A systematic analysis to perform robust Glioblastoma subtypes deconvolution from single cell data towards extracting prognostic biomarkers for clinical associations

Behnoush Rostami, Christa Ventresca, Mahnoor Gondal

Abstract:

Glioblastoma is an incurable and poorly understood tumor, with diverse genetic, epigenetic, and developmental factors contributing to its progression. However, their precise characterization remains a challenge. The molecular characteristics of glioblastoma are unique and the cancer can be divided into four subtypes. The disambiguation of these subtypes is difficult with only bulk data. An approach using single cell technology allows us to better understand cellular heterogeneity within an individual at the single cell level. However, as single cell data can be expensive to generate and usually contains fewer samples, it is necessary to bridge the gap between single cell and bulk data to classify glioblastoma patients more accurately when analyzing bulk data.

In this study, we develop a unified model of glioblastoma cell states and genetic diversity using single-cell RNA-sequencing as well as bulk genetic expression analysis from The Cancer Genome Atlas (TCGA). As glioblastoma has four subtypes, this study categorizes the malignant cells within the four subtypes and makes use of TCGA data on these subtypes towards identifying putative biomarkers for potential treatment. Taken together, we utilized single cell data to find features that can be mapped back to bulk data for glioblastoma robust classification for clinical evaluations.

Keywords: *Glioblastoma, single-cell RNA-sequencing, glioblastoma subtypes*

Introduction:

Heterogeneity of tumors is one of the biggest challenges in cancer diagnosis and treatment. Generally, it can manifest as varying outcomes or therapeutic responses among tumors, depending on their stage, genetic lesions, or expression profiles.[1] Furthermore, cells from the same tumor may possess distinct mutations or epigenetic states.[2] The idea that intratumoral heterogeneity is correlated with treatment failure and disease recurrence is gaining more widespread recognition.[3]

One of the deadliest human malignancies, glioblastoma, is an archetypal example of a heterogeneous cancer.[4] In most cases, conventional and targeted therapies are not able to induce long-term remissions due to intratumoral heterogeneity and redundant signaling routes.[5] Cellular niches in these tumors exhibit distinct phenotypic characteristics, including resistance to radiation-induced DNA damage,[6] transient quiescence and self-renewal,[7] and adaptation to hypoxia.[8] Glioblastomas can be classified genetically and transcriptionally based on DNA and RNA profiles of bulk tumors.[9] However, studying bulk tumors does not provide a clear understanding of the relationship between genetic, transcriptional, and functional sources of intratumoral heterogeneity.

Glioblastoma is the most common primary brain malignancy in adults and one of the most aggressive cancers. There is just a 4.6-month median survival rate for the general patient population. Patients with optimal treatment have a median survival of 14 months and 2-year survival rate of 26%.[4] A combination of surgery and radiotherapy has been the primary treatment since the 1970s. Recent targeted chemotherapy approaches, such as the alkylating agent temozolomide,[4] have also been used, though with modest success. Poor drug delivery in the brain and the inability to debulk extensive tumors contribute significantly to the lack of effective treatment options.

In glioblastoma, heterogeneity is one of the main reasons for therapeutic failure. Genetic, epigenetic, and microenvironmental factors influence the cellular program and cause heterogeneity. Transcriptional subtypes show a variety of layers of heterogeneity. At least three subtypes of glioblastoma have been identified based on bulk expression profiles, namely proneural (TCGA-PN), classical (TCGA-CL), and mesenchymal (TCGA-MES). Therefore, certain gene mutations are more prevalent in certain expression-based subtypes than in others. For example, PDGFRA alterations are more frequent in TCGA-PN glioblastoma, while EGFR alterations are more prevalent in TCGA-CL. Multiregion tumor sampling also shows that multiple subtypes can coexist in different regions of the same tumor. It is shown that subtypes can change over time and with treatment, and single-cell RNA-sequencing (scRNA-seq) suggests that distinct cells within a tumor carry over programs from distinct subtypes.[11, 12, 13]

In light of this, single-cell technology holds a lot of promise for identifying distinct glioblastoma subtypes within a single sample. Single-cell analysis involves dissociating individual cells from each other, and processing each of them individually. This allows researchers to identify the specific subtype for each nuclei.[2] However, it is much more expensive than bulk analysis, meaning that not every clinic will be able to afford the analysis.

In this study, we use single-cell RNA-seq glioblastoma data to employ genetic markers of the different cell types within a tumor, as well as whether they are malignant or benign. We then use these markers and apply them to bulk RNA-seq glioblastoma data, to deconvolve the different subtypes within the bulk data. We can then use those features to identify putative biomarkers to associate with clinical survival outcomes in glioblastoma patients. In our analysis we were able to develop a classification method to robustly identify glioblastoma subtypes using only 90 transcription factors. Among the top features for the glioblastoma classification we identified *CDX2*[14] and *RUNX1*[15] as known biomarkers for glioblastoma. Our survival analysis demonstrates that these two transcription factors are enough to predict patient survival in case the subtype is not known. We also observed the importance of incorporating patients sex into the model towards reaching better clinical associations.[16] This satisfies our main goal which is to make it possible for clinics to keep costs low by performing bulk RNA-sequencing of samples, but still extract valuable information about the tumor and its heterogeneity that can be used during treatment.

Methods:

Single cell data extraction

Single-cell data was downloaded from Tumor Immune Single-cell Hub 2 (TISCH2). Single cell data was obtained from a previous study on glioblastoma (Glioma_GSE131928_10x).[17] The data is from 28 patients (20 adult and 8 pediatric) and has been pre-batch corrected and generally pre-processed.

Bulk data extraction

The transcriptomic bulk data used in this analysis was downloaded from The Cancer Genome Atlas (TCGA) using gdc client (GDC Data Transfer Tool Client), whereas the clinical data for the analysis was downloaded using R package called *TCGAbiolinks*.

Deconvolution methods

To make robust estimation of subtypes we employed *granulator*, an R package with integrated state of the art methods for cell type deconvolution, including ordinary least squares, non-negative least squares,[18] quadratic programming with non-negativity and sum-to-one constraint,[19] quadratic programming without constraints, re-weighted least squares,[20] support vector regression,[21] and linear mixing model.[22] We then assigned the subtype to a patient in the bulk using the maximum proposition the sample had of each subtype. To note, we did not include qprogwc allocations because the method failed to produce useful results without data. The six methods have unique strengths and weaknesses, therefore we only assigned a subtype confidently if it was consistent for more than 3 methods. This ensured robust allocation of subtypes.

Clustering methods

Single-cell RNA-seq

Seurat was used to cluster the data based on expression of marker genes to identify cell types.[23] Seurat uses a weighted nearest neighbors model for clustering where relationships between cells are weighted by the number of neighbors that they share. Feature extraction was then performed to extract gene expression data for the different cell types within the data by using the commands “FindAllMarkers” and “AverageExpression”. FindAllMarkers extracts the fold change data for genes within different clusters, while AverageExpression finds the average gene expression for each gene in the different clusters. In our analysis, we used all genes, the top 50 genes (based on absolute value of fold change), and the bottom 50 genes to compare results. We decided that “all genes” was giving the best result and is recommended in various deconvolution benchmarking papers,[24] therefore, we based most of our analysis on “all genes” markers list.

Bulk RNA-seq

To cluster the bulk dataset into subtypes, we have investigated five different machine learning (ML) clustering algorithms and compared their performance using metrics such as Adjusted Rand index (ARI), Jaccard score, Normalized Mutual Information (NMI) score, Fowlkes-Mallows (FM) score, and also their computation time. The clustering methods that we used at this step include K-means clustering (k=3, k=4), Hierarchical clustering, Gaussian Mixture model (GMM), DBSCAN, and spectral clustering.

Network construction using WGCNA

To construct networks to cluster gene expressions we used *Py.WGCNA* which is a correlation network analysis library in python.[25] This analysis starts with clustering highly correlated genes into clusters called modules. Clustering is accomplished by calculating the correlation between pairs of genes to create a co-expression matrix. Next, based on soft power thresholding the correlation matrix a co-expression network is produced. Finally, using hierarchical clustering of the network and dynamic tree cut processing, this analysis identified co-expressed modules of genes.

Feature selection

We employed the *Dorothea* package in R to help extract 1335 transcription factors list.[26] To select important transcription factors for classifying subtypes in glioblastoma we used the *Boruta*

algorithm for feature selection.[27] The results from the feature selection were then evaluated using random forest method. We compared results generated selecting all features compared to only confirmed features from Boruta for the random forest. The final AUC (Area under the ROC Curve) values were then compared. The varImpPlot function was used to extract the top 30 important features.

BayesPrism for biomarker identification

BayesPrism pipeline was employed to calculate the cell type specific expression of gene per sample.[28] The expression values were incorporated into BayesPrism for the 30 important features extracted from the feature selection step for each cell type. We observed *CBX2* and *RUNX1* were present in the selected features list and we confirmed from literature that these were potential biomarkers for glioblastoma.

Survival analysis

For survival analysis we employed the *survival* and the *survminer* packages in R. Kaplien Meier curves were generated from multiple scenarios to arrive at the best possible model, including (i) only subtype, (ii) only *RUNX1* status, (iii) only *CBX2* status, (iv) *CBX2* and *RUNX1* status, (v) *CBX2* and *RUNX1* status and biological sex, and finally (vi) *CBX2* and *RUNX1* and subtype and biological sex. The *CBX1* and *RUNX1* status as positive or negative was estimated using their mean expression values from the bulk data as threshold.

Results:

Identifying cell type specific markers from single-cell data

The initial analysis revealed nine different cell types present in the single-cell sample. From all of these cells, 30% were identified as immune cell types, 66% were malignant cells, and 3% were other cell types. The cell types identified included AC-like malignant cells, CD8Tex, M1, malignant cells, MES-like malignant cells, monocyte, NPC-like malignant cells, oligodendrocytes, and OPC-like malignant cells. The results of this clustering are shown in Figure 1, Table 1, Table 2, and Figure 2.

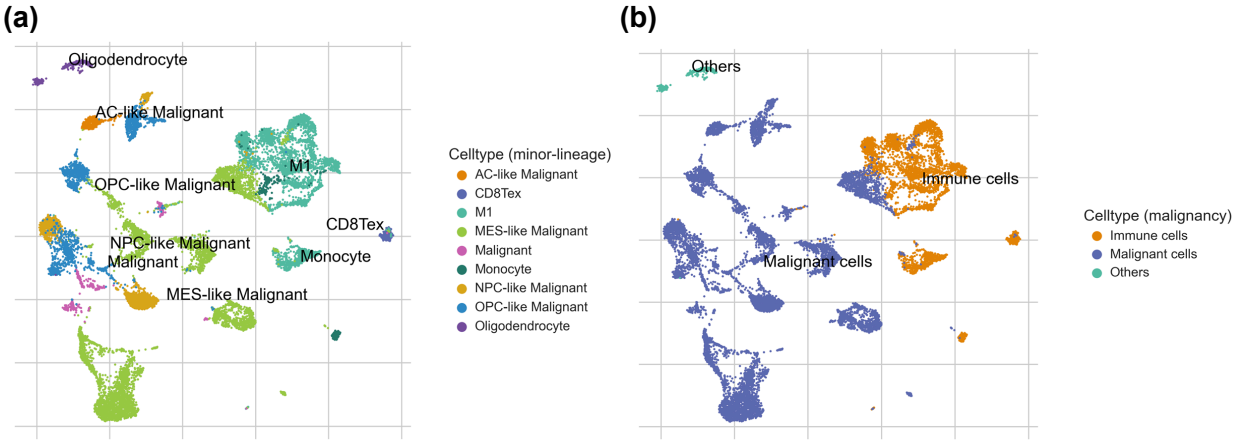

Figure 1. UMAP of cell types in the single cell data.[17] (a)Clusters colored by minor lineage of cell type, (b) Clusters colored by malignancy.

Table 1. Number and percentage of malignant and immune cells within the single-cell samples.

	Immune Cells	Malignant Cells	Other
Number of Cells	4166	8998	394

Percentage	0.30727	0.66366	0.02906

Table 2. Number and percentage of different cell types within the single cell samples.

	AC-like Malignant	CD8Tex	M1	Malignant	MES-like Malignant	Monocyte	NPC-like Malignant	Oligo-dendrocyte	OPC-like Malignant
Number of Cells	262	197	3679	367	5203	290	1368	394	1798
Percentage	0.01932438	0.01453017	0.27135271	0.02706889	0.38375867	0.02138959	0.10089984	0.02906033	0.13261543

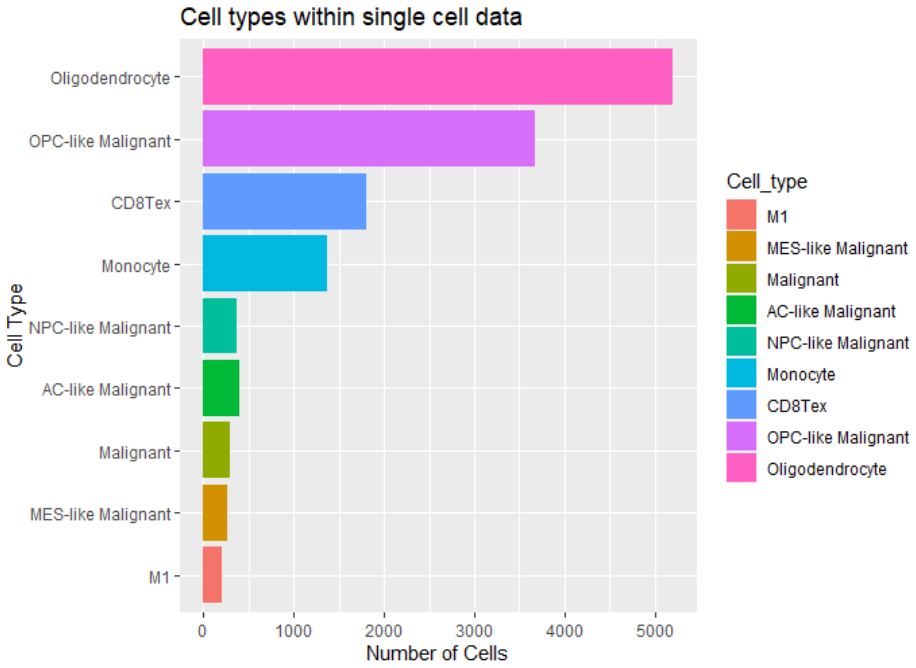

Figure 2. Bar graph of cell types within single cell data.

Evaluating seven deconvolution methods to label bulk data

To reach a robust identification of the subtypes in the bulk data we employed seven different deconvolution methods. Each method reached a similar conclusion for the MES (mesenchymal) subtype, however, there were discrepancies between the methods for the other subtypes. We therefore assigned each subtype only if at least 3 methods reached the same conclusion. This enabled us to make a robust assignment of each subtype to each patient. The resultant allocations had 111 MES subtype (mesenchymal-like), 9 AC (astrocyte-like), 43 OPC (oligodendrocyte-progenitor-like), and 2 NPC (neural progenitor-like) subtype (Figure 3).

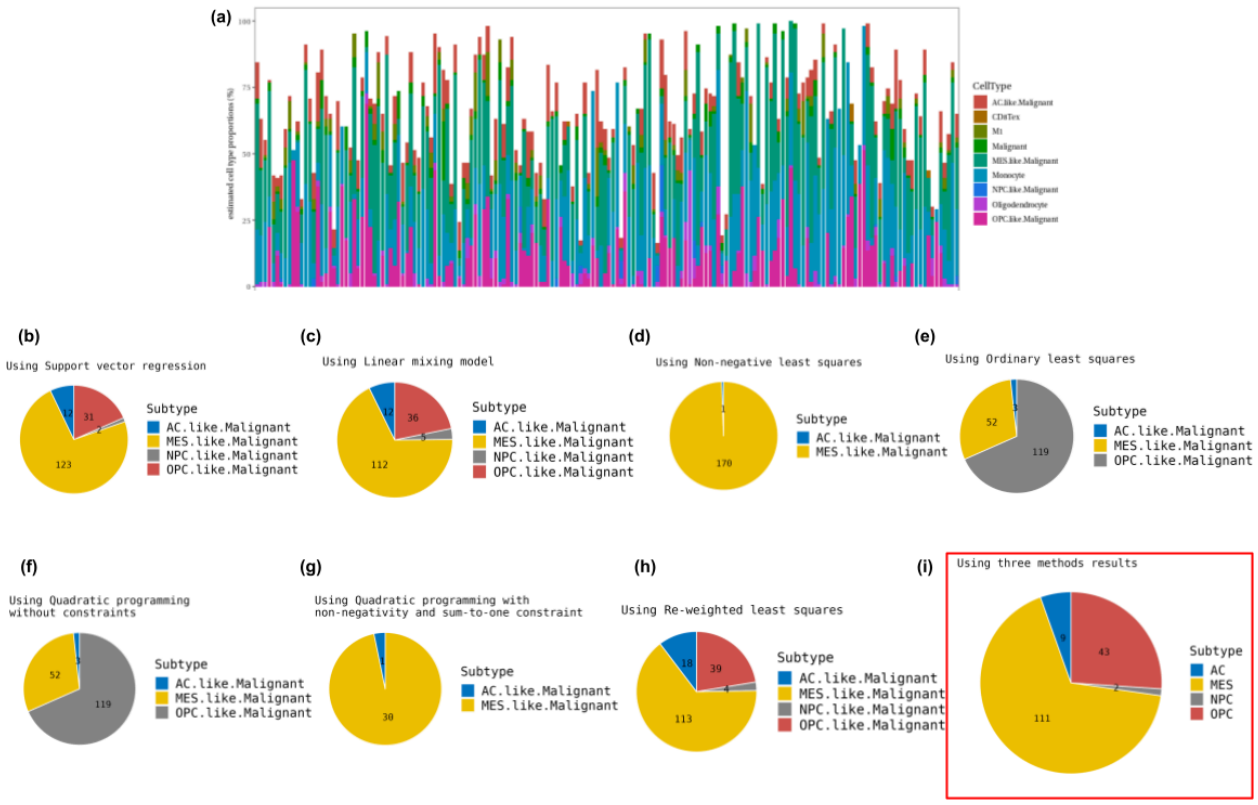

Figure 3. Deconvolution of bulk data. (a) Bar graph showing composition of cell types within the bulk RNA-seq samples using support vector regression (svr) method. Color indicates cell type; individuals are shown on the x-axis. (b-h) - Pie charts showing the composition of the four malignant cell types based on different deconvolution methods. Panel i represents using three methods simultaneously.

Using robust clustering and network construction methods to validate bulk sample classification

In this research we used five different unsupervised algorithms to cluster the bulk dataset into different subtypes. The first method is Hierarchical clustering with 'Ward' method and 'Euclidean' metric. Using this algorithm, the data is clustered into three main clusters with some outlier clustering as shown in Figure 4. Hierarchical clustering has some benefits such as not requiring a predetermined number of clusters. It also produces a Hierarchical dendrogram showing the clusters. Nevertheless, in order to determine the results, some parameters must be predetermined, such as the method and metric of the linkage function.

The second method that we investigated for clustering the bulk data is K-means clustering. To use this method, we first need to determine the optimum number of clusters. The Elbow plot is one of the most popular methods to determine this optimal value of k in a K-means algorithm. To obtain the Elbow-plot, we calculate the values of distortion and inertia for each k in the range 2 to 6.

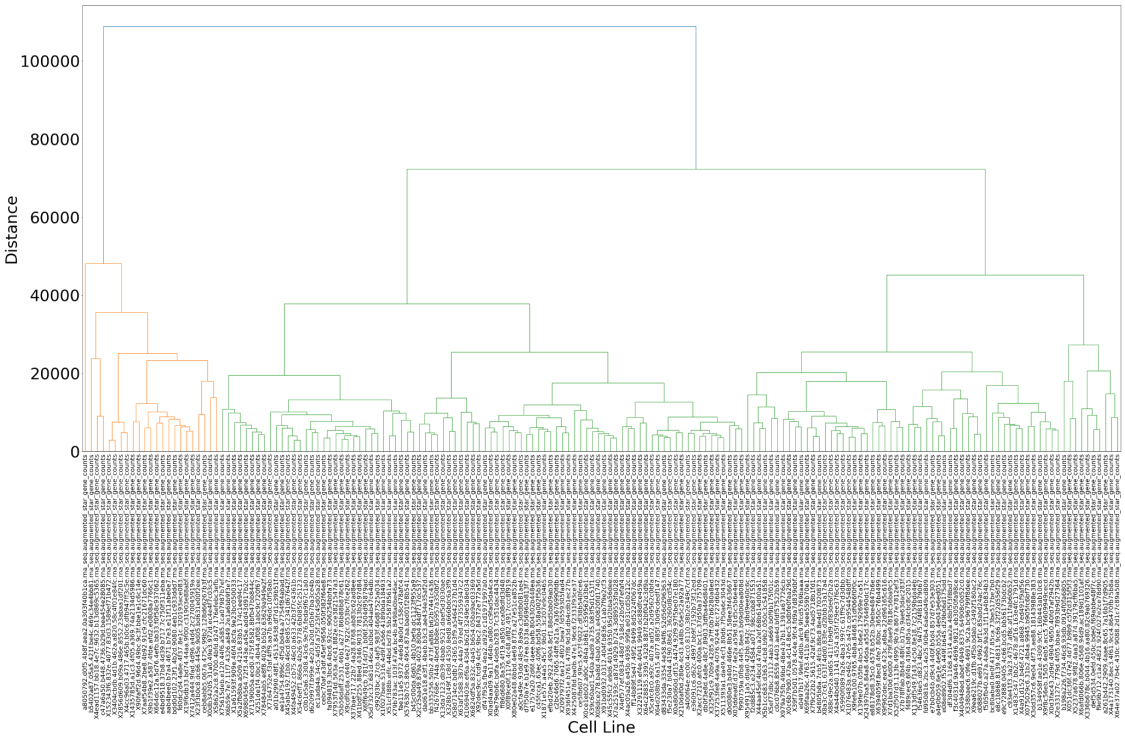

Figure 4. Cluster Dendrogram produced by Hierarchical clustering. The Dendrogram shows 3 main clusters with some outlier clustering.

The distortion value is calculated as the average of the squared distances (typically Euclidean) from the cluster centers of the respective clusters. In addition, inertia is calculated by adding together the squared distances between samples and their closest cluster centers. According to Figure 5 (a), the optimal number of clusters for the K-means algorithm would be k=3. We also performed Silhouette analysis to provide a visual representation of how close each cluster point is to its neighboring clusters for k in range of 2 to 5 as shown in Figure 5 (b). The thickness of the Silhouette plot provides information about the size of clusters and clusters with below average Silhouette scores indicate that the number of clusters is not optimal. Based on that, as shown in Figure 5 (b), k=4 is an optimal value for the K-means clustering algorithm. Based on these two analyses, we decided to do the K-means clustering with both 3 and 4 clusters and compare their performance as shown in Figure 5 (e, f). The third method that we used for clustering in this study is a Gaussian Mixture Model (GMM) in which data points are assigned to clusters based on their probabilities and a Gaussian distribution is applied to each cluster. Unlike K-means, a GMM can cluster overlapping clusters. The clusters using the GMM algorithm is shown in Figure 5 (g). Since K-Means, Hierarchical, and GMM clustering all fail in creating clusters of arbitrary shapes, we decided to investigate Density-Based Spatial Clustering of Applications with Noise (DBSCAN). According to DBSCAN, a point belongs to a cluster if it is close to many other points from that cluster. The clusters using DBSCAN are shown in Figure 5 (h). The last algorithm that was studied in this research is spectral clustering which is a technique based on graph theory and identifies groups of nodes based on the edges connecting them in a graph (Figure 5 (i)).

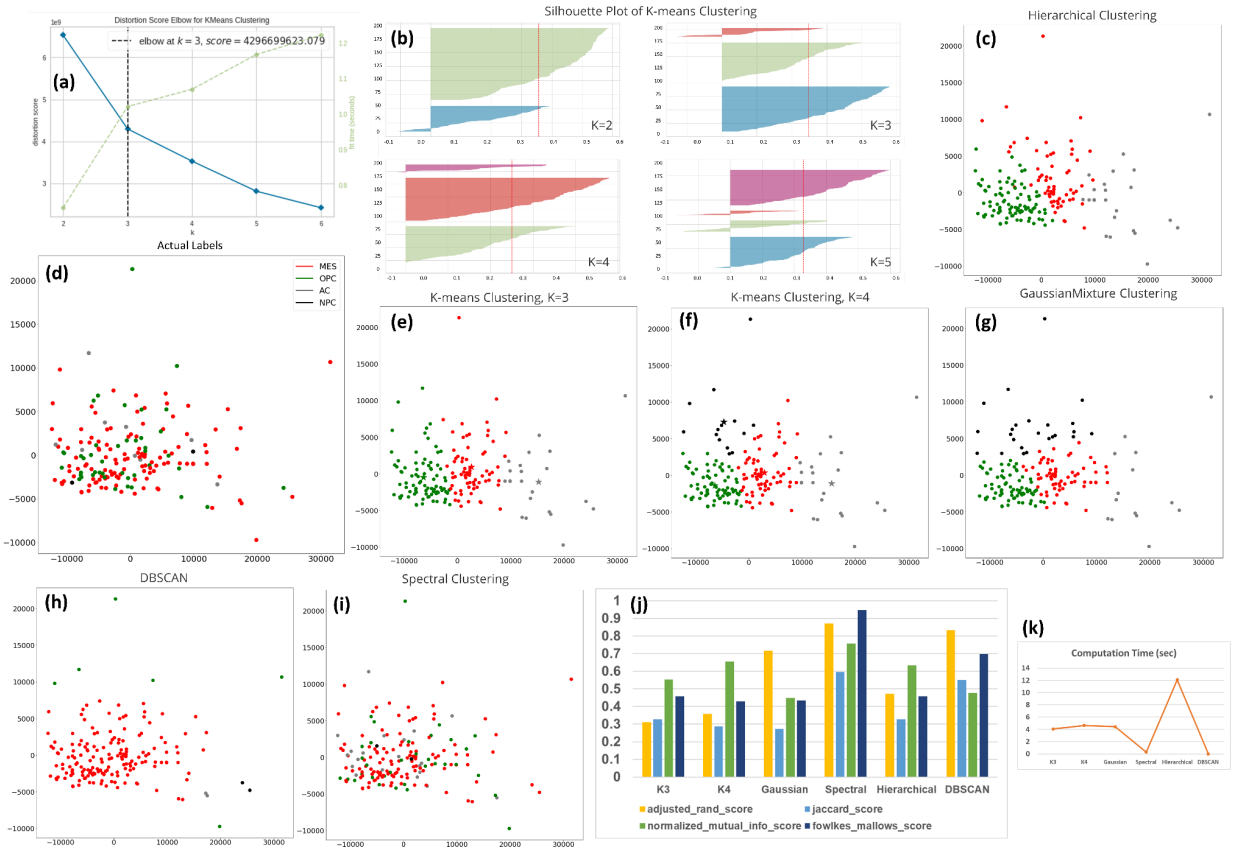

Figure 5. Five different clustering methods used for clustering the bulk dataset, (a) the Elbow-plot to find the optimum k for K-means algorithm. (b) Silhouette analysis to provide a visual representation to choose an optimal value for the number of clusters for K-means clustering. (c) Hierarchical clustering, (d) actual labels of 174 samples, (e) K-means clustering k=3, (f) K-means clustering k=4, (g) Gaussian Mixture Model clustering, (h) DBSCAN, (i) Spectral clustering, (j) performance comparison graph, (k) computation time comparison graph.

To compare the performance of the studied ML algorithms we used ARI, Jaccard score, NMI score, FM score as presented in Figure 5 (j). ARI, Jaccard, NMI, and FM can be defined as follows:

$$ARI(U,V) = \frac{2(N_{00}N_{11} - N_{01}N_{10})}{(N_{00}+N_{01})(N_{01}+N_{11}) + (N_{00}+N_{10})(N_{10}+N_{11})}$$

$$Jaccard(U,V) = \frac{N_{11}}{N_{11}+N_{10}+N_{01}}$$

$$NMI(U,V) = \frac{I(U;V)}{H(U)+H(V)}$$

$$FM(U,V) = \sqrt{\left(\frac{N_{11}}{N_{11}+N_{01}}\right)\left(\frac{N_{11}}{N_{11}+N_{10}}\right)}$$

Actual Labels: $U = \{U_1, U_2, \cdots, U_R\}$

Predicted Labels: $V = \{V_1, V_2, \cdots, V_c\}$

N_{11}: the number of pairs that are in the same partition in both U and V;

N_{00}: the number of pairs that are in different partitions in both U and V;

N_{01}: the number of pairs that are in the same partition in U but in different partitions in V;

N_{10}: the number of pairs that are in different partitions in U but in the same partition in V.

$I(U;V)$ is the mutual information between U and V

$H(.)$ is the entropy of partitions, in which H(U) and H(V) are calculated

Based on this comparison, spectral clustering provides the best performance in clustering the bulk dataset into the subtypes with ARI=0.87, Jaccard=0.6, NMI=0.75, and FM=0.95. Moreover, the computation time for these methods is calculated as shown in Figure 5 (k). In terms of

computational time, DBSCAN and spectral clustering take the least amount of time, while Hierarchical clustering takes the most.

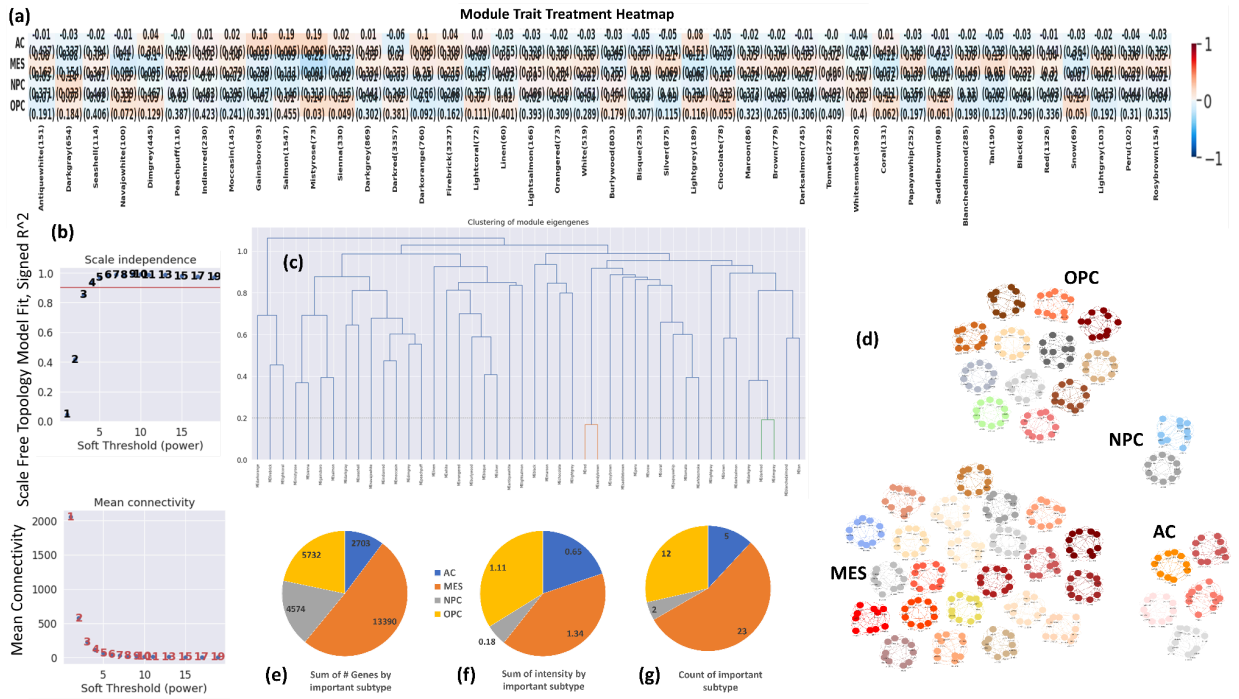

Figure 6. *Py.WGCNA* analysis, (a) module-trait relationship heatmap for different subtypes and gene modules, (b) soft-thresholding power network topologies, (c) the clustering of the gene modules after removing outliers, (d) co-expression networks of genes for 42 modules, labeled by various colors were constructed and clustered based on their important subtypes, (e) the number of genes in each subtype, (f) the sum of correlation intensity of modules with the same important subtype, (g) the number of modules for each important subtype.

As mentioned in the methods section, to construct networks to cluster gene expressions, we used a python version of *WGCNA* which is a correlation network analysis library. By implementing this software package, we could perform network construction, module detection, and clustering networks based on their important subtypes. The networks were constructed, and 42 modules were identified. In order to define the adjacency matrix, soft threshold power of 5 was used based on the approximate scale-free topology criterion. Figure 6 shows the results of performing *Py.WGCNA* analysis. Figure 6 (a) shows the calculated module-trait relationship heatmap for different subtypes and 42 modules. Same figure, part (b) shows various soft-thresholding power network topologies calculated to find modules. The numbers in the plots refer to the soft thresholding powers. It indicates that with a soft-thresholding power of 5, the approximate scale-free topology can be achieved. After removing outliers and low expressed genes, the clustering of the gene modules is produced as shown in part (c). Gene co-expression networks of 42 modules were constructed and clustered according to their important subtypes, as shown in part (d). The number of genes in each subtype, the sum of correlation intensity of modules with the same important subtype, and the number of modules related to each subtype are respectively plotted in Figure 6 (e-g). In total, 23 modules show MES as their important and high correlation density subtype, whereas this number is 12, 5, and 2 for OPC, AC, and NPC.

Using the correlation between module eigengenes and all subtypes, modules related to the subtypes are identified as shown in Figure 7. Each module's corresponding network is also shown next to it.

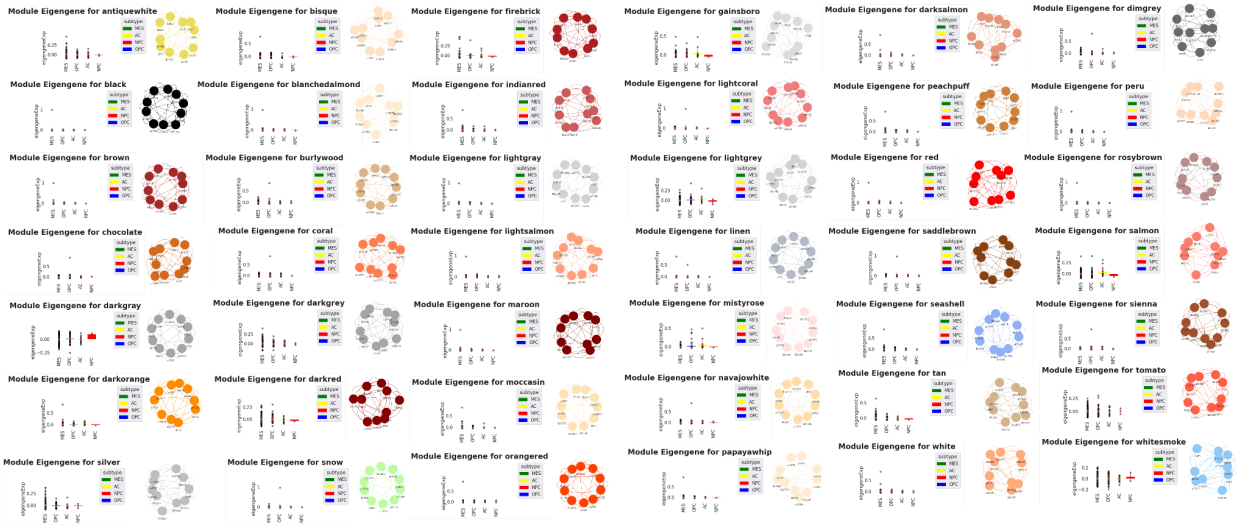

Figure 7. Module eigenes for 42 modules and their corresponding networks.

Extracting important features for putative biomarker discovery

To extract relevant features per subtype we employed a feature selection method called *Boruta* which helped us reach 99 confirmed features for classifying each subtype. We used a random forest classification method to assess the selected features. The results showed that with all features (i.e., 1335 transcription factors) included into the model the resultant AUC value was 0.7727; when we used non-rejected features list of 199 features, the AUC increased to 0.7879; and when we only used the confirmed features list of 99 features the AUC came out to be 0.7879 again. The results therefore conclude that for our classifier 99 transcription factors are enough to predict the glioblastoma subtype when subtype information is not available.

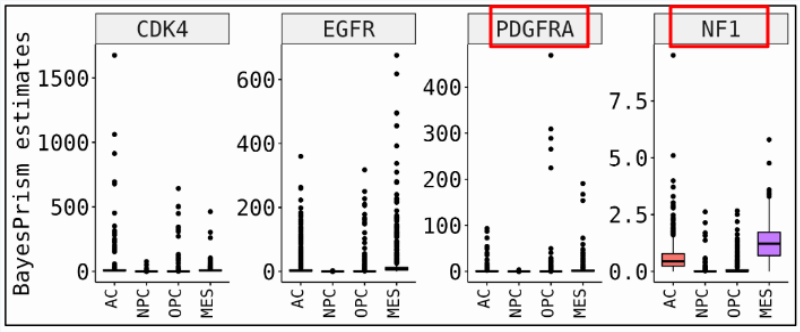

Figure 8. Assessing the expression of each subtype biomarkers in BayesPrism results. Two of the four biomarkers seem to have the correct expression per subtype labels.

We next wanted to see what the expression is of the features per cell type. For that, we used the *BayesPrism* pipeline. *BayesPrism* estimates the contribution from each gene to each cell type through an intricate deconvolution and expression estimation methodology. The method was able to assign 2 of out 4 known biomarkers for each subtype correctly which gave us some confidence in the applicability of the method (Figure 8). We then extracted the top 30 features from the previous feature selection method and observed their expression per cell type from

BayesPrism results (Figure 9). The expression and feature selected show *CDX2*[14] and *RUNX1*[15] as promising prognostic candidates for treating glioblastoma because of their potential evaluations gathered through existing literature.

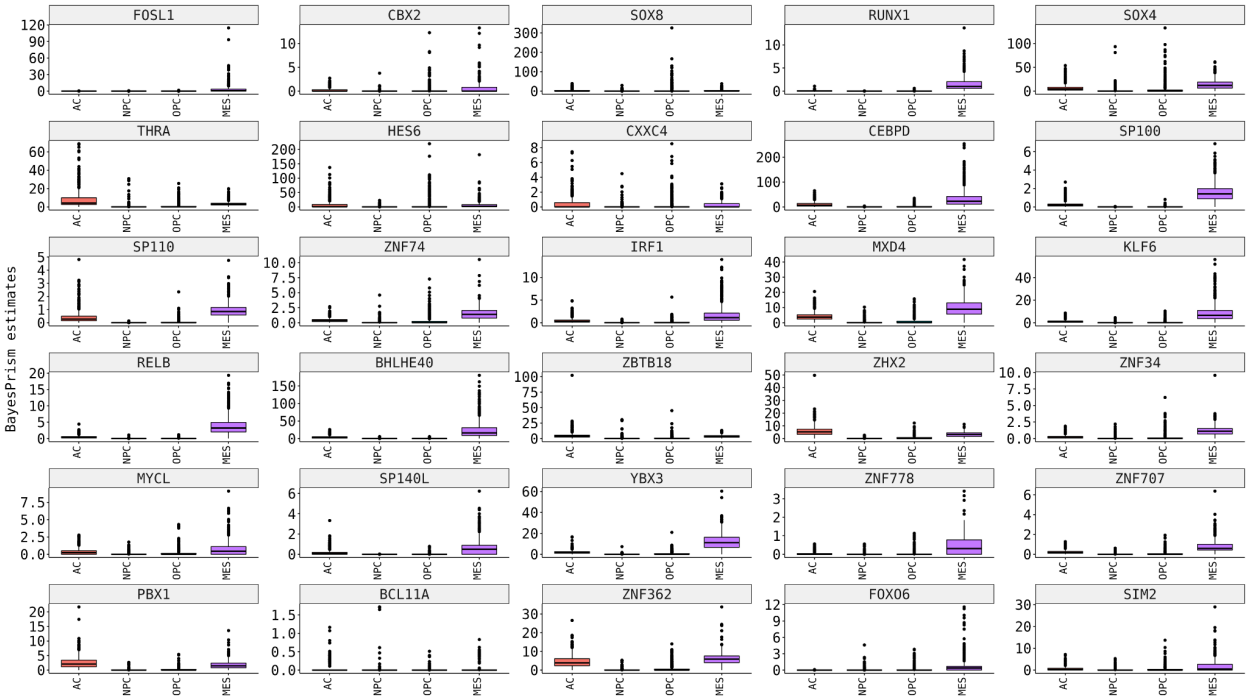

Figure 9. Top 30 features expression from the feature selection process expression using *BayesPrism* estimates for each subtype label.

Employing biomarkers and subtypes to perform survival analysis

In order to make confident clinical associations based on subtypes annotated in the bulk data we used six different scenarios i.e., (i) only subtype, (ii) only *RUNX1* status, (iii) only *CBX2* status, (iv) *CBX2* and *RUNX1* status, (v) *CBX2* and *RUNX1* status and biological sex, and finally (vi) *CBX2* and *RUNX1* and subtype and biological sex. The survival analysis performed with only subtype information was able to generate a p-value of 0.0062 alluding that subtype information differences can lead to different clinical outcomes in the patient (Figure 10). This also promotes the importance of subtype identifications. We also generated Kalien-Miere curves for only *CBX2* and *RUNX1*, which showed that the two genes separately were not able to affect the outcome significantly (p-value of 0.097 and 0.19, respectively). However, evaluating both biomarkers together allowed the p value to become significant (p value = 0.04). We evaluated different parameters in the survival regression model, and we observed biological sex to be a good predictor and adding biological sex to the *CBX2* and *RUNX1* model made the model significant with a p-value of 0.024. We also evaluated the presence of subtype with *CBX2*, *RUNX1* and biological sex and the model generated a p-value of 0.0061 which is significant. This highlights that biological sex is a strong predictor in the clinical outcome of glioblastoma alongside the two biomarkers *CBX2* and *RUNX1*. This was very revealing because it in line with published studies which have noticed a similar association of biological sex with survival in glioblastoma patients specifically.[16]

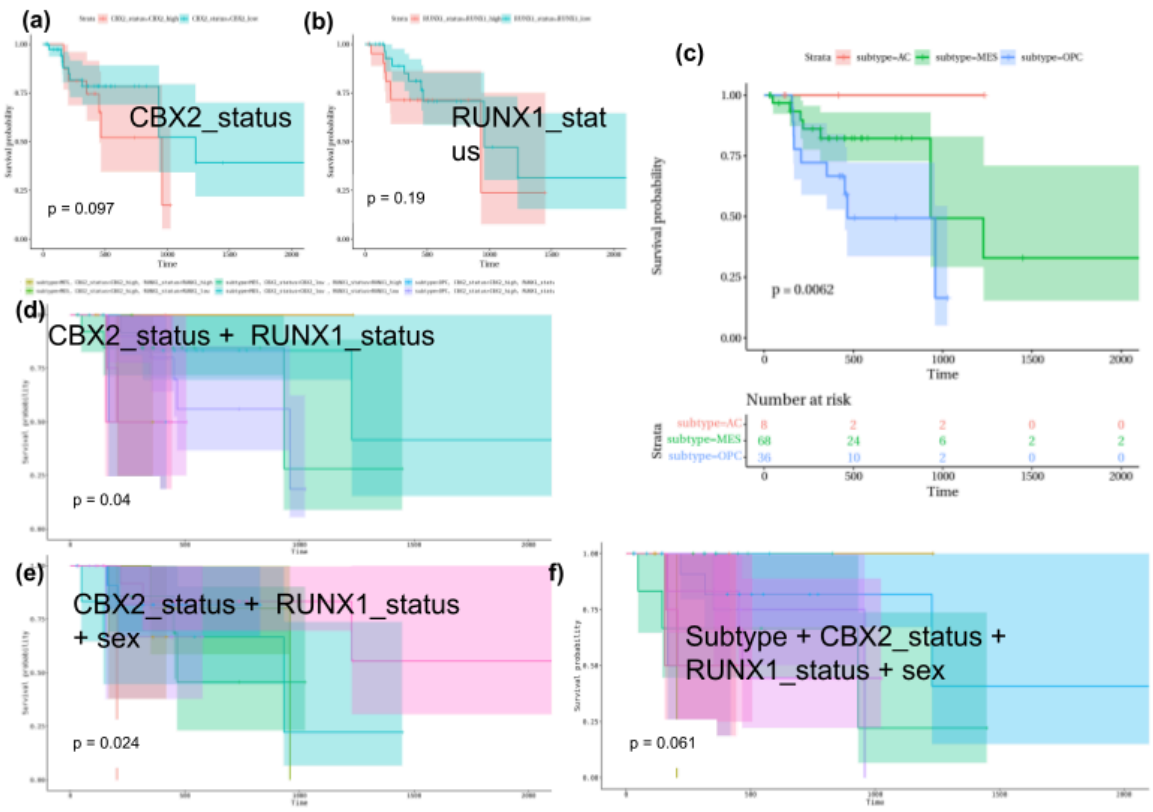

Figure 10. Kalien-Miere curves for (a) only *CBX2* status, (b) only *RUNX1* status, (c) only subtype, (d) *CBX2* and *RUNX1* status, (e) *CBX2* and *RUNX1* status and biological sex, and finally (f) *CBX2* and *RUNX1* and subtype and biological sex

Conclusion:

In conclusion, we have developed and evaluated a robust pipeline to systematically analyze glioblastoma patient data in order to group them into the most suitable subtypes. Single cell data is very expensive to generate, especially for large-scale clinical studies, but it can give us very fine grain information which is not available in bulk. We have, therefore, used single cell expression markers to deconvolve bulk glioblastoma samples and then we robustly evaluated the subtypes through several clustering and network-based methods. Once we were confident that our subtypes are reliable for downstream tasks, we extracted the features most important for assigning subtypes to patients. These 99 confirmed transcription factors showed an AUC score of 0.7879 to classify glioblastoma patients. These features are important because they can help classify glioblastoma subtypes in cases where subtype information is not available. We also extracted, unbiasedly, two important transcription factors *CDX2*[14] and *RUNX1*[15] which are known to be important biomarkers for glioblastoma. Our survival analysis further showed that these biomarkers together can significantly affect survival of glioblastoma patients. We also saw from our analysis that biological sex plays an important role in survival outcomes. This was very interesting because we were able to confirm this finding from existing literature highlighting the importance of biological sex in studying patient survival outcomes, especially for glioblastoma[16]. Taken together, our approach gives a roadmap into merging single cell resolution information with analysis of bulk patient survival outcome to not only map patients to subtypes but also evaluate putative biomarkers for application in clinical settings. Our approach can, therefore, also be extended to other cancer types as well.

References:

1. Distant metastasis occurs late during the genetic evolution of pancreatic cancer | Nature. Accessed December 8, 2022. https://www.nature.com/articles/nature09515
2. Future medical applications of single-cell sequencing in cancer | SpringerLink. Accessed December 8, 2022. https://link.springer.com/article/10.1186/gm247
3. Tumour heterogeneity in the clinic | Nature. Accessed December 8, 2022. https://www.nature.com/articles/nature12627
4. Stupp R, Mason WP, van den Bent MJ, et al. Radiotherapy plus Concomitant and Adjuvant Temozolomide for Glioblastoma. *N Engl J Med*. 2005;352(10):987-996. doi:10.1056/NEJMoa043330
5. Coactivation of Receptor Tyrosine Kinases Affects the Response of Tumor Cells to Targeted Therapies | Science. Accessed December 8, 2022. https://www.science.org/doi/full/10.1126/science.1142946
6. Bhat KPL, Balasubramaniyan V, Vaillant B, et al. Mesenchymal differentiation mediated by NF-κB promotes radiation resistance in glioblastoma. *Cancer Cell*. 2013;24(3):331-346. doi:10.1016/j.ccr.2013.08.001
7. Chen J, Young SM, Allen C, et al. Identification of a Small Molecule Yeast TORC1 Inhibitor with a Multiplex Screen Based on Flow Cytometry. *ACS Chem Biol*. 2012;7(4):715-722. doi:10.1021/cb200452r
8. Li Z, Bao S, Wu Q, et al. Hypoxia-inducible factors regulate tumorigenic capacity of glioma stem cells. *Cancer Cell*. 2009;15(6):501-513. doi:10.1016/j.ccr.2009.03.018
9. Verhaak RGW, Hoadley KA, Purdom E, et al. Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1. *Cancer Cell*. 2010;17(1):98-110. doi:10.1016/j.ccr.2009.12.020
10. Louis DN, Perry A, Reifenberger G, et al. The 2016 World Health Organization Classification of Tumors of the Central Nervous System: a summary. *Acta Neuropathol (Berl)*. 2016;131(6):803-820. doi:10.1007/s00401-016-1545-1
11. Patel AP, Tirosh I, Trombetta JJ, et al. Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science*. 2014;344(6190):1396-1401. doi:10.1126/science.1254257
12. Sottoriva A, Spiteri I, Piccirillo SGM, et al. Intratumor heterogeneity in human glioblastoma reflects cancer evolutionary dynamics. *Proc Natl Acad Sci U S A*. 2013;110(10):4009-4014. doi:10.1073/pnas.1219747110
13. Wang J, Jenjaroenpun P, Bhinge A, et al. Single-cell gene expression analysis reveals regulators of distinct cell subpopulations among developing human neurons. *Genome Res*. 2017;27(11):1783-1794. doi:10.1101/gr.223313.117
14. Li J, Xu Z, Zhou L, Hu K. Expression profile and prognostic values of Chromobox family members in human glioblastoma. *Aging*. 2022;14(4):1910-1931. doi:10.18632/aging.203912
15. Tuo Z, Zhang Y, Wang X, et al. RUNX1 is a promising prognostic biomarker and related to immune infiltrates of cancer-associated fibroblasts in human cancers. *BMC Cancer*. 2022;22:523. doi:10.1186/s12885-022-09632-y
16. Tian M, Ma W, Chen Y, et al. Impact of gender on the survival of patients with glioblastoma. *Biosci Rep*. 2018;38(6):BSR20180752. doi:10.1042/BSR20180752
17. Neftel C, Laffy J, Filbin MG, et al. An integrative model of cellular states, plasticity and genetics for glioblastoma. *Cell*. 2019;178(4):835-849.e21. doi:10.1016/j.cell.2019.06.024
18. Deconvolution of Blood Microarray Data Identifies Cellular Activation Patterns in Systemic Lupus Erythematosus | PLOS ONE. Accessed December 8, 2022. https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0006098
19. Gong T, Szustakowski JD. DeconRNASeq: a statistical framework for deconvolution of heterogeneous tissue samples based on mRNA-Seq data. *Bioinforma Oxf Engl*. 2013;29(8):1083-1085. doi:10.1093/bioinformatics/btt090
20. Monaco G, Lee B, Xu W, et al. RNA-Seq Signatures Normalized by mRNA Abundance Allow Absolute Deconvolution of Human Immune Cell Types. *Cell Rep*. 2019;26(6):1627-1640.e7. doi:10.1016/j.celrep.2019.01.041
21. Newman AM, Liu CL, Green MR, et al. Robust enumeration of cell subsets from tissue expression profiles. *Nat Methods*. 2015;12(5):453-457. doi:10.1038/nmeth.3337
22. dtangle: accurate and robust cell type deconvolution | Bioinformatics | Oxford Academic. Accessed December 8, 2022. https://academic.oup.com/bioinformatics/article/35/12/2093/5165376

23. Integrated analysis of multimodal single-cell data: Cell. Accessed December 8, 2022. https://www.cell.com/cell/fulltext/S0092-8674(21)00583-3?_returnURL=https%3A%2F%2Flinkinghub.elsevier.com%2Fretrieve%2Fpii%2FS0092867421005833%3Fshowall%3Dtrue
24. Benchmarking of cell type deconvolution pipelines for transcriptomics data | Nature Communications. Accessed December 8, 2022. https://www.nature.com/articles/s41467-020-19015-1
25. Rezaie N, Reese F, Mortazavi A. PyWGCNA: A Python package for weighted gene co-expression network analysis. Published online August 23, 2022:2022.08.22.504852. doi:10.1101/2022.08.22.504852
26. Benchmark and integration of resources for the estimation of human transcription factor activities. Accessed December 8, 2022. https://genome.cshlp.org/content/29/8/1363.short
27. Degenhardt F, Seifert S, Szymczak S. Evaluation of variable selection methods for random forests and omics data sets. *Brief Bioinform*. 2019;20(2):492-503. doi:10.1093/bib/bbx124
28. Chu T, Wang Z, Pe'er D, Danko CG. Cell type and gene expression deconvolution with BayesPrism enables Bayesian integrative analysis across bulk and single-cell RNA sequencing in oncology. *Nat Cancer*. 2022;3(4):505-517. doi:10.1038/s43018-022-00356-3