© 2016 USCAP, Inc All rights reserved 0893-3952/16 $32.00

Pan-cancer transcriptome analysis reveals a gene expression signature for the identification of tumor tissue origin

Qinghua Xu[1,6], Jinying Chen[1,6], Shujuan Ni[2,3,4,6], Cong Tan[2,3,4,6], Midie Xu[2,3,4], Lei Dong[2,3,4], Lin Yuan[5], Qifeng Wang[2,3,4] and Xiang Du[2,3,4]

[1]Canhelp Genomics, Hangzhou, Zhejiang, China; [2]Department of Oncology, Shanghai Medical College, Fudan University, Shanghai, China; [3]Department of Pathology, Fudan University Shanghai Cancer Center, Shanghai, China; [4]Institute of Pathology, Fudan University, Shanghai, China and [5]Pathology Center, Shanghai General Hospital, School of Medicine, Shanghai Jiaotong University, Shanghai, China

Carcinoma of unknown primary, wherein metastatic disease is present without an identifiable primary site, accounts for ~3–5% of all cancer diagnoses. Despite the development of multiple diagnostic workups, the success rate of primary site identification remains low. Determining the origin of tumor tissue is, thus, an important clinical application of molecular diagnostics. Previous studies have paved the way for gene expression-based tumor type classification. In this study, we have established a comprehensive database integrating microarray- and sequencing-based gene expression profiles of 16 674 tumor samples covering 22 common human tumor types. From this pan-cancer transcriptome database, we identified a 154-gene expression signature that discriminated the origin of tumor tissue with an overall leave-one-out cross-validation accuracy of 96.5%. The 154-gene expression signature was first validated on an independent test set consisting of 9626 primary tumors, of which 97.1% of cases were correctly classified. Furthermore, we tested the signature on a spectrum of diagnostically challenging tumors. An overall accuracy of 92% was achieved on the 1248 tumor specimens that were poorly differentiated, undifferentiated or from metastatic tumors. Thus, we have identified a 154-gene expression signature that can accurately classify a broad spectrum of tumor types. This gene panel may hold a promise to be a useful additional tool for the determination of the tumor origin.

Modern Pathology (2016) **29,** 546–556; doi:10.1038/modpathol.2016.60; published online 18 March 2016

Cancer of unknown primary, also known as occult primary tumors, is a heterogeneous group of tumors whose primary site cannot be found when the cancer has metastasized.[1] Per 100 000 individuals, the incidence varies from 5 to 7 cases in Europe, 7 to 12 cases in the USA, and 18 to 19 cases in Australia.[2] The latest data show that cancer of unknown primary accounts for ~3–5% of all newly diagnosed cancers,[2] and it is the fourth leading cause of cancer-related death worldwide.[3,4] Generally, the prognosis of patients with carcinoma of unknown primary site is poor for those receiving empiric chemotherapy. The median survival period is 3–9 months, even when newer combination treatment regiments are administered.[5] Hence, cancer of unknown primary remains an important clinical problem that generates frustration among surgeons, oncologists, and pathologists, in addition to the uncertainty and stress it imposes on patients. Identification of the primary site can ease the patient's anxiety and improve long-term survival with the help of more specific therapy.[2,6]

In current clinical practice, patients with carcinoma of unknown primary should inform doctors of their medical history and receive detailed physical examination, laboratory testing, digital imaging, and endoscopic examination. Positron emission tomography–computed tomography, the most efficient imaging test to depict the tumor tissue of origin, can only detect 24–53% of primary lesions of cancer of unknown origin.[7] Histological examination, particularly immunohistochemistry, is the cornerstone to identify the tumor of origin. However, even with the best experts and the most advanced technology, the primary site can be identified in only 20–30% of

Correspondence: Dr Q Wang, MD or Professor X Du, MD, PhD, Department of Pathology, Fudan University Shanghai Cancer Center, No. 270 Dong An Road, Shanghai 200032, China.
E-mail: wangqifeng19821982@126.com or dx2008cn@163.com

[6]These authors contributed equally to this work.

Received 11 January 2016; revised 14 February 2016; accepted 15 February 2016; published online 18 March 2016

patients with cancer of unknown primary,[8] and the results can be subjective.

This clinical need has resulted in a quest for better and more accurate identification of the primary site of tumors. To address this need, several studies have demonstrated that the expression levels of tens to hundreds of genes can be used as a 'molecular fingerprint' to classify a multitude of tumor types. Varadhachary *et al*[9] and Talantov *et al*[10] presented a reverse transcription polymerase chain reaction-based method that measures the expression of 10 signature genes among six tumor types. Ma *et al*[11] developed a similar method based on 92 genes to classify 32 tumor types. Tothill *et al*[12] reported a 79-gene panel to discriminate among 13 tumor types. Instead of measuring conventional gene expression, Rosenfeld *et al*[13] analyzed microRNA expression to classify tumor samples.

With the rapid evolution of microarray technology over the last decade, there have been tremendous efforts invested in the field of cancer research using standardized genome-wide microarrays. Considering the large amount of high-quality, publicly available gene expression data sets, the integrative analysis of genomic data, in which data from multiple studies are combined to increase the sample size and avoid laboratory-specific bias, has the potential to yield new biological insights that are not possible from a single study.[14]

In the present study, we established a comprehensive gene expression database containing the genome-wide expression profiles of more than 16 000 tumor samples representing 22 common human cancer types. By using an innovative analytical method, we aimed to develop a gene expression signature to aid in the identification of tumor origin.

Materials and methods

Sample Collection and Data Curation

The gene expression data sets of 16 674 tumor samples with histologically confirmed origins were collected from public data repositories (eg, ArrayExpress, Gene Expression Omnibus, and The Cancer Genome Atlas Data Portal) and curated to form a comprehensive pan-cancer transcriptome database.

Array-based gene expression profiling of 7048 tumor samples was mainly conducted on three different platforms of Affymetrix oligonucleotide microarray: GeneChip Human Genome U133A Array, U133A 2.0 Array, and U133 Plus 2.0 Array. Data from raw CEL files were pre-processed using the single-channel array normalization method with default parameters. Although different opinions exist concerning data pre-processing, the single-channel array normalization method was considered as most suitable for personalized-medicine workflows. Rather than processing microarray samples as groups, which can introduce biases and present logistical challenges, the single-channel array normalization method can normalize each sample individually by modeling and removing probe- and array-specific background noise using only internal array data.[15] We further used the alternative CDF files from BrainArray Resource (http://brainarray.mbni.med.umich.edu/) to summarize the probe level intensities directly to the Entrez gene IDs. Probes mapping to multiple genes and other problems associated with old generations of Affymetrix probe designs were thereby excluded.[16]

Sequencing-based gene expression profiling of 9626 tumor samples were generated on the Illumina HiSeq 2000 RNA sequencing platform and kindly provided by The Cancer Genome Atlas pan-cancer analysis working group at Synapse website (https://www.synapse.org/).[17] The gene expression profile consists of transcriptomic data for 20 501 unique genes. The clinical information for selected samples was retrieved from the 'Clinical Biotab' section of the data matrix based on the Biospecimen Core Resource IDs of the patients.

Gene Signature Identification

Gene expression data analysis was performed using R software and packages from the Bioconductor project.[18–20] To identify a gene expression signature, we used the support vector machine—recursive feature elimination algorithm for feature selection and classification modeling.[21] For multi-class classification, a one-*versus*-all approach was used whereby multiple binary classifiers are first derived for each tumor type. The results are reported as a series of probability scores for each of the 22 tumor types. The probability score was estimated as an indicator of the certainty of a classification made by the gene expression signature. The probability score ranges from 0 (low certainty) to 100 (high certainty) and sum to 100 across the 22 primary tumor types. A threshold of probability score equal to 50 was established to indicate the confidence of a single classification. When the probability score fell below 50, the samples were considered 'unclassifiable cases'. When the probability score was above 50, the tumor type with the highest probability score was considered the tumor of origin. An example of gene expression signature classification is shown in the Supplementary Figure 1.

Signature Performance Assessment

For each specimen, the predicted primary site of the tumor was compared with the reference diagnosis. A true-positive result was indicated when the predicted tumor type matched the reference diagnosis. When the predicted tumor type and reference diagnosis did not match, the specimen was considered a false positive. For each tissue on the panel, sensitivity was defined as the ratio of true-positive results to the total

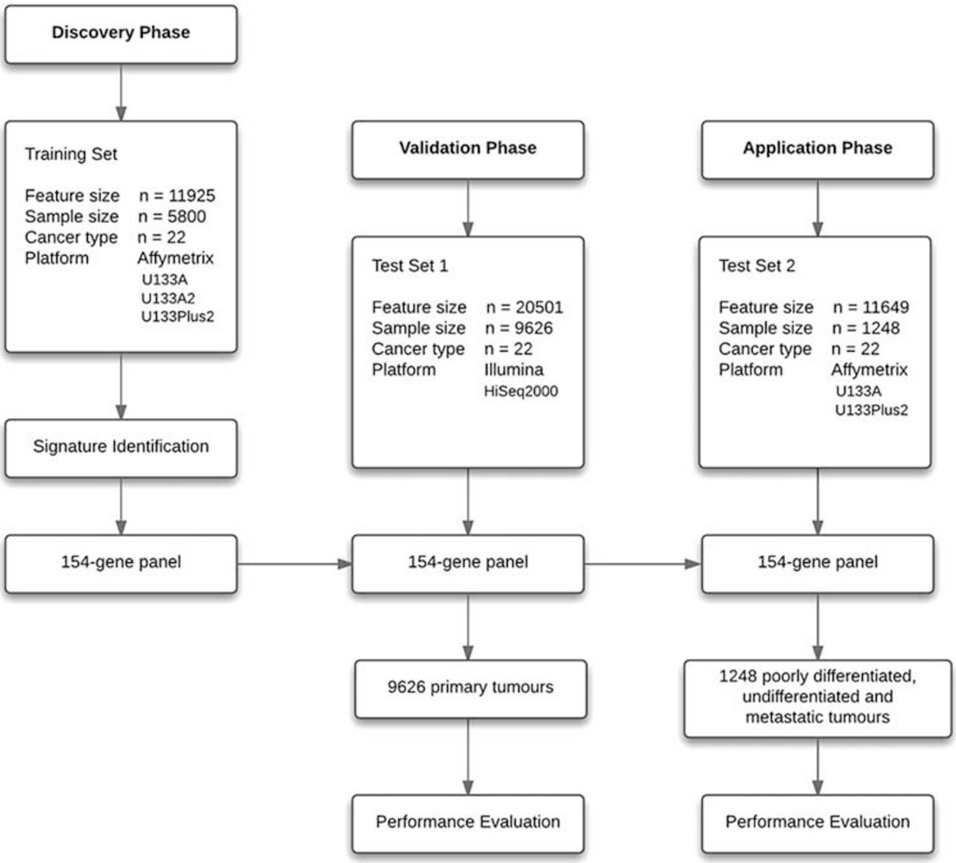


Figure 1 Flow diagram of gene expression signature identification and performance evaluation.

positive samples analyzed, while specificity was defined as the ratio (1 – false positive)/(total tested – total positive). The diagnostic odds ratio was calculated as a combination of the sensitivity and specificity as described by Glas *et al*.[22]

Results

Establishment of Pan-Cancer Transcriptome Database

To create a cancer transcriptome database for tumor primary site identification, the following issues were primarily considered. First, our database should span the tumor sites to be as large as possible. Second, within each tumor type, all possible histological subtypes should be covered. In addition, to mimic the performance of the candidate gene expression signature to identify the tumor origin in carcinoma of unknown primary, metastatic cancers, poorly differentiated tumors, and undifferentiated tumors should also be included. Thus, a systematic search of major biological data repositories—eg, ArrayExpress, Gene Expression Omnibus, and The Cancer Genome Atlas project—was performed to collect the gene expression profiling data sets of different tumor types.

Overall, we accumulated the gene expression profiles of 16 674 tumor samples to form a comprehensive pan-cancer transcriptome database. The carcinomas originated from 22 major tissue types, including adrenal gland, brain, breast, cervix, colorectal, endometrium, gastroesophagus, head and neck, kidney, liver, lung, lymphoma, melanoma, mesothelioma, neuroendocrine, ovary, pancreas, prostate, sarcoma, testis, thyroid, and urinary. The database also contains patient demographic data and clinical information. To identify a reliable gene expression signature, we adopted a training-validation approach in this study. First, the gene expression profiles of 5800 primary tumors with histologically confirmed origins were retrieved from the database and curated to form a large training set. Next, two independent validation sets were formed: one is composed of sequencing-based gene expression profiles of 9626 tumor specimens with histologically confirmed origins (test set 1) and the other is composed of gene expression profiles of 1248 tumor specimens that were poorly differentiated, undifferentiated or from metastatic tumors (test set 2). Figure 1 depicts three different phases of our study design and Table 1 summarizes the clinical characteristics of the samples in the study.

Table 1 Summary of sample information

	Training set		*Test set 1*		*Test set 2*	
Cancer type	n	%	n	%	n	%
Adrenal	55	0.95	79	0.82	44	3.53
Brain	446	7.69	708	7.36	26	2.08
Breast	542	9.34	1218	12.65	142	11.38
Cervix	113	1.95	310	3.22	19	1.52
Colorectal	439	7.57	434	4.51	96	7.69
Endometrium	262	4.52	201	2.09	15	1.2
Gastroesophagus	530	9.14	196	2.04	19	1.52
Head and neck	254	4.38	566	5.88	34	2.72
Kidney	256	4.41	1020	10.6	55	4.41
Liver	222	3.83	469	4.87	34	2.72
Lung	285	4.91	1130	11.74	190	15.22
Lymphoma	366	6.31	48	0.5	30	2.4
Melanoma	163	2.81	554	5.76	72	5.77
Mesothelioma	100	1.72	87	0.9	40	3.21
Neuroendocrine	209	3.6	187	1.94	22	1.76
Ovary	225	3.88	266	2.76	87	6.97
Pancreas	134	2.31	183	1.9	24	1.92
Prostate	458	7.9	550	5.71	41	3.29
Sarcoma	169	2.91	265	2.75	216	17.31
Testis	136	2.34	156	1.62	17	1.36
Thyroid	238	4.1	572	5.94	12	0.96
Urinary	198	3.41	427	4.44	13	1.04
Total	5800	100	9626	100	1248	100

Gene Selection and Functional Annotation

The training set consisted of 5800 samples covering more than 95% of solid tumors by incidence, with 55–542 specimens per tumor class that encompass a range of intratumor heterogeneity. After data normalization and annotation steps, a matrix of 12 000 unique genes in 5800 samples (≈70 million data points) was prepared for downstream bioinformatics analyses. Extracting a subset of informative genes from such high-dimension genomic data is a critical step for gene expression signature identification. Although many algorithms have been developed, the support vector machine—recursive feature elimination approach is considered one of the best gene selection algorithms. For each tumor type, we used the support vector machine—recursive feature elimination approach to: (1) evaluate and rank the contributions of each gene toward the optimal separation of a specific cancer type from other tumor types; (2) select the top 10-ranked genes as the most differentially expressed genes for this tumor type; and (3) repeat this process for each tumor types, and obtain 22 lists of the top 10 gene set. After removing redundant features, 154 unique genes were obtained. Full list of the 154 candidate genes with respect to each tumor types were provided in Table 2.

We further investigated whether these candidate genes revealed biological features known to be relevant to different cancers. Kyoto Encyclopedia of Genes and Genomes pathway enrichment analysis was performed using the GeneCodis bioinformatics tool (http://genecodis.dacya.ucm.es/).[23] As shown in Table 3, a diverse group of gene families is represented in the 154-gene list. The most significantly enriched gene categories are those involved in specific biological processes, including tyrosine metabolism, fat digestion and absorption, cytokine–cytokine receptor interaction, extracellular matrix–receptor interaction, and gastric acid secretion. Even more interestingly, genes described in oncogenic pathways such as those of bladder cancer, melanoma, and prostate cancer were also significantly overrepresented, reflecting their differential involvement in a range of tumor classes.

Leave-One-Out Internal Cross-Validation

As an initial step, we assessed the performance of the classifier using leave-one-out cross-validation within the training set. Leave-one-out cross-validation simulates the performance of a classification algorithm on unseen samples. With leave-one-out cross-validation, the algorithm is repeatedly retrained, leaving out one sample in each round and testing each sample on a classifier that was trained without this sample. The 154-gene expression signature showed an overall accuracy of 96.5% (5597 of 5800; 95% CI 96.0 to 97.0%) with notable variation between different cancer types. Sensitivities ranged from 89.7% (endometrium) to 100% (neuroendocrine). Using this internal validation of the training set, these data provide a preliminary estimate of classification performance.

Independent Validation in Primary Tumors Profiled with Next-Generation Sequencing

The final classification model of the 154-gene expression signature was established using the entire training set and then applied to an independent validation set comprising 9626 primary tumor samples profiled with next-generation sequencing (test set 1). Representation from 22 sites ranged from 48 (lymphoma) to 1218 (breast). The 154-gene expression signature estimated 9100 (94.5%) of 9626 samples with probability scores above 50 as 'valid classification'. Among these 9100 valid cases, the 154-gene expression signature showed 97.1% overall agreement with the reference diagnosis (8839 of 9100; 95% CI 96.8 to 97.5%). Figure 2 shows a matrix of the relationship of the test results compared with the reference diagnoses. Sensitivities for the 22 main cancer types ranged from 84.2% (gastroesophagus) to 100% (prostate). Specificities ranged from 99.4% (gastroesophagus) to 100% (mesothelioma, neuroendocrine and thyroid). The detailed sensitivity and specificity are listed in Table 4. A total of 526 cases (5.5%) were considered 'unclassifiable' by the 154-gene expression signature, with probability scores below 50. Cervix, urinary, sarcoma, head and neck, gastroesophagus, and endometrium were the most common biopsy sites

Table 2 List of selected 154 candidate genes and related tumor types

Gene symbol	*Description*	*Related tumor type*
ACPP	Acid phosphatase, prostate	Liver and prostate
ACTC1	Actin, alpha, cardiac muscle 1	Urinary
ACTG2	Actin, gamma 2, smooth muscle, enteric	Gastroesophagus, mesothelioma, and urinary
AGR2	Anterior gradient 2, protein disulfide isomerase family member	Ovary
ALDH1A2	Aldehyde dehydrogenase 1 family member A2	Mesothelioma
APOBEC3B	Apolipoprotein B mRNA editing enzyme, catalytic polypeptide-like 3B	Adrenal
APOD	Apolipoprotein D	Pancreas
ASPN	Asporin	Head and neck
ATP1B1	ATPase, Na^+/K^+ transporting, beta 1 polypeptide	Kidney and urinary
AZGP1	Alpha-2-glycoprotein 1, zinc-binding	Breast
C4BPA	Complement component 4-binding protein, alpha	Lung
C7	Complement component 7	Ovary
CA12	Carbonic anhydrase XII	Kidney
CALB2	Calbindin 2	Mesothelioma
CARTPT	CART prepropeptide	Neuroendocrine
CCL18	Chemokine (C-C motif) ligand 18	Lymphoma
CDH1	Cadherin 1, type 1	Sarcoma
CDH17	Cadherin 17, LI cadherin (liver–intestine)	Colorectal
CEACAM5	Carcinoembryonic antigen-related cell adhesion molecule 5	Breast, colorectal, and endometrium
CEACAM6	Carcinoembryonic antigen-related cell adhesion molecule 6 (non-specific Cross-reacting antigen)	Lung and urinary
CHGA	Chromogranin A	Neuroendocrine
CHGB	Chromogranin B	Neuroendocrine and pancreas
CHI3L1	Chitinase 3-like-1	Brain, sarcoma, and urinary
CHRNA3	Cholinergic receptor, nicotinic alpha 3	Neuroendocrine
CKB	Creatine kinase, brain	Liver
CLDN11	Claudin 11	Ovary
CLDN18	Claudin 18	Gastroesophagus
CLU	Clusterin	Brain
COL11A1	Collagen, type XI, alpha-1	Brain and endometrium
CPB1	Carboxypeptidase B1	Adrenal and pancreas
CXCL14	Chemokine (C-X-C motif) ligand 14	Liver
CXCL5	Chemokine (C-X-C motif) ligand 5	Liver and pancreas
CYP17A1	Cytochrome *P*450 family 17 subfamily A member 1	Adrenal
DBH	Dopamine beta-hydroxylase (dopamine beta-monooxygenase)	Neuroendocrine
DCT	Dopachrome tautomerase	Melanoma
DDX3Y	DEAD (Asp-Glu-Ala-Asp) box helicase 3, Y-linked	Cervix and testis
DLK1	Delta-like 1 homolog (*Drosophila*)	Neuroendocrine and testis
DMBT1	Deleted in malignant brain tumors 1	Lymphoma
EFEMP1	EGF-containing fibulin-like extracellular matrix protein 1	Mesothelioma
EGFL6	EGF-like-domain, multiple 6	Lung
EGFR	Epidermal growth factor receptor	Lung
EPCAM	Epithelial cell adhesion molecule	Colorectal, liver, lymphoma, and mesothelioma
ESR1	Estrogen receptor 1	Endometrium
FABP1	Fatty acid-binding protein 1, liver	Colorectal
FABP4	Fatty acid-binding protein 4, adipocyte	Breast and lung
FAM107A	Family with sequence similarity 107 member A	Brain
FOXE1	Forkhead box E1	Thyroid
GATA3	GATA-binding protein 3	Breast and colorectal
GCG	Glucagon	Pancreas
GFAP	Glial fibrillary acidic protein	Brain
GJA1	Gap junction protein alpha-1	Cervix
GPM6B	Glycoprotein M6B	Brain and melanoma
GPX3	Glutathione peroxidase 3	Thyroid
GREM1	Gremlin 1, DAN family BMP antagonist	Gastroesophagus
HBB	Hemoglobin subunit beta	Brain and sarcoma
HLA-DQA1	Major histocompatibility complex, class II, DQ alpha-1	Cervix, lymphoma, mesothelioma, sarcoma, and testis
ID4	Inhibitor of DNA binding 4, dominant-negative helix-loop-helix protein	Thyroid
IGFBP2	Insulin-like growth factor binding protein 2	Brain
IGFBP7	Insulin-like growth factor binding protein 7	Kidney
IGJ	Joining chain of multimeric IgA and IgM	Lung
INSM1	Insulinoma-associated 1	Neuroendocrine
ISL1	ISL LIM homeobox 1	Neuroendocrine
KCNJ16	Potassium channel, inwardly rectifying subfamily J, member 16	Kidney and thyroid
KLK2	Kallikrein-related peptidase 2	Prostate
KLK3	Kallikrein-related peptidase 3	Liver, prostate, and testis
KRT1	Keratin 1, type II	Melanoma
KRT13	Keratin 13, type I	Head and neck, melanoma, and urinary
KRT14	Keratin 14, type I	Breast

Table 2 (Continued)

Gene symbol	*Description*	*Related tumor type*
KRT15	Keratin 15, type I	Head and neck
KRT19	Keratin 19, type I	Adrenal, head and neck, lymphoma, mesothelioma, and urinary
KRT20	Keratin 20, type I	Colorectal
KRT4	Keratin 4, type II	Head and neck
KRT7	Keratin 7, type II	Head and neck
L1TD1	LINE-1 type transposase domain containing 1	Testis
LGALS4	Lectin, galactoside-binding, soluble, 4	Colorectal
LIPF	Lipase, gastric	Gastroesophagus
LUM	Lumican	Endometrium and ovary
MAB21L2	Mab-21-like 2 (*C. elegans*)	Gastroesophagus
MGP	Matrix Gla protein	Endometrium and ovary
MITF	Microphthalmia-associated transcription factor	Melanoma
MLANA	Melan-A	Melanoma
MMP1	Matrix metallopeptidase 1	Head and neck
MMP12	Matrix metallopeptidase 12	Endometrium and ovary
MMP3	Matrix metallopeptidase 3	Head and neck
MS4A1	Membrane-spanning 4-domains, subfamily A, member 1	Lymphoma
MSLN	Mesothelin	Mesothelioma
MSMB	Microseminoprotein-beta	Prostate
MSX1	Msh homeobox 1	Endometrium
MT3	Metallothionein 3	Adrenal
NKX2-1	NK2 homeobox 1	Thyroid
NKX3-1	NK3 homeobox 1	Prostate
NPTX2	Neuronal pentraxin II	Adrenal
NPY1R	Neuropeptide Y receptor Y1	Kidney
OGN	Osteoglycin	Sarcoma
OR51E2	Olfactory receptor family 51 subfamily E member 2	Prostate
PAPPA	Pregnancy-associated plasma protein A, pappalysin 1	Sarcoma
PAX3	Paired box 3	Melanoma
PCDH7	Protocadherin 7	Ovary
PCP4	Purkinje cell protein 4	Prostate
PEG3	Paternally expressed 3	Ovary and testis
PHOX2B	Paired-like homeobox 2b	Neuroendocrine
PI15	Peptidase inhibitor 15	Lymphoma
PIGR	Polymeric immunoglobulin receptor	Gastroesophagus
PIP	Prolactin-induced protein	Breast
PLA2G2A	Phospholipase A2 group IIA	Liver and prostate
POSTN	Periostin, osteoblast-specific factor	Thyroid
POU3F3	POU class 3 homeobox 3	Kidney
PRRX1	Paired-related homeobox 1	Endometrium
PTGDS	Prostaglandin D2 synthase	Liver
PTN	Pleiotrophin	Brain and sarcoma
PTX3	Pentraxin 3	Sarcoma
RGS4	Regulator of G-protein signaling 4	Cervix
RPS11	Ribosomal protein S11	Testis
RPS4Y1	Ribosomal protein S4, Y-linked 1	Cervix, head and neck, kidney, ovary, prostate, and testis
S100A2	S100 calcium-binding protein A2	Urinary
S100A8	S100 calcium-binding protein A8	Cervix, lymphoma, mesothelioma, and sarcoma
S100P	S100 calcium-binding protein P	Urinary
SCG5	Secretogranin V	Pancreas
SCGB1A1	Secretoglobin, family 1A, member 1 (uteroglobin)	Lung
SCGB2A2	Secretoglobin, family 2A, member 2	Breast
SERPINA3	Serpin peptidase inhibitor, clade A (alpha-1 antiproteinase, antitrypsin), member 3	Brain, breast, liver, and mesothelioma
SERPINA5	Serpin peptidase inhibitor, clade A (alpha-1 antiproteinase, antitrypsin), member 5	Adrenal
SERPINB3	Serpin peptidase inhibitor, clade B (ovalbumin), member 3	Cervix
SERPINB4	Serpin peptidase inhibitor, clade B (ovalbumin), member 4	Cervix
SFN	Stratifin	Sarcoma
SFRP1	Secreted frizzled-related protein 1	Cervix
SFTPB	Surfactant protein B	Lung
SFTPC	Surfactant protein C	Lung
SFTPD	Surfactant protein D	Lung
SLC26A3	Solute carrier family 26 (anion exchanger), member 3	Colorectal
SLC26A4	Solute carrier family 26 (anion exchanger), member 4	Thyroid
SLC2A3	Solute carrier family 2 (facilitated glucose transporter), member 3	Testis
SLC3A1	Solute carrier family 3 (amino acid transporter heavy chain), member 1	Kidney
SPINK1	Serine peptidase inhibitor, Kazal type 1	Gastroesophagus and pancreas
SPP1	Secreted phosphoprotein 1	Kidney and lymphoma
SST	Somatostatin	Pancreas
STAR	Steroidogenic acute regulatory protein	Adrenal

Table 2 (Continued)

Gene symbol	*Description*	*Related tumor type*
SULT2A1	Sulfotransferase family 2A member 1	Adrenal
TACSTD2	Tumor-associated calcium signal transducer 2	Colorectal, lymphoma, and urinary
TG	Thyroglobulin	Thyroid
TH	Tyrosine hydroxylase	Adrenal and neuroendocrine
THBS4	Thrombospondin 4	Gastroesophagus
TM4SF4	Transmembrane 4 L six family member 4	Liver and pancreas
TPO	Thyroid peroxidase	Thyroid
TRPM1	Transient receptor potential cation channel, subfamily M, member 1	Melanoma
TRPS1	Trichorhinophalangeal syndrome I	Breast
TSHR	Thyroid-stimulating hormone receptor	Thyroid
TSPAN8	Tetraspanin 8	Breast, colorectal, and gastroesophagus
TSPYL5	TSPY-like 5	Endometrium
TTR	Transthyretin	Pancreas and testis
TYR	Tyrosinase	Melanoma
TYRP1	Tyrosinase-related protein 1	Melanoma
VEGFA	Vascular endothelial growth factor A	Kidney
XIST	X-inactive-specific transcript (non-protein coding)	Cervix, endometrium, gastroesophagus, head and neck, ovary, and prostate

among those unclassifiable cases. Diagnostic odds ratios for all the 22 tumor types were significantly >1, indicating that each class reported by the 154-gene expression signature provides significant discrimination and performance.

Independent Validation in Metastatic and Undifferentiated Tumors

The 154-gene expression signature was further validated in the test set 2 comprising 1248 tumor specimen samples. For the test set 2, we particularly enriched for tumor metastatic specimens with known primary sites or primary tumors with poor differentiation because these probably reflect the clinical circumstance of carcinoma of unknown primary. Representation from 22 sites ranged from 12 (thyroid) to 216 (sarcoma). The 154-gene expression signature estimated 1077 (86.3%) of 1248 samples with probability scores above 50 as 'valid classification'. Among these 1077 valid cases, the 154-gene expression signature showed 92% overall agreement with the reference diagnosis (991 of 1077; 95% CI 90.2 to 93.6%). Figure 3 shows a matrix of the relationship of the test results compared with the reference diagnoses. Sensitivities for the 22 main tumor types ranged from 38.9% (pancreas) to 100% (adrenal, brain, head and neck, liver, neuroendocrine, and testis). Specificities ranged from 98.0% (lung) to 100% (adrenal, brain, cervix, mesothelioma, neuroendocrine, pancreas, and prostate). The detailed sensitivity and specificity are listed in Table 4. One hundred seventy-one (13.7%) cases were considered 'unclassifiable' by the 154-gene expression signature, with probability scores below 50. Prostate, kidney, pancreas, urinary, adrenal, and melanoma were the most common biopsy sites among those unclassifiable cases. Diagnostic odds ratios for all 22 tumor types were significantly >1.

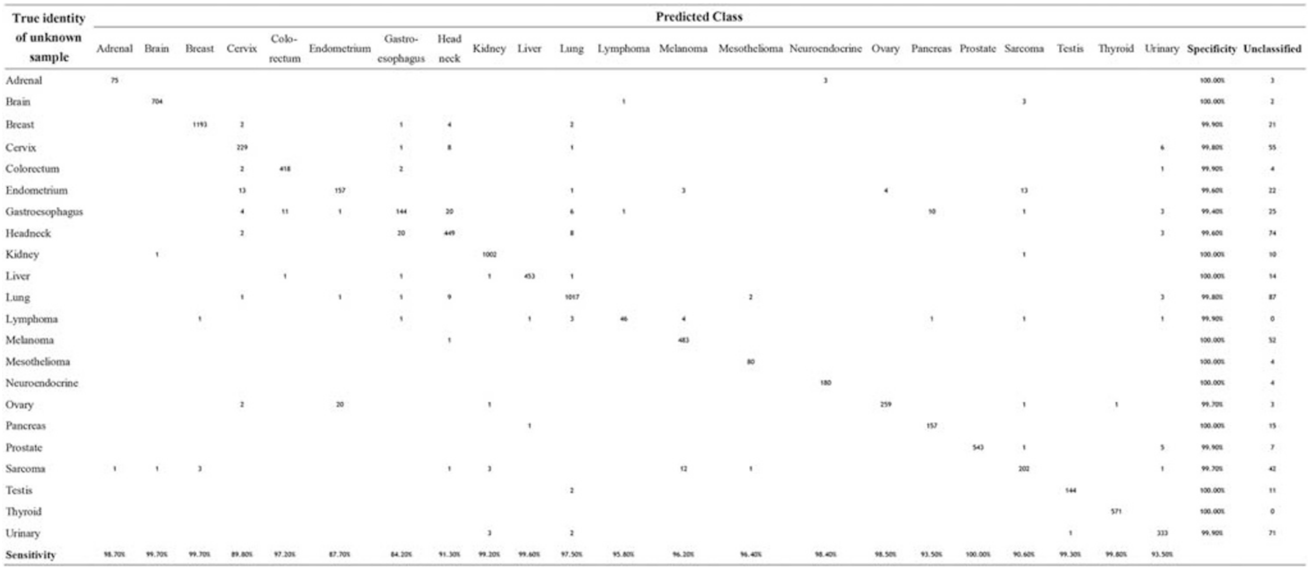
Discussion

Owing to great advancements in high-throughput microarray technologies and the comprehensive efforts of systematic cancer genomics projects, we were able to utilize large genomic data sets for our study. We report here the creation of a pan-cancer gene expression database from more than 160 000 human tumor samples and demonstrate that multiclass tumor classification is feasible by comparing an unknown sample to this reference database. The 154-gene expression signature demonstrated an overall accuracy of 96.5% for 22 tumor types by cross-validation of the training set, and 97.1% in an independent test set of 9626 primary tumors profiled with the next-generation sequencing. Furthermore, we tested the signature on a spectrum of diagnostically challenging tumors. An overall accuracy of 92% was achieved on the 1248 tumor specimens that were poorly differentiated, undifferentiated, or from metastatic tumors.

Several investigations have reported multigene algorithms and results that demonstrate the promise of gene expression-based signatures in tumor origin identification. Unlike many studies in which samples were often dominated by well-differentiated primary cancers, our approach directly exploited undifferentiated metastatic tumor samples for the validation of our 154-gene expression signature. In a clinical scenario, the uncertainty of tumors' origin usually arises within the context of metastatic and/or poorly differentiated to undifferentiated malignancies, and some of the previously published gene expression-based signatures have shown decreased performance with less-differentiated tumors. In this study, we show that the 154-gene expression signature could reliably identify the tumor origin in 92% of the 1077 tumor samples tested. This accuracy is comparable to other gene expression-based signatures with reported accuracies in the range of 79–91%.[24–26] The performance of this test also

Table 3 The top Kyoto Encyclopedia of Genes and Genomes pathways enriched in the 154-gene list

Kyoto Encyclopedia of Genes and Genomes pathways	*No. of genes*	P*-value*	*Genes*
Tyrosine metabolism	6	2.10E − 08	*TYRP1*, *DCT*, *TYR*, *DBH*, *TH*, and *TPO*
Bladder cancer	4	2.80E − 05	*EGFR*, *CDH1*, *VEGFA*, and *MMP1*
Rheumatoid arthritis	5	3.89E − 05	*MMP3*, *CXCL5*, *HLA-DQA1*, *VEGFA*, and *MMP1*
Autoimmune thyroid disease	4	4.97E − 05	*TG*, *HLA-DQA1*, *TPO*, and *TSHR*
Protein digestion and absorption	4	4.24E − 04	*ATP1B1*, *SLC3A1*, *CPB1*, and *COL11A1*
Pathways in cancer	7	6.68E − 04	*KLK3*, *NKX3-1*, *EGFR*, *MITF*, *CDH1*, *VEGFA*, and *MMP1*
Fat digestion and absorption	3	8.86E − 04	*LIPF*, *FABP1*, and *PLA2G2A*
Pancreatic secretion	4	1.00E − 03	*ATP1B1*, *CPB1*, *PLA2G2A*, and *SLC26A3*
Melanogenesis	4	1.00E − 03	*TYRP1*, *MITF*, *DCT*, and *TYR*
Cytokine–cytokine receptor interaction	6	1.13E − 03	*CXCL5*, *CCL18*, *CXCL14*, *EGFR*, *VEGFA*, and *TPO*
Endocrine and other factor-regulated calcium reabsorption	3	1.39E − 03	*ATP1B1*, *KLK2*, and *ESR1*
Focal adhesion	5	1.98E − 03	*SPP1*, *EGFR*, *THBS4*, *VEGFA*, and *COL11A1*
Cell adhesion molecules	4	2.45E − 03	*HLA-DQA1*, *CLDN11*, *CLDN18*, and *CDH1*
Chemokine signaling pathway	3	2.74E − 03	*CXCL5*, *CCL18*, and *CXCL14*
Complement and coagulation cascades	3	3.13E − 03	*C4BPA*, *SERPINA5*, and *C7*
ECM–receptor interaction	3	3.13E − 03	*SPP1*, *THBS4*, and *COL11A1*
Melanoma	3	3.55E − 03	*EGFR*, *MITF*, and *CDH1*
PPAR signaling pathway	3	3.86E − 03	*FABP4*, *FABP1*, and *MMP1*
Gastric acid secretion	3	4.18E − 03	*ATP1B1*, *KCNJ16*, and *SST*
Prostate cancer	3	4.68E − 03	*KLK3*, *NKX3-1*, and *EGFR*
Amoebiasis	3	1.09E − 02	*SERPINB3*, *SERPINB4*, and *COL11A1*
Hepatitis C	3	2.20E − 02	*EGFR*, *CLDN11*, and *CLDN18*
Phagosome	3	2.38E − 02	*THBS4*, *SFTPD*, and *HLA-DQA1*



True identity of unknown sample	Predicted Class																							
	Adrenal	Brain	Breast	Cervix	Colo-rectum	Endometrium	Gastro-esophagus	Head neck	Kidney	Liver	Lung	Lymphoma	Melanoma	Mesothelioma	Neuroendocrine	Ovary	Pancreas	Prostate	Sarcoma	Testis	Thyroid	Urinary	**Specificity**	**Unclassified**
Adrenal	75														3								100.00%	3
Brain		704										1							3				100.00%	2
Breast			1193	2			1	4			2												99.90%	21
Cervix				229			1	8			1											6	99.80%	55
Colorectum				2	418		2															1	99.90%	4
Endometrium				13		157					1		3			4			13				99.60%	22
Gastroesophagus				4	11	1	144	20			6	1					10		1			3	99.40%	25
Headneck				2			20	449			8											3	99.60%	74
Kidney		1							1002										1				100.00%	10
Liver					1		1		1	453	1												100.00%	14
Lung				1		1	1	9			1017			2								3	99.80%	87
Lymphoma			1				1			1	3	46	4				1		1			1	99.90%	0
Melanoma								1					483										100.00%	52
Mesothelioma														80									100.00%	4
Neuroendocrine															180								100.00%	4
Ovary				2		20			1							259			1		1		99.70%	3
Pancreas										1							157						100.00%	15
Prostate																		543	1			5	99.90%	7
Sarcoma	1	1	3					1	3				12	1					202			1	99.70%	42
Testis											2									144			100.00%	11
Thyroid																					571		100.00%	0
Urinary									3		2									1		333	99.90%	71
Sensitivity	98.70%	99.70%	99.70%	89.80%	97.20%	87.70%	84.20%	91.30%	99.20%	99.60%	97.50%	95.80%	96.20%	96.40%	98.40%	98.50%	93.50%	100.00%	90.60%	99.30%	99.80%	93.50%		

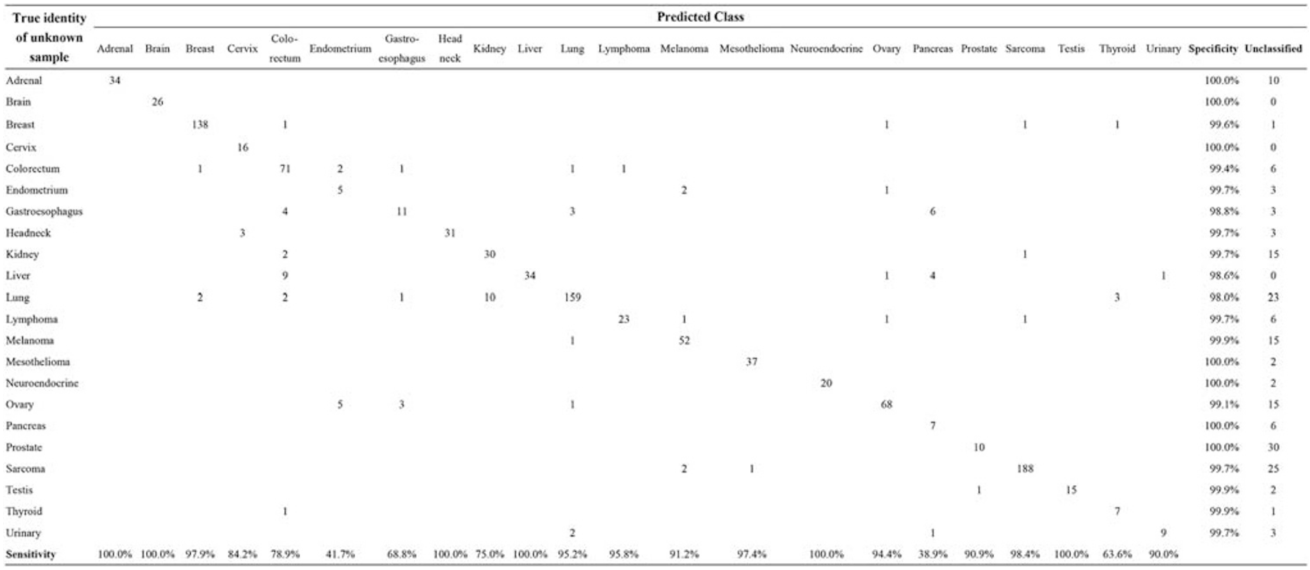
Figure 2 Confusion matrix by tumor type of the test set 1. Reference diagnoses are shown across the top row, and 154-gene expression signature predictions are shown along the left-hand column. The matrix shows the direct relationship between each adjudicated reference diagnosis *versus* the molecular classifier prediction, including reproducible patterns of classification and misclassification.

compares favorably with current clinical practice standards such as immunohistochemistry, which has shown 75% accuracy in metastatic samples using a predetermined panel of 10 antibodies.[27]

It is noteworthy that the expression patterns of several genes among the 154-gene panel have been observed previously by other methods to be relatively tissue specific for certain types of carcinomas —eg, *KLK3* has been identified as the gene encoding prostate-specific antigen, which has long been known as an important tumor marker used in the diagnosis and monitoring of prostate cancer. Originally, it was thought that prostate-specific antigen was only produced by the cells of the prostate gland. Recently, it has been shown that elevated levels of prostate-specific antigen are also observed in some breast and gynecologic cancers.[28,29] In addition, overexpression of the *EGFR* gene occurs across a

Table 4 Performance characteristics of the 154-gene expression signature in two test sets

	Test set 1			*Test set 2*		
Class	n	*Sensitivity (%)*	*Specificity (%)*	n	*Sensitivity (%)*	*Specificity (%)*
Adrenal	76	98.7	100	34	100	100
Brain	706	99.7	100	26	100	100
Breast	1197	99.7	99.9	141	97.9	99.6
Cervix	255	89.8	99.8	19	84.2	100
Colorectal	430	97.2	99.9	90	78.9	99.4
Endometrium	179	87.7	99.6	12	41.7	99.7
Gastroesophagus	171	84.2	99.4	16	68.8	98.8
Head and neck	492	91.3	99.6	31	100	99.7
Kidney	1010	99.2	100	40	75	99.7
Liver	455	99.6	100	34	100	98.6
Lung	1043	97.5	99.8	167	95.2	98
Lymphoma	48	95.8	99.9	24	95.8	99.7
Melanoma	502	96.2	100	57	91.2	99.9
Mesothelioma	83	96.4	100	38	97.4	100
Neuroendocrine	183	98.4	100	20	100	100
Ovary	263	98.5	99.7	72	94.4	99.1
Pancreas	168	93.5	100	18	38.9	100
Prostate	543	100	99.9	11	90.9	100
Sarcoma	223	90.6	99.7	191	98.4	99.7
Testis	145	99.3	100	15	100	99.9
Thyroid	572	99.8	100	11	63.6	99.9
Urinary	356	93.5	99.9	10	90	99.7
Overall	9100	95.8	99.9	1077	86.5	99.6



True identity of unknown sample	Predicted Class																							
	Adrenal	Brain	Breast	Cervix	Colo-rectum	Endometrium	Gastro-esophagus	Head neck	Kidney	Liver	Lung	Lymphoma	Melanoma	Mesothelioma	Neuroendocrine	Ovary	Pancreas	Prostate	Sarcoma	Testis	Thyroid	Urinary	**Specificity**	**Unclassified**
Adrenal	34																						100.0%	10
Brain		26																					100.0%	0
Breast			138		1											1			1		1		99.6%	1
Cervix				16																			100.0%	0
Colorectum			1		71	2	1				1	1											99.4%	6
Endometrium						5							2			1							99.7%	3
Gastroesophagus					4		11				3						6						98.8%	3
Headneck				3				31															99.7%	3
Kidney					2				30										1				99.7%	15
Liver					9					34						1	4					1	98.6%	0
Lung			2		2		1		10		159										3		98.0%	23
Lymphoma												23	1			1			1				99.7%	6
Melanoma											1		52										99.9%	15
Mesothelioma														37									100.0%	2
Neuroendocrine															20								100.0%	2
Ovary						5	3				1					68							99.1%	15
Pancreas																	7						100.0%	6
Prostate																		10					100.0%	30
Sarcoma													2	1					188				99.7%	25
Testis																		1		15			99.9%	2
Thyroid					1																7		99.9%	1
Urinary											2						1					9	99.7%	3
Sensitivity	100.0%	100.0%	97.9%	84.2%	78.9%	41.7%	68.8%	100.0%	75.0%	100.0%	95.2%	95.8%	91.2%	97.4%	100.0%	94.4%	38.9%	90.9%	98.4%	100.0%	63.6%	90.0%		

Figure 3 Confusion matrix by tumor type of the test set 2. Reference diagnoses are shown across the top row, and 154-gene expression signature predictions are shown along the left-hand column. The matrix shows the direct relationship between each adjudicated reference diagnosis *versus* the molecular classifier prediction, including reproducible patterns of classification and misclassification.

wide range of different cancers, including brain, colorectal, lung, esophageal, cervical cancers, and sarcoma.[30–35] *CDH1* and *VEGFA* have been reported among the highly significant markers in colorectal, gastric, and liver cancers.[36–41]

The 154-gene expression signature shows clear promise in identifying the tumor's origin, but it is not perfect. For diagnostically challenging tumors, systematic errors were noted in the classes of endometrial and pancreatic tumors (58 and 61% misclassified, respectively). Among the seven misclassified endometrial cancers, five were predicted to be ovarian cancer. Given the current controversies over the ontogeny of female genital tract cancers,[42–45] molecular profiling with the 154-gene expression signature may reflect this biologic intersection and provide additional insight into the origin of these tumors. Among the 11 misclassified pancreatic cancers, six were predicted to have originated from the gastroesophagus, and four from the liver. It is known that pancreatic cancer has a complex and heterogeneous genetic base, which is often identified as esophageal cancer.[46] Indeed, pancreatic cancer is the most difficult type of carcinoma of unknown primary to identify using our method as well as all published methods.[8,24,47–50]

Additional research is needed to successfully translate the 154-gene signature from gene expression microarray to real-time reverse transcription polymerase chain reaction assays, thus allowing broader access and utilization in the clinical setting. In routine practice, most diagnostic materials are formalin-fixed and paraffin-embedded; thus, it will be highly interesting to assess the usefulness of the 154-gene signature in formalin-fixed and paraffin-embedded samples. Future translational research should focus on the development and validation of the real-time polymerase chain reaction-based gene expression test using formalin-fixed and paraffin-embedded samples.

In conclusion, this study describes the development and validation of a gene expression-based signature to assist in the identification of the origin of tumor tissue. We foresee its application in cases of poorly differentiated or undifferentiated metastatic tumors and in cases where histology alone fails to suggest a specific primary site of origin. Further studies evaluating the impact of gene expression-based test results on therapy choice and treatment outcome for patients with carcinoma of unknown primary are warranted.

Acknowledgments

The results shown here are, in part, based on data from multiple previously published studies. We acknowledge the investigators and patients who contributed to the acquisition and analysis of the data used in this study. This work was partially supported by research funding from National Natural Science Foundation of China (Grant no. 81472220), Shanghai Science and Technology Development Fund (the Domestic Science and Technology Cooperation Project, No. 14495800300) and Canhelp Genomics. We thank Yang Yang, Xinming Zhang, Yi Cai, and Minzhe Fang for excellent technical and operational assistance.

Disclosure/conflict of interest

QX and JC are employees of Canhelp Genomics. No other potential conflicts of interest were disclosed by the authors.

References

1 Stella GM, Senetta R, Cassenti A *et al.* Cancers of unknown primary origin: current perspectives and future therapeutic strategies. J Transl Med 2012;10:12.

2 Richardson A, Wagland R, Foster R *et al.* Uncertainty and anxiety in the cancer of unknown primary patient journey: a multiperspective qualitative study. BMJ Support Palliat Care 2015;5:366–372.

3 Pavlidis N, Fizazi K. Cancer of unknown primary (CUP). Crit Rev Oncol Hematol 2005;54:243–250.

4 Kamposioras K, Pentheroudakis G, Pavlidis N. Exploring the biology of cancer of unknown primary: breakthroughs and drawbacks. Eur J Clin Invest 2013;43:491–500.

5 Kurahashi I, Fujita Y, Arao T *et al.* A microarray-based gene expression analysis to identify diagnostic biomarkers for unknown primary cancer. PLoS One 2013;8:e63249.

6 Hyphantis T, Papadimitriou I, Petrakis D *et al.* Psychiatric manifestations, personality traits and health-related quality of life in cancer of unknown primary site. Psychooncology 2013;22:2009–2015.

7 Reske SN, Kotzerke J. FDG-PET for clinical use. Results of the 3rd German Interdisciplinary Consensus Conference, 'Onko-PET III', 21 July and 19 September 2000. Eur J Nucl Med 2001;28:1707–1723.

8 Horlings HM, van Laar RK, Kerst JM *et al.* Gene expression profiling to identify the histogenetic origin of metastatic adenocarcinomas of unknown primary. J Clin Oncol 2008;26:4435–4441.

9 Varadhachary GR, Talantov D, Raber MN *et al.* Molecular profiling of carcinoma of unknown primary and correlation with clinical evaluation. J Clin Oncol 2008;26:4442–4448.

10 Talantov D, Baden J, Jatkoe T *et al.* A quantitative reverse transcriptase-polymerase chain reaction assay to identify metastatic carcinoma tissue of origin. J Mol Diagn 2006;8:320–329.

11 Ma XJ, Patel R, Wang X *et al.* Molecular classification of human cancers using a 92-gene real-time quantitative polymerase chain reaction assay. Arch Pathol Lab Med 2006;130:465–473.

12 Tothill RW, Kowalczyk A, Rischin D *et al.* An expression-based site of origin diagnostic method designed for clinical application to cancer of unknown origin. Cancer Res 2005;65:4031–4040.

13 Rosenfeld N, Aharonov R, Meiri E *et al.* MicroRNAs accurately identify cancer tissue origin. Nat Biotechnol 2008;26:462–469.

14 Rhodes DR, Yu J, Shanker K *et al.* Large-scale meta-analysis of cancer microarray data identifies common transcriptional profiles of neoplastic transformation and progression. Proc Natl Acad Sci USA 2004;101:9309–9314.

15 Piccolo SR, Sun Y, Campbell JD *et al.* A single-sample microarray normalization method to facilitate personalized-medicine workflows. Genomics 2012;100: 337–344.

16 Dai M, Wang P, Boyd AD *et al.* Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data. Nucleic Acids Res 2005;33:e175.

17 Omberg L, Ellrott K, Yuan Y *et al.* Enabling transparent and collaborative computational analysis of 12 tumor types within The Cancer Genome Atlas. Nat Genet 2013;45:1121–1126.

18 Ihaka R, Robert GR. A language for data analysis and graphics. J Comput Graph Stat 1996;5: 299–314.

19 Reimers M, Carey VJ. Bioconductor: an open source framework for bioinformatics and computational biology. Methods Enzymol 2006;411:119–134.

20 Chang C, Lin C. LIBSVM: a library for support vector machines. Acm Trans Intell Syst Technol 2011;2: 21–27.
21 Guyon I, Weston J, Barnhill S *et al.* Gene selection for cancer classification using support vector machines. Mach Learn 2002;46:389–422.
22 Glas AS, Lijmer JG, Prins MH *et al.* The diagnostic odds ratio: a single indicator of test performance. J Clin Epidemiol 2003;56:1129–1135.
23 Tabas-Madrid D, Nogales-Cadenas R, Pascual-Montano A. GeneCodis3: a non-redundant and modular enrichment analysis tool for functional genomics. Nucleic Acids Res 2012;40:W478–W483.
24 Monzon FA, Lyons-Weiler M, Buturovic LJ *et al.* Multicenter validation of a 1,550-gene expression profile for identification of tumor tissue of origin. J Clin Oncol 2009;27:2503–2508.
25 Kerr SE, Schnabel CA, Sullivan PS *et al.* Multisite validation study to determine performance characteristics of a 92-gene molecular cancer classifier. Clin Cancer Res 2012;18:3952–3960.
26 Weiss LM, Chu P, Schroeder BE *et al.* 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.
27 Park SY, Kim BH, Kim JH *et al.* Panels of immunohistochemical markers help determine primary sites of metastatic adenocarcinoma. Arch Pathol Lab Med 2007;131:1561–1567.
28 Mashkoor FC, Al-Asadi JN, Al-Naama LM. Serum level of prostate-specific antigen (PSA) in women with breast cancer. Cancer Epidemiol 2013;37:613–618.
29 Kucera E, Kainz C, Tempfer C *et al.* Prostate specific antigen (PSA) in breast and ovarian cancer. Anticancer Res 1997;17:4735–4737.
30 Devarakonda S, Morgensztern D, Govindan R. Genomic alterations in lung adenocarcinoma. Lancet Oncol 2015;16:e342–e351.
31 Furnari FB, Cloughesy TF, Cavenee WK *et al.* Heterogeneity of epidermal growth factor receptor signalling networks in glioblastoma. Nat Rev Cancer 2015;15: 302–310.
32 Giampieri R, Aprile G, Del Prete M *et al.* Beyond RAS: the role of epidermal growth factor receptor (EGFR) and its network in the prediction of clinical outcome during anti-EGFR treatment in colorectal cancer patients. Curr Drug Targets 2014;15:1225–1230.
33 Teng HW, Wang HW, Chen WM *et al.* Prevalence and prognostic influence of genomic changes of EGFR pathway markers in synovial sarcoma. J Surg Oncol 2011;103:773–781.
34 Li Q, Tang Y, Cheng X *et al.* EGFR protein expression and gene amplification in squamous intraepithelial lesions and squamous cell carcinomas of the cervix. Int J Clin Exp Pathol 2014;7:733–741.
35 Li JC, Zhao YH, Wang XY *et al.* Clinical significance of the expression of EGFR signaling pathway-related proteins in esophageal squamous cell carcinoma. Tumor Biol 2014;35:651–657.
36 Li YX, Lu Y, Li CY *et al.* Role of CDH1 promoter methylation in colorectal carcinogenesis: a meta-analysis. DNA Cell Biol 2014;33:455–462.
37 Jing H, Dai F, Zhao C *et al.* Association of genetic variants in and promoter hypermethylation of CDH1 with gastric cancer. Medicine (Baltimore) 2014;93:e107.
38 Liu F, Li H, Chang H *et al.* Identification of hepatocellular carcinoma-associated hub genes and pathways by integrated microarray analysis. Tumori 2015;101: 206–214.
39 Angelescu C, Burada F, Ioana M *et al.* VEGF-A and VEGF-B mRNA expression in gastro-oesophageal cancers. Clin Transl Oncol 2013;15:313–320.
40 Zhang H, Yang R. Resveratrol inhibits VEGF gene expression and proliferation of hepatocarcinoma cells. Hepatogastroenterology 2014;61:410–412.
41 Kjaer-Frifeldt S, Fredslund R, Lindebjerg J *et al.* Prognostic importance of VEGF-A haplotype combinations in a stage II colon cancer population. Pharmacogenomics 2012;13:763–770.
42 Samartzis EP, Noske A, Dedes KJ *et al.* ARID1A mutations and PI3K/AKT pathway alterations in endometriosis and endometriosis-associated ovarian carcinomas. Int J Mol Sci 2013;14:18824–18849.
43 Seidman JD, Zhao P, Yemelyanova A. 'Primary peritoneal' high-grade serous carcinoma is very likely metastatic from serous tubal intraepithelial carcinoma: assessing the new paradigm of ovarian and pelvic serous carcinogenesis and its implications for screening for ovarian cancer. Gynecol Oncol 2011;120: 470–473.
44 Kurman RJ, Shih IeM. Molecular pathogenesis and extraovarian origin of epithelial ovarian cancer—shifting the paradigm. Hum Pathol 2011;42:918–931.
45 Wiegand KC, Shah SP, Al-Agha OM *et al.* ARID1A mutations in endometriosis-associated ovarian carcinomas. N Engl J Med 2010;363:1532–1543.
46 Jones S, Zhang X, Parsons DW *et al.* Core signaling pathways in human pancreatic cancers revealed by global genomic analyses. Science 2008;321:1801–1806.
47 Ojala KA, Kilpinen SK, Kallioniemi OP. Classification of unknown primary tumors with a data-driven method based on a large microarray reference database. Genome Med 2011;3:63.
48 Monzon FA, Medeiros F, Lyons-Weiler M *et al.* Identification of tissue of origin in carcinoma of unknown primary with a microarray-based gene expression test. Diagn Pathol 2010;5:3.
49 van Laar RK, Ma XJ, de Jong D *et al.* Implementation of a novel microarray-based diagnostic test for cancer of unknown primary. Int J Cancer 2009;125:1390–1397.
50 Dumur CI, Lyons-Weiler M, Sciulli C *et al.* Interlaboratory performance of a microarray-based gene expression test to determine tissue of origin in poorly differentiated and undifferentiated cancers. J Mol Diagn 2008;10:67–77.

Supplementary Information accompanies the paper on Modern Pathology website (http://www.nature.com/modpathol)