OXFORD

Briefings in Bioinformatics, 2022, **23(5)**, 1–13

https://doi.org/10.1093/bib/bbac294

Problem Solving Protocol

Contrastive learning-based computational histopathology predict differential expression of cancer driver genes

Haojie Huang, Gongming Zhou, Xuejun Liu, Lei Deng, Chen Wu, Dachuan Zhang and Hui Liu

Correspondence authors: Hui Liu, School of Computer Science and Technology, Nanjing Tech University, 211816, Nanjing, China. E-mail: hliu@njtech.edu.cn; Dachuan Zhang, The third affiliated hospital of Soochow University, 213100, Changzhou, China. E-mail: zhangdachuan@suda.edu.cn

Abstract

Motivation: Digital pathological analysis is run as the main examination used for cancer diagnosis. Recently, deep learning-driven feature extraction from pathology images is able to detect genetic variations and tumor environment, but few studies focus on differential gene expression in tumor cells.

Results: In this paper, we propose a self-supervised contrastive learning framework, HistCode, to infer differential gene expression from whole slide images (WSIs). We leveraged contrastive learning on large-scale unannotated WSIs to derive slide-level histopathological features in latent space, and then transfer it to tumor diagnosis and prediction of differentially expressed cancer driver genes. Our experiments showed that our method outperformed other state-of-the-art models in tumor diagnosis tasks, and also effectively predicted differential gene expression. Interestingly, we found the genes with higher fold change can be more precisely predicted. To intuitively illustrate the ability to extract informative features from pathological images, we spatially visualized the WSIs colored by the attention scores of image tiles. We found that the tumor and necrosis areas were highly consistent with the annotations of experienced pathologists. Moreover, the spatial heatmap generated by lymphocyte-specific gene expression patterns was also consistent with the manually labeled WSIs.

Keywords: Whole slide image, Contrastive learning, Differential gene expression, Cancer driver gene, Computational histopathology, Convolutional neural network

Introduction

With the advancement of scanner and imaging technology, pathology glass slides are increasingly digitized and computational histopathology analysis has emerged as a new standard of diagnostic workflow. Digital histopathology promotes the efficiency and accuracy of pathologists in disease diagnosis and clinical grading.

By virtue of the power of automatic feature extraction, deep learning is increasingly used to extract features from pathology images for multiple subsequent applications. Some deep learning-based models used pathological images to predict the prognosis of cancer patients, such as colorectal cancer [1, 2], hepatocellular carcinoma [3], pancreatic ductal adenocarcinoma [4] and even pan-Cancer [5]. Some studies focused on classification of tumor subtypes and treatment efficacy by deep computational pathology analysis [6–9]. For example, Li *et al.* [10] proposed a classification model without manual annotations by fusing multi-scale features extract from unlabeled patches tiled from whole slide images (WSIs). Saltz *et al.* [11] studied a variety of cancer types and proposed a computational staining method based on deep learning, which can deconvolve the spatial structure of tumor infiltrating lymphocytes. Gheisari *et al.* [12] combined convolution deep belief network and features encoding from WSIs to classify neuroblastoma.

The aforementioned studies promoted the application of digital pathology to routine clinical diagnosis and prognosis. However, these studies were limited to the diagnostic and prognostic value; the potential of digital pathological images has not been fully mined. The development of high-throughput sequencing technology

Haojie Huang is an undergraduate student at School of Computer Science and Engineering, Central South University, Changsha, China.
Gongming Zhou is an undergraduate student at School of Computer Science and Engineering, Central South University, Changsha, China.
Xuejun Liu is a professor at School of Computer Science and Technology, Nanjing Tech University, Nanjing, China. His research interests include data mining and deep learning.
Lei Deng is a professor at School of Computer Science and Engineering, Central South University, Changsha, China. His research interests include data mining, bioinformatics and systems biology.
Chen Wu is an experienced oncologist at The third affiliated hospital of Soochow University, 213100, Changzhou, China. His research interest includes cancer immune-therapy.
Dachuan Zhang is an experienced pathologist at The third affiliated hospital of Soochow University, 213100, Changzhou, China. His research interest includes Tumor immune microenvironment.
Hui Liu is a professor at School of Computer Science and Technology, Nanjing Tech University, Nanjing, China. His research interests include Bioinformatics and Deep Learning.

Received: April 27, 2022. **Revised:** June 16, 2022. **Accepted:** June 29, 2022

generated large-scale multi-omics datasets. Based on the assumption that specific micromolecular patterns will be reflected in histological morphology and cell phenotype, some studies begun the exploration of pathological image features to infer molecular patterns and gene-level information. For example, Kather *et al.* [13] used deep neural network to predict a wide range of gene mutations from histology images, which verified and quantified the relationship between genotype and phenotype in cancer cells. Yu *et al.* [14] found that, in most cancers, there is a general association between histological morphology features and genetic mutations. Chen *et al.* [15] combined the histopathology features and molecular patterns to improve the diagnosis and prognosis of cancer patients. Coudray *et al.* [16] further predicted gene mutations based on the histopathological features derived for the classification task of non-small cell lung cancer subtypes by using deep convolution network. In addition, a lot of studies have demonstrated the applicability of computational pathology to predict microsatellite instability [17–20] and detect mitosis [21–24]. Ash *et al.* [25] proposed ImageCCA that adopted autoencoder to extract features of pathological images, and then applied sparse canonical correlation analysis to reveal associations between cell infrastructure and gene sets. These pioneer studies greatly extended the application of pathological images.

To further explore the relationship between gene expression profiles and histopathological features, we propose a new framework, HistCode, to exploit the histopathology images by self-supervised contrastive learning to infer the differential gene expression in tumor cells. More precisely, we applied adversarial contrastive learning to extract tile-level features, and then aggregated the features to produce the slide-level latent representation by gated-attention pooling. The slide-level representations were transferred to multiple downstream tasks, including tumor diagnosis and differential gene expression prediction. For systematically evaluating our proposed method, we conducted sufficient experiments to demonstrate the expressiveness of computational histopathological features, including tumor diagnosis, quantitative analysis of differential gene expression and spatial localization of lymphocytes. The experimental results fully verify that our model is superior to state-of-the-art models in tumor diagnosis task. We visualized the WSIs by the attention scores of tiles, and found that the tumor and necrosis area are highly consistent with the annotation of pathological experts, which strongly solidified the capacity of HistCode to extract the informative features from pathological images. Next, we verified the ability of HistCode to predict differential gene expression. Interestingly, we observed that the genes with higher fold change can be more precisely predicted. The activation map generated by lymphocyte marker gene expression level was also consistent with the labeled slide by an experienced pathologist. To the best of our knowledge, we are the first to apply computational pathology for differential gene expression analysis. We believe our work would yield inspiring insights into digital pathological analysis and greatly extend its clinical application.

Materials and methods

Whole slide image

All hematoxylin and eosin-stained digital slides (frozen tissue) were downloaded from TCGA via the Genomic Data Commons Data Portal. We collected two solid tumors, breast and lung cancer, in our study, because there are enough samples for us to establish solid model, as well as for systematic evaluation. From TCGA-BRCA project, we gathered 1979 WSIs from 1094 breast cancer patients, including 1580 tumor samples and 399 normal samples. We also collected the breast cancer slides from CPTAC-BRCA project, which had 642 WSIs from 134 participants. This CPTAC-BRCA dataset was used as an independent test set to verify the generalizability of our model. The lung cancer dataset comes from the project TCGA-LUSC and TCGA-LUAD. There are in total 2168 WSIs from 1010 patients, including 1577 tumor WSIs and 591 normal WSIs.

Only the slide with a magnification greater than 20* were included in this study. The slide-level diagnosis provided by TCGA database were used as ground truth for classification labels.

Slide annotation

A pathologist with more than 10-year experience was asked to manually annotate a few slides. The annotated slides were used for the validation of spatial localization of tumor and necrosis area, as well as tumor infiltrating lymphocytes. During annotation, the pathologist was blinded with regard to any molecular or clinical feature.

RNA-seq dataset

The RNA-seq datasets from TCGA are used for differential expression analysis. We selected the patients with both tumor and normal pathology images, and obtained 153 matched samples for breast cancer. Genome-wide expression profiles in fragment per kilobase million with upper-quartile normalization (FPKM-UQ) were used for differential gene expression analysis.

Cancer driver genes

In this study, we focused on the cancer-driven genes, as their differential expression is the drive force for the generation and proliferation of tumor cells. The driver genes is obtained from the intogen compendium of cancer driver genes curated by Martínez-Jiménez *et al.* [26]. From a total of 568 driver genes, we chose the ones that have been reported high cumulative number of nonsynonymous mutations. As a result, we filtered out top 200 most mutational driver genes included in our differential expression analysis.

Image processing

Due to the high dimensionality of the WSIs (up to 100 000×100 000 pixels), we tessellated slides into small-size squares (tiles) so that they were ready as input of deep learning model. The slides were read using python library openslide tool [27]. We first converted the image from RGB space to HSV space, and applied median filter in the saturation channel to detect foreground tissue area with threshold 8. Next, we used the Otsu algorithm [28] to select areas that contained enough tissue cells. After exclusion of white background, the slide was tessellated to 128 × 128 μm(256 × 256 pixels) tiles. Of note, we only stacked the coordinates of each tile and the slide metadata using the hdf5 hierarchical data format. Next, those tiles where the surface area covered by tissue cells is less than 100 were removed. Finally, we got 14 705 914 tiles from 1677 TCGA-BRCA slides, 5888 085 tiles from 552 CPTAC-BRCA slides, and 12 769 300 tiles from 1983 TCGA-Lung slides. To eliminate the influence of different staining condition and imaging protocol of different WSI datasets, we applied color normalization on tiles.

Differential expression analysis

We calculated the fold change of each gene in tumor relative to normal tissues. Because the absolute expression levels of different genes vary in large scope, especially when the gene expression level of normal tissue is close to 0, fold change give rise to instability. To overcome this problem, we scaled the fold change by $\log_{10}(fc+1)$ [29] for each gene per patient. Apart from the quantitative prediction of differential expression level, we also cast the problem into binary classification task. If the $|\log_2(fc)|$ is more than 1.5 [30], the label of corresponding gene was set to 1, and 0 otherwise.

HistCode framework

Our HistCode framework included three modules, as shown in Figure 1. First, the self-supervised contrastive learning is used to learn latent representation of tiles. Next, we leveraged the gated-attention pooling to aggregate the tile-level features to build slide-level features. In the downstream tasks, we transferred the slide-level features to tumor diagnosis and differential gene expression prediction.

Contrastive learning for feature extraction

As we have only slide-level labels (tumor or normal), supervised learning is not applicable to extract features of tiles. We adopted self-supervised pretraining on large-scale unannotated tiles to obtain tile-level features. In this study, the adversarial contrastive learning, AdCo [31], is adopted for pretraining. We have also tested another contrastive learning methods, SimCLR, and found that AdCo achieved a better performance in downstream tasks.

Formally, denoting the input tile by x_i, we used Convolutional Neural Network (CNN) as the backbone network f_θ to transform tiles into latent representation $h_i = f_\theta(x_i)$, which is then projected into embedding q_i by a multilayer projection. Embedding is a low-dimensional vector describing the original object. The contrastive learning is to train the network parameter θ to discriminate query sample q_i from a set K of negative samples. The contrastive loss was defined as follows:

$$L = -\frac{1}{N}\sum_{i=1}^{N} log \frac{exp(q_i q{i}^{'}/\tau)}{exp(q_i q_i^{'}/\tau) + \sum_{k=1}^{K} exp(q_i m_k/\tau)}, \qquad (1)$$

where q_i' is the embedding of augmentation of the same instance x_i, which is considered as the positive sample for the query q_i, and τ is a positive value of temperature.

Inspired by adversarial learning, AdCo aims to generate challenging negative samples to be distinguished from query samples. Mathematically, AdCo takes the embedding of all samples as negative adversaries and updates the negative samples to maximize the contrastive loss, so that the adversarial negative samples were pushed closer to query samples. A memory bank is maintained to stack the negative samples. So, we have the following min-max problem:

$$\theta^*, \mathcal{M}^* = \mathbf{arg}\min_{\theta}\max_{\mathcal{M}} L(\theta, \mathcal{M}) \qquad (2)$$

in which $\mathcal{M}$ represents the set of dynamically updated adversarial negative samples, which can be regarded as a set of model parameters. During the training process, the network parameter θ is updated along the descending direction of the gradient, while $\mathcal{M}$ is updated along the ascending direction of the gradient. So, the parameter θ and $\mathcal{M}$ is alternately updated as below:

$$\theta \leftarrow \theta - \lambda_\theta \frac{\partial L(\theta, \mathcal{M})}{\partial \theta} \qquad (3)$$

$$m_k \leftarrow m_k - \lambda_{m_k} \frac{\partial L(\theta, \mathcal{M})}{\partial m_k} \qquad (4)$$

As the size of total tiles is too large to be loaded into memory, we randomly select a percentage of tiles from each slide and stacked them in the memory bank. Rather than a fixed percentage, we set a percentage list 0.10, 0.15, 0.25, 0.5, from which we randomly select a value as the percentage of tiles chosen as negative samples. This is reasonable because pathology tiles have relatively low information density compared with natural picture, and random sampling a portion of tiles accounts for most information and meanwhile reduce the computational overhead. More importantly, this stochastic sampling increased the difficulty of the task and thus helped to reduce overfitting. The final goal of the solution is to achieve the saddle point. The adversarial negative samples forced the encoder to capture essential information of each image to discriminate it from others.

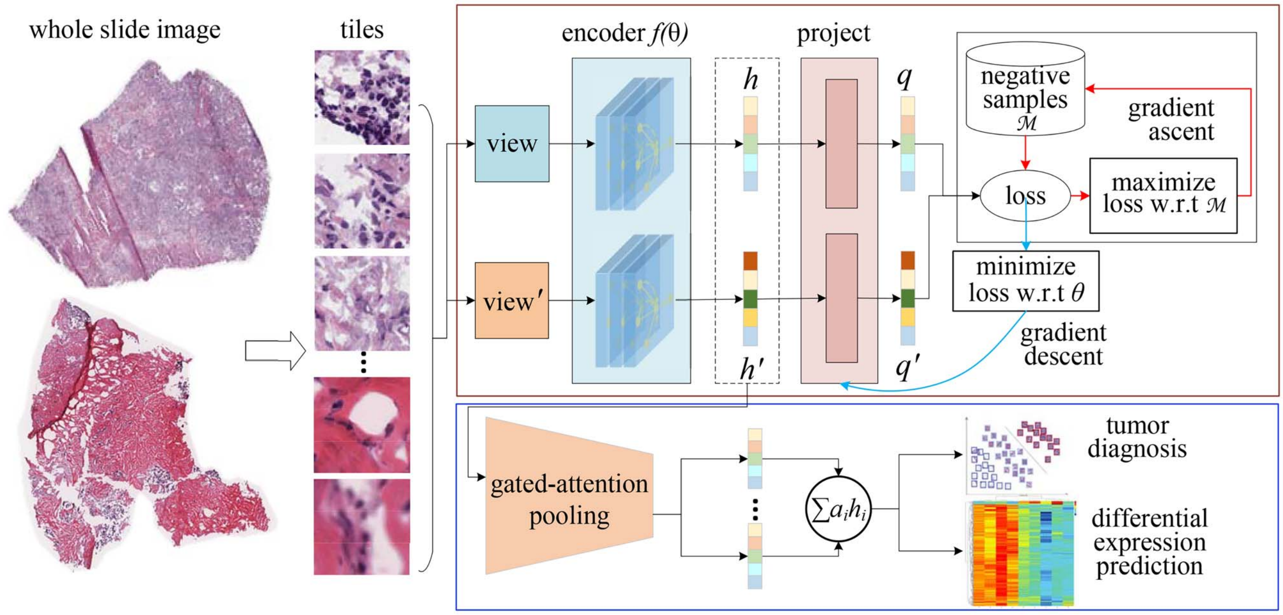

Figure 1. Illustrative flowchart of HistCode framework. There were three steps, slide preprocessing, contrastive learning-based pretraining (dark red box) transfer learning to downstream tasks (blue box). The cyan lines represent backward propagation messages, and the red lines represent the generation process of adversarial negative samples.

In our implementation, ResNet50 was used as the backbone network. We used only the first four main layers (the output of the last layer is 1024), and load the pretrained weights on ImageNet. Two fully connected layers were added to the model. In the training process, SGD [32] was used as optimizer. The learning rate of the backbone network is set to 0.03. The weight decay is set to 0.0001, and the momentum is set to 0.9.

Feature aggregation

For downstream tasks, we need to aggregate the tile-level features to derive the slide-level features. In addition to widely adopted max-pooling and mean-pooling, we also introduced gated-attention pooling [33] to aggregate tile-level features. Let $H = \{h_1, ..., h_L\}$ be L tile-level embedding of a slide, the gated-attention pooling is actually the instance-level weighted average pooling:

$$z = \sum_{i=1}^{L} a_i h_i \tag{5}$$

in which

$$a_i = \frac{exp(w^T(\mathbf{tanh}(Vh_i) \odot sigm(Uh_i)))}{\sum_{j=1}^{L} exp(w^T(\mathbf{tanh}(Vh_j) \odot sigm(Uh_j)))}, \tag{6}$$

where U, V and w^T are all learnable parameters, and w^T is used as a classifier layer that can transfer the feature vector into raw attention score, $\odot$ is an element-wise multiplication and *sigm*() is the sigmoid nonlinearity. The gating mechanism introduces a learnable nonlinearity that potentially removes the troublesome linearity in **tanh**. The major difference between gated-attention pooling and max/mean pooling was that it introduced learnable parameters that were iteratively updated during the training process, making the model highly flexible and adaptable to different downstream tasks. More importantly, the attention weights after training reflected the importance of tile-level features to downstream task, which makes the model more interpretable.

Tumor diagnosis

Given the learned slide-level representation, we applied transfer learning to conduct downstream tasks. For the tumor diagnosis task, we used a fully connection layer and a softmax layer. When the slide features are fed into the fully connected layer and the outputs are logits, then they can be used as $\widehat{y_i}$ to calculate the cross entropy loss as follows:

$$L_{TD} = -\sum_{i=1}^{S}[y_i \log \widehat{y_i} + (1 - y_i) \log(1 - \widehat{y_i})], \tag{7}$$

where S is the total number of slide included in the tumor diagnosis task and y_i is the real slide label. The softmax layer outputs the probability indicating that the input slide included tumor tissues or not. Note that during the downstream task, only the parameters of the gated-attention pooling and downstream linear layers were updated, while the parameters in the contrastive learning module were frozen.

Differential expression prediction

Since the differential gene expression is actually the change of transcriptional level in tumor cells compared with normal cells, we chose a number of tiles that was most possibly located in tumor and normal tissues to predict differential gene expression. The attention weights learned in the tumor diagnosis task were indicative of whether the tiles contain tumor cells or not; we leveraged the learned attention weights to select tiles. More precisely, we sorted the tiles according to their attention weights in descending order, and then selected the l highest tiles and l lowest tiles. In our study, l takes 100. Next, the embedding of the l highest and lowest tiles were averaged, respectively. Finally, the two averaged embeddings were concatenated. Taking the concatenated embedding as input to predict gene differential expression level.

We train a prediction model for each gene independently, as done by a few prior works [13, 29]. For each gene of interest, a linear model includes only a fully connected layer, and one output node was trained to predict its differential expression levels on patient-level samples. The linear model includes only a fully connected layer with 1024 input nodes and 1 output node. The mean squared error (MSE) was used as loss function:

$$L_{DE} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \widehat{y_i})^2, \tag{8}$$

where y_i is the RNA-seq derived differential expression level, and $\widehat{y_i}$ is the predicted one and N is the total number of patients included in the differential expression task. In our practice, we found that a single fully connected layer can achieve good performance in the downstream tasks.

Results

Tumor diagnosis

For tumor diagnosis, we performed a classification task to predict the slide-level labels. For both breast and lung cancer datasets, we adopted 5-fold cross validation to evaluate model performance. For each fold, the slides were split into three subsets at the level of patients, and 80% slides were used for training, 10% for validation and 10% for testing. The model performance was reported by the mean predicted accuracy on testing set of 5-fold cross-validation.

On TCGA-LUNG cohort, our method achieved accuracy 0.963 and ROC-AUC 0.976. Moreover, we compared our method with three other competitive methods, including MIL-RNN [34], ABMIL [33] and DSMIL [10]. As shown in Table 1, our method outperformed these competitive methods by at least 4% accuracy. The results show that HistCode successfully extract the features from pathological images for tumor diagnosis.

Also, we explored the impact of different tile-level feature aggregation strategies. We presented the ROC curve and precision-recall curves of three aggregation strategies on the TCGA-LUNG cohort in Figure 2. It can be found that the gated-attention pooling acquired significantly better performance than mean-pooling, while max pooling performed relatively poor.

Table 1. Performance comparison of HistCode and competitive methods on TCGA-LUNG cohort

Method	Accuracy	AUC
MIL-RNN[34]	0.8619	0.9107
ABMIL[32]	0.9000	0.9551
DSMIL[35]	0.9268	0.9633
Max pooling	0.8930	0.8846
Mean pooling	0.9114	0.9073
HistCode	0.9630	0.9765

On TCGA-BRCA cohort, our model achieved ROC-AUC 0.965 and accuracy 0.960, as shown in Figure 3. To verify the generalizability of our model, we trained the model on TCGA-BRCA cohort, and then used it to predict the slide labels of CPTAC-BRCA cohort. On CPTAC-BRCA cohort, our model obtained ROC-AUC value 0.962 and accuracy 0.931. This result strongly verified the robustness of our method.

Differential gene expression prediction

Based on the attention weights derived from the tumor diagnosis, we chose a number of tiles with highest and lowest attention weights to established the slide-level feature for differential expression prediction (see method), so that the resulting features integrated both tumor and normal tiles.

For the regression task of differential expression levels, we used Pearson and Spearman correlation coefficient as evaluation metrics. On the TCGA-BRCA cohort, almost all genes yield significant prediction results. The average Pearson correlation coefficient of all test genes is 0.185 (P-value<0.01). Also, we observed inter-gene variation of the coefficients. Among the 200 tested genes, the Pearson coefficient of 84 genes is more than 0.20, and that of 39 genes is more than 0.40. The distribution of coefficients was shown in Figure 4 (a). To verify the reliability of the differential expression prediction, we compared our prediction results with random baseline. For this purpose, we generated random numbers within the 5–95% range of the real differential expression (logarithm transformed fold changes) across all samples for each gene, and then calculated the correlation coefficient between the random baseline and the real differential expression levels. As shown in Figure 4 (b), we noticed the correlation coefficients followed nearly Gaussian distribution (μ=0 and σ=0.15), which is far from the results of our prediction results. The statistical test verified the difference between our prediction and random generation is strongly significant (P-value=2.39e-17, Wilcoxon rank sum test). We have also tested the remaining 335 cancer driver genes included in the intogen compendium; we observed similar results, as shown in Figure S1.

We went further to run hypotheses test for each gene to check whether the predicted differential expression levels were significantly different from random baseline. The bilateral Wilcoxon test with Benjamin–Hochberg multiple testing correction was conducted for each gene.

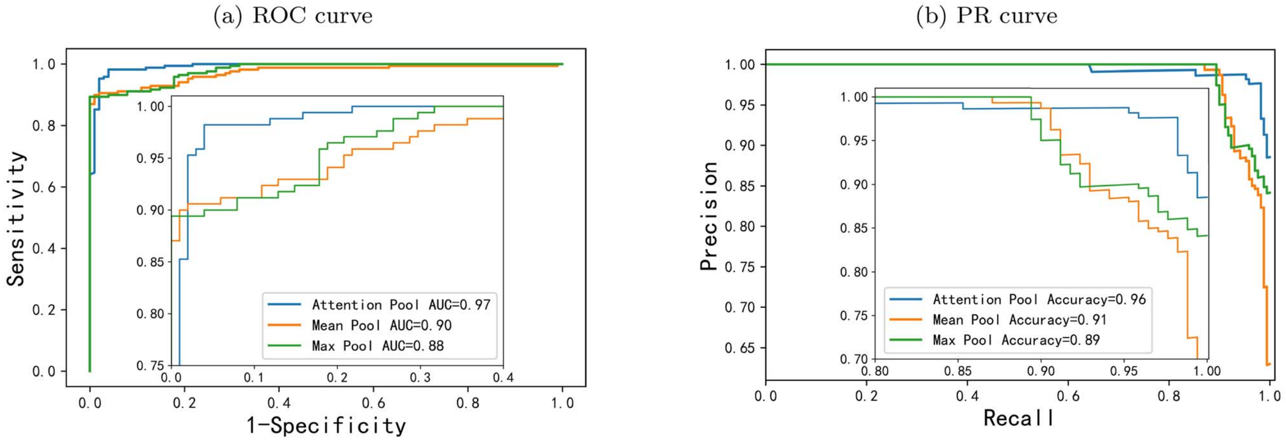

Figure 2. ROC curves and precision-recall curves achieved by our method for tumor diagnosis task on TCGA-LUNG cohort. It shows that gated-attention pooling of tile-level features improves the performance, compared with max-pooling and mean-pooling.

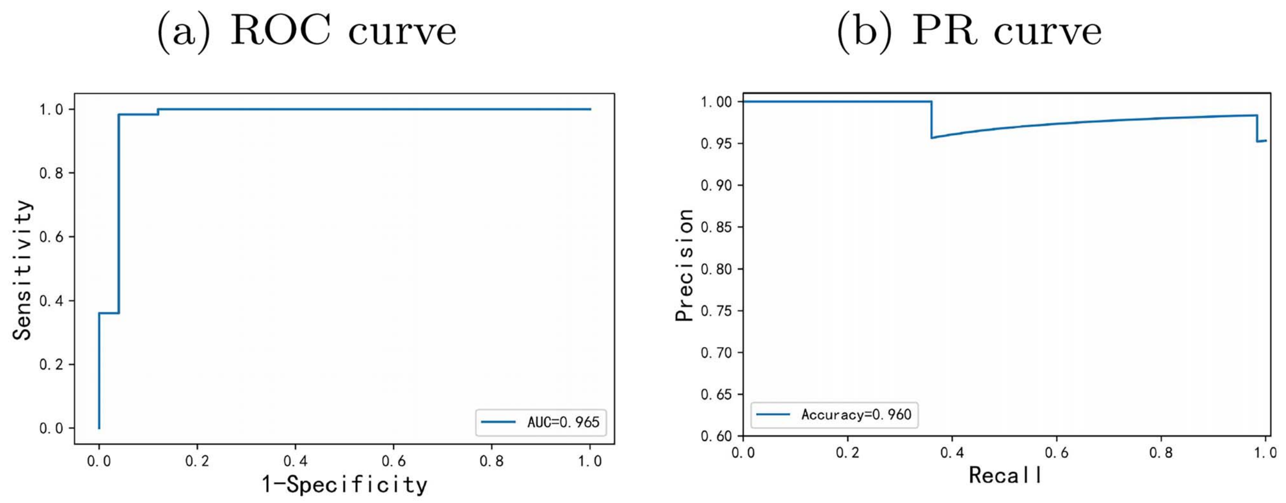

Figure 3. ROC curves and precision-recall curves achieved by our method for tumor diagnosis task on TCGA-BRCA cohort.

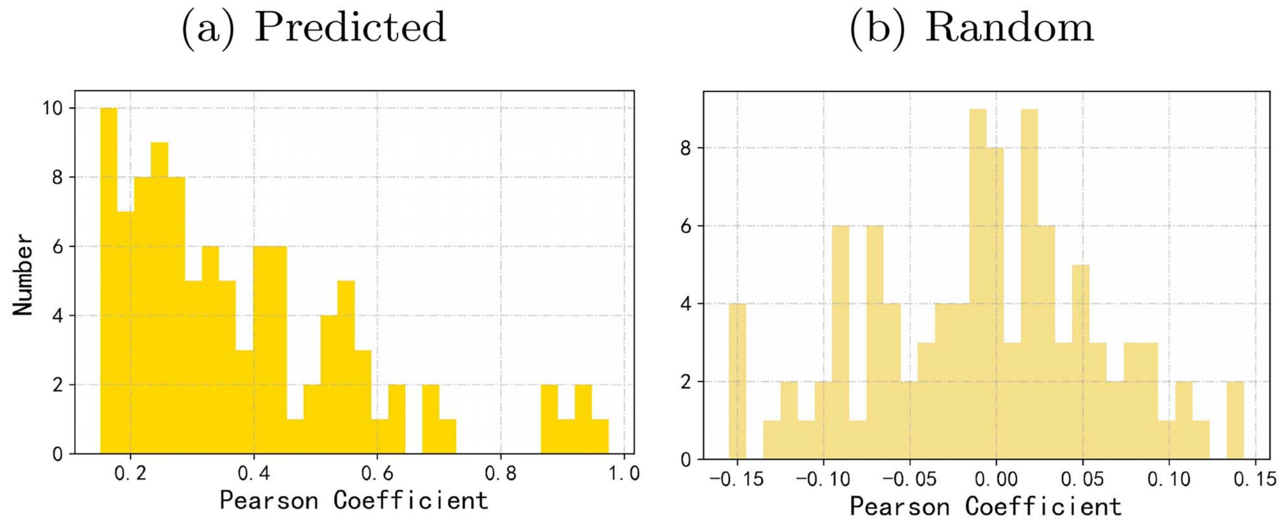

Figure 4. The frequency distribution of cancer driver genes with respect to correlation coefficients between the predicted differential expression levels and real ones, compared with random guess. The differential expression levels were predicted by slide-level pathological features on TCGA-BRCA cohort.

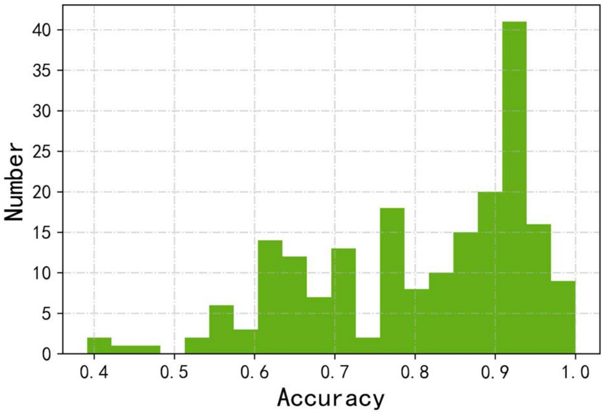

Figure 5. The frequency histogram of cancer driver genes with respect to their prediction accuracy of differential expression. The differential expression was cast to a binary classification task, and the prediction was based on the slide-level pathological features on TCGA-BRCA cohort.

Among the 200 cancer driver genes, 88% (n=176) genes were verified (P-value<0.05).

Moreover, we cast the differential expression prediction to binary classification problem, by discretizing fold change levels. If the $|\log_2(fc)|$ was more than 1.5, the gene was regarded as significantly upregulated or downregulated and its label was set to 1, and 0 otherwise. Upon the same slide-level features and linear prediction model, we just replaced the output layer by two nodes and used softmax activation function for classification task. For most of 200 cancer driver genes, their differential expression can be significantly predicted, as shown in Figure 5. The predictive accuracy 0.9 account for more than 33 percent genes(n=66). There are 60 percent genes(n=119), and their prediction accuracy was more than 0.8. This result also supported that our method achieved a notable performance in predicting differential gene expression using computational pathological features.

Contrastive learning and attentive pooling improved performance

To verify whether the contrastive learning-based feature extraction improves the performance of downstream task, we compared our method with supervised model without pretraining. For this purpose, we used the ResNet50 network pretrained on ImageNet as a backbone to extract feature from tiles, and gated-attention pooling was applied for feature aggregation. For the differential expression prediction task, we reported Pearson and Spearmen coefficients in Figure 6. When gated-attention pooling was applied, the AdCo-based contrastive learning performed better than ResNet50-based feature extraction (AdCo+attention vs ResNet50+ attention).

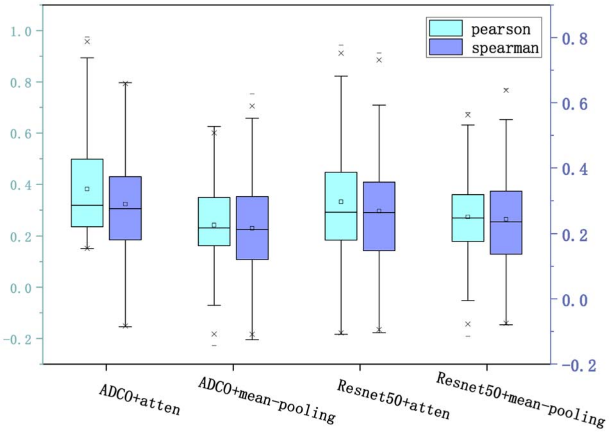

Figure 6. Comparison of the models with pretraining versus without pretraining, as well as gated-attention pooling versus mean-pooling. The performance was evaluated on the differential expression prediction task.

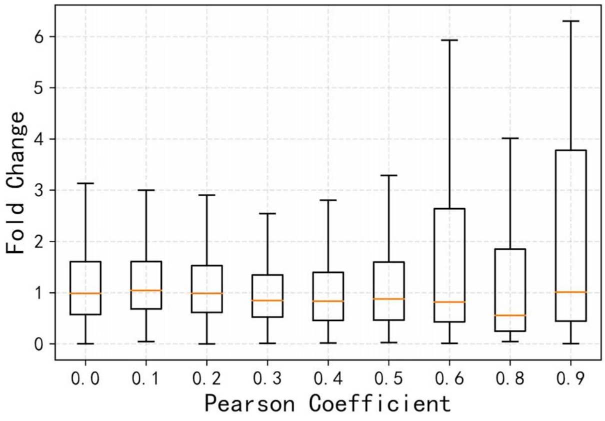

Figure 7. Boxplot of fold changes with respect to predictive performance of differentail expression. The cancer driver genes were binned according to the correlation coefficients of predicted and real differential expression levels.

Meanwhile, we investigated the impact of gated-attention mechanism. First, in the tumor diagnosis classification task, we have validated that attention-pooling outperformed mean-pooling and max-pooling. Here, we verified its improvement of performance on regression task. Without loss of generality, we took into account contrastive learning-derived and supervised learning-derived tile-level feature. As shown in Figure 6, attention-pooling achieved better prediction accuracy than mean-pooling, for both AdCo-based and ResNet50-based features.

Fold-change level positively correlated to prediction accuracy

With the assumption that molecular pattern underlies the histological morphology, we speculated that the change of expression level can be reflected in the pathological feature. More intuitively, greater change of gene expression level leads to a more significant pathological feature change that can be captured by our model. To verify this viewpoint, we checked whether the fold-change levels are correlated to the prediction accuracy of differential expression. We drew boxplot of fold-change levels with respect to Pearson coefficients, as shown in Figure 7. Overall, the prediction accuracy (Pearson coefficient) is positively correlated to the fold-change level. In particular, for the driver genes with prediction accuracy greater than 0.2, the mean fold changes was 2.602, while for the genes with prediction accuracy less than 0.2, the mean fold change was only 2.198. For those genes with prediction accuracy greater

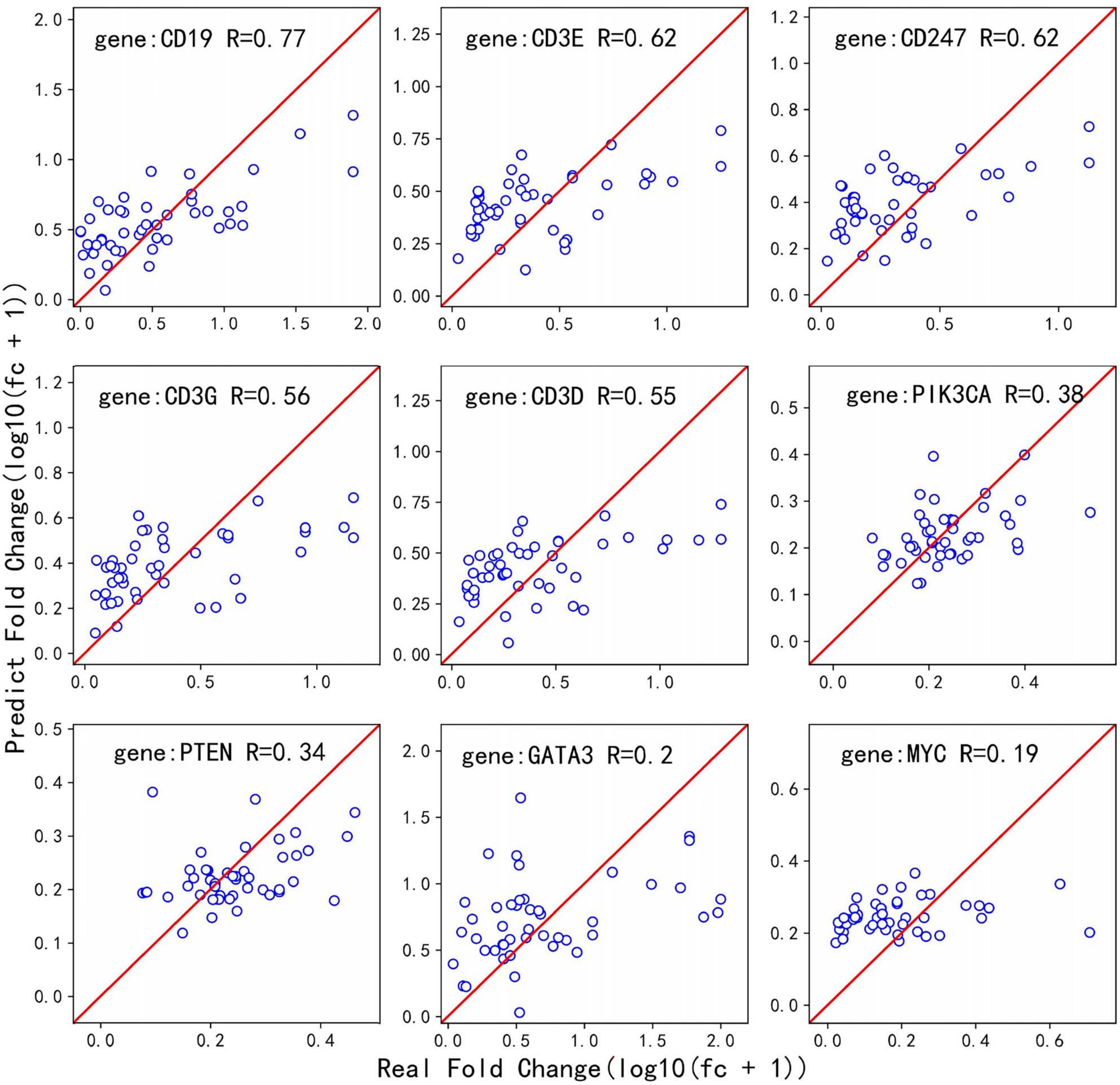

Figure 8. Scatter plots of predicted and real differential expression levels for four frequently mutated genes in breast cancer and five immune-related genes. For each gene, the Pearson coefficient was also shown.

than 0.5, the mean fold-change level reached 3.241. We have also tested the remaining 335 cancer driver genes included in the intogen compendium; we observed similar results— the fold-change levels were positively correlated to prediction accuracy, as shown in Figure S2. Therefore, we drew the conclusion that our model actually captured the underlying molecular features dominated by gene expression patterns.

Frequently mutated and immune-related genes were well predicted

For further verification of our model, we checked whether the differential expression of frequently mutated driver genes are well predicted or not. Besides, as tumor immune microenmironment reflected the interaction between immune system and tumor evolution, we also considered the immune-related genes. From the list of 10 frequently mutated genes in breast cancer [35], we got four genes, PIK3CA, MYC, PTEN and GATA3, which overlapped with compendium of cancer driver genes. We also selected four genes, CD3D, CD3E, CD3G, CD247, that encode the subunits of T lymphocyte glycoprotein CD3 receptor [36]. For B cell population, its marker gene CD19 was included.

The results are shown in the Figure 8. It can be seen that the differential expression patterns of these frequently mutated and immune-related genes can be significantly predicted. In particular, the genes related to immune cells achieved the average Pearson correlation coefficient 0.624. For CD19 gene, the predicted fold changes showed strongly positive correlation to RNA-seq results.

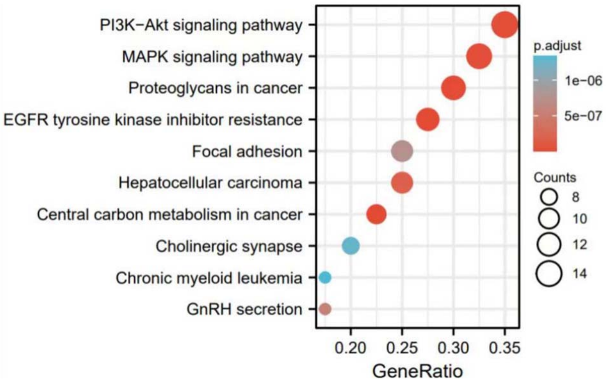

Figure 9. Enrichment analysis on KEGG signaling pathway based on top 50 genes of which differential expression were predicted.

For TCGA-BRCA cohort, we inspected the biological functions of top five best predicted genes among all 200 driver genes, including Elf3, SRGAP3, FGFR2, BRCA2 and FANCD2. Among them, BRCA2 (*P*-value=2.3e-7) is the notorious susceptibility gene to breast cancer and ovarian cancer [37], and involved in cell cycle control, gene transcription regulation, DNA damage repair, apoptosis and other important processes [38]. Also, FGFR2 (*P*-value=4.8*e*-4) plays an important role in regulating cell proliferation, survival, migration and differentiation, and can induce mitosis and promote the occurrence of cancer [39]. Elf3 (*P*-value=1.2e-6) plays an important role in development, differentiation and transformation [40]. SRGAP3 (*P*-value=7.9*e*-6) regulates cytoskeleton and participates in cell migration [41]. FANCD2(*P*-value=5.7e-6) protein is necessary to ensure efficient replication of common fragile sites [42].

Finally, we selected 50 most predicted genes for enrichment analysis. The results are shown in Figure 9. Among the top 10 enriched signaling pathways, we found the PI3K pathway, MAPK pathway and EGFR signaling that have been reported abnormal activation in breast cancer.

Interpretability and spatial localization

Human-readable interpretability of the extracted feature from pathology images can validate that the predictive power of our model kept in line with the well-known morphology annotated by experienced pathologist. Our slide-level prediction of tumor diagnosis was made by identifying and aggregating tile-level features that are of high diagnostic importance (high attention score). The derived attention scores are highly indicative of localization of tumor area. To visualize and interpret the relative importance of each region of the WSIs, we generated the attention heatmap by normalization of attention scores and spatial deconvolution of tiles to original slide. Fine-grained attention heatmap were evaluated by aligning to annotated tumor and necrosis, normal areas by pathologist. As shown in Figure 10, the attention heatmap reflected the spatial localization of highly diagnostic tissues, which were coincident with the boundary of tumor and necrosis and normal tissues delineated by the pathologist. Visiting tiles of both high and low attention scores within a slide would convey more human-readable pathological feature. So, we visualized a few tiles selected according attention scores, and found that the tiles with low attention scores are mostly normal tissue, while those with high attention scores were dominantly covered by tumor cells.

Beyond the spatial localization of tumor tissue, differential gene expression also allowed us to visualize the spatial distribution of immune infiltrating cells. For this purpose, we chose the immune cell marker genes for further analysis. As T lymphocytes and B lymphocytes infiltrating to tumor tissue play the main anticancer effect, we selected several genes specific to T and B lymphocytes. For T lymphocyte, we chose CD3D and CD247 genes that encode T lymphocyte receptor glycoprotein CD3. For B lymphocyte, the CD19 gene was chosen. Besides, we used another tumor suppressor gene PTEN for comparison. For each gene, we generated the slide-level heatmap using the attention scores of each tiles derived from the differential expression prediction models specific to this gene. Meanwhile, we asked the pathologist to manually label the tumor area, which was used as a reference boundary of the spatial distribution of immune infiltrating cells.

As shown in Figure 11, the heatmap of CD3D and CD247 marker genes indicated that T lymphocyte was mostly located at the tumor area. For the CD19 gene reflecting B lymphocyte localization, we observed similar spatial distribution to T lymphocyte. The observation was consistent to the fact that immune cells infiltrated to solid tumor to kill cancer cells by recognition of neoantigen. In contrast, the PTEN expression showed quite a different spatial distribution. Interestingly, we found that PTEN gene showed high expression levels in a normal tissue compared with to a tumor tissue.

Discussion and Conclusion

Genetic testing has been an important clinical examination for tumor subtyping and targeted drug delivery. However, due to technology and cost reasons, genetic testing has not become clinically spread, especially in developing countries. The rapid advancement of digital pathology motivated the application of computational pathology to infer genetic mutations, microsatellite instability and tumor microenvironment. Following the rationale that the change of cell and tissue phenotype is driven by the variation of gene expression pattern, we have explored the computational pathology-based prediction of differential expression of cancer driver genes. In general, the hematoxylin and eosin-stained digital slides contained both tumor and normal tissues; this allows us to infer differentially expressed genes from pathological feature from single WSI. In fact, we have shown that the fold-change levels positively correlated to prediction accuracy. This indicated that dramatic variation of underlying molecule expression pattern

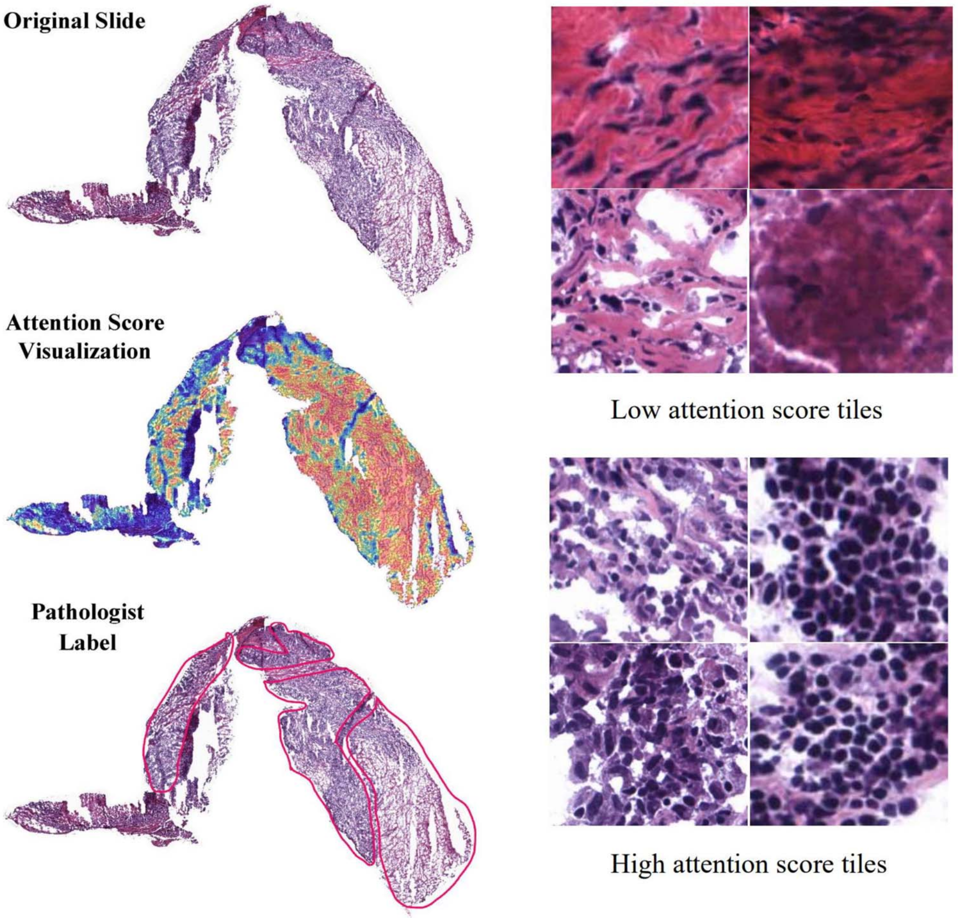
Original Slide
Attention Score Visualization
Pathologist Label
Low attention score tiles
High attention score tiles

Figure 10. Spatial localization of tumor area of implied by attention mechanism compared with pathologist annotations. The heatmap (left middle) was generated by spatial deconvolution of tiles to original slide (left top), and each tile was colored according to its normalized attention score. The area circled by the red lines were tumor and necrosis area (left bottom) annotated by an experienced pathologist. The right column showed some representative tiles with the highest and lowest attention scores.

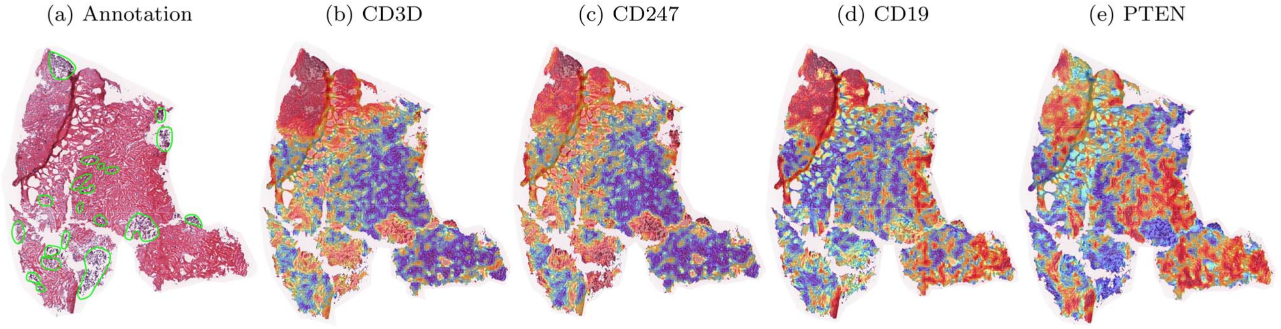

Figure 11. Spatial localization of tumor area and cells labeled by marker genes. The tumor area was labeled by an experienced pathologist. The localization of T lymphocytes and B lymphocytes was predicted by their marker genes.

would be more reflected in phenotypic features, which in turn allows us to infer differential expression.

The inference from pathological feature to molecule patterns mainly depends on the powerful feature extraction capacity of deep learning. Although weakly supervised deep learning has been applied in a few studies for computational pathology, self-supervised contrastive learning showed superior performance in representation learning, but has not been exploited in mining digital pathology. Our study verified that the contrastive learning-based pretraining significantly improved the performance in downstream classification (tumor diagnosis) and regression (differential expression) tasks. We speculate that contrastive learning can capture fine-grained information in learning to acquire similar features for tile-level positive pair but dissimilar from a set of negative tiles. In contrast, weakly supervised learning may only extract information for slide-level instance classification. Here, we use the adversarial contrastive learning for feature extraction; however, it needs a huge memory bank to store a large number of negative samples. How to reduce hardware consumption will be a key step toward wide range application.

Different from the discovery of novel cancer driver genes [43, 44], we focus on predicting the differential expression of genes of interest. For each cancer driver gene, we train a prediction model for each gene independently, as done by a few prior works [13, 29]. We chose the cancer driver genes from intogen database [26] with high cumulative frequencies of non-synonymous mutations because these genes are highly clinical value and may account for a larger proportion of carcinogenic causes. For other non-driver genes, we have tested several immune-related genes, and found our model achieved high performance for these immune-related genes. The bulk RNA-seq dataset is the standard for differential expression analysis, which has been widely used to identify upregulated and downregulated genes in tumor cells compared with normal cells. However, bulk sequencing lost the spatial localization of tumor and normal tissues. The spatial deconvolution of marker gene expression to histopathology images greatly helps to visualize the spatial distribution of tumor and normal cells, especially the immune infiltrating cells. Looking forward, computational pathology would promote the spatial visualization of tumor immune microenvironment.

Although our method achieved a promising performance on downstream tasks, we think our model can be improved at least in two aspects. First, in practice, we found that the adversarial contrastive learning needs a huge memory bank required to store a large number of negative samples. The requirement for long training time and large memory leads to the application failure of contrastive learning-based pretraining to large-scale dataset. Second, we found that our model did not perform well for all driver genes. As shown in Figure 4a, the correlation coefficient of some driver genes were about 0.2. On one hand, we think the weak predictive power on these genes may be attributed to the features extracted from pathological images. The extracted feature may not contain related information of these genes. On the other hand, some genes encode signal transduction proteins that have very low concentrations in cells and have no direct phenotypic effect on the cell and tissue morphology. So, these genes are difficult to be predicted in nature from pathological feature. In fact, our result is consistent with previous studies [29, 45]. To address the problems, we consider using more advanced self-supervised learning model for feature extraction. Also, we will check the biological functions of driver genes and further screened the protein-coding genes that are significantly reflected in pathological features.

In summary, we developed a self-supervised contrastive learning framework, HistCode, to infer differential gene expression from pathology images. Our experiments showed that contrastive learning-based pretraining effectively improved the downstream task, including tumor diagnosis and differential expression prediction. We have also shown informative spatial visualization of tumor and specific gene expression by leveraging the tile-level attention scores learned by our model. We believe that our study would yield inspiring insight into computational pathology.

Key Points

- Self-supervised contrastive learning was applied to large-scale unlabeled digital pathology images and extracted tile-level features, which were then aggregated to build the slide-level features via gated-attention pooling. Adversarial negative samples were generated to pose challenges that drove the self-supervised learning to capture informative representation from large-scale unlabeled tiles.
- The computational pathological features have been shown to be highly predictive of both tumor diagnosis and differential gene expressions. Interestingly, the prediction accuracy was positively correlated to fold-change level, which indicated that dramatic variation of underlying molecule expression pattern would be more reflected in phenotypic features.
- We explored the model interpretability via spatial deconvolution, and colored each tile according to its normalized attention scores. The spatial localization of high attention-scored tiles showed high consistence to the distribution of tumor tissues and immune infiltrating cells annotated by an experienced pathologist.

Author contributions statement

H.H. and H.L. conceived the main idea and the framework of the manuscript. H.L. collected the datasets. H.H. drafted the manuscript. H.H. and G.Z. performed the experiments. X.L. and L.D. helped to improve the idea. C.W. reviewed drafts of the paper. D.Z. annotated the pathology images. H.L. revised the manuscript. H.L. supervised the study and provided funding. All authors read and commented on the manuscript.

Acknowledgments

We would like to thank The Cancer Genome Atlas (TCGA) for providing free WSI datasets and RNA-seq data.

Data availability

Source code and all datasets used in this study are available at https://github.com/hoarjour/HistCode/

Funding

This work was supported by the National Natural Science Foundation of China under grants No. 62072058 and No.61972422.

References

1. Kather JN, Krisam J, Charoentong P, *et al.* Predicting survival from colorectal cancer histology slides using deep learning: A retrospective multicenter study. *PLoS Med* 2019;**16**(1):e1002730.
2. Bychkov D, Linder N, Turkki R, *et al.* Deep learning based tissue analysis predicts outcome in colorectal cancer. *Scientific reports* 2018;**8**(1):1–11.
3. Saillard C, Schmauch B, Laifa O, *et al.* Predicting survival after hepatocellular carcinoma resection using deep learning on histological slides. *Hepatology* 2020;**72**(2):2000–13.
4. Jie J, Wismans LV, Mustafa DAM, *et al.* Robust deep learning model for prognostic stratification of pancreatic ductal adenocarcinoma patients. *Iscience* 2021;**24**(12):103415.
5. Cheerla A, Gevaert O. Deep learning with multimodal representation for pancancer prognosis prediction. *Bioinformatics* 2019;**35**(14):i446–54.
6. Hou L, Samaras D, Kurc TM, *et al.* Patch-based convolutional neural network for whole slide tissue image classification. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, 2424–33.
7. Noorbakhsh J, Farahmand S, Soltanieh-ha M, *et al.* Deep learning-based cross-classifications reveal conserved spatial behaviors within tumor histological images. *Nat Commun* 2020;**11**(1):1–14.
8. Mobadersany P, Yousefi S, Amgad M, *et al.* Predicting cancer outcomes from histology and genomics using convolutional networks. *Proc Natl Acad Sci* 2018;**115**(13):E2970–9.
9. Ström P, Kartasalo K, Olsson H, *et al.* Artificial intelligence for diagnosis and grading of prostate cancer in biopsies: a population-based, diagnostic study. *Lancet Oncol* 2020;**21**(2): 222–32.
10. Li B, Li Y, Eliceiri KW. Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning. *Conf Comput Vis Pattern Recognit Workshops* 2021;**2021**:14318–28.
11. Saltz J, Gupta R, Hou L, *et al.* Spatial organization and molecular correlation of tumor-infiltrating lymphocytes using deep learning on pathology images. *Cell Rep* 2018;**23**(1):181–93.
12. Gheisari S, Catchpoole DR, Charlton A, *et al.* Convolutional deep belief network with feature encoding for classification of neuroblastoma histological images. *Journal of pathology informatics* 2018;**9**(1):17.
13. Kather JN, Heij LR, Grabsch HI, *et al.* Pan-cancer image-based detection of clinically actionable genetic alterations. *Nature Cancer* 2020;**1**(8):789–99.
14. Yu F, Jung AW, Torne RV, *et al.* Pan-cancer computational histopathology reveals mutations, tumor composition and prognosis. *Nature Cancer* 2020;**1**(8):800–10.
15. Chen RJ, Lu MY, Wang J, *et al.* Pathomic fusion: an integrated framework for fusing histopathology and genomic features for cancer diagnosis and prognosis. *IEEE Trans Med Imaging* 2022;**41**(4):757–70.
16. Coudray N, Ocampo PS, Sakellaropoulos T, *et al.* Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning. *Nat Med* 2018;**24**(10):1559–67.
17. Kather JN, Pearson AT, Halama N, *et al.* Deep learning can predict microsatellite instability directly from histology in gastrointestinal cancer. *Nat Med* 2019;**25**(7):1054–6.
18. Echle A, Grabsch HI, Quirke P, *et al.* Clinical-grade detection of microsatellite instability in colorectal tumors by deep learning. *Gastroenterology* 2020;**159**(4):1406–16.
19. Cao R, Yang F, Ma S-C, *et al.* Development and interpretation of a pathomics-based model for the prediction of microsatellite instability in colorectal cancer. *Theranostics* 2020; **10**(24):11080.
20. Yamashita R, Long J, Longacre T, *et al.* Deep learning model for the prediction of microsatellite instability in colorectal cancer: a diagnostic study. *Lancet Oncol* 2021;**22**(1):132–41.
21. Albarqouni S, Baur C, Achilles F, *et al.* Aggnet: deep learning from crowds for mitosis detection in breast cancer histology images. *IEEE Trans Med Imaging* 2016;**35**(5):1313–21.
22. Tellez D, Balkenhol M, Otte-Höller I, *et al.* Whole-slide mitosis detection in h&e breast histology using phh3 as a reference to train distilled stain-invariant convolutional networks. *IEEE Trans Med Imaging* 2018;**37**(9):2126–36.
23. Rao S. Mitos-rcnn: A novel approach to mitotic figure detection in breast cancer histopathology images using region based convolutional neural networksarXiv preprint arXiv:1807.01788. 2018.
24. Li C, Wang X, Liu W, *et al.* Weakly supervised mitosis detection in breast histopathology images using concentric loss. *Med Image Anal* 2019;**53**:165–78.
25. Ash JT, Darnell G, Munro D, *et al.* Joint analysis of expression levels and histological images identifies genes associated with tissue morphology. *Nat Commun* 2021;**12**:1609.
26. Martínez-Jiménez F, Muiños F, Sentís I, *et al.* A compendium of mutational cancer driver genes. *Nat Rev Cancer* 2020;**20**(10): 555–72.
27. Goode A, Gilbert B, Harkes J, *et al.* Openslide: A vendor-neutral software foundation for digital pathology. *Journal of pathology informatics* 2013;**4**(1):27.
28. Otsu N. A threshold selection method from gray-level histograms. *IEEE Trans Syst Man Cybern* 1979;**9**(1):62–6.
29. Schmauch B, Romagnoni A, Pronier E, *et al.* A deep learning model to predict rna-seq expression of tumours from whole slide images. *Nat Commun* 2020;**11**(1):1–15.
30. Liao Y, Yin G, Wang X, *et al.* Identification of candidate genes associated with the pathogenesis of small cell lung cancer via integrated bioinformatics analysis. *Oncol Lett* 2019;**18**(4): 3723–33.
31. Qianjiang H, Wang X, Wei H, *et al.* Adco: Adversarial contrast for efficient learning of unsupervised representations from self-trained negative adversaries. In: *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, June 19–25*. Computer Vision Foundation/IEEE, 2021, 1074–83.
32. Bottou L. Large-scale machine learning with stochastic gradient descent. In: *Proceedings of COMPSTAT'2010*. Physica-Verlag HD, 2010, 177–86.
33. Ilse M, Tomczak J, Welling M. Attention-based deep multiple instance learning. In: *International conference on machine learning*. PMLR, 2018, 2127–36.
34. Campanella G, Hanna MG, Geneslaw L, *et al.* Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nat Med* 2019;**25**(8):1301–9.
35. Nik-Zainal S, Davies H, Staaf J, *et al.* Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature* 2016;**534**(7605):47–54.
36. Loh EY, Lanier LL, Turck CW, *et al.* Identification and sequence of a fourth human t cell antigen receptor chain. *Nature* 1987;**330**(6148):569–72.
37. Wooster R, Bignell G, Lancaster J, *et al.* Identification of the breast cancer susceptibility gene brca2. *Nature* 1995;**378**(6559): 789–92.

38. Wang F, Fang Q, Ge Z, *et al.* Common brca1 and brca2 mutations in breast cancer families: a meta-analysis from systematic review. *Mol Biol Rep* 2012;**39**(3):2109–18.
39. Wesche J, Haglund K, Haugsten EM. Fibroblast growth factors and their receptors in cancer. *Biochem J* 2011;**437**(2):199–213.
40. Schedin PJ, Eckel-Mahan KL, McDaniel SM, *et al.* Esx induces transformation and functional epithelial to mesenchymal transition in mcf-12a mammary epithelial cells. *Oncogene* 2004;**23**(9):1766–79.
41. Yang Y, Marcello M, Endris V, *et al.* Megap impedes cell migration via regulating actin and microtubule dynamics and focal complex formation. *Exp Cell Res* 2006;**312**(12): 2379–93.
42. Madireddy A, Kosiyatrakul ST, Boisvert RA, *et al.* Fancd2 facilitates replication through common fragile sites. *Mol Cell* 2016;**64**(2):388–404.
43. Schulte-Sasse R, Budach S, Hnisz D. Integration of multiomics data with graph convolutional networks to identify new cancer genes and their associated molecular mechanisms. *Nat Mach Intell* 2021;**3**:513–26.
44. Peng W, Tang Q, Dai W, *et al.* Improving cancer driver gene identification using multi-task learning on graph convolutional network. *Brief Bioinform* 2022;**23**:bbab432.
45. Lu M Y, Williamson D F K, Chen T Y, *et al.* Data-efficient and weakly supervised computational pathology on whole-slide images. *Nature biomedical engineering* 2021;**5**(6):555–70.