


RESEARCH ARTICLE

Open Access



Automatic discovery of 100-miRNA signature for cancer classification using ensemble feature selection

Alejandro Lopez-Rincon^{1*} , Marlet Martinez-Archundia², Gustavo U. Martinez-Ruiz³, Alexander Schoenhuth⁴ and Alberto Tonda⁵

Abstract

Background: MicroRNAs (miRNAs) are noncoding RNA molecules heavily involved in human tumors, in which few of them circulating the human body. Finding a tumor-associated signature of miRNA, that is, the minimum miRNA entities to be measured for discriminating both different types of cancer and normal tissues, is of utmost importance. Feature selection techniques applied in machine learning can help however they often provide naive or biased results.

Results: An ensemble feature selection strategy for miRNA signatures is proposed. miRNAs are chosen based on consensus on feature relevance from high-accuracy classifiers of different typologies. This methodology aims to identify signatures that are considerably more robust and reliable when used in clinically relevant prediction tasks. Using the proposed method, a 100-miRNA signature is identified in a dataset of 8023 samples, extracted from TCGA. When running eight-state-of-the-art classifiers along with the 100-miRNA signature against the original 1046 features, it could be detected that global accuracy differs only by 1.4%. Importantly, this 100-miRNA signature is sufficient to distinguish between tumor and normal tissues. The approach is then compared against other feature selection methods, such as UFS, RFE, EN, LASSO, Genetic Algorithms, and EFS-CLA. The proposed approach provides better accuracy when tested on a 10-fold cross-validation with different classifiers and it is applied to several GEO datasets across different platforms with some classifiers showing more than 90% classification accuracy, which proves its cross-platform applicability.

Conclusions: The 100-miRNA signature is sufficiently stable to provide almost the same classification accuracy as the complete TCGA dataset, and it is further validated on several GEO datasets, across different types of cancer and platforms. Furthermore, a bibliographic analysis confirms that 77 out of the 100 miRNAs in the signature appear in lists of circulating miRNAs used in cancer studies, in stem-loop or mature-sequence form. The remaining 23 miRNAs offer potentially promising avenues for future research.

Keywords: MicroRNAs, miRNA, Feature selection, Machine learning, Classifiers, Dataset

Background

Cancer is difficult to diagnose and classify at early stages, and is one of the top leading causes of death worldwide [1]. Therefore, several attempts have been made to identify possible biomarkers for cancer detection. MicroRNAs (miRNAs) represent a class of small noncoding RNA

molecules, with a critical role in the post-transcriptional regulation of gene expression. miRNAs also act on several cellular processes, such as cell differentiation, cell cycle progression, and apoptosis. Additionally, in tumors, some miRNAs can function as oncogenes, while others suppress tumors [2]. Succeeding the earliest evidence of miRNA involvement in human cancer by Croce et al. [3], various studies have demonstrated that miRNA expressions are deregulated in human cancer through a variety of mechanisms [4]. Since ectopic modulation of specific miRNAs compromise the hallmarks of cancer, several efforts have

*Correspondence: alejandro.lopez@iscpi.fr

¹Division of Pharmacology, Utrecht Institute for Pharmaceutical Sciences, Faculty of Science, Utrecht University, David de Wied building, Universiteitsweg 99, 3584 CG Utrecht, The Netherlands
Full list of author information is available at the end of the article



been spent to generate scaffold-mediated miRNA-based delivery systems trying to demonstrate the potential of miRNA-mediated therapies.

In comparison to invasive methods currently used for cancer diagnosis, there is an ongoing debate on the use of circulating miRNAs as possible biomarkers due to the fact that they can be detected directly from biological fluids, such as blood, urine, saliva and pleural fluid [5]. MiRNAs possess other qualities of good candidate biomarkers such as: a) they are useful for the identification of cancer types, b) their availability of high-quality measurement techniques for miRNAs and c) they present good conservation between practical and preclinical models [6].

Several studies have shown the properties of miRNAs as oncogenes and tumor suppressors genes [7–9]. Since then, techniques such as microarray (Affymetrix, Agilent) and sequencing techniques (Illumina), have been proposed for their identification [10]. In the context of increasing availability of data, it is of utmost practical importance to build databases of miRNA expressions data for cancer research [11–13] and to extract features that could be used as cancer biomarkers [14–16]. For example, the expression levels of miRNA *hsa-miR-21* change for different cancer types such as: squamous cell lung carcinoma [17], astrocytoma [18], breast cancer [19], and gastric cancer [20]. Following this idea, the scientific community is currently looking for miRNA signatures (a subset of miRNAs), representing the minimal number of miRNAs to be measured for discriminating between different stages and types of cancer.

Thousands of miRNAs have been identified, and currently miRBase (v22.1) contains 1917 stem-loop sequences, and 2657 mature sequences for human microRNA [13]. Although a classification of cancer tumor type is possible using isomirs [21], not all of the miRNAs listed are available in every study, and only a few of them have been shown to work as circulating biomarkers [6]. Obtaining a minimal list of miRNAs able to correctly classify tumors is of utmost practical importance, because it would reduce the measurements needed and improve the likelihood of validation across multiple studies.

Several approaches in the literature propose the use of machine learning techniques for feature selection involving miRNAs. For example, feature selection for identifying miRNA targets [22], for prediction of specific biomarkers for tumor origin [23] and to learn subset of features for tumor classification [24]. In this study, the objective was to use feature selection and to uncover a small miRNAs signature with the aim to correctly classify cancer tumor types, and distinguish between normal and tumor tissue reducing the necessary features by an order of magnitude.

We propose an ensemble feature selection method, starting from a subset of The Cancer Genome Atlas dataset (TCGA) [25], containing 8023 cases, with 28

different types of cancer, and 1046 different stem-loop miRNA expressions (miRBase V16¹, summarized in Table 10). Typically, classifiers trained on a dataset do not use the whole set of available features to separate classes, but only a subset which could be ordered by relative importance, with a different meaning given to the list by the specific technique, pushing for simpler models. Using 8 state-of-the-art classifiers implemented in the `scikit-learn` toolbox [26], the most relevant miRNAs are extracted to be used as features for cancer classification. The top k features in the list are then evaluated as a potential reduced signature for classification. In this work, after preliminary tests, we select $k = 100$ to reduce the original features by an order of magnitude. Because other feature selection methods require the user to specify a desired number of features, this also allows for a fair and meaningful comparison with these methods.

The obtained 100-miRNA signature is first tested to classify the initial TCGA dataset, and later applied on 14 Gene Expression Omnibus (GEO) datasets obtained with different platforms (Affymetrix Multispecies Array miRNA-1, miRNA-2 and miRNA-3, Illumina 2000, and Agilent-021827 Human miRNA Microarray V3), for different cancer tumor types (Prostate, Liver, Breast, Esophageal, Head and Neck Squamous and Lung). A summary of this validation is presented in Fig. 1. Furthermore, the proposed methodology is compared to popular feature selection methods in bioinformatics, such as Univariate Feature Selection, Recursive Feature Elimination, Genetic Algorithms, Least Absolute Shrinkage and Selection Operator, Random Selection, Elastic Net and Ensemble Feature Selection with Complete Linear Aggregation. Next, we use the same signature to try to distinguish molecular subtypes in breast cancer, both for the TCGA dataset and a set of GEO datasets. Finally, the 100 miRNAs included in the signature are evaluated through a meta-analysis based on the medical literature. Because this meta-analysis reveals known relationships between features selected by our approach, relative to the type of cancer considered, it has the potential to yield insight into the biological processes and relationships combinedly affecting miRNAs and cancer.

Results

Feature selection and validation on the TCGA dataset

Table 1 compares the classification accuracy on a 10-fold cross-validation for each classifier, using the full 1046 features, and then employing the reduced 100-miRNA signature. It is interesting to notice how the accuracy is, for most cases, unchanged, providing empirical evidence that a 100-miRNA signature is enough to obtain good classification results, with a small statistically significant (T-test, $p < 0.05$) difference of 1.4%.

¹<http://mirbase.org/pub/mirbase/16/>

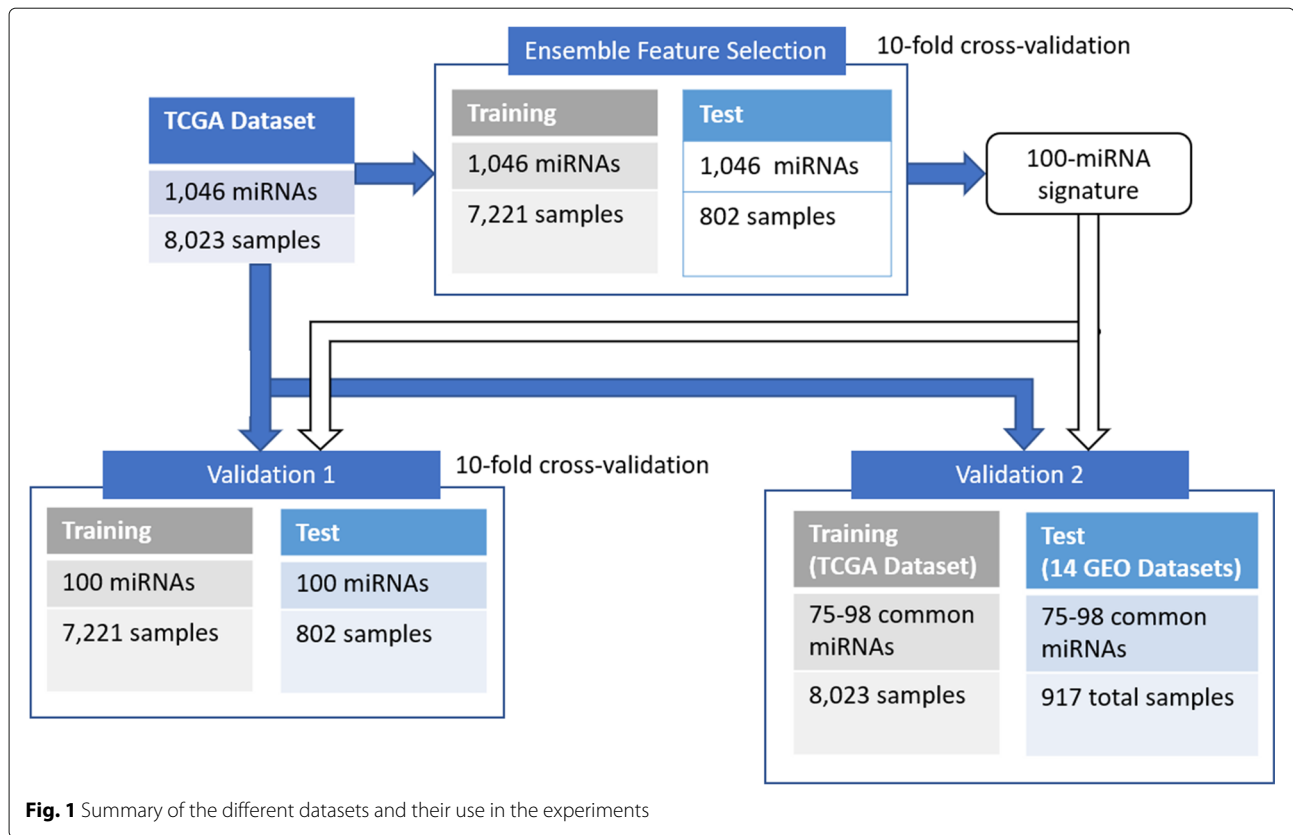


Figure 2 shows a heatmap comparing the relative frequency of the overall top 100 most frequent miRNA features, for each considered classifier. As expected, not all classifiers used the same features to separate the types of cancer, and thus, evaluating their consensus is more robust than just relying upon a single algorithm, as it is commonly accepted in the field of machine learning [27]. It is interesting to notice that while the most common

biomarkers appear among the top for most classifier, others make use of only a few. For example, Bagging and Ridge do not use the vast majority of the features exploited by other techniques to discriminate between classes. A further difference between the two classifiers is that features used by Bagging that also appear in the top 100 are clearly important for the classifier, being used in almost 100% of its 10 runs; while it is noticeable how

Table 1 Accuracy of classifiers used in the experiments on the TCGA dataset

Classifier	Accuracy (10-fold CV)				Hyper parameters	Feature selection method
	1046 Features		100 Features			
	avg	std	avg	std		
Gradient Boosting	0.9398	0.0076	0.9359	0.0086	300 predictors	Decision Trees
Random Forest	0.9351	0.0071	0.9324	0.0073	300 predictors	Decision Trees
Logistic Regression	0.9178	0.0096	0.9237	0.0067	-	Coefficients
Passive Aggressive	0.9117	0.0104	0.8831	0.0115	-	Coefficients
SGD	0.91	0.0074	0.9035	0.0152	-	Coefficients
SVC	0.9211	0.0122	0.9154	0.0065	Linear kernel	Coefficients
Ridge	0.8971	0.0138	0.8305	0.0062	-	Coefficients
Bagging	0.9151	0.0120	0.9110	0.0077	300 predictors	Decision Trees
Average	0.918463	-	0.9044	-	-	-

In the case a classifier is not using standard values for its hyperparameters, the relevant variations are summarized in the corresponding column

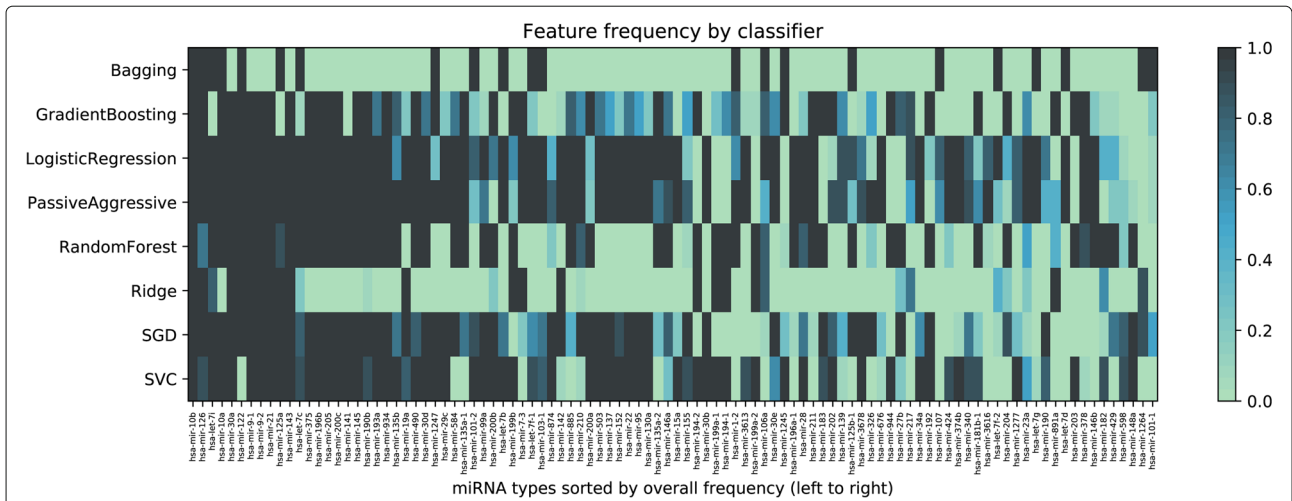


Fig. 2 Heatmap with the frequency of the overall top 100 most frequent features, divided by classifier. Features are sorted from overall most to least frequent, from left to right, using information from the whole ensemble. For example, the most frequent is mir-10b, that is considered important by all classifiers. Color intensity is computed using information from instances of the same classifier, only. This shows the different importance that different classifiers assign to each feature

Ridge probably bases its discrimination on features that do not appear among the top 100. This would also explain why Ridge is the only algorithm that presents a decrease in performance when using the 100-miRNA signature. It's important to note that, while the results emerging from the heatmap suggest that this is indeed the case, Ridge's

decision boundaries should be analyzed more in-depth, for each class and multiple instances, in order to have absolute certainty, a task that is outside of the scope of the current work. Figure 3 shows the difference between 1046 features and 100 features for each cancer type and classifier.

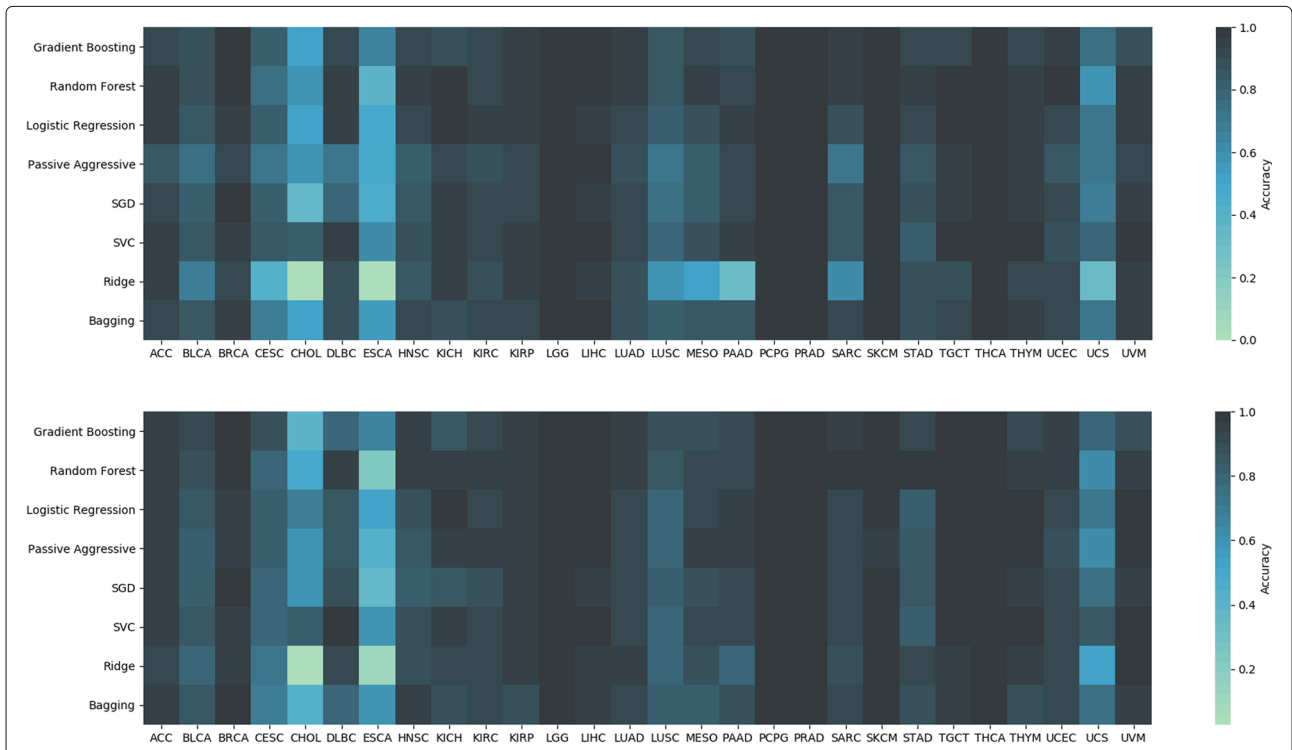


Fig. 3 Heatmap of the accuracy by cancer type, by classifier using the 1046 features (top) and the 100-miRNA signature (bottom)

Normal vs tumor tissue classification

We compared Tumor Tissue (TT) vs Normal Tissue (NT) in a 10-cross fold validation, using stratified cross-validation to maintain the proportions for the two classes inside the folds. The global score and the classification accuracy by class are reported in Table 2. All of the classifiers have fair quality for differentiating between normal tissue and tumor tissue, except Ridge, which is more sensitive to the unbalanced number of examples.

Comparison to established feature selection methods

Several feature selection techniques have been proposed for microarray data [28]. The most effective approaches include Univariate Feature Selection (UFS), Recursive Feature Elimination (RFE), Elastic Net (EN), Genetic Algorithms (GALGO), Least Absolute Shrinkage and Selection Operator (LASSO) and Ensemble Feature Selection with Complete Linear Aggregation (EFS-CLA). UFS aims at finding the best features, scoring them using univariate statistical tests, such as the ANOVA F-value [29], and ultimately taking the k features with the highest scores. RFE runs several times a machine learning algorithm capable of scoring features, such as SVC, iteratively removing the feature with the lowest score [30] until it reaches the user-specified k features. EN simply runs the machine learning algorithm Elastic Net [31], and takes the k highest-scored features. As Elastic Net is trying to balance accuracy and weight size in a linear model, exploiting L1 and L2 regularization, it is a popular choice for feature selection in bio-informatics [32, 33], because it tends to create sparse models with few weights different from zero. LASSO is a regression analysis method, performing variable selection and regularization to improve prediction accuracy and interpretability of the statistical model it produces [34], so it can be easily used for feature selection, only. All considered feature selection methods are implemented in the machine learning package `scikit-learn`, already used in the previous experiments. GALGO is a genetic algorithms-based feature selection library in R that ranks the features using

several calls to a classifier and choosing the features that appear the most after evolving a subset several times [35]. EFS-CLA is a method that uses instances of SVM with several calls to a subsample of the data, ranks the features by weight value and reduces a percentage at each iteration [36].

As some of these techniques require the user to specify the number of features k to be taken, to provide a comparison with the approach presented in this paper, we have selected $k = 100$ features using all the formerly described feature selection methods and compared classification accuracy on the considered classifiers with a 10-fold cross validation. For RFE, we have decided to use SVC, as not only it is commonly adopted for feature selection in bioinformatics [30, 37], but also represents a good compromise between accuracy and speed of convergence on our specific dataset. For EN, we have chosen the `ElasticNetCV` scikit-learn method, which exploits a 3-fold cross-validation to automatically adapt the internal parameter α , balancing the importance of L1 and L2 regularization in the model. For the same reasons, the `LassoCV` scikit-learn method is selected for LASSO. For EFS-CLA, we use percentage of reduction $E = 20\%$, 40 as SVM calls per step, and $k=100$. Finally, we add a random selection of 100 features, as a baseline reference to portray the efficiency of the feature selection algorithms.

From the results presented in Table 3, it is immediately clear that the 100 features selected by UFS are much less informative than the ones found by the proposed approach. RFE performs better, especially when considering SVC as the classifier used for the cross validation, but overall the performance for the other classifiers is lower. It must also be noted that, among all the methods, RFE is the most computationally expensive, as it calls the considered classifier, SVC in this case, $N - k = 1,046 - 100 = 946$ times, where N is the original number of features. All feature selection algorithms, as expected, perform much better than the baseline random selection of features.

A qualitative analysis of the features selected by each method shows that the highest-scoring ones are easily

Table 2 Accuracy for each classifier in a 10-fold cross-validation for the comparison between Tumor Tissue (TT) and Normal Tissue (NT) for 1046 and 100 features

Classifier	100-NT	100-TT	1046-NT	1046-TT	100-Global	1046-Global
Gradient Boosting	0.8612	0.9944	0.8707	0.9950	0.9846	0.9859
Random Forest	0.8091	0.9978	0.7256	0.9985	0.9839	0.9785
Logistic Regression	0.8423	0.9908	0.8659	0.9764	0.9799	0.9683
Passive Aggressive	0.7177	0.9798	0.8123	0.9728	0.9606	0.9611
SGD	0.8060	0.9902	0.7445	0.9936	0.9767	0.9754
SVC(linear)	0.8517	0.9892	0.8218	0.9771	0.9791	0.9657
Ridge	0.2997	0.9981	0.5994	0.9923	0.9470	0.9635
Bagging	0.8028	0.9953	0.7792	0.9966	0.9812	0.9807

Table 3 Comparison between different feature selection techniques and the proposed ensemble method for $k = 100$, on the TCGA dataset

Classifier	Random	GALGO	EFS-CLA	UFS	EN	LASSO	RFE	EFS
Gradient Boosting	0.8588	0.8782	0.8871	0.9028	0.9208	0.9315	0.9309	0.9359
Random Forest	0.8515	0.8787	0.8824	0.8929	0.9224	0.9341	0.9288	0.9324
Logistic Regression	0.8015	0.8295	0.8832	0.8813	0.8988	0.8996	0.9088	0.9237
Passive Aggressive	0.6986	0.7235	0.8111	0.8091	0.8406	0.8424	0.8506	0.8831
SGD	0.7278	0.764	0.8446	0.8334	0.8649	0.8648	0.8824	0.9035
SVC	0.8077	0.8348	0.8706	0.885	0.9049	0.9008	0.9103	0.9154
Ridge	0.6534	0.6614	0.7422	0.7504	0.7753	0.7751	0.7954	0.8305
Bagging	0.822	0.8382	0.8562	0.8719	0.8889	0.9078	0.9061	0.911
Global Average	0.7777	0.8010	0.8472	0.8534	0.8771	0.8820	0.8892	0.9044
Calls to Classifier	-	60,000	480	-	-	10	946	80

found by all considered approaches. In particular, from the 100 features found by our approach, 8 are in common with Random, 11 with GALGO, 29 with EFS-CLA, 38 are common to the group obtained through UFS, 44 are shared with the group found by LASSO, 48 again are found by EN, and 54 are in common with RFE.

Cross-Platform validation on gEO datasets

As different datasets present distinctive sets of miRNAs, it is important to assess the performance of the signature we identified on unseen data. Using the methodology previously described, the proposed approach is validated on the 14 GEO datasets. Each run of a classifier on a dataset was repeated 10 times, to compensate possible random elements that appear during the training phase of specific algorithms, e.g. RandomForest. It is worth noticing how this validation presents considerable challenges. As

we are dealing with different platforms, not all of the 100 features in the signature were available everywhere. For most GEO datasets 98 were available, while for GSE62182 featured 75 of them. Furthermore, despite the transformation needed to bring the samples of the GEO datasets in the TCGA dataset space, samples measured by platforms used in the GEO datasets might prove particularly difficult to tackle for classifiers trained on TCGA samples, as most GEO datasets use microarray technology while TCGA uses sequencing. The properties of the used GEO datasets are summarized in Table 4.

Figure 4 shows the outcomes of the validation for all classifiers. In spite of the difficulties, most algorithms yielded good classification results, with Logistic and SGD in particular featuring over 93% average accuracy on all GEO datasets. Several classifiers, on the other hand, show poor performance on specific datasets, probably

Table 4 Summary of the used GEO datasets, and the number of features in common with our 100-miRNA signature

Dataset ID	Platform	Tumor Type	#Samples	Total Feats.	Common Feats.	Reference
GSE34496	GPL8786	HNSC	44	847	98	-
GSE36802	GPL8786	PRAD	21	847	98	[38]
GSE67138	GPL8786	LIHC	57	847	98	-
GSE67139	GPL8786	LIHC	115	847	98	-
GSE45604	GPL14613	PRAD	50	2143	98	[39]
GSE48088	GPL14613	BRCA	33	2143	98	[40]
GSE55856	GPL14613	ESCA	108	2143	98	[41]
GSE86277	GPL14613	BRCA	72	2143	98	[42]
GSE116182	GPL14613	LIHC	64	2143	98	-
GSE86278	GPL16384	BRCA	49	3,242	98	[42]
GSE86281	GPL16384	BRCA	50	3,242	98	[42]
GSE31164	GPL10850	LIHC	110	851	98	[43]
GSE105134	GPL10850	BRCA	50	851	98	-
GSE62182	GPL11154	LUAD	94	3,242	75	[44]

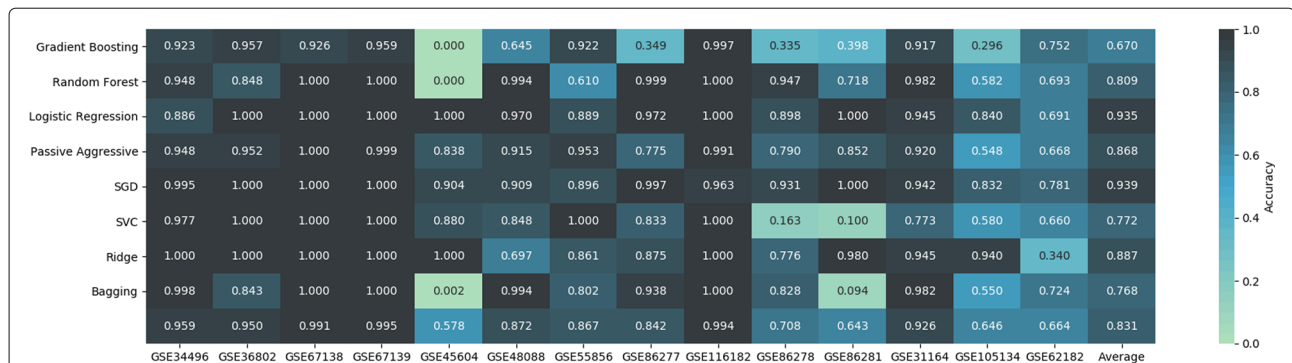


Fig. 4 Results with the 100 selected features in the GEO datasets, using a 10-fold cross-validation. From the average accuracy and standard deviation, SGD proves to be significantly better than the rest using a Kolmogorov-Smirnov test ($p < 0.05$)

due to the way their decision boundaries for that specific class were learned on the TCGA dataset. In this sense, dataset GSE45604 proves to be the overall hardest to classify correctly for most algorithms. GSE86277, GSE86278 and GSE86281, deal with different molecular subtypes of BRCA, that could explain some of the performance issues. Finally the average performance in GSE62182, is because the classifiers have problems differentiating LUAD and LUSC. In general, however, different algorithms seem to have difficulties for different classes and datasets, which suggests that an ensemble approach for classification could compensate local issues.

To the best of our knowledge, the most similar work in literature that we can compare our results to is Telonis et al. [21], where isoform quantification was adopted to classify three of the GEO datasets used in this study (GSE36802, GSE67138, GSE67139), training SVC on a TCGA-derived dataset. For GSE36802, [21] reports an accuracy of 76%, that is surpassed by all of the classifiers. Considering GSE67138, for which an accuracy of 91% is reported, all the algorithms in our case perform better. Finally, for GSE67139, a 96% accuracy, again all the algorithms outperform that value. It must be noted, however, that even this comparison is made difficult by differences in how data was treated: for example, [21] reduced the number of classes to 6 and tested on 4 different types of tumors. In our study, we keep all 28 classes for testing.

Tumor subtype

To further test our approach, we use the 100-miRNA signature to classify tumor subtypes. As a comparison with GEO datasets is important for our validation, we select molecular subtype in breast cancer (BRCA), as it's the only tumor class for which molecular subtype information is available in the GEO datasets. From the information in [45, 46], we are able to label 764 of the 777 BRCA samples in the TCGA dataset in 5 different subtypes (Luminal A, Luminal B, Triple-negative/basal-like, HER2-enriched and Normal-like). More information on the subtypes can

be found in [47]. Next, we calculate the accuracy in a 10-fold cross validation for the 1046 TCGA features and the 100-miRNA signature, with results reported in Tables 5 and 6 respectively.

The best classification results are obtained for subtypes Triple-Negative Breast Cancer (TNBC) and Luminal A (LumA), due to the scarcity of samples for other subtypes (especially Normal and Her2). Luminal B (LumB) presents considerable similarities to LumA, and the classifiers have difficulty separating the two subtypes using the data at our disposal. For these reasons, and the practical concern that TNBC is the subtype of BRCA with the worst prognosis, we decide to tackle the issue as a binary classification problem, separating TNBC from the other classes. TNBC is a subtype of cancer where the cells have tested negative for estrogen receptors (ER), hormone epidermal growth factor receptor 2 (Her2), and progesterone receptors (PR). This subtype of cancer has limited treatment options and poor prognosis, as hormone therapies or targeted drugs do not work on it. Results of the binary classification problem on TCGA are reported in Table 7.

Table 5 Molecular subtype classification accuracy of Breast Cancer for the 1046 features

	Normal	LumA	LumB	TNBC	Her2	Global
#Samples	33	399	139	135	58	764
Gradient Boosting	0.1818	0.9348	0.5396	0.9333	0.5172	0.7987
Random Forest	0.0606	0.9724	0.4532	0.9630	0.0345	0.7657
Logistic Regression	0.1212	0.8747	0.5540	0.9259	0.4483	0.7606
Passive Aggressive	0.1515	0.8622	0.5612	0.9111	0.4483	0.7539
SGD	0.3030	0.9073	0.4604	0.9556	0.4655	0.7752
SVC	0.2727	0.8797	0.5252	0.9185	0.5345	0.7697
Ridge	0.1515	0.7293	0.4317	0.3704	0.2759	0.5524
Bagging	0.3333	0.9298	0.5108	0.9704	0.4310	0.7973
Average	0.1970	0.8863	0.5045	0.8685	0.3944	0.7467

Table 6 Molecular subtype classification accuracy of Breast Cancer for the 100 features

	Normal	LumA	LumB	TNBC	Her2	Global
#Samples	33	399	139	135	58	764
Gradient Boosting	0.2424	0.9248	0.5324	0.9333	0.5517	0.7975
Random Forest	0.2121	0.9599	0.4029	0.9704	0.2069	0.7712
Logistic Regression	0.2727	0.8997	0.4892	0.9037	0.5517	0.7727
Passive Aggressive	0.3939	0.8546	0.4460	0.8667	0.5000	0.7358
SGD	0.4545	0.8897	0.4460	0.8444	0.4310	0.7475
SVC	0.5152	0.8446	0.5108	0.9037	0.5517	0.7581
Ridge	0.0606	0.9474	0.4388	0.8593	0.3966	0.7594
Bagging	0.2727	0.9173	0.4964	0.9481	0.3793	0.7777
Average	0.3030	0.9048	0.4703	0.9037	0.4461	0.7650

Finally, we test the binary subtype classification of BRCA for the GEO datasets, using just the 100-miRNA signature. We create a single dataset composed of 4 series (GSE86281, GSE86277, GSE86278, GSE46823), with 2 classes: TNBC, featuring 139 samples, and all other molecular subtypes (LumA, LumB, and Her2), with 32 samples in total. Using the stem-loop sequences from platform GPL14613, and GPL1368, we use the 98 common stem-loop miRNAs of the 100 in the signature signature for the classification. In Table 8, we show the results of the classification in a 10-fold cross validation, and the accuracy by class.

Discussion

The results of the five experiments performed with the 100-miRNA signature (Tumor Type Classification, Tumor Tissue vs Normal Tissue, GEO datasets, BRCA subtype in TCGA, and BRCA subtype in GEO datasets), are reported in Table 9. All classifiers show high levels of accuracy over all trials, with the validation on the GEO datasets (both tumor type and subtype classification) proving to be the hardest task.

As miRNAs have been shown to regulate approximately 30% of the human genes, and because their dysregulation has been associated with the development and progression of cancer, miRNAs have been found to have the potential to play a critical role in computational oncology. Nevertheless, their analysis and their employment in clinically relevant settings still faces various, specific technical challenges: a) the extremely small size of the miRNAs leads to diverse complications for example with respect to hybridization techniques, b) there is a lack of specificity in detection because of the high similarity of several miRNA family members, and c) the low expression of various miRNAs requires detection methods of utmost sensitivity [48]. To date, most new miRNAs are discovered through cloning, despite these methods being time-consuming, low-throughput, and being biased toward the discovery of abundant miRNAs [49, 50].

Nevertheless, we can conclude from our results that the extracted 100-miRNA signature is able to reliably classify the 28 different types of cancer in the TCGA dataset, and distinguish between normal and tumor tissue. In addition, it is sufficiently stable to be applicable across platforms, such as the ones such as the ones used in the ten GEO datasets and which show a good accuracy in differentiating TNBC from other molecular subtypes of BRCA. Looking ahead into the possibility of classifying tumor types using miRNAs, we need to consider circulating miRNAs, and their relationship to cancer studies.

For the miRNAs included in the signature, we performed a bibliographic meta-analysis of specialized literature. The proposed meta-analysis is mainly based on 5 surveys of circulating miRNAs for cancer studies [6, 7, 51–53]. Out of the 100 miRNAs in the signature, 77 appear as circulatory miRNAs, either in their stem-loop form or mature sequence. The complete list for the 100-miRNAs is reported in Annex A of the online Additional file 1, in Fig. 5 shows the expression levels by type of cancer of the top 50 miRNAs.

Table 7 TNBC classification from the other molecular subtypes in the TCGA dataset, using 1046 features and 100 signature

	TNBC-100	TNBC-1046	Other-100	Other-1046	Global-100	Global-1046
#Samples	135	135	629	629	764	764
Gradient Boosting	0.9111	0.8963	0.9857	0.9857	0.9725	0.9699
Random Forest	0.8889	0.8815	0.9905	0.9905	0.9725	0.9712
Logistic Regression	0.8963	0.9630	0.9793	0.9587	0.9647	0.9593
Passive Aggressive	0.8815	0.9630	0.9714	0.9523	0.9556	0.9540
SGD	0.8000	0.8222	0.9809	0.9841	0.9490	0.9555
SVC	0.8444	0.8963	0.9666	0.9825	0.9451	0.9673
Ridge	0.8000	0.7259	0.9825	0.9237	0.9503	0.8888
Bagging	0.8444	0.8963	0.9793	0.9825	0.9555	0.9673
Average	0.8583	0.8806	0.9795	0.9700	0.9582	0.9542

Table 8 Molecular subtype classification of Breast Cancer to separate TNBC from other breast cancer subtypes using the 100-miRNA signature, on the GEO dataset

	TNBC	Other	Global
#Samples	139	44	183
Gradient Boosting	0.9353	0.7500	0.8909
Random Forest	0.9424	0.6136	0.8634
Logistic Regression	0.9065	0.6590	0.8476
Passive Aggressive	0.8561	0.7045	0.8197
SGD	0.9065	0.5227	0.8145
SVC	0.8561	0.7727	0.8355
Ridge	0.8993	0.6136	0.8300
Bagging	0.9496	0.7727	0.9070
Average	0.9065	0.6761	0.8511

Across all surveys analyzed, *hsa-miR-21*, included in our signature in stem-loop form, appears to be the most commonly over-expressed miRNA for all classes of tumors, as we would expect of a known oncomarker. In Annex B of the Additional file 1, we present a detailed analysis of the top 50 miRNAs in the signature, showing cancer study type, reference and circulating sample type used for measuring the expression. 23 miRNAs in the signature do not appear in the surveys, but they are mentioned in recent research papers, as promising research leads whose role may need further corroboration (we put the mature sequence as they appear in the study): *miR-211* [54], *miR-135a* [55], *miR-3678-3p* [56], *miR-204* [57], *miR-1228* [58], *miR-374b* [59], *miR-424* [60], *miR-217-5p* [60], *miR-3613-5p* [61], *miR-124* [62], *miR-1277-5p* [63], *miR-190* [64], *miR-934* [65], *miR-490* [66], *miR-1247* [67], *miR-199b* [68], *miR-135a* [55], *miR-503* [69], *miR-584* [70], *miR-137-3p* [71], and *miR-103* [72].

Table 9 Comparison of the 8 classifiers, for the different experiments with the 100-miRNA signature

Classifier	TT vs		TCGA	GEO	Global
	TCGA	NT	GEO (Subtype)	(Subtype)	
Gradient Boosting	0.9359	0.9846	0.6697	0.9725	0.8909
Random Forest	0.9324	0.9839	0.8085	0.9725	0.8634
Logistic Regression	0.9237	0.9799	0.9351	0.9647	0.8476
Passive Aggressive	0.8831	0.9606	0.8678	0.9556	0.8197
SGD	0.9035	0.9767	0.9393	0.9490	0.8145
SVC	0.9154	0.9791	0.7724	0.9451	0.8355
Ridge	0.8305	0.9470	0.8867	0.9503	0.8300
Bagging	0.9110	0.9812	0.7682	0.9555	0.9070

Logistic Regression was the best across all experiments, and Ridge has the worst accuracy

Interestingly, *hsa-mir-135a-1* and *hsa-mir-135a-2*, located inside chromosomes 3 and 12, respectively, generate the same mature active sequence [73]. In the same manner, *hsa-mir-124-1*, *hsa-mir-124-2*, and *hsa-mir-124-3*, generate the same mature sequence *hsa-miR-124-5p*, and *miR-124* is known as a tumor suppressor in head and neck squamous cell carcinoma [74], hepatocellular carcinoma [75] and breast cancer [76]. All of them were identified by our feature selection approach, indicting the presence of miRNA pathways shared across different tumor types. Targeting these miRNA pathways with anti-miRNA-based approaches such as infection with viral particles (having antisense sequence against the specific miRNA) or even drug design of small molecules inhibitors of miRNAs (SMIRs) which can be considered potential anti-tumoral therapy. On the other hand, the down regulation of tumor suppressor miRNAs also contributes to the acquisition of malignant features. For example, by ectopic expression of *hsa-miR-944* which decreases malignant features in gastric [77], colorectal [78] and endometrial [79] cancers. Strikingly, *miR-944* and other understudied miRNAs could have been detected by our approach analyzing 28 different types of cancer, suggesting that they could play a key role in the biology of cancer. Future works will include further analyses of the 100-miRNA signature, crossing the information with genetic sources, assessing measures of gene quality and biomarker stability, using tools such as sigQC [80].

Conclusions

miRNAs fine-tune the regulation of the transcriptome [81, 82]. Alterations in miRNA expression profiles are associated with several diseases, such as cancer. On the other hand, the altered miRNA expression profiles present in cancer could be used as prognostic and/or diagnostic markers. In summary, several miRNA signatures are associated with clinically relevant factors [83, 84]. Therefore, our miRNA signature, which we obtained by using data from different types of cancers, can highlight the presence of so far underestimated miRNA's such as *miR-944*, and overall has the potential to be used in the frame of microarray based assays, as a potential building block in clinical decision support. Of course, further experimental validation on cancer patient samples will be required to weigh the biological significance of the signature in terms of diagnosing, treating and prognosing the outcome of cancer.

In this study, we developed a new machine-learning approach to obtain a robust, reduced miRNA signature, from a TCGA dataset containing 28 different types of cancer. When tested against other datasets, our system provided good classification accuracy using only the reduced 100-feature signature, despite significant differences in the platforms used to gather the data. A further meta-analysis

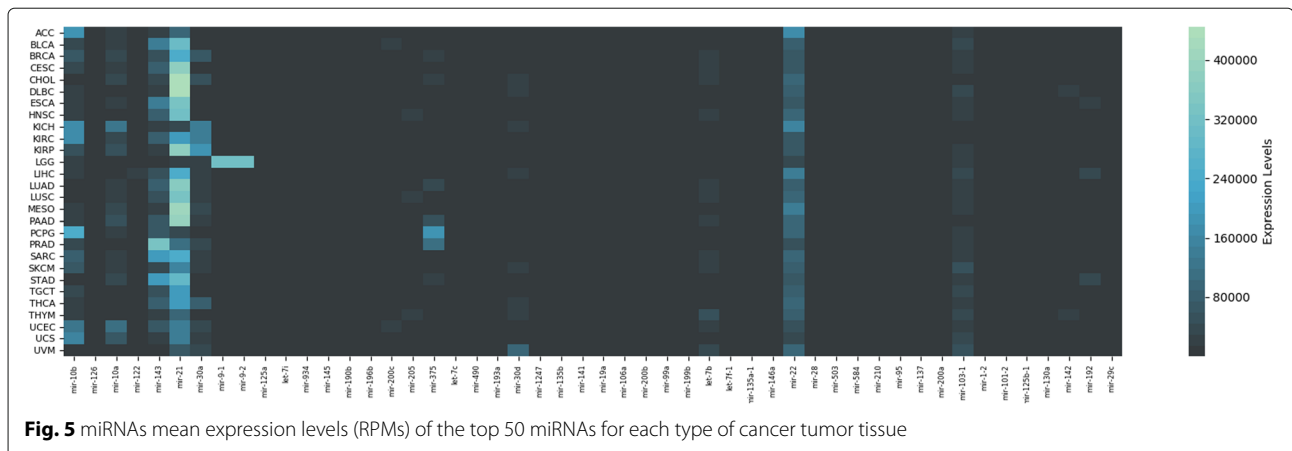


Fig. 5 miRNAs mean expression levels (RPMs) of the top 50 miRNAs for each type of cancer tumor tissue

of literature on the miRNA in the identified signature showed both well-known oncogenic and underestimated miRNA types. The results of this work could potentially be used to uncover new, promising leads of research for a better understanding of miRNA behavior. Furthermore, personal-directed anti-tumoral therapy could be achieved by measurement of the specific, minimal miRNA signature, identified in this work.

Methods

Ensemble feature selection

As the objective is to discover and validate a reduced list of miRNAs to be used as a signature for tumor classification, we need to select features that could optimally assist in distinguishing between different cancer types and tumor tissue. In this sense, popular approaches used for feature selection range from univariate statistical considerations, to iterated runs of the same classifier with a progressively reduced number of features in order to assess the contribution of the features to the overall result. As the considered problem is particularly complex, relying upon simple statistical analyses might not suffice. Furthermore, features extracted using an iterative method on one classifier are likely to work well only for that specific classifier. Following the idea behind *ensemble feature selection* [36, 37, 85], we propose the use of multiple algorithms to obtain a more robust and general predictive performance. An ensemble approach has the advantage of obtaining features that will be effective across several classifiers, with a better likelihood of being more representative of the data, and not just of the inner workings of a single classifier.

For this purpose, we train a set of classifiers in order to extract a sorted list of the most relevant features from each. Intuitively, as a feature considered important by the majority of classifiers in the set is also likely to be relevant for our objective, then information from all classifiers is compiled to find the most common relevant features. Starting from a comparison of 22 different state-of-the-art classifiers on the considered dataset, presented in [86], a

subset of those classifiers was selected considering both; high accuracy and a way to extract the relative importance of the features from the trained classifier. After preliminary tests to set algorithms' hyperparameters, 8 classifiers were chosen, all featuring an average accuracy higher than 90% on a 10-fold cross-validation: Bagging [87], Gradient Boosting [88], Logistic Regression [89], Passive Aggressive [90], Random Forest [91], Ridge [92], SGD (Stochastic Gradient Descent on linear models) [93], SVC (Support Vector Machines Classifier with a linear kernel) [94]. All considered classifiers are implemented in the `scikit-learn` Python toolbox.

Overall, the selected classifiers fall into two broad typologies: those exploiting ensembles of classification trees [95] (Bagging, Gradient Boosting, Random Forest), and those optimizing the coefficients of linear models to separate classes (Logistic Regression, Passive Aggressive, Ridge, SGD, SVC). Depending on classifier typology, there are two different ways of extracting relative feature importance. For classifiers based on classification trees, the features used in the splits are counted and sorted by frequency, from the most to the least common. For classifiers based on linear models, the values of the coefficients associated to each feature can be used as a proxy of their relative importance, sorting coefficients from the largest to the smallest in absolute value. As the two feature extraction methods return heterogeneous numeric values, only the relative sorting of features provided by each classifier was considered. Furthermore, we decide to extract the top 100 most relevant features as a reduction of about an order of magnitude, so we assign to each feature f a simple score $s_f = N_f/N_c$, where N_f is the number of times that specific feature appears among the top 100 of a specific classifier instance, while N_c is the total number of classifiers instances used; for instance, a feature appearing among the 100 most relevant in 73% of the classifiers used would obtain a score $s_f = 0.73$. We select 100 features because we wanted to compress the dataset at least 90%, thus from 1046 we reduce it to

100. In order to increase the generality of our results, each selected classifier was run 10 times, using a 10-fold stratified cross-validation, so that each fold preserves the percentage of samples of each class in the original dataset. Thus, $N_c = 80$ (8 types of classifiers, run 10 times each). The complete procedure is summarized by Algorithm 1. Different approaches to the aggregation of heterogeneous feature importance from various sources are also possible (see for example [36, 37, 85]), such as assigning to each feature a weight proportional to its relative importance. However, most alternatives would require adding and tuning extra parameters, so we decided to opt for a simpler approach.

TCGA dataset

The data was downloaded from the TCGA Data Portal², on September 1, 2016. The used data is miRNA-SEQ

files (*.mirna.quantification.txt) a total of 1046 miRNA expression features for each sample in format mirbase V16 for stem-loop sequences³. We consider the read per million (RPM) values in the file and we remove all of the samples where the item does not meet the study protocol as stated in the *file annotations*. In summary, the dataset used in the following experiments includes 28 types of tumors, 1046 miRNA features, and 8023 patient samples. Information on the dataset is summarized in Table 10. We standardized the data by removing the mean and scaling to unit variance (specifying that we had learned the standardization on the training set, and applied it to the test set, so that knowledge of the whole dataset did not bias the performance on the test set). In addition, we created a second dataset that differentiates between normal tissue (NT) and tumor tissue (TT) that consists of 8657 samples; 8023 TT and 634 NT.

Algorithm 1: Ensemble feature selection.

```

1 Normalize dataset on each of the  $F$  features, Divide
  dataset in  $N$  folds, Select  $K$  classifiers, Choose the
  number of features in the signature  $S$ ;
2 for each fold  $n$  of  $N$  do
3   for each classifier  $k$  of  $K$  do
4     Train classifier  $k_n$  on all folds minus  $n$ , using all
      features;
      Test classifier  $k_n$  on fold  $n$ ;
      Obtain sorted list  $l_{kn}$  of features from  $k_n$ ;
      Assign weight  $w_{fkn}$  to each  $f$  of the  $F$  features;
5   for each feature  $f$  of  $F$  do
6     if  $f$  is among the top  $S$  features in  $l_{kn}$  then
7        $w_{fkn} = 1$ 
8     else
9        $w_{fkn} = 0$ 
10   $N_c = N \cdot K$ ;
11  for each miRNA feature  $f$  do
12     $N_t = \sum_n \sum_k w_{fkn}$ ;
13     $s_f = N_t / N_c$ ;
14  Select  $S$ -feature signature, from features with highest
     $s_f$ ;
15  for each fold  $n$  of  $N$  do
16    for each classifier  $k$  of  $K$  do
17      Train classifier  $k_n$  on all folds minus  $n$ , using
        signature;
        Test classifier  $k_n$  on fold  $n$ ;
18  Compare performance of classifiers using all features
    and signature;
```

Geo datasets

To validate our results, we use 14 datasets from the GEO repository⁴, from 5 different platforms. We use 2 types of miRNA discovery technologies: microarrays and sequencing. miRNAs expression levels are platform and technology dependent [96–98]. Therefore, we need to consider if the information is in stem-loop or mature sequence and then calculate the contributions to make a direct comparison.

In the TCGA dataset, stem-loop sequences were directly measured in raw read counts. When reading a mature sequence, the protocol that was followed assigns a read count to it, and then randomly assigns a read count to one of the stem-loop sequences that share the same mature sequence [99].

GPL8786, gPL10850

Affymetrix Multispecies miRNA-1 Array (GPL8786) and Agilent-021827 Human miRNA Microarray V3 (GPL10850) cannot read stem-loop sequences, so the corresponding GEO datasets only show information for mature sequences. Thus, in order to perform a fair comparison, we consider the raw read count for stem-loop sequences as a linear function of the read counts of the mature sequences. If we call the read counts of a specific stem-loop sequence X_i , for *hsa-mir-10b* we have for example:

$$X_{hsa-mir-10b} = a_0 \cdot X_{hsa-miR-10b} + a_1 \cdot X_{hsa-miR-10b*} \quad (1)$$

Where a_0 and a_1 are two coefficients to be set. The mapping between the values of two different platforms $P1$ and $P2$ can then be written as:

²<https://tcga-data.nci.nih.gov/docs/publications/tcga/>

³<ftp://mirbase.org/pub/mirbase/16/genomes/hsa.gff>

⁴<https://www.ncbi.nlm.nih.gov/gds>

Table 10 Summary of the TCGA dataset used in the study

Tumor Type	Acronym	Tumor Tissue	Normal Tissue	Class
Adrenocortical carcinoma	ACC	80	0	0
Bladder Urothelial Carcinoma	BLCA	411	19	1
Breast invasivex carcinoma	BRCA	777	87	2
Cervical squamous cell carcinoma	CESC	306	3	3
Cholangiocarcinoma	CHOL	36	9	4
Lymphoid Neoplasm Diffuse Large B-cell Lymphoma	DLBC	47	0	5
Esophageal carcinoma	ESCA	187	13	6
Head and Neck squamous cell carcinoma	HNSC	487	44	7
Kidney Chromophobe	KICH	66	25	8
Kidney renal clear cell carcinoma	KIRC	260	71	9
Kidney renal papillary cell carcinoma	KIRP	291	34	10
Lower Grade Glioma	LGG	528	0	11
Liver hepatocellular carcinoma	LIHC	374	50	12
Lung adenocarcinoma	LUAD	458	46	13
Lung squamous cell carcinoma	LUSC	341	45	14
Mesothelioma	MESO	86	0	15
Pancreatic adenocarcinoma	PAAD	154	4	16
Pheochromocytoma and Paraganglioma	PCPG	184	3	17
Prostate adenocarcinoma	PRAD	494	52	18
Sarcoma	SARC	260	0	19
Skin Cutaneous Melanoma	SKCM	450	2	20
Stomach adenocarcinoma	STAD	399	45	21
Testicular Germ Cell Tumors	TGCT	156	0	22
Thyroid carcinoma	THCA	513	59	23
Thymoma	THYM	124	2	24
Uterine Corpus Endometrial Carcinoma	UCEC	417	21	25
Uterine Carcinosarcoma	UCS	57	0	26
Uveal Melanoma	UVM	80	0	27
Total		8023	634	

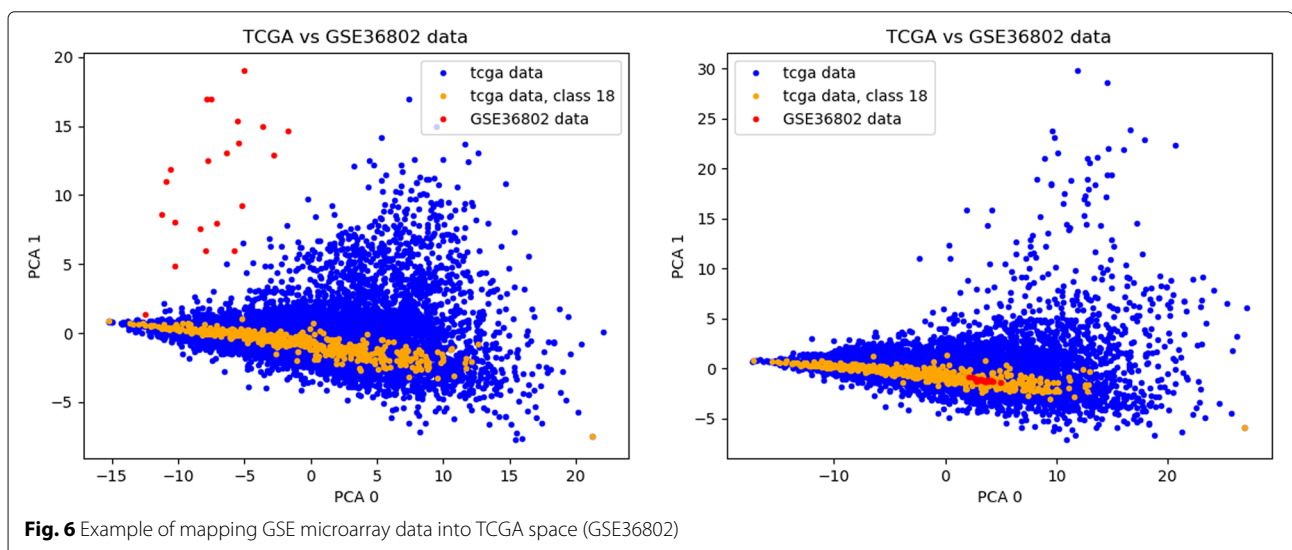
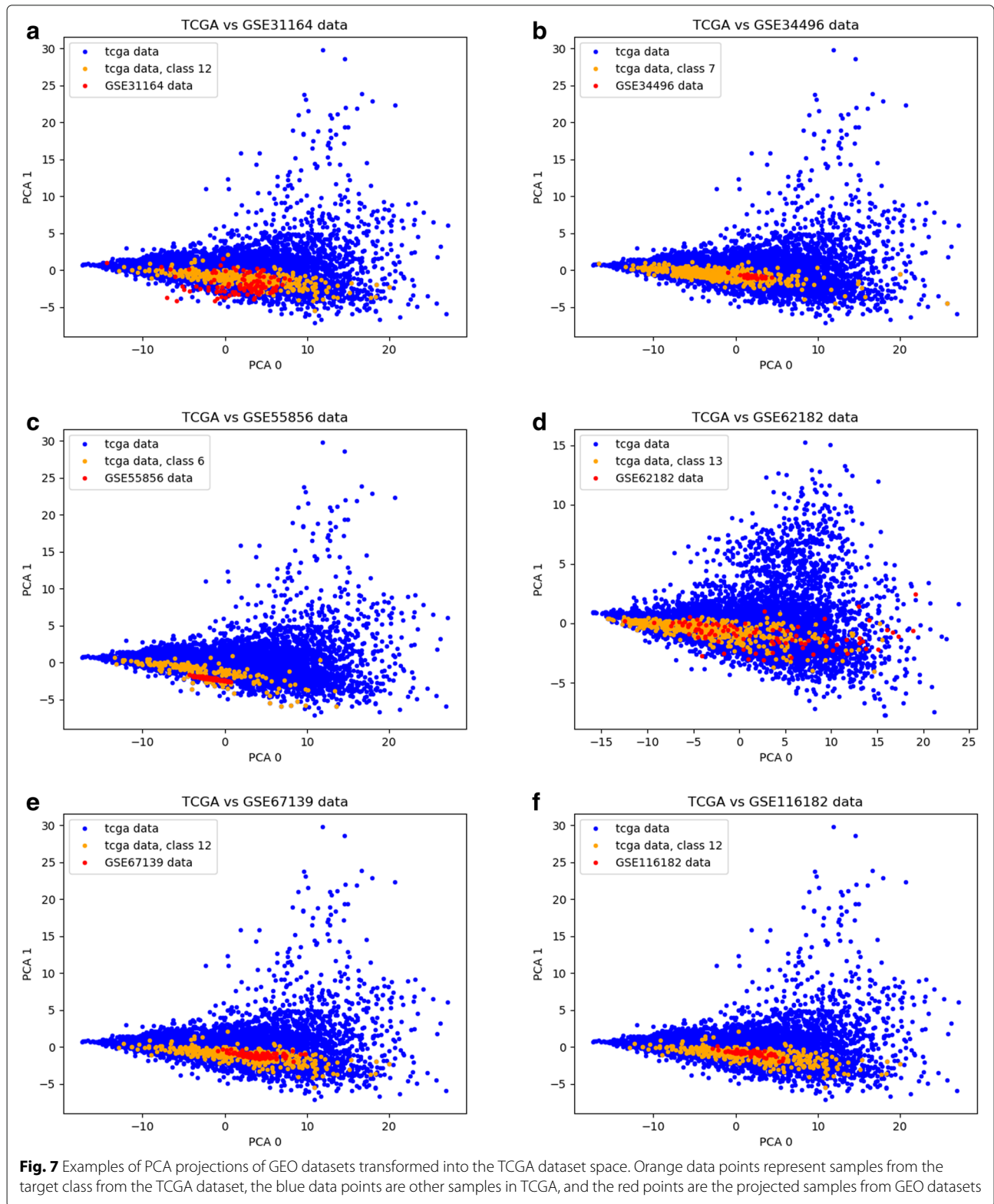


Fig. 6 Example of mapping GSE microarray data into TCGA space (GSE36802)



$$X_{hsa-mir-10b}^{P1} = a_2 \cdot X_{hsa-mir-10b}^{P2} \quad (2)$$

To reduce the problem, we consider only relationships between a stem-loop sequence and its most common corresponding mature sequence e.g hsa-mir-10b to hsa-miR-10b, disregarding hsa-miR-10b*. From Eq. 1 and 2 we then have:

$$\begin{aligned} X_{hsa-mir-10b}^{P1} &= a_2 \cdot X_{hsa-mir-10b}^{P2} \\ X_{hsa-mir-10b}^{P1} &= a_2 \cdot (a_0 \cdot X_{hsa-miR-10b}^{P2} + a_1 \cdot X_{hsa-miR-10b*}^{P2}) \\ X_{hsa-mir-10b}^{P1} &= a_2 \cdot a_0 \cdot X_{hsa-miR-10b}^{P2} \\ X_{hsa-mir-10b}^{P1} &= a_{hsa-miR-10b}^P \cdot X_{hsa-miR-10b}^{P2} \end{aligned}$$

where a_i^P becomes the only coefficient to be found, and it represents the transformation between platforms for that specific sequence. A different linear function will be found for each pair of platforms, as we assume that each machine will have unique properties.

For GPL8786 GEO datasets, we consider the linear gene expression values given by the function `rmasummary` from the Matlab bioinformatics toolbox, which is a normalized robust multi-array average procedure, as a z-score [100, 101]. The equation of a z-score is:

$$Z = \frac{(X - \mu)}{\sigma} \quad (3)$$

where X is the value of a feature; μ and σ are the average and the standard deviation for a feature. Next, by considering the linear expression values as z-scores, the GEO datasets are mapped to corresponding intensities in the TCGA dataset space, by solving for X :

$$X_i = \left(Z_i \cdot (\sigma_i^{TCGA}) + \mu_i^{TCGA} \right) \cdot a_i^P \quad (4)$$

where X_i is the intensity of miRNA i in the TCGA dataset space, Z_i is the linear gene expression value given by the scaled `rmasummary` summary function, μ_i^{TCGA} and σ_i^{TCGA} are the average value and the standard deviation for miRNA i , both computed on the original TCGA dataset, and a_i^P is a scale value, dependent on the platform. The value a_i^P is computed using a subset of all the GEO datasets from the same platform, by minimizing the error between actual class and predicted class, using a model trained in the TCGA dataset with Root Mean Squared Error (RMSE).

$$RMSE = \sqrt{\frac{\sum_{s=1}^S Predicted_s(TCGA, a^P) - Actual_s(TCGA)}{S}} \quad (5)$$

where S is the total number of samples in the dataset, and a^P is a vector containing the values of a_i^P for each feature i . A state-of-the-art numerical optimizer [102] is applied to this task, to find the 98 parameters represented by a^P .

For GPL10850 we use the MatLab function `agferead` from the Bioinformatics Toolbox and use the value of `gTotalGeneSignal` as value for each of the probes and calculate the contributions and a_i^P as for GPL8786.

GPL14613, gPL16384

Affymetrix Multispecies miRNA-2 Array (GPL14613) and Affymetrix Multispecies miRNA-3 Array (GPL16384) measure the stem-loop sequences directly, and denote them by `hp_hsa`. The linear relationship between the TCGA dataset and the corresponding subset of GEO datasets is thus represented by Eq. 2, and the a_i^P parameters to be found are reduced to the a_{2i}

As remarked by Telonis et al. [21], for these datasets, not all the types of cancer are available, or present the necessary quality standards. Thus, we reduce our analysis to 6 different types of cancer; Prostate, Liver, Breast, Esophageal, Head and Neck Squamous Cell and Lung. For the sequencing data, extra mapping is not necessary besides the sample normalization (platform GPL11154), and we use only stem-loop sequences.

Using this procedure, we are able to map the GEO repository measurements into the TCGA dataset space as seen in Fig. 6. Other examples are shown in Fig. 7, where plots were created using the first two dimensions of a Principal Component Analysis (PCA) computed on the TCGA dataset and applied to the GEO datasets, to provide a comparison between the cancer type in each GEO and the corresponding class in TCGA. Remarkably, samples from GEO datasets are often considerably close to samples of the corresponding class in TCGA. During validation, we selected the common features between each GEO dataset and the 100-miRNA signature obtained using the ensemble approach. The accuracy of the classification algorithms was then evaluated by training them on the TCGA dataset and testing them on each GEO dataset. A summary of the experiments is presented in Fig. 1.

Additional file

Additional file 1: Annex A-List of the top 100 most relevant features identified by the proposed methodology, in order of importance.
Annex B-Table comparing the top 50 most frequent features extracted by the machine learning algorithms with existing biomarkers references in literature.
Annex C- Figure with all PCA projections of GEO datasets.
Annex D- Explanation of n-fold cross-validation. (PDF 505 kb)

Abbreviations

ACC: Adrenocortical carcinoma; BLCA: Bladder Urothelial carcinoma; BRCA: Breast invasive carcinoma; CESC: Cervical squamous cell carcinoma; CHOL: Cholangiocarcinoma; DLBC: Lymphoid neoplasm diffuse large B-cell lymphoma; EFS-CLA: Ensemble feature selection with complete linear aggregation; EN: Elastic net; ESCA: Esophageal carcinoma; GEO: Gene expression omnibus; HNSC: Head and neck squamous cell carcinoma; KICH: Kidney chromophobe; KIRC: Kidney renal clear cell carcinoma; KIRP: Kidney renal papillary cell carcinoma; LASSO: Least absolute shrinkage and selection operator; LGG: Lower grade glioma; LIHC: Liver hepatocellular carcinoma;

LumA: Luminal A; LUAD: Lung adenocarcinoma; LumB: Luminal B; LUSC: Lung squamous cell carcinoma; MESO: Mesothelioma; miRNA: microRNA; NT: Normal tissue; PAAD: Pancreatic adenocarcinoma; PCA: Principal component analysis; PCPG: Pheochromocytoma and paraganglioma; RMSE: Root mean squared error; PRAD: Prostate adenocarcinoma; RFE: Recursive feature elimination; RPM: Read per million; SARC: Sarcoma; SGD: Stochastic gradient descent; SKCM: Skin cutaneous melanoma; STAD: Stomach adenocarcinoma; SVC: Support vector machines classifier; TCGA: The cancer genome atlas; TGCT: Testicular germ cell tumors; THCA: Thyroid carcinoma; THYM: Thymoma; TNBC: Triple negative breast cancer; TT: Tumor tissue; UCEC: Uterine corpus endometrial carcinoma; UCS: Uterine carcinosarcoma; UFS: Univariate feature selection; UVM: Uveal melanoma

Acknowledgements

The results published here are based upon data generated by The Cancer Genome Atlas and GSM.

Authors' contributions

ALR suggested the problem, wrote, built the datasets and coded. MMA and GMR helped with the writing, miRNAs concepts, and the bibliographic analysis. AS wrote and conceived the cross-platform validation, and comparison to other methods. AT got funding, wrote, and coded. All authors have read and approved the final manuscript.

Funding

Not applicable.

Availability of data and materials

The code and the datasets are available at <https://github.com/steppenwolf0/miRNAs100>

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

¹Division of Pharmacology, Utrecht Institute for Pharmaceutical Sciences, Faculty of Science, Utrecht University, David de Wied building, Universiteitsweg 99, 3584 CG Utrecht, The Netherlands. ²Laboratorio de Modelado Molecular, Bioinformática y diseño de fármacos. Departamento de Posgrado. Escuela Superior de Medicina del Instituto Politécnico Nacional (IPN), Mexico City, Mexico. ³Faculty of Medicine, National Autonomous University of Mexico; Federico Gomez Children's Hospital of Mexico, Mexico City, Mexico. ⁴Life Sciences and Health, CWI, Amsterdam, Netherlands. ⁵UMR 782 GMPA, Université Paris-Saclay, INRA, AgroParisTech, Thiverval-Grignon, France.

Received: 15 March 2019 Accepted: 22 August 2019

Published online: 18 September 2019

References

- Ferlay J, Soerjomataram I, Dikshit R, Eser S, Mathers C, Rebelo M, Parkin DM, Forman D, Bray F. Cancer incidence and mortality worldwide: sources, methods and major patterns in globocan 2012. *Int J Cancer*. 2015;136(5):359–86.
- Tanase C, OGREZEANU I, BADIU C. Molecular Pathology of Pituitary Adenomas: Elsevier Insights; 2012, p. 130.
- Calin GA, Dumitru CD, Shimizu M, Bichi R, Zupo S, Noch E, Aldler H, Rattan S, Keating M, Rai K, et al. Frequent deletions and down-regulation of micro-rna genes mir15 and mir16 at 13q14 in chronic lymphocytic leukemia. *Proc Natl Acad Sci*. 2002;99(24):15524–9.
- Peng Y, Croce CM. The role of microRNAs in human cancer. *Signal Transduct Target Ther*. 2016;1:15004.
- Sauter ER, Patel N. Body fluid micro (mi) rnas as biomarkers for human cancer. *J Nucleic Acids Investig*. 2011;2(1):1.
- He Y, Lin J, Kong D, Huang M, Xu C, Kim T-K, Etheridge A, Luo Y, Ding Y, Wang K. Current state of circulating microRNAs as cancer biomarkers. *Clin Chem*. 2015;61(9):1138–1155. <https://doi.org/10.1373/clinchem.2015.241190>.
- Calore F, Lovat F, Garofalo M. Non-coding rnas and cancer. *Int J Mol Sci*. 2013;14(8):17085–110.
- Ferracin M, Veronese A, Negrini M. Micromarkers: miRNAs in cancer diagnosis and prognosis. *Expert Rev Mol Diagn*. 2010;10(3):297–308.
- Fabbri M. Non-coding RNAs and Cancer: Springer Science + Business Media, LCC; 2014. https://doi.org/10.1007/978-1-4614-8444-8_10.
- Liu B, Li J, Cairns MJ. Identifying miRNAs, targets and functions. *Brief Bioinform*. 2012;15(1):1–19.
- Akhtar MM, Micolucci L, Islam MS, Olivieri F, Procopio AD. Bioinformatic tools for microRNA dissection. *Nucleic Acids Res*. 2015;44(1):24–44.
- Bhattacharya A, Ziebarth JD, Cui Y. Somamir: a database for somatic mutations impacting microRNA function in cancer. *Nucleic Acids Res*. 2012;41(D1):977–82.
- Kozomara A, Griffiths-Jones S. mirbase: integrating microRNA annotation and deep-sequencing data. *Nucleic Acids Res*. 2010;39(suppl_1):152–7.
- Bartels CL, Tsongalis GJ. MicroRNAs: novel biomarkers for human cancer. *Clin Chem*. 2009;55(4):623–31.
- Cortez MA, Bueso-Ramos C, Ferdin J, Lopez-Berestein G, Sood AK, Calin GA. MicroRNAs in body fluids—the mix of hormones and biomarkers. *Nat Rev Clin Oncol*. 2011;8(8):467.
- Iorio MV, Croce CM. MicroRNA dysregulation in cancer: diagnostics, monitoring and therapeutics. a comprehensive review. *EMBO Mol Med*. 2012;4(3):143–59.
- Gao W, Shen H, Liu L, Xu J, Xu J, Shu Y. Mir-21 overexpression in human primary squamous cell lung carcinoma is associated with poor patient prognosis. *J Cancer Res Clin Oncol*. 2011;137(4):557–66.
- Zhi F, Chen X, Wang S, Xia X, Shi Y, Guan W, Shao N, Qu H, Yang C, Zhang Y, et al. The use of hsa-mir-21, hsa-mir-181b and hsa-mir-106a as prognostic indicators of astrocytoma. *Eur J Cancer*. 2010;46(9):1640–9.
- Yan L-X, Huang X-F, Shao Q, Huang M-Y, Deng L, Wu Q-L, Zeng Y-X, Shao J-Y. MicroRNA mir-21 overexpression in human breast cancer is associated with advanced clinical stage, lymph node metastasis and patient poor prognosis. *Rna*. 2008;14(11):2348–60.
- Wang D, Fan Z, Liu F, Zuo J. Hsa-mir-21 and hsa-mir-29 in tissue as potential diagnostic and prognostic biomarkers for gastric cancer. *Cell Physiol Biochem*. 2015;37(4):1454–62.
- Telonis AG, Magee R, Loher P, Chervoneva I, Londin E, Rigoutsos I. Knowledge about the presence or absence of miRNA isoforms (isomirs) can successfully discriminate amongst 32 tcga cancer types. *Nucleic Acids Res*. 2017;45(6):2973–85.
- Yousef M, Allmer J, Khalifa W. Feature selection for microRNA target prediction comparison of one-class feature selection methodologies. Conference Paper. DSpace@ZTECH. 2016. <https://doi.org/10.5220/0005701602160225>.
- Tang W, Wan S, Yang Z, Teschendorff AE, Zou Q. Tumor origin detection with tissue-specific miRNA and dna methylation markers. *Bioinformatics*. 2017;34(3):398–406.
- Piao Y, Piao M, Ryu KH. Multiclass cancer classification using a feature subset-based ensemble from microRNA expression profiles. *Comput Biol Med*. 2017;80:39–44.
- Weinstein JN, Collisson EA, Mills GB, Shaw KRM, Ozenberger BA, Ellrott K, Shmulevich I, Sander C, Stuart JM, Network CGAR, et al. The cancer genome atlas pan-cancer analysis project. *Nat Genet*. 2013;45(10):1113.
- Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, et al. Scikit-learn: Machine learning in python. *J Mach Learn Res*. 2011;12(Oct):2825–30.
- Altman N, Krzywinski M. Points of Significance: Ensemble methods: bagging and random forests. *Nat Publ Group*. 2017;14(10):933–4. Part of Springer Nature.
- Hira ZM, Gillies DF. A review of feature selection and feature extraction methods applied on microarray data. *Adv Bioinforma*. 2015;2015:1–13.
- Lazo AV, Rathie P. On the entropy of continuous probability distributions (corresp.) *IEEE Trans Inf Theory*. 1978;24(1):120–2.
- Guyon I, Weston J, Barnhill S, Vapnik V. Gene selection for cancer classification using support vector machines. *Mach Learn*. 2002;46(1-3):389–422.
- Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. *J Stat Softw*. 2010;33(1):1.

32. Sokolov A, Carlin DE, Paull EO, Baertsch R, Stuart JM. Pathway-based genomics prediction using generalized elastic net. *PLoS Comput Biol*. 2016;12(3):1004790.
33. Basu A, Mitra R, Liu H, Schreiber SL, Clemons PA. Rwen: Response-weighted elastic net for prediction of chemosensitivity of cancer cell lines. *Bioinformatics*. 2018;1:8.
34. Tibshirani R. Regression shrinkage and selection via the lasso. *J R Stat Soc Ser B Methodol*. 1996;58(1):267–88.
35. Trevino V, Falciani F. Galgo: an R package for multivariate variable selection using genetic algorithms. *Bioinformatics*. 2006;22(9):1154–6.
36. Abeel T, Helleputte T, Van de Peer Y, Dupont P, Saeys Y. Robust biomarker identification for cancer diagnosis with ensemble feature selection methods. *Bioinformatics*. 2009;26(3):392–8.
37. Seijo-Pardo B, Porto-Diaz I, Bolon-Canedo V, Alonso-Betanzos A. Ensemble feature selection: Homogeneous and heterogeneous approaches. *Knowl-Based Syst*. 2017;118:124–39. <https://doi.org/10.1016/j.knosys.2016.11.017>.
38. Lin P-C, Chiu Y-L, Banerjee S, Park K, Mosquera JM, Giannopoulou E, Alves P, Tewari AK, Gerstein MB, Beltran H, et al. Epigenetic repression of mir-31 disrupts androgen receptor homeostasis and contributes to prostate cancer progression. *Cancer Res*. 2013;73(3):1232–44.
39. Casanova-Salas I, Rubio-Briones J, Calatrava A, Mancarella C, Masiá E, Casanova J, Fernández-Serra A, Rubio L, Ramírez-Backhaus M, Armiñán A, et al. Identification of mir-187 and mir-182 as biomarkers of early diagnosis and prognosis in patients with prostate cancer treated with radical prostatectomy. *J Urol*. 2014;192(1):252–9.
40. Peña-Chilet M, Martínez MT, Pérez-Fidalgo JA, Peiró-Chova L, Oltra SS, Tormo E, Alonso-Yuste E, Martínez-Delgado B, Eroles P, Climent J, et al. MicroRNA profile in very young women with breast cancer. *BMC Cancer*. 2014;14(1):529.
41. Jang H-J, Lee H-S, Burt BM, Lee GK, Yoon K-A, Park Y-Y, Sohn BH, Kim SB, Kim MS, Lee JM, et al. Integrated genomic analysis of recurrence-associated small non-coding RNAs in oesophageal cancer. *Gut*. 2017;66(2):215–25.
42. Romero-Cordoba SL, Rodriguez-Cuevas S, Bautista-Pina V, Maffuz-Aziz A, D'Ippolito E, Cosentino G, Baroni S, Iorio MV, Hidalgo-Miranda A. Loss of function of mir-342-3p results in mct1 over-expression and contributes to oncogenic metabolic reprogramming in triple negative breast cancer. *Sci Rep*. 2018;8(1):12252.
43. Murakami Y, Tamori A, Itami S, Tanahashi T, Toyoda H, Tanaka M, Wu W, Brojigin N, Kaneoka Y, Maeda A, et al. The expression level of mir-18b in hepatocellular carcinoma is associated with the grade of malignancy and prognosis. *BMC Cancer*. 2013;13(1):99.
44. Vucic EA, Thu KL, Pikor LA, Enfield KS, Yee J, English JC, MacAulay CE, Lam S, Jurisica I, Lam WL. Smoking status impacts microRNA mediated prognosis and lung adenocarcinoma biology. *BMC Cancer*. 2014;14(1):778.
45. Network CGA, et al. Comprehensive molecular portraits of human breast tumours. *Nature*. 2012;490(7418):61.
46. Colaprico A, Silva TC, Olsen C, Garofano L, Cava C, Garolini D, Sabedot TS, Malta TM, Pagnotta SM, Castiglioni I, et al. TCGAbiolinks: an R/bioconductor package for integrative analysis of TCGA data. *Nucleic Acids Res*. 2015;44(8):71.
47. Weiss M. Your guide to the breast cancer pathology report. *Breastcancer.org*. 2016. <https://www.breastcancer.org>.
48. Li X, Ni M, Zhang C, Ma W, Zhang Y. A convenient system for highly specific and sensitive detection of miRNA expression. *RNA*. 2014;20(2):252–9.
49. Chen Y, Gelfond JA, McManus LM, Shireman PK. Reproducibility of quantitative RT-PCR array in miRNA expression profiling and comparison with microarray analysis. *BMC Genomics*. 2009;10(1):407.
50. Li W, Ruan K. MicroRNA detection by microarray. *Anal Bioanal Chem*. 2009;394(4):1117–24.
51. Larrea E, Sole C, Manterola L, Goicoechea I, Armesto M, Arestin M, Caffarel MM, Araujo AM, Araiz M, Fernandez-Mercado M, et al. New concepts in cancer biomarkers: circulating miRNAs in liquid biopsies. *Int J Mol Sci*. 2016;17(5):627.
52. Cheng G. Circulating miRNAs: roles in cancer diagnosis, prognosis and therapy. *Adv Drug Deliv Rev*. 2015;81:75–93.
53. Wang J, Zhang K-Y, Liu S-M, Sen S. Tumor-associated circulating microRNAs as biomarkers of cancer. *Molecules*. 2014;19(2):1912–38.
54. Margue C, Reinsbach S, Philippidou D, Beaume N, Walters C, Schneider JG, Nashan D, Behrmann I, Kreis S. Comparison of a healthy mirnome with melanoma patient mirnomes: are microRNAs suitable serum biomarkers for cancer? *Oncotarget*. 2015;6(14):12110.
55. Koga Y, Yasunaga M, Takahashi A, Kuroda J, Moriya Y, Akasu T, Fujita S, Yamamoto S, Baba H, Matsumura Y. MicroRNA expression profiling of exfoliated colonocytes isolated from feces for colorectal cancer screening. *Cancer Prev Res*. 2010;3(11):1435–42.
56. Giulietti M, Occhipinti G, Principato G, Piva F. Identification of candidate miRNA biomarkers for pancreatic ductal adenocarcinoma by weighted gene co-expression network analysis. *Cell Oncol*. 2017;40(2):181–92.
57. Mengual L, Lozano JJ, Ingelmo-Torres M, Gazquez C, Ribal MJ, Alcaraz A. Using microRNA profiling in urine samples to develop a non-invasive test for bladder cancer. *Int J Cancer*. 2013;133(11):2631–41.
58. Tan Y, Ge G, Pan T, Wen D, Chen L, Yu X, Zhou X, Gan J. A serum microRNA panel as potential biomarkers for hepatocellular carcinoma related with hepatitis B virus. *PLoS ONE*. 2014;9(9):107986.
59. Summerer I, Unger K, Braselmann H, Schuettrumpf L, Maihoefer C, Baumeister P, Kirchner T, Niyazi M, Sage E, Specht H, et al. Circulating microRNAs as prognostic therapy biomarkers in head and neck cancer patients. *Br J Cancer*. 2015;113(1):76.
60. Giraldez MD, Lozano JJ, Ramirez G, Hijona E, Bujanda L, Castells A, Gironella M. Circulating microRNAs as biomarkers of colorectal cancer: results from a genome-wide profiling and validation study. *Clin Gastroenterol Hepatol*. 2013;11(6):681–8.
61. Matamala N, Vargas MT, González-Cámpora R, Miñambres R, Arias JI, Menéndez P, Andrés-León E, Gómez-López G, Yanowsky K, Calvete-Candenas J, et al. Tumor microRNA expression profiling identifies circulating microRNAs for early breast cancer detection. *Clin Chem*. 2015;61(8):1098–106.
62. Medina-Villaamil V, Martínez-Breijo S, Portela-Pereira P, Quindós-Varela M, Santamarina-Cainzos I, Antón-Aparicio L, Gómez-Veiga F. Circulating microRNAs in blood of patients with prostate cancer. *Actas Urol Esp (Engl Ed)*. 2014;38(10):633–9.
63. Zheng X-H, Cui C, Ruan H-L, Xue W-Q, Zhang S-D, Hu Y-Z, Zhou X-X, Jia W-H. Plasma microRNA profiling in nasopharyngeal carcinoma patients reveals mir-548q and mir-483-5p as potential biomarkers. *Chin J Cancer*. 2014;33(7):330.
64. Scheffer A-R, Holdenrieder S, Kristiansen G, von Ruecker A, Müller SC, Ellinger J. Circulating microRNAs in serum: novel biomarkers for patients with bladder cancer? *World J Urol*. 2014;32(2):353–8.
65. Tsuchiya N, Ogata H, Okusaka T, Nakagama H. Method for detecting pancreatic cancer and detection kit. Google Patents. US Patent APP. 14/410,408. 2015. <https://www.google.com>.
66. Jiang Y, Luan Y, Chang H, Chen G. The diagnostic and prognostic value of plasma microRNA-125b-5p in patients with multiple myeloma. *Oncol Lett*. 2018;16(3):4001–7.
67. Wang J, Raimondo M, Guha S, Chen J, Diao L, Dong X, Wallace MB, Killary AM, Frazier ML, Woodward TA, et al. Circulating microRNAs in pancreatic juice as candidate biomarkers of pancreatic cancer. *J Cancer*. 2014;5(8):696.
68. Montalbo R, Izquierdo L, Ingelmo-Torres M, Lozano JJ, Capitán D, Alcaraz A, Mengual L. Prognostic value of circulating microRNAs in upper tract urinary carcinoma. *Oncotarget*. 2018;9(24):16691.
69. Shin VY, Ng EK, Chan VW, Kwong A, Chu K-M. A three-miRNA signature as promising non-invasive diagnostic marker for gastric cancer. *Mol Cancer*. 2015;14(1):202.
70. Wang H, Peng R, Wang J, Qin Z, Xue L. Circulating microRNAs as potential cancer biomarkers: the advantage and disadvantage. *Clin Epigenetics*. 2018;10(1):59.
71. Hsu C-M, Lin P-M, Wang Y-M, Chen Z-J, Lin S-F, Yang M-Y. Circulating miRNA is a novel marker for head and neck squamous cell carcinoma. *Tumor Biol*. 2012;33(6):1933–42.
72. Jiang X, Du L, Duan W, Wang R, Yan K, Wang L, Li J, Zheng G, Zhang X, Yang Y, et al. Serum microRNA expression signatures as novel noninvasive biomarkers for prediction and prognosis of muscle-invasive bladder cancer. *Oncotarget*. 2016;7(24):36733.
73. Tribollet V, Barenton B, Kroiss A, Vincent S, Zhang L, Forcet C, Cerutti C, Perian S, Allili N, Samarut J, et al. mir-135a inhibits the invasion of cancer cells via suppression of α err. *PLoS ONE*. 2016;11(5):0156445.

74. Zhao Y, Ling Z, Hao Y, Pang X, Han X, Califano JA, Shan L, Gu X. Mir-124 acts as a tumor suppressor by inhibiting the expression of sphingosine kinase 1 and its downstream signaling in head and neck squamous cell carcinoma. *Oncotarget*. 2017;8(15):25005.
75. Cai QQ, Dong YW, Wang R, Qi B, Guo JX, Pan J, Liu YY, Zhang CY, Wu XZ. Mir-124 inhibits the migration and invasion of human hepatocellular carcinoma cells by suppressing integrin α v expression. *Sci Rep*. 2017;7:40733.
76. Wang Y, Chen L, Wu Z, Wang M, Jin F, Wang N, Hu X, Liu Z, Zhang C-Y, Zen K, et al. mir-124-3p functions as a tumor suppressor in breast cancer by targeting cbl. *BMC Cancer*. 2016;16(1):826.
77. Pan T, Chen W, Yuan X, Shen J, Qin C, Wang L. mir-944 inhibits metastasis of gastric cancer by preventing the epithelial–mesenchymal transition via macc1/met/akt signaling. *FEBS Open Bio*. 2017;7(7):905–14.
78. Wen L, Li Y, Jiang Z, Zhang Y, Yang B, Han F. mir-944 inhibits cell migration and invasion by targeting macc1 in colorectal cancer. *Oncol Rep*. 2017;37(6):3415–22.
79. He Z, Xu H, Meng Y, Kuang Y. mir-944 acts as a prognostic marker and promotes the tumor progression in endometrial cancer. *Biomed Pharmacother*. 2017;88:902–10.
80. Dhawan A, Barberis A, Cheng W-C, Domingo E, West C, Maughan T, Scott J, Harris AL, Buffa FM. sigQC: A procedural approach for standardising the evaluation of gene signatures. <https://doi.org/10.1101/203729>. <https://www.biorxiv.org/content/10.1101/203729v2>.
81. Catalanotto C, Cogoni C, Zardo G. MicroRNA in control of gene expression: an overview of nuclear functions. *Int J Mol Sci*. 2016;17(10):1712.
82. Muniyappa M, Dowling P, Henry M, Meleady P, Doolan P, Gammell P, Clynes M, Barron N. MiRNA-29a regulates the expression of numerous proteins and reduces the invasiveness and proliferation of human carcinoma cell lines. *Eur J Cancer*. 2009;45(17):3104–18.
83. Lamberti M, Capasso R, Lombardi A, Di Domenico M, Fiorelli A, Feola A, Perna AF, Santini M, Caraglia M, Ingrosso D. Two different serum miRNA signatures correlate with the clinical outcome and histological subtype in pleural malignant mesothelioma patients. *PLoS ONE*. 2015;10(8):0135331.
84. Sathipati SY, Ho S-Y. Identifying the miRNA signature associated with survival time in patients with lung adenocarcinoma using miRNA expression profiles. *Sci Rep*. 2017;7(1):7507.
85. Saeys Y, Abeel T, Van de Peer Y. Robust feature selection using ensemble feature selection techniques. In: *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, vol. 5212. Springer LINK; 2008. p. 313–25.
86. Rincon AL, Tonda A, Elati M, Schwander O, Piwowarski B, Gallinari P. Evolutionary optimization of convolutional neural networks for cancer miRNA biomarkers classification. *Appl Soft Comput*. 2018. <https://doi.org/10.1016/j.asoc.2017.12.036>.
87. Breiman L. Pasting small votes for classification in large databases and on-line. *Mach Learn*. 1999;36(1-2):85–103.
88. Friedman JH. Greedy function approximation: a gradient boosting machine. *Ann Stat*. 2001;29(5):1189–232.
89. Cox DR. The regression analysis of binary sequences. *J R Stat Soc Ser B Methodol*. 1958;20(2):215–32.
90. Crammer K, Dekel O, Keshet J, Shalev-Shwartz S, Singer Y. Online passive-aggressive algorithms. *J Mach Learn Res*. 2006;7(Mar):551–85.
91. Breiman L. Random forests. *Mach Learn*. 2001;45(1):5–32.
92. Tikhonov AN. On the stability of inverse problems. *Cr Acad Sci Urss*. 1943;39:195–8. Downloaded from Science Open.com.
93. Zhang T. Solving large scale linear prediction problems using stochastic gradient descent algorithms. In: *Proceedings of the Twenty-first International Conference on Machine Learning*. New York: ACM; 2004. p. 116.
94. Hearst MA, Dumais ST, Osman E, Platt J, Scholkopf B. Support vector machines. *IEEE Intell Syst Appl*. 1998;13(4):18–28.
95. Breiman L, Friedman J, Stone CJ, Olshen RA. *Classification and Regression Trees*: Chapman and Hall/ CRC press; 1984, p. 368.
96. Leshkowitz D, Horn-Saban S, Parmet Y, Feldmesser E. Differences in microRNA detection levels are technology and sequence dependent. *RNA*. 2013;19(4):527–38.
97. Del Vecovo V, Meier T, Inga A, Denti MA, Borlak J. A cross-platform comparison of affymetrix and agilent microarrays reveals discordant miRNA expression in lung tumors of c-raf transgenic mice. *PLoS ONE*. 2013;8(11):78870.
98. Bassani N, Ambrogi F, Biganzoli E. Assessing agreement between miRNA microarray platforms. *Microarrays*. 2014;3(4):302–21.
99. Chu A, Robertson G, Brooks D, Mungall AJ, Biorl I, Coope R, Ma Y, Jones S, Marra MA. Large-scale profiling of microRNAs for the cancer genome atlas. *Nucleic Acids Res*. 2015;44(1):3.
100. Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, Speed TP. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*. 2003;4(2):249–64.
101. Cheadle C, Vawter MP, Freed WJ, Becker KG. Analysis of microarray data using z score transformation. *J Mol Diagn*. 2003;5(2):73–81.
102. Hansen N, Müller SD, Koumoutsakos P. Reducing the time complexity of the derandomized evolution strategy with covariance matrix adaptation (cma-es). *Evol Comput*. 2003;11(1):1–18.

Publisher's Note

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

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions

