



# A Novel Method for Cancer Subtyping and Risk Prediction Using Consensus Factor Analysis

Duc Tran<sup>1</sup>, Hung Nguyen<sup>1</sup>, Uyen Le<sup>2</sup>, George Bebis<sup>1</sup>, Hung N. Luu<sup>3,4</sup> and Tin Nguyen<sup>1\*</sup>

<sup>1</sup> Department of Computer Science and Engineering, University of Nevada, Reno, NV, United States, <sup>2</sup> NTT Hi-Tech Institute, Nguyen Tat Thanh University, Ho Chi Minh City, Vietnam, <sup>3</sup> Division of Cancer Control and Population Sciences, Hillman Cancer Center, University of Pittsburgh Medical Center, Pittsburgh, PA, United States, <sup>4</sup> Department of Epidemiology, University of Pittsburgh Graduate School of Public Health, Pittsburgh, PA, United States

## OPEN ACCESS

### Edited by:

Timothy I. Shaw,  
St. Jude Children's Research Hospital,  
United States

### Reviewed by:

Shengli Li,  
University of Texas Health Science  
Center at Houston, United States  
Prashanth N. Suravajhala,  
Birla Institute of Scientific  
Research, India

### \*Correspondence:

Tin Nguyen  
tinn@unr.edu

### Specialty section:

This article was submitted to  
Cancer Genetics,  
a section of the journal  
Frontiers in Oncology

**Received:** 31 January 2020

**Accepted:** 27 May 2020

**Published:** 24 June 2020

### Citation:

Tran D, Nguyen H, Le U, Bebis G,  
Luu HN and Nguyen T (2020) A Novel  
Method for Cancer Subtyping and  
Risk Prediction Using Consensus  
Factor Analysis.  
Front. Oncol. 10:1052.  
doi: 10.3389/fonc.2020.01052

Cancer is an umbrella term that includes a range of disorders, from those that are fast-growing and lethal to indolent lesions with low or delayed potential for progression to death. One critical unmet challenge is that molecular disease subtypes characterized by relevant clinical differences, such as survival, are difficult to differentiate. With the advancement of multi-omics technologies, subtyping methods have shifted toward data integration in order to differentiate among subtypes from a holistic perspective that takes into consideration phenomena at multiple levels. However, these integrative methods are still limited by their statistical assumption and their sensitivity to noise. In addition, they are unable to predict the risk scores of patients using multi-omics data. Here, we present a novel approach named Subtyping via Consensus Factor Analysis (SCFA) that can efficiently remove noisy signals from consistent molecular patterns in order to reliably identify cancer subtypes and accurately predict risk scores of patients. In an extensive analysis of 7,973 samples related to 30 cancers that are available at The Cancer Genome Atlas (TCGA), we demonstrate that SCFA outperforms state-of-the-art approaches in discovering novel subtypes with significantly different survival profiles. We also demonstrate that SCFA is able to predict risk scores that are highly correlated with true patient survival and vital status. More importantly, the accuracy of subtype discovery and risk prediction improves when more data types are integrated into the analysis. The SCFA software and TCGA data packages will be available on Bioconductor.

**Keywords:** multi-omics integration, risk score prediction, cancer subtyping, survival analysis, factor analysis

## 1. INTRODUCTION

After 20 years of cancer screening, the chance of a person being diagnosed with prostate or breast cancer has nearly doubled (1–4). However, this has only marginally reduced the number of patients with advanced disease, suggesting that screening has resulted in the substantial harm of excess detection and over-diagnosis. At the same time, 30–50% of patients with non-small cell lung cancer (NSCLC) develop recurrence and die after curative resection (5), suggesting that a subset of patients would have benefited from more aggressive treatments at early stages. Although not routinely recommended as the initial course of treatment, adjuvant and neoadjuvant chemotherapy have been shown to significantly improve the survival of patients with advanced early-stage disease (6–8). The ability to prognosticate outcomes would allow us to manage these diseases better: patients

whose cancer is likely to advance quickly or recur would receive the necessary treatment. The important challenge is to discover the molecular subtypes of disease and subgroups of patients (9–12).

Cluster analysis has been a basic tool for subtype discovery using gene expression data. These include hierarchical clustering (HC), neural networks (13–17), mixture model (18–20), matrix factorization (21, 22), and graph-theoretical approaches (23–25). Arguably, the state-of-the-art approach in this area is Consensus Clustering (CC) (26, 27), which is a resampling-based methodology of class discovery and cluster validation (28–30). However, these approaches are not able to combine multiple data types. Although analyses on a single data type could reveal some distinct characteristics for different subtypes, it is not sufficient to explain the mechanism that happens across multiple biological levels.

With the advancement of multi-omics technologies, recent subtyping methods have shifted toward multi-omics data integration. The goal is to differentiate among subtypes from a holistic perspective, that can take into consideration phenomena at various levels (e.g., transcriptomics, proteomics, epigenetics). These methods can be grouped into three categories: simultaneous data decomposition methods, joint statistical models, and similarity-based approaches. Methods in the first category (data decomposition) include md-modules (31), intNMF (32), and LRAcluster (33). These methods assume that there exist molecular patterns that are shared across multiple types of data. Therefore, these methods aim at finding a low dimensional representation of the high-dimensional multi-omics data that retains those patterns. For example, both md-modules and intNMF utilize a joint non-negative matrix factorization to simultaneously factorize the data matrices of multiple data types. In their design, the basis vectors are shared across all data types while the coefficient matrices vary from data type to data type. These two methods, md-modules and intNMF, only differ in the way they iteratively estimate the coefficient matrices. Another method is LRAcluster, which applies the low-rank approximation and singular vector decomposition to generate low dimensional representations of the data and then performs k-means clustering to identify the subtypes. These methods strongly rely on the assumption that all molecular signals can be linearly and simultaneously reconstructed.

Methods in the second category (statistical modeling) include BCC (34), MDI (35), iClusterBayes (36), iClusterPlus (37), and iCluster (38, 39). These methods assume that each data type follows a mixture of distributions and then integrate multiple types of data using a joint statistical model. The parameters of the mixture models are estimated by maximizing the likelihood of observed data. These methods strongly depend on the correctness of their statistical assumptions. Also, due to a large number of parameters and iterations involved, the computation complexity of statistical methods is usually extensive. Therefore, these methods often rely on pre-processing and gene filtering to ease the computational burden.

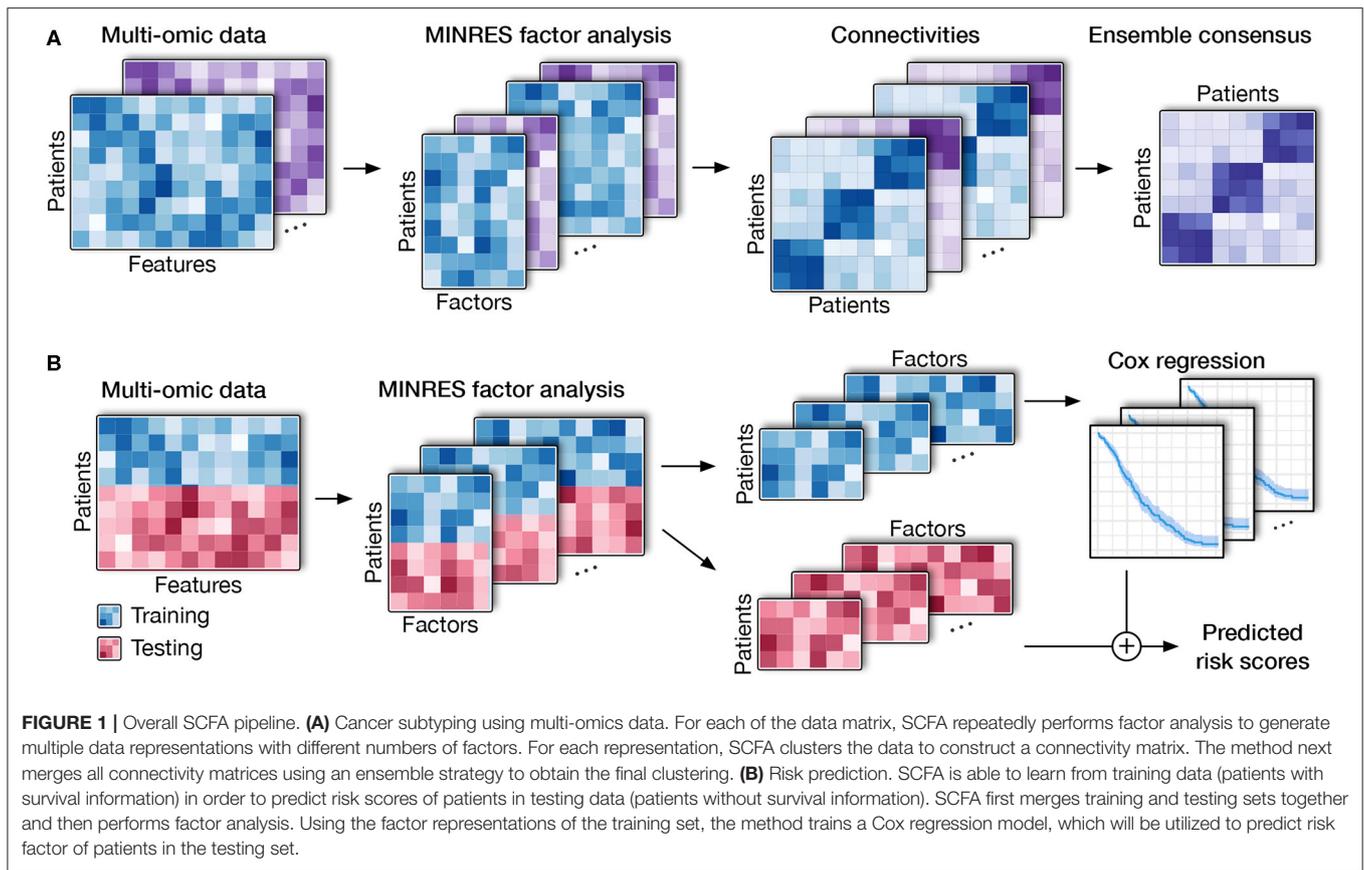
Methods in the third category (similarity-based) typically construct the pair-wise connectivity between patients (that represents how often the patients are grouped together) for

each data type and then integrate multiple data types by fusing the individual connectivity matrices. As these methods perform data integration in the sample space, their computational complexity depends mostly on the number of patients, not the dimensions of features/genes. Therefore, these methods are capable of performing subtyping on a genomic scale. Methods in this category include SNF (40), rMKL-DR (41), NEMO (42), CIMLR (43), and PINS (44, 45). SNF creates a patient-to-patient network by fusing connectivity matrices and then partitions the network using spectral clustering (46). rMKL-DR projects samples into a lower-dimensional subspace and then partitions the patients using k-means. NEMO follows a similar strategy with the difference is that it incorporates only partial data into the integrative analysis. Though powerful, these methods do not account for the noise and unstable nature of quantitative assays. PINS and CIMLR follow two different strategies to address noise and instability. PINS introduces Gaussian noise to the data in order to obtain subtypes that are robust against data perturbation. CIMLR combines multiple gaussian kernels per data type to measure the similarity between each pair of samples. The resulted similarity matrix is then subjected to dimension reduction and k-means to determine the subtypes. Though powerful, the similarity metrics used in these methods (i.e., Gaussian kernel, Euclidean distance) make them susceptible to noise and the “curse of dimensionality” (47) from the high-dimensional multi-omics data.

Here we propose a novel approach, named Subtyping via Consensus Factor Analysis (SCFA), that follows a three-stage hierarchical process to ensure the robustness of the discovered subtypes. First, the method uses an autoencoder to filter out genes with an insignificant contribution in characterizing each patient. Second, it applies a modified factor analysis to generate a collection of factor representations of the high-dimensional multi-omics data. Finally, it utilizes a consensus ensemble to find subtypes that are shared across all factor representations. The software package also includes a model based on Cox regression and Elastic net that is able to predict the risk scores of new patients. In an extensive analysis using 7,973 samples related to 30 different cancer diseases, we demonstrate that our method outperforms other state-of-the-art methods in discovering subtypes with significantly different survival profiles. We also demonstrate that data integration indeed improves the subtyping procedure as subtypes obtained from multi-omics data have more significant Cox *p*-values than subtypes obtained from individual data types. Finally, we demonstrate that the method is able to predict the risk factor of new patients with high accuracy.

## 2. METHODS

The high-level workflow of SCFA is shown in **Figure 1**. The framework consists of two main modules: disease subtyping (**Figure 1A**) and risk assessment (**Figure 1B**). The input of the subtyping module is a list of data matrices (e.g., mRNA, methylation, miRNA) in which rows represent patients while columns represent genes/features. For each matrix, the method



first performs a filtering step using an autoencoder and then repeatedly performs factor analysis (48) to represent the data with different numbers of factors. By representing data with different numbers of factors, we can improve on situations where the projected data do not accurately represent the original data due to noise. Using an ensemble strategy, SCFA combines all of the factor representations to determine the final subtypes.

In the second module, SCFA focuses on predicting the risk scores of patients with unknown survival information. In this module, SCFA combines factor analysis with Cox regression (49, 50) and elastic net (51) to build a prediction model. The method first performs factor analysis on both training (patients with survival information) and testing data (patients without survival information) and then builds a Cox regression model, which can be used to predict the risk scores of patients from the testing data. By default, our software package includes data obtained from The Cancer Genome Atlas (TCGA) that can be used as the training data by default. However, users are free to provide new training data. Using the training data, users can train the model and then predict the risk score of new patients using molecular data.

In the following sections, we will describe in detail the techniques used in the SCFA framework: (i) dimension reduction and factor analysis, (ii) the ensemble strategy for subtyping, and (iii) Cox model and elastic net for risk assessment.

## 2.1. Dimension Reduction and Factor Analysis

Both modules start with dimension reduction and factor analysis. The purpose of dimension reduction is to remove features/genes that play no role in differentiating between patients. This technique was originally introduced in our scDHA method for single-cell analysis (52). Briefly, we utilize a non-negative kernel autoencoder which consists of two components: encoder and decoder. The encoder aims at representing the data in a low dimensional space whereas the decoder tries to reconstruct the original input from the compressed data. By forcing the weights of the network to be non-negative, we capture the positive correlation between the original features and the representative features. Selecting features with high variability in weights would result in a set of features that are informative, non-redundant, and capable of representing the original data.

After the filtering step using the non-negative autoencoder, we perform another dimension reduction step using Factor Analysis (FA) (48). In general, factor analysis aims at minimizing the difference of feature-feature correlation matrix between the latent space and original data. Correlation is a standardized metric, where it takes into account the number of observations and variance of the features during the calculation process. This makes factor analysis robust against scaling and high number of dimensions compared to traditional decomposition such as principle component analysis (PCA), which uses

Euclidean distance as the distance metric. To further improve the performance of factor analysis, we adjust the objective of FA to maintain the patient-patient correlation.

Starting with the original correlation matrix, FA finds  $k$  (number of factors) largest principle components and tries to reproduce the original matrix using those principal components (model matrix). FA iteratively fits the model matrix to the original matrix using optimization algorithms. In our model, we employ the Minimum Residual (MINRES) optimization because it copes better with the small and medium sample size of the input data (53). Also, instead of preserving the relationship between variables, we aim to maintain the overall patient-patient relationships by preserving their Pearson correlations in the representations. By changing the objective, the computational power required is significantly lower as the number of patients (in the scale of hundreds) is much lower than the number of features (in the scale of tens of thousands). Moreover, maintaining the distance between patients in the low dimensional representation would be more beneficial for our desired applications. To avoid overfitting, we repeatedly perform factor analysis with different numbers of factors, resulting in multiple representations of each input matrix. In the clustering module (**Figure 1A**), all factor representations of all data types (data matrices) are combined using an ensemble strategy to determine the subtypes. In the risk prediction module (**Figure 1B**), the factor representations of the training data are combined to build the prediction model.

## 2.2. Subtyping Using Consensus Ensemble

Given a collection of factor representations from all data types, we aim at finding patient subgroups that are consistently observed together in all representations (**Figure 1A**). For each representation, we first determine the optimal number of clusters using two indices: (i) the ratio of *between sum of squares* over the *total sum of squares*, and (ii) the increase of *within sum of squares* when the number of cluster increases (52). After the optimal number of clusters is determined, we use k-means to cluster the underlying factor representation to build a connectivity matrix. To avoid the convergence to a local minimum, we perform k-means clustering using multiple starting points and choose the results with the smallest sum of square error. This process is repeated for all of the representations to obtain a collection of connectivity matrices for all data types.

Finally, we use the Weighted-based meta-clustering algorithm (54) to combine all clustering results from each data representation to determine the final subtyping. In short, the meta-clustering first calculates the weight for each pair of patients regarding their chance to be grouped together. Next, it assigns a weight for each patient by accumulating the weights of all pairs containing this patient. It then computes the weighted cluster-to-cluster similarity from all connectivity matrices. Finally, it partitions the cluster-to-cluster similarity matrix using hierarchical clustering to determine the final subtypes.

## 2.3. Risk Score Prediction

The goal of this module is to calculate the risk score of new patients using their molecular data. This is a supervised learning

method that learns from a training set in order to predict the risk scores each patient in the testing set. More specifically, the training set consists of a set of patients with molecular data (e.g., mRNA, methylation, miRNA) and known survival information while the testing set consists of patients with only molecular data. By default, we provide TCGA datasets in our package as training data, but users are free to provide training data if necessary. Using the training data, this module will train the Cox regression model that can be used to predict the risk scores of new patients. Below is the description of the method for one data type and for multi-omics data.

Given a single data type as input, we merge the testing data with training data and then perform dimension reduction and factor analysis to generate multiple representations of this data. For each representation, we use the training data to train the Cox regression model. This model aims at estimating a coefficient  $\beta_i$  for each corresponding predictor  $x_i$  of the input data. After the model is trained, the risk scores for new patients can be calculated as  $\exp(\sum_{i=1}^n \beta_i x_i)$ , where  $n$  is the number of features in the factor representation. In the Cox model, the risk score is defined as  $\frac{h(t)}{h_0(t)}$ , where  $h(t)$  is the expected hazard at time  $t$ , and  $h_0(t)$  is the baseline hazard when all the predictors are equal zero. Patients with a higher risk score are likely to suffer the event of interest (e.g., vital status or disease recurrence) earlier than the one with a lower risk score. Here we use elastic net (51) implemented in the R-package “glmnet” (55) to fit the model to better cope with the dynamic number of predictors. Elastic net linearly combines Lasso and Ridge penalty during the training process to select only the most relevant predictors that have important effects on the response (the risk scores in this case). We use five-fold cross-validation to select the parameters for the model. The final risk score for each patient is the geometric average of the risk scores resulted from all representations.

In the case of multi-omics data, we repeat the same process (described above) for each data type. We perform factor analysis to produce multiple representations, resulting in a collection of representations from all data types. For a new patient, each representation will produce an estimated risk score. The final risk score for the patient is calculated as the geometric average of all predictions from all representations.

## 3. RESULT

Here we assess the performance of SCFA using data obtained from 7,973 patients related to 30 different cancer diseases downloaded from The Cancer Genome Atlas (TCGA). For each of the 30 cancer datasets, we downloaded mRNA, miRNA, and methylation data. We also downloaded the clinical data for these patients, which includes vital status and survival information. Using clinical information, we assess the ability of SCFA in both unsupervised subtyping and supervised risk prediction.

### 3.1. Subtyping on 30 TCGA Datasets

Here we compare the performance of SCFA with four state-of-the-art methods: Consensus Clustering (CC) (26, 27), Similarity Network Fusion (SNF) (40), Cancer Integration via Multikernel

**TABLE 1** | Cox *p*-values of subtypes identified by SCFA, CC, SNF, iClusterBayes (iCB), and CIMLR for 30 TCGA datasets.

	SCFA	CC	SNF	iCB	CIMLR
ACC	3.4e-03	5.4e-04	4.3e-05	9.2e-04	3.4e-01
BLCA	7.2e-03	1.1e-01	1.1e-01	5.1e-01	4.7e-01
BRCA	3.2e-04	2.9e-02	1.2e-01	2.7e-02	4.9e-03
CESC	9.4e-03	5.8e-02	5.1e-01	2e-02	1.9e-01
DLBC	4.3e-06	5.1e-01	7.5e-01	2.9e-01	7.4e-01
ESCA	7.3e-05	7.7e-01	3.9e-01	7.9e-01	5.6e-01
GBM	2.3e-03	3.2e-01	2.1e-02	1.1e-01	8.1e-02
GBMLGG	5.8e-14	1.6e-04	4.8e-14	8e-02	6.4e-10
HNSC	4e-02	5e-01	3.7e-01	3.7e-01	4e-01
KICH	2.3e-13	8.7e-01	7e-01	6.9e-01	4.6e-01
KIPAN	1.4e-19	9.3e-08	2.1e-07	1.6e-09	9.8e-05
KIRP	1.7e-03	4.5e-01	5.3e-03	3e-03	1.9e-02
LAML	5.8e-04	3.9e-02	1.7e-03	9e-01	1.4e-04
LGG	6.5e-15	6.6e-07	1.6e-14	1.1e-01	8.3e-15
MESO	1.6e-04	3.1e-01	4.2e-04	3.7e-02	1.1e-02
PAAD	6.9e-04	1.1e-02	7.4e-04	2.3e-03	2e-03
SARC	3.3e-03	2.4e-01	4.4e-02	4.3e-02	5.6e-02
SKCM	1.6e-03	6.3e-01	4.8e-01	8.4e-03	7.4e-05
STES	3.9e-02	2e-01	1.6e-01	4.1e-03	3.4e-02
THCA	7.8e-03	7.9e-01	6.2e-01	7.8e-01	8.6e-03
THYM	8.1e-04	1.5e-01	9.7e-02	9e-03	1.2e-01
UCEC	6.5e-03	8.9e-02	1.8e-02	5.9e-02	4.6e-02
UCS	3.4e-02	1.6e-01	8.6e-01	9.6e-01	3.6e-01
UVM	1.3e-06	6.1e-04	1.7e-04	6.6e-02	5.8e-04
CHOL	3.1e-01	7.9e-02	5.7e-01	9.1e-01	3.4e-01
COAD	4.7e-01	5.8e-01	1.3e-01	2.2e-01	5.6e-01
KIRC	1e-01	8.3e-01	6.9e-01	8.3e-01	9.1e-02
LIHC	3.8e-01	8.8e-01	3.3e-01	9.3e-02	1.9e-01
OV	4.2e-01	6.1e-01	4.4e-01	4.6e-01	5.4e-01
TGCT	3.9e-01	7.4e-01	8.4e-01	7.1e-01	8.4e-01
#Significant	24	8	12	11	13

The cells highlighted in yellow have Cox *p*-values smaller than 5%. In each row, cells highlighted in green have the most significant *p*-value. SCFA outperforms other methods by having significant *p*-values in most datasets (24 out of 30 datasets).

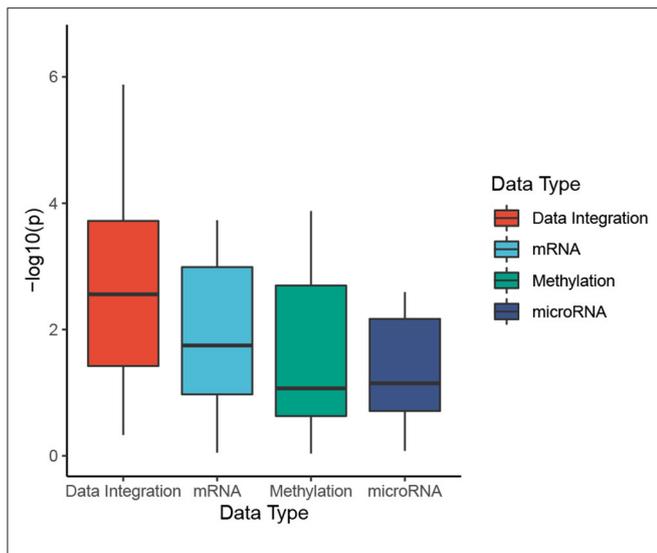
LeaRning (CIMLR) (43), and iClusterBayes (iCB) (36). CC is a resampling-based approach, while SNF and CIMLR are graph-theoretical approaches. The fourth method, iClusterBayes is a model-based approach and is the enhanced version iClusterPlus. These methods were selected to represent three distinctively different subtyping strategies. Among these methods, CC is the only method that cannot integrate multiple data types. For CC, we concatenate the three data types for the integrative analysis. We demonstrate that SCFA outperforms these methods in identifying subtypes with significantly different survival profiles.

Note that here we focus on unsupervised learning, in which each dataset is partitioned independently without using any external information. For example, when analyzing the glioblastoma multiforme (GBM) dataset, we use only the molecular data (mRNA, miRNA, and methylation) of this dataset

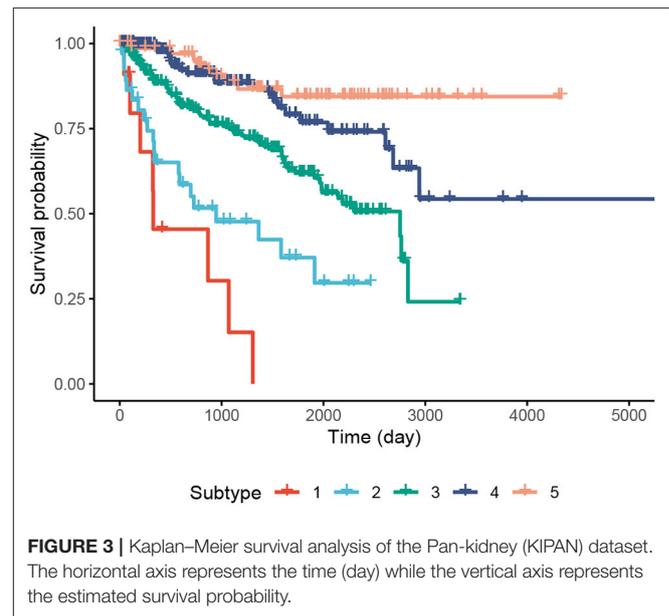
to determine the subtypes. For each cancer dataset, we first use each of the five methods (SCFA, CC, SNF, CIMLR, and iClusterBayes) to integrate the molecular data (mRNA, miRNA, and methylation) in order to determine patient subgroups. For each method, we calculate the Cox *p*-value that measures the statistical significance in survival differences between the discovered subtypes. The Cox *p*-values of subtypes discovered by the five methods for the 30 datasets are shown in **Table 1**. Among the 30 datasets, there are 6 datasets (CHOL, COAD, KIRC, LIHC, OV, and TGCT) for which no method is able to identify subtypes with significant survival differences. In the remaining 24 datasets, SCFA is able to obtain significant Cox *p*-values in all of them while CC, SNF, iClusterBayes, and CIMLR have significant *p*-values in only 8, 12, 11, and 13 datasets, respectively. Also, SCFA has the most significant *p*-values in 19 out of 24 datasets. Regarding time complexity, SCFA, CC, SNF, and CIMLR are able to analyze each dataset in minutes, whereas iClusterBayes can take up to hours to analyze a dataset.

To better understand the usefulness of data integration, we also calculated the Cox *p*-values obtained from individual data types and compared them to Cox *p*-values obtained from data integration (when mRNA, miRNA, and methylation are analyzed together). For each dataset, we perform subtyping using SCFA for each data type and report the Cox *p*-value of the discovered subtypes. The distributions of Cox *p*-values for data integration and for individual data types using SCFA are shown in **Figure 2**. Among 30 cancer datasets, the Cox *p*-values obtained from data integration has the median  $-\log_{10}(p)$  of 2.6, compared to 1.7, 1.1, and 1.1 from gene expression, methylation and miRNA data. Interestingly, subtypes discovered using gene expression data have significantly different survival in 18 over 30 datasets, compared to 10 and 14 of methylation and miRNA data, respectively. The figure also shows that the Cox *p*-values obtained from gene expression data are more significant than those obtained from methylation and miRNA data ( $p = 0.046$  using one-sided Wilcoxon test). However, we note that miRNA and methylation also provide valuable information in data integration, when all data types are analyzed together. As shown in **Figure 2**, the Cox *p*-values obtained from data integration are more significant than those of any individual data type (including mRNA) with a one-sided Wilcoxon test *p*-value of 0.004. This means that each of the three data types provides meaningful contributions to the data integration. To understand how other methods perform with respect to each data type, we also plot the distributions of Cox *p*-values obtained from each data type using CC, SNF, iClusterBayes, and CIMLR (**Figure S1**). CC is the only method that produces comparable Cox *p*-values across the three data types. SNF and CIMLR perform better using miRNA, while iClusterBayes favors mRNA and miRNA data.

There are four important clinical variables that are available in more than 10 TCGA datasets: age (21 datasets), gender (25 datasets), cancer stages (24 datasets), and tumor grades (12 datasets). To understand the association between these variables and the discovered subtypes, we perform the following analyses: (1) Fisher's exact test to assess the association between gender (male and female) and the discovered subtypes; (2) ANOVA test to assess the age difference between the discovered subtypes;



**FIGURE 2** | Cox  $p$ -values of subtypes identified by SCFA. To better understand the usefulness of data integration, we calculate the Cox  $p$ -values obtained from individual data types and compared them to Cox  $p$ -values obtained from data integration (when mRNA, miRNA, and methylation are analyzed together). The horizontal axis shows the data types while the vertical axis shows the minus  $\log_{10}$   $p$ -values. Overall the Cox  $p$ -values obtained from data integration are significantly smaller than those obtained from individual data types ( $p = 0.004$  using one-sided Wilcoxon test).



**FIGURE 3** | Kaplan–Meier survival analysis of the Pan-kidney (KIPAN) dataset. The horizontal axis represents the time (day) while the vertical axis represents the estimated survival probability.

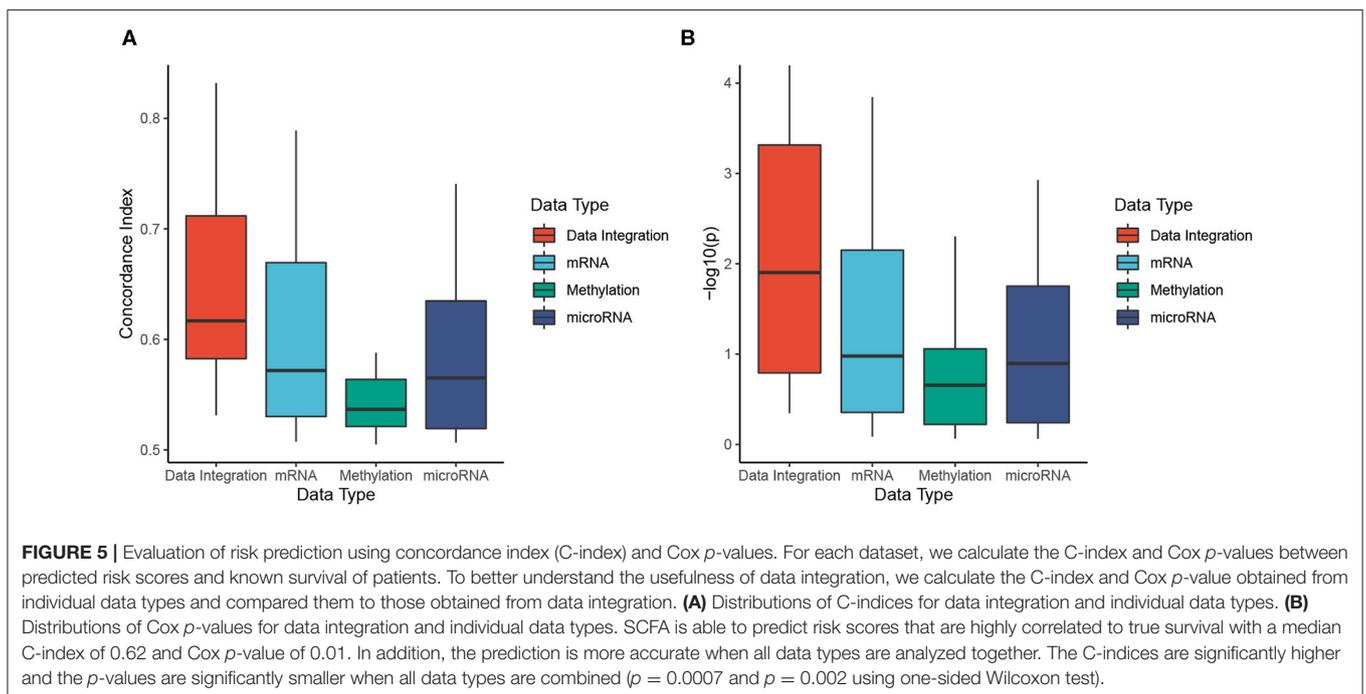
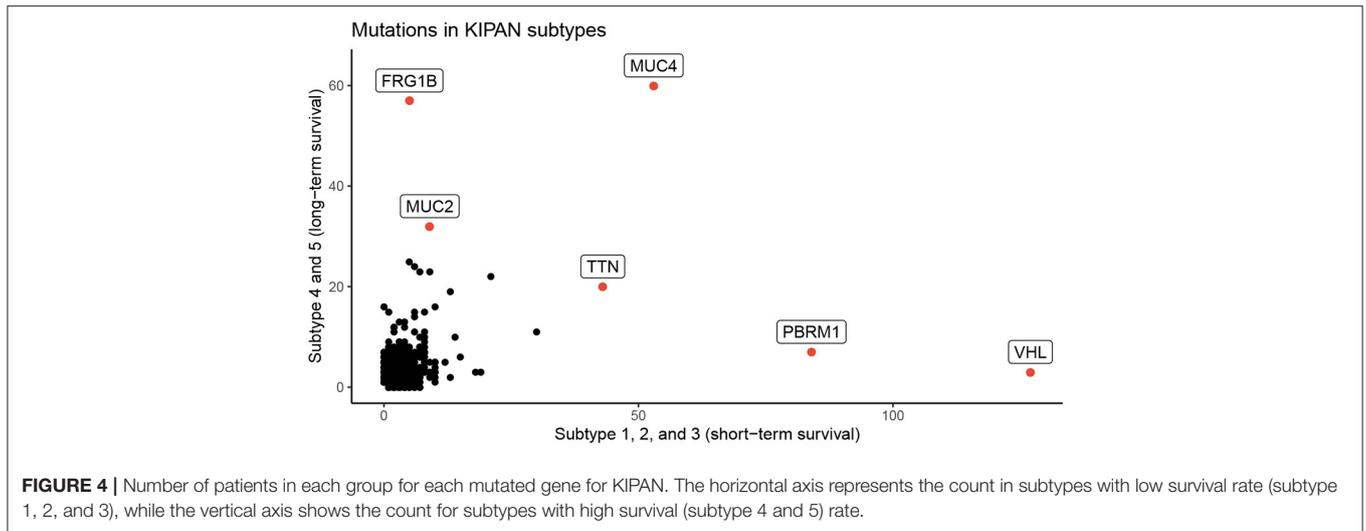
and finally (3) calculate the agreement between the discovered subtypes and known cancer stages/tumor grades using Adjusted Rand Index (ARI) and Normalized Mutual Information (NMI). **Figure S2** and **Tables S1, S2** show the  $p$ -values obtained for gender and age. Overall, the four methods, SCFA, CC, SNF, and CIMLR, are not biased toward gender with only some significant  $p$ -values (**Table S1**). In contrast, iClusterBayes is subject to gender bias with significant  $p$ -values in 12 out of 25 datasets (**Table S1**). The  $p$ -values of iClusterBayes are significantly smaller than those of other methods ( $p = 0.0007$  using one-sided Wilcoxon test). Regarding age, all methods have comparable  $p$ -values (**Table S2**). **Figure S3** and **Table S3** show the ARI values that represent the agreement between the discovered subtypes and known cancer stages and tumor grades. The median ARI of SCFA and SNF are comparable and they are higher than those of CC, iClusterBayes, and CIMLR. Regarding tumor grade, the ARI values of SCFA are higher than the rest. **Figure S4** and **Table S4** show the NMI values. SCFA has higher NMI values in both comparisons. However, the low NMI and ARI values show that there is a low agreement between the discovered subtypes and known stages/grades.

We perform an in-depth analysis for the Pan-kidney (KIPAN) dataset. For this dataset, SCFA discovers five subtypes, each with a very different survival probability (**Figure 3**). Subtype 1 has the lowest survival probability while Subtype 5 has the highest survival probability. All patients of Subtype 1 die within 3 years whereas 85% of patients in Subtype 5 survive at the end of the study (after 15 years). We also perform variant analysis to look for mutations that are highly abundant in the short-term survival

groups (Subtypes 1, 2, and 3) but not in the long-term survival groups (Subtypes 4 and 5), and vice versa. In **Figure 4**, each point represents a gene and its coordinates represent the number of patients having at least a variant in that gene in each group. In principle, we would look for mutated genes in the top left and the bottom right corners. From this figure, we can identify four notable markers: VHL, PBRM1, MUC4, and FRG1B. Among these, MUC4 has been reported to be associated with exophytic growth of clear cell renal cell carcinoma (56) while VHL linked to a primary oncogenic driver in kidney cancers (57). PBRM1 is also a major clear cell renal cell carcinoma (ccRCC) gene (58). See **Supplementary Section 2** and **Figures S5–S8** for details.

### 3.2. Risk Score Prediction Using Multi-Omics Data

We also use the same set of data to demonstrate the ability of SCFA in predicting risk score of each patient. For each of the TCGA datasets, we randomly split the data into two equal sets of patients: a training set and a testing set. We use the training set to train the model and then predict the risk for patients in the testing set. The predicted risk scores are then compared with the true vital status and survival information using Cox  $p$ -value and concordance index (C-index) (59). Concordance index represents the probability that, for a pair of randomly chosen patients, the patient with higher predicted risk will experience death event before the other patient. On the other hand, Cox  $p$ -value measures how significant the difference in survival when correlating with predicted risk scores. This process is repeated 20 times for each dataset, and the average C-index and  $-\log_{10}(p)$  for each dataset are calculated using results from these 20 runs. We note that some datasets do not have enough patients with either event (survive or death), which leads to errors for Cox regression. For that reason, we removed five datasets (DLBC, KIRP, TGCT,



THYM, UCEC) from the analysis, and report survival prediction for only 25 datasets without errors.

**Figure 5** shows the distributions of C-indices and Cox  $p$ -values (in minus  $\log_{10}$  scale), while **Table 2** shows the exact values calculated for each dataset. We calculate the C-index and Cox  $p$ -value obtained from individual data types and compared them to those obtained from data integration (when mRNA, miRNA, and methylation are analyzed together). As shown in **Figure 5A**, the accuracy of the prediction using data integration is generally higher than the accuracy obtained from individual data types. Predictions using data integration have a median C-index of 0.62, compared to 0.57, 0.54, and 0.57 when using mRNA, methylation, and miRNA, respectively. Similar results are

also observed in the evaluation using Cox  $p$ -values (**Figure 5B**). The Cox  $p$ -values obtained from data integration has the median  $-\log_{10}(p)$  of 1.9, compared to 1.0, 0.7, and 0.9 for mRNA, methylation, and miRNA. The results demonstrate that we can potentially predict the risk score of each patient using only molecular data. More importantly, the prediction using multi-omics data is generally more accurate than using individual data types.

## 4. CONCLUSION

In this article, we presented a novel method (SCFA) for disease subtyping and risk assessment using multi-omics data. The

**TABLE 2** | Risk score prediction evaluated by concordance index (C-index) and Cox *p*-values.

Dataset	C-index				-log <sub>10</sub> ( <i>p</i> )			
	Integration	mRNA	Methylation	microRNA	Integration	mRNA	Methylation	microRNA
ACC	0.78	0.79	0.59	0.72	3.32	3.84	0.66	2.73
BLCA	0.59	0.55	0.55	0.54	2.44	1.1	0.9	0.73
BRCA	0.62	0.55	0.52	0.51	1.38	0.77	0.28	0.14
CESC	0.68	0.63	0.54	0.64	3.42	2.15	1.4	2.02
CHOL	0.56	0.56	0.51	0.55	0.38	0.36	0.2	0.24
COAD	0.56	0.52	0.51	0.57	0.52	0.09	0.09	0.48
ESCA	0.53	0.52	0.5	0.51	0.35	0.09	0.18	0.06
GBM	0.55	0.51	0.53	0.53	2.44	0.3	1.04	1.12
GBMLGG	0.77	0.79	0.72	0.73	14.1	11.56	4.83	5.14
HNSC	0.59	0.59	0.51	0.55	1.41	1.81	0.22	0.48
KICH	0.68	0.6	0.63	0.57	1.35	0.62	2.3	1.31
KIPAN	0.79	0.77	0.73	0.74	24.42	14.53	11.65	20.54
KIRC	0.58	0.59	0.54	0.6	0.79	1.24	0.5	0.94
LAML	0.63	0.61	0.56	0.59	2.45	1.94	1.06	1.16
LGG	0.77	0.78	0.73	0.73	14.02	11.44	5.21	7.53
LIHC	0.62	0.53	0.55	0.57	1.9	0.36	0.86	0.9
MESO	0.72	0.69	0.53	0.63	4.46	3.72	0.22	2.93
OV	0.54	0.51	0.53	0.51	0.41	0.12	0.72	0.14
PAAD	0.71	0.67	0.56	0.59	3.35	2.58	0.79	1.75
SARC	0.62	0.57	0.53	0.53	1.19	0.98	0.19	0.26
SKCM	0.61	0.53	0.53	0.52	2.32	0.55	0.32	0.24
STES	0.54	0.51	0.52	0.51	0.4	0.11	0.29	0.16
THCA	0.66	0.53	0.54	0.51	1.26	0.44	0.33	0.57
UCS	0.58	0.53	0.51	0.51	0.68	0.15	0.06	0.08
UVM	0.83	0.67	0.69	0.72	2.62	1.14	2.87	1.33

contribution of SCFA is two-fold. First, it utilizes a robust dimension reduction procedure using autoencoder and factor analysis to retain only essential signals. Second, it allows researchers to predict risk scores of patients using multi-omics data—the attribute that is missing in current state-of-the-art subtyping methods.

To evaluate the developed method, we examined data obtained from 7,973 patients related to 30 cancer diseases downloaded from The Cancer Genome Atlas (TCGA). SCFA was compared against four state-of-the-art subtyping methods, CC, SNF, iClusterBayes, and CIMLR. We demonstrate that SCFA outperforms existing approaches in discovering novel subtypes with significantly different survival profiles. We also demonstrate that the method is capable of exploiting complementary signals available in different types of data in order to improve the subtypes. Indeed, the Cox *p*-values obtained from data integration are more significant than those obtained from individual data types.

To further demonstrate the usefulness of the developed method, we also performed a risk assessment using molecular data. We demonstrate that SCFA is able to predict risk scores that are highly correlated with vital status and survival probability. The correlation between predicted risk scores and survival

information has a median of 0.62 and can be as high as 0.83. More importantly, we demonstrate that the risk prediction becomes more accurate when more data types are involved.

## DATA AVAILABILITY STATEMENT

TCGA datasets were downloaded from <http://firebrowse.org/>. The docker contains the environment and scripts used in this article are available at: <http://scfa.tinnguyen-lab.com>. The current version of SCFA can be found at: <https://github.com/duct317/SCFA>. The SCFA software and TCGA data package will be available in the next release of Bioconductor.

## AUTHOR CONTRIBUTIONS

DT and TN conceived of and designed the approach. DT implemented the method in R, performed the data analysis and computational experiments. HN and UL helped with data preparation and some data analysis. DT, HL, GB, and TN wrote the manuscript. All authors reviewed and approved the manuscript.

## FUNDING

This work was partially supported by NASA under grant number 80NSSC19M0170, and by Research Enhancement Grant at UNR to TN. This work was also partially supported by the UPMC Start-up Fund to HL. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors

and do not necessarily reflect the views of any of the funding agencies.

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fonc.2020.01052/full#supplementary-material>

## REFERENCES

- Esserman LJ, Thompson IM, Reid B, Nelson P, Ransohoff DF, Welch HG, et al. Addressing overdiagnosis and overtreatment in cancer: a prescription for change. *Lancet Oncol.* (2014) 15:e234–42. doi: 10.1016/S1470-2045(13)70598-9
- Esserman L, Shieh Y, Thompson I. Rethinking screening for breast cancer and prostate cancer. *J Am Med Assoc.* (2009) 302:1685–92. doi: 10.1001/jama.2009.1498
- Jemal A, Siegel R, Ward E, Hao Y, Xu J, Thun MJ. Cancer statistics, 2009. *CA Cancer J Clin.* (2009) 59:225–49. doi: 10.3322/caac.20006
- Seidman H, Mushinski MH, Gelb SK, Silverberg E. Probabilities of eventually developing or dying of cancer—United States, 1985. *CA Cancer J Clin.* (1985) 35:36–56. doi: 10.3322/canjclin.35.1.36
- Uramoto H, Tanaka F. Recurrence after surgery in patients with NSCLC. *Transl Lung Cancer Res.* (2014) 3:242–9. doi: 10.3978/j.issn.2218-6751.2013.12.05
- Booth CM, Shepherd FA. Adjuvant chemotherapy for resected non-small cell lung cancer. *J Thoracic Oncol.* (2006) 1:180–7. doi: 10.1097/01243894-200602000-00016
- NSCLC Meta-analysis Collaborative Group. Preoperative chemotherapy for non-small-cell lung cancer: a systematic review and meta-analysis of individual participant data. *Lancet.* (2014) 383:1561–71. doi: 10.1016/S0140-6736(13)62159-5
- Felip E, Rosell R, Maestre JA, Rodriguez-Paniagua JM, Moran T, Astudillo J, et al. Preoperative chemotherapy plus surgery versus surgery plus adjuvant chemotherapy versus surgery alone in early-stage non-small-cell lung cancer. *J Clin Oncol.* (2010) 28:3138–45. doi: 10.1200/JCO.2009.27.6204
- Collisson EA, Bailey P, Chang DK, Biankin AV. Molecular subtypes of pancreatic cancer. *Nat Rev Gastroenterol Hepatol.* (2019) 16:207–20. doi: 10.1038/s41575-019-0109-y
- Dienstmann R, Vermeulen L, Guinney J, Kopetz S, Tejpar S, Tabernero J. Consensus molecular subtypes and the evolution of precision medicine in colorectal cancer. *Nat Rev Cancer.* (2017) 17:79–92. doi: 10.1038/nrc.2016.126
- Yam C, Mani SA, Moulder SL. Targeting the molecular subtypes of triple negative breast cancer: understanding the diversity to progress the field. *Oncologist.* (2017) 22:1086–93. doi: 10.1634/theoncologist.2017-0095
- Lehmann BD, Bauer JA, Chen X, Sanders ME, Chakravarthy AB, Shyr Y, et al. Identification of human triple-negative breast cancer subtypes and preclinical models for selection of targeted therapies. *J Clin Invest.* (2011) 121:2750–67. doi: 10.1172/JCI45014
- Kohonen T. The self-organizing map. *Proc IEEE.* (1990) 78:1464–80. doi: 10.1109/5.58325
- Tamayo P, Slonim D, Mesirov J, Zhu Q, Kitarawan S, Dmitrovsky E, et al. Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proc Natl Acad Sci USA.* (1999) 96:2907–12. doi: 10.1073/pnas.96.6.2907
- Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science.* (1999) 286:531–7. doi: 10.1126/science.286.5439.531
- Herrero J, Valencia A, Dopazo J. A hierarchical unsupervised growing neural network for clustering gene expression patterns. *Bioinformatics.* (2001) 17:126–36. doi: 10.1093/bioinformatics/17.2.126
- Luo F, Khan L, Bastani F, Yen IL, Zhou J. A dynamically growing self-organizing tree (DGSOT) for hierarchical clustering gene expression profiles. *Bioinformatics.* (2004) 20:2605–17. doi: 10.1093/bioinformatics/bth292
- McLachlan GJ, Bean R, Peel D. A mixture model-based approach to the clustering of microarray expression data. *Bioinformatics.* (2002) 18:413–22. doi: 10.1093/bioinformatics/18.3.413
- Ghosh D, Chinnaiyan AM. Mixture modelling of gene expression data from microarray experiments. *Bioinformatics.* (2002) 18:275–86. doi: 10.1093/bioinformatics/18.2.275
- Jiang D, Tang C, Zhang A. Cluster analysis for gene expression data: a survey. *IEEE Trans Knowledge Data Eng.* (2004) 16:1370–86. doi: 10.1109/TKDE.2004.68
- Brunet JP, Tamayo P, Golub TR, Mesirov JP. Metagenes and molecular pattern discovery using matrix factorization. *Proc Natl Acad Sci USA.* (2004) 101:4164–9. doi: 10.1073/pnas.0308531101
- Gao Y, Church G. Improving molecular cancer class discovery through sparse non-negative matrix factorization. *Bioinformatics.* (2005) 21:3970–5. doi: 10.1093/bioinformatics/bti653
- Sharan R, Shamir R. CLICK: a clustering algorithm with applications to gene expression analysis. In: *Proceedings of the 8th International Conference on Intelligent Systems for Molecular Biology.* Vol. 8. La Jolla, CA (2000) p. 16.
- Hartuv E, Shamir R. A clustering algorithm based on graph connectivity. *Inform Process Lett.* (2000) 76:175–81. doi: 10.1016/S0020-0190(00)00142-3
- Ben-Dor A, Shamir R, Yakhini Z. Clustering gene expression patterns. *J Comput Biol.* (1999) 6:281–97. doi: 10.1089/106652799318274
- Wilkerson MD, Hayes DN. ConsensusClusterPlus: a class discovery tool with confidence assessments and item tracking. *Bioinformatics.* (2010) 26:1572–3. doi: 10.1093/bioinformatics/btq170
- Monti S, Tamayo P, Mesirov J, Golub T. Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data. *Mach. Learn.* (2003) 52:91–118. doi: 10.1023/A:1023949509487
- Dudoit S, Fridlyand J. A prediction-based resampling method for estimating the number of clusters in a dataset. *Genome Biol.* (2002) 3:1–21. doi: 10.1186/gb-2002-3-7-research0036
- Ben-Hur A, Elisseeff A, Guyon I. A stability based method for discovering structure in clustered data. In: *Pacific Symposium on Biocomputing.* Vol. 7 (2001). p. 6–17. doi: 10.1142/9789812799623\_0002
- Tseng GC, Wong WH. Tight clustering: a resampling-based approach for identifying stable and tight patterns in data. *Biometrics.* (2005) 61:10–6. doi: 10.1111/j.0006-341X.2005.031032.x
- Zhang S, Liu CC, Li W, Shen H, Laird PW, Zhou XJ. Discovery of multi-dimensional modules by integrative analysis of cancer genomic data. *Nucleic Acids Res.* (2012) 40:9379–91. doi: 10.1093/nar/gks725
- Chalish P, Fridley BL. Integrative clustering of multi-level 'omic data based on non-negative matrix factorization algorithm. *PLoS ONE.* (2017) 12:e0176278. doi: 10.1371/journal.pone.0176278
- Wu D, Wang D, Zhang MQ, Gu J. Fast dimension reduction and integrative clustering of multi-omics data using low-rank approximation: application to cancer molecular classification. *BMC Genomics.* (2015) 16:1022. doi: 10.1186/s12864-015-2223-8
- Lock EF, Dunson DB. Bayesian consensus clustering. *Bioinformatics.* (2013) 29:2610–6. doi: 10.1093/bioinformatics/btt425

35. Kirk P, Griffin JE, Savage RS, Ghahramani Z, Wild DL. Bayesian correlated clustering to integrate multiple datasets. *Bioinformatics*. (2012) 28:3290–7. doi: 10.1093/bioinformatics/bts595
36. Mo Q, Shen R, Guo C, Vannucci M, Chan KS, Hilsenbeck SG. A fully Bayesian latent variable model for integrative clustering analysis of multi-type omics data. *Biostatistics*. (2017) 19:71–86. doi: 10.1093/biostatistics/kxx017
37. Mo Q, Wang S, Seshan VE, Olshen AB, Schultz N, Sander C, et al. Pattern discovery and cancer gene identification in integrated cancer genomic data. *Proc Natl Acad Sci USA*. (2013) 110:4245–50. doi: 10.1073/pnas.1208949110
38. Shen R, Olshen AB, Ladanyi M. Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. *Bioinformatics*. (2009) 25:2906–12. doi: 10.1093/bioinformatics/btp543
39. Shen R, Mo Q, Schultz N, Seshan VE, Olshen AB, Huse J, et al. Integrative subtype discovery in glioblastoma using iCluster. *PLoS ONE*. (2012) 7:e35236. doi: 10.1371/journal.pone.0035236
40. Wang B, Mezlini AM, Demir F, Fiume M, Tu Z, Brudno M, et al. Similarity network fusion for aggregating data types on a genomic scale. *Nat Methods*. (2014) 11:333–37. doi: 10.1038/nmeth.2810
41. Speicher NK, Pfeifer N. Integrating different data types by regularized unsupervised multiple Kernel learning with application to cancer subtype discovery. *Bioinformatics*. (2015) 31:i268–75. doi: 10.1093/bioinformatics/btv244
42. Rappoport N, Shamir R. NEMO: cancer subtyping by integration of partial multi-omic data. *Bioinformatics*. (2019) 35:3348–56. doi: 10.1093/bioinformatics/btz058
43. Ramazzotti D, Lal A, Wang B, Batzoglou S, Sidow A. Multi-omic tumor data reveal diversity of molecular mechanisms that correlate with survival. *Nat Commun*. (2018) 9:4453. doi: 10.1038/s41467-018-06921-8
44. Nguyen T, Tagett R, Diaz D, Draghici S. A novel approach for data integration and disease subtyping. *Genome Res*. (2017) 27:2025–39. doi: 10.1101/gr.215129.116
45. Nguyen H, Shrestha S, Draghici S, Nguyen T. PINSPlus: A tool for tumor subtype discovery in integrated genomic data. *Bioinformatics*. (2019) 35:2843–6. doi: 10.1093/bioinformatics/bty1049
46. Von Luxburg U. A tutorial on spectral clustering. *Stat Comput*. (2007) 17:395–416. doi: 10.1007/s11222-007-9033-z
47. Bellman R. *Dynamic Programming*. Princeton, NJ: Princeton University Press (1957).
48. Harman HH, Jones WH. Factor analysis by minimizing residuals (minres). *Psychometrika*. (1966) 31:351–68. doi: 10.1007/BF02289468
49. Breslow NE. Analysis of survival data under the proportional hazards model. *Int Stat Rev*. (1975) 43:45–57. doi: 10.2307/1402659
50. Cox DR. Regression models and life-tables. *J R Stat Soc Ser B*. (1972) 34:187–220. doi: 10.1111/j.2517-6161.1972.tb00899.x
51. Zou H, Hastie T. Regularization and variable selection via the elastic net. *J R Stat Soc Ser B*. (2005) 67:301–20. doi: 10.1111/j.1467-9868.2005.00503.x
52. Tran D, Nguyen H, Tran B, La Vecchia C, Luu HN, Nguyen T. Fast and precise single-cell data analysis using hierarchical autoencoder. *bioRxiv*. (2019) 799817:1–10. doi: 10.1101/799817
53. Li CH. The Performance of ML, DWLS, and ULS estimation with robust corrections in structural equation models with ordinal variables. *Psychol Methods*. (2016) 21:369–87. doi: 10.1037/met0000093
54. Wan S, Kim J, Won KJ. SHARP: Single-cell RNA-seq hyper-fast and accurate processing via ensemble random projection. *bioRxiv*. (2018) 461640:1–25. doi: 10.1101/461640
55. Simon N, Friedman J, Hastie T, Tibshirani R. Regularization paths for Cox's proportional hazards model via coordinate descent. *J Stat Softw*. (2011) 39:1–3. doi: 10.18637/jss.v039.i05
56. Shinagare AB, Vikram R, Jaffe C, Akin O, Kirby J, Huang E, et al. Radiogenomics of clear cell renal cell carcinoma: preliminary findings of the cancer genome atlas-renal cell carcinoma (TCGA-RCC) imaging research group. *Abdominal Imaging*. (2015) 40:1684–92. doi: 10.1007/s00261-015-0386-z
57. Thomas GV, Tran C, Mellinghoff IK, Welsbie DS, Chan E, Fueger B, et al. Hypoxia-inducible factor determines sensitivity to inhibitors of mTOR in kidney cancer. *Nat Med*. (2006) 12:122–7. doi: 10.1038/nm1337
58. Varela I, Tarpey P, Raine K, Huang D, Ong CK, Stephens P, et al. Exome sequencing identifies frequent mutation of the SWI/SNF complex gene PBRM1 in renal carcinoma. *Nature*. (2011) 469:539–42. doi: 10.1038/nature09639
59. Harrell FE, Califf RM, Pryor DB, Lee KL, Rosati RA. Evaluating the yield of medical tests. *J Am Med Assoc*. (1982) 247:2543–6. doi: 10.1001/jama.1982.03320430047030

**Conflict of Interest:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2020 Tran, Nguyen, Le, Bebis, Luu and Nguyen. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.