Identification of Prognostic Pathogenic Genes and Clinical Trends in Individuals with Glioblastoma

Author: Kevin Liu

Introduction

Glioblastoma (GBM) is a common cancer of the central nervous system and affected individuals has shown poor prognosis despite the currently available treatment options. [1] Furthermore, immune-related genes have been shown to play a prominent role in tumor gene expression. [1] Previously, Liang et al. identified 24 immune genes related to the prognosis of GBM using genes from the ImmPort database. [1-2] Through the screening process of prognostic GBM genes, Liang et al. categorized the 24 immune genes based on each gene's hazard ratio and the CCL1, LPA, and SH2D1B genes were considered prognostic pathogenic genes, having high hazard ratios relative to the remaining 21 prognostic protection genes. [1] Additionally, it was found that CCL1 is expressed at higher levels in males relative to females while BMP1 and OSMR are expressed at higher levels in those with ages greater than 50 years old. [1]

In this analysis, we hypothesize that the differential gene expression of prognostic pathogenic genes between normal tissue samples and tumor tissue samples will be the main driving factor influencing our principal components (PCs). Therefore, this analysis aims to identify the prognostic pathogenic genes using PCA, examine the gene expression patterns and interrelationships of both samples from normal tissue and GBM tissue via hierarchical clustering, and replicate the clinical findings of differential gene expression based on age group and sex by Liang et al.

Methods, Results, and Discussion

Prior to conducting our main analysis, we conducted an exploratory analysis on the sample donor demographics to identify any potential sources of bias within our data. This was accomplished by constructing bar plots to illustrate the distribution of tumor tissue sample donors in terms of race, sex, age group, and the number of sample donors for all tissue sample types per tissue source site. Additionally, we assessed the distribution of age at index, age at diagnosis, and survival time in years since indexing for tumor tissue sample donors by sample subtype via box plots. Based on such analyses, we find that our dataset is biased towards having a majority of sample donors who are white, male, and age greater than 50 years. We Additionally observe that Henry Ford Hospital contributed the largest number of samples and that donors with the PN sample subtype are typically younger in terms of age at index and age of diagnosis; these individuals also appear to have a longer survival time in years, which is expected given their young age and earlier diagnosis (figures not shown, see knitted HTML/PDF files).

Since the authors' analyses were based on the TCGA-GBM dataset, which is readily available to the public, we attempted to explore the utility of PCA weights in the screening of prognostic pathogenic genes among the 24 immune genes in GBM instead of the usage of a hazard ratio. The usage of PCA is justified here as we anticipate that genes with the highest relevance to the etiology of GBM to demonstrate the highest relevance to the PCs that explain the most variance in our RNA-seq read count data between each of the individuals.

As seen in Figure 1, it is evident that CCL1, LPA, and SH2D1B are consistently giving most extreme weights for PC1 and PC2. Specifically, the weights of CCL1, LPA, and SH2D1B are the only three negative weights in PC1 and they have the most positive weights in PC2. This is consistent with the identified prognostic pathogenic genes by Liang et al. and supports our hypothesis that PCA weights can identify prognostic pathogenic genes that are expected to contribute the most to the PCs yielding the highest explained variances.

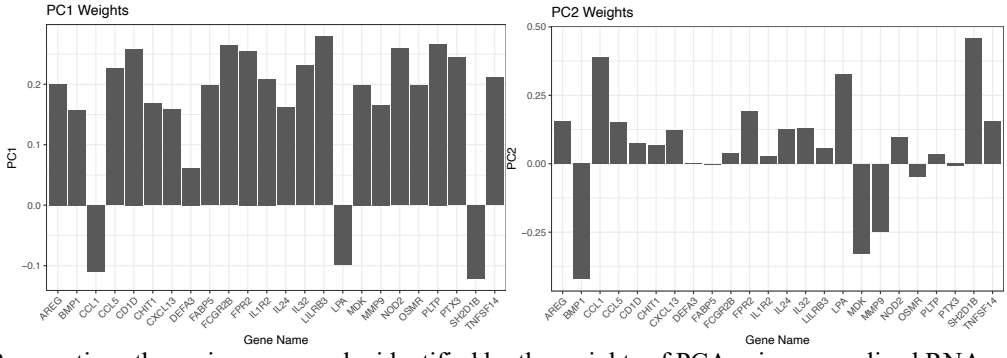

Figure 1. Prognostic pathogenic genes can be identified by the weights of PCA using normalized RNA-seq read counts of the 24 immune genes.

Subsequently, we utilized hierarchical clustering to explore the 24 immune gene expression patterns in terms of how each of the genes are related to one another and whether sample subtypes can be clustered based on the normalized RNA-seq read counts of the 24 immune genes for each sample donor. Here, hierarchical clustering is favored over k-means clustering as hierarchical clustering outputs a dendrogram that is more informative of the interrelationships between genes and samples rather than an unstructured set of flat clusters returned by k-means. More specifically, Euclidean distance was applied to the clustering as it is the most common distance metric and complete linkage was used to encourage the most compact clusters from hierarchical clustering.

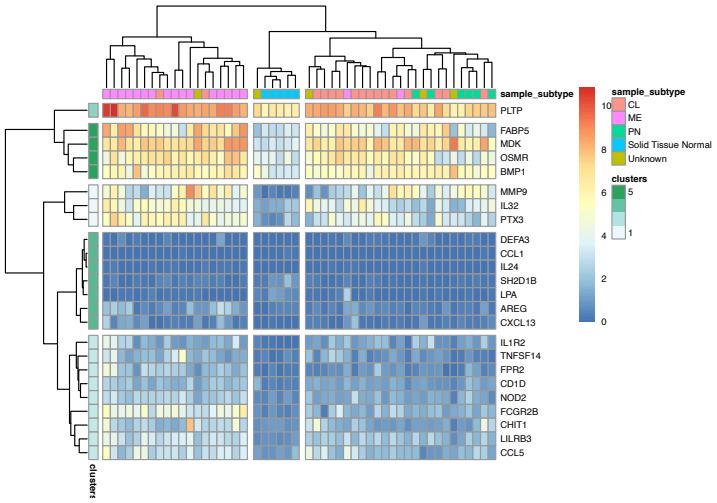

Figure 2. Hierarchical clustering of both samples and genes shows differential expression patterns between samples from normal tissue and GBM tissue. Each row represents an immune gene and each column represents a sample.

Based on the hierarchical clustering shown in Figure 2, we see that samples of normal tissue have much lower expression levels of almost all genes except DEFA3, CCL1, IL24, AREG, and CXCL13 while these samples express higher levels of SH2D1B and LPA relative to GBM tissue samples. We also see that when clustering on samples (i.e., columns), the ME subtype and Solid Tissue Normal subtype are nicely clustered together in their respective clusters. In this case, the clustering of samples were cut into 3 clusters as doing so results in clusters that correspond to the (mostly) ME subtype, (mostly) Solid Tissue Normal subtype, and all remaining subtypes that are mixed together; clustering of genes were cut into 5 clusters as these clusters convey the most relevance between genes without leaving too many single genes in their own clusters; it is evident that doing so captures the best gene expression patters that distinguish the normal tissue samples from the tumor tissue samples (for more detailed analysis methodology and explanations, please refer to the knitted HTML/PDF files).

Finally, we attempted to replicate the observed trends identified by Liang et al. and verified that males express higher levels of CCL1 than females and both BMP1 and OSMR are expressed at higher levels in those with ages greater than 50 years old (figures not shown, see knitted HTML/PDF files).

Taken together, we conclude that interpretation of PCA weights is a valid method for the identification of the prognostic pathogenic genes and that hierarchical clustering using Euclidean distances and complete linkage sufficiently captures the interrelationships between both gene expression patterns and sample subtypes. We also successfully validated the clinical findings of differential gene expression based on age group and sex by Liang et al. Nonetheless, it is encouraged to review our knitted HTML/PDF documents of the present analysis, which provides a detailed walkthrough of the entire analysis with all results and figures shown.

References

1. Liang P, Chai Y, Zhao H, Wang G. Predictive Analyses of Prognostic-Related Immune Genes and Immune Infiltrates for Glioblastoma. Diagnostics (Basel). 2020;10(3):177. Published 2020 Mar 24. doi:10.3390/diagnostics10030177
2. Bhattacharya S, Dunn P, Thomas CG, et al. ImmPort, toward repurposing of open access immunological assay data for translational and clinical research. Sci Data. 2018;5:180015. Published 2018 Feb 27. doi:10.1038/sdata.2018.15