Reconstructing Spatial Transcriptomics at the Single-cell Resolution with BayesDeep

Xi Jiang[1,2], Lei Dong[1], Shidan Wang[1], Zhuoyu Wen[1], Mingyi Chen[3], Lin Xu[1,4] *, Guanghua Xiao[1] * and Qiwei Li[5] *

[1] Quantitative Biomedical Research Center, Peter O'Donnell Jr. School of Public Health, The University of Texas Southwestern Medical Center, Dallas, Texas, U.S.A.

[2] Department of Statistics and Data Science, Southern Methodist University, Dallas, Texas, U.S.A.

[3] Department of Pathology, The University of Texas Southwestern Medical Center, Dallas, Texas, U.S.A

[4] Department of Pediatrics, Division of Hematology/Oncology, University of Texas Southwestern Medical Center, Dallas, Texas, U.S.A

[5] Department of Mathematical Sciences, The University of Texas at Dallas, Richardson, Texas, U.S.A

*Correspondence to:

Lin Xu. Email: lin.xu@utsouthwestern.edu

Guanghua Xiao. Email: guanghua.xiao@utsouthwestern.edu

Qiwei Li. Email: qiwei.li@utdallas.edu

Conflict of Interests: The authors declare no potential conflicts of interest.

Abstract

Spatially resolved transcriptomics (SRT) techniques have revolutionized the characterization of molecular profiles while preserving spatial and morphological context. However, most next-generation sequencing-based SRT techniques are limited to measuring gene expression in a confined array of spots, capturing only a fraction of the spatial domain. Typically, these spots encompass gene expression from a few to hundreds of cells, underscoring a critical need for more detailed, single-cell resolution SRT data to enhance our understanding of biological functions within the tissue context. Addressing this challenge, we introduce BayesDeep, a novel Bayesian hierarchical model that leverages cellular morphological data from histology images, commonly paired with SRT data, to reconstruct SRT data at the single-cell resolution. BayesDeep effectively model count data from SRT studies *via* a negative binomial regression model. This model incorporates explanatory variables such as cell types and nuclei-shape information for each cell extracted from the paired histology image. A feature selection scheme is integrated to examine the association between the morphological and molecular profiles, thereby improving the model robustness. We applied BayesDeep to two real SRT datasets, successfully demonstrating its capability to reconstruct SRT data at the single-cell resolution. This advancement not only yields new biological insights but also significantly enhances various downstream analyses, such as pseudotime and cell-cell communication.

Keywords

Spatially resolved transcriptomics; Cellular morphological features; Single-cell resolution; Bayesian hierarchical model.

INTRODUCTION

Understanding the spatial distribution of transcript expression provides valuable insights into biological function and histopathology[1]. The advent of single-cell RNA sequencing (scRNA-seq) techniques has profoundly transformed our understanding of gene expression regulation in different cell lineages or types. However, tissue dissociation in scRNA-seq leads to the loss of the spatial context of gene expression, which is essential for comprehending the cell-cell and cell-environment interaction mechanism[2]. Recent advancements in spatially resolved transcriptomics (SRT) have enabled the exploration of cellular transcription in conjunction with the spatial organization and morphological features of cells. Next-generation sequencing (NGS)-based spatial molecular profiling platforms, such as Spatial Transcriptomics[3] (ST) and 10x Visium[4] (i.e., an improved platform by 10x Genomics), utilize spatial barcodes to capture RNA molecules, synthesize their complementary DNA molecules, and subsequently perform sequencing. These techniques facilitate the measurement of whole-genome expression across thousands of spatial locations, referred to as 'spots', on the tissue section. The development of SRT technology has provided valuable clinical and biological insights into various areas, including tumor heterogeneity, brain function, and sepsis pathophysiology[5-7].

One of the primary constraints inherent to NGS-based SRT techniques is their spatial resolution. Gene expression in these platforms is measured on an array of spots, with a typical diameter of 100 μm for ST and 55 μm for 10x Visium[8]. Consequently, depending on the platform and the type of biological tissue being analyzed, each spot's area may encompass a heterogeneous population of cells, ranging from a few to as many as 200 cells. Furthermore, SRT techniques solely measure gene expression within the area of the confined array spots, thus covering part of the total cells observed within the entire tissue section. To illustrate, the area covered by measured spots accounts

for approximately 1/3 of the total tissue region for the 10x Visium platform and around 20% for ST platforms (see Supplementary Figure S1 for a detailed explanation). The inherent limitation in spatial resolution and the scope of regions could lead to substantial information loss when analyzing SRT data. Such limitations may, in turn, impact the applicability of these techniques in facilitating the discovery of deeper insights in various biomedical studies, specifically in investigating gene expression patterns at the cellular level.

Several computational methods have been developed to enhance the spatial resolution of SRT data, addressing these limitations. For instance, BayesSpace[8] employs a Bayesian approach to improve gene expression at the sub-spot resolution. However, despite this advancement, sub-spots generated by BayesSpace may still encompass multiple cells, yielding only a modest improvement in spatial resolution. Furthermore, BayesSpace cannot predict gene expression outside the spotted regions. In parallel, a series of approaches, such as TESLA[9], XFuse[10], ST-Net[11], and HisToGene[12], have been developed to reconstruct gene expression at super-resolution. These deep-learning-based techniques utilize molecular information from SRT data and morphological information from the paired histology image. Notably, while these methods enhance gene expression prediction at the pixel level, which may complicate interpretation, none directly account for cellular spatial organization in their analyses.

In response, we introduce a novel Bayesian methodology, BayesDeep, for deeply resolving gene expression for all "real" cells by integrating the molecular profile from SRT data and the morphological information extracted from the paired histology image. Specifically, BayesDeep builds upon a regularized negative binomial regression model with grouped observations. The response variable is the spot-resolution gene expression measurements in terms of counts; and the explanatory variables are a range of cellular features extracted from the paired histology image,

including cell type and nuclei-shape descriptors. Following the estimation of regression coefficients, BayesDeep predicts the gene expression of all cells based on their cellular features, regardless of whether they are within or beyond spot regions. The model robustness is achieved by regularization using a spike-and-slab prior distribution to each regression coefficient. We validated the accuracy of gene expression prediction at both spot and single-cell resolution using simulated and real SRT datasets. Additionally, we demonstrated that the single-cell-resolution spatial molecular profiles characterized by BayesDeep enable in-depth investigations on cell-type specific differential expression analysis, cell-cell communication analysis, and pseudo-time analysis. Overall, our findings highlight the effectiveness of BayesDeep in reconstructing molecular profiles at the single-cell resolution from NGS-based SRT data, which is typically limited to spot resolution, paving the way for advanced research in pseudotime and cell-cell communication analysis.

RESULTS

Overview of BayesDeep

BayesDeep builds upon a Bayesian negative binomial regression model to recover gene expression at the single-cell resolution from NGS-based SRT data. The schematic diagram is depicted in Figure 1.

BayesDeep integrates three distinct modalities from a standard NGS-based SRT experiment: the molecular, image, and geospatial profiles (see Figure 1A). The molecular profile refers to the spot-resolution gene expression data denoted by an N-by-P count matrix $\boldsymbol{Y}$, where N is the number of spots and P is the number of genes. The image profile corresponds to the detailed morphological

context of the paired histology image in terms of a set of cellular features. We use an M-by-L design matrix X to denote the image profile, where M is the number of all observed cells and L is the number of cellular features, which may include cell types, nuclei-shape characteristics, and any other relevant explanatory features that can be quantified at large scale. The geospatial profile reveals the spatial relationship between the N spots and M cells, which can be defined by an N-by-M binary matrix G, with one signifying that a cell is within the barcoded area of a spot.

The spot-resolution gene expression matrix Y and the single-cell-resolution morphological features of those cells within spot regions serve as a reference for recovering the single-cell-resolution gene expression of all M cells, whether within or beyond spot regions. The model is specified in Figure 1B. We first modeled the observed read count for a specific gene within a spot using a negative binomial (NB) distribution. Then, the underlying spot-resolution relative gene expression in the NB mean is assumed to be the average of single-cell-resolution relative expression across all cells within the spot. Next, we considered the logarithm of each cell's relative expression as a linear combination of covariates that includes a scalar of one for the intercept and L measurable explanatory variables that pertain to that cell. A spike-and-slab prior model is applied for each covariate coefficient. On one hand, this feature selection scheme improves the robustness of our model. On the other hand, the corresponding coefficient matrix B uncovers significant associations between gene expression and cellular characteristics, illustrated in Figure 1C, thereby potentially offering valuable biological insights. With the reconstructed single-cell-resolution spatial molecular profile Θ, we can undertake several pivotal downstream analyses, as depicted in Figure 1D. These analyses allow for the differentiation of cell populations, the exploration of tumorigenesis through pseudotime analysis, and the dissection of ligand-receptor signaling pathways vital for cell-cell communication.

Model Validation

<u>Model validation on simulated data.</u> We designed a simulation study to validate the accuracy of BayesDeep. We selected a connected region of $N = 500$ spots from the human breast cancer 10x Visium data (displayed as green circles in Figure 2A) to generate the spot-resolution molecular profile. The single-cell-resolution image profile X, including cell types and nuclei-shape features and the locations of cells, were from the nuclei identification results by HD-staining[13] for the paired histology image of the SRT data. As introduced in the Method section and observed in real SRT data analysis, the covariate coefficient matrix B is highly sparse, indicating that many explanatory variables do not contribute to gene expression for most genes. To replicate this condition, we generated the coefficients, ensuring nearly half were zeros, reflecting the observed zero-inflation. Based on the model assumption of BayesDeep, we first generated single-cell-resolution relative gene expression matrix Θ on the selected region, and then the spot-resolution gene expression matrix Y on all spots within. The data-generating procedure is detailed in section S1 in Supplementary Materials.

We evaluated the performance of BayesDeep in recovering the single-cell-resolution molecular profile Θ, against two other methods, TESLA[9] and Gaussian process (GP)[14], by measuring the Pearson correlation coefficient $\rho \in [-1,1]$ between the logged estimated and actual gene expression at the single-cell resolution, $\hat{\theta}_{mj}$'s and θ_{mj}'s. Using relative gene expression allows us to discern which cells exhibit high gene expression and which demonstrate lower expression. A high positive ρ value signifies the successful model outcome and robust overall performance. As depicted in Figure 2A, the results demonstrate that BayesDeep, when employing regularization, significantly surpasses the comparative methods, with a median ρ of 0.928. In contrast, TESLA and GP failed to reconstruct the single-cell-resolution molecular profile Θ, as indicated by their

respective ρ values approaching zero. To further examine the efficacy of the regularization approach, we computed the root mean square error (RMSE) between the estimated and actual covariate coefficients, $\hat{\beta}_{lj}$'s and β_{lj}'s, with lower RMSE values denoting more precise estimations. Specifically, in scenarios where regularization was omitted, a standard normal prior was applied in place of the spike-and-slab prior to each covariate coefficient β_{lj}. These results demonstrate that the incorporation of regularization significantly refines the precision of estimations for non-zero coefficients and substantially improves accuracy for truly zero coefficients (Supplementary Figure S2). These findings support the implementation of a regularized prior within our BayesDeep framework.

Model validation on real SRT data at the spot resolution. In this study, we utilized the human breast cancer[15] and human prostate cancer 10x Visium data[16] in the real data analysis to demonstrate the efficacy of BayesDeep in restoring gene expression within manually masked spots. For each SRT data, we randomly selected $P = 100$ genes with non-zero read counts in at least 50% spots. We assessed BayesDeep against TESLA and GP, through two distinct masking scenarios for spot-resolution gene expression imputation. In scenario 1, 40% of spots were randomly masked, while in scenario 2, we divided the domain into two halves and masked all spots on one side (refer to Figure 2B, with masked spots marked in orange and unmasked spots marked in green). The second scenario aimed to challenge the model's predictive power over a separate region that does not overlap with those masked spots. The gene expression of those remaining unmasked spots Y served as the main input of each method. For BayesDeep, we also incorporated the morphological information X, detailed in the Method section, from the corresponding histology image.

Figure 2B shows the correlation coefficient ρ for comparing observed and imputed spot-resolution gene expression in the masked spots within the human breast cancer 10x Visium data. BayesDeep

achieved a median correlation of 0.585 in scenario 1, significantly higher than that of the competing methods (TESLA: $\rho = 0.358$; GP: $\rho = 0.391$). In scenario 2, the performance of BayesDeep is comparable to scenario 1, whereas TESLA and GP failed to impute gene expression for the entirely masked half, as evidenced by negligible correlation values. This failure is attributed to their spatial dependence assumption in reconstructing gene expression, resulting in inadequate performance when the masked and observed regions do not intersect. Supplementary Figure S3 shows similar results through the analysis on human prostate cancer 10x Visium data. These results show the capability of BayesDeep to predict gene expression at the spot resolution, both in the same and new regions, by leveraging single-cell-resolution morphological features.

Model validation on real SRT data at the single-cell resolution. We extended our study to assess the performance of BayesDeep on the prediction of single-cell-resolution gene expression when ground truth is available at the single-cell resolution. We validated the model using the mouse visual cortex STARmap data[17], which was collected by an imaging-based SRT platform. It comprises measurements of 1,020 genes across $M = 1,207$ cells, categorized into $Q = 15$ distinct cell types distributed across seven layers labeled in the original study. In our analysis, we filtered out genes that are expressed in less than 30% cells or the highest read count is less than ten. With these filtering criteria, we kept $P = 77$ genes for the following analysis. We then constructed the spot-resolution molecular profile Y by overlaying a square lattice grid across the entire domain, with each square unit representing a "spot" of 750×750 pixels, as illustrated in Figure 2C. This yielded $N = 105$ spots, each containing more than one cell, covering the entire domain without inter-spot gaps. The read count y_{ij} of gene j for spot i was calculated by aggregating the read counts from all cells within that spot. Owing to the absence of a corresponding

histology image for the data, we utilized the cell type and layer information provided along with the SRT data in the original study as explanatory variables X to inform BayesDeep.

Figure 2C visualizes the correlation coefficient ρ between the imputed and true relative expression across all cells for all genes. BayesDeep achieved a median correlation of approximately 0.336, similar to TESLA (median $\rho = 0.331$) but significantly outperforming GP (median $\rho = 0.065$). Additionally, Figure 2C illustrates the actual *versus* predicted relative expression for four representative genes, demonstrating a notable agreement between the actual and predicted gene expression patterns. These findings affirm the capability of BayesDeep to reconstruct the expression patterns of the underlying cells from spot-resolution gene expression accurately.

Application to Human Breast Cancer 10x Visium Data

We applied BayesDeep to reconstruct the single-cell-resolution molecular profiles for the SRT data from a human breast cancer study[15]. The data includes $N = 2,518$ spots and 17,651 genes. The gene expression was measured on a section of the human breast with invasive ductal carcinoma *via* the 10x Visium platform, along with annotation from pathologists as a reference (H&E-stained image with five annotated tissue regions in Figure 3A). After applying HD-Staining[13] to the histology image of breast cancer tissue, we identified $M = 156,235$ cells within seven categories: macrophage, ductal epithelium, karyorrhexis, tumor cell, lymphocyte, red blood cell, and stromal cell. Supplementary Table S1 provided ten shape features used in this study. We applied BayesDeep to reconstruct the single-cell-resolution molecular profiles on the top $P = 2,000$ highly variable genes, compared its performance to a competing method, TESLA, and further utilized the generated higher-resolution molecular profiles on several downstream analyses to reveal biological insights.

The gene expression at the single-cell resolution generated by BayesDeep offers a more detailed view of the spatial transcriptome landscape within the cellular environment. We chose both a tumor and an immune-related region as illustrative examples to demonstrate BayesDeep's capacity to enhance our understanding of gene expression in regions with high cell type heterogeneity (Figure 3A). For the example tumor region, we observed two distinct small areas densely populated with tumor cells. We further examined and presented the original spot-resolution expression from SRT data and reconstructed expression using BayesDeep and TESLA of three breast cancer-related genes (Figure 3B): First of all, *AZGP1*, which is primarily expressed in breast epithelial cells, plays a multifaceted role associated with cancer cachexia, carcinogenesis, and tumor differentiation[18]. Secondly, *MUCL1*, a breast-specific gene predominantly expressed in breast cancer, serves as a vital biomarker for tumor progression and metastasis[19,20]. Last but not least, *TOP2A* serves as a notable proliferation marker, demonstrating high expression in various subtypes of breast cancers[21]. The detailed expression pattern is relatively difficult to observe at the spot resolution (spot-resolution relative expression levels in Figure 3B) due to the low resolution. However, the expression reconstructed by BayesDeep for these three example genes distinctly exhibits the expression pattern consistent with the distribution of tumor cells, i.e., high expression on tumor cells and low expression on other cells. Conversely, for the gene expression reconstructed by TESLA, no similar expression pattern is observed in the two tumor areas for all three genes, indicating its limitation in recovering the comprehensive cellular expression. Violin plots in Figure 3B show the higher expression on tumor cells and ductal epithelium for BayesDeep, which aligns with the gene functions and their association with breast cancer. In contrast, the predicted expression from TESLA shows no difference among different cell types. Another observed region in Figure 3A is an immune region predominantly comprised of lymphocytes, which covers around

56 spots in ST data. In this region, many lymphocytes encircle a small tumor area. We selected three example genes related to immune functions for validation purposes. *CORO1A* is an identified immunity gene signature in breast cancer cohort studies[22]. *IGHG2* encodes the constant region of the immunoglobulin gamma-2 heavy chain, which enables antigen binding activity and immunoglobulin receptor binding activity and is involved in several processes, such as activating immune response[23]. *TRBC1* encodes T cell receptor β chain constant region 1, which is partially expressed in subsets of T cells[24,25]. No obvious expression patterns can be observed at the spot resolution due to the limited resolution (Figure 3C). However, through single-cell-resolution reconstruction *via* BayesDeep, we observed that these three genes exhibited high expression in the surrounding immune area. Again, gene expression recovered by TESLA shows no difference among various cell types and spatial regions. In summary, BayesDeep-predicted single-cell-resolution gene expression exhibits significantly stronger associations with the cell type information, unveiling detailed gene expression patterns not captured by the ST data alone.

In addition, BayesDeep quantifies the association between gene expression and the cellular features extracted from the histology image by estimating the covariate coefficient matrix ***B.*** In Figure 4A, the heatmap displays the coefficients of seven distinct cell types and ten nuclei-shape descriptors for the top 2,000 highly variable genes. It is noticed that a substantial number of the estimated covariate coefficients β_{lj}'s are non-zero and vary across cell types for most genes, which highlights unique expression patterns for different cell types. On the other hand, certain genes are characterized by all β_{lj}'s being zero, implying homogenous expression across cell types. In contrast, the covariate coefficients of those shape descriptors excluding solidity are notably sparse, with approximately 87.0% of those coefficients being zero. There exists an association between gene expression and the solidity of the cell nucleus for around 50.1% of genes. It is important to

note that the orientation covariate (i.e., the third column from the right of the heatmap shown in Figure 4A) is a negative control. This covariate is expected to be independent of gene expression, given that the measurement of nuclear orientation relies on the tissue's placement on the slide. We observed that the estimated coefficients for the orientation covariate are exactly zero for all genes, indicating the validity of BayesDeep results.

To demonstrate that BayesDeep-reconstructed single-cell-resolution gene expression data can improve downstream analysis, we conducted a comparative analysis between BayesDeep and TESLA across three tasks: (i) distinguishing different cell populations, (ii) explicating tumorigenesis through pseudotime analysis and (iii) elucidating cell-cell communication patterns. First, we employed Seurat for clustering and dimensional reduction analysis on the BayesDeep-generated single-cell-resolution gene expression. The results demonstrated that cells of the same cell type co-localized well on the UMAP plot and exhibited clear separation among distinct cell types (Figure 4B left panel). In contrast, we performed a similar analysis on the gene expression generated by TESLA. However, the results showed that TESLA could not effectively separate many cell types (Figure 4B right panel). These findings underscore the capability of BayesDeep to distinguish different cell populations. Second, we employed Moncle3[26,27] to perform a pseudotime analysis on tumor cells extracted from BayesDeep and TESLA results. In such analysis, we incorporated *CD44* and *CD24* as breast cancer stem cell markers, as previously reported[28-30]. We designated the cells with the highest *CD44* and *CD24* expression as the root of the trajectory to assess cancer progression over pseudotime (Supplementary Figure S4). Predicted results based on BayesDeep indicated that cancer cells from earlier pseudotime are predominantly enriched in the conventional tumor region. In contrast, those from later pseudotime are primarily concentrated within the necrotic tumor regions[31] (Fig. 4C, left panel). This trajectory pattern is consistent with

the timing order of tumor evolution. In contrast, the pseudotime derived from TESLA-generated predictive results did not show a biologically meaningful trend (Figure 4C, right panel). Overall, our results demonstrate the consistent superiority of BayesDeep in explicating tumorigenesis through the pseudotime analysis. Third, because BayesDeep can identify various cell types more accurately than TESLA, we hypothesized that BayesDeep could offer more details of cell-cell communications among these cell types than TESLA. To corroborate this hypothesis, we applied CellChat[32] to define the cell-cell communication landscape based on the cell clusters identified by pathologist annotations. Both BayesDeep and TESLA can achieve abundant cell-cell interactions, but BayesDeep can reveal more cellular communications than TESLA (Figure 4D). Interestingly, the majority ($1,001/1,486 = 67.3\%$) of cell-cell communications revealed by BayesDeep overlapped with those identified by TESLA (Figure 4D). However, some detailed cell-cell interactions revealed by BayesDeep and TESLA are different (Figure 4E): for example, BayesDeep could identify more immune-related, especially macrophage-related cellular communications than TESLA. In general, consistent with our expectations, BayesDeep can reveal more cell-cell communications than TESLA.

Application to Human Prostate Cancer 10x Visium Data

To assess the adaptability of BayesDeep across diverse tissue types, we conducted analysis utilizing another SRT data derived from human prostate cancer tissue[16]. The data includes $N = 4,371$ distinct spots for 17,651 genes. The gene expression measurement was obtained from a section of invasive carcinoma within the human prostate (H&E-stained image with six annotated tissue regions in Figure 4A), utilizing the 10x Visium platform. Subsequently, we applied the HD-Staining technique to identify nuclei on the histology image of this tissue. This image analysis process led to the segmentation of $M = 259,257$ individual cells, which were systematically

categorized into six classes: macrophage, karyorrhexis, tumor cell, lymphocyte, red blood cell, and stromal cell. We utilized BayesDeep to reconstruct high-resolution molecular profiles for the top 2,000 highly variable genes at the single-cell resolution. We then assessed its performance against the TESLA method and leveraged the resulting detailed molecular profiles in several downstream analyses.

To examine the gene expression reconstructed at the single-cell resolution by BayesDeep and TESLA, we selected two example regions with a high cell-type mixture - a tumor and an immune-related region (in Figure 5A). For the example tumor region depicted in Figure 5A, we observe that tumor cells have circular patterns around each empty region. We examined the expression of three prostate cancer-related genes (Figure 5B). *ADGRF1* is an adhesion-G protein-coupled receptor and has an essential function in cancer[33]. Aberrant expression and mutation of G protein-coupled receptors and their signaling partners, G proteins, have been well documented in many cancer[34]. *SPON2* (tumor cell-derived spondin 2) is an extracellular matrix glycoprotein, and overexpression of *SPON2* has been shown to promote tumor cell migration[35]. *TMEFF2* encodes a transmembrane protein containing an epidermal growth factor-like motif and two follistatin domains, highly expressed in prostate cancer samples [36]. At the spot resolution in Figure 5B, the expression patterns of these genes remain obscure. However, the gene expression reconstructed by BayesDeep shows strong expression patterns. These three example genes have higher expression in tumor cells. In contrast, the gene expression reconstructed by TESLA does not display differential expression on tumor cells. The violin plots in Figure 5B provide evidence of the differential expression among cell types for BayesDeep, in line with the three example genes' cancer-related functions. For another selected region, as depicted in Figure 5A, we find an immune region in the center primarily composed of lymphocytes. For our analysis, we selected three

immune-related genes. *CD24* is a cell surface glycosyl-phosphatidylinositol–anchored protein expressed on various cell types, including developing T and most B lymphocytes[37]. *CD47* is an immunoglobulin superfamily pentatransmembrane protein ubiquitously expressed in hematopoietic cells, including lymphocytes[38]. *CXCR4* is the receptor for the CXC chemokine stromal-derived-factor-1, which has essential functions on immunological functions and T lymphocyte trafficking[39]. At the spot resolution, differential expression in the small immune region is not readily observable since the region is only covered by around five spots. However, through single-cell-resolution reconstruction with BayesDeep, we can clearly see that these three genes exhibit high expression within the immune area. And violin plots show their high expression within immune cells, such as lymphocytes or macrophage. In contrast, TESLA fails to detect gene expression changes that occur only in a small region.

The inference of covariate coefficients in BayesDeep reflects the correlation between single-cell-resolution molecular profiles and extracted cell characteristics from histology images. Figure 6A shows a heatmap of coefficients for six cell types and ten nuclei-shape covariates among the top 2,000 highly variable genes. Most genes exhibit non-zero coefficients for at least one cell type, indicating the differential expression of these cell types. For nuclei-shape features, the estimated coefficients are highly sparse, with around 90.4% being zero. The estimated coefficients on the covariate for negative control and orientation are exactly zero for all genes, further validating the efficiency of BayesDeep.

Furthermore, we compared BayesDeep and TESLA on their impact on downstream analysis. First, through clustering and dimensional reduction analysis of the prostate cancer single-cell-resolution results generated by BayesDeep and TESLA, we found that the BayesDeep UMAP plot, but not the TESLA UMAP plot, revealed clear clustering results for various cell types (Figure 6B). Second,

in the pseudotime analysis of prostate cancer, we used *ITGA6* and *ALCAM* as prostate cancer stem cell markers, as previously reported[28-30] (Supplementary Figure S5). Trajectory analysis of the prostate cancer cells, extracted from BayesDeep results, revealed an even distribution of cancer cells with varying pseudotime throughout the prostate tumor. In contrast, TESLA results indicated a preference for prostate cancer cells with earlier pseudotime to be enriched at the tumor periphery (Figure 6C). Third, we assessed BayesDeep's ability to identify cell-cell communication in prostate cancer tissue compared to TESLA. BayesDeep identified 281 ligand-receptor interacting pairs, more than twice the number identified by TESLA (Figure 6D). In summary, these findings demonstrated that BayesDeep could outperform TESLA in multiple downstream analysis tasks and, therefore, might provide deeper insights into the underlying molecular mechanisms that regulate tumor microenvironment and tumorigenesis.

DISCUSSION

Here, we introduced BayesDeep, a Bayesian hierarchical model for the reconstruction of gene expression at the single-cell resolution for NGS-based SRT data. BayesDeep links the cellular morphological information extracted from histology images and the spot-resolution molecular profiles derived from SRT data, enabling the inference of single-cell-resolution molecular profiles. Specifically, BayesDeep employs a negative binomial distribution to model spot-resolution gene expression and considers the latent normalized gene expression as an average of single-cell-resolution gene expression across all cells within a given spot. Our findings demonstrated the superior performance of BayesDeep compared to other competing methods for gene expression reconstruction in both simulation study and real SRT data.

Furthermore, by applying BayesDeep to two real SRT data, we unveiled detailed expression patterns that cannot be captured at the spot resolution, which offers invaluable insights for further exploration of the subsequent biomedical research. Firstly, cell populations can be separated to characterize different cell types or subpopulations based on the predicted single-cell-resolution molecular profiles, which can be further applied to marker gene identification and automated annotation[40]. Secondly, pseudotime analysis on the reconstructed high-resolution molecular profiles helps to understand the temporal progression of cellular states or trajectories within the tumor microenvironment, which further aids in unraveling the dynamic nature of tumors, providing valuable information on how cells change and evolve over time[41]. Last but not least, reconstructed molecular profiles at the single-cell resolution enable the examination of cell-cell communication patterns, which further contributes to a deeper understanding of cellular behavior, intercellular signaling dynamics, and their implications in clinical studies[32]. In summary, the downstream analysis indicates the significance of BayesDeep in exploring the tumor microenvironment and advancing the study of molecular biology.

There are several important future extensions for BayesDeep. First, the performance of BayesDeep highly relies on the selection of cellular features included in the model. On the simulated data where the actual gene expression is artificially generated and thus totally determined by the included covariates, BayesDeep can reconstruct single-cell-resolution gene expressions highly consistent with the actual expression. However, for the model validation on mouse visual cortex STARmap data, the limited cellular characteristic data available may constrain a comprehensive reconstruction of gene expression, thus resulting in relatively low consistency with the actual gene expression. Therefore, improving nuclei segmentation and classification methods offers more accurate cell biological and morphological characteristics and, in turn, further promotes the

performance of BayesDeep. Second, it is necessary to optimize the computational efficiency of BayesDeep for large-scale studies, especially for larger data that are becoming increasingly common with the advancement of SRT technologies. In addition, there is potential for the BayesDeep model to integrate external data or other types of omics data, such as scRNA-seq data, to enhance gene expression inference. Integrating other data types could provide a more comprehensive understanding of the cellular environment and the complex interactions within cells. Lastly, we anticipate that BayesDeep could be extended to generate molecular profiles for three-dimensional tissue regions. We have demonstrated BayesDeep's ability to extrapolate gene expression in a region with a similar cellular environment but without SRT data. With the available three-dimensional image profiles, BayesDeep can be adjusted to build molecular profiles at the single-cell resolution on three-dimensional tissue sections and provide more detail and context to cellular interactions and tissue architecture. These future directions could further boost the performance and generalizability of BayesDeep.

METHODS

In this section, we define the spot-resolution molecular profile of NGS-based SRT data generated from ST or the improved 10x Visium platform and the single-cell-resolution image profile derived from the AI-reconstructed histology image paired with the SRT data. Then, we describe the geospatial profile that captures the spatial relationship between spots and cells. Following this, we detail the Bayesian statistical model, BayesDeep, employed for the reconstruction of expression for each gene of interest at the single-cell resolution. For quick reference, all data and parameter notations introduced in this paper are summarized in Supplementary Table S2.

<u>Data Preparation</u>

<u>The spot-resolution molecular profile</u>. We denote the SRT molecular profile (i.e., gene expression measurements in terms of counts) as an N-by-P matrix $\boldsymbol{Y} = [y_{ij}]_{N \times P}$, where each entry $y_{ij} \in \mathbb{N}$ represents the read count for gene j ($j = 1, \cdots, P$, with P being the total number of genes) observed at spot i ($i = 1, \cdots, N$, with N being the total number of spots). These N spots are regularly arrayed on a two-dimensional square or triangle lattice, with their spatial coordinates given by an N-by-2 matrix $\boldsymbol{T}^Y = [t_{ir}^Y]_{N \times 2}$, where each row $\boldsymbol{t}_i^Y = (t_{i1}^Y, t_{i2}^Y) \in \mathbb{R}^2$ records the x and y-coordinates of spot i in the designated domain.

<u>The single-cell-resolution image profile.</u> We denote the image profile, derived from the histology image paired with the SRT data, as an M-by-L matrix $\boldsymbol{X} = [x_{ml}]_{M \times L}$, where each entry $x_{ml} \in \mathbb{R}$ represents a measurement for explanatory variable l ($l = 1, \cdots, L$, with L being the total number of explanatory variables) observed for cell m ($m = 1, \cdots, M$, with M being the total number of cells). Our study utilizes L single-cell-resolution morphological features as the explanatory variables, extracted using the histology-based digital (HD)-staining model[13]. Specifically, the HD-staining

model is a deep-learning model based on the mask regional convolutional neural network (Mask R-CNN) architecture[42], which is trained to segment the nucleus of various cell types, such as immune, tumor, and stromal cells. HD-staining then computes ten geometric shape features for each identified cell nucleus, including filled area, net area, convex area, extent, perimeter, solidity, eccentricity, major axis length, minor axis length, and orientation. The definitions of these shape features can be found in Table S3 of a recent study[43]. Supposing there are Q different types of cells coded in Q dummy variables, the total number of explanatory variables is $L = Q + 10$. The HD-staining model was initially trained using histology images from lung adenocarcinoma patients in the National Lung Screening Trial (NLST) study (https://biometry.nci.nih.gov/cdas/nlst/), wherein nuclei of six different cell types were manually annotated by well-experienced pathologists. Although the original model was trained by using data specific to lung cancer, it has been improved and validated to adapt to histology images from various cancer types, including breast cancer, prostate cancer, and other carcinomas. To represent cell nuclei locations, we use an M-by-2 matrix $\boldsymbol{T}^X = [t_{mr}^X]_{M \times 2}$, where each row $\boldsymbol{t}_m^X = (t_{m1}^X, t_{m2}^X) \in \mathbb{R}^2$ records the x and y-coordinates of cell m in the designated domain.

The spot-cell geospatial profile. In the context of NGS-based SRT technologies, spots are defined as circular regions comprising barcoded mRNA capture probes, where gene expression is quantified within a given tissue section[3]. The SRT data provides both the physical spot diameter d and its corresponding length in pixels. Assuming that the x and y-coordinates in spot-resolution geospatial profile T^Y represent the locations of spot centers, then we can identify whether cell m is located within spot i by a direct evaluation of the condition $\sqrt{(t_{m1}^X - t_{i1}^Y)^2 + (t_{m2}^X - t_{i2}^Y)^2} \leq d/2$. This allows us to construct the spot-cell geospatial profile, denoted by an N-by-M binary matrix $\boldsymbol{G} = [g_{im}]_{N \times M}$, which reflects whether a given cell m is within the barcoded area of a given

spot i (i.e., $g_{im} = 1$) or not (i.e., $g_{im} = 0$). Notably, the coverage of measured areas in NGS-based SRT techniques is relatively limited, encompassing approximately 38% of the overall tissue section. As a result, most cells do not fall within the boundaries of any of the defined spots. In our real data analysis, for instance, 97,135 out of $M = 156,115$ cells for the human breast cancer 10x Visium data, and 252,765 out of $M = 352,818$ cells for the human prostate cancer 10x Visium data, are not covered by the measured area of any spots. This limitation in the scope of measured regions can lead to substantial information loss when analyzing SRT data. This motivates us to develop BayesDeep to address this challenge and enhance the analysis of SRT data.

Model Description

BayesDeep is essentially a negative binomial regression model with regularization for handling grouped observations under the Bayesian framework. Its primary objective is to utilize the spot-level molecular profile, single-cell-resolution image profile, and spot-cell geospatial profile to reconstruct the single-cell-resolution molecular profile, which can be represented by an M-by-P matrix $\boldsymbol{\Theta} = \left[\theta_{mj}\right]_{M \times P}$, where each entry $\theta_{mj} \in \mathbb{R}^+$ is the predictive relative expression for gene j within cell m. It is worth noting that BayesDeep focuses on reconstructing relative gene expression at the single-cell resolution for one gene at a time (i.e., a specific column vector in $\boldsymbol{\Theta}$), which allows for efficient parallel processing of multiple genes.

Modeling the spot-resolution molecular profile. We start by modeling the over-dispersed spot-resolution gene expression matrix $\boldsymbol{Y}$ using a negative binomial (NB) distribution. NB-based models are widely used for analyzing sequence count data[44-47] due to their ability to accommodate inherent over-dispersion. Specifically, we postulate that each observed read count follows an NB distribution:

$$y_{ij} \,|\lambda_{ij}, \phi_j \sim \text{NB}\big(s_i \lambda_{ij}, \phi_j\big).$$

Here the NB(ν, ϕ) distribution is parameterized in terms of its mean $\nu \in \mathbb{R}^+$ and dispersion $1/\phi \in \mathbb{R}^+$, with the probability mass function given by $f(y|\nu,\phi) = \frac{\Gamma(y+\phi)}{y!\Gamma(\phi)}\left(\frac{\phi}{\nu+\phi}\right)^{\phi}\left(\frac{\nu}{\nu+\phi}\right)^{y}$. This parameterization provides the flexibility to characterize an unknown mean-variance structure, with the variance calculated $\nu + \nu^2/\phi$. A small value of ϕ indicates a high variance to mean ratio (i.e., $\frac{\nu+\nu^2/\phi}{\nu} = 1 + \nu/\phi$), while a large value approaching infinity reduces the NB distribution to a Poisson distribution with the same mean and variance. The NB mean is further decomposed of two multiplicative components, 1) the size factor, denoted as s_i, and 2) the spot-resolution relative expression for gene j observed at spot i, denoted as λ_{ij}. Such a multiplicative characterization of the NB or Poisson mean is typical in both the frequentist[48-50] and the Bayesian[51,52] literature when modeling sequence count data. The set of n size factors is represented as $\boldsymbol{s} = [s_i]_{N\times1}$, capturing a range of biological and technical variabilities across samples, such as reverse transcription efficiency, amplification/dilution efficiency, and sequencing depth[53]. To ensure identifiability between these two parameters, we set s_i proportional to the summation of the total read counts across all genes at spot i [54]. With a constraint of $\prod_{i=1}^{N} s_i = 1$, we compute $s_i = \sum_{j=1}^{P} y_{ij} / \prod_{i=1}^{n}\sum_{j=1}^{P} y_{ij}$ [44]. It is noteworthy that alternative methods for estimating $\boldsymbol{s}$ are available, such as setting each s_i to be proportional to the upper quartiles of non-zero read counts across all genes at spot i [55], or even modeling $\boldsymbol{s}$ through a Dirichlet process mixture model with mean constraint[56]. To complete the prior model specification, we place a common gamma prior on all dispersion parameters, denoted as $\boldsymbol{\phi} = [\phi_j]_{P\times1}$, that is, $\phi_j \sim \text{Ga}\big(a_\phi, b_\phi\big)$, where a_ϕ and b_ϕ are fixed

hyperparameters. We recommend choosing small values, such as $a_\phi = b_\phi = 0.1$, to maintain a weakly informative setting.

Linking the spot-resolution and single-cell-resolution molecular profiles. We posit that the spot-resolution relative expression for gene j (i.e., λ_{ij}) can be derived as an average of single-cell-resolution relative expression (i.e., θ_{mj}'s) across all cells within the given spot i. This relationship can be formally expressed as:

$$\lambda_{ij} = \sum_{m=1}^{M} g_{im}\theta_{mj} \Big/ \sum_{m=1}^{M} g_{im},$$

where the indicator variable g_{im} in the spot-cell geospatial profile $\boldsymbol{G}$ takes the value one if cell m is located within the barcoded area of spot i, and zero otherwise. The denominator in the above formulae, $\sum_{m=1}^{M} g_{im}$, represents the total number of cells within spot i observed in the paired histology images. From an alternative perspective, this assumption can be interpreted as the spot-resolution gene expression being an average of cell-type-specific gene expression within the same spot, weighted by the respective cell-type proportions. Notably, this foundational assumption finds validations in various cell-type deconvolution algorithms designed for SRT data[57-59]. From a statistical viewpoint, this assumption enables the application of a regression model tailored to accommodate grouped observations[60,61].

Linking the single-cell-resolution molecular and image profiles. In our pursuit of predicting the single-cell-resolution gene expression θ_{mj} 's, we leverage morphological information encompassing cell type and nuclei-shape features of each cell m. Specifically, we adopt a linear model to capture the relative expression for gene j within cell m:

$$\log \theta_{mj} = \beta_{0j} + \sum_{l=1}^{L} \beta_{lj}\, x_{ml}.$$

The choice of a log-link function is driven by the necessity for maintaining θ_{mj} values in the positive domain, thereby ensuring a positive NB mean. In this formulation, β_{0j} is the baseline expression for gene j shared by all cells. Note that $\exp(\beta_{0j})$ can also be interpreted as a scaling factor that adjusts for gene-specific effects. As previously introduced, we use the M-by-L matrix X to present observations from L explanatory variables extracted from the histology image, including cell type information, cell nuclei-shape descriptors, and other important single-cell-resolution measurements relevant to the analysis. Given the coding of Q cell types into Q dummy variables, we added a constraint $\sum_{q=1}^{Q} \beta_{qj} = 0$ to avoid identifiability issues arising from the sum of the components. Consequently, $\exp(\beta_{lj})$, corresponding to cell type Q, represents the mean relative expression for gene j across all cells. Each column in the L-by-P coefficient matrix $B = \left[\beta_{lj}\right]_{L \times P}$, denoted as $\boldsymbol{\beta}_j = \left[\beta_{lj}\right]_{L \times 1}$ with $\beta_{lj} \in \mathbb{R}$, describes the effect of the L explanatory variables on the logarithm of the relative expression across cells for gene j. Hence, we can use the coefficient matrix B to explore the association between the image and molecular profiles at the single-cell resolution. In practice, for a specific gene j, it is likely that only a limited number of explanatory variables account for its expression. For instance, the expression of some housekeeping genes exhibits small variability among different tissues, cell types, or samples[62]. Under such circumstances, their corresponding coefficients β_{lj}'s tend to be zeros. On the other hand, some pairs of explanatory variables might be highly correlated, leading to an inflation of regression coefficients and potentially harming prediction performance. We employ a

regularization mechanism to prevent over-fitting and reduce the potential multicollinearity by specifying a spike-and-slab prior[63] on each β_{lj}. This prior is a mixture of distributions:

$$\beta_{lj}|\gamma_{lj} \sim (1-\gamma_{lj})\delta_0(\beta_{lj}) + \gamma_{lj}\text{Norm}(0, \sigma_\beta^2).$$

The spike component $\delta_0(\beta_{lj})$ is a point mass distribution at $\beta_{lj} = 0$, while the slab component is a normal distribution centered at zero. If the auxiliary binary variable $\gamma_{lj} = 1$, then the probability of $\beta_{lj} = 0$ is zero, indicating that explanatory variable l is relevant for explaining the relative expression for gene j. Conversely, $\gamma_{lj} = 0$ restricts that $\beta_{lj} = 0$, indicating that explanatory variable l has no contribution to the gene j's expression. This prior setting enables us to identify significant associations between gene expression and explanatory variables, which, in our case, are the morphological features. The variance of the slab component $\sigma_\beta^2 \in \mathbb{R}^+$ is a fixed hyperparameter set to one. We complete the model specification by placing an independent Bernoulli distribution on γ_{lj}, i.e., $\gamma_{lj} \sim \text{Bern}(\pi_\gamma)$, where $\pi_\gamma \in (0,1)$ is a fixed hyperparameter that indicates the percentage of explanatory variables included in the final model *a priori*. We set $\pi_\gamma = 0.5$ to incorporate relatively weak information.

Full data likelihood and posterior. The model parameter space consists of 1) the dispersion parameter $\boldsymbol{\phi}$ that accounts for the over-dispersion commonly observed in gene expression data, 2) the coefficient matrix $\boldsymbol{B}$ that quantifies the relationship between gene expression as measured in SRT data and the morphological features extracted from the paired histology image, and 3) the L-by-P selection matrix $\boldsymbol{\Gamma} = [\gamma_{lj}]_{L \times P}$ that indicates the significant association in the coefficient matrix $\boldsymbol{B}$. The complete data likelihood can be written as:

$$f(Y|\boldsymbol{\phi}, B, \boldsymbol{\Gamma}) = \prod_{j=1}^{P} \prod_{i=1}^{N} \text{NB}\left(s_i \frac{\sum_{m=1}^{M} g_{im} \exp\left(\sum_{l=1}^{L} \beta_{lj} x_{ml}\right)}{\sum_{m=1}^{M} g_{im}}, \phi_j\right).$$

With the prior specifications detailed above, the full posterior distribution can be written as:

$$\pi(\boldsymbol{\phi}, B, \boldsymbol{\Gamma}|Y) = f(Y|\boldsymbol{\phi}, B, \boldsymbol{\Gamma}) \times \pi(B|\boldsymbol{\Gamma})\pi(\boldsymbol{\Gamma})\pi(\boldsymbol{\phi})$$

$$= \prod_{j=1}^{P} \prod_{i=1}^{N} \text{NB}\left(s_i \frac{\sum_{m=1}^{M} g_{im} \exp\left(\sum_{l=1}^{L} \beta_{lj} x_{ml}\right)}{\sum_{m=1}^{M} g_{im}}, \phi_j\right)$$

$$\times \prod_{j=1}^{P} \prod_{l=1}^{L} \left(\left[(1 - \gamma_{lj})\delta_0(\beta_{lj}) + \gamma_{lj}\text{Norm}(\beta_{lj}; 0, \sigma_\beta^2)\right]\text{Bern}(\gamma_{lj}; \pi_\gamma)\right) \prod_{j=1}^{P} \text{Ga}(\phi_j; a_\phi, b_\phi)$$

Model Fitting

Posterior sampling *via* MCMC algorithms. We explore the posterior distribution *via* a Markov chain Monte Carlo (MCMC) algorithm based on stochastic search variable selection[64,65]. Specifically, we iteratively update each parameter using a Metropolis-Hasting (MH) algorithm. We note that this algorithm is sufficient to guarantee ergodicity for our model. Full details regarding the implementation of the MCMC algorithm are available in Section S2 in Supplementary Materials.

Posterior inference. We obtain posterior inference by post-processing the MCMC samples following the burn-in phase. Let $\phi_j^{(u)}$, $\beta_{lj}^{(u)}$ and $\gamma_{lj}^{(u)}$ denote the posterior samples in the u-th iteration after burn-in, where $u = 1, \dots, U$. Our primary focus is on the selection of the important explanatory variables for each gene j, *via* the selection matrix $\boldsymbol{\Gamma}$. An effective approach to summarize the posterior distributions of these binary parameters is by computing the marginal posterior probability of inclusion (PPI):

$$\text{PPI}(\gamma_{lj}) = \frac{1}{U} \sum_{u=1}^{U} \gamma_{lj}^{(u)}.$$

Then, we set $\hat{\gamma}_{lj} = 1$ if its corresponding PPI exceeds a pre-specified threshold. When choosing the threshold, we recommend either opting for a threshold of 0.5, leading to a median model, or following a procedure that controls the expected Bayesian false discovery rate [66]. For each dispersion parameter ϕ_j and each coefficient β_{lj}, we estimate them by calculating their posterior means by averaging over all their respective MCMC samples after burn-in,

$$\hat{\phi}_j = \frac{1}{U} \sum_{u=1}^{U} \phi_j^{(u)} \text{ and } \hat{\beta}_{lj} = \frac{1}{U} \sum_{u=1}^{U} \beta_{lj}^{(u)}.$$

Additionally, a quantile estimation or credible interval for each parameter of interest can be obtained from MCMC samples.

Predictive inference. Our primary goal is to reconstruct gene expression at the single-cell resolution by estimating the matrix $\boldsymbol{\Theta} = \left[\theta_{mj} \right]_{M \times P}$. On the basis of the MCMC samples on the coefficient matrix $\boldsymbol{B} = \left[\beta_{lj} \right]_{L \times P}$, we predict each θ_{mj} by Monte Carlo simulation. Specifically, at each iteration u after burn-in, we compute the relative gene expression for gene j within cell m as follows:

$$\theta_{mj}^{(u)} = \exp \left(\sum_{l=1}^{L} \beta_{lj}^{(u)} x_{ml} \right).$$

Subsequently, we sample the gene expression for gene j within spot i using an NB distribution,

$$y_{ij}^{(u)} \sim \text{NB} \left(s_i \frac{\sum_{m=1}^{M} g_{im} \theta_{mj}^{(u)}}{\sum_{m=1}^{M} g_{im}}, \phi_j^{(u)} \right).$$

Consequently, both single-cell-resolution relative gene expression $\hat{\theta}_{mj}$ and spot-resolution gene expression $\hat{y}_{ij}$ can be estimated by summarizing their corresponding MCMC samples. For instance, their predictive means can be approximated as follows:

$$\hat{\theta}_{mj} = \frac{1}{U}\sum_{u=1}^{U} \theta_{mj}^{(u)} \text{ and } \hat{y}_{ij} = \left\lceil \frac{1}{U}\sum_{u=1}^{U} y_{ij}^{(u)} \right\rceil,$$

where $\lceil \cdot \rceil$ denotes the ceiling function.

Data availability

The authors analyzed two publicly available SRT data. Raw count matrices, images, and spatial information for two SRT data from 10x Visium are accessible on the 10x Genomics website at https://support.10xgenomics.com/spatial-gene-expression/datasets.

Code availability

An open-source implementation of the BayesDeep algorithm in R/C++ is available at https://github.com/Xijiang1997/BayesDeep.

Acknowledgments

Competing Interests statement

The authors declare no competing interests.

A

Spot-resolution molecular profile   **Spot-cell geospatial profile**   **Single-cell-resolution image profile**

Gene expression measurements $\boldsymbol{y}_j$ for gene j

Spot coordinates   Nucleus coordinates

Histology image   Nuclei characterization

Spot-cell spatial relationship $\boldsymbol{G}$

Nucleus characteristics $\boldsymbol{X}$ for all M cells

Cell type   Shape features

B

The BayesDeep Model

Modeling spot-resolution molecular profile: $y_{ij}|\lambda_{ij},\phi_j \sim \mathrm{NB}(s_i\lambda_{ij},\phi_j)$ for $i = 1,\cdots,n$

Linking spot- and single-cell-resolution molecular profiles : $\boldsymbol{\lambda}_j = \boldsymbol{G}\boldsymbol{\theta}_j/\boldsymbol{G}\mathbf{1}_{M\times 1}$

Linking single-cell-resolution molecular and image profiles : $\log(\boldsymbol{\theta}_j) = \boldsymbol{X}\boldsymbol{\beta}_j$

Selecting features with the spike-and-slab prior: $\beta_{lj}|\gamma_{lj} \sim \gamma_{lj}\delta_0 + (1-\gamma_{lj})\mathrm{Normal}(0,\sigma_\beta^2)$ for $l = 1,\cdots,L$

C

BayesDeep Outputs

Association between single-cell-resolution image and molecular profiles $\widehat{\boldsymbol{B}}$

$\widehat{\boldsymbol{\beta}}_j \rightarrow$

Cell type

Shape features

Gene 1 Gene 2 ...... Gene j ...... Gene P

Single-cell-resolution molecular profile $\widehat{\boldsymbol{\theta}}_j = \exp(\boldsymbol{X}\widehat{\boldsymbol{\beta}}_j)$

D

BayesDeep enhances downstream analyses

Cell populations

Cell type 1
Cell type 5
Cell type 2
Cell type 4
Cell type 3
Cell type 6
Cell type 7
UMAP2
UMAP1

Pseudotime

Pseudotime
0   20

Cell-cell communication

Cell type 3
Cell type 2
Cell type 4
Cell type 1
Cell type 5
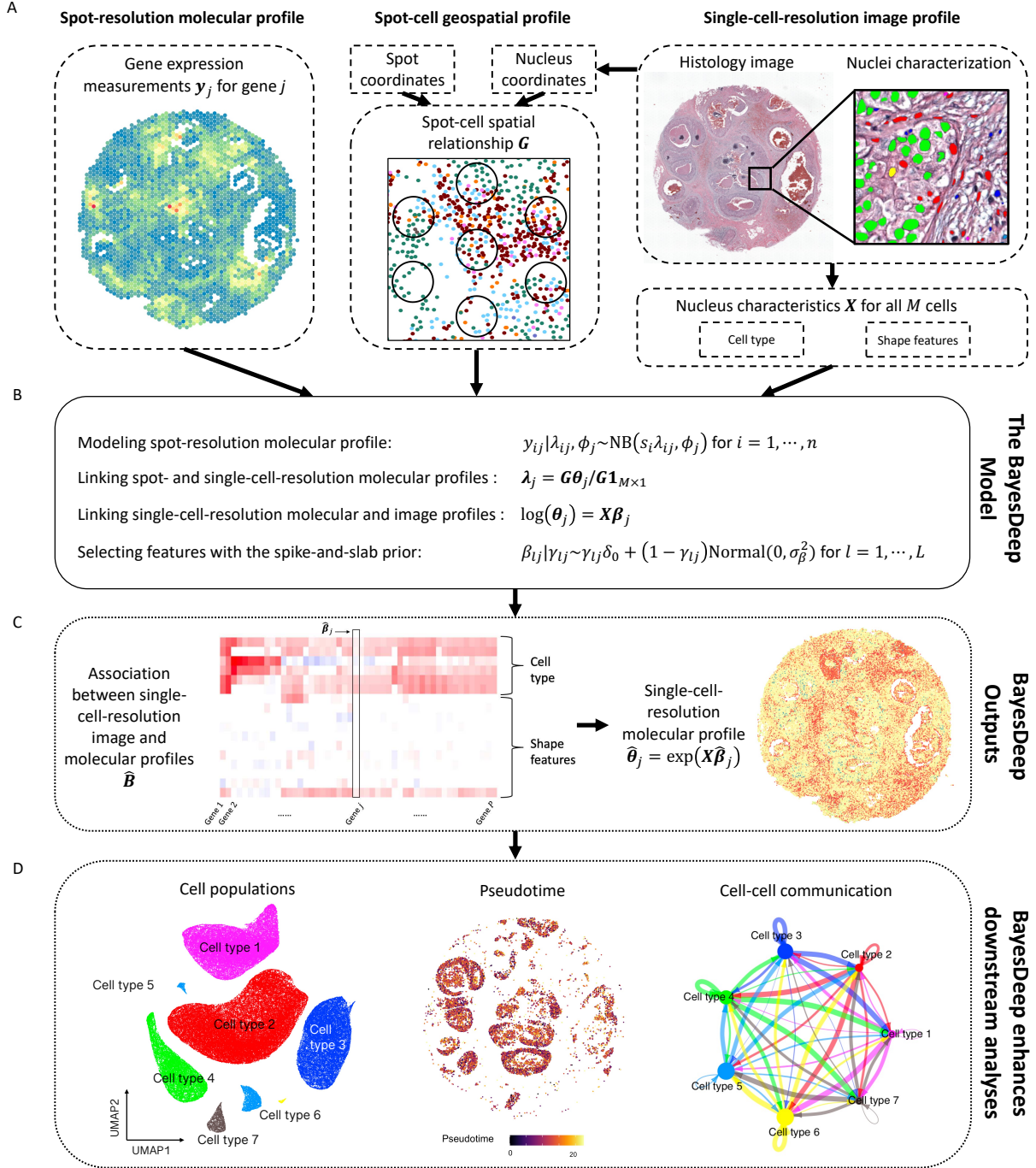Cell type 7
Cell type 6

Figure 1. Flowchart of the proposed BayesDeep: A. BayesDeep integrates the spot-resolution molecular profile $\boldsymbol{Y}$ from NGS-based SRT data, the single-cell-resolution image profile $\boldsymbol{X}$ from the paired AI-reconstructed histology image, and the spot-cell geospatial profile $\boldsymbol{G}$ to recover gene

expression at the single-cell resolution $\boldsymbol{\Theta}$. B. The hierarchical formulation of the BayesDeep model is based on a Bayesian regularized negative binomial regression model with grouped observations. C. BayesDeep estimates the association between the single-cell-resolution molecular and image profiles $\boldsymbol{B}$ and predicts the single-cell-resolution molecular profile $\boldsymbol{\Theta}$. D. Several downstream analyses can be enhanced based on the availability of the single-cell-resolution molecular profile $\boldsymbol{\Theta}$, including identifying distinct cell populations, elucidating the process of tumorigenesis *via* pseudotime analysis, and exploring the mechanisms of cell-cell communication.
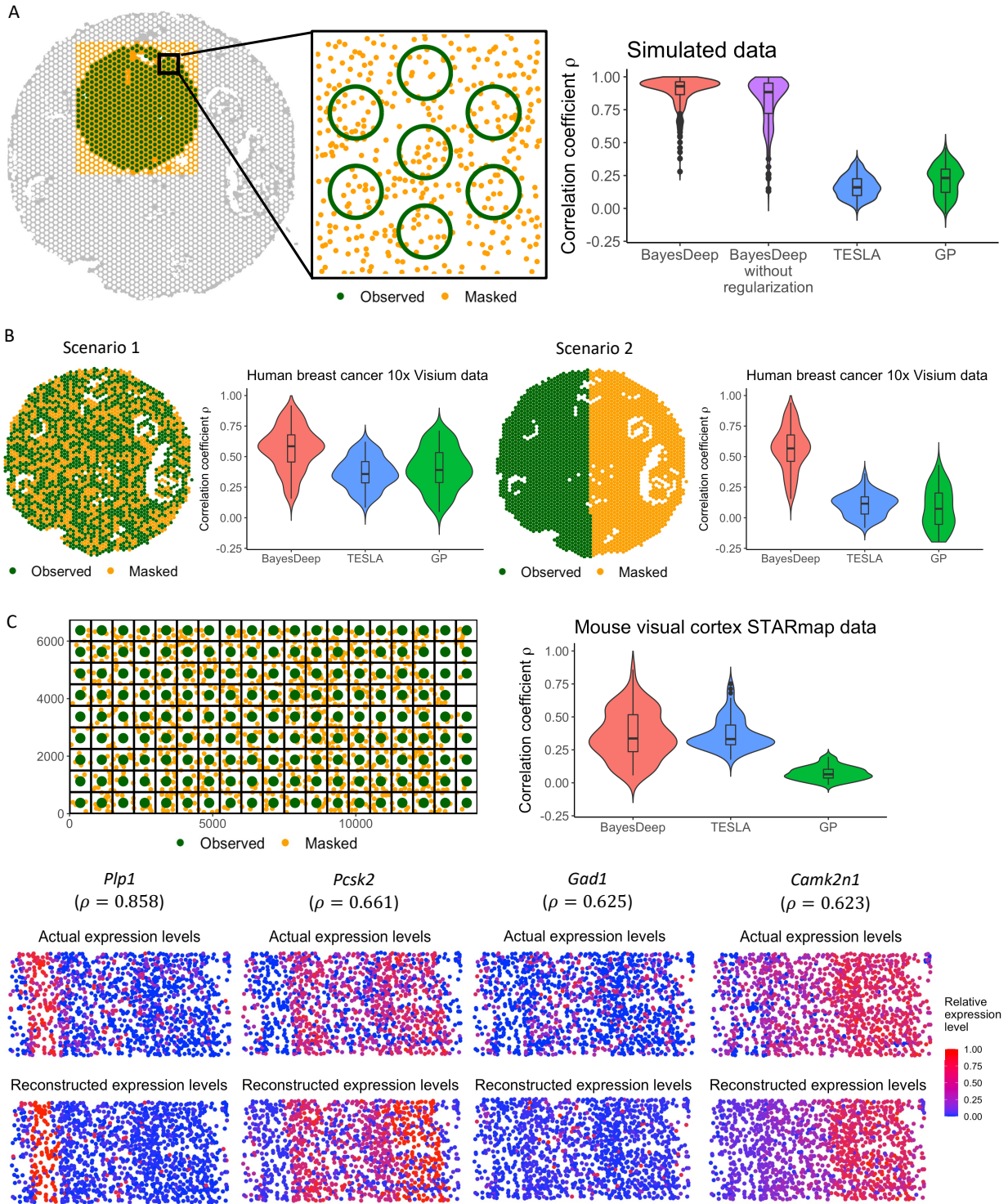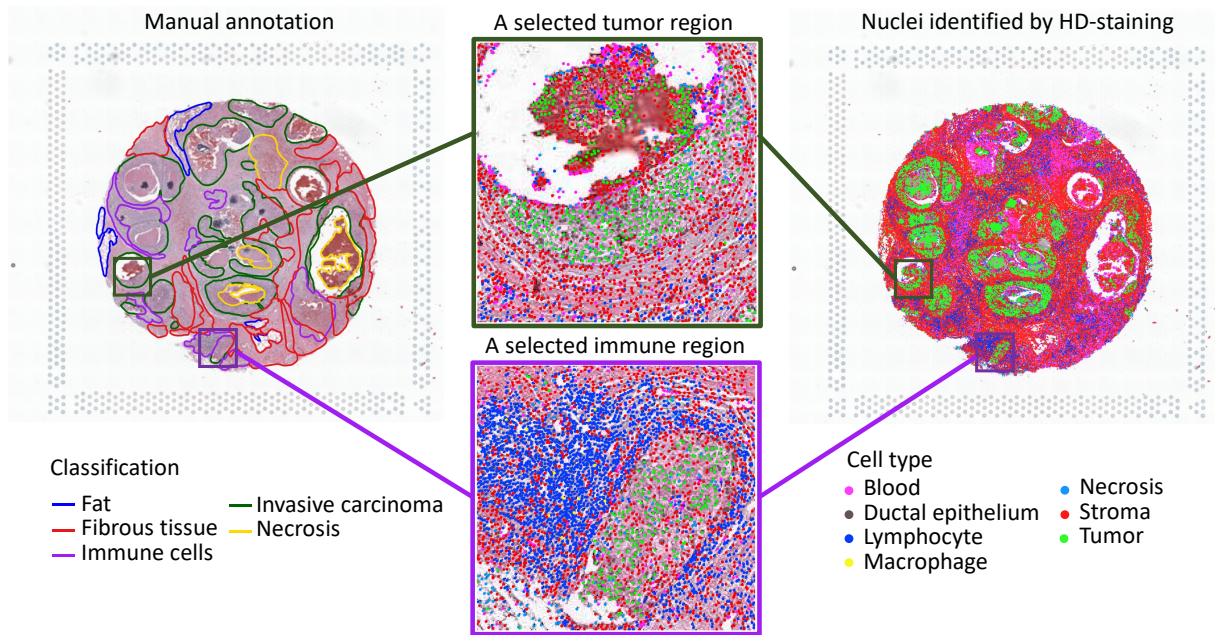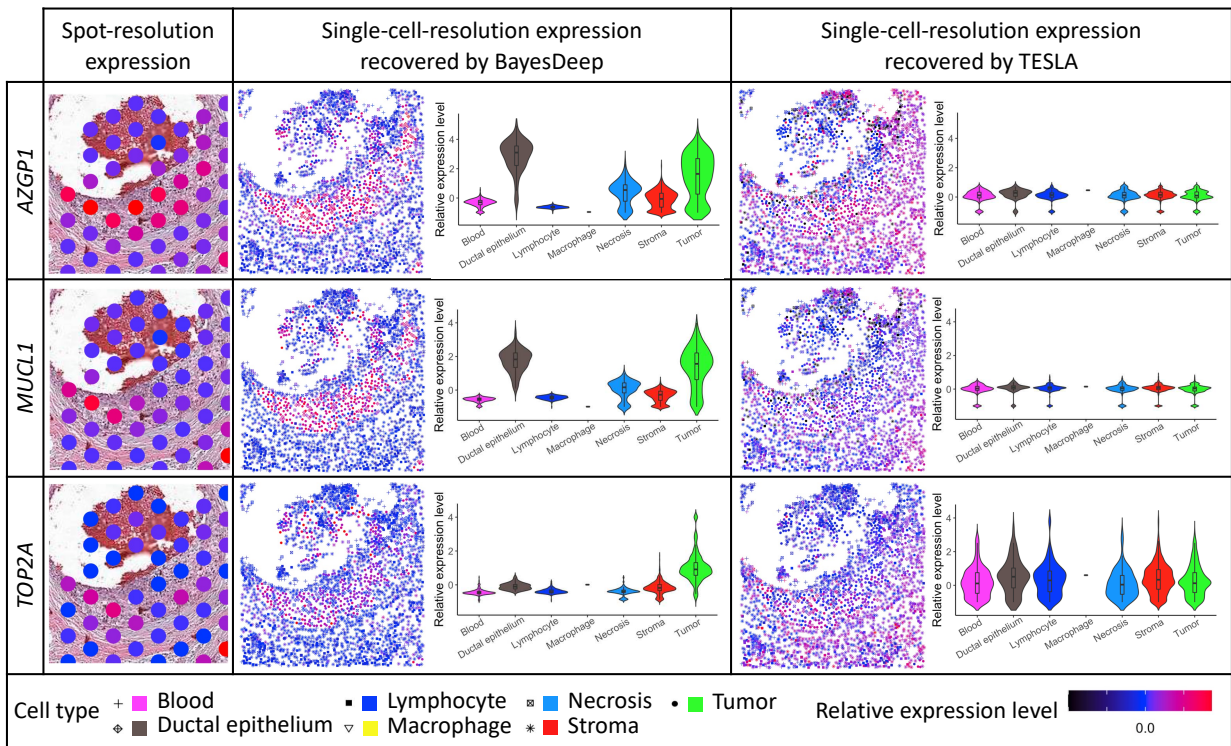
Figure 2. Overview of model validation, including the validation settings and evaluation results in terms of the Pearson correlation coefficients ρ between the actual and predicted gene expression

for BayesDeep, TESLA, and Gaussian Process (GP), respectively. The validation is stratified into three distinct data: A. Simulated data at the spot resolution; B. Human breast cancer 10x Visium data at the spot resolution; C. Mouse visual cortex STARmap data at the single-cell resolution.

A

Manual annotation

A selected tumor region

Nuclei identified by HD-staining

A selected immune region

Classification

— Fat
— Fibrous tissue
— Immune cells
— Invasive carcinoma
— Necrosis

Cell type

- Blood
- Ductal epithelium
- Lymphocyte
- Macrophage
- Necrosis
- Stroma
- Tumor

B

	Spot-resolution expression	Single-cell-resolution expression recovered by BayesDeep	Single-cell-resolution expression recovered by TESLA
AZGP1			
MUCL1			
TOP2A			

Cell type
+ Blood
⊕ Ductal epithelium
■ Lymphocyte
▽ Macrophage
⊠ Necrosis
✳ Stroma
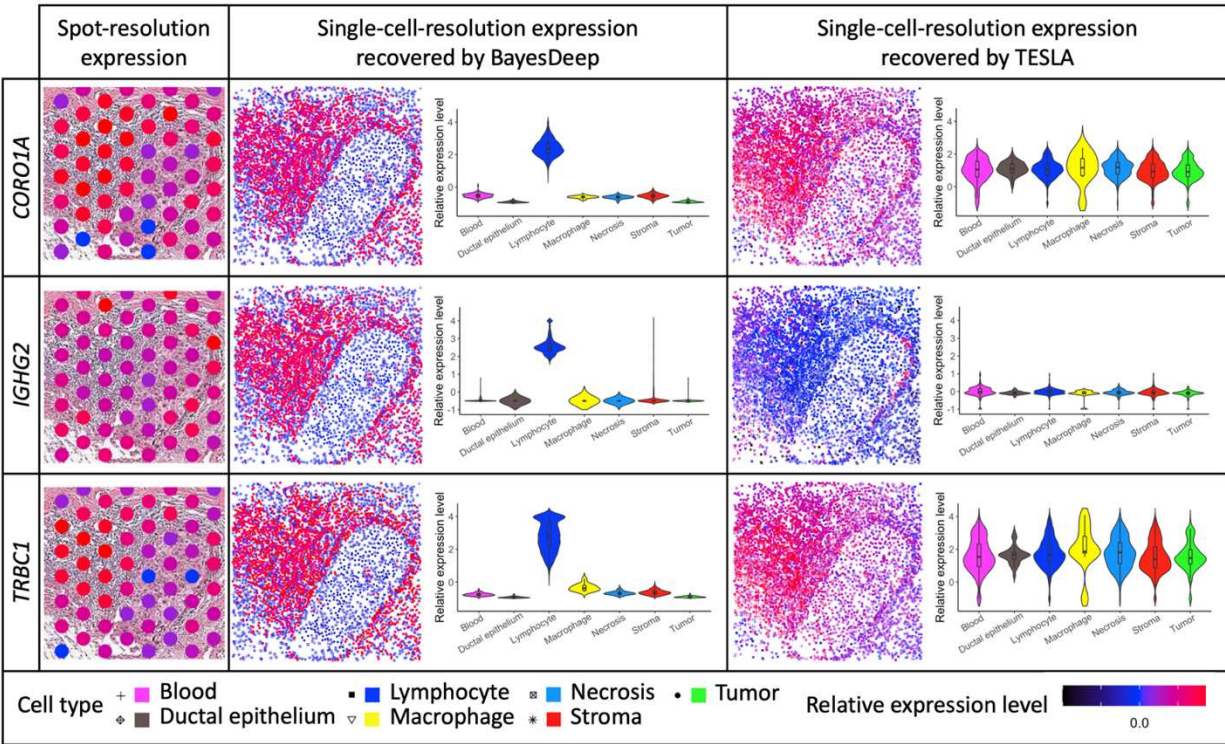● Tumor

Relative expression level
0.0

Figure 3. Real data analysis on the human breast cancer 10x Visium data: A. Manual annotation and nuclei identification results by applying HD-staining to the paired histology image; and the two selected tumor and immune regions for an illustrative comparison between BayesDeep and TESLA. B. The actual expression at the spot resolution and the predicted expression at the single-cell resolution by BayesDeep and TESLA on the selected tumor region for genes *AZGP1*, *MUCL1*, and *TOP2A*. Violin plots display gene expression on the selected tumor region across different cell types. C. The actual expression at the spot resolution and the predicted expression at the single-cell resolution by BayesDeep and TESLA on the selected immune region for genes *CORO1A*, *IGHG2*, and *TRBC1*. Violin plots display gene expression on the selected tumor region across different cell types.
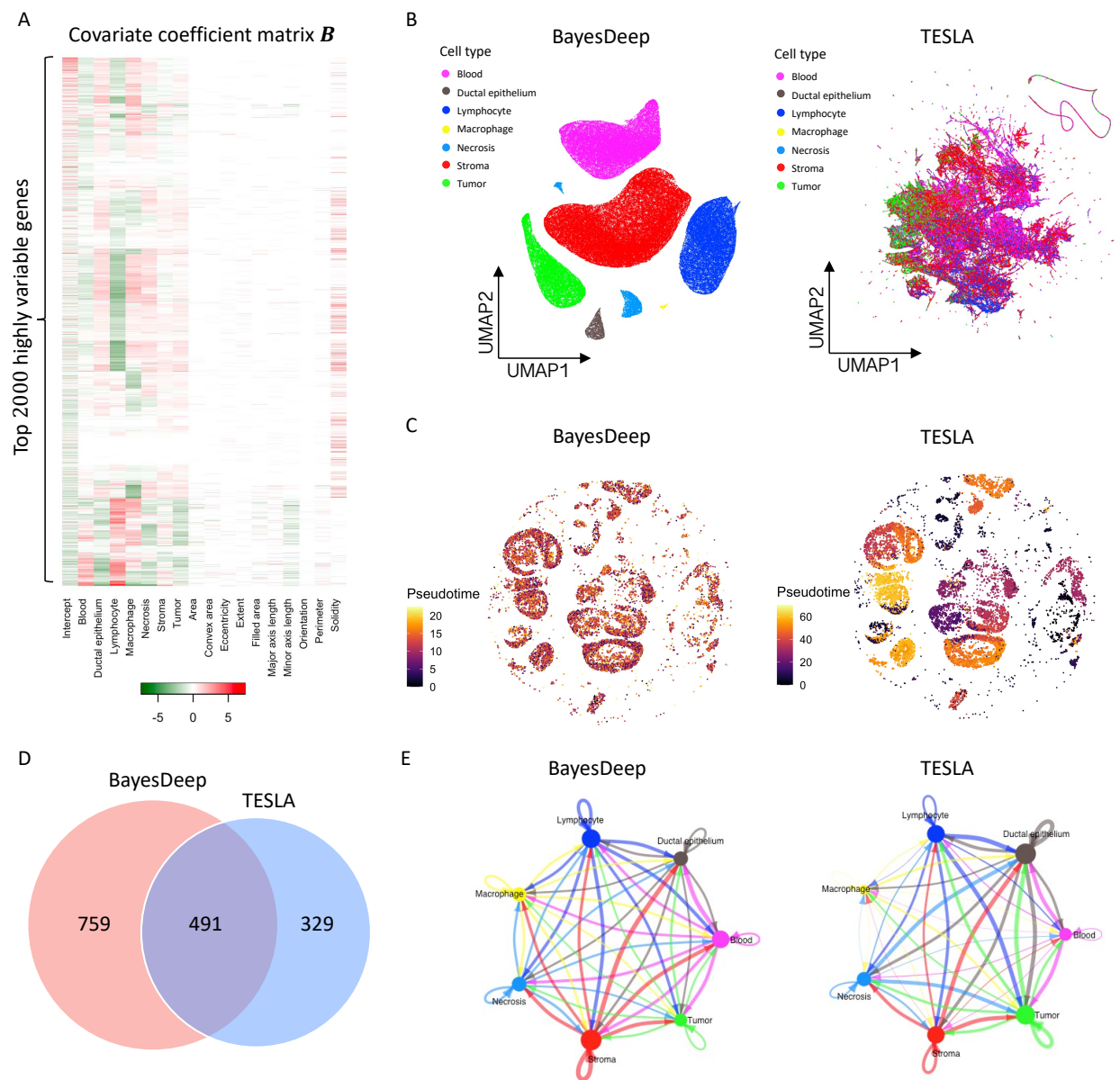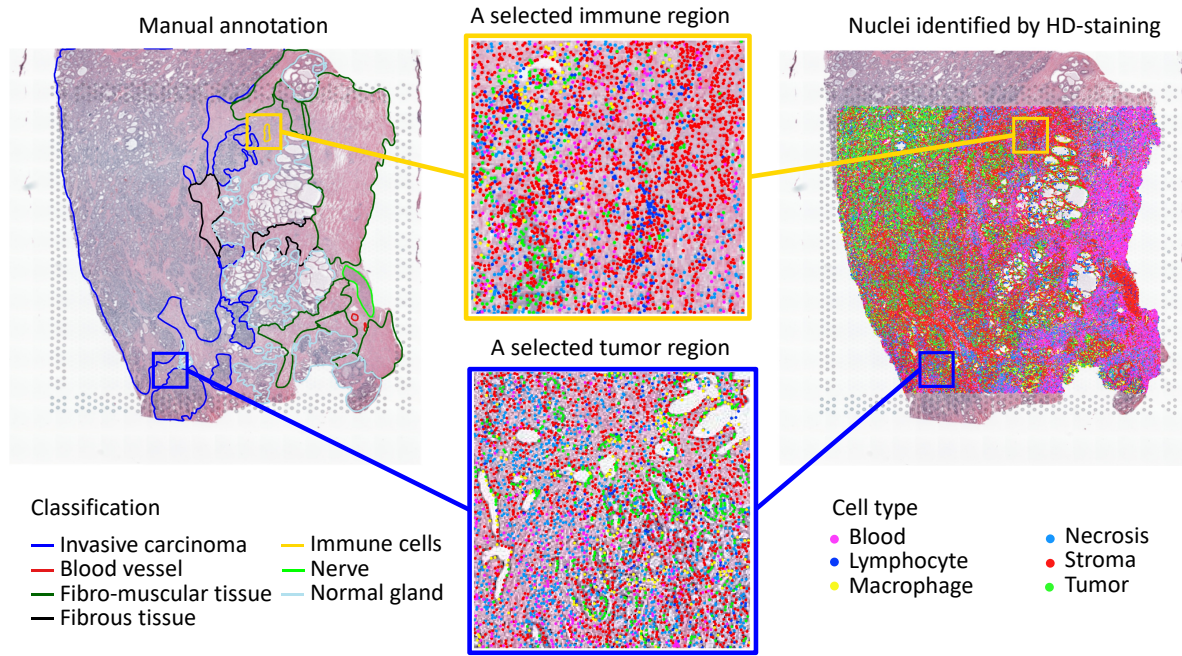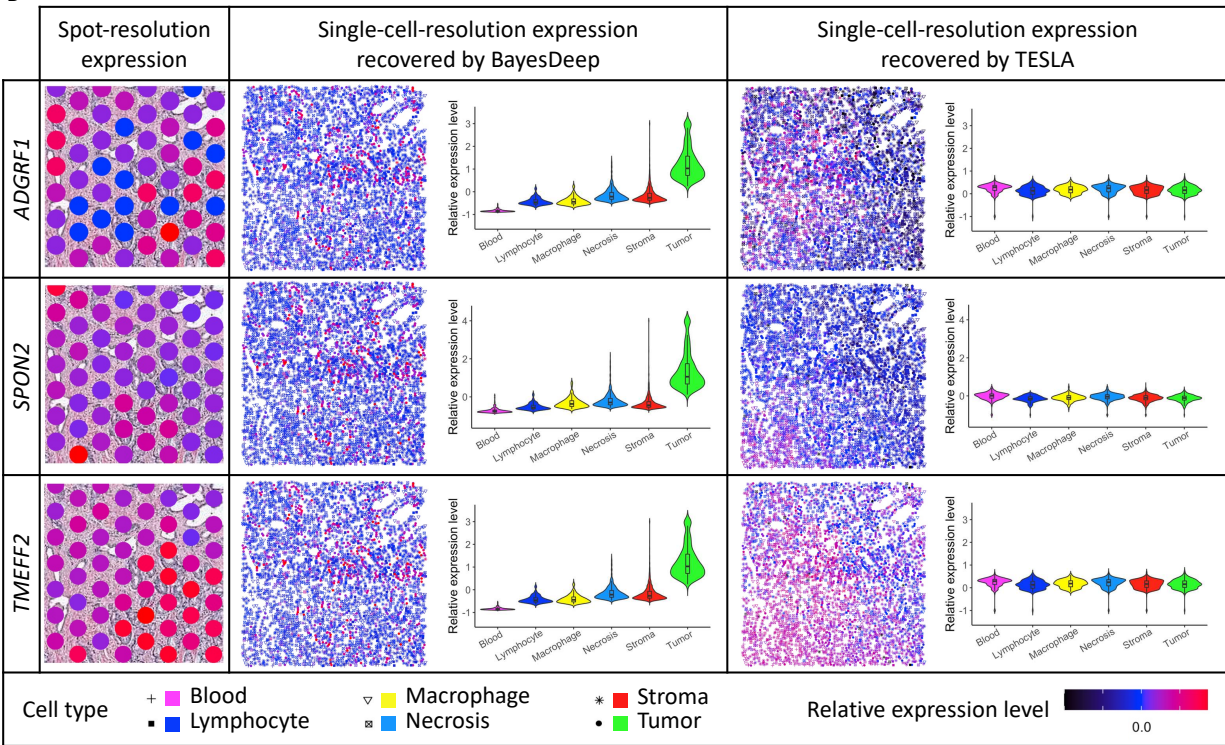
Figure 4. Real data analysis (continued) and downstream analysis on the human breast cancer 10x Visium data: A. Heatmap of the covariate coefficient matrix ***B*** estimated by BayesDeep, indicating the association between gene expression and morphological features extracted from the paired histology image. B. Cell population analysis on the BayesDeep- and TESLA-generated single-cell-resolution gene expression. C. Pseudotime analysis on tumor cells extracted from the cell

population analysis. D and E. Cell-cell interactions inferred from the BayesDeep- and TESLA-

generated single-cell-resolution gene expression, and their overlap.

A

Manual annotation

A selected immune region

Nuclei identified by HD-staining

A selected tumor region

Classification

— Invasive carcinoma
— Blood vessel
— Fibro-muscular tissue
— Fibrous tissue
— Immune cells
— Nerve
— Normal gland

Cell type
● Blood
● Lymphocyte
● Macrophage
● Necrosis
● Stroma
● Tumor

B

	Spot-resolution expression	Single-cell-resolution expression recovered by BayesDeep	Single-cell-resolution expression recovered by TESLA
ADGRF1			
SPON2			
TMEFF2			

Cell type
+ Blood
■ Lymphocyte
▽ Macrophage
⊠ Necrosis
✳ Stroma
● Tumor

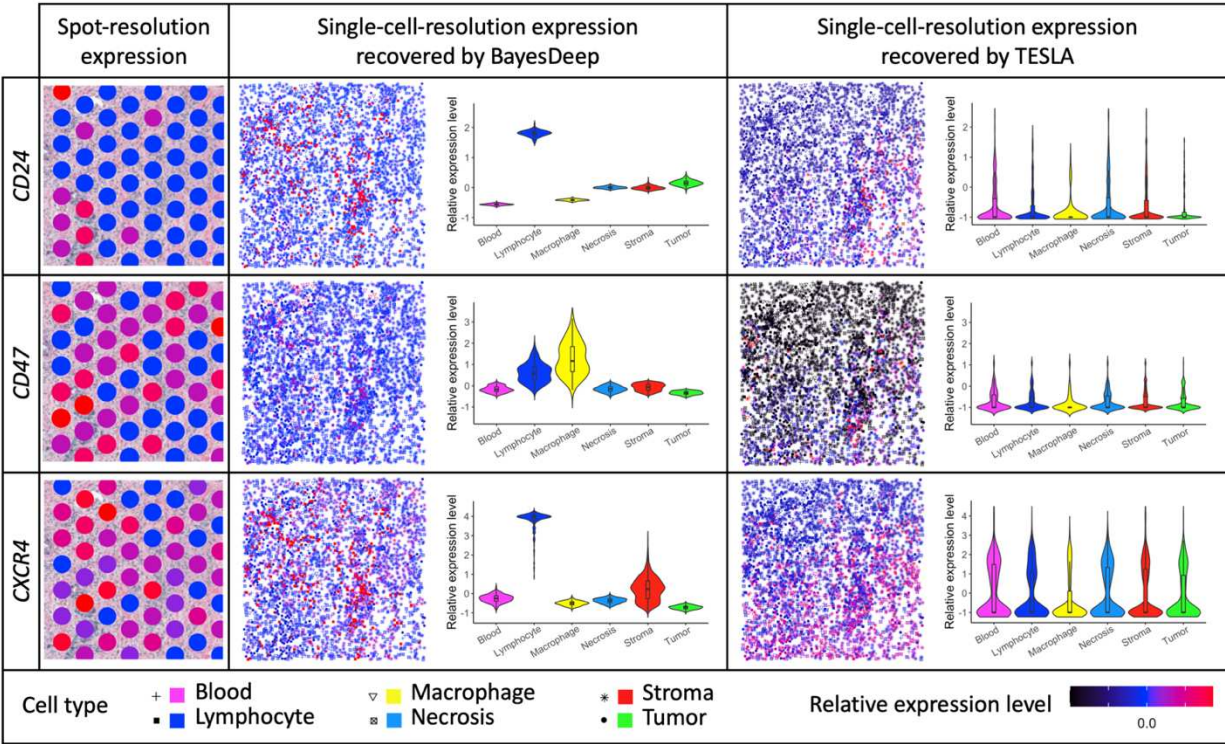Relative expression level

0.0

C

Figure 5. Real data analysis on the human prostate cancer 10x Visium data: A. Manual annotation and nuclei identification results by applying HD-staining to the paired histology image and the two selected tumor and immune regions for an illustrative comparison between BayesDeep and TESLA. B. The actual expression at the spot resolution and the predicted expression at the single-cell resolution by BayesDeep and TESLA on the selected tumor region for genes *ADGRF1*, *SPON2*, and *TMEFF2*. Violin plots display gene expression on the selected tumor region across different cell types. C. The actual expression at the spot resolution and the predicted expression at the single-cell resolution by BayesDeep and TESLA on the selected immune region for genes *CD24*, *CD47*, and *CXCR4*. Violin plots display gene expression on the selected tumor region across different cell types.
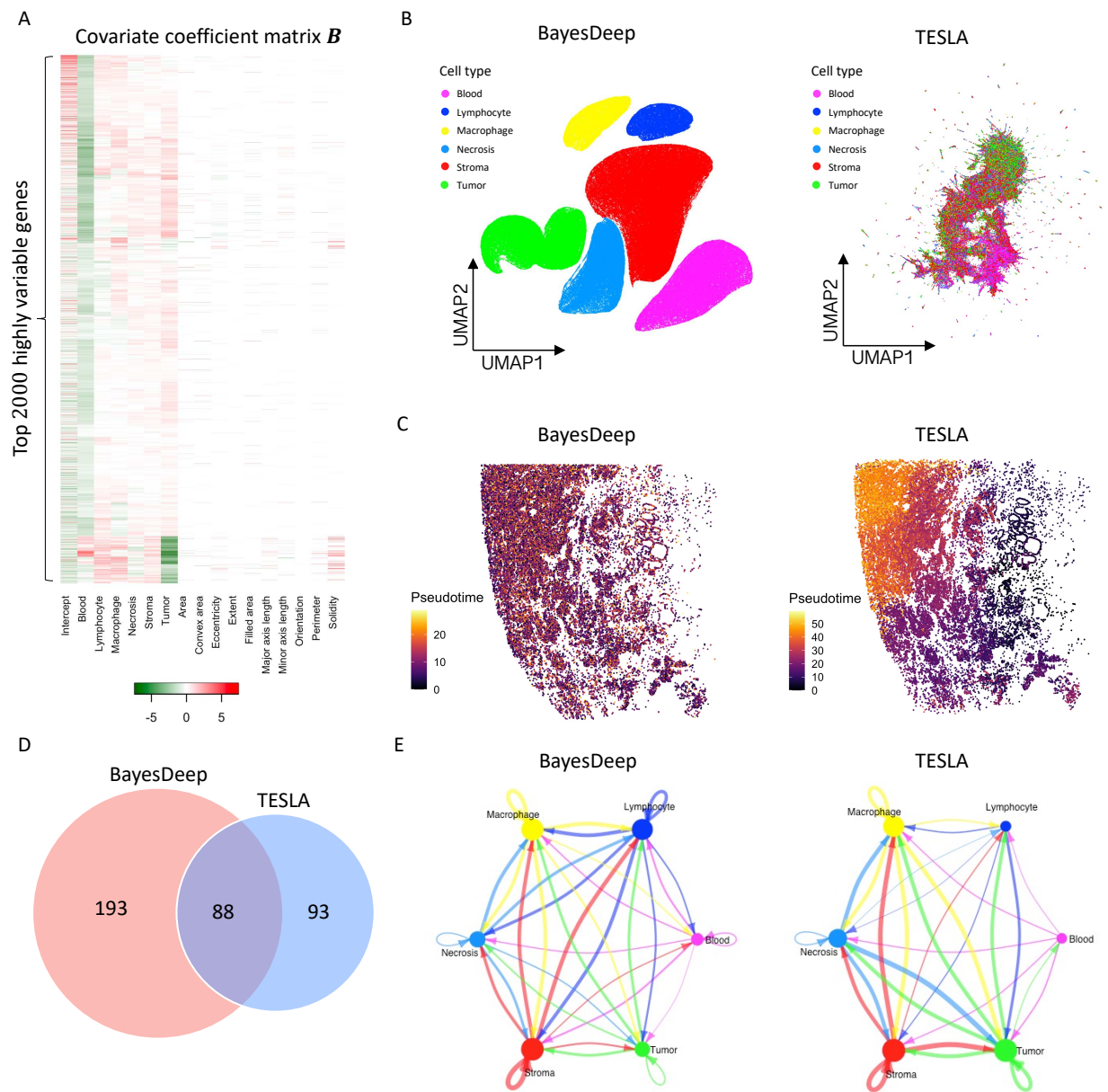
Figure 6. Real data analysis (continued) and downstream analysis on the human prostate cancer 10x Visium data: A. Heatmap of the covariate coefficient matrix ***B*** estimated by BayesDeep, indicating the association between gene expression and morphological features extracted from the paired histology image. B. Cell population analysis on the BayesDeep- and TESLA-generated single-cell-resolution gene expression. C. Pseudotime analysis on tumor cells extracted from the

cell population analysis. D and E. Cell-cell interactions inferred from the BayesDeep- and TESLA-

generated single-cell-resolution gene expression and their overlap.

References

1   Liao, J., Lu, X., Shao, X., Zhu, L. & Fan, X. Uncovering an organ's molecular architecture at single-cell resolution by spatially resolved transcriptomics. *Trends in biotechnology* **39**, 43-58 (2021).

2   Gruene, M. *et al.* Laser printing of three-dimensional multicellular arrays for studies of cell–cell and cell–environment interactions. *Tissue Engineering Part C: Methods* **17**, 973-982 (2011).

3   Ståhl, P. L. *et al.* Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science* **353**, 78-82, doi:10.1126/science.aaf2403 (2016).

4   10x Genomics. *Visium Spatial Gene Expression*, <https://www.10xgenomics.com/products/spatial-gene-expression> (2023).

5   Satija, R., Farrell, J. A., Gennert, D., Schier, A. F. & Regev, A. Spatial reconstruction of single-cell gene expression data. *Nat Biotechnol* **33**, 495-502, doi:10.1038/nbt.3192 (2015).

6   Lein, E., Borm, L. E. & Linnarsson, S. The promise of spatial transcriptomics for neuroscience in the era of molecular cell typing. *Science* **358**, 64-69, doi:10.1126/science.aan6827 (2017).

7   Zhang, L. *et al.* Clinical and translational values of spatial transcriptomics. *Signal Transduction and Targeted Therapy* **7**, 111 (2022).

8   Zhao, E. *et al.* Spatial transcriptomics at subspot resolution with BayesSpace. *Nat Biotechnol* **39**, 1375-1384, doi:10.1038/s41587-021-00935-2 (2021).

9   Hu, J. *et al.* Deciphering tumor ecosystems at super resolution from spatial transcriptomics with TESLA. *Cell systems* **14**, 404-417. e404 (2023).

10   Bergenstråhle, L. *et al.* Super-resolved spatial transcriptomics by deep data fusion. *BioRxiv*, 2020.2002. 2028.963413 (2020).

11   He, B. *et al.* Integrating spatial gene expression and breast tumour morphology via deep learning. *Nature biomedical engineering* **4**, 827-834 (2020).

12   Pang, M., Su, K. & Li, M. Leveraging information in spatial transcriptomics to predict super-resolution gene expression from histology images in tumors. *bioRxiv*, 2021.2011. 2028.470212 (2021).

13   Wang, S. *et al.* Computational Staining of Pathology Images to Study the Tumor Microenvironment in Lung Cancer. *Cancer Res* **80**, 2056-2066, doi:10.1158/0008-5472.CAN-19-1629 (2020).

14   Williams, C. K. & Rasmussen, C. E. *Gaussian processes for machine learning*. Vol. 2 (MIT press Cambridge, MA, 2006).

15   10x Genomics. *Human Breast Cancer: Ductal Carcinoma In Situ, Invasive Carcinoma (FFPE)*, <https://www.10xgenomics.com/resources/datasets/human-breast-cancer-ductal-carcinoma-in-situ-invasive-carcinoma-ffpe-1-standard-1-3-0> (2023).

16   10x Genomics. *Human Prostate Cancer, Adenocarcinoma with Invasive Carcinoma (FFPE)*, <https://www.10xgenomics.com/resources/datasets/human-prostate-cancer-adenocarcinoma-with-invasive-carcinoma-ffpe-1-standard-1-3-0> (2023).

17   Wang, X. *et al.* Three-dimensional intact-tissue sequencing of single-cell transcriptional states. *Science* **361**, doi:10.1126/science.aat5691 (2018).

18   Huang, C.-y. *et al.* Decreased expression of AZGP1 is associated with poor prognosis in primary gastric cancer. *PloS one* **8**, e69155 (2013).

19   Skliris, G. *et al.* Expression of small breast epithelial mucin (SBEM) protein in tissue microarrays (TMAs) of primary invasive breast cancers. *Histopathology* **52**, 355-369 (2008).

20   Valladares-Ayerbes, M. *et al.* Diagnostic accuracy of small breast epithelial mucin mRNA as a marker for bone marrow micrometastasis in breast cancer: a pilot study. *Journal of cancer research and clinical oncology* **135**, 1185-1195 (2009).

21   An, X. *et al.* The prognostic significance of topoisomerase II alpha protein in early stage luminal breast cancer. *BMC cancer* **18**, 1-10 (2018).

22   Yang, B. *et al.* An assessment of prognostic immunity markers in breast cancer. *NPJ Breast Cancer* **4**, 35 (2018).

23   De Jong, B. G. *et al.* Human IgG2‑and IgG4‑expressing memory B cells display enhanced molecular and phenotypic signs of maturity and accumulate with age. *Immunology and cell biology* **95**, 744-752 (2017).

24   Chen, M. *et al.* Analysis of the Expression of the TRBC1 in T lymphocyte tumors. *Indian Journal of Hematology and Blood Transfusion* **37**, 271-279 (2021).

25   Horna, P., Shi, M., Olteanu, H. & Johansson, U. Emerging role of t-cell receptor constant β chain-1 (TRBC1) expression in the flow cytometric diagnosis of t-cell malignancies. *International journal of molecular sciences* **22**, 1817 (2021).

26   Trapnell, C. *et al.* The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nature biotechnology* **32**, 381-386 (2014).

27   Qiu, X. *et al.* Reversed graph embedding resolves complex single-cell trajectories. *Nature Methods* **14**, 979-982 (2017).

28   Murar, M. & Vaidya, A. Cancer stem cell markers: premises and prospects. *Biomarkers in Medicine* **9**, 1331-1342 (2015).

29   Hadjimichael, C. *et al.* Common stemness regulators of embryonic and cancer stem cells. *World journal of stem cells* **7**, 1150 (2015).

30   Zhao, W., Li, Y. & Zhang, X. Stemness-related markers in cancer. *Cancer translational medicine* **3**, 87 (2017).

31   Matusiak, M. *et al.* A spatial map of human macrophage niches reveals context-dependent macrophage functions in colon and breast cancer. *Research Square* (2023).

32   Jin, S. *et al.* Inference and analysis of cell-cell communication using CellChat. *Nature communications* **12**, 1088 (2021).

33   Abdulkareem, N. M. *et al.* A novel role of ADGRF1 (GPR110) in promoting cellular quiescence and chemoresistance in human epidermal growth factor receptor 2‐positive breast cancer. *The FASEB Journal* **35**, e21719 (2021).

34   Gad, A. A. & Balenga, N. The emerging role of adhesion GPCRs in cancer. *ACS pharmacology & translational science* **3**, 29-42 (2020).

35   Huang, C. *et al.* Tumor cell-derived SPON2 promotes M2-polarized tumor-associated macrophage infiltration and cancer progression by activating PYK2 in CRC. *Journal of Experimental & Clinical Cancer Research* **40**, 1-17 (2021).

36   Gery, S., Sawyers, C. L., Agus, D. B., Said, J. W. & Koeffler, H. P. TMEFF2 is an androgen-regulated gene exhibiting antiproliferative effects in prostate cancer cells. *Oncogene* **21**, 4739-4746 (2002).

37   Li, O., Zheng, P. & Liu, Y. CD24 expression on T cells is required for optimal T cell proliferation in lymphopenic host. *The Journal of experimental medicine* **200**, 1083-1089 (2004).

38   Komori, S. *et al.* CD47 promotes peripheral T cell survival by preventing dendritic cell– mediated T cell necroptosis. *Proceedings of the National Academy of Sciences* **120**, e2304943120 (2023).

39   Caruz, A. *et al.* Genomic organization and promoter characterization of human CXCR4 gene 1. *FEBS letters* **426**, 271-278 (1998).

40   Jovic, D. *et al.* Single‐cell RNA sequencing technologies and applications: A brief overview. *Clinical and Translational Medicine* **12**, e694 (2022).

41   Ji, Z. & Ji, H. TSCAN: Pseudo-time reconstruction and evaluation in single-cell RNA-seq analysis. *Nucleic acids research* **44**, e117-e117 (2016).

42   He, K., Gkioxari, G., Dollár, P. & Girshick, R. Mask R-CNN. *In Proceedings of the IEEE international conference on computer vision.* 2961-2969.

43   Fernández, E. *et al.* SAFARI: shape analysis for AI-segmented images. *BMC Medical Imaging* **22**, 1-7 (2022).

44   Li, Q., Zhang, M., Xie, Y. & Xiao, G. Bayesian Modeling of Spatial Molecular Profiling Data via Gaussian Process. *Bioinformatics*, doi:10.1093/bioinformatics/btab455 (2021).

45   Jiang, S. *et al.* A Bayesian zero-inflated negative binomial regression model for the integrative analysis of microbiome data. *Biostatistics* **22**, 522-540 (2021).

46   Li, Q., Cassese, A., Guindani, M. & Vannucci, M. Bayesian negative binomial mixture regression models for the analysis of sequence count and methylation data. *Biometrics* **75**, 183-192 (2019).

47   Zhang, M., Li, Q. & Xie, Y. A Bayesian hierarchical model for analyzing methylated RNA immunoprecipitation sequencing data. *Quantitative Biology* **6**, 275-286 (2018).

48   Cameron, A. C. & Trivedi, P. K. *Regression analysis of count data*. Vol. 53 (Cambridge university press, 2013).

49   Witten, D. M. Classification and clustering of sequencing data using a Poisson model. (2011).

50   Li, J., Witten, D. M., Johnstone, I. M. & Tibshirani, R. Normalization, testing, and false discovery rate estimation for RNA-sequencing data. *Biostatistics* **13**, 523-538 (2012).

51   Banerjee, S., Carlin, B. P. & Gelfand, A. E. *Hierarchical modeling and analysis for spatial data*. (Chapman and Hall/CRC, 2003).

52   Airoldi, E. M. & Bischof, J. M. Improving and evaluating topic models and other models of text. *Journal of the American Statistical Association* **111**, 1381-1403 (2016).

53   Lu, J. *et al.* scRNA‐seq data analysis method to improve analysis performance. *IET nanobiotechnology* (2023).

54   Sun, S., Zhu, J. & Zhou, X. Statistical analysis of spatial expression patterns for spatially resolved transcriptomic studies. *Nat Methods* **17**, 193-200, doi:10.1038/s41592-019-0701-7 (2020).

55   Bullard, J. H., Purdom, E., Hansen, K. D. & Dudoit, S. Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC bioinformatics* **11**, 1-13 (2010).

56   Li, Q., Guindani, M., Reich, B. J., Bondell, H. D. & Vannucci, M. A Bayesian mixture model for clustering and selection of feature occurrence rates under mean constraints. *Statistical Analysis and Data Mining: The ASA Data Science Journal* **10**, 393-409 (2017).

57   Cable, D. M. *et al.* Cell type-specific inference of differential expression in spatial transcriptomics. *Nature Methods* **19**, 1076-1087 (2022).

58   Cable, D. M. *et al.* Robust decomposition of cell type mixtures in spatial transcriptomics. *Nat Biotechnol*, doi:10.1038/s41587-021-00830-w (2021).

59   Ma, Y. & Zhou, X. Spatially informed cell-type deconvolution for spatial transcriptomics. *Nature biotechnology* **40**, 1349-1359 (2022).

60   Jarque, C. M. Efficient grouping of observations in regression analysis. *International Economic Review*, 709-718 (1981).

61   Prais, S. J. & Aitchison, J. The grouping of observations in regression analysis. *Revue de l'Institut International de Statistique*, 1-22 (1954).

62   Silver, N., Best, S., Jiang, J. & Thein, S. L. Selection of housekeeping genes for gene expression studies in human reticulocytes using real-time PCR. *BMC molecular biology* **7**, 1-9 (2006).

63   Bai, R., Rockova, V. & George, E. I. Spike-and-slab meets LASSO: A review of the spike-and-slab LASSO. *arXiv preprint arXiv:2010.06451* (2020).

64   Brown, P. J., Vannucci, M. & Fearn, T. Multivariate Bayesian variable selection and prediction. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **60**, 627-641 (1998).

65   George, E. I. & McCulloch, R. E. Approaches for Bayesian variable selection. *Statistica sinica*, 339-373 (1997).

66   Newton, M. A., Noueiry, A., Sarkar, D. & Ahlquist, P. Detecting differential gene expression with a semiparametric hierarchical mixture method. *Biostatistics* **5**, 155-176 (2004).