A Novel Method to Efficiently Highlight Nonlinearly Expressed Genes

Qifei Wang¹, Haojian Zhang¹, Yuqing Liang¹, Heling Jiang¹, Siqiao Tan², Feng Luo³, Zheming Yuan¹ and Yuan Chen¹*

¹ Hunan Engineering & Technology Research Center for Agricultural Big Data Analysis & Decision-Making, Hunan Agricultural University, Changsha, China, ² School of Information Science and Technology, Hunan Agricultural University, Changsha, China, ³ School of Computing, Clemson University, Clemson, SC, United States

For precision medicine, there is a need to identify genes that accurately distinguish the physiological state or response to a particular therapy, but this can be challenging. Many methods of analyzing differential expression have been established and applied to this problem, such as t-test, edgeR, and DEseq2. A common feature of these methods is their focus on a linear relationship (differential expression) between gene expression and phenotype. However, they may overlook nonlinear relationships due to various factors, such as the degree of disease progression, sex, age, ethnicity, and environmental factors. Maximal information coefficient (MIC) was proposed to capture a wide range of associations of two variables in both linear and nonlinear relationships. However, with MIC it is difficult to highlight genes with nonlinear expression patterns as the genes giving the most strongly supported hits are linearly expressed, especially for noisy data. It is thus important to also efficiently identify nonlinearly expressed genes in order to unravel the molecular basis of disease and to reveal new therapeutic targets. We propose a novel nonlinearity measure called normalized differential correlation (NDC) to efficiently highlight nonlinearly expressed genes in transcriptome datasets. Validation using six real-world cancer datasets revealed that the NDC method could highlight nonlinearly expressed genes that could not be highlighted by t-test, MIC, edgeR, and DEseq2, although MIC could capture nonlinear correlations. The classification accuracy indicated that analysis of these genes could adequately distinguish cancer and paracarcinoma tissue samples. Furthermore, the results of biological interpretation of the identified genes suggested that some of them were involved in key functional pathways associated with cancer progression and metastasis. All of this evidence suggests that these nonlinearly expressed genes may play a central role in regulating cancer progression.

Keywords: RNA sequencing, maximal information coefficient, differential expressed gene, gene selection, normalized differential correlation
INTRODUCTION

Identifying and characterizing biomarkers that accurately reflect a physiological state (normal or diseased) or response to a particular drug or therapy is challenging. High-quality biomarkers are especially important for cancer detection and the development of safe, effective treatments. Such biomarkers are also the ultimate goal of many next-generation sequencing studies (Xiao et al., 2017; American Association for the Advancement of Science, 2018).

Many methods for analyzing differential expression in RNA sequencing (RNA-seq) data with the aim of finding genes that are differentially expressed across groups of samples have been established, such as t-test, limma (Gentleman et al., 2004), edger (Robinson et al., 2010), and DEseq2 (Love et al., 2014). However, these methods tend to consider only the linear relationship (differential expression) between gene expression and phenotype, but may overlook nonlinear relationships (as shown in Figure 1) because gene expression may differ between population groups or indeed between individuals, and can also vary as a patient’s status changes. As shown in Figure 1, both high and low levels of IGLC1 gene expression are found in non-tumor samples. The IGLC1 gene would thus generally be overlooked in cancer studies, with it giving a large $p$-value in the t-test (0.48). However, the expression of this gene is actually very useful to discriminate between control subjects and patients affected by prostate cancer.

The maximal information coefficient (MIC) was proposed to capture a wide range of associations of two variables, in both linear and nonlinear relationships (Reshef et al., 2011). Owing to its generality, MIC is becoming widely accepted in scientific research (Zhang et al., 2014), and is also used to analyze large biological datasets (Rau et al., 2013; Wang and Zhao, 2015; Wang et al., 2016). However, even with MIC, it is difficult to identify genes with a nonlinear expression pattern as the genes giving the most strongly supported hits are linearly expressed. This makes MIC less practical for the exploration of nonlinear informative genes in next-generation sequencing datasets such as those obtained by RNA-seq.

We developed a novel nonlinearity measure, normalized differential correlation (NDC), which could efficiently find nonlinearly expressed genes in RNA-seq datasets. We verified its validity on six real-world, publicly available cancer RNA-seq datasets (for details on the datasets, see Table 1). The results showed that it could discover most genes that were nonlinearly expressed across groups of samples, even though these genes could not be identified by popular differential expression analysis approaches. We also validated these genes using classification performance, gene function analysis, and survival analysis. As expected, the results confirmed that the expression of these genes is related to the phenotype.

DATASETS AND METHODS

Datasets

The details of the six gene expression RNA-seq datasets are summarized in Table 1. Level 3 data of TCGA were downloaded from the UCSC Xena platform (Goldman et al., 2019). For each dataset, the patient samples do not contain paracarcinoma tissue were excluded.

Methods

MIC can capture dependence between pairs of variables, including both functional and nonfunctional relationships. However, the ApproxMaxMI method provided by Reshef et al. (2011) results in a larger MIC score for paired variables under finite-sample conditions (Chen et al., 2016). Here, we use the improved algorithm ChiMIC to calculate the MIC value (Chen et al., 2016). The NDC score for a pair of data series $x$ (gene) and $y$ (phenotype) is defined as follows:

$$NDC(x,y) = \frac{MIC(x,y) - R^2(x,y) - Thre_{MIC}}{|R(x,y)|}$$  \hspace{1cm} (1)

Here, $R^2(x,y)$ and $|R(x,y)|$ are the determination coefficient and absolute value of the correlation coefficient, respectively. For a discrete phenotype, $R^2(x,y)$ and $R(x,y)$ are effective only for a binary class. $Thre_{MIC}$ is the confidence limit at level $p = 0.05$ based on a given sample size, can be calculated as follows:

Step 1: Expression data of one gene ($x_i$) are selected at random from the expression matrix and the sample order is shuffled;

| TABLE 1 | Six binary-class RNA-seq datasets. |
| Cancer types | No. of patients | No. of gene |
| Breast Cancer (BRCA) | 114 | 20,530 |
| Kidney Clear Cell Carcinoma (KIRC) | 72 | 20,530 |
| Liver Cancer (LIHC) | 50 | 20,530 |
| Lung Adenocarcinoma (LUAD) | 58 | 20,530 |
| Lung Cancer (LUNG) | 109 | 20,530 |
| Lung Squamous Cell Carcinoma (LUSC) | 51 | 20,530 |
Step 2: The MIC value between initial categorization labels (normal or diseased) and shuffled expression data is calculated.

Step 3: This process is repeated 1,000 times or more and these MIC values are sorted in ascending order. The MIC value at the 95% fractile is denoted as ThreMIC.

The NDC score is the normalized difference between nonlinear measure (MIC) and linear measure ($R^2$) by $|R|$. Because the MIC value is suspected of being large under finite-sample conditions, we use $\text{MIC}(x,y) - \text{Thre}_{\text{MIC}}$ to rule this out. We use the nonlinear score ("$\text{MIC}(x,y) - \text{Thre}_{\text{MIC}}$") minus the linear score ($R^2$) to exclude the linearly expressed genes. Lastly, we use $R(x,y)$ as the background normalized to exclude the genes that are irrelevant to the categorization labels. Therefore, a higher NDC score indicates a strong nonlinear association, but a weak linear correlation.

In Figures 2A–D, the genes are ranked by the NDC score, as well as the corresponding result in Figure 3. In other results, we first rank the genes in descending order of NDC score; then, we further sort the 1,000 most highly ranked genes in descending order by MIC value to obtain the most important nonlinear genes (top genes).

RESULTS AND DISCUSSION

Identification of Nonlinearly Expressed Informative Genes

Four popular differential expression analysis methods, namely, $t$-test, edgeR, DEseq2, and MIC, were selected here for comparison with the NDC measure. We took the typical linear method, $t$-test, as a benchmark. Figures 2A–F illustrate and compare the rank of each gene between the $t$-test and the other methods for lung squamous cell carcinoma (LUSC). As shown in Figure 2C, most of the top genes highlighted by MIC are linearly expressed. There

![Figure 2](https://example.com/figure2.png)

FIGURE 2 | Comparison of gene ranking by $t$-test, edgeR, DEseq2, maximal information coefficient (MIC), and normalized differential correlation (NDC) for the lung squamous cell carcinoma (LUSC) dataset. In (A–F), the color indicating data density; The red and black circles are just approximate region to visualize that, there are some genes can be highlighted by DNC methods, but cannot be highlighted by $t$-test. (G) overlaps of the top 100 genes selected by five methods.
are also some genes (red circled area) that can be detected by MIC but not by t-test, edgeR, and DEseq2; these can be defined as nonlinearly expressed genes. However, these nonlinearly expressed genes cannot be identified as important by MIC as it confers a low ranking on them. As indicated in Figures 2D–F, only the NDC measure can efficiently identify the importance of these nonlinearly expressed genes (such as the black circled area). We also used the rank correlation coefficient to quantify the results by different methods, as shown in Table 2. Figure 2G illustrates the overlaps of the top 100 genes selected by five methods. There are many overlaps among the top 100 genes selected by the t-test, edgeR, DEseq2, and MIC methods, but no overlap between the NDC measure and the other methods. The same results were also obtained for five other datasets (Figure 3). We therefore deduce that the NDC measure is able to identify important nonlinearly expressed genes, while the other four reference methods could only discover linearly expressed ones.

Figure 4 illustrates the top gene highlighted by the NDC measure for each dataset. The expression patterns of these genes are clearly the same as in Figure 1. From a data-driven perspective, these genes have high power to discriminate between the two classes. However, they achieve a low ranking with the reference methods. For example, with regard to PSAT1 for BRCA (Figure 4A), the rankings by t-test, DESeq2, edgeR, and MIC are 8,583rd, 4,642nd, 4,292nd, and 947th, respectively. The details of the rankings by the reference methods for the top genes highlighted by NDC are shown in Table 3.

Besides the expression patterns shown in Figure 4, another pattern was identified in KIRC, as shown in Figure 5. We found that the expression level of ERAP2 was stratified into four levels. Essentially, ERAP2 is a differentially expressed gene here, with 27 out of 72 patients having lower ERAP2 expression levels. ERAP2 is a proteolytic enzyme that acts in the endoplasmic reticulum, where it plays a central role in the trimming of peptides for
presentation by MHC class I molecules (MHC I) (Serwold et al., 2002; Saveanu et al., 2005). Forloni et al. (2010) reported that MHC I and ERAP2 are under the control of NF-κB through enhancer A in human neuroblastoma cells. Gadalla et al. (2013) also found that ERAP2 interacts with epithelial cell adhesion molecule (EpCAM) in breast cancer cells; EpCAM is a well-known epithelial and cancer cell “marker” (Tandon et al., 1990; Spizzo et al., 2006; Baeuerle and Gires, 2007). These patients in this study were also divided into two groups regarding their expression levels of ERAP2; this indicated that the expression of ERAP2 was strongly associated with the tumor grade (p-value of chi-squared test of 0.0084). Those in the group with high expression of ERAP2 have poorly differentiated KIRC (Table 4). All of this evidence suggests that ERAP2 may play a regulatory role in cancer. However, only the NDC method was able to identify the importance of ERAP2 (ranking: 14th); in contrast, all of the reference methods failed to identify the importance of this gene, with the rankings by t-test, DESeq2, edger, and MIC being 10,567th, 4,819th, 4,382, and 2,561th, respectively.

G1: well differentiated, G2: moderately differentiated, G3: poorly differentiated, G4: undifferentiated. Seven patients (TCGA-B8-5552, TCGA-CJ-5679, TCGA-CJ-5680, TCGA-CJ-5681, TCGA-CW-5591, TCGA-CZ-5456, TCGA-CZ-5469) are excluded due to inconsistent expression patterns.

### Classification Accuracy
To measure the impact of the nonlinearly expressed genes selected by NDC on the classification, we examined the classification accuracy on the six datasets with a supported vector classifier (SVC). SVC was performed using LIBSVM (Chang and Lin, 2011), which is available at http://www.csie.

| Gene symbol | t-test | edgeR | DESeq2 | MIC |
|-------------|--------|-------|--------|-----|
| PSAT1       | 8,583  | 4,642 | 4,282  | 947 |
| CCDC146     | 10,504 | 4,738 | 5,525  | 1,081 |
| MANEAL      | 8,983  | 6,521 | 7,713  | 216 |
| GJB1        | 13,072 | 6,049 | 5,969  | 1,336 |
| EFS         | 11,207 | 7,742 | 8,559  | 957 |
| SCN9A       | 13,660 | 9,275 | 7,842  | 1,091 |

### Table 4 | Contingency table for expression of ERAP2 and tumor grade in KIRC.

|          | Low expression | High expression |
|----------|----------------|-----------------|
| G1 or G2 | 16             | 10              |
| G3 or G4 | 10             | 29              |

FIGURE 4 | Expression pattern of top gene highlighted by the normalized differential correlation (NDC) measure for each dataset. (A) BRCA, (B) KIRC, (C) LIHC, (D) LUAD, (E) LUNG, (F) LUSC.

FIGURE 5 | Expression pattern of ERAP2 in KIRC. Tumor samples and paracarcinoma tissue samples from the same patient are connected by a black line.
Rigorous 10-fold cross-validation was used to evaluate the performance of the various gene selection algorithms. The dataset was randomly divided into 10 subsets. For each run, 9 out of 10 subsets were selected to be the training set and the remaining samples were used as test samples. For a fair comparison between gene selection algorithms, the training and test sets for each run were kept the same for all algorithms. In this section, in contrast to the case in the other sections, we could obtain 10 gene subsets for each method on each dataset.

Figure 6 illustrates the 10-fold cross-validation prediction accuracy using the top 10, 20, 30, 40, and 50 genes selected by NDC, as well as by t-test, edgeR, DEseq2, and MIC. The results reveal comparable accuracy for the five methods. This indicates that analysis of these nonlinearly expressed genes has sufficient ability to distinguish samples from cancerous tissue and from paracarcinoma tissue, via a machine learning approach.

Biological Interpretation of the Nonlinearly Expressed Genes
Enrichment Analysis of Pathways and Biological Processes
We further validated these top 100 genes selected by five methods using the LUSC dataset as an example, according to Metascape (A Gene Annotation & Analysis Resource) (Zhou et al., 2019), as shown in Figure 7. We found that these nonlinearly expressed genes selected by NDC (Figure 7E) were enriched in the ten terms, and these terms have no overlap between other terms enriched by reference methods. However, for the glutathione conjugation term, Fletcher et al. (2015) reported that glutathione-S-transferases (GSTs) catalyze the conjugation of glutathione with toxic oxidant compounds and were associated with acute and chronic inflammatory lung diseases. For the glycerophospholipid metabolism term, phosphoenolpyruvate carboxykinase (PEPCK) has been shown to promote cancer cell survival under conditions of nutrient deprivation, a typical feature in solid cancers, as well as cancer growth (Méndez-Lucas et al., 2014; Leithner et al., 2015; Montal et al., 2015; Vincent et al., 2015). Recently, Leithner et al. (2018) further showed that mitochondrial PEPCK (PEPCK-M) mediates the synthesis of glycerol phosphate from noncarbohydrate precursors, and that PEPCKM is needed to maintain the levels of glycerophospholipids, major constituents of biomembranes, in starved lung cancer cells.

Survival Analysis
For the nonlinearly expressed genes, as shown in Figure 9A, the patients were divided into two groups depending on their expression pattern. In the case of the PPL gene in the LUSC dataset, it showed higher expression in tumor tissue for 26 patients and in paracarcinoma tissue for 25 patients. We further compared the overall survival of the two groups of patients, with the results indicating that some nonlinearly expressed genes were strongly
associated with overall survival, such as PPL (hub gene for WGCNA), SCN9A (hub gene for WGCNA and the first ranked gene for NDC), EPS8 (second ranked gene for NDC), and MTAP (fourth ranked gene for NDC) for the LUSC dataset. As shown in Figures 9B–E, patients with the high expression of PPL in tumor tissue compared with the level in paracarcinoma tissue had significantly shorter overall survival (p = 0.033). SCN9A and EPS8 were the top 2 ranked genes selected by the NDC method, exhibiting similar patterns of expression (as shown in Figure 8). The expression of SCN9A was not significantly associated with overall survival (p = 0.113), but that of EPS8 was p = 0.038. It has been reported that EPS8 is strongly associated with tumor progression and metastasis (Wang et al., 2009; Mitra et al., 2011; Jeganathan et al., 2016; Fang et al., 2017). Furthermore, we found that PPL, SCN9A, and EPS8 are functionally related, as shown in Figure 9F. EPS8 is a crucial molecule that mediates EGFR-induced
activation of Akt (Innocenti, 2003) and PPL proteins as a localization signal in AKT1-mediated signaling by interacting with AKT1 (van den Heuvel et al., 2002; Wang et al., 2008). Akt activation is well known to be associated with the development of many tumors (Steelman et al., 2008; Jin et al., 2019), and it can also prevent FoxO3a, a tumor suppressor, from binding to EPS8. Moreover, it was reported that EGFR, via a U0126-sensitive ERK1/2 pathway, controls the transcriptional upregulation of SCN9A to promote cellular invasion in NSCLC cell lines (Campbell et al., 2013). MTAP, a tumor suppressor gene, has been reported to be associated with the overall survival of NSCLC patients, which is consistent with our results (p = 0.04) (Su et al., 2014). Furthermore, the expression pattern of MTAP is strongly correlated with the vital status of LUSC patients (p = 0.002). A total of 28 out of 36 patients with the high expression of MTAP in tumor tissue compared with the level in paracarcinoma tissue died, but only 4 out of 15 with a low MTAP expression level in tumor tissue did so.

**CONCLUSION**

In this paper, we propose a novel nonlinearity measure named NDC to efficiently identify nonlinearly expressed genes in transcriptome datasets. Validation using six real-world cancer datasets revealed that the NDC method could identify nonlinearly expressed genes that were overlooked by t-test, MIC, edgeR, and DEseq2, although MIC could capture nonlinear correlations. The results regarding the classification accuracy indicate that these genes have sufficient ability to distinguish cancer and paracarcinoma tissue samples. Moreover, the results of biological interpretation of these genes also suggest that some of them are involved in key functional pathways associated with cancer progression and metastasis. All of this evidence suggests that these nonlinearly expressed genes may play central roles in regulating cancer progression. Interestingly, as shown in Figures 4 and 5, analysis of these nonlinearly expressed genes proved that genes could be expressed in different patterns in different patients. This explains the need for the development of precision medicine, but also the challenges associated with this. Not all of the top-ranked nonlinearly expressed genes or hub genes found here have previously been reported to correlate with cancer progression and metastasis, but the NDC method suggests their importance as informative genes. The approach presented here suggests that these genes warrant attention as potential targets for therapy and disease risk predictors, as well as for their ability to achieve a clinical diagnosis and evaluate therapeutic efficacy.

**DATA AVAILABILITY STATEMENT**

Publicly available datasets were analyzed in this study. This data can be found here: https://xenabrowser.net/datapages/
and the NDC algorithm can be found here: https://github.com/chenyuan0510/normalized-differential-correlation-NDC.git.

AUTHOR CONTRIBUTIONS

All authors contributed to the methodology proposal, analysis of results, writing, and review of the manuscript.

REFERENCES

American Association for the Advancement of Science. (2018). sponsored collection) precision medicine and cancer immunology in China. Science 359 (6375), 598. doi: 10.1126/science.359.6375.598-b

Baeuerle, P. A., and Gires, O. (2007). EpCAM (CD326) can be found here: https://github.com/chenyuan0510/normalized-differential-correlation-NDC-.git.

Wang et al. Highlight Nonlinearly Expressed Genes Efficiently

FUNDING

This work was supported by the National Natural Science Foundation of China (61701177); Hunan Provincial Natural Science Foundation (2018J3225); Science foundation open project of Hunan Provincial Key Laboratory of Crop Germplasm Innovation and Utilization (18KFXM08); and Hunan Provincial Research learning and innovative experimental project for college students (SCX1826).

Mol. Cancer

with non-small cell lung cancer. Biochem. Biophys. Rep. 18, 100625. doi: 10.1016/j.bbrep.2019.100625

Langfelder, P., and Horvath, S. (2008). WGCNA: An R package for weighted correlation network analysis. BMC Bioinf. 9 (1), 559. doi: 10.1186/1471-2105-9-559

Leitner, K., Krähenbühl, A., Trötzmüller, M., Moustafa, T., Kofeler, H. C., Wohlkönig, C., et al. (2015). PCK2 activation mediates an adaptive response to glucose depletion in lung cancer. Oncogene 34 (8), 1044–1050. doi: 10.1038/onc.2014.47

Leitner, K., Triebel, A., Trötzmüller, M., Hinteregger, B., and Krähenbühl, A. (2018). The glycerol backbone of phospholipids derives from noncarbohydrate precursors in starved lung cancer cells. Proc. Natl. Acad. Sci. U. S. A. 115 (24), 6225–6230. doi: 10.1073/pnas.1719871115

Lim, W. T., Zhang, W., Miller, C., Watters, J., Gao, F., Viswanathan, A., et al. (2017). PTEN and phosphorylated AKT expression and prognosis in early- and late-stage non-small cell lung cancer. Oncol. Rep. 17 (4), 853–857. doi: 10.3892/or.17.4.853

Love, M. I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-Seq data with DESeq2. Genome Biol. 15 (12), 550. doi: 10.1186/s13059-014-0550-8

Méndez-Lucas, A., Hyroshlová, P., Novellasdemunt, L., Viñals, F., and Perales, J. C. (2014). Mitochondrial phosphoenolpyruvate carboxykinase (PEPCK-M) is a pro-survival, endoplasmic reticulum (ER) stress response gene involved in tumor cell adaptation to nutrient availability. J. Biol. Chem. 289 (32), 22090–22102. doi: 10.1074/jbc.M114.566927

Mitra, S., Lee, J. S., Cantrell, M., and Van, D. B. C. L. (2011). C-Jun N-terminal kinase 2 (JNK2) enhances cell migration through epithelial growth factor substrate 8 (EPS8). J. Biol. Chem. 286 (17), 15287–15297. doi: 10.1074/jbc.M109.094441

Montal, E., Dewi, R., Bhalla, K., Ou, L., Hwang, B., Ropell, A., et al. (2015). PECK coordinates the regulation of central carbon metabolism to promote cancer cell growth. Mol. Cell 60 (4), 571–583. doi: 10.1016/j.molcel.2015.09.025

Rau, C. D., Nicholas, W., Orozco, L. D., Brian, B., James, W., and Lusis, A. J. (2013). Maximal information component analysis: a novel non-linear network analysis method. Front. Genet. 4, 28. doi: 10.3389/fgene.2013.00028

Reshef, D. N., Reshef, Y. A., Finucane, H. K., Grossman, S. R., Mcvean, G., Turnbaugh, P. J., et al. (2011). Detecting Novel Associations in Large Data Sets. Science 334 (6062), 1518–1524. doi: 10.1126/science.1205438

Robinson, M. D., McCarthy, D. J., and Smyth, G. K. (2010). edgR: a bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26 (1), 139–140. doi: 10.1093/bioinformatics/btp616

Saveau, L., Carroll, O., Lindo, V., Del Val, M., Lopez, D., epelletier, Y., et al. (2005). Concentrated peptide trimming by human ERAP1 and ERAP2 aminopeptidase complexes in the endoplasmic reticulum. Nat. Immunol. 6 (7), 689. doi: 10.1038/nm1208

Serwold, T., Gonzalez, F., Kim, J., Jacob, R., and Shastri, N. (2002). ERAAP customizes peptides for MHC class I molecules in the endoplasmic reticulum. Nature 419 (6906), 480. doi: 10.1038/nature01074

Singh, D., Febo, P. G., Ross, K., Jackson, D. G., Manola, J., Ladd, C., et al. (2002). Gene expression correlates of clinical prostate cancer behavior. Cancer Cell 1 (2), 203–209. doi: 10.1016/S1535-6108(02)00030-2

Spizzio, G., Went, P., Dinh Hofer, S., Obrist, P., Moch, H., Baeuerle, P. A., et al. (2006). Overexpression of epithelial cell adhesion molecule (Ep-CAM) is an independent prognostic marker for reduced survival of patients with epithelial...
ovarian cancer. *Gynecol. Oncol.* 103 (2), 483–488. doi: 10.1016/j.ygyno.2006.03.035

Steelman, L. S., Stadelman, K. M., Chappell, W. H., Horn, S., Basecke, J., Cervello, M., et al. (2008). Akt as a therapeutic target in cancer. *Expert Opin. Ther. Targets* 12 (9), 1139–1165. doi: 10.1517/14728222.12.9.1139

Su, C. Y., Chang, Y. C., Chan, Y. C., Lin, T. C., Huang, M. S., Yang, C. J., et al. (2014). MTAP is an independent prognosis marker and the concordant loss of MTAP and p16 expression predicts short survival in non-small cell lung cancer patients. *Eur. J. Surg. Oncol.* 40 (9), 1143–1150. doi: 10.1016/j.ejso.2014.04.017

Tandon, A. K., Clark, G. M., Chamness, G. C., and McGuire, W. L. (1990). Association of the 323/A3 surface glycoprotein with tumor characteristics and behavior in human breast cancer. *Cancer Res.* 50 (11), 3317–3321. doi: 10.1097/00002820-199006000-00011

van den Heuvel, A. P., de Vries-Smits, A. M., van Weeren, P. C., Dijkers, P. F., de Bruyn, K. T., Riedl, J. A., et al. (2002). Binding of protein kinase B to the plakin family member periplakin. *J. Cell Sci.* 115 (20), 3957–3966. doi: 10.1242/jcs.00069

Vincent, E. E., Sergushichev, A., Griss, T., Gingras, M. C., Samborska, B., Ntumbane, T., et al. (2015). Mitochondrial phosphoenolpyruvate carboxykinase regulates metabolic adaptation and enables glucose-independent tumor growth. *Mol. Cell.* 60 (2), 195–207. doi: 10.1016/j.molcel.2015.08.013

Wang, J., Yuan, Y., Zhou, Y., Guo, L., Zhang, L., Kuai, X., et al. (2008). Protein interaction data set highlighted with human Ras-MAPK/P13K signaling pathways. *J. Proteome Res.* 7 (9), 3879–3889. doi: 10.1021/pr8001645

Wang, H., Patel, V., Miyazaki, H., Gutkind, J. S., and Yeudall, W. A. (2009). Role for EPS8 in squamous carcinogenesis. *Carcinogenesis* 30 (1), 165–174. doi: 10.1093/carcin/bgn252

Wang, S., and Zhao, Y. (2015). Analysing large biological data sets with an improved algorithm for MIC. *Int. J. Data Min. Bioinform.* 13 (2), 158. doi: 10.1504/IJDMB.2015.071548

Wang, C., Dai, D., Li, X., Wang, A., and Zhou, X. (2016). SuperMIC: analyzing large biological datasets in bioinformatics with maximal information coefficient. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 14 (4), 783–795. doi: 10.1109/TCBB.2016.2550430

Xiao, C. L., Chen, Y., Xie, S. Q., Chen, K. N., Wang, Y., Han, Y., et al. (2017). Mecat: fast mapping, error correction, and de novo assembly for single-molecule sequencing reads. *Nat. Methods* 14, 1072–1074. doi: 10.1038/nmeth.4432

Zhang, Y., Jia, S., Huang, H., Qiu, J., and Zhou, C. (2014). A Novel algorithm for the precise calculation of the maximal information coefficient. *Sci. Rep.* 4, 6662. doi: 10.1038/srep06662

Zhou, Y., Zhou, B., Pache, L., Chang, M., Khodabakhshi, A. H., Tanaseichuk, O., et al. (2019). Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. *Nat. Commun.* 10 (1), 1523. doi: 10.1038/s41467-019-09234-6

Conflicts of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2020 Wang, Zhang, Liang, Jiang, Tan, Luo, Yuan and Chen. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.