Arnab Sadhu
Balaram Bhattacharyya
Tathagato Mukhopadhyay

SINGLE-SHOT DETERMINATION OF DIFFERENTIAL GENE NETWORK ON MULTIPLE DISEASE SUBTYPES

Abstract *A differential gene expressional network determines the prominent genes under altered phenotypes. The traditional approach requires $n(n-2)/2$ comparisons for n phenotypes. We present a direct method for determining a differential network under multiple phenotypes. We explore the non-discrete nature of gene expression as a pattern in a fuzzy rough set. An edge between a pair of genes represents a positive region of a fuzzy similarity relationship upon a phenotypic change. We apply a weight-ranking formula and obtain a directed ranked network; we label this as a phenotype interwoven network. Those nodes with large in-degree connectivity bubble up as significant genes under respective phenotypic changes. We tested the method on six diseases and achieved good corroboration with the results of previous studies in the two-step approach. The subgraphs of the isolated genes achieved good significance upon validation through an information theoretic approach. The top-ranking genes determined in all of our case studies are in consonance with the findings of the respective wet-lab tests.*

Keywords phenotype interwoven network, fuzzy rough set-based attribute selection, gene interaction network, differential co-expression

Citation Computer Science 23(2) 2022: 247–276

1. Introduction

A cell is a fundamental functional and biological unit in all living organisms. Individual organisms are characterized by their respective deoxyribonucleic acid (DNA) sequences – the primary constituent of their cells. Genes are subsequences of DNA that are distributed throughout each cell. Parts of the genes emulate to form m-RNA through the process of transcription and ultimately produce sequences of amino acids through translation – this results in proteins. The process by which the information content of a gene is transformed into proteins is defined as *gene expression*. The actual genetic encoding for an organism is termed its *genotype*, and the resulting physical characteristics are known as its *phenotype*. For a cell to develop and function properly, it must turn on the right gene at the right moment. Cellular diseases like cancer are caused by malfunctions in cells that effect genetic and epigenetic changes. These phenotypic changes occur from the abnormal expression of cancer-related genes (such as oncogenes or tumor suppressor genes). The DNA microarray experiment is an epoch-making technology that measures the expression profiles of several thousand genes from a relatively small number of samples and, thus, makes it possible to explore the genetic causes of anomalies occurring in the functioning of the human body. The samples are experimental cells under a specific phenotypic state.

Gene expression data opens up possibilities for looking for aberrations at the molecular level and correlate a set of genes with phenotypic changes. A set of genes is considered to have significant attributes if they exhibit expression patterns under diseased samples that are distinct from those of normal samples [22]. These are called molecular markers. These genes are of great importance in diagnosis as well as in the medication perspective. First, the accurate classification of normal and contaminated cells is important for disease diagnosis. For this reason, classifiers are built with expression values of those genes as their attributes. Second, those genes are analyzed for target-specific drug discovery and personalized medicines. Typically, the number of samples is too small as compared to the number of genes in a gene expression dataset. This often introduces ‘overfitting’ to such classifiers. Moreover, there are often redundant as well as noisy attributes that introduce error in the classifiers. Finding significant genes among many is, thus, a challenging task in view of the high dimensionality of the source data.

Methods for isolating such pivotal genes can be broadly classified into two categories: differential expression analysis, and differential co-expression analysis. The former [7, 8, 19, 47] finds individual genes that are deferentially expressed under an altered phenotype, while the latter [14, 32] focuses on isolating a geneset through building a co-regulation network. Genes are isolated from alterations in differential networks. In the present work, we aim to achieve the same task through building a single differential network that encompasses multiple phenotypic states.

A co-expression network of genes is an undirected graph where nodes correspond to genes and edge-weight represents an index of the co-regulation of a corresponding pair. Determining the index is crucial in a gene co-expression network. Key changes in an index under an altered phenotype condition indicate changes in regulation. Phenotype-specific co-expression networks are formed across the normal and diseased samples from which the genes are isolated by using the differential network. The differentially co-expressed genes that are thus identified are able to distinguish diseased samples from normal ones. Usually, the Pearson correlation coefficient (PCC) is considered to be an index [46] that works poorly with the existence of a nonlinear correlation. The expressions of genes are continuous in nature and rarely have perfect linear relationships among them. The mutual information approach [17] does not specifically assume the correlation pattern to be linear nor nonlinear, but the discretization of gene expression values might lead to a loss of crucial information [46]. We address these issues using a fuzzy rough set [13]. Moreover, all of the prevailing methods involve the development of multiple phenotype-specific networks in order to obtain the differential network. A geneset that is isolated for binary classification might not always distinguish all of the phenotypic subsets. For example, a geneset G_1 that is able to correctly distinguish phenotype subset P_1 from P_2 may not be able to isolate P_1 from P_3 (or vice-versa). A single network of genes across all phenotypes would, thus, be a viable alternative.

We have constructed an object set of samples with genes as general attributes and phenotypic states as decision attributes and model it as a fuzzy rough set that incorporates continuous patterns of the expression of genes. We compute the edge-weight index by employing a fuzzy equivalence relationship and its positive region with respect to the decision attribute. We thus build a single interwoven network that is comprised of gene-phenotype relationships – this eventually forms a weighted clique. We convert it into a ranked network [45] in order to obtain an ordered set of genes from their in-degree and call the ranked network that was built a phenotype interwoven network (*PINK*). A higher ranking implies a greater influence in disease progression. *PINK* props up the genes that feature high cardinality in connection with any phenotypic changes. The method explains the relationships among the genes by a fuzzy equivalence relationship and, hence, can reflect both linear and nonlinear interactions while implicitly encompassing the continuous nature of the expression data. This paper makes the following salient contributions:

- The method builds a single network that finds differentially expressed genes under multiple phenotypes.
- The method takes the inherent nonlinearity of the interactions into account.
- As we employ a fuzzy rough set into the method, quantizing the gene expression data is not required – this makes it robust against noise.
- We report those novel marker genes that have substantial wet-lab support.

The rest of the paper is organized as follows. We brief the problem statement in Section 2. In Section 3, we present the detailed method and algorithm; then, we demonstrate the method on a real data excerpt in Section 4. In Section 5, we present our results, analysis, comparisons, and discussion. We conclude this work in Section 6.

2. Problem statement

We consider two genes (g_1 and g_2) that possess expression levels in multiple samples (as can be seen in Table 1). The samples are either N (i.e., normal) or T (i.e., tumorous) phenotypes.

Table 1
Sample expression data

	Samples				
	T1	T2	N1	N2	N3
g_1	2	1	3	1	–2
g_2	3	2	3	1	–1

Figures 1a and 1b represent their corresponding expression levels. It is apparent that both genes failed to individually separate the two phenotypes with their expression patterns.

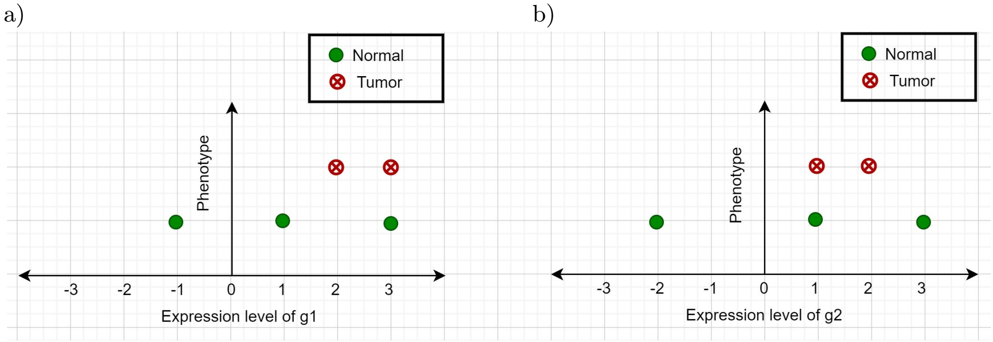

Figure 1. Expression pattern of g_1 and g_2 individually

In order to distinguish the phenotypes from the collective expression patterns of the genes, we take the co-expression graph of the gene pair (Fig. 2). It is apparent from Figure 2 that the phenotypes are separable through a single straight line; hence, the discriminating power of the genes increase when taken collectively. This is due to the differential co-expression pattern of the gene pair. This is the reason why we consider a network of genes where the edges represent the collective behavior of a gene pair. Gene pairs often possess a nonlinear correlation, which is better measured in terms of information gain(IG). Let $H(g)$ be the entropy of gene g and $H(p)$ be the entropy of the phenotype. We can find that the conditional entropies are $H(p|g_1) = H(p|g2) = 0.40$; however, the bi-variate conditional entropy is $H(p|g_1, g_2) = 2.059$ (hence, $IG = 2.059 - 0.4 = 1.66$). If the expression levels are displayed as fractions, then they must be converted into integers to compute IG. This conversion causes a loss of crucial information, and we cannot isolate genes that have an actual high discriminating power.

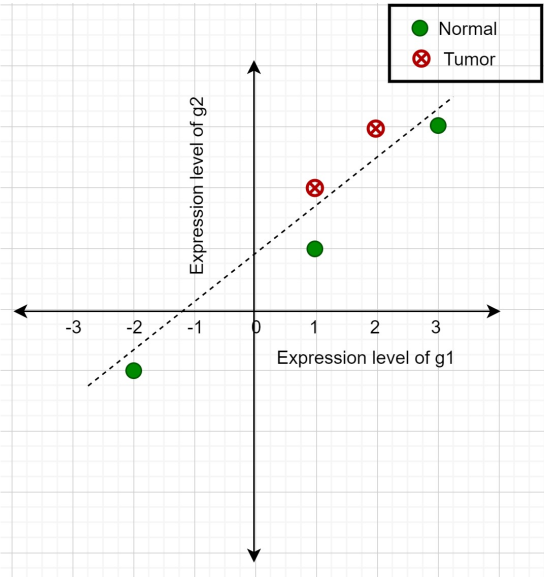

Figure 2. Co-expression patterns of g_1 and g_2

So, the existence of a nonlinear correlation and the non-discrete nature of the expression data yield errors while computing the collective expressional changes in a network of genes. We combat all of these issues by employing a fuzzy rough set. The proposed method takes fractional real expression values as inputs without any necessary conversions. It does not suffer in the presence of inherent non-linearity either.

3. Method

Let $A = \{g_1, g_2, \ldots, g_m\}$ be the set of genes, $U = \{S_1, S_2, \ldots, S_n\}$ be the set of samples, and $\mathbb{P} = \{P_1, P_2, \ldots, P_t\}$ be the set of phenotypes such that there exists a surjective mapping $U \rightarrow \mathbb{P}$. We present the system in Figure 3.

We aim to evaluate the pairwise differential co-regulation of the genes across the phenotypic changes. We emphasize finding the intra-phenotype similarity and inter-phenotype dissimilarity among the genes by using a common index. We attempt to compute the index from the expression values of the genes across sthe amples. Considering the samples to be tuples and the genes to be attributes, our task is to evaluate the pair-wise relationships among the attributes across the tuples. As the expression values are continuous, a fuzzy rough set (FRS) is appropriate for quantifying the relationships in terms of the similarity measure between gene pairs.

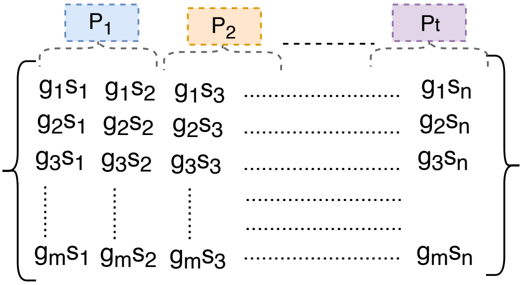

Figure 3. Sample-phenotype mapping

In the framework of FRS, we define the samples as tuples, the gene pairs as conditional attributes, and the sample phenotype as the decision attribute. We formulate the gene interaction by using the fuzzy equivalence relationship. An equivalence relationship organizes the groups of tuples into disjoint classes that are equivalent under the relationship. In the current scenario, we consider the pair to achieve a good fuzzy similarity if the samples from different phenotypes can be grouped into equivalent classes under the relationship between a pair of genes. Taking $\mathbb{R}$ as a family of fuzzy equivalence relationships that are associated with the set of conditional attribute $\mathbb{C}$ over sample set S, we model the system in Figure 3 into FRS (as can be seen in Figure 4).

FRS model
Input: Dataset $\mathbb{E}^{m\times n}$, where m is number of genes and n is total number of samples
Procedure:
$\mathbb{C} = (\{g_\alpha, g_\beta\})\forall\alpha \in 1, 2, \ldots, m-1; \beta = \alpha + 1, \ldots, m;$
$U = (\{S_1, S_2, \ldots, S_n\})$, $A = (\{g_1, g_2, \ldots, g_m\})$
$\mathbb{P} = (P_1, P_2, \ldots, P_t);$
Decision attribute $D = \mathbb{P}(S_i)$
$I = (U, A)$
Output: Fuzzy decision system $FDS = (U, \mathbb{R} \cup \mathrm{D})$

Figure 4. Modeling information system in FRS

With the transformed system in FRS, our task is to compute the pair-wise similarity among the genes that are present in conditional attribute subset $\mathbb{C}$. Let $\mathbb{C}_{eq}$ be the set of equivalence classes that are generated under the relationship of subset pair $\mathbb{C}$. The set of equivalent classes that are based on D (U/D) are given by the following equation:

$$(U/D) = (\{S_i|\mathbb{P}(S_i) = P_1\}, \ldots, \{S_i|\mathbb{P}(S_i) = P_t\}). \tag{1}$$

$\mathbb{C}_{eq}$ may not necessarily match (U/D). We measure the goodness of the fuzzy similarity by evaluating the positive region of the fuzzy similarity relationship with the actual decision attribute (D). We denote $sim(\mathbb{R})$ as the fuzzy similarity relationship. Either positive region ($POS_{sim(\mathbb{R})}D$) or the P-lower approximation is the union of all of the equivalence classes in $\mathbb{R}$ that are contained by the target set (U/D). The lower approximation denotes the complete set of objects in U/D that can be classified without any ambiguity. We proceed with the computation of the fuzzy similarity relationship and its positive region with the decision attribute.

3.1. Computation of fuzzy similarity relationship

For a pair of g_α and g_β genes in $\mathbb{C}$, we find their similarity relationship $sim(\mathbb{R})$ as follows:

$$sim(\mathbb{R}) = \cap\{\mathrm{R} : \mathrm{R} \in \mathbb{R}\}, \tag{2}$$

where R_α and R_β correspond to g_α, and the g_β in $\mathbb{C}$ is computed by using Lukasiewicz's t-norm ($TL(x, y) = max(x + y - 1, 0)$):

$$R_k(S_i, S_j) = 1 - |C_k(S_i) - C_k(S_j)|, \tag{3}$$

where $k \in \{\alpha, \beta\}$.

3.2. Computation of positive region

The positive region of $sim(\mathbb{R})$ with D is as follows:

$$POS_{sim(\mathbb{R})}D = \cup^{t}_{(\lambda=1)} sim(\mathbb{R})_* \mathrm{D}_\lambda, \tag{4}$$

where $sim(\mathbb{R})_*(\mathrm{D}_\lambda)(\mathrm{S})$ is the fuzzy lower approximation of $sim(\mathbb{R})$, which we compute by using

$$sim(\mathbb{R})_*(\mathrm{D}_\lambda)(\mathrm{S_i}) = \cap\{\mathrm{neg(R)(D_\lambda)(S_i)}\}. \tag{5}$$

We compute fuzzy dissimilarity relationship $neg(R)(D_\lambda)$ (that is, the negative relationship) from the samples that are outside set D_λ as follows:

$$neg(R)(D_\lambda)(S_i) = \{1 - sim(\mathbb{R}(\mathrm{i}, \mathrm{j}))|\mathbb{P}(\mathrm{S_i}) = \mathrm{D}_\lambda, \mathbb{P}(\mathrm{S_j}) \neq \mathrm{D}_\lambda\}. \tag{6}$$

3.3. Network formation

The network is comprised of undirected edges among all possible pairs of genes. The weight of an edge (g_α, g_β) is the fraction of the positive region of D in the universe of the discourse (U). We compute the positive region of D over attribute set $\mathbb{C}$ by using Eq. (4). Thus weight of the edge (g_α, g_β) is as follows:

$$\omega(\alpha, \beta) = \gamma(\mathbb{C}, D) = \frac{|POS_{(sim(\mathbb{R})}D|}{|U|}. \tag{7}$$

This forms a weighted, undirected, and fully connected network of genes. We now apply the rank-based method [46] on the network to obtain the phenotype interwoven network ($PINK$). For each gene, the weights of the out-degree edges are sorted in descending order. We retain only the top κ(user-defined) edges in the order of edge weight ω as follows:

$$PINK(\alpha,\beta) = \begin{cases} 1, & \text{if } g_\beta \in top_\kappa(\alpha), \forall \beta \neq \alpha \\ 0, & \text{otherwise} \end{cases}. \tag{8}$$

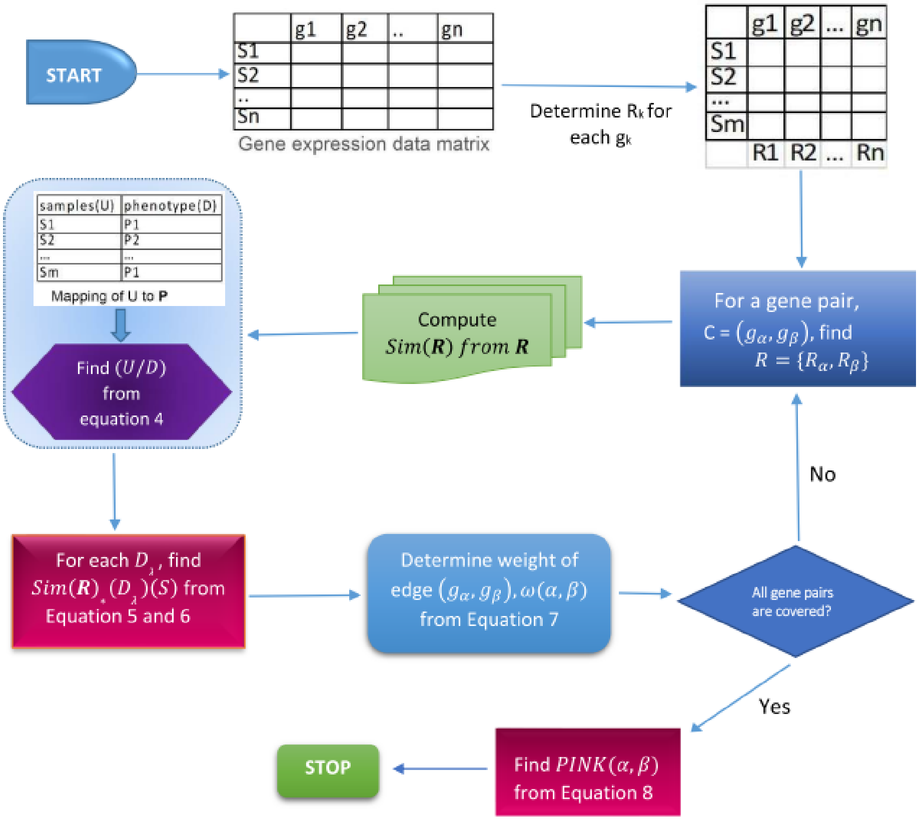
START
g1
g2
..
gn
S1
S2
..
Sn
Gene expression data matrix
Determine R_k for each g_k
g1
g2
...
gn
S1
S2
...
Sm
R1
R2
...
Rn
samples(U)
phenotype(D)
S1
P1
S2
P2
...
...
Sm
P1
Mapping of U to P
Find (U/D) from equation 4
Compute $Sim(\boldsymbol{R})$ $from$ $\boldsymbol{R}$
For a gene pair, $C = (g_\alpha, g_\beta)$, find $R = \{R_\alpha, R_\beta\}$
No
For each D_λ, find $Sim(\boldsymbol{R})_*(D_\lambda)(S)$ from Equation 5 and 6
Determine weight of edge $(g_\alpha, g_\beta), \omega(\alpha,\beta)$ from Equation 7
All gene pairs are covered?
Yes
STOP
Find $PINK(\alpha,\beta)$ from Equation 8

Figure 5. Flowchart of algorithm

The out-degree of each node in PINK is κ, while the corresponding in-degree can be up to $m-1$. The in-degree of a gene indicates its order of impact in the network. The steps that are followed for building $PINK$ are illustrated via a flowchart in Figure 5.

3.4. Algorithm

The pseudo code for the whole method is stated in Algorithm 1.

Algorithm 1 Build *PINK*

```
function Build_PINK($\mathbb{D}$)    ▷ $\mathbb{D}^{m\times n}$ is the dataset, $m$ is the number of genes, and $n$ is the total number of samples
    $R_\alpha[i,j] = 1 - |\mathbb{D}[\alpha,i] - \mathbb{D}[\alpha,j]| \forall (i,j) \leq n, i \neq j, \alpha = 1,2,\ldots,m$;
    for $\alpha = 1$ to $m-1$ do
        for $\beta = \alpha + 1$ to $m$ do
            $sim\mathbb{R} = \min(\mathrm{R}_\alpha, \mathrm{R}_\beta)$
            for $\lambda = 1$ to $t$ do
                $\mathrm{POS}_{sim\mathbb{R}} + = \mathrm{sim}_*\mathbb{R}\_\mathrm{P}(\mathrm{sim}\mathbb{R}, P_\lambda)$
            end for
            $\omega(g_\alpha, g_\beta) = POS_{sim\mathbb{R}}/|U|$;
        end for
    end for
    $PINK$ = build_ranked_net($\omega$);
    return($PINK$);
end function
function find_sim$_*\mathbb{R}$_P(sim$\mathbb{R}$,$P_\lambda$)
    for $i = 1$ to $|P_\lambda|$ do
        for $j = 1$ to $n$ do
            negR$[i,j] = 1 - sim\mathbb{R}$;
        end for
    end for
    $P'_\lambda = U \setminus P_\lambda$;
    for $i = 1$ to $|P_\lambda|$ do
        for $j = 1$ to $|P'_\lambda|$ do
            tempR[i,j] = negR[i,j];
        end for
    end for
    for $i = 1$ to $|P_\lambda|$ do
        sim$_*R\_P[i]$ = min(tempR[i]);
    end for
    return(sim$_*R\_P_\lambda$);
end function
function form_ranked_net($\omega$)
    for $\alpha = 1$ to $m-1$ do
        sort $\omega(\alpha,:)$ in decreasing order;
        for $\beta = \alpha + 1$ to $m$ do
            if ($g_\beta \in top_\kappa(\alpha)$) then
                $ranked\_net(\alpha,\beta) = 1$;
            else
                $ranked\_net(\alpha,\beta) = 0$;
            end if
        end for
    end for
    return(ranked_net);
end function
```

For m number of genes, the method finds an interwoven network of a size of $m \times m$. The computation of the fuzzy lower approximation has a complexity of $O(mn^2)$ for n number of samples where $n<<m$. The ranked network formulation possesses the complexity of $O(m^2 \log m\theta)$ where $\theta<<mn$. Therefore, the total run time is $O(m^3n^2) \approxeq O(m^3)$. The total space complexity is $O(m^2)$, as *PINK* is only a single square matrix of an order of m. The space requirement is independent of the number of phenotypes. This shows the applicability of *PINK* in large datasets that are comprised of multiple phenotypes.

4. Working of algorithm on real data excerpt

Here, we present a demonstration of building up a network from a subset of real data. We take an excerpt from the $GDS3257$ lung cancer gene expression dataset [30]. The excerpt contains the expression data of nine genes taken from four normal samples and four cancerous samples (Tab. 2). The expression data is normalized to $[0, 1]$. The phenotype partitions are $N = N_1, N_2, N_3, N_4$ and $C = C_1, C_2, C_3, C_4$. Decision attribute $D \equiv \{D_N, D_C | D_N \in N, D_C \in C\}$.

Table 2
Lung cancer data excerpt (N for normal, and C for cancerous)

Gene	N1	N2	N3	N4	C1	C2	C3	C4
TOX3	0.23	0.25	0.19	0.18	0.64	0.68	0.78	0.81
SPP1	0.28	0.20	0.17	0.10	0.78	0.73	0.86	0.80
COL10A1	0.11	0.15	0.14	0.13	0.77	0.60	0.65	0.82
GREM1	0.13	0.16	0.13	0.13	0.19	0.28	0.31	0.67
JAM2	0.85	0.90	0.80	0.65	0.46	0.39	0.18	0.36
AGER	0.93	0.87	0.98	0.89	0.19	0.53	0.19	0.33
SFTPC	0.98	0.96	0.97	0.99	0.66	0.78	0.26	0.69
CRIP1	0.79	0.75	0.79	0.96	0.62	0.69	0.70	0.65
CEACAM6	0.65	0.68	0.78	0.44	0.97	0.93	0.90	0.92

Step 1: We proceed to compute Lukasiewicz's similarity (R) using Equation 3 as follows:

$$\begin{aligned} R_{TOX3}(N_1, N_2) &= 1 - |TOX3(N_1) - TOX3(N_2)| \\ &= 1 - |0.23 - 0.25| = 1 - 0.02 = 0.98 \end{aligned}$$

$$\begin{aligned} R_{TOX3}(N_1, C_1) &= 1 - |TOX3(N_1) - TOX3(C_1)| \\ &= 1 - |0.23 - 0.64| = 1 - 0.41 = 0.59. \end{aligned}$$

In the same manner, the similarity matrix R_{TOX3} for gene TOX3 and the R_{SPP1} for gene SPP1 are as follows:

$$R_{TOX3} = \begin{pmatrix} 1 & 0.98 & 0.96 & 0.95 & 0.59 & 0.55 & 0.45 & 0.42 \\ & 1 & 0.94 & 0.93 & 0.61 & 0.57 & 0.48 & 0.44 \\ & & 1 & 0.99 & 0.55 & 0.51 & 0.41 & 0.38 \\ & & & 1 & 0.54 & 0.50 & 0.41 & 0.37 \\ & & & & 1 & 0.96 & 0.86 & 0.83 \\ & & & & & 1 & 0.90 & 0.87 \\ & & & & & & 1 & 0.96 \\ & & & & & & & 1 \end{pmatrix}$$

$$R_{SPP1} = \begin{pmatrix} 1 & 0.92 & 0.89 & 0.82 & 0.50 & 0.55 & 0.42 & 0.47 \\ & 1 & 0.97 & 0.90 & 0.42 & 0.48 & 0.34 & 0.40 \\ & & 1 & 0.93 & 0.40 & 0.45 & 0.31 & 0.37 \\ & & & 1 & 0.32 & 0.37 & 0.24 & 0.30 \\ & & & & 1 & 0.94 & 0.92 & 0.98 \\ & & & & & 1 & 0.87 & 0.92 \\ & & & & & & 1 & 0.94 \\ & & & & & & & 1 \end{pmatrix}.$$

Step 2: Then, we determine $sim(\mathbb{R})$ for the two genes from matrices R_{TOX3} and R_{SPP1} by using the $sim(\mathbb{R}) = \cap\{\mathrm{R} : \mathrm{R} \in \mathbb{R}\}$ formula:

$$sim(\mathbb{R}) = \begin{pmatrix} 1 & 0.92 & 0.89 & 0.80 & 0.50 & 0.55 & 0.42 & 0.42 \\ & 1 & 0.94 & 0.90 & 0.42 & 0.48 & 0.34 & 0.40 \\ & & 1 & 0.93 & 0.40 & 0.45 & 0.31 & 0.37 \\ & & & 1 & 0.32 & 0.37 & 0.24 & 0.30 \\ & & & & 1 & 0.94 & 0.86 & 0.83 \\ & & & & & 1 & 0.87 & 0.87 \\ & & & & & & 1 & 0.94 \\ & & & & & & & 1 \end{pmatrix}.$$

Step 3: We compute fuzzy dissimilarity relationship $neg(R)(D_C)$ for C using Equation 6 followed by the fuzzy lower approximation $sim_*(\mathbb{R})\mathrm{D_N})$ of $sim(\mathbb{R})$ for N and $sim(\mathbb{R})(\mathrm{D_C})$ for C using Equation 5:

$$sim(\mathbb{R})(\mathrm{D_N}) = [0.45, 0.53, 0.56, 0.63]$$

$$sim(\mathbb{R})(\mathrm{D_C}) = [0.50, 0.45, 0.58, 0.58].$$

Step 4: Finally, we compute the positive similarity region of D in $sim(\mathbb{R})$ using Equation 4 and obtain the edge weight ($\omega(TOX3, SPP1)$) between TOX3 and SPP1 using Equation 7.

Similarly, we compute the edge weights among all gene pairs and obtain the complete matrix ω:

$$\omega = \begin{pmatrix} 0 & 4.28 & 4.2 & 3.62 & 3.79 & 4.2 & 3.79 & 3.62 & 3.66 \\ & 0 & 4.57 & 4.24 & 4.33 & 4.58 & 4.34 & 4.22 & 4.22 \\ & 0 & 0 & 4.11 & 4.20 & 4.35 & 4.31 & 4.11 & 4.13 \\ & & & 0 & 2.82 & 3.79 & 2.47 & 1.59 & 2.11 \\ & & & & 0 & 4.06 & 3.03 & 2.83 & 3.30 \\ & & & & & 0 & 3.81 & 3.79 & 3.92 \\ & & & & & & 0 & 2.29 & 2.66 \\ & & & & & & & 0 & 1.65 \\ & & & & & & & & 0 \end{pmatrix}.$$

From matrix ω, we get *PINK* in Figure 6 using Equation 8 by taking the top-three outgoing edges ($\kappa = 3$):

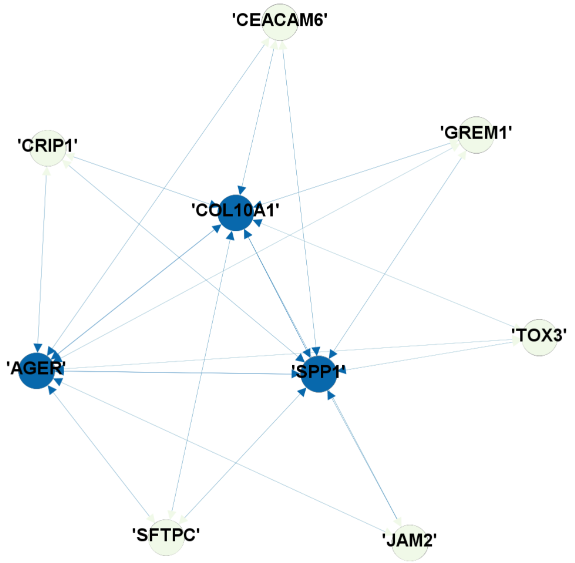

Figure 6. *PINK* from lung cancer data excerpt

.

$\omega(\alpha, \beta)$ signifies the goodness of the fuzzy similarity between g_α and g_β under the altered decision attribute. A higher $\omega(\alpha, \beta)$ suggests that the set of equivalent classes that is generated by the lower approximation of the fuzzy similarity relationship between g_α and g_β is more similar to the original class distribution. The top-ranked genes thus have distinct patterns of behavior under the altered phenotypes. Hence, the phenotypic status of the cell sample under observation can be determined by the pattern of the top-ranked genes.

The group of highly connected genes in *PINK* (Fig. 6) are COL10A1, SPP1, and AGER. The hierarchical clustering on the heatmap (Fig. 7) of this group of genes is the same as the sample pheno.

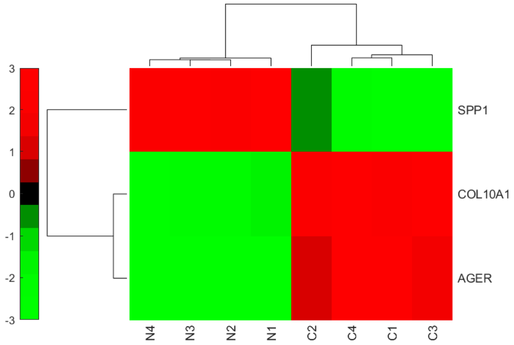

Figure 7. Heatmap of top-three connected genes
(note: expression values are normalized in $[-3, 3]$ for better visualization)

5. Results and discussion

We conducted our experiment on the six expression datasets that are shown in Table 3.

Table 3
Dataset description

Dataset	Gene count	Sample count
ALL-AML ([22])	7128	ALL: 27; AML: 11
Colon Cancer ([3])	2000	Cancer: 40; Normal: 22
OSCC ([50])	858	Normal: 40; OSCC: 40
BLCA	43,148	Primary bladder cancer: 165 Normal tissue surrounding cancer: 58

SRBCT ([29])	2308	Ewing family(EWS): 23; Burkitt lymphoma(BL): 8;
		Neuroblastoma(NB): 12; Rhabdomyosarcoma(RMS): 21;
		Test samples: 25

SLE ([6])	49,576	Strep: 12; Straph: 40; Still: 31; PSLE: 82; ASLE: 28; Control: 81

The top-ranking genes are isolated from *PINK* in each of the six case studies. We present comparisons and corroborations with the findings of several previous computational studies here, followed by their biological significance.

5.1. Test of validation

To verify the correctness of the results from *PINK*, we conducted a set of experiments on two benchmark datasets: ALL-AML, and colon cancer (as extensive works are reported on these datasets in the literature).

ALL-AML is one of the most extensively studied and referred-to datasets in the literature of gene expression mining [9, 22, 23, 28]. We present a list of the top-ten-ranking genes in Table 4.

Table 4
Isolated genes from *PINK* along with citations of matching found in ALL-AML dataset

Top-ranking genes	Reported in	Wet-lab tests
ACADM	[22, 28]	–
Zyxin	[22, 23, 28]	[5]
hdlc1	[22]	–
GLUL	–	[18]
LYN-Vy1	[22, 28]	–
SFTPA1	–	–
TCRA	–	[49]
MB-1	[28]	–
RBP P48	[22]	–
CD19	–	[53]

For a quantitative analysis of the performance of PINK in selecting significant genes, we conducted a voting-based classification that was similar to that which can be found in [22]. We trained the classifier by employing the isolated genes as attributes and recorded the prediction strength (PS) [22] of the selected features in both the training and test datasets in Table 5.

Table 5
Performance of classifier in ALL-AML data

Number of top-ranked genes, k	Classifying accuracy in training data		Classifying accuracy in test data	
	Median PS	#sample <0.3 PS	Median PS	#sample <0.3 PS
20	0.89	0/38	0.72	4/34
30	0.9	0/38	0.74	4/34
40	0.9	0/38	0.73	3/34
50	0.89	0/38	0.66	4/34
60	0.88	1/38	0.67	4/34
80	0.86	1/38	0.64	4/34

It is apparent from Table 5 that $k = 40$ produced an optimum result. Our predictor with $k = 50$ outperformed the 50-gene predictor of Golub et al. During their cross validation, Golub et al. left two training samples as uncertain (i.e., $PS<0.3$), while our predictor assigned all of the 34 samples correctly with certainty. On the test data, Golub et al. missed five samples ($PS<0.3$) as unclassified, while our predictor left only three samples. Our method's accuracy was also on par with two other highly respected methods that were applied on this dataset; viz., Furey et al. [21] and Guyon et al. [23]. In each case, the number of misclassified samples in the test data was above three.

Four widely found genes in the previous studies [22, 23, 28] (namely, ACADM, Zyxin, hdlc1, and LYN-v-y-1) placed in the top five of the *PINK* rankings. Among the newly isolated genes, GLUL, TCRA, and CD19 were reported to be significant in wet-lab tests [18, 49, 53].

For the colon cancer dataset, we list the top-ten-ranking genes that were isolated by *PINK* in Table 6.

Table 6

List of genes isolated from *PINK* along with citations in colon cancer dataset

Top-ranking mRNA	Reported in	Wet-lab tests
Hsa.692 (CRP)	–	–
Hsa.627 (MONAP)	[28]	–
Hsa.8147 (Human desmin gene)	[1, 28]	–
Hsa.36689 (GCAP-II)	[28]	[25]
Hsa.11673 (GTP-BINDING NUCLEAR PROTEIN RAN)	–	–
Hsa.1832 (MYOSIN REGULATORY LIGHT CHAIN 2)	[16, 20]	–
Hsa.37937 (MYOSIN HEAVY CHAIN)	[28]	[40]
Hsa.1131 (TROPOMYOSIN)	[16, 20]	[44]
Hsa.1130 (Tropomyosin isoform)	–	[35]
Hsa.3306 (hnRNP)	[1]	–

We employed the top-ranking genes to build a classifier for the purpose of a performance analysis. Following the validation procedures of Li et al. [33] and Cho et al. [16], we studied the average performance over 100 random partitions into 50 training and 12 test samples. Table 7 shows the performance of the SVM classifier trained with k top-ranked genes. Columns 3 and 4 explain that the misclassification in the test data was lower as compared to the earlier reports. In contrast with Cho et al. [15], the top-ranked genes also consistently participated in the building of the classifier in the proposed method.

Table 7
Performance of SVM classifier in colon cancer dataset

k	Classifying accuracy in		Misclassification in test data	Comparison of misclassification
	training data	test data		
7	0.88	0.85	1.8 ± 0.34	2.90 ± 0.13 [33]
10	0.88	0.83	2.04 ± 0.54	2.15 ± 1.2 [16]
15	0.95	0.85	1.8 ± 0.21	2.04 ± 0.14 [33]
30	0.98	0.81	2.28 ± 0.42	2.57 ± 1.76 [15]

We noticed the best performance with $k = 15$. Figure 8 illustrates a heatmap of the top-15 genes; 57 out of 62 samples were correctly classified through hierarchical clustering based on these 15 features.

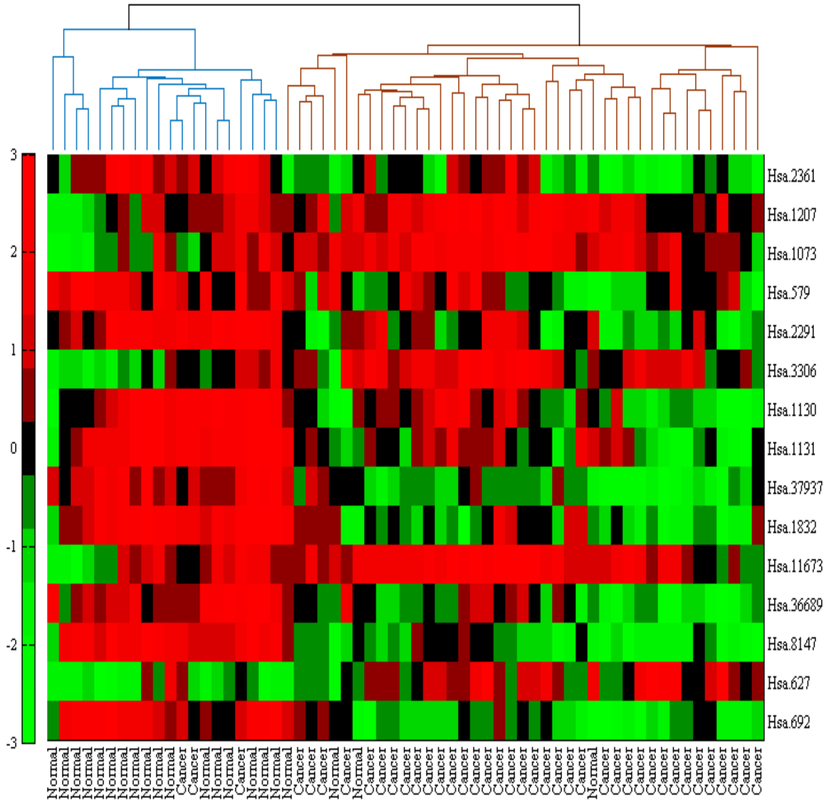

Figure 8. Gene expression heatmap of top-15 genes isolated by *PINK* in colon cancer data – two major sample clusters can be identified merely by visual inspection

Out of the top-ten-ranking genes listed in Table 6, a total of six matched with the previous studies [1, 16, 20, 28]. Four of the top-ranked genes coincided with the wet-lab tests [25, 35, 40, 44] as well. The results of these well-explored datasets show that, even though it was designed for network formation that encompasses multiple phenotypes, the present method can isolate potential genes in a similar fashion to the previous studies.

5.2. Comparison with differential network-based methods

We selected two datasets (OSCC [50] and BLCA [14]) for the purpose of corroborating with the previous findings in the two-step approach of a differential co-expression network. Oral squamous cell carcinoma (OSCC), the most common type of oral cancer, causes damage to oral epithelial cells. It is the major cause of mortality in those patients that suffer from head and neck cancers. We show a list of the top-ten-ranking genes in *PINK* for the OSCC dataset in Table 8. The p-values of the isolated genes are very significant in five-fold cross validation Li et al. [32] employed weighted gene co-expression network analysis (WGCNA) on the same dataset and isolated two sets of novel miRNA associated with OSCC. They reported the let-7c gene as a hub that was ranked third by *PINK*. The same gene was also reported as differentially expressed in [26] and as a tumor-suppressor molecule in [37]. The miR-410 molecule was the second-most-connected component found by Li et al.; it was also reported as differentially co-regulated by Shiah et al. [50]. We obtained the same molecule as being ranked fourth (Tab. 8).

Table 8
Isolated genes from *PINK* along with citations of matching found by WGCNA method and wet-lab tests in OSCC dataset

Genes isolated by PINK	p-value	Identified by WGCNA [32]	Wet-lab tests
miR-21	$9.53e-23$	–	–
miR-30a	$4.20e-18$	–	–
miR-let-7c	$8.08e-22$	✓	[26, 37]
miR-410	$1.55e-14$	✓	–
miR-1267	$1.50e-16$	✓	–
miR-125b	$1.08e-12$	–	[24]
miR-503	$1.54e-20$	–	[54]
miR-7	$3.04e-19$	–	–
miR-99a	$2.05e-13$	–	–
miR-136	$1.23e-13$	–	–

Bladder urothelial carcinoma (BLCA) is one of the most common neoplasms in urological systems. On the dataset of BLCA, we built *PINK* with 165 primary bladder cancers and 58 surrounding normal tissue samples. Table 9 lists the top-ten connected genes. Chen et al. [14] employed the WGCNA model and identified seven co-expressed modules that were related to urothelial bladder cancer (BLCA). Four of the ten isolated genes in *PINK* were present in the modules that were identified by the authors.

Table 9

Isolated genes from *PINK* along with citations of matching found by WGCNA method and wet-lab tests in BLCA dataset

Genes isolated by PINK	p-value	Identified by WGCNA [14]	Wet-lab tests
SLC12A8	$6.84e-15$	✓	–
RPL27A	$1.35e-11$	–	–
HOXB8	$4.59e-09$	–	[39]
DCN	$6.44e-10$	✓	–
SERPINE1	$2.84e-08$	–	[42]
ALDHIL1	$3.42e-09$	–	–
CLIP3	$1.01e-14$	✓	–
TPST1	$1.51e-14$	✓	–
SLC2A3	$1.98e-08$	✓	–
NPHS2	$1.29e-08$	–	–

From the findings of the OSCC and BLCA datasets, it is worth noting that two independent techniques of studies (viz., WGCNA and *PINK*) had distinctly separate methodologies that corroborated each other.

5.3. PINK with datasets of multiple phenotypes

Finally, we apply *PINK* on the datasets of co-existent multiple phenotypes (i.e., $t > 2$) to identify the set of genes that are able to conjugately isolate a subclass of the disease. We selected two datasets (SRBCT and SLE), which were comprised of four and six phenotypes, respectively.

5.3.1. SRBCT dataset

First, we put on a demonstration of how *PINK* achieves similar results to the differential network-based approach in much fewer straightforward steps. Usually, differential networks can isolate genes to distinguish one pair of phenotypes at a time. The task gets complicated as the number of phenotypes increases. On the SRBCT dataset, we exhibited the process of building *PINK* for $t > 2$ by breaking it into $(t-1)$-phases. In the first phase, we built $PINK_{(EWS+BL)}$ and isolated the top-ranking genes (as shown in the first column of Table 10). These genes can be used as an attribute vector for discriminating the EWS samples from the BL samples. We added NB and RMS samples consecutively in Phases 2 and 3. The set of top-ranking genes for the corresponding phases are displayed in Table 10.

Table 10

Top-ranking genes through different phases in SRBCT dataset

Top-ranking genes in		
$PINK_{(EWS+BL)}$	$PINK_{(EWS+BL+NB)}$	$PINK_{(EWS+BL+NB+RMS)}$
CAV1	FGFR4 *	CITED4*
FCGRT	TLE2*	MEST* ([52])
WAS	HOXB7*	MYL4* ([29, 52])
PTPN13	DAPK1*	FCGRT ([20, 29, 41, 52])
DDR2	TNFAIP6*	FVT1 ([20, 29, 41, 52])
KIAA0467	PTPN13	PTPN13 ([29, 52])
FVT1	FVT1	OLFM1* ([29, 41, 52]
MAPK7	GYG2*	TLE2 ([29, 52])
CTNNA1	GSTM5*	FGFR4 ([20, 29, 41, 52])

Note: * indicates that gene is introduced as top-connected in specific phase and was absent in earlier phase(s); citations in last column indicate substantiation of corresponding genes

The same task of isolating phenotype-specific feature attributes following a differential network-based approach would have required us to first build co-expression networks for all of the phenotypes individually. Then, differential networks are required to be extracted for $\binom{4}{2}$ possible pairs of phenotypes; viz., $DN_{EWS\backslash BL}$, $DN_{EWS\backslash NB}$, $DN_{EWS\backslash RMS}$, etc. These differential networks had to be compared to obtain the second-order differential network (such as $DN_{EWS\backslash BL\backslash NB}$, etc.). So, *PINK* definitely simplified the process.

The scatterplots (Fig. 9) explain the appearances of the genes in Table 10. For instance, CAV1 was only up-regulated for the EWS samples (Fig. 9a); it is ranked first in $PINK_{(EWS+BL)}$. Similarly, FGFR4 was specifically up-regulated in the NB phenotypic samples (Fig. 9b), and CITED4 was only up-regulated in the RMS samples (Fig. 9c). Eight of the top rankings were identified by at least one of the previous methods [20, 29, 41, 52] that were applied on this dataset.

We tested whether the top-connected genes could correctly classify the blind test samples into the tumor subclasses. For this, we trained an artificial neural network (ANN) that consisted of ten neurons in the hidden layer. We partitioned training samples randomly into three groups by keeping 15% for testing, another 15% for validation, and the remaining 70% for training the neural network. We used the ANN classifier that was trained with the top-ten connected genes to classify 25 blind test samples. By applying majority voting to categorize the samples, we were able to diagnose 22 samples correctly with high confidence. This number was slightly improved as compared to the 18 samples that were reported in the earlier work [29].

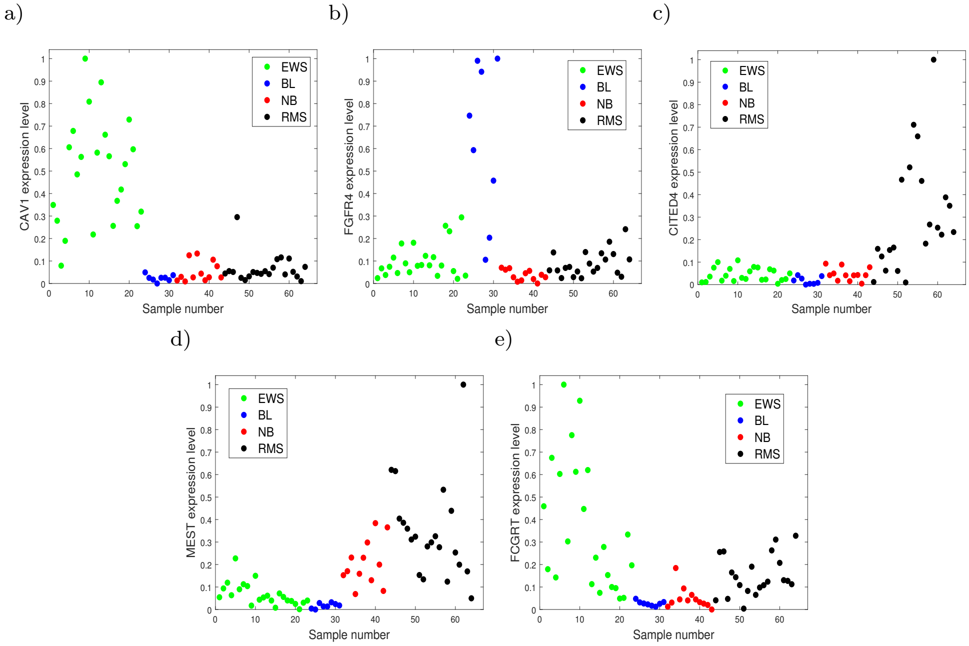

Figure 9. Scatterplots of expression values of selected isolated genes in SRBCT dataset: a) CAV1; b) FGFR4; c) CITED4; d) MEST; e) FCGRT

5.3.2. Dataset: SLE

The SLE dataset contained whole blood RNA from six different inflammatory and infectious diseases. We applied the proposed method on this dataset to find specific transcripts in order to discriminate these inflammatory and infectious diseases. We built a single *PINK* that consisted of six phenotypes ($t = 6$). Table 11 lists the top-connected genes.

Table 11
List of genes isolated from PINK in SLE dataset

Top-ranking genes	p-value	t-score	Reported in	Wet-lab tests
CSAD	$3.22e-36$	–15.73	–	[43]
OAS3	$2.68e-09$	–6.25	[6]	[31]
RSAD2	$1.87e-48$	–20.06	–	[38]
IFI44L	$9.12e-45$	–19.99	[10]	[48]
SPATS2L	$2.02e-45$	–19.84	–	–
OAS2	$1.51e-42$	17.85	[6, 10]	[31]
IFITM3	$3.46e-41$	–17.70	[36, 55]	–
IFI27	$1.15e-30$	–15.08	[10]	[27]
SERPING1	$5.33e-41$	–17.75	–	[34]
PLSCR1	$3.46e-04$	–3.65	–	–

To evaluate the statistical significance of the top-ranking genes, we performed a t-test on them with a five-fold cross validation. As the main target phenotype of the dataset was SLE, a one-vs.-all approach was executed by taking the ASLE and PSLE samples as one class and the rest of the disease phenotypes as another. The p-values and t-scores of the isolated genes were quite significant (as is shown in Table 11).

5.4. Test of validation of results

In this section, we compare the significance of our results with the state-of-the-art methods. We first present a quantitative comparison, followed by an information theoretic approach in order to find the information gain by the top-ranked genes that are isolated by *PINK*.

5.4.1. Quantitative analysis

Our motivation was to isolate a set of genes that work as biomarkers for a corresponding disease. Biomarkers are used to classify diseased and normal samples or the different subtypes of a disease. In order to compare the accuracy of the classifier in the aforementioned task, we trained a classifier with a gene set that was isolated by *PINK* and did the same with the existing state-of-the-art methods. We compared the accuracy with four different methods in total for four datasets where the lists of the isolated genes by different methods were available. Neighborhood component analysis (NCA) is a well-studied method for feature selection. WGCNA was particularly coined by Li et al. [32], and isolated markers are available for the BLCA dataset.

Table 12

Performance evaluation of classifier trained with isolated genes

Dataset	Method	N_g	Classifier	Accuracy [%]
OSCC	NCA	12	Linear SVM	96.0
	PINK	10		100.0
BLCA	NCA	10	Medium Gaussian SVM	74.0
	WGCNA	14		76.7
	PINK	10		88.8
SRBCT	PCA and ANN [29]	96	Naive Bayes	98.4
			Medium Gaussian SVM	100.0
	Shrunken centroids [52]	43	Naive Bayes	96.9
			Medium Gaussian SVM	98.4
	PINK	32	Naive Bayes	96.9
			Medium Gaussian SVM	100.0
SLE	NCA	11	Medium Gaussian SVM	77.0
	Random Forest [2]	17		69.0
	Statistical approach [10]	10		66.8
	PINK	10		85.0

Note: N_g is number of isolated genes employed as attributes in classifier; principal component analysis is abbreviated "PCA"

Principal component analysis (PCA) and an artificial neural network-based model (ANN) with 96 genes was used by Khan et al. [29] on the SRBCT dataset. Successively shrunken centroids of the gene expression model were proposed by Tibshirani et al. [52] for the same dataset. The random forest algorithm and statistical methods (fold change, the t-test p-value, and the false discovery rate p-value) were applied on the SLE dataset by [2] and [10], respectively. We used five-fold cross validation to evaluate the model's performance. To avoid any bias, we employed the same classifier while using different gene-selection methods on a single dataset. We used the naive Bayes and support vector machine (SVM) methods with linear and medium Gaussian kernels as classifiers. Table 12 displays the comparative results. It is apparent from these results that *PINK* helped the classifiers attain the same level of accuracy as its prevalent methods but with much smaller sets of biomarkers.

5.4.2. Information gain

We assess the validity of *PINK* by measuring the significance of its top-ranking genes by using an information theoretic approach. The synergy value (S) [11] between a pair of genes is the gain in the discriminating power between phenotypes when both are considered jointly as compared to that of the individuals. We tested the significance of isolated genes by their discriminating power. We built the complete graph $\mathbb{G}$ with genes as nodes and $1/S$ as the corresponding edge weight. Subsequently, we found the minimum spanning tree (MST) of $\mathbb{G}$. As $1/S$ is assigned as the edge weight, the MST represents the maximal total synergistic value. Let T be the weight of the MST. We constructed a subgraph ($\mathbb{G}'$) that consisted of the n top-ranking genes along with their connectives ($\{n' | \exists edge(n' \rightarrow n)\}$). The MST of $\mathbb{G}'$ was extracted as well. Let T' be the weight of this MST. We took ratio $\frac{T'}{T}$ as the significance (Ψ) of *PINK*. The impact of the MST of *PINK* for all of the datasets is shown in Table 13. In all of the datasets, Ψ was greater than 0.9 with $n = 10$. This implies that the subgraph that was formed with only the top-ten-ranking genes in *PINK* contained most of the high-synergy-value pairs and carried more than 90% of the knowledge that the complete graph had.

Table 13
Significance of MST with top-ten-ranked genes of *PINK*

Dataset	Ψ
ALL-AML	0.931
Colon cancer	0.916
BLCA	0.951
OSCC	0.943
SRBCT	0.947
SLE	0.938

5.5. Biological significance of selected genes

In this section, we explore the biological significance of the isolated genes. Most of the isolated genes have corroborations in wet-lab tests such as northern blot analysis, qRT-PCR, etc. Tables 4–11 contain references to the corresponding wet-lab reports. Some of the isolated genes are part of the significant biological pathways in the KEGG pathway database (https://www.kegg.jp/kegg/pathway.html) as shown in Table 14.

Table 14

KEGG pathway report of isolated genes from *PINK*

Dataset	Gene	Relevant KEGG Pathways
ALL-AML	GLUL	Necroptosis
	LYN-Vy1	B cell receptor signaling
	RBP P48	Viral carcinogenesis
	CD19	Hematopoietic cell lineage, B cell receptor signaling, PI3K-Akt signaling, Epstein-Barr virus infection, Primary immunodeficiency
OSCC	miR-30a	Proteoglycans in cancer (map05205), MicroRNAs in cancer (map05206)
	miR-let-7c	MicroRNAs in cancer (map05206)
	miR-125b	MicroRNAs in cancer (map05206)
	miR-7	MicroRNAs in cancer (map05206)
BLCA	SERPINE1	HIF-1 signaling pathway (map04066), Cellular senescence (map04218)
SRBCT	PTPN13	Apoptosis (map04210)
	MYL4	Apelin signaling pathway (map04371)
	FGFR4	MAPK signaling pathway (map04010), Pathways in cancer (map05200), Endocytosis (map04144)
	CAV1	Focal adhesion (map04510), Endocytosis (map04144)
SLE	CSAD	Taurine and hypotaurine metabolism (map00430), Metabolic (map0110)
	RSAD2	Hepatitis C (map05160), Influenza A (map05164)
	SERPING1	Complement and coagulation cascades (map04610), Pertussis (map05133)

Among the newly isolated genes in ALL-AML, GLUL has been found to play role in removing NH_4^+ by incorporating it into glutamine; this is integral to dex-induced catabolism in B-ALL cells [18]. CD19 was reported to be a biomarker for B cell development by Wang et al. [53].

In colon cancer, Myo1a (Hsa.37937) was reported to be a reason for increased tumorigenicity [40]. Fibroblast TM1 (Hsa.1131) was reported as differentially expressed in [44]. Northern blot analysis showed that tropomyosin isoform (Hsa.1130) was preferentially associated with colon cancer in [35].

In the case of OSCC, miR-125b was reported to cause malignancy in oral cells in [24]. miR-503 was reported to be a suppressor gene in squamous cell carcinoma in [54].

In BLCA, Hox genes were found to be associated with the development of bladder cancer by Real Time-PCR [39]. SERPINE1 is a part of the HIF-1 signaling pathway. This pathway has been found to enhance the malignant nature of bladder cancer cells [42]; it also plays a role in tumor suppression, as it is a part of the cellular senescence pathway.

In the case of SRBCT, the MEST gene that is ranked second was discovered as potential marker that is specific for RMS by RT-PCR [4]. Figure 9d shows that it is particularly up-regulated in RMS samples. In the same study [4], FCGRT was reported to be target genes for EWS samples. Figure 9e shows that FCGRT is up-regulated – particularly in EWS samples. PPTN13 is a part of the apoptosis pathway, which is responsible for cell growth and death. FGFR4 is part of MAPK signaling and cancer pathways. CAV1 is part of the focal adhesion pathway. This pathway is reported to play an important role in Ewing sarcoma [12]. Figure 9a shows that CAV1 is highly expressed in EWS samples.

In the case of SLE, [31] confirmed that the roles of OAS3 and OAS2 as mediators in the innate immune response to infection may be important. The CSAD gene is a part of the taurine and hypotaurine metabolism pathway. This is one of the most affected pathways because of SLE [43]. IFITM3 was reported to be significantly up-regulated in patients with SLE [55]. RT-PCR also found that transcription levels of the IFI27 gene was significantly increased in SLE patients [27]. The SERPING1 gene is part of the complement and coagulation cascades pathway. A study made on this pathway discovered that it contributes to the severity of SLE [34]. PLSCR1 mRNA was found to be significantly increased in SLE samples when compared to normal ones [51].

6. Conclusion

In the mining of microarray data, the precise identification of genes that have discriminating expression patterns across different classes is the most coveted task. This helps in disease diagnosis and targets specific drug discovery. Studying the differential network is a well-established approach for isolating the otherwise expressed genes as they are related to a phenotype, but this requires the generation of separate networks for the phenotypes under study and then comparing them.

In the present study, we propose a phenotype interwoven network (*PINK*) method that accomplishes the task in a single network for multiple phenotypes, thereby setting aside the cumbersome and tedious steps of comparing multiple networks. The method encompasses both linear and nonlinear correlations between gene pairs; it also considers any inherent fuzziness in expression data and does not require any sort of defuzzification. As a result, several of the new top-ranking genes reported here have found place in prominent wet-lab tests, and a few have been corroborated with previous works. Moreover, *PINK* unfolds the connectivity profiles of isolated genes that might help in studying gene co-expression and co-regulation patterns under phenotypic changes.

Declaration

Conflicts of interest The authors do not have any known sources of conflicts of interests.

References

[1] Alladi S.M., Shinde Santosh P., Ravi V., Murthy U.S.: Colon cancer prediction with genetic profiles using intelligent techniques, *Bioinformation*, vol. 3(3), pp. 130–133, 2008.

[2] Almlöf J.C., Alexsson A., Imgenberg-Kreuz J., Sylwan L., Bäcklin C., Leonard D., Nordmark G., Tandre K., *et al.*: Novel risk genes for systemic lupus erythematosus predicted by random forest classification, *Scientific Reports*, vol. 7(1), pp. 1–11, 2017.

[3] Alon U., Barkai N., Notterman D.A., Gish K., Ybarra S., Mack D., Levine A.J.: Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays, *Proceedings of the National Academy of Sciences*, vol. 96(12), pp. 6745–6750, 1999.

[4] Baer C., Nees M., Breit S., Selle B., Kulozik A.E., Schaefer K.L., Braun Y., Wai D., Poremba C.: Profiling and functional annotation of mRNA gene expression in pediatric rhabdomyosarcoma and Ewing's sarcoma, *International Journal of Cancer*, vol. 110(5), pp. 687–694, 2004. doi: 10.1002/ijc.20171.

[5] Bernusso V.A., Machado-Neto J.A., Pericole F.V., Vieira K.P., Duarte A.S., Traina F., Hansen M.D., Saad S.T.O., Barcellos K.S.: Imatinib restores VASP activity and its interaction with Zyxin in BCR–ABL leukemic cells, *Biochimica et Biophysica Acta (BBA) – Molecular Cell Research*, vol. 1853(2), pp. 388–395, 2015.

[6] Berry M.P.R., Berry M., Graham C.M., McNab F.W., Xu Z., Bloch S.A.A., Oni T., Wilkinson K.A., Banchereau R., Skinner J., Wilkinson R.J., Quinn C., Blankenship D., Dhawan R., Cush J.J., Mejias A., Ramilo O., Kon O.M., Pascual V., Banchereau J., Chaussabel D., O'Garra A., Bloch S.: An interferon-inducible neutrophil-driven blood transcriptional signature in human tuberculosis, *Nature*, vol. 466(7309), pp. 973–977, 2010. doi: 10.1038/nature09247.

[7] Bhattacharyya R.: Cohesion: A concept and framework for confident association discovery with potential application in microarray mining, *Applied Soft Computing*, vol. 11(1), pp. 592–604, 2011.

[8] Bhattacharyya R.: Analyzing deviation pattern in strongly-correlated genes through core cluster mining, *Information Sciences*, vol. 251, pp. 47–62, 2013.

[9] Bhattacharyya R., Bhattacharyya B.: Gene Expression Mining for Cohesive Pattern Discovery. In: M. Elloumi, J. Küng, M. Linial, R. Murphy, K. Schneider, C. Toma (eds.), *Bioinformatics Research and Development*, *Communications in Computer and Information Science*, vol. 13, pp. 221–234, Springer, Berlin–Heidelberg. doi: 10.1007/978-3-540-70600-7_17.

[10] Bing P.F., Xia W., Wang L., Zhang Y.H., Lei S.F., Deng F.Y.: Common Marker Genes Identified from Various Sample Types for Systemic Lupus Erythematosus, *PLOS ONE*, vol. 11(6), p. e0156234, 2016. doi: 10.1371/journal.pone.0156234.

[11] Chatterjee P., Pal N.R.: Discovery of synergistic genetic network: A minimum spanning tree-based approach, *Journal of Bioinformatics and Computational Biology*, vol. 14(1), pp. 2–4, 2016. doi: 10.1142/S0219720016500037.

[12] Chaturvedi A., Hoffman L.M., Jensen C.C., Lin Y.C., Grossmann A.H., Randall R.L., Lessnick S.L., Welm A.L., Beckerle M.C.: Molecular dissection of the mechanism by which EWS/FLI expression compromises actin cytoskeletal integrity and cell adhesion in Ewing sarcoma, *Molecular Biology of the Cell*, vol. 25(18), pp. 2695–2709, 2014.

[13] Chen D., Zhang L., Zhao S., Hu Q., Zhu P.: A novel algorithm for finding reducts with fuzzy rough sets, *IEEE Transactions on Fuzzy Systems*, vol. 20(2), pp. 385–389, 2012.

[14] Chen Z., Liu G., Hossain A., Danilova I.G., Bolkov M.A., Liu G., Tuzankina I.A., Tan W.: A co-expression network for differentially expressed genes in bladder cancer and a risk score model for predicting survival, pp. 1–11, 2019.

[15] Cho J.H., Lee D., Park J.H., Lee I.B.: New gene selection method for classification of cancer subtypes considering within-class variation, *FEBS Letters*, vol. 551(1–3), pp. 3–7, 2003.

[16] Cho J.H., Lee D., Park J.H., Lee I.B.: Gene selection and classification from microarray data using kernel machine, *FEBS Letters*, vol. 571(1–3), pp. 93–98, 2004.

[17] Deb S., Mahanta P., Bhattacharyya D.K., Dutta M.A.: Subspace module extraction from MI-based co-expression network, *International Journal of Bioinformatics Research and Applications*, vol. 14(3), pp. 207–234, 2018.

[18] Dyczynski M., Vesterlund M., Björklund A.C., Zachariadis V., Janssen J., Gallart-Ayala H., Daskalaki E., *et al.*: Metabolic reprogramming of acute lymphoblastic leukemia cells in response to glucocorticoid treatment, *Cell Death & Disease*, vol. 9(9), p. 846, 2018.

[19] El Akadi A., Amine A., El Ouardighi A., Aboutajdine D.: A two-stage gene selection scheme utilizing MRMR filter and GA wrapper, *Knowledge and Information Systems*, vol. 26(3), pp. 487–500, 2011.

[20] Fu L.M., Fu-Liu C.S.: Evaluation of gene importance in microarray data based upon probability of selection, *BMC Bioinformatics*, vol. 6(1), pp. 1–11, 2005.

[21] Furey T.S., Cristianini N., Duffy N., Bednarski D.W., Schummer M., Haussler D.: Support vector machine classification and validation of cancer tissue samples using microarray expression data, *Bioinformatics*, vol. 16(10), pp. 906–914, 2000.

[22] Golub T.R., Slonim D.K., Tamayo P., Huard C., Gaasenbeek M., Mesirov J.P., Coller H., *et al.*: Molecular classification of cancer: class discovery and class prediction by gene expression monitoring, *Science*, vol. 286(5439), pp. 531–537, 1999.

[23] Guyon I., Weston J., Barnhill S., Vapnik V.: Gene selection for cancer classification using support vector machines, *Machine Learning*, vol. 46(1–3), pp. 389–422, 2002.

[24] Henson B.J., Bhattacharjee S., O'Dee D.M., Feingold E., Gollin S.M.: Decreased expression of miR-125b and miR-100 in oral cancer cells contributes to malignancy, *Genes, Chromosomes and Cancer*, vol. 48(7), pp. 569–582, 2009.

[25] Hill O., Cetin Y., Cieslak A., Mägert H.J., Forssmann W.G.: A new human guanylate cyclase-activating peptide (GCAP-II, uroguanylin): precursor cDNA and colonic expression, *Biochimica et Biophysica Acta (BBA) – Protein Structure and Molecular Enzymology*, vol. 1253(2), pp. 146–149, 1995.

[26] Hui A.B., Lenarduzzi M., Krushel T., Waldron L., Pintilie M., Shi W., Perez-Ordonez B., *et al.*: Comprehensive MicroRNA profiling for head and neck squamous cell carcinomas, *Clinical Cancer Research*, vol. 16(4), pp. 1129–1139, 2010.

[27] Ishii T., Onda H., Tanigawa A., Ohshima S., Fujiwara H., Mima T., Katada Y., *et al.*: Isolation and Expression Profiling of Genes Upregulated in the Peripheral Blood Cells of Systemic Lupus Erythematosus Patients, *DNA Research*, vol. 12(6), pp. 429–439, 2005. doi: 10.1093/dnares/dsi020.

[28] Jing L., Ng M.K., Zeng T.: Novel hybrid method for gene selection and cancer prediction, *World Academy of Science, Engineering and Technology*, vol. 62(89), pp. 482–489, 2010.

[29] Khan J., Wei J.S., Ringnér M., Saal L.H., Ladanyi M., Westermann F., Berthold F., *et al.*: Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks, *Nature Medicine*, vol. 7(6), pp. 673–679, 2001.

[30] Landi M.T., Dracheva T., Rotunno M., Figueroa J.D., Liu H., Dasgupta A., Mann F.E., *et al.*: Gene Expression Signature of Cigarette Smoking and Its Role in Lung Adenocarcinoma Development and Survival, *PLoS ONE*, vol. 3(2), p. e1651, 2008. doi: 10.1371/journal.pone.0001651.

[31] Leisching G., Wiid I., Baker B.: OAS1, 2, and 3: Significance During Active Tuberculosis?, *The Journal of Infectious Diseases*, vol. 217(10), pp. 1517–1521, 2018. doi: 10.1093/infdis/jiy084.

[32] Li J., Zhou D., Qiu W., Shi Y., Yang J.J., Chen S., Wang Q., Pan H.: Application of Weighted Gene Co-expression Network Analysis for Data from Paired Design, *Scientific Reports*, vol. 8(1), pp. 1–8, 2018. doi: 10.1038/s41598-017-18705-z.

[33] Li Y., Campbell C., Tipping M.: Bayesian automatic relevance determination algorithms for classifying gene expression data, *Bioinformatics*, vol. 18(10), pp. 1332–1339, 2002.

[34] Liang Y., Xie S.B., Wu C.H., Hu Y., Zhang Q., Li S., Fan Y.G., *et al.*: Coagulation cascade and complement system in systemic lupus erythematosus, *Oncotarget*, vol. 9(19), pp. 14862–14881, 2018. doi: 10.18632/oncotarget.23206.

[35] Lin J.L.C., Geng X., Bhattacharya S.D., Yu J.R., Reiter R.S., Sastri B., Glazier K.D., *et al.*: Isolation and sequencing of a novel tropomyosin isoform preferentially associated with colon cancer, *Gastroenterology*, vol. 123(1), pp. 152–162, 2002.

[36] Maertzdorf J., McEwen G., Weiner J., Tian S., Lader E., Schriek U., Mayanja-Kizza H., Ota M., Kenneth J., Kaufmann S.H.: Concise gene signature for point of care classification of tuberculosis, *EMBO Molecular Medicine*, vol. 8(2), pp. 86–95, 2016. doi: 10.15252/emmm.201505790.

[37] Manikandan M., Rao A.K.D.M., Arunkumar G., Manickavasagam M., Rajkumar K.S., Rajaraman R., Munirajan A.K.: Oral squamous cell carcinoma: microRNA expression profiling and integrative analyses for elucidation of tumourigenesis mechanism, *Molecular Cancer*, vol. 15(1), p. 28, 2016.

[38] Manzanillo P.S., Shiloh M.U., Portnoy D.A., Cox J.S.: Mycobacterium tuberculosis activates the DNA-dependent cytosolic surveillance pathway within macrophages, *Cell Host and Microbe*, vol. 11(5), pp. 469–480, 2012. doi: 10.1016/j.chom.2012.03.007.

[39] Marra L., Cantile M., Scognamiglio G., Perdonà S., La Mantia E., Cerrone M., Gigantino V., *et al.*: Deregulation of HOX B13 expression in urinary bladder cancer progression, *Current Medicinal Chemistry*, vol. 20(6), pp. 833–839, 2013.

[40] Ouderkirk J.L., Krendel M.: Non-muscle myosins in tumor progression, cancer cell invasion, and metastasis, *Cytoskeleton*, vol. 71(8), pp. 447–463, 2014.

[41] Pal N.R., Aguan K., Sharma A., Amari S.i.: Discovering biomarkers from gene expression data for predicting cancer subgroups using neural networks and relational fuzzy clustering, *BMC Bioinformatics*, vol. 8(1), p. 5, 2007.

[42] Peixoto A., Fernandes E., Gaiteiro C., Lima L., Azevedo R., Soares J., Cotton S., *et al.*: Hypoxia enhances the malignant nature of bladder cancer cells and concomitantly antagonizes protein *O*-glycosylation extension, *Oncotarget*, vol. 7(39), pp. 63138–63157, 2016. doi: 10.18632/oncotarget.11257.

[43] Perl A., Hanczko R., Lai Z.W., Oaks Z., Kelly R., Borsuk R., Asara J.M., Phillips P.E.: Comprehensive metabolome analyses reveal N-acetylcysteine-responsive accumulation of kynurenine in systemic lupus erythematosus: implications for activation of the mechanistic target of rapamycin, *Metabolomics*, vol. 11(5), pp. 1157–1174, 2015. doi: 10.1007/s11306-015-0772-0.

[44] Prasad G., Meissner S., Sheer D.G., Cooper H.L.: A cDNA encoding a muscle-type tropomyosin cloned from a human epithelial cell line: identity with human fibroblast tropomyosin TM1, *Biochemical and biophysical research communications*, vol. 177(3), pp. 1068–1075, 1991.

[45] Roy S., Bhattacharyya D.K., Kalita J.K.: Reconstruction of gene co-expression network from microarray data using local expression patterns, *BMC Bioinformatics*, vol. 15(7), pp. 1–14, 2014.

[46] Ruan J., Dean A.K., Zhang W.: A general co-expression network-based approach to gene expression analysis: comparison and applications, *BMC Systems Biology*, vol. 4(1), pp. 1–21, 2010.

[47] Sadhu A., Bhattacharyya B.: Common Subcluster Mining in Microarray Data for Molecular Biomarker Discovery, *Interdisciplinary Sciences: Computational Life Sciences*, vol. 11(3), 2019. doi: 10.1007/s12539-017-0262-3.

[48] Sambarey A., Devaprasad A., Mohan A., Ahmed A., Nayak S., Swaminathan S., *et al.*: Unbiased Identification of Blood-based Biomarkers for Pulmonary Tuberculosis by Modeling and Mining Molecular Interaction Networks, *EBioMedicine*, vol. 15, pp. 112–126, 2017. doi: 10.1016/j.ebiom.2016.12.009.

[49] Sandberg Y., Kallemeijn M.J., Dik W.A., Tielemans D., Wolvers-Tettero I.L.M., van Gastel-Mol E.J., Szczepanski T., *et al.*: Lack of common TCRA and TCRB clonotypes in CD8+/TCR$\alpha\beta$+ T-cell large granular lymphocyte leukemia: a review on the role of antigenic selection in the immunopathogenesis of CD8+ T-LGL, *Blood Cancer Journal*, vol. 4(1), pp. e172–e172, 2014. doi: 10.1038/bcj.2013.70.

[50] Shiah S.G., Hsiao J.R., Chang W.M., Chen Y.W., Jin Y.T., Wong T.Y., Huang J.S., *et al.*: Downregulated miR329 and miR410 promote the proliferation and invasion of oral squamous cell carcinoma by targeting Wnt-7b, *Cancer Research*, vol. 74(24), pp. 7560–7572, 2014. doi: 10.1158/0008-5472.CAN-14-0978.

[51] Suzuki E., Amengual O., Atsumi T., Oku K., Hashimoto T., Kataoka H., Horita T., *et al.*: Increased expression of phospholipid scramblase 1 in monocytes from patients with systemic lupus erythematosus, *The Journal of Rheumatology*, vol. 37(8), pp. 1639–1645, 2010.

[52] Tibshirani R., Hastie T., Narasimhan B., Chu G.: Diagnosis of multiple cancer types by shrunken centroids of gene expression, *Proceedings of the National Academy of Sciences*, vol. 99(10), pp. 6567–6572, 2002.

[53] Wang K., Wei G., Liu D.: CD19: a biomarker for B cell development, lymphoma diagnosis and therapy, *Experimental Hematology & Oncology*, vol. 1(1), pp. 2–7, 2012.

[54] Wu J., Gao F., Xu T., Deng X., Wang C., Yang X., Hu Z., *et al.*: miR-503 suppresses the proliferation and metastasis of esophageal squamous cell carcinoma by triggering autophagy via PKA/mTOR signaling, *International Journal of Oncology*, vol. 52(5), pp. 1427–1442, 2018.

[55] Ye H., Wang X., Wang L., Chu X., Hu X., Sun L., Jiang M., *et al.*: Full high-throughput sequencing analysis of differences in expression profiles of long noncoding RNAs and their mechanisms of action in systemic lupus erythematosus, *Arthritis Research & Therapy*, vol. 21(1), pp. 1–17, 2019. doi: 10.1186/s13075-019-1853-7.

Affiliations

Arnab Sadhu
Visva-Bharati University, Department of Computer and System Sciences, Santiniketan, West Bengal 731235, India; arnabsadhu30@gmail.com

Balaram Bhattacharyya
Visva-Bharati University, Department of Computer and System Sciences, Santiniketan, West Bengal 731235, India; balaramb@gmail.com

Tathagato Mukhopadhyay
Visva-Bharati University, Department of Computer and System Sciences, Santiniketan, West Bengal 731235, India; toto.mukh@gmail.com

Received: 16.06.2021
Revised: 04.10.2021
Accepted: 13.11.2021