IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

SPECIAL SECTION ON FEATURE REPRESENTATION AND LEARNING METHODS WITH APPLICATIONS IN LARGE-SCALE BIOLOGICAL SEQUENCE ANALYSIS

Received September 2, 2019, accepted September 15, 2019, date of publication September 26, 2019,
date of current version October 11, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2944132

Gene Expression Difference Between Primary and Metastatic Renal Cell Carcinoma Using Patient-Derived Xenografts

YANG JIANG[1], XIAOYONG PAN[2,3], YU-HANG ZHANG[4], TAO HUANG[4], AND YUFEI GAO[5]

[1]Department of Gastrointestinal and Colorectal surgery, China-Japan Union Hospital of Jilin University, Jilin 130033, China
[2]Institute of Image Processing and Pattern Recognition, and Key Laboratory of System Control and Information Processing, Shanghai Jiao Tong University, Ministry of Education of China, Shanghai 200240, China
[3]IDLab, Department for Electronics and Information Systems, Ghent University, 9000 Ghent, Belgium
[4]Shanghai Institute of Nutrition and Health, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai 200031, China
[5]Department of Neurosurgery, China–Japan Union Hospital of Jilin University, Jilin 130033, China

Corresponding author: Yufei Gao (gaoyf@jlu.edu.cn)

This work was supported in part by the Department of Science and Technology of Jilin Province under Grant 20180101136JC, in part by the Department of Finance of Jilin Province under Grant 2018SCZ030, in part by the Excellent Talents Training Plan China–Japan Union Hospital of Jilin University under Grant YXZN-201803, in part by the Medical and Health Project of Jilin Province under Grant 20190304047YY, in part by the International Cooperation Projects of Jilin Province under Grant 20160414047GH, in part by the National Natural Science Foundation of China under Grant 31701151, in part by the National Key Research and Development Program of China under Grant 2018YFC0910403, in part by the Shanghai Municipal Science and Technology Major Project under Grant 2017SHZDZX01, in part by the Shanghai Sailing Program under Grant 16YF1413800, and in part by the Youth Innovation Promotion Association of Chinese Academy of Sciences (CAS) under Grant 2016245.

ABSTRACT Metastasis is the leading cause of cancer-related death. A small proportion of tumor cells can spread to other tissues through the lymph system or bloodstream and colonize in a new microenvironment. However, not all tumors from the primary site can metastasize. What is the difference between metastatic and primary tumors? Can such difference be preserved in widely used patient-derived xenografts (PDX)? To answer these questions, we analyzed the single-cell gene expression profiles of 36 cells from PDX of metastatic renal cell carcinoma (mRCC), 47 cells from PDX of primary RCC (pRCC), and 35 cells from parental mRCC (pt-mRCC). First, the gene expression patterns of PDX-mRCC and PDX-pRCC were compared, and the PDX-mRCC signatures were generated. Such signatures reflected the difference between metastatic and primary tumors. Second, the pt-mRCC were tested on whether they can be correctly classified into the PDX-mRCC class rather than PDX-pRCC. We found that pt-mRCC were very similar with PDX-mRCC. Our results prove that the PDX is a great research model for metastatic tumors since it preserved the essences for tumor metastasis. Our results justify the applications of PDX in metastatic tumor studies.

INDEX TERMS Tumor metastasis, patient-derived xenografts, gene expression, Monte Carlo feature selection, support vector machine.

I. INTRODUCTION

Renal cell carcinoma (RCC) is a malignant proliferative disease involving the abnormal proliferation and invasion of mixed renal cells [1], [2]. Originating from the proximal convoluted tubule, RCC has been reported to be one of the most common subtypes of kidney cancer, accounting for more than 90% of all renal cancers [3]. Globally, kidney and renal pelvis cancers are responsible for more than 15.6 new cases and approximately 3.9 new deaths per 100,000 men and women [4]. According to 2014 statistics, more than 480,000 people suffered from such diseases in the United States [3]. Furthermore, in 2017, more than 64,000 people were estimated to have been diagnosed with such disease, considering the increasing morbidity in the past 15 years. Although the 5-year survival rate of such disease has increased by more than 74% with the development of clinical treatment, RCC is still a great threat to human health [3].

The associate editor coordinating the review of this manuscript and approving it for publication was Leyi Wei.

Metastasis is one of the major pathogenic behaviors for almost all types of cancers [5], [6]. As a malignant tumor subtype, RCC metastasis has been widely identified in clinical practice [7], [8]. Generally, there are three major ways for cancer to metastasize and spread in the body: (1) direct spread into tissues around the primary situs (2) movement into the lymph system, and (3) movement into the bloodstream [9]. During metastasis, not all primary tumor cells can successfully transfer and colonize into a new microenvironment. Therefore, all metastatic cells are derived from the original primary tumor tissues and are further screened and filtered by metastatic tumor microenvironment. According to recent publications, the distinctive expression pattern between metastatic and primary tumor tissues have already been identified in various tumor types, including RCC [10], [11]. Early in 2003, a systematic study [12] on primary and metastatic tumor tissues confirmed that the expression profiles of metastatic tumor tissues can only reflect those of only a small subgroup of primary tumor tissues, validating the distinctive and alternative expression patterns during tumor metastasis.

Considering the alternative expression pattern of these tumor cell subgroups, for a long time, researchers have tried to establish an applicable tumor model to identify such differences. The patient-derived xenograft (PDX) mouse model was created by implanting the patient's clinical tumor tissues (either primary or metastatic) into an immune-deficient mouse; it has been widely used in scientific studies on multiple tumor types, including breast cancer, colorectal cancer, pancreatic cancer, and renal carcinoma [13]–[16]. Implanted under aseptic conditions, the PDX mouse model is able to reflect the genetic background and basic biological characteristics of patients' original tumor tissues to the maximum extent [15]. Therefore, PDX mouse models are gradually being used in tumor studies. However, various studies have confirmed that during the implantation and development of tumor xenografts, the implanted tumor tissues may lose some of the hereditary information due to the distinctive environmental stress [17]. Therefore, the genetic background of PDX mouse model containing tumors may not be exactly the same with the original tumor tissues. Such controllable distinctions may be acceptable in studies focusing only on primary or metastatic tumors. However, PDX mouse models are also widely used in comparative studies between primary and metastatic tumor tissues, in which the proportion of genetic information loss may be quite significant [18]. Up to now, no direct studies have validated whether the PDX model can accurately reflect the expression distinction of primary and metastatic tumor tissues. Therefore, we summarized the expression profiles from Kim et al.'s study [18] on RCC and explored the distinctive expression pattern between primary and metastatic PDX mouse model-containing tissues.

In this study, based on the expression pattern of metastasis and primary tumor retrieved from Kim et al.'s study, we established a systematic computational method to identify the core differential expression pattern in primary and metastatic tumor tissues. On one hand, we qualitatively identified a group of effective genes that may have different expression patterns in primary tumor tissues relative to the metastatic ones and quantitatively set up a rule for such distinction. On the other hand, the correspondence between our training sets and test sets reflected the expression consistency between clinical tumor tissues (in situ or metastatic) and PDX tumor tissues, implying that the PDX mouse model may be a quite accurate and efficient mouse model for tumor research.

II. MATERIALS AND METHODS

A. DATASET

We downloaded single-cell RNA sequencing data of 24,866 genes in 36 cells from PDX of metastatic renal cell carcinoma (mRCC), 47 cells from PDX of primary RCC (pRCC), and 35 cells from parental mRCC (pt-mRCC) from the Gene Expression Omnibus with accession number of GSE73121 [18]. We investigated the expression difference of 24,866 expressed genes in PDX-mRCC and PDX-pRCC and tested whether the PDX-mRCC signature can be used to identify the mRCC samples. The PDX-mRCC and PDX-pRCC formed the training dataset, and the pt-mRCC served as the test dataset. First, the PDX-mRCC signatures were generated by comparing the differential expression of PDX-mRCC and PDX-pRCC. Second, the pt-mRCC were tested on whether they can be correctly classified into the PDX-mRCC class rather than PDX-pRCC.

B. FEATURE SELECTION

To identify highly related genes for PDX-mRCC and PDX-pRCC cells, we used a two-stage feature selection method. First, Monte Carlo feature selection (MCFS) [19] was applied to rank all available genes. Second, incremental feature selection (IFS) [20] was utilized on the ranked features to identify genes with strong discriminative power for cells of PDX-mRCC and PDX-pRCC.

MCFS [19] is used to rank the input features because it is good at dealing with high-dimensional datasets such as the training dataset in this study. To date, it has been applied to analyze different biological problems [21]–[23]. It consists of multiple decision tree classifiers. Each decision tree is built using m features randomly selected from original M features ($m << M$) and a bootstrap set from the original training set. For each feature subset, p decision trees are grown on p bootstrap sets consisting of features from this feature subset. The above process is repeated t times. In the end, we yield t feature subsets and a total of $p \times t$ decision trees. MCFS estimates the relative importance (RI) as an importance score for each feature according to weighted accuracy of each decision tree and the overall number of splits made on that feature in all nodes of all trees. In this study, MCFS software package downloaded from http://www.ipipan.eu/staff/m.draminski/mcfs.html is used to produce the ranked feature list in descending order of RI scores of features.

In addition, the MCFS method can output the most important features, called informative features. These features are always top ranking features in the feature list by setting a critical value of RI scores, which is determined by a permutation test on class labels and one-sided Student's t-test [24]. The obtained informative features are deemed to be essential for classification.

However, different classification algorithm needs different feature subspace to construct an optimal classifier. Thus, the informative features produced by MCFS method are not always optimal for each classification algorithm. In view of this, we selected a certain number of features with the best performance for a give classification algorithm using the IFS method [20]. Given a ranked feature list with M features from MCFS, denoted as $F = [f_1, f_2, \ldots, f_M]$, here we only kept M features of the original 24,866 genes with RI scores greater than 0. IFS first constructs a series of feature subsets according to the rank of each feature, with each feature subset having an additional feature than the former. Then, it collects the samples consisting of features from individual feature subsets. Support vector machine (SVM) was used to assess the classification performance on samples using 10-fold cross-validation [25]–[28]. In the end, we obtained an SVM classifier with the best performance, whose input features are called optimal features.

C. SVM

SVM seeks a separating hyperplane with maximum margin between samples from different classes in the feature space. In many cases, however, data samples are neither linearly nor perfectly separable. Thus, soft-margin SVM is proposed as it allows misclassification errors and maps the original features into a higher dimensional space using kernel tricks, in which the data samples are linearly separable. SVMs are widely and successfully used in many biological problems [29]–[33], especially for binary classification problems.

A tool, named ''SMO'', in Weka [34] was employed in this study because it implements a type of SVM algorithm optimized by the sequential minimal optimization (SMO) [35]. For convenience, this tool was performed with its default parameters. In detail, the parameter C was 1.0 and the kernel was a polynomial function. The Weka can be downloaded at https://www.cs.waikato.ac.nz/ml/weka/downloading.html.

D. RULE LEARNING

As metnioned in Section II-B, MCFS method can produce some informative features, which are some top featuers in the feature list. From these informative features, the Johnson Reducer algorithm [36] was adopted to extract a reduced feature subset that can give similar performance comparing with using all informative features. Then, a rule algorithm, Repeated Incremental Pruning to Produce Error Reduction (RIPPER) algorithm [37], was applied to construct classification rules. The detailed procedures for constructing rules via RIPPER algorithm are shown in **Figure 1**. The rules produced by RIPPER algorithm contain two parts: (I) conditions, listed at the left-hand-side of the rule; (II) result, listed at the right-hand-side of the rule. For instance, a rule can be ''IF Gene1 >= 0.2 AND Gene2 <= −1.3 THEN mRCC''. The MCFS program used in this study integrated the Johnson Reducer algorithm and RIPPER algorithm. Thus, it can directly output the classification rules, which were extensively analyzed in this study.

```
Initialize a set E to be the training set
Choose a class C that contains least instances
    Initialize a rule R to have an empty left-hand side that predicts C
    Split E into growing and pruning sets
    While there are positive samples (instances of C) in the growing set, or the description length
    (DL) is 64 bits greater than the smallest DL found so far, or the error rate is greater than 50%
        Until R is perfect (or no more attributes to add)
            For each attribute a not included in R, and for each value v,
                Consider a = v to add to the left-hand side of R
                Choose the a and v that have the highest Foil's information gain
                Add a = v to R
                Prune R using reduced error pruning
        Remove the instances covered by R from the growing set
Global optimization strategy is applied to further prune the rule.
```

FIGURE 1. The whole procedures of Repeated Incremental Pruning to Produce Error Reduction (RIPPER) algorithm for extracting classification rules [23].

E. PERFORMANCE EVALUATION

In this study, we performed 10-fold cross-validation on the training set and also evaluated the training model on a test set consisting of pt-mRCC samples. As a binary classification problem, we compared the predicted and real labels, and four values were counted. They were true positive (TP), true negative (TN), false negative (FN), and false positive (FP). Based on these values, four measurements, sensitivity (SN), specificity (SP), prediction accuracy (ACC), and Matthew's correlation coefficient (MCC) [38]–[40], can be calculated to evaluate the prediction ability of the classifier:

$$\begin{cases} SN = \dfrac{TP}{TP + FN} \\ SP = \dfrac{TN}{TN + FP} \\ ACC = \dfrac{TP + TN}{TP + TN + FP + FN} \\ MCC = \dfrac{TP \times TN - FP \times FN}{\sqrt{(TN+FN)\times(TN+FP)\times(TP+FN)\times(TP+FP)}} \end{cases} \tag{1}$$

III. RESULTS

In this study, we employed several advanced computational methods to analyze the single-cell RNA sequencing data in PDX-mRCC and PDX-pRCC. The whole procedures are illustrated in Figure 2.

A. RESULTS OF MCFS METHOD

We first used MCFS method to rank all features in the feature list by descending order of their RI scores. The top 3,525 features/genes with corresponding RI scores greater

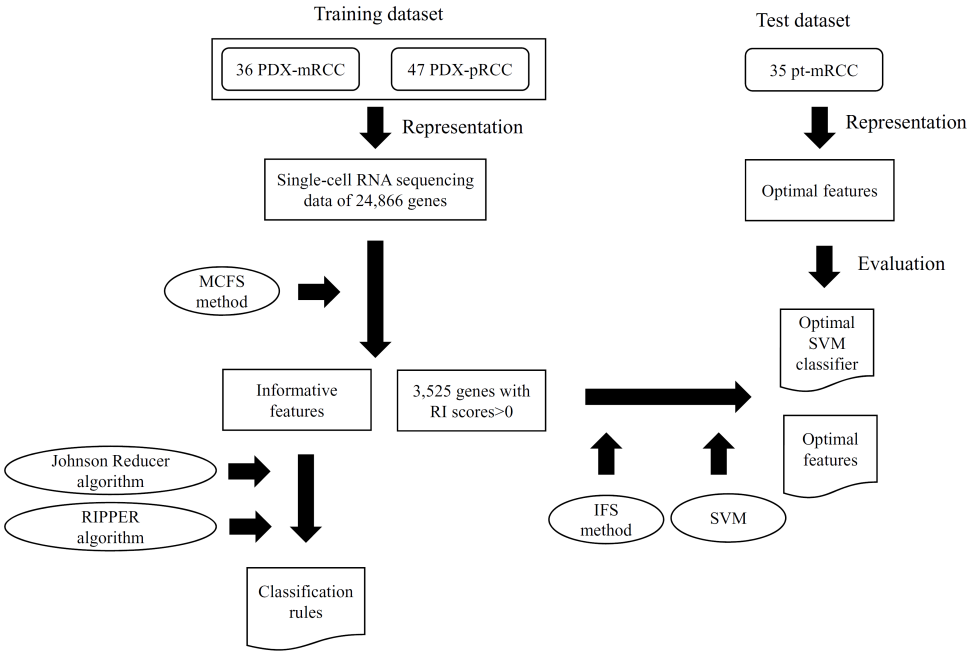


FIGURE 2. Whole procedures for analyzing single-cell RNA sequencing data in PDX-mRCC and PDX-pRCC. PDX-mRCC and PDX-mRCC samples constituted a training dataset, which were represented by single-cell RNA sequencing data of 24,866 genes. The training dataset was analyzed by the MCFS method, producing 31 informative features and a feature list containing 3,525 genes with RI scores larger than zero. Classification rules were constructed via Johnson Reducer and RIPPER algorithms with informative features. For the feature list, IFS method using SVM as the prediction engine was applied to extract optimal features and construct an optimal SVM classifier. The optimal SVM classifier was finally applied on pt-mRCC samples to evaluate its performance.

TABLE 1. Two produced classification rules for distinguishing PDX-mRCC and PDX-pRCC samples.

Rules	Criteria	Cancer
Rule1	MT2A >= 10.138	mRCC
Rule2	others	pRCC

than 0 are listed in Table S1. These features would be used in the IFS method and the rest features were discarded.

In addition, the MCFS method also provided 31 informative features, which were exact 31 top features in the feature list. Based on these 31 informative features, we further produced two rules listed in Table 1 using the Johnson reducer algorithm and RIPPER algorithm implemented in the MCFS program. To indicate the effectiveness of rules yielded by these two algorithms on 31 informative features, we tested them via 10-fold cross-validation three times, resulting in an SN of 0.963, SP of 1.000, ACC of 0.984, MCC of 0.968. It is quite effective.

B. RESULTS OF IFS METHOD WITH SVM

We also tried to select optimal feature subspace for SVM via the IFS method. We first selected 3,525 genes with RI scores >0, which were calculated via MCFS method. Second, a series of feature subsets with step 1 on the 3,525 genes were constructed. Third, the SVM classifier built on samples in the training dataset consisting of features from each constructed feature subset was evaluated by 10-fold cross-validation. The predicted results were counted as SNs, SPs, ACCs, and MCCs as mentioned in Section II-E, and they are all listed in Table S2. SVM could correctly classify all samples when the top four genes were used.

Furthermore, we evaluated the above SVM classifiers trained on PDX-mRCC and PDX-pRCC samples represented by different numbers of features to classify pt-mRCC samples in the test dataset. The results are provided in Table S3. We obtained an accuracy of 100% when the top 838 genes were used. Thus, we can construct an optimal SVM classifier on these 838 genes, which was trained on PDX-mRCC and PDX-pRCC and used to correctly classify pt-mRCC samples.

However, analyzing these 838 genes is an impossible task. Thus, it is necessary to reduce them. To this end, a curve was plotted in Figure 3 to show the trends of these SNs yielded by the SVM classifiers with 4-500 top genes. The SN was 0.829 when the top eight genes were used. Thus, we believed that 8 is a critical value for selecting most important genes to correctly classify pt-mRCC samples by the model trained on PDX-mRCC and PDX-pRCC samples.

C. COMPARISON OF IFS METHOD WITH C4.5 DECISION TREE

In this study, we selected SVM as the classification algorithm to construct the optimal classifier. In fact, we also tried another classic machine learning algorithm, C4.5, a classic decision tree [41]. For quickly implementing this algorithm,

TABLE 2. Detailed information of the top eight genes.

Rank	Ensembl ID	Gene symbol	Description	RI score
1	ENSG00000125148	MT2A	Metallothionein 2A	0.808
2	ENSG00000173110	HSPA6	Heat Shock Protein Family A (Hsp70) Member 6	0.763
3	ENSG00000260549	MT1L	Metallothionein 1L, Pseudogene	0.709
4	ENSG00000205426	KRT81	Keratin 81	0.634
5	ENSG00000164400	CSF2	Colony Stimulating Factor 2	0.614
6	ENSG00000163430	FSTL1	Follistatin Like 1	0.537
7	ENSG00000143156	NME7	NME/NM23 Family Member 7	0.515
8	ENSG00000092421	SEMA6A	Semaphorin 6A	0.453

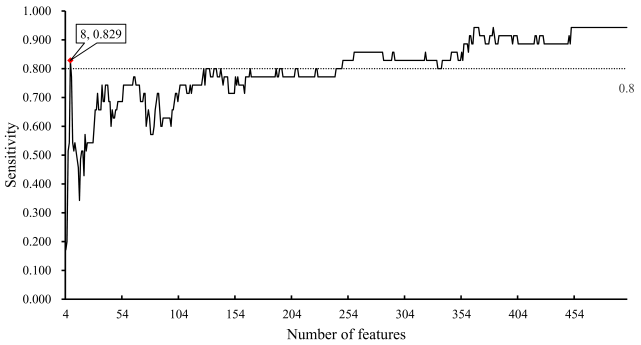


FIGURE 3. The trends of sensitivity on pt-mRCC samples corresponding to the number of features that were used to build a SVM classifier on PDX-mRCC and PDX-pRCC samples. It can be observed that the SVM classifier with top eight features can yield the sensitivity higher than 0.8. Thus, top eight features were deemed most important.

we employed the tool, called ''J48'', in Weka and executed it with its default parameters. For each of feature subsets constructed in the IFS method, a C4.5 classifier was constructed on PDX-mRCC and PDX-pRCC samples and evaluated via 10-fold cross-validation. Table S4 lists the evaluation results, including SNs, SPs, ACCs and MCCs. It can be observed that evaluation results were identical whenever how many features were selected. The MCC was 0.927, which was inferior to the perfect classification yielded by the optimal SVM classifier. Furthermore, we also tested the performance of these C4.5 classifiers on pt-mRCC samples, yielding an identical SN of 0.886, listed in Table S5, which was also lower than 1.000 that was produced by the optimal SVM classifier. All of these indicate that selection of SVM as the classification algorithm was a proper choice.

IV. DISCUSSION

As we have mentioned above, based on the expression profiles from Kim et al.'s study, we screened out eight functional genes (Table 2) that have distinctive expression patterns in primary and metastatic tumor tissues derived from PDX mouse model tumors. Based on the detailed quantitative gene expression level, we also set up two rules (Table 1) for further distinction on the two groups of PDX tumor tissues. All distinctive functions of the eight genes and rules could be validated by recent publications, reflecting the efficacy and accuracy of our results. Furthermore, such distinctions could be confirmed to be derived from the original primary and metastatic tumor tissues from the patients by test set results. These results confirmed that, on one hand, the PDX mouse model can directly reflect the expression profile distinction between primary and metastatic tumor tissues and, on the other hand, the genes and rules we screened out may accurately distinguish the two subgroups of tumor tissues, providing a novel computational tool for tumor studies. The detailed analysis of each gene and rule can be seen below.

A. ANALYSIS OF OPTIMAL DIFFERENTIALLY EXPRESSED GENES

MT2A (**ENSG00000125148**), encoding a metallothionein protein, has been widely reported to contribute to heavy metal binding and is transcriptionally regulated by both heavy metals and glucocorticoids [42]. Two studies [43], [44] that investigated the expression profile and drug sensitivity of primary and metastatic tumors by using a PDX mouse model confirmed that the expression distinction of such gene between primary and metastatic tumor tissues can still be identified in their respective PDX mouse model-containing tumor tissues, validating that such core distinctive expression marker of tumor metastasis can be stably expressed and identified in the PDX mouse model. Similarly, the next gene, named **MT1L** (**ENSG00000260549**), has also been widely reported to participate in glucocorticoid-associated biological processes [45]. An early study [45] on metallothionein proteins confirmed that such gene may contribute to the invasion and metastasis of renal carcinoma, indicating its differential expression pattern in tumors in vivo. As for its distinctive functions on tumors in the PDX mouse model, glucocorticoid-associated genes have been widely reported to maintain the differential expression pattern in a PDX mouse model [43], [44]. Therefore, the biological function of such gene may also be distinctive in primary and metastatic PDX tumor tissues.

Apart from such glucocorticoid-associated genes, the **HSPA6** (**ENSG00000173110**) gene has also been deemed to contribute to the identification of primary and metastatic tumor tissues in the PDX mouse model. As a member of heat shock protein family, HSPA6 has been reported to interact with HSP70-2 and has differential expression patterns in primary and metastatic clinical tumor samples [46], [47]. As for its expression profiles in the PDX mouse model, although

there are no direct reports on the expression pattern of such gene in patient-derived mouse model, the specific contribution [48], [49] to metastasis and invasion of heat shock protein family members, including this gene, in PDX models indicates that such gene may still be a proper identifier contributing to the distinction of primary and metastatic tumor-derived mouse models. **KRT81** (**ENSG00000205426**), encoding a basic protein to form the keratin of multiple tissues all over the human body, is functionally related to various tumor subtypes [50], [51]. As for its differential expression pattern in primary and metastatic tumor tissues, KRT81 has been confirmed to be associated with the survival and invasion ability of tumor cells in a new tumor microenvironment (such as the novel metastatic focus) [52]. As for its functional retention during PDX implantation and development, another study [53] on similar transplantation process confirmed that the differential expression pattern of such gene in different tissues, such as primary and metastatic tissues, can be stably retained.

CSF2 (**ENSG00000164400**) encodes a cytokine that controls the production, differentiation, and function of granulocytes and macrophages in the tumor microenvironment [54]. As for its distinctive expression pattern in primary and metastatic tumor lesions, considering the negative regulatory functions of such gene on tumor invasion and metastasis, it is quite reasonable to speculate that such gene may have differential expression patterns in primary and metastatic tumor tissues [55], [56]. As for the retention characteristics of such gene in the PDX mouse model, recent publications confirmed that, as one of the components of tumor microenvironment, myeloid-derived suppressor cells have a quite stable expression pattern in different tumor microenvironments, indicating that such gene may also be a stable biomarker for the identification of primary and metastatic tumor tissues in the PDX mouse model [57], [58]. **FSTL1** (**ENSG00000163430**), encoding an activing-binding protein, has also been predicted to have a differential expression pattern in primary and metastatic tumor-derived PDX mouse models. According to recent publications, FSTL1 has been widely reported to participate in tumor metastasis, indicating its differential expression pattern between primary and metastatic tumor tissues in vivo [59], [60]. By regulating the immune dysfunction of certain metastatic focus, such gene may maintain its unique differential expression pattern in the PDX mouse model after implantation due to the lack of immune selection. Such results have been verified by another individual study [61].

The **NME7** (**ENSG00000143156**) gene encodes a non-metastatic expressed nucleoside diphosphate kinase [62]. Only two studies on the NME protein family confirmed that such gene may directly regulate the metastatic biological functions of tumors with different expression profiles in primary and metastatic tumor tissues [63], [64]. As for the retention ability of the expression pattern in the PDX mouse model, based on an evolutionary study, such gene has a quite stable expression pattern in multiple vertebrate cell microenvironments [64]. Therefore, considering the stable expression pattern of such gene in alternative microenvironments, it is quite reasonable to speculate that such gene may also have a distinctive expression pattern in primary and metastatic tumor tissues in the PDX mouse model. The last gene **SEMA6A** (**ENSG00000092421**) has been generally reported to participate in normal granule cell migration [65]. As for its specific contribution on tumor metastasis, such gene has been widely reported to contribute to the progression, invasion, and metastasis of multiple tumor subtypes by interacting with semaphorins and their receptors [66], [67]. A novel study [68] on OTX2-driven stem cell confirmed that implanting patient-derived tumors in mouse model may not affect the original expression pattern of SEMA6A (a ligand for Plexin-A2), indicating that such gene can definitely distinguish primary and metastatic tumor tissues in PDX mouse model.

B. ANALYSIS OF OPTIMAL RULES FOR DISTINCTION

Apart from such qualitative distinctive genes, we also screened out two potential quantitative rules (Table 1) for the identification of metastatic tumor-derived mouse model. Only one unique functional gene, called **MT2A** (**ENSG00000125148**), is involved with the rules. As we have analyzed above, the differential expression patterns of such gene in primary and metastatic tumor-derived PDX mouse model have been verified by two experimental studies [43], [44]. Considering that the up-regulation of such gene directly promotes the invasion and metastasis of various tumor subtypes, it is quite reasonable to speculate that with expression level higher than 10.138 (number summarized from our training set), the test sample may be more reasonable to be derived from metastatic tumor tissues compared with primary ones [69].

Taken together, we screened out eight qualitative identifiers for the recognition of PDX mouse model derived from primary or metastatic tumor tissues. These genes have been confirmed to have differential expression patterns in primary and metastatic tumor tissues and may maintain their expression pattern during implantation. As for the quantitative rules, we only screened out a unique rule involving a qualitative gene that we have just predicted above, MT2A, which has also been validated by recent publications. In summary, based on the expression profile provided by Kim et al.'s study [18], we not only identified a group of stably expressed genes in the PDX model (which may distinguish metastatic tumor samples from primary tumor samples) but also established a quantitative rule for further validation on tumor metastatic tissues derived from the PDX mouse model.

REFERENCES

[1] Cancer Genome Atlas Research Network, "Comprehensive molecular characterization of clear cell renal cell carcinoma," *Nature*, vol. 499, pp. 43–49, Jul. 2013.

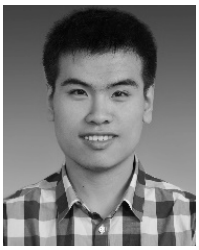
[2] A. J. Adeniran, B. Shuch, and P. A. Humphrey, "Hereditary Renal Cell Carcinoma Syndromes," *Amer. J. Surgical Pathol.*, vol. 39, pp. e1–e18, Dec. 2015.

[3] E. Jonasch, J. Gao, and W. K. Rathmell, "Renal cell carcinoma," *Proc. BMJ*, vol. 349, p. 4797, Nov. 2014.
[4] D. Su, L. Stamatakis, E. A. Singer, and R. Srinivasan, "Renal cell carcinoma: Molecular biology and targeted therapy," *Current Opinion Oncol.*, vol. 26, pp. 321–327, May 2014.
[5] M. Ghaouti, K. Znati, A. Jahid, F. Zouaidia, Z. Bernoussi, Y. El Fakir, and N. Mahassini, "A gallbladder tumor revealing metastatic clear cell renal carcinoma: Report of case and review of literature," *Diagnostic Pathol.*, vol. 8, p. 4, Jan. 2013.
[6] A. Afriansyah, A. R. Hamid, C. A. Mochtar, and R. Umbas, "Targeted therapy for metastatic renal cell carcinoma," *Acta Medica Indonesian*, vol. 48, pp. 335–347, Oct. 2016.
[7] S. G. D. Gangadaran, "Current management options in metastatic renal cell cancer," *Oncol. Rev.*, vol. 11, p. 339, Jun. 2017.
[8] G. A. Cirkel, P. Hamberg, S. Sleijfer, O. J. L. Loosveld, M. W. Dercksen, and M. Los, "Alternating treatment with Pazopanib and Everolimus vs continuous pazopanib to delay disease progression in patients with metastatic clear cell renal cell cancer the ROPETAR randomized clinical trial," *JAMA Oncol.*, vol. 3, pp. 501–508, Apr. 2017.
[9] "Biological therapy for metastatic renal cell cancer," *Lancet*, vol. 337, pp. 522–523, Mar. 1991. [Online]. Available: https://www.thelancet.com/journals/lancet/article/PII0140-6736(91)91302-B/fulltext
[10] B. Davidson, V. M. Abeler, M. Førsund, A. Holth, Y. Yang, Y. Kobayashi, L. Chen, G. B. Kristensen, I.-M. Shih, and T.-L. Wang, "Gene expression signatures of primary and metastatic uterine leiomyosarcoma," *Hum. Pathol.*, vol. 45, pp. 691–700, Apr. 2014.
[11] H. Zhao, Y. Li, S. Wang, Y. Yang, J. Wang, X. Ruan, Y. Yang, K. Cai, B. Zhang, P. Cui, J. Yan, Y. Zhao, E. K. Wakeland, Q. Li, S. Hu, and X. Fang, "Whole transcriptome RNA-seq analysis: Tumorigenesis and metastasis of melanoma," *Gene*, vol. 548, pp. 234–243, Sep. 2014.
[12] B. Weigelt, A. M. Glas, L. F. Wessels, A. T. Witteveen, J. L. Peterse, and L. J. van't Veer, "Gene expression profiles of primary breast tumors maintained in distant metastases," *Proc. Nat. Acad. Sci. U S A*, vol. 100, pp. 15901–15905, Dec. 2003.
[13] X. Zhang and M. T. Lewis, "Establishment of patient-derived xenograft (PDX) models of human breast cancer," *Current Protocols Mouse Biol.*, vol. 3, pp. 21–29, Mar. 2013.
[14] M. Chiron, R. G. Bagley, J. Pollard, P. K. Mankoo, C. Henry, L. Vincent, C. Geslin, N. Baltes, and D. A. Bergstrom, "Differential antitumor activity of aflibercept and bevacizumab in patient-derived xenograft models of colorectal cancer," *Mol. Cancer Therapeutics*, vol. 13, pp. 1636–1644, Jun. 2014.
[15] I. Lohse, A. Borgida, P. Cao, M. Cheung, M. Pintilie, T. Bianco, S Holter, E. Ibrahimov, R. Kumareswaran, R. G. Bristow, M.-S. Tsao, S. Gallinger, and D. W. Hedley, "BRCA1 and BRCA2 mutations sensitize to chemotherapy in patient-derived pancreatic cancer xenografts," *Brit. J. Cancer*, vol. 113, pp. 425–432, Jul. 2015.
[16] K. M. Miles, M. Seshadri, E. Ciamporcero, R. Adelaiye, B. Gillard, P. Sotomayor, K. Attwood, L. Shen, D. Conroy, F. Kuhnert, A. S. Lalani, G. Thurston, R. Pili, "Dll4 blockade potentiates the anti-tumor effects of VEGF inhibition in renal cell carcinoma patient-derived xenografts," *PLoS ONE*, vol. 9, Nov. 2014, e112371.
[17] C. Laurent, D. Gentien, S. Piperno-Neumann, F. Némati, A. Nicolas, B. Tesson, L. Desjardins, P. Mariani, A. Rapinat, X. Sastre-Garau, J. Couturier, P. Hupé, L. de Koning, T. Dubois, S. Roman-Roman, M.-H. Stern, E. Barillot, J. W. Harbour, S. Saule, and D. Decaudin, "Patient-derived xenografts recapitulate molecular features of human uveal melanomas," *Mol. Oncol.*, vol. 7, pp. 625–636, Jun. 2013.
[18] K.-T. Kim, H. W. Lee, H.-O. Lee, H. J. Song, D. E. Jeong, S. Shin, H. Kim, Y. Shin, D.-H. Nam, B. C. Jeong, D. G. Kirsch, K. M. Joo, and W.-Y. Park, "Application of single-cell RNA sequencing in optimizing a combinatorial therapeutic strategy in metastatic renal cell carcinoma," *Genome Biol.*, vol. 17, p. 80, Apr. 2016.
[19] M. Dramiński, A. Rada-Iglesias, S. Enroth, C. Wadelius, J. Koronacki, and J. Komorowski, "Monte Carlo feature selection for supervised classification," *Bioinformatics*, vol. 24, pp. 110–117, Jan. 2008.
[20] H. Liu and R. Setiono, "Incremental feature selection," *Appl. Intell.*, vol. 9, no. 3, pp. 217–230, 1998.
[21] L. Chen, J. Li, Y.-H. Zhang, K. Feng, S. Wang, Y. Zhang, T. Huang, X. Kong, and Y.-D. Cai, "Identification of gene expression signatures across different types of neural stem cells with the Monte-Carlo feature selection method," *J. Cell Biochem.*, vol. 119, pp. 3394–3403, Apr. 2018.
[22] M. Kruczyk, H. Zetterberg, O. Hansson, S. Rolstad, L. Minthon, A. Wallin, K. Blennow, J. Komorowski, and M. G. Andersson, "Monte Carlo feature selection and rule-based models to predict AlzheimerâĂŹs disease in mild cognitive impairment," *J. Neural Transmiss.*, vol. 119, pp. 821–831, Jul. 2012.
[23] D. Wang, J.-R. Li, Y.-H. Zhang, L. Chen, T. Huang, and Y.-D. Cai, "Identification of differentially expressed genes between original breast cancer and xenograft using machine learning algorithms," *Genes*, vol. 9, no. 3, p. 155, 2018.
[24] M. K. Dramiński, M. Kierczak, J. Koronacki, and J. Komorowski, "Monte Carlo feature selection and interdependency discovery in supervised classification," in *Advances in Machine Learning II*, vol. 2. Berlin, Germany: Springer, 2010, pp. 371–385.
[25] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *Proc. Int. joint Conf. Artif. Intell.*, 1995, pp. 1137–1145.
[26] L. Chen, S. Wang, Y.-H. Zhang, J. Li, Z.-H. Xing, J. Yang, T. Huang, and Y.-D. Cai, "Identify key sequence features to improve CRISPR sgRNA efficacy," *IEEE Access*, vol. 5, pp. 26582–26590, 2017.
[27] X. Zhao, L. Chen, Z.-H. Guo, and T. Liu, "Predicting drug side effects with compact integration of heterogeneous networks," *Current Bioinf.*, to be published.
[28] X. Zhao, L. Chen, and J. Lu, "A similarity-based method for prediction of drug side effects with heterogeneous information," *Math. Biosci.*, vol. 306, pp. 136–144, Dec. 2018.
[29] X.-Y. Pan and H.-B. Shen, "Robust prediction of B-factor profile from sequence using two-stage SVR based on random forest feature selection," *Protein Peptide Lett.*, vol. 16, no. 12, pp. 1447–1454, 2009.
[30] X. Pan and K. Xiong, "PredcircRNA: Computational classification of circular RNA from other long non-coding RNA using hybrid features," *Mol. Biosyst.*, vol. 11, no. 8, pp. 2219–2226, Aug. 2015.
[31] L. Chen, X. Pan, Y.-H. Zhang, X. Kong, T. Huang, and Y.-D. Cai, "Tissue differences revealed by gene expression profiles of various cell lines," *J. Cellular Biochem.*, vol. 120, pp. 7068–7081, May 2019.
[32] L. Chen, X. Pan, X. Hu, Y.-H. Zhang, S. Wang, T. Huang, and Y.-D. Cai, "Gene expression differences among different MSI statuses in colorectal cancer," *Int. J. Cancer*, vol. 143, no. 7, pp. 1731–1740, Oct. 2018.
[33] H. Cui and L. Chen, "A binary classifier for the prediction of EC numbers of enzymes," *Current Proteomics*, vol. 16, no. 5, pp. 381–389, 2019.
[34] I. H. Witten and E. Frank, Eds., *Data Mining Practical Machine Learning Tools and Techniques*, 2nd ed. San Francisco, CA, USA: Morgan, Kaufmann, 2005.
[35] J. Platt, "Sequential minimal optimizaton: A fast algorithm for training support vector machines," Microsoft, Redmon, WA, USA, Tech. Rep. MSR-TR-98-14,1998.
[36] D. S. Johnson, "Approximation algorithms for combinatorial problems," *J. Comput. Syst. Sci.*, vol. 9, no. 3, pp. 256–278, 1974.
[37] W. W. Cohen, "Fast effective rule induction," in *Proc. 12th Int. Conf. Mach. Learn.*, 1995, pp. 115–123.
[38] B. W. Matthews, "Comparison of the predicted and observed secondary structure of T4 phage lysozyme," *Biochim. Biophys. Acta-Protein Struct.*, vol. 405, no. 2, pp. 442–451, Oct. 1975.
[39] L. Chen, C. Chu, Y.-H. Zhang, M. Zheng, L. Zhu, X. Kong, and T. Huang, "Identification of drug-drug interactions using chemical interactions," *Current Bioinform.*, vol. 12, no. 6, pp. 526–534, 2017.
[40] T. Wang, L. Chen, and X. Zhao, "Prediction of drug combinations with a network embedding method," *Combinat. Chem. High Throughput Screening*, vol. 21, no. 10, pp. 789–797, 2018.
[41] R. Quinlan, *C4.5: Programs for Machine Learning*. San Mateo, CA, USA: Morgan Kaufmann Publishers, 1993.
[42] R. Giacconi, C. Cipriano, E. Muti, L. Costarelli, C. Maurizio, and V. Saba, "Novel -209A/G MT2A polymorphism in old patients with type 2 diabetes and atherosclerosis: Relationship with inflammation (IL-6) and zinc," *Biogerontology*, vol. 6, p. 407, Dec. 2005.
[43] B. Hong, Y. Yang, S. Guo, S. Duoerkun, X. Deng, D. Chen, S. Yu, W. Qian, Q. Li, Q. Li, K. Gong, and N. Zhang, "Intra-tumour molecular heterogeneity of clear cell renal cell carcinoma reveals the diversity of the response to targeted therapies using patient-derived xenograft models," *Oncotarget*, vol. 8, pp. 49839–49850, Jul. 2017.
[44] H. Zhao, R. Nolley, A. M. W. Chan, E. B. Rankin, and D. M. Peehl, "Cabozantinib inhibits tumor growth and metastasis of a patient-derived xenograft model of papillary renal cell carcinoma with MET mutation," *Cancer Biol. Therapy*, vol. 18, pp. 863–871, Aug. 2016.

[45] G. Hellemans, A. Soumillion, P. Proost, J. Van Damme, H. Van Poppel, and L. Baert, "Metallothioneins in human kidneys and associated tumors," *Nephron*, vol. 83, no. 4, pp. 331–340, 1999.
[46] S. Singh and A. Suri, "Targeting the testis-specific heat-shock protein 70-2 (HSP70-2) reduces cellular growth, migration, and invasion in renal cell carcinoma cells," *Tumor Biol.*, vol. 35, pp. 12695–12706, Dec. 2014.
[47] P. Kuballa, A.-L. Baumann, K. Mayer, U. Bär, H. Burtscher, and U. Brinkmann, "Induction of heat shock protein HSPA6 (HSP70B') upon HSP90 inhibition in cancer cell lines," *FEBS Lett.*, vol. 589, pp. 1450–1458, Jun. 2015.
[48] L. Mo, R. E. Bachelder, M. Kennedy, P. H. Chen, J.-T. Chi, A. Berchuck, G. Cianciolo, and S. V. Pizzo, "Syngeneic murine ovarian cancer model reveals that ascites enriches for ovarian cancer stem-like cells expressing membrane GRP78," *Mol Cancer Ther*, vol. 14, pp. 747–756, Mar. 2015.
[49] H. Jin, X. Cheng, Y. Pei, J. Fu, Z. Lyu, H. Peng, Q. Yao, Y. Jiang, L. Luo, and H. Zhuo, "Identification and verification of transgelin-2 as a potential biomarker of tumor-derived lung-cancer endothelial cells by comparative proteomics," *J. Proteomics*, vol. 136, pp. 77–88, Mar. 2016.
[50] M. Campayo, A. Navarro, N. Vinolas, R. Tejero, C. Muñoz, T. Diaz, M. L. Cabanas, J. M. Gimferrer, P. Gascon, J. Ramirez, M. Monzo, "A dual role for KRT81: A miR-SNP associated with recurrence in non-small-cell lung cancer and a novel marker of squamous cell lung carcinoma," *PLoS ONE*, vol. 6, Jul. 2011, Art. no. e22509.
[51] E. M. Noll *et al.*, "CYP3A5 mediates basal and acquired therapy resistance in different subtypes of pancreatic ductal adenocarcinoma," *Nature Med.*, vol. 22, pp. 278–287, Feb. 2016.
[52] N. Nanashima, K. Horie, T. Yamada, T. Shimizu, and S. Tsuchida, "Hair keratin KRT81 is expressed in normal and breast cancer cells and contributes to their invasiveness," *Oncol. Rep.*, vol. 37, pp. 2964–2970, May 2017.
[53] C. F. de Larrea, A. Navarro, R. Tejero, N. Tovar, T. Díaz, M. T. Cibeira, L. Rosiñol, G. Ferrer, M. Rovira, M. Rozman, M. Monzó, and J. Bladé, "Impact of MiRSNPs on survival and progression in patients with multiple myeloma undergoing autologous stem cell transplantation," *Clin Cancer Res*, vol. 18, pp. 704–3697, Jul. 2012.
[54] C. Y. Sasaki, P. Ghosh, and D. L. Longo, "Recruitment of RelB to the Csf2 promoter enhances RelA-mediated transcription of granulocyte-macrophage colony-stimulating factor," *J. Biol. Chem.*, vol. 286, pp. 1093–1102, Jan. 2011.
[55] X.-F. Liu, L. Xiang, Y. Zhang, K. G. Becker, T. K. Bera, and I. Pastan, "CAPC negatively regulates NF-κB activation and suppresses tumor growth and metastasis," *Oncogene*, vol. 31, pp. 1673–1682, Mar. 2012.
[56] E. J. Fertig, E. Lee, N. B. Pandey, and A. S. Popel, "Analysis of gene expression of secreted factors associated with breast cancer metastases in breast cancer subtypes," *Sci. Rep.*, vol. 5, p. 12133, Jul. 2015.
[57] O. Draghiciu, J. Lubbers, H. W. Nijman, and T. Daemen, "Myeloid derived suppressor cells-An overview of combat strategies to increase immunotherapy efficacy," *Oncoimmunology*, vol. 4, Jan. 2015, Art. no. e954829.
[58] S. M. Tham, K. H. Ng, S. H. Pook, K. Esuvaranathan, and R. Mahendran, "Tumor and microenvironment modification during progression of murine orthotopic bladder cancer," *Clin. Developmental Immunol.*, vol. 2011, Oct. 2011, Art. no. 865684.
[59] C. Kudo-Saito, T. Fuwa, K. Murakami, and Y. Kawakami, "Targeting FSTL1 prevents tumor bone metastasis and consequent immune dysfunction," *Cancer Res.*, vol. 73, pp. 6185–6193, Oct. 2013.
[60] C. Kudo-Saito, "FSTL1 promotes bone metastasis by causing immune dysfunction," *Oncoimmunology*, vol. 2, Nov. 2013, Art. no. e26528.
[61] M. C. Lau, K. Y. Ng, T. L. Wong, M. Tong, T. K. Lee, X. Y. Ming, S. Law, N. P. Lee, A. L. Cheung, Y. R. Qin, K. W. Chan, W. Ning, X. Y. Guan, and S. Ma, "FSTL1 promotes metastasis and chemoresistance in esophageal squamous cell carcinoma through NFκB-BMP signaling cross-talk," *Cancer Res.*, vol. 77, pp. 5886–5899, Nov. 2017.
[62] C.-H. Wang, N. Ma, Y.-T. Lin, C.-C. Wu, M. Hsiao, and F. L. Lu, "A shRNA functional screen reveals Nme6 and Nme7 are crucial for embryonic stem cell renewal," *Stem Cells*, vol. 30, pp. 2199–2211, Oct. 2012.
[63] M. Boissan and M.-L. Lacombe, "Learning about the functions of NME/NM_{23}: Lessons from knockout mice to silencing strategies," *Naunyn-Schmiedeberg's Arch. Pharmacol.*, vol. 384, pp. 421–431, Oct. 2011.
[64] T. Desvignes, P. Pontarotti, C. Fauvel, and J. Bobe, "Nme protein family evolutionary history, a vertebrate perspective," *BMC Evol. Biol.*, vol. 9, p. 256, Oct. 2009.
[65] J. Renaud and A. Chédotal, "Time-lapse analysis of tangential migration in Sema6A and PlexinA2 knockouts," *Mol. Cellular Neurosci.*, vol. 63, pp. 49–59, Nov. 2014.
[66] G. Neufeld, Y. Mumblat, T. Smolkin, S. Toledano, I. Nir-Zvi, K. Ziv, and O. Kessler, "The role of the semaphorins in cancer," *Cell Adhes. Migration*, vol. 10, pp. 652–674, Nov. 2016.
[67] V. A. Potiron, J. Roche, and H. A. Drabkin, "Semaphorins and their receptors in lung cancer," *Cancer Lett.*, vol. 273, pp. 1–14, Jan. 2009.
[68] Said Assou, T. Anahory, V. Pantesco, T. L. Carrour, F. Pellestor, B. Klein, L. Reyftmann, H. Dechaud, J. De Vos, and S. Hamamah, "The human cumulus-oocyte complex gene-expression profile," *Hum. Reproduction*, vol. 21, pp. 1705–1719, Jul. 2006.
[69] J. M. Arriaga, E. M. Levy, A. I. Bravo, S. M. Bayo, M. Amat, M. Aris, A. Hannois, L. Bruno, M. P. Roberti, F. S. Loria, A. Pairola, E. Huertas, J. Mordoh, and M. Bianchini "Metallothionein expression in colorectal cancer: Relevance of different isoforms for tumor progression and patient survival," *Hum. Pathol.*, vol. 43, pp. 197–208, Feb. 2012.

YANG JIANG received the Ph.D. degree from Jilin University, in 2006. He is currently a Professor in gastrointestinal surgery with the China–Japan Hospital of Jilin University. His research interests are in clinical and basic study of gastrointestinal tumors. In recent years, he has published 59 research articles, including 17 articles indexed by SCI, and cumulative impact factor >38.0. Moreover, he has hosted over eight research projects, including fund from the Science and Technology Department of Jilin Province and others.

XIAOYONG PAN received the Ph.D. degree in major bioinformatics from Copenhagen University, Denmark, in 2017, and the master's degree from Shanghai Jiao Tong University, in 2011. He held a postdoctoral position with the Erasmus Medical Center, Rotterdam, The Netherlands, from 2016 to 2018. He is currently an Assistant Professor with Shanghai Jiao Tong University. His research interests include bioinformatics, deep learning, and electronic health record. He is also a Lead Guest Editor of IEEE Access.



YU-HANG ZHANG was born in Jinzhou, Liaoning, China, in 1992. He received the B.S. degree in medical laboratory from Shanghai Jiaotong University Medical School, in 2014, and the Ph.D. degree in genetics from the Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, University of Chinese Academy of Sciences, in 2019. He is the author of more than 40 articles. His research interests include machine learning, liquid biopsy, and tumor immunotherapy. He was a recipient of merit student of the University of Chinese Academy of Sciences, in 2017.

TAO HUANG received the B.S. degree in bioinformatics from the Huazhong University of Science and Technology, Wuhan, China, in 2007, and the Ph.D. degree in bioinformatics from the Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai, China, in 2012.

Since 2014, he has been an Associate Professor and the Director of the Bioinformatics Core Facility, Institute of Health Sciences, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai. From 2012 to 2014, he was a Postdoctoral Fellow at the Department of Genetics and Genomics Sciences, Icahn School of Medicine at Mount Sinai, New York City, NY, USA. He has published over 100 articles. His works have been cited for over 3000 times with an h-index of 26 and an i10-index of 64. His research interests include bioinformatics, computational biology, systems genetics, and big data research. He has been a Reviewer of over 20 journals and an Editor/Guest Editor for seven journals and books.

YUFEI GAO is currently a Professor and the Director of the Neurosurgery Department, China–Japan Union Hospital of Jilin University, Changchun, China. He heads the Neuro-oncology Engineering Laboratory of Jilin Province, China–Japan Union Hospital. He is the Vice-Chairman of the Neurosurgery Branch of Jilin Provincial Medical Doctor Association. His clinical and research interests lie in the molecular mechanism of occurrence and development in central nervous system tumors. He is an expert in the pituitary surgery and skull base surgery through Endoscopic technique. In recent years, he has published 66 research articles, including 28 articles indexed by SCI, and cumulative impact factor >50.0. He has hosted over ten research projects, including the National Natural Science Foundation of China, the China Postdoctoral Fund from the Ministry of Education, funding from the Science and Technology Department of Jilin Province, and others.