RNA-Seq Accurately Identifies Cancer Biomarker Signatures to Distinguish Tissue of Origin[1]

Iris H. Wei*, Yang Shi†, Hui Jiang†, Chandan Kumar-Sinha‡,§ and Arul M. Chinnaiyan‡,§,¶,#,**

*University of Michigan Department of Surgery, University of Michigan Medical School, Ann Arbor, MI, USA 48109; †University of Michigan Department of Biostatistics, University of Michigan Medical School, Ann Arbor, MI, USA 48109; ‡Michigan Center for Translational Pathology, University of Michigan Medical School, Ann Arbor, MI, USA 48109; §University of Michigan Department of Pathology, University of Michigan Medical School, Ann Arbor, MI, USA 48109; ¶Comprehensive Cancer Center, University of Michigan Medical School, Ann Arbor, MI, USA 48109; #University of Michigan Department of Urology, University of Michigan Medical School, Ann Arbor, MI, USA 48109; **Howard Hughes Medical Institute, University of Michigan Medical School, Ann Arbor, MI, USA 48109

Abstract

Metastatic cancer of unknown primary (CUP) accounts for up to 5% of all new cancer cases, with a 5-year survival rate of only 10%. Accurate identification of tissue of origin would allow for directed, personalized therapies to improve clinical outcomes. Our objective was to use transcriptome sequencing (RNA-Seq) to identify lineage-specific biomarker signatures for the cancer types that most commonly metastasize as CUP (colorectum, kidney, liver, lung, ovary, pancreas, prostate, and stomach). RNA-Seq data of 17,471 transcripts from a total of 3,244 cancer samples across 26 different tissue types were compiled from in-house sequencing data and publically available International Cancer Genome Consortium and The Cancer Genome Atlas datasets. Robust cancer biomarker signatures were extracted using a 10-fold cross-validation method of log transformation, quantile normalization, transcript ranking by area under the receiver operating characteristic curve, and stepwise logistic regression. The entire algorithm was then repeated with a new set of randomly generated training and test sets, yielding highly concordant biomarker signatures. External validation of the cancer-specific signatures yielded high sensitivity (92.0% ± 3.15%; mean ± standard deviation) and specificity (97.7% ± 2.99%) for each cancer biomarker signature. The overall performance of this RNA-Seq biomarker-generating algorithm yielded an accuracy of 90.5%. In conclusion, we demonstrate a computational model for producing highly sensitive and specific cancer biomarker signatures from RNA-Seq data, generating signatures for the top eight cancer types responsible for CUP to accurately identify tumor origin.

Neoplasia (2014) 16, 918–927

Introduction

Metastatic cancer of unknown primary origin (CUP) is an important clinical dilemma, comprising 3% to 5% of all new cancer cases [1,2]. Without a firm histologic diagnosis, the clinical management of these patients varies widely [3], and despite protocol-driven guidelines, outcomes remain poor. Median survival is 6 to 9 months [4], with a 5-year survival rate of only 10% [5,6]. The role of chemotherapy in the treatment of occult primary tumors is primarily palliative and does not improve long-term survival; the National Comprehensive Cancer

Address all correspondence to: Arul M. Chinnaiyan, MD, PhD, University of Michigan Medical School, 1400 E. Medical Center Dr. 5316 CCC, Ann Arbor, MI 48109–5940. E-mail: arul@med.umich.edu

[1] Conflict of Interests: None.

Received 12 August 2014; Revised 23 September 2014; Accepted 23 September 2014

1476-5586/14
http://dx.doi.org/10.1016/j.neo.2014.09.007

Network (NCCN) panel encourages CUP patient enrollment in clinical trials when possible [7]. An accurate method for distinguishing tumor origin to tailor personalized therapies is therefore critical to the successful management of these malignancies.

Thus far, there have been a number of studies focused on identifying unique signatures that distinguish among different cancer types, using immunohistochemistry [8–11], cytogenetic studies [12–14], comparative microarray analysis [15–22], combined microarray and quantitative polymerase chain reaction techniques [23,24], bead-based miRNA profiling [25], and more recently, limited, high-throughput sequencing data combined with microarray [26]. Currently, only qRT-PCR [23,24] and microarray-based [19,22] assays are commercially available for use, with diagnostic accuracies ranging from 74 – 85%.

Compared to traditional microarray technology, transcriptome sequencing (RNA-Seq) possesses a number of advantages, including unlimited genome coverage and discovery potential, a greater than 8000-fold dynamic range for quantifying gene expression levels, and the ability to identify splice variants, unmapped genes, and unrecognized non-coding RNAs [27,28]. The rapidly decreasing cost of high-throughput sequencing methods has also improved the accessibility of these techniques for clinical application and allowed for the generation of large-scale datasets to robustly interrogate such clinical problems as CUP.

In a review of all published autopsies performed on CUP patients who died from cancer progression from 1944 to 2000, a primary tumor was successfully identified post-mortem in 73% of the 884 cases [29]. The most common tissues of origin were lung (27%), pancreas (24%), kidney (6%), colorectum (6%), stomach (5%), liver (5%), ovary (3%), and prostate (3%) [4,30–38]. Our objective was therefore to identify lineage-specific biomarker signatures for each of these cancers, using a large, multi-cancer RNA-Seq database to distinguish tissues of origin from among different cancer types.

Material and Methods

Multi-Cancer RNA-Seq Database

Paired-end RNA-Seq data for 364 cancer samples from 22 different tissue types were used to compile a multi-cancer gene expression dataset as previously described [39]. This dataset was then supplemented with publically available RNA-Seq cancer data accessed from the International Cancer Genome Consortium [40] and The Cancer Genome Atlas [41]. This included four additional cancer types (acute myeloid leukemia, endometrial cancer, head and neck squamous cancer, and lung cancer). The dataset was restricted to those transcripts commonly annotated across all three datasets. The final composite data matrix was comprised of gene expression

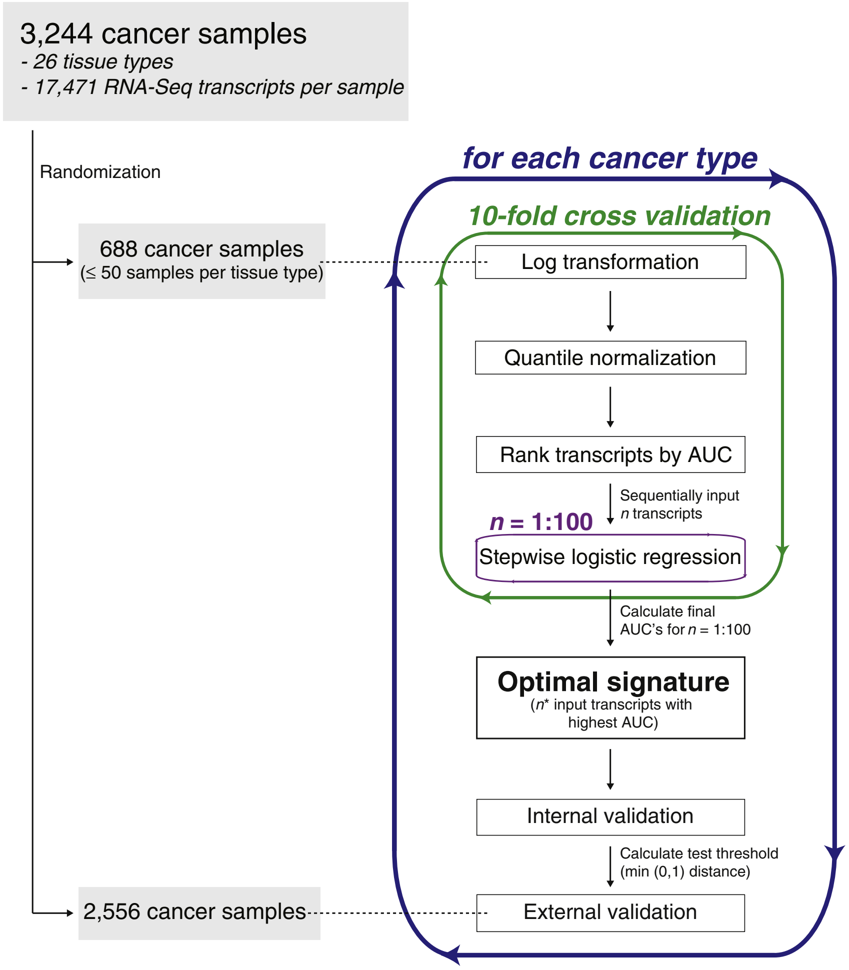

Figure 1. Algorithm for extracting optimal cancer biomarker signatures from RNA-Seq dataset. AUC, area under the receiver operating characteristic curve.

readouts for 17,471 transcripts from 3,244 cancer samples (139 cancer cell lines and 3,105 tissues) across 26 cancer types.

Model for Deriving an Optimal Biomarker Signature

R language [42] was used to program an algorithm to derive an optimal biomarker signature for a cancer type of interest (Figure 1). The 3,244 samples were randomly allocated to the training or test sets (Table 1). A maximum of 50 samples per tissue type were assigned to the training set (688 samples) and the remainder to the test set (2,556 samples). The training set was used to generate the optimal biomarker signatures for each cancer type, while the test set was reserved for final, external validation. Biomarker signatures were generated for the eight tissue types that account for approximately 80% of CUP cases (colorectum, kidney, liver, lung, ovary, pancreas, prostate, and stomach).

Transcript normalization. Transcript reads were normalized with log transformation followed by quantile normalization to account for variations between and within datasets, such as differences in the amount of starting material and reported transcript units. The entire 688-sample training set was 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 was used to train the model on 9-fold and test each signature on the remaining 1-fold.

Univariate transcript analysis. Within each 9-fold training subset, area under the receiver operating characteristic (ROC) curve values (AUC) [43] were calculated for each of the 17,471 transcripts. The transcripts were then sorted by decreasing AUC.

Stepwise logistic regression. To generate an optimal signature, an iterative approach was used, rather than inputting all 17,471 transcripts into the model, which would increase computational burden and the likelihood of overfitting. The top 100 transcripts, based on univariate AUC rank, were introduced into a stepwise logistic regression model, to determine the optimal signature at each iteration for n, between 1 and 100 input transcripts. Logistic regression was performed in both directions to optimize the Akaike information criterion (AIC) [44–46] so that at each step, it was calculated whether the current signature would be improved not only by adding the next variable but also by discarding any of the variables present within the currently optimized signature. The final signatures were used to calculate the predicted likelihood of each sample in the remaining 1-fold being of that cancer type, given n input transcripts.

Table 1. Allocation of Cancer Samples to RNA-Seq Training and Test Sets

Cancer Type	All	Training Set	Test Set
Adrenal gland	3	3	0
Acute myeloid leukemia	174	50	124
Bladder	70	50	20
Breast	864	50	814
Cervix	8	8	0
Colorectum	244	50	194
Endometrium	333	50	283
Germ cell	1	1	0
Kidney	24	24	0
Liver	15	15	0
Head and neck	263	50	213
Lung	348	50	298
Lymphoma	11	11	0
Medulloblastoma	1	1	0
Melanoma	136	50	86
Merkel cell	3	3	0
Myeloproliferative neoplasm	9	9	0
Neuroblastoma	2	2	0
Neuroepithelioma	1	1	0
Oropharynx	4	4	0
Ovary	418	50	368
Pancreas	76	50	26
Prostate	154	50	104
Rhabdomyosarcoma	1	1	0
Salivary gland	4	4	0
Stomach	77	50	27
Total	*3244*	*688*	*2556*

Biomarker signature selection. A final "cross-validated AUC" was determined for each signature generated from n transcripts, based on the calculated predictions for each sample compared to their true identities. The optimal biomarker signature was determined to be the one generated from the top n^* number of transcripts that yielded the highest cross-validated AUC. The entire 688-sample training set was then used as the input training set to generate a final, optimal biomarker signature based on the top n^* transcripts.

Internal validation. Each cancer biomarker signature was internally validated by using the entire 688-sample training set as the input. Each sample received a predicted value, m, between 0 and 1, indicating likelihood of the sample being the cancer type of interest. The predicted values were then used to generate ROC curves for each signature. Optimal score thresholds, k, (above which was defined as "positive" for that cancer type and below which was "negative") were calculated by selecting the point on the ROC curve with the minimum distance from (0,1), which represents a perfect test of 100% sensitivity and specificity [47].

External validation. Each cancer biomarker signature was then externally validated against the reserved 2,556-sample test set using the optimal score thresholds. Overall sensitivity and specificity were calculated for each cancer signature.

Duplicate cancer predictions. Each of the 2,556 cancer samples in the reserved test set was tested using each of the eight cancer biomarker signatures. Those samples that predicted positive for more than one cancer type were assigned the cancer type that had the highest *relative* predicted value, defined as $[m - k]/[1 - k]$.

Additional analysis. Graphs were plotted using GraphPad Prism. The heat map was generated with Cluster 3.0 [48] and visualized using TreeView [49]. Statistical analysis was performed using R and GraphPad Prism.

Results

The results of our biomarker-generating model are shown in Figure 2. For all eight cancer types, the maximum, cross-validated AUC was obtained within the first 100 input transcripts. Interestingly, cross-validated AUC plots of the colorectal, lung, pancreas, and stomach cancer samples yielded prominent peaks, beyond which the inclusion of additional transcripts worsened the biomarker signature's accuracy. Conversely, liver and ovarian cancer samples yielded relatively flat curves of near-perfect cross-validated AUC's, suggesting that these cancer types have such unique biomarker profiles that many highly accurate signatures may be generated. Optimal signatures for each tissue type were objectively determined by selecting the number of input transcripts, n^*, that corresponded with the maximum cross-validated AUC (Figure 2, red points).

Next, using the entire 688-sample training set as the input test set, the final list of transcripts was generated for each cancer biomarker signature, by performing stepwise logistic regression of the top n^* transcripts. The entire model was then run again using a new, random allocation of training and test samples. The final biomarker signatures for each cancer type were concordant with the signatures generated

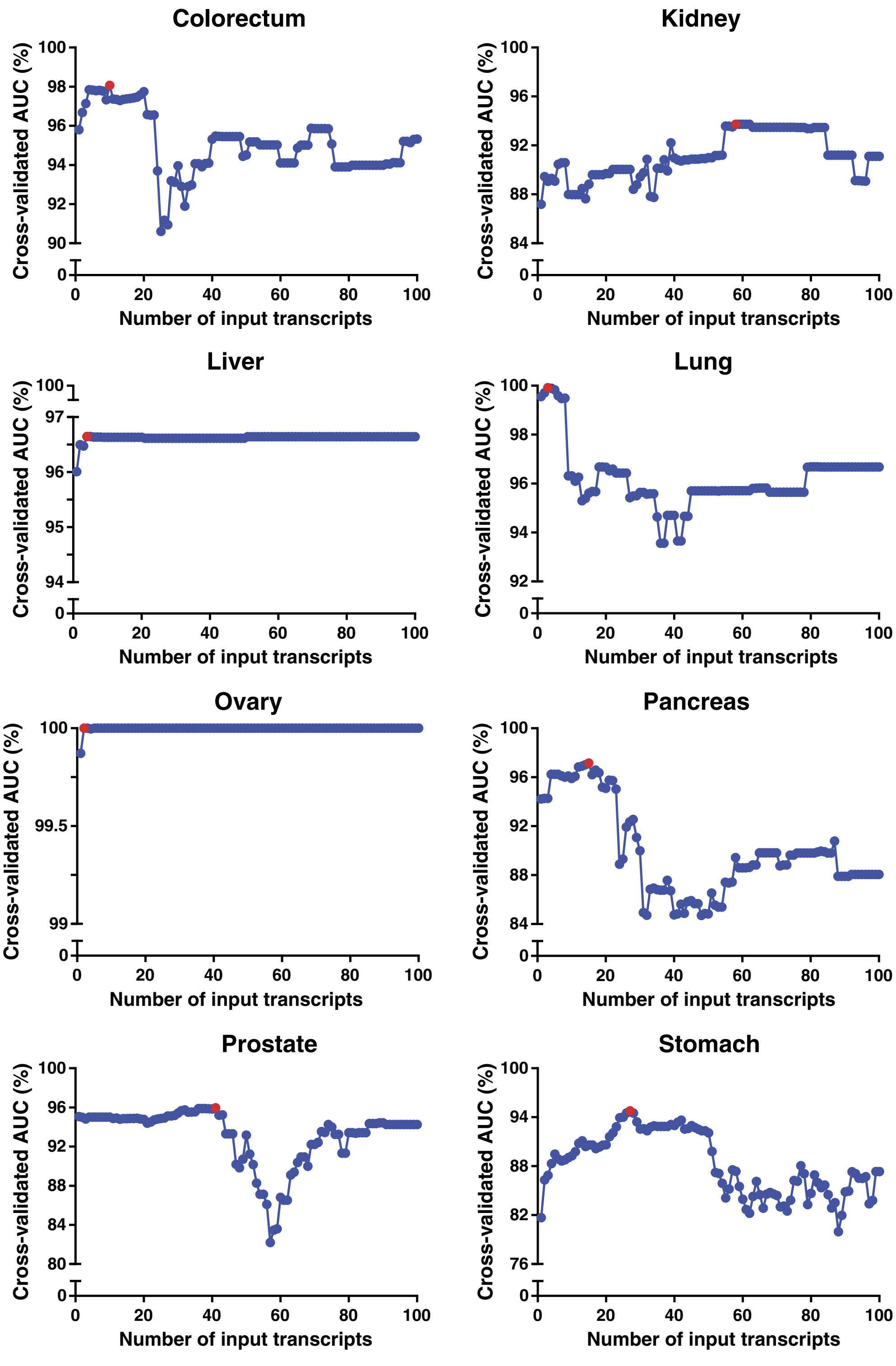

Figure 2. Scatter plots of 10-fold cross-validation method to determine optimal biomarker signatures for 8 different cancer types. Points highlighted in red indicate the highest, cross-validated AUC for each cancer type. AUC, area under the receiver operating characteristic curve.

from the first randomization (Table 2), with an overall cosine similarity measurement of 0.53 [50].

ROC curves were then generated for each biomarker signature, yielding high AUC’s (Figure 3); for comparison, lines of identity are shown, representing a random test that has no prognostic value. From each cancer signature ROC curve, threshold cut-offs minimizing the distance to (0,1) were calculated to use in subsequent external validation testing (Figure 3, red points).

The cancer biomarker signatures were then externally validated using the reserved 2,556-sample RNA-Seq test set. Each sample was tested against each of the 8 biomarker signatures and predicted to be positive or negative for that cancer type based on the threshold cut-

Table 2. Cancer Biomarker Signatures Generated from Two Separate Randomizations

Cancer type	Randomization #1	Randomization #2
Colorectum	*ATOH1* *FAM55A* *FAM55D* *NOX1*	*CDX1* *CDX2* *NOX1*
Kidney	*CTSL3* *DPYS* *FXYD2* *GLYAT* *NR1H4* *OR2T10* *RBP5* *SLC17A3* *TMEM174*	*CTSL3* *DPYS* *FXYD2* *SLC17A3* *SLC39A5* *TMEM174*
Liver	*APOH* *IGFBP1*	*ALB* *IGFBP1*
Lung	*NAPSA* *SCGB3A2* *SFTPB*	*AQP4* *NAPSA* *SCGB3A2* *SFTPB* *TBX4*
Ovary	*BEST1* *BGLAP*	*BEST1* *IER5L*
Pancreas	*ANKHD1-EIF4EBP3* *NKX3-1* *NKX3-2* *NKX6-1* *PALM2-AKAP2* *PPAN-P2RY11* *STON1-GTF2A1L* *TNFSF12-TNFSF13*	*ANKHD1-EIF4EBP3* *NKX3-2* *NKX6-1*
Prostate	*FEV* *KLK3* *MMP26* *NKX3-1* *OR51F2* *OR51T1* *PRAC* *SI*	*KLK15* *KLK2* *KLK3* *MYBPC1* *PRAC*
Stomach	*FAM166A* *GKN1* *LIPF* *MUC17* *OTC* *PRSS3* *SI* *TM4SF5* *USH1C*	*CTSE* *FAM166A* *GKN1* *GKN2* *HNF4A* *OTC* *SI* *TM4SF5* *USH1C* *VIL1*

Transcripts highlighted in red are common between the two signatures.

offs calculated previously. The true identities of each of the 2,556 samples were then compared to the predicted identities, and high sensitivity (92.0% ± 3.15%; mean ± standard deviation) and specificity (97.7% ± 2.99%) were demonstrated for each biomarker signature (Table 3). A heat map representation of the 43 transcripts comprising the 8 biomarker signatures for all 3,244 samples illustrated the strength with which each cancer signature successfully distinguished samples from among the different cancer types (Figure 4). Despite similarities in expression patterns across different cancers, the transcript signatures were nonetheless highly specific, requiring elevated expression across all the biomarker transcripts for a given cancer type (Table 3). Of the 3,244 samples, 2.9% had positive predictions for more than one cancer. These samples with duplicate predictions were assigned the final identity of the cancer type with the highest relative predicted value, yielding an overall accuracy of 90.5% (Table 4).

Discussion

In this study, we demonstrate an effective and efficient model for extracting highly sensitive and specific cancer biomarker signatures from a large RNA-Seq dataset. This technique yielded transcript signatures for the top eight cancer types that cause metastatic CUP. The robustness of the final signatures was demonstrated by external validation, through testing of a large test set randomly allocated *a priori*. By reserving these 2,556 samples solely for external validation, this test set represents a large dataset of "clinical unknowns" (i.e., identities unknown to the training model). The high sensitivities and specificities achieved with each of the cancer biomarker signatures therefore represent a realistic approximation of the accuracy with which the signatures may be used to predict tissue origin of a CUP sample in the clinical setting.

Many of the cancer biomarkers identified in this study have previously been well-characterized in their respective cancer types. The prostate cancer signature includes *KLK3*, which is responsible for encoding prostate-specific antigen, the serine protease used as a serum marker in prostate cancer screening and disease monitoring [51], as well as *PRAC*, which is known to be highly expressed in prostate cancers [52,53]. Similarly, *NKX3-1* is an androgen-regulated homeobox gene, which transcriptionally regulates oxidative damage response and is required for prostate stem cell maintenance; aberrant expression has been found to correlate strongly with prostate cancer progression [54–59]. In addition, a recent study demonstrated that immunohistochemical staining with the kidney biomarker *FXYD2*, a Na-K-ATPase regulator, is highly sensitive and specific for renal cell carcinoma [60]. Similarly, *NOX1* was highly expressed in our colorectal cancer samples, as confirmed in prior studies, which stimulates mitogenesis and angiogenesis though a *ROS*-mediated mechanism; *NOX1* expression has also been found to correlate strongly with activating *KRAS* mutations, which are present in approximately 50% of colorectal tumors [61,62]. *IGFBP1*, is a hepatocyte-derived secreted protein required for normal liver regeneration by inhibiting proapoptotic signals [63], with overexpression previously identified in hepatocellular cancers [64], as well as in our study. The lung biomarker *NAPSA* is a well characterized proteinase expressed in type II pneumocytes [65–67], whose expression has high sensitivity and specificity for distinguishing primary lung adenocarcinoma from metastatic pulmonary lesions from other primaries [67]. The pancreas biomarker *NKX6-1* is a transcriptional regulator that has been shown to play an important role in beta cell differentiation during pancreatic development [68–71]. Similarly, *GKN1* is highly expressed in the gastric epithelium, providing protection to the antral mucosa and promoting healing after injury; it also acts as a tumor suppressor and is down-regulated compared to normal gastric tissues [72–76] but in our model was still significantly overexpressed as compared to other cancer types.

Our cancer signatures also identified transcripts that have not previously been associated with the cancer types of interest. Although not yet characterized in ovarian cancer, *BEST1*, which forms calcium-activated chloride channels across epithelial cells to promote cell proliferation [77], has been shown to be up-regulated in colon cancers [78]. Similarly, hypermethylation of *DPYS*, a gene important in pyrimidine metabolism, has been identified in prostate, colon, and breast cancers, as well as melanomas [79–81], but in our study, high expression was most sensitive and specific for kidney cancer. Interestingly, the transmembrane glycoprotein *SI* was highly expressed in both prostate and gastric cancer samples. Mutations in SI have previously been identified in head and neck, colorectal, and ovarian cancers, and in a recent study, *SI* mutations resulted in significant gene enrichment in oxidative phosphorylation, glycolysis/gluconeogenesis, and B-cell receptor signaling pathways, for promoting malignant progression in chronic lymphocytic leukemia [82].

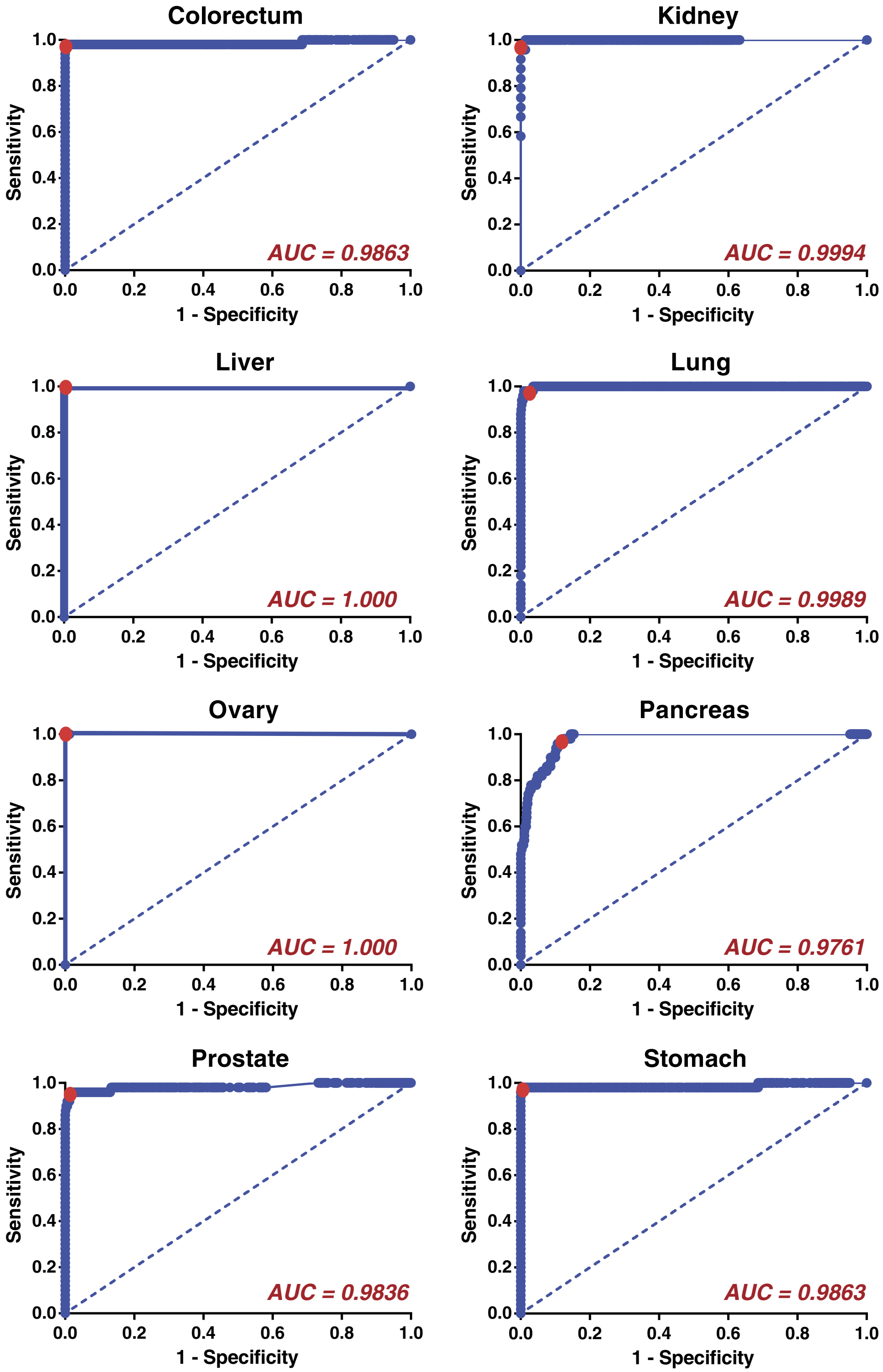

Figure 3. Internal validation of eight cancer-specific biomarker signatures yields high area under the ROC curve values. The entire 688-sample RNA-Seq training set was used as the test set for each cancer signature. Dotted lines indicate lines of identity. Points of minimum distance to (0,1) are highlighted in red. ROC, receiver operating characteristic; AUC, area under the ROC curve.

In our pancreatic cancer signature, our model also identified several gene fusions not previously associated with this disease, including *ANKHD1-EIF4EBP3*, a readthrough transcript of the neighboring cell survival scaffolding gene *ANKHD1* and the downstream translational repressor *EIF4EBP3*, both of which are effectors of the *RAS/MAPK* pathway [83,84], which is known to play a critical role in the development and progression of pancreatic cancer [85–88]. The prostate biomarker *NKX3-1* and related family member *NKX3-2* also comprised the pancreas signature. While the role of *NKX3-2* in pancreatic cancer has not yet been characterized, its role in chondreogenesis and skeletal development has been well studied, acting as a transcriptional repressor downstream of *SHH* through

Table 3. External Validation of Eight Cancer-Specific Biomarker Signatures Using 2,556-Sample RNA-Seq Test Set

	Colorectum	Kidney	Liver	Lung	Ovary	Pancreas	Prostate	Stomach
TP	174	0	0	274	360	24	95	24
TN	2355	2533	2552	2222	2188	2389	2424	2305
FP	7	23	4	36	0	141	28	224
FN	20	0	0	24	8	2	9	3
Sensitivity (%)	*89.7*	-	-	*91.9*	*97.8*	*92.3*	*91.3*	*88.9*
Specificity (%)	*99.7*	*99.1*	*99.8*	*98.4*	*100*	*94.4*	*98.9*	*91.1*

TP, true positive; TN, true negative; FP, false positive; FN, false negative.

Table 4. Overall Performance of RNA-Seq Biomarker Generating Algorithm for Predicting Tissue of Origin in 2,556 Cancer Samples

	Samples		%
TP	946	Sensitivity	95.0
TN	1366	Specificity	87.6
FP	194	PPV	83.0
FN	50	NPV	96.5
		Accuracy	90.5

Samples with duplicate cancer predictions were assigned the identity with the highest predicted value. TP, true positive; TN, true negative; FP, false positive; FN, false negative; PPV, positive predictive value; NPV, negative predictive value.

interactions with the signal transduction protein *SMAD4* [89–92]. *SHH* and its related hedgehog-signaling pathways are well-known mediators of pancreatic carcinogenesis and are the targets of many new therapeutics [88,93–95]. Similarly, inactivation of the tumor suppressor *SMAD4* plays a critical role in the development of pancreatic cancer and correlates with increased tumor aggressiveness and poor prognosis [96–99]. It is important to note that a common difficulty encountered in the analysis of pancreatic adenocarcinoma tissues is frequent contamination by a dense, desmoplastic stroma that characteristically surround these tumor cells and can occupy up to 90% of a tumor sample's content [100]. However, in our study of 76 pancreatic cancer samples, we were nonetheless able to extract an 8-transcript signature to distinguish pancreatic samples from other cancer types with high sensitivity and specificity.

As compared to other studies focused on distinguishing tissue of origin for CUP, our study has multiple strengths. We analyzed a large number of cancer samples from 26 different tissue types (3,244 samples as compared to the previous studies analyzing fewer than 800 samples) [15–19,23–26]. In addition, we used multiple validation methods to strengthen our biomarker signatures, specifically reserving 2,556 samples for external validation testing, to yield an overall accuracy of 90.5%. This is as compared to previously reported classification accuracies of 76% to 89% [15–17,19,23,24]. Finally, the use of RNA-Seq expression data has a number of potential advantages over microarray techniques, as previously outlined, including wide genome coverage, which allowed us to identify several new biomarkers, such as *BEST1* in ovarian cancer and the gene fusion *ANKHD1-*

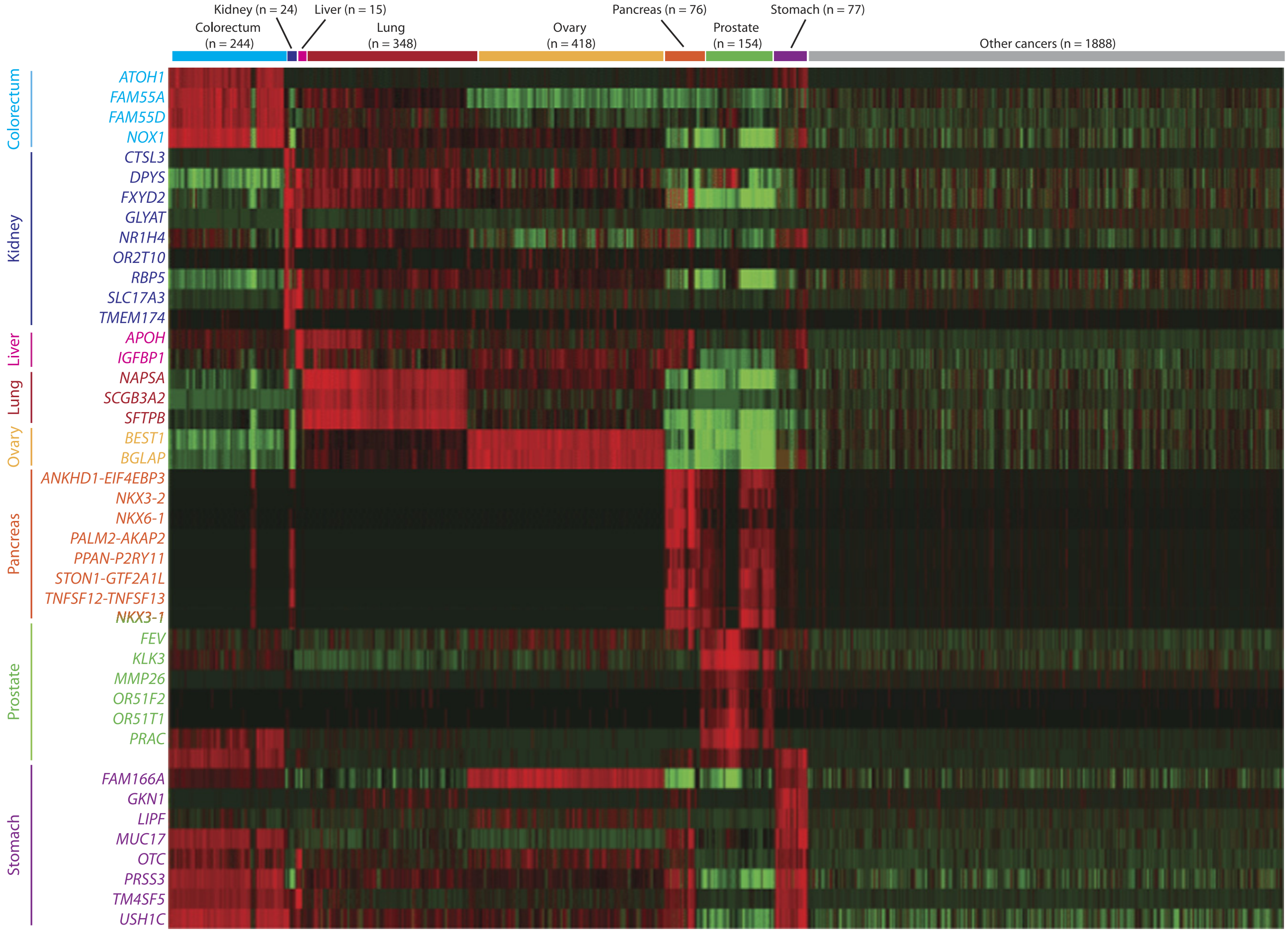

Figure 4. RNA-Seq heat map of 8 cancer-specific biomarker signatures (rows) across all 3,244 cancer samples (columns).

EIF4EBP3 in pancreatic cancer. To our knowledge, this is the first CUP study using large-scale RNA-Seq data for both training and validation to demonstrate a highly accurate model for cancer prediction.

One of the limitations of our study is that, although RNA-Seq allows for the capability to detect unmapped genes, in this proof-of-concept study, we limited our analysis to only annotated transcripts. While computationally more intensive, a dataset comprised of chromosomal positions rather than annotated genes would allow for additional discovery of novel biomarkers and could potentially improve the accuracy. In addition, in our dataset, there were insufficient kidney and liver samples to allocate to our test set for external validation; however, the kidney and liver biomarker signatures nonetheless yielded strong specificities of > 99%.

We have demonstrated the strength of this model in its ability to accurately and efficiently distinguish samples of one type (i.e., cancer type of interest) from another (i.e., heterogeneous group of other cancer types). While this study focused specifically on deriving lineage-specific cancer signatures by RNA-Seq, this model may be applied to any large dataset to query other clinical questions, such as deriving a signature that distinguishes premalignant and cancerous lesions from inflammatory and other benign conditions. Similarly, given a dataset of patients with known clinical outcomes, our model could be used to derive biomarker signatures that identify patients who would respond to a given therapy or identify patients with worse outcomes compared to clinically and histologically-matched cohorts who should be targeted for aggressive treatment and surveillance.

Conclusions

In this study, we introduced a computational model that successfully extracted accurate, lineage-specific cancer signatures for the top eight tissue types that contribute to CUP using RNA-Seq. Through external validation of a large dataset, we have shown how these signatures may be used to accurately identify tumors of unknown origin, demonstrating the translational potential of not only our cancer biomarker signatures but also the model itself, which may be applied to other clinical queries.

References

[1] Pimiento JM, Teso D, Malkan A, Dudrick SJ, and Palesty JA (2007). Cancer of unknown primary origin: a decade of experience in a community-based hospital. *Am J Surg* **194**, 833–837 [discussion 837–838].

[2] Pavlidis N and Pentheroudakis G (2010). Cancer of unknown primary site: 20 questions to be answered. *Ann Oncol* **21**(Suppl. 7), vii303–vii307.

[3] Shaw PH, Adams R, Jordan C, and Crosby TD (2007). A clinical review of the investigation and management of carcinoma of unknown primary in a single cancer network. *Clin Oncol (R Coll Radiol)* **19**, 87–95.

[4] Pavlidis N, Briasoulis E, Hainsworth J, and Greco FA (2003). Diagnostic and therapeutic management of cancer of an unknown primary. *Eur J Cancer* **39**, 1990–2005.

[5] Hainsworth JD and Greco FA (1993). Treatment of patients with cancer of an unknown primary site. *N Engl J Med* **329**, 257–263.

[6] Blaszyk H, Hartmann A, and Bjornsson J (2003). Cancer of unknown primary: clinicopathologic correlations. *APMIS* **111**, 1089–1094.

[7] National Comprehensive Cancer Network clinical Practice Gidelines in Oncology. Occult Primary (Cancer of Unknown Primary [CUP]). Version 3; 2014 [nccn.org].

[8] Werling RW, Yaziji H, Bacchi CE, and Gown AM (2003). CDX2, a highly sensitive and specific marker of adenocarcinomas of intestinal origin: An immunohistochemical survey of 476 primary and metastatic carcinomas. *Am J Surg Pathol* **27**, 303–310.

[9] Kaufmann O and Dietel M (2000). Thyroid transcription factor-1 is the superior immunohistochemical marker for pulmonary adenocarcinomas and large cell carcinomas compared to surfactant proteins A and B. *Histopathology* **36**, 8–16.

[10] Dennis JL, Hvidsten TR, Wit EC, Komorowski J, Bell AK, Downie I, Mooney J, Verbeke C, Bellamy C, and Keith WN, et al (2005). Markers of adenocarcinoma characteristic of the site of origin: development of a diagnostic algorithm. *Clin Cancer Res* **11**, 3766–3772.

[11] Kaufmann O, Fietze E, Mengs J, and Dietel M (2001). Value of p63 and cytokeratin 5/6 as immunohistochemical markers for the differential diagnosis of poorly differentiated and undifferentiated carcinomas. *Am J Clin Pathol* **116**, 823–830.

[12] Motzer RJ, Rodriguez E, Reuter VE, Bosl GJ, Mazumdar M, and Chaganti RS (1995). Molecular and cytogenetic studies in the diagnosis of patients with poorly differentiated carcinomas of unknown primary site. *J Clin Oncol* **13**, 274–282.

[13] Atkin NB and Baker MC (1982). Specific chromosome change, i(12p), in testicular tumours? *Lancet* **2**, 1349.

[14] Ilson DH, Motzer RJ, Rodriguez E, Chaganti RS, and Bosl GJ (1993). Genetic analysis in the diagnosis of neoplasms of unknown primary tumor site. *Semin Oncol* **20**, 229–237.

[15] Bloom G, Yang IV, Boulware D, Kwong KY, Coppola D, Eschrich S, Quackenbush J, and Yeatman TJ (2004). Multi-platform, multi-site, microarray-based human tumor classification. *Am J Pathol* **164**, 9–16.

[16] Ramaswamy S, Tamayo P, Rifkin R, Mukherjee S, Yeang CH, Angelo M, Ladd C, Reich M, Latulippe E, and Mesirov JP, et al (2001). Multiclass cancer diagnosis using tumor gene expression signatures. *Proc Natl Acad Sci U S A* **98**, 15149–15154.

[17] Shedden KA, Taylor JM, Giordano TJ, Kuick R, Misek DE, Rennert G, Schwartz DR, Gruber SB, Logsdon C, and Simeone D, et al (2003). Accurate molecular classification of human cancers based on gene expression using a simple classifier with a pathological tree-based framework. *Am J Pathol* **163**, 1985–1995.

[18] Tothill RW, Kowalczyk A, Rischin D, Bousioutas A, Haviv I, van Laar RK, Waring PM, Zalcberg J, Ward R, and Biankin AV, et al (2005). An expression-based site of origin diagnostic method designed for clinical application to cancer of unknown origin. *Cancer Res* **65**, 4031–4040.

[19] Rosenfeld N, Aharonov R, Meiri E, Rosenwald S, Spector Y, Zepeniuk M, Benjamin H, Shabes N, Tabak S, and Levy A, et al (2008). MicroRNAs accurately identify cancer tissue origin. *Nat Biotechnol* **26**, 462–469.

[20] Ojala KA, Kilpinen SK, and Kallioniemi OP (2011). Classification of unknown primary tumors with a data-driven method based on a large microarray reference database. *Genome Med* **3**, 63.

[21] Ross DT, Scherf U, Eisen MB, Perou CM, Rees C, Spellman P, Iyer V, Jeffrey SS, Van de Rijn M, and Waltham M, et al (2000). Systematic variation in gene expression patterns in human cancer cell lines. *Nat Genet* **24**, 227–235.

[22] Horlings HM, van Laar RK, Kerst JM, Helgason HH, Wesseling J, van der Hoeven JJ, Warmoes MO, Floore A, Witteveen A, and Lahti-Domenici J, et al (2008). Gene expression profiling to identify the histogenetic origin of metastatic adenocarcinomas of unknown primary. *J Clin Oncol* **26**, 4435–4441.

[23] Ma XJ, Patel R, Wang X, Salunga R, Murage J, Desai R, Tuggle JT, Wang W, Chu S, and Stecker K, et al (2006). Molecular classification of human cancers using a 92-gene real-time quantitative polymerase chain reaction assay. *Arch Pathol Lab Med* **130**, 465–473.

[24] Talantov D, Baden J, Jatkoe T, Hahn K, Yu J, Rajpurohit Y, Jiang Y, Choi C, Ross JS, and Atkins D, et al (2006). A quantitative reverse transcriptase-polymerase chain reaction assay to identify metastatic carcinoma tissue of origin. *J Mol Diagn* **8**, 320–329.

[25] Lu J, Getz G, Miska EA, Alvarez-Saavedra E, Lamb J, Peck D, Sweet-Cordero A, Ebert BL, Mak RH, and Ferrando AA, et al (2005). MicroRNA expression profiles classify human cancers. *Nature* **435**, 834–838.

[26] Quon G and Morris Q (2009). ISOLATE: a computational strategy for identifying the primary origin of cancers using high-throughput sequencing. *Bioinformatics* **25**, 2882–2889.

[27] Sultan M, Schulz MH, Richard H, Magen A, Klingenhoff A, Scherf M, Seifert M, Borodina T, Soldatov A, and Parkhomchuk D, et al (2008). A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. *Science* **321**, 956–960.

[28] Wang Z, Gerstein M, and Snyder M (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet* **10**, 57–63.

[29] Pentheroudakis G, Golfinopoulos V, and Pavlidis N (2007). Switching benchmarks in cancer of unknown primary: from autopsy to microarray. *Eur J Cancer* **43**, 2026–2036.

[30] Briasoulis E and Pavlidis N (1997). Cancer of unknown primary origin. *Oncologist* **2**, 142–152.

[31] Abbruzzese JL, Abbruzzese MC, Lenzi R, Hess KR, and Raber MN (1995). Analysis of a diagnostic strategy for patients with suspected tumors of unknown origin. *J Clin Oncol* **13**, 2094–2103.

[32] Stewart JF, Tattersall MH, Woods RL, and Fox RM (1979). Unknown primary adenocarcinoma: incidence of overinvestigation and natural history. *Br Med J* **1**, 1530–1533.

[33] Le Chevalier T, Cvitkovic E, Caille P, Harvey J, Contesso G, Spielmann M, and Rouesse J (1988). Early metastatic cancer of unknown primary origin at presentation. A clinical study of 302 consecutive autopsied patients. *Arch Intern Med* **148**, 2035–2039.

[34] Hamilton CS and Langlands AO (1987). ACUPS (adenocarcinoma of unknown primary site): a clinical and cost benefit analysis. *Int J Radiat Oncol Biol Phys* **13**, 1497–1503.

[35] Kirsten F, Chi CH, Leary JA, Ng AB, Hedley DW, and Tattersall MH (1987). Metastatic adeno or undifferentiated carcinoma from an unknown primary site–natural history and guidelines for identification of treatable subsets. *Q J Med* **62**, 143–161.

[36] Nystrom JS, Weiner JM, Wolf RM, Bateman JR, and Viola MV (1979). Identifying the primary site in metastatic cancer of unknown origin. Inadequacy of roentgenographic procedures. *JAMA* **241**, 381–383.

[37] Moertel CG, Reitemeier RJ, Schutt AJ, and Hahn RG (1972). Treatment of the patient with adenocarcinoma of unknown origin. *Cancer* **30**, 1469–1472.

[38] Osteen RT, Kopf G, and Wilson RE (1978). In pursuit of the unknown primary. *Am J Surg* **135**, 494–497.

[39] Kothari V, Wei I, Shankar S, Kalyana-Sundaram S, Wang L, Ma LW, Vats P, Grasso CS, Robinson DR, and Wu YM, et al (2013). Outlier kinase expression by RNA sequencing as targets for precision therapy. *Cancer Discov* **3**, 280–293.

[40] International Cancer Genome Consortium; 2014. http://www.icgc.org.

[41] The Cancer Genome Atlas; 2014. http://cancergenome.nih.gov.

[42] R language; 2014. http://www.R-project.org.

[43] Sing T, Sander O, Beerenwinkel N, and Lengauer T (2005). ROCR: visualizing classifier performance in R. *Bioinformatics* **21**, 3940–3941.

[44] Everitt BS and Hothorn T (2010). A handbook of statistical analyses using R. CRC PressINC; 2010 [Vol.].

[45] Akaike H (1974). A new look at the statistical model identification. *IEEE Trans Autom Control* **19**, 716–723.

[46] McCullagh P and Nelder JA (1989). Generalized linear model. Chapman & Hall; 1989 [Vol.].

[47] Perkins NJ and Schisterman EF (2006). The inconsistency of "optimal" cutpoints obtained using two criteria based on the receiver operating characteristic curve. *Am J Epidemiol* **163**, 670–675.

[48] de Hoon MJ, Imoto S, Nolan J, and Miyano S (2004). Open source clustering software. *Bioinformatics* **20**, 1453–1454.

[49] Eisen M (2002). TreeView. http://rana.lbl.gov/EisenSoftware.htm.

[50] Pesquita C, Faria D, Falcao AO, Lord P, and Couto FM (2009). Semantic similarity in biomedical ontologies. *PLoS Comput Biol* **5**, e1000443.

[51] Eeles RA, Kote-Jarai Z, Giles GG, Olama AAA, Guy M, Jugurnauth SK, Mulholland S, Leongamornlert DA, Edwards SM, and Morrison J, et al (2008). Multiple newly identified loci associated with prostate cancer susceptibility. *Nat Genet* **40**, 316–321.

[52] Edwards S, Campbell C, Flohr P, Shipley J, Giddings I, Te-Poele R, Dodson A, Foster C, Clark J, and Jhavar S, et al (2005). Expression analysis onto microarrays of randomly selected cDNA clones highlights HOXB13 as a marker of human prostate cancer. *Br J Cancer* **92**, 376–381.

[53] Liu XF, Olsson P, Wolfgang CD, Bera TK, Duray P, Lee B, and Pastan I (2001). PRAC: A novel small nuclear protein that is specifically expressed in human prostate and colon. *Prostate* **47**, 125–131.

[54] Bhatia-Gaur R, Donjacour AA, Sciavolino PJ, Kim M, Desai N, Young P, Norton CR, Gridley T, Cardiff RD, and Cunha GR, et al (1999). Roles for Nkx3.1 in prostate development and cancer. *Genes Dev* **13**, 966–977.

[55] Bowen C, Bubendorf L, Voeller HJ, Slack R, Willi N, Sauter G, Gasser TC, Koivisto P, Lack EE, and Kononen J, et al (2000). Loss of NKX3.1 expression in human prostate cancers correlates with tumor progression. *Cancer Res* **60**, 6111–6115.

[56] Ellwood-Yen K, Graeber TG, Wongvipat J, Iruela-Arispe ML, Zhang J, Matusik R, Thomas GV, and Sawyers CL (2003). Myc-driven murine prostate cancer shares molecular features with human prostate tumors. *Cancer Cell* **4**, 223–238.

[57] He WW, Sciavolino PJ, Wing J, Augustus M, Hudson P, Meissner PS, Curtis RT, Shell BK, Bostwick DG, and Tindall DJ, et al (1997). A novel human prostate-specific, androgen-regulated homeobox gene (NKX3.1) that maps to 8p21, a region frequently deleted in prostate cancer. *Genomics* **43**, 69–77.

[58] Nelson WG, De Marzo AM, and Isaacs WB (2003). Prostate cancer. *N Engl J Med* **349**, 366–381.

[59] Wang X, Julio MKD, Economides KD, Walker D, Yu H, Halili MV, Hu YP, Price SM, Abate-Shen C, and Shen MM (2009). A luminal epithelial stem cell that is a cell of origin for prostate cancer. *Nature* **461**, 495–500.

[60] Gaut JP, Crimmins DL, Lockwood CM, McQuillan JJ, and Ladenson JH (2013). Expression of the Na+/K+− transporting ATPase gamma subunit FXYD2 in renal tumors. *Mod Pathol* **26**, 716–724.

[61] Laurent E, McCoy Iii JW, Macina RA, Liu W, Cheng G, Robine S, Papkoff J, and Lambeth JD (2008). Nox1 is over-expressed in human colon cancers and correlates with activating mutations in K-Ras. *Int J Cancer* **123**, 100–107.

[62] Kamata T (2009). Roles of Nox1 and other Nox isoforms in cancer development. *Cancer Sci* **100**, 1382–1388.

[63] Leu JI, Crissey MAS, and Taub R (2003). Massive hepatic apoptosis associated with TGF-β1 activation after Fas ligand treatment of IGF binding protein-1–deficient mice. *J Clin Invest* **111**, 129–139.

[64] Borlak J, Meier T, Halter R, Spanel R, and Spanel-Borowski K (2005). Epidermal growth factor-induced hepatocellular carcinoma: Gene expression profiles in precursor lesions, early stage and solitary tumours. *Oncogene* **24**, 1809–1819.

[65] Chuman Y, Bergman A, Ueno T, Saito S, Sakaguchi K, Alaiya AA, Franzen B, Bergman T, Arnott D, and Auer G, et al (1999). Napsin A, a member of the aspartic protease family, is abundantly expressed in normal lung and kidney tissue and is expressed in lung adenocarcinomas. *FEBS Lett* **462**, 129–134.

[66] Hirano T, Auer G, Maeda M, Hagiwara Y, Okada S, Ohira T, Okuzawa K, Fujioka K, Franzen B, and Hibi N, et al (2000). Human tissue distribution of TA02, which is homologous with a new type of aspartic proteinase, napsin A. *Jpn J Cancer Res* **91**, 1015–1021.

[67] Ueno T, Linder S, and Elmberger G (2003). Aspartic proteinase napsin is a useful marker for diagnosis of primary lung adenocarcinoma. *Br J Cancer* **88**, 1229–1233.

[68] Habener JF, Kemp DM, and Thomas MK (2005). Minireview: transcriptional regulation in pancreatic development. *Endocrinology* **146**, 1025–1034.

[69] Sander N, Sussel L, Conners J, Scheel D, Kalamaras J, Dela Cruz F, Schwitzgebel V, Hayes-Jordan A, and German M (2000). Homeobox gene Nkx6.1 lies downstream of Nkx2.2 in the major pathway of β-cell formation in the pancreas. *Development* **127**, 5533–5540.

[70] Schwitzgebel VM, Scheel DW, Conners JR, Kalamaras J, Lee JE, Anderson DJ, Sussel L, Johnson JD, and German MS (2000). Expression of neurogenin3 reveals an islet cell precursor population in the pancreas. *Development* **127**, 3533–3542.

[71] Yang L, Li S, Hatch H, Ahrens K, Cornelius JG, Petersen BE, and Peck AB (2002). In vitro trans-differentiation of adult hepatic stem cells into pancreatic endocrine hormone-producing cells. *Proc Natl Acad Sci U S A* **99**, 8078–8083.

[72] Moss SF, Lee JW, Sabo E, Rubin AK, Rommel J, Westley BR, May FEB, Gao J, Meitner PA, and Tavares R, et al (2008). Decreased expression of gastrokine 1 and the trefoil factor interacting protein TFIZ1/GKIM2 in gastric cancer: influence of tumor histology and relationship to prognosis. *Clin Cancer Res* **14**, 4161–4167.

[73] Oien KA, McGregor F, Butler S, Ferrier RK, Downie I, Bryce S, Burns S, and Keith WN (2004). Gastrokine I is abundantly and specifically expressed in superficial gastric epithelium, down-regulated in gastric carcinoma, and shows high evolutionary conservation. *J Pathol* **203**, 789–797.

[74] Xing R, Li W, Cui J, Zhang J, Kang B, Wang Y, Wang Z, Liu S, and Lu Y (2012). Gastrokine 1 induces senescence through p16/Rb pathway activation in gastric cancer cells. *Gut* **61**, 43–52.

[75] Yoon JH, Kang YH, Choi YJ, Park IS, Nam SW, Lee JY, Lee YS, and Park WS (2011). Gastrokine 1 functions as a tumor suppressor by inhibition of epithelial-mesenchymal transition in gastric cancers. *J Cancer Res Clin Oncol* **137**, 1697–1704.

[76] Yoon JH, Song JH, Zhang C, Jin M, Kang YH, Nam SW, Lee JY, and Park WS (2011). Inactivation of the Gastrokine 1 gene in gastric adenomas and carcinomas. *J Pathol* **223**, 618–625.

[77] Kunzelmann K, Kongsuphol P, Aldehni F, Tian Y, Ousingsawat J, Warth R, and Schreiber R (2009). Bestrophin and TMEM16-Ca2+ activated Cl− channels with different functions. *Cell Calcium* **46**, 233–241.

[78] Spitzner M, Martins JR, Soria RB, Ousingsawat J, Scheidt K, Schreiber R, and Kunzelmann K (2008). Eag1 and bestrophin 1 are up-regulated in fast-growing colonic cancer cells. *J Biol Chem* **283**, 7421–7428.

[79] Chung W, Kwabi-Addo B, Ittmann M, Jelinek J, Shen L, Yu Y, and Issa JPJ (2008). Identification of novel tumor markers in prostate, colon and breast cancer by unbiased methylation profiling. *PLoS One* , 3.

[80] Furuta J, Nobeyama Y, Umebayashi Y, Otsuka F, Kikuchi K, and Ushijima T (2006). Silencing of peroxiredoxin 2 and aberrant methylation of 33 CpG islands in putative promoter regions in human malignant melanomas. *Cancer Res* **66**, 6080–6086.

[81] Vasiljević N, Wu K, Brentnall AR, Kim DC, Thorat MA, Kudahetti SC, Mao X, Xue L, Yu Y, and Shaw GL, et al (2011). Absolute quantitation of DNA methylation of 28 candidate genes in prostate cancer using pyrosequencing. *Dis Markers* **30**, 151–161.

[82] Rodríguez D, Ramsay AJ, Quesada V, Garabaya C, Campo E, Freije JMP, and López-Otín C (2013). Functional analysis of sucrase-isomaltase mutations from chronic lymphocytic leukemia patients. *Hum Mol Genet* **22**, 2273–2282.

[83] Prakash T, Sharma VK, Adati N, Ozawa R, Kumar N, Nishida Y, Fujikake T, Takeda T, and Taylor TD (2010). Expression of conjoined genes: another mechanism for gene regulation in eukaryotes. *PLoS One* **5**, e13284.

[84] Poulin F, Brueschke A, and Sonenberg N (2003). Gene fusion and overlapping reading frames in the mammalian genes for 4E-BP3 and MASK. *J Biol Chem* **278**, 52290–52297.

[85] Almoguera C, Shibata D, Forrester K, Martin J, Arnheim N, and Perucho M (1988). Most human carcinomas of the exocrine pancreas contain mutant c-K-ras genes. *Cell* **53**, 549–554.

[86] Bos JL (1989). ras Oncogenes in human cancer: a review. *Cancer Res* **49**, 4682–4689.

[87] Hingorani SR, Petricoin Iii EF, Maitra A, Rajapakse V, King C, Jacobetz MA, Ross S, Conrads TP, Veenstra TD, and Hitt BA, et al (2003). Preinvasive and invasive ductal pancreatic cancer and its early detection in the mouse. *Cancer Cell* **4**, 437–450.

[88] Morton JP, Mongeau ME, Klimstra DS, Morris JP, Yie CL, Kawaguchi Y, Wright CVE, Hebrok M, and Lewis BC (2007). Sonic hedgehog acts at multiple stages during pancreatic tumorigenesis. *Proc Natl Acad Sci U S A* **104**, 5103–5108.

[89] Kim DW and Lassar AB (2003). Smad-dependent recruitment of a histone deacetylase/Sin3A complex modulates the bone morphogenetic protein-dependent transcriptional repressor activity of Nkx3.2. *Mol Cell Biol* **23**, 8704–8717.

[90] Lefebvre V and Smits P (2005). Transcriptional control of chondrocyte fate and differentiation. *Birth Defects Res C Embryo Today* **75**, 200–212.

[91] Murtaugh LC, Zeng L, Chyung JH, and Lassar AB (2001). The chick transcriptional repressor Nkx3.2 acts downstream of Shh to promote BMP-dependent axial chondrogenesis. *Dev Cell* **1**, 411–422.

[92] Zeng L, Kempf H, Murtaugh LC, Sato ME, and Lassar AB (2002). Shh establishes an Nkx3.2/Sox9 autoregulatory loop that is maintained by BMP signals to induce somitic chondrogenesis. *Genes Dev* **16**, 1990–2005.

[93] Katoh Y and Katoh M (2009). Hedgehog target genes: mechanisms of carcinogenesis induced by aberrant hedgehog signaling activation. *Curr Mol Med* **9**, 873–886.

[94] Lee CJ, Dosch J, and Simeone DM (2008). Pancreatic cancer stem cells. *J Clin Oncol* **26**, 2806–2812.

[95] Thayer SP, Di Magliano MP, Heiser PW, Nielsen CM, Roberts DJ, Lauwers GY, Qi YP, Gysin S, Fernández-del Castillo C, and Yajnik V, et al (2003). Hedgehog is an early and late mediator of pancreatic cancer tumorigenesis. *Nature* **425**, 851–856.

[96] Hahn SA, Schutte M, Shamsul Hoque ATM, Moskaluk CA, Da Costa LT, Rozenblum E, Weinstein CL, Fischer A, Yeo CJ, and Hruban RH, et al (1996). DPC4, a candidate tumor suppressor gene at human chromosome 18q21.1. *Science* **271**, 350–353.

[97] Hezel AF, Kimmelman AC, Stanger BZ, Bardeesy N, and DePinho RA (2006). Genetics and biology of pancreatic ductal adenocarcinoma. *Genes Dev* **20**, 1218–1249.

[98] Rozenblum E, Schutte M, Goggins M, Hahn SA, Panzer S, Zahurak M, Goodman SN, Sohn TA, Hruban RH, and Yeo CJ, et al (1997). Tumor-suppressive pathways in pancreatic carcinoma. *Cancer Res* **57**, 1731–1734.

[99] Wilentz RE, Iacobuzio-Donahue CA, Argani P, McCarthy DM, Parsons JL, Yeo CJ, Kern SE, and Hruban RH (2000). Loss of expression of Dpc4 in pancreatic intraepithelial neoplasia: evidence that DPC4 inactivation occurs late in neoplastic progression. *Cancer Res* **60**, 2002–2006.

[100] Mahadevan D and Von Hoff DD (2007). Tumor-stroma interactions in pancreatic ductal adenocarcinoma. *Mol Cancer Ther* **6**, 1186–1197.