Dynamic Quantum Clustering of Gliomas RNA-seq Identifies Diagnostic Separation and Survival Gradients

Fereshteh Jahanbani, PharmD, PhD[1,2], Rajan Maynard[1,2], Steven J. Schrodi*[,3,4,5,6,7,8,9], Marvin Weinstein*[,10]

[1]Division of Immunology and Rheumatology, Department of Medicine, Stanford University School of Medicine, Stanford, CA, USA
[2]4C Program, Stanford University, Stanford, CA, USA
[3]Department of Medical Genetics, School of Medicine and Public Health, University of Wisconsin-Madison, Madison, WI, USA
[4]McPherson Eye Research Institute, University of Wisconsin-Madison, Madison, WI, USA
[5]Center for Genomic Science Innovation, University of Wisconsin-Madison, Madison, WI, USA
[6]Center for Precision Medicine, University of Wisconsin-Madison, Madison, WI, USA
[7]Comprehensive Diabetes Center, University of Wisconsin-Madison, Madison, WI, USA
[8]Quantitative Biology Initiative, University of Wisconsin-Madison, Madison, WI, USA
[9]Computation and Informatics in Biology and Medicine, University of Wisconsin-Madison, Madison, WI, USA
[10]SLAC National Accelerator Laboratory, Menlo Park, CA, USA

*Correspondence: schrodi@wisc.edu; niv@slac.stanford.edu

Highlights

- Significant clusters discovered among glioma tumors using 554 RNAs. Overlaying histology labels on these clusters showed 97% discrimination accuracy between low-grade gliomas and glioblastomas.
- Using 90 RNAs, three separate low-grade glioma clusters are identified with markedly different progression-free survival times.
- The mesenchymal/invasion-extracellular matrix axis plays a substantial role in the clustering and survival gradients align with expression profiles along this biological axis.

Abstract

Public RNAseq sample sets can refine per-tumor diagnosis and risk, but heterogeneous biology and analytic drift often obscure structure. Dynamic Quantum Clustering (DQC), an unsupervised geometry preserving method requiring no clinical labels or preset cluster counts, addresses both challenges. Applied to RNAseq from 692 TCGA gliomas (524 low-grade gliomas (LGG), 168 glioblastomas (GBM); 20,057 protein coding genes), DQC produced two dominant clusters with 90.9% post-hoc diagnostic concordance and clear survival time separation. Filtering genes by intercluster mean differences yielded a 554 gene subset that improved accuracy to 97.3%. Rank ordering these genes identified ~90 genes that, under DQC, produced three LGG pure subclusters with ordered, but different survival outcomes and one GBM rich cluster (PPV 97.1%)--the RNA-based clustering without clinical information thereby inherently reveals molecular groupings which mirror critically important clinical features. Comparing these clusters defined four non overlapping gene modules and assigned four “BioCoords” per tumor. DQC with Biocoords recapitulated the LGG-to-GBM continuum with a mesenchymal/invasion-extracellular matrix axis exhibiting a monotonic survival gradient, illustrating how geometry-aware unsupervised learning can translate bench and computational discovery into meaningful biology-based patient stratification and prognosis.

Keywords

Clustering, glioma, glioblastoma, cancer genetics, transcriptomics, diagnostics

Introduction

Public consortia now host large numbers of deeply profiled human samples (e.g., TCGA, GTEx1, dbGaP2), creating an opportunity to convert complex transcriptomic data into clinically useful stratifiers.[3-5] In practice, three obstacles blunt this promise:(i) biological heterogeneity and spatial sampling constraints in small or site-limited biopsies obscure signal[6]; (ii) label uncertainty in clinical annotations and interobserver variability in histology contribute to noise[7]; and (iii) analytic drift, arising from batch/platform effects, tumor purity and stromal/immune admixture, and methods that over-weight high-variance axes—can yield embeddings that track artifacts (proliferation, purity, batch) rather than patient-relevant biology.[8-10]

These constraints motivate clinical data-independent, geometry-first analyses in which data self-organize before interpretation. Conventional unsupervised

approaches[11,12] such as PCA/SVD with k-means or t-SNE are valuable but can be sensitive to tuning choices and may compress away nonlinear, orthogonal structure.[13,14] Dynamic Quantum Clustering (DQC) offers a complementary approach: it constructs a smooth, data-derived potential over expression space and lets samples flow toward nearby high-density clusters, producing neighborhoods without labels or a preset number of clusters.[15] This preserves both discrete boundaries and continuous gradients, helping separate clinically informative biology from confounders.[15-17]

Diffuse gliomas, ranging from lower-grade glioma (LGG) to glioblastoma (GBM)[18-21] are an ideal test bed. GBM is particularly challenging to treat because of marked intratumoral heterogeneity, diffuse infiltration that limits resectability, resistance to chemoradiation, and poor drug penetration across the blood–brain barrier.[22-27] Most tumors recur rapidly and median survival remains ~15 months in contemporary series.[28] Clinical decisions therefore hinge on (1) correctly distinguishing closely related entities, (2) timely risk assessment to guide surveillance and intervention, and (3) anticipating evolution, especially the likelihood that an LGG will adopt GBM-like behavior, so surgery, radiation, systemic therapy, and trials can be timed for maximal benefit. Histology, imaging, and targeted DNA panels remain indispensable yet can be indeterminate in gray zones and often reflect where a tumor was, not where it is heading. RNA, by reporting active programs, complements them with a real-time view of state. Understanding the molecular pathophysiology of gliomas offers to reveal specific etiological mechanisms, motivate prognostics, and determine underlying subtypes which may play a role in treatment response and disease management.

Here we apply DQC to bulk RNA-seq from 692 TCGA gliomas without using clinical labels during structure discovery and then connect the learned geometry to biology and outcome strictly post hoc. Three results follow. First, DQC on all 20,057 protein-coding genes recovers the dominant LGG–GBM manifold and two high-density clusters, with progression-free survival (PFS) differences that align with cluster identity after labels are overlaid. Second, an unsupervised refinement to 554 transcripts preserves the manifold while sharpening diagnostic boundaries; further t-test–based ranking (p-values used for ordering only) yields a compact ~90-gene solution that resolves four clusters, three LGG subclusters with distinct prognosis and one GBM-rich cluster. Third, we translate the 90-gene separation into four non-redundant, biologically interpretable coordinates ("BioCoords"). Re-running DQC on the resulting 692×4 BioCoords matrix recapitulates the LGG→GBM trajectory and cluster structure with ~5,000× dimensionality reduction (20,057→4). Among the axes, the mesenchymal/invasion-loaded A–B coordinate shows the clearest monotonic association with risk: higher A–B values align with a post-hoc gradient of shorter PFS/OS, and within-cluster splits along A–B reveal coherent increases in expression of EMT/ECM genes (e.g., *VIM*, *PDPN*, *TIMP1*, *PLA2G2A*), indicating that A–B orders tumors along an invasion-loaded continuum even inside LGG-pure clusters.

Together, this geometry-aware, label-agnostic workflow, unsupervised discovery with DQC, targeted refinement, translation to compact BioCoords, and post-hoc linkage to biology and outcome, provides a general template for turning public RNA-seq into histology-independent diagnostic and prognostic stratifiers despite heterogeneity, sampling constraints, and experimental drift. The same framework extends to pan-cancer and non-oncology settings and to longitudinal designs (bulk or single-cell,

including cfRNA) to track patient-specific movement on the manifold and anticipate state transitions in real time.

Methods

Cohort and preprocessing

We analyzed TCGA glioma bulk RNA-seq data comprising 692 tumors with matched clinical metadata (histology and time-to-event variables for progression-free and overall survival). Expression matrices were restricted to 20,057 protein-coding genes, $\log_2$-scaled, and L2-normalized per sample (each sample vector scaled to unit L2 norm). Principal components (PC1–PC3) were used solely as a visualization space for displaying DQC evolution frames; all DQC computations operated on the full feature space appropriate to each analysis.

All clustering and feature selection steps were performed without clinical labels. Histology (LGG/GBM) and survival annotations (PFS/OS) from TCGA were overlaid only after each unsupervised DQC run to compute concordance metrics and to relate clusters/coordinates to outcomes.

1. Dynamic Quantum Clustering (DQC) hyperparameters and visualization

DQC organizes samples into intrinsic neighborhoods by constructing a smooth, data-derived potential over the expression space and allowing samples to move toward nearby high-density regions of the data. Operationally, an initially diffuse cloud contracts onto density ridges and subsequently settles into high-density clusters. This geometry-preserving evolution makes the clustering process auditable across successive frames. An intuitive description of the DQC potential function and evolution dynamics, together with illustrative model examples, is provided in **Note S1** and **Figures S1, S2,** and **S3** (panels **S3-1** to **S3-8** and **Supplemental Animation DQC_Evolution_Dynamics**). We applied the identical DQC settings across four inputs: (i) the full 20,057 protein-coding genes, (ii) the unsupervised 554-gene “elbow” panel (Section 2), (iii) rank-ordered top-k subsets (k = 140, 120, 100, 90; Section 3), and (iv) a four-coordinate embedding derived from compact non-overlapping gene modules (“BioCoords”; Section 4). For each run we recorded representative frames at initialization (“cloud”), at an intermediate “ridge” stage, and at convergence (“clusters”). Principal-component axes (PC1–PC3) were used only for display; DQC itself operated in the feature space supplied to each run. No clinical labels (histology or outcomes) were used during DQC or feature selection; labels were overlaid only after clustering for post-hoc evaluation of diagnostic concordance and survival linkage. Robustness was assessed by repeating runs with doubled PCA dimensionality in the visualization step; partitions were qualitatively unchanged, indicating that conclusions reflect the input features rather than visualization settings.

For all downstream analyses (elbow 554-gene panel, top-k panels, and BioCoords), we fixed the PCA input dimensionality at 16 components (the value selected and validated during the initial setup) and held all other DQC hyperparameters constant; no higher-dimensional PCA inputs were tested at later stages to maintain consistency.

2. Elbow-based unsupervised feature selection

To derive a label-free gene set aligned with the Step-1 DQC structure, for each gene we computed the absolute difference in mean normalized expression between the two unlabeled high-density clusters (|Δμ|), then ranked genes to form a monotonically increasing |Δμ| curve. We identified the elbow by (i) visually locating the bend and (ii) independently confirming it with a simple first-derivative scan that flagged the first index at which the slope became near-vertical; the two estimates coincided (**Figure 2A**). On the $\log_2$-scaled, L2-normalized matrix, this elbow occurred at |Δμ| ≈ 0.0036, retaining 554 genes. We then re-ran DQC on the resulting 692 tumor samples by 554 genes matrix, and only after clustering overlaid TCGA histology and survival for post hoc evaluation.

Differential expression and pathway analysis. To interpret the biology of the two 554-gene DQC clusters, we performed differential expression on the same normalized matrix using Welch's t-test and applied Benjamini–Hochberg (BH) correction across all tested genes, declaring significance at FDR<0.05. Volcano plots displayed robust $\log_2$ fold-change against $-\log_{10}$ p. Canonical pathway enrichment was assessed with Ingenuity Pathway Analysis (IPA), using BH-adjusted statistics within IPA; exported fields included pathway name, $-\log_{10}$(BH p-value), ratio, z-score, and member molecules. These analyses served purely as post-hoc interpretation and did not influence DQC, elbow selection, or any clustering step.

3. Rank-ordered top-k panels

To test whether progressively smaller gene panels preserve the unsupervised DQC structure (i.e., clear cluster separation with graded patterns across clusters), we ranked the 544 genes by a Welch t-test comparing the two DQC clusters from the elbow step. The resulting p-values were used solely as monotonic ranking scores (feature ordering; no hypothesis testing). Clustering quality was assessed by repeating DQC across the rank-ordered top-k panels with the same fixed settings. We assembled top-k panels (k = 140, 120, 100, 90) and re-ran DQC on each reduced matrix under identical DQC settings established in Step-1, including the same input dimensionality (16 principal components); no runs with increased dimensionality were performed at later stages to maintain consistency. Robustness was assessed by repeating runs with doubled PCA dimensionality in the visualization step; partitions were qualitatively unchanged, indicating that conclusions reflect the input features rather than visualization settings.

After clustering, we post hoc overlaid TCGA histology to summarize cluster purity and assessed clinical separation by Kaplan–Meier analyses for progression-free and overall survival using two-sided log-rank tests (reporting KM medians and/or means with 95% CIs). All uses of histology and survival were strictly post hoc.

4. Construction of four compact coordinates ("BioCoords")

To better understand the specific biological modules contributing to the clustering structure, we compressed the 90-gene DQC separation into four low-dimensional, label-free coordinates using gene modules. These modules were derived from pairwise comparison among the four DQC clusters (labeled A, B. C, and D). Starting from the 692 tumor samples by 20,057 $\log_2$-scaled genes, L2-normalized expression matrix, we considered four comparisons that cover the observed A→D→C→B ordering with minimal redundancy: A–B and D–B (two distances along the LGG→GBM direction) and

A–C and C–D (two steps along the LGG-internal gradient). The two omitted pairs (A–D and B–C) added little independent structure. All 20,057 protein-coding genes were scored by a two-sample t-test computed on the two clusters being compared; the resulting p-values were used strictly to rank genes (ordering only; no hypothesis testing, no multiple-testing adjustment at this step). From each ranked list we took the top 30 genes, then removed overlaps by retaining a shared gene only in the comparison where it ranked highest. This yielded four non-overlapping modules of 24 (A–B), 25 (A–C), 27 (D–B), and 27 (C–D) genes (103 total).

For each module, we extracted the corresponding columns from the 692 tumor samples by 20,057 gene matrix and standardized each gene across tumors by column-wise z-scoring (subtract that gene's cohort mean; divided by its cohort standard deviation). This places genes on a common scale so that no high-variance gene dominates the module average; genes with zero variance would be excluded by design (none occurred). We then averaged the standardized gene columns within a module to obtain a single per-tumor coordinate for that module. Concatenating the four module coordinates produced 692 tumor samples by 4 "BioCoords" matrix. Similar methods for visualization based on biological axes have previously been described.[29-33]

We ran DQC on this BioCoords matrix to test whether the progression-like geometry and separation seen with the 90-gene matrix were preserved in this reduced representation. All DQC runs, feature rankings, and coordinate constructions were performed without histology or outcome labels; labels were overlaid only after clustering/coordinate construction for evaluation. Importantly, plotting tumors in the BioCoords space enables targeted interrogation of geometry with respect to specific biological programs: because each coordinate is defined by an explicit, non-overlapping gene module, one can visualize how pathway-level activities (e.g., mesenchymal/ECM remodeling or developmental HOX signals) shift sample positions and influence cluster boundaries in this low-dimensional map.

For post hoc clinical/label evaluation, after constructing BioCoords (and after DQC), we overlaid TCGA histology to compute concordance metrics and evaluated associations of clusters and BioCoords with PFS/OS using Kaplan–Meier analyses (days from diagnosis to event or last follow-up; two-sided log-rank tests; samples with missing times excluded).

5. General statistics, software, and reproducibility

Differential expression and IPA pathway analyses controlled multiple testing by Benjamini–Hochberg FDR at q=0.05. In contrast, the t-test p-values used in Section 3 and Section 4 were employed strictly for ranking (ordering only) and were not used for inference or subjected to multiple-testing correction. DQC and the elbow-based gene selection were fully unsupervised and did not use histology or outcome labels; the elbow cutoff (~0.0036 on $\log_2$, L2-normalized data) was set by a visual/first-derivative knee criterion (Kneedle gave a similar vicinity but was not the final rule). PCA was used only for visualization of DQC frames. Analyses were performed in Maple for DQC and standard Python/R environments (statistics/plotting) with IPA for canonical pathway enrichment. Exact preprocessing ($\log_2$ scaling, L2 normalization), DQC hyperparameters, random seeds, and scripts for ranking, de-duplication, coordinate

construction, and post hoc evaluation are provided in **Extended Data/Code** to enable reproduction.

Results

All clustering and gene-selection steps were performed without histology or outcome labels; TCGA histology and progression-free survival (PFS)/overall survival (OS) were applied post hoc for evaluation of how clinical features match to clustering patterns.

1. DQC on all protein-coding genes reveals two dominant clusters of samples

Applying DQC to 692 TCGA gliomas using all 20,057 protein-coding genes ($\log_2$-scaled, L2-normalized per sample) and no clinical labels or parameter fine-tuning, the evolution frames (shown in the first three PCs for display only) already partition the initial heterogeneous cloud into two major groups (**Figure 1A, Table S1**). As evolution proceeds, the cloud samples contract and by intermediate frame into an extended one-dimensional shape. At convergence DQC resolves two high-density clusters plus six isolated points (**Figure 1B–C**). Post hoc overlaying the TCGA clinical metadata showed cluster 1 was LGG-enriched (469/473 LGG, ~99.2%) and cluster 2 GBM-enriched (161/213 GBM, ~76%); three LGG and three GBM were isolated points. Using histology solely as an external reference, overall diagnostic accuracy of the unsupervised DQC assignment was 630/692 (90.9%) (**Figure 1D**). Kaplan–Meier curves stratified by cluster demonstrated longer PFS in the LGG-enriched cluster versus the GBM-enriched cluster, OS showed the same direction (**Figure 1E**). These Step-1 observations motivated a targeted refinement to preserve the learned geometry while sharpening subtype separation.

2. Gene refinement by quantitative elbow determination sharpens subtype stratification

To denoise features, we ranked genes by the absolute difference in mean normalized expression between the two Step-1 clusters, producing a monotonically decreasing curve. We then visually identified the elbow of this curve and independently confirmed it with a simple numerical first-derivative scan. Specifically, we marked the first point at which the slope became near-vertical and verified that this point coincided with the visual bend (**Figure 2A**). This procedure yielded a 554-gene panel (**Table S2**). Re-running DQC on the 692×554 matrix preserved the global manifold and reduced mixed-label samples at the LGG/GBM boundary. (**Figure 2B**; **Fig. S4**, panels **S4-1—S4-3**). Only after clustering was complete did we overlay TCGA histology to quantify concordance between DQC structure and clinical features. Of 524 LGG, 510 mapped to the DQC LGG-enriched cluster1 and 14 to the DQC GBM-enriched cluster2; of 168 GBM, 163 mapped to the DQC cluster2 and 5 to the DQC cluster1, yielding 97.3% overall accuracy (673/692) with PPV 99.0% for LGG and 92.1% for GBM (**Figure 2B**). Representative frames of the DQC evolution illustrate the contraction of the diffuse data cloud into two stable endpoint clusters while preserving the underlying manifold structure (**Fig. S4**). As a sensitivity check, the PCA dimensionality was doubled and produced near-identical partitions, supporting the robustness of the elbow cutoff and feature set. Because DQC is unsupervised, we then performed a post hoc differential analysis between the two DQC-defined clusters to interpret the biological programs that distinguish them.

Post-hoc differential expression and pathway context

After DQC clustering was completed (without labels), we contrasted the two clusters on the same normalized matrix for biological interpretation. For each gene, we computed a two-sample Welch t-test, where corresponding p-values were Benjamini–Hochberg (BH) adjusted across all genes, and DEGs were defined at FDR < 0.05 (**Figure 2C**; **Table S2**). Ingenuity Pathway Analysis (BH-adjusted) of the DEG set highlighted extracellular-matrix and invasion programs (e.g., matrix metalloprotease activation/degradation, collagen pathways), leukocyte adhesion/diapedesis and vascular transmigration, PI3K-linked motility, and pro-inflammatory signaling (IL-17, NF-κB, PI3-family) (**Figure S5**). Consistent with these enrichments, the GBM-enriched cluster showed higher expression of ECM/adhesion and inflammatory genes (*POSTN*, *MMP7*, *CHI3L1*, *PLA2G2A*, SAA1, *CAPN6*) together with a pronounced developmental HOX signal (*HOXA5/6*, *HOXD11*, *HOXC13*, *HOXA-AS3*) and the glioma-associated membrane protein EMP3; in contrast, the LGG-enriched cluster showed relative up-regulation of neuronal/neurite and neurodevelopmental markers (*NEUROD6*, *SPHKAP*), along with *PRLHR*, *CSMD3*, and *DEFB119* (**Table S4**). These post-hoc analyses were used solely for biological interpretation and not for clustering, gene filtering, or model construction, and they support that the unsupervised DQC clusters correspond to a GBM-like state characterized by ECM remodeling/inflammation and an LGG-like state retaining neuronal features (**Table S4** and **Figure S5**). IPA-curated annotations of differentially expressed genes, including cellular localization and functional context, are provided in **Table S3**.

Post hoc clinical linkage

For assessment of clinical outcomes, we used TCGA time-to-event metadata to generate Kaplan–Meier curves stratified by the two DQC clusters under the 554-gene result: progression-free survival (PFS) was markedly shorter in the GBM-enriched cluster2 and prolonged in the LGG-enriched cluster1 (**Figure 2E**), with overall survival showing the same direction of effect. Together, these findings indicate that an unsupervised, geometry-preserving refinement to a few hundred transcripts improves diagnostic stratification and aligns with clinical outcomes without the use of clinical labels (histology, PFS/OS) in the clustering process.

3. t-test–ranked reduction to ~90 genes preserves diagnostic utility and refines the prognostic substructure

Building on the 554-gene manifold, we asked whether a smaller panel could retain diagnostic separation and expose additional biological insights. To select a reduced set of features, we computed a two-sample t-test for each gene between the two DQC clusters and used the resulting p-values to rank the genes (i.e., the p-values are utilized as a summary metric to quantify the difference between the clusters in a semi-standardized manner, not to test hypotheses or declare statistical significance). Using ranking, we systematically evaluated DQC-derived geometry changes on sets of the top genes (top-k subsets: k = 140, 120, 100, 90) (**Table S2**). This aided with assessment of both robustness of the clustering and feature importance.

Top-140 genes analysis yielding three clusters

With the use of the top 140 ranked genes, DQC produced three clusters (Cluster A = 450, Cluster B = 177, Cluster C = 65). Post hoc, A and C were entirely LGG, while B contained all 168 GBM plus 9 LGG (GBM PPV = 94.9%, 168/177). Kaplan–Meier analyses showed graded outcomes across clusters (log-rank P<0.001), with mean PFS/OS (days) of A 811/1053, C 390/660, and B 275/445, indicating two LGG subclusters with distinct prognosis and a GBM-rich cluster with the poorest survival (**Table S5-1**).

Top-100 genes analysis yielding four clusters

Geometry refined the data to four clusters (A = 421, B = 173, C = 65, D = 33). Post hoc, A/C/D were 100% LGG; GBM-rich B held all 168 GBM + 5 LGG (GBM PPV = 97.1%, error = 2.9%). Survival remained ordered (log-rank P<0.001), with mean PFS/OS (days) A 835/1062, C 397/602, D 448/1044, B 263/430 (**Table S5-2**).

Top-90 genes analysis (compact panel) yields four clusters

The four-cluster structure persisted (A = 421 LGG; B = 168 GBM + 5 LGG; C = 65 LGG; D = 33 LGG). LGG clusters remained pure; GBM-rich B retained PPV = 97.1% (168/173). Prognostic separation was consistent (log-rank P<0.001). Mean PFS/OS (days) were A 836/1063, C 406/692, D 458/871, B 263/430, reinforcing a reproducible LGG substructure with distinct clinical outcomes (**Table S5-3**). Performance deteriorated below ~90 genes and clustering ultimately collapsed, indicating a stable “sweet spot” near 90–100 genes.

Together, these results show that selecting genes based on t-test p-values ranking (for ordering only) progressively denoises the data, preserves the DQC manifold, improves diagnostic precision, especially for the GBM-rich cluster, and reveals clinically meaningful LGG subtypes (**Figure 3A**). All clinical labels (histology and PFS/OS) were applied strictly post hoc for evaluation (**Methods**). Full survival definitions and model specifications are provided in **Methods/Extended Data**.

4. Translating the 90-gene separation into four compact biological coordinates for risk stratification

The 90-gene DQC solution cleanly separates tumors into four clusters with distinct prognosis: A/D/C are pure LGG with progressively shorter PFS/OS (A > D > C), while B is almost entirely GBM (plus 5 LGG) (**Figure 3A**, **Figure S6-1**). Cluster A shows the longest average PFS/OS (836/1063 days), followed by cluster D (458/871 days) and cluster C (406/692 days), whereas the GBM-rich cluster B exhibits the poorest outcomes (262/430 days). Kaplan-Meier analysis confirms this ordered survival separation across the four clusters (**Figure 3B**).

The high purity of the LGG clusters and the GBM-rich cluster suggests that the underlying structure may be summarized by compact gene modules. We therefore asked whether a limited set of gene module-defined axes could summarize these contrasts and allow each tumor to be positioned along them using expression alone.

Defining four non-redundant LGG–GBM contrasts

To span the observed ordering A → D → C → B, we selected four contrasts: A–B and D–B (LGG→GBM at different distances from GBM (near vs far from GBM) and A–C and C–D (the LGG-internal gradient from longer- to shorter-survival groups). The remaining

pairwise comparisons (A–D and B–C) were largely compositional combinations of these and did not introduce meaningful independent structure.

Module construction from all protein-coding genes

For each contrast, all 20,057 protein-coding genes were scored using a two-sample t-test between the corresponding clusters. As the resultant p-values collapsed discrimination signal into a single metric, the resulting p-values were used strictly for ranking the importance ranking of each gene in separating clusters (no statistical inference was performed from these p-values). Because ranking was performed on clusters defined in the higher-dimensional space and not evaluated for statistical significance, this step serves to prioritize genes that reflect the observed geometry rather than to infer differential expression. The top 30 genes per cluster contrast were selected (totaling 120 candidates), and duplicates across contrasts were resolved by retaining each gene only in the contrast where it ranked highest. This yielded four non-overlapping modules of 24, 25, 27, and 27 genes, each representing a compact biological module coordinate axis.

From modules to four BioCoords

For each gene module, we retrieved its genes' expression values from the 692×20,057 matrix, and z-scored each gene across the 692 tumors (subtract the cohort mean and divide by the cohort SD) to place genes on a common scale (genes with zero variance would be excluded; none occurred). We then averaged the standardized values within the module to obtain one score per tumor. Repeating this for all four modules yielded four scores per tumor; stacking them forms a 692×4 BioCoords matrix (four compact coordinates per tumor) that summarizes each tumor's position in the overall pattern (**Figure 3C**).

Validation of the four-coordinate representation

In the 90 gene analyses, tumors organize along a continuous progression-like path that resolves into four clusters (A-D), with clear separation of the long-survival LGG extreme (cluster A), intermediate LGG states (clusters D and C), and the GBM-rich poor-survival extreme (cluster B) (**Figure 3A**). On the other hand, in the reduced BioCoord space the relative positions of clusters are preserved while the separation between the LGG sub-clusters (C and D) becomes somewhat less pronounced, indicating that four coordinates retain the dominant structure while modestly compressing finer distinctions (**Figure 3C**). Only after clustering was TCGA histology and survival overlaid on the data.

Among the four coordinates, the A–B (mesenchymal) axis shows the clearest monotonic risk gradient, functions as the dominant progression-like coordinate along the LGG→GBM continuum. The distribution of tumors along this axis reveals a continuous shift from long-survival LGG to GBM-rich tumors, with intermediate LGG states occupying transitional regions (**Figure 3D**). By contrast, the remaining axes (A-C, D-B, and C-D) provide orthogonal biological refinements that capture additional structure within LGG but show weaker associations with survival.

Thus, four compact modules recapitulate the progression and cluster structure, place each tumor along interpretable biological directions, and retain prognostic ordering while reducing dimensionality from 90 features to 4 (≈22.5× reduction), and

from the original 20,057 genes to 4 biological coordinates (which includes only 103 genes). All label overlays were strictly post hoc.

Linking biological networks to DQC geometry with BioCoords

To understand the biological meaning of the BioCoord axes, we examined the top-ranked genes contributing to each coordinate. Each axis corresponds to a distinct biological program that explains both the spatial organization of tumors in DQC space and their associated clinical behavior (full gene lists in **Table S1**).

The A–B axis represents a mesenchymal and invasion-related program (24 genes: ANXA2P2, CASP6, CLIC1, EMP3, GNG5, HOXA6, HOXB9, HOXC11, HOXC13, HOXC9, HOXD11, IGF2BP3, IGFBP2, MAGED4B, PABPC3, PAX3, PDPN, PLA2G2A, RAB42, SLC43A3, TGIF1, TIMP1, TUBA1C, VIM). This axis concentrates canonical extracellular matrix (ECM) remodeling and invasion genes together with HOX-driven transcriptional reprogramming. Higher values along this axis correspond to increasingly invasive tumor states and are associated with poorer clinical outcomes. The A–C axis captures developmental and lineage identity (25 genes: DCTD, EN1, EVX2, GNG12, HOXA-AS2, HOXA-AS3, HOXA1, HOXA2, HOXA3, HOXA4, HOXA5, HOXA7, HOXA10, HOXC4, HOXC6, HOXC10, HOXD9, HOXD10, HOXD13, ISL2, OSR2, OTP, PTGFRN, PYGL, RAET1K). This axis is dominated by HOX cluster genes and neural patterning factors, consistent with progressive erosion of lineage-specific programs as tumors transition away from more differentiated states.

The D–B axis reflects coordinated disruption across multiple cellular systems (27 genes: ARMC10, CHCHD2, CKS1B, COMMD10, EIF3CL, FAM226B, GUCA2A, HSP90B2P, KCNB1, LIMS3, MRPL36, MYH14, PGAM4, PIPSL, PPIA, PRELID1, RAP1B, RHAG, RPL13AP6, RPS3A, RPS7, SGCD, SLC25A21-AS1, SPAG4, TMEM183B, UQCRHL, ZFAND2A). These genes link mitochondrial oxidative phosphorylation, translational machinery, proteostasis, and neuronal identity, suggesting increasing metabolic stress and loss of specialized cellular function. The C–D axis represents a second layer of coordinated program collapse (27 genes: C1QTNF2, CCDC150, CCDC34, CDCA7L, CHAF1B, DNAJB12, DNAJC12, DNAJC2, HAP1, HOXA-AS2, HOXA-AS3, HOXA2, HOXA3, HOXA5, HOXA10, HOXA10-AS, HOXB3, HOXB7, HOXD10, HOXD12, HOXD13, LIG1, MDGA2, SKP2, SPOCK1, SUSD5, ZNF473). This axis integrates developmental regulation (HOX genes), chaperone/co-chaperone systems (DNAJ family), chromatin replication machinery, and neuronal adhesion pathways, consistent with progressive breakdown of coordinated cellular programs. Together, these results demonstrate that each BioCoord axis corresponds to a biologically interpretable network, allowing tumor position in DQC space to be directly linked to underlying molecular programs driving progression.

Survival structure across biological coordinate space

To better understand how clinical outcome relates to the reduced biological coordinate system, we next visualized PFS directly in BioCoord space by plotting each tumor using its mesenchymal coordinate, developmental coordinate, and PFS value as the three displayed variables (**Figure 4A–C**).

These plots show that survival is not organized along the mesenchymal axis alone. Instead, in the LGG-dominant region, tumors occupy an extended structure in the joint mesenchymal–developmental plane, across which both the average and variance

of PFS vary continuously. In the earliest frame (**Figure 4A**), tumors are broadly dispersed, with substantial heterogeneity in both biological position and survival. As DQC evolution proceeds (**Figure 4B**), this broad cloud contracts into a more organized two-dimensional sheet structure, making the ordering of tumors by survival increasingly apparent. By the late frame (**Figure 4C**), the point cloud resolves into compact endpoint groups with distinct average PFS values.

This organization is informative in two ways. First, it indicates that glioma progression may be better represented as movement along a curved biological manifold than as a shift along a single linear axis. Second, it suggests that as tumors progress from healthier LGG-like states toward more aggressive states, not only does average survival decrease, but the range of possible outcomes might also narrow. In the LGG compartment, tumors are distributed over a broader survival surface, whereas the GBM-rich endpoint is more compact and associated with uniformly poor outcomes. Residual variation within the endpoint groups might reflect biological factors not fully captured by the four-BioCoord representation.

Together, these observations support the view that the BioCoords capture a substantial fraction of the progression-related structure in the data and provide a compact framework for visualizing how tumor biology and clinical outcome vary jointly across the LGG→GBM continuum.

Discussion

Public RNA-seq archives contain immense, largely untapped diagnostic and prognostic information, but extracting clinically coherent structure from these data has been hindered by heterogeneity, technical noise, and the dependence of most analytic pipelines on curated labels. Here, we demonstrate that Dynamic Quantum Clustering (DQC), a geometry-preserving, unsupervised approach, can systematically recover meaningful biological manifolds and prognostic stratification directly from unannotated expression profiles. Without using histological or outcome labels, DQC organized 692 TCGA gliomas into morphologically and clinically concordant groups, achieving 91–97% post-hoc diagnostic agreement and recapitulating key features of the LGG→GBM continuum in both raw and reduced representations.

Several principles emerge. First, manifold geometry, rather than variance magnitude alone, can carry clinically relevant signal: by letting samples flow along the data-derived potential rather than compressing variance as in PCA or UMAP, DQC preserves subtle nonlinear relationships that parallel histopathological boundaries and continuous grade transitions. Stepwise, unsupervised gene refinement from ~20,000 protein-coding genes to a compact 90-gene panel sharpened diagnostic boundaries and revealed LGG subclusters with ordered progression-free and overall survival, indicating that much of the diagnostic and prognostic information in diffuse gliomas is encoded in a relatively low-dimensional, geometry-preserving representation of the transcriptome. The identified 90-gene panel, and its translation into four orthogonal BioCoords, provides interpretable axes that remain faithful to the high-dimensional geometry while allowing intuitive visualization and hypothesis testing, yielding a conceptually simple, reproducible representation of glioma state that appears robust to sampling variation and analytic drift, although this will require external validation.

Importantly, this four-coordinate representation preserves both the global progression structure and the separation of clinically meaningful tumor groups while compressing finer distinctions between closely related LGG subclusters. In particular, although clusters C and D become less sharply separated in the reduced BioCoord space, their relative positioning and prognostic ordering are retained, indicating that the dominant biological structure can be captured with substantial dimensionality reduction while modestly smoothing higher-resolution variation.

Post-hoc pathway enrichment (for example, IPA canonical pathways) offered useful biological context for differentially expressed signals between DQC clusters, but it inherits well-known limitations: curated pathways overlap extensively; gene sets differ in size and specificity; and term ranking can be biased by literature density and database curation rather than cohort-specific effect sizes. Such factors can bias enrichment analyses away from key cohort-specific programs and toward heavily studied, tissue- or disease-centric pathways in public databases. By design, our framework avoids these priors during structure discovery: the geometry is learned in an unsupervised, data-driven way and only then summarized by four compact BioCoords, with pathway annotations used strictly for interpretation rather than feature selection or model fitting.

Translating the 90-gene manifold into four non-overlapping BioCoords further clarified the underlying biology. In this embedding, one axis enriched for mesenchymal and ECM-remodeling programs functioned as a pseudo-temporal progression coordinate: the A–B BioCoord mirrored a mesenchymal/ECM-invasion gradient with monotonic deterioration in PFS and OS and aligned with established mesenchymal transition and microenvironmental remodeling signatures described in GBM. Higher values on this axis are aligned with increased expression of invasion- and EMT-associated genes and with shorter PFS and OS, even within LGG-pure clusters. The remaining coordinates captured orthogonal processes linked to developmental identity loss, mitochondrial and proteostatic collapse, and chromatin reorganization, all hallmarks of malignant progression. Notably, these axes stratified outcomes even within histologically pure LGG clusters, suggesting that conventional grade underestimates ongoing molecular evolution. The DQC-derived geometry thus unifies known subtype markers (for example, IDH-related neuronal signatures and HOX-associated mesenchymal programs) within a single unsupervised continuum.

When clinical outcome is projected into this coordinate system, prognosis does not align strictly with a single dominant axis but instead occupies a structured, low-dimensional surface defined jointly by the mesenchymal and developmental coordinates. In this space, LGG tumors span a broader region with variable survival, whereas GBM-rich tumors occupy a more compact region associated with uniformly poor outcomes. This organization suggests that outcome reflects the combined influence of multiple biological programs rather than a single progression axis. Furthermore, the evolution of the DQC representation reveals that this structure is progressively uncovered rather than imposed: an initially diffuse distribution of tumors contracts into a lower-dimensional sheet and ultimately into compact endpoint groups with distinct survival characteristics. This behavior supports a model in which glioma progression involves both continuous transitions and discrete state consolidation along an underlying biological manifold.

Methodologically, our approach differs from more commonly used workflows that combine linear factorization (for example, PCA) or non-linear embeddings (for example, t-SNE, UMAP) with density-based clustering or graph-based community detection. Those methods are powerful but can be sensitive to tuning choices, batch structure, and high-variance confounders such as tumor purity and proliferation. In contrast, DQC constructs a smooth potential directly from the data and follows the evolution of sample positions as they flow toward density peaks or ridges, preserving both discrete boundaries and continuous gradients. In the TCGA glioma setting, this yielded clusters and trajectories that were robust to changes in visualization dimensionality and to substantial feature reduction, suggesting that the recovered structure reflects underlying biology rather than artifacts of a particular embedding or distance metric. More broadly, the DQC→elbow→top-k→BioCoords pipeline provides a reproducible template for distilling large public RNA-seq resources into compact, auditably derived genomic coordinates suitable for downstream modeling, integration with DNA-level features, or prospective validation in independent cohorts.

In clinical terms, DQC-derived coordinates could serve as quantitative phenotypes for glioma risk stratification and longitudinal tracking. Projecting serial or cfRNA profiles onto the learned manifold may allow inference of directionality along the mesenchymal or metabolic axes, signaling malignant transformation or therapeutic response before conventional imaging changes occur. Because the workflow depends solely on expression geometry, it is adaptable to other cancers and potentially to immune, neurodegenerative, or inflammatory contexts where state transitions define pathology.

Several limitations should be acknowledged. First, all analyses were performed on retrospective bulk TCGA data, which are subject to variable tumor purity, spatial sampling bias, and cohort-specific technical effects; prospective validation in independent glioma datasets and, ideally, in clinically annotated prospective series will be essential before any translational deployment. Second, although our workflow is fully label-agnostic during structure discovery, the biological and clinical interpretation relies on post-hoc overlays of histology, survival, and pathway annotations; this raises the possibility that unmeasured covariates or treatment differences contribute to some of the observed gradients. Third, DQC itself has hyperparameters and modeling choices (for example, kernel width, density-smoothing scale, preprocessing, and normalization) that, while held fixed and stress-tested here, may need adaptation or additional regularization in other disease contexts or for single-cell and cfRNA applications. Finally, our BioCoords were derived from TCGA-specific contrasts and gene rankings, and although the modules are mechanistically interpretable, their exact composition and thresholds may not transfer directly across platforms, ancestries, or sequencing protocols without recalibration.

Despite these caveats, the present study establishes that geometry-preserving, unsupervised clustering coupled to compact, biologically grounded genomic coordinates can robustly reconstruct histology-like classes, expose prognostic heterogeneity within histologic categories, and suggest invasion- and lineage-linked axes of risk in glioma.

The same strategy can, in principle, be extended to pan-cancer and non-oncology cohorts, as well as to longitudinal bulk, single-cell, or cfRNA designs,

where monitoring patient-specific movement on a learned manifold could enable genuinely dynamic, genomics-informed disease staging and treatment planning.

Diffuse gliomas remain clinically challenging because histology and targeted DNA panels do not fully capture the dynamic biology that drives recurrence and malignant transformation. A geometry-aware, label-agnostic workflow based on Dynamic Quantum Clustering applied to bulk RNA-seq recovered key features of the LGG→GBM continuum, refined diagnostic boundaries, and exposed prognostic substructure without using clinical labels during discovery. Progressive reduction from 20,057 genes to a 90-gene panel preserved and sharpened this geometry and, when translated into four non-overlapping BioCoords, yielded compact, biologically interpretable axes that stratify prognosis and could be implemented as practical assays for histology-independent risk stratification and longitudinal tracking.

Figure Legends

Figure 1. Dynamic Quantum Clustering (DQC) of TCGA gliomas: unsupervised structure with post hoc clinical overlay.

1A-C (top): Representative frames from the DQC evolution, shown in the first three principal components, using all 20,057 protein-coding genes (n=692), DQC was run without clinical labels. The data point colors shown here are assigned post hoc for evaluation only: magenta (LGG) / blue (GBM) reflect TCGA histology overlays applied after clustering; axes (PC1-3) are for display.

1D: Post hoc confusion matrix comparing final DQC clustering assignments to TCGA histology. Table depicts a comparison of clustering results to histologic diagnoses. “Isolated” denotes samples not assigned to either major clustering groups at convergence.

1E: Post hoc Kaplan-Meier curves for TCGA progression-free survival (PFS), stratified by DQC clustering. The GBM-enriched cluster shows significantly shorter PFS compared to the LGG-enriched cluster (log-rank $P<0.001$). Both histology labels and PFS/OS annotations were used only for post hoc evaluation of the unsupervised DQC results; they were not used to determine the clustering.

Figure 2. Quantitative elbow for gene refinement and subsequent improved DQC-defined subtype stratification (labels used only post hoc).

2A: Ranked absolute mean difference in normalized expression between the two Step-1 DQC basins (basins obtained from expression alone). A smooth fit (blue) and its first derivative (red) identify an elbow at ~0.0036, retaining 554 genes. TCGA histology and PFS/OS are applied post hoc in panels B-D.

2B: Post hoc comparison to TCGA histology after re-running DQC on the 692x554 matrix (log_2-scaled, L2-normalized): 510 of the 515 samples within the LGG-enriched cluster were diagnosed as LGG via histology. 163 of the 177 samples within the GBM-enriched cluster were diagnosed as GBM via histology. Of the total sample set, 19 samples were discrepant and 673 were concordant between the clustering and the

histology determination. Overall accuracy is 97.3% (673/692); PPV is 99.0% for LGG and 92.1% for GBM.
2C: IPA-derived volcano plot of differentially expressed genes between the two DQC clusters (554-gene refinement). X-axis: log-fold change (DQC-derived cluster 1 / cluster 2). Y-axis: $-\log_{10}$ p (Welch t-test on normalized expression); BH-FDR = 0.05. Red points are higher in the cluster predominantly aligning post hoc with GBM; green points are higher in the cluster predominantly aligning post hoc with LGG. Histology is used only for post hoc interpretation.
2D: Post hoc survival linkage (Kaplan-Meier PFS from TCGA clinical metadata) stratified by the two DQC clusters. Earlier progression is observed in the DQC GBM-enriched cluster (blue), while there is delayed progression in the DQC LGG-enriched cluster (magenta), consistent with subtype sharpening under the refined gene panel.

Figure 3. Compression of the 90-gene DQC structure into biological coordinates for risk stratification

3A: DQC-based clustering using the 90-gene signature. An unsupervised reduction of the 554-gene set, driven by rank-ordering of genes by between-cluster contrast, yielded a 90-gene panel. DQC resolved tumors into four clusters (cluster A–cluster D) without the use of clinical labels.
3B: Post hoc survival analysis. Kaplan–Meier curves demonstrate that the DQC-defined clusters, identified purely through gene expression, align closely with clinical outcomes. A progressive decrease in both PFS and OS is observed following the cluster A→D→C→B geometric ordering.
3C: Low-dimensional biological embedding. Projection of tumors into a reduced "BioCoord" space, defined by averaging z-scored expression values within contrast-derived gene modules, preserves the global geometry of the dataset. This 3D representation captures the same LGG→GBM progression structure observed in the full 20,057-gene space. The clusters occupy an elongated trajectory (A→D→C→B) consistent with a dominant mesenchymal/invasion-associated gradient.
3D: Coordinate-based risk stratification. Distribution of tumors along the dominant BioCoord (mesenchymal/invasion axis). Lower coordinate values are enriched for long-surviving LGG samples (red), shortened surviving LGG (yellow), whereas higher values correspond to short survival GBM-rich tumors (blue), which might suggest progression-aligned axis for molecular risk assessment.

Figure 4: Progression-free survival maps onto a low-dimension biological surface defined by mesenchymal and developmal BioCoords.

4A: Early frame of the DQC evolution. Showing tumors plotted in a three-dimensional space defined by the mesenchymal BioCoord, the developmental BioCoord, and PFS. In this initial state, tumors occupy a broad cloud, with wide dispersion in both biological coordinates and survival.
4B: Intermediate frame of DQC evolution. As the data contract in BioCoord space, the broad cloud organizes into a two-dimensional sheet, revealing a clearer relationship between biological position and PFS.

4C: Late frame of DQC evolution. The data resolve into compact endpoint groups with distinct average PFS values, while preserving residual within-group spread. Across the evolution, the LGG-dominant region forms an extended structure in the mesenchymal–developmental plane, whereas the GBM-rich region is more compact and associated with shorter survival. Together, these plots show that outcome is distributed over a curved biological manifold rather than along a single axis.

Resource availability

Software availability:

https://github.com/zanderteller/dqm

Acknowledgments

The results here are in whole or part based upon data generated by the TCGA Research Network: https://www.cancer.gov/tcga.

Funding provided by start-up funds from the University of Wisconsin-Madison. Research for the study was conducted under the 4C program at Stanford University.

Declaration of Interests

The authors report no conflicts of interest.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work, AI technologies were used to identify areas of the manuscript which were unclear and the authors subsequently rewrote sections of the manuscript to improve clarity.

References

1. https://gtexportal.org/home/

2. https://dbgap.ncbi.nlm.nih.gov/home/

3. Asim MN, Ibrahim MA, Asif T, Dengel A. (2025) RNA sequence analysis landscape: A comprehensive review of task types, databases, datasets, word embedding methods, and language models. *Heliyon* 11(2):e41488. DOI: 10.1016/j.heliyon.2024.e41488

4. The Cancer Genome Atlas Research Network., Weinstein, J., Collisson, E. et al. (2013) The Cancer Genome Atlas Pan-Cancer analysis project. *Nat Genet* 45, 1113–1120. https://doi.org/10.1038/ng.2764

5. GTEx Consortium. Human genomics. (2015) The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science*. 348(6235):648-60. doi: 10.1126/science.1262110.

6. Cui W, Xue H, Wei L, Jin J, Tian X, Wang Q. (2021) High heterogeneity undermines generalization of differential expression results in RNA-Seq analysis. *Hum Genomics*. 15(1):7. doi:10.1186/s40246-021-00308-5.

7. Echle, A., Rindtorff, N.T., Brinker, T.J. et al. (2021) Deep learning in cancer pathology: a new generation of clinical biomarkers. *Br J Cancer* 124: 686–696. https://doi.org/10.1038/s41416-020-01122-x

8. Yu, Y., Mai, Y., Zheng, Y. et al. (2024) Assessing and mitigating batch effects in large-scale omics studies. *Genome Biol* 25: 254. https://doi.org/10.1186/s13059-024-03401-9.

9. Roth, A., Khattra, J., Yap, D. et al. (2014) PyClone: statistical inference of clonal population structure in cancer. *Nat Methods* 11: 396–398. https://doi.org/10.1038/nmeth.2883.

10. Wang Z, Cao S, Morris JS, Ahn J, Liu R, Tyekucheva S, Gao F, Li B, Lu W, Tang X, Wistuba II, Bowden M, Mucci L, Loda M, Parmigiani G, Holmes CC, Wang W. (2018) Transcriptome Deconvolution of Heterogeneous Tumor Samples with Immune Infiltration*. iScience* 9:451-460. doi: 10.1016/j.isci.2018.10.028.

11. Chaudhry, M., Shafi, I., Mahnoor, M., Vargas, D. L. R., Thompson, E. B., & Ashraf, I. (2023). A Systematic Literature Review on Identifying Patterns Using Unsupervised Clustering Algorithms: A Data Mining Perspective. *Symmetry* 15(9): 1679. https://doi.org/10.3390/sym15091679.

12. McInnes et al., (2018). UMAP: Uniform Manifold Approximation and Projection. *Journal of Open Source Software*, 3(29), 861, https://doi.org/10.21105/joss.00861.

13. Wani, A. A. (2025). Comprehensive review of dimensionality reduction algorithms: challenges, limitations, and innovative solutions. *PeerJ Computer Science* 11: e3025. https://doi.org/10.7717/peerj-cs.3025

14. Wang Y., Huang H., Rudin C., Shaposhnik Y. (2021) Understanding how dimension reduction tools work: An empirical approach to deciphering t-SNE, UMAP, TriMap, and PaCMAP for data visualization. *J Machine Learning Res* 22:1-73. https://doi.org/10.5555/3546258.3546459.

15. Weinstein M, Horn D. Dynamic quantum clustering: A method for visual exploration of structures in data. *Phys Rev E Stat Nonlin Soft Matter Phys.* 2009;80(6 Pt 2): Article 066117. https://doi.org/10.1103/PhysRevE.80.066117.

16. Weinstein M., Meirer F., et al. (2013) Analyzing big data with dynamic quantum clustering. *arXiv* preprint arXiv:1310.2700. https://doi.org/10.48550/arXiv.1310.2700.

17. Roche, K.E., Weinstein, M., Dunwoodie, L.J. et al. (2018) Sorting Five Human Tumor Types Reveals Specific Biomarkers and Background Classification Genes. *Sci Rep* 8: 8180. https://doi.org/10.1038/s41598-018-26310-x.

18. Galbraith K, Snuderl M. Molecular Pathology of Gliomas. *Surg Pathol Clin.* 2021 Sep;14(3):379-386. doi: 10.1016/j.path.2021.05.003.

19. Nafe R, Porto L, Samp PF, You SJ, Hattingen E. Adult-type and Pediatric-type Diffuse Gliomas : What the Neuroradiologist Should Know. *Clin Neuroradiol.* 2023 Sep;33(3):611-624. doi: 10.1007/s00062-023-01277-z.

20. Mazumder H, Lin HY, Baddoo M, Gałan W, Polania-Villanueva D, Hicks C, Otohinoyi D, Peruzzi F, Madeja Z, Belancio VP, Flemington EK, Reiss K, Rak M. Human endogenous retroviruses (HERVs) associated with glioblastoma risk and prognosis. *Cancer Gene Ther.* 2025 Jun;32(6):622-632. doi: 10.1038/s41417-024-00868-3.

21. Ma S, Guo Z, Wang B, Yang M, Yuan X, Ji B, Wu Y, Chen S. A Computational Framework to Identify Biomarkers for Glioma Recurrence and Potential Drugs Targeting Them. *Front Genet.* 2022 Jan 17;12:832627. doi: 10.3389/fgene.2021.832627.

22. Ispirjan M, Marx S, Freund E, Fleck SK, Baldauf J, Roessler K, Schroeder HWS, Bekeschus S. Markers of tumor-associated macrophages and microglia exhibit high intratumoral heterogeneity in human glioblastoma tissue. *Oncoimmunology.* 2024 Dec 31;13(1):2425124. doi: 10.1080/2162402X.2024.2425124.

23. Inano, R., Oishi, N., Kunieda, T. et al. Visualization of heterogeneity and regional grading of gliomas by multiple features using magnetic resonance-based clustered images. *Sci Rep* 6, 30344 (2016). https://doi.org/10.1038/srep30344

24. Linninger A, Hartung GA, Liu BP, Mirkov S, Tangen K, Lukas RV, Unruh D, James CD, Sarkaria JN, Horbinski C. Modeling the diffusion of D-2-hydroxyglutarate from IDH1 mutant gliomas in the central nervous system. *Neuro Oncol.* 2018 Aug 2;20(9):1197-1206. doi: 10.1093/neuonc/noy051.

25. Berger, A., Tzarfati, G.G., Serafimova, M. et al. Clinical and prognostic implications of rim restriction following glioma surgery. *Sci Rep* 12, 12874 (2022). https://doi.org/10.1038/s41598-022-16717-y

26. Zhao WJ, Ou GY, Lin WW. Integrative Analysis of Neuregulin Family Members-Related Tumor Microenvironment for Predicting the Prognosis in Gliomas. *Front Immunol.* 2021 May 13;12:682415. doi: 10.3389/fimmu.2021.682415.

27. Miniere HJM, Hormuth DA 2nd, Lima EABF, Farhat M, Panthi B, Langshaw H, Shanker MD, Talpur W, Thrower S, Goldman J, Ty S, Chung C, Yankeelov TE. A data

assimilation framework for predicting the spatiotemporal response of high-grade gliomas to chemoradiation. *BMC Cancer*. 2025 Jul 29;25(1):1239. doi: 10.1186/s12885-025-14557-3.

28. Bähr O, Herrlinger U, Weller M, Steinbach JP. Very late relapses in glioblastoma long-term survivors. *J Neurol.* 2009 Oct;256(10):1756-8. doi: 10.1007/s00415-009-5167-6.

29. Liberzon A, Birger C, Thorvaldsdóttir H, Ghandi M, Mesirov JP, Tamayo P. The Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell Syst.* 2015 Dec 23;1(6):417-425. doi: 10.1016/j.cels.2015.12.004.

30. Brunet J-P. Tamayo P, Golub TR, Mesirov JP. Metagenes and molecular pattern discovery using matrix factorization. *Proc Natl Acad Sci U.S.A.* 2004 Mar 23; 101(12):4164-9. https://doi.org/10.1073/pnas.0308531101

31. Tomfohr J, Lu J, Kepler TB. Pathway level analysis of gene expression using singlular value decomposition. *BMC Bioinformatics* 2005 Sept 12; 225. doi: 10.1186/1471-2105-6-225.

32. Antonino Zito, Xavier Escribà Montagut, Gabriela Scorici, Axel Martinelli, Murodzhon Akhmedov, Ivo Kwee, PLAID: ultrafast single-sample gene set enrichment scoring, *Bioinformatics* Dec 2025; 41(12): btaf621. https://doi.org/10.1093/bioinformatics/btaf621

33. McDermaid A, Monier B, Zhao J, Liu B, Ma Q. Interpretation of differential gene expression results of RNA-seq data: review and integration. *Brief Bioinform*. 2019 Nov 27;20(6):2044-2054. doi: 10.1093/bib/bby067.

Supplemental Information

Supplemental Notes

Supplemental Note S1-1. Intuitive description of Dynamic Quantum Clustering (DQC)

The TCGA glioma analysis revealed that extended structures emerging during DQC evolution, even when applied to noisy high-dimensional expression data, correspond to clinically meaningful patterns such as progression-free survival (PFS) and overall survival (OS). To help readers understand why this occurs, we provide an intuitive explanation of the DQC procedure using a simple two-dimensional example. A full mathematical description of the algorithm is available in the original publications [1,2].

The explanation below avoids mathematical formalism and instead illustrates the geometric intuition underlying the algorithm.

Supplemental Note S1-2. Model dataset used for illustration

To illustrate how DQC reveals hidden structure in complex data, we generated a two-dimensional dataset intended to mimic simplified RNA-seq measurements for two genes. The dataset contains four components:

1. **Three compact Gaussian clusters**, each representing a group of tumors with similar expression values. Differences among points within each cluster arise only from statistical variability.
2. **One extended distribution**, constructed by concatenating many overlapping Gaussian distributions arranged along a vertical trajectory. This extended structure represents a set of tumors whose gene expression changes smoothly along an underlying biological axis (for example, disease progression).

The combined dataset therefore contains both compact clusters and continuous structure, similar to patterns commonly observed in transcriptomic data.

Figure S1 shows the simulated dataset.

Supplemental Note S1-3. Construction of the DQC potential function

DQC begins by constructing a potential function from the data.
This function can be understood as a smooth energy-like landscape defined across the data space.

Importantly:

- Regions containing many nearby samples correspond to lower values of the potential function.
- Regions with few samples correspond to higher values.

In intuitive terms, the potential surface highlights areas where samples accumulate.

To visualize this surface, the potential is plotted as a three-dimensional surface above the two-dimensional data plane. The third dimension does not represent an additional data variable; it simply displays the value of the potential function. For visualization, the hyperbolic tangent of the potential (tanh(V)) is shown to suppress large amplitudes away from the data and emphasize the topology near dense regions.

Figure S2 illustrates this potential surface together with the original data points.

In this representation:

- Compact clusters appear as circular depressions in the landscape.
- Sparse regions may contain multiple small local minima reflecting sampling noise.
- Extended data structures generate elongated valleys in the potential surface.

Supplemental Note S1-4. Evolution of samples during DQC

After constructing the potential surface, DQC evolves the positions of the data points according to a rule derived from quantum mechanical dynamics (the Hamiltonian). In practice, this rule simply determines how each point moves across the potential landscape during the algorithm.

An intuitive way to understand this process is to imagine each data point moving across the landscape while being influenced not only by its immediate surroundings but also by the broader structure of the landscape.

Two properties distinguish this evolution from classical gradient descent:

1. Non-local influence

Each point responds to a weighted average of the potential in its neighborhood rather than only the local slope at a single location. As a result, motion reflects the broader geometry of the data rather than small fluctuations in the surface.

2. Reduced sensitivity to small local minima

Because the evolution averages information across nearby regions, very small local depressions in the potential become effectively invisible. This prevents the algorithm from fragmenting data into many artificial micro-clusters, a common issue in high-dimensional datasets.

Together, these properties allow DQC to reveal large-scale structures such as extended trajectories or major clusters that reflect the intrinsic organization of the data.

Supplemental Note S1-5. Example of DQC evolution

Figure S3 shows successive frames from a DQC animation applied to the model dataset. The red points represent the data samples evolving along the potential surface.

Key observations include:

- Initially separated clusters move toward stable positions within their respective valleys.
- The extended distribution gradually contracts along its principal axis.
- Small local irregularities in the potential do not trap points because the evolution responds to the broader landscape.

By later frames, the extended distribution collapses into a narrow linear structure representing a single dominant underlying axis in the data. This example illustrates how DQC can reveal hidden low-dimensional structure embedded within complex datasets. The full animation can be seen in this file (**DQC_Evolution_Dynamics**).

Supplementary Figures

Figure S1. Simulated dataset used to illustrate DQC behavior.

Two-dimensional model dataset composed of three compact Gaussian clusters and one

extended distribution generated from overlapping Gaussian components arranged along a vertical trajectory. The simulated data are used to illustrate how DQC distinguishes compact groupings from continuous structure.

Figure S2. Potential function generated from the model dataset. The DQC potential function is visualized as a three-dimensional surface above the two-dimensional data plane. Lower regions correspond to areas containing higher densities of samples. Data points are shown in red.

Figure S3. Evolution of the model dataset during DQC.
Panels S3-1 to S3-8) Selected frames from the DQC animation showing how samples move across the potential surface over the course of the algorithm. Compact clusters converge to stable basins, whereas the extended distribution contracts into a lower-dimensional structure that reflects the dominant organization of the data.

Figure S4. DQC evolution of the 554-gene panel from diffuse cloud to two final clusters.
Representative frames from Dynamic Quantum Clustering (DQC) applied to the 692×554 expression matrix, displayed in the first three principal components for visualization only. **Left**, frame 1: the initial diffuse cloud shows separation of the two dominant sample groupings. **Middle**, frame 11: the cloud contracts onto a lower-dimensional manifold while preserving the global LGG–GBM organization. **Right**, final frame: the data resolve into two compact endpoint clusters corresponding to the LGG-enriched and GBM-enriched states identified in the main analysis. Histology labels were overlaid only after clustering for post hoc evaluation.

Figure S5. Pathway enrichment highlights ECM remodeling, inflammation, and invasion programs distinguishing DQC-defined clusters. Ingenuity Pathway Analysis (IPA) of differentially expressed genes between the DQC-defined GBM-enriched and LGG-enriched clusters (FDR < 0.05) identified significant enrichment of extracellular matrix (ECM) organization and degradation pathways, including matrix metalloprotease activity and collagen-related processes. Additional enriched pathways include leukocyte adhesion and diapedesis, vascular transmigration, PI3K-associated motility, and pro-inflammatory signaling (e.g., IL-17, NF-κB, and PI3-family pathways). Pathways are ranked by Benjamini–Hochberg–adjusted p values. These results provide biological context for the unsupervised DQC separation, indicating that the GBM-enriched cluster is associated with ECM remodeling, invasion, and inflammatory signaling, whereas the LGG-enriched cluster retains more neuronal and developmental features.

Figure S6. Compression of the 90-gene DQC structure into biological coordinates for risk stratification DQC-based clustering using the 90-gene signature. An unsupervised reduction of the 554-gene set, driven by rank-ordering of genes by between-cluster contrast, yielded a 90-gene panel. Dynamic Quantum Clustering (DQC) resolved tumors into four clusters (A–D) without the use of clinical labels. While the

GBM-enriched Cluster B notably shows five LGG samples that cluster biologically with the GBM group, suggesting an aggressive molecular phenotype. The clusters occupy an elongated trajectory (A→D→C→B) consistent with a dominant mesenchymal/invasion-associated gradient.

Supplementary Tables

Table S1. Mean normalized expression of 20,057 protein-coding genes across DQC-predicted LGG and GBM groups from the initial whole-transcriptome analysis.

Table S2. Ranked genes between the Step-1 DQC GBM-enriched and LGG-enriched clusters defining the 554-gene panel.

Table S3. IPA-curated annotations for genes ranked by log2 fold change between the Step-1 DQC GBM-enriched and LGG-enriched clusters within the 554-gene panel.

Table S4. IPA canonical pathways ranked by Benjamini–Hochberg adjusted p values for genes distinguishing the Step-1 DQC GBM-enriched and LGG-enriched clusters.

Table S5. DQC cluster membership and survival summaries for top-k gene panels
Notes: Clusters are derived by DQC without labels. TCGA histology and PFS/OS were applied post hoc for evaluation only. PFS/OS values are reported as mean days.

The rank-ordered gene panels preserve DQC structure and refine LGG prognostic substructure. To assess whether a reduced gene set could preserve diagnostic separation while revealing finer biological structure, genes were ranked by two-sample t-test p-values between the two DQC-defined clusters (used for ordering only), and DQC was evaluated across progressively smaller top-k subsets (140, 100, and 90 genes; **Table S5**).

At 140 genes, DQC resolved three clusters, including two LGG-only groups with distinct survival and one GBM-enriched group. At 100 genes, the geometry refined to four clusters, with three LGG-only groups and one GBM-enriched cluster, indicating improved diagnostic separation. At 90 genes, this four-cluster structure and its associated survival ordering were preserved, supporting a stable low-dimensional representation of the data.

Together, these results indicate that progressive feature reduction denoises the dataset while preserving the DQC manifold, maintaining robust LGG–GBM separation, and

exposing reproducible prognostic substructure within LGG. Performance deteriorated below ~90 genes, suggesting a stable “sweet spot” near 90–100 genes.

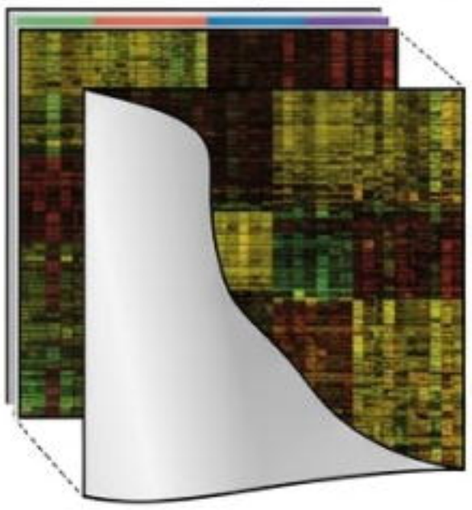
Supplemental references

1. Weinstein M, Horn D. Dynamic quantum clustering: A method for visual exploration of structures in data. *Phys Rev E Stat Nonlin Soft Matter Phys*. 2009;80(6 Pt 2): Article 066117. https://doi.org/10.1103/PhysRevE.80.066117.

2. Weinstein M., Meirer F., *et al.* (2013) Analyzing big data with dynamic quantum clustering. arXiv preprint arXiv:1310.2700. https://doi.org/10.48550/arXiv.1310.2700.

UNSUPERVISED 90-GENE DQC MAPS A CONTINUOUS TRAJECTORY OF MESENCHYMAL PROGRESSION IN GLIOMA FOR RISK STRATIFICATION

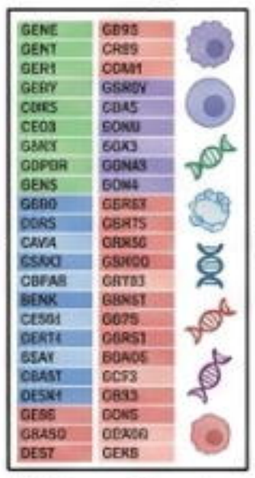
20,057-Gene Expression Matrix



Sieve

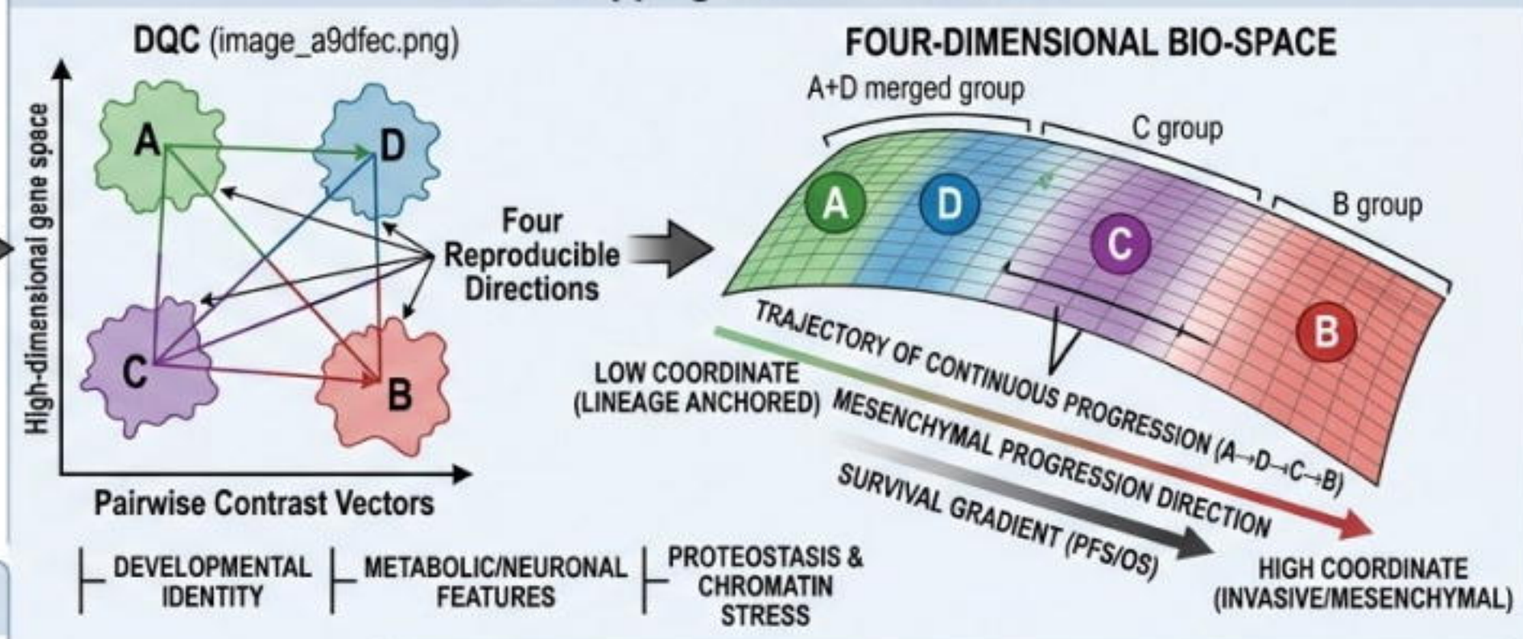
t-test ranked reduction

90-Gene Signature

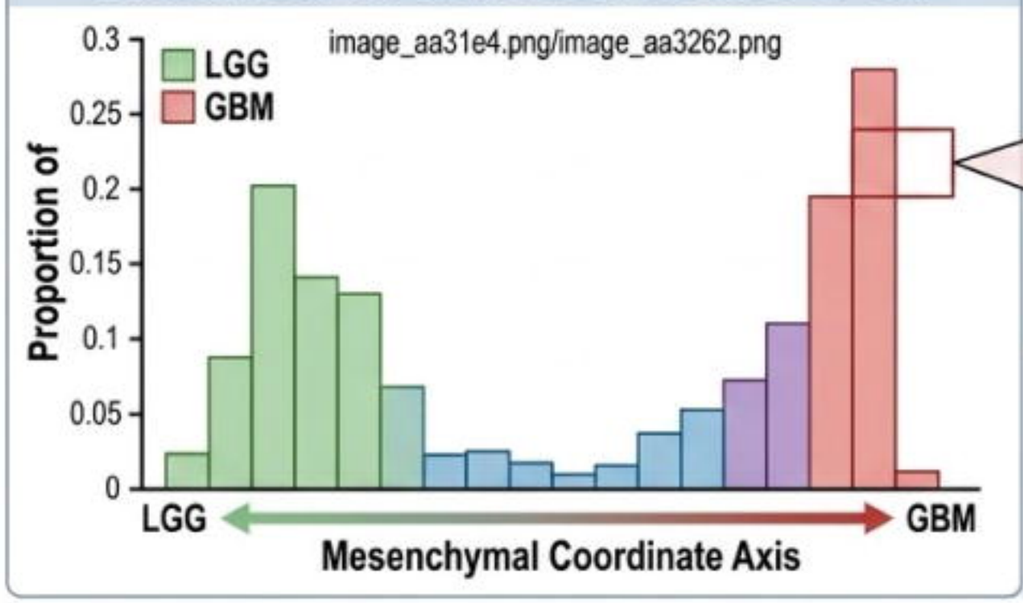


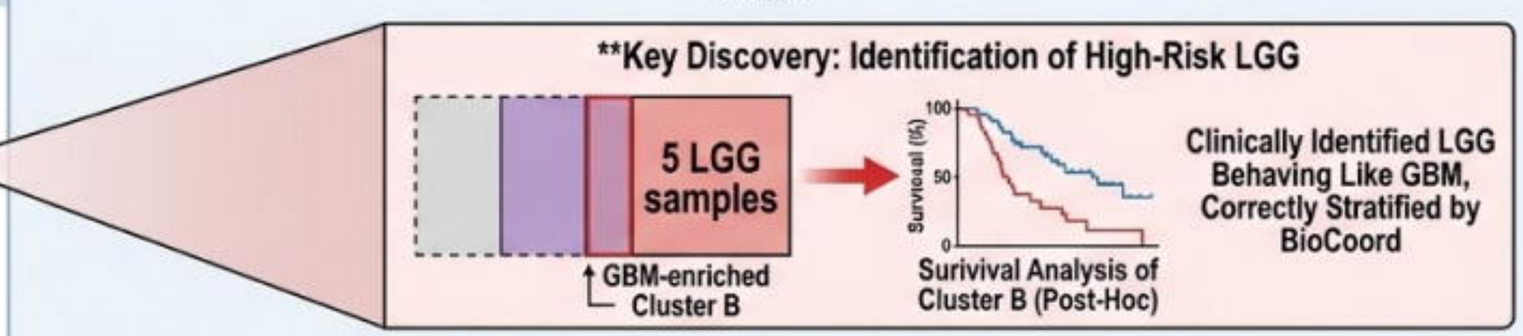
90-Gene Signature

DQC Mapping & Axis Construction



CLINICAL UTILITY & THE GRADE-BIOLOGY GAP





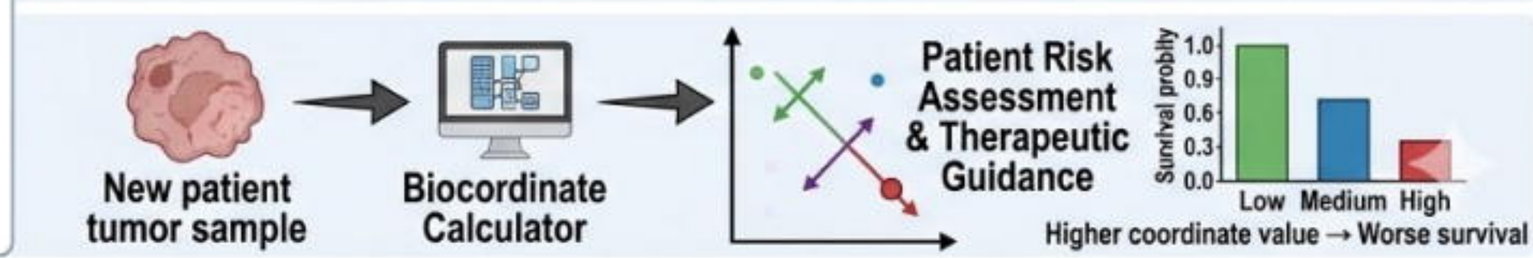
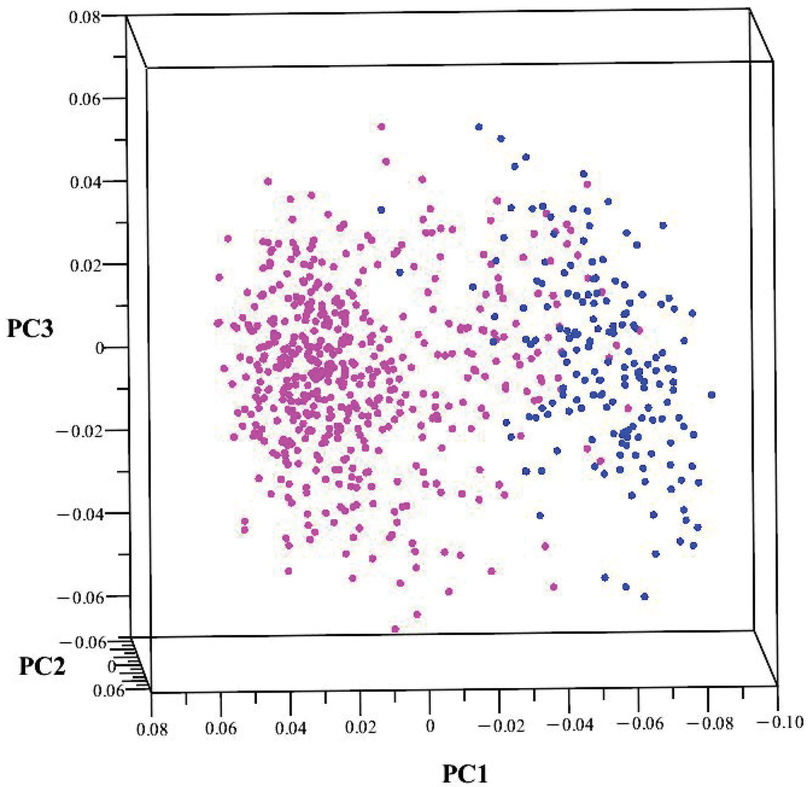
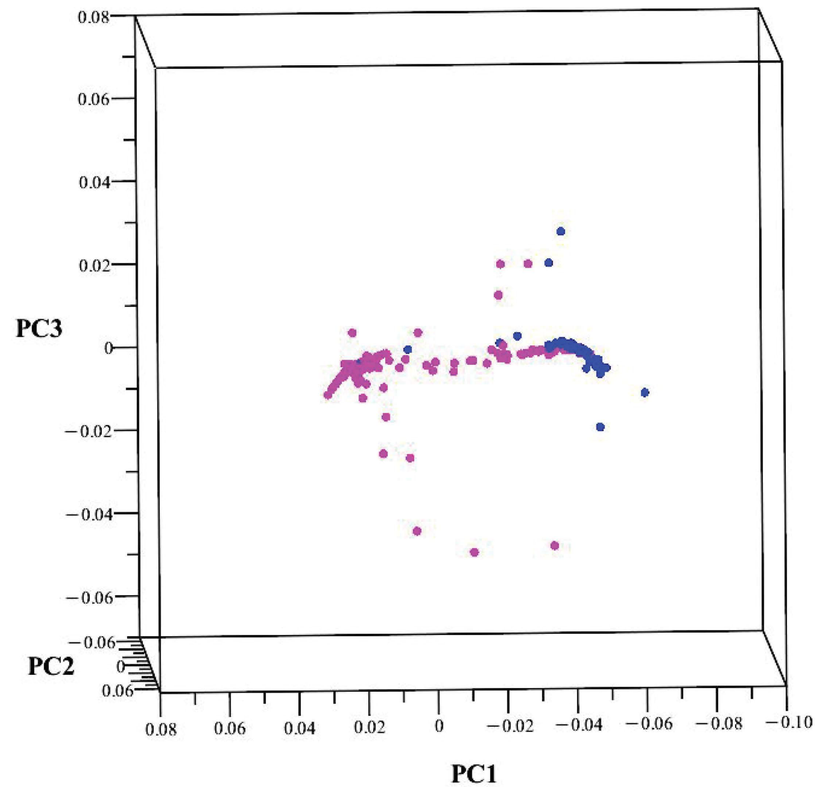


Figure 1

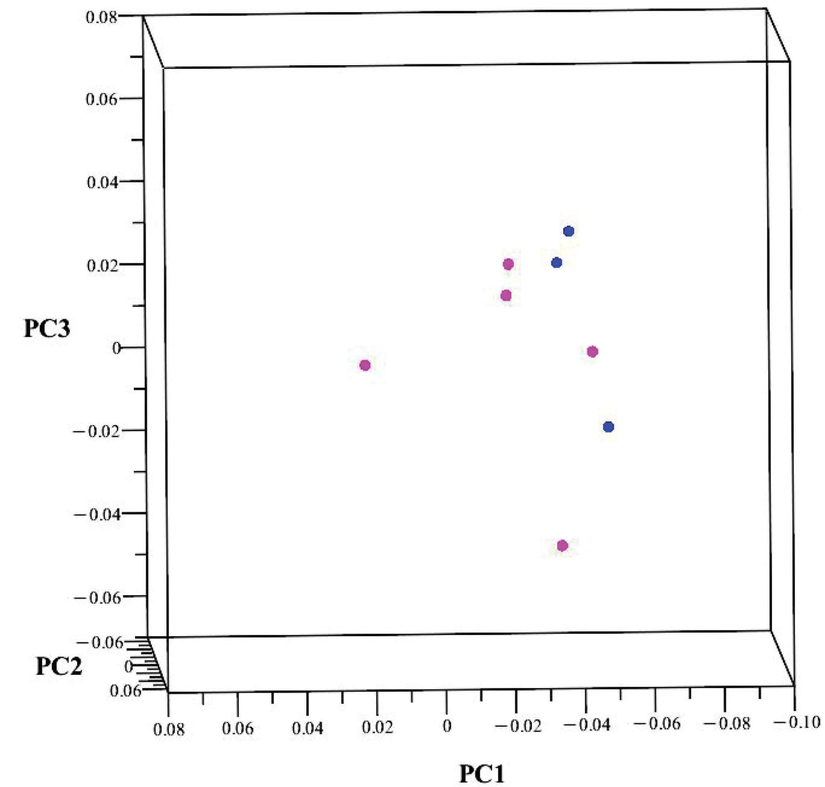
A.



B.



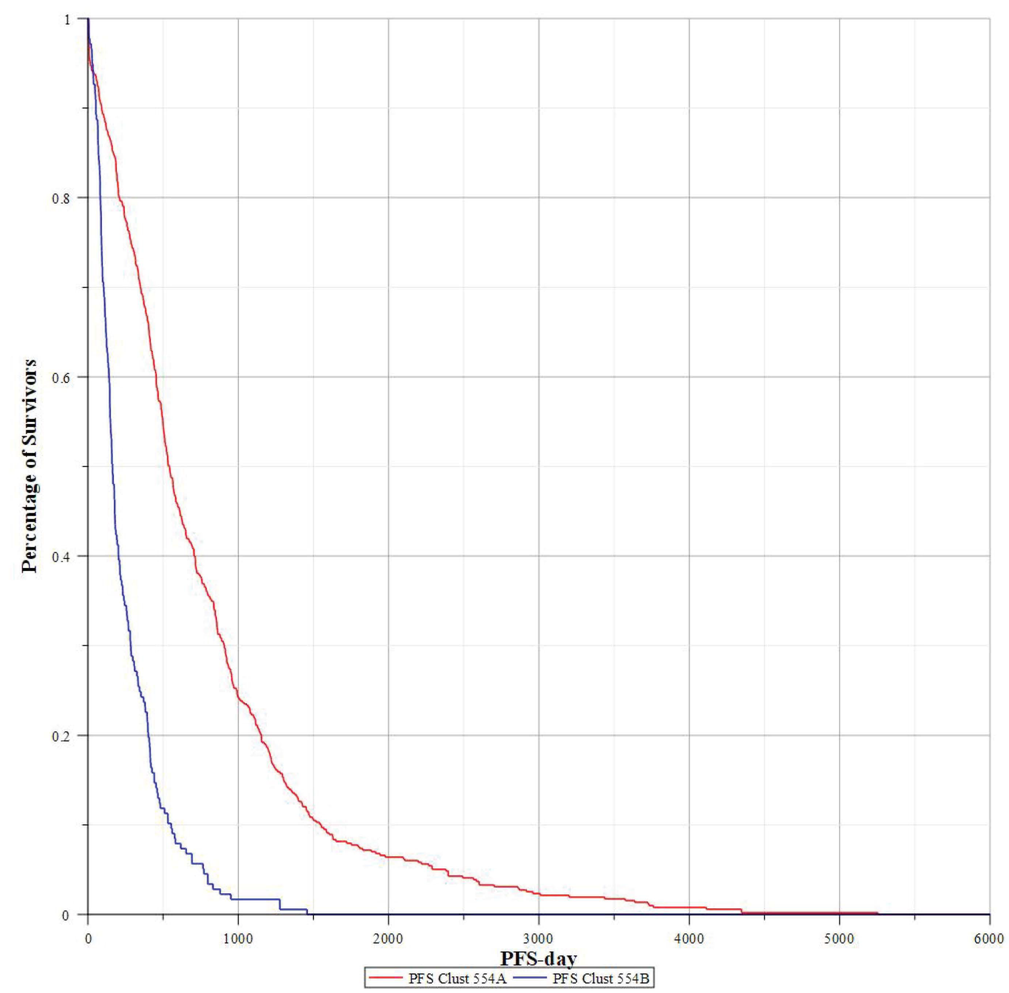
C.



D.

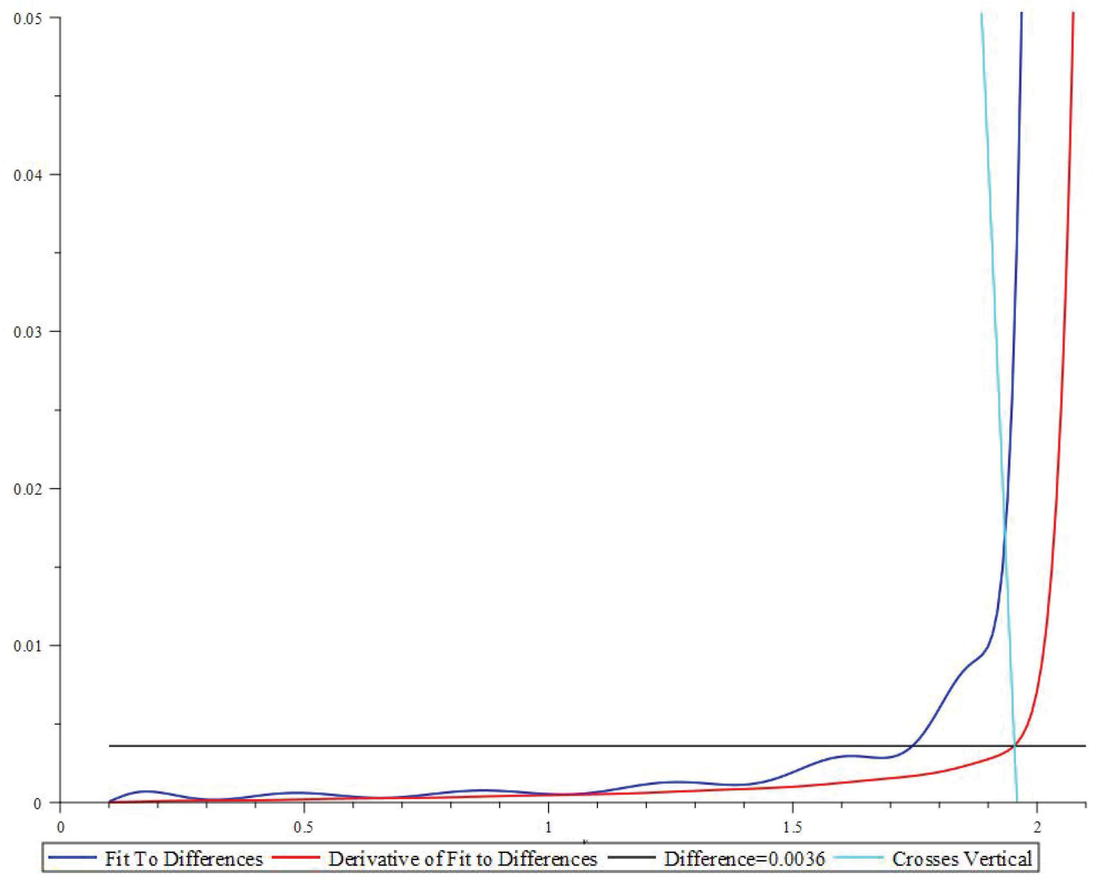
	Pred. LGG	Pred. GMB	Isolated
True LGG	469	52	3
True GBM	4	161	3

E.



A.

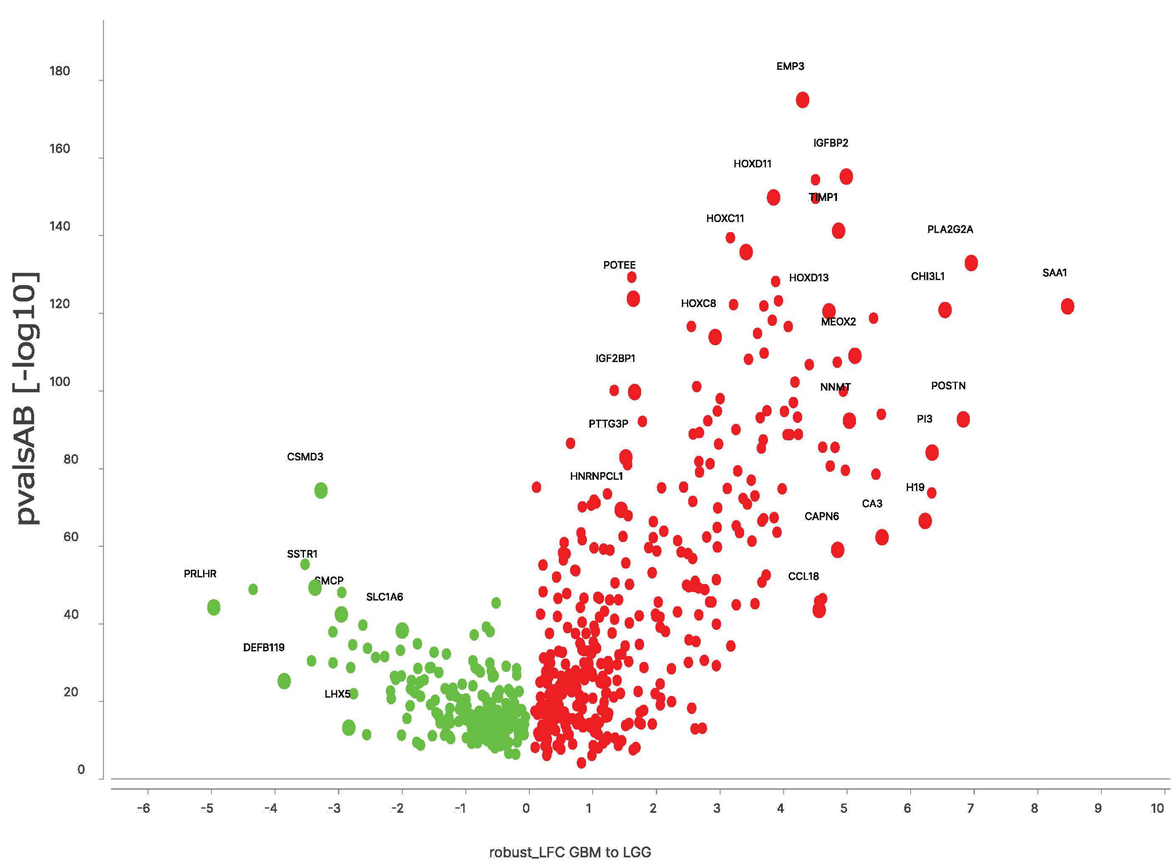
Figure 2



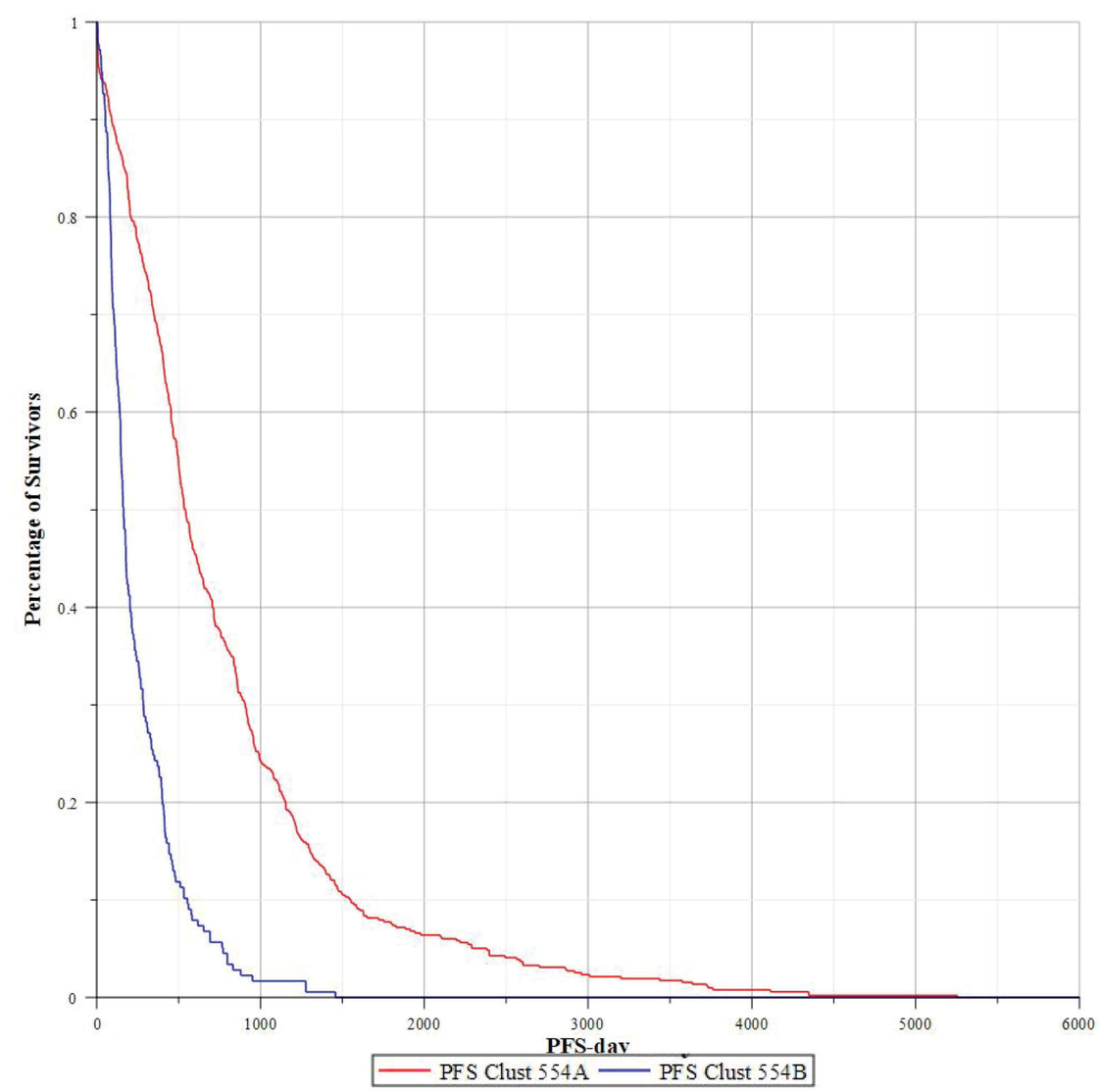
B.

	Pred. LGG	Pred. GMB
True LGG	510	14
True GBM	5	163

C.

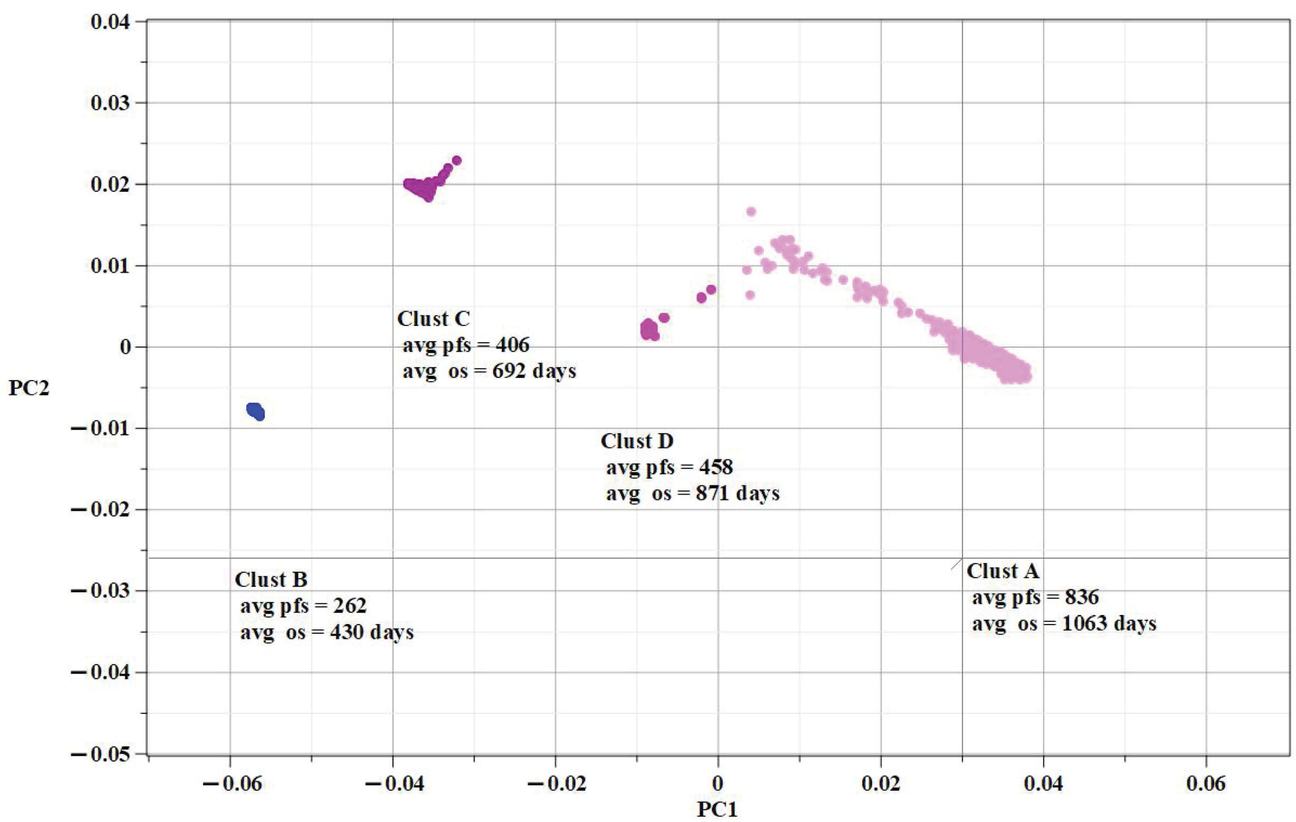


D.

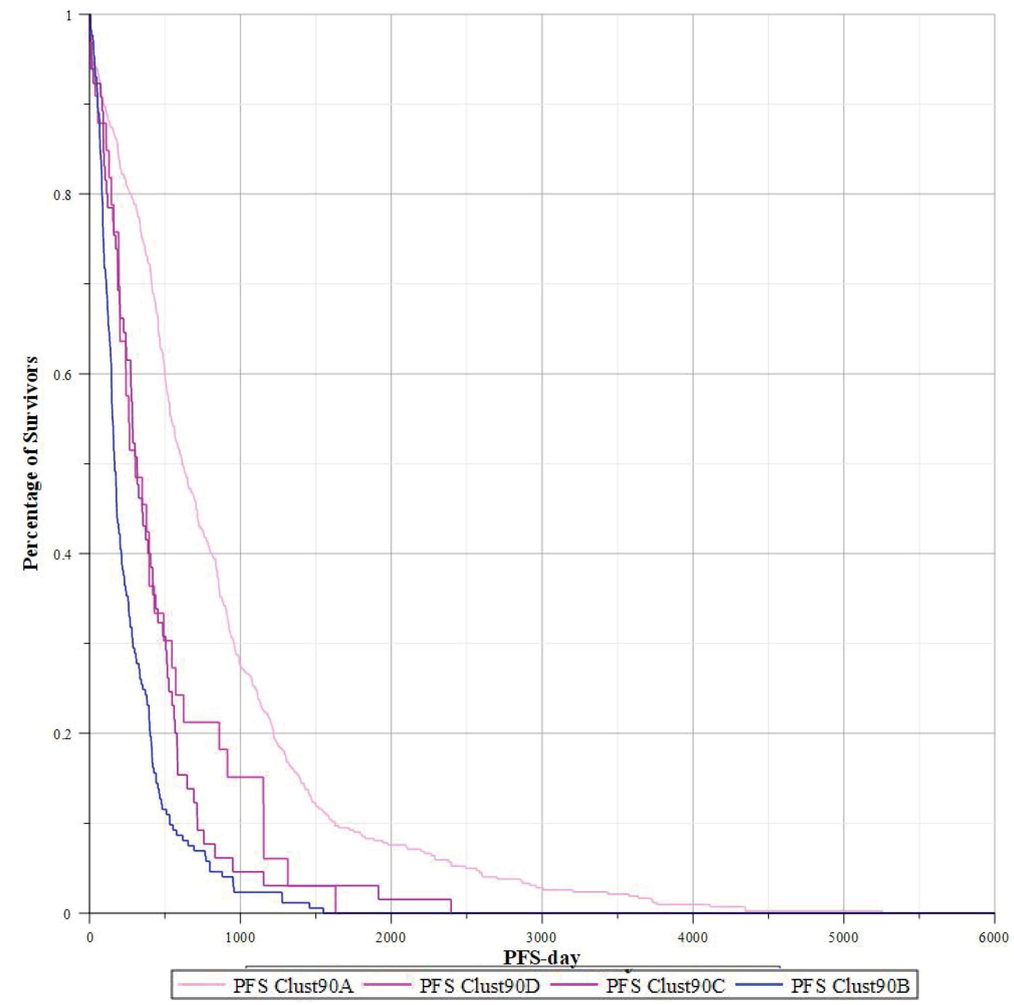


A.

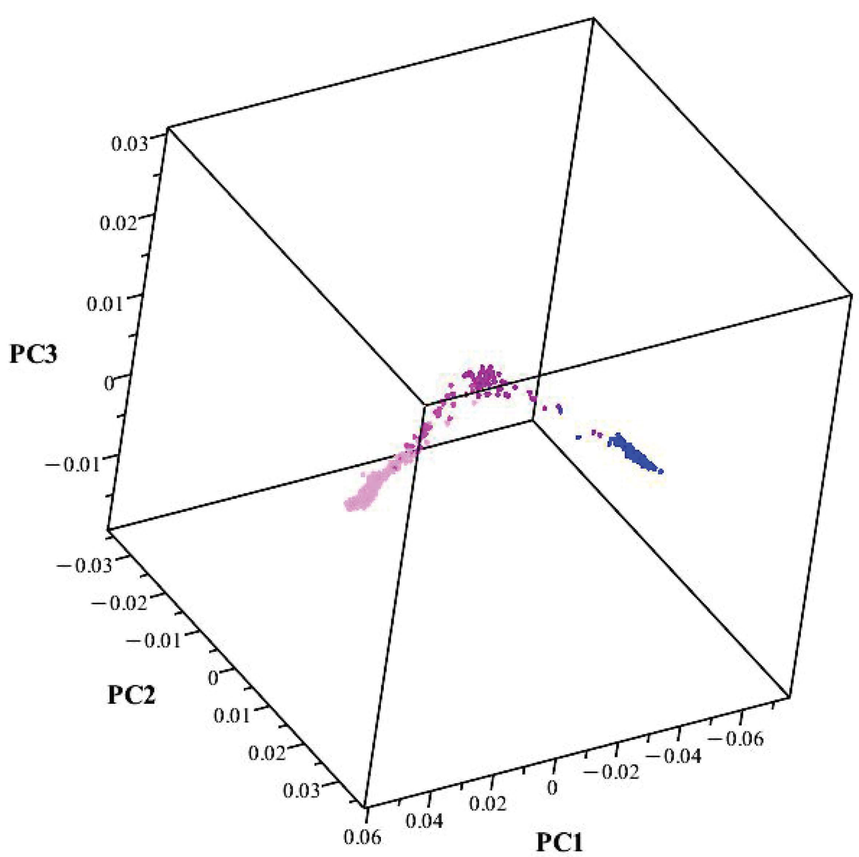
Figure 3



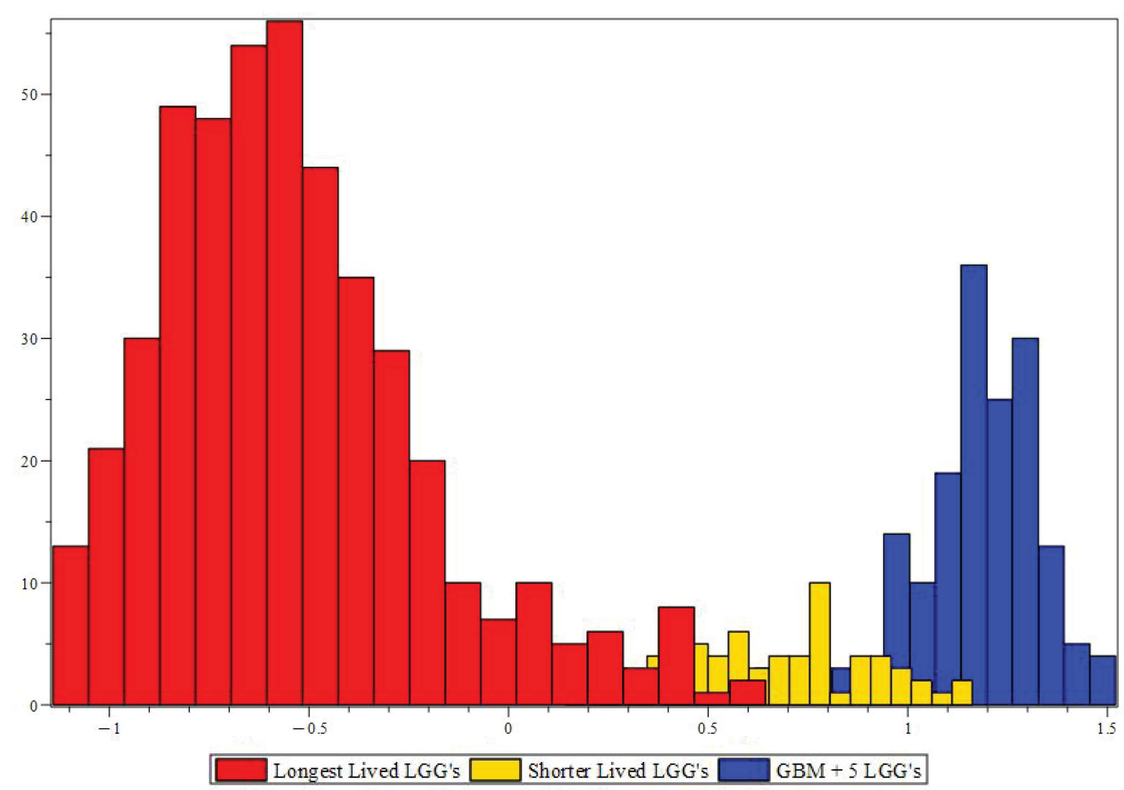
B.



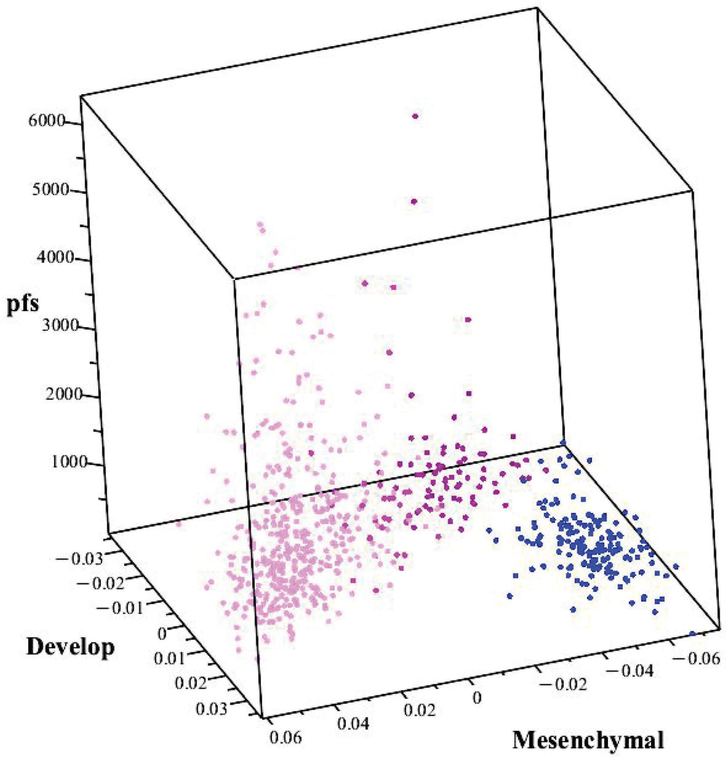
C.



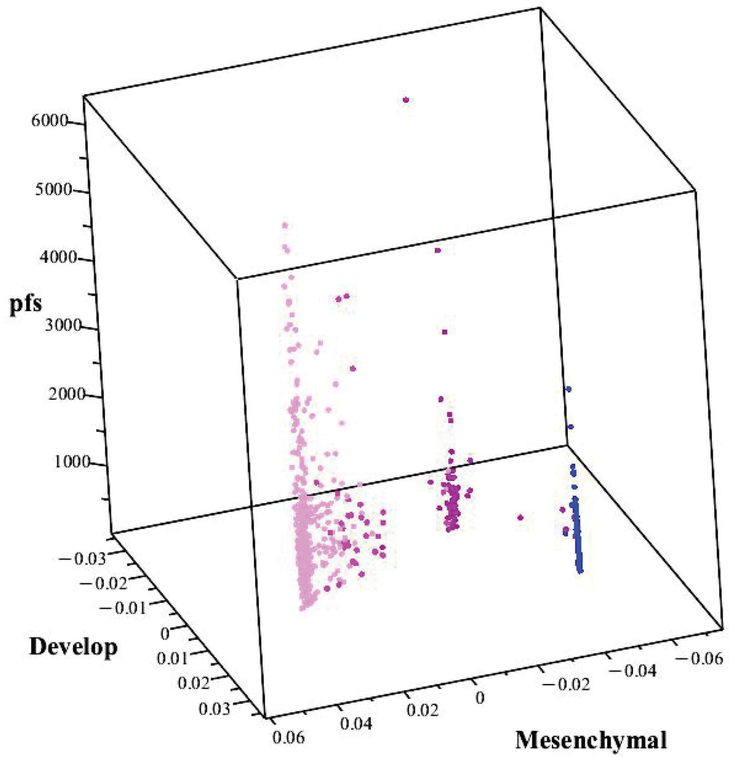
D.



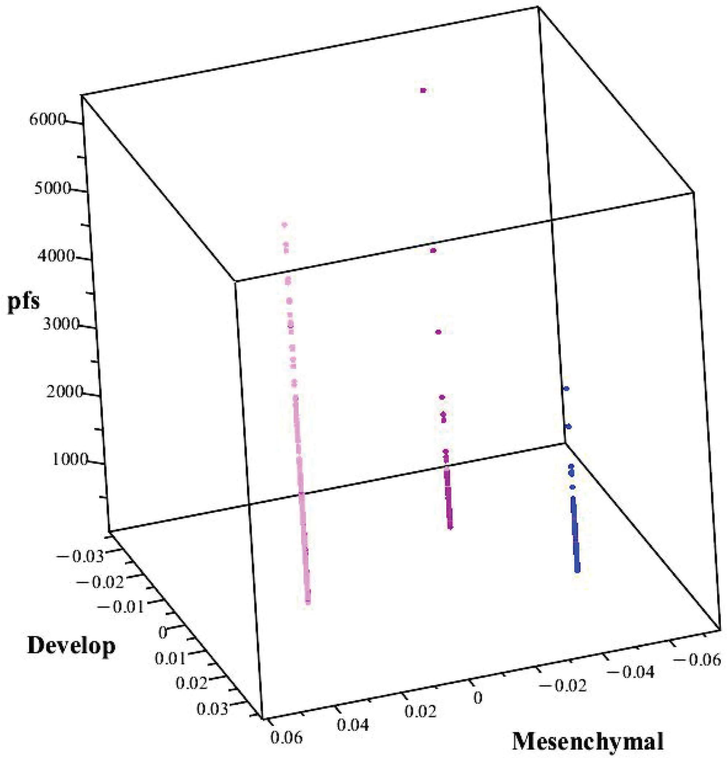
A. Figure 4


6000
5000
4000
pfs
3000
2000
1000
−0.03
−0.02
−0.01
0
0.01
0.02
0.03
Develop
0.06
0.04
0.02
0
−0.02
−0.04
−0.06
Mesenchymal

B.


6000
5000
4000
pfs
3000
2000
1000
−0.03
−0.02
−0.01
0
0.01
0.02
0.03
Develop
0.06
0.04
0.02
0
−0.02
−0.04
−0.06
Mesenchymal

C.


6000
5000
4000
pfs
3000
2000
1000
−0.03
−0.02
−0.01
0
0.01
0.02
0.03
Develop
0.06
0.04
0.02
0
−0.02
−0.04
−0.06
Mesenchymal