Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

GENESO a framework for pan-cancer classification and marker gene discovery by symmetrical occlusion method using deep learning

Xing Zhao
Chinese University of Hong Kong
Huating Wang
Chinese University of Hong Kong
Hao Sun
haosun@cuhk.edu.hk

Chinese University of Hong Kong

Research Article

Keywords: Pan-cancer classification, Marker gene identification, Deep neural network, Long Short Term Memory, Occlusion, GENESO

Posted Date: December 7th, 2023

DOI: https://doi.org/10.21203/rs.3.rs-3550032/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Version of Record: A version of this preprint was published at BMC Bioinformatics on August 8th, 2024.
See the published version at https://doi.org/10.1186/s12859-024-05870-y.

GENESO: a framework for pan-cancer classification and marker gene discovery by symmetrical occlusion method using deep learning

Xing Zhao
xingz@link.cuhk.edu.hk
Department of Orthopaedics and Traumatology, The Chinese University of Hong Kong
Hong Kong, China

Huating Wang
huating.wang@cuhk.edu.hk
Department of Orthopaedics and Traumatology, The Chinese University of Hong Kong
Hong Kong, China

Hao Sun*
haosun@cuhk.edu.hk
Department of Chemical Pathology, The Chinese University of Hong Kong
Hong Kong, China

ABSTRACT

Quantitative measurement of RNA expression levels through RNA-Seq is an ideal replacement for conventional cancer diagnosis via microscope examination. Currently, cancer-related RNA-Seq studies focus on two aspects: classifying the status and tissue of origin of a sample and discovering marker genes. Existing studies typically identify marker genes by statistically comparing healthy and cancer samples. However, this approach overlooks marker genes with low expression level differences and may be influenced by experimental results. In this paper, we introduce "GENESO," a novel framework for pan-cancer classification and marker gene discovery using the symmetrical occlusion method in conjunction with deep learning. we first trained a baseline deep LSTM neural network capable of distinguishing the origins and statuses of samples utilizing RNA-Seq data. Then, we propose a novel marker gene discovery method called "Symmetrical Occlusion". It collaborates with the baseline LSTM network, mimicking the "gain of function" and "loss of function" of genes to quantitatively evaluate their importance in pan-cancer classification. By identifying the genes of utmost importance, we then isolate them to train new neural networks, resulting in higher performance LSTM models that utilize only a reduced set of highly relevant genes. The baseline neural network achieves an impressive validation accuracy of 96.59% in pan-cancer classification. Through the integration of SO and selecting the top 33% of genes ranked by importance, the accuracy of the second network reaches an accuracy of 98.30%, while using 67% fewer genes. Notably, our method excels in identifying marker genes that are not differentially expressed. Moreover, we assessed the feasibility of our method using single-cell RNASeq data, employing known marker genes as a validation test.

CCS CONCEPTS

• **Applied computing** → *Bioinformatics*; **Computational genomics**; • **Computing methodologies** → **Neural networks**.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
Conference'17, July 2017, Washington, DC, USA
© 2023 Association for Computing Machinery.
ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

KEYWORDS

Pan-cancer classification, Marker gene identification, Deep neural network, Long Short Term Memory, Occlusion, GENESO

ACM Reference Format:
Xing Zhao, Huating Wang, and Hao Sun*. 2023. GENESO: a framework for pan-cancer classification and marker gene discovery by symmetrical occlusion method using deep learning. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Cancer, being a leading global cause of mortality, has spurred significant research efforts to advance diagnosis and treatment techniques [1]. Cancer prediction primarily revolves around identifying factors like susceptibility, recurrence, and prognosis. Meanwhile, cancer detection aims to accurately classify tumor types and discern markers for each cancer, especially in its early stages. However, the conventional clinical method for cancer detection, relying on manual searches for morphological tumor characteristics through microscope, demands extensive training to distinguish nuances and is susceptible to human error. Moreover, cancers often exhibit heterogeneity, leading to undetected structures when examined using conventional microscopes [2]. Another challenge arises from the fact that conventional classification methods, based on the analysis of "Differentially Expressed Genes" (DEG) and other statistical algorithms, might overlook interactions between genes with low expression level variations [3]. Consequently, the development of a method for quantitatively evaluating the significance of genes across all cancer types for pan-cancer classification remains an unsolved problem.

To address these challenges, we present "GENESO," a novel framework for pan-cancer classification and marker gene discovery, employing the symmetrical occlusion method through deep learning. In summary, our framework utilizes a deep learning model on NGS RNASeq data from clinical samples to perform pan-cancer classification, predicting the samples' status and tissue-of-origin, while simultaneously evaluating the importance of each gene through the symmetrical occlusion method.

2 METHOD

2.1 Dataset Preprocessing

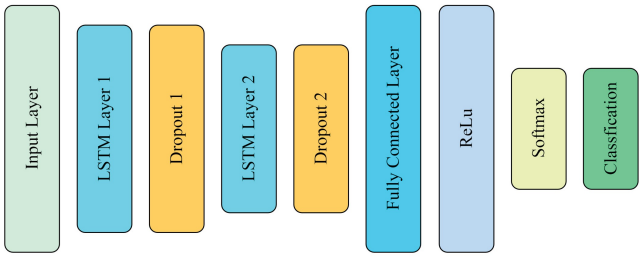
As tumors exhibit considerable heterogeneity, sub-populations of tumor cells may display distinct genotypes and phenotypes within a primary tumor or even among tumors of the same histopathological subtype [4, 5]. To ensure the comprehensiveness of our dataset, we acquired RNA-Seq data from various public sources, detailing the tissue of origin and status of the samples, as shown in **Additional File.1**. The dataset includes paired tumor and normal samples, along with some rare tumors. The dataset contains paired tumor and normal samples as well as some rare tumors. For instance, Petrini et al.'s dataset comprises 286 thymus tumor samples from patients of diverse sex, age, tumor types, and stages [6]. We also collected RNA-Seq data on rare tumors such as paraganglioma and sarcoma from Snezhkina et al. and Lesluyes et al., respectively [7, 8]. To organize the dataset, we first categorized the RNA-Seq data based on the tissue of origin, further dividing it according to the status of each sample (e.g., cancer, normal), resulting in 28 classes. We opted not to divide the dataset into finer classes to maintain an adequate number of samples in each class. All RNA-Seq data was aligned against the human reference genome GRCh.38 using "Bowtie2," while "featurecounts" was employed to quantify the number of mapped reads for each corresponding gene, excluding genes on sex chromosomes without coverage across all classes [9, 10]. No batch effect correction is made to the read count to avoid introducing systematic bias.

2.2 Expression Level Normalization

Various methods have been employed for the normalization of read count data in previous studies, with the most common approaches being "Transcripts Per Kilobase Million" (TPM) and "Reads Per Kilobase Million" (RPKM). These methods are designed to mitigate the bias introduced by gene length by taking transcript length into account and performing a multiplication of a constant on the result as shown in **Eq.1**.

$$
\begin{aligned}
RPKM &= 10^9 \times \frac{Reads\ mapped\ to\ gene}{Total\ reads\ \times\ Gene\ length} \\
TPM &= 10^6 \times \frac{\frac{Reads\ mapped}{Gene\ length}}{\sum \frac{Readsmapped}{Genelength}} \\
NRC &= \frac{Reads\ mapped}{Total\ mapped\ reads}
\end{aligned}
\tag{1}
$$

However, when it comes to pan-cancer classification using neural networks, neither TPM nor RPKM prove to be suitable. Instead, we employ "Normalized Counts" (NRC), which eliminates the constant number and gene length from the normalization process. There are three reasons why NRC is more suitable for neural network processing: **(1)** Redundant constant number: The Z-score normalization process within the neural network renders constants in the equations of TPM and RPKM unnecessary. **(2)** Suitability for cross-sample comparisons: Extensive research has demonstrated that TPM and RPKM are more appropriate for comparing transcript expression within a single sample rather than facilitating cross-sample comparisons [11–14]. **(3)** Gene length influence in neural network: During neural network prediction, genes are compared across multiple samples. Consequently, the influence of gene length remains consistent throughout all samples, enabling us to remove it without compromising the classification performance.



Name	Parameters
Input Layer	Size: 1x57584 (Z-score normalization)
LSTM Layer 1	Size: 120 LSTM cells (sequence mode)
Dropout 1	0.25 dropout probability
LSTM Layer 2	Size: 80 LSTM cells (last mode)
Dropout 2	0.25 dropout probability
Fully Connected Layer	Size: 1x28

Figure 1: Neural network architecture

2.3 Neural network architecture

$$
\begin{aligned}
i_t &= \sigma(W_x i x_t + W_h i h_(t-1) + W_c i c_(t-1) + b_i) \\
f_t &= \sigma(W_x f x_t + W_h f h_(t-1) + W_c f c_(t-1) + b_f) \\
c_t &= f_t c_(t-1) + i_t tanh(W_x c x_t + W_h c h_(t-1) + b_c) \\
o_t &= \sigma(W_x o x_t + W_h o h_(t-1) + W_c o c_t + b_0) \\
h_t &= o_t tanh(c_t)
\end{aligned}
\tag{2}
$$

As stipulated in **Eq.2**, the building block of LSTM layer, a LSTM cell has three inputs and two outputs: sequence input x_t, cell state input from previous LSTM cell $c_(t-1)$, hidden state input from previous LSTM cell $h_(t-1)$, hidden state output h_t, and cell state output c_t . In another word, hidden state h_t contains the "short-term memory" while cell state c_t stores the "long-term memory" or "global memory".

As shown in **Fig.1**, the first layer of the neural network functions as the input layer by normalizing the input data using "Z-score" normalization. This step is crucial to ensure optimal network performance as it allows the network to compare gene expression levels across samples without being affected by differences in scale. Following the input layer, the neural network consists of two LSTM layers: the first with 120 cells and the second with 80 cells. This design enables the first layer to abstract normalized inputs, extract high-level information, and outperform convolutional neural networks that leverage multiple convolutional layers [15]. To prevent over-fitting and promote the identification of general patterns in the dataset, a dropout layer with a dropout probability of 30% is

applied after each LSTM layer [16]. At the end of the neural network, a "fully connected" (FC) layer is utilized to link the LSTM layer and the output layer, which generates the predicted label for the network. The softmax and classification layer outputs the prediction labels based on the FC layer output, identifying the label with the highest prediction score.

2.4 Marker Gene Identification

In order to identify marker genes and quantitatively evaluate their for pan-cancer classification, a novel identification algorithm named "Symmetrical Occlusion" (SO) is introduced in this study. This method draws inspiration from "CNN occlusion" (CO), a feature identification method used in neural network-based image classification, but it overcomes the limitations of CO in signature gene identification.

2.4.1 CNN occlusion method and its drawbacks. CO method provides an intuitive and interpretable approach for identifying regions of importance within an image, and it has been widely utilized in image classification [17, 18]. The theory of CO hypothesis that blocking or occluding a region that is crucial for image classification typically results in a sharp decrease in the prediction score, which indicates the probability of the input image belonging to the corresponding class. This change can rapidly and accurately identify the important regions for the neural network. However, applying CO directly to marker gene identification poses challenges due to differences in data structure between images and NRC. While an image is a 2D matrix of pixels, gene expression levels are stored in a vector based on gene position. Additionally, adjacent pixels in an image often exhibit relationships, whereas neighboring elements in gene expression levels are independent. Furthermore, pixel values in an image are fixed between 0 and 255, while gene expression ranges differ between samples.

2.4.2 Marker gene identification. SO is a method for determining the importance of individual genes in pan-cancer classification using a neural network. This is achieved by mimicking the "gain of function" and "loss of function" of genes, which involves increasing or decreasing the expression level compared to the original expression level, respectively. While the gene expression level may exceed their biological limitations, which is biologically impossible, it helps in quantifying their importance in the classification process.

The SO method utilizes a multi-step process to quantify the importance of a specific gene in pan-cancer classification. First, a baseline LSTM neural network using the full list of genes as input (**Fig.2A**). Next, the importance of a specific gene is quantified using occlusion by simulating its "loss of function" and "gain of function" effects. For the "loss of function" simulation, the expression level of the gene in a reference sample from the validation dataset is reduced in a step-wise manner. This generates a set of pseudo samples which have lower expression levels the gene of interest. These pseudo samples are then input into the baseline LSTM neural network, which outputs prediction scores for each pseudo sample which will be used to quantify the importance of the gene in pan-cancer classification (**Fig.2B**). Finally, this process is then repeated for "gain of function" simulations, where the expression level of the gene is increased.

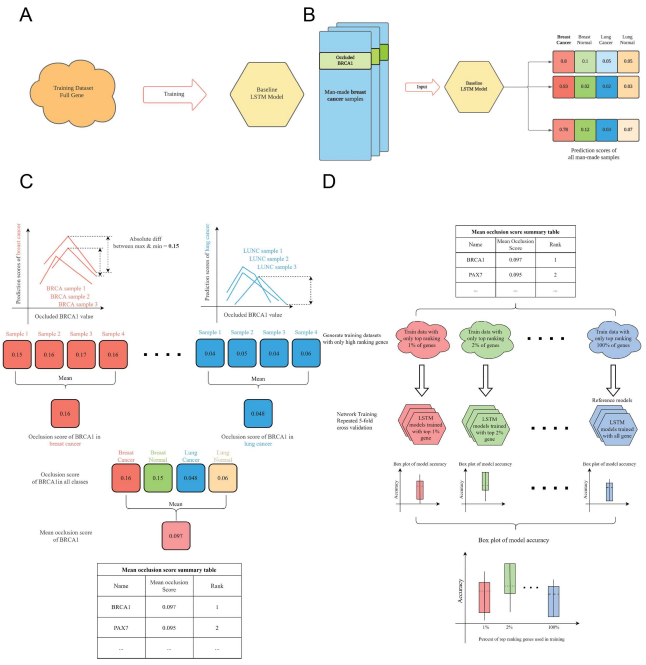


Figure 2: (A) A baseline LSTM model is trained first for the occlusion. (B) Man-made samples which generated by occluding the genes of samples in validation dataset is forwarded to baseline model which outputs prediction scores for each man-made sample. (C) For each gene, plots between the occlusion value and its corresponding in all 28 classes are drawn and the absolute difference between maximum and minimum is calculated. The occlusion score of specific class is the average of absolute differences of multiple samples to ensure robustness. The mean occlusion score of a gene is the average of its occlusion score which is used for compiling a summary table. (D) New neural networks will be trained using different percent genes to find the gene combination to achieve optimal performances.

2.4.3 Neural Network Optimization. After all pseudo samples have been processed by the baseline LSTM neural network, their prediction scores are combined and the absolute value between the maximum and minimum prediction score is calculated for each pseudo sample. A higher absolute value indicates a stronger influence of the gene on the neural network's classification of the pseudo sample into its corresponding class. For example, the gene BRCA1 may have a higher influence on the neural network's classification of a sample into the "breast cancer" class than the "lung cancer" class, resulting in a higher absolute value (**Fig.2C**). To summarize the results and quantitatively measure the importance of a gene in a specific class, the "occlusion score" is introduced and obtained by calculating the average of the absolute values. Next, to measure its importance in pan-cancer classification, the mean value of the occlusion scores of this gene in across classes (e.g. breast cancer, breast normal, lung cancer) is calculated and stored for further use. In order to improve the pan-cancer classification accuracy of the

LSTM neural network using fewer genes, the genes in the summary table are ranked by their mean occlusion score. New training datasets are generated by selecting different portions of the top-ranking genes, and new LSTM neural networks are trained using five-fold cross-validation (**Fig.2D**). The network with the highest validation accuracy is then selected as the final output. This approach allows for the identification of a smaller subset of genes that can still achieve high classification accuracy, making it more efficient and cost-effective for future studies.

3 RESULT

3.1 Overview

In this study, an LSTM neural network is developed for pan-cancer classification and "symmetrical occlusion" algorithm is proposed to identify the most important genes in classification while improving network performance with fewer genes. Our dataset consisted of 3524 samples, classified into 28 classes, with 22 classes consisting of paired tumor and normal samples from the same tissue of origin. After five-fold cross-validation, the network achieved an accuracy of 96.59% using NRC as a normalization method with the complete set of genes. In comparison, the accuracy of the network trained with the complete gene set of TPM and RPKM was 91.84% and 89.53%, respectively. This demonstrates that NRC is more suitable for gene expression quantification in pan-cancer classification.

As shown in **Fig.3**, by sorting genes according to their mean occlusion scores, we were able to increase the validation accuracy to 98.30% using only the top-ranking 33% genes (**Fig.3A**). Two example confusion matrices are provided in **Additional Figure.2**. In another attempt, we constructed training dataset unique top-ranking genes sorted by their occlusion scores in each individual class, resulting in a validation accuracy between 94% and 96% (**Fig.3B**). It demonstrates that our proposed framework can improve pan-cancer classification accuracy and reduce the number of genes required for accurate classification.

3.2 Performance comparison in pan-cancer classification

In this paper, LSTM achieved the best performance in validation accuracy and prediction classes compared to previously reported works on pan-cancer classification. As shown in **Table.1**, Mostavi et al. achieved accuracy of 92.5% by implementing various CNN networks featuring a vector of 7100 selected genes and a reshaped 100x71 matrix as input [19]. Similarly, Zhao et al. constructed a network featuring 1D inception architecture and achieved accuracy of 92.89%, while de Guia et al. and Khalifa et al. converted the gene expression vector into a matrix for CNN training and obtained accuracy of 95.65% and [20–22]. In addition, compared to these CNN-based methods, our LSTM neural network can distinguish 28 classes of various tissue of origin and status, which is significantly higher than others. For example, Sun et al. implemented a model to distinguish normal samples and tumor samples without tissue of origin (binary classification) and achieved an accuracy of 96.0%, their second model for classifying 11 tumor classes with tissue of origin achieved an accuracy of 98.6% [23]. Similarly, Khalifa et al. achieved an accuracy of 96.9% with five tumor classes and no normal class [21].

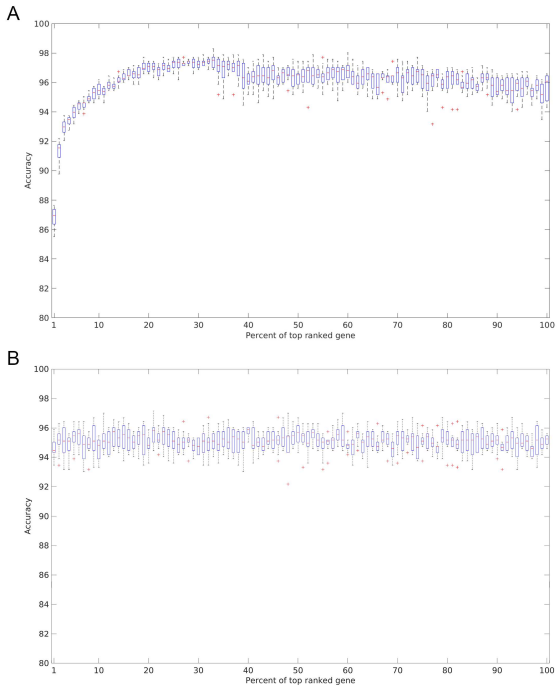


Figure 3: Comparison of validation accuracy of neural network using different gene selection criteria. (A) Box plot of validation accuracy of neural networks trained using top-ranking genes ranked by the mean occlusion score in all classes. (B) Box plot of validation accuracy of neural networks trained using top-ranking genes of each subclass with duplicates removed.

3.3 Performance comparison in metastasized cancer

To compare the classification performance against Sun et.al., metastasized colorectal cancer samples from Kim et.al. were utilized as test dataset for direct comparison [24, 25]. Notably, our accuracy is much higher than the reported work by with the same dataset. Our study achieved an 88.89% prediction accuracy, with 16 out of 18 metastasized colorectal cancer samples correctly classified as “colorectal cancer”. The remaining two samples were classified as "colorectal normal" and "liver normal", respectively.

3.4 Performance comparison in identification marker gene

Selecting appropriate marker genes is crucial in improving pan-cancer classification performance. We compared the performance of several published marker gene selection methods on the same LSTM network, using marker genes selected according to their original method. Additionally, we compared the occlusion score and tissue specificity entropy score and found that occlusion method is superior to entropy method in identifying pan-cancer marker genes.

Table 1: Performance comparison with existing methods

Method	Number of classes	Accuracy	Precision	Recall	F1-score	Comment
LSTM (full gene list)	22+6	94.03	85.67	83.02	84.32	22 paired classes + 6 unpaired
LSTM (Top 33% gene)	22+6	97.59	97.55	94.63	96.07	22 paired classes + 6 unpaired
Zhao et.al.	32	92.89	95.55	95.25	95.40	32 tumor types
Mostavi et.al.	33+1	95.70	93.60	91.90	92.74	33 tumor types + 1 normal
de Guia et.al.	33	95.65	95.55	95.69	95.62	34 tumor types
Sun et.al.	10	96.70	99.30	96.90	98.09	11 tumor types
Khalifa et.al.	5	96.90	94.96	95.09	95.02	5 tumor types

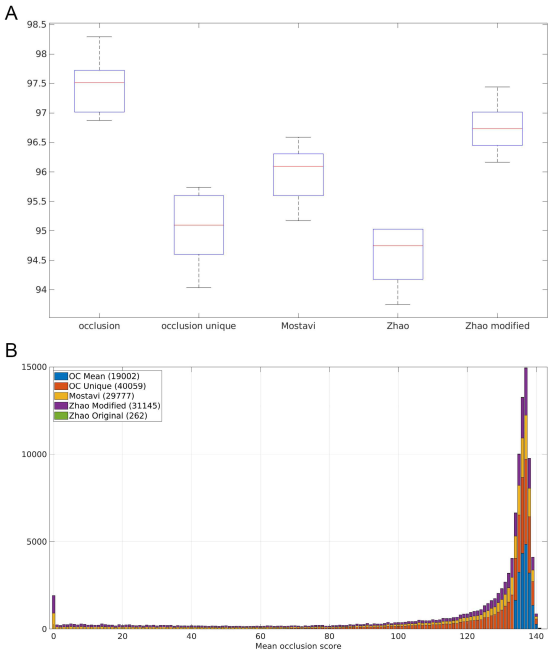


Figure 4: Box plot of neural network validation accuracy using different marker gene selection method. (A) validation accuracy of networks trained using gene list. (B) Stacked histogram of mean occlusion score of genes selected by corresponding method. The number in the parenthesis is the number of the genes.

In the study conducted by Mostavi et al., a fixed threshold was used to select with an FPKM mean or standard deviation above the threshold [19]. However, as this method has two criteria, it is difficult to control the number of selected gene and thus remained unchanged. Consequently, a gene list containing 29777 genes was used for preparing LSTM neural network dataset. In the work done by Zhao et.al's study, the top 40 genes with the highest difference between the median expression of each gene in the in-class sample relative to the out-of-class samples were selected [22]. A total of 1120 genes were selected using this method. To make it comparable to the occlusion method, the scope was widened to increase the number of marker genes selected.

As shown in **Fig.4A**, LSTM network trained with dataset containing genes selected by occlusion method obtained the highest median validation accuracy of 97.51% with fewest number of genes. The LSTM network trained using modified criteria by selecting the top 33% unique genes from individual 28 classes (occlusion unique) obtained a median validation accuracy of 95.10%. In comparison, LSTM network trained with dataset containing genes selected by Mostavi's method obtained median validation accuracy of 96.09%; Original method from Zhao et.al. has the lowest median validation accuracy of 94.74% and increased to 96.73% after widening the criteria. Further analysis suggests that the mean occlusion score of the genes is an important factor in determining the validation accuracy of the LSTM network.

As shown in**Fig.4B**, the stacked histogram demonstrates that the occlusion method selects genes with the highest mean occlusion score and the lowest standard deviation, leading to the highest validation accuracy. The modified occlusion method selects unique genes with high occlusion scores from individual classes and selects the greatest number of genes but with low mean occlusion score, resulting in lower median validation accuracy. Mostavi's method and modified method from Zhao et.al. selected similar number of genes, with Mostavi's method is slightly outperformed due to its fewer genes with high mean occlusion score. In summary, the mean occlusion score of the selected genes has a significant impact on the LSTM neural network's accuracy, genes with high mean occlusion score improves the accuracy of the network while genes with low mean occlusion score have detrimental effect.

Genes expressed only in specific cancer type or tissue, and not in others have very high tissue specificity and may be a marker gene of cancer. In order to quantify this specificity, Schug et.al introduced a tissue specificity score based on Shannon entropy which is commonly seen in information theories [26]. Later on, a superior method known as "ROKU entropy" was published which improved in sequencing genes according to tissue specificity across tissues [27]. To check if the occlusion score in this study is correlated with the entropies. Tissue specificity scores of Shannon entropy and ROKU entropy are obtained using software "TSPEX" [28]. The maximum entropy value of Shannon entropy and ROKU entropy output by TSPEX is $log_2 N \approx 8.43$ (N=28), indicating complete tissue specific. The gene expression matrix for all 28 classes is constructed using the median of FPKM for all samples, as required by the software.

As shown in **Fig.5**, there is only a weak positive correlation between Shannon entropy and occlusion score and a weak positive

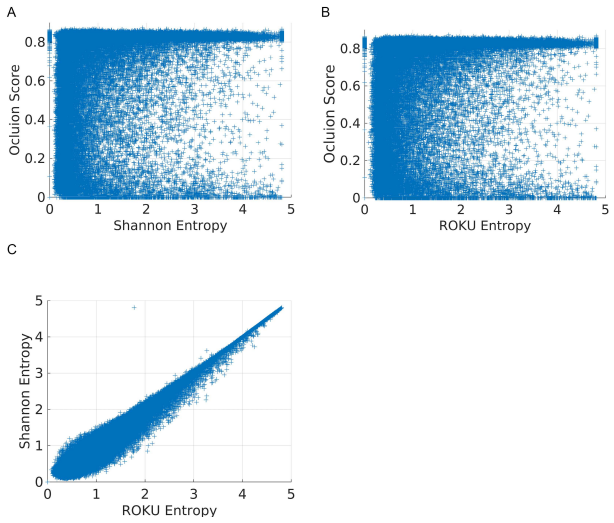


Figure 5: Scatter plots of Shannon tissue specificity entropy, ROKU entropy and occlusion score. (A) Scatter plot between Shannon entropy and occlusion score, the correlation coefficient is 0.1893. (B) Scatter plot between ROKU entropy and Occlusion score, the correlation coefficient is 0.1630. (C) Scatter plot between ROKU entropy and Shannon entropy, the correlation coefficient is 0.9859.

correlation between ROKU entropy and occlusion score. Meanwhile, there is a strong positive correlation between Shannon entropy and ROKU entropy. This demonstrates that despite lower tissue specificity some genes are still considered important by the occlusion method. Most of all, this proves that occlusion method is looking beyond the tissue specificity of each gene when evaluating its importance in pan-cancer classification.

To further investigate the difference between tissue specificity score and mean occlusion score, understanding why some genes are not ideal markers for pan-cancer classification despite their high tissue specificity. Investigations were first conducted on genes with a high tissue specificity score but low mean occlusion score. As shown in **Table.2**, gene GFAP is a classical marker of astrocytoma which expressed only in brain tissue [29, 30]. Thus, GFAP has one of the highest tissue specificity scores. However, this extremely high tissue specificity also prevents GFAP from being an ideal pan-cancer classification marker which leads to very low mean occlusion scores. The same phenomenon was also observed on genes such as KRT4, LIPF, TG, PGC which also almost exclusively expressed in specific tissue, making them less suitable for pan-cancer classification [29, 31–33].

Meanwhile, investigations were also conducted on genes with high tissue specificity and high mean occlusion score. As shown in **Table.3**, microRNA MIR663B has both a high mean occlusion score as well as high tissue specificity score. Closer examination demonstrates that despite MIR663B is expressed in several tissues but its expression level in specific tissue such as brain is significantly higher than the rest [34–36].

In comparison, **Table.4** demonstrates genes which have a high mean occlusion score but low tissue specificity score. This was due to their broad expression in many tissues but lower standard deviation of expression level. Despite their low tissue specificity, some research have reported their correlation with cancer. For example, lncRNA ADPGK-AS1 is related to osteosarcoma, colorectal cancer, pancreatic cancer and breast cancer via different pathway [37–40]. LncRNA DNM1P35 is reported to be a novel prognostic factor for kidney cancer [41]. STK24-AS1 could be used to predict patient survival rates of colon cancer patients [42].

This result demonstrates that occlusion method is superior to tissue specificity method in identifying ideal pan-cancer marker genes. The tissue specificity method, like other statistical methods, may overlook genes with low expression level differences. Please refer to **Additional File 3** for the complete table.

3.5 Classification performance in Single Cell RNA-Seq

The performance of the proposed deep-learning method applying on single cell type classification using "single cell RNA Sequencing" (scRNA-Seq) data was evaluated from two perspectives. First, we aimed to verify whether the LSTM network could accurately classify cell types. Second, we seek to further validate the proposed occlusion method on maker gene identification, which has been widely used for cell annotation.

For the first perspective, a separate LSTM neural network was trained on human muscle acquired from public sources using the same workflow. The network obtained validation accuracy of 96.0%, it indicates that the network can accurately identify cell type using scRNA-Seq data. For the second perspective, 100 randomly selected "Muscle Stem Cells" (MuSC) cells were extracted from validation dataset. Then, the expression level of gene PAX7 which is one of the marker genes of MuSC cells and only expressed in muSC was reduced.

As shown in **Fig.6A**, when the expression level of PAX7 remains unchanged, majority of the MuSC cells are correctly classified. As the expression level of PAX7 decreases during the occlusion, from top to the bottom, many of these cells are begin to be mis-classified as non-muSC, indicating that PAX7 is essential for the identification of muSC. Meanwhile, as shown in**Fig.6B**, reducing the expression level of gene Dido1 which is randomly selected, made no impact on the prediction result of another 100 randomly selected muSC cells.

To further illustrate that PAX7 is a marker gene for MuSC cells. The expression level of PAX7 was increased on 100 randomly selected non-muSC cells. As shown in **Fig.7A**, from bottom to the top, most of these non-MuSC cells would be classified as "muSC" as the expression level of PAX7 is increased. In comparison, increasing the expression level of another random selected gene Cryz, had no effect on the classification result, indicating that this gene has much less influence on the prediction result.

This result highlights the potential of the LSTM neural network for cell type classification using scRNA-Seq data. Moreover, by conducting occlusion on both marker genes of muSC and non-muSC cells, we had demonstrated the effectiveness of occlusion method in

Table 2: Example genes with high tissue specificity score but low mean occlusion score

Gene ID	Gene Name	Mean Occlusion Score	Shannon Entropy	Roku Entropy	Tissue Specificity
ENSG00000170477.13	KRT4	0.0016	4.65	4.65	enriched in stomach
ENSG00000096088.16	PGC	0.0011	3.51	3.51	enriched in stomach
ENSG00000131095.13	GFAP	0.0010	4.78	4.79	enriched in brain

Table 3: Example genes with high tissue specificity score and high mean occlusion score

Gene ID	Gene Name	Mean Occlusion Score	Shannon Entropy	Roku Entropy
ENSG00000221288.1	MIR663B	0.868	4.807	4.807
ENSG00000199204.1	Y_RNA	0.867	4.807	4.807
ENSG00000230142.2	LINC01075	0.860	4.807	4.807

Table 4: Example genes with high mean occlusion score but low tissue specificity score

Gene ID	Gene Name	Mean Occlusion Score	Shannon Entropy	Roku Entropy
ENSG00000224418.1	STK24-AS1	0.8595	0.2365	0.4062
ENSG00000246877.1	DNM1P35	0.8560	0.3091	0.5346
ENSG00000260898.6	ADPGK-AS1	0.8535	0.2941	0.4468

capturing and identifying marker genes. Overall, these findings contribute to the development of more accurate and efficient methods for single cell classification and annotation.

3.6 GO term analysis on identified marker genes

The top 33% of genes, ranked by occlusion score, were subjected to "Gene Ontology" (GO) term analysis using DAVID to confirm their relevance to cancer [43]. As shown in **Fig.8**, the most significant molecular function identified was "olfactory receptor activity", which has been linked to perception of smell and cancer by prior publications. For instance, Shibel et.al. reported that olfactory receptor OR5H2 regulates the proliferation of endometrial cancer cells through the IGF1 signaling pathway [44]. Furthermore, PSGR, a prostate-specific G protein-coupled receptor, has been found to be upregulated in prostate cancer. Another study by Webber et.al. reported that olfactory receptor OR10H1 is primarily expressed in the human bladder cancer [45]. Pathway analysis also revealed olfactory transduction to be the most significant pathway, followed by "micro RNAs in cancer" and JAK-STAT signaling. The JAK-STAT pathway is known to promote tumour genesis, and its inhibition can impede cancer cell growth [46]. Overall, these findings suggest that the selected genes may be implicated in cancer-related processes. The complete result can be found in **Additional File 4**.

3.7 Analysis of pseudogenes among the top-ranking genes

Pseudogenes were previously believed to be non-functional copies of their parental genes as the result of retrotransposition of mRNA or genomic duplication and carry no biological functions. But recent studies have revealed that they play various roles in diverse physiological and pathological processes, especially in cancer [47]. It has been found that pseudogenes are differentially expressed among established tumor subtypes [48, 49]. Han et.al. had reported that pseudogenes are significantly related to patient survival in kidney cancer [50].

Among the top 33% genes ranked by mean occlusion score, 6108 genes are pseudogenes, to examine if there's any correlation between our result and theirs, 2616 pseudogenes been reported by Han et.al. to be differentially expressed in 4 types of tumors have been selected to check their corresponding ranking in our result. Firstly, corresponding ensemble gene IDs of the 2616 pseudogenes were extracted through the REST API from ensemble.org. Secondly, the gene ID were double checked to ensure its consistency in the genomic location given by Han et.al.. In total, 974 out of 2616 pseudogenes have valid gene ID, moreover, 34.85% of identified pseudogenes from "Glioblastoma" (GBM), 27.76% of identified pseudogenes from "Breast Cancer" (BRCA), 27.67% of identified pseudogenes from "Lung Squamous Cell Carcinoma" (LUSC) and 10.47% of identified pseudogenes from "Uterine Corpus Endometrial Carcinoma" (UCEC) are among the top 33% gene selected by occlusion method. Therefore, we hypothesis that pseudogenes may be a major contributor in pan-cancer classification.

To further investigate the contribution of pseudogenes in pan-cancer classification, a depletion study was conducted. New LSTM neural networks were trained using updated training datasets with all pseudogenes removed from the top 33% gene. However, the networks trained using datasets without pseudogenes had an average accuracy 0.5% lower than the networks trained using datasets with the same number of genes that contain 17% of the top-ranking genes. This suggests that pseudogenes may have a limited impact on the accuracy of pan-cancer classification but should still be included to achieve maximum performance.



Figure 6: Prediction result comparison after simulating "loss of function" of marker gene and randomly selected gene. (A) Prediction result of 100 randomly selected MuSC cells, after reducing expression level of MuSC marker gene PAX7 below the baseline, many MuSC cells are classified as other cell types. (B) Prediction result of another 100 randomly selected MuSC cells, the expression level of a randomly selected gene Dido1 is reduced; No MuSC cell is misclassified.

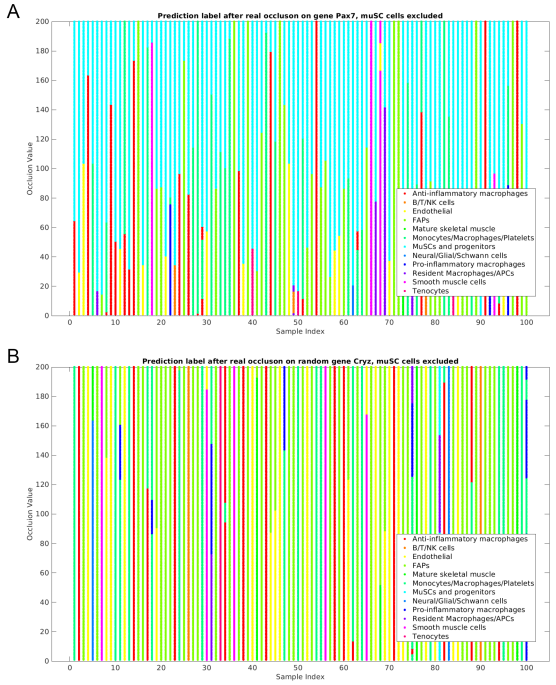


Figure 7: Prediction result comparison after simulating "gain of function" of marker gene and randomly selected gene. (A) Prediction label of 100 randomly selected non-MuSC cells, increasing the expression level of PAX7 above the baseline, many non-MuSC cells being classified as MuSC cells. (B) Prediction label of 100 randomly selected non-MuSC cells,increasing the expression level of gene Cryz above the baseline, little non-MuSC cells being classified as MuSC cells.

4 DISCUSSION

This paper demonstrates the effectiveness of the NRC normalization method and the LSTM neural network in pan-cancer classification and marker gene prediction using both RNASeq and scRNA-Seq data. Additionally, the proposed occlusion algorithm has been shown to effectively identify marker genes and improve classification accuracy of the neural network with fewer genes.

Furthermore, the occlusion algorithm has the potential to uncover gene-gene interactions by testing combinations of candidate genes, although this approach may be computationally demanding due to the large number of possible combinations. Overall, these findings demonstrate the versatility and potential of the occlusion algorithm for various applications in genomics research.

In addition to cell type classification using scRNA-Seq data, the LSTM neural network has the potential to detect novel cell types that are not included in the training dataset. This is supported by the fact that the network would misclassify novel cell types but with a low prediction score. For example, a trained LSTM neural network would misclassify all MuSC cells as other cell types when the MuSC cells are excluded from training dataset but fed into network for prediction. There is a significant difference between the prediction

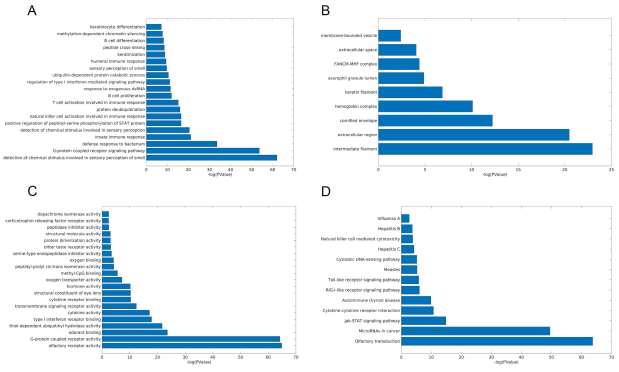


Figure 8: GO term analysis result of top 33 percent genes with highest mean occlusion score carried out by DAVID. (A) Biological Process (BP). (B) Cellular Component (CC). (C) Molecular Function (MF). (D) Pathway

results of misclassification and the true positive results, which are 0.4~0.6 and around 1, respectively. This difference is leveraged to detect novel cell types and the detection accuracy can reach 85~88% when applying machine learning classifiers such as linear discriminator and SVM.

5 DECLARATIONS

5.1 Ethics approval and consent to participate

Not Applicable

5.2 Consent for publication

Not Applicable

5.3 Availability of data and materials

All data generated or analysed during this study are publicly available using their accession numbers in the supplementary files. The code is available from corresponding author on reasonable request.

5.4 Competing interests

The authors declares no competing interests

5.5 Funding

This work was supported by General Research Funds (GRF) from the Research Grants Council (RGC), University Grants Committee of the Hong Kong Special Administrative Region, China (Project Codes: 2141109, 2141157, 2141261 to H.S.)

5.6 Authors' contributions

X.Z. conducted the design and programming of the software as well as drafting of the manuscript. H.S. and H.T. supervised the study. All authors reviewed the manuscript.

5.7 Acknowledgements

The authors thank the reviewers for their valuable suggestions and Yuanyuan Wei from Department of Biomedical Engineering, The Chinese University of Hong Kong for her help in polishing the manuscript.

REFERENCES

[1] Changfa Xia, Xuesi Dong, He Li, Maomao Cao, Dianqin Sun, Siyi He, Fan Yang, Xinxin Yan, Shaoli Zhang, and Ni Li. Cancer statistics in china and united states, 2022: profiles, trends, and determinants. *Chinese medical journal*, 135(05):584–590, 2022.

[2] N. Tanaka, D. Kaczynska, S. Kanatani, C. Sahlgren, P. Mitura, A. Stepulak, A. Miyakawa, P. Wiklund, and P. Uhlen. Mapping of the three-dimensional lymphatic microvasculature in bladder tumours using light-sheet microscopy. *Br J Cancer*, 118(7):995–999, 2018.

[3] J. J. Chen, S. J. Wang, C. A. Tsai, and C. J. Lin. Selection of differentially expressed genes in microarray data analysis. *Pharmacogenomics J*, 7(3):212–20, 2007.

[4] R. Fisher, L. Pusztai, and C. Swanton. Cancer heterogeneity: implications for targeted therapeutics. *British Journal of Cancer*, 108(3):479–485, 2013.

[5] Bert Vogelstein, Nickolas Papadopoulos, Victor E Velculescu, Shibin Zhou, Luis A Diaz Jr, and Kenneth W Kinzler. Cancer genome landscapes. *science*, 339(6127):1546–1558, 2013.

[6] Iacopo Petrini, Paul S. Meltzer, In-Kyu Kim, Marco Lucchi, Kang-Seo Park, Gabriella Fontanini, James Gao, Paolo A. Zucali, Fiorella Calabrese, Adolfo Favaretto, Federico Rea, Jaime Rodriguez-Canales, Robert L. Walker, Marbin Pineda, Yuelin J. Zhu, Christopher Lau, Keith J. Killian, Sven Bilke, Donna Voeller, Sivanesan Dakshanamurthy, Yisong Wang, and Giuseppe Giaccone. A specific missense mutation in gtf2i occurs at high frequency in thymic epithelial tumors. *Nature Genetics*, 46(8):844–849, 2014.

[7] A. V. Snezhkina, E. N. Lukyanova, A. R. Zaretsky, D. V. Kalinin, A. V. Pokrovsky, A. L. Golovyuk, G. S. Krasnov, M. S. Fedorova, E. A. Pudova, S. L. Kharitonov, N. V. Melnikova, B. Y. Alekseev, M. V. Kiseleva, A. D. Kaprin, A. A. Dmitriev, and A. V. Kudryavtseva. Novel potential causative genes in carotid paragangliomas. *BMC Med Genet*, 20(Suppl 1):48, 2019.

[8] Tom Lesluyes, Jessica Baud, Gaëlle Pérot, Céline Charon-Barra, Axel You, Isabelle Valo, Céline Bazille, Florence Mishellany, Agnès Leroux, Sophie Renard-Oldrini, Philippe Terrier, Axel Le Cesne, Marick Laé, Sophie Piperno-Neumann, Sylvie Bonvalot, Agnès Neuville, Françoise Collin, Philippe Maingon, Jean-Michel Coindre, and Frédéric Chibon. Genomic and transcriptomic comparison of post-radiation versus sporadic sarcomas. *Modern pathology : an official journal of the United States and Canadian Academy of Pathology, Inc*, 32(12):1786–1794, 2019.

[9] Ben Langmead and Steven L Salzberg. Fast gapped-read alignment with bowtie 2. *Nature methods*, 9(4):357–359, 2012.

[10] Yang Liao, Gordon K Smyth, and Wei Shi. featurecounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*, 30(7):923–930, 2014.

[11] Ana Conesa, Pedro Madrigal, Sonia Tarazona, David Gomez-Cabrero, Alejandra Cervera, Andrew McPherson, Michał Wojciech Szcześniak, Daniel J Gaffney, Laura L Elo, and Xuegong Zhang. A survey of best practices for rna-seq data analysis. *Genome biology*, 17(1):1–19, 2016.

[12] Marie-Agnès Dillies, Andrea Rau, Julie Aubert, Christelle Hennequet-Antier, Marine Jeanmougin, Nicolas Servant, Céline Keime, Guillemette Marot, David Castel, and Jordi Estelle. A comprehensive evaluation of normalization methods for illumina high-throughput rna sequencing data analysis. *Briefings in bioinformatics*, 14(6):671–683, 2013.

[13] Shanrong Zhao, Zhan Ye, and Robert Stanton. Misuse of rpkm or tpm normalization when comparing across samples and sequencing protocols. *Rna*, 26(8):903–909, 2020.

[14] Yingdong Zhao, Ming-Chung Li, Mariam M. Konaté, Li Chen, Biswajit Das, Chris Karlovich, P. Mickey Williams, Yvonne A. Evrard, James H. Doroshow, and Lisa M. McShane. Tpm, fpkm, or normalized counts? a comparative study of quantification measures for the analysis of rna-seq data from the nci patient-derived models repository. *Journal of Translational Medicine*, 19(1):269, 2021.

[15] A. Mohamed A. Graves and G. Hinton. Speech recognition with deep recurrent neural networks. *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 6645–6649, 2013.

[16] Pierre Baldi and Peter J Sadowski. Understanding dropout. *Advances in neural information processing systems*, 26:2814–2822, 2013.

[17] Houjing Huang, Dangwei Li, Zhang Zhang, Xiaotang Chen, and Kaiqi Huang. Adversarially occluded samples for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5098–5107, 2018.

[18] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13*, pages 818–833. Springer, 2014.

[19] M. Mostavi, Y. C. Chiu, Y. Huang, and Y. Chen. Convolutional neural network models for cancer type prediction based on gene expression. *BMC Med Genomics*, 13(Suppl 5):44, 2020.

[20] 2019.

[21] Nour Eldeen M. Khalifa, Mohamed Hamed N. Taha, Dalia Ezzat Ali, Adam Slowik, and Aboul Ella Hassanien. Artificial intelligence technique for gene expression by tumor rna-seq data: A novel optimized deep learning approach. *IEEE Access*, 8:22874–22883, 2020.

[22] Y. Zhao, Z. Pan, S. Namburi, A. Pattison, A. Posner, S. Balachander, C. A. Paisie, H. V. Reddi, J. Rueter, A. J. Gill, S. Fox, K. P. S. Raghav, W. F. Flynn, R. W. Tothill, S. Li, R. K. M. Karuturi, and J. George. Cup-ai-dx: A tool for inferring cancer tissue of origin and molecular subtype using rna gene-expression data and artificial intelligence. *EBioMedicine*, 61:103030, 2020.

[23] K. Sun, J. Wang, H. Wang, and H. Sun. Genect: a generalizable cancerous status and tissue origin classifier for pan-cancer biopsies. *Bioinformatics*, 34(23):4129–4130, 2018.

[24] F. Fan, D. Chen, Y. Zhao, H. Wang, H. Sun, and K. Sun. Rapid preliminary purity evaluation of tumor biopsies using deep learning approach. *Comput Struct Biotechnol J*, 18:1746–1753, 2020.

[25] S. K. Kim, S. Y. Kim, J. H. Kim, S. A. Roh, D. H. Cho, Y. S. Kim, and J. C. Kim. A nineteen gene-based risk score classifier predicts prognosis of colorectal cancer patients. *Mol Oncol*, 8(8):1653–66, 2014.

[26] J. Schug, W. P. Schuller, C. Kappen, J. M. Salbaum, M. Bucan, and Jr. Stoeckert, C. J. Promoter features related to tissue specificity as measured by shannon entropy. *Genome Biol*, 6(4):R33, 2005.

[27] K. Kadota, J. Ye, Y. Nakai, T. Terada, and K. Shimizu. Roku: a novel method for identification of tissue-specific genes. *BMC Bioinformatics*, 7:294, 2006.

[28] Antonio P. Camargo, Adrielle A. Vasconcelos, Mateus B. Fiamenghi, Gonçalo A. G. Pereira, and Marcelo F. Carazzolle. tspex: a tissue-specificity calculator for gene expression data. 2020.

[29] L. Fagerberg, B. M. Hallström, P. Oksvold, C. Kampf, D. Djureinovic, J. Odeberg, M. Habuka, S. Tahmasebpoor, A. Danielsson, K. Edlund, A. Asplund, E. Sjöstedt, E. Lundberg, C. A. Szigyarto, M. Skogs, J. O. Takanen, H. Berling, H. Tegel, J. Mulder, P. Nilsson, J. M. Schwenk, C. Lindskog, F. Danielsson, A. Mardinoglu, A. Sivertsson, K. von Feilitzen, M. Forsberg, M. Zwahlen, I. Olsson, S. Navani, M. Huss, J. Nielsen, F. Ponten, and M. Uhlén. Analysis of the human tissue-specific expression by genome-wide integration of transcriptomics and antibody-based proteomics. *Mol Cell Proteomics*, 13(2):397–406, 2014.

[30] E. J. van Bodegraven, J. V. van Asperen, P. A. J. Robe, and E. M. Hol. Importance of gfap isoform-specific analyses in astrocytoma. *Glia*, 67(8):1417–1433, 2019.

[31] M. O. Duff, S. Olson, X. Wei, S. C. Garrett, A. Osman, M. Bolisetty, A. Plocik, S. E. Celniker, and B. R. Graveley. Genome-wide identification of zero nucleotide recursive splicing in drosophila. *Nature*, 521(7552):376–9, 2015.

[32] John Lonsdale, Jeffrey Thomas, Mike Salvatore, Rebecca Phillips, Edmund Lo, Saboor Shad, Richard Hasz, Gary Walters, Fernando Garcia, Nancy Young, Barbara Foster, Mike Moser, Ellen Karasik, Bryan Gillard, Kimberley Ramsey, Susan Sullivan, Jason Bridge, Harold Magazine, John Syron, Johnelle Fleming, Laura Siminoff, Heather Traino, Maghboeba Mosavel, Laura Barker, Scott Jewell, Dan Rohrer, Dan Maxim, Dana Filkins, Philip Harbach, Eddie Cortadillo, Bree Berghuis, Lisa Turner, Eric Hudson, Kristin Feenstra, Leslie Sobin, James Robb, Phillip Branton, Greg Korzeniewski, Charles Shive, David Tabor, Liqun Qi, Kevin Groch, Sreenath Nampally, Steve Buia, Angela Zimmerman, Anna Smith, Robin Burges, Karna Robinson, Kim Valentino, Deborah Bradbury, Mark Cosentino, Norma Diaz-Mayoral, Mary Kennedy, Theresa Engel, Penelope Williams, Kenyon Erickson, Kristin Ardlie, Wendy Winckler, Gad Getz, David DeLuca, Daniel MacArthur, Manolis Kellis, Alexander Thomson, Taylor Young, Ellen Gelfand, Molly Donovan, Yan Meng, George Grant, Deborah Mash, Yvonne Marcus, Margaret Basile, Jun Liu, Jun Zhu, Zhidong Tu, Nancy J. Cox, Dan L. Nicolae, Eric R. Gamazon, Hae Kyung Im, Anuar Konkashbaev, Jonathan Pritchard, Matthew Stevens, Timothèe Flutre, Xiaoquan Wen, Emmanouil T. Dermitzakis, Tuuli Lappalainen, Roderic Guigo, Jean Monlong, Michael Sammeth, Daphne Koller, Alexis Battle, Sara Mostafavi, Mark McCarthy, Manual Rivas, Julian Maller, Ivan Rusyn, Andrew Nobel, Fred Wright, Andrey Shabalin, Mike Feolo, Nataliya Sharopova, et al. The genotype-tissue expression (gtex) project. *Nature Genetics*, 45(6):580–585, 2013.

[33] Fredrik Pontén, Karin Jirström, and Matthias Uhlen. The human protein atlas—a tool for pathology. *The Journal of Pathology: A Journal of the Pathological Society of Great Britain and Ireland*, 216(4):387–393, 2008.

[34] Huihua Cai, Yong An, Xuemin Chen, Donglin Sun, Tongbing Chen, Yan Peng, Feng Zhu, Yong Jiang, and Xiaozhou He. Epigenetic inhibition of mir-663b by long non-coding rna hotair promotes pancreatic cancer cell proliferation via up-regulation of insulin-like growth factor 2. *Oncotarget*, 7(52):86857, 2016.

[35] Mulong Du, Danni Shi, Lin Yuan, Pengchao Li, Haiyan Chu, Chao Qin, Changjun Yin, Zhengdong Zhang, and Meilin Wang. Circulating mir-497 and mir-663b in plasma are potential novel biomarkers for bladder cancer. *Scientific Reports*, 5(1):10437, 2015.

[36] S. Hong, Z. Yan, H. Wang, L. Ding, Y. Song, and M. Bi. mir-663b promotes colorectal cancer progression by activating ras/raf signaling through downregulation of tnk1. *Hum Cell*, 33(1):104–115, 2020.

[37] XF Luo, XJ Wu, X Wei, AG Wang, SH Wang, and JL Wang. Lncrna adpgk-as1 regulated cell proliferation, invasion, migration and apoptosis via targeting mir-542-3p in osteosarcoma. *Eur Rev Med Pharmacol Sci*, 23(20):8751–8760, 2019.

[38] HY Jiang and ZJ Wang. Adpgk-as1 promotes the progression of colorectal cancer via sponging mir-525 to upregulate fut1. *Eur Rev Med Pharmacol Sci*, 24(5):2380–2386, 2020.

[39] Suzhen Song, Weihua Yu, Sen Lin, Mingbao Zhang, Teng Wang, Shuang Guo, and Hongbo Wang. Lncrna adpgk-as1 promotes pancreatic cancer progression through activating zeb1-mediated epithelial–mesenchymal transition. *Cancer biology & therapy*, 19(7):573–583, 2018.

[40] Jiahui Yang, Weizhu Wu, Minhua Wu, and Jinhua Ding. Long noncoding rna adpgk-as1 promotes cell proliferation, migration, and emt process through regulating mir-3196/otx1 axis in breast cancer. *In Vitro Cellular & Developmental Biology-Animal*, 55(7):522–532, 2019.

[41] Jukun Song, Juxiang Peng, Chen Zhu, Guohui Bai, Yongda Liu, Jianguo Zhu, and Jianguo Liu. Identification and validation of two novel prognostic lncrnas in kidney renal clear cell carcinoma. *Cellular Physiology and Biochemistry*, 48(6):2549–2562, 2018.

[42] Lijie Yang, Tiecheng Yang, Huaqiao Wang, Tingting Dou, Xiaochang Fang, Liwen Shi, Xuanfei Li, and Maohui Feng. Dnmbp-as1 regulates nhlrc3 expression by sponging mir-93-5p/17-5p to inhibit colon cancer progression. *Frontiers in Oncology*, 12:765163, 2022.

[43] Da Wei Huang, Brad T Sherman, Qina Tan, Joseph Kir, David Liu, David Bryant, Yongjian Guo, Robert Stephens, Michael W Baseler, H Clifford Lane, et al. David bioinformatics resources: expanded annotation database and novel algorithms to better extract biology from large gene lists. *Nucleic acids research*, 35(suppl_2):W169–W175, 2007.

[44] R. Shibel, R. Sarfstein, K. Nagaraj, L. Lapkina-Gendler, Z. Laron, M. Dixit, S. Yakar, and H. Werner. The olfactory receptor gene product, or5h2, modulates endometrial cancer cells proliferation via interaction with the igf1 signaling pathway. *Cells*, 10(6), 2021.

[45] L. Weber, W. A. Schulz, S. Philippou, J. Eckardt, B. Ubrig, M. J. Hoffmann, A. Tannapfel, B. Kalbe, G. Gisselmann, and H. Hatt. Characterization of the olfactory receptor or10h1 in human urinary bladder cancer. *Front Physiol*, 9:456, 2018.

[46] S. Bose, S. Banerjee, A. Mondal, U. Chakraborty, J. Pumarol, C. R. Croley, and A. Bishayee. Targeting the jak/stat signaling pathway using phytocompounds for cancer prevention and therapy. *Cells*, 9(6), 2020.

[47] L. Xiao-Jie, G. Ai-Mei, J. Li-Juan, and X. Jiang. Pseudogene in cancer: real functions and promising signature. *J Med Genet*, 52(1):17–24, 2015.

[48] Y. Pan, C. Sun, M. Huang, Y. Liu, F. Qi, L. Liu, J. Wen, J. Liu, K. Xie, H. Ma, Z. Hu, and H. Shen. A genetic variant in pseudogene e2f3p1 contributes to prognosis of hepatocellular carcinoma. *J Biomed Res*, 28(3):194–200, 2014.

[49] Y. H. Loh, Q. Wu, J. L. Chew, V. B. Vega, W. Zhang, X. Chen, G. Bourque, J. George, B. Leong, J. Liu, K. Y. Wong, K. W. Sung, C. W. Lee, X. D. Zhao, K. P. Chiu, L. Lipovich, V. A. Kuznetsov, P. Robson, L. W. Stanton, C. L. Wei, Y. Ruan, B. Lim, and H. H. Ng. The oct4 and nanog transcription network regulates pluripotency in mouse embryonic stem cells. *Nat Genet*, 38(4):431–40, 2006.

[50] L. Han, Y. Yuan, S. Zheng, Y. Yang, J. Li, M. E. Edgerton, L. Diao, Y. Xu, R. G. W. Verhaak, and H. Liang. The pan-cancer analysis of pseudogene expression reveals biologically and clinically relevant tumour subtypes. *Nat Commun*, 5:3963, 2014.

Received 20 June 2023; revised 01 August 2023; accepted 20 July 2009