Identification of miRNA Biomarkers for Diverse Cancer Types Using Statistical Learning Methods at the Whole-Genome Scale

Jnanendra Prasad Sarkar 1,2†, Indrajit Saha 3,4*, Adrian Lancucki 4†, Nimisha Ghosh 5, Michał Własnowolski 6, Grzegorz Bokota 7,8, Ashmita Dey 2, Piotr Lipinski 4 and Dariusz Plewczynski 6,8*

1 Data, Analytics & AI, Larsen & Toubro Infotech Ltd., Pune, India, 2 Department of Computer Science & Engineering, Jadavpur University, Kolkata, India, 3 Department of Computer Science and Engineering, National Institute of Technical Teachers’ Training and Research, Kolkata, India, 4 Computational Intelligence Research Group, Institute of Computer Science, University of Wrocław, Wrocław, Poland, 5 Department of Computer Science and Information Technology, SOA University, Bhubaneswar, India, 6 Faculty of Mathematics and Information Science, Warsaw University of Technology, Warsaw, Poland, 7 Institute of Informatics, University of Warsaw, Warsaw, Poland, 8 Centre of New Technologies, University of Warsaw, Warsaw, Poland

*Correspondence:
Indrajit Saha
indrajit@nitttrkol.ac.in
Dariusz Plewczynski
d.plewczynski@cent.uw.edu.pl
† These authors have contributed equally to this work

Genome-wide analysis of miRNA molecules can reveal important information for understanding the biology of cancer. Typically, miRNAs are used as features in statistical learning methods in order to train learning models to predict cancer. This motivates us to propose a method that integrates clustering and classification techniques for diverse cancer types with survival analysis via regression to identify miRNAs that can potentially play a crucial role in the prediction of different types of tumors. Our method has two parts. The first part is a feature selection procedure, called the stochastic covariance evolutionary strategy with forward selection (SCES-FS), which is developed by integrating stochastic neighbor embedding (SNE), the covariance matrix adaptation evolutionary strategy (CMA-ES), and classifiers, with the primary objective of selecting biomarkers. SNE is used to reorder the features by performing an implicit clustering with highly correlated neighboring features. A subset of features is selected heuristically to perform multi-class classification for diverse cancer types. In the second part of our method, the most important features identified in the first part are used to perform survival analysis via Cox regression, primarily to examine the effectiveness of the selected features. For this purpose, we have analyzed next generation sequencing data from The Cancer Genome Atlas in form of miRNA expression of 1,707 samples of 10 different cancer types and 333 normal samples. The SCES-FS method is compared with well-known feature selection methods and it is found to perform better in multi-class classification for the 17 selected miRNAs, achieving an accuracy of 96%. Moreover, the biological significance of the selected miRNAs is demonstrated with the help of network analysis, expression analysis using hierarchical clustering, KEGG pathway analysis, GO enrichment analysis, and...
1. INTRODUCTION

MicroRNAs (miRNAs) belong to the non-coding RNA family. They consist of 19–25 nucleotides and play an important role in the regulation of gene silencing. These non-coding RNAs are present in every eukaryotic cell and can also be encoded by a viral genome (Ray and Maiti, 2015; Bruscella et al., 2017). The miRNAs are formed by RNA polymerase II in the cell nucleus and are then transferred to the cytoplasm (Bartel, 2009) for biological activities such as cell cycle control, apoptosis, and oncogenesis. They interact with the complementary strand of mRNAs and lead to the degradation of the corresponding miRNAs; they also interfere with protein production by suppressing protein synthesis (Valencia-Sanchez et al., 2006). A miRNA molecule can bind one or more targets, thus forming a complex underlying regulatory network. These networks have a profound impact on cancer signaling pathways (Wang et al., 2017). Previously, low-throughput and high-cost technologies were the main obstacle to answering systems-level biological questions. However, recent advancements in next generation sequencing (NGS) have enabled researchers to address such complex problems (van Dijk et al., 2014). Moreover, new sequencing technologies and genomic datasets have helped us to gain better understanding of the biological complexities related to genomic abnormalities in cancer. The considerable achievements in sequencing techniques have made high-throughput techniques a fundamental platform for miRNA, RNA, and DNA research. Generally, miRNAs are involved in a wide range of diseases, including neurological disease, heart disease, and cancer (Giza et al., 2014; Paul et al., 2018). In many cases of cancer in humans, dysregulation of miRNA expression has been observed, and it is well-known that miRNAs can serve as potential cancer biomarkers (Lu et al., 2005; Jacobsen et al., 2013; Wong et al., 2017). In this regard, scientific communities are also trying to understand the role of miRNAs in paring with mRNAs (Zhang et al., 2014; Shrestha et al., 2017), in different cancer types by ranking miRNAs (Li et al., 2014) to elucidate their effects and drug resistance (Ma et al., 2010; Li and Yang, 2014; Cheerla and Gevaert, 2017).

To reduce the time taken for clinical trials, and to provide better and more accurate treatments while avoiding unnecessary interventions, the proper selection of miRNAs as biomarkers is crucial. For this purpose, miRNAs are often used as features in statistical learning methods viz. clustering, classification, and regression in order to identify potential biomarkers (Song et al., 2016; Yang et al., 2017; Yokoi et al., 2017). Song et al. (2016) performed a clustering analysis on breast cancer data in order to find miRNAs that could be prognostic biomarkers; these miRNAs are up-regulated in this type of cancer and are linked to local relapse, distant metastasis, and poor clinical outcomes. Similarly, Yang et al. (2017) used clustering to find miRNA biomarkers for breast cancer. The identified miRNAs have higher specificity and sensitivity than single-gene biomarkers. On the other hand, Yokoi et al. (2017) used a classification task to inform the development of a predictive model to distinguish patients with ovarian cancer tumors from healthy subjects. In this study, eight miRNAs were found as biomarkers for ovarian cancer. Jacob et al. (2017) conducted a study on colon cancer and identified 16 miRNA signatures, which act as prognostic biomarkers at cancer stages II and III. Apart from the aforementioned works, regression analysis has been used to predict the survival rates of patients with different types of cancer (Liang et al., 2018). Liang et al. (2018) used Cox regression analysis for pancreatic cancer and identified five miRNAs as independent prognostic factors. The progress in this regard can be found in the literature (e.g., Peng and Croce, 2016; Hosseinahli et al., 2018).

Generally, in miRNA-based cancer studies, statistical learning algorithms viz. clustering, classification, and regression are used separately for different cancer types, as described above and in the literature (Akhtar et al., 2015; Ang et al., 2016; Li et al., 2016; Lin and Lane, 2017). However, to leverage the advantages of the different algorithms, it may be useful to integrate them into a single method for identifying potential biomarker miRNAs. Besides, little work exists on multi-class classification of diverse cancer types using NGS data. These two facts motivated us to develop the method described in this paper, which can not only classify 10 types of cancer (bladder, breast, colon, glioblastoma, head and neck squamous cell, kidney renal clear cell, lung adenocarcinoma, lung squamous cell, ovarian, and uterine corpus endometrial carcinoma) but also find putative miRNAs that are highly associated with these cancer types. The proposed two-part wrapper-based feature selection method, referred to as the stochastic covariance evolutionary strategy with forward selection (SCES-FS), uses stochastic neighbor embedding (SNE) (Hinton and Roweis, 2003) in conjunction with the covariance matrix adaptation evolutionary strategy (CMA-ES) (Hansen et al., 2003) and a simple classifier, either random forest (RF) (Breiman, 2001), support vector machine (SVM) (Cortes and Vapnik, 1995), naive Bayes (NB) classifier (George and Langley, 1995), K-nearest-neighbors (K-NN) classifier (Altman, 1992), or decision tree (DT) (Quinlan, 1986), in the first part. Here SNE is used to reorder the features by performing an implicit clustering, such that neighboring features are highly correlated. Then, from these clusters of features, a subset of features is selected randomly in order to perform the multi-class...
classification task for the diverse cancer types. However, as the features are randomly selected, this classification task is treated as an underlying optimization problem for CMA-ES to find the features automatically. Hence, the final set of features/miRNAs is obtained by using forward selection (Whitney, 1971).

In the second part of the method, survival analysis is performed using Cox regression (Cox, 1972), where the levels of miRNA expression and corresponding clinical data are used. The experiment is conducted with data collected from The Cancer Genome Atlas (TCGA)\(^1\) for 10 different types of cancer. The performance of the proposed wrapper-based feature selection method is compared with the following methods in terms of classification accuracy, with the top 17 miRNAs selected as putative biomarkers: ensemble SVM-recursive feature elimination (ESVM-RFE) (Anaissi et al., 2016), the least absolute shrinkage and selection operator (LASSO) (Tibshirani, 1996), the non-dominated sorting genetic algorithm II-based stacked ensemble (NSGA-II-SE) (Saha et al., 2017), the SVM-wrapped multi-objective genetic algorithm (MOGA) (Mukhopadhyay and Maulik, 2013), SVM-based novel recursive feature elimination (SVM-nRFE) (Peng et al., 2009), SVM recursive feature elimination (SVM-RFE) (Guyon et al., 2002), conditional mutual information (CMIM) (Fleuret, 2004), interaction capping (ICAP) (Jakulin, 2005), smoothly clipped absolute deviation (SCAD) (Fan and Li, 2001), joint mutual information (JMI) (Bennasar et al., 2015), conditional infomax feature extraction (CIFE) (Brown et al., 2012), minimum redundancy maximum relevance (mRMR) (Peng et al., 2005), feature selection with Cox regression (FSCOX) (Kim et al., 2014), double-input symmetrical relevance (DISR) (Brown et al., 2012), signal-to-noise ratios (SNRs) (Mishra and Sahu, 2011), and the Wilcoxon rank-sum test (RankSum) (Troyanskaya et al., 2002). Thereafter, the significance of the 17 selected miRNAs to 10 different cancer types is determined using Cox regression analysis. Finally, survival analysis, network analysis, expression analysis using hierarchical clustering in the form of heatmaps, KEGG pathway analysis (Kanehisa and Goto, 2000), gene ontology (GO) enrichment analysis (Kuleshov et al., 2016), and protein-protein interaction (PPI) network analysis (Szklarczyk et al., 2019) are performed to assess the biological significance of the selected miRNAs. Additionally, a web-based cancer predictor application is developed to predict 10 different types of cancer given the expression of 17 miRNAs.

2. MATERIALS AND METHODS

In this section, we briefly describe SNE (Hinton and Roweis, 2003), CMA-ES (Hansen, 2006), and Cox regression analysis (Cox, 1972). More details about classification techniques and feature selection methods are given in the Supplementary Material\(^2\) for this article. This section also describes the proposed method, which consists of two parts. The first part is the wrapper-based SCES-FS. In the second part, the selected features, such as expression of miRNAs and clinical data, are used in survival analysis to assess the importance of the selected miRNAs in different types of cancer and to evaluate the effectiveness of the selected miRNAs. Figure 1 shows the flowchart of the proposed method.

2.1. Stochastic Neighbor Embedding

Let \(\mathcal{X} = \{x_1, x_2, \ldots, x_n\}\) denote a set of \(n\) observations, where \(x_i \in \mathbb{R}^d\). SNE (Hinton and Roweis, 2003) constructs a low-dimensional embedding that recreates \(\mathcal{X}\) in a space of lower dimension as \(\mathcal{X}' = \{x'_1, x'_2, \ldots, x'_n\}\), where \(x'_i \in \mathbb{R}^d\). In SNE, both \(\mathcal{X}\) and \(\mathcal{X}'\) are represented as discrete probability distributions \(P\) and \(Q\), where

\[
p_{ij} = \frac{\exp\left(-\frac{\|x_i - x_j\|^2}{2 \text{var}_i}\right)}{\sum_{k \neq i} \exp\left(-\frac{\|x_i - x_k\|^2}{2 \text{var}_i}\right)}, \quad q_{ij} = \frac{\exp\left(-\|x'_i - x'_j\|^2\right)}{\sum_{k \neq j} \exp\left(-\|x'_i - x'_k\|^2\right)},
\]

that model pairwise distances between data points. The values of \(\text{var}_i \in \mathbb{R}\) are adjusted in such a way that the entropies of all distributions \(P_i\) are equal.

The mismatch between \(P\) and \(Q\) is reduced through minimization of Kullback-Leibler (KL) divergence objective \(KL(P||Q) = \sum_i \sum_j p_{ij} \log \frac{p_{ij}}{q_{ij}} = \sum_i KL(P_i||Q_i)\), by altering \(Q\) with gradient-based optimization methods. Optimization is difficult due to the existence of multiple local optima, and entirely different embeddings may be obtained with different initial \(Q\) distributions.

\[\]
2.2. Covariance Matrix Adaptation Evolution Strategy

Evolutionary strategies are black-box optimization algorithms which belong to a broader group of evolutionary algorithms. In such methods, a set of candidate solutions is maintained. In successive iterations of the procedure, these candidate solutions are perturbed and evaluated, and in each iteration the best solution is left unchanged and carried over to the next set of candidate solutions. CMA-ES (Hansen and Ostermeier, 1996) is an evolutionary strategy where the set of candidate solutions is modeled and sampled from a multivariate Gaussian distribution $\mathcal{N}(\mathbf{m}, \mathbf{C})$. The covariance matrix $\mathbf{C}$ represents pairwise relationships between attributes. The objective function of CMA-ES maximizes two likelihoods: (a) the likelihood of having the best individuals in previous iterations, and (b) the likelihood of taking the best search steps in previous iterations. At the end of each iteration, (a) guides updates of the mean $\mathbf{m}$ as a weighted average of $\mu$ best solutions, $\mathbf{m}^{(g+1)} = \sum_{i=0}^{\mu} w_i x_i^{(g+1)}$, where $x_i^{(g+1)}$ is the $i$th best solution in iteration $g + 1$ and $w_i$ is its weight, and (b) updates the covariance matrix $\mathbf{C}$ as follows:

$$
\mathbf{C}^{(g+1)} = (1 - c_1 - c_\mu) \mathbf{C}^{(g)} + c_\mu \mathbf{p}_c^{(g+1)} \mathbf{p}_c^{(g+1)T} \\
+ c_{\mu} \sum_{i=1}^{\mu} w_i \left( x_i^{(g+1)} - \mathbf{m}^{(g)} \right) \left( x_i^{(g+1)} - \mathbf{m}^{(g)} \right)^T.
$$

Finally, $\mathbf{p}_c^{(g+1)} \in \mathbb{R}^D$ is a vector that amplifies the updates in favorable directions:

$$
\mathbf{p}_c^{(g+1)} = (1 - c_p) \mathbf{p}_c^{(g)} + c_p \frac{\mathbf{m}^{(g+1)} - \mathbf{m}^{(g)}}{\sigma^{(g)}},
$$

where $c_1, c_\mu, c_p \in \mathbb{R}$ are weights, $\sigma^{(g)} \in \mathbb{R}$ is an adaptive step size, which is dependent on the iteration, and $z \in \mathbb{R}$ is a normalizing constant. Details of the parameters of CMA-ES can be found in Hansen and Ostermeier (1996) and Hansen et al. (2003).

2.3. Cox Regression Analysis

The Cox regression model (Cox, 1972) is a proportional hazards regression model in which the hazard ratio is constant but other contents have the same baseline hazard function. Based on this assumption, the survival function is calculated as

$$
S(t) = \exp(-H_0(t) \exp(X\beta)) = S_0(t) \exp(X\beta),
$$

where $H_0(t)$ represents the cumulative baseline hazard function at time $t$ and $S_0(t) = \exp(-H_0(t))$ is the baseline survival function; $H_0(t)$ is taken to be Breslow’s estimator (Breslow, 1974), which is the most widely used and given by

$$
\hat{H}_0(t) = \sum_{t_i \leq t} \hat{h}_0(t_i).
$$

As the Cox model is based on the proportional hazards assumption, it is represented as

$$
h(t, x_i) = h_0(t) \exp(x_i \beta)
$$

for an given instance $i = 1, 2, \ldots, n$, where the baseline hazard function $h_0(t)$ can be an arbitrary negative function of time, and $x_i = (x_{i1}, x_{i2}, \ldots, x_{id})$ is the corresponding covariate vector for instance $i$ and is the coefficient vector. The Cox model is a semi-parametric algorithm where the baseline hazard function $h_0(t)$ is unspecified. For any two instances $x_1$ and $x_2$, the hazard ratio is given by

$$
h(t, x_1) \overline{h(t, x_2)} = \frac{h_0(t) \exp(x_1 \beta)}{h_0(t) \exp(x_2 \beta)} = \exp[(x_1 - x_2)\beta].
$$

This means that the hazard ratio is independent of the baseline hazard function.

2.4. Wrapper-Based Feature Selection Integrating SNE and CMA-ES

The first part of our proposed method performs the task of miRNA selection for diverse cancer types, which is considered a multi-class classification problem here. SNE is used to reorder the features by performing an implicit clustering such that neighboring features are highly correlated. Then, the underlying multi-class classification task is performed using well-known classifiers and treated as an optimization problem for which CMA-ES is used to find the miRNAs automatically. The miRNAs thus found are further refined using forward selection (FS). Therefore, we call this wrapper-based feature selection method as stochastic covariance evolutionary strategy with forward selection (SCES-FS).

Algorithm 1 presents the SCES-FS method in detail. It starts with the dataset $X = \{x_1, x_2, \ldots, x_n\}$, which denotes a set of $n$ observations with $x_i \in \mathbb{R}^D$, the class label $Y$, the population size $\lambda$, the maximum number of generations $N$, the classifier $A$, and the number of runs $R$ as inputs. In the dataset, each feature is characterized by expression levels of the samples. The features of the original dataset $X$ are reordered using SNE in the ConstructEmbedding step, producing the embedding dataset $X'$ whose size is the same as that of the original dataset. The parameters are initialized in the SetParamsCMAES step. The individuals/vectors in CMA-ES are encoded as a simple threshold weight vector, with a single weight for each miRNA. An individual $x \in \mathbb{R}^H$ encodes a weight vector $w \in \mathbb{R}^H$ and a threshold $t \in \mathbb{R}$, where $H \leq D$ is the number of weights. Only those features whose weights exceed the threshold are eventually selected into a feature set $S$:

$$
S = \{i | 1, \ldots, D: closest(x_i) = j \wedge w_i \geq t\},
$$

where $t \in \mathbb{R}$ is a threshold that is chosen carefully. The population of vectors is drawn in the DrawPopulationCMAES step from $\mathcal{N}(\mathbf{m}, \sigma^2 \mathbf{C})$, where $\mathbf{m}$ is the mean, $\mathbf{C}$ is the covariance matrix, and $\sigma$ is the step-size control parameter of CMA-ES. In the ConstructFeatureSets step, each individual $x_i$ is translated to a feature set $S_i$ according to Equation (8).

In the ScoreFeatureSet step, each feature set is evaluated by training a classifier on a dataset restricted to only the selected features. The objective function of CMA-ES combines
the accuracy obtained from the classifier for each individual x and
the size of the feature set S as follows:
\[ f(x) = \text{accuracy}(S) - \alpha \max(0, |S| - \Upsilon) \]
where \( D \) is the dimension of the dataset, \( \Upsilon \) is the target number of
features, and \( \alpha \in \mathbb{R}_+ \) balances the penalty for the excessive
number of features. The objective function is designed in this way
so that higher classification accuracy can be achieved for a small
number of features. The optimization with CMA-
ES allows inter-feature relationships with covariance matrices.
Finally, the parameters are updated according to CMA-ES update
rules in the UpdateParamsCMAES step, and the best set of
features/miRNAs in a particular run are kept in \( S_{\text{RBest}} \).

Algorithm 1 Pseudo-code of the SCES-FS
Input: \( X, \lambda, \lambda, N, A, R \) (dataset, class label, population size,
maximum number of generations, classifier, number of runs)
Output: \( S_{\text{best}} \) (feature subset)
1: Initialize a NULL list, \( L \)
2: for \( i \leftarrow 1 \) to \( R \) do
3: \( \mathcal{X}' \leftarrow \text{ConstructEmbedding}(X) \) // With SNE
4: \( \theta \leftarrow \text{SetParamsCMAES}(\mu, \sigma, m, C) \)
5: for \( g \leftarrow 1 \) to \( N \) do
6: \( \mathcal{P} \leftarrow \text{DrawPopulationCMAES}(\mathcal{X'}, \lambda, \theta) \)
7: \( S \leftarrow \text{ConstructFeatureSets}(\mathcal{P}) \)
8: \( \text{ScoreFeatureSet}(S, \mathcal{X'}, A, \mathcal{Y'}) \)
9: \( S_{\text{RBest}} \leftarrow \text{BestFeatureSet}(S, S_{\text{RBest}}) \)
10: \( \theta \leftarrow \text{UpdateParamsCMAES}(\theta) \)
11: end for
12: \( L \leftarrow L \cup S_{\text{RBest}} \)
13: end for
14: \( L' \leftarrow \text{RankFrequency}(L) \)
15: \( S_{\text{best}} \leftarrow \text{ForwardSelection}(L', X', A, \mathcal{Y'}) \)
16: return \( S_{\text{best}} \)

2.5. Preparation of the Final Set of Features
The SCES-FS is random in nature to reduce the probability of
returning sub-optimal solutions. Therefore, a single run of SCES-
FS does not guarantee a reliable solution. To overcome this
limitation, SCES-FS is executed up to a maximum number of
runs, \( R \). In each run, a set of best features/miRNAs are collected
into a list, \( L \). After completion of the maximum number of runs,
RankFrequency sorts the cumulative set of features in descending
order according to the frequency of occurrence in each run
and produces a modified list, \( L' \). Thereafter, ForwardSelection
applies the forward feature selection method, using the classifier
to evaluate the feature set iteratively to obtain the best feature
set, \( S_{\text{best}} \).

2.6. Justification for miRNA Selection
Because of the random nature of the algorithm, there is a chance of
false positives or false negatives occurring in the selection
of miRNAs. To reduce the probability of having false positives
or false negatives in the selected set of miRNAs, the SCES-FS
algorithm is run 50 times, and then the miRNAs are ranked based
on their frequencies of occurrence in 50 different sets of features.
Thus, a stable set of miRNAs is selected on the basis of maximum
classification accuracy, which is computed by considering the
miRNAs cumulatively from the top of the list. By this process,
17 distinct miRNAs are selected.

This procedure does not, however, ensure the absence of false
positives or false negatives, so additional measures are taken.
The presence of false positives is made unlikely by the sorting
procedure. In fact, since a given false positive is supposed to occur
less frequently than all the true positives, it will appear in the
tail after sorting. Consequently, it is likely that this false positive
will be filtered out when selecting the final list of miRNAs. On
the other hand, the presence of false negatives is related to the
choice of the number of runs and the corresponding expected
number of miRNAs belonging to the sorted list. In this regard,
a mathematical argument is given below to justify that the SCES-FS
does not exhibit random behavior.

Suppose that at each run the SCES-FS randomly selects 10
miRNAs from the whole collection of miRNAs. We first compute
the expected number of distinct miRNAs reported after all the
runs, and then we compare this number with the results of our
experiments. We have the following parameters: \( D = 199 \) is
the number of miRNAs, \( S = 10 \) is the number of miRNAs
returned after each run, and \( R = 50 \) is the number of runs. For
\( i = 1, \ldots, D, \) let \( V_i \) be a Bernoulli-distributed indicator variable,
where \( V_i = 1 \) if the miRNA \( m_i \) never shows up. The probability
that \( m_i \) is selected in one run is \( S/D \), so the probability that it is
never selected is \( 1 - S/D \). Since each of the \( R \) runs is independent,
the following equation can be written:
\[ \mathbb{E}[V_i] = \Pr(V_i = 1) = \left(1 - \frac{S}{D}\right)^R. \]

Let \( V = \sum_{i=1}^{D} V_i \) be the random variable that counts the number of
miRNAs that do not belong to the final set of miRNAs reported
at least once. By linearity of the expectation, we obtain the equation
\[ \mathbb{E}[V] = \sum_{i=1}^{D} \mathbb{E}[V_i] = D \left(1 - \frac{S}{D}\right)^R. \]

Hence, the number of expected miRNAs reported at least once is
\[ \mathbb{E}[(D - V)] = D - \mathbb{E}[V]. \]

Substituting the above-mentioned values for the parameters, we obtain that
- the expected number of miRNAs reported at least once after
50 iterations is 183; and
- the number of new miRNAs added in a further iteration would
be 1.

As the number of sorted miRNAs in our experiment is 39 (see the
Supplementary Material), this difference suggests that 50 runs
are enough to conclude that all the true positives are included in the sorted list and so false negatives are unlikely.

### 2.7. Cox Regression Analysis for Evaluating miRNAs

The second part of the proposed method is for the evaluation of selected miRNAs using Cox regression analysis. The primary objective of this stage is to assess the importance of the miRNAs selected in the first part of the method, and this is done using Cox regression analysis for survival in 10 different cancer types. Expression data of the selected miRNAs and the associated clinical data are used for the Cox regression analysis. In the clinical data, the vital status of each patient, indicating whether the patient is still alive or has died, and the number of days since the last followup are taken into account when performing the survival analysis. Based on the expression levels and the clinical data of the selected miRNAs in each cancer type, the Cox coefficient, hazard ratio, and $p$-value are computed. A higher value of the Cox coefficient signifies greater importance of that miRNA to the respective cancer type. Moreover, the up- and down-regulation of all the selected miRNAs are observed to understand their behavior with respect to that particular cancer type based on change in expression in tumor and normal samples.

### 2.8. Complexity Analysis

Let $D$ be the number of features and $n$ the number of samples in the input dataset. As the available approximations may considerably lower the overall complexity, we discuss the complexity of each building block separately. A single step of SNE requires computing relations between all data points. We embed a transposed dataset, making an optimization step in $O(D^2)$ time. Computing SNE usually involves performing principal component analysis for preliminary dimension reduction, though its cost is negligible. The internal complexity of CMA-ES is estimated as $O(D^2)$, due to sampling and updating of the covariance matrix. The matrix needs to be factorized, which can be done by eigen decomposition in $O(D^3)$ time. Factorization does not happen in every generation, which gives $O(D^3)$ amortized time. Empirical evidence suggests that the sufficient number of objective function evaluations usually scales sub-quadratically with $D$ (Ros and Hansen, 2008). The computation time is similar, since the vast majority of it is time spent training similar classifiers. On the other hand, the time complexity of Cox regression analysis is $O(nD^2)$ for a single run (Kelley, 1999).
et al., 2002) (see section 1 for the full names of these methods), as well as the results with all features (i.e., without feature selection).

### 3.1. Dataset Preparation and Parameter Setting

The miRNA expression and clinical datasets of bladder, breast, colon, glioblastoma, head and neck squamous cell, kidney renal clear cell, lung adenocarcinoma, lung squamous cell, ovarian, and uterine corpus endometrial carcinoma were obtained from TCGA. These 10 cancer types have also been studied previously (see, e.g., Jacobsen et al., 2013). Moreover, Hoadley et al. (2018) found that the characteristics of certain cancers out of 33 types provided in TCGA are overlapping in nature. As a result, 10–15 distinct groups of cancer were reported in Hoadley et al. (2018), which are similar to the cancer types studied in the present article. Our choice of cancer types was based on: (1) careful review of the literature, (2) the availability of tissue-specific tumor and normal samples to avoid the class imbalance problem in classification, and (3) the availability of common miRNA expression data for different cancer types and their corresponding clinical data. Thus, 10 cancer types were selected for our study. The expression data were generated using an Illumina high-throughput sequencing machine in the form of read counts of 199 miRNAs, normalized to reads per million, while the clinical data contain gender, age, days since last followup, and vital status. After removing miRNAs that contain more than 60% zeros in each cancer type and taking the miRNAs common to all the cancer types, we found 199 miRNAs and considered their expression in different cancer types. The number of samples and other details for each cancer type are given in Table 1; we also included 333 normal samples in the analysis with the expression of same miRNAs. For convenience, the expression datasets containing reads per million are further normalized onto a log_2 scale. These processed datasets can be downloaded from the Supplemental Material website.

To construct the final ranking of miRNAs, the SCES-FS algorithm was run 50 times. Five-fold cross-validation was applied during each classification to avoid the issue of overfitting or underfitting. The parameters used in the experiments are shown in Table 2 and were obtained either experimentally or from the literature (Latinne et al., 2001; Oshiro et al., 2012; Hansen, 2016).

### 3.2. Experimental Outcomes

The problem of finding miRNAs that can correctly distinguish different cancer types was posed as a multi-class classification task using expression data of miRNAs, and the importance of the selected miRNAs was evaluated using Cox regression analysis with the help of clinical data. The classification results of SCES-FS using the classifiers RF, SVM, NB, K-NN, and DT over 50 runs are reported in Table 3 and compared with the results of ESVM-RFE, LASSO, NSGA-II-SE, MOGA, SVM-nRFE, SVM-RFE, CMIM, ICAP, SCAD, JMI, CIFE, mRMR, FSCOX, DISR, SNR, and RankSum, as well as the results with all features (i.e., without feature selection). These results show that SCES-FS with RF achieved the highest classification accuracy, 96.881, with a standard deviation of 0.039, while the accuracy results of SCES-FS with SVM, NB, K-NN, and DT are 96.332 ± 0.194, 96.251 ± 0.168, 96.132 ± 0.369, and 94.232 ± 0.057, respectively, still better than those of the other existing feature selection methods.

It is to be noted that our results have been verified with FSCOX, which computes important miRNAs via Cox regression on all miRNAs and subsequently uses

---

**TABLE 3** | Number of features and classification accuracy of feature selection methods for five classifiers with five-fold cross-validation.

| Method        | Number of features | RF          | SVM          | NB          | K-NN         | DT          |
|---------------|--------------------|-------------|--------------|-------------|--------------|-------------|
| SCES-FS       | 17                 | 96.881 ± 0.029 | 96.332 ± 0.194 | 96.251 ± 0.168 | 96.132 ± 0.369 | 94.232 ± 0.057 |
| ESVM-RFE      | 22                 | 95.684 ± 0.031 | 95.902 ± 0.193 | 92.672 ± 0.161 | 91.382 ± 0.369 | 90.429 ± 0.056 |
| LASSO         | 48                 | 95.601 ± 0.038 | 95.547 ± 0.191 | 92.582 ± 0.164 | 91.251 ± 0.367 | 90.241 ± 0.051 |
| NSGA-II-SE    | 26                 | 95.587 ± 0.033 | 95.537 ± 0.195 | 92.538 ± 0.166 | 91.183 ± 0.366 | 90.229 ± 0.052 |
| MOGA          | 24                 | 95.391 ± 0.026 | 95.531 ± 0.194 | 92.293 ± 0.161 | 90.338 ± 0.362 | 89.142 ± 0.056 |
| SVM-RFE       | 26                 | 95.224 ± 0.032 | 95.321 ± 0.191 | 92.281 ± 0.167 | 90.106 ± 0.361 | 88.993 ± 0.051 |
| SVM-RFE       | 28                 | 95.048 ± 0.038 | 95.159 ± 0.199 | 92.116 ± 0.165 | 89.889 ± 0.366 | 88.691 ± 0.053 |
| CMIM          | 28                 | 94.299 ± 0.037 | 92.029 ± 0.183 | 90.683 ± 0.161 | 89.374 ± 0.369 | 89.161 ± 0.052 |
| ICAP          | 27                 | 93.721 ± 0.031 | 92.951 ± 0.192 | 90.874 ± 0.163 | 90.643 ± 0.362 | 87.057 ± 0.063 |
| SCAD          | 25                 | 91.972 ± 0.034 | 90.839 ± 0.194 | 90.003 ± 0.165 | 89.918 ± 0.366 | 87.485 ± 0.058 |
| JMI           | 28                 | 91.718 ± 0.031 | 90.602 ± 0.196 | 89.986 ± 0.166 | 88.639 ± 0.369 | 87.057 ± 0.051 |
| CIFE          | 32                 | 90.886 ± 0.034 | 89.072 ± 0.199 | 88.389 ± 0.162 | 87.261 ± 0.362 | 86.205 ± 0.056 |
| mRMR          | 28                 | 91.063 ± 0.032 | 89.753 ± 0.195 | 87.402 ± 0.161 | 87.254 ± 0.361 | 85.208 ± 0.059 |
| FSCOX         | 23                 | 89.298 ± 0.038 | 88.529 ± 0.198 | 87.505 ± 0.169 | 87.287 ± 0.368 | 85.498 ± 0.058 |
| DISR          | 29                 | 89.286 ± 0.031 | 88.276 ± 0.196 | 87.580 ± 0.167 | 87.321 ± 0.369 | 85.858 ± 0.053 |
| SNR           | 30                 | 87.866 ± 0.035 | 86.712 ± 0.193 | 85.749 ± 0.163 | 85.364 ± 0.365 | 84.042 ± 0.059 |
| RankSum       | 32                 | 86.633 ± 0.033 | 85.556 ± 0.199 | 85.466 ± 0.166 | 84.669 ± 0.366 | 84.322 ± 0.055 |
| Without feature selection | 199 | 86.428 ± 0.036 | 85.294 ± 0.191 | 85.183 ± 0.162 | 84.552 ± 0.367 | 84.118 ± 0.063 |

---

3http://www.nitttrkol.ac.in/indrajit/projects/mirna-prediction-multicells/
them for the classification. It has been observed that the overall classification accuracy of FSCOX is less than the accuracy attained by SCES-FS as reported in Table 3. Other omics-based analyses, such as protein array, copy number variation (CNV), and methylation studies, also have potential applications in pan-cancer classification, as explained in Zhang et al. (2015, 2016, 2019), where the overall accuracy achieved in the range of 93–97%. The cancer dataset used in Zhang et al. (2015, 2016, 2019) was carefully selected for protein array, CNV, and methylation studies and is not directly suitable for experimenting with miRNAs, as it creates a class imbalance problem when selecting miRNA expression data for both tumor and normal cases. However, the overall accuracy of our method is higher than 96%, which is on the higher side of the accuracy range reported by Zhang et al. (2015, 2016, 2019), suggesting that our selected miRNAs can also be considered potential markers for pan-cancer classification.

The results of SCES-FS have been further validated by survival analysis, including Cox regression, network analysis, expression analysis using hierarchical clustering in the form of heatmaps, KEGG pathway analysis, GO enrichment analysis, and PPI network analysis, as described in the following.

### 3.2.1. Survival Analysis

Table 4 reports the results of the Cox regression analysis, i.e., the Cox coefficient and hazard ratio values, for 10 cancer types in order to evaluate the importance of the 17 selected miRNAs with respect to these cancer types. The Cox regression analysis was performed by integrating the miRNA expression and clinical data to assess the effect of miRNA expression on cancer survival. Higher Cox coefficient and hazard ratio values indicate greater influence of the miRNA on the cancer type. For example, the miRNA hsa-mir-375 has the highest Cox coefficient, 0.9882, and hazard ratio, 1.7879, for the GBM cancer type. To help visualize the importance of each miRNA on the cancer type, a circos plot of Table 4 is shown in Figure 2. In the figure, a broader band signifies a stronger association between the miRNA and the particular cancer type. Table 5 summarizes the cancer types that are most highly associated with the selected miRNAs. In this table, for each miRNA, the cancer type for which that miRNA has the highest Cox coefficient is shown, along with the associated \( p \)-value, false discovery rate (FDR), and up- or down-regulation of the miRNA expression in that cancer type. Similarly, Supplementary Table 2 gives the up- and down-regulations along with the corresponding FDR values of the selected miRNAs for all cancer types. The up- and down-regulations of the miRNAs are computed after evaluating the change in expression of the selected miRNAs between tumor samples and normal samples or the population for a particular cancer type using the ANOVA test; this is discussed further in section 3.2.3. Here, the change in expression is considered significant if the \( p < 0.05 \), and the up- or down-regulation is computed if the change in population expression is positive or negative, respectively. In summary, we find that hsa-mir-205, hsa-mir-10a, hsa-mir-196b, hsa-mir-10b, hsa-mir-375, hsa-mir-143, hsa-let-7c, hsa-mir-107, hsa-mir-378, hsa-mir-133a, hsa-mir-1, hsa-mir-30c,
hsa-mir-16, hsa-mir-30a, hsa-let-7i, hsa-mir-24, and hsa-mir-95 are highly associated with GBM, UCEC, OV, BLCA, COAD, and BRCA.

3.2.2. Network Analysis
For the 17 selected miRNAs, miRTarBase (Huang et al., 2020) was used to find their targets in order to elucidate their role in the different cancer types. To identify the most correlated targets, we computed the Pearson correlation between the expression values of miRNAs and mRNAs obtained from TCGA for the 10 cancer types and took the negative correlation value used in Zhou et al. (2015) as indicating strong association. The top five negatively correlated mRNAs associated with each of the 17 miRNAs are reported in Table 6; the rest are reported in Supplementary Material. To construct the interaction network, the miRNAs and their targets were ranked based on the cumulative negative correlation score and their presence in different cancer types as indicated by the association number. These results are reported in Supplementary Table 3. For example, hsa-mir-16 and its target, PHYHIP, are related to six cancer types and the cumulative negative correlation score is −4.142. Similarly, hsa-mir-24 is correlated with C1QTNF6 in another six cancer types, with a cumulative negative correlation score of −3.906. The subset of such targeted mRNAs is used to construct the interaction network shown in Figure 3. The network reveals that the miRNAs hsa-mir-205, hsa-mir-10a, hsa-mir-107, hsa-mir-378, and hsa-mir-16 and their targets {CYR61, STARD8, TNFSF8}, {TTYH3, CARHSP1, LILRA2}, {CPEB3, TGFBR3, FGF2}, {NME4, NWD1, ORAI2}, and {PHYHIP, CPEB3, CTDSP1} play a crucial role in different types of cancer. The red, blue, pink, dark green, light green, and black edges in Figure 3 signify that the number of cancer types associated with the corresponding pair of miRNA and target mRNA is 6, 5, 4, 3, 2, and 1, respectively. The targets of the miRNAs are investigated further using KEGG pathway, GO enrichment, and PPI network analysis in the following subsections in order to see their impact on the different types of cancer.

3.2.3. Expression Analysis
Expression analysis was conducted using a one-way ANOVA test in order to evaluate the statistical significance of the differential expression of the 17 selected miRNAs. Alternative techniques can also be used (Conesa et al., 2016; Costa-Silva et al., 2017; Crow et al., 2019). To perform the test, the population of patients...
TABLE 5 | Cancer type most strongly associated with each selected miRNA, based on Cox coefficient.

| miRNA    | Cox coefficient | Hazard ratio | Cancer type | p-value   | FDR       | Regulation (up ↑ /down ↓) | PubMed ID |
|----------|-----------------|--------------|-------------|-----------|-----------|---------------------------|-----------|
| hsa-mir-205 | 0.9233         | 2.5177       | GBM         | 5.14E-03  | 5.46E-03  | ↓                          | 23054677 |
| hsa-mir-10a  | 0.4881         | 1.6293       | GBM         | 6.17E-24  | 1.87E-23  | ↓                          | 20444541 |
| hsa-mir-196b | 0.4581         | 1.5811       | UCEC        | 2.84E-48  | 1.21E-47  | ↑                          |          |
| hsa-mir-10b  | 0.5080         | 1.6621       | UCEC        | 4.71E-03  | 4.71E-03  | ↓                          |          |
| hsa-mir-375  | 0.9882         | 1.7879       | GBM         | 1.80E-14  | 2.36E-14  | ↓                          |          |
| hsa-mir-143  | 0.1504         | 1.1623       | OV          | 3.38E-57  | 1.79E-56  | ↓                          |          |
| hsa-let-7c   | 0.1731         | 1.1890       | BLCA        | 3.85E-38  | 1.31E-37  | ↓                          |          |
| hsa-mir-107  | 0.5363         | 1.7096       | GBM         | 5.42E-24  | 1.87E-23  | ↑                          | 24213470 |
| hsa-mir-378  | 0.5511         | 1.7352       | GBM         | 5.04E-20  | 7.79E-20  | ↓                          | 29088758 |
| hsa-mir-133a | 0.4422         | 1.5561       | OV          | 2.78E-20  | 4.73E-20  | ↑                          | 24944666 |
| hsa-mir-1    | 0.3702         | 1.3078       | GBM         | 9.18E-08  | 1.04E-07  | ↑                          |          |
| hsa-mir-30c  | 0.5053         | 1.6575       | COAD        | 4.48E-17  | 6.34E-17  | ↓                          |          |
| hsa-mir-16   | 0.1409         | 1.1513       | GBM         | 5.42E-24  | 1.87E-23  | ↑                          | 25864039 |
| hsa-mir-30a  | 0.8083         | 2.2441       | COAD        | 2.88E-56  | 9.80E-56  | ↓                          | 22287560 |
| hsa-let-7i   | 0.4383         | 1.5501       | BRCA        | 1.33E-43  | 2.83E-43  | ↑                          | 26378051 |
| hsa-mir-24   | 0.4885         | 1.6298       | GBM         | 5.42E-24  | 1.87E-23  | ↑                          | 25864039 |
| hsa-mir-96   | 0.2872         | 1.3326       | GBM         | 6.59E-24  | 1.87E-23  | ↑                          | 28156650 |

was divided into tumor and normal groups for a given miRNA in a particular cancer type. As a result of ANOVA, significant ($p < 0.05$) changes in expression were observed between the tumor and normal groups for the 17 selected miRNAs. For example, the $p$-values of hsa-mir-10a, hsa-mir-196b, hsa-mir-10b, hsa-mir-375, hsa-mir-143, hsa-let-7c, hsa-mir-107, hsa-mir-378, hsa-mir-133a, and hsa-mir-30c were 6.17E-24, 2.84E-48, 4.71E-03, 1.80E-14, 3.38E-57, 4.71E-03, 1.87E-14, 3.85E-38, 1.31E-37, and 3.85E-38 respectively. Additionally, box plots of the selected miRNAs in tumor and normal samples are provided in Supplementary Figure 3. To investigate the relationship between the expression levels of miRNAs for each cancer type, hierarchical clustering was performed on the tumor and normal samples of the 17 selected miRNAs. The results are shown in Figure 4, from which the change in expression levels of the miRNAs between tumor and normal samples is evident. In the cluster plots, red indicates high expression levels and green low expression levels; black corresponds to not significantly expressed samples.

3.2.4. KEGG Pathway Analysis

In order to perform KEGG pathway analysis for the 17 selected miRNAs in 10 cancer types, their targets were identified based on negative correlation values as described in section 3.2.2. Similar to the KEGG pathway analysis, GO enrichment analysis was also performed with the targets of the selected miRNAs using the Enrichr tool (Kuleshov et al., 2016), to assess the significance of the roles the selected miRNAs play in different biological activities. The results of different analyses for biological processes, molecular functions, and cellular components are reported in Table 8 and in Supplementary Tables 4, 5, respectively; the details of the enrichment analysis are also given in Supplementary Material. In Table 8, we see that various biological process GO terms which are related to the targets of the 17 selected miRNAs have important roles in cancer development. For example, GO:0023051 is linked with regulation of signaling, which is involved in the development of colorectal cancer. Similarly, GO:005120 is linked with regulation of the RNA metabolic process of bladder urothelial carcinoma,
| miRNA       | BLCA  | BRCA  | COAD  | GBM   | HNSC  | KIRC  | LUAD  | LUSC  | OV    | UCEC  |
|------------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
|            | mRNA  | score | mRNA  | mRNA  | mRNA  | mRNA  | mRNA  | mRNA  | mRNA  | mRNA  |
| ZEB2       | 7.45  | -5.29 | 17.95 | -9.95 | -9.52 | 5.28  | 8.49  | 6.98  | 4.15  | 5.82  |
| hsa-mir-205| SYT1  | 7.02  | 2.94  | 5.93  | 6.46  | 4.61  | 4.61  | 4.58  | 3.88  | 2.10  |
| RCA2       | 6.99  | -2.70 | 8.99  | 7.89  | 8.23  | 4.25  | 5.74  | 7.97  | 4.18  | 5.13  |
| LRRK2      | 6.99  | -2.63 | 8.78  | 7.84  | 6.51  | 5.14  | 2.57  | 5.23  | 3.77  | 6.49  |
| NACC2      | 5.33  | -2.39 | 9.33  | 8.28  | 4.54  | 5.45  | 5.06  | 3.16  | 6.29  | 5.75  |
| CLC4       | 5.29  | -2.67 | 7.97  | 11.11 | 9.14  | 5.00  | 9.99  | 4.98  | 2.79  | 5.18  |
| COL6A2     | 5.04  | -3.89 | 2.03  | 9.68  | 9.42  | 2.82  | 3.57  | 2.16  | 3.55  | 2.95  |
| RP1A       | 4.79  | -2.90 | 8.82  | 7.18  | 2.85  | 2.67  | 6.12  | 2.77  | 2.01  | 4.77  |
| TGF3       | 4.77  | -3.54 | 5.92  | 8.79  | 5.12  | 6.03  | 2.65  | 3.13  | 9.89  | 3.34  |
| MN1B       | 4.08  | -2.77 | 7.19  | 9.91  | 6.23  | 5.41  | 4.68  | 2.79  | 3.78  | 2.87  |
| PPP3CB     | 3.86  | -1.81 | 9.61  | 4.39  | 5.43  | 1.43  | 3.55  | 1.78  | 3.53  | 2.64  |
| SODC       | 3.78  | -1.60 | 7.60  | 9.89  | 6.23  | 6.81  | 2.57  | 2.14  | 2.73  | 2.73  |
| CTGF       | 4.17  | -3.35 | 7.90  | 4.97  | 4.97  | 1.15  | 1.77  | 1.81  | 1.86  | 1.86  |
| FN1        | 4.09  | -5.02 | 8.81  | 5.00  | 5.38  | 3.38  | 5.58  | 3.67  | 2.50  | 2.50  |
| TNS1       | 6.09  | -1.81 | 9.92  | 4.97  | 4.97  | 1.15  | 2.43  | 2.24  | 2.16  | 2.16  |
| SHD12      | 4.01  | -1.86 | 9.58  | 4.97  | 4.97  | 1.15  | 1.77  | 1.81  | 1.86  | 1.86  |
| SAMD4A     | 3.99  | -5.71 | 5.94  | 5.94  | 4.97  | 4.97  | 1.15  | 2.43  | 2.24  | 2.16  |
| OTUB1      | 5.34  | 4.22  | 7.90  | 4.97  | 4.97  | 1.15  | 2.43  | 2.24  | 2.16  | 2.16  |
| CAPZA1     | 4.88  | -5.34 | 5.94  | 5.94  | 4.97  | 4.97  | 1.15  | 2.43  | 2.24  | 2.16  |
| SYNO2L     | 4.88  | -4.69 | 9.51  | 3.97  | 3.97  | 1.81  | 1.77  | 1.81  | 1.86  | 1.86  |
| STXB2P1    | 4.79  | -4.68 | 9.48  | 4.97  | 4.97  | 1.15  | 2.43  | 2.24  | 2.16  | 2.16  |
| KCNA7      | 4.65  | -4.84 | 9.92  | 4.97  | 4.97  | 1.15  | 2.43  | 2.24  | 2.16  | 2.16  |
| YWHAZ      | 4.49  | -5.84 | 9.81  | 5.38  | 5.38  | 4.23  | 7.94  | 3.43  | 2.68  | 2.68  |
| SLC22A1    | 4.16  | -5.88 | 9.80  | 6.14  | 6.14  | 4.04  | 6.88  | 3.32  | 2.32  | 2.32  |
| EFHD2      | 3.93  | -5.84 | 9.79  | 7.97  | 7.97  | 4.81  | 7.94  | 3.43  | 2.68  | 2.68  |
| MRPL12     | 3.86  | -5.72 | 9.67  | 5.80  | 5.80  | 4.77  | 6.74  | 3.31  | 2.54  | 2.54  |
| PCGF3      | 3.73  | -5.56 | 8.67  | 5.54  | 5.54  | 4.28  | 6.74  | 3.31  | 2.54  | 2.54  |

(Continued)
| miRNA  | BLCA | BRCA | COAD | GBM  | HNSC | KIRC | LUAD | LUSC | OV  | UCEC |
|-------|------|------|------|------|------|------|------|------|-----|------|
| hsa-mir-107 | GLP2R | −7.83 | CAV1 | −8.34 | SMARCA5−9.64 | REL | −4.65 | CPEB3 | −5.23 | FOXC1 | −8.31 | RS1 | −8.88 | RS1 | −8.75 | SUN2 | −4.14 | CPEB1 | −5.76 |
|       | PER1 | −7.67 | FG2 | −8.05 | TMEB18A−9.63 | PKR1 | −3.59 | CHRM1 | −4.70 | PLAG1 | −8.31 | SH3GL2 | −8.68 | ALDH3B1 | −6.89 | ERN1 | −2.97 | CYSLTR2 | −3.71 |
|       | CPEB1 | −7.62 | FOXO1 | −7.82 | CDC42E5−9.52 | ZBTB38 | −3.43 | AMOT | −4.09 | CPEB3 | −8.13 | TGFBR3 | −8.45 | LATS2 | −6.77 | VCAN | −2.96 | FG2 | −3.41 |
|       | NFIA | −7.32 | CST | −7.71 | PURA | −9.42 | CAV1 | −3.36 | TGFBR3 | −3.51 | SH3GL2 | −8.09 | CAV1 | −8.26 | PRKCE | −6.36 | PAG1 | −2.85 | SLCO2A1 | −2.91 |
|       | DMFK | −7.00 | KLFL4 | −7.62 | ADOAR3 | −3.08 | NFA | −3.36 | CKMT1A | −7.93 | FG2 | −8.15 | PAG1 | −6.34 | VWAH | −2.61 | DAPK1 | −2.73 |
|       | ENO1 | −5.57 | PTBP1 | −6.96 | TMEM154−9.76 | NWD1 | −6.56 | SPPRNH1−6.16 | PKD2 | −5.07 | BYSL | −6.95 | MELK | −4.10 | KLHL7 | −3.51 | MYOC2 | −3.84 |
|       | TTC4 | −5.48 | RABEP2 | −6.75 | MYADM | −9.69 | RTNI | −6.27 | ENAH | −5.83 | PML | −4.96 | ALDOA | −6.91 | PRMT1 | −3.81 | IGDC3 | −3.48 | IGFS3 | −3.45 |
| hsa-mir-378 | MRPL37 | −5.37 | BCSC | −6.68 | OPA3 | −9.57 | CYP2U1 | −5.99 | MARVELD1−5.48 | LDHA | −4.71 | LDHA | −6.78 | PTOV1 | −3.50 | ENAH | −3.25 | NLG2N | −3.35 |
|       | CDK4 | −5.27 | HIST1H2B2−6.65 | UGT8 | −9.55 | WDR5B | −5.63 | FBLM1 | −5.46 | ORA2 | −4.67 | P4HB | −6.71 | ENO1 | −3.27 | IGFS3 | −3.19 | MSC | −3.27 |
|       | FGF2 | −4.70 | PLAG1 | −8.24 | SH3GL2 | −3.51 | HSPA1A | −5.39 | STOML1 | −4.50 | PAICS | −6.70 | DCTPP1 | −3.19 | VAMP4 | −3.16 | YPEL1 | −3.26 |
|       | MGAT2 | −3.99 | MEG3 | −5.74 | TMEM59 | −9.63 | NR3C1 | −4.30 | PI3R | −2.49 | PRDM16 | −7.24 | FERM2 | −3.94 | PRDM16 | −3.79 | VEGFA | −2.21 | PI3R | −2.42 |
|       | DEK | −3.52 | PKR | −5.39 | UGT2B10−9.30 | ANGEL2 | −4.28 | PIAS2 | −2.40 | CMTM4 | −6.94 | ANGPT4 | −3.84 | MLEC | −3.72 | ATFPH | −1.90 | BCL3 | −2.28 |
|       | CDK5R1 | −3.49 | PER2 | −5.30 | CDC42 | −9.08 | FAM160B1−5.35 | RBMXL1 | −2.28 | KCNQ1 | −6.81 | ZEB1 | −3.75 | ANGPT4 | −3.53 | CIOA1 | −1.79 | MYPN | −2.19 |
|       | SUPT16H | −3.43 | NGFR | −7.77 | SEOB1 | −6.51 | MMP14 | −3.21 | CIAO1 | −2.23 | ERBB2 | −6.77 | GRID1 | −3.45 | ARHGAP3−5.90 | KRT7 | −1.72 | KCNQ1 | −2.08 |
|       | TCTE1D2 | −3.37 | NR3C1 | −4.65 | MYL12A | −8.03 | UBA2 | −3.20 | AFTPH | −2.23 | PNP | −6.67 | ZFP28 | −3.30 | ZEB1 | −3.41 | NR2C2 | −1.64 | NFAM1 | −1.83 |
|       | NCAPG | −6.57 | PTBP1 | −5.38 | BMP7 | −9.88 | SLCOA1 | −5.20 | LHX4 | −3.32 | VEGFA | −5.09 | RFC5 | −5.15 | EFTUD2 | −4.78 | IFF4 | −2.33 | CDK8 | −2.85 |
|       | PTBP1 | −6.35 | ATP13A1 | −4.69 | CAST | −9.73 | C10orf7 | −5.03 | FANCI | −3.32 | ANGPT4L | −4.62 | KIF4A | −4.92 | IQGAP3 | −4.72 | MYEF2 | −1.98 | PPB | −2.81 |
|       | RCC2 | −6.23 | ODA2O | −4.54 | CEBPA | −9.72 | SYL3 | −4.94 | SFXN1 | −3.29 | NETO2 | −4.58 | MAD2L1 | −4.88 | DSG2 | −3.99 | TWF1 | −1.91 | MTHFS | −2.72 |
|       | UHRF1 | −6.21 | KIAA1522 | −4.54 | GOLGA7 | −6.97 | LTFL1 | −4.80 | BR33P | −3.27 | KAT2 | −4.53 | MTHFD2 | −4.83 | EFG1 | −3.80 | FOLR1 | −1.90 | MACROD1−2.44 |
|       | SFXN1 | −6.17 | RCC3 | −4.51 | OAT | −9.63 | WDR11 | −4.78 | NPCAPD3 | −3.01 | HPs4 | −4.44 | SPG24 | −3.72 | YHDF2 | −1.86 | CRELD2 | −2.24 |
|       | PAM | −4.93 | NRBP1 | −3.56 | SOCS3 | −9.78 | MARCKSL1−5.82 | SNAI2 | −3.65 | VIM | −7.03 | TOMM5 | −4.76 | BIRC5 | −4.04 | ETS1 | −3.27 | PSMD7 | −4.26 |
|       | CXL11 | −4.92 | RUNX2 | −3.11 | ARF3 | −9.49 | VASH1 | −5.63 | SPPRINE1−3.30 | LHPL1 | −6.51 | RPS3 | −3.70 | MYBL2 | −4.03 | MTDH | −3.15 | UBE2I | −4.23 |
|       | SNAIL2 | −3.79 | TGB | −9.98 | FOXA1 | −9.48 | ETS1 | −5.40 | SLCA7−3.22 | SH3GL1 | −6.40 | MRPS16 | −4.43 | SLCA7 | −3.91 | ITGB3 | −3.15 | NDUF2A−3.64 |
|       | ADAM9 | −3.84 | CASP3 | −3.92 | UBE2 | −9.40 | DIL4 | −5.37 | CTSC | −3.04 | IFN4 | −6.33 | BIRC5 | −4.33 | ECT2 | −3.85 | RRM2 | −2.99 | LLPH | −3.63 |
|       | MGAT2 | −3.79 | PTIF2 | −2.82 | CADPS2 | −8.43 | SOX1 | −4.98 | SLCA7−3.03 | BIRC5 | −6.17 | UKT1 | −4.29 | RRM2 | −3.81 | CASP3 | −2.99 | PPRPI−1.38 | −5.38 |

**TABLE 6** Continued
with an FDR-corrected $p$-value of $1.30E-03$, which is $<0.05$. Other GO terms for biological processes, such as GO:0051171, GO:0048519, and GO:0009653, are linked with the regulation of nitrogen compound metabolic process, negative regulation of biological process, and anatomical structure morphogenesis, respectively, for different cancer types.

As with the biological processes, GO terms for molecular functions were also found to be significant in the development of various types of cancer, as reported in Supplementary Table 4. For example, GO:0008134 is linked with transcription factor binding, which plays an important role in BRCA, GBM, KIRC, LUAD, OV, and UCEC, with respective FDR-corrected $p$-values $8.10E-03$, $8.31E-06$, $6.80E-06$, $4.70E-03$, $3.76E-05$, and $3.40E-03$ all being $<0.05$. Other important GO terms such as GO:0044877, GO:0003723, GO:0003676, and GO:0008092 are found to be linked to protein-containing complex binding, RNA binding, nucleic acid binding, and cytoskeletal protein binding, respectively, for different cancer types. Supplementary Table 5 reports the significant GO terms for cellular components in various cancer types. GO:0070013 is linked to intracellular organelle lumen, which has FDR-corrected $p$-values within the 5% significance level for seven different cancer types, namely BLCA, BRCA, COAD, GBM, HNSC, KIRC, and LUSC. In addition, GO:0043227, GO:0005654, and GO:0044444 are associated with membrane-bounded organelle, nucleoplasm, and cytoplasmic part, respectively, which are critical processes in the progression of different types of cancer. The relationship between these GO terms and important activities in cancer development have also been cross-validated in other studies (Waldman et al., 1997; Dhillon et al., 2007; He et al., 2013; Reimand et al., 2013; McClurg and Robson, 2015). Taken together, all of these evidences point to the potential importance of the 17 selected miRNAs in the development of various types of cancer.

### 3.2.6. Protein-Protein Interaction Network Analysis

In PPI networks, a node and an edge signify the interaction of a given protein and the protein-protein association. Here, related proteins share common functions, although they do not necessarily physically interact with each other. In our study, to perform the PPI network analysis, 170 sets of targets relating to 17 miRNAs in 10 different cancer types were used to compute the PPI networks using the STRING database (Szklarczyk et al., 2019). Then, the 170 interaction networks were further analyzed to rank the proteins based on the degree of their nodes and their presence in 10 cancer types. The top 30 proteins are reported in Table 9, while the rest are given in Supplementary Material. For example, MYC has degrees 34, 28, 41, 85, 33, 133, 54, 25, 38, and 32 with respect to BLCA, BRCA, COAD, GBM, HNSC, KIRC, LUAD, LUSC, OV, and UCEC, respectively. Therefore, the total degree of MYC is 503. Similarly, other proteins have certain degrees in different cancer types. Moreover, their presence in the different cancer types is indicated by the association count; for example, MYC has an association count of 10. The proteins in Table 9 are used to construct the final consolidated PPI network shown in Figure 5, which represents the associations between the top 30 proteins and 10 different cancer types. The average node

| miRNA | score | mRNA | score | mRNA | score | mRNA | score | mRNA | score | mRNA | score | mRNA | score | mRNA | score |
|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| hsa-mir-71 | 4.30 | DUSP1 | 1.86 | DLG2 | 1.56 | DLG3 | 1.56 | DLG1 | 1.56 | DLG2 | 1.56 | DLG3 | 1.56 | DLG1 | 1.56 |
| hsa-mir-24 | 1.86 | DUSP1 | 1.56 | DLG2 | 1.26 | DLG3 | 1.26 | DLG1 | 1.26 | DLG2 | 1.26 | DLG3 | 1.26 | DLG1 | 1.26 |
| hsa-mir-95 | 1.30 | DUSP1 | 1.00 | DLG2 | 0.75 | DLG3 | 0.75 | DLG1 | 0.75 | DLG2 | 0.75 | DLG3 | 0.75 | DLG1 | 0.75 |
degree in this network is 13.7, with PPI enrichment $p < 1.0 \times 10^{-16}$. In Figure 5, the significant proteins are MYC, PTEN, CDK1, BRCA1, AKT1, PIK3R1, etc. MYC is the most important protein, known to be oncogenic for breast cancer. Similarly, AKT1 is very significant for breast, lung, and colon cancers. PIK3R1 also plays an important role in breast cancer. The detailed list of PPI networks is provided in Supplementary Material. In summary, the PPI network analysis suggests that the 17 selected miRNAs can be considered important biomarkers with respect to diagnosis of 10 different cancer types.

### 3.3. Web Application for Prediction of 10 Cancer Types

A web version of the predictor was developed for use of the scientific and medical communities to aid in cancer diagnosis. The application uses the random forest and the expression of 17 miRNAs to predict the occurrence of 10 different types of cancer. The random forest was chosen as it outperformed other classifiers in our experiments. To predict the type of cancer for a particular patient or set of patients, the expression values of 17 miRNAs in those patients need to be uploaded in a specific format, instructions are provided on the website. The application will return and display the predicted cancer types for the given set of patients.

### 4. CONCLUSIONS

We have proposed a statistical learning method for feature selection that integrates clustering, classification, and regression to find putative miRNAs by solving a multi-class classification problem on 10 cancer types. The first part of the method is a wrapper-based feature selection algorithm, called stochastic covariance evolutionary strategy with forward selection (SCES-FS), which involves stochastic neighbor embedding (SNE),
FIGURE 4 | Hierarchical clustering results of the differentially expressed miRNAs for the BLCA, BRCA, COAD, GBM, HNSC, KIRC, LUAD, LUSC, OV, and UCEC datasets. Red indicates high expression levels, green low expression levels, and black not significantly expressed samples.
| miRNA    | Pathway                                | BLCA  | BRCA  | COAD  | GBM   | HNSC  | KIRC  | LUAD  | LUSC  | OV    | UCEC  |
|----------|----------------------------------------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| hsa05206 | MicroRNAs in cancer                    | 2.20E-03 | –     | 1.60E-04 | 1.73E-05 | 2.23E-02 | 1.02E-06 | 2.20E-04 | 1.70E-04 | 2.00E-03 | 2.40E-03 |
| hsa05202 | Transcriptional misregulation in cancer | 1.33E-02 | –     | 1.62E-02 | –     | –     | 1.11E-02 | 2.15E-02 | –     | 9.20E-03 | –     |
| hsa-mir-205 | Bladder cancer                           | – | 1.06E-02 | – | – | 3.54E-02 | – | – | – | 8.30E-03 | –     |
| hsa-mir-10a | Proteoglycans in cancer                  | – | – | 2.38E-02 | – | 1.73E-02 | – | – | – | 8.30E-03 | –     |
| hsa-mir-10a | Adherens junction                       | 2.13E-02 | – | – | – | – | 2.80E-02 | – | – | – | –     |
| hsa05016 | Huntington's disease                    | 1.27E-02 | – | – | – | – | – | – | 4.64E-02 | – | –     |
| hsa04218 | Cellular senescence                     | – | – | – | 3.01E-02 | – | – | – | – | – | –     |
| hsa-mir-10a | Thermogenesis                           | 1.27E-02 | – | – | – | – | – | – | – | – | –     |
| hsa04510 | Focal adhesion                          | 1.27E-02 | – | – | – | – | – | – | – | – | –     |
| hsa04213 | Longevity regulating pathway—multiple species | 1.27E-02 | – | – | – | – | – | – | – | – | –     |
| hsa05100 | Colorectal cancer                       | 2.24E-07 | 2.30E-06 | 2.31E-07 | 7.00E-04 | 2.36E-07 | 4.54E-06 | 8.48E-05 | 2.38E-06 | 2.24E-07 | 2.20E-07 |
| hsa01952 | Endocrine resistance                    | 2.77E-07 | 5.58E-07 | – | – | 2.92E-07 | 5.08E-06 | 8.58E-05 | 5.75E-07 | – | –     |
| hsa05215 | Prostate cancer                         | 2.99E-06 | 3.00E-06 | – | – | 2.51E-06 | 3.81E-05 | – | 3.09E-06 | 3.42E-06 | –     |
| hsa05161 | Hepatitis B                             | – | 3.00E-06 | – | 1.52E-05 | 3.61E-07 | – | 8.58E-06 | 3.09E-06 | 5.15E-07 | –     |
| hsa04915 | Estrogen signaling pathway              | 2.15E-06 | 2.88E-06 | – | – | 2.26E-06 | 3.27E-05 | – | 2.97E-06 | – | –     |
| hsa05196 | Epstein-Barr virus infection            | – | 3.28E-02 | – | 2.40E-03 | 2.90E-02 | 1.10E-04 | – | – | 1.19E-02 | –     |
| hsa04550 | Signaling pathways regulating pluripotency of stem cells | – | – | 1.50E-03 | – | 1.48E-02 | – | – | – | 1.19E-02 | –     |
| hsa-mir-10b | MicroRNAs in cancer                     | 2.50E-03 | – | – | – | – | – | – | – | 1.19E-02 | –     |
| hsa04110 | Cell cycle                              | – | 9.10E-03 | – | – | – | 2.80E-04 | – | – | – | –     |
| hsa04914 | Progesterone-mediated oocyte maturation | – | 1.34E-02 | – | – | 4.95E-02 | – | – | – | – | –     |
| hsa01521 | EGFR tyrosine kinase inhibitor resistance | 7.00E-04 | 1.09E-02 | – | – | – | – | – | 7.50E-04 | 2.80E-03 | –     |
| hsa04550 | Signaling pathways regulating pluripotency of stem cells | 4.02E-02 | 1.09E-02 | – | – | – | – | – | – | – | –     |
| hsa-mir-375 | HIF-1 signaling pathway                | – | – | 2.32E-02 | – | – | – | – | 4.60E-03 | – | –     |
| hsa05165 | Human papillomavirus infection          | – | – | 2.32E-02 | 8.70E-03 | – | – | – | – | – | –     |
| hsa05224 | Breast cancer                           | – | – | 6.00E-03 | 1.90E-03 | – | – | – | – | – | –     |
| hsa05230 | Central carbon metabolism in cancer     | – | 1.00E-03 | – | – | 4.09E-02 | – | – | – | – | –     |
| hsa05213 | Endometrial cancer                      | – | 1.00E-03 | – | – | – | – | – | – | – | –     |
| hsa05206 | MicroRNAs in cancer                     | – | 1.00E-03 | – | – | – | – | – | – | – | –     |
| hsa05205 | Proteoglycans in cancer                 | – | 1.00E-03 | – | – | – | – | – | – | – | –     |
| hsa05161 | Hepatitis B                             | – | 1.00E-03 | – | – | – | – | – | – | – | –     |
TABLE 7 | Continued

| miRNA     | Pathway                                      | BLCA     | BRCA     | COAD     | GBM      | HNSC     | KIRC     | LUAD     | LUSC     | OV       | UCEC     |
|-----------|----------------------------------------------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|
| hsa05206: MicroRNAs in cancer | –                     | 1.68E-02 | 1.86E-02 | 2.08E-04 | 1.50E-03 | 1.10E-04 | –        | –        | –        | –        | –        |
| hsa04115: p53 signaling pathway | –                     | –        | –        | 2.03E-03 | 1.28E-03 | 1.63E-03 | 4.31E-02 | –        | –        | –        | –        |
| hsa04110: Cell cycle | –                     | 6.10E-03 | –        | –        | –        | 4.31E-02 | –        | –        | –        | –        | –        |
| hsa05022: Small cell lung cancer | –                     | 4.05E-02 | –        | –        | –        | 4.10E-04 | –        | –        | –        | –        | –        |
| hsa04121: Apoptosis – multiple species | –                     | 4.05E-02 | –        | –        | –        | 4.31E-02 | –        | –        | –        | –        | –        |
| hsa05200: Pathways in cancer | 2.50E-03 | 1.30E-03 | –        | 1.31E-02 | –        | –        | 2.40E-03 | 4.70E-04 | 7.36E-05 | –        | –        |
| hsa01521: EGFR tyrosine kinase inhibitor resistance | 5.50E-03 | 1.40E-03 | –        | 3.10E-03 | –        | –        | 5.20E-03 | 5.60E-03 | 3.50E-04 | –        | –        |
| hsa05165: Human papillomavirus infection | 8.60E-03 | 1.30E-03 | 1.41E-02 | –        | –        | –        | 8.80E-03 | 2.30E-03 | –        | –        | –        |
| hsa04151: PI3K-Akt signaling pathway | 5.50E-03 | 1.40E-03 | –        | –        | –        | –        | 5.60E-03 | 2.70E-03 | –        | –        | –        |
| hsa05224: Breast cancer | 5.50E-03 | –        | –        | 1.25E-02 | –        | –        | –        | –        | –        | –        | –        |
| hsa03013: RNA transport | 1.49E-02 | –        | –        | –        | –        | –        | 3.11E-02 | 2.04E-02 | –        | –        | –        |
| hsa03010: Ribosome | –        | –        | –        | –        | –        | –        | 6.60E-04 | 2.04E-02 | –        | –        | –        |
| hsa01100: Metabolic pathways | –        | –        | –        | –        | –        | –        | 2.56E-02 | 2.04E-02 | –        | –        | –        |
| hsa00010: Glycolysis/Gluconeogenesis | –        | –        | –        | –        | –        | –        | 2.56E-02 | 2.04E-02 | –        | –        | –        |
| hsa05224: Breast cancer | 5.50E-03 | –        | –        | 1.25E-02 | –        | –        | –        | –        | –        | –        | –        |
| hsa05206: MicroRNAs in cancer | 1.33E-02 | 1.50E-03 | 1.70E-05 | 1.60E-06 | 4.70E-03 | 2.21E-02 | 1.92E-05 | 1.40E-06 | 7.20E-03 | –        | –        |
| hsa05215: Prostate cancer | 7.80E-03 | 1.60E-04 | –        | 3.90E-03 | 7.90E-04 | 2.21E-02 | –        | 8.00E-04 | 7.20E-03 | –        | –        |
| hsa05212: Pancreatic cancer | 5.50E-04 | –        | 4.20E-03 | –        | 5.50E-04 | 2.21E-02 | –        | –        | –        | 6.20E-03 | –        |
| hsa01524: Platinum drug resistance | 3.30E-03 | –        | 2.71E-02 | –        | 1.70E-03 | –        | –        | –        | 6.20E-03 | –        | –        |
| hsa05205: Proteoglycans in cancer | –        | –        | 3.10E-03 | –        | –        | –        | 8.34E-05 | 8.00E-04 | –        | –        | –        |
| hsa03030: DNA replication | 1.34E-10 | 1.83E-07 | –        | –        | 6.25E-06 | 7.60E-03 | 6.10E-09 | 1.16E-05 | –        | –        | –        |
| hsa03430: Mismatch repair | 3.80E-04 | 7.10E-03 | –        | –        | 3.10E-04 | –        | 4.00E-04 | 1.75E-02 | –        | –        | –        |
| hsa04110: Cell cycle | 2.56E-04 | 9.20E-03 | –        | –        | 2.10E-04 | –        | 8.74E-08 | –        | –        | –        | –        |
| hsa03015: miRNA surveillance pathway | 3.64E-02 | –        | –        | –        | –        | –        | 1.01E-02 | –        | –        | –        | –        |
| hsa05166: HTLV-I infection | 4.42E-02 | –        | –        | –        | –        | –        | –        | –        | –        | –        | –        |
| hsa05026: MicroRNAs in cancer | 1.79E-05 | –        | 2.90E-03 | –        | 2.04E-05 | –        | –        | 1.21E-02 | –        | –        | –        |
| hsa05200: Pathways in cancer | –        | 8.81E-05 | –        | 2.40E-03 | 2.07E-02 | 3.40E-03 | –        | –        | –        | –        | –        |
| hsa05211: Renal cell carcinoma | –        | 2.30E-03 | –        | 2.09E-02 | 6.40E-03 | –        | 1.42E-02 | –        | –        | –        | –        |
| hsa04141: Protein processing in endoplasmic reticulum | –        | 2.30E-03 | –        | 1.93E-02 | –        | 1.04E-02 | 2.56E-02 | –        | –        | –        | –        |
| hsa04380: Osteoclast differentiation | 2.71E-02 | –        | 5.10E-03 | –        | –        | –        | –        | 9.20E-03 | –        | –        | –        |
| hsa04510: Focal adhesion | 1.90E-03 | 5.80E-03 | –        | –        | –        | –        | –        | –        | –        | –        | –        |
| hsa04010: MAPK signaling pathway | 1.31E-02 | 6.30E-03 | –        | –        | –        | –        | –        | –        | –        | –        | –        |
| hsa05200: Pathways in cancer | 1.58E-02 | –        | –        | –        | –        | 7.42E-05 | –        | –        | –        | –        | –        |
| hsa04151: PI3K-Akt signaling pathway | –        | 2.60E-03 | –        | –        | –        | –        | –        | 3.20E-03 | –        | –        | –        |
| hsa05206: MicroRNAs in cancer | –        | 6.30E-03 | –        | –        | –        | 7.40E-04 | –        | –        | –        | –        | –        |
TABLE 7 | Continued

| miRNA         | Pathway                                | BLCA   | BRCA | COAD | GBM    | HNSC   | KIRC   | LUAD   | LUSC   | OV    | UCEC   |
|---------------|----------------------------------------|--------|------|------|--------|--------|--------|--------|--------|-------|--------|
| hsa04110: Cell cycle |                                        | 4.39E-06 | –     | –    | 3.90E-03 | –      | 5.81E-05 | 2.30E-03 | 1.93E-06 | –     | –      |
| hsa05206: MicroRNAs in cancer |                                     | –      | –    | –    | 2.10E-04 | 2.28E-02 | 6.20E-04 | –      | 5.70E-03 | –     | –      |
| hsa-mir-30a: Viral carcinogenesis |                                      | 5.50E-04 | –     | –    | –      | –      | 1.80E-03 | –      | –      | 2.14E-02 | –      |
| hsa05130: Pathogenic *Escherichia coli* infection |                            | 1.24E-02 | –     | –    | 2.28E-02 | –      | –      | 5.70E-03 | –     | –      | –      |
| hsa03013: RNA transport |                                         | –      | –    | –    | –      | –      | –      | 2.00E-03 | 5.70E-03 | –     | –      |
| hsa05206: MicroRNAs in cancer |                                        | –      | –    | 9.80E-03 | –      | 1.97E-02 | –      | 1.80E-04 | –     | –      | –      |
| hsa04550: Signaling pathways regulating pluripotency of stem cells | 1.20E-03 | –     | –    | –    | –      | –      | 6.20E-04 | –      | –      | –      |
| hsa-let-7i: HIF-1 signaling pathway |                                        | 1.20E-03 | –     | –    | –      | –      | 6.80E-04 | –      | –      | –      |
| hsa05130: Pathogenic *Escherichia coli* infection |                                        | –      | –    | 1.60E-03 | –      | 1.97E-02 | –      | –      | –      | –      |
| hsa05225: Hepatocellular carcinoma |                                         | 1.01E-02 | –     | –    | –      | –      | –      | –      | –      | –      | –      |
| hsa04110: Cell cycle |                                        | 1.29E-06 | 1.57E-07 | –    | –      | 9.85E-09 | 5.43E-05 | 4.19E-10 | 4.40E-10 | –     | –      |
| hsa03030: DNA replication |                                        | 2.13E-07 | 3.59E-06 | –    | –      | 1.10E-04 | –      | 1.20E-03 | 1.03E-07 | –     | –      |
| hsa-mir-24: Cellular senescence |                                      | 4.00E-03 | 4.00E-03 | –    | –      | 1.10E-04 | 1.50E-04 | 9.90E-04 | 4.20E-04 | –     | –      |
| hsa04115: p53 signaling pathway |                                        | 7.49E-05 | 1.30E-03 | –    | –      | 8.40E-04 | –      | 1.20E-03 | –     | –      | –      |
| hsa00240: Pyrimidine metabolism |                                         | 4.50E-04 | –     | –    | –      | –      | –      | –      | 4.28E-05 | –     | –      |
| hsa05211: Renal cell carcinoma |                                        | 4.20E-03 | 4.93E-02 | –    | –      | –      | –      | 4.30E-03 | 4.10E-03 | –     | –      |
| hsa05220: Chronic myeloid leukemia |                                        | –      | –    | 4.93E-02 | –      | –      | 2.28E-02 | –      | –      | –     | –      |
| hsa05219: Bladder cancer |                                        | –      | –    | 4.93E-02 | –      | –      | 4.90E-03 | –      | –      | –     | –      |
| hsa05206: MicroRNAs in cancer |                                        | –      | –    | 4.93E-02 | –      | –      | 2.79E-02 | –      | –      | –     | –      |
| hsa05203: Viral carcinogenesis |                                        | –      | –    | 4.93E-02 | –      | –      | 1.67E-02 | –      | –      | –     | –      |
**TABLE 8** | Five significant GO biological processes for each of the 17 selected miRNAs in 10 cancer types.

| miRNA    | GO biological process                                           | BLCA | BRCA | COAD | GBM | HNSC | KIRC | LUAD | LUSC | OV | UCEC |
|----------|-----------------------------------------------------------------|------|------|------|-----|------|------|------|------|----|------|
| GO:0051239 | regulation of multicellular organismal process                 | 1.30E-04 | –     | 4.34E-05 | –   | –    | –    | 4.70E-04 | –   | –   | – |
| GO:0023051 | regulation of signaling                                        | –    | –     | 6.76E-05 | –   | 6.10E-04 | –    | 4.70E-04 | –   | –   | – |
| hsa-mir-205 | regulation of cell communication                                | –    | –     | 6.76E-05 | –   | 7.30E-04 | –    | 4.70E-04 | –   | –   | – |
| GO:0060255 | regulation of macromolecule metabolic process                  | –    | –     | 3.87E-07 | –   | 7.21E-06 | –    | –    | –   | 7.59E-05 | – |
| GO:0051171 | regulation of nitrogen compound metabolic process              | –    | –     | 3.87E-07 | –   | 2.61E-05 | –    | –    | –   | 7.59E-05 | – |
| GO:0006139 | nucleobase-containing compound metabolic process               | –    | 8.72E-05 | –     | –   | –    | –    | –    | 2.90E-04 | 1.10E-04 | – |
| hsa-mir-10a | cellular component organization or biogenesis                 | 2.33E-02 | –     | –     | –   | 1.39E-06 | –    | –    | –   | –   | – |
| GO:0034641 | cellular nitrogen compound metabolic process                   | 2.33E-02 | –     | –     | –   | 7.28E-05 | –    | –    | –   | –   | – |
| GO:0016043 | cellular component organization                                | 2.33E-02 | –     | –     | –   | 7.28E-05 | –    | –    | –   | –   | – |
| hsa-mir-10b | gene expression                                                 | –    | 1.70E-04 | –     | –   | –    | –    | –    | 8.00E-04 | –   | – |
| GO:0048523 | negative regulation of cellular process                        | 8.93E-06 | –     | 1.92E-07 | 2.60E-04 | –    | 6.77E-05 | 1.30E-04 | 2.62E-06 | 1.20E-04 | – |
| GO:0048519 | negative regulation of biological process                      | 2.58E-05 | –     | 1.92E-07 | 2.60E-04 | –    | 6.77E-05 | –    | 5.64E-06 | –   | – |
| hsa-mir-196b | response to oxygen levels                                      | –    | –     | 5.06E-06 | –   | 4.43E-05 | 1.18E-05 | 1.20E-04 | –   | –   | – |
| GO:1901700 | response to oxygen-containing compound                         | –    | –     | –     | –   | 1.00E-04 | –    | –    | 1.11E-05 | 3.28E-05 | 4.54E-05 |
| GO:0051173 | positive regulation of nitrogen compound metabolic process     | 2.58E-06 | –     | 1.62E-06 | –   | –    | –    | –    | 1.10E-04 | –   | – |
| GO:0009653 | anatomical structure morphogenesis                              | 2.70E-04 | –     | 3.12E-05 | –   | –    | –    | –    | 1.20E-04 | –   | – |
| GO:1901564 | organonitrogen compound metabolic process                      | –    | –     | 2.80E-04 | 9.30E-03 | 9.50E-03 | –    | –    | –   | –   | – |
| hsa-mir-10a | cell development                                                | 2.50E-04 | –     | 1.10E-04 | –   | –    | –    | –    | –   | –   | – |
| GO:1903047 | mitotic cell cycle process                                      | 3.30E-03 | –     | –     | –   | 2.65E-02 | –    | –    | –   | –   | – |
| GO:0044237 | cellular metabolic process                                      | –    | –     | 7.10E-04 | –   | 7.00E-03 | –    | –    | –   | –   | – |
| GO:0048522 | positive regulation of cellular process                         | –    | 1.25E-05 | 5.30E-03 | –   | –    | –    | 5.46E-07 | –   | 5.80E-04 | 5.50E-03 |
| GO:0031325 | positive regulation of cellular metabolic process               | –    | 1.51E-05 | 7.60E-03 | –   | –    | –    | 5.82E-06 | –   | –   | – |
| hsa-mir-375 | positive regulation of intracellular signal transduction       | 5.80E-03 | –    | –     | –   | –    | –    | –    | 5.60E-04 | –   | – |
| GO:0051173 | positive regulation of nitrogen compound metabolic process     | –    | 1.51E-05 | –     | –   | 3.61E-02 | –    | –    | –   | –   | – |
| GO:0006139 | cellular component organization or biogenesis                  | –    | 8.40E-03 | –     | –   | –    | –    | –    | 5.60E-04 | –   | – |
| GO:0042373 | cellular metabolic process                                      | –    | 2.84E-02 | –     | –   | –    | –    | –    | 2.30E-03 | –   | – |
| GO:0008152 | metabolic process                                               | –    | 2.84E-02 | –     | –   | –    | –    | –    | 8.40E-03 | –   | – |
| hsa-mir-143 | positive regulation of protein localization to membrane        | –    | 2.84E-02 | –     | –   | –    | –    | –    | –   | –   | – |
| GO:2032950 | positive regulation of cellular protein localization            | –    | 2.84E-02 | –     | –   | –    | –    | –    | –   | –   | – |
| GO:1903829 | positive regulation of cellular protein localization            | –    | 2.84E-02 | –     | –   | –    | –    | –    | –   | –   | – |
| GO:0066807 | nitrogen compound metabolic process                             | –    | 2.84E-02 | –     | –   | –    | –    | –    | –   | –   | – |
| miRNA     | GO biological process                                                                 | BLCA    | BRCA    | COAD    | GBM     | HNSC    | KIRC    | LUAD    | LUSC    | OV      | UCEC    |
|-----------|----------------------------------------------------------------------------------------|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|
| GO:0071840| cellular component organization or biogenesis                                         | –       | –       | 4.50E-03| –       | 6.50E-04| –       | –       | 7.68E-06| –       | –       |
| GO:0016043| cellular component organization                                                        | –       | –       | 5.70E-03| –       | 6.50E-04| –       | –       | 1.49E-05| –       | –       |
| hsa-let-7c| positive regulation of macromolecule metabolic process                                | –       | –       | –       | 3.95E-09| –       | –       | –       | 1.20E-04| 1.50E-03| –       |
| GO:0090304| nucleic acid metabolic process                                                         | 4.30E-04| –       | –       | –       | –       | –       | –       | 4.41E-05| –       | –       |
| GO:0051252| regulation of RNA metabolic process                                                   | 1.30E-03| –       | –       | –       | –       | –       | –       | 2.16E-02| –       | –       |
| GO:0048522| positive regulation of cellular process                                                | 5.54E-07| 5.57E-06| 5.00E-03| 8.70E-03| –       | –       | –       | 1.44E-07| 1.40E-04| –       |
| GO:0048519| negative regulation of biological process                                             | –       | –       | –       | 8.70E-03| 9.40E-04| 3.38E-02| 2.88E-06| –       | –       | –       |
| hsa-mir-107| negative regulation of nitrogen compound metabolic process                            | 5.54E-07| –       | –       | –       | –       | 2.31E-06| –       | 6.70E-05| –       | –       |
| GO:0048518| positive regulation of biological process                                             | 1.17E-06| 1.07E-05| –       | –       | –       | –       | –       | 2.42E-07| –       | –       |
| GO:0080090| regulation of primary metabolic process                                                | 1.31E-06| –       | –       | –       | –       | –       | –       | 9.80E-04| –       | –       |
| GO:0044237| cellular metabolic process                                                            | –       | 1.07E-02| –       | –       | –       | –       | –       | 4.46E-05| 2.30E-04| –       |
| hsa-mir-378| organic substance biosynthetic process                                                | –       | 1.07E-02| –       | –       | –       | –       | –       | 4.46E-05| –       | –       |
| GO:0044249| cellular biosynthetic process                                                         | –       | 1.07E-02| –       | –       | –       | –       | –       | 4.46E-05| –       | –       |
| GO:0034645| macromolecule biosynthetic process                                                    | –       | 1.07E-02| –       | –       | –       | –       | –       | 1.25E-02| –       | –       |
| GO:0009058| biosynthetic process                                                                  | –       | 1.07E-02| –       | –       | –       | –       | –       | 1.46E-05| –       | –       |
| GO:0071495| cellular response to endogenous stimulus                                              | 4.66E-05| –       | –       | –       | –       | –       | –       | 2.21E-05| 7.70E-04| 8.10E-03|
| GO:0017071| cellular response to oxygen-containing compound                                        | 4.40E-04| –       | –       | 3.10E-04| 7.70E-04| –       | –       | –       | –       | –       |
| hsa-mir-133a| cellular response to organonitrogen compound                                          | 7.30E-04| –       | –       | 8.70E-03| 3.10E-04| –       | –       | –       | –       | –       |
| GO:0048518| positive regulation of biological process                                             | –       | –       | 5.00E-03| 8.70E-03| –       | –       | –       | 9.23E-05| –       | –       |
| GO:0065008| regulation of biological quality                                                      | –       | –       | 1.50E-03| –       | –       | –       | –       | 7.70E-04| –       | –       |
| GO:0007049| cell cycle                                                                             | 1.76E-09| 8.75E-06| –       | –       | 6.45E-08| –       | 2.12E-06| 4.40E-04| 9.80E-04| –       |
| GO:001576 | chromosome organization                                                                | 1.52E-07| 3.59E-05| –       | –       | 9.12E-07| –       | 5.41E-07| 7.92E-05| –       | –       |
| hsa-mir-1 | DNA-dependent DNA replication                                                          | 4.12E-08| 1.40E-04| –       | –       | –       | –       | –       | 6.25E-07| –       | –       |
| GO:0023402| cell cycle process                                                                     | 4.03E-07| –       | –       | –       | 2.18E-06| –       | –       | 1.70E-03| –       | –       |
| GO:0006259| DNA metabolic process                                                                  | –       | 1.70E-04| –       | 5.77E-06| –       | 5.41E-07| –       | –       | –       | –       |
| GO:0010033| response to organic substance                                                         | –       | 1.03E-02| 5.58E-05| –       | –       | –       | –       | 1.97E-05| –       | –       |
| GO:0062555| regulation of macromolecule metabolic process                                        | 3.20E-04| –       | –       | 1.20E-04| –       | –       | –       | –       | –       | –       |
| hsa-mir-30c| cellular response to chemical stimulus                                                | –       | 1.03E-02| 2.00E-03| –       | –       | –       | –       | –       | –       | –       |
| GO:0016912| vesicle-mediated transport                                                            | –       | 1.03E-02| –       | 3.60E-03| –       | –       | –       | –       | –       | –       |
| GO:0023766| immune system process                                                                  | –       | 1.03E-02| –       | –       | –       | –       | –       | 2.06E-05| –       | –       |
| GO:0051239| regulation of multicellular organism process                                         | 4.64E-05| 5.02E-08| –       | –       | –       | –       | –       | –       | –       | –       |
| GO:0048731| system development                                                                    | 4.10E-04| –       | –       | –       | –       | –       | –       | 5.33E-08| –       | –       |
| hsa-mir-16 | regulation of cell differentiation                                                    | 4.10E-04| –       | –       | –       | –       | –       | –       | 2.61E-08| –       | –       |
| GO:0050793| regulation of developmental process                                                   | –       | 1.11E-07| –       | –       | –       | –       | –       | 2.61E-08| –       | –       |

(Continued)
| miRNA         | GO biological process                                      | FDR-corrected p-value |
|--------------|------------------------------------------------------------|-----------------------|
|              |                                                            | BLCA    | BRCA  | COAD  | GBM   | HNSC  | KIRC  | LUAD  | LUSC  | OV    | UCEC  |
| GO:2000026   | regulation of multicellular organismal development         |         | 1.29E-07 |       |       |       |       |       |       | 1.50E-03 |       |
| GO:0071840   | cellular component organization or biogenesis              |         |         |       | 1.30E-04 | 6.20E-03 | 1.02E-06 |       |       | 2.04E-02 |       |
| GO:1903047   | mitotic cell cycle process                                  |         | 1.99E-02 |       |       |       |       |       |       | 6.20E-03 | 8.40E-06 |       |
| hsa-mir-30a  | nucleic acid metabolic process                              |         |         |       | 6.20E-04 |       |       |       |       |       | 2.73E-02 |       |
|              | regulation of vascular smooth muscle cell proliferation     |         |         |       |         |       |       |       |       |       |       |       |
| GO:0071900   | regulation of protein serine/threonine kinase activity      |         |         |       |         |       |       |       |       |       |       |       |
| GO:0061061   | muscle structure development                                |         |         |       |         |       |       |       |       |       |       |       |
| hsa-let-7i   | regulation of vascular smooth muscle cell proliferation     | 1.05E-02 |         |       |         |       |       |       |       |       |       |       |
| GO:0070749   | cell cycle                                                  | 1.13E-13 | 1.08E-08 |       | 2.6E-11 | 1.14E-08 | 1.16E-13 |       |       | 2.32E-09 |       |
| GO:0000278   | mitotic cell cycle                                          | 1.72E-13 | 1.08E-08 |       | 4.26E-13 | 7.08E-05 | 2.78E-10 | 1.17E-12 |       |       |       |
| hsa-mir-24   | mitotic cell cycle process                                   | 1.16E-11 | 4.58E-08 |       | 1.04E-10 | 7.08E-05 | 6.94E-10 | 1.17E-12 |       |       |       |
| GO:0024402   | cell cycle process                                          | 3.15E-11 |         |       | 5.20E-11 | 9.55E-05 | 2.75E-07 | 9.89E-13 |       |       |       |
| GO:0044772   | mitotic cell cycle phase transition                          | 2.15E-10 | 4.58E-08 |       |         | 4.86E-10 | 1.51E-11 |       |       |       |       |
| GO:0060255   | regulation of macromolecule metabolic process               | 3.22E-06 | 2.03E-05 | 8.31E-07 | 6.89E-05 | 1.30E-03 | 5.67E-06 | 2.05E-05 | 1.88E-06 | 1.03E-05 | 9.40E-04 |
| GO:0080090   | regulation of primary metabolic process                     | 3.22E-06 | 2.03E-05 | 8.31E-07 |       | 5.80E-05 | 2.96E-05 | 1.19E-05 | 9.40E-04 |       |       |
| hsa-mir-95   | regulation of nitrogen compound metabolic process           | 3.22E-06 | 2.03E-05 | 8.31E-07 |       | 9.97E-06 | 5.80E-05 | 2.96E-05 | 1.19E-05 |       |       |
| GO:0050789   | regulation of biological process                            | 9.66E-07 |         | 1.93E-07 | 6.89E-05 | 1.30E-03 | 9.97E-06 |       |       | 4.16E-06 |       |
| GO:0065007   | biological regulation                                       | 3.22E-06 | 8.31E-07 | 6.89E-05 | 1.30E-03 |       |       |       |       | 1.19E-05 | 9.40E-04 |
TABLE 9 | Association of top 30 proteins in 10 cancer types for the 17 selected miRNAs through their targets.

| TF     | BLCA | BRCA | COAD | GBM | HNSC | KIRC | LUAD | LUSC | OV | UCEC |
|--------|------|------|------|-----|------|------|------|------|----|------|
| MYC    | 34   | 28   | 41   | 85  | 33   | 133  | 54   | 25   | 38 | 32   |
| VEGFA  | 23   | 10   | 15   | 53  | 23   | 18   | 36   | 36   | 48 | 39   |
| AKT1   | 17   | 59   | 50   | 16  | 54   | 32   | 0    | 16   | 17 | 8    |
| RRm2   | 17   | 28   | 0    | 11  | 21   | 23   | 30   | 32   | 10 | 10   |
| CDK1   | 23   | 31   | 10   | 13  | 21   | 0    | 21   | 30   | 0  | 24   |
| CDKN1A | 20   | 19   | 17   | 15  | 15   | 18   | 20   | 17   | 8  | 10   |
| UHRF1  | 18   | 29   | 1    | 7   | 21   | 23   | 25   | 14   | 0  | 5    |
| CHEK1  | 24   | 22   | 0    | 0   | 21   | 0    | 24   | 28   | 0  | 22   |
| H2AFX  | 32   | 16   | 9    | 0   | 10   | 0    | 16   | 21   | 8  | 24   |
| MCM10  | 20   | 22   | 0    | 0   | 19   | 11   | 20   | 23   | 0  | 18   |
| POLD1  | 33   | 26   | 0    | 0   | 15   | 14   | 16   | 30   | 0  | 0    |
| IL6    | 11   | 12   | 7    | 29  | 8    | 0    | 15   | 15   | 14 | 16   |
| RHOA   | 11   | 15   | 9    | 0   | 0    | 8    | 30   | 18   | 22 | 12   |
| PCNA   | 20   | 26   | 0    | 0   | 0    | 0    | 25   | 40   | 0  | 17   |
| DTL    | 13   | 18   | 0    | 3   | 12   | 13   | 13   | 23   | 10 | 16   |
| CCNF   | 6    | 18   | 5    | 14  | 7    | 18   | 18   | 14   | 19 | 0    |
| BRCA1  | 27   | 12   | 0    | 0   | 25   | 12   | 7    | 25   | 12 | 0    |
| CDC42  | 8    | 0    | 22   | 17  | 0    | 13   | 12   | 23   | 0  | 19   |
| PTEN   | 15   | 10   | 7    | 9   | 3    | 0    | 25   | 25   | 18 | 0    |
| WWHAZ  | 5    | 12   | 13   | 9   | 27   | 4    | 11   | 13   | 7  | 3    |
| PAICS  | 7    | 19   | 0    | 5   | 9    | 0    | 23   | 27   | 1  | 4    |
| PIK3R1 | 7    | 19   | 7    | 7   | 10   | 0    | 16   | 8    | 9  | 5    |
| UBAS2  | 11   | 0    | 8    | 10  | 0    | 16   | 18   | 16   | 0  | 8    |
| CTGF   | 11   | 12   | 10   | 8   | 5    | 4    | 16   | 13   | 5  | 0    |
| KIF4A  | 17   | 15   | 0    | 0   | 10   | 8    | 15   | 11   | 10 | 0    |
| MTO1R  | 9    | 8    | 9    | 0   | 9    | 14   | 7    | 9    | 8  | 8    |
| UBE2C  | 13   | 13   | 0    | 0   | 11   | 8    | 12   | 15   | 0  | 11   |
| KIF2C  | 16   | 15   | 0    | 0   | 11   | 0    | 16   | 12   | 9  | 0    |
| KIF18B | 11   | 11   | 0    | 0   | 11   | 8    | 11   | 12   | 0  | 11   |
| CHAF1B | 10   | 13   | 0    | 4   | 12   | 0    | 12   | 9    | 5  | 2    |

covariance matrix adaptation evolutionary strategy (CMA-ES), forward selection (FS), and a classification technique. Features are reordered using SNE by performing clustering of highly correlated features. From the result of the clustering, a subset of features is randomly selected to perform multi-class classification on 10 cancer types. Although the features are randomly selected, the underlying classification task is treated as an optimization problem for CMA-ES in order to find the features automatically. Thereafter, a final set of features/miRNAs is obtained using forward selection. The results of the first part of SCES-FS have been compared with the results of some well-known feature selection methods, including ESVM-RFE, LASSO, NSGA-II-SE, MOGA, SVM-nRFE, SVM-RFE, CMIM, ICAP, SCAD, JMI, CIFE, mRMR, FSCOX, DISR, SNRs, and RankSum, as well as the result with all features, in terms of classification accuracy. The SCES-FS method selected 17 putative miRNAs associated with 10 cancer types and achieved higher classification accuracy than the other methods. Using the 17 selected miRNAs, a web-based multi-class cancer predictor application has been developed. These selected miRNAs are used in the second part of the proposed method, which employs Cox regression analysis to examine their importance with respect to survival of particular types of cancer. The analysis uses data on expression of the 17 selected miRNAs together with clinical data. A high Cox coefficient value signifies the importance of an miRNA for a particular cancer type. For example, it is found that hsa-mir-375 has the highest Cox coefficient, 0.9882, for the glioblastoma multiforme cancer type. Similar results were obtained for other miRNAs, associated with different cancer types. The up- and down-regulations of the 17 selected miRNAs have been computed based on the ANOVA test. Furthermore, network analysis, expression analysis using hierarchical clustering, KEGG pathway analysis, GO enrichment analysis, and PPI network analysis have been performed to assess the biological significance of the selected miRNAs. The network analysis revealed the association of different cancer types with each pair of miRNA and its target mRNA. The hierarchical clustering
analysis demonstrated the effective changes in expression levels of the miRNAs between tumor and normal samples. Both the KEGG and GO enrichment analyses reveals the significant pathways and biological functions in different cancer types. Moreover, using PPI networks, key cancer regulators such as MYC, VEGFA, AKT1, CDKN1A, RHOA, and PTEN are identified. All these evidences suggest that our selected miRNAs play key roles in the development of 10 different types of cancer. A future research direction is the integration of multi-omics data for finding effective regulators pan-cancer biomarkers.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: https://portal.gdc.cancer.gov/ and http://www.nitttrkol.ac.in/indrajit/projects/mirna-prediction-multicalss/.

AUTHOR CONTRIBUTIONS

JPS, IS, and AL conceived and designed the experiments. JPS, IS, AL, NG, AD, and PL performed the experiments. JPS, IS, MW, and AD wrote the manuscript. JPS, IS, GB, and DP corrected and edited the manuscript. All authors read and approved the final manuscript.

FUNDING

This work was supported by the Polish National Science Centre (2014/15/B/ST6/05082, 2019/35/O/ST6/02484), the Foundation for Polish Science (TEAM to DP), and a grant from the Department of Science and Technology, India (Indo-Polish/Polish-Indo project no. DST/INT/POL/P-36/2016). The work was also supported by grant 1U54DK107967-01 Nucleome Positioning System for Spatiotemporal Genome Organization and Regulation within the 4DNucleome NIH.
program, and was partially supported by the Academy of Development as the Key to Strengthen Human Resources of the Polish Society co-financed by the European Union under the European Social Fund. This research was also partially supported under the RENOIR Project of the European Union Horizon 2020 Research and Innovation Programme under the Marie Skłodowska-Curie grant agreement no. 691152 and by the Ministry of Science and Higher Education of Poland under grant nos. W3/H2020/2016 and 329025/P/zH1/2016.

**SUPPLEMENTARY MATERIAL**

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.00982/full#supplementary-material
