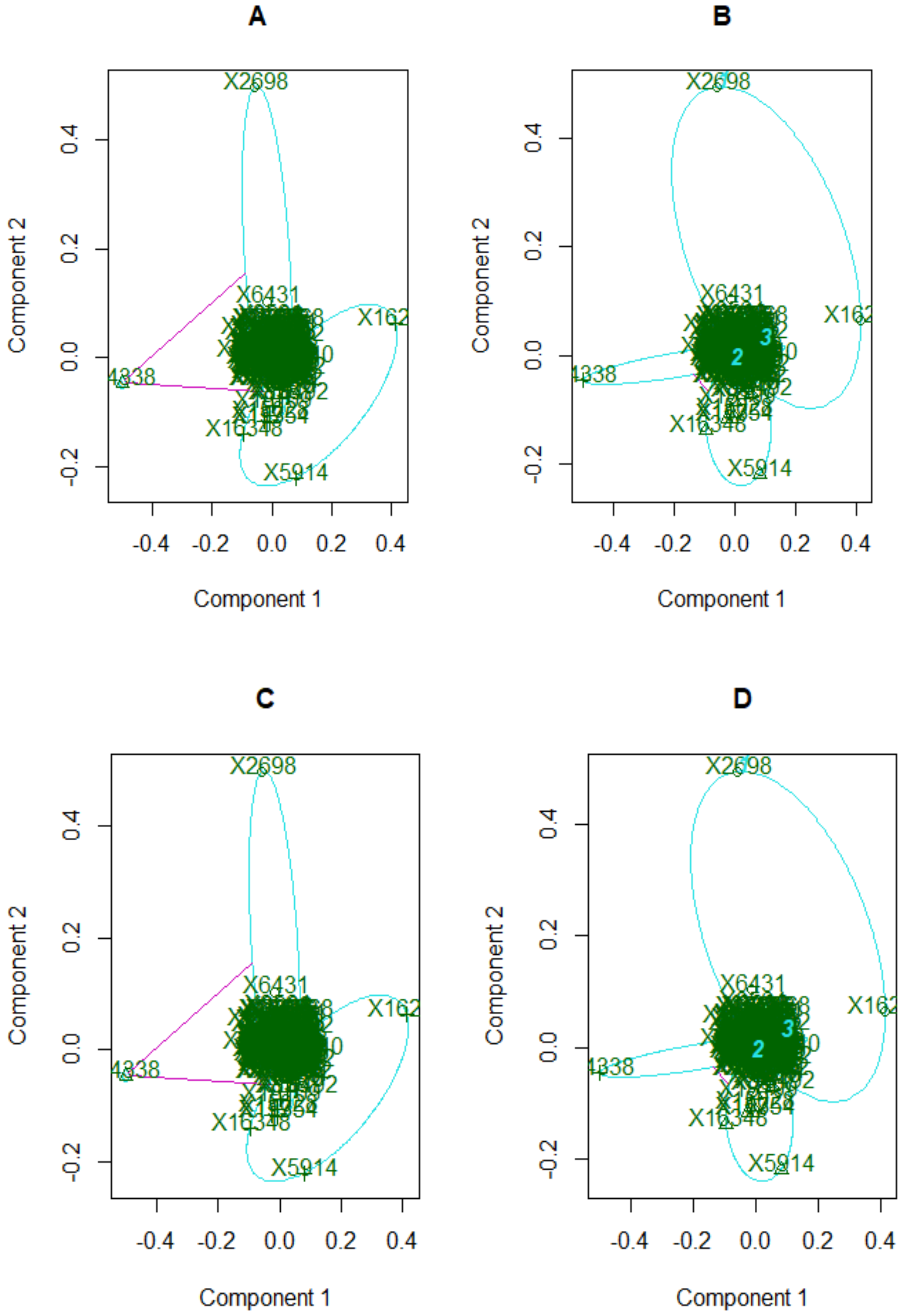


Fig 4. The formation of clusters by pam and k-means methods. There is not much of a difference in the formation clusters. A variation does occur when genes are used especially in the size of the cluster, but the conclusion is not different. (A) 3-means Cluster based on Genes (B) 3-medoids Cluster based on Genes (C) 3-means Cluster based of T-cell Status (D) 3-medoids Cluster based of T-cell Status.

The above figure shows the cluster formation plots based on the principal component analysis. For the formation of these plots two methods (kmeans, pam) were used. Pam is a more robust version of kmeans method. If you observe the cluster there is not much change in the formation of clusters especially in the case of formations based on T-cell status. There is a shift in size when genes were used to form the clusters but itself formation is not different. The cluster in 3b show that memory cells are much closer to effector cells and as previously mentioned much more similar than they are to naïve cells. The gene clusters unfortunately do not reveal any information as it is too cluttered. There are a few genes that seem to be at the very far ends of each cluster indicating they have a strong influence on the differentiation of T-cells.

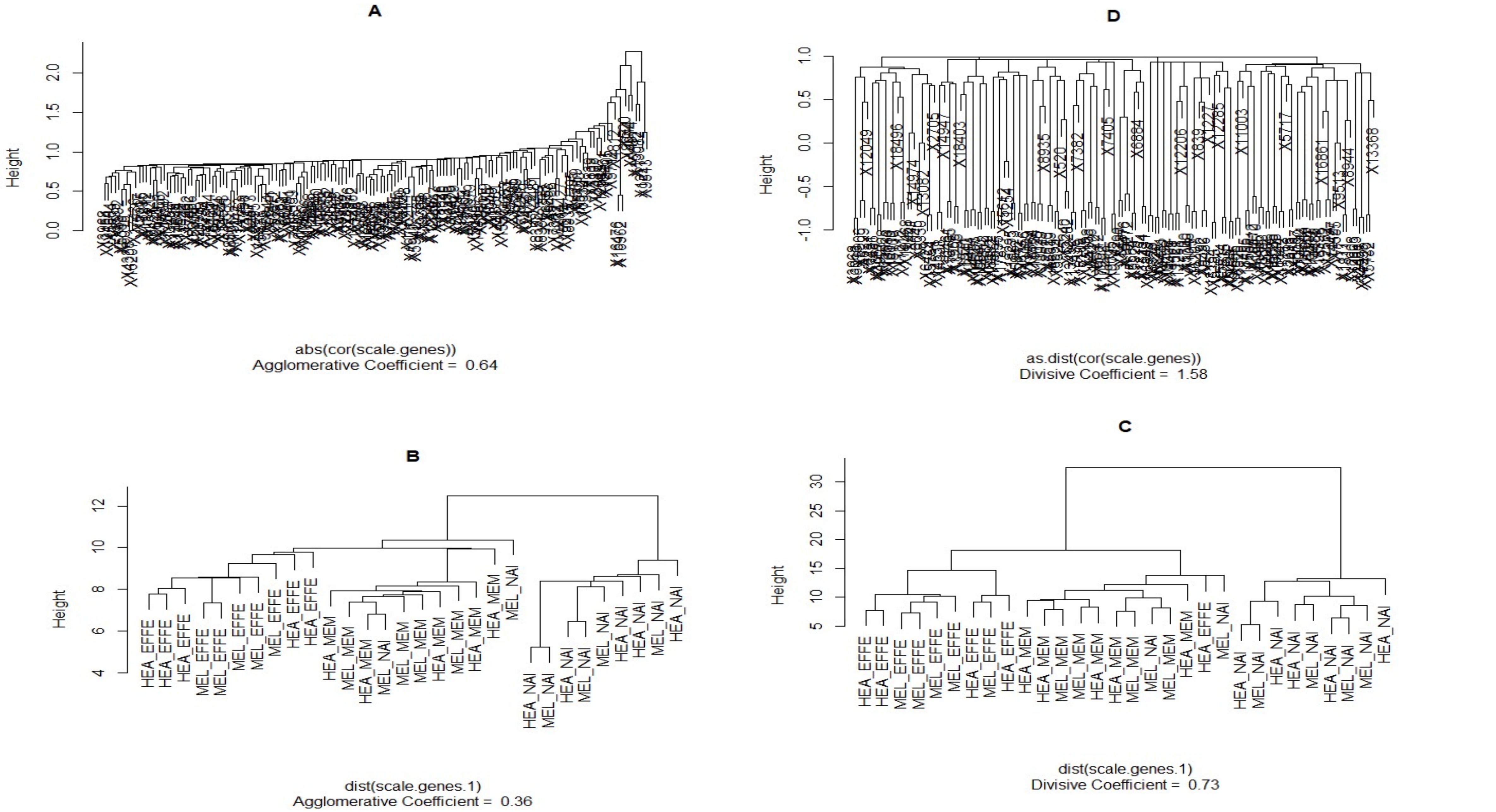


Fig 5. (A) Phylogenetic Tree Based on Expressed Genes Using Agnes. (B) B. Phylogenetic Tree based on Cell Status using Agnes. (C) Phylogenetic Tree based on Cell Status using Diana. (D) Phylogenetic Tree Based on Expressed Genes Using Diana.

In hierarchical clustering two methods were chosen to form the clusters which were eventually presented as phylogenetic trees. The two methods were Agnes and Diana, and their cluster formation reliability is indicated by agglomerative coefficient and divisive coefficients. Agnes is agglomerative method and Diana is divisive method. The formation of clusters when it comes to gene names has considerable number of branches and does not reveal much in regard to T-cell differentiation. The formation of clusters and the length of some branches is significantly much different by what is produced by Agnes in comparison to Diana.

In regard to the phylogenetic trees 4c and 4d the formation of branches is considerably similar by what is produced by both methods. There is a presence of an effector cell in the cluster of memory cell (healthy) which has been placed close to melanoma naïve cells an observation in the Diana tree. This is something not observed in the Agnes agglomerative method. Another difference is the branch length of naïve cells vary within its cluster. In conclusion thought both trees reveal that effector cells are much closer in relation to memory cells. The tree produced by Holmes et al 2015 is much closer to what is produced by Diana Method as in their tree an effector cell is also close to a naïve cell which is something observed in our tree. Finally, the naïve cells are closer to memory cells than they are to effector cells which has been a consistent observation from all other clustering methods.

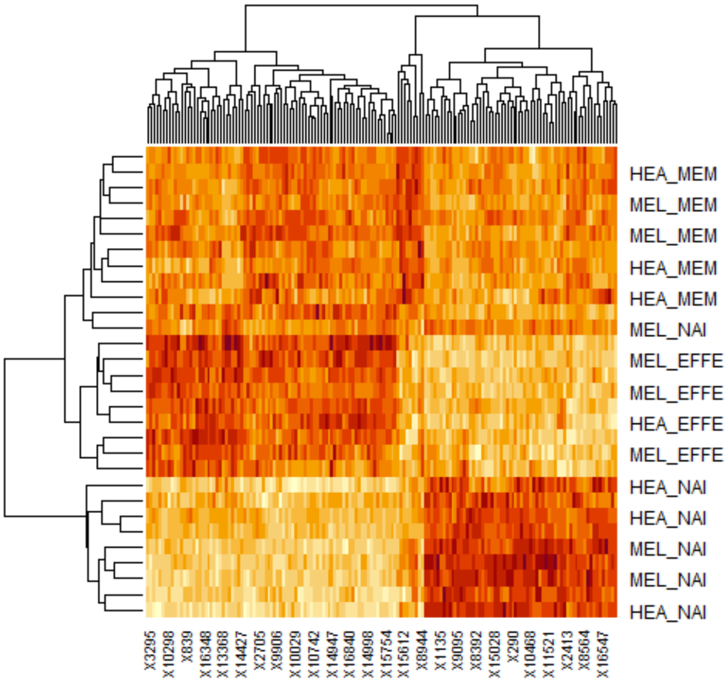
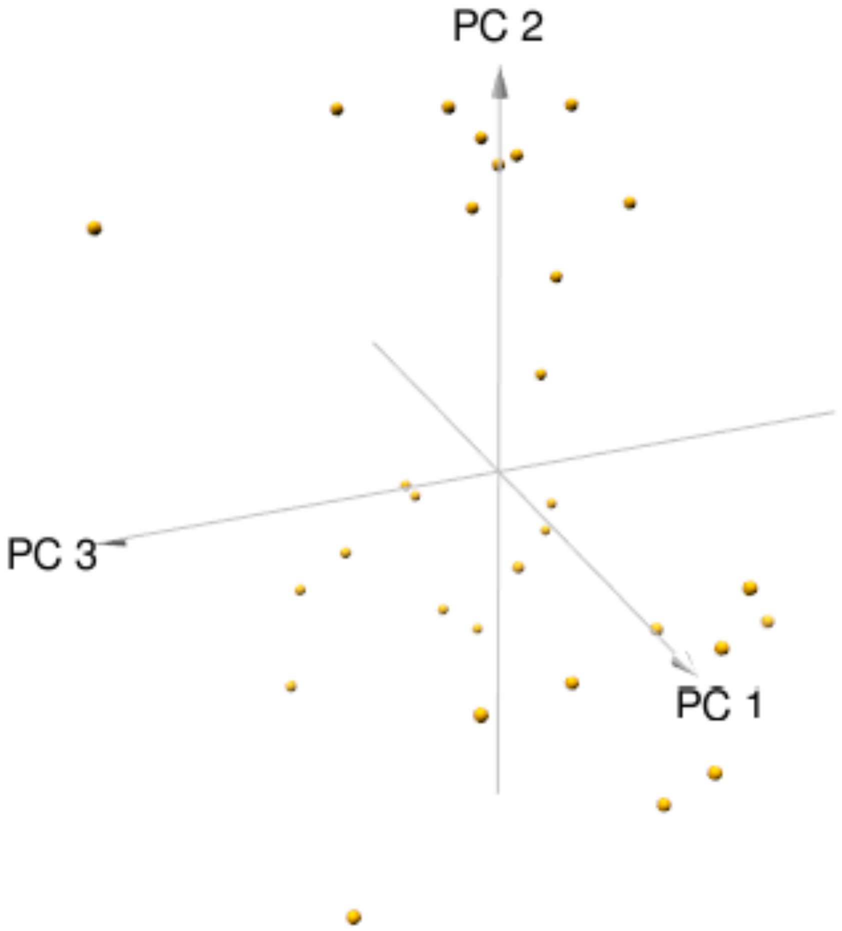


Fig 6. A heat map where the bottom part shows the genes expressed and on the right are T-cells with their status

The heat map reveals that the genes expressed in effector cells are not the same that are being expressed in naïve cells. However, similar genes are being expressed in memory cells indicating that genetically their structure seems to be very similar. What is also interesting is that the genes being expressed in naïve cells presented in a deep brownish red color are being expressed with high intensity and even though these genes are being expressed in memory cells as well, the degree of expression is lower. In conclusion, our results agree with Holmes et al study 2015 that the naïve cells because they are closer in relation to memory cells the naïve cells will first take on intermediate memory cell state which then become true memory cells. The memory cells are very close to effector cells in-terms of gene expression and thus will differentiate into them much faster than naive cells.

Supplementary Image for Problem 2



Supplementary figure 1: 3d PCA plot for T-cells