Genetic Differences Between Glioblastoma Multiforme and Low-Grade Glioma Populations: How Does Copy Number Variation Distinguish Each Disease?

GitHub Link:
https://github.com/Ninada-U/Analyzing-CNVs-between-GBM-and-LGG/

Website Link:
https://siqihuang47.github.io/dsc180b_visualization/

Noah Inada
ninada@ucsd.edu

Yifei Ning
y3ning@ucsd.edu

Siqi Huang
sih041@ucsd.edu

1. Introduction

1.1 Abstract

The goal of this project is to identify genetic differences between glioblastoma multiforme (GBM) and low-grade glioma (LGG) populations. This is important because these genetic differences may be useful for diagnosing these diseases in people. This was done by comparing each population's copy number variation focal scores. Our project's question is: How does copy number variation of GBM and LGG distinguish each disease? Our hypothesis is that the GBM population will have more copy number variation because GBM is more malignant, and so has more genetic mutations.

In comparing the focal scores it was found that the GBM population had 1.5 times more duplications and deletions than the LGG population, confirming our hypothesis. And, there were specific sets of outlying genes in the GBM population that distinguish it from the LGG population - but only on a cross-case analysis level. Practically, it is not applicable to use them as markers for diagnosis. This project confirms the hypothesis that GBM has more genetic variation in its copy number variation, but does not support the approach to use copy number variation as a method of diagnosing either disease.

1.2 Background

Gliomas are a type of tumor in the brain or the spinal cord occurring from glial cells. The glial cells most prone to developing tumors are astrocytes, a specific type of brain cell whose main function is to support neurons and form scar tissue in damaged areas [1]. Gliomas that occur from astrocytes are called astrocytomas. GBM exhibits more genetic abnormalities than other astrocytomas and is the most aggressive and fatal of all glial tumors. [2] The causes of GBM are largely unknown, but research suggests that approximately 5 percent of all glioblastomas are caused by hereditary conditions, with some of the cases being from people with Neurofibromatosis type 1 (NF1), Turcot syndrome and Li-Fraumeni syndrome all being genetic syndromes associated with increased susceptibility to cancer. [3] Because there are multiple syndromes associated with glioblastomas, and hereditary causes comprise only a small percentage of the entire affected population, it is difficult to predict whether glioblastomas will occur in an individual or not, and how dangerous they are in those harboring mutations. [4] Another kind of glioma is a low-grade glioma, being low-grade because it is much less severe than a high-grade glioma like GBM. [5]

Copy number variants (CNVs) are abnormal numbers of copies of a gene. Such mutations can occur from deletions, insertions, duplications or any combination among these three. Cancer cells are typically related to a set of multiple mutated genes, so CNVs are a good indicator for some diseases like cancers.

Focal scores summarize copy number variation. In our data, a score of one indicates a gene duplication, a score of zero indicates that there was no change in copy number, and a score of negative one indicates a gene deletion. [6]

2. Data

2.1 Data Overview

Our data is from The Cancer Genome Atlas (TCGA) which is an online repository of cancer data. It was started by a diverse group of researchers to consolidate and analyze cancer data. Their data portal is called the Genomic Data Commons Data Portal (GDC). The GDC has over 84,000 cancer cases spanning across 67 primary cancer sites like the lung, breast, and brain. This project uses all of GDC's GBM and LGG cases, amounting to 617 and 507 cases respectively. Each case has both CNV data and clinical data. The clinical data includes gender, when they were diagnosed, and if/when they passed away. [7]

2.2 Data Ingestion

The R analytics package we use, TCGAbiolinks, is capable of making queries directly to the GDC. Depending on the query, it can use a list of case IDs. Because the R code gathers the data itself, much of the data ingestion pipeline we had previously made that scraped for URLs, downloads the data, unzips and zips, and renames and organizes the files, became redundant. That being said, when the R package does download the data, it can be stored for quicker later retrieval.

2.3 Concerns

As stated on the TCGA website, the data "will remain publicly available for anyone in the research community to use." Thus, our project is likely not infringing on legal access rights. However, it is possible to identify a person based on their genetic information, so there is a responsibility to respect the privacy of the people behind our data.

3. Methodology

3.1 Survival Plots

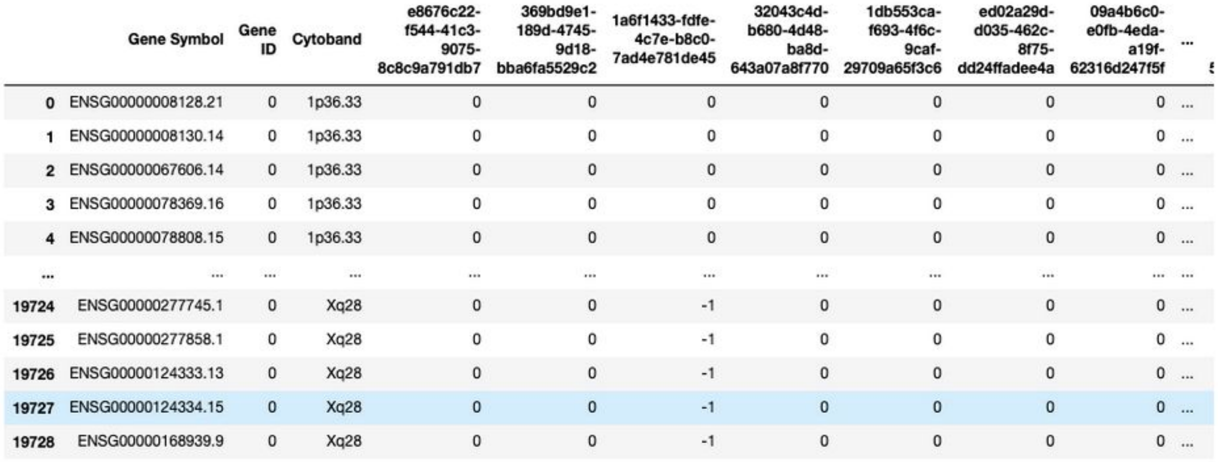
We created survival plots that confirmed our background research that GBM is more lethal than LGG. We created two sets of survival plots and present an estimate of survival probability depending on the days from cancer diagnostics. We also compared how gender and ethnicity affect the survival rate within each population.

By using the TCGAbiolinks R package, we used the 'days_to_death' and 'vital' data fields of the clinical data to create a survival plot. The x-axis is the days-since-diagnosis, and the y-axis is the survival probability. These plots can help us recognize if there is a large difference in mortality between the GBM and LGG populations.
The guidelines linked in the references describe the TCGAbiolinks R package [8]. There are instructions for installing the package, ingesting data, and visualizing the data. We specified our data query to fetch the relative data for our project, and used the 'TCGAanalyze_survival' function to visualize the survival plots of the two populations.

3.2 Focal Scores

A CNV focal score is a value between negative one and one that indicates if the gene was deleted or copied. A negative one means it was deleted, a zero means nothing changed, and a one means it was copied. Because these scores are calculated across many samples, decimal values are an indicator of 'what usually happens'. When our data file was generated, a noise-cutoff of .3 was used, meaning that values less than -.3 are classified as -1, values between -.3 and .3 are classified as 0, and values above .3 are classified as 1.

To retrieve our data, we used the TCGAbiolinks R package to fetch the gene level copy number scores. All of the data we needed comes in one text file. This file can also just be downloaded manually by navigating the GDC data portal.



	Gene Symbol	Gene ID	Cytoband	e8676c22-f544-41c3-9075-8c8c9a791db7	369bd9e1-189d-4745-9d18-bba6fa5529c2	1a6f1433-fdfe-4c7e-b8c0-7ad4e781de45	32043c4d-b680-4d48-ba8d-643a07a8f770	1db553ca-f693-4f6c-9caf-29709a65f3c6	ed02a29d-d035-462c-8f75-dd24ffadee4a	09a4b6c0-e0fb-4eda-a19f-62316d247f5f	...
0	ENSG00000008128.21	0	1p36.33	0	0	0	0	0	0	0	...
1	ENSG00000008130.14	0	1p36.33	0	0	0	0	0	0	0	...
2	ENSG00000067606.14	0	1p36.33	0	0	0	0	0	0	0	...
3	ENSG00000078369.16	0	1p36.33	0	0	0	0	0	0	0	...
4	ENSG00000078808.15	0	1p36.33	0	0	0	0	0	0	0	...
...	...	...	...	...	...	...	...	...	...	...	...
19724	ENSG00000277745.1	0	Xq28	0	0	-1	0	0	0	0	...
19725	ENSG00000277858.1	0	Xq28	0	0	-1	0	0	0	0	...
19726	ENSG00000124333.13	0	Xq28	0	0	-1	0	0	0	0	...
19727	ENSG00000124334.15	0	Xq28	0	0	-1	0	0	0	0	...
19728	ENSG00000168939.9	0	Xq28	0	0	-1	0	0	0	0	...

Fig 3a. Focal Score Data

In our data file, every row represents a protein-coding gene, and every column represents a case. So, each cell in the table indicates, in all of the cases, if that gene was deleted (-1), copied (1), or stayed the same (0).

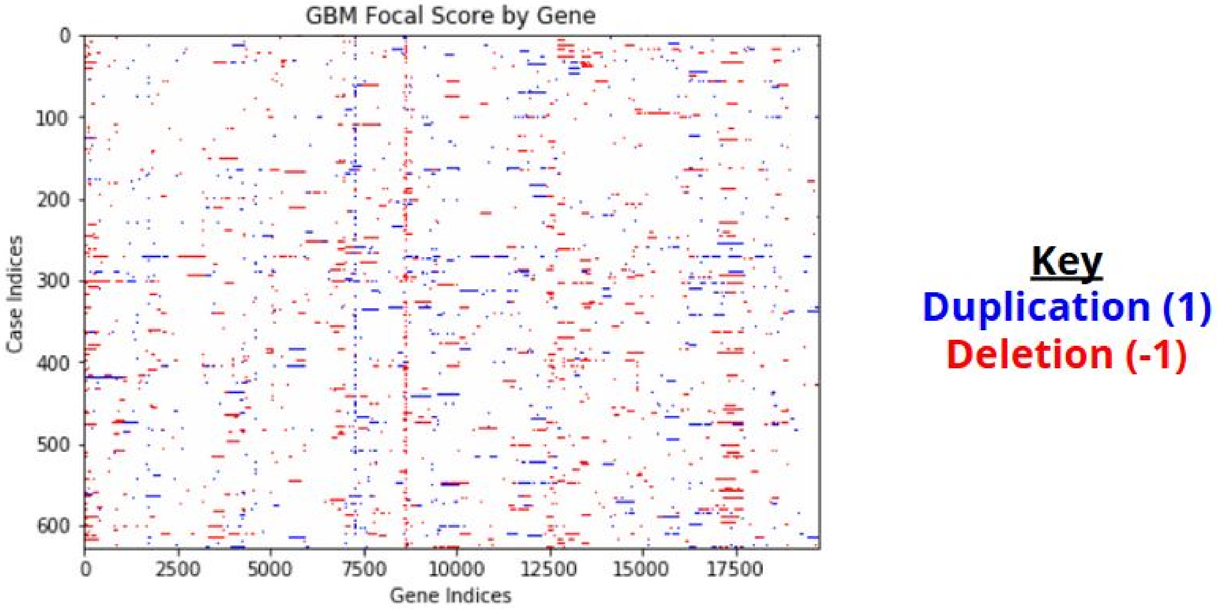


Fig 3b. Focal Score Plot

We plotted this data table, assigning a color to each data value. This classification makes it easier to interpret the plot. We generated plots of the focal scores for both of our populations. They show, for each population, how copies of each protein-coding gene varies in each person.

3.3 Focal Score Proportions

We summarized the focal score plots for our analysis by finding the proportion of cases of duplications and deletions for each gene and plotting both lists of proportions. By doing so, the number of cases in a gene where there is copy number variance can be easily compared with other genes. To interpret the plot: the higher a blue dot is, the more the gene was duplicated across the cases. The lower a red dot is, the more the gene was deleted. We also used cytoband data included in the focal scores file to delineate which genes belong to which chromosomes.

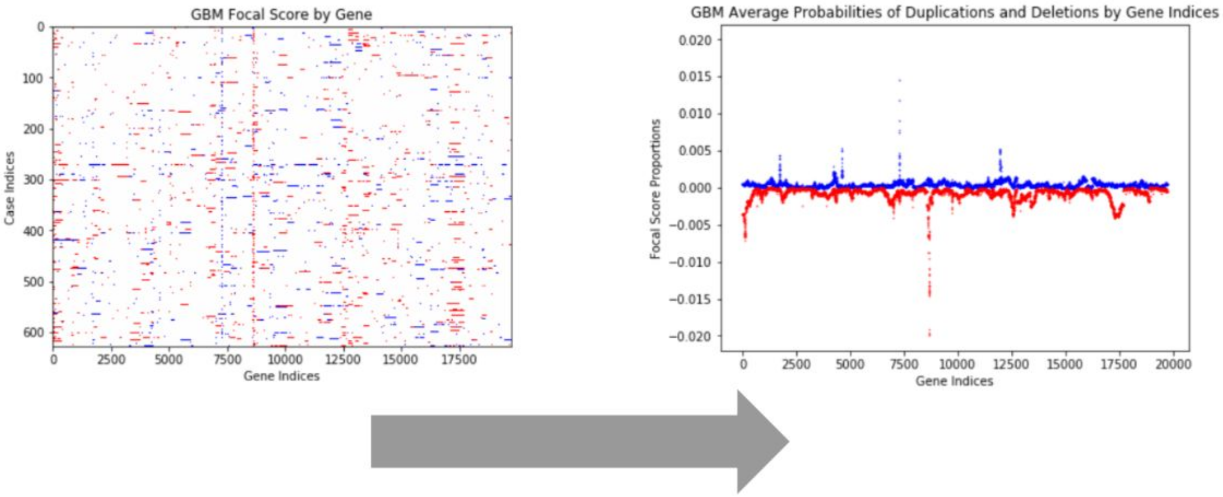


Fig 3c. Summarizing Focal Score Plots

4. Exploratory Data Analysis

4.1 Survival Rate Analysis Across Genders

To further establish that GBM is statistically more dangerous and thus has a lower survival rate, two plots are used to distinguish the differences between GBM and LGG across gender. Within each population, the survival rates are compared across males and females. The y-axis in the plots represents the survival probability. The x-axis represents the days passed since the diagnosis of the disease (either GBM or LGG). Each plot has a line graph showing the proportion of surviving patients as time goes on, and a chart annotating how many people compose the proportion. As shown below the left image is the survival analysis plot for LGG, and the right image is the analysis plot for GBM.

Comparing the graphs, numbers of survivors with GBM and LGG tend to decline at an increasing rate as time passes. Such a declining trend also holds for both males and females. For the population with LGG, their survival probability varies much more in the latter days and the declining trend is much more smooth with a larger variance. Visually, the males and females in the population develop a similar pattern: no significant differences across gender. On the other hand, the GBM survival plot shows a much more aggressive declining trend. As for different genders in the population with GBM, females tend to have a higher probability of survival than the males in the population, with respect to the whole range of days.

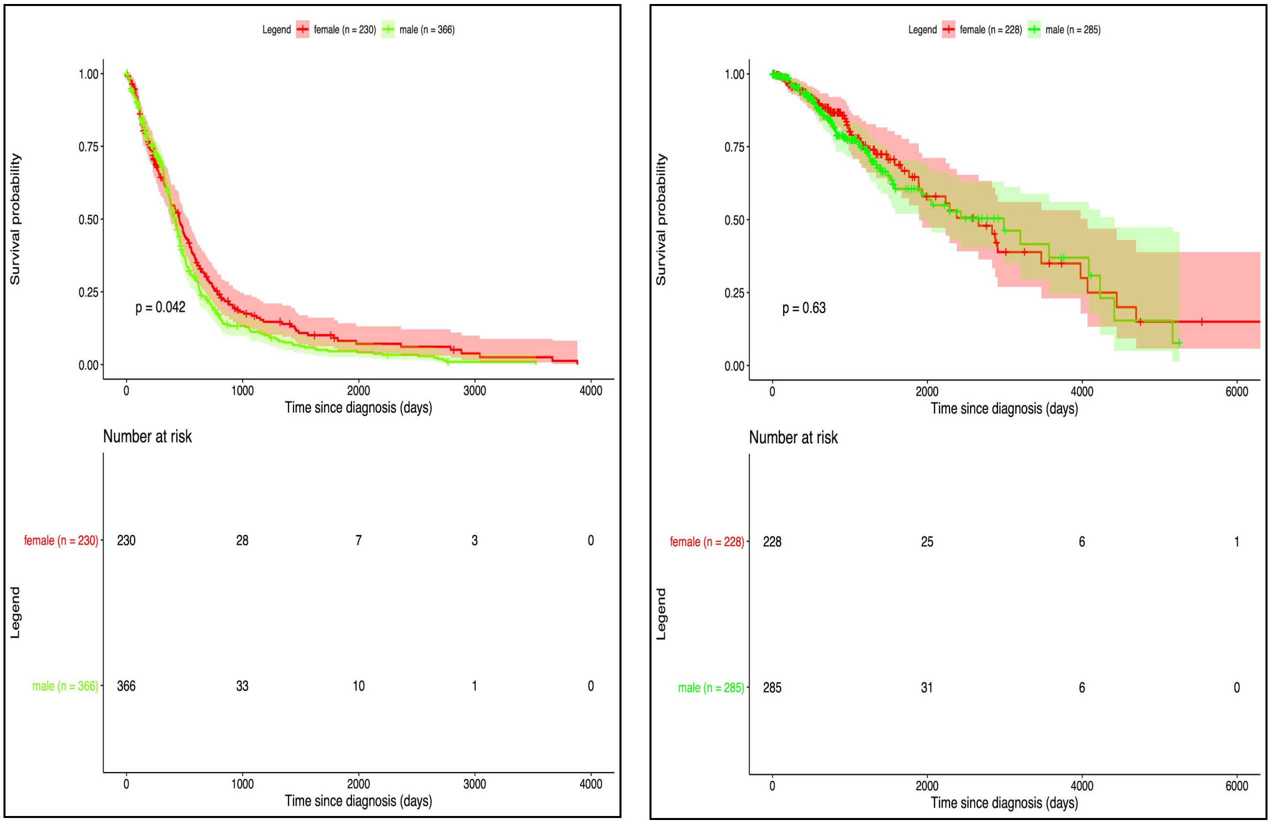


Fig 4a. Survival plots of GBM (left) and LGG (right) populations across genders

Conclusively, GBM appears to be more lethal - there is a much steeper decline in survival probability within just the first 1,000 days. After 1,00 days, both male and female people who have GBM have less than a .25 probability of survival, while male and female people in the LGG population have above a .75 probability at 1,000 days. Also, per the annotation chart, there are no survivors after 4,000 days (11 years), but there are still 12 LGG patients living.

4.2 Survival Rate Analysis Across Ethnicities

Aside from the fact that differences exist across different genders, a similar comparison across ethnicities will be drawn between the LGG and GBM populations. Three categories of ethnicities are compared: ‘hispanic or latino’, ‘not hipanic or latino', and ‘not reported’.

In the survival plot for GBM, the number of cases are so small for the group ‘hispanic or latino’ (13) and the group ‘not reported’ (93). Most cases belong to ‘non-hispanic or latino’. It is not appropriate to draw conclusions from limited numbers of cases, and thus we should only consider the general trend for all. Though visually, ‘hispanic or latino’ and ‘not hipanic or latino’ are somewhat different, their differences are not significant as the p-value equals 0.64.

In the survival plot for LGG, likewise, the number of cases are so limited for the group ‘hispanic or latino’ and the group ‘not reported’. It is not persuasive enough to draw conclusions about these two groups. Again, the general pattern of the ‘not hispanic or latino’ population follows the previous findings.

Conclusively, regardless of the ethnicity, the survival probability plot for either GBM or LGG conforms the general pattern in section 4.1: the survival rate decreases at an increasing rate as time goes by. The decreasing rate is much larger in GBM than in LGG, further attesting our hypothesis that GBM is more malignant compared with LGG regardless of ethnicities.

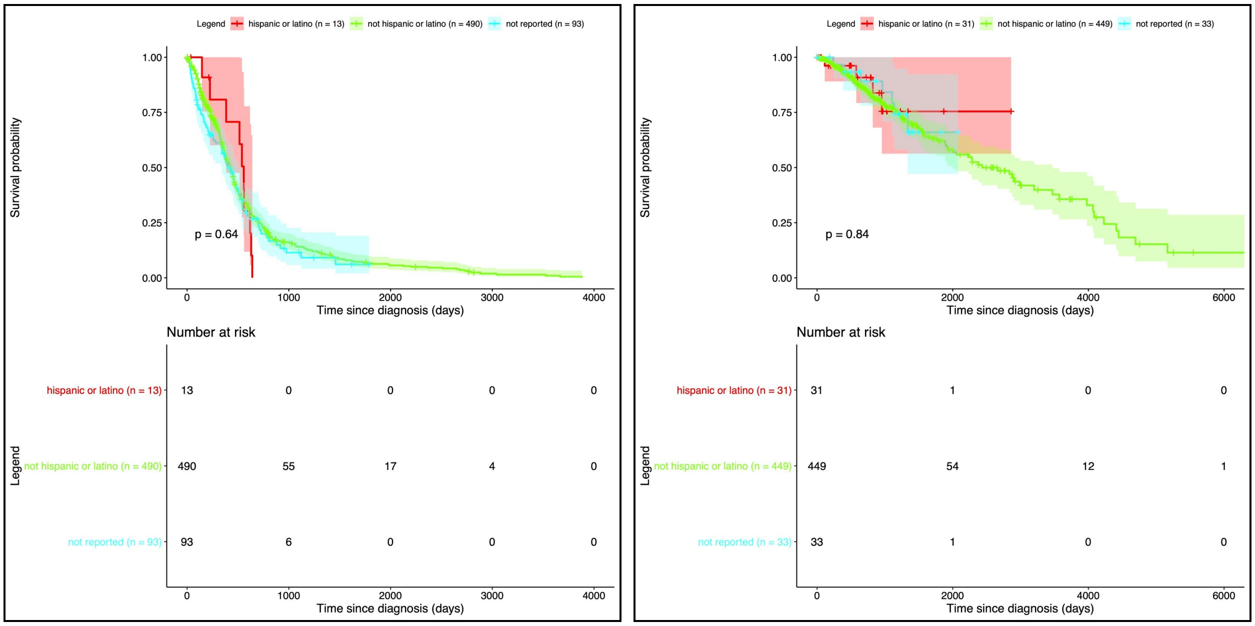


Fig 4a. Survival plots of GBM (left) and LGG (right) populations across ethnicities

4.3 Focal Score EDA

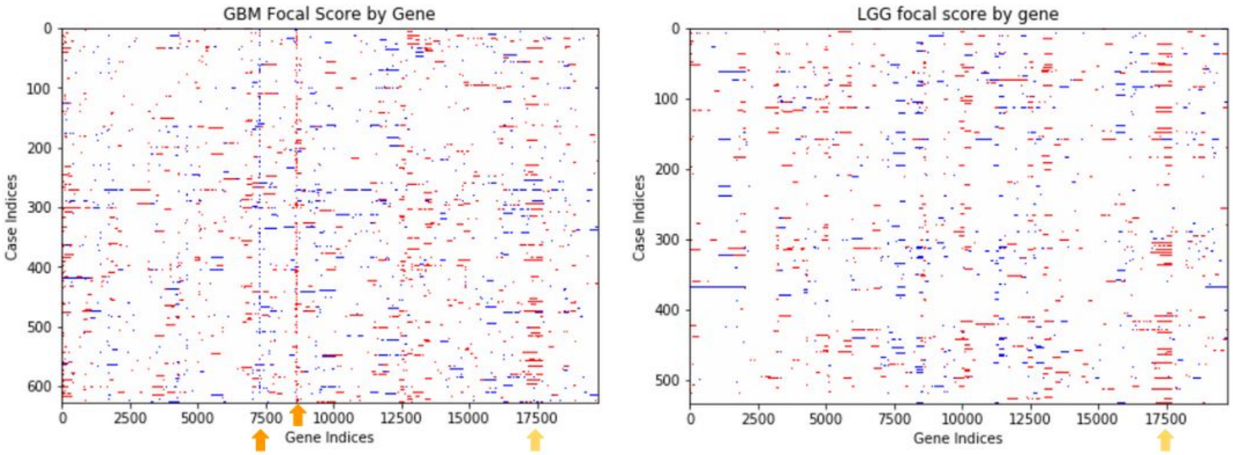
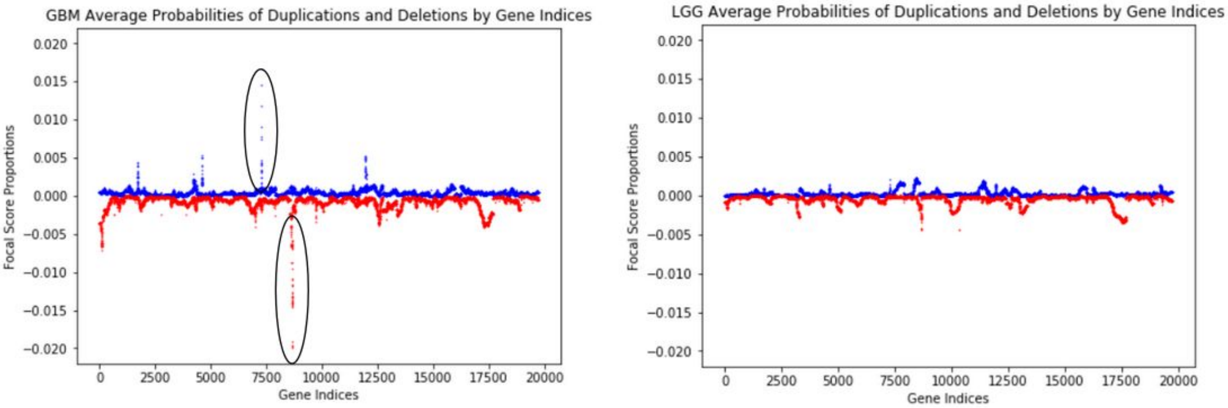


Fig 4b. Focal Score Plots

In our exploratory data analysis, the first things were noticed were the GBM vertical lines of duplications and deletions, indicated by orange arrows, and we also noticed that there are many deletions in both GBM and LGG to the right side of the graphs at the yellow arrows. We look more into these observations in our analysis.

5. Analysis


GBM Average Probabilities of Duplications and Deletions by Gene Indices

LGG Average Probabilities of Duplications and Deletions by Gene Indices

Fig 5a. Focal Score Proportions

Here are the focal score proportions of GBM on the left and LGG on the right. The GBM outliers aforementioned in the EDA are indicated with ovals. The most outlying deletion gene is **ENSG0000026454** and the most outlying deletion gene is **ENSG00000146648**. The first is involved in coding proteins and transferase activity, and the second is involved in transferring pentosyl groups. [9][10] We did not determine how these activities relate with GBM and LGG, but the outlying presence or absence of these genes seem to indicate that they can serve as statistical indicators to distinguish GBM from LGG.

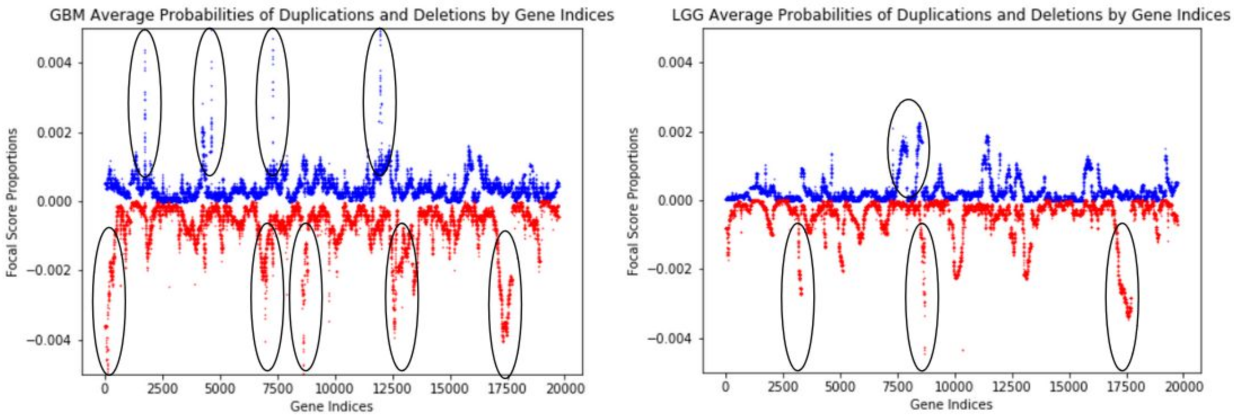


Fig 5b. Zoomed Focal Score Proportions

Figure 5b shows the data with a more zoomed perspective. The GBM population has more outliers than the LGG population. There is also a general rise and fall pattern in the proportions of both populations - duplications or deletions will 'stack', and then 'fall'. We suspect this is due to duplications and deletions across multiple genes.

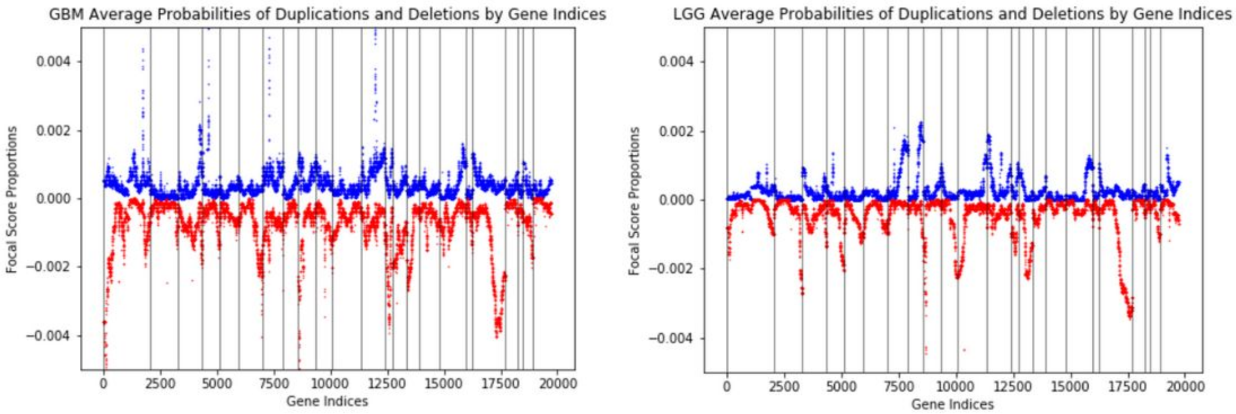


Fig 5c. Delineated Focal Score Proportions

Figure 5c includes lines from the cytoband data where a new chromosome begins to delineate the chromosomes. Where a new chromosome begins, there is often a fall in proportions. This seems to show that duplications and deletions will span across genes, but not into the successive chromosome.

The average number of gene duplications and deletions were calculated for each population. The GBM population has about 7.8 duplicates per gene and 14.9 deletions per gene, while the LGG population has about 5.2 duplications per gene and 9.3 deletions per gene. The GBM population has 1.5x more duplications and deletions than the LGG population.

6. Conclusion

The GBM population does have more copy number variation than the LGG population, confirming our hypothesis. GBM does have outliers to distinguish it from LGG, answering our project's goal. However, because the proportion differences were only .01, so based on this analysis, a duplication or deletion at those genes may not be a reliable indicator for diagnosis.

7. Evaluation

Our focal scores did not take into account how many duplications or how many deletions actually occurred - only if any duplication or deletion occurred. A further investigation could include approximately how many times the gene was copied or deleted. Second, having more cases to study could improve a future analysis. Finally, this project only focuses on how to distinguish LGG and GBM using just CNVs. Including more data types than Copy Number Variation could improve a future analysis where clustering methods could be used.

8. References

1. **https://www.mayoclinic.org/diseases-conditions/glioma/symptoms-causes/syc-20350251**
2. **https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5563115/**
3. **https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2940552/**
4. **https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2761018/**
5. **https://www.hopkinsmedicine.org/health/conditions-and-diseases/gliomas**
6. **https://docs.gdc.cancer.gov/Data/Bioinformatics_Pipelines/CNV_Pipeline/**
7. **https://portal.gdc.cancer.gov/**
8. **https://www.bioconductor.org/packages/3.3/bioc/vignettes/TCGAbiolinks/inst/doc/tcgaBiolinks.html?fbclid=IwAR0VWPSCh9PikN4t_1IPJRdH6-SuSoX6ro-o1oNM7FoKNWSkDRUeG2phCjk#tcgaanalyze_dea-tcgaanalyze_leveltab-differential-expression-analysis-dea**
9. **http://uswest.ensembl.org/Homo_sapiens/Gene/Summary?g=ENSG00000146648;r=7:55019021-55211628**
10. **https://www.genecards.org/cgi-bin/carddisp.pl?gene=ENSG00000264545**