



Integrative Analysis of Multi-Genomic Data for Kidney Renal Cell Carcinoma

Ashwinder Singh¹ · Neelam Goel¹ · Yogita²

Received: 20 January 2019 / Revised: 21 July 2019 / Accepted: 24 July 2019 / Published online: 8 August 2019
© International Association of Scientists in the Interdisciplinary Areas 2019

Abstract

Accounting for nine out of ten kidney cancers, kidney renal cell carcinoma (KIRC) is by far the most common type of kidney cancer. In view of limited and ineffective available therapies, understanding the genetic basis of disease becomes important for better diagnosis and treatment. The present studies are based on a single type of genomic data. These studies do not consider interactions between genomic data types and their underlying biological relationships in the disease. However, the current availability of multiple genomic data and the possibility of combining it have facilitated a better understanding of the cancer's characterization. But high dimensionality and the existence of complex interactions (within and between genomic data types) are the two main challenges of integrative methods to analyze cancer effectively. In this paper, we propose a method to build an integrative model based on Bayesian model averaging procedure for improved prediction of clinical outcome in cancer survival. The proposed method initially uses dimensionality reduction techniques to generate low-dimensional latent features for the predictive models and then incorporates interactions between them. It defines the latent features using principal components and their sparse version. It compares the predictive performance of models based on these two latent features on real data. These models also validate several ccRCC-specific cancer biomarkers previously reported in the literature. Applied on kidney renal cell carcinoma (KIRC) dataset of The Cancer Genome Atlas (TCGA), the method achieves better prediction with sparse principal components model by including latent feature interactions as compared to without including them.

Keywords Integrated analysis · Genomics · miRNA · mRNA · Copy number alteration · Kidney cancer

1 Introduction

Kidney cancer or renal cell carcinoma (RCC) has been ranked as the seventh leading cancer type among men in western communities. The incidence of RCC steadily rises by 2–4% each year [1]. RCC is a collection of various histological subtypes such as clear cell renal cell carcinomas (ccRCC), papillary renal cell carcinomas (pRCC), and chromophobe renal cell carcinomas (crRCC). Among them, ccRCC is the most common (70–85%) and lethal subtype [1]. Surgical and targeted therapies exist to treat the kidney cancer and they are also successful in improving the patient's overall survival [2]. But most patients ultimately grow resistance toward these treatments and surrender to the disease.

Besides multiple discussions on cancer evolution and progression by various studies [3], cancer at its core is characterized by somatic copy number alterations and unique gene expression profiles. Therefore, there is a need to thoroughly

Electronic supplementary material The online version of this article (<https://doi.org/10.1007/s12539-019-00345-8>) contains supplementary material, which is available to authorized users.

✉ Neelam Goel
erneelam@pu.ac.in
Ashwinder Singh
ashwinder389@gmail.com
Yogita
yogitathakran@nitm.ac.in

¹ University Institute of Engineering and Technology, Panjab University, Chandigarh 160014, India

² Department of Computer Science, National Institute of Technology, Meghalaya 793003, India

understand the ccRCC disease for building reliable prognostic and therapeutic strategies by incorporating genomic data. Earlier research based on a single type of genomic data has reported a number of molecular alterations in ccRCC at the mRNA, miRNA, and DNA (copy number alterations) level. Most of the ccRCC cases have shown alterations in the short arm of chromosome 3 and 30–56% have the VHL (Von Hippel-Lindau) gene mutated [4]. However, these studies have not produced sufficient results as they lack in exploring the complex mechanism of multiple genomic processes in human diseases. Now, with emerging technologies for genome profiling, multiple genomic data types are available and analytic methods for integrating these data types provide a better understanding of cancer evolution and progression. This results in identification of targets and clinical prediction of cancer by incorporating the necessary interactions between the data types. Many recent studies have shown benefits of the integrated approach in ccRCC [5–10].

However, these analytical methods for multiple data types are facing two challenges: high dimensionality of data and the presence of complex correlations and interactions both within and between platform-specific features [11]. The proposed method is driven by the dataset from TCGA (The Cancer Genome Atlas) Pan-Cancer Survival Prediction Challenge project that contains different molecular types of KIRC (kidney renal cell carcinoma). In the proposed method, principal component analysis and its sparse version are the machine learning approaches used to overcome the first challenge of high dimensionality. The second challenge is handled through modeling the interactions by taking the product of principal component score vectors [12]. Additionally, it also finds important genomic variables that are linked to ccRCC progression.

To the best of our knowledge, very few studies of integrative analysis for ccRCC are available [5–10] and none of them have incorporated multi-level interaction effects, within and between the molecular data types when fitting the integrative model for RCC. So the proposed work contributes significantly in ccRCC research by providing a unique methodology that contains data type interaction effects at different levels. The method achieves better prediction with sparse principal components model by including latent feature interactions as compared to without including them.

2 Related Work

Earlier research on various types of cancer such as gene expression profiles in breast cancer [13], miRNA in lung cancer [14], copy number alterations in ovarian cancer [15], etc., was mainly focussed on single type of genomic data to derive biomarkers of prognostic significance or improve the clinical outcome of cancer. Although these studies

helped in important discoveries, they were limited to one type of molecular data. However, a thorough and comprehensive understanding of cancer development and its biological mechanism requires the examination of the interplay between different layers of genomic data. This has motivated current research to integrate diverse types of genomic data. These studies have revealed many benefits of the integrated approach in different cancers [16, 17]. With the similar focus, various integrative studies in KIRC were conducted.

Dondeti et al. [5] identified potentially important targets in ccRCC by combining copy number and gene expression data. Two important chromosome 5q oncogenes are discovered whose overexpression play a sufficient role in promoting tumorigenesis in ccRCC. An integrated molecular analysis of ccRCC by Sato et al. [6] identified new mutated genes and pathways that are involved in the pathogenesis of ccRCC. Gene expression, DNA methylation, and copy number data for more than 100 ccRCC samples were analyzed using different sequencing techniques.

Multiple datasets of miRNA expression related to ccRCC were incorporated into an integrative framework by Chen et al. [7]. The study discovered 14 unique molecular pathways that have an important role in the production of ccRCC tumor. Integrative analysis for analyzing mRNA and miRNA interactions together was performed to build a predictive model for survival outcome by Chekouo et al. [8]. The Bayesian model proposed by them also identifies cancer biomarkers specific to KIRC progression.

A study by Butz et al. [9] integrated mRNA, microRNA, and protein expression data of ccRCC using pathway analysis. They identified three new potential biomarkers that are linked to kidney cancer. Similarly, the work by Bluysen et al. [10] reviews the recent findings in the integrative studies of ccRCC. It discusses how significant technological advances led to the availability of different genomic data and helped in understanding the complex pathology of ccRCC and its molecular mechanism.

3 Dataset

In this study, the proposed method is tested using KIRC dataset [18] from TCGA Pan-Cancer Survival Prediction Challenge project [19]. The project home page can be accessed on Synapse (<http://dx.doi.org/10.7303/syn1710282>). The data available for each cancer type on the website contain core sample sets, comprising overall survival time, different types of molecular data, etc. The core tumor sample set is used in this study. The core data contain the survival time, gene expressions, micro-RNA expressions, and copy number alterations for tumor samples of patients diagnosed with KIRC. Survival data contain entries about overall survival

time (to death) for each patient in days. Three genomic data types used in the study are as follows:

1. *mRNA Expression*: Messenger RNA or mRNA is the key molecule to enable gene expression for the production of proteins. For sequencing mRNA data, RNA sequencing (RNA-Seq) is used. Illumina HiSeq 2000 is the instrument used in RNA sequencing of the data used in this study.
2. *microRNA (miRNA) Expression*: miRNAs constitute a recently discovered class of short non-coding RNAs of around 22 nucleotides that have crucial roles in regulating the gene expression [20]. microRNA sequencing (miRNA-Seq) is used for sequencing miRNA data. It is a type of RNA-Seq, which is also known as small RNA sequencing as it constitutes small RNAs. Illumina Genome Analyzer/HiSeq 2000 is used as a tool or platform for performing small RNA sequencing for the miRNA data used in this study.
3. *Somatic CNAs (Copy Number Alterations)*: Also referred as CNV (copy number variation), somatic CNA is a phenomenon in which parts of the genes are duplicated or deleted. An SNP array is a type of DNA microarray that is used to detect mutations in the genomic sequence. Chip-based methods for SNP arrays such as comparative genomic hybridization can detect genomic alterations leading to the loss of heterozygosity (LOH). Such a chip-based method or platform by Affymetrix, known as the Genome-Wide Human SNP Array 6.0 is used in this study for the detection of copy number variations.

The four types of data (three genomic types and survival time) were taken for 243 patients. Initially, predictors/features with zero variance from the three genomic data types were eliminated, leaving 795 features in miRNA, 20,203 in mRNA, and 69 alterations in sCNA. The dataset is summarized in Table 1.

4 Proposed Method

The proposed method initially analyzes data that are obtained from different data types by reducing the dimensionality using dimension reduction techniques. The resulting data are then integrated into a single statistical

model by incorporating between and within interactions among data types, to predict clinical outcome and identify the clinically relevant genes. The diagram representing the proposed method is shown in Fig. 1.

Let X_1, \dots, X_K be the $n \times l_1, \dots, n \times l_K$ matrices and Y be the $n \times 1$ vector. These matrices represent the values of K groups (genomic data types) used in the present model with l_1, \dots, l_K genomic features along with the responses (clinical outcomes) vector taken from a random sample of n observations. The aim of the function is to predict the values in Y from the K groups of features and the interactions among them.

A conceptual model integrating the interactions within and between the groups of features can be written as

$$Y = t_1(X_1) + t_2(X_2) + \dots + t_k(X_k) \tag{1}$$

$$+ s_{11}(X_1 \times X_1) + s_{22}(X_2 \times X_2) + \dots + s_{KK}(X_K \times X_K) \tag{2}$$

$$+ s_{12}(X_1 \times X_2) + s_{13}(X_1 \times X_3) + \dots + s_{(K-1)K}(X_{K-1} \times X_K) \tag{3}$$

+ e ,

where “ $A \times B$ ” is a matrix in which the i th row value corresponds to the Cartesian product of the values of the i th rows of A and B (i.e., the values of the interaction terms for observation i), and “ $A \times A$ ” is a matrix in which the i th row value corresponds to all pairwise products of the values in the i th row of A (so that there are no second-order terms in the model), for $i = 1, \dots, n$ [11]. Here, consider $\{t(\cdot), s(\cdot)\}$ be two functions of a data matrix X , which are defined below, and e is an $n \times 1$ vector of error terms. The model terms are as follows:

- Term (1) denotes data type-specific effects modeled as main effects for each data type.
- Term (2) denotes within data type interaction effects and it represents interactions among the features from the similar data type.
- Term (3) denotes between data type interaction effects and it represents interactions among the features from the different data types.

Now, to fit the above model, one needs to define the functions $s(\cdot)$ and $t(\cdot)$. The function $t_k(X_k)$ can be defined

Table 1 Dataset used in the study

mRNA	miRNA	Somatic CNA
Illumina HiSeq 2000 RNA sequencing V2, > 20,000 genes	Illumina Genome HiSeq 2000 miRNA sequencing platform, > 700 microRNAs	Affymetrix Genome-wide human SNP Array 6.0, ~ 80 arm or focal alterations

The first row displays the molecular type and the second row displays the platform along with the number of expressions in the data

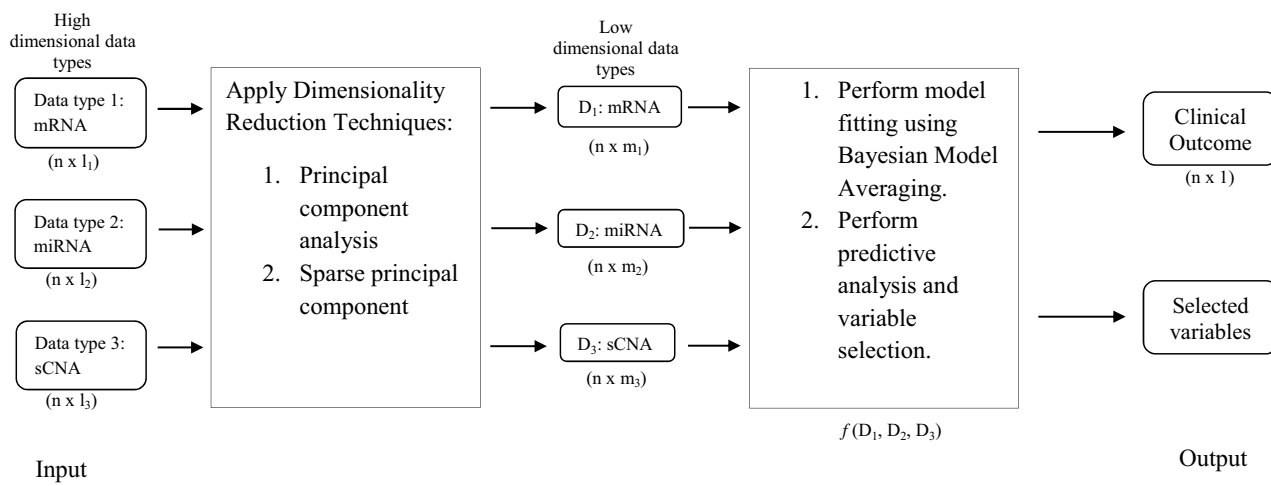


Fig. 1 High-dimensional data predictors from multiple genomic data like copy number and gene expression are converted into low-dimensional data predictors using dimensionality reduction techniques. Subsequently, within and between interactions of obtained predictors

are used to perform predictive analysis using BMA for the required clinical outcome like survival time, and then variable selection procedure is performed

as a linear function $X_k \alpha_k$, i.e., $t_k(X_k) = X_k \alpha_k$. The function $s_{kp}(X_k \times X_p)$ can be defined as a linear function $(X_k \times X_p) \delta_{kp}$, i.e., $s_{kp}(X_k \times X_p) = (X_k \times X_p) \delta_{kp}$, where δ_{kp} is a vector of parameters having the identical length as $(X_k \times X_p)$ and α_k is $l_k \times 1$, for $k, p \in \{1, \dots, K\}$. With these definitions, the model can be written as,

$$Y = \alpha_o + \sum_k \sum_{j=1}^{l_k} \alpha_{kj} X_{kj} + \sum_k \sum_{0 < i < j \leq l_k} \gamma_{kji} X_{kj} X_{ki} + \sum_{k > p} \sum_{j=1}^{l_k} \sum_{i=1}^{l_p} \eta_{kpji} X_{kj} X_{pi} + e, \tag{4}$$

where X_{kj} is the j th column of X_k , α_o the intercept, α_{kj} the member of α_k , and γ_{kji} and η_{kpji} are members of δ_{kp} for $k, p \in \{1, \dots, K\}$ [11]. Now if there are $\bar{l} = l_1 + l_2 + \dots + l_k$, then Eq. (4) will have $\bar{l} + \bar{l}(\bar{l} - 1)/2$ values, which will even surpass the total observations n . In such a scenario, taking higher-order interactions will further increase the number of values polynomially, which may result in unstable model fitting. The KIRC dataset has data for $n = 243$ patients with $\bar{l} = 21,607$ predictors, that leads to $(21,607)(21,606)/2 = 233,420,421$ possible two-way interactions!

To simplify this, the dimensionality of input is reduced that will cover the maximum information in the data with lesser dimensions. If R is the dimension reduction technique that projects the higher dimensional features of data matrix X_k for $k = 1, \dots, K$ into lower dimensional latent features matrix H_k ($n \times h_k$) containing latent feature scores, such that h_k is less than l_k , then R can be defined as follows:

$$R(X_{k1}, \dots, X_{kl_k}) \equiv \{H_{k1}, \dots, H_{kh_k}\}$$

using R , the new feature set is h_k which is of far lower dimensions (tens) than l_k (thousands). Therefore, the model equation can be rewritten with new functions constituting lower dimensional latent feature scores and their interactions, such as $\bar{t}_k(H_k) = H_k \alpha_k$ and $\bar{s}_{kp}(H_k \times H_p) = (H_k \times H_p) \delta_{kp}$, for $k, p \in \{1, \dots, K\}$.

With these definitions, the model can be rewritten as,

$$Y = \bar{\alpha}_o + \sum_k \sum_{j=1}^{l_k} \bar{\alpha}_{kj} H_{kj} + \sum_k \sum_{0 < i < j \leq l_k} \bar{\gamma}_{kji} H_{kj} H_{ki} + \sum_{k > p} \sum_{j=1}^{l_k} \sum_{i=1}^{l_p} \bar{\eta}_{kpji} H_{kj} H_{pi} + \bar{e}, \tag{5}$$

where H_{kj} is the j th column of H_k , $\bar{\alpha}_o$ the intercept, $\bar{\alpha}_{kj}$ the main effect of the j th latent feature of the k th variable group, $\bar{\gamma}_{kji}$ the interaction effect between the i th and the j th latent feature from the k th variable group, and $\bar{\eta}_{kpji}$ is the interaction effect between the j th latent feature from the k th variable group and the i th latent feature for the p th variable group, for $k, p \in \{1, \dots, K\}$ [11].

Therefore, for the fitting model (5), a dimensionality reduction approach R is needed which can be applied on each given data matrix X_k of actual features, resulting in an $n \times h_k$ matrix H_k such that $h_k \leq l_k$ for $k = 1, \dots, K$ and $\bar{h} + \bar{h}(\bar{h} - 1)/2$ is less than n . Therefore, H_1, \dots, H_k will constitute of h_1, \dots, h_k latent scores for n units. If we compare the number of predictors and their interactions in this model from the model (4), there is a substantial reduction in the number of main effects and interaction effects, e.g., if $h_1 = 6, h_2 = 5, h_3 = 6$, then model (5) would contain 17 main effects and $17(16)/2 = 136$ interaction effects and

overall 153 effects. This makes the model more scalable, interaction preserving and it also maintains the predictive accuracy.

Fitting the Model: After applying the above dimension reduction approach that maps the actual feature space from X to H space containing latent scores, next step is to perform model fitting (B) and variable selection procedure to evaluate the parameters in (5).

4.1 Dimensionality Reduction Techniques

For performing the function R in this work, principal component analysis (PCA) technique and its sparse version are used. PCA is a technique used to transform variables from higher dimensional space into lower dimensional space, where transformed variables are a linear combination of original variables. These new set of transformed variables is called principal components. In this case, new features will be of the form, $H_{kj} = \rho_{kj1} X_{k1} + \dots + \rho_{kjk} X_{kk}$ for $j = 1, \dots, h_k$, which also helps in finding the dependencies in terms of the original features. Therefore, each principal component is a weighted average of all variables (e.g., genes) with a weight (called loading coefficient) assigned to each variable. The sparse version will take the loadings of variables that are ineffective in PCA as zero, which in turn helps in the variable selection process.

Various papers [21–25] demonstrate the use of dimensionality reduction techniques in case of datasets with large number of dimensions or features to reduce the number of computations and simplify the handling of data. One such technique is PCA, which is quite commonly used for dimensionality reduction in bioinformatics [12], and it can be implemented using singular value decomposition on matrix X_k for $k = 1, \dots, K$, where h_k is the rank of the decomposition. It results in orthogonal components that are non-collinear and capture most of the information of original dataset. These principal components are ordered as per the maximum possible variance of the component, with the first having the maximum possible variance and so on. Different methods are available to specify the number of principal components to be retained [26]. In this work, we have used the method of “scree plot test”. The expected pattern in a scree plot includes a steep curve which is followed by a bend and ends with a horizontal line. Those components (or factors) are retained in the steep curve, which are before the first point that starts the flat line trend. The sparse version of PCA has indicated various advantages over traditional PCA in cancer research [27].

4.2 Model Fitting Using Bayesian Model Averaging

Here, the model Eq. (5) will be fitted with the obtained latent feature scores from the dimensionality reduction technique. Bayesian model averaging (BMA) procedure is selected as (B) to be used for the fitting model (5) on latent features. The typical model selection includes selecting a model from a class of models, and then continues as if the selected model had generated the data. But this leads to overconfident decisions and inferences. Compared to these regular modeling methods which overlook model uncertainty, BMA considers uncertainty and makes inferences by averaging over the posterior distributions of a range of possible models, weighted by their posterior model probabilities. This helps in selecting the most appropriate model for a given outcome variable as it has been shown that BMA gives a better predictive performance for new observations than fitting a single assumed to be the best model [28].

The BMA algorithm assigns a posterior probability to each model and for each variable included in a given model, the probability that the coefficient (or parameter value) for a given variable is non-zero is returned. Therefore, either the model with the highest posterior probability can be selected or a model that contains every variable for which the probability that the coefficient is non-zero is above some threshold can be selected. In this work, the model with the highest posterior probability is selected.

4.3 Selection of Significant Variables

Having obtained the model equation, the significant variables are selected from it. A list of significant variables is as follows. Dimensionality reduction technique R applied on X_k for each $k = 1, \dots, K$ data type, generates a set of h_k latent feature score vectors making H_k matrices, which are linear combinations of the original column vectors X_{k1}, \dots, X_{kk} such that $H_{kj} = \rho_{kj1} X_{k1} + \dots + \rho_{kjk} X_{kk}$ for $j = 1, \dots, h_k$. Depending on the contribution of variables in the linear combination, R assigns higher or lower loading to that variable. The model selection process B results in a set of indices $\mathcal{L} \subset \{(k, j) : j = 1, \dots, h_k, k = 1, \dots, K\}$ such that the set of latent features $\mathbb{N} \equiv \{H_{kj} : (k, j) \in \mathcal{L}\}$ is preserved in the model, where values of \mathbb{N} can occur either as main effects or as part of an interaction. Now, the variables X_{k1}, \dots, X_{kk} are ordered as per their contributions. The maximum magnitude of the loadings that are assigned to each variable across all the latent features from group k retained by B is taken as a contribution.

Consider $\{x_{k1}^2, \dots, x_{kk}^2\}$ be the vector of maximum loading magnitudes arranged in non-increasing order. Now square all the components individually and divide each of them by the

sum of these squared components resulting in $y_{k_j} = x_{k_j}^2 / (x_{k_1}^2 + \dots + x_{k_k}^2)$ for $j = 1, \dots, l_k$. Next, consider the variables associated with $(x_{k_1}^2, \dots, x_{k_z}^2)$ to be significant if $z = \min \{g : y_{k_1} + \dots + y_{k_z} > \tau\}$, for some threshold $\tau \in (0, 1)$. If $\tau = 0.8$, the procedure will select variables with squared maximum loading magnitude that constitutes minimum of 80% of the sum of squared maximum loading magnitudes for the variables in that group.

5 Results and Discussion

The present study is motivated by the challenges associated with analyzing the multi-genomic dataset for KIRC, available from TCGA Pan-Cancer Survival Prediction Challenge project. The focus of this study is on integrating gene expression and copy number data from the KIRC study. In this study, the outcome of interest is overall survival time acquired from $n = 243$ patient samples, where survival time is the time from initial diagnosis to death. The objective is to integrate the data from three genomic data types, such as mRNA, miRNA, and sCNA to predict the patient’s (log-transformed) survival time and to identify genes of biological significance in KIRC.

After removal of zero variance features, along with the survival time, the input data consists of feature matrices for three genomic data types as follows: $l_{cnv} = 69$, $l_{miRNA} = 795$, and $l_{mRNA} = 20,203$, summing up to $\bar{l} = 21,607$ features. On application of principal components (PC) and sparse principal components (SPC) on these matrices, each technique selected to keep five sCNA features, six mRNA features, and five miRNA latent features, i.e., $h_{cnv} = 5$, $h_{mRNA} = 6$, and $h_{miRNA} = 5$. Subsequently, these latent features and their interactions are used for fitting the model, wherein the best model is selected using the Bayesian model averaging procedure. The fitted linear regression model [29] is then used to predict the response variable, i.e., survival time for both with and without interaction effects. Thereafter, statistical results for linear regression model are calculated. To choose the best predictive model for latent features with and without the inclusion of interactions, a tenfold cross-validation procedure is used. The procedure splits the data into ten equal-sized parts (folds). Then one part is retained for predicting the response time and remaining nine parts are used for model fitting. This is repeated ten times for each part of the data and the resulting mean squared error of prediction (MSEP) is computed by taking an average of mean squared errors over all the parts. Further an alternative

method is taken to test the prediction accuracy of the models. Here, an independent dataset test is performed where the data to be tested are never exposed during the model development process. The dataset is split according to 80/20 rule, i.e., 80% of dataset form the training set and 20% form the test set. The obtained models from BMA (with and without interactions) are first trained and then tested using the corresponding split datasets. Subsequently, the mean square error (MSE) is calculated for measuring the prediction error in the models.

Now, to perform the significant variable selection from the latent features that remain in the model, a threshold of eight is set, such that the energy retained is 80%. Finally, a list of genes obtained from the individual models for PC and SPC with the inclusion of interactions is prepared to check their biological roles.

Implementation of this work is performed in R language. PCA technique is implemented using the singular value decomposition (SVD) algorithm of the standard R package. SPC is implemented using the R language package ‘PMA’ [30], which executes the algorithm described in Witten et al. [30]. BMA is implemented using the R language package BMA [31].

5.1 Experimental Results

The Bayesian model averaging procedure selected regression with high posterior probability, which resulted in the selection

Table 3 Linear regression statistics obtained from principal components and sparse principal components

Regression statistics	
Principal components (PC)	
Multiple R	0.185772
R^2	0.034511
Adjusted R^2	0.030505
Standard error	1.452905
Observations	243
Sparse principal components (SPC)	
Multiple R	0.255895
R^2	0.065482
Adjusted R^2	0.057695
Standard error	1.432388
Observations	243

Bold values are used to indicate better model fitting in SPC model than PC model

Table 2 Selected models by principal components and sparse principal components

Principal components (PC)	$\hat{y} = 6.25 + 0.0028CNV_3mRNA_3$
Sparse principal components (SPC)	$\hat{y} = 6.36 - 73.90 CNV_5mRNA_6 + 62.96 CNV_2miRNA_2$

Fig. 2 a and b: Variables with continuous horizontal bands are selected for our models as these are included regularly in BMA exploration models, sorted in non-increasing order from left to right as per their posterior probabilities. Numbers following data type names (*CN* copy number, *mR* mRNA, *miR* = miRNA) list the latent features with interaction effects for principal components (PC) and sparse principal components (SPC)



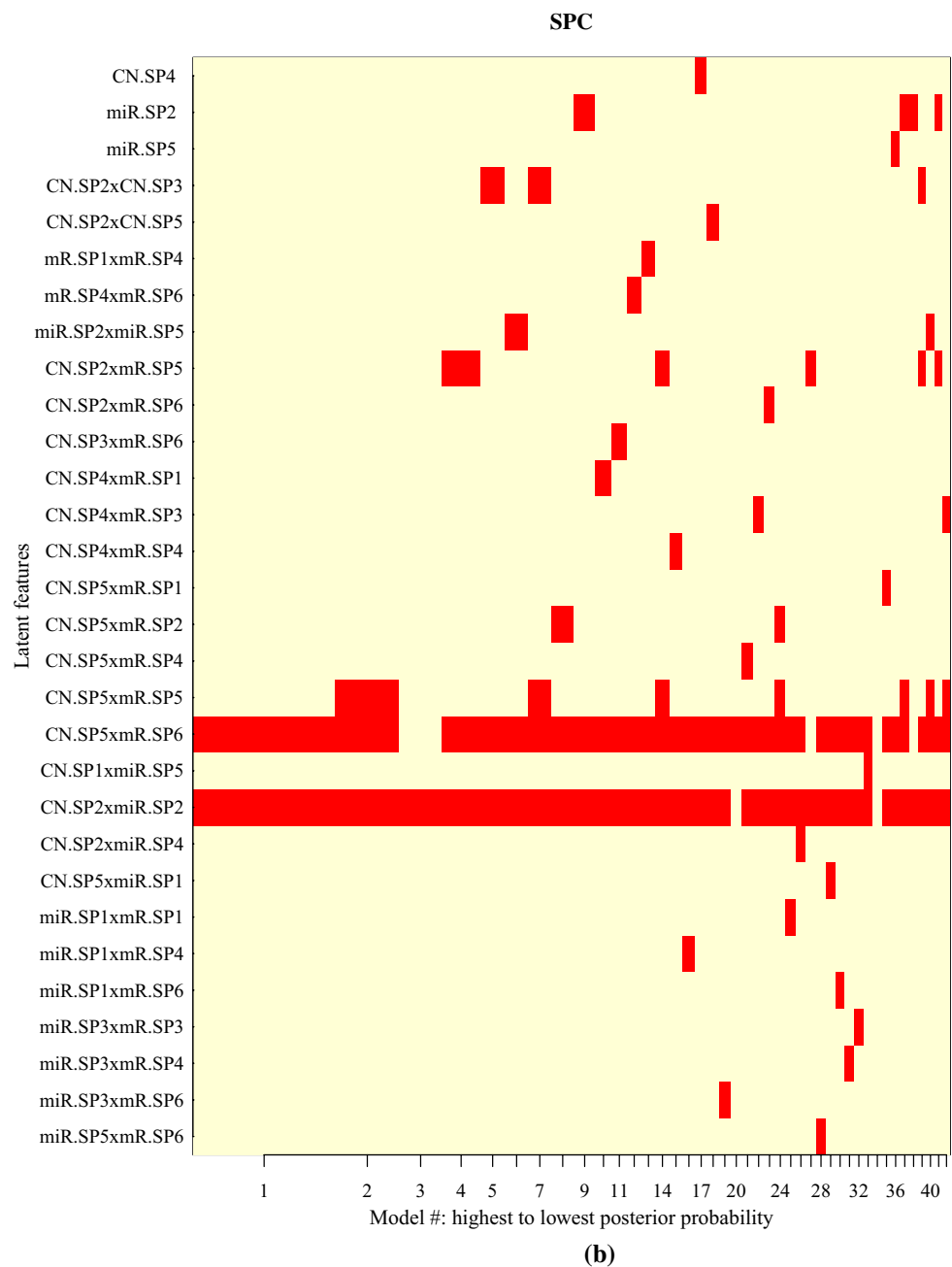
of CNV and mRNA interaction as the variable in case of PC. In the case of SPC, CNV miRNA interaction and CNV mRNA interaction as the variables were selected. These equations are given in Table 2.

It can be observed that the model obtained from PC is a linear regression model with only one input variable whereas model from SPC is a multiple linear regression model with two input variables, one negatively correlated and one positively correlated. The linear regression statistics obtained for these two models are displayed in Table 3. Table 3 indicates that the standard residual error in PC model is somewhat higher than

SPC model. Latent variables selected are also different in both cases. Moreover, higher adjusted R^2 , multiple R , and R^2 values in SPC than in PC show a better model fitting in SPC.

Low R^2 values found in both the models indicate the inherent greater amount of unexplainable variation. Still the conclusion can be drawn that when multiple variables are included for a regression model, latent features with interactions have a reliable and statistically significant role in the models and, therefore, leads to better predictive and variable selection results. Additionally, Akaike Information Criterion (AIC) is used to compare both the models. AIC considers both the

Fig. 2 (continued)



fitness of the model and the number of parameters used. We obtained AIC values 854 and 860 for SPC and PC models, respectively. A lower AIC value in case of SPC model that contains more parameters indicates a better fit.

The high probability fitted model and corresponding variable(s) selected by BMA for PC and SPC-based latent features are shown in Fig. 2. Each square displays a matrix, in which a variable is denoted by each row, and selected model in the BMA exploration is denoted by each column. The selected models are sorted in non-increasing order from left to right as per their posterior probabilities. The rectangle in the matrix is red if the variable is present in the model

and white otherwise. The model’s posterior probability is relative to the thickness of each column. The plots make it easy to see the variables picked by the most probable models in BMA. The continuous horizontal bands show the consistency and convergence of model selection, representing that a variable appears regularly in the BMA exploration. It can be seen in Fig. 2 that under PC decomposition, only one interaction between CNV and mRNA represented by label CN.P3 × mRNA.P3 is selected and it is present consistently in the selected models, signifying that it is selected as a result of convergence. However, for SPC decomposition, two interactions, one between CNV and mRNA, and

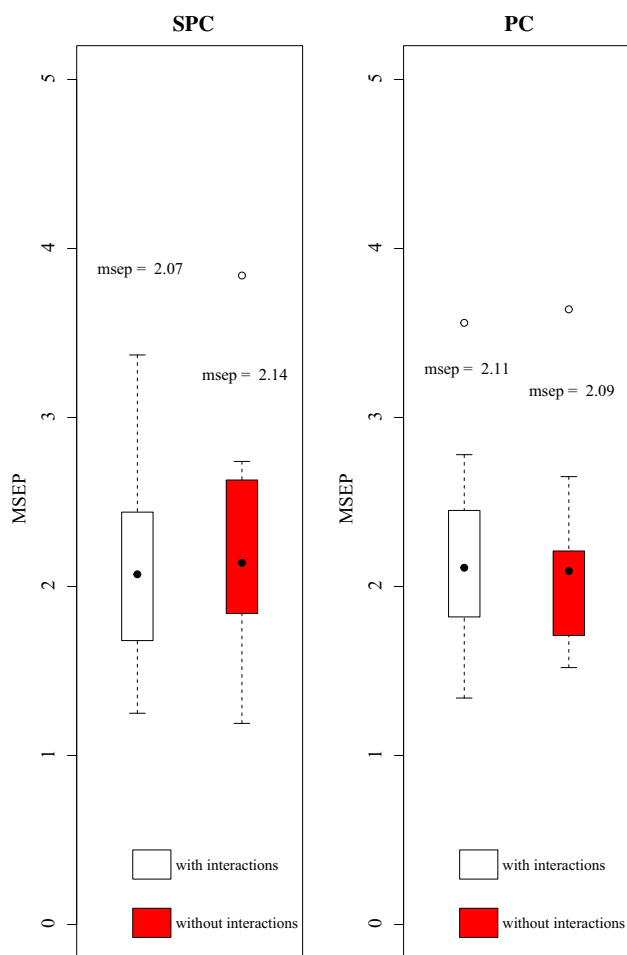


Fig. 3 Boxplots of mean squared error prediction resulted from tenfold cross validation for the proposed method in case of SPC- and PC-based dimensionality reduction models with and without the inclusion of latent feature interactions. Obtained MSEP is shown above each box

Table 4 MSE values achieved in case of 80/20 split test for PC and SPC based models with and without interactions

Model	MSE value
PC model with interactions	49.80
SPC model with interactions	57.22
PC model without interactions	46.60
SPC model without interactions	51.72

one between CNV and miRNA are selected, represented by labels CN.SP5 × mR.SP6 and CN.SP2 × miR.SP2, respectively. It can be noted that the first interaction is negatively associated while second is positively associated with the outcomes, and both appear consistently in the selected models. If only mRNA and CNV data types would have been considered for regression, then we would have 20,203 + 69

+ 20,272(20,271)/2 = 205,497,264 independent variables in the regression model, which is computationally impractical. This makes the proposed method advantageous and efficient to deal with interactions.

Figure 3 shows the boxplots indicating (with filled circles and in text) the MSEP achieved by tenfold cross validation, performed on PC and SPC models with and without the inclusion of interaction effects. It can be observed that in the case of PC model, when including the interaction terms MSEP was higher, as compared to PC model without interaction terms. However, MSEP is lower in the case of SPC model that includes interaction terms as compared to without interaction terms.

In the case of train/test split, datasets are split as per 80/20 rule. Out of total 243 observations, 194 were retained as training dataset and 49 as testing dataset. Table 4 shows the MSE values obtained from the PC and SPC models with and without interactions. Better fitting and prediction accuracy in PC models are observed.

Plots of variable selection procedure conducted for the SPC model are shown in Fig. 4a–d. It illustrates sorted loading magnitudes for the 69 copy number alterations, 795 miRNA expression levels, and 20,203 mRNA expression levels, for the terms or components obtained from the selected linear model. Filled circles in blue correspond to the selected variables on the application of variable selection for each term at the threshold level of $\gamma = 0.8$, while black ones are not selected.

The list of all the selected features obtained from PC and SPC model and the model equations for these models with and without the inclusion of interaction effects can be found in Supplementary file 1.

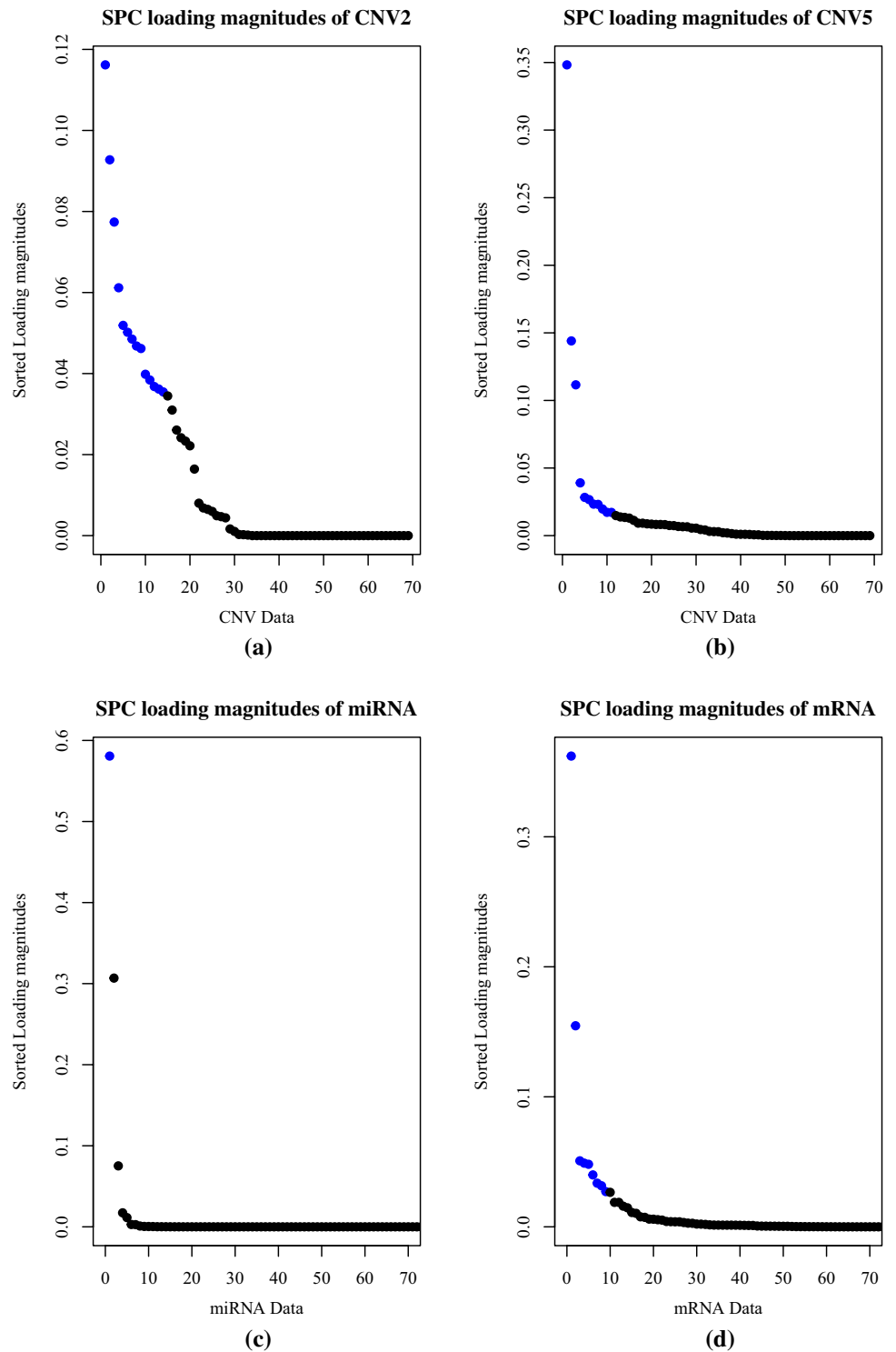
5.2 Biological Significance

From the variable selection process, a list of genes (for probes associated with expression and copy number) is prepared together with miRNAs for the PC and SPC models. This list is used to find the genes of biological significance by referring the published work. A majority of the selected variables are found to be associated with the KIRC cancer.

Inactivation of the Von Hippel-Lindau (VHL) tumor suppressor gene has been found to be responsible for the majority of ccRCC cases [32] and the proposed method has identified VHL (Von Hippel-Lindau Tumor Suppressor Like) gene from the PC model. Additionally, the proposed method has identified significantly mutated genes from the models that are associated with the pathogenesis of ccRCC. It is evident from the findings of genes such as BAP1, SETD2, TCEB1, TET2 in mRNA variable analysis, which are reported in [6] as significant mutations in ccRCC.

Further, various new genes such as ACHANK, CUL7, MLL2 that are reported in [10] that have played a potential

Fig. 4 Plots of variable selection conducted for SPC model



role in renal cell tumorigenesis, are identified by the models. The miRNAs variable selected by SPC model discovered miR-21 and miR-10b, which are stated in [18] to have strong regulatory interactions with ccRCC. Alterations of chromosomal regions in ccRCC have resulted in new candidate tumor suppressor genes (TSGs) and oncogenes.

The proposed model identified copy number alteration (or CNV) at 1q24.1 that is reported in [32] as a potential risk factor for RCC. Other significant regions that were stated for ccRCC in [18] are at 3p25.3, 6q26, 9p23 for oncogenes VHL, QKI, and PTPRD, respectively. Supplementary file 2 lists all the significant markers based on the cited literature that are found from both the models.

5.3 Discussion

The present study is motivated by TCGA pan-cancer survival prediction project [19] that provides open access to well curated, computable datasets to analyze TCGA data for improving prognostic models. Unlike other such community-based project [33] that mainly deals with a specific type of cancer and data, the above project is chosen as it includes different types of cancer and their molecular data types from TCGA.

We have proposed a method for integrative analysis of different genomic data types available for KIRC dataset from the Pan-Cancer Survival Prediction project. The method incorporates interactions within and between these data types to build the model that predicts survival time of patients and identify significant tumor biomarkers. The model has the ability to simultaneously model all type of relations in the data in a single model and may be used for clinical diagnosis in future with further improvement in accuracy.

For a fair comparison with a work which employs similar predictive evaluation metric to measure the performance of linear regression-based survival model, we used the latent feature decomposition (LFD) study [11] applied on glioblastoma multiforme (GBM) dataset.

The LFD study integrated the data from four genomic platforms—mRNA, miRNA, DNA methylation, and CNV. It used several dimensionality reduction techniques to build the survival model and reported principal components and sparse principal components techniques to achieve the best results.

In case of both PC and SPC-based model in LFD, obtained mean squared error of prediction with interactions is 1.20. This seems to suggest that model fitting is better in LFD than the model used in this study. However, the proposed model brings new insights with reliable accuracy and variable selection into the integrative study based on KIRC.

Some of the limitations in the integrative study involve time intensive calculations for large-scale datasets while modeling interactions. Therefore, above discussion suggests that there is a plenty of room for methodological improvements in the study by incorporating more data types, other dimensionality reduction techniques and/or model selection criteria.

6 Conclusion and Future Work

In this paper, high-dimensional genomic (sCNAs) and transcriptomic (mRNA and miRNA) data from TCGA KIRC dataset are integrated, to predict survivals and identify significant genes whose expression levels affect the

clinical outcome. Incorporating interactions among different genomic data types and using the dimensionality reduction techniques helps not only to reduce the large computations but also leads to an effective way of making predictions and identifying significant variables from the original featured dataset. The proposed method used two-dimension reduction techniques, PCA and SPCA to generate the latent features that were used to build the predictive models and carry out the variable selection. Among the methods, SPCA with a lower MSEP of 2.07 than MSEP of 2.11 with PCA performs better for the prediction on including interactions. However, both the models help in achieving improved and convenient model fitting in the BMA procedure with lesser computations and also included interaction effects for identifying potential markers in the integrative study of KIRC dataset. As future work, the proposed method can be extended to include more biological data types like DNA methylation and their interactions that may improve the predictive power of the model. In addition, we are planning to use other dimensionality reduction techniques that may lower the prediction error and lead to more sophisticated modeling for the proposed method.

Compliance with Ethical Standards

Conflict of Interest No competing financial or other interests exist.

References

1. Cairns P (2011) Renal cell carcinoma. *Cancer Biomark* 9:461–473. <https://doi.org/10.3233/cbm-2011-0176>
2. Vera-Badillo FE, Templeton AJ, Duran I, Ocana A, De Gouveia P, Aneja P, Knox JJ, Tannock IF, Escudier B, Amir E (2015) Systemic therapy for non-clear cell renal cell carcinomas: a systematic review and meta-analysis. *Eur Urol* 67:740–749. <https://doi.org/10.1016/j.eururo.2014.05.010>
3. Kashyap D, Tuli HS, Sak K, Garg VK, Goel N, Punia S, Chaudhary A (2019) Role of reactive oxygen species in cancer progression. *Curr Pharmacol Rep* 5:79–86. <https://doi.org/10.1007/s40495-019-00171-y>
4. López JI (2013) Renal tumors with clear cells. A review. *Pathol Res Pract* 209:137–146. <https://doi.org/10.1016/j.prp.2013.01.007>
5. Dondeti VR, Wubbenhorst B, Lal P, Gordan JD, Andrea DK, Attiyeh EF, Simon MC, Nathanson KL (2012) Integrative genomic analyses of sporadic clear cell renal cell carcinoma define disease subtypes and potential new therapeutic targets. *Cancer Res* 72:112–121. <https://doi.org/10.1158/0008-5472.can-11-1698>
6. Sato Y, Yoshizato T, Shiraishi Y, Maekawa S, Okuno Y, Kamura T, Shimamura T, Sato-Otsubo A, Nagae G, Suzuki H, Nagata Y, Yoshida K, Kon A, Suzuki Y, Chiba K, Tanaka H, Niida A, Fujimoto A, Tsunoda T, Morikawa T, Maeda D, Kume H, Sugano S, Fukayama M, Aburatani H, Sanada M, Miyano S, Homma Y, Ogawa S (2013) Integrated molecular analysis of clear-cell renal cell carcinoma. *Nat Genet* 45:860–867. <https://doi.org/10.1038/ng.2699>
7. Chen J, Zhang D, Zhang W, Tang Y, Yan W, Guo L, Shen B (2013) Clear cell renal cell carcinoma associated microRNA expression

- signatures identified by an integrated bioinformatics analysis. *J Transl Med* 11:169. <https://doi.org/10.1186/1479-5876-11-169>
8. Chekouo T, Stingo FC, Doecke JD, Do K-A (2015) miRNA-target gene regulatory networks: a bayesian integrative approach to biomarker selection with application to kidney cancer. *Biometrics* 71:428–438. <https://doi.org/10.1111/biom.12266>
 9. Butz H, Szabó PM, Nofech-Mozes R, Rotondo F, Kovacs K, Mirham L, Girgis H, Boles D, Patocs A, Yousef GM (2014) Integrative bioinformatics analysis reveals new prognostic biomarkers of clear cell renal cell carcinoma. *Clin Chem* 60:1314–1326. <https://doi.org/10.1373/clinchem.2014.225854>
 10. Bluyssen HAR, Wesoly J, Rydzanicz M, Wrzesin T (2013) Genomics and epigenomics of clear cell renal cell carcinoma: recent developments and potential applications. *Cancer Lett* 341:111–126. <https://doi.org/10.1016/j.canlet.2013.08.006>
 11. Gregory KB, Momin AA, Coombes KR, Baladandayuthapani V (2014) Latent feature decompositions for integrative analysis of multi-platform genomic data. *IEEE/ACM Trans Comput Biol Bioinform* 11:984–994. <https://doi.org/10.1109/TCBB.2014.2325035>
 12. Ma S, Dai Y (2011) Principal component analysis based methods in bioinformatics studies. *Brief Bioinform* 12:714–722. <https://doi.org/10.1093/bib/bbq090>
 13. Sotiropoulos C, Wirapati P, Loi S, Harris A, Fox S, Smeds J, Nordgren H, Farmer P, Praz V, Haibe-kains B, Desmedt C, Larsimont D, Cardoso F, Peterse H, Nuyten D, Buyse M, De Van Vijver MJ, Bergh J, Piccart M, Delorenzi M (2006) Gene expression profiling in breast cancer: understanding the molecular basis of histologic grade to improve prognosis. *J Natl Cancer Inst* 98:262–272. <https://doi.org/10.1093/jnci/djj052>
 14. Yanaihara N, Caplen N, Bowman E, Seike M, Kumamoto K, Yi M, Stephens RM, Okamoto A, Yokota J, Tanaka T, Calin GA, Liu C, Croce CM, Harris CC (2006) Unique microRNA molecular profiles in lung cancer diagnosis and prognosis. *Cancer Cell* 9:189–198. <https://doi.org/10.1016/j.ccr.2006.01.025>
 15. Engler DA, Gupta S, Growdon WB, Drapkin RI, Nitta M, Petra A, Allred SF, Gross J, Deavers MT, Kuo W, Karlan BY, Bo R, Orsulic S, Gershenson DM, Birrer MJ, Gray JW, Mohapatra G (2012) Genome wide DNA copy number analysis of serous type ovarian carcinomas identifies genetic markers predictive of clinical outcome. *PLoS One* 7(2):e30996. <https://doi.org/10.1371/journal.pone.0030996>
 16. Gligorijević V, Pržulj N (2015) Methods for biological data integration: perspectives and challenges. *J R Soc Interface*. <https://doi.org/10.1098/rsif.2015.0571>
 17. Wang W, Baladandayuthapani V, Morris JS, Broom BM, Manayam G, Do K-A (2012) iBAG: integrative Bayesian analysis of high-dimensional multiplatform genomics data. *Bioinformatics* 29:149–159. <https://doi.org/10.1093/bioinformatics/bts655>
 18. Network CGAR (2013) Comprehensive molecular characterization of clear cell renal cell carcinoma. *Nature* 499:43–49. <https://doi.org/10.1038/nature12222.COMPREHENSIVE>
 19. Yuan Y, Van Allen EM, Omberg L, Wagle N, Amin-Mansour A, Sokolov A, Byers LA, Xu Y, Hess KR, Diao L, Han L, Huang X, Lawrence MS, Weinstein JN, Stuart JM, Mills GB, Garraway LA, Margolin AA, Getz G, Liang H (2014) Assessing the clinical utility of cancer genomic and proteomic data across tumor types. *Nat Biotechnol* 32:644–652. <https://doi.org/10.1038/nbt.2940>
 20. Kashyap D, Tuli HS, Garg VK, Goel N, Bishayee A (2018) Oncogenic and tumor-suppressive roles of microRNAs with special reference to apoptosis: molecular mechanisms and therapeutic potential. *Mol Diagn Ther* 22:179–201. <https://doi.org/10.1007/s40291-018-0316-1>
 21. Abualigah LM, Khader AT, Hanandeh ES (2018) A combination of objective functions and hybrid Krill herd algorithm for text document clustering analysis. *Eng Appl Artif Intell* 73:111–125. <https://doi.org/10.1016/j.engappai.2018.05.003>
 22. Abualigah LM, Khader AT, Hanandeh ES (2018) Hybrid clustering analysis using improved krill herd algorithm. *Appl Intell* 48:4047–4071. <https://doi.org/10.1007/s10489-018-1190-6>
 23. Abualigah LM, Khader AT (2017) Unsupervised text feature selection technique based on hybrid particle swarm optimization algorithm with genetic operators for the text clustering. *J Supercomput* 73:4773–4795. <https://doi.org/10.1007/s11227-017-2046-2>
 24. Qasim Abualigah LM, Hanandeh ES (2015) Applying genetic algorithms to information retrieval using vector space model. *Int J Comput Sci Eng Appl* 5:19–28. <https://doi.org/10.5121/ijcse.2015.5102>
 25. Abualigah LMQ (2019) Feature selection and enhanced krill herd algorithm for text document clustering. Springer International Publishing, Switzerland. <https://doi.org/10.1007/978-3-030-10674-4>
 26. Jolliffe IT (1986) Choosing a subset of principal components or variables. *Principal component analysis: Springer Series in Statistics*, 2nd edn. Springer, New York, pp 111–149. <https://doi.org/10.1007/978-1-4757-1904-8>
 27. Hsu Y-L, Huang P-Y, Chen D-T (2014) Sparse principal component analysis in cancer research. *Transl Cancer Res* 3:182–190. <https://doi.org/10.3978/j.issn.2218-676X.2014.05.06>
 28. Raftery AE, Madigan D, Hoeting JA (1997) Bayesian model averaging for linear regression models. *J Am Stat Assoc* 92:179–191. <https://doi.org/10.1080/01621459.1997.10473615>
 29. Goel N, Karir P, Garg VK (2017) Role of DNA methylation in human age prediction. *Mech Ageing Dev* 166:33–41. <https://doi.org/10.1016/j.MAD.2017.08.012>
 30. Witten DM, Tibshirani R, Hastie T (2009) A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics* 10:515–534. <https://doi.org/10.1093/biostatistics/kxp008>
 31. Raftery AE, Painter IS, Volinsky CT (2005) BMA: an R Package for Bayesian model averaging. *R News* 5:2–8
 32. Nyhan MJ, Sullivan GCO, Mckenna SL (2008) Role of the VHL (von Hippel-Lindau) gene in renal cancer: a multifunctional tumour suppressor. *Biochem Soc Trans* 36:472–478. <https://doi.org/10.1042/BST0360472>
 33. Guinney J, Wang T, Laajala TD, Winner KK, Bare JC, Neto EC, Khan SA, Peddinti G, Airola A, Pahikkala T, Mirtti T, Yu T, Bot BM, Shen L, Abdallah K, Norman T, Friend S, Stolovitzky G, Soule H, Sweeney CJ, Ryan CJ, Scher HI, Sartor O, Xie Y, Aittokallio T, Zhou FL, Costello JC (2017) Prediction of overall survival for patients with metastatic castration-resistant prostate cancer: development of a prognostic model through a crowd-sourced challenge with open clinical trial data. *Lancet Oncol* 18(1):132–142. [https://doi.org/10.1016/s1470-2045\(16\)30560-5](https://doi.org/10.1016/s1470-2045(16)30560-5)