JMIR Publications
The Leading eHealth Publisher

Unsupervised multiple correspondence analysis is a relevant tool for investigating associations between prognostic factors in gliomas

Maria Eduarda Goes Job, Heidge Fukumasu, Tathiane Maistro Malta, Pedro Luiz Porfirio Xavier

Submitted to: Interactive Journal of Medical Research
on: August 21, 2024

Disclaimer: **© The authors. All rights reserved.** This is a privileged document currently under peer-review/community review. Authors have provided JMIR Publications with an exclusive license to publish this preprint on it's website for review purposes only. While the final peer-reviewed paper may be licensed under a CC BY license on publication, at this stage authors and publisher expressively prohibit redistribution of this draft paper other than for review purposes.

Table of Contents

Unsupervised multiple correspondence analysis is a relevant tool for investigating associations between prognostic factors in gliomas

Maria Eduarda Goes Job[1]; Heidge Fukumasu[1]; Tathiane Maistro Malta[2]; Pedro Luiz Porfirio Xavier[1]

[1]Laboratory of Comparative and Translational Oncology (LOCT) Department of Veterinary Medicine School of Animal Science and Food Engineering, University of Sao Paulo Pirassununga BR
[2]Cancer Epigenomics Laboratory Department of Clinical Analysis, Toxicology and Food Sciences School of Pharmaceutical Sciences of Ribeirao Preto, University of Sao Paulo Ribeirão Preto BR

Corresponding Author:
Pedro Luiz Porfirio Xavier
Laboratory of Comparative and Translational Oncology (LOCT)
Department of Veterinary Medicine
School of Animal Science and Food Engineering, University of Sao Paulo
Avenida Duque de Caxias, 225
Jardim Elite
Pirassununga
BR

Abstract

Background: Multiple Correspondence Analysis (MCA) is an unsupervised data science methodology that aims to identify and represent associations between categorical variables. Gliomas are an aggressive type of cancer characterized by diverse molecular and clinical features that serve as key prognostic factors. Thus, advanced computational approaches are essential to enhance analysis and interpretation of the associations between clinical and molecular features in gliomas.

Objective: This study aims to apply MCA to identify associations between glioma prognostic factors and also explore their associations with stemness phenotype.

Methods: Clinical and molecular data from 448 brain tumor patients were obtained from The Cancer Genome Atlas (TCGA). The mDNA stemness index, derived from DNA methylation patterns, was built using a one-class logistic regression (OCLR). Associations between variables were evaluated using the chi-square test with k degrees of freedom, followed by analysis of the adjusted standardized residuals. MCA was employed to uncover associations between glioma prognostic factors and stemness.

Results: Our analysis revealed significant associations among molecular and clinical characteristics in gliomas. Additionally, we demonstrated the capability of MCA to identify associations between stemness and these prognostic factors. Our results exhibited a strong association between higher mDNA stemness index and features related to poorer prognosis, demonstrating the utility of MCA as an analytical tool for elucidating potential prognostic factors.

Conclusions: MCA proves to be a valuable tool for understanding the complex interdependence of prognostic markers in gliomas. MCA facilitates the exploration of large-scale datasets and enhances the identification of significant associations.

(JMIR Preprints 21/08/2024:65645)
DOI: https://doi.org/10.2196/preprints.65645

Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v
Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

Original Manuscript

Unsupervised Multiple Correspondence Analysis is a relevant tool for investigating associations between prognostic factors in Gliomas

Maria Eduarda G. Job[1], Heidge Fukumasu[1], Tathiane M. Malta[2], Pedro Luiz P. Xavier[1*]

[1] Laboratory of Comparative and Translational Oncology (LOCT), Department of Veterinary Medicine, School of Animal Science and Food Engineering, University of Sao Paulo, Pirassununga, Brazil

[2] Cancer Epigenomics Laboratory, Department of Clinical Analysis, Toxicology and Food Sciences, School of Pharmaceutical Sciences of Ribeirao Preto, University of Sao Paulo, Ribeirao Preto, Brazil.

* Correspondence:
Pedro Luiz P. Xavier
porfirioxavier@usp.br

Abstract

Background:

Multiple Correspondence Analysis (MCA) is an unsupervised data science methodology that aims to identify and represent associations between categorical variables. Gliomas are an aggressive type of cancer characterized by diverse molecular and clinical features that serve as key prognostic factors. Thus, advanced computational approaches are essential to enhance analysis and interpretation of the associations between clinical and molecular features in gliomas.

Objective:

This study aims to apply MCA to identify associations between glioma prognostic factors and also explore their associations with stemness phenotype.

Methods:

Clinical and molecular data from 448 brain tumor patients were obtained from The Cancer Genome Atlas (TCGA). The mDNA stemness index, derived from DNA methylation patterns, was built using a one-class logistic regression (OCLR). Associations between variables were evaluated using the chi-square test with k degrees of freedom, followed by analysis of the adjusted standardized residuals. MCA was employed to uncover associations between glioma prognostic factors and stemness.

Results:

Our analysis revealed significant associations among molecular and clinical characteristics in gliomas. Additionally, we demonstrated the capability of MCA to identify associations between stemness and these prognostic factors. Our results exhibited a strong association between higher mDNA stemness index and features related to poorer prognosis, demonstrating the utility of MCA as an analytical tool for elucidating potential prognostic factors.

Conclusions:

MCA proves to be a valuable tool for understanding the complex interdependence of prognostic markers in gliomas. MCA facilitates the exploration of large-scale datasets and enhances the identification of significant associations.

Keywords: Brain tumors, Bioinformatics, Stemness, Correspondence Analysis

Introduction

Cancer is a dynamic and heterogeneous disease characterized by several hallmarks controlling and contributing to its development and progression [1]. Cancer research continually generates large scales of data encompassing clinical information, genomic and transcriptomic profiles, prognostic and diagnostic markers, and therapeutic targets [2]. To manage this complexity, different approaches have been employed to study and associate all these variables, aiming to reduce the dimensionality and enhance data interpretation and decision-making process. Several features used to study and classify the different types of cancer are based on categorical variables. For instance, the most widely used cancer staging system, TNM, is based on categorical variables, where “T” refers to the size of the primary tumor, “N” refers to the number of lymph nodes affected by cancer, and “M” refers to absence or presence of metastasis [3]. Thus, these biological and clinical variables interact, and their associations can be measured and diagnosticated using statistical tests such as Fisher’s exact test and Chi-square tests. However, these approaches could not provide a global and comprehensive picture of the associations between these variables, particularly in datasets with a large number of categorical variables. Therefore, employing multivariate and visual analysis methods can significantly improve the analysis and interpretation of associations between clinical and molecular cancer phenotypes.

Brain tumors are a particularly aggressive type of cancer, mostly due to local tissue damage and highly invasive growth. Gliomas, which originate from neuroglial stem cells or progenitor cells, account for 30% of primary brain tumors and 80% of malignant brain tumors [4]. This heterogeneous disease is histologically classified based on anaplasia criteria and predominant cell type such as oligodendroglioma, astrocytoma, glioblastoma (GBM) [5]. Nevertheless, as further investigation aimed to elucidate the neuropathological mechanisms of gliomas, new variables are considered for characterizing this cancer tumor, leading to reclassifications based on mutational profiles, clinical data, and epigenetic factors [6]. This scenario resulted in different prognosis predictions, diagnosis determination, and treatment responses, contributing to an increasingly complex and stratified understanding of gliomas.

Stemness is a key phenotype of cancer stem cells, related to tumor initiation and progression, therapy resistance, and metastasis [7]. Cancer stem cells (CSCs) are referred to as a subpopulation of tumor cells able to self-renew and differentiate into distinct cell lineages, enabling those cells to adapt to different environmental situations [8]. Moreover, recent studies have demonstrated associations between stemness features and different histologic classifications or prognostic factors of gliomas [9], [10], [11]. Therefore, providing a comprehensive visualization of the associations between clinical features and stemness in brain tumors could be valuable for identifying and determining potential prognostic and therapeutic markers.

Multiple Correspondence Analysis (MCA) is an unsupervised data science methodology that aims to observe and represent associations between variables disposed in contingency tables, visualizing these associations in a two-dimensional perceptual map. This approach allows for the simultaneous visualization of the relationship between two or more characteristics [12]. MCA shares general characteristics and it is an extension of Principal Component Analysis (PCA) which is effective in reducing data dimensionality. Thus, MCA can significantly reduce the workload and simplify statistical analysis in healthy research [13]. The results of MCA are typically interpreted in a two-dimensional map, where the relative positions of categories of each variable and their distribution along the dimensions are analyzed. Categories that cluster together and are closer are more likely to be associated, providing key insights into the relationship [14]. Despite its applicability, rigor, and success in other disciplines such as Geography, Epidemiology, and Human Physiology, MCA remains underutilized in Oncology research and few studies are applying [12], [14], [15], [16].

Thus, the main contribution of this study is to highlight Multiple Correspondence Analysis (MCA) as a powerful tool for overcoming the barrier of representing the heterogeneity and complexity of cancer variables, particularly in glioma. By employing MCA, we aimed to gain a deeper understanding of the interdependence between Stemness and prognostic factors. Our findings revealed associations among molecular and clinical characteristics and prognostic factors, as previously described by the literature [17], [18]. Additionally, we demonstrated the capability of MCA to identify associations between stemness and these prognostic factors. Our results exhibited a strong association between higher stemness index and features related to poorer prognosis, demonstrating the utility of MCA as an analytical tool for elucidating oncological heterogeneity and may also offers a valuable strategy for therapeutic decision-making.

Methods

Dataset of the tumor samples

Clinical and molecular information of a total of 448 brain tumor patients were obtained from The Cancer Genome Atlas (TCGA). We tailored the dataset to contain only qualitative information, with 12 variables: cancer type, histology, grade , patient’s vital status, IDH (isocitrate dehydrogenase) status, co-deletion of chromosomes 1p and 19q arms, MGMT (Methylguanine methyltransferase) gene methylation, TERT expression, gain of chromosome 19 and 20, chromosome 7 gain and chromosome 10 loss, ATRX (Alpha Thalassemia/Mental Retardation Syndrome X-linked) status, and glioblastoma transcriptome subtypes.

mDNA Stemness index

The mDNAsi based on DNA methylation was built using a one-class logistic regression (OCLR) [19] on the pluripotent stem cell samples (ESC and iPSC) from the PCBC dataset [20], [21]. The algorithm was built and validated as described in the original paper [22]. The mDNAsi was applied in 381 samples from the TCGA database. Malta's model presented a high correlation among other cancer stem cell signatures, providing significant insights into the biological and clinical features of pan-cancer. The workflow to generate the mDNAsi is available at [22].

Multiple Correspondence analysis

Multiple correspondence analyses were conducted in the RStudio 4.3.1 environment using the packages FactoMineR [23] and cabootcrs for creating matrices for MCA analyses. Contingency tables for the categorical variables were generated, and associations between variables were assessed using a chi-square test with k degrees of freedom. This was followed by the analysis of the adjusted standardized residuals. The Chi-squared test evaluates whether the observed associations between categorical variables are randomly associated ($p < 0.05$). Adjusted standardized residuals higher than 1.96 indicate a significant association between variables in the matrix. To perform MCA, the categorical variables should not be randomly associated. To create the perceptual map, inertia was determined as the total chi-square divided by the number of samples, resulting in the number of associations in the dataset. MCA was performed based on the binary matrices and row and column profiles were determined to demonstrate the influence of each category of variables on the others. Matrices were defined based on the row and column profiles. Eigenvalues were then extracted to represent the number of dimensions that could be captured in the analysis. Finally, the x- and y-axis coordinates of the perceptual map were determined, allowing the category of the variables to be represented and established. In MCA, the spatial distance between categories of different variables reflects their associations. Categories with high coordinates that are close in space are directly associated, while categories presenting high coordinates but opposing coordinates are inversely associated.

Results

MCA can identify associations between different variables of gliomas and patient vital status

To determine the suitability of glioma variables for Multiple Correspondence Analysis (MCA), we first evaluated whether categorical glioma variables were randomly or non-randomly associated. This involved creating individual contingency tables for each pair of glioma variables (*Supplementary Tables 1-13*). Then, we applied chi-squared tests for each contingency table to confirm non-random associations (*p-value < 0.05*). Based on the Chi-squared test, the results indicated that only two categorical variables, Gender and DAXX expression, were randomly associated, suggesting no significant association patterns between these variables and the others. Consequently, Gender and DAXX expression were excluded from further analysis.

In the subsequent analysis, we observed and measured the strength of associations between the patient vital status (0 – alive; 1 – dead) and different factors including cancer type, histology, grade, IDH status, 1p19q codeletion, MGMT promoter methylation, gain of Chr7 and loss of Chr10 (7+/10-), co-gain of Chr 19 and 20 (19+/20+), TERT expression, ATRX Status, and transcriptome subtype, aiming to determine whether MCA could identify associations between prognostic factors for this disease. We used adjusted standardized residuals (ASR) to assess these associations, considering a category of each variable to be associated with either alive or dead vital status when the ASR values were higher than 1.96. Patients' vital status classified as dead were associated with poorer prognostics factors such as glioblastomas, grade 4, IDH wild type, non-codeleted 1p19q, unmethylated MGMT promoter, gain of Chr7 and loss of Chr10, expression of TERT, ATRX wild type, and classical and mesenchymal transcriptome subtypes (*Table 1*). In contrast, patients classified as alive were linked to favorable prognostic variables, including oligoastrocytomas and oligodendrogliomas, grade 2, IDH mutant, codeleted 1p19q, methylated MGMT promoter, absence of combined Chr7+/Chr10-, lack of TERT expression, ATRX mutant, and the proneural and neural transcriptome subtypes (*Table 1*). Histological classification, grade, IDH status, and chr7+/chr10- were the most strongly associated features with patient vital status. These associations were further illustrated in a heatmap (*Figure 1A-D*).

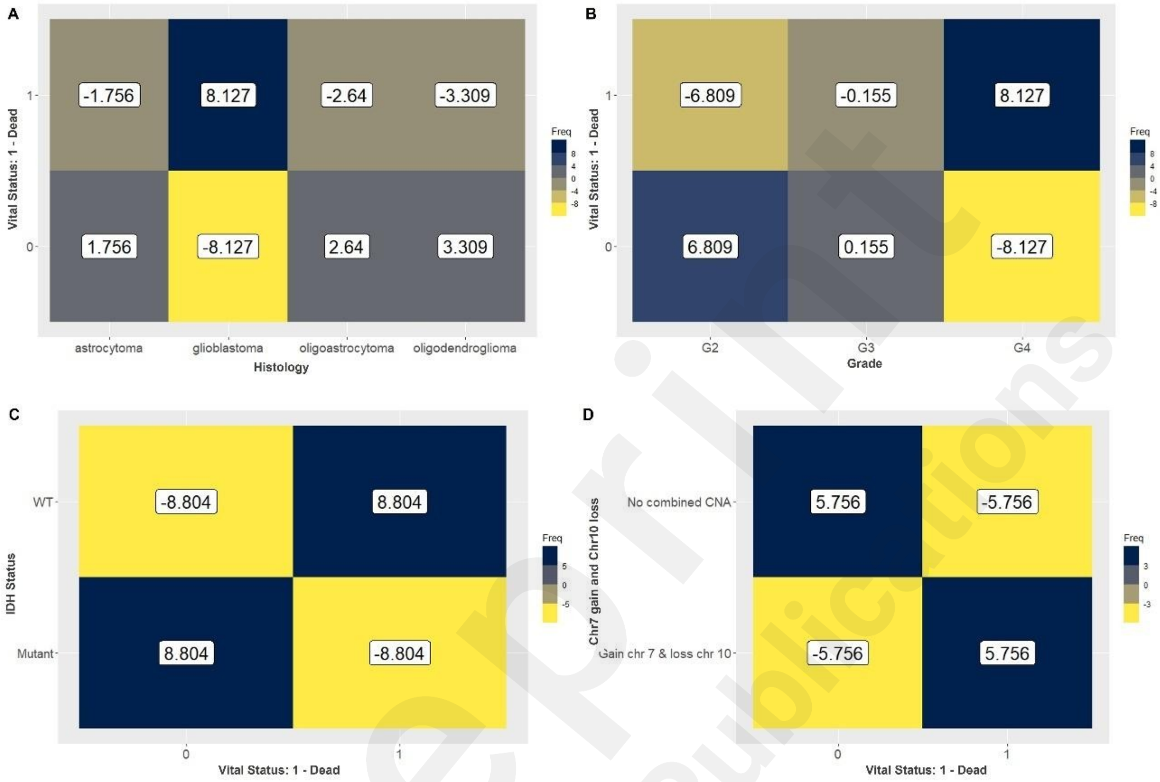


Figure 1. Heatmap exhibiting the values of the adjusted standardized residuals. Categories of variables with values higher than 1.96 are associated. We could observe a strong association of (**A**) Glioblastoma (8.127), (**B**) grade 4 (8.127), (**C**) IDH wild type (8.804), and (**D**) Chr7+/Chr10- (5.756) with dead vital status. Favorable prognostic factors including (**A**) oligoastrocytoma, oligodendroglioma, (**B**) grade 2, (**C**) IDH mutant, and (**D**) no combined CNA were associated with alive vital status.

Table 1. Table exhibiting the values of the adjusted standardized residuals. Categories of variables with values higher than 1.96 are considered associated. We could observe a strong association between poorer prognostic factors and dead vital status. In contrast, better prognostic factors were associated with alive vital status.

	Patient Vital Status		
Glioma Variables	**Alive**	**Dead**	**Categories are associated with:**
Glioblastoma	-	8.127	Dead

Oligoastrocytoma	2.64	-	Alive
Oligodendroglioma	3.309	-	Alive
Astrocytoma	1.756	-	Not associated
Grade 2	6.809	-	Alive
Grade 3	0.155		Not associated
Grade 4	-	8.127	Dead
IDH Wild Type	-	8.804	Dead
IDH Mutant	8.804		Alive
1p/19q codeletion	5.265	-	Alive
1p/19q non-codeletion	-	5.265	Dead
Methylated MGMT Promoter	5.26	-	Alive
Unmethylated MGMT Promoter	-	5.26	Dead
No combined Chr7+/Chr10-	5.756	-	Alive
Chr7+/Chr10-	-	5.756	Dead
Not Expressed TERT	3.078	-	Alive
Expressed TERT	-	3.078	Dead
ATRX mutant	2.311	-	Alive
ATRX Wild Type	-	2.311	Dead
Proneural Subtype	4.122	-	Alive
Neural Subtype	3.593	-	Alive
Mesenchymal Subtype	-	4.635	Dead
Classical Subtype	-	4.852	Dead

Using MCA, we observed that dimension 1 (x-axis) accounted for 33.71% of the variance, while dimension 2 (y-axis) accounted for 14.08%. The inertia (sum of the variances) for these two dimensions was 47.79%. The variance of the overall dimensions (17 dimensions) for the combinations of the variables is illustrated in *Supplementary Figure 1*. The main idea was to present the percentage of explained variance for each dimension and not the influence of individual variables. The total inertia (sum of the variances) was 1.41.

The results obtained from the MCA were visualized in a two-dimensional perceptual map (*Figure 2*), highlighting the associations between the categories of each variable. The coordinates of each category are detailed in *Table 2*. The perceptual map reveals that categories such as glioblastoma, Unmethylated MGMT promoter, IDH wild type, chr7 gain and chr 10 loss, grade 4, glioblastoma ATRX wild type, TERT expression, non-codel 1p.19q, classical (CL) and mesenchymal (ME) transcriptome subtypes are closely associated with dead vital status, appearing along the positive x-axis (dimension 1). Conversely, categories like oligoastrocytomas and oligodendrogliomas, grade 2, IDH mutant, codel 1p19q, methylated MGMT promoter, no combined CNA, no expression of TERT, ATRX mutant, and proneural (PN) and neural (NE) transcriptome subtypes are closely associated with alive vital status, appearing along the negative X-axis (dimension 1) (*Figure 2*).

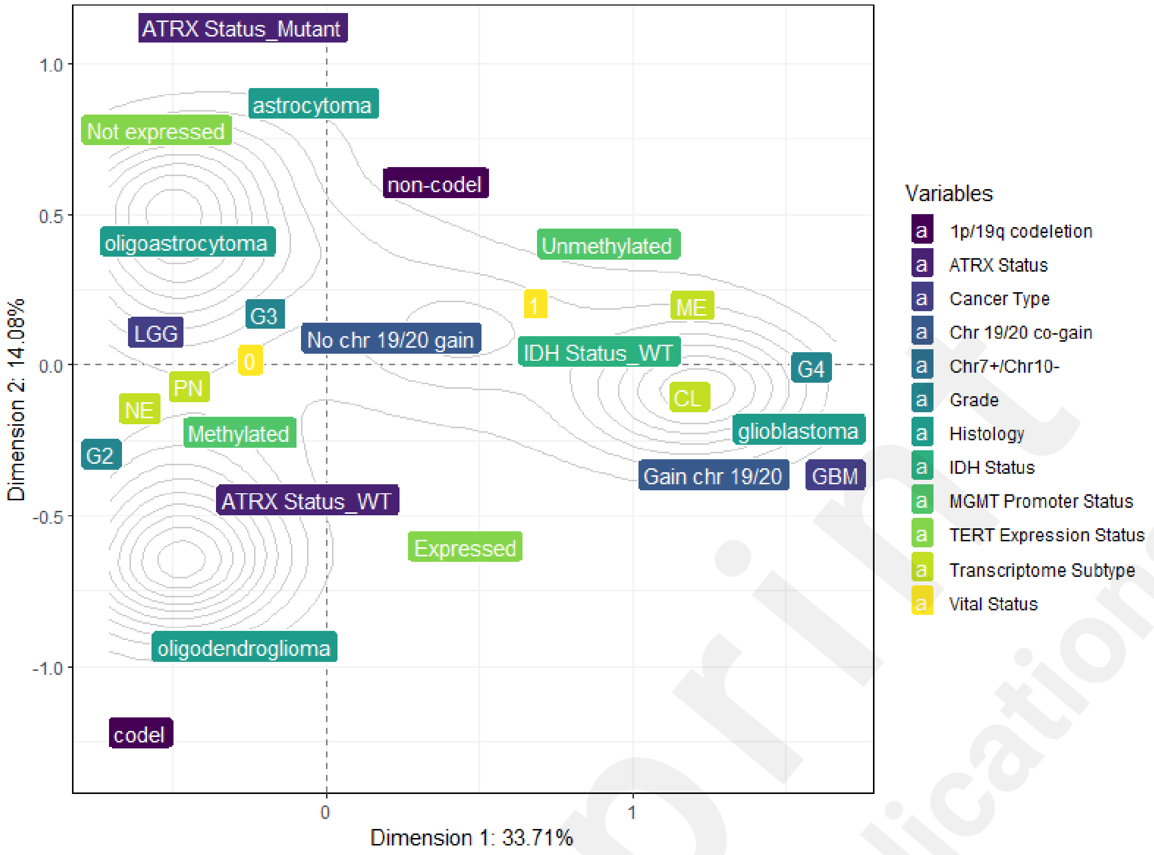


Figure 2. MCA two-dimensional perceptual map demonstrating the association between the categories of each categorical variable. Categories that are closely clustered are strongly associated with each other. Categories such as glioblastoma, Unmethylated MGMT promoter, IDH wild type, chr7 gain and chr 10 loss, grade 4, glioblastoma ATRX wild type, TERT expression, non-codel 1p.19q, classical (CL) and mesenchymal (ME) transcriptome subtypes are closely associated with dead vital status (1), appearing along the positive x-axis (dimension 1).

These findings highlight the utility and capacity of MCA in reducing data dimensionality and demonstrate that, in gliomas, variables interact cohesively. MCA allows us to further visualize these interactions on a global perceptual map, organizing the characteristics into distinct clusters that correspond to different prognostic profiles.

Table 2. Coordinates of each categories compounding the perceptual map.

Category	Dim 1 (X-axis)	Dim 2 (Y-Axis)
Glioblastoma	1.6650830	-0.0896760
Low-grade glioma	-0.4723301	0.0254382
astrocytoma	-0.2672355	0.9527631
glioblastoma	1.6650830	-0.0896760
oligoastrocytoma	-0.5334711	0.3276318
oligodendroglioma	-0.6011671	-0.9346433
Grade 2	-0.6611308	-0.1971919
Grade 3	-0.2970898	0.2320783
Grade 4	1.6650830	-0.0896760
0 – Alive	-0.3185609	-0.0551369
1 - Dead	0.7544862	0.1305874
IDH Mutant	-0.6734117	-0.0548104
IDH WT	1.1888626	0.0967641
1p/19q codel	-0.6877365	-13.034.766
1p/19q non-codel	0.2750946	0.5213906
Methylated	-0.3429710	-0.1087842
Unmethylated	1.0048449	0.3187185
Chr7+/Chr10-	1.4087248	-0.0210234
No combined Chr7+/Chr10-	-0.4205758	0.0062766
Chr 19/20 co-gain	1.4900007	-0.1295089
No Chr 19/20 co-gain	-0.0843397	0.0073307
Expressed TERT	0.3715020	-0.6845760
Not expressed TERT	-0.4690682	0.8643636
ATRX Mutant	-0.6448249	1.0773395
ATRX WT	0.2693572	-0.4500279
Classical	1.2675815	-0.0217510
Mesenchymal	1.0920361	0.2687642
Neural	-0.5475482	-0.0650952
Proneural	-0.5971662	-0.0604168

MCA can associate an epigenetic stemness index (mDNAsi) as a prognostic factor in Gliomas

After demonstrating that MCA effectively reduces dimensionality and identifies associations between prognostic factors and clinical data in the glioma database, we proceeded to explore whether MCA could also associate these variables with stemness phenotype. For this analysis, we updated our database by including mDNA stemness index (mDNAsi) as a new variable, categorized into low, intermediate, and high levels of stemness. These categories were based on the DNA methylation index related to tumor pathology and clinical outcomes, as previously studied by Malta et al., 2018 [22]

First, we evaluated whether the categorical glioma variables were randomly or non-randomly associated with mDNAsi by creating individual contingency tables for each pair of glioma variables and applying Chi-squared tests (*Supplementary Tables 14*). All the variables were found to be suitable for MCA. Then, using ASR values to evaluate the strength of these associations, our results indicated strong associations between high mDNAsi levels and poor prognostic and clinical factors. Higher mDNAsi levels were associated with glioblastoma, IDH wild-type, absence of 1p19q co-deletion, unmethylated MGMT promoter, TERT Expression, grade 3 and 4, patient's vital status as dead, chromosome 7 gain and 10 loss (Chr7+/Chr10-), chromosomes 19/20 co-gain, ATRX wildtype and mesenchymal and classical transcriptome subtypes (*Table 3*). Conversely, intermediate and lower levels of mDNAsi were associated with characteristics related to favorable prognosis, including oligodendroglioma, IDH mutant, 1p19q co-deletion, methylation of MGMT promoter, absence of TERT expression, grade 2, patient's vital status as alive, no combined copy number alteration, absence of chromosomes 19/20 co-gain, ATRX mutant, and proneural and neural transcriptome subtypes (*Table 3*).

Table 3. Table exhibiting the values of the adjusted standardized residuals. Categories of variables with values higher than 1.96 are considered associated. We could observe a strong association between poorer prognostic factors and higher stemness index (mDNAsi). In contrast, better prognostic factors were associated with lower stemness index.

		mDNAsi		
Glioma Variables	**Low**	**Intermediate**	**High**	**Categories are associated with:**
Glioblastoma	-	-	8.507	High
Oligoastrocytoma	-	-	-	Not Associated
Oligodendroglioma	3.949	-	-	Low
Astrocytoma	-		2.832	High
G2	3.279	4.057	-	Low and Intermediate
G3	-	-	2.392	High
G4	-	-	8.507	High
IDH Wild Type	-	-	15.904	High
IDH Mutant	8.743	7.057	-	Low and intermediate
1p/19q codeletion	5.772	2.102	-	Low and intermediate
1p/19q non-codeletion	-	-	7.964	High
Methylated MGMT Promoter	5.944	3.961	-	Low and intermediate
Unmethylated MGMT Promoter	-	-	9.983	High
No combined	6.436	5.927	-	Low and

Chr7+/Chr10-				intermediate
Chr7+/Chr10-	-	-	12.433	High
Not Expressed TERT	-	3.216	-	Intermediate
Expressed TERT	-	-	3.351	High
ATRX mutant	-	3.505	-	Intermediate
ATRX Wild Type	-	-	4.949	High
Proneural Subtype	8.476	-	-	Low
Neural Subtype	-	4.218	-	Intermediate
Mesenchymal Subtype	-	-	4.771	High
Classical Subtype	-	-	10.981	High

Using MCA, dimension 1 (x-axis) accounted for 28.7% of the variance, while dimension 2 (y-axis) accounted for 14.39%. The inertia (sum of the variances) for these two dimensions was 43.09%. The variance of the overall dimensions (18 dimensions) for the combinations of the variables is illustrated in *Supplementary Figure 2*. The total inertia (sum of the variances) was 1.5. The two-dimensional perceptual map exhibited the associations between the categories of each variable (*Figure 3*). The perceptual map reveals that categories such as glioblastoma, Unmethylated MGMT promoter, IDH wild type, chr7 gain and chr 10 loss, grade 4, glioblastoma ATRX wild type, TERT expression, non-codel 1p.19q, classical (CL) and mesenchymal (ME) transcriptome subtypes are closely associated with high mDNAsi, appearing along the positive x-axis (dimension 1). Conversely, categories like oligoastrocytomas and oligodendrogliomas, grade 2, IDH mutant, codel 1p19q, methylated MGMT promoter, no combined CNA, no expression of TERT, ATRX mutant, and proneural (PN) and neural (NE) transcriptome subtypes are closely associated with alive vital status, appearing along the negative X-axis (dimension 1) (*Figure 3*).

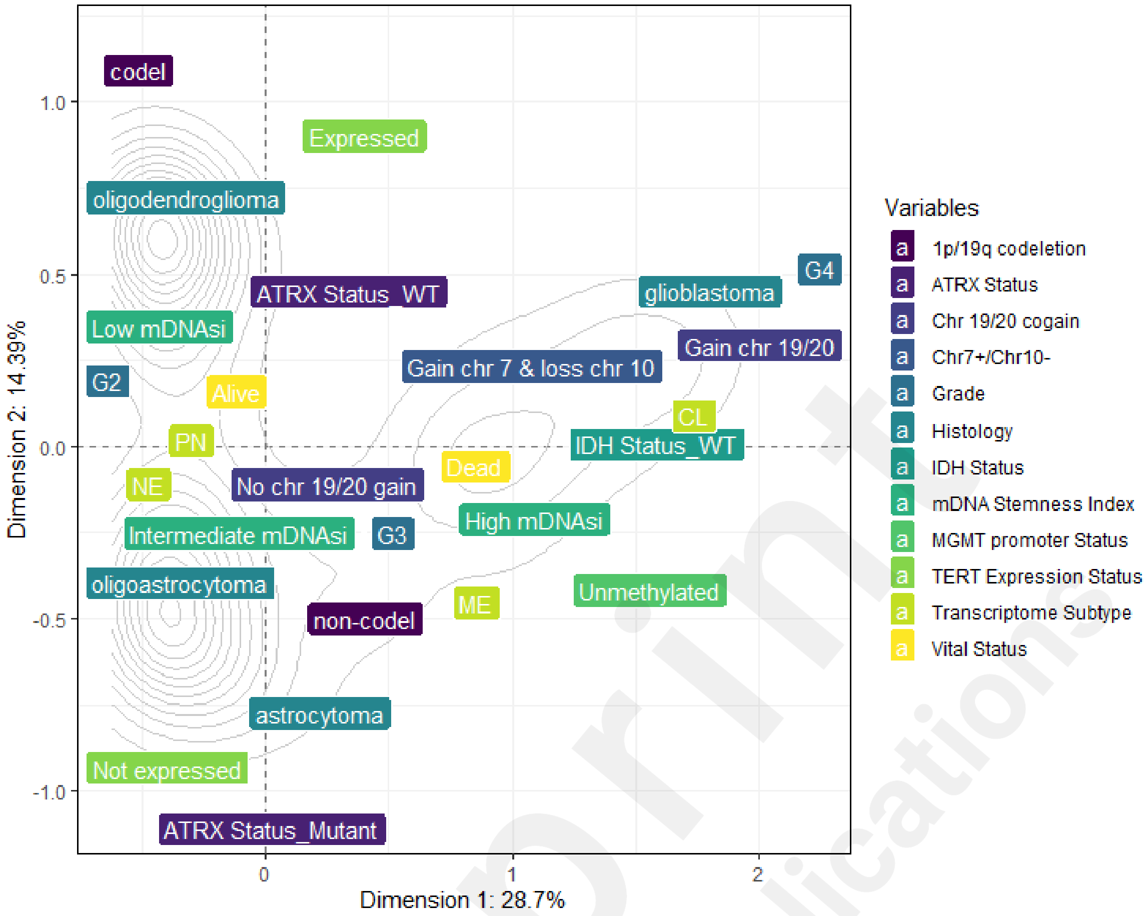


Figure 3. MCA two-dimensional perceptual map demonstrating the association between the categories of each categorical variable. Categories that are closely clustered are strongly associated with each other. Categories such as glioblastoma, Unmethylated MGMT promoter, IDH wild type, chr7 gain and chr 10 loss, grade 4, glioblastoma ATRX wild type, TERT expression, non-codel 1p.19q, classical (CL) and mesenchymal (ME) transcriptome subtypes are closely associated with high mDNAsi, appearing along the positive x-axis (dimension 1).

Discussion

Multiple efforts have been made to explore the diversity of oncologic diseases, with significant contributions from genetics, cell and tissue biology, as well as computational and experimental technologies, providing a wealth of information on cancer manifestations [24]. In the field of glioma research, emerging approaches have sought to clarify tumor pathology and grading through the introduction of novel types and subtypes, as well as by identifying molecular markers and genetic mutations that contribute to predicting diagnosis and prognosis [25]. However, it also results in an accumulation of extensive datasets, presenting challenges in interpretation and visualization regarding the associations between prognostic factors. In this study, we employed Multiple Correspondence Analysis, an unsupervised data science approach, to establish statistical associations between different qualitative variables of gliomas. This method was able to reduce data dimensionality and represent it on a two-dimensional perceptual map, revealing associations between various established glioma prognostic factors, including histological classification, IDH status, MGMT promoter methylation, and transcriptome subtypes. Furthermore, we associated these clinical and prognostic variables with an epigenetic-based stemness index (mDNAsi), demonstrating that higher stemness levels were associated with poorer prognostic factors, providing a useful tool to associate prognostic markers in brain tumors.

Several clinical and molecular factors are considered in predicting the prognosis and survival of brain tumors, more specifically for gliomas. Beyond histological classification and tumor grade, genetic and molecular biomarkers have been incorporated as potential prognostic indicators. Thus, we first evaluated the ability of MCA to associate these consolidated prognostic variables with the patient's vital status. Our findings demonstrate that MCA effectively clusters poor prognostic factors with dead vital status. Subsequently, we applied MCA to explore the association between high stemness levels (mDNAsi) and characteristics related to poor prognosis. Stemness has been considered an important phenotype in glioma malignancy and is potentially associated with classical genetic alterations, such as the gain of chromosome 7. Chromosome 7 harbors some key genes related to stemness, including *EGFR*, *MET*, and *HOXA*. A study of 86 glioblastomas reported that *EGFR* amplification occurs with higher probability in samples that have a gain of chromosome 7 (82.1%) compared to those without it (66.7%) [26]. In addition, *EGFR* amplification is more prevalent in IDH-wildtype diffuse gliomas (66.0%) and GBM (85.5%) [27], which are also associated with poorer prognostic factors, consistent with our findings. High mDNAsi has been previously linked to EGFR mutations [22]. The HOXA and MET loci, also located on chromosome 7, have been implicated in stemness-related pathways. Notably, studies have demonstrated interactions between chromosome 7 gain and the expression of a stem cell-related HOX signature in glioblastomas [28]. Analysis of the MET gene at 7q31.2 revealed that gain occurs in 47% of primary and 44% of secondary glioblastomas, suggesting that this genetic alteration contributes to the pathogenesis of both glioblastoma subtypes [29]

Overall, relatively few studies have employed MCA to explore associations with cancer phenotypes. Previous studies have applied MCA to different approaches, such as analyzing prognosis low rectal cancer surgery [30], investigating the association between some types of cancer in rural or urban areas [15], examining the association between Traditional Chinese Medicine Syndrome and histopathology of colorectal cancer [31], assessing clinically relevant demographic variables across multiple gastrointestinal cancers [32], and the relationship between types of diagnostic classification in breast cancer [33]. Our study also highlights the utility of MCA in investigating associations within the context of cancer. MCA enables to investigation of the pattern among many categorical factors in gliomas, providing a powerful computational approach to identify and test prognostic variables. It was possible to visually and quantitatively represent the associations, which facilitates the identification of distinct patient clusters based on shared prognostic characteristics. Our findings were consistent with previous literature and emphasized stemness as an important phenotype for gliomas.

Our study has inherent limitations. First, as a retrospective analysis of TCGA data, it is subject to selection bias. Furthermore, we associated all the prognostic variables with patient vital status, which may not be the most optimal variable for determining prognosis. The absence of therapy data is another limitation of this study. Moreover, an intrinsic limitation of MCA is that retaining only two or three dimensions may not sufficiently capture all the significant features in the data. In our analysis, the percentage of explained inertia was approximately 40%. While there is not an accepted threshold for adequately explained inertia, common guidelines recommend retaining dimensions that represent over 70% of the inertia [34]. However, explained inertia in the range of 40%-60% is often considered informative, and the interpretability and relevance of the patterns revealed by the dimensions are frequently more important than the exact percentage of inertia explained [35].

In conclusion, our findings suggest that MCA is a valuable tool for understanding the interdependence between prognostic markers in gliomas. MCA facilitates the exploration of a large-scale dataset and enhances the identification of associations. Considering the advances in computational oncology and the emergence of new oncological features, such as stemness phenotype, incorporating MCA into cancer research as an approach to exploring the complex heterogeneity of the oncologic field becomes a powerful tool for simplifying data management. It contributes to researchers statistically identifying associations between variables within extensive databases and improves the visual representation, leading to a deeper understanding of cancer findings.

Acknowledgments

This study has been supported by grants from the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) and from the Sao Paulo Research Foundation (FAPESP), Brazil (2018/00583-0, 2022/06305-7, 2022/09378-5, 2023/05099-7, 2023/07358-0).

Conflict of interest

None declared.

References

[1] D. Hanahan, “Hallmarks of Cancer: New Dimensions,” *Cancer Discov*, vol. 12, no. 1, pp. 31–46, 2022, doi: 10.1158/2159-8290.CD-21-1059.

[2] I. Dagogo-Jack and A. T. Shaw, “Tumour heterogeneity and resistance to cancer therapies,” *Nat Rev Clin Oncol*, vol. 15, no. 2, pp. 81–94, 2018, doi: 10.1038/nrclinonc.2017.166.

[3] J. Brierley *et al.*, “Global Consultation on Cancer Staging: promoting consistent understanding and use,” *Nat Rev Clin Oncol*, vol. 16, no. 12, pp. 763–771, 2019, doi: 10.1038/s41571-019-0253-x.

[4] M. Weller *et al.*, “Glioma,” *Nat Rev Dis Primers*, vol. 1, no. July, 2015, doi: 10.1038/nrdp.2015.17.

[5] D. N. Louis *et al.*, “The 2007 WHO classification of tumours of the central nervous system,” Aug. 2007. doi: 10.1007/s00401-007-0243-4.

[6] D. N. Louis *et al.*, “The 2021 WHO classification of tumors of the central nervous system: A summary,” *Neuro Oncol*, vol. 23, no. 8, pp. 1231–1251, 2021, doi: 10.1093/neuonc/noab106.

[7] A. Z. Ayob and T. S. Ramasamy, “Cancer stem cells as key drivers of tumour progression,” *J Biomed Sci*, vol. 25, no. 1, pp. 1–18, 2018, doi: 10.1186/s12929-018-0426-4.

[8] E. Batlle and H. Clevers, “Cancer stem cells revisited,” *Nat Med*, vol. 23, no. 10, pp. 1124–1134, 2017, doi: 10.1038/nm.4409.

[9] Q. Wang *et al.*, “Tumor Evolution of Glioma-Intrinsic Gene Expression Subtypes Associates with Immunological Changes in the Microenvironment,” *Cancer Cell*, vol. 32, no. 1, pp. 42-56.e6, Jul. 2017, doi: 10.1016/j.ccell.2017.06.003.

[10] B. Ortensi, M. Setti, D. Osti, and G. Pelicci, “Cancer stem cell contribution to glioblastoma invasiveness,” *Stem Cell Res Ther*, vol. 4, no. 1, pp. 1–11, 2013, doi: 10.1186/scrt166.

[11] J. Tan *et al.*, “Molecular Subtypes Based on the Stemness Index Predict Prognosis in Glioma Patients,” *Front Genet*, vol. 12, Mar. 2021, doi: 10.3389/fgene.2021.616507.

[12] N. Sourial *et al.*, “Correspondence analysis is a useful tool to uncover the relationships among categorical variables,” *J Clin Epidemiol*, vol. 63, no. 6, pp. 638–646, 2010, doi: 10.1016/j.jclinepi.2009.08.008.

[13] B. H. Li, Z. Q. Sun, and S. F. Dong, “Correspondence analysis and its application in oncology,” *Commun Stat Theory Methods*, vol. 39, no. 7, pp. 1229–1236, 2010, doi: 10.1080/03610920902871446.

[14] P. S. Costa, N. C. Santos, P. Cunha, J. Cotter, and N. Sousa, “The use of multiple correspondence analysis to explore associations between categories of qualitative variables in healthy ageing,” *J Aging Res*, vol. 2013, 2013, doi: 10.1155/2013/302163.

[15] D. Florensa *et al.*, “The Use of Multiple Correspondence Analysis to Explore Associations between Categories of Qualitative Variables and Cancer Incidence,” *IEEE J Biomed Health Inform*, vol. 25, no. 9, pp. 3659–3667, 2021, doi: 10.1109/JBHI.2021.3073605.

[16] A. Van Horn *et al.*, “Using multiple correspondence analysis to identify behaviour patterns associated with overweight and obesity in Vanuatu adults,” *Public Health Nutr*, vol. 22, no. 9, pp. 1533–1544, 2019, doi: 10.1017/S1368980019000302.

[17] D. N. Louis *et al.*, “The 2021 WHO classification of tumors of the central nervous system: A

summary," *Neuro Oncol*, vol. 23, no. 8, pp. 1231–1251, Aug. 2021, doi: 10.1093/neuonc/noab106.

[18] P. Śledzińska, M. G. Bebyn, J. Furtak, J. Kowalewski, and M. A. Lewandowska, "Prognostic and predictive biomarkers in gliomas," Oct. 01, 2021, *MDPI*. doi: 10.3390/ijms221910373.

[19] A. Sokolov, E. O. Paull, and J. M. Stuart, "One-class detection of cell states in tumor subtypes," *Pacific Symposium on Biocomputing*, pp. 405–416, 2016, doi: 10.1142/9789814749411_0037.

[20] N. Salomonis *et al.*, "Integrated Genomic Analysis of Diverse Induced Pluripotent Stem Cells from the Progenitor Cell Biology Consortium," *Stem Cell Reports*, vol. 7, no. 1, pp. 110–125, Jul. 2016, doi: 10.1016/j.stemcr.2016.05.006.

[21] K. Daily *et al.*, "Molecular, phenotypic, and sample-associated data to describe pluripotent stem cell lines and derivatives," *Sci Data*, vol. 4, Mar. 2017, doi: 10.1038/sdata.2017.30.

[22] T. M. Malta *et al.*, "Machine Learning Identifies Stemness Features Associated with Oncogenic Dedifferentiation," *Cell*, vol. 173, no. 2, pp. 338-354.e15, Apr. 2018, doi: 10.1016/j.cell.2018.03.034.

[23] S. Lê, J. Josse, and F. Husson, "FactoMineR: An R Package for Multivariate Analysis," *J Stat Softw*, vol. 25, no. 1, pp. 1–18, 2008, doi: 10.18637/jss.v025.i01.

[24] D. Hanahan, "Hallmarks of Cancer: New Dimensions," Jan. 01, 2022, *American Association for Cancer Research Inc.* doi: 10.1158/2159-8290.CD-21-1059.

[25] D. N. Louis *et al.*, "The 2021 WHO classification of tumors of the central nervous system: A summary," *Neuro Oncol*, vol. 23, no. 8, pp. 1231–1251, Aug. 2021, doi: 10.1093/neuonc/noab106.

[26] S. N. McNulty *et al.*, "Beyond sequence variation: assessment of copy number variation in adult glioblastoma through targeted tumor somatic profiling," *Hum Pathol*, vol. 86, pp. 170–181, Apr. 2019, doi: 10.1016/j.humpath.2018.12.004.

[27] H. Wang *et al.*, "Clinical roles of EGFR amplification in diffuse gliomas: a real-world study using the 2021 WHO classification of CNS tumors," *Front Neurosci*, vol. 18, 2024, doi: 10.3389/fnins.2024.1308627.

[28] S. Kurscheid *et al.*, "Chromosome 7 gain and DNA hypermethylation at the HOXA10 locus are associated with expression of a stem cell related HOX-signature in glioblastoma," *Genome Biol*, vol. 16, no. 1, Jan. 2015, doi: 10.1186/s13059-015-0583-7.

[29] D. Pierscianek *et al.*, "MET gain in diffuse astrocytomas is associated with poorer outcome," *Brain Pathology*, vol. 23, no. 1, pp. 13–18, Jan. 2013, doi: 10.1111/j.1750-3639.2012.00609.x.

[30] R. Mancini *et al.*, "Tumor Regression Grade After Neoadjuvant Chemoradiation and Surgery for Low Rectal Cancer Evaluated by Multiple Correspondence Analysis: Ten Years as Minimum Follow-up," *Clin Colorectal Cancer*, vol. 17, no. 1, pp. e13–e19, Mar. 2018, doi: 10.1016/j.clcc.2017.06.004.

[31] T. Wu *et al.*, "Correspondence analysis between traditional Chinese medicine (TCM) syndrome differentiation and histopathology in colorectal cancer," *Eur J Integr Med*, vol. 7, no. 4, pp. 342–347, Aug. 2015, doi: 10.1016/j.eujim.2015.07.003.

[32] R. J. Kramer *et al.*, "Unsupervised clustering using multiple correspondence analysis reveals clinically-relevant demographic variables across multiple gastrointestinal cancers," *Surgical Oncology Insight*, vol. 1, no. 1, p. 100009, Mar. 2024, doi: 10.1016/j.soi.2024.100009.

[33] M. Nadjib Bustan, M. Arif Tiro, and S. Annas, "Correspondence Analysis of Breast Cancer Diagnosis Classification," in *Journal of Physics: Conference Series*, Institute of Physics Publishing, Jun. 2019. doi: 10.1088/1742-6596/1244/1/012030.

[34] N. T. Higgs, "Practical and Innovative Uses of Correspondence Analysis," *Journal of the Royal Statistical Society. Series D (The Statistician)*, vol. 40, no. 2, pp. 183–194, 1991, doi: 10.2307/2348490.

[35] F. Husson, S. Lê, and J. Pagès, *Exploratory multivariate analysis by example using R*. CRC Press, 2011.

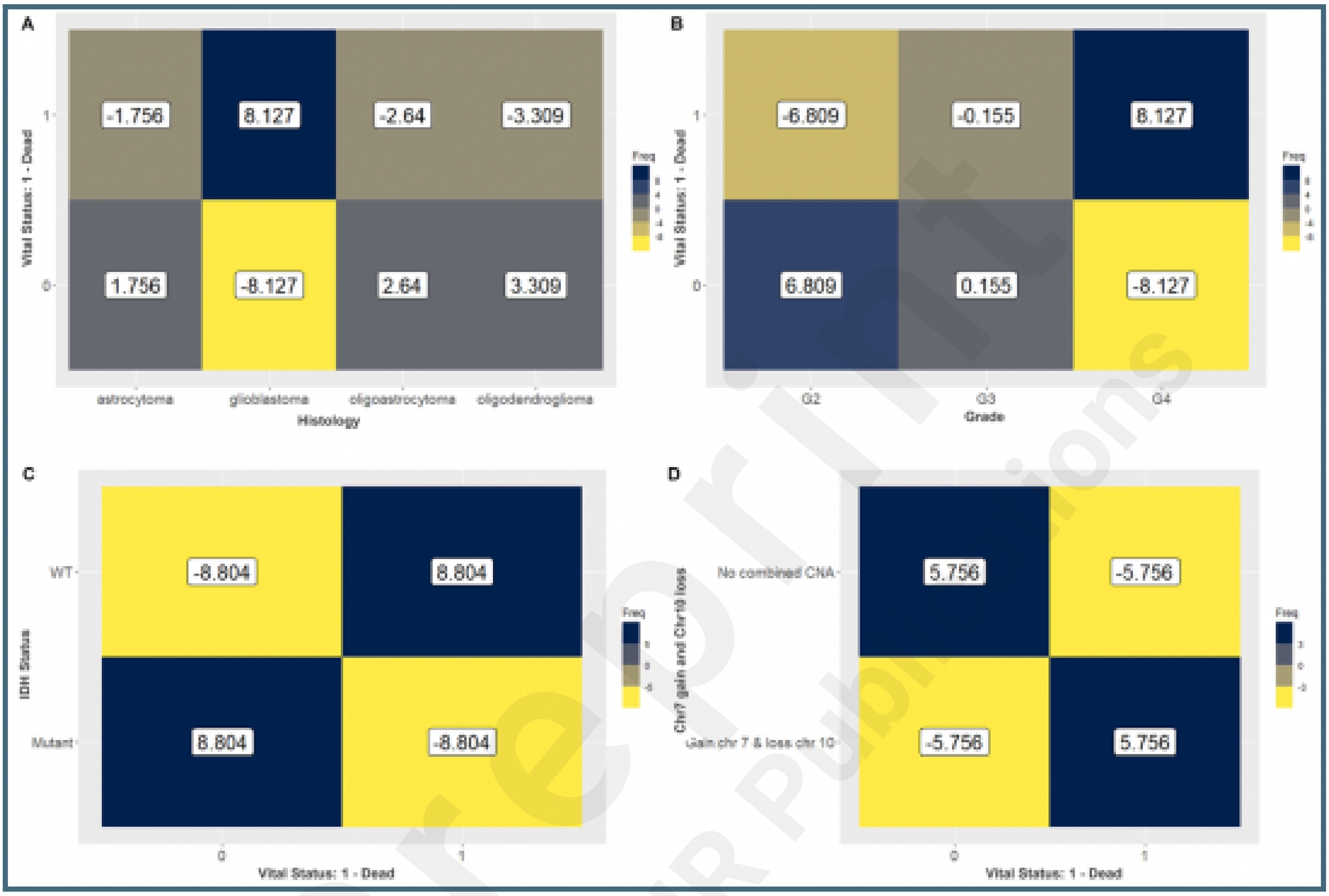
Abbreviations

GMB: Glioblastomas
CSCs: Cancer Stem Cells
MCA: Multiple Correspondence Analysis
PCA: Principal Component Analysis
TCGA: The Cancer Genome Atlas
IDH: Isocitrate Dehydrogenase
MGMT: Methylguanine Methyltransferase
TERT: Telomerase Reverse Transcriptase
ATRX: Alpha Thalassemia/Mental Retardation Syndrome X-linked
mDNAsi: Methylation DNA Stemness Index
OCLR: One-Class Logistic regression
PCBC: Progenitor Cell Biology Consortium
ESC: Embryonic Stem Cell
iPSC: Induced Pluripotent Stem Cell
ASR: Adjusted Standardized Residuals
CL: Classical
ME: Mesenchymal
PR: Proneural
NE: Neural

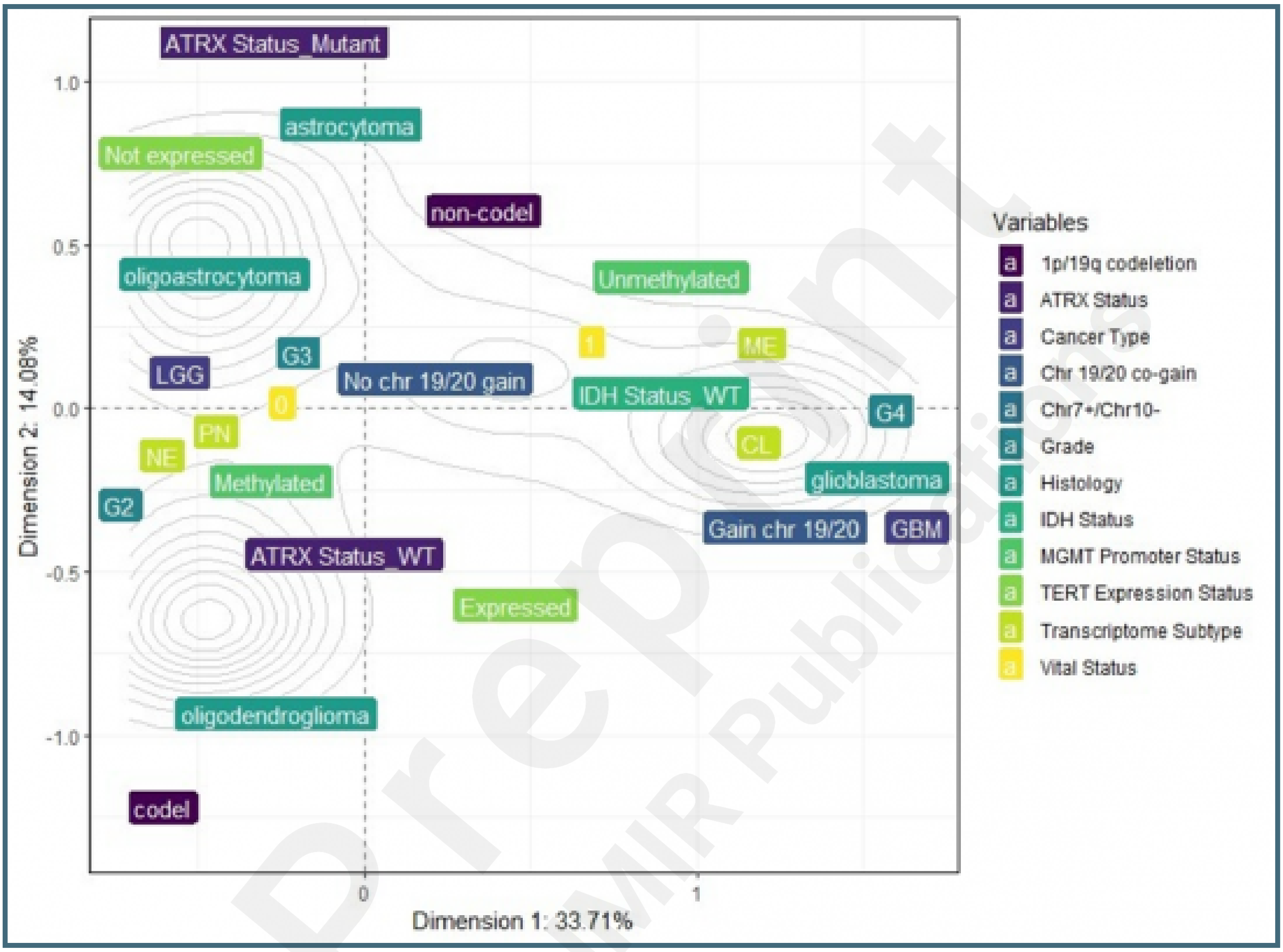
Supplementary Files

Figures

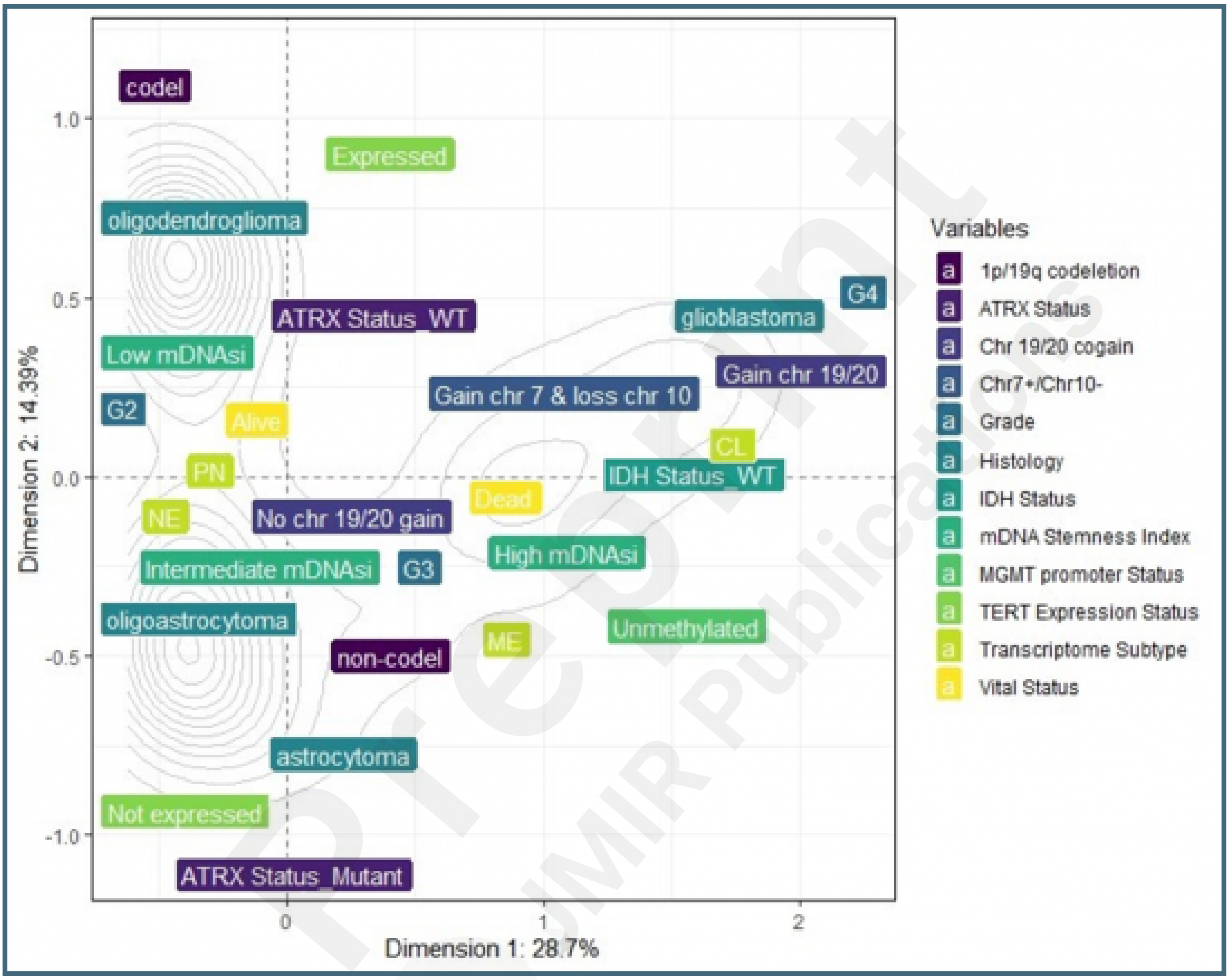
Heatmap exhibiting the values of the adjusted standardized residuals. Categories of variables with values higher than 1.96 are associated. We could observe a strong association of (A) Glioblastoma (8.127), (B) grade 4 (8.127), (C) IDH wild type (8.804), and (D) Chr7+/Chr10- (5.756) with dead vital status. Favorable prognostic factors including (A) oligoastrocytoma, oligodendroglioma, (B) grade 2, (C) IDH mutant, and (D) no combined CNA were associated with alive vital status.



MCA two-dimensional perceptual map demonstrating the association between the categories of each categorical variable. Categories that are closely clustered are strongly associated with each other. Categories such as glioblastoma, Unmethylated MGMT promoter, IDH wild type, chr7 gain and chr 10 loss, grade 4, glioblastoma ATRX wild type, TERT expression, non-codel 1p.19q, classical (CL) and mesenchymal (ME) transcriptome subtypes are closely associated with dead vital status (1), appearing along the positive x-axis (dimension 1).



MCA two-dimensional perceptual map demonstrating the association between the categories of each categorical variable. Categories that are closely clustered are strongly associated with each other. Categories such as glioblastoma, Unmethylated MGMT promoter, IDH wild type, chr7 gain and chr 10 loss, grade 4, glioblastoma ATRX wild type, TERT expression, non-codel 1p.19q, classical (CL) and mesenchymal (ME) transcriptome subtypes are closely associated with high mDNAsi, appearing along the positive x-axis (dimension 1).



Multimedia Appendixes

Percentage of explained variances of the overall dimensions.
URL: http://asset.jmir.pub/assets/f0941e4ae31ed85aaaddd688c47015cd.png

Percentage of explained variances of the overall dimensions.
URL: http://asset.jmir.pub/assets/a94dd644eb17d00937f89b905a036f49.png

Individual contingency tables for each pair of glioma variables.
URL: http://asset.jmir.pub/assets/478d21ba3cc275610a0ec00a5585dfac.xlsx

Individual contingency tables for each pair of glioma variables.
URL: http://asset.jmir.pub/assets/efd4928e4d4dafbf84c711df3230384d.xlsx

Individual contingency tables for each pair of glioma variables.
URL: http://asset.jmir.pub/assets/f4713233bbafafba82e082d1b86ee34b.xlsx

Individual contingency tables for each pair of glioma variables.
URL: http://asset.jmir.pub/assets/69ee50f09e026da1c98f16136933ca3f.xlsx

Individual contingency tables for each pair of glioma variables.
URL: http://asset.jmir.pub/assets/fb4b897367506c00335c862a963ecbfe.xlsx

Individual contingency tables for each pair of glioma variables.
URL: http://asset.jmir.pub/assets/a0ea877051a95cd6185d7650180c4d5e.xlsx

Individual contingency tables for each pair of glioma variables.
URL: http://asset.jmir.pub/assets/e9781c301a44ea3a84072d8131d3e959.xlsx

Individual contingency tables for each pair of glioma variables.
URL: http://asset.jmir.pub/assets/718ae64b71a7fc7daf34d500d8ce5584.xlsx

Individual contingency tables for each pair of glioma variables.
URL: http://asset.jmir.pub/assets/0b1e67cc7d6f915bd64fcfe59bc01e48.xlsx

Individual contingency tables for each pair of glioma variables.
URL: http://asset.jmir.pub/assets/f0fac134af45bac5db13037dd5955c47.xlsx

Individual contingency tables for each pair of glioma variables.
URL: http://asset.jmir.pub/assets/2dad978cd0e4dd20a8dfac0318224130.xlsx

Individual contingency tables for each pair of glioma variables.
URL: http://asset.jmir.pub/assets/f24c15ca56218faffb9d296517d6ce22.xlsx

Individual contingency tables for each pair of glioma variables.
URL: http://asset.jmir.pub/assets/7061c8b82c37f6fb1972e5eb336ca185.xlsx

Individual contingency tables for each pair of glioma variables.
URL: http://asset.jmir.pub/assets/b28691e492782340de538950dde66800.xlsx