Bioinformatics, 34(3), 2018, 398–406
doi: 10.1093/bioinformatics/btx622
Advance Access Publication Date: 3 October 2017
Original Paper

Genome analysis

Tumor origin detection with tissue-specific miRNA and DNA methylation markers

Wei Tang[1,2], Shixiang Wan[1], Zhen Yang[3], Andrew E. Teschendorff[3,4,*] and Quan Zou[1,*]

[1]School of Computer Science and Technology, [2]Department of Biological Engineering, School of Chemical Engineering, Tianjin University, Tianjin 300050, China, [3]Key Laboratory of Computational Biology, CAS-MPG Partner Institute for Computational Biology, Shanghai 200031, China and [4]Statistical Cancer Genomics, Paul O'Gorman Building, UCL Cancer Institute, University College London, London WC1E 6BT, UK

*To whom correspondence should be addressed.

Associate Editor: Cenk Sahinalp

Received on January 6, 2017; revised on September 20, 2017; editorial decision on September 22, 2017; accepted on September 27, 2017

Abstract

Motivation: A clear identification of the primary site of tumor is of great importance to the next targeted site-specific treatments and could efficiently improve patient's overall survival. Even though many classifiers based on gene expression had been proposed to predict the tumor primary, only a few studies focus on using DNA methylation (DNAm) profiles to develop classifiers, and none of them compares the performance of classifiers based on different profiles.
Results: We introduced novel selection strategies to identify highly tissue-specific CpG sites and then used the random forest approach to construct the classifiers to predict the origin of tumors. We also compared the prediction performance by applying similar strategy on miRNA expression profiles. Our analysis indicated that these classifiers had an accuracy of 96.05% (Maximum–Relevance–Maximum–Distance: 90.02–99.99%) or 95.31% (principal component analysis: 79.82–99.91%) on independent DNAm datasets, and an overall accuracy of 91.30% (range 79.33–98.74%) on independent miRNA test sets for predicting tumor origin. This suggests that our feature selection methods are very effective to identify tissue-specific biomarkers and the classifiers we developed can efficiently predict the origin of tumors. We also developed a user-friendly webserver that helps users to predict the tumor origin by uploading miRNA expression or DNAm profile of their interests.
Availability and implementation: The webserver, and relative data, code are accessible at http://server.malab.cn/MMCOP/.
Contact: zouquan@nclab.net or a.teschendorff@ucl.ac.uk
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

The primary site of cancers can remain ambiguous, or fail to be identified even after thorough physical examinations, such as full blood count, biochemistry and histological evaluation of biopsy material involving immunohistochemistry (Kwak *et al.*, 2010). Patients diagnosed with these cancers with an elusive origin site are often associated with a very low median survival time of 9–12 months in average after diagnosis (Daugaard *et al.*, 2009). Although the survival rate depends on various factors such as cancer cell type, location, cancer spread, treatment received and response to treatment, the high mortality of patients mainly reflects the misclassification tumors (Monzon *et al.*, 2010; Søkilde *et al.*, 2014). Indeed, many cases remain undiagnosed or mistakenly diagnosed, so therapy cannot be matched to the specific disease. This is particularly relevant for cancers that respond well to specific chemotherapies or hormone drugs. Therefore, it is of importance to accurately identify the primary origins of these tumor samples for ensuring a subsequent efficient specific treatment.

The genetic expression of underlie cancer pathogenesis are rapidly being studied, which provides novel insights in tumor biology as well as in potential cancer biomarkers. Recent studies have also demonstrated that a comprehensive, enterprise-wide effort to map the genetic alterations of patients' tumors is feasible, which could provide important information for timely patient care, and may help shape the future of cancer therapy (Heinzelmann *et al.*, 2011; Kurahashi *et al.*, 2013; Zehir *et al.*, 2017). Tumor origin classification based on gene expression has been much proposed as a clinical application to predict the primary origin of cancers (Budhu 2008; Heinzelmann *et al.*, 2011; Rosenfeld *et al.*, 2008).

The expression profiles of micro (mi)RNAs which are small noncoding RNAs (Bartel 2009; Wang *et al.*, 2016) that regulate the expression of genes involved in biological processes such as cell proliferation, death and differentiation (Bartel 2009; Hayashita *et al.*, 2005; Hwang and Mendell 2006; Meng *et al.*, 2016) also had been used to identify primary sites of cancers. For example, Rolf *et al.* (Søkilde *et al.*, 2014) developed an miRNA-based classifier involving feature selection embedded in the Least Absolute Shrinkage and Selection Operator (LASSO) classification algorithm. This classifier demonstrated a high overall accuracy of 88% (confidence interval (CI) 75–94%) at predicting the origin site of cancers. Unfortunately, stomach and esophagus samples could not be separated by this classifier, one of the reasons could be that the histologies of these two tissues are pretty similar (Søkilde *et al.*, 2014) and gastroesophageal junction adenocarcinomas were also similar to samples of stomach cancer. Thus, some studies also suggested the 'stomach' class includes both stomach cancers and gastroesophageal junction adenocarcinomas (Rosenfeld *et al.*, 2008). Besides, the cost to maintain the state of preservation the examined biological tissues urges the development of new biomarkers for identifying cancer sites.

As an important class of regulatory mechanism, DNA methylation (DNAm) is also central to numerous biological processes, such as regulating gene expression (e.g. embryonic development, X-chromosome inactivation, genomic imprinting and preservation of chromosome stability). Given such many cellular processes in which the DNAm could involve, it is not surprising that the abnormal methylation may result in devastating consequences, such as common human disease (e.g. cancers, neurodevelopmental and degenerative disorders, autoimmune diseases) also highly tissue-specific, thus could be helpful in the detection and prediction of tumors' origin. A recent study has demonstrated that DNAm profiles can also accurately determine the occult original site of cancers (Moran *et al.*, 2016). Another recent research proposed a probabilistic method named CancerLocator has also achieved promising results on determining the presence and predicting location of tumors for several tissues with exploiting methylation profiles of cell-free DNA (Kang *et al.*, 2017). Thus, predicting the origin of tumors by means of DNAm profiles will become a new trend and additional tool to help predict the tumor origin.

In this study, we firstly adopted novel feature selection strategies, which incorporating two levels to identify tissue-specific DNAm of CpG sites. Then a random forest algorithm was used to construct the classifiers which can identify the site of tumor origin with high specificity on the basis of the DNAm profile of the cancers. Another novelty of this study is that a similar prediction pipeline was also applied on the miRNA expression profiles to evaluate the performance difference between these two profile types. We select a large number of datasets from The Cancer Genome Atlas (TCGA) with a total of 5379 DNAm profiles and 6602 miRNA expression profiles to develop the classifiers, representing 14 commonly recognized sites of origin in the differential diagnosis of cancers, respectively. Our classifiers based on Illumina 450K DNAm profiles and miRNA expression are available through the Methylation and MiRNA Cancer Origin Predictor (MMCOP: http://server.malab.cn/MMCOP/) webserver, which enables researchers to predict the origin site of tumor samples of their interests. We also tested seven DNAm datasets in GEO to further validate the performance of our algorithm and webserver.

2 Materials and methods

2.1 Flowchart and data collection

Figure 1 is a flowchart of this study, including the algorithm flow. We randomly split the selected datasets from TCGA into two groups, one group was used as the training sets for the feature selection and classifier construction; another group, along with seven datasets in GEO, were used as the independent testing datasets to evaluate the performance of our classifiers and webserver.

For data quality control, we conducted a strict review of each dataset to only select those data that met the requirements of our study. One of the inclusion criteria for subsequent feature selection of miRNAs and DNAm CpGs is that the datasets should have a sufficient number of samples for both case and control (≥ 5) groups. Therefore, a total of 6602 samples for miRNA-based profiles, including 6045 tumor samples, and a total of 5379 samples for DNAm-based profiles, including 4668 tumor samples, were collected through The Cancer Genome Atlas pilot (TCGA) project (https://tcga-data.nci.nih.gov/tcga/). The samples based on miRNA expression profiles were sequenced by the BCGSC (IlluminaHiSeq_miRNAseq) sequencing platform, which enables highly sensitive and specific detection of common human miRNAs. The DNAm-based profiles samples were obtained from the Infinium

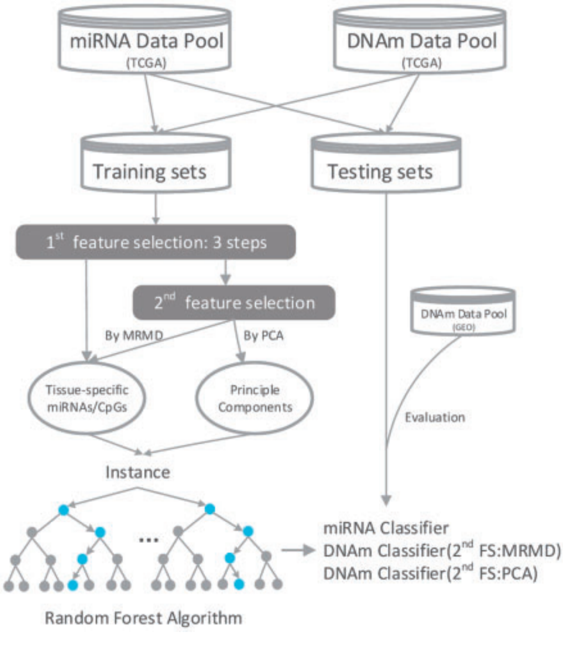

Fig. 1. Schematic overview of the workflow of data analysis, and the development of three classifiers. Note: PCA, principal component analysis; MRMD, maximum–relevance–maximum–distance; DNAm, DNA methylation; FS, feature selection

Table 1. Number of samples per tissue for miRNAs expression and DNA methylation profiles, training sets, testing sets

		miR-based datasets(n)						DNAm-based datasets(n)					
Primary Site	Histogy	NT	TN	Total tissue samples	Sets used in FS	Sets used in training RF	Testing sets	NT	TN	Total tissue samples	Sets used in FS	Sets used in training RF	Testing sets
Bladder	BLCA(Bladder Urothelial Carcinoma)	19	406	425	222	203	203	19	201	220	119	100	101
Breast	BRCA(Breast Invasive Carcinoma)	90	1066	1156	623	533	533	81	652	733	407	326	326
Bile Duct	CHOL(Cholangiocarcinoma)	9	36	45	27	18	18	*	*	*	*	*	*
Colorectal	COAD(Colon Adenocarcinoma)	8	441	449	228	220	221	81	652	733	407	326	326
Esophagus	ESCA(Esophageal Carcinoma)	13	173	186	99	86	87	15	126	141	78	63	63
Head and Neck	HNSC (Head and Neck Squamous Cell Carcinoma)	44	518	562	303	259	259	45	405	450	247	202	203
Kidney Chromophobe	KICH (Kidney Chromophobe)	25	66	91	58	33	33	*	*	*	*	*	*
Kidney Renal Clear Cell	KIRC (Kidney Renal Clear Cell Carcinoma)	71	512	583	327	256	256	160	299	459	310	150	149
Kidney Renal Papillary Cell	KIRP (Kidney Renal Papillary Cell Carcinoma)	*	*	*	*	*	*	45	196	241	143	98	98
Liver	LIHC(Liver Hepatocellular Carcinoma)	51	370	421	236	185	185	47	176	223	135	88	88
Lung(Lung Squamous Cell)	LUSC (Lung Squamous Cell Carcinoma)	*	*	*	*	*	*	41	275	316	179	138	137
Lung	LUAD (Lung Adenocarcinoma)	45	510	555	300	255	255	32	399	431	232	200	199
Pancreas	PAAD (Pancreatic Adenocarcinoma)	*	*	*	*	*	*	10	146	156	83	73	73
Prostate	PRAD (Prostate Adenocarcinoma)	51	480	531	291	240	240	48	278	326	187	139	139
Stomach	STAD (Stomach Adenocarcinoma)	41	434	475	258	217	217	*	*	*	*	*	*
Thyroid	THCA (Thyroid Carcinoma)	57	501	558	307	250	251	53	489	542	297	244	245
Uterus	UCEC (Uterine Corpus Endometrial Carcinoma)	33	532	565	299	266	266	34	374	408	221	187	187
Total		557	6045	6602	3578	3021	3024	711	4668	5379	3045	2334	2334

Note: miR, miRNAs; DNAm, DNA methylation; NT, normal tissue sample; TN, tumor tissue sample; *This tissue has no corresponding dataset

HumanMethylation450 platform, which allows for the assessment of methylation status of more than 480 000 cytosines distributed over the entire genome in 12 samples in parallel (Dedeurwaerder *et al.*, 2011). MiRNA expression and DNAm comprised 14 clinically relevant histologies, covering a broad selection of solid tumors. The details of all tissue samples, including tumor status and histopathologic details used for constructing the classifier are provided in Table 1. To avoid the potential overfitting issue, we divided the tumor samples of each tissue equally into two groups at random, one group, along with the normal samples, were used to conduct feature selection, and then the selected features will be used for classifier training on the tumor samples, while another was used to as the totally independent dataset to test the classifier (Table 1).

2.2 Data preprocessing and normalization

The preprocessing analysis of datasets was performed with the Linear Models for Microarray and RNA-seq Data package (Limma) http://www.bioconductor.org/packages/release/bioc/html/limma.html) (Ritchie *et al.*, 2015), embedded in the R environment (http://www.r-project.org/). For miRNA-based datasets, we selected miRNA isoform expression data because all isomiRs are from a specific miRNA locus and provide information about mature miRNA expression. The maximum miRNA expression value was selected if there were multiple isoforms for a given miRNA in each sample. For each tissue type of the selected dataset, we removed miRNAs or CpGs with more than 30% missing sample values (NA). The remaining missing values were imputed using the impute.knn function. The miRNA expression values were logarithmically transformed with base 2 and quantile normalized. For DNAm datasets, the absolute methylation values representing the methylated intensity of every CpG were calculated using BMIQ_1.4 (Beta MIxture Quantile dilation) (Teschendorff *et al.*, 2013) to correct the type II probe bias.

2.3 Feature selection and classifier construction

2.3.1 First-level feature selection

For miRNA- and DNAm-based samples, first-level feature selection was conducted by Limma to identify tissue-specific miRNAs/DNAms and to reduce the considerable redundancy of original data. Three different steps of analysis were conducted to select miRNAs/CpGs showing: (1) differential expression/methylation values in a given normal tissue compared with other normal tissue types [one versus all, threshold: $P \leq 0.05$ (miRNAs), $P \leq 0.01$ (DNAms)]; (2) no different levels of the value in a given cancer tissue compared with corresponding normal tissues [one versus all, threshold: $P \geq 0.3$ (miRNAs), $P \geq 0.5$ (DNAms)]; and (3) differential expression/methylation values for the corresponding cancer type compared with other tumor tissues [one versus all, threshold: $P \leq 0.05$ (miRNAs), $P \leq 0.01$ (DNAms)].

Table 2. The number of selected miRNAs and CpGs from the feature selection and the performance of three classifiers for the miRNA expression and DNA methylation profiles

Primary site	No. of first FS miRNAs	miRNA-based classifier		No. of first FS CpGs	No. of second FS CpGs (by MRMD)	No. of second FS components (by PCA)	DNAm-based (by MRMD) classifier		DNAm-based (by PCA) classifier	
		Training_ Acc (%)	Testing_ Acc (%)				Training_ Acc (%)	Testing_ Acc (%)	Training_ Acc (%)	Testing_ Acc (%)
Bladder	23	90.15	87.47	430	267	428	96.50	95.89	91.50	91.00
Breast	25	92.21	91.83	1567	971	573	98.47	99.40	96.17	98.16
Bile Duct	11	86.11	79.33	—[a]	—[a]	—[a]	—[a]	—[a]	—[a]	—[a]
Colorectal	13	84.32	83.40	6780	9	998	99.99	99.99	98.31	99.66
Esophagus	16	80.81	81.45	453	72	451	96.03	90.02	72.22	79.82
Head and neck	26	93.44	94.01	1029	467	588	97.28	93.27	92.57	93.06
Kidney chromophobe	28	98.48	98.74	—[a]	—[a]	—[a]	—[a]	—[a]	—[a]	—[a]
Kidney renal clear cell	31	92.97	97.16	2716	334	151	97.67	95.59	97	98.67
Kidney renal papillary cell	—[a]	—[a]	—[a]	1199	341	378	96.94	96.36	90.82	98.89
Liver	36	97.84	98.41	1489	666	448	98.30	99.74	94.32	98.93
Lung (lung squamous cell)	—[a]	—[a]	—[a]	535	366	534	86.96	83.08	86.23	87.23
Lung	25	91.18	91.37	639	309	638	96.50	95.24	92.75	97.26
Pancreas	—[a]	—[a]	—[a]	2342	738	167	93.15	97.64	96.58	97.99
Prostate	22	98.13	97.62	1742	672	467	99.28	99.99	96.76	99.91
Stomach	19	87.33	88.76	—[a]	—[a]	—[a]	—[a]	—[a]	—[a]	—[a]
Thyroid	32	98.20	98.45	4115	127	99	99.80	99.83	99.99	99.83
Uterus	20	88.53	90.15	742	520	405	97.59	98.63	94.39	93.96
Average accuracy		91.41	91.30				96.75	96.05	92.83	95.31

Note: FS, feature selection; Acc, accuracy.

[a] This tissue has no corresponding datasets; the details of CpGs selected by the MRMD are available in the Supplementary Material.

Due to the much higher dimensionality of DNAm profiles than that of the miRNA expression profiles, we set a much stricter threshold for DNAm profiles to efficiently reduce the data redundancy. The first step aims to select those miRNAs or CpGs whose mean values were significantly different between a given normal tissue type and other normal tissue types. Because the goal of this study is to predict tumor original sites, the selection of biomarkers with differential expression among different normal tissues is necessary to identify tissue-specific miRNAs or CpGs. Those miRNAs/CpGs with false discovery rate (FDR)-adjusted P-values $\leq 0.05/0.01$ were extracted as candidates showing significant differential expression or methylation value. For dataset of a given tissue including both tumor samples and corresponding normal samples, we also need to ensure that the tissue-specific biomarkers did not show differential expression or different methylated values among the cancer and normal samples of the same tissue. The second step confirmed this. The thresholds were set as 0.3 and 0.5, respectively, so miRNAs/CpGs with FDR-adjusted P-values $\geq 0.3/0.5$ were considered to have normal miRNA expression or DNAm. The third step identified miRNAs or CpGs showing differential expression/methylation among different tumor tissue types to discriminate cancer types from each other. The thresholds were set as the same as the first step. The main concern that we set the threshold for the second step as 0.3 and 0.5 for miRNA expression profiles and DNAm profiles, respectively, is for decreasing the amount of calculation, making the developed webserver friendlier and also trying to avoid potential overfitting issue. Furthermore, for DNAm profiles, the subsequent second-level feature selection will examine each CpGs selected from first-level feature selection and then find out the optimal number of CpGs which will be regarded as features.

Together, we referred to these three steps to identify tissue-specific biomarkers as first-level feature selection. The intersection of miRNAs and CpGs selected using these three steps was regarded as preliminary features of first-level feature selection.

2.3.2 Second-level feature selection

The number of miRNAs selected from the first-level feature selection ranged from 11 to 32 (Table 2 and Supplementary Material; Supplementary Table S1). Our subsequent miRNA-based classifier demonstrated that these selected miRNAs were tissue-specific biomarkers sufficiently capable of predicting the tumor origin with a high level of accuracy. It was therefore not necessary to apply second-level feature selection to the miRNAs selected in the first level. However, DNAm profiles included more than 400 000 CpGs even after the preprocessing procedure. Consequently, those CpGs selected by first-level feature selection were still large in number and redundant, ranging from 430 to 6780 (Table 2 and Supplementary Material; Supplementary Table S2) for each tissue. Therefore, we proposed a second-level feature selection to further identify tissue-specific CpGs.

Here, we present two methods for the second-level feature selection. One of the second-level feature selection methods called Maximum–Relevance–Maximum–Distance (MRMD, http://lab.malab.cn/soft/MRMD/index_en.html) (Zou *et al.*, 2016) which selects features with strong correlations and lowest redundancy features. In MRMD, Pearson's correlation coefficient is used to measure the relevance and Euclidean distance is used to calculate the redundancy. Pearson's correlation coefficient showed the close relationship between features and labels, while the distance between features was used to present data redundancy. With the increasing of

Pearson's correlation coefficient, the relevance between feature and target class will become higher. The larger the distance of feature is, the lower the redundancy of sub-feature set will become. The feature with large sum of relevance and distance would be selected into the ultimate sub-feature set. Finally, the sub-feature set that selected by MRMD will have lowest redundancy and strongest relevance with target class. We used MRMD to find out the optimal number of features. First, MRMD will rank all the feature candidates according to the calculated Pearson's correlation coefficient and Euclidean distance, then MRMD will use top-ranked features to construct a simple classifier to evaluate the classification accuracy. After all the feature candidates were ranked and the accuracy were computed, a top-ranked feature list with highest accuracy will be selected out as the final features.

Another feature selection method named principal component analysis (PCA, embedded in the Dimensionality Reduction part of scikit-learn, http://scikit-learn.org/stable/index.html) (Pedregosa *et al.*, 2011) was also used to conduct the second-level feature selection. PCA is a statistical procedure that uses orthogonal transformation to obtain a set of linearly uncorrelated variables, principal components, from observations of possibly correlated variables. We selected the principle components according to cumulative percentage of total variation. In general, let $\lambda_1, \lambda_2, \ldots, \lambda_n$ be the eigenvalues of Σ (sorted in decreasing order), so that λ_j is the eigenvalue corresponding to the eigenvector u_j. Then if we retain k principal components, the percentage of variance retained is given by:

$$\frac{\sum_{j=1}^{k} \lambda_j}{\sum_{j=1}^{n} \lambda_j}$$

Here, we selected the number of principle components which retains the cumulative percentage of total variation more than 95%. 95% is also a commonly used threshold in determining the number of selected components in PCA (Bro and Smilde, 2014; Hirsch, 2016).

2.3.3 Classifier construction

MiRNAs selected from first-level feature selection and CpGs selected from second-level feature selection were used as tissue-specific biomarkers of each class. All classes were combined for the further construction of a random forest model. Because the selection of relevant biomarkers (e.g. genes, miRNAs, CpGs) for sample classification (e.g. to differentiate between patients with and without cancer) is common to most genomics studies, another main objective of this study was the identification of small biomarker sets that could be used for clinical diagnostic purposes. This would require the possible smallest set of biomarkers capable of achieving a high prediction performance, thus excluding 'redundant' biomarkers (Díaz-Uriarte and De Andres, 2006). Considering the unique characteristics of this research and the properties of genomics data, classification algorithms suitable for both two-class and multi-class problems, or when the number of variables exceeds that of observations, and those that avoid overfitting would be of great interest. Random forest is one such algorithm that has been shown to have a high performance in many classification cases base on gene expression microarray (Breiman, 2001; Statnikov *et al.*, 2008). We therefore adopted random forest after feature selection step for miRNAs and CpGs. For a more comprehensive evaluation on the algorithm, we also compare the random forest with other two benchmark classifiers (SVM and KNN).

Because the number of minority class samples (a given tissue class) were very small compared with that of majority class samples (other tissue classes), this would cause imbalance problem. To address this problem, which may seriously impact on classifier performance, we adopted an under-sampling (Al-Shahib *et al.*, 2005) method to randomly sample a subset from the majority class to form a balanced dataset with the corresponding minority class. Each tissue and individual model, with a balanced dataset, was trained to discriminate a given tissue from all other tissues (one versus all). For example, in the case of miRNA expression profiles, we had a total of 203 bladder urothelial carcinomas, and a total of 2818 other tissue samples. We therefore randomly selected 203 samples from the 2818 to construct a balanced dataset by combining with the 203 bladder samples (see columns 7 and 12 in Table 1). Each individual model was trained with a 5-fold cross-validation.

We also developed a Java-based MMCOP webserver to enable users to predict tumor origin sites by uploading miRNA expression data or DNAm profiling data. This webserver supports miRNA-based classifier and DNAm-based classifier (second-level feature selection is MRMD). Our web-server also supports the prediction of the datasets which contain not many missing miRNAs or CpGs. For the further validation of performance of webserver, we also used several DNAm datasets in GEO (GSE67116, GSE85845, GSE69914, GSE38268, GSE61446, GSE49149 and GSE45187) to test the webserver (Supplementary Material; Supplementary Table S6).

3 Results

3.1 Sample selection

To determine which tissue should be included to construct the classifier, we focused on those cancers most commonly detected by light microscopy. Most of these samples (~90%) are adenocarcinomas, with ~60% moderately to well differentiated, and ~30% poorly differentiated. Common adenocarcinoma origins include the lung, pancreas, breast, prostate, stomach, liver and colon. The remaining 10% of these samples are squamous cell carcinomas, mostly arising from head and neck tumors, which are often poorly or even undifferentiated (Greco and Hainsworth, 2006). To ensure a comprehensive representation of major carcinoma types defined by their anatomic tissue or organ of origin, we selected several major carcinomas (bladder, breast, colon, lung, stomach, kidney, liver and uterus tumors). Thus, for miRNA-based and DNAm-based samples, we respectively selected 6602 miRNA samples and 5379 DNAm samples for 14 tissue types covering most cancer types. Table 1 lists the 14 tissues and histologies.

3.2 Feature selection and tissue-specific miRNAs expression and CpGs methylation

The key to constructing a classifier that performs well at predicting the tumor origin site is to use true tissue-specific features. We therefore adopted different strategies for the selection of highly tissue-specific biomarkers from different dataset types. Another consideration of feature selection was the feature size of different datasets. For miRNA-based datasets, with only ~1800 common miRNAs (only 419 miRNAs remained after data preprocessing), we adopted a one-level feature selection that not only ensured the optimal identification of tissue-specific miRNAs, but also selected appropriate amounts of miRNAs (~11–32) for each tissue (Table 2). However, because DNAm datasets covered the entire genome, we further adopted a second feature selection (MRMD and PCA) to filter out redundant CpGs. The optimal number of features selected out for each tissue was determined by the automated searching model of MRMD and PCA, since more complex models

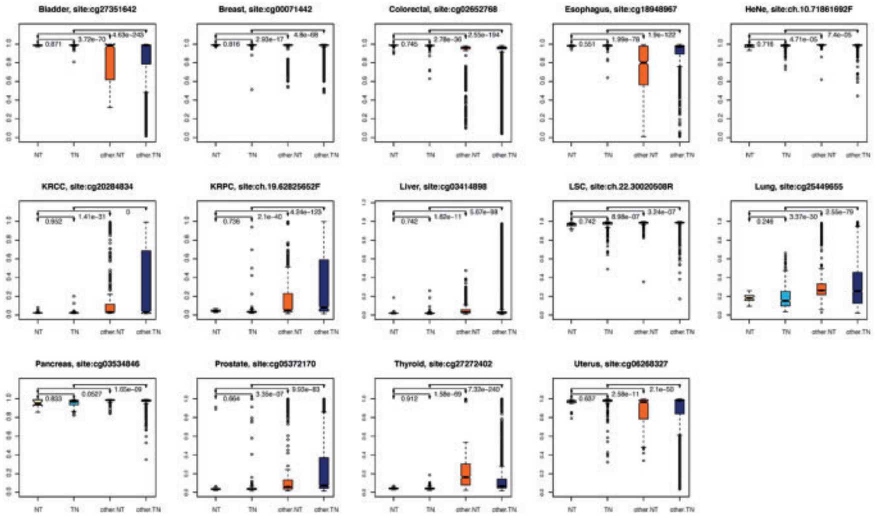

Fig. 2. Boxplot of beta-value of 14 top tissue-specific CpGs of 14 histologies in the training set. Since the boxplot is showing the comparison of a single CpG site, the *P*-value between two boxes was calculated by the *t*-test. *Note*: HeNe, head and neck; KRCC, kidney renal clear cell; KRPC, kidney renal papillary cell; LSC, lung squamous cell; NT, normal tissue sample; TN, tumor tissue sample

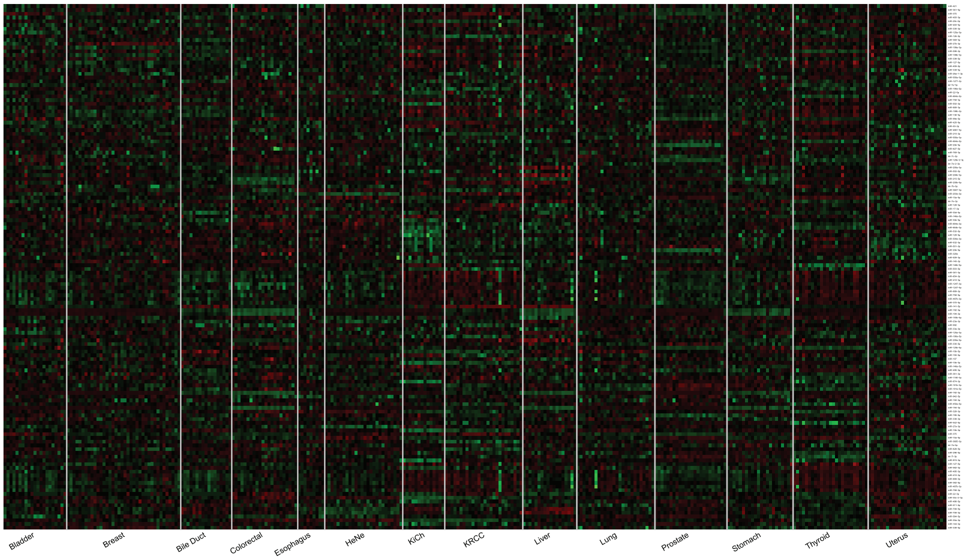

Fig. 3. Heatmap of expression of tumor tissue-specific miRNAs (rows) across 400 samples (columns) that represent the 14 histologies in the training set. These 400 samples were minified in equal proportion from the total 3578 samples. The heatmap shows median normalized log2 data for every miRNA selected in each tissue. *Note*: HeNe, head and neck; KiCh, kidney chromophobe; KRCC, kidney renal clear cell

would have included additional features with no corresponding increase in classifier performance. The selected miRNAs with high tissue-specific discriminatory potential from the first feature selection are listed in Table 2, which also includes the number of automated searching features from both feature selection levels. Detailed information of CpGs selected after the second feature selection is available in the Supplementary Material (Supplementary Table S2).

To verify the rationality of our feature selection method, we constructed a box plot (Fig. 2) for the top CpGs of 14 DNAm tissues and a heat-map (Fig. 3) of selected tissue-specific miRNAs for the miRNA-based profiles. These two figures show that some tissues are easy to distinguish from others because of their strong and differentially methylated CpG or differentially expressed tissue-specific miRNA signatures.

3.3 Classifier performance evaluation

Indeed, the performance of a predictor depends on the quality and the number of selected features. In this case, the optimal number of tissue-specific miRNAs was obtained from the first feature selection, while the best performance of the DNAm-based classifier was obtained using CpGs selected by the automated searching model of MRMD and PCA. A random forest algorithm of this balanced dataset was then used to train the classifier with the optimal selection of tissue-specific biomarkers, thus generating an individual model. For a more comprehensive evaluation on the classification, we also compared the random forest method with other two benchmark classifiers (SVM and KNN). After examining all possible values of each method's hyper parameters, we report only the best prediction results for each of three classifiers (random forest, SVM and KNN). All the training went through a 5-fold cross-validation phase. The

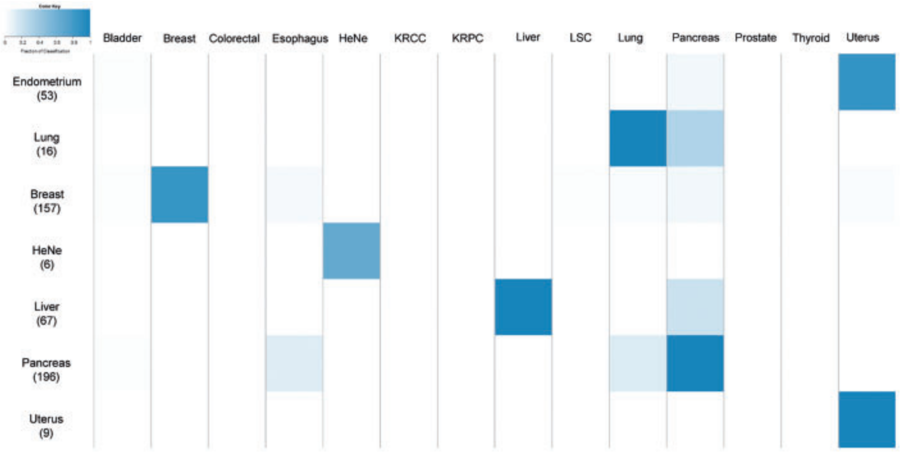

Fig. 4. The heatmap of performance for GEO datasets that are used for validation of the webserver. *Note*: HeNe, head and neck; KRCC, kidney renal clear cell; KRPC, kidney renal papillary cell; LSC, lung squamous cell

results are available in the Supplementary Material; Supplementary Tables S3–S5. We also added false positives (fp) and false negatives (fn) in the attachment tables to offer more useful information on cancer diagnosis. Both for miRNA expression profiles and DNAm profiles, the results showed that the random forest classifier outperforms other two classifiers (SVM and KNN). The analysis below is all based on the random forest classifier.

The 5-fold cross-validation training accuracy and the testing accuracy are shown in Table 2. Those tissues of origin correctly predicted by either a miRNA-based or DNAm-based classifier account for the majority in all cases, with overall testing accuracies of 91.30% (miRNA-based, CI 79.33–98.74%), 96.05% (DNAm-based by MRMD, CI 83.08–99.99%) and 95.31% (DNAm-based by PCA, CI 79.82–99.83%) (Table 2).

In terms of the biology bias, the first-level feature selection is largely based on miRNA differential expression and DNA differential methylation analysis. However, the second-level feature selection, either by MRMD or PCA, is mainly based on the algorithm from the mathematical meaning. In fact, the datasets we used for detailed analysis represent two main data regimes: the first is that the dimension (D) of data is smaller than the sample size (n) (miRNA expression profiles: $D = 419$, $n = 3578$); the second is that the high-dimensional dataset with an ambient dimension (D) may be the same or substantially larger than the sample size (n) (DNAm profiles: $D = 395\,515$, $n = 3045$). For the first data regime ($D < n$), many machine learning-based classifiers have been developed to predict the tumor origin. Our algorithm, which combines first-level feature selection (miRNA differential expression analysis) and random forest to construct the miRNA-based classifier, has a similar sensitivity to other cancer classification methods. Our miRNA-based classifier has a high prediction accuracy 91.41% (CI 80.81–98.48%) by 5-fold cross-validation of initial training datasets, and a high testing accuracy of 91.30% (CI 79.33–98.74%). The classifier based on the LASSO algorithm proposed by Rolf *et al.* (Søkilde *et al.*, 2014) had a relatively lower overall prediction accuracy (88% accuracy; CI 75–94%) on 15 tissues, particularly for the colorectal tissue, the LASSO classifier had an overall accuracy of 76.47%, while our miRNA-based classifier had a higher performance on predicting this tissue (83.40%) (Table 2). Similarly, our miRNA-based classifier had an accuracy of 87.47% for predicting bladder tissue, while the *K* nearest neighbor-based miRNA classifier reported by Rosenfeld *et al.* (2008) had zero sensitivity to bladder cancer. However, there's one tissue (bile duct) which is inherently difficult to classify correctly. The testing accuracy for this tissue using our miRNA-based classifier was 79.33%. This is lower than accuracies of other tissues but superior to the immunohistochemistry marker-based method of Park *et al.* (2007) which identified cholangiocarcinoma (bile duct) with an accuracy of 28.00%, and other machine-learning methods, such as the method proposed by Rolf et al (Søkilde *et al.*, 2014) which had also a lower performance (78.00%) on predicting bile duct. All of these had demonstrated our miRNA-based classifier had a similar or higher prediction performance when compared with other reported machine-learning methods.

The second data regime ($D \geq n$) has been observed with the rapid development of data collection technology, enabling more observations to be collected (larger n), and more variables to be measured, such as the dimensions (larger D) (Negahban *et al.*, 2009). One example of this data regime is DNAm intensity data collected from Infinium 450K array. To process these data more efficiently, we adopted two widely-used methods (MRMD and PCA) as second-level feature selection to identify tissue-specific CpGs from first-level selection (differential methylation analysis). Our two DNAm-based classifiers were shown to have much higher accuracy levels than the miRNA-based classifier in some tissues (Table 2). The 5-fold cross-validation of initial training datasets achieved accuracy levels of 96.75% (DNAm-based by MRMD, CI 86.96–99.99%) and 92.83% (DNAm-based by PCA, CI 72.22–99.99%) and a very high testing accuracy of 96.05% (DNAm-based by MRMD, CI 83.08–99.99%) and 95.31% (DNAm-based by PCA, CI 79.82–99.83%) on the independent datasets (Table 2). Indeed, in these two DNAm-based classifiers, most tissues had quite high testing accuracy (>90%), including those which are difficult to be identified by miRNA expression profiling, such as uterus (98.63% in MRMD and 93.96% in PCA), and bladder (95.89% in MRMD and 91.00% in PCA). Particularly, the prediction accuracy of colorectal tissue by DNAm profiles, achieving 99.99% by MRMD and 99.66% by PCA, are much higher than the classifier based on miRNA expression profiles, which only has an accuracy of 83.40%. However, it is possible that the testing samples and training sample coming from a same database (TCGA) which may cause kind of technical artifact on such high prediction accuracy. It is possible that we may get a lower prediction accuracy if we choose another samples coming from a totally different database. Indeed, we also considered this possibility by evaluating our algorithm on seven datasets coming from GEO database, which is totally different from

TCGA, and the results showed our algorithm still works quite good on these testing samples.

Overall, the DNAm-based classifiers' performances on predicting the origins of tumors are much higher than the miRNA-based classifier (except for liver tissue). The DNAm-based classifier (MRMD) and DNAm-based classifier (PCA) can also complement each other with their respective advantages on predicting some tissues. For example, the DNAm-basd classifier (PCA) could predict the kidney, lung, pancreas more accurately than the MRMD, while has a relative lower prediction performance on other tissues compared with the DNAm-based classifier (MRMD) (Table 2).

We also used seven independent datasets from GEO (Supplementary Material; Supplementary Table S6) to test our webserver, and the performances of GEO datasets were also displayed in Figure 4. The confusion matrix of webserver prediction results was also available the Supplementary Material; Supplementary Table S7. Among these seven datasets, GSE67116 contains 53 metastasis samples, the original site of these samples is endometrium, which located in uterus. And the performance of our classifier showed there are 46 samples were correctly classified with an accuracy of 86.79%. This result, along with other six GEO datasets' results had further demonstrated that our methods have a very efficient prediction.

4 Conclusion

Patients with cancers which has an elusive site often present with a relative low survive rate and survival time, since the optimal tumor treatment depends much on the correct identification of origin of tumors. With rapid developments of sequencing technology, large amounts of sequencing data have been generated and are becoming readily available, which also facilitates the development of machine learning methods based on the mRNA expression, or miRNA expression profiles to improve the prediction of the tumor origin (Horlings *et al.*, 2008; Kurahashi *et al.*, 2014; Rosenfeld *et al.*, 2008; Søkilde *et al.*, 2014; Tothill *et al.*, 2005). However, as displayed in this study and other reported research, some tissues are still difficult to be predicted just by the miRNA expression profiles. On the other hand, the DNAm, which are characterized by their highly tissue-specific expression, have been reported to be useful for classification of tumor types (Assié *et al.*, 2014; Network 2013, 2014) and for carcinoma of unknown primary origin recently (Moran *et al.*, 2016).

In this study, we developed novel feature selection methods to identify those tissue-specific CpGs and used random forest to construct the classifiers. The comparisons with other two benchmark classifiers (SVM and KNN) also demonstrated that the random forest is a better choice. We also apply this method on the miRNA expression profiles to compare the prediction performance with DNAm profiles. Our subsequent constructed classifiers for miRNA-based and DNAm-based datasets are both demonstrated with promising results on predicting tumor origins on a spectrum of diagnostically well-characterized tissues. We also test our classifiers on the metastases samples from GEO datasets with also achieving a promising result. The experiment results also demonstrate that the classifiers based on the DNAm profiles have a higher prediction performance than miRNA-based classifiers.

One of the main challenges in machine-learning-based classifier development is the identification of an appropriate set of features to train a classifier to accurately identify each class. For DNAm profiles, our two-level feature selection process ensured an adequate number of selected biomarkers to construct an accurate classifier. Figures 2 and 3 show that those biomarkers selected from first- or second-level feature selection had strong heterogeneous tissue-specific signatures. In addition, those selected biomarkers (miRNAs or CpGs) may have specific biological meanings, which are worth exploring for further research.

Taken together, our findings show that our three classifier types in this study (miRNA-based, DNAm-based by MRMD and DNAm-based by PCA) can efficiently predict tumor sites on well-characterized samples, which may help improve the diagnosis and treatment of patients, and also the performance of DNAm-based classifiers are better than that of miRNA-based classifiers. Through our webserver (MMCOP), users may predict the unidentified tumor sites by uploading or pasting miRNA expression profiles or DNAm profiles of some diseases. For most patients with advanced-stage tumors, treatments are becoming increasingly specific, and an adjunct genomics diagnostic regimen could enable a more directed clinical evaluation. We believe that our classifiers, as well as those based on relevant biomarkers such as mRNA and proteins, combined with additional clinical investigation will advance and promote the rational and specific therapy.

Funding

This study was supported by the Natural Science Foundation of China (Nos. 61771331, 31571359 and 31401120).

Conflict of Interest: none declared.

References

Al-Shahib,A. *et al.* (2005) Feature selection and the class imbalance problem in predicting protein function from sequence. *Applied Bioinformatics*, **4**, 195–203.

Assié,G. *et al.* (2014) Integrated genomic characterization of adrenocortical carcinoma. *Nature Genetics*, **46**, 607–612.

Bartel,D.P. (2009) MicroRNAs: target recognition and regulatory functions. *Cell*, **136**, 215–233.

Breiman,L. (2001) Random forests. *Machine Learning*, **45**, 5–32.

Bro,R., and Smilde,A.K. (2014) Principal component analysis. *Analytical Methods*, **6**, 2812–2831.

Budhu,A. (2008) Identification of metastasis-related microRNAs in hepatocellular carcinoma. *Hepatology*, **47**, 897–907.

Daugaard,D. *et al.* (2009) Tumors of unknown origin. In: *Textbook of Medical Oncology*. Informa, London, pp. 313–322.

Dedeurwaerder,S. *et al.* (2011) Evaluation of the Infinium Methylation 450K technology.

Greco,F.A., and Hainsworth,J.D. (2006) Cancer of unknown primary site. In, *Oncology*. Springer, New York;. pp. 1119–1132.

Hayashita,Y. *et al.* (2005) A polycistronic microRNA cluster, miR-17-92, is overexpressed in human lung cancers and enhances cell proliferation. *Cancer Research*, **65**, 9628–9632.

Heinzelmann,J. *et al.* (2011) Specific miRNA signatures are associated with metastasis and poor prognosis in clear cell renal cell carcinoma. *World Journal of Urology*, **29**, 367–373.

Hirsch,M.S. (2016) *Genitourinary Pathology, an Issue of Surgical Pathology Clinics*. Elsevier Health Sciences.

Horlings,H.M. *et al.* (2008) Gene expression profiling to identify the histogenetic origin of metastatic adenocarcinomas of unknown primary. *Journal of Clinical Oncology*, **26**, 4435–4441.

Hwang,H., and Mendell,J. (2006) MicroRNAs in cell proliferation, cell death, and tumorigenesis. *British Journal of Cancer*, **94**, 776–780.

Kang,S. *et al.* (2017) CancerLocator: non-invasive cancer diagnosis and tissue-of-origin prediction using methylation profiles of cell-free DNA. *Genome Biology*, **18**, 53.

Kurahashi,I. *et al.* (2013) A microarray-based gene expression analysis to identify diagnostic biomarkers for unknown primary cancer. *PLoS One*, **8**, e63249.

Kwak,E.L. *et al.* (2010) Anaplastic lymphoma kinase inhibition in non–small--cell lung cancer. *New England Journal of Medicine*, **363**, 1693–1703.

Meng,F. *et al.* (2016) Psmir: a database of potential associations between small molecules and miRNAs. *Scientific Reports*, **6**, 19264.

Monzon,F.A. *et al.* (2010) Identification of tissue of origin in carcinoma of unknown primary with a microarray-based gene expression test. *Diagnostic Pathology*, **5**, 3.

Moran,S. *et al.* (2016) Epigenetic profiling to classify cancer of unknown primary: a multicentre, retrospective analysis. *The Lancet Oncology*, **17**, 1386–1395.

Negahban,S. *et al.* (2009) A unified framework for high-dimensional analysis of $M $-estimators with decomposable regularizers. In: *Advances in Neural Information Processing Systems*. pp. 1348–1356.

Network,C.G. (2013) A. Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia. *New England Journal of Medicine*, **368**, 2059–2074.

Park,S.-Y. *et al.* (2007) Panels of immunohistochemical markers help determine primary sites of metastatic adenocarcinoma. *Archives of Pathology & Laboratory Medicine*, **131**, 1561–1567.

Pedregosa,F. *et al.* (2011) Scikit-learn: machine learning in Python. *Journal of Machine Learning Research*, **12**, 2825–2830.

Ritchie,M.E. *et al.* (2015) limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, **43**, e47.

Rosenfeld,N. *et al.* (2008) MicroRNAs accurately identify cancer tissue origin. *Nature Biotechnology*, **26**, 462–469.

Søkilde,R. *et al.* (2014) Efficient identification of miRNAs for classification of tumor origin. *The Journal of Molecular Diagnostics*, **16**, 106–115.

Statnikov,A. *et al.* (2008) A comprehensive comparison of random forests and support vector machines for microarray-based cancer classification. *BMC Bioinformatics*, **9**, 1.

Teschendorff,A.E. *et al.* (2013) A beta-mixture quantile normalization method for correcting probe design bias in Illumina Infinium 450 k DNA methylation data. *Bioinformatics*, **29**, 189–196.

Tothill,R.W. *et al.* (2005) An expression-based site of origin diagnostic method designed for clinical application to cancer of unknown origin. *Cancer Research*, **65**, 4031–4040.

Wang,J. *et al.* (2016) Identification of associations between small molecule drugs and miRNAs based on functional similarity. *Oncotarget*, **7**, 38658.

Zehir,A. *et al.* (2017) Mutational landscape of metastatic cancer revealed from prospective clinical sequencing of 10, 000 patients. *Nature Medicine*, **23**, 703–713.

Zou,Q. *et al.* (2016) A novel features ranking metric with application to scalable visual and bioinformatics data classification. *Neurocomputing*, **173**, 346–354.