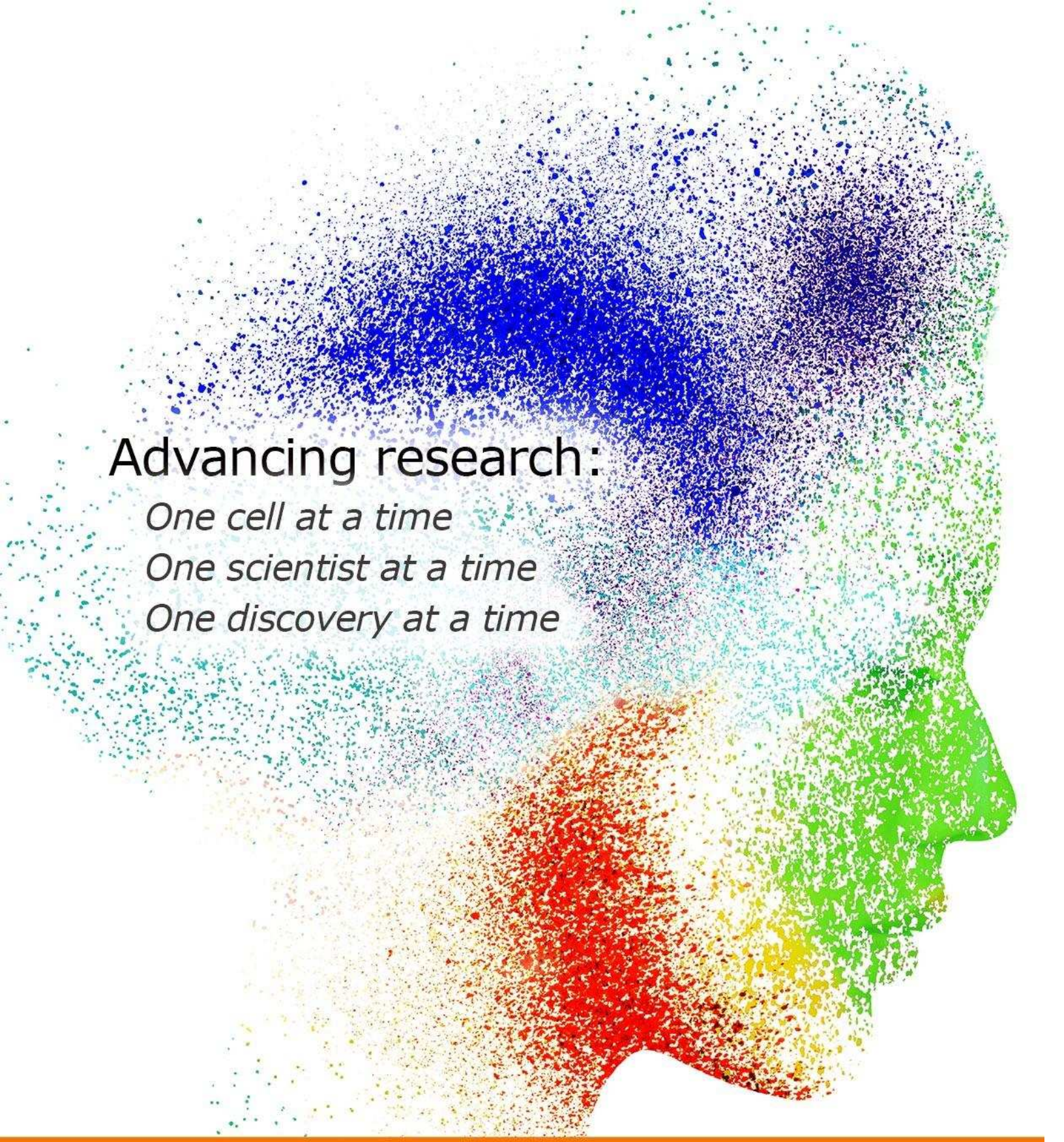


**Proven solutions that further science**

BD Accuri™ C6 Plus
BD FACSCelesta™
BD LSRFortessa™

**Discover more>**

BD

www.bdbiosciences.com/us/go/research-solutions

# A naive Bayes algorithm for tissue origin diagnosis (TOD-Bayes) of synchronous multifocal tumors in the hepatobiliary and pancreatic system

Weiqin Jiang[1,*], Yifei Shen[2,3,*], Yongfeng Ding[4,*], Chuyu Ye[2,3], Yi Zheng[1], Peng Zhao[1,5],Lulu Liu[1], Zhou Tong[1],Linfu Zhou[6], Shuo Sun[2,3], Xingchen Zhang[2,3], Lisong Teng[4,5],Michael P. Timko[7], Longjiang Fan[2,3#], Weijia Fang[1,5#]

[1]Cancer Biotherapy Center, First Affiliated Hospital, Zhejiang University, China.

[2]Institute of Bioinformatics &IBM Bio-computational Laboratory, Zhejiang University, China.

[3]Research Center for Air Pollution and Health, Zhejiang University, Hangzhou 310058, China

[4]Department of Surgical Oncology, First Affiliated Hospital, Zhejiang University, China.

[5]Key Laboratory of Precision Diagnosis & Treatment for Hepatobiliary & Pancreatic Tumor of Zhejiang Province, First Affiliated Hospital, Zhejiang University, China.

[6]Medical Biotechnology Laboratory, Zhejiang University, China.

[7]Departments of Biology and Public Health Science, University of Virginia, Charlottesville, VA 22904, USA.

*These authors contributed equally to this work.

#Correspondence to:



This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process which may lead to differences between this version and the Version of Record. Please cite this article as an 'Accepted Article', doi: 10.1002/ijc.31054

Weijia Fang

Email: weijiafang@zju.edu.cn

Tel:0086-571-87236858

Longjiang Fan

Email: fanlj@zju.edu.cn

Tel: 0086-571-88982730

**CONFLICTS OF INTEREST**

The authors have no conflicts of interest to disclose.

**Keywords:**

Synchronous multifocal tumors; Tissue origin; RNA-Seq; Naive Bayes algorithm; Hepatobiliary and pancreatic system

**Novelty and Impact:**

We have developed an algorithm that allows the rapid and accurate tissue origin diagnosis (TOD-Bayes) of synchronous multifocal tumors across tissues based on the use of genome-wide gene data sets. We demonstrated the utility of this algorithm by accurately identifying the tissue clonal origin of synchronous multifocal tumors in a Chinese population. We also show that the TOD-Bayes can be expanded for use in the precise diagnosis of the cancer of unknown primary (CUP).

**Abstract**

Synchronous multifocal tumors are common in the hepatobiliary and pancreatic system but because of similarities in their histological features, oncologists have difficulty in identifying their precise tissue clonal origin through routine histopathological methods. To address this problem and assist in more precise diagnosis, we developed a computational approach for tissue origin diagnosis based on naive Bayes algorithm (TOD-Bayes) using ubiquitous RNA-Seq data. Massive tissue-specific RNA-Seq data sets were first obtained from The Cancer Genome Atlas (TCGA) and ~1,000 feature genes were used to train and validate the TOD-Bayes algorithm. The accuracy of the model was > 95% based on 10-fold cross validation by the data from TCGA. A total of 18 clinical cancer samples (including 6 negative controls) with definitive tissue origin were subsequently used for external validation and 17 of the 18 samples were classified correctly in our study (94.4%). Furthermore, we included as cases studies seven tumor samples, taken from two individuals who suffered from synchronous multifocal tumors across tissues, where the efforts to make a definitive primary cancer diagnosis by traditional diagnostic methods had failed. Using our TOD-Bayes analysis, the two clinical test cases were successfully diagnosed as pancreatic cancer (PC) and cholangiocarcinoma (CC), respectively, in agreement with their clinical outcomes. Based on our findings, we believe that the TOD-Bayes algorithm is a powerful novel methodology to accurately identify the

tissue origin of synchronous multifocal tumors of unknown primary cancers using RNA-Seq data and an important step towards more precision-based medicine in cancer diagnosis and treatment.

## Introduction

Synchronous multifocal tumors across tissues are common in clinic, most of which are metastatic, and a small number of which are multiple primary tumors[1-4]. Failure to make a definitive tissue origin diagnosis is a main cause of poor prognosis for these patients. As a result, doctors are often faced with the dilemma of deciding on what course of clinical management is best: immediate surgery or system treatment (e.g., chemotherapy and targeted therapy)[5-8]. In system treatment, tumors derived from different tissues require different chemotherapy regimens, and even targeted therapy of identical oncogenic mutations requires knowledge of tissue of origin[9, 10]. The accurate identification of the tissue clonal origin is the premise of precision medicine in cancer treatment.

The hepatobiliary and pancreatic system has similar cellular origins in the embryo. Cancers in this system have very similar anatomical and histological features, making identification of tissue clonal origin challenging. Traditional identification methods based on histopathology and immunohistochemistry (IHC) are often unsuccessful especially in case of synchronous multifocal tumors across tissues. For example, it is very difficult to identify intrahepatic cholangiocarcinoma (ICC) and metastatic pancreatic cancer (PC) by liver biopsy, because different primary tumors share the same biomarkers and possible expression changes in the biomarkers can occur when the tumor status shifts from primary to metastatic[11, 12]. In addition, a further complication is that about 5% of the primary cancers in the liver are combined hepatocellular-cholangiocarcinoma (cHCC-CC), composed of cells with histopathological features of both hepatocellular carcinoma (HCC) and cholangiocarcinoma (CC)[13-15]. Therefore, finding new and accurate methods to identify tissue origin of synchronous multifocal tumors in the hepatobiliary and pancreatic system is of critical need.

Researchers have been looking for other more reliable means to identify the origin of tissue of synchronous multifocal tumors. Initially, cytogenetic studies were carried out in the hopes that this approach would be beneficial[16-19]. However,

synchronous multifocal tumors are cytogenetically heterogeneous and cannot be distinguished by several chromosomal aberrations. Recently, researchers also tried to analyze tissue origin of metastatic deposits in the setting of synchronous multiple malignancies by massively parallel sequencing platform[4, 20]. Through efforts of The Cancer Genome Atlas (TCGA) Project (https://tcga-data.nci.nih.gov), six different omics datasets (i.e., DNA copy number variation, DNA methylation, mRNA expression, microRNA expression, protein expression, and somatic point mutation) of 34 different cancer types have become available. Our previous study showed that patterns of copy number variation (CNV) varied across tissue types, and subtyping of the tumors from different types based on the genomic CNVs from TCGA revealed a correlation with tissue[21]. However, recent research from Hoadley *et al.* confirmed that subtyping of tumors based on mRNA expression data (RNA-seq) had the most significant correlation with tissue origin[22].

Up to now, several bioinformatics methods (e.g., decision trees[23, 24], support vector machines[25-28] and others[29] ) have been used to analyze tissue origin of tumors using RNA-seq data. These studies are all based on microarray-based gene expression signatures to identify a 'molecular fingerprint' including tens to hundreds of genes to discriminate cancers of different tissue origin. For some types of carcinomas, the accuracy of these methods were <90%. Many of these studies avoided incorporating cancers of the hepatobiliary and pancreatic system[24, 27] because of the difficulties associated with determining the tissue origin in this system. Given this limitation, we postulated that genome-wide assessment of gene expression would be a more suitable method for accurately determining tissue clonal origin of cancer, especially given the large amount of data currently available in public databases. However, such an approach was limited by lacking of appropriate bioinformatics tools.

Therefore, we developed a naive Bayes model based algorithm, dubbed TOD-Bayes, using the top ~1,000 feature genes in the hepatobiliary and pancreatic system present in the RNA-Seq data from TCGA. By using an innovative

bioinformatics tool, we aimed to develop a genome-wide algorithm to accurately identify tissue origin of synchronous multifocal tumors across tissues.

## Materials and Methods

### Sample collection and RNA-Seq data curation

A total of 18 formalin-fixed paraffin-embedded (FFPE) cancer tissue specimens found with definitive tissue origins in the hepatobiliary and pancreatic system (3PCs, 4CCs, 5HCCs and 6 mCRCs, metastatic colorectal cancers in liver) were collected for RNA-Seq analysis and use in external validation of our generated computational method. The proportion of cancer cells in each tumor specimens was independently reviewed and evaluated by three histopathologists to confirm the tumor cell content when possible. Only samples with > 50% tumor cells were included in the analysis. Total RNAs were isolated from these samples and pair-end sequenced by Illumina HiSeq 4000. Tissue samples from two patients who suffered from synchronous multifocal tumors across tissues where definitive primary cancer diagnosis by traditional histopathology or IHC methods in the hepatobiliary and pancreatic system were not successful in diagnosis were selected for analysis as case studies of the application of the algorithm. Nucleic acid extraction and sequencing was performed as described above.

Written informed consent was obtained from the patient for genomic examination and analyses of the samples. The Internal Review Board of The First Affiliated Hospital, Zhejiang University approved the genetic analysis of the patients.

### Clinical history of two case study cancer samples

Case study 1 is a patient with a ~ 6 month history of recurrent abdominal pain and jaundice and a higher than normal level of carbohydrate antigen 199 (CA-199) (486 U/mL; upper limit of 37 U/mL). Computerized tomography revealed three separate cellular masses in his pancreatic tail, upper common bile duct, and omentum, respectively. Case study 2 was a patient found to have two synchronous nodules in the

left liver and pancreas during a routine health examination. The patient reported no visible symptoms and had no abnormal tumor markers. Both patients agreed to surgery and following the procedures three tumors were removed in the first patient (one pancreatic tumor, one common bile duct tumor, and one omentum tumor), and four tumors were found in the second patient (one liver tumor, one pancreatic tumor, and two mesenteric lymph nodes tumors). Histopathological and immunohistochemical analysis of these two cases revealed moderately differentiated adenocarcinomas with positive cytokeratin 7 (CK7) or 19 (CK19). Whether the two cases are multiple primary cancers or metastatic cancers, and the respective tissue origin of these masses are all equivocal, whether by clinical manifestation, radiology, or traditional histopathology and IHC methods.

**Statistical and bioinformatic analyses of RNA-Seq data**

RNA-Seq reads were mapped to the reference genome (hg19) using MapSplice[32]. Gene expression was quantified for the transcript algorithms corresponding to TCGA GAF 2.13, using RSEM4 and normalized within-sample to a fixed upper quartile. The publicly available RNA-Seq cancer data sets were downloaded from TCGA including three cancer types (PC: 179; CC: 36; HCC: 374) (details in Supplementary materials).

*Tissue-dependent clustering:* Consensus Cluster Plus R-package[34] was used to identify clusters in the data using 1000 iterations, 80% sample re-sampling from 2 to 20 clusters using hierarchical clustering with average inner Linkage and final Linkage and Pearson correlation as the similarity metric (details in Supplementary materials).

**The TOD-Bayes algorithm for inferring origin of synchronous multifocal tumors**

The naive Bayes algorithm employs a simplified version of the Bayes formula to decide to which class a novel instance belongs[35]. In this study, we used the naive Bayes algorithm to decide synchronous multifocal tumors belong to which cancer types. The posterior probability of each class is calculated, given the feature values

present in the instance; the instance is assigned the class with the highest probability. The equation below shows the naive Bayes formula.

$$p(C_i|v_1, v_2, \dots, v_n) = \frac{p(C_i)\prod_{j=1}^{n} p(v_j|C_i)}{p(v_1, v_2, \dots, v_n)}$$

The left side of the equation is the posterior probability of class $C_i$ given the feature values, $<v_1, v_2, \dots, v_n>$, observed in the instance to be classified. The denominator of the right side of the equation is often omitted because it is a constant which is easily computed if one requires that the posterior probabilities of the classes sum to one.

In addition, we used an effective filtering approach termed correlation-based feature selection (CFS) [36], to identify the genes highly correlated with the class but not correlated with each other for the naive Bayes. As shown in the equation below, CFS evaluates a subset of features by considering the individual predictive ability of each feature along with the degree of redundancy between them [36].

$$\mathrm{CFS}_s = \frac{K\bar{r}_{cf}}{\sqrt{k + k(k-1)\bar{r}_{ff}}}$$

$\mathrm{CFS}_s$ is the score of a feature subset *S* containing *k* features, $\bar{r}_{cf}$ is the average feature to class correlation, and $\bar{r}_{ff}$ is the average feature to feature correlation.

Three steps were included in the TOD-Bayes algorithm based analysis for identifying the clonal origin of the tumor samples in the hepatobiliary and pancreatic system (Figure 1A). **Step 1**: To discriminate whether the samples could be excluded from the hepatobiliary and pancreatic system, we calculated the item-consensus value of each sample based on the analysis of consensus cluster. If the item-consensus value

was >90%, we accepted that the sample belonged to the hepatobiliary and pancreatic system[37], otherwise, the sample was included in the "Others" category of cancer types. **Step 2**: We then identified the liver origin cancer and the pancreas and biliary duct origin of the cancer. To accomplish this we selected the most informative genes for class detection, generated a dataset including the variable genes, and measured by median absolute deviation (MAD). We then used the Consensus Clustering method[34] for unsupervised class discovery of the cancer samples from TCGA. After that, we used an effective filter approach, correlation-based feature selection (CFS) [36](Figure 1), to get the feature genes highly correlated with the class, yet uncorrelated with each other. Transcript reads were normalized with log transformation followed by quantile normalization to account for variations (e.g., differences in the amount of starting material and reported transcript units) between and within datasets. 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 method based on naive Bayes model was used to train the algorithm on 9-fold and test it on the remaining 1-fold. The accuracy of the gene expression signature-generating computational algorithm was calculated based on this algorithm (Figure 1B). After that, we tested our clinical cancer samples in the TOD-Bayes. **Step 3**: To further investigate the origins of the pancreas and biliary duct cancer types, we consensus clustered the pancreas and biliary duct cancer samples. Then we applied the method mentioned in Step 2 to further identify the tissue origin of the cancer samples. In addition, the 10-fold cross-validation method based on naive Bayes was used to calculate the accuracy of TOD-Bayes. A detailed pipeline of TOD-Bayes, including some R packages used by this study, is available at our lab website (http://ibi.zju.edu.cn/bioinplant/tools/).

## Results

### The RNA-Seq data generated and used by this study

In order to develop an accurate method to predict the clonal origin of tumor samples, publicly available RNA-Seq data from 589 known/identified samples of HCC, PC, and CC were obtained from TCGA for this study. The number of samples was sufficient for classical analysis of 10-fold cross-validation (see details in the next section). To further validate our method, a total of 18 samples, including 12 cancer samples of HCC (5), PC (3) and CC (4) and 6 mCRCs in liver as negative control were collected and sequenced by this study (Table 1).An average of 20 Gb of RNA-Seq raw data were generated for each sample (Supplementary Table 1, Supplementary Figure 1).

As case studies, synchronous multifocal hepatobiliary and pancreatic tumors (three in patient 1 and four in patient 2, respectively) were collected from two patients for which the clonal origin of the tumors could not be determined by traditional diagnostic approaches and the transcriptomes sequenced to an average of 20 Gb RNA-Seq data (Supplementary Table 1 and Supplementary Figure 1). Their RNA-Seq data were then used to predict the tissue clonal origin of each multifocal tumor using the algorithm developed by this study.

### The development and accuracy of TOD-Bayes for inferring the origin of synchronous multifocal tumors

Using RNA-Seq data from TCGA as training data, a computational algorithm for tissue origin diagnosis of cancer based on naive Bayes method(TOD-Bayes), was developed to identify the origin of the tumor samples by tissue-specific gene expression levels of hepatocellular carcinoma, pancreatic carcinoma, and cholangiocarcinoma (details see Methods). For the purpose of selecting the most informative genes for classification detection, we generated a dataset including the top ~10,000 most variable expressed genes (actual count: 9,987) measured by MAD (Figure 1) as expressed tag genes. We then used the Consensus Clustering method for unsupervised class discovery of the cancer samples from TCGA. The Consensus

Clustering method involves subsampling from a set of items and determines clustering of specified cluster counts (*k*). This resulted in the identification of two main classes among the three cancer types (Figure 2 upper A-D). We used different cluster counts (from 2 to 6) to consensus cluster the data from TCGA. The data can be correctly clustered based on the cancer types when $k = 2$ in the liver, biliary duct and pancreas samples (Figure 2A in upper panel). Based on the cumulative distribution functions (CDF) of the consensus matrices and the relative change in area under the CDF curve for $k = 2$ and other counts, a similar curve or no significant area change for $k = 2$ comparing to 3 or more higher counts were observed. We found 97.3% of the samples in the first class were HCC types and 92.6% of the samples in the second class were PC and CC types. Furthermore, we also generated the cluster-consensus for the two clusters which showed that both of the two clusters had over 98.7% of the cluster-consensus. These results suggested the HCC tumor type have a distinct gene expression patterns from the PC and CC types. Thus, we were able to discriminate HCC tumors from PC and CC tumors at the second step (Figure 1). In addition, we generated the item consensus for each sample. We selected the samples with >95% of the item-consensus in each cluster for the further analysis.

After correlation-based feature selection, we got ~1000 feature genes (actual count: 943) (Supplementary Table 2) from the expressed tag gene sets for TOD-Bayes. The accuracy of the TOD-Bayes for the two cancer types was calculated based on 10-fold cross-validation method in naive Bayes algorithm (Supplementary Table 3A). The percentage of the correctly classified instances was 96.6% (569 among the total 589 TCGA samples). The results showed that the precision of TOD-Bayes was >98.1% in the classification of HCC, and >95.7% in the classification of CC and PC (Supplementary Table 4).

To further investigate the origins of PC and CC types, we consensus clustered the PC and CC cancer samples from TCGA at the second step (Figure 1). This identified subclasses among the two cancer types (Figure 2A-D lower panel). We subsequently applied different cluster counts (from 2 to 8) to consensus cluster the data from TCGA and found that the data correctly predicts the cancer types when $k = 6$ in the CC and

PC samples (Figure 2A lower panel). Again, based on the cumulative distribution functions (CDF) of the consensus matrices and the relative change in area under the CDF curve for $k$ =6 and other counts, a similar curve or no significant area change for $k$ =6 comparing to 7 or more higher counts were observed. We found that 93.8% of the CC samples could be classified into one subcategory. The PC samples were divided in several subcategories. So we classified the cancer types in two categories (CC type or PC type) based on the cancer type. The accuracy of the TOD-Bayes method for the CC and PC types was calculated based on 10-fold cross-validation method in naive Bayes algorithm. The percentage of the correctly classified instances was >94.8% (204 among the total 215 CC and PC samples) (Supplementary Table 3B).

**External validation of the TOD-Bayes**

We used RNA-Seq data generated specifically for this study in external validation testing in order to determine whether our algorithm could accurately identify tissue clonal origins in the general Chinese population. We sequenced the transcriptomes from 12 cancer samples including 5 HCC samples, 4 CC samples and 3 PC samples and calculated the gene expression levels (Table 1). At the same time, we also sequenced the transcriptomes from 6 mCRCs in the liver as negative controls for the validation. Based on the feature genes, a heatmap including the 18 clinical samples with 589 TCGA samples was constructed by hierarchical clustering (Figure 3). A complex phylogeny was observed among tissues in the hepatobiliary and pancreatic system as previous studies and failed to provide a clear tissue origin diagnosis. We tested TOD-Bayes using all the samples including the negative control. Then, we used the method to identify the cancer samples to calculate the accuracy rate of our method in naive Bayes algorithm.

To determine whether the samples could be excluded from the hepatobiliary and pancreatic system, we calculated the item-consensus of each sample based on the two clusters in the analysis of consensus cluster. We found that all 12 cancer samples had>0.95of the item-consensus value from the classified clusters. In contrast, the

item-consensus results of the 6 mCRC samples had < 0.60 from any of the classified clusters, which are all significantly lower than the cut-off item-consensus value (0.90).Therefore, the liver metastasis from colon cancer samples (Supplementary Figure 1) were not included in the further analysis. These results suggested that the TOD-Bayes method was able to exclude the cancer types which did not belong to the liver cancer types, biliary duct cancer type and pancreas cancer types.

As described in the TOD-Bayes protocol, we first identified the HCC and the PC/CC origin cancer at the first step. Among the total 12 cancer samples, the percentage of the correctly classified instances was >91.7% (11 /12 samples) (Supplementary Table 5; for a detailed accuracy by class see Supplementary Table 6). The one outlier was an intrahepatic CC (ICC) sample that was misjudged as HCC (see Discussion for details). Among the 18 clinical samples (including 6 negative controls) used for external validation, 17 samples were classified correctly. Taken together the overall accuracy of TOD-Bayes was 94.4% in the external validation for the 18 samples (Supplementary Table 5). Overall, these findings provide strong support that our method can accurately identify the origin of tumors in the hepatobiliary and pancreatic system using RNA-Seq data.

**Two case studies of synchronous multifocal tumors**

We further applied our algorithm in two clinical cases involving patients who suffered from synchronous multifocal tumors of the hepatobiliary and pancreatic system. We first discriminated whether the cancer samples belonged to the hepatobiliary and pancreatic system using all of the 7 tested samples from 2 patients (Step 1). The results showed that the origin of all the cancer samples in the two case studies were from hepatobiliary and pancreatic system (i.e., the 0.90 item-consensus cut-off value). Therefore, we used the TOD-Bayes to predict their origin.

In Case study 1, the patient suffered from synchronous multifocal tumors in three tissues. As shown in Figure 2, we found the data can be correctly classified based on the cancer types when $k = 2$ and 6 in HCC, CC and PC samples. So we used $k = 2$ and 3 to consensus cluster the samples at first step, and $k = 6$ and 7 to consensus cluster

the samples in case studies at second step. All three of the patient's tumors were classified as a type of PC by consensus clustering analysis and the cut-off value of item-consensus (Figure 4A, B), indicating that the patient likely had pancreatic cancer that metastasized to the biliary duct and omentum. We subsequently applied TOD-Bayes to identify the origin of three tumor tissues in this case, and all of the tumor tissues were discriminated as PC consistent with the consensus clustering analysis. Taken together, our data suggested that the tissue clonal origin of the three metastatic tumors in Case study1 was the pancreas.

In Case study 2, the patient was found to have synchronous multifocal tumors in four sites. Based on the prediction results of consensus clustering analysis all four of the patient's samples were CC type with the values of item-consensus of 0.937 (Figure 4A, B). This suggested that the patient initially had ICC which subsequently metastasized to the pancreas and mesenteric lymph nodes. When we applied TOD-Bayes to identify the tissue clonal origin of the four tumor samples in Case 2, our result also indicated that the four samples were all classified as CC type. Accordingly, we suggested that the tissue clonal origin of the four metastatic tumors in Case study 2 were in the intrahepatic biliary duct (Supplementary Figure 1).

TOD-Bayes also provides a brief report for each sample that includes data quality and identification of tumor tissue origin for the target sample (an example of report see Supplementary table 7). A summary of TOD-Bayes reports for the three tumor tissues in Case 1 is illustrated in Figure 5 (a summary for all 25 tumor tissue samples by this study are provided as Supplementary Figure 1). According to the reports, at least 57 million RNA-Seq reads have been generated and mapped into the reference genome for each tumor tissue samples of Case study 1, which covered all of the 1000 feature genes used by TOD-Bayes. The three samples of Case study 1 were all identified as CC or PC in the Step 1-2 (0.979-0.996 of the item-consensus values for $k$ = 2) and further determined as PC in Step 3 (0.942-0.959 for $k$ = 6) by TOD-Bayes.

**Discussion**

In this study, we described the development of an effective and efficient computational tool based on Bayesian classification (TOD-Bayes) to accurately identify the tissue clonal origin of synchronous multifocal tumors in hepatobiliary and pancreatic system. The overall accuracy was >95.0% for internal verification based on TCGA data. External validation based on analyzing RNA-Seq data from 18 Chinese cancer samples with four different types, achieved an overall accuracy of >94.4% (17 among 18 samples). We also applied the TOD-Bayes algorithm to judge the tissue origins of two patients who suffered from synchronous multifocal tumors in the hepatobiliary and pancreatic system but failed to make a definitive primary cancer diagnosis by traditional methods. The results showed both were metastatic cancers. The first patient died after half one year with multiple liver metastasis. If his initial diagnosis is pancreatic cancer with multiple metastases, surgery may be avoided and a more suitable systemic treatment could be initiated. The second patient accepted our tissue origin analysis one month after surgery, and was diagnosed as ICC with lymph node metastasis (invading pancreas). After chemotherapy treatment with gemcitabine and platinum, the patient continues to survive to the present. The two patients' clinical outcomes are consistent with our tissue origin diagnosis, implying the possibility of application of the TOD-Bayes algorithm in clinic.

In the case of synchronous multifocal tumors of the hepatobiliary and pancreatic system, the pathological and IHC methods tend to lose its effectiveness. In general, regardless of cell morphology or the presence of IHC markers, it remains difficult to accurately distinguish primary ICC and metastatic PC in liver[38]. Cytokeratin (CK) 7, CK19, and CK20 proteins are often detected in both PC and CC [12, 39] and human pancreatic cancer fusion #2 (HPC2) proteins are observed in 80% of PCs, and 32% of CCs. Similarly, N-cadherin was observed in 27% of PCs, and 58% of CCs, separately[40]. For these reasons we chose to first divide CC and PC into group (as described elsewhere in the text). Furthermore, prior studies aimed at tissue origin diagnosis of cancers using gene expression signatures in the hepatobiliary and pancreatic system are extremely limited and many have significant experimental

shortcomings, such as insufficient sample size from a single tumor(only 13, 10 and 17sample size for HCC, CC and PC, separately[27]), failure to include CC[28, 41] and low sensitivity for PC(38.9%)[28]. To our knowledge, this is the first RNA-Seq based computational algorithm to specifically identify tissue origin of cancers in the hepatobiliary and pancreatic system.

TOD-Bayes algorithm is based on RNA-Seq analysis but not microarray probes or markers based. We used a wide-genomic gene set including 10,000 most variable expressed genes in the consensus clustering step in TOD-Bayes. Based on the results of CFS, we selected ~1000 feature genes from the 10,000 gene set, instead of tens to hundreds of the individual marker genes for the tissue origin diagnosis of cancers. Gene expression is regulated by a complex cascade of events. Genetic and epigenetic events in cognate binding partners[42], competitive endogenous RNAs[43] and upstream regulators[44] can all contribute to aberrant expression of oncogenes, which suggesting that the diagnose of tumor tissue origin based on limited marker genes' expression were not all-inclusive and even not so accurate. In other words, compared with previous individual gene expression signatures depended methods, our algorithm used genome-wide information including more gene expression regulation factors, through which we can identify the tissue origin of cancers more accurately.

In addition, we used two steps rather than one to predict the tissue origin of tumors, basing on both unsupervised and supervised algorithms. Just as we have shown, a complex phylogeny was observed among tissues in the hepatobiliary and pancreatic system and failed to provide a clear tissue origin diagnosis based on traditional hierarchical clustering (Figure 3). We used consensus cluster method (unsupervised algorithm) to reclassify the known origin cancer samples at first. After that, we used a naive Bayes algorithm-based method (supervised algorithm) to identify the tissue origin of synchronous multifocal tumors in the hepatobiliary and pancreatic system. The naive Bayes algorithm is very well suited for examining gene expression profiles, because computation of the posterior distribution is all that is required for making the desired inferences, such as the computation of quantiles,

standard deviations, credible sets, and predictions[45]. Because of the data structure of gene expression profiles, frequentist inference using parametric algorithms does not appear feasible, and computing variances and other quantities based on asymptotic theory does not appear tenable. Thus, the naive Bayes algorithm appears to be better suited for dealing with these types of problems. Furthermore, most previous studies only included publicly available data for internal as well as external validation tests, leading to an overfitting and reducing the reliability of the algorithm.

We have performed a comparison of our protocol with the current algorithm-based methods such as that reported in the paper by Wei et al[41], which had a good performance for the tissue origin diagnosis. To do this, and be as accurate in our comparison as possible, we used the same data set and the same methodology (i.e., stepwise logistic regression-based method) described by Wei et al. By comparison, the accuracy of our TOD-Bayes method was slightly higher than that Wei et al found using the model based on the biomarker signatures in the liver (95.7% vs 93.6%) and pancreas (95.0% vs 92.0%); in the bile duct there was no significant difference between the methods (94.4% vs 95.0%) (details in Supplementary materials). In light of our findings that the TOD-Bayes method provides improvement in tissue origin diagnosis, and the potential exists that both independent methods can be used simultaneously to cross check patient diagnosis and thereby improving clinical outcomes.

In our study, we sequenced 18 cancer samples from Chinese population (four different cancer types included) rather than simply using the public data to validate the accuracy of our computational method. Because most of the cancer samples we used as the training dataset in TCGA were from Western populations. The accuracy of our method was >94.4% in identifying tissue origin of tumors with histologically confirmed origin in the hepatobiliary and pancreatic system. So our results suggested that we could accurately identify the tissue origin of the cancer samples from Chinese population on our algorithm. However, one of the eighteen clinical samples with histologically confirmed origin failed to be identified correctly in our study. It was an

ICC sample which was misjudged as HCC. In this research, the content of the tumor cells in our cancer samples were >50%. We suggested the heterogeneity in the bulk sample might have caused the incorrect classification of this sample. It also suggested that better sample preparation, such as microdissection, might improve the precision of our method. However, we could not rule out the possibility that this patient might have a cHCC-CC cancer. Additionally, another three of the four ICC samples were all accurately classified as CC type, suggesting that our algorithm had high cancer specificity and could conquer the problem of the influence of carcinoma adjacent tissues.

This new algorithm has a good expansibility because it can be trained on data from different types of cancer, making it useful in the identification of different kinds of tissue origins in addition to those of the hepatobiliary and pancreatic system. Additionally, if we included training data across more tissues, the method could be used for determining the origins of cancers of unknown primary (CUP) tissues. Based on the consensus cluster method we first calculated the value of the item-consensus for each sample to discriminate whether the sample belonged to target tumor types or "Others". Then we identified the accurate tissue origin (target tumor types included in the method) using the naive Bayes based algorithm. CUP accounts for approximately 5% of all newly diagnosed cancers[46]. As precision medicine plays a larger role in the clinical management of cancer, precise diagnosis for CUP is more urgent. The existence of type "Others" transforms the method from a "closed system" to an "open system", and in this way we can reduce the false positive rate and increase the applicability of the algorithm.

Obviously, the FFPE samples (used by this study) are always a challenge for RNA analysis because of their tendency to exhibit high degradation levels. However, they are undoubtedly the most accessible samples in clinics. Fortunately, the TOD-Bayes algorithm is both less platform-specific and less sample type limited, thereby allowing easier integration of data from multiple laboratories. Therefore, we

anticipate it having broad clinical application in synchronous multifocal tumors diagnosis in particular involving unknown primary cancers.

**Accession Numbers**

The sequence data from this article can be found in the GenBank/EMBL databases with BioProject ID: PRJNA353768.

**Tables**

**Table 1.** The numbers of samples from TCGA and clinical samples used in this study. Details about RNA-Seq data from the samples see Supplementary Table 1 and Supplementary Figure 1. HCC: hepatocellular carcinoma, PC: pancreatic cancer, CC: cholangiocarcinoma, mCRC: metastatic colorectal cancer in liver.

| Sample type | Cancer type | Sample number | Expressed genes* | Expressed tag genes* | Feature genes* | Origin |
|---|---|---|---|---|---|---|
| TCGA | HCC | 374 | 24,384 | 9,864 | 943 | Weinstein *et al.*[47] |
| | PC | 179 | 24,629 | 9,944 | 943 | Weinstein *et al.*[47] |
| | CC | 36 | 25,108 | 9,869 | 943 | Weinstein *et al.*[47] |
| | Total/average | 589 | 24,503 | 9,889 | 943 | |
| Clinical samples | HCC | 5 | 24,639 | 9,807 | 943 | This study |
| | PC | 3 | 23,729 | 9,855 | 943 | This study |
| | CC | 4 | 25,186 | 9,842 | 943 | This study |
| | mCRC† | 6 | 24,104 | 9,832 | 943 | This study |
| | Case study 1 | 3 | 23,619 | 9,880 | 943 | This study |
| | Case study 2 | 4 | 23,779 | 9,821 | 943 | This study |
| | Total/average | 25 | 24,229 | 9,835 | 943 | |

*Average number of expressed genes, expressed tag genes and expressed feature genes; expressed tag genes were used for consensus clustering analysis; feature genes were used for TOD-Bayes

†Negative control

**Legends**

**Figure 1.** Overview of TOD-Bayes algorithm for identifying the tissue clonal origin of synchronous multifocal tumors in the hepatobiliary and pancreatic system. (A) Work flow to identify the clonal origin of synchronous multifocal tumors in the hepatobiliary and pancreatic system and (B) TOD-Bayes algorithm and internal/external validation. HCC: hepatocellular carcinoma, PC: pancreatic cancer, CC: cholangiocarcinoma, MAD: Median Absolute Deviation; CFS: Correlation-based Feature Selection.

**Figure 2.** The best cluster counts ($k$) estimation for clustering HCC from PC and CC samples ($k$ = 2, upper Fig.2A-D) and PC from CC samples ($k$ = 6, lower Fig. 2A-D) based on RNA-Seq data from TCGA. **Upper A-D:**(A) The heatmap of cluster consensus matrix for $k$ = 2 for HCC, PC and CC samples. The matrix has items as both rows and columns and where consensus values range from 0 (never clustered together) to 1 (always clustered together) marked by white to dark blue. The HCC samples were clustered together (dark blue) from PC and CC samples (white); (B) The cumulative distribution functions (CDF) of the consensus matrices for $k$ =2 and other counts (indicated by different colors), estimated by a histogram of 100 bins. A similar curve of $k$ =2 to other counts was observed; (C) The cluster assignment of samples (columns) for $k$ =2 and other counts (rows) by color. The different colors correspond to the different consensus matrix class assignments; (D) The relative change in area under the CDF curve comparing $k$ and $k - 1$. A very small relative change for $k$ =4 from 3 was observed. **Lower A-D:**(A)The heatmap of the cluster consensus matrix for $k$ = 6 for PC and CC samples. PC samples were clustered into five classes while CC samples to another one; (B) The cumulative distribution functions (CDF) of the consensus matrices for $k$ =6 and other counts. A similar curve of$k$ =6 to $k$ = 7 or 8 but significant not to $k$ = 2 was observed; (C)The cluster assignment of samples (columns) for $k$ =6 and other counts (rows) by different colors;

(D) The relative change in area under the CDF curve comparing $k$ and $k-1$. A very small relative change for $k$ =7 from 6 was observed.

**Figure 3.** The heatmap of hierarchical clustering tree of the 18 clinical samples for external validation with 589 samples of the hepatobiliary and pancreatic system from TCGA based on the feature genes.

**Figure 4.** The clustering locations for the samples of two case studies with multifocal tumors. (A) The results of cluster analysis for the samples of two case studies with the best consensus count ($k$=2) for identification of HCC, CC and PC samples. The item-consensus values are indicated in Y-axis. Item-consensus values are the mean consensus of an item with all items in a particular cluster. An item has $k$ item-consensus values corresponding to each cluster at a particular $k$. These values are depicted in bar plots for each $k$. Samples are stacked bars. Item-consensus values are indicated by the heights of the colored portion of the bars, whose color corresponds to the common color scheme. Bars' rectangles are ordered by increasing value from bottom to top. (B) The results of cluster analysis for the samples of two case studies with the best consensus count ($k$= 6) for identification of CC and PC samples.

**Figure 5.** Summary of TOD-Bayes reports for three tumor tissues in Case study 1. (A) The data quality indicates how many sequencing data were generated and used for tissue origin diagnosis for the three tissues in Case study 1, including the total number of sequencing reads mapped into the human reference genome, the number of genes being mapped or detected, and the number of expressed tag genes used in TOD-Bayes analysis. (B) The results of tissue origin diagnosis for the three samples in Case study 1. The tissue origins (triangle arrow) and their item-consensus values predicted by TOD-Bayes in Steps 1-2 (upper) and Step 3(lower) were shown. C& P: the subgroup including CC and PC; C: CC; P: PC; H: HCC; O: others

**Author's Contributions**

Conception and design: Weijia Fang and Weiqin Jiang

Acquisition of data: Weiqin Jiang, Yi Zheng, Peng Zhao, Lulu Liu, Zhou Tong, Lisong Teng, Linfu Zhou

Analysis and interpretation of data: Yifei Shen, Yongfeng Ding, Shuo Sun, Xingchen Zhang, Weiqin Jiang, Michael P. Timko and Longjiang Fan

Writing, review, and/or revision of the manuscript: Weiqin Jiang, Yifei Shen, Yongfeng Ding, Longjiang Fan, Weijia Fang and Michael P. Timko

Administrative, technical, or material support: Longjiang Fan, Weijia Fang, Chuyu. Ye and Linfu Zhou

Study supervision: Longjiang Fan and Weijia Fang

**Acknowledgments**

The work was supported by the Natural Science Foundation of Zhejiang Province (No. LY15H160013), National Natural Science Foundation of China (No. 81472210), the Key Subject of Clinical Medical Engineering of Zhejiang Province (No. G3221), the Major Scientific Project of Zhejiang (No. N2016C03G1121060), the National Natural Science Foundation of China (No. 81272676), National Science and Technology Major Project of the Ministry of Science and Technology of China (No. 2013ZX09506015), Medical Scientific Project of Zhejiang Province (No. 2013KYB087) and Zhejiang Medical Association (No. N20130053).

## References

1. Liu Z, Liu C, Guo W, Li S, Bai O. Clinical analysis of 152 cases of multiple primary malignant tumors in 15,398 patients with malignant tumors. *PLoS One* 2015;**10**: e0125754.

2. Mendez LE, Atlass J. Triple synchronous primary malignancies of the colon, endometrium and kidney in a patient with Lynch syndrome treated via minimally invasive techniques. *Gynecologic oncology reports* 2016;**17**: 29-32.

3. Hale CS, Lee L, Mittal K. Triple synchronous primary gynecologic carcinomas: a case report and review of the literature. *International journal of surgical pathology* 2011;**19**: 552-5.

4. De Mattos-Arruda L, Bidard FC, Won HH, Cortes J, Ng CK, Peg V, Nuciforo P, Jungbluth AA, Weigelt B, Berger MF, Seoane J, Reis-Filho JS. Establishing the origin of metastatic deposits in the setting of multiple primary malignancies: the role of massively parallel sequencing. *Mol Oncol* 2014;**8**: 150-8.

5. Thompson WM, Oddson TA, Kelvin F, Daffner R, Postlethwait RW, Rice RP. Synchronous and metachronous squamous cell carcinomas of the head, neck and esophagus. *Gastrointestinal radiology* 1978;**3**: 123-7.

6. Abbruzzese JL, Abbruzzese MC, Lenzi R, Hess KR, Raber MN. Analysis of a diagnostic strategy for patients with suspected tumors of unknown origin. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology* 1995;**13**: 2094-103.

7. Shaw PH, Adams R, Jordan C, Crosby TD. A clinical review of the investigation and management of carcinoma of unknown primary in a single cancer network. *Clinical oncology (Royal College of Radiologists (Great Britain))* 2007;**19**: 87-95.

8. Steeg PS. Targeting metastasis. *Nat Rev Cancer* 2016;**16**: 201-18.

9. Mao M, Tian F, Mariadason JM, Tsao CC, Lemos R, Jr., Dayyani F, Gopal YN, Jiang ZQ, Wistuba, II, Tang XM, Bornman WG, Bollag G, et al. Resistance to BRAF inhibition in BRAF-mutant colon cancer can be overcome with PI3K inhibition or demethylating agents. *Clinical cancer research : an official journal of the American Association for Cancer Research* 2013;**19**: 657-67.

10. Mayers JR, Torrence ME, Danai LV, Papagiannakopoulos T, Davidson SM, Bauer MR, Lau AN, Ji BW, Dixit PD, Hosios AM, Muir A, Chin CR, et al. Tissue of origin dictates branched-chain amino acid metabolism in mutant Kras-driven cancers. *Science* 2016;**353**: 1161-5.

11. Lok T, Chen L, Lin F, Wang HL. Immunohistochemical distinction between intrahepatic cholangiocarcinoma and pancreatic ductal adenocarcinoma. *Hum Pathol* 2014;**45**: 394-400.

12. Lau SK, Prakash S, Geller SA, Alsabeh R. Comparative immunohistochemical profile of hepatocellular carcinoma, cholangiocarcinoma, and metastatic adenocarcinoma. *Hum Pathol* 2002;**33**: 1175-81.

13. Jarnagin WR, Weber S, Tickoo SK, Koea JB, Obiekwe S, Fong Y, DeMatteo RP, Blumgart LH, Klimstra D. Combined hepatocellular and cholangiocarcinoma: demographic, clinical, and prognostic factors. *Cancer* 2002;**94**: 2040-6.

14. Yano Y, Yamamoto J, Kosuge T, Sakamoto Y, Yamasaki S, Shimada K, Ojima H, Sakamoto M, Takayama T, Makuuchi M. Combined hepatocellular and cholangiocarcinoma:

a clinicopathologic study of 26 resected cases. *Japanese journal of clinical oncology* 2003;**33**: 283-7.

15. Wu CH, Yong CC, Liew EH, Tsang LL, Lazo M, Hsu HW, Ou HY, Yu CY, Chen TY, Huang TL, Concejero AM, Chen CL, et al. Combined Hepatocellular Carcinoma and Cholangiocarcinoma: Diagnosis and Prognosis After Resection or Transplantation. *Transplantation proceedings* 2016;**48**: 1100-4.

16. Ilson DH, Motzer RJ, Rodriguez E, Chaganti RS, Bosl GJ. Genetic analysis in the diagnosis of neoplasms of unknown primary tumor site. *Seminars in oncology* 1993;**20**: 229-37.

17. Motzer RJ, Rodriguez E, Reuter VE, Bosl GJ, Mazumdar M, Chaganti RS. Molecular and cytogenetic studies in the diagnosis of patients with poorly differentiated carcinomas of unknown primary site. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology* 1995;**13**: 274-82.

18. Sandberg AA, Bridge JA. Updates on the cytogenetics and molecular genetics of bone and soft tissue tumors. gastrointestinal stromal tumors. *Cancer genetics and cytogenetics* 2002;**135**: 1-22.

19. Fasanella KE, Krasinskas A, Schoedel KE, Sasatomi E, Slivka A, Whitcomb DC, Sanders M, Nodit L, Raab S, McGrath KM, Ohori NP, Khalid A. DNA mutational differences in cytological specimens from pancreatic cancer and cholangiocarcinoma. *Pancreatology : official journal of the International Association of Pancreatology (IAP) [et al]* 2010;**10**: 429-33.

20. Navin N, Kendall J, Troge J, Andrews P, Rodgers L, McIndoo J, Cook K, Stepansky A, Levy D, Esposito D, Muthuswamy L, Krasnitz A, et al. Tumour evolution inferred by single-cell sequencing. *Nature* 2011;**472**: 90-4.

21. Jiang W, Ding Y, Shen Y, Fan L, Zhou L, Li Z, Zheng Y, Zhao P, Liu L, Tong Z, Fang W, Wang W. Identifying the clonal origin of synchronous multifocal tumors in the hepatobiliary and pancreatic system using multi-omic platforms. *Oncotarget* 2016.

22. Hoadley KA, Yau C, Wolf DM, Cherniack AD, Tamborero D, Ng S, Leiserson MD, Niu B, McLellan MD, Uzunangelov V, Zhang J, Kandoth C, et al. Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin. *Cell* 2014;**158**: 929-44.

23. Dennis JL, Hvidsten TR, Wit EC, Komorowski J, Bell AK, Downie I, Mooney J, Verbeke C, Bellamy C, Keith WN, Oien KA. Markers of adenocarcinoma characteristic of the site of origin: development of a diagnostic algorithm. *Clin Cancer Res* 2005;**11**: 3766-72.

24. Shedden KA, Taylor JM, Giordano TJ, Kuick R, Misek DE, Rennert G, Schwartz DR, Gruber SB, Logsdon C, Simeone D. Accurate molecular classification of human cancers based on gene expression using a simple classifier with a pathological tree-based framework. *American Journal of Pathology* 2003;**163**: 1985-95.

25. Tothill RW, Kowalczyk A, Rischin D, Bousioutas A, Haviv I, van Laar RK, Waring PM, Zalcberg J, Ward R, Biankin AV, Sutherland RL, Henshall SM, 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.

26. Tothill RW, Li J, Mileshkin L, Doig K, Siganakis T, Cowin P, Fellowes A, Semple T, Fox S, Byron K, Kowalczyk A, Thomas D, et al. Massively-parallel sequencing assists the diagnosis and guided treatment of cancers of unknown primary. *J Pathol* 2013;**231**: 413-23.

27. Tothill RW, Shi F, Paiman L, Bedo J, Kowalczyk A, Mileshkin L, Buela E, Klupacs R, Bowtell D, Byron K. Development and validation of a gene expression tumour classifier for cancer of unknown primary. *Pathology* 2015;**47**: 7-12.

28. Xu Q, Chen J, Ni S, Tan C, Xu M, Dong L, Yuan L, Wang Q, Du X. Pan-cancer transcriptome analysis reveals a gene expression signature for the identification of tumor tissue origin. *Modern pathology : an official journal of the United States and Canadian Academy of Pathology, Inc* 2016;**29**: 546-56.

29. Horlings HM, van Laar RK, Kerst JM, Helgason HH, Wesseling J, van der Hoeven JJ, Warmoes MO, Floore A, Witteveen A, Lahti-Domenici J, Glas AM, Van't Veer LJ, et al. Gene expression profiling to identify the histogenetic origin of metastatic adenocarcinomas of unknown primary. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology* 2008;**26**: 4435-41.

30. Andrews S. FastQC: A quality control tool for high throughput sequence data. *Reference Source* 2010.

31. Dai M, Thompson RC, Maher C, Contrerasgalindo R, Kaplan MH, Markovitz DM, Omenn G, Fan M. NGSQC: cross-platform quality analysis pipeline for deep sequencing data. *Bmc Genomics* 2010;**11 Suppl 4**: 1-9.

32. Wang K, Singh D, Zeng Z, Coleman SJ, Huang Y, Savich GL, He X, Mieczkowski P, Grimm SA, Perou CM. MapSplice: accurate mapping of RNA-seq reads for splice junction discovery. *Nucleic acids research* 2010;**38**: e178-e.

33. Team RDC. R: A Language and Environment for Statistical Computing. *Computing* 2011;**14**: 12-21.

34. Wilkerson MD, Hayes DN. ConsensusClusterPlus: a class discovery tool with confidence assessments and item tracking. *Bioinformatics* 2010;**26**: 1572-3.

35. Murphy KP. Naive bayes classifiers. *University of British Columbia* 2006.

36. Hall MA. Correlation-based feature selection for machine learning: The University of Waikato, 1999.

37. Monti S, Tamayo P, Mesirov J, Golub T. Consensus Clustering: A Resampling-Based Method for Class Discovery and Visualization of Gene Expression Microarray Data. *Machine Learning* 2003;**52**: 91-118.

38. Yachida S, Iacobuzio-Donahue CA. The pathology and genetics of metastatic pancreatic cancer. *Archives of pathology & laboratory medicine* 2009;**133**: 413-22.

39. Tsuji M, Kashihara T, Terada N, Mori H. An immunohistochemical study of hepatic atypical adenomatous hyperplasia, hepatocellular carcinoma, and cholangiocarcinoma with alpha-fetoprotein, carcinoembryonic antigen, CA19-9, epithelial membrane antigen, and cytokeratins 18 and 19. *Pathology international* 1999;**49**: 310-7.

40. Hooper JE, Morgan TK, Grompe M, Sheppard BC, Troxell ML, Corless CL, Streeter PR. The novel monoclonal antibody HPC2 and N-cadherin distinguish pancreatic ductal adenocarcinoma from cholangiocarcinoma. *Hum Pathol* 2012;**43**: 1583-9.

41. Wei IH, Shi Y, Jiang H, Kumar-Sinha C, Chinnaiyan AM. RNA-Seq accurately identifies cancer biomarker signatures to distinguish tissue of origin. *Neoplasia* 2014;**16**: 918-27.

42. Wang X, Haswell JR, Roberts CW. Molecular pathways: SWI/SNF (BAF) complexes are frequently mutated in cancer--mechanisms and potential therapeutic insights. *Clin Cancer Res* 2014;**20**: 21-7.

43. Sumazin P, Yang X, Chiu HS, Chung WJ, Iyer A, Llobet-Navas D, Rajbhandari P, Bansal M, Guarnieri P, Silva J, Califano A. An extensive microRNA-mediated network of RNA-RNA interactions regulates established oncogenic pathways in glioblastoma. *Cell* 2011;**147**: 370-81.

44. Chen J, Alvarez M, Talos F, Dhruv H, Rieckhof G, Iyer A, Diefes K, Aldape K, Berens M, Shen M. Identification of Causal Genetic Drivers of Human Disease through Systems-Level Analysis of Regulatory Networks. *Cell* 2014;**159**: 402-14.

45. Ibrahim JG, Chen MH, Gray RJ. Bayesian Models for Gene Expression With DNA Microarray Data. *Journal of the American Statistical Association* 2002;**97**: 88-99.

46. Pavlidis N. Forty years experience of treating cancer of unknown primary. *Acta oncologica (Stockholm, Sweden)* 2007;**46**: 592-601.

47. Weinstein JN, Collisson EA, Mills GB, Shaw KRM, Ozenberger B, Ellrott K, Shmulevich I, Sander C, Stuart JM. The cancer genome atlas pan-cancer analysis project. *Nature Genetics* 2013;**45**: 1113-20.

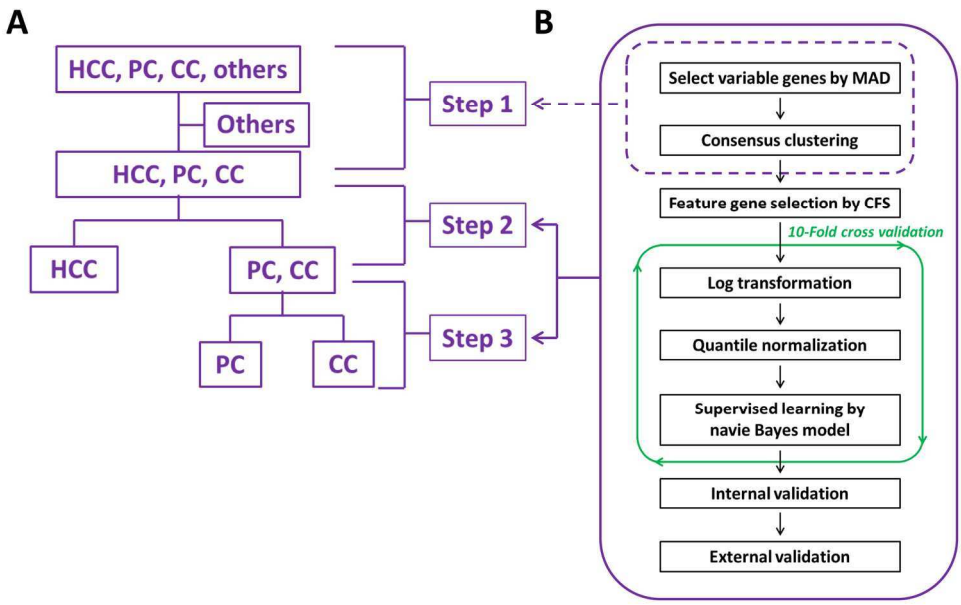


Figure 1. Overview of TOD-Bayes algorithm for identifying the tissue clonal origin of synchronous multifocal tumors in the hepatobiliary and pancreatic system. (A) Work flow to identify the clonal origin of synchronous multifocal tumors in the hepatobiliary and pancreatic system and (B) TOD-Bayes algorithm and internal/external validation. HCC: hepatocellular carcinoma, PC: pancreatic cancer, CC: cholangiocarcinoma, MAD: Median Absolute Deviation; CFS: Correlation-based Feature Selection.

180x109mm (300 x 300 DPI)

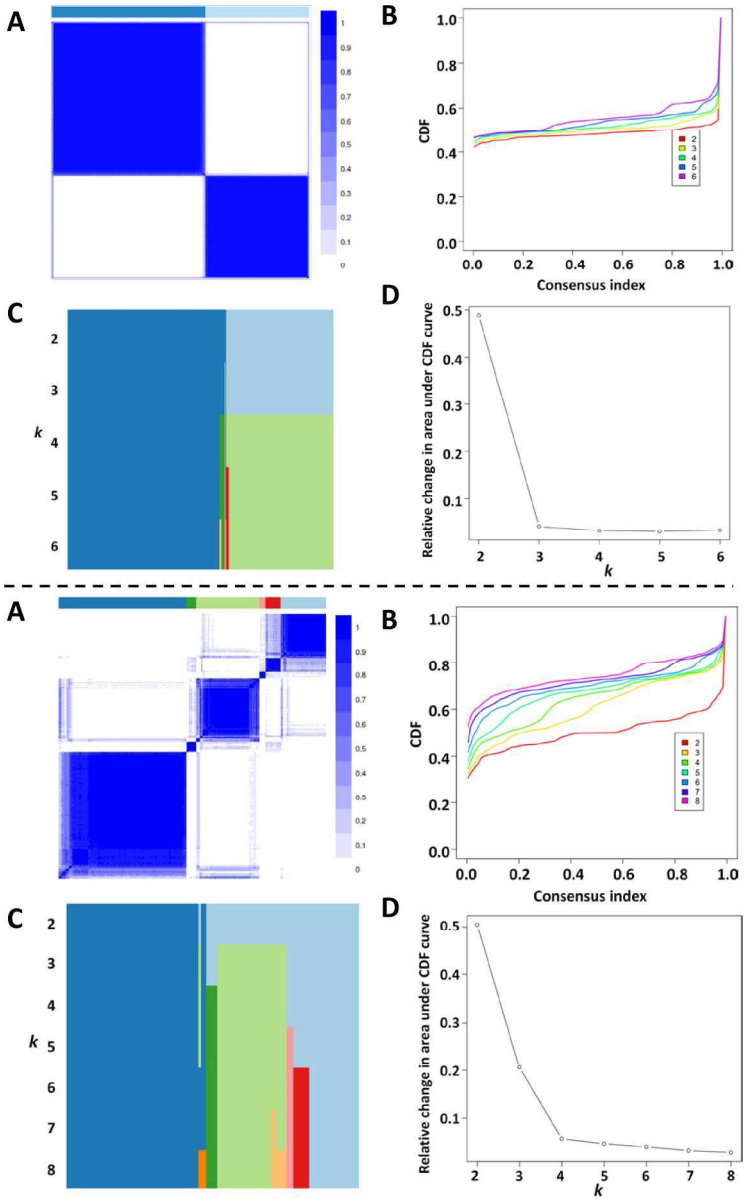


Figure 2. The best cluster counts (k) estimation for clustering HCC from PC and CC samples (k = 2, upper Fig.2A-D) and PC from CC samples (k = 6, lower Fig. 2A-D) based on RNA-Seq data from TCGA.

158x250mm (300 x 300 DPI)

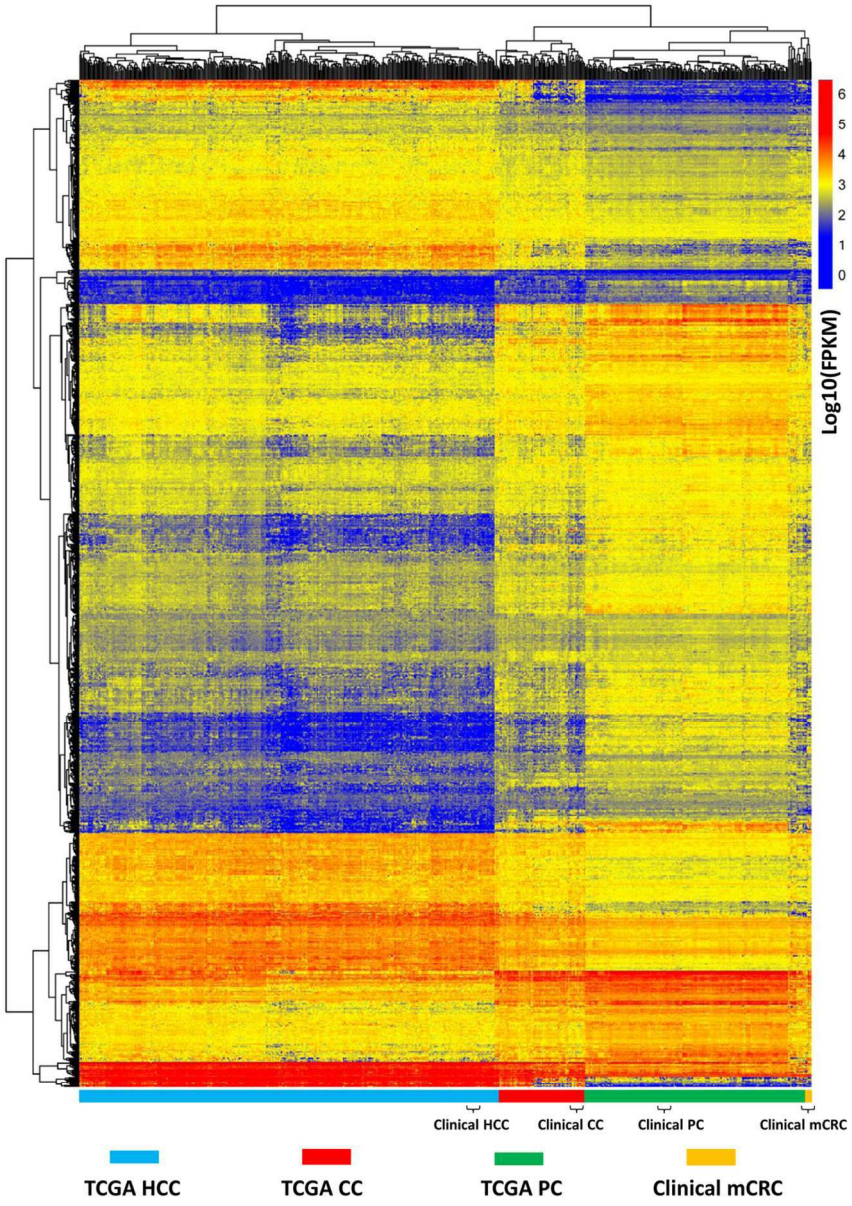


Figure 3. The heatmap of hierarchical clustering tree of the 18 clinical samples for external validation with 589 samples of the hepatobiliary and pancreatic system from TCGA based on the feature genes.

99x137mm (300 x 300 DPI)

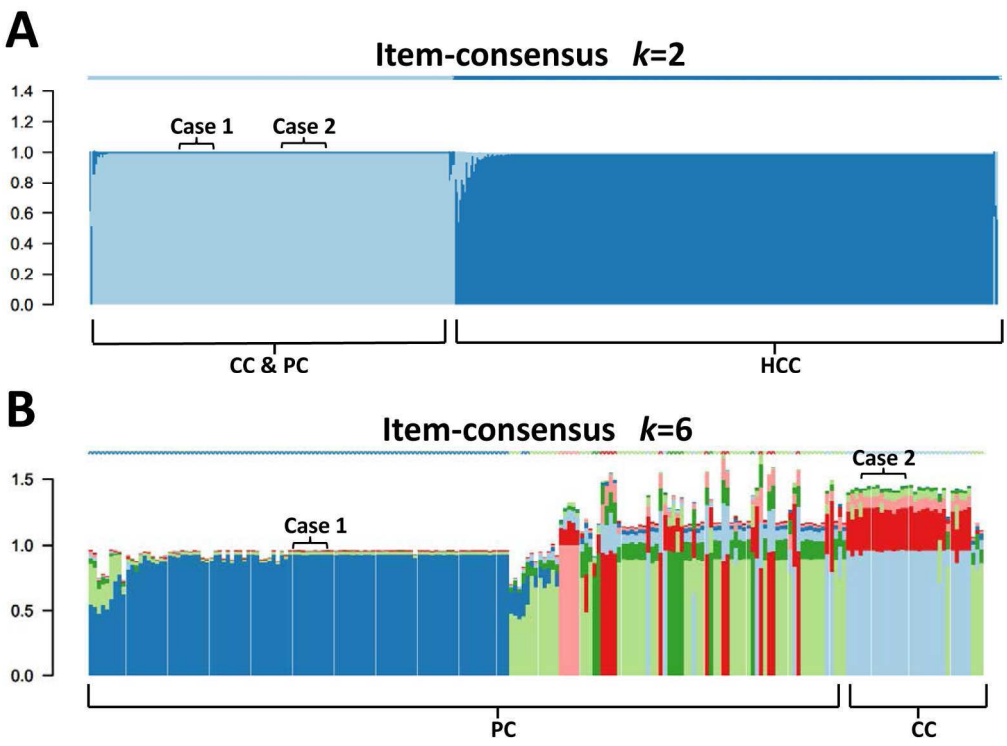


Figure 4. The clustering locations for the samples of two case studies with multifocal tumors.

176x132mm (300 x 300 DPI)

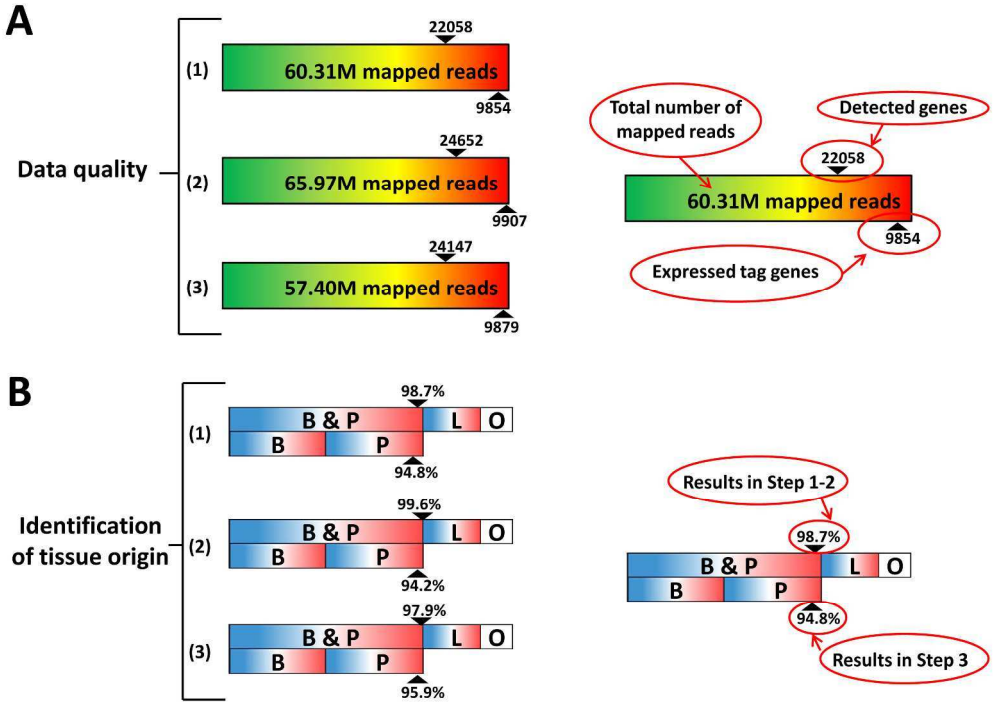


Figure 5. Summary of TOD-Bayes reports for three tumor tissues in Case study 1.

246x179mm (300 x 300 DPI)