nature biotechnology

Brief Communication

https://doi.org/10.1038/s41587-023-02019-9

Inferring super-resolution tissue architecture by integrating spatial transcriptomics with histology

Received: 14 April 2023

Accepted: 4 October 2023

Published online: 2 January 2024

Check for updates

Daiwei Zhang[1] ✉, Amelia Schroeder[1], Hanying Yan[1], Haochen Yang[1], Jian Hu[2], Michelle Y. Y. Lee[1,3], Kyung S. Cho[4], Katalin Susztak[5], George X. Xu[6], Michael D. Feldman[7], Edward B. Lee[6], Emma E. Furth[6], Linghua Wang[4] & Mingyao Li[1,6] ✉

Spatial transcriptomics (ST) has demonstrated enormous potential for generating intricate molecular maps of cells within tissues. Here we present iStar, a method based on hierarchical image feature extraction that integrates ST data and high-resolution histology images to predict spatial gene expression with super-resolution. Our method enhances gene expression resolution to near-single-cell levels in ST and enables gene expression prediction in tissue sections where only histology images are available.

The rapid advancement of spatial transcriptomics (ST) technologies has made it possible to measure gene expression within the original tissue context[1–4], enabling researchers to characterize spatial gene expression patterns[5–7], study cell–cell communications[8,9] and resolve the spatiotemporal order of cellular development[10]. Despite the availability of many ST platforms, none of them provides a comprehensive solution. An ideal ST platform should offer single-cell resolution, cover the entire transcriptome, capture a large tissue area and be cost-effective. While generating such ST data with existing platforms remains challenging, computational approaches can be employed to reconstruct such data in silico.

Popular experimental methods for ST include in situ sequencing or hybridization-based technologies, such as STARmap[11], seqFISH[12–14] and MERFISH[15,16], and spatial barcoding followed by next-generation sequencing-based technologies, such as 10x Visium, SLIDE-seqV2 (ref. 17) and Stereo-seq[18]. These platforms differ in their spatial resolution and gene coverage. In situ sequencing or hybridization-based methods typically have a higher spatial resolution and sensitivity but relatively lower multiplexity for genes, whereas sequencing-based methods cover the entire transcriptome but have a lower spatial resolution, which limits their ability in studying detailed gene expression patterns.

Previous studies have shown that gene expression patterns are correlated with histological image features, suggesting the possibility of predicting gene expression from histology[19–21]. However, these existing methods do not fully utilize the rich cellular information provided by high-resolution histology images. In practice, a pathologist examines a histology image hierarchically. In this process, the first step is to identify a region of interest through the examination of high-level image features that capture the global tissue structure. After a region of interest is identified, low-level image features that reflect the local cellular structure of the tissue are examined. In this Brief Communication, to mimic this process, we propose to use a hierarchical image feature extraction approach that aims to capture both local and global tissue structures. We further develop a super-resolution gene expression prediction model by leveraging high-resolution tissue information obtained from hierarchically extracted image features. The resulting gene expression enables cell type annotation with a near-single-cell

[1]Statistical Center for Single-Cell and Spatial Genomics, Department of Biostatistics, Epidemiology and Informatics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA. [2]Department of Human Genetics, School of Medicine, Emory University, Atlanta, GA, USA. [3]Graduate Group in Genomics and Computational Biology, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA. [4]Department of Genomic Medicine, The University of Texas MD Anderson Cancer Center, Houston, TX, USA. [5]Renal, Electrolyte, and Hypertension Division, Department of Medicine, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA. [6]Department of Pathology and Laboratory Medicine, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA. [7]Department of Pathology and Laboratory Medicine, School of Medicine, Indiana University, Indianapolis, IN, USA. ✉e-mail: Daiwei.Zhang@Pennmedicine.upenn.edu; mingyao@pennmedicine.upenn.edu

resolution. We have implemented these procedures in iStar (Inferring Super-resolution Tissue ARchitecture).

An overview of iStar is shown in Fig. 1a. Our method employs a hierarchical vision transformer[22–24] (HViT) that has been pretrained on publicly available hematoxylin-and-eosin-stained histology image datasets using self-supervised learning (SSL)[25,26]. The HViT initially extracts histology features at a 16 × 16-pixel scale to capture fine-grained tissue characteristics, followed by 256 × 256-pixel scale to capture global tissue structures. Subsequently, a feed-forward neural network, trained through weakly supervised learning, uses these features to predict superpixel-level gene expressions. This model divides the gene expression measurement at a given spot for each gene into multiple values, assigning one to each superpixel, facilitated by the histology features at every superpixel. Additionally, the model predicts superpixel-level gene expressions outside the spots as well as in external tissue sections, as long as histology images are available.

To assess the accuracy of iStar in super-resolution gene expression prediction, we applied it to a simulated dataset derived from a Xenium breast cancer dataset generated by 10x Genomics[27]. The Xenium dataset comprises subcellular ST data for 313 genes, measured in two consecutively cut tissue sections from a single patient. To simulate Visium data, we binned the Xenium gene expressions based on Visium's spot size and layout. We assessed prediction accuracy for both in-sample and out-of-sample predictions. For in-sample prediction, super-resolution gene expression prediction was performed on section 2's pseudo-Visium data. For out-of-sample prediction, the pseudo-Visium data from section 1 were used as the training data, and super-resolution gene expression prediction was performed on section 2 using only its histology image as the input. We compared the prediction accuracy of iStar to that of the state-of-the-art method XFuse[28], and visually, iStar's predictions more closely match the ground truth as measured by Xenium compared to XFuse (Fig. 1b). To numerically evaluate the performance, we calculated the root mean square error (RMSE) and structural similarity index measure[29] (SSIM) between the predicted and ground truth gene expressions for each gene. iStar outperformed XFuse for virtually all genes across all resolutions (Fig. 1c, Extended Data Figs. 1–3 and Supplementary Figs. 1–5). iStar not only enhances the resolution of gene expression within the measured spots but also predicts high-resolution gene expression outside the measured spots, such as tissue gaps between spots and adjacent tissue sections where only the histology image is available. We further assessed iStar's ability to predict single-cell-level gene expression. As shown in Fig. 1d and Extended Data Figs. 4 and 5, iStar-predicted gene expression resembles that measured by Xenium.

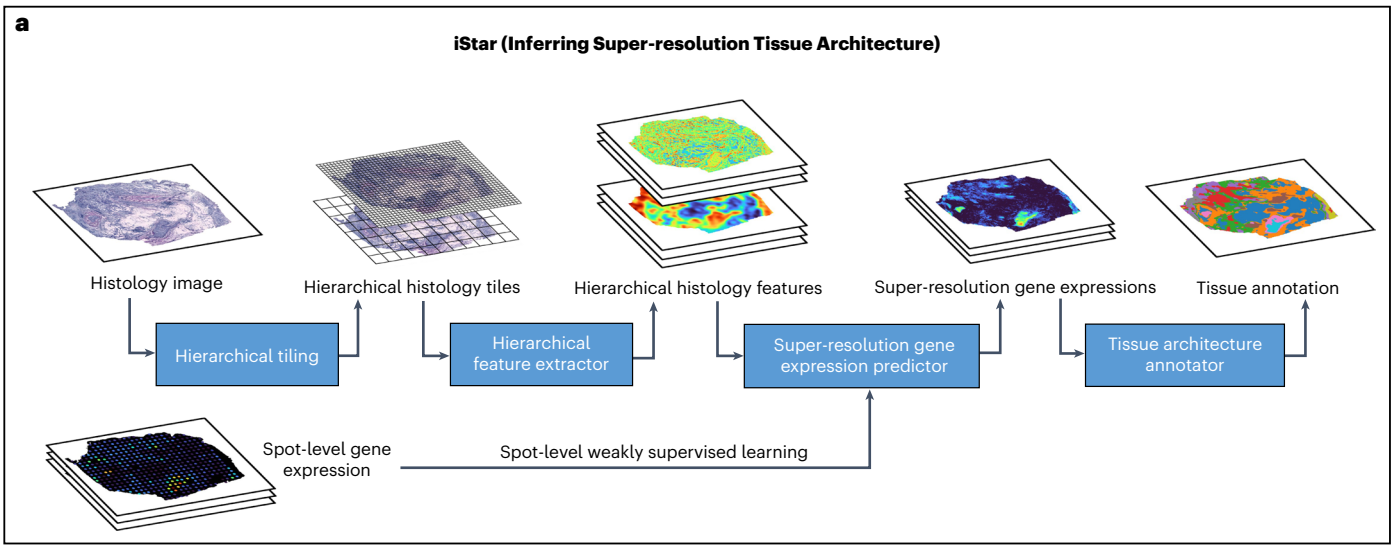
Next, we assessed iStar's capability for high-resolution annotation of tissue architecture of multiple tissue sections. In contrast to existing frameworks, which often involve challenging image registration tasks, we show that iStar can bypass the image registration step. To illustrate this capability, we used the breast cancer data in Fig. 1 as an example, assuming pseudo-Visium training data were available for section 1 but not for section 2. To perform super-resolution gene expression prediction for the two sections, we concatenated their histology images into one and treated it as a single image in downstream analyses to perform super-resolution gene expression prediction for both sections. We used the second last layer of the feed-forward neural network as features for the *k*-means algorithm[30], and the resulting segmentation highly agreed with the manual annotation and successfully separated invasive cancer (brown cluster), ductal carcinoma in situ (DCIS) #1 (gray cluster), and DCIS #2 (cyan cluster) from the rest of the tissue (Fig. 2a and Supplementary Fig. 6). By contrast, segmentation using super-resolution gene expression predicted by XFuse failed to separate DCIS #2 from invasive cancer or DCIS #2 from DCIS #1. Moreover, iStar was able to annotate tissue regions outside the spot-covered tissue area. Finally, the annotation for section 2 closely resembled that of section 1, demonstrating iStar's consistency across multiple samples.

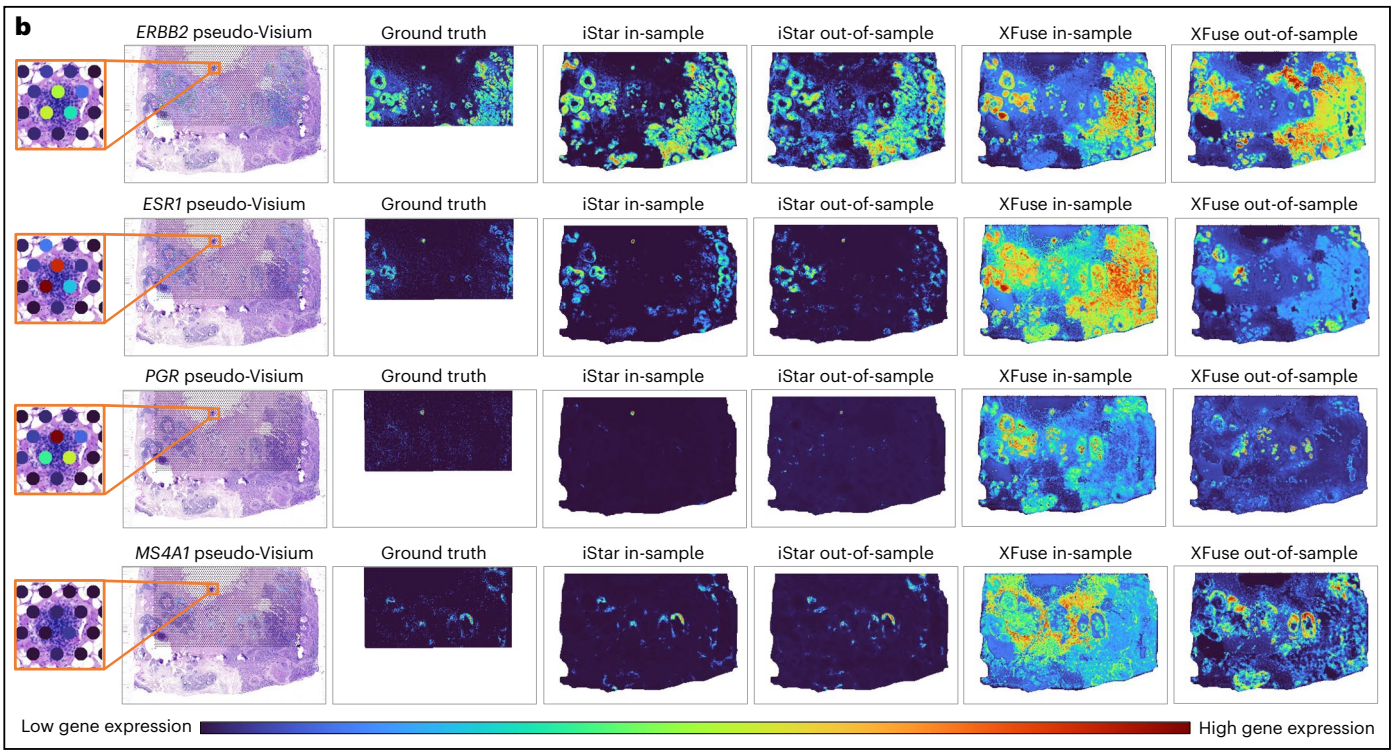
To evaluate iStar's generalizability in super-resolution tissue segmentation and annotation, we applied it to another HER2$^+$ breast cancer dataset[31] (denoted as HER2ST) generated using the legacy ST technology[32], which has a lower spatial resolution than Visium (Fig. 2b). We considered three consecutively cut tissue sections from subject H, where manual annotation was provided for only one section in the original publication. To segment all three sections, we carried out multi-sample tissue segmentation using the same approach as in Fig. 2a and found iStar showed strong agreement with the coarse manual annotation while providing increased granularity (Fig. 2b and Supplementary Figs. 7–9). Moreover, the three sections exhibited similar structures, demonstrating the consistency of iStar across samples. After segmenting the tissue, we conducted cell-type annotation at the superpixel level (Fig. 2c) and inferred cell types on the basis of predicted gene expressions of marker genes[33]. The cell type annotation yielded cell type proportion estimates within each tissue cluster, enabling the evaluation of cell type enrichment. As shown in Fig. 2c, Clusters 9 (cyan), 6 (pink), 4 (purple) and 3 (red) closely matched with the invasive and in situ cancer regions based on the manual annotation and were enriched with cancer epithelial cells. Furthermore, clusters 8 (yellow) and 5 (brown) were enriched with B cells and T cells, as expected from the manual annotation. Figure 2c and Supplementary Fig. 10 visualize the superpixels annotated as B cells, T cells, cancer epithelial cells and other cell types. The underlying biological relevance of each tissue cluster is also hinted at by the most overexpressed genes in the cluster. For example, *FABP4* (encodes fatty acid binding protein) was enriched in cluster 1 (orange), *CD8A* (a lineage marker of T cell) was enriched in cluster 5 (brown) and *MS4A1* (a lineage marker of B cell) was enriched in cluster 8 (olive) (Supplementary Fig. 11).

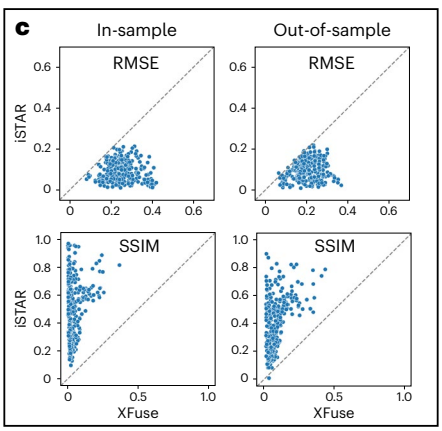
Further examination of the iStar unsupervised segmentation revealed intratumoral heterogeneity, as shown in Fig. 2d, where the refined annotation was provided by a board-certified pathologist (E.E.F.). Overall, superpixel-level cell type annotation provides biologically meaningful interpretations of the automatically detected tissue clusters, closely aligned with manual labels while revealing fine tissue

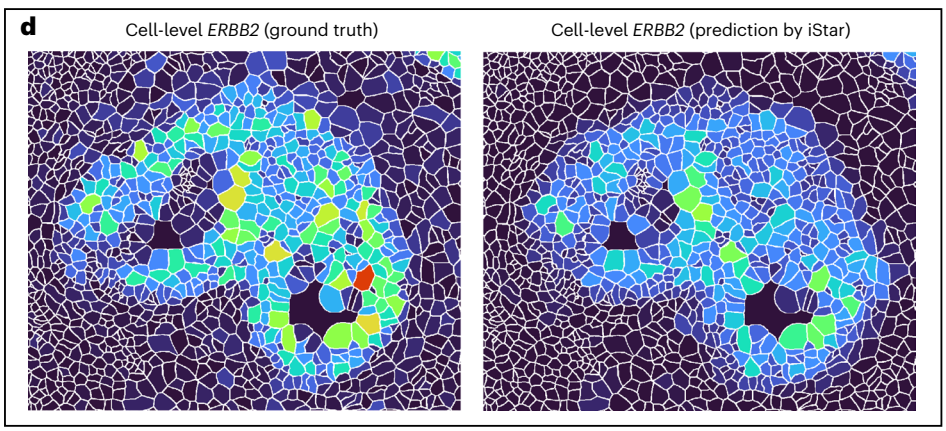
Fig. 1 | Workflow and super-resolution gene expression prediction accuracy of iStar. a, Model summary of iStar. The histology image is hierarchically divided into tiles, which are then converted into hierarchical histology image features. These features, combined with the spot-level gene expression data, are then utilized to predict super-resolution gene expression. Finally, tissue architecture is inferred on the basis of the super-resolved gene expression prediction. **b**,**c**, Evaluation of super-resolution gene expression prediction accuracy using the Xenium breast cancer dataset, which includes two consecutively cut tissue sections. The Xenium data served as the ground truth and were used to simulate spot-level gene expression based on the spot size and layout of Visium. For in-sample prediction, both model training and prediction were performed using section 2's pseudo-Visium data. For out-of-sample prediction, section 1 was used for model training and section 2 was used for prediction in which only its histology image was used as input. Visual comparison (**b**) between iStar and XFuse. *ERBB2*, *ESR1* and *PGR* are genes that encode biomarkers for breast cancer prognosis, while *MS4A1* is a B cell marker gene. Super-resolution gene expressions are visualized at the scale of 8× resolution enhancement. Visualizations of additional genes at the scale of 8× resolution enhancement are shown in Supplementary Figs. 1–5. Visualizations at other resolutions are available in Supplementary Fig. 6 and at ref. 40. Numerical comparison (**c**) between iStar and XFuse for 128× resolution enhancement of gene expression. The degree of resolution enhancement is defined as the number of superpixels in the super-resolution prediction divided by the number of spots in the training data. Each dot represents one of the 313 genes. Additional numerical evaluations are reported in Supplementary Fig. 7. **d**, Predicted single-cell-level gene expression, which was computed from the predicted superpixel-level gene expression using the cell segmentation masks provided in the dataset. Additional examples are shown in Supplementary Fig. 9.

https://doi.org/10.1038/s41587-023-02019-9

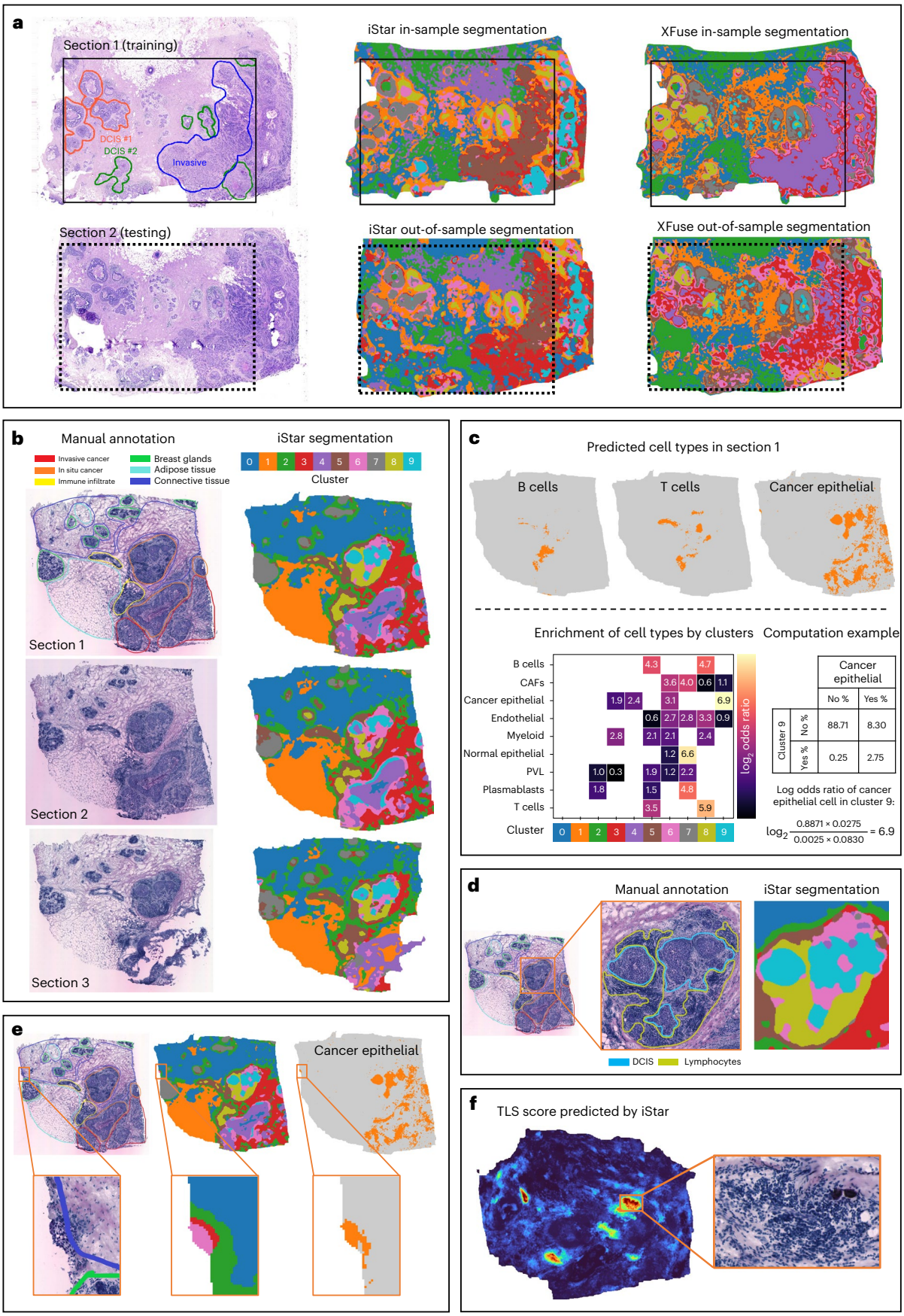








https://doi.org/10.1038/s41587-023-02019-9



structures. Notably, iStar was even able to detect a positive surgical margin (Fig. 2e) missed in the original manual annotation, and the validity of this cancer region was confirmed by E.E.F., demonstrating that iStar can identify small regions of interest that are neglected during the initial manual annotation. Our findings in this dataset suggest that iStar can accurately annotate tissue architecture even for the legacy ST

Fig. 2 | Tissue annotation using iStar. a, Comparison of unsupervised tissue segmentation by iStar and XFuse with manual annotation of one of the two consecutively cut tissue sections of a breast cancer patient in the Xenium dataset. The model was trained using the pseudo-Visium spot-level gene expression simulated from section 1 (in-sample) of the Xenium data. Section 2 was treated as the out-of-sample section, and its super-resolution gene expression was predicted using only its histology image. Super-resolution was performed with 128× resolution enhancement. **b**, Tissue architecture annotation of a breast cancer tissue in the HER2ST breast cancer dataset (three consecutively cut tissue sections in subject H). Super-resolution was performed with 128× resolution enhancement. **c**, iStar assigned biologically meaningful labels to the tissue clusters by performing superpixel-level cell type inference, followed by a cell type enrichment analysis, where depletion, that is, negative enrichment, was not shown in the heatmap. CAFs, cancer-associated fibroblasts; PVL, perivascular-like. **d**, iStar's unsupervised tissue segmentation revealed intratumoral heterogeneity that agreed with the pathologist's manual annotation. **e**, A small cancer region detected by iStar that was missed in manual annotation provided in the original publication. **f**, Detection of TLSs in the HER2ST breast cancer dataset (subject G) by iStar. The TLS score was calculated as the mean of the standardized super-resolution gene expressions of the TLS marker genes presented in Supplementary Table 1.

platform. The identification of biologically relevant genes within tissue clusters further supports the potential utility of iStar in uncovering new insights into tissue biology and diseases.

Next, we show that iStar can be utilized to detect multicellular structures, such as tertiary lymphoid structures (TLSs), which are clusters of highly organized immune cells formed in non-lymphoid tissues, often found at sites of inflammation, including a variety of solid tumors[34,35]. The presence of TLSs has been shown to be associated with positive clinical outcomes and responses to immunotherapy[36–38]. However, the manual detection of TLSs using the spot-resolution Visium data is labor-intensive and imprecise, due to the small size and the fine-grained characteristics of TLSs. To demonstrate the ability of iStar for automatic TLS detection, we analyzed three consecutively cut tissue sections of another patient (subject G) in the HER2ST dataset. To detect TLSs, we curated a list of unique TLS marker genes (Supplementary Table 1) and computed TLS gene signature scores by standardizing and averaging the predicted gene expression of the TLS marker genes (Fig. 2f). We identified multiple TLSs, all of which were confirmed by a board-certified pathologist (E.E.F.), and the TLS marker gene expressions are shown in Supplementary Fig. 12. By contrast, the original HER2ST study[31] detected several TLSs, but the analyses were based on low-resolution spot-level gene expression, resulting in a much lower resolution compared to our results (Extended Data Fig. 6).

In addition to the two breast cancer datasets analyzed above, we also analyzed one additional breast cancer dataset generated using Visium by 10x Genomics. As shown in Supplementary Fig. 13, iStar revealed fine-grained tissue structures. Although we have primarily focused on the applications to breast cancer in this study, iStar is a generic tool that can be applied to various diseased or healthy tissue types. To demonstrate iStar's capability in analyzing healthy tissues, we conducted benchmarking evaluations using a Xenium dataset generated from mouse brain by 10x Genomics. The benchmarking was designed similarly to the experiments for the Xenium-derived pseudo-Visium breast cancer dataset. As shown in Extended Data Figs. 7 and 8a and Supplementary Figs. 14–17, iStar achieved high accuracy in this evaluation across all resolutions and outperformed XFuse. In addition, our super-resolution gene expression-based segmentation (Extended Data Fig. 8d), compared with that by XFuse (Extended Data Fig. 8e), revealed fine-grained tissue structures that match closely with the Allen Brain Atlas annotation (Extended Data Fig. 8b,c).

Finally, to demonstrate iStar's broad applicability to diverse cancer and healthy tissue types, we applied it to additional Visium datasets generated from mouse brain (Extended Data Fig. 9), mouse kidney (Extended Data Fig. 10a), prostate cancer (Extended Data Fig. 10b and Supplementary Fig. 18), colorectal cancer (Extended Data Fig. 10c) and kidney cancer (Extended Data Fig. 10d). In all applications, iStar was able to characterize tissue architecture with high resolution. For example, iStar accurately detected TLSs that aligned well with pathologist's manual annotation in kidney cancer (Extended Data Fig. 10d).

In summary, we have presented iStar, a method for rapid annotation of super-resolution tissue architecture based on ST data generated from platforms that lack single-cell resolution. This holds implications for practical studies, as existing ST platforms lack either single-cell resolution or whole-transcriptome coverage. However, iStar allows us to generate ST data that cover the entire transcriptome with near-single-cell resolution (Supplementary Fig. 19). A key step of iStar is to leverage the high-resolution histology image obtained from the same ST tissue section to reconstruct the unobserved super-resolution gene expression. Through the analysis of several datasets across multiple cancer types and healthy tissues, we have demonstrated that the super-resolution gene expressions predicted by iStar are accurate. These predictions not only preserve the original gene expression at the spot level (Supplementary Figs. 20 and 21) but also have practical applications in various tissue architecture inference tasks. Moreover, we have shown that iStar can perform out-of-sample prediction for tissue sections where only histology images are available. iStar is computationally efficient, with the end-to-end analysis of the Xenium-derived pseudo-Visium breast cancer data taking only 9 min (Supplementary Table 2). By contrast, XFuse took 1,969 min to analyze the same data, which was 218 times slower. This advantage in computational efficiency allows iStar to generate virtual ST data from a large number of consecutively cut tissue sections with histology images, enabling a comprehensive characterization of gene expression variations in 3D tissues[39].

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41587-023-02019-9.

References

1. Burgess, D. J. Spatial transcriptomics coming of age. *Nat. Rev. Genet.* **20**, 317 (2019).
2. Asp, M., Bergenstrahle, J. & Lundeberg, J. Spatially resolved transcriptomes-next generation tools for tissue exploration. *Bioessays* **42**, e1900221 (2020).
3. Crosetto, N., Bienko, M. & van Oudenaarden, A. Spatially resolved transcriptomics and beyond. *Nat. Rev. Genet.* **16**, 57–66 (2015).
4. Moor, A. E. & Itzkovitz, S. Spatial transcriptomics: paving the way for tissue-level systems biology. *Curr. Opin. Biotechnol.* **46**, 126–133 (2017).
5. Hu, J. et al. SpaGCN: integrating gene expression, spatial location and histology to identify spatial domains and spatially variable genes by graph convolutional network. *Nat. Methods* **18**, 1342–1351 (2021).
6. Sun, S., Zhu, J. & Zhou, X. Statistical analysis of spatial expression patterns for spatially resolved transcriptomic studies. *Nat. Methods* **17**, 193–200 (2020).
7. Svensson, V., Teichmann, S. A. & Stegle, O. SpatialDE: identification of spatially variable genes. *Nat. Methods* **15**, 343–346 (2018).

Methods

The algorithm of iStar

The algorithm of iStar consists of three components: the histology feature extractor, super-resolution gene expression predictor and tissue architecture annotator.

Histology feature extractor

To facilitate the processing of histology images with different resolutions, we first rescale each image such that the size of one pixel is $0.5 \times 0.5\ \mu m^2$. This rescaling ensures a 16×16-pixel tile corresponds to $8 \times 8\ \mu m^2$, which is about the size of a single cell. To simplify the subsequent tiling procedure, we pad the rescaled image so that its height and width are both divisible by 256.

Next, we partition the whole image into image tiles hierarchically such that the large (high-level) tiles reflect the global tissue structure, whereas the small (low-level) tiles within a large tile reflect the local fine-grained cellular structure of the tissue. Let $X \in \mathbb{R}^M \times \mathbb{R}^N \times \mathbb{R}^3$ be the RGB-channel histology image with height M and width N. We first partition X into a $(M/256)$-row, $(N/256)$-column rectangular grid of 256×256-pixel image tiles: $X = [X_{m_1 n_1}]_{m_1=1,n_1=1}^{M/256,N/256}$, where each $X_{m_1 n_1} \in \mathbb{R}^{256} \times \mathbb{R}^{256} \times \mathbb{R}^3$. Next, each 256×256-pixel image tile is further partitioned into a 16-row, 16-column rectangular grid of 16×16-pixel image tiles: $X_{m_1 n_1} = [X_{m_1 n_1 m_2 n_2}]_{m_2=1,n_2=1}^{16,16}$, where each $X_{m_1 n_1 m_2 n_2} \in \mathbb{R}^{16} \times \mathbb{R}^{16} \times \mathbb{R}^3$.

To extract hierarchical histology features, we use an HViT architecture[22–24] that consists of a local vision transformer (ViT) $f_1 \circ f_2$ and a global ViT f_0. First, within each 256×256-pixel image tile, the local ViT maps each 16×16-pixel subtile into a low-level local feature vector of length C_2, that is,

$$z_{m_1 n_1 m_2 n_2} = f_2(X_{m_1 n_1 m_2 n_2}) \in \mathbb{R}^{C_2},$$

and then maps all the 256 low-level local feature vectors within the 256×256-pixel image tile into a high-level local feature vector of length C_1, that is,

$$z_{m_1 n_1} = f_1\left([z_{m_1 n_1 m_2 n_2}]_{m_2=1,n_2=1}^{16,16}\right) \in \mathbb{R}^{C_1}.$$

Next, to model long-range dependencies of histology features within the whole image, the global ViT maps all the high-level local features within the whole image into high-level global features of the same dimension:

$$[t_{m_1,n_1}]_{m_1=1,n_1=1}^{M/256,N/256} = f_0\left([z_{m_1 n_1}]_{m_1=1,n_1=1}^{M/256,N/256}\right) \in \mathbb{R}^{M/256} \times \mathbb{R}^{N/256} \times \mathbb{R}^{C_1}.$$

After this hierarchical histology feature extraction procedure, we have

1. the high-level global feature image $T = [t_{m_1,n_1}]_{m_1=1,n_1=1}^{M/256,N/256}$, which is an image of size $(M/256) \times (N/256)$ with C_1 channels,
2. the low-level local feature image $Z = [z_{m_1 n_1 m_2 n_2}]_{m_1=1,n_1=1,m_2=1,n_2=1}^{M/256,N/256,16,16}$, which is an image of size $(M/16) \times (N/16)$ with C_2 channels, and
3. the original RGB image, which is an image of size $M \times N$ with three channels.

To align these feature images, we use bicubic interpolation to resize each image into the desired size $M' \times N'$ and stack the channels of the resized images, which results in a combined histology feature image $H = [h_{mn}]_{m=1,n=1}^{M',N'}$ of size $M' \times N'$ with $C_1 + C_2 + 3$ channels, where each $h_{mn} \in \mathbb{R}^{C_1+C_2+3}$ is the histology feature vector at pixel (m, n). In our implementation, we set $C_1 = 192$ and $C_2 = 384$. For the image size, we varied (M', N') among $(M/16, N/16)$, $(M/32, N/32)$, $(M/64, N/64)$ and $(M/128, N/128)$.

To train the histology feature extractor, we optimize the ViTs through SSL. Because of the benefits of transfer learning on ViTs[41], the model is pretrained on publicly available histology datasets. In this step, since only histology images are needed and no gene expression data or image-level labels are required, many publicly available histology datasets, such as The Cancer Genome Atlas, the Genotype-Tissue Expression project and the kidney biopsies in Holscher et al.[42], are suitable for pretraining the model. Moreover, for the choice of the SSL algorithm, any SSL method for analyzing image data, such as DINO[25] or BEIT[26], is suitable for our purpose. In our implementation, we adopted the pretrained model in Chen et al.[22], which uses DINO to train the ViTs hierarchically on the The Cancer Genome Atlas data. In our experiments, we found that the pretrained model was able to capture the histology characteristics well and thus decided to skip the fine-tuning step to improve computation efficiency.

Super-resolution gene expression predictor

Once the histology feature images have been extracted, we use them to predict super-resolution gene expression. The histology features at every superpixel contain information on not only its local cellular characteristics but also its global relationships to other regions in the whole image. Thus the gene expression predictor does not need to explicitly model the spatial dependencies, since correlation between superpixels are already reflected in the similarity between their high-level global histology features (that is, the C_1 channels in the combined histology feature image H), even if the superpixels are physically far away from each other. Therefore, when predicting the gene expression at a superpixel, the input of the predictor only includes the histology features at this superpixel, and no convolution, attention, or any other mechanisms with spatial awareness are needed, which substantially reduces the computation cost.

To train the super-resolution gene expression predictor, since the model output is at the superpixel level but the training data are at the spot level, we adopt a weakly supervised learning framework. We model the gene expression observed at each spot as the sum of the superpixels' gene expression inside that spot. This model design mimics the data collection procedure of sequencing-based ST platforms, which barcodes and combines all the transcripts inside a spot into a sample and sends it for next-generation sequencing. To express the loss function, let S be the number of spots in the whole image, K be the number of genes to predict, g_k be the gene expression prediction model for gene k, y_{ks} be the observed gene expression for gene k at spot s, $\mathcal{M}_s$ be the spot mask of s (that is, the collection of superpixels covered by spot s), and h_{mn} be the histology feature vector at superpixel (m, n). Then the weakly supervised loss function is

$$\mathcal{L} = \sum_{k=1}^{K} \sum_{s=1}^{S} \left(y_{ks} - \sum_{(m,n) \in \mathcal{M}_s} g_k(h_{mn}) \right)^2.$$

Superpixels outside the spot masks, including the between-spot gaps and the background image, are excluded during model training. After model training, the predicted gene expression for gene k at superpixel (m,n) is $\hat{y}_{kmn} = g_k(h_{mn})$, which gives us the gene expression image $\hat{Y}_k = [\hat{y}_{kmn}]_{m=1,n=1}^{M',N'}$. Furthermore, if cell segmentation masks are provided, single cell-level gene expression can be obtained using the predicted superpixel-level gene expressions, where the former is computed as a weighted sum of the latter, with the weight equal to the proportion of the superpixel that overlaps with the cell mask[43]. In our experiments, we only predicted the union of the top 1,000 most highly variable genes in each dataset and the marker genes for the user-defined structures (for example, TLS), since lowly variables genes had low signal-to-noise ratios and would introduce extra noise to the model training procedure.

The only two exceptions are the benchmark experiments using the Xenium breast cancer dataset (313 genes) and the Xenium mouse brain dataset (248 genes), in which case we predicted all the genes due to their small number and the need for method evaluation.

For network architecture of the gene expression prediction model, we use a feed-forward neural network with 4 hidden layers and 256 nodes per hidden layer. The leaky rectified linear unit (ReLU)[44] is used as the activation function for the hidden layers. The output layer is a linear layer with 256 input nodes and K output nodes, and the outputs are activated by an exponential linear unit (ELU)[45] to ensure that the predicted gene expressions are non-negative.

Tissue architecture annotator

Once obtaining the super-resolution gene expression, we segment the tissue by clustering the superpixels using their gene expression information. First, we obtain gene expression embeddings by reducing the dimension of the predicted gene expression vector, where each superpixel is treated as a sample and each gene as a feature. Although any dimension reduction technique (for example, principal component analysis[46] or uniform manifold approximation and projection[47]) can obtain gene expression embeddings from the predicted super-resolution gene expression, we recommend treating the intermediate values in the second-last feed-forward layer of the gene expression prediction model as the gene expression embeddings, since they are not only low-dimensional (256 in our setting) but also linearly related to the predicted gene expression vectors, and using these precomputed values does not incur any additional computational cost. Next, to promote spatial contiguousness in segmentation, we smooth the gene expression embeddings by a Gaussian filter, an approach that is similar in spirit to the sliding-window method for cell neighborhood identification[48]. Then we treat the smoothed gene expression embedding vector at every superpixel as a sample and cluster all the superpixels with the k-means algorithm[30]. This procedure partitions the tissue into functionally distinct regions in an unsupervised manner based on their gene expression profiles.

To assign biologically meaningful interpretations to the tissue regions in the segmentation, we perform cell type inference at the superpixel level, where we treat each superpixel as an artificial cell and infer its cell type using its predicted gene expressions along with a marker gene reference panel. Recall that the total number of genes in the model is K. Let T be the total number of candidate cell types. For each cell type $t \in \{1, \dots, T\}$, suppose we have a list of marker gene indices $\mathcal{A}_t$, which is a subset of $\{1, \dots, K\}$. For example, in our experiments with breast cancer data, we used the marker gene lists provided by Wu et al.[33]. For each marker gene $k \in \mathcal{A}_t$, we standardize its predicted super-resolution gene expression image $\hat{Y}_k = [\hat{y}_{kmn}]_{m=1,n=1}^{M',N'} \in \mathbb{R}^{M'N'}$ into the range of $[0.0, 1.0]$ and obtain $\tilde{Y}_k = [\tilde{y}_{kmn}]_{m=1,n=1}^{M',N'} \in \mathbb{R}^{M'N'}$, where $\tilde{y}_{kmn} = (\hat{y}_{kmn} - \min \hat{Y}_k)/(\max \hat{Y}_k - \min \hat{Y}_k)$. Then for each superpixel (m, n), we compute the score for cell type t by averaging the standardized gene expressions of all its marker genes: $u_{tmn} = |\mathcal{A}_t|^{-1} \sum_{k \in \mathcal{A}_t} \tilde{y}_{kmn}$ where $|\mathcal{A}_t|$ is the number of genes in $\mathcal{A}_t$. To infer the cell type of superpixel (m, n), let $t_{mn}^{\max} = \arg\max_{1 \le t \le T} u_{tmn}$ be the cell type with the maximal score and $u_{mn}^{\max} = \max_{1 \le t \le T} u_{tmn}$ be the score of this cell type. Given a predetermined threshold $u_{\text{threshold}} \in [0, 1]$, if $u_{mn}^{\max} > u_{\text{threshold}}$, then the cell type of superpixel (m, n) is predicted to be $t_{mn}^{\max}$; otherwise, the cell type of this superpixel is unclassified. In our experiments, we set $u_{\text{threshold}} = 0.1$ and found it effective in most cases. For a demonstration of the effects of $u_{\text{threshold}}$ on cell type inference, see Supplementary Fig. 22. While this score-based approach was used for cell type inference in our experiments, any cell type annotation tool serves the purpose. For example, when a well-annotated single-cell RNA-sequencing reference panel is available, the cell types can be annotated by methods such as SingleR[49] or ItClust[50]. Finally, to combine the superpixel-level predicted cell types with the tissue clusters obtained through unsupervised segmentation, an enrichment analysis is applied to every cell type–tissue cluster pair, which elucidates the biological activities inside each cluster by examining the cell types overrepresented in the cluster.

In addition to the above-described unsupervised tissue annotation procedure, iStar also allows annotation with user-defined tissue structures. In this procedure, a score image is produced to reflect the intensity of the user-defined structure across the tissue. The computation of the user-defined structure score is similar to the computation of the cell type score described in the previous paragraph. Given a list of user-defined gene indices $\mathcal{A}$, which is a subset of $\{1, \dots, K\}$, for the structure of interest (for example, TLS; Supplementary Table 1), for every gene $k \in \mathcal{A}$, we first standardize its predicted super-resolution gene expression image $\hat{Y}_k \in \mathbb{R}^{M'N'}$ into the range of $[0.0, 1.0]$ and obtain the standardized image $\tilde{Y}_k = (\hat{Y}_k - \min \hat{Y}_k)/(\max \hat{Y}_k - \min \hat{Y}_k) \in \mathbb{R}^{M'N'}$. Then we compute the score image for the user-defined structure by averaging the standardized gene expression images of all the marker genes: $U = |\mathcal{A}|^{-1} \sum_{k \in \mathcal{A}} \tilde{Y}_k$ where $|\mathcal{A}|$ is the number of genes in $\mathcal{A}$. The resulting score image $U \in \mathbb{R}^{M'N'}$ reflects the activity of the user-defined structure in the tissue.

Benchmark data generation

To evaluate the super-resolution gene expression prediction accuracy, we generated spot-level pseudo-Visium data using pixel-level Xenium data[27]. The pixel size of the Xenium gene expression images was $0.2 \times 0.2\ \mu m^2$, and we rescaled the pixel size to $0.5 \times 0.5\ \mu m^2$. The gene expression measurements in Xenium were binned into spots on the basis of the spot size, shape and layout of Visium: a hexagonal grid of disk-shaped spots with a spot diameter of 55 µm and a center-to-center distance of 100 µm. For the ground truth, we binned the Xenium gene expressions into a rectangular grid of superpixels, where the size of the superpixels varied among $8 \times 8\ \mu m^2$, $16 \times 16\ \mu m^2$, $32 \times 32\ \mu m^2$ and $64 \times 64\ \mu m^2$, depending on the experimental settings.

Evaluation criteria for super-resolution gene expression prediction accuracy

To evaluate the accuracy of the predicted super-resolution gene expressions, for each gene, we treated both the ground truth and the predicted gene expression as images, where the image intensity was standardized into the range of $[0.0, 1.0]$. Then the prediction accuracy was measured by the RMSE and the SSIM[29]. To compute the RMSE, the ground truth and the predicted gene expression images were flattened into vectors, and the RMSE was equal to the Euclidean distance between the two vectors. The RMSE is a straightforward and fast metric for assessing the prediction accuracy of any outcomes that can be vectorized, but for image data, the RMSE ignores the spatial contexts within the images[51]. Thus, in addition to the RMSE, we also computed the SSIM to evaluate the similarity between the spatial structures of the ground truth and the predicted gene expression images. The SSIM is an image similarity metric that is widely used for super-resolution tasks in computer vision[52,53] and medical imaging[54]. A higher SSIM indicates a higher degree of similarity between two images. In our context, the SSIM captures both global trends and the fine-grained spatial structures in the super-resolution gene expression images. Our experiments showed that iStar outperformed XFuse as measured by both the RMSE and SSIM.

In addition to RMSE and SSIM, Pearson correlation coefficient (PCC), which is an uncommon metric for super-resolution tasks[55,56], was employed in some previous works[28,57] on ST as an evaluation criterion for gene expression prediction accuracy. However, these works studied ST at low resolutions, where the number of spatial units (that is, superpixels or spots) was no more than 2,000, and the size of the spatial units was around 100 µm. By contrast, the prediction

accuracy in our experiments was evaluated at a much higher resolution, where the number of superpixels was as large as 10^6 and their size was as small as 8 μm. Due to the high image resolution and high noise magnitude in the ground truth, PCC is sensitive to outlying noisy superpixels, especially for sparsely expressed genes. In our experiments, compared to RMSE and SSIM, PCC had difficulties in differentiating superior and inferior super-resolution predictions when the resolution was high. As the resolution decreased, the noise level in the ground truth also decreased, which led to sharpened contrast between the accuracy of iStar and XFuse as measured by PCC. Furthermore, more spatially variables genes, which were associated with higher signal-to-noise ratio, produced higher PCCs and greater differences in PCC between iStar and XFuse, which again indicates the sensitivity of PCC to the noise level in the ground truth. Overall, PCC had limited power in differentiating super-resolution prediction accuracy for high-resolution, high-noise gene expression images. On the other hand, when the resolution and noise level were low, PCC produced results similar to those by RMSE and SSIM (Supplementary Figs. 23 and 24).

Computational efficiency

Computational efficiency was another area where iStar outperformed XFuse. In the benchmark experiments, iStar was approximately 200 times faster than XFuse. The end-to-end analysis of a typical dataset by iStar usually finishes within 10 min, while XFuse took about a day. The detailed runtimes for training and prediction are reported in Supplementary Table 2. Experiments were conducted on an NVIDIA GeForce RTX 2080 Ti graphics card.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

We analyzed the following publicly available ST datasets: (1) 10x Xenium human breast cancer data (https://www.10xgenomics.com/products/xenium-in-situ/preview-dataset-human-breast); (2) 10x Xenium mouse brain data (https://www.10xgenomics.com/resources/datasets/fresh-frozen-mouse-brain-replicates-1-standard); (3) human HER2-positive breast cancer ST data reported in Anderson et al. (https://github.com/almaan/her2st); (4) 10x Visium human breast cancer data (https://www.10xgenomics.com/resources/datasets/human-breast-cancer-visium-fresh-frozen-whole-transcriptome-1-standard); (5) 10x Visium human colorectal cancer data (https://www.10xgenomics.com/resources/datasets/human-colorectal-cancer-whole-transcriptome-analysis-1-standard-1-2-0); (6) 10x Visium human prostate cancer data (https://www.10xgenomics.com/resources/datasets/human-prostate-cancer-adenocarcinoma-with-invasive-carcinoma-ffpe-1-standard-1-3-0); (7) human prostate cancer data reported in Erickson et al. (https://doi.org/10.17632/svw96g68dv.1); (8) human clear cell renal cell carcinoma primary tumors reported in Meylan et al. (GSE175540); (9) 10x Visium mouse kidney data (https://www.10xgenomics.com/resources/datasets/adult-mouse-kidney-ffpe-1-standard-1-3-0); (10) 10x Visium mouse brain coronal cut data (https://www.10xgenomics.com/resources/datasets/mouse-brain-coronal-section-2-ffpe-2-standard); (11) 10x Visium mouse brain sagittal cut posterior data (https://www.10xgenomics.com/resources/datasets/mouse-brain-serial-section-2-sagittal-posterior-1-standard); (12) 10x Visium mouse brain olfactory bulb data (https://www.10xgenomics.com/resources/datasets/adult-mouse-olfactory-bulb-1-standard-1). Details of the datasets analyzed in this paper are described in Supplementary Table 3. Gene expression visualizations for other spatial resolutions in the 10x Xenium breast cancer and mouse brain data are available at https://zenodo.org/doi/10.5281/zenodo.10071636. Source data are provided with this paper.

Code availability

The iStar algorithm was implemented in Python and is available on GitHub at https://github.com/daviddaiweizhang/istar.

References

41. Steiner, A. et al. How to train your ViT? Data, augmentation, and regularization in vision transformers. *Trans. Mach. Learn Res.* https://openreview.net/pdf?id=4nPswr1KcP (2022).
42. Hölscher, D. L. et al. Next-generation morphometry for pathomics-data mining in histopathology. *Nat. Commun.* **14**, 470 (2023).
43. Rappez, L. et al. SpaceM reveals metabolic states of single cells. *Nat. Methods* **18**, 799–805 (2021).
44. Xu, B., Wang, N., Chen, T. & Li, M. Empirical evaluation of rectified activations in convolutional network. Preprint at *arXiv* https://doi.org/10.48550/arXiv.1505.00853 (2015).
45. Clevert, D.-A., Unterthiner, T. & Hochreiter, S. Fast and accurate deep network learning by exponential linear units (ELUs). In *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*. (ICLR, 2016).
46. Ringnér, M. What is principal component analysis? *Nat Biotechnol.* **26**, 303–304 (2008).
47. McInnes, L., Healy, J. & Melville, J. UMAP: uniform manifold approximation and projection for dimension reduction. Preprint at *arXiv* https://doi.org/10.48550/arXiv.1802.03426 (2018).
48. Schurch, C. M. et al. Coordinated cellular neighborhoods orchestrate antitumoral immunity at the colorectal cancer invasive front. *Cell* **182**, 1341–1359 e1319 (2020).
49. Aran, D. et al. Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. *Nat. Immunol.* **20**, 163–172 (2019).
50. Hu, J. et al. Iterative transfer learning with neural network for clustering and cell type classification in single-cell RNA-seq analysis. *Nat. Mach. Intell.* **2**, 607–618 (2020).
51. Lu, Y. The level weighted structural similarity loss: a step away from MSE. *In Proc. AAAI Conference on Artificial Intelligence* **33**, 9989–9990 (2019).
52. Lai, W.-S., Huang, J.-B., Ahuja, N. & Yang, M.-H. Deep laplacian pyramid networks for fast and accurate super-resolution. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition* 624–632 (2017).
53. Dahl, R., Norouzi, M. & Shlens, J. Pixel recursive super resolution. *Proc. IEEE International Conference on Computer Vision* 5439–5448 (2017).
54. Masutani, E. M., Bahrami, N. & Hsiao, A. Deep learning single-frame and multiframe super-resolution for cardiac MRI. *Radiology* **295**, 552–561 (2020).
55. Anwar, S., Khan, S. & Barnes, N. A deep journey into super-resolution: a survey. *ACM Comput. Surv.* **53**, 1–34 (2020).
56. Wang, Z., Chen, J. & Hoi, S. C. Deep learning for image super-resolution: a survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **43**, 3365–3387 (2020).
57. Ma, Y. & Zhou, X. Spatially informed cell-type deconvolution for spatial transcriptomics. *Nat. Biotechnol.* **40**, 1349–1359 (2022).

Acknowledgements

M.L. was supported by the following NIH grants: R01GM125301, R01EY030192, R01HL150359, R01HG013185 and P01AG066597. E.B.L. was supported by NIH grant P01AG066597. L.W. was supported in part by NIH grant R01CA266280, the Cancer Prevention and Research Institute of Texas (CPRIT) award RP200385, the University Cancer Foundation via the Institutional Research Grant Program at the University of Texas MD Anderson Cancer Center, the Andrew Sabin Family Foundation, and the Break Through Cancer Foundation.

We thank M. Meylan and W. H. Fridman for sharing the kidney cancer histology image data. We also thank Erickson, Lamb and Lundberg for sharing the prostate cancer histology image and clone annotation data.

Author contributions

This study was conceived of and led by M.L. D.Z. designed the model and algorithm, implemented the iStar software and led data analyses with input from M.L., E.E.F., L.W., K.S., G.X.X., M.D.F. and E.B.L. E.E.F. examined histology images. A.S., H.Y., M.Y.Y.L., K.S.C. and J.H. helped with data analyses. L.W. provided marker genes for TLS detection and interpreted results related to cancer. D.Z. and M.L. wrote the paper with feedback from the other co-authors.

Competing interests

M.L. receives research funding from Biogen Inc. unrelated to the current manuscript. The other authors declare no competing financial interests.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41587-023-02019-9.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41587-023-02019-9.

Correspondence and requests for materials should be addressed to Daiwei Zhang or Mingyao Li.

Peer review information *Nature Biotechnology* thanks the anonymous reviewers for their contribution to the peer review of this work.