STAN, a computational framework for inferring spatially informed transcription factor activity across cellular contexts

Linan Zhang[1†], April Sagan[2,3†], Bin Qin[4,5], Baoli Hu[4,5], Hatice Ulku Osmanbeyoglu[2,3,6,7*]
[1]Department of Applied Mathematics, School of Mathematics and Statistics, Ningbo
[2]University, Ningbo, Zhejiang 315211, China
[3]Department of Biomedical Informatics, University of Pittsburgh School of Medicine, Pittsburgh, PA 15206, USA
UPMC Hillman Cancer Center, University of Pittsburgh, Pittsburgh, PA 15232, USA
[4]Department of Neurological Surgery, University of Pittsburgh School of Medicine, Pittsburgh, Pennsylvania.
[5]Division of Pediatric Neurosurgery, UPMC Children's Hospital of Pittsburgh, Pittsburgh, Pennsylvania.
[6]Department of Bioengineering, University of Pittsburgh School of Engineering, Pittsburgh, PA 15219, USA
[7]Department of Biostatistics, University of Pittsburgh School of Public Health, Pittsburgh, PA 15261, USA

*Correspondence should be addressed to H.U.O. (osmanbeyogluhu@pitt.edu), University of Pittsburgh School of Medicine, 5051 Centre Ave, Pittsburgh, PA 15213, USA; Phone: 412-623-7789; ORCID ID: 0000-0002-4972-4347
[†]The first two authors contributed equally to this work

Abstract
Transcription factors (TFs) drive significant cellular changes in response to environmental cues and intercellular signaling. Neighboring cells influence TF activity and, consequently, cellular fate and function. Spatial transcriptomics (ST) captures mRNA expression patterns across tissue samples, enabling characterization of the local microenvironment. However, these datasets have not been fully leveraged to systematically estimate TF activity governing cell identity. Here, we present STAN (Spatially informed Transcription factor Activity Network), a linear mixed-effects computational method that predicts spot-specific, spatially informed TF activities by integrating curated TF-target gene priors, mRNA expression, spatial coordinates, and morphological features from corresponding imaging data. We tested STAN using lymph node, breast cancer, and glioblastoma ST datasets to demonstrate its applicability by identifying TFs associated with specific cell types, spatial domains, pathological regions, and ligand-receptor pairs. STAN augments the utility of ST to reveal the intricate interplay between TFs and spatial organization across a spectrum of cellular contexts.

Elaborately orchestrated transcriptional programs distinguish specialized cell types/states and define their functionality. Combinations of transcription factors (TFs) drive these programs through interactions with *cis*-regulatory elements, such as promoters and enhancers, to control cellular identity and functional state. Nearby cells are critical for determining context-specific TF activity. TFs are typically expressed at low levels and are regulated post-transcriptionally and/or post-translationally. Further, cellular processes depend on the expression levels (and activities) of proteins, notably TFs, which can be distinct from mRNA levels.

TF activities are commonly inferred from their target gene expression. Various methods to investigate TF activity have been developed based on bulk (1-3) and single-cell omics (4) datasets. Early methods using single-cell omics relied on single-cell RNA-seq (scRNA-seq) data and/or identifying TF motifs in annotated promoter regions (4-10). These methods primarily depend on co-expression of TFs and their potential target genes (e.g., SCENIC tool (4)). With the availability of single-cell epigenomic datasets such as those generated by single-cell sequencing assays for transposase-accessible chromatin, various TF inference methods have focused on context-specific accessible regions of chromatin that

may be bound by TFs either upstream or downstream of target genes. Recently, SPaRTAN (Single-cell Proteomic and RNA based Transcription factor Activity Network) was developed, leveraging single-cell proteomic and corresponding scRNA-seq datasets obtained using cellular indexing of transcriptomes and epitopes by sequencing, to link surface proteins with TFs (11). However, these approaches overlook relationships between other signaling proteins and TFs within individual cells, as well as the influence of surrounding cells on these interactions and how cell identity is affected by the local environment.

Genome-wide spatial transcriptomics (ST) platforms such as 10x Genomics' Visium ST, DBiTseq (12), and Slide-seq (13) enable transcriptome profiling of cells in their native context and provide information about how cell identity is influenced by locality. For example, Visium measures up to 5,000 spots, each containing up to 10 cells. ST has been used to study cancer (14-18), Alzheimer's disease, periodontitis (19,20), and healthy tissues (21,22). One advantage of ST data is the additional generated information of co-registered images, which contain both morphological and functional patterns. Multiple computational approaches have been introduced to analyze ST data, including the characterization of spatial gene expression patterns, spatially distributed differentially expressed genes, and spatial cell-cell communication patterns (23-34). Despite these advances, no existing methods integrate spatial and histological data with gene expression and *cis*-regulatory information to systematically estimate spot-specific TF activities.

Here, we describe a new linear mixed-effects model computational framework, STAN (Spatially informed Transcription factor Activity Network), for inferring spatially informed TF activity (**Figure 1**). This framework leverages curated TF-target gene priors, spot-specific gene expression, spatial coordinates, and morphological features extracted from corresponding hematoxylin and eosin (H&E) histological data obtained with ST. Briefly, STAN views expression of TF-target genes as a proxy of their activities. Spatial coordinates and morphological features extracted from corresponding imaging data are used to estimate spatially informed activity of TFs. Widely used ST platforms, such as Visium, measure the entire transcriptome in low-resolution spots, limiting their ability to detect the activity of cell-type-specific TFs. Therefore, STAN incorporates cell-type deconvolution approaches (e.g., Cell2location (35), DestVI (36), Tangram (37), Stereoscope (38), RCTD (39), and others) to identify cell-type-specific TFs. We have successfully applied and tested STAN using lymph node, breast cancer, and glioblastoma ST datasets to demonstrate its utility by identifying TFs associated with specific cell types, spatial domains (e.g., germinal centers [GCs]), pathological regions, and ligand-receptor pairs.

Online Methods

Data and pre-processing

We obtained three publicly available genome-wide ST datasets comprising one human lymph node sample and one glioblastoma (both from 10x Genomics website, along with six breast cancer samples derived from multiple patients from Wu et al. (18) (**Supplementary Table 1**). These datasets represents both structured (e.g. lymph node) and unstructured (e.g. solid tumors) tissues. For each ST dataset, we collected spot-level expression data, spatial coordinates of the spots, the associated image of the tissue slice. Four ST samples obtained from Wu et al. (18) included scRNA-seq data from the same patient. We also downloaded these processed scRNA-seq data with cell type annotations (18) (GSE176078).

We obtained predicted cell-type abundance scores and GC annotations for the lymph node data from Kleshchevnikov et al. (35) and pathological annotations for the breast cancer dataset from Wu et al. (18). We obtained a gene set resource comprising TF-target gene priors from hTFtarget (40) and retained only those TFs identified in the Human Transcription Factor database (41) to generate the TF-target gene prior matrix, which delineates a candidate set of associations between TFs and target genes.

We then analyzed the ST data using Scanpy (version 1.9.3) (42) and anndata (version 0.9.2) (43). Genes expressed in fewer than five spots were filtered, and low-quality spots were filtered based on unique molecular identifier (UMI) counts (dataset-specific threshold, **Supplementary Table 2**). For each ST dataset, the TF-target gene prior matrix

(referred to as $\mathbf{D}$) and spot-level gene expression matrix (referred to as $\mathbf{Y}$) underwent a sequential filtering process. Initially, genes expressed in less than 20% of the filtered spots were eliminated. Subsequently, only mutual genes present in both matrices were retained. Further refinement involved removing genes associated with fewer than five remaining TFs. Finally, TFs associated with fewer than 10 remaining genes were removed to address insufficient target gene representation for reliable analysis. For each filtered gene expression matrix, we normalized each spot by total counts over all genes in that dataset. The normalized counts were then square root transformed to stabilize the variance.

Training sample-specific STAN models

STAN is designed to leverage *cis*-regulatory information (e.g., TF-target gene priors), gene expression, spatial information, and morphological features extracted from corresponding H&E staining imaging data acquired using ST to enable inference of spatially informed, spot-specific TF activities. **Figure 1** illustrates our overall STAN framework. We let $\mathbf{D} \in \mathrm{R}^{N \times Q}$ and $\mathbf{Y} \in \mathrm{R}^{N \times M}$ denote the TF-target gene prior matrix and the normalized spot-level gene expression matrix where *N*, *M*, and *Q* represent the number of genes, spots, and TFs, respectively. Subsequently, we formulated a regression problem aimed at learning the weight matrix $\mathbf{W} \in \mathrm{R}^{Q \times M}$ between TFs and spots to predict TF-target gene expression through the equation:

$$\mathbf{DW} \approx \mathbf{Y}. \tag{1}$$

To include spatial features in Equation (1), we defined a Gaussian kernel matrix $\mathbf{K} \in \mathrm{R}^{M \times M}$ as follows. For the ith spot, we first cropped the image to a square of size $(2a+1)$ centered at the spot and then calculated the pixel intensity of the spot as the mean RGB values across all pixels in the cropped image. The spatial coordinate of the spot, (x_i, y_i), as well as the RGB pixel intensity of the spot (r_i, g_i, b_i), forms a 5-dimentional vector $\mathrm{v}^{(i)} = (x_i, y_i, \omega\, r_i, \omega\, g_i, \omega\, b_i)$, which contains both the spatial and morphological information of the ith spot (44), and ω is a weight parameter multiplied to its morphological components. Next, we normalized each component in v to have zero mean and unit variance across all spots. We let b be the parameter that determines the bandwidth of the Gaussian kernel. Given two spots indexed i and j, the(i,j)th element of $\mathbf{K}$ is given by:

$$\boldsymbol{K}_{i,j} = \exp\left(-\frac{\left\|\mathrm{v}^{(i)} - \mathrm{v}^{(j)}\right\|_2^2}{2b^2}\right).$$

Note that the kernel $\mathbf{K}$ is symmetric (i.e., $\mathbf{K} = \mathbf{K}^T$).

We assumed that the TF activity matrix $\mathbf{W}$ can be split into two matrices,

$$\mathbf{W} = \mathbf{W}_{\mathrm{sd}} + \mathbf{W}_{\mathrm{si}} \tag{2}$$

where $\mathbf{W}_{\mathrm{sd}}$ is spatially and morphologically dependent, and $\mathbf{W}_{\mathrm{si}}$ is spatially and morphologically independent. With the kernel $\mathbf{K}$ defined, to find the optimal solution $\mathbf{W}^*$ of Equation (1), which has the minimal residual while reserving as many spatial features as possible, we solved Equation (1) using a spatially weighted ridge regression:

$$\min_{\mathbf{W}} \frac{1}{2}\|\mathbf{DW} - \mathbf{Y}\|_F^2 + \frac{\lambda_1}{2}\|\mathbf{W}_{\mathrm{sd}}\mathbf{K}\|_F^2 + \frac{\lambda_2}{2}\|\mathbf{W}_{\mathrm{si}}\|_F^2, \tag{3}$$
$$\text{subject to } \mathbf{W} = \mathbf{W}_{\mathrm{sd}} + \mathbf{W}_{\mathrm{si}},$$

where F denotes the Frobenius norm, and λ_1, λ_2 are positive hyperparameters. The nonspatial component $\mathbf{W}_{\mathrm{si}}$ is L_2-regularized, while the L_2-regularizer of the spatial component $\mathbf{W}_{\mathrm{sd}}$ is weighted by the kernel $\mathbf{K}$. Equation (3) can be solved efficiently via a low rank approximation of the spanning set of **K**.

The analytic solution of STAN can be derived as follows. Let F be the objective function of the optimization problem (3), that is:

$$F(\mathbf{W}_{\mathrm{sd}}, \mathbf{W}_{\mathrm{si}}) = \frac{1}{2}\|\mathbf{D}(\mathbf{W}_{\mathrm{sd}} + \mathbf{W}_{\mathrm{si}}) - \mathbf{Y}\|_F^2 + \frac{\lambda_1}{2}\|\mathbf{W}_{\mathrm{sd}}\mathbf{K}\|_F^2 + \frac{\lambda_2}{2}\|\mathbf{W}_{\mathrm{si}}\|_F^2.$$

The gradients of F with respect $\mathbf{W}_{\mathrm{sd}}$ and $\mathbf{W}_{\mathrm{si}}$ are:

$$\nabla_{\mathrm{sd}} F(\mathbf{W}_{\mathrm{sd}}, \mathbf{W}_{\mathrm{si}}) = \mathbf{D}^T(\mathbf{D}(\mathbf{W}_{\mathrm{sd}} + \mathbf{W}_{\mathrm{si}}) - \mathbf{Y}) + \lambda_1 \mathbf{W}_{\mathrm{sd}}\mathbf{K}\mathbf{K}^T,$$

$$\nabla_{\mathrm{si}} F(\mathbf{W}_{\mathrm{sd}}, \mathbf{W}_{\mathrm{si}}) = \mathbf{D}^T(\mathbf{D}(\mathbf{W}_{\mathrm{sd}} + \mathbf{W}_{\mathrm{si}}) - \mathbf{Y}) \ + \ \lambda_2 \mathbf{W}_{\mathrm{si}}.$$

The optimal solutions $\mathbf{W}_{\mathrm{sd}}^*$ and $\mathbf{W}_{\mathrm{si}}^*$ that minimize F satisfy $\nabla_{\mathrm{sd}} F(\mathbf{W}_{\mathrm{sd}}^*, \mathbf{W}_{\mathrm{si}}^*) = \nabla_{\mathrm{si}} F(\mathbf{W}_{\mathrm{sd}}^*, \mathbf{W}_{\mathrm{si}}^*) = 0$, that is:

$$\mathbf{D}^T(\mathbf{D}(\mathbf{W}_{\mathrm{sd}}^* + \mathbf{W}_{\mathrm{si}}^*) - \mathbf{Y}) \ + \ \lambda_1 \mathbf{W}_{\mathrm{sd}}^* \mathbf{K}\mathbf{K}^T = 0, \tag{4}$$

$$\mathbf{D}^T(\mathbf{D}(\mathbf{W}_{\mathrm{sd}}^* + \mathbf{W}_{\mathrm{si}}^*) - \mathbf{Y}) \ + \ \lambda_2 \mathbf{W}_{\mathrm{si}}^* = 0. \tag{5}$$

Comparing Equations (4) and (5), we can observe that:

$$\lambda_2 \mathbf{W}_{\mathrm{si}}^* = \lambda_1 \mathbf{W}_{\mathrm{sd}}^* \mathbf{K}\mathbf{K}^T. \tag{6}$$

Combining Equations (2) and (6), the optimal solution $\mathbf{W}^*$ of Equation (1) can be simplified as:

$$\mathbf{W}^* = \mathbf{W}_{\mathrm{sd}}^* + \mathbf{W}_{\mathrm{si}}^* = \mathbf{W}_{\mathrm{sd}}^* \left(\mathbf{I} + \frac{\lambda_1}{\lambda_2} \mathbf{K}\mathbf{K}^T\right),$$

where $\mathbf{I}$ is the identify matrix. Thus, it is sufficient to solve $\mathbf{W}_{\mathrm{sd}}^*$.

From Equation (4):

$$\mathbf{D}^T \mathbf{D} \mathbf{W}_{\mathrm{sd}}^* (\lambda_2 \mathbf{I} + \lambda_1 \mathbf{K}\mathbf{K}^T) + \ \lambda_1 \lambda_2 \mathbf{W}_{\mathrm{sd}}^* \mathbf{K}\mathbf{K}^T = \ \lambda_2 \mathbf{D}^T \mathbf{Y}. \tag{7}$$

Reformulating the matrix products in Equation (7) yields an equivalent linear system:

$$\left((\lambda_2 \mathbf{I} + \lambda_1 \mathbf{K}\mathbf{K}^T) \otimes (\mathbf{D}^T \mathbf{D}) + \mathbf{I} \otimes (\mathbf{K}\mathbf{K}^T)\right) \mathrm{vec}(\mathbf{W}_{\mathrm{sd}}^*) \ = \ \mathrm{vec}(\lambda_2 \mathbf{D}^T \mathbf{Y}), \tag{8}$$

where $\otimes$ is the Kronecker multiplication and vec is the vectorization of a matrix, and we make use of the symmetric properties of $\mathbf{I}$ and $\mathbf{K}$. From Equation (8) we have:

$$\mathrm{vec}(\mathbf{W}_{\mathrm{sd}}^*) = \ \left((\lambda_2 \mathbf{I} + \lambda_1 \mathbf{K}\mathbf{K}^T) \otimes (\mathbf{D}^T \mathbf{D}) + \mathbf{I} \otimes (\mathbf{K}\mathbf{K}^T)\right)^{-1} \mathrm{vec}(\lambda_2 \mathbf{D}^T \mathbf{Y}),$$

which can be solved efficiently using singular value decomposition (SVD).

For comparison, we also solved Equation (1) using the Ridge regression:

$$\min_{\mathbf{W}} \frac{1}{2} \|\mathbf{D}\mathbf{W} - \mathbf{Y}\|_F^2 \ + \ \frac{\lambda}{2} \|\mathbf{W}\|_F^2, \tag{9}$$

which does not model any spatial features, and the solution from Equation (9) was considered as the baseline. We evaluated performance with 10-fold cross-validation and computed the Pearson correlation coefficient (PCC) between the predicted and measured gene expression profiles on held-out spots. Finally, we used the trained **W** to infer the TF activities for each spot. To cluster gene expression and STAN-predicted TF activities (**W**, scaled to unit variance and zero mean), we employed the Leiden algorithm (51). Additionally, a neighbor graph of spots was generated and embedded using uniform manifold approximation and projection (UMAP) (49,50).

Deconvoluting ST data

We employed GraphST (version 1.1.1) (45) to perform cell-type deconvolution on human breast cancer samples, integrating both scRNA-seq and ST data (18). GraphST is a graph-based self-supervised contrastive learning method which makes use of both spatial information and gene expression profiles for spatially informed clustering, batch integration, and cell-type deconvolution. Utilizing the shared samples (**Supplementary Table 3**), we applied the GraphST deconvolution pipeline with the recommended settings. Following training of the graphical model, we derived a mapping matrix wherein each element represents the mapping probability of a cell type in each spot.

Inferring cell-type-specific TFs

Widely used ST platforms, such as Visium, measure the whole transcriptome in low-resolution spots that may capture multiple cells. To identify cell-type-specific TFs using STAN, we conducted association analysis between TFs and cell types by predicting the activity of each TF individually from spot-specific cell type proportions. These proportions were estimated using deconvolution methods (e.g., Cell2location (35), GraphST (45), DestVI (36), Tangram (37), Stereoscope (38), RCTD (39), and others), to integrate spatial RNA-seq data with reference cell-type signatures obtained from scRNA-seq profiles of the same tissue type. Briefly, we let $\mathbf{A}$ denote each spot-specific cell type matrix, where each row of $\mathbf{A}$ represents a spot and each column a cell type. For the ith TF, we let w_i be a column of $\mathbf{W}$, which is the inferred activity of the TF across all spots. Then, by using a linear regression model (the ordinary least squares method), we inferred the cell-type-specific score x_i of the

TF via $\mathbf{A}x_i = w_i$, where each component of x_{ij} is the TF score for the jth cell type. We evaluated the significance of effect size (coefficient) for each cell type in the regression model using a Student's t-distribution under the null hypothesis. Following false discovery rate (FDR) correction across TFs, we identified a significant set of TFs whose activities are associated with each cell type.

Identifying pathological region-specific TFs

To identify pathological region-specific TFs, we initially scaled the inferred TF activity matrix to unit variance and zero mean. Subsequently, for each region and each TF, we conducted a Wilcoxon rank-sum test to compare the inferred TF activity within and outside the region. Finally, we adjusted *P*-values using the Benjamini-Hochberg (BH) procedure.

Linking TFs to ligands and receptors

We obtained a list of human ligand-receptor interaction pairs from the CellTalkDB database (46), which includes 3,398 available ligand-receptor interaction pairs from among 815 ligands and 780 receptors. We began by determining the average normalized mRNA counts of neighboring spots for the ligands and receptors individually calculated for each spot. Subsequently, we computed the PCC between the mean mRNA levels of neighboring spots and the inferred TF activities. Highly correlated TF-ligand, TF-receptor, and ligand-receptor pairs were selected based on thresholds (PCC > 0.6), and PCCs were visualized in clustered heatmaps.

Implementing decoupleR

We compared STAN with decoupleR (version 1.5.0) (47), which can be used to extract biological activities (e.g. transcription factor activities) from expression-level data. At the recommended settings for TF activity inference, we ran decoupleR for the filtered human lymph node sample (**Supplementary Table 2**).

Immunofluorescence (IF)

For IF staining, OCT frozen brain tumor sections were thawed at room temperature for 30 min, rinsed and rehydrated with TBS 3 times, fixed the tissue for 10 minutes in acetone .After blocking with normal horse serum(2.5%), the sections were incubated with indicated primary antibodies overnight at 4°C. The specific dilution of primary antibodies used are as follows: rabbit SOX2(1:200;EMD Millipore AB5603),rat CD44(1:100;Biolegend 103002),mouse VIM(1:100;Biolegend 677801). After three washes with TBST(0.1% Tween), the slices were incubated with goat anti-rabbit IgG H&L Alexa Fluor 488 (1:500; Invitrogen A11008) , donkey anti-rat IgG H&L Alexa Fluor 594 (1:500; Invitrogen A21209) and goat anti-mouse IgG H&L Alexa Fluor 647 (1:500; Invitrogen A32728) at RT for 1h. After washing with TBST three times, slides were mounted with DAPI (Abcam ab104139) and imaged using a Leica DMi8 confocal microscope. Frozen glioblastoma (GBM) samples from The Cancer Genome Atlas (TCGA) collections were described previously (48).

Statistical analysis and visualization

Statistical analysis and visualization were performed in Python (version 3.11.5) using packages including SciPy (version 1.10.1) and statsmodels (version 0.14.0). For general data analysis and manipulation, pandas (version 2.0.3) and NumPy (version 1.22.4) were used. Graphs were generated using Python packages matplotlib (version 3.7.2) and seaborn (version 0.12.2) and R (version 4.4.0) packages ggplot2 (version 3.5.1) and ComplexHeatmap (version 2.20.0).

For comparisons of inferred TF activities between groups, we performed Wilcoxon rank-sum test and determined the direction of shifts by comparing the mean of the two populations. We corrected raw *P*-values for multiple hypothesis testing using Bonferroni and FDR (BH) methods.

Results

Inferring TF activity according to cell type, physical location, and ligand/receptor associations

STAN is a toolbox for analyzing spot-based ST datasets to estimate spot-specific TF activities and identify TFs associated with specific cell types, spatial domains, pathological regions, and ligands/receptors (**Figure 1**). Methodologically, STAN takes spot-specific morphological information, spatial coordinates, and gene expression ($\mathbf{Y}$), together with TF-target gene priors ($\mathbf{D}$), as an input and then models gene expression in terms of spatially coherent TF activities (**Figure 1A**). To achieve this capability, we used a kernel regression model, where we created a spot-specific TF activity matrix, $\mathbf{W}$**,** that is decomposed into two terms: one required to follow a spatial pattern ($\mathbf{W}_{\mathrm{sd}}$) generated using a kernel matrix ($\mathbf{K}$) and another that is unconstrained but regularized using the L_2- norm ($\mathbf{W}_{\mathrm{si}}$). The kernel matrix enforces similarity in spot-specific TF activity in nearby physical spaces and/or morphologically. We constructed $\mathbf{K}$ using spatial proximity between spots (Euclidean distance) and morphological similarity (measured as the similarity between pixel intensity for H&E-stained images). The model then fits the expression of each gene as the sum of the activities for a set of TFs.

A key limitation of most grid-based ST profiling methods is the lack of single-cell resolution, as spatial RNA-seq measurements combine multiple cells (e.g., 1–10 cells). Hence, inferred spot specific activity of TFs are from multiple cells. To overcome this limitation, we used deconvolution methods to estimate cell type proportions for each spot by the computational integration of spatial RNA-seq with the reference transcriptome signatures of cell types obtained from scRNA-seq profiles (**Figure 1B**). We then performed association analysis between TFs and cell types using linear regression to predict the activity of each TF individually from the cell type proportions based on deconvolution methods (**Figure 1C**). We also performed downstream analysis to identify TFs whose activities differentially vary across pathological regions and spatial domains (**Figure 1D**) and to elucidate ligands/receptors whose gene expression is associated with TFs in neighboring spots (**Figure 1E**).

Identifying GCs and cell-type-specific TFs in the human lymph node

We first trained STAN using a publicly available Visium ST dataset from a human lymph node (**Supplementary Figure 1A**). For statistical evaluation, we computed the PCC between predicted and measured gene expression profiles on held-out spots using 10-fold cross-validation. We obtained significantly better performance than the baseline model without spatial/morphological information based on Ridge regression ($P < 10^{-16}$, one-sided Wilcoxon rank-sum test; **Figure 2A**). In particular, the increased benefit of spatial information and prediction performance was inversely correlated with the total UMI count per spot (PCC = -0.2097), suggesting that our method of pooling information from neighboring spots was more effective than utilizing spots with low UMI counts.

Clustering of STAN-predicted TF activities identified ten major clusters and revealed that the corresponding spatial distribution showed highly distinct features (**Figure 2B**). The spatial distribution based on mRNA expression correlated well with TF activity clusters but appeared to be noisier and less precise (**Supplementary Figure 1B**). For example, the human lymph node is characterized by dynamic microenvironments with many spatially interlaced cell populations, such as GCs. Histological examination of the Visium human lymph node sample revealed multiple GCs (35) (**Figure 2C**); more than half annotated GC regions were part of Cluster 0 (**Figure 2B**; **Supplementary Table 4)**. We next assessed TF-GC associations by comparing TF activities between spots in a given GC region vs. those in all other regions. We identified several known TFs associated with GCs in the lymph node sample (**Figure 2D**), including BACH2, which is critical to the formation and survival of GCs (49), CREB1 (50), and E2F1 (51). We found fewer associations using TF mRNA expression levels directly (**Figure 2E**, **Supplementary Figure 1C–E**) or when using decoupleR (47) without leveraging spatial and morphological features (**Supplementary Figure 2A–B).**

To identify cell-type-specific TFs, we next used cell type proportions obtained from the Cell2location deconvolution approach (35). Briefly, Kleshchevnikov et al. analyzed the Visium dataset of the human lymph node and spatially mapped a comprehensive atlas of reference cell types derived by integration of scRNA-seq datasets from human secondary lymphoid organs. To explore the associations between TFs and cell types, we first computed correlations between spot-specific TF activities and cell type proportions. For several known cell-type-specific TFs, we observed high correlation. For example, FOXP3 activity was highly correlated with the regulatory T (Treg) cell type proportion (PCC = 0.5904) (**Figure 2F**), whereas STAT2 activity was highly correlated with cells expressing interferon-β (IFN-β) (PCC = 0.7992) (**Figure 2G**). Indeed, FOXP3 is a master TF that regulates Treg cell development and function (52), and STAT2 is essential for immune responses involving Type 1 IFNs, including IFN-β (53). Correlations between FOXP3/STAT2 activity and cell type proportion were higher than those between *FOXP3/STAT2* mRNA expression and cell type proportion (PCC = 0.3223 and 0.3421, respectively; **Figure 2F–G**), highlighting the utility of our method for detecting biologically meaningful patterns in ST data. In addition, correlations between FOXP3/STAT2 activity and cell type proportion are higher with inference from STAN than from decoupleR (PCC = 0.4428 and 0.4289, respectively; **Figure 2F–G**), indicating the advantage of incorporating spatial and morphological features in TFA inference.

To systematically identify cell-type-specific TFs, we next modeled the relationship between estimated cell type proportions and TF activities using linear regression. This model learns a coefficient for each cell type and for each TF, which we interpreted as the TF score for each cell type across spots. We identified several novel and known TF-cell type relationships in the lymph node sample (**Figure 2H**). These inferred TFs are known regulators of several lymph node cell types including FOXP3 (52) for Tregs; ETS1 for NK cells (54); SPI1 (PU1) (55,56) for macrophages and dendritic cells; BACH2 (57) for B and T cells; IRF1, IRF2, and STAT2 (58) (53) for TIM^+ T cells and cells expressing IFN-β. We also identified TFs involved in the control of cell cycle progression, such as MAX, KMT2A, E2F, GTF2B, and FOXM1, specifically associated with proliferating B cells but not other B cell states nor other nonproliferative cell types. Importantly, these known relationships were not identified by TF mRNA levels alone (**Supplementary Figure 1E**) or by decoupleR identification of cell-type-specific TF activities using only single-cell gene expression measurements.

TF activity patterns across distinct pathological regions in breast cancer

To showcase its applicability to analyze a complex cancer tissue, we tested STAN on a publicly available breast cancer ST dataset (n=6) (18). Leveraging pathological annotations (**Figure 3A** and **Supplementary Figure 3A-B**) from both triple-negative breast cancer (TNBC, n=4) and estrogen-receptor-positive breast cancer (ER+, n=2) samples, STAN unveiled TF activity patterns across distinct pathological regions in these types of cancer (18) (**Figure 3B** and **Supplementary Figures 4–5)**. We first compared TF activity patterns associated with pathological regions across all breast cancer samples (18), three of which appeared in at least four of the six samples: stroma, lymphocytes, and the combination of invasive cancer, stroma, and lymphocytes (**Figure 3B–C**). We observed that TFs associated with lymphocyte and stroma regions for all samples showed consisted activity patterns. For the combined invasive cancer, stroma, and lymphocyte region, we observed heterogeneous TF activity profiles between the TNBC and ER+ samples.

In **Figure 3D**, we show TFs associated with different pathological regions for one TNBC sample as a representative relationship. In lymphocyte-rich regions with and without invasive cancer regions, we observed increased activity of IRF, STAT family TFs, FOXP3 (**Figure 3E**), and EBF1. These TFs are important regulators of B and T cell function and differentiation. We also found that FOXA1 activity was high in invasive cancer regions, the stroma, and lymphocytes (**Figure 3E**). FOXA1 plays pivotal roles in breast development, is crucial for mammary morphogenesis (21,23), mediates full ER activity (24), and directly

interacts with the TF GATA3. Notably, corresponding TF gene expression patterns did not typically match TF activity, except for some TFs whose encoding gene was more highly expressed, such as FOXA1 (**Figure 3E**).

Four ST samples obtained from Wu et al. (18) included scRNA-seq data from the same patient. We inferred cell-type-specific TF activities by fitting the model to the scRNA-seq data without spatial information, computed cell-type-specific TF scores, and calculated Spearman's rank correlation coefficient for the TF scores for each cell type between corresponding ST and scRNA-seq data (**Figure 3F**). The medium-to-high correlation for most cell types and samples was consistent between scRNA-seq and ST-based inference, despite the ST data being sparser.

Linking ligands and receptors to TFs in glioblastoma

Cell behavior is intricately influenced by signals from their surrounding microenvironment, often in the form of ligands, which are extracellular protein signals produced by neighboring cells. These ligands interact with receptors expressed on recipient cells, triggering a cascade of molecular events that can profoundly impact gene expression programs and modulate the activity of TFs. To further explore this interplay between ligands, receptors, and TFs, we leveraged STAN to analyze Visium ST data obtained from one human glioblastoma sample (**Supplementary Figure 6**). We employed inferred TF activities to model intercellular signaling pathways originating from ligands expressed within the local neighborhood of cellular regions and to explore the complex communication networks within the tumor microenvironment.

We began by determining the average normalized mRNA counts of neighboring spots for a selection of ligands extracted from the CellTalkDB database (46), individually calculated for each spot. Subsequently, we examined the correlation between the mean mRNA levels of neighboring spots and the inferred TF activity and identified patterns in the interplay between ligand and receptor expression, TF activity, and the cellular microenvironment (**Figure 4A**) through a two-way clustering of ligands, receptors, and TFs based on pairwise PCCs. For example, our analysis revealed a significant correlation (PCC > 0.6) between STAN-predicted SOX2 and HIF1A activity and expression levels of RACK1, CD44 receptors and VIM (vimentin), VEGFA and DKK1 ligands (**Figure 4A-B**). SOX2 has been extensively implicated in promoting malignancy in glioblastoma, with its regulatory roles extending to CD44, DKK1, and VIM, among other factors (59). Additionally, RACK1, acting as a scaffolding protein, modulates the expression of pluripotency TFs such as Nanog, Oct4, and SOX2 in human hepatocellular carcinoma (60). To further explore SOX2 coupling with CD44 and VIM, we performed immunofluorescence (IF) on glioblastoma patient specimens. Indeed, we found co-expression of SOX2, CD44 and VIM at the protein level (**Figure 4C**).

Discussion

ST datasets provide rich information about cellular diversity and heterogeneity, as well as the dynamics of complex biological systems. One advantage of ST data is the additional information of co-registered images, which contain both morphological and functional patterns and distance information. Here, we present STAN, a broadly applicable tool that leverages parallel spot-specific ST data and imaging data with *cis*-regulatory information (e.g., TF-target gene priors) to predict spatially informed TF activities. While several methods for inferring TF activities use TF expression to predict gene expression, they do not consider how cell identity is influenced by neighboring cells.

We utilized STAN to uncover TF relationships across diverse tissue types, from lymph nodes to more complex tissues such as breast cancer and glioblastoma. Applying STAN to ST datasets (i) deciphers critical TF regulators underlying cell identities, spatial domains (e.g., GCs), and pathological regions (e.g., stroma vs. tumor); (ii) determine whether a given pathological region/spatial domain has different and/or common regulators across disease subtypes (e.g., stroma in TNBC vs. ER+ breast cancer); (iii) identify similar and/or different TFs associated with spatial domains or cell types across healthy individuals

and those manifesting a disease; and (iv) links ligands and receptors to TFs for elucidating potential signaling pathways and regulatory networks involved in cellular communication and tissue microenvironment interactions.

The identified TFs from our analysis of lymph node, glioblastoma, and breast cancer data are clinically and biologically relevant. The distinctive activity patterns of these TFs may offer valuable insights for diagnostic and prognostic purposes in the management of breast cancer, glioblastoma, and other cancer types. Within a tumor, several subpopulations of cancer cells can differ from each other in their context-specific TFs, which drive their functions due to residing in distinct tumor microenvironments. Accurately identifying context-specific TFs, such as those associated with malignant cells in regions with a higher degree of infiltrating immune cells compared to those without, is crucial. Further, we used STAN to delineate TF, ligand and receptor relationships. For example, our analysis revealed a significant correlation (PCC > 0.6) between STAN-predicted SOX2 activity and expression levels of CD44 receptors and VIM (vimentin) ligand. We validated co-expression of SOX2, VIM and CD44 in the protein level using independent glioblastoma patient specimens. Future validation experiments can evaluate the role of SOX2 coupling with VIM and CD44 in glioblastoma. SOX2 expression can be evaluated in the setting of VIM and CD44 modulation in cultured glioblastoma cells. Ultimately, in vivo validation of the role of SOX2, VIM and CD44 can be studied using glioblastoma mouse models through silencing of SOX2 and/or VIM, CD44 and measuring the functional activity of glioblastoma cells. Overall, STAN is applicable to any disease type (e.g., autoimmune diseases) or biological system (e.g., immune system) and will enable researchers to leverage spot-based ST datasets to predict context-specific TF activities.

The method we describe does have some limitations. First, measurements from capture locations (spots) often include mixtures of multiple cells, leading to analytical challenges in dissecting cellular disposition, particularly in complex systems, such as cancerous tissues. However, the STAN framework can be extended modeling of emerging single-cell resolution ST, as we will report elsewhere. Second, STAN relies on curated TF target-gene interactions from various sources (e.g., literature-curated data, ChIP-seq peaks, TF binding site motifs) that are often noisy, incomplete, and not context-specific. Our framework can be extended using emerging spatial epigenome-transcriptome co-profiling (61) for more accurate representation of both promoter and enhancer regions as performed in the context of patient-specific predictive regulatory models (62,63). In addition, STAN does not account for directionality in the TF-target gene interaction matrix. Thus, negative values of inferred TF activities should be interpreted based on whether the TF acts as an activator or repressor. Our model assumes a TF either induces or represses its target(s), but some TFs may switch roles based on co-factors, potentially confounding data interpretation. Lastly, we use a fixed gene-target representation, inferring TF activity by correlating it with target expression in a linear model, which does not capture the more complex combinatorics of TF binding. Despite these limitations, STAN will enable more robust and context-specific analysis of cellular regulatory states in heterogenous and complex tissue types.

Data Availability

The 10x Genomics datasets were downloaded using the Scanpy function scanpy.datasets.visium_sge with sample IDs V1_Human_Lymph_Node and Parent_Visium_Human_Glioblastoma. The cell type and geminal center annotations were downloaded from the Cell2location study by Kleshchevnikov et al. (35) (https://github.com/vitkl/cell2location_paper/tree/master/notebooks/selected_results/lymph_nodes_analysis). The ST data from Wu et al. (18), including pathological annotations, were accessed through the Zenodo data repository (https://doi.org/10.5281/zenodo.4739739). The scRNA-seq data from Wu et al. (18) were obtained from the Gene Expression Omnibus (GEO) dataset under accession number GSE176078. A list of human ligand-receptor interaction pairs was obtained from the CellTalkDB database (46) (https://github.com/ZJUFanLab/CellTalkDB/blob/master/database/human_lr_pair.rds).

Code Availability

A Python package code for STAN is available at https://github.com/osmanbeyoglulab/STAN. Code used to produce the results in this paper, including for data pre-processing, cell type deconvolution, fitting models, and downstream analyses, are available as Jupyter notebooks in the same GitHub repository.

Acknowledgements

We thank Jing Hong Wang, Anthony R. Cillo, and Harinder Singh for helpful discussions.

Funding

HUO was supported by the National Institutes of Health (R35GM146989). LZ was supported by the National Natural Science Foundation of China (12101342) and the Ningbo Yongjiang Talent Introduction Programme (2022A-222-G). Data analyses were supported by the University of Pittsburgh Center for Research Computing and the Extreme Science and Engineering Discovery Environment (XSEDE) supported by the National Science Foundation (OCI-1053575) using the Bridges2 system supported by the National Science Foundation (ACI-1445606) at the Pittsburgh Supercomputing Center.

Conflict of Interest Statement

None declared.

REFERENCES

1. Osmanbeyoglu, H.U., Toska, E., Chan, C., Baselga, J. and Leslie, C.S. (2017) Pancancer modelling predicts the context-specific impact of somatic mutations on transcriptional programs. *Nat Commun*, **8**, 14249.
2. Osmanbeyoglu, H.U., Pelossof, R., Bromberg, J.F. and Leslie, C.S. (2014) Linking signaling pathways to transcriptional programs in breast cancer. *Genome Res*, **24**, 1869-1880.
3. Balwierz, P.J., Pachkov, M., Arnold, P., Gruber, A.J., Zavolan, M. and van Nimwegen, E. (2014) ISMARA: automated modeling of genomic signals as a democracy of regulatory motifs. *Genome Res*, **24**, 869-884.
4. Aibar, S., Gonzalez-Blas, C.B., Moerman, T., Huynh-Thu, V.A., Imrichova, H., Hulselmans, G., Rambow, F., Marine, J.C., Geurts, P., Aerts, J. *et al.* (2017) SCENIC: single-cell regulatory network inference and clustering. *Nat Methods*, **14**, 1083-1086.
5. Van de Sande, B., Flerin, C., Davie, K., De Waegeneer, M., Hulselmans, G., Aibar, S., Seurinck, R., Saelens, W., Cannoodt, R., Rouchon, Q. *et al.* (2020) A scalable SCENIC workflow for single-cell gene regulatory network analysis. *Nat Protoc*, **15**, 2247-2276.
6. Chan, T.E., Stumpf, M.P.H. and Babtie, A.C. (2017) Gene Regulatory Network Inference from Single-Cell Data Using Multivariate Information Measures. *Cell Syst*, **5**, 251-267 e253.
7. Matsumoto, H., Kiryu, H., Furusawa, C., Ko, M.S.H., Ko, S.B.H., Gouda, N., Hayashi, T. and Nikaido, I. (2017) SCODE: an efficient regulatory network inference algorithm from single-cell RNA-Seq during differentiation. *Bioinformatics*, **33**, 2314-2321.
8. Fiers, M., Minnoye, L., Aibar, S., Bravo Gonzalez-Blas, C., Kalender Atak, Z. and Aerts, S. (2018) Mapping gene regulatory networks from single-cell omics data. *Brief Funct Genomics*, **17**, 246-254.
9. Holland, C.H., Tanevski, J., Perales-Paton, J., Gleixner, J., Kumar, M.P., Mereu, E., Joughin, B.A., Stegle, O., Lauffenburger, D.A., Heyn, H. *et al.* (2020) Robustness and applicability of transcription factor and pathway analysis tools on single-cell RNA-seq data. *Genome Biol*, **21**, 36.
10. Hu, X., Hu, Y., Wu, F., Leung, R.W.T. and Qin, J. (2020) Integration of single-cell multi-omics for gene regulatory network inference. *Comput Struct Biotechnol J*, **18**, 1925-1938.
11. Ma, X., Somasundaram, A., Qi, Z., Hartman, D.J., Singh, H. and Osmanbeyoglu, H.U. (2021) SPaRTAN, a computational framework for linking cell-surface receptors to transcriptional regulators. *Nucleic Acids Res*, **49**, 9633-9647.
12. Liu, Y., Yang, M., Deng, Y., Su, G., Enninful, A., Guo, C.C., Tebaldi, T., Zhang, D., Kim, D., Bai, Z. *et al.* (2020) High-Spatial-Resolution Multi-Omics Sequencing via Deterministic Barcoding in Tissue. *Cell*, **183**, 1665-1681 e1618.

13. Rodriques, S.G., Stickels, R.R., Goeva, A., Martin, C.A., Murray, E., Vanderburg, C.R., Welch, J., Chen, L.M., Chen, F. and Macosko, E.Z. (2019) Slide-seq: A scalable technology for measuring genome-wide expression at high spatial resolution. *Science*, **363**, 1463-1467.
14. Berglund, E., Maaskola, J., Schultz, N., Friedrich, S., Marklund, M., Bergenstrahle, J., Tarish, F., Tanoglidi, A., Vickovic, S., Larsson, L. *et al.* (2018) Spatial maps of prostate cancer transcriptomes reveal an unexplored landscape of heterogeneity. *Nat Commun*, **9**, 2419.
15. Stahl, P.L., Salmen, F., Vickovic, S., Lundmark, A., Navarro, J.F., Magnusson, J., Giacomello, S., Asp, M., Westholm, J.O., Huss, M. *et al.* (2016) Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science*, **353**, 78-82.
16. Thrane, K., Eriksson, H., Maaskola, J., Hansson, J. and Lundeberg, J. (2018) Spatially Resolved Transcriptomics Enables Dissection of Genetic Heterogeneity in Stage III Cutaneous Malignant Melanoma. *Cancer Res*, **78**, 5970-5979.
17. Ji, A.L., Rubin, A.J., Thrane, K., Jiang, S., Reynolds, D.L., Meyers, R.M., Guo, M.G., George, B.M., Mollbrink, A., Bergenstrahle, J. *et al.* (2020) Multimodal Analysis of Composition and Spatial Architecture in Human Squamous Cell Carcinoma. *Cell*, **182**, 1661-1662.
18. Wu, S.Z., Al-Eryani, G., Roden, D.L., Junankar, S., Harvey, K., Andersson, A., Thennavan, A., Wang, C., Torpy, J.R., Bartonicek, N. *et al.* (2021) A single-cell and spatially resolved atlas of human breast cancers. *Nat Genet*, **53**, 1334-1347.
19. Navarro, J.F., Croteau, D.L., Jurek, A., Andrusivova, Z., Yang, B., Wang, Y., Ogedegbe, B., Riaz, T., Stoen, M., Desler, C. *et al.* (2020) Spatial Transcriptomics Reveals Genes Associated with Dysregulated Mitochondrial Functions and Stress Signaling in Alzheimer Disease. *iScience*, **23**, 101556.
20. Lundmark, A., Gerasimcik, N., Bage, T., Jemt, A., Mollbrink, A., Salmen, F., Lundeberg, J. and Yucel-Lindberg, T. (2018) Gene expression profiling of periodontitis-affected gingival tissue by spatial transcriptomics. *Sci Rep*, **8**, 9370.
21. Asp, M., Salmen, F., Stahl, P.L., Vickovic, S., Felldin, U., Lofling, M., Fernandez Navarro, J., Maaskola, J., Eriksson, M.J., Persson, B. *et al.* (2017) Spatial detection of fetal marker genes expressed at low level in adult human heart tissue. *Sci Rep*, **7**, 12941.
22. Maniatis, S., Aijo, T., Vickovic, S., Braine, C., Kang, K., Mollbrink, A., Fagegaltier, D., Andrusivova, Z., Saarenpaa, S., Saiz-Castro, G. *et al.* (2019) Spatiotemporal dynamics of molecular pathology in amyotrophic lateral sclerosis. *Science*, **364**, 89-93.
23. Forcato, M., Romano, O. and Bicciato, S. (2021) Computational methods for the integrative analysis of single-cell data. *Brief Bioinform*, **22**, 20-29.
24. Palla, G., Spitzer, H., Klein, M., Fischer, D., Schaar, A.C., Kuemmerle, L.B., Rybakov, S., Ibarra, I.L., Holmberg, O., Virshup, I. *et al.* (2022) Squidpy: a scalable framework for spatial omics analysis. *Nat Methods*, **19**, 171-178.
25. Argelaguet, R., Arnol, D., Bredikhin, D., Deloro, Y., Velten, B., Marioni, J.C. and Stegle, O. (2020) MOFA+: a statistical framework for comprehensive integration of multi-modal single-cell data. *Genome Biol*, **21**, 111.
26. Jin, S., Zhang, L. and Nie, Q. (2020) scAI: an unsupervised approach for the integrative analysis of parallel single-cell transcriptomic and epigenomic profiles. *Genome Biol*, **21**, 25.
27. Gayoso, A., Steier, Z., Lopez, R., Regier, J., Nazor, K.L., Streets, A. and Yosef, N. (2021) Joint probabilistic modeling of single-cell multi-omic data with totalVI. *Nat Methods*, **18**, 272-282.
28. Zuo, C. and Chen, L. (2021) Deep-joint-learning analysis model of single cell transcriptome and open chromatin accessibility data. *Brief Bioinform*, **22**.
29. Duren, Z., Chen, X., Zamanighomi, M., Zeng, W., Satpathy, A.T., Chang, H.Y., Wang, Y. and Wong, W.H. (2018) Integrative analysis of single-cell genomics data by coupled nonnegative matrix factorizations. *Proc Natl Acad Sci U S A*, **115**, 7723-7728.
30. Welch, J.D., Kozareva, V., Ferreira, A., Vanderburg, C., Martin, C. and Macosko, E.Z. (2019) Single-Cell Multi-omic Integration Compares and Contrasts Features of Brain Cell Identity. *Cell*, **177**, 1873-1887 e1817.
31. Welch, J.D., Hartemink, A.J. and Prins, J.F. (2017) MATCHER: manifold alignment reveals correspondence between single cell transcriptome and epigenome dynamics. *Genome Biol*, **18**, 138.
32. Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck, W.M., 3rd, Hao, Y., Stoeckius, M., Smibert, P. and Satija, R. (2019) Comprehensive Integration of Single-Cell Data. *Cell*, **177**, 1888-1902 e1821.
33. Kim, H.J., Lin, Y., Geddes, T.A., Yang, J.Y.H. and Yang, P. (2020) CiteFuse enables multi-modal analysis of CITE-seq data. *Bioinformatics*, **36**, 4137-4143.

34. Hao, Y., Hao, S., Andersen-Nissen, E., Mauck, W.M., 3rd, Zheng, S., Butler, A., Lee, M.J., Wilk, A.J., Darby, C., Zager, M. *et al.* (2021) Integrated analysis of multimodal single-cell data. *Cell*, **184**, 3573-3587 e3529.
35. Kleshchevnikov, V., Shmatko, A., Dann, E., Aivazidis, A., King, H.W., Li, T., Elmentaite, R., Lomakin, A., Kedlian, V., Gayoso, A. *et al.* (2022) Cell2location maps fine-grained cell types in spatial transcriptomics. *Nat Biotechnol*, **40**, 661-671.
36. Lopez, R., Li, B., Keren-Shaul, H., Boyeau, P., Kedmi, M., Pilzer, D., Jelinski, A., Yofe, I., David, E., Wagner, A. *et al.* (2022) DestVI identifies continuums of cell types in spatial transcriptomics data. *Nat Biotechnol*, **40**, 1360-1369.
37. Biancalani, T., Scalia, G., Buffoni, L., Avasthi, R., Lu, Z., Sanger, A., Tokcan, N., Vanderburg, C.R., Segerstolpe, A., Zhang, M. *et al.* (2021) Deep learning and alignment of spatially resolved single-cell transcriptomes with Tangram. *Nat Methods*, **18**, 1352-1362.
38. Andersson, A., Bergenstrahle, J., Asp, M., Bergenstrahle, L., Jurek, A., Fernandez Navarro, J. and Lundeberg, J. (2020) Single-cell and spatial transcriptomics enables probabilistic inference of cell type topography. *Commun Biol*, **3**, 565.
39. Cable, D.M., Murray, E., Zou, L.S., Goeva, A., Macosko, E.Z., Chen, F. and Irizarry, R.A. (2021) Robust decomposition of cell type mixtures in spatial transcriptomics. *Nat Biotechnol*.
40. Zhang, Q., Liu, W., Zhang, H.M., Xie, G.Y., Miao, Y.R., Xia, M. and Guo, A.Y. (2020) hTFtarget: A Comprehensive Database for Regulations of Human Transcription Factors and Their Targets. *Genomics Proteomics Bioinformatics*, **18**, 120-128.
41. Lambert, S.A., Jolma, A., Campitelli, L.F., Das, P.K., Yin, Y., Albu, M., Chen, X., Taipale, J., Hughes, T.R. and Weirauch, M.T. (2018) The Human Transcription Factors. *Cell*, **175**, 598-599.
42. Wolf, F.A., Angerer, P. and Theis, F.J. (2018) SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol*, **19**, 15.
43. Virshup, I., Bredikhin, D., Heumos, L., Palla, G., Sturm, G., Gayoso, A., Kats, I., Koutrouli, M., Scverse, C., Berger, B. *et al.* (2023) The scverse project provides a computational ecosystem for single-cell omics data analysis. *Nat Biotechnol*, **41**, 604-606.
44. Hainmueller, J. and Hazlett, C. (2014) Kernel Regularized Least Squares: Reducing Misspecification Bias with a Flexible and Interpretable Machine Learning Approach. *Political Analysis*, **22**, 143-168.
45. Long, Y., Ang, K.S., Li, M., Chong, K.L.K., Sethi, R., Zhong, C., Xu, H., Ong, Z., Sachaphibulkij, K., Chen, A. *et al.* (2023) Spatially informed clustering, integration, and deconvolution of spatial transcriptomics with GraphST. *Nat Commun*, **14**, 1155.
46. Shao, X., Liao, J., Li, C., Lu, X., Cheng, J. and Fan, X. (2021) CellTalkDB: a manually curated database of ligand-receptor interactions in humans and mice. *Brief Bioinform*, **22**.
47. Badia, I.M.P., Velez Santiago, J., Braunger, J., Geiss, C., Dimitrov, D., Muller-Dott, S., Taus, P., Dugourd, A., Holland, C.H., Ramirez Flores, R.O. *et al.* (2022) decoupleR: ensemble of computational methods to infer biological activities from omics data. *Bioinform Adv*, **2**, vbac016.
48. Hu, B., Wang, Q., Wang, Y.A., Hua, S., Sauve, C.G., Ong, D., Lan, Z.D., Chang, Q., Ho, Y.W., Monasterio, M.M. *et al.* (2016) Epigenetic Activation of WNT5A Drives Glioblastoma Stem Cell Differentiation and Invasive Growth. *Cell*, **167**, 1281-1295 e1218.
49. Hu, Q., Xu, T., Zhang, W. and Huang, C. (2022) Bach2 regulates B cell survival to maintain germinal centers and promote B cell memory. *Biochem Biophys Res Commun*, **618**, 86-92.
50. Ohl, K., Schippers, A. and Tenbrock, K. (2018) CD11c-Specific Deletion Reveals CREB as a Critical Regulator of DC Function during the Germinal Center Response. *J Immunol Res*, **2018**, 8947230.
51. Beguelin, W., Rivas, M.A., Calvo Fernandez, M.T., Teater, M., Purwada, A., Redmond, D., Shen, H., Challman, M.F., Elemento, O., Singh, A. *et al.* (2017) EZH2 enables germinal centre formation through epigenetic silencing of CDKN1A and an Rb-E2F1 feedback loop. *Nat Commun*, **8**, 877.
52. Deng, G., Song, X. and Greene, M.I. (2020) FoxP3 in T(reg) cell biology: a molecular and structural perspective. *Clin Exp Immunol*, **199**, 255-262.
53. Hu, X., Li, J., Fu, M., Zhao, X. and Wang, W. (2021) The JAK/STAT signaling pathway: from bench to clinic. *Signal Transduct Target Ther*, **6**, 402.
54. Taveirne, S., Wahlen, S., Van Loocke, W., Kiekens, L., Persyn, E., Van Ammel, E., De Mulder, K., Roels, J., Tilleman, L., Aumercier, M. *et al.* (2020) The transcription factor ETS1 is an important regulator of human NK cell development and terminal differentiation. *Blood*, **136**, 288-298.

55. Oikawa, T., Yamada, T., Kihara-Negishi, F., Yamamoto, H., Kondoh, N., Hitomi, Y. and Hashimoto, Y. (1999) The role of Ets family transcription factor PU.1 in hematopoietic cell differentiation, proliferation and apoptosis. *Cell Death Differ*, **6**, 599-608.
56. Huang, Q., Liu, J., Wu, S., Zhang, X., Xiao, Z., Liu, Z. and Du, W. (2021) Spi-B Promotes the Recruitment of Tumor-Associated Macrophages via Enhancing CCL4 Expression in Lung Cancer. *Front Oncol*, **11**, 659131.
57. Igarashi, K., Kurosaki, T. and Roychoudhuri, R. (2017) BACH transcription factors in innate and adaptive immunity. *Nat Rev Immunol*, **17**, 437-450.
58. Minamino, K., Takahara, K., Adachi, T., Nagaoka, K., Iyoda, T., Taki, S. and Inaba, K. (2012) IRF-2 regulates B-cell proliferation and antibody production through distinct mechanisms. *Int Immunol*, **24**, 573-581.
59. Berezovsky, A.D., Poisson, L.M., Cherba, D., Webb, C.P., Transou, A.D., Lemke, N.W., Hong, X., Hasselbach, L.A., Irtenkauf, S.M., Mikkelsen, T. *et al.* (2014) Sox2 promotes malignancy in glioblastoma by regulating plasticity and astrocytic differentiation. *Neoplasia*, **16**, 193-206, 206 e119-125.
60. Cao, J., Zhao, M., Liu, J., Zhang, X., Pei, Y., Wang, J., Yang, X., Shen, B. and Zhang, J. (2019) RACK1 Promotes Self-Renewal and Chemoresistance of Cancer Stem Cells in Human Hepatocellular Carcinoma through Stabilizing Nanog. *Theranostics*, **9**, 811-828.
61. Zhang, D., Deng, Y., Kukanja, P., Agirre, E., Bartosovic, M., Dong, M., Ma, C., Ma, S., Su, G., Bao, S. *et al.* (2023) Spatial epigenome-transcriptome co-profiling of mammalian tissues. *Nature*, **616**, 113-122.
62. Tao, Y., Ma, X., Palmer, D., Schwartz, R., Lu, X. and Osmanbeyoglu, H.U. (2022) Interpretable deep learning for chromatin-informed inference of transcriptional programs driven by somatic alterations across cancers. *Nucleic Acids Res*, **50**, 10869-10881.
63. Osmanbeyoglu, H.U., Shimizu, F., Rynne-Vidal, A., Alonso-Curbelo, D., Chen, H.A., Wen, H.Y., Yeung, T.L., Jelinic, P., Razavi, P., Lowe, S.W. *et al.* (2019) Chromatin-informed inference of transcriptional programs in gynecologic and basal breast cancers. *Nat Commun*, **10**, 4369.

FIGURES

Figure 1: Overview of the STAN workflow. (A) As input, our model accepts input as a normalized gene expression matrix, spatial locations, and pixel intensity of the corresponding image for each spot, along with a binary matrix indicating the genes each TF targets as obtained using curated database TFs. Spot-specific TF activity levels are inferred by training a spatially regularized linear regression model to predict gene expression using TF-target gene relationships. **(B)** Deconvolution analysis infers spot-specific cell type/state proportions based on reference scRNA-seq data (signature matrix). Downstream analysis includes **(C)** inferring cell-type-specific TFs, **(D)** identifying TF activity patterns across distinct pathological regions and spatial domains, and **(E)** identifying ligands/receptors whose gene expression correlates with TF activity in neighboring spots.

Figure 2: STAN identifies GCs and cell-type-specific TFs in a human lymph node. (A) STAN and Ridge regression models (without spatial information) predict spot-specific expression of held-out genes. Pearson correlations between predicted and actual gene expression for all spots are shown. For each method and each sample, the correlation was computed using 10-fold cross-validation of held-out genes. Each point represents one spot in the dataset and is colored based on the spot's UMI count. STAN outperforms Ridge for all spots (dots) above the grey diagonal line. **(B)** UMAP (left) and spatial plot (right) of spots clustered based on inferred TF activity using the Leiden algorithm. **(C)** Locations of GCs. **(D)** Inferred TF activity of the top 3 ranked TFs in the GCs (outlined). **(E)** mRNA level of the top 3 ranked TFs in the GCs (outlined). **(F)** Spatial maps of spot-level proportions of regular T cells (left), expression of *FOXP3* (2nd column), inferred activity of FOXP3 using STAN (3rd column), and inferred activity of STAT2 using decoupleR (right). **(G)** Spatial maps of spot-level proportions of IFN-β cells (left), expression of *STAT2* (2nd column), inferred activity of STAT2 using STAN (3rd column), and inferred activity of STAT2 using decoupleR (right). **(H)** Median TF activity scores of TFs across different cell types/states are shown. A TF-cell type pair was selected if the TF activity score was >4, the adjusted *P*-value was <1e-2, and r^2 was > 0.4 for the cell type.

Figure 3: TF activity patterns across distinct pathological regions in breast cancer. (A) The H&E-stained image (left) and pathological annotation (right) of a triple negative breast cancer sample from Wu et al. (18) (Sample ID: TNBC_3). **(B)** The mean inferred TF activity differences between a given pathological region [stroma regions (top), lymphocyte regions (middle), and invasive cancer + stroma + lymphocyte regions (bottom)] vs. all other regions across samples. The dot size indicates $-\log_{10}$(FDR) by Wilcoxon rank-sum test. **(C)** Heatmaps showing Pearson correlation coefficients between mean TF activities in all applicable samples for stroma (top), lymphocytes (bottom left), and invasive cancer + stroma + lymphocytes (bottom right). **(D)** The mean inferred TF activity differences between a given pathological region vs. all other regions are shown. The dot size indicates $-\log_{10}$(FDR) by Wilcoxon rank-sum test. Rows and columns are hierarchically clustered. The list of TFs includes the top 5 TFs ranked by *P*-value associated with each pathological region. **(E)** Spatial feature plots showing FOXP3, FOXA1, IRF2, and STAT2 TF activities (top) and mRNA levels (bottom). **(F)** Correlation between cell-type-specific scores based on ST (inferred by STAN) and scRNA-seq (inferred by Ridge regression).

Figure 4: STAN links ligands/receptors to TFs in glioblastoma. (A) Clustered heatmap showing the Pearson correlation coefficient between inferred TF activities and ligand/receptor expression. **(B)** Spatial feature plots displaying mRNA expression of *VEGFA* and *CD44* (ligands) (top), inferred activity of *HIF1A* and *SOX2* (middle), and mRNA expression of *RACK1* and *VIM* (receptors) (bottom). **(C)** Representative images of glioblastoma patient tumors with co-staining for SOX2, CD44, and VIM.

Figure 1

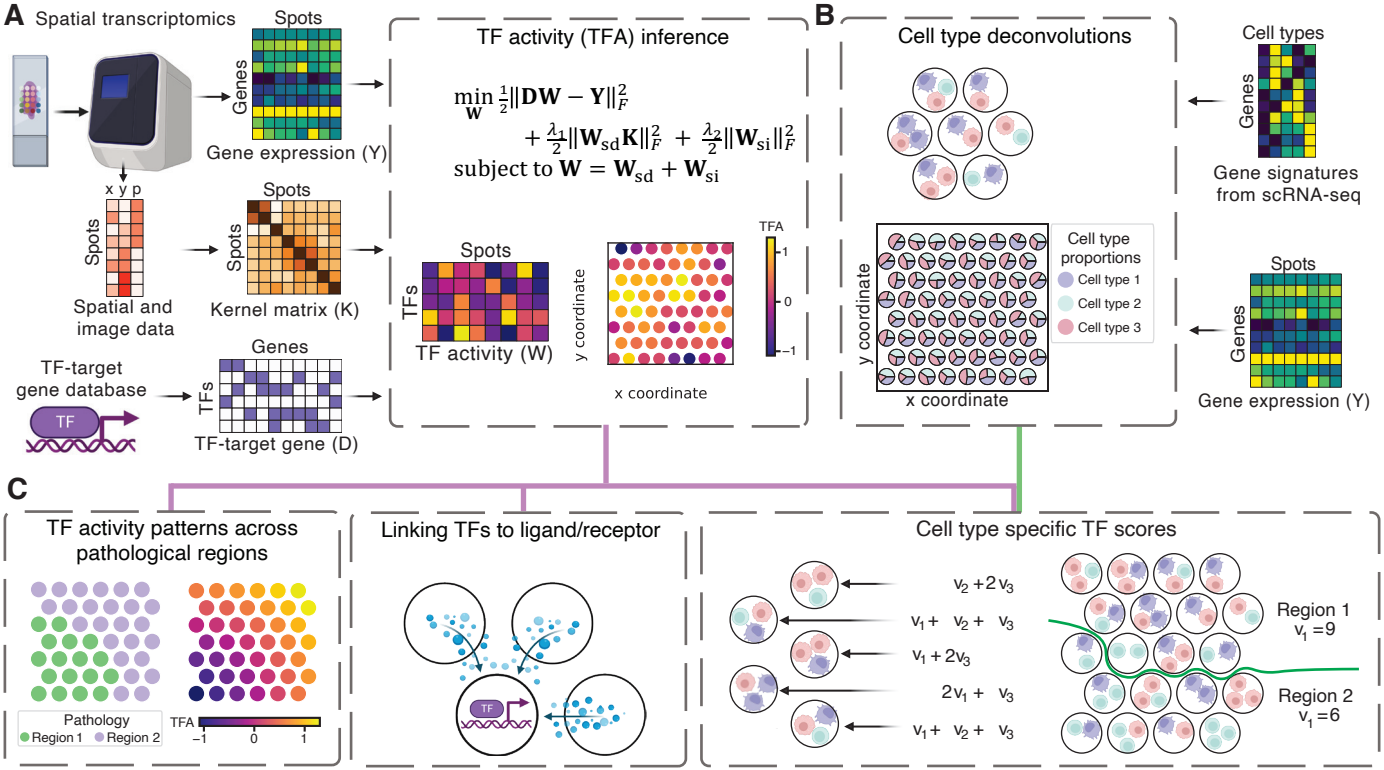

A
Spatial transcriptomics
Spots
Genes
Gene expression (Y)
x y p
Spots
Spatial and image data
Spots
Spots
Kernel matrix (K)
TF-target gene database
TF
Genes
TFs
TF-target gene (D)
TF activity (TFA) inference
$\min_{\mathbf{W}} \frac{1}{2}\|\mathbf{DW}-\mathbf{Y}\|_F^2 + \frac{\lambda_1}{2}\|\mathbf{W}_{sd}\mathbf{K}\|_F^2 + \frac{\lambda_2}{2}\|\mathbf{W}_{si}\|_F^2$
subject to $\mathbf{W} = \mathbf{W}_{sd} + \mathbf{W}_{si}$
Spots
TFs
TF activity (W)
y coordinate
x coordinate
TFA
1
0
−1
B
Cell type deconvolutions
Cell types
Genes
Gene signatures from scRNA-seq
y coordinate
x coordinate
Cell type proportions
Cell type 1
Cell type 2
Cell type 3
Spots
Genes
Gene expression (Y)
C
TF activity patterns across pathological regions
Pathology
Region 1
Region 2
TFA
−1
0
1
Linking TFs to ligand/receptor
TF
Cell type specific TF scores
$v_2 + 2v_3$
$v_1 + v_2 + v_3$
$v_1 + 2v_3$
$2v_1 + v_3$
$v_1 + v_2 + v_3$
Region 1
$v_1 = 9$
Region 2
$v_1 = 6$

Figure 2

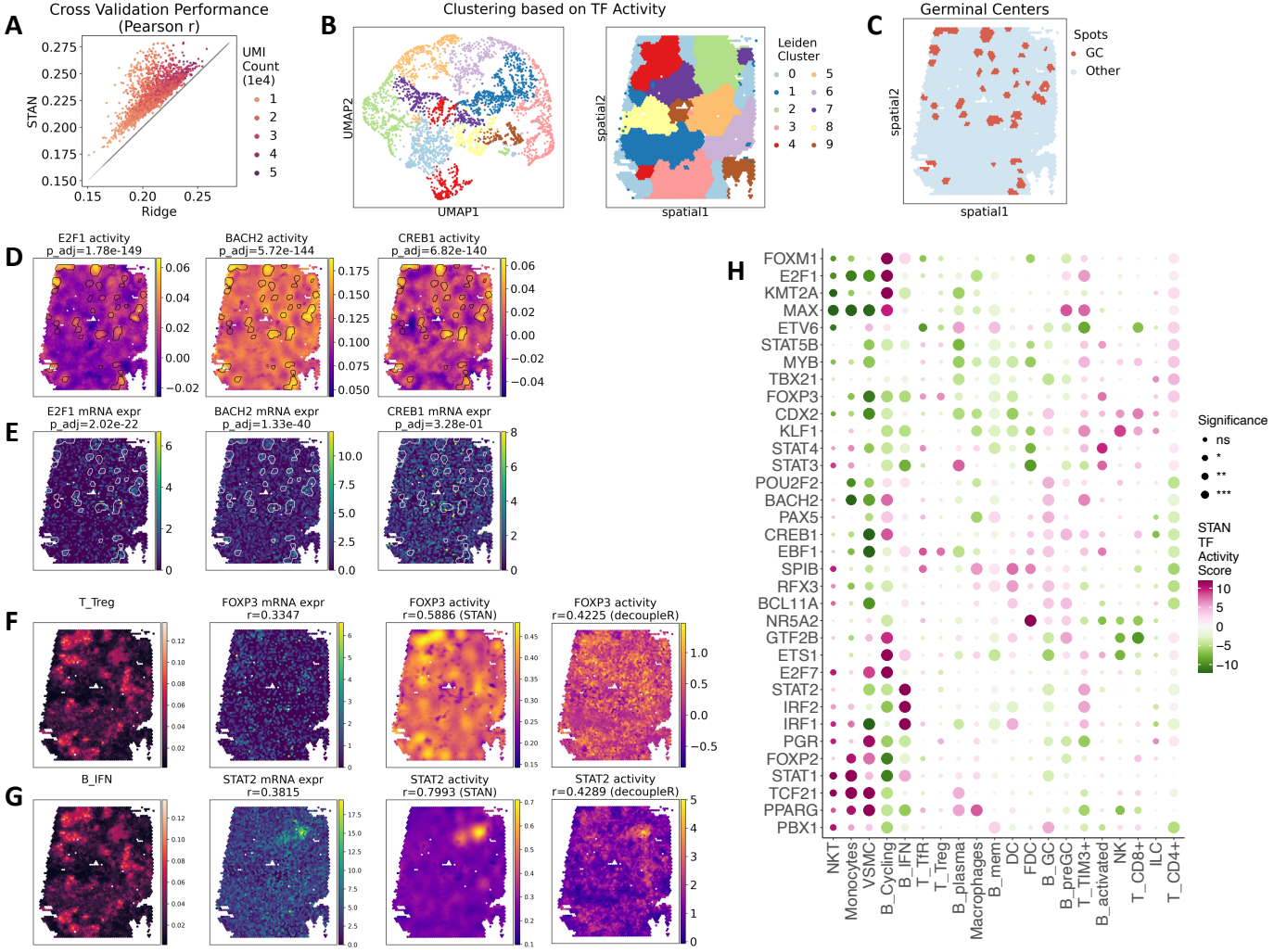

A
Cross Validation Performance
(Pearson r)
STAN
Ridge
UMI Count (1e4)
B
Clustering based on TF Activity
UMAP1
UMAP2
spatial1
spatial2
Leiden Cluster
C
Germinal Centers
Spots
GC
Other
D
E2F1 activity
p_adj=1.78e-149
BACH2 activity
p_adj=5.72e-144
CREB1 activity
p_adj=6.82e-140
E
E2F1 mRNA expr
p_adj=2.02e-22
BACH2 mRNA expr
p_adj=1.33e-40
CREB1 mRNA expr
p_adj=3.28e-01
F
T_Treg
FOXP3 mRNA expr
r=0.3347
FOXP3 activity
r=0.5886 (STAN)
FOXP3 activity
r=0.4225 (decoupleR)
G
B_IFN
STAT2 mRNA expr
r=0.3815
STAT2 activity
r=0.7993 (STAN)
STAT2 activity
r=0.4289 (decoupleR)
H
FOXM1
E2F1
KMT2A
MAX
ETV6
STAT5B
MYB
TBX21
FOXP3
CDX2
KLF1
STAT4
STAT3
POU2F2
BACH2
PAX5
CREB1
EBF1
SPIB
RFX3
BCL11A
NR5A2
GTF2B
ETS1
E2F7
STAT2
IRF2
IRF1
PGR
FOXP2
STAT1
TCF21
PPARG
PBX1
NKT
Monocytes
VSMC
B_Cycling
B_IFN
T_TfR
T_Treg
B_plasma
Macrophages
B_mem
DC
FDC
B_GC
B_preGC
T_TIM3+
B_activated
NK
T_CD8+
ILC
T_CD4+
Significance
ns
*
**
***
STAN TF Activity Score

Figure 3

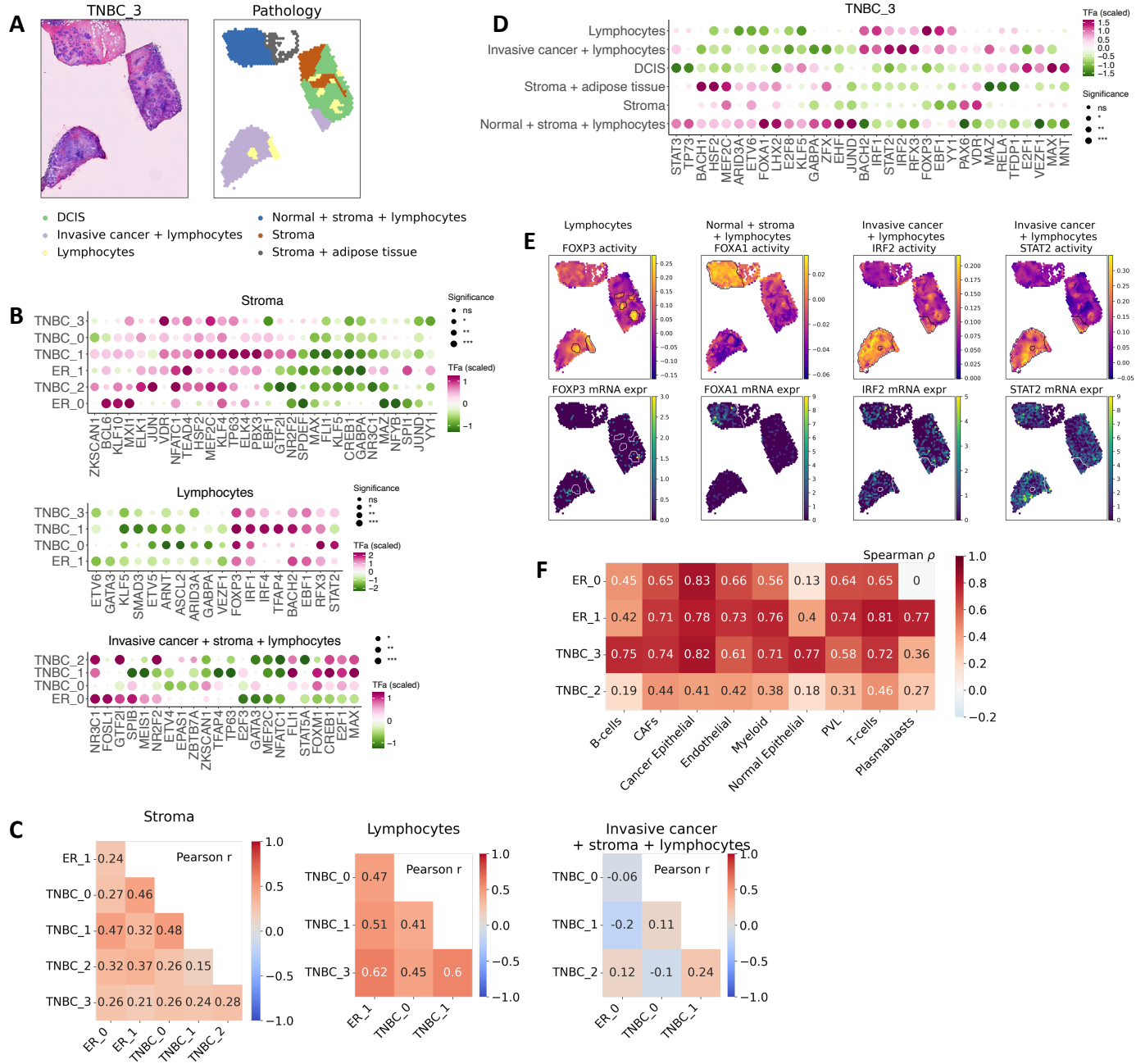
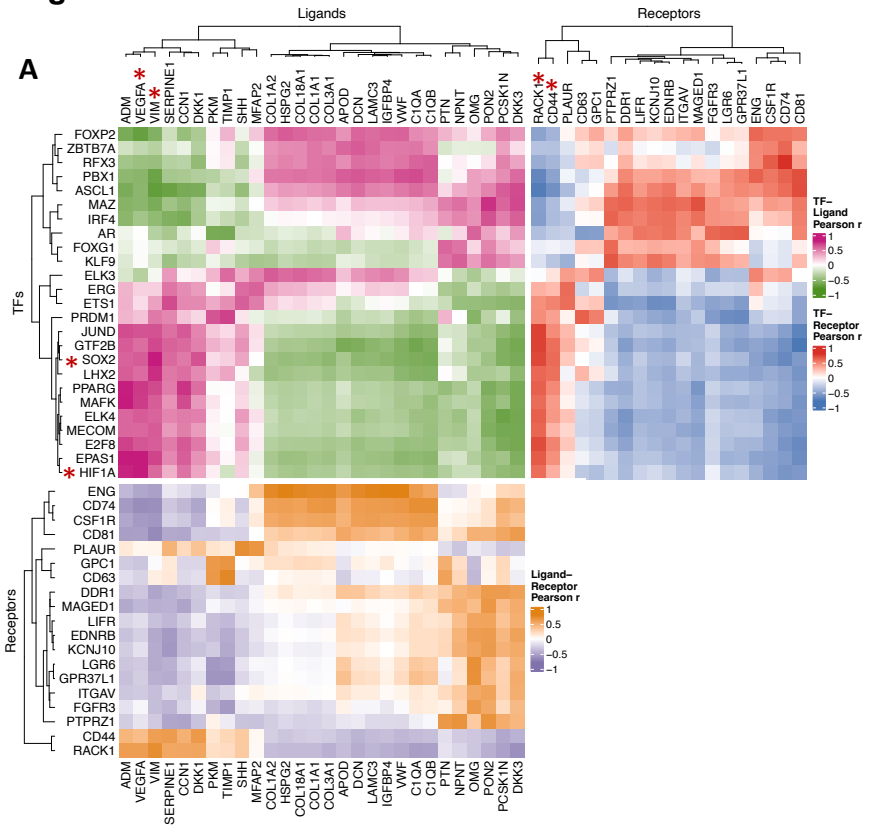
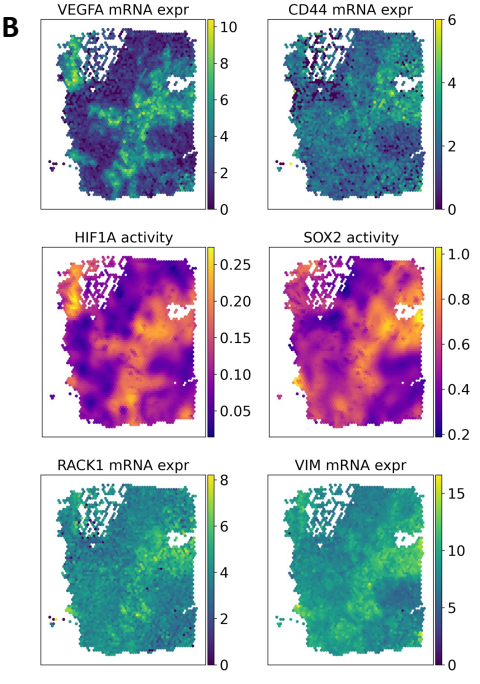
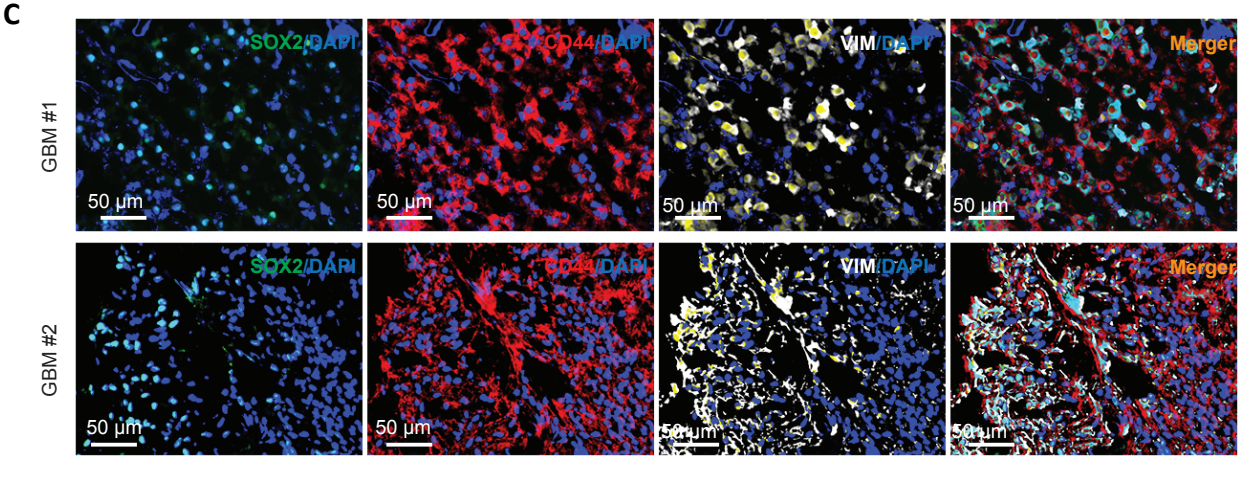

A
TNBC_3
Pathology
DCIS
Invasive cancer + lymphocytes
Lymphocytes
Normal + stroma + lymphocytes
Stroma
Stroma + adipose tissue
B
Stroma
TNBC_3
TNBC_0
TNBC_1
ER_1
TNBC_2
ER_0
ZKSCAN1 BCL6 KLF10 MXI1 ELK1 JUN VDR NFATC1 TEAD4 HSF2 MEF2C KLF4 TP63 ELK4 PBX3 EBF1 GTF2I NR2F2 SPDEF MAX FLI1 KLF5 CREB1 GABPA NR3C1 MAZ NFYB SP1 JUND YY1
Significance
ns
*
**
***
TFa (scaled)
Lymphocytes
ETV6 GATA3 KLF5 SMAD3 ETV5 ARNT ASCL2 ARID3A GABPA VEZF1 FOXP3 IRF1 IRF4 TFAP4 BACH2 EBF1 RFX3 STAT2
Invasive cancer + stroma + lymphocytes
NR3C1 FOSL1 GTF2I SPIB MEIS1 NR2F2 ETV4 EPAS1 ZBTB7A ZKSCAN1 TFAP4 TP63 E2F3 GATA3 MEF2C NFATC1 FLI1 STAT5A FOXM1 CREB1 E2F1 MAX
C
Stroma
Pearson r
ER_1 0.24
TNBC_0 0.27 0.46
TNBC_1 0.47 0.32 0.48
TNBC_2 0.32 0.37 0.26 0.15
TNBC_3 0.26 0.21 0.26 0.24 0.28
ER_0 ER_1 TNBC_0 TNBC_1 TNBC_2
Lymphocytes
TNBC_0 0.47
TNBC_1 0.51 0.41
TNBC_3 0.62 0.45 0.6
ER_1 TNBC_0 TNBC_1
Invasive cancer + stroma + lymphocytes
TNBC_0 -0.06
TNBC_1 -0.2 0.11
TNBC_2 0.12 -0.1 0.24
ER_0 TNBC_0 TNBC_1
D
TNBC_3
Lymphocytes
Invasive cancer + lymphocytes
DCIS
Stroma + adipose tissue
Stroma
Normal + stroma + lymphocytes
STAT3 TP73 BACH1 HSF2 MEF2C ARID3A ETV6 FOXA1 LHX2 E2F8 KLF5 GABPA ZFX EHF JUND BACH2 IRF1 STAT2 IRF2 RFX3 FOXP3 EBF1 YY1 PAX6 VDR MAZ RELA TFDP1 E2F1 VEZF1 MAX MNT
TFa (scaled)
Significance
ns
*
**
***
E
Lymphocytes
FOXP3 activity
Normal + stroma + lymphocytes
FOXA1 activity
Invasive cancer + lymphocytes
IRF2 activity
Invasive cancer + lymphocytes
STAT2 activity
FOXP3 mRNA expr
FOXA1 mRNA expr
IRF2 mRNA expr
STAT2 mRNA expr
F
Spearman ρ
ER_0 0.45 0.65 0.83 0.66 0.56 0.13 0.64 0.65 0
ER_1 0.42 0.71 0.78 0.73 0.76 0.4 0.74 0.81 0.77
TNBC_3 0.75 0.74 0.82 0.61 0.71 0.77 0.58 0.72 0.36
TNBC_2 0.19 0.44 0.41 0.42 0.38 0.18 0.31 0.46 0.27
B-cells CAFs Cancer Epithelial Endothelial Myeloid Normal Epithelial PVL T-cells Plasmablasts

Figure 4


A
Ligands
Receptors
TFs
Receptors
TF–Ligand Pearson r
TF–Receptor Pearson r
Ligand–Receptor Pearson r
ADM VEGFA VIM SERPINE1 CCN1 DKK1 PKM TIMP1 SHH MFAP2 COL1A2 HSPG2 COL18A1 COL1A1 COL3A1 APOD DCN LAMC3 IGFBP4 VWF C1QA C1QB PTN NPNT OMG PON2 PCSK1N DKK3
RACK1 CD44 PLAUR CD63 GPC1 PTPRZ1 DDR1 LIFR KCNJ10 EDNRB ITGAV MAGED1 FGFR3 LGR6 GPR37L1 ENG CSF1R CD74 CD81
FOXP2 ZBTB7A RFX3 PBX1 ASCL1 MAZ IRF4 AR FOXG1 KLF9 ELK3 ERG ETS1 PRDM1 JUND GTF2B SOX2 LHX2 PPARG MAFK ELK4 MECOM E2F8 EPAS1 HIF1A
ENG CD74 CSF1R CD81 PLAUR GPC1 CD63 DDR1 MAGED1 LIFR EDNRB KCNJ10 LGR6 GPR37L1 ITGAV FGFR3 PTPRZ1 CD44 RACK1
ADM VEGFA VIM SERPINE1 CCN1 DKK1 PKM TIMP1 SHH MFAP2 COL1A2 HSPG2 COL18A1 COL1A1 COL3A1 APOD DCN LAMC3 IGFBP4 VWF C1QA C1QB PTN NPNT OMG PON2 PCSK1N DKK3


B
VEGFA mRNA expr
CD44 mRNA expr
HIF1A activity
SOX2 activity
RACK1 mRNA expr
VIM mRNA expr


C
GBM #1
GBM #2
SOX2/DAPI
CD44/DAPI
VIM/DAPI
Merger
50 µm