Data Integration with SUMO Detects Latent Relationships Between Patients in Lower-Grade Gliomas

Karolina Sienkiewicz[1,7], Jinyu Chen[2,7], Ajay Chatrath[3], John T Lawson[1,4], Nathan C Sheffield[1,3,4,5,6], Louxin Zhang[2], and Aakrosh Ratan[1,5,6,*]

[1]Center for Public Health Genomics, University of Virginia, Charlottesville, VA 22908, USA

[2]Department of Mathematics and Computational Biology Program, National University of Singapore, Singapore 119076

[3]Department of Biochemistry and Molecular Genetics, University of Virginia, Charlottesville, VA, 22908, USA

[4]Department of Biomedical Engineering, University of Virginia, Charlottesville, VA, 22908, USA

[5]Department of Public Health Sciences, University of Virginia, Charlottesville, VA 22908, USA

[6]University of Virginia Cancer Center, Charlottesville, VA 22908, USA

[7]These authors contributed equally to this work.

[*]Correspondence should be addressed to Aakrosh Ratan, ratan@virginia.edu

Abstract

Joint analysis of multiple genomic data types can facilitate the discovery of complex mechanisms of biological processes and genetic diseases. We present a novel data integration framework based on non-negative matrix factorization that uses patient similarity networks. Our implementation supports continuous multi-omic datasets for molecular subtyping and handles missing data without using imputation, making it more efficient for genome-wide assays in large cohorts.

Applying our approach to gene expression, microRNA expression, and methylation data from patients with lower grade gliomas, we identify a subtype with a significantly poorer prognosis. Tumors assigned to this subtype are hypomethylated genome-wide with a gain of AP-1 occupancy in the demethylated distal enhancers. These tumors' genomic profiles are similar to Grade IV gliomas: they are enriched for somatic chr7 gain, chr10 loss, and other molecular events that have yet to be used in the diagnosis of lower-grade gliomas as per the current WHO guidelines.

Introduction

Biotechnologies for large-scale molecular studies of genetic diseases have advanced significantly. High throughput assays are now available to measure RNA expression, DNA methylation, and metabolite concentration in multiple tissues [1]. Given that each assay reveals a snapshot of certain cellular aspects of a disease, integrative analysis of multiple assays is often necessary for a complete understanding of its molecular etiology and important for discovering the molecular subtypes and biomarkers of the disease [2].

Molecular typing through clustering has traditionally focused on individual data types, primarily gene expression. In a few studies with multiple data types, the subtypes were generated from the different data types individually and subsequently integrated by domain experts [3–5]. Discordant results and disagreements in such analyses can be difficult to interpret and resolve. Another popular strategy is to concatenate feature matrices from multiple data types and then operate on the resulting matrix as a single data type. This approach allows the use of existing clustering techniques but requires cross-data type normalization and feature selection in individual data types before concatenation, possibly biasing the results. More sophisticated methods such as those implemented in iCluster [6], iClusterPlus [7], and Bayesian consensus clustering [8] model the probabilistic distribution of each data type and infer subtypes by maximization of the likelihood of the observed data. However, these methods require a feature selection step and make strong assumptions about the data.

The more recent methods for clustering on multi-omic data focus on similarity or distances between samples in-lieu of clustering on the feature matrices. For example, PINS [9] creates an average connectivity matrix based on the sample connectivity observed in the different data types. It then clusters using a method that depends on the level of agreement between the data types. Additionally, it perturbs the original data by adding Gaussian noise and chooses the number of clusters such that the output clusters are robust to this noise. Another popular method, Similarity Network Fusion (SNF) [10] creates a fused network of patients using a metric fusion technique and then partitions the data using spectral clustering. A more recent method, NEMO [11] calculates an average similarity matrix and then detects the clusters using spectral clustering. A comprehensive review of multi-omic and multi-view methods for the detection of subtypes is presented in Rappoport and Shamir [12].

The existing approaches have a few limitations. First, all approaches mentioned above but NEMO require that data be available for every sample and every data type, which is unlikely in most biological studies. For data with missing values, these methods need to impute missing values. But the imputation process is often computationally challenging for genome-wide analyses. Secondly, most of the methods rely on randomization to overcome computational challenges in some portion of their algorithm. Though the randomization approach can assist in finding a solution that avoids over-fitting, it also has implications for the robustness of the method. Different invocations of such a method on the same input data may produce different clustering outcomes. Lastly, statistical methods have the advantage of being able to include biological knowledge as priors. However, they often assume a parametric normal or gamma distribution of the data to make the parameter estimation tractable. Such an assumption is often not realistic and again leads to poor performance, as demonstrated in a recent comprehensive assessment of the methods for drug response prediction [13].

Here, we present a data integration framework based on non-negative matrix factorization (NMF) and showcase an implementation called SUMO (`https://github.com/ratan-lab/sumo`) that can integrate continuous data from multiple data types to infer molecular subtypes. SUMO handles missing data effectively and produces clusters that are robust to perturbations. Throughout the study, whenever appropriate, we compare SUMO v0.2.5 to LRAcluster v1.0 [14], MCCA v1.1 [15], NEMO v0.1 [11], PINSPlus v2.0 [9], and SNF v2.3 [10]. We use a recent benchmark [12] to show that SUMO is consistently among the best methods in identifying groups of patients with significantly differential prognosis and enrichment of clinical associations. Using simulation, we also compare SUMO to the other methods in the ability to cluster noisy datasets, to respond to perturbations, and to handle missing information.

As an application of our approach, we apply SUMO to multi-omic datasets from patients diagnosed with lower-grade glioma. Diffuse low-grade and intermediate-grade gliomas together make up the lower-grade gliomas (World Health Organization grades II and III), a diverse group of primary brain tumors with highly variable clinical behavior. Mutations in IDH, TP53, and ATRX and codeletion of chromosome arms 1p and 19q (1p/19q codeletion) have been identified as clinically relevant markers of lower-grade gliomas [16], and as of the 2016 edition of the WHO classification, gliomas are classified based not only on histopathologic appearance but also on these molecular markers [17]. Several studies have associated IDH mutations with a more favorable course of the disease, and have identified multiple subtypes with a poor clinical course [16, 18]. We identify a single cluster of patients with a significantly differential prognosis with SUMO. Patients of this cluster are enriched for genome-wide hypomethylation, somatic chr7 gain, chr10 loss, and other molecular events that have yet to be used in the diagnosis of lower-grade gliomas as per the current WHO guidelines.

Results

Method overview

The NMF technique aims to explain the observed data using a small number of basis components by factoring the data into the product of two non-negative matrices; one represents the basis components and the other contains mixture coefficients [19, 20]. NMF has been successfully used as a clustering method in image and pattern recognition [21–24], text-mining [25–28], and bioinformatics [29–34]. Symmetric NMF is a variant where the decomposition is done on a symmetrical matrix that contains pairwise similarity values between the data points, instead of being done directly on the data points [35]. Symmetric NMF improves clustering quality compared to the traditional formulation and forms the basis of our approach [36].

Similar to NEMO and SNF, we preprocess, transform, and standardize the data before calculating the similarity between the samples for each data type separately. If all data types are measured for all n samples, the similarity between samples of the i^{th} data type form a $n \times n$ symmetric matrix A_i. We then tri-factorize $A_i \approx HS_iH^T$, where H is a non-negative $n \times r$ matrix, S_i is a $r \times r$ non-negative matrix, and r ($\ll n$) is the desired number of clusters. H in this decomposition is shared among the various data types and is a representation of the n samples in a r-dimensional subspace accounting for the adjacencies observed in all data types. Each row in H represents a sample, and each column in H denotes a cluster. If H is sparse, as is typically the case in NMF, a sample is assigned to the cluster corresponding to the column in which the sample has the maximum value.

Multiplicative updates are used to solve the above factorization. Since the solution is sensitive to the initial conditions, we run the solver multiple times on several subsets of samples using different initial conditions and use consensus clustering to assign the final labels and infer the optimal number of clusters (see Method section for details).

SUMO exhibits improved performance with noisy and incomplete data

We performed several simulations to compare the accuracy of the various methods on noisy datasets with varying sample sizes and a varying fraction of missing data. We first generated a ‘ground truth’ feature matrix consisting of 200 samples and 400 features, with two distinctly separable clusters. We then simulated feature matrices of two different data types by adding different levels of Gaussian noise to this ground truth to conduct three sets of simulation experiments.

Fig S1 shows the experimental setup for the first simulation where we increase the noise in one data type while keeping a moderate amount of noise in the other data type. We generate 100 datasets for each amount of added noise and run all methods, comparing the resulting clusters to the ground truth using the adjusted Rand index (ARI). The results in Fig 1A show that all methods exhibit a median decrease in accuracy with an increase in noise. SUMO has the highest median ARI and the least variance (Fig S2) for all levels of noise.

Next, Fig S3 shows the experimental setup to study the impact of the sample size on the accuracy of the various tools. We again created two data types, one with a small amount of Gaussian noise, and another data type with a higher amount of Gaussian noise. Fig S4 shows the ARI of the resulting classification as an increasing fraction of samples are removed from each data type. SUMO, LRAcluster, and MCCA all score a median ARI of 1.00 for all sample sizes studied in this experiment.

Lastly, using the same setup as the second experiment, we compared SUMO to NEMO in their ability to classify samples accurately with missing data. Other methods do not handle missing data, and so were not included in this comparison. In this experiment, we removed a random fraction of samples from one data type, while preserving the data in the other data type. SUMO shows a higher median ARI compared to NEMO for most data points (Fig 1B).

Performance of SUMO on a recent benchmark

We compared SUMO to several other methods using a recently published benchmark [12]. The benchmark consists of methylation, gene expression, and miRNA expression data from 10 cancers sequenced as part of the TCGA project. As in the original benchmark, we evaluate each method for its ability to identify a subtype that shows significantly differential survival, and is enriched for clinical annotations. We chose or calculated parameters for the methods as suggested by the authors, without considering the survival and clinical parameters that are used for assessment.

Fig. 2 depicts the performance of the various methods on the data from the different cancer types. With respect to survival, SUMO had the total best prognostic value (sum of $-log_{10}$ p-values $= 18.88$), with MCCA being the second best with 17.48. However, the sum of p-values can be biased due to outliers, so we also counted the number of datasets for which a method's solution obtains significantly different survival (p-value < 0.05) (Table 1). As with the original benchmark, we also evaluated if at least one of the clusters were enriched for at least one of the clinical labels. p-values for the logrank test were calculated using permutation tests [37], enrichment for discrete parameters was calculated using the χ^2 test for independence, and enrichment for numeric parameters was calculated using the Kruskal-Wallis test. The p-values for clinical enrichment were corrected using Bonferroni correction.

Based on the results in Table 1, SUMO outperformed the other approaches, finding at least one cluster with significantly different survival in 7 out of the 10 cancers analyzed. For colorectal cancer and lung squamous cell carcinoma, none of the methods identified a subtype that showed significant differential survival. Subtypes for those cancers may be confounded

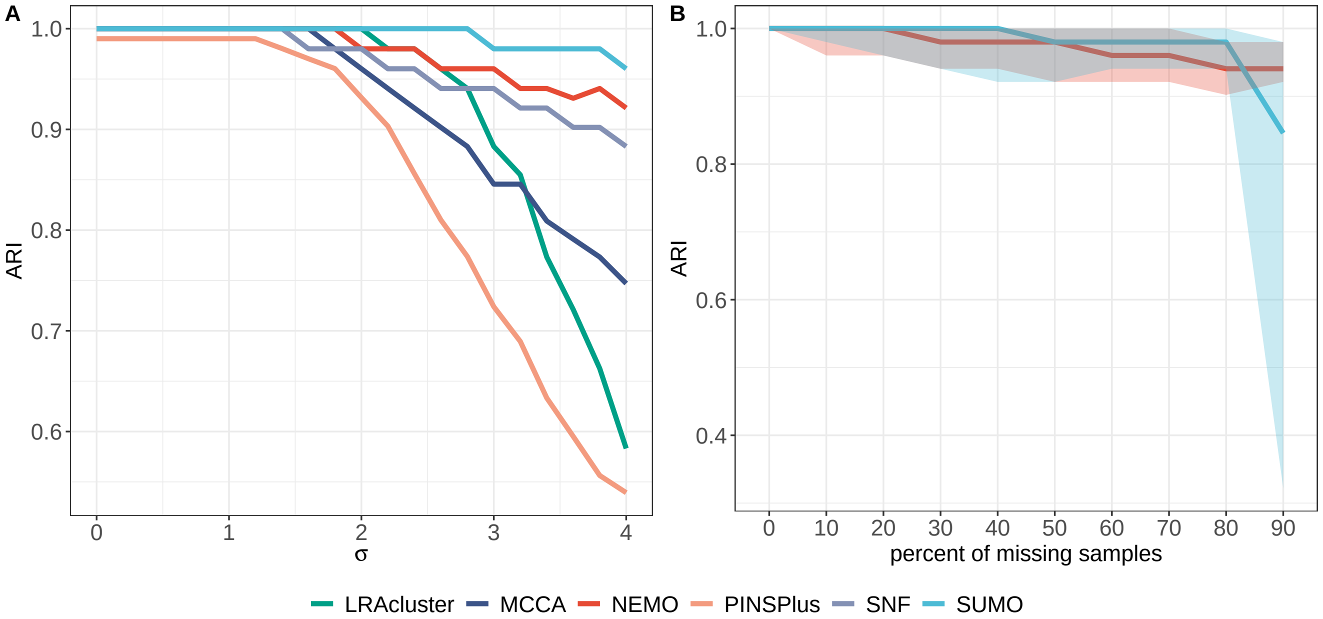

Figure 1: **Accuracy of the six methods on noisy data and missing values.** (A) Datasets were created by adding different amounts of noise to a 'ground truth' to simulate two distinct data types. The first data type is simulated by adding random noise from a Gaussian $\mathcal{N}(\mu = 0, \sigma = 1.5)$ distribution, while the noise in the second data type is from a Gaussian distribution ($\mathcal{N}(\mu = 0)$ where the standard deviation is varied $\sigma \in (0, 4)$). We report the median ARI of the classification at each data point for 100 repetitions. (B) Simulated datasets were created by removing the random fraction of samples from a random data type while keeping corresponding sample data in the other data type. We plot the ARI for 100 repetitions at each data point.

due to unknown covariates or may not exist at all, as suggested in Ma et al. [38], who found no evidence to support the existence of discrete transcriptional subtypes in colorectal cancer. SUMO is the only method to find a subgroup of patients in ovarian cancer with a significant differential survival (Fig S5A). This group of patients with poor prognosis is enriched for patients with mesenchymal tumors (Fig S5B) that are known to lead to worse outcomes [39].

All methods identified at least one cluster in Glioblastoma (GBM) with significantly differential prognosis. We used this GBM dataset to investigate the reproducibility and robustness of the methods, i.e. whether the p-values for the logrank test or the number of enriched clinical parameters would change if we changed the seed to the random number generator used by the methods and the assessment calculations. We ran each method 10 times using random seeds and found that the methods were stable to different extents on this data (Fig.S6). NEMO gave the same result in each of the 10 runs, while SUMO showed small deviations in the p-values for survival, but the remaining methods showed variation in both the p-value of the logrank test and the chi-square test used to assess the enrichment of clinical parameters. Specifically, the

Method	**Number of cancers with differential survival**	**Number of cancers with clinical enrichment**
LRAcluster	5	9
MCCA	5	8
NEMO	6	8
PINSPlus	4	6
SNF	4	7
SUMO	**7**	**7**

Table 1: **Summary of results from the benchmark analysis.** We report the number of cancers for which at least one cluster had significantly different prognosis (first column) and that had at least one enriched clinical label (second column).

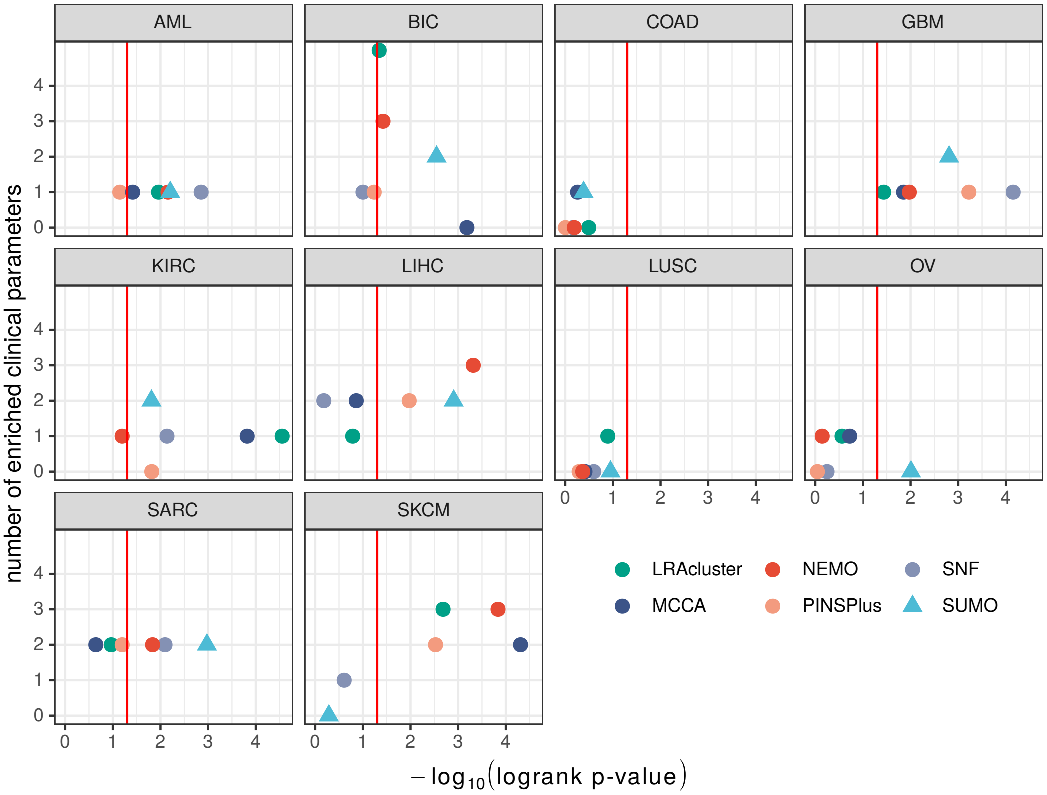

Figure 2: **Benchmark results for the TCGA datasets.** The vertical line indicates p-value$= 0.05$ for the logrank test, which are shown on the x-axis. The y-axis shows the number of clinical labels that were found to be enriched in at least one of the detected subtypes. SUMO results are shown using a triangle.

results for PINSPlus varied significantly in terms of survival and enrichment of clinical labels.

SUMO analysis of TCGA-LGG identifies a cluster of patients with poor-prognosis

Several integrative approaches have been applied to understand the molecular heterogeneity and subtypes in gliomas. The largest study of diffuse grade II-III-IV gliomas to date used TumorMap [40] to integrate gene expression and DNA methylation data from around 1000 patients and identified IDH status as the primary driver of two macro-clusters [41]. The authors concluded that the IDH mutant gliomas were further composed of three coherent subgroups: (1) the Codel group, consisting of LGGs with 1p/19q codeletion; (2) the G-CIMP-low group, including gliomas without 1p/19q codeletion with relatively low genome-wide DNA methylation; and (3) the G-CIMP-high group, including gliomas without 1p/19q codeletion with higher global levels of DNA methylation. They also concluded that the IDH wild type gliomas segregated into three subgroups: (1) Classic-like, exhibiting classical gene expression signature, (2) Mesenchymal-like, enriched for mesenchymal subtype tumors, and (3) PA-like, enriched for tumors with molecular similarity to grade I pilocytic astrocytomas.

We decided to apply SUMO to subtype the lower-grade gliomas as a case study with the intent to evaluate the robustness and relevance of known glioma subtypes. We ran SUMO on the processed Level 3 gene expression, DNA methylation, and miRNA expression data for the TCGA-LGG cohort downloaded from the UCSC Xena platform [42]. We evaluated the solutions with 2 to 19 clusters according to the proportion of ambiguously clustered pairs (PAC) [43] and the cophenetic correlation [44] (See Methods for details). The PAC values suggest that the patients can be partitioned into 2 or 5 clusters, with both solutions being stable (Fig. 3A). Here, we compare our solution with 2 clusters to the findings in Ceccarelli et al. [41], and then present the solution with 5 clusters in greater detail.

Fig 3B shows the Kaplan-Meier survival analysis for the 2 clusters identified by SUMO. Patients assigned to the group with worse prognosis have a median survival of 1262 days compared to a median survival of 2988 days for patients

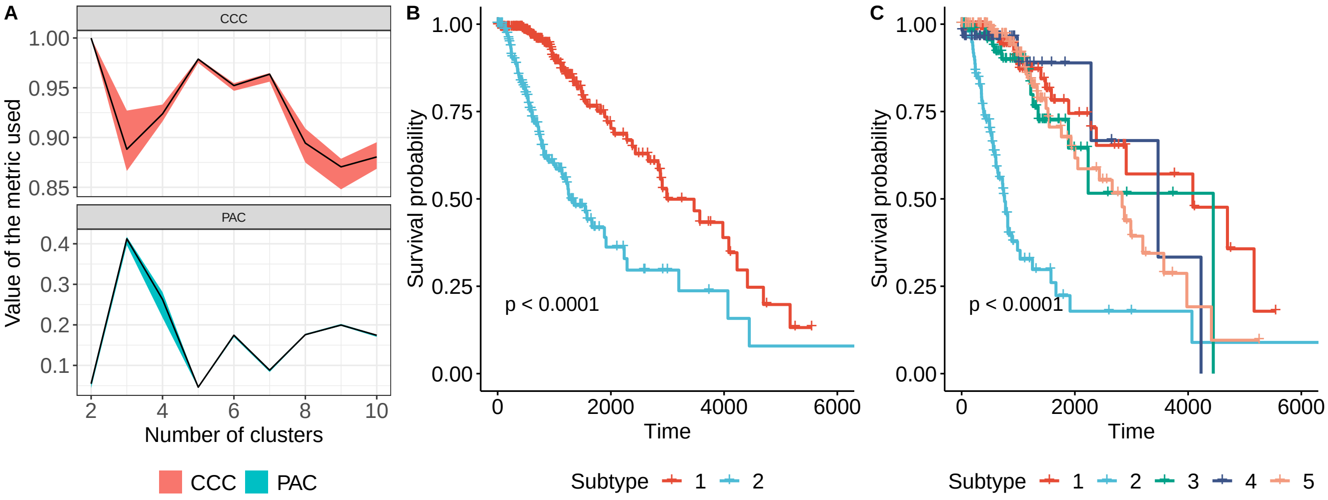

Figure 3: **SUMO detects a single cluster showing differential prognosis in TCGA-LGG.** (A) shows the two metrics used to decide the optimal number of clusters for LGG dataset. We use the proportion of ambiguously clustered pairs (PAC) (lower is better) and the cophenetic correlation (CCC) (higher is better) to select 2 and 5 as the optimal number of clusters. (B) and (C) shows the KM analysis of the subtypes detected by SUMO when 2 and 5 clusters are selected respectively.

assigned to the other subtype. The cluster of patients that show better prognosis include a majority of IDH mutant LGGs with 1p/19q codeletion and the majority of the IDH mutant LGG without 1p/19q codeletion with higher global levels of DNA methylation. SUMO assigns all IDH wild type patients and a subset of the IDH mutants to the subtype that exhibits a poor clinical course, and is significantly associated with higher aneuploidy (Wilcoxon rank sum test W=24243, p-value=0.0003), lower global methylation (Wilcoxon rank sum test W=58218, p-value=2.2×10^{-16}), a higher age of diagnosis (Wilcoxon rank sum test W=24674, p-value=2.92×10^{-5}) and a higher neoplasm grade based on histology (OR 2.45 (95% CI, 1.69 to 3.54)). Fig S7 summarizes the association of the 2 clusters with mutations, clinical phenotypes, and existing supervised classifications.

Fig 3C shows the Kaplan-Meier survival analysis for the 5 clusters as identified by SUMO. Patients assigned to Subtype 2 show a significant differential prognosis with a median survival of 758 days. Subtype 2 includes most samples (76 out of 80) that were labeled as Classic-like, Mesenchymal-like and C-GIMP low, and reported to have a poor clinical course in Ceccarelli et al. [41]. Subtype 2 also contains 18 of the 26 IDH wild type samples (labeled in Ceccarelli et al. [41] as PA-like) that were identified as having a favorable clinical course compared to other IDH wild type samples based on methylation analysis. To understand the reason for this difference, we compared the similarity between the PA-like samples assigned to Subtype 2 to (a) other samples in Subtype 2, and (b) PA-like samples assigned to other subtypes by SUMO. We determined that the PA-like samples in Subtype 2 are similar to the PA-like samples assigned to other clusters based on methylation data, consistent with Ceccarelli et al. [41]. However, the PA-like samples assigned to Subtype 2 show greater affinity to the other samples within Subtype 2 when the information from gene expression and miRNA expression were used (Fig S8).

In order to investigate if the subtypes detected by SUMO were enriched for other clinical and molecular events, we conducted enrichment analyses with the clinical phenotypes and GISTIC thresholded gene copy-number calls from UCSC Xena, along with molecular data from Ceccarelli et al. [41] and the somatic variants generated by the MC3 working group [45]. Subtype 2 is enriched for patients who are IDH wild-type and who were significantly older at the age of diagnosis (Tukey HSD test; p-value < 0.05 for all pairwise comparisons). Subtype 2 is also enriched for grade III tumors (OR 6.28 (95% CI, 3.40 to 11.59)) and significantly enriched for anaplastic Astrocytomas (p-value $< 10^{-5}$); it is also enriched for samples with a high percentage of aneuploidy (Tukey HSD pvalues < 0.05 for all pairwise comparisons), high ESTIMATE stromal score (Tukey HSD pvalues < 0.05 for all pairwise comparisons) and high ESTIMATE combined score (Tukey HSD pvalues < 0.05 for all pairwise comparisons). This is consistent with results that suggest that the ESTIMATE scores correlate with DNA copy number-based tumor purity and high ESTIMATE scores in LGG are associated with poor outcome [46, 47].

Fig 4 summarizes some of these associations in an oncoplot. Interestingly, Subtype 2 is enriched for point mutations and amplifications of the epidermal growth factor receptor (EGFR) oncogene on Chromosome (Chr) 7. Somatic aberrations in EGFR including amplification and activating point mutations occur in $\sim 57\%$ of Grade IV gliomas but are relatively uncommon in LGGs [48]. However, 55 of the 109 patients assigned to Subtype 2 show Chr 7 gain (and hence amplification of EGFR) and Chr 10 loss, which leads to deletion of the PTEN gene, a known tumor suppressor. These chromosomal aberrations together with global hypomethylation are features unique to this subtype. As per the WHO guidelines from 2016, Chr7 gain and/or Chr10 loss are not considered in the diagnosis of Grade II/III gliomas, though other studies have suggested that these events are clinically relevant, and their inclusion in the diagnostic criterion could lead to the reclassification of several LGGs into GBMs [49].

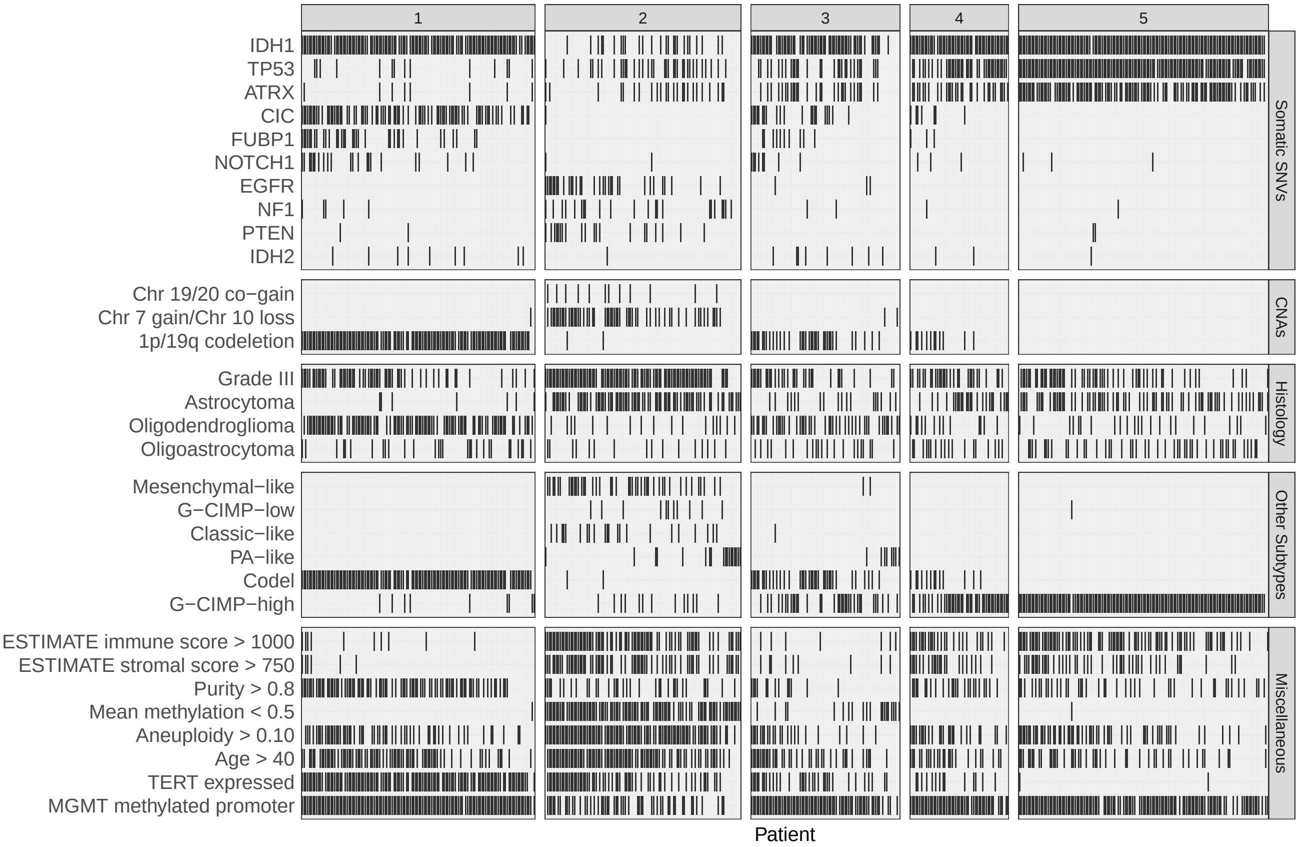

Figure 4: **Oncoplot showing enrichment of molecular and clinical features in the various subtypes.**

Since tumors are a complex milieu of numerous cell types, we hypothesized that the microenvironment plays an important role in the determination of these subtypes. To investigate this, we downloaded the xCell scores corresponding to enrichment of 64 different immune and stromal cell types in these TCGA samples [50]. Hierarchical clustering of the mean enrichment scores for the various cell types in Fig 5A shows that the cellular profile of Subtype 2 tumors is more similar to GBMs than to the other LGGs. More importantly, astrocytomas assigned to Subtype 2 have higher enrichment scores for astrocytes, similar to those calculated for GBM samples, and significantly higher than astrocytomas assigned to the other subtypes (Fig 5B). xCell scores are calculated using gene expression, but we observe similar results on analysis of methylation data using MIRA [51]. Subtype 2 samples show lower methylation and higher regulatory activity at astrocyte-specific elements (Fig 5C) compared to the other subtypes. These differences in cellular population also manifest in principal component analysis of gene expression and methylation data when we consider the LGG and GBM samples together (GBMLGG dataset from UCSC Xena). In PCA analyses of expression and methylation (Fig S9), the first principal component shows the similarities between Subtype 2 and the GBM samples. These findings along with the observed chromosomal aberrations suggest that LGGs assigned to Subtype 2 should be treated more aggressively and potentially reclassified as GBM.

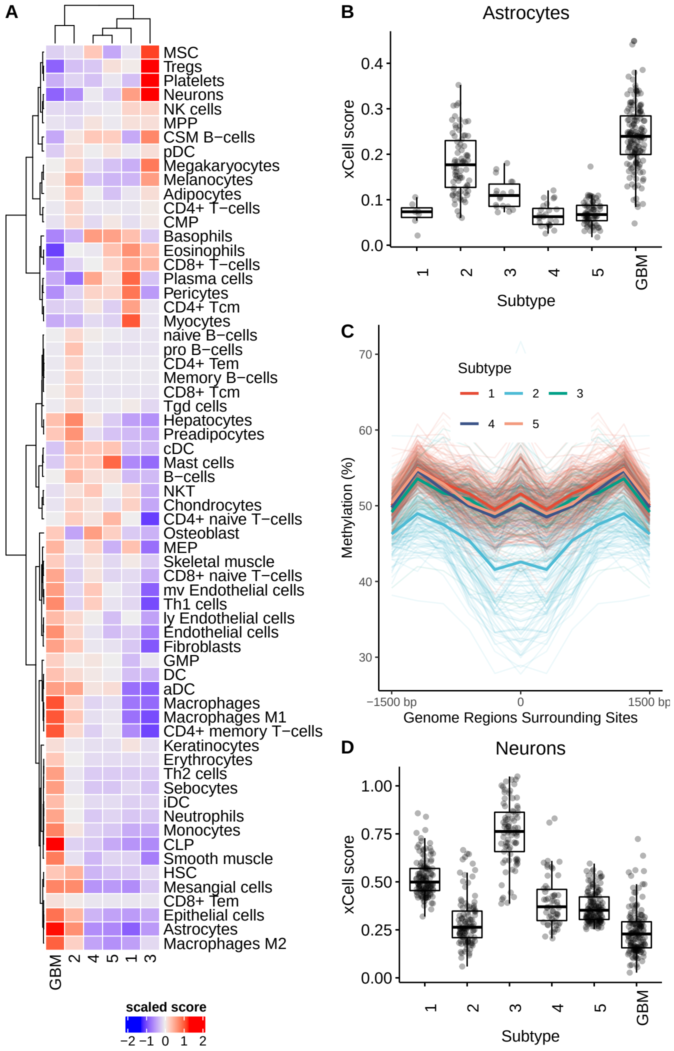

Figure 5: **Subtype 2 shows similarities to GBM, and Subtype 3 is enriched for neurons.**(A) is a heatmap that shows the mean xCell enrichment scores for the LGG subtypes and GBM corresponding to 64 cell types, with Subtype 2 and GBM sharing enrichment of several cellular populations. (B) Astrocytomas assigned to Subtype 2 show higher xCell scores compared to astrocytomas that are assigned to the other LGG subtypes. (C) Tumors assigned to Subtype 2 show lower methylation and higher regulatory activity at astrocyte-specific elements. The mean methylation levels are shown using dark line. (D) Tumors in Subtype 3 are enriched for neuronal cells.

Our enrichment analyses show that global hypomethylation is a hallmark of Subtype 2 tumors. In order to investigate this further, we used ELMER [52] in an unsupervised mode to compare Subtype 2 tumors to the other LGGs. ELMER identified 16,822 distal probes that were hypomethylated in Subtype 2 samples (adjusted p-value < 0.01 and methylation difference between means of the groups > 0.3). For 382 of those probes, their methylation status was inversely proportional to the expression of a putative target gene. These target genes are enriched for biological processes such as extracellular matrix (ECM) organization and molecular functions such as kinase binding (Fig 6A). ECM is known to be an important determinant of glioma invasion and kinase binding is activated in gliomas [53–55]. Fig 6B shows the motifs that are enriched around the 382 probes that are identified as putative distal enhancers. The motifs that show the highest enrichment correspond to the Fos and Jun transcription factor gene families. Fos genes encode leucine zipper proteins that can dimerize with proteins of the JUN family, thereby forming the early response transcription factor complex AP-1. As such, the FOS proteins have been implicated as regulators of cell proliferation, differentiation, and transformation [56]. More specifically, we find that the expression of FOSL1, which contributes to regulation of placental development is significantly higher in Subtype 2 tumors, and higher expression of the gene is associated with worse prognosis [57]. These results are in agreement with other published studies that show that AP-1 binds to demethylated regions in G-CIMP-low tumors, but we find this to be true for all samples assigned to Subtype 2 [58].

Since members of a TF family have very similar DNA binding domains, it is challenging to identify the TF that binds *in-vivo* to a region containing a motif. But we instead searched for cases where the motif occupancy of hypomethylated enhancers accompanied an increase in expression for at least one member of that TF family. Furthermore, we checked to see if the expression of the TF was significantly correlated with survival (logrank test, p-value < 0.00001). We again found an enrichment of AP-1 containing enhancers, which is a common feature of many cancer types. Interestingly, we found that TGIF1 expression was highly correlated with the degree of enhancer hypomethylation even for motifs where we did not expect TGIF1 to bind. Fig 6E and Fig 6F show that expression of TGIF1 is higher for Subtype 2 tumors and higher expression of the gene is predictive of worse prognosis. It is possible that these correlations are due to indirect effects caused by TF networks. TGIF1 is involved in regulation of cell development and maturation, and other studies have included TGIF1 in prognostic gene sets for Glioblastoma though the role of TGIF1 in gliomas is not clear [59].

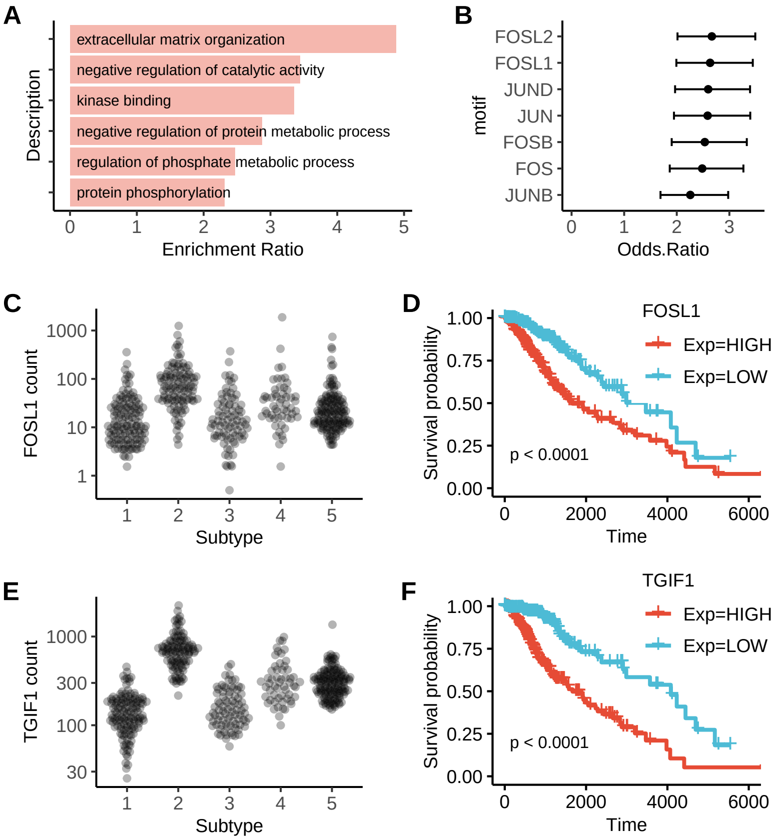

Figure 6: **Hypomethylation in Subtype 2.** (A) Enrichment of molecular function and biological processes in genes which are regulated via hypomethylation of distal enhancers in Subtype 2. (B) Motif enrichment analysis of regions around putative distal enhancers probes that regulate expression and are hypomethylated in Subtype 2 tumors. We show the 95% confidence interval of the odds ratio, and only show motifs which have a lower odds ration > 1.5. FOSL1 occupied enhancer elements show significant hypomethylation in Subtype 2 tumors. FOSL1 gene is upregulated in Subtype 2 (C), and higher expression is associated with worse survival (D). TGIF1 expression is highly correlated to enhancer hypomethylation. TGIF1 is upregulated in Subtype 2 and higher expression is associated with worse survival.

We also find significant enrichment of clinical and molecular features in other subtypes. Subtype 1 is enriched for Oligodendrogliomas (p-value $< 1.0 \times 10^{-5}$), mutations in the TERT promoter and high expression of TERT (Tukey HSD test; p-value < 0.05 for all pairwise comparisons), high tumor purity (Tukey HSD test; p-value < 0.05 for all pairwise comparisons), 1p/9q co-deletion, and mutations in CIC, a known tumor suppressor. 128 of the 130 patients in Subtype 1 have a methylated promoter for MGMT (post hoc test of residuals for χ^2 test, p-value: $< 1.0 \times 10^{-5}$). MGMT promoter methylation is associated with better response to alkylating chemotherapy, suggesting that patients assigned to Subtype 1 are more likely to respond to temozolomide [60] .

Subtype 3 is enriched for the neural (NE) subtype detected in previous gene-expression studies [61]. The NE subtype has previously been related to the tumor margin where increased normal neural tissue is likely to be detected [62]. Consistent with this hypothesis, we find that the tumors assigned to Subtype 3 have lower tumor purity (Tukey HSD test; p-value < 0.05 for all pairwise comparisons except with Subtype 5) and high enrichment score for neurons (Fig 5C. Subtype 4 and Subtype 5 are both enriched for G-CIMP high samples, although Subtype 5 is enriched for mutations in ATRX (post hoc test of residuals for χ^2 test, p-value: $< 10^{-5}$), and shows a higher enrichment for Mast cells which are known to induce release of selective inflammatory cytokines such as IL-4 with anti-glioma activity leading to improved prognosis [63].

Discussion

We present an approach to integrate multi-omic data and use it to subtype LGG through the integration of gene expression, DNA methylation and miRNA expression data. Our method is based on symmetric NMF and can be easily extended for various applications. For instance, we develop an implementation, SUMO, for unsupervised learning by regularization of the cluster indicator matrix using the Frobenius norm. It can be modified into a semi-supervised learning to classify samples after the inclusion of priors based on a phenotype of interest, as suggested in other studies [64, 65]. Additionally, we find that the primary LGG tumors show a significant difference in survival based on histological grade within existing subtypes (Fig S11). Such a semi-supervised framework will allow for integration of clinical observations with molecular information.

SUMO improves on existing methods in its ability to handle noisy and missing data. We compared SUMO to several existing methods for integrative clustering. SUMO produces consistently reproducible results on a recently published benchmark. The benchmark uses differential survival and enrichment of a small number of clinical labels in the resulting clusters as metrics for assessment of subtyping methods. However, it is important to remember that subtypes of a disease that are biologically different can lead to similar survival. For example, we find that PA-like samples from Ceccarelli et al. [41] get classified by SUMO primarily into two groups based on gene expression and miRNA-expression, even though the two groups are not significantly different in terms of survival. SUMO focuses on the integration of continuous data types such as expression, methylation, and metabolomics. Sparse and noisy data types such as somatic mutations can be included for integration after limiting the features to those that have a known role in the disease. Alternatively, such data types can be converted into continuous data types by use of network propagation techniques and then included as input to SUMO [66].

We applied SUMO for the detection of subtypes in lower-grade gliomas, and identified a single subtype with differential prognosis compared to the other subtypes. We show that this subtype includes all previously studied groups of patients with features that are associated with a poor outcome. Like GBM, gain of chr7, loss of chr10 and global hypomethylation appear to be hallmarks of this subtype, and our analyses suggest that LGGs assigned to Subtype 2 should be treated more aggressively and potentially reclassified as GBM. This subtype should also be analyzed separately in clinical trials as its molecular differences may make it susceptible to different drugs with respect to the other LGG subtypes. It is also an open question as to whether or not this subtype regrows/recurs faster after neurosurgical resection compared to the other subtypes. Additionally, we also found that the hypomethylated distal enhancers in this subtype are enriched for AP-1 binding. This has been shown to be a feature of G-CIMP-low tumors, but we find it to be characteristic of most Subtype 2 tumors. We also identified TGIF1 expression to be inversely proportional to the global hypomethylation, and predictive of prognosis, even though its role in glioma is not clear.

A common post hoc analysis to molecular subtyping is identification of feature or sets of features that can be used as markers or surrogates for the various subtypes. SUMO includes a mode to build a tree-based model that can predict the importance of each feature for each of the detected subtypes. For example, we identified an clinically relevant subtype of LGG with differential prognosis compared to the other subtypes. According to our analysis, the non-CpG island methylation probes in the proximity to the gene CLCF1 are the best marker for the subtype. Fig S12 shows the beta values of the samples for the three methylation probes that have the highest explanatory values for the classifier.

In summary, SUMO is as a molecular subtyping method that can handle noise and missing data that commonly exist in genomic datasets and can be extended for other applications. Our study suggests that NMF-based multi-omic integration is a promising approach that can be applied to a wide range of biomedical datasets and can provide valuable biological insight.

Methods

Our approach is based on non-negative matrix factorization, where the factorization is jointly performed on the similarity matrices calculated for all data types separately. After removal of outliers and data normalization, we first transform the feature matrix from the i^{th} data type into a similarity matrix A_i between the samples and then tri-factorize A_i into HS_iH^T, where H is the cluster indicator matrix that is shared across all data types. The objective function used for computing the tri-factorization accounts for the missing samples and the difference in sample size for the various assays. Lastly, to produce a robust clustering, we run the solver multiple times and apply consensus clustering to obtain the final clusters. Now, we describe these steps in details.

Data preprocessing

Data preprocessing involves (a) filtration, (b) transformation, and (c) normalization of each data type separately. The filtering process removes features that are not informative; for example, we removed genes that had zero counts in most samples. Even though our approach can handle missing values, removing features and samples with a large fraction of missing values ($> 10\%$) often speeds up computation and improves the classification if it does not remove a significant fraction of samples.

The transformation process is data-dependent. We use a variance-stabilizing transform to convert abundance in count datasets, for example as in RNA-seq, to yield a matrix of values that are approximately homoscedastic (with constant variance along the range of mean values). This had an additional advantage of reducing the effect of outliers in the dataset. We use M-values over beta values to transform methylation datasets [67]. If batch information is known, we use ComBat [68] to adjust for batch effects in this step.

In the normalization step, we perform feature standardization to make the value of each feature in the data have zero-mean and unit variance.

The construction of similarity networks and matrices

Let n be the number of patient samples s, that are found in the dataset of every data type and let t be the number of data types e.g., gene expression or DNA methylation. In this step, we construct a similarity network N, which we represent as a set of $n \times n$ similarity matrices $\{A_1, A_2, \cdots, A_t\}$, where $A_k(i,j) = (a_{ij}(k))$ and k is used as an index for the data type. $a_{ij}(k)$ represents the similarity between two samples s_i and s_j calculated from the features of the k^{th} data type , $k = 1, \cdots, t$.

For each data type k, we assume its data is represented in a matrix (f_{ij}) containing n sample rows and p feature columns. We calculate A_k as a radial basis function of the Euclidean distance $\rho(i,j) = \sum_{m=1}^{p}(f_{im} - f_{jm})^2$ between the samples x_i and x_j:

$$A(i,j) = exp\left(-\frac{\rho^2(i,j)}{\mu\epsilon_i\epsilon_j}\right)$$

where μ is a hyperparameter with a default value of 0.5 and ε_i represents the average distance between x_i and its K nearest neighbors:

$$\varepsilon_i = \frac{\sum_{j=1}^{K}\rho(i,j)}{K}.$$

We set the number of nearest neighbors K equal to 10% of the samples in the data type. The selection of this parameter can effect the results, and we recommend setting it to $\frac{\#\ samples}{\#\ clusters}$ if the number of clusters is known.

The Euclidean distance is appropriate for normalized count datasets, such as those that arise from gene expression or DNA methylation data. However, depending on the data type and the application, different distances or similarity metrics may better represent sample relationships. For example, cosine similarity has been shown to be a better metric for calculation of similarity between single cells in the single-cell sequencing for transposase accessible chromatin (scATAC-seq) [69].

Joint tri-factorization of the similarity matrices

Each matrix A_i of the multiplex network N is symmetric and non-negative. We tri-factorize $A_1, A_2, \cdots, A_t$ as follows:

$$A_i \approx HS_iH^T, \ i = 1, \cdots, t,$$

in which H is a $n \times r$ matrix shared across the data types and r is the desired number of clusters such that $r \ll n$ (Fig S10B).

We compute the above tri-factorization by minimizing the following objective function:

$$\mathcal{L} = \sum_{i=1}^{t} \lambda_i \left\| W_i \circ (A_i - HS_iH^T) \right\|_F^2 + \eta \left\| H \right\|_F^2 \tag{1}$$

where $\circ$ denotes entry-wise multiplication for matrices, and H and S_i are both constrained to be non-negative. The first term of the objective function measures the divergences between A_i and HS_iH^T using the Frobenius norm in each data type. For each data type, measurements may be not available for all the n samples, thus leading to missing entries in the matrix A_i. We use W_i to remove the missing values, where

$$W_i(x,y) = \begin{cases} 1 & \text{if } x \text{ is connected with } y \text{ in data type } i \\ 0 & \text{otherwise} \end{cases}$$

Then we add an another factor $\lambda_i = n_i^{-2}$ to account for the imbalance in the number of entries among $A_i (i = 1, ..., t)$, where n_i is the number of samples for the i^{th} data type.

The second term of the objective function is used to enforce sparsity on the matrix H, hopefully leading to a non-overfitted result and the hyperparameter η is used to balance the contribution of these two terms.

Note that the cost function in Eqn. 1 is convex in either but not both H and S_i. The following multiplicative updates are used to solve the optimization problem given in Eqn. 1 [70].

$$S_i \leftarrow S_i \circ \frac{H^T(W_i \circ A_i)H}{H^T(W_i \circ (HS_iH^T))H}$$

$$H \leftarrow H \circ \frac{\sum_i \lambda_i (W_i \circ A_i) HS_i}{\sum_i \lambda_i (W_i \circ (HS_iH^T))HS_i + 0.5\eta H}$$

As the algorithm iterates using the updates, H and S_i converge to a local minimum of the cost function. We apply above rules iteratively while alternating fixed matrices, keeping track of objective function value $\mathcal{L}^{(i)}$ until it satisfies

$$\frac{|\mathcal{L}^{(i+1)} - \mathcal{L}^{(i)}|}{\mathcal{L}^{(i+1)}} < \varepsilon$$

where ε is a predefined threshold, or the maximum number of allowed iterations are reached.

Since the solution is relatively sparse, we can assign each sample (represented by a row in H) to the cluster corresponding to the column that contains the maximum value, as depicted in Fig S10C. In practice, the solution can be sensitive to the initial conditions. We discuss the details of this in the implementation details, but briefly, we run the above solver multiple times and then use consensus clustering to get the final assignments.

Derivation of multiplicative-update rules

For the objective function Eqn. 1, when we update matrix S_i, matrices H and S_j $(j \neq i)$ should be fixed, thus it would be an optimization problem about the matrix S_i, that is,

$$\min \ \|W_i \circ (A_i - HS_iH^T)\|_F^2, \text{ subject to } S_i \geq 0. \tag{2}$$

The corresponding Lagrange function of Eq. (2) is

$$\mathcal{L}(S_i) = tr\Big(\big(W \circ (A_i - HS_iH^T)\big)^T \big(W \circ (A_i - HS_iH^T)\big)\Big) - tr(B_i^T S_i),$$

where $B_i \geq 0$ is the Lagrange multiplier for S_i, and $tr(\cdot)$ represent the trace of matrix X. Then

$$\frac{\partial \mathcal{L}(S_i)}{\partial S_i} = -2H^T\big(W_i \circ (A_i - HS_iH^T)\big)H - B_i.$$

Let $\frac{\partial \mathcal{L}(S_i)}{\partial S_i} = 0$, thus

$$H^T\big(W_i \circ (HS_iH^T)\big)H - H^T(W_i \circ A_i)H = \frac{1}{2}B_i,$$

and

$$(S_i)_{jk} \cdot (B_i)_{jk} = 0,$$

thus S_i satisfies

$$\Big(H^T\big(W_i \circ (HS_iH^T)\big)H - H^T(W_i \circ A_i)H\Big)_{jk} \cdot (S_i)_{jk} = 0.$$

We obtain the update formula for S_i as follows:

$$S_i \leftarrow S_i \circ \frac{H^T(W_i \circ A_i)H}{H^T\big(W_i \circ (HS_iH^T)\big)H},$$

where $\circ$ and $\div$ denote entry-wise multiplication and division for matrices, respectively.

Similarly, when we update matrix H,

$$\frac{\partial \mathcal{L}(H)}{\partial H} = -4\sum_{i=1}^{t} \lambda_i\big(W_i \circ (A_i - HS_iH^T)\big)HS_i + 2\eta H - B_0,$$

where $B_0 \geq 0$ is the Lagrange multiplier for H. Thus, H satisfies the following equations:

$$\Big(\sum_{i=1}^{t} \lambda_i\big(W_i \circ (HS_iH^T)\big)HS_i + 0.5\eta H - \sum_{i=1}^{t} \lambda_i(W_i \circ A_i)HS_i\Big)_{jk} \cdot (H)_{jk} = 0;$$

Then, we obtain the following update formulas for H:

$$H \leftarrow H \circ \frac{\sum_{i=1}^{t} \lambda_i(W_i \circ A_i)HS_i}{\sum_{i=1}^{t} \lambda_i\big(W_i \circ (HS_iH^T)\big)HS_i + 0.5\eta H}.$$

Implementation details

SUMO (`https://github.com/ratan-lab/sumo`) is specifically designed to integrate multi-omic data for molecular subtyping. It consists of four subroutines. It allows the user to construct the multiplex network from normalized feature matrices (*sumo prepare*), tri-factorize the multiplex network to assign samples to the desired number of clusters (*sumo run*), compares the assignments to another classification using multiple metrics (*sumo evaluate*), and detect the importance of each feature towards each cluster (*sumo interpret*), which facilitate the discovery of biomarkers and molecular signatures.

SUMO is available in the form of a command-line tool on GitHub (https://github.com/ratan-lab/sumo) and at The Python Package Index (https://pypi.org/project/python-sumo/).

Support for missing data

Biomedical studies measure a large number of molecular parameters. Almost every dataset has missing entries. Most methods for molecular subtyping require perfect data. This implies that that both samples and features that have missing entries have to be removed or the missing entries are imputed in the pre-processing stage. SUMO takes a different approach. It scales the calculated distance between a pair of samples by the number of common features available for both samples. If sufficient overlap (by default at least 10% of features) is not found, the distance is set to *NA* (not available). A missing value in an adjacent matrix A_i is equivalent to a missing edge between two nodes in the multiplex network and is masked during factorization as we describe in the last section.

Consensus clustering

As we mention in the last section, our iterative solution using multiplicative rules is sensitive to the initial conditions. Both initialization and convergence speeds are important factors to consider when formulating the appropriate factorization algorithms [71]. Our method utilizes an SVD based initialization approach to set the initial H to be the average similarity matrix across all data types. This method reduces residual error and provides faster convergence than using random initialization. However, we still have to set S_i randomly; as such, the algorithm does not guarantee convergence to a local minimum. Here, we set the diagonal entries of each S_i to be absolute singular values, that are derived from the SVD decomposition of the corresponding A_i matrix. We repeat the factorization n times, each time including 95% of the total samples in calculating the cluster assignments from H and a residual error RE_i for that run. We create a consensus matrix from these n assignments that is weighted to incorporate the residual error (RE) of each factorization in a dataset with t data type as follows.

$$C = \frac{\sum_{x=1}^{n} C(x) * weight(x)}{\sum_{x=1}^{n} weight(x)},$$

where

$$M = max_i RE(i), 1 \leq i \leq n$$

$$N = min_i RE(i), 1 \leq i \leq n$$

$$weight(x) = \frac{M - RE(x)}{M - N}$$

$$RE(x) = \sum_{i=1}^{t} \lambda_i \left\| W_i \circ (A_i - H S_i H^T) \right\|_F^2$$

We use the Normalized Cut clustering algorithm [72] on this consensus matrix to assign the final cluster labels.

Estimating the optimal number of clusters

Estimation of an optimal rank for NMF is a challenging problem. It is common to compare several solutions based on a clustering metric [73]. We implement two popular metrics that leverage the consensus matrix to help the user in the determination of stable solutions to the factorization. The first metric is the cophenetic correlation coefficient (CCC) [74]. It measures the Pearson correlation between sample distances and its hierarchical clustering. A higher CCC value is considered better. The second metric is the proportion of ambiguously clustered pairs (PAC), which is defined as the proportion of the consensus matrix values in $(0.1, 0.9)$ range. Based on our experiments, we recommend investigating factorization rank values for which the PAC score is less than 0.1, and the CCC value is high (typically > 0.95). Increasing the number of repetitions of the solver can assist in identification of the optimal number of clusters, but as we show in Fig S13 using the acute myeloid leukemia (AML) dataset from benchmark data [12], we can identify one of the stable solutions in a small number of repetitions. Similarly, we use the same dataset to show in Fig S14 that the trends observed in the PAC curve and the CCC curve are preserved for a wide range of values corresponding to the number of samples that are removed in each iteration $[0, 0.1]$. In the current default setting, we run 60 repetitions of the solver. With each run we randomly remove 5% of the samples, while making sure that each sample will be clustered at least once. We then use random subsets of 50 runs to create multiple weighted consensus matrices as described in previous section. While only one of the matrices is utilized to call sample labels, the CCC and PAC metrics are calculated for every one of them, providing a robust assessment of stability of factorization results.

Identification of biomarkers

Once the subtypes are assigned, a frequent challenge is to identify a set of features that correlate with the cluster separation. These can be used as markers for the assignment of future samples and can aid in understanding the differences between the groups. To this end, we first train a gradient boosting classifier implemented in LightGBM [75]. We use 80% of the features for training this model while performing hyperparameter optimization of the model using a random search with 5-fold cross-validation to avoid overfitting. When we have this model, we calculate the Shapley values of all features for each identified cluster. The features with a Shapley value greater than 1 are considered to be important in driving separation of that cluster.

Acknowledgments

This work was supported by NIH award (P30CA044579) to the UVA Cancer Center, and by NIGMS award GM128636 (NCS). JY Chen and LX Zhang were partially supported by the Singapore Research Foundation (Grant No. NRF2016NRF-NSFC001–026). JTL was supported by NIH training grant (NLM; 5T32LM012416) and the UVA Cancer Center.

References

[1] Gaye Lightbody, Valeriia Haberland, Fiona Browne, Laura Taggart, Huiru Zheng, Eileen Parkes, and Jaine K Blayney. Review of applications of high-throughput sequencing in personalized medicine: barriers and facilitators of future progress in research and clinical application. *Briefings in bioinformatics*, 20:1795–1811, September 2019. ISSN 1477-4054. doi: 10.1093/bib/bby051.

[2] Vinay Prasad, Tito Fojo, and Michael Brada. Precision oncology: origins, optimism, and potential. *The Lancet. Oncology*, 17:e81–e86, February 2016. ISSN 1474-5488. doi: 10.1016/S1470-2045(15)00620-8.

[3] Cancer Genome Atlas Network. Comprehensive molecular characterization of human colon and rectal cancer. *Nature*, 487:330–337, July 2012. ISSN 1476-4687. doi: 10.1038/nature11252.

[4] Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. *Nature*, 490:61–70, October 2012. ISSN 1476-4687. doi: 10.1038/nature11412.

[5] Cancer Genome Atlas Research Network. Comprehensive genomic characterization of squamous cell lung cancers. *Nature*, 489:519–525, September 2012. ISSN 1476-4687. doi: 10.1038/nature11404.

[6] Ronglai Shen, Adam B Olshen, and Marc Ladanyi. Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. *Bioinformatics (Oxford, England)*, 25:2906–2912, November 2009. ISSN 1367-4811. doi: 10.1093/bioinformatics/btp543.

[7] Qianxing Mo, Sijian Wang, Venkatraman E Seshan, Adam B Olshen, Nikolaus Schultz, Chris Sander, R Scott Powers, Marc Ladanyi, and Ronglai Shen. Pattern discovery and cancer gene identification in integrated cancer genomic data. *Proceedings of the National Academy of Sciences of the United States of America*, 110:4245–4250, March 2013. ISSN 1091-6490. doi: 10.1073/pnas.1208949110.

[8] Eric F Lock and David B Dunson. Bayesian consensus clustering. *Bioinformatics (Oxford, England)*, 29:2610–2616, October 2013. ISSN 1367-4811. doi: 10.1093/bioinformatics/btt425.

[9] Tin Nguyen, Rebecca Tagett, Diana Diaz, and Sorin Draghici. A novel approach for data integration and disease subtyping. *Genome research*, 27:2025–2039, December 2017. ISSN 1549-5469. doi: 10.1101/gr.215129.116.

[10] Bo Wang, Aziz M Mezlini, Feyyaz Demir, Marc Fiume, Zhuowen Tu, Michael Brudno, Benjamin Haibe-Kains, and Anna Goldenberg. Similarity network fusion for aggregating data types on a genomic scale. *Nature methods*, 11: 333–337, March 2014. ISSN 1548-7105. doi: 10.1038/nmeth.2810.

[11] Nimrod Rappoport and Ron Shamir. Nemo: cancer subtyping by integration of partial multi-omic data. *Bioinformatics (Oxford, England)*, 35:3348–3356, September 2019. ISSN 1367-4811. doi: 10.1093/bioinformatics/btz058.

[12] Nimrod Rappoport and Ron Shamir. Multi-omic and multi-view clustering algorithms: review and cancer benchmark. *Nucleic acids research*, 46:10546–10562, November 2018. ISSN 1362-4962. doi: 10.1093/nar/gky889.

[13] Jinyu Chen and Louxin Zhang. A survey and systematic assessment of computational methods for drug response prediction. *Briefings in Bioinformatics*, 2020.

[14] Dingming Wu, Dongfang Wang, Michael Q Zhang, and Jin Gu. Fast dimension reduction and integrative clustering of multi-omics data using low-rank approximation: application to cancer molecular classification. *BMC genomics*, 16 (1):1022, 2015.

[15] Daniela M Witten and Robert J Tibshirani. Extensions of sparse canonical correlation analysis with applications to genomic data. *Statistical applications in genetics and molecular biology*, 8:Article28, 2009. ISSN 1544-6115. doi: 10.2202/1544-6115.1470.

[16] Cancer Genome Atlas Research Network, Daniel J Brat, Roel G W Verhaak, Kenneth D Aldape, W K Alfred Yung, Sofie R Salama, Lee A D Cooper, Esther Rheinbay, C Ryan Miller, Mark Vitucci, Olena Morozova, A Gordon Robertson, Houtan Noushmehr, Peter W Laird, Andrew D Cherniack, Rehan Akbani, Jason T Huse, Giovanni Ciriello, Laila M Poisson, Jill S Barnholtz-Sloan, Mitchel S Berger, Cameron Brennan, Rivka R Colen, Howard Colman, Adam E Flanders, Caterina Giannini, Mia Grifford, Antonio Iavarone, Rajan Jain, Isaac Joseph, Jaegil Kim, Katayoon Kasaian, Tom Mikkelsen, Bradley A Murray, Brian Patrick O'Neill, Lior Pachter, Donald W Parsons, Carrie Sougnez, Erik P Sulman, Scott R Vandenberg, Erwin G Van Meir, Andreas von Deimling, Hailei Zhang, Daniel Crain, Kevin Lau, David Mallery, Scott Morris, Joseph Paulauskis, Robert Penny, Troy Shelton, Mark Sherman, Peggy Yena, Aaron Black, Jay Bowen, Katie Dicostanzo, Julie Gastier-Foster, Kristen M Leraas, Tara M Lichtenberg, Christopher R Pierson, Nilsa C Ramirez, Cynthia Taylor, Stephanie Weaver, Lisa Wise, Erik Zmuda, Tanja Davidsen, John A Demchok, Greg Eley, Martin L Ferguson, Carolyn M Hutter, Kenna R Mills Shaw, Bradley A Ozenberger, Margi Sheth, Heidi J Sofia, Roy Tarnuzzer, Zhining Wang, Liming Yang, Jean Claude Zenklusen, Brenda Ayala, Julien Baboud, Sudha Chudamani, Mark A Jensen, Jia Liu, Todd Pihl, Rohini Raman, Yunhu Wan, Ye Wu, Adrian Ally, J Todd Auman, Miruna Balasundaram, Saianand Balu, Stephen B Baylin, Rameen Beroukhim, Moiz S Bootwalla, Reanne Bowlby, Christopher A Bristow, Denise Brooks, Yaron Butterfield, Rebecca Carlsen, Scott Carter, Lynda Chin, Andy Chu, Eric Chuah, Kristian Cibulskis, Amanda Clarke, Simon G Coetzee, Noreen Dhalla, Tim Fennell, Sheila Fisher, Stacey Gabriel, Gad Getz, Richard Gibbs, Ranabir Guin, Angela Hadjipanayis, D Neil Hayes, Toshinori Hinoue, Katherine Hoadley, Robert A Holt, Alan P Hoyle, Stuart R Jefferys, Steven Jones, Corbin D Jones, Raju Kucherlapati, Phillip H Lai, Eric Lander, Semin Lee, Lee Lichtenstein, Yussanne Ma, Dennis T Maglinte, Harshad S Mahadeshwar, Marco A Marra, Michael Mayo, Shaowu Meng, Matthew L Meyerson, Piotr A Mieczkowski, Richard A Moore, Lisle E Mose, Andrew J Mungall, Angeliki Pantazi, Michael Parfenov, Peter J Park, Joel S Parker, Charles M Perou, Alexei Protopopov, Xiaojia Ren, Jeffrey Roach, Thaís S Sabedot, Jacqueline Schein, Steven E Schumacher, Jonathan G Seidman, Sahil Seth, Hui Shen, Janae V Simons, Payal Sipahimalani, Matthew G Soloway, Xingzhi Song, Huandong Sun, Barbara Tabak, Angela Tam, Donghui Tan, Jiabin Tang, Nina Thiessen, Timothy Triche, David J Van Den Berg, Umadevi Veluvolu, Scot Waring, Daniel J Weisenberger, Matthew D Wilkerson, Tina Wong, Junyuan Wu, Liu Xi, Andrew W Xu, Lixing Yang, Travis I Zack, Jianhua Zhang, B Arman Aksoy, Harindra Arachchi, Chris Benz, Brady Bernard, Daniel Carlin, Juok Cho, Daniel DiCara, Scott Frazer, Gregory N Fuller, JianJiong Gao, Nils Gehlenborg, David Haussler, David I Heiman, Lisa Iype, Anders Jacobsen, Zhenlin Ju, Sol Katzman, Hoon Kim, Theo Knijnenburg, Richard Bailey Kreisberg, Michael S Lawrence, William Lee, Kalle Leinonen, Pei Lin, Shiyun Ling, Wenbin Liu, Yingchun Liu, Yuexin Liu, Yiling Lu, Gordon Mills, Sam Ng, Michael S Noble, Evan Paull, Arvind Rao, Sheila Reynolds, Gordon Saksena, Zack Sanborn, Chris Sander, Nikolaus Schultz, Yasin Senbabaoglu, Ronglai Shen, Ilya Shmulevich, Rileen Sinha, Josh Stuart, S Onur Sumer, Yichao Sun, Natalie Tasman, Barry S Taylor, Doug Voet, Nils Weinhold, John N Weinstein, Da Yang, Kosuke Yoshihara, Siyuan Zheng, Wei Zhang, Lihua Zou, Ty Abel, Sara

Sadeghi, Mark L Cohen, Jenny Eschbacher, Eyas M Hattab, Aditya Raghunathan, Matthew J Schniederjan, Dina Aziz, Gene Barnett, Wendi Barrett, Darell D Bigner, Lori Boice, Cathy Brewer, Chiara Calatozzolo, Benito Campos, Carlos Gilberto Carlotti, Timothy A Chan, Lucia Cuppini, Erin Curley, Stefania Cuzzubbo, Karen Devine, Francesco DiMeco, Rebecca Duell, J Bradley Elder, Ashley Fehrenbach, Gaetano Finocchiaro, William Friedman, Jordonna Fulop, Johanna Gardner, Beth Hermes, Christel Herold-Mende, Christine Jungk, Ady Kendler, Norman L Lehman, Eric Lipp, Ouida Liu, Randy Mandt, Mary McGraw, Roger Mclendon, Christopher McPherson, Luciano Neder, Phuong Nguyen, Ardene Noss, Raffaele Nunziata, Quinn T Ostrom, Cheryl Palmer, Alessandro Perin, Bianca Pollo, Alexander Potapov, Olga Potapova, W Kimryn Rathmell, Daniil Rotin, Lisa Scarpace, Cathy Schilero, Kelly Senecal, Kristen Shimmel, Vsevolod Shurkhay, Suzanne Sifri, Rosy Singh, Andrew E Sloan, Kathy Smolenski, Susan M Staugaitis, Ruth Steele, Leigh Thorne, Daniela P C Tirapelli, Andreas Unterberg, Mahitha Vallurupalli, Yun Wang, Ronald Warnick, Felicia Williams, Yingli Wolinsky, Sue Bell, Mara Rosenberg, Chip Stewart, Franklin Huang, Jonna L Grimsby, Amie J Radenbaugh, and Jianan Zhang. Comprehensive, integrative genomic analysis of diffuse lower-grade gliomas. *The New England journal of medicine*, 372:2481–2498, June 2015. ISSN 1533-4406. doi: 10.1056/NEJMoa1402121.

[17] David N Louis, Arie Perry, Guido Reifenberger, Andreas von Deimling, Dominique Figarella-Branger, Webster K Cavenee, Hiroko Ohgaki, Otmar D Wiestler, Paul Kleihues, and David W Ellison. The 2016 world health organization classification of tumors of the central nervous system: a summary. *Acta neuropathologica*, 131:803–820, June 2016. ISSN 1432-0533. doi: 10.1007/s00401-016-1545-1.

[18] Jeanette E Eckel-Passow, Daniel H Lachance, Annette M Molinaro, Kyle M Walsh, Paul A Decker, Hugues Sicotte, Melike Pekmezci, Terri Rice, Matt L Kosel, Ivan V Smirnov, Gobinda Sarkar, Alissa A Caron, Thomas M Kollmeyer, Corinne E Praska, Anisha R Chada, Chandralekha Halder, Helen M Hansen, Lucie S McCoy, Paige M Bracci, Roxanne Marshall, Shichun Zheng, Gerald F Reis, Alexander R Pico, Brian P O'Neill, Jan C Buckner, Caterina Giannini, Jason T Huse, Arie Perry, Tarik Tihan, Mitchell S Berger, Susan M Chang, Michael D Prados, Joseph Wiemels, John K Wiencke, Margaret R Wrensch, and Robert B Jenkins. Glioma groups based on 1p/19q, idh, and tert promoter mutations in tumors. *The New England journal of medicine*, 372:2499–2508, June 2015. ISSN 1533-4406. doi: 10.1056/NEJMoa1407279.

[19] Pentti Paatero and Unto Tapper. Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values. *Environmetrics*, 5(2):111–126, 1994.

[20] D D Lee and H S Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401:788–791, October 1999. ISSN 0028-0836. doi: 10.1038/44565.

[21] David Guillamet and Jordi Vitria. Non-negative matrix factorization for face recognition. In *Catalonian Conference on Artificial Intelligence*, pages 336–344. Springer, 2002.

[22] Johannes Leuschner, Maximilian Schmidt, Pascal Fernsel, Delf Lachmund, Tobias Boskamp, and Peter Maass. Supervised non-negative matrix factorization methods for maldi imaging applications. *Bioinformatics (Oxford, England)*, 35:1940–1947, June 2019. ISSN 1367-4811. doi: 10.1093/bioinformatics/bty909.

[23] Ryuichi Maruyama, Kazuma Maeda, Hajime Moroda, Ichiro Kato, Masashi Inoue, Hiroyoshi Miyakawa, and Toru Aonishi. Detecting cells using non-negative matrix factorization on calcium imaging data. *Neural networks : the official journal of the International Neural Network Society*, 55:11–19, July 2014. ISSN 1879-2782. doi: 10.1016/j.neunet.2014.03.007.

[24] Nicolas Sauwen, Marjan Acou, Diana M Sima, Jelle Veraart, Frederik Maes, Uwe Himmelreich, Eric Achten, and Sabine Van Huffel. Semi-automated brain tumor segmentation on multi-parametric mri using regularized non-negative matrix factorization. *BMC medical imaging*, 17:29, May 2017. ISSN 1471-2342. doi: 10.1186/s12880-017-0198-4.

[25] Lin Chen, Kirsten Vallmuur, and Richi Nayak. Injury narrative text classification using factorization model. *BMC medical informatics and decision making*, 15 Suppl 1:S5, 2015. ISSN 1472-6947. doi: 10.1186/1472-6947-15-S1-S5.

[26] Yifeng Li and Alioune Ngom. The non-negative matrix factorization toolbox for biological data mining. *Source code for biology and medicine*, 8:10, April 2013. ISSN 1751-0473. doi: 10.1186/1751-0473-8-10.

[27] Yuan Luo, Yu Xin, Ephraim Hochberg, Rohit Joshi, Ozlem Uzuner, and Peter Szolovits. Subgraph augmented non-negative tensor factorization (santf) for modeling clinical narrative text. *Journal of the American Medical Informatics Association : JAMIA*, 22:1009–1019, September 2015. ISSN 1527-974X. doi: 10.1093/jamia/ocv016.

[28] Dingding Wang, Mitsunori Ogihara, Carlos Gallo, Juan A Villamar, Justin D Smith, Wouter Vermeer, Gracelyn Cruden, Nanette Benbow, and C Hendricks Brown. Automatic classification of communication logs into implementation stages via text analysis. *Implementation science : IS*, 11:119, September 2016. ISSN 1748-5908. doi: 10.1186/s13012-016-0483-6.

[29] Hong-Qiang Wang, Chun-Hou Zheng, and Xing-Ming Zhao. jnmfma: a joint non-negative matrix factorization meta-analysis of transcriptomics data. *Bioinformatics (Oxford, England)*, 31:572–580, February 2015. ISSN 1367-4811. doi: 10.1093/bioinformatics/btu679.

[30] 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. *Nucleic acids research*, 40:9379–9391, October 2012. ISSN 1362-4962. doi: 10.1093/nar/gks725.

[31] Belhassen Bayar, Nidhal Bouaynaya, and Roman Shterenberg. Probabilistic non-negative matrix factorization: theory and application to microarray data analysis. *Journal of bioinformatics and computational biology*, 12:1450001, February 2014. ISSN 1757-6334. doi: 10.1142/S0219720014500012.

[32] Yun Cai, Hong Gu, and Toby Kenney. Learning microbial community structures with supervised and unsupervised non-negative matrix factorization. *Microbiome*, 5:110, August 2017. ISSN 2049-2618. doi: 10.1186/s40168-017-0323-1.

[33] Stéphane Chrétien, Christophe Guyeux, Bastien Conesa, Régis Delage-Mouroux, Michèle Jouvenot, Philippe Huetz, and Françoise Descôtes. A bregman-proximal point algorithm for robust non-negative matrix factorization with possible missing values and outliers - application to gene expression analysis. *BMC bioinformatics*, 17 Suppl 8:284, August 2016. ISSN 1471-2105. doi: 10.1186/s12859-016-1120-8.

[34] Xue Jiang, Han Zhang, Zhao Zhang, and Xiongwen Quan. Flexible non-negative matrix factorization to unravel disease-related genes. *IEEE/ACM transactions on computational biology and bioinformatics*, 16:1948–1957, 2019. ISSN 1557-9964. doi: 10.1109/TCBB.2018.2823746.

[35] Da Kuang, Chris Ding, and Haesun Park. Symmetric nonnegative matrix factorization for graph clustering. In *Proceedings of the 2012 SIAM international conference on data mining*, pages 106–117. SIAM, 2012.

[36] Jinyu Chen and Shihua Zhang. Discovery of two-level modular organization from matched genomic data via joint matrix tri-factorization. *Nucleic Acids Research*, 46(12):5967–5976, 06 2018. ISSN 0305-1048. doi: 10.1093/nar/gky440. URL `https://doi.org/10.1093/nar/gky440`.

[37] Nimrod Rappoport and Ron Shamir. Inaccuracy of the log-rank approximation in cancer data analysis. *Molecular systems biology*, 15:e8754, August 2019. ISSN 1744-4292. doi: 10.15252/msb.20188754.

[38] Siyuan Ma, Shuji Ogino, Princy Parsana, Reiko Nishihara, Zhirong Qian, Jeanne Shen, Kosuke Mima, Yohei Masugi, Yin Cao, Jonathan A Nowak, Kaori Shima, Yujin Hoshida, Edward L Giovannucci, Manish K Gala, Andrew T Chan, Charles S Fuchs, Giovanni Parmigiani, Curtis Huttenhower, and Levi Waldron. Continuity of transcriptomes among colorectal cancer subtypes based on meta-analysis. *Genome biology*, 19:142, September 2018. ISSN 1474-760X. doi: 10.1186/s13059-018-1511-4.

[39] Cancer Genome Atlas Research Network. Integrated genomic analyses of ovarian carcinoma. *Nature*, 474:609–615, June 2011. ISSN 1476-4687. doi: 10.1038/nature10166.

[40] Yulia Newton, Adam M Novak, Teresa Swatloski, Duncan C McColl, Sahil Chopra, Kiley Graim, Alana S Weinstein, Robert Baertsch, Sofie R Salama, Kyle Ellrott, Manu Chopra, Theodore C Goldstein, David Haussler, Olena Morozova, and Joshua M Stuart. Tumormap: Exploring the molecular similarities of cancer samples in an interactive portal. *Cancer research*, 77:e111–e114, November 2017. ISSN 1538-7445. doi: 10.1158/0008-5472.CAN-17-0580.

[41] 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, TCGA Research Network, Houtan Noushmehr, Antonio Iavarone, and Roel G W Verhaak. Molecular profiling reveals biologically discrete subsets and pathways of progression in diffuse glioma. *Cell*, 164:550–563, January 2016. ISSN 1097-4172. doi: 10.1016/j.cell.2015.12.028.

[42] Mary Goldman, Brian Craft, Mim Hastie, Kristupas Repečka, Akhil Kamath, Fran McDade, Dave Rogers, Angela N. Brooks, Jingchun Zhu, and David Haussler. The ucsc xena platform for public and private cancer genomics data visualization and interpretation, 2019.

[43] Yasin Șenbabaoğlu, George Michailidis, and Jun Z Li. Critical limitations of consensus clustering in class discovery. *Scientific reports*, 4:6207, August 2014. ISSN 2045-2322. doi: 10.1038/srep06207.

[44] Lucie N Hutchins, Sean M Murphy, Priyam Singh, and Joel H Graber. Position-dependent motif characterization using non-negative matrix factorization. *Bioinformatics (Oxford, England)*, 24:2684–2690, December 2008. ISSN 1367-4811. doi: 10.1093/bioinformatics/btn526.

[45] Kyle Ellrott, Matthew H Bailey, Gordon Saksena, Kyle R Covington, Cyriac Kandoth, Chip Stewart, Julian Hess, Singer Ma, Kami E Chiotti, Michael McLellan, Heidi J Sofia, Carolyn Hutter, Gad Getz, David Wheeler, Li Ding, MC3 Working Group, and Cancer Genome Atlas Research Network. Scalable open science approach for mutation calling of tumor exomes using multiple genomic pipelines. *Cell systems*, 6:271–281.e7, March 2018. ISSN 2405-4712. doi: 10.1016/j.cels.2018.03.002.

[46] Kosuke Yoshihara, Maria Shahmoradgoli, Emmanuel Martínez, Rahulsimham Vegesna, Hoon Kim, Wandaliz Torres-Garcia, Victor Treviño, Hui Shen, Peter W Laird, Douglas A Levine, Scott L Carter, Gad Getz, Katherine Stemke-Hale, Gordon B Mills, and Roel G W Verhaak. Inferring tumour purity and stromal and immune cell admixture from expression data. *Nature communications*, 4:2612, 2013. ISSN 2041-1723. doi: 10.1038/ncomms3612.

[47] Jun Su, Wenyong Long, Qianquan Ma, Kai Xiao, Yang Li, Qun Xiao, Gang Peng, Jian Yuan, and Qing Liu. Identification of a tumor microenvironment-related eight-gene signature for predicting prognosis in lower-grade gliomas. *Frontiers in genetics*, 10:1143, 2019. ISSN 1664-8021. doi: 10.3389/fgene.2019.01143.

[48] Cameron W Brennan, Roel G W Verhaak, Aaron McKenna, Benito Campos, Houtan Noushmehr, Sofie R Salama, Siyuan Zheng, Debyani Chakravarty, J Zachary Sanborn, Samuel H Berman, Rameen Beroukhim, Brady Bernard, Chang-Jiun Wu, Giannicola Genovese, Ilya Shmulevich, Jill Barnholtz-Sloan, Lihua Zou, Rahulsimham Vegesna, Sachet A Shukla, Giovanni Ciriello, W K Yung, Wei Zhang, Carrie Sougnez, Tom Mikkelsen, Kenneth Aldape, Darell D Bigner, Erwin G Van Meir, Michael Prados, Andrew Sloan, Keith L Black, Jennifer Eschbacher, Gaetano Finocchiaro, William Friedman, David W Andrews, Abhijit Guha, Mary Iacocca, Brian P O'Neill, Greg Foltz, Jerome Myers, Daniel J Weisenberger, Robert Penny, Raju Kucherlapati, Charles M Perou, D Neil Hayes, Richard Gibbs, Marco Marra, Gordon B Mills, Eric Lander, Paul Spellman, Richard Wilson, Chris Sander, John Weinstein, Matthew Meyerson, Stacey Gabriel, Peter W Laird, David Haussler, Gad Getz, Lynda Chin, and TCGA Research Network. The somatic genomic landscape of glioblastoma. *Cell*, 155:462–477, October 2013. ISSN 1097-4172. doi: 10.1016/j.cell.2013.09.034.

[49] Damian Stichel, Azadeh Ebrahimi, David Reuss, Daniel Schrimpf, Takahiro Ono, Mitsuaki Shirahata, Guido Reifenberger, Michael Weller, Daniel Hänggi, Wolfgang Wick, Christel Herold-Mende, Manfred Westphal, Sebastian Brandner, Stefan M Pfister, David Capper, Felix Sahm, and Andreas von Deimling. Distribution of egfr amplification, combined chromosome 7 gain and chromosome 10 loss, and tert promoter mutation in brain tumors and their potential for the reclassification of idhwt astrocytoma to glioblastoma. *Acta neuropathologica*, 136:793–803, November 2018. ISSN 1432-0533. doi: 10.1007/s00401-018-1905-0.

[50] Dvir Aran, Zicheng Hu, and Atul J Butte. xcell: digitally portraying the tissue cellular heterogeneity landscape. *Genome biology*, 18:220, November 2017. ISSN 1474-760X. doi: 10.1186/s13059-017-1349-1.

[51] John T Lawson, Eleni M Tomazou, Christoph Bock, and Nathan C Sheffield. Mira: an r package for dna methylation-based inference of regulatory activity. *Bioinformatics (Oxford, England)*, 34:2649–2650, August 2018. ISSN 1367-4811. doi: 10.1093/bioinformatics/bty083.

[52] Tiago C Silva, Simon G Coetzee, Nicole Gull, Lijing Yao, Dennis J Hazelett, Houtan Noushmehr, De-Chen Lin, and Benjamin P Berman. Elmer v.2: an r/bioconductor package to reconstruct gene regulatory networks from dna methylation and transcriptome profiles. *Bioinformatics (Oxford, England)*, 35:1974–1977, June 2019. ISSN 1367-4811. doi: 10.1093/bioinformatics/bty902.

[53] R H Goldbrunner, J J Bernstein, and J C Tonn. Ecm-mediated glioma cell invasion. *Microscopy research and technique*, 43:250–257, November 1998. ISSN 1059-910X. doi: 10.1002/(SICI)1097-0029(19981101)43:3⟨250::AID-JEMT7⟩3.0.CO;2-C.

[54] R H Goldbrunner, J J Bernstein, and J C Tonn. Cell-extracellular matrix interaction in glioma invasion. *Acta neurochirurgica*, 141:295–305; discussion 304–5, 1999. ISSN 0001-6268. doi: 10.1007/s007010050301.

[55] Valéria Pereira Ferrer, Vivaldo Moura Neto, and Rolf Mentlein. Glioma infiltration and extracellular matrix: key players and modulators. *Glia*, 66:1542–1565, August 2018. ISSN 1098-1136. doi: 10.1002/glia.23309.

[56] Shwetal Mehta and Costanza Lo Cascio. Developmentally regulated signaling pathways in glioma invasion. *Cellular and molecular life sciences : CMLS*, 75:385–402, February 2018. ISSN 1420-9071. doi: 10.1007/s00018-017-2608-8.

[57] Kaiyu Kubota, Lindsey N Kent, M A Karim Rumi, Katherine F Roby, and Michael J Soares. Dynamic regulation of ap-1 transcriptional complexes directs trophoblast differentiation. *Molecular and cellular biology*, 35:3163–3177, September 2015. ISSN 1098-5549. doi: 10.1128/MCB.00118-15.

[58] Camila Ferreira de Souza, Thais S Sabedot, Tathiane M Malta, Lindsay Stetson, Olena Morozova, Artem Sokolov, Peter W Laird, Maciej Wiznerowicz, Antonio Iavarone, James Snyder, Ana deCarvalho, Zachary Sanborn, Kerrie L McDonald, William A Friedman, Daniela Tirapelli, Laila Poisson, Tom Mikkelsen, Carlos G Carlotti, Steven Kalkanis, Jean Zenklusen, Sofie R Salama, Jill S Barnholtz-Sloan, and Houtan Noushmehr. A distinct dna methylation shift in a subset of glioma cpg island methylator phenotypes during tumor recurrence. *Cell reports*, 23:637–651, April 2018. ISSN 2211-1247. doi: 10.1016/j.celrep.2018.03.107.

[59] Quan Cheng, Chunhai Huang, Hui Cao, Jinhu Lin, Xuan Gong, Jian Li, Yuanbing Chen, Zhi Tian, Zhenyu Fang, and Jun Huang. A novel prognostic signature of transcription factors for the prediction in patients with gbm. *Frontiers in genetics*, 10:906, 2019. ISSN 1664-8021. doi: 10.3389/fgene.2019.00906.

[60] Andreana L Rivera, Christopher E Pelloski, Mark R Gilbert, Howard Colman, Clarissa De La Cruz, Erik P Sulman, B Nebiyou Bekele, and Kenneth D Aldape. Mgmt promoter methylation is predictive of response to radiotherapy and prognostic in the absence of adjuvant alkylating chemotherapy for glioblastoma. *Neuro-oncology*, 12:116–121, February 2010. ISSN 1523-5866. doi: 10.1093/neuonc/nop020.

[61] Roel G W Verhaak, Katherine A Hoadley, Elizabeth Purdom, Victoria Wang, Yuan Qi, Matthew D Wilkerson, C Ryan Miller, Li Ding, Todd Golub, Jill P Mesirov, Gabriele Alexe, Michael Lawrence, Michael O'Kelly, Pablo Tamayo, Barbara A Weir, Stacey Gabriel, Wendy Winckler, Supriya Gupta, Lakshmi Jakkula, Heidi S Feiler, J Graeme Hodgson, C David James, Jann N Sarkaria, Cameron Brennan, Ari Kahn, Paul T Spellman, Richard K Wilson, Terence P Speed, Joe W Gray, Matthew Meyerson, Gad Getz, Charles M Perou, D Neil Hayes, and Cancer Genome Atlas Research Network. Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in pdgfra, idh1, egfr, and nf1. *Cancer cell*, 17:98–110, January 2010. ISSN 1878-3686. doi: 10.1016/j.ccr.2009.12.020.

[62] Brian J Gill, David J Pisapia, Hani R Malone, Hannah Goldstein, Liang Lei, Adam Sonabend, Jonathan Yun, Jorge Samanamud, Jennifer S Sims, Matei Banu, Athanassios Dovas, Andrew F Teich, Sameer A Sheth, Guy M McKhann,

Michael B Sisti, Jeffrey N Bruce, Peter A Sims, and Peter Canoll. Mri-localized biopsies reveal subtype-specific differences in molecular and cellular composition at the margins of glioblastoma. *Proceedings of the National Academy of Sciences of the United States of America*, 111:12550–12555, August 2014. ISSN 1091-6490. doi: 10.1073/pnas.1405839111.

[63] S Benedetti, B Pirola, B Pollo, L Magrassi, M G Bruzzone, D Rigamonti, R Galli, S Selleri, F Di Meco, C De Fraja, A Vescovi, E Cattaneo, and G Finocchiaro. Gene therapy of experimental brain tumors using neural progenitor cells. *Nature medicine*, 6:447–450, April 2000. ISSN 1078-8956. doi: 10.1038/74710.

[64] Jaegul Choo, Changhyun Lee, Chandan K Reddy, and Haesun Park. Weakly supervised nonnegative matrix factorization for user-driven clustering. *Data mining and knowledge discovery*, 29(6):1598–1621, 2015.

[65] Tao Li, Chris Ding, and Michael I Jordan. Solving consensus and semi-supervised clustering problems using nonnegative matrix factorization. In *Seventh IEEE International Conference on Data Mining (ICDM 2007)*, pages 577–582. IEEE, 2007.

[66] Lenore Cowen, Trey Ideker, Benjamin J Raphael, and Roded Sharan. Network propagation: a universal amplifier of genetic associations. *Nature reviews. Genetics*, 18:551–562, September 2017. ISSN 1471-0064. doi: 10.1038/nrg.2017.38.

[67] Pan Du, Xiao Zhang, Chiang-Ching Huang, Nadereh Jafari, Warren A Kibbe, Lifang Hou, and Simon M Lin. Comparison of beta-value and m-value methods for quantifying methylation levels by microarray analysis. *BMC bioinformatics*, 11:587, November 2010. ISSN 1471-2105. doi: 10.1186/1471-2105-11-587.

[68] W Evan Johnson, Cheng Li, and Ariel Rabinovic. Adjusting batch effects in microarray expression data using empirical bayes methods. *Biostatistics (Oxford, England)*, 8:118–127, January 2007. ISSN 1465-4644. doi: 10.1093/biostatistics/kxj037.

[69] Stanley Cai, Georgios K Georgakilas, John L Johnson, and Golnaz Vahedi. A cosine similarity-based method to infer variability of chromatin accessibility at the single-cell level. *Frontiers in genetics*, 9:319, 2018. ISSN 1664-8021. doi: 10.3389/fgene.2018.00319.

[70] H. Sebastian Seung Daniel D. Lee. Algorithms for non-negative matrix factorization, 2001.

[71] Christos Boutsidis and Efstratios Gallopoulos. SVD based initialization: A head start for nonnegative matrix factorization. *Pattern Recognit.*, 41(4):1350–1362, 2008. doi: 10.1016/j.patcog.2007.09.010.

[72] and Jianbo Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):888–905, 2000.

[73] Renaud Gaujoux and Cathal Seoighe. A flexible r package for nonnegative matrix factorization. *BMC bioinformatics*, 11:367, July 2010. ISSN 1471-2105. doi: 10.1186/1471-2105-11-367.

[74] Jean-Philippe Brunet, Pablo Tamayo, Todd R Golub, and Jill P Mesirov. Metagenes and molecular pattern discovery using matrix factorization. *Proceedings of the National Academy of Sciences of the United States of America*, 101:4164–4169, March 2004. ISSN 0027-8424. doi: 10.1073/pnas.0308531101.

[75] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. Lightgbm: A highly efficient gradient boosting decision tree. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 3149–3157, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

Supporting information

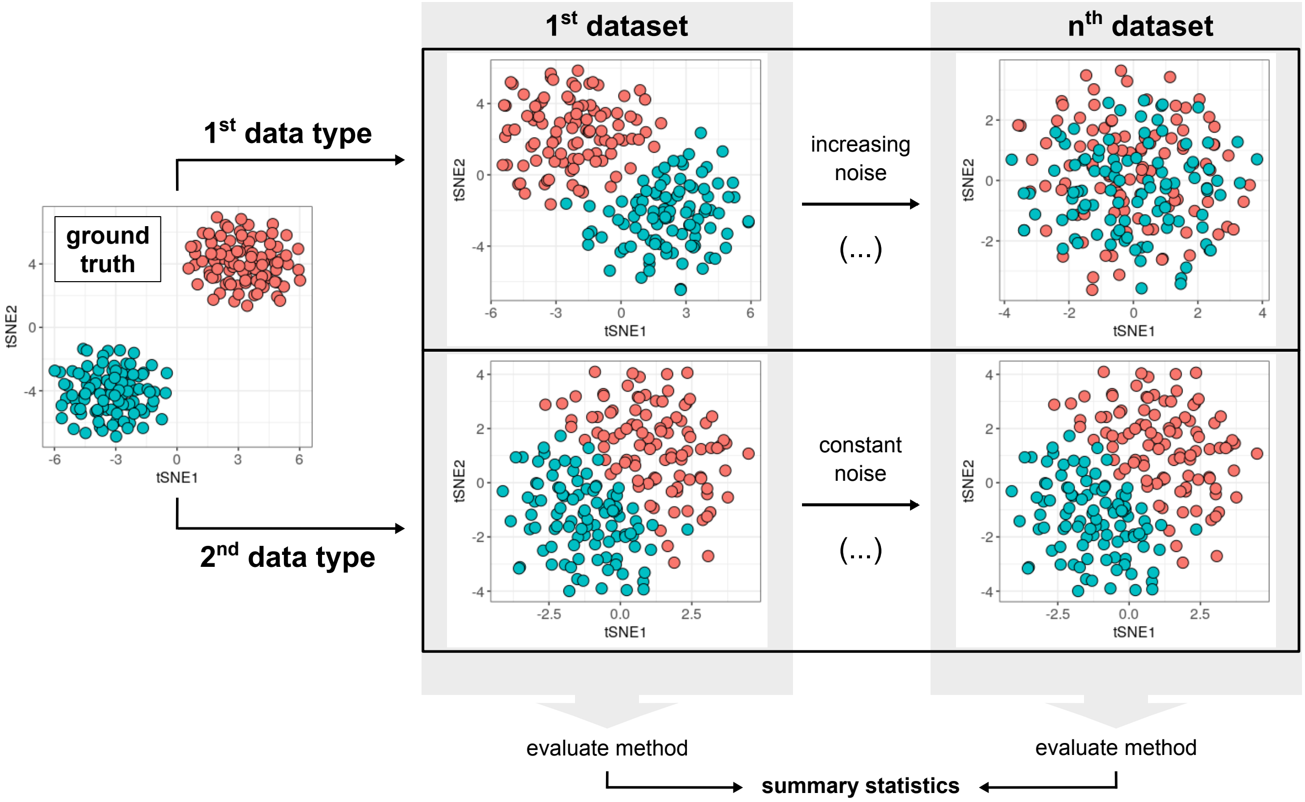

Figure S1: **Experimental setup to compare the accuracy of the various methods on noisy data.** During simulation, noise in one of the data types remains constant (generated from Gaussian distribution $\mathcal{N}(\mu = 0, \sigma = 1.5)$). Another data type contains data with increasing noise, randomly generated from the Gaussian distribution as a function of standard deviation, and constant mean. Here we show two sampling points for following sets of parameters $(\mu, \sigma) \in \{(0, 1), (0, 4)\}$.

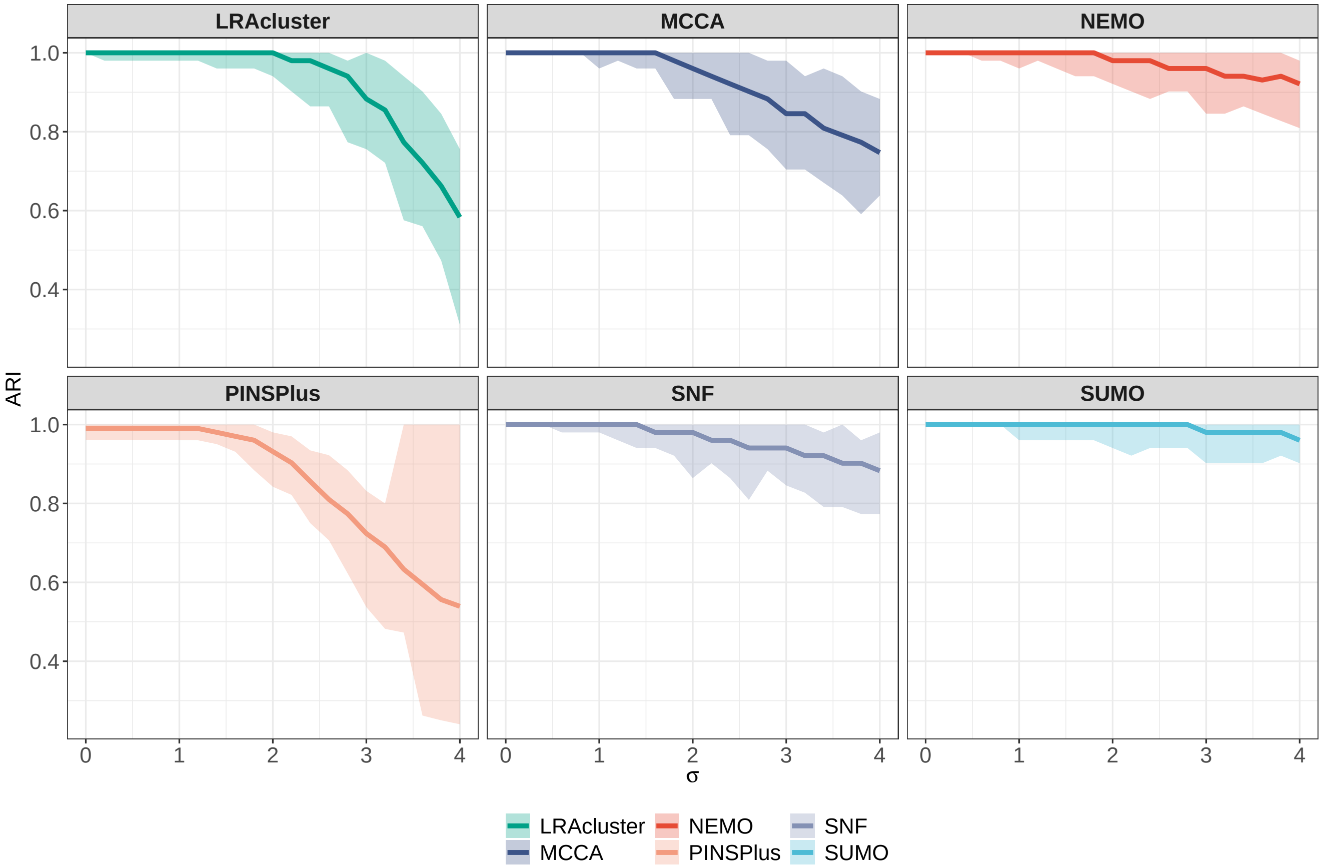

Figure S2: **Adjusted Rand Index (ARI) from running the various methods on simulated noisy datasets.** Datasets were created by adding either adding $\mathcal{N}(\mu = 0, \sigma = 1.5)$ noise (constant layer) or $\mathcal{N}(\mu = 0)$ with standard deviation $\sigma \in (0, 4)$ (layer with varying amount of noise). We report ARI of the classification at each data point for 100 repetitions.

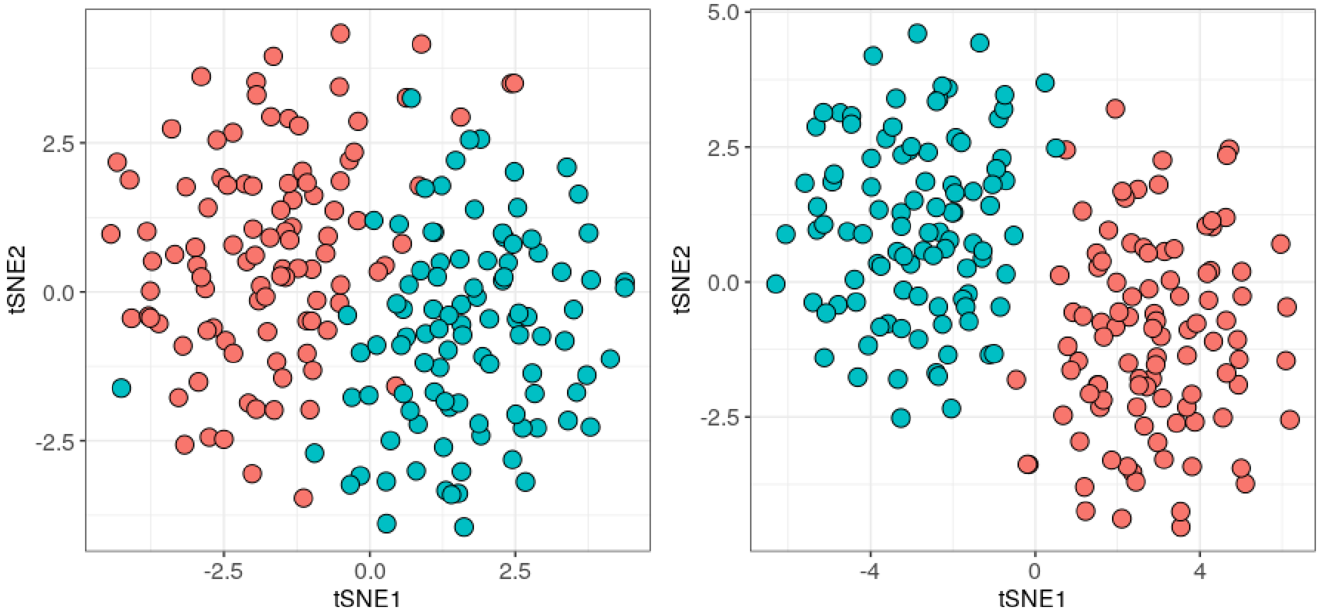

Figure S3: **Experimental setup to compare the accuracy of the various methods with varying sample size.** Cluster separability in stability simulations show here using tSNE. Data type shown in the left panel contains noise from $\mathcal{N}(\mu = 0, \sigma = 1.5)$ distribution, while the one on the right has noise from $\mathcal{N}(\mu = 1, \sigma = 1)$ distribution.

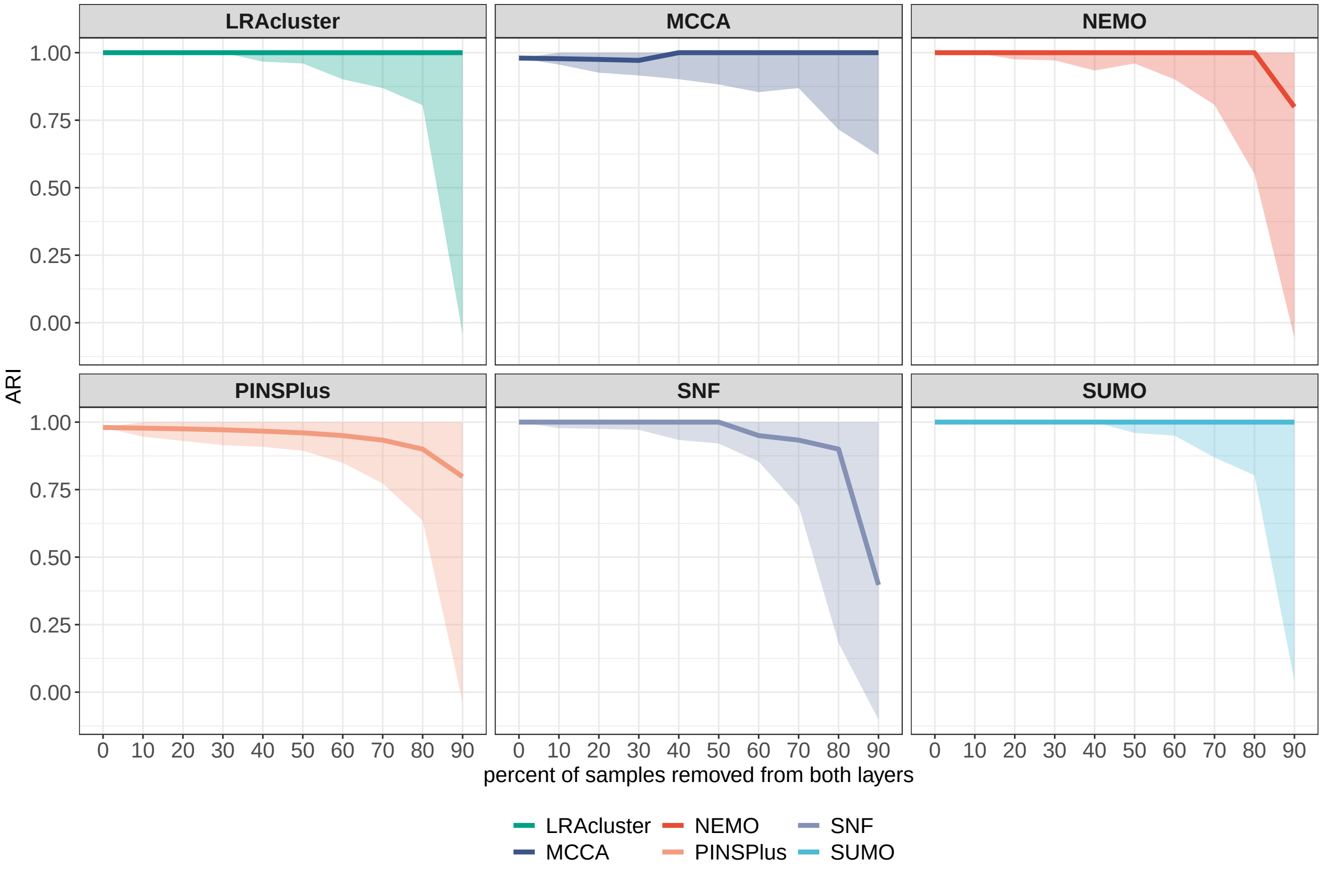

Figure S4: **Effect of sample size on the accuracy of the various methods.** Datasets were created by removing the same random fraction of samples from both data type. We plot the ARI scores for 100 repetitions at each data point.

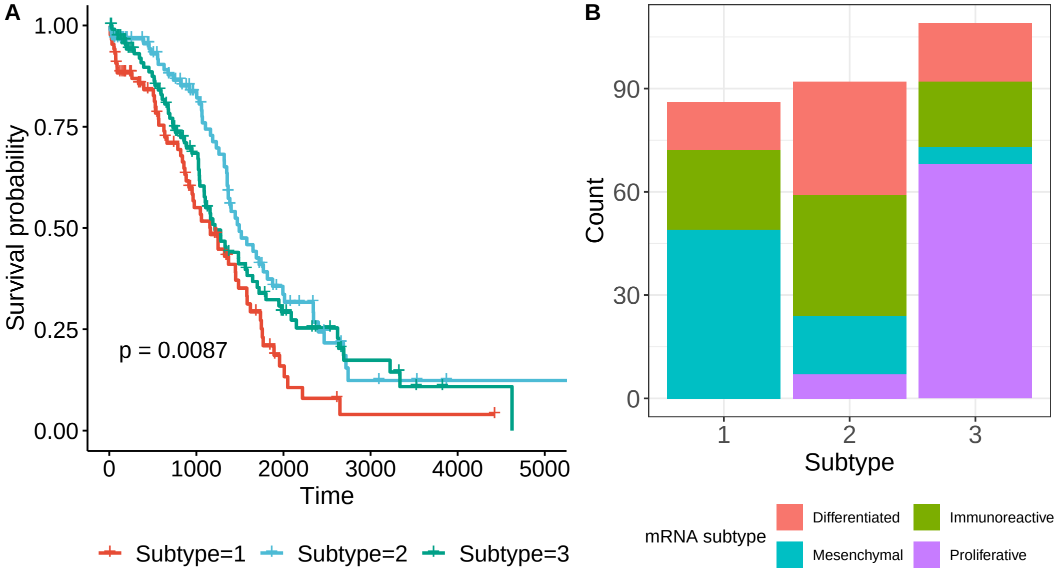

Figure S5: **SUMO identifies a subgroup of patients with significant differential survival in the TCGA-OV dataset.** (A) shows the KM analysis of the clusters identified by SUMO, and (b) shows the distribution of the patients based on the subtypes identified by using mRNA dataset.

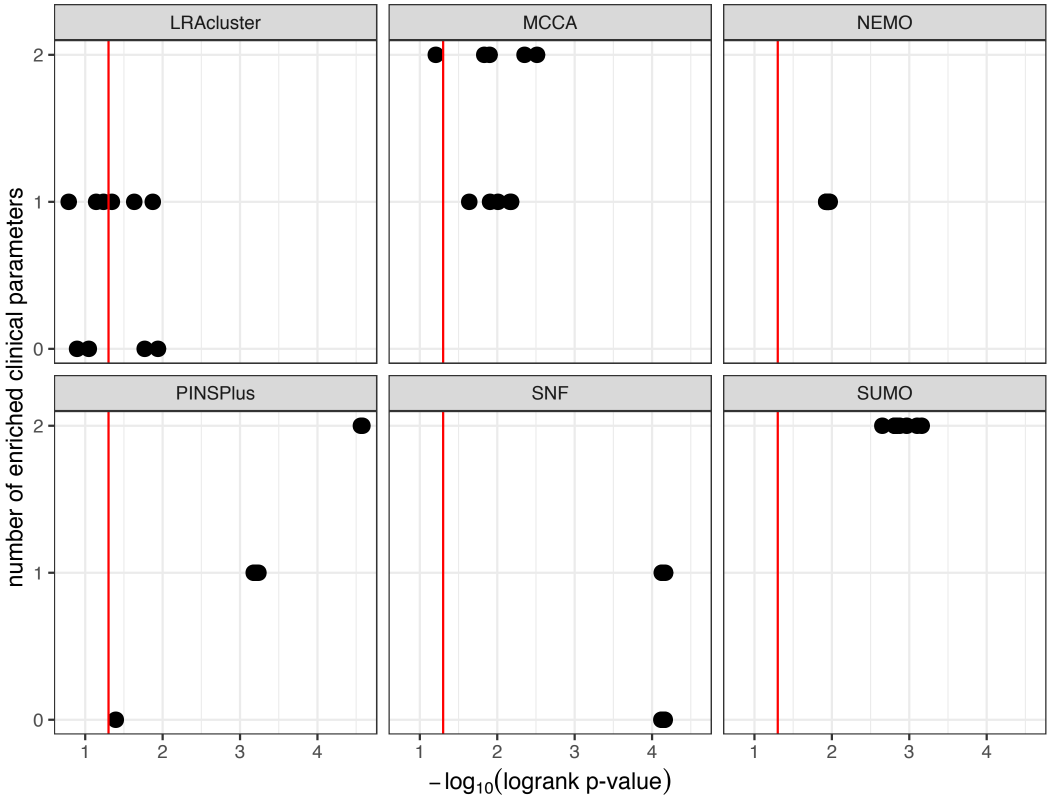

Figure S6: **Results of benchmark evaluation using GBM dataset.** Vertical line indicates p-value equal 0.05. Each method was run 10 times on this dataset using random seeds as input.

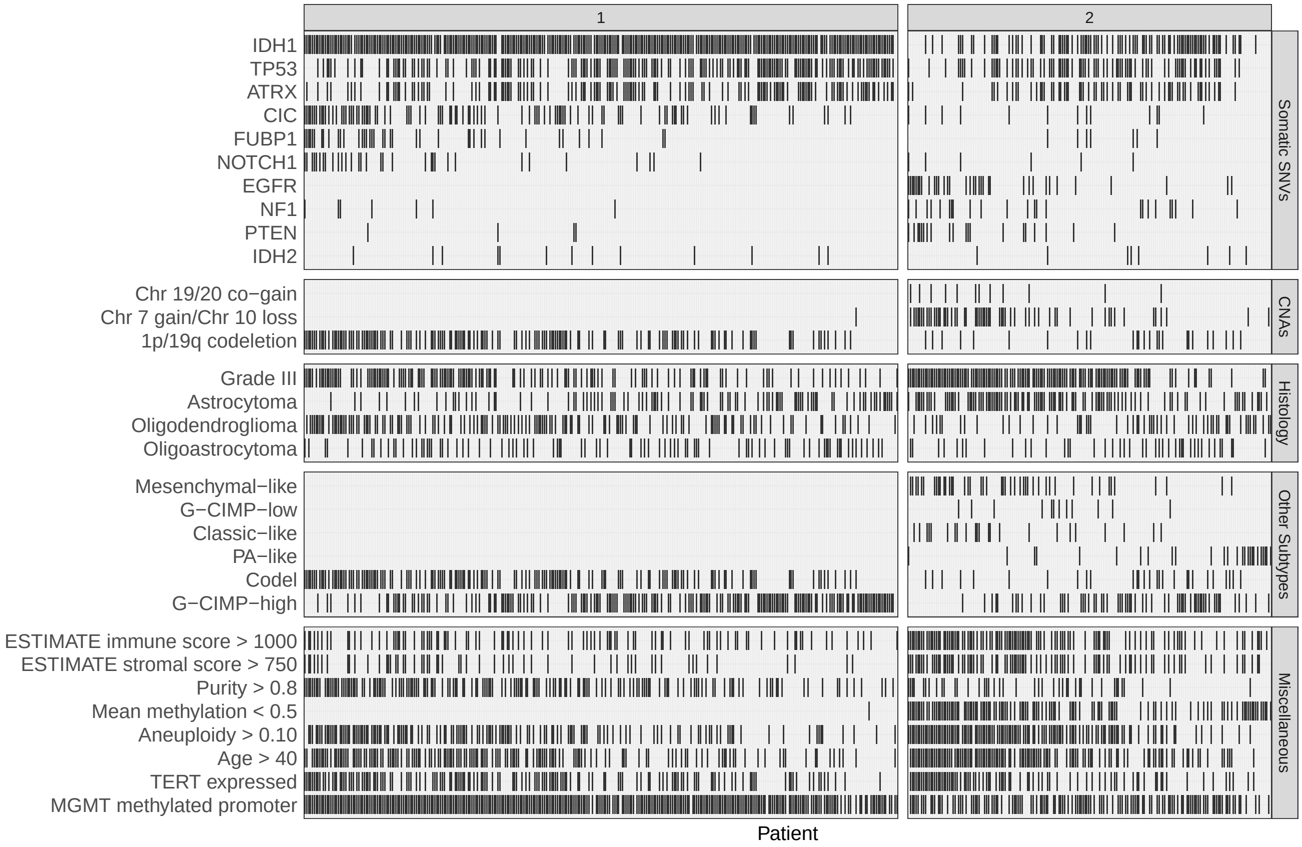

Figure S7: **Association of the 2 subtypes identified by SUMO with mutations, clinical phenotypes, and existing supervised classifications.**

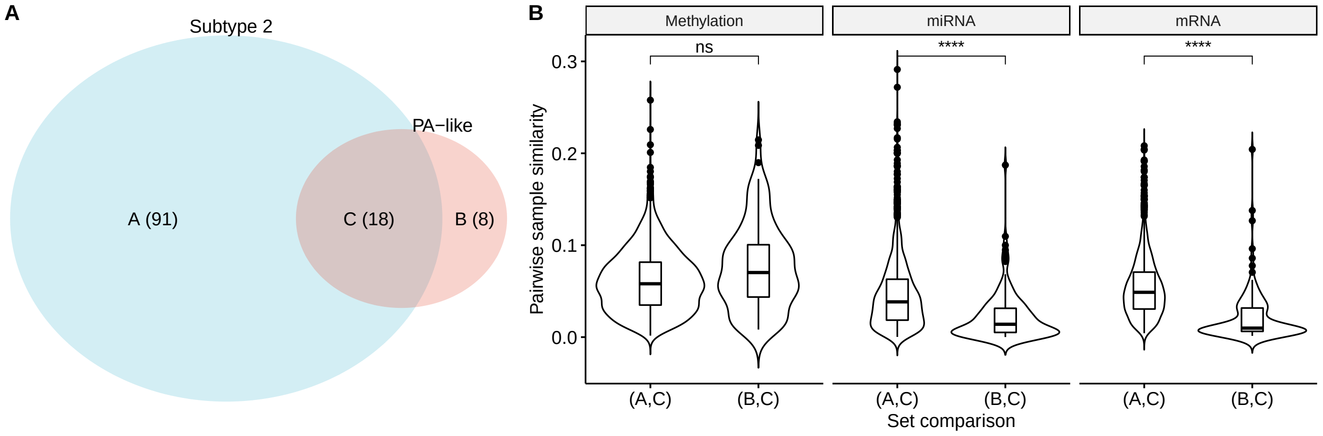

Figure S8: **PA-like samples in subtype 2 compared to other samples in subtype 2 and PA-like samples assigned to other subtypes.** (A) 18 of the 26 PA-like samples are assigned to Subtype 2 by SUMO. (B) The distribution of pairwise similarity between samples, calculated from euclidean distances. Label A,B,C in the x-axis point to sets of samples in (A). PA-like samples assigned to subtype 2 are more similar to other samples assigned to Subtype 2, when data from mRNA and miRNA are included in the analysis.

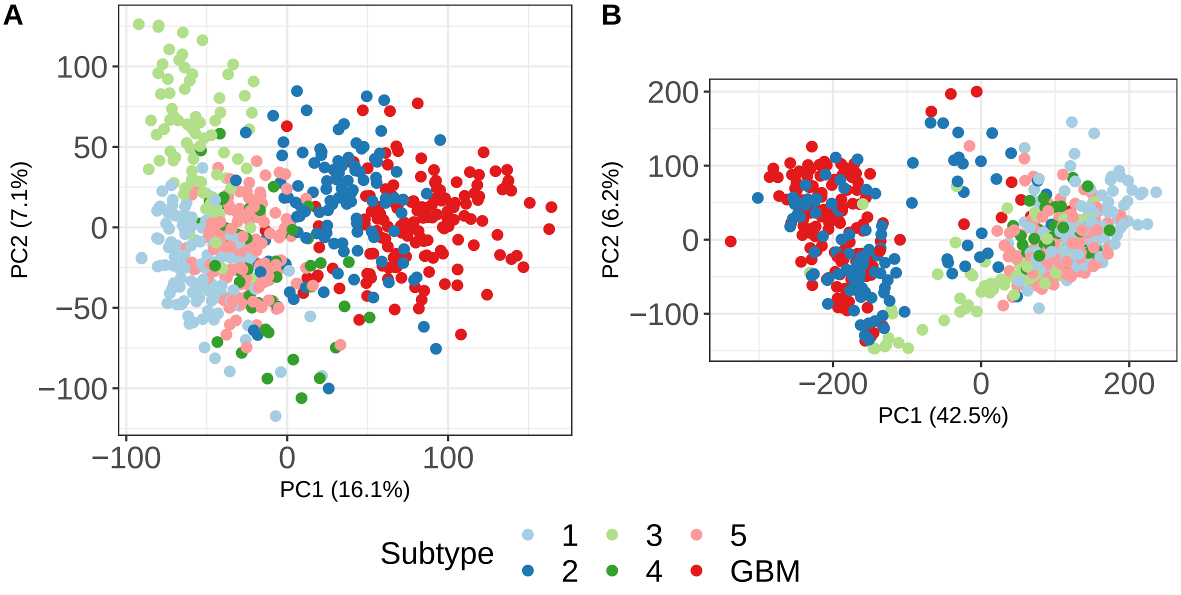

Figure S9: **Principal component analysis** We show the top 2 principal components from (A) expression, (B) methylation of TCGA-GBMLGG which shows the similarities between Subtype 2 and GBMs.

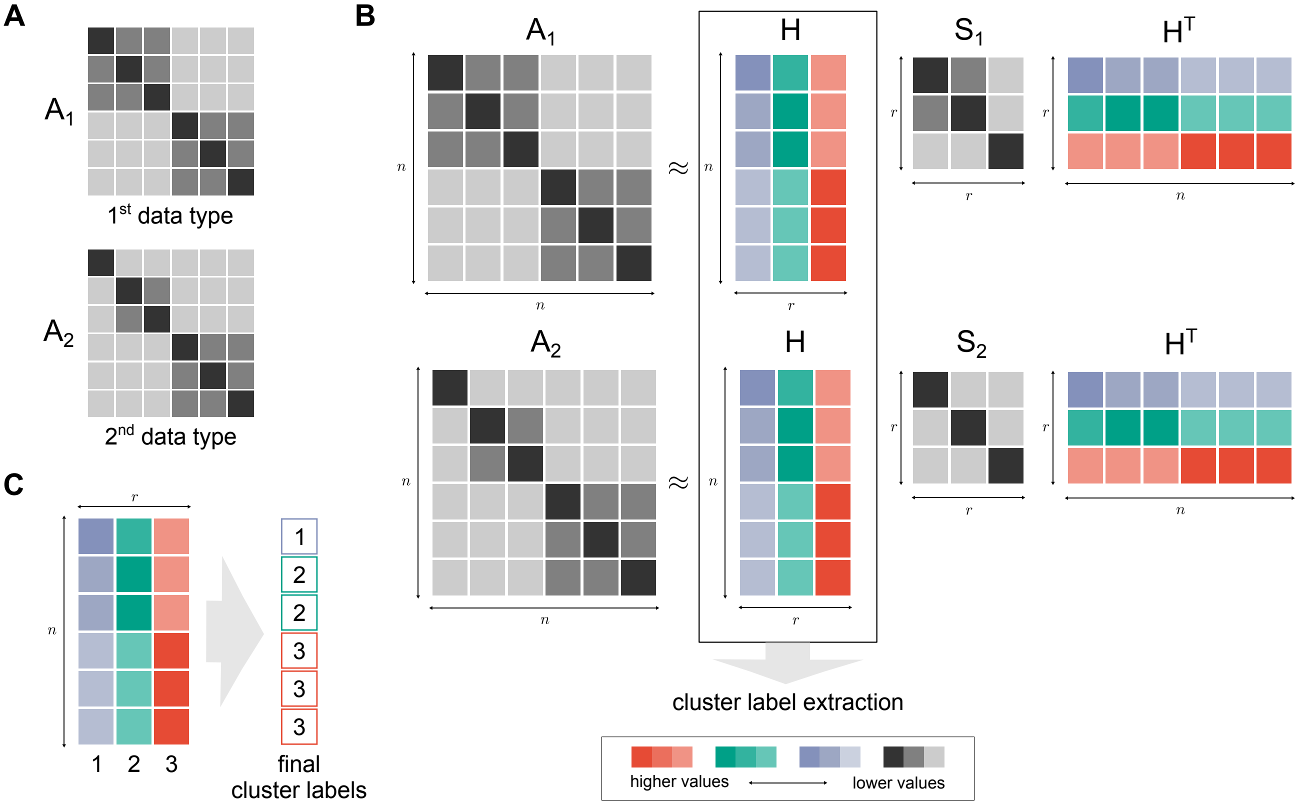

Figure S10: **Illustrative description of factorization.** (A) Two similarity matrices A_1 and A_2 display complementary sample-sample similarity in both data types. (B) Each similarity matrix is tri-factorized in such a way that H matrix is shared across the data types and afterward used for cluster label extraction. Data type specific S_i matrices display relationships between clusters. (C) Final cluster labels for samples are extracted by inspecting columns containing row-wise maximum values of H matrix.

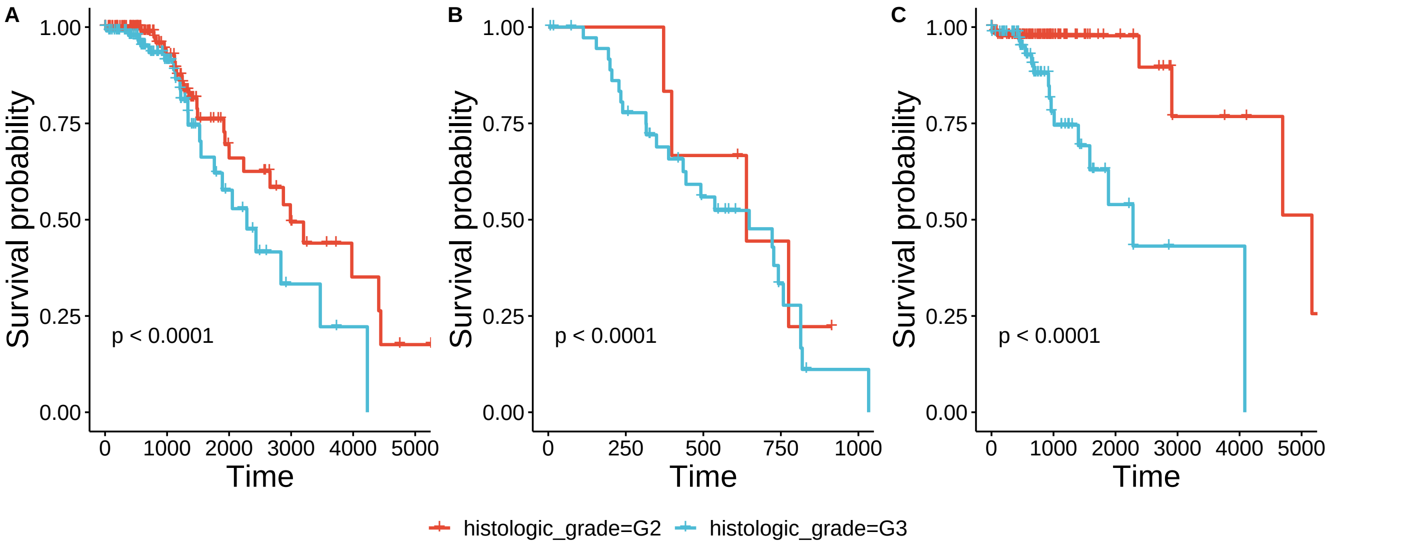

Figure S11: **There is a significant difference in survival within the subtypes based on tumor grade** Here we show the samples assigned to (A) G-CIMP-high, (B) Mesenchymal-like, and (C) Codel subtypes and show that KM analysis finds significant differences in survival based on the assigned tumor grade.

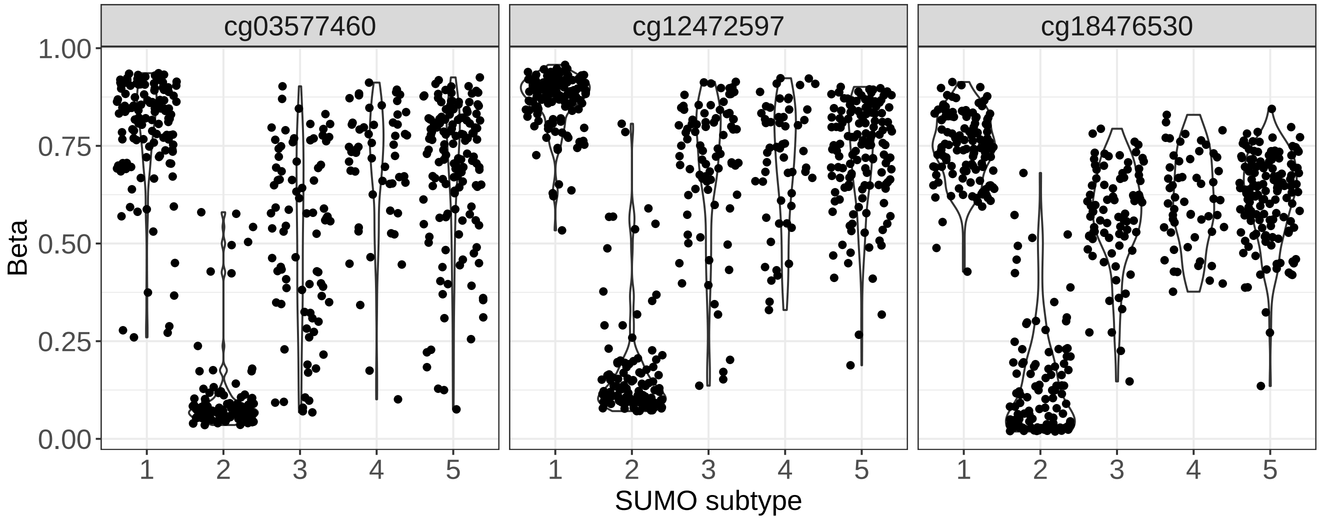

Figure S12: **The top three features with the potential to be biomarkers for Subtype 2.** Here we show the violin plot of the beta values for the probes with the highest predictive values, as identified on running the interpretation module in SUMO.

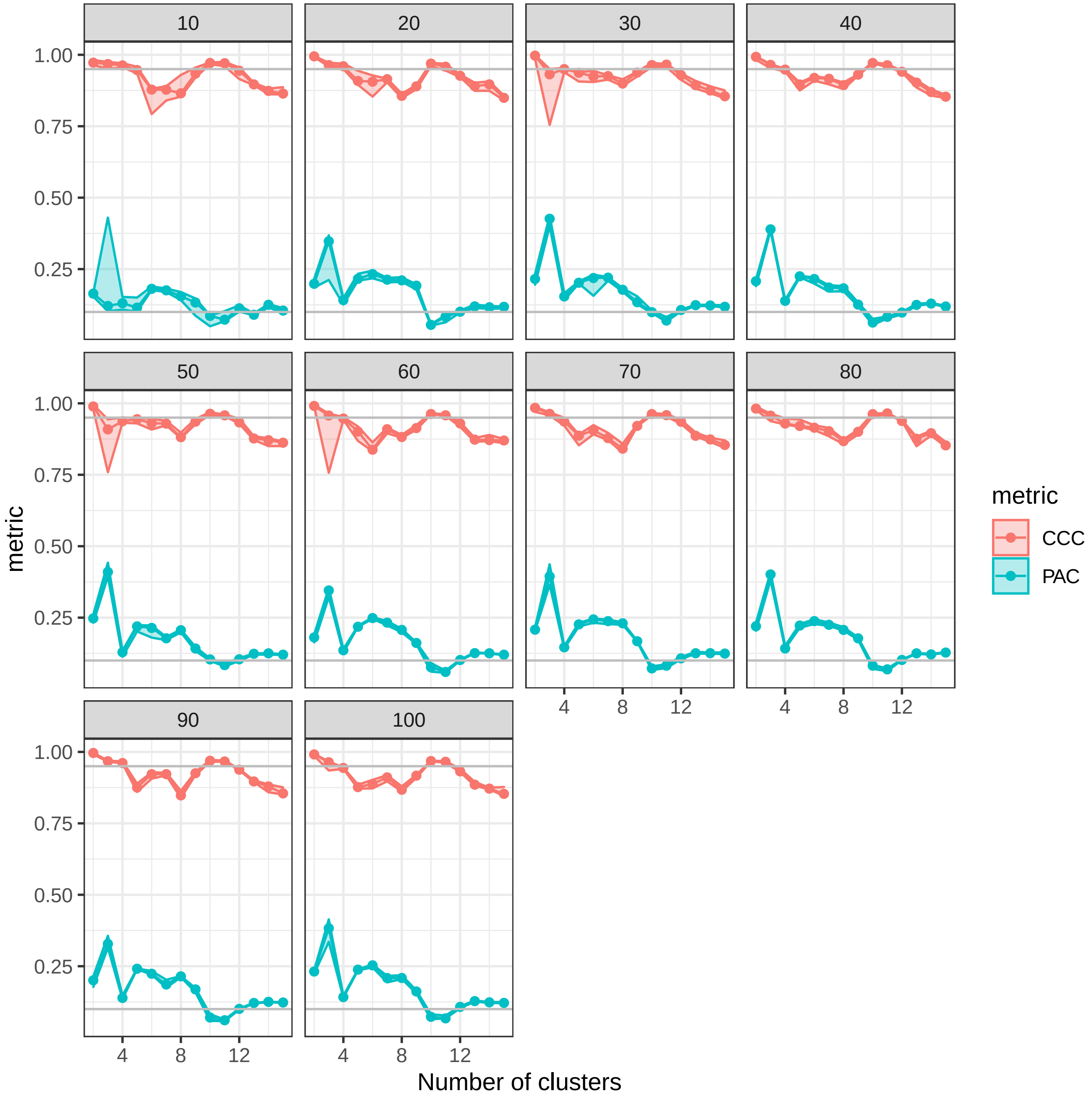

Figure S13: **SUMO can identify one of the stable solutions after 20-30 repetitions.** Here, each facet shows the PAC and CCC curves (the minimum, median and the maximum value of those metrics are shown for each "number of clusters") as the number of repetitions of the solver is increased. SUMO identifies either 10 or 11 as the optimal number of clusters when a small number of repetitions are run. As the number of repetitions increase, both 10 and 11 emerge as equally stable solutions. The horizontal lines correspond to values of 0.1 and 0.95.

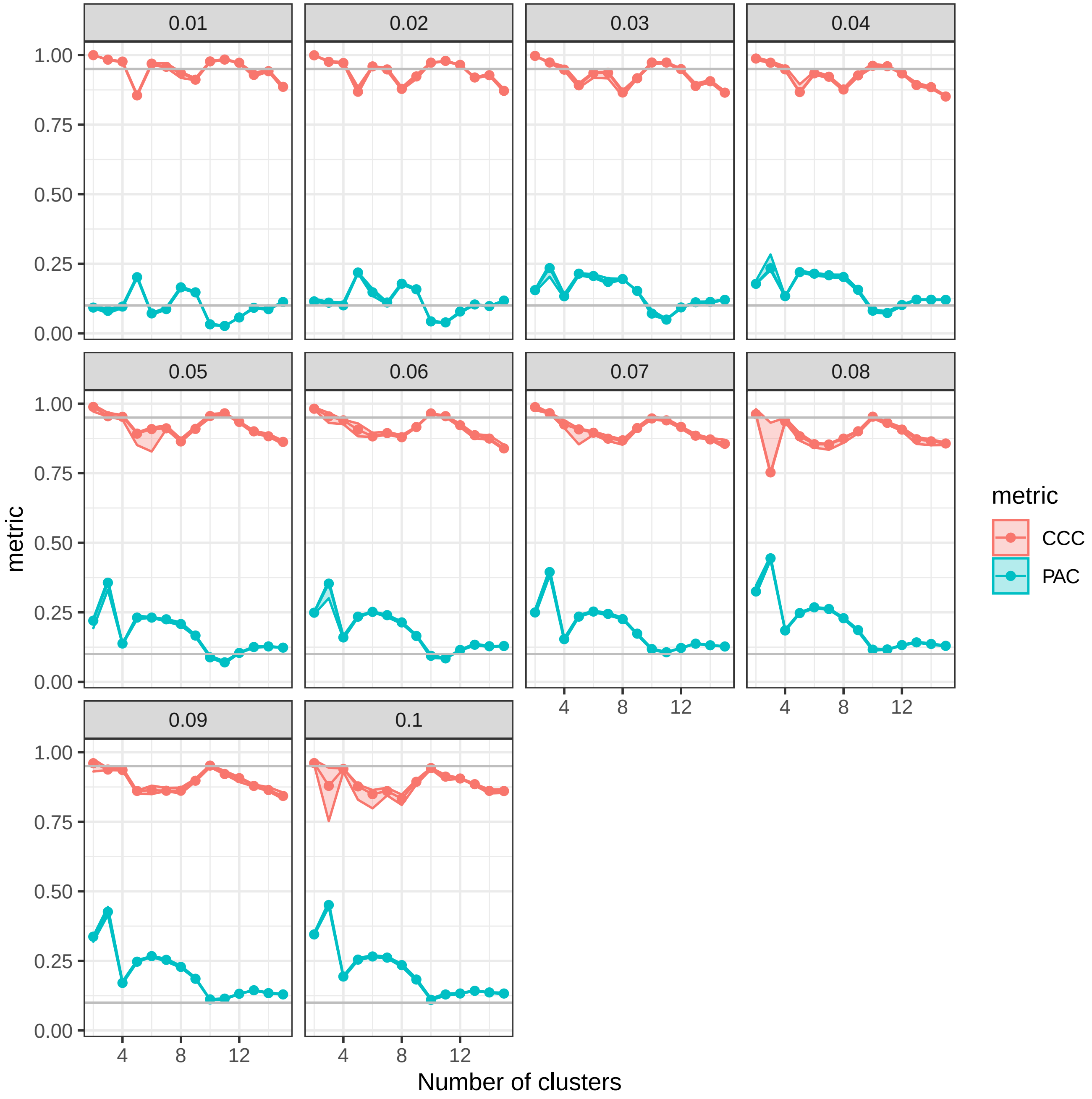

Figure S14: **SUMO is stable for a wide fraction of samples that are removed in a single repetition.** Here, each facet shows the PAC and CCC curves (the minimum, median and the maximum value of those metrics are shown for each "number of clusters") as the fraction of samples that are removed in each of repetitions of the solver is varied. SUMO identifies either 10 or 11 as the optimal number of clusters as the fraction is change from 1% to 10% of the samples. The horizontal lines correspond to values of 0.1 and 0.95.