BioMedInformatics

MDPI

Article

Genetic Optimization in Uncovering Biologically Meaningful Gene Biomarkers for Glioblastoma Subtypes

Petros Paplomatas, Ioanna-Efstathia Douroumi, Panagiotis Vlamos and Aristidis Vrahatis *

Bioinformatics and Human Electrophysiology Laboratory, Department of Informatics, Ionian University, 49100 Corfu, Greece; p.paplomatas@ionio.gr (P.P.); stadour@hotmail.com (I.-E.D.); vlamos@ionio.gr (P.V.)
* Correspondence: aris.vrahatis@ionio.gr

Abstract: Background: Glioblastoma multiforme (GBM) is a highly aggressive brain cancer known for its challenging survival rates; it is characterized by distinct subtypes, such as the proneural and mesenchymal states. The development of targeted therapies is critically dependent on a thorough understanding of these subtypes. Advances in single-cell RNA-sequencing (scRNA-seq) have opened new avenues for identifying subtype-specific gene biomarkers, which are essential for innovative treatments. Methods: This study introduces a genetic optimization algorithm designed to select a precise set of genes that clearly differentiate between the proneural and mesenchymal GBM subtypes. By integrating differential gene expression analysis with gene variability assessments, our dual-criterion strategy ensures the selection of genes that are not only differentially expressed between subtypes but also exhibit consistent variability patterns. This approach enhances the biological relevance of identified biomarkers. We applied this algorithm to scRNA-seq data from GBM samples, focusing on the discovery of subtype-specific gene biomarkers. Results: The application of our genetic optimization algorithm to scRNA-seq data successfully identified significant genes that are closely associated with the fundamental characteristics of GBM. These genes show a strong potential to distinguish between the proneural and mesenchymal subtypes, offering insights into the molecular underpinnings of GBM heterogeneity. Conclusions: This study introduces a novel approach for biomarker discovery in GBM that is potentially applicable to other complex diseases. By leveraging scRNA-seq data, our method contributes to the development of targeted therapies, highlighting the importance of precise biomarker identification in personalized medicine.

Keywords: genetic optimization; feature selection; single-cell RNA-seq; glioblastoma; proneural; mesenchymal

Citation: Paplomatas, P.; Douroumi, I.-E.; Vlamos, P.; Vrahatis, A. Genetic Optimization in Uncovering Biologically Meaningful Gene Biomarkers for Glioblastoma Subtypes. *BioMedInformatics* **2024**, *4*, 811–822. https://doi.org/10.3390/biomedinformatics4010045

Academic Editors: Jörn Lötsch and Burghardt Wittig

Received: 15 January 2024
Revised: 23 February 2024
Accepted: 5 March 2024
Published: 8 March 2024

1. Introduction

Studying diseases at the molecular level, particularly complex cancers like glioblastoma multiforme (GBM), has been pivotal in uncovering their complex causes. Among the various forms of cancer, GBM stands out due to its heterogeneity, with subtypes such as the mesenchymal and proneural states exhibiting distinct molecular profiles and clinical behaviors [1]. The advent of advanced gene sequencing technologies, such as single-cell RNA sequencing (scRNA-seq), has provided critical insights into these conditions, leading to more precise diagnoses and customized treatments. The technique examines gene activity in individual cells, capturing the intrinsic differences within tissue populations, which is crucial in understanding the diverse nature of GBM subtypes. This method allows for a nuanced measurement of gene expression variation across single cells, revealing intricate disease mechanisms specific to GBM and its subtypes. However, analyzing the detailed data provided by scRNA-seq presents its own set of challenges, particularly in the context of the highly variable and complex nature of GBM. Identifying important genes across various cell types can be daunting, due to the presence of errors and artifacts in raw

scRNA-seq data. Therefore, meticulous data cleaning is imperative before any effective use of this information can be made.

In the realm of scRNA-seq data analysis, several critical steps lay the groundwork for meaningful biological discoveries. Dimensionality reduction techniques, which are crucial for managing the substantial data generated by scRNA-seq, play a pivotal role in simplifying the inherent complexity of high-dimensional data while safeguarding the integrity of cell populations. This preservation is vital for accurate cell type identification and subsequent analyses, highlighting the indispensability of dimensionality reduction [2,3]. In addition to traditional dimensionality reduction methods, feature selection is a critical technique in scRNA-seq [4]. This approach focuses on identifying key genes, which process is equally essential as it enables more targeted and efficient analysis. Advanced methods significantly enhance the purity of cell clustering and the accuracy of lineage reconstruction, demonstrating the critical role of feature selection in mitigating noise and improving the precision of scRNA-seq data analysis [5].

While dimensionality reduction and feature selection are fundamental in refining the complexity of scRNA-seq data for more accurate analysis, the application of genetic algorithms (GAs) introduces a complementary, evolutionary-based approach to optimizing this analysis process.

Genetic algorithms (GAs) are adaptive metaheuristic search algorithms classified as evolutionary computing algorithms, which use techniques inspired by natural evolution. They are efficient tools for solving optimization problems. GA is a discrete and non-linear process that is not mathematically guided, wherein optima evolve from one generation to another without mathematical formulation. Integration among GA parameters, including mutation and crossover rates in addition to population, is vital for a successful GA search. GA implementation operates on a binary chromosome representation, where each gene is denoted by a bit in the chromosome. The presence or absence of a gene in the feature set is represented by a 1 or 0, respectively. The GA initiates with a population of randomly generated chromosomes and iteratively evolves this population using genetic operators such as crossover and mutation [6,7].

Moreover, the analysis of scRNA-seq data frequently involves comparing transcript abundance across various conditions or cell types to identify differentially expressed genes (DEGs). These DEGs are pivotal in unveiling the dynamics of gene regulation and the cellular heterogeneity inherent in scRNA-seq data. Serving as more than mere markers of cellular diversity, DEGs are crucial in unraveling the intricate gene expression landscape within individual cells [8–10]. This makes them invaluable for advancing biomedical research and the development of personalized medicine, as they offer deep insights into gene regulation, development, and disease.

Furthermore, the significance of variance in scRNA-seq data cannot be overstated. It plays a pivotal role in identifying DEGs and profoundly impacts the overall analysis. The effectiveness of variance-driven approaches in integrating scRNA-seq data leads to the discovery of new cell types and markers. These findings underscore the importance of capturing hidden variations for robust analysis [11,12]. Such insights highlight the necessity of comprehensively understanding and accounting for variance in scRNA-seq data. High-variability genes (HVGs) play a pivotal role in scRNA-seq analysis, as they exhibit significant variation in expression levels across individual cells within a sample [13]. This variability extends beyond mere technical noise or experimental errors, often reflecting genuine biological differences among cells. The identification of HVGs is crucial, as these genes provide insights into cellular heterogeneity and are instrumental in unraveling the underlying biological processes and cellular states [14]. In the context of scRNA-seq data, understanding gene variance is integral to deciphering cellular heterogeneity. Genes with high variance typically indicate a diversity of biological processes or cell states, thus facilitating the identification of distinct cell types or states in complex samples. Conversely, genes exhibiting low variance generally show uniform expression across cells, which could indicate housekeeping functions or consistent expression regardless of cell type or state [15].

Despite the significance of DEGs and HVGs in scRNA-seq analysis, relying solely on these metrics can sometimes be misleading. High-variability genes (HVGs), for instance, may not always reflect true biological variation but could be influenced by technical noise, such as dropout events, potentially leading to false conclusions about cellular heterogeneity [16]. Similarly, DEGs may not fully capture the complexity of gene regulation dynamics, as they might overlook subtle but biologically relevant changes in gene expression [17]. Therefore, a balanced approach that considers both DEGs and HVGs, along with additional validation methods, is crucial for accurate scRNA-seq data interpretation.

To effectively navigate the complexities of scRNA-seq data, particularly for intricate cancers like glioblastoma multiforme (GBM), the GeneSelector framework has been developed. It addresses the heterogeneity seen in GBM subtypes such as the mesenchymal and proneural categories. The framework begins with comprehensive preprocessing, enhancing data quality by removing noise and artifacts. Central to GeneSelector is a sophisticated genetic algorithm (GA) that not only emphasizes gene variance but also incorporates differentially expressed genes (DEGs). Through a well-constructed fitness function, this approach allows for identifying genes that are both biologically significant and highly variable, thus aiding in the discovery of potential GBM biomarkers. This paper delves into the multifaceted methodology of GeneSelector, showcasing its effectiveness in providing a thorough and biologically pertinent analysis of scRNA-seq data in the context of GBM's cellular complexity.

2. Methodology

In the GeneSelector framework, scRNA-seq data preprocessing plays an integral part, beginning with the Seurat package for quality control and normalization. Initial steps involve filtering cells based on gene count thresholds to remove potential technical noise and non-informative signals [18]. The pipeline then systematically filters the genes to retain those genes expressed in a sufficient number of cells, enhancing signal clarity. Seurat's normalization method is applied, in which the feature counts for each cell are first normalized by dividing them by the total counts for that cell and then scaled using the scale factor. Following this process, these values undergo a natural logarithm transformation using the log1p function [19].

The genetic algorithm (GA) feature selection phase is enhanced with a sophisticated selection operator using the "GA" package in R [20]. This operator selects the top 50% of individuals from the population, based on performance, ensuring the retention of high-quality genetic combinations. Additionally, the best individual from each generation is always carried over to the new population, making the algorithm elitist. This approach, combined with a custom initialization giving each gene a 1% chance of selection, ensures a balance between diversity and the preservation of superior solution gene sets.

The fitness function is a pivotal component of our genetic algorithm (GA), which has been meticulously designed to assess genes for their expression variability, statistical significance, and changes in expression. It plays a critical role in highlighting the algorithm's capacity to discern genes characterized by both considerable variance and biological importance [21]. By leveraging a tour-selection approach alongside carefully calibrated crossover and mutation rates, our GA is strategically optimized to enhance genetic diversity and facilitate thorough exploration. This multifaceted evaluation mechanism ensures the identification of genes that are not only significantly differentially expressed but also demonstrate substantial variance and relevance to biological processes, thereby reinforcing the algorithm's effectiveness in uncovering biologically significant gene markers.

In each generation of the analysis, key metrics such as variance, p-value, and log fold change (logFC) are evaluated to determine gene significance. The algorithm employs a weighted scoring system for these metrics: assigning 50% to variance 30% to logFC, and 20% to the p-value score. Once these scores are computed, they undergo normalization to ensure uniformity and comparability across the dataset. Subsequently, an average of these normalized, weighted scores is calculated. This systematic approach aids in identifying

genes that are not only statistically significant but also demonstrate significant biological variance and expression changes, ensuring a comprehensive assessment of gene relevance.

The DEGs are identified using the Wilcoxon rank sum test, a non-parametric approach that allows us to detect statistically significant differences in gene expression between subtypes. This method is complemented by variance analysis to pinpoint HVGs, focusing on genes that exhibit significant expression variability across samples, which is indicative of their potential regulatory roles in glioblastoma.

Building upon this evaluation, the genetic algorithm then applies the principles of natural selection and evolution through its crossover and mutation processes. The process of crossover entails the blending of genetic information from two parent individuals to produce a descendant, whereby certain portions of the parental genetic sequences are chosen at random for exchange. This influences the genetic composition of the offspring. Similarly, the mutation process affects every new individual by introducing random changes to their genetic makeup. By randomly flipping elements in their genetic code, variations in the traits that are passed on for future classification purposes are created. Each individual has a certain probability of undergoing these alterations, which introduces diversity and adaptability into the population, fostering the evolution of more effective solutions over successive generations.

To provide comprehensive insights into our algorithmic methodology and ensure reproducibility, we have made our full codebase, datasets, and a detailed table of the 92 gene markers identified available on GitHub. Interested parties can access these resources at https://github.com/PaplomatasP/GeneSelector, accessed on 2 February 2024. This repository encompasses the genetic algorithm scripts, data preprocessing and analysis procedures, and detailed findings concerning the 92 gene markers. By sharing these resources, we aim to facilitate the replication of our study, validation of our results, and further investigation into the identified genetic markers by our peers and the scientific community at large.

3. Results

This study's results highlight the effectiveness of the GeneSelector framework in identifying critical biomarkers in glioblastoma multiforme (GBM) through a novel genetic algorithm (GA)-based feature selection process. The integration of differentially expressed genes (DEGs) and highly variable genes (HVGs) within this framework facilitates the discovery of genes that are both biologically significant and exhibit high variance, a necessary step in understanding GBM's complexity.

Initially, the clear separation between the two GBM subtypes (see Figure 1) is apparent, implying that in-depth gene-centric analysis holds the promise of uncovering substantial insights regarding their interrelation. Employing the well-established dimensionality reduction method of UMAP, we condensed the gene feature space into a 2D representation, providing a clearer perspective on our samples. An initial investigation into the selection of genes revealed a pivotal observation: the actual overlap between DEGs and HVGs is limited (Figure 2). This discrepancy highlights the inherent differences in gene selection with each method and suggests that a considerable number of relevant genes might be overlooked if only one selection method is employed. By utilizing both DEGs and HVGs within a genetic algorithm framework, the possibility arises of harnessing a more complete representation of the genomic landscape. This dual-method approach increases the chances of identifying genes that are both biologically significant and exhibit high variability, which may be crucial in understanding complex disease mechanisms, such as those found in GBM. Moreover, the integration of DEGs and HVGs within a GA provides a strategic balance; while DEGs offer insights into differential expression under various conditions, HVGs reveal intrinsic expression variability that can pinpoint cellular heterogeneity. The combination within a GA thus enriches the feature selection process, yielding a robust set of candidate biomarkers and providing a comprehensive genetic profile that is essential for advancing precision medicine.

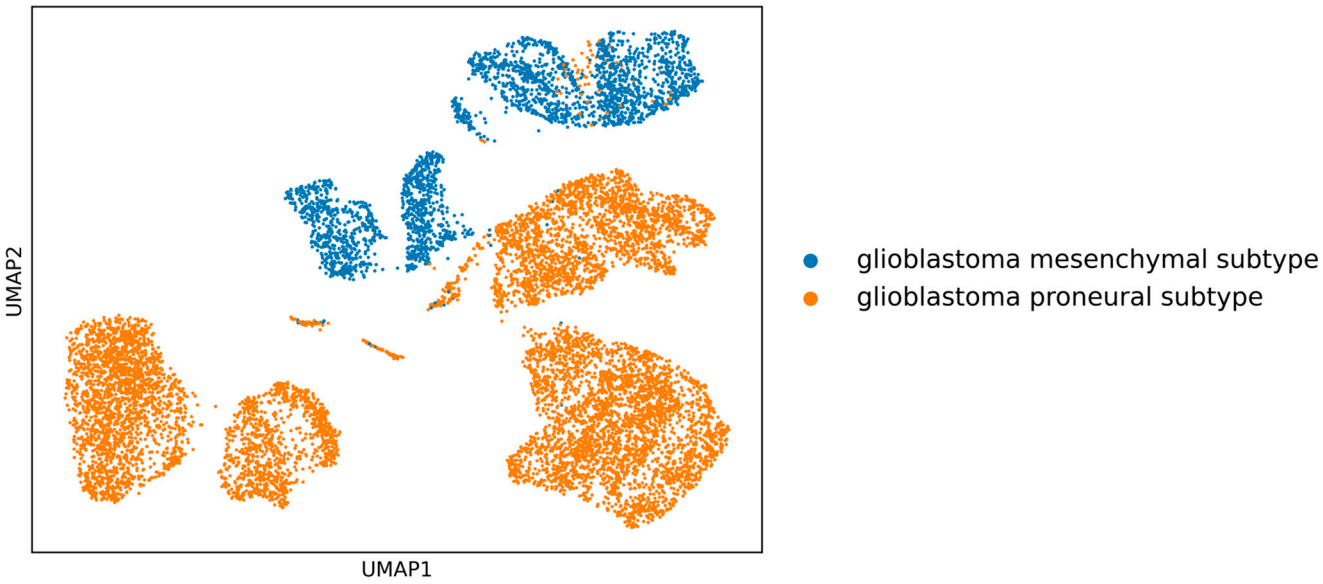

Figure 1. The UMAP 2D plot illustrates the spatial distribution of cells within a glioblastoma multiforme (GBM) sample. Each dot on the plot represents an individual cell, and their positions in the two-dimensional space have been determined using uniform manifold approximation and projection (UMAP) for dimensionality reduction. The cells are color-coded based on their GBM subtypes, with mesenchymal subtypes shown in vibrant orange and proneural subtypes shown in deep blue. The distinct separation between the two GBM subtypes is evident, suggesting that a thorough gene-based analysis has the potential to reveal significant insights into their relationship.

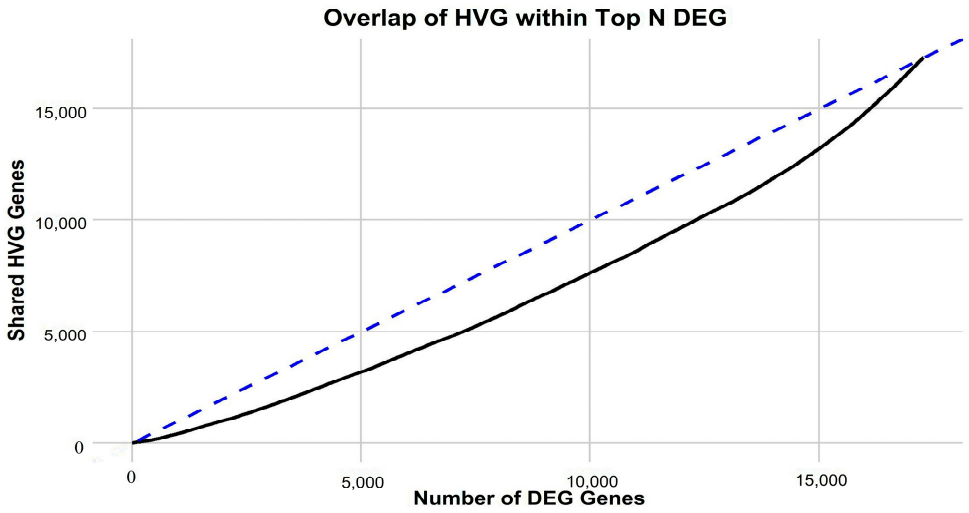

Figure 2. This graph depicts the relationship between the number of DEGs and the corresponding overlap with HVGs. The solid black line represents the actual shared genes between the two categories, while the dashed blue line indicates the hypothetical maximum overlap if all DEGs were also HVGs. The widening gap between these lines underscores the distinct nature of gene selection by DEGs and HVGs, justifying the integration of both sets into the GA for a comprehensive biomarker discovery process.

Motivated by the limited convergence between differentially expressed genes (DEGs) and highly variable genes (HVGs) (Figure 2), a genetic algorithm (GA) was utilized to amalgamate these distinct gene sets effectively. The trajectory of the GA's performance, illustrated in Figure 3, reveals a consistent improvement in gene selection quality within the population over successive generations. The mean fitness score climbs steadily, indicating a concentrated effort by the GA to identify genes that have significant biological relevance.

This is complemented by the observation that the best individual fitness score in each generation surpasses the mean, evidencing the algorithm's evolutionary advantage in identifying and retaining the most promising gene candidates. As the GA progresses, it reaches a point of stability, a steady state wherein further improvements in fitness scores begin to plateau. This stabilization signals that the algorithm is nearing an optimal gene set, thereby highlighting the GA's adeptness at narrowing down the most biologically pertinent features for disease understanding and biomarker discovery.

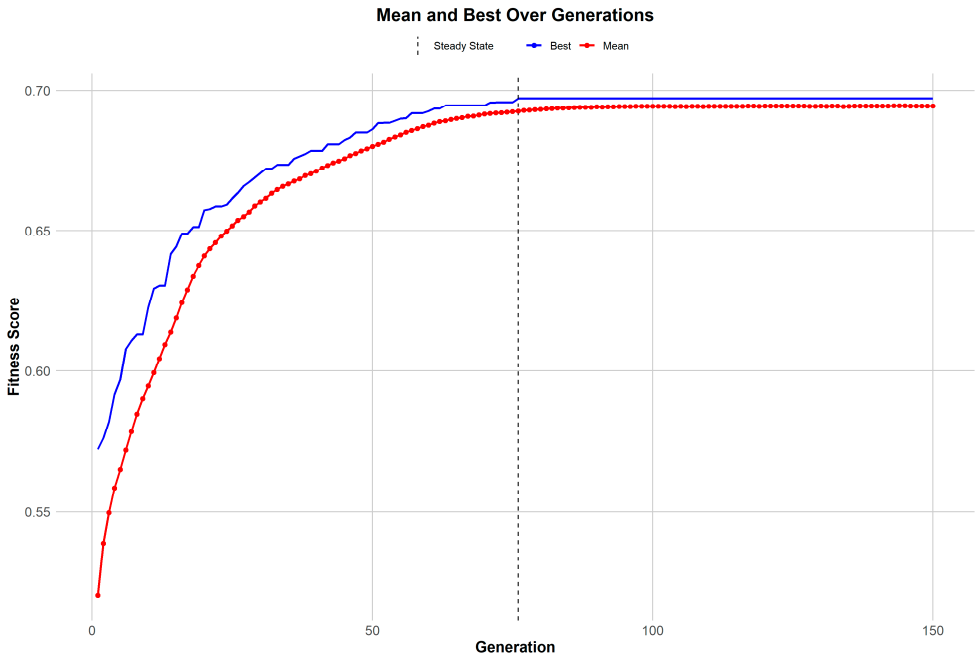

Figure 3. This plot illustrates the optimization process of the genetic algorithm (GA) in selecting gene features across successive generations. The red line indicates the mean fitness score of the population at each generation, revealing how the average quality of gene selection improves over time. The blue line represents the best fitness score achieved by any individual in the population, which provides insight into the peak performance of the GA at each stage. The vertical dashed line marks the "steady state", the point beyond which further iterations yield diminishing improvements, suggesting the convergence of the GA towards an optimal set of genes.

In the exploration of the genetic landscape pertaining to glioblastoma multiforme (GBM) subtypes, particularly the mesenchymal and proneural categories, a panel of 92 genes was rigorously examined. The genes were selected based on their functional roles within cancer biology, their expression profiles within GBM subtypes, and any documented correlations with patient prognoses or responses to treatment. The selection process was informed by reliance on the established literature due to the absence of specific gene expression-level data.

The gene ATM, known for its critical role in DNA repair and cell cycle control, was identified as a key player, with its mutations and signaling pathways being recurrent themes across various cancer types, including GBM [22]. Similarly, the gene HDAC6, a member of the histone deacetylase family, was noted for its significance in chromatin remodeling and gene expression regulation, with ongoing research investigating its potential as a therapeutic target in GBM [23]. Furthermore, the gene ANGPTL2 was highlighted for its involvement in angiogenesis, a process quintessential to cancer progression that is specifically noted for its potential impact on the vascularization of GBM tumors.

Additionally, STAG2 [24] was recognized for its integral role within the cohesin complex and its contribution to chromosomal stability, a pivotal aspect in the pathogenesis of cancer. The gene LAMB2, which encodes a component of the extracellular matrix, was noted for its influence on cellular differentiation, migration, and survival—factors

pertinent to the pathology of GBM. The importance of these genes was emphasized by their functional attributes within the realms of cancer biology. However, it was acknowledged that a more precise determination of genes related to the mesenchymal and proneural GBM subtypes would necessitate a comprehensive analysis of gene expression data, along with a review of the current scientific literature specific to GBM subtypes. Such an approach would provide a contextually richer understanding, taking into account the presence of these genes in GBM stem cells and their roles within the tumor microenvironment.

For the mesenchymal subtype, those genes implicated in mesenchymal differentiation, angiogenesis, inflammatory responses, and extracellular matrix remodeling were given particular attention. Although the provided gene list did not include specific data on differential expression or functional studies, two genes, ANGPTL2 and HDAC6, were repeatedly referenced due to their known biological functions and potential relevance to the mesenchymal transition in GBM [25]. In contrast, the proneural GBM subtype, which is characterized by those genes involved in neural development and oligodendrocyte lineage transcriptional programs, presented a different challenge. Without specific expression data, the association of genes with this subtype was less clear. Nonetheless, genes like OLIG2 and PDGFRA, which were not listed but are known to be associated with the proneural subtype, were discussed for their roles in neural progenitor cell development and maintenance.

In summary, the discussion elucidated the complexity of selecting and analyzing genes in relation to GBM subtypes. It underscored the need for a multifaceted approach that combines bioinformatics analyses with experimental validation and a literature review. This would enhance the identification of genes most pertinent to the mesenchymal and proneural subtypes of GBM, thereby advancing the understanding of their respective biological underpinnings and contributing to the development of subtype-specific therapeutic strategies.

The final 92 genes that resulted from the genetic optimization-based pipeline were examined in an enrichment analysis based on gene ontologies (Figure 4). The analysis was conducted using WebGestal [26], a popular tool for the interpretation of gene lists derived from large-scale -omics studies. It was found that the gene ontologies associated with these genes displayed a variety related to glioblastoma. The outcomes of the enrichment analysis demonstrate a variety in those gene ontologies related to the complex biological processes associated with glioblastoma multiforme (GBM) subtypes. Indicatively, the ontology term "glycerophospholipid metabolic process" is related to the metabolic pathways that are crucial for maintaining cellular membrane integrity and the signaling, processes that are often altered in cancer cells, including those in GBM.

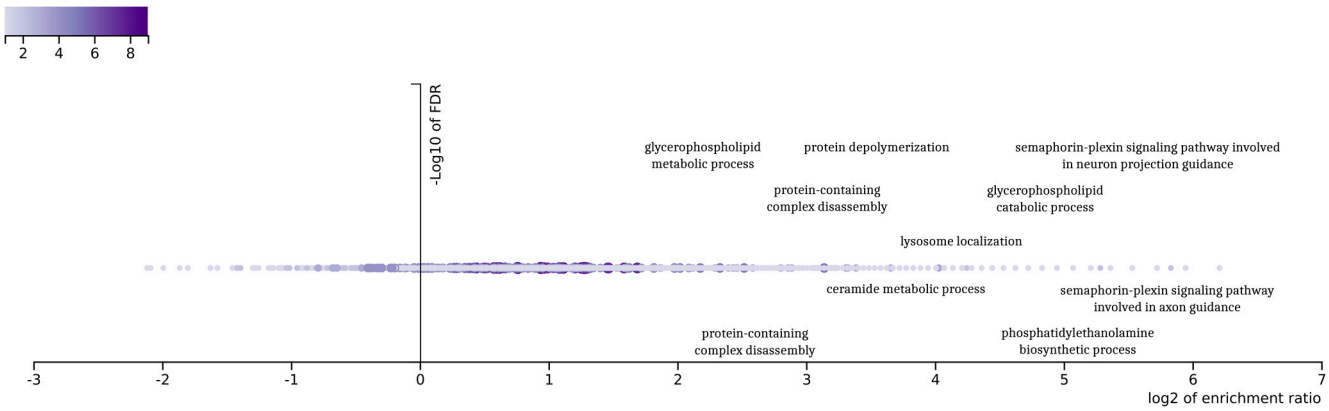

Figure 4. This figure illustrates the outcomes of a comprehensive enrichment analysis using gene ontologies (GOs) for the set of 92 genes obtained through a genetic optimization-based pipeline.

Furthermore, the term "semaphorin-plexin signaling pathway" is particularly noteworthy due to its involvement in neural development, which aligns with the characteristics of the proneural subtype of GBM. This pathway is known to play a significant role in neural

patterning and development, making it a point of interest for further investigation in the context of GBM. Additionally, "protein-containing complex disassembly" and "protein depolymerization" are GO terms that may highlight important aspects of protein regulation and degradation in GBM. These processes could be pivotal in understanding the aberrant protein dynamics within GBM cells.

The GO term "lysosome localization" also emerges as relevant, potentially indicating alterations in the cellular trafficking and degradation pathways within the GBM tumor microenvironment [27]. This aspect could provide insights into how GBM cells manage cellular waste and recycling, which are critical for their survival and proliferation. Given the objective of achieving a more comprehensive cellular analysis, the analytical phase of this research focused on isolating the most significant differentially expressed genes (DEGs) that differentiate between the proneural and mesenchymal states of glioblastoma multiforme (GBM). Their expression profiles across all cell samples were then examined (refer to Figure 5). It was observed that these genes exhibited a higher level of expression in cells derived from patients with the mesenchymal subtype of GBM, while a lower level of expression was noted in cells from the proneural subtype.

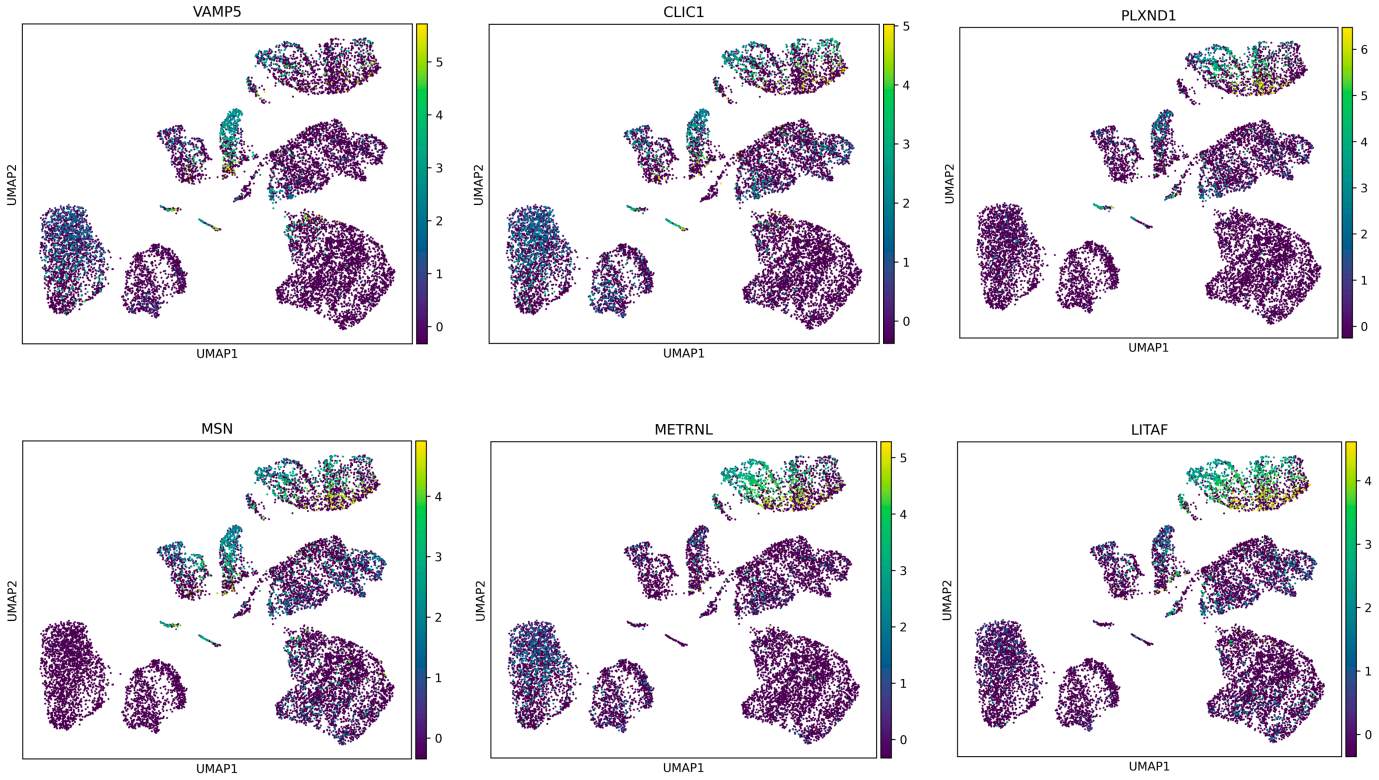

Figure 5. Uniform manifold approximation and projection (UMAP) visualization depicting individual cells, each represented as a distinct dot. The cells are colored based on their expression profiles, which have been normalized, logarithmized, and scaled, focusing on the six most significantly differentially expressed genes distinguishing the proneural from mesenchymal subtypes in glioblastoma. This color-coded representation offers a detailed comparative insight into the gene expression variations between these two glioblastoma subtypes.

This in-depth analysis at the cellular level highlighted the potential of these DEGs to interpret the differences between the two GBM subtypes. The findings provided insights into the molecular mechanisms underlying the distinct pathological features of the proneural and mesenchymal GBM subtypes. The approach enabled a detailed understanding of the gene expression variations that contribute to the unique characteristics of each subtype, thereby offering a foundation for future research aimed at targeted therapeutic strategies.

In our study, a set of key genes, including VAMP5, CLIC1, PLXND1, MSN, METRNL, and LITAF, were identified as having higher mRNA expression levels in mesenchymal cells compared to the proneural cells of glioblastoma multiforme (GBM). This differential expression suggests a potential overproduction of the respective proteins in the mesenchymal subtype.

Among these, VAMP5 (vesicle-associated membrane protein 5) and CLIC1 (chloride intracellular channel 1) are particularly noteworthy. VAMP5 plays a role in vesicular transport and membrane fusion, processes that are crucial for cellular trafficking and signaling. Its elevated expression in mesenchymal cells may contribute to the altered cellular dynamics characteristic of this GBM subtype [28]. Meanwhile, CLIC1 is involved in chloride ion transport and cell cycle control, and its upregulation could be linked to the enhanced proliferative capacity and invasiveness observed in mesenchymal GBM cells. PLXND1 (Plexin D1), another gene from the list, is integral to the semaphorin signaling pathway, which is implicated in angiogenesis and cellular migration. Its increased expression in mesenchymal cells aligns with the subtype's aggressive nature, marked by enhanced angiogenesis and invasiveness.

MSN (moesin) and METRNL (meteorin-like) are also of interest. MSN is part of the ERM (ezrin, radixin, moesin) protein family, playing a role in cytoskeletal rearrangement and cellular morphology, potentially influencing cell motility and invasion in mesenchymal GBM. METRNL, which is implicated in immunometabolic responses, could contribute to the unique inflammatory microenvironment of the mesenchymal subtype [29]. Finally, LITAF (lipopolysaccharide-induced tumor necrosis factor-alpha factor), with its role in inflammatory response regulation, may reflect the mesenchymal subtype's pro-inflammatory traits.

Furthermore, the mesenchymal gene signature in glioblastoma multiforme (GBM) is associated with aggressive tumor behavior and poor patient outcomes. Elevated expression of VAMP5, CLIC1, PLXND1, MSN, METRNL, and LITAF genes has been observed within this subtype, indicating their potential involvement in driving disease progression. These genes are implicated in various cellular processes, such as vesicle trafficking (VAMP5), ion transport (CLIC1), cell signaling (PLXND1), cytoskeletal organization (MSN), and immune regulation (METRNL and LITAF). Their collective impact likely contributes to the invasive and treatment-resistant nature of mesenchymal GBM. Understanding the specific roles of these genes within the mesenchymal subtype could offer insights into novel therapeutic strategies aimed at mitigating disease aggressiveness and improving patient outcomes.

The elevated mRNA expression of these genes in mesenchymal cells compared to the proneural cells of GBM provides valuable insight into the molecular distinctions driving the pathophysiology of these subtypes. This understanding could pave the way for targeted therapeutic strategies tailored to the unique characteristics of each GBM subtype.

In our study, the "Pathways in Cancer" KEGG pathway map (see Figure 6) plays a pivotal role in elucidating the molecular mechanisms underlying the transition from proneural to mesenchymal glioblastoma (GBM) cells. The above map highlights those genes with the most significant differential expression (DEGs) by enclosing them within red boxes, providing a visual representation of their importance within the broader context of cancer-related pathways. Notably, these DEGs are portrayed as central players in cancer biology, suggesting their potential roles as key regulators in GBM progression. Furthermore, we observe 23 genes displaying considerable activity. This observation underscores the substantial involvement of the "Pathways in Cancer" KEGG pathway in the transition between GBM cell subtypes, reaffirming the relevance of our experimental results in shedding light on the underlying molecular processes that drive GBM pathogenesis.

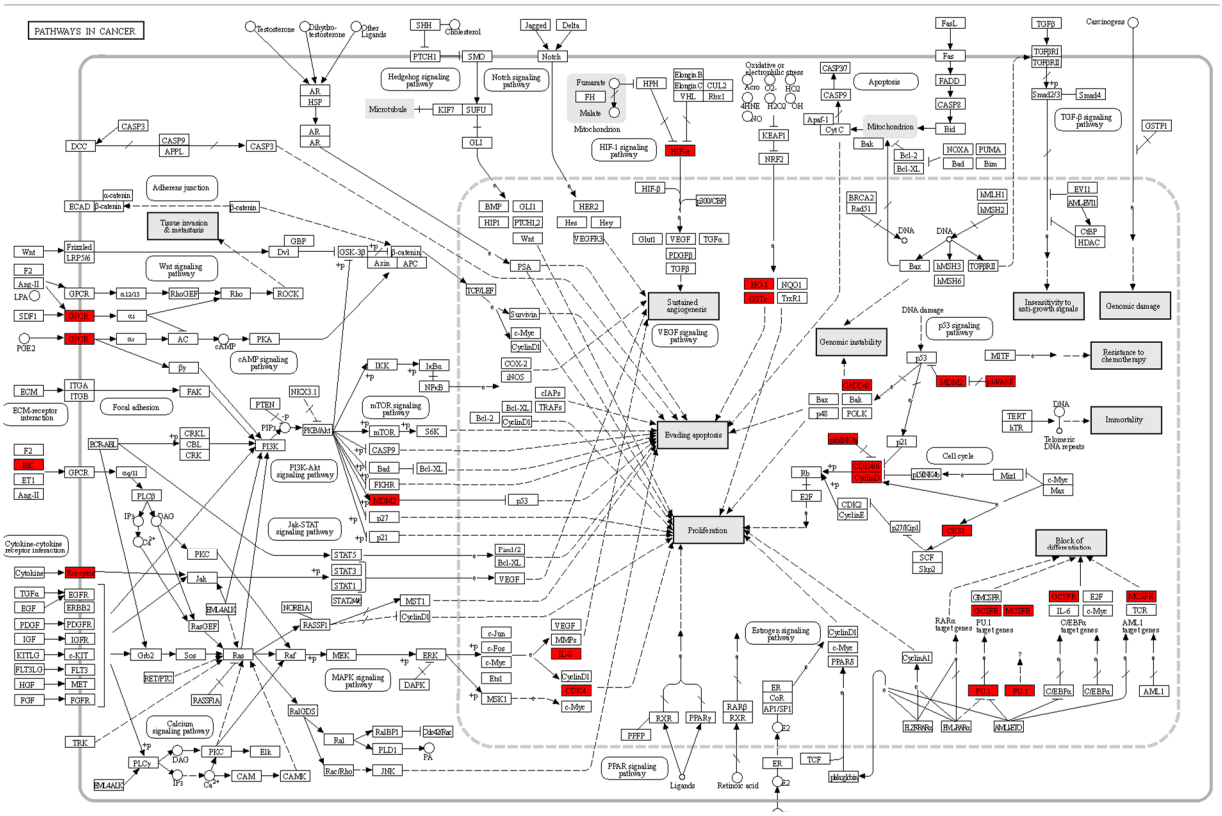

Figure 6. The "Pathways in Cancer" KEGG pathway map, highlighting key DEG genes. Arrows within the map typically represent the direction of biochemical reactions or biological processes. The genes with the most significant differential expression when transitioning from proneural to mesenchymal GBM cells are enclosed within red boxes, illustrating their positions and roles within the broader context of cancer-related pathways. Observations reveal the activity of 23 genes within the pathway, indicating a considerable level of pathway activity, as per the results of the experimental study.

4. Conclusions

This study marks a significant advancement in the field of glioblastoma multiforme (GBM) research, employing a novel genetic optimization algorithm to identify key gene biomarkers for proneural and mesenchymal GBM subtypes. By effectively integrating differential gene expression and gene variability, this approach not only identifies genes specific to GBM subtypes but also ensures their biological relevance. The successful application of this methodology to single-cell RNA-sequencing data underscores its potential, not only in enhancing our understanding of GBM but also in contributing to the broader realm of biomarker discovery for various complex diseases. This work opens new avenues for targeted therapy development in GBM and potentially in other cancers, demonstrating the power of innovative genomic techniques in precision medicine.

Author Contributions: In the preparation of this research article, the contributions of each author were delineated as follows. P.P. and A.V. were instrumental in conceptualizing the study. The methodology was developed by A.V. and P.P. was responsible for the software development, playing a key role in the project. Validation was jointly conducted by P.P., I.-E.D. and A.V., ensuring the accuracy of the research findings. P.P. took the lead in the formal analysis of data, while also spearheading the investigation process. P.V. provided essential resources for the study. The crucial task of data curation was handled by P.P. The original draft of the manuscript was prepared by P.P., with P.V. and A.V. contributing significantly to the review and editing process. I.-E.D. was responsible for the visualization aspects of the research. The entire project was supervised by A.V., who also managed the project administration. Funding acquisition was undertaken by P.V., adhering to the principles of

the CRediT taxonomy. Authorship was confined to those who have substantially contributed to the work reported, ensuring a fair and accurate representation of individual contributions. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: This study has received formal ethical approval from the Ionian University's Research Ethics and Deontology Committee under protocol number 3600.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Verhaak, R.G.W.; Hoadley, K.A.; Purdom, E.; Wang, V.; Wilkerson, M.D.; Miller, C.R.; Ding, L.; Golub, T.; Jill, P.; Alexe, G.; et al. Integrated Genomic Analysis Identifies Clinically Relevant Subtypes of Glioblastoma Characterized by Abnormalities in PDGFRA, IDH1, EGFR, and NF1. *Cancer Cell* **2010**, *17*, 98–110. [CrossRef]
2. Xiang, R.; Wang, W.; Yang, L.; Wang, S.; Xu, C.; Chen, X. A Comparison for Dimensionality Reduction Methods of Single-Cell RNA-seq Data. *Front. Genet.* **2021**, *12*, 646936. [CrossRef]
3. Sun, S.; Zhu, J.; Ma, Y.; Zhou, X. Accuracy, robustness and scalability of dimensionality reduction methods for single-cell RNA-seq analysis. *Genome Biol.* **2019**, *20*, 269. [CrossRef]
4. Paplomatas, P.; Krokidis, M.G.; Vlamos, P.; Vrahatis, A.G. An Ensemble Feature Selection Approach for Analysis and Modeling of Transcriptome Data in Alzheimer's Disease. *Appl. Sci.* **2023**, *13*, 2353. [CrossRef]
5. Chen, B.; Herring, A.C.; Lau, K.S. pyNVR: Investigating factors affecting feature selection from scRNA-seq data for lineage reconstruction. *Bioinformatics* **2018**, *35*, 2335–2337. [CrossRef] [PubMed]
6. Feng, J.; Niu, X.; Zhang, J.; Wang, J.H. Gene selection and classification of scRNA-seq data combining information gain ratio and genetic algorithm with dynamic crossover. *Wirel. Commun. Mob. Comput.* **2022**, *2022*, 9639304. [CrossRef]
7. Katoch, S.; Chauhan, S.S.; Kumar, V. A review on genetic algorithm: Past, present, and future. *Multimed. Tools Appl.* **2021**, *80*, 8091–8126. [CrossRef] [PubMed]
8. Zou, J.; Deng, F.; Wang, M.; Zhang, Z.; Liu, Z.; Zhang, X.; Hua, R.; Chen, K.; Zou, X.; Hao, J. scCODE: An R package for data-specific differentially expressed gene detection on single-cell RNA-sequencing data. *Briefings Bioinform.* **2022**, *23*, bbac180. [CrossRef]
9. Sekula, M.; Gaskins, J.; Datta, S. Detection of Differentially Expressed Genes in Discrete Single-Cell RNA Sequencing Data Using a Hurdle Model With Correlated Random Effects. *Biometrics* **2019**, *75*, 1051–1062. [CrossRef] [PubMed]
10. Anders, S.; McCarthy, D.J.; Chen, Y.; Okoniewski, M.; Smyth, G.K.; Huber, W.; Robinson, M.D. Count-based differential expression analysis of RNA sequencing data using R and Bioconductor. *Nat. Protoc.* **2013**, *8*, 1765–1786. [CrossRef]
11. Zhang, H.; Lee, C.A.A.; Li, Z.; Garbe, J.R.; Eide, C.R.; Petegrosso, R.; Kuang, R.; Tolar, J. A multitask clustering approach for single-cell RNA-seq analysis in Recessive Dystrophic Epidermolysis Bullosa. *PLoS Comput. Biol.* **2018**, *14*, e1006053. [CrossRef]
12. Yip, S.H.; Sham, P.C.; Wang, J. Evaluation of tools for highly variable gene discovery from single-cell RNA-seq data. *Briefings Bioinform.* **2019**, *20*, 1583–1589. [CrossRef]
13. Sun, C.; Wang, L.; Wang, H.; Huang, T.; Yao, W.; Li, J.; Zhang, X. Single-cell RNA-seq highlights heterogeneity in human primary Wharton's jelly mesenchymal stem/stromal cells cultured in vitro. *Stem Cell Res. Ther.* **2020**, *11*, 149. [CrossRef]
14. Lee, D.; Cheng, A.; Ucar, D. A robust statistical framework to detect multiple sources of hidden variation in single-cell transcriptomes. *bioRxiv* **2017**, 151217. [CrossRef]
15. Stuart, T.; Satija, R. Integrative single-cell analysis. *Nat. Rev. Genet.* **2019**, *20*, 257–272. [CrossRef]
16. Liu, J.; Zeng, W.; Kan, S.; Li, M.; Zheng, R. CAKE: A flexible self-supervised framework for enhancing cell visualization, clustering and rare cell identification. *Briefings Bioinform.* **2024**, *25*, bbad475. [CrossRef]
17. Le Priol, C.; Azencott, C.-A.; Gidrol, X. Detection of genes with differential expression dispersion unravels the role of autophagy in cancer progression. *PLoS Comput. Biol.* **2023**, *19*, e1010342. [CrossRef]
18. Chen, G.; Ren, M.; Lv, C.; Shi, T. Low Quality Cells Should Be Removed from Single-Cell RNA-Seq Data Analysis. Available online: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3307902 (accessed on 30 December 2018). [CrossRef]
19. Hafemeister, C.; Satija, R. Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. *Genome Biol.* **2019**, *20*, 296. [CrossRef] [PubMed]
20. Scrucca, L. On some extensions to GA package: Hybrid optimisation, parallelisation and islands evolution. *R J.* **2017**, *9*, 187–206. [CrossRef]
21. Chatzilygeroudis, K.I.; Vrahatis, A.G.; Tasoulis, S.K.; Vrahatis, M.N. Feature Selection in single-cell RNA-seq data via a Genetic Algorithm. In Proceedings of the Learning and Intelligent Optimization: 15th International Conference (LION 15), Athens, Greece, 20–25 June 2021; Revised Selected Papers 15. Springer: Berlin/Heidelberg, Germany, 2021; pp. 66–79.

22. Hashimoto, T.; Urushihara, Y.; Murata, Y.; Fujishima, Y.; Hosoi, Y. AMPK increases expression of ATM through transcriptional factor Sp1 and induces radioresistance under severe hypoxia in glioblastoma cell lines. *Biochem. Biophys. Res. Commun.* **2022**, *590*, 82–88. [CrossRef] [PubMed]
23. Auzmendi-Iriarte, J.; Saenz-Antoñanzas, A.; Mikelez-Alonso, I.; Carrasco-Garcia, E.; Tellaetxe-Abete, M.; Lawrie, C.H.; Sampron, N.; Cortajarena, A.L.; Matheu, A. Characterization of a new small-molecule inhibitor of HDAC6 in glioblastoma. *Cell Death Dis.* **2020**, *11*, 417. [CrossRef]
24. Waldman, T.; Kim, J.S.; Xu, W.; Yang, T.; Ya, A.; Tallon, L.; Jin, F. *STAG2 Mutations Regulate 3D Genome Organization, Chromatin Loops, and Polycomb Signaling in Glioblastoma Multiforme*; Research Square: Durham, NC, USA, 2023.
25. Zhang, H.; Huang, Y.; Yang, E.; Gao, X.; Zou, P.; Sun, J.; Tian, Z.; Bao, M.; Liao, D.; Ge, J.; et al. Identification of a Fibroblast-Related Prognostic Model in Glioma Based on Bioinformatics Methods. *Biomolecules* **2022**, *12*, 1598. [CrossRef]
26. Liao, Y.; Wang, J.; Jaehnig, E.J.; Shi, Z.; Zhang, B. WebGestalt 2019: Gene Set Analysis Toolkit with Revamped UIs and APIs. *Nucleic Acids Res.* **2019**, *47*, W199–W205. [CrossRef] [PubMed]
27. Jacobs, A.K.; Maghe, C.; Gavard, J. Lysosomes in glioblastoma: Pump up the volume. *Cell Cycle* **2020**, *19*, 2094–2104. [CrossRef] [PubMed]
28. Yin, L.; Xu, Y.; Yin, J.; Cheng, H.; Xiao, W.; Wu, Y.; Ji, D.; Gao, S. Construction and validation of a risk model based on the key SNARE proteins to predict the prognosis and immune microenvironment of gliomas. *Front. Mol. Neurosci.* **2023**, *16*, 1304224. [CrossRef] [PubMed]
29. Luksik, A.S.; Yazigi, E.; Shah, P.; Jackson, C.M. CAR T Cell Therapy in Glioblastoma: Overcoming Challenges Related to Antigen Expression. *Cancers* **2023**, *15*, 1414. [CrossRef]