12-1-2020

CODC: a Copula-based model to identify differential coexpression

Sumanta Ray
Centrum Wiskunde & Informatica

Snehalika Lall
Indian Statistical Institute, Kolkata

Sanghamitra Bandyopadhyay
Indian Statistical Institute, Kolkata

npj | Systems Biology and Applications

www.nature.com/npjsba

ARTICLE OPEN

CODC: a Copula-based model to identify differential coexpression

Sumanta Ray[1,3]✉, Snehalika Lall[2,3] and Sanghamitra Bandyopadhyay[2]✉

Differential coexpression has recently emerged as a new way to establish a fundamental difference in expression pattern among a group of genes between two populations. Earlier methods used some scoring techniques to detect changes in correlation patterns of a gene pair in two conditions. However, modeling differential coexpression by means of finding differences in the dependence structure of the gene pair has hitherto not been carried out. We exploit a copula-based framework to model differential coexpression between gene pairs in two different conditions. The Copula is used to model the dependency between expression profiles of a gene pair. For a gene pair, the distance between two joint distributions produced by copula is served as differential coexpression. We used five pan-cancer TCGA RNA-Seq data to evaluate the model that outperforms the existing state of the art. Moreover, the proposed model can detect a mild change in the coexpression pattern across two conditions. For noisy expression data, the proposed method performs well because of the popular scale-invariant property of copula. In addition, we have identified differentially coexpressed modules by applying hierarchical clustering on the distance matrix. The identified modules are analyzed through Gene Ontology terms and KEGG pathway enrichment analysis.

npj Systems Biology and Applications (2020)6:20 ; https://doi.org/10.1038/s41540-020-0137-9

INTRODUCTION

Microarray-based gene coexpression analysis has been demonstrated as an emerging field that offers opportunities to the researcher to discover coregulation pattern among gene expression profiles. Genes with similar transcriptomal expression are more likely to be regulated by the same process. Coexpression analysis seeks to identify genes with similar expression patterns, which can be believed to associate with the common biological process[1–3]. Recent approaches are interested to find the differences between coexpression pattern of genes in two different conditions[4,5]. This is essential to get a more informative picture of the differential regulation pattern of genes under two phenotype conditions. Identifying the difference in coexpression patterns, which is commonly known as differential coexpression, is no doubt a challenging task in computational biology. Several computational studies exist for identifying a change in gene coexpression patterns across normal and disease states[6–9]. Finding differentially coexpressed (DC) gene pairs, gene clusters, and dysregulated pathways between normal and disease states is most common[6,10–13]. Another way for identifying DC gene modules is to find gene cluster in one condition, and test whether these clusters show a change in coexpression patterns in another condition significantly[8,10].

For example, CoXpress[10] utilizes hierarchical clustering to model the relationship between genes. The modules are identified by cutting the dendrogram at some specified level. It used a resampling technique to validate the modules coexpressed in one condition but not in the other. Another approach called DiffCoex[11] utilized a statistical framework to identify DC modules. DiffCoex proposed a score to quantify differential coexpression between gene pairs and transform this into dissimilarity measures to use in clustering. A popularly used tool WGCNA (Weighted Gene Coexpression Network Analysis) is exploited to group genes into DC clusters[14]. Another method called DICER (Differential Correlation in Expression for meta-module Recovery)[15] also identifies gene sets whose correlation patterns differ between disease and control samples. Dicer not only identifies the differentially coexpressed module, but it goes one step beyond and identifies metamodules or a class of modules where a significant change in coexpression patterns is observed between modules, while the same patterns exist within each module.

In another approach, Ray and Maulik[16] proposed a multi-objective framework called DiffCoMO to detect differential coexpression between two stages of HIV-1 disease progression. Here, the algorithm operates on two objective functions that simultaneously optimize the distances between two correlation matrices obtained from two microarray data of HIV-infected individuals.

Most of the methods proposed some scoring technique to capture the differential coexpression pattern and utilized some searching algorithm to optimize it. Here, we have proposed CODC Copula-based model to identify **D**ifferential **C**oexpression of genes under two different conditions. Copula[17,18] produces a multivariate probability distribution from multiple uniform marginal distribution. It was extensively used in high-dimensional data applications. In the proposed method, first, a pairwise dependency between gene expression profile is modeled using an empirical copula. As the marginals are unknown, so we used empirical copula to model the joint distribution between each pair of gene expression profiles. To investigate the difference in coexpression pattern of a gene pair across two conditions, we compute a statistical distance between the joint distributions. We hypothesized that the distance between two joint distributions can model the differential coexpression of a gene pair between two conditions. To investigate this fact, we have performed a simulation study that provides the correctness of our method. We have also validated the proposed method by applying it in real-life datasets. For this, we have used five pan-cancer RNA-Seq

[1]Centrum Wiskunde & Informatica, Life Sciences & Health, 1098 XG Amsterdam, The Netherlands. [2]Machine Intelligence Unit, Indian Statistical Institute, Kolkata, India. [3]These authors contributed equally: Sumanta Ray, Snehalika Lall. ✉email: Sumanta.Ray@cwi.nl; sanghami@gmail.com

data from TCGA: Breast-invasive carcinoma [BRCA], Head and Neck squamous carcinoma [HNSC], Liver hepatocellular carcinoma [LIHC], Thyroid carcinoma [THCA] and Lung adenocarcinoma [LUAD], which are publicly available in the TCGA data portal (https://tcga-data.nci.nih.gov/docs/publications/tcga/).

RESULTS

Dataset preparation

We have evaluated the performance of the proposed method in five RNA-seq expression data downloaded from the TCGA data portal (https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga). We have downloaded a matched pair of tumor and normal samples from five pan-cancer datasets: Breast-invasive carcinoma (BRCA, #samples = 112), head and neck squamous cell carcinoma (HNSC, #samples = 41), liver hepatocellular carcinoma (LIHC, #samples = 50), thyroid carcinoma (THCA, #samples = 59), and Lung Adenocarcinoma (LUAD, #samples = 58). For preprocessing the dataset, we first take those genes that have raw read count greater than two in at least four cells. The filtered data matrix is then normalized by dividing each UMI (Unique Molecular Identifiers) count by the total UMI counts in each cell, and subsequently, these scaled counts are multiplied by the median of the total UMI counts across cells[19]. The top 2000 most variable genes were selected based on their relative dispersion (variance/mean) with respect to the expected dispersion across genes with similar average expression. Transcriptional responses of the resulting genes were represented by the log2(fold change) of gene expression levels from paired tumor and normal samples. A brief description of the datasets used in this paper is summarized in Table 1. Figure 1a, b represents box and violin plot of the average expression value of samples for each dataset.

Table 1. Tumor types and number of TCGA RNA-seq samples used in the analysis.

Sl No.	Cancer type	# matched pair samples
1	Breast-invasive carcinoma (BRCA)	112
2	Head and neck squamous cell carcinoma (HNSC)	51
3	Liver hepatocellular carcinoma (LIHC)	50
4	Thyroid carcinoma (THCA)	59
5	Lung adenocarcinoma (LUAD)	58

Detection of DC gene pair

Differential coexpression between a gene pair is modeled as a statistical distance between the joint distributions of their expression profiles in a paired sample. Joint distribution is computed by using empirical copula that takes the expression profile of a gene as marginals in normal and tumor samples. The K–S distance, computed between the joint distribution, served as differential coexpression score between a gene pair. The score for a gene pair (g_i, g_j) can be formulated as DC_Copula(g_i, g_j) = KS_dist(e.c$(g_i^{\text{tumor}}, g_j^{\text{tumor}})$, e.c$(g_i^{\text{normal}}, g_j^{\text{normal}})$), where KS-dist represents Kolmogorov–Smirnov (K–S) distance between two joint probability distributions, e.c represents empirical copula, and g_i^P represents the expression profile of gene g_i at phenotype P. For each RNA-seq data, we have computed the DC_Copula matrix, from which we identify differentially coexpressed gene pairs.

To know how the magnitude of differential coexpression is changing with the score, we plot the distribution of correlation values of gene pairs with their scores in Fig. 2. The figure also shows the number of gene pairs having positive and negative correlations in each stage (normal/tumor). It can be noticed from the figure that high scores produce differentially coexpressed gene pairs having a higher positive and negative correlation. We collected the gene pairs having the score greater than 0.56 and plot the correlations values in Fig. 3. This figure shows plots of all gene pairs having a positive correlation in normal and the negative correlation in tumor (shown in panel a) and vice versa (shown in panel b). The density of the correlation values is shown in panels c and d for each case. In Fig. 4, we create a visualization of top differentially coexpressed gene pairs in BRCA data, which show a strong positive correlation in tumor stage and negative correlation in normal stage. The figure shows a heatmap of a binary matrix constructed from the expression data of those gene pairs in tumor and normal stages. The expression values showing the same pattern for a gene pair are assumed 1, while 0 represents a nonmatching pattern. From the figure, it is quite understandable

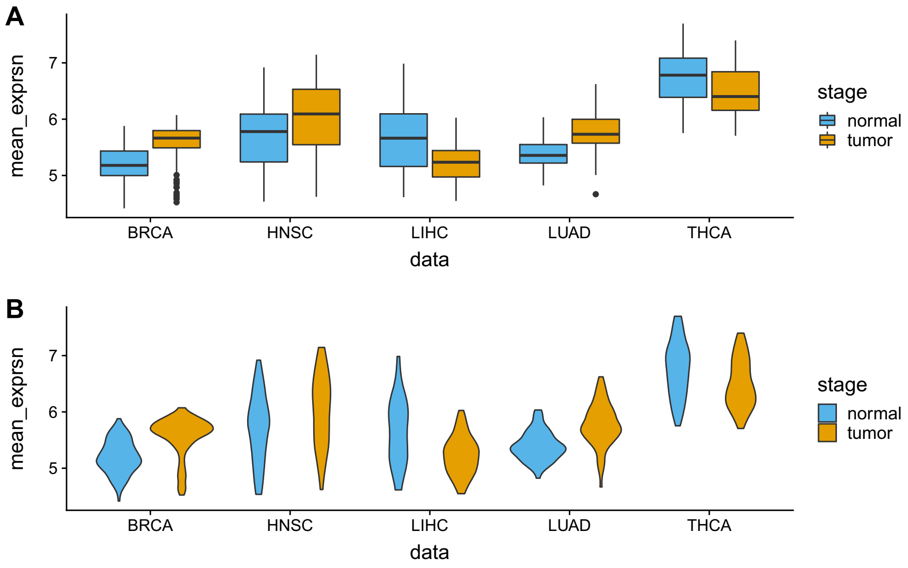

Fig. 1 Description of TCGA data used in the analysis: Panel-A and -B describes box and violin plots of mean expression values of the used datasets.

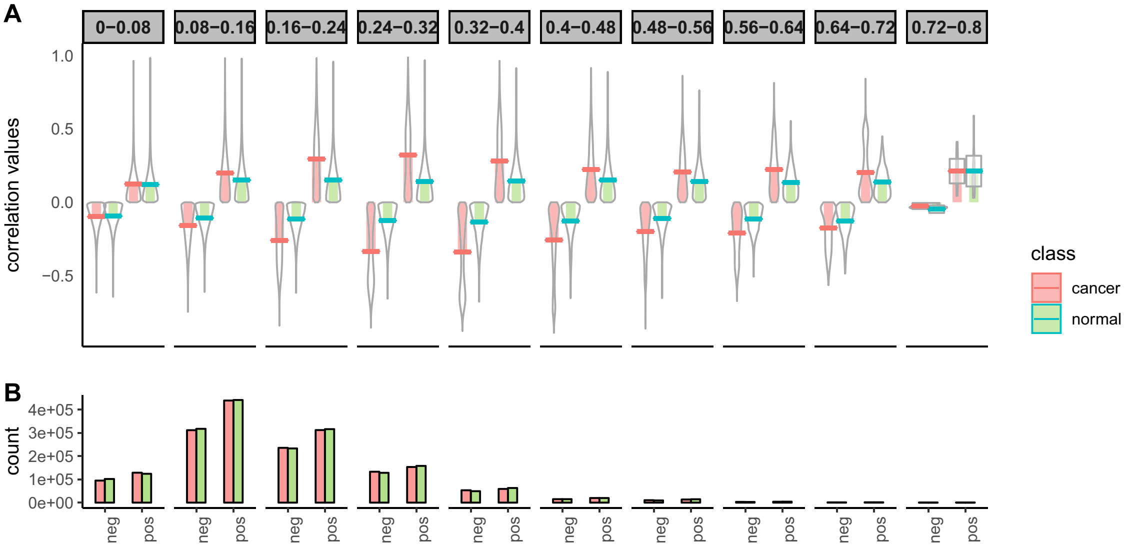

Fig. 2 The figure shows the distribution of correlation values in normal and cancer samples of BRCA data with the DC_Copula score. **a** shows the distribution for different DC_Copula scores. Here, four pirate plots are shown in each facet, two for positive and two for negative correlations. The violins in each facet represent the distribution of positive and negative correlations of gene pairs in normal and cancer samples. **b** shows a bar plot representing the number of positive and negatively correlated gene pairs in normal and cancer samples in each facet.

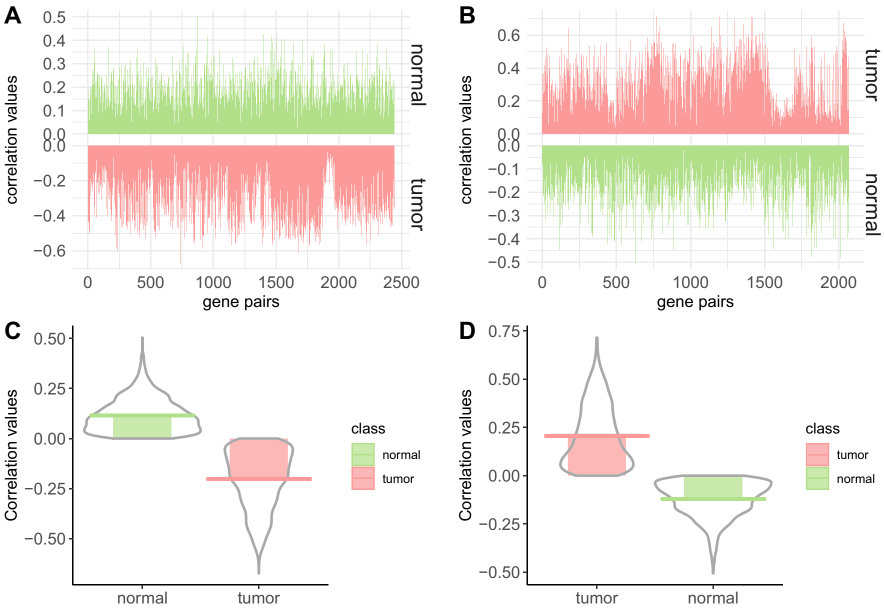

Fig. 3 The figure shows visualizations of gene pairs having DC_copula score greater than 0.56. **a**, **b** show the visualization of correlation values of gene pairs having a positive correlation in normal and negative correlation in tumor and vice versa, respectively. **c**, **d** represent the distribution of correlation values according to **a**, **b**, respectively.

that most of the entry in the normal stage is 0 (nonmatch) while in tumor stage, it is 1 (match). For other datasets, the plots are shown in Supplementary Fig. 2.

Stability performance of CODC

To prove the stability of CODC, we have performed the following analysis:

First, we add Gaussian noise to the original expression data of normal and cancer samples to transform these into noisy datasets. We use the rnorm function of R to create normally distributed noise with mean 0 and standard deviation 1, and we add this into the input data. We have utilized BRCA data for this analysis.

First, we compute the K–S distance and then obtain DC_Copula matrix for both original and noisy datasets. Let us denote these two matrices as D and D'.

The usual way is to pick a threshold t for D (or D') and extract the gene pair (i,j) for which $D(i,j)(orD'(i,j)) \geq t$. First, we set t as the maximum of D and D', and then decrease it continuously to extract the gene pairs. For each t, we observe the number of common gene pairs obtained from D and D'. Figure 5 shows the proportion of common genes selected from D and D' for different threshold selection and different level of noise. Theoretically, CODC produces D with scores no more than D' (see the above section for details). So, it is quite obvious that the number of common genes increased with a lower threshold value. From the

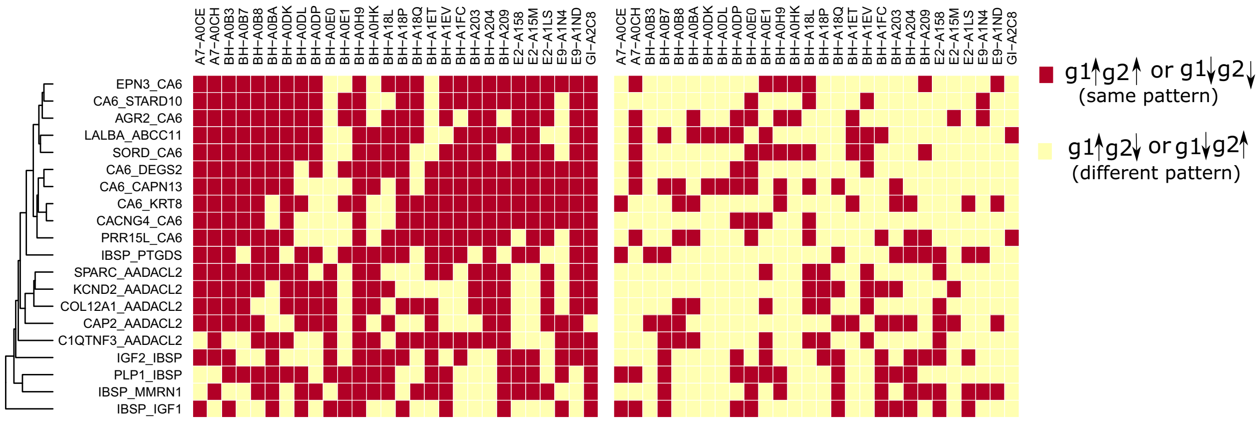

Fig. 4 The figure shows a heatmap representation of a binary matrix constructed from the expression matrix of top differentially coexpressed gene pairs in normal and tumor stages. Expression values of a gene pair showing the same pattern are indicated as 1, and showing a different pattern is indicated as 0 in the matrix. The columns represent differentially coexpressed gene pairs, while rows are the samples of BRCA data.

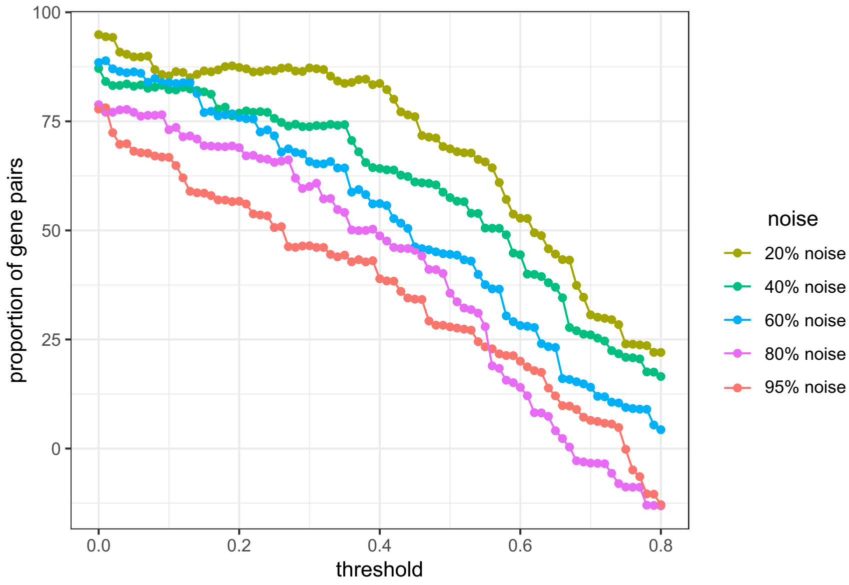

Fig. 5 Performance of CODC in noisy datasets.

property (see the above section for details), it can be noticed that the scores in *D* get preserved in *D*′. So, it is expected that obtained gene pairs from original data are also preserved in noisy data. Figure 5 shows the evidence for this case. As can be seen from the figure that even the noise label is 80%, for threshold value above 0.25, more than 55% of the gene pairs are common between noisy and original datasets.

Detection of differentially coexpressed modules

Detection of DC gene modules is performed by using hierarchical clustering on the DC matrix. Here, the differential coexpression score obtained from each gene pair is treated as the similarity measure between genes. The distance between a gene pair is formulated as dist_copula(g_i, g_j) = 1 − DC_Copula(g_i, g_j). For each dataset, modules are extracted using average linkage hierarchical clustering by using the dist_copula as a dissimilarity measure between a pair of a gene. For BRCA and HNSC data, we have identified 15 modules, for LIHC data 14 modules, for LUAD 21 modules, and for THCA 22 modules are identified. For studying the relationship between the modules, we have identified module eigengene networks for each dataset. According to ref. [14] module, eigengene represents a summary of the module expression profiles. Here, module eigengene network signifies coexpression relationship among the identified modules in two stages. We create visualizations of the module eigengene network for normal and tumor stages in Fig. 6. The upper triangular portion of the correlation matrix represents the correlation between module eigengenes for normal samples, whereas the lower triangular portion represents the same for tumor samples. This figure shows the heatmap for BRCA, HNSC, and LUAD datasets. It is clear from Fig. 6 that most of the modules show differential coexpression pattern in normal and tumor stages. For a differentially coexpressed module, it is expected that it shows an opposite correlation pattern in two different phenotype conditions. Here, the correlation pattern between two modules is represented as the correlation between the module eigengenes. In the heatmap of Fig. 6, we can observe that in all three datasets, the correlation pattern between most of the MEs in normal and tumor stages has the opposite direction. For example, from panel a, it can be noticed that for BRCA data, the modules have a negative correlation in the normal phase while showing a strong positive

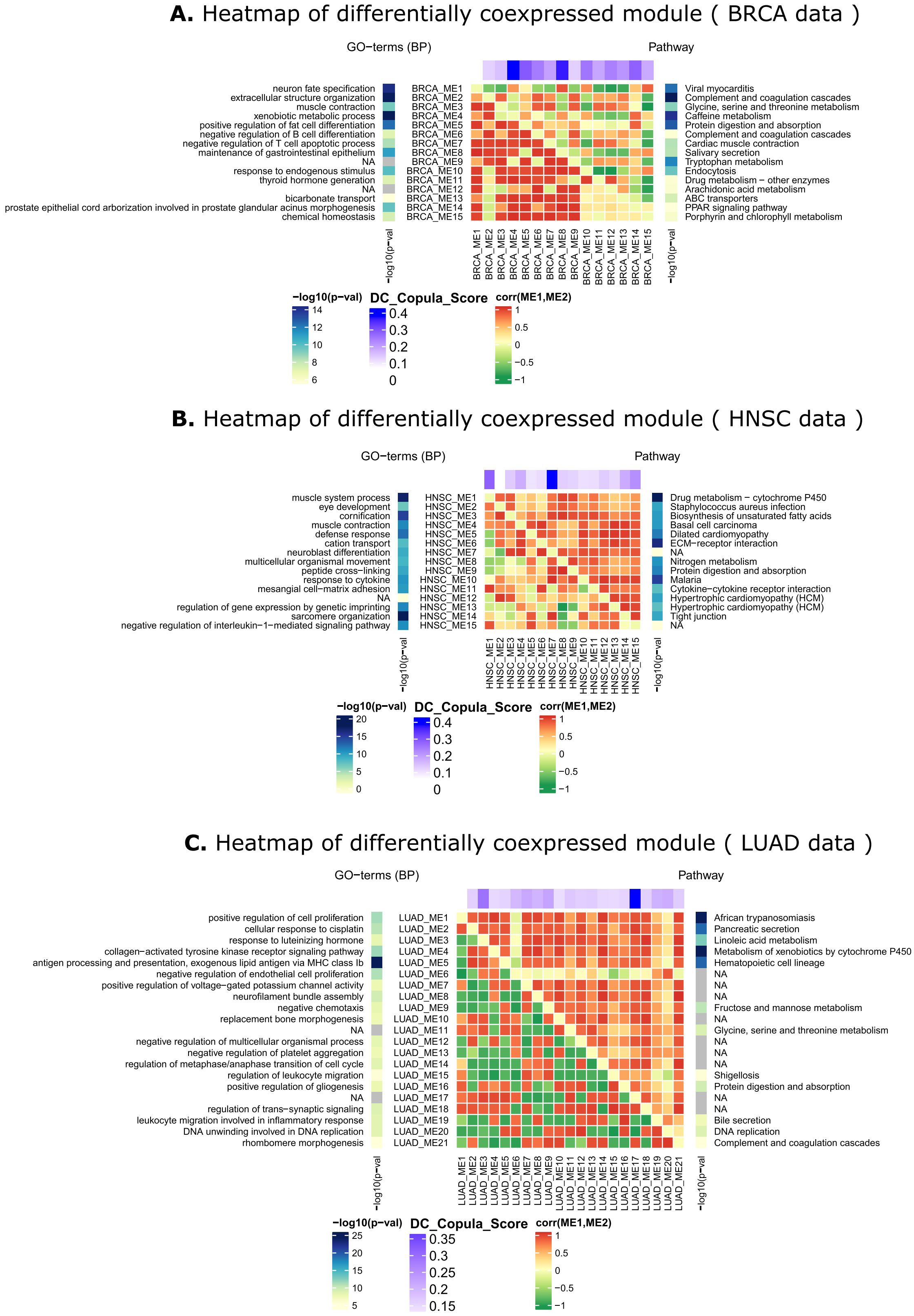

Fig. 6 Heatmap of differentially coexpressed modules. Here the heatmap is shown for module eigengenes.The upper triangular portion of the matrix represents correlations of module eigengenes in normal samples, whereas the lower triangular portion signifies the same for tumor samples. Left and right sidebar of the heatmap represents −log(p value) of significantly enriched GO terms and pathways, respectively. "NA" stands for unavailability of significant pathway or GO terms. The upper annotation bar of the heatmap shows the DC_copula score of the module. **a** Shows the heatmap for BRCA data, whereas (**b**, **c**) Show heatmap HNSC and LUAD data.

correlation in tumor phase. For the HNSC dataset, the opposite case is observed. Modules have a strong positive correlation in normal phases while having a negative correlation in tumor phase. In Supplementary Fig. 2, the visualization of all datasets is given.

Comparisons with competing methods

For comparison purpose, we have taken three competing techniques, such as Diffcoex, coXpress, and DiffCoMO, and compared them with our proposed method. All these methods are extant DC based, which look for gene modules with altered coexpression between two classes. DiffCoEx performed hierarchical clustering on the distance matrix complied from correlation matrices of two phenotype stages. CoXpress detects a correlation module in one stage and finds the alternation of the correlation pattern within the module in other classes. DiffCoMO uses the multiobjective technique to detect differential coexpression modules between two phenotype stages. We have made two approaches for comparing our proposed method with competing methods. We first compare the efficacy of these methods for detecting differential coexpressed gene pairs, and next compare the modules identified in each case. For the first case, we take the top 1000 gene pairs having high DC_Copula scores from the DC matrix, and perform classification using normal and tumor samples. The expression ratio of each DC gene pair from the expression matrix was taken and compiled a $n \times 1000$, where n represents the number of samples in each data. For the other three methods, we have also selected the same number of differentially coexpressed gene pairs for the classification task. Table 2 shows some parameters we have used for the selection of the gene pairs. For CoXpress, first, we have used “cluster.gene” and “cuttree” function with default parameters provided in the R package of CoXpress to get the gene clusters according to the similarity of their expression profiles. These groups are then examined by the coXpress R function to identify the differentially coexpressed modules by comparing with the t statistics generated by randomly resampling the dataset 10,000 times for each group. We have taken the top 10 modules based on the robustness parameter, which tells the number of times that the group was differentially coexpressed in 1000 randomly resampled data. Now, we have selected 1000 gene pairs randomly from those modules. For the DiffCoEx method, we collected the DC gene pairs before partitioning them in modules. We used the code available in the supplementary file of the original paper of DiffCoEx, to get the distance score matrix that is used in the hierarchical clustering for module detection. We sort the score of the distance matrix and pick the top 1000 gene pairs based on the scores. For DiffCoMO, we use the default parameters to cluster the network to obtained differentially coexpressed modules. As it utilized the multiobjective method, all the Pareto optimal solutions of the final generation are taken as selected modules. We then choose 1000 gene pairs randomly from the identified modules. Classification is performed by treating normal and tumor samples as class labels. A toy example of the comparison is shown in Fig. 7. Note that all these methods are meant for differentially coexpressed module detection. So, for comparison, we collected the DC gene pairs

Table 2. Table shows the different parameters/threshold we have used for selecting differentially coexpressed gene pairs for other methods.

Method	No of gene pairs selected	Parameters used
CoXpress	1000	Used cluster.gene and cutree function with corr.coef threshold 0.6 and cutting height of hierarchical tree $h = 0.4$. Robustness parameter threshold $= 800$
DiffCoEX	1000	Used Spearman correlation to compute adjacency matrix for each phenotype condition. Use default soft-thresholding parameter $\beta = 6$ for computation of distance score matrix
DiffCoMO	1000	No of modules (population size) is taken as 50 and the number of generation is 200

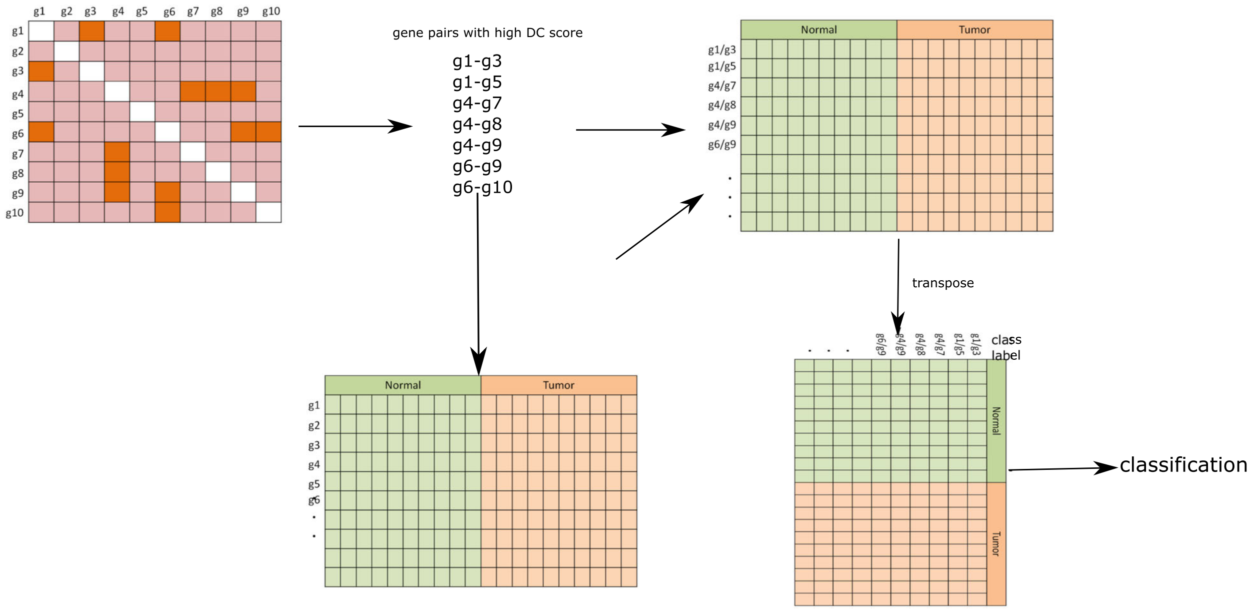

Fig. 7 A toy example of performing classification on differentially coexpressed gene pairs. From the DC matrix, the top gene pairs are selected based on DC_copula score. The expression ratio is computed for each gene pair for normal and tumor samples. The final matrix is then transposed and subsequently, classification is performed using normal and tumor samples as class labels.

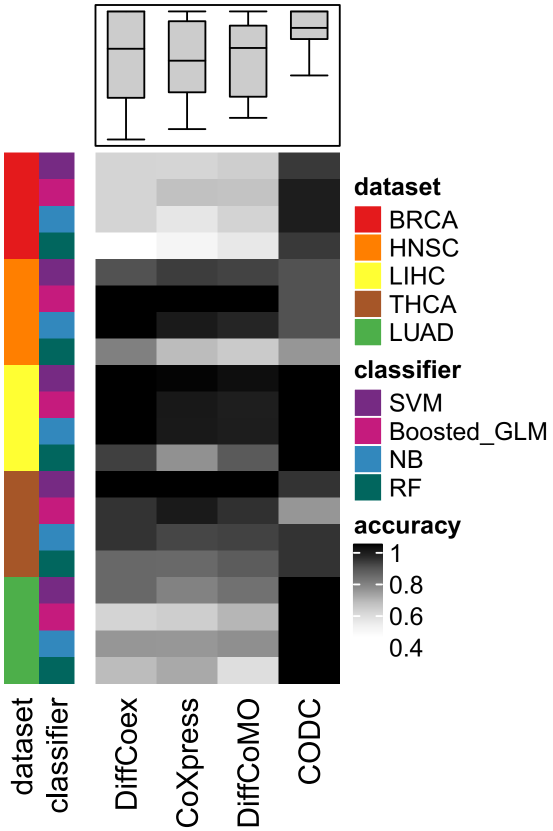

Fig. 8 Performance comparison with state-of-the-art:Comparison of classification accuracy for five datasets with four classifiers Boosted -GLM, Naive Bayes, Random Forest, and SVM.

before partitioning them in modules. We train four classifiers Boosted GLM, Naive Bayes, Random Forest, and SVM with the data and take the classification accuracy. The classification results are shown in Fig. 8. It can be noticed from the figure that for most of the dataset, the proposed method achieved high accuracy compared with the other methods.

To assess the performance of all the methods for detecting differential coexpression modules, we check the distribution of the correlation score of gene pairs within top modules in normal and tumor samples. Extant methods do a comparison by computing the absolute change in correlation value between a pair of a gene within a module. The problem for this type of comparison is that the score ignores a small change in differential coexpression. It also fails to consider the gene pair having a low score and correlation of opposite sign in two conditions. For example, it emphasized the gene pair with correlation value 0.2 in normal and 0.7 in the tumor (here the score is 0.5) rather than the gene pair whose correlation value is −0.2 in normal and 0.2 in the tumor (here the score is 0.4). So, for comparison, it is required to investigate the number of gene pairs having correlation values of an opposite sign over −1 to +1. So, for all identified modules, we calculate the correlation score of each gene pair in two different samples (normal and cancer) and plot the frequency polygon in Fig. 9. To investigate whether the gene pairs within the modules show a good balance in positive and negative correlations, we have computed the correlation score for all the identified modules of DiffCoMO, DiffCoEx, and CoXpress. Figure 9 shows the comparisons of the correlation scores. It is noticed from the figure that gene pairs within the identified modules of the proposed method show good balance in positive and negative correlation values. DiffCoMO and DiffCoEX have also achieved the same, whereas most of the gene pairs within the coXpress modules shifted toward positive correlation in both tumor and normal samples. In Fig. 9a, we have also shown the boxplot of the correlation values obtained from different methods. As can be seen from the figure, the median line of correlation values for the proposed method is nearer to 0, which signifies good distribution of correlation scores in normal and tumor samples over −1 to +1. Thus, the proposed method can able to detect differentially coexpressed gene pairs having correlation values well distributed between −1 and +1.

To compare CODC with the other methods, we tested its performance in a simulated dataset also. To create the simulated data, we have used HNSC RNA-seq expression dataset. We create the simulated data as follows:

1. For each gene g_i in the normal sample, we simulated the expression profile of sample s_j as $X_{ij} = \mathcal{N}(m_i, \sigma_i^2)$ where m_i represents mean expression value of gene g_i across all 51 HNSC normal samples, and σ^2 represents their variance.
2. Similarly, we simulated the expression profile of tumor sample s'_j as $Y_{i,j} = \mathcal{N}(m'_i, \sigma_i'^2)$, where m'_i is mean of the expression value of gene g'_i and $\sigma_i'^2$ is the variance. Here, we assume that a gene pair is truly differentially coexpressed if the following condition holds: DC_score(g_i, g_j) > 0.5 and the correlation between g_i and g_j has opposite sign in tumor and normal stage.
3. We then add different levels of Gaussian noise to the Y_{ij} to create different noisy expression data (Y'_{ij}) from the simulated tumor samples. We use rnorm function of R to produce normally distributed noise with mean 0 and standard deviation 1.

Now, we compute differentially coexpressed gene pairs between simulated normal (X_{ij}) and noisy sample (Y'_{ij}) and compare them with the underlying true differentially coexpressed gene pairs. We compute the proportion of matched gene pairs and plot the results against all the different noise levels in Fig. 10. We have done this analysis for all competing methods. To select the differentially coexpressed gene pairs from simulated, normal, and noisy tumor samples, we use DC_Copula threshold as 0.6. For other competing methods, we use the threshold and other parameters, same as the previous analysis, which are provided in Table 2.

Pathway analysis

To compare functional enrichment of identified modules, we have utilized KEGG pathway enrichment analysis. We defined enrichment score of a pathway in a module as $\text{pathway\_score} = \frac{n}{m},$ where n is the fraction of the pathway genes in the module and m is the fraction of the pathway genes in the dataset. We compare the pathway score for the modules identified for DiffCoEx, DiffCoMO, CoXpress, and the proposed method. For comparison purpose, we have utilized the HNSC dataset. For identifying modules in other competing methods, we have utilized the the parameters provided in Table 2. For coXpress, we take the modules with robustness parameter value greater than 760, which produces 19 clusters. For DiffcoEx, we have used the default parameters for cutreeDynamic function (cutHeight = 0.996, minClusterSize = 20), which produces 42 clusters. Figure 11 shows the result. The Y axis represents CCDF (complementary cumulative distribution function), which represents how often the number of modules is above a certain value of pathway score. It is clear from the figure that more modules for the proposed method achieved a high pathway score compared with other competing methods. In Fig. 6, we have shown heatmaps of differentially coexpressed modules for BRCA, HNSC, and for LUAD data. The heatmap also provides pathways and GO terms significantly enriched with the modules. The p value for KEGG pathway and GO enrichment is computed by using the hypergeometric test with

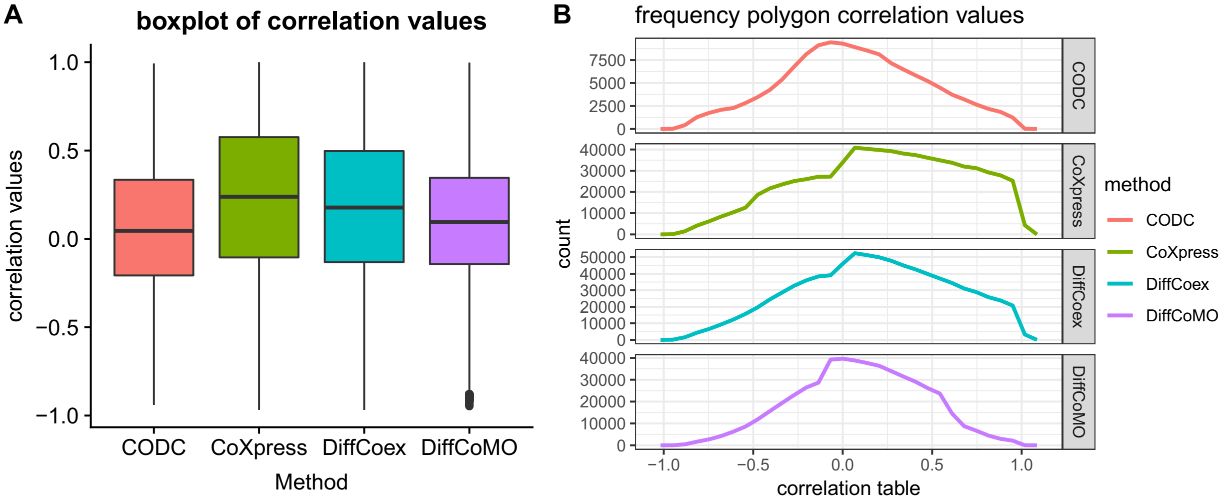

Fig. 9 Distribution of correlation scores of the gene pairs in normal and tumor stage. Each facet shows the distribution for different methods.

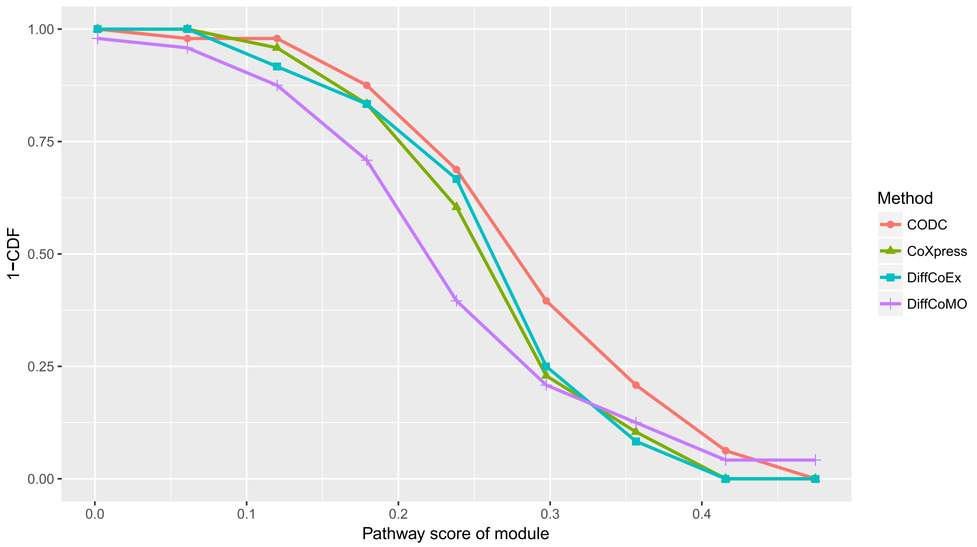

Fig. 10 Figure shows the percentage of matching between identified gene pairs with true differentially coexpressed gene pairs of simulated data. Results are shown for different noise levels and for different methods.

0.05 FDR corrections. We have utilized GOstats, kegg.db, and GO.db R package for that. It can be seen from Fig. 6a that some pathways such as "Complement and coagulation cascades", "Proximal tubule bicarbonate reclamation", "Caffeine metabolism", "Protein digestion and absorption", "Tryptophan metabolism", and "ABC transporters", are strongly associated with the identified modules of BRCA. "Tryptophan metabolism" has eminent evidence to link with malignant progression in breast cancer[20]. In ref. [21], the association between ABC transporters with breast carcinoma has been established. From panel b, it can be seen that drug metabolism–cytochrome P450[22] ECM–receptor interaction[23], "Nitrogen metabolism", and "Protein digestion and absorption" are significantly associated with the modules of HNSC data. Some pathways such as "Drug metabolism–cytochrome P450" and "ECM–receptor interaction" have strong evidence associated with the head and neck squamous cell carcinomas[22,23]. Similarly from panel c, it can be noticed that pathways such as "Metabolism of xenobiotics by cytochrome P450", "Pancreatic secretion", and "Linoleic acid metabolism" are significantly associated with modules of LUAD data. Among them, there exists strong evidence for pathways: "Metabolism of xenobiotics by cytochrome P450"[24], "Pancreatic secretion"[25], and "Linoleic acid metabolism"[26] to be associated with lung carcinoma.

DISCUSSION

In this paper, we have proposed CODC, a copula-based model to detect differential coexpression of genes in two different samples. CODC seeks to identify the dependency between expression patterns of a gene pair in two conditions separately. The Copula is used to model the dependency in the form of two joint distributions. K–S distance between two joint distributions is treated as differential coexpression score of a gene pair. We have compared CODC with three competing methods DiffCoex, CoXpress, and DiffCoMO in five pan-cancer RNA-Seq data of TCGA. CODC's ability for delineating a minor change of coexpression in two different samples makes it unique and suitable for differential coexpression analysis. The scale-invariant property of copula inherited into CODC to make it robust against noisy expression data. It is advantageous for detecting the minor change in correlation across two different conditions, which is the most desirable feature of any differential coexpression analysis.

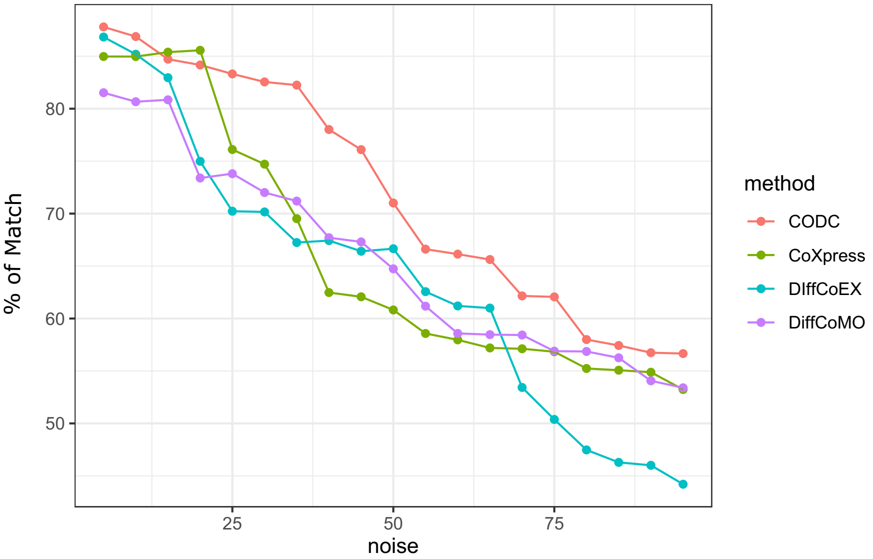

Fig. 11 Distribution of pathway score for each of the comparing methods. The figure shows the fraction of identified modules having a pathway score above a certain value.

Under the premise that the differential coexpressed genes are likely to be important biomarkers, we demonstrate that CODC identifies those that achieve better accuracy for classifying samples. Moreover, CODC goes a step further from the pairwise analysis of genes, and seeks modules wherein differential coexpression is prevalent among each pair of genes. We have also analyzed the identified modules enriched with different biological pathways, and highlighted some of these such as "Complement and coagulation cascades", "Tryptophan metabolism", "Drug metabolism–cytochrome P450", and "ECM–receptor interaction".

We have evaluated the efficacy of CODC on five different pan-cancer datasets to effectively extract differential coexpression gene pairs. Besides that, we have also compared different methods for detecting differentially coexpressed modules in those data. It is worth mentioning that CODC improves upon the competing methods. We have also proved that the scale-invariant property of copula makes CODC more robust for detecting differential coexpression in noisy data. The most important part of the DC analysis is to reveal changes in gene correlation that would not be detected by traditional DE analysis. CODC uses copula for measuring gene–gene dependency, and copula is a multivariate measure, so it can be easily extendible to use in the measurement of dependence structure of multiple genes.

METHODS

In this section, we have briefly introduced the proposed method.

Modeling differential coexpression using Copula

Differential coexpression is simply defined as the change in coexpression patterns of a gene pair across two conditions. A straightforward method to measure this is to take the absolute difference of correlations between two gene expression profiles in two conditions. For a gene pair *gene*$_i$ and *gene*$_j$, this can be formally stated as $\mathrm{DC\_Score}_{i,j}^{p1,p2} = |\mathrm{Sim}(x_i, x_j)^{p1} - \mathrm{Sim}(x_i, x_j)^{p2}|$, where p_1, p_2 are two different phenotype conditions, and x_i,x_j represent expression profile of *gene*$_i$ and *gene*$_j$, respectively. Here $\mathrm{Sim}(x_i, x_j)^p$ signifies Pearson correlation between x_i and x_j for phenotype p.

In the statistical analysis, a simple way to measure the dependence between the correlated random variable is to use copulas[27]. Copula is extensively used in high-dimensional data applications to obtain joint distributions from a random vector, easily by estimating their marginal functions.

Copulas can be described as a multivariate probability distribution function for which the marginal distribution of each variable is uniform. For a bivariate case, copula is a function: C: $[0, 1]^2 \rightarrow [0, 1]$, and can be defined as $C(x, y) = P(X \leq x, Y \leq y)$, for $0 \leq x, y \leq 1$, where X and Y are uniform random variables. Let, $Y1$ and $Y2$ be the random vectors whose marginals are uniformly distributed in [0, 1] and having marginal distribution F_{Y1} and F_{Y2}, respectively. By Sklar's theorem[28], we have the following: there exists a copula C such that $F(y_1, y_2) = C(F_{Y1}(y_1), F_{Y2}(y_2))$, for all y_1 and y_2 in the domain of F_{Y1} and F_{Y2}. In other words, there exists a bivariate copula that represents the joint distribution as a function of its marginals. For the multivariate case, the copula (C) function can be represented as

$$F_Y(y_1, y_2, \dots, y_n) = C(F_1(y_1), F_2(y_2), \dots, F_n(y_n)), \tag{1}$$

where (Y_1, Y_2, …, Y_n) be the random vectors whose marginals are $F_1(y_1)$, $F_2(y_2)$, …, $F_n(y_n)$. The converse of the theorem is also true. Any copula function with individual marginals $F_i(y_i)$ as the arguments, represents valid joint distribution function. Assuming that $F(Y_1, Y_2, \dots, Y_n)$ has nth-order partial derivatives, the relation between the joint probability-density function and the copula-density function, say c, can be obtained as

$$\begin{aligned} f(y_1, y_2, \dots, y_n) &= \tfrac{\partial^n (F(Y_1, Y_2, \dots, Y_n))}{\partial Y_1 \partial Y_2 \dots \partial Y_n} \\ &= \tfrac{\partial^n (C(F_1(y_1), F_2(y_2), \dots, F_n(y_n))}{\partial Y_1 \partial Y_2 \dots \partial Y_n} \\ &= cF_1(y_1), F_2(y_2), \dots, F_n(y_n) \prod_i f_i(y_i) \end{aligned} \tag{2}$$

where, we define

$$c(y_1, \dots, y_n) = \frac{\partial^n C(y_1, \dots, y_n)}{\partial y_1 \cdots \partial y_n}. \tag{3}$$

So, Copula is also known as joint distribution-generating function with a separate choice of marginals. Hence, different families (parametric and nonparametric) of copulas exist, which model different types of dependence structure. The example includes Farlie–Gumbel–Morgenstern n family (parametric), Archimedean Copula (parametric), Empirical Copula (nonparametric), Gaussian (parametric), and t (parametric). Empirical copulas are governed by the empirical distribution functions, which try to estimate the underlying probability distribution from given observations.

Empirical Copula is defined as follows:

Let Y_1, Y_2, …, Y_n be the random variables with marginal cumulative distribution function $F_1(y_1)$, $F_2(y_2)$, …, $F_n(y_n)$, respectively.

The empirical estimate of (F_i, $i = 1, \dots, n$), based on a sample, $\{y_{i1}, y_{i2}, \dots, y_{im}\}$, of size m is given by

$$\hat{F}_i(y) = \frac{1}{m} \sum_{j=1}^{m} 1_{\{Y_{ij} \leq y\}}, [i = 1, \dots, n] \tag{4}$$

The *Empirical Copula* of $Y_1, Y_2, \dots, Y_n$ is then defined as

$$\hat{C}(u_1, u_2, \dots, u_n) \\ = \tfrac{1}{m}\sum_{j=1}^{m} 1\{\widehat{F}_1(y_{1,j}) \le u_1, \widehat{F}_2(y_{2,j}) \le u_2, \dots, \widehat{F}_n(y_{n,j}) \le u_n\}, \qquad (5)$$

for $u_i \in [0, 1]$, $[i = 1, \dots, n]$. Here, we model the dependence between each pair of gene expression profile using empirical copulas. As we were unaware of the distributions of expression profiles, so empirical copulas are the only choice here. Notably, we have estimated joint empirical copula density from the marginals of each gene expression profile. We have used beta-kernel estimation to determine the copula density directly from the given data. The smoothing parameters are selected by minimizing the asymptotic mean-integrated squared error (AMISE) using the Frank copula as the reference copula. The input to the copula-density estimator is of size $n \times 2$, where n is the number of samples in different datasets. For each pair of samples, we estimate the empirical copula density using beta-kernel estimator. We have shown some marginal normal contour plots of copula density during the estimation process of the BRCA dataset in Supplementary Fig. 1. To model the differential coexpression of a gene pair, we have measured a statistical distance between two joint distributions provided by the copulas. We have utilized the K–S test to quantify the distance between two empirical distributions. The value of d statistic represents the distance here. Thus, the distance obtained for a gene pair is treated as a differential coexpression score.

To check whether the distance between the joint distribution perfectly models the differential coexpression, we have performed an analysis. To show the concordance between the DC_Score with the proposed distance, we have performed the following analysis. We create a 20 × 20 matrix M, whose rows (i) and columns (j) annotated with correlation values from −1 to +1 with 0.1 spacing. We create two pairs of marginals (F_{x1}, F_{x2}) and (F_{y1}, F_{y2}) having correlations i and j, respectively. For the generation of marginals, we have used mvnorm function of MASS R-package and set the "empirical" parameter of mvnorm as "TRUE". This generates (F_{x1}, F_{x2}) or (F_{y1}, F_{y2}) with an exact correlation of i or j. Next, we compute joint distributions using copula function $F_{X1X2} = C(F_{x1}, F_{x2})$, $F_{Y1Y2} = C(F_{y1}, F_{y2})$ and finally compute KS distance between F_{X1X2} and F_{Y1Y2}. Each entry of (i,j) in M is filled with this distance value. We generate M 100 times following the same method. Now, to visualize the matrices, each row is represented as a series of boxplots in Fig. 12. For a fixed row, the DC_Score will increase from left to right along the column as it ranges from correlation value −1 to +1. Each facet in the figure corresponds to a row/column in the matrix, which represents 20 sets of 100 distances corresponding to the correlations ranging from −1 to +1 with a spacing of 0.1. Considering each facet of the plot, it can be noticed that distances are gradually increasing with the increase in the DC_Score. For example, considering the second facet (corr value = −0.9), the distances increased from left to right gradually. So, it is evident from the figure that there exists a strong correlation between the distance and DC_Score, which signifies that the proposed method can model the difference in coexpression patterns.

STABILITY OF CODC

CODC is stable under noisy expression data. This is because of the popular "non-parametric", "distribution-free", or "scale-invariant" nature of the copula[29]. The properties can be written as follows: let C_{XY} be a copula function of two random variables X and Y. Now, suppose α and β are two functions of X and Y, respectively. The relation of $C_{(\alpha(X),\beta(Y))}$ and C_{XY} can be written as follows:

- **Property 1:** If α and β are strictly increasing functions, then the following is true:

$$C_{\alpha(X)\beta(Y)}(u, v) = C_{XY}(u, v) \qquad (6)$$

- **Property 2:** If α is strictly increasing and β is strictly decreasing, then the following holds:

$$C_{\alpha(X)\beta(Y)}(u, v) = u - C_{XY}(u, 1 - v) \qquad (7)$$

- **Property 3:** If α is strictly decreasing and β is a strictly increasing function, then we have

$$C_{\alpha(X)\beta(Y)}(u, v) = v - C_{XY}(v, 1 - u) \qquad (8)$$

- **Property 4:** If both α and β are strictly decreasing functions, then the following holds:

$$C_{\alpha(X)\beta(Y)}(u, v) = u + v - 1 - C_{XY}(1 - u, 1 - u) \qquad (9)$$

These properties of copula are used to prove that the distance measure used in CODC is approximately scaled invariant. Theoretical proofs are described below, and the simulation result is given later in section (Stability performance of CODC). The proof is as follows: we know that the K–S statistic for a cumulative distribution function F(x) can be expressed as

$$D = \sup_x |H_n(x) - F(x)|,$$

where H_n is an empirical distribution function for n i.i.d observation $X_i \le x$, and sup corresponds to supremum function. The two-sample K–S test used in CODC can be described similarly

$$D = \sup_{x,y} |(H_n^1(x, y) - F(x, y)) - (H_n^2(x, y) - F(x, y))| \\ = \sup_{x,y} |(H_n^1(x, y) - H_n^2(x, y))|, \qquad (10)$$

where H_n^1, H_n^2 are denoted as the joint empirical distribution for two samples taken from normal and cancer, respectively. Now the D statistic can be written as

$$D = \sup_{x,y} |(H_n^1(x, y) - H_n^2(x, y))| \\ = \sup_{x,y} |C(F_1(x), F1(y)) - C(F_2(x), F2(y))| \\ = \sup_{x,y} |C_{XY}(u, v) - C_{XY}(\tilde{u}, \tilde{v})|, \qquad (11)$$

where C(.) is copula function and $u = F_1(x)$, $v = F_1(y)$, $\tilde{u} = F_2(x)$, $\tilde{v} = F_2(y)$ are uniform marginals of joint distributions H_n^1 and H_n^2.

Let us assume that both α and β functions are strictly increasing. Then from Eqs. (6) and (11), the distance D between $H_n^1(\alpha(x), \beta(y))$ and $H_n^2(\alpha(x), \beta(y))$ has the form

$$D = \sup_{x,y} |H_n^1(\alpha(x), \beta(y)) - H_n^2(\alpha(x), \beta(y))| \\ = \sup_{x,y} |C(F_1(\alpha), F1(\beta)) - C(F_2(\alpha), F2(\beta))| \\ = \sup_{x,y} |C_{\alpha(x),\beta(y)}(u, v) - C_{\alpha(x),\beta(y)}(\tilde{u}, \tilde{v})| \\ = \sup_{x,y} |C_{XY}(u, v) - C_{XY}(\tilde{u}, \tilde{v})| \qquad (12) \\ [\text{By using the property in Eq.}(6)] \\ = \sup_{x,y} |C(F_1(x), F1(y)) - C(F_2(x), F2(y))| \\ = \sup_{x,y} \left|(H_n^1(x, y) - H_n^2(x, y))\right|.$$

Now, if α is strictly increasing and β is strictly decreasing, then D can be written as

$$D' = \sup_{x,y} |H_n^1(\alpha(x), \beta(y)) - H_n^2(\alpha(x), \beta(y))| \\ = \sup_{x,y} |C(F_1(\alpha), F1(\beta)) - C(F_2(\alpha), F2(\beta))| \\ = \sup_{x,y} |C_{\alpha(x),\beta(y)}(u, v) - C_{\alpha(x),\beta(y)}(\tilde{u}, \tilde{v})| \\ = \sup_{x,y} |u - C_{XY}(u, 1 - v) - \tilde{u} - C_{XY}(\tilde{u}, 1 - \tilde{v})| \\ [\text{By using the property in Eq. (7)}] \\ = \sup_{x,y} |(u - \tilde{u}) + C_{XY}(\tilde{u}, 1 - \tilde{v}) - C_{XY}(u, 1 - v)| \qquad (13) \\ = \sup_{x,y} |(u - \tilde{u}) + C_{XY}(\tilde{u}, \tilde{m}) - C_{XY}(u, m)| \\ \ge \sup_{x,y} |[C_{XY}(\tilde{u}, \tilde{m}) - C_{XY}(u, m)]| \\ \ge \sup_{x,y} |[C_{XY}(u, m) - C_{XY}(\tilde{u}, \tilde{m})]| \\ = \sup_{x,y} |H_n^1(x, y) - H_n^2(x, y)| \\ = D$$

Similarly, for strictly increasing β and strictly decreasing α, the distance D' between $H_n^1(\alpha(x), \beta(y))$ and $H_n^2(\alpha(x), \beta(y))$ can be shown to satisfy the relation

$$D' = \sup_{x,y} |H_n^1(\alpha(x), \beta(y)) - H_n^2(\alpha(x), \beta(y))| \\ \ge \sup_{x,y} |H_n^1(x, y) - H_n^2(x, y)| \qquad (14) \\ = D.$$

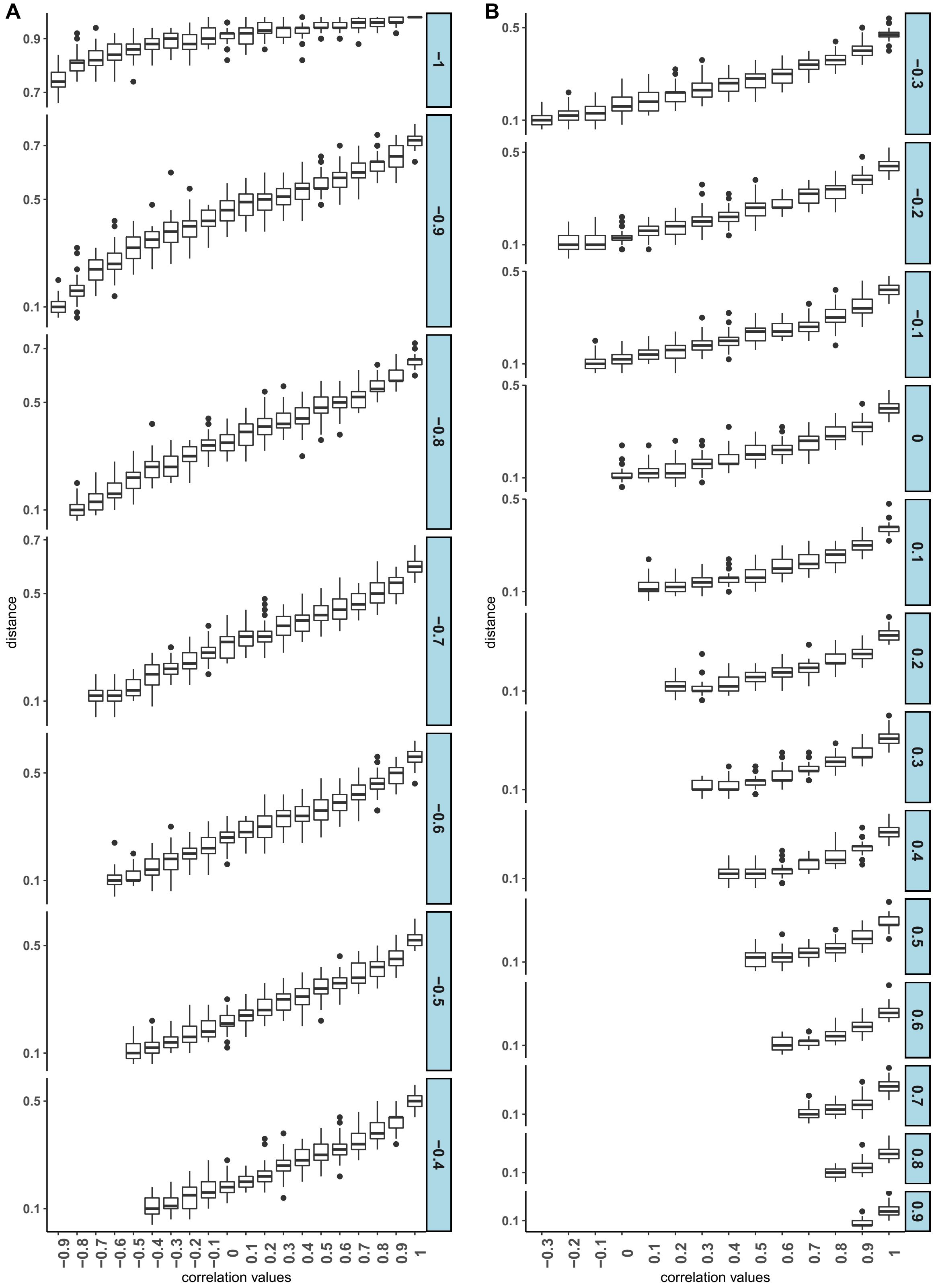

Fig. 12 Boxplot showing the dependency between DC_Score and K–S distance, between two joint distributions. a shows the distances for the facets from correlation −1 to −0.4 and **b** shows the same for correlation −0.3 to +1.

Finally, let us consider that both α and β are strictly decreasing functions. The distance D' can be described as

$$\begin{aligned} D' &= \sup_{x,y}|H_n^1(\alpha(x),\beta(y)) - H_n^2(\alpha(x),\beta(y))| \\ &= \sup_{x,y}|C(F_1(\alpha),F1(\beta)) - C(F_2(\alpha),F2(\beta))| \\ &= \sup_{x,y}|C_{\alpha(x),\beta(y)}(u,v) - C_{\alpha(x),\beta(y)}(\tilde{u},\tilde{v})| \\ &= \sup_{x,y}|u+v-1+C_{XY}(1-u,1-v) \\ &\quad -\tilde{u}+\tilde{v}-1+C_{XY}(1-\tilde{u},1-\tilde{v})| \\ &\quad [\text{By using the property in Eq. (8)}] \\ &= \sup_{x,y}|(u-\tilde{u})+(v-\tilde{v})+C_{XY}(1-u,1-v) \\ &\quad -C_{XY}(1-\tilde{u},1-\tilde{v})| \\ &\geq \sup_{x,y}|C_{XY}(1-u,1-v)-C_{XY}(1-\tilde{u},1-\tilde{v}|)| \\ &= \sup_{x,y}|C_{XY}(m,n)-c_{XY}(\tilde{m},\tilde{n})| \\ &= \sup_{x,y}|H_n^1(x,y)-H_n^2(x,y)|. \end{aligned} \tag{15}$$

Thus, the value of D' between two joint distributions $H_n^1(\alpha(x), \beta(y))$ and $H_n^2(\alpha(x), \beta(y))$ is the same as that when we add Gaussian noise to the original expression data of normal and cancer samples to transform these into noisy datasets of D that represents the distances $H_n^1(x, y)$ and $H_n^2(x, y)$ when both α and β are increasing functions. For other cases of α and β, D' attains at least the value of D. So, the distance for two random variables $\alpha(X)$ and $\beta(Y)$ is equal or at least that of the random variables X and Y. CODC treats the distance D as differential coexpression score; thus, it remains the same (or at least equal) under any transformation of X and Y.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this article.

DATA AVAILABILITY

Datasets are available in TCGA data portal.

CODE AVAILABILITY

https://github.com/Snehalikalall/CODC (https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga).

Received: 29 August 2019; Accepted: 18 March 2020;
Published online: 19 June 2020

REFERENCES

1. Ralston, A. & Shaw, K. Gene expression regulates cell differentiation. *Nat. Education* **1**, 127 (2008).
2. Yang, Y. et al. Gene co-expression network analysis reveals common system-level properties of prognostic genes across cancer types. *Nat. Commun.* **5**, 3231 (2014).
3. Eisen, M. B., Spellman, P. T., Brown, P. O. & Botstein, D. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl Acad. Sci.* **95**, 14863–14868 (1998).
4. Ideker, T. & Krogan, N. Differential network biology. *Mol. Syst. Biol.* 8, **565** (2011).
5. Ray, S. & Bandyopadhyay, S. Discovering condition specific topological pattern changes in coexpression network: an application to hiv-1 progression. *IEEE/ACM Trans. Comput. Biol. Bioinformatics* **11** (2015).
6. Cho, S., Kim, J. & Kim, J. Identifying set-wise differential co-expression in gene expression microarray data. *BMC Bioinformatics* **10**, 109 (2009).
7. Kostka, D. & Spang, R. Finding disease specific alterations in the co-expression of genes. *Bioinformatics* **20**, i194–i199 (2004).
8. Lai, Y., Wu, B., Chen, L. & Zhao, H. A statistical method for identifying differential gene–gene co-expression patterns. *Bioinformatics* **20**, 3146–3155 (2004).
9. Kostka, D. & R, R. S. Finding disease specific alterations in the co-expression of genes. *Bioinformatics* **20**, i194–199 (2005).
10. Watson, M. Coxpress: differential co-expression in gene expression data. *BMC Bioinformatics* **7**, 509 (2006).
11. Tesson, B., Breitling, R. & Jansen, R. Diffcoex: a simple and sensitive method to find differentially coexpressed gene modules. *BMC Bioinformatics* **11**, 497 (2010).
12. Fang, G. et al. Subspace differential coexpression analysis: problem definition and a general approach. *Biocomputing* **2010**, 145–156 (2009).
13. Wu, G. & Stein, L. A network module-based method for identifying cancer prognostic signatures. *Genome Biology* **13**, https://doi.org/10.1186/gb-2012-13-12-r112 (2012).
14. Langfelder, P. & Horvath, S. Wgcna: an R package for weighted correlation network analysis. *BMC Bioinformatics* **9**, 559 (2008).
15. Amar, D., Safer, H. & Shamir, R. Dissection of regulatory networks that are altered in disease via differential co-expression. *Plos Comput. Biol.* **9**, e1002955 (2013).
16. Ray, S. & Maulik, U. Identifying differentially coexpressed module during hiv disease progression: a multiobjective approach. *Scientific Rep.* **7**, 86 (2017).
17. Nelsen, R. B. *An Introduction to Copulas* (Springer Science & Business Media, 2007).
18. Embrechts, P. Copulas: a personal view. *J. Risk Insurance* **76**, 639–650 (2009).
19. Zheng, G. X. et al. Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.* **8**, 14049 (2017).
20. Juhász, C. et al. Tryptophan metabolism in breast cancers: molecular imaging and immunohistochemistry studies. *Nuclear Med. Biol.* **39**, 926–932 (2012).
21. Hashimoto, K. et al. Activated pi3k/akt and mapk pathways are potential good prognostic markers in node-positive, triple-negative breast cancer. *Annal. Oncol.* **25**, 1973–1979 (2014).
22. Shatalova, E. G., Klein-Szanto, A. J., Devarajan, K., Cukierman, E. & Clapper, M. L. Estrogen and cytochrome p450 1b1 contribute to both early-and late-stage head and neck carcinogenesis. *Cancer Prevention Res.* **4**, 107–115 (2011).
23. Kuang, J., Zhao, M., Li, H., Dang, W. & Li, W. Identification of potential therapeutic target genes and mechanisms in head and neck squamous cell carcinoma by bioinformatics analysis. *Oncology Lett.* **11**, 3009–3014 (2016).
24. Anttila, S., Raunio, H. & Hakkola, J. Cytochrome p450–mediated pulmonary metabolism of carcinogens: regulation and cross-talk in lung carcinogenesis. *Am. J. Respiratory Cell Mol. Biol.* **44**, 583–590 (2011).
25. Gonlugur, U., Mirici, A. & Karaayvaz, M. Pancreatic involvement in small cell lung cancer. *Radiol. Oncol.* **48**, 11–19 (2014).
26. Barhoumi, R., Mouneimne, Y., Chapkin, R. S. & Burghardt, R. C. Effects of fatty acids on benzo [a] pyrene uptake and metabolism in human lung adenocarcinoma a549 cells. *PloS ONE* **9**, e90908 (2014).
27. Nelsen, R. B. Introduction. In *An Introduction to Copulas*, 1–4 (Springer, 1999).
28. Sklar, A. Random variables, joint distribution functions, and copulas. *Kybernetika* **9**, 449–460 (1973).
29. Nelsen, R. B. Properties and applications of copulas: A brief survey. In *Proceedings of the first brazilian conference on statistical modeling in insurance and finance* (University Press USP Sao Paulo, 2003).

ACKNOWLEDGEMENTS

We would like to thank Dr. Debarka Sengupta, IIIT, Delhi for insightful discussions. S.R. acknowledges support from ERCIM Alain Bensoussan Fellowship programme grant. S. B. acknowledges support from J.C. Bose Fellowship [SB/S1/JCB-033/2016 to S.B.] by the DST, Govt. of India; SyMeC Project grant [BT/Med-II/NIBMG/SyMeC/2014/Vol. II] was given to the Indian Statistical Institute by the Department of Biotechnology (DBT), Govt. of India.

AUTHOR CONTRIBUTIONS

All authors contributed equally to this work.

COMPETING INTERESTS

The authors declare no competing interests.

ADDITIONAL INFORMATION

Supplementary information is available for this paper at https://doi.org/10.1038/s41540-020-0137-9.

Correspondence and requests for materials should be addressed to S.R. or S.B.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.