Multi-project and Multi-profile joint Non-negative Matrix Factorization for cancer omic datasets

D.A Salazar[a], N. Pržulj[b,c], C.F. Valencia[a]

[a] *School of Industrial Engineering, University of Los Andes, 111711, Bogota, Colombia*
[b] *Barcelona Supercomputing Center (BSC), 08034, Barcelona, Spain*
[c] *Department of Computer Science, University College London, WC1E 6BT, London, UK*

Abstract

Motivation: The integration of multi-omic data using machine learning methods has been focused on solving relevant tasks such as predicting sensitivity to a drug or subtyping patients. Recent integration methods, such as joint Non-negative Matrix Factorization (jNMF), have allowed researchers to exploit the information in the data to unravel the biological processes of multi-omic datasets.
Results: We present a novel method called Multi-project and Multi-profile joint Non-negative Matrix Factorization (M&M-jNMF) capable of integrating data from different sources, such as experimental and observational multi-omic data. The method can generate co-clusters between observations, predict profiles and relate latent variables. We applied the method to integrate low-grade glioma omic profiles from The Cancer Genome Atlas (TCGA) and Cell Line Encyclopedia (CCLE) projects. The method allowed us to find gene clusters mainly enriched in cancer-associated terms. We identified groups of patients and cell lines similar to each other by comparing biological processes. We predicted the drug profile for patients, and we identified genetic signatures for resistant and sensitive tumors to a specific drug.
Availability and implementation: Source code repository is publicly available at *https:/bitbucket.org/dsalazarb/mmjnmf/*
Supplementary information: Supplementary data are available at *Bioinformatics* online.

Preprint submitted to Bioinformatics *February 15, 2022*

1. Introduction

Data fusion has become an area of interest in biological sciences [1] because it is possible to integrate data from different sources to describe and uncover new properties of an individual. For instance, consider a type of cancer known as low-grade glioma, a subtype of brain cancer caused by somatic mutations in glial cells. We can measure many molecules to obtain partial knowledge of the disease for that cancer, but a greater understanding of the system comes when a model integrates all the interactions between different sources.

Many machine learning strategies have been used to better understand the interactions of the various data sources. In general, these methods are focused on tasks such as drug repurposing, molecular interactions prediction, variable importance identification, etc [2, 3, 4]. Among these methods, non-negative matrix factorization (NMF), which factorizes a non-negative input matrix X into low-rank matrices known as the base matrix (W) and the coefficient matrix (H), have been used to integrate various types of data to solve the tasks mentioned above and others [5, 6, 7, 8].

NMF methods have quite interesting properties for capturing patterns since they integrate a sparse and part-based representation of the data captured by two non-negative low-dimensional matrices (base and coefficient matrix). However, despite their usefulness, variants such as tri-factorization of non-negative matrices (NMTF) or joint factorization of non-negative matrices (jNMF) have taken a further step to include data from different sources and generate patterns or clusters based on this information, in addition to the possibility of predicting new links between the objects of study (patients, genes, or diseases) [8, 9]. For instance, [10] used a sparse version of jNMF to integrate miRNA and gene profiles of ovarian cancer. As a result, they identified enrichment co-clusters and groups of patients with significantly different survival characteristics.

Furthermore, these methods have been used to stratify patients, predict driver genes, and repurpose drugs. [6] used the NMTF strategy to integrate somatic mutations from The Cancer Genome Atlas (TCGA), molecular interactions from BioGRID and KEGG database, and chemical drug data from DrugBank [6]. They found three groups of ovarian cancer patients significantly separable by survival. They used these patient groups to identify two new genes (ADAM32 and REG1P) related to cell proliferation and tumor progression.

Although NMTF integrates relational matrices, it does not predict a concentration, dose, or expression for a particular molecule. A variant of NMF that allows factorizing non-relational matrices (X_I) is jNMF, which factors input matrices into a common base matrix (W) and individual coefficient matrices (H_I) with the same clustering properties of the NMF method. For example, [11] integrated six types of cell line profiles obtained from the Cancer Cell Line Encyclopedia (CCLE) database. As a result, they identified a greater sensitivity to PLX4720 when there is a mutation in BRAF and activation of MITF. Furthermore, jNMF can predict omic profiles and incorporate prior knowledge as a type of constraint that improves the interpretation of results, as [9] proposed. They used a data set of protein-RNA interactions predicting the interactions for 26 of 31 proteins with an AUC greater than 0.71.

Since jNMF allows researchers to solve different tasks in a single model, we propose that the data integration can use datasets from different projects simultaneously, for example, TCGA and CCLE. By using this information, we can explore different scientific tasks of interest, such as the identification of suitable cell lines for studying certain types of tumors [12] or the prediction of the degree of sensitivity that tumors may have based on the information of epigenetic and genetic expression of both projects [13].

In this paper, we present a new variant of jNMF to integrate omic profiles of observational data (TCGA), experimental data (CCLE), and biological knowledge to identify clusters for genes and miRNA, to co-cluster cell lines and patients, and to predict the drug sensitivity profile for tumors.

2. Methods

2.1. Datasets

The omic profiles for the observational dataset, i.e., low-grade glioma (LGG) tumors, were downloaded from The Cancer Genome Atlas (TCGA) project using the TCGA-Assembler v2.0.6 tool [14]. The omic profiles for the experimental dataset, i.e., cancer cell lines, were obtained from the Cancer Cell Line Encyclopedia (CCLE) project [15]. The common omic profiles between the two projects were gene expression, miRNA expression, and copy number variation (Supplementary Section S1). The drug sensitivity profile, which contains AUC values, was downloaded exclusively for the CCLE project. To ensure a positive input profiles, we scaled the values per columns using the formula ${}^{x_{ij}-X_{min}}\!/_{X_{max}-X_{min}}$ where x_{ij} is the i^{th} observation in the

j^{th} column of the matrix X. X_{max} and X_{min} is the maximum and the minimum value of the j^{th} column, respectively (detail pre-processing steps in Supplementary Section S1).

Each project required that the individuals (patients or cell lines) have all the profiles. In addition, for both projects, the pairs of omic profiles must have the same set of molecules (genes, miRNAs, or drugs).

2.2. Biological prior knowledge

The biological constraints incorporated in jNMF by [9] to improve the clustering of clusters (co-clustering) were Θ_I and R_{IJ}, where I and J are the identifiers for matrices. The former constraint refers to the intra-variable relationships, and the latter corresponds to inter-variable relationships. In Table 1, we summarized both constraints (detail description in Supplementary Section S2).

In the case of Θ_I constraints, we employed the notation $\Theta_{gene}^{(t)}$, where the superscript (t) corresponds to the number of constraints associated to this profile, when $(t) \geq 1$. For instance, we have four different matrices on genes, genetic interactions (Θ_{gene}^{1}), protein-protein interactions (Θ_{gene}^{2}), metabolic interactions (Θ_{gene}^{3}), and co-expression profiles (Θ_{gene}^{4}) which are described in the Table 1. The Θ_I constraints matrix have a square structure, e.g., the Θ_{gene} $(No.\,genes \times No.\,genes)$ constraint matrix corresponds to a binary matrix where an association gene-gene is categorized as 1, and 0 otherwise. The same is true for the Θ_{miRNA} matrix (Table 1).

The R_{IJ} constraints may have a square or a rectangular shape because they contained the association between two types of variables. For instance, $R_{drug-miRNA}$ $(No.\,drugs \times No.\,miRNAs)$ constraints relate drug with a miRNA. As Θ_I constraint, R_{IJ} constraints is a binary matrix.

2.3. Methods of joint factorization of non-negative matrices

2.3.1. Joint Non-negative Matrix Factorization

The standard method of joint non-negative matrix factorization (jNMF) approximates a set of non-negative input matrices $X_I \in R^{(n \times m_I)}$ for $I = 1, \ldots, M$, where I represents matrices of different measurements of many features (m_I) for the same objects, e.g. patients (n). The estimation of these matrices consists of finding non-negative low-rank approximations of each matrix, such that $X_I \approx WH_I$, where $W \in R^{(n \times k)}$ is a base matrix, and $H_I \in R^{(k \times m_I)}$ are the coefficient matrices for each I. Here, the coefficient matrices are particular for each input matrix, and the base matrix is unique

Table 1: Summary of constraints Θ_I and R_{IJ}. The number of nodes is lower than the variables used in the proposed factorization problem because there is no prior knowledge for all the variables, e.g., there are 314 miRNA, which 312 have evidence. Therefore, the dimension of Θ_{miRNA} constraint is 314×314. In $Theta^t_{gene}$ constraint the subscript (t) is equal to four.

Constraint	Description	No. Nodes	No. Edges	Edge density	Reference
Θ^t_{gene}	Genetic interactions	8585	848542	0.01151445	BioGRID v3.5 [16] STRINGdb v9.1 [17] KEGG graphite v.1.32.0 [18] limma v3.42.2 [19]
Θ_{miRNA}	miRNA-miRNA synergism	312	80678	0.8314577	CancerNet
Θ_{drug}	Drug-drug interactions	64	2866	0.7108135	DrugBank v5.0
$R_{miRNA-gene}$	miRNA-target interactions	13336	101659	0.0005716461	miRNet v2.0
$R_{miRNA-drug}$	miRNA-drug associations	70	86	0.01780538	miRNet v2.0

to the entire set of input matrices. Therefore, the matrix W allows for the integration of the data. The low-dimensional rank k guarantees a simpler latent structure that is interpretable as a separation into shared classes among all dimensions. For this reason, it is possible to cluster patients, cell lines, and molecules into groups. The jNMF method finds the matrices W and H_I that minimize:

$$\sum_{I=1}^{M} \|X_I - WH_I\|_F^2 \tag{1}$$

where $\|\cdot\|_F^2$ is the Frobenius norm of a matrix, that is, the sum of all squared elements. As W has dimensions $n \times k$, we can use the columns k as a latent structure to separate objects into shared groups. Similarly, H_I has dimensions $k \times m_I$, then it is possible to use k as a latent structure to group variables of different I in a common cluster, which is called co-cluster (Section 2.4).

2.3.2. Multi-project and Multi-profile joint Non-negative Matrix Factorization

We can solve the jNMF problem for observational and experimental data sets separately to obtain low-rank interpretations for each. However, given that both datasets shared most dimensions, although measured over different kinds of objects (e.g., patients and cell lines), we explored a pair-wise integration approach on which matrices H_I are shared for different objects, with an individual base matrix for observational data (W_{obs}) and for experimental data (W_{exp}).

In general, let X_I ($I = 1, \ldots, M_X$) and Y_I ($I = 1, \ldots, M_Y$) be the non-negative input matrices or profiles corresponding to observational and experimental data, respectively. As there are profiles that can be observed for one or both datasets, let $\mathcal{L}$ be the set of profiles that are common to both datasets, $\mathcal{L} = \{1, \ldots, |\mathcal{L}|\}$, where $|\mathcal{L}|$ is the number of elements in $\mathcal{L}$. For the matrices that are in one dataset, but not in the other, let $\mathcal{I}$ be the set of profiles in $\{1, \ldots, M_X\}$ that are specific to observational data, and $\mathcal{J}$ the set of profiles that are specific to experimental data, but starting in $M_X + 1$, that is, $\mathcal{J} = \{M_X + 1, \ldots, (M_X + 1) + M_Y - |\mathcal{L}|\}$. When all matrices are shared, $\mathcal{I}$ and $\mathcal{J}$ are empty sets. The total number of matrices is $2|\mathcal{L}| + |\mathcal{I}| + |\mathcal{J}|$ (Figure 1). Note that this approach accepts unobserved matrices in one of the datasets so that they could be estimated. If there are no observed matrices X_J, where $J \in \mathcal{J}$, then these could be estimated as $\hat{X}_J = W_{obs} H_J$.

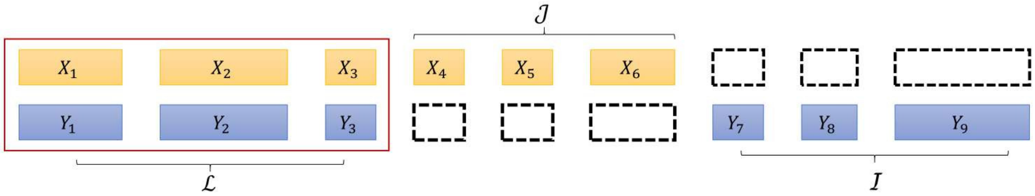

Figure 1: Representation of the pairing of observational and experimental profiles.

Accordingly, we propose a Multi-project and a Multi-profile joint Non-negative Matrix Factorization (M&M-jNMF) to the solution of the simultaneous non-negative factorization of matrices as the H_L for $L \in \mathcal{L}$, H_I^X for $I \in \mathcal{I}$, H_J^Y for $J \in \mathcal{J}$, W_{obs} and W_{exp} that minimizes the following expression:

$$\begin{aligned} &\sum_{L \in \mathcal{L}} \left(\|X_L - W_{obs} H_L\|_F^2 + \|Y_L - W_{exp} H_L\|_F^2 \right) \\ &+ \sum_{I \in \mathcal{I}} \|X_I - W_{obs} H_I^X\|_F^2 + \sum_{J \in \mathcal{J}} \|Y_J - W_{exp} H_J^Y\|_F^2 \end{aligned} \tag{2}$$

Maintaining constant the matrices H_L for both datasets implies that the basis (representative centers of each cluster) for objects are the same; therefore, it is possible to cluster the groups of patients and cell lines (co-clusters). These co-clusters create a new integration on which different sets of individuals may be related among several dimensions of measurements.

2.3.3. M&M-jNMF with prior knowledge constraints

Besides the four terms that define the objective function in Equation 2, we considered constraints on matrices H_I that can included prior knowledge in the model (Section 2.2) and, at the same time, work as regularization terms that help to achieve sparse and stable solutions [9].

In terms of Equation 2, the set $\mathcal{L}$ corresponds to three omic profiles (Section 2.1), whereas $\mathcal{I}$ is empty for TCGA data (X_I) and $\mathcal{J}$ corresponds to the drug profile (Y_D) that is only observed for CCLE data (Y_I) (Figure 2).

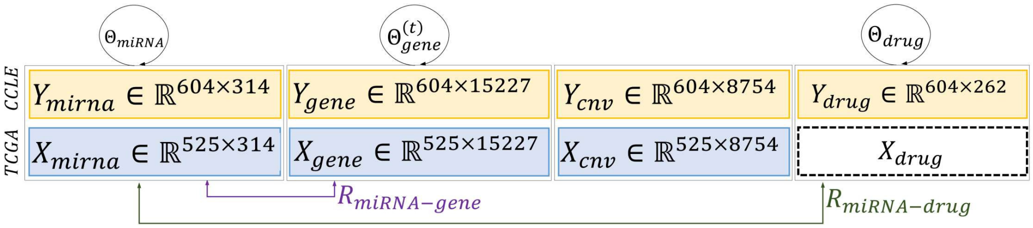
Θ_{miRNA}
$\Theta^{(t)}_{gene}$
Θ_{drug}
CCLE
TCGA
$Y_{mirna} \in \mathbb{R}^{604\times314}$
$Y_{gene} \in \mathbb{R}^{604\times15227}$
$Y_{cnv} \in \mathbb{R}^{604\times8754}$
$Y_{drug} \in \mathbb{R}^{604\times262}$
$X_{mirna} \in \mathbb{R}^{525\times314}$
$X_{gene} \in \mathbb{R}^{525\times15227}$
$X_{cnv} \in \mathbb{R}^{525\times8754}$
X_{drug}
$R_{miRNA-gene}$
$R_{miRNA-drug}$

Figure 2: Scheme of the integration of different data sources (TCGA and CCLE). M&M-jNMF requires the same variables among projects; therefore, the dimensions of each profile are equal to the number of samples (patients or cell lines) × the number of variables (genes, miRNAs, or drugs). Omic profiles are associated with their respective Θ_I and R_{IJ} constraints. The drug sensitivity profile for patients is not available.

Therefore, we proposed the solution to the following optimization problem (Equation 3):

$$\begin{aligned}
&min\, F(W_{tcga}, W_{ccle}, H_1, \ldots, H_I, H_D) = \\
&\sum_{I\in\mathcal{L}} \left[\|X_I - W_{tcga}H_I\|_F^2 + \|Y_I - W_{ccle}H_I\|_F^2\right] + \|Y_D - W_{ccle}H_D\|_F^2 \\
&-\lambda_1 \sum_{I\in\mathcal{L}} \sum_t Tr(H_I {\Theta^{(t)}}_I H_I^T) - \lambda_2 \sum_{(I,J\in\mathcal{L}:\, I\neq J)} Tr(H_I R_{IJ} H_J^T) \\
&-\lambda_1 \sum_t Tr(H_D \Theta_D^{(t)} H_D^T) - \lambda_2 \sum_{J\in\mathcal{L}} Tr(H_D R_{DJ} H_J^T) \\
&+\gamma_1 \|W_{tcga}\|_F^2 + \gamma_2 \|W_{ccle}\|_F^2 + \delta_1 \sum_I \sum_j \|h_j^I\|_1^2
\end{aligned} \tag{3}$$

where $\|.\|_F^2$, $\|.\|_1^2$, and $Tr\,(\cdot)$ denote Frobenius norm, L_1 norm and trace, respectively. The index D corresponds to the drug profile. In Equation 3, the first three terms correspond to the individual factorization of the observational and experimental data. The next four terms are associated with prior knowledge, also known as regularization graphs, which are explained in Section 2.2. The last three terms are regularization that controls the sparsity on the H_I matrices and the scale on the W_{tcga} and W_{ccle} matrices.

2.3.4. Multiplicative update rule algorithm

The objective function $F(\cdot)$ described in Equation 3 is not convex for all parameters simultaneously. In particular, for the optimization problem for jNMF proposed by [9], the solution implies an iterative procedure that updates by the group of H_I matrices, or for W at each step, with other variables fixed while the others are updated. These alternating algorithms are NP-problems that do not guarantee a global optimal but a local optimal [20].

We developed and implemented the multiplicative update rules (MUR) algorithm as described by [9] (Equations 4 to 8):

$$(w_{tcga})_{ij} \leftarrow (w_{tcga})_{ij} \times \frac{(\sum_I X_I H_I^T)_{ij}}{(\sum_I W_{tcga} H_I H_I^T + \gamma_1 W_{tcga})_{ij}} \tag{4}$$

$$(w_{ccle})_{ij} \leftarrow (w_{ccle})_{ij} \times \frac{(\sum_I Y_I H_I^T)_{ij}}{(\sum_I W_{ccle} H_I H_I^T + \gamma_1 W_{ccle})_{ij}} \tag{5}$$

$$h_{ij}^I \leftarrow h_{ij}^I \times \frac{(W_{tcga}^T X_I + W_{ccle}^T Y_I + \lambda_1/2 \sum_t H_I(\Theta_I + (\Theta^{(t)})^T) + \lambda_2/2 \sum_{I \neq J} H_J R_{IJ}^T)_{ij}}{((W_{tcga}^T W_{tcga} + W_{ccle}^T W_{ccle} + \delta_1 e_{(K \times K)}) H_I)_{ij}} \tag{6}$$

$$h_{ij}^I \leftarrow h_{ij}^I \times \frac{(W_{tcga}^T X_I + W_{ccle}^T Y_I + \lambda_1/2 \sum_t H_I(\Theta_I + (\Theta^{(t)})^T) + \lambda_2/2 \sum_{I \neq J} H_J R_{IJ}^T)_{ij}}{((W_{tcga}^T W_{tcga} + W_{ccle}^T W_{ccle} + \delta_1 e_{(K \times K)}) H_I)_{ij}} \tag{7}$$

$$h_{ij}^D \leftarrow h_{ij}^D \times \frac{(W_{ccle}^T Y_D + \lambda_1/2 \sum_t H_D(\Theta_D + (\Theta^{(t)})^T) + \lambda_2/2 \sum_{D \neq J} H_J R_{DJ}^T)_{ij}}{((W_{ccle}^T W_{ccle} + \delta_1 e_{K \times K}) H_D)_{ij}} \tag{8}$$

In Equation 7 and 8, $e_{K \times K}$ is a matrix of $K \times K$ dimensions, where the element's value is set to 1. The stop criterion for the algorithm was proposed by [9], where a relative measure was calculated between the results of two consecutive iterations; in our case, τ, the stopping threshold was set to 10^{-7}. The formulation of the stop criterion is $\frac{F_t - F_{t+1}}{F_0 - F_{t+1}} \leq \tau$, where F indicates the objective function evaluated at iteration 0, t or $t+1$ with their respective matrices (W_{tcga}, W_{ccle} and H_I).

2.3.5. Hyperparameters selection

From Equation 3, a total of six hyperparameters were defined: k, λ_1, λ_2, γ_1, γ_2 and δ_1. For k, we set values of 30, 60, and 90 which represent a range where information can be concentrated ($k = 30$) or dispersed ($k = 90$). For the other hyperparameters, we set them between the range values of 0 to 10. Using this range, we could explore the strength of the penalty, i.e., strong (10) or null (0). In the case of the hyperparameters γ_1, γ_2 and δ_1 a value equal to zero nullified the term to be penalized, while high values generated values close to zero in the W_{tcga}, W_{ccle} and H_I matrices. For the hyperparameters λ_1 and λ_2, a high value gives much importance to the multiplying terms (prior knowledge) since they are subtracting in the objective function.

We performed two iterations of the MUR algorithm, and we calculated four metrics using the model outputs to choose the best set of hyperparameters. These metrics include:

1. the Sum of Squares of the Residuals, RSS, ($\|X_I - W_{tcga}H_I\|_F^2$ or $\|Y_I - W_{ccle}H_I\|_F^2$)
2. the Cophenetic correlation coefficient (ρ) calculated by [21]. Among MUR runs, it may not converge to the same solution. So for several runs, this metric reflects the probability that observations i and j are grouped in the same cluster. Therefore, this coefficient measures the reproducibility of the assignment of the observations in each cluster [21].
3. measures of cluster enrichment: the ratio of enriched gene clusters, the number of enrichment terms identified, and the number of patient groups.
4. an adjusted version of R^2 which was defined as:

$$R^2_{adjusted} = 1 - \frac{\|X_I - W_{tcga}H_I\|_F^2 \times N_I}{\|X_I\|_F^2 \times [N_I - k\,(p\,+\,m_I)]} \tag{9}$$

where m_I is the number of variables, p is the number of samples, and N_I is defined by $p \times m_I$ for the profile I; the parameter $k\,(p\,+\,m_I)$ refers to the estimated number of parameters.

We chose the optimal set of hyperparameters that meet the following criteria: the sum of squares of the residuals was as small as possible, and $R^2_{adjusted}$ and ρ were close to 1. In addition, the ratio of enriched gene clusters must be close to 1, the number of enriched terms must be as large as possible, and the number of patient groups should contain a representative sample of patients since this allows the finding of molecular markers between these groups.

2.4. Co-cluster assignment rule

The matrices H_I and W contain the latent structure to cluster molecules and objects (patients or cell lines), respectively [10, 7]. We found that using the standard assignment method, which assigns a molecule to a cluster (k) if its value inside this cluster exceeds a threshold, can incur redundant clusters by including molecules with high weights in several clusters. Therefore, we first detected the maximum values of each molecule in each cluster, then using the 50th quartile of this set of values, we choose the highest values to be included in cluster k. We repeated the same procedure with the second maximum value of each molecule, but we used the 75th quartile to select molecules (Supplementary Section S3). In addition, the clusters of each omic profile can be grouped to obtain a co-cluster. For example, the first co-cluster will gather the clusters assembled in the first cluster of each profile, e.g., gene, miRNA, CNV, and drug profiles.

It is also possible to determine co-clusters for objects (patient or cell line) by using W_{tcga} and W_{ccle}. Using matrix W, the column k where the maximum value for object i is found corresponds to the assignment cluster for that object. Since some clusters contained very few samples, which led to a problematic comparison, we decided to reassign these samples to clusters containing more samples whether the value of the cluster for a particular sample was close to the mean of another cluster (Supplementary Section S3). We used the terms groups and clusters interchangeably.

2.5. Matrix comparison between TCGA and CCLE

In evolutionary theory, the quantitative traits of a population expressed as G-matrix (covariance matrix) have been used to determine the inheritance of genetic or phenotypic traits between populations. Therefore, the comparison of G-matrices of two particular populations requires a similarity metric. PCASimilarity measures the degree to which the eigenvectors of both matrices span the same space and considers the amount of variation that each population has in that direction (eigenvalues). PCASimilarity can take a value of 1 when there is a high similarity between the two matrices; otherwise, it will take a value close to or equal to 0 [22].

Accordingly, we measured the degree of similarity between the co-variance matrices of omic profiles for the patient and cell lines clusters (Section 2.4) by calculating the PCASimilarity score. We calculated the PCASimilarity score using Equation 10.

$$PCASimilarity(A, B) = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} \lambda_i^A \lambda_j^B cos^2(\Lambda_i^A, \Lambda_j^B)}{\sum_{i=1}^{n} \lambda_i^A \lambda_j^B} \tag{10}$$

where A and B are input matrices, λ_i^A and Λ_i^A are the ith eigenvalue, and the ith principal component of matrix A, respectively. This procedure is analogous to matrix B.

2.6. Biological interpretation

Using ClusterProfiler package v3.14.3 [23] we performed a Gene Ontology analysis and a signaling and metabolic pathways analysis using the KEGG database. In addition, for the CNV and miRNA profiles, we conducted a literature review. For miRNA analysis, we used mirNet v2.0 [24].

cBioportal was employed to understand and compare the groups of patients and cell lines [25, 26]. In addition, we reviewed the literature on comparison between TCGA and CCLE data [27, 12]. For Kaplan-Meier curves and biological, clinical classifications, we employed the results from [28] and [29], respectively.

For Significance Analysis Microarrays, we used the samr package v3.0. This method uses repeated permutations of the profiles to determine if any gene or miRNA are significantly in two unpaired conditions [30]. In addition, the method uses the False Discovery Rate and $q - value$ method.

2.7. Synthetic data and evaluation metrics

We created two artificial datasets, S_1 and S_2, which have three common profiles ($\mathcal{L}$). We generated the base matrices for the two sources from a uniform distribution $(0, 1)$ with $n \times k$ dimensions, where n is the observations and k the range of the matrix W_S. Similarly, H_I were obtained from a uniform distribution $(0, 1)$ with $k \times m_I$ dimensions, where m_I is the number of variables in each matrix (Supplementary Section S4). Then, we calculated the original matrices as $X_S = W_S H_I + \epsilon$ where ϵ is an error term. Finally, we created Θ and R constraints matrices as a sparse binary matrix with appropriate dimensions for each profile.

We defined associations of observations as pairs between two samples in a cluster; for example, in cluster k, there are ten pairs or associations if there are five patients. To measure the capability of our method to detect the original associations in the predicted matrices, we calculated the metrics F1-Score, Recall, and Precision.

2.8. Implementation

We generated a project in Spyder v4.2.1 using Python v3.6 to implement the algorithm. We used R v4.0.4 to download the data and create the constraint matrices.

3. Results

3.1. Simulation study

We evaluated the ability of the M&M-jNMF algorithm to identify the original dimensionality (k) of the W and H^T matrices. Also, we identified the associations between objects or between variables. As defined in Section 2.7, we defined a dimensionality of $k = 5$, and we pre-defined the clusters for objects and variables. We evaluated different sets of hyperparameters and compared their performance using the metrics $R^2_{adjusted}$, F1-Score, Recall, and Precision (Supplementary File S1). When comparing the metrics at different k evaluated, we found that the method performed well at $K = 6$. However, the ranges (k) greater than 20 did not obtain good approximations since the values of the metrics were below 0.7. This result indicated that the information is not well distributed in the low-rank matrices obtained. When we used k close to the real one, the method correctly represented the original X matrices because we found that $R^2_{adjusted}$ was greater than 0.7. For the classification metrics, we found that the precision was above 0.9 for the values of $k = 5$, 6, and 7, indicating that at other k, the erroneous associations increase. The Recall followed similar behavior and showed that our method correctly detected the original associations in k close to 5 (Figure 3, and Supplementary Figure F1).

3.2. Hyperparameter selection for M&M-jNMF method

In Figure 2, we show a representation of the observational data (TCGA), experimental (CCLE) data, and the constraints used as input for the integration with the M&M-jNMF method. To ensure a representative latent structure of input matrices, we added the hyperparameters (γ_1, γ_2, and δ)

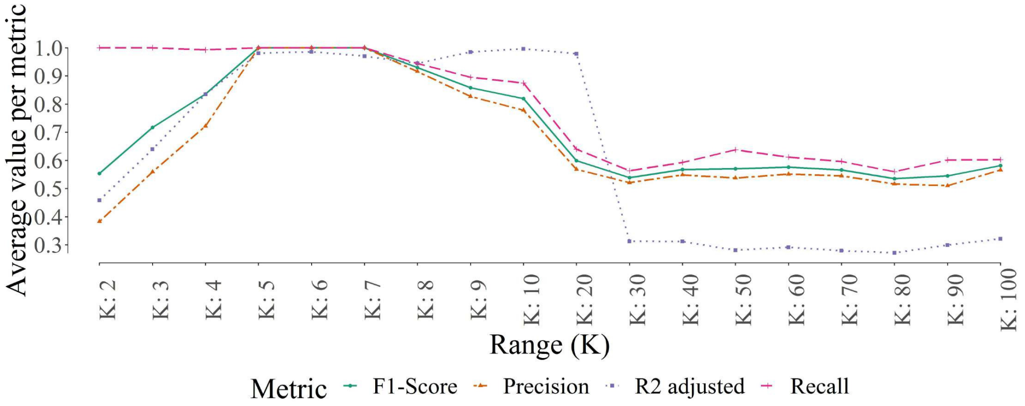

Figure 3: Performance of the M&M-jNMF method to determine the correct dimensionality, and to identify the correct clusters of variables. At large ranges (k), the performance of clustering decreases, while k close to the real one improves the approximation of the original datasets and clusters.

which control the scale of the values of the W_{tcga} and W_{ccle} matrices, and the sparsity in the H_I matrix (Equation 3). In addition, the hyperparameters (λ_1 and λ_2) control the importance of the prior knowledge represented in the Θ and R constraints. As we described in Section 3.2, we selected the best set of hyperparameters from a defined range of values (Supplementary File S2).

The best set of hyperparameters that we found was: $k = 60$, $\gamma_1 = \gamma_2 = 3.5 \times 10^{-6}$, $\delta_1 = 3.5 \times 10^{-3}$ and $\lambda_1 = \lambda_2 = 10$. Using these hyperparameters, we obtained an $RSS_{tcga} = 32890.19$ and $RSS_{ccle} = 46693.95$ that were among the lowest of the other hyperparameter sets tested, $R^2_{adjusted}$ and ρ was over 0.88, which means there was a good representation of the original matrices and stability of the clusters, respectively (Supplementary Figure F2). In addition, we applied the rule reassignment defined in Section 2.4 on W_{tcga}, and we identified 7 patient clusters (Supplementary File S3). The convergence time for MUR was approximately one-half hour using a 2.20GHz Intel Corei7 processor with 16GB RAM.

Concerning Θ and R constraints, although we observed an additive effect of these constraints on the objective function, they had no relevant effect on the clustering of the molecules (data not shown). However, when λ_1 and λ_2 values were greater than 100, the objective function converges very slowly. Thus, we believe our constraint matrices are very sparse and require more information on the associations between molecules, or perhaps they work better with lower-dimensional data as was the case in [9].

3.3. Gene enrichment analysis of gene clusters shows significant biological processes related to glioma

We performed an enrichment analysis (Section 2.6) on the gene clusters. As other studies have shown, the clusters obtained from variants of jNMF are enriched in biological terms and are associated with different cancer processes [31, 5]. For the 60 gene clusters, 53 clusters were enriched in biological processes (BP), 44 clusters were enriched in molecular functions (MF), and 54 clusters were enriched in cellular components (CC) ($p - value < 0.05$, Supplementary File S4).

We identified 49 gene clusters highly enriched in 185 KEGG terms ($p - value < 0.05$, Supplementary Figure F3, and Supplementary File S4). In these enrichment terms, we identified five main categories, which we manually generated according to their relationship. The first group contains amino acids, fatty acids and, simple and complex glycans. The second group contains groups of neurotransmitters and related biological processes, such as calcium signaling pathways. A third group contains signaling pathways involved or related to cancer, such as the cAMP signaling pathway, cGMP-PKG signaling pathway, p53 signaling pathway, and cell cycle. The fourth group contains terms related to immune system response, for example, inflammatory mediators and genes related to infection processes. The last group contains terms related to extracellular matrix terms, such as focal adhesion, axon guidance, and extracellular matrix-receptor interaction (Supplementary File S4).

3.4. M&M-jNMF method clusters patients into relevant clinical groups

From the W matrix, the objects can be assigned into groups or clusters. For example, [31] used the W matrix from a non-constrained jNMF solution to cluster patients; these groups were similar to the existing clinical categories of ovarian cancer.

Similarly, we obtained 7 groups (I-VII) of patients using the W_{tcga} matrix (Figure 4). We found that these groups have a significant separation for survival curves (log-rank test $p-value = 8.83 \times 10^{-224}$). In addition, we compared them to the clinical classifications for LGG to obtain a deeper analysis [29]. The clinical classifications contain a molecular and epigenetic status (IDH status, MGMT promoter status, TERT promoter status, and ATRX status). Interestingly, we identified that our groups mainly separate into clinically defined groups, but our method was able to identify new groups. The survClust method obtained similar results [32], where it found five groups, 3 of them correlated with the three clinical LGG classifications. In our case, the groups represent the LGG clinic classification, where group I corresponds to patients with IDH mutation, 1p/19q co-deletion, MGMT methylated, and ATRX wild type. These molecular characteristics have been associated with higher rates of survival [29]. Groups II and IV represent mainly IDHmut-non-codel, and groups III and V represent a mixture of IDHmut-non-codel and IDH wild-type subtypes. At the same time, groups VI and VII represent mostly IDH wild-type patients. Finally, groups I-V represent mostly MGMT methylated promoter (Figure 5, and Supplementary File S5).

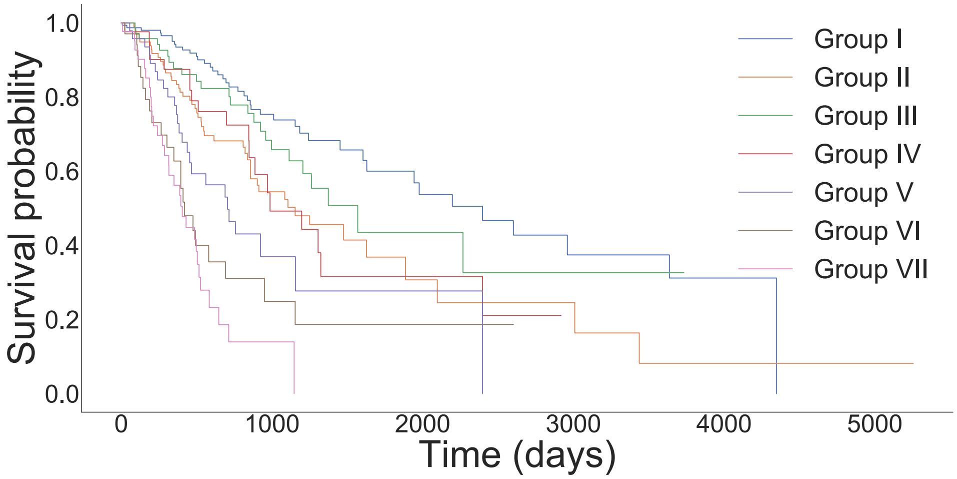

Figure 4: Progression-free interval plot for cluster patients. The number of patients per cluster is: I (164), II (103), III (81), IV (46), V (48), VI (38), and VII (44) (Log-rank test $p - value = 8.83e^{-224}$).

Using cBioportal tool [26, 25], we found 9366 genes expressed differentially between the cluster of patients ($p - value < 0.05$, Supplementary File S6, and Supplementary Section S5). Among these genes (Figure 5), we highlight TRIM67 (Group I), ADAMTS20 (Group II), TESPA1 (Group III), TPTEP1 (Group IV), GJB1 (Group V), POSTN (Group VI), and MEOX2 (Group VII) which had a deferentially level of expression than the other groups. These genes have been related with the progression of apoptosis

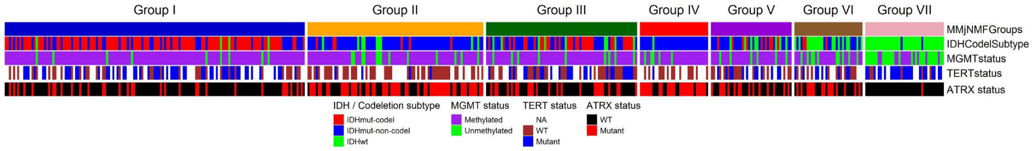

Figure 5: Comparison of M&M-jNMF clusters and clinical and molecular classification for LGG.

[33], angiogenesis in cancer [34], chemoresistance [35], radioresistance [36], or as a prognostic biomarkers [37].

Since our method generates groups of patients by their molecular characteristics according to the integrated omic profiles, we wanted to show the differences between these groups. We were interested in the difference between groups VI and VII, despite having a very similar survival curve. Using cBioportal, we found that there are differences in genes, proteins, and DNA methylation. For example, a high expression of the FBLIM1 gene, a low expression of the CCDND1 protein, and high methylation of the AQP4 gene in group VI. These genes and this protein may enhance drug sensitivity [38] or reduce tumor cell viability [39], i.e., increase the probability of survival.

3.5. There are metabolic and signaling similarities between patients and cell lines

Some studies have compared TCGA and CCLE projects to identify cell lines for pre-clinical and pharmacological purposes [12, 15]. Despite this, there are many differences which difficult this comparison between primary tumors and cell lines because the former contains a mixture of cells (tumor cells, immune cells, and stromal cells), and the latter has a more significant number of genomic alterations [40, 41]. [27] proposed a methodology using weights to match cell lines and tumors according to the similarity in different contexts such as signaling pathways or mutations. Despite these differences, it is possible to identify possible biological traits between these two projects [40, 27]. Similarly to [27], we propose a metric to compare between groups of cell lines and groups of tumors in specific contexts such as alterations in signaling and metabolic pathways and gene ontology enrichment.

Therefore, we compared the 7 groups of patients (I-VII) and the 9 groups of cell lines (1-9) (Section 2.4, and Supplementary File S3). For that, we calculated a PCASimilarity score to find similar biological traits between cell lines and tumors (Supplementary File S7).

This strategy allowed us to characterize similarities in different biological processes such as molecular functions (MF), cellular components (CC), biological processes (BP), and KEGG pathways. As expected, there is much diversity among cell lines and patients (Figure 6, and Supplementary Section S6). We identified that the proportion of enriched terms with a PCASimilarity greater than 0.85 corresponded in most cases to associations of group 5 and each of the tumor groups (29%, test for equality of proportions $p-value < 0.001$). We depicted these results in Figure 6. For example, group I of tumors has 34 similar enriched terms with group 5 of cell lines, including DNA mismatch repair, covalent chromatin modification, and GTPase regulator activity. Whereas, for example, for group I of tumors vs. group 8 of cell lines, only 17 terms were found. Therefore, cell line group 5 has a higher similarity for all groups of tumors, but to a lesser degree for tumor group V, which had only 16 similar enriched terms (Supplementary Section S6). This result is relevant because this group includes glioma cell lines (5.6%) whose sample type is entirely from primary tumors (Supplementary File S8 and Supplementary Section S6). Thus, we found similar results obtained by [40] who compared CCLE cell lines with all TCGA cancer types. They used gene expression profiles and found that LGG had a high correlation with glioma cell lines. In addition, our method agreed with the results obtained by [40] because 25 cell lines found by their analysis (correlation coefficient > 0.48) correspond to cell lines that we classified in group 5. For other omic profiles (CNV and miRNA), group 5 of cell lines has more terms related to tumors groups than the other cell lines groups (Supplementary Figure F4).

3.6. Drug repurposing is also associated with specific genetic and miRNA signatures

We estimated the patient drug sensitivity profile as $X_{Drug} = W_{tcga}H_{Drug}$ (Supplementary File S9). Since we used the drug sensitivity profile of cell lines (AUC), then the calculated profile for LGG tumors is an approximation of the potency and efficacy of the 262 drugs for tumors. In the CCLE project, low AUC values correspond to sensitivity to a drug, or also a reduction in cell viability [42, 43].

The predicted matrix contains the degree of drug sensitivity a tumor may have. For each drug, we defined regions where the degree of sensitivity corresponds to resistance or sensitivity. For column j^{th} in this profile, we defined a resistant tumor for observations whose sensitivity value is above

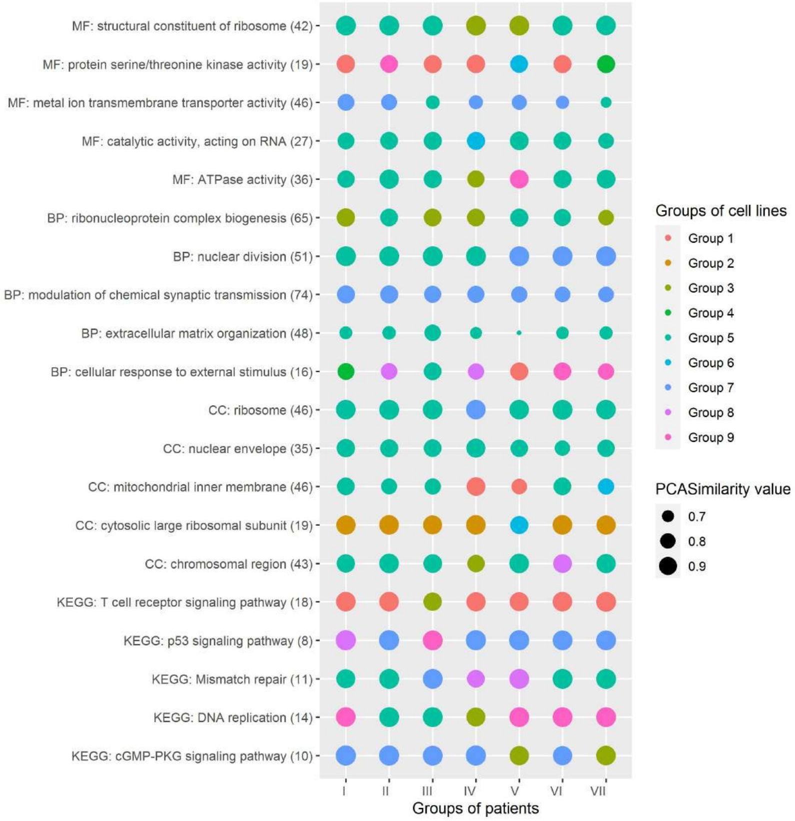

Figure 6: Similarity score between CCLE and TCGA matrices. The PCASimilarity score must be close to one for the correlation matrices to be similar. Here we present enriched gene clusters according to their relation or importance to the biological processes of cancer.

the 75th quartile, while we considered them as sensitive if their value is below the 25th quartile.

We performed a Significance Analysis of Microarrays for miRNA and gene profiles between tumors classified as resistant or sensible to find gene or miRNA signatures that could be related to the drug sensitivity (Section 2.6). In Supplementary File S10, we reported genes and miRNAs differentially expressed between resistant and sensitive tumors.

We used the patterns in the differential expression of genes and miRNAs to assess whether there is a relationship between these patterns and drug sensitivity. We realized that these signatures might contain: (i) molecules related to the mechanism of drug sensitivity, (ii) molecules indirectly related to mechanisms that compromise the viability of the tumor, or (iii) molecules characteristic of other biological processes of the tumor (Supplementary Section S7). We analyzed two drugs, Temozolomide (TMZ) and Shikonin. For TMZ, we analyzed the first ten genes and miRNAs with the most significant fold change of expression (resistant vs. sensitive) to identify known markers for sensitivity to this drug. For Shikonin, we performed enrichment analysis on the low/high expressed genes, and we analyzed the context in which the tumor might be sensitive to this drug.

Firstly, we analyzed the predicted drug sensitivity profile for TMZ, a standard drug in glioma therapy, because there is evidence of mechanisms of sensitivity [44]. Interestingly, the patterns of gene and miRNA expression profiles agree with experimental evidence (see TMZ in Supplementary File S11). We found 4 of the ten miRNAs analyzed associated with increased sensitivity to TMZ; these are miR-34a, miR-301a, miR-146a, and miR-126-3p [45, 46]. In the case of the genes, we highlight the low expression of FBXO44, which its inhibition induces replication stress and DNA strand breaks in cancer cells, reducing tumor growth [47]. In addition, we found a high expression of the gene CREB3L1 in sensible tumors; this gene is a suppressor of metastasis, which is involved in the Unfolded Protein Response (UPR). Its functional activation in this response generates a cytoprotective effect, but if it fails to mediate, it leads to apoptosis [48]. Recently, the expression of this gene has been correlated with a better prognosis in low and high-grade gliomas [49]. This pattern is an example of how genes and miRNAs can be associated directly with TMZ sensitivity or indirectly by impairing tumor viability.

Secondly, we analyzed the predicted Shikonin sensitivity profile because we found that its genes have the greatest difference between resistant and sensible tumor groups. Shikonin is a compound extracted from the root of *Lithospermum erythrorhizon*. The active compound has an anti-cancer and anti-adipogenic effect. The molecular mechanism involves the suppression of Tumor Necrosis Factor-alpha (TNF-α), decreased phosphorylated levels of EGFR, ERK1/2, and protein tyrosine kinases [50, 51]. The anti-glioma effect has been suggested to interfere with endoplasmatic reticulum (ER) stress-mediated tumor apoptosis [51]. The high expressed genes show that the sensible tumor may be related with an active process in the endoplasmic reticulum ($p - value = 2.13 \times 10^{-3}$), e.g., STAB1, RAB13, REEP4 genes in 7, with a low expression of genes related to neurotransmitter processes ($p - value = 6.33 \times 10^{-9}$), e.g., SLC17A7, SYN2 genes in Figure 7. Apparently, the expression of miRNAs is associated with an aggressive tumor type, which has a high expression of onco-miRNAs such as miR-18a, miR-19a, miR-21, miR-155, miR-196a, miR-210, among other ($p - value = 3.5 \times 10^{-7}$), and low expression of miR-379 cluster ($p - value = 8.53 \times 10^{-55}$), and miR-212 cluster ($p - value = 9.97 \times 10^{-7}$). The former is a cluster with an important role in glioblastoma, also known as C14MC, which has been related positively to prognosis in glioma [52]. However, the pattern includes the high expression of miR-200c/miR-141 cluster ($p - value = 3.80 \times 10^{-6}$) in sensitive tumors. For this miRNA cluster, it has been recently evidenced that it has anti-oncogenic roles in glioma; exactly, its target genes are Moesin, VEGF, HIF-1α, MMP2, ZEB1, which participate in the processes of progression and metastasis in cancer [53]. In general, low expression of this miRNA is associated with high categories of glioma [54]. Therefore, the pattern found by us indicates that the increase of this miRNA may have a combined effect on the anti-angiogenic mechanism of Shikonin. Transfection of miR-200c in glioma cells has shown a cytotoxic effect of radiotherapy since attenuation of EGFR-mediated signaling-associated pro-survival signaling, and impaired DNA damage repair has been observed [55]. For this reason, a first-line therapy such as radiotherapy in glioma could be accompanied by Shikonin treatment when this pattern is present in a tumor (Figure 7).

4. Discussion and conclusion

In this study, we proposed M&M-jNMF as a new method for integrating omic data from different projects (TCGA and CCLE) or between different types of cancer of the same project to compare similarities in metabolisms or signaling pathways. Based on jNMF, this method allowed us to integrate omic profiles and perform clustering and co-clustering between molecules and between patients and cell lines. An essential advantage of the proposed method is that we integrated the two projects but maintaining the difference between them, i.e., each project is allowed to have its basis matrix (W_{tcga} and W_{ccle}), but at the same time, it is integrated according to the omic profiles (H_I). Because of this, we identified clusters enriched in ontological terms related to cancer and stratified patients and cell lines. The latter allowed us to compare clusters between cell lines and patients to match them and propose functional cell lines for the pre-clinical phase study of drugs.

We applied the M&M-jNMF method to omic LGG data obtained from the TCGA and cancer cell lines from the CCLE projects. We identified 7 groups

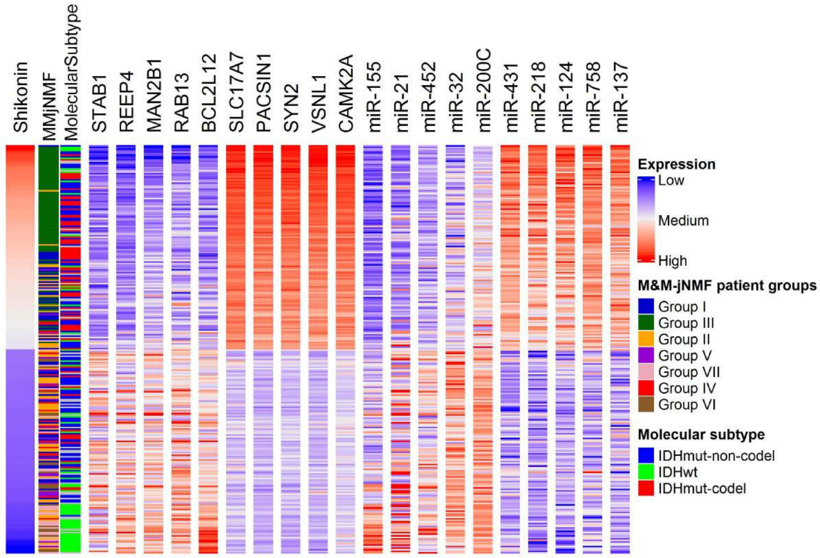

Figure 7: miRNA and gene markers for the sensitivity profile of Shikonin. For the Shikonin bar, red means resistant tumor group and blue represents the sensible tumor group.

for patients and 9 groups for cell lines. For the patient groups, we evidenced that they are similar to the clinical classifications currently available for this type of cancer. Our method identified some groups with marked differences in survival time or the presence of molecular markers such as MGMT promoter or ATRX status. However, the distribution of the groups is very similar to the clinical classification of the IDH status and 1p/19q co-deletion (Figure 4).

In addition, we identified biological similarities between cell line groups and tumor groups. The importance of this result is because the search for cell lines that match patient's tumors is complex. After all, the cell lines have a very high mutation rate [15]. For this reason, we decided to employ another strategy to identify patterns in the signaling pathways by comparing the similarity between the groups of cell lines and patients obtained by our method [27]. We used the PCASimilarity score, which yielded a set of cell lines with similar gene expression patterns between group 5 cell lines and most patient groups. This group contains some glioma-specific cell lines (Supplementary Section S6).

Finally, we performed a repositioning of 262 drugs, considering that we predicted the possible response to drug-only treatment. Thanks to this, we identified genetic patterns to establish when a tumor might be sensitive to a drug. In general, these signatures correspond in some cases to molecules that may indicate a direct association with the sensitization mechanism, e.g., in the case of TMZ. Nevertheless, also, some signatures may indicate mechanisms by which the tumor may be vulnerable and favor treatment, as we showed to Shikonin (Figure 7).

Our method can identify new strategies to address drug repositioning issues, identify clusters given their omic profiles, and search for cell lines suitable for pre-clinical drug testing.

Funding

This work has been supported by the Ministry of Science, Technology and Innovation of Colombia grant No. 785. The European Research Council (ERC) Consolidator Grant 770827 and the Spanish State Research Agency AEI 10.13039/501100011033 grant number PID2019-105500GB-I00.

Conflict of Interest: none declared.

References

[1] Giovanna Nicora, Francesca Vitali, Arianna Dagliati, Nophar Geifman, and Riccardo Bellazzi. Integrated Multi-Omics Analyses in Oncology: A Review of Machine Learning Methods and Tools. *Frontiers in Oncology*, 10:1030, jun 2020.

[2] Indhupriya Subramanian, Srikant Verma, Shiva Kumar, Abhay Jere, and Krishanpal Anamika. Multi-omics Data Integration, Interpretation, and Its Application. *Bioinformatics and Biology Insights*, 14:1–24, jan 2020.

[3] Marinka Zitnik, Francis Nguyen, Bo Wang, Jure Leskovec, Anna Goldenberg, and Michael M Hoffman. Machine learning for integrating data in biology and medicine: Principles, practice, and opportunities. *Information Fusion*, 50:71–91, oct 2019.

[4] Sijia Huang, Kumardeep Chaudhary, and Lana X. Garmire. More Is Better: Recent Progress in Multi-Omics Data Integration Methods. *Frontiers in Genetics*, 8:1–12, jun 2017.

[5] F Vitali, S Marini, D Pala, A Demartini, S Montoli, A Zambelli, and R Bellazzi. Patient similarity by joint matrix trifactorization to identify subgroups in acute myeloid leukemia. *JAMIA Open*, 0(0):1–12, 2018.

[6] Vladimir Gligorijevic, Noel Malod-Dognin, and Natasa Przulj. Patient-specific data fusion for cancer stratification and personalised treatment. *Biocomputing 2016*, 21:321–332, 2016.

[7] Zi Yang and George Michailidis. A non-negative matrix factorization method for detecting modules in heterogeneous omics multi-modal data. *Bioinformatics*, 32(1):1–8, 2015.

[8] Marinka Zitnik and Blaz Zupan. Data Fusion by Matrix Factorization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(1):41–53, jan 2015.

[9] Lihua Zhang and Shihua Zhang. A General Joint Matrix Factorization Framework for Data Integration and Its Systematic Algorithmic Exploration. *IEEE Transactions on Fuzzy Systems*, 28(9):1971–1983, sep 2020.

[10] Shihua Zhang, Qingjiao Li, Juan Liu, and Xianghong Jasmine Zhou. A novel computational framework for simultaneous integration of multiple types of genomic data to identify microrna-gene regulatory modules. *Bioinformatics*, 27(13):401–409, 2011.

[11] Naoya Fujita, Shinji Mizuarai, Katsuhiko Murakami, and Kenta Nakai. Biomarker discovery by integrated joint non-negative matrix factorization and pathway signature analyses. *Scientific Reports*, 8(1):1–10, 2018.

[12] K. Yu, B. Chen, D. Aran, J. Charalel, C. Yau, D. M. Wolf, L. J. van ‘t Veer, A. J. Butte, T. Goldstein, and M. Sirota. Comprehensive transcriptomic analysis of cell lines as models of primary tumors across 22 tumor types. *Nature Communications*, 10(1):3574, dec 2019.

[13] Theodore Sakellaropoulos, Konstantinos Vougas, Sonali Narang, Filippos Koinis, Athanassios Kotsinas, Alexander Polyzos, Tyler J. Moss, Sarina Piha-Paul, Hua Zhou, Eleni Kardala, Eleni Damianidou, Leonidas G. Alexopoulos, Iannis Aifantis, Paul A. Townsend, Mihalis I. Panayiotidis, Petros Sfikakis, Jiri Bartek, Rebecca C. Fitzgerald, Dimitris Thanos, Kenna R. Mills Shaw, Russell Petty, Aristotelis Tsirigos,

and Vassilis G. Gorgoulis. A Deep Learning Framework for Predicting Response to Therapy in Cancer. *Cell Reports*, 29(11):3367–3373, dec 2019.

[14] Lin Wei, Zhilin Jin, Shengjie Yang, Yanxun Xu, Yitan Zhu, and Yuan Ji. TCGA-assembler 2: software pipeline for retrieval and processing of TCGA/CPTAC data. *Bioinformatics*, 34(9):1615–1617, may 2018.

[15] Mahmoud Ghandi, Franklin W. Huang, Judit Jané-Valbuena, Gregory V. Kryukov, Christopher C. Lo, E. Robert McDonald, Jordi Barretina, Ellen T. Gelfand, Craig M. Bielski, Haoxin Li, Kevin Hu, Alexander Y. Andreev-Drakhlin, Jaegil Kim, Julian M. Hess, Brian J. Haas, François Aguet, Barbara A. Weir, Michael V. Rothberg, Brenton R. Paolella, Michael S. Lawrence, Rehan Akbani, Yiling Lu, Hong L. Tiv, Prafulla C. Gokhale, Antoine de Weck, Ali Amin Mansour, Coyin Oh, Juliann Shih, Kevin Hadi, Yanay Rosen, Jonathan Bistline, Kavitha Venkatesan, Anupama Reddy, Dmitriy Sonkin, Manway Liu, Joseph Lehar, Joshua M. Korn, Dale A. Porter, Michael D. Jones, Javad Golji, Giordano Caponigro, Jordan E. Taylor, Caitlin M. Dunning, Amanda L. Creech, Allison C. Warren, James M. McFarland, Mahdi Zamanighomi, Audrey Kauffmann, Nicolas Stransky, Marcin Imielinski, Yosef E. Maruvka, Andrew D. Cherniack, Aviad Tsherniak, Francisca Vazquez, Jacob D. Jaffe, Andrew A. Lane, David M. Weinstock, Cory M. Johannessen, Michael P. Morrissey, Frank Stegmeier, Robert Schlegel, William C. Hahn, Gad Getz, Gordon B. Mills, Jesse S. Boehm, Todd R. Golub, Levi A. Garraway, and William R. Sellers. Next-generation characterization of the Cancer Cell Line Encyclopedia. *Nature*, 569(7757):503–508, 2019.

[16] Rose Oughtred, Chris Stark, Bobby-Joe Breitkreutz, Jennifer Rust, Lorrie Boucher, Christie Chang, Nadine Kolas, Lara O'Donnell, Genie Leung, Rochelle McAdam, Frederick Zhang, Sonam Dolma, Andrew Willems, Jasmin Coulombe-Huntington, Andrew Chatr-aryamontri, Kara Dolinski, and Mike Tyers. The BioGRID interaction database: 2019 update. *Nucleic Acids Research*, 47(D1):D529–D541, 2019.

[17] Andrea Franceschini, Damian Szklarczyk, Sune Frankild, Michael Kuhn, Milan Simonovic, Alexander Roth, Jianyi Lin, Pablo Minguez, Peer

Bork, Christian von Mering, and Lars J. Jensen. STRING v9.1: protein-protein interaction networks, with increased coverage and integration. *Nucleic Acids Research*, 41(D1):D808–D815, nov 2012.

[18] Gabriele Sales, Enrica Calura, Duccio Cavalieri, and Chiara Romualdi. graphite - a Bioconductor package to convert pathway topology to gene network. *BMC Bioinformatics*, 13(1):20, 2012.

[19] Matthew E. Ritchie, Belinda Phipson, Di Wu, Yifang Hu, Charity W. Law, Wei Shi, and Gordon K. Smyth. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, 43(7):e47, apr 2015.

[20] Stephen A. Vavasis. On the Complexity of Nonnegative Matrix Factorization. *SIAM Journal on Optimization*, 20(3):1364–1377, jan 2010.

[21] J.-P. Brunet, P. Tamayo, T. R. Golub, and J. P. Mesirov. Metagenes and molecular pattern discovery using matrix factorization. *Proceedings of the National Academy of Sciences*, 101(12):4164–4169, mar 2004.

[22] Scott J. Steppan, Patrick C. Phillips, and David Houle. Comparative quantitative genetics: Evolution of the G matrix. *Trends in Ecology and Evolution*, 17(7):320–327, 2002.

[23] Guangchuang Yu, Li-Gen Wang, Yanyan Han, and Qing-Yu He. clusterProfiler: an R Package for Comparing Biological Themes Among Gene Clusters. *OMICS: A Journal of Integrative Biology*, 16(5):284–287, may 2012.

[24] Le Chang, Guangyan Zhou, Othman Soufan, and Jianguo Xia. miRNet 2.0: network-based visual analytics for miRNA functional analysis and systems biology. *Nucleic Acids Research*, 48(W1):W244–W251, jul 2020.

[25] J. Gao, B. A. Aksoy, U. Dogrusoz, G. Dresdner, B. Gross, S. O. Sumer, Y. Sun, A. Jacobsen, R. Sinha, E. Larsson, E. Cerami, C. Sander, and N. Schultz. Integrative Analysis of Complex Cancer Genomics and Clinical Profiles Using the cBioPortal. *Science Signaling*, 6(269):pl1–pl1, apr 2013.

[26] Ethan Cerami, Jianjiong Gao, Ugur Dogrusoz, Benjamin E. Gross, Selcuk Onur Sumer, Bülent Arman Aksoy, Anders Jacobsen, Caitlin J.

Byrne, Michael L. Heuer, Erik Larsson, Yevgeniy Antipin, Boris Reva, Arthur P. Goldberg, Chris Sander, and Nikolaus Schultz. The cBio Cancer Genomics Portal: An Open Platform for Exploring Multidimensional Cancer Genomics Data: Figure 1. *Cancer Discovery*, 2(5):401–404, may 2012.

[27] Rileen Sinha, Nikolaus Schultz, and Chris Sander. Comparing cancer cell lines and tumor samples by genomic profiles. *bioRxiv*, page 28159, 2015.

[28] Jianfang Liu, Tara Lichtenberg, Katherine A. Hoadley, Laila M. Poisson, Alexander J. Lazar, Andrew D. Cherniack, Albert J. Kovatich, Christopher C. Benz, Douglas A. Levine, Adrian V. Lee, Larsson Omberg, Denise M. Wolf, Craig D. Shriver, Vesteinn Thorsson, Samantha J. Caesar-Johnson, John A. Demchok, Ina Felau, Melpomeni Kasapi, Martin L. Ferguson, Carolyn M. Hutter, Heidi J. Sofia, Roy Tarnuzzer, Zhining Wang, Liming Yang, Jean C. Zenklusen, Jiashan (Julia) Zhang, Sudha Chudamani, Jia Liu, Laxmi Lolla, Rashi Naresh, Todd Pihl, Qiang Sun, Yunhu Wan, Ye Wu, Juok Cho, Timothy DeFreitas, Scott Frazer, Nils Gehlenborg, Gad Getz, David I. Heiman, Jaegil Kim, Michael S. Lawrence, Pei Lin, Sam Meier, Michael S. Noble, Gordon Saksena, Doug Voet, Hailei Zhang, Brady Bernard, Nyasha Chambwe, Varsha Dhankani, Theo Knijnenburg, Roger Kramer, Kalle Leinonen, Yuexin Liu, Michael Miller, Sheila Reynolds, Ilya Shmulevich, Vesteinn Thorsson, Wei Zhang, Rehan Akbani, Bradley M. Broom, Apurva M. Hegde, Zhenlin Ju, Rupa S. Kanchi, Anil Korkut, Jun Li, Han Liang, Shiyun Ling, Wenbin Liu, Yiling Lu, Gordon B. Mills, Kwok Shing Ng, Arvind Rao, Michael Ryan, Jing Wang, John N. Weinstein, Jiexin Zhang, Adam Abeshouse, Joshua Armenia, Debyani Chakravarty, Walid K. Chatila, Ino de Bruijn, Jianjiong Gao, Benjamin E. Gross, Zachary J. Heins, Ritika Kundra, Konnor La, Marc Ladanyi, Augustin Luna, Moriah G. Nissan, Angelica Ochoa, Sarah M. Phillips, Ed Reznik, Francisco Sanchez-Vega, Chris Sander, Nikolaus Schultz, Robert Sheridan, S. Onur Sumer, Yichao Sun, Barry S. Taylor, Jioajiao Wang, Hongxin Zhang, Pavana Anur, Myron Peto, Paul Spellman, Christopher Benz, Joshua M. Stuart, Christopher K. Wong, Christina Yau, D. Neil Hayes, Joel S. Parker, Matthew D. Wilkerson, Adrian Ally, Miruna Balasundaram, Reanne Bowlby, Denise Brooks,

Rebecca Carlsen, Eric Chuah, Noreen Dhalla, Robert Holt, Steven J.M. Jones, Katayoon Kasaian, Darlene Lee, Yussanne Ma, Marco A. Marra, Michael Mayo, Richard A. Moore, Andrew J. Mungall, Karen Mungall, A. Gordon Robertson, Sara Sadeghi, Jacqueline E. Schein, Payal Sipahimalani, Angela Tam, Nina Thiessen, Kane Tse, Tina Wong, Ashton C. Berger, Rameen Beroukhim, Andrew D. Cherniack, Carrie Cibulskis, Stacey B. Gabriel, Galen F. Gao, Gavin Ha, Matthew Meyerson, Steven E. Schumacher, Juliann Shih, Melanie H. Kucherlapati, Raju S. Kucherlapati, Stephen Baylin, Leslie Cope, Ludmila Danilova, Moiz S. Bootwalla, Phillip H. Lai, Dennis T. Maglinte, David J. Van Den Berg, Daniel J. Weisenberger, J. Todd Auman, Saianand Balu, Tom Bodenheimer, Cheng Fan, Katherine A. Hoadley, Alan P. Hoyle, Stuart R. Jefferys, Corbin D. Jones, Shaowu Meng, Piotr A. Mieczkowski, Lisle E. Mose, Amy H. Perou, Charles M. Perou, Jeffrey Roach, Yan Shi, Janae V. Simons, Tara Skelly, Matthew G. Soloway, Donghui Tan, Umadevi Veluvolu, Huihui Fan, Toshinori Hinoue, Peter W. Laird, Hui Shen, Wanding Zhou, Michelle Bellair, Kyle Chang, Kyle Covington, Chad J. Creighton, Huyen Dinh, Harsha Vardhan Doddapaneni, Lawrence A. Donehower, Jennifer Drummond, Richard A. Gibbs, Robert Glenn, Walker Hale, Yi Han, Jianhong Hu, Viktoriya Korchina, Sandra Lee, Lora Lewis, Wei Li, Xiuping Liu, Margaret Morgan, Donna Morton, Donna Muzny, Jireh Santibanez, Margi Sheth, Eve Shinbro, Linghua Wang, Min Wang, David A. Wheeler, Liu Xi, Fengmei Zhao, Julian Hess, Elizabeth L. Appelbaum, Matthew Bailey, Matthew G. Cordes, Li Ding, Catrina C. Fronick, Lucinda A. Fulton, Robert S. Fulton, Cyriac Kandoth, Elaine R. Mardis, Michael D. McLellan, Christopher A. Miller, Heather K. Schmidt, Richard K. Wilson, Daniel Crain, Erin Curley, Johanna Gardner, Kevin Lau, David Mallery, Scott Morris, Joseph Paulauskis, Robert Penny, Candace Shelton, Troy Shelton, Mark Sherman, Eric Thompson, Peggy Yena, Jay Bowen, Julie M. Gastier-Foster, Mark Gerken, Kristen M. Leraas, Tara M. Lichtenberg, Nilsa C. Ramirez, Lisa Wise, Erik Zmuda, Niall Corcoran, Tony Costello, Christopher Hovens, Andre L. Carvalho, Ana C. de Carvalho, José H. Fregnani, Adhemar Longatto-Filho, Rui M. Reis, Cristovam Scapulatempo-Neto, Henrique C.S. Silveira, Daniel O. Vidal, Andrew Burnette, Jennifer Eschbacher, Beth Hermes, Ardene Noss, Rosy Singh, Matthew L. Anderson, Patricia D. Castro, Michael Ittmann, David Huntsman, Bernard Kohl, Xuan Le, Richard Thorp, Chris Andry, Elizabeth R. Duffy, Vladimir Lyadov, Oxana Paklina, Galiya Setdikova, Alexey Shabunin, Mikhail Tavobilov, Christopher McPherson, Ronald Warnick, Ross Berkowitz, Daniel Cramer, Colleen Feltmate, Neil Horowitz, Adam Kibel, Michael Muto, Chandrajit P. Raut, Andrei Malykh, Jill S. Barnholtz-Sloan, Wendi Barrett, Karen Devine, Jordonna Fulop, Quinn T. Ostrom, Kristen Shimmel, Yingli Wolinsky, Andrew E. Sloan, Agostino De Rose, Felice Giuliante, Marc Goodman, Beth Y. Karlan, Curt H. Hagedorn, John Eckman, Jodi Harr, Jerome Myers, Kelinda Tucker, Leigh Anne Zach, Brenda Deyarmin, Hai Hu, Leonid Kvecher, Caroline Larson, Richard J. Mural, Stella Somiari, Ales Vicha, Tomas Zelinka, Joseph Bennett, Mary Iacocca, Brenda Rabeno, Patricia Swanson, Mathieu Latour, Louis Lacombe, Bernard Têtu, Alain Bergeron, Mary McGraw, Susan M. Staugaitis, John Chabot, Hanina Hibshoosh, Antonia Sepulveda, Tao Su, Timothy Wang, Olga Potapova, Olga Voronina, Laurence Desjardins, Odette Mariani, Sergio Roman-Roman, Xavier Sastre, Marc Henri Stern, Feixiong Cheng, Sabina Signoretti, Andrew Berchuck, Darell Bigner, Eric Lipp, Jeffrey Marks, Shannon McCall, Roger McLendon, Angeles Secord, Alexis Sharp, Madhusmita Behera, Daniel J. Brat, Amy Chen, Keith Delman, Seth Force, Fadlo Khuri, Kelly Magliocca, Shishir Maithel, Jeffrey J. Olson, Taofeek Owonikoko, Alan Pickens, Suresh Ramalingam, Dong M. Shin, Gabriel Sica, Erwin G. Van Meir, Hongzheng Zhang, Wil Eijckenboom, Ad Gillis, Esther Korpershoek, Leendert Looijenga, Wolter Oosterhuis, Hans Stoop, Kim E. van Kessel, Ellen C. Zwarthoff, Chiara Calatozzolo, Lucia Cuppini, Stefania Cuzzubbo, Francesco DiMeco, Gaetano Finocchiaro, Luca Mattei, Alessandro Perin, Bianca Pollo, Chu Chen, John Houck, Pawadee Lohavanichbutr, Arndt Hartmann, Christine Stoehr, Robert Stoehr, Helge Taubert, Sven Wach, Bernd Wullich, Witold Kycler, Dawid Murawa, Maciej Wiznerowicz, Ki Chung, W. Jeffrey Edenfield, Julie Martin, Eric Baudin, Glenn Bubley, Raphael Bueno, Assunta De Rienzo, William G. Richards, Steven Kalkanis, Tom Mikkelsen, Houtan Noushmehr, Lisa Scarpace, Nicolas Girard, Marta Aymerich, Elias Campo, Eva Giné, Armando López Guillermo, Nguyen Van Bang, Phan Thi Hanh, Bui Duc Phu, Yufang Tang, Howard Colman, Kimberley Evason, Peter R. Dottino, John A. Martignetti, Hani Gabra, Hartmut Juhl, Teniola Akeredolu, Serghei Stepa, Dave Hoon, Keunsoo Ahn, Koo Jeong Kang, Felix Beuschlein, Anne Breggia, Michael Birrer, Debra Bell, Mitesh Borad, Alan H. Bryce, Erik Castle, Vishal Chandan, John Cheville, John A. Copland, Michael Farnell, Thomas Flotte, Nasra Giama, Thai Ho, Michael Kendrick, Jean Pierre Kocher, Karla Kopp, Catherine Moser, David Nagorney, Daniel O'Brien, Brian Patrick O'Neill, Tushar Patel, Gloria Petersen, Florencia Que, Michael Rivera, Lewis Roberts, Robert Smallridge, Thomas Smyrk, Melissa Stanton, R. Houston Thompson, Michael Torbenson, Ju Dong Yang, Lizhi Zhang, Fadi Brimo, Jaffer A. Ajani, Ana Maria Angulo Gonzalez, Carmen Behrens, Jolanta Bondaruk, Russell Broaddus, Bogdan Czerniak, Bita Esmaeli, Junya Fujimoto, Jeffrey Gershenwald, Charles Guo, Christopher Logothetis, Funda Meric-Bernstam, Cesar Moran, Lois Ramondetta, David Rice, Anil Sood, Pheroze Tamboli, Timothy Thompson, Patricia Troncoso, Anne Tsao, Ignacio Wistuba, Candace Carter, Lauren Haydu, Peter Hersey, Valerie Jakrot, Hojabr Kakavand, Richard Kefford, Kenneth Lee, Georgina Long, Graham Mann, Michael Quinn, Robyn Saw, Richard Scolyer, Kerwin Shannon, Andrew Spillane, Jonathan Stretch, Maria Synott, John Thompson, James Wilmott, Hikmat Al-Ahmadie, Timothy A. Chan, Ronald Ghossein, Anuradha Gopalan, Douglas A. Levine, Victor Reuter, Samuel Singer, Bhuvanesh Singh, Nguyen Viet Tien, Thomas Broudy, Cyrus Mirsaidi, Praveen Nair, Paul Drwiega, Judy Miller, Jennifer Smith, Howard Zaren, Joong Won Park, Nguyen Phi Hung, Electron Kebebew, W. Marston Linehan, Adam R. Metwalli, Karel Pacak, Peter A. Pinto, Mark Schiffman, Laura S. Schmidt, Cathy D. Vocke, Nicolas Wentzensen, Robert Worrell, Hannah Yang, Marc Moncrieff, Chandra Goparaju, Jonathan Melamed, Harvey Pass, Natalia Botnariuc, Irina Caraman, Mircea Cernat, Inga Chemencedji, Adrian Clipca, Serghei Doruc, Ghenadie Gorincioi, Sergiu Mura, Maria Pirtac, Irina Stancul, Diana Tcaciuc, Monique Albert, Iakovina Alexopoulou, Angel Arnaout, John Bartlett, Jay Engel, Sebastien Gilbert, Jeremy Parfitt, Harman Sekhon, George Thomas, Doris M. Rassl, Robert C. Rintoul, Carlo Bifulco, Raina Tamakawa, Walter Urba, Nicholas Hayward, Henri Timmers, Anna Antenucci, Francesco Facciolo, Gianluca Grazi, Mirella Marino, Roberta Merola, Ronald de Krijger, Anne Paule Gimenez-Roqueplo, Alain Piché, Simone Chevalier, Ginette McKercher, Kivanc Birsoy, Gene Barnett, Cathy Brewer, Carol Farver, Theresa Naska, Nathan A. Pennell, Daniel Raymond, Cathy Schilero, Kathy Smolenski, Felicia Williams, Carl Morrison, Jeffrey A. Borgia, Michael J. Liptay, Mark Pool, Christopher W. Seder, Kerstin Junker, Larsson Omberg, Mikhail Dinkin, George Manikhas, Domenico Alvaro, Maria Consiglia Bragazzi, Vincenzo Cardinale, Guido Carpino, Eugenio Gaudio, David Chesla, Sandra Cottingham, Michael Dubina, Fedor Moiseenko, Renumathy Dhanasekaran, Karl Friedrich Becker, Klaus Peter Janssen, Julia Slotta-Huspenina, Mohamed H. Abdel-Rahman, Dina Aziz, Sue Bell, Colleen M. Cebulla, Amy Davis, Rebecca Duell, J. Bradley Elder, Joe Hilty, Bahavna Kumar, James Lang, Norman L. Lehman, Randy Mandt, Phuong Nguyen, Robert Pilarski, Karan Rai, Lynn Schoenfield, Kelly Senecal, Paul Wakely, Paul Hansen, Ronald Lechan, James Powers, Arthur Tischler, William E. Grizzle, Katherine C. Sexton, Alison Kastl, Joel Henderson, Sima Porten, Jens Waldmann, Martin Fassnacht, Sylvia L. Asa, Dirk Schadendorf, Marta Couce, Markus Graefen, Hartwig Huland, Guido Sauter, Thorsten Schlomm, Ronald Simon, Pierre Tennstedt, Oluwole Olabode, Mark Nelson, Oliver Bathe, Peter R. Carroll, June M. Chan, Philip Disaia, Pat Glenn, Robin K. Kelley, Charles N. Landen, Joanna Phillips, Michael Prados, Jeffry Simko, Karen Smith-McCune, Scott VandenBerg, Kevin Roggin, Ashley Fehrenbach, Ady Kendler, Suzanne Sifri, Ruth Steele, Antonio Jimeno, Francis Carey, Ian Forgie, Massimo Mannelli, Michael Carney, Brenda Hernandez, Benito Campos, Christel Herold-Mende, Christin Jungk, Andreas Unterberg, Andreas von Deimling, Aaron Bossler, Joseph Galbraith, Laura Jacobus, Michael Knudson, Tina Knutson, Deqin Ma, Mohammed Milhem, Rita Sigmund, Andrew K. Godwin, Rashna Madan, Howard G. Rosenthal, Clement Adebamowo, Sally N. Adebamowo, Alex Boussioutas, David Beer, Thomas Giordano, Anne Marie Mes-Masson, Fred Saad, Therese Bocklage, Lisa Landrum, Robert Mannel, Kathleen Moore, Katherine Moxley, Russel Postier, Joan Walker, Rosemary Zuna, Michael Feldman, Federico Valdivieso, Rajiv Dhir, James Luketich, Edna M. Mora Pinero, Mario Quintero-Aguilo, Carlos Gilberto Carlotti, Jose Sebastião Dos Santos, Rafael Kemp, Ajith Sankarankuty, Daniela Tirapelli, James Catto, Kathy Agnew, Elizabeth Swisher, Jenette Creaney, Bruce Robinson, Carl Simon Shelley, Eryn M. Godwin, Sara Kendall, Cassaundra Shipman, Carol Bradford, Thomas Carey, Andrea Haddad, Jeffey Moyer, Lisa Peterson, Mark Prince, Laura Rozek, Gregory Wolf, Rayleen Bowman, Kwun M. Fong, Ian Yang, Robert Korst, W. Kimryn Rathmell, J. Leigh Fantacone-Campbell, Jeffrey A. Hooke, Albert J. Kovatich, Craig D. Shriver, John DiPersio, Bettina Drake, Ramaswamy Govindan, Sharon Heath, Timothy Ley, Brian Van Tine,

Peter Westervelt, Mark A. Rubin, Jung Il Lee, Natália D. Aredes, Armaz Mariamidze, and Hai Hu. An Integrated TCGA Pan-Cancer Clinical Data Resource to Drive High-Quality Survival Outcome Analytics. *Cell*, 173(2):400–416.e11, 2018.

[29] Michele Ceccarelli, Floris P. Barthel, Tathiane M. Malta, Thais S. Sabedot, Sofie R. Salama, Bradley A. Murray, Olena Morozova, Yulia Newton, Amie Radenbaugh, Stefano M. Pagnotta, Samreen Anjum, Jiguang Wang, Ganiraju Manyam, Pietro Zoppoli, Shiyun Ling, Arjun A. Rao, Mia Grifford, Andrew D. Cherniack, Hailei Zhang, Laila Poisson, Carlos Gilberto Carlotti, Daniela Pretti Da Cunha Tirapelli, Arvind Rao, Tom Mikkelsen, Ching C. Lau, W. K.Alfred Yung, Raul Rabadan, Jason Huse, Daniel J. Brat, Norman L. Lehman, Jill S. Barnholtz-Sloan, Siyuan Zheng, Kenneth Hess, Ganesh Rao, Matthew Meyerson, Rameen Beroukhim, Lee Cooper, Rehan Akbani, Margaret Wrensch, David Haussler, Kenneth D. Aldape, Peter W. Laird, David H. Gutmann, Samreen Anjum, Harindra Arachchi, J. Todd Auman, Miruna Balasundaram, Saianand Balu, Gene Barnett, Stephen Baylin, Sue Bell, Christopher Benz, Natalie Bir, Keith L. Black, Tom Bodenheimer, Lori Boice, Moiz S. Bootwalla, Jay Bowen, Christopher A. Bristow, Yaron S.N. Butterfield, Qing Rong Chen, Lynda Chin, Juok Cho, Eric Chuah, Sudha Chudamani, Simon G. Coetzee, Mark L. Cohen, Howard Colman, Marta Couce, Fulvio D'Angelo, Tanja Davidsen, Amy Davis, John A. Demchok, Karen Devine, Li Ding, Rebecca Duell, J. Bradley Elder, Jennifer M. Eschbacher, Ashley Fehrenbach, Martin Ferguson, Scott Frazer, Gregory Fuller, Jordonna Fulop, Stacey B. Gabriel, Luciano Garofano, Julie M. Gastier-Foster, Nils Gehlenborg, Mark Gerken, Gad Getz, Caterina Giannini, William J. Gibson, Angela Hadjipanayis, D. Neil Hayes, David I. Heiman, Beth Hermes, Joe Hilty, Katherine A. Hoadley, Alan P. Hoyle, Mei Huang, Stuart R. Jefferys, Corbin D. Jones, Steven J.M. Jones, Zhenlin Ju, Alison Kastl, Ady Kendler, Jaegil Kim, Raju Kucherlapati, Phillip H. Lai, Michael S. Lawrence, Semin Lee, Kristen M. Leraas, Tara M. Lichtenberg, Pei Lin, Yuexin Liu, Jia Liu, Julia Y. Ljubimova, Yiling Lu, Yussanne Ma, Dennis T. Maglinte, Harshad S. Mahadeshwar, Marco A. Marra, Mary McGraw, Christopher McPherson, Shaowu Meng, Piotr A. Mieczkowski, C. Ryan Miller, Gordon B. Mills, Richard A. Moore, Lisle E. Mose, Andrew J. Mungall, Rashi Naresh, Theresa Naska, Luciano Neder, Michael S. Noble, Ardene

Noss, Brian Patrick O'Neill, Quinn T. Ostrom, Cheryl Palmer, Angeliki Pantazi, Michael Parfenov, Peter J. Park, Joel S. Parker, Charles M. Perou, Christopher R. Pierson, Todd Pihl, Alexei Protopopov, Amie Radenbaugh, Nilsa C. Ramirez, W. Kimryn Rathmell, Xiaojia Ren, Jeffrey Roach, A. Gordon Robertson, Gordon Saksena, Jacqueline E. Schein, Steven E. Schumacher, Jonathan Seidman, Kelly Senecal, Sahil Seth, Hui Shen, Yan Shi, Juliann Shih, Kristen Shimmel, Hugues Sicotte, Suzanne Sifri, Tiago Silva, Janae V. Simons, Rosy Singh, Tara Skelly, Andrew E. Sloan, Heidi J. Sofia, Matthew G. Soloway, Xingzhi Song, Carrie Sougnez, Camila Souza, Susan M. Staugaitis, Huandong Sun, Charlie Sun, Donghui Tan, Jiabin Tang, Yufang Tang, Leigh Thorne, Felipe Amstalden Trevisan, Timothy Triche, David J. Van Den Berg, Umadevi Veluvolu, Doug Voet, Yunhu Wan, Zhining Wang, Ronald Warnick, John N. Weinstein, Daniel J. Weisenberger, Matthew D. Wilkerson, Felicia Williams, Lisa Wise, Yingli Wolinsky, Junyuan Wu, Andrew W. Xu, Lixing Yang, Liming Yang, Travis I. Zack, Jean C. Zenklusen, Jianhua Zhang, Wei Zhang, Jiashan Zhang, Erik Zmuda, Houtan Noushmehr, Antonio Iavarone, and Roel G.W. Verhaak. Molecular Profiling Reveals Biologically Discrete Subsets and Pathways of Progression in Diffuse Glioma. *Cell*, 164(3):550–563, 2016.

[30] V. G. Tusher, R. Tibshirani, and G. Chu. Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences*, 98(9):5116–5121, apr 2001.

[31] Shihua Zhang, Chun-chi Liu, Wenyuan Li, Hui Shen, Peter W Laird, and Xianghong Jasmine Zhou. Discovery of multi-dimensional modules by integrative analysis of cancer genomic data. 40(19):9379–9391, 2012.

[32] Arshi Arora, Adam B. Olshen, Venkatraman E. Seshan, and Ronglai Shen. Pan-cancer identification of clinically relevant genomic subtypes using outcome-weighted integrative clustering. *Genome Medicine*, 12(1):110, dec 2020.

[33] Rui Liu, Yajuan Chen, Tao Shou, Jing Hu, Jingbo Chen, and Chen Qing. TRIM67 promotes NF-κB pathway and cell apoptosis in GA-13315-treated lung cancer cells. *Molecular Medicine Reports*, 20(3):2936–2944, jul 2019.

[34] Saran Kumar, Nithya Rao, and Ruowen Ge. Emerging Roles of ADAMTSs in Angiogenesis and Cancer. *Cancers*, 4(4):1252–1299, nov 2012.

[35] Ting Tang, Ling-Xing Wang, Mei-Li Yang, and Rong-Mou Zhang. lncRNA TPTEP1 inhibits stemness and radioresistance of glioma through miR-106a-5p-mediated P38 MAPK signaling. *Molecular Medicine Reports*, 22(6):4857–4867, sep 2020.

[36] Soon Young Park, Yuji Piao, Kang Jin Jeong, Jianwen Dong, and John F. de Groot. Periostin (POSTN) Regulates Tumor Resistance to Antiangiogenic Therapy in Glioma Models. *Molecular Cancer Therapeutics*, 15(9):2187–2197, sep 2016.

[37] Gaelle Tachon, Konstantin Masliantsev, Pierre Rivet, Christos Petropoulos, Julie Godet, Serge Milin, Michel Wager, Pierre-Olivier Guichet, and Lucie Karayan-Tapon. Prognostic significance of MEOX2 in gliomas. *Modern Pathology*, 32(6):774–786, jun 2019.

[38] Yu-Long Lan, Xun Wang, Jia-Cheng Lou, Xiao-Chi Ma, and Bo Zhang. The potential roles of aquaporin 4 in malignant gliomas. *Oncotarget*, 8(19):32345–32355, may 2017.

[39] Danfeng Zhang, Dawei Dai, Mengxia Zhou, Zhenxing Li, Chunhui Wang, Yicheng Lu, Yiming Li, and Junyu Wang. Inhibition of Cyclin D1 Expression in Human Glioblastoma Cells is Associated with Increased Temozolomide Chemosensitivity. *Cellular Physiology and Biochemistry*, 51(6):2496–2508, 2018.

[40] Silvia Domcke, Rileen Sinha, Douglas A. Levine, Chris Sander, and Nikolaus Schultz. Evaluating cell lines as tumour models by comparison of genomic profiles. *Nature Communications*, 4(1):2126, oct 2013.

[41] Dvir Aran, Marina Sirota, and Atul J. Butte. Systematic pan-cancer analysis of tumour purity. *Nature Communications*, 6(1):8971, dec 2015.

[42] Steven M. Corsello, Rohith T. Nagari, Ryan D. Spangler, Jordan Rossen, Mustafa Kocak, Jordan G. Bryan, Ranad Humeidi, David Peck, Xiaoyun Wu, Andrew A. Tang, Vickie M. Wang, Samantha A. Bender, Evan Lemire, Rajiv Narayan, Philip Montgomery, Uri Ben-David, Colin W.

Garvie, Yejia Chen, Matthew G. Rees, Nicholas J. Lyons, James M. McFarland, Bang T. Wong, Li Wang, Nancy Dumont, Patrick J. O'Hearn, Eric Stefan, John G. Doench, Caitlin N. Harrington, Heidi Greulich, Matthew Meyerson, Francisca Vazquez, Aravind Subramanian, Jennifer A. Roth, Joshua A. Bittker, Jesse S. Boehm, Christopher C. Mader, Aviad Tsherniak, and Todd R. Golub. Discovering the anticancer potential of non-oncology drugs by systematic viability profiling. *Nature Cancer*, 1(2):235–248, feb 2020.

[43] Nikita Pozdeyev, Minjae Yoo, Ryan Mackie, Rebecca E. Schweppe, Aik Choon Tan, and Bryan R. Haugen. Integrating heterogeneous drug sensitivity data from cancer pharmacogenomic studies. *Oncotarget*, 7(32):51619–51625, aug 2016.

[44] Chia-Hung Chien, Wei-Ting Hsueh, Jian-Ying Chuang, and Kwang-Yu Chang. Dissecting the mechanism of temozolomide resistance and its association with the regulatory roles of intracellular reactive oxygen species in glioblastoma. *Journal of Biomedical Science*, 28(1):18, dec 2021.

[45] Muhammad Babar Khan, Rosamaria Ruggieri, Eesha Jamil, Nhan L. Tran, Camila Gonzalez, Nancy Mugridge, Steven Gao, Jennifer MacDiarmid, Himanshu Brahmbhatt, Jann N. Sarkaria, John Boockvar, and Marc Symons. Nanocell-mediated delivery of miR-34a counteracts temozolomide resistance in glioblastoma. *Molecular Medicine*, 27(1):28, dec 2021.

[46] Wenzheng Luo, Dongming Yan, Zhenyu Song, Xuqiang Zhu, Xianzhi Liu, Xueyuan Li, and Shanshan Zhao. miR-126-3p sensitizes glioblastoma cells to temozolomide by inactivating Wnt/β-catenin signaling via targeting SOX2. *Life Sciences*, 226:98–106, jun 2019.

[47] Jia Z. Shen and Charles Spruck. Targeting FBXO44/SUV39H1 elicits tumor cell-specific DNA replication stress and viral mimicry. *Cell Stress*, 5(3):37–39, mar 2021.

[48] P. Mellor, L. Deibert, B. Calvert, K. Bonham, S. A. Carlsen, and D. H. Anderson. CREB3L1 Is a Metastasis Suppressor That Represses Expression of Genes Regulating Metastasis, Invasion, and Angiogenesis. *Molecular and Cellular Biology*, 33(24):4985–4995, dec 2013.

[49] Li-qiang Liu, Li-fei Feng, Cheng-rui Nan, and Zong-mao Zhao. CREB3L1 and PTN expressions correlate with prognosis of brain glioma patients. *Bioscience Reports*, 38(3), jun 2018.

[50] Fiza Singh, Dayuan Gao, Mark G Lebwohl, and Huachen Wei. Shikonin modulates cell proliferation by inhibiting epidermal growth factor receptor signaling in human epidermoid carcinoma cells. *Cancer Letters*, 200(2):115–121, oct 2003.

[51] Xiaoqin Ma, Meixiang Yu, Chenxia Hao, and Wanhua Yang. Shikonin induces tumor apoptosis in glioma cells via endoplasmic reticulum stress, and Bax/Bak mediated mitochondrial outer membrane permeability. *Journal of Ethnopharmacology*, 263:113059, dec 2020.

[52] Subhashree Nayak, Meghali Aich, Anupam Kumar, Suman Sengupta, Prajakta Bajad, Parashar Dhapola, Deepanjan Paul, Kiran Narta, Suvendu Purkrait, Bharati Mehani, Ashish Suri, Debojyoti Chakraborty, Arijit Mukhopadhyay, and Chitra Sarkar. Novel internal regulators and candidate miRNAs within miR-379/miR-656 miRNA cluster can alter cellular phenotype of human glioblastoma. *Scientific Reports*, 8(1):7673, dec 2018.

[53] Lilei Peng, Jie Fu, and Yang Ming. The miR-200 family: multiple effects on gliomas. *Cancer Management and Research*, 10:1987–1992, jul 2018.

[54] Yuanyuan Qin, Weilong Chen, Bingjie Liu, Lei Zhou, Lu Deng, Wanxiang Niu, Dejun Bao, Chuandong Cheng, Dongxue Li, Suling Liu, and Chaoshi Niu. MiR-200c Inhibits the Tumor Progression of Glioma via Targeting Moesin. *Theranostics*, 7(6):1663–1673, 2017.

[55] Taeryool Koo, Bong Jun Cho, Dan Hyo Kim, Ji Min Park, Eun Jung Choi, Hans H. Kim, David J Lee, and In Ah Kim. MicroRNA-200c increases radiosensitivity of human cancer cells with activated EGFR-associated signaling. *Oncotarget*, 8(39):65457–65468, sep 2017.