Group – 5: Prasanth Kumar Thuthika, Tejaswini Repala

Somatic Variant Analysis in Mouse Glioma Tumor Samples Using Whole Genome Sequencing Data and Genome Analysis Tool Kit (GATK)

Introduction

Somatic variant calling is a pivotal process in genomics research, enabling the identification of genetic mutations that arise in somatic cells during an individual's lifetime. These mutations, which are not inherited, often play significant roles in tumor development, progression, and response to therapies. By analysing somatic variants, researchers can uncover critical insights into the genetic underpinnings of diseases like cancer, facilitating the development of targeted therapies and personalized medicine approaches in clinical settings.

The Genome Analysis Toolkit (GATK) is a widely-used computational toolkit for processing high-throughput sequencing data. Renowned for its accuracy and efficiency, GATK provides robust workflows for variant discovery, including tools specifically designed for somatic variant calling.

Objective

The objective of this project is to conduct a comprehensive somatic variant analysis on two types of glioma tumors samples in mouse model. Utilizing whole genome sequencing (WGS) data, this study aims to identify and characterize somatic mutations that distinguish the tumor types. The analysis focuses on detecting single nucleotide variants (SNVs) and insertions/deletions (indels), which are key mutation types with potential implications for tumor behaviour and progression. The variants are filtered and then annotated to identify their genomic location, functional effects, and involvement in key biological pathways.

Data

The data was obtained from NCBI-SRA, Illumina short-read whole genome paired-end sequencing data. Three samples were selected: Control, Tumor-1 (H3.3 S31A glioma tumor), and Tumor-2 (H3.3 K27M glioma tumor).

Data Source	NCBI SRA
Accession Number	PRJNA961166
Samples	Control Tumor-1: H3.3 S31A glioma tumor Tumor-2: H3.3 K27M glioma tumor
Type	WGS
Reads	Paired-end

Reference Data

Genome: Mus musculus GRCm39

Annotation: Mus musculus.GRCm39.109.gff3

Tools

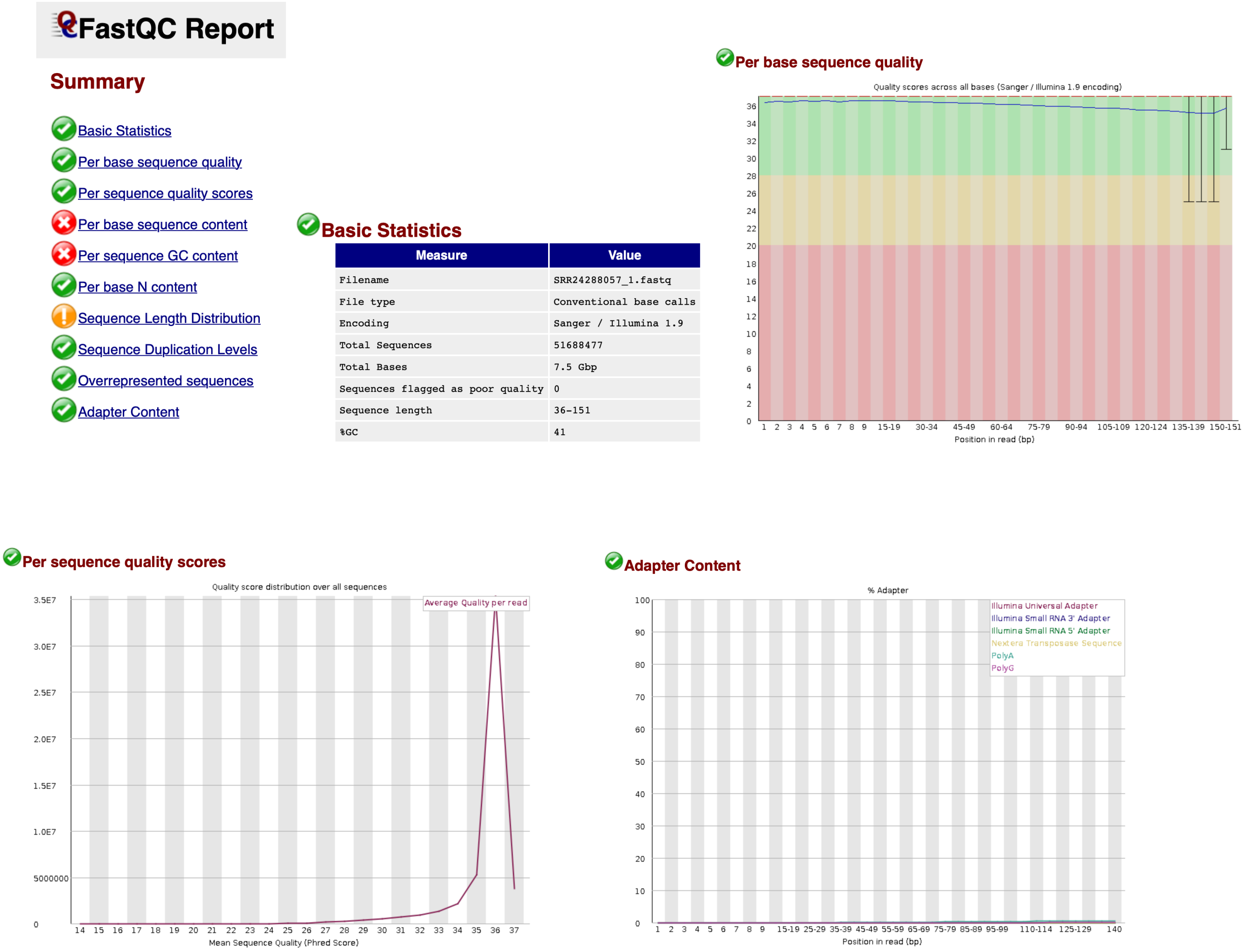
FASTQC, Trim Galore, BWA, Picard, GATK, IGV, ANNOVAR, ClueGO

Workflow

The proposed workflow for variant calling begins with a FASTQC quality check to assess the data. Next, trimming is performed using Trim Galore to remove any adapter contamination and low-quality reads. The cleaned reads are then mapped to the reference genome using the BWA tool. Following this, duplicates are marked, and base quality recalibration is performed to ensure accurate variant calling. Variant calling is conducted using GATK Mutect, followed by filtering using GATK's filtering options. The filtered variants are then annotated using ANNOVAR, and functional enrichment analysis is performed using Gene Ontology (GO).

FASTQC

The initial FASTQC results indicate that the read quality is good, and there is no adapter contamination in the files. Since the quality is acceptable, no additional trimming or quality filtering steps were performed. Below are the FASTQC reports for one of the samples, with similar patterns observed for the remaining sample files.



FastQC Report

Summary

- ✓ Basic Statistics
- ✓ Per base sequence quality
- ✓ Per sequence quality scores
- ✗ Per base sequence content
- ✗ Per sequence GC content
- ✓ Per base N content
- ! Sequence Length Distribution
- ✓ Sequence Duplication Levels
- ✓ Overrepresented sequences
- ✓ Adapter Content

✓ Basic Statistics

Measure	Value
Filename	SRR24288057_1.fastq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	51688477
Total Bases	7.5 Gbp
Sequences flagged as poor quality	0
Sequence length	36-151
%GC	41

✓ Per base sequence quality

✓ Per sequence quality scores

✓ Adapter Content

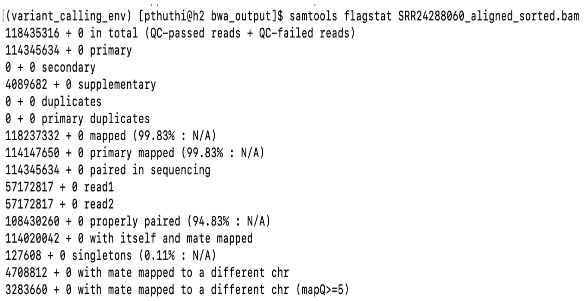
Figures showing the FASTQC results for one of the samples

Trim Galore

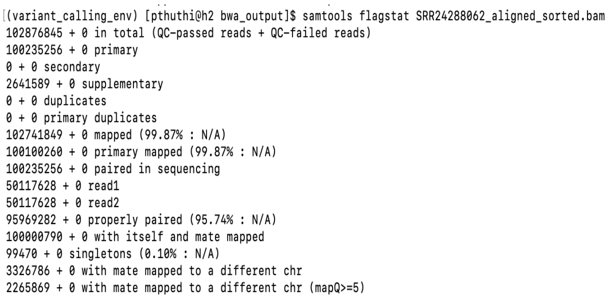
The reads and the quality of the data are good, and there is no adapter contamination therefore, trimming was not performed.

Reference Mapping

The reads were mapped to the reference genome (Mus musculus GRCm39), using BWA and the mapping statistics indicate good alignment for all samples, with over 99% of the reads successfully mapped to the reference genome. Among these, 90.6% of the paired-end reads were properly aligned, indicating good pairing and accurate mapping. Anomalies were minimal, with only 0.07% singleton reads and less than 7% multi-chromosomal mappings, ensuring the integrity and reliability of the mapped data for subsequent analyses.



```
[(variant_calling_env) [pthuthi@h2 bwa_output]$ samtools flagstat SRR24288060_aligned_sorted.bam
118435316 + 0 in total (QC-passed reads + QC-failed reads)
114345634 + 0 primary
0 + 0 secondary
4089682 + 0 supplementary
0 + 0 duplicates
0 + 0 primary duplicates
118237332 + 0 mapped (99.83% : N/A)
114147650 + 0 primary mapped (99.83% : N/A)
114345634 + 0 paired in sequencing
57172817 + 0 read1
57172817 + 0 read2
108430260 + 0 properly paired (94.83% : N/A)
114020042 + 0 with itself and mate mapped
127608 + 0 singletons (0.11% : N/A)
4708812 + 0 with mate mapped to a different chr
3283660 + 0 with mate mapped to a different chr (mapQ>=5)
```



```
[(variant_calling_env) [pthuthi@h2 bwa_output]$ samtools flagstat SRR24288062_aligned_sorted.bam
102876845 + 0 in total (QC-passed reads + QC-failed reads)
100235256 + 0 primary
0 + 0 secondary
2641589 + 0 supplementary
0 + 0 duplicates
0 + 0 primary duplicates
102741849 + 0 mapped (99.87% : N/A)
100100260 + 0 primary mapped (99.87% : N/A)
100235256 + 0 paired in sequencing
50117628 + 0 read1
50117628 + 0 read2
95969282 + 0 properly paired (95.74% : N/A)
100000790 + 0 with itself and mate mapped
99470 + 0 singletons (0.10% : N/A)
3326786 + 0 with mate mapped to a different chr
2265869 + 0 with mate mapped to a different chr (mapQ>=5)
```

Figures showing the mapping statistics for the two tumor variant samples

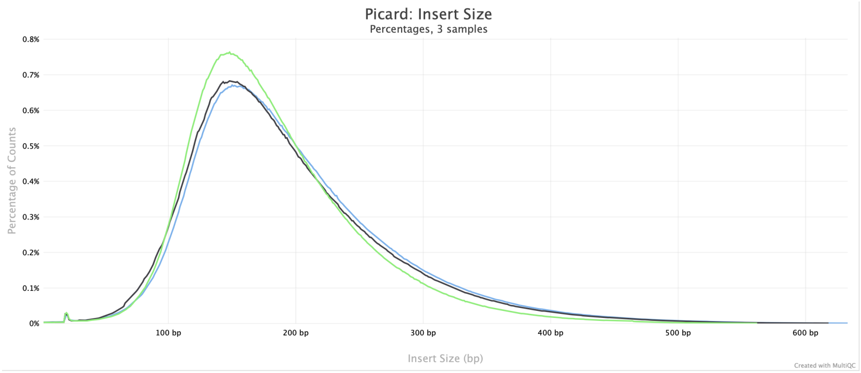
Marking Duplicates

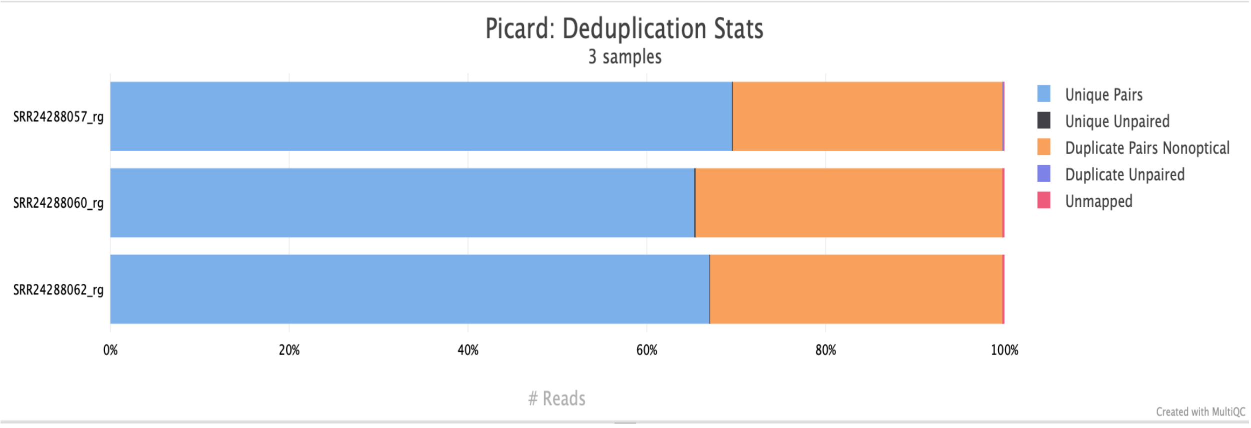
Involves identifying and marking duplicate reads generated during PCR amplification or optical duplicates produced during sequencing. Using the Picard tool, duplicate reads are flagged but not removed, ensuring that all data remains available for downstream analysis. This process reduces biases in variant calling by preventing the overrepresentation of duplicate reads, which could falsely inflate the confidence of certain mutations. Properly marking duplicates is especially important in whole-genome sequencing studies, as it ensures accurate variant allele frequency calculate ions and improves the reliability of downstream analyses.

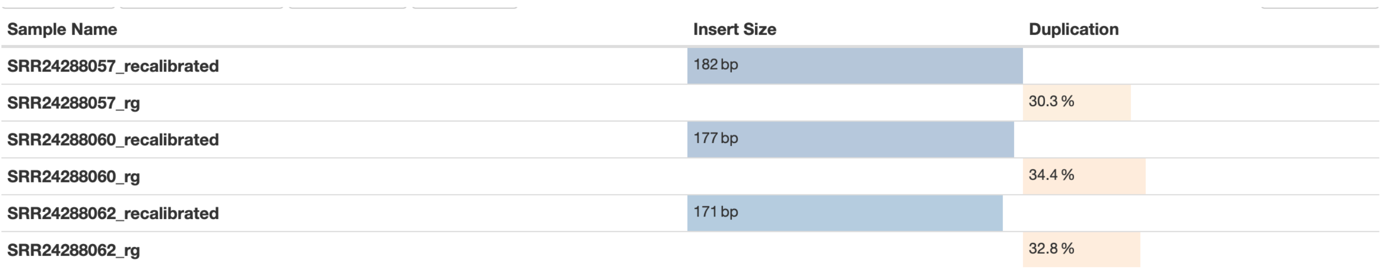
From the Picard duplication metrics, we observe the percentage of each type of duplicate. The plot shows that 60-65% of the reads are unique, while approximately 35% are non-optical duplicates. The insert size distribution is consistent, with minimal variation, indicating uniform fragment sizes across the dataset.

Base Quality Recalibration Score

Base quality score recalibration (BQSR) is a crucial step performed with GATK's BaseRecalibrator tool. It adjusts the quality scores assigned to sequencing reads to correct for systematic errors introduced by the sequencing machine. By using a database of known variant sites, this step identifies and accounts for machine-specific biases, sequence context effects, and other artifacts. Correcting these errors ensures that base quality scores more accurately reflect the probability of a sequencing error. This, in turn, improves the accuracy of variant calling, as reliable quality scores are critical for distinguishing true mutations from sequencing artifacts.







Sample Name	Insert Size	Duplication
SRR24288057_recalibrated	182 bp	
SRR24288057_rg		30.3 %
SRR24288060_recalibrated	177 bp	
SRR24288060_rg		34.4 %
SRR24288062_recalibrated	171 bp	
SRR24288062_rg		32.8 %

Figures showing the results generated by Picard

Somatic Variant Calling

Variant Calling was performed using GATK's Mutect. The results of the somatic variant calling are presented below. IGV was used to visualize the variants.

➢ 3027 unique_H3.3 S31A glioma tumor_variants

➢ 2690 unique_H3.3 K27M glioma tumor_variants

➢ 205 shared_variants

H3.3 S31A glioma tumor	**H3.3 K27M glioma tumor**
➢ **Total Variants: 3,232**	➢ **Total Variants: 2,895**
➢ **SNPs: 2,614**	➢ **SNPs: 2,336**
➢ **Insertions/Deletions: 448**	➢ **Insertions/Deletions: 398**
➢ **Chromosomes with high Variant Count:**	➢ **Chromosomes with high Variant Count:**
Chr_1: 258 variants	**Chrom_8: 232 variants**
Chr_5: 257 variants	**Chrom_9: 190 variants**
Chr_8: 235 variants	

Variant Annotation

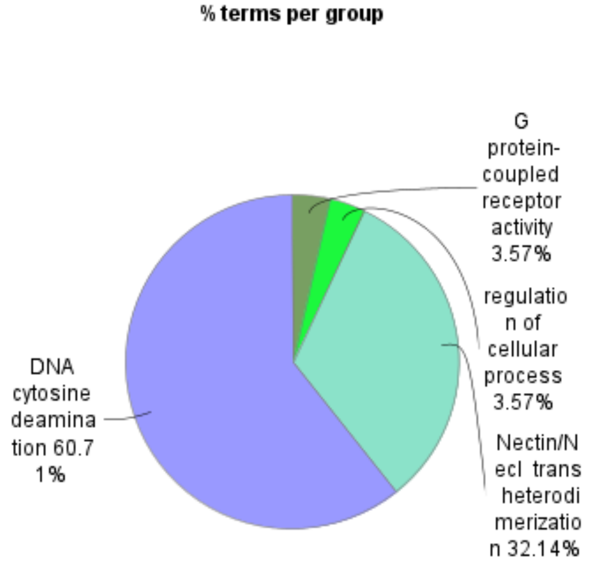
The filtered variants were annotated using ANNOVAR with the Mus_musculus.GRCm39.109.gff3 annotation file. The below table shows the results of annotated variants

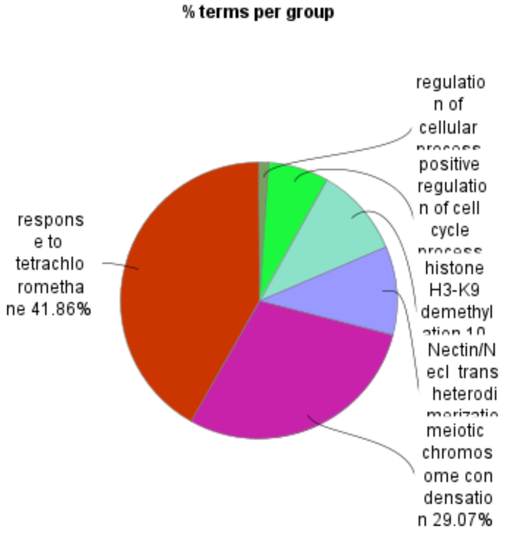
High Impact Genes

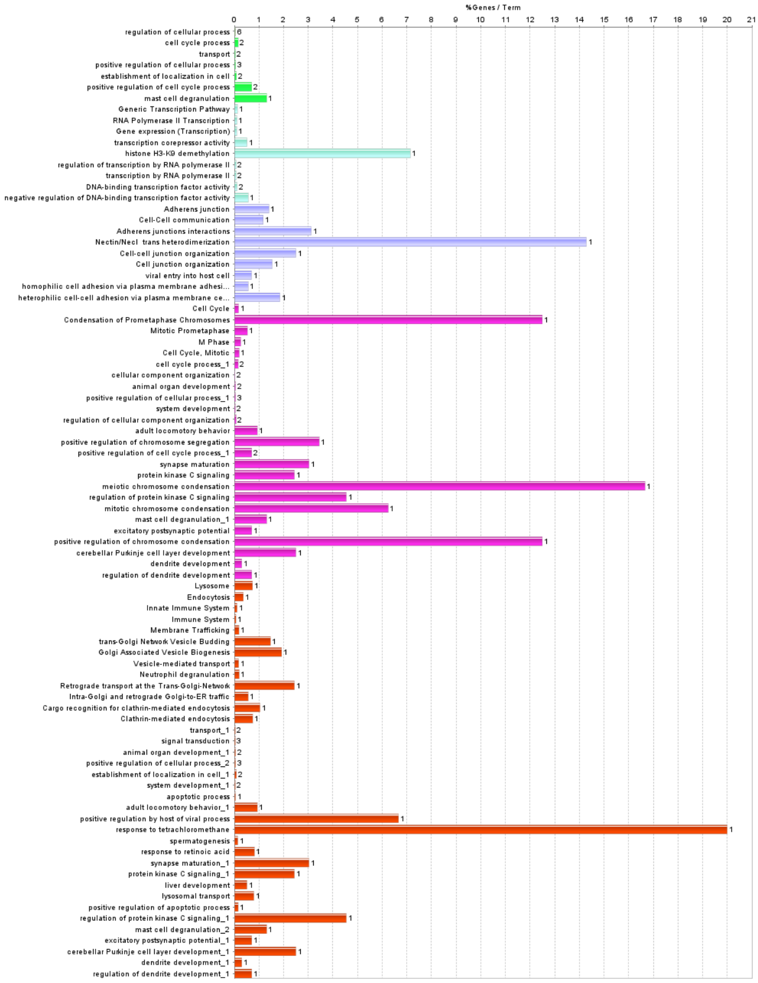
H3.3 S31A glioma tumor	H3.3 K27M glioma tumor
Gm10354 Igf2r Mrgpra9 Ncapd2 Nectin4 Zfp710	Apobec3 Nectin4 Vmn2r43

Functional Enrichment Analysis using ClueGO

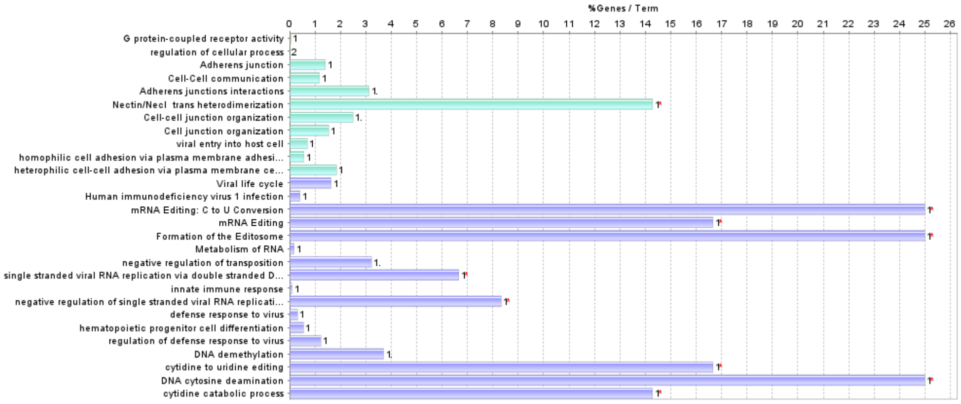
Category	H3.3 S31A glioma tumor	H3.3 K27M glioma tumor
UTR3	3	5
UTR5	1	1
Exonic	15	3
Intergenic	303	216
Intronic	150	119
ncRNA Exonic	2	3
NcRNA intronic	10	10
Splicing	1	1
Upstream	6	7
Nonframeshift Substitution	3	0
Nonsynonymous SNV	6	2
Synonymous SNV	6	1







The gene IDs were retrieved using gene names through Ensembl BioMart, and functional enrichment analysis was performed using the ClueGO tool. The results revealed key processes in two variants of Glioma tumors: H3.3 S31A glioma tumor was dominated by DNA cytosine deamination (60.71%), a process critical for epigenetic modifications and immune responses, such as RNA editing and antiviral defense. H3.3 K27M glioma tumor highlighted the response to tetrachloromethane (41.86%), indicating detoxification and stress-response mechanisms, along with histone H3-K9 demethylation (29.07%), which suggests chromatin remodeling and transcriptional activation. These findings underscore the interplay of immune activation, epigenetic changes, and stress adaptation mechanisms in glioma tumor biology



Figures showing the functional enrichment using ClueGO

Discussions and Conclusions

This study analyzed somatic variants in two glioma subtypes, H3.3 S31A and H3.3 K27M, using whole genome sequencing and bioinformatics tools. H3.3 S31A had a higher mutation burden (3,232 variants) than H3.3 K27M (2,895 variants), with key differences in chromosomal distribution and high-impact genes such as Igf2r and Vmn2r43. Functional enrichment revealed distinct pathways: DNA cytosine deamination for epigenetic regulation in H3.3 S31A, and stress response mechanisms like detoxification in H3.3 K27M. These findings highlight critical biological processes and potential therapeutic targets in glioma tumorigenesis.

Github: https://github.iu.edu/pthuthi/Project.git

References:

1. A three-caller pipeline for variant analysis of cancer whole-exome sequencing dataLiu et al.
https://pmc.ncbi.nlm.nih.gov/articles/PMC5428716/?t

2. Whole Genome Sequencing of the Mutamouse Model Reveals Strain- and Colony-Level Variation, and Genomic Features of the Transgene Integration SiteMeier et al.
https://www.nature.com/articles/s41598-019-50302-0?t

3. Optimized pipeline of MuTect and GATK tools to improve the detection of somatic single nucleotide polymorphisms in whole-exome sequencing data - BMC BioinformaticsValle et al.
https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-016-1190-7?t

4. An optimized GATK4 pipeline for Plasmodium falciparum whole genome sequencing variant calling and analysisJA;
https://pubmed.ncbi.nlm.nih.gov/37420214/

5. Comparative analysis of somatic variant calling on matched FF and FFPE WGS samples
de Schaetzen van Brienen L;Larmuseau M;Van der Eecken K;De Ryck F;Robbe P;Schuh A;Fostier J;Ost P;Marchal K;
https://pubmed.ncbi.nlm.nih.gov/32631411/