CODC: A copula based model to identify differential coexpression

Sumanta Ray[1◑], Snehalika Lall[2◑], Sanghamitra Bandyopadhyay[2¤],

1 Centrum Wiskunde & Informatica, Life Sciences & Health, 1098 XG Amsterdam, The Netherlands
2Machine Intelligence Unit, Indian Statistical Institute,Kolkata, India

◑These authors contributed equally to this work.
* Sumanta.Ray@cwi.nl

Abstract

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 mean 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 pair 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 which 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. Additionally, 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.

Introduction

Microarray based gene coexpression analysis has been demonstrated as an emerging field which 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 change in gene coexpression patterns across normal and disease states [6, 6–9]. Finding differentially coexpressed (DC) gene pairs, gene clusters, and dysregulated pathways between normal and disease states are 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, 14].

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 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 [15]. Another method called DICER(Differential Correlation in Expression for meta-module Recovery) [16] also identifies gene sets whose correlation patterns differ between disease and control samples. Dicer not only identifies the differentially coexpressed module but it goes step beyond and identifies meta-modules or 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 et al [17] proposed a multiobjective framework called DiffCoMO to detect differential coexpression between two stages of HIV-1 disease progression. Here, the algorithm operates on two objective functions which 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 **Co**pula based model to identify **D**ifferential **C**oexpression of genes under two different conditions. 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 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 TCGA data portal (https://tcga-data.nci.nih.gov/docs/publications/tcga/?).

Materials and methods

In this section, we have briefly introduced the proposed method which is based on copula function.

Modeling differential coexpression using Copula

Differential coexpression is simply defined as the change of 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:

$DC\_Score_{i,j}^{p1,p2} = |Sim(x_i, x_j)^{p1} - Sim(x_i, x_j)^{p2}|$, 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 $Sim(x_i, x_j)^p$ signifies Pearson correlation between x_i and x_j for phenotype p.

In the statistical analysis, the simple way to measure the dependence between the correlated random variable is to use copulas [18]. 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 variable. Let, $X1$ and $X2$ be the random vectors whose marginals are uniformly distributed in $[0, 1]$ and having marginal distribution F_{X1} and F_{X2} respectively. By Sklar's Theorem [19] we have the following: there exists a copula C such that $F(x1, x2) = C(F_{X1}(x1), F_{X2}(x2))$, for all $x1$ and $x2$ in the domain of F_X1 and F_X2. In other words, there exists a bivariate copula which represents the joint distribution as a function of its marginals. For the multivariate case the copula (C) function can be represented as:

$$F_X(X_1, X_2, \cdots, X_n) = C(F_1(x_1), F_2(x_2), \cdots, F_n(x_n))$$

, where $X_1, X_2, \cdots, X_n$, be the random vectors whose marginals are $F_1(x_1), F_2(x_2), \cdots, F_n(x_n)$. The converse of the theorem is also true. Any copula function with individual marginals $F_i(x_i)$ as the arguments, represents valid joint distribution function with marginals $F_i(x_i)$. So, Copula is also known as joint distribution generating function with a separate choice of marginals. Hence, different families (parametric and non-parametric) of copulas exist which model different types of dependence structure. The example includes Farlie-Gumbel-Morgenstern family (parametric), Archimedean Copula (parametric), Empirical Copula (non-parametric), Gaussian (parametric), t (parametric) etc. Empirical copulas are governed by the empirical distribution functions, which tries to estimate the underlying probability distribution from given observations.

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. Particularly, we have used joint empirical distributions to estimate the marginals of each gene expression profile. For two different phenotype conditions, the dependency between gene expression profiles is measured by empirical copulas. To model the differential coexpression of a gene pair, we have measured a statistical distance between two joint distribution provided by the copulas. We have utilized the Kolmogorov-Smirnov (K-S) test to quantify the distance between two empirical distributions. 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) corresponds to 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. 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 boxplot in the figure 1. 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 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 the proposed method can able to model the difference in coexpression patterns.

Stability of CODC

CODC is stable under noisy expression data. This is because of the popular "nonparametric", "distribution-free" or "scale-invariant" nature of the copula [20]. 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:
$$C_{\alpha(X)\beta(Y)}(u,v) = C_{XY}(u,v) \tag{1}$$
- **Property 2:** If α is strictly increasing and β is strictly decreasing, then the following holds:
$$C_{\alpha(X)\beta(Y)}(u,v) = u - C_{XY}(u,1-v) \tag{2}$$
- **Property 3:** If α is strictly decreasing and β is strictly increasing function, then we have
$$C_{\alpha(X)\beta(Y)}(u,v) = v - C_{XY}(v,1-u) \tag{3}$$
- **Property 4:** If α and β both are strictly decreasing function then the following holds:
$$C_{\alpha(X)\beta(Y)}(u,v) = u + v - 1 - C_{XY}(1-u,1-u) \tag{4}$$

These properties of copula are used to prove that the distance measure used in CODC is approximately scaled invariant. Theoretical proves are described below and simulation result is given later in section . The proof is as follows: we know the Kolmogorov-Smirnov 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 $X_i \leq x$, and sup corresponds to supremum function. The two sample K-S test is used in CODC can be described similarly:

$$\begin{aligned} D &= sup_{x,y}|(H_n^1(x,y) - F(x,y)) - (H_n^2(x,y) - F(x,y))| \\ &= sup_{x,y}|(H_n^1(x,y) - H_n^2(x,y))| \end{aligned} \tag{5}$$

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:

$$\begin{aligned} D &= sup_{x,y}|(H_n^1(x,y) - H_n^2(x,y))| \\ &= sup_{x,y}|C(F_1(x), F1(y)) - C(F_2(x), F2(y))| \\ &= sup_{x,y}|C_{XY}(u,v) - C_{XY}(\tilde{u},\tilde{v})| \end{aligned} \tag{6}$$

, 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 distribution H_n^1 and H_n^2.

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

$$\begin{aligned}
D &= sup_{x,y}|H_n^1(\alpha(x), \beta(y)) - H_n^2(\alpha(x), \beta(y))| \\
&= sup_{x,y}|C(F_1(\alpha), F1(\beta)) - C(F_2(\alpha), F2(\beta))| \\
&= sup_{x,y}|C_{\alpha(x),\beta(y)}(u, v) - C_{\alpha(x),\beta(y)}(\tilde{u}, \tilde{v})| \\
&= sup_{x,y}|C_{XY}(u, v) - C_{XY}(\tilde{u}, \tilde{v})| \\
&\text{[By using the property in equation 1]} \\
&= sup_{x,y}|C(F_1(x), F1(y)) - C(F_2(x), F2(y))| \\
&= sup_{x,y}|(H_n^1(x, y) - H_n^2(x, y))|
\end{aligned} \tag{7}$$

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

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

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:

$$\begin{aligned}
D' &= sup_{x,y}|H_n^1(\alpha(x), \beta(y)) - H_n^2(\alpha(x), \beta(y))| \\
&\geq sup_{x,y}|H_n^1(x, y) - H_n^2(x, y)| \\
&= D
\end{aligned} \tag{9}$$

Finally let us consider α and β both are strictly decreasing function. The distance

D' can be described as:

$$\begin{aligned}
D' &= sup_{x,y}|H_n^1(\alpha(x),\beta(y)) - H_n^2(\alpha(x),\beta(y))| \\
&= sup_{x,y}|C(F_1(\alpha),F1(\beta)) - C(F_2(\alpha),F2(\beta))| \\
&= sup_{x,y}|C_{\alpha(x),\beta(y)}(u,v) - C_{\alpha(x),\beta(y)}(\tilde{u},\tilde{v})| \\
&= sup_{x,y}|u+v-1+C_{XY}(1-u,1-v) \\
&- \tilde{u}+\tilde{v}-1+C_{XY}(1-\tilde{u},1-\tilde{v})| \\
&[\text{By using the property in equation 3}] \qquad (10) \\
&= Sup_{x,y}|(u-\tilde{u})+(v-\tilde{v})+C_{XY}(1-u,1-v) \\
&- C_{XY}(1-\tilde{u},1-\tilde{v})| \\
&\geq sup_{x,y}|C_{XY}(1-u,1-v) - C_{XY}(1-\tilde{u},1-\tilde{v}|)| \\
&= sup_{x,y}|C_{XY}(m,n) - c_{XY}(\tilde{m},\tilde{n})| \\
&= sup_{x,y}|H_n^1(x,y) - H_n^2(x,y)|
\end{aligned}$$

Thus the value of D' between two joint distribution $H_n^1(\alpha(x),\beta(y))$ and $H_n^2(\alpha(x),\beta(y))$ is the same as that of D which represents the distance $H_n^1(x,y)$ and $H_n^2(x,y)$ when both α and β are increasing function. For other cases of α and β, D' attains at least the value of D. So, the distance for two random variable $\alpha(X)$ 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.

Results

Dataset preparation

We have evaluated the performance of the proposed method in five RNAseq expression data. We have downloaded matched pair of tumor and normal samples from five pan cancer data sets: 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 counts 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 [21]. 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 article is summarized in table 1. Figure 2-panel(A) and Panel-B represent box and violin plot of average expression value of samples for each dataset.

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 which takes expression profile of a gene as marginals in normal and tumor sample. The K-S distance, computed between the joint distribution is served as differential coexpression score between a gene pair.

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

The score for a gene pair (g_i, g_j) can be formulated as: $DC\_Copula(g_i, g_j) = KS\_dist(e.c(g_i^{tumor}, g_j^{tumor}), e.c(g_i^{normal}, g_j^{normal}))$, where KS-dist represents Kolmogorov-Smirnov distance between two joint probability distribution, e.c represents empirical copula, 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 figure 3. 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 figure 4. This figure shows plots of all gene pairs having a positive correlation in normal and the negative correlation in tumor (shown in the panel-A) and vice-versa (shown in the panel-(B)). The density of the correlation values is shown in panel-C and panel-D for each case. In figure 5 we create a visualization of top differentially coexpressed gene pairs in BRCA data which shows a strong positive correlation in tumor stage and negative correlation in normal stage. The figure shows a heatmap of binary matrix constructed from the expression data of those gene pairs in tumor and normal stages. When the expression values showing the same pattern for a gene pair it is assumed 1, while 0 representing a non-matching pattern. From the figure, it is quite understandable that most of the entry in the normal stage is 0 (non-match) while in tumor stage is 1 (match). For other datasets, the plots are shown in supplementary figure.

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 sample to transform these into noisy datasets. We have utilized BRCA data in this analysis.

compute the K-S distance and obtain DC_copula matrix for 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 $D(i,j)(orD'(i,j)) \geq t$. Now, we first we set t as the maximum of D and D', and then decreases it continuously to extract the gene pairs. For each t we observe the common gene pairs obtained from D and D'. The rationale behind this is to observe the performance of CODC in noisy data. Figure 6 shows the proportion of common genes selected from D and D' for different threshold selection. Theoretically, CODC produces D with scores no more than D' (from sec). So, it is quite natural that common genes get increased with the lower threshold. From the property of section 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 6 shows the evidence for this case. As can be seen from the figure for threshold value above 0.25 more than 60% 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 among the modules we have identified module eigengene networks for each dataset. According to [15] 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 figure 7. 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 dataset. It is clear from figure 7 most of the modules show differential coexpression pattern in normal and tumor stage. From panel-A it can be noticed that for BRCA data the modules have a negative correlation in normal stage while showing a strong positive correlation in tumor stage. For HNSC dataset the opposite case is observed. Modules have a strong positive correlation in normal stages while having a negative correlation in tumor stage. In supplementary figures, the visualization of all datasets is given.

Comparisons with state-of-the-art

For comparison purpose, we have taken three state-of-the-art 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 detect correlation module in one stage and find the alternation of the correlation pattern within the module in other class. 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 state-of-the art. 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 top 1000 gene pairs having high DC_copula scores from the DC matrix, and perform classification using normal and tumor samples. We take expression ratio of each DC gene pairs from the expression matrix and compiled an $n \times 1000$, where n represents number of samples in each data. Classification is performed by treating normal and tumor samples as class label. A toy example of the comparison is shown in figure 8. Please note that all these methods are meant for differentially coexpressed module detection. So, for comparison, we collected the DC gene pairs 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 the figure 9. It can be noticed from the figure that for most of the dataset proposed method achieved high accuracy compare to the other methods.

To assess the performance of all the methods for detecting differential coexpression modules, we check the distribution of 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 differentially coexpression. It also fails to consider the gene pair having a low score but 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 pairs in two different samples (normal and cancer) and plot frequency polygon in figure 10. 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 10 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 towards positive correlation in both tumor and normal samples. In figure 10-(b) 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 to +1.

Pathway analysis

To compare functional enrichment of identified modules we have utilized KEGG pathway analysis. We defined the pathway score of a module as a proportion of gene within the module enriched with a certain pathway. We compare the pathway score for the modules identified for DiffCoEx, DiffCoMO, CoXpress and the proposed method. Figure 11 shows the result. It is clear from the figure that more modules for the proposed method achieved high pathway score compare to other state-of-the-arts. In figure 7 we have shown heatmaps of differentially coexpressed modules for BRCA, HNSC and for LUAD data. The heatmap also provides pathways and G-terms significantly enriched with the modules. The p-value for KEGG pathway enrichment and GO enrichment is computed by using the hypergeometric test with 0.05 FDR corrections. We have utilized GOstats, kegg.db and GO.db R package for that. It can be seen from figure 7, panel-A 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’ have eminent evidence to linked with malignant progression in breast cancer [22]. In [23] the association between ABC transporters with breast carcinoma has been established. From panel-B it can be seen that Drug metabolism-cytochrome P450 [24] ECM-receptor interaction [25], ’Nitrogen metabolism’, ’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 [24] [25]. Similarly from panel-C it can notice that pathways such as 'Metabolism of xenobiotics by cytochrome P450', 'Pancreatic secretion', 'Linoleic acid metabolism' is significantly associated with modules of LUAD data. Among them there exist strong evidence for pathways: 'Metabolism of xenobiotics by cytochrome P450', [26], 'Pancreatic secretion' [27], and 'Linoleic acid metabolism' [28] to be associated with lung carcinoma.

Conclusion

In this article, 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. Kolmogorov-Smirnov distance between two joint distributions is treated as differential coexpression score of a gene pair. We have compared CODC with three state-of-the-arts DiffCoex, CoXpress and DiffCoMO in five pan-cancer RNA-Seq data of TCGA. CODC's ability for delineating 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.

Under the premise that the differential coexpressed genes are likely to be important bio-markers, we demonstrate that CODC identifies those which achieve better accuracy for classifying samples. Moreover, CODC goes a step further from the pairwise analysis of genes and seeks modules wherein differential coexpression are 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', 'ECM-receptor interaction'.

We have evaluated the efficacy of CODC on 5 different pan-cancer dataset 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 state-of-the-arts. We have also proved that the scale-invariant property of copula makes CODC more robust for detecting differential coexpression in noisy data.

Acknowledgment

We would like to acknowledge 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] given to the Indian Statistical Institute by the Department of Biotechnology (DBT), Govt. of India.

References

1. Ralston A, Shaw K. Gene expression regulates cell differentiation. Nature education. 2008;1(1):127.

2. Yang Y, Han L, Yuan Y, Li J, Hei N, Liang H. Gene co-expression network analysis reveals common system-level properties of prognostic genes across cancer types. Nature communications. 2014;5:3231.

3. Eisen MB, Spellman PT, Brown PO, Botstein D. Cluster analysis and display of genome-wide expression patterns. Proceedings of the National Academy of Sciences. 1998;95(25):14863–14868.

4. Ideker T, Krogan N. Differential network biology. Molecular Systems Biology. 2011;565(8).

5. Ray S, Bandyopadhyay S. Discovering condition specific topological pattern changes in coexpression network: an application to HIV-1 progression. IEEE/ACM Transactions on Computational Biology and Bioinformatics. 2015;11(4).

6. Cho S, Kim J, Kim J. Identifying set-wise differential co-expression in gene expression microarray data. BMC Bioinformatics. 2009;10(109).

7. Kostka D, Spang R. Finding disease specific alterations in the co-expression of genes. Bioinformatics. 2004;20(1):i194–i199.

8. Lai Y, Wu B, Chen L, Zhao H. A statistical method for identifying differential gene–gene co-expression patterns. Bioinformatics. 2004;20(17):3146–3155.

9. Kostka D, R RS. Finding disease specific alterations in the co-expression of genes. Bioinformatics. 2005;20(Sup 1):i194–199.

10. Watson M. CoXpress: differential co-expression in gene expression data. BMC Bioinformatics. 2006;7(509).

11. Tesson B, Breitling R, Jansen R. DiffCoEx: a simple and sensitive method to find differentially coexpressed gene modules. BMC Bioinformatics. 2010;11(497).

12. Fang G, Kuang R, Pandey G, Steinbach M, et al, CM. Subspace differential coexpression analysis: problem definition and a general approach. Pacific Symposium on Biocomputing; 2010. p. 145–156.

13. Wu G, Stein L. A network module-based method for identifying cancer prognostic signatures. Genome Biology. 2012;13(R112):DOI: 10.1186/gb–2012–13–12–r112.

14. Watson M. CoXpress: differential co-expression in gene expression data. BMC Bioinformatics. 2006;7(509).

15. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics. 2008;9(559).

16. Amar D, Safer H, Shamir R. Dissection of Regulatory Networks that Are Altered in Disease via Differential Co-expression. Plos Comp Bio. 2013;9(3):e1002955.

17. Ray S, Maulik U. Identifying differentially coexpressed module during HIV disease progression: A multiobjective approach. Scientific reports. 2017;7(1):86.

18. Nelsen RB. Introduction. In: An Introduction to Copulas. Springer; 1999. p. 1–4.

19. Sklar A. Random variables, joint distribution functions, and copulas. Kybernetika. 1973;9(6):449–460.

20. Nelsen RB. Properties and applications of copulas: A brief survey. In: Proceedings of the first brazilian conference on statistical modeling in insurance and finance. Citeseer; 2003. p. 10–28.

21. Zheng GX, Terry JM, Belgrader P, Ryvkin P, Bent ZW, Wilson R, et al. Massively parallel digital transcriptional profiling of single cells. Nature communications. 2017;8:14049.

22. Juhász C, Nahleh Z, Zitron I, Chugani DC, Janabi MZ, Bandyopadhyay S, et al. Tryptophan metabolism in breast cancers: molecular imaging and immunohistochemistry studies. Nuclear medicine and biology. 2012;39(7):926–932.

23. Hashimoto K, Tsuda H, Koizumi F, Shimizu C, Yonemori K, Ando M, et al. Activated PI3K/AKT and MAPK pathways are potential good prognostic markers in node-positive, triple-negative breast cancer. Annals of oncology. 2014;25(10):1973–1979.

24. Shatalova EG, Klein-Szanto AJ, Devarajan K, Cukierman E, Clapper ML. Estrogen and cytochrome P450 1B1 contribute to both early-and late-stage head and neck carcinogenesis. Cancer Prevention Research. 2011;4(1):107–115.

25. 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 letters. 2016;11(5):3009–3014.

26. Anttila S, Raunio H, Hakkola J. Cytochrome P450–mediated pulmonary metabolism of carcinogens: regulation and cross-talk in lung carcinogenesis. American journal of respiratory cell and molecular biology. 2011;44(5):583–590.

27. Gonlugur U, Mirici A, Karaayvaz M. Pancreatic involvement in small cell lung cancer. Radiology and oncology. 2014;48(1):11–19.

28. Barhoumi R, Mouneimne Y, Chapkin RS, Burghardt RC. Effects of fatty acids on benzo [a] pyrene uptake and metabolism in human lung adenocarcinoma A549 cells. PloS one. 2014;9(3):e90908.

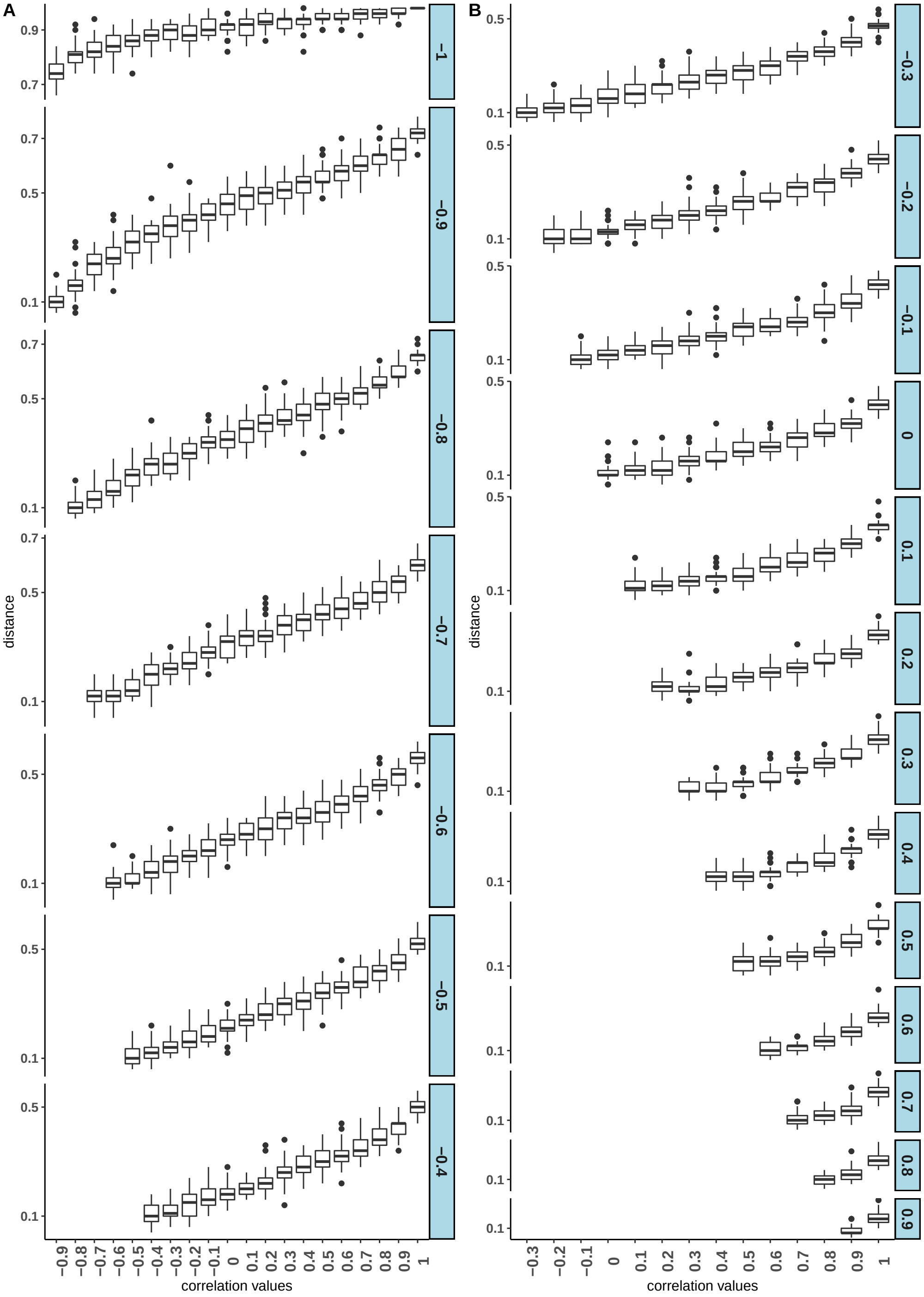

Fig 1. Boxplot showing the dependency between *DC_Score* and K-S distance between two joint distributions. Panel-A shows the distances for the facets from correlation -1 to -0.4 and Panel-B shows the same for correlation -0.3 to +1.

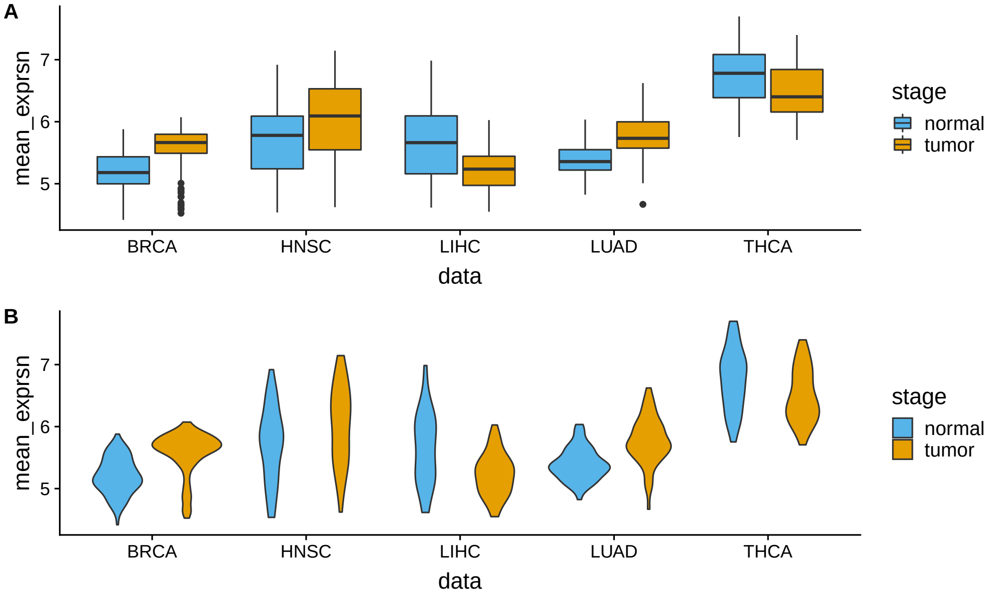

Fig 2. The figure the describes box (panel-A) and violin plots (panel-B) of mean expression values of the data used.

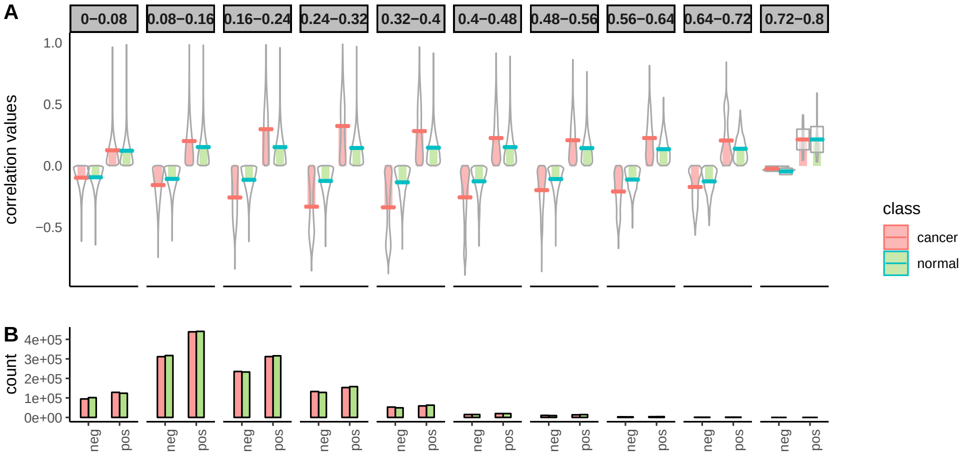

Fig 3. The figure shows the distribution of correlation values in normal and cancer samples of BRCA data with the *DC_Copula* score. Panel-A shows the distribution for different *DC_Copula* scores. Here, 4 pirate plots are shown in each facet, two for positive and two for negative correlations. The violins in each facet represent the distribution of positive and negative correlations of gene pair in normal and cancer samples. Panel-B shows a bar plot representing the number of positive and negatively correlated gene pairs in normal and cancer samples in each facet.

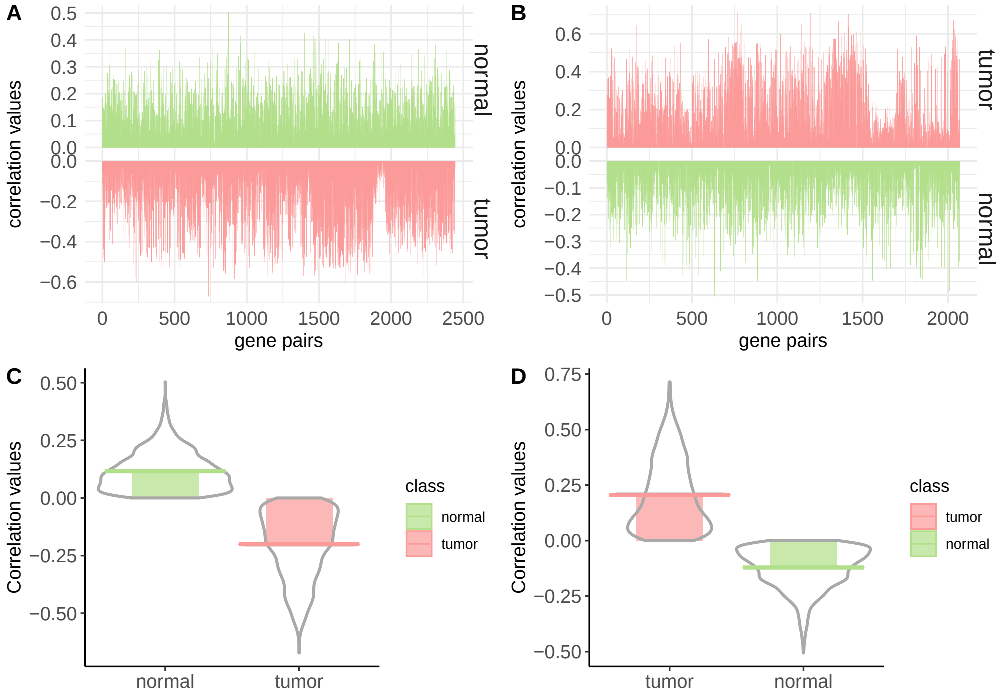

Fig 4. The figure shows visualizations of gene pairs having *DC_copula* score greater than 0.56. Panel-A and Panel-B show the visualization of correlation values of gene pair having a positive correlation in normal and negative correlation in tumor and vice-versa, respectively. Panel-C and Panel-D represent the distribution of correlation values according to panel-A and panel-B respectively.

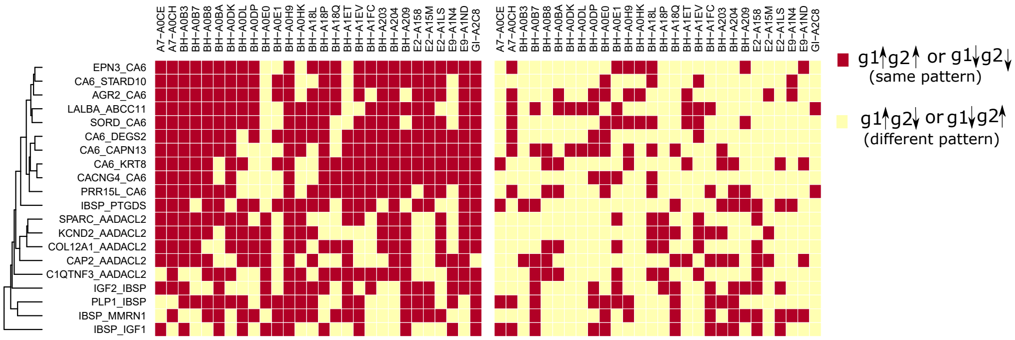

Fig 5. The figure shows a heatmap representation of binary matrix constructed from the expression matrix of top differentially coexpressed gene pairs in normal and tumor stages. Expression values of a gene pair showing the same pattern are indicated as '1' and showing a different pattern is indicated as '0' in the matrix. The columns representing differentially coexpressed gene pairs while rows are the samples of BRCA data.

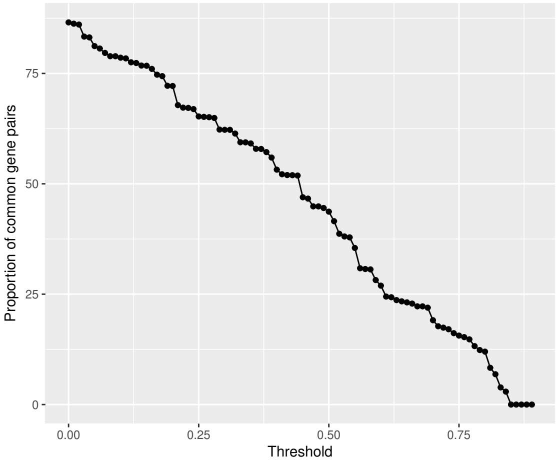

Fig 6. The proportion of common gene pairs obtained from noisy and original dataset with different threshold values.

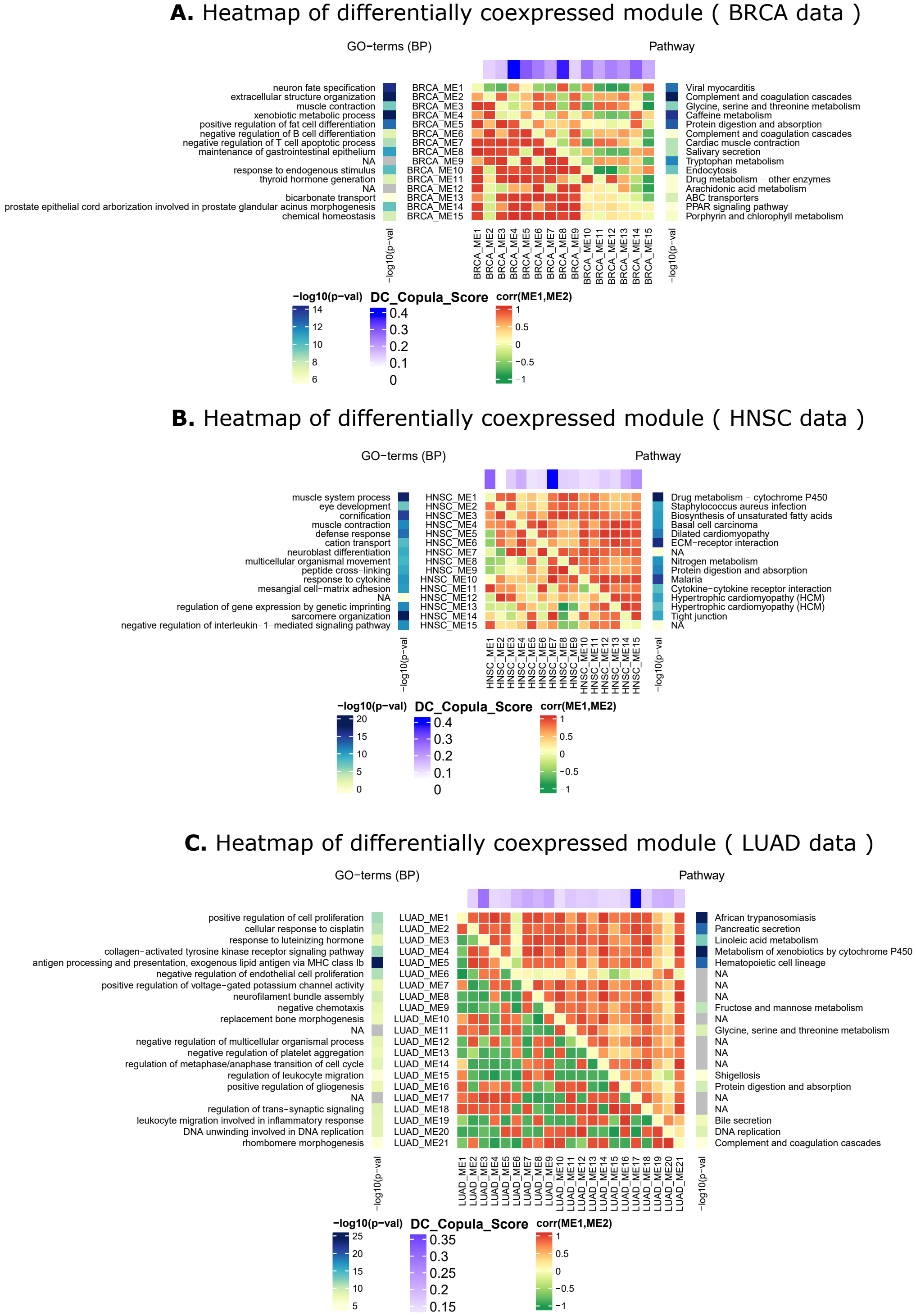

Fig 7. Heatmap of differentially coexpressed modules. Here the heatmap is shown for module eigengenes.The upper triangular portion of the matrix represents correlations of module eigengenes in normal samples whereas lower triangular portion signifies the same for tumor samples. Left and right sidebar of the heatmap represents -log(p-value) of significantly enriched GO-terms and pathway, respectively. Upper annotation bar of the heatmap shows the *DC_copula* score of the module. Panel-A shows the heatmap for BRCA data, whereas panel-B and Panel-C show heatmap HNSC and LUAD data.

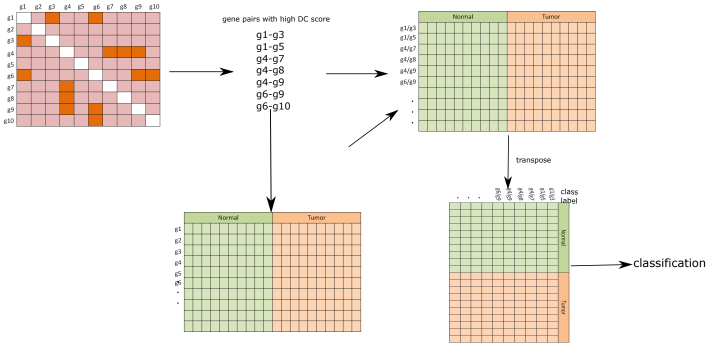

Fig 8. A toy example of performing classification on differentially coexpressed gene pairs. From the DC matrix top gene pairs are selected based on the DC score. Expression ratio is computed for each gene pairs for normal and tumor samples. The final matrix is then transposed and subsequently, classification is performed using normal and tumor sample ass class label.

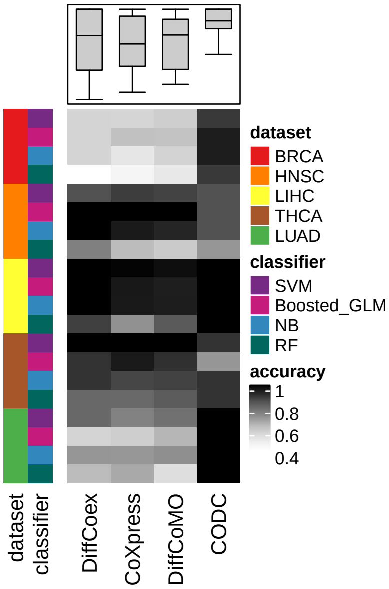

Fig 9. Comparison of classification accuracy for five datasets with four classifiers BGM, Naive-Bayes, Random Forest and SVM.

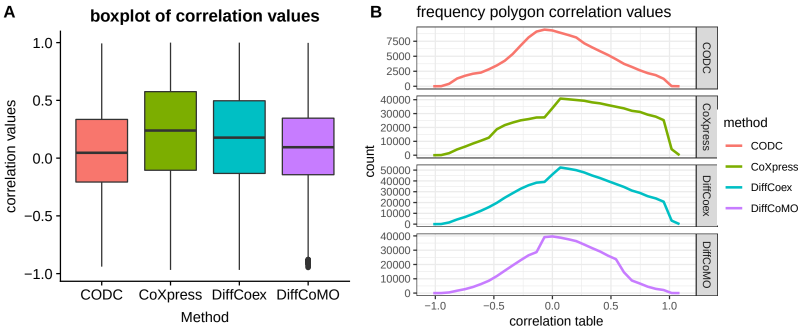

Fig 10. Distribution of correlation scores of the gene pairs in normal and tumor stage. Each facet shows the distribution for the different method.

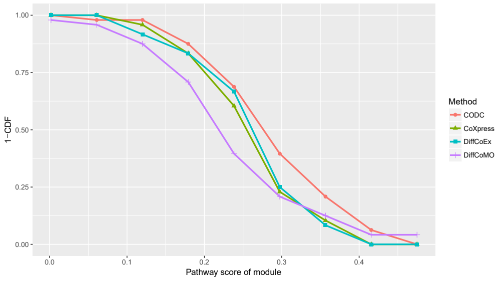

Fig 11. Distribution of pathway score for each of the comparing method. The figure shows the fraction of identified modules having a pathway score above a certain value.