**cBioPortal Dataset**

**Introduction**

This project involves using a open source genomics database and performing queries on the data. MongoDB will be the primary database used to make queries on the data.

**Database Information and Objective**

<https://www.cbioportal.org/study/summary?id=lgg_tcga>

<https://bit.ly/3MLtjFB> ⇒ Shortened URL with Query Information

All data was collected from the Brain Lower Grade Glioma (TCGA< Firehose Legacy) repository from the cBioPortal Dataset. From this repository, I specifically selected Oligoastrocytoma, Anaplastic Astrocytoma, Oligodendroglioma which represented 376 samples (134 Oligoastrocytoma, 130 Anaplastic Astrocytoma, Oligodendroglioma)

I also selected the genetic information for IDH1 gene expression from these which was expressed in 77.6% frequency in all the mutations throughout these 3 cancer lines. The IDH1 gene information from the data also contains phenotypic clinical data for this gene.

**My overall objective is to demonstrate a difference between the 3 cancer cell using this IDH1 gene (and its associated phenotype data) and analyze the clinical data provided for the samples.**

**IDH1 Gene Information**

IDH mutations can dominantly inhibit WT-IDH (Isocitrate Dehydrogenase) when heterozygous through the formation of enzymatically inactive heterodimers ([**Zhao et al., 2009**](https://www.frontiersin.org/articles/10.3389/fnmol.2021.722396/full#B104)). It was shown by [**Uhm (2010)**](https://www.frontiersin.org/articles/10.3389/fnmol.2021.722396/full#B82) that IDH mutations lead to the acquisition of a new enzymatic function that catalyzes the formation of D-2HG from alpha-KG. 2-HG can inhibit alpha-KG dependent dioxygenases and cause epigenetic alterations ([**Xu et al., 2011**](https://www.frontiersin.org/articles/10.3389/fnmol.2021.722396/full#B97) ). It can also stimulate the activity of EGLN leading to decreased HIF levels. This in turn allows tumor proliferation in low oxygen conditions ([**Zhao et al., 2009**](https://www.frontiersin.org/articles/10.3389/fnmol.2021.722396/full#B104) ; **[Koivunen et al., 2012](https://www.frontiersin.org/articles/10.3389/fnmol.2021.722396/full" \l "B48)**).

**Cancer Cell Types and Relation to IDH1**

**Oligodendroglioma**

Oligodendroglioma is genetically defined as a tumor confirmed to harbor either an IDH1 or IDH2 mutation along with co-deletion of chromosome arms 1p and 19q. Histologically, oligodendroglial tumors show sheets of isomorphic round nuclei with a clear cytoplasm—the classic “fried egg” appearance. Grade III oligodendroglioma show a worse prognosis than grade II tumors due to the presence of anaplastic features such as nuclear atypia, necrosis, microvascular proliferation, high cell density and number of mitotic figures.[1]

**Anaplastic Astrocytoma**

Anaplastic astrocytomas are aggressive glial cancers that present poor prognosis and high recurrence. Heterozygous IDH1 R132H mutations are common in adolescent and young adult anaplastic astrocytomas. In a majority of cases, the IDH1 R132H mutation is unique to the tumor, although rare cases of anaplastic astrocytoma have been described in patients with mosaic IDH1 mutations[2]

**Oligoastrocytoma**

Oligoastrocytoma emerges from both astocytes and oligodendrocyte cell lines.

Anaplastic oligoastrocytoma is a brain tumor that forms when two types of cells in the brain, called oligodendrocytes and astrocytes, rapidly increase in number to form a mass. These brain cells are known as glial cells, which normally protect and support nerve cells in the brain. Because an oligoastrocytoma is made up of a combination of two cell types, it is known as a mixed glioma[3]

Similarly, high frequencies of ***IDH1*** mutations were found in oligodendrogliomas (79%) and oligoastrocytomas (94%). Analyses of multiple biopsies from the same patient (51 cases) showed that there were no cases in which an ***IDH1*** mutation occurred after the acquisition of either a ***TP53*** mutation or loss of 1p/19q, suggesting that ***IDH1*** mutations are very early events in gliomagenesis and may affect a common glial precursor cell population [4]

**Initial Data Wrangling**

**Importing Data**

I will be using MongoDB Atlas (a cloud version of MongoDB).

TCGA files are tsv files and I used a tsv to json website to convert all necessary files to JSON format.

There are 3 tsv files that were downloaded from the TCGA website

1. Clinical Data ⇒ Contains the clinical data information for each of the samples (196 samples)
2. Clinical Data Secondary ⇒ Contains additional clinical data (196 samples)
3. IDH1 ⇒ Contains genetic and phenotypic IDH1 information related to samples (156 samples)

Creating a new database in MongoDB and creating a new collection inside the database

Atlas atlas-c25k3z-shard-0 [primary] ClusterTRGN> use FinalProject

switched to db FinalProject

Atlas atlas-c25k3z-shard-0 [primary] FinalProject> db.createCollection("Clinical\_Data")

{ ok: 1 }

Atlas atlas-c25k3z-shard-0 [primary] FinalProject> show collections

Clinical\_Data

Atlas atlas-c25k3z-shard-0 [primary] FinalProject> db.createCollection("Clinical\_Data\_Secondary")

{ ok: 1 }

Atlas atlas-c25k3z-shard-0 [primary] FinalProject> db.createCollection("IDH1")

{ ok: 1 }

Atlas atlas-c25k3z-shard-0 [primary] FinalProject> show collections

Clinical\_Data

Clinical\_Data\_Secondary

IDH1

Importing will be done on the local computer (since the files are there) using the following command. This command is repeated multiple times to import files to their respective collection:

mongoimport --uri mongodb+srv://bambat:<PASSWORD>@clustertrgn.eyagghm.mongodb.net/<DATABASE> --collection <COLLECTION> --type <FILETYPE> --file <FILENAME>

2023-04-19T18:53:26.856-0700 196 document(s) imported successfully. 0 document(s) failed to import.

Here are the results of that command:

Atlas atlas-c25k3z-shard-0 [primary] FinalProject> show collections

Clinical\_Data

Clinical\_Data\_Secondary

IDH1

Note that although there were 196 samples in the Clinical Data and Secondary Clinical Data tsv, the IDH1 had 156 samples. I attempted to inner join all these collections into 1 large collection (with 156 samples). However, the sample IDs for all the tsvs are different. Unfortunately, I will not be able to combine them into one collection. Instead, since each collection has unique headers, I will be querying from the relavent collection(s) for each of the 5 queries.

**Query # 1**

**Is there a difference in incidence of Anaplastic Astrocytoma between Males and Females?**

Anaplastic Astrocytoma is known for being an especially difficult brain cancer to treat due to its advanced stage. I am interested to see if there is a sex difference in prevalance of this specific cancer. “A male: female ratio 1.18:1 in low-grade astrocytomas is there. In anaplastic astrocytoma, there is a substantial male dominance, with a male: female incidence being 1.87:1” [5].

The Clinical\_Data contain cancer type and sex information (for different TCGA patients). So, I’ll be querying on the Clinical\_Data collection

Here is the query for the Clinical\_Data collection:

db.Clinical\_Data.countDocuments({"Cancer Type Detailed": "Anaplastic Astrocytoma", "Sex": "Male"})

db.Clinical\_Data.countDocuments({"Cancer Type Detailed": "Anaplastic Astrocytoma", "Sex": "Female"})

Here are the results for those queries:

Atlas atlas-c25k3z-shard-0 [primary] FinalProject> db.Clinical\_Data.countDocuments({"Cancer Type Detailed": "Anaplastic Astrocytoma", "Sex": "Male"})

34

Atlas atlas-c25k3z-shard-0 [primary] FinalProject> db.Clinical\_Data.countDocuments({"Cancer Type Detailed": "Anaplastic Astrocytoma", "Sex": "Female"})

26

Although it is a small sample size, the results from our query show there are slightly more males than females in this dataset (though its not at the 1.87:1 ratio that was suggested by the study)

**Query #2**

**What is the number of patients with Anaplastic Astrocytoma, Oligodendrogliooma, and Oligoastrocytoma with a "TMB (nonsynonymous)" of greater than or equal to 1?**

“Defective DNA repair leads to higher tumor mutational burden (TMB) which is defined as the total number of nonsynonymous mutations per megabase (Mb) of coding regions of a tumor genome” [7]

“TMB was higher for the group of mutant genes that are frequently mutated in glioblastomas (GBMs) and lower for the group of mutant genes that are frequently mutated in lower-grade gliomas (LGGs). Patients with a higher TMB exhibited shorter overall survival” [8]

I am interested to see if there is a difference in outcomes for those with a TMB value of greater than 1. A greater than 1 value indicates more aggressive form of tumor (although other variables like age, comorbidities etc do play a role).

Anaplastic Astrocytoma should have the highest TMB value followed by the other 2 cancer lines

Here is the query on the Clinical\_Data collection to answer:

db.Clinical\_Data.count({"TMB (nonsynonymous)": {$gte: 1}, "Cancer Type Detailed": "Anaplastic Astrocytoma"})

db.Clinical\_Data.count({"TMB (nonsynonymous)": {$gte: 1}, "Cancer Type Detailed": "Oligodendroglioma"})

db.Clinical\_Data.count({"TMB (nonsynonymous)": {$gte: 1}, "Cancer Type Detailed": "Oligoastrocytoma"})

Here are the results for the query:

Atlas atlas-c25k3z-shard-0 [primary] FinalProject> db.Clinical\_Data.count({"TMB (nonsynonymous)": {$gte: 1}, "Cancer Type Detailed": "Anaplastic Astrocytoma"})

27

Atlas atlas-c25k3z-shard-0 [primary] FinalProject> db.Clinical\_Data.count({"TMB (nonsynonymous)": {$gte: 1}, "Cancer Type Detailed": "Oligodendroglioma"})

13

Atlas atlas-c25k3z-shard-0 [primary] FinalProject> db.Clinical\_Data.count({"TMB (nonsynonymous)": {$gte: 1}, "Cancer Type Detailed": "Oligoastrocytoma"})

21

These are the expected results.

**Query # 3**

**For the cancer types (Oligoastrocytoma, Oligodendroglioma, and Anaplastic Astrocytoma), what is the average value for the “Fraction Genome Altered” column for each of the cancer types?**

“In this [cancer stage] system, oligodendrogliomas and oligoastrocytomas are usually grade II or grade III tumors. Grade II tumors are considered low-grade tumors, which generally grow at a slower rate than grade III tumors. Grade II tumors may evolve over time into grade III tumors. Grade III tumors are anaplastic, or malignant tumors”[6]

Based on this data, I hypothesize that the Anaplastic Astrocytoma (Stage III minimum) would have the highest fraction of the genome altered. Since the Oligoastrocytoma and the Oligodendroglioma are not anaplastic, they would be considered lower stage cancers, and therefore, would have a lower fraction of genome altered.

I used the following code to update the Fraction Genome Altered column to a float variable in the collection:

db.Clinical\_Data.updateMany(

{},

[{$set: {"Fraction Genome Altered": { $toDouble: "$Fraction Genome Altered" }}}])

To answer this quesiton, the following is the MongoDB query on the Clinical\_Data collection:

db.Clinical\_Data.aggregate([

{

$group: {

\_id: "$Cancer Type Detailed",

avgFractionGenomeAltered: { $avg: "$Fraction Genome Altered" }

}

}

])

Here are the results of the query:

[

{

\_id: 'Anaplastic Astrocytoma', avgFractionGenomeAltered: 0.16760666666666665

},

{ \_id: 'Oligoastrocytoma', avgFractionGenomeAltered: 0.0939 },

{ \_id: 'Oligodendroglioma', avgFractionGenomeAltered: 0.106325 }

]

Across all the samples, the Anaplastic Astrocytoma contained the highest average fraction of genome altered . These are expected results.

**Query #4**

**What are the average allele frequencies for Anaplastic Astrocytoma for those that had a diagnosis age above or below 50?**

Out of all 3 cancer types, Anaplastic Astrocytoma is the most aggressive form. Age plays a significant role in many cancers. I wanted to see if there is a difference in the minor allele frequency for Anaplastic Astrocytoma with respect to age

Here is the query for above 50 and below 50:

db.IDH1.aggregate([

{ $match: { "Cancer Type Detailed": "Anaplastic Astrocytoma", "Diagnosis Age": { $lt: 50 } } },

{ $group: { \_id: null, avg\_allele\_freq: { $avg: "$Allele Freq (T)" } } }

])

The following are the results of the for 50 and above:

Atlas atlas-c25k3z-shard-0 [primary] FinalProject> db.IDH1.aggregate([ { $match: { "Cancer Type Detailed": "Anaplastic Astrocytoma", "Diagnosis Age": { $gte: 50 } } }, { $group: { \_id: null, avg\_allele\_freq: { $avg: "$Allele Freq (T)" } } }])

[ { \_id: null, avg\_allele\_freq: 0.3271428571428571 } ]

The following are the results for below 50:

Atlas atlas-c25k3z-shard-0 [primary] FinalProject> db.IDH1.aggregate([ { $match: { "Cancer Type Detailed": "Anaplastic Astrocytoma", "Diagnosis Age": { $gte: 50 } } }, { $group: { \_id: null, avg\_allele\_freq: { $avg: "$Allele Freq (T)" } } }])

[ { \_id: null, avg\_allele\_freq: 0.3403225806451613 } ]

**Query #5**

**What are the counts for each option in the First Symptom Longest Duration field for each of the 3 cancer types?**

db.Clinical\_Data.aggregate([

{ $match: { "First symptom longest duration": { $in: ["0 - 30 Days", "31 - 90 Days", "91 - 180 Days", "> 181 Days"] } } },

{ $group: { \_id: { First\_Symptom\_Longest\_Duration: "$First symptom longest duration", cancer: "$Cancer Type Detailed" }, count: { $sum: 1 } } },

{ $project: { First\_Symptom\_Longest\_Duration: "$\_id.First\_Symptom\_Longest\_Duration", cancer: "$\_id.cancer", count: 1, \_id: 0 }

}])

Here is the output to the code:

[

{

count: 32,

First\_Symptom\_Longest\_Duration: '0 - 30 Days',

cancer: 'Anaplastic Astrocytoma'

},

{

count: 12,

First\_Symptom\_Longest\_Duration: '31 - 90 Days',

cancer: 'Oligoastrocytoma'

},

{

count: 8,

First\_Symptom\_Longest\_Duration: '> 181 Days',

cancer: 'Oligoastrocytoma'

},

{

count: 5,

First\_Symptom\_Longest\_Duration: '91 - 180 Days',

cancer: 'Oligodendroglioma'

},

{

count: 37,

First\_Symptom\_Longest\_Duration: '0 - 30 Days',

cancer: 'Oligoastrocytoma'

},

{

count: 15,

First\_Symptom\_Longest\_Duration: '> 181 Days',

cancer: 'Oligodendroglioma'

},

{

count: 6,

First\_Symptom\_Longest\_Duration: '91 - 180 Days',

cancer: 'Oligoastrocytoma'

},

{

count: 11,

First\_Symptom\_Longest\_Duration: '> 181 Days',

cancer: 'Anaplastic Astrocytoma'

},

{

count: 5,

First\_Symptom\_Longest\_Duration: '31 - 90 Days',

cancer: 'Anaplastic Astrocytoma'

},

{

count: 9,

First\_Symptom\_Longest\_Duration: '31 - 90 Days',

cancer: 'Oligodendroglioma'

},

{

count: 25,

First\_Symptom\_Longest\_Duration: '0 - 30 Days',

cancer: 'Oligodendroglioma'

},

{

count: 6,

First\_Symptom\_Longest\_Duration: '91 - 180 Days',

cancer: 'Anaplastic Astrocytoma'

}

]

Overall, all 3 cancer types had the highest count of First Symptom Longest duration in the 0-30 Day range. Also, it seems that the Anaplastic Astrocytoma and Oligoastrocytoma had the highest counts from 0-30 Days and 31-90 Days.

It seems that for all these cancers. First Symptom duration often appears mostly commonly in the 0-30 range.

**RShiny App**

For my RShiny app, I was interested to see if I can combine Queries 3 and 5 into a useful graph. I selected the First\_Symptom\_Longest\_Duration and the Allele Frequency (T) Column. I was curious to see if there are any trends across the samples

I took the IDH1 tsv file and modified it to remove any irrelavent columns. I also removed any samples that had NULL values in the First\_Symptom\_Longest\_Duration and the Allele Frequency (T) Column.

So, I created dropdown menu to select between the Cancer Type

Below are screenshots of the app:

A picture containing diagram

Description automatically generated

A picture containing diagram

Description automatically generated

A picture containing diagram

Description automatically generated

Anaplastic Astrocytoma has the highest Allele Frequency for greater than >181 Days. Overall, Oligoastrocytoma and the Oligodendroglioma had many more samples in the 0-30 Day and 31-90 range for the First Symptom Longest Duration

**References**

[1] Bou Zerdan, M., & Assi, H. I. (2021, September 15). *Oligodendroglioma: A Review of Management and Pathways*. Frontiers. Retrieved April 26, 2023, from <https://www.frontiersin.org/articles/10.3389/fnmol.2021.722396/full>

[2] Lee, S., Kambhampati, M., Almira-Suarez, M. I., Ho, C.-Y., Panditharatna, E., Berger, S. I., Turner, J., Van Mater, D., Kilburn, L., Packer, R. J., Myseros, J. S., Vilain, E., Nazarian, J., & Bornhorst, M. (2019, December 16). *Somatic mosaicism of Idh1 r132h predisposes to anaplastic astrocytoma: A case of two siblings*. Frontiers. Retrieved April 26, 2023, from <https://www.frontiersin.org/articles/10.3389/fonc.2019.01507/full#:~:text=Anaplastic> astrocytomas are aggressive glial,and young adult anaplastic astrocytomas.

[3] U.S. Department of Health and Human Services. (n.d.). *Anaplastic oligoastrocytoma - about the disease*. Genetic and Rare Diseases Information Center. Retrieved April 26, 2023, from <https://rarediseases.info.nih.gov/diseases/10637/anaplastic-oligoastrocytoma>

[4] Watanabe T, Nobusawa S, Kleihues P, Ohgaki H. IDH1 mutations are early events in the development of astrocytomas and oligodendrogliomas. Am J Pathol. 2009 Apr;174(4):1149-53. doi: 10.2353/ajpath.2009.080958. Epub 2009 Feb 26. PMID: 19246647; PMCID: PMC2671348.

[5] Kapoor M, Gupta V. Astrocytoma. [Updated 2022 Oct 3]. In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; 2023 Jan-. Available from: <https://www.ncbi.nlm.nih.gov/books/NBK559042/>

[6] *Oligodendroglioma oligoastrocytoma 200526 - American brain tumor ...* (n.d.). Retrieved April 26, 2023, from <https://www.abta.org/wp-content/uploads/2018/03/Oligodendroglioma_Oligoastrocytoma_2020_web_en.pdf>

[7] Asmann, Y.W., Parikh, K., Bergsagel, P.L. *et al.* Inflation of tumor mutation burden by tumor-only sequencing in under-represented groups. \*npj Precis. Onc.\***5**, 22 (2021). <https://doi.org/10.1038/s41698-021-00164-5>

[8]Wang, L., Ge, J., Lan, Y. *et al.* Tumor mutational burden is associated with poor outcomes in diffuse glioma. *BMC Cancer* **20**, 213 (2020). <https://doi.org/10.1186/s12885-020-6658-1>

Cerami et al. The cBio Cancer Genomics Portal: An Open Platform for Exploring Multidimensional Cancer Genomics Data. Cancer Discovery. May 2012 2; 401. [PubMed](https://www.ncbi.nlm.nih.gov/pubmed/22588877).

Gao et al. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. Sci. Signal. 6, pl1 (2013). [PubMed](https://www.ncbi.nlm.nih.gov/pubmed/23550210).