Journal of Biomedical Informatics 97 (2019) 103254

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Genome-wide analysis of multi-view data of miRNA-seq to identify miRNA biomarkers for stomach cancer

Namrata Pant[a,1,2], Somnath Rakshit[b,c,1,2], Sushmita Paul[a], Indrajit Saha[b,*,2]

[a] Department of Bioscience and Bioengineering, Indian Institute of Technology, Jodhpur, India
[b] Department of Computer Science and Engineering, National Institute of Technical Teachers' Training and Research, Kolkata, India
[c] Laboratory of Functional and Structural Genomics, Center of New Technologies, University of Warsaw, Warsaw, Poland

ARTICLE INFO

Keywords:
Ensemble ranking
Feature selection
miRNA
Next generation sequencing
Stomach cancer

ABSTRACT

Stomach cancer is one of the leading causes of cancer-related deaths worldwide. More than 80% diagnosis of this cancer occur at later stages leading to low 5-year survival rate. This emphasizes the need to have better prognostic techniques for stomach cancer. In this regard, the Next-Generation Sequencing of whole genome and multi-view approach to omics may reveal the underlying molecular complexity of stomach cancer using high throughput expression data of miRNA. Generally, miRNAs are small, non-coding RNAs, which cause down-regulation of target mRNAs. They also show differential expression for a specific biological condition like stage or histological type of stomach cancer, highlighting their importance as potential biomarkers. Analyzing miRNA expression data is a challenging task due to the existence of large number of miRNAs and less sample size. A small set of miRNAs will be helpful in designing efficient diagnostic and prognostic tool. In this regard, here a computational framework is proposed that selects different sets of miRNAs for five different categories of clinical outcomes viz. condition, clinical stage, age, histological type, and survival status. First, the miRNAs are ranked using four feature ranking methods. These ranks are used to find an ensemble rank based on adaptive weight. Second, the top 100 miRNAs from each category are used to find the miRNAs that are common to all categories as well as miRNAs that belong to only one category. Finally, the results have been validated quantitatively and through biological significance analysis.

1. Introduction

Stomach cancer or gastric cancer, is the fifth most common cancer among men and seventh most common cancer among women in India [1]. Both environmental as well as genetic factors lead to the on set and progression of the disease. Stomach cancer usually does not cause any deterministic early symptoms, making it difficult to be diagnosed at early stages. The worldwide statistics reflect the fact that most people with stomach cancer are diagnosed after the cancer has already spread to other parts of the body. This leads to the 5-year survival rate for people with stomach cancer being 31%[3]. With the advancement in medicine and health care services, the worldwide incidence and mortality of this disease has declined over the past few years, but poor prognosis still persists. Here lies the importance of looking into the problem from molecular perspective.

Whole-genome or whole-transcriptome analysis using Next-Generation Sequencing (NGS) technology has identified genetic and epigenetic modifications in various types of cancers. NGS platforms available today like Roche-454, ABI/SOLiD3 and Illumina/Solexa are able to extract large set of information from genomic sequences [2]. The importance of microRNAs (miRNAs) as regulators of oncogenesis and potential biomarkers for stomach cancer has already been investigated [3]. miRNAs are a class of small (~22 nucleotide long) RNAs that are involved in post-transcriptional regulation of gene expression [4]. They are known to regulate a number of cellular processes like differentiation, growth and metabolism [5]. miRNAs bind to complementary regions in the target mRNAs [6] and result in negative regulation of the mRNA expression [7]. Abnormal miRNA expression is, therefore, able to reflect the changes in gene expression in diseased condition like stomach cancer and can therefore be regarded as

[*] Corresponding author.
E-mail address: indrajit@nitttrkol.ac.in (I. Saha).
[1] Joint first authors.
[2] Contributed equally.
[3] https://seer.cancer.gov/statfacts/html/stomach.html.

https://doi.org/10.1016/j.jbi.2019.103254
Received 24 March 2019; Received in revised form 3 July 2019; Accepted 17 July 2019
Available online 25 July 2019

potential biomarker for the same. miRNA expression data consists of large number of features (miRNAs) which makes it difficult to be analyzed using traditional linear statistical methods only. Therefore, a suitable computational method is required for selecting set of miRNAs from such data that are significantly altered in stomach cancer and can be utilized for further analyses. In the past, various methods have been proposed for ranking of miRNAs using statistical and machine learning techniques [8,9]. Each of these methods use different criteria for feature selection like t-test or information gain. However, none of them takes into account the different clinical categories of the samples in miRNA expression data or ensemble of methods for ranking of miRNAs. Most of the studies have been conducted using single view of data.

In this study, a computational framework is proposed, which analyzes the miRNA expression of 524 miRNAs by categorizing the samples on the basis of condition, age group, clinical stage, histological type and survival status by using the information from different feature ranking techniques. Hence, the proposed approach is multi-view based study. Here, the computational and biological motivations of this work are to take the advantages of the different feature selection methods in order to rank the miRNAs based on ensemble of ranks and to find sets of miRNAs that are effective overall for stomach cancer and different clinical outcomes. Therefore, the miRNAs that are present in all five categories may help in designing efficient prognostic tool. In this regard, at first, the miRNAs are ranked individually using four well known feature ranking methods viz. Conditional Mutual Information Maximisation (CMIM) [10], Double Input Symmetrical Relevance (DISR) [11], Interaction Capping (ICAP) [12] and Conditional Informative Feature Extraction (CIFE) [13] for each category. Second, such ranks are considered to compute the Weighted Ensemble of Ranks (WER) in order to make the final ranking of miRNAs. Next, the top 100 miRNAs across all five categories are used to plot a Venn diagram to find the miRNAs that are common to all five categories and the miRNAs that are present in each category. The common miRNAs are then used for classification across all five categories and their classification accuracy is compared with the individial feature selection methods. Furthermore, the selected set of miRNAs are analyzed by miRNA-Gene-Transcription Factor (TF) network, PPI network, expression analysis, KEGG and GO enrichment analysis to see their role in cancer pathways.

2. Material and method

In this section, brief description of the feature selection methods, the miRNA expression data, its categorization, and the methods used to identify the miRNA biomarkers are described.

2.1. A brief description of the feature selection methods

Conditional Mutual Information Maximisation (CMIM) [10] is a fast feature selection method that is based on conditional mutual information. It iteratively selects features that maximize their mutual information based on the class to predict. This is done conditionally to the response to the already-picked features. CMIM does not select any feature that is similar to a feature that has been already picked since it does not bring much additional information about the class that it predicts even though it may be individually powerful. Feature selection in CMIM is based on conditional mutual information as shown in Eq. (1) where X, Y and Z are finite random variables, J_{CMIM} is the conditional mutual information and $H(X)$ is the entropy of a random variable X.

$$J_{CMIM}(X;Y|Z) = H(X|Z) - H(X|Z, Y) \tag{1}$$

Double Input Symmetrical Relevance (DISR) [11] is a filter-based method to select feature variables from large dimensional datasets. It is based on the double input symmetric relevance. The idea behind this approach is that a set of variables can together give more information than the sum of information given by each variable individually. This criterion can be used to select the subset amongst a finite number of alternative subsets which returns the maximum amount of information about the output class. Considering two random variables as X, Y and a joint probability distribution $P(X, Y)$, the symmetrical relevance $SR(X, Y)$ is defined in Eq. (2). Here, I is the conditional mutual information of the random variables and H is the entropy.

Table 1
Categorisation of samples in five categories based on clinical information.

Category	Group	Number of patients	Number of miRNAs
Condition	Tumour	231	524
	Normal	33	
Age	Group I	82	
	Group II	177	
Survival Status	Living	202	
	Deceased	62	
Clinical Stage	Group I	146	
	Group II	114	
Histological Type	Stomach Adenocarcinoma	173	
	Stomach Intestinal Adenocarcinoma	91	

$$SR(X, Y) = \frac{I(X, Y)}{H(X, Y)} \tag{2}$$

Using SR in Eq. (2), the resulting criterion for DISR is given in Eq. (3).

$$J_{DISR} = \arg\max_{X_i \in X_s}\{\sum_{X_j \in X_s} SR(X_{i,j};Y)\} \tag{3}$$

Interaction Capping (ICAP) [12] is also a filter-based method that selects the optimum attributes from the data. It relies on maximizing feature interaction. Here, Naive Bayes classifier is considered which assumes independence between attributes. It assumes all attributes and selects the attribute pairs with highest interaction information. Then it constructs a joint attribute using the selected pairs of attributes such that each distinct pair of values of original attributes maps to a distinct value of a new attribute. Its scoring criterion is defined in Eq. (4):

$$J_{ICAP}(X_i) = I(X_i;Y) - \sum_{X_j \in S} max[0, I(X_i;X_j) - I(X_i;X_j|Y)] \tag{4}$$

Here, J_{ICAP} stands for Joint Mutual Information (JMI) and I is the mutual information, for attributes X and Y.

Conditional Informative Feature Extraction (CIFE) [13] is another information maximization based feature selection method that combines the class relevance factor as well as the redundancy factor to depict the information structure. It is based on the theorem that if the communication of any two features is not affected by other features, the joint class-relevant information equals the sum of the individual feature information minus the total pairwise redundancies. The objective function of CIFE is defined in Eq. (5),

$$J_{CIFE} = \arg\max_{J}\{I(Y^{(i)};c) - \sum_{j=1}^{i-1} R_c(Y^{(j)};Y^{(c)})\} \tag{5}$$

where I is the joint class-relevant information, Y is the target feature, R is the redundancy factor, c is the true underlying feature, i is the total number of features, and J_{CIFE} is the calculated combined parameter.

2.2. Data preparation

The expression data of miRNA used in this study is taken from the Cancer Genome Atlas (TCGA) [14]. The data is used in reads per million (RPM) log2 normalized form. Initially, the data contained 2588 miRNAs for 264 patients. After removing the miRNAs with more than 60% zero expression, 524 miRNAs have been obtained. This data is then categorized into five categories on the basis of clinical outcomes that

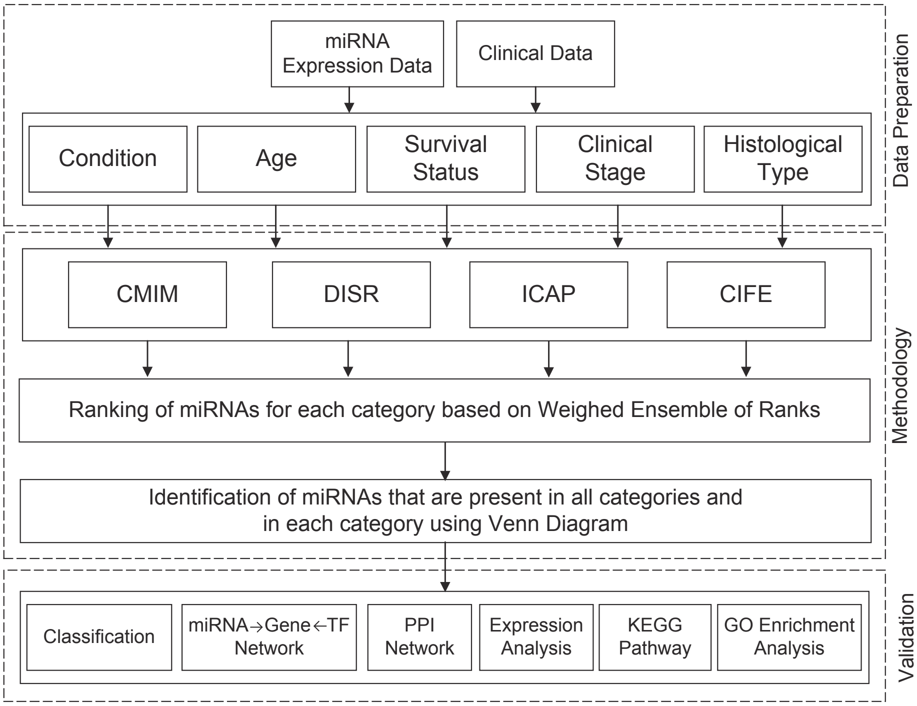


Fig. 1. Steps of the proposed framework to rank the miRNAs and identify the most relevant miRNAs for common and five individual categories.

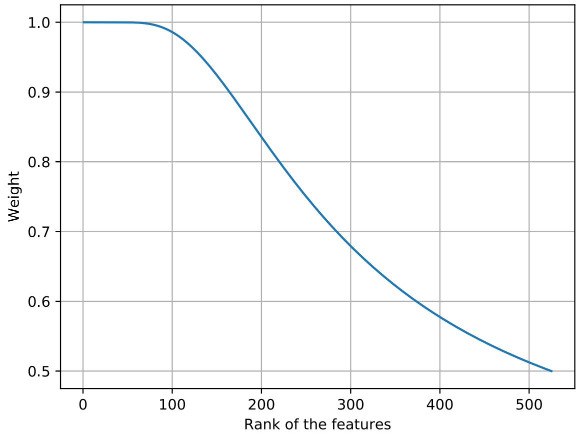


Fig. 2. Line chart showing the distribution of Eq. (7) for the total number of miRNAs in our dataset.

are also obtained from TCGA. For each category, the patients have been divided into two groups. In condition category, the patients are categorized according to the available clinical information about the disease condition of the patient viz. Tumour and Normal. On the other hand, in case of age category, the patients with age less than 60 are placed in Group I and those with age greater than or equal to 60 are placed in Group II. It is done by observing that the average age of the 264 patients is 60.4835 years. Similarly, in survival status category, the patients are grouped according to their vital status as 'Living' or 'Deceased'. In Clinical Stage category, the samples belonging to Clinical Stage I & II of stomach cancer are kept in Group I, while those belonging to Clinical Stage III & IV are kept in Group II. While in case of Histological Type category, the samples are divided on the basis of histological type of stomach cancer, stomach adenocarinoma or stomach intestinal adenocarcinoma. The statistics of samples in each category is mentioned in Table 1. Moreover, the refined dataset is provided in the supplementary website.

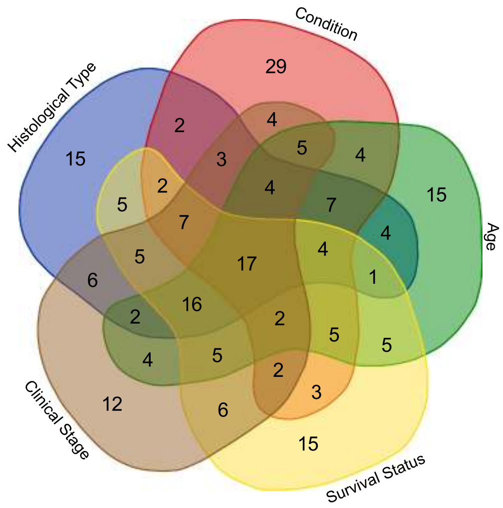


Fig. 3. Venn diagram used to determine the miRNAs that are present in all categories and in each category.

Table 2
Number of miRNAs obtained using each feature selection method for common and five individual categories.

Category	WER	CMIM	DISR	CIFE	ICAP	EFS
Common	17	8	22	43	6	1
Condition	29	30	57	33	32	32
Age	15	32	16	5	30	32
Survival Status	15	28	17	4	28	39
Clinical Stage	12	33	12	3	36	31
Histological Type	15	34	13	6	30	38

2.3. Method

This section describes the proposed framework to identify the sets of miRNAs for common and different clinical outcomes of stomach cancer.

Table 3
Selected miRNAs common in five categories.

miRNA	Avg. WER	Experiment	PMID
hsa-miR-196b-5p	3.9877	qRT-PCR	28053808
hsa-miR-205-5p	3.9736		
hsa-miR-215-5p	3.9531	qRT-PCR	26716895
hsa-miR-708-5p	3.9519	qRT-PCR	27322246
hsa-miR-217	3.9463	RT-PCR/MTT Assay	25869101
hsa-miR-194-5p	3.9142	Dual luciferase reporter assay	30405772
hsa-miR-192-3p	3.9020	qPCR	24981590
hsa-miR-210-3p	3.8789		
hsa-miR-429	3.8539	qRT-PCR	28521484
hsa-miR-31-3p	3.8473	Immunohistochemistry	19175831
hsa-miR-375	3.8458	qRT-PCR	21557705
hsa-miR-196a-5p	3.8194	qRT-PCR	28440445
hsa-miR-33b-5p	3.7922	Luciferase reporter assay	28436711
hsa-miR-511-5p	3.7888		
hsa-miR-200b-5p	3.7855	qRT-PCR	23851184
hsa-miR-146a-5p	3.7759	qPCR	24805774
hsa-miR-375	3.8458	qRT-PCR	21557705
hsa-miR-196a-5p	3.8194	qRT-PCR	28440445
hsa-miR-33b-5p	3.7922	Luciferase reporter assay	28436711
hsa-miR-511-5p	3.7888		
hsa-miR-200b-5p	3.7855	qRT-PCR	23851184
hsa-miR-146a-5p	3.7759	qPCR	24805774
hsa-miR-31-5p	3.6572		

Table 4
Selected miRNAs in condition category.

miRNA	WER	Experiment	PMID
hsa-miR-21-5p	4.0000	Microarray analysis/PCR	26824898
hsa-miR-202-5p	3.9912	qRT-PCR	30013369
hsa-miR-204-5p	3.9891	qRT-PCR	30013369
hsa-miR-378a-5p	3.9682		
hsa-miR-29b-2-5p	3.9377	qPCR/ Western Blot	30405849
hsa-let-7i-3p	3.8636		
hsa-miR-625-3p	3.8320		
hsa-miR-149-5p	3.8176	qRT-PCR	23144691
hsa-miR-874-3p	3.7114		
hsa-miR-29a-3p	3.6073	qRT-PCR	25889078
hsa-miR-101-5p	3.5486	Luciferase assay/ qRT-PCR	26460960
hsa-miR-539-5p	3.5258		
hsa-miR-940	3.4976	qRT-PCR	26456959
hsa-miR-26b-5p	3.4799	qRT-PCR	26172537
hsa-miR-27a-5p	3.4536	qRT-PCR	27409164
hsa-miR-3928-3p	3.4504		
hsa-miR-582-3p	3.4314		
hsa-miR-136-5p	3.4202	qRT-PCR	29541241
hsa-miR-582-5p	3.3968	qRT-PCR	29228422
hsa-miR-491-5p	3.3585	qRT-PCR/MTT assay	28358374
hsa-miR-30e-5p	3.3559		
hsa-miR-365a-3p	3.3546	qRT-PCR	24384510
hsa-miR-642a-5p	3.3123		
hsa-miR-24–1-5p	3.3080	qRT-PCR	24886316
hsa-miR-628-5p	3.2815		
hsa-miR-15b-3p	3.2747	qRT-PCR	18449891
hsa-miR-195-3p	3.2672	QF-RT-PCR	27097947
hsa-miR-337-3p	3.2494	qRT-PCR	24422944
hsa-miR-497-5p	3.2435		

Table 5
Selected miRNAs in age category.

miRNA	WER	Experiment	PMID
hsa-miR-937-3p	3.9987	qRT-PCR	29060929
hsa-miR-431-5p	3.9855		
hsa-miR-3131	3.9671		
hsa-miR-1229-3p	3.8826		
hsa-miR-584-5p	3.8710	Luciferase assay	28431583
hsa-miR-376a-5p	3.8665	qRT-PCR	30522118
hsa-miR-496	3.7244		
hsa-miR-214-5p	3.7190	qRT-PCR	23834902
hsa-miR-3934-5p	3.7064	qRT-PCR	29483646
hsa-miR-3613-5p	3.6450	qRT-PCR	24384510
hsa-miR-335-5p	3.6432	qRT-PCR	29215918
hsa-miR-505-5p	3.6400	qRT-PCR	30525214
hsa-miR-32-3p	3.6176		
hsa-miR-760	3.5969	qRT-PCR	24097871
hsa-miR-99a-3p	3.5776	qRT-PCR	27994509

Table 6
Selected miRNAs in survival status category.

miRNA	WER	Experiment	PMID
hsa-miR-188-3p	3.9722		
hsa-miR-514a-3p	3.8921	TaqMan assay	25167801
hsa-miR-744-3p	3.8733		
hsa-miR-3170	3.7953	Microarray analysis/qRT-PCR	22112324
hsa-miR-3614-3p	3.7948	qRT-PCR	24384510
hsa-miR-550a-5p	3.7542		
hsa-miR-200a-5p	3.7483		
hsa-miR-551a	3.7281	qRT-PCR	23248648
hsa-miR-589-3p	3.6884		
hsa-miR-363-3p	3.6688	qRT-PCR	30013369
hsa-miR-221-5p	3.6592		
hsa-miR-676-3p	3.6181	qRT-PCR	24616567
hsa-miR-504-5p	3.5922		
hsa-miR-320d	3.5767	qRT-PCR	29113415
hsa-miR-3687	3.5717	Cox regression analysis	30864737

Table 7
Selected miRNAs in clinical stage category.

miRNA	WER	Experiment	PMID
hsa-miR-323b-3p	4.0000		
hsa-miR-135b-3p	3.8877		
hsa-miR-142-5p	3.8573	Microarray analysis/qRT-PCR	21343377
hsa-miR-369-5p	3.8354		
hsa-miR-424-5p	3.7116	Luciferase/ FISH/MTT assay	28893265
hsa-miR-3065-3p	3.6882		
hsa-miR-200c-5p	3.6129	qRT-PCR	22954417
hsa-miR-16–2-3p	3.5972	Microarray analysis/qRT-PCR	22112324
hsa-miR-671-5p	3.5871	qRT-PCR	25897338
hsa-miR-1301-3p	3.5856		
hsa-miR-1226-3p	3.5636		
hsa-miR-218-5p	3.5285	qRT-PCR	22860003

In this regard, first, the steps followed to rank the miRNAs using various well known feature ranking methods are discussed. Second, the ensemble rank is computed using the ranks generated in the previous step. Third, the miRNAs that play an important role in all categories and in each category are identified. The entire framework for this approach is shown in Fig. 1.

2.3.1. *Ranking miRNAs using feature selection methods*

Four feature ranking methods viz. CMIM, DISR, ICAP and CIFE are used in the proposed framework. The steps used to rank the miRNAs by means of four feature ranking methods are described henceforth. The four feature ranking methods are used to rank all 524 miRNAs individually. Thereafter, each miRNA is assigned a rank by each of the method and thus, we obtain a rank matrix of size 524 × 4. This rank matrix is then used in the next stage to find the weighted ensemble of ranks for each miRNA.

2.3.2. *Computation of weighted ensemble of ranks*

From the earlier step, four ranks from four feature selection methods are obtained as a 524 × 4 matrix. Using these matrix, the Weighted Ensemble of Ranks (WER_j) for each miRNA is computed as in Eq. (6),

Table 8
Selected miRNAs in histological type category.

miRNA	WER	Experiment	PMID
hsa-miR-4326	3.9948		
hsa-miR-147b	3.9845		
hsa-miR-7-5p	3.9472		
hsa-miR-29b-1-5p	3.8976	Luciferase assay	30405849
hsa-miR-23b-5p	3.7740	HITS-CLIP	28903436
hsa-miR-708-3p	3.7634		
hsa-miR-501-3p	3.7447	qRT-PCR	28903436
hsa-miR-99a-5p	3.7358	qRT-PCR	28903436
hsa-miR-1306-3p	3.6335		
hsa-miR-20b-3p	3.6011		
hsa-miR-3648	3.5775		
hsa-miR-216a-5p	3.5722		

Table 9
Classification accuracy using Random Forest by considering the common set of miRNAs.

Category	WER	CMIM	DISR	CIFE	ICAP	EFS
Condition	0.939	0.920	0.970	0.943	0.920	0.807
Age	0.668	0.629	0.660	0.683	0.591	0.587
Survival Status	0.738	0.754	0.727	0.735	0.739	0.643
Clinical Stage	0.581	0.580	0.523	0.542	0.538	0.581
Histological Type	0.663	0.610	0.682	0.655	0.598	0.569
Sum Score	**3.590**	3.494	3.562	3.559	3.387	3.187

$$WER_j = \frac{\sum_{i=1}^{n}(W_i * R_i)}{A_j} \tag{6}$$

where i and j signify the number of feature selection methods (in this case, $n = 4$) and number of miRNAs. Here, the objective is to assign the weight (W_i) in such a way so that higher ranked miRNA gets higher weight as compared to the low ranked miRNAs. The weight, W_i is computed as in Eq. (7),

$$W_i = \frac{1}{1 + e^{1-\frac{N}{R_i}}} \tag{7}$$

where N is the total number of samples, R_i is the rank provided by each of the four feature ranking methods and W_i is the weight corresponding to the rank. W_i is varying within the range of [0.5, 1] and this is shown in Fig. 2.

2.3.3. Identification of miRNAs

In this step, the WER of each miRNA is used to rank the miRNAs. Thereafter, the 100 miRNAs with the highest WER for each of the five categories are selected to prepare a Venn diagram. From the Venn diagram, the list of miRNAs that are common to all the five categories and the miRNAs that are present in each of the five categories is obtained. These six sets of miRNAs are further used for quantitative and biological validation.

3. Experimental results

This section describes the experimental setup and the obtained results.

3.1. Experimental testbed

The feature ranking methods have been implemented in Matlab R2017a as they are widely used in various feature ranking applications while other computations have been done using Pandas 0.24 and Numpy 1.14 in Python 3.6.5. An Intel i5 processor with 4 cores and 8 GB RAM has been used for all computational purposes. The results of WER is compared with Conditional Mutual Information Maximisation (CMIM) [10], Double Input Symmetrical Relevance (DISR) [11], Interaction Capping (ICAP) [12], Conditional Informative Feature Extraction (CIFE) [13] and Ensemble Feature Selection (EFS) [15].

3.2. Results

3.2.1. Selected top miRNAs in different categories

The WER technique has identified six sets of miRNAs, one with miRNAs that are present in all categories and the other five with miRNAs for each category. The detailed results of the ranks of all miRNAs are present in supplementary Table S1. Out of the selected 100 miRNAs with the highest WER in each category, 17 are found to be common as depicted in Fig. 3 and 29 are found to be exclusively related with the condition category. Similarly, age, survival status, clinical stage and histological type have 15, 15, 12 and 15 exclusive miRNAs respectively. Moreover, the number of miRNAs obtained using each feature selection method for common and five individual categories is shown in Table 2, while the Venn diagram of the top 100 miRNAs using

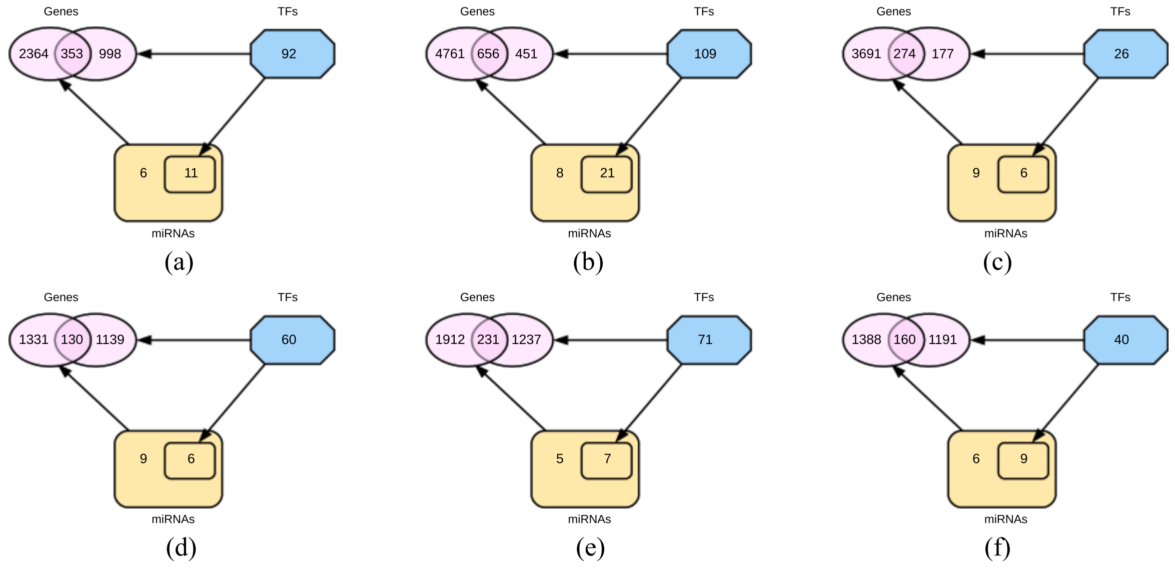


Fig. 4. Network diagrams showing the overlap in the miRNA, Genes and TFs for miRNAs in (a) common, (b) condition, (c) age, (d) survival status, (e) clinical stage and (f) histological type.

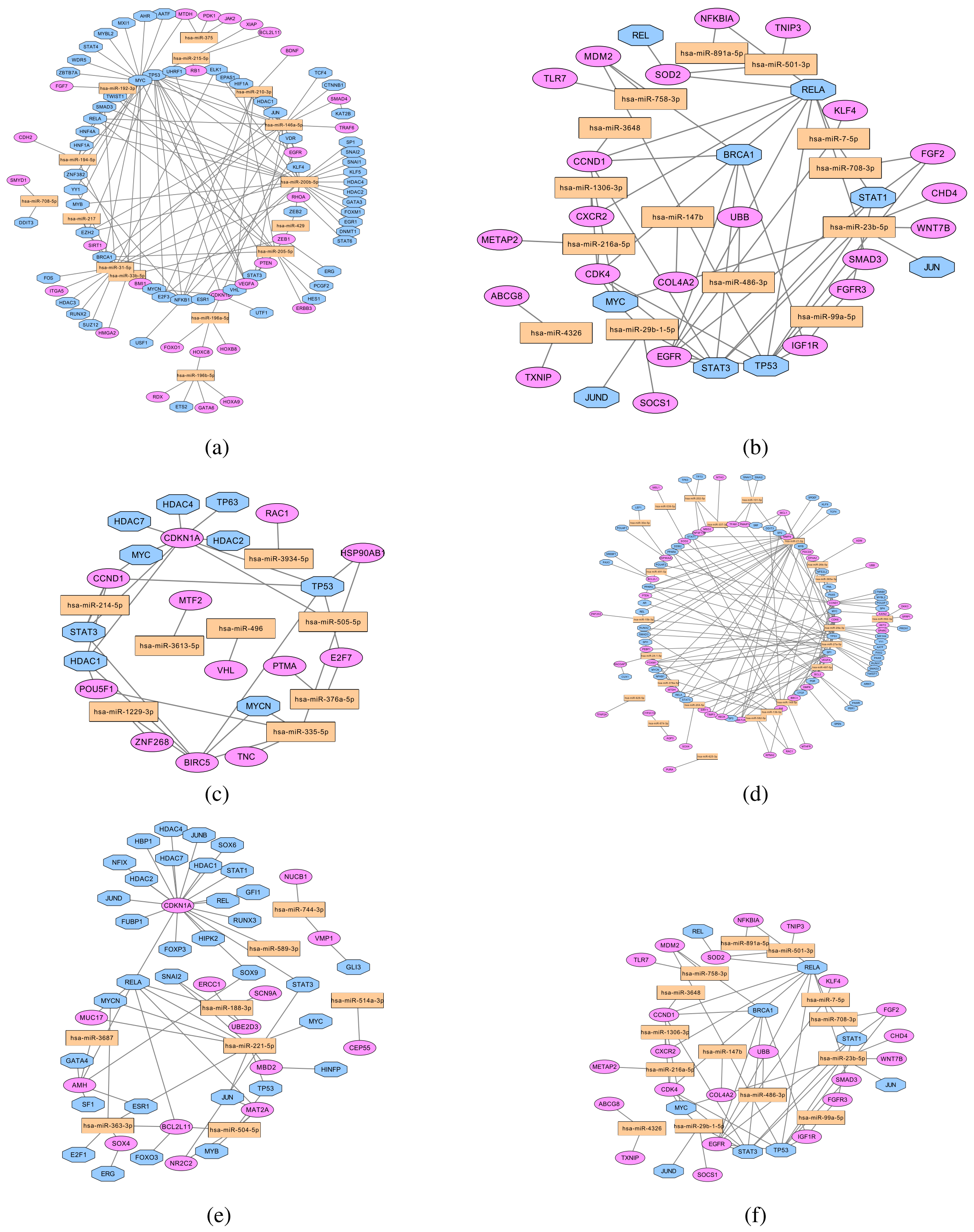
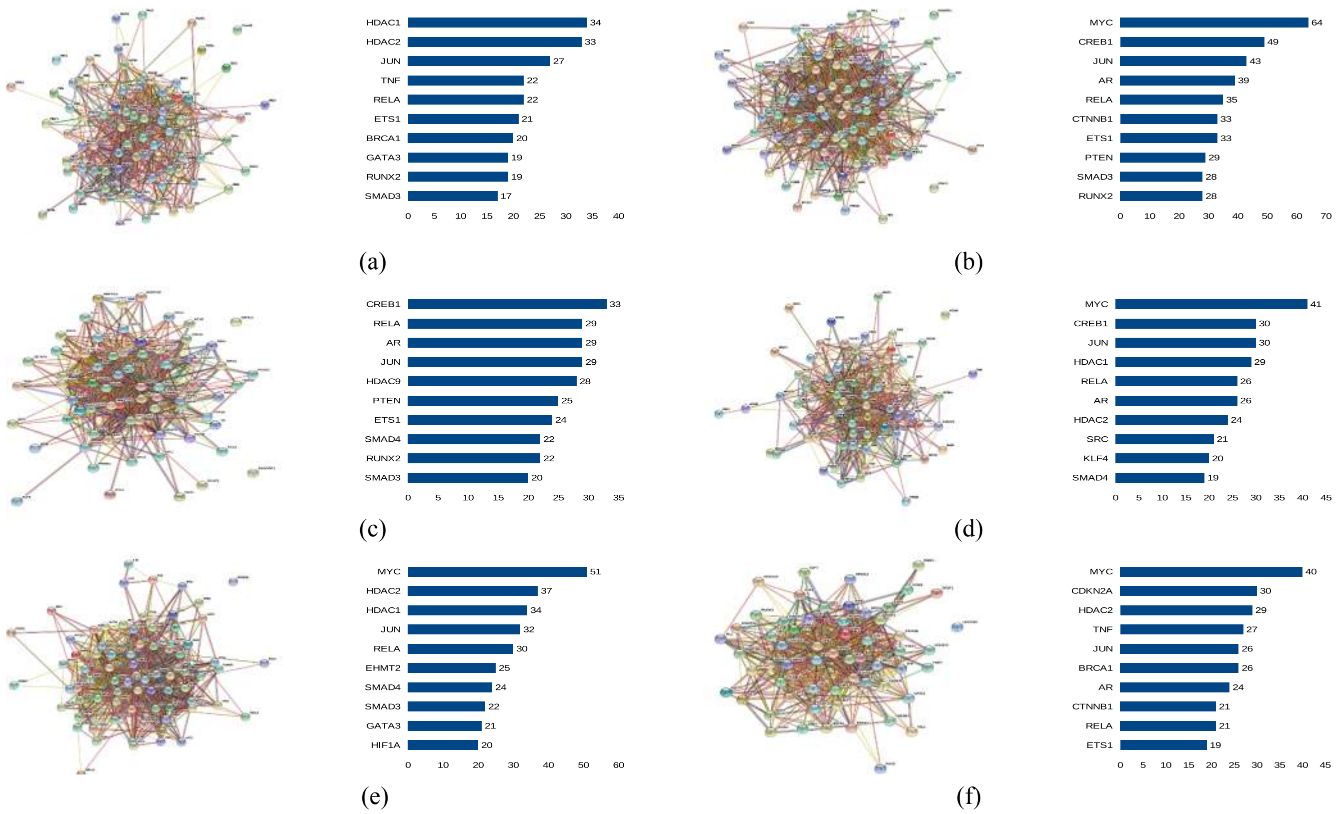


Fig. 5. Network diagrams showing the overlap in the miRNA, Genes and TFs for miRNAs in (a) common, (b) condition, (c) age, (d) survival status, (e) clinical stage and (f) histological type.

other methods individually have been shown in supplementary Fig. S1.

In order to show the importance of the miRNAs in each clinical category of stomach cancer, an extensive literature survey has been conducted and found most of them have been reported in the literature while others can be investigated further in wet lab. A short description about the role of top miRNAs from each category in stomach cancer is given below:

1. **hsa-miR-196b-3p:** This miRNA is found to be significantly upregulated in stomach cancer tissues by quantitative Real-Time Polymerase Chain Reaction (qRT-PCR) analysis. This alteration in expression influences the epithelial-mesenchymal transition in stomach cancer.
2. **hsa-miR-21-5p:** This miRNA has been found to play a role in prediction of recurrence of stomach cancer in the patients. It is found to be overexpressed in the group of patients which have a history of recurrence of stomach cancer. Microarray technique has been used to fetch the altered expression of miR-21-5p in the patients which is further confirmed with qRT-PCR analysis.
3. **hsa-miR-937-3p:** It has been reported that this miRNA inhibits the

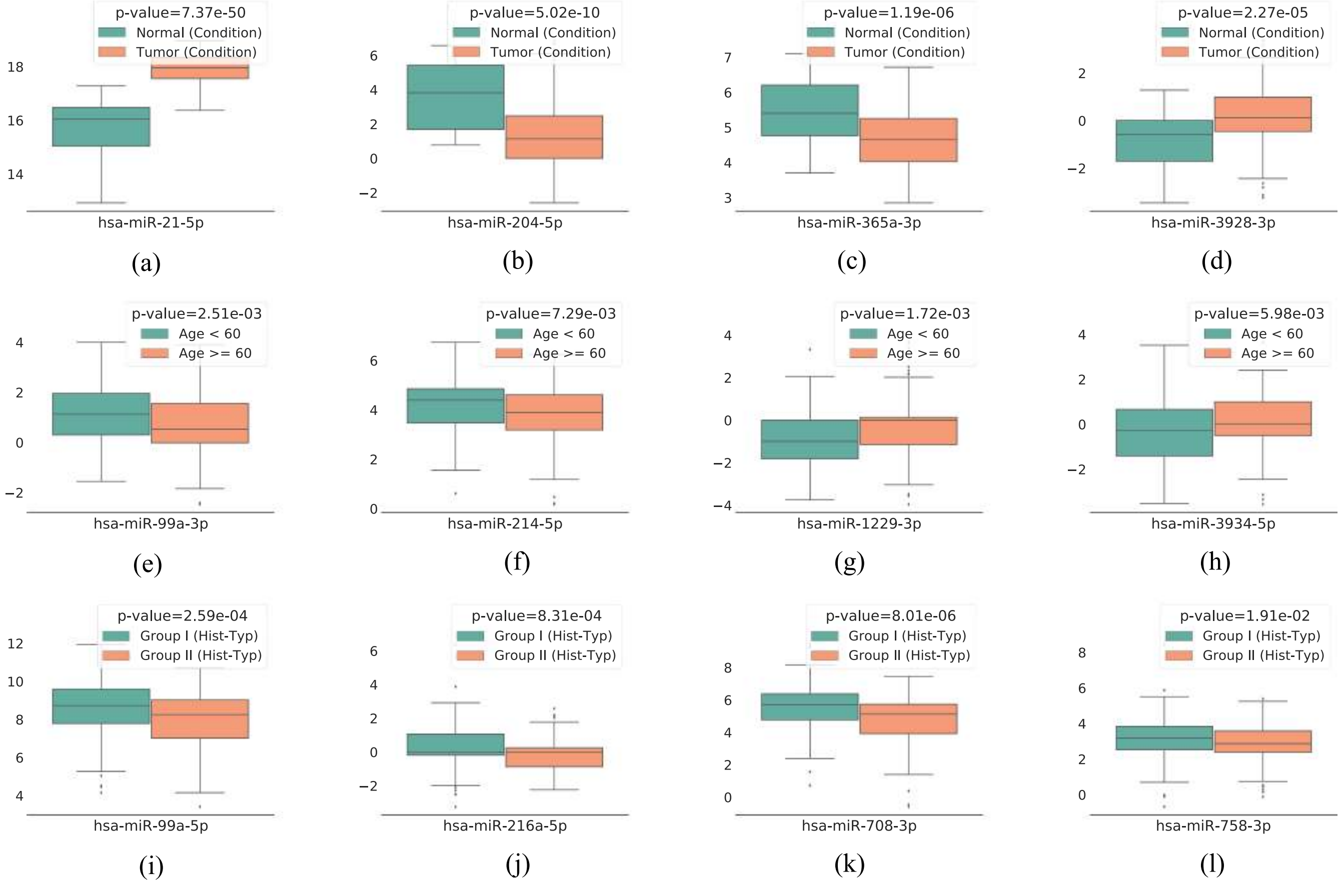


(a) (b)

(c) (d)

(e) (f)

Fig. 6. PPI Networks of the transcription factors for miRNAs in (a) common, (b) condition, (c) age, (d) survival status, (e) clinical stage and (f) histological type.



(a) (b) (c) (d)

(e) (f) (g) (h)

(i) (j) (k) (l)

Fig. 7. Boxplots showing the change in expression values for the top 4 miRNAs based on p-value in (a)-(d): condition, (e)-(h): age and (i)-(l): histological type.

Table 10
Significant KEGG pathway terms for common and exclusive miRNAs in all categories.

ID	Common	Condition	Age	Survival Status	Clinical Stage	Histological Type
hsa04152					✓	
hsa04022					✓	
hsa04010	✓			✓		
hsa04150				✓		✓
hsa05200	✓				✓	
hsa04151	✓				✓	
hsa05205		✓	✓		✓	
hsa04015	✓	✓				
hsa04014				✓	✓	
hsa04071	✓			✓	✓	
hsa04350			✓			✓
hsa04310	✓	✓		✓	✓	✓

ID	Description
hsa04152	AMPK signaling pathway
hsa04022	cGMP-PKG signaling pathway
hsa04010	MAPK signaling pathway
hsa04150	mTOR signaling pathway
hsa05200	Pathways in cancer
hsa04151	PI3K-Akt signaling pathway
hsa05205	Proteoglycans in cancer
hsa04015	Rap1 signaling pathway
hsa04014	Ras signaling pathway
hsa04071	Sphingolipid signaling pathway
hsa04350	TGF-beta signaling pathway
hsa04310	Wnt signaling pathway

Table 11
Significant GO-Biological Process terms for common and exclusive miRNAs in all categories.

ID	Common	Condition	Age	Survival Status	Clinical Stage	Histological Type
GO:0002756	✓	✓	✓	✓	✓	✓
GO:0035666	✓	✓	✓	✓	✓	✓
GO:0035872	✓			✓	✓	
GO:0038095	✓	✓	✓	✓	✓	✓
GO:0038096	✓	✓	✓	✓	✓	✓
GO:0038123	✓	✓	✓	✓	✓	✓
GO:0038124	✓	✓	✓	✓	✓	✓
GO:0048011		✓	✓			
GO:0048015	✓	✓		✓	✓	✓

ID	Description
GO:0002756	Myd88-independent toll-like receptor signaling pathway
GO:0035666	TRIF-dependent toll-like receptor signaling pathway
GO:0035872	Nucleotide-binding domain, leucine rich repeat containing receptor signaling pathway
GO:0038095	Fc-epsilon receptor signaling pathway
GO:0038096	Fc-gamma receptor signaling pathway involved in phagocytosis
GO:0038123	Toll-like receptor TLR1:TLR2 signaling pathway
GO:0038124	Toll-like receptor TLR6:TLR2 signaling pathway
GO:0048011	Neurotrophin TRK receptor signaling pathway
GO:0048015	Phosphatidylinositol-mediated signaling

metastasis and proliferation of stomach cancer by inhibiting PI3K/AKT signalling pathway and is a potential biomarker for its treatment. The results are experimentally validated using qRT-PCR analysis.

4. **hsa-miR-514a-3p:** This miRNA is found to be significantly upregulated in stomach cancer by microarray analysis as well as TaqMan miRNA assay.
5. **hsa-miR-142-5p:** This miRNA is found to be significantly downregulated in stomach cancer patients and is known as a potential predictor of recurrence of the disease by regulating genes involved in Wnt, TP53 and MAPK pathways. This finding has been confirmed using qRT-PCR analysis.
6. **hsa-miR-29b-1-5p:** It has been confirmed using Luciferase assay that miR-29b negatively regulates the gene MMP2 resulting in suppression of cell migration and tumour growth in stomach cancer. It is, therefore, regarded as important therapeutic and diagnostic target of stomach cancer.

All the miRNAs obtained in different categories, along with their average WER are listed in Tables 3–8. Also, the PubMed IDs and experiments used to validate the role of these miRNAs in stomach cancer are also provided in tables.

3.2.2. *Comparison of classification accuracy with different methods*

Next, the identified common set of miRNAs, i.e., 17, 8, 22, 43, 6 and 1 for WER, CMIM, DISR, CIFE, ICAP and EFS respectively are used to perform the classification task for each category using Random Forest. The classification is performed using five fold cross validation and the results are reported in the form of classification accuracy in Table 9. It is observed from the table that the cumulative classification score for WER is 3.590 while the other methods viz. CMIM, DISR, CIFE, ICAP and EFS provides 3.494, 3.562, 3.559, 3.387 and 3.187 respectively for all categories. The results suggest that the selected common miRNAs using WER show better classification accuracy while considering multiple clinical outcomes to judge the condition of patients.

3.2.3. Biological network analysis

In order to perform the biological network analysis for each out of six sets of miRNAs, target genes are fetched using miRTarBase.[4] Thereafter, these gene targets are fed in TRRUST [5] database to find related Transcription Factors (TFs). For these transcription factors, target miRNAs are fetched using the database TransmiR v2.0.[6] It is found that most of these miRNAs are present in our sets of miRNAs. For each set of miRNAs, different number of TFs is found to be associated as depicted in Fig. 4 indicating that these TFs are involved in the regulation of miRNAs which in turn are involved in the regulation of the target genes. These TFs are also directly involved in the regulation of genes. A network of miRNAs, genes and TFs is plotted to visualize the integrated molecular relationship using Cytoscape [16]. In order to construct miRNA-Gene-TF regulatory network for common miRNAs, a small set of genes, i.e. 353, is identified based on the intersection of all targeted genes of common miRNAs, i.e. 2717 genes and the genes that are targeted by common miRNAs and their TFs, i.e. 1460 genes. Thereafter, for such small set of genes, their targeted miRNAs and the TFs are identified in order to construct the miRNA-Gene-TF network. The same procedure is also followed for the other five sets of miRNAs and such regulatory networks are shown in Fig. 5 and in supplementary Fig. S2 for better visualization. The selected miRNAs are found to be associated with a number of important TFs, for example hsa-miR-146a-5p is found to be interacting with STAT3 in Fig. 5 (a) which is over-expressed in stomach cancer and is regarded as a potential target for its treatment [17]. Moreover, the detailed interaction results of miRNA-Gene, TF-Gene and TF-miRNA are provided in supplementary Tables S2 and S3.

The predicted TFs for each set of miRNAs are further used to prepare the protein–protein interaction (PPI) network using STRING[7] database. Fig. 6 represents the PPI networks along with a bar plot representing the degree of connections for top 10 TFs. It is seen that the nodes are densely connected implying high functional relatedness with a PPI enrichment value less than 10^{-16} for all the six networks. Many TFs in the networks are known to be involved in stomach cancer, like MYC, which can be regarded as biomarker for stomach cancer prognosis as well as clinical stratification [18]. Other important nodes identified by the network like CREB1 [19] and ETS1 [20] are also known to play significant role in stomach cancer progression. The detailed diagram of the PPI networks and the degree of all nodes are present in supplementary Figs. S3 and S4 for better visualization.

3.2.4. Expression analysis of miRNAs

Box plots are used to represent the variable expression of the miRNAs in two different groups for each clinical category. Two-sample t-test has been performed with two groups of each category for selected six different sets of miRNAs. The low p-value in the box plots in Fig. 7 denotes that the miRNAs exclusive to each category obtained by WER have efficiently separated the samples into two groups with significant difference. All the box plots of exclusive miRNAs for all categories are provided in supplementary Fig. S5.

3.2.5. Enrichment analysis of miRNAs

To further validate each set of miRNAs, KEGG pathway and GO enrichment analysis has been performed. For identifying the potential pathways, DIANA-miRPath v3.0 [21] has been used. The KEGG pathway terms obtained in different categories are listed in Table 10. It is seen that significant pathway terms related to stomach cancer are present in the results. One of such terms is Wnt Signalling pathway, which can be triggered by *Helicobacter pylori* infection which is one of the major causes of stomach cancer [22]. The abberant activation of this pathway leads to the development of tumorigenic stem cell-like subpopulation [23]. Other important pathways for stomach cancer like PI3K-Akt signaling pathway, Rap1 signaling pathway, cGMP-PKG signaling pathway, etc are also reported in the table. Such sets of miRNAs are also subjected to Gene Ontology enrichment analysis using Enrichr [24]. The selected miRNAs are found to be associated with biological processes that are specific for stomach cancer, listed in Table 11, for example, toll-like receptor TLR1:TLR2 signaling pathway and MyD88-independent toll-like receptor signaling pathway which are activated as a result of *H. pylori* infection and play a crucial role in innate immunity system at the time of infection [25]. Detailed information about all enriched pathways is provided in supplementary Tables S5 and S6. Also, the annotation ratio and GO Enrichment terms for cellular component and molecular function are present in supplementary Tables S7 and S8.

4. Conclusion

Despite several advancements in medical and healthcare sciences, the problem of early detection of stomach cancer prevails. NGS technology has provided enormous information regarding this disease but for drawing better inferences from the available data, one needs to have a set of important features to consider. Keeping this in mind, a multi-view method for selecting important set of miRNAs from miRNA expression data has been proposed in this study. The method used here considers five different clinical categories of the patients, to identify important miRNAs common in all categories and specific to each category. Four well-known feature ranking methods have been utilized to give a unique weighted ensemble rank to the miRNAs. The obtained sets of miRNAs have been analyzed using several methods and the results prove their importance in the molecular epidemiology of the disease.

These sets of miRNAs can be further analyzed for fetching their associations with stomach cancer using several prediction models [26,27]. Identification of such potential miRNA biomarkers and predicting their association with diseases is emerging as an important research area due to its cost effectiveness [27], as well as its wide application in diagnosis, prognosis and treatment of the disease [28].

Author Contributions

NP, SR, SP and IS have conceived and designed the experiments. SR and IS have performed the experiments. NP, SR, SP and IS have scripted the manuscript. NP, SR, SP and IS have corrected and edited the manuscript. All authors read and approved the final manuscript.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

This work was partially supported by the grant from Department of

[4] https://bio.tools/mirtarbase.

[5] https://www.grnpedia.org/trrust/.

[6] http://www.cuilab.cn/transmir.

[7] https://string-db.org/.

Science and Technology, India (Grant No. DST/INT/POL/P-36/2016) to IS and the seed grant program of Indian Institute of Technology, Jodhpur, India (Grant No. I/SEED/SPU/20160010) to SP while NP was supported by a MHRD fellowship.

Appendix A. Supplementary material

The code, data and supplementary materials are available online at http://www.nitttrkol.ac.in/indrajit/projects/mirna-stomachcancer/. Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.jbi.2019.103254.

References

[1] C. Servarayan Murugesan, K. Manickavasagam, A. Chandramohan, Gastric cancer in India: Epidemiology and Standard of Treatment, Updates Surg. 70 (2018) 233–239.

[2] E. Mardis, The impact of next-generation sequencing technology on genetics, Trends Genet. 24 (2008) 133–141.

[3] H.S. Liu, H.S. Xiao, MicroRNAs as potential biomarkers for gastric cancer, World J. Gastroenterol. 20 (2014).

[4] Y. Liu, J. Luo, P. Ding, Inferring microRNA Targets Based on Restricted Boltzmann Machines, IEEE J. Biomed. Health Informat. 23 (2019) 427–436.

[5] X. Chen, D. Xie, Q. Zhao, Z.-H. You, MicroRNAs and complex diseases: from experimental results to computational models, Briefings Bioinformat. 20 (2019) 515–539.

[6] H. Liu, S. Zhou, J. Guan, Identifying mammalian microRNA Targets Based on Supervised Distance Metric Learning, IEEE J. Biomed. Health Informat. 17 (2013) 427–435.

[7] Y. Zhang, D.-H. Guan, R.-X. Bi, J. Xie, C.-H. Yang, Y.-H. Jiang, Prognostic value of microRNAs in gastric cancer: a meta-analysis, Oncotarget 33 (2017).

[8] J. Ang, A. Mirzal, H. Haron, H. Hamed, Supervised, unsupervised, and semi-supervised feature selection: A review on gene selection, IEEE/ACM Trans. Comput. Biol. Bioinf. 13 (2016) 971–989.

[9] X. Zhu, Y. Wang, Y. Li, Y. Tan, G. Wang, Q. Song, A new unsupervised feature selection algorithm using similarity-based feature clustering, Computat. Intell. (2018).

[10] F. Fleuret, Fast binary feature selection with conditional mutual information, J. Machine Learn. Res. 5 (2004) 1531–1555.

[11] P. Meyer, G. Bontempi, On the use of variable complementarity for feature selection in cancer classification, Proc. Appl. Evol. Comput. 3907 (2006) 91–102.

[12] A. Jakulin, Machine learning based on attribute interactions, Fakulteta za racunalništvo in informatiko, Univerza v Ljubljani (2005).

[13] D. Lin, X. Tang, Conditional infomax learning: an integrated framework for feature extraction and fusion, Proceedings of European Conference on Computer Vision, vol. 3951, 2006, pp. 68–82.

[14] T.C.G.A.R. Network, The Cancer Genome Atlas Pan-Cancer analysis project, Nat. Genet. 45 (2018).

[15] U. Neumann, N. Genze, D. Heider, Efs: an ensemble feature selection tool implemented as r-package and web-application, BioData Min. 10 (2017).

[16] P. Shannon, A. Markiel, O. Ozier, N. Baliga, J. Wang, D. Ramage, N. Amin, B. Schwikowski, T. Ideker, Cytoscape: a software environment for integrated models of biomolecular interaction networks, Genome Res. 11 (2003) 2498–2504.

[17] M. Hajimoradi, Z. Mohammad Hassan, M. Ebrahimi, M. Soleimani, M. Bakhshi, J. Firouzi, F.S. Samani, STAT3 is overactivated in gastric cancer stem-like cells, Cell J. 17 (2016) 617–628.

[18] C. Souza, M. Leal, D. Calcagno, S.E. Costa, B. Borges, R. Montenegro, A. Dos Santos, S. Dos Santos, H. Ribeiro, P. Assumpcao, M. Arruda Cardoso Smith, R. Burbano, MYC deregulation in gastric cancer and its clinicopathological implications, PLoS One 8 (2013).

[19] M. Rao, Y. Zhu, X. Cong, Q. Li, Knockdown of CREB1 inhibits tumor growth of human gastric cancer in vitro and in vivo, Oncol. Rep. 37 (2017).

[20] Y. Yu, Y. Zhang, W. Zhang, L. Shen, P. Hertzog, T. Wilson, D. Xu, Ets1 as a marker of malignant potential in gastric carcinoma, World J. Gastroenterol 9 (2003).

[21] I.S. Vlachos, K. Zagganas, M.D. Paraskevopoulou, G. Georgakilas, D. Karagkouni, T. Vergoulis, T. Dalamagas, A.G. Hatzigeorgiou, DIANA-miRPath v3. 0: deciphering microRNA function with experimental support, Nucleic Acids Res. 43 (2015).

[22] D.B. Polk, R.M. Peek Jr, Helicobacter pylori: gastric cancer and beyond, Nat. Rev. Cancer 10 (2010) 233–239.

[23] N. Takebe, P.J. Harris, R.Q. Warren, S.P. Ivy, Targeting cancer stem cells by inhibiting Wnt, Notch, and Hedgehog pathways, Nature Reviews, Clin. Oncol. 8 (2011) 97.

[24] W. Yan, S. Wang, Z. Sun, Y. Lin, S. Sun, J. Chen, W. Chen, Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool, BMC Bioinformat. 14 (2013) 128.

[25] K. Echizen, O. Hirose, Y. Maeda, M. Oshima, Inflammation in gastric cancer: Interplay of the COX-2/prostaglandin E2 and Toll-like receptor/MyD88 pathways, Cancer Sci. 107 (2018) 391–397.

[26] X. Chen, L. Wang, J. Qu, N.-N. Guan, J.-Q. Li, Predicting miRNA-disease association based on inductive matrix completion, Bioinformatics 34 (2018) 4256–4265.

[27] X. Chen, D. Xie, L. Wang, Q. Zhao, Z.-H. You, H. Liu, BNPMDA: bipartite network projection for MiRNA-disease association prediction, Bioinformatics 34 (2018) 3178–3186.

[28] K. Sawaki, M. Kanda, Y. Kodera, Review of recent efforts to discover biomarkers for early detection, monitoring, prognosis, and prediction of treatment responses of patients with gastric cancer, Expert Rev. Gastroenterol. Hepatol. 12 (2018) 657–670.