Briefings in Bioinformatics, 00(00), 2020, 1–13

doi: 10.1093/bib/bbaa031
Problem Solving Protocol

TOD-CUP: a gene expression rank-based majority vote algorithm for tissue origin diagnosis of cancers of unknown primary

Yifei Shen, Qinjie Chu, Xinxin Yin, Yinjun He, Panpan Bai, Yunfei Wang, Weijia Fang, Michael P. Timko, Longjiang Fan and Weiqin Jiang

Corresponding authors: Weiqin Jiang, Department of Medical Oncology, First Affiliated Hospital, Zhejiang University, China. Tel./Fax: +86(571)87236621; E-mail: WeiqinJiang@zju.edu.cn; Longjiang Fan, Department of Medical Oncology, First Affiliated Hospital, Zhejiang University, China. Tel./Fax: +86(571)88982730; E-mail: fanlj@zju.edu.cn

Abstract

Gene expression profiling holds great potential as a new approach to histological diagnosis and precision medicine of cancers of unknown primary (CUP). Batch effects and different data types greatly decrease the predictive performance of biomarker-based algorithms, and few methods have been widely applied to identify tissue origin of CUP up to now. To address this problem and assist in more precise diagnosis, we have developed a gene expression rank-based majority vote algorithm for tissue origin diagnosis of CUP (TOD-CUP) of most common cancer types. Based on massive tissue-specific RNA-seq data sets (10 553) found in The Cancer Genome Atlas (TCGA), 538 feature genes (biomarkers) were selected based on their gene expression ranks and used to predict tissue types. The top scoring pairs (TSPs) classifier of the tumor type was optimized by the TCGA training samples. To test the prediction accuracy of our TOD-CUP algorithm, we analyzed (1) two microarray data sets (1029 Agilent and 2277 Affymetrix/Illumina chips) and found 91% and 94% prediction accuracy, respectively, (2) RNA-seq data from five cancer types derived from 141 public metastatic cancer tumor samples and achieved 94% accuracy and (3) a total of 25 clinical cancer samples (including 14 metastatic cancer samples) were able to classify 24/25 samples correctly (96.0% accuracy). Taken together, the TOD-CUP algorithm provides a powerful and robust means to accurately identify the tissue origin of 24 cancer types across different data platforms. To make the TOD-CUP algorithm easily accessible for clinical application, we established a Web-based server for tumor tissue origin diagnosis (http://ibi. zju.edu.cn/todcup/).

Key words: cancer of unknown primary (CUP); tissue origin diagnosis; RNA-seq; gene expression rank; majority vote algorithm

Yifei Shen is a researcher of the Department of Medical Oncology, First Affiliated Hospital, Zhejiang University and the Department of Bioinformatics and Computational Biology, The University of Texas MD Anderson Cancer Center, USA.
Qinjie Chu is a researcher of the Institute of Bioinformatics, Zhejiang University, China.
Xinxin Yin is a graduate student of the Institute of Bioinformatics, Zhejiang University, China.
Yinjun Hu is a researcher of the College of Medicine, Zhejiang University, China.
Panpan Bai is a graduate student of the Institute of Bioinformatics, Zhejiang University, China.
Yunfei Wang is a part of Zhejiang Sheng Ting Biotechnology Co., China.
Weijia Fang is an associate professor of the Department of Medical Oncology, First Affiliated Hospital, Zhejiang University, China.
Michael P. Timko is a professor of the Department of Biology & Public Health Sciences, University of Virginia, USA.
Longjiang Fan is a professor of the Department of Medical Oncology, First Affiliated Hospital and Institute of Bioinformatics, Zhejiang University, China.
Weiqin Jiang is an associate professor of the Department of Medical Oncology, First Affiliated Hospital, Zhejiang University, China.
Submitted: 8 December 2019; **Received (in revised form):** 19 January 2020

Table 1. Summary of the method for TOD-CUP

Method	Cancer type number	Sample number	Accuracy (%)	Data type	Public available	Reference
Varadhachary *et al.* (2008)	6	120	87	RNA	No	[11]
Talantov *et al.* (2006)	6	205	78	RNA	No	[12]
Ma *et al.* (2006)	32	119	82	RNA	No	[13]
Tothill *et al.* (2005)	13	229	89	RNA	No	[14]
Jiang *et al.* (2018)	3	589	95	RNA	No	[15]
Rosenfeld *et al.* (2008)	22	400	90	miRNA	No	[16]
Morgan *et al.* (2016)	38	10 481	97	Methylation	No	[17]
Penson *et al.* (2019)	22	7791	74	DNA	No	[18]
TOD-CUP	24	10 553	97	RNA	YES	This study

Introduction

Cancer of unknown primary (CUP) is a diagnosis given by clinicians when malignant (cancer) cells/tumors are found in the body, but the site of origin is not known on presentation. As such, CUPs represent a heterogeneous group of cancers constituting 3–5% of all human malignancies with an annual incidence of approximately 7–12 per 100 000 [1, 2]. Without the ability to make a definitive tissue origin diagnosis, tissue-specific first-line therapy cannot be applied, and the prognosis of patients with CUP is often poor with a median overall survival of 6 months even when empiric combination chemotherapy is administered [3, 4]. Recently, next-generation sequencing technology has facilitated the use of biomarker-based personalized therapy for CUP [5, 6]. However, even targeted therapy requires the knowledge of tissue origin of cancer, for example the same BRAF inhibitor has different efficacy for patients with BRAF mutation in melanoma and colorectal cancer [7, 8]. Hence, tissue origin diagnosis and genetic mutation analysis are the same importance for precision medicine of tumor, and CUP remains a dilemma that generates frustration among pathologists, oncologists and surgeons and undoes the levels of uncertainty and stress on patients [9].

Current clinical practice for patients with CUP involves detailed physical examination, laboratory testing, digital imaging and endoscopic examination, but even the most experienced physicians using the most advanced (immumo) histological and imaging technologies can only identify the primary site in 20–30% of patients with CUP [10]. Even in these cases, the results can be highly subjective. Clearly, there is a clinical need for better and more accurate identification of the primary site of tumors.

With the increased availability to acquire high-throughput genomic and transcriptomic data, a large number of biomarkers based on differential gene expression have been identified and used in the tissue origin identification of CUP. Such studies have utilized anywhere from tens to hundreds of differentially expressed genes to classify different tumor types (Table 1). For example, Varadhachary *et al.* [11] and Talantov *et al.* [12] used reverse transcriptase-polymerase chain reaction-based methods employing 10 genes to classify 6 cancer types. Ma *et al.* [13] used a 92 gene set to identify the tissue origin of 32 cancer types, and Tothill *et al.* [14] developed a panel of 79 genes to classify 13 cancer types. At the higher end of the spectrum, Jiang *et al.* [15] developed a naïve Bayes algorithm based on ∼1000 genes in the hepatobiliary and pancreatic system for tissue origin diagnosis. Rosenfeld *et al.* [16], Morgan *et al.* [17] and Penson *et al.* [18] analyzed microRNA expression, DNA methylation and DNA sequence data to classify tumor samples.

Although previous studies have developed multiple diagnostic methods for gene expression-based tumor-type classification, few methods could be widely applied to identify the primary site successfully. Because of the batch effects and differences in the nature of the data acquired by RNA-seq- and microarray-based methodologies, the predictive ability of biomarkers based on gene expression profiles often decreased greatly in interlaboratory validation. As far as we known, there are still no methodologies or computational programs (software) publicly available that could accurately predict the tissue origin of CUP. To address this need, we developed an innovative gene expression rank-based algorithm for tissue origin diagnosis for CUP (TOD-CUP). We also established the first publicly available Web-based server to assist clinicians and other interested parties in tumor origin diagnosis.

Materials and Methods

Training data set: The Cancer Genome Atlas RNA-seq data

The publicly available RNA-seq cancer data sets of different cancer types were downloaded from The Cancer Genome Atlas (TCGA) database [19]. We reclassified and filtered some samples based on required strategies for clinical treatment. We combined the stomach adenocarcinoma and esophageal adenocarcinoma into esophagogastric adenocarcinoma cancer type (a combined total of 486 samples, Figure **1A**) since they share similar pathology and clinical treatment types, and esophageal squamous carcinoma (83 samples, Figure 1**A**) were defined as a new cancer type for the therapy of esophageal squamous carcinoma separate from that of esophagogastric adenocarcinoma. For the same reason, intrahepatic cholangiocarcinoma and hilar cholangiocarcinoma were combined into cholangiocarcinoma, and liver cancer only consisted of hepatocellular carcinoma (HCC). Most cases of leukemia, brain tumor, retinoblastoma and osteocarcinoma can be correctly diagnosed in pathology using immunohistochemistry; therefore, these four cancer types, which also have information collected by TCGA, were not included in this study. Samples without effective clinical information were not included in our training data. For example, there are 47 skin samples and 12 uterus samples without clinical diagnosis information that were not included in our training data. In total, 10 553 TCGA RNA-seq samples originating from 24 cancer types were utilized in our work. Detailed information on the number of samples of each cancer type can be found in Figure 1**A** and Supplementary Table S1, available online at https://academic.oup.com/bib.

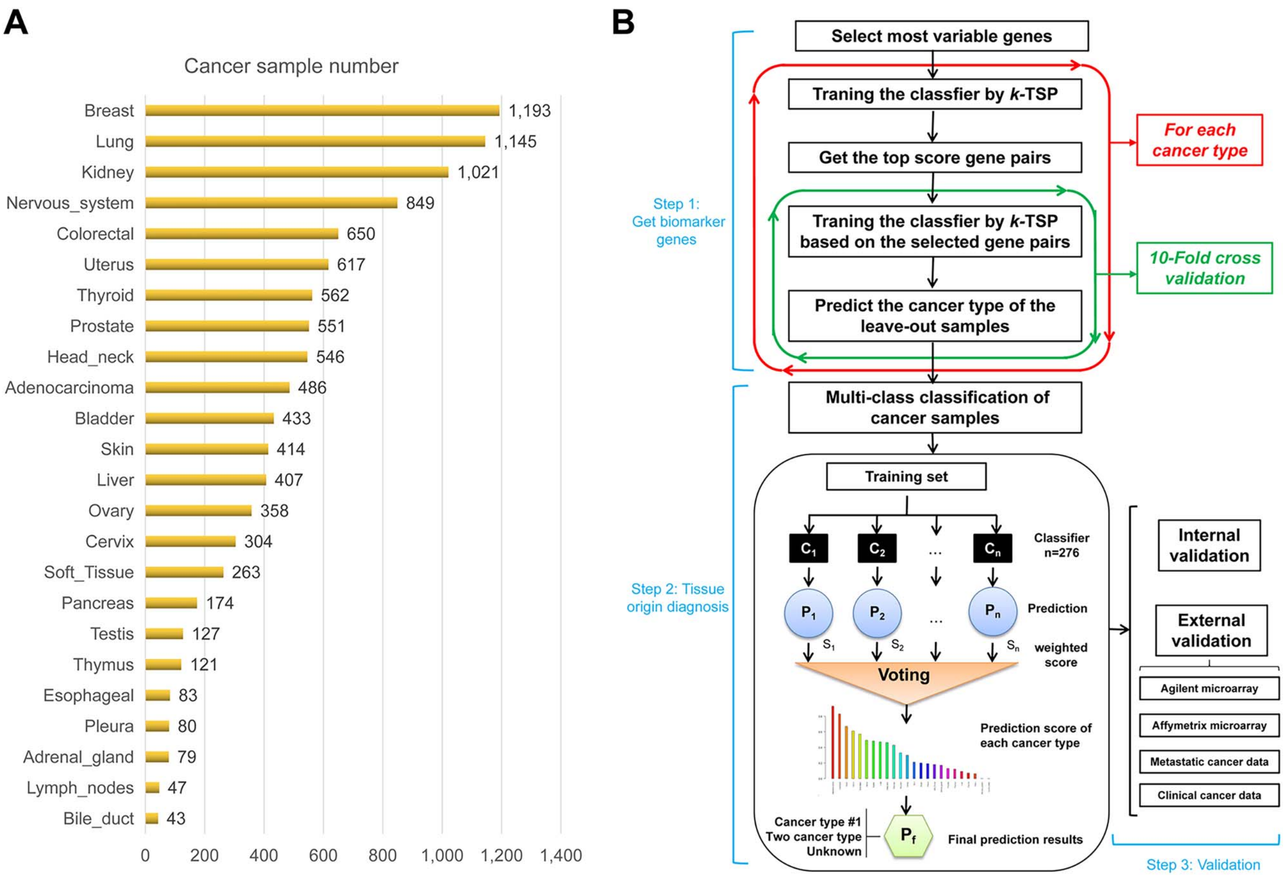


Figure 1. The training RNA-seq data set and the TOD-CUP algorithm. **A**. The training data (TCGA) of 24 cancer type and sample number used in this study. Adenocarcinoma: esophagogastric adenocarcinoma. **B**. The TOD-CUP algorithm (Step 1–2) and validation for inferring origin of CUP (Step 3).

External validation data: public microarray data

To validate the accuracy and robustness of our algorithm across different platforms, we downloaded RNA-seq and microarray data sets from different projects employing different sequencing platforms for external validation. These are as follows: 1029 TCGA Agilent microarray platform-generated samples; 347 Affymetrix microarray platform-generated samples; 1788 Illumina microarray platform-generated samples from the Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) project; and 133 metastatic cancer RNA-seq samples (Table 2).

Among the external validation microarray samples are breast cancer data set 1 [20] that includes 86 breast cancer samples generated by the Affymetrix GPL96 platform; breast cancer data set 2 that includes 1904 breast cancer samples from the METABRIC project [21] generated by the Illumina HumanHT-12 platform; a colorectal cancer data set [22] of 192 colorectal cancer samples generated by the Affymetrix GPL570 platform; a liver cancer data set of 64 liver cancer samples generated by the Affymetrix GPL13158 platform; and a thyroid cancer data set [23] of 31 thyroid cancer samples generated by the Affymetrix GPL570 platform.

External validation data: public metastatic cancer RNA-seq data

Among the metastatic cancer RNA-seq samples used for external validation are the following: metastatic breast cancer data set 1 [24] containing 78 Korean breast cancer samples generated by the Illumina HiSeq 2000 platform; breast cancer data set 2 [25] which included three matched primary, two nodal and three liver metastatic breast tumor samples generated by the Illumina HiSeq 2000 platform; the liver cancer data set [26] which included 10 Chinese liver cancer samples generated by the Illumina HiSeq 2000 platform; the skin cancer data set [27] which included 10 metastatic melanoma cancer samples generated by the Roche 454 Titanium platform; the kidney cancer data set [28] of 27 kidney cancer samples generated by the Illumina HiSeq 2500 platform; and the prostate cancer data set [29] which included eight prostate cancer samples generated by the Illumina Ovation platform.

Clinical validation data: RNA-seq data of clinical samples generated internally by this study

A total of 25 formalin-fixed paraffin-embedded (FFPE) well-characterized clinical patient cancer tissue specimens were obtained from the First Affiliated Hospital, Zhejiang University and used for RNA-Seq analysis as part of the clinical validation of our TOD-CUP algorithm (see Supplementary Table S2 available online at https://academic.oup.com/bib). Among the 25 cancer samples, 11 are primary cancer samples (five liver cancers, three pancreatic cancers, one head-neck cancer, one lung cancer and one breast cancer). Among the 25 cancer samples, 14 are metastatic cancer samples including six metastatic colorectal cancer samples in the liver, two metastatic colorectal cancer samples in the lung, three metastatic breast cancer

Table 2. External validation data sets generated by microarray (1029 TCGA samples, 2277 public samples) and RNA-seq (141 public metastatic cancer samples, 25 clinical samples)

Data type	Cancer type	Number	Data platform	Platform company	Study accession	Reference
TCGA microarray data	Breast	588	G450A	Agilent	TCGA	Weinstein *et al.* [19]
	Colorectal	244				
	Kidney	85				
	Lung	112				
		Total: 1029				
Microarray data sets	Breast 1	86	GPL96	Affymetrix	GSE25011	Hatzis *et al.* [20]
	Colorectal	192	GPL570	Affymetrix	GSE21510	Tsukamoto *et al.* [22]
	Liver	64	GPL13158	Affymetrix	GSE116174	Unpublished (2018)
	Thyroid	31	GPL570	Affymetrix	GSE3467	He *et al.* [23]
	Breast 2	1904	HuamanHT-12	Illumina	METABRIC	Curtis *et al.* [21]
		Total:2277				
Metastatic cancer RNA-seq data sets	Breast 1	78	HiSeq 2000	Illumina	ERP010142	Lee *et al.* [24]
	Liver	10	HiSeq 2000	Illumina	SRP058626	Zhang *et al.* [26]
	Kidney	27	HiSeq 2500	Illumina	SRP069243	Sciacovelli *et al.* [28]
	Prostate	8	Ovation	Illumina	SRP029603	Sowalsky *et al.* [29]
	Breast 2	8	HiSeq 2000	Illumina	SRP043470	McBryan *et al.* [25]
	Skin	10	454 Titanium	Roche	SRP003173	Valsesia *et al.* [27]
		Total: 141				
Clinical samples	Liver	5	HiSeq 4000	Illumina		Jiang *et al.* [15]
	Pancreas	3				
	Metastatic colorectal in liver	6				
	Lung	1				This study
	Head_neck	1				
	Breast	1				
	Metastatic colorectal in lung	2				
	Metastatic breast in liver	3				
	Metastatic liver in lung	2				
	Metastatic lung in adrenal gland	1				
		Total: 25				

samples in the liver, two metastatic liver cancer samples in the lung and one metastatic lung cancer samples in the adrenal gland (Table 2). Three trained histopathologists reviewed and evaluated the proportion of cancer cells to confirm the tumor cell content when possible. Total RNAs were isolated from each of the samples and used to generate pair-end sequence reads on Illumina HiSeq 4000 platform. MapSplice was used to map RNA-Seq reads to the human reference genome (hg19). RSEM was used to quantify gene expression level. This study was approved by the Research Ethical Committee of the First Affiliated Hospital, College of Medicine, Zhejiang University (Reference Number: 2018-999-1), and patients provided written informed consent to have their information used in the study.

The TOD-CUP algorithm

Top scoring pairs (TSP) classifier was introduced by Geman *et al.* [30] for the classification of gene expression data based entirely on relative gene expression values, specifically pairwise comparisons between two gene expression levels. In essence, the program exploits discriminating information contained in the R matrix by focusing on "marker gene pairs" (i, j), for which there is a significant difference in the probability of the event across the N samples from class C_1 to C_2. The quantities of interest are $p_{ij}\ (C_m) = \text{Prob}\ (R_i < R_j \mid Y = C_m)$, $m = \{1, 2\}$. These probabilities are estimated by the relative frequencies of occurrences of $Ri < Rj$ within profiles and over samples. Letting Δij denote the "score" of the gene pair (i, j), where $\Delta ij = |p_{ij}(C_1) - p_{ij}(C_2)|$, the method computes the score Δij for every pair of genes $i, j \in \{1, \ldots, P\}, i \neq j$. Pairs of genes with high scores are viewed as most informative for classification. For each top-scoring gene pair (i, j), the method computes the "average rank difference" γ_{ij} in class C_m, defined as

$$\gamma_{ij}\ (C_m) = \frac{\sum_{n\in C_m}\left(R_{i,n} - R_{j,n}\right)}{|C_m|}, m = \{1, 2\}\,.$$

Based on the original TSP algorithm, k-TSP, an ensemble method uses K pairs of genes for classifying gene expression data [31]. When $k = 1$, this algorithm, referred to simply as TSP, necessarily selects a unique pair of genes. More generally, both TSP and k-TSP may be seen as special cases of a new classification methodology based on the concept of "relative expression reversals."

However, both TSP and k-TSP are designed for binary classification problems. Therefore, in this study, we developed a weighted ensemble k-TSP algorithm for multiclass classification in TOD-CUP across 24 cancer types. Three steps were included in the TOD-CUP algorithm-based analysis for identifying the clonal origin of the tumor samples (Figure 1B).

Step 1: identification of biomarker genes based on RNA-seq data

We first selected the top 5000 most informative genes measured by median absolute deviation (MAD) to generate a data set including the variable genes. We identified the biomarker genes for each cancer type based on a "one-versus-others approach." Given multiple cancer types $T = \{T_1, T_2, \ldots, T_X\}$, the one-versus-others approach decomposes the original problem into a set of M two-class problems. For each cancer type $x = 1, \ldots, X$, we trained the classifier and identified the top score gene pairs as the biomarker genes based on k-TSP method for distinguishing between the individual cancer type T_x and the composite cancer types consisting of all other classes. To select the number of pairs, we measured the accuracy on the training set by calculating the area under the receiver operating characteristic curve (AUC) for each possible number of pairs K.

To evaluate the performance of the biomarker genes in each cancer type, the samples were then divided into 10 randomly generated subsets, each with an equal proportion of samples of the cancer type of interest. A 10-fold cross-validation was used to train the algorithm on 9-fold and test it on the remaining 1-fold. The selected biomarker gene pairs were used to train classifier. For each sample, 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 cancer type, recall was defined as the ratio of true positive/(true positive + false negative), while precision was defined as the ratio of true positive/(true positive + false positive).

Step 2: multiclass classification of cancer samples

A weighted ensemble learning method was developed for extending binary to multiclass cancer type classification. A binary cancer type classifier is constructed for each distinct pair of classes cancer type T_x, $T_y \in T$, $T_x \neq T_y$, using only the training samples for those cancer types. Consequently, this approach generates $X(X-1)/2$ binary cancer type classifiers ($X = 24$) (Figure 1B). We used all of the identified genes in Step 1, after removing the redundant ones, to train each classifier. In this scheme, the cancer type classifiers were combined by weighted voting which is based on the score from the prediction results of each classifier. To calculate the weighted score, we combined the votes of individual TSPs contained in each k-TSP classifier. We aggregated the individual TSP votes and computed a final consensus of all TSP votes based on specific combination rules. The consensus is the count of the votes taking into account the order of the features in each TSP. And the score were further used in the weighted majority voting method of each cancer type. Finally, for each sample, we obtained a score of each cancer type. To further increase the precision of the method, the final prediction results (P_f) have three states (Figure 1B). If the highest score among all the cancer type is lower than an unknown-cutoff-score (0.6), the result is placed in the "unknown" category (see Supplementary Table S3 available online at https://academic.oup.com/bib). The unknown-cutoff-score was used to exclude the prediction results which were not reliable enough. To determine the unknown-cutoff-score, we calculated the number of "unknown" sample in different corresponding the unknown-cutoff-score from 0 to 1 based on TCGA RNA-seq cancer data sets (see Supplementary Table S3 available online at https://academic.oup.com/bib). Because all the TCGA samples are well-characterized in tissue type, based on the results, we selected the unknown-cutoff-score as 0.6 to minimize the unknown sample number which had well-characterized tissue origin. If the highest score was closer than a two-cancer-type-cutoff-score (0.05) with the second highest score, the result is placed into the "two-cancer-type candidates" category (see Supplementary Table S4 and Supplementary Figure S1 available online at https://academic.oup.com/bib). The two-cancer-type-cutoff-score was used to avoid the misclassification between the top two candidate cancer types among the prediction results. To determine the two-cancer-type-cutoff-score, we calculated the number of "two-cancer-type" sample and TOD-CUP method precision in different corresponding the two-cancer-type-cutoff-scores based on TCGA RNA-seq cancer data sets (see Supplementary Table S4 available online at https://academic.oup.com/bib). Based on the results, the method had the highest increased precision level when the two-cancer-type-cutoff-score was 0.05. Finally, if the highest score is greater than 0.6 and 0.05 higher than the second score, the final result is the cancer type that has the highest score among all the cancer types. We then used the curated TOD-CUP algorithm to classify each sample among all the TCGA samples separately for internal validation.

Step 3: independent validation based on RNA-seq and microarray data

To evaluate the precision and robustness of our TOD-CUP algorithm, we used both RNA-seq and microarray data sets to perform the validation analysis. The microarray data used in this analysis were generated from by different groups using different sequencing platforms, including TCGA data sets (Agilent platform), GEO data sets (Affymetrix Human Genome U133A Array platform) and METABRIC data sets (Illumina bead chip platform) (Table 2).

The RNA-seq data of clinical samples used in this analysis were generated by our group, which included both primary and metastatic cancers.

Statistical analysis

We used a class-proportional random predictor to determine the number of correct classifications that would be expected by chance for multiclass prediction. For the permutation tests, 1000 permutations were performed on the data set. Associated P values were calculated based on the likelihood that the observed classification accuracy could be arrived at by chance [32]. In the previous studies, three schemes have been used to extend binary classifier TSP to multiclass classifiers [one-versus-one, one-versus-others and hierarchical classification (HC) schemes] [31]. The results showed that the HC-k-TSP performed best out of all three schemes [31]. To further compare the performance between the TOD-CUP algorithm and the HC-k-TSP in multiclass problems, we performed the classification analysis based on the HC-k-TSP using all the TCGA data. In brief, the HC k-TSP scheme is a sequential procedure in which a binary classifier is associated with each internal node of a binary decision tree and a class label is assigned to each leaf of the tree. The classifier $c1$ at the root is designed to distinguish between the largest class and the other classes combined ("composite class 1"); it is trained using all of the training samples. If $c1$ chooses the largest class, the procedure terminates and this becomes the final prediction. Otherwise, if $c1$ chooses composite class 1, the second classifier, $c2$, is applied, which is dedicated to separating the second largest class from "composite class 2," consisting of all classes combined except the largest and second largest; *c*2 is trained from all examples whose class labels belong to composite class 1. This procedure iterates until all the leaves in the decision-tree are labeled with a unique class.

Software implementation and Website development

TOD-CUP was developed within a Web framework with its backend based on R and PHP and is hosted at http://ibi.zju.edu.cn/todcup/. This Web framework minimizes inherent dependencies on specific hardware, software packages and libraries, and file-system attributes. Users are provided with a detailed application guide that includes several step-by-step tutorials.

Results

Summary of RNA-seq and microarray data used in this study

Publicly available RNA-seq data from 10 553 samples were obtained from TCGA for this study (Table 2). To further validate the accuracy and robustness of our method across different sequencing and analytical platforms, we also downloaded RNA-seq and microarray data sets from different projects and platforms for external validation. This included 1029 Agilent microarray platform generated samples from TCGA, 347 Affymetrix microarray platform generated samples, 1788 METABRIC Illumina microarray platform generated samples and 133 metastatic cancer RNA-seq samples. In addition, a total of 25 cancer samples (11 primary cancer samples and 14 metastatic cancer samples) were obtained and *de novo* sequenced in this study (Table 2).

Development and performance evaluation of the TOD-CUP algorithm

To select the most informative genes for classification detection, a data set of containing the top 5000 most variably expressed genes as measured by MAD was initially created (Figure 1). We then identified the biomarker genes for each of the 24 cancer types based on the one-versus-others approach. Given multiple cancer types $T = \{T_1, T_2, \ldots, T_x\}$, for each cancer type $x = 1, \ldots, X$, we trained the classifier and identified the top scoring gene pairs based on *k*-TSP method between individual cancer type T_x and the composite samples of all other cancer types. We identified between 8 and 40 biomarker genes from each cancer type (see Supplementary Table S5 available online at https://academic.oup.com/bib). After removing the redundant genes, a total of 538 biomarker genes were identified for inferring the origin of synchronous tumors among 24 cancer types (see Supplementary Table S8 available online at https://academic.oup.com/bib).

To evaluate the performance of the biomarker genes in each cancer type, a 10-fold cross-validation method based on *k*-TSP method was used to train the algorithm on 9-fold and test it on the remaining 1-fold (Figure 1**B**). The accuracy of the biomarker gene pair-generating computational algorithm was calculated based on this algorithm for each cancer type. Based on the results of 10-fold cross validation, 17 cancer types have accuracy higher than 95% among all the cancer types, and the lung cancer type has lowest accuracy at 86.7% (see Supplementary Figure S2 available online at https://academic.oup.com/bib).

To accurately identify the tissue origin of each sample among the 24 different cancer types, we developed a weighted ensemble learning method for multiclass cancer type classification based on *k*-TSP method. This approach generates a binary cancer type classifier for each pair of cancer types. We used all of the identified 538 biomarker genes to train each classifier. Among the 24 cancer types, a total of 276 binary classifiers were generated (Figure 1**B**). To further improve the accuracy of the method, the cancer type classifiers were combined by weighted voting based on the score from the prediction results of each *k*-TSP classifier. Finally, for each sample, we obtained a score of each cancer type. Based on the analysis of the prediction score of each misclassified sample, we found two patterns: (1) misclassified samples with a very low prediction score, indicating that it was difficult to classify this sample to any of the current cancer types using our algorithm and (2) misclassified samples in which the prediction scores were very close between the highest and second highest scored cancer type. This latter category accounted for most of the misclassifications.

To further increase the precision of the method, the final prediction results (P_f) were used to generate three states (Figure 1**B**). First, if the highest score among all the cancer types is lower than 0.6, the result was placed into the "unknown" category. Second, if the highest score is very close to the second highest score (within 0.05 or less than one vote among all the cancer type), the result belongs to the "two-cancer-type candidates" category. Finally, if the highest score is greater than 0.6 and 0.05 higher than the second score, the result is the cancer type with the highest score among all the cancer types.

Our analysis above shows that the TOD-CUP algorithm significantly improved the accuracy of multiclass classification. The total accuracy of resolving TOD-CUP increased from 93.5 to 97.5% compared with a simple majority vote algorithm (see Supplementary Figure S3 available online at https://academic.oup.com/bib). The ability to accurately predict 16 cancer types among the 24 cancer types analyzed increased, and the diagnosis of six cancer types increased more than 5% (e.g. for bile duct, cervix, adrenal gland, head-neck, bladder and pancreas) using the TOD-CUP algorithm compared with simple majority vote algorithm (see Supplementary Figure S3 available online at https://academic.oup.com/bib).

Using the TOD-CUP algorithm to classify each sample of the TCGA samples as internal validation, we achieved an average accuracy in the prediction of cancer type of 97.5%. In seven cancer types (adenocarcinoma, adrenal gland, nervous system, prostate, testis, thymus and thyroid), the accuracy was 100%, and in 11 others (i.e. kidney, skin, ovary, breast, colorectal, liver, lung, pleura, uterus, pancreas and bladder), it was between 95 and 100%. The recall (i.e. ability to reconfirm prior definition) was 100% in seven cancer types (esophageal, lymph nodes, ovary, prostate, testis, thymus and thyroid) and between 95 and 100% in 13 others (colorectal, nervous system, soft tissue, kidney, adrenal gland, liver, breast, pleura, adenocarcinoma, pancreas, uterus, bile duct and skin) (Tables 3A and 3B).

We further used a class-proportional random predictor to determine the number of correct classifications that would be expected by chance for multiclass prediction. Thousand permutations were performed on the data set of each cancer type, and the results showed that the prediction accuracy was highly statistically significant when compared with class-proportional random prediction ($P < 0.001$). To further compare the performance between the TOD-CUP algorithm and the HC-*k*-TSP in multiclass problems. Among the total 10 553 TCGA cancer samples, 8545 (81.0%) samples were corrected classified by the HC-*k*-TSP algorithm (see Supplementary Table S6 available online at https://academic.oup.com/bib), which is much lower than

Table 3A. The multiclassification results of 24 cancer types by the TOD-CUP algorithm based on TCGA RNA-seq data: the precision and recall of each cancer type in multiclassification results

Cancer type	Precision (%)	Recall (%)
Adenocarcinoma	100.0	97.1
Adrenal_gland	100.0	98.7
Bile_duct	88.6	97.7
Bladder	96.3	91.5
Breast	99.6	97.8
Cervix	89.1	95.4
Colorectal	98.9	99.5
Esophageal	92.2	100.0
Head_neck	87.9	95.6
Kidney	99.8	98.8
Liver	98.9	98.0
Lung	98.9	94.1
Lymph_nodes	88.7	100.0
Nervous_system	100.0	99.3
Ovary	99.7	100.0
Pancreas	96.5	97.1
Pleura	98.7	97.5
Prostate	100.0	100.0
Skin	99.7	96.6
Soft_Tissue	75.8	98.9
Testis	100.0	100.0
Thymus	100.0	100.0
Thyroid	100.0	100.0
Uterus	97.6	97.1

Table 3B. The multiclassification results of 24 cancer types by the TOD-CUP algorithm based on TCGA RNA-seq data: the confusion matrix of multiclassification of 24 cancer types

	Adenocarcinoma	Adrenal_gland	Bile_duct	Bladder	Breast	Cervix	Colorectal	Esophageal	Head_neck	Kidney	Liver	Lung	Lymph_nodes	Nervous_system	Ovary	Pancreas	Pleura	Prostate	Skin	Soft_Tissue	Testis	Thymus	Thyroid	Uterus
Adenocarcinoma	435																							
Adrenal_gland		77																						
Bile_duct			31								4													
Bladder				339					1	4		6				2								
Breast					1139				3	1		1												
Cervix				5	6	221			12			4												
Colorectal	3			1		2	642									1								
Esophageal	7							83																
Head_neck				15	1	1			449			42							3					
Kidney										1006										1				1
Liver			1								357	2												1
Lung				1	2				4			872					1							2
Lymph_nodes	1			1								1	47			1			2					
Nervous_system														837										
Ovary	1														355									
Pancreas							2					4				165								
Pleura												1					76							
Prostate																		551						
Skin																			388	1				
Soft_Tissue		1		14	17	2	1		3	7	4	2		6		1	1		9	257				14
Testis																					127			
Thymus																						121		
Thyroid																							561	
Uterus						9			1			3								1				569
Two-cancer-type	37	1	11	57	28	69	5		73	3	42	205		6	3	4	2		12	3			1	30
Unknown	2											2												

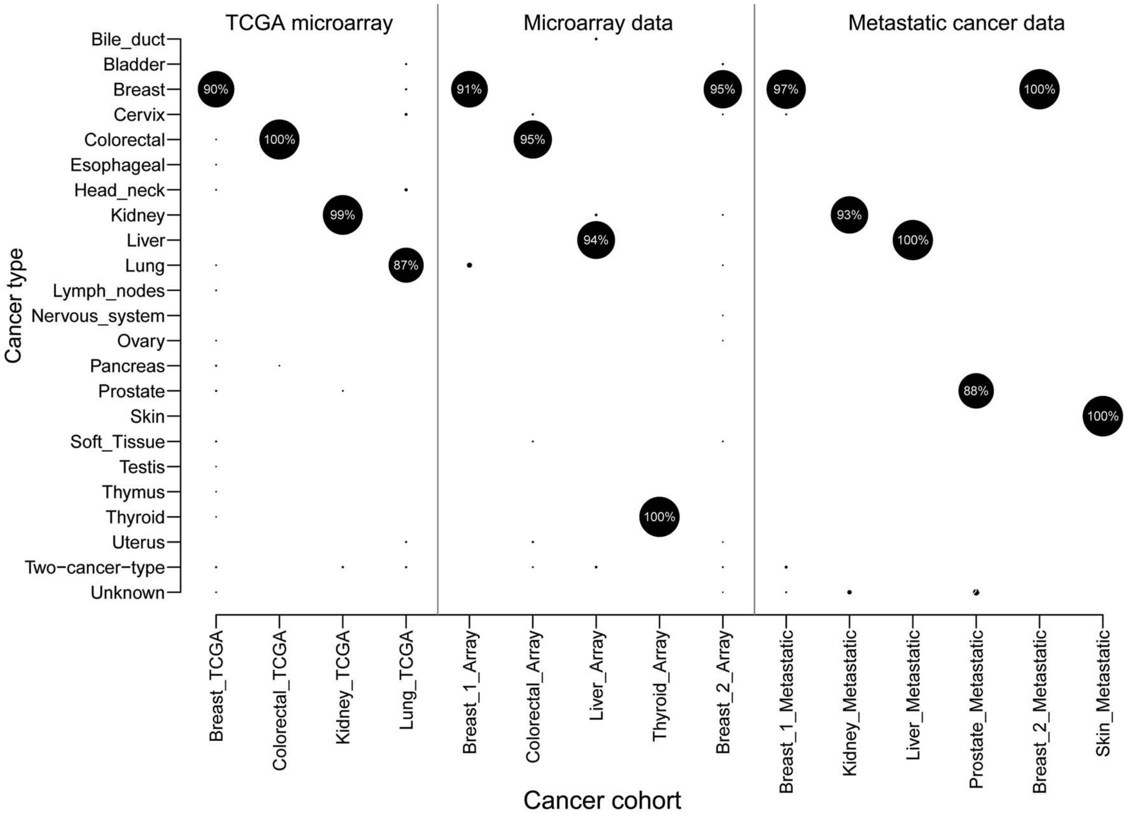


Figure 2. Prediction accuracy of samples within external validation data sets (TCGA and other microarray data sets, metastatic cancer RNA-seq data sets). Bubble size corresponds to the percentage of samples from the cohort predicted to have a given cancer type. (x-axis: the cancer cohort of external validation data sets; y-axis: the predicted cancer types).

the TOD-CUP method (97.5%). The results showed that the HC-*k*-TSP algorithm was highly accurate (>95%) in prediction in some cancer types (e.g. breast, colorectal, kidney, liver, lung, nervous_system, prostate and thyroid) but had very low accuracy (<50%) in other cancer types (Adenocarcinoma, bile_duct, bladder, cervix, esophageal, head_neck and ovary). The TOD-CUP algorithm had better performance than the HC-*k*-TSP in the tissue origin prediction for the 24 cancer types.

External validation of the TOD-CUP algorithm: microarray data sets

To assess the potential clinical performance of the TOD-CUP algorithm, we carried out an external validation using two main microarray data sets. The first set is TCGA microarray gene expression data derived from the same samples used to generate the TCGA RNA-seq data sets but obtained using microarray technology. This data set serves to evaluate the robustness of TOD-CUP algorithm in analyzing microarray data types. The second data set is microarray data from tumor samples of different cancer types generated using different expression profiling platforms in several different research projects. This second data set is used to validate the ability of the TOD-CUP algorithm to handle varied data input types.

Four main cancer types were included in TCGA microarray data set, including 588 breast cancer samples, 244 colorectal cancer samples, 85 kidney cancer samples and 112 lung cancer samples. All TCGA microarray data were generated by Agilent G450A platform. The accuracy of tissue origin diagnoses in breast, colorectal, kidney and lung cancer was 90.3, 99.6, 98.8 and 86.6%, respectively (Figure 2).

In addition, for the validation of the TOD-CUP algorithm, we analyzed a curated data set that included four cancer types analyzed on four different microarray platforms (Figure 2). The first breast cancer data set [20] included 86 breast cancer samples generated by the Affymetrix GPL96 platform. We were able to correctly predict cancer cell types in 78/86 samples (90.7% accuracy). In the second breast cancer data set (i.e. 1904 breast cancer samples from METABRIC project [21] generated by the Illumina HumanHT-12 platform), we predicted 1788/1904 as breast cancer type and 21 as two-cancer-type. Upon further investigation, of the 21 two-cancer-type, all had breast cancer type as the first or second ranked classification among all the cancer types. The total accuracy of prediction in the METABRIC data set is 95.0%. The colorectal cancer data set [22], which included 192 colorectal cancer samples generated by the Affymetrix GPL570 platform, yielded 180/192 correctly predicted colorectal cancer type classifications for a total accuracy of 94.8%. This data set had two samples predicted as two-cancer-type classifications of which both samples had the colorectal cancer type as the highest predicted score among all the cancer types. The liver cancer data set which included 64 liver cancer samples generated by the Affymetrix GPL13158 platform gave 58/64 correctly predicted liver cancer type classifications (93.8% total accuracy) and two samples as two-cancer-type classifications. The thyroid cancer data set [23] (a group of 31 thyroid cancer samples generated by the Affymetrix GPL570 platform) had all samples correctly predicted as thyroid cancer types (100% accuracy).

External validation of the TOD-CUP algorithm: metastatic cancer RNA-seq data sets

To further evaluate the performance of the TOD-CUP algorithm for tissue of origin diagnosis of the metastatic cancer, we curated an RNA-seq data set of metastatic cancer tumor samples from five different cancer types (breast, skin, kidney, liver and prostate) (Figure 2). Four sequencing platforms (Illumina HiSeq 2000, Illumina Hiseq 2500, Illumina Ovation and Roche 454 Titanium) were used to generate these data. In the two breast cancer data sets, one used biopsies samples and the other FFPE samples.

Metastatic breast cancer data set 1 [24] generated by the Illumina HiSeq 2000 platform included 78 Korea breast cancer samples. The TOD-CUP algorithm correctly predicted 73/78 (97.4%) samples as breast cancer type and three samples as two-cancer-type. Metastatic breast cancer data set 2 [25] included three matched primary, two nodal and three liver metastatic breast tumor samples generated by the Illumina HiSeq 2000 platform. All eight primary and metastatic samples were predicted correctly. The liver cancer data set [26] generated by the Illumina HiSeq 2000 platform was comprised of 10 Chinese liver cancer samples with venous metastases of HCC. All 10 liver cancer samples were correctly predicted. The 10 metastatic melanoma cancer samples, generated by the Roche 454 Titanium platform, comprising the skin cancer data set [27] were also all correctly predicted as skin cancer type. The kidney cancer data set [28], generated by the Illumina HiSeq 2500 platform, contained 27 kidney cancer samples taken from patients with HLRCC (hereditary leiomyomatosis and renal cell cancer) metastatsis to the mediastinum. Here, 25/27 (92.6%) of the samples were correctly predicted as kidney cancer data type, with the other two samples classified as unknown cancer type because of the low predicted score. The prostate cancer data set [29] which was generated by the Illumina Ovation platform and included eight prostate cancer samples obtained from the posterior iliac crest was 87.5% accurate with 7/8 samples correctly predicted as prostate cancer data type. The lone remaining sample was classified as unknown cancer type because of the low predicted score.

Clinical samples validation of the TOD-CUP algorithm

We further used primary and metastatic cancer samples obtained from clinical patients for tissue clonal origins identification in the general Chinese population using TOD-CUP algorithm. For the primary cancers, we generated transcriptomic data from 11 cancer samples (i.e. five liver, three pancreatic and one each of head-neck, lung and breast cancer samples) (Table 4). At the same time, we also sequenced the transcriptomes from 14 metastatic cancer samples (including six metastatic colorectal cancers in the liver, two metastatic colorectal cancers in the lung, three metastatic breast cancers in the liver, two metastatic liver cancers in the lung and one metastatic lung cancer in the adrenal gland). The TOD-CUP algorithm was used to analyze all the data, including the primary and metastatic cancer samples, and we calculated the prediction accuracy of our method relative to the clinically defined cancer type.

Among the 11 primary clinical cancer samples, all of the TOD-CUP classification results are the same as those in the clinical diagnosis report. Among the 14 metastatic cancer clinical samples, 13 of the classification results are identical to that in the clinical diagnosis report. One metastatic colorectal cancer in the liver sample was misclassified as a liver cancer type. Taken together, the overall accuracy of the TOD-CUP algorithm-based classification was 100% for the 11 primary cancer samples and 92.9% for metastatic cancer samples. The clinical samples validation provides strong support that the TOD-CUP algorithm can accurately identify the tissue origin of CUP using RNA-seq data.

In summary, using a combination of differential gene expression data sets derived from different cancer cell types generated using different experimental methodologies and platforms, we provide clear and convincing evidence that our TOD-CUP algorithm is robust and accurate in its ability to yield a tissue origin diagnosis of the metastatic cancer. Using multiple forms of microarray and RNA-seq data obtained from primary cancer samples, as well as metastatic cancer samples, our TOD-CUP algorithm accurately predicted the tissue origin of tumor in both general Chinese population samples and those obtained from individuals of mixed ethnic origins. Importantly, we demonstrated the excellent performance of the TOD-CUP algorithm in a range of clinical samples generated using standard methods, thereby underscoring it broad applicability.

A Web-based TOD-CUP server for tumor tissue origin diagnosis

To allow the TOD-CUP algorithm for tumor tissue origin diagnosis to be easily accessed and readily applied in a broad range of clinical settings, we developed an online server that is publicly available (http://ibi.zju.edu.cn/todcup/) (Figure 3**A**). Users are able to select the types of the data sets (e.g. RNA-seq or microarray) to be analyzed. They are also able to select the cancer type(s) they wish to include in their analysis should they be interested in evaluating a selective range of candidate cancer types in their test samples. Alternatively, users have the option of analyzing all 24 cancer types included in the platform by simply selecting “Select_all_cancer_type” option. The user is then able to directly upload the gene expression data into the “Gene expression data input” frame to start analysis (Figure 3**B**). To improve accuracy, we recommended that users include all cancer types in the prediction analysis. However, if users already know their candidate cancer types, they could directly select the target cancer types to perform the analysis. For example, if the tumors were just found in the hepatobiliary and pancreatic system, the users can select the Bile_duct, Liver and Pancreas as the candidate cancer types for the prediction. The final cancer type prediction results will not be influenced by the selected cancer types.

The results’ report generated by the Web-based TOD-CUP analysis includes three types of content (Figure 3**C**). First, the user will be presented with the biomarker gene number identified in the cancer sample data among all the 538 genes used in the TOD-CUP method. Second, the user will be provided the final diagnostic classification of the tested sample. Three different outcomes will appear: (i) if the first-ranked cancer type have a high enough cancer type score and the score is not closed with the second-ranked cancer type, it will give the first-ranked cancer type at the results part; (ii) if the first-ranked cancer type have a high cancer type score but the score is closed with the second-ranked cancer type, it will give the “two-cancer-type” results like “first-ranked cancer type, but can’t exclude second-ranked cancer type”; and (iii) if the first-ranked cancer type have a very low cancer type score, it will give the “unknown” results like “first-ranked cancer type, but need further examination.” Although it will also report the cancer type which had highest cancer type score, the results will not be credible in this case.

Table 4. Clinical samples' validation of the TOD-CUP based on RNA-seq data sets. A. Clinical samples' validation of the TOD-CUP based on primary cancer samples. B. Clinical samples' validation of the TOD-CUP based on metastatic cancer samples

A. Primary cancer samples

Patient ID	Tissue origin of primary cancer	Predicted cancer type	Results
P1	Liver	Liver	√
P2	Liver	Liver	√
P3	Liver	Liver	√
P4	Liver	Liver	√
P5	Liver	Liver	√
P6	Pancreas	Pancreas	√
P7	Pancreas	Pancreas	√
P8	Pancreas	Pancreas	√
P9	Lung	Lung	√
P10	Head_neck	Head_neck	√
P11	Breast	Breast	√

B. Metastatic cancer samples

Patient ID	Tissue origin of metastatic cancer	Tissue of samples collected from	Predicted cancer type	Results
M1	Colorectal	Liver	Colorectal	√
M2	Colorectal	Liver	Colorectal	√
M3	Colorectal	Liver	Colorectal	√
M4	Colorectal	Liver	Colorectal	√
M5	Colorectal	Liver	Colorectal	√
M6	Colorectal	Liver	Liver	×
M7	Colorectal	Lung	Colorectal	√
M8	Colorectal	Lung	Colorectal	√
M9	Breast	Liver	Breast	√
M10	Breast	Liver	Breast	√
M11	Breast	Liver	Breast	√
M12	Liver	Lung	Liver	√
M13	Liver	Lung	Liver	√
M14	Lung	Adrenal gland	Lung	√

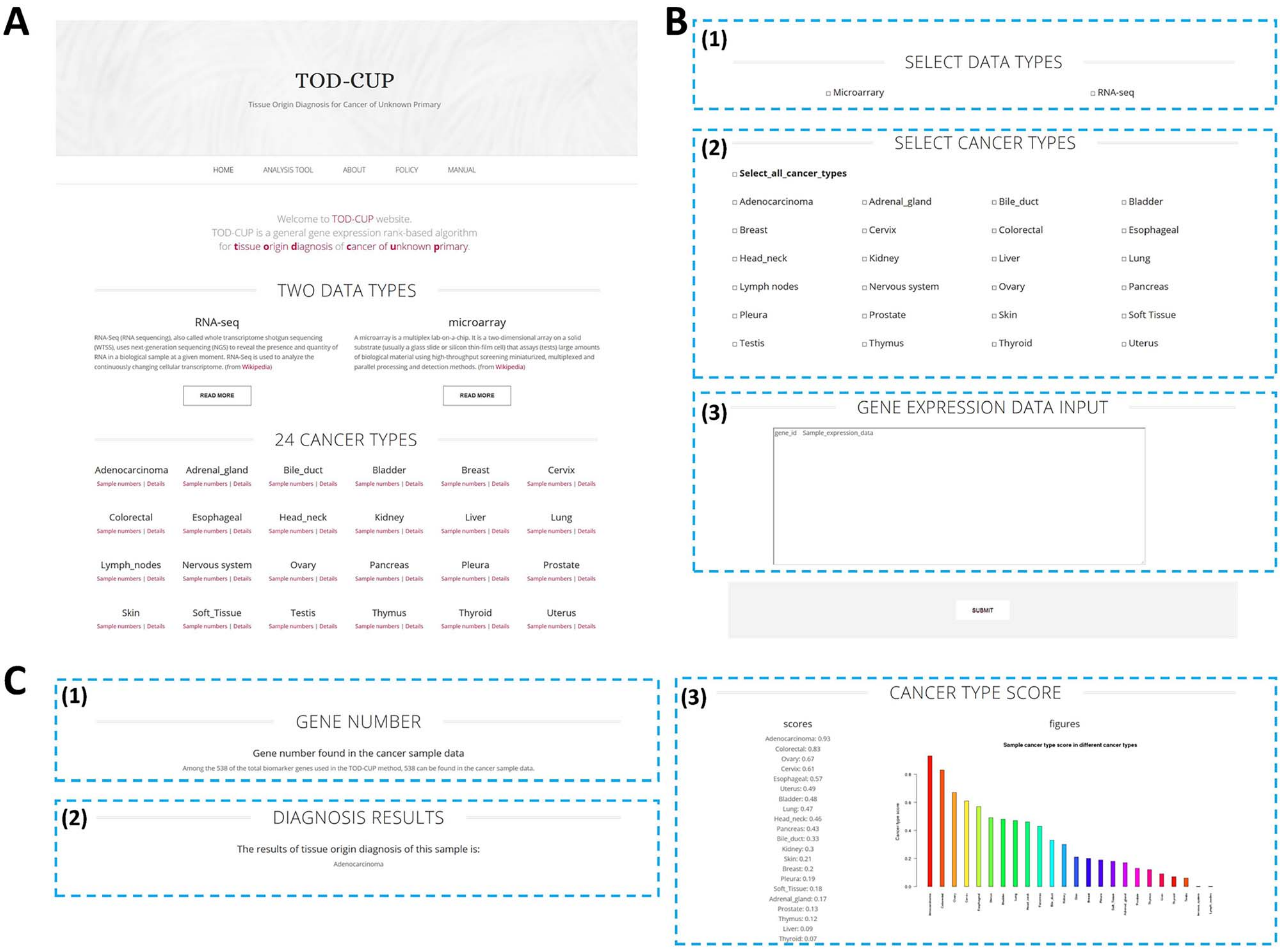
The final information presented to the user will be the predicted cancer type score of each cancer type of the cancer sample, ranked from the highest to the lowest and a bar plot to visualize the cancer type score results. An example report resulting from a TOD-CUP algorithm analysis will also be included in the "About" section on the Website.

Discussion

In this study, we developed an effective and efficient computational tool based on gene expression rank to accurately identify the tissue clonal origin of tumors in 24 cancer types across different data types. External validation based on analyzing microarray and transcriptomic data from primary and metastatic cancer tumor samples, TOD-CUP algorithm has a higher success rate in predicting the tissue origin of multiple cancer types comparing with previous studies [11, 13, 14]. Previous studies had shown that gene expression patterns remain consistent with tissue of origin, both in cell lines [33] and tumor samples [34–36]. Therefore, gene expression data may enable an accurate identification of the tissue origin of a tumor, implying that the gene expression data could be developed into a clinically useful diagnostic test. The results of Ramaswamy *et al.*'s study [35] further indicated that many cancers retain their tissue of origin identity throughout metastatic evolution, suggesting that gene expression-based approaches to the diagnosis of clinically problematic metastases of unknown primary origin [37] are feasible.

The TOD-CUP algorithm also has a good performance on chemotherapy-treated patient samples. Among the TCGA samples, there are 69 chemotherapy-treated samples (0.65% of the total 10 553 samples) which included skin, bladder, kidney, head-neck, breast, lung, thyroid, colorectal and nervous system cancer samples (see Supplementary Table S7 available online at https://academic.oup.com/bib). In the 69 chemotherapy samples, 68 (68/69, 98.55%) samples were corrected predicted by TOD-CUP algorithm and only one lung cancer sample was misclassified as soft tissue cancer type. The results suggested that the chemotherapy-treated samples might also share the same transcriptional characteristics with the untreated samples. Therefore, the chemotherapy treatments have limited effect on the prediction performance of the TOD-CUP algorithm.

The TOD-CUP algorithm is based on the gene expression rank in samples making the method less platform-specific and less sample-type limited. We used massive tissue-specific RNA-seq data from TCGA as the training data to identify 538 feature genes across 24 cancer types. The results of the external validation employing 3306 microarray data sets clearly demonstrates that the classifier trained by TCGA RNA-seq data sets accurately predicts the tissue origin of tumor samples data from both RNA-seq and microarray data types. In other words, compared with previous methods based on gene expression signatures [11–15],



Welcome to TOD-CUP website.
TOD-CUP is a general gene expression rank-based algorithm
for **t**issue **o**rigin **d**iagnosis of **c**ancer of **u**nknown **p**rimary.

TWO DATA TYPES

RNA-seq

RNA-Seq (RNA sequencing), also called whole transcriptome shotgun sequencing (WTSS), uses next-generation sequencing (NGS) to reveal the presence and quantity of RNA in a biological sample at a given moment. RNA-Seq is used to analyze the continuously changing cellular transcriptome. (from Wikipedia)

microarray

A microarray is a multiplex lab-on-a-chip. It is a two-dimensional array on a solid substrate (usually a glass slide or silicon thin-film cell) that assays (tests) large amounts of biological material using high-throughput screening miniaturized, multiplexed and parallel processing and detection methods. (from Wikipedia)

24 CANCER TYPES

Adenocarcinoma Sample numbers | Details	Adrenal_gland Sample numbers | Details	Bile_duct Sample numbers | Details	Bladder Sample numbers | Details	Breast Sample numbers | Details	Cervix Sample numbers | Details
Colorectal Sample numbers | Details	Esophageal Sample numbers | Details	Head_neck Sample numbers | Details	Kidney Sample numbers | Details	Liver Sample numbers | Details	Lung Sample numbers | Details
Lymph_nodes Sample numbers | Details	Nervous system Sample numbers | Details	Ovary Sample numbers | Details	Pancreas Sample numbers | Details	Pleura Sample numbers | Details	Prostate Sample numbers | Details
Skin Sample numbers | Details	Soft_Tissue Sample numbers | Details	Testis Sample numbers | Details	Thymus Sample numbers | Details	Thyroid Sample numbers | Details	Uterus Sample numbers | Details

SELECT DATA TYPES

SELECT CANCER TYPES

GENE EXPRESSION DATA INPUT

GENE NUMBER

Gene number found in the cancer sample data

Among the 538 of the total biomarker genes used in the TOD-CUP method, 538 can be found in the cancer sample data.

DIAGNOSIS RESULTS

The results of tissue origin diagnosis of this sample is:

Adenocarcinoma

CANCER TYPE SCORE

scores

Adenocarcinoma: 0.93
Colorectal: 0.83
Ovary: 0.67
Cervix: 0.61
Esophageal: 0.57
Uterus: 0.49
Bladder: 0.48
Lung: 0.47
Head_neck: 0.46
Pancreas: 0.43
Bile_duct: 0.33
Kidney: 0.3
Skin: 0.21
Breast: 0.2
Pleura: 0.19
Soft_Tissue: 0.18
Adrenal_gland: 0.17
Prostate: 0.13
Thymus: 0.12
Liver: 0.09
Thyroid: 0.07

figures

Figure 3. The TOD-CUP Web-based analysis server for tumor tissue origin diagnosis. **A**. The home page of TOD-CUP Web server. **B**. The analysis page of TOD-CUP Web server. (1) Select data types which need to be analyzed, i.e. microarray or RNA-seq data. (2) Select cancer types which you are interested in. If option "Select_all_cancer_type" is selected, the analysis will include all the 24 cancer types. (3) Upload the gene expression data into the "Gene expression data input" frame to start analysis. **C**. The diagnosis results of TOD-CUP. (1) The biomarker gene number found in the cancer sample data. (2) The diagnosis results of the sample. (3) The cancer type score of each cancer type.

our algorithm, which employs gene expression rank information, more accurately identifies the tissue origin of cancers and is independent of the batch effect and data types effects of other methods.

In addition, to further increase the precision of our approach, we introduced two categories ("unknown" and "two cancer types") to represent samples that cannot be classified into any of the 24 cancer types. The use of "Two cancer types" as a classification significantly increased the precision of prediction for many cancer types including bile-duct, bladder, cervix and head-neck types. After further investigation, we found that the bladder, cervix, head-neck and lung cancer types were all related to squamous-cell carcinoma, which could be a reason underlying misclassification among these cancer types. The similarity of tissue origin might also be the cause for difficulty in classification between liver and bile duct cancer types. As constructed, the TOD-CUP algorithm can accurately decide whether to give a definitive result or to remain ambiguous in distinguishing between two top-ranked candidate cancer types to avoid giving misleading diagnosis in the analysis.

Most of the cancer samples used in our study as the training data set and as the external validation of the accuracy of our computational method were obtained from public databases, which are largely taken from western populations of mixed ethnicity. To ensure the broadest applicability and accuracy of our algorithm, rather than simply rely on these data, we also sequenced 25 cancer samples (11 primary cancer samples and 14 metastatic cancer samples) specifically from individuals of Chinese ethnicity collected locally. The accuracy of our method was high regardless of the (known or unknown) ethnic origin of the individual from which the sample was derived.

While it is difficult to know exactly what leads to misclassification in our analysis, we were able to gain some insights specifically on this matter in our analysis of the cancer samples collected locally. The results suggested that based on our algorithm, the accuracy for the tissue origin of the cancer samples was higher than 96%. All of the samples obtained were derived from histologically confirmed origin in six cancer types. However, one of the 14 clinical metastatic cancer samples with histologically confirmed origin failed to be identified correctly in our study. It was a liver metastatic colon cancer sample which was misjudged as liver cancer. It is likely that normal liver tissue contamination in the bulk sample might relate with the incorrect classification for this sample. Additionally, another five of the six liver metastatic colorectal cancer samples were all accurately classified as colorectal cancer type, suggesting that our algorithm had high cancer specificity and could conquer the problem of the influence of carcinoma adjacent tissues. To avoid the influence of the normal tissue to the prediction results, it is recommended to use some software, such as "estimate" [38], to infer the tumor purity of each sample before performing the prediction analysis. We will address the problem for low tumor purity samples in our future work. In addition, the development of novel liquid biopsy methods could help to detect extremely low circulating cancer cells in the blood of patients [39–43]. Anyway, any new experimental technologies will be helpful for us to find new computational methods to detect accurately the tissue origin of metastatic cancers.

Despite the low rate of false predictions (~2.5%) presently obtained with using the TOD-CUP algorithm, we strongly feel that the method offers a beneficial and easily applied alternative to pathology alone to assist clinicians identify tumor origins more objectively and precisely. Thus, the combination of histology and gene expression-based technologies offers the best case scenario for providing patents diagnosed with CUP, the best possible information for personalized treatment.

Key Points

- A gene expression rank-based majority vote algorithm was developed for the tissue origin diagnosis of cancers of unknown primary (TOD-CUP) of most common cancer types.
- The TOD-CUP algorithm provides a powerful and robust means to accurately identify the tissue origin of 24 cancer types across different data platforms.
- The TOD-CUP algorithm could be easily accessible for clinical application through a Web-based server for tumor tissue origin diagnosis (http://ibi.zju.edu.cn/todcup/).

Supplementary data

Supplementary data mentioned in the text are available to subscribers in *BIOLRE* online.

Author's contributions

Conception and design: W. Jiang, L. Fan.

Development of methodology: Y. Shen, W. Jiang, L. Fan.

Acquisition of data (acquired and managed patients, provided facilities, etc.): W. Jiang, Y. He, Y. Wang, W. Fang,

Analysis and interpretation of data (e.g. statistical analysis, biostatistics, computational analysis): Y. Shen, Q. Chu, X. Yin, P. Bai, W. Jiang, L. Fan.

Writing, review and/or revision of the manuscript: Y. Shen, W. Jiang, L. Fan, M. P. Timko.

Study supervision: L. Fan, W. Jiang.

Funding

National Natural Science Foundation of China (81802350).

References

1. 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**:366–72.
2. Varadhachary GR, Raber MN. Cancer of unknown primary site. *N Engl J Med* 2014;**371**:757–65.
3. Pavlidis N, Fizazi K. Cancer of unknown primary (CUP). *Crit Rev Oncol Hematol* 2005;**54**:243–50.
4. Pentheroudakis G, Stoyianni A, Pavlidis N. Cancer of unknown primary patients with midline nodal distribution: midway between poor and favourable prognosis? *Cancer Treat Rev* 2011;**37**:120–6.
5. Kato S, Krishnamurthy N, Banks KC, *et al.* Utility of genomic analysis in circulating tumor DNA from patients with carcinoma of unknown primary. *Cancer Res* 2017;**77**:4238–46.
6. Ross JS, Wang K, Gay L, *et al.* Comprehensive genomic profiling of carcinoma of unknown primary site: new routes to targeted therapies. *JAMA Oncol* 2015;**1**:40–9.
7. Flaherty KT, Puzanov I, Kim KB, *et al.* Inhibition of mutated, activated BRAF in metastatic melanoma. *N Engl J Med* 2010;**363**:809–19.
8. Mao M, Tian F, Mariadason JM, *et al.* Resistance to BRAF inhibition in BRAF-mutant colon cancer can be overcome with PI3K inhibition or demethylating agents. *Clin Cancer Res* 2013;**19**:657–67.
9. Thomas H, Ilias P, Dimitrios P, *et al.* Psychiatric manifestations, personality traits and health-related quality of life in cancer of unknown primary site. *Psychooncology* 2013;**22**:2009–15.
10. Horlings HM, van Laar RK, Kerst J-M, *et al.* Gene expression profiling to identify the histogenetic origin of metastatic adenocarcinomas of unknown primary. *J Clin Oncol* 2008;**26**:4435–41.
11. 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**:4442–8.
12. 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**:320–9.
13. Ma X-J, 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**:465–73.
14. 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**:4031–40.
15. Jiang W, Shen Y, Ding Y, *et al.* A naive Bayes algorithm for tissue origin diagnosis (TOD-Bayes) of synchronous multifocal tumors in the hepatobiliary and pancreatic system. *Int J Cancer* 2018;**142**:357–68.
16. Rosenfeld N, Aharonov R, Meiri E, *et al.* MicroRNAs accurately identify cancer tissue origin. *Nat Biotechnol* 2008;**26**:462.
17. Moran S, Martínez-Cardús A, Sayols S, *et al.* Epigenetic profiling to classify cancer of unknown primary: a multicentre, retrospective analysis. *Lancet Oncol* 2016;**17**:1386–95.
18. Penson A, Camacho N, Zheng Y, *et al.* Development of genome-derived tumor type prediction to inform clinical cancer care. *JAMA Oncol* 2019.
19. Weinstein JN, Collisson EA, Mills GB, *et al.* The cancer genome atlas pan-cancer analysis project. *Nat Genet* 2013;**45**:1113.
20. Hatzis C, Sun H, Yao H, *et al.* Effects of tissue handling on RNA integrity and microarray measurements from resected breast cancers. *J Natl Cancer Inst* 2011;**103**:1871–83.

21. Curtis C, Shah SP, Chin S-F, *et al*. The genomic and transcriptomic architecture of 2000 breast tumours reveals novel subgroups. *Nature* 2012;**486**:346.
22. Tsukamoto S, Ishikawa T, Iida S, *et al*. Clinical significance of osteoprotegerin expression in human colorectal cancer. *Clin Cancer Res* 2011;**17**:2444–50.
23. He H, Jazdzewski K, Li W, *et al*. The role of microRNA genes in papillary thyroid carcinoma. *Proc Natl Acad Sci* 2005;**102**:19075–80.
24. Lee J-H, Zhao X-M, Yoon I, *et al*. Integrative analysis of mutational and transcriptional profiles reveals driver mutations of metastatic breast cancers. *Cell Discov* 2016;**2**: 16025.
25. McBryan J, Fagan A, McCartan D, *et al*. Transcriptomic profiling of sequential tumors from breast cancer patients provides a global view of metastatic expression changes following endocrine therapy. *Clin Cancer Res* 2015;**21**: 5371–9.
26. Zhang H, Ye J, Weng X, *et al*. Comparative transcriptome analysis reveals that the extracellular matrix receptor interaction contributes to the venous metastases of hepatocellular carcinoma. *Cancer Genet* 2015;**208**:482–91.
27. Valsesia A, Rimoldi D, Martinet D, *et al*. Network-guided analysis of genes with altered somatic copy number and gene expression reveals pathways commonly perturbed in metastatic melanoma. *PLoS One* 2011;**6**:e18369, e18369.
28. Sciacovelli M, Gonçalves E, Johnson TI, *et al*. Fumarate is an epigenetic modifier that elicits epithelial-to-mesenchymal transition. *Nature* 2016;**537**:544.
29. Sowalsky AG, Xia Z, Wang L, *et al*. Whole transcriptome sequencing reveals extensive unspliced mRNA in metastatic castration-resistant prostate cancer. *Mol Cancer Res* 2015;**13**:98–106.
30. Geman D, d'Avignon C, Naiman DQ, *et al*. Classifying gene expression profiles from pairwise mRNA comparisons. *Stat Appl Genet Mol Biol* 2004;**3**:1–19.
31. Tan AC, Naiman DQ, Xu L, *et al*. Simple decision rules for classifying human cancers from gene expression profiles. *Bioinformatics* 2005;**21**:3896–904.
32. Hair JF, Black WC, Babin BJ, *et al*. *Multivariate Data Analysis*. NJ: Prentice hall Upper Saddle River, 1998.
33. Ross DT, Scherf U, Eisen MB, *et al*. Systematic variation in gene expression patterns in human cancer cell lines. *Nat Genet* 2000;**24**:227.
34. Khan J, Wei JS, Ringner M, *et al*. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nat Med* 2001;**7**:673.
35. Ramaswamy S, Tamayo P, Rifkin R, *et al*. Multiclass cancer diagnosis using tumor gene expression signatures. *Proc Natl Acad Sci* 2001;**98**:15149–54.
36. Su AI, Welsh JB, Sapinoso LM, *et al*. Molecular classification of human carcinomas by use of gene expression signatures. *Cancer Res* 2001;**61**:7388–93.
37. Hainsworth JD, Greco FA. Treatment of patients with cancer of an unknown primary site. *N Engl J Med* 1993;**329**:257–63.
38. Yoshihara K, Shahmoradgoli M, Martínez E, *et al*. Inferring tumour purity and stromal and immune cell admixture from expression data. *Nat Commun* 2013;**4**:2612.
39. Riethdorf S, O'Flaherty L, Hille C, *et al*. Clinical applications of the CellSearch platform in cancer patients. *Adv Drug Deliv Rev* 2018;**125**:102–21.
40. Song Y, Zhu Z, An Y, *et al*. Selection of DNA aptamers against epithelial cell adhesion molecule for cancer cell imaging and circulating tumor cell capture. *Anal Chem* 2013;**85**:4141–9.
41. Martin JA, Phillips JA, Parekh P, *et al*. Capturing cancer cells using aptamer-immobilized square capillary channels. *Mol BioSyst* 2011;**7**:1720–7.
42. Sheng W, Chen T, Kamath R, *et al*. Aptamer-enabled efficient isolation of cancer cells from whole blood using a microfluidic device. *Anal Chem* 2012;**84**:4199–206.
43. Ahmadyousefi Y, Malih S, Mirzaee Y, *et al*. Nucleic acid aptamers in diagnosis of colorectal cancer. *Biochimie* 2019;**156**:1–11.