The Journal of Molecular Diagnostics, Vol. 16, No. 1, January 2014

ELSEVIER

the Journal of Molecular Diagnostics

jmd.amjpathol.org

Efficient Identification of miRNAs for Classification of Tumor Origin

Rolf Søkilde,* Martin Vincent,* Anne K. Møller,† Alastair Hansen,‡ Poul E. Høiby,* Thorarinn Blondal,* Boye S. Nielsen,* Gedske Daugaard,† Søren Møller,* and Thomas Litman*

From Exiqon A/S, Vedbæk; the Department of Oncology,† State University Hospital, Copenhagen; and the Department of Pathology,‡ Herlev University Hospital, Herlev, Denmark*

Accepted for publication October 1, 2013.

Address correspondence to Thomas Litman, Ph.D., Molecular Biomedicine, Leo Pharma A/S, Ballerup, Denmark DK-2750. E-mail: tlitman@hotmail.com.

Carcinomas of unknown primary origin constitute 3% to 5% of all newly diagnosed metastatic cancers, with the primary source difficult to classify with current histological methods. Effective cancer treatment depends on early and accurate identification of the tumor; patients with metastases of unknown origin have poor prognosis and short survival. Because miRNA expression is highly tissue specific, the miRNA profile of a metastasis may be used to identify its origin. We therefore evaluated the potential of miRNA profiling to identify the primary tumor of known metastases. Two hundred eight formalin-fixed, paraffin-embedded samples, representing 15 different histologies, were profiled on a locked nucleic acid—enhanced microarray platform, which allows for highly sensitive and specific detection of miRNA. On the basis of these data, we developed and cross-validated a novel classification algorithm, least absolute shrinkage and selection operator, which had an overall accuracy of 85% (CI, 79%—89%). When the classifier was applied on an independent test set of 48 metastases, the primary site was correctly identified in 42 cases (88% accuracy; CI, 75%—94%). Our findings suggest that miRNA expression profiling on paraffin tissue can efficiently predict the primary origin of a tumor and may provide pathologists with a molecular diagnostic tool that can improve their capability to correctly identify the origin of hitherto unidentifiable metastatic tumors and, eventually, enable tailored therapy. *(J Mol Diagn 2014, 16: 106—115; http://dx.doi.org/10.1016/j.jmoldx.2013.10.001)*

Although most patients with cancer present with a primary tumor (at its site of origin), 10% to 15% of all cancers are diagnosed as metastases, and one-third of these may have a site of origin, which remains elusive, even after thorough physical and radiological examination, blood tests, and histological evaluation.[1] Thus, metastatic cancer of unknown primary (CUP) origin accounts for 3% to 6% of all cancer diagnoses and represents the seventh most frequent type of cancer, ranking below cancers of the lung, prostate, breast, cervix, colon, and stomach. Because effective cancer treatment depends on early identification of the primary tumor, patients with CUP origin have a poor prognosis with a median survival of 3 to 6 months and a 1-year survival rate of <25%. In addition, many patients with CUP origin are diagnosed with poorly differentiated adenocarcinomas, which make morphological and immunohistochemical interpretation difficult. Thus, an unrecognized number of patients may be misclassified for tumor origin, and these patients could benefit from improved molecular classification.[2] Cancer classification that is based on gene expression profiling by DNA microarrays was reported in 1999 for leukemia by Golub et al[3] and, subsequently, has been extended to include categorization of solid tumors.[4—8]

miRNAs constitute a recently discovered class of tissue-specific, small, noncoding RNAs, which regulate the expression

Supported by a grant from the Danish National Advanced Technology Foundation (048-2008-1).

Disclosure: At the time of this study, R.S., M.V., T.B., P.E.H., B.S.N., S.M., and T.L. were employed at Exiqon A/S, which has a direct financial interest in the subject matter discussed.

Current address of R.S., Institute of Clinical Sciences, Department of Oncology, Lund University, Lund, Sweden; of M.V., Department of Mathematical Sciences, University of Copenhagen, Copenhagen, Denmark; of P.E.H., Dako Denmark A/S, Glostrup, Denmark; of B.S.N., Bioneer, Hørsholm, Denmark; of S.M., Novo A/S, Hellerup, Denmark; and of T.L., Molecular Medicine, Leo Pharma A/S, Ballerup, Denmark.

http://dx.doi.org/10.1016/j.jmoldx.2013.10.001

of genes involved in many biological processes, including development, differentiation, apoptosis, and carcinogenesis.[9,10] That miRNAs are promising molecular biomarkers for classification of cancer has previously been suggested by Lu et al,[11] Volinia et al,[12] and work from Rosetta Genomics[13–16] and was recently reviewed by Di Leva and Croce.[17]

Besides their tissue specificity, a main advantage of miRNAs as biomarkers is their short size, which renders them more stable in formalin-fixed, paraffin-embedded (FFPE) material compared with mRNA.[18,19] By applying a microarray platform based on locked nucleic acid (LNA)-modified detection probes,[20] which enable highly sensitive and specific detection of >2000 miRNAs, we identified tissue-specific miRNA signatures for 35 tumors and histologies, of which 15 were selected for classification.

In this study, we evaluate the potential of miRNA expression profiling to identify the primary tumor in patients with cancer. To this end, we have developed a multiclass classification algorithm, which can identify the site of tumor origin with high specificity on the basis of the miRNA profile of the metastasis. We here describe the development of this classifier, which is based on a comprehensive miRNA expression data set.

Materials and Methods

Tumor Samples

More than 1100 FFPE tumor (both primary and metastases) and normal adjacent tissue samples were procured from the National Disease Research Interchange (Philadelphia, PA), Cytomyx (Lexington, MA), Proteogenex (Culver City, CA), and our in-house tissue bank. Every sample was obtained with a copy of its anonymized pathological report, and both the pathology information and an H&E section of each preparation was reviewed by a pathologist (A.H.) to ascertain the diagnosis, origin, and tumor percentage of the sample. Inclusion criteria for subsequent RNA extraction and miRNA expression analysis were $>0.5\text{-mm}^2$ tumor size, <25% normal adjacent tissue, <20% necrosis or hemorrhage, and confirmed histology. In the pilot phase of the project, we collected 408 samples from 35 different tumor histologies to cover a broad selection of solid tumors, whereas for the classifier, we narrowed down the list of included tissues to 15, to represent only the clinically most relevant histologies to identify tumors of unknown origin (Table 1). All demographic metadata were deposited in a database and are available in Supplemental Table S1. For validation of the classifier, an independent set of 48 metastases with known origin was collected from the National Disease Research Interchange and our in-house tissue-bank.

RNA Isolation

Total RNA was extracted from 20-μm FFPE sections with the High Pure miRNA Isolation Kit (Roche Applied Science, Mannheim, Germany) according to the manufacturer's instructions. After elution in 40 μL of RNase free water, the RNA concentration (A260 nm) and purity (A260/280 and A260/230 ratios) were assessed with a Nanodrop ND-1000

Table 1 Number of Samples per Tissue, TP, Mean PPV, and Sensitivity (with CIs) of the Classification, Assessed by Fivefold Cross-Validation of the Classifier

Tissue	Histology	Samples (*n*)	TP	Mean PPV (%)	Mean sensitivity (%)
Adrenal	ACC	8	6	100	75 (41–93)
Bile duct	Cholangiocarcinoma	18	14	100	78 (55–91)
Colorectal	Adenocarcinoma, mucinous adenocarcinoma	17	13	77	76 (53–90)
EG junction*	Adenocarcinoma, signet cell, mucinous adenocarcinoma, (squamous excluded)	20	17	83	85 (64–95)
Germ cell tumor	Nonseminoma, seminoma, embryonal carcinoma, yolk sac carcinoma	7	7	83	100 (65–100)
GIST†	Gastrointestinal stromal tumor	5	4	100	80 (38–99)
Kidney	Papillary cell carcinoma, clear cell carcinoma	20	18	87	90 (70–97)
Lung	Adenocarcinoma (squamous excluded)	20	18	86	90 (70–97)
Lymphoma	B cell, large cell, marginal zone Hodgkin's	13	12	95	93 (67–100)
Melanoma	Malignant melanoma	9	9	100	100 (70–100)
Ovary	Serous, mucinous, endometrioid adenocarcinoma, clear cell	20	13	90	65 (43–82)
Pancreas	Ductal adenocarcinoma, mucinous noncystic	20	16	80	80 (58–92)
Prostate†	Adenocarcinoma	5	4	100	80 (38–99)
Thyroid†	Papillary, Hürthle cell, follicular carcinoma	6	6	100	100 (61–100)
Urinary bladder	Transitional cell carcinoma, papillary and nonpapillary	20	19	83	95 (76–100)
Total		208	176		

*The EG junction class combines samples from esophagus and gastric cancers.
†For some tissue types, the number of samples is relatively low; therefore, the validation results for these histologies should be interpreted with caution.
ACC, adrenal cortical carcinoma; EG, esophagogastric; PPV, positive predictive value; TP, true positive count.

spectrophotometer (Thermo Scientific, Wilmington DE). The RNA was stored at −80°C until further analysis.

Microarray Profiling

For microarray analysis, we applied a common reference design in which the reference sample contains a mixture of total RNA to represent all tissue types in the study. This allows for both one- and two-channel data analysis, as described in detail by Søkilde et al.[21] In the present study, we applied the two-channel ratio analysis, because this permits comparison across different array versions. One microgram of total RNA from each sample was labeled by using the miRCURY LNA microRNA Power labeling Kit (Exiqon, Vedbæk, Denmark), according to a two-step protocol as follows: calf intestinal alkaline phosphatase was applied to remove terminal 5′ phosphates, and fluorescent labels were attached enzymatically to the 3′ end of the miRNAs. Sample-specific RNA was labeled with Hy3 (green) fluorophore, whereas the common reference RNA pool was labeled with the Hy5 (red).

The Hy3- and Hy5-labeled RNA samples were mixed and co-hybridized to miRCURY LNA Arrays version Dx10 and version 11 (Exiqon), which contain Tm-normalized capture probes that target miRNAs from human, mouse, and rat, as registered in miRBase version 19.0 at the Sanger Institute.[22] Hybridization was performed overnight for 16 hours at 65°C in a Tecan HS4800 hybridization station (Tecan, Männedorf, Switzerland). After washing and drying, the microarray slides were scanned under ozone-free conditions (ozone level < 2.0 ppb to minimize bleaching of the fluorescent dyes) in a G2565BA Microarray Scanner System (Agilent, Santa Clara, CA). The resulting images were quantified with Imagene software version 8.0 (BioDiscovery, El Segundo, CA), and both automatic quality control (flagging of poor spots by the software) and manual, visual inspection were performed to ensure the highest possible data quality.

Quantitative Real-Time PCR

The expression levels of 39 selected miRNAs were validated by quantitative real-time PCR to apply the miRCURY LNA Universal RT microRNA PCR system and SYBR Green master mix according to the manufacturer's instructions (Exiqon). The results are shown in Supplemental Figure S1.

Data Preprocessing and Normalization

All low-level analyses were performed in the R environment, including importing and preprocessing of the data with the use of the LIMMA package (*http://www.bioconductor.org/packages/2.13/bioc/html/limma.html*, last accessed August 29, 2013). Mean pixel intensities were used to calculate signal (foreground) spot intensities, and median pixel intensities were applied to estimate background intensity. After excluding flagged spots from the analysis, the normexp background correction method, with offset equal to 10, was applied.[23] For intraslide normalization, the global Lowess (Locally Weighted Scatterplot Smoothing) regression algorithm was applied, and $\log_2$ ratios of four intraslide replicates were averaged. All expression data were deposited in the Rosetta Resolver (Rosetta Biosoftware, Hoddesdon, UK) data management and analysis system.

Feature Selection and Classification

A miRNA expression database was built to identify miRNAs with high discriminatory power between tumor histologies. Three approaches for feature selection (that is, filters, wrappers, and embedded methods) are commonly used.[24] Here, we have applied both filtering and a wrapper; differentially expressed miRNAs were identified by running a one versus one, as well as a one versus all *t*-tests for each histology, followed by ranking of the most significant candidate miRNAs. In addition, the feature selection embedded in the least absolute shrinkage and selection operator (LASSO) classification algorithm was applied. The LASSO classifier was originally described by Tibshirani[25] and is based on a multinomial logistic model, which is fitted by using L1 regularization.[26] The regularization parameter is chosen by evaluating the results of a cross-validation along the entire regularization path. To solve the L1 regularized optimization problem we used the glmnet algorithm.[27] The classifier was built on $\log_2$ ratio data from the 208 samples and 15 cancer classes listed in Table 1.

We tested and fivefold cross-validated the LASSO algorithm and have listed its model coefficient, a measure of discriminatory potential, in Supplemental Table S2. For the present multiclass classification task, we found that LASSO performed on par with or even better than other classification algorithms, such as K nearest neighbor and linear discriminant analysis (data not shown).

Statistical Analysis

All calculations and statistical tests were done in the free software environment for statistical computing and graphics R version 2.9.2 (*http://www.r-project.org*, last accessed August 29, 2013). For microarray analysis, the open source package for R, Bioconductor, was used (*http://www.bioconductor.org*). Confidence intervals were calculated with the Wilson method by using the R binom library, and the following script: binom.confint(x, n, conf.level $= 0.95$, methods $=$ wilson), where $x =$ number of successes and $n =$ number of independent trials.

Results

Sample Selection

To obtain as comprehensive a data set as possible for constructing the microarray tumor database, we initially profiled 1129 samples that spanned most tumor sites and covered 35 major histological subtypes. When considering which tissue classes to include in the final classifier, we focused on those metastatic cancers that are most frequently found, that is, at autopsy, in CUP origin. Greater than 75% of all CUP cases are adenocarcinomas and poorly differentiated carcinomas, of which the most common primary sites (when determined) are pancreas (25%), lung (20%), stomach, colorectum, and hepatobiliary tract (8% to 12% each), and kidney (5%). Squamous cell carcinomas account for 10% to 15%, most of which arise from head and neck tumors, whereas melanoma represents 4% of all CUP cases. These relative frequencies, however, should be interpreted with caution, because the epidemiology of CUP is changing due to both improved medical imaging technology and lifestyle habits; therefore, different studies report dissimilar frequencies of primary sites.[28] On the basis of the above considerations, our selection of tissues includes the major carcinoma (12 of 15 histologies), as well as melanoma, germ cell tumors (clear cell tumors), and lymphoma (small cell neoplasms), because these can be difficult to distinguish from poorly differentiated carcinoma. Finally, taking into account that in the clinical setting FFPE material is readily available and, thus, represents an important resource for molecular profiling, and that miRNAs are stable in FFPE blocks and straightforward to extract,[29] we decided to develop the classifier on FFPE material. Table 1 lists the 15 tissues and histologies (columns 1 and 2), which were included in the training set that consisted of 208 FFPE samples (199 primary tumors and 9 metastases). A detailed summary of all patient demographic data can be found in Supplemental Table S1, and the expression data are deposited in Gene Expression Omnibus (*http://www.ncbi.nlm.nih.gov/geo*; accession number GSE50894).

Tissue-Specific miRNA Expression

The distribution of tissue-specific miRNAs (ie, those miRNAs that were preferentially expressed in samples originating from one tissue compared with all other tissues) is summarized in the heatmap (Figure 1).

From the heatmap it is evident that some histologies are easy to distinguish from the rest because of a strong and homogeneous tissue-specific miRNA signature [adrenal, lymphoma, germ cell, prostate, gastrointestinal stromal tumor (GIST), and melanoma], whereas other tissue origins are more difficult to classify accurately, mainly because of heterogeneity within the group (ovary, lung) or because of high similarity to related tissue types [colorectal and esophagogastric (EG) junction].

Feature Selection

Because selection of the candidate biomarkers is crucial for performance of the classifier, we took several different approaches to identify the best possible tissue-specific markers. The first and simplest approach was to run one-against-one and one-against-all comparisons for each tissue, identifying differentially expressed miRNAs by *t*-tests. However, because running multiple two-sample *t*-tests can result in an increased risk of committing a type I error (false positive), we also applied analysis of variance to compare all 15 means (of the different histologies) in one test. Yet, because filtering-based methods, such as *t*-test and analysis of variance, do not provide a cross-validation option for optimization of the set of discriminatory features, we decided for an embedded approach, namely the LASSO method, which integrates feature selection within the classifier construction. With this method 132 miRNAs with high tissue discriminatory potential were identified; these are listed in Supplemental Table S2, which is a data matrix showing each feature's LASSO model coefficient for the particular tissue of interest.

Finally, we made a literature search for tissue-specific miRNAs and compared these with our top candidate discriminatory miRNAs. There was, not surprisingly, a high degree of overlap between the miRNAs identified in our study and those reported previously as having high predictive ability for cancer classification.[12,15,16] The overlapping

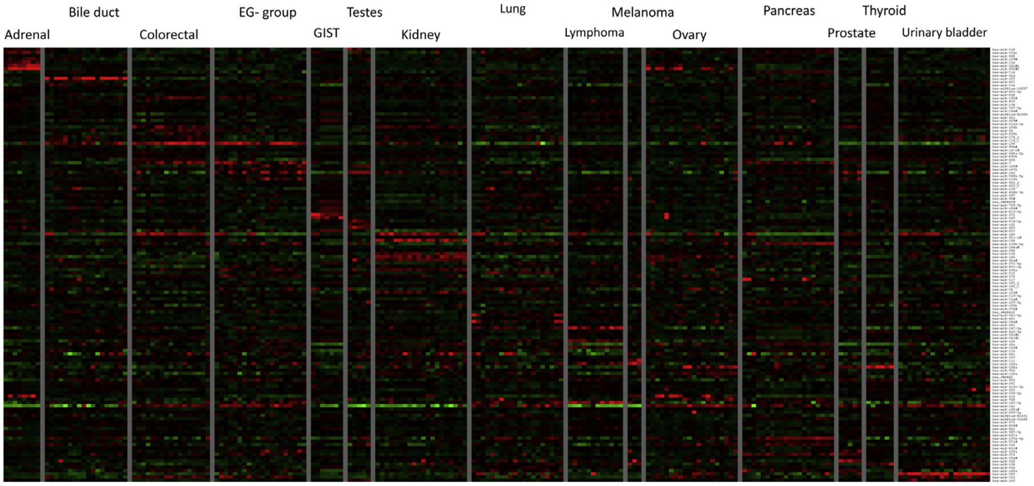

Figure 1 Expression of cancer-tissue specific miRNAs (rows) across 208 samples (columns) that represent the 15 histologies in the training set. The heatmap shows median normalized $\log_2$ data for the top 5 to 10 miRNAs, identified by LASSO's embedded feature selection algorithm, per class. For a detailed view of individual miRNAs and their classification score, please see Supplemental Table S2.

Table 2 The miRNAs That Can Be Used for Identification of Tumor Origin

Tissue	miRNA
ACC	hsa-miR-129*, hsa-miR-136, hsa-miR-202*, hsa-miR-218, hsa-miR-376c, hsa-miR-488
Bile duct; cholangiocarcinoma	hsa-miR-23a, hsa-miR-122, hsa-miR-214, hsa-miR-452, hsa-miR-616
Colorectal; adenocarcinoma, mucinous adenocarcinoma	hsa-miR-26b*, hsa-miR-95, hsa-miR-99b* hsa-miR-134, hsa-miR-192*, hsa-miR-194 hsa-miR-196b, hsa-miR-220b, hsa-miR-224 hsa-miR-433, hsa-miR-491-5p, hsa-miR-516a-3p hsa-miR-629*, hsa-miR-767-3p, hsa-miR-890
EG junction; adenocarcinoma, signet cell, mucinous adenocarcinoma (squamous excluded)[†]	hsa-miR-7, hsa-miR-16-1*, hsa-miR-96* hsa-miR-124, hsa-miR-133b, hsa-miR-143 hsa-miR-145*, hsa-miR-147b, hsa-miR-450b-3p[‡] hsa-miR-323, hsa-miR-504, hsa-miR-548a-3p hsa-miR-548b-5p, hsa-miR-647, hsa-miR-892b
Germ cell tumor; nonseminoma, seminoma, embryonal carcinoma, yolk sac carcinoma	hsa-miR-154*, hsa-miR-367, hsa-miR-372 hsa-miR-423-3p, hsa-miR-769-3p
GIST	hsa-miR-132, hsa-miR-574-3p, hsa-miR-603
Kidney; papillary cell carcinoma, clear cell carcinoma	hsa-miR-10b, hsa-miR-30a*, hsa-miR-92a-1* hsa-miR-105, hsa-miR-148a*, hsa-miR-196a hsa-miR-199b-5p, hsa-miR-204, hsa-miR-210 hsa-miR-340, hsa-miR-491-3p, hsa-miR-557
Lung; adenocarcinoma (squamous excluded)	hsa-miR-23a*, hsa-miR-34b*, hsa-miR-34c-5p hsa-miR-96, hsa-miR-126*, hsa-miR-129-3p hsa-miR-185, hsa-miR-193b, hsa-miR-212 hsa-miR-217, hsa-miR-219-5p, hsa-miR-601
Lymphoma; B cell, large cell, marginal zone Hodgkin's	hsa-miR-10a, hsa-miR-27b, hsa-miR-142-5p hsa-miR-153, hsa-miR-155, hsa-miR-155* hsa-miR-451, hsa-miR-541*, hsa-miR-615-5p hsa-miR-641
Melanoma	hsa-miR-146a, hsa-miR-150*, hsa-miR-211 hsa-miR-541*
Ovary; serous, mucinous, endometrioid adenocarcinoma, clear cell	hsa-miR-92b, hsa-miR-130a, hsa-miR-130a* hsa-miR-135a, hsa-miR-141, hsa-miR-142-3p hsa-miR-330-5p, hsa-miR-499-5p, hsa-miR-514 hsa-miR-519c-3p, hsa-miR-522, hsa-miR-572 hsa-miR-592, hsa-miR-708, hsa-miR-923
Pancreas; ductal adenocarcinoma, mucinous noncystic	hsa-miR-199a-3p, hsa-miR-221*, hsa-miR-335 hsa-miR-431*, hsa-miR-454*, hsa-miR-582-3p hsa-miR-801, hsa-miR-892a
Prostate; adenocarcinoma	hsa-miR-99a*, hsa-miR-133a, hsa-miR-363 hsa-miR-375, hsa-miR-924
Thyroid; papillary, Hürthle cell, follicular carcinoma	hsa-miR-138
Urinary bladder; transitional cell carcinoma, papillary and nonpapillary	hsa-miR-148a, hsa-miR-149, hsa-miR-203 hsa-miR-205, hsa-miR-934

[†] The EG junction class combines samples from esophagus and gastric cancers.

[‡] miR-323 was previously named miR-453.

ACC, adrenal cortical carcinoma.

miRNAs are also indicated in Supplemental Table S2. The miRNAs that can be used for classification of tumor origin are listed in Table 2.

Classifier Performance

Many different algorithms are available for multiclass cancer classification and feature selection, such as K nearest neighbor,[30] genetic algorithm,[6] linear discriminant analysis,[31] support vector machine,[32] recursive feature elimination,[4] nearest shrunken centroids,[12] decision trees,[15,33] and artificial neural networks.[34,35]

One of the main objectives of this study was to combine feature selection and multiclass classification into one pipeline. The pipeline should be able to integrate identification of highly informative features useful for classification with cross-validation of the results. This dual function is not offered by most other commonly used algorithms, which is why we decided to remodel the LASSO algorithm for this purpose.[27] Specifically, we wished to optimize the model to obtain as high sensitivity (and accuracy) on all 15 tumor classes as possible. This is illustrated in Supplemental Figure S2, which shows the performance of the LASSO classifier as a function of the regularization parameter. The optimal value of this parameter was determined to be 4.1, because more complex models would entail more miRNAs without a corresponding gain in performance.

The results of the fivefold cross-validation of the LASSO classifier are given in Table 3, which is a confusion matrix, showing the number of correct classifications along the diagonal. The correct tissue of origin was predicted in the majority of cases (176 of 208 samples tested) with an overall accuracy of 85% (CI, 79%−89%). Typically, the false-positive calls were because of similarities in histology which caused cross-reactivity; for example, three gastroesophageal (EG junction) samples were wrongly predicted

Table 3 Confusion Matrix of Classification Results That Show the Number of Correct Classifications Along the Diagonal and the Number of Mis-Classifications Off the Diagonal (Based on Fivefold Cross-Validation of the LASSO Classifier)

	True class														
Predicted class	Adrenal gland	Cholangiocarcinoma	Colorectal	EG junction	Germ cell tumor	GIST	Kidney	Lung	Lymphoma	Melanoma	Ovary	Pancreas	Prostate	Thyroid	Urothelial
Adrenal gland	6*	0	0	0	0	0									
Cholangiocarcinoma	0	14*	0	0	0	0									
Colorectal	0	0	13*	3	0	0									
EG junction	0	1	2	17*	0	0									
Germ cell tumor	0	0	0	0	7*	0	0	0	0	0	1	0	0	0	1
GIST	0	0	0	0	0	4*	0	0	0	0	0	0	0	0	0
Kidney	0	1	0	0	0	1	18*	0	0	0	1	1	0	0	0
Lung	0	2	0	0	0	0	0	18*	0	0	1	0	0	0	0
Lymphoma	0	0	0	0	0	0	0	1	12*	0	0	0	0	0	0
Melanoma	0	0	0	0	0	0	0	0	0	9*	0	0	0	0	0
Ovary	1	0	0	0	0	0	1	0	0	0	13*	0	0	0	0
Pancreas	1	0	2	0	0	0	1	0	0	0	1	16*	0	0	0
Prostate	0	0	0	0	0	0	0	0	0	0	0	0	4*	0	0
Thyroid	0	0	0	0	0	0	0	0	0	0	0	0	0	6*	0
Urothelial	0	0	0	0	0	0	0	1	1	0	1	0	1	0	19*

*The number of correct classifications.

as colorectal. We were not able to separate stomach cancers from esophageal adenocarcinomas, based on their miRNA profile; which is why we decided to pool these two, rather similar histologies, which is consistent with other, recent miRNA profiling studies.[15,16]

Validation on Metastatic Samples

Except for melanoma, the LASSO classifier was built on primary tumors. Therefore, it was important to validate its performance in an independent test set, consisting of metastases ($n = 48$) to different sites, including liver, lymph nodes, and omentum, to ensure that overfitting to the original training data were not an issue. The results of the validation are summarized in Table 4. During the optimization of the classifier, we discovered that even though the validation samples all contained <25% normal surrounding tissue, the signal from especially the liver, classified most metastases to the liver as cholangiocarcinoma. Therefore, it was necessary to add the rule to the classifier that the site of metastasis cannot be classified as the primary tumor (ie, metastasis to the liver is excluded from being identified as a primary liver tumor). The prediction of the LASSO classifier was correct in 42 of 48 cases (accuracy, 88%; CI, 75%—94%), in either the first (33 cases) or the second (nine cases) classification attempt. Thus, the classification of the independent test set that consisted of metastatic samples only showed that the performance of the LASSO classifier was comparable with the estimates from the fivefold cross-validation. The same trend of misclassification of the digestive system is seen for the metastatic samples, as for the primary tumors. Unfortunately, it has not been possible to test metastases from all of the histological classes and to all metastatic locations because of limited availability of metastatic samples.

Discussion

CUP represents a well-recognized and important clinical problem, because optimal treatment selection depends on a correct identification of the site of origin, which is per definition occult in a patient presenting with CUP. Therefore, many attempts have been made to improve diagnostic pathology workup of CUP, ranging from purely immunohistochemical schemes for subtyping the tumor,[36] over combined classification approaches,[35] to proteomic analysis[37] and machine learning algorithms that are based on large-scale mRNA microarray profiling[4,7,32,38,39] or on reverse transcription-PCR data.[6,31,40] Recently, miRNAs, which are characterized by their highly tissue-specific expression, have also been reported as useful for classification of tumor types[11,12] and for carcinoma of unknown primary origin.[13,15,16]

In this study, we have applied an LNA-enhanced microarray platform to generate miRNA expression profiles from 208 FFPE samples that represent 15 different tumor histologies. The miRNA data were used to successfully develop and validate a novel classification scheme, based on the LASSO algorithm, which integrates feature selection within the classifier construction.[27] The accuracy of the LASSO algorithm was 85% (CI, 79%—89%) when assessed by fivefold cross-validation on the initial training set, and 88% (CI, 75%—94%) when applied on an independent test set of 48 metastases. Thus, the present approach has approximately the same sensitivity as other multiclass cancer classification methods.[6,15] Where the LASSO method shows its strength, is its approximately equal sensitivity to all of the classes in the classifier. Other methods may have poor performance on a few classes; for example, the combined tree and K nearest neighbor—based miRNA classifier reported by Rosenfeld et al[15] has zero sensitivity to bladder cancer, whereas our LASSO algorithm detects this histology with a mean sensitivity of 95% (CI, 76%—100%).

Table 4 Validation of the LASSO Classifier on an Independent Test Set of 48 Metastatic Samples

True class	Correct	Metastasis site	First prediction	Percent (%)	Second prediction	Percent (%)
Colorectal	Second	Pelvis	EG junction*	31	Colorectal	22
Colorectal	First	Adrenal gland	Colorectal	52	Ovary	16
Colorectal	First	Liver	Colorectal	74	Ovary	11
Colorectal	First	Liver	Colorectal	51	EG junction	20
Colorectal	First	Liver	Colorectal	81	Ovary	7
Colorectal	First	Liver	Colorectal	70	Ovary	9
Colorectal	No	Liver	Pancreas	30	Kidney	17
Colorectal	First	Lung	Colorectal	54	EG junction	14
Colorectal	First	Liver	Colorectal	61	EG junction	11
Colorectal	Second	Omentum	Pancreas	35	Colorectal	22
Colorectal	Second	Liver	EG junction	30	Colorectal	19
Colorectal	Second	Omentum	EG junction	40	Colorectal	37
Colorectal	First	Lung	Colorectal	53	EG junction	17
Colorectal	First	Pending	Colorectal	56	EG junction	23
EG junction	First	Lymph node	EG junction	17	Pancreas	15
EG junction	First	Lymph node	EG junction	54	Lymphoma	19
EG junction	First	Lymph node	EG junction	46	Colorectal	14
Pancreas	First	Lymph node	Pancreas	81	Lung	3
Pancreas	Second	Omentum	EG junction	19	Pancreas	17
Pancreas	Second	Abdominal wall	Lung	17	Pancreas	16
Pancreas	No	Liver	Colorectal	51	EG junction	22
Pancreas	No	Omentum	EG junction	40	Colorectal	31
Ovary	First	Bowel	Ovary	37	Urothelial carcinoma	23
Ovary	First	Colon	Ovary	31	Lung	22
Ovary	Second	Colon	Pancreas	60	Ovary	13
Ovary	First	Colon	Ovary	94	Thyroid	4
Ovary	No	Colon	EG junction	46	Pancreas	9
Ovary	First	Gastric wall	Ovary	38	Thyroid	38
Ovary	First	Omentum	Ovary	33	Lung	18
Ovary	First	Omentum	Ovary	37	Pancreas	26
Ovary	First	Omentum	Ovary	45	Thyroid	12
Ovary	First	Omentum	Ovary	70	Kidney	19
Ovary	No	Omentum	Urothelial	49	Lung	32
Ovary	First	Omentum	Ovary	83	Pancreas	6
Ovary	Second	Omentum	Cholangiocarcinoma	19	Ovary	16
Ovary	First	Omentum	Ovary	55	Thyroid	17
Ovary	First	Omentum	Ovary	47	Thyroid	18
Ovary	Second	Pelvis	Lung	39	Ovary	16
Ovary	First	Pending	Ovary	31	Lung	13
Ovary	First	Pending	Ovary	54	Thyroid	16
Kidney	First	Lung	Kidney	51	Cholangiocarcinoma	14
Kidney	First	Lymph node	Kidney	61	Cholangiocarcinoma	10
Kidney	First	Adrenal gland	Kidney	76	Melanoma	4
Kidney	First	Pancreas	Kidney	96	EG junction	1
Kidney	First	Pancreas	Kidney	42	Ovary	29
Lung	First	Lymph node	Lung	44	Kidney	12
Lung	No	Lymph node	Urothelial	48	Cholangiocarcinoma	30
Urothelial	First	Colon	Urothelial	75	Pancreas	13

Correct indicates whether the classifier was correct in either its first or second prediction. The percentages are calculated by the LASSO algorithm and indicate the likelihood of a correct classification of the particular tissue.

*The EG junction class combines samples from esophagus and gastric cancers.

Identifying the algorithm that is best suited for clinical use is an ongoing and controversial discussion. It has been argued that black box machine learning classifiers, such as support vector machine and artificial neural networks, are not as transparent as, for example, decision trees for practical use by pathologists.[35] However, despite their intuitive and visual appeal, decision trees are not without limitations. If they become over-complex, they do not generalize the data well, and there is no backtracking option, meaning that a local (erroneous) optimal solution will prevent one from reaching the global optimal solution (eg, the correct classification will be missed once a wrong path is followed down a branch).[41] In this respect, it is interesting that the binary decision tree originally proposed by Rosenfeld et al[15] for miRNA classification of cancer tissue has undergone substantial structural changes in the follow-up study by Rosenwald et al[16] and Meiri et al,[13] resulting in a more complex tree (with 12 branch points for some class labels) and more than half of the 48 miRNAs reported in the original study replaced by other, tissue-specific miRNAs. This adjustment of tree structure probably reflects both the altered tissue selection and that several different tree designs may co-exist.

A recent article by Centeno et al[35] suggests a hybrid, decision tree model, which incorporates both immunohistochemistry (IHC) and expression data for optimal separation of four types of carcinoma. However, one should bear in mind that interpretation of IHC staining is subjective; therefore, it can be difficult to determine a positive from a negative. As Gown's fourth law of immunohistochemistry laconically states, "All that turns brown is not positive."[42,p30] A meta-analysis performed by Anderson and Weiss[43] showed that IHC only provides correct tissue identification in 65.6% of metastatic cancers, and recent studies by Weiss et al[44] and by Oien et al[45] both conclude molecular profiling outperforms classification by IHC, in particular in cases with poorly differentiated tumors. This underscores the need for improved identification of the origin of metastases, which are inherently more difficult to classify than their corresponding, and often more well differentiated, primary tumor.

We believe that the LASSO algorithm offers the best of both worlds, that is, the performance of the complex machine learning algorithm together with the intuitive understanding of the simpler classifiers, because it is powerful and easy to train, the model complexity (number and type of features) can be easily controlled, over-fitting is restricted by a penalty term, and data interpretation is simple; the readout is the likelihood of a correct classification. Other conventional methods, such as linear discriminant analysis and K nearest neighbor, resulted in less accurate classification (data not shown) of this data compared with LASSO.

Because we were able to identify the origin of metastatic tumors by their miRNA profile is consistent with the paradigm that the genetic makeup of a primary tumor is retained in the distant metastases.[5,13,31] Several of the identified tissue-specific miRNAs are involved in differentiation, so if the miRNA signature is retained in the metastases, it should be possible to identify its tissue of origin, unless the cancer is so dedifferentiated that all molecular marks of its primary origin are lost. This brings up the question whether a real CUP represents an entity of its own, with a CUP-specific rather than a primary tissue-specific molecular signature.[4,46,47]

Some tissues are inherently difficult to classify correctly, for example, pancreas cancer, which is often poorly differentiated or dedifferentiated, and lung cancer with many possible histologies. In our validation study, the classifier was able to correctly label pancreas as the primary site in three of five cases, which is not impressive but still better than what could be achieved in the commercial CupPrint follow-up study,[39] in which none of the three pancreas cancers could be identified. In addition, Park et al[48] found that with IHC markers the sensitivity toward the combined group of pancreas cancer and cholangiocarcinoma was quite low (28%).

We discovered that a main limitation to this type of study is the identification of the superimposed host tissue (the site of the metastasis) signature, typically liver or lymph node, rather than the metastasis signature; in particular, when the amount of host tissue is large compared with the metastasis. Specifically, a primary liver cancer could not readily be distinguished from a liver metastasis because of the overlaid, strong liver-associated miRNA signature. Therefore, when optimizing the classifier, we had to make the assumption that a metastasis to the liver cannot be primary liver cancer and that a lymph node metastasis is not a lymphoma. A likely solution to the problem of contaminating surrounding tissue is to apply laser capture microdissection as suggested by Chen et al[49] for miRNA analysis in intrahepatic cholangiocarcinoma.

In conclusion, our study suggests that miRNA expression profiling on FFPE tissue, followed by an efficient multiclass classification algorithm, in this case LASSO, can efficiently predict the primary origin of a tumor. Thus, it may provide pathologists with an adjunct molecular diagnostic tool that either alone or in combination with other relevant biomarkers, such as mRNA and proteins, for example, automated IHC, can improve their capability to correctly identify the origin of metastatic tumors, and eventually, to advance and expedite rational, specific therapy of patients with metastatic disease.

Finally, we do acknowledge that even in the light of these encouraging results, some caution is required; to translate the present discovery work and proof of concept into clinical utility requires a reduction in test complexity, migration of the microarray analysis to a quantitative reverse transcription PCR platform, further trimming of the number of discriminatory miRNAs, and prospective clinical trials. Such trials will determine the significance of miRNA profiling in molecular cancer diagnostics.

Acknowledgments

We thank Drs. Bogumil Kaczkowski and Peder Worning for initial support vector machine classification of the data and Gitte Friis and Stine Jørgensen for excellent technical assistance.

Supplemental Data

Supplemental material for this article can be found at *http://dx.doi.org/10.1016/j.jmoldx.2013.10.001*.

References

1. Greco FA: Cancer of unknown primary site. Am Soc Clin Oncol Educ Book 2013, 2013:175−181

2. Daugaard D, Møller A, Petersen B: Tumors of unknown origin. In: Edited by Cavalli F, Kaye S, Hansen H, Armitage J, Piccart-Gebhart M. Textbook of Medical Oncology. London, Informa, 2009, pp 313−322
3. Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, Bloomfield CD, Lander ES: Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 1999, 286:531−537
4. Ramaswamy S, Tamayo P, Rifkin R, Mukherjee S, Yeang CH, Angelo M, Ladd C, Reich M, Latulippe E, Mesirov JP, Poggio T, Gerald W, Loda M, Lander ES, Golub TR: Multiclass cancer diagnosis using tumor gene expression signatures. Proc Natl Acad Sci U S A 2001, 98:15149−15154
5. Buckhaults P, Zhang Z, Chen YC, Wang TL, St Croix B, Saha S, Bardelli A, Morin PJ, Polyak K, Hruban RH, Velculescu VE, Shih IeM: Identifying tumor origin using a gene expression-based classification map. Cancer Res 2003, 63:4144−4149
6. Ma XJ, Patel R, Wang X, Salunga R, Murage J, Desai R, Tuggle JT, Wang W, Chu S, Stecker K, Raja R, Robin H, Moore M, Baunoch D, Sgroi D, Erlander M: Molecular classification of human cancers using a 92-gene real-time quantitative polymerase chain reaction assay. Arch Pathol Lab Med 2006, 130:465−473
7. Kurahashi I, Fujita Y, Arao T, Kurata T, Koh Y, Sakai K, Matsumoto K, Tanioka M, Takeda K, Takiguchi Y, Yamamoto N, Tsuya A, Matsubara N, Mukai H, Minami H, Chayahara N, Yamanaka Y, Miwa K, Takahashi S, Takahashi S, Nakagawa K, Nishio K: A microarray-based gene expression analysis to identify diagnostic biomarkers for unknown primary cancer. PLoS One 2013, 8:e63249
8. Greco FA, Lennington WJ, Spigel DR, Hainsworth JD: Molecular profiling diagnosis in unknown primary cancer: accuracy and ability to complement standard pathology. J Natl Cancer Inst 2013, 105: 782−790
9. Esquela-Kerscher A, Slack FJ: Oncomirs - microRNAs with a role in cancer. Nat Rev Cancer 2006, 6:259−269
10. Iorio MV, Croce CM: microRNA involvement in human cancer. Carcinogenesis 2012, 33:1126−1133
11. Lu J, Getz G, Miska EA, varez-Saavedra E, Lamb J, Peck D, Sweet-Cordero A, Ebert BL, Mak RH, Ferrando AA, Downing JR, Jacks T, Horvitz HR, Golub TR: MicroRNA expression profiles classify human cancers. Nature 2005, 435:834−838
12. Volinia S, Calin GA, Liu CG, Ambs S, Cimmino A, Petrocca F, Visone R, Iorio M, Roldo C, Ferracin M, Prueitt RL, Yanaihara N, Lanza G, Scarpa A, Vecchione A, Negrini M, Harris CC, Croce CM: A microRNA expression signature of human solid tumors defines cancer gene targets. Proc Natl Acad Sci U S A 2006, 103:2257−2261
13. Meiri E, Mueller WC, Rosenwald S, Zepeniuk M, Klinke E, Edmonston TB, Werner M, Lass U, Barshack I, Feinmesser M, Huszar M, Fogt F, Ashkenazi K, Sanden M, Goren E, Dromi N, Zion O, Burnstein I, Chajut A, Spector Y, Aharonov R: A second-generation microRNA-based assay for diagnosing tumor tissue origin. Oncologist 2012, 17:801−812
14. Pentheroudakis G, Pavlidis N, Fountzilas G, Krikelis D, Goussia A, Stoyianni A, Sanden M, St Croix B, Yerushalmi N, Benjamin H, Meiri E, Chajut A, Rosenwald S, Aharonov R, Spector Y: Novel microRNA-based assay demonstrates 92% agreement with diagnosis based on clinicopathologic and management data in a cohort of patients with carcinoma of unknown primary. Mol Cancer 2013, 12:57
15. Rosenfeld N, Aharonov R, Meiri E, Rosenwald S, Spector Y, Zepeniuk M, Benjamin H, Shabes N, Tabak S, Levy A, Lebanony D, Goren Y, Silberschein E, Targan N, Ben-Ari A, Gilad S, Sion-Vardy N, Tobar A, Feinmesser M, Kharenko O, Nativ O, Nass D, Perelman M, Yosepovich A, Shalmon B, Polak-Charcon S, Fridman E, Avniel A, Bentwich I, Bentwich Z, Cohen D, Chajut A, Barshack I: MicroRNAs accurately identify cancer tissue origin. Nature Biotechnol 2008, 26:462−469
16. Rosenwald S, Gilad S, Benjamin S, Lebanony D, Dromi N, Faerman A, Benjamin H, Tamir R, Ezagouri M, Goren E, Barshack I, Nass D, Tobar A, Feinmesser M, Rosenfeld N, Leizerman I, Ashkenazi K, Spector Y, Chajut A, Aharonov R: Validation of a microRNA-based qRT-PCR test for accurate identification of tumor tissue origin. Mod Pathol 2010, 23:814−823
17. Di Leva G, Croce CM: miRNA profiling of cancer. Curr Opin Genet Dev 2013, 23:3−11
18. Liu A, Tetzlaff MT, Vanbelle P, Elder D, Feldman M, Tobias JW, Sepulveda AR, Xu X: MicroRNA expression profiling outperforms mRNA expression profiling in formalin-fixed paraffin-embedded tissues. Int J Clin Exp Pathol 2009, 2:519−527
19. Siebolts U, Varnholt H, Drebber U, Dienes HP, Wickenhauser C, Odenthal M: Tissues from routine pathology archives are suitable for microRNA analyses by quantitative PCR. J Clin Pathol 2009, 62: 84−88
20. Castoldi M, Benes V, Hentze MW, Muckenthaler MU: miChip: a microarray platform for expression profiling of microRNAs based on locked nucleic acid (LNA) oligonucleotide capture probes. Methods 2007, 43:146−152
21. Søkilde R, Kaczkowski B, Barken K, Mouritzen P, Møller S, Litman T: MicroRNA expression analysis by LNA enhanced microarrays. Edited by Gusev Y. MicroRNA Profiling in Cancer: A Bioinformatics Perspective. Singapore, Pan Stanford Publishing, 2009, pp 23−46
22. Kozomara A, Griffiths-Jones S: miRBase: integrating microRNA annotation and deep-sequencing data. Nucleic Acids Res 2011, 39: D152−D157
23. Ritchie ME, Silver J, Oshlack A, Holmes M, Diyagama D, Holloway A, Smyth GK: A comparison of background correction methods for two-colour microarrays. Bioinformatics 2007, 23: 2700−2707
24. Ma S, Huang J: Penalized feature selection and classification in bioinformatics. Brief Bioinform 2008, 9:392−403
25. Tibshirani R: Regression shrinkage and selection via the lasso. J Royal Stat Soc 1996, 58:267−288
26. Efron B, Hastie T, Johnstone I, Tibshirani R: Least angle regression. Ann Statist 2004, 32:409−499
27. Friedman J, Hastie T, Tibshirani R: Regularization Paths for Generalized Linear Models via Coordinate Descent. J Stat Softw 2010, 33: 1−22
28. Pentheroudakis G, Golfinopoulos V, Pavlidis N: Switching benchmarks in cancer of unknown primary: from autopsy to microarray. Eur J Cancer 2007, 43:2026−2036
29. Li J, Smyth P, Flavin R, Cahill S, Denning K, Aherne S, Guenther SM, O'Leary JJ, Sheils O: Comparison of miRNA expression patterns using total RNA extracted from matched samples of formalin-fixed paraffin-embedded (FFPE) cells and snap frozen cells. BMC Biotechnol 2007, 7:36
30. van Laar RK, Ma XJ, de JD, Wehkamp D, Floore AN, Warmoes MO, Simon I, Wang W, Erlander M, van't Veer LJ, Glas AM: Implementation of a novel microarray-based diagnostic test for cancer of unknown primary. Int J Cancer 2009, 125:1390−1397
31. Talantov D, Baden J, Jatkoe T, Hahn K, Yu J, Rajpurohit Y, Jiang Y, Choi C, Ross JS, Atkins D, Wang Y, Mazumder A: A quantitative reverse transcriptase-polymerase chain reaction assay to identify metastatic carcinoma tissue of origin. J Mol Diagn 2006, 8:320−329
32. Tothill RW, Kowalczyk A, Rischin D, Bousioutas A, Haviv I, van Laar RK, Waring PM, Zalcberg J, Ward R, Biankin AV, Sutherland RL, Henshall SM, Fong K, Pollack JR, Bowtell DD, Holloway AJ: An expression-based site of origin diagnostic method designed for clinical application to cancer of unknown origin. Cancer Res 2005, 65:4031−4040
33. Shedden KA, Taylor JM, Giordano TJ, Kuick R, Misek DE, Rennert G, Schwartz DR, Gruber SB, Logsdon C, Simeone D, Kardia SL, Greenson JK, Cho KR, Beer DG, Fearon ER, Hanash S: Accurate molecular classification of human cancers based on gene expression using a simple classifier with a pathological tree-based framework. Am J Pathol 2003, 163:1985−1995

34. Dennis JL, Oien KA: Hunting the primary: novel strategies for defining the origin of tumours. J Pathol 2005, 205:236−247
35. Centeno BA, Bloom G, Chen DT, Chen Z, Gruidl M, Nasir A, Yeatman TY: Hybrid model integrating immunohistochemistry and expression profiling for the classification of carcinomas of unknown primary site. J Mol Diagn 2010, 12:476−486
36. Oien KA: Pathologic evaluation of unknown primary cancer. Semin Oncol 2009, 36:8−37
37. Bloom GC, Eschrich S, Zhou JX, Coppola D, Yeatman TJ: Elucidation of a protein signature discriminating six common types of adenocarcinoma. Int J Cancer 2007, 120:769−775
38. Su AI, Welsh JB, Sapinoso LM, Kern SG, Dimitrov P, Lapp H, Schultz PG, Powell SM, Moskaluk CA, Frierson HF Jr., Hampton GM: Molecular classification of human carcinomas by use of gene expression signatures. Cancer Res 2001, 61:7388−7393
39. Horlings HM, van Laar RK, Kerst JM, Helgason HH, Wesseling J, van der Hoeven JJ, Warmoes MO, Floore A, Witteveen A, Lahti-Domenici J, Glas AM, van't Veer LJ, de JD: Gene expression profiling to identify the histogenetic origin of metastatic adenocarcinomas of unknown primary. J Clin Oncol 2008, 26:4435−4441
40. Varadhachary GR, Talantov D, Raber MN, Meng C, Hess KR, Jatkoe T, Lenzi R, Spigel DR, Wang Y, Greco FA, Abbruzzese JL, Hainsworth JD: Molecular profiling of carcinoma of unknown primary and correlation with clinical evaluation. J Clin Oncol 2008, 26:4442−4448
41. Geurts P, Irrthum A, Wehenkel L: Supervised learning with decision tree-based methods in computational and systems biology. Mol Biosyst 2009, 5:1593−1605
42. Voigt J, Mathieu M, Bibeau F: The advent of immunohistochemistry in carcinoma of unknown primary site: a major progress. In: Edited by Fizazi K. Carcinoma of an Unknown Primary Site. New York, Taylor & Francis, 2006, pp 25−33
43. Anderson GG, Weiss LM: Determining tissue of origin for metastatic cancers: meta-analysis and literature review of immunohistochemistry performance. Appl Immunohistochem Mol Morphol 2010, 18:3−8
44. Weiss LM, Chu P, Schroeder BE, Singh V, Zhang Y, Erlander MG, Schnabel CA: Blinded comparator study of immunohistochemical analysis versus a 92-gene cancer classifier in the diagnosis of the primary site in metastatic tumors. J Mol Diagn 2013, 15:263−269
45. Oien KA, Dennis JL: Diagnostic work-up of carcinoma of unknown primary: from immunohistochemistry to molecular profiling. Ann Oncol 2012, 23(suppl 10):x271−x277
46. Pentheroudakis G, Briasoulis E, Pavlidis N: Cancer of unknown primary site: missing primary or missing biology? Oncologist 2007, 12:418−425
47. Varadhachary G: New strategies for carcinoma of unknown primary: the role of tissue of origin molecular profiling. Clin Cancer Res 2013, 19:4027−4033
48. Park SY, Kim BH, Kim JH, Lee S, Kang GH: Panels of immunohistochemical markers help determine primary sites of metastatic adenocarcinoma. Arch Pathol Lab Med 2007, 131:1561−1567
49. Chen L, Yan HX, Yang W, Hu L, Yu LX, Liu Q, Li L, Huang DD, Ding J, Shen F, Zhou WP, Wu MC, Wang HY: The role of microRNA expression pattern in human intrahepatic cholangiocarcinoma. J Hepatol 2009, 50:358−369