OXFORD

Briefings in Bioinformatics, 2024, **25(1)**, 1–10

https://doi.org/10.1093/bib/bbad464

Problem Solving Protocol

THItoGene: a deep learning method for predicting spatial transcriptomics from histological images

Yuran Jia, Junliang Liu, Li Chen, Tianyi Zhao and Yadong Wang

Corresponding authors: Tianyi Zhao, School of Medicine and Health, Harbin Institute of Technology, Harbin 150040, China. E-mail: zty2009@hit.edu.cn; Yadong Wang, School of Medicine and Health, Harbin Institute of Technology, Harbin 150040, China. E-mail: ydwang@hit.edu.cn

Abstract

Spatial transcriptomics unveils the complex dynamics of cell regulation and transcriptomes, but it is typically cost-prohibitive. Predicting spatial gene expression from histological images via artificial intelligence offers a more affordable option, yet existing methods fall short in extracting deep-level information from pathological images. In this paper, we present THItoGene, a hybrid neural network that utilizes dynamic convolutional and capsule networks to adaptively sense potential molecular signals in histological images for exploring the relationship between high-resolution pathology image phenotypes and regulation of gene expression. A comprehensive benchmark evaluation using datasets from human breast cancer and cutaneous squamous cell carcinoma has demonstrated the superior performance of THItoGene in spatial gene expression prediction. Moreover, THItoGene has demonstrated its capacity to decipher both the spatial context and enrichment signals within specific tissue regions. THItoGene can be freely accessed at https://github.com/yrjia1015/THItoGene.

Keywords: spatial transcriptomics; histopathological images; transformer; capsule network

INTRODUCTION

With the development of many spatial transcriptomics techniques that can align high-resolution hematoxylin and eosin (H&E)-stained histological images with high-throughput ribonucleic acid (RNA) sequencing, we have been able to systematically characterize cell types and states while preserving positional information [1–6]. Spatial transcriptomics technology has revolutionized our understanding of tissue structure and function by correlating cellular functionality, morphology and location with gene expression. It offers a novel perspective that facilitates the discovery of information pertaining to cell–cell interactions and molecular signaling. In the field of oncology, spatial transcriptomics enables unbiased analysis of cellular states, positions and potential interactions within the tumor microenvironment, allowing for the exploration of transcriptional differences among different pathological features of tumors [7–9]. This is of great significance for comprehensively understanding the molecular mechanisms underlying tumor development, achieving more accurate molecular subtyping of patients and discovering spatially related biomarkers. Unfortunately, the high cost of data generation limits the application of spatial transcriptome technology in large-scale studies [10].

High-resolution histological images contain rich yet underutilized biomedical signals. Previous studies have shown that abnormal gene expression or mutations often affect cell morphology, structure, and distribution, leading to changes in histological features [11, 12]. By utilizing methods such as computer vision, we can establish associations between pathological images and the occurrence and development of diseases, thereby enabling quantitative analysis of biological features such as cellular structures, protein expression and gene expression [13–15]. Jain *et al.* [16] proposed Image2TMB, a multi-scale deep learning model for detecting global morphological changes caused by aggregated mutations in individual tumor cells in routine histopathology images. MOMA is a scalable deep learning method for predicting alterations in clinical molecular events in cancer genomics, proteomics and patient prognosis developed by Yu *et al.* [17]. Wagner *et al.* [18] combined a pre-trained Transformer encoder with a Transformer network for patch aggregation to predict biomarkers from colorectal cancer histology.

Furthermore, existing methodologies have demonstrated the ability to learn specific molecular characterization from histological patterns. HE2RNA [19] employs deep learning to reconstruct transcriptomic maps from histological images, showcasing its

Yuran Jia is a PhD candidate at the School of Computer Science and Technology at the Harbin Institute of Technology of China. His research interests include bioinformatics.

Junliang Liu is a master candidate at the School of Computer Science and Technology at the Harbin Institute of Technology of China. His research interests include bioinformatics.

Li Chen is a postdoctoral researcher at the School of Life Sciences, Westlake University, with a focus on bioinformatics.

Tianyi Zhao is an associate professor at the School of Medicine and Health at the Harbin Institute of Technology of China. His research interests include molecular characterization of complex diseases, biological network analysis and machine learning.

Yadong Wang is a professor at the College of School of Medicine and Health at the Harbin Institute of Technology of China. His research interests are bioinformatics, machine learning and algorithms.

Received: October 15, 2023. **Revised:** November 10, 2023. **Accepted:** November 18, 2023

utility in clinical diagnosis. Pathological images furnish valuable information regarding tissue structure, cellular morphology and pathological features, while spatial transcriptomics techniques reveal the spatial distribution of genes and their expression patterns within tissues. Integrating these two approaches can offer a more comprehensive understanding and provide deeper insights for pathological research and disease diagnosis. Compared to the expensive ST technology, extracting molecular features from tissue section images stained with H&E is a faster and more cost-effective alternative [10, 20]. ST-Net [21] successfully combines cellular morphology with gene expression, enabling prediction from histology to spatial transcriptomics. Li *et al.* [22] developed a method named HisToGene based on the Vision Transformer (ViT), which can predict super-resolution gene expression in tumor histology images. However, ST-Net only extracts image features through convolutional neural networks (CNNs) and does not incorporate spatial positional information of gene expression during modeling, it fails to learn the associations between gene expression in spatially adjacent positions. HisToGene improves upon the existing ViT by considering both the positional information of gene expression and histological features. However, since ViT requires compressing the image into sequences for input despite the addition of spatial positional encoding in the enhanced ViT, there is still a certain degree of information loss. His2ST [10] addresses the issue of spatial relationship loss in HisToGene by explicitly learning the neighborhood relationships of histological features using a combination of CNNs and graph convolutional networks. Nevertheless, histopathology images contain a large number of morphologically diverse cellular landscapes and complex pathological features. Current methods lack the necessary flexibility and depth to interpret the complex structures found in histopathology images and are unable to effectively resolve detailed features at super-resolution, resulting in insufficient accuracy in predicting spatial gene expression.

To overcome the limitations of existing methods, we propose a novel deep learning framework called THItoGene. THItoGene is designed to accurately predict gene expression at a spatial resolution in pathology images by integrating spatial information and histological multi-view neighborhood features. Comprehensive benchmarking on spatial transcriptomic data generated from different tumors demonstrates the exceptional predictive capability of THItoGene. Moreover, THItoGene has exhibited its capability in recognizing significant alterations in gene expression within pathologist-annotated spatial domains and predicting the enrichment of specific genes in complex biological tissues.

MATERIALS AND METHODS

Overview of THItoGene

THItoGene is specifically designed to provide a comprehensive representation of the intricate relationships between gene expression, spatial location and histopathological images (Figure 1). To characterize super-resolution histopathology images, THItoGene first leverages the multidimensional attention mechanism of dynamic convolution to extract comprehensive and detailed local visual information from the initial input image. Compared to conventional convolutions, dynamic convolutional networks are able to effectively capture complex, detailed features and enhance the perception of specific cellular landscape textures through more flexible tuning capabilities. The Efficient-CapsNet module captures the spatial relationships and distribution patterns of cells through the dynamic routing mechanism, learning the spatial arrangement and interaction information of different types of cells, thereby enhancing the understanding and prediction of gene expression.

THItoGene then fuses the spot location information embedding and the depth image features learned from the histological images to perform global modeling via the ViT module to comprehend the correlation between different regions. Finally, THItoGene constructs a neighborhood network based on the relative spatial positions between spots and adaptively characterizes the relationship between the spatial position of spots and gene expression using the graph attention network (GAT) module.

Datasets

In the evaluation process of THItoGene, we utilized spatial transcriptomics data from two different tumors generated by the 10X Genomics platform: human HER2 positive breast cancer (HER2+) [23] and human cutaneous squamous cell carcinoma (cSCC) [24] datasets. In the HER2+ dataset, a total of 36 sections from eight breast cancer patients were included. We retained 32 sections with a spot count greater than 180 to assess the performance of THItoGene. In the cSCC dataset, there were 12 tissue sections from four patients included.

For the spatial transcriptomic gene expression data of the two tumors, we filtered out genes that were expressed in fewer than 1000 spots across all tissue sections and only considered the top 1000 highly variable genes expressed in the sections [22]. We then normalized the gene feature counts within each spot by dividing them by the total feature counts of all genes, multiplied by 1 000 000, and applied a natural logarithmic transformation. After processing, the HER2+ dataset contained 9612 spots and 785 genes, while the cSCC dataset contained 6630 spots and 134 genes. When processing the pathological images, we divided the entire image into N patches of size (3∗W∗H), where W and H represent the width and height of each patch. To match the diameter of the spots, we set W and H to 112 pixels.

Dynamic convolution module

Compared to conventional convolution, which employs a fixed convolution kernel to process various locations of a histological image uniformly, dynamic convolution has the capability to adaptively adjust the size of the convolution kernel according to different spatial locations. This allows for the enhanced capturing and representation of the multi-scale fine features found in diverse cellular and tissue structures within the histological image. The process can be illustrated as follows:

$$Y = (\alpha_{w1} W_1 + \alpha_{w2} W_2 + \cdots + \alpha_{wn} W_n) * X \tag{1}$$

where $X \in R^{w \times h \times c_{in}}$ and $Y \in R^{w \times h \times c_{out}}$ represent the input image and the extracted features, respectively; $W_i \in R^{k \times k \times c_{in}}$ represents the ith convolutional kernel; $\alpha_{wi} \in R$ denotes the dynamic attention coefficients; $*$ stands for convolutional operations.

Omni-dimensional Dynamic Convolution [25] is a design of dynamic convolution that learns different attention weights along the four dimensions of convolutional kernel space, including kernel size ($k \times k$), input channel number(c_{in}), output channel number(c_{out}), and the number of convolutional kernels(n), in order to better adapt to variations in input features:

$$\begin{aligned} Y = (&\alpha_{w1} \odot \alpha_{k1} \odot \alpha_{c1} \odot \alpha_{o1} \odot W_1 + \ldots \\ &+ \alpha_{wn} \odot \alpha_{kn} \odot \alpha_{cn} \odot \alpha_{on} \odot W_n) * X \end{aligned} \tag{2}$$

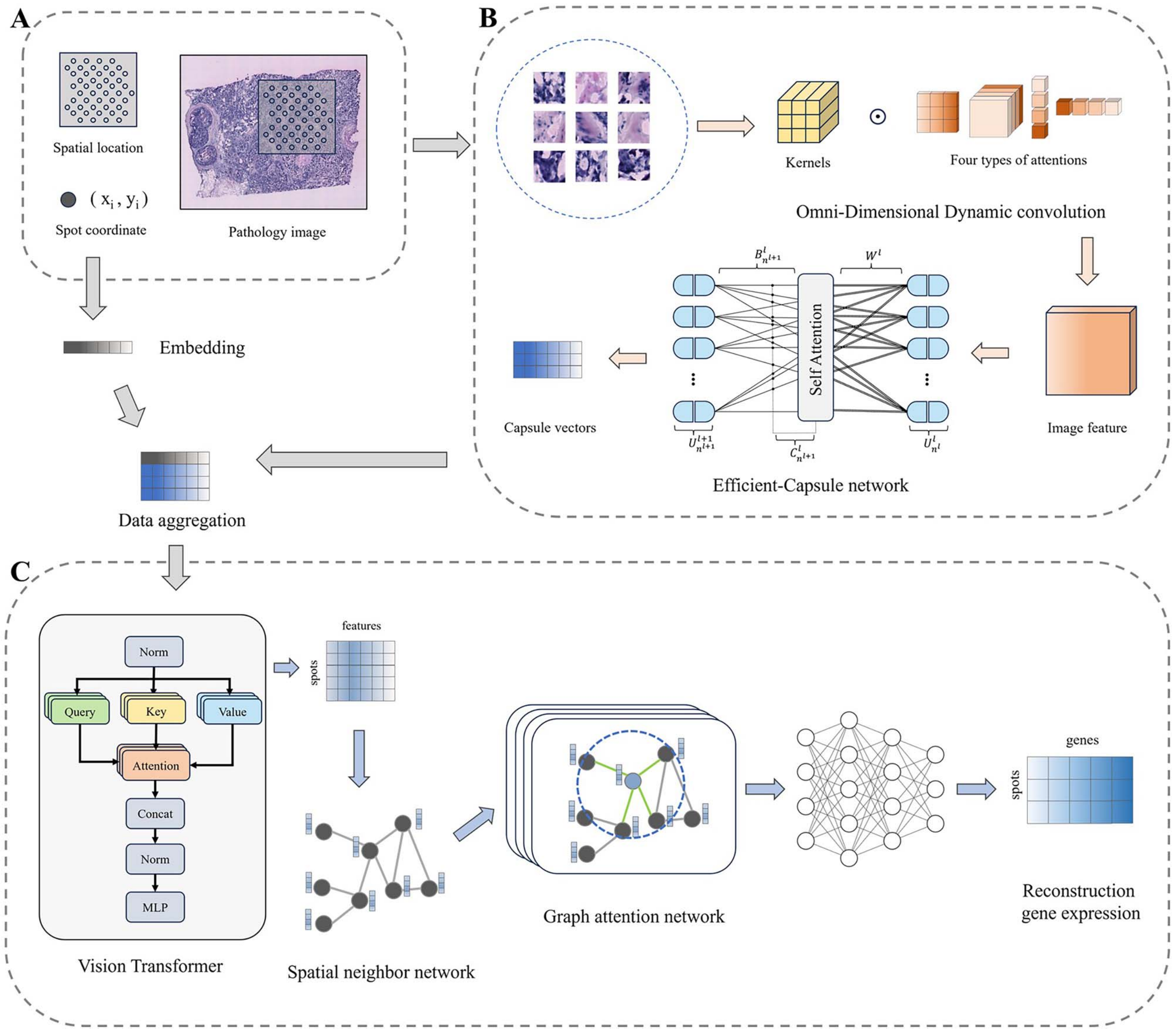

Figure 1. Overview of THItoGene. (**A**) Segmentation of the histological image into patches corresponding to each spot's location and creation of position embeddings for the spatial coordinates of each spot. (**B**) Extraction of deep molecular features from patches utilizing the dynamic convolution module and the Efficient-Capsule module. (**C**) Global modeling of position embeddings and pathology features using the ViT module to understand cross-positional association relationships in the images, followed by adaptive learning of the relationship between the spatial position of the neighborhood of the spot and gene expression based on the GAT module.

In this equation, $\alpha_{wi} \in R$ represents the attention coefficient of the convolutional kernel W_i. $\alpha_{ki} \in R^{k\times k}$, $\alpha_{ci} \in R^{c_{in}}$ and $\alpha_{oi} \in R^{c_{out}}$, respectively, denote the attention weights computed along the spatial dimensions of the convolutional kernel W_i, the input channel dimension, and the output channel dimension. The symbol $\odot$ represents element-wise multiplication across different dimensions of the kernel space. The α_{wi}, α_{ki} , α_{ci} and α_{oi} are computed using the multi-head attention module $\varphi_i(x)$. In $\varphi_i(x)$, the input x is first compressed into a feature vector of length c_{in} through the Global Average Pooling operation. Then, the compressed feature vector is mapped to a lower-dimensional space using a fully connected (FC) layer. Finally, for each attention head branch, there is an output FC layer of size $n \times 1, k \times k$, $c_{\mathrm{in}} \times 1$ and $c_{\mathrm{out}} \times 1$, as well as a Sigmoid function, used to generate normalized attention coefficients α_{wi}, α_{ki}, α_{ci} and α_{oi}.

Efficient-CapsNet module

The Efficient-CapsNet [26] module is an optimized version of the Capsule Network that utilizes self-attention mechanisms to effectively route capsules, significantly reducing the number of model parameters. By employing dynamic routing mechanisms, the Efficient-CapsNet enables communication and collaboration among capsules. The dynamic routing allocates weights between capsules based on their similarities, thereby determining the final output feature representation. This mechanism helps capture spatial relationships and hierarchical structures among features, allowing for the learning of multi-level feature representations in pathological images and modeling the spatial relationships between features. In the Efficient-CapsNet, the input is first mapped to a high-dimensional space:

$$F^{l+1}\left(X^{l}\right) = \mathrm{Relu}\left(\mathrm{Conv}_{k\times k}\left(X^{l}\right)\right) \tag{3}$$

The shape of X^l is represented as $W \times H \times C$, where W, H and C denote the width, height, and number of channels of the convolved image, respectively. The convolutional kernel has a size of $k \times k$ and a stride of $s = 1$. Subsequently, C convolutional kernels, each of the same size as in Eq. (4), are applied to each channel independently, resulting in the primary capsule layer denoted as $S^l_{n,d}$, where n and d represent the number of primary capsules in the lth layer and the dimensionality of a single capsule, respectively.

After this, it is activated by the squash operation and routed to the whole by the self-attention routing algorithm:

$$u^l_n = \mathbf{squash}\left(s^l_n\right) = \left(1 - \frac{1}{e^{\|s^l_n\|}}\right)\frac{s^l_n}{\|s^l_n\|} \tag{4}$$

$$U^l_{n^{l+1}} = u^l_n \times W^l \tag{5}$$

$$s^{l+1}_n = \left(U^l_{n^{l+1}}\right)^T \times \left(C^l_{n^{l+1}} + B^l_{n^{l+1}}\right) \tag{6}$$

where s^l_n and u^l_n represent the capsules before and after activation, respectively; $B^l_{n^{l+1}}$ represents the log priors matrix; $U^l_{n^{l+1}}$ represents the prediction of all primary capsules in the lth layer, and W^l is used for the prediction's weight matrix. $C^l_{n^{l+1}}$ represents the matrix of all coupling coefficients generated after self-attention, and it is summed with $B^l_{n^{l+1}}$ with to get the final routing weights. The coupling coefficient $C^l_{n^{l+1}}$ can be expressed as:

$$C^l_{n^{l+1}} = \frac{\mathbf{exp}\left(\sum_{n^l} \frac{U^l_{n^{l+1}} \times \left(U^l_{n^{l+1}}\right)^T}{\sqrt{d^l}}\right)}{\sum_{n^{l+1}} \mathbf{exp}\left(\sum_{n^l} \frac{U^l_{n^{l+1}} \times \left(U^l_{n^{l+1}}\right)^T}{\sqrt{d^l}}\right)} \tag{7}$$

where d^l represents the dimension of each capsule in the lth layer, and $U^l_{n^{l+1}}$ comes from the calculation of Eq. (6).

ViT module

Subsequently, we encode the spatial location information coordinates(x, y) of spots into the feature matrices of $E_x \in R^{1\times d}$ and $E_y \in R^{1\times d}$, respectively, and fuse them with the visual features of the pathology images, $S^{l+1} \in R^{n\times d}$, and use ViT to capture the long-range dependencies in the images:

$$S^{l+2} = \mathbf{concat}\left(S^{l+1}, E_x, E_y\right) \tag{8}$$

where $S^{l+1} \in R^{n\times d}$ is the output of the capsule network, n and d are the number and dimension of the output capsule, respectively; and then the attention of each spot is learned through the multi-head attention layer of the Transformer, which assigns different weights to different image patches:

$$\mathrm{Attention}\left(Q, K, V\right) = \mathbf{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{9}$$

$$\mathbf{head}_i = \mathrm{Attention}\left(XW^Q_i, XW^K_i, XW^v_i\right) \tag{10}$$

$$H(X) = \mathbf{concat}\left(\mathbf{head}_1, \mathbf{head}_2, \ldots, \mathbf{head}_m\right) \times W^h \tag{11}$$

where Q, K, V represent 'Query', 'Key' and 'Value', respectively; W^Q_i、W^K_i and W^v_i are the weight matrices for transforming inputs into Q, K, V; The term $\frac{QK^T}{\sqrt{d_k}}$ represents the similarity between Query, Key and the Attention coefficient can be obtained by weighting and summing the values of each V by the weight coefficients calculated by Softmax. W^h is the weight matrix of multiple Attention heads, and m is the number of Attention heads.

Graph attention network module

To explore the correlation between gene expression and spatial location, THItoGene uses GAT to explicitly learn the interactions between spatially located neighboring spots. GAT could pass information in a specific network structure, learn different attention values based on the importance of each neighbor of the node, and then use these weights to aggregate the features of the neighbors.

We refer to His2ST [10] for the nearest neighbor graph construction. The spatial locations of the Euclidean distances are used as edges in the nearest neighbor graph by calculating the Euclidean distances. Specifically, we select four nearest neighbors for each spot to construct the nearest neighbor graph $G = (V, E)$. $N = |V|$ denotes the number of spots, and E denotes the edges connected to the nearest neighbors. The Euclidean distance between every two spots (i, j) is calculated as follows:

$$d\left(i, j\right) = \sqrt{\left(i_x - j_x\right)^2 + \left(i_y - j_y\right)^2} \tag{12}$$

We then use GAT to aggregate the node features in the network. Here, the importance of the jth node with respect to the ith node is calculated as follows:

$$e_{ij} = f\left(W_h h_i, W_h h_j\right) \tag{13}$$

where f is a single-layer feed-forward neural network, W_h is a learnable parameter matrix which transforms the input features into a hierarchical feature representation between spots, and $h_i \in R^M$ is the feature representation of the ith node. Here, THItoGene unfolds the features output from ViT into a one-dimensional vector of length M. In order to regulate the influence between different nodes, the importance weights are normalized among all neighboring nodes of v_i:

$$\alpha_{ij} = \frac{\mathrm{exp}\left(\mathrm{LeakyRelu}\left(e_{ij}\right)\right)}{\sum_{t\in\Omega\left(v_i\right)} \mathrm{exp}\left(\mathrm{LaekyReLU}\left(e_{it}\right)\right)} \tag{14}$$

where $\Omega\left(v_i\right)$ represents the neighboring nodes of v_i and the normalized attention coefficients are used to compute a linear combination of the features of the neighboring nodes:

$$h_{\Omega(v_i)} = \sum_{t\in\Omega(v_i)} \alpha_{it} h_t \tag{15}$$

Finally, we connect the current node feature representation with the feature representations of its neighboring nodes to update the embedding:

$$h = \mathbf{Elu}\left(\mathbf{concat}\left(h_*, h_{\Omega(*)}\right)\ \omega + \beta\right) \tag{16}$$

where ω and β are the parameterized weight and bias matrices, respectively.

Evaluated metrics and criteria

In this study, we employ the Pearson correlation coefficient (PCC) to compare the spatial gene expression predicted by THItoGene with the observed gene expression, in order to assess their level of correlation. The PCC has a range of values between −1 and 1.

It is calculated by dividing the covariance of the two variables by the product of their respective SDs:

$$\mathrm{PCC} = \frac{\mathbf{Cov}\left(X_{\mathrm{true}}, X_{\mathrm{pred}}\right)}{\sigma\left(X_{\mathrm{true}}\right) \cdot \sigma\left(X_{\mathrm{pred}}\right)} \tag{17}$$

where **Cov** denotes covariance; $X_{\mathbf{true}}$ and $X_{\mathbf{pred}}$ denote the original gene expression and the gene expression obtained by prediction, respectively. σ () denotes SD.

When evaluating the performance of spatial clustering, we employ the Adjusted Rand Index (ARI) to measure the similarity between the clustering results and the true pathological annotation regions. It has a range of values between −1 and 1, where a value closer to 1 indicates a higher consistency between the clustering results and the true labels. Conversely, a value closer to 0 suggests that the similarity between the clustering results and the true labels is random, while a value closer to −1 indicates a complete opposition between the clustering results and the true labels. The ARI can be expressed as follows:

$$\mathrm{ARI} = \frac{\Sigma_{ij}\binom{n_{ij}}{2} - \left[\Sigma_i\binom{n_i}{2}\Sigma_j\binom{n_j}{2}\right] / \binom{n}{2}}{\frac{1}{2}\left[\Sigma_i\binom{n_i}{2} + \Sigma_j\binom{n_j}{2}\right] - \left[\Sigma_i\binom{n_i}{2}\Sigma_j\binom{n_j}{2}\right] / \binom{n}{2}} \tag{18}$$

where n_i and n_j denote the number of samples appearing in the ith prediction cluster and the jth true cluster, respectively; n_{ij} denotes the number of overlapping samples between the ith prediction cluster and the jth true cluster.

RESULT

THItoGene can improve the prediction accuracy of spatial resolution gene expression

In order to quantitatively evaluate the gene expression prediction performance of THItoGene, we first applied it to the HER2+ and cSCC datasets using leave-one-out cross-validation. We compared THItoGene to two other recently developed state-of-the-art methods for predicting spatial gene expression, HisToGene and His2ST. The comparison was based on calculating the PCC between predicted and observed gene expression in each tissue section. This allowed us to evaluate how accurately each method could predict actual spatial gene expression patterns. By comparing the PCCs, we aimed to determine which of the three methods provided the most accurate spatial gene expression predictions overall. Since HisToGene [22] does not give the models, we retrained HisToGene based on the parameters provided by the authors. For each section in the HER2+ and cSCC datasets, we utilized the remaining sections to train THItoGene and evaluated the correlation between predicted gene expression and actual gene expression on the current section. As shown in Figure 2A, among all the sections in the HER2+ dataset, THItoGene predicts the highest PCC between spatially resolved gene expression and actual gene expression. Notably, HisToGene has 11 sections (A2–A6, E1–F3) among the 32 sections in the HER2+ dataset with average correlations floating around the near-zero level. Among them, the average correlation of E3 and F1 sections was negative, which might be caused by its failure to capture the association relationship between deep histologic features and spot gene expression.

THItoGene exhibited the highest PCC across all tissue sections in the cSCC dataset (as shown in Figure 2B). Notably, it showed the greatest improvement over existing methods in predicting gene expression correlation for section P10_ST_rep2. These results provide strong evidence that THItoGene can more accurately forecast the relationship between spatial and actual gene expression. By incorporating adaptive super-resolution image signatures, THItoGene effectively reveals intricate cellular landscapes in histology images, enabling more reliable and precise gene expression prediction. In summary, THItoGene's use of enhanced image features leads to considerable benefits in predicting spatial gene expression compared to current methods.

Investigating the impact of each module in THItoGene on the predicted gene expression results

In order to comprehend the reasons behind the enhanced performance of THItoGene compared to other methods, we evaluated the contribution of each module to THItoGene. To achieve this, we removed key modules of THItoGene for the ablation experiment, as shown in Figure 3. Keeping all modules intact leads to the highest correlation between predicted and observed gene expression. Notably, excluding the Efficient-CapsNet module caused average PCC reductions of 0.095 and 0.076 for the HER2+ and cSCC datasets, only slightly higher than HisToGene's performance. This suggests the Efficient-CapsNet module plays a key role in enabling THItoGene to uncover the relationship between histopathology and gene expression. Furthermore, removing the ViT, dynamic convolution or GAT modules also decreased performance to varying degrees. In summary, the results of the ablation experiment demonstrate the necessity of preserving the integrity of all modules in THItoGene to ensure optimal performance in predicting gene expression.

THItoGene accurately predicts tumor-related genes

We investigated whether THItoGene's predicted gene expression reflects the true status of tumor-associated genes. In the HER2+ dataset, we analyzed the correlation between observed and predicted gene expression, calculating both correlation coefficients and *P*-values for each spot. We then derived the average −log10 *P*-value across all genes. These genes were ranked in descending order of their −log10 *P*-values, as detailed in Supplementary Table 1. We focused on the top four genes with the highest −log10 *P*-values (FN1, SCD, IGKC and FASN), which are visualized in Figure 4. THItoGene yielded correlation coefficients for these genes of 0.747, 0.711, 0.672 and 0.452, respectively—outperforming existing methods. Notably, these top-ranked genes have established associations with breast cancer, with studies indicating that the expression patterns of FN1 and IGKC correlate with patient survival and clinical outcomes [27, 28].

To demonstrate the scalability of THItoGene, we conducted same experiments on the cSCC dataset (Supplementary Table 2 and Supplementary Figure 1). THItoGene enables accurate prediction of key genes associated with cutaneous squamous cell carcinoma. This is in line with findings by Nindl *et al.* [29] and Zhu *et al.* [30] have both identified increased expression levels of NDRG1 and PI3 in cutaneous squamous cell carcinoma compared to normal skin.

THItoGene effectively analyzes the changes in gene expression within the spatial region

The spatial gene expression predicted by THItoGene is capable of accurately restoring specific spatial domains. We performed quantitative analysis to evaluate the spatial clustering performance of gene expression predicted by THItoGene, using a set of six pathologist-annotated tissue sections (B1, C1, D1, E1, F1 and

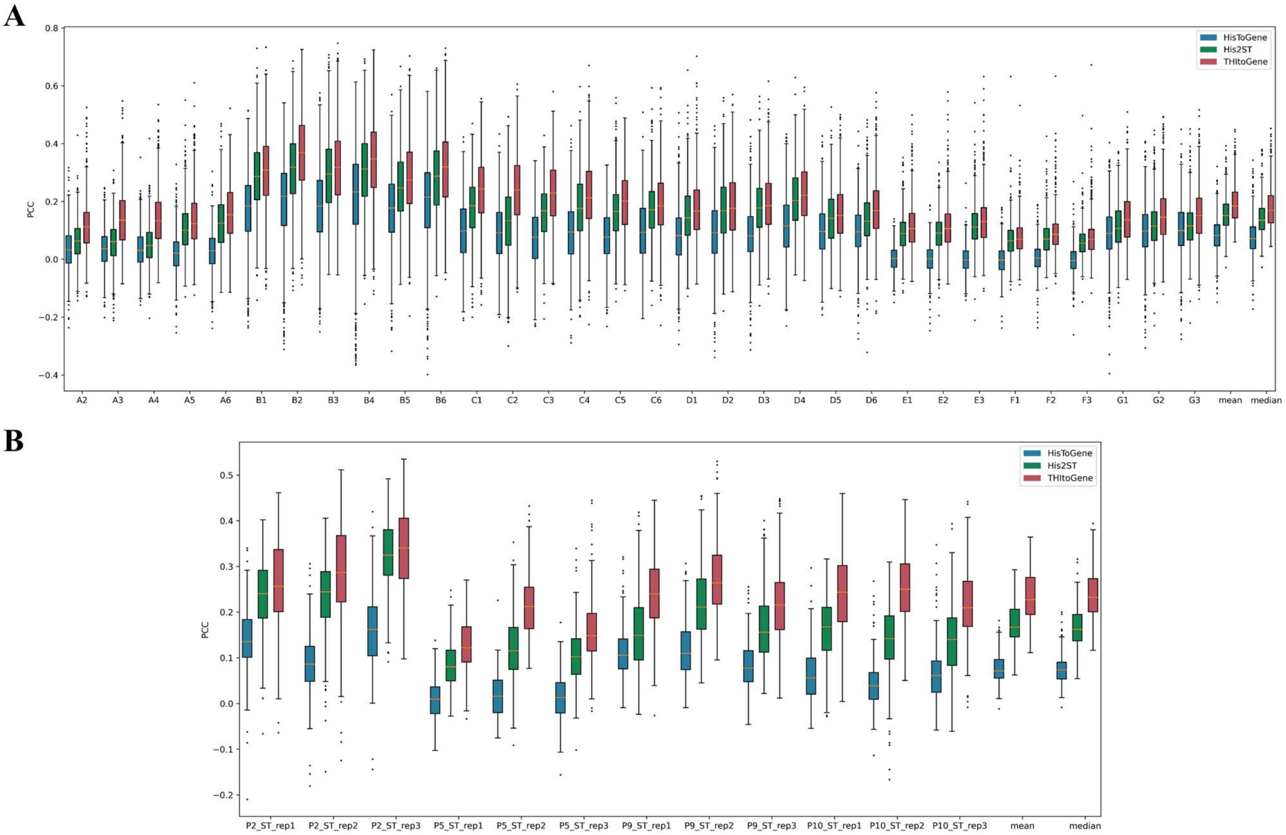

Figure 2. Comparative evaluation of spatial transcriptome prediction methods. (**A**) Boxplot representation of PCC distributions comparing observed and predicted gene expressions in the HER2+ dataset. (**B**) Similar PCC distribution comparisons for the cSCC dataset. The predictions are made by three different methods: THItoGene, His2ST and HisToGene.

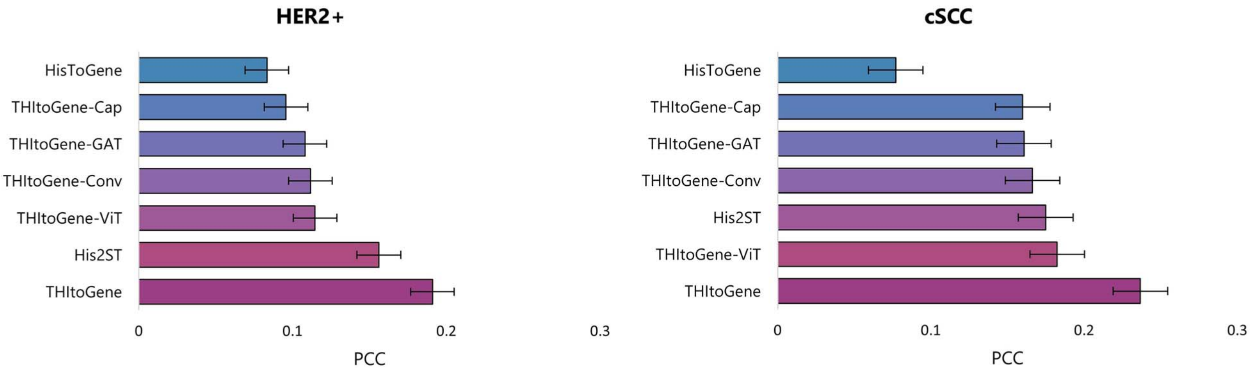

Figure 3. Ablation experiments on the HER2+ and cSCC datasets. The data showed the average PCC for THItoGene (after removing specific modules), HisToGene and His2ST in each dataset.

G2) from the HER2+ dataset. We used the spatial transcriptomics annotations generated by pathologists as the gold standard to assess the gene expression predicted by THItoGene, HisToGene and His2ST using K-means clustering (Figure 5). Compared to other methods, THItoGene is able to effectively identify the predefined spatial structures and achieve significant improvements. Specifically, in the B1 section, THItoGene (ARI = 0.327) achieved a similar ARI to HisToGene (ARI = 0.329) and significantly outperformed His2ST (ARI = 0.250). Across other sections, THItoGene consistently reconstructs spatial-specific gene expression more clearly. Interestingly, in all six sections, the clustering results obtained from the predicted gene expression by THItoGene outperformed those obtained from the observed gene expression. This discrepancy may stem from the inherent variability and noise in observed gene expression, affected by technical factors such as sequencing depth, batch effects and background noise, which can compromise clustering accuracy.

Compared to existing methods, THItoGene has the ability to capture the correlation between subtle differences in histopathological phenotypes and gene expression through dynamic convolutional adaptive learning. This allows for a better understanding of the spatial gene regulation patterns reflected by different histopathological features. Therefore, the gene expression data generated by THItoGene is superior in spatial domain identification and better reflects the true biological characteristics of tissues.

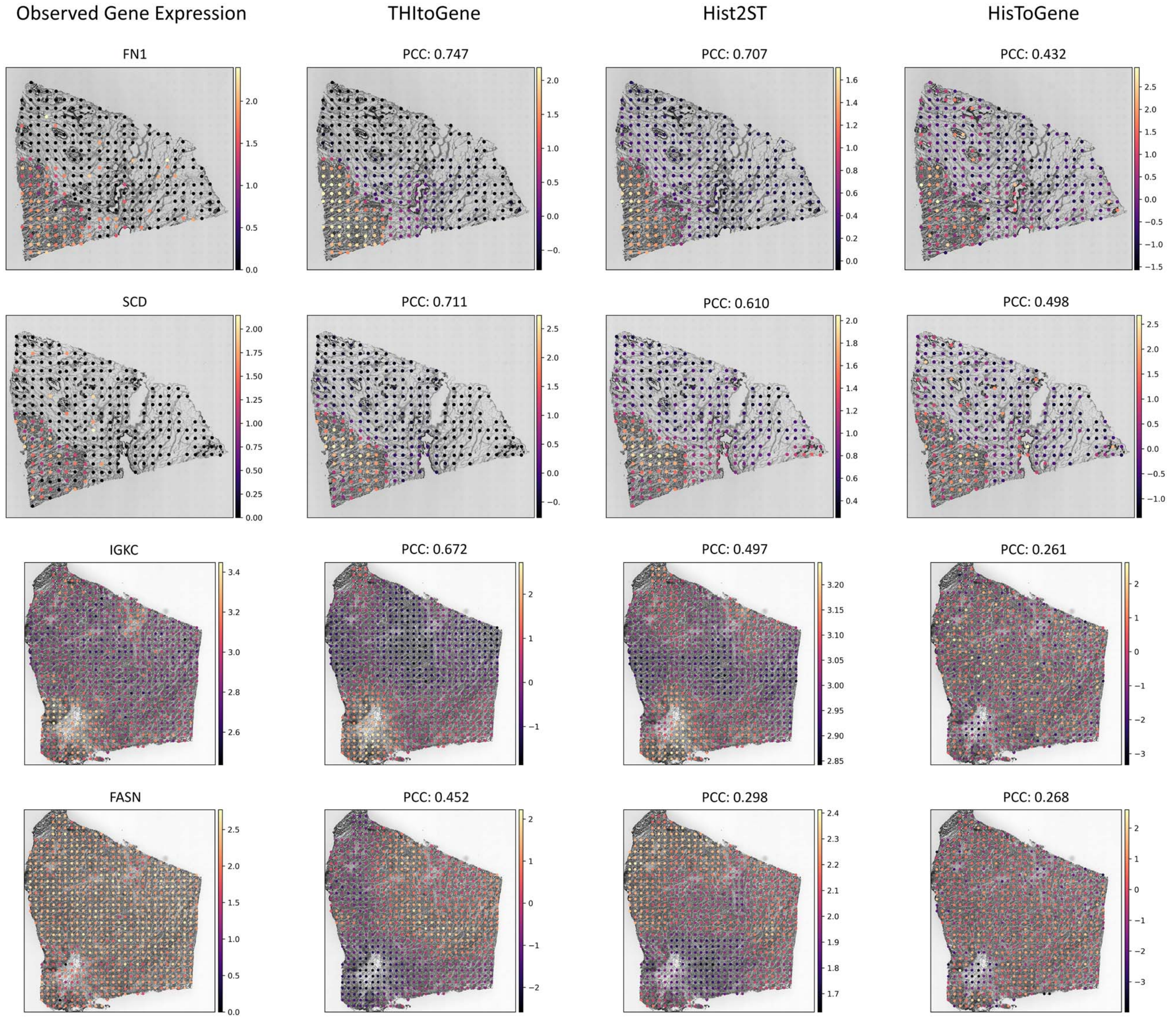

Figure 4. The top four predicted genes with the highest mean −log10 *P*-value in the HER2+ dataset. The first column displays the observed gene expression, while the last three columns show gene expression predictions from three different methods. We derived *P*-values for each gene by calculating the PCC between predicted and observed gene expressions. For visualization, we selected the section that exhibited the highest −log10 *P*-value for each gene.

THItoGene can predict differentially enriched molecular signals in complex tissue structures

We tested the ability of THItoGene to accurately predict context-specific signals within the complex structures of tissues. Specifically, we applied THItoGene to a spatial transcriptomics (ST) dataset of the human dorsolateral prefrontal cortex (DLPFC) obtained using 10x Visium technology [31]. This dataset included 12 sections from three different DLPFC samples. The DLPFC is functionally stratified into six layers, each with distinct neuronal compositions and roles in cognition, emotion regulation and decision-making [32]. The layered gene expression within the cortex is pivotal for the brain's complex architecture. In our analysis, we assessed the accuracy of THItoGene in predicting the expression of genes enriched in specific layers by comparing it with His2ST and HisToGene, as shown in Figure 6 and detailed in Supplementary Table 3. THItoGene's predictions more closely mirrored the actual gene expression patterns associated with different cortical layers. For instance, HPCAL1, recently identified by Kristen *et al.* [31] as a layer-specific marker gene in DLPFC's layer 2, was precisely predicted by THItoGene in the corresponding brain regions.

DISCUSSION

Spatial transcriptomics is an emerging technique that holds great promise for elucidating tissue development, understanding cell fate, advancing disease research, and facilitating drug discovery [33, 34]. Despite its promise, the high cost of spatial transcriptomics techniques limits their widespread application [35]. Computational predictions of spatial gene expression from histological images have emerged as a cost-effective alternative, yet existing methods often fail to capture the full complexity of molecular signatures. In this paper, we introduce THItoGene, a deep learning framework tailored for decoding spatial gene expression from pathology images. THItoGene utilizes histological images as input and employs dynamic convolutional and capsule networks to

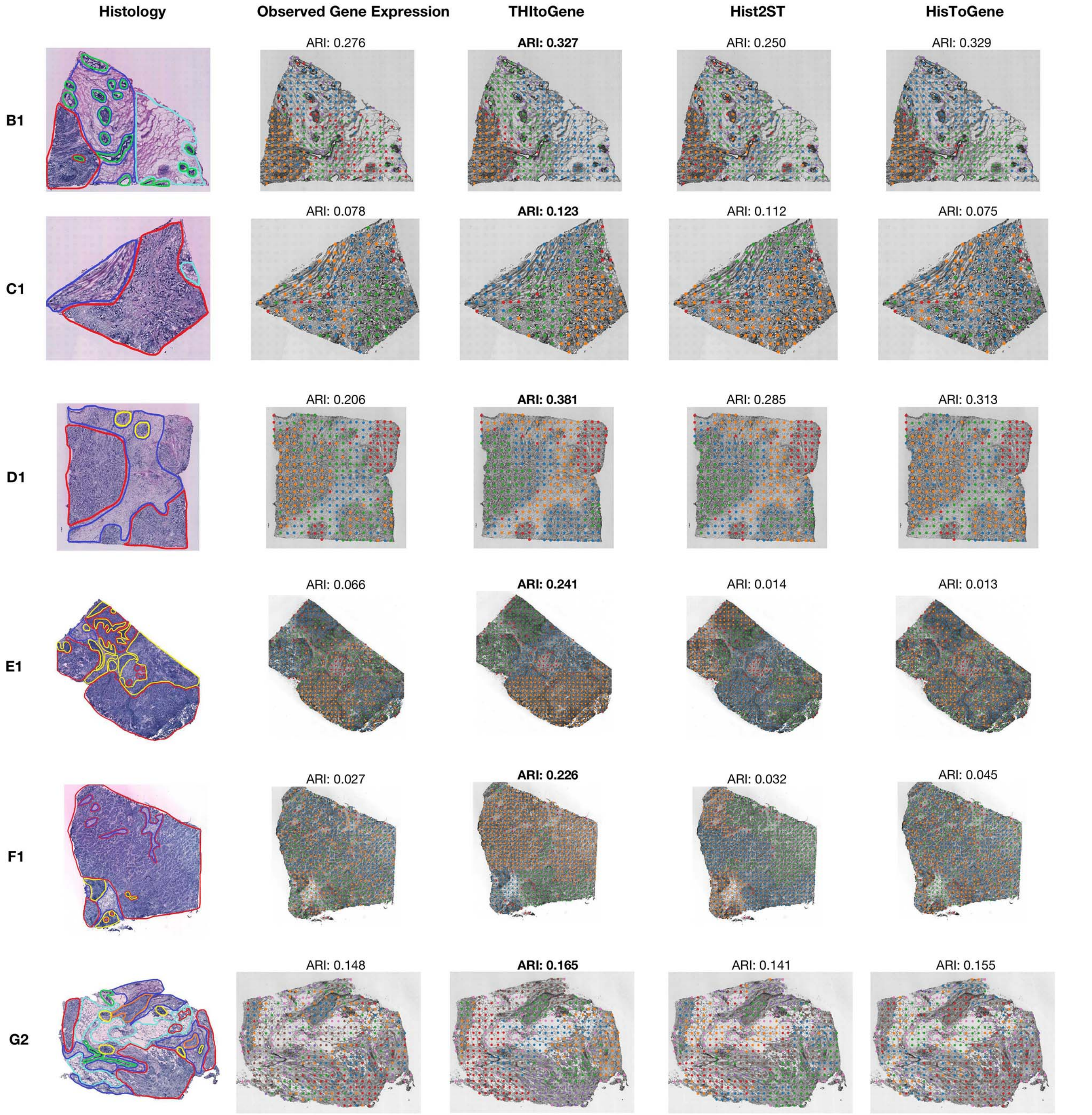

Figure 5. The spatial clustering analysis compares gene expression predictions from three different methods: THItoGene, Hist2ST and HisTGene. Histological images annotated by the pathologist are displayed in the first column. The second column presents the observed gene expression clustering results, while the last three columns show the clustering outcomes for gene expression as predicted by different methods.

capture signals of potential molecular features within histological samples. In addition, it integrates spatial location information by utilizing ViT and GAT to establish a deep connection between spatial location and gene expression, thereby deciphering cellular landscapes in histopathology images.

However, THItoGene has not yet achieved the capability to resolve gene expression at the cellular or subcellular level with higher resolution. In future studies, our aim is to augment the feature characterization potential of THItoGene by employing active machine learning approaches. Furthermore, with the emergence of new technologies and the availability of more data, we anticipate that future research will enable us to achieve higher resolution prediction of spatial gene expression and enhance the overall performance of THItoGene.

THItoGene has demonstrated the ability to extract valuable molecular information from high-resolution histologic images containing hundreds of thousands of cells. By analyzing H&E-stained pathology images, the framework captures tissue structure, cell morphology, and lesion details, correlating them with gene expression. This is particularly valuable for research scenarios with scarce resources or tissue samples. THItoGene can facilitate the analysis of tumorigenesis and progression, enhance

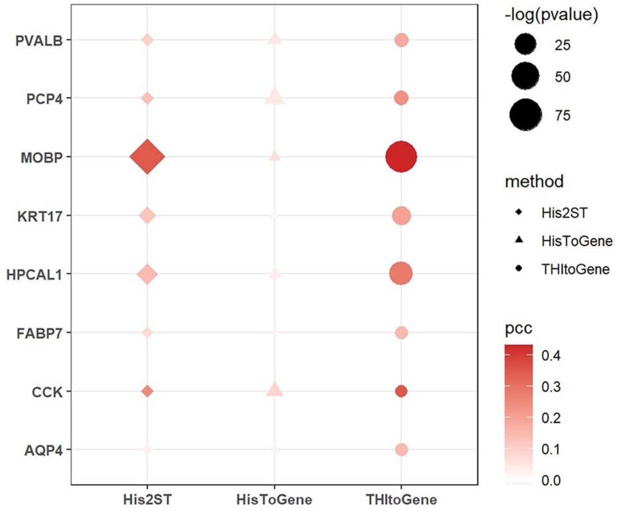

Figure 6. The results for layer marker genes predicted using three different methods within the spatial transcriptomics dataset of DLPFC.

the understanding of complex interactions and spatial effects in the tumor microenvironment, and offer a robust tool for precision medicine in resource-limited settings.

Key Points

- We introduce THItoGene, a deep learning framework designed to predict spatially resolved gene expression profiles from histopathological images.
- THItoGene outperforms other methods in accurately predicting spatial gene expression and has demonstrated effectiveness across different tumor tissue sections.
- THItoGene is able to decipher spatial domains and enrichment signals within tissue regions.

SUPPLEMENTARY DATA

Supplementary data are available online at https://academic.oup.com/bib.

ACKNOWLEDGEMENTS

We would like to thank the anonymous reviewers for their valuable suggestions.

FUNDING

Natural Science Foundation of China (62102116); Interdisciplinary Research Foundation of HIT.

AUTHORS’ CONTRIBUTIONS

Y.J., L.C. and J.L. completed the article writing. Y.J. and J.L. completed the development and experimental validation of the model and toolkit. T.Z. and Y.W. participated in discussions and oversaw the entire project. All the authors reviewed the manuscript.

DATA AVAILABILITY

The human HER2-positive breast cancer datasets are available for download at https://github.com/almaan/her2st/. The human cutaneous squamous cell carcinoma dataset can be accessed through the GEO database under accession number GSE144240. The human dorsolateral prefrontal cortex dataset can be found at http://research.libd.org/spatialLIBD/.

CODE AVAILABILITY

The source code for THItoGene is available in: https://github.com/yrjia1015/THItoGene. Trained models are available for download in: https://www.synapse.org/#!Synapse:syn52503858/files/

REFERENCES

1. Achim K, Pettit JB, Saraiva LR, *et al.* High-throughput spatial mapping of single-cell RNA-seq data to tissue of origin. *Nat Biotechnol* 2015;**33**:503–9.
2. Alon S, Goodwin DR, Sinha A, *et al.* Expansion sequencing: spatially precise in situ transcriptomics in intact biological systems. *Science* 2021;**371**:eaax2656.
3. Chen A, Liao S, Cheng MN, *et al.* Spatiotemporal transcriptomic atlas of mouse organogenesis using DNA nanoball-patterned arrays. *Cell* 2022;**185**:1777–1792.e21.
4. Chen KH, Boettiger AN, Moffitt JR, *et al.* Spatially resolved, highly multiplexed RNA profiling in single cells. *Science* 2015;**348**:6233.
5. Rodriques SG, Stickels RR, Goeva A, *et al.* Slide-seq: a scalable technology for measuring genome-wide expression at high spatial resolution. *Science* 2019;**363**:1463–7.
6. Stahl PL, Salmen F, Vickovic S, *et al.* Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science* 2016;**353**:78–82.
7. Dirkse A, Golebiewska A, Buder T, *et al.* Stem cell-associated heterogeneity in glioblastoma results from intrinsic tumor plasticity shaped by the microenvironment. *Nat Commun* 2019;**10**:1787.
8. Moncada R, Barkley D, Wagner F, *et al.* Integrating microarray-based spatial transcriptomics and single-cell RNA-seq reveals tissue architecture in pancreatic ductal adenocarcinomas. *Nat Biotechnol* 2020;**38**:333–42.
9. Wang Q, Zhi Y, Zi MX, *et al.* Spatially resolved transcriptomics technology facilitates cancer research. *Adv Sci* 2023;**10**(30):2302558.
10. Zeng YS, Wei ZY, Yu WJ, *et al.* Spatial transcriptomics prediction from histology jointly through transformer and graph neural networks. *Brief Bioinform* 2022;**23**(5):bbac297.
11. McGranahan N, Swanton C. Clonal heterogeneity and tumor evolution: past, present, and the future. *Cell* 2017;**168**:613–28.
12. Rizvi NA, Hellmann MD, Snyder A, *et al.* Mutational landscape determines sensitivity to PD-1 blockade in non-small cell lung cancer. *Science* 2015;**348**:124–8.
13. Huang Z, Bianchi F, Yuksekgonul M, *et al.* A visual–language foundation model for pathology image analysis using medical twitter. *Nat Med* 2023;**29**(9):1–10.
14. Echle A, Rindtorff NT, Brinker TJ, *et al.* Deep learning in cancer pathology: a new generation of clinical biomarkers. *Br J Cancer* 2021;**124**:686–96.
15. Fu Y, Jung AW, Torne RV, *et al.* Pan-cancer computational histopathology reveals mutations, tumor composition and prognosis. *Nature Cancer* 2020;**1**:800–10.

16. Jain MS, Massoud TF. Predicting tumour mutational burden from histopathological images using multiscale deep learning. *Nat Mach Intell* 2020;**2**:356–62.
17. Tsai PC, Lee TH, Kuo KC, *et al.* Histopathology images predict multi-omics aberrations and prognoses in colorectal cancer patients. *Nat Commun* 2023;**14**:2102.
18. Wagner SJ, Reisenbüchler D, West NP, *et al.* Transformer-based biomarker prediction from colorectal cancer histology: a large-scale multicentric study. *Cancer Cell* 2023;**41**:1650–1661.e4.
19. 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):3387.
20. Shmatko A, Laleh NG, Gerstung M, *et al.* Artificial intelligence in histopathology: enhancing cancer research and clinical oncology. *Nature Cancer* 2022;**3**:1026–38.
21. He B, Bergenstrahle L, Stenbeck L, *et al.* Integrating spatial gene expression and breast tumour morphology via deep learning. *Nature Biomedical Engineering* 2020;**4**:827–34.
22. Pang M, Su K, Li M. Leveraging information in spatial transcriptomics to predict super-resolution gene expression from histology images in tumors. bioRxiv. 2021.
23. Andersson A, Larsson L, Stenbeck L, *et al.* Spatial deconvolution of HER2-positive breast cancer delineates tumor-associated cell type interactions. *Nat Commun* 2021;**12**(1):6012.
24. Ji AL, Rubin AJ, Thrane K, *et al.* Multimodal analysis of composition and spatial architecture in human squamous cell carcinoma. *Cell* 2020;**182**:497–514.e22.
25. Li C, Zhou A, Yao A. Omni-dimensional dynamic convolution. arXiv. 2022; 2209.07947.
26. Mazzia V, Salvetti F, Chiaberge M. Efficient-CapsNet: capsule network with self-attention routing. *Sci Rep* 2021;**11**:14634.
27. Chen Y, Chen L, Hong DY, *et al.* Baicalein inhibits fibronectin-induced epithelial-mesenchymal transition by decreasing activation and upregulation of calpain-2. *Cell Death Dis* 2019;**10**(5):341.
28. Schmidt M, Hellwig B, Hammad S, *et al.* A comprehensive analysis of human gene expression profiles identifies stromal immunoglobulin κ C as a compatible prognostic marker in human solid tumors. *Clin Cancer Res* 2012;**18**:2695–703.
29. Dang C, Gottschling M, Manning K, *et al.* Identification of dysregulated genes in cutaneous squamous cell carcinoma. *Oncol Rep* 2006;**16**:513–9.
30. Wei W, Chen Y, Xu J, *et al.* Identification of biomarker for cutaneous squamous cell carcinoma using microarray data analysis. *J Cancer* 2018;**9**:400–6.
31. Maynard KR, Collado-Torres L, Weber LM, *et al.* Transcriptome-scale spatial gene expression in the human dorsolateral prefrontal cortex. *Nat Neurosci* 2021;**24**:425–36.
32. Zeng HK, Shen EH, Hohmann JG, *et al.* Large-scale cellular-resolution gene profiling in human neocortex reveals species-specific molecular signatures. *Cell* 2012;**149**:483–96.
33. Palla G, Fischer DS, Regev A, *et al.* Spatial components of molecular tissue biology. *Nat Biotechnol* 2022;**40**:308–18.
34. Tian LY, Chen F, Macosko EZ. The expanding vistas of spatial transcriptomics. *Nat Biotechnol* 2023;**41**:773–82.
35. Alexandrov T, Saez-Rodriguez J, Saka SK. Enablers and challenges of spatial omics, a melting pot of technologies. *Mol Syst Biol* 2023;**19**:e10571.